MASARYK UNIVERSITY
FACULTY OF SCIENCES
DEPARTMENT OF BOTANY AND ZOOLOGY

Habilitation thesis

BRNO 2018 **NATÁLIA MARTÍNKOVÁ**

MASARYK UNIVERSITY
FACULTY OF SCIENCES
DEPARTMENT OF BOTANY AND ZOOLOGY

Modelling in phylogenetic framework

Habilitation thesis

Natália Martínková

Brno 2018

Bibliographic Entry

Author:	Mgr. Natália Martínková, Ph.D. Faculty of Science, Masaryk University Department of Botany and Zoology

Title of Thesis:	Modelling in phylogenetic framework
Field of Study:	Zoology
Academic Year:	2018/19
Number of Pages:	x + 298
Keywords:	Phylogenetic analysis; Maximum likelihood; Bayesian inference; Molecular dating; Codivergence; Comparative phylogenetics; Outlier analysis

Acknowledgement

Collaboration provides inspiration, drive, and increases achievements in scientific research. This thesis contains research that would have been impossible without collaboration. I thank all colleagues who worked with me in addressing numerous research questions. Discussions with Jan Zima, Maarit Jaarola, Jeremy B. Searle, Gerald Heckel, Jiří C. Moravec, Kamil S. Jaroň and Jiří Pikula have been very formative during progressing stages of my career. My friends Alexandra Zahradníková, Jr., Danka Haruštiaková, Nancy R. Irwin, Mabel Giménez and Pavel Škrabánek enriched not only my personal but also my professional life. I thank Ladislav Dušek for his patience and support during my pedagogical career.

Brno 9 Oct, 2018 .....

Natália Martínková

Contents

List of mathematical symbols

$\mathbf{A}$	matrix of associations between organisms
$\mathbf{b}$	vector of pairwise phylogenetic distances
$\mathbf{b}'$	vector of pairwise phylogenetic distances of species that share a binary trait
$\mathbf{B}$	phylogenetic distance matrix
C	author contribution
$\mathbf{c}$	vector of the smallest phylogenetic distances for each species
$\mathbf{c}'$	vector of the smallest phylogenetic distances of species that share a binary trait
$\mathbf{C}$	phylogenetic distance matrix
$\mathbf{D}$	fourth-corner matrix
DIAS	discrete interval accumulative score in the SigHunt analysis
f	frequency
g	generation
k	number of categories partitioning research for a publication
K	phylogenetic signal statistic
i	index
M	substitution model
$\mathbf{M}$	matrix
n	number of units
N	nearest neighbour interchange
NRI	net relatedness index
NTI	nearest taxon index
p	probability
P	population
ParaFitGlobal	global statistic of phylogeny codivergence in the ParaFit analysis
R_i	list of authors contributing to the i-th category
s	selection coefficient
S	sequence data
T	phylogenetic tree

T	matrix transposition
W	global stastistic of matrix congruence from the CADM test
α	significance level or parameter of the Γ distribution
β	parameter of the Γ distribution
B	bootstrap sampling
$\boldsymbol{\varepsilon}$	vector of differences between the vector of trait values and the phylogenetic mean of the trait value at the root of the tree
$\hat{\boldsymbol{\varepsilon}}$	vector of trait values corrected for the variance-covariance of the tree using the generalized least-squares
λ	parameter of the exponential distribution
μ	mean
ρ	rate of evolution
σ	standard deviation
ϕ	function evaluating contribution of a specific author

Structure and goals of the habilitation thesis

This thesis includes 17 publications in journals with an impact factor and one reviewed conference paper. Included publications represent a selection of my work that is closely related to the topic of statistical modelling in zoology. The publications are organized into five topics, with each topic corresponding to a goal of the thesis.

1. Testing hypotheses using phylogenetic inference from genetic data to evaluate taxon divergence and evolutionary history.
2. Molecular dating to model the time line of phylogenetic divergence at intra- and inter-specific levels.
3. Testing phylogenetic congruence in search of functional relationships between groups of taxa.
4. Predictive modelling of species traits in the context of genetic diversity.
5. Identifying biologically relevant information in data structure with detection of outliers.

The habilitation thesis contains an Introduction that provides a wider knowledge base for application of the methods and processes in my collated research. The Introduction updates the current information in individual topics with respect to the accumulated progress in research since the time of the publication. In some topics, the new advances explain the justification for the practical applications of the analytical tools in my publication. I demonstrate the concepts using diagnostic graphs, and I use simulation models to present a streamlined picture of the described effects.

The diagnostic graphs originate directly from the archived analyses files, the results of which are included in the respective publications. They explain and justify the applied analytical procedure. The purpose of the diagnostic graphs is to guide a detailed commentary into the problem of data analyses in phylogenetic framework and to critically evaluate the obtained results.

Validating any result is a common practice in experimental biology, and the practice needs to be applied to the data analysis as well. Simulations help to provide data for testing the effect the analytical approach should reveal, and thus validate the results and help in their interpretation. Statistical methods return a number if the user manages to input data in the proper format. However, the more advanced methods challenge the user intuition in

spotting incorrect results. A method might report a result with precision to several decimal places that is several orders of magnitude incorrect. The diagnostics of the analysis run in conjunction with simulations of the expected process enable the user to better differentiate the correct results from analysis artefacts.

Chapter 1

Introduction

Genetic databases contain data from varied types of studies, where the respective authors used DNA sequences and related information to test their specific hypothesis. As a result, the databases accumulate information on model organisms and other taxa that were considered relevant with respect to the multitude of different hypotheses in the scientific endeavours of the last decades. Given time, the genetic databases accumulate DNA sequences with great diversity both in terms of sampled taxa and genes. The amount of the available genetic information enables the researchers to ask new questions that were unfeasible at the time of the original research that generated the data. Using publically available data is becoming routine not only as an addition to the original data, but datasets from the public domain can now fully support new research questions on their own. Ease in obtaining data increases demand for creativity in designing novel hypotheses and inspires the ability to quickly and cheaply test the hypothesis in a pilot study.

The keystone hypotheses stemming from genetic information relate to differentiation of biological diversity in form of reconstruction of phylogenetic relationships. Phylogenetic relationships provide information on the succession of divergences and relatedness of taxon groups. With help from additional information, a phylogeny can be transformed to reflect time and rate of speciation events, providing a biological timeline corresponding to geological or climate changes. In interactions between organisms, assessing codivergence between phylogenetic trees addresses the character of associations between organisms through time. A phylogeny facilitates comparing species traits, because similarity between species is affected by their shared evolutionary history. To discover rare mechanisms that may influence biological variability, the outlier analysis indicates directions for research focus. A phylogeny provides a foundation for studying processes affecting biodiversity in the web of interactions between organisms through time and space.

1.1 Testing hypotheses using phylogenetic inference from genetic data

Phylogenetic information in DNA sequences can be extracted using statistical methods based on multiple sequence alignment or on alignment-free methods. Alignment-free methods provide undisputable advantages in allowing the analysis without additional information on gene content or location and appear to be ideal tools for the genomic era. However, methods based on oligonucleotide composition, complexity measures or metrics derived from information theory perform worse than statistical methods that use multiple sequence alignment (Bogusz & Whelan, 2017; Chan, Bernard, Poirion, Hogan, & Ragan, 2014; Philippe et al., 2011). To construct a multiple sequence alignment means to find positions in DNA (or amino-acid) sequences that share evolutionary history. An alignment is then a matrix, where the rows represent individual sequences and columns correspond to homologous positions in DNA sequences. When we construct an alignment from sequences of multiple genes, the homology should be estimated per gene. In practice, we first align gene sequences and then concatenate the individual alignments into a chimeric supermatrix (Martínková & Moravec, 2012; Pečnerová & Martínková, 2012; Pečnerová, Moravec, & Martínková, 2015).

Unlike numeric matrices, the cells in an alignment contain nucleotide bases. A transition from one value in a matrix cell to another is a qualitative change reflecting a substitution of one nucleotide for another. A challenge to quantify a change from one qualitative value to the next requires a modelling approach. In phylogenetics, the substitution model quantifies the probability of a specific substitution. For example, when a cytosine mutates to a guanine, the probability of the change depends primarily on the frequency of guanines in the sequence and the rate of cytosine to guanine mutations. The substitution models can then be used to calculate genetic distances among sequences or to reconstruct their phylogenetic relationships.

Current trends strongly prefer methods of phylogenetic reconstruction based on maximum likelihood (ML) or Bayesian inference (BI) due to the reliability of the approaches in estimating the correct phylogenetic relationships under variable conditions. Other methods common in the past, such as parsimony or neighbour-joining clustering, do not scale well with increasing sequence length and number of taxa in the dataset (Bogusz & Whelan, 2017). Although the ML and BI methods are also challenged with increasing dataset size, they are sufficiently robust and allow flexibility in algorithm design that facilitated studies of thousands of taxa with thousands of genes (Dobrin, Zwickl, & Sanderson, 2018).

Maximum likelihood searches the tree space using heuristics for a tree that will have the highest probability p of observing the sequence data S under the condition of the given phylogenetic tree T and the applied substitution model M. We call this conditional probability $p(S|T,M)$ a likelihood. A tree with the highest likelihood will best reflect the evolutionary relationships between the sequences in the dataset.

Phylogenetic reconstruction using ML is computationally expensive in the attempts to obtain node support using a bootstrap analysis. One bootstrap replicate imitates the analysis of the full phylogenetic reconstruction, and thus the bootstrap analysis has linear time complexity. First, the bootstrap analysis creates a pseudoreplicate of the DNA sequence alignment, in which the columns are sampled from the original alignment with repetitions

up to the total alignment length. The pseudoreplicate then contains identical traits (columns of the alignment) but in different frequencies than those in the original data. In the bootstrap analysis, each pseudoreplicate is subsequently analysed with ML and the analysis with 100 bootstrap replicates will take 100 times as long as the analysis with the original alignment. The computational price lead to decrease in utility of the ML analysis with the increase in available sequence data. Nowadays, the ML analysis has rejuvenated owing to increased speed in bootstrap analysis (Stamatakis, 2014) and development of alternative measures for node support estimation (Anisimova & Gascuel, 2006).

Bayesian inference evaluates the probability that we will observe a specific phylogeny and its substitution model given the data in the alignment $p(T,M|S)$. In BI, we need to collate information about the model and specify its prior distribution before the analysis. The advantage of the BI is that after the data are collected, the BI directly provides statistical support for the hypothesis in form of the posterior probability (Nascimento, dos Reis, & Yang, 2017). The posterior probability is the probability that the hypothesis (tree and model) is correct given the data (alignment).

The BI uses Markov chain Monte Carlo (MCMC) to search the tree space. We can imagine the tree space as a multidimensional space with all trees that are possible for a given number of taxa (Fig. 1.1) (Hillis, Heath, & St John, 2005). The MCMC searches the tree space with respect to the priors defined by the user by slightly modifying the tree and evaluating whether to accept or reject the tree modification. Each tree and model modification will be drawn from a prior for the substitution model parameters, tree topology and branch lengths. The user chooses a substitution model and the distributions from which the model parameters will be drawn. The tree topology can be influenced by priors that assign a taxon or a group of monophyletic taxa as an outgroup, thereby specifying the root of the ingroup and restricting the allowed topology changes during MCMC sampling. Alternatively, monophyly of a group deep in the tree can be assigned as a prior, reducing the tree space that will be searched by the MCMC. The last category of priors in BI is the prior for branch lengths.

Length of branches in a phylogenetic tree is a linear combination of mutation rate and time. Short branches represent evolutionary scenarios when the divergence occurred fast with respect to the mutation rate of the respective locus. In BI, the mutation rate can freely vary across the tree and the branch lengths are drawn randomly from a prior distribution. Program MrBayes (Ronquist et al., 2012) uses an exponential distribution for the branch length prior with the default value for its parameter $\lambda = 10$ (Fig. 1.2, red line). The default prior is called uninformative, because it has similar probability density for a relatively wide interval of possible branch lengths. In theory, an uninformative prior should have little influence on the posterior and the posterior should be data-driven (Nelson, Andersen, & Brown, 2015). In practice, the priors could in some cases override information in the data in sampling the posterior.

The distribution of the posterior sample of branch lengths may mimic the prior distribution in situations when the data are not informative (Fig. 1.3). In case of phylogeny of voles from the subfamily Arvicolinae (Rodentia: Cricetidae) (Martínková & Moravec, 2012), the data were uninformative probably because of large amount of missing sequences in the multilocus alignment. In several loci, relatively few species had homologous DNA sequences resulting in more than 70% of missing data in the alignment. Matrices of aligned

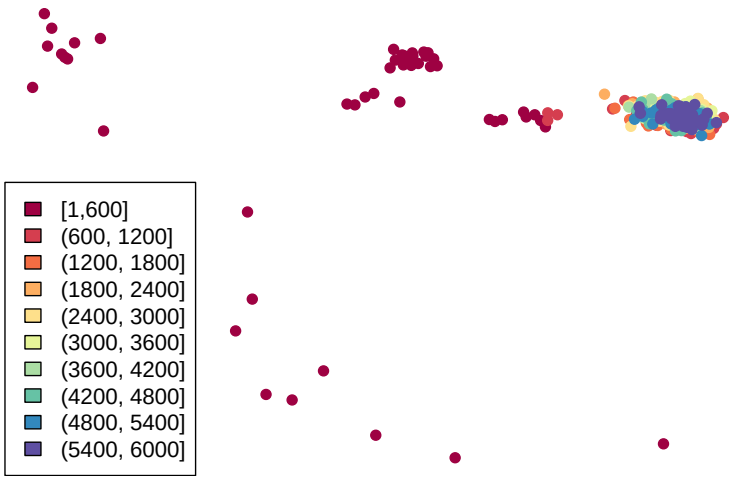


Figure 1.1: **Markov chain Monte Carlo sampling of the tree space.** Phylogenetic trees occupy a multidmensional tree space, where their distances correspond to differences in tree topology and branch lengths (Hillis, Heath, & St John, 2005). The tree space here is reduced to two dimensions with multidimensional scaling for visualisation. Dots show MCMC sampling of the tree space, with colours representing the progression of the algorithm through the saved trees.

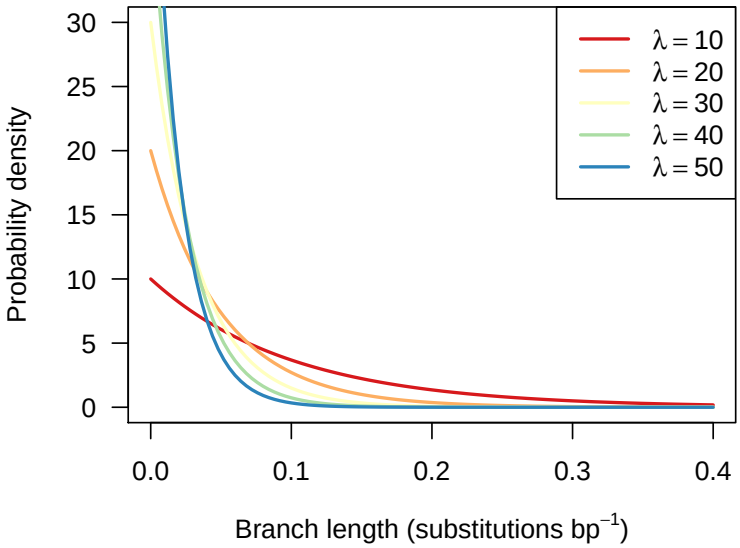


Figure 1.2: **Probability density of the exponential distribution.** Exponential distribution is used as a prior for branch lengths in Bayesian inference with default value of parameter $\lambda = 10$. By increasing the parameter value, each generation of Markov chain Monte Carlo modifies the phylogeny in Bayesian inference by drawing low values for branch lengths with higher probabilities.

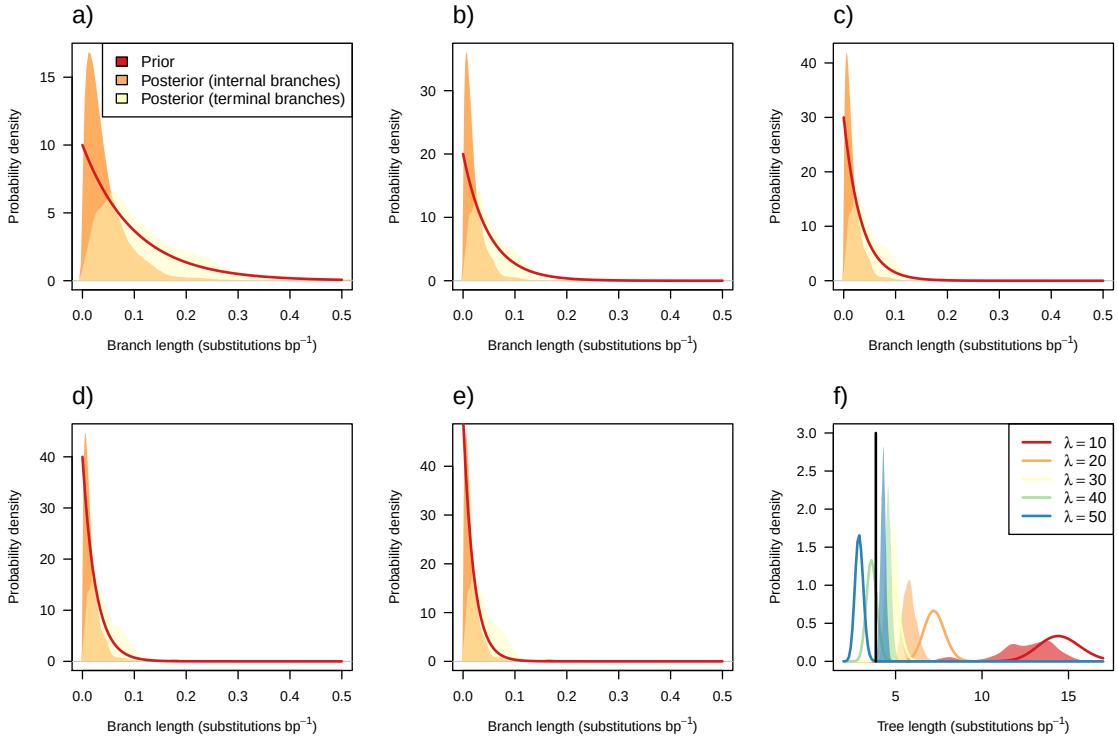


Figure 1.3: **Influence of priors on the posterior in Bayesian inference.** Line - probability density of the prior distribution, polygon - probability density of the posterior distribution estimated from real data (Martínková & Moravec, 2012). Branch lengths are displayed separately for internal branches and for terminal branches leading to tree tips. a) Prior for branch lengths is given as an exponential distribution for $\lambda = 10$, b) $\lambda = 20$, c) $\lambda = 30$, d) $\lambda = 40$, e) $\lambda = 50$. f) Prior for tree length is not explicitly defined, but it is implicitly determined by the prior for the branch lengths as a Γ distribution with parameters $\alpha = 2n - 3$ and $\beta = \lambda$, where n is the number of sequences in the dataset. Vertical black line - maximum likelihood estimate of the tree length.

DNA sequences that contain missing data should not be confused with sparse matrices, characterised by high frequency of zeroes. The large amount of missing data coupled with their non-random distribution across the alignment matrix may produce inconsistent results of phylogenetic reconstruction (Xi, Liu, & Davis, 2016). Reconstruction of phylogenies from multilocus sequence data requires an efficient computational algorithm due to high number of traits that are included in the analysis, and potential problems when the result may contain statistically supported incorrect nodes. To analyse relationships between arvicoline voles from the alignment with large amount of missing data, we chose BI. Compared to ML, BI is more robust against missing data and the results contain fewer artefacts (Dwivedi & Gadagkar, 2009; Simmons, 2012), although the BI could become trapped in the tree space that contains unrealistically long trees (Ekman & Blaalid, 2011; Marshall, 2010). Despite expected challenges, including multiple loci in phylogenetic reconstruction improves estimation of relationships and better approximates true divergence processes and their progression (Martínková & Moravec, 2012).

My work in multilocus phylogenetics addressed the influence of branch length priors on tree topology and the overall length of the tree in BI. The tree length is a sum of all branch lengths in a phylogeny. If the posterior contains samples from a prior exponential distribution, the posterior will reflect a convolution of exponential densities of all branches (Nelson et al., 2015). A phylogeny contains $2n-3$ branches for n sequences. Convolution of $2n-3$ exponential densities with parameters $\lambda = 10$ is a Γ distribution with parameters $\alpha = 2n-3$ and $\beta = \lambda$. What follows is that there exists an implicit prior on the tree length in BI that is sensitive to the number of sequences in the alignment and the branch length prior (Nelson et al., 2015; Rannala, Zhu, & Yang, 2012). The branch length prior strongly influences the theoretical tree lengths sampled in the posterior (Fig. 1.3f). Using the example of the vole phylogeny (Martínková & Moravec, 2012), the implicit prior for the tree length included the ML estimate of the tree length with the highest probability for $\lambda \in \{30, 40\}$.

Except the tree length, the branch length prior influences also the tree topology and node support (Ekman & Blaalid, 2011; Kolaczkowski & Thornton, 2009; Martínková & Moravec, 2012; Nelson et al., 2015). We observed a paradoxical situation in BI of the vole phylogeny, when the node support decreased with increasing accuracy of the tree length estimation (Fig. 3 in paper 2.1.2; Martínková & Moravec, 2012). Selection of the branch length prior requires good justification, because the prior has an important influence on the final phylogeny. Whenever possible, an uninformative prior with the parameter $\lambda = 10$ should be used.

We have shown that the BI analysis of alignments with a large amount of missing data may result in unrealistically long phylogenies. An opposite situation can arise in an analysis of divergent sequences, where the dataset contains information on insertions and deletions, the indels (Dwivedi & Gadagkar, 2009; Kopecna et al., 2014; Wallace et al., 2012). Indels are inheritable traits that appear in sequences during DNA replication when the DNA polymerase slips along a DNA strand. As the offspring inherits the indels, they carry phylogenetic information that could be informative for phylogeny reconstruction. However, phylogenetic information in indels is lost in default uses of the phylogeny reconstruction algorithms, and the indels are treated as missing data. The reason for the omission is in the model of evolution that quantifies the qualitative change in the DNA sequence. Common substitution models used in the ML and BI methods assume nucleotide changes, and calculate the likelihood from DNA sequence nucleotide composition and rates of substitutions. To include the information in indels, the dataset needs to be partitioned with indels encoded as a binary morphological character. The evolution in the indels is then modelled with an F81 model modified for binary data (McGuire, Denham, & Balding, 2001).

In phylogenetic reconstruction of aquaporins (Wallace et al., 2012), we utilised phylogenetic information in indels and we obtained an unrealistically short tree given that the dataset contained both protostomes and deuterostomes. The tree length estimated from the BI was short compared to the ML estimate, but also with respect to the implicit prior on the tree length (Fig. 1.4). The reason why the BI tree was too short is probably the specifics of the model. The F81 model assumes an equal mutation rate at all sites. When the alignment contains sites that mutate at a higher or slower rate, the BI tends to systematically underestimate branch lengths (Nascimento et al., 2017). Different mutation rate at

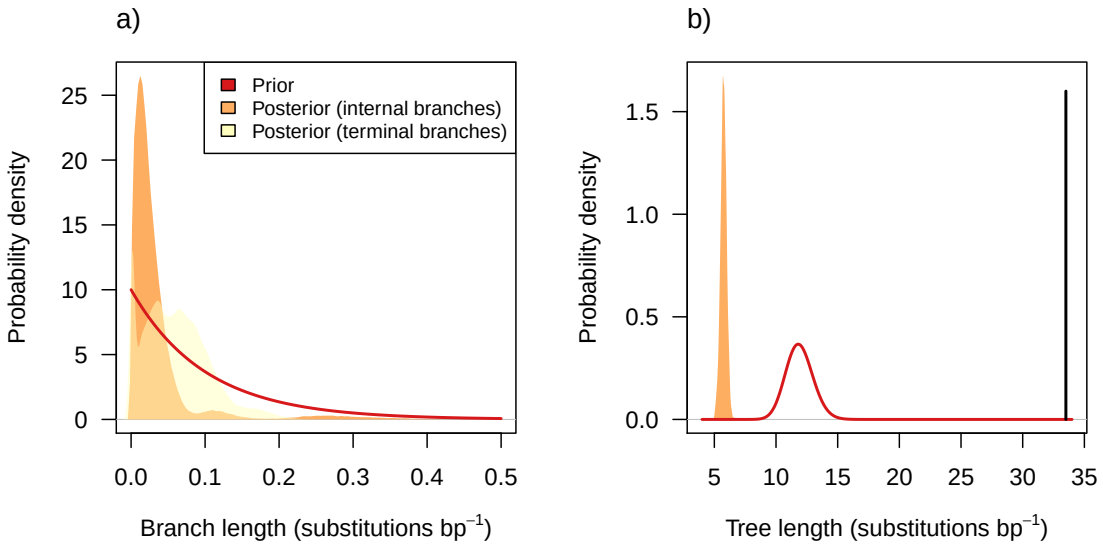


Figure 1.4: **Influence of a prior on the posterior in Bayesian inference with a mixed model including DNA sequences and binary encoded indels.** Line - probability density of the prior distribution, polygon - probability density of the posterior estimated from real data (Wallace et al., 2012). Branch lengths are displayed separately for internal branches and for terminal branches leading to tree tips. a) Prior for branch lengths is given as an exponential distribution for $\lambda = 10$, b) Prior for tree length is not explicitly defined, but it is implicitly determined by the prior for the branch lengths as a Γ distribution with parameters $\alpha = 2n - 3$ and $\beta = \lambda$, where n is the number of sequences in the dataset. Vertical black line - maximum likelihood estimate of the tree length.

different alignment positions is biologically obvious. The most straightforward mechanism generating different mutation rates for indels involves survival of carriers of mutations in coding versus non-coding genomic regions. In non-coding sequences, indels will persist, and thus appear to mutate, more often than in protein coding sequences, because insertions or deletions of nucleotide bases in coding regions might lead to a damaged gene product through frame shift or truncation of the coding sequence. Current implementations of the evolutionary models for indels ignore variable mutation rates across sites. The caveat on how to obtain a solution that can still utilise the robust BI is to estimate the tree topology with BI, and then to calculate the branch lengths in the ML framework on a fixed topology (Kopecna et al., 2014; Wallace et al., 2012).

We have demonstrated that multilocus alignments with missing data are informative in reconstructing phylogenetic relationships (Martínková & Moravec, 2012; Pečnerová & Martínková, 2012; Pečnerová et al., 2015). The amount of missing data present in the sequence alignments would be prohibitive in other statistical analyses and the datasets will need to be pruned or the missing data will be imputed to allow statistical evaluation. In phylogenetics, we work with the temporal information inherent to the biological systems and their continuity, and we are able to mine the information about taxon divergence also from incomplete data matrices. Missing data nevertheless lead to loss of diversifying sensitivity in additive elements of the tree likelihood calculations. In other words, likelihood

will be identical for all possible scenarios of species divergence if the clade topology is assessed from missing data (Sanderson, McMahon, & Steel, 2011). Trees with different topology that have identical likelihood are called a terrace. When the alignment includes complete data for all taxa and loci, a terrace will contain a single tree. That is; there will be a single phylogeny with the given likelihood. In alignments with missing data, small terraces can exist and they will ideally contain one tree (Dobrin et al., 2018; Martínková & Moravec, 2012; Sanderson, McMahon, Stamatakis, Zwickl, & Steel, 2015; Sanderson et al., 2011). The terrace size generally increases in datasets where the missing data affect multiple loci or taxa (Dobrin et al., 2018); more so if the missing data are distributed in the alignment non-randomly (Xi et al., 2016).

Occurrence of the terraces in the tree space based on alignments with missing data is influential in interpretation of the result. First, the reported topology estimated with ML or BI can belong to a terrace, meaning that there exist multiple trees that represent an equally correct result, have the same likelihood, as the reported tree. In extreme published cases, $10^{23} - 10^{388}$ alternative tree topologies reside on the terrace (Dobrin et al., 2018). The topological differences between trees from a terrace should be addressed in interpreting the phylogeny (Martínková & Moravec, 2012; Pečnerová et al., 2015).

Except for the tree topology, terraces influence node support. A monophyletic group can have high node support in bootstrap analysis in the ML framework if the clade is present in trees on a terrace (Sanderson et al., 2015). In BI, the terraces affect node support in a different way. While the trees on a terrace will have an identical likelihood, the priors influence their posterior probability. The BI that uses exponential prior for branch lengths (e.g. in Ronquist et al., 2012) will place a taxon with missing data as a sister to long branches in the tree with higher probability (Kolaczkowski & Thornton, 2009; Sanderson et al., 2015). Differences in posterior probability of trees from one terrace are given by a complex interplay of priors, branch lengths and structure of missing data on topology of the trees on the terrace (Sanderson et al., 2015).

There are no universal guidelines on how to address phylogenetic inference of large and complex datasets. Each specific application of the phylogeny reconstruction methods on concatenated data either from multiple loci or multiple traits requires sophisticated search for balance between accuracy and precision of the solution. Fortunately, recent advances in computational phylogenetics begin to suggest new algorithms that consider terraces. Heuristics that search the tree space will soon be modified in such a way that they will be robust against effects of terraces on phylogenetic reconstruction (Biczok et al., 2018; Chernomor, von Haeseler, & Minh, 2016).

1.2 Modelling the time line of phylogenetic divergence

In the last chapter, we addressed a problem of branch length estimation and the potential inconsistencies under different scenarios. Correct estimation of the tree topology but also branch lengths gains importance when we interpret the phylogeny in terms of evolutionary change through time. Since branch lengths reflect change in data, they approximate the biological phenomenon of mutation accumulation in the DNA sequences through time. We can extract the time component inherent to the phylogenetic reconstruction and estimate the divergence times by decomposing the branch lengths. If the branch lengths are a product of mutation rate and time, calibrating mutation rate will enable transformation of the branch lengths to time. We speak about molecular clocks (Kumar, 2005; Zuckerkandl & Pauling, 1962). Molecular dating analyses transform the node heights to time units and we will be able to place the divergence events into relation to the history of Earth. The potential of knowing when evolutionary scenarios happened is tantalising.

High variability between biological entities and processes complicates the elegant molecular clock hypothesis. Mutation rate varies between different organisms, different loci and when we observe mutation rate through time, it changes even within one lineage (dos Reis, Donoghue, & Yang, 2016; Ho, Phillips, Cooper, & Drummond, 2005). The differences in estimates of the mutation rate can be small with a small effect on the interpretation of the results. For example, alternative molecular dating analyses can differentiate whether Phoenicians or Vikings transported house mice along the European coasts (Gabriel, Jóhannesdóttir, Jones, & Searle, 2010). Unfortunately, the application of the molecular clock hypothesis and the time estimation from molecular data can seduce even a single team into vastly different mutation rate estimates and skew the interpretations of such results. As an example, two very close biological events were dated to times that differed by more than five orders of magnitude using overlapping datasets within a year of scientific advances. Specifically, the split of a pathogen from its sister taxon was dated to tens of millions years ago (Palmer, Drees, Foster, & Lindner, 2018), whereas the same pathogen intra-specific divergence was dated to thousands years ago (Drees et al., 2017). Molecular dating is one of the most used and misused analyses in molecular genetics. Availability of clickable software in combination with mathematical complexity of the underlying statistics can lead and mislead biologists through parameter distributions, hyperparameter distributions and transition kernels to sensational discoveries. Ascertaining that the molecular dating discovery is not only sensational but also correct requires additional information about the biological system that the analyst suitably incorporates into the analyses (cf. Aghova et al., 2018).

The differences in mutation rates between evolutionary lineages will be notable on a phylogeny as long branches that exceed the average root-to-tip distances. (In chapter 1.1, we addressed the shortening or elongation of the whole tree, the tree length. Here, we discuss length change in individual branches.) Unless the long branches are a sampling or analysis artefact (Kolaczkowski & Thornton, 2009), they have a biological meaning. In some organisms, such as viruses, an elevated mutation rate might be characteristic for the whole clade (Smith et al., 2009). In others, long braches may indicate fast evolutionary processes that might be adaptive (Fink, Excoffier, & Heckel, 2007).

Loci that encode proteins are limited in where the mutations can persist. In effect, we

observe that the protein-coding loci mutate in a way that is compatible with survival. That does not mean that mutations do not occur elsewhere in the gene. It means that carriers of such mutations did not survive long enough to be sampled. Only a mutation that enables the organism to survive to adulthood and to reproduce will have a chance to spread in a population and to persist through time. On the other hand, mutations that are beneficial to their carriers and increase the number of their offspring and their survival will have a selective advantage. Beneficial mutations will thus have a higher probability in persisting in a population through time. Looking at genetic diversity from the present situation back through time, we can imagine that we observe mutations that had been sieved through evolution in the past. Old and new beneficial mutations will remain in populations more likely, and old deleterious mutations will not be visible in today's genetic information due to their carriers not passing their genes onto the offspring.

To better demonstrate the process of mutation loss with time, we can simulate a model situation. Let's have a population of individuals that carry mutations with selection coefficient s, where $s < 0$ means the mutation is deleterious, $s = 0$ the mutation is neutral and $s > 0$ represents beneficial mutations. Empirical research shows that about 30% of mutations are deleterious, 42% neutral and only 18% of mutations bring the carrier a selective advantage (Barrick & Lenski, 2013). Starting from these conditions, the mutations become lost in time through genetic drift and selection (Fig. 1.5). We speak about mutation rate decay (Barrick & Lenski, 2013; Ho et al., 2005).

We can use Markov chains to simulate the basic principle of the mutation rate decay, when the relative decrease in mutation rate will develop from the starting conditions through genetic drift and selection. Let's assume that, for the purpose of simulating the molecular rate decay, the frequency of appearance of a mutation with the specific selection coefficient is equal to frequency of the mutation in the starting population. A population P with n individuals is characterised as $P = \{p_1, p_2, \ldots, p_n\}$, where $p = 1$ in individuals that carry the mutation and $p = 0$ in individuals without the mutation. The number of individuals with a mutation with a given selection coefficient follows information in Box 1 in Barrick & Lenski (2013) that lists the frequency of mutations f for different selection coefficients s. We start the simulation with a population P_1 in the first generation g_1. Development of the frequency of the mutation in the populations can be simulated as a Markov process:

$$f_{g+1} = \begin{cases} \frac{1}{n} \sum B\left(P_{g+1}\right), & \text{if } f_g\left(1+s\right) < 1 \\ 1, & \text{if } f_g\left(1+s\right) \geq 1 \end{cases}, \tag{1.1}$$

where B is bootstrap samplings of the population P_{g+1}. The population P_{g+1} is created as a permutation of $\|f_g\left(1+s\right)n\|$ individuals with the mutation and the remaining individuals without the mutation up to the total of n individuals. We can see that no lethal mutations occur in the simulated populations after the initial generations (Fig. 1.5, red line) and the deleterious mutations also quickly disappear from the populations (Fig. 1.5, orange lines). Selectively neutral and beneficial mutations persist in populations, although not universally (Fig. 1.5, blue and green lines), and some mutations disappear from the simulated populations through random drift. Multiple other factors such as demography, population fragmentation or associations between mutations in the genome will influence persistence and loss of mutations in a real biological system and in more advanced models (Ho et

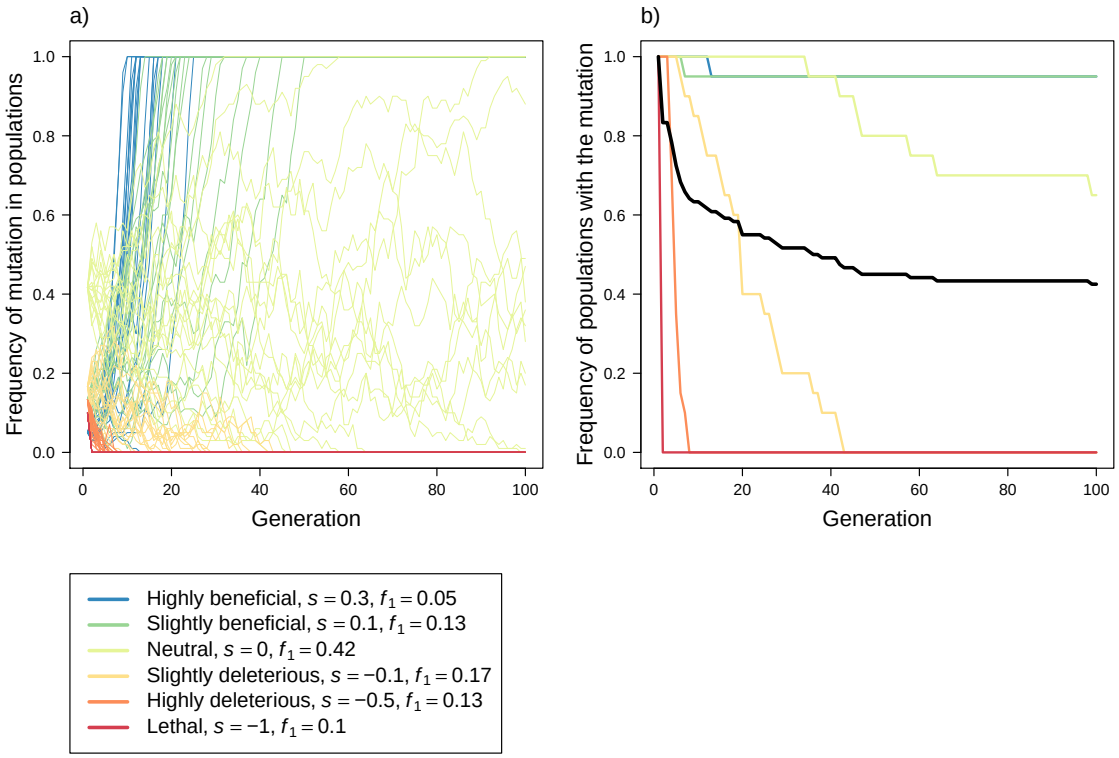


Figure 1.5: **Simulation of the principle of the mutation rate decay.** a) Persistance of individual mutations modelled with Markov chains in 20 populations consisting of 100 individuals each. b) Frequency of populations in which the mutation persists. Black line shows the total proportion of populations that have the starting mutation with a certain selection coefficient s at the given time. Coloured lines demonstrate that the loss of mutations occurs predominantly in selectively deleterious mutations, and that the decay is fastest in early generations.

al., 2005). Despite limitations, our simple simulation shows that when we estimate the mutation rate from old divergences, we will evaluate fewer mutations than how many truly appeared in the populations early after their split. When we categorised mutations into six groups, from lethal with $s = -1$ to selectively beneficial with $s = 0.3$, we lost more than half of all initial mutations in 100 generations (Fig. 1.5, black line).

Because of high variability of the mutation rate between lineages and through time, extrapolating the mutation rate outside of the range of data used to estimate it is risky. To avoid the errors, we need to find information that will provide a rough time frame for the possible divergence in the group. Using the terminology of BI, we need to specify the priors. Subsequent molecular dating analysis will be based on the most accurate and reliable calibration that can be found.

There are two ways how to calibrate the molecular clock - by calibrating tree tips or by calibrating the internal nodes of the phylogeny. Calibrating the molecular clock from the tree tips assumes that the sequences originated at different times (Martínková et al., 2013; Murray et al., 2016). Depth of the dated tree tips (difference between the oldest and the youngest sequence in the dataset) should reach at least one tenth of the expected time frame between the tips and the root of the tree. This means that we will not be able to use DNA sequences obtained from museum collections several decades old for dating divergence of taxa that could have occurred hundred thousands to millions of years ago. Instead, we will have to find suitable fossil material several thousands to tens of thousands years old, estimate the age of the fossils using radiocarbon dating or a similar method, isolate DNA from the fossils, sequence it and use those sequences for the tip-dated calibration of the molecular clock (Martínková et al., 2013).

Calibration of internal nodes is less demanding on technological applications than the tip-dated calibration, but we should verify the suitability of the node height limits for the specific divergence. Paleontological, geological or biogeographic information can guide the node calibration. Nodes in a phylogeny characterise points of divergence of two lineages. That is, a node represents a hypothetical most recent common ancestor of the two lineages. A fossil organism known to be a common ancestor of a certain monophyletic group can be used as a suitable calibration anchor, a so-called time constraint. Some paleontological data may contradict one another with respect to the phylogeny (Aghova et al., 2018; Near, Bolnick, & Wainwright, 2005). For example, a fossil considered an ancestor for a genus would be dated as older than a fossil ancestor of a family. We need to identify the conflicts using cross-validation (Near & Sanderson, 2004). In addition to fossil calibration, we can use geological (e.g. island emergence above the sea level or ice-sheet retreat (Martínková, McDonald, & Searle, 2007)) or biogeographic data (e.g. formation of land bridges between continents (Pečnerová et al., 2015)). Geological and biogeographic calibration enables the use the molecular clock hypothesis in groups and at times when the fossil record is rare.

We can demonstrate the effect of additional information on molecular dating using the example of the Irish population of stoats (*Mustela erminea*; Carnivora: Mustelidae) (Martínková et al., 2007). Using additional information in molecular dating requires accepting a set of assumptions about the biological system that should be firmly anchored in reality. For example, the analysis of genetic diversity in Irish stoats indicates a single colonisation of the island followed by a population expansion. That means that the number

of first colonists was relatively small and the influence of the founder effect and genetic drift resulted in low genetic diversity shortly after the colonisation. The subsequent population expansion means that many offspring, and thus carriers of mutations, survived and thus we observe an increase in genetic variability. The burst of diversification will look like a star tree. In Irish stoats, we observe a star tree in the DNA sequence data (Martínková et al., 2007). When we identified the pattern of colonisation, we can consider the time frame. Arrival of stoats to Ireland was unlikely before the ice-sheets retreated 40 thousand years ago. On the other hand, the oldest stoat fossils from Ireland are dated to 13 thousand years ago. The two dates set the youngest and the oldest possible date for the colonisation. The time limits represent a uniform prior for the tree root of mitochondrial DNA sequences of stoats from Ireland (Fig. 1.6, extent of the blue field on the x axis). We chose the prior for mutation rate as uninformative (Fig. 1.6, extent of the blue field on the y axis). Using the Bayesian coalescence analysis, we estimate the posterior distribution of the tree height and mutation rate based on the coalescence of the DNA sequences given the priors. The root height thus reflects the age of the Irish stoat population. The obtained mutation rate estimates the rate of evolution in stoats during late Pleistocene glaciations and can be applied to other stoat populations in which we expect genetic changes at similar time frames. The intraspecific mutation rate in stoats equals to about 14% change per million years, which is an order of magnitude faster than mutation rates dating speciation events (Ho et al., 2005; Pečnerová et al., 2015; Smith et al., 2009). We need to bear in mind that the estimated mutation rate cannot be used in million years old divergences, as it will decay with time (Fig. 1.5).

Mastering the molecular dating analyses requires good statistical knowledge of the analyst, but equally imporant is the ability to cooperate with specialists in other fields. Namely, biologists planning molecular dating analyses should work with paleontologists, radiologists and geologists to obtain reliable data for the molecular clock calibration. Avoiding the pitfalls of molecular dating is easiest with correctly identified fossil remains that represent ancestors of the studied contemporary taxa, reliably estimated ages of the fossils from radiocarbon dating and properly chosen geological and climate events for constraining the nodes of the dated phylogeny.

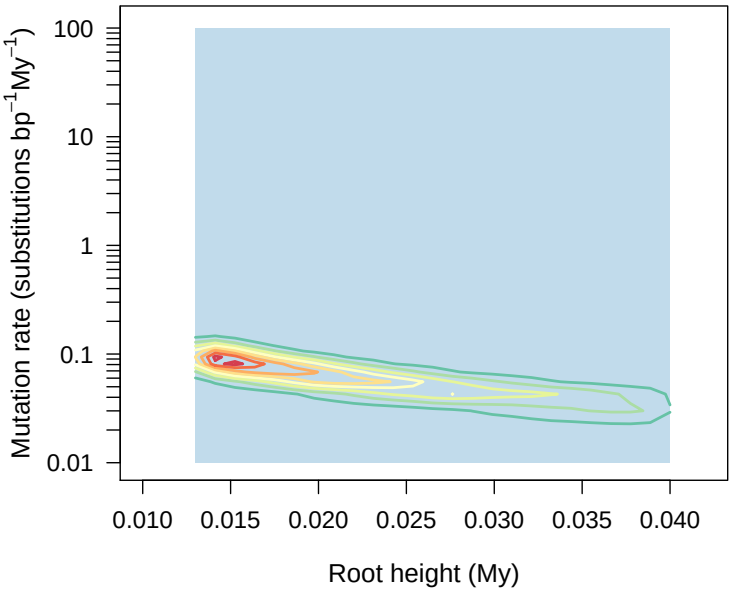


Figure 1.6: **Relationship between the prior and the posterior in molecular dating using Bayesian coalescence analysis.** Blue shading - priors for root height and mutation rate. The root height prior for the Irish population of stoats was set as a uniform distribution in range $[0.013, 0.04]$ million years (My). The mutation rate prior is set as uninformative as a uniform distribution in range $[0, 150]$ substitutions per base pair (bp) per My. Contour graphs - the area of the highest density in the posterior sample (red to green colour).

1.3 Phylogenetic congruence in search of functional relationships

During the course of evolution, organisms influence one another, sometimes to the extent, where a change in one organism leads to a reciprocal change in another. We speak of coevolution (Krasnovyd, Vetešník, Gettová, Civáňová, & Šimková, 2017; Simoes et al., 2016). Similarity in topology of phylogenetic trees of two groups of associated organisms is called codivergence. Where the coevolution assumes mutual influence of the pairs of organisms, codivergence does not imply causality in change. Two trees codiverge, when we observe a corresponding divergence in both of the associated clades in the two phylogenies.

The character of the association may differ, but we study codivergence usually between parasites and their hosts. Following fundamental principles of parasitism, we would expect that random selection of hosts might seldom occur in generalists, but would still be restricted to a certain taxonomic group. For example, a virus causing rabies, RABV, may infect many mammalian species, but plants are resistant to the infection (Marston et al., 2018). Most parasite-host relationships will reflect variable specificity of the parasite and differences in susceptibility of the host. Their interactions through time will modify not only the ability of the parasite to infect the host, but also protective mechanisms of the host. Herein, we will assume non-random associations. Phylogenetic codivergence between parasites and hosts can be defined in situations, when we observe a speciation event in hosts co-occurring with a speciation event in parasites. The direction of the interaction is not relevant for the purposes of codivergence.

We can test associations between parasites and hosts in phylogenetic context using comparisons of matrices of pairwise phylogenetic distances (Legendre et al., 2002; Martinez-Aquino, 2016). The strictest comparison in assessing codivergence is a test whether the matrix of phylogenetic distances between parasites is congruent with a similar matrix of the hosts. Two matrices $\mathbf{B}$ and $\mathbf{C}$ are congruent, when there exists a matrix $\mathbf{M}$, for which we can equate:

$$\mathbf{C} = \mathbf{M}^{\mathrm{T}}\mathbf{B}\mathbf{M}, \tag{1.2}$$

where T means transposition. Illustratively, we can understand the essence of matrix congruence as if we compared two photographs of the same place taken at different weather conditions.

Congruence between phylogenetic trees can be tested with a permutation test of congruence between distance matrices (CADM; Campbell et al., 2011). The CADM test calculates a global statistic W that is defined over the interval $[0, 1]$. $W = 0$ means that two matrices are fully incongruent, and $W = 1$ means that matrices are fully congruent. The statistic W provides an intuitive guide to the extent to which two matrices are congruent. Statistical significance of W can be tested with a permutation test that evaluates a null hypothesis that the two matrices are incongruent ($W = 0$). A non-significant CADM test means that we can expect at least partial congruence between matrices of phylogenetic distances of parasites and hosts.

The CADM test can be used to test the global congruence between a group of parasites and a group of hosts with the caveat that the associations must be strictly paired. This

means that in the CADM test, each parasite infects exactly one host and each host has exactly one parasite. In natural applications, we often observe that multiple parasites infect one host or that one parasite infects multiple host species. In host-parasite systems with multiple associations per taxon, ParaFit can test the global signal for codivergence. ParaFit uses a permutation test to evaluate whether the associations between parasites and hosts are random (Legendre et al., 2002).

The ParaFit method calculates a fourth-corner statistic using matrix algebra:

$$\mathbf{D} = \mathbf{C}\mathbf{A}^{\mathrm{T}}\mathbf{B}, \tag{1.3}$$

where **B** is matrix of phylogenetic distances between parasites, modified with principal coordinate analysis, **C** is a similar matrix obtained for the host relationships and **A** is an association matrix, where value 1 means association between the corresponding taxa and 0 means no association. Matrix **D** is thus a linear combination of phylogenetic information in the parasite and host trees weighted through their associations (Legendre et al., 2002). ParaFit calculates a statistic ParaFitGlobal, which is a sum of squares of the values in matrix **D**.

Statistical significance of codivergence between the two phylogenies given their associations does not fully suffice in explaining the biological phenomenon. Let's use an example with random simulated trees (Fig. 1.7). We tested the signal for codivergence in a set of 100 random trees compared to the starting tree. We found several trees with significantly non-random codivergence, where each tip has exactly one association in both compared trees using ParaFit (Fig. 1.7c, orange colour). The non-random codivergence is surprising in light of tree distance calculated as the weighted, normalised Robinson & Foulds distance (RF; Robinson & Foulds, 1981). The RF distance between random trees with non-random associations reached more than 50% of maximum RF distance weighted to the total tree length.

Applying a biologically more realistic model, we can simulate the divergence of two trees with evolving host-parasite interactions using a Markov chain:

$$T_i = N\left(N\left(T_{i-1}\right)\right), \tag{1.4}$$

where T is a phylogenetic tree, $i \in \mathbb{N}^+$ up to the total length of the simulation, T_0 is a starting tree and N is the nearest neighbour interchange. We observed statistically non-random codivergence with the starting tree in 13 simulated phylogenies from a series of 100 trees, where each tree differed from a previous tree by two nearest neighbour interchanges (Fig. 1.7d).

In most simulated paired trees of parasites and hosts, in which we observed non-random codivergence for $p(\text{ParaFitGlobal}) \leq 0.05$, we also observed that the matrices of phylogenetic distances were congruent at $p(W) \leq 0.05$ (Fig. 1.7d, f). This is to be expected. Our simulations showed a surprising result, where some randomly codiverging trees tested with ParaFit can be significantly congruent as tested with CADM (Fig. 1.7c, e), indicating a partial congruence in some parts of the two trees.

Our simulations demonstrate that we can observe higher than random congruence even in relatively divergent trees. Phylogenetic codivergence and congruence thus merits a more detailed study.

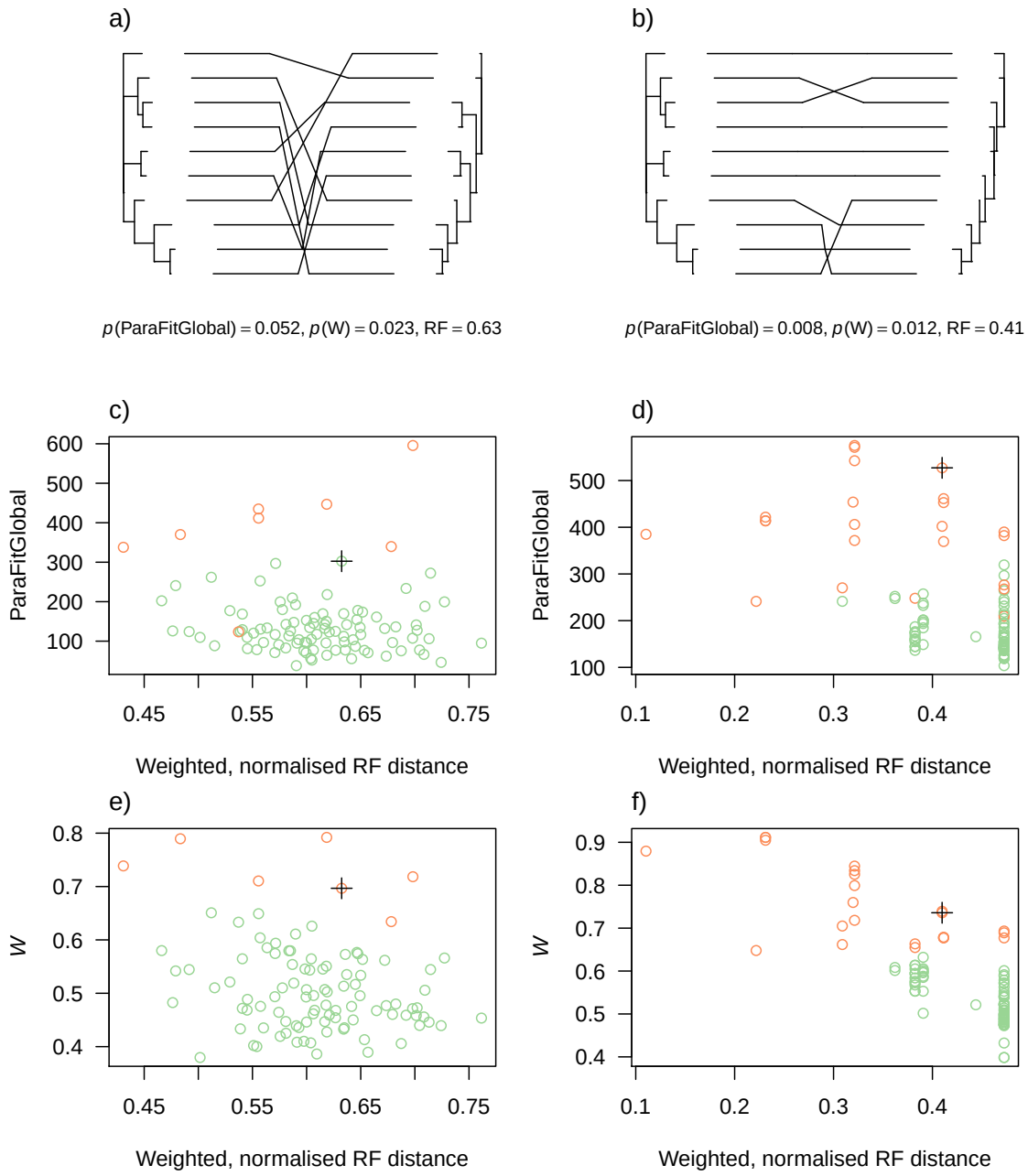


Figure 1.7: **Codivergence and congruence of simulated phylogenies.** a) Tanglegram between randomly generated trees shows that codivergence between them cannot be distinguished from random. b) Example of codiverging phylogenetic trees. c-f) Weighted, normalised Robinson & Foulds distances (RF; Robinson & Foulds, 1981) between random (c, e) and step-wise diverging (d, f) phylogenies in relation to the statistic ParaFitGlobal (Legendre, Desdevises, & Bazin, 2002) (c, d) and CADM test statistic W (Campbell, Legendre, & Lapointe, 2011) (e, f). Orange colour - ParaFitGlobal is significantly higher than in random associations between the trees (c, d), W is significantly higher than in fully incongruent matrices (e, f); green colour - the test were not significant at $\alpha = 0.05$; cross - example of associations displayed in panels a) and b).

We studied codivergence of New World arenaviruses with their rodent hosts from subfamilies Sigmodontinae and Neotominae (Rodentia: Cricetidae) (Irwin, Bayerlová, Missa, & Martínková, 2012). The relationship was historically believed to represent codivergence based on the fact that arenaviruses in the New and Old World infect rodents from different families and the infections may be asymptomatic (Bowen, Peters, & Nichol, 1997). Even though empirical data rejected the hypothesis of long-term codivergence of arenaviruses with their hosts (Jackson & Charleston, 2004), the opinion lingered in predominantly biomedical literature. Studying all known arenaviruses from South and North America in the context of all available hosts from the respective rodent subfamilies with DNA sequence data, we found signal for codivergence only in some groups of the parasites. In these groups, phylogenetic clustering of arenaviral infection on the host phylogeny (Chapter 1.4) in conjunction with geographical modelling showed that the parasites utilise related hosts that are locally available (Irwin et al., 2012).

We found no codivergence when comparing a molecular phylogeny with a similarity tree derived from skull morphological data in tree squirrels from the tribe Sciurini (Rodentia: Sciuridae) (Pečnerová et al., 2015). Instead, the skull morphology corresponded to the current taxonomy of tree squirrels at the genus level and was likely driven by adaptation to different food sources. The molecular phylogeny indicated a step-wise divergence associated with lineage colonisation of new regions.

Interpreting presence (Irwin et al., 2012) or absence (Pečnerová et al., 2015) of statistically significant signal for codivergence of two phylogenies requires knowledge of the studied system and search for covariates that may influence the respective evolutionary histories. By applying the codivergence analysis, we were able to find signals indicating interesting functional variations in the diverse groups of organisms.

1.4 Predictive modelling of species traits in the context of genetic diversity

Characteristics of organisms, their traits, have a heritable component. The concept of trait heritability forms a firm foundation of current systematics. For example, we speak of a synapomorphy if a monophyletic group shares a trait and we expect that the trait was present also in their ancestor. Traits change over the course of evolution. Qualitative traits arise or become lost and quantitative change their size or range. The biological concept of trait inheritance poses a complication in statistical modelling, because we cannot assume that the observations are independent. We need to quantify the extent to which phylogenetic relatedness influences trait variability and to account for the relatedness in the analyses.

When studying binary traits in the phylogenetic context, we can use metrics derived in community ecology that assess phylogenetic similarity of communities. Specifically, the net relatedness index and the nearest taxon index are useful to model evolution of a binary trait on the phylogeny.

Net relatedness index (NRI) shows a scaled and transformed difference between mean phylogenetic distance of the species with the given trait and the overall mean phylogenetic distance in the tree (Webb, Ackerly, McPeek, & Donoghue, 2002):

$$\mathrm{NRI} = -1\frac{\mu(\mathbf{b}') - \mu(\mathbf{b})}{\sigma(\mathbf{b})}, \tag{1.5}$$

where $\mathbf{b}$ is a vector of pairwise phylogenetic distances, $\mathbf{b}'$ is a vector of pairwise phylogenetic distance of the species that share the trait of interest, μ is a mean and σ is a standard deviation. The multiplicaiton by -1 in the equation 1.5 simplifies the interpreration of NRI. Values greater than zero indicate that the trait is distributed on the tree in a phylogenetically closely related group and the values less the zero mean that the trait is overdispersed on the tree. Values of NRI ≈ 0 mean that the species with the trait are randomly related with respect to the observed diversity of the studied system. We test the statistical significance whether NRI $\neq 0$ using permutations. The model that permutes the trait value across the whole phylogeny without topological limitations is the one applicable for comparative phylogenetics.

The other community ecology index useful in comparative phylogenetics is the nearest taxon index (NTI). We calculate the NTI similarly to the NRI (Webb et al., 2002). The NTI also compares observed and expected relatedness scaled to the standard deviation of the phylogenetic distances, but it does not work with all pairwise distances. Instead, for each taxon, the NTI searches for the smallest phylogenetic distance that signifies its nearest neighbour. We can define NTI as:

$$\mathrm{NTI} = -1\frac{\mu(\mathbf{c}') - \mu(\mathbf{c})}{\sigma(\mathbf{c})}, \tag{1.6}$$

where $\mathbf{c}$ is a vector of the smallest phylogenetic distances for each species in the phylogeny and $\mathbf{c}'$ is a vector of the smallest phylogenetic distances of the species that share the trait of interest.

The utility of the indices NRI and NTI in comparative phylogenetics is in their ability to estimate the conservatism of a given trait (Fountain-Jones et al., 2017; Webb et al., 2002).

When the trait is conservative across the whole studied diversity, both indices will show high values. When the trait is present predominantly in closely related taxa, the NRI will decrease, but the NTI will remain high. In community ecology, high values of NRI and NTI mean selection of evolutionary conservative traits driven by environmental conditions (Fountain-Jones et al., 2017). We used an analogous interpretation in studying the trait distribution in a model of host-parasite interaction in fungi and viruses infecting mammals.

We found an opposite relationship in hibernating bats (Chiroptera: Vespertilionidae, Miniopteridae, Rhinolophidae) infected with fungus *Pseudogymnoascus destructans* (Ascomycota: Pseudeurotiaceae) than that reported by Fountain-Jones et al. (2017). Infected bats were significantly clustered on the phylogenetic tree (NRI > 0), but the sister taxa were randomly affected (NTI ≈ 0) (Zukal et al., 2014). We expect that the result reflects non-representative sampling that concentrated on the representatives of the genus *Myotis*. *Myotis* species are the most frequent bats in Holarctic hibernacula and due to their species diversity, *Myotis* strongly influence the phylogenetic composition of bat communities in hibernacula (Horáček, Bartonička, Lučan, & Czech Bat Conservation Trust, 2014).

Group B of New World arenaviruses are important human pathogens (Charrel & de Lamballerie, 2003). Hosts of the group B arenaviruses are randomly distributed on the phylogenetic tree of all possible rodent hosts from South and North America. Hosts of other groups of arenaviruses, not pathogenic to humans, are closely related (Irwin et al., 2012). Together with information on host-parasite codivergence (Chapter 1.3), the absence of phylogenetic clustering on the host tree means that group B arenaviruses are capable of frequent host switching.

For quantitative traits, we can study evolution of the trait using phylogenetic signal. We consider a trait to show phylogenetic signal when two phylogenetically related taxa will be more similar to one another than two taxa randomly selected from a phylogeny (Blomberg, Garland jr., & Ives, 2003; Losos, 2008; Pyron, 2015). The phylogenetic signal is thus covariance of the quantitative trait in different taxa that can be explained by their shared evolutionary history (Revell, Harmon, & Collar, 2008). Whether we assess body size, food preferences or susceptibility to infection, we assume that the mean of the given trait in a species will be determined by heritable processes. Closely related species will have similar mean values of the studied trait.

Phylogenetic signal as defined by Blomberg et al. (2003) is calculated as a ratio of mean squared error of the trait given the phylogenetically corrected mean of the trait (trait value at the root of the phylogeny) to the mean squared error of the trait transformed with the variance-covariance matrix derived from the phylogeny:

$$\frac{\mu\left(\boldsymbol{\varepsilon}^2\right)}{\mu\left(\hat{\boldsymbol{\varepsilon}}^2\right)}, \tag{1.7}$$

where $\boldsymbol{\varepsilon}$ is a vector of differences between the vector of trait values and the phylogenetic mean of the trait value at the root of the tree and $\hat{\boldsymbol{\varepsilon}}$ is a vector of trait values corrected for the variance-covariance of the tree using the generalized least-squares. The values of the phylogenetic signal calculated with equation 1.7 will be dependent on the tree size, topology and length, making universal cross-comparisons difficult. Therefore, we now express phylogenetic signal as a statistic K that is a ratio of the observed to expected ratios

in equation 1.7 (Blomberg et al. 2003). The expected values are calculated assuming a Brownian motion model of the trait evolution.

When $K \approx 1$, the variable changes along the phylogeny as expected under the Brownian motion model of trait evolution. Departures from one in either direction indicate biologically meaningful interpretation in terms of adaptive evolution. For $K < 1$, closely related species have trait values less similar than expected due to their relatedness and thus they might be influenced by alternative adaptive processes. The adaptation to a similar selection pressure affecting multiple taxa may drive trait similarity within a clade with $K > 1$ (Blomberg et al., 2003), and a trait with $K > 1$ could be correlated with phylogenetic niche conservatism (Losos, 2008).

We found statistically significant phylogenetic signal for $K < 1$ in pathogen load in hibernating bats (Zukal et al., 2016). The fungus *P. destructans* causes disease white-nose syndrome (WNS) in hibernating bats inhabiting Holarctic biogeographic ecozone, where the Nearctic bats often experience drastic population declines following pathogen invasion in the hibernacula, but Palearctic bats are able to tolerate the infection better (Lorch et al., 2011; Wibbelt et al., 2010; Zukal et al., 2014; Zukal et al., 2016). The pattern with the highly susceptible species in the Nearctic and the tolerant species in the Palearctic can be observed in multiple affected genera with representatives in both regions, such as *Myotis*, *Eptesicus*, or *Pipistrellus/Perimyotis* species group. However, species from one genus vary in infection intensity also in the Palearctic (Pikula et al., 2017; Zukal et al., 2016).

We can demonstrate the scale of interspecific variation in fungal load by simulating the expected fungal load at tree tips with a model of evolution following Brownian motion and comparing the simulated data with the empirical data (Fig. 1.8). The Brownian motion evolutionary model assumes that the rate of evolution ρ is stable through time and thus the change in a trait value depends on the rate of evolution and varies with respect to a stochastic component drawn from a normal distribution with mean equal to zero and variance equal to ρ^2 (Butler & King, 2004). With progressing time, the lineages will accumulate more random changes in the trait values leading to their divergence from one another. The density plots from simulated trait values in derived taxa (*Myotis*) will show greater variance than taxa that branched from the common ancestor early (*Rhinolophus*) on the phylogeny (Fig. 1.8, green colour).

The variable susceptibility of different species that cannot be attributed to the relatedness of taxa might thus reflect other influences on infection intensity such as species ecology, behaviour or environmental conditions during hibernation. Relationship between quantitative traits in a phylogenetic context can be studied with the phylogenetic generalized least-squares (PGLS; Ives, Midford, & Garland jr., 2007; Revell, 2010). In PGLS, the inferred intercept and slope of the regression are corrected with respect to the fact that the data are hierarchically structured and co-vary to a certain extent due to the relatedness of the taxa (Pikula et al., 2017; Zukal et al., 2016). The ordinary least-squares method is extended in the phylogenetic context by adding a variance-covariance structure derived from a phylogeny to the regression (Symonds & Blomberg, 2014). The variance-covariance structure reflects the non-independence of the samples in PGLS, where the samples (trait values in species) will be influenced by their shared evolutionary history. The variance-covariance matrix in the PGLS represents the expected covariance of residuals in the regression (Symonds & Blomberg, 2014). That means that in using PGLS,

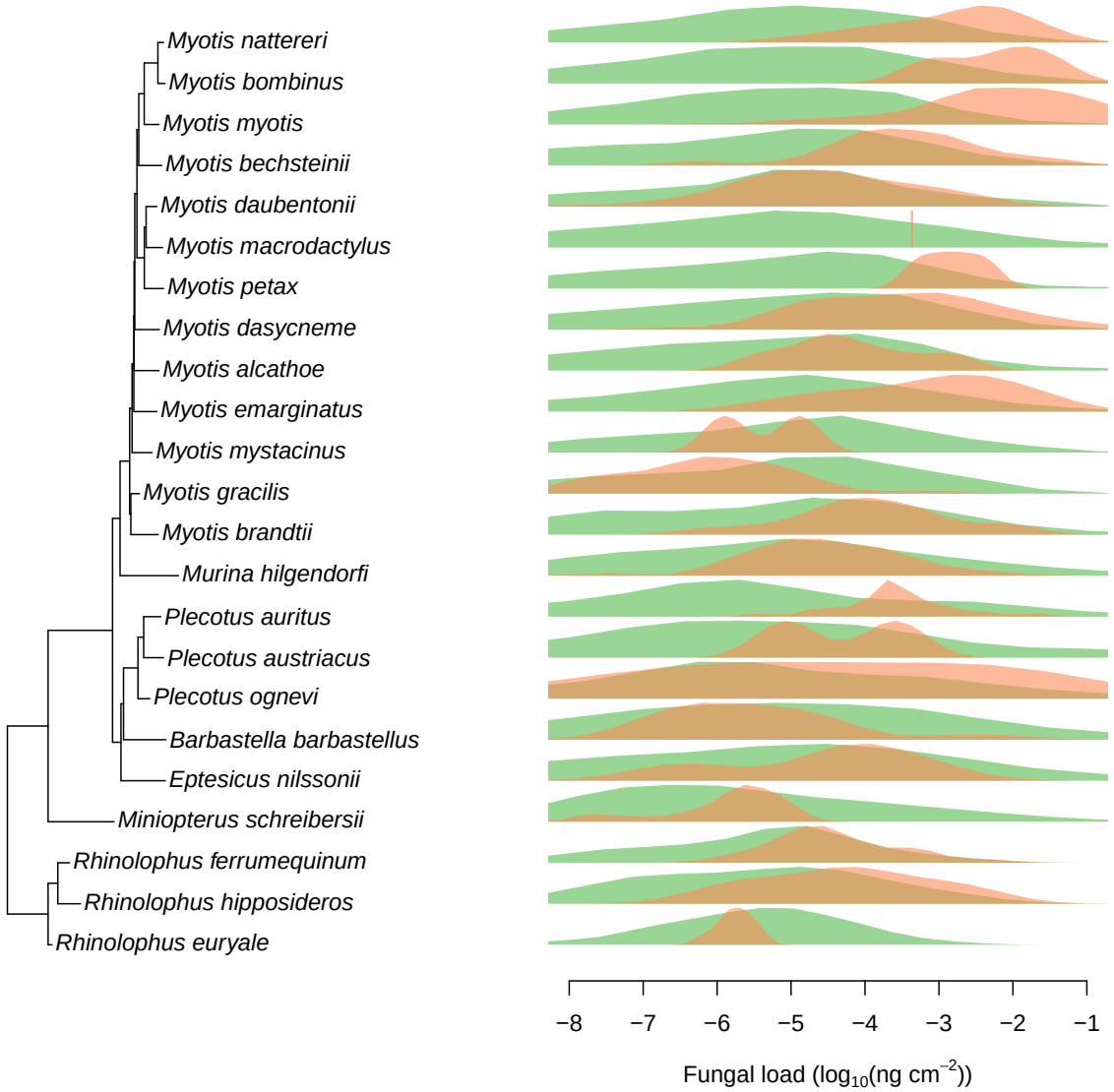


Figure 1.8: **Trait evolution modelled with Brownian motion.** The fungal load trait values were modelled on a multilocus phylogeny of hibernating Palearctic bats infected with *Pseudogymnoascus destructans* (Zukal et al., 2014). The rate of evolution $\rho = 3.8$ and fungal load at the root equal to 3.62×10^{-6} ng cm^{-2} used in simulations were derived from Zukal et al. (2016) and Pikula et al. (2017). Green - density plots from 100 simulations of the expected fungal load at tree tips, orange - density of the empirical fungal load in each species.

we assume phylogenetic signal in the residuals from the regression, not necessarily the variables themselves (Revell, 2010).

The current regression methods in the phylogenetic context take into account intraspecific variability of the trait value (Ho, & Ané, 2014; Ives et al, 2010). This is an important advancement, because including within species variation in a model improves the fit between the model and the observations. We applied the PGLS to study the relationship between infection intensity measures that characterize disease progression from a superficial fungal infection to development of the WNS (Zukal et al., 2016). By correcting for shared evolutionary history between taxa and trait variability within species, we showed that the skin lesions diagnostic of WNS develop with increasing fungal load more slowly than expected (Zukal et al., 2016), but severity of the disease increases faster if the bat had more skin lesions (Pikula et al., 2017).

1.5 Using outliers for identification of biologically relevant information in data structure

Each biological system is variable and statistical methods assume that the values of a trait follow a frequency distribution, often the normal distribution. Trends are then estimated in relation to the parameters of the distributions such as their mean. Statistical analyses transform data in such a way that we minimise observed variance, we explain the variance or we categorise data into classes with minimum variance. In each of these applications, we focus on the overall information content in the data and we consider outliers as undesirable. The outliers would either be viewed as errors in measurement or unique aberrations likely to be unsuccessful in the course of evolution. Since they may have a strong influence on results of the statistical analyses, the outliers are deleted from the datasets upon identification. While deleting measurement errors is always warranted, some outliers might carry information about an alternative mechanism with a biological foundation. If we had a sufficiently large sample size (hundreds to thousands of measurements), outliers generated by an alternative mechanism would form a separate cluster in ordination analyses and we would detect them during data explorations as relevant and worth of further investigation. Datasets in biology are often smaller and the power of statistical tests cannot distinguish difference in means of common and rare phenomena. Rare phenomena that generate outliers in small datasets are long overlooked or they may appear as descriptions of individual occurrences (case studies; cf. Pikula et al., 2017).

We used the outlier analysis in two applications to detect observations generated by an unusual and rare mechanism. We designed a method SigHunt to search for genomic islands in eukaryotic genomes. The rare mechanism that generates the genomic islands is the horizontal gene transfer. The second method of outlier analysis is EEO that can be used to identify exceptional behaviour in hibernating bats. We expect the exceptional behaviour during hibernation in moribund bats with WNS, but lethal cases of WNS are rare in the Palearctic (Pikula et al., 2017).

Genomic islands are regions in the genome that the organism obtained from unrelated organisms through horizontal gene transfer. Horizontal gene transfer can equip organisms with pre-evolved traits that may contain metabolic pathways enabling utilisation of new resources. We hypothesised that the genome of the fungal pathogen *P. destructans* might contain genomic islands based on the fact that the fungus belongs to a group of predominantly saprophytic organisms. An ability to infect live tissues biochemically differs from dead tissue decomposition. A parasite thus needs to produce compounds facilitating invasive growth or protecting the parasite against the immune system of the host. Obtaining a pre-evolved metabolic pathway for utilising new resources is advantageous, but the process can occur only between organisms that are at least temporarily in close contact to expedite exchange of genetic information. Bacteria use conjugation or transduction as common mechanisms of horizontal gene transfer. In eukaryotes, transposons or retrotransposons enable transfer of genetic material between organisms that do not normally produce viable offspring (Schaack, Gilbert, & Feschotte, 2010).

We developed the method SigHunt that detects horizontally transferred genes in eukaryotic genomes using composition of the DNA molecules (Jaron, Moravec, & Martínková, 2014). We built the method on the assumption that molecular mechanisms of DNA repli-

cation and repair are heritable and in the long term lead to similarities in DNA composition in related organisms (Blokzijl, Janssen, van Boxtel, & Cuppen, 2018). Related organisms will have similar frequencies of nucleotide bases or short oligonucleotides in their genomes (Karlin & Burge, 1995). We call a set of frequencies of oligonucleotides a genomic signature. When a genomic region transfers from one organism to an unrelated organism, it will initially retain the genomic signature of the original genome that could be distinguished from the host genomic signature. Resolution of genomic signatures will be higher and they will be more sensitive to differences between taxa for longer oligonucleotides, because we will estimate oligonucleotide frequencies for more categories. The number of possible oligonucleotides in a DNA sequence is 4^n, where n is the oligonucleotide length.

We chose $n = 4$ for calculating genomic signatures in SigHunt and our vectors characterising a genome contained frequencies of 256 tetranucleotides. Not all tetranucleotides are informative for differentiating home genomic sequence from the horizontally transferred genes. To find genomic islands in a eukaryotic genome, we selected those tetranucleotides that had the smallest variance in the home genome and differed the most from genomes of other organisms that could have been sources for the horizontal gene transfer. SigHunt calculates genomic signatures for the selected tetranucleotides in sliding windows of the DNA sequence. Based on a kernel density estimate of the frequencies in a chromosome, SigHunt finds in which windows the tetranucleotide frequencies fall outside of the credibility intervals (Fig. 1.9). Identification of the genomic island assumes that in horizontally transferred regions, the genomic sequence will contain more tetranucleotides with extreme frequencies (Jaron et al., 2014).

Research of hibernating bat behaviour tests the hypothesis that infection with fungus *P. destructans* and development of WNS cause increased arousal frequency during the hibernation period and increased flight activity. Moribund infected bats from the Nearctic have been observed to dramatically change their behaviour during hibernation (Reeder et al., 2012), but similar changes in activity were not observed in infected bats that survived (Lilley et al., 2016). Since Palearctic bats with WNS survive better and we do not observe an associated population decline in any species, their winter behaviour should mimic that of Nearctic WNS survivors. We expect that the behaviour of most bats during hibernation will show regular patterns of torpor-arousal as per Lilley et al. (2016). In rare cases, bats with WNS die in the Palearctic (Pikula et al., 2017), and those should be detectable on recordings of behaviour as increased arousal frequency as per Reeder et al. (2012).

We developed the method EEO (extraction of exceptional observations; Škrabánek & Martínková, 2017) that identifies rare exceptional events based on their differentiation from the common observations. The core principle of the EEO method is identification of a point when we observe a sharp increase in sorted row sums of the distance matrix between observations. We calculated the pairwise distances between observations as Mahalanobis distances, because those are independent of the directionality of the change or units of the observations. The distance matrix can be calculated from normalised data with Euclidean distance for large datasets where computational expenses of Mahalanobis distances would be prohibitive. The EEO method identifies outliers as those observations that have sum of distances to all other observations higher than the threshold. Compared to other outlier detection methods, the EEO is suitable for small datasets starting from $10^1 - 10^2$ of observations and it is capable to effectively seach for outliers in multidimensional space.

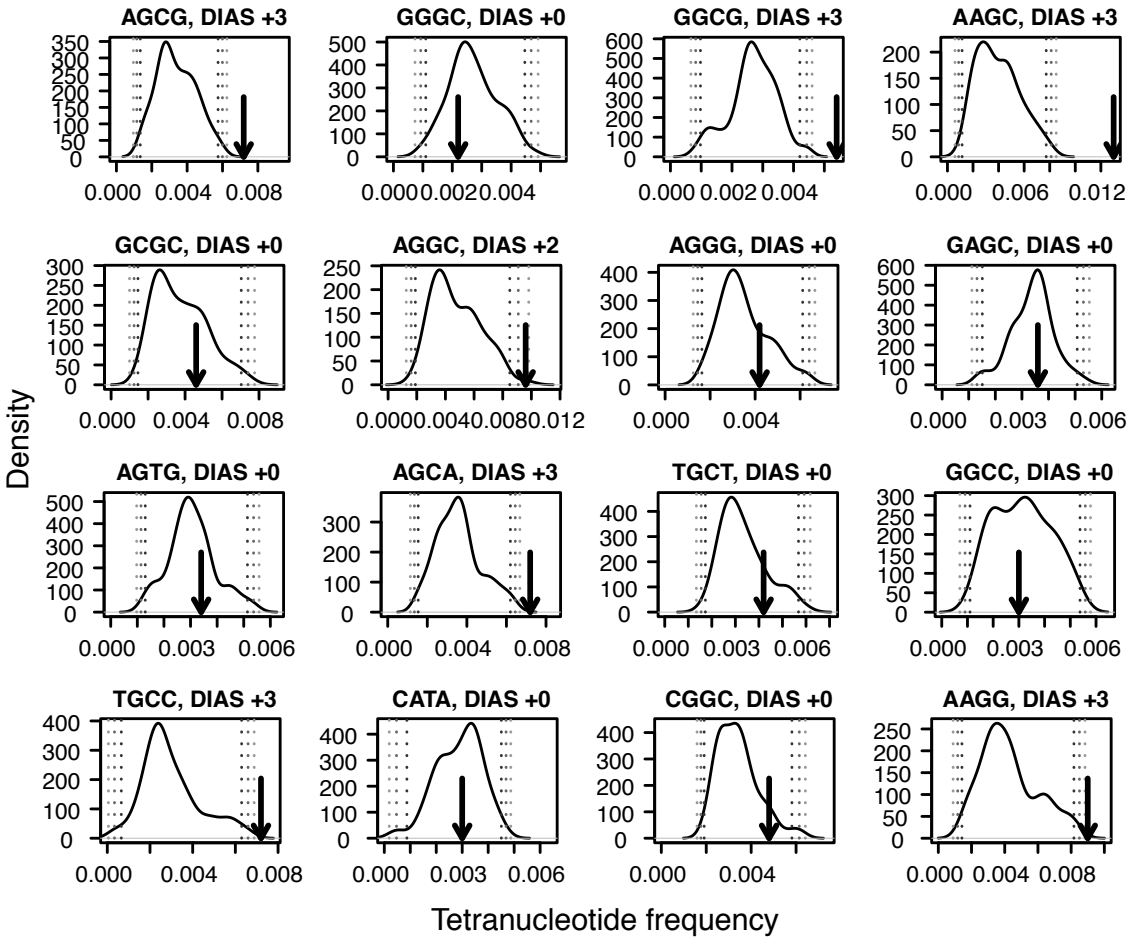


Figure 1.9: **Searching for genomic islands in eukaryotic genomes using SigHunt.** SigHunt calculates frequencies of selected tetranucleotides in overlapping adjacent regions of genomic sequence about $10^3 - 10^4$ bp long and estimates a kernel density for a longer sliding window of the genomic sequence (about $10^5 - 10^6$ bp). The short window can be considered a genomic island when its tetranucleotide freqencies (arrows) exceed marginal values. In the worked example, the marginal values (dashed lines) are defined as kernel density credibility intervals with probability $\alpha \in \{0.05, 0.025, 0.01\}$. If the tetranucleotide frequency is outside of the marginal values a discrete interval accumlative score (DIAS) increases by 1, 2 and 3, respectively. The final value DIAS $= 21$ means that the genomic signature of the tested DNA sequence considerably differs from its genomic context.

In studying biodiversity, we are able to address exciting questions about the nature. Phylogenetic and outlier analyses described in this thesis provide powerful tools to address complex questions with genetic, ecological or behavioural data. We use phylogenetic analysis to estimate the relationships between taxa and we are able to infer the age of their divergence through the molecular dating analysis. Once we have a strong phylogenetic hypothesis with good support, we are able to elucidate relationships between organisms with a codivergence analysis and we can characterise functional relationships between traits in light of the evolving systems. With outlier analyses, we can identify rare phenomena in biological datasets and submit those for further testing.

Chapter 2

Author contributions

Research questions that investigate global biodiversity in light of ecological, behavioural or epidemiological characteristics require an extensive access to data and expertise. Following successful cooperation across regions and specialisations, research teams become numerous in phylogenetic studies. The listed publications resulted predominantly from collaborations with colleagues and students from the Czech and foreign universities and research institutions. I derive estimation of individual contributions from the published author contributions paragraphs. The author contributions list k categories, in which the team members cooperated in the published research (e.g. conceptualisation, study design, material collection, laboratory experiments, statistical analysis, writing, etc.). Where the published papers did not contain the author contribution box, the number of categories and individual contributions follow a documented input towards the research questions and the publication. Each category i had equal weight for calculating the percent contribution of individual authors, and in each category the contribution was equally divided between all included authors R_i. Author contribution C was then the sum of all partial contributions in all listed categories in the given publication for the given author:

$$C = \frac{100}{k} \sum_{i=1}^{k} \frac{\phi(R_i)}{|R_i|}, \tag{2.1}$$

where R_i is a list of researchers contributing to the i-th category, $|R_i|$ is the length of the list of researchers, i.e. the number of authors contributing in the i-th category, and

$$\phi(R_i) = \begin{cases} 1, & \text{if 'NM'} \in R_i \\ 0, & \text{otherwise} \end{cases}. \tag{2.2}$$

For example, the paper 2.2.3 on phylogeography of Orkney voles (Martínková et al., 2013) had 17 authors and the author box lists nine categories of contributions to the research topic (Table 2.1). As each category has equal weight, it represents a fraction of $\frac{100}{9} = 11.1\%$. During the course of the research, I collected material in the field together with seven other authors, performed laboratory experiments with two others, analysed data with two others and co-wrote the paper with all authors. My calculated contribution equals to $C = 11.1 \times \left(\frac{1}{8} + \frac{1}{3} + \frac{1}{3} + \frac{1}{16}\right) \approx 9\%$.

The contribution discussions have an undeniable potential to sour professional relationships, but they are required in this type of work. My intention was thus to reflect published

Table 2.1: Example of calculating my author contribution in paper 2.2.3 (Martínková et al. 2013) using equation 2.1 for $k = 9$ categories. R_i - list of authors contributing to the given category, $|R_i|$ - number of authors contributing to the category, $\phi(R_i)$ - a function evaluating whether I contributed to the category (equation 2.2).

Category i	Contributing authors R_i	$\lvert R_i \rvert$	ϕR_i
Project design	KMD, JBS	2	0
Field material collection	TC, MF, GH, **NM**, MP, MaP, JPQ, JBS	8	1
Museum material collection	RB, TC, KMD	3	0
Laboratory experiments	RB, TC, **NM**	3	1
Lab support	KMD, ARH, GH, JBS, SB, TH	6	0
Data analyses	RB, TC, **NM**	3	1
Analyses support	RS, KMD, LE, POH, ARH, GH, SYH, JBS	7	0
Writing	JBS	1	0
Writing support	**NM** + 15 authors	16	1

or documented facts as objectively as possible. In the following section, I declare the putative shift in interpreting the percentage of my contributions, where the strict adherence to the equation 2.1 unjustly affected my students or colleagues.

2.1 Phylogenetic analyses papers

2.1.1 Jaarola M., **Martínková N.**, Gündüz İ., Brunhoff C., Zima J., Nadachowski A., Amori G., Bulatova N. S., Chondropoulos B., Fraguedakis-Tsolis S., González-Esteban J., López-Fuster M. L., Kandaurov A. S., Kefelioğlu H., da Luz Mathias M., Villate I., Searle J. B. 2004. Molecular phylogeny of the speciose vole genus *Microtus* (Arvicolinae, Rodentia) inferred from mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution* 33: 647-663.

Author contribution: 16%, I contributed to material collection, laboratory experiments, data analysis and manuscript preparation.

Main innovation of the paper: The article reports phylogeny reconstruction of voles from the genus *Microtus* based on a mitochondrial gene for cytochrom *b*. The analyses compare results from neighbour-joining, maximum parsimony, ML and BI trees. Phylogeny based on a single gene was unable to resolve deep divergencies, but it showed good support for *Microtus* and *Chionomys* and subgenera *Microtus* and *Terricola*.

2.1.2 **Martínková N.**, Moravec J. 2012. Multi-locus phylogeny of arvicoline voles (Arvicolini, Rodentia) shows small tree terrace size. *Folia Zoologica* 61: 254-267.

Author contribution: 67%, I designed the study from the publically available DNA sequence data and I contributed to the statistical analyses and writing.

Main innovation of the paper: A study inspired by the paper 2.1.1 (Jaarola et al., 2004) that attempts to solve analytical problems in the older study. The paper resolves phylogenetic relationships between voles of the tribe Arvicolini that were previously problematic.

We achieved the improvement by using both mitochondrial and nuclear sequences and we tested the impact of missing data on results reliability and stability. Further, we tested the impact of branch length priors on BI posterior node support with the alignment with missing data. We found that the vole phylogeny from an alignment that contained over 70% of missing data belongs to a small terrace; that is there exists a small number of alternative trees that have the same likelihood as the consensus tree from the posterior sample of the BI. In other statistical analyses, missing data need to be either imputed or samples or variables containing missing data are omitted from the analyses, and thus practice calls for less than 5% of missing data. Relative robustness of phylogenetic analyses from matrices with missing data shows the remarkable information value of historical processes during DNA sequence evolution.

2.1.3 **Martínková N.**, Zima J., Jaarola M., Macholán M., Spitzenberger F. 2007. The origin and phylogenetic relationships of *Microtus bavaricus* based on karyotype and mitochondrial DNA sequences. *Folia Zoologica* 56: 39-49.

Author contribution: 37%, I co-designed the study together with J. Zima and I contributed to the laboratory experiments, statistical analyses and writing.

Main innovation of the paper: *Microtus bavaricus* belongs among voles with small distribution ranges that inhabit montane and submontane regions in Europe. We succeeded in obtaining a karyotype as well as DNA sequences of the species. We used neighbour-joining, maximum parsimony, ML and BI phylogenetic analyses to show that *M. bavaricus* is a young, recently diverged species from the *Microtus multiplex* species group.

2.1.4 Pečnerová P., **Martínková N.** 2012. Evolutionary history of tree squirrels (Rodentia, Sciurini) based on multilocus phylogeny reconstruction. *Zoologica Scripta* 41: 211-219.

Author contribution: 50%, although I contributed to all categories that constitute this work, my real share was lower. The paper is in essence the Bachelor thesis of P. Pečnerová. I designed the topic and I supervised the student throughout her work in data collection, analyses and writing.

Main innovation of the paper: Tree squirrels of the tribe Sciurini are distributed in the Holarctic and Neotropic biogeographic ecozones. Most species of tree squirrels and three out of five genera belonging to the tribe can be found in the Neotropic. Using publically available sequences of eight genes, we found well supported phylogenetic divergence between genera *Sciurus* and *Tamiasciurus* based on the BI phylogeny. Other genera from the tribe Sciurini were paraphyletic with respect to *Sciurus* and the relationships between species were unresolved. Alternative approaches of phylogenetic reconstruction from multilocus data with supermatrices or supertrees showed similar results. We concluded that the unresolved relationships were a result of explosive radiation and incomplete lineage sorting of tree squirrels in the Neotropic.

2.1.5 Kandemir İ., Sözen M., Matur F., Kankiliç T., **Martínková N.**, Çolak F., Özkurt S. Ö., Çolak E. 2012. Phylogeny of species and cytotypes of mole rats (Spalacidae) in Turkey inferred from mitochondrial cytochrome *b* gene sequences. *Folia Zoologica* 61: 25-33.

Author contribution: 22%, I worked on the statistical analyses and manuscript preparation and helped with the laboratory experiments and the study design.

Main innovation of the paper: Mole rats of the genus *Nannospalax* belong to a group of underground rodents needing a taxonomic revision. We studied the relationships between mole rat populations in Turkey using neighbour-joining, maximum parsimony, ML and BI phylogenetic methods in conjunction with information on their karyotypes. Further, we altered the prior for branch lengths in the BI analysis to ascertain that the posterior credibility interval of the tree length included its ML estimate. We found that *Nannospalax* populations with low number of chromosomes ($2n \in \{36, 38, 40\}$) usually form monophyletic groups. Populations with high numbers of chromosomes ($2n \geq 52$) were paraphyletic on a phylogeny contructed from partial sequences of a single mitochondrial locus.

2.1.6 Vallo P., Benda P., **Martínková N.**, Kaňuch P., Kalko E. K. V., Červený J., Koubek P. 2011. Morphologically uniform bats *Hipposideros aff. ruber* (Hipposideridae) exhibit high mitochondrial genetic diversity in southeastern Senegal. *Acta Chiropterologica* 13: 79-88.

Author contribution: 14%, I contributed towards study design, statistical analyses and writing.

Main innovation of the paper: Morphological and ecological differentiation sometimes does not coincide with phylogenetic divergence. We identified genetically distinct groups in Afrotropic bats from the *Hipposideros ruber* species complex (Chiroptera, Hipposideridae) in the phylogenetic reconstructions with maximum parsimony, ML and BI. Using sample classification based on the phylogeny, we were able to morphologically differentiate groups that may belong to separate taxa. We then tested alternative phylogenetic hypotheses, and the tests revealed that the group differentiation is not robust in the available data.

2.1.7 Kopecna O., Kubickova S., Cernohorska H., Cabelova K., Váhala J., **Martínková N.**, Rubes J. 2014. Tribe-specific satellite DNA in non-domestic Bovidae. *Chromosome Research* 22: 277-291.

Author contribution: 12%, I reconstructed the phylogenetic relationships and wrote the pertaining sections in the manuscript.

Main innovation of the paper: Satellite DNA in the centromers consists of repetitive sequences, where individual repetitions are relatively long and may contain signal of past speciation processes. We found high diversity in satellite DNA sequences in representatives of Bovidae (Artiodactyla). The resulting alignment incorporated many indels. We mined the phylogenetic information in indels by recoding the indels as a binary morphological character in a partitioned phylogenetic analyses with ML and BI. The information in indels revealed potential problem with homology of the satellite sequences in some taxa, but including indels in the phylogenetic reconstruction helped to resolve relationships between

satellite DNA sequences of the subfamily Antilopinae.

2.1.8 Wallace I. S., Shakesby A. J., Hwang J. H., Choi W. G., **Martínková N.**, Douglas A. E., Roberts D. M. 2012. *Acyrthosiphon pisum* AQP2: A multifunctional insect aquaglyceroporin. *BBA Biomembranes* 1818: 627-635.

Author contribution: 11%, I reconstructed phylogenetic relationships and wrote the pertaining sections in the manuscript.

Main innovation of the paper: With help from phylogenetic analysis, we were able to identify the newly discovered gene *ApAQP2* as an aquaporin-encoding gene in insects. We approached the task by designing the dataset representative of the diverse family of aquaporins. We included aminoacid sequences of known aquaporins in arthropods and chordates, and we analysed the data in a mixed model with indels coded as a binary partition. The BI analysis may become trapped in the tree space with short branches when analysing divergent sequences. The issue affected our analyses and the aquaporin phylogeny was too short. We corrected the branch lengths by estimating them with ML on the tree topology fixed to the consensus from the BI posterior trees.

2.2 Molecular dating papers

2.2.1 **Martínková N.**, McDonald R. A., Searle J. B. 2007. Stoats (*Mustela erminea*) provide evidence of natural overland colonisation of Ireland. *Proceedings of the Royal Society B-Biological Series* 274: 1387-1393.

Author contribution: 63%, my contribution included material collections from European museums and private collections, all laboratory experiments, data analyses and their interpretation, manuscript preparation.

Main innovation of the paper: Unless molecular dating of recent divergencies addresses mutation rate decay, the divergence times of events correlated with Pleistocene glaciations could be overestimated by orders of magnitude. To avoid the pitfalls, we estimated the mutation rate specifically for the model organism and expected time frame. We used time constraint priors for root height of one stoat population and we set them based on known paleontological data and geological events. Using Bayesian coalescence analysis, we estimated mutation rate for the given population and we applied the rate to date stoat populations in Europe. We inferred information on age and population size changes of island and continental populations of stoats.

2.2.2 Seifertová M., Bryja J., Vyskočilová M., **Martínková N.**, Šimková A. 2012. Multiple Pleistocene refugia and post-glacial colonization in the European chub (*Squalius cephalus*) revealed by combined use of nuclear and mitochondrial markers. *Journal of Biogeography* 39: 1024-1040.

Author contribution: 12%, I joined the project in the phase of manuscript review, when a reviewer requested additional analyses. I performed the molecular dating analysis of the available data and I wrote the respective sections of the manuscript.

Main innovation of the paper: Similarly as in stoats in paper 2.2.1 (Martínková et al., 2007), we expected the Pleistocene glaciations to influence distribution range changes and thus genetic diversity in fish. We used paleontological and biogeographic information to constrain the heights of the tree nodes. We constrained the tree root with the first known fossil of the European chub from Germany and a fossil of its ancestor from Greece. Other constrained nodes were ancestral to star-like phylogenies that characterise growing populations. We expected the sudden population growth coincided with climate amelioration after glaciations. The time constraints at different depth of the tree enabled us to estimate the divergence times for the European chub lineages.

2.2.3 **Martínková N.**, Barnett R., Cucchi T., Struchen R., Pascal M., Pascal M., Fischer M. C., Higham T., Brace S., Ho S. Y. W., Quéré J.-P., O'Higgins P., Excoffier L., Heckel G., Hoelzel A. R., Dobney K. M., Searle, J. B. 2013. Divergent evolutionary processes associated with colonization of offshore islands. *Molecular Ecology* 22: 5205-5220.

Author contribution: 9%, I collected all material from the Orkney Islands and participated in material collection in France, I performed laboratory experiments on the recent material, analysed and interpreted the data and co-wrote the paper.

Main innovation of the paper: Molecular dating that uses paleontological and geological time constraints assumes that the chosen events influenced directly and strongly the population of the model organism. We chose a direct, tip-dated molecular clock calibration reflecting the observed genetic changes in time to date the origin of the Orkney voles and the time of colonisation of the archipelago. We used fossil voles that were radiocarbon-dated and sequenced to calibrate the mutation rate with the Bayesian coalescence analysis. The precise and accurate mutation rate allowed us to test colonisation scenarios of islands in the North Sea and we interpreted the results with respect to the archeological findings from the Neolithic.

2.3 Codivergence papers

2.3.1 Irwin N. R., Bayerlová M., Missa O., **Martínková N.** 2012. Complex patterns of host switching in New World arenaviruses. *Molecular Ecology* 21: 4137-4150.

Author contribution: 36%, I participated in study design, supervised the student M. Bayerlová, contributed towards data analyses and manuscript preparation.

Main innovation of the paper: Using publically available data, we addressed a long standing assumptions that arenaviruses coevolved with their hosts. We reconstructed phylogenetic relationships between New World arenaviruses and all hosts and related taxa with available DNA sequences using ML and BI phylogenetic analyses. We computed contribution of individual associations to the overall codivergence between the parasite and host phylogenies using the ParaFit analysis and we tested clustering of arenavirus hosts on the rodent phylogeny with indices used in community ecology. We found that host species of some groups of arenaviruses are randomly distributed across rodent diversity in the Americas. Together with geographical modelling, our results suggest that medically-

important strains utilise host-switching between geographically, not phylogenetically close hosts.

2.3.2 Pečnerová P., Moravec J. C., **Martínková N.** 2015. A skull might lie: Modeling ancestral ranges and diet from genes and shape of tree squirrels. *Systematic Biology* 64: 1074-1088.

Author contribution: 23%, I designed the study, supervised the student P. Pečnerová, contributed to analyses and manuscript preparation.

Main innovation of the paper: We expanded the research questions about speciation explosion of tree squirrels in the Nearctic established in paper 2.1.4 (Pečnerová & Martínková, 2012). Here, we applied an innovative approach of contrasting codivergence of a molecular genetic phylogeny of tree squirrels with a tree characterising morphological similarity between taxa and compared that to a tree of food preferrences of the tree squirrels. We then modelled geographic spread of speciating lineages in the Nearctic. We found striking dissimilarities between the trees, where the current taxonomy corresponds to the morphological, not the phylogenetic relationships. Convergent evolution of skull shape in tree squirrels is likely a result of specialisation and diversification in utilising specific food sources, when the ancestors of contemporary tree squirrels crossed to South America after formation of the Isthmus of Panama.

2.4 Comparative phylogenetics papers

2.4.1 Zukal J., Bandouchova H., Bartonicka T., Berkova H., Brack V., Brichta J., Dolinay M., Jaron K. S., Kovacova V., Kovarik M., **Martínková N.**, Ondracek K., Rehak Z., Turner G. G., Pikula J. 2014. White-nose syndrome fungus: a generalist pathogen of hibernating bats. *PLoS ONE* 9: e97224.

Author contribution: 13%, I participated in study design, field research, data analyses and writing, and I supervised the student M. Dolinay.

Main innovation of the paper: White-nose syndrome leads to increased mortality in hibernating bats in the Nearctic, but Palearctic bats survive the infection better. We reconstructed a multilocus phylogeny of vespertilionid and minioperid bats in the ML framework from an alignment with 64% of missing data. We used comparative phylogenetics to study the interaction between the parasite and its hosts. We modelled clustering of the infected species on the bat phylogeny. We found that the probability that sister species will be infected decreases with improved sampling and that the pathogenic fungus is not species-specific. Ecological similarity of infected species varies as modelled with NRI and NTI indices on a neighbour joining tree based on the matrix characterising ecological and behavioural traits of hibernating bats. Using our models, we predicted *P. destructans* infection in five bat species, previously expected to be uninfected and even not susceptible. Research in subsequent years confirmed most of our predictions both in Palearctic (Zukal et al., 2016) and in Nearctic bat species (Bernard et al., 2015).

2.4.2 Zukal J., Bandouchova H., Brichta J., Cmokova A., Jaron K. S., Kolarik M., Kovacova V., Kubátová A., Nováková A., Orlov O., Pikula J., Presetnik P., Šuba J., Zahradníková A. Jr., **Martínková N.** 2016. White-nose syndrome without borders: *Pseudogymnoascus destructans* infection confirmed in Asia. *Scientific Reports* 6: 19829.

Author contribution: 17%, I co-designed the study and participated in material collection, data analyses and writing.

Main innovation of the paper: We found high prevalence of the *P. destructans* infection and the WNS in Palearctic hibernating bats. Fungal load at the wing surface of infected animals is comparable to that in the Nearctic animals that die of the disease. Number of UV fluorescent lesions diagnostic of WNS increases with increasing fungal load on the wing surface with and without the correction for phylogenetic non-independence of the data. Fungal load shows significant phylogenetic signal, but variance in number of UV fluorescent lesions is not significantly explained by the phylogeny. Our results indicate that in some (unrelated) bat species intensive fungal infection might not progress to development of numerous lesions and thus to severe disease. In conjunction with histopathological data, this means that the Palearctic bats are able to tolerate *P. destructans* infection.

2.4.3 Pikula J., Amelon S. K., Bandouchova H., Bartonička T., Berkova H., Brichta J., Hooper S., Kokurewicz T., Kolarik M., Köllner B., Kovacova V., Linhart P., Piacek V., Turner G. G., Zukal J., **Martínková N.** 2017. White-nose syndrome pathology grading in Nearctic and Palearctic bats. *PLoS ONE* 12: e0180435.

Author contribution: 33%, I contributed to all aspects of the study with the exception of design of the method scoring disease severity of WNS from histopathology of the wing punch biopsy. I worked on material collection, evaluation of the histopathologic slides, preparation of materials for naïve personnel, I analysed the data and co-wrote the manuscript.

Main innovation of the paper: The only method that unambiguously diagnoses WNS is a histopathologic evaluation of skin infected with *P. destructans*. We used UV light transillumination to detect the skin lesions in the field and to navigate the biopsy sampling of bat wings (Turner et al., 2014). We developed a methodology, how to use the biopsy punch to assess the WNS severity. From the analytical perspective, we validated the new method for use by experienced pathologists as well as by naïve evaluators who passed basic training. We analysed the relationships between the non-destructive methods of WNS research and the histopathological findings and we evaluated the relationship between the newly developed histopathology score to the quantitative measures of infection intensity in the phylogenetic context.

2.5 Outlier analyses papers

2.5.1 Jaron K. S., Moravec J. C., **Martínková N.** 2014. SigHunt: Horizontal gene transfer finder optimized for eukaryotic genomes. *Bioinformatics* 30: 1081-1086.

Author contribution: 17%, students K. S. Jaron and J. C. Moravec drove the development of the method detecting genomic islands in eukaryotic genomes and its software implementation. I tested the algorithms on model organisms, validated the results and wrote the manuscript.

Main innovation of the paper: DNA sequence variability along the chromosome complicates detection of horizontally transferred genes in eukaryotic genomes. We designed SigHunt, a method that uses local genomic signatures from informative tetranucleotides, i.e DNA composition from oligonucleotides that showed small variance in the host genome and were different from genomes of putative source organisms. We classify a genomic island as the DNA sequence region where multiple frequencies of tetranucleotides are outliers with respect to the respective frequencies in the surrounding genomic context. We quantify the outlier score as a discrete interval accumulative score from extreme values of the cumulative distribution function of the tetranucleotide frequency. That is a recently acquired genomic island will have markedly different DNA composition.

2.5.2 Škrabánek P., **Martínková N.** 2017. Extraction of outliers from imbalanced sets. In: Martínez de Pisón F., Urraca R., Quintián H., Corchado E. (Eds.) *Hybrid Artificial Intelligent Systems. HAIS 2017. Lecture Notes in Computer Science*, vol. 10334. pp. 402-412. Springer, Cham, DOI: 10.1007/978-3-319-59650-1_34.

Author contribution: 50%, the method is based on research by P. Škrabánek, with my minor contribution to all aspects of the study, including method development, analyses and writing.

Main innovation of the paper: We can identify exceptional observations using statistical methods if we have sufficiently large dataset, the exceptional observations are sufficiently similar to one another and we measured variables that characterise the exceptional observations sufficiently well to differentiate them from the normal observations. In biology, the requirements for statistical power to distinguish the exceptional observations pose a requirement that we know the exceptional observation exists and we know how to search for it. In other words, we test a hypothesis. The hypothesis-based research risks that we will omit processes fundamentally different from the focus of the study that have the potential to revolutionize the field. When we expect unknown processes to govern a biological system, we can apply outlier analysis to detect the observations generated by a rare mechanism. We designed a method, extraction of exceptional observations (EEO), that can detect individual samples in small, imbalanced sets that markedly differ from other samples. We use a threshold value where pairwise distance row sums rapidly increase as an indicator for outliers in a multidimensional variable space. The EEO method can single out observations that merit detailed study and evaluation of potential differences in mechanisms that generated the outlier observations.

References

[1] Aghova, T., Kimura, Y., Bryja, J., Dobigny, G., Granjon, L., & Kergoat, G. J. (2018). Fossils know it best: using a new set of fossil calibrations to improve the temporal phylogenetic framework of murid rodents (Rodentia: Myomorpha: Muroidea: Muridae). *Mol Phylogenet Evol, 128*, 98-111. doi: 10.1016/j.ympev.2018.07.017

[2] Anisimova, M., & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol, 55*, 539-552.

[3] Barrick, Jeffrey E., & Lenski, Richard E. (2013). Genome dynamics during experimental evolution. *Nat Rev Genet, 14*(12), 827-839. doi: 10.1038/nrg3564

[4] Biczok, R., Bozsoky, P., Eisenmann, P., Ernst, J., Ribizel, T., Scholz, F., . . . Stamatakis, A. (2018). Two C++ Libraries for Counting Trees on a Phylogenetic Terrace. *Bioinformatics*. doi: 10.1093/bioinformatics/bty384

[5] Bernard, R. F., Foster, J. T., Willcox, E. V., Parise, K. L., McCracken, G. F. (2015). Molecular detection of the causative agent of white-nose syndrome on Rafinesque's big-eared bats (*Corynorhinus rafinesquii*) and two species of migratory bats in the southeastern USA. *J Wildl Dis, 51*, 519-22. doi: 10.7589/2014-08-202

[6] Blokzijl, F., Janssen, R., van Boxtel, R., & Cuppen, E. (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med, 10*, 33. doi: 10.1186/s13073-018-0539-0

[7] Blomberg, S. P., Garland jr., T., & Ives, A.R. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution, 57*, 717-745.

[8] Bogusz, M., & Whelan, S. (2017). Phylogenetic tree estimation with and without alignment: New distance methods and benchmarking. *Syst Biol, 66*(2), 218-231. doi: 10.1093/sysbio/syw074

[9] Bowen, M. D., Peters, C. J., & Nichol, S. T. (1997). Phylogenetic analysis of the Arenaviridae: patterns of virus evolution and evidence for cospeciation between arenaviruses and their rodent hosts. *Mol Phylogenet Evol, 8*, 301-316. doi: 10.1006/mpev.1997.0436

[10] Butler, M. A., & King, A. A. (2004). Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *The American Naturalist, 164*, 683-695. doi: 10.1086/426002

[11] Campbell, V., Legendre, P., & Lapointe, F.-J. (2011). The performance of the congruence among distance matrices (CADM) test in phylogenetic analyses. *BMC Evol Biol, 11*, 64. doi: 10.1186/1471-2148-11-64

[12] Chan, C. X., Bernard, G., Poirion, O., Hogan, J. M., & Ragan, M. A. (2014). Inferring phylogenies of evolving sequences without multiple sequence alignment. *Scientific Reports, 4*, 6504. doi: 10.1038/srep06504

[13] Charrel, R. N., & de Lamballerie, X. (2003). Arenaviruses other than Lassa virus. *Antiviral Research, 57*(1-2), 89-100. doi: 10.1016/s0166-3542(02)00202-4

[14] Chernomor, O., von Haeseler, A., & Minh, B. Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. *Syst Biol, 65*(6), 997-1008. doi: 10.1093/sysbio/syw037

[15] Dobrin, B. H., Zwickl, D. J., & Sanderson, M. J. (2018). The prevalence of terraced treescapes in analyses of phylogenetic data sets. *BMC Evol Biol, 18*(1), 46. doi: 10.1186/s12862-018-1162-9

[16] dos Reis, M., Donoghue, P. C., & Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. *Nat Rev Genet, 17*(2), 71-80. doi: 10.1038/nrg.2015.8

[17] Drees, K. P., Lorch, J. M., Puechmaille, S. J., Parise, K. L., Wibbelt, G., Hoyt, J. R., ... Foster, J. T. (2017). Phylogenetics of a fungal invasion: origins and widespread dispersal of white-nose syndrome. *mBio, 8*, e01941-01917. doi: 10.1128/mBio.01941-17

[18] Dwivedi, B., & Gadagkar, S. R. (2009). Phylogenetic inference under varying proportions of indel-induced alignment gaps. *BMC Evol Biol, 9*, 211. doi: 10.1186/1471-2148-9-211

[19] Ekman, S., & Blaalid, R. (2011). The devil in the details: interactions between the branch-length prior and likelihood model affect node support and branch lengths in the phylogeny of the Psoraceae. *Syst Biol, 60*, 541-561. doi: 10.1093/sysbio/syr022

[20] Fink, S., Excoffier, L., & Heckel, G. (2007). High variability and non-neutral evolution of the mammalian *avpr1a* gene. *BMC Evol Biol, 7*, 176. doi: 10.1186/1471-2148-7-176

[21] Fountain-Jones, N. M., Pearse, W. D., Escobar, L. E., Alba-Casals, A., Carver, S., Davies, T. J., . . . Craft, M. E. (2017). Towards an eco-phylogenetic framework for infectious disease ecology. *Biol Rev Camb Philos Soc*. doi: 10.1111/brv.12380

[22] Gabriel, S. I., Jóhannesdóttir, F., Jones, E. P., & Searle, J. B. (2010). Colonization, mouse-style. *BMC Biology, 8*, 131. doi: 10.1186/1741-7007-8-131

[23] Hillis, D. M., Heath, T. A., & St John, K. (2005). Analysis and visualization of tree space. *Syst Biol, 54*(3), 471-482. doi: 10.1080/10635150590946961

[24] Ho, S. Y., Phillips, M. J., Cooper, A., & Drummond, A. J. (2005). Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol, 22*(7), 1561-1568. doi: 10.1093/molbev/msi145

[25] Ho, L. s. T., & Ané, C. (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst Biol, 63*, 397-408. doi: 10.1093/sysbio/syu005

[26] Horáček, I., Bartonička, T., Lučan, R. K., & Czech Bat Conservation Trust. (2014). Macroecological characteristics of bat geomycosis in the Czech Republic: Results of five years of monitoring. *Vespertilio, 17*, 65-77.

[27] Irwin, N. R., Bayerlová, M., Missa, O., & Martínková, N. (2012). Complex patterns of host switching in New World arenaviruses. *Mol Ecol, 21*(16), 4137-4150. doi: 10.1111/j.1365-294X.2012.05663.x

[28] Ives, A.R., Midford, P. E., & Garland jr., T. (2007). Within-species measurement error in phylogenetic comparative methods. *Syst. Biol., 56*, 252-270.

[29] Jackson, A. P., & Charleston, M. A. (2004). A cophylogenetic perspective of RNA-virus evolution. *Mol Biol Evol, 21*(1), 45-57. doi: 10.1093/molbev/msg232

[30] Jaron, K. S., Moravec, J. C., & Martínková, N. (2014). SigHunt: Horizontal gene transfer finder optimized for eukaryotic genomes. *Bioinformatics, 30*, 1081-1086.

[31] Karlin, S., & Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet, 11*, 283-290.

[32] Kolaczkowski, B., & Thornton, J. W. (2009). Long-branch attraction bias and inconsistency in Bayesian phylogenetics. *PLoS ONE, 4*, e7891. doi: 10.1371/

[33] Kopecna, O., Kubickova, S., Cernohorska, H., Cabelova, K., Vahala, J., Martinkova, N., & Rubes, J. (2014). Tribe-specific satellite DNA in non-domestic Bovidae. *Chromosome Res, 22*(3), 277-291. doi: 10.1007/s10577-014-9401-4

[34] Krasnovyd, V, Vetešník, L, Gettová, L, Civáňová, K, & Šimková, A. (2017). Patterns of parasite distribution in the hybrids of non-congeneric cyprinid fish species: Is asymmetry in parasite infection the result of limited coadaptation? *Int. J. Parasitol., 47*, 471-483.

[35] Kumar, S. (2005). Molecular clocks: four decades of evolution. *Nat Rev Genet, 6*(8), 654-662.

[36] Legendre, P., Desdevises, Y., & Bazin, E. (2002). A statistical test for host-parasite coevolution. *Syst Biol, 51*, 217-234.

[37] Lilley, T. M., Johnson, J. S., Ruokolainen, L., Rogers, E. J., Wilson, C. A., Schell, S. M., . . . Reeder, D. M. (2016). White-nose syndrome survivors do not exhibit frequent arousals associated with *Pseudogymnoascus destructans* infection. *Front Zool, 13*, 12. doi: 10.1186/s12983-016-0143-3

[38] Lorch, J. M., Meteyer, C. U., Behr, M. J., Boyles, J. G., Cryan, P. M., & Hicks, A. C. (2011). Experimental infection of bats with *Geomyces destructans* causes white-nose syndrome. *Nature, 480*, 376-378.

[39] Losos, J. B. (2008). Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecol Lett, 11*, 995-1003. doi: 10.1111/j.1461-0248.2008.01229.x

[40] Marshall, D. C. (2010). Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. *Syst Biol, 59*(1), 108-117. doi: 10.1093/sysbio/syp080

[41] Marston, D. A., Banyard, A. C., McElhinney, L. M., Freuling, C. M., Finke, S., de Lamballerie, X., . . . Fooks, A. R. (2018). The lyssavirus host-specificity conundrum-rabies virus-the exception not the rule. *Curr Opin Virol, 28*, 68-73. doi: 10.1016/j.coviro.2017.11.007

[42] Martinez-Aquino, A. (2016). Phylogenetic framework for coevolutionary studies: a compass for exploring jungles of tangled trees. *Curr Zool, 62*, 393-403. doi: 10.1093/cz/zow018

[43] Martínková, N., Barnett, R., Cucchi, T., Struchen, R., Pascal, M., Pascal, M., . . . Searle, J. B. (2013). Divergent evolutionary processes associated with colonization of offshore islands. *Mol Ecol, 22*(20), 5205-5220. doi: 10.1111/mec.12462

[44] Martínková, N., McDonald, R. A., & Searle, J. B. (2007). Stoats (*Mustela erminea*) provide evidence of natural overland colonization of Ireland. *Proc Biol Sci, 274*(1616), 1387-1393. doi: 10.1098/rspb.2007.0334

[45] Martínková, N., & Moravec, J. C. (2012). Multilocus phylogeny of arvicoline voles (Arvicolini, Rodentia) shows small tree terrace size. *Folia Zool., 61*, 254-267.

[46] McGuire, G., Denham, M. C., & Balding, D. J. (2001). Models of sequence evolution for DNA sequences containing gaps. *Mol Biol Evol, 18*, 481-490. doi: 10.1093/oxfordjournals.molbev.a003827

[47] Murray, G. G., Wang, F., Harrison, E. M., Paterson, G. K., Mather, A. E., Harris, S. R., . . . Welch, J. J. (2016). The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol Evol, 7*(1), 80-89. doi: 10.1111/2041-210X.12466

[48] Nascimento, F. F., dos Reis, M., & Yang, Z. (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nat Ecol Evol, 1*, 1446-1454. doi: 10.1038/s41559-017-0280-x

[49] Near, T. J., Bolnick, D. I., & Wainwright, P. C. (2005). Fossil calibrations and molecular divergence time estimates in centrarchid fishes (Teleostei: Centrarchidae). *Evolution, 59*, 1768-1782. doi: 10.1554/05-030.1.s1

[50] Near, T. J., & Sanderson, M. J. (2004). Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection. *Philos Trans R Soc Lond B Biol Sci, 359*, 1477-1483. doi: 10.1098/rstb.2004.1523

[51] Nelson, B. J., Andersen, J. J., & Brown, J. M. (2015). Deflating trees: improving Bayesian branch-length estimates using informed priors. *Syst Biol, 64*, 441-447. doi: 10.1093/sysbio/syv003

[52] Palmer, J. M., Drees, K. P., Foster, J. T., & Lindner, D. L. (2018). Extreme sensitivity to ultraviolet light in the fungal pathogen causing white-nose syndrome of bats. *Nat Commun, 9*, 35. doi: 10.1038/s41467-017-02441-z

[53] Pečnerová, P., & Martínková, N. (2012). Evolutionary history of tree squirrels (Rodentia, Sciurini) based on multilocus phylogeny reconstruction. *Zoologica Scripta, 41*, 211-219. doi: 10.1111/j.1463-6409.2011.00528.x

[54] Pečnerová, P., Moravec, J. C., & Martínková, N. (2015). A skull might lie: Modeling ancestral ranges and diet from genes and shape of tree squirrels. *Syst Biol, 64*, 1074-1088. doi: 10.1093/sysbio/syv054

[55] Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., & Baurain, D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol., 9*, e1000602. doi: 10.1371/ journal.pbio.1000602

[56] Pikula, J., Amelon, S. K., Bandouchova, H., Bartonicka, T., Berkova, H., Brichta, J., . . . Martínková, N. (2017). White-nose syndrome pathology grading in Nearctic and Palearctic bats. *PLoS ONE, 12*, e0180435. doi: 10.1371/journal.pone.0180435

[57] Pyron, R. A. (2015). Post-molecular systematics and the future of phylogenetics. *Trends Ecol Evol, 30*, 384-389. doi: 10.1016/j.tree.2015.04.016

[58] Rannala, B., Zhu, T., & Yang, Z. (2012). Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Mol Biol Evol, 29*, 325-335. doi: 10.1093/molbev/msr210

[59] Reeder, D. M., Frank, C. L., Turner, G. G., Meteyer, C.U., Kurta, A., Britzke, E. R., . . . Blehert, D.S. (2012). Frequent arousal from hibernation linked to severity of infection and mortality in bats with white-nose syndrome. *PLoS ONE, 7*, e38920. doi: 10.1371/journal.pone.0038920.g001

[60] Revell, L. J. (2010). Phylogenetic signal and linear regression on species data. *Methods Ecol Evol, 1*, 319-329. doi: 10.1111/j.2041-210X.2010.00044.x

[61] Revell, L. J., Harmon, L. J., & Collar, D. C. (2008). Phylogenetic signal, evolutionary process, and rate. *Syst Biol, 57*, 591-601. doi: 10.1080/10635150802302427

[62] Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences, 53*(1), 131-147. doi: 10.1016/0025-5564(81)90043-2

[63] Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., . . . Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol, 61*, 539-542. doi: 10.1093/sysbio/sys029

[64] Sanderson, M. J., McMahon, M. M., Stamatakis, A., Zwickl, D. J., & Steel, M. (2015). Impacts of terraces on phylogenetic inference. *Syst Biol, 64*, 709-726. doi: 10.1093/sysbio/syv024

[65] Sanderson, M. J., McMahon, M. M., & Steel, M. (2011). Terraces in phylogenetic tree space. *Science, 333*, 448-450. doi: 10.1126/science.1206357

[66] Schaack, S., Gilbert, C., & Feschotte, C. (2010). Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol Evol, 25*, 537-546. doi: 10.1016/j.tree.2010.06.001

[67] Simmons, Mark P. (2012). Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. *Cladistics, 28*, 208-222. doi: 10.1111/j.1096-0031.2011.00375.x

[68] Simoes, M., Breitkreuz, L., Alvarado, M., Baca, S., Cooper, J. C., Heins, L., . . . Lieberman, B. S. (2016). The evolving theory of evolutionary radiations. *Trends Ecol Evol, 31*, 27-34. doi: 10.1016/j.tree.2015.10.007

[69] Škrabánek, P., & Martínková, N. (2017). Extraction of outliers from imbalanced sets. In F. J. Martínez de Pisón, R. Urraca, H. Quintián & E. Corchado (Eds.), *Lecture Notes in Artificial Intelligence. Hybrid Artificial Intelligent Systems: 12th International Conference, HAIS 2017, La Rioja, Spain, June 21-23, 2017, Proceedings.* (Vol. 10334, pp. 402-412). Cham: Springer.

[70] Smith, G. J., Bahl, J., Vijaykrishna, D., Zhang, J., Poon, L. L., Chen, H., . . . Guan, Y. (2009). Dating the emergence of pandemic influenza viruses. *Proc Natl Acad Sci U S A, 106*, 11709-11712. doi: 10.1073/pnas.0904991106

[71] Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*, 1312-1313. doi: 10.1093/bioinformatics/btu033

[72] Symonds, M. R. E., & Blomberg, S. P. (2014). A Primer on Phylogenetic Generalised Least Squares. In L. Garamszegi (Ed.), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. Berlin, Heidelberg: Springer.

[73] Turner, G. G., Meteyer, C. U., Barton, H., Gumbs, J. F., Reeder, D. M., Overton, B., . . . Blehert, D. S. (2014). Nonlethal screening of bat-wing skin with the use of ultraviolet fluorescence to detect lesions indicative of white-nose syndrome. *J Wildl Dis, 50*, 566-73. doi: 10.7589/2014-03-058

[74] Wallace, I. S., Shakesby, A. J., Hwang, J. H., Choi, W. G., Martínková, N., Douglas, A. E., & Roberts, D. M. (2012). *Acyrthosiphon pisum* AQP2: a multifunctional insect aquaglyceroporin. *Biochim Biophys Acta, 1818*, 627-635. doi: 10.1016/j.bbamem.2011.11.032

[75] Webb, C. O., Ackerly, D. D., McPeek, M. A., & Donoghue, M. J. (2002). Phylogenies and Community Ecology. *Annual Review of Ecology and Systematics, 33*, 475-505. doi: 10.1146/annurev.ecolsys.33.010802.150448

[76] Wibbelt, G., Kurth, A., Hellmann, D., Weishaar, M., Barlow, A., & Veith, M. (2010). White-nose syndrome fungus (*Geomyces destructans*) in bats, Europe. *Emerg. Infect. Dis., 16*, 1237-1243.

[77] Xi, Z., Liu, L., & Davis, C. C. (2016). The impact of missing data on species tree estimation. *Mol Biol Evol, 33*, 838-860. doi: 10.1093/molbev/msv266

[78] Zuckerkandl, E., & Pauling, L. (1962). Molecular disease, evolution, and genic heterogeneity. In M. Kasha & B. Pullman (Eds.), *Horizons in biochemistry* (pp. 167-187). New York: Academic Press.

[79] Zukal, J., Bandouchova, H., Bartonička, T., Berková, H., Brack, V., Brichta, J., . . . Pikula, J. (2014). White-nose syndrome fungus: A generalist pathogen of hibernating bats. *PLoS ONE, 9*, e97224.

[80] Zukal, J., Bandouchova, H., Brichta, J., Cmokova, A., Jaron, K. S., Kolarik, M., . . . Martínková, N. (2016). White-nose syndrome without borders: *Pseudogymnoascus destructans* infection confirmed in Asia. *Scientific Reports, 6*, 19829. doi: 10.1038/srep19829

Index

Paper 2.1.1

Jaarola M., **Martínková N.**, Gündüz İ., Brunhoff C., Zima J., Nadachowski A., Amori G., Bulatova N. S., Chondropoulos B., Fraguedakis-Tsolis S., González-Esteban J., López-Fuster M. L., Kandaurov A. S., Kefelioğlu H., da Luz Mathias M., Villate I., Searle J. B. 2004. Molecular phylogeny of the speciose vole genus *Microtus* (Arvicolinae, Rodentia) inferred from mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution* 33: 647-663.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 33 (2004) 647–663
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Molecular phylogeny of the speciose vole genus *Microtus* (Arvicolinae, Rodentia) inferred from mitochondrial DNA sequences

Maarit Jaarola[a,*], Natália Martínková[b,c], İslam Gündüz[a,d], Cecilia Brunhoff[a], Jan Zima[b], Adam Nadachowski[e], Giovanni Amori[f], Nina S. Bulatova[g], Basil Chondropoulos[h], Stella Fraguedakis-Tsolis[h], Jorge González-Esteban[i], María José López-Fuster[j], Andrei S. Kandaurov[k], Haluk Kefelioğlu[d], Maria da Luz Mathias[l], Idoia Villate[i], Jeremy B. Searle[m]

[a] *Department of Cell and Organism Biology, Genetics Building, Lund University, Sölvegatan 29, SE-223 62 Lund, Sweden*
[b] *Institute of Vertebrate Biology, Academy of Science of the Czech Republic, Studenec 122, 675 02 Koněšín, Czech Republic*
[c] *Biodiversity Research Group, Department of Zoology, Charles University, Viničná 7, 128 44 Praha 2, Czech Republic*
[d] *Department of Biology, Faculty of Arts and Sciences, Ondokuz Mayis University, TR-55139 Kurupelit, Samsun, Turkey*
[e] *Institute of Systematics and Evolution of Animals, Polish Academy of Sciences, Sławkowska 17, 31-016 Kraków, Poland*
[f] *CNR—Institute of Ecosystem Studies, Via A. Borelli 50, 00161 Rome, Italy*
[g] *Severtzov Institute of Ecology and Evolution, Russian Academy of Sciences, 33 Leninsky Prospect, 117071 Moscow, Russia*
[h] *Section of Animal Biology, Department of Biology, University of Patra, GR-26 500 Rion, Patras, Greece*
[i] *Sociedad de Ciencias Aranzadi, Alto de Zorroaga, E20014 San Sebastián, Spain*
[j] *Department of Animal Biology, Faculty of Biology, University of Barcelona, Avinguda Diagonal 645, 08028-Barcelona, Spain*
[k] *Institute of Zoology, Georgian Academy of Science, 31 Chavchavadze Av., Tbilisi, Georgia*
[l] *Centre for Environmental Biology and Department of Animal Biology, Faculty of Science, University of Lisbon, Campo Grande, 1749-016 Lisbon, Portugal*
[m] *Department of Biology, University of York, PO Box 373, York YO10 5YW, UK*

Received 22 January 2004; revised 30 June 2004
Available online 16 September 2004

Abstract

Voles of the genus *Microtus* represent one of the most speciose mammalian genera in the Holarctic. We established a molecular phylogeny for *Microtus* to resolve contentious issues of systematic relationships and evolutionary history in this genus. A total of 81 specimens representing ten *Microtus* species endemic to Europe as well as eight Eurasian, six Asian and one Holarctic species were sequenced for the entire cytochrome *b* gene (1140 bp). A further 25 sequences were retrieved from GenBank, providing data on an additional 23, mainly Nearctic, *Microtus* species. Phylogenetic analysis of these 48 species generated four well-supported monophyletic lineages. The genus *Chionomys*, snow voles, formed a distinct and well-supported lineage separate from the genus *Microtus*. The subgenus *Microtus* formed the strongest supported lineage with two sublineages displaying a close relationship between the *arvalis* species group (common voles) and the *socialis* species group (social voles). Monophyly of the Palearctic pitymyid voles, subgenus *Terricola*, was supported, and this subgenus was also subdivided into two monophyletic species groups. Together, these groupings clarify long-standing taxonomic uncertainties in *Microtus.* In addition, the "Asian" and the Nearctic lineages reported previously were identified although the latter group was not supported. However, relationships among the main *Microtus* branches were not resolved, suggesting a rapid and potentially simultaneous radiation of a widespread ancestor early in the history of the genus. This and subsequent radiations discernible in the cytochrome *b* phylogeny, show the considerable potential of *Microtus* for analysis of

* Corresponding author. Fax: +46 46 14 78 74.
E-mail address: Maarit.Jaarola@cob.lu.se (M. Jaarola).

1055-7903/$ - see front matter
doi:10.1016/j.ympev.2004.07.015

historical and ecological determinants of speciation in small mammals. It is evident that speciation is an ongoing process in the genus and that the molecular data provides a vital insight into current species limits as well as cladogenic events of the past.

Keywords: Voles; *Microtus*; Arvicolinae; Mitochondrial DNA; Cytochrome *b*; Holarctic; Phylogeny; Speciation

1. Introduction

Voles of the genus *Microtus* Schrank (1798) are ecologically diverse and constitute the dominant herbivorous small mammals in many habitats in the Northern Hemisphere. Most species prefer open grasslands such as meadows and pastures but some species also occupy forests and highland habitats (Getz, 1985; Hoffmann and Koeppl, 1985; Mitchell-Jones et al., 1999; Nowak, 1999). *Microtus* represents one of the most speciose mammalian genera in the Holarctic, accounting for nearly 50 percent of the species of Arvicoline rodents (voles and lemmings) (e.g., Musser and Carleton, 1993). The genus represents one of the best known cases of rapid mammalian radiation resulting in about 65 extant species distributed throughout the Palearctic and Nearctic regions (Musser and Carleton, 1993; Chaline et al., 1999; Nowak, 1999). The shrew genus *Sorex* (Soricidae, Insectivora) is the only other mammalian genus that displays a comparable diversity across the Holarctic region, but *Sorex* is much older than *Microtus* (see Fumagalli et al., 1999).

The genus *Microtus* is apparently derived from the fossil genus *Allophaiomys*, which itself appears to have descended from *Mimomys* (Chaline and Graf, 1988; Conroy and Cook, 1999; Garapich and Nadachowski, 1996). Palaeontological data suggest that *Allophaiomys* radiated independently in northern Eurasia, central Asia-Himalayas and North America (Brunet-Lecomte and Chaline, 1991; Chaline et al., 1999). Until recently, the appearance of *Allophaiomys* was dated to approximately 2 million years ago (Mya) (Chaline and Graf, 1988), but a new finding of *Allophaiomys* from China traces the origin of the lineage back to 2.3–2.4 Mya (Zheng and Zhang, 2000). The majority of extant *Microtus* species, however, do not appear in the fossil record until Middle Pleistocene about 0.7–0.5 Mya (Chaline et al., 1999; Rabeder, 1986; Richmond, 1996) and it has even been suggested that some species trace their origin to the last glaciation (e.g., Brunet-Lecomte and Chaline, 1990; Chaline and Graf, 1988).

The genus *Microtus* displays a number of features that makes it ideal for evolutionary studies of speciation and the role of Quaternary glacial cycles on diversification. However, the phylogenetic relationships within *Microtus* and its closest relatives are uncertain and difficulties remain both in delimiting species and defining subgenera (e.g., Musser and Carleton, 1993; Nadachowski and Zagorodnyuk, 1996; Zagorodnyuk, 1990).

Current species' boundaries and phylogenetic relationships in *Microtus* rely mainly on morphology and karyotypes but these taxonomic characters have not been sufficient for solving all systematic questions in *Microtus.* The treatment of this genus is consequently characterized by inconsistency and lack of consensus (Musser and Carleton, 1993). For example, studies of dental and skull characters in *Microtus* demonstrate great intra-specific variability, a high incidence of adaptive convergence and many pairs of sibling species (Chaline et al., 1999; Chaline and Graf, 1988; Nadachowski and Zagorodnyuk, 1996; Zakrzewski, 1985). Karyotype evolution on the other hand appears largely uncoupled from morphological evolution (Baskevich, 1996; Suchentrunk et al., 1998). The *Microtus* karyotype varies between $2n = 17$–62 (Zagorodnyuk, 1990; Zima and Král, 1984) and exhibits one of the highest rates of karyotypic change in mammals (Maryama and Imai, 1981; Modi, 1987). Although some phylogenetic relationships can be deduced, especially from G-banded karyotypes (Mazurok et al., 2001; Meier et al., 1985, 1996; Modi, 1987; Orlov et al., 1983; Zagorodnyuk, 1990), the overall picture is that of extensive karyotypic variation among closely related species with no apparent phylogenetic trends (e.g., Akhverdyan et al., 1999; Chaline et al., 1999; Macholán et al., 2001; Modi, 1987). The systematic uncertainties in *Microtus* are perhaps best illustrated by the fact that the number of recognized species varies widely in different accounts (e.g., Gromov and Polyakov, 1992; Musser and Carleton, 1993; Nowak, 1999; Panteleyev, 1998; Zagorodnyuk, 1990) and that several new species have been proposed over recent years (Golenishchev et al., 2003; Kefelioğlu and Kryštufek, 1999; Kryštufek and Kefelioğlu, 2001; Yigit and Colak, 2002). Other unresolved issues concern the delineation and validity of higher-order relationships and species groups such as *Chionomys*, *Volemys*, *Lasiopodomys*, *Blanfordimys*, and *Terricola* that are alternately given either generic or subgeneric rank (e.g., Gromov and Polyakov, 1992; Musser and Carleton, 1993; Nowak, 1999).

Attempts to reconstruct the *Microtus* phylogeny using molecular approaches have been made using allozymes (e.g., Chaline and Graf, 1988; Gill et al., 1987; Graf, 1982; Mezhzherin et al., 1993, 1995; Suchentrunk et al., 1998), restriction enzyme analysis of mitochondrial DNA (DeBry, 1992) and RAPD analysis (Potapov et al., 1999) based on a limited number of *Microtus* species. However, the most comprehensive molecular

phylogenetic studies to date have been carried out by Conroy and Cook (1999, 2000a) and Conroy et al. (2001) using mitochondrial cytochrome *b* gene sequences. These studies comprised the North American endemics as well as some Asian and Eurasian *Microtus* species. In addition, Mazurok et al. (2001) analyzed cytochrome *b* sequences of four species of the *M. arvalis* group and Haring et al. (2001) inferred a phylogeny for species of the *M. multiplex* species group by analysis of mitochondrial D-loop sequences. Phylogeographic surveys of North American and Eurasian *Microtus* have demonstrated extensive cytochrome *b* variation within species (Brunhoff et al., 2003; submitted; Conroy and Cook, 2000b; Haynes et al., 2003; Jaarola and Searle, 2002, 2004). The results even suggest the existence of hitherto unidentified, cryptic species in *Microtus* (Jaarola and Searle, 2002, 2004; Hellborg, 2004), thereby pointing out the importance of within-species sampling for phylogeny reconstruction in *Microtus*.

The aim of this study is to further the understanding of phylogenetic relationships and evolutionary history in *Microtus* voles. We establish a molecular phylogeny for Palearctic *Microtus* including all the currently recognized European species (Mitchell-Jones et al., 1999) except *M. bavaricus*, as well as eight Eurasian and six Asian species. For this purpose we use DNA sequence analysis of the entire cytochrome *b* gene (1140 bp) since it evolves rapidly over the expected divergence times and because it enables us to incorporate previously published data on Nearctic species (Conroy and Cook, 1999, 2000a; Conroy et al., 2001). The combined phylogeny includes 48 out of 65 species of *Microtus*, with multiple representatives of many of the species. This near-comprehensive phylogeny of the genus allows us not only to resolve long-standing taxonomic controversies in *Microtus* but also to provide a phylogenetic tool to help understand speciation and species' radiations in small mammals.

2. Materials and methods

2.1. Samples

A total of 81 specimens representing 25 *Microtus* species were analyzed for variation in the mitochondrial cytochrome *b* gene (Table 1). All 25 species except for the Holarctic *M. oeconomus* are Palearctic; 10 are endemic to Europe, eight occur in Eurasia and six are restricted to Asia (cf. Gromov and Polyakov, 1992; Mitchell-Jones et al., 1999; Musser and Carleton, 1993). The only European species not analyzed is *M. bavaricus* that was considered extinct (Mitchell-Jones et al., 1999), but recently rediscovered in the northern Tyrol (Hutterer, 2001). The samples contain two species from the genus *Chionomys*, until recently often classified as *Microtus*, as well as six subgenera of *Microtus* (Table 1). Up to five individuals per taxon were surveyed to account for intra-specific variation. When possible, conspecific individuals were chosen from geographically distant localities. A total of 17 sequences representing *M. agrestis*, *M. oeconomus*, and *M. arvalis* have been published previously (Brunhoff et al., 2003; Haynes et al., 2003; Jaarola and Searle, 2002); their GenBank accession numbers are given in Table 1. In addition, 24 sequences representing 19 North American endemics and five Asian/Palearctic *Microtus* species (including one species that we sequenced, *M. gregalis*) were retrieved from GenBank (Conroy and Cook, 1999, 2000a; Conroy et al., 2001); for accession numbers see Table 2. A second *M.* (*Volemys*) *kikuchii* sequence was obtained from the whole mtDNA sequence in GenBank (AF348082, Lin et al., 2002).

2.2. DNA extraction, PCR amplification and sequencing

Total genomic DNA was extracted from frozen or ethanol preserved tail tips, ears, kidneys or liver. A few samples consisted of skulls collected from owl and raptor pellets. The Qiagen Dneasy Tissue kit was used for all types of samples. Pure mtDNA was isolated from two *M. rossiaemeridionalis* samples according to Jaarola and Tegelström (1995).

The primers used in this study are given in Table 3. For most specimens, the complete mitochondrial cytochrome *b* gene (1140 bp) was amplified in a single PCR reaction using the L14727-SP and H-15195-SP/H-ISO-SP primers complementary to glutamate and threonine/proline tRNA sequences, respectively. Alternatively, two to four separate amplifications that produced overlapping fragments were carried out. Due to the presence of nuclear copies of the cytochrome *b* gene in several *Microtus* species (DeWoody et al., 1999; Jaarola and Searle, 2004; Jaarola et al., in prep.), the majority of amplifications involved only *Microtus*- or species-specific primers designed by us (Table 3).

PCR amplification was carried out using Ampli*Taq* Gold DNA polymerase (Applied Biosystems). The PCR protocol for tissue samples consisted of an initial 7 min denaturation step at 95 °C, 30–35 cycles of denaturation at 94 °C for 1 min, annealing at 49 or 50 °C for 1 min and extension at 72 °C for 1–2 min, and a final 10-min extension step at 72 °C. PCR products were purified using the Qiagen QIAquick kit. The PCR protocol for skull samples is described in Jaarola and Searle (2004).

We sequenced each DNA fragment in both directions using a combination of PCR primers and internal primers (Table 3). Cycle sequencing reactions were carried out using the BigDye Terminator cycle sequencing kit (Applied Biosystems). Amplifications and sequencing reactions were performed in a PTC-200 thermal cycler (MJ Research). Sequencing products were purified using

Table 1
Microtus species, specimens, locations, and GenBank accession numbers of cytochrome *b* sequences

Subgenus (species group)	Species	Common name	No.	Location	Country	Accession Nos.
Agricola (*agrestis*)	*Microtus agrestis*	Field vole	1	Novosibirsk	Russia	AY167149
			2	Bonn	Germany	AY167210
			3	Sion, Valais	Switzerland	AY167160
			4	Pyrenees	Spain	AY167187
Microtus (*arvalis*)	*M. arvalis arvalis*	Common vole	1	Mantet, Pyrenees	Spain	AY220789
			2	Trento	Italy	AY220766
			3	Nuijamaa	Finland	AY220770
			4	Lauwersee	Netherlands	AY220778
Microtus (*arvalis*)	*M. arvalis obscurus*	(Altai vole)	1	Crimea	Ukraine	AY220762
			2	Kavka River, Serov	Russia	AY220764
			3	Neiva River	Russia	AY220765
			4	Sisian	Armenia	AY220761
			5	Ninotsminda	Georgia	AY220760
Agricola (*agrestis*)	*M. cabrerae*	Cabrera's vole	1	Alandron	Portugal	AY513788
			2	Idanha-a-Velha	Portugal	AY513789
Terricola (*subterraneus*/*majori*)	*M. daghestanicus*	Daghestan pine vole	1	Beniani	Georgia	AY513790
			2	Bağdaşan	Turkey	AY513791
			3	Handere	Turkey	AY513792
Microtus (*socialis*)	*M. dogramacii*		1	Ortaköy-Aksaray	Turkey	AY513793
			2	Amasya	Turkey	AY513794
			3	Boyalı Köyü-Amasya*	Turkey	AY513795
Terricola (*duodecimcostatus*)	*M. duodecimcostatus*	Mediterranean pine vole	1	Setùbal	Portugal	AY513796
			2	Algarve	Portugal	AY513797
Terricola (*savii*)	*M. felteni*	Balkan pine vole		Mt. Pelister, Begova Češma	Macedonia	AY513798
Terricola (*savii*)	*M. gerbei*	Pyrenean pine vole	1	Arrós, Vall d'Aran	Spain	AY513799
			2	Riba	Spain	AY513800
			3	Hecho	Spain	AY513801
			4	Hecho	Spain	AY513802
Stenocranius	*M. gregalis*	Narrow-headed vole	1	Yamal Peninsula	Russia	AY513803
Microtus (*socialis*)	*M. guentheri*	Guenther's vole	1	Gravia	Greece	AY513804
			2	Aqrabat	Syria	AY513805
			3	Locality unknown	Israel	AY513806
			4	Locality unknown	Israel	AY513807
Neodon	*M. juldaschi*	Juniper vole		Mazarsay	Kyrgyzstan	AY513808
Microtus (*arvalis*)	*M. kirgisorum*	Tien Shan vole	1	Balkash	Kyrgyzstan	AY513809
			2	Balkash	Kyrgyzstan	AY513810
Terricola (*multiplex*)	*M. liechtensteini*	Liechtenstein's pine vole		Anhovo	Slovenia	AY513811
Terricola (*duodecimcostatus*)	*M. lusitanicus*	Lusitanian pine vole	1	Burgos	Spain	AY513812
			2	Melgar de Fernamental	Spain	AY513813
Terricola (*subterraneus*/*majori*)	*M. majori*	Major's pine vole		Damar	Turkey	AY513814
Terricola (*multiplex*)	*M. multiplex*	Alpine pine vole	1	Staffarda, Piedmont	Italy	AY513815
			2	Trento	Italy	AY513816
			3	Lillaz	Italy	AY513817
			4	Méribel	France	AY513818
Pallasiinus (*oeconomus*)	*M. oeconomus*	Root vole (Tundra vole)	1	Ivvavik Nat. Park	Canada	AY220028
			2	Krasnoyarsk	Russia	AY220018
			3	Hamningberg	Norway	AY219988
			4	Texel	Netherlands	AY220006
Microtus (*arvalis*)	*M. rossiaemeridionalis*	Sibling vole	1	Kauhava	Finland	AY513819
			2	Svalbard	Norway	AY513820
			3	Gerede, Istanbul	Turkey	AY513821

Table 1 (*continued*)

Subgenus (species group)	Species	Common name	No.	Location	Country	Accession Nos.
			4	Erciyes Mt., Kayseri	Turkey	AY513822
			5	Kangal-Sivas	Turkey	AY513823
Terricola (*savii*)	*M. savii*	Savi's pine vole	1	Viterbo	Italy	AY513824
			2	Torino, Piedmont	Italy	AY513825
			3	Cerano, Piedmont	Italy	AY513826
			4	Fiume Freddo	Italy	AY513827
			5	Fiume Freddo	Italy	AY513828
Microtus (*socialis*)	*M. socialis*	Social vole	1	Iori River valley	Georgia	AY513829
			2	Iori River valley	Georgia	AY513830
			3	Reine	Iran	AY513831
Terricola (*subterraneus/majori*)	*M. subterraneus*	Common pine vole	1	Seli	Greece	AY513832
			2	Glocknerhaus	Austria	AY513833
			3	Çiğlikara	Turkey	AY513834
			4	Çiğlikara	Turkey	AY513835
			5	Güzyurdu	Turkey	AY513836
Terricola (*multiplex*)	*M. tatricus*	Tatra vole	1	Tretie Roháčske pleso lake	Slovakia	AY513837
			2	Smutná dolina valley	Slovakia	AY513838
			3	Veľká studená dolina valley*	Slovakia	AY513839
Terricola (*duodecimcostatus*)	*M. thomasi*	Thomas's vole	1	Agios Stefanos	Greece	AY513840
			2	Ano Kastritsi	Greece	AY513841
			3	Kyparissia	Greece	AY513842
			4	Itea	Greece	AY513843
			5	Trebinje, Herzegovina	Bosnia	AY513844
	Chionomys nivalis	Snow vole	1	Trento	Italy	AY513845
			2	Trento	Italy	AY513846
			3	Prvé Roháčske pleso lake	Slovakia	AY513847
			4	Queralbs, Girona	Spain	AY513848
			5	Saleh, As Suwayda	Syria	AY513849
	C. roberti	Robert's vole	1	Altındere Vadisi	Turkey	AY513850
			2	Datvisi	Georgia	AY513851

Subgenus and species group designation follows Musser and Carleton (1993) whose classification is largely based on the reclassification of Zagorodnyuk (1990).

* Type localities.

standard protocols and run in an ABI 310 or 3100 automated DNA sequencer (Applied Biosystems).

2.3. Phylogenetic analysis

Sequences were aligned and ambiguous bases resolved by eye using Sequencher v. 3.1.1 (Gene Codes Corp.). Nucleotide and amino acid composition was analyzed using MacClade v. 4.05 (Maddison and Maddison, 2000). Frequencies of transitions and transversions were estimated from maximum parsimony trees using MacClade.

The phylogenetic relationships among haplotypes were reconstructed using neighbor-joining (NJ), maximum parsimony (MP) and maximum likelihood (ML) algorithms implemented in PAUP* v. 4.0b10 (Swofford, 2002) as well as the Bayesian approach (Huelsenbeck et al., 2001) using the program MrBayes 3 (Ronquist and Huelsenbeck, 2003). The parsimony analyses were carried out ten times with the heuristic search approach using the TBR swapping algorithm, steepest descent option, random addition and 100–1000 replicates. Strict and 50% majority consensus trees were constructed from multiple equally parsimonious MP trees. The hierarchical likelihood ratio test (hLRT) and the Akaike information criterion (AIC) implemented in the computer program MODELTEST v. 3.06 (Posada and Crandall, 1998), were used to identify the most appropriate model of DNA substitution for our data. The model selected was the general time reversible model, GTR (Yang, 1994) with a gamma distributed shape parameter (α) of 0.8722 and the proportion of invariable sites (I) equaling 0.5320. This model, with parameters determined by MODELTEST, as well as several simpler substitution models, was implemented in the NJ and ML analyses. The ML tree search was conducted as described for MP but with the "as is" addition replicate (i.e. alphabetically by species). Relative stability of NJ and MP trees was assessed with bootstrap analysis using 10 000 and 1000 replicates, respectively. Bootstrapping of the ML tree could not be carried out due to the excessive computer capacity required.

Table 2
Microtus cytochrome *b* sequences retrieved from GenBank (Conroy and Cook, 1999, 2000a; Conroy et al., 2001; Lin et al., 2002)

Species	Accession Nos.
Microtus abbreviatus	AF163890
M. californicus	AF163891
M. canicaudus	AF163892
M. chrotorrhinus	AF163893
M. fortis	AF163894
M. gregalis	AF163895
M. guatemalensis	AF410262
M. kikuchii	AF163896
M. kikuchii	AF348082
M. longicaudus	AF187230
M. mexicanus	AF163897
M. middendorffi	AF163898
M. miurus	AF163899
M. montanus	AF119280
M. montebelli	AF163900
M. oaxacensis	AF410260
M. ochrogaster	AF163901
M. oregoni	AF163903
M. pennsylvanicus	AF119279
M. pinetorum	AF163904
M. quasiater	AF410259
M. richardsoni	AF163905
M. townsendii	AF163906
M. umbrosus	AF410261
M. xanthognathus	AF163907

We conducted Bayesian phylogenetic analyses using the GTR + I + G model with unequal base frequencies. Model parameters were estimated as part of the analysis. Altogether four independent runs, each with four Markov chain Monte Carlo (MCMC), were performed. All Bayesian analyses were initiated with random starting trees and run for two million generations. Trees were sampled every 10 generations. Log-likelihood scores of trees were plotted against generation time to determine the "burn-in" period and ensure that equilibrium log-likelihood values for different runs approached similar mean values (Huelsenbeck et al., 2002; Ronquist and Huelsenbeck, 2003). After discarding burn-in trees, we generated 50% majority rule consensus trees in PAUP for each single run and compared posterior probabilities for convergence among runs.

Since the correlation between and significance of tree support values as estimated by standard, nonparametric bootstrap values as opposed to posterior probabilities is not well understood and intensely debated (e.g., Huelsenbeck et al., 2002; Suzuki et al., 2002; Erixon et al., 2003), we follow the conservative approach of Leaché and Reeder (2002). Thus, only bootstrap values of $\geqslant 70\%$ (corresponding to 95% CI) and posterior probabilities of $\geqslant 95\%$ were considered significant.

Table 3
Primers used for PCR amplification and sequencing of the cytochrome *b* gene in *Microtus*

Primer	Sequence (5′–3′)	Reference
L14724B	CGAGATCTGAAAAACCATCGTTG	Kocher et al. (1989)
L14727-SP	GACAGGAAAAATCATCGTTG	Jaarola and Searle (2002)
L14841M	CCATCAAATATTTCATCATGATGAAA	Jaarola and Searle (2002)
L15162M2	GCTACGTACTTCCATGAGGACAAATATC	Jaarola and Searle (2002)
L15162Marv	G(CT)TACGT(CT)CTTCCATGAGGCCAAATATC	Haynes et al. (2003)
L15162MO	CTTCCATGAGGCCAAATATC	Brunhoff et al. (2003)
L15408M	GCAGACAAAATCCCGTTCCA	Jaarola and Searle (2002)
L15408Marv	GCAGACAAAATCCCATTCCA	Haynes et al. (2003)
L15408-SP	GCAGACAAA AT(TC)CC(AG)TT(TC)CA	Present study
H15177-SP2	AGGAGGTTTGT(AG)ATGACTG	Present study
H15177-SP3	A(AG)GAGGTTTGT(AG) ATNACTG	Present study
H15177Marv	AAGAGATTTGTAAT(CT)ACTG	Present study
H15177MO	AGGAGGTTTGTGATTACTG	Brunhoff et al. (2003)
H 15319Marv	AAAGGTGGACTAATACGAGG	Haynes et al. (2003)
H15348A-SP	GTTGGA(CT)CCTGTTTCGTG	Jaarola and Searle (2002)
H15408M	TGGAACGGGATTTTGTCTGC	Jaarola and Searle (2002)
H15408MO	TGGAATGGGATTTTGTCTGT	Brunhoff et al. (2003)
H15497-SP	T(AG)TAATT(AG)TCNGGGTCTCC	Present study
H15497-SP2	TGTAATT(AG)TCGGGGTCTCC	Present study
H15549M	AAGAGGAAATACCATTCTGGTTTAA	Jaarola and Searle (2002)
H15576M	GACCGTAAAATGGCGTAGG	Jaarola and Searle (2002)
H15576MO	GATCGTAGGATGGCGTAGG	Brunhoff et al. (2003)
H15915	AACTGCAGTCATCTCCGGTTTACAAGAC	Irwin et al. (1991)
H15915-SP	TTCATTACTGGTTTACAAGAC	Jaarola and Searle (2002)
H-ISO-M	AAGTAGTTTAATTAGAATGTCAG	Haynes et al. (2003)
H-PRO	AAGTAGTTTAATTAGAATATCAG	Brunhoff et al. (2003)
H-ISO-SP	AGTAGTTTAATTAGAATGTCAGC	Jaarola and Searle (2002)

SP: *Microtus* spp.; M: *M. agrestis*; Marv: *M. arvalis*; and MO: *M. oeconomus*.

Various combinations of lemmings (*Dicrostonyx*, *Lemmus*), *Clethrionomys rutilus* and *C. glareolus* and *Arvicola terrestris* were tested as outgroups. Since the position of *Arvicola* proved unstable and often tended to occur within *Microtus*, we discarded this option. The two *Chionomys* species analyzed, *C. nivalis* and *C. roberti*, were also tested but were too closely related to *Microtus* to function as outgroups. Our findings corroborate the conclusion of Conroy and Cook (1999, 2000a) that *Clethrionomys* is a sister taxon to *Microtus*, and we therefore used *C. glareolus* and/or *C. rutilus* as outgroups.

2.4. Tests of sequence saturation and a cytochrome b clock

To diagnose sequence saturation and homoplasy at the third position, we constructed a scatter plot of uncorrected pairwise transitions and transversion frequencies versus corrected pairwise divergences. Sequence divergence was corrected with the Jukes–Cantor (JC, Jukes and Cantor, 1969) model as well as a maximum likelihood model based on the GTR + G + I model of substitution. Total and net divergence (Dxy and Da) between species was estimated according to Nei (1987). We also tested for a molecular clock by comparing log likelihood scores of ML trees constructed with and without a molecular clock constraint (Felsenstein, 1988) in PAUP.

3. Results

GenBank accession numbers for the 64 new 1140 bp cytochrome *b* sequences representing 22 species are given in Table 1 (AY513788–AY513851) together with accession numbers of 17 sequences, representing three additional species, that we have published previously. We are confident that the new sequences represent the mitochondrial cytochrome *b* gene and not nuclear pseudogenes since they closely match previously reported *Microtus* sequences (Brunhoff et al., 2003; Conroy and Cook, 2000a,b; Conroy et al., 2001; Haynes et al., 2003; Jaarola and Searle, 2002) and because they did not display any of the anomalies typical for nuclear copies (cf. Bensasson et al., 2001; Mirol et al., 2000). We did, however, also obtain pseudogenes for the cytochrome *b* gene in a number of samples and taxa despite using only *Microtus*-specific primers. These findings will be reported elsewhere (Jaarola et al., unpublished).

3.1. Sequence composition and variation

The total, aligned data matrix included 100 cytochrome *b* haplotypes derived from 106 sequences representing 48 currently recognized species. A total of 504 (44%) variable sites were observed and 456 of these were informative for the parsimony analyses. More than one type of nucleotide substitution was observed in 199 (17%) sites and 83 (7%) of these displayed all four nucleotides. The majority of polymorphic sites were at third positions (364, 72%), followed by first positions (115, 23%) and second positions (25, 5%). Most substitutions were transitions (78%). The light strand nucleotide composition was characterized by a deficit of guanines (13%) similar to that described in North American *Microtus* species (Conroy and Cook, 2000a,b) as well as other mammals (e.g., Irwin et al., 1991). All but eight of the 86 variable amino acid residues and ten of the 136 amino acid replacements were located within the variable matrix and transmembrane region of cytochrome *b* (cf. Irwin et al., 1991; McClellan and McCracken, 2001).

Inter-specific distances varied between 4.2% and 18.0% using the JC model, whereas maximum likelihood distances estimated under the GTR + G + I model ranged from 4.5% to 51.6%. Corresponding intra-specific distances ranged up to 6.2% and 7.2% (*M. agrestis*). *M. arvalis*, *M. agrestis*, *M. daghestanicus*, *M. guentheri*, *M. oeconomus*, *M. savii*, *M. subterraneus*, and *C. nivalis* showed intra-specific divergences of 4–7%, values similar to net distance estimates between closely related species such as *M. duodecimcostatus–M. lusitanicus*, *M. dogramacii–M. guentheri*, and *M. liechtensteini–M. multiplex*. Some saturation occurred at third position transitions for divergences at the subgenus and genus levels (not shown), similar to that described in other studies of rodent cytochrome *b* (e.g., Yang and Yoder, 1999).

3.2. Phylogenetic results

The four phylogenetic methods used (MP, NJ, ML, and Bayesian) displayed trees with very similar topologies (Figs. 1–3). All methods discriminated two, well-supported major lineages corresponding to the two subgenera *Microtus* and *Terricola*. Each of the two subgenera was further divided into two well-supported sublineages. The two *Microtus* lineages previously described by Conroy and Cook (2000a) and Conroy et al. (2001), the "Asian" and the Nearctic, were also observed although the latter group was not supported. The two species of the genus *Chionomys* formed a highly supported branch outside the genus *Microtus*. Monophyly of the genus *Microtus* was supported although the bootstrap support depended much on the position of *M. gregalis*.

We obtained 48 MP trees (3744 steps, CI = 0.214) when using *C. rutilus* as outgroup, and 156 trees (3792 steps, CI = 0.212) when *C. glareolus* and *C. rutilus* were used as outgroups. The strict consensus tree of the 48 trees is given in Fig. 1. The same tree topology was recovered in the two consensus trees except for the position of *M. gregalis*. The 48 MP trees rooted with *C. rutilus* only

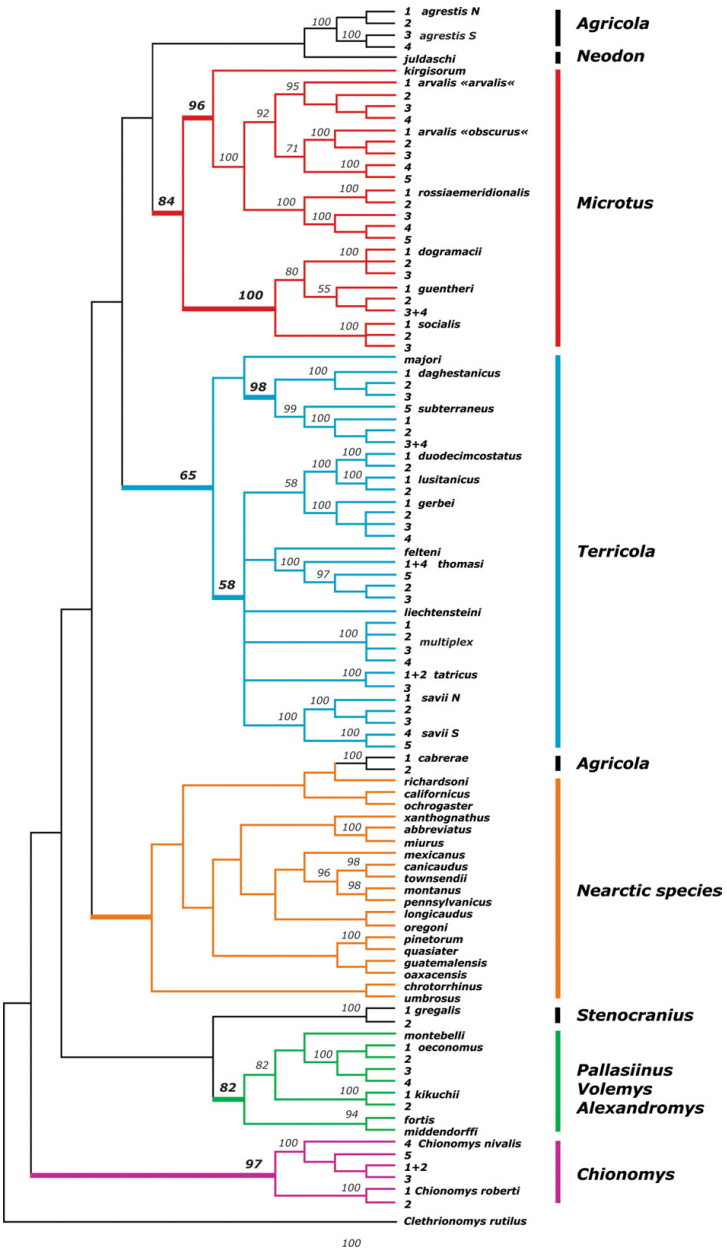


Fig. 1. Strict consensus tree of 48 maximum parsimony (MP) trees from cytochrome *b* haplotypes for 46 *Microtus* and two *Chionomys* species rooted with *Clethrionomys rutilus*. Numbers above branches denote percentage bootstrap resampling support (>50%) from 1000 replications.

differed because of unresolved intra-specific branches of *M. multiplex*, *M. gerbei*, *M. socialis*, and *M. dogramacii*. However, the importance of intra-specific sampling was clearly illustrated by the fact that removal of this intra-specific variation for one or several of these taxa resulted in fewer MP trees but also a decreased resolution of some

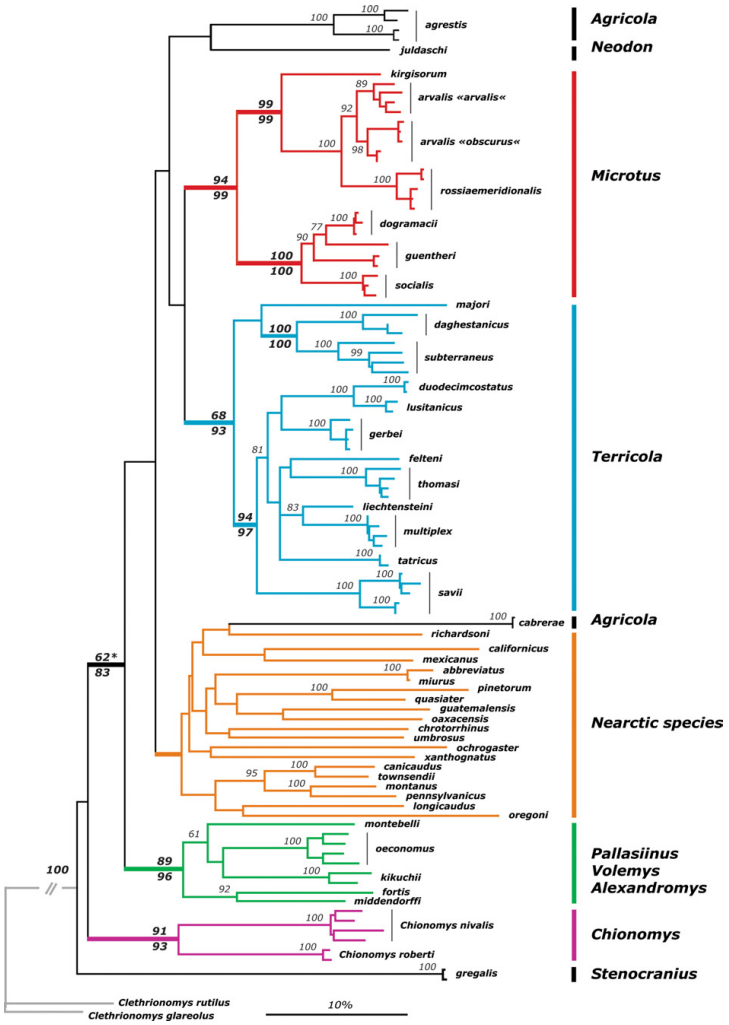


Fig. 2. Maximum likelihood (ML) tree based on GTR + G + I distances and rooted with *Clethrionomys glareolus* and *C. rutilus*. The tree shows the inferred phylogenetic relationships among 100 cytochrome *b* haplotypes representing 46 *Microtus* and two *Chionomys* species. Bootstrap resampling support from 10,000 iterations for a neighbor-joining (NJ) tree based on the JC model is listed above main branches. Estimates below branches show corresponding bootstrap support values when *M. gregalis*, *M. majori* and *M. cabrerae* are removed from the phylogenetic analyses. Only values greater than 50 percent are shown. *Percentage bootstrap support for *Microtus* monophyly in a NJ tree with *M. gregalis* included in *Microtus*. Subgenus designation follows Musser and Carleton (1993) (see Table 1).

of the deeper branches in the tree. The MP trees clearly contained an excessive degree of homoplasy, but all major lineages, except for the Nearctic, as well as many sublineages exhibited high bootstrap values, and relatively few MP trees were generated. Phylogenetic analyses using only first and second positions did not provide enough resolution (data not shown).

The NJ algorithm recovered the same topology independent of substitution model used, except for statistically unsupported differences in the positions of Nearctic species and basal taxa such as *M. agrestis*, *M. cabrerae*, *M. gregalis*, and *M. juldaschi*. The bootstrap values increased with the simplicity of the substitution model, the JC model generating the highest estimates.

The ML tree (Fig. 2) based on the GTR + G + I model showed the same topology as the MP and NJ trees. The ML tree constructed under a molecular clock constraint differed significantly from the unconstrained ML tree

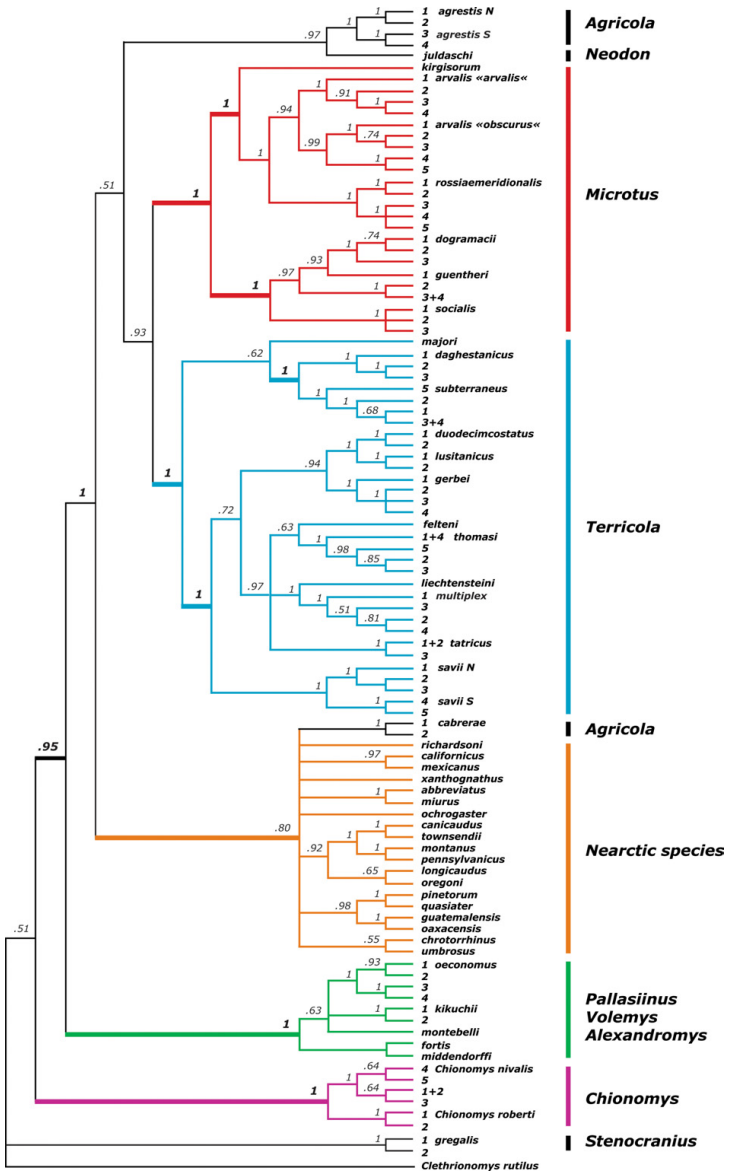


Fig. 3. Fifty percent majority rule consensus tree of 190, 000 trees from a Bayesian analysis of cytochrome *b* haplotypes for 46 *Microtus* and two *Chionomys* species rooted with *Clethrionomys rutilus*. Numbers above branches represent posterior probability values (>0.50).

when outgroups were included ($\chi^2 = 140.4$, df = 100, $P < 0.001$), but the difference was not significant when outgroups were excluded ($\chi^2 = 120.5$, df = 98, $P > 0.05$).

The Bayesian analyses reached stationarity around 100,000 generations (equaling 10,000 saved trees), so that the last 190,000 trees were used to compute a majority rule consensus tree for each run. The four independent analyses converged on similar log-likelihood values and the mean ln L score for the posterior distribution of trees was −17674. Topology of consensus trees, values of posterior probabilities and parameter estimates were highly similar in all four analyses. The

marginal probabilities of the nucleotide frequencies were similar to those obtained with MODELTEST. The average gamma distribution shape parameter (α) was 0.80 and the proportion of invariable sites (I) averaged 0.48. The lineages and sublineages also supported by standard, nonparametric bootstrapping values $\geqslant 70\%$, calculated using MP and/or NJ (Figs. 1 and 2), all showed posterior probabilities of 100% (Fig. 3). In addition, significance levels around 95% were obtained for a few branches not supported by bootstrapping (Fig. 3).

The long branches of *M. agrestis*, *M. juldaschi* and, especially, *M. gregalis* and *M. cabrerae* were basal and only formed non-significant associations that were dependent on the taxa present and the substitution model used. The apparent relationship of the Iberian endemic *M. cabrerae* with the Nearctic species represents an example of this type of clustering (Figs. 1 and 3). The Bayesian analyses supported the grouping of *M. agrestis* with *M. juldaschi* with a significance level of 97% (Fig. 3), but this was not supported by bootstrapping (Figs. 1 and 2). The position of *M. gregalis* was especially unstable. Within the *Terricola* lineage, the position of the single *M. majori* sequence could not be determined. Removal of the unstable taxa *M. gregalis*, *M. cabrerae*, and *M. majori* increased the bootstrap support for most major lineages in the NJ tree so that all except the Nearctic branch generated bootstrap values over 90% as well as high (83%) support for *Microtus* monophyly (Fig. 1).

4. Discussion

4.1. The cytochrome b tree and Microtus taxonomy

Although there are many issues relating to details of particular relationships, overall there is much concordance between the *Microtus* cytochrome *b* tree and current taxonomy based on morphology and karyotype. Four monophyletic lineages with good bootstrap support and 100% posterior probabilities were identified (Figs. 1–3). First, the two species of snow voles, genus *Chionomys*, formed a lineage separate from *Microtus*. Second, species of the subgenus *Microtus* formed a well-supported lineage with two highly supported sublineages representing the *arvalis* species group (common voles) and the *socialis* species group (social voles). Third, monophyly of the Palearctic pitymyid voles, i.e. subgenus *Terricola* was supported. Fourth, the "Asian" lineage previously reported by Conroy and Cook (2000a) and Conroy et al. (2001) was identified. All the Nearctic species clustered as previously described by Conroy and Cook (2000a) and Conroy et al. (2001) but this grouping was not well supported. The following discussion will deal with phylogenetic relationships, biogeographic scenarios and species delimitations in *Microtus*. Data on molecular evolution involving estimation of a molecular clock and datings will be reported elsewhere (Jaarola et al., unpublished). Our data support the analyses of Conroy and Cook (2000b) suggesting rapid cytochrome *b* evolution in *Microtus*.

4.2. Genus Chionomys (snow voles)

Chionomys nivalis and *C. roberti* form a distinct and well-supported clade separate from *Microtus* (Figs. 1–3). Thus, our cytochrome *b* results corroborate the ranking of *Chionomys* as a genus separate from *Microtus*. The genus *Chionomys* was previously included in *Microtus* but recent analyses have suggested separation (reviewed in Musser and Carleton, 1993; Nadachowski, 1991). According to palaeontological data, *Chionomys* represents an early split, 1.3–1.5 Mya, from *Allophaiomys* or, possibly, *Mimomys*, the ancestor of *Allophaiomys* (Horáček, pers. comm.). The distant relationship in cytochrome *b* between *C. nivalis* and *C. roberti* is also in accord with fossil data (cf. Nadachowski, 1991).

4.3. Genus Microtus

The relationship among the main *Microtus* branches was not resolved and the data indicate that Eurasian species are not basal to North American endemics as suggested by Conroy and Cook (2000a). This "hard" polytomy is likely to indicate an early burst of rapid diversification of a widespread ancestor resulting in the nearly simultaneous appearance and radiation of major *Microtus* lineages in Eurasia, central Asia-Himalayas and North America as advocated by palaeontologists (Brunet-Lecomte and Chaline, 1991; Chaline et al., 1999). We cannot, however, entirely dismiss the added effects of a high degree of homoplasy at third positions in the cytochrome *b* gene confounding the resolution of basal relationships (cf. Conroy and Cook, 1999; Reed and Sperling, 1999). Furthermore, some relationships among subgenera were indicated by Bayesian inference (Fig. 3) but not supported by bootstrapping of NJ or MP trees (Figs. 1 and 2).

4.4. Subgenus Microtus

The *Microtus* subgenus represents the strongest supported lineage in the cytochrome *b* tree, displaying a close relationship between the *arvalis* species group (common voles) and the *socialis* species group (social voles) *sensu* Zagorodnyuk (1990). The close phylogenetic relationships among the morphologically cryptic species in the *arvalis* group are concordant with the cytochrome *b* studies of Haynes et al. (2003) and Mazurok et al. (2001) involving members of this group. Mazurok et al. (2001) also included *M. transcaspicus* but since the sequence was not submitted to GenBank, we could not include it in our analyses.

The systematics of social voles has proven much more complex than previously thought (see Kefelioğlu and Kryštufek, 1999; Kryštufek and Kefelioğlu, 2001). Only recently a number of new species have been suggested (Golenishchev et al., 2003; Kefelioğlu and Kryštufek, 1999; Kryštufek and Kefelioğlu, 2001; Yigit and Colak, 2002). Besides the established species, *M. socialis* and *M. guentheri*, we also included sequences of the newly described *M. dogramacii* (Kefelioğlu and Kryštufek, 1999) in our analyses. The cytochrome *b* data demonstrate a recent divergence of *M. dogramacii* ($2n = 48$) from *M. guentheri* ($2n = 54$). The NJ, ML and Bayesian trees indicated a paraphyletic relationship (Figs. 2 and 3) but the MP analyses supported reciprocal monophyly (Fig. 1). However, the situation is more complex in that our *M. guentheri* specimens from Syria and Israel may be more appropriately considered *M. irani* (cf. Kefelioğlu and Kryštufek, 1999; Kryštufek and Kefelioğlu, 2001). The distribution of *M. irani* remains, however, uncertain (Kryštufek and Kefelioğlu, 2001; Mitchell-Jones et al., 1999). Overall, it is clear that a much more detailed molecular investigation of the phylogenetic relationships among social voles is warranted.

4.5. Subgenus Terricola (ground voles)

The cytochrome *b* tree supports Brunet-Lecomte and Chaline (1992), Chaline and Graf (1988), and Zagorodnyuk (1989) in the separation of pitymyine forms into Nearctic (*Pitymys*) and Palearctic (*Terricola*) components (Figs. 1–3). The monophyly of the subgenus *Terricola* was recently questioned by Kryštufek et al. (1996) who claimed that it constituted an artificial group of unrelated convergently evolved species with no shared apomorphies. However, our molecular data demonstrate that *Terricola* species do share a common ancestor. The Bayesian analyses yielded strong supported for *Terricola* (Fig. 3), but the bootstrap support was relatively low (Figs. 1 and 2). Removal of *M. majori* from the analyses, however, increased the bootstrap values drastically for the whole group as well as the two subgroups within *Terricola* (Fig. 2). Analysis of additional *M. majori* sequences might stabilize the group, but the results may also indicate that the Asian endemic *M. majori* represents a separate evolutionary lineage (see below).

The cytochrome *b* data strongly support two monophyletic species groups within *Terricola*: one subgroup with *M. subterraneus* and *M. daghestanicus*, species with ranges extending to Asia Minor, and the other subgroup consisting of the European endemics. The cytochrome *b* phylogeny does not agree fully with the prevailing species groups within *Terricola* (cf. Table 1). However, extensive karyotypic and morphological polymorphism in *Terricola* has spawned many alternative classifications (e.g., Chaline et al., 1999; Kratochvíl and Král, 1974) and our data imply that yet another systematic revision is necessary.

Microtus majori, *M. subterraneus*, and *M. daghestanicus* are believed to have diverged recently (reviewed in Macholán et al., 2001). Our data fully support a sister relationship between *M. subterraneus* and *M. daghestanicus* but the position of *M. majori* remains uncertain (see above)—even in the Bayesian analyses. This result is somewhat surprising since Macholán et al. (2001) found a close allozymic relationship between *M. majori* and *M. subterraneus*. Our results are, however, in accordance with Zagorodnyuk (1990) who considered *M. majori* the sole member of its own species group.

Species' relationships in the European subgroup of *Terricola* are poorly resolved, although, again, the Bayesian analyses support some groupings not recognized by MP or NJ bootstrap analysis (Fig. 3). Since these species are so closely related, the hard polytomy observed cannot be ascribed to saturation in cytochrome *b*. Instead, strong bootstrap support above and below unresolved polytomies indicate a rapid radiation involving nearly simultaneous diversification of many lineages (cf. Conroy and Cook, 1999; Lessa and Cook, 1998). Such a surge of speciation could have occurred during a single or, more likely, a few consecutive glacial periods by geographic isolation of small and genetically differentiated populations in different glacial refugia as suggested by Chaline (1987). The relative importance of Mediterranean peninsulas as opposed to more northern mountain areas as "speciation traps" deserves further attention (cf. Chaline, 1987; Bilton et al., 1998; Martínková and Dudich, 2003). In this context it is noteworthy that all taxa in this subgroup except *M. savii* (see below) seem to harbor little intra-specific variation.

4.6. Subgenera Pallasiinus, Alexandromys, and Volemys (the "Asian" lineage)

The taxonomic validity of the subgenera *Pallasiinus*, *Alexandromys*, and *Volemys* was not supported by our cytochrome *b* analysis since their representatives formed a strongly supported monophyletic group. This "Asian" group was previously described by Conroy and Cook (2000a). Our addition of within-species sequences and more species to the *Microtus* tree has significantly increased the support for this lineage. The Japanese *M. montebelli* and Taiwanese *M. kikuchii* are sister species to the Holarctic *M. oeconomus* and constitute examples of allopatric speciation on islands. The results are supported by chromosome data. Thus, pairing of the X and Y chromosome in *Microtus* meiosis seems to be a rare lineage-specific phenomenon that can be used in reconstructing systematic relationships in the genus (Mekada et al., 2002; Megías-Nogales et al., 2002). To

date, X–Y pairing is only reported for the three species forming the Asian lineage (see Mekada et al., 2002) as well as *Chionomys nivalis* (Megías-Nogales et al., 2002) and two species representing the subgenera *Lasiopodomys* and *Neodon* not included here (Gu et al., 1999; Mekada et al., 2002).

4.7. Subgenera Agricola, Neodon, and Stenocranius

Microtus agrestis, *M. cabrerae*, *M. juldaschi*, and *M. gregalis* were placed basal in the phylogenetic analyses and did not form significant associations with other species. For example, the association of *M. cabrerae* with the Nearctic species was not supported by either bootstrapping or posterior probabilities (Figs. 1–3) but most probably due to long-branch attraction (e.g., Hendy and Penny, 1989). These four, basal species represent three subgenera—*Agricola*, *Neodon*, and *Stenocranius*—characterized by few and ancestral species. Our data suggest that *M. agrestis* and *M. cabrerae* should not both be placed in the subgenus *Agricola* since they do not show any sister relationship. Especially the *M. cabrerae* lineage seems to be either very old or has undergone accelerated evolution in cytochrome *b*. Actually, *M. cabrerae* displays morphological characters that are archaic (Gromov and Polyakov, 1992), and Chaline (1972) described a new subgenus, *Iberomys*, for the fossil vole *Microtus* (*Iberomys*) *brecciensis*, a direct ancestor of *M. cabrerae*. Altogether, our data strongly support the classification of *M. cabrerae* in the separate subgenus *Iberomys*.

Microtus juldaschi belongs to the subgenus *Neodon* also containing *M. irene* and *M. sikimensis* (Musser and Carleton, 1993; Zagorodnyuk, 1990). *Neodon* as well as the subgenera *Blanfordimys* and *Phaiomys* (not analyzed) are considered old Pleistocene relicts that probably descended directly from the *Allophaiomys* stock (Nadachowski and Garapich, 1998; Nadachowski and Zagorodnyuk, 1996). Consequently, the position of *M. juldaschi* in the phylogenetic tree is expected to be basal, in line with our result.

Both our data and those of Conroy and Cook (2000a) indicate that the subgenus *Stenocranius* is a polyphyletic and artificial group as the Asian *M. gregalis* does not cluster with the North American *M. miurus* and *M. abbreviatus*. Nor does *M. gregalis* cluster with *M. middendorffi* as suggested by the alternative classification of Zagorodnyuk (1990). Thus, the morphological similarity of these species is most probably due to adaptive convergence as suggested by Chaline et al. (1999) and Conroy and Cook (2000a). *M. gregalis* is by far the most divergent *Microtus* species in our cytochrome *b* data set and its position is unclear even in relation to *Chionomys* (cf. Figs. 1–3). According to palaeontological data, *M. gregalis* represents an early split from the *Allophaiomys* stock (Rekovets and Nadachowski, 1995). The support for *Microtus* as a bona fide taxonomic group is mainly influenced by the instability of *M. gregalis*. Thus, in order to fully evaluate monophyly of the genus *Microtus*, the position of *M. gregalis* needs to be determined. In addition, the position of the genus *Arvicola* needs to be evaluated.

4.8. Species limits in Microtus?

Our data confirm that the genus *Microtus* contains many closely related species as well as many species that are characterized by extensive intra-specific variation (Fig. 2). Consequently, there is an overlap between inter- and intra-specific cytochrome *b* distances around 4–8%, similar to what has been described in other mammals (e.g., Avise, 2000; Bradley and Baker, 2001). These data corroborate the notion that genetic distances and/or reciprocal monophyly in mtDNA are highly uncertain criteria for delimiting closely related species (Bradley and Baker, 2001; Hudson and Coyne, 2002; Hudson and Turelli, 2003; Nichols, 2001; Rosenberg and Nordborg, 2002). Moreover, the overlap in cytochrome *b* divergence between and within currently recognized species demonstrates that speciation is an ongoing process in *Microtus* and that many taxa and species groups offer a huge potential for molecular, evolutionary research, particularly with regards to phylogeography and speciation (cf. Barraclough and Nee, 2001).

Examples of sister species that show low cytochrome *b* divergence, but do represent widely accepted species include *M. duodecimcostatus—lusitanicus* and *M. arvalis—rossiaemeridionalis* (Fig. 1). The cytochrome *b* distance of 4–5% between *M. duodecimcostatus* and *M. lusitanicus* is among the lowest recorded for *Microtus* species, the exception being *M. miurus—M. abbreviatus* that differ by only 1.5%. However, the latter two taxa are probably conspecific (Conroy and Cook, 2000a). The taxonomic rank of *M. lusitanicus* has varied in recent years but it is now considered a species (reviewed in Spitzenberger et al., 2000), its ranking validated by sterility of male F_1 hybrids between *M. duodecimcostatus* and *M. lusitanicus*.

The sibling species *M. arvalis* ($2n = 46$) and *M. rossiaemeridionalis* ($2n = 54$) are closely related, demonstrating a divergence of only 6–8% in cytochrome *b*. However, hybrids between the two taxa are sterile (Meier et al., 1985). The status of the eastern '*obscurus*' taxon in the *arvalis* group is unclear. While some authors regard this taxon as a species, *M. obscurus*, separate from the western *M. arvalis*, other authors describe the taxa as two karyotypic forms, *M. arvalis* '*obscurus*' and *M. arvalis* '*arvalis*'. We have used the latter classification since hybridization between the two taxa appears to occur in the wild (Bulatova, unpublished) and hybridization studies have shown that F_1, F_2 and subsequent backcrosses are fertile, albeit with lowered fertility (Malygin and Panteleichuk, 2003). The cytochrome *b* data imply a re-

cent split between '*arvalis* ' and '*obscurus*'; the divergence is only 2–4%. For references and a recent discussion based on cytochrome *b* data see Haynes et al. (2003).

Other species with unclear status accompanied by low cytochrome *b* divergence include *M. liechtensteini* and *M. atticus. M. liechtensteini* seems to represent a cytochrome *b* lineage distinct from *M. multiplex* (Figs. 1–3). Our observation is consistent with the results obtained by Haring et al. (2001) who conducted a more extensive study of the Alpine voles of the *M. multiplex* complex using mitochondrial D-loop sequences. However, additional data on hybridization and fertility of offspring in the *M. multiplex* complex are needed to fully evaluate the taxonomic ranking of *M. liechtensteini.* The *M. thomasi* samples 1–3 represent the karyotype form '*atticus*' previously ascribed to a separate species, *M. atticus*. Our data corroborate the present ranking of '*atticus*' as a mere karyotype form of *M. thomasi* (references in Tsekoura et al., 2002).

Species with high intra-specific cytochrome *b* variation include *M. savii*, *M. agrestis*, *M. daghestanicus*, *M. oeconomus*, *M. subterraneus*, and *C. nivalis* (Fig. 1). The subspecies *M. savii savii* in northern and central Italy and *M. s. brachycercus* in southern Italy differ in karyotype (Galleni et al., 1992) and because the male F_1 hybrids are sterile, Galleni et al. (1994) suggested that the two subspecies should be elevated to species rank. Our data set is limited, but indicates a recent mtDNA split between southern and central-northern *M. savii*, with cytochrome *b* divergences of 4–5%. Recent molecular data on *M. agrestis* demonstrate a south-north split in Europe, with a net divergence of 5.2% in cytochrome *b* and 0.7% in the X and Y chromosome, all three genealogies exhibiting reciprocal monophyly. These data suggest that the southern *M. agrestis* represent a new, morphologically and karyotypically cryptic species (Jaarola and Searle, 2002, 2004; Hellborg, 2004).

Another species with high intra-specific variation is *M. daghestanicus* that exhibits a highly divergent haplotype from Georgia differing by 4% from the two Turkish haplotypes. Georgian specimens are sometimes ascribed to *M. nasarovi*, but *M. daghestanicus* and *M. nasarovi* differ in karyotype ($2n = 52/54$ and 38, respectively), habitat preference and distribution range (Bukhnikashvili and Kandaurov, 2002). Since the Georgian *M. daghestanicus* originated from the eastern part of the country, i.e. outside the distribution range of *M. nasarovi*, and the Turkish specimens both carried a $2n = 54$ karyotype (Macholán, pers. comm.), the large cytochrome *b* divergence observed between Georgian and Turkish *M. daghestanicus* is likely to reflect intra-specific variation.

Microtus oeconomus is divided into four divergent cytochrome *b* lineages differing by net distances up to 3.5% (Brunhoff et al., 2003). The largely allopatric distribution of these lineages and inferences on their late Quaternary history is presented in Brunhoff et al. (2003, submitted). Yet another species that is characterized by very high cytochrome *b* distances between haplotypes is *M. subterraneus*; one of the haplotypes from Turkey differed from the other by 6–7%. Finally, *C. nivalis*, a species with fragmented distribution over Europe and Middle East mountain ranges, exhibited intra-specific distances up to 4%. Studies of morphological characters (Kryštufek, 1999; Nadachowski, 1991) and allozymes (Filippucci et al., 1991) have also demonstrated much diversity in this species.

5. Conclusions

'Small mammals' are far more speciose than their larger relatives, and it is interesting to speculate on the reasons for this (Searle, 1996). However, before the present study, there have been few detailed attempts to obtain molecular phylogenies for particularly species-rich genera of small mammals. Our study, through its coverage of almost all European and North American species, as well as many from Asia, provides a fascinating insight into *Microtus*, the most speciose genus of Arvicoline rodents and one of the most speciose genera of all mammals (Nowak, 1999). The cytochrome *b* phylogeny demonstrates species' radiations at a variety of temporal and spatial scales. In a temporal sense, an early rapid radiation about 2 Mya appears to have generated the major subgenera, which then subsequently radiated further to generate the variety of extant species. Moreover, the currently recognized species are not static forms, but are clearly differentiating and contain cryptic entities that may in many cases best be considered species (e.g., within *M. agrestis*: Jaarola and Searle, 2004). In a spatial sense, the radiations of subgenera and equivalent groupings of *Microtus* have occurred in different geographic areas as evidenced by the current geographic ranges of major cytochrome *b* lineages.

Now that a molecular phylogeny of *Microtus* is available, there is an opportunity to use it to understand the ecological and historical circumstances that lead to speciation of small mammals (cf. Searle, 1996). The phylogeny can also be used for a range of future comparative studies in ecology, behaviour, physiology, parasitology etc. It will, for instance, be possible to use the phylogeny for coevolutionary studies (e.g., *Microtus* and their parasites; Wickström, 2004) and for analyses of trait evolution (e.g., mating systems).

A further, practical achievement of this study is to resolve a variety of long-standing taxonomic uncertainties in the genus *Microtus* since the molecular results are clear-cut and provide an objective basis for taxonomic revision. Our data show that there is a justification for subdividing the genus into subgenera, as proposed by previous workers, but there are also groupings of species within subgenera (Figs. 1–3) and an appropriate nomen-

clature will need to be developed, building on that previously established.

Acknowledgments

We are grateful to M. Akhverdyan, P. Basset, P. Benda, F. Catzeflis, V. Fedorov, D. Frynta, H. Hauffe, S. Haynes, V. Haukisalmi, H. Henttonen, R.A. Ims, B. Kryštufek, M. Macholán, J. Michavx, P. Munclinger, F. Poitevin, A. Polyakov, K. Sperling, C. Tez, N. Yoccoz, and D. Ziak for providing samples. We also want to thank C. Conroy, K. Fredga, F. Golenishchev, and I. Horáček for valuable advice, discussions and comments on a previous version of the manuscript and R.S. Hoffmann and two anonymous reviewers for comments on the manuscript. Financial support was received from the Swedish Institute (Guest Scholarship to I.G.), the Swedish Research Council, the Nilsson-Ehle Foundation, the Erik Philip-Sörensen's Foundation, the Grant Agency of the Czech Republic (GAČR 206/01/0562), and the Ministry of Education, Youth and Sport of the Czech Republic (MSMT 311100004) and from the Russian Foundation of Basic Research and INTAS (01-2163).

References

Akhverdyan, M.R., Lyapunova, E.A., Vorontsov, N.N., Teslenko, S.V., 1999. Intrapopulation autosomal polymorphism in the common vole *Microtus arvalis* of the Transcaucasian region. Russ. J. Genet. 35, 1452–1463.

Avise, J.C., 2000. Phylogeography: the History and Formation of Species. Harvard University Press, Cambridge, MA.

Barraclough, T.G., Nee, S., 2001. Phylogenetics and speciation. Trends Ecol. Evol. 16, 391–399.

Baskevich, M.I., 1996. On the karyological differentiation in Caucasian populations of common vole (Rodentia, Cricetidae, *Microtus*). Zool. Zh. 75, 297–308 (in Russian with English summary).

Bensasson, D., Zhang, D.-X., Hartl, D.L., Hewitt, G.M., 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol. Evol. 16, 314–321.

Bilton, D.T., Mirol, P.M., Mascheretti, S., Fredga, K., Zima, J., Searle, J.B., 1998. Mediterranean Europe as an area of endemism for small mammals rather than a source for northwards postglacial colonization. Proc. R. Soc. Lond. Ser. B 265, 1219–1226.

Bradley, R.D., Baker, R.J., 2001. A test of the genetic species concept: cytochrome *b* sequences and mammals. J. Mammal. 82, 960–973.

Brunet-Lecomte, P., Chaline, J., 1990. Relations phylogénétiques et évolution des campagnols souterrains d'Europe (*Terricola*, Arvicolidae, Rodentia). C.R. Acad. Sci. Paris, Série II 311, 745–750.

Brunet-Lecomte, P., Chaline, J., 1991. Morphological evolution and phylogenetic relationships of the European ground voles (Arvicolinae, Rodentia). Lethaia 24, 45–53.

Brunet-Lecomte, P., Chaline, J., 1992. Morphological convergences versus biochemical divergences in the holarctic ground voles: *Terricola* and *Pitymys* (Arvicolidae, Rodentia). N. Jb. Geol. Paläont. Mh. 12, 721–734.

Brunhoff, C., Galbreath, K.E., Fedorov, V.B., Cook, J.A., Jaarola, M., 2003. Holarctic phylogeography of the root vole (*Microtus oeconomus*): implications for late Quaternary biogeography of high latitudes. Mol. Ecol. 12, 957–968.

Bukhnikashvili, A., Kandaurov, A., 2002. The annotated list of mammals of Georgia. Proc. Inst. Zool. Acad. Sci. Georgia, Metsniereba, Tbilisi 21, 319–340.

Chaline, J., 1972. Les rongeurs du Pléistocène moyen et supérieur de France. Cahiers Paléont., CNRS, Paris, pp. 1–410.

Chaline, J., 1987. Arvicolid data (Arvicolidae, Rodentia) and evolutionary concepts. Evol. Biol. 21, 237–310.

Chaline, J., Brunet-Lecomte, P., Montuire, S., Viriot, L., Courant, F., 1999. Anatomy of the arvicoline radiation (Rodentia): palaeogeographical, palaeoecological history and evolutionary data. Ann. Zool. Fenn. 36, 239–267.

Chaline, J., Graf, J.-D., 1988. Phylogeny of the Arvicolidae (Rodentia): biochemical and paleontological evidence. J. Mammal. 69, 22–33.

Conroy, C.J., Cook, J.A., 1999. MtDNA evidence for repeated pulses of speciation within arvicoline and murid rodents. J. Mamm. Evol. 6, 221–245.

Conroy, C.J., Cook, J.A., 2000a. Molecular systematics of a holarctic rodent (*Microtus*: Muridae). J. Mammal. 81, 344–359.

Conroy, C.J., Cook, J.A., 2000b. Phylogeography of a post-glacial colonizer: *Microtus longicaudus* (Rodentia: Muridae). Mol. Ecol. 9, 165–175.

Conroy, C.J., Hortelano, Y., Cervantes, F.A., Cook, J.A., 2001. The phylogenetic position of southern relictual species of *Microtus* (Muridae: Rodentia) in North America. Mamm. Biol. 66, 332–344.

DeBry, R.W., 1992. Biogeography of New World taiga-dwelling *Microtus* (Mammalia: Arvicolidae): a hypothesis test that accounts for phylogenetic uncertainty. Evolution 46, 1347–1357.

DeWoody, A.J., Chesser, R.K., Baker, R.J., 1999. A translocated mitochondrial cytochrome *b* pseudogene in voles (Rodentia: *Microtus*). J. Mol. Evol. 48, 380–382.

Erixon, P., Svennblad, B., Britton, T., Oxelman, B., 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst. Biol. 52, 665–673.

Felsenstein, J., 1988. Phylogenies from molecular sequences—inference and reliability. Annu. Rev. Genet. 22, 521–565.

Filippucci, M.G., Fadda, V., Kryštufek, B., Simson, S., Amori, G., 1991. Allozyme variation and differentiation in *Chionomys nivalis* (Martins, 1842). Acta Theriol. 36, 47–62.

Fumagalli, L., Taberlet, P., Stewart, D.T., Gielly, L., Hausser, J., Vogel, P., 1999. Molecular phylogeny and evolution of *Sorex* shrews (Soricidae: Insectivora) inferred from mitochondrial DNA sequence data. Mol. Phylogenet. Evol. 11, 222–235.

Galleni, L., Stanyon, R., Tellini, A., Giordano, G., Santini, L., 1992. Karyology of the Savi pine vole, *Microtus savii* (De Sélys-Longchamps, 1838) (Rodentia, Arvicolidae): G-, C-, DA/DAPI-, and *Alu*I-bands. Cytogenet. Cell Genet. 59, 290–292.

Galleni, L., Tellini, A., Stanyon, R., Cicalò, A., Santini, L., 1994. Taxonomy of *Microtus savii* (Rodentia, Arvicolidae) in Italy: cytogenetic and hybridization data. J. Mammal. 75, 1040–1044.

Garapich, A., Nadachowski, A., 1996. A contribution to the origin of *Allophaiomys* (Arvicolidae, Rodentia) in Central Europe: the relationships between *Mimomys* and *Allophaiomys* from Kamyk (Poland). Acta Zool. Cracov. 39, 179–184.

Getz, L.L., 1985. Habitats. In: Tamarin, R.H.(Ed.), Biology of New World *Microtus*. American Society of Mammalogists. Special Publication No. 8, pp. 286–309.

Gill, A., Petrov, B., Zivkovic, S., Rimsa, D., 1987. Biochemical comparison in Yugoslavian rodents of the families Arvicolidae and Muridae. Z. Säugetierk. 52, 247–256.

Golenishchev, F.N., Malikov, V.G., Nazari, F., Vaziri, A.S., Sablina, O.V., Polyakov, A.V., 2003. New species of *guentheri* group (Rodentia, Arvicolinae, *Microtus*) from Iran. Russ. J. Theriol. 1, 117–123.

Graf, J.-D., 1982. Génétique biochimique, zoogéographie et taxonomie des Arvicolidae (Mammalia, Rodentia). Rev. Suisse Zool. 89, 749–787.

Gromov, I.M., Polyakov, I.Y., 1992. Fauna of the USSR, vol. 3, Voles (*Microtinae*). Brill Publishing Company, Leiden.

Gu, W., Wang, T., Zhu, B., 1999. Study on the morphology of sex chromosomes pairing of the synaptonemal complex in Mandarin vole (*Microtus mandarinus*). Acta Theriol. Sin. 19, 150–154 (in Chinese).

Haring, E., Herzig-Straschil, B., Spitzenberger, F., 2001. Phylogenetic analysis of Alpine voles of the *Microtus multiplex* complex using the mitochondrial control region. J. Zool. Syst. Evol. Res. 38, 231–238.

Haynes, S., Jaarola, M., Searle, J.B., 2003. Phylogeography of the common vole (*Microtus arvalis*) with particular emphasis on the colonization of the Orkney archipelago. Mol. Ecol. 12, 951–956.

Hellborg, L., 2004. Evolutionary studies of the mammalian Y chromosome. Ph.D, thesis, Uppsala University.

Hendy, M.D., Penny, D., 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38, 297–309.

Hoffmann, R.S., Koeppl, J.W., 1985. Zoogeography. In: Tamarin, R.H. (Ed.), Biology of New World *Microtus*. American Society of Mammalogists. Special Publication No. 8, pp. 84–0115.

Hudson, R.R., Coyne, J.A., 2002. Mathematical consequences of the genealogical species concept. Evolution 56, 1557–1565.

Hudson, R.R., Turelli, M., 2003. Stochasticity overrules the three-times rule: genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. Evolution 57, 182–190.

Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51, 673–688.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294, 2310–2314.

Hutterer, R., 2001. Bavarian pine vole still alive. Oryx 35, 185.

Irwin, D.M., Kocher, T.D., Wilson, A.C., 1991. Evolution of the cytochrome *b* gene of mammals. J. Mol. Evol. 32, 128–144.

Jaarola, M., Searle, J.B., 2002. Phylogeography of field voles (*Microtus agrestis*) in Eurasia inferred from mitochondrial DNA sequences. Mol. Ecol. 11, 2613–2621.

Jaarola, M., Searle, J.B., 2004. A highly divergent mitochondrial DNA lineage of *Microtus agrestis* in southern Europe. Heredity 92, 229–234.

Jaarola, M., Tegelström, H., 1995. Colonization history of north European field voles (*Microtus agrestis*) revealed by mitochondrial DNA. Mol. Ecol. 4, 299–310.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, pp. 21–132.

Kefelioğlu, H., Kryštufek, B., 1999. The taxonomy of *Microtus socialis* group (Rodentia: Microtinae) in Turkey, with the description of a new species. J. Nat. Hist. 33, 289–303.

Kocher, T.D., Thomas, W.K., Meyer, S.A., Edwards, S.V., Pääbo, S., Villablanca, F.X., Wilson, A.C., 1989. Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. Proc. Natl. Acad. Sci. USA 86, 6196–6200.

Kratochvíl, J., Král, B., 1974. Karyotypes and relationships of Palaearctic 54-chromosome *Pitymys* species (Microtidae, Rodentia). Zool. Listy 23, 289–302.

Kryštufek, B., 1999. Snow voles, genus *Chionomys*, of Turkey. Mammalia 63, 323–339.

Kryštufek, B., Griffiths, H.I., Vohralík, V., 1996. The status and use of *Terricola* Fatio, 1867 in the taxonomy of Palaearctic pine voles' (*Pitymys*) (Rodentia, Arvicolinae). Bull. Inst. R. Sci. Nat. Belg. Biol. 66, 237–240.

Kryštufek, B., Kefelioğlu, H., 2001. Redescription of *Microtus irani*, the species limits and a new social vole from Turkey. Bonn. Zool. Beitr. 50, 1–14.

Leaché, A.D., Reeder, T.W., 2002. Molecular systematics of the eastern fence lizard (*Sceloporus undulatus*): A comparison of parsimony, likelihood and Bayesian approaches. Syst. Biol. 51, 44–68.

Lessa, E.P., Cook, J.A., 1998. The molecular phylogenetics of tuco-tucos (genus *Ctenomys*, Rodentia: Octodontidae) suggests an early burst of speciation. Mol. Phylogenet. Evol. 9, 88–99.

Lin, Y.-H., Waddell, P.J., Penny, D., 2002. Pika and vole mitochondrial genomes increase support for both rodent monophyly and Glires. Gene 294, 119–129.

Macholán, M., Filippucci, M.G., Zima, J., 2001. Genetic variation and zoogeography of pine voles of the *Microtus subterraneus/majori* group in Europe and Asia Minor. J. Zool. 255, 31–42.

Maddison, D.R., Maddison, W.P., 2000. MacClade 4: Analysis of phylogeny and character evolution, Version 4.0. Sinauer Associates, Sunderland, MA.

Malygin, V.M., Panteleichuk, T.M.S.L., 2003. Efficiency of reproductive isolation mechanisms in six species of common voles (*Microtus*, Rodentia). In: Kryokov, A.P., Yakimenko, L.V. (Eds.), Problems of Evolution, vol. V. Russian Academy of Sciences, Far East Division, Vladivostok, Dalnauka, pp. 198–206 (in Russian with English summary).

Martínková, N., Dudich, A., 2003. The fragmented distribution range of *Microtus tatricus* and its evolutionary implications. Folia Zool. 52, 11–22.

Maryama, T., Imai, H.T., 1981. Evolutionary rate of the mammalian karyotype. J. Theor. Biol. 90, 111–121.

Mazurok, N.A., Rubtsova, N.V., Isaenko, A.A., Pavlova, M.E., Slobodyanyuk, S.Y., Nesterova, T.B., Zakian, S.M., 2001. Comparative chromosome and mitochondrial DNA analyses and phylogenetic relationships within common voles (*Microtus*, Arvicolidae). Chrom. Res. 9, 107–120.

Megías-Nogales, B., Marchal, J.A., Acosta, M.J., Bullejos, M., Díaz de la Guardia, R., Sánchez, A., 2002. Sex chromosome pairing in two Arvicolidae species: *Microtus nivalis* and *Arvicola sapidus*. Hereditas 138, 114–121.

McClellan, D.A., McCracken, K.G., 2001. Estimating the influence of selection on the variable amino acid sites of the cytochrome *b* protein functional domains. Mol. Biol. Evol. 18, 917–925.

Meier, M.N., Golenishchev, F.N., Radjabli, S.I., Sablina, O.V., 1996. Voles (subgenus *Microtus* Schrank) of Russia and adjacent territories. Proc. Zool. Inst. Russ. Acad. Sci. 232, 1–320 (in Russian).

Meier, M.N., Radjabli, S.I., Bulatova, N. Sh., Golenishchev, F.N., 1985. Karyological peculiarities and probable relations of common voles of the group '*arvalis*' (Rodentia, Cricetidae, *Microtus*). Zool. Zh. 64, 417–428 (in Russian with English summary).

Mekada, K., Koyasu, K., Harada, M., Narita, Y., Shrestha, K.C., Oda, S.-I., 2002. Karyotype and X–Y chromosome pairing in the Sikkim vole (*Microtus* (*Neodon*) *sikimensis*). J. Zool. 257, 417–423.

Mezhzherin, S.V., Morozov-Leonov, S.Yu., Kuznetsova, I.A., 1995. Biochemical variation and genetic divergence in Palearctic voles (Arvicolidae): subgenus *Terricola*, true lemmings *Lemmus* Link 1795, pied lemmings *Dicrostonyx* Gloger 1841, steppe lemmings *Lagurus* Gloger 1842, mole voles *Ellobius* Fischer von Waldheim 1814. Genetika 31, 788–797 (in Russian with English summary).

Mezhzherin, S.V., Zykov, A.E., Morozov-Leonov, S.Y.u., 1993. Biochemical variation and genetic divergence in Palearctic voles (Arvicolidae): meadow voles *Microtus* Schrank 1798, snow voles *Chionomys* Miller 1908, water voles *Arvicola* Lacepede 1799. Genetika 29, 28–41 (in Russian with English summary).

Mirol, P.M., Mascheretti, S., Searle, J.B., 2000. Multiple nuclear pseudogenes of mitochondrial cytochrome *b* in *Ctenomys* (Caviomorpha, Rodentia) with either great similarity to or high divergence from the true mitochondrial sequence. Heredity 84, 538–547.

Mitchell-Jones, A.J., Amori, G., Bogdanowicz, W., Kryštufek, B., Reijnders, P.J.H., Spitzenberger, F., Stubbe, M., Thissen, J.B.M., Vohralík, V., Zima, J., 1999. The Atlas of European Mammals. Poyser, London.

Modi, W.S., 1987. Phylogenetic analyses of chromosomal banding patterns among the Nearctic Arvicolidae (Mammalia: Rodentia). Syst. Zool. 36, 109–136.

Musser, G.G., Carleton, M.D., 1993. Family Muridae. In: Wilson, D.E., Reeder, D.M. (Eds.), Mammal Species of the World: A Taxonomic and Geographic Reference, second ed. Smithsonian Institution Press, Washington, pp. 510–756.

Nadachowski, A., 1991. Systematics, geographic variation, and evolution of snow voles (*Chionomys*) based on dental characters. Acta Theriol. 36, 1–45.

Nadachowski, A., Garapich, A., 1998. *Allophaiomys* evolutionary stage in extant *Microtus*. Paludicola 2, 91–94.

Nadachowski, A., Zagorodnyuk, I., 1996. Recent *Allophaiomys*-like species in the Palaearctic: Pleistocene relicts or a return to an initial type? Acta Zool. Cracov. 39, 387–394.

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nichols, R., 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16, 358–364.

Nowak, R.M., 1999. Walker's Mammals of the World, sixth ed. The Johns Hopkins University Press, Baltimore, ML.

Orlov, V.N., Yatsenko, V.N., Malygin, V.M., 1983. Karyotype homology and species phylogeny in a group of field mice (Cricetidae, Rodentia). Doklady Biol. Sci. 269, 217–219.

Panteleyev, P.A., 1998. The Rodents of the Palaearctic Fauna: Composition and Areas. Russian Academy of Sciences, Moscow.

Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Potapov, S.G., Orlov, V.N., Kovalskaya, Yu.M., Malygin, V., M, Ryskov, A.P., 1999. Genetic differentiation in the voles of the tribe Arvicolini (Cricetidae, Rodentia) using DNA taxonprint and RAPD-PCR. Russ. J. Genetics 35, 403–410.

Rabeder, G., 1986. Herkunft und frühe Evolution der Gattung *Microtus* (Arvicolidae, Rodentia). Z. Säugetierk. 51, 350–367.

Reed, R.D., Sperling, F.A.H., 1999. Interaction of process partitions in phylogenetic analysis: an example from the swallowtail butterfly genus *Papilio*. Mol. Biol. Evol. 16, 286–297.

Rekovets, L., Nadachowski, A., 1995. Pleistocene voles (*Arvicolidae*) of the Ukraine. Paleont. Evolutió 28-29, 145–245.

Richmond, G.M., 1996. The INQUA-approved provisional Lower-Middle Pleistocene boundary. In: Turner, C. (Ed.), The Early Middle Pleistocene in Europe. Balkema, Rotterdam.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Rosenberg, N.A., Nordborg, M., 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3, 380–390.

Searle, J.B., 1996. Speciation in small mammals. Symp. Zool. Soc. Lond. 69, 143–156.

Spitzenberger, F., Brunet-Lecomte, P., Nadachowski, A., Bauer, K., 2000. Comparative morphometrics of the first lower molar in *Microtus* (*Terricola*) cf. *liechtensteini* of the Eastern Alps. Acta Theriol. 45, 471–483.

Suchentrunk, F., Markov, G., Haiden, A., 1998. On gene pool divergence of the two karyotypically distinct sibling vole species *Microtus arvalis* and *M. rossiaemeridionalis* (Arvicolidae, Rodentia). Folia Zool. 47, 103–114.

Suzuki, Y., Glazko, G.V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99, 16138–16143.

Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4. Sinauer Associates, Sunderland, MA.

Tsekoura, N., Fraguedakis-Tsolis, S., Chondropoulos, B., Markakis, G., 2002. Morphometric and allozyme variation in central and southern Greek populations of *Microtus* (*Terricola*) *thomasi*. Acta Theriol. 47, 137–149.

Wickström, L.M., 2004. Phylogeny, phyletic coevolution and phylogeography of anoplocephaline cestodes in mammals. Ph.D. thesis, University of Helsinki, Finnish Forest Research Institute, Research Papers 918.

Yang, Z., 1994. Estimating the pattern of nucleotide substitutions. J. Mol. Evol. 39, 105–111.

Yang, Z., Yoder, A.D., 1999. Estimation of the transition/transversion rate bias and species sampling. J. Mol. Evol. 48, 274–283.

Yigit, N., Colak, E., 2002. On the distribution and taxonomic status of *Microtus guentheri* (Danford and Alston, 1880) and *Microtus lydius* Blackler, 1916 (Mammalia: Rodentia) in Turkey. Turk. J. Zool. 26, 197–204.

Zagorodnyuk, I.V., 1989. Taksonomiya, rasprostranenie i morfologicheskaya izmenchivost'polevok roda *Terricola* vostochnoi Evropy. [Taxonomy, distribution and morphological variation of the *Terricola* voles in East Europe]. Vestnik Zool. 5, 3–14 (in Russian with English summary).

Zagorodnyuk, I.V., 1990. Kariotipicheskaya izmenchivost'i sistematika serykh polevok (Rodentia, Arvicolini). Soobshchenie1. Vidovoi sostav i khromosomnye chisla. [Karyotypic variability and systematics of the gray voles (Rodentia, Arvicolini). Communication 1. Species composition and chromosomal numbers]. Vestnik Zool. 2, 26–37 (in Russian).

Zakrzewski, R.J., 1985. The fossil record. In: Tamarin, R.H.(Ed.), Biology of New World *Microtus*. American Society of Mammalogists. Special Publication No. 8, pp. 1–51.

Zheng, S.-H., Zhang, Z.-Q., 2000. Late Miocene–Early Pleistocene micromammals from Wenwanggou of Lingtai, Gansu, China. Vertebrata PalAsiatica 38, 58–71.

Zima, J., Král, B., 1984. Karyotypes of European mammals II. Acta Sc. Nat. Brno 18, 1–62.

Paper 2.1.2

Martínková N., Moravec J. 2012. Multi-locus phylogeny of arvicoline voles (Arvicolini, Rodentia) shows small tree terrace size. *Folia Zoologica* 61: 254-267.

Folia Zool. – 61 (3–4): 254–267 (2012)

Multilocus phylogeny of arvicoline voles (Arvicolini, Rodentia) shows small tree terrace size

Natália MARTÍNKOVÁ[1,2]* and Jiří MORAVEC[2]

[1] *Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, v.v.i., Květná 8, 603 65 Brno, Czech Republic; e-mail: martinkova@ivb.cz*
[2] *Institute of Biostatistics and Analyses, Masaryk University, Kamenice 3, 625 00 Brno, Czech Republic; e-mail: jirka678@seznam.cz*

Received 4 January 2012; Accepted 25 June 2012

Abstract. We combined mitochondrial (*cyb*, control region, *coi*, *nd4*) and nuclear (*irbp*, *ghr*, *sry*, *lcat*) DNA sequence data to infer phylogenetic relationships of arvicoline voles. The concatenated supermatrix contained 72.8 % of missing data. From this dataset, Bayesian inference showed close relationships of *Arvicola* and *Chionomys*, *Proedromys* with *Lasiopodomys* and *Microtus gregalis*, *Phaiomys* with *Neodon* and *M. clarkei*. Genus *Microtus* formed a supported group with *Blanfordimys* and *N. juldaschi*. The gene partition taxon sets were explained in the multilocus phylogeny in such a way that the resulting Bayesian inference tree represented a unique solution on a terrace in the tree space. This means that although the supermatrix contained a large proportion of missing data, it was informative in retrieving a phylogeny with a unique optimality score, tree likelihood.

Key words: divergence, evolutionary history, supertree, supermatrix, phylogenetic tree terrace, *Microtus*, Arvicolinae

Introduction

Arvicoline voles (Rodentia, Arvicolini) are a young group of small rodents distributed on the northern hemisphere. They started to diverge probably as recently as two to three million years ago, but in the short time frame, they speciated into one of the most speciose mammalian groups (Wilson & Reeder 2005). Today, the species show rapid temporal changes in genetic composition of populations (Bryja et al. 2007, Oliver et al. 2009, Rudá et al. 2010; but see Spaeth et al. 2009) and fast karyotype reorganisation (Mazurok et al. 2001, Mekada et al. 2002, Sitnikova et al. 2007, Mitsainas et al. 2010) coupled with gene reorganisation between mitochondrial and nucleotide genomes (Triant & DeWoody 2007, 2008). Populations of arvicoline voles diverge quickly when they become fragmented in refugia, leading at times to speciation in these areas, and refugia become speciation traps (Martínková & Dudich 2003, Martínková et al. 2007, Tougard et al. 2008, Kryštufek et al. 2009, Haring et al. 2011). Multiple vole species co-occur in many regions and habitat partitioning was found with sympatric occurrence (Jurdíková et al. 2000, Santos et al. 2011). Yet, they are morphologically similar with small differences between distantly related taxa (Fraguedakis-Tsolis et al. 2009).

Reconstruction of phylogenetic relationships in arvicoline voles is complicated by morphological similarities and rapid karyotype rearrangements, making the group an ideal model for molecular genetic studies. Both mitochondrial (mtDNA) and nuclear (nucDNA) markers have been extensively used to resolve relationships within the group and to study evolutionary history of different taxa. The relative merit of information value of mtDNA and nucDNA markers overlaps in arvicoline rodents. MtDNA sequences provide valuable information in phylogeny reconstruction at specific and intrageneric level, contributing to resolving phylogeographic histories of taxa (e.g. Conroy & Cook 2000a, Jaarola & Searle 2002, Fink et al. 2004, Brunhoff et al. 2006, Fan et al. 2011), identification of putative cryptic species

* *Corresponding Author*

(Kefelioğlu & Kryštufek 1999, Hellborg et al. 2005, Castiglia et al. 2008, Conroy & Neuwald 2008, Weksler et al. 2010) and phylogenetic placement of taxa with unstable position based on other data (Macholán et al. 2001, Jaarola et al. 2004, Martínková et al. 2007, Kryštufek et al. 2009, Bannikova et al. 2010). The phylogenetic signal of mtDNA and its ability to resolve relationships decrease at higher level of phylogenies that exhibits as a rapid burst of diversification (Jaarola et al. 2004). However, also nuclear markers, either DNA sequence data or AFLP markers (Galewski et al. 2006, Abramson et al. 2009, Fink et al. 2010), fail to fully resolve the signal of the rapid diversification of voles at the base of the tree. Incongruence of sampling between studies further complicates interpretation of relationships within the group (Bužan et al. 2008, Haring et al. 2011).

Here, we combine available mitochondrial and nuclear DNA sequence markers to reconstruct phylogenetic relationships of arvicoline rodents and to assess stability of the retrieved model. We use Bayesian inference analysis of a concatenated supermatrix and SuperTriplets supertree reconstruction to estimate parent trees. These we then use to establish the size of the terrace where trees will have the same likelihood with the dataset with large content of missing data.

Material and Methods

Sequences were downloaded from GenBank for 74 species belonging to genera *Arvicola*, *Blanfordimys*, *Chionomys*, *Lasiopodomys*, *Microtus*, *Neodon*, *Phaiomys* and *Proedromys* (Table 1). Herewith, we accept species designation of Wilson & Reeder (2005) with the addition of recently established species *Microtus gromovi* (Bannikova et al. 2010) and *Proedromys liangshanensis* (Liu et al. 2007). Alignments were constructed in Geneious 5.4 (Drummond et al. 2011) with sequences from mitochondrial genes for cytochrome *b* (*cyb*), control region (CR), cytochrome *c* oxidase subunit I (*coi*), NADH dehydrogenase subunit 4 (*nd4*) and nuclear genes for interphotoreceptor retinoid-binding protein, exon 1 (*irbp*), growth hormone receptor, exon 10 (*ghr*), sex-determining region Y (*sry*), lecithin: cholesterol acyl transferase, exons 2 through 5 (*lcat*). The alignments were reduced to contain at least three sequences at every base. In the *sry* gene, the microsatellite $(TC)_n(TG)_n$ (Acosta et al. 2010) could not be aligned unambiguously, and the region was deleted. The data and results are available through TreeBASE (http://purl.org/phylo/treebase/phylows/study/TB2:S12667).

Two additional loci have good taxonomic sampling. However, both loci in the *avpr1a* gene, its upstream region and exon 1, were previously documented to be disparate with the species tree (Fink et al. 2007, 2010), and some sequences of the *avpr1a* exon 1 were shared between distantly related species (Fink et al. 2007). To reduce conflict between gene trees in our analyses, we chose to omit these loci.

Optimal substitution model was estimated in MrModeltest 2.3 (Nylander 2004) with Akaike Information Criterion (AIC) and applied to the individual gene alignments and partitions of the concatenated alignment. Where the parameters of the selected model were extreme, a simpler model was used. Bayesian Inference (BI) was conducted in MrBayes 3.1 (Ronquist & Huelsenbeck 2003) with Markov chain Monte Carlo (MCMC) parameters set to 2-5 million steps sampled every 1000^{th}, five to six chains in two runs with chain temperature 0.08-0.12 and chain swapping attempted once every third generation. The MCMC runs were optimised to mix and ideally to finish with average standard deviation of split frequencies below 0.01, potential scale reduction factor for model parameters approaching 1.000 and proportion of successful chain stage swaps between 0.4 and 0.7. BI is robust in recovering the correct tree topology, but it might fail to establish appropriate branch lengths with default branch length prior (Marshall 2010). The 95 % credibility intervals of the tree lengths from BI were compared to the tree length obtained from maximum likelihood (ML) analysis. The ML tree was calculated in RAxML 7.2 (Stamatakis 2006). The trees were re-rooted to midpoint root to allow for uncertainty in the phylogenetic position of *Arvicola* (Galewski et al. 2006, Bužan et al. 2008, Abramson et al. 2009, Bannikova et al. 2009).

Divergence events between all taxa were investigated in two ways; from a combination of gene trees and directly from the concatenated supermatrix. The gene trees were combined into a SuperTriplets supertree (Ranwez et al. 2010). The method breaks down the gene trees to their smallest components containing three taxa, where any two taxa are more closely related than either is to the third. Supertree then contains medians of relationships from the triplets as they were found in the gene trees. Edge support in the SuperTriplets analysis is the proportion of triplets that support a given edge.

The supermatrix was resolved with partitioned BI ran for 6 million generations with five heated and one cold MCMC, chain temperature set to 0.09 and one chain swap attempted every third generation. Partition rates were allowed to vary.

Table 1. *Accession numbers of DNA sequences used in this study to reconstruct phylogenetic relationships.*

Species	Distrib.	mitochondrial				nuclear			
		cyb	CR	*coi*	*nd4*	*irbp*	*ghr*	*sry*	*lcat*
Arvicola									
A. amphibius (=terrestris)	Europe	AF119269	AY948543	AY332681	AF128938	AY277407	AM392380	FN433499	GQ267511
A. sapidus	Europe	FJ539345	FJ502319					FN433498	
Blanfordimys									
B. afghanus	Asia	EF599109							
B. bucharensis	Asia	AM392369					AM392392		
Chionomys									
C. gud	Asia	EU700087							GQ267512
C. nivalis	Europe	AY513845	AF267284	AY332687		AM919424	AM392378	FN433493	AH005248
C. roberti	Asia	AY513850							
Lasiopodomys									
L. brandtii	Asia						GQ142008		FN401371
L. mandarinus	Asia	FJ986322				AM919413	AM392396		GQ267514
Microtus									
M. abbreviatus	N America	AF163890							
M. agrestis	Europe	FJ619786	AF267270		AF128940	AM919427	AM910792	FN433492	
M. anatolicus	Asia	FJ767740							
M. arvalis	Europe	AY220789	AF267285	AY332683		AM919416	AM392386	FN433490	GQ267517
M. bavaricus	Europe	GQ243218	AF267277						
M. brachycercus	Europe	AY513828							
M. breweri	N America							FN433501	
M. cabrerae	Europe	AY513788							
M. californicus	N America	AF163891						FN433510	
M. canicaudus	N America	AF163892							
M. chrotorrhinus	N America	AF163893				AM919403	AM392383	FN433503	
M. clarkei	Asia	AY641526							
M. daghestanicus	Asia	AY513790					GQ142009		GQ267518
M. dogramacii	Asia	AY513795							
M. duodecimcostatus	Europe	AJ717744					AM392400	FN433496	
M. evoronensis	Asia		HM135862						
M. felteni	Europe	AY513798							
M. fortis	Asia	AF163894	JF261174	JF261174	JF261174				
M. gerbei	Europe	AY513799							
M. gregalis	Asia	AF163895					GQ142007		GQ267513
M. gromovi	Asia	FJ986319	HM135891						
M. guatemalensis	C America	AF410262							
M. guentheri	Eurasia	AY513804				AM919420	AM392397	FN433488	
M. ilaeus	Asia	AY513809							
M. irani	Asia	FJ767748							

M. kikuchii	Asia	AF348082	AF348082	AF348082	AF348082	AM919410	AM392385		
M. levis	Eurasia	DQ015676	DQ015676	DQ015676	DQ015676			FN433489	
M. liechtensteini	Europe	AY513811	AF267281						
M. limnophilus	Asia	FJ986323				AM919426			
M. longicaudus	N America	AF119267			AF128936	AM919414	AM392379		
M. lusitanicus	Europe	AY513812						FN433497	
M. majori	Europe	AY513814				AM919409	AM910796		
M. maximowiczii	Asia	FJ986311	HM135863	HM137737					
M. mexicanus	N America	AF163897	AF251260						
M. middendorffii	Asia	HM119493	HM135899	HM137752		AM919419	AM392390		GQ267516
M. miurus	N America	GU809171							
M. mongolicus	Asia	FJ986305		HM137754					
M. montanus	N America	AF119280	GU394082		AF128946			FN433513	
M. montebelli	Asia	AF163900				AM919421	AM910793		
M. mujanensis	Asia		HM135854						
M. multiplex	Europe	AY513815	AF267276						
M. oaxacensis	C America	AF410260							
M. ochrogaster	N America	AF163901				AM919423	AM392389	FN433506	
M. oeconomus	Holarctic	AB372207	AJ616853			AM919418	AM392388	FN433491	GQ267515
M. oregoni	N America	AF163903						FN433516	
M. pennsylvanicus	N America	AF119279	AY369777	JF443830	AF128945	AM919415	AF540633		
M. pinetorum	N America	AF163904							
M. quasiater	C America	AF410259							
M. richardsoni	N America	AF163905				AM919404	AM392387	FN433517	
M. sachalinensis	Asia	FJ986318	HM135900						
M. savii	Europe	AY513824						FN433495	
M. schelkovnikovi	Asia	AM910619				AM919408	AM910794		
M. socialis	Asia	AY513829				FM162055	FM162073		
M. subterraneus	Europe	AY513832	AF267271	AY332685					
M. tatricus	Europe	AY513837	AF267267						
M. thomasi	Europe	AY513840	AY560558			AM919422	AM910797	FN433494	
M. townsendii	N America	AF163906							
M. transcaspicus	Asia					AM919405	AM910795		
M. umbrosus	C America	AF410261							
M. xanthognathus	N America	AF163907							
Neodon									
N. irene	Asia	AM392370				AM919412	AY294924		
N. juldaschi	Asia	AY513808							
N. sikimensis	Asia					AY163593			
Phaiomys									
Ph. leucurus	Asia	AM392371				AM919400	AM392394		
Proedromys									
Pr. liangshanensis	Asia	FJ463038	FJ463038	FJ463038	FJ463038				

Given the fractional nature of the supermatrix, multiple distinct trees were likely to display the same set of subtrees representing taxa sampled per gene. Such trees will have the same log-likelihood for a partitioned analysis (Sanderson et al. 2011). A set of trees that display the same set of subtrees for sampled loci and have the same log-likelihood is called a terrace. The trees from a terrace are derived from each other by nearest-neighbour interchange (NNI) rearrangement, and, in a dataset with considerable missing data content, the terraces might contain many trees. The size of terraces for our dataset was assessed with perl scripts from the PhyloTerraces package (Sanderson et al. 2011). SuperTriplets supertree and BI tree based on the supermatrix were used as parent trees. Terrace identification requires binary (fully resolved) trees. The BI tree was resolved by accepting all relationships resolved in the posterior sample. The SuperTriplets supertree was resolved using relationships from the *cyb* tree. In terrace identification analysis, the parent tree was broken to subtrees, where each subtree represented relationships of taxa sampled for the respective gene as resolved in the parent tree. All relationships in the subtrees were further characterized by triplets. From these, all alternative parent trees that contain the subtrees were constructed.

Results

The dataset contained 1143 base-pairs (bp) long alignment with 68 taxa for *cyb*, CR alignment was 1025 bp long and contained 25 taxa, *coi* was 1545 bp long with 12 taxa, *nd4* was 1378 bp long with 9 taxa, *irbp* was 1181 bp long with 24 taxa, *ghr* was 911 bp long with 27 taxa, *sry* was 908 bp long with 19 taxa and *lcat* was 590 bp long with 10 taxa. The concatenated supermatrix had 72.8 % missing data composed of missing sequences of individual genes, alignment gaps and unknown nucleotides.

Substitution models selected by AIC for each gene were GTR + Γ + I for *cyb*, HKY + Γ + I for CR, GTR + Γ + I for *coi*, GTR + Γ for *nd4*, HKY + Γ for *irbp*, GTR + Γ + I for *ghr*, HKY + I for *sry* and HKY + I for *lcat*. As the GTR model requires estimation of the rate matrix, a simpler HKY model was tested for *coi*, *nd4* and *ghr* genes. The difference between log-likelihood based on selected and tested model was 10.8 for *coi*, 1.8 for *nd4* and 1.1 for *ghr* and the HKY model was used for *nd4* and *ghr* genes. The resulting trees were similar, and MCMC convergence was faster with the simpler model. The results from the simpler models are reported (Table 2, Figs. 1-2).

The 95 % credibility interval (CI) of the BI tree length based on the partitioned concatenated dataset was 7.89-15.12 with default rate parameter of the exponential branch length prior (λ = 10.0). This CI of the BI tree length did not contain the tree length 3.86 estimated from the ML analysis. Increasing the rate by increments of 10.0 to the final value of 50.0, the tree length decreased to 4.0-4.59, but we did not further test the branch length prior because of decrease in node support for high λ. Node support improved with optimisation of the branch length prior when λ = 20.0, and subsequently decreased (Fig. 3). We further analyse the BI tree with the highest average node support.

The BI supermatrix phylogeny re-rooted with midpoint root showed two initial groups (Fig. 4). The root separated genera *Arvicola* and *Chionomys* from *Microtus*, *Blanfordimys*, *Phaiomys*, *Proedromys* and *Lasiopodomys*. *Proedromys liangshanensis* was a sister species to a well-supported group containing *Lasiopodomys* and *Microtus gregalis*. Similarly, *Phaiomys leucurus* was a sister taxon to a supported group with unresolved internal relationships containing *Neodon irene*, *N. sikimensis* and *M. clarkei*. Remaining taxa formed a monophyletic group with high Bayesian posterior probability (BPP). It contained species currently attributed to genus *Microtus* not mentioned above, genus *Blanfordimys* and *N. juldaschi*. The first group that diverged at this level was the subgenus *Microtus* (*Alexandromys*) predominantly distributed in the eastern Palaearctic. *Microtus fortis* group was well differentiated from the basal species of *Alexandromys*, *M. kikuchii*, *M. montebelli* and *M. oeconomus*. We confirmed position of *M. gromovi* as a distinct taxon rather than a subspecies of *M. maximowiczii*.

Further notable group consisted of *N. juldaschi* with *Blanfordimys afghanus* and *B. bucharensis*. *M. agrestis* was a sister species to this group, but the relationship was unsupported. Nearctic species formed an unsupported group with *M. cabrerae*. Within the Nearctic group, three pairs of sister taxa had significant node support. Subgenera *Microtus* (*Terricola*) and *Microtus* (*Microtus*) were sister groups that were most derived in the BI phylogeny (Fig. 4). In the subgenus *Microtus*, *M. arvalis* group and *M. socialis/guentheri* group were separated, but classification of *M. schelkovnikovi* to the *M. socialis/guentheri* group had low support (BPP = 0.85). Subgenus *Terricola* was separated to the eastern and western clade, where the eastern clade containing *M. majori*, *M. subterraneus* and *M. daghestanicus* had BPP = 0.94.

Table 2. *Substitution models used for separate gene tree analyses and partitions of the supermatrix. Model parameters are means estimated from sampled posterior distribution of the gene trees after burn-in.* κ *– transition/ transversion rate ratio,* r *– substitution rate,* f *– base frequency,* α *– shape parameter of the* Γ *distribution, I – proportion of invariable sites, n/a – not available.*

Parameter	*cyb* GTR + Γ + I	CR HKY + Γ + I	*coi* GTR + Γ + I	*nd4* HKY + Γ	*irbp* HKY + Γ	*ghr* HKY + I	*sry* HKY + I	*lcat* HKY + I
κ	n/a	3.0452	n/a	14.1872	4.0786	4.8725	3.3133	6.5095
r(A ↔ C)	0.0176	n/a	0.0236	n/a	n/a	n/a	n/a	n/a
r(A ↔ G)	0.4088	n/a	0.4862	n/a	n/a	n/a	n/a	n/a
r(A ↔ T)	0.0402	n/a	0.0515	n/a	n/a	n/a	n/a	n/a
r(C ↔ G)	0.0065	n/a	0.0070	n/a	n/a	n/a	n/a	n/a
r(C ↔ T)	0.4714	n/a	0.4152	n/a	n/a	n/a	n/a	n/a
r(G ↔ T)	0.0555	n/a	0.0165	n/a	n/a	n/a	n/a	n/a
f(A)	0.3713	0.3262	0.3006	0.3421	0.2205	0.2611	0.2909	0.2015
f(C)	0.3596	0.2586	0.2736	0.3032	0.2809	0.2877	0.2640	0.2728
f(G)	0.0774	0.1141	0.1478	0.0976	0.2909	0.2291	0.2166	0.2766
f(T)	0.1918	0.3012	0.2780	0.2572	0.2077	0.2221	0.2285	0.2491
α	0.6046	1.1420	8.9089	0.2172	0.4326	n/a	n/a	n/a
I	0.4968	0.3503	0.6647	n/a	n/a	0.7059	0.3133	0.6735

The SuperTriplets supertree agreed with the BI tree in distinguishing groups *Arvicola*, *Microtus* (*Terricola*), *Microtus* (*Microtus*), *Microtus* (*Pitymys*) with *M. guatemalensis*, *Blanfordimys* with *N. juldaschi* and a separate group including taxa from the subgenus *Microtus* (*Alexandromys*) (data available at TreeBASE). The other groups were distorted due to different position of *Chionomys nivalis*, *N. sikimensis* and *Lasiopodomys brandtii*. In the supertree, *C. nivalis* was placed as a basal taxon after diversification of *Arvicola*. *Neodon sikimensis* formed a polyphyly with *C. gud* and *C. roberti*. *Lasiopodomys brandtii* was placed within the group containing Nearctic species. Phylogenetic terraces where the trees belonged to were small. The terrace with the BI tree used as the parent tree consisted of a single tree, and the terrace with the supertree consisted of 15 trees.

Discussion

Tree space of the concatenated dataset

We found that by optimising branch length prior of the Bayesian inference analysis of the multilocus phylogeny of arvicoline rodents with missing data comprising nearly 73 % of the concatenated supermatrix we were able to retrieve a phylogeny that is unique on a terrace. This means that there are no trees with alternative topology that would explain sets of taxa from individual gene partitions that would be present in the BI phylogeny. For the supertree approach, the terrace size was also small, and it contained 15 trees with alternative topology that explained the relationships of gene tree datasets as depicted on the supertree.

The topology of our phylogeny that well explained gene sets in the final trees was not reflected in similar congruence in branch length estimations. Our Bayesian phylogeny with highest node support was longer than the ML tree as is known to occur in partitioned datasets (Marshall et al. 2006, Brown et al. 2010, Marshall 2010). The cause of this discrepancy was recently identified to be a branch length prior that places too much probability density on large tree lengths (Rannala et al. 2012). The default on branch lengths in MrBayes assigns independent and identical exponential priors for individual branch lengths. As the default initial value for branch lengths is 0.1, the MCMC starts from very long trees for large datasets with many taxa, which is often unrealistic. The prior then places too much influence on the posterior. The effect is exacerbated for large datasets where there are partitions with low variability or correlations in rate variation or substitution models in the posterior (Rannala et al. 2012). This seems to be the case in our dataset. We optimised the branch length prior in our partitioned analyses, assuming that by setting the branch length exponential prior mean closer to mean branch length estimated from the ML tree length, the posterior tree length would be more similar to the ML tree length (Zhang et al. 2012). This is not a suitable approach in the Bayesian framework (c.f. Zhang et al. 2012). Our data showed that decreasing the tree length in this way lead to decrease of node support for high values of rate parameter of the exponential distribution of the uncorrelated branch length prior. Using compound Dirichlet priors on branch lengths in

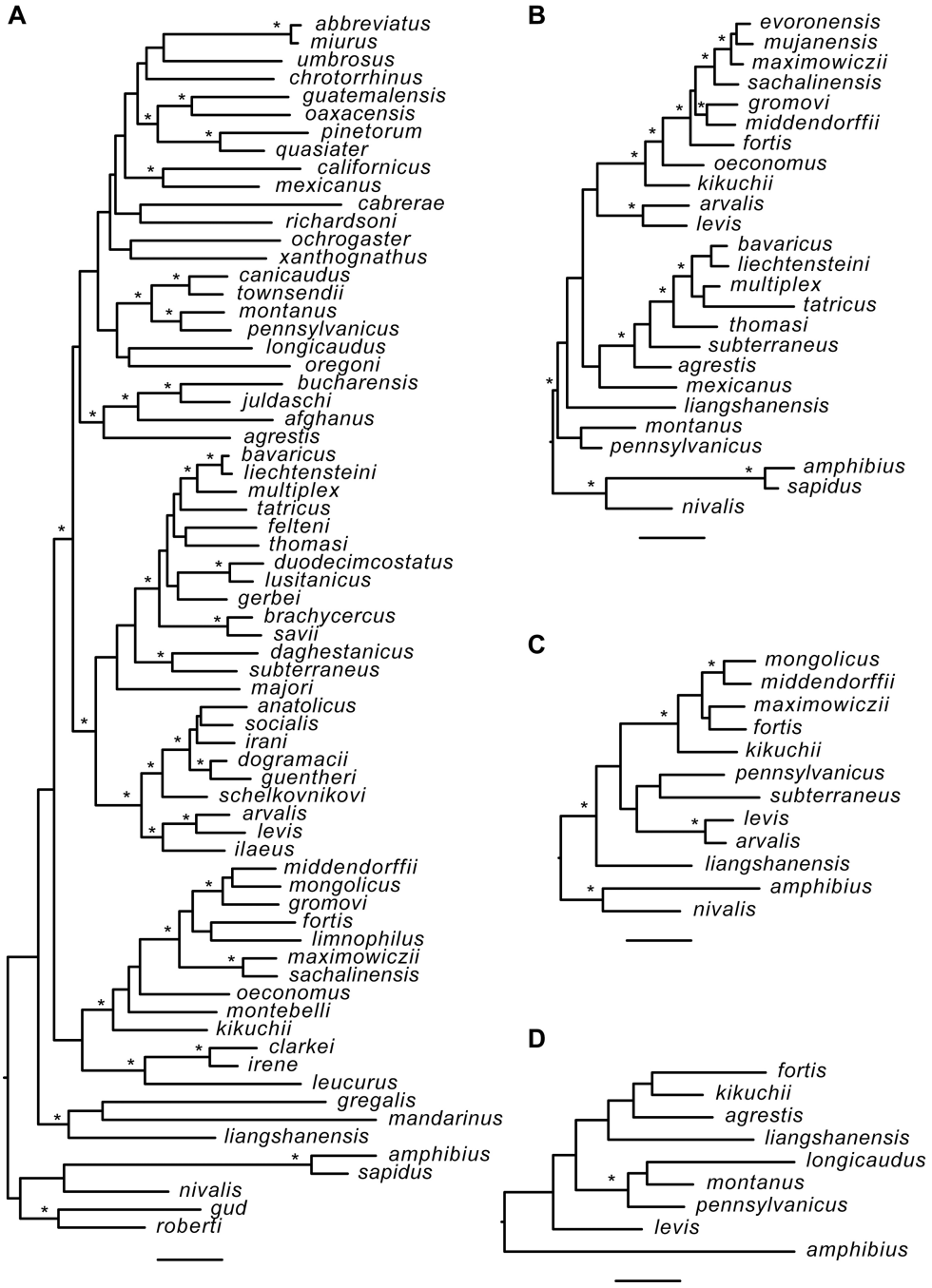


Fig. 1. *Bayesian inference phylogenetic trees based on complete mitochondrial sequences for* cyb *(A), and partial sequences for control region (B),* coi *(C) and* nd4 *(D) genes. All relationships are shown that were compatible with the consensus tree from the posterior sample of trees after burn-in. Scale bars indicate 0.1 substitutions per site, asterisk denotes nodes with Bayesian posterior probability (BPP) ≥ 0.95.*

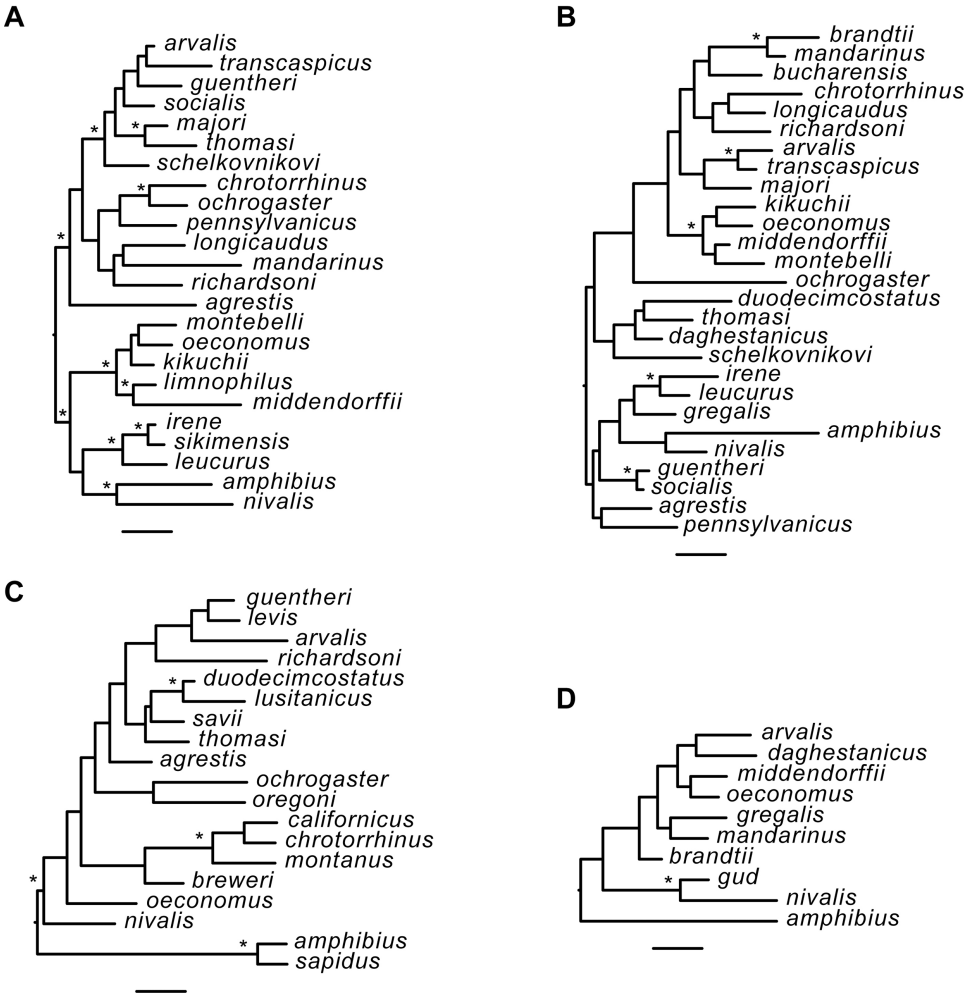


Fig. 2. *Bayesian inference phylogenetic trees based on nuclear sequences for partial* irbp *(A),* ghr *(B),* sry *(C) and* lcat *(D) genes. All relationships are shown that were compatible with the consensus tree from the posterior sample of trees after burn-in. Scale bars indicate 0.01 substitutions per site, asterisk denotes nodes with BPP ≥ 0.95.*

modified MrBayes 3.1 (Zhang et al. 2012), we were not able to obtain a tree with CI of tree length that would include the ML estimate and the average BPP remained lower than in our optimal tree.

Arvicoline phylogeny

Phylogenetic position of *Arvicola* within subfamily Arvicolinae is unstable in studies utilising mtDNA (Bužan et al. 2008, Bannikova et al. 2009) in comparison with studies that use nucDNA (Galewski et al. 2006, Abramson et al. 2009). We also observed this in gene trees. Our supermatrix results show that by rooting the tree of the Arvicolini tribe *sensu* Galewski et al. (2006) with midpoint root, *Arvicola* forms a supported group with *Chionomys* at the base of the tree.

Microtus gregalis represented a phylogenetic enigma. In early molecular phylogenies, it was placed distantly from other supposedly related species at the base of the phylogeny of *Microtus*, but its basal position was unsupported (Conroy & Cook 2000b, Conroy et al. 2001, Jaarola et al. 2004). It was later retrieved as a sister taxon to *Chionomys* based on mtDNA (Bužan &

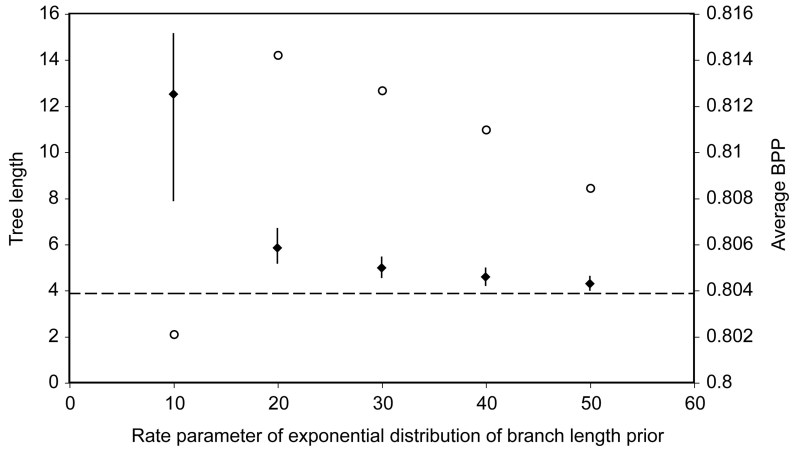


Fig. 3. *Changes of the tree length (black diamonds, primary axis) and average Bayesian posterior probability of node support (empty circles, secondary axis) with increasing shape parameter of the exponential distribution of uncorrelated branch lengths prior. Dashed line indicates tree length estimated from maximum likelihood analysis with GTR + Γ substitution model for each partition. Tree length is given with 95 % credibility interval.*

Kryštufek 2008), but nucDNA grouped *M. gregalis* with *Lasiopodomys* (Abramson et al. 2009). This grouping elucidates towards rapid karyotype rearrangements between species, as *M. gregalis* has 36 chromosomes (Martínková et al. 2004), whereas *L. mandarinus* chromosome number varies between 47 and 52 (Liu et al. 2010). If the ancestral karyotype of the group was $2n = 54$, the rearrangements leading to *M. gregalis* that branches close to the base of the tree were extensive (c.f. Lemskaya et al. 2010). The phylogenetic position of *M. gregalis* in the vicinity of *Lasiopodomys* was confirmed once *Lasiopodomys* was sequenced for the mtDNA (Bannikova et al. 2010). Our combined phylogeny placed *M. gregalis* close to the base of the trees in a well-supported group with *L. mandarinus* and *L. brandtii* in accordance with recent studies (Abramson et al. 2009, Bannikova et al. 2010). The sister taxon of the group is *Proedromys liangshanensis* that was described recently (Liu et al. 2007). Based on phylogeny of complete mtDNA, *Pr. liangshanensis* is a sister species to *Microtus* (Hao et al. 2011). Our analysis included more comprehensive sampling, and we found that the species forms a supported sister relationship with *Lasiopodomys* and *M. gregalis*.

The genus *Neodon* was polyphyletic with *N. juldaschi* grouping with *Blanfordimys* deeper in the phylogeny than other species attributed to *Neodon*, *N. irene* and *N. sikimensis*. The latter two species consistently belonged to the *Phaiomys*/*Neodon* lineage (Galewski et al. 2006, Robovský et al. 2008, Bannikova et al. 2009, 2010) that included also *M. clarkei* in our analyses. Its relationship with *Neodon* was not resolved, forming a strongly supported trichotomy, where BPP for the monophyly of *Neodon* within this lineage was as low as 0.41.

The phylogenetic position of *M. agrestis* was unstable in the gene trees, and it was placed as an unsupported sister taxon to the *N. juldaschi*/*Blanfordimys* group in our multilocus phylogeny. *Microtus agrestis* split from the common ancestors of *Microtus* early in the radiation of the genus, but its closest relatives might not be identifiable today similarly as in the case of *M. cabrerae*. Interestingly, *M. agrestis* from the Iberian peninsula, where *M. cabrerae* is also distributed, forms a phylogenetic lineage distinct from other *M. agrestis* populations. This divergence is present both in mitochondrial and nuclear phylogenies and might represent a cryptic species that was not formally described to date (Jaarola & Searle 2004, Hellborg et al. 2005). The erratic placement of *M. agrestis* in different gene trees and unsupported position of *M. cabrerae* with North American *Microtus* indicates that these species represent relicts of a very early colonization of Arvicolini to Western Europe. Phylogenies of Arvicolidae improve with more comprehensive sampling (Bužan et al. 2008), and based on the fact that we analysed majority of species in the tribe Arvicolini, we are confident to state that the close relatives of *M. agrestis* and *M. cabrerae* are extinct today and their phylogenetic position is influenced more by stochastic processes in DNA sequence evolution such as saturation or

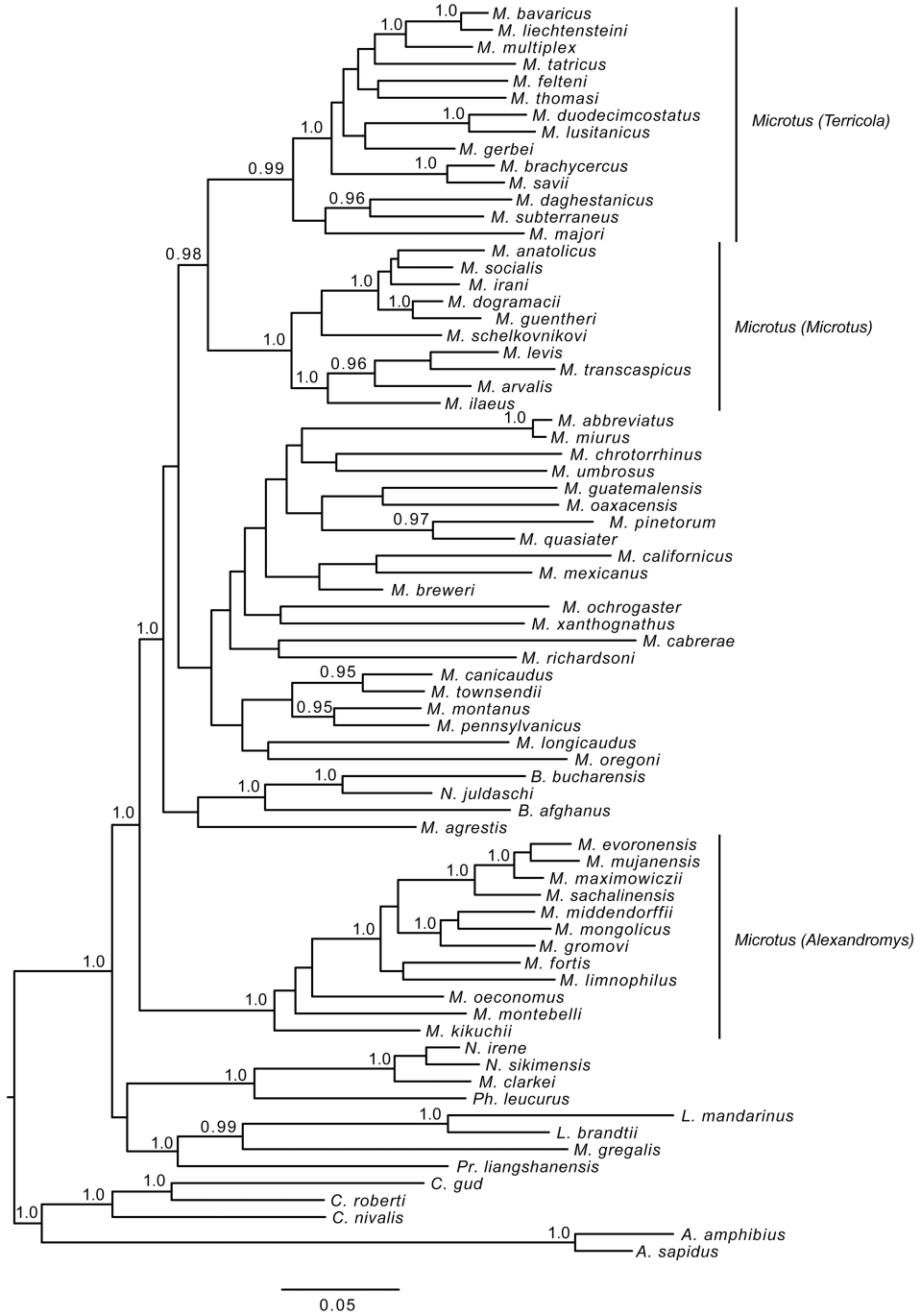


Fig. 4. *Bayesian inference phylogenetic tree based on concatenated sequence of the genes depicted in Figs. 1–2. All relationships are shown that were compatible with the consensus tree from the posterior sample of trees after burn-in. Scale bar is in substitutions per site, BPP ≥ 0.95 is shown.*

convergence and by computational artefacts in phylogeny reconstruction.

Within *Microtus*, we retrieved three groups that represent subgenera recognised by Wilson & Reeder (2005) with minor changes. Subgenus *Alexandromys* was supported without *M. clarkei* as per Wilson & Reeder (2005), *Microtus* without *M. cabrerae* and *Terricola* with *M. tatricus*. The species within subgenera showed relationships established in previous studies (Jaarola et al. 2004, Martínková et al. 2007, Kryštufek et al. 2009, Bannikova et al. 2010, Haring et al. 2000, 2011).

Nearctic *Microtus* consistently suffer from lack of resolution of many relationships (Conroy & Cook 2000b, Conroy et al. 2001, Jaarola et al. 2004, Fink et al. 2010), although recent studies show that these taxa are also strongly influenced by rapid diversification in speciation traps. Weksler et al. (2010) found *M. abbreviatus* from Wrangell Mts. to be divergent and potentially merit species status, and Conroy & Neuwald (2008) distinguished two species within *M. californicus*. In our multilocus study, phylogenetic position of *M. breweri* was particularly unstable. This was probably in lieu of the fact that only the *sry* gene sequence was available for this species.

Species with small ranges were often part of rapidly differentiating groups. This leads to an assumption that geographic isolation in small refugia triggers diversification on both molecular and morphological levels. In *Microtus*, the results of phylogeography couple with species phylogenies where phylogeography nowadays indicates regions and populations that might give rise to new species in the future.

Acknowledgements

We appreciate M. Macholán's invitation to participate in the issue of Folia Zoologica in tribute to Jan Zima's anniversary. N. Martínková is thankful to J. Zima for his guidance and support at the beginning of her career. The bioinformatic analyses were conducted at the computational cluster of the Institute of Vertebrate Biology and at the MetaCentrum. The access to the MetaCentrum computing facilities was provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" LM2010005 funded by the Ministry of Education, Youth, and Sports of the Czech Republic. This study was supported by the Grant Agency of the Academy of Sciences of the Czech Republic, grant no. IAA600930609 and with institutional support RVO: 68081766.

Literature

Abramson N.I., Lebedev V.S., Tesakov A.S. & Bannikova A.A. 2009: Supraspecies relationships in the subfamily Arvicolinae (Rodentia, Cricetidae): an unexpected result of nuclear gene analysis. *Mol. Biol. 43: 834–846. (in Russian)*

Acosta M.J., Marchal J.A., Romero-Fernández I., Megías-Nogales B., Modi W.S. & Sánchez Baca A. 2010: Sequence analysis and mapping of the *sry* gene in species of the subfamily Arvicolinae (Rodentia). *Sex. Dev. 4: 336–347.*

Bannikova A.A., Lebedev V.S. & Golenishchev F.N. 2009: Taxonomic position of Afghan vole (subgenus *Blanfordimys*) by the sequence of the mitochondrial *cytb* gene. *Russ. J. Genet. 45: 91–97.*

Bannikova A.A., Lebedev V.S., Lissovsky A.A., Matrosova V., Abramson N.I., Obolenskaya E.V. & Tesakov A.S. 2010: Molecular phylogeny and evolution of the Asian lineage of vole genus *Microtus* (Rodentia: Arvicolinae) inferred from mitochondrial cytochrome *b* sequence. *Biol. J. Linn. Soc. 99: 595–613.*

Brown J.M., Hedtke S.M., Lemmon A.R. & Lemmon E.M. 2010: When trees grow too long: investigating the causes of highly inaccurate Bayesian branch-length estimates. *Syst. Biol. 59: 145–161.*

Brunhoff C., Yoccoz N.G., Ims R.A. & Jaarola M. 2006: Glacial survival or late glacial colonization? Phylogeography of the root vole (*Microtus oeconomus*) in north-west Norway. *J. Biogeogr. 33: 2136–2144.*

Bryja J., Charbonnel N., Berthier K., Galan M. & Cosson J.-F. 2007: Density-related changes in selection pattern for major histocompatibility complex genes in fluctuating populations of voles. *Mol. Ecol. 16: 5084–5097.*

Bužan E.V. & Kryštufek B. 2008: Phylogenetic position of *Chionomys gud* assessed from a complete cytochrome *b* gene. *Folia Zool. 57: 274–282.*

Bužan E.V., Kryštufek B., Hänfling B. & Hutchinson W.F. 2008: Mitochondrial phylogeny of Arvicolinae using comprehensive taxonomic sampling yields new insights. *Biol. J. Linn. Soc. 94: 825–835.*

Castiglia R., Annesi F., Aloise G. & Amori G. 2008: Systematics of the *Microtus savii* complex (Rodentia, Cricetidae) via mitochondrial DNA analyses: paraphyly and pattern of sex chromosome evolution. *Mol. Phylogenet. Evol. 46: 1157–1164.*

Conroy C.J & Cook J.A. 2000a: Phylogeography of a post-glacial colonizer: *Microtus longicaudus* (Rodentia: Muridae). *Mol. Ecol. 9: 165–175.*

Conroy C.J. & Cook J.A. 2000b: Molecular systematics of a holarctic rodent (*Microtus*: Muridae). *J. Mammal. 81: 344–359.*

Conroy C.J., Horteland Y., Cervantes F.A. & Cook J.A. 2001: The phylogenetic position of southern relictual species of *Microtus* (Muridae: Rodentia) in North America. *Mamm. Biol. 66: 332–344.*

Conroy C.J. & Neuwald J.L. 2008: Phylogeographic study of the California vole, *Microtus californicus*. *J. Mammal. 89: 755–767.*

Drummond A.J., Ashton B., Buxton S., Cheung M., Cooper A., Duran C., Field M., Heled J., Kearse M., Markowitz S., Moir R., Stones-Havas S., Sturrock S., Thierer T. & Wilson A. 2011: Geneious v5.4. *Available from http://www.geneious.com/*

Fan Z., Liu S., Liu Y., Zhang X. & Yue B. 2011: How Quaternary geologic and climatic events in the southeastern margin of the Tibetan Plateau influence the genetic structure of small mammals: inferences from phylogeography of two rodents, *Neodon irene* and *Apodemus latronum*. *Genetica 139: 339–351.*

Fink S., Excoffier L. & Heckel G. 2004: Mitochondrial gene diversity in the common vole *Microtus arvalis* shaped by historical divergence and local adaptations. *Mol. Ecol. 13: 3501–3514.*

Fink S., Excoffier L. & Heckel G. 2007: High variability and non-neutral evolution of the mammalian *avpr1a* gene. *BMC Evol. Biol. 7, 176.*

Fink S., Fischer M.C., Excoffier L. & Heckel G. 2010: Genomic scans support repetitive continental colonization events during the rapid radiation of voles (Rodentia: *Microtus*): the utility of AFLPs versus mitochondrial and nuclear sequence markers. *Syst. Biol. 59: 548–572.*

Fraguedakis-Tsolis S.E., Chondropoulos B.P., Stamatopoulos C.V. & Giokas S. 2009: Morphological variation of the five vole species of the genus *Microtus* (Mammalia, Rodentia, Arvicolinae) occurring in Greece. *Acta Zool. 90: 254–264.*

Galewski T., Tilak M., Sanchez S., Chevret P., Paradis E. & Douzery E.J.P. 2006: The evolutionary radiation of Arvicolinae rodents (voles and lemmings): relative contribution of nuclear and mitochondrial DNA phylogenies. *BMC Evol. Biol. 6, 80.*

Hao H., Liu S., Zhang X., Chen W., Song Z., Peng H., Liu Y. & Yue B. 2011: Complete mitochondrial genome of a new vole *Proedromys liangshanensis* (Rodentia: Cricetidae) and phylogenetic analysis with related species: are there implications for the validity of the genus *Proedromys*? *Mitochondr. DNA 22: 28–34.*

Haring E., Herzig-Straschil B. & Spitzenberger F. 2000: Phylogenetic analysis of Alpine voles of the *Microtus multiplex* complex using the mitochondrial control region. *J. Zool. Syst. Evol. Res. 38: 231–238.*

Haring E., Sheremetyeva I.N. & Kryukov A.P. 2011: Phylogeny of Palearctic vole species (genus *Microtus*, Rodentia) based on mitochondrial sequences. *Mamm. Biol. 76: 258–267.*

Hellborg L., Gündüz İ. & Jaarola M. 2005: Analysis of sex-linked sequences supports a new mammal species in Europe. *Mol. Ecol. 14: 2025–2031.*

Jaarola M., Martínková N., Gündüz İ., Brunhoff C., Zima J., Nadachowski A., Amori G., Bulatova N.S., Chondropoulos B., Fraguedakis-Tsolis S., González-Esteban J., López-Fuster M.J., Kandaurov A.S., Kefelioğlu H., da Luz Mathias M., Villate I. & Searle J.B. 2004: Molecular phylogeny of the speciose vole genus *Microtus* (Arvicolinae, Rodentia) inferred from mitochondrial DNA sequences. *Mol. Phylogenet. Evol. 33: 647–663.*

Jaarola M. & Searle J.B. 2002: Phylogeography of field voles (*Microtus agrestis*) in Eurasia inferred from mitochondrial DNA sequences. *Mol. Ecol. 11: 2613–2621.*

Jaarola M. & Searle J.B. 2004: A highly divergent mitochondrial DNA lineage of *Microtus agrestis* in southern Europe. *Heredity 92: 228–234.*

Jurdíková N., Žiak D. & Kocian Ľ. 2000: Habitat requirements of *Microtus tatricus*: macrohabitat and microhabitat. In: Urban P. (ed.), Výskum a ochrana cicavcov na Slovensku 4. *ŠOP COPK, Banská Bystrica.*

Kefelioğlu H. & Kryštufek B. 1999: The taxonomy of *Microtus socialis* group (Rodentia: Microtinae) in Turkey, with the description of a new species. *J. Nat. Hist. 33: 289–303.*

Kryštufek B., Bužan E.V., Vohralík V., Zareie R. & Özkan B. 2009: Mitochondrial cytochrome *b* sequence yields new insight into the speciation of social voles in south-west Asia. *Biol. J. Linn. Soc. 98: 121–128.*

Lemskaya N., Romanenko S.A., Golenishchev F.N., Rubtsova N.V., Sablina O.V., Serdukova N.A., O'Brian P.C.M., Fu B., Yiğit N., Ferguson-Smith M.A., Yang F. & Graphodatsky A.S. 2010: Chromosomal evolution

of Arvicolinae (Cricetidae, Rodentia). III. Karyotype relationships of ten *Microtus* species. *Chromosome Res. 18: 459–471.*

Liu S., Sun Z., Zeng Z. & Zhao E. 2007: A new vole (Cricetidae: Arvicolinae: *Proedromys*) from the Liangshan Mountains of Sichuan Province, China. *J. Mammal. 88: 1170–1178.*

Liu H., Yan N. & Zhu B. 2010: Two new karyotypes and bandings in *Microtus mandarinus faeceus* (Rodentia). *Hereditas 147: 123–126.*

Macholán M., Filippucci M.G. & Zima J. 2001: Genetic variation and zoogeography of pine voles of the *Microtus subterraneus/majori* group in Europe and Asia Minor. *J. Zool. 255: 31–42.*

Martínková N. & Dudich A. 2003: The fragmented distribution range of *Microtus tatricus* and its evolutionary implications. *Folia Zool. 52: 11–22.*

Martínková N., Nová P., Sablina O.V., Graphodatsky A.S. & Zima J. 2004: Karyotypic relationships of the Tatra vole (*Microtus tatricus*). *Folia Zool. 53: 279–284.*

Martínková N., Zima J., Jaarola M., Macholán M. & Spitzenberger F. 2007: The origin and phylogenetic relationships of *Microtus bavaricus* based on karyotype and mitochondrial DNA sequences. *Folia Zool. 56: 39–49.*

Marshall D.C. 2010: Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. *Syst. Biol. 59: 108–117.*

Marshall D.C., Simon C. & Buckley T.R. 2006: Accurate branch length estimation in partitioned Bayesian analyses requires accommodation of among-partition rate variation and attention to branch length priors. *Syst. Biol. 55: 993–1003.*

Mazurok N.A., Rubtsova N.V., Isaenko A.A., Pavlova M.E., Slobodyanyuk S.Y., Nesterova T.B. & Zakian S.M. 2001: Comparative chromosome and mitochondrial DNA analyses and phylogenetic relationships within common voles (*Microtus*, Arvicolidae). *Chromosome Res. 9: 107–120.*

Mekada K., Koyasu K., Harada M., Narita Y., Shrestha K.C. & Oda S.-I. 2002: Karyotype and X-Y chromosome pairing in the Sikkim vole (*Microtus* (*Neodon*) *sikimensis*). *J. Zool. 257: 417–423.*

Mitsainas G.P., Rovatsos M.T. & Giagia-Athanasopoulou E.B. 2010: Heterochromatin study and geographical distribution of *Microtus* species (Rodentia, Arvicolinae) from Greece. *Mamm. Biol. 75: 261–269.*

Nylander J.A.A. 2004: MrModeltest v2. Program distributed by the author. *Evolutionary Biology Centre, Uppsala University.*

Oliver M.K., Lambin X., Cornulier T. & Piertney S.B. 2009: Spatio-temporal variation in the strength and mode of selection acting on major histocompatibility complex diversity in water vole (*Arvicola terrestris*) metapopulations. *Mol. Ecol. 18: 80–92.*

Rannala B., Zhu T. & Yang Z. 2012: Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Mol. Biol. Evol. 29: 325–335.*

Ranwez V., Criscuolo A. & Douzery E.J.P. 2010: SuperTriplets: a triplet-based supertree approach to phylogenomics. *Bioinformatics 26: i115–i123.*

Robovský J., Řičánková V. & Zrzavý J. 2008: Phylogeny of Arvicolinae (Mammalia, Cricetidae): utility of morphological and molecular data sets in a recently radiating clade. *Zool. Scr. 37: 571–590.*

Ronquist F. & Huelsenbeck J.P. 2003: MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics 19: 1572–1574.*

Rudá M., Žiak D., Kocian Ľ. & Martínková N. 2010: Low genetic variability in a mountain rodent, the Tatra vole. *J. Zool. 281: 118–124.*

Sanderson M.J., McMahon M.M. & Steel M. 2011: Terraces in phylogenetic tree space. *Science 333: 448–450.*

Santos S.M., da Luz Mathias M. & Mira A.P. 2011: The influence of local, landscape and spatial factors on the distribution of the Lusitanian and the Mediterranean pine voles in a Mediterranean landscape. *Mamm. Biol. 76: 133–142.*

Sitnikova N.A., Romanenko S.A., O'Brien P.C.M., Perelman P.L., Fu B., Rubtsova N.V., Serdukova N.A., Golenishchev F.N., Trifonov V.A., Ferguson-Smith M.A., Yang F. & Graphodatsky A.S. 2007: Chromosomal evolution of Arvicolinae (Cricetidae, Rodentia). I. The genome homology of tundra vole, field vole, mouse and golden hamster revealed by comparative chromosome painting. *Chromosome Res. 15: 447–456.*

Spaeth P.A., van Tuinen M., Chan Y.L., Terca C. & Hadly E.A. 2009: Phylogeography of *Microtus longicaudus* in the tectonically and glacially dynamic Central Rocky Mountains. *J. Mamm. 90: 571–584.*

Stamatakis A. 2006: RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics 22: 2688–2690.*

Tougard C., Brunet-Lecomte P., Fabre M. & Montuire S. 2008: Evolutionary history of two allopatric *Terricola* species (Arvicolinae, Rodentia) from molecular, morphological, and palaeontological data. *Biol. J. Linn. Soc. 93: 309–323.*

Triant D.A. & Dewoody J.A. 2007: Extensive mitochondrial DNA transfer in a rapidly evolving rodent has been mediated by independent insertion events and by duplications. *Gene 401: 61–70.*

Triant D.A. & Dewoody J.A. 2008: Molecular analyses of mitochondrial pseudogenes within the nuclear genome of arvicoline rodents. *Genetica 132: 21–33.*

Weksler M., Lanier H.C & Olson L.E. 2010: Eastern Beringian biogeography: historical and spatial genetic structure of singing voles in Alaska. *J. Biogeogr. 37: 1414–1431.*

Wilson D.E. & Reeder D.M. 2005: Mammal species of the world. A taxonomic and geographic reference (3rd ed.). *John Hopkins University Press, Baltimore.*

Zhang C., Rannala B. & Yang Z. 2012: Robustness of compound Dirichlet priors for Bayesian inference of branch lengths. *Syst. Biol. 61: 779–784.*

Paper 2.1.3

Martínková N., Zima J., Jaarola M., Macholán M., Spitzenberger F. 2007. The origin and phylogenetic relationships of *Microtus bavaricus* based on karyotype and mitochondrial DNA sequences. *Folia Zoologica* 56: 39-49.

Folia Zool. – 56(1): 39–49 (2007)

The origin and phylogenetic relationships of *Microtus bavaricus* based on karyotype and mitochondrial DNA sequences

Natália MARTÍNKOVÁ[1,2], Jan ZIMA[1], Maarit JAAROLA[3,*], Miloš MACHOLÁN[4] and Friederike SPITZENBERGER[5]

[1] *Department of Population Biology, Institute of Vertebrate Biology of the ASCR, v.v.i., Studenec 122, 675 02 Koněšín, Czech Republic; e-mail: martinkova@brno.cas.cz, jzima@ivb.cz*

[2] *Department of Biology, University of York, PO Box 373, YO10 5YW York, United Kingdom*

[3] *Department of Cell and Organism Biology, Genetics Building, Lund University, Sölvegatan 29, SE-223 62 Lund, Sweden; e-mail: m_jaarola@zbs.bialowieza.pl*

[4] *Laboratory of Mammalian Evolutionary Genetics, Institute of Animal Physiology and Genetics of the ASCR, v.v.i., Veveří 97, 602 00 Brno, Czech Republic; e-mail: macholan@iach.cz*

[5] *Natural History Museum, Burgring 7, A-1010 Vienna, Austria; e-mail: friederike.spitzenberger@nhm-wien.ac.at*

[*] *Present address: Mammal Research Institute, Polish Academy of Sciences, ul. Waszkiewicza 1c, 17-230 Białowieża, Poland*

Received 2 August 2006; Accepted 2 March 2007

A b s t r a c t . Geographic isolation of small populations in refugia during late Pleistocene glaciations resulted in population differentiation that in some cases lead to speciation. We report the karyotype of *Microtus bavaricus*, an evolutionary young and threatened rodent endemic to the Alps. Our results show that the karyotype of *M. bavaricus* is almost identical to that of *M. liechtensteini* (2N = 46, NF = 54). A close relationship between the two species was also supported by phylogenetic analysis of complete mitochondrial DNA sequences for the cytochrome *b* gene. The cytochrome *b* divergence between *Microtus bavaricus* and *M. liechtensteini* was 1.7 %, the lowest estimate observed among the 14 currently recognised species of Eurasian pine voles (subgenus *Terricola*).

Key words: *Terricola*, molecular divergence, glaciation

Introduction

Vole species of the genus *Microtus* (Arvicolinae, Rodentia) differ considerably in age and various evolutionary stages of speciation can be observed within the genus. The Eurasian pine voles, subgenus *Terricola*, include species groups that are especially suitable for analysis of recent divergence events (J a a r o l a et al. 2004). Geographic isolation of small populations in refugia during the late Pleistocene glaciations could have served as "speciation traps" for several of these young taxa, thus promoting speciation (C h a l i n e 1987, M a r t í n k o v á & D u d i c h 2003).

The Bavarian pine vole, *Microtus bavaricus* (König, 1962), is an endemic species of the Alps with an extremely restricted range and rather enigmatic phylogenetic relationships. In fact, its distribution area covers only six known localities in the Innsbruck Alps in Bavaria, Germany (*terra typica* at Garmisch-Partenkirchen), and northern Tyrol in Austria (K ö n i g 1982, S p i t z e n b e r g e r 2002, C a r l e t o n & M u s s e r 2005). Because of the species' distributional pattern, it was suggested that *M. bavaricus* survived the last glacial period in a refugium situated in the northern Alps (K r a t o c h v í l 1970,

Spitzenberger 2002). Since the original morphological description by König (1962), affinities of *M. bavaricus* to the two other lineages of pine voles endemic to the Alps and some nearby mountain ranges, *M. multiplex* and *M. liechtensteini*, have been indicated (Kratochvíl 1970, Spitzenberger 2002) and confirmed by both morphological (Spitzenberger et al. 2000) and molecular genetic analysis (Haring et al. 2000). This group, the *M. multiplex* complex, including *M. multiplex*, *M. liechtensteini* and *M. bavaricus*, is characterised by low morphological divergence and *M. multiplex* and *M. liechtensteini* were occasionally considered to be conspecific (Krapp 1982). This opinion found particular support in a finding of a single natural hybrid in the area of parapatric contact between the two taxa (Storch & Winking 1977). The F1-hybrid from Calliano, Trento province in Italy showed karyotype characteristics of both parental species and had an intermediate diploid number of chromosomes (2N = 47).

M. multiplex and *M. liechtensteini* can be distinguished by their parapatric distribution patterns and differences in karyotype: *M. multiplex* is distributed in the western parts of the Alps and certain adjacent mountain ranges (eastern margins of Massif Central, northern Apennines), whereas *M. liechtensteini* occurs in the eastern and north-eastern Alps and the western Dinaric mountains (Mitchell – Jones et al. 1999). The diploid number of chromosomes differs between the two taxa (2N = 48 in *M. multiplex*, 2N = 46 in *M. liechtensteini*) and the two karyotypes can also be distinguished also by other details in the morphology of individual chromosomes (Zima & Král 1984). However, the molecular divergence between *M. multiplex* and *M. liechtensteini* falls within the 4–8 % cytochrome *b* range that includes both inter- and intra-specific divergence in *Microtus* (Jaarola et al. 2004).

The aim of the present paper is to report on the hitherto unknown karyotype of *M. bavaricus*. We have also examined sequences of the mitochondrial cytochrome *b* gene for this species, in order to estimate the taxonomic position and phylogenetic relationships of *M. bavaricus* to other species of pine voles (subgenus *Terricola*). Analysis of cytochrome *b* sequences enabled us to utilize the extensive data set of publicly available sequences of closely related *Terricola* species.

Material and Methods

Chromosomes

The karyotype was studied in a male of *M. bavaricus* collected in the Rofangebirge in northern Tyrol, Austria. It was collected on 3 August, 2004 by Simon Engelberger northwest of Steinberg/Rofan in an open spruce forest near a small brook, a tributary to the Ampelsbach River (47° 32' N, 11° 45' E, 1100 m a.s.l.). The voucher skin and skull (NMW65362) are stored in the Mammal Collection of the Natural History Museum in Vienna. Mitotic chromosomes were prepared from bone marrow cells obtained from short-term culture, using the standard technique with hypotonic treatment and fixation in a mixture of ethanol and acetic acid. The karyotype was then analyzed by conventional Giemsa staining.

Cytochrome *b* sequences

Complete or partial sequences of the mitochondrial gene for cytochrome *b* were obtained for 14 individuals of *Microtus bavaricus, M. tatricus*, *M. majori* and *M. liechtensteini*

Table 1. Individuals voucher numbers, sample localities, cytochrome *b* haplotypes and GenBank accession numbers for sequenced specimens of *Microtus*, subgenus *Terricola*.

Species	Voucher No.	Locality	Haplotype	Accession No.
M. bavaricus	NMW65362	Steinberg am Rofan, Northern Tyrol, Austria	bavaricus 1	DQ841693
	NMW8072	Garmisch-Partenkirchen, Bavaria, Germany	bavaricus 2	DQ841694
	NMW26592	Steinberg am Rofan, Northern Tyrol, Austria	bavaricus 3	DQ841695
M. tatricus	NM-179	Prvé Roháčske pleso Lake, Western Tatra Mts, Slovakia	tatricus 4	DQ841696
	NM-546	Prvé Roháčske pleso Lake, Western Tatra Mts, Slovakia	tatricus 5	DQ841697
	NM-182	Prvé Roháčske pleso Lake, Western Tatra Mts, Slovakia	tatricus 6	DQ841698
	NM-194	Rakytovská dolina Valley, Veľká Fatra Mts, Slovakia	tatricus 7	DQ841699
	NM-195	Rakytovská dolina Valley, Veľká Fatra Mts, Slovakia	tatricus 7	DQ841699
	NM-202	Dolný Harmanec, Veľká Fatra Mts, Slovakia	tatricus 8	DQ841700
	NM-76	Tretie Roháčske pleso Lake, Western Tatra Mts, Slovakia	tatricus 9	DQ841701
	NM-84	Smutná dolina Valley, Western Tatra Mts, Slovakia	tatricus 10	DQ841702
M. majori	TU601	Damar, Turkey	majori 2	DQ841703
	MM388	Hopa, Turkey	majori 3	DQ841704
M. liechtensteini	CR43	Croatia	liechtensteini 2	EF379100

NMW – National Museum, Vienna, Austria; NM – collection of N. Martínková; TU, MM – collection of M. M a c h o l á n ; CR – collection of Heikki H e n t t o n e n .

(Table 1). Additional sequences were downloaded from GenBank (Accession Numbers in Fig. 2; Jaarola et al. 2004, Galewski et al. 2006, Tougard et al., in press). The haplotype names used here are identical to the original references except for *brachycercus 1–3* (*cf.* Carleton & Musser 2005) that were named *savii 1–3* in Jaarola et al. (2004).

Primers used for amplification and sequencing of the cytochrome *b* gene, PCR and sequencing conditions are described in detail in Jaarola et al. (2004). Sequences were completed in Sequencher 4.2 (GeneCodes) and manually aligned in BioEdit 5.0 (Hall 1999).

Sequence composition and nucleotide diversity (π) were calculated in DnaSP 4.1 (Rozas et al. 2003). Total (Dxy) and net (Da) divergence was estimated in MEGA 3.1 (Kumar et al. 2004) using Kimura 2-parameter distances (Kimura 1980). Substitution model was assessed in ModelTest 3.7 (Posada & Crandall 1998) using the Akaike Information Criterion. The model selected, GRT+I+Γ (Tavaré 1986; I = 0.59, α = 1.5008), was used to estimate a maximum likelihood (ML) phylogenetic tree in PAUP* 4.0b10 (Swofford 2003). Bootstrap support of taxonomic units was calculated from 10,000 neighbour-joining (NJ) and 100 maximum parsimony (MP) parametric replicates. Bayesian posterior probabilities were estimated in MrBayes 3.1 (Ronquist & Huelsenbeck 2003) from 2,000,000 generations sampled every 1000th generation excluding a burn-in of 200,000 steps.

Results

Chromosomes

The male diploid complement of *M. bavaricus* contains 46 chromosomes (Fig. 1). The largest pair consists of two subtelocentric chromosomes with tiny, but clearly visible short arms. The second largest chromosome is an odd submetacentric. There are three meta- or submetacentric chromosomes of medium size. The other chromosomes are acrocentric, and their size decreases gradually. The largest acrocentric pair approximately equals in size to the larger arm of the odd submetacentric chromosome. The number of chromosomal arms is therefore 54, however, additional small short arms are apparent in at least two pairs of acrocentric chromosomes. The odd large submetacentric element can be identified as the X chromosome. The Y chromosome is one of the three elements with the meta- or submetacentric position of the centromere. In most metaphases, one of these chromosomes was slightly larger than the other two, and this could be considered as the Y chromosome.

Cytochrome *b*

Altogether 11 complete cytochrome *b* haplotypes (1143 base pairs; bp) were submitted to public databases (GenBank Accession Numbers: DQ841693-DQ841704, EF379100). Additionally, two *M. bavaricus* individuals yielded partial cytochrome *b* sequences, of 1052 (DQ841695) and 650 bp (DQ841694). Phylogenetic analyses were based on the complete cytochrome *b* alignment of 38 haplotypes of European endemic *Terricola* species, including one available complete *M. bavaricus* sequence, and 11 haplotypes of *Terricola* species with distribution areas covering Europe and Asia Minor. Within- and between-species divergences were estimated from all available *Terricola* sequences using a 650 bp alignment of 54

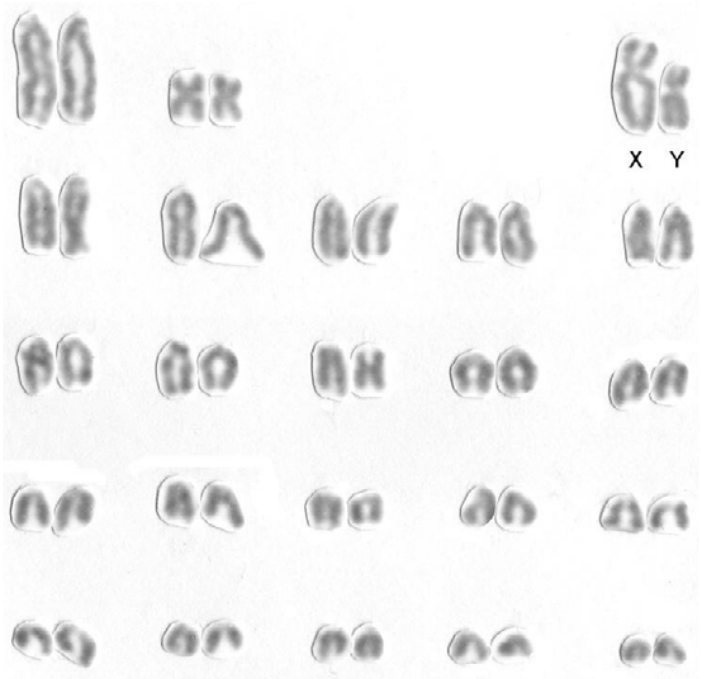


Fig. 1 . Karyotype of a male *Microtus bavaricus* from Rofangebirge, northern Tyrol, Austria analysed by conventional Giemsa staining.

sequences. Phylogenetic analysis of this alignment, including three *M. bavaricus* and two *M. liechtensteini*, shows that *M. bavaricus* and *M. liechtensteini* are reciprocally monophyletic (data not shown).

For the 1143 bp cytochrome *b* alignment, a total of 371 (32 %) polymorphic sites were observed and 325 (28 %) of those were parsimony informative. Phylogenies inferred with ML, MP, NJ and Bayesian methods had similar topologies. The phylogenetic analysis shows that *M. bavaricus* is most closely related to *M. liechtensteini* and the result is supported by high bootstrap support and posterior probability values (Fig. 2). Monophyly of the *M. multiplex* complex had bootstrap support of 71 % in the NJ tree, 75 % in the MP trees and a Bayesian posterior probability of 1.0. Total and net divergence between *M. bavaricus* and *M. liechtensteini* was estimated at 2.3 % and 1.7 %, respectively, which is the lowest between-species divergence reported within the subgenus *Terricola* (Table 2). Despite the limited sample size, nucleotide diversity of the *M. bavaricus*/*liechtensteini* group was within the range of intraspecific nucleotide diversity in other *Terricola* species (Fig. 3).

Discussion

The karyotype of the *M. bavaricus* male studied is almost identical with those reported from various parts of the range of *M. liechtensteini* (e.g., Petrov & Živković 1971, Storch & Winking 1977). The autosome sets are apparently quite similar, and the only difference between the karyotypes reported for individual populations of *M. liechtensteini* and *M. bavaricus* can be found in the size and the centromeric position of

Table 2. Total (Dxy; below the diagonal) and net (Da; above the diagonal) divergence (Kimura 2-parameter distances) between species of the *Terricola* subgenus based on partial cytochrome *b* gene sequences (650 bp). Number of sequences in parenthesis. Numbers in bold represent total and net divergence between *Microtus bavaricus* and *M. liechtensteini*.

	Microtus													
Dxy/Da	*bav.*	*brach.*	*dagh.*	*duod.*	*felt.*	*gerbei*	*liecht.*	*lusit.*	*majori*	*multipl.*	*savii*	*subt.*	*tatr.*	*thom.*
M. bavaricus (3)		0.090	0.076	0.074	0.070	0.058	**0.017**	0.078	0.089	0.058	0.098	0.067	0.076	0.074
M. brachycercus (3)	0.099		0.109	0.094	0.108	0.092	0.077	0.094	0.114	0.100	0.044	0.098	0.102	0.094
M. daghestanicus (3)	0.092	0.129		0.097	0.078	0.073	0.080	0.100	0.090	0.078	0.115	0.052	0.088	0.077
M. duodecimcostatus (4)	0.088	0.112	0.123		0.083	0.066	0.066	0.030	0.082	0.072	0.101	0.091	0.083	0.080
M. felteni (1)	0.072	0.114	0.092	0.095		0.067	0.076	0.083	0.101	0.070	0.106	0.078	0.077	0.061
M. gerbei (5)	0.065	0.104	0.092	0.083	0.072		0.053	0.069	0.075	0.067	0.090	0.069	0.068	0.059
M. liechtensteini (2)	**0.023**	0.087	0.098	0.082	0.080	0.063		0.069	0.077	0.056	0.087	0.066	0.074	0.070
M. lusitanicus (3)	0.084	0.104	0.117	0.045	0.086	0.078	0.077		0.085	0.075	0.106	0.094	0.083	0.079
M. majori (3)	0.092	0.121	0.106	0.096	0.103	0.082	0.083	0.090		0.084	0.125	0.085	0.093	0.099
M. multiplex (5)	0.065	0.111	0.097	0.088	0.075	0.077	0.065	0.083	0.090		0.107	0.075	0.068	0.066
M. savii (2)	0.101	0.051	0.129	0.114	0.107	0.096	0.092	0.111	0.127	0.113		0.098	0.109	0.092
M. subterraneus (6)	0.089	0.124	0.085	0.123	0.098	0.094	0.090	0.117	0.106	0.099	0.118		0.086	0.073
M. tatricus (10)	0.081	0.112	0.105	0.098	0.080	0.077	0.081	0.089	0.098	0.076	0.113	0.109		0.075
M. thomasi (5)	0.088	0.113	0.103	0.104	0.073	0.077	0.086	0.095	0.112	0.083	0.104	0.105	0.090	

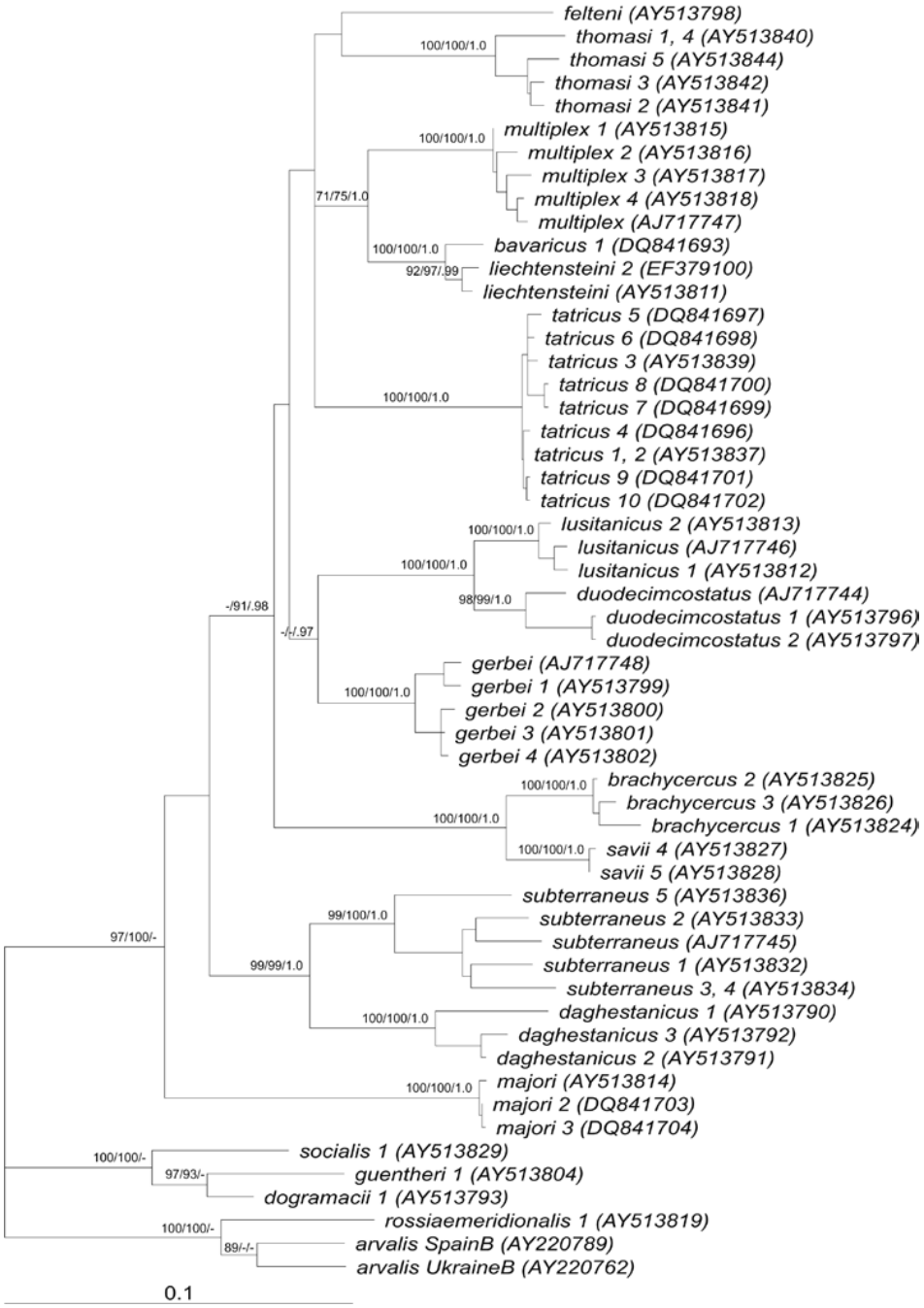


Fig. 2. Maximum likelihood phylogenetic tree (-ln L = 7584.58) based on the GTR+I+Γ substitution model showing the inferred phylogenetic relationships among 49 *Terricola* cytochrome *b* haplotypes (1143 bp). The tree was rooted with representatives of *Microtus sensu stricto* subgenus. Numbers above branches represent bootstrap support based on neighbour-joining, maximum parsimony analysis and Bayesian posterior probability, respectively. Only values for major branches greater than 70 % and 0.95, respectively, are shown.

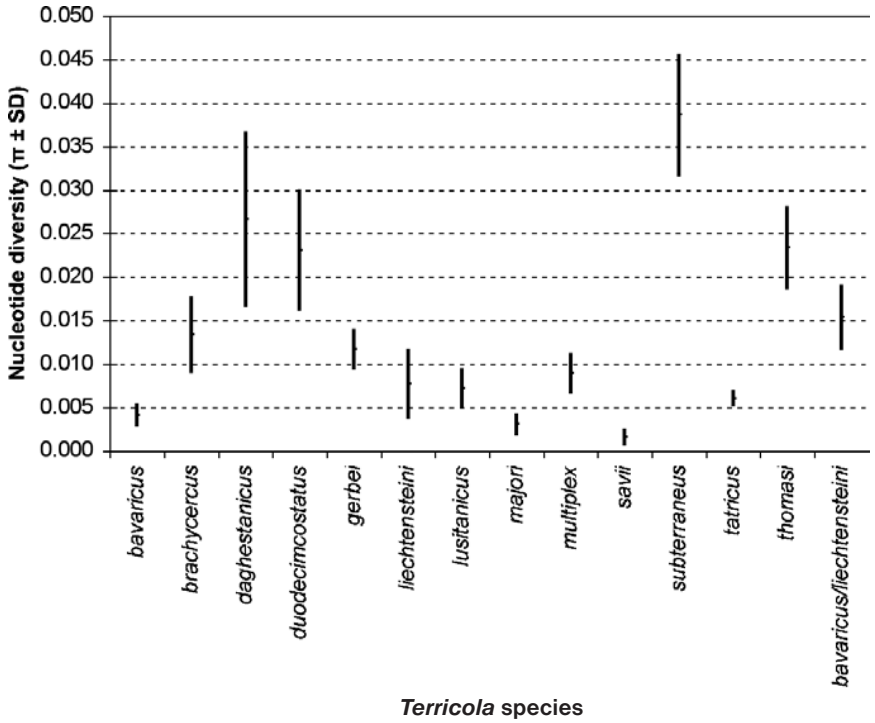


Fig. 3. Comparison of variation of within-species nucleotide diversity (π ± SD) in the *Terricola* subgenus and between-species divergence of the *Microtus bavaricus/liechtensteini* group based on partial cytochrome *b* gene sequences (650 bp).

the Y chromosome. The Y chromosome observed in *M. bavaricus* is almost identical to that of a population of *M. liechtensteini* from northern Velebit Mts in Croatia studied by Petrov & Živković (1974). On the other hand, the Y chromosome of males from *M. liechtensteini* populations from southern Austria (Defereggen-Gebirge, Karnische Alpen) was subtelocentric and distinctly larger than a similar metacentric pair of small autosomes (Král et al. 1978). Submetacentric and subtelocentric morphs of the Y chromosome were found also in populations of *M. liechtensteini* from northern Italy (Trento and Belluno provinces; Storch & Winking 1977). The distribution of individual morphs of the Y chromosome within the range of *M. liechtensteini* has no distinct geographic pattern, and it probably yields no phylogenetic information. Similar variation in the sex chromosomes was reported also among populations of *M. multiplex* (Graf & Meylan 1980, Brunet–Lecomte & Volobouev 1994). The karyotype structure in the studied male of *M. bavaricus* thus indicates its close relatedness to populations of *M. liechtensteini*, and no distinct specific features were observed.

The phylogenetic inference derived from complete cytochrome *b* sequences confirms the close relationship of *M. bavaricus* and *M. liechtensteini* reported by Haring et al. (2000) based on mitochondrial control region sequences. Similarly, in accordance with previous phylogenetic analyses (Haring et al. 2000, Jaarola et al. 2004), our data demonstrate that *M. multiplex, M. liechtensteini* and *M. bavaricus* represent a well supported monophyletic lineage.

The primary divergence within the *M. multiplex* complex occurs between *M. multiplex sensu stricto* and the sister species *M. liechtensteini* and *M. bavaricus*. The total and net divergence between *M. bavaricus* and *M. liechtensteini* is the lowest observed between any pine vole species, indicating a very recent origin of the two taxa.

The branching pattern within the *M. multiplex* complex, namely the sister relationship of *M. multiplex* to closely related but reciprocally monophyletic *M. bavaricus* and *M. liechtensteini*, suggests that the ancestral population of the complex survived the last glaciations at the rims of the ice sheet covering the Alps and/or in the unglaciated mountainous areas. The ancestral population became divided into two glacial refugia, one situated probably in the southwestern and western Alps, while the second refugium occurred in the south, east and north of the Alpine main ridge. This geographic isolation caused speciation of *M. multiplex* and subsequently lead to a split between *M. liechtensteini* and *M. bavaricus*. Additionally, the range of *M. liechtensteini* occurring north of the main alpine ridge became fragmented leading to the apparent isolation of the contemporary populations of *M. liechtensteini* in Niedere Tauern, Salzburg and Totes Gebirge, Styria (Spitzenberger 2002). A similar scenario can also be proposed for the origin of the current *M. bavaricus* populations.

Our cytochrome *b* results and the control region sequence analyses reported by Haring et al. (2000) are congruent with respect to the phylogenetic relationships of the *M. multiplex* complex as well as other species of pine voles. Specifically, the data show that *M. tatricus* and *M. subterraneus* are not closely related to the *M. multiplex* complex and that they belong to other lineages of pine voles (*cf.* Kratochvíl 1970, Jaarola et al. 2004). The topology of the phylogenetic tree presented here indicates an apparent discord between molecular and chromosomal data (see Zima & Král 1984 for review) since monophyletic lineages contain species with distinctly divergent karyotypes. For example, the sister groups of the *M. multiplex* complex with 2N = 46-48 are *M. tatricus* (2N = 32), *M. felteni* (2N = 54), and *M. thomasi* (2N = 40-44), whereas *M. gerbei* (2N = 54) is a sister group of the Iberian species, *M. duodecimcostatus* and *M. lusitanicus* (2N = 62). Altogether, the data strongly suggest that morphological, chromosomal and molecular evolution has proceeded independently during pine vole evolution, and that each evolutionary process has had its own specific rate. This model may be applied for divergence between species as well as between populations within single species. The low mitochondrial DNA variability observed in many pine vole species today suggests that chromosomal evolution in this subgenus could have been facilitated by small historical population sizes.

We conclude that both our chromosomal and mitochondrial DNA findings demonstrate a close affinity between *M. bavaricus* and *M. liechtensteini*. The pattern of evolutionary divergence between populations of this group must have been rather complex in the past, and some populations probably survived the last glaciation period in refugia situated in the northern Alps.

Acknowledgement

We are grateful to Heikki Henttonen for providing a sample of *Microtus liechtensteini* and to two anonymous reviewers for their helpful comments on the earlier version of the manuscript. This study was supported by a grant from the Ministry of Education of the Czech Republic (no. LC06073) and the Carl Tryggers Foundation.

LITERATURE

Brunet-Lecomte P. & Volobouev V. 1994: Comparative morphometry and cytogenetics of *Microtus* (*Terricola*) *multiplex* (Arvicolidae, Rodentia) of the western French Alps. *Z. Säugetierkd. 59: 116–125.*

Carleton M.D. & Musser G.G. 2005: Superfamily Muroidea. In: Wilson D.E. & Reeder D.M. (eds), Mammal species of the world. A taxonomic and geographic reference. 3rd ed. *The Johns Hopkins University Press, Baltimore: 894–1531.*

Chaline J. 1987: Arvicolid data (Arvicolidae, Rodentia) and evolutionary concepts. *Evol. Biol. 21: 237–310.*

Galewski T., Tilak M.-K., Sanchez S., Chevret P., Paradis E. & Douzery E. J. P. 2006: The evolutionary radiation of Arvicolinae rodents (voles and lemmings): relative contribution of nuclear and mitochondrial DNA phylogenies. *BMC Evol. Biol. 6: 80.*

Graf J.-D. & Meylan A. 1980: Polymorphisme chromosomique et biochemique chez *Pitymys multiplex* (Mammalia, Rodentia). *Z. Säugetierkd. 45: 133–148.*

Hall T.A. 1999: BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser. 41: 95–98.*

Haring E., Herzig-Straschil B. & Spitzenberger F. 2000: Phylogenetic analysis of Alpine voles of the *Microtus multiplex* complex using the mitochondrial control region. *J. Zool. Syst. Evol. Research 38: 231–238.*

Jaarola M., Martínková N., Gündüz İ., Brunhoff C., Zima J., Nadachowski A., Amori G., Bulatova N., Chondropoulos B., Fraguedakis-Tsolis S., González-Esteban J., Lopez-Fuster M.J., Kandaurov A., Kefelioğlu H., Mathias M.L., Villate I. & Searle J.B. 2004: Molecular phylogeny of the speciose vole genus *Microtus* (Arvicolinae, Rodentia) inferred from mitochondrial DNA sequences. *Mol. Phylogenet. Evol. 33: 647–663.*

Kimura M. 1980: A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol. 16: 111–120.*

König C. 1962: Eine neue Wühlmaus aus der Umgebung Garmisch-Partenkirchen (Oberbayern): *Pitymys bavaricus. Senckenbergiana biol. 43: 1–10.*

König C. 1982: *Microtus bavaricus* (König, 1962) – Bayerische Kurzohrmaus. In: Niethammer J. & Krapp F. (eds), Handbuch der Säugetiere Europas 2/1. *Akademische Verlagsges., Wiesbaden: 447–451.*

Král B., Zima J. & Herzig-Straschil B. 1978: Karyotype analysis of voles of the genus *Pitymys* from southern Austria. *Folia Zool. 27: 129–133.*

Krapp F. 1982: *Microtus multiplex* (Fatio, 1905) – Alpen-Kleinwühlmaus. In: Niethammer J. & Krapp F. (eds), Handbuch der Säugetiere Europas 2/1. *Akademische Verlagsges., Wiesbaden: 419–428.*

Kratochvíl J. 1970: *Pitymys*-Arten aus der Hohen Tatra (Mamm., Rodentia). *Acta Sc. Nat. Brno 4(12): 1–63.*

Kumar S., Tamura K. & Nei M. 2004: MEGA3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment. *Briefings in Bioinformatics 5: 150–163.*

Martínková N. & Dudich A. 2003: The fragmented distribution range of *Microtus tatricus* and its evolutionary implications. *Folia Zool. 52: 11–22.*

Mitchell-Jones A.J., Amori G., Bogdanowicz W., Kryštufek B., Reijnders P.J.H., Spitzenberger F., Stubbe M., Thissen J.B.M., Vohralík V. & Zima J. 1999: The atlas of European mammals. *Academic Press, London.*

Petrov B. & Živković S. 1971: Zur Kenntnis der *Pitymys liechtensteini* Wettstein, 1972 (Rodentia, Mammalia) in Jugoslawien. *Arhiv bioloških nauka, Beograd 23: 31–32.*

Petrov B. & Živković S. 1974: Der taxonomische Status einiger Vertreter der Untergattung *Pitymys* (Rodentia, Mammalia) in Jugoslawien im Lichte der Daten über ihren Karyotyp. In: Kratochvíl J. & Obrtel R. (eds), Symposium Theriologicum II. *Academia, Praha: 283–290.*

Posada D. & Crandall K.A. 1998: Modeltest: testing the model of DNA substitution. *Bioinformatics 14: 817–818.*

Ronquist F. & Huelsenbeck J. P. 2003: MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics 19: 1572–1574.*

Rozas J., Sánchez-DelBarrio J.C., Messeguer X. & Rozas R. 2003: DnaSP, DNA polymorphism analysis by the coalescent and other methods. *Bioinformatics 19: 2496–2497.*

Spitzenberger F. 2002: Bayerische Kurzohrmaus *Microtus bavaricus* (König, 1962); Illyrische Kurzohrmaus *Microtus liechtensteini* (Wettstein, 1927). In: Spitzenberger F. (ed.), Die Säugetierfauna Österreichs. *Grüne Reihe des Bundesministeriums für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft, Band 13, Graz: 441–449.*

Spitzenberger F., Brunet-Lecomte P., Nadachowski A. & Bauer K. 2000: Comparative morphometrics of the first lower molar *in Microtus (Terricola)* cf. *liechtensteini* of the Eastern Alps. *Acta Theriol. 45: 471–483.*

Storch G. & Winking H. 1977: Zur Systematik der *Pitymys multiplex-Pitymys liechtensteini*-Gruppe (Mammalia, Rodentia). *Z. Säugetierkd. 42: 78–88*.

Swofford D.L. 2003: PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. *Sinauer Associates, Sunderland, Massachusetts*.

Tavaré S. 1986: Some probabilistic and statistical problems on the analysis of DNA sequences. *Lec. Math. Life Sci. 17: 57–86*.

Tougard C., Brunet-Lecomte P., Fabre M. & Montuire S. 2007: Evolutionary history of two allopatric *Terricola* species (Arvicolinae, Rodentia) from molecular, morphological and palaeontological data. *Biol. J. Linn. Soc., in press*.

Zima J. & Král B. 1984: Karyotypes of European mammals II. *Acta Sc. Nat. Brno 18(8): 1–62*.

Paper 2.1.4

Pečnerová P., **Martínková N.** 2012. Evolutionary history of tree squirrels (Rodentia, Sciurini) based on multilocus phylogeny reconstruction. *Zoologica Scripta* 41: 211-219.

Zoologica Scripta

Evolutionary history of tree squirrels (Rodentia, Sciurini) based on multilocus phylogeny reconstruction

Patrícia Pečnerová & Natália Martínková

Submitted: 6 October 2011
Accepted: 13 December 2011
doi:10.1111/j.1463-6409.2011.00528.x

Pečnerová, P. & Martínková, N. (2012). Evolutionary history of tree squirrels (Rodentia, Sciurini) based on multilocus phylogeny reconstruction. —*Zoologica Scripta*, *41*, 211–219.
Tree squirrels of the tribe Sciurini represent a group with unresolved phylogenetic relationships in gene trees. We used partial sequences of mitochondrial genes for 12S rRNA, 16S rRNA, cytochrome *b* and d-loop, and nuclear *irbp*, *c-myc* exon 2 and 3 and *rag1* genes to reconstruct phylogenetic relationships within the tribe, maximizing the number of analysed species. Bayesian inference analysis of the concatenated sequences revealed common trends that were similar to those retrieved with supertree reconstruction. We confirmed congruence between phylogeny and zoogeography. The first group that diverged from a common ancestor was genus *Tamiasciurus*, followed by Palaearctic *Sciurus* and Indomalayan *Rheithrosciurus macrotis*. Nearctic and Neotropical *Sciurus* species formed a monophyletic group that included *Microsciurus* and *Syntheosciurus*. Neotropical Sciurini were monophyletic with a putative exception of *Syntheosciurus brochus* that was included in a polychotomy with Nearctic *Sciurus* in supertree analyses. Our data indicate that Sciurini tree squirrels originated in the northern hemisphere and ancestors of contemporary taxa attained their current distribution through overland colonization from the nearest continent rather than through trans-Pacific dispersal.
Corresponding author: *Natália Martínková, Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, v.v.i., Květná 8, 603 65 Brno, Czech Republic. E-mail: martinkova@ivb.cz and Institute of Biostatistics and Analyses, Masaryk University, Kamenice 3, 625 00 Brno, Czech Republic. E-mail: martinkova@ivb.cz*
Patrícia Pečnerová, Department of Botany and Zoology, Masaryk University, Kotlářská 2, 602 00 Brno, Czech Republic. E-mail: pata.pecnerova@gmail.com

Introduction

Tree squirrels of the tribe Sciurini are sciuromorph rodents that favour arboreal habitat. They are common in temperate forests of the northern hemisphere, where they feed on seeds, often storing them in underground food caches. In the tropical forests, Sciurini tree squirrels are found in Borneo and in Central and South America (Nowak 1999).

Five genera belong to the tribe, *Microsciurus*, *Sciurus*, *Syntheosciurus*, *Rheithrosciurus* and *Tamiasciurus*. Most of the 37 currently recognized species of the tribe are distributed in Central America and northern part of South America, and only four species inhabit the large area of Eurasia, one of which is found in Borneo (Nowak 1999; Wilson & Reeder 2005).

Distribution of Sciurini squirrels exhibits latitudinal gradient in species richness (Wiens *et al.* 2006; Mittelbach *et al.* 2007; Fuhrman *et al.* 2008) with higher number of species occurring in the tropics than in more northern latitudes. In Sciuridae, the high species richness in equatorial regions on different continents is probably a result of higher diversification rate in the tropics based on higher lineage 'birth' and 'death' rates (Roth & Mercer 2008). Other hypotheses that explain the species richness gradient from tropics to the poles assume that lineages in the tropics are older or occupy larger areas facilitating lineage diversification through historical and spatial processes that influence populations (Stebbins 1974; Rosenzweig 1995; Chown & Gaston 2000; Fine 2001; Stephens & Wiens 2003; Wiens & Donoghue 2004; Wiens *et al.* 2006).

The earliest known well-preserved fossil of the family Sciuridae is *Douglassciurus jeffersoni* from North America that lived about 36 million years ago (Mya; Emry & Thorington 1982; Mercer & Roth 2003). Diversification of the tribe Sciurini began later with divergence of the North American genus *Tamiasciurus* about 13 Mya (Mercer &

Roth 2003). This date, based on multilocus molecular dating (Mercer & Roth 2003), coincides with the oldest *Sciurus* fossil, *S. olsoni*, which was dated to the Clarendonian period (approximately 10–13 Mya) in Nevada (Emry *et al.* 2005). However, Oshida *et al.* (2009) suggested that Eurasia might be the most probable place of early divergence of the genus *Sciurus* with subsequent dispersal through Beringia to North America. If the tribe Sciurini originated in North America, this scenario would assume that common ancestors of *Sciurus* dispersed to Eurasia after diversification of *Tamiasciurus*, diverged, and subsequently, as *Sciurus*, returned to North America conditioned by opening of the Bering Strait. Alternatively, as Eurasian Sciurini squirrels are not monophyletic based on mitochondrial DNA phylogeny (Oshida *et al.* 2009), their ancestors could have crossed from North America to Eurasia multiple times. In both cases, the extension of the genus *Sciurus* proceeded to Central America and, after formation of the Isthmus of Panama about three Mya, also to South America (Fig. 1A). In tropical parts of Central and South America, the Sciurini squirrels reach the highest levels of species richness.

Support for a second scenario assuming Eurasian origin of the tribe Sciurini can be found again in the fossil record. The earliest fossil with evident traits of modern squirrels was *Palaeosciurus* from the Oligocene period in France that was most likely a ground dweller (Thorington & Ferrell 2006; Michaux *et al.* 2008). It is considered a putative ancestor of Sciurini and Pteromyini (Thorington & Santana 2007). In this scenario, the group originated in Eurasia and colonized the Americas twice. First, the ancestors of contemporary *Tamiasciurus* and the second time the ancestors of *Sciurus* sensu lato crossed Beringia and further diversified in the Americas (Fig. 1B).

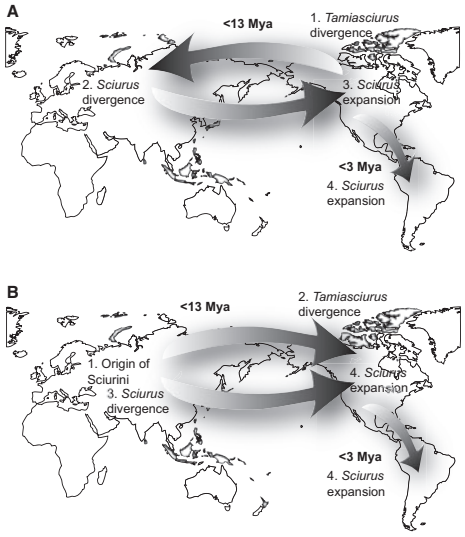


Fig. 1 Colonization and diversification scenarios of the tribe Sciurini. —A. Origin of Sciurini in North America. —B. Origin of Sciurini in Eurasia. The indicated dates are based on molecular dating analysis by Mercer & Roth (2003) and opening of the Isthmus of Panama for passage to South America.

This explains the evolutionary history of the genera with known fossil record, *Sciurus* and *Tamiasciurus*, but not the other currently recognized genera belonging to Sciurini. *Microsciurus*, *Rheithrosciurus* and *Syntheosciurus* are distributed in Central and South America, and in south-east Asia (Wilson & Reeder 2005), and their direct fossil ancestors are not known from the area of their current distribution. The tropical regions are generally unsuitable for fossil preservation (Tappen 1994); hence, the timing of colonization is open to speculation. The lineage of ancestors of *Rheithrosciurus* occupying Eurasia is either extinct without known fossil remains or the genus supposedly colonized Borneo in an independent long-distance colonization event (Mercer & Roth 2003). Genera *Microsciurus* and *Syntheosciurus* from South America are considered late colonizers together with the extension of the genus *Sciurus* following the opening of the Isthmus of Panama (Mercer & Roth 2003; Roth & Mercer 2008).

Our intention was to establish the phylogenetic relationships between species of the tribe Sciurini in a comprehensive way, using supertree reconstruction algorithms and Bayesian inference analysis of a supermatrix of multiple gene alignments. We analysed all species and loci that had sufficient number and length of available DNA sequences to increase robustness of the analyses. We confirmed the origin of the group in the northern hemisphere with two colonization events across Beringia. Sciurini squirrels from the Neotropic ecozone including species from *Sciurus*, *Microsciurus* and *Syntheosciurus* formed a monophyletic group, indicating recent rapid diversification in South and Central America, but *Rheithrosciurus* diverged close to the base of Sciurini phylogeny indicating an overland colonization of Borneo from Asia.

Materials and methods

We analysed partial sequences of eight loci (12S rRNA, 16S rRNA, *mt-cyb*, d-loop, *irbp*, *c-myc* exon 2, *c-myc* exon 3, *rag1*), both mitochondrial and nuclear, that were publicly available (Wettstein *et al.* 1995; Oshida *et al.* 1996; Matthee & Robinson 1997; Barratt *et al.* 1999; Bentz & Montgelard 1999; Arbogast *et al.* 2001; Montgelard *et al.* 2002; Harrison *et al.* 2003; Lance *et al.* 2003; Mercer & Roth 2003;

Steppan *et al.* 2004; Lee *et al.* 2008; Blanga-Kanfi *et al.* 2009; Oshida *et al.* 2009; Chavez *et al.* 2011; B. R. Barber, unpubl. data; R. D. Bradley, G. Ceballos, P. Manzano, F. M. Mendez-Harclerode, M. L. Haynie & D. H. Walker, unpubl. data; A. A. Neverov & D. V. Volokhov, unpubl. data; P. D. Sudman & M. S. Hafner, unpubl. data; L. J. Uimaniemi, M. I. Orell & P. O. Reunanen unpubl. data; N. Yaekashiwa & H. B. Tamate, unpubl. data). We used data sets with partially overlapping species content, comprising in total of 19 species of the tribe Sciurini and two outgroup taxa. Samples of all recently recognized genera of the tribe (*Microsciurus, Rheithrosciurus, Sciurus, Syntheosciurus, Tamiasciurus*) were included. Gene data sets included multiple sequences per species (data not shown). For final analyses, the gene data sets were purged to include one sequence per species (see below for details). Accession numbers of sequences in the final data set are available in Table S1.

We aligned the DNA sequences in GENEIOUS 4.7 (Biomatters Ltd., Auckland, New Zealand; Drummond *et al.* 2009). To estimate phylogenetic relationships for each locus separately, we calculated Bayesian inference analysis for every data set in MRBAYES 3.1.2 (Huelsenbeck & Ronquist 2001), utilizing substitution models selected by the Bayesian Information Criterion in MODELTEST 3.7 (Table 1; Posada & Crandall 1998). To optimize chain convergence, we ran five Markov chains Monte Carlo (MCMC) for 2 million generations, sampling trees every 1000th generation. Chain heating parameter was 0.1, and 1 chain swap was attempted every 3rd generation. The burn-in fraction was set to 30%. The outgroup taxa included *Glaucomys volans* and *Pteromys volans*, which were previously identified as sister taxa to Sciurini (e.g. Montgelard *et al.* 2002). The gene trees were assessed visually for intraspecific variability, and single representatives of each species were chosen. For the supertrees approach, the gene trees were purged to include one tree leaf per species, and for the supermatrix approach, we used the respective sequence to concatenate the chimeric sequence to represent the species.

To estimate phylogenetic relationships between all sampled taxa, we used both supermatrix and supertree approaches. The supermatrix contained concatenated sequences of the eight loci, and we inferred the Bayesian phylogenetic tree using partitions corresponding to the loci. The substitution models for each partition were the same as were used for the gene tree inferences, where the model parameters were unlinked. The MCMC ran for 4 million generations, and the chain temperature was 0.08.

We used the gene trees to reconstruct supertrees that would include all taxa with available DNA sequence data. We applied six methods: SuperTriplets (Ranwez *et al.* 2010), MinCut (Semple & Steel 2000), modified MinCut supertrees (Page 2002), standard matrix representation with parsimony (MRP; Baum 1992; Ragan 1992), Purvis-MRP (Purvis-MRP; Purvis 1995) and veto supertree reconstruction (Scornavacca *et al.* 2008). The supertree reconstruction methods combine information from gene trees where each individual gene tree contains a subsample of taxa.

The SuperTriplets method searches for a supertree that is based on decomposition of the source trees to the simplest trees, the triplets (Ranwez *et al.* 2010). The resulting supertree contains medians of triplet relationships as resolved in the source trees. We conducted the analysis in the program SuperTriplets (Ranwez *et al.* 2010).

The MinCut (Semple & Steel 2000) and modified MinCut (Page 2002) are robust to contradictions in the source trees. The algorithms convert the source trees into a graph of edges with allocated weights. The weights are assigned based on the number of trees where an edge occurs in one cluster. The supertree is constructed from the graph after execution of a minimum cut that discards relationships with the lower than threshold weight. The MinCut method does not include information uncontradicted in the source trees, which modified MinCut algorithm accounts for (Page 2002). We worked with these algorithms in the Supertree software (Page 2002).

Table 1 Proportion of missing data in gene alignments and substitution models for each analysed locus with model parameter means estimated from their Bayesian posterior probability distribution

Locus	Species	Sequences	Missing data (%)	Model	κ	α	I
12S rRNA	10	16	25.6	GTR + Γ + I	N/A	0.529	0.448
16S rRNA	9	13	34.6	GTR + Γ	N/A	0.203	N/A
mt-cyb	14	365	25.8	GTR + Γ + I	N/A	0.243	0.329
d-loop	8	378	66.6	GTR + Γ + I	N/A	86.6	0.423
irbp	15	16	5.3	GTR + Γ	N/A	0.149	N/A
c-myc exon 2	5	5	5.7	HKY	4.475	N/A	N/A
c-myc exon 3	6	6	1.3	HKY	5.389	N/A	N/A
rag1	6	7	34.9	HKY + I	6.049	N/A	0.634

κ – transition/transversion rate ratio, α – shape parameter of the Γ distribution, I – proportion of invariable sites, N/A – not available.

Matrix representation with parsimony and Purvis-MRP are consensus methods, where taxa relationships defined in the phylogenetic trees are translated into a binary matrix. We constructed the matrix representation in r8s 1.70 (Sanderson 2003) and analysed it using maximum parsimony in PAUP* 4b10 (Sinauer Associates, Inc., Sunderland, MA) with 10 heuristic search replicates and TBR branch swapping algorithm. Maximum number of the swapped trees was limited to 10 000. The final tree was constructed as a 50% majority rule consensus.

In the veto method, as implemented in PhySIC_IST, the resulting supertree is a combination of the relationships agreed upon by all source trees (Ranwez *et al.* 2007; Scornavacca *et al.* 2008) as opposed to the relationships most favoured by the source trees in the other supertree reconstruction algorithms. We used the veto supertree method without the source tree correction preprocess and with the correction of the source trees where less frequent triplets are dropped if they contradict more frequent ones.

Results

Gene trees

The eight analysed loci represented 9065 base pairs (bp) of genomic sequence. *T. hudsonicus* was the only ingroup species represented in all gene data sets. The alignment of partial 12S rRNA gene was 851 bp long and contained ten taxa, 16S rRNA 1058 bp for nine taxa, *mt-cyb* 1140 bp for 14 taxa, d-loop 1254 bp for eight taxa, *irbp* 1180 bp for 15 taxa, *c-myc* exon 2 674 bp for five taxa, *c-myc* exon 3 956 bp for six taxa, and the alignment for the partial *rag1* gene was 2141 bp long for six taxa. Six gene alignments contained multiple sequences per species, including sequences that were shorter than the overall gene alignment. This resulted in missing data content (alignment gaps) ranging from 1.3% in the *c-myc* exon 3 alignment to 66.6% in the d-loop alignment (Table 1).

Presented gene trees were purged of intraspecific variability to include one representative from within each species clade. Six of eight phylogenetic trees yielded by Bayesian inference analyses of gene alignments revealed a similar basic pattern in the distribution of taxa with successive divergence of *Tamiasciurus*, followed by Palaearctic, Nearctic and Neotropical species of Sciurini (Fig. 2). The gene tree for 12S rRNA showed *Tamiasciurus* included in a poorly resolved group of New World taxa (Fig. 2A) and the d-loop tree depicted a basal polychotomy of Sciurini tree squirrels (Fig. 2D). Unresolved polychotomies were present in all data sets, but several relationships were supported [Bayesian posterior probability (BPP) ≥ 0.95] in individual gene trees.

Sciurus lis and *S. vulgaris* were confirmed as sister taxa based on the 12S rRNA, *mt-cyb* and d-loop gene trees. According to the *mt-cyb* gene tree, *S. anomalus* diverged after split of the ancestor of *S. lis* and *S. vulgaris*. The other Palaearctic genus, represented by *Rheithrosciurus macrotis* (12S rRNA, 16S rRNA and *irbp*), diverged early but without significant support of its relationship to Palaearctic *Sciurus*. New World taxa formed supported monophyletic groups in five data sets (16S rRNA, *mt-cyb*, *c-myc* exon 2, *c-myc* exon 3 and *rag1*). In the *irbp* gene tree, the support for monophyly of the New World group was low (BPP = 0.69). In addition to the differentiation of the New World clade, Neotropical taxa were supported as a monophylum in *mt-cyb* and *c-myc* exon 3 trees and occurred as a distinct, unsupported lineage in *irbp* (BPP = 0.45) and *rag1* (BPP = 0.77) trees. Relationships among the Nearctic and Neotropical species were unresolved (Fig. 2).

To create the concatenated data set, one sequence per species per gene was chosen. The selected sequence was as long as possible for the given gene and its position with respect to other species is shown on the gene trees (Fig. 2). The concatenated data set had 64.4% of missing data. Bayesian phylogenetic tree based on the concatenated data set distinguished two main groups, *Tamiasciurus* and a lineage including *Sciurus*, *Rheithrosciurus*, *Microsciurus* and *Syntheosciurus* (Fig. 3A). In *Tamiasciurus*, the phylogeny showed significant relationships, where *T. hudsonicus* was a sister species to a lineage including *T. douglasii* and *T. mearnsi*. Within the latter group, the New World tree squirrels formed a significantly supported monophyletic group with the Nearctic taxa branching first. The analysis retrieved a monophyletic lineage containing all Neotropical taxa, but relationships within the group were mostly unresolved. Of five pairs of sister taxa apparent in the tree, only *S. vulgaris* with *S. lis* and *S. aestuans* with *S. ignitus* were significantly supported (Fig. 3).

The diverse methods of supertree reconstruction partially coincided in the branching pattern with the Bayesian tree inferred from the concatenated sequence data set (Figs 3B and S1). In particular, SuperTriplets, modified MinCut, MRP and veto supertree without source tree correction preprocessing showed successive diversification of *Tamiasciurus*, Palaearctic/Indomalayan, Nearctic and Neotropical species with *Sy. brochus* included in early radiation of the New World taxa (Figs 3B and S1B,C,E). The MinCut supertree differed in showing successive diversification of *Tamiasciurus* species at the base of the Sciurini clade (Fig. S1A) and veto supertree with source tree correction preprocessing included *Tamiasciurus* in a polychotomy with *S. lis* and *S. vulgaris* (Fig. S1F). The Purvis-MRP supertree showed heretic relationships of the ingroup taxa that were not present in the other trees (Fig. S1D).

The veto supertrees eliminated several taxa because of conflict in source trees. In the veto supertree without

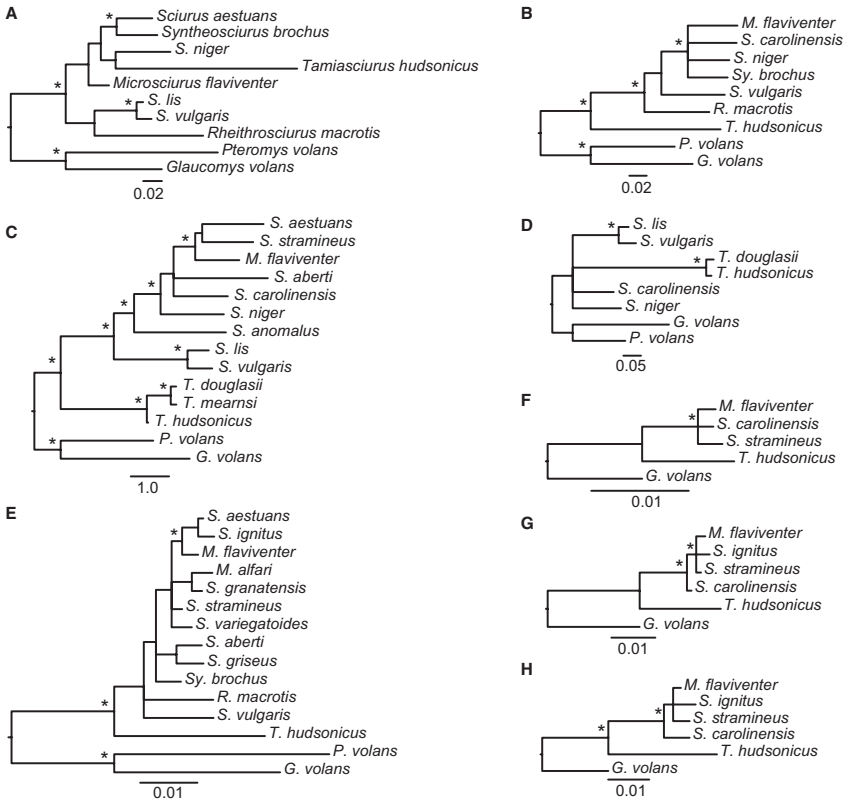


Fig. 2 Bayesian phylogenetic trees of Sciurini based on the mitochondrial and nuclear gene sequences. —A. 12S rRNA. —B. 16S rRNA. —C. *mt-cyb*. —D. d-loop. —E. *irbp*. —F. *c-myc* exon 2. —G. *c-myc* exon 3. —H. *rag1*. The trees were purged from phylogenies that included multiple sequences per species. See text for details. The stars above edges indicate significantly supported relationships (BPP ≥ 0.95). All scale bars show substitutions site^{-1}.

preprocessing, *M. flaviventer* and *S. carolinensis* were excluded (Fig. S1E). The veto supertree with preprocessing did not include five taxa, *S. aestuans*, *S. carolinensis*, *S. niger*, *Sy. brochus* and *R. macrotis* (Fig. S1F).

Discussion

Most trees showed one common pattern in Sciurini phylogeny, where *Tamiasciurus* diverged first in the radiation of Sciurini and the other ingroup species were organized according to their geographic distribution rather than taxonomic status. In wider phylogenetic context, *Tamiasciurus* is consistently placed at the root of the tribe Sciurini (Hafner *et al.* 1994; Mercer & Roth 2003; Herron *et al.* 2004; Steppan *et al.* 2004; Villalobos & Cervantes-Reza 2007; Roth & Mercer 2008). We included three species belonging to the genus, and we confirmed close relationships of *T. douglasii* and *T. mearnsi* with *T. hudsonicus* (Arbogast *et al.* 2001; Herron *et al.* 2004; Chavez *et al.* 2011).

Three Palaearctic *Sciurus* taxa (*S. anomalus*, *S. lis* and *S. vulgaris*) and *R. macrotis* from Borneo branched close to the base of the trees. The New World squirrels formed a monophyletic group. Within the New World clade, there was a gradual transition from the Nearctic to the Neotropical species. The Nearctic species (*S. aberti*, *S. carolinensis*, *S. griseus* and *S. niger*) were paraphyletic with respect to the monophyletic clade comprising the Neotropical species (*M. alfari*, *M. flaviventer*, *S. aestuans*, *S. granatensis*, *S. ignitus*, *S. stramineus* and *S. variegatoides*). *Syntheosciurus*

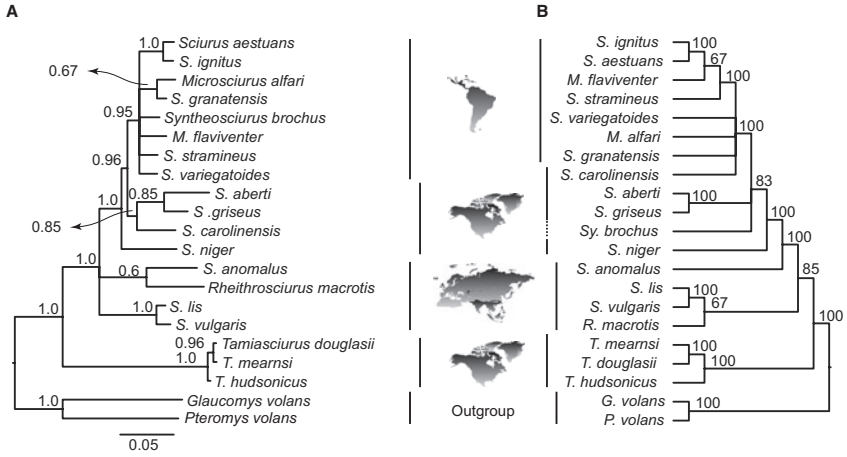


Fig. 3 Multilocus phylogenies of the Sciurini tribe based on information from four mitochondrial and four nuclear loci as depicted in Fig. 2. —A. Bayesian inference phylogenetic tree based on the concatenated supermatrix. The numbers above edges indicate Bayesian posterior probability and the scale bar represents substitutions site^{-1}. —B. SuperTriples supertree based on gene trees in Fig. 2. The numbers above edges indicate percentage of triplets that support the edge in the SuperTriplets analysis. The dashed vertical bar indicates a Neotropical species that grouped with the Nearctic species. Maps of the biogeographic ecozones are not to scale.

brochus was included in the Neotropical group in the Bayesian analysis of the concatenated data set, but it was placed among Nearctic taxa in the supertrees. This discrepancy between supermatrix and supertrees approaches was probably a consequence of variable placement of *Sy. brochus* with respect to *S. aestuans* in the 12S rRNA and the *irbp* gene trees.

Rheithrosciurus is an island species from Borneo, geographically isolated from the nearest representatives of Sciurini that occur in Japan and north-east China (Nowak 1999; Wilson & Reeder 2005). In our study, *R. macrotis* branched early in the trees and diverged after the split of ancestors of *Tamiasciurus*, possibly together with initial diversification of *Sciurus* in Eurasia (see also Mercer & Roth 2003). We believe that *R. macrotis* diverged from a common ancestor as early as the Palaearctic species of the genus *Sciurus* and colonized Borneo overland from south-east Asia. During the evolution of Sciuridae in the last 36 million years, Borneo fauna benefited from multiple land connections to continental Asia (Hall 2001), and given historical biogeography of other arboreal animals from Borneo (examples in Metcalfe *et al.* 2001), the area was forested at least during some of those periods. In light of our reconstruction of the phylogenetic position of *R. macrotis*, the overland colonization hypothesis seems more plausible than trans-Pacific colonization from the Americas.

Two species of *Microsciurus*, which we included in this study, form a close relationship with Central and South American species of the genus *Sciurus*. *M. flaviventer* was included in a supported group with several Neotropical taxa in the *mt-cyb*, *irbp* and *c-myc* exon 3 gene trees. Both *M. flaviventer* and *M. alfari* were sampled in the *irbp* gene tree, but this tree supported monophyly of Sciurini and a group including *M. flaviventer*, *S. aestuans* and *S. ignitus*. Other relationships were not significant. The Bayesian phylogeny inferred from the concatenated data set and the supertrees showed *M. flaviventer* close to the base of the Neotropical clade and *M. alfari* deeper in the group. The *irbp* gene tree showed an unsupported sister relationship of *M. alfari* with *S. granatensis* that was also present in the analysis by Villalobos & Cervantes-Reza (2007). We treat this pattern with caution because *M. alfari* was sampled for the *irbp* gene only, and this gene tree was overall poorly resolved.

Phylogenetic position of the genus *Syntheosciurus* varies in trees obtained with different methods. In the Bayesian tree based on the concatenated data set, *Sy. brochus* belonged to a basal polychotomy of the Neotropical clade, and in four supertrees, it formed a polychotomy with different Nearctic species. The position of *Sy. brochus* was based on the 12S, 16S rRNA and *irbp* gene trees, where the 12S rRNA gene tree showed its supported sister relationship with *S. aestuans*, but in the *irbp* gene tree, the two

species were separate and the tree indicated a more basal position of *Sy. brochus* in a polychotomy with North American *S. aberti* and *S. griseus*. In the 16S rRNA gene tree, *Sy. brochus* belonged to a polychotomy that included also *M. flaviventer*, *S. carolinensis* and *S. niger*. Previously suggested sister relationship with *S. variegatoides* (Villalobos & Cervantes-Reza 2007) was not confirmed in any of our trees.

Evolutionary history

The colonization pathways of Sciurini tree squirrels started most likely in the northern hemisphere. *Tamiasciurus* was a basal group of Sciurini in all our analyses with the exception of the 12S rRNA gene tree. The genus *Tamiasciurus* is distributed in North America, and its basal position in the phylogeny indicates origin of the group in the northern hemisphere. This is supported by the fossil record as known fossils attributed to the tribe were found in North America and Eurasia, but not in tropical regions (Emry & Thorington 1982; Emry *et al.* 2005). Our results do not conclusively distinguish whether the tribe originated in Eurasia or North America. Both continents could be the regions where early diversification of the tribe occurred, the alternative scenarios differing in the directions of the following colonization events across Beringia (Fig. 1).

First, phylogeny of Sciurini tree squirrels indicated that the ancestor of the tribe first diverged in North America and subsequently colonized Eurasia (Oshida *et al.* 2009). After further differentiation in Eurasia, during which time the ancestors of *Rheithrosciurus* entered Sundaland, the *Sciurus* squirrels returned to North America and spread southwards, which manifests in our multilocus phylogeny as a supported monophyletic group of New World *Sciurus* with *Microsciurus* and *Syntheosciurus*.

Second, broader analyses of Sciuridae showed that the sister group of Sciurini was Pteromyini, and members of this tribe are primarily distributed in the Palaearctic region (Montgelard *et al.* 2002; Mercer & Roth 2003; Herron *et al.* 2004; Steppan *et al.* 2004; Michaux *et al.* 2008; Roth & Mercer 2008). If an ancestor of Sciurini was distributed in Eurasia, two eastward colonization events across Beringia would be needed to explain the phylogeny we present here; *Tamiasciurus* first, followed by ancestors of New World *Sciurus*.

Fossil record indicates that *Sciurus* originated in North America, supporting the latter colonization scenario. The earliest *Sciurus* is known from late Miocene in the US, but a representative of the genus appeared as late as Pliocene in Europe (Emry *et al.* 2005), making the oldest Nearctic tree squirrel fossil up to 10 million years older than the oldest Palaearctic one. If the known fossils represented the true diversification history of the group, we should observe a similar branching pattern in the Old World as we see in the Neotropical squirrels. A more recent colonization of Eurasia should appear as a monophyletic lineage with short branches that split deep in the phylogenetic tree. Instead, the Old World taxa are basal. They do not form a monophyletic group, and therefore, we are reluctant to accept a single westward colonization event of a common ancestor of Old World Sciurini squirrels from North America. Multiple independent lineages could explain the observed trees. If multiple lineages entered Eurasia at approximately the same time, they would diversify into the four species, *R. macrotis*, *S. anomalus* and a lineage that differentiated into *S. vulgaris* and *S. lis*. But such a scenario is speculative, as the timing of the diversification would need to occur after the split of the ancestors of *Tamiasciurus*, but before *Sciurus* began to diversify in North America. Rather, we assume that the genus *Sciurus* first diverged in Eurasia. The squirrels then spread eastwards to North America and colonized the continent as far south as Central America. Expansion to South America occurred last, in time with the formation of the Isthmus of Panama (as dated in Mercer & Roth 2003). The species of *Microsciurus* and possibly also *Syntheosciurus* originated in tropical South America after that, and the tribe attained its highest diversification in the region.

The fact that Sciurini tree squirrels from South and Central America form a monophyletic group of closely related taxa in our analyses shows that the latitudinal species richness gradient of the tree squirrels formed recently, probably due to the remarkably high diversification rate suggested by Roth & Mercer (2008). Phylogeography of species from the tribe Sciurini from higher latitudes illustrates that periods of geographic isolation exacerbate intraspecific differentiation. For example, *S. vulgaris* lineage endemic to Calabria in southern Italy differs from other *S. vulgaris* populations in Europe by about 2% in *mt-cyb* gene sequence data, whereas diversity elsewhere in Europe is about 0.3% (Grill *et al.* 2009). Grill *et al.* (2009) explained the difference as a result of an isolation of the Calabrian population during Pleistocene glaciations. Phylogeographic structure of *S. niger* indicates a rapid and recent population expansion, co-occurring with size and coat diversification attributed to postglacial range expansion after the last glaciation (Moncrief *et al.* 2010). Pleistocene glacial cycles probably played a critical role in differentiation of *T. hudsonicus* and *T. douglasii*, where their genetic admixture at present is limited because of different habitat utilization (Arbogast *et al.* 2001; Chavez *et al.* 2011). At the same time, population viability simulations show that a very small number of tree squirrels can establish a new population (Wood *et al.* 2007). We propose that

localized foci established from a limited number of founders that rapidly adapted to specific environmental conditions could have facilitated explosive diversification of Sciurini tree squirrels in Central and South America.

Acknowledgements

We thank Peter Vallo and two anonymous reviewers for commenting on the earlier draft of this manuscript. The analyses were conducted on a computational cluster of the Institute of Vertebrate Biology. This study was funded from grant number AV0Z60930519 provided by the Academy of Sciences of the Czech Republic.

References

Arbogast, B. S., Browne, R. A. & Weigl, P. D. (2001). Evolutionary genetics and Pleistocene biogeography of North American tree squirrels (*Tamiasciurus*). *Journal of Mammalogy*, *82*, 302–319.

Barratt, E. M., Gurnell, J., Malarky, G., Deaville, R. & Bruford, M. W. (1999). Genetic structure of fragmented populations of red squirrel (*Sciurus vulgaris*) in the UK. *Molecular Ecology*, *8*, S55–S63.

Baum, B. R. (1992). Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, *41*, 3–10.

Bentz, S. & Montgelard, C. (1999). Systematic position of the African dormouse *Graphiurus* (Rodentia, Gliridae) assessed from cytochrome *b* and 12S rRNA mitochondrial genes. *Journal of Mammalian Evolution*, *6*, 67–83.

Blanga-Kanfi, S., Miranda, H., Penn, O., Pupko, T., DeBry, R. W. & Huchon, D. (2009). Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. *BMC Evolutionary Biology*, *9*, 71.

Chavez, A. S., Saltzberg, C. J. & Kenagy, G. J. (2011). Genetic and phenotypic variation across a hybrid zone between ecologically divergent tree squirrels (*Tamiasciurus*). *Molecular Ecology*, *20*, 3350–3366.

Chown, S. L. & Gaston, K. J. (2000). Areas, cradles and museums: the latitudinal gradient in species richness. *Trends in Ecology and Evolution*, *15*, 311–315.

Drummond, A. J., Ashton, B., Cheung, M., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Thierer, T. & Wilson, A. (2009). *Geneious v4.7*. Available via http://www.geneious.com/.

Emry, R. J. & Thorington, R. W., Jr (1982). Descriptive and comparative osteology of the oldest fossil squirrel, *Protosciurus* (Rodentia, Sciuridae). *Smithsonian Contributions to Paleobiology*, *47*, 1–35.

Emry, R. J., Korth, W. W. & Bell, M. A. (2005). A tree squirrel (Rodentia, Sciuridae, Sciurini) from the Late Miocene (Clarendonian) of Nevada. *Journal of Vertebrate Paleontology*, *25*, 228–235.

Fine, P. V. A. (2001). An evaluation of the geographic area hypothesis using the latitudinal gradient in North American tree diversity. *Evolutionary Ecology Research*, *3*, 413–428.

Fuhrman, J. A., Steele, J. A., Hewson, I., Schwalbach, M. S., Brown, M. V., Green, J. L. & Brown, J. H. (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 7774–7778.

Grill, A., Amori, G., Aloise, G., Lisi, I., Tosi, G., Wauters, L. A. & Randi, E. (2009). Molecular phylogeography of European *Sciurus vulgaris*: refuge within refugia? *Molecular Ecology*, *18*, 2687–2699.

Hafner, M. S., Barkley, L. J. & Chupasko, J. M. (1994). Evolutionary genetics of New World tree squirrels (tribe Sciurini). *Journal of Mammalogy*, *75*, 102–109.

Hall, R. (2001). Cenozoic reconstructions of SE Asia and the SW Pacific: changing patterns of land and sea. In I. Metcalfe, J. M. B. Smith, M. Morwood & I. D. Davidson (Eds) *Faunal and Floral Migrations and Evolution in SE Asia – Australasia* (pp. 35–56). Lisse: Balkema.

Harrison, R. G., Bogdanowicz, S. M., Hoffmann, R. S., Yensen, E. & Sherman, P. W. (2003). Phylogeny and evolutionary history of the ground squirrels (Rodentia, Marmotinae). *Journal of Mammalian Evolution*, *10*, 249–276.

Herron, M. D., Castoe, T. A. & Parkinson, C. L. (2004). Sciurid phylogeny and the paraphyly of Holarctic ground squirrels (*Spermophilus*). *Molecular Phylogenetics and Evolution*, *31*, 1015–1030.

Huelsenbeck, J. P. & Ronquist, F. (2001). MrBayes: Bayesian inference of phylogeny. *Bioinformatics*, *17*, 754–755.

Lance, S. L., Maldonado, J. E., Bocetti, C. I., Pattee, O. H., Ballou, J. D. & Fleischer, R. C. (2003). Genetic variation in natural and translocated populations of the endangered Delmarva fox squirrel (*Sciurus niger cinereus*). *Conservation Genetics*, *4*, 707–718.

Lee, M.-Y., Park, S.-K., Hong, Y.-J., Kim, Y.-J., Voloshina, I., Myslenkov, A., Saveljev, A. P., Choi, T.-Y., Piao, R.-Z., An, J.-H., Lee, M.-H., Lee, H. & Min, M.-S. (2008). Mitochondrial genetic diversity and phylogenetic relationships of Siberian flying squirrel (*Pteromys volans*) populations. *Animal Cells and Systems*, *12*, 269–277.

Matthee, C. A. & Robinson, T. J. (1997). Molecular phylogeny of the springhare, *Pedetes capensis*, based on mitochondrial DNA sequences. *Molecular Biology and Evolution*, *14*, 20–29.

Mercer, J. M. & Roth, V. L. (2003). The effects of Cenozoic global change on squirrel phylogeny. *Science*, *299*, 1568–1572.

Metcalfe, I., Smith, J. M. B., Morwood, M. & Davidson, I. (2001). *Faunal and Floral Migration and Evolution in SE Asia-Australasia*. Lisse: Balkema.

Michaux, J., Hautier, L., Simonin, T. & Vianey-Liaud, M. (2008). Phylogeny, adaptation and mandible shape in Sciuridae (Rodentia, Mammalia). *Mammalia*, *72*, 286–296.

Mittelbach, G. G., Schemske, D. W., Cornell, H. V., Allen, A. P., Brown, J. M., Bush, M. B., Harrison, S. P., Hurlbert, A. H., Knowlton, N., Lessios, H. A., McCain, C. M., McCune, A. R., McDade, L. A., McPeek, M. A., Near, T. J., Price, T. D., Ricklefs, R. E., Roy, K., Sax, D. F., Schluter, D., Sobel, J. M. & Turelli, M. (2007). Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecology Letters*, *10*, 315–331.

Moncrief, N. D., Lack, J. B. & Van den Bussche, R. A. (2010). Eastern fox squirrel (*Sciurus niger*) lacks phylogeographic structure: recent range expansion and phenotypic differentiation. *Journal of Mammalogy*, *91*, 1112–1123.

Montgelard, C., Bentz, S., Tirard, C., Verneau, O. & Catzeflis, F. M. (2002). Molecular systematics of Sciurognathi (Rodentia): the mitochondrial cytochrome *b* and 12S rRNA genes support the Anomaluroidea (Pedetidae and Anomaluridae). *Molecular Phylogenetics and Evolution*, *22*, 220–233.

Nowak, R. M. (1999). *Walker's Mammals of the World*, 6th edn. Baltimore: The Johns Hopkins University Press.

Oshida, T., Masuda, R. & Yoshida, M. C. (1996). Phylogenetic relationships among Japanese species of the family Sciuridae (Mammalia, Rodentia), inferred from nucleotide sequences of mitochondrial 12S ribosomal RNA genes. *Zoological Science*, *13*, 615–620.

Oshida, T., Arslan, A. & Noda, M. (2009). Phylogenetic relationships among the Old World *Sciurus* squirrels. *Folia Zoologica*, *58*, 14–25.

Page, R. D. M. (2002). Modified mincut supertrees. In: R. Guigó & D. Gusfield (Eds) *Proceedings of the Second International Workshop on Algorithms in Bioinformatics. Lecture Notes in Computer Science*, Vol. *2452* (pp. 537–551). London: Springer.

Posada, D. & Crandall, K. A. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics*, *9*, 817–818.

Purvis, A. (1995). A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology*, *44*, 251–255.

Ragan, M. A. (1992). Matrix representation in reconstructiong phylogenetic-relationships among the Eukaryotes. *Biosystems*, *28*, 47–55.

Ranwez, V., Berry, V., Criscuolo, A., Fabre, P. H., Guillemot, S., Scornavacca, C. & Douzery, E. J. P. (2007). PhySIC: a veto supertree method with desirable properties. *Systematic Biology*, *56*, 798–817.

Ranwez, V., Criscuolo, A. & Douzery, E. J. P. (2010). Supertriplets: a triplet-based supertree approach to phylogenomics. *Bioinformatics*, *26*, i115–i123.

Rosenzweig, M. L. (1995). *Species Diversity in Space and Time*. New York: Cambridge University Press.

Roth, V. L. & Mercer, J. M. (2008). Differing rates of macroevolutionary diversification in arboreal squirrels. *Current Science*, *95*, 857–861.

Sanderson, M. J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, *19*, 301–302.

Scornavacca, C., Berry, V., Lefort, V., Douzery, E. J. P. & Ranwez, V. (2008). PhySIC_IST: cleaning source trees to infer more informative supertrees. *BMC Bioinformatics*, *9*, 413.

Semple, C. & Steel, M. (2000). A supertree method for rooted trees. *Discrete Applied Mathematics*, *105*, 147–158.

Stebbins, G. L. (1974). *Flowering Plants: Evolution Above the Species Level*. Cambridge: The Belknap Press of Harvard University Press.

Stephens, P. R. & Wiens, J. J. (2003). Explaining species richness from continents to communities: the time-for-speciation effect in emydid turtles. *American Naturalist*, *161*, 112–128.

Steppan, S. J., Storz, B. L. & Hoffmann, R. S. (2004). Nuclear DNA phylogeny of the squirrels (Mammalia: Rodentia) and the evolution of arboreality from *c-myc* and *rag1*. *Molecular Phylogenetics and Evolution*, *30*, 703–719.

Tappen, M. (1994). Bone weathering in the tropical rain forest. *Journal of Archaeological Science*, *21*, 667–673.

Thorington, R. W., Jr & Ferrell, K. (2006). *Squirrels: The Animal Answer Guide*. Baltimore: Johns Hopkins University Press.

Thorington, R. W., Jr & Santana, E. (2007). How to make a flying squirrel: *Glaucomys* anatomy in phylogenetic perspective. *Journal of Mammalogy*, *88*, 882–896.

Villalobos, F. & Cervantes-Reza, F. (2007). Phylogenetic relationships of Mesoamerican species of the genus *Sciurus* (Rodentia: Sciuridae). *Zootaxa*, *1525*, 31–40.

Wettstein, P. J., Strausbauch, M., Lamb, T., States, J., Chakraborty, R., Jin, L. & Riblet, R. (1995). Phylogeny of six *Sciurus aberti* subspecies based on nucleotide sequences of cytochrome *b*. *Molecular Phylogenetics and Evolution*, *4*, 150–162.

Wiens, J. J. & Donoghue, M. J. (2004). Historical biogeography, ecology, and species richness. *Trends in Ecology and Evolution*, *19*, 639–644.

Wiens, J. J., Graham, C. H., Moen, D. S., Smith, S. A. & Reeder, T. W. (2006). Evolutionary and ecological causes of the latitudinal diversity gradient in hylid frogs: treefrog trees unearth the roots of high tropical diversity. *American Naturalist*, *168*, 579–596.

Wilson, D. E. & Reeder, D. M. (Eds) (2005). *Mammal species of the world. A Taxonomic and Geographic Reference*, 3rd edn. Baltimore: John Hopkins University Press.

Wood, D. J. A., Koprowski, J. L. & Lurz, P. W. W. (2007). Tree squirrel introduction: a theoretical approach with population viability analysis. *Journal of Mammalogy*, *88*, 1271–1279.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Supertrees based on gene trees obtained from four mitochondrial and four nuclear loci.

Table S1. Accession numbers of sequences of Sciurini tree squirrels and the outgroup taxa used for the reconstruction of phylogenetic relationships.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Paper 2.1.5

Kandemir İ., Sözen M., Matur F., Kankiliç T., **Martínková N.**, Çolak F., Özkurt S. Ö., Çolak E. 2012. Phylogeny of species and cytotypes of mole rats (Spalacidae) in Turkey inferred from mitochondrial cytochrome *b* gene sequences. *Folia Zoologica* 61: 25-33.

Folia Zool. – 61 (1): 25–33 (2012)

Phylogeny of species and cytotypes of mole rats (Spalacidae) in Turkey inferred from mitochondrial cytochrome *b* gene sequences

İrfan KANDEMİR[1], Mustafa SÖZEN[2], Ferhat MATUR[2]*, Teoman KANKILIÇ[3], Natália MARTÍNKOVÁ[4], Faruk ÇOLAK[2], Sakir Ö. ÖZKURT[5] and Ercument ÇOLAK[1]

[1] *Department of Biology, Faculty of Science, Ankara University, 06100-Tandoğan, Ankara, Turkey; e-mail: ikandemir@gmail.com, allactaga@hotmail.com*
[2] *Department of Biology, Faculty of Arts and Sciences, Zonguldak Karaelmas University, 67100-Zonguldak, Turkey; e-mail: ferhat.matur@gmail.com, spalaxtr@hotmail.com, farukcolak@gmail.com*
[3] *Department of Biology, Faculty of Arts and Sciences, Nigde University, Niğde, Turkey; e-mail: spalax06@hotmail.com*
[4] *Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, v.v.i., Květná 8, 603 65 Brno, Czech Republic; e-mail: martinkova@ivb.cz*
[5] *Department of Biology Education, Faculty of Education, Ahi Evran University, Kırşehir, Turkey; e-mail: onderÖzkurt64@gmail.com*

Received 4 October 2010; Accepted 26 August 2011

Abstract. We described the genetic variation of cytochrome *b* gene sequences of blind mole rats in Turkey. We examined 47 individuals belonging to nine cytotypes of three superspecies *Nannospalax leucodon*, *N. xanthodon* and *N. ehrenbergi* in the 402bp gene sequence of cytochrome *b*. Phylogenetic analyses showed that relationships between cytotypes were well supported, but deeper divergence between species showed insignificant relationships. Cytotypes of *N. xanthodon* with low diploid number of chromosomes from western Turkey formed a monophyletic group distinct from the populations with higher number of chromosomes (2n = 56-60). The monophyly of *N. xanthodon* was supported with respect to *N. leucodon* (2n = 56) in the Bayesian and maximum likelihood phylogenies. The divergence between two analyzed cytotypes of *N. ehrenbergi* (2n = 52, 2n = 56) was 9.4 %, and the Kilis cytotype (2n = 52) appeared as the basal branch of the whole analysed dataset. *N. ehrenbergi* cytotypes were paraphyletic and they formed unsupported relationships with previously described *N. galili* (2n = 52), *N. golani* (2n = 54), *N. carmeli* (2n = 58) and *N. judaei* (2n = 60) from Israel. The results of this study showed that the *Nannospalax* species complex most likely represents more species than currently recognized, especially in *N. xanthodon*. We suggest that cytotypes of *N. xanthodon* and *N. ehrenbergi* from Turkey should be investigated in detail as possible candidates for being separate species.

Key words: *Nannospalax*, molecular phylogeny, chromosomal form, Anatolia, Thrace

Introduction

The mole rats are adapted for subterranean life. They are distributed in the Palaearctic region, throughout eastern and southeastern Europe, Anatolia, the Caucasus, and the Middle East up to northeastern Africa (Topachevskii 1969, Wilson & Reeder 1993). Their evolutionary history and taxonomic status are difficult to ascertain and molecular, karyological and morphological studies are needed to determine the phylogenetic relationships of mole rats in Turkey (Nevo et al. 1995, Suziki et al. 1996, Sözen et al. 2000, Kankılıç et al. 2005, Ivanitskaya et al. 2008). For several decades, scientists agree that taxonomy of mole rats (Spalacinae, Rodentia) needs a modern revision based on chromosome and molecular genetic data coupled with morphology, physiology and behavior

* *Corresponding Author*

(Savić & Nevo 1990, Kryštufek & Vohralík 2009). The genus *Spalax* differs from representatives of the genus *Nannospalax* in having a less variable diploid number (2n = 60, 62 in *Spalax* versus 2n = 36-60 in *Nannospalax*) and a higher number of subtelocentric chromosomes (NF = 116-124 in *Spalax* versus NF = 72-98 in *Nannospalax* cytotypes) (Lyapunova et al. 1971, Savić & Nevo 1990, Németh et al. 2009), and shows slow chromosomal evolution rate (Kryštufek & Vohralík 2009). Kryštufek & Vohralík (2009) and Németh et al. (2009) ranked *Nannospalax* species as superspecies. Here we use superspecies order for Turkish mole rats instead of species.

According to Wilson & Reeder (1993) and Yiğit et al. (2006), three species, *Nannospalax leucodon*, *N. xanthodon* (senior synonym of *N. nehringi*; Kryštufek & Vohralík 2009) and *N. ehrenbergi*, occur in Turkey. *N. leucodon* is found in the Turkish Thrace, and *N. ehrenbergi* in southeastern Turkey. *N. xanthodon* has a wide distribution area extending over Anatolia.

Karyological studies revealed 17 cytotypes in Turkey: one cytotype (2n = 56) in *N. leucodon*, 11 (2n = 36, 38, 40, 48, 50, 52, 54, 56, 58, 60 and 60R) in *N. xanthodon* and five (2n = 48, 52, 54, 56, 58) in *N. ehrenbergi* (see reviews in Sözen et al. 1999, 2006a, 2011, Coşkun et al. 2006, Kankılıç et al. 2007, 2010). Additionally, Nevo et al. (1994, 1995) recorded the 2n = 62 form, but Ivanitskaya et al. (2008) reported that the 2n = 62 forms should be eliminated from the list of Turkish mole rats. Later Arslan et al. (2011) studied C- and AgNOR banding patterns of three cytotypes (2n = 40, 58 and 60) from southern Anatolia. These karyological results show that one of the most complex chromosomal diversity within the distribution range of the genus *Nannospalax* is found in Turkey. These results complicate the taxonomical status of the cytotypes in *Nannospalax* in Turkey.

While taxonomic status of cytotypes of mole rats in Turkey has been under discussion, Nevo et al. (2001) raised four cytoypes (2n = 52, 54, 58 and 60) in Israel to species level; *Nannospalax galili* (2n = 52), *N. golani* (2n = 54), *N. carmeli* (2n = 58) and *N. judaei* (2n = 60). Arslan et al. (2010) studied mitochondrial DNA (mtDNA) variation between three cytotypes (2n = 40, 58 and 60) in southern Anatolia, and they showed well-supported lineages in the phylogenetic tree. But there is no detailed study based on mtDNA sequences on the genetic structure of cytotypes of mole rats in Western Turkey. This study focused on phylogenetic relationships among cytotypes of *N. xanthodon* in Western Turkey, and *N. leucodon* in the Turkish Thrace and *N. ehrenbergi* in southeast of Turkey inferred from mtDNA cytochrome *b* gene sequences. We aimed to highlight the genetic relationships between and among the cytotypes of the species in Turkey, and also the recently described species of *Nannospalax* from Israel.

Material and Methods

Sampling for molecular studies

A total of 47 mole rat individuals were used in the molecular studies. These samples came from three previously recognized species, namely *N. leucodon* (N = 3), *N. xanthodon* (N = 40; cytotypes 2n = 36, 38, 40, 50, 56, 60) and *N. ehrenbergi* (N = 4; cytotypes 2n = 52 and 56) (Fig. 1). Karyotypes were prepared according to Ford & Hamerton (1956). Skins and skulls were deposited to the Department of Biology of Zonguldak Karaelmas and Ankara University, Turkey.

DNA isolation, amplification, and sequencing

We sequenced all 47 samples for a part of the cytochrome *b* locus. Total genomic DNA was isolated from muscle tissues following the techniques reported by Doyle & Doyle (1987) known as CTAB method. Fragment of the cytochrome *b* gene was amplified using polymerase chain reaction in 25 µl reaction volume and each reaction included 1 µl of each primer (20 pmol) (L14724 and H15154; Irwin et al. 1991,

Fig. 1. *Sampled sites in Turkey and respective mole rat cytotypes.*

Smith & Patton 1993), 4 µl of dNTPs, 2.5 µl of 10× Taq Buffer with $(NH_4)_2SO_4$, 1.5 µl of $MgCl_2$ (25 mM) and 0.25 µl of Taq DNA Polymerase (5 units/µl, Fermentas, Ontario, Canada). The PCR cycling conditions first included four cycles of 1 min of denaturation at 95 °C, 1 min of annealing at 40 °C, and 1 min extension at 72 °C, followed by 33 cycles using annealing temperature at 50 °C. Before the sequencing reaction

Table 1. *Samples used in the present study.*

Species	Map code	2n	Locality	Genbank accession number	Voucher name	Coordinates	
N. leucodon	1	56Tr	Kırklareli	FJ656299	TR-KIRKLARELİ 5097	41°25′34.05″ N	27°7′9.09″ E
	1	56	Kırklareli	FJ656300	TR-KIRKLARELİ 5094	41°25′34.05″ N	27°7′9.09″ E
	1	56	Kırklareli	FJ656301	TR-KIRKLARELİ 5095	41°25′34.05″ N	27°7′9.09″ E
N. xanthodon	2	36	Aydın	FJ656275	TR-AYDIN 6218	37°51′46.19″ N	27°49′27.67″ E
	2	36	Aydın	FJ656276	TR-AYDIN 6256	37°51′46.19″ N	27°49′27.67″ E
	2	36	Aydın	FJ656277	TR-AYDIN 6222	37°51′46.19″ N	27°49′27.67″ E
	2	36	Aydın	FJ656278	TR-AYDIN 6207	37°51′46.19″ N	27°49′27.67″ E
	2	36	Aydın	FJ656279	TR-AYDIN 6250	37°51′46.19″ N	27°49′27.67″ E
	2	36	Aydın	FJ656280	TR-AYDIN 6532	37°51′46.19″ N	27°49′27.67″ E
	2	36	Aydın	FJ656281	TR-AYDIN 5258	37°51′46.19″ N	27°49′27.67″ E
	3	38	Çanakkale	FJ656259	TR-GOKCEADA 4959	40°9′30.61″ N	25°50′30.05″ E
	3	38	Çanakkale	FJ656260	TR-GOKCEADA 4957	40°9′30.61″ N	25°50′30.05″ E
	3	38	Çanakkale	FJ656261	TR-GOKCEADA 4955	40°1′18.10″ N	26°25′26.44″ E
	3	38	Çanakkale	FJ656262	TR-GOKCEADA 4954	40°1′18.10″ N	26°25′26.44″ E
	3	38	Çanakkale	FJ656263	TR-GOKCEADA 4958	40°1′18.10″ N	26°25′26.44″ E
	3	38	Manisa	FJ656282	TR-MANISA 5214	39°6′41.11″ N	27°40′14.72″ E
	3	38	Manisa	FJ656283	TR-MANISA 5211	39°6′41.11″ N	27°40′14.72″ E
	3	38	Balıkesir	FJ656284	TR-BALIKESIR 6120	39°22′14.33″ N	27°59′25.34″ E
	3	38	Balıkesir	FJ656285	TR-BALIKESIR 6161	39°22′14.33″ N	27°59′25.34″ E
	3	38	Manisa	FJ656286	TR-MANISA 5210	39°10′24.88″ N	27°51′18.40″ E
	3	38	Manisa	FJ656287	TR-MANISA 6131	39°10′24.88″ N	27°51′18.40″ E
	3	38	Manisa	FJ656288	TR-MANISA 6153	39°10′24.88″ N	27°51′18.40″ E
	3	38	Izmir	FJ656289	TR-IZMIR 6138	38°27′40.06″ N	27°12′51.77″ E
	4	40	Isparta	FJ656264	TR-ISPARTA 6202	37°43′7.76″ N	30°56′55.64″ E
	4	40	Isparta	FJ656265	TR-ISPARTA 6225	37°43′7.76″ N	30°56′55.64″ E
	4	40	Isparta	FJ656266	TR-ISPARTA 6265	37°43′7.76″ N	30°56′55.64″ E
	4	40	Isparta	FJ656267	TR-ISPARTA 6266	37°43′7.76″ N	30°56′55.64″ E
	4	40	Konya	FJ656268	TR-KONYA 6268	37°30′42.12″ N	31°24′3.10″ E
	4	40	Konya	FJ656269	TR-KONYA 5346	37°30′42.12″ N	31°24′3.10″ E
	4	40	Konya	FJ656270	TR-KONYA 6203	37°30′42.12″ N	31°24′3.10″ E
	4	40	Konya	FJ656271	TR-KONYA 4734	37°30′42.12″ N	31°24′3.10″ E
	4	40	Konya	FJ656272	TR-KONYA 6267	37°30′42.12″ N	31°24′3.10″ E
	5	50	Aydın	FJ656273	TR-AYDIN 5255	38°15′15.61″ N	28°23′15.30″ E
	5	50	Aydın	FJ656274	TR-AYDIN 5256	38°15′15.61″ N	28°23′15.30″ E
	6	56	Uşak	FJ656290	TR-USAK 6224	38°40′28.91″ N	29°14′15.28″ E
	6	56	Manisa	FJ656291	TR-MANISA 6141	38°28′1.76″ N	28°38′36.98″ E
	6	56	Isparta	FJ656292	TR-ISPARTA 4784	37°47′46.80″ N	30°54′23.74″ E
	6	56	Isparta	FJ656294	TR-ISPARTA 6257	37°47′46.80″ N	30°54′23.74″ E
	6	56	Karabük	FJ656297	TR-KARABUK 4858	41°13′21.30″ N	32°43′16.30″ E
	7	60	Manisa	FJ656293	TR-MANISA 6143	38°43′17.87″ N	28°52′13.76″ E
	7	60	Kütahya	FJ656295	TR-KUTAHYA 3683	39°24′5.99″ N	29°16′14.63″ E
	7	60	Kütahya	FJ656296	TR-KUTAHYA 6119	39°24′5.99″ N	29°16′14.63″ E
	7	60R	Kastamonu	FJ656298	TR-KASTAMONU 5607	41°42′25.42″ N	33°35′24.63″ E
N. ehrenbergi	8	56	Osmaniye	FJ656304	TR-OSMANIYE 5301	37°4′34.75″ N	36°14′22.08″ E
	8	56	Osmaniye	FJ656305	TR-OSMANIYE 5302	37°4′34.75″ N	36°14′22.08″ E
	9	52	Kilis	FJ656302	TR-KILIS 5120	36°47′34.56″ N	37°15′27.46″ E
	9	52	Kilis	FJ656303	TR-KILIS 5121	36°47′34.56″ N	37°15′27.46″ E

the PCR fragments were cleaned with Nucleospin extract kit (Macherey-Nagel, Düren, Germany) and later sequencing reactions of both DNA strands were commercially performed using Big Dye Terminator v. 3.1 sequencing chemistry on an ABI 3100 Genetic Analyzer (Applied Biosystems, California, USA). All specimens were deposited to the GenBank under the accession numbers FJ656259-FJ656305 (Table 1).

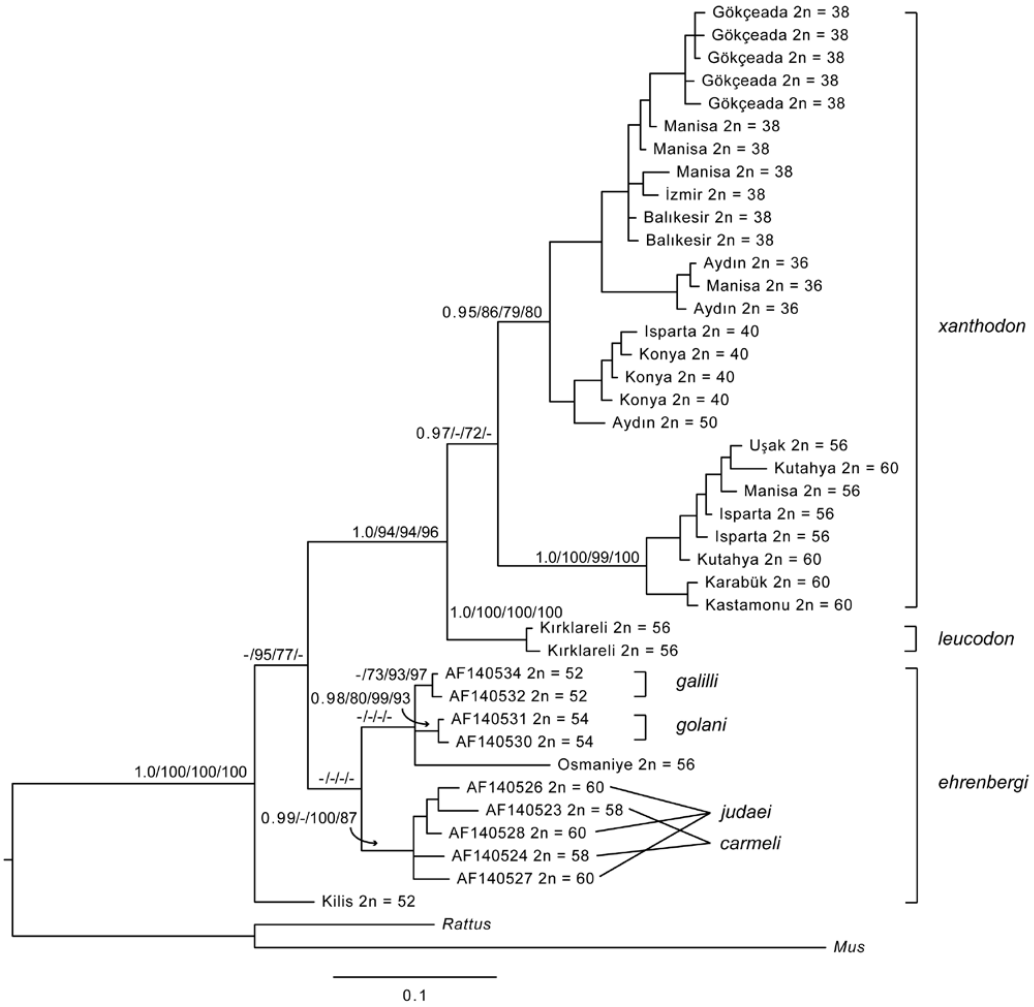


Fig. 2. *Bayesian inference phylogenetic tree based on* Nannospalax *partial sequences of the mitochondrial cytochrome* b *gene (402 bp). Node support from Bayesian posterior probabilities and NJ, ML and MP bootstrapping is shown for main groups. '-' indicates non-significant support, < 70 % for bootstrap analyses and < 0.95 for Bayesian posterior probability.*

Phylogenetic analyses

Previously published cytochrome *b* sequences of *Nannospalax* were obtained from GenBank (AF140523-AF140534; Nevo et al. 1999), and two other sequences for *Mus musculus* (NC010339) and *Rattus norvegicus* (EU273707) were added as an outgroup. The dataset was aligned in Clustal X 2.0.9 (Larkin et al. 2007) and had the total length 402 base-pairs.

Genetic divergence was estimated in *MEGA* 4 (Tamura et al. 2007). Neighbour-joining (NJ) and maximum parsimony (MP) phylogenetic analyses were executed in PAUP 4.0b10 (Swofford 1999), using HKY + Γ substitution model for the NJ analysis selected by Bayesian Information Criterion in Modeltest 3.7 (Posada & Crandall 1998), which was the simplest suitable model, and 10000 bootstrap replicates. Maximum likelihood (ML) tree was obtained from RAxML 7.4 (Stamatakis 2006) with GTR + Γ model and 10000 bootstrap replicates. Bayesian Inference (BI) analysis was run in MrBayes 3.1.2 (Ronquist & Huelsenbeck 2003) with 12 Markov chains Monte

Carlo for two million generations in two independent runs. The chain swapping was successful in 68-72 % of attempts indicating effective chain mixing. We used default prior settings with the exception of the branch length. That was limited to brlenspr = unconstrained: exp(20) to ensure that the posterior 95 % confidence interval of the tree length included the tree length obtained from the maximum likelihood analysis (*cf.* Marshall 2010). The prior setting did not influence tree topology compared to the default setting (data not shown). A median-joining network (Bandelt et al. 1999) was also reconstructed for the sequenced mtDNA region. Finally, tree topologies were compared by using the Shimodaira-Hasegawa test in PAUP 4.0b10 (Swofford 1999).

Results

The 47 sequences represented 42 unique haplotypes, which were used in the phylogenetic analyses. The 402 bp long dataset contained 120 parsimony informative sites and 31 singleton mutations. Support for monophyly of the studied taxa based on partial cytochrome *b* sequences was different for specific analyses (Fig. 2). Total tree length in the MP analysis was 364 steps and the groupings reflected the genetic distances between cytotypes and species.

Nucleotide diversity within cytotypes ranged from 0 to 0.03 substitutions per site (Table 2). Within *N. xanthodon*, the greatest diversity was found in the 2n = 38 cytotype that was also represented by the highest number of individuals in our dataset.

The genetic distances were calculated between all taxa used in the present study using uncorrected *p*-distance (Table 3). The distance between four Israeli *Nannospalax* species ranged between 0.019 and 0.058 substitutions per site. The cytotypes attributed to *N. xanthodon* showed deeper divergence ranging between 0.024 and 0.105. The greatest genetic distance between studied cytotypes was between *N. carmeli* and *N. xanthodon* 2n = 40. Smaller genetic distances tended to be found between the chromosomal cytotypes with the highest diploid chromosome numbers. The genetic distance between the two cytotypes of *N. ehrenbergi* in Turkey was 0.087, which is greater than the genetic distance between the two most distantly related Israeli species (*N. golani* and *N. carmeli*) (Table 3).

Table 2. *Within group nucleotide diversity for partial sequences of the mitochondrial cytochrome* b *gene (402 bp) for identified* Nannospalax *cytotypes.*

Cytotypes	N	Nucleotide diversity	Standard Error
N. galili 2n = 52[1]	2	0.0017	0.0016
N. golani 2n = 54[1]	2	0.0017	0.0016
N. judaei 2n = 60[1]	3	0.0216	0.0058
N. carmeli 2n = 58[1]	2	0.0265	0.0066
N. ehrenbergi 2n = 52[2]	2	0.000	0.000
N. ehrenbergi 2n = 56[2]	2	0.000	0.000
N. xanthodon 2n = 36[2]	7	0.0021	0.0012
N. xanthodon 2n = 38[2]	13	0.0192	0.0044
N. xanthodon 2n = 40[2]	9	0.0072	0.0028
N. xanthodon 2n = 50[2]	2	0.000	0.000
N. xanthodon 2n = 56[2]	5	0.0033	0.0023
N. xanthodon 2n = 60[2]	4	0.0108	0.0038
N. leucodon 2n = 56[2]	3	0.0030	0.0058

[1] Nevo et al. 1999, [2] this study.

Table 3. *Pairwise uncorrected* p*-distance for partial sequences of the mitochondrial cytochrome* b *gene (402 bp) for identified* Nannospalax *cytotypes (below diagonal) and their standard errors (above diagonal).*

	1	2	3	4	5	6	7	8	9	10	11	12	13
N. galili 2n = 52[1]		0.007	0.010	0.010	0.013	0.012	0.014	0.014	0.015	0.014	0.014	0.015	0014
N. golani 2n = 54[1]	0.019		0.011	0.010	0.013	0.013	0.015	0.015	0.015	0.014	0.015	0.015	0.014
N. judaei 2n = 60[1]	0.052	0.056		0.005	0.012	0.014	0.014	0.014	0.015	0.014	0.014	0.014	0.014
N. carmeli 2n = 58[1]	0.056	0.058	0.022		0.012	0.014	0.015	0.014	0.015	0.014	0.014	0.014	0.014
N. ehrenbergi 2n = 52[2]	0.080	0.078	0.076	0.076		0.013	0.014	0.015	0.015	0.015	0.015	0.015	0.015
N. ehrenbergi 2n = 56[2]	0.068	0.066	0.096	0.098	0.087		0.014	0.015	0.015	0.014	0.015	0.015	0.015
N. xanthodon 2n = 40[2]	0.119	0.118	0.120	0.125	0.104	0.108		0.010	0.007	0.011	0.012	0.014	0.013
N. xanthodon 2n = 38[2]	0.119	0.118	0.117	0.123	0.113	0.117	0.059		0.010	0.010	0.012	0.014	0.013
N. xanthodon 2n = 50[2]	0.119	0.118	0.117	0.121	0.104	0.112	0.029	0.053		0.012	0.012	0.014	0.013
N. xanthodon 2n = 36[2]	0.100	0.099	0.107	0.107	0.102	0.097	0.062	0.052	0.065		0.013	0.014	0.013
N. leucodon 2n = 56[2]	0.101	0.107	0.106	0.114	0.111	0.116	0.080	0.080	0.076	0.089		0.014	0.013
N. xanthodon 2n = 56[2]	0.123	0.113	0.117	0.120	0.121	0.124	0.104	0.105	0.100	0.102	0.099		0.005
N. xanthodon 2n = 60[2]	0.122	0.116	0.115	0.119	0.121	0.121	0.102	0.104	0.098	0.099	0.098	0.024	

[1]Nevo et al. 1999, [2] this study.

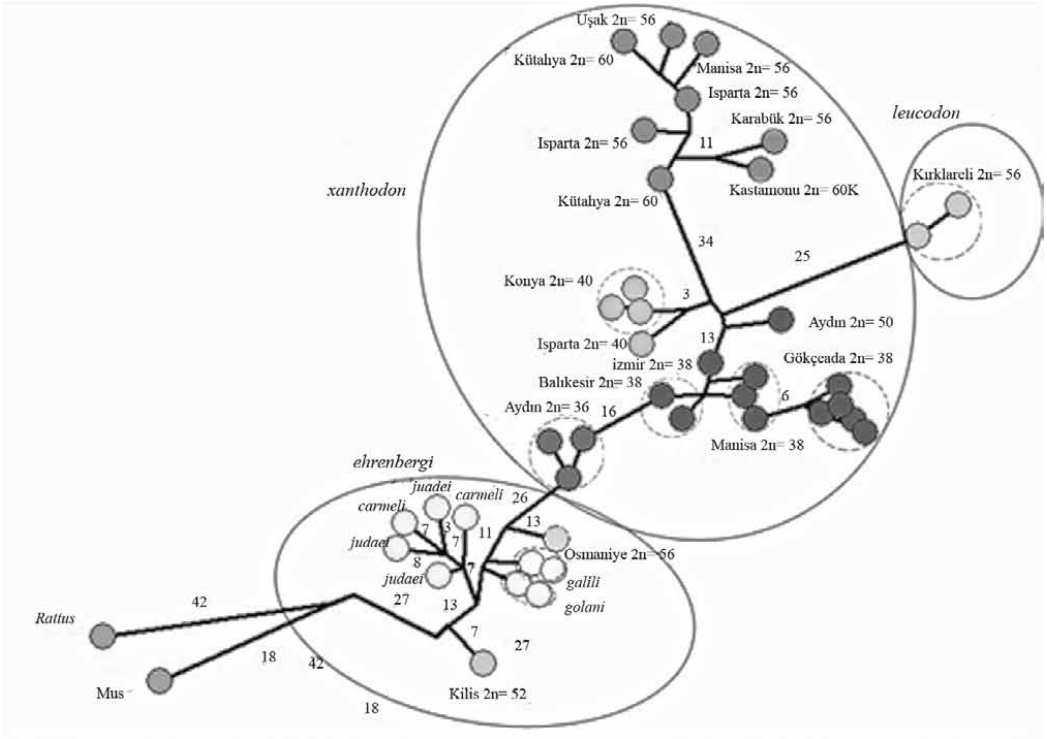


Fig. 3. *The 42 sequences appear as three different species and five clades on median joining network congruent with Fig. 2. The samples from identical localities are indicated with small circles. The node size is not proportional to the haplotype frequency. The numbers on branch show differences between two neighboring haplotypes.*

We conducted four separate sets of analyses. The data were analyzed using Bayesian inference, neighbour-joining, maximum likelihood and maximum parsimony analyses. The trees showed similar topologies (Fig. 2). The result of the Shimodaira-Hasegawa test showed that only NJ tree topology was significantly different from the other trees (Table 4). This result was caused by the placement of the *N. ehrenbergi* clade. In all trees, cytotypes with low diploid chromosome numbers ($2n \leq 50$) grouped as one supported clade.

Table 4. *Results of Shimodaira-Hasegawa (SH) tests for all trees. The NJ tree topology was found different from ML (* $p < 0.05$).*

Tree	–ln L	Diff –ln L	*P*
Bayesian inference	227.988	0.434	0.52
Maximum likelihood	227.553	(best)	
Maximum parsimony	228.626	1.073	0.29
Neighbour-joining	231.095	3.542	0.01*

The other *N. xanthodon* cytotypes (2n = 56 and 60) occurred as a separate clade, but monophyly of *N. xanthodon* was supported only in BI and ML analyses (Fig. 2). *N. leucodon* formed a monophyletic group with *N. xanthodon*. The other taxa showed a basal polytomy in most of our analyses. Both cytotypes of *N. ehrenbergi* were paraphyletic with respect to the Israeli taxa. *N. carmeli* and *N. judaei* formed a monophyletic group, but the taxa were not reciprocally monophyletic within this group.

We found similar topology in the network analysis (Fig. 3). The *N. xanthodon* cytotypes formed two putative groups with low ($2n \leq 50$) and high ($2n \geq 56$) chromosome numbers. The former group was more closely related to *N. leucodon* as well as to the group including *N. ehrenbergi* and the Israeli taxa (Fig. 3).

Discussion

Our phylogenetic analysis showed monophyly of *N. xanthodon* and *N. leucodon*, but paraphyletic

relationships of the two *N. ehrenbergi* sequences with Israeli taxa. *N. xanthodon* sequences formed two reciprocally monophyletic groups that included cytotypes with low (2n ≤ 50) and high (2n ≥ 56) chromosome numbers. Their sister species was *N. leucodon*. These results are similar to recent findings of Arslan et al. (2010) on a smaller dataset including three cytotypes and Kryštufek et al. (2012) that were based on a longer alignment. The taxa *N. ehrenbergi*, *N. galili*, *N. golani*, *N. carmeli* and *N. judaei* diverged in a basal polytomy wherein only a single group including *N. carmeli* and *N. judaei* was consistently retrieved. While Reyes et al. (2003) also showed an undifferentiated clade of *N. carmeli* and *N. judaei* samples, our analyses did not support their clear relationship of *N. galili* and *N. golani*. This was distorted by position of *N. ehrenbergi* 2n = 56 as a sister taxon with unresolved relationship to the Israeli species.

Cytotypes of *N. xanthodon* with low diploid number (2n = 36, 38, 40 and 50) formed a monophyletic group. This may be a result of specific evolutionary pathways. For example, Matur & Sözen (2005) stated that the River Sakarya separated 2n = 52 and 60 cytotypes of *N. xanthodon* in Bilecik province, and the river might act as a barrier between these cytotypes. But in our analyses, 2n = 52 *N. xanthodon* was not analysed and we cannot confirm this scenario. According to our results, 2n = 38 from Gökçeada (Aegean island) differentiated from the 2n = 38 populations in the mainland indicating mole rat diversification in island isolation.

Monophyly of specific cytotypes was present in all cytotypes belonging to the *N. xanthodon* 2n ≤ 50 group, but the cytotypes were not likewise differentiated in the 2n ≥ 56 group. Rather, the northern populations of *N. xanthodon*, Karabük (2n = 56N) and Kastamonu (2n = 60R), were differentiated from the other samples of 2n = 56 and 60. This was suggested by Sözen (2004) and Sözen et al. (2006b), who claimed that the cytotypes from Northern Anatolia (2n = 54N, 56N, 58N and 60R) differentiated from other cytotypes. Ivanitskaya et al. (2008) showed that the 2n = 60 population in Kastamonu is different from the 2n = 60 population in central Anatolia using classical and molecular cytogenetic techniques. Matur et al. (2010) studied chromosomal evolution of Turkish mole rats inferred from G- and C- banding of 11 cytotypes. They found that cytotypes from Kastamonu (2n = 60R) and Karabük (2n = 56R) formed a clade together with 2n = 54 cytotype not included in our study. This northern clade has independent evolutionary pathway, and 2n = 60R might be considered an ancestral karyotype (Matur et al. 2010).

There is a conflict in the ancestral population of Thracian mole rats (*N. leucodon* 2n = 56). They grouped with *N. xanthodon* in our trees. Recently, Matur et al. (2011) proposed the 2n = 60 cytotype as an ancestral for all Anatolian cytotypes. Nevertheless, this process needs more detailed studies in order to verify the ancestral karyotype of the *N. xanthodon/ leucodon* group.

In our phylogenetic analyses, *N. carmeli* (2n = 58) and *N. judaei* (2n = 60) did not appear in separate clusters. Reyes et al. (2003) found similar results from the sequence studies of the mitochondrial control region. In their study, *N. golani* and *N. galili* were well separated, whereas *N. judaei* (2n = 60) and *N. carmeli* (2n = 58) formed a single cluster. Nevo et al. (1999) argued that this is due to the fact that they represent young species. The paraphyly of the two cytotypes of *N. ehrenbergi* (2n = 52 and 2n = 56) from Kilis and Osmaniye further complicate this in our study. Both cytotypes formed unsupported relationships and polychotomies. Tentatively, *N. ehrenbergi* from Osmaniye grouped with *N. galili* and *N. golani*, and *N. ehrenbergi* from Kilis was a basal taxon in the tree (Fig. 2).

Further comprehensive and detailed multidisciplinary research combining morphology, karyology, physiology, behaviour, mtDNA and nuclear DNA phylogeny should be applied to clarify the taxonomic status and phylogenetic relationships of cytotypes of *N. leucodon*, *N. xanthodon*, and *N. ehrenbergi* in Turkey. In addition, research on intra-population variation should also be considered for better understanding of mole rats in Turkey.

Acknowledgements

This study was supported by the Scientific and Technical Research Council of Turkey (TUBITAK 101T084 and 106T225), Zonguldak Karaelmas University (2004-13-06-08, 2008-13-06-01) and Academy of Sciences of the Czech Republic (AV0Z60930519). The authors thank Ahmet Turkyılmaz from REFGEN Biotechnology Inc. for performing the sequencing. Authors thank to Zeynep Yuksel Sezen and Joan E. Johnston for careful editing the manuscript and two anonymous reviewers and the editor for helpful and detailed comments.

Literature

Arslan E., Gulbahce E., Arıkoglu H., Arslan A., Bužan E.V. & Kryštufek B. 2010: Mitochondrial divergence between three cytotypes of the Anatolian mole rat *Nannospalax xanthodon* (Mammalia: Rodentia). *Zool. Middle East. 50: 27–34.*

Arslan A., Akan Ş. & Zima J. 2011: Variation in C-heterochromatin and NOR distribution among chromosomal races of mole rats (Spalacidae) from Central Anatolia,Turkey. *Mamm. Biol. 76: 28–35.*

Bandelt H.J., Forster P. & Röhl A. 1999: Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol. 16: 37–48.*

Coşkun Y., Ulutürk S. & Yürümez G. 2006: Chromosomal, diversity in mote-rats of the species *Nannospalax ehrenbergi* (Rodentia: Spalacidae) from South Anatolia, Turkey. *Mamm. Biol. 71: 244–250.*

Doyle J. & Doyle L. 1987: A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull. 19: 11–15.*

Ford C.E. & Hamerton J.L. 1956: A colchicine hypotonic citrate, squash sequence for mammalian chromosomes. *Stain Technol. 31: 247–251.*

Irwin D., Kocher T. & Wilson A. 1991: Evolution of the cytochrome *b* gene of mammals. *J. Mol. Evol. 32: 128–144.*

Ivanitskaya E., Sözen M., Rashkovetsky L., Matur F. & Nevo E. 2008: Discrimination of 2n = 60 *Spalax leucodon* cytotypes (Spalacidae, Rodentia) in Turkey by means of classical and molecular cytogenetic techniques. *Cytogenet. Genome Res. 122: 139–149.*

Kankılıç T., Çolak E., Çolak R. & Yiğit N. 2005: Allozyme variation in *Spalax leucodon* Nordmann, 1840 (Rodentia: Spalacidae) in the area between Ankara and Beyşehir. *Turk. J. Zool. 29: 377–384.*

Kankılıç T., Kankılıç T., Çolak R., Çolak E. & Karataş A. 2007: Karyological comparison of populations of the *Spalax leucodon* Nordmann, 1840 superspecies (Rodentia: Spalacidae) in Turkey. *Zool. Middle East. 42: 15–24.*

Kankılıç T., Kankılıç T., Seker P.S., Çolak R., Selvi E. & Çolak E. 2010: Contributions to the karyology and distribution areas of cytotypes of *Nannospalax leucodon* (Rodentia: Spalacidae) in Western Anatolia. *Acta Zool. Bulg. 62: 161–167.*

Kryštufek B. & Vohralík V. 2009: Mammals of Turkey and Cyprus. Rodentia II: Cricetinae, Muridae, Spalacidae, Calomyscidae, Capromyida, Hystricidae, Castoridae. *Annales Majora, Koper.*

Kryštufek B., Ivanitskaya E., Arslan A., Arslan E. & Bužan E.V. 2012: Evolutionary history of mole rats (genus *Nannospalax*) inferred from mitochondrial cytochrome *b* sequence. *Biol. J. Linn. Soc. 105: 446–455.*

Larkin M.A., Blakshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin. F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J. & Higgins D.G. 2007: Clustal W and clustal X version 2.0. *Bioinformatics 23: 2947–2948.*

Lyapunova E.A., Vorontsov N.N. & Martynova L. 1971: Cytological differentiation of burrowing mammals in the Palaeartic. *In Syposium Theriologicum II. Prague. 203–205.*

Marshall D.C. 2010: Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. *Syst. Biol. 59: 108–117.*

Matur F. & Sözen M. 2005: A karyological study on subterrranean mole rats of the *Spalax leucodon* Nordmann, 1840 (Mammalia: Rodentia) superspecies around Bilecik province in Turkey. *Zool. Middle East 36: 5–10.*

Matur F., Çolak F., Sevindik M. & Sözen M. 2010: Chromosomal evolution of blind mole rats. *12th International Conference on Rodent Biology. Rodens et Spatium. Zonguldak, Turkey, 19-23 July 2010, Abstract Book: 46.*

Matur F., Çolak F., Sevindik M. & Sözen M. 2011: Chromosome differentiation of four 2n = 50 chromosomal forms of Turkish mole rat, *Nannospalax nehringi*. *Zool. Sci. 28: 61–67.*

Németh A., Révay T., Hegyeli Z., Farkas J., Czabán D., Rózsás A. & Csorba G. 2009: Chromosomal forms and risk assessment of *Nannospalax* (superspecies *leucodon*) (Mammalia: Rodentia) in the Carpathian Basin. *Folia Zool. 58: 349–361.*

Nevo E., Filippucci M.G., Redi C., Korol A. & Beiles A. 1994: Chromosomal speciation and adaptive radiation of mole-rats in Asia Minor correlated with increased ecological stress. *Proc. Natl. Acad. Sci. USA 91: 8160–8164.*

Nevo E., Filippucci M.G., Redi C., Simson S., Heth G. & Beiles A. 1995: Karyotype and genetic evolution in speciation of subterranean mole rats of the genus *Spalax* in Turkey. *Biol. J. Linn. Soc. 54: 203–229.*

Nevo E., Beiles A. & Spradling T. 1999: Molecular evolution of cytochrome *b* of subterranean mole rats, *Spalax ehrenbergi* superspecies, in Israel. *J. Mol. Evol. 49: 215–226.*

Nevo E., Ivanitskaya E. & Beiles A. 2001: Adaptive radiation of blind subterranean mole rats: naming and revisiting the four sibling species of the *Spalax ehrenbergi* superspecies in Israel: *Spalax galili* (2n = 52), *S. golani* (2n = 54), *S. carmeli* (2n = 58) and *S. judaei* (2n = 60). *Bachkhuys Publishers, Leiden, The Netherlands.*

Posada D. & Crandall K.A. 1998: Modeltest: testing the model of DNA substitution. *Bioinformatics 14: 817–818.*

Reyes A., Nevo E. & Saccone C. 2003: DNA sequence variation in the mitochondrial control region of subterranean mole rats, *Spalax ehrenbergi* superspecies, in Israel. *Mol. Biol. Evol. 20: 622–632.*

Ronquist F. & Huelsenbeck J.P. 2003: MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics 191: 572–1574.*

Savić I. & Nevo E. 1990: The Spalacidae: evolutionary history, speciation, and population biology. In: Nevo E. & Reig O.A. (eds.), Evolution of subterranean mammals at the organismal and molecular levels. *Alan R. Liss, Inc., New York: 129–153.*

Smith M.F. & Patton J.L. 1993: The diversification of South American murid rodents: evidence from mitochondrial DNA sequence data for the akodontine tribe. *Biol. J. Linn. Soc. 50: 149–177.*

Sözen M. 2004: A karyological study on subterranean mole rats of the *Spalax leucodon* Nordmann, 1840 superspecies in Turkey. *Mamm. Biol. 69: 420–429.*

Sözen M., Çolak E., Yiğit N., Özkurt Ş. & Verimli R. 1999: Contributions to the karyology and taxonomy of the genus *Spalax* Güldenstaedt, 1770 (Mammalia: Rodentia) in Turkey. *Z. Säugetierkd. 64: 210–219*.

Sözen M., Yiğit N. & Çolak E. 2000: A study on karyotypic evolution of the genus *Spalax* Güldenstaedt, 1770 (Mammalia: Rodentia) in Turkey. *Isr. J. Zool. 46: 239–242*.

Sözen M., Matur F., Çolak E., Özkurt Ş. & Karataş A. 2006a: Some karyological records and a new chromosomal form for *Spalax* (Mammalia: Rodentia) in Turkey. *Folia Zool. 55: 247–256.*

Sözen M., Sevindik M. & Matur F. 2006b: Karyological and some morphological characteristics of *Spalax leucodon* Nordmann, 1840 (Mammalia: Rodentia) superspecies around Kastamonu province, Turkey. *Turk. J. Zool. 30: 205–219.*

Sözen M., Çataklı K., Eroğlu F., Matur F. & Sevindik M. 2011: Distribution of chromosomal forms of *Nannospalax nehringi* (Satunin, 1898) (Rodentia: Spalacidae) in Çankırı and Çorum provinces, Turkey. *Turk. J. Zool. 35: 367–374.*

Stamatakis A. 2006: RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics 22: 2688–2690.*

Suzuki H., Wakana S., Yonekawa H., Moriwaki K., Sakurai S. & Nevo E. 1996: Variations in ribosomal DNA and mitochondrial DNA among chromosomal species of subterranean mole rats. *Mol. Biol. Evol. 13: 85–92.*

Swofford D.L. 1999: PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). *Sinauer Associates, Sunderland, Massachusetts.*

Tamura K., Dudley J., Nei M. & Kumar S. 2007: MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol. 24: 1596–1599.*

Topachevskii W. 1969: Fauna USSR: mammals, Spalacidae. Vol. 3, No. 3. *Nauka, Leningrad. (in Russian)*

Wilson D.E. & Reeder D.M. 1993: Mammal species of the world: a taxonomic and geographic reference. 2nd edition. *Smithsonian Institution Press, Washington and London.*

Yiğit N., Çolak E., Sözen M. & Karataş A. 2006: Rodents of Turkey. *Meteksan, Ankara.*

Paper 2.1.6

Vallo P., Benda P., **Martínková N.**, Kaňuch P., Kalko E. K. V., Červený J., Koubek P. 2011. Morphologically uniform bats *Hipposideros aff. ruber* (Hipposideridae) exhibit high mitochondrial genetic diversity in southeastern Senegal. *Acta Chiropterologica* 13: 79-88.

Acta Chiropterologica, 13(1): 79–88, 2011
PL ISSN 1508-1109 © Museum and Institute of Zoology PAS
doi: 10.3161/150811011X578633

Morphologically uniform bats *Hipposideros* aff. *ruber* (Hipposideridae) exhibit high mitochondrial genetic diversity in southeastern Senegal

PETER VALLO[1, 9], PETR BENDA[2, 3], NATÁLIA MARTÍNKOVÁ[1, 4], PETER KAŇUCH[1, 5], ELISABETH K. V. KALKO[6, 7], JAROSLAV ČERVENÝ[1, 8], and PETR KOUBEK[1, 8]

[1]*Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, v.v.i., Květná 8, 603 65 Brno, Czech Republic*
[2]*Department of Zoology, National Museum (Natural History), Václavské náměstí 68, 115 79 Praha 1, Czech Republic*
[3]*Department of Zoology, Charles University, Viničná 7, 128 44 Praha 2, Czech Republic*
[4]*Institute of Biostatistics and Analyses, Masaryk University, Kamenice 3, 625 00 Brno, Czech Republic*
[5]*Institute of Forest Ecology, Slovak Academy of Sciences, Štúrova 2, 960 53 Zvolen, Slovak Republic*
[6]*Institute of Experimental Ecology, University of Ulm, Albert-Einstein Allee 11, 89069 Ulm, Germany*
[7]*Smithsonian Tropical Research Institute, Balboa, Republic of Panama*
[8]*Department of Forest Protection and Game Management, Faculty of Forestry and Wood Sciences, Czech University of Life Sciences, Kamýcká 129, 165 21 Praha 6, Czech Republic*
[9]*Corresponding author: E-mail: vallo@ivb.cz*

Two mitochondrial lineages of bats that are morphologically attributed to *Hipposideros ruber* have been shown to occur sympatrically in southeastern Senegal. We studied genetic diversity in these bats in the Niokolo Koba National Park using sequences of mitochondrial cytochrome *b* gene to determine the taxonomic status of the two genetic forms, and included skull morphology for comparison. Detailed multidimensional analysis of skull measurements indicated slight morphological differences between the two genetic forms. Exploration of peak frequency of the constant-frequency echolocation signals in a local population of *Hipposideros* aff. *ruber* was not available for both groups. Phylogenetic comparison with other available West African representatives of *H.* aff. *ruber* revealed paraphyletic relationship of the two Senegalese forms, with the less abundant form from Senegal forming a monophyletic group with that from Benin. Based on genetic divergence and sympatric occurrence, the two forms from Senegal might represent cryptic species. However, absence of nuclear gene flow between them is yet to be investigated to demonstrate their reproductive isolation.

Key words: cytochrome *b*, *Hipposideros caffer* complex, cryptic species, phylogeny

INTRODUCTION

Genetic differences in living organisms may precede phenotypic differences, making genetically distinct forms difficult or even impossible to detect by traditional morphological means (Yoder *et al.*, 2000; Jacobs *et al.*, 2006). Such forms are believed to belong to the same species until additional evidence shows that they represent independent evolutionary units. The existence of these so-called cryptic species (Mayr, 1996; Lincoln *et al.*, 1998) has been revealed for many taxonomic groups (Avise, 2004; Bickford *et al.*, 2007). Although careful examination of morphology, ecology or behaviour helps to discover cryptic species, molecular genetics has contributed enormously in the last two decades to discovering cryptic forms within traditionally recognised taxa (Avise, 2004; Bickford *et al.*, 2007).

The presence of cryptic species is a rather common phenomenon in bats, and molecular data play an important role in their recognition and formal systematic acknowledgement (Jones, 1997; Mayer and von Helversen, 2001; Baker and Bradley, 2006; Ibáñez *et al.*, 2006). Differences in call characteristics of echolocation signals have been likewise useful to recognize distinct forms deserving taxonomic recognition. The European vespertilionid *Pipistrellus pipistrellus*/*P. pygmaeus* complex is a classic example where differences in peak frequencies of search calls led to discovery of two distinct sonotypes, which have been subsequently confirmed by molecular methods to represent two species (Barratt *et al.*, 1997).

The genus *Hipposideros* Gray, 1831 represents the most speciose group within the Palaeotropical family of Old World leaf-nosed bats (Hipposideridae). It contains a high number of phonic types, some of which have been subsequently confirmed as new cryptic species by molecular analyses, e.g., *H. khakhouayensis* (Guillén-Servent and Francis, 2006) and *H. khasiana* (Thabah *et al.*, 2006) from Southeast Asia. Distinct phonic types exist also within African species of *Hipposideros*, although none or only limited molecular justification appeared to date to confirm their distinct taxonomic statuses. The first is the case of phonic types of African *H. commersoni* (E. Geoffroy, 1813), which were revealed by Pye (1972) and which have been recently raised to species rank as *H. gigas* (Wagner, 1845) and *H. vittatus* (Peters, 1852) (Simmons, 2005). The second example is the complex of forms pertaining to *H. caffer* (Sundevall, 1846). Currently recognised species of this complex, *H. caffer* and *H. ruber* (Noack, 1893), are generally assumed to differ in peak frequencies of their echolocation calls (Pye, 1972; Fenton, 1986; Heller, 1992; Jones *et al.*, 1993). However, the first study of the *H. caffer* complex using molecular genetic tools (Vallo *et al.*, 2008) showed that this group is composed of more than two evolutionary units that might represent separate species. As the metric characters traditionally used to distinguish between *H. caffer* and *H. ruber* overlap, importance of echolocation frequencies for taxonomy should be re-evaluated in relation to molecular phylogeny. Differences in echolocation parameters can be good indicators of cryptic species. Particularly with respect to social communication between conspecific individuals, differences in echolocation appear to be more valid for closely related *Hipposideros* bats than the concept of adaptation to feeding on hearing insects (Barratt *et al.*, 1997; Jones *et al*., 1997; Bogdanowicz *et al.*, 1999; Guillén-Servent *et al*., 2000; Sedlock and Weyandt, 2009).

Our field work in the Niokolo Koba National Park (NKNP), Senegal, yielded a large number of leaf-nosed bats tentatively identified as *H. ruber*. These bats represent a distinct phylogenetic lineage restricted to West Africa, ranging from Senegal to Benin ('lineage D' of Vallo *et al.*, 2008), and are probably not closely related to *H. ruber* s. str., which was described from Tanzania. Although samples of *Hipposderos* aff. *ruber* bats from NKNP showed no obvious morphological differences, in the subsample used by Vallo *et al.* (2008), one haplotype significantly differed from the main mitochondrial lineage of the Senegalese population. This interesting intrapopulational pattern inspired us to investigate phylogenetic relationships in *H.* aff. *ruber* from Senegal using sequences of the mitochondrial gene for cytochrome *b* (cyt*b*). We studied genetic diversity of populations in the NKNP to evaluate the relevance of two sympatric mitochondrial lineages for the systematics of the *H. caffer* complex. We further analysed sequences of male-specific nuclear zinc finger region on the Y chromosome (*zfy* — Page *et al.*, 1987) in the two mitochondrial lineages to obtain independent evidence for reproductive separation. Additionally, we used skull morphometrics in relation to patterns revealed on the genetic level. As the divergent haplotype within Senegalese *H.* aff. *ruber* came from the village of Dar Salam at the northwestern border of the NKNP, we also explored echolocation calls in the local population to determine whether calls differ in correspondence to genetic lineages. We hypothesised that possible differences in echolocation frequencies and morphology would support presence of cryptic species suggested by genetic divergence.

Materials and Methods

Sampling

Bats were netted at six localities in the NKNP, SE Senegal, between 2004 and 2007 (Fig. 1 and Appendix). Captured specimens were weighted and external measurements were recorded. A subset of 52 specimens was collected and preserved in ethanol for further study. Tissue samples (spleen) were taken from collected specimens. Skulls were extracted from ethanol-preserved vouchers for morphological analysis. All biological material including voucher specimens was deposited at the Institute of Vertebrate Biology of the Academy of Sciences, Brno, Czech Republic. For comparison, we included additional samples and GenBank sequences of West African *H.* aff. *ruber* from Benin, Ghana, and Ivory Coast (Lim *et al.*, 2007; Vallo *et al.*, 2008), and another hipposiderid bat, *Asellia tridens*, as an outgroup (Benda and Vallo, 2009 — see Appendix).

Morphological Analysis

One external and 15 skull dimensions of collected specimens were measured following Benda and Vallo (2009) and using mechanical callipers with a precision of 0.02 mm: LAt = forearm length, LCr = greatest length of skull including premaxillae, LOc = occipito-canine length, LCc = condylo-canine length, LaZ = zygomatic width, LaI = width of interorbital constriction, LaInf = rostral width between foramina infraorbitalia, LaN = neurocranium width, LaM = mastoid width, ANc = height of neurocranium, LBT = largest horizontal length of tympanic bulla, CC = width across upper canines at crowns, M^3M^3 = width across third upper molars, CM^3 = length of upper tooth-row from front of canine to back of third molar, LMd = condylar length of mandible, ACo = height of coronoid

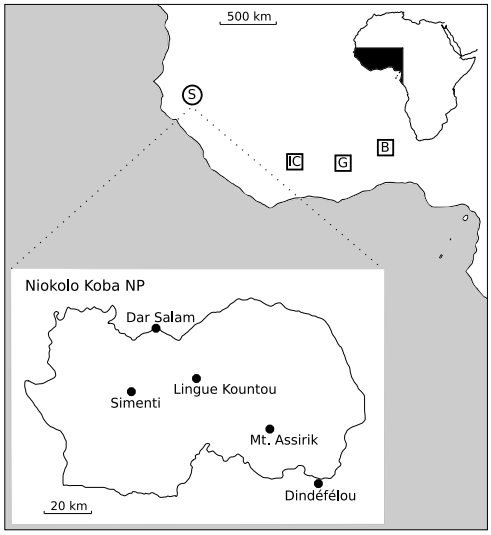


FIG. 1. Distribution of localities sampled in Niokolo Koba NP, southeastern Senegal, and of all West African populations sampled; Senegal (circle; S), and Ivory Coast, Ghana, and Benin (squares; IC, G, B, respectively)

process, CM_3 = length of lower tooth-row from front of canine to back of third molar. Analysis of variance and canonical discriminant analysis using raw data values were employed to reveal morphological differences in our dataset. Statistical analyses were performed using Statistica 6.0 software (StatSoft, Tulsa, OK, USA).

Analysis of Echolocation Calls

Echolocation calls of individuals captured at Dar Salam were recorded in time expansion (10×) and heterodyne mode with a bat-detector Pettersson D240x (Pettersson Elektronik AB, Uppsala, Sweden; sampling rate 307 kHz) with built-in microphone linked to a Sony MiniDisc MZ. Sound recordings were saved in uncompressed digital format (.wav). Animals were flown individually in a volary made of mosquito net (2 × 2 × 2 m). Echolocation calls were recorded when bats were flying towards the microphone only, thus compensation for the Doppler effect should have the same rate in the small constant-sized volary. Peak frequency of the constant frequency (CF) component of calls with maximum energy derived from power spectra was extracted from time-expanded (10×) sequences of approx. 1 s (FFT size 512 samples, Hanning window) with BatSound 3.31 software (Pettersson Elektronik AB). We measured peak frequencies of four consecutive, randomly selected calls from the sonograms of each individual and took their average as the final value.

DNA Processing

Total genomic DNA was extracted from ethanol-preserved tissue with DNeasy Tissue Kit (Qiagen, Halden, Germany) according to the manufacturer's protocol. Complete cyt*b* gene was amplified using universal primers L14724 and H15915 (Irwin *et al.*, 1991). Each PCR reaction contained 0.8 µM of each primer, 0.2 mM dNTP, 1U of HotMaster Taq DNA polymerase with 5 µl of corresponding 10× buffer (Eppendorf, Hamburg, Germany), and 2–5 µl of extracted DNA in 50 µl volume. Initial denaturation at 94°C for 3 min was followed by 35 cycles of denaturation for 40 s at 94°C, annealing for 40 s at 50°C, and extension for 90 s at 65°C, with final extension at 65°C for 5 min. Partial sequences of *zfy* gene were amplified using primers 33X5YF and LGL 331 (Trujillo *et al.*, 2009) and the same PCR protocol. The resulting PCR products were purified with QIAquick PCR Purification Kit (Qiagen) and sequenced commercially (Macrogen, Seoul, Korea) with the same primers using Big-Dye Terminator sequencing chemistry (Applied Biosystems, Foster City, CA, USA) on ABI 3730xl sequencer. Sequences were assembled and edited in Sequencher 4.6 (Gene Codes, Ann Arbor, MI, USA). Sequences were submitted to GenBank with accession numbers HQ343240–HQ343266 (Appendix).

Phylogenetic Analysis

Sequences of *Hipposideros* aff. *ruber* from Senegal were aligned in BioEdit 7.0 (Hall, 1999). Polymorphism within the Senegalese sequence dataset was assessed using DnaSP 4.0 (Rozas *et al.*, 2003) and a median-joining network of cyt*b* sequences was constructed in Network 4.2 (Fluxus Technology, Clare, UK). Based on this initial analysis, the cyt*b* dataset was reduced to haplotypes representing main haplogroups to facilitate reconstruction of phylogenetic trees and additional sequences including the outgroup were added for subsequent analyses (Appendix).

Phylogenetic trees were reconstructed in PAUP* 4.10b (Sinauer Associates, Sunderland, MA, USA) using maximum parsimony (MP) and maximum likelihood (ML) methods. In both methods, tree space was heuristically searched with tree bisection-reconnection swapping algorithm on 100 random sequence additions. Hasegawa-Kishino-Yano evolutionary model with gamma-distributed among-site rate variation (HKY85 + Γ — Hasegawa *et al.*, 1985; Yang, 1996) was used in ML analysis. This five-parameter model was suggested by the program Modeltest 3.7 (Posada and Crandall, 1998) as the 3rd best model under the Akaike Information Criterion (AIC), and was chosen in preference to two more complex models with eight parameters in order to reduce variance in parameter estimates. Reliability of branching pattern was assessed by bootstrapping using 1000 replicates in both analyses. Phylogeny was further estimated using Bayesian inference in MrBayes 3.1.2 (Ronquist and Huelsenbeck, 2003) with the same model. We used two independent simultaneous Metropolis-coupled MCMC runs of four chains running for 10^6 generations, sampled every 100th generation, starting from random trees. The first 2,500 sampled trees were discarded as burn-in. A 50% majority rule consensus tree was constructed from remaining trees with posterior probabilities representing confidence estimates of topology. Templeton test (Templeton, 1983) and Shimodaira-Hasegawa test (SH test — Shimodaira and Hasegawa, 1999) with RELL re-sampling algorithm and 1,000 bootstrap replicates were used to compare tree topologies. Sequence divergences were based on pairwise Kimura two-parameter genetic distances (K2P — Kimura, 1980).

RESULTS

In the Senegalese dataset, 29 haplotypes of cyt*b* (1,140 bp) were identified among 52 analysed individuals, resulting in high haplotype diversity (H = 0.966, SD = 0.01). As most haplotypes were closely related, nucleotide diversity was low (π = 0.009, SD = 0.002). Median-joining network showed diversification into two groups within Senegalese *H.* aff. *ruber*, denoted here with respect to the original lineage D by Vallo *et al.* (2008) as D1 and D2 (Fig. 2). Group D1 comprised 27 haplotypes representing 47 specimens. Group D2 was formed by two haplotypes from five specimens. Groups D1 and D2 differed by 2.4–3.6% K2P-distance. Genetic structure was not related to distribution of sampling sites, and bats of groups D1 and D2 occurred syntopically at Dar Salam, Simenti and Lingue Kountou (Appendix). Sequences of *zfy* (1,316 bp) from four specimens of group D1 and from two specimens of group D2 were identical.

We recorded echolocation calls of 16 specimens from Dar Salam. Variation in peak frequencies within the four measured consecutive calls of individual bats was very low (around 0.1 kHz). Given the low intra-individual variation in peak frequency and similar recording conditions in the constant-sized volary leading to similar recording bias (i.e., equal compensation for the Doppler effect), we compared average peak frequencies among individuals. Peak frequencies ranged from 130 kHz to 139 kHz (Appendix). Genetic results showed that all of the recorded specimens belonged to group D1, hence precluding a comparison with group D2 as originally expected. Frequencies were related to sex, as males called at significantly higher frequencies (134–139 kHz, n = 8) than females (130–135 kHz, n = 8 — Mann-Whitney test, U = 4, P < 0.01).

Univariate analysis of forearm length and skull morphometrics did not reveal distinct morphotypes. Measurements of the two haplogroups D1 and D2 mostly overlapped (Table 1); they differed only in CM^3 (ANOVA, F = 5.81, *d.f.* = 48, P < 0.05) and slightly in ANc (ANOVA, F = 3.78, *d.f.* = 48, P = 0.058). Males and females of groups D1 and D2 did not differ in size. Canonical discriminant analysis of skull dimension revealed LCo, LCc, LaZ, LaI, CC, ANc, CM^3 and CM_3 as the most important variables distinguishing between groups D1 and D2 along the 1st canonical axis (CV1), which explained 71.0% of variance (Fig. 3).

Phylogenetic relationships were reconstructed among 11 cyt*b* sequences of *H.* aff. *ruber*: six from Senegal (four from haplogroup D1 and two from haplogroup D2), two from Ivory Coast, two from Ghana, and one from Benin. Heuristic search under MP criterion yielded a single best tree 269 steps long (Fig. 4) showing five well-supported (bootstrap support ≥ 70%, Bayesian posterior probability ≥ 0.95) phylogenetic lineages, which corresponded to the respective geographic origin of samples, and included the two distinct haplogroups from Senegal: Senegal D1, Senegal D2, Benin, Ghana, and Ivory Coast. Genetic divergences among the five lineages ranged 2.0–5.4% (Table 2). Divergences within the lineages except those from Senegal were small (up to 0.1%). Monophyletic relationship was supported between lineages Senegal D2 and Benin, and between lineages Ivory Coast and Ghana, respectively. Senegal D1 was placed as sister lineage to a clade

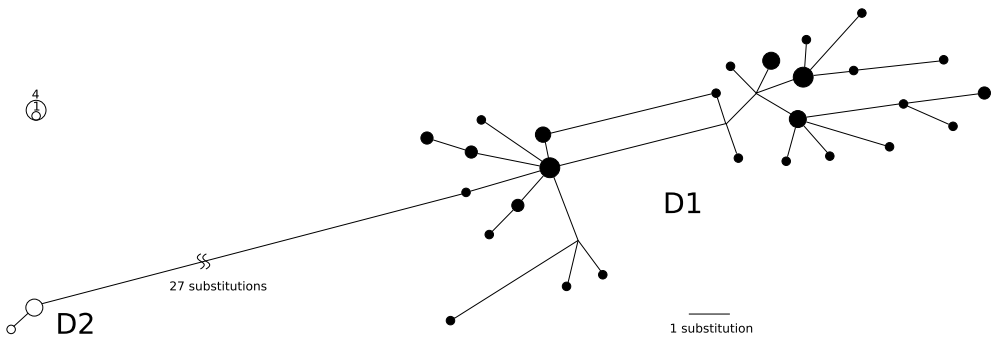


FIG. 2. Median-joining network of 29 haplotypes found in the population of NKNP. Size of nodes is proportional to frequency of particular haplotypes. Black and white circles denote haplogroups D1 and D2, respectively, and this scheme is consistent with Figs. 3 and 4

TABLE 1. Forearm length and skull measurements (in mm) of the examined specimens. See Materials and Methods for abbreviations of measurements

Measurements	Haplogroup D1					Haplogroup D2				
	n	$\bar{x}$	Min	Max	SD	*n*	$\bar{x}$	Min	Max	SD
LAt	47	47.88	44.20	49.80	1.100	4	47.15	46.30	48.30	0.780
LCr	45	18.92	18.27	19.62	0.298	5	18.77	18.28	19.11	0.338
LOc	45	18.68	17.78	19.23	0.294	5	18.56	18.14	18.86	0.264
LCc	45	16.34	15.83	16.83	0.238	5	16.23	16.03	16.37	0.149
LaZ	45	10.51	10.02	10.86	0.222	5	10.57	10.28	10.75	0.195
LaI	45	2.88	2.59	3.09	0.120	5	2.82	2.60	2.98	0.143
LaInf	45	5.07	4.81	5.42	0.114	5	5.06	4.98	5.18	0.088
LaN	45	8.31	7.75	8.68	0.176	5	8.27	7.85	8.73	0.370
LaM	45	9.89	9.52	10.28	0.154	5	9.80	9.69	9.97	0.121
ANc	45	5.85	5.29	6.38	0.223	5	6.07	5.80	6.67	0.353
CC	45	4.98	4.73	5.24	0.134	5	4.93	4.75	5.06	0.138
M^3M^3	45	7.15	6.68	7.42	0.159	5	7.11	6.82	7.25	0.168
CM^3	45	7.08	6.78	7.29	0.123	5	6.95	6.84	7.05	0.076
LMd	45	12.38	11.82	12.74	0.197	5	12.41	12.31	12.51	0.091
ACo	45	3.04	2.74	3.31	0.142	5	3.03	2.93	3.23	0.117
CM_3	45	7.65	7.37	7.93	0.138	5	7.62	7.55	7.68	0.048
LBT	45	3.36	3.07	3.55	0.092	5	3.34	3.28	3.48	0.080

Senegal D2 + Benin, but the statistical support for this group was low. ML tree (-ln*L* = 2741.72205) and Bayesian consensus tree showed the same supported groups as the MP tree: Senegal D2 + Benin and Ivory Coast + Ghana, and their supported sister relationship. However, they also revealed a rather unclear pattern of paraphyletic haplotypes belonging to the group Senegal D1 placed basally in the tree. Thus, both monophyly of Senegal D1 and its phylogenetic position considering the monophyly were tested using Templeton and SH tests. Constrained ML tree with forced monophyly of Senegal D1 (-ln*L* = 2743.04876) did not differ significantly from the original ML tree (Templeton test: diff. length = 4, z = -1.1547, P = 0.2482; SH test: diff. -ln*L* = 1.32671, P = 0.168) and monophyly of Senegal D1 thus could not be rejected. Position of other lineages in the constrained tree was otherwise

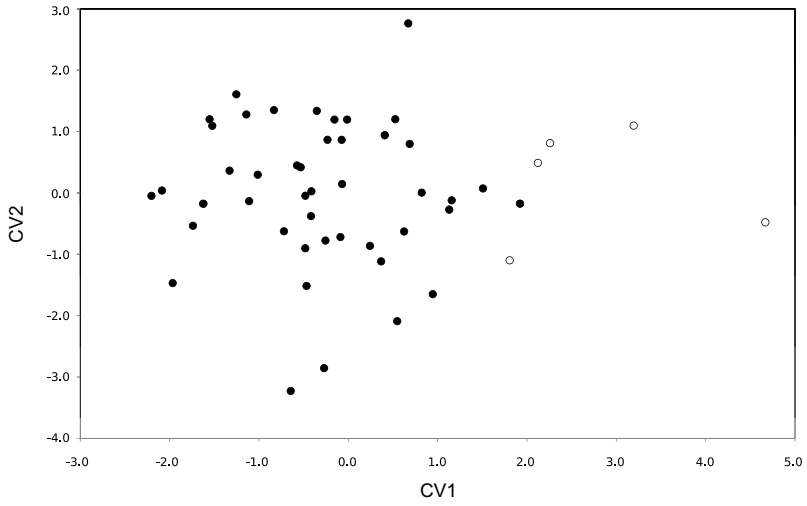


FIG. 3. Plot of main canonical variables CV1 and CV2 from canonical discriminant analysis of 15 skull measurements. Symbols as in Fig. 2

identical to the original ML tree. Although ML analysis favoured sister relationship between Senegal D2 + Benin and Ivory Coast + Ghana (73% bootstrap support) in contrast to MP analysis that pointed towards sister relationship of Senegal D1 to Senegal D2 + Benin (67% bootstrap support), these two phylogenetic hypotheses did not differ from each other significantly (Templeton test: diff. length = 3, z = -0.9045, P = 0.37; SH test: diff. -ln$L$ = 0.71613, P = 0.29). Consequently, mutual positions of the lineages Senegal D1, Senegal D2 + Benin and Ivory Coast + Ghana remained unresolved.

TABLE 2. Pairwise genetic divergences among phylogenetic lineages of West African *H.* aff. *ruber*. K2P — Kimura two-parameter distance (%)

K2P	Senegal D2	Senegal D1	Benin	Ivory Coast
Senegal D1	2.6–3.6			
Benin	2.0–2.1	3.2–4.1		
Ivory Coast	5.1–5.2	4.0–4.3	5.3–5.4	
Ghana	4.4–4.7	4.5–5.1	5.2–5.3	2.8–3.0

DISCUSSION

Analysis of mitochondrial DNA sequences of *H.* aff. *ruber* revealed two distinct genetic lineages with sequence divergence of 2.4–3.6% co-occurring in the Niokolo Koba NP, southeastern Senegal. Only five specimens from our collection belonged to lineage D2. Haplotypes of both lineages were present at three localities in the NKNP. Absence of the lineage D2 at two other localities may have been due to low sample size, although habitat preferences or competition between the two lineages may also be a relevant explanation. As previous results indicated existence of both lineages at Dar Salam (Vallo *et al.*, 2008), we compared echolocation calls in the local population. The differences in call frequencies might indicate presence of cryptic species (Jones, 1997; Mayer and von Helversen, 2001). Our results showed peak echolocation frequencies of CF calls to range 130–139 kHz, which is fairly similar to the combined range of cryptic forms of *H. bicolor* (128.0–144.5 kHz — Kingston *et al.*, 2001) and *H. ridley* (65–72 kHz — Francis *et al.*, 1999). However, genetic analysis revealed that all 16 individuals, in which we successfully recorded and analysed echolocation frequencies, belonged to the more abundant lineage D1. Therefore, the original hypothesis of congruence between echolocation and genetic diversity in Senegalese *H.* aff. *ruber* could

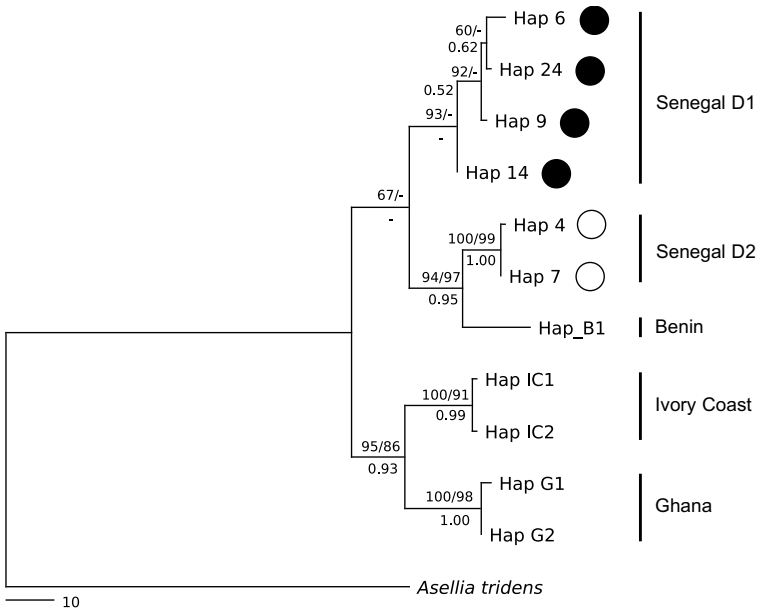


FIG. 4. Maximum parsimony tree depicting phylogenetic relationships in West African *H.* aff. *ruber*. Bootstrap support for MP and ML ≥ 70% (above branches) and posterior probabilities of BA ≥ 0.95 (below braches) are considered significant. Symbols as in Fig. 2

not be tested. Distribution of echolocation frequencies within lineage D1 was related to sex, with males calling at significantly higher frequencies than females. Similar relationship of echolocation frequency and sex has been shown in *H. ruber* from Equatorial Guinea (Guillén-Servent *et al.*, 2000). The frequency range of 130–139 kHz found in the bat population from Dar Salam should be regarded as intraspecific variation.

Morphological analysis showed only slight differences in skull dimensions. Overall, bats of the two lineages remain indistinguishable by morphological traits that are commonly used for species identification. Although the canonical discriminant analysis showed rather sufficient differences for separation of the lineages in some (mainly rostral) dimensions, the ranges of measured values overlap to such an extent that competent determination of taxa is not possible.

Given the values of sequence divergence and sympatric occurrence, the two lineages from Senegal could be considered separate species based on known limits of divergence between cryptic species (Baker and Bradley, 2006). Similar values of interspecific divergence (2.4–3.9%) have been found between cryptic forms of *Scotophilus dinganii* (A. Smith, 1833) from South Africa (Jacobs *et al.*, 2006). Furthermore, similar divergence (3.9–4.1%) was documented between two *Hipposideros* species from Southeast Asia, *H. khakhouayensis* and *H. rotalis* (Guillén-Servent and Francis, 2006). Sequence divergence values alone, however, have to be interpreted with caution when deciding on an appropriate taxonomic rank of genetically separate populations. Lausen *et al.* (2008) clearly showed that divergence values in mitochondrial sequences led to false classification of lineages within North American *Myotis lucifugus* as distinct species, as they discovered substantial nuclear gene flow among these lineages although mitochondrial sequence divergence ranged up to 5%. Similar mitochondrial differentiation not reflected in nuclear markers was found in *Myotis capaccinii* (Bilgin *et al.*, 2008). It is therefore important to document reproductive isolation of the two genetically distinct forms in Senegal to confirm their status as two species. Our use of a paternally inherited *zfy* gene as an independent phylogenetic marker to maternally inherited cyt*b* gene (Lim *et al.*, 2008; Trujillo *et al.*, 2009) did not provide a conclusive answer because all obtained *zfy* sequences were identical. On one hand, this uniformity may indicate extensive, male-mediated gene flow between haplogroups D1 and D2. On the other hand, it could also result from low mutation rate of this nuclear gene and a relatively recent split of both lineages. More informative estimation of relationships between the two Senegalese mitochondrial forms could be achieved using analysis of microsatellites, as this would detect gene flow between both lineages and thus provide necessary data for a taxonomic conclusion.

The basal node of phylogenetic tree remained unresolved suggesting a rapid radiation of the three basal lineages Senegal D1, Senegal D2 + Benin, and Ghana + Ivory Coast. Moreover, a sister relationship of the two Senegalese lineages was not supported and the lineage D2 turned out to be closely related to the lineage from Benin. According to this phylogenetic structure and considering geographical distribution of the lineages, the sympatric occurrence of the two lineages from Senegal, D1 and D2, may be explained as a secondary contact of two formerly isolated populations. The genetic distance, reaching over 5% among the sampled West African populations, further suggests that even lineages from Ghana and Ivory Coast might be regarded as cryptic species. Unlike the Senegalese lineages D1 and D2, these populations do not occur in sympatry and the genetic differences could be explained by the isolation by distance. On the other hand, conflict between geographic distance between sampled localities in Ghana and Benin (ca. 300 km), where the corresponding divergence exceeds 5%, and Senegal and Benin (ca. 2,000 km), with divergence only 2%, indicate a rather more complicated relationship. A broader sampling throughout West Africa would thus be needed to better understand the indicated phylogeographic pattern and resolve the question of cryptic diversity in *H.* aff. *ruber*.

Acknowledgements

We thank Adam Konečný, Josef Bryja and other colleagues from the Institute of Vertebrate Biology AS CR, v.v.i., Brno, for their assistance in capturing of bats in Senegal. Field work in the NKNP and collecting of bat specimens was approved and supervised by the Direction des Parcs Nationaux du Sénégal, Dakar, and we thank the director Col. Mame Balla Gueye for his kind support. Further we thank Burton Lim (Royal Ontario Museum, Toronto, Canada) for providing tissue samples of bats from Ivory Coast, and Christian Drosten (University of Bonn Medical Centre, Germany) and his team for acquiring samples from Ghana within the project 01KI0701 by the German Federal Ministry of Education and Research. This study was supported by the grant No. IAA6093404 of the Grant Agency of the Academy of Sciences of the Czech Republic and grant No. MK00002327201 of the Ministry of Culture of the Czech Republic. We also thank Jakob Fahr (University of Ulm, Germany) for critical comments and suggestions on the manuscript.

Literature Cited

Avise, J. C. 2004. Molecular markers, natural history, and evolution, 2nd edition. Sinauer Associates, Sunderland, Massachusetts, 684 pp.

Baker, R. J., and R. D. Bradley. 2006. Speciation in mammals and the genetic species concept. Journal of Mammalogy, 87: 643–662.

Barratt, E. M., R. Deaville, T. M. Burland, M. W. Bruford, G. Jones, P. A. Racey, and R. K. Wayne. 1997. DNA answers the call of pipistrelle bat species. Nature, 387: 138–139.

Benda, P., and P. Vallo. 2009. Taxonomic revision of the genus *Triaenops* (Mammalia: Chiroptera: Hipposideridae) with description of a new species from southern Arabia and definitions of new genus and tribe. Folia Zoologica, 53 (Monograph 1): 1–45.

Bickford, D., D. J. Lohman, N. S. Sodhi, P. K. L. Ng, R. Meier, K. Winker, K. K. Ingram, and I. Das. 2007. Cryptic species as a window on diversity and conservation. Trends in Ecology and Evolution, 22: 148–155.

Bilgin, R., A. Karataş, E. Çoraman, and J. Morales. 2008. The mitochondrial and nuclear genetic structure of *Myotis capaccinii* (Chiroptera: Vespertilionidae) in the Eurasian transition, and its taxonomic implications. Zoologica Scripta, 37: 253–262.

Bogdanowicz, W., M. B. Fenton, and K. Daleszczyk. 1999. The relationships between echolocation calls, morphology and diet in insectivorous bats. Journal of Zoology (London), 247: 381–393.

Fenton, M. B. 1986. *Hipposideros caffer* (Chiroptera: Hipposideridae) in Zimbabwe: morphology and echolocation calls. Journal of Zoology (London), 210: 347–353.

Francis, C. M., D. Kock, and J. Habersetzer. 1999. Sibling species of *Hipposideros ridleyi* (Mammalia, Chiroptera, Hipposideridae). Senckenbergiana Biologica, 79: 255–270.

Guillén-Servent, A., and C. M. Francis. 2006. A new species of bat of the *Hipposideros bicolor* group (Chiroptera: Hipposideridae) from Central Laos, with evidence of convergent evolution with Sundaic taxa. Acta Chiropterologica, 8: 39–61.

Guillén-Servent, A., C. Ibáñez, and J. Juste. 2000. Variation in the echolocation calls of *Hipposideros ruber* in the Gulf of Guinea: An exploration of the adaptive meaning of the constant frequency value in rhinolophoid bats. Journal of Evolutionary Biology, 13: 70–80.

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series, 41: 95–98.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution, 22: 160–174.

Heller, K.-G. 1992. The echolocation calls of *Hipposideros ruber* and *Hipposideros caffer*. Pp. 75–77, *in* Prague studies in mammalogy (I. Horáček and V. Vohralík, eds.). Charles University Press, Prague, xxi + 245 pp.

Ibáñez, C., J. L. García-Mudarra, M. Ruedi, B. Stadelmann, and J. Juste. 2006. The Iberian contribution to cryptic diversity in European bats. Acta Chiropterologica, 8: 277–297.

Irwin, D. M., T. D. Kocher, and A. C. Wilson. 1991. Evolution of the cytochrome *b* gene of mammals. Journal of Molecular Evolution, 32: 128–144.

Jacobs, D. S., G. N. Eick, M. C. Schoeman, and C. A. Mathee. 2006. Cryptic species in an insectivorous bat, *Scotophilus dinganii*. Journal of Mammalogy, 87: 161–170.

Jones, G. 1997. Acoustic signals and speciation: the roles of natural and sexual selection in the evolution of cryptic species. Advanced Study in Behaviour, 26: 317–354.

Jones, G., M. Morton, P. M. Hughes, and R. M. Budden. 1993. Echolocation, flight morphology and foraging strategies of some West African hipposiderid bats. Journal of Zoology (London), 230: 385–400.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution, 16: 111–120.

Kingston, T., M. C. Lara, G. Jones, Z. Akbar, T. H. Kunz, and C. J. Schneider. 2001. Acoustic divergence in two cryptic *Hipposideros* species: a role for social selection? Proceedings of the Royal Society of London, 268: 1381–1386.

Lausen, C. L., I. Delisle, R. M. R. Barclay, and C. Strobeck. 2008. Beyond mtDNA: nuclear gene flow suggests taxonomic over-splitting in the little brown bat (*Myotis lucifugus*). Canadian Journal of Zoology, 86: 700–713.

Lim, B. K., M. D. Engstrom, J. W. Bickham, and J. C. Patton. 2007. Molecular phylogeny of New World sheath-tailed bats (Emballonuridae: Diclidurini) based on loci from the four genetic transmission systems in mammals. Biological Journal of the Linnean Society, 93: 189–209.

Lincoln, R., G. Boxshall, and P. Clark. 1998. A dictionary of ecology, evolution and systematics, 2nd edition. Cambridge University Press, Cambridge, United Kingdom, 371 pp.

Mayer, F., and O. von Helversen. 2001. Cryptic diversity in European bats. Proceedings of the Royal Society of London, 268B: 1825–1832.

Mayr, E. 1996. What is a species, and what is not? Philosophy of Science, 63: 262–277.

Page, D. C., R. Mosher, E. M. Simpson, E. M. C. Fisher, G. Mardon, J. Pollack, B. McGillivray, A. de la Chapelle, and L. G. Brown. 1987. The sex-determining region of the human Y chromosome encodes a finger protein. Cell, 51: 1091–1104.

Posada, D., and K. A. Crandall. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics, 14: 817–818.

Pye, J. D. 1972. Bimodal distribution of constant frequencies in some hipposiderid bats (Mammalia: Hipposideridae). Journal of Zoology (London), 166: 323–335.

Ronquist, J., and J. J. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19: 1572–1574.

Rozas, J., J. C. Sánchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics, 19: 2496–2497.

Sedlock, J. L., and S. E. Weyandt, 2009. Genetic divergence between morphologically and acoustically cryptic bats: novel niche partitioning or recent contact? Journal of Zoology (London), 279: 388–395.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular Biology and Evolution, 16: 1114–1116.

Simmons, N. 2005. Order Chiroptera. Pp. 312–529, *in* Mammal species of the World: a taxonomic and geographic reference (D. E. Wilson and D. M. Reeder, eds.). The Johns Hopkins University Press, Baltimore, 2142 pp.

Templeton, A. R. 1983. Phylogenetic inference from restriction

endonuclease cleavage site maps with particular reference to the humans and apes. Evolution, 37: 221–244.

Thabah, A., S. J. Rossiter, T. Kingston, S. Zhang, S. Parsons, K. Mya Mya, A. Zubaid, and G. Jones. 2006. Genetic divergence and echolocation call frequency in cryptic species of *Hipposideros larvatus* s.l. (Chiroptera: Hipposideridae) from the Indo-Malayan region. Biological Journal of the Linnean Society, 88: 119–130.

Trujillo, R. G., J. C. Patton, D. A. Schlitter, and J. W. Bickham. 2009. Molecular phylogenetics of the bat genus *Scotophilus* (Chiroptera: Vespertilionidae): perspectives from paternally and maternally inherited genomes. Journal of Mammalogy, 90: 548–560.

Vallo, P., A. Guillén-Servent, P. Benda, D. B. Pires, and P. Koubek. 2008. Variation of mitochondrial DNA in the *Hipposideros caffer* complex (Chiroptera: Hipposideridae) and its taxonomic implications. Acta Chiropterologica, 10: 193–206.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology and Evolution, 11: 367–372.

Yoder, A. D., R. M. Rasoloarison, S. M. Goodman, J. A. Irwin, S. Atsalis, M. J. Ravosa, and J. H. Ganzhorn. 2000. Remarkable species diversity in Malagasy mouse lemurs (Primates, *Microcebus*). Proceedings of the National Academy of Sciences of the USA, 97: 1325–1330.

Received 25 February 2010, accepted 10 February 2011

Appendix

List of specimens included in this paper. Echo — average peak frequency of CF echolocation call, Acc. number — accession number in the GenBank database

Sample	Haplotype	Haplogroup	Echo (kHz)	Country	Locality	Acc. number	Source of sequence
IVB S8	Hap_1	D1	–	Senegal	Mt. Assirik	HQ343240	This study
IVB S95	Hap_9	D1	–	Senegal	Simenti	–	This study
IVB S112	Hap_2	D1	–	Senegal	Lingué Kountou	HQ343241	This study
IVB S119	Hap_3	D1	–	Senegal	Lingué Kountou	EU934478	Vallo *et al*. (2008)
IVB S139	Hap_6	D1	–	Senegal	Lingué Kountou	HQ343243	This study
IVB S218	Hap_8	D1	–	Senegal	Simenti	HQ343244	This study
IVB S253	Hap_6	D1	–	Senegal	Simenti	HQ343245	This study
IVB S272	Hap_9	D1	–	Senegal	Simenti	EU934481	Vallo *et al*. (2008)
IVB S273	Hap_10	D1	–	Senegal	Simenti	EU934482	Vallo *et al*. (2008)
IVB S275	Hap_11	D1	–	Senegal	Simenti	EU934483	Vallo *et al*. (2008)
IVB S278	Hap_5	D1	–	Senegal	Simenti	HQ343246	This study
IVB S280	Hap_12	D1	–	Senegal	Simenti	HQ343247	This study
IVB S283	Hap_13	D1	–	Senegal	Simenti	HQ343249	This study
IVB S285	Hap_14	D1	–	Senegal	Simenti	EU934484	Vallo *et al*. (2008)
IVB S290	Hap_15	D1	–	Senegal	Simenti	HQ343250	This study
IVB S291	Hap_11	D1	–	Senegal	Simenti	–	This study
IVB S341	Hap_16	D1	–	Senegal	Simenti	HQ343251	This study
IVB S342	Hap_9	D1	–	Senegal	Simenti	–	This study
IVB S362	Hap_17	D1	–	Senegal	Simenti	HQ343252	This study
IVB S695	Hap_18	D1	–	Senegal	Dar Salam	HQ343253	This study
IVB S701	Hap_14	D1	–	Senegal	Dar Salam	–	This study
IVB S702	Hap_19	D1	–	Senegal	Dar Salam	HQ343254	This study
IVB S803	Hap_5	D1	–	Senegal	Lingué Kountou	–	This study
IVB S819	Hap_3	D1	–	Senegal	Dindéfélou	–	This study
IVB S820	Hap_20	D1	–	Senegal	Dindéfélou	EU934485	Vallo *et al*. (2008)
IVB S821	Hap_21	D1	–	Senegal	Dindéfélou	HQ343255	This study
IVB S825	Hap_20	D1	–	Senegal	Dindéfélou	–	This study
IVB S899	Hap_5	D1	–	Senegal	Dar Salam	–	This study
IVB S900	Hap_22	D1	–	Senegal	Dar Salam	HQ343256	This study
IVB S1374	Hap_5	D1	–	Senegal	Dar Salam	EU934479	Vallo *et al*. (2008)
IVB S1377	Hap_27	D1	–	Senegal	Dar Salam	HQ343260	This study
IVB S1538	Hap_28	D1	138	Senegal	Dar Salam	HQ343261	This study
IVB S1539	Hap_11	D1	137	Senegal	Dar Salam	–	This study
IVB S1540	Hap_9	D1	134	Senegal	Dar Salam	–	This study
IVB S1541	Hap_14	D1	139	Senegal	Dar Salam	–	This study
IVB S1551	Hap_23	D1	132	Senegal	Dar Salam	HQ343262	This study

APPENDIX. Continued

Sample	Haplotype	Haplogroup	Echo (kHz)	Country	Locality	Acc. number	Source of sequence
IVB S1554	Hap_23	D1	136	Senegal	Dar Salam	–	This study
IVB S1555	Hap_24	D1	137	Senegal	Dar Salam	HQ343257	This study
IVB S1561	Hap_29	D1	133	Senegal	Dar Salam	HQ343263	This study
IVB S1654	Hap_24	D1	133	Senegal	Dar Salam	–	This study
IVB S1655	Hap_24	D1	136	Senegal	Dar Salam	–	This study
IVB S1657	Hap_25	D1	134	Senegal	Dar Salam	HQ343258	This study
IVB S1660	Hap_11	D1	131	Senegal	Dar Salam	–	This study
IVB S1662	Hap_5	D1	131	Senegal	Dar Salam	–	This study
IVB S1663	Hap_14	D1	135	Senegal	Dar Salam	–	This study
IVB S1664	Hap_14	D1	135	Senegal	Dar Salam	–	This study
IVB S1665	Hap_26	D1	130	Senegal	Dar Salam	HQ343259	This study
IVB S132	Hap_4	D2	–	Senegal	Lingué Kountou	HQ343242	This study
IVB S281	Hap_7	D2	–	Senegal	Simenti	HQ343248	This study
IVB S286	Hap_7	D2	–	Senegal	Simenti	–	This study
IVB S403	Hap_7	D2	–	Senegal	Dar Salam	–	This study
IVB S1400	Hap_7	D2	–	Senegal	Dar Salam	EU934480	Vallo *et al*. (2008)
NMP 91879	Hap_B1	–	–	Benin	Tagayé	EU934476	Vallo *et al*. (2008)
ROM 100516	Hap_IC1	–	–	Ivory Coast	Sibabli	EF584226	Lim *et al*. (2008)
ROM 100518	Hap_IC2	–	–	Ivory Coast	Sibabli	HQ343264	This study
IVB PV59	Hap_G1	–	–	Ghana	Buoyem	HQ343265	This study
IVB PV56	Hap_G2	–	–	Ghana	Buoyem	HQ343266	This study
	Asellia tridens	–	–	Egypt	–	FJ457617	Benda and Vallo (2009)

Paper 2.1.7

Kopecna O., Kubickova S., Cernohorska H., Cabelova K., Váhala J., **Martínková N.**, Rubes J. 2014. Tribe-specific satellite DNA in non-domestic Bovidae. *Chromosome Research* 22: 277-291.

Chromosome Res
DOI 10.1007/s10577-014-9401-4

Tribe-specific satellite DNA in non-domestic Bovidae

Olga Kopecna · Svatava Kubickova · Halina Cernohorska · Katerina Cabelova · Jiri Vahala · Natalia Martinkova · Jiri Rubes

Received: 4 November 2013 / Revised: 2 January 2014 / Accepted: 10 January 2014

Abstract Satellite sequences present in the centromeric and pericentric regions of chromosomes represent useful source of information. Changes in satellite DNA composition may coincide with the speciation and serve as valuable markers of phylogenetic relationships. Here, we examined satellite DNA clones isolated by laser microdissection of centromeric regions of 38 bovid species and categorized them into three types. Sat I sequences from members of Bovini/Tragelaphini/Boselaphini are similar to the well-documented 1.715 sat I DNA family. Sat I DNA from Caprini/Alcelaphini/Hippotragini/Reduncini/Aepycerotini/Cephalophini/Antilopini/Neotragini/Oreotragini form the second group homologous to the common 1.714 sat I DNA. The analysis of sat II DNAs isolated in our study confirmed conservativeness of these sequences within Bovidae. Newly described centromeric clones from *Madoqua kirkii* and *Strepsiceros strepsiceros* were similar in length and repetitive tandem arrangement but showed no similarity to any other satellite DNA in the GenBank database. Phylogenetic analysis of sat I sequences isolated in our study from 38 bovid species enabled the description of relationships at the subfamily and tribal levels. The maximum likelihood and Bayesian inference analyses showed a basal position of sequences from Oreotragini in the subfamily Antilopinae. According to the Bayesian inference analysis based on the indels in a partitioned mixed model, Antilopinae satellite DNA split into two groups with those from Neotragini as a basal tribe, followed by a stepwise, successive branching of Cephalophini, Aepycerotini and Antilopini sequences. In the second group, Reduncini sequences were basal followed by Caprini, Alcelaphini and Hippotragini.

Keywords Bovidae · satellite DNA · centromeric repeats · phylogeny · FISH · laser microdissection

Abbreviations

BAC	Bacterial artificial chromosome
BI	Bayesian inference
BRU-PCR	Basic repeat unit obtained by PCR
DOP-PCR	Degenerate oligonucleotide primed polymerase chain reaction
FISH	Fluorescence in situ hybridization
LINE	Long interspersed nuclear element

Responsible editor: Walther Traut

Electronic supplementary material The online version of this article (doi:10.1007/s10577-014-9401-4) contains supplementary material, which is available to authorized users.

O. Kopecna (✉) · S. Kubickova · H. Cernohorska · K. Cabelova · J. Rubes
Department of Genetics and Reproduction, Veterinary Research Institute,
Hudcova 70, 621 00 Brno, Czech Republic
e-mail: kopecna@vri.cz

J. Vahala
Dvur Kralove Zoo,
544 01 Dvur Kralove n. L., Czech Republic

N. Martinkova
Institute of Biostatistics and Analyses, Masaryk University,
Brno, Czech Republic

Published online: 23 January 2014

LTR	Long terminal repeat
MCMC	Markov chains Monte Carlo
ML	Maximum likelihood
NCBI	National Center for Biotechnology Information
NFA	Numbers of autosomal arms
RFLP	Restriction fragment length polymorphism
sat	Satellite
SINE	Short interspersed nuclear element
SNP	Single nucleotide polymorphism

Introduction

The family Bovidae (order Artiodactyla, i.e. the even-toed ungulates) comprises approximately 140 species (Nowak 1999). Many of these are of economic importance (e.g. cattle, sheep and goats), while several species are endangered or threatened with extinction (http://www.iucnredlist.org). The chromosomal complement of domestic cattle ($2n=60$) is considered to reflect the ancestral condition forming the basis from which all the recent bovid karyotypes have been derived (Wurster and Benirschke 1968; Buckland and Evans 1978; Gallagher and Womack 1992). Although diploid chromosome numbers vary from 30 to 60 among species, the autosomal arm number remains relatively constant (NFA=56–58). This reflects the dominance of Robertsonian fusions (i.e. centric fusions of two acrocentric chromosomes) in the karyotype evolution of Bovidae (reviewed in Robinson and Ropiquet 2011).

A substantial proportion of the eukaryote genome consists of constitutive heterochromatin, a genomic fraction that includes LINEs, SINEs, satellite DNAs and other repetitive sequences. Satellite sequences, organized in tandem, usually reside in the centromeric and pericentric regions of chromosomes (D'Aiuto et al. 1997). DNA in the centromeric region appears quite complex (Lee et al. 1997), and no sequence conservation can be found among a wide range of species. Only closely related species share homologous satellite sequences (Jobse et al. 1995), but even among the most closely related species, satellite DNAs can differ in nucleotide sequence, copy number and the composition of satellite families (Ugarković and Plohl 2002). In the process of satellite DNA evolution, newer satellite sequences may be derived from preexisting satellite DNA sequences that replace or coexist with the old satellites via 'three-phase' evolutionary processes (Nijman and Lenstra 2001). Changes in satellite DNA composition may coincide with speciation and serve as valuable markers of phylogenetic relationships (Jobse et al. 1995; Saffery et al. 1999; Chaves et al. 2000, 2005), or can occur subsequent to speciation permitting the retrieval of species-specific profiles (Ugarković and Plohl 2002).

Bovine satellites constitute ~25 % of the total nuclear DNA content (Vaiman et al. 1999) but are quite heterogeneous. Macaya et al. (1978) isolated eight major satellite DNA fractions from the bovine genome DNA, some of which are related to each other. Moreover, certain sequence motifs have been found in different bovid tribes. For example, Chaves et al. (2000, 2005) hybridized satellite I sequences from sheep (1.714) and cattle (1.715) to the metaphase chromosomes of 15 species (representative of seven tribes). They document the existence of two major satellite clades in Bovidae—the Bovinae and a composite comprising the Caprinae/Alcelaphinae/Hippotraginae. Their findings of evolutionarily older variants of cattle satellite I being preserved in sheep satellite I provide support for an evolutionary mechanism as proposed by Nijman and Lenstra (2001). Kopecna et al. (2012) analysed fluorescence in situ hybridization (FISH) patterns of the bovine satellite I probe (1.715 DNA family) in different bovid tribes. Their results were in most cases in accordance with the results of other authors with the exception of three tribes. The authors concluded that the determination of phylogenetic relationships based only on FISH analysis could be misleading and that a more rigorous comparison that entailed the analysis of the main motif sequence of the repetitive DNA could be more informative. On the other hand, the evolutionarily young satellite IV sequence represents 4.3 % of the bovine genome and has no resemblance to other satellite DNAs (Skowronski et al. 1984). Two studies that relied on this fraction have been reported—Modi et al. (2004) and Adega et al. (2006). Both sets of authors analysed satellite IV DNA by Southern blotting and FISH in species of the Bovini and Tragelaphini. In contrast to the work done by Modi et al. (2004), Adega et al. (2006) found bovine satellite IV DNA in Tragelaphini. Modi et al. (1996) analysed the presence of six highly repeated DNA families isolated from cattle in a comparative FISH study comprising 46 species of the order Artiodactyla. These authors found two of the repeated families (*Pst* family and the 1.715 satellite I family)

present in all pecoran ruminants. Different restriction patterns of the 1.715 sequence were observed in different taxonomic families suggesting that independent concerted evolution events have homogenized different motifs in different lineages.

Most of the investigations referred above were conducted using caprine or bovine satellite DNAs separated from genomic DNA by density-gradient centrifugation. This approach ignores their chromosomal localization. In sharp contrast, the isolation of satellite DNAs by laser microdissection of centromeric regions has the advantage of chromosomal specificity (Pauciullo et al. 2006; Rubes et al. 2008; Louzada et al. 2008; Cernohorska et al. 2011, 2012). Moreover, it represents an easy and rapid method for the isolation of satellite DNAs of a larger number of species required for extensive studies.

This study extends our preliminary work based on the FISH analysis and distribution of tribe-specific repeats representing a pool of different satellite DNA families that coexist in centromeres (Kopecna et al. 2012). The aim of the study was to isolate various types of satellite DNAs and determine their co-localization in centromeric regions by FISH. Sequences representing the main motifs of the tribe-specific sat I were used for hybridizations across various bovid tribes and provided data for the phylogenetic analysis.

Materials and methods

Chromosome preparation

Material from 38 species representing 12 bovid tribes was used in this study. Blood samples from 33 taxa were taken from captive-born animals held in the zoological gardens in Dvur Kralove, Plzen, Liberec and Olomouc (Czech Republic) and cultured according to the standard protocols. Fibroblast cell cultures were established from a *Tetracerus quadricornis* female specimen held in the Menagerie du Jardin des Plantes, Paris. Fibroblast cell cultures from *Alcelaphus lichtensteinii* and *Oreotragus oreotragus* and cell suspensions of *Neotragus moschatus* and *Raphicerus sharpei* were obtained from Evolutionary Genomics Group, Department of Botany and Zoology, University of Stellenbosch, South Africa. Cell cultures were grown and harvested using conventional procedures (Gallagher and Womack 1992). Classification of Bovidae was carried out following the newest taxonomy by Groves and Grubb (2011). In case the name of species was changed, we present the former name in brackets (Wilson and Reeder 2005).

Isolation of tribe-specific satellite DNAs

The centromeric regions of autosomes of species representing each bovid tribe were microdissected by the MicroLaser system (Carl Zeiss MicroImaging GmbH, Munich, Germany). The pooled DNA was amplified by DOP-PCR without pretreatment as described by Kubickova et al. (2002). Amplification products were cloned into a pDrive vector (Qiagen, Hilden, Germany). Species-specific clones were selected by DOT BLOT hybridization (Pauciullo et al. 2006), fluorescently labelled and checked for specificity by FISH. Plasmid DNA was subsequently isolated and sequenced. Sequences comprised satellite DNA but were not long enough to represent the whole basic repeat unit. Therefore, primers amplifying the 5′- and 3′-flanking regions were designed. A simplified version of the inverse PCR was performed on isolated untreated genomic DNA. The primers were chosen with the emphasis on the distance between primers being as short as possible. This permitted the retrieval of almost full-length satellite DNA basic repeat units (Cernohorska et al. 2012). A simplified version of inverse PCR generates amplification product from untreated genomic DNA only in the case of repetitive sequences organized in tandem. In the case of Hippotragini, Alcelaphini and Tragelaphini, primers selected from one representative species of each tribe were used for the inverse PCR in other members of the same tribe. The amplification products representing the basic repeat unit obtained by PCR (BRU-PCR) were cloned and plasmids isolated. We chose a subset of clones on their Hae III RFLP patterns (most common and most dissimilar) for sequencing using BigDye terminator chemistry on an automated sequencer. The sequences were compared to those in the GenBank database using BLASTN searches. BLAST2 was used to assess sequence homologies. DNA sequences were screened for interspersed repeats by RepeatMasker. One species of each tribe was selected, and its most frequent clone was labelled with Orange-dUTP (Vysis, Richmond, UK) and used in the FISH analysis.

Preparation of sat II probe

The sat II probe for FISH on Antilopinae species was prepared from ovine genomic DNA according to the sequence from the GenBank (NCBI accession number AF245169). The primers were designed to amplify the region 24–385 of this sequence and generated a 362-bp PCR product. The sat II probe for FISH on Bovinae species was constructed from bovine genomic DNA using primers designed to amplify the region 241–584 of the bovine sequence (NCBI accession number M36668). The PCR products were cloned into pDrive Cloning Vectors (Qiagen) and recombinant plasmids were labelled by Green-dUTP (Vysis, Richmond, UK) using Nick Translation Reagent Kit (Vysis, Richmond, UK). Labelled clones were hybridized in double-colour FISH together with tribe-specific sat I to chromosome preparations of the selected species. The bovine satellite II primers were also used for the isolation and sequencing of sat II DNA from *Nyala angasii* (formerly *Tragelaphus angasii*).

Moreover, *Nanger dama*, *Gazella leptoceros*, *Antidorcas marsupialis* and *Antilope cervicapra* sat II sequences were obtained from each species genomic DNA using microdissection, cloning and inverse PCR (described above). These four Antilopini species were chosen for the isolation of sat II DNA on the basis of high intratribe variability of sat I sequences found in our study. The isolated sat II sequences were compared with sat II sequences available in the GenBank database using BLASTN searches.

Fluorescence in situ hybridization

Metaphase spreads for FISH analysis were prepared from lymphocytes or fibroblast cell cultures using standard methanol/acetic acid (3:1) fixation. Centromeric probes were hybridized as described by Kopecna et al. (2012). Slides were washed in 0.4xSSC/0.3 % Igepal at 73 °C, counterstained and mounted with 4',6-diamidino-2-phenylindole (DAPI)/Antifade. The FISH preparations were evaluated using Olympus BX 51 and BX 60 microscopes that were equipped with the necessary fluorescence filters and automated pad shifts. Good quality mitoses were scanned by CCD camera and evaluated by image analysis (ISIS 3, MetaSystems, Altlussheim, Germany).

Phylogenetic analysis

DNA sequences were pretreated to include the tandem repeat sequence. Two sets of analyses were performed. First, one sequence representing the most frequent clone per species that originated from this study was included. The dataset contained 38 sequences. Second, intraspecific variation was included with the most common and the most dissimilar clones per species (1–5 sequences) for a total of 100 sequences. Three species were added from the GenBank database. The datasets were analysed separately.

The sequences were aligned in Clustal Omega (Sievers et al. 2011; Goujon et al. 2010) using two combined iterations of the guide tree and the hidden Markov model algorithms.

The optimal substitution model was estimated from the Bayesian information criterion comparison of tree likelihoods of alternative substitution models with a fixed tree topology in jModeltest 2.1 (Darriba et al. 2012; Guindon and Gascuel 2003). The rate heterogeneity specific to alignment sites was modelled with the Γ distribution to avoid co-estimation of the α parameter of the Γ distribution and the proportion of invariable sites.

The maximum likelihood (ML) phylogeny was constructed in RAxML 7.4 (Stamatakis 2006) with rapid bootstrapping, followed by the full-likelihood search. Optimal number of bootstrap replicates was estimated with the automatic majority-rule bootstopping procedure (Pattengale et al. 2010).

Bayesian inference (BI) was used to reconstruct the phylogenetic trees in MrBayes 3.2 (Ronquist et al. 2012). To ensure convergence, two runs were employed with one cold and five heated Markov chains Monte Carlo (MCMC) for 2 million generations, sampled every 1,000th step and MCMC temperature set to 0.09. Discarded burn-in fraction represented the initial 700 trees. The hypothesis of monophyly of the subfamily Antilopinae was tested on the first dataset with Bayes factors from marginal likelihoods in Tracer 1.5 (Rambaut and Drummond 2007). The phylogenetic signal of insertions and deletions (indels) was investigated by coding the alignment gaps as binary characters and analysed as a partitioned mixed model together with nucleotide sequence information in BI.

Midpoint root was used. A node was considered supported when its bootstrap support was ≥70 and Bayesian posterior probability ≥0.95. Interpretation of

nodes with lesser reliability was considered speculative, and nodes with support <50 % were collapsed.

Results and discussion

Sequence analysis

Satellite DNAs were isolated from 38 species representative of the 12 bovid tribes (9 tribes in Antilopinae and 3 in Bovinae). A minimum of two clones were sequenced from each specimen, and the sequence data representing the respective BRU-PCR were deposited to GenBank under accession numbers KF787894–KF787988. The Cephalophini, Neotragini and Oreotragini were represented by a single species; at least two species were examined for the remainder: Alcelaphini (3), Antilopini (7), Boselaphini (2), Bovini (3), Caprini (3), Hippotragini (6), Reduncini (4) and Tragelaphini (6). The monotypic tribe Aepycerotini was represented by *Aepyceros melampus*. Sequences from three specimens were compared.

The isolated centromeric sequences were compared to those available in GenBank and on the basis of ascertained sequence homologies were categorized into three types: sat I, sat II and unclassified.

1. Sat I sequences

   All sat I sequences isolated in our study (GenBank: KF787894–KF787981) showed sequence similarity to the 1.714 or 1.715 sat DNA family. Sequence homologies of sat I clones within a tribe were ascertained using BLAST2. Sat I clones showed sequence similarities of >80 % in most of the tribes, but significant variations were found in Reduncini and Antilopini (Table 1). *Madoqua kirkii* (tribe Antilopini) showed low sequence similarity to other Antilopini species (70 %) and is therefore assessed separately below. The variability of the BRU-PCR among species within a tribe exceeded moderately the variability detected in the repeat units of a single species or for that matter, in a single specimen. Examples of intraspecific and intratribe variation are given in Table 4 for Hippotragini and Table 5 for Antilopini (see Supplementary), both tribes with a high number of analysed species. Hippotragini showed the highest average values of sequence similarity contrary to Antilopini with the lowest values. The most common BRU-PCR was ~800 bp in length and was found in Caprini, Alcelaphini, Hippotragini, Reduncini, Aepycerotini, Cephalophini, Antilopini and Neotragini. By contrast, that of the Bovini and Tragelaphini was ~1,400 bp. BRU-PCR of unique length occurred in Boselaphini (630 bp in *T. quadricornis*, 2,200 bp in *Boselaphus tragocamelus*) and Oreotragini (1,300 bp). These lengths dissimilar to the common form were found in all clones isolated from the distinct species. In addition to the ~1,300 bp long clone, representing the main satellite DNA of *O. oreotragus*, a minor BRU-PCR of ~800 bp and properties of sat I was detected. The atypical ~2,200 bp BRU-PCR found in *B. tragocamelus* is the result of an interspersed block of LTR elements (400 bp) that were revealed by Repeat Masker. The sequences of the most frequently encountered BRU-PCR from a specific tribal representative were chosen to determine intertribe orthologies (Table 3).

   It is possible to group the sat I sequences obtained in this investigation on the basis of their length and mutual homology and the degree of similarity shared with sequences of species archived in GenBank. The first group included those sequences isolated from members of the subfamily Bovinae: Bovini/Tragelaphini (1,400 bp) and Boselaphini (630 bp in *T. quadricornis*, 2,200 bp in *B. tragocamelus*). These showed sequence similarity 67–95 % to sequences of 1.715 sat I subfamily that were documented by

Table 1 Sequence similarity of all sat I clones from pairwise comparisons within a distinct tribe. Minimum, maximum and average values

	Sequence similarity (%)		
Tribe	Min	Max	Average
Caprini	82	95	87
Alcelaphini	91	98	94
Hippotragini	88	99	93
Reduncini	71	96	83
Aepycerotini	90	94	92
Cephalophini	83	96	84
Antilopini	67	96	76
Neotragini	87	90	88
Oreotragini	98	99	98
Tragelaphini	83	93	88
Bovini	81	97	84
Boselaphini	85	96	90

several authors (Modi et al. 1996; Chaves et al. 2005; Kopecna et al. 2012). On the other hand, the satellite sequences isolated from the subfamily Antilopinae: Caprini/Alcelaphini/Hippotragini/Reduncini/Aepycerotini/Cephalophini/Antilopini/Neotragini are ~800 bp long and together with Oreotragini (1,300 bp) form the second group that are similar (sequence similarity 68–96 %) to the common 1.714 sat I DNA (Jobse et al. 1995; Chaves et al. 2005). The sat I sequences of both groups were characterized by high similarity which was exclusive to members within each group. The last finding corresponds with the dissimilarity between 1.714 and 1.715 sat I. We found no significant sequence similarity between sequences of 1.714 and 1.715 sat I subfamilies available in the GenBank using MEGABLAST, only less strict BLASTN searches revealed a partial similarity.

The affiliation of our sat I sequences isolated across the whole family Bovidae either to 1.174 or 1.715 DNA subfamilies supports the well-established division of Bovidae into two subfamilies, Bovinae and Antilopinae (Hassanin and Douzery 1999; Matthee and Davis 2001; Hassanin and Douzery 2003; Ropiquet et al. 2009; Groves and Grubb 2011).

2. Sat II sequences

Satellite DNA II was isolated from four species of the Antilopini (*N. dama*, *A. marsupialis*, *G. leptoceros* and *A. cervicapra*), and the sequence data representing the respective BRU-PCR were deposited to GenBank under accession numbers KF787982–KF787985. BRU-PCR length was ~700 bp. We designated these sequences as sat II DNA on the basis of high sequence similarity (72–76 %) with ovine sat II DNA published by Buckland in 1985 (GenBank: X03117) and the absence of any internal repetition. We supplemented the published bovine sat II sequences with newly isolated orthologous sequences from *N. angasii* (GenBank: KF787986). This in turn demonstrated high sequence similarity to species within the subfamily Bovinae—82 % similarity with *Bos taurus* and *Bubalus bubalis*. In sharp contrast to highly conserved sat II DNA in Bovinae, sat II isolated from the four Antilopini species included in our investigation varied considerably (from 70 to 97 %) and showed sequence similarity to ovine sat II DNA (72–76 %). A rather high level of sequence similarity (73 %) exists between sat II DNAs of both subfamilies that consequently provide little phylogenetic signal. The absence of any internal repetition and similar length of repeat units described in Bovidae were found also in Cervidae (Qureshi and Blake 1995), indicating the conservativeness of satellite DNA II which, however, does not concern the sequence composition. Comparison of sat II sequences using BLASTN searches revealed existing sequence similarity only in species within the same family. We found only one case of interfamily sequence similarity (70 %) between sheep (GenBank: XO3117) and *Odocoileus virginianus* (GenBank: U49916) sat II DNA. These findings support the hypothesis that evolution of the satellite II DNA family in Bovidae and Cervidae occurred mainly by base substitution from an ancestral 700-bp tandem repeat (Tanaka et al. 1999).

3. Unclassified sequences

The unclassified centromeric clones were isolated from *M. kirkii* (GenBank: KF787987) and *Strepsiceros strepsiceros* (formerly *Tragelaphus strepsiceros*; GenBank: KF787988). These were similar in length (2,900 bp in *M. kirkii* cf 2,500 bp in *S. strepsiceros*), their repetitive tandem arrangement, and they also contained interspersed blocks of LTR elements (350 bp in *M. kirkii* and 220 bp in *S. strepsiceros*). These new repeats showed 86 % sequence similarity within these two species, but no similarity to any other satellite DNA in the GenBank database was found. BLASTN searches suggest high sequence orthology (82 % in *M. kirkii*, 84 % in *S. strepsiceros*) only to the "*Muntiacus muntjak vaginalis* BAC clone SatM5 centromeric sequence" (GenBank: EU433566). Our unclassified centromeric clones lie in the region, which does not represent muntjac satellite DNA described on SatM5 clone by Cheng et al. (2009) and are not present as repetitions in muntjac. Whereas these sequences probably remained preserved in this shape in most of the bovid species, they were amplified during the evolution to high copy number forming the main centromeric DNA in *M. kirkii* and substantial satellite DNA in *S. strepsiceros*.

FISH analysis

The most common sat I clone of each specific tribal representative was labelled with Orange-dUTP (Vysis,

Richmond, UK) and used as a probe for FISH analysis of species within and across various tribes. Table 2 summarizes the hybridization patterns obtained from tribe-specific sat I probes to different chromosomal groups (i.e. acrocentric and biarmed autosomes, the X and Y chromosomes) in 36 bovid species. The sat I probe gave large hybridization signals on the acrocentric autosomes, while the signals were much weaker in the

Table 2 FISH patterns with various tribe-specific sat I DNA probes to the different chromosome groups (acrocentric and biarmed autosomes, X and Y chromosomes) in the 36 Bovidae species analysed

Tribe	Species analysed (2*n*)	Acrocentric autosomes	Biarmed autosomes	X	Y
Caprini	*Ammotragus lervia* (58)	+/–	–	–	–
	Ovis aries (54)	+	+/–	–	–
	Ovibos moschatus (48)	+	–	+	–
Alcelaphini	*Damaliscus phillipsi* (38)	+	+	+	–
	Connochaetes taurinus (58)	+	–	+	–
	Alcelaphus lichtensteinii (40)	+	+	+	♀
Hippotragini	*Hippotragus niger* (60)	+		+	–
	Hippotragus equinus (60)	+		+	–
	Adax nasomaculatus (58)	+	+	+	–
	Oryx dammah (58)	+	–	+	–
	Oryx gazella (56)	+	+	+	–
	Oryx leucoryx (58)	+	–	+	–
Reduncini	*Redunca fulvorufula* (56)	+	+	+	–
	Kobus leche (48)	+	+	+	–
	Kobus megaceros (52)	+	+/–	+	–
	Kobus ellipsiprymnus (50)	+	+/–	+/–	–
Aepycerotini	*Aepyceros melampus* (60)	+		+	–
Cephalophini	*Cephalophus natalensis* (60)	+		–	♀
Antilopini	*Nanger dama* (38)	+/–	+/–	+	+
	Gazella leptoceros (32♀, 33♂)	+/–	+/–	+	+
	Antilope cervicapra (32♀, 33♂)	+	+/–	–	+
	Antidorcas marsupialis (56)	+	–	+	–
	Madoqua kirkii (46)	+	–	+	–
Neotragini	*Neotragus moschatus* (56)	+	+/–	+/–	–
Oreotragini	*Oreotragus oreotragus* (60)	+		+	♀
Tragelaphini	*Tragelaphus spekii* (30)	+	–	+	–
	Strepsiceros strepsiceros (32♀, 31♂)	+	–	+	–
	Ammelaphus imberbis (38)	+	+/–	+	+
	Tragelaphus euryceros (34♀, 33♂)	+	–	+	+
	Nyala angasii (56♀, 55♂)	+	+	+	–
	Taurotragus oryx (32♀, 31♂)	+	+	+	–
Bovini	*Bubalus bubalis* (50), river type	+	+/–	+	–
	Bos bonasus (60)	+		–	–
	Syncerus caffer (52)	+/–	+/–	+	–
Boselaphini	*Tetracerus quadricornis* (38)	+		+	♀
	Boselaphus tragocamelus (46)	+	+/–	+	–

(+) positive FISH signals on all chromosomes; (+/–) positive FISH signals only on several chromosomes or weak signal; (–) apparent absence of FISH signals; (♀) female, no Y chromosome; note 2*n*=60, species have no biarmed chromosomes

biarmed autosomes demonstrating a reduction in the size of the block of repeats following centric fusions in derived karyotypes (Modi et al. 1996). A reduction of the heterochromatin is considered to reflect the age of the fusion event—the more recent, the greater the quantity of heterochromatin still remaining (Iannuzzi et al. 1987; Chaves et al. 2000; Di Meo et al. 2006; Rubes et al. 2008; Kopecna et al. 2012). All acrocentric X chromosomes were FISH positive—the exception being the Caprini. Biarmed X chromosomes showed no sat I hybridization. Strikingly, however, in *N. dama*, *Tragelaphus spekii*, *Ammelaphus imberbis* (formerly *Tragelaphus imberbis*) and *B. tragocamelus* (all species with X autosome translocations), sat I was localized to the boundary between the two chromosomal compartments. The Y chromosomes failed to show any hybridization to the sat I probe in the majority of species, although there were exceptions within Antilopini and Tragelaphini. The hybridization motifs of tribe-specific satellite DNAs on autosomes and sex chromosomes were thoroughly described in Antilopini (Cernohorska et al. 2011) and Tragelaphini (Rubes et al. 2008). In the study of Cabelova et al. (2012), distribution of sat I sequences on Y chromosomes was well documented in representatives of nine bovid tribes.

The intensity of hybridization signals of the sat II probe was relatively weak when compared to those obtained with the sat I probe—the exception being the biarmed autosomes where intensities of both satellite probes were similar.

The unclassified centromeric clones isolated from *M. kirkii* and *S. strepsiceros* genomic DNA hybridized reciprocally resulting in identical patterns due to the high sequence orthology. Within Bovidae, these clones gave positive FISH results with only Tragelaphini and the two Antilopini species (*M. kirkii* and *Antidorcas marsupialis*). The hybridization patterns obtained with the unclassified clones were similar to sat I probes (except for *M. kirkii*), although signal intensity differed.

Results of FISH using tribe-specific sat I probes on representatives of the 12 tribes of the family Bovidae are shown in Table 3. In our experiments, the sat I probe specific for a particular tribe always resulted in positive in situ hybridization in all tested species of the tribe in question. The finding that Bovini and Tragelaphini species have similar sat I structure and homology is confirmed by positive FISH results. Similarly, positive cross-hybridization reflects significant sequence similarity of sat I of Bovini, Tragelaphini and Boselaphini. As anticipated, no positive FISH signals were detected when species of the subfamily Bovinae (Bovini/Tragelaphini/Boselaphini) were hybridized to those of the subfamily Antilopinae. The Antilopinae sat I (~800 bp in Caprini, Alcelaphini, Hippotragini, Reduncini, Aepycerotini, Cephalophini, Antilopini and Neotragini or 1,300 bp in Oreotragini) had substantial sequence similarity as demonstrated in the intensity levels of cross-hybridization signals when compared among Antilopinae specimens. This is consistent with our phylogenetic analysis of sat sequences that grouped the taxa into two distinct evolutionary clades, the Bovinae and Antilopinae (see below). However, within subfamilies, no correlation exists between the intensity of FISH signal detected and the degree of sequence similarity. An obvious explanation may be that this reflects differences in copy number of repeat units in the centromeric regions of the compared species. Moreover, our satellite DNA shows that intraspecific variability ranging from 70 to 96 % exists among repetitive units. Clone sequences represent one of many variable repeat units. A probe with high homology to one subunit can show lower homology to other subunits, lowering hybridization efficiency (Cabelova et al. 2012).

Co-localization of the sat I, sat II and unclassified repeat families

Double-colour FISH with tribe-specific sat I probes and bovine sat II probe was conducted on the chromosomes of several species of the subfamily Bovinae. Sat I and sat II sequences were localized to corresponding sites in all species. Sat I was found at centromeric regions, while sat II mapped to the outer boundaries of the centromeric regions, partly overlapping with sat I. In our FISH experiments on various Bovinae species, we obtained similar results as Tanaka et al. (1999, 2000) did on species of the genus *Bubalus*.

The co-localization of the tribe sat I probes and ovine sat II probes was similarly examined in species of the subfamily Antilopinae. Both sat I and II usually hybridized in the centromeric regions of chromosomes in Alcelaphini, Hippotragini, Reduncini, Aepycerotini, Cephalophini, Neotragini and the two Antilopini species (*A. cervicapra*, *G. leptoceros*). A unique hybridization pattern of sat I FISH signals at the centromeric regions and sat II at the pericentromeric regions was noted in Caprini and *A. marsupialis*, respectively. D'Aiuto et al.

Table 3 Results of FISH with tribe-specific sat I probes on various species representing 12 bovid tribes and intertribe sequence similarity (percentage values) obtained by comparison of the most frequent clones in representatives of distinct tribes

Tribe, species analysed	Origin (tribe, species) of centromeric DNA probe													
	Caprini *A. lervia*	Alcelaphini *D. phillipsi*	Hippotragini *H. niger*	Reduncini *R. fulvorufula*	Aepycerotini *A. melampus*	Cephalophini *C. natalensis*	Antilopini *G. dama*	Neotragini *A. marsupialis*	Oreotragini *M. kirkii*	Tragelaphini *N. moschatus*	Bovini *O .oreotragus*	Boselaphini *T. spekei*	*B. bubalis*	*T. quadricornis*
Caprini	+++	+++	+++	++	+	+ to ++	0	0	0	+	++	0	0	0
Ammotragus lervia	100	80	83	76	75	76	69	68	66	69	72	no hom.	no hom.	no hom.
Alcelaphini	+++	++++	++ to +++	0 to +	+	+	0	0 to +	0 to +	0	+	0	0	0
Damaliscus phillipsi	80	100	80	73	73	74	67/68	69	65	72	70	no hom.	no hom.	no hom.
Hippotragini	++++	++++	++++	+++ to ++++	+ to ++	+++ to ++++	0 to +	0 to +	0 to +	++ to +++	+	0	0	0
Hippotragus niger	83	80	100	74	75	76	70/68	70	67	73	74/71	no hom.	no hom.	no hom.
Reduncini	+++ to ++++	++	++ to +++	++++	+ to ++	+	0	0 to +	0 to +	+ to 0	+	0	0	0
Redunca fulvorufula	76	73	74	100	75	76	66/68	69	65	71	71	no hom.	no hom.	no hom.
Aepycerotini	++ to +++	+++	+++	+++ to ++++	++++	++	0	++++	+	++	0 to +	0	0	0
Aepyceros melampus	75	73	75	75	100	77	68/71	72	67	70	71	no hom.	no hom.	no hom.
Cephalophini	+++	+	++++	+	+	++++	0 to +	0	0	0 to +	+	0	0	0
Cephalophus natalensis	76	74	76	76	77	100	69	72	67	71	73/72	no hom.	no hom.	no hom.
Antilopini	0 to +	0	0	0	+	0	++++	++++	++ to +++	0	0	0	0	0
Nanger dama	69	67/68	70/68	66/68	68/71	69	100	77	70/67	67	66/65	no hom.	no hom.	no hom.
	+ to ++	0 to +	0 to +	0 to +	++	0 to +	+++	++++	+++	0 to +	0	0	0	0
Antidorcas marsupialis	68	69	70	69	72	72	77	100	69	no hom.	69/67	no hom.	no hom.	no hom.
	+	0	0 to +	0	0 to +	0	++	++	++++	++	0	0	0	0
Madoqua kirkii	66	65	67	65	67	67	70/67	69	100	no hom.	66/65	no hom.	no hom.	no hom.
Neotragini	++	+	++ to +++	++	0	+	0	0	0	++++	0	0	0	0
Neotragus moschatus	69	72	73	71	70	71	67	no hom.	no hom.	100	no hom.	no hom.	no hom.	no hom.
Oreotragini	++++	++	++ to +++	++ to +++	++ to +++	+++	+	+	+	++	++++	0	0	0
Oreotragus oreotragus	72	0	74/71	71	71	73/72	66/65	69/67	66/65	no hom.	100	no hom.	no hom.	no hom.
Tragelaphini	0	0	0	0	0	0	0	0	0	0	0	++++	+ to ++	++
Tragelaphus spekii	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	100	68	68
Bovini	0	0	0	0	0	0	0	0 to +	0	0	0	+ to ++	++++	++
Bubalus bubalis	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	68	100	74
Boselaphini	0	0	0	0	0	0	0	0	0	0	0	+	++	+++

Table 3 (continued)

Tribe, species analysed	Origin (tribe, species) of centromeric DNA probe													
	Caprini *A. lervia*	Alcelaphini *D. phillipsi*	Hippotragini *H. niger*	Reduncini *R. fulvorufula*	Aepycerotini *A. melampus*	Cephalophini *C. natalensis*	Antilopini *G. dama*	Neotragini *A. marsupialis*	Oreotragini *M. kirkii*	Tragelaphini *N. moschatus*	Bovini *O .oreotragus*	Boselaphini *T. spekei*	*B. bubalis*	*T. quadricornis*
Tetracerus quadricornis	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	no hom.	68	74	100

(+) positive FISH signals, number of + corresponds to the intensity of signals; (0) apparent absence of FISH signals; (no hom.) no homology found using BLAST2 programme; (67/68) two blocks with different sequence similarity values found in sequences using BLAST2 programme

(1997) found this specific pattern in one Caprini species (*Ovis aries*). The converse was observed in *N. dama* where sat II localized to the centromeric regions and sat I to the pericentromeric regions, respectively. In gazelles, the distribution of sat I and sat II DNAs was described in more details in the study by Cernohorska et al. (2012).

Three satellite DNA probes (sat I, sat II and unclassified) hybridized to discrete chromosomal domains in *M. kirkii* (Fig. 1). Signals of the unclassified centromeric clone from *M. kirkii* were bordered by sat I in the pericentromeric areas of chromosomes, and sat II was placed on the opposite proximal side. FISH of the three satellite DNA probes on *O. oreotragus* chromosomes revealed that the 1,300 bp sat I hybridized to the centromeric regions of all chromosomes, whereas the 800 bp sat I clone gave strong signals on the distal portions of the centromeric regions of half of the chromosomes (Fig. 2). The patterns found for ovine sat II were similar to those detected with the 800 bp clone, but these were present on all chromosomes and with lower hybridization intensity. In our FISH experiments, the hybridization signals appeared as one compact dot in centromeres, while at the pericentromeric regions lying on the separate chromatids, the signals split into two discrete dots. This phenomenon was observed regardless of the type of satellite DNA used.

Phylogenetic analysis

Species-level phylogeny of Bovidae sat I DNA sequences

The alignment of the sequences from 38 species isolated in this study was 1,854 bp long, where 47.3 % of the alignment consisted of gaps. The ML analysis converged after 550 bootstrap replicates, and bootstrapping was stopped. Convergence of the runs of the BI analysis was demonstrated with the final values of the average standard deviation of split frequencies <0.01, the potential scale reduction factor approaching 1.000 for all model parameters, and the 95 % highest posterior density interval of the tree length that included its ML estimate. These results show that the analyses found the global optimum in the tree space, and the trees were reliable and suitable for interpretation.

Both ML and BI trees were in accord in resolving the relationships between investigated sat I DNA sequences (Fig. 3). Here, the sequences from subfamilies

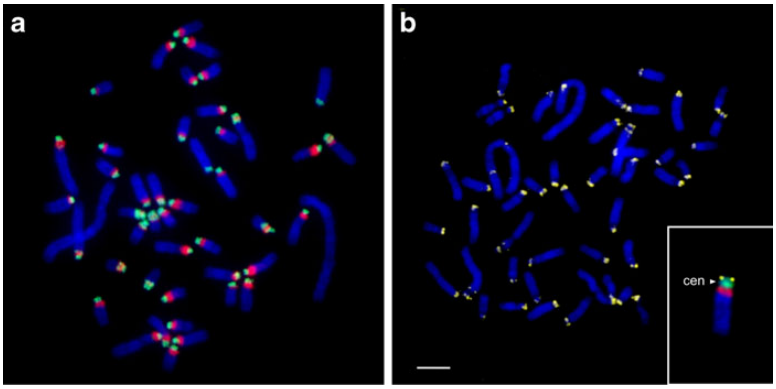


Fig. 1 **a** Co-hybridization of sat I clone (*red*) and the unclassified centromeric clone (*green*), both isolated from *Madoqua kirkii* to metaphase chromosomes of *M kirkii*. **b** FISH with sat II probe derived from *Antidorcas marsupialis* to chromosomes of *M. kirkii*. **Inset** shows co-hybridization of sat I (*red*), sat II (*yellow*) and the unclassified centromeric clone (*green*) to a chromosome of *M. kirkii*. Chromosomes were counterstained with DAPI (blue). *Scale bar* represents 5 μm

Antilopinae and Bovinae were clearly separated with midpoint rooting of the phylogenies. This is in agreement with the current phylogeny of Bovidae (Decker et al. 2009; Hassanin et al. 2012; Ropiquet et al. 2009), indicating, at this level, congruence between bovid phylogeny based on sat I DNA sequences and other markers. Within the subfamilies, sequences from all tribes with multiple sampled members were monophyletic, albeit their relationships were often unresolved. Tribes in Bovinae diverged fast after differentiation of the subfamily (Ropiquet et al. 2009; Hassanin et al. 2012; Bibi 2013), and the sat I DNA sequences provided no resolution between them. Tribal level structure in Antilopinae based on our data consistently showed basal position of Oreotragini sequences in the group, followed by a rapid divergence of sequences representing other tribes as shown in the Bayesian phylogeny. This shows better resolution compared to the previous knowledge of the group where in phylogenies reconstructed from other genomic sequences Oreotragini showed variable positions at deeper divergences of Antilopinae (Ropiquet et al. 2009; Bibi 2013). Aepycerotini were unstable with no supported affinity in our analyses. The phylogenetic position of the sole representative of the tribe, *Aepyceros melampus*, is notoriously difficult to ascertain from DNA sequence data. It is either in an unsupported

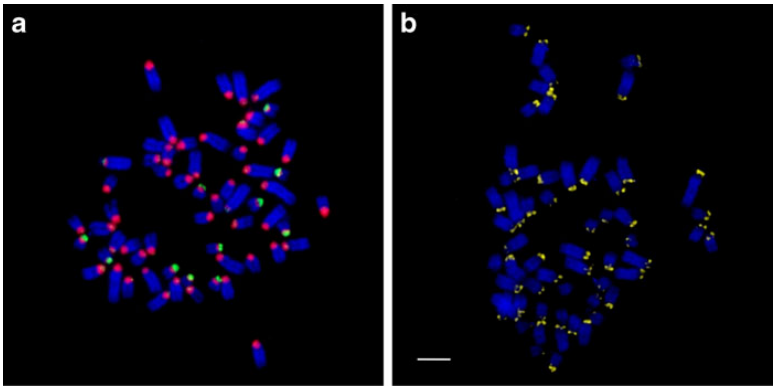


Fig. 2 **a** Double-colour FISH with the 1,300 bp sat I clone (*red*) and the 800 bp sat I clone (*green*), both isolated from *Oreotragus oreotragus* to metaphase chromosomes of *O. oreotragus*. **b** FISH with ovine sat II to chromosomes of *O. oreotragus*. Chromosomes were counterstained with DAPI (*blue*). *Scale bar* represents 5 μm

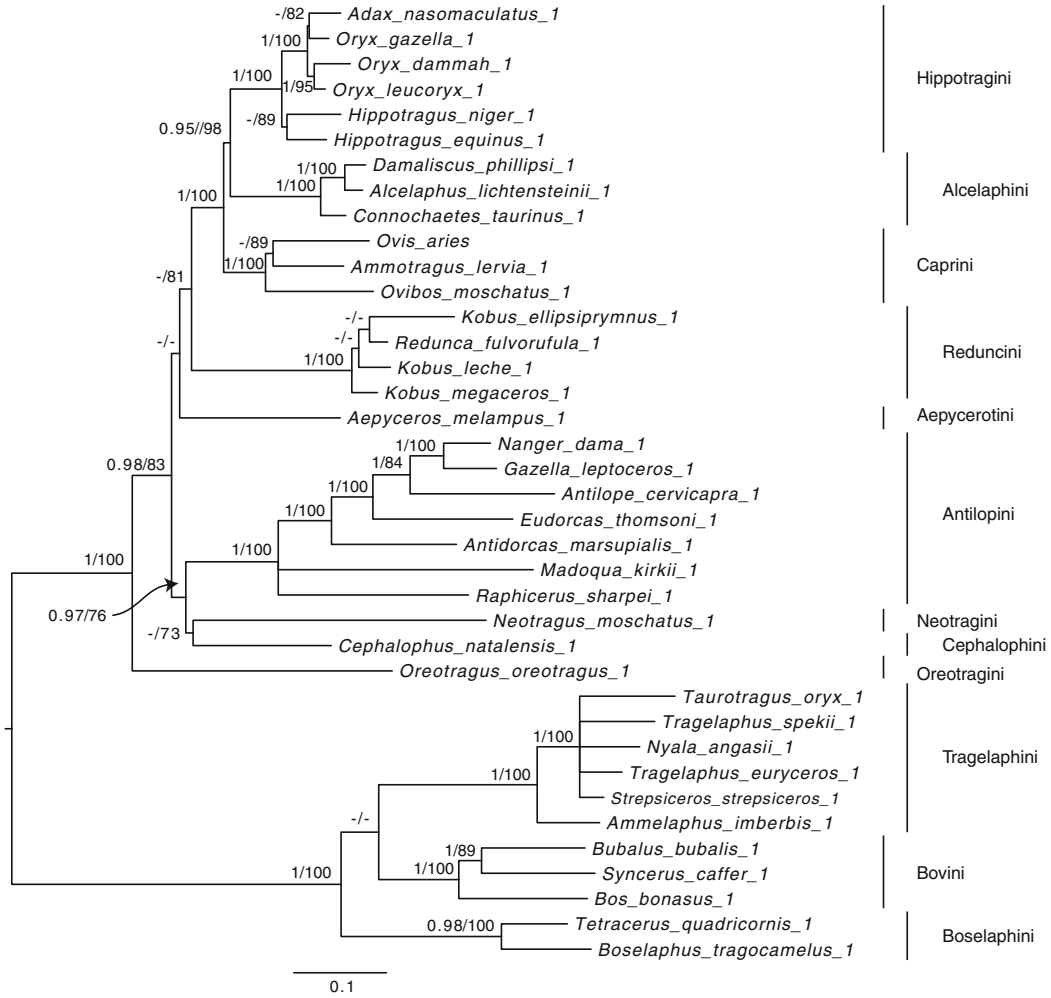


Fig. 3 Bayesian phylogenetic tree based on species representative sat I DNA sequences isolated in this study. Numbers above branches represent Bayesian posterior probability/maximum likelihood bootstrap support. Nodes indicated with a *dash* were unsupported in the analysis (see "Materials and methods" for details)

position (Decker et al. 2009; Ropiquet et al. 2009; Hassanin et al. 2012) or it forms a sister relationship with Neotragini (Bibi 2013; Hassanin et al. 2012). In recent analyses, including this study, it branches close to the base of the Antilopinae, indicating that it is a relict from an early divergence of the group. Sat I DNA sequences from members of Antilopini, Cephalophini and Neotragini formed a group, wherein the ML analysis distinguished a sister relationship of Cephalophini and Neotragini sequences that was absent in the BI phylogeny. A close relationship between Cephalophini and Neotragini seems to be unresolved in mitochondrial phylogenies (Hassanin et al. 2012) or Neotragini form a sister relationship with Aepycerotini (Bibi 2013). This conflict in placement of Neotragini between mitochondrial and sat DNA phylogenies indicates the need for thorough investigation of nuclear markers to ascertain the origin of the group. Similarly, sequences belonging to the tribe Reduncini formed a basal group with respect to those from the remaining tribes in the ML tree only.

Both analyses showed that Hippotragini and Alcelaphini sat I DNA sequences were closely related and sister to Caprini (Fig. 3).

Sat I DNA phylogeny with intraspecific variation in Bovidae

The alignment that included within-species variation in the sat I DNA sequences consisted of 41 species, 100 sequences, and it was 2,058 bp long where 52.9 % of sites were gaps. Majority-rule-based bootstopping criterion indicated 400 replicates as sufficient for the estimation of the ML phylogeny. The BI analysis with this data did not converge with the given MCMC settings, and the number of generations was increased to 3 million with the proportional increase in burn-in fraction.

The subfamily- and the tribal-level topology of the sequences on the trees that included intraspecific variation in the sat I DNA (see Supplementary Figs. 4 and 5) was comparable to that described above. The main difference was that the sequences from Oreotragini species were no longer basal to those from the Antilopinae. In trees with intraspecific variation, the Antilopinae sat I DNA sequences exhibited a rapid burst of differentiation that affected representatives from all the tribes except Caprini, Alcelaphini and Hippotragini that formed a supported group.

Interestingly, throughout the tree, the alternative sequences obtained from the same individuals did not form sister relationships within several tribes. Antilopini, Bovini, Caprini, Hippotragini and Tragelaphini conspecific sequences generally did not group together. This could be explained by parallel differentiation of sat DNA at variable loci. The sequences from the tribe Reduncini displayed a different topology that leads to an alternative interpretation. Here, two lineages formed, and they exhibited long branches similar to those distinguishing sequences from some other tribes. One of these lineages is likely driven by a long indel. Analyses, which included phylogenetic information from indels, placed this lineage at the base of the whole tree in a sister relationship with sat I DNA from *Tetracerus quadricornis* (see Supplementary Fig. 6). While this is a computational artefact caused by apparent lack of shared evolutionary history between the sequences that share large indels, the case of Reduncini sat I DNA demonstrated the putative heterogeneity of the dataset. The sequences were identified here based on similarity, which is consistent with the evolutionary divergence only in some cases. In Reduncini, the two lineages represent disparate evolutionary histories of multiple repeat units.

Elsewhere in the Antilopinae tree, the phylogenetic information from indels stabilized the relationships between tribal sat I DNA sequences. The sequences at the subfamily level split into two groups, and each group had a distinct ladder-like topology, meaning that tribal sat DNA branched from the common ancestral sequence successively. In the first group, Neotragini sat I DNA sequences were basal, followed by a ladder-like differentiation of sequences from Cephalophini, Aepycerotini and Antilopini. Similarly, in the second group, sat I DNA sequences from Reduncini were basal, and then those from Caprini, Alcelaphini and Hippotragini diverged (see Supplementary Fig. 6). These relationships were indicated albeit often unresolved in an otherwise highly resolved phylogeny based upon 40,843 genome-wide SNP constructed by Decker et al. (2009).

The phylogenetic relationships of sat DNA sequences within family Bovidae showed relationships roughly consistent with the current taxonomy (Groves and Grubb 2011). Where our phylogeny diverged from expectations, the phenomena could have resulted from rapid divergence in early evolution of a group, incomplete lineage sorting or analyses of paralogous repeat units.

In conclusion, we isolated and examined so far unpublished sat I sequences from non-domestic species representative of all tribes of the family Bovidae. All of them showed high sequence similarity to 1.714 or 1.715 satellite I DNA and, with few exceptions, were of similar length. Affiliation of our sat I sequences to ovine 1.714 or bovine 1.715 sat DNA corresponds with their assignment to the respective subfamily (Antilopinae vs. Bovinae). Our sat I sequences showed sequence variability sufficient for providing successful phylogenetic analysis. However, high intraspecific variability enabled the differentiation only at the subfamily and tribal levels. On the other hand, sat II DNA sequences isolated in our study are more conservative and lack the variability essential for phylogenetic studies. Apart from sat I and sat II sequences, we found new, as yet undescribed centromeric clones in *M. kirkii* and *S. strepsiceros* that showed no similarity to any other known satellite DNA. To sum up, the present investigation extends what is known of the structure, distribution and phylogenetic spread of satellite repeats in Bovidae and particularly non-domestic species.

Acknowledgments We thank Terence J. Robinson and Anne Ropiquet for kindly providing several samples and for useful comments on a draft of the manuscript. This work was supported by the Ministry of Agriculture of the Czech Republic (project MZE 0002716202), the Grant Agency of the Czech Republic (grant P506/10/0421) and by the project "CEITEC—Central European Institute of Technology" (CZ.1.05/1.1.00/02.0068) from European Regional Development Funds.

References

Adega F, Chaves R, Guedes-Pinto H, Heslop-Harrison JS (2006) Physical organization of the 1.709 satellite IV DNA family in Bovini and Tragelaphini tribes of the Bovidae: sequence and chromosomal evolution. Cytogenet Genome Res 114:140–146

Bibi F (2013) A multi-calibrated mitochondrial phylogeny of extant Bovidae (Artiodactyla, Ruminantia) and the importance of the fossil record to systematics. BMC Evol Biol 13:166

Buckland RA (1985) Sequence and evolution of related bovine and caprine satellite DNAs. J Mol Biol 186:25–30

Buckland RA, Evans HJ (1978) Cytogenetic aspects of phylogeny in the Bovidae. I. G-banding. Cytogenet Cell Genet 21:42–63

Cabelova K, Kubickova S, Cernohorska H, Rubes J (2012) Male-specific repeats in wild Bovidae. J Appl Genet 53:423–433

Cernohorska H, Kubickova S, Vahala J, Robinson TJ, Rubes J (2011) Cytotypes of Kirk's dik-dik (Madoqua kirkii, Bovidae) show multiple tandem fusions. Cytogenet Genome Res 132:255–263

Cernohorska H, Kubickova S, Vahala J, Rubes J (2012) Molecular insights into X; BTA5 chromosome rearrangements in the tribe Antilopini (Bovidae). Cytogenet Genome Res 136:188–198

Chaves R, Guedes-Pinto H, Heslop-Harrison J, Schwarzacher T (2000) The species and chromosomal distribution of the centromeric alpha-satellite I sequence from sheep in the tribe Caprini and other Bovidae. Cytogenet Cell Genet 91:62–66

Chaves R, Guedes-Pinto H, Heslop-Harrison JS (2005) Phylogenetic relationships and the primitive X chromosome inferred from chromosomal and satellite DNA analysis in Bovidae. Proc Biol Sci 272:2009–2016

Cheng YM, Li TS, Hsieh LJ, Hsu PC, Li YC, Lin CC (2009) Complex genomic organization of Indian muntjac centromeric DNA. Chromosome Res 17:1051–1062

D'Aiuto L, Barsanti P, Mauro S, Cserpan I, Lanave C, Ciccarese S (1997) Physical relationship between satellite I and II DNA in centromeric regions of sheep chromosomes. Chromosome Res 5:375–381

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772

Decker JE, Pires JC, Conant GC et al (2009) Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. Proc Natl Acad Sci U S A 106:18644–18649

Di Meo GP, Perucatti A, Chaves R, Adega F, De Lorenzi L, Molteni L, De Giovanni A, Incarnato D, Guedes-Pinto H, Eggen A, Iannuzzi L (2006) Cattle rob(1;29) originating from complex chromosome rearrangements as revealed by both banding and FISH-mapping techniques. Chromosome Res 14:649–655

Gallagher DS Jr, Womack JE (1992) Chromosome conservation in the Bovidae. J Hered 83:287–298

Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R (2010) A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res 38:W695–W699

Groves C, Grubb P (2011) Ungulate taxonomy. The Johns Hopkins University Press, Baltimore

Guindon S, Gascuel O (2003) A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. Syst Biol 52:696–704

Hassanin A, Douzery EJ (1999) Evolutionary affinities of the enigmatic saola (*Pseudoryx nghetinhensis*) in the context of the molecular phylogeny of Bovidae. Proc Biol Sci 7:893–900

Hassanin A, Douzery EJ (2003) Molecular and morphological phylogenies of Ruminantia and the alternative position of the Moschidae. Syst Biol 52:206–228

Hassanin A, Delsuc F, Ropiquet A, Hammer C, Jansen van Vuuren B, Matthee C, Ruiz-Garcia M, Catzeflis F, Areskoug V, Nguyen TT, Couloux A (2012) Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. C R Biol 335:32–50

Iannuzzi L, Di Berardino D, Gustavsson I, Ferrara L, Di Meo GP (1987) Centromeric loss in translocations of centric fusion type in cattle and water buffalo. Hereditas 106:73–81

Jobse C, Buntjer JB, Haagsma N, Breukelman HJ, Beintema JJ, Lenstra JA (1995) Evolution and recombination of bovine DNA repeats. J Mol Evol 41:277–283

Kopecna O, Kubickova S, Cernohorska H, Cabelova K, Vahala J, Rubes J (2012) Isolation and comparison of tribe-specific centromeric repeats within Bovidae. J Appl Genet 53:193–202

Kubickova S, Cernohorska H, Musilova P, Rubes J (2002) The use of laser microdissection for the preparation of chromosome-specific painting probes in farm animals. Chromosome Res 10:571–577

Lee C, Wevrick R, Fisher RB, Ferguson-Smith MA, Lin CC (1997) Human centromeric DNA. Hum Genet 100:291–304

Louzada S, Paço A, Kubickova S, Adega F, Guedes-Pinto H, Rubes J, Chaves R (2008) Different evolutionary trails in the related genomes *Cricetus cricetus* and *Peromyscus eremicus* (Rodentia, Cricetidae) uncovered by orthologous satellite DNA repositioning. Micron 39:1149–1155

Macaya G, Cortadas J, Bernardi G (1978) An analysis of the bovine genome by density-gradient centrifugation. Preparation of the dG+dC-rich DNA components. Eur J Biochem 84:179–188

Matthee CA, Davis SK (2001) Molecular insights into the evolution of the family Bovidae: a nuclear DNA perspective. Mol Biol Evol 18:1220–1230

Modi WS, Gallagher DS, Womack JE (1996) Evolutionary histories of highly repeated DNA families among the Artiodactyla (Mammalia). J Mol Evol 42:337–349

Modi WS, Ivanov S, Gallagher DS (2004) Concerted evolution and higher-order repeat structure of the 1.709 (satellite IV) family in bovids. J Mol Evol 58:460–465

Nijman IJ, Lenstra JA (2001) Mutation and recombination in cattle satellite DNA: a feedback model for the evolution of satellite DNA repeats. J Mol Evol 52:361–371

Nowak RM (1999) Order Artiodactyla. In: Walker's Mammals of the World, vol 2, 6th edn. The Johns Hopkins University Press, Baltimore, 1051–1238

Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A (2010) How many bootstrap replicates are necessary? J Comput Biol 17:337–354

Pauciullo A, Kubickova S, Cernohorska H, Petrova K, Di Berardino D, Ramunno L, Rubes J (2006) Isolation and physical localization of new chromosome-specific centromeric repeats in farm animals. Vet Med 51

Qureshi SA, Blake RD (1995) Sequence characteristics of a cervid DNA repeat family. J Mol Evol 40:400–404

Rambaut A, Drummond AJ (2007) Tracer v1.4, Available from http://beast.bio.ed.ac.uk/Tracer

Robinson TJ, Ropiquet A (2011) Examination of hemiplasy, homoplasy and phylogenetic discordance in chromosomal evolution of the Bovidae. Syst Biol 60:439–450

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542

Ropiquet A, Li B, Hassanin A (2009) SuperTRI: a new approach based on branch support analyses of multiple independent data sets for assessing reliability of phylogenetic inferences. C R Biol 332:832–847

Rubes J, Kubickova S, Pagacova E, Cernohorska H, Di Berardino D, Antoninova M, Vahala J, Robinson TJ (2008) Phylogenomic study of spiral-horned antelope by cross-species chromosome painting. Chromosome Res 16:935–947

Saffery R, Earle E, Irvine DV, Kalitsis P, Choo KH (1999) Conservation of centromere protein in vertebrates. Chromosome Res 7:261–265

Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins D (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539

Skowronski J, Pluciennicz A, Bednarek A, Jaworski J, Bovine 1.709 satellite (1984) Recombination hotspots and dispersed repeated sequences. J Mol Biol 177:399–416

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690

Tanaka K, Matsuda Y, Masangkay JS, Solis CD, Anunciado RVP, Namikawa T (1999) Characterization and chromosomal distribution of satellite DNA sequences of the water buffalo (*Bubalus bubalis*). J Hered 90:418–422

Tanaka K, Matsuda Y, Masangkay JS, Solis CD, Anunciado RV, Kuro-o M, Namikawa T (2000) Cytogenetic analysis of the tamaraw (*Bubalus mindorensis*): a comparison of R-banded karyotype and chromosomal distribution of centromeric satellite DNAs, telomeric sequence, and 18S-28S rRNA genes with domestic water buffaloes. J Hered 91:117–121

Ugarković D, Plohl M (2002) Variation in satellite DNA profiles-causes and effects. EMBO J 21:5955–5959

Vaiman D, Billault A, Tabet-Aoul K, Schibler L, Vilette D, Oustry-Vaiman A, Soravito C, Cribiu EP (1999) Construction and characterization of a sheep BAC library of three genome equivalents. Mamm Genome 10:585–587

Wilson DE, Reeder DM (2005) Mammal species of the world. Johns Hopkins University Press, Baltimore

Wurster DH, Benirschke K (1968) Chromosome studies in the superfamily Bovoidea. Chromosoma 25:152–171

Paper 2.1.8

Wallace I. S., Shakesby A. J., Hwang J. H., Choi W. G., **Martínková N.**, Douglas A. E., Roberts D. M. 2012. *Acyrthosiphon pisum* AQP2: A multifunctional insect aquaglyceroporin. *BBA Biomembranes* 1818: 627-635.

Biochimica et Biophysica Acta 1818 (2012) 627–635

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbamem

Acyrthosiphon pisum AQP2: A multifunctional insect aquaglyceroporin

Ian S. Wallace [a,1], Ally J. Shakesby [b], Jin Ha Hwang [a], Won Gyu Choi [a], Natália Martínková [b,c], Angela E. Douglas [b,d], Daniel M. Roberts [a,*]

[a] Department of Biochemistry & Cellular, and Molecular Biology, The University of Tennessee, Knoxville, Knoxville, TN, 37996–0840, USA
[b] Department of Biology, University of York, York, YO10 5DD, UK
[c] Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, v.v.i., Květná 8, 603 65 Brno, Czech Republic
[d] Department of Entomology, Comstock Hall, Cornell University, Ithaca, NY 14850, USA

ARTICLE INFO

Article history:
Received 31 August 2011
Received in revised form 19 November 2011
Accepted 28 November 2011
Available online 8 December 2011

Keywords:
Aphid
Aquaporins
Buchnera aphidicola
Osmoregulation
Polyols
Symbiosis

ABSTRACT

Annotation of the recently sequenced genome of the pea aphid (*Acyrthosiphon pisum*) identified a gene *ApAQP2* (ACYPI009194, Gene ID: 100168499) with homology to the Major Intrinsic Protein/aquaporin superfamily of membrane channel proteins. Phylogenetic analysis suggests that ApAQP2 is a member of an insect-specific clade of this superfamily. Homology model structures of ApAQP2 showed a novel array of amino acids comprising the substrate selectivity-determining "aromatic/arginine" region of the putative transport pore. Subsequent characterization of the transport properties of ApAQP2 upon expression in *Xenopus* oocytes supports an unusual substrate selectivity profile. Water permeability analyses show that the ApAQP2 protein exhibits a robust mercury-insensitive aquaporin activity. However unlike the water-specific ApAQP1 protein, ApAQP2 forms a multifunctional transport channel that shows a wide permeability profile to a range of linear polyols, including the potentially biologically relevant substrates glycerol, mannitol and sorbitol. Gene expression analysis indicates that ApAQP2 is highly expressed in the insect bacteriocytes (cells bearing the symbiotic bacteria *Buchnera*) and the fat body. Overall the results demonstrate that ApAQP2 is a novel insect aquaglyceroporin which may be involved in water and polyol transport in support of the *Buchnera* symbiosis and aphid osmoregulation.

1. Introduction

Major Intrinsic Proteins (MIPs) are an ancient class of integral membrane protein channels that facilitate the selective bidirectional transport of water and uncharged solutes across biological membranes [1]. The aquaporins represent the best characterized transporters in this family, and these proteins play diverse roles in physiology by mediating the bulk movement of water driven by osmotic and pressure gradients [1–3]. Members of the MIP superfamily share a signature topology and pore architecture, referred to as the "hourglass fold". This topology consists of six transmembrane α-helices related by a pseudo two-fold symmetry [4]. Additionally, two highly conserved loops between transmembrane α-helices 2 and 3, and transmembrane α-helices 5 and 6 form two smaller α-helical segments (the "NPA" boxes) which fold back into the protein and pack with the six transmembrane α-helices, forming a 7th pseudo transmembrane helix [4]. The pore formed by the packing of these helices resembles an hourglass with the transport selectivity determined by a narrow constriction, the "aromatic/arginine" (ar/R) selectivity filter [4,5]. The ar/R selectivity filter is defined by four residues (one each from transmembrane helices 2 and 5, and two from the interhelical loop containing the second NPA box) that mediate transport selectivity based on solute size and hydrophobicity. Vertebrate MIPs generally fall into two broad transport classes, water-selective aquaporins and multifunctional aquaglyceroporins that differ in the amino acid composition of the ar/R [3].

While the transport behavior and physiology of vertebrate aquaporins have been the subject of intensive study [3], the structure, function, and physiology of invertebrate MIPs are less well studied [2]. In the case of insects (e.g. bed bugs, mosquitoes, aphids, plant hoppers) that ingest large volumes of vertebrate blood or plant sap, aquaporins have been implicated in the bulk water movement across various cellular membranes in the gut, participating in volume and osmotic homeostasis, and fluid excretion [2,6–9]. For example, a water-selective aquaporin, ApAQP1, expressed in both the stomach and closely juxtaposed distal intestine of the plant-sap feeding pea aphid (*Acyrthosiphon pisum*) has been proposed to serve an osmoregulatory role in water cycling [9].

Analysis of the recently sequenced genome of *A. pisum* [10] revealed the presence of a second MIP gene encoding an aquaporin-like protein (ApAQP2) that is represented among ESTs obtained from both whole aphids [11] and isolated bacteriocytes [12]. Bacteriocytes are specialized aphid cells that house and maintain intracellular symbiotic γ-

* Corresponding author. Tel.: +1 865 974 4070; fax: +1 865 974 6306.
E-mail address: drobert2@utk.edu (D.M. Roberts).
[1] Present address: The Energy Biosciences Institute, 130 Calvin lab MC5230, Berkeley, CA, 94720, USA.

0005-2736/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbamem.2011.11.032

proteobacteria *Buchnera aphidicola* (reviewed in [13]). The symbiosis has a nutritional basis: *Buchnera* provide the insect with essential amino acids that are in short supply in the aphid diet of plant phloem sap [14,15]. In this study, molecular modeling and functional analysis reveal that ApAQP2 is a multifunctional aquaglyceroporin channel that shows unique ar/R pore structure and exhibits permeability to a wide range of physiologically relevant polyols and is localized to both bacteriocytes and the fat body (an insect organ that functions in energy storage and immunity). The potential significance of this second aquaporin channel for the physiology of the insect, including its symbiosis with intracellular bacteria, is discussed.

2. Materials and methods

2.1. ApAQP2 cDNA cloning

Acyrthosiphon pisum clone LL01 was maintained on broad bean *Vicia faba* cv. The Sutton, grown at 20 °C with a 16 h light/8 h dark cycle. Aphids were homogenized in ice-cold TRIzol reagent (Invitrogen) and RNA was extracted following manufacturer's instructions (Invitrogen). To remove contaminating DNA, the RNA was incubated with RNAse-free DNAseI (Roche) for 30 min at 37 °C then at 75 °C for 15 min before purification with the RNAeasy minikit (Qiagen) using the RNA cleanup protocol.

For *ApAQP2* cDNA cloning, first strand cDNA was synthesized from RNA using Superscript II reverse transcriptase (Invitrogen) following the manufacturer's instructions. *ApAQP2* was amplified in 1× PCR buffer, 1 mM $MgSO_4$, 0.2 mM of each dNTP, 1 U KOD Hot Start Polymerase (Toyobo), 2 μL template (ca. 0.5 μg) and 1 μM gene-specific primers (Supplemental Table 1) with the following conditions: 31 cycles of 94 °C for 1 min, 35 °C for 1 min, and 72 °C for 2 min 30 s. PCR products were purified by electrophoresis in a 1% (w/v) agarose gel and were cloned into the pCR®-Blunt II-TOPO® plasmid vector (Invitrogen) following the manufacturer's instructions. Cloned PCR products were identified as *ApAQP2* by sequencing using an Applied Biosystems 3130 Genetic Analyzer (University of York Technology Facility, York, U.K.).

2.2. Real time Q-PCR analysis of ApAQP2 transcript abundance

Embryo, fat body, gut, bacteriocyte and head tissues were dissected from 7-day-old final instar larvae using fine pins and scissors, and the RNA was extracted from these tissues and parallel samples of whole aphids, as described earlier. cDNAs were generated from isolated RNA samples using Superscript II reverse transcriptase (Invitrogen) and p(dN_6) random hexamers (Roche). Control reactions without reverse transcriptase (−RTase) were included in all assays. The abundance of *ApAQP2* transcripts was determined by Q-PCR with an ABI Prism 7900 Sequence Detection System (Applied Biosystems), using the comparative Ct method. The reaction mixtures contained 1× Power SYBR Green MasterMix (Applied Biosystems), 0.1 μM gene-specific primers (Supplemental Table 1), and 2 μL cDNA template in a final reaction volume of 25 μL. Thermal cycling conditions were 2 min at 50 °C, 10 min at 95 °C followed by 40 cycles of 15 s at 95 °C and 1 min at 60 °C. The assays included a dissociation curve (95 °C for 30 s followed by a 60–95 C temperature ramp in increments of 0.5 °C for 1 min each), which confirmed that all detectable fluorescence was derived from specific products. All experimental samples were assayed in triplicate, with template free and -RTase controls. The relative expression of ApAQP2 was assessed by determining the threshold cycle (Ct), and each transcript was normalized to the expression of the ribosomal protein L32 transcript (*RPL32*) standardized to the expression level of the transcript in the whole aphid body.

2.3. Expression and functional analyses of ApAQP2 in Xenopus laevis oocytes

The *ApAQP2* open reading frame (ORF) was amplified by PCR from the pCR Blunt-II TOPO vector (Invitrogen) with ExTaq polymerase and gene-specific primers (Supplemental Table 1) containing *Bam*HI restriction sites, and was cloned into the *Bgl*II site of the pXβG-FLAG vector by the approach described in [16]. The final expression construct contained *ApAQP2* open reading frame translationally fused to an N-terminal FLAG epitope tag.

Capped cRNA was produced from *Xba*I-linearized Flag-*ApAQP2*/pXβG and soybean *nodulin 26*/pXβG constructs by using the AmpliCap-Max T3 High Yield Message Maker Kit (Epicentre Technologies). Stage V and VI *X. laevis* oocytes were microinjected and cultured as described previously [9,16]. Experimental oocytes were injected with 46 nL of 1 ng/nL of each test cRNA, and negative control oocytes were injected with an equivalent volume of sterile RNase-free water. Expression of ApAQP2 and nodulin 26 proteins in oocytes was quantified by Western blot analysis of oocyte lysates using an anti-FLAG epitope monoclonal antibody (Stratagene) for ApAQP2-injected oocytes, and an antibody against soybean nodulin 26 for nodulin 26-injected oocytes as previously described [16–18]. The osmotic water permeability (P_f) of *Xenopus* oocytes was determined as previously described [16,18]. Oocytes were placed in a 15 °C bath solution containing diluted (30%) frog Ringers solution [16–18], and serial images of the oocytes were collected as they began to swell in response to hypoosmotic challenge. The rate of oocyte swelling ($d[V/V_0]/dt$) determined by video microscopy was used to calculate the osmotic permeability coefficient (P_f) using the following equation:

$$P_f = \frac{V_0/S_0(d[V/V_0]/dt)}{\left(\frac{S_{real}}{S_{sphere}}\right)V_w(\mathrm{osm_{in}}-\mathrm{osm_{out}})}$$

where V_0 is the initial volume and S_o is the initial surface area of the oocyte, osm_{in} is the osmolarity on the inside of the oocyte, osm_{out} is the osmolarity of the bathing solution, V_w is the partial molar volume of water (18 cm^3/mol), S_{real} is the actual oocyte surface area, and S_{sphere} is the theoretical oocyte surface area assuming a perfect sphere. A value of S_{real}/S_{sphere} of 9 was used in all calculations to correct for the increase in oocyte plasma membrane area resulting from the presence of folds and microvilli [18].

The effect of mercurials on transport was investigated by preincubating *ApAQP2*-injected oocytes in frog Ringers solution supplemented with 1 mM $HgCl_2$ for 5 min prior to assay. The viability of control and ApAQP2-expressing oocytes was verified by measurement of the resting oolemma membrane potentials as described in [19]. All oocytes possessed inwardly negative resting potentials between −22 and −25 mV, consistent with the reported membrane potential values of viable oocytes [20].

Xenopus oocytes expressing ApAQP2 were assayed for solute permeability by two different methods. Nonradiolabeled solute uptake assays were performed by measuring solute-induced swelling under isoosmotic conditions as previously described [16]. The oocytes were placed into a bath solution containing isoosmotic frog Ringers solution with the NaCl component replaced with 200 mM test solute. In this assay, solute uptake results in an inwardly-directed osmotic gradient that leads to water uptake and oocyte swelling. Solute uptake is reported as an oocyte swelling rate [$d(V/V_0)/dt$], determined by video microscopy.

Radiolabeled glycerol and mannitol uptake assays were performed by a method modified from [16]. Twelve oocytes were incubated in 150 μL of 5 mM HEPES NaOH pH 7.6, 86 mM NaCl, 2 mM KCl, 5 mM $MgCl_2$, 0.6 mM $CaCl_2$ supplemented with 20 mM ^{3}H-glycerol (14 μCi/mL) for glycerol uptake assays, or with 20 mM ^{14}C-mannitol (1.4 μCi/mL) for mannitol uptake assays. All radioisotopic assays were conducted for 10 min at 25 °C. The oocytes were rinsed four times

with 10 mL of radioisotope-free, ice-cold assay buffer, and separated into three groups of four oocytes in scintillation vials. The oocytes were lysed in 300 μL of 10% (w/v) SDS and isotope uptake was quantified by as in [16].

2.4. Molecular modeling

A homology model of ApAQP2 was constructed using the Molecular Operating Environment software (MOE2009.10; Chemical Computing Group, Montreal, Canada). The crystal structure of human AQP4 (pdb 3GD8 [21]) was chosen as the structural template because of its high resolution (1.8 Å) and the high level of amino acid sequence identity (33%) to the ApAQP2 amino acid sequence within transmembrane helical regions that form the transport pore. Similar results were obtained with other aquaporin structural templates (data not shown). ApAQP2 was aligned with the AQP4 template by using the MOE structural alignment tool. Homology models were constructed by using the homology modeling facility in MOE and the CHARMM27 force field. An ensemble of ten possible structures for ApAQP2 was generated and ranked by packing score. The model with the most favorable packing score (3.0373) was energy-minimized using the CHARMM27 force field and distance-dependent dielectric down to an energy gradient of 10^{-5} kcal/mol/Å^2. The stereochemical quality of the final model was assessed by using Ramachandran plot analysis and the Protein Report structural analysis function in the MOE Protein Structure Evaluation package as previously described [22] to determine disallowed bond angles, bond lengths, and side-chain rotamers. The final ApAQP2 model possessed two residue outliers in the Ramachandran plot. Both residues were found in the unconserved and unstructured region of the C loop which does not contribute to the formation of the aquaporin fold and transport pore. Pore diameters of homology models were calculated using the HOLE2.0 program [23] of the energy minimized homology model structure using the simple.rad van der Waals radius file.

2.5. Phylogenetic analysis

The protein sequences were aligned using ClustalX [24], checked manually in BioEdit 5.0.9 [25] and truncated to remove the predicted N and C terminal cytoplasmic tails, which could not be aligned with confidence. Bayesian inference (BI) and maximum likelihood (ML) analyses were conducted, following selection of the WAG + Γ + I model [26], using the optimal instantaneous rate matrix estimated in ProtTest 1.4 [27]. The gamma shape parameter α and proportion of invariable sites were estimated as part of the analysis. Eight rate categories were used for rate heterogeneity estimation. The BI analysis was run in MrBayes 3.1.2 [28]. Data were processed in two partitions. In the first partition, gaps in the amino-acid alignment were treated as missing data, and phylogenetic information present in the gaps was then encoded as a binary dataset of the same length and analysed as the second partition. Each Metropolis-coupled Markov chain Monte Carlo (MCMC) run was three million generations long, sampled every 1000th step, and the first 30% of sampled trees were discarded as burn-in. The runs were considered converged when average standard deviation of split frequencies was less than 0.01 and potential scale reduction factor approached 1.0. The observed 95% confidence interval of the tree length did not include ML tree length estimate and branch lengths were subsequently calculated using ML approach on fixed BI tree topology in RAxML 7.2.6 [29]. The ML analysis was performed in PhyML 3.0 [30] with BIONJ selected as the starting tree and NNI tree topology search algorithm. Alignment gaps were treated as missing data and bootstrap support was estimated from 100 parametric replicates. Midpoint rooting was used in the analyses. Posterior probabilities ≥ 0.95 in BI analyses and bootstrap support ≥ 70% in ML analyses were designated significant. Analyses were conducted on computational clusters at the Institute of Vertebrate Biology AS CR, Brno, Czech Republic and Bioportal, University of Oslo, Norway.

3. Results

3.1. Sequence and phylogenetic analysis of pea aphid aquaporin genes

Detailed examination of the pea aphid genome [Acyr 2.0 primary assembly (NCBI)] resulted in the identification of three gene loci encoding proteins with aquaporin/MIP homology. The first locus corresponds to the aquaporin gene *ApAQP1* (ACYPI006387, Gene ID: 100165436 on scaffold NW_003383975). *ApAQP1* expression generates two isoforms by alternative splicing of the 5′ exons encoding proteins of 272 and 250 amino acids, respectively. The dominant isoform expressed in the gut is isoform-1 (272 amino acids, NP_001139376.1), which has been demonstrated previously to function as a Hg-sensitive water-selective aquaporin involved in water cycling and osmoregulation [9]. A second locus (Gene ID: 100573582, scaffold NW_003384268) encodes an aquaporin-like protein in which the amino-terminal 212 amino acids is identical to ApAQP1 isoform-1. However this protein diverges from the canonical aquaporin sequence with the 60 carboxyl terminal residues of ApAQP1 replaced by 41 amino acids that lack aquaporin/MIP pore forming determinants (the second NPA box and the LE_1 and LE_2 residues of the ar/R selectivity filter), suggesting that it is incapable of forming a functional aquaporin/MIP channel.

Another aquaporin-like gene (ACYPI009194, Gene ID: 100168499, scaffold NW_003383567), which we refer to as ApAQP2, encodes two transcriptional variants that encode separate protein isoforms: isoform-1 (308 amino acids [Fig. 1]) and isoform-2 (275 amino acids), which lacks the 33 N-terminal amino acids in isoform-1. Published EST data indicate that both isoforms of this gene are expressed in aphids [11]. This study focused on isoform-1, building on evidence that it is expressed in isolated bacteriocytes [12]. It contains all of the sequence and structural characteristics of the MIP channel family, including six predicted transmembrane domains and two canonical NPA boxes, all components of the prototypical "hourglass fold" of the aquaporin/MIP superfamily (Fig. 1).

The phylogenetic position of the ApAQP2 protein sequence was investigated by comparison with 60 animal MIP sequences using Bayesian (BI) and maximum likelihood (MI) methods. The tree topologies obtained with the two methods were very similar, assigning all but one of the sequences to one of four significantly supported clades, termed A–D (Fig. 2 shows the BI tree). Clade A contained a number of functionally characterized water-specific aquaporins from both mammals and insects. This clade contains the three insect subfamilies of aquaporins, BIBs, DRIPs, and PRIPs [2]. The previously characterized *A. pisum* aquaporin ApAQP1 [9] clusters within the PRIP subfamily. Clade C represents a group of functionally characterized mammalian-like aquaglyceroporins that includes AQP3, AQP7, and AQP9 [31–33]. Clade D MIPs are similar to human AQP11 and AQP12, which have been termed "superaquaporins", and yet have poorly defined transport properties [34].

ApAQP2 was assigned to clade B, which differs from the other clades in that all members are of insect origin. Since ApAQP2 belongs to an insect-specific clade, distinct from other MIP family members with canonical aquaporin or aquaglyceroporin activities, structural and functional analyses were undertaken to model the putative transport pore and determine its transport selectivity.

3.2. Molecular modeling of ApAQP2

The structural properties of the putative transport pore of ApAQP2 were investigated by molecular modeling by using the Molecular Operating Environment (MOE) software. A homology model of the ApAQP2 protein was constructed utilizing the 1.8 Å X-ray structure (PDB ID 3GD8 [21]) of human AQP4 as a modeling template (Fig. 3).

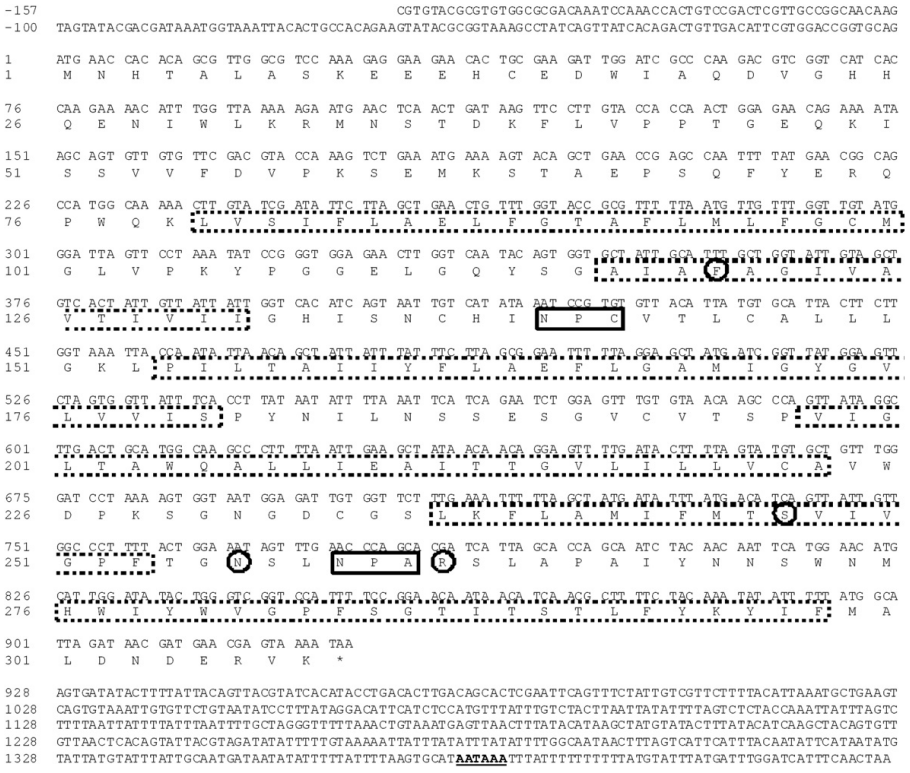


Fig. 1. Amino acid sequence of ApAQP2: The full length cDNA sequence of ApAQP2 showing the open reading frame and deduced amino acid sequence. The predicted transmembrane domains are highlighted in dashed boxes, the NPC and NPA motifs in solid boxes, and the residues of the proposed ar/R selectivity filter in solid circles. The position of a potential polyadenylation site in the 3′-untranslated region is underlined.

The structural alignment of the ApAQP2 homology model with the modeling template was excellent with an average carbon backbone root mean square deviation (rmsd) of 0.762 Å. All elements of the aquaporin fold were apparent which allowed modeling of the predicted pore determinant regions.

The selectivity-determining ar/R region consists of a tetrad of residues from transmembrane α-helices 2 (H2) and 5 (H5), as well as two residues from the 2nd NPA helical loop E (LE_1 and LE_2) [22]. In the human AQP4 X-ray structure, these residues are Phe 77, His 201, Ala 210, and Arg 216, respectively. This collection of ar/R residues, particularly a conserved His at the H5 position, is characteristic of water-selective aquaporins. In the AQP4 ar/R structure, water is bound by four hydrogen bonds collectively contributed by the sides chains of His 201, Arg 216, as well as by the backbone carbonyl of Ala210 [21]. These residues provide the necessary structural features to facilitate rapid water transport while the small ar/R diameter excludes larger solutes, resulting in a high conductance water-specific channel.

The ApAQP2 homology model suggests that the Phe residue at H2 and the Arg residue at LE_2 are conserved with respect to human AQP4 (Fig. 3B). Comparison of the predicted ar/R residues of Clade B MIPs suggests that these two aquaporin-like ar/R residues are conserved in other subfamily members as well (Table 1). However, ApAQP2 contains a Ser substitution at H5, which is a highly conserved His in AQP4 and other water-specific aquaporins. The presence of a Ser, Cys or Ala in the place of His is a characteristic feature of the Clade B insect MIPs (Table 1). At the LE_1 position, ApAQP2 possesses an unusual Asn residue, whereas water-selective aquaporins and other Clade B MIPs possess a smaller amino acid, typically Ala, Gly or Cys (Table 1). Modeling of the predicted ApAQP2 pore diameter using the HOLE program suggests that the ar/R region still forms a size constriction, albeit with a larger pore diameter compared to AQP4 (Fig. 3C). Overall, the novel ar/R region of ApAQP2 is atypical of water-specific aquaporins and mammalian aquaglyceroporins, suggesting that this protein may exhibit different functional properties and substrate selectivity compared to these well-characterized MIP channels.

3.3. Functional analysis of ApAQP2 transport

To investigate the functional properties of ApAQP2, the protein was expressed in *X. laevis* oocytes and subjected to water and solute transport analyses. Assays of ApAQP2-injected oocytes indicated that expression of this channel increases the osmotic water permeability (P_f) of the oocyte plasma membrane 15-fold, a clear indication of aquaporin activity (Fig. 4A). Western blot analysis confirmed that this increase in P_f corresponded with the expression of the ApAQP2 channel (Fig. 4B). The effect of $HgCl_2$, a classical aquaporin inhibitor, was also investigated. Unlike many aquaporins, including the water-selective ApAQP1 [9], the P_f of ApAQP2 oocytes treated with 1 mM $HgCl_2$ was not significantly different ($p = 0.265$) from untreated controls (Fig. 4C).

Due to the unusual composition of the ApAQP2 ar/R region, a broader range of test solutes was assayed to investigate the channel substrate selectivity. Similar to the well-characterized soybean

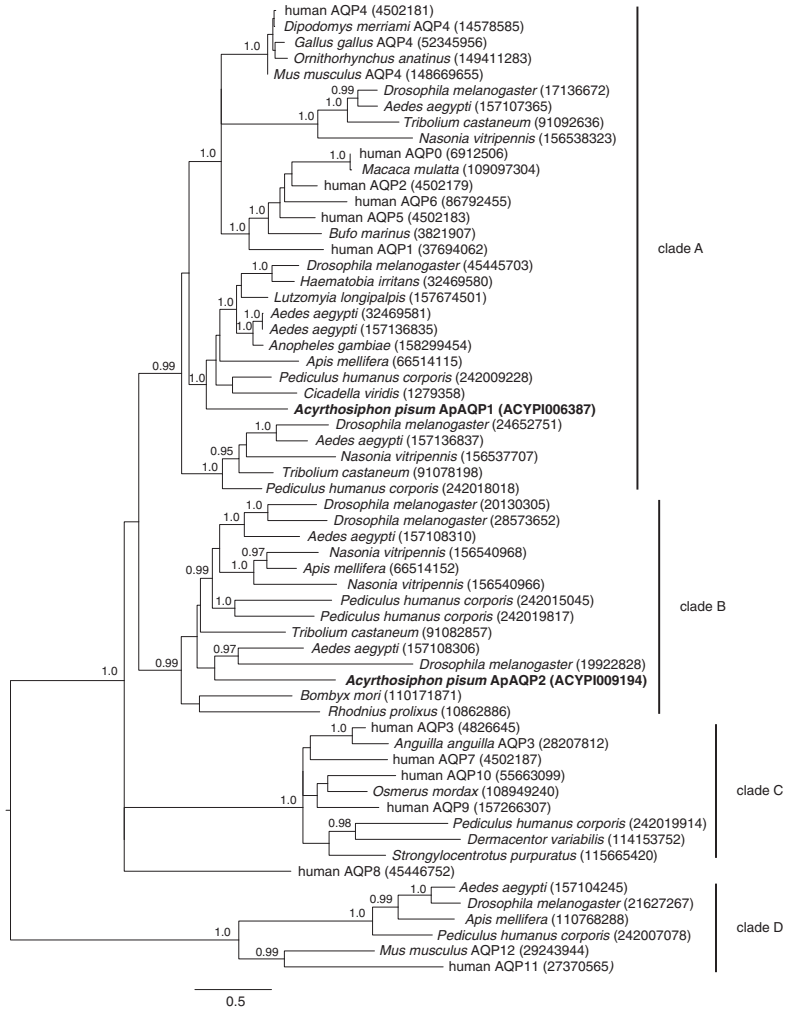


Fig. 2. Phylogenetic analysis of aphid aquaporins: A phylogenetic tree of animal aquaporin sequences was generated by Bayesian inference based on the Blosum62 rate matrix with gamma distribution of rate variation across sites. Numbers above edges denote statistically significant posterior probability (≥0.95). Genus and species names are indicated in italics with NCBI protein sequence identifiers indicated in parentheses. The sequences comprising the four phylogenetically supported clades (A–D) are labeled by vertical bars in the right margin. The sequence identifiers for ApAQP2, as well as the water-specific ApAQP1 [9] are shown in bold.

nodulin-26 aquaglyceroporin (GmNod26 in Table 1) [18,35], ApAQP2-expressing oocytes showed an increased permeability to ^{3}H-glycerol (Fig. 5A). However, ApAQP2-expressing oocytes also exhibited a 10-fold increase in permeability to the larger polyol mannitol, while the nodulin 26-expressing oocytes effectively excluded this solute (Fig. 5B).

Oocyte solute-dependent swelling assays were also performed to test ApAQP2 permeability to a wider range of polyols and model substrates. In agreement with the data from radioisotopic uptake assays, ApAQP2-expressing oocytes were permeable to mannitol as well as the epimeric six carbon alditols galacitol and sorbitol (Fig. 6). Additionally, ApAQP2-expressing oocytes were permeable to four carbon (erythritol) and five carbon (arabinitol, ribitol, xyltiol) alditols. The measured permeabilities for each solute were very similar, and the ApAQP2-expressing oocytes exhibited very little transport preference for alditol chain length or hydroxyl group stereochemistry. ApAQP2-expressing oocytes were also capable of transporting the uncharged test compounds formamide and urea, which are also common substrates for many aquaglyceroporins. However, ApAQP2-expressing oocytes were not permeable to the cyclic six-carbon polyol inositol. Overall, these results indicate a unusual substrate profile for ApAQP2 which forms a high-conductance, Hg-insensitive water channel that is also permeated by a broad but defined range of linear polyols.

3.4. Expression analysis of ApAQP2 in A. pisum

Previous studies have indicated that *ApAQP2* is expressed in whole aphid samples [11] and bacteriocytes [12]. To quantitate more precisely the expression pattern of *ApAQP2*, transcript levels in dissected pea aphid organs was investigated using real-time Q-PCR expression

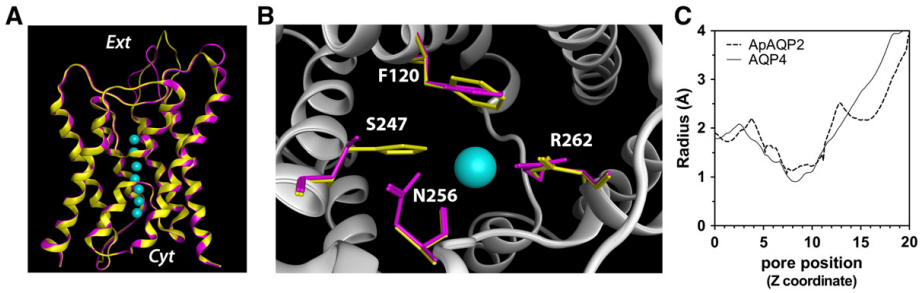


Fig. 3. Homology modeling analysis of ApAQP2 pore-forming residues: A homology model of ApAQP2 generated with the Molecular Operating Environment (MOE) software using the experimental AQP4 structure (pdb 3GD8) as a structural template. A. Superimposition of the structural model of ApAQP2 (fushia) and the AQP4 structure (yellow). The water molecules in the AQP4 pore are indicated by light blue spheres to show the position of the transport pore. The relative positions of the extracellular space (ex) and the cytosol (cyt) are indicated. B. The residues comprising the ar/R region of ApAQP2 (fushia) are shown viewed perpendicular to the transport pore axis from the extracellular face of the protein superimposed upon the corresponding residues in the AQP4 structure (yellow). Each ApAQP2 ar/R amino acid is labeled with the single letter amino acid designation as well as the residue index. The ar/R positions proceed counterclockwise as follows: H2, H5, LE_1, LE_2. The position of the transport substrate (water) bound to the ar/R region is indicated by the light blue sphere. C. Comparison of the ar/R regions of the AQP4 and ApAQP2 structures with HOLE. The calculated pore radius along the z-coordinate of the pore across the ar/R region is shown.

analysis (Fig. 7). ApAQP2 expression levels were significantly over-represented in the fat body (5.1-fold) and bacteriocytes (19.3-fold), and under-represented in the aphid gut.

Table 1
Conservation of pore-forming residues of insect clade B MIP sequences.

Protein[a]	Ar/R residues[b]				NPA motifs[c]	
	H2	H5	LE_1	LE_2	NPA1	NPA2
Selective aquaporins						
HsAQP4	F	H	A	R	NPA	NPA
ApAQP1	F	H	A	R	NPA	NPA
Aquaglyceroporins						
GmNod26	W	V	A	R	NPA	NPA
HsAQP3	F	G	Y	R	NPA	NPA
PfAQP	W	G	F	R	NLA	NPS
Glyceroporins						
EcGlpF	W	G	F	R	NPA	NPA
Clade B insect MIPs						
ApAQP2	F	S	N	R	NPC	NPA
AQP-Bom2	F	S	A	R	NPS	NPA
AaMIP1	F	A	A	R	NPA	NPA
AaMIP2	F	S	A	R	NPS	NPA
PhcMIP1	F	A	C	R	NPA	NPA
PhcMIP2	F	A	G	R	NPS	NPA
NvMIP1	F	A	C	R	NPA	NPV
NvMIP2	F	C	C	R	NPA	NPA
TcMIP	F	S	A	R	NPA	NTA
AmMIP	F	A	C	R	NPA	NPA
DmCG4019	F	A	G	R	NPA	NPA
DmCG17664	F	S	A	R	NPA	NPV
RpMIP	F	S	A	R	NPV	NPV

[a] The protein sequence name for each MIP isoform is indicated. Genus and species designations are as follows: Hs, *Homo sapiens*; Gm, *Glycine* max; Pf, *Plasmodium falciparum*; Ec, *Escherichia coli*; Ap, *Acrythosiphon pisum*; Rp, *Rhodnius prolixus*; Aa, *Aedes aegypti*;Phc, *Pediculus humanus corporis*; Nv, *Nassonia vitripennis*; Tc, *Tribolium castaneum*; Am, *Apis melifera*; Dm, *Drosophila melanogaster*. Accession number for these proteins are as follows: HsAQP4 (NP_001641), ApAQP1 (NP_001139376), GmNod26 (CAA28471), HsAQP3 (NP_004916), PfAQP (AAN35922); EcGlpF, (BAE77383); ApAQP2 (NP_001139377), AQP-Bom2 (BAE97427), AaMIP1 (XP_001650170), AaMIP2 (XP_001650168), PhcMIP1 (XP_002428189), PhcMIP2 (XP_002430355), NvMIP1 (XP_001601253), NvMIP2 (XP_001601231), TcMIP (XP_970728), AmMIP (XP_624194), DmCG4019 (NP_611813), DmCG17664 (NP_788433), and RpMIP (CAC13959).

[b] The residues comprising the helix 2 (H2), helix 5 (H5), loop E position 1 (LE_1) and loop E position 2 (LE_2) residues of the ar/R region are indicated for each sequence. A representative sequence alignment showing the position of each ar/R residue is shown in Supplementary Fig. 1.

[c] The residues comprising the N-terminal (NPA1) and C-terminal (NPA2) NPA motifs are shown for each sequence.

4. Discussion

This study has identified an aphid aquaglyceroporin gene, *ApAQP2*, that is preferentially expressed in the bacteriocytes and fat body of the pea aphid *Acyrthosiphon pisum*. *ApAQP2* is a member of an insect-specific clade (clade B MIPs) that is phylogenetically distinct from other known animal aquaporin and aquaglyceroporin sequences. In accordance with this observation, structural analysis of an ApAQP2 protein homology model indicates it contains novel substitutions within the proposed selectivity-determining ar/R region. Similar to water-selective aquaporins, Clade B MIPs contain a conserved Phe at the H2 position and the invariant Arg residue at the LE_2 position of the ar/R region. However, the Clade B ar/R possesses small neutral hydrophilic amino acids (Ser or Ala) at the H5 position that results in an overall wider and more hydrophilic selectivity filter compared to other insect and mammalian aquaporins (Table 1). ApAQP2 is unique among other clade B MIPs and most aquaporins since it contains an unusual Asn residue at the LE_1 position. This positions an additional side chain within the ar/R selectivity filter that could potentially create new hydrogen bonding contacts for transported solutes.

Functional analysis of ApAQP2 in *Xenopus* oocytes indicates that the protein possesses a multifunctional aquaglyceroporin activity capable of transporting both water and neutral polyol substrates. ApAQP2 is highly permeable to water, inducing a 15-fold increase in the P_f of the oolemma upon expression in *Xenopus* oocytes. However, unlike most aquaporins, including the water-selective aphid aquaporin ApAQP1 [9], ApAQP2 water permeability was not sensitive to the common aquaporin channel blocker $HgCl_2$. Although an unusual property, aquaporin channels that are insensitive to mercury ions have been documented [36,37]. The inability of Hg^{2+} to block these transporters could be the result of the absence a Cys residue near the transport pore.

While ApAQP2 transports glycerol in a manner similar to established animal (e.g., AQP3 [31]), higher plant (e.g., nodulin 26 [35]), protist (e.g., PfAQP [38]) and bacterial (e.g., GlpF [39]) aquaglyceroporins, it differs significantly from these proteins with regard to its polyol transport behavior. For example, the well characterized *E. coli* GlpF shows limited ability to transport longer polyol substrates. *E. coli* GlpF transports the five carbon polyol ribitol at a rate similar to the established biological substrate glycerol, but exhibits strong size and stereoselectivity preferences for other polyols, with epimers of ribitol (e.g., xylitol and arabinitol) showing low permeability, and the six carbon compounds mannitol and sorbitol showing complete impermeability [40]. Additionally, a multifunctional aquaglyceroporin isolated from *Plasmodium falciparum* (PfAQP) exhibited

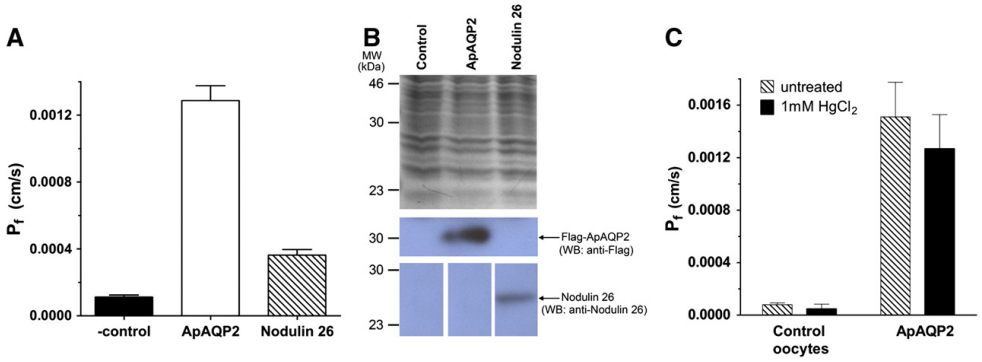


Fig. 4. Water permeability analysis of *Xenopus* oocytes expressing ApAQP2: A. *Xenopus* oocytes were injected 46 ng of *ApAQP2* cRNA (white bar) or soybean *nodulin 26* cRNA (hatched bar) and osmotic water permeability (P_f) was determined by the oocyte swelling assay. Oocytes injected with sterile RNase free-water were used as negative controls (black bar). Error bars represent SEM (n = 10 oocytes). B. Western blot analyses of oocyte lysates (20 μg total protein per lane). Top panel, Coomassie blue stained SDS-PAGE gel, middle panel, Western blot with anti-FLAG antibody, bottom panel, Western blot with anti-nodulin 26 antibody [55]. The positions of the molecular weight markers are indicated to the left of each panel. C. Negative control and ApAQP2-expressing oocytes were pre-treated with 1 mM $HgCl_2$ (solid black bars) prior to the standard water permeability assay described in Materials and methods. Untreated oocytes (hatched bars) were used as controls. Error bars represent SEM (n = 6–9 oocytes).

robust permeability to five carbon polyols with a preference for arabitol and xylitol. However, PfAQP was impermeable to the five carbon stereoisomer ribitol [38] and was also impermeable to the larger six carbon polyol mannitol [41]. In contrast, ApAQP2 shows strong permeability to a wide variety of linear polyols (C3 to C6) with little apparent preference for hydroxyl group stereochemistry, indicating a fundamental difference between in the transport pores of these classes of aquaglyceroporins.

Vertebrate aquaglyceroporins are a structurally and functionally conserved class of MIPs that cluster into a distinct phylogenetic group (Clade C) with a defined ar/R selectivity filter (Table 1). Interestingly, clade C MIPs are poorly represented in invertebrates whereas a diverse array of Clade B MIPs is represented in all insect species examined. This suggests that insects have acquired structurally and functionally distinct aquaglyceroporins compared to their vertebrate animal counterparts. It is unknown whether the broad array of polyol transport properties exhibited by ApAQP2 is shared by other clade B MIPs. Transport behavior of another clade B MIP, encoded by *AQP-Bom2* from the silkworm *Bombyx mori*, shows that it is able to transport water, glycerol and urea [42], but its selectivity and ability to transport larger polyols (e.g., mannitol or sorbitol) have not yet been assessed.

The variety of substrates transported by ApAQP2 also raises the question of the physiological function of water and polyol transport through this multifunctional channel. Based on Q-PCR expression analysis, *ApAQP2* is highly expressed in aphid bacteriocytes as well as in fat body cells. Both of these cell types are bathed in the hemolymph and actively engage in exchange of metabolites and solutes with this circulatory fluid. The osmotic pressure of the aphid hemolymph is tightly regulated within narrow limits, thereby protecting the fat body, bacteriocytes and other internal organs from the high and variable osmotic pressure of plant phloem sap ingested by the insect into the gut lumen [43,44]. One potential role for the robust aquaporin activity of ApAQP2 could be the maintenance of osmotic equilibrium between the hemolymph and bacteriocytes or fat body cells through rapid adjustments in water content.

With respect to aphid physiology, the polyol transport activity of ApAQP2 could also participate in carbohydrate metabolism and transport related to stress biology as well as the symbiosis with the *Buchera* bacteria. The accumulation of polyols including mannitol, sorbitol and erythritol in insects [45–47] has been linked to the ability of these compounds to protect proteins and membrane structures from denaturation and disruption in response to osmotic and temperature (cold or heat) stresses. In the case of aphids, both mannitol (*Aphis gossypii* [48]) and sorbitol (*Acyrthosiphon pisum* [49]) accumulate to high concentrations, particularly in response to heat stress, and may play a role in thermotolerance [49,50]. Polyols have been proposed to be synthesized from fructose precurors in the hemolymph via a ketone reductase activity isolated from whole aphids [50], although the discovery of a putative aldose reductase gene (ACYPI005685) expressed in the bacteriocytes and fat body ([51],

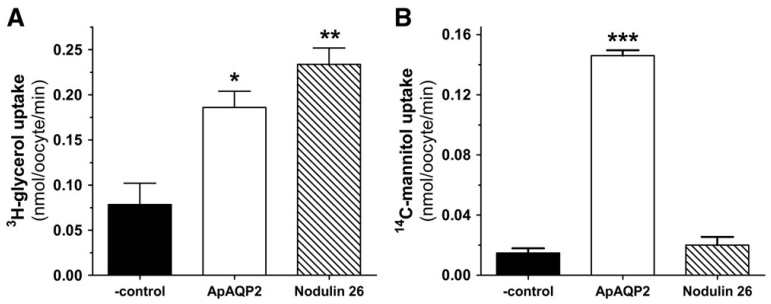


Fig. 5. Analysis of glycerol and mannitol permeabilities of ApAQP2 by radioisotopic substrate uptake: *Xenopus* oocytes injected with *ApAQP2* (white bars) or *nodulin-26* (hatched bars) cRNAs were assayed for uptake of A. ^{3}H-glycerol or B. ^{14}C-mannitol. Oocytes injected with sterile RNase-free water (black bars) were used as negative controls. Error bars represent SEM (n = 4). Asterisks over individual bars signify that transport rates are significantly higher than control oocytes (*p<0.05; **p<0.01; ***p<0.001).

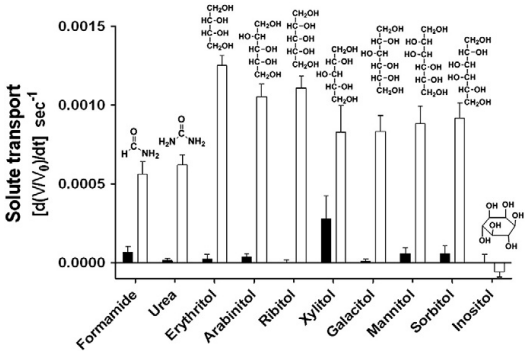


Fig. 6. ApAQP2 permeability to a series of polyol transport substrates: *Xenopus* oocytes were injected with *ApAQP2* cRNA, and were assayed for permeability to various solutes by video microscopy using an oocyte swelling assay (white bars). Oocytes injected with sterile RNase-free water (black bars) served as a negative control. Assays were conducted by immersion of oocytes in an isoosmotic Ringer's solution containing the test solutes shown above each bar. The swelling rates were measured as a change in oocyte volume [$(dV/V_0)/dt$] due to the uptake of each test solute followed by the osmotically-driven uptake of water. Error bars represent SEM (n = 10–24 oocytes).

Douglas, unpublished data) suggests additional potential sites of polyol biosynthesis. Since polyols are largely impermeable to lipid bilayers, the presence of a mannitol or a sorbitol facilitator such as ApAQP2 may be essential for the transport and partitioning of these critical protective polyols within bacteriocytes and fat body cells.

With respect to the bacteriocytes, ApAQP2 may have a symbiotic function. Most of the bacteriocyte cytoplasm is occupied by *Buchnera* cells, each of which is enclosed individually in a membrane of aphid origin ("the symbiosome membrane"). All metabolites required by the *Buchnera* and products of *Buchnera* metabolism are, of necessity, transported across the symbiosome membrane. The *Buchnera* genome is much reduced (0.64 Mb) and of small gene content (620 genes) [52], but it has retained the genes for a GlpF-like glycerol faciliator and a mannitol phosphotransferase system MtlAD, suggesting conservation of a polyol transport system. The potential significance of mannitol in the symbiosis is further underscored *in silico* analysis of the *Buchnera* metabolic network which suggests that it could represent an important carbon source for the endosymbiont [53,54]. In addition, the ability to take up and partition polyols between the aphid bacteriocyte cell cytosol, the internal space of the symbiosome, and *Buchnera* cells, may aid in osmoregulation. Based on the permeability profile of ApAQP2, and the observation that its expression is enriched in bacteriocytes, a potential transport function of these physiologically relevant polyols can be postulated. The determination of the ApAQP2 subcellular localization in bacteriocytes, including the determination of whether it is a symbiosome membrane protein, will provide valuable insight into these potential functions of this unusual aquaglyceroporin.

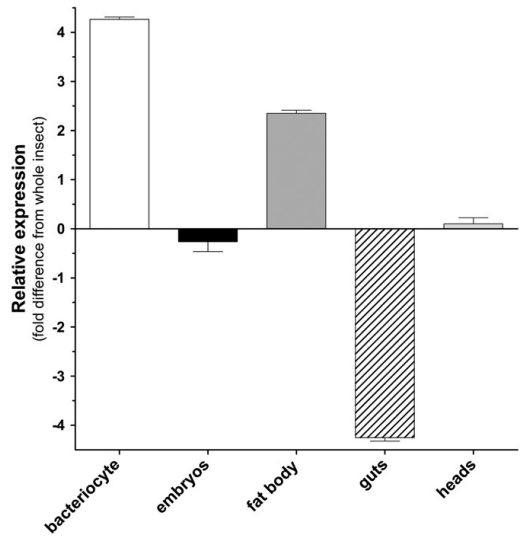


Fig. 7. Quantitative PCR analysis of *ApAQP2* gene expression in aphid tissues: Expression of the *ApAQP2* gene was normalized to the *RPL32* reference gene and standardized to the *ApAQP2* transcript expression level over the entire aphid body. Significant differences in expression (*) were determined by multiple t-tests with Bonferroni correction for five tests (critical probability $\alpha = 0.01$).

In summary, this study shows that the pea aphid aquaporin ApAQP2 is a multifunctional MIP that exhibits unprecedented substrate selectivity to both water and a wide range of potentially relevant linear polyols. Transport analyses suggest potential physiological functions of the protein in osmoregulation, as well as carbon nutrition of the insect and its symbiotic bacteria, and also provide a basis to investigate the transport properties of related MIPs in other insects. As a final note, it is also important to recognize that ApAQP2 is apparently expressed as two transcript variants that encode different protein isoforms that have the same core aquaporin-like pore structure but differ in the length of their cytosolic amino terminal regions. Further investigation is needed to determine the biological significance of these alternative splice variants in *Acyrthosiphon pisum* physiology.

Supplementary materials related to this article can be found online at doi:10.1016/j.bbamem.2011.11.032.

Acknowledgements

The authors acknowledge the assistance of Tian Li in the molecular modeling analyses. This study is supported in part by the National Science Foundation grant MCB-0618075, and a BBSRC Research Fellowship (BB/C520898) and the Sarkaria Institute of Insect Physiology and Toxicology.

References

[1] C. Hachez, F. Chaumont, Aquaporins: a family of highly regulated multifunctional channels, Adv. Exp. Med. Biol. 679 (2010) 1–17.
[2] E.M. Campbell, A. Ball, S. Hoppler, A.S. Bowman, Invertebrate aquaporins: a review, J. Comp. Physiol. B 178 (2008) 935–955.
[3] L.S. King, D. Kozono, P. Agre, From structure to disease: the evolving tale of aquaporin biology, Nat. Rev. Mol. Cell Biol. 5 (2004) 687–698.
[4] T. Walz, Y. Fujiyoshi, A. Engel, The AQP structure and functional implications, Handb. Exp. Pharmacol. 190 (2009) 31–56.
[5] B.L. de Groot, H. Grubmuller, Water permeation across biological membranes: mechanism and dynamics of aquaporin-1 and GlpF, Science 294 (2001) 2353–2357.
[6] L. Duchesne, J.F. Hubert, J.M. Verbavatz, D. Thomas, P.V. Pietrantonio, Mosquito (Aedes aegypti) aquaporin, present in tracheolar cells, transports water, not glycerol, and forms orthogonal arrays in Xenopus oocyte membranes, Eur. J. Biochem. 270 (2003) 422–429.
[7] M. Echevarria, R. Ramirez-Lorca, C.S. Hernandez, A. Gutierrez, S. Mendez-Ferrer, E. Gonzalez, J.J. Toledo-Aral, A.A. Ilundain, G. Whittembury, Identification of a new water channel (Rp-MIP) in the Malpighian tubules of the insect Rhodnius prolixus, Pflugers Arch. 442 (2001) 27–34.
[8] F. Le Caherec, S. Deschamps, C. Delamarche, I. Pellerin, G. Bonnec, M.T. Guillam, D. Thomas, J. Gouranton, J.F. Hubert, Molecular cloning and characterization of an insect aquaporin functional comparison with aquaporin 1, Eur. J. Biochem. 241 (1996) 707–715.
[9] A.J. Shakesby, I.S. Wallace, H.V. Isaacs, J. Pritchard, D.M. Roberts, A.E. Douglas, A water-specific aquaporin involved in aphid osmoregulation, Insect Biochem. Mol. 39 (2009) 1–10.
[10] I.A.G. Consortium, Genome sequence of the pea aphid Acyrthosiphon pisum, PLoS Biol. 8 (2010) e1000313.

[11] B. Sabater-Munoz, F. Legeai, C. Rispe, J. Bonhomme, P. Dearden, C. Dossat, A. Duclert, J.P. Gauthier, D.G. Ducray, W. Hunter, P. Dang, S. Kambhampati, D. Martinez-Torres, T. Cortes, A. Moya, A. Nakabachi, C. Philippe, N. Prunier-Leterme, Y. Rahbe, J.C. Simon, D.L. Stern, P. Wincker, D. Tagu, Large-scale gene discovery in the pea aphid Acyrthosiphon pisum (Hemiptera), Genome Biol. 7 (2006) R21.
[12] A. Nakabachi, S. Shigenobu, N. Sakazume, T. Shiraki, Y. Hayashizaki, P. Carninci, H. Ishikawa, T. Kudo, T. Fukatsu, Transcriptome analysis of the aphid bacteriocyte, the symbiotic host cell that harbors an endocellular mutualistic bacterium, Buchnera, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 5477–5482.
[13] A.E. Douglas, Phloem-sap feeding by animals: problems and solutions, J. Exp. Bot. 57 (2006) 747–754.
[14] E.A. Gunduz, A.E. Douglas, Symbiotic bacteria enable insect to use a nutritionally inadequate diet, P R Soc B 276 (2009) 987–991.
[15] N.A. Moran, J.P. McCutcheon, A. Nakabachi, Genomics and evolution of heritable bacterial symbionts, Annu. Rev. Genet. 42 (2008) 165–190.
[16] I.S. Wallace, D.M. Roberts, Distinct transport selectivity of two structural subclasses of the nodulin-like intrinsic protein family of plant aquaglyceroporin channels, Biochemistry 44 (2005) 16826–16834.
[17] J.F. Guenther, N. Chanmanivone, M.P. Galetovic, I.S. Wallace, J.A. Cobb, D.M. Roberts, Phosphorylation of soybean nodulin 26 on serine 262 enhances water permeability and is regulated developmentally and by osmotic signals, Plant Cell 15 (2003) 981–991.
[18] R.L. Rivers, R.M. Dean, G. Chandy, J.E. Hall, D.M. Roberts, M.L. Zeidel, Functional analysis of nodulin 26, an aquaporin in soybean root nodule symbiosomes, J. Biol. Chem. 272 (1997) 16256–16261.
[19] E.D. Vincill, K. Szczyglowski, D.M. Roberts, GmN70 and LjN70. Anion transporters of the symbiosome membrane of nodules with a transport preference for nitrate, Plant Physiol. 137 (2005) 1435–1444.
[20] N. Dascal, The use of Xenopus oocytes for the study of ion channels, CRC Crit. Rev. Biochem. 22 (1987) 317–387.
[21] J.D. Ho, R. Yeh, A. Sandstrom, I. Chorny, W.E. Harries, R.A. Robbins, L.J. Miercke, R.M. Stroud, Crystal structure of human aquaporin 4 at 1.8 A and its mechanism of conductance, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 7437–7442.
[22] I.S. Wallace, D.M. Roberts, Homology modeling of representative subfamilies of Arabidopsis major intrinsic proteins. Classification based on the aromatic/arginine selectivity filter, Plant Physiol. 135 (2004) 1059–1068.
[23] O.S. Smart, J.M. Goodfellow, B.A. Wallace, The pore dimensions of gramicidin A, Biophys. J. 65 (1993) 2455–2460.
[24] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins, The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res. 25 (1997) 4876–4882.
[25] T.A. Hall, BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/Nt, Nucleic Acids Symp. Ser. 41 (1999) 95–98.
[26] S. Whelan, N. Goldman, A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, Mol. Biol. Evol. 18 (2001) 691–699.
[27] F. Abascal, R. Zardoya, D. Posada, ProtTest: selection of best-fit models of protein evolution, Bioinformatics 21 (2005) 2104–2105.
[28] F. Ronquist, J.P. Huelsenbeck, MrBayes 3: Bayesian phylogenetic inference under mixed models, Bioinformatics 19 (2003) 1572–1574.
[29] A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, Bioinformatics 22 (2006) 2688–2690.
[30] S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst. Biol. 52 (2003) 696–704.
[31] M. Echevarria, E.E. Windhager, G. Frindt, Selectivity of the renal collecting duct water channel aquaporin-3, J. Biol. Chem. 271 (1996) 25079–25082.
[32] K. Ishibashi, M. Kuwahara, Y. Gu, Y. Tanaka, F. Marumo, S. Sasaki, Cloning and functional expression of a new aquaporin (AQP9) abundantly expressed in the peripheral leukocytes permeable to water and urea, but not to glycerol, Biochem. Biophys. Res. Commun. 244 (1998) 268–274.
[33] K. Ishibashi, M. Kuwahara, Y. Kageyama, A. Tohsaka, F. Marumo, S. Sasaki, Cloning and functional expression of a second new aquaporin abundantly expressed in testis, Biochem. Biophys. Res. Commun. 237 (1997) 714–718.
[34] K. Ishibashi, S. Hara, S. Kondo, Aquaporin water channels in mammals, Clin. Exp. Nephrol. 13 (2009) 107–117.
[35] R.M. Dean, R.L. Rivers, M.L. Zeidel, D.M. Roberts, Purification and functional reconstitution of soybean nodulin 26. An aquaporin with water and glycerol transport properties, Biochemistry 38 (1999) 347–353.
[36] M.J. Daniels, T.E. Mirkov, M.J. Chrispeels, The plasma membrane of Arabidopsis thaliana contains a mercury-insensitive aquaporin that is a homolog of the tonoplast water channel protein TIP, Plant Physiol. 106 (1994) 1325–1333.
[37] H. Hasegawa, T. Ma, W. Skach, M.A. Matthay, A.S. Verkman, Molecular cloning of a mercurial-insensitive water channel expressed in selected water-transporting tissues, J. Biol. Chem. 269 (1994) 5497–5500.
[38] E. Beitz, S. Pavlovic-Djuranovic, M. Yasui, P. Agre, J.E. Schultz, Molecular dissection of water and glycerol permeability of the aquaglyceroporin from Plasmodium falciparum by mutational analysis, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 1153–1158.
[39] C. Maurel, J. Reizer, J.I. Schroeder, M.J. Chrispeels, M.H. Saier Jr., Functional characterization of the Escherichia coli glycerol facilitator, GlpF, in Xenopus oocytes, J. Biol. Chem. 269 (1994) 11869–11872.
[40] D. Fu, A. Libson, L.J. Miercke, C. Weitzman, P. Nollert, J. Krucinski, R.M. Stroud, Structure of a glycerol-conducting channel and the basis for its selectivity, Science 290 (2000) 481–486.
[41] T. Zeuthen, B. Wu, S. Pavlovic-Djuranovic, L.M. Holm, N.L. Uzcategui, M. Duszenko, J.F. Kun, J.E. Schultz, E. Beitz, Ammonia permeability of the aquaglyceroporins from Plasmodium falciparum, Toxoplasma gondii and Trypansoma brucei, Mol. Microbiol. 61 (2006) 1598–1608.
[42] N. Kataoka, S. Miyake, M. Azuma, Aquaporin and aquaglyceroporin in silkworms, differently expressed in the hindgut and midgut of Bombyx mori, Insect Mol. Biol. 18 (2009) 303–314.
[43] A.E. Douglas, The nutritional physiology of aphids, Adv. Insect. Physiol. 31 (2003) 73–140.
[44] T.L. Wilkinson, D.A. Ashford, J. Pritchard, A.E. Douglas, Honeydew sugars and osmoregulation in the pea aphid Acyrthosiphon pisum, J. Exp. Biol. 200 (1997) 2137–2143.
[45] L. Lalouette, V. Kostal, H. Colinet, D. Gagneul, D. Renault, Cold exposure and associated metabolic changes in adult tropical beetles exposed to fluctuating thermal regimes, FEBS J. 274 (2007) 1759–1767.
[46] M.R. Michaud, J.B. Benoit, G. Lopez-Martinez, M.A. Elnitsky, R.E. Lee, D.L. Denlinger, Metabolomics reveals unique and shared metabolic changes in response to heat shock, freezing and desiccation in the Antarctic midge, Belgica antarctica, J. Insect Physiol. 54 (2008) 645–655.
[47] D. Doucet, V.K. Walker, W. Qin, The bugs that came in from the cold: molecular adaptations to low temperatures in insects, Cell. Mol. Life Sci. 66 (2009) 1404–1418.
[48] D.L. Hendrix, M.E. Salvucci, Polyol metabolism in homopterans at high temperatures: accumulation of mannitol in aphids (Aphididae: Homoptera) and sorbitol in whiteflies (Aleyrodidae: Homoptera), Comp. Biochem. Phys. A 120 (1998) 487–494.
[49] G. Burke, O. Fiehn, N. Moran, Effects of facultative symbionts and heat stress on the metabolome of pea aphids, ISME J. 4 (2010) 242–252.
[50] M.E. Salvucci, G.R. Wolfe, D.L. Hendrix, Purification and properties of an unusual NADPH-dependent ketose reductase from the silverleaf whitefly, Insect Biochem. Mol. 28 (1998) 357–363.
[51] A. Poliakov, C.W. Russell, L. Ponnala, H.J. Hoops, Q. Sun, A.E. Douglas, K.J. van Wijk, Large-scale label-free quantitative proteomics of the pea aphid-Buchnera symbiosis, Mol. Cell. Proteomics 10 (2011) M110.007039.
[52] S. Shigenobu, H. Watanabe, M. Hattori, Y. Sakaki, H. Ishikawa, Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp APS, Nature 407 (2000) 81–86.
[53] G.H. Thomas, J. Zucker, S.J. Macdonald, A. Sorokin, I. Goryanin, A.E. Douglas, A fragile metabolic network adapted for cooperation in the symbiotic bacterium Buchnera aphidicola, BMC Syst. Biol. 3 (2009).
[54] S.J. Macdonald, G.H. Thomas, A.E. Douglas, Genetic and metabolic determinants of nutritional phenotype in an insect-bacterial symbiosis, Mol. Ecol. 20 (2011) 2073–2084.
[55] Y. Zhang, D.M. Roberts, Expression of soybean nodulin 26 in transgenic tobacco. targeting to the vacuolar membrane and effects on floral and seed development, Mol. Biol. Cell 6 (1995) 109–117.

Paper 2.2.1

Martínková N., McDonald R. A., Searle J. B. 2007. Stoats (*Mustela erminea*) provide evidence of natural overland colonisation of Ireland. *Proceedings of the Royal Society B-Biological Series* 274: 1387-1393.

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2007) **274**, 1387–1393
doi:10.1098/rspb.2007.0334
Published online 3 April 2007

Stoats (*Mustela erminea*) provide evidence of natural overland colonization of Ireland

Natália Martínková[1,2], Robbie A. McDonald[3] and Jeremy B. Searle[1,*]

[1]*Department of Biology (area 2), University of York, PO Box 373, York YO10 5YW, UK*
[2]*Institute of Vertebrate Biology, Academy of Science of the Czech Republic, Studenec 122, 675 02 Konešin, Czech Republic*
[3]*Quercus, School of Biological Sciences, Queen's University Belfast, 97 Lisburn Road, Belfast BT9 7BL, UK*

The current Irish biota has controversial origins. Ireland was largely covered by ice at the Last Glacial Maximum (LGM) and may not have had land connections to continental Europe and Britain thereafter. Given the potential difficulty for terrestrial species to colonize Ireland except by human introduction, we investigated the stoat (*Mustela erminea*) as a possible cold-tolerant model species for natural colonization of Ireland at the LGM itself. The stoat currently lives in Ireland and Britain and across much of the Holarctic region including the high Arctic. We studied mitochondrial DNA variation (1771 bp) over the whole geographical range of the stoat (186 individuals and 142 localities), but with particular emphasis on the British Isles and continental Europe. Irish stoats showed considerably greater nucleotide and haplotype diversity than those in Britain. Bayesian dating is consistent with an LGM colonization of Ireland and suggests that Britain was colonized later. This later colonization probably reflects a replacement event, which can explain why Irish and British stoats belong to different mitochondrial lineages as well as different morphologically defined subspecies. The molecular data strongly indicate that stoats colonized Ireland naturally and that their genetic variability reflects accumulation of mutations during a population expansion on the island.

Keywords: cytochrome *b*; D-loop; last glacial maximum; mitochondrial DNA; phylogeography

1. INTRODUCTION

Although Britain and Ireland are neighbouring islands, the difference in their current biota is astonishing. Britain has a fauna and flora that is somewhat restricted but broadly similar to the nearby areas of continental Europe. By contrast, Ireland is very species poor, but has some unexpected forms among those present. Most noteworthy are those that otherwise occur in southwest Europe, collectively known as the 'Lusitanian element', including species such as the Kerry slug (*Geomalacus maculosus*) and Mackay's Heath (*Erica mackaiana*; Corbet 1961). Moore (1987) articulated these unusual features of the biogeography of Ireland as 'the Irish question', recognizing that there is a need to understand how the Irish fauna and flora could have developed in this way.

The unusual nature of the Irish biota, and particularly its distinctiveness from the British, is particularly seen clearly in mammals (Yalden 1999). For instance, in mainland Britain, non-commensal small mammals are represented by three species of vole, three species of shrew, three species of mouse and two small carnivores, all of which apparently colonized naturally at around the end of the last glaciation. Only four of these eleven species are found in Ireland, three with a long-standing presence (the pygmy shrew (*Sorex minutus*), the wood mouse (*Apodemus sylvaticus*) and the stoat (*Mustela erminea*)) and the fourth (the bank vole; *Clethrionomys glareolus*) is a twentieth century introduction. Furthermore, phylogeographic studies have shown that the pygmy shrews in Ireland have greater molecular similarity with populations in northern Iberia than those in Britain (Mascheretti *et al.* 2003). A comparable result has been obtained with the pine marten (*Martes martes*; Davison *et al.* 2001), providing further examples of the Lusitanian element, as applied to genetic forms within species of mammal. Mountain hares (*Lepus timidus*) in Ireland also have more molecular similarity to populations in continental Europe than to those in Britain (Hamill *et al.* 2006).

Few of the species currently present in Britain and Ireland would have been present at the Last Glacial Maximum (LGM; 19–23 kyr ago), when both landmasses were predominantly covered by ice (Andersen & Borns 1997; Mix *et al.* 2001), so colonization has been largely since that time. This colonization to Britain was over a land bridge to continental Europe which persisted until 7500 years ago at the very latest (Yalden 1999). The restricted fauna of Ireland clearly suggests that there was not such a long-lasting land connection between Ireland and either Britain and/or continental Europe. Indeed, while there would have been a land connection between Ireland, Britain and continental Europe at the LGM (owing to lowered sea levels; Andersen & Borns 1997), there are doubts whether Ireland was connected to either landmass at all thereafter (Lambeck & Purcell 2001). If there was no land bridge, that would explain the restricted fauna and flora in Ireland (Stuart & Van Wijngaarden-Bakker 1985). It would also suggest that the organisms which do live there have, to a large extent, been brought by

* Author for correspondence (jbs3@york.ac.uk).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2007.0334 or via http://www.journals.royalsoc.ac.uk.

Received 22 December 2006
Accepted 9 March 2007

Table 1. Primers and PCR conditions. (The new primers designed for this study include the indication of the position of their 3′ end within the dog complete mitochondrial sequence (Kim *et al.* 1998). T_a, annealing temperature; Cb, cytochrome *b*; D, partial D-loop (and flanking tRNA sequences); F, forward; R, reverse.)

primer name	sequence (5′–3′)	T_a	reference
Cb-M1-F	CTC ACA TGG AAT CTA ACC ATG AC	56	Kurose *et al.* (2000)
Cb-L14474-F	CTT TAT CTG CCT ATT CCT ACA CG	56	this study
Cb-H14549-R	CTA TGA ATG CAG TTG CTA TGA C	55	this study
Cb-Mustela07-F	TTC ATC ATT TCA GCA CTA GCA GCA GTC	55	Fleming & Cook (2002)
Cb-Mustela06-R	GTG GAA TGG GAT TTT GTC AGA GTC GGA	55	Fleming & Cook (2002)
Cb-L14897-F	GCC CTA TTC CTT ATT CTA ACA C	55	this study
Cb-H14939-R	GGC TGG GAT ATA GTT GTC TGG	56	this study
D-Cbz-F	ATG AAT TGG AGG ACA ACC AGT	55	Kurose *et al.* (1999)
Cb-MR1-R	TCT TCC TTG AGT CTT AGG GAG	56	Kurose *et al.* (2000)
D-L15454-F	CTG ACA TTC TAA CTA AAC TAT TCC	55	this study
D-H15777-R	TGA AGT AAG AAC CAG ATG CCA G	55	this study
D-MER-R	CGC GGG TGG TGT ATA AAT AT	55	Kurose *et al.* (1999)

humans who first appeared in Ireland in the Mesolithic (Woodman *et al.* 1997). The Lusitanian element would therefore suggest a particular importance of cultural links in early human history between Ireland and southwest Europe in introducing species to Ireland (Corbet 1961).

The Irish question can be fully answered only by the consideration of the colonization history of a wide range of species of animals and plants. However, given the uncertainty surrounding natural overland colonization, it is worthwhile to study the best candidates for this and establish that it occurred at all. From the palaeontological record, two such candidates for natural colonization are the mountain hare and the stoat. Single fossil specimens of the mountain hare and the stoat have been dated to the Late Glacial period, i.e. after the LGM but before the known occurrence of Mesolithic people in Ireland (Woodman *et al.* 1997; McCormick 1999). Clearly, if these two species occurred in Ireland before the records of people, it suggests natural colonization. Both these cases underline the uncertainty of the presence of a land bridge to Ireland after the LGM. Given their cold tolerance, the stoat and mountain hare are species that may have been able to survive LGM conditions in Ireland (Stuart & Van Wijngaarden-Bakker 1985). If these species did manage to colonize Ireland during the LGM, then the speculation about whether or not there was a land bridge subsequent to the LGM is irrelevant to their colonization, i.e. they would already have been there. Indeed, when we suggest 'colonization' at the LGM, it is also possible that the species were already present in Ireland at this time and the colonization was an even earlier event.

While single radiocarbon-dated records of the two species are encouraging, further evidence of natural colonization is desirable. Here, we adopt a complementary phylogeographic approach, i.e. we examine whether the molecular variation in modern specimens is consistent with natural colonization, as opposed to human introduction. Our study is based on the stoat because we believe it to be a particularly good model, and unlike the mountain hare (Hamill *et al.* 2006), it has not been subjected to any previous phylogeographic analysis relating to Ireland (Kurose *et al.* 2005).

Stoats are found over a very wide range of temperature conditions from warm temperate to arctic (King 1991). They currently occur in the high Arctic of Greenland and Canada, feeding on lemmings (Gilg *et al.* 2003), and at the LGM, it is known from fossils that lemmings survived in Ireland (Woodman *et al.* 1997), providing a potential food supply. Although fossils of stoats have yet to be found from the LGM in Ireland, the fossil record in continental Europe indicates their probable presence because the species is one of the glacial faunal elements, occurring in assemblages together with mammoth, woolly rhinoceros, spotted hyena and reindeer (Sommer & Benecke 2004). A stoat vertebra has also been found in a deposit believed to be *ca* 15 kyr old on the island of Andøya off northern Norway (Fjellberg 1978), at a time when Fennoscandia still had substantial ice cover (Andersen & Borns 1997).

One of the great advantages of the stoat for phylogeographic analysis is its wide and continuous contemporary distribution in Ireland, Britain and continental Europe. Not only it is thermally adaptable but the species is also generalized and common; in Britain, for instance, the stoat is the most abundant wild carnivore (Harris *et al.* 1995). Although there have been some recent introductions of the stoat to Shetland and New Zealand for biological control (King 1991), there is no reason to believe there have been accidental or deliberate introductions of the stoat to Britain or Ireland, or movements around continental Europe. Therefore, there is every expectation that the stoats present during the Late Glacial (based on the radiocarbon-dated specimens) survived in Ireland through to the present-day without any extinctions or introductions. Here, we examine whether the molecular data are consistent with these expectations of natural colonization based on a comparison of Irish and British stoats within the context of the European and Holarctic ranges of the species.

2. MATERIAL AND METHODS

(a) *Sample collection, processing and sequencing*

A total of 197 tissue and skin samples collected from stoats from 153 localities in Eurasia and Greenland successfully yielded sequences. Details of these samples, which largely came from museum skin collections, are given in the electronic supplementary material. DNA isolation, PCR amplification and sequencing were described in detail previously (Martínková & Searle 2006). The full set of primers used for amplification and sequencing of mitochondrial genes for cytochrome *b* (*cyt b*), Thr-tRNA, Pro-tRNA and partial D-loop are listed in the table 1. We

took particular care to ensure authenticity of sequences from the museum samples using methods derived from laboratory protocols for handling ancient DNA (Martínková & Searle 2006).

(**b**) ***Data analysis***

Chromatogram contigs were assembled in SEQUENCHER v. 4.2 (Gene Codes) and sequences were aligned manually with BIOEDIT v. 7.0.5.2 (Hall 1999). Haplotypes of mitochondrial (mt) DNA sequences of stoats were identified by DAMBE v. 4.2.13 (Xia & Xie 2001), and haplotype and nucleotide diversity and total and net divergence were calculated with DNASP v. 4.00.5 (Rozas *et al.* 2003). For network construction, the 49 *cyt b* haplotypes that were obtained from our 197 specimens were combined with published haplotypes from Russia, Japan and North America (14 GenBank sequences, 12 new haplotypes and 13 new localities). From the 197 stoats, we were able to obtain concatenated mtDNA sequences from 186 individuals, yielding 76 haplotypes. Median-joining networks were constructed in NETWORK v. 4.1.1.0 (Bandelt *et al.* 1999) using either the complete *cyt b* sequences or the concatenated sequence of the *cyt b*, Thr-tRNA, Pro-tRNA and the partial D-loop excluding an ambiguous part of the T_nC_n stretch. The sequences reported in this paper have been deposited in the GenBank database (Accession numbers: EF088939–EF089135).

(**c**) ***Analysis of population size changes***

Mismatch distributions were constructed from the stoat sequence data with ARLEQUIN v. 3.00 (Excoffier *et al.* 2005) and compared with those expected under the sudden population expansion model (Rogers & Harpending 1992) using goodness-of-fit statistics based on the sum of square deviations from 10 000 bootstrap replicates (Schneider & Excoffier 1999). The population expansion hypothesis was further tested with ARLEQUIN by Fu's (1997) *Fs* index of selective neutrality that is sensitive also to demographic population expansion.

(**d**) ***Estimation of mutation rate***

The mutation rate of the whole mtDNA concatenated region of the stoat was estimated under a molecular clock assumption from the Irish haplotype dataset by Bayesian coalescent analysis in BEAST v. 1.3 (Drummond & Rambaut 2006). The lower constraint of the age of the Irish population was set to 13 kyr ago, which is the age of the Late Glacial fossil found in Ireland (Woodman *et al.* 1997). This relates to a stoat from Killavullen Cave, Co. Cork that Woodman *et al.* (1997) dated as $10\,680 \pm 110$ ^{14}C years old. In terms of calendar years, this is *ca* 12 750–12 900 years old based on the IntCal04 radiocarbon calibration curve (Reimer *et al.* 2004). The upper age constraint was selected as the time when Ireland was fully covered by an ice-sheet 40 kyr ago (Bowen *et al.* 2002), an event also supported by a gap in the Irish fossil record (Woodman *et al.* 1997). The mutation rate was then calculated from a range of plausible alternative trees in the Markov chain Monte Carlo (MCMC) algorithm search in which root height fits the constraint range. The MCMC was run for 50 million steps and the first 10% were discarded as burn-in.

(**e**) ***Bayesian molecular dating***

As the mismatch distribution and Fu's (1997) *Fs* indicated that the Irish, British and continental European populations of stoat exhibited independent demographic histories, their ages were estimated using an exponential growth model. They were calculated by Bayesian coalescent analysis using MCMC in BEAST with the previously estimated mutation rate and the age constraints on the Irish population. BEAST records the age of a particular node in various trees after the MCMC convergence in which the node appears. Hence, the age estimate is not constrained to a precisely defined evolutionary history (tree topology and branch lengths), but rather explores the available tree space making the method particularly suitable for dating recent divergence events. Five independent MCMC runs were made with 10 million steps each, sampled every 1000th step. The results of all runs were then combined to obtain the ages of the most recent common ancestors of Irish, British and European stoat populations, respectively, with the first 10% discarded as burn-in.

3. RESULTS

We analysed complete *cyt b* sequences from a total of 211 stoats (166 localities), which yielded 61 haplotypes (figure 1*a* and electronic supplementary material). All British *cyt b* haplotypes (which came from stoats collected throughout the mainland and in Islay and Shetland) formed a monophyletic lineage (figure 1*a*), 85% of which shared the same haplotype (no. 39). Similarly, Irish stoats (also collected throughout the country and in the Isle of Man) formed a monophyletic but distinct lineage with a single dominant haplotype (no. 17). There is little indication of genetic structure in continental Eurasia. Some of the haplotypes identified in Alaska (nos 56–60) were similar to those from Siberia and Japan (no. 44, nos 49–55, no. 61). Fleming & Cook (2002) identified this 'Beringian' clade, and we found that it forms part of an extensive continental Eurasian lineage. Other North American haplotypes remain distinct and consistent with Fleming and Cook's 'Continental (American)' and 'British Columbia islands' clades (figure 1*a*).

We identified 76 haplotypes in 186 stoats (142 localities) from a concatenated sequence of mtDNA, comprising 1771 bp of *cyt b*, Thr-tRNA, Pro-tRNA and the partial D-loop, excluding an ambiguous part of the T_nC_n stretch (figure 1*b* and electronic supplementary material). British and Irish populations retained their distinctiveness and each remained monophyletic (figure 1*b*).

Considering the concatenated sequence, we found markedly lower nucleotide and haplotype diversity in the British population than in Irish or continental European populations. Nucleotide ($\pi \pm$s.d.) and haplotype ($h \pm$s.d.) diversity were as follows: Irish population: $\pi = 0.0026 \pm 0.0002$, $h = 0.96 \pm 0.01$, $n = 52$; British population: $\pi = 0.0007 \pm 0.0001$, $h = 0.58 \pm 0.08$, $n = 53$; and continental European population: $\pi = 0.0031 \pm 0.0002$, $h = 0.97 \pm 0.01$, $n = 74$. The total (Dxy) and net (Da) divergences between Irish and continental European populations of 0.43 and 0.15%, respectively, were similar in magnitude to the divergences between Britain and continental Europe ($Dxy = 0.45\%$ and $Da = 0.26\%$). However, the British and Irish populations were more divergent from each other ($Dxy = 0.61\%$ and $Da = 0.45\%$) than either was from the continental European population. There is no reason to believe that the greater genetic diversity of Irish stoats is the result of multiple colonization events; mismatch distributions of all populations significantly fitted the sudden population

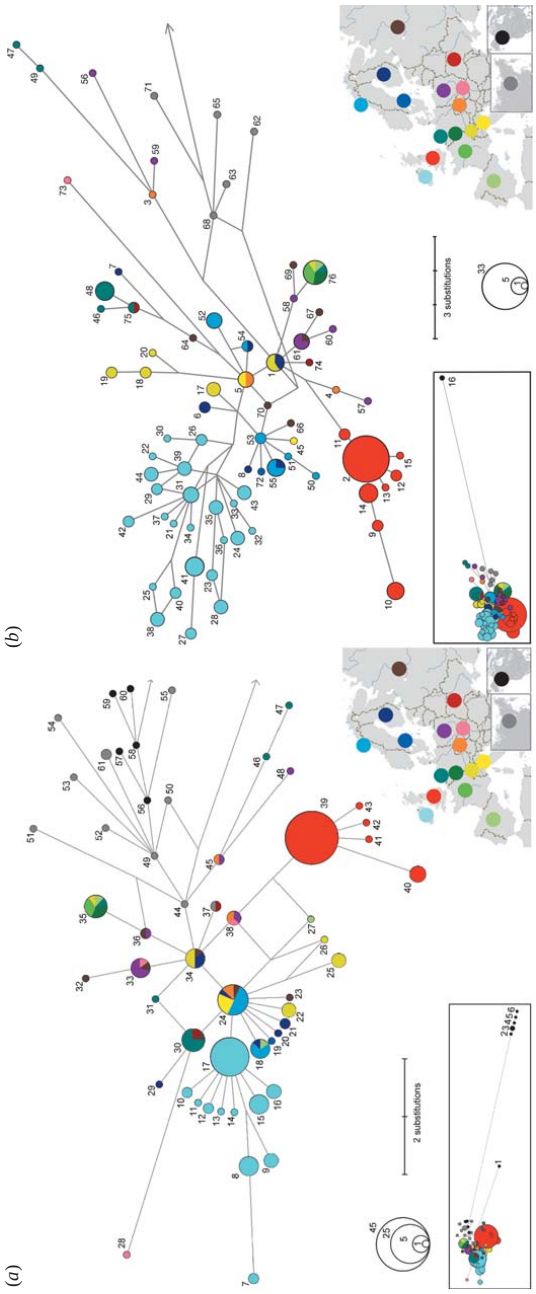


Figure 1. The relatedness and diversity of genetic lineages of the stoat, depicted by median-joining networks of sequences obtained from individuals throughout the species range. (*a*) *cyt b* sequences (1140 bp), the arrow indicates the connection to Beringian and North American *cyt b* haplotypes (inset), haplotype no. 44 is separated from haplotype no. 1 by 14 substitutions and from haplotype no. 2 by 37 substitutions, respectively. (*b*) Concatenated sequences of cyt *b*, Thr-tRNA, Pro-tRNA and partial D-loop, excluding the ambiguous part of the T_nC_n region (1771 bp), the arrow indicates the connection to the Greenland haplotype no. 16 (inset) separated from haplotype no. 68 by 50 substitutions. Colour coding is used to distinguish individual European countries, Asia and North America. Further details of the numbered haplotypes are available in the electronic supplementary material.

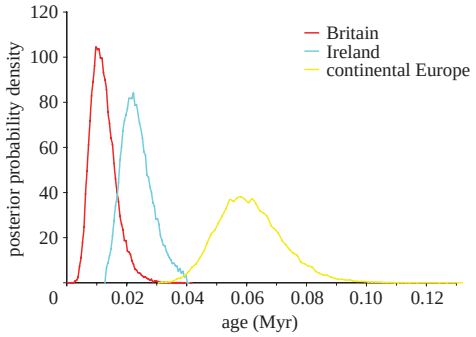


Figure 2. Estimates of the age of the Irish, British and continental European stoat populations. These are shown as posterior probability density functions estimated by the Bayesian coalescent analysis using a mutation rate of 0.07244 $bp^{-1}\,Myr^{-1}$ and age constraints of the Irish population: 13–40 kyr. Five independent runs of MCMC were performed for 10 million steps each, with the first 1 million steps discarded as burn-in.

expansion model, which describes growth of a single founder population (Ireland: sum of square deviations SSD = 0.0021, $p = 0.46$; Britain: SSD = 0.0131, $p = 0.56$; continental Europe: SSD = 0.00067, $p = 0.87$). Also consistent with population expansion, Fu's (1997) *Fs* values were negative and significantly different from zero for the Irish ($Fs = -10.188$, $p = 0$) and continental European lineages ($Fs = -28.548$, $p = 0$) but non-significant for the British lineage ($Fs = -2.207$, $p = 0.13$).

The mutation rate of the concatenated mtDNA region consistent with the age constraints of the Irish population sampled over a range of plausible alternative trees was 7.244×10^{-8} substitutions $site^{-1}\,yr^{-1}$ (14.5% per Myr). This high rate is expected when measuring within-species divergences younger than 1 Myr (Ho *et al.* 2005). Analysis with this mutation rate showed that the age of the most recent common ancestor of the Irish lineage is greater than that for the British lineage, though non-significantly so (figure 2). With 95% confidence, the most recent common ancestor of the Irish lineage is 14 270–34 050 years old (mean estimate, 23 510), whereas that of the British lineage is 4858–20 710 years old (mean estimate, 12 240).

4. DISCUSSION

The molecular diversity data that we obtained are extraordinary. Even though the landmass of Britain is 2.5 times larger than Ireland and is located closer to continental Europe, we found stoats to be far less genetically variable in Britain than Ireland. Indeed, for the concatenated *cyt b* and D-loop dataset, the haplotype and nucleotide diversities for Ireland are similar to those for continental Europe, while those in Britain are considerably lower.

These data argue strongly against the likelihood of introduction of stoats to Ireland. Generally, populations arising from introductions show lower genetic diversity than those that have been colonized naturally. The numbers of the individuals colonizing tend to be smaller and introductions tend to be more recent than natural colonizations. We have, for instance, found low levels of genetic diversity among stoats introduced to New Zealand (N. Martínková *et al.* 2007, unpublished data) and bank voles introduced to Ireland (P. Stuart *et al.* 2007, unpublished data). On the basis of the molecular data, it would seem positively perverse to suggest that stoats in Ireland are an introduction while those in Britain are a natural colonization (and, to our knowledge, there has *never* been any suggestion that stoats were introduced to Britain).

Therefore, the molecular data are supportive of the fossil evidence for a natural colonization of Ireland by stoats. We applied a Bayesian method to date the origins of the Irish and British populations of stoats from the molecular data, using the palaeontological record for Ireland to set age constraints. The Irish and British populations fit into separate monophyletic lineages. The constraints determined the 95% CI that we obtained for the age of the Irish lineage: 14 270–34 050 years, which is, of course, compatible with an LGM origin for the lineage. These dates are also compatible with colonization over a Late Glacial land bridge from continental Europe to southwest England and Ireland, which is the most probable time for such a land bridge if it existed (Lambeck & Purcell 2001).

The estimate of the age of the British lineage is 4858–20 710 years old, suggesting a more recent colonization than the Irish one. Again, given that the British colonization is natural, it is difficult to argue that an earlier colonization of Irish stoats should be an introduction. Clearly, the dating of the British lineage is consistent with the natural colonization of Britain by stoats before the flooding of the English Channel.

Some of our most interesting data relate to stoats from the Isle of Man, an island in the Irish Sea, approximately halfway between Ireland and Britain. The Isle of Man has a similar range of non-commensal small mammals as Ireland, with the same three species (wood mouse, pygmy shrew and stoat) out of the eleven found in Britain (Yalden 1999). However, the Isle of Man was connected longer to Britain than to Ireland (Innes *et al.* 2004). Hence, the Manx stoats might be expected to be more similar to the British than the Irish. In fact, in terms of mtDNA, the Manx stoats fall within the Irish lineage. Morphologically, the Irish and Manx stoats are also more similar. They are classified together as a distinct subspecies, *Mustela erminea hibernica*, on the basis of pelage characteristics (Miller 1912).

A scenario can be developed which best explains our molecular data for stoats in the British Isles. Following Stuart & Van Wijngaarden-Bakker (1985), we believe that stoats are sufficiently cold tolerant to have survived close to the British–Irish ice sheet at the LGM, and therefore were present on the exposed landmass of Ireland, having colonized at that time or earlier. Thus, we propose that the progenitors of the Irish mtDNA lineage and morphotype were cold tolerant. Consistent with that, mtDNA haplotypes from high latitudes (Scandinavia) and altitudes (Swiss and Italian Alps) are among those most closely related to modern Irish haplotypes (figure 1). We believe that these cold-tolerant stoats followed the retreating ice at the end of the LGM and spread throughout Britain, Ireland and the Isle of Man. The stoats in Ireland would have become isolated with rising sea level after the LGM and would have increased in population size from a small number of founders (hence the signal for population expansion). At a later time, the Isle of Man population of stoats would also have become separated from the British.

However, Ireland and the Isle of Man would have become islands long before Britain was itself separated from continental Europe. They perhaps became islands early enough that cold-sensitive species were totally unable to colonize them overland, hence the paucity of species in both Ireland and the Isle of Man.

There is a need to explain why British stoats are currently a different mtDNA lineage and morphotype from those found in Ireland and the Isle of Man. We believe that the continued land bridge with continental Europe is the key. This land bridge would have been available to stoats during the warm periods of the Late Glacial and Postglacial. For instance, during the Postglacial period between the end of the Younger Dryas (11 400 years ago; Walker *et al.* 2003) and the flooding of the English Channel (7500 years ago at the very latest; Yalden 1999), the climate in Britain would have been similar to today (Andersen & Borns 1997). Therefore, we suggest that an mtDNA lineage and morphotype of stoats adapted to such warm conditions would have been able to enter Britain and replace the cold-tolerant lineage that was originally present and continues to survive in the isolated populations of Ireland and the Isle of Man. In this way, we can not only explain the morphological and mitochondrial differences between British and Irish stoats, but also clarify why the British mtDNA lineage is younger and less variable than the mtDNA lineage in Ireland.

There is evidence from the literature for replacement processes of the type we suggest. A phylogeographic study by Piertney *et al.* (2005) indicated partial replacement of one mtDNA lineage of water voles (*Arvicola terrestris*) by another in Britain, and ancient DNA studies have demonstrated sequential replacement of mtDNA lineages in brown bears (*Ursus arctos*) occupying Beringia (Barnes *et al.* 2002). There is also evidence that different mtDNA haplotypes may adapt individuals to different temperature conditions (Fontanillas *et al.* 2005). Different external morphologies may also of course be related to different temperature adaptations. Within the crow species (*Corvus corone*), the all-black carrion crow (*Corvus corone corone*) is expanding in Britain at the expense of the grey-and-black hooded crow (*C. c. cornix*) apparently in response to climate change (Cook 1975). The hooded crow is now found only in the north of Scotland. However, in a clear parallel with what we suggest for the stoat, it is the hooded crow (i.e. the form that is being displaced) that is found in Ireland, even though the climate in Ireland is similar to southern Britain.

In terms of the Irish question, we do not claim to have 'answered' it. There is a need to study a wide variety of species before there will be a comprehensive understanding of the peculiarities of Ireland's biota. Human introductions have undoubtedly been very important in generating the Irish fauna and flora, and it is actually very difficult to demonstrate convincingly that any particular terrestrial species did manage to colonize naturally. We believe that the weight of evidence strongly favours such a natural colonization by the stoat and that this colonization was very early, probably around the LGM. Our data suggest that there is no need to propose a Late Glacial land bridge to explain the natural occurrence of stoats in Ireland.

Our studies of the stoat provide another very interesting example of a terrestrial mammal with strong genetic differences between Irish and British populations to add to pine martens (Davison *et al.* 2001), pygmy shrews (Mascheretti *et al.* 2003) and mountain hares (Hamill *et al.* 2006). However, in the case of the stoat, we suggest a new explanation for this discrepancy between Britain and Ireland: that there was originally the same genetic type in Britain and Ireland, and that the much long-lasting connection with continental Europe allowed a replacement event in Britain. Replacement may be important in other examples of within-species genetic differences between Ireland and Britain. It could also explain some of the differences in the species lists for the two countries.

We are grateful to all people who provided samples for this study as listed in the electronic supplementary material. We thank I. Barnes, D. Förster, İ. Gündüz, J. Provan, A. Rambaut and B. Shapiro for their technical advice, S. Martínek for help with the figures and S. Bearhop, A. Douglas, J. Herman, C. Maggs and J. Provan for their comments on earlier versions of the manuscript. This work was supported by the Natural Environment Research Council. R.M. is supported by the Quercus partnership between Queen's University Belfast and the Environment & Heritage Service, Northern Ireland.

REFERENCES

Andersen, B. G. & Borns Jr, H. W. 1997 *The Ice Age world*. Oslo, Norway: Scandinavian University Press.

Bandelt, H.-J., Forster, P. & Röhl, A. 1999 Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48.

Barnes, I., Matheus, P., Shapiro, B., Jensen, D. & Cooper, A. 2002 Dynamics of Pleistocene population extinctions in Beringian brown bears. *Science* **295**, 2267–2270. (doi:10.1126/science.1067814)

Bowen, D. Q., Philips, F. M., McCabe, A. M., Knutz, P. C. & Sykes, G. A. 2002 New data for the Last Glacial Maximum in Great Britain and Ireland. *Quaternary Sci. Rev.* **21**, 89–101. (doi:10.1016/S0277-3791(01)00102-0)

Cook, A. 1975 Changes in carrion-hooded crow hybrid zone and possible importance of climate. *Bird Study* **22**, 165–168.

Corbet, G. B. 1961 Origin of the British insular races of small mammals and of the 'Lusitanian' fauna. *Nature* **191**, 1037–1040. (doi:10.1038/1911037a0)

Davison, A., Birks, J. D. S., Brookes, R. C., Messenger, J. E. & Griffiths, H. I. 2001 Mitochondrial phylogeography and population history of pine martens *Martes martes* compared with polecats *Mustela putorius*. *Mol. Ecol.* **10**, 2479–2488. (doi:10.1046/j.1365-294X.2001.01381.x)

Drummond, A. J. & Rambaut, A. 2006 BEAST v1.0, http://evolve.zoo.ox.ac.uk/beast/.

Excoffier, L., Laval, G. & Schneider, S. 2005 ARLEQUIN ver. 3.0: an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* **1**, 47–50.

Fjellberg, A. 1978 Fragments of Middle Weichselian fauna on Andøya, north Norway. *Boreas* **7**, 39.

Fleming, M. A. & Cook, J. A. 2002 Phylogeography of endemic ermine (*Mustela erminea*) in southeast Alaska. *Mol. Ecol.* **11**, 795–807. (doi:10.1046/j.1365-294X.2002.01472.x)

Fontanillas, P., Dépraz, A., Giorgi, M. S. & Perrin, N. 2005 Nonshivering thermogenesis capacity associated to mitochondrial DNA haplotypes and gender in the greater white-toothed shrew, *Crocidura russula*. *Mol. Ecol.* **14**, 661–670. (doi:10.1111/j.1365-294X.2004.02414.x)

Fu, Y.-X. 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925.

Gilg, O., Hanski, I. & Sittler, B. 2003 Cyclic dynamics in a simple vertebrate predator–prey community. *Science* **302**, 866–868. (doi:10.1126/science.1087509)

Hall, T. A. 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98.

Hamill, R. M., Doyle, D. & Duke, E. J. 2006 Spatial patterns of genetic diversity across European subspecies of the mountain hare, *Lepus timidus* L. *Heredity* **97**, 355–365. (doi:10.1038/sj.hdy.6800880)

Harris, S., Morris, P., Wray, S. & Yalden, D. 1995 *A review of British mammals: population estimates and conservation status of British mammals other than cetaceans*. Peterborough, UK: Joint Nature Conservation Committee.

Ho, S. Y. W., Phillips, M. J., Cooper, A. & Drummond, A. J. 2005 Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol. Biol. Evol.* **22**, 1561–1568. (doi:10.1093/molbev/msi145)

Innes, J. B., Chiverrell, R. C., Blackford, J. J., Davey, P. J., Gonzalez, S., Rutherford, M. M. & Tomlinson, P. R. 2004 Earliest Holocene vegetation history and island biogeography of the Isle of Man, British Isles. *J. Biogeogr.* **31**, 761–772. (doi:10.1111/j.1365-2699.2003.01048.x)

Kim, K. S., Lee, S. E., Jeong, H. W. & Ha, J. H. 1998 The complete nucleotide sequence of the domestic dog (*Canis familiaris*) mitochondrial genome. *Mol. Phylogenet. Evol.* **10**, 210–220. (doi:10.1006/mpev.1998.0513)

King, C. M. 1991 Stoat. In *The handbook of British mammals* (eds G. B. Corbet & S. Harris), pp. 377–387, 3rd edn. Oxford, UK: Blackwell.

Kurose, N., Masuda, R. & Yoshida, M. C. 1999 Phylogeographic variation in two mustelines, the least weasel *Mustela nivalis* and the ermine *M. erminea* of Japan, based on mitochondrial DNA control region sequences. *Zool. Sci.* **16**, 971–977. (doi:10.2108/zsj.16.971)

Kurose, N., Abramov, A. V. & Masuda, R. 2000 Intrageneric diversity of the cytochrome *b* gene and phylogeny of Eurasian species of the genus *Mustela* (Mustelidae, Carnivora). *Zool. Sci.* **17**, 673–679. (doi:10.2108/zsj.17.673)

Kurose, N., Abramov, A. V. & Masuda, R. 2005 Comparative phylogeography between the ermine *Mustela erminea* and the least weasel *M. nivalis* of Palaearctic and Nearctic regions, based on analysis of mitochondrial DNA control region sequences. *Zool. Sci.* **22**, 1069–1078. (doi:10.2108/zsj.22.1069)

Lambeck, K. & Purcell, A. P. 2001 Sea-level change in the Irish Sea since the Last Glacial Maximum: constraints from isostatic modelling. *J. Quaternary Sci.* **16**, 497–506. (doi:10.1002/jqs.638)

Martínková, N. & Searle, J. B. 2006 Amplification success rate of DNA from museum skin collections: a case study of stoats from 18 museums. *Mol. Ecol. Notes* **6**, 1014–1017. (doi:10.1111/j.1471-8286.2006.01482.x)

Mascheretti, S., Rogatcheva, M. B., Gündüz, I., Fredga, K. & Searle, J. B. 2003 Phylogeography of the Eurasian pygmy shrew *Sorex minutus*: clues on its mode of colonisation of Ireland. *Proc. R. Soc. B* **270**, 1593–1599. (doi:10.1098/rspb.2003.2406)

McCormick, F. 1999 Early evidence for wild animals in Ireland. In *The Holocene history of the European vertebrate fauna. Modern aspects of research* (ed. N. Benecke), pp. 355–371. Rahden, Germany: Verlag Marie Leidorf GmbH.

Miller, G. S. 1912 *Catalogue of the mammals of western Europe*. London, UK: British Museum.

Mix, A. C., Bard, E. & Schneider, R. 2001 Environmental processes of the ice age: land, oceans, glaciers (EPILOG). *Quaternary Sci. Rev.* **20**, 627–657. (doi:10.1016/S0277-3791(00)00145-1)

Moore, P. D. 1987 Snails and the Irish question. *Nature* **328**, 381–382. (doi:10.1038/328381a0)

Piertney, S. B., Stewart, W. A., Lambin, X., Telfer, S., Aars, J. & Dallas, J. F. 2005 Phylogeographic structure and postglacial evolutionary history of water voles (*Arvicola terrestris*) in the United Kingdom. *Mol. Ecol.* **14**, 1435–1444. (doi:10.1111/j.1365-294X.2005.02496.x)

Reimer, P. J. *et al*. 2004 IntCal04 terrestrial radiocarbon age calibration, 0–26 cal ky BP. *Radiocarbon* **46**, 1029–1058.

Rogers, A. R. & Harpending, H. 1992 Population-growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9**, 552–569.

Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. 2003 DnaSP, DNA polymorphism analysis by coalescent and other methods. *Bioinformatics* **19**, 2496–2497. (doi:10.1093/bioinformatics/btg359)

Schneider, S. & Excoffier, L. 1999 Estimation of demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* **152**, 1079–1089.

Sommer, R. & Benecke, N. 2004 Late- and post-glacial history of the Mustelidae in Europe. *Mammal Rev.* **34**, 249–284. (doi:10.1111/j.1365-2907.2004.00043.x)

Stuart, A. J. & Van Wijngaarden-Bakker, L. 1985 Quaternary vertebrates. In *The Quaternary history of Ireland* (eds K. J. Edwards & W. P. Warren), pp. 221–250. London, UK: Academic Press.

Walker, M. J. C., Coope, G. R., Sheldrick, C., Turney, C. S. M., Lowe, J. J., Blockley, S. P. E. & Harkness, D. D. 2003 Devensian lateglacial environmental changes in Britain: a multi-proxy environmental record from Llanilid, South Wales, UK. *Quaternary Sci. Rev.* **22**, 475–520. (doi:10.1016/S0277-3791(02)00247-0)

Woodman, P., McCarthy, M. & Monaghan, N. 1997 The Irish quaternary fauna project. *Quaternary Sci. Rev.* **16**, 129–159. (doi:10.1016/S0277-3791(96)00037-6)

Xia, X. & Xie, Z. 2001 DAMBE: data analysis in molecular biology and evolution. *J. Hered.* **92**, 371–373. (doi:10.1093/jhered/92.4.371)

Yalden, D. 1999 *The history of British mammals*. London, UK: Poyser.

Paper 2.2.2

Seifertová M., Bryja J., Vyskočilová M., **Martínková N.**, Šimková A. 2012. Multiple Pleistocene refugia and post-glacial colonization in the European chub (*Squalius cephalus*) revealed by combined use of nuclear and mitochondrial markers. *Journal of Biogeography* 39: 1024-1040.

Journal of Biogeography (J. Biogeogr.) (2012) 39, 1024–1040

ORIGINAL ARTICLE

Multiple Pleistocene refugia and post-glacial colonization in the European chub (*Squalius cephalus*) revealed by combined use of nuclear and mitochondrial markers

Mária Seifertová[1], Josef Bryja[1,2], Martina Vyskočilová[1], Natália Martínková[2,3] and Andrea Šimková[1]*

[1]*Department of Botany and Zoology, Faculty of Science, Masaryk University, Kotlářská 2, 61137 Brno, Czech Republic,* [2]*Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, Květná 8, 60365 Brno, Czech Republic,* [3]*Institute of Biostatistics and Analyses, Masaryk University, Kamenice 3, 62500 Brno, Czech Republic*

*Correspondence: Andrea Šimková, Department of Botany and Zoology, Faculty of Science, Masaryk University, Kotlářská 2, 61137 Brno, Czech Republic.
E-mail: simkova@sci.muni.cz

ABSTRACT

Aim To analyse patterns of nuclear and mitochondrial genetic variation in the European chub, *Squalius cephalus* (Linnaeus, 1758), in order to understand the evolutionary history of this species and to test biogeographical hypotheses for the existence of co-distributed European freshwater fish species.

Location Rivers in Europe (Finland, Poland, Czech Republic, France, Bulgaria, Spain, Italy).

Methods We genotyped 12 polymorphic microsatellite markers derived from 310 individuals collected from across the distribution of *S. cephalus* in Europe (including a total of 15 populations) and sequenced mitochondrial DNA (mtDNA) from a subset of 75 individuals. Sequences of mtDNA cytochrome *b* were analysed using both phylogenetic (median-joining networks) and population genetic methods (tests for demographic history, mismatch distributions, Bayesian coalescent analysis). Geographical structure in microsatellite loci was examined using a distance method (F_{ST}), factorial correspondence analysis (FCA) and a Bayesian clustering method (STRUCTURE).

Results The mtDNA network showed a clear split into four different haplogroup lineages: Western (separated into Atlantic and Danubian sublineages), Eastern, Aegean (occurring in two distinct sublineages in the Balkans and in Spain) and Adriatic. Our results indicate recent population expansion in the Eastern and Western Atlantic lineages and the admixture of two previously separate sublineages (Atlantic and Danubian) in the Western lineage. Bayesian structure analysis as well as FCA results roughly corresponded to the mtDNA-based structure, separating the sampled individuals into almost non-overlapping groups.

Main conclusions Our results support hypotheses suggesting origins of extant lineages of freshwater fishes in multiple refugia and the subsequent post-glacial colonization of Europe via different routes. We confirmed the previously proposed two-step expansion scenario from the Danube refuge, the existence of a secondary (Atlantic) refuge during the last glaciation (probably in the Rhone River) and population expansion of this lineage. Conspicuous divergences among Mediterranean populations reflect their different origin, as well as their low contribution to the recent genetic pool of chub in central Europe.

Keywords
Cytochrome *b*, Europe, freshwater fishes, glacial refugia, microsatellites, phylogeography, population structure.

http://wileyonlinelibrary.com/journal/jbi
doi:10.1111/j.1365-2699.2011.02661.x

Journal of Biogeography

INTRODUCTION

The most important processes influencing the evolution of biota in Europe throughout the past 2 million years (Myr) were the climatic oscillations that have driven periodic movements of ice sheets (Hewitt, 1999). Repeated range contractions were followed by range expansions from climatically more favourable regions (so-called glacial refugia), these events leaving a signature within the genome of many species (Hewitt, 1996). In freshwater fishes, evolutionary history is closely related to the biotic and geological evolution of a region, because their dispersal is dependent on direct connections between hydrographic basins, and because the history of basin interconnections reflects the underlying geological development of landscapes (Lundberg, 1993). Recent phylogeographical studies have brought many new insights into the post-glacial history of European freshwater fish species. Studies have often demonstrated the existence of evolutionarily distinct groups within each species, which suggests multiple refugia during glacial cycles (e.g. Nesbo *et al.*, 1999; Kotlik & Berrebi, 2001; Gum *et al.*, 2005; Hänfling *et al.*, 2009). Generally, the number of refugia defines the number of recent major genetic lineages, and the geographical distribution of these lineages provides clues to the routes of post-glacial colonization (Hänfling *et al.*, 2009).

The European chub, *Squalius cephalus* (Linnaeus, 1758), belonging to the Cyprinidae family and the Leuciscinae subfamily, has become a convenient model organism for the testing of biogeographical hypotheses concerning European freshwater ecosystems, especially due to its occurrence throughout most of Europe and a postulated Mesopotamian origin during the Pliocene (Durand *et al.*, 1999a, 2000; Zardoya & Doadrio, 1999). Moreover, chub distribution in Europe is not likely to be influenced by intentional introduction because of its low importance in commercial aquaculture, and, similar to other primary freshwater fishes, this species has a low capacity for trans-watershed dispersal (Zardoya & Doadrio, 1999). Thus, the distribution of this fish species can be postulated to closely reflect both the geological and climatic historical development of hydrographic basins (Ketmaier *et al.*, 2004), including episodes of isolation and interconnection processes (Durand *et al.*, 1999a; Sanjur *et al.*, 2003).

Durand *et al.* (1999a, 2000) identified four highly divergent mitochondrial DNA (mtDNA) phylogeographical lineages of chub in Europe – Western, Eastern and two Mediterranean lineages (Adriatic and Aegean) – that are probably descendants resulting from the rapid radiation of a widespread ancestor. This marked population diversification provides strong evidence of the eradication of *S. cephalus* from most of Europe during Pleistocene glacial maxima and its survival in four refugia, i.e. the Adriatic side of the Balkans, eastern Greece (Aegean rivers), the southern tributaries of the Danube, and the peripheries of the Black and Caspian seas. Two sources of post-glacial recolonization of central and northern Europe have been proposed on the basis of mtDNA analyses (Durand *et al.*, 1999a). First, the Western (Danube) lineage reached central and western Europe in two steps. This lineage entered western Europe during the Riss–Würm interglacial period (*c.* 130–110 ka) and survived the next glacial period in western European rivers (the Rhine and the Rhone); the next expansion started from these refugia at the end of the Würm period (10 ka). Second, expansion of the Eastern lineage (Ponto–Caspian) into northern Europe was in a single step from the Caspian refuge (Durand *et al.*, 1999a).

Nevertheless, the reconstruction of species history often depends on the markers studied. The majority of the postulated hypotheses elucidating the phylogeography and post-glacial colonization routes of freshwater fishes in the Palaearctic are inferred only from the geographical distribution of mtDNA variation (Hänfling & Brandl, 1998; Durand *et al.*, 1999a, 2000; Laroche *et al.*, 1999; Nesbo *et al.*, 1999; Bernatchez, 2001; Kotlik & Berrebi, 2001; Salzburger *et al.*, 2003). To date, only a few studies have addressed historical processes using nuclear markers (Triantafyllidis *et al.*, 2002; Makinen *et al.*, 2006) or used a comprehensive approach that combined both highly polymorphic microsatellite and mitochondrial markers (Gum *et al.*, 2005; Barluenga *et al.*, 2006; Bryja *et al.*, 2010a). Although mtDNA or allozymes were previously considered to be almost ideal for genealogical and evolutionary studies of animal populations (Avise *et al.*, 1987; Avise, 1991), microsatellites have, in the last two decades, become a more convenient molecular tool for inferring finer levels of population structure and dynamics for a species across its native range (Estoup *et al.*, 1998). First, the principal advantage of microsatellites lies in their high level of polymorphism, high mutation rate, and ability to distinguish fine-scale population genetic structure. Second, microsatellites (as nuclear markers) are biparentally inherited, whereas mtDNA is maternally inherited without recombination and reflects only the matrilineal history (Zhang & Hewitt, 2003). Third, studies concentrating only on mtDNA have often been criticized because the mitochondrial genome acts as a single genetic locus, providing only a single 'gene tree' which might not accurately reflect the 'organismal tree' (Degnan, 1993). The accuracy of mtDNA gene trees is compromised especially through hybridization that may lead to the introgression of mitochondrial genomes, which is not a rare process in nature (e.g. Bryja *et al.*, 2010b; Rodriguez *et al.*, 2010). However, using only microsatellites might not be optimal for phylogeographical inference mainly due to fast mutation rates, complicated evolutionary relationships among alleles, variable mutation rates among organisms and even between varieties, and the questionable neutrality of some microsatellite sequences (Zhang & Hewitt, 2003).

The aim of the present study was to clarify the evolutionary history of *S. cephalus* in Europe by integrating phylogeographical patterns of nuclear and mtDNA. We sampled 15 populations and assayed geographical variation across both microsatellites (12 loci; 310 samples) and mitochondrial sequence data [cytochrome *b* (cyt *b*); 75 samples]. We expected that microsatellite markers would reveal a pattern of variation on a more recent temporal scale, while analysis of mtDNA sequences would provide a historical perspective

helpful for the interpretation of microsatellite results. In particular, this study addressed the following three objectives: (1) to re-analyse the phylogeographical structure of chub populations in Europe, previously analysed using only mtDNA sequences; (2) to compare the patterns of population structure inferred from microsatellites and mtDNA to detect possible past hybridization and admixture zones; and (3) to investigate whether the phylogeographical distribution of chub populations analysed using microsatellites is consistent with the hypothesis of multiple-refugia origin, followed by the differential post-glacial colonization of central and northern Europe, as previously predicted for European freshwater fishes. In addition, we included the populations from Mediterranean areas (Adriatic, Iberian and Aegean populations), where the taxonomy of *S. cephalus* and closely related taxa is controversial (Kottelat & Freyhof, 2007). This may eventually contribute to a taxonomic revision of the Mediterranean chub.

MATERIALS AND METHODS

Sample collection and DNA extraction

A geographically and genetically representative total of 310 specimens of *S. cephalus* were sampled (Fig. 1) at 15 sites across Europe by electro-fishing during the years 2004–06 (Table 1; for more details see Seifertová *et al.*, 2008). In the Mediterranean sampling areas, Kottelat & Freyhof (2007) recently described several new species of the genus *Squalius* on the basis of morphological differences. In this study, however, we consider all collected individuals as *S. cephalus* in accordance with the studies of Durand *et al.* (1999a) and Zardoya & Doadrio (1999), and hereafter we use the term '*S. cephalus* complex' to strengthen its unclear taxonomical situation. Fin clips were taken from each individual and kept in absolute ethanol until the extraction of total genomic DNA using a DNeasy® Tissue Kit (Qiagen, Hilden, Germany).

Phylogeographical analysis of mtDNA

Amplification of an 840-bp fragment of cyt *b* was performed using the primers LACB (Schmidt & Gold, 1993) and HBCB (Dowling *et al.*, 2002) in five individuals of each population. Polymerase chain reaction (PCR) protocols and sequencing procedures are described in Seifertová *et al.* (2008). Nucleotide sequences obtained from our samples were aligned with available cyt *b* sequences of European chub retrieved from GenBank (see Appendix S1 in Supporting Information) in BioEdit 7.0.5.3. (Hall, 1999) using ClustalW multiple alignment.

Phylogenetic analysis of mtDNA sequences was not performed in the present study because such analyses have already been described elsewhere (i.e. Durand *et al.*, 1999a, 2000; Seifertová *et al.*, 2008). Haplotype relationships were obtained with the median-joining (MJ) algorithm in the software package network 4.5.1.6 (Bandelt *et al.*, 1999) using an equal transition/transversion ratio. The cyt *b* haplotypes and their frequencies were identified using DnaSP 5 (Librado & Rozas, 2009). The frequencies of the published haplotypes were retrieved from the original studies (Durand *et al.*, 1999a,b).

Genetic diversity and historical demography of major mtDNA lineages

DnaSP 5 software was used to estimate the genetic diversity of the main phylogenetic lineages observed in this study in *S. cephalus*, i.e. number of sequences (n), number of polymorphic sites (N_p), number of haplotypes (N_h), haplotype diversity (H_d), nucleotide diversity (π, expressed as percentages, i.e. 0.001 = 0.1%) and average number of nucleotide differences (k).

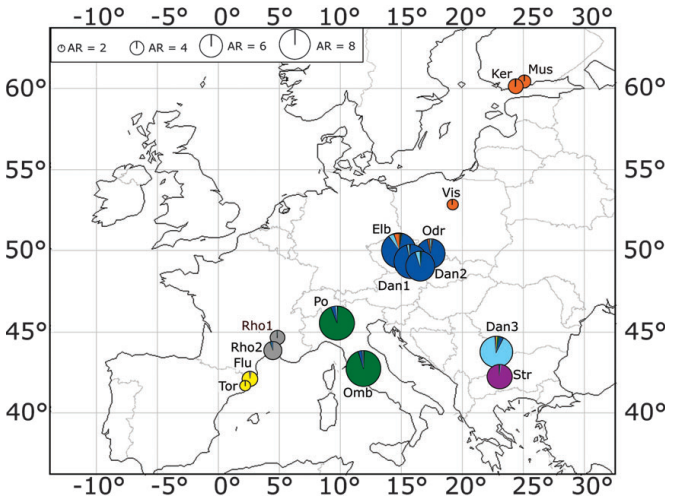


Figure 1 Sampled localities and geographical distribution of nuclear genetic variation in European chub (*Squalius cephalus*). Pie diagrams on the map correspond to the proportion of chub population in each of seven genetic clusters defined with the model-based clustering method (for $K = 7$) implemented in structure (Pritchard *et al.*, 2000). The size of circles represents the mean allelic richness of 12 microsatellite loci corrected for sample size.

Table 1 Characteristics of *Squalius cephalus* samples in this study, including mitochondrial DNA (mtDNA), phylogenetic lineages, population sampled, geographical location and hydrographic system, sample size (n) for microsatellite analyses, and symbols and numbers (in parentheses) of mtDNA haplotypes (five individuals from each locality were sequenced).

mtDNA lineage	Population ID	Location	River	Basin	n	mtDNA haplotypes
Eastern	Mus	Finland	Mustijoki	Baltic	21	H2 (5)
	Ker	Finland	Keravanjoki	Baltic	21	H2 (5)
	Vis	Poland	Mroga/Vistula	Baltic	21	H2 (5)
Western						
Danube	Odr	Czech Republic	Oder	Baltic	20	H18 (5)
	Elb	Czech Republic	Elbe	North Sea	18	H16 (2), H21 (2), H23 (1)
	Dan1	Czech Republic	Jihlava/Danube	Black Sea	19	H16 (2), H23 (1), H21 (1), H24 (1)
	Dan2	Czech Republic	Svitava/Danube	Black Sea	19	H16 (2), H21 (2), H22 (1)
	Dan3	Bulgaria	Vidbol/Danube	Black Sea	21	H16 (3), H17 (1), H18 (1)
Atlantic	Rho1	France	Le Buech/Rhone	Mediterranean Sea	20	H12 (5)
	Rho2	France	L'Arc/Rhone	Mediterranean Sea	19	H12 (5)
Aegean						
Balkan	Str	Bulgaria	Struma	Aegean	23	H19 (4), H20 (1)
Spain	Flu	Spain	Ser/Fluvia	Mediterranean Sea	23	H25 (5)
	Tor	Spain	Santa Coloma/Tordera	Mediterranean Sea	21	H25 (5)
Adriatic	Po	Italy	Ticino/Po	Adriatic	22	H7 (1), H29 (3), H30 (1)
	Omb	Italy	Merse/Ombrone	Tyrrhenian Sea	22	H29 (2), H31 (2), H32 (1)

Furthermore, we analysed the population history of clades by using an array of statistics that were originally introduced as tests of neutrality. We used the D^* and F^* tests (Fu & Li, 1993) in DnaSP 5 to analyse the population history of individual lineages. The critical values obtained by simulations in Fu and Li's tests were used to determine statistical significance (significant values at $P < 0.05$). Fu's F_S (Fu, 1997), one of the most powerful neutrality tests for detecting expansion in non-recombining genomic regions, and Tajima's D (Tajima, 1989) were estimated in Arlequin 3.5 (Excoffier *et al.*, 2005) and their significance was tested using 1000 simulations. A significant departure from the null hypothesis can be interpreted as the result of demographic history (population decrease or expansion), population structure and/or selection. Negative values of Fu's F_S and Tajima's D indicate population size expansion, while significantly positive values of F_S indicate a deficit of rare haplotypes and significant positive values of Tajima's D often suggest an admixture (Rand, 1996).

The hypothesis of recent population growth was evaluated in Arlequin 3.5 by calculating the mismatch distribution of pairwise differences (Rogers & Harpending, 1992) between all haplotypes within the main genetic lineages. A unimodal distribution is assumed to be the signature of a recent population size and, in a geographical context, range expansion, in contrast to multimodal or ragged distribution, which is typical in populations with stable demography (Rogers & Harpending, 1992). A bimodal distribution often results from the admixture of two previously separate lineages. We tested whether the data fitted the sudden demographic expansion model using Harpending's raggedness index (r_g; Harpending, 1994) and the sum of squared deviations (SSD) between the observed and expected mismatch from 2000 parametric bootstrap replicates.

Molecular dating

Molecular dating using Bayesian coalescent analysis requires reliable time constraints within the ingroup for dating events that are expected to be younger than 1–2 Ma, because of the possible decay of the evolutionary rate during the time period (Ho *et al.*, 2005). In this study, we used three time constraints, a fossil record, and two biogeographical assumptions based on previous knowledge. The constraint based on the fossil record was applied to the root of the tree. The lower limit was set to 0.7 Ma based on a *S. cephalus* fossil found in Germany, and the upper limit was 6.56 Ma as dated for a *S.* aff. *cephalus* fossil from Greece (Böhme & Ilg, 2003). Two star-like patterns were observed in the MJ network. One was found in the Dniester drainage area and one in the Rhone drainage. Both regions were previously identified as glacial refugia for chub (Durand *et al.*, 1999a). We assume that the star-like pattern observed in our data represents a sudden population expansion following climate amelioration. To provide for the uncertainty of the evolutionary history, we included the Riss–Würm interglacial and Würm glacial periods (0.01–0.13 Ma) as the time constraint for sequences of the central haplotypes of the respective stars and sequences derived from them by 1 bp.

The Bayesian coalescent analysis was run in beast 1.6.1 (Drummond & Rambaut, 2007), using all sequences generated in this study and related published sequences of the *S. cephalus* species complex, which amounted to 300 sequences in total. All time constraints were applied as priors with the uniform distribution. The Markov chain Monte Carlo (MCMC) was run for 10–20 million generations, sampling the chain every 1000th step. MCMC convergence was estimated from the likelihood trace in Tracer 1.4 (Rambaut & Drummond, 2007) and effective sample size for all parameters was > 100. We used

both a strict molecular clock and a relaxed clock model, with lognormal distribution for constant population and Yule process tree priors. We compared the four alternative models with Bayes factors, estimating marginal likelihoods for the models.

Microsatellite genotyping and intra-population variability

A total of 310 samples were genotyped at 12 polymorphic microsatellite loci (N7F8, N7G5, N7K4, LC03, LC04, E01, LC93, LC32, LC293, LC27, LC166 and LC288). Details about PCR conditions, primer sequences, fluorescent dyes and fragment analysis are included in Vyskočilová *et al.* (2007). The frequency of null alleles (NF) for each locus and population was computed using FREENA (Chapuis & Estoup, 2007).

Observed (H_O) and non-biased expected heterozygosities (H_E; Nei, 1978) were calculated for each locus and for each population using GENETIX 4.05.2 (Belkhir *et al.*, 2004). GENEPOP 3.4 (Raymond & Rousset, 1995) was used to test for deviation from the Hardy–Weinberg equilibrium (HWE) for each locus and population, and to test linkage disequilibrium between all pairs of loci among populations and within each population using the Fisher exact test. Probability values (P) for each pairwise comparison were estimated by a Markov chain method (1000 de-memorization, 100 batches and 1000 iterations). Significance levels of all multiple statistical tests were corrected using the false discovery rate (FDR) approach (Benjamini & Hochberg, 1995) implemented in the QVALUE package of the software R (Storey *et al.*, 2004). The allelic richness (AR) corrected by the rarefaction method was calculated using the FSTAT 2.9.3.2 program (Goudet, 1995) in order to investigate differences in the number of alleles among populations independent of sample size.

Population structure and differentiation

To quantify genetic structure, we calculated pairwise F_{ST}^{ENA} values in the FREENA software, using the so-called ENA method, because null alleles are known to overestimate the genetic differentiation between pairs of populations. This method efficiently corrects F_{ST} estimates for the positive bias introduced by the presence of null alleles (Chapuis & Estoup, 2007). Confidence intervals (95% CI) for mean F-statistics were generated by bootstrap resampling across loci.

Considering that chub have a large geographical distribution and long evolutionary history, the relevance of molecular information of microsatellites was tested. We compared F_{ST} and R_{ST} (Slatkin, 1995) values with the aim of analysing the main causes of population differentiation in microsatellites (i.e. whether the differentiation is caused by drift or by mutations). The calculation of these values shares equal assumptions when differentiation is caused solely by drift, whereas R_{ST} is expected to be larger than F_{ST} under the contribution of stepwise-like mutations (Hardy *et al.*, 2003). We used SPAGEDI 1.3 (Hardy & Vekemans, 2002) to test the null hypothesis of there being no contribution of stepwise mutations to genetic differentiation. We permuted allele sizes, i.e. we produced the variable pR_{ST}, and compared it with the real R_{ST} values also calculated in SPAGEDI. If the real R_{ST} is significantly higher than pR_{ST}, stepwise mutations have played a significant role in genetic differentiation. One thousand random permutations of allele sizes provided a distribution of pR_{ST} values, 95% CIs covering the 25th to the 975th ordered values, and P-values testing if $R_{ST} > pR_{ST}$ (Hardy *et al.*, 2003).

The pattern of gene frequency differentiation between populations was represented by a factorial correspondence analysis (FCA) implemented in GENETIX 4.05.2 (Belkhir *et al.*, 2004). Additionally, the program STRUCTURE 2.2 (Pritchard *et al.*, 2000) was used to perform an individual-based Bayesian cluster analysis, which allowed us to infer and delineate the most probable number of genetically homogeneous groups of sampled individuals (K) with no a priori assumptions of population structure, and to assign individuals to the inferred groups. We performed a series of 10 independent runs for each K ranging from 1 to 10 using 500,000 MCMC replications with a burn-in period of 50,000 chains. Admixture models with correlated allele frequencies were used. To infer the most appropriate number of genetic groups in our dataset, we used the ad hoc statistic ΔK method proposed by Evanno *et al.* (2005). Additionally, we forced the assignment of individuals to clusters at values of K beyond the number considered to maximize the posterior probability of the data $P(X|K)$. This approach can be used to reconstruct the hierarchical relationships among populations, as well as to distinguish between processes that are likely to shape this structure (e.g. Flanders *et al.*, 2009; Bryja *et al.*, 2010a). We utilized the program CLUMPP 1.1.1 (Jakobsson & Rosenberg, 2007) in order to assess the average pairwise similarity (H) of runs with the Greedy algorithm and 10,000 random input orders of the 10 independent STRUCTURE runs. Results from CLUMPP were imported into DISTRUCT 1.1 (Rosenberg, 2004) for graphical representation.

RESULTS

Phylogeographical structure and historical demography based on mtDNA

Using 75 individuals from 15 populations sequenced for mtDNA, 18 different cyt *b* haplotypes of 840 bp were identified (GenBank accession numbers EU791864–EU791885). Detailed information on the cyt *b* sequences is included in Seifertová *et al.* (2008). To perform phylogeographical and historical demography analyses, we completed our dataset of cyt *b* with 24 sequences retrieved from GenBank (AJ006887–AJ006902, AJ002319, AJ002342, AJ002343, AJ002344, AJ002346 and AJ002347) and corresponding frequency data from Durand *et al.* (1999a,b). A total of 275 individuals (Appendix S2) were used for network calculation and the final dataset comprised 33 haplotypes of 600-bp cyt *b* sequences with 91 polymorphic sites (Appendix S1). No indels or stop codons were found, suggesting that no nuclear copies of cyt *b* were included. The MJ network showed a clear split into four different

haplogroups (Fig. 2), which correspond to four previously described phylogenetic lineages, i.e. Western, Eastern, Aegean and Adriatic (*sensu* Durand *et al.*, 1999a). Two of these lineages can be further subdivided to sublineages that are geographically separated, increasing the number of mtDNA clusters to six. First, haplotype 25 (yellow in Fig. 2; hereafter called Aegean–Spain) of the Aegean lineage was observed in the Ebro, Fluvia and Tordera rivers in Spain, while remaining haplotypes of this lineage (violet in Fig. 2; Aegean–Balkan) occur in the Evros and Struma rivers in the Balkan Peninsula (Fig. 2). Second, the Western lineage can be divided into Western–Atlantic (grey in Fig. 2) and Western–Danubian sublineages (blue in Fig. 2). The distribution of these two sublineages overlaps in the Elbe and Rhine rivers (Fig. 2). Two common haplotypes (H2 from the Eastern lineage and H12 from the Western–Atlantic lineage) were found in 142 out of 275 (51.64%) chub individuals. As a consequence, the Western–Atlantic lineage and partially also the Eastern lineage displayed a star-like pattern of haplotype distribution, indicating the recent descent of the satellite haplotypes from the ancestral haplotypes H12 and H2 (Fig. 2).

Measures of mitochondrial DNA diversity, estimates of demographic parameters and neutrality tests calculated for each lineage are shown in Table 2. Because the Aegean lineage comprised the haplotypes from the Aegean area (Aegean–Balkan sublineage) and also from the Iberian area (Aegean–Spain sublineage), the analyses were performed separately for both sublineages. The highest values of diversity were observed for the Adriatic lineage. The Western–Atlantic sublineage showed significant negative D^* and F^* values as well as a large and significant negative value of F_S (Table 2). The mismatch distributions for Eastern (Fig. 3a), Western–Danubian (Fig. 3b) and Western–Atlantic lineages (Fig. 3c) were unimodal, which is a typical sign of populations having undergone a recent expansion. A bimodal pattern for the Aegean–Balkan lineage (Fig. 3d) was observed, which may suggest the admixture of two previously separate lineages in the Aegean region (Valqui *et al.*, 2010). A ragged pattern of mismatch distribution was observed for the Adriatic lineage (Fig. 3e), indicating that populations are at demographic equilibrium (Rogers & Harpending, 1992).

The model with a relaxed uncorrelated molecular clock with lognormal distribution and Yule process tree prior had the highest likelihood amongst the tested models. Comparison of marginal likelihoods with Bayes factors did not strongly favour the model over the one assuming the same clock and constant population size tree prior (Bayes factor = 5.5). We further used the latter model to avoid overparameterization. The median mutation rate with this model was 0.0631 mutations per lineage per site per Myr, which represents a sequence divergence of *c.* 13% Myr^{-1}. The most recent common ancestor of the paraphyletic Western–Danubian lineage was 0.16 Ma [95% highest posterior density (HPD) 0.03–0.36]. The Western–Atlantic lineage was included within the group and was 0.04 Ma (HPD: 0.01–0.09). The Eastern lineage was 0.09 (HPD: 0.02–0.19) Ma, the Aegean–Spain 0.03 Ma (HPD: 0.003–0.09), the Aegean–Balkan 0.14 Ma (HPD: 0.02–0.33) and the Adriatic lineage 0.18 Ma (HPD: 0.03–0.39) (Fig. 4). The time of population expansion, estimated from the τ parameter of the mismatch distribution analysis ($t = \tau/2\mu_{locus}$), was 0.04 Ma (95% CI: 0.005–0.05) for the Eastern lineage, 0.02 Ma (95% CI: 0.01–0.03) for the Western–Danubian lineage, and 0.04 Ma (95% CI: 0.006–0.04) for the

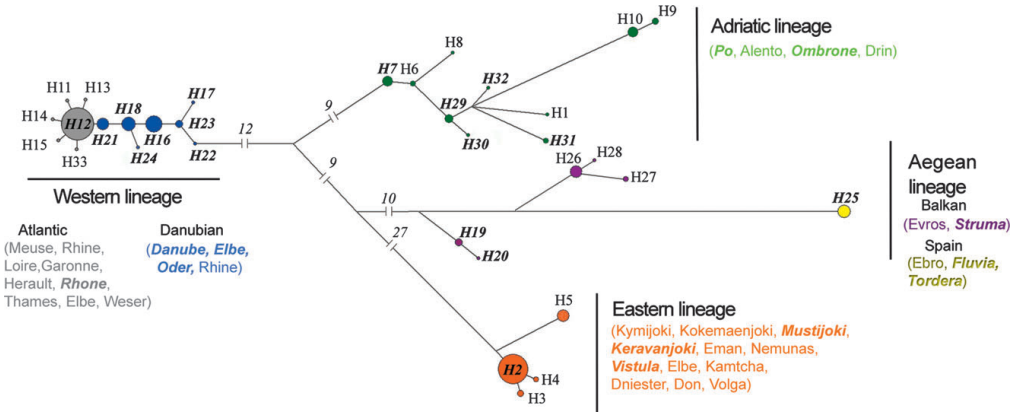


Figure 2 Network constructed using mitochondrial DNA cytochrome *b* gene sequences of European chub (*Squalius cephalus*). The haplotypes are represented by circles, with diameters proportional to their frequencies (for detailed information see Table 1). The haplotypes and rivers included in the sampling within this study are shown in bold italic. Branch lengths correspond to the number of mutation steps between haplotypes. In the case of long branches, the number of mutation steps is represented by a number above the branches. Sequences obtained from GenBank are included in the network with their respective frequencies (as given in the original publication; see the text for more details). Colour legend of mtDNA lineages: orange, Eastern; blue, Western–Danubian; grey, Western–Atlantic; violet, Aegean–Balkan; yellow, Aegean–Spain; green, Adriatic.

Table 2 Measures of mitochondrial DNA diversity, estimates of demographic parameters and neutrality tests calculated for four lineages of *Squalius cephalus* identified in European rivers in this study. The results of the observed mismatch distribution against a sudden expansion distribution included the raggedness r_g statistic and the sum of squared deviations (SSD).

	Eastern	Western–Danubian	Western–Atlantic	Aegean–Balkan	Aegean–Spain	Adriatic
n	79	51	83	19	12	31
N_P	6	6	5	13	0	20
N_h	4	7	6	5	1	10
H_d (SD)	0.330 (0.063)	0.740 (0.035)	0.118 (0.048)	0.637 (0.105)	–	0.862 (0.034)
π (SD)	0.170 (0.038)	0.193 (0.020)	0.020 (0.008)	0.715 (0.152)	–	0.999 (0.074)
k	1.020	1.161	0.120	4.292	–	5.983
D^*	1.139	−1.414	**−4.137**** ($P = 0.002$)	0.727	–	0.370
F^*	0.757	−1.256	**−4.027**** ($P = 0.002$)	0.788	–	0.551
F_S	1.329	−1.445	**−8.450***** ($P < 0.001$)	3.126	–	1.232
D	−0.375	−0.329	**−1.934*** ($P = 0.03$)	0.566	–	0.675
SSD	0.057	0.006	0.0002	0.105	–	0.021
r_g	0.411	0.087	0.602	0.189	–	0.037

n, number of sequences; N_P, number of polymorphic sites; N_h, number of haplotypes; H_d, haplotype diversity; π, nucleotide diversity (expressed as percentages, i.e. 0.001 = 0.1%); k, average number of nucleotide differences; SD, standard deviation; F_S, Fu's statistic; D, Tajima's D-test. The significant tests are shown with asterisks as follows: $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

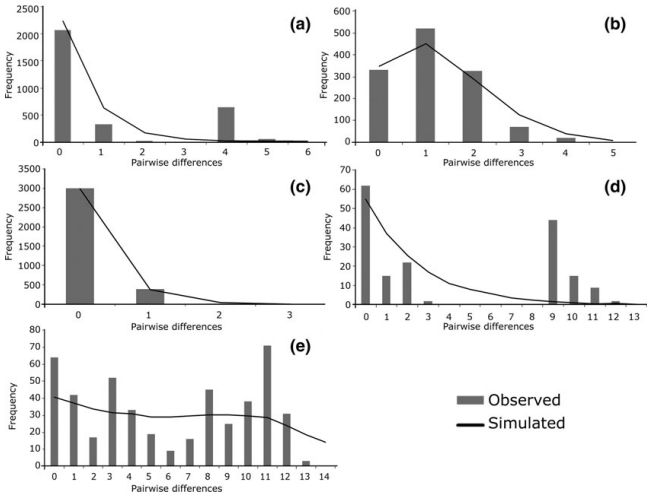


Figure 3 Mismatch distribution among mitochondrial DNA (mtDNA) haplotypes from five major *Squalius cephalus* mtDNA lineages observed in Europe: (a) Eastern lineage, (b) Western–Danubian lineage, (c) Western–Atlantic lineage, (d) Aegean–Balkan lineage, (e) Adriatic lineage. The expected frequency is based on the sudden expansion model. Grey bars indicate the observed values and black lines show the expected distribution under the sudden expansion model.

Western–Atlantic lineage. The remaining lineages did not exhibit the signal of a single recent population expansion in the mismatch analyses.

Genetic diversity and structure assessed by microsatellites

A total of 257 different alleles were detected at the 12 microsatellite loci in the 310 chub individuals (2–52 alleles per locus; Appendix S2). The non-random association of genotypes at two loci (N7F8, LC93) in the Flu locality at the 0.05 level after FDR correction was found. The mean frequencies of null alleles per population were lower than 5% in all but one locus (the only exception was locus E01 with an average of 6.61% of null alleles per population; Appendix S2). A high frequency of null alleles (> 15%) was detected only in the Elb (15.8%), Rho1 (22.9%) and Rho2 (33.6%) populations at the locus E01 and in the Omb population (17.1%) at the locus

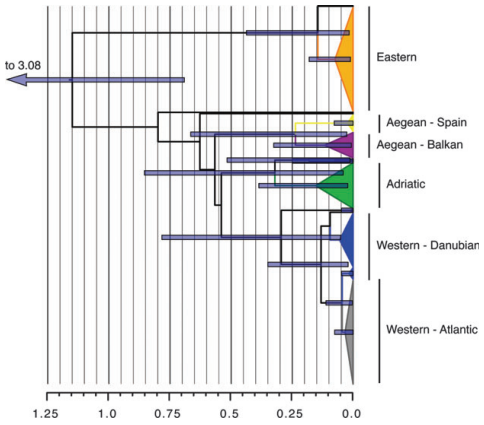


Figure 4 Chronogram of the European chub (*Squalius cephalus*) cytochrome *b* sequences inferred from the Bayesian coalescent analysis assuming an uncorrelated relaxed clock with lognormal distribution and constant population size. Node bars represent the 95% highest posterior density intervals of node ages; collapsed lineages are colour-coded according to Fig. 2. Sequences not included in collapsed lineages were downloaded from GenBank. The scale bar is calibrated to million years ago (Ma).

LC166. Deviation from HWE expectation was found for the three populations Dan3, Rho1 and Rho2, which could be the result of the presence of null alleles at locus E01 in the Rho1 and Rho2 (see above) populations and at locus LC32 in the Dan3 population (14.4%). When these two loci were removed from the dataset, no deviation from HWE for these populations was observed. The mean observed and expected heterozygosities and allelic richness estimated by the rarefaction method for the smallest sample size (17 diploid individuals) for each population are shown in Table 3. Mean AR per locus was highly correlated with H_E (Spearman correlation, $r = 0.964$, $P < 0.001$). The highest values of genetic diversity were found in the populations of Adriatic (Po and Omb) and Western–Danubian lineages (Elb, Odr, Dan1, Dan2 and Dan3) (Fig. 1).

Microsatellites revealed a considerable level of genetic differentiation over all populations (global $F_{ST}^{ENA} = 0.287$; $P = 0.0001$, 95% CI = 0.224–0.373). Pairwise F_{ST}^{ENA} values ranged from 0.017 (Dan2 and Elb; Dan1 and Dan2) to 0.560 (Mus and Tor) and most of them (94.3%) were significantly higher than zero, suggesting a distinct genetic structure of European chub populations (Table 4). The multilocus R_{ST} value was significantly higher than mean pR_{ST}. This difference was also significant at 5 out of 12 loci when analysed separately per locus (Appendix S2), suggesting the important role of stepwise mutations in population differentiation. When we analysed differences between pairwise R_{ST} and pairwise pR_{ST}, we found important effects of mutations only in pairs of population samples originating from different mtDNA lineages (six groups were considered; Eastern, Western–Danubian, Western–Atlantic, Adriatic, Aegean–Balkan, Aegean–Spain). This difference was significant in 49 out of 89 comparisons (55%). However, none of the 16 pairwise comparisons within mtDNA lineages showed significant differences, suggesting that genetic differences within lineages were caused predominately by drift (Table 4).

Table 3 The measures of microsatellite genetic diversity in 15 populations of *Squalius cephalus* in European rivers.

Population	H_O	H_E	AR
Mus	0.413	0.402	3.441
Ker	0.500	0.509	3.776
Vis	0.524	0.478	3.041
Odr	0.612	0.635	7.293
Elb	0.694	0.729	8.530
Dan1	0.679	0.707	8.585
Dan2	0.617	0.662	7.670
Dan3*	0.647	0.688	8.453
Str	0.601	0.589	6.509
Po	0.708	0.716	8.888
Omb	0.687	0.702	8.902
Rho1*	0.491	0.533	3.857
Rho2*	0.535	0.574	4.722
Flu	0.522	0.519	3.803
Tor	0.258	0.283	2.864
Mean	0.566	0.582	6.022

H_O, observed heterozygosity; H_E, expected heterozygosity, which represents a non-biased estimate according to Nei (1978); AR, mean allelic richness per locus.
*Significant departure ($P < 0.05$) from the Hardy–Weinberg equilibrium (after false discovery rate correction for multiple comparisons).

Using a FCA of microsatellite genotypes, the sampled individuals were clearly separated into the four different groups with almost no overlap (Fig. 5a). Eastern populations (i.e. Mus, Ker and Vis) were separated from all remaining samples on the first axis (explaining 21.06% of total variance). Two Iberian populations (i.e. Flu and Tor) and the Bulgarian population from the Aegean basin (i.e. Str) were separated from the remaining populations on the second axis (explaining 21.06% of variance). The Czech (Odr, Elb, Dan1, Dan2), Italian (Po, Omb), French (Rho1, Rho2) and Bulgarian Danubian (Dan3) populations clustered together in the analysis using the whole dataset. The third axis (14.84% of variance) only deepened the differences among the above-described groups and did not yield the separation of further populations (not shown). When we performed more detailed analysis using only the samples from the last-mentioned large group, three clusters became visible, i.e. Italian, French and Czech+Bulgarian (Fig. 5b).

Bayesian analyses in STRUCTURE based on the highest ΔK value, i.e. ln $P(X/K)$, revealed seven population clusters (Appendix S3). The seven clusters mostly corresponded to geographical areas, with the exception of two closely situated Bulgarian localities (Dan3 and Str), which formed separate clusters (Figs 1 & 6). When analysing the individual assignments beyond the best K, we found that the first split ($K = 2$)

Table 4 Pairwise F_{ST}^{ENA} values (below the diagonal) and pairwise R_{ST} values (above the diagonal, ANOVA approach) calculated for all microsatellite loci across European populations of *Squalius cephalus*.

	Mus	Ker	Vis	Odr	Elb	Dan1	Dan2	Dan3	Str	Po	Omb	Rho1	Rho2	Flu	Tor
Mus		0.138	0.124	0.323	0.303	0.364	0.387	0.318	**0.500**	**0.602**	**0.541**	**0.622**	**0.675**	**0.724**	**0.885**
Ker	0.133		0.261	0.417	0.415	0.464	0.482	0.392	**0.590**	**0.680**	**0.626**	**0.707**	**0.749**	**0.781**	**0.909**
Vis	0.181	0.148		0.308	0.256	0.306	0.363	0.323	0.309	**0.528**	**0.500**	0.516	**0.562**	**0.609**	**0.764**
Odr	0.400	0.358	0.344		0.006	0.049	0.006	−0.003	**0.377**	**0.274**	0.254	0.246	0.244	**0.424**	**0.666**
Elb	0.340	0.296	0.292	$0.034^{n.s.}$		0.008	0.016	0.023	**0.347**	**0.237**	**0.225**	0.220	**0.235**	**0.384**	**0.651**
Dan1	0.377	0.332	0.330	0.042	$0.018^{n.s.}$		0.006	0.063	**0.338**	0.100	0.126	0.098	0.102	0.228	**0.489**
Dan2	0.403	0.358	0.357	0.029	$0.017^{n.s.}$	$0.017^{n.s.}$		0.014	**0.401**	0.167	0.162	0.178	0.171	**0.360**	**0.626**
Dan3	0.394	0.349	0.363	0.073	$0.035^{n.s.}$	0.039	0.045		**0.363**	**0.296**	**0.283**	0.309	**0.310**	**0.470**	**0.690**
Str	0.463	0.407	0.428	0.302	0.252	0.262	0.281	0.269		**0.561**	**0.586**	**0.591**	**0.631**	**0.645**	**0.791**
Po	0.384	0.336	0.357	0.135	0.094	0.116	0.121	0.094	0.236		0.034	0.105	0.104	0.217	**0.488**
Omb	0.371	0.330	0.345	0.127	0.087	0.107	0.111	0.087	0.249	0.028		0.103	0.107	0.255	**0.521**
Rho1	0.468	0.413	0.418	0.167	0.143	0.150	0.159	0.175	0.359	0.203	0.203		−0.007	0.180	0.563
Rho2	0.447	0.391	0.393	0.132	0.119	0.130	0.129	0.157	0.336	0.181	0.177	$0.022^{n.s.}$		0.239	**0.660**
Flu	0.441	0.392	0.410	0.267	0.216	0.229	0.247	0.261	0.413	0.273	0.275	0.338	0.325		0.262
Tor	0.560	0.504	0.529	0.451	0.397	0.404	0.432	0.436	0.545	0.429	0.433	0.529	0.510	0.149	

For F_{ST}^{ENA} values, a value significantly different from zero was tested by bootstrapping over loci. n.s., not significant ($P > 0.05$) after false discovery rate correction. Significant differences between pairwise R_{ST} and pairwise pR_{ST} (i.e. important effect of mutations on population differentiation) are shown in bold. F_{ST}^{ENA}, a measure of population differentiation corrected for the presence of null alleles (Chapuis & Estoup, 2007); R_{ST}, allele size-based measure of population differentiation assuming stepwise mutation process (Slatkin, 1995); pR_{ST}, R_{ST} computed after allele size permutation (Hardy *et al.*, 2003).

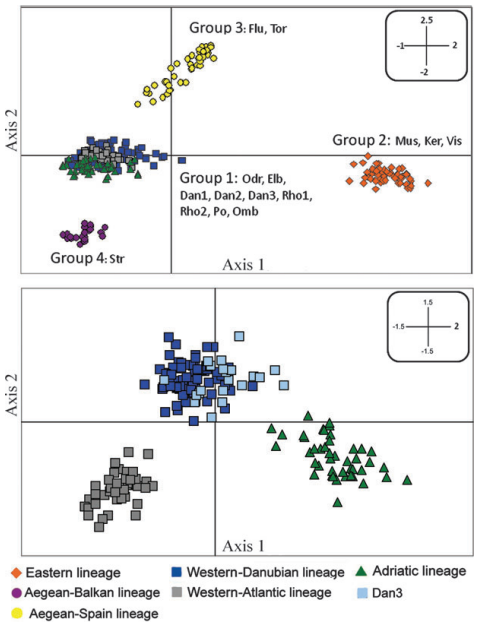


Figure 5 Two-dimensional factorial correspondence analysis (FCA) of nuclear microsatellite variation: (a) analysis using the whole European chub (*Squalius cephalus*) dataset (the first axis explains 21.06% of total variance and the second axis explains 17.66% of total variance); (b) detailed analysis using only the subset of individuals clustered in Group 1 in (a) (the first axis explains 29.74% of total variance and the second axis explains 24.78% of total variance). Different symbols represent main genetic groups.

separated eastern populations (Mus, Ker and Vis) from all others; with increasing K (from 2 to 10) the populations from Mediterranean areas were successively assigned to separate groups (Fig. 6). The Iberian populations (Flu and Tor) already started to separate at $K = 2$. The southern Bulgarian population from the Struma Basin (Str) formed a cluster with the Italian populations (Po and Omb) at $K = 4$ and 5, whereas this Bulgarian population was separated from Italian populations and formed a separate group at $K = 6$. The French populations (Rho1 and Rho2) were separated from the Czech populations (Odr, Elb, Dan1 and Dan2) and Bulgarian population (Dan3) at $K = 5$. The Bulgarian population from the lower Danube (Dan3) formed a single cluster at $K = 7$. Further increasing K up to 10 did not introduce new information, except partial separation of the population from the Oder River (Odr) at $K = 8$.

Neither FCA nor STRUCTURE allowed us to correct efficiently for null alleles at microsatellite loci, but both analyses also provided very similar results when run using datasets in which the loci with null alleles (E01 and LC166) were excluded.

DISCUSSION

In the present study, the range-wide genetic diversity and population structure of *S. cephalus* were, for the first time, investigated using a combination of mtDNA and nuclear markers. Comparison of the phylogeographical history of chub (estimated by cytochrome *b* sequences) with its recent population structure (assessed by nuclear microsatellites) allowed us to obtain a better estimate of the chub's evolutionary history. A notable result of our study is evidence of at least seven genetically separated subpopulations of *S. cephalus* in Europe. Their genetic differentiation may be

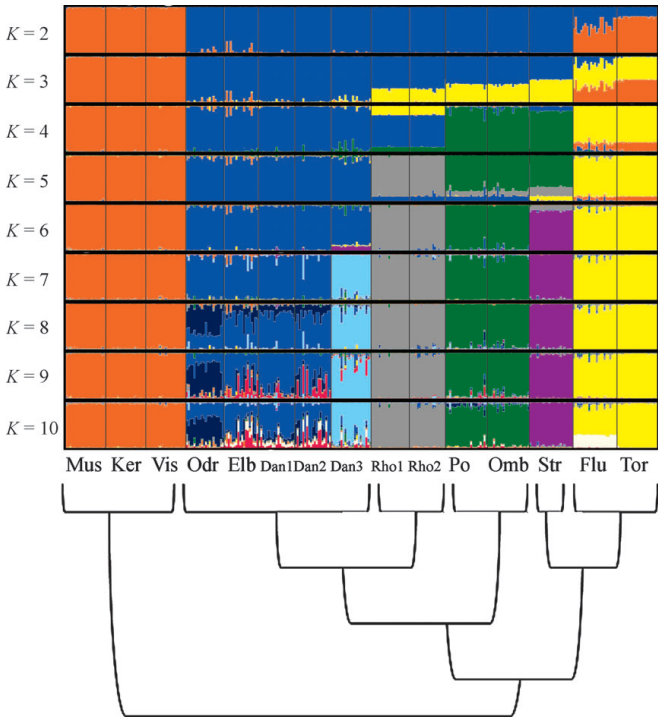


Figure 6 Bayesian population assignments of 310 European chub (*Squalius cephalus*) individuals using 12 microsatellite loci considering a given *K* (from 2 to 10) in STRUCTURE. Black lines separate individuals sampled from different geographical locations. The phylogenetic tree indicates a schematic topology of mitochondrial DNA sequences based on Bayesian inference in BEAST.

caused by a combination of different factors, i.e. post-glacial colonization from different refugia or recent evolutionary processes such as random drift or isolation by distance.

Population-genetic structure of chub in Europe – comparison of mtDNA and nuclear markers

Hitherto, there have been few studies of European freshwater fish species using both nuclear and mitochondrial markers to investigate the population structure from the phylogeographical perspective (e.g. Gum *et al.*, 2005; Barluenga *et al.*, 2006; Bryja *et al.*, 2010a). Using only mitochondrial markers can lead to biased estimates of species phylogeny (because of sampling only one locus), while using both microsatellites and mtDNA sequences allows us to obtain a more precise view of species history. Thus, this latter approach is recommended in order to increase the performance of phylogenetic studies (e.g. Rodriguez *et al.*, 2010).

Our sampling covered most of the geographical area of chub distribution in Europe. In accordance with previous studies (Durand *et al.*, 1999a, 2000), our mtDNA analysis, including both published cyt *b* sequences and those obtained by this study, confirmed the existence of four main mitochondrial lineages of *S. cephalus*, i.e. Western, Eastern, Adriatic and Aegean lineages. Network analysis further divided two of them into geographically distinct sublineages: the Western lineage into Atlantic (occurring mainly in France and Germany) and Danubian (occurring mainly in the Danube Basin, the Elbe and the Oder) sublineages; and the Aegean lineage into Balkan (the Evros and Struma rivers) and Spanish (rivers in northern Spain) sublineages, increasing the number of mtDNA groups to six. Likewise, microsatellite analysis revealed the strong genetic differentiation of chub populations even on a relatively small geographical scale, which is, in most cases, in congruence with mitochondrial data (e.g. the strict separation of Danube and Struma populations in Bulgaria). In addition, microsatellites, due to their high variability and fast mutation rate, showed the substantial divergence between hydrologically connected populations. This seems to be a common feature of European freshwater species, and it suggests the important role of genetic drift or local selection in shaping the genetic structure of non-migratory fishes (Triantafyllidis *et al.*, 2002; Gum *et al.*, 2005; Bryja *et al.*, 2010a). Almost all pairwise comparisons of populations showed highly significant values ($P < 0.001$) of F_{ST}. The only exceptions were several pairs

Table 5 Overview of the results of selected phylogeographical studies of freshwater European fish species with similar environmental preferences and distribution in Europe.

Species	References	Marker	Phylogeographical structure	Suggested refugia
European chub (*Squalius cephalus*), Cyprinidae	Durand *et al.* (1999a, 2000), this study	mtDNA, microsatellites	1. Western lineage (two-step expansion scenario)	Southern tributaries of the Danube (secondary Atlantic refugium in Rhine/Rhone drainages)
			2. Eastern lineage	Periphery of the Black and Caspian seas
			3. Aegean lineage	Eastern Greece (Aegean rivers)
			4. Adriatic lineage	Adriatic side of the Balkans
European bitterling (*Rhodeus amarus*), Cyprinidae	Bohlen *et al.* (2006), Bryja *et al.* (2010a)	mtDNA, microsatellites	1. Danubian cluster (Western lineage)	Refugium in the lower Danube
			2. Dnieper–Dniester–Vistula cluster (Eastern lineage)	North of the Black Sea (colonization of NE Europe)
			3. Aegean cluster	Several refugia around the Aegean Sea
Varione (*Telestes souffia*), Cyprinidae	Salzburger *et al.* (2003)	mtDNA	1. Italian lineage	Refugium in Italy
			2. Alpine lineage (Rhone/Var and Danube/Rhine sublineages)	Danubian refugium
Barbel (*Barbus barbus*), Cyprinidae	Kotlik & Berrebi (2001)	mtDNA	1. Lineage: E Bulgaria and N Anatolia	N/W Black Sea refugium
			2. Lineage: W and central Europe (two-step expansion scenario)	Danubian refugium (secondary Atlantic refugium in southern France – Rhine/Rhone)
European grayling (*Thymallus thymallus*), Salmonidae	Gum *et al.* (2005), Gum (2009)	mtDNA, microsatellites	1. Lineage N/NE Europe	North of the Black or Caspian Sea basin
			2a. Lineage SW/W – Rhine/Main drainage and England	Persistence within the Rhine/Main system during the last major Pleistocene glaciation
			2b. Central and E Europe from Elbe drainage to the Vistula, in S Scandinavia and Lithuania	Ponto-Caspian refugium
			3. Danube drainages	Danubian refugium
			4. Adriatic lineage	Adriatic refugium
European catfish (*Silurus glanis*), Siluridae	Krieg *et al.* (2000), Triantafyllidis *et al.* (2002)	mtDNA, microsatellites	No pattern of geographical structuring	One glacial refugium around the Ponto-Caspian region (colonization of Central and S Europe)
Spined loach (*Cobitis taenia*), Cobitidae	Culling *et al.* (2006)	mtDNA	1. Lineage: N/W Europe	Three refuges in the Ponto-Caspian area: 1. colonization of Europe westward to England
			2. Lineage: N, E and N/W Europe	2. Colonization of Europe spreading northward into Russia and moving west
			3. Lineage: lower southern Bug River	3. Nearby Black Sea refugium did not contribute to the colonization of Europe
European perch (*Perca fluviatilis*), Percidae	Nesbo *et al.* (1999)	mtDNA	1. Lineage: S Europe	Danubian refugium
			2. Lineage: E Europe	Refugium in SE Europe – Black/Caspian Sea drainages
			3. Lineage: W Europe and Poland	Refugium in W European Rivers – position of this refugium is unclear
			4. Lineage: Norway	Refugium in NE Europe – ice-dammed lakes east of the ice sheet
Bullhead (*Cottus gobio*), Cottidae	Englbrecht *et al.* (2000), Hänfling *et al.* (2002)	mtDNA, microsatellites	1. Lineage: Danube, Elbe, Rhine, Main	Danubian refugium
			2. Lineage: Oder, Vistula	Ponto/Caspian refugium
			3–7. Lineage: Lower Rhine, Seine, Adour, England	Persistence within the British Isles and within the drainages of the Rivers Elbe and Main

mtDNA, mitochondrial DNA.

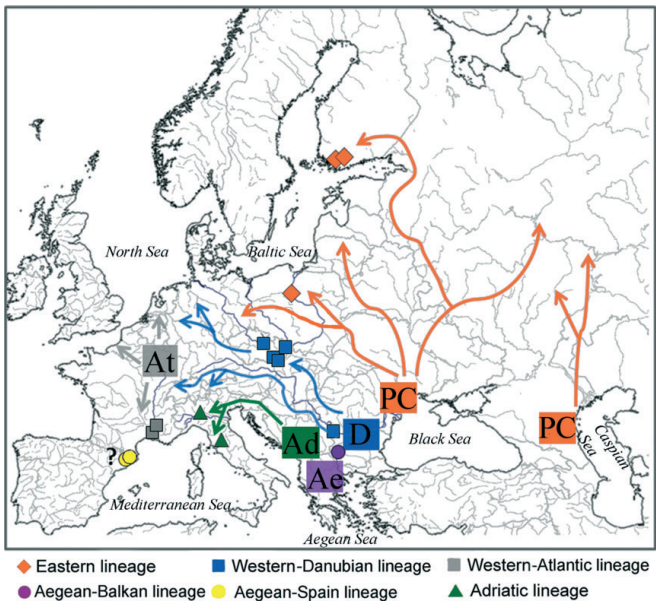


Figure 7 Geographical location of sampling sites and schematic overview of the refugia suggested for the European chub (*Squalius cephalus*). Arrows indicate possible colonization pathways. Symbols refer to the affiliation of haplotypes of a population to one of the mitochondrial DNA lineages. PC, Ponto-Caspian refugium; D, Danubian refugium; Ad, Adriatic refugium; Ae, Aegean refugium; At, secondary Atlantic refugium. The presence of a possible Iberian refugium is indicated by a question mark.

of central European populations, which could be explained by short divergence time and/or by short geographical distances allowing gene flow among them. Based on the allele permutation test (Hardy *et al.*, 2003), we indicated that, in most cases, R_{ST} provides a better description of population structure than F_{ST}. This result suggests that genetic differences among basins may be caused also by mutations at microsatellite loci, while the differences within basins are caused predominately by drift.

The analysis of population genetic structure based on nuclear markers revealed seven major groups of chub populations in Europe, which is in line with the mitochondrial evidence supporting the multiple-refugia assumption (Durand *et al.*, 1999a, 2000). However, when using Bayesian analysis on microsatellite data in STRUCTURE, the pattern observed for $K = 4$ (Fig. 5) was not completely congruent with the separation of chub populations into four main mtDNA phylogenetic lineages, i.e. the Aegean population from the Struma River was associated with geographically close populations of the Adriatic region, while Iberian populations belonging to the Aegean mtDNA lineage formed a separate group. At $K = 6$, however, the STRUCTURE model showed exactly the same differentiation as the network analysis of mtDNA. The best model of genetic structure using microsatellites ($K = 7$) additionally showed the subdivision of Danube Basin populations into two groups (i.e. central European and Balkan) (Fig. 1).

Roughly congruent phylogeographical patterns in mtDNA and nuclear markers were also found in other freshwater fishes, e.g. European grayling (Gum, 2009), bullhead (Englbrecht *et al.*, 2000; Hänfling *et al.*, 2002) and burbot (Barluenga *et al.*, 2006). However, contrasting results between both marker types have also occasionally been found (e.g. Bryja *et al.*, 2010a). This discordance may indicate past evolutionary processes undetectable with mtDNA, such as hybridization among highly divergent lineages in contact zones, perhaps resulting in mtDNA introgression or replacement.

Inferences on the evolutionary history of *Squalius cephalus* in Europe and the phylogeography of chub populations in the Mediterranean region

Primary freshwater fishes, due to their dependence upon water routes, most reliably reflect the historical geographical attributes of a region, including connections among hydrographic basins, and the process of isolation and interconnection among rivers and lakes (Levy *et al.*, 2009). On the basis of the comprehensive phylogenetic and biogeographical analyses of the Leuciscinae subfamily performed by Perea *et al.* (2010), the chub (*S. cephalus*) belongs to the Euroasiatic (Central European and Anatolian) group of the genus *Squalius*. The diversification of this group is dated in the Upper Miocene (9.16 Ma) and the existence of the four contemporary phylogenetic lineages

of chub in Europe is explained as the result of an ancient allopatric fragmentation that occurred during the late Pliocene (3–2.5 Ma) (Durand *et al.*, 2000). Therefore, the current distribution and phylogenetic structure of chub in Europe may be connected with several historical events, such as the Messian Lago Mare stage, the rise of humidity at the beginning of the Pliocene in the Mediterranean region, and the glaciations and uplift of the Alps during the Pleistocene (Perea *et al.*, 2010). Nevertheless, it is generally accepted that the phylogeographical structure of widely distributed European freshwater fish species is a result of Pleistocene glacial periods. During the extremely cold last glaciation (115–10 ka), most central and western European basins were covered by ice. Consequently, most freshwater fish species disappeared from these areas and survived in several different refugia, including the lower Danube, around the Black Sea, and in the Aegean region (see Table 5 summarizing phylogeographical studies of freshwater fish species). The subsequent colonization of central and western Europe from these refugia explains the recent origin and expansion of lineages in central and western Europe. However, these lineages could not have reached the Mediterranean peninsulas – probably as a result of the uplift of the Alps during the Pleistocene – explaining the divergence of Mediterranean populations. It is highly probably that Alpine mountains prevented Mediterranean populations from moving northwards; therefore, the populations from the Mediterranean area did not contribute to the post-Pleistocene colonization of non-Mediterranean Europe (Durand *et al.*, 1999a; Perea *et al.*, 2010).

The evolutionary history of the European chub is consistent with this scenario. Being unable to survive in periglacial areas of Europe, the species colonized the northern part of its distribution recently. Our network analysis indicated at least two foci that formed star-like phylogenies, typical for sudden population expansion, such as those that would accompany post-glacial colonization. Additionally, both nuclear and mitochondrial markers showed the high divergence of Mediterranean (Adriatic, Aegean–Balkan, and Aegean–Spain) lineages of chub from non-Mediterranean lineages. This indicates their long independent evolutionary history linked to the different geological history of freshwater bodies in the Mediterranean region. Likewise, we observed high nucleotide diversity for both lineages, which is typical for freshwater species living in non-glaciated areas (Durand *et al.*, 1999a) and long-term demographic stability (i.e. multimodal mismatch distribution).

Durand *et al.* (1999a) proposed the existence of two southern refugia for chub in the Mediterranean area, i.e. the Adriatic side of the Balkans and Aegean rivers (Fig. 7). As was previously mentioned, microsatellite markers showed slightly inconsistent results regarding the relationships among Mediterranean populations contrary to mtDNA. This could contribute to a better understanding of the evolutionary history of chub in this region. First, the link between the Struma population belonging to the Aegean lineage and the populations of Adriatic lineage revealed by nuclear markers in this study (see STRUCTURE analysis for $K = 4$; Fig. 6) may indicate the location of the refugium for the Italian populations on the Adriatic side of the Balkans (*sensu* Durand *et al.*, 1999a) or may be a result of trans-Adriatic exchanges between the western Balkans and eastern Italy during the Würmian regression (Fig. 7) (Levy *et al.*, 2009). Generally, the ichthyofauna of northern Italy and the northern rivers of the Balkan Peninsula show high affinities, which is explained by more recent events, such as expansion of the Po River during the Last Glacial Maximum (*c.* 20 ka) as a result of a drop in sea level (Perea *et al.*, 2010). Second, the close phylogenetic affinity of Iberian populations of chub to the Aegean lineage was revealed by analyses of mitochondrial cyt *b* (Zardoya & Doadrio, 1999; Durand *et al.*, 2000; Sanjur *et al.*, 2003; Seifertová *et al.*, 2008), but strong differentiation of Iberian populations at nuclear microsatellites may indicate their independent histories as well as the possible existence of a Pleistocene glacial refugium in the Iberian Peninsula. The fact that the Aegean lineage has the oldest most recent common ancestor further supports its complicated history, which may include, for example, strong founder events followed by lineage sorting. It could also be the result of ancient introgression or hybridization with several highly endemic *Squalius* species, mostly with parapatric distributions in this region, and of very limited connections between adjacent areas in France where the Western–Atlantic mtDNA lineage is distributed (Fig. 1). This supports the study of Durand *et al.* (2000), which suggested that hybridization between the chub and endemic *Squalius* species restricted to Mediterranean areas may be considered a widespread evolutionary phenomenon. Therefore, additional analyses of different DNA markers are necessary in the future to elucidate the phylogenetic relationships between the Iberian and Aegean populations.

Biogeography of Central/Western/Eastern European populations and post-glacial colonization of non-Mediterranean Europe by chub

The general scenario of the colonization of non-Mediterranean Europe by freshwater fishes comprises two main routes, i.e. the 'eastern colonization route' including the northward spread of lineages from the Ponto-Caspian refugia, and the 'western colonization route' often occurring in two waves (Durand *et al.*, 1999a; Kotlik & Berrebi, 2001; Bryja *et al.*, 2010a; review in Table 5). Partitioning of the Eastern populations of European chub in both analyses (mtDNA and microsatellites) clearly confirmed the survival of the Eastern lineage during Pleistocene glaciations in a separate refugium. Durand *et al.* (1999a) postulated the location of this refugium to be on the periphery of the Black and Caspian seas (Fig. 7), whence this Eastern lineage probably entered the Baltic areas as far as the Oder and northern Elbe during the Holocene. Low mtDNA and microsatellite variability, partial star-like phylogeny with a young most recent common ancestor, and unimodal mismatch distribution observed for the Eastern lineage reflects the recent origin of the colonization of the Baltic drainages. The Ponto-Caspian region was also identified as a separate

refugium for other freshwater fishes with distribution and environmental preferences similar to chub, e.g. spined loach (*Cobitis taenia*), grayling (*Thymallus thymallus*), bitterling (*Rhodeus amarus*), perch (*Perca fluviatilis*) and catfish (*Silurus glanis*) (Table 5).

Likewise, confirmation of the existence of the Western–Danubian lineage by both markers is in accordance with the previously reported importance of the Danubian refugium (lower Danube) for the survival of European freshwater fish fauna during glacial periods, e.g. barbell (*Barbus barbus*), varione (*Telestes souffia*), bitterling, perch and grayling (Table 5, Fig. 7). However, not all freshwater fish had refugia in the lower Danube area. For example, recent dispersion from a single glacial refugium around the Ponto-Caspian region is proposed for the European catfish, and several refugia in the Ponto-Caspian area are proposed as sources for the colonization of central Europe by the spined loach, as the Danube is not part of its current distribution (Table 5).

The observed isolation of the Rhone populations from the ancestral Western lineage, using mtDNA and microsatellites in the present study, supports the two-step expansion scenario from the Danubian refuge and the existence of a secondary glacial (Atlantic) refugium probably located in the Rhone Basin during the Last Glacial Maximum. The present results show that the Western–Atlantic lineage recently underwent rapid population expansion as indicated by demographic analyses and the star-like shape of the haplotype network. This is also supported by the greater reduction in mtDNA genetic diversity in this lineage relative to the Western–Danubian lineage. The predominance of a single haplotype (i.e. H12) found over a wide geographical area is more compatible with non-equilibrium conditions such as previous habitat reduction or with population bottlenecks due to glacial episodes. A similar pattern of low genetic diversity within the Western lineage identified using mtDNA was observed in other fish species, e.g. burbot (*Lota lota*) (Barluenga *et al.*, 2006). Therefore, it seems that reduction of genetic diversity could represent a general pattern in western European rivers.

The next typical feature of post-glacial dispersion common for European freshwater fishes is that the main mtDNA lineages came into secondary contact in the same central European areas (Bernatchez, 2001; Kotlik & Berrebi, 2001; Gum *et al.*, 2005; Bryja *et al.*, 2010a). Our mtDNA and microsatellite data do not provide evidence for secondary contact between Western and Eastern lineages in central Europe, which is in contrast with Durand *et al.* (1999a), who reported the Western–Atlantic, Western–Danubian and Eastern mtDNA lineages of chub co-occurring in the Elbe Basin. The absence of signs of admixture could be caused by the low number of sites sampled in the present study in the areas of potential contact zones (i.e. only one locality is sampled from the Elbe River). Following previous studies (Durand *et al.*, 1999a, 2000), the populations of the Baltic Basin belong to the Eastern lineage. This assumption was not confirmed in our study, as both markers revealed a Western origin for the Oder population belonging to the Baltic Basin. Human factors (e.g. accidental transport related to aquaculture) or geographical proximity to Western lineage populations are the most plausible explanations for this finding. The same pattern, i.e. the Danubian origin of fish in the upper Oder River, was recently also found in the European bitterling (*Rhodeus amarus*) (J. Bryja & M. Reichard, unpublished data).

Similar to the work of Durand *et al.* (1999a), analyses of nuclear as well as mtDNA data revealed no introgression between two Bulgarian populations (the River Struma and the lower Danube). This regional clustering could reflect different colonization events or may be the result of limited connection between these areas due to geographical barriers to dispersal. In the bitterling (Bryja *et al.*, 2010a), population structure analysis using microsatellites has shown separation of the south-eastern Danubian population in Bulgaria from other Danubian populations in central Europe, which is in accordance with our results (see Fig. 6, $K = 7$). However, the Danubian population of *R. amarus* in Bulgaria formed a cluster with the geographically proximate Bulgarian population from the River Struma, which contrasts with our results.

Taxonomic implication

Our comprehensive sampling revealed six major lineages within the European chub. Comparing these data with previously published patterns for congeneric species, we found that some of these taxa corresponded to lineages as defined in this paper, whereas others remained separated (Fig. 4). Namely, *Squalius orientalis* was included in the Eastern lineage, *Squalius laietanus* in the Aegean–Spain, *Squalius* cf. *orpheus* in the Aegean–Balkan, *Squalius prespensis* and *Squalius albus* in the Adriatic, and *Squalius vardarensis* in the Western–Danubian (Zardoya & Doadrio, 1999; Durand *et al.*, 2000; Freyhof *et al.*, 2005; Perea *et al.*, 2010). The Western–Atlantic lineage did not contain any previously published sequences attributed to a different species. Additional taxa formed branches separate from lineages recognized here (Zardoya & Doadrio, 1999; Durand *et al.*, 2000; Perea *et al.*, 2010).

Our Bayesian coalescent analysis showed that the model assuming a phylogenetic dataset that used the Yule process to model species diversification and extinction was better, but not strongly favoured, compared with the model that assumes an intra-specific dataset with constant population size. This supports our treatment of all included populations as one species. Given the complexity of information available today and the lack of consensus in the literature, we believe that a thorough systematic revision of the *S. cephalus* complex is warranted to establish possible taxonomic ranks or synonymy of the multiple taxa.

ACKNOWLEDGEMENTS

We thank all anonymous referees for comments that greatly improved the manuscript, and Pavel Jurajda, Markéta Ondračková, Teodora Trichkova, Milen Vassilev, Alexis Ribas Salvador, Antoni Arrizabalaga, André Gilles, René Chappaz,

Paolo Galli, Jussi Pennanen, Miroslaw Przybylski and Grziegorz Zieba for their help with fish sampling in the different sampling localities and the opportunity to use their laboratories. This study was funded by the Czech Science Foundation, project no. 524/07/0188. Travel costs for the different field studies and material collections were funded by the Research Project of the Masaryk University, Brno, project no. MSM 0021622 416 and also supported by the Ichthyoparasitology Centre of Excellence, project no. LC 522, funded by the Ministry of Education, Youth and Sports of the Czech Republic. J.B. was supported by the Biodiversity Research Centre (LC06073). The bioinformatic analyses (Bayesian coalescent analysis) were conducted at the computational cluster of the Institute of Vertebrate Biology, AS CR, Brno. We are very grateful to Matthew Nicholls for the English revision of the draft.

REFERENCES

Avise, J.C. (1991) Ten unorthodox perspectives on evolution prompted by comparative population genetic findings on mitochondrial DNA. *Annual Review of Genetics*, **25,** 45–69.

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A. & Saunders, N.C. (1987) Intraspecific phylogeography – the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, **18,** 489–522.

Bandelt, H.J., Forster, P. & Rohl, A. (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16,** 37–48.

Barluenga, M., Sanetra, M. & Meyer, A. (2006) Genetic admixture of burbot (Teleostei: *Lota lota*) in Lake Constance from two European glacial refugia. *Molecular Ecology*, **15,** 3583–3600.

Belkhir, K., Borsa, P., Goudet, J., Chikhi, L. & Bonhomme, F. (2004) *Genetix version. 4.0, logiciel sous Windows TM pour la genetique des populations*. Laboratoire Genome et Populations, CNRS UPR 9060, Université de Montpellier II, Montpellier, France.

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B: Methodological*, **57,** 289–300.

Bernatchez, L. (2001) The evolutionary history of brown trout (*Salmo trutta* L.) inferred from phylogeographic, nested clade, and mismatch analyses of mitochondrial DNA variation. *Evolution*, **55,** 351–379.

Bohlen, J., Šlechtová, V., Bogutskaya, N. & Freyhof, J. (2006) Across Siberia and over Europe: phylogenetic relationships of the freshwater fish genus *Rhodeus* in Europe and the phylogenetic position of *R. sericeus* from the River Amur. *Molecular Phylogenetics and Evolution*, **40,** 856–865.

Böhme, M. & Ilg, A. (2003) *fosFARbase*. Available at: http://www.wahre-staerke.com/ (accessed 25 May 2011).

Bryja, J., Smith, C., Konečný, A. & Reichard, M. (2010a) Range-wide population genetic structure of the European bitterling (*Rhodeus amarus*) based on microsatellite and mitochondrial DNA analysis. *Molecular Ecology*, **19,** 4708–4722.

Bryja, J., Granjon, L., Dobigny, G., Patzenhauerová, H., Konečný, A., Duplantier, J.M., Gauthier, P., Colyn, M., Durnez, L., Lalis, A. & Nicolas, V. (2010b) Plio-Pleistocene history of West African Sudanian savanna and the phylogeography of the *Praomys daltoni* complex (Rodentia): the environmental/geography/genetic interplay. *Molecular Ecology*, **19,** 4783–4799.

Chapuis, M.P. & Estoup, A. (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, **24,** 621–631.

Culling, M.A., Janko, K., Boron, A., Vasil'ev, V.P., Cote, I.M. & Hewitt, G.M. (2006) European colonization by the spined loach (*Cobitis taenia*) from Ponto-Caspian refugia based on mitochondrial DNA variation. *Molecular Ecology*, **15,** 173–190.

Degnan, S.M. (1993) The perils of single-gene trees – mitochondrial versus single-copy nuclear-DNA variation in white-eyes (Aves, Zosteropidae). *Molecular Ecology*, **2,** 219–225.

Dowling, T.E., Tibbets, C.A., Minckley, W.L. & Smith, G.R. (2002) Evolutionary relationships of the Plagopterins (Teleostei: Cyprinidae) from cytochrome *b* sequences. *Copeia*, **3,** 665–678.

Drummond, A.J. & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7,** 214.

Durand, J.D., Persat, H. & Bouvet, Y. (1999a) Phylogeography and postglacial dispersion of the chub (*Leuciscus cephalus*) in Europe. *Molecular Ecology*, **8,** 989–997.

Durand, J.D., Templeton, A.R., Guinand, B., Imsiridou, A. & Bouvet, Y. (1999b) Nested clade and phylogeographic analyses of the chub, *Leuciscus cephalus* (Teleostei, Cyprinidae), in Greece: implications for Balkan peninsula biogeography. *Molecular Phylogenetics and Evolution*, **13,** 566–580.

Durand, J.D., Unlu, E., Doadrio, I., Pipoyan, S. & Templeton, A.R. (2000) Origin, radiation, dispersion and allopatric hybridization in the chub (*Leuciscus cephalus*). *Proceedings of the Royal Society B: Biological Sciences*, **267,** 1687–1697.

Englbrecht, C.C., Freyhof, J., Nolte, A., Rassmann, K., Schliewen, U. & Tautz, D. (2000) Phylogeography of the bullhead *Cottus gobio* (Pisces: Teleostei: Cottidae) suggests a pre-Pleistocene origin of the major central European populations. *Molecular Ecology*, **9,** 709–722.

Estoup, A., Rousset, F., Michalakis, Y., Cornuet, J.M., Adriamanga, M. & Guyomard, R. (1998) Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). *Molecular Ecology*, **7,** 339–353.

Evanno, G., Regnaut, S. & Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14,** 2611–2620.

Excoffier, L., Laval, G. & Schneider, S. (2005) Arlequin (version 3.0): an integrated software package for population

genetics data analysis. *Evolutionary Bioinformatics*, **1,** 47–50.

Flanders, J., Jones, G., Benda, P., Dietz, Ch., Zhang, S., Li, G., Sharifi, M. & Rossiter, S.J. (2009) Phylogeography of the greater horseshoe bat, *Rhinolophus ferrumequinum*: contrasting results from mitochondrial and microsatellite data. *Molecular Ecology*, **18,** 306–318.

Freyhof, J., Lieckfeldt, D., Pitra, C. & Ludwig, A. (2005) Molecules and morphology: evidence for introgression of mitochondrial DNA in Dalmatian cyprinids. *Molecular Phylogenetics and Evolution*, **37,** 347–354.

Fu, Y.X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147,** 915–925.

Fu, Y.X. & Li, W.H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133,** 693–709.

Goudet, J. (1995) FSTAT (version 1.2): a computer program to calculate F-statistics. *Journal of Heredity*, **86,** 485–486.

Gum, B. (2009) Conservation genetics and management implications for European grayling, *Thymallus thymallus*: synthesis of phylogeography and population genetics. *Fisheries Management and Ecology*, **16,** 37–51.

Gum, B., Gross, R. & Kuehn, R. (2005) Mitochondrial and nuclear DNA phylogeography of European grayling (*Thymallus thymallus*): evidence for secondary contact zones in Central Europe. *Molecular Ecology*, **14,** 1707–1725.

Hall, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41,** 95–98.

Hänfling, B. & Brandl, R. (1998) Genetic and morphological variation in a common European cyprinid, *Leuciscus cephalus* within and across central European drainages. *Journal of Fish Biology*, **52,** 706–715.

Hänfling, B., Hellemans, B., Volckaert, F.A.M. & Carvalho, G.R. (2002) Late glacial history of the cold-adapted freshwater fish *Cottus gobio*, revealed by microsatellites. *Molecular Ecology*, **11,** 1717–1729.

Hänfling, B., Dumpelmann, C., Bogutskaya, N.G., Brandl, R. & Brandle, M. (2009) Shallow phylogeographic structuring of *Vimba vimba* across Europe suggests two distinct refugia during the last glaciation. *Journal of Fish Biology*, **75,** 2269–2286.

Hardy, O.J. & Vekemans, X. (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2,** 618–620.

Hardy, O.J., Charbonnel, N., Fréville, H. & Heuertz, M. (2003) Microsatellite allele sizes: a simple test to assess their significance on genetic differentiation. *Genetics*, **163,** 1467–1482.

Harpending, H.C. (1994) Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Human Biology*, **66,** 591–600.

Hewitt, G.M. (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society*, **58,** 247–276.

Hewitt, G.M. (1999) Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, **68,** 87–112.

Ho, S.Y.W., Phillips, M.J., Cooper, A. & Drummond, A.J. (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular Biology and Evolution*, **22,** 1561–1568.

Jakobsson, M. & Rosenberg, N.A. (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23,** 1801–1806.

Ketmaier, V., Bianco, P.G., Cobollia, M., Krivokapic, M., Caniglia, R. & De Matthaeis, E. (2004) Molecular phylogeny of two lineages of Leuciscinae cyprinids (*Telestes* and *Scardinius*) from the peri-Mediterranean area based on cytochrome *b* data. *Molecular Phylogenetics and Evolution*, **32,** 1061–1071.

Kotlik, P. & Berrebi, P. (2001) Phylogeography of the barbel (*Barbus barbus*) assessed by mitochondrial DNA variation. *Molecular Ecology*, **10,** 2177–2185.

Kottelat, M. & Freyhof, J. (2007) *Handbook of European freshwater fishes*. Kottelat, Cornol, Switzerland.

Krieg, F., Triantafyllidis, A. & Guymard, R. (2000) Mitochondrial DNA variation in European populations of *Silurus glanis*. *Journal of Fish Biology*, **56,** 713–724.

Laroche, J., Durand, J.D., Bouvet, Y., Guinand, B. & Brohon, B. (1999) Genetic structure and differentiation among populations of two cyprinids, *Leuciscus cephalus* and *Rutilus rutilus*, in a large European river. *Canadian Journal of Fisheries and Aquatic Sciences*, **56,** 1659–1667.

Levy, A., Doadrio, I. & Almada, V.C. (2009) Historical biogeography of European leuciscins (Cyprinidae): evaluating the Lago Mare dispersal hypothesis. *Journal of Biogeography*, **36,** 55–65.

Librado, P. & Rozas, J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25,** 1451–1452.

Lundberg, J.G. (1993) African–South American freshwater fish clades and continental drift, problems with a paradigm. *Biotic relationships between Africa and South America* (ed. by P. Goldblatt), pp. 156–198. Yale University Press, New Haven, CT.

Makinen, H.S., Cano, J.M. & Merila, J. (2006) Genetic relationships among marine and freshwater populations of the European three-spined stickleback (*Gasterosteus aculeatus*) revealed by microsatellites. *Molecular Ecology*, **15,** 1519–1534.

Nei, M. (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89,** 583–590.

Nesbo, C.L., Fossheim, T., Vollestad, L.A. & Jakobsen, K.S. (1999) Genetic divergence and phylogeographic relationships among European perch (*Perca fluviatilis*) populations reflect glacial refugia and postglacial colonization. *Molecular Ecology*, **8,** 1387–1404.

Perea, S., Bohme, M., Zupancic, P., Freyhof, J., Sanda, R., Ozulug, M., Abdoli, A. & Doadrio, I. (2010) Phylogenetic

relationships and biogeographical patterns in Circum-Mediterranean subfamily Leuciscinae (Teleostei, Cyprinidae) inferred from both mitochondrial and nuclear data. *BMC Evolutionary Biology*, **10,** 265.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155,** 945–959.

Rambaut, A. & Drummond, A.J. (2007) *Tracer v1.4*. Available at: http://beast.bio.ed.ac.uk/Tracer (accessed 20 May 2011).

Rand, D.M. (1996) Neutrality tests of molecular markers and the connection between DNA polymorphism, demography, and conservation biology. *Conservation Biology*, **10,** 665–671.

Raymond, M. & Rousset, F. (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86,** 248–249.

Rodriguez, F., Pérez, T., Hammer, S.E., Albornoz, J. & Dominguez, A. (2010) Integrating phylogeographic patterns of microsatellite and mtDNA divergence to infer the evolutionary history of chamois (genus *Rupicapra*). *BMC Evolutionary Biology*, **10,** 222.

Rogers, A.R. & Harpending, H. (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, **9,** 552–569.

Rosenberg, N.A. (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes*, **4,** 137–138.

Salzburger, W., Brandstatter, A., Gilles, A., Parson, W., Hempel, M., Sturmbauer, C. & Meyer, A. (2003) Phylogeography of the vairone (*Leuciscus souffia*, Risso 1826) in Central Europe. *Molecular Ecology*, **12,** 2371–2386.

Sanjur, O.I., Carmona, J.A. & Doadrio, I. (2003) Evolutionary and biogeographical patterns within Iberian populations of the genus *Squalius* inferred from molecular data. *Molecular Phylogenetics and Evolution*, **29,** 20–30.

Schmidt, T.R. & Gold, J.R. (1993) Complete sequence of the mitochondrial cytochrome *b* gene in the cherryfin shiner, *Lythrurus roseipinnis* (Teleostei, Cyprinidae). *Copeia*, **3,** 880–883.

Seifertová, M., Vyskočilová, M., Morand, S. & Šimková, A. (2008) Metazoan parasites of freshwater cyprinid fish (*Leuciscus cephalus*): testing biogeographical hypotheses of species diversity. *Parasitology*, **135,** 1417–1435.

Slatkin, M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139,** 457–462.

Storey, J.D., Taylor, J.E. & Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society B: Methodological*, **66,** 187–205.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123,** 585–595.

Triantafyllidis, A., Krieg, F., Cottin, C., Abatzopoulos, T.J., Triantaphyllidis, C. & Guyomard, R. (2002) Genetic structure and phylogeography of European catfish (*Silurus glanis*) populations. *Molecular Ecology*, **11,** 1039–1055.

Valqui, J., Hartl, G.B. & Zachos, F.E. (2010) Non-invasive genetic analysis reveals high levels of mtDNA variability in the endangered South-American marine otter (*Lontra felina*). *Conservation Genetics*, **11,** 2067–2072.

Vyskočilová, M., Šimková, A. & Martin, J.F. (2007) Isolation and characterization of microsatellites in *Leuciscus cephalus* (Cypriniformes, Cyprinidae) and cross-species amplification within the family Cyprinidae. *Molecular Ecology Notes*, **7,** 1150–1154.

Zardoya, R. & Doadrio, I. (1999) Molecular evidence on the evolutionary and biogeographical patterns of European cyprinids. *Journal of Molecular Evolution*, **49,** 227–237.

Zhang, D.X. & Hewitt, M. (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, **12,** 536–584.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Haplotypes of cytochrome *b* (600 bp) in *Squalius cephalus* analysed in this study.

Appendix S2 Geographical distribution and accession numbers of cytochrome *b* haplotypes (Table S2.1), and genetic variability across microsatellite loci genotyped (Table S2.2) in *Squalius cephalus.*

Appendix S3 Results of the STRUCTURE analysis for $K = 1$–10 for *Squalius cephalus*.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

BIOSKETCH

This work is a part of the PhD study of **Mária Seifertová** on host–parasite interactions in a geographical context using fishes as model hosts. The research team included specialists working on molecular ecology, phylogeny and phylogeography, and modelling of DNA sequence evolution.

Author contributions: M.S., J.B. and A.Š. conceived the ideas; M.S. and A.Š. collected the data; M.S., J.B., N.M. and M.V. analysed the data; M.S., J.B., N.M. and A.Š. wrote the manuscript.

Editor: Brett Riddle

Paper 2.2.3

Martínková N., Barnett R., Cucchi T., Struchen R., Pascal M., Pascal M., Fischer M. C., Higham T., Brace S., Ho S. Y. W., Quéré J.-P., O'Higgins P., Excoffier L., Heckel G., Hoelzel A. R., Dobney K. M., Searle, J. B. 2013. Divergent evolutionary processes associated with colonization of offshore islands. *Molecular Ecology* 22: 5205-5220.

MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 5205–5220 doi: 10.1111/mec.12462

Divergent evolutionary processes associated with colonization of offshore islands

NATÁLIA MARTÍNKOVÁ,*†[1] ROSS BARNETT,*‡§[1] THOMAS CUCCHI,§¶**[1] RAHEL STRUCHEN,†† MARINE PASCAL,‡‡ MICHEL PASCAL,‡‡ MARTIN C. FISCHER,††§§ THOMAS HIGHAM,¶¶ SELINA BRACE,*** SIMON Y. W. HO,††† JEAN-PIERRE QUÉRÉ,‡‡‡ PAUL O'HIGGINS,§§§ LAURENT EXCOFFIER,††§§ GERALD HECKEL,††§§ A. RUS HOELZEL,‡ KEITH M. DOBNEY§¶[2] and JEREMY B. SEARLE*¶¶¶[2]

*Department of Biology, University of York, Wentworth Way, York YO10 5DD, UK, †Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, Květná 8, Brno 603 65, Czech Republic, ‡School of Biological and Biomedical Sciences, Durham University, South Road, Durham, DH1 3LE, UK, §Department of Archaeology, Durham University, South Road, Durham, DH1 3LE, UK, ¶Department of Archaeology, University of Aberdeen, St Mary's, Elphinstone Road, Aberdeen AB24 3UF, UK, **Muséum national d'histoire naturelle, case postale 56 (bâtiment d'anatomie comparée), 55 rue Buffon, F-75231 Paris Cedex 05, Paris, France, ††Computational and Molecular Population Genetics (CMPG), Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, Bern CH-3012, Switzerland, ‡‡Équipe Écologie des Invasions Biologiques, INRA, Campus de Beaulieu 35 000, Rennes Cedex, France, §§Swiss Institute of Bioinformatics, Genopode, Lausanne CH-1015, Switzerland, ¶¶Oxford Radiocarbon Accelerator Unit, Research Laboratory for Archaeology and the History of Art (RLAHA), Dyson Perrins Building, South Parks Road, Oxford OX1 3QY, UK, ***School of Biological Sciences, Royal Holloway, University of London, Egham TW20 0EX, UK, †††School of Biological Sciences, University of Sydney, Sydney, NSW 2006, Australia, ‡‡‡UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), INRA, Campus international de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez cedex, France, §§§Centre for Anatomical and Human Sciences, Hull York Medical School (HYMS), University of York, John Hughlings Jackson Building, York YO10 5DD, UK, ¶¶¶Department of Ecology and Evolutionary Biology, Corson Hall, Cornell University, Ithaca, NY 14853-2701, USA*

Abstract

Oceanic islands have been a test ground for evolutionary theory, but here, we focus on the possibilities for evolutionary study created by offshore islands. These can be colonized through various means and by a wide range of species, including those with low dispersal capabilities. We use morphology, modern and ancient sequences of cytochrome *b* (*cytb*) and microsatellite genotypes to examine colonization history and evolutionary change associated with occupation of the Orkney archipelago by the common vole (*Microtus arvalis*), a species found in continental Europe but not in Britain. Among possible colonization scenarios, our results are most consistent with human introduction at least 5100 BP (confirmed by radiocarbon dating). We used approximate Bayesian computation of population history to infer the coast of Belgium as the possible source and estimated the evolutionary timescale using a Bayesian coalescent approach. We showed substantial morphological divergence of the island populations, including a size increase presumably driven by selection and reduced microsatellite variation likely reflecting founder events and genetic drift. More surprisingly, our results suggest that a recent and widespread *cytb* replacement event in the continental source area purged *cytb* variation there, whereas the ancestral diversity is largely retained in the colonized islands as a genetic 'ark'. The replacement event in the continental *M. arvalis* was probably triggered

Correspondence: Jeremy B. Searle, Fax: +1-607-255-8088; E-mail: jeremy.searle@cornell.edu

[1]These authors contributed equally.

[2]These authors contributed equally.

by anthropogenic causes (land-use change). Our studies illustrate that small offshore islands can act as field laboratories for studying various evolutionary processes over relatively short timescales, informing about the mainland source area as well as the island.

Keywords: demographic analysis, genetic replacement, island colonization, *Microtus arvalis*, phylogeography

Received 5 May 2013; revision received 2 July 2013; accepted 9 July 2013

Introduction

Marine islands were pivotal settings for the development of evolutionary theory (Wallace 1880) and continue to inspire evolutionary biologists today (Grant 2008). They tend to be ecologically simple and distinctive compared with mainland settings, with fewer competitors and predators, a different food supply and environmental conditions and a smaller living space, all factors that promote rapid and perhaps substantial evolutionary change relative to mainland source populations. Studies on islands have therefore provided insights into evolutionary processes such as species radiations (Emerson 2002) and the generation of evolutionary novelty (Lachaise *et al*. 2000).

Marine islands are also interesting for the varied ways in which they can be colonized. First, for islands separating from a mainland, some species may already be present at island formation. Second, islands may be colonized by natural 'sweepstake' dispersal (Simpson 1940). Finally, species may be introduced into islands by humans.

Many of the classic systems for the study of island evolution, particularly species radiations, have involved oceanic islands or archipelagos colonized by sweepstake dispersal, for example Darwin's finches on the Galapagos and Hawaiian drosophila. However, here we consider offshore islands, which can be colonized by all three mechanisms, providing rich opportunities for evolutionary studies, notably of poor dispersers such as small mammals (Berry 1996).

Offshore islands around the Atlantic coast of Europe were formed by end-glacial sea level rise, and thus, small mammals currently found on them may have already been present prior to island separation, and perhaps even before the last glacial maximum (LGM; *c*. 20 000 BP: Ehlers & Gibbard 2004), if local conditions were compatible with survival (Beirne 1952; Martínková *et al*. 2007). Alternatively, the small mammals could have been introduced by humans, for example, within bedding/fodder for livestock (Corbet 1961). Sweepstake dispersal, by rafting on floating masses of vegetation (Morrison *et al*. 1992) or walking over ice (White & Searle 2008), is also a possibility.

In this study, we aim to distinguish between these different modes of colonization of a European offshore archipelago (Orkney) by a small mammal (the common vole *Microtus arvalis*) and to document evolutionary change in that species postcolonization. *Microtus arvalis* is a ground-living, grass-eating small rodent that is common in agricultural meadows (Niethammer & Krapp 1982). The species is widespread in western continental Europe, although it is absent from Britain and Fennoscandia (Fig. 1a). Therefore, the whole British landmass separates the *M. arvalis* populations in continental Europe and on Orkney. The only species of *Microtus* in mainland Britain and other islands off mainland Britain is *Microtus agrestis*, which, like *M. arvalis*, has a wide continental European range (Shenbrot & Krasnov 2005). The two species belong to unrelated lineages within *Microtus* (Fink *et al*. 2010; Martínková & Moravec 2012).

The colonization history of a population can be inferred through comparisons with potential source populations. If the presence of the Orkney vole is explained by glacial survival or sweepstake dispersal, then surviving representatives of the source populations are no longer available, because there are no extant populations suitably close to Orkney. However, human introduction can come from a distance, and the most likely source area is represented by the closest coastal part of the continental European range of the same mitochondrial lineage as found on Orkney, that is, northern France and Belgium (Haynes *et al*. 2003; Tougard *et al*. 2008). For this reason, we focus on this potential source area in our analyses and consider whether such a human introduction is feasible or whether one of the other modes of colonization is more likely.

To explore the colonization history of the Orkney vole and the evolutionary change since colonization, we compare populations from Orkney and the potential source area using mitochondrial DNA sequences and microsatellites. We also use archaeological deposits of the species from Orkney for ancient DNA (aDNA) sequencing and radiocarbon dating (two Orkney vole bones have previously been radiocarbon-dated [Hedges *et al*. 1987] using technology now outdated), to obtain a more detailed temporal dimension to the colonization history and to

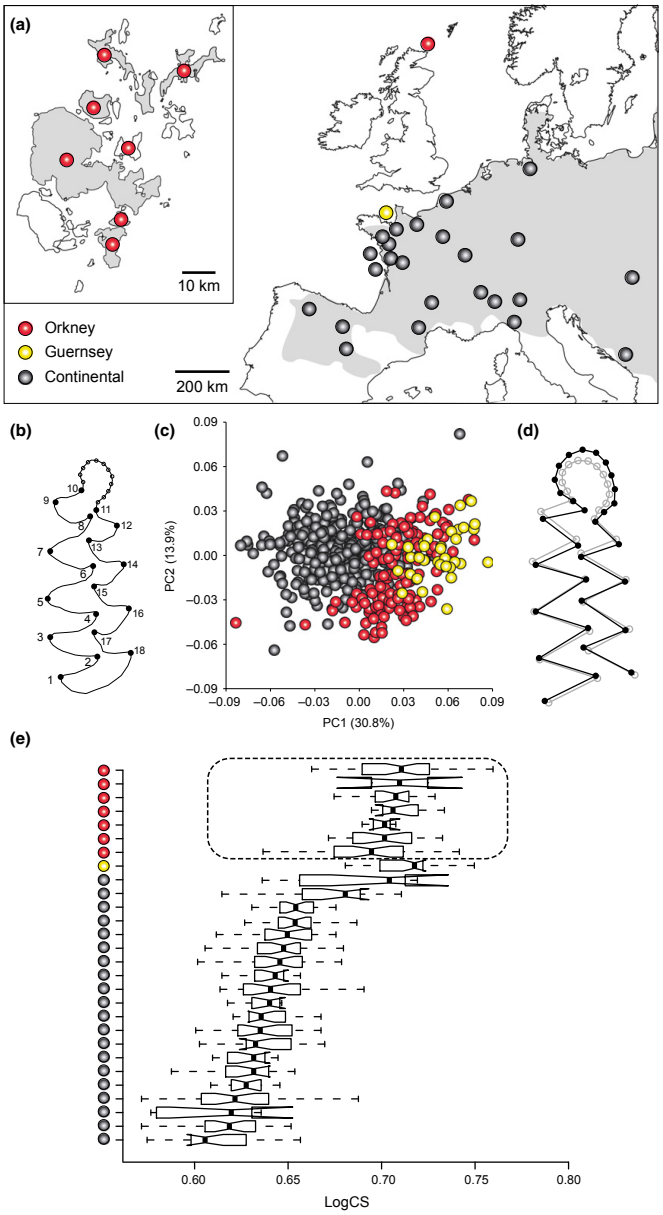


Fig. 1 The morphological variation of the first lower molar (M_1) among modern *Microtus arvalis* populations. (a) Sampling localities within its western and central European range (Shenbrot & Krasnov 2005) (in grey). Note that the three Spanish localities and the three Italian localities are treated as single entries in Fig. 1e and Table S1 (Supporting information). (b) Position of the 18 landmarks and 12 semi-landmarks on the M_1 occlusal view. (c) Scatter plot of the first two principal component scores depicting overall M_1 shape variation. (d) Diagram showing the M_1 shape change associated with the PC1 scores by 0.1 units in the positive direction (black) against the mean shape (grey). (e) Box plot showing log-transformed centroid size variation of the M_1.

support a demographic analysis. Finally, we use geometric morphometrics of a molar tooth as a marker for morphological evolution. This work provides a more sophisticated assessment of colonization and evolution of the Orkney vole than has previously been possible (Berry & Rose 1975; Corbet 1986; Haynes *et al*. 2003; Thaw *et al*. 2004; Gorman & Reynolds 2008). More generally, this case study provides new insights into not only short-term

Table 1 Microsatellite comparison of *Microtus arvalis* in Orkney and continental Europe, including diversity indices for all populations sampled and approximate Bayesian computation selection of most likely continental population for the Orkney colonization. For map of localities, see Fig. S1 (Supporting information)

Locality	Country/Orkney Island	Latitude	Longitude	Number of specimens	mtDNA lineage	A_R	H	Δ_C
Stalhille	Belgium	51.21	3.07	26	Western-North	5.69 (1.90)	0.72 (0.15)	2.03
Pihen lès Guînes	France	50.87	1.79	22	Western-North	7.02 (2.08)	0.78 (0.17)	2.24
Daubeuf	France	49.78	0.52	20	Western-North	6.65 (2.51)	0.73 (0.25)	2.24
Alflen	Germany	50.18	7.04	22	Western-North	6.62 (1.66)	0.80 (0.09)	2.27
Clérmont-Ferrand	France	45.78	3.08	23	Western-North	6.06 (2.21)	0.70 (0.23)	2.62
Avallon	France	47.85	3.90	22	Western-North	6.74 (1.84)	0.78 (0.13)	2.63
Aiffres	France	46.29	−0.41	22	Western-South	7.59 (3.34)	0.73 (0.27)	2.66
Fressenneville	France	50.07	1.58	24	Western-North	6.84 (2.23)	0.73 (0.23)	2.71
Thaon	France	49.26	−0.46	24	Western-North	7.19 (2.76)	0.74 (0.24)	2.72
Ste Marie du Mont	France	49.38	−1.23	20	Western-North	6.72 (2.88)	0.69 (0.28)	2.91
Schiltach	Germany	48.30	8.34	20	Western-North	5.30 (1.58)	0.74 (0.11)	2.94
Veurne	Belgium	51.07	2.66	26	Western-North	4.09 (0.95)	0.65 (0.11)	3.02
Baie d'Aiguillon	France	46.30	1.17	15	Western-South	7.76 (3.38)	0.74 (0.26)	—
St Jean le Thomas	France	48.73	−1.51	10	Western-North	4.07 (2.13)	0.59 (0.24)	—
Dinteloord	Netherlands	51.64	4.37	12	Central	4.68 (1.59)	0.66 (0.24)	—
Heerenveen	Netherlands	52.96	5.93	13	Central	6.04 (2.08)	0.75 (0.23)	—
Harray Stenness	Mainland, Orkney	59.02	−3.20	23	Western-North	5.01 (2.31)	0.62 (0.26)	—
Settiscarth	Mainland, Orkney	59.05	−3.10	26	Western-North	4.35 (1.88)	0.60 (0.27)	—
St Ola	Mainland, Orkney	58.94	−2.95	19	Western-North	3.79 (1.94)	0.50 (0.31)	—
Whitemill Bay	Sanday, Orkney	59.30	−2.55	19	Western-North	2.11 (1.60)	0.23 (0.26)	—
Wind Wick	S Ronaldsay, Orkney	58.76	−2.94	20	Western-North	1.83 (1.14)	0.21 (0.26)	—
Grimness	S Ronaldsay, Orkney	58.82	−2.91	20	Western-North	2.17 (1.21)	0.24 (0.26)	—
Loch of Swartmill	Westray, Orkney	59.29	−2.92	21	Western-North	2.16 (1.50)	0.26 (0.31)	—
Ness	Westray, Orkney	59.24	−2.87	23	Western-North	1.73 (1.19)	0.19 (0.29)	—

A_R: mean allelic richness over loci. H: mean heterozygosity over loci; values in parentheses are standard deviations. Δ_C: average distance between the observed and 1000 simulated summary statistics computed over the three Mainland Orkney populations for each continental population (c) to find the most likely source population for the colonization of Orkney by *M. arvalis* (smallest value). Only samples of 19 or more individuals were used for this analysis. The populations on Mainland Orkney were considered to best represent the colonized area, because of retention of high diversity (see also Fig. 3).

evolutionary changes in offshore islands, but also the mainland areas with which they are being compared.

Materials and methods

Specimens

Details of all *Microtus arvalis* specimens used in this study are listed in Tables 1 and 2 and Tables S1–S3 (Supporting information) and mapped in Figs 1 and 2 and Figs S1 and S2 (Supporting information). Geometric morphometrics were applied to 553 modern specimens from 27 localities, while 651 *M. arvalis* specimens from 70 localities were used for the analysis of modern DNA (125 specimens were used in both *cytb* and microsatellite studies). Ancient DNA analysis was conducted on 37 archaeological specimens from eight localities. Nineteen of these were among 23 archaeological specimens used for radiocarbon dating, calibrated using Calib Rev. 5.0.1 (Stuiver & Reimer 1993) and IntCal04 (Reimer *et al.* 2004), with dates before present (BP) standardized to 1950 AD.

Geometric morphometrics

Modern *M. arvalis* from populations in Orkney and continental Europe (Fig. 1a) were compared by geometric morphometrics of dental phenotype. Digital photographs of the first lower molar (M_1) were used to record two-dimensional Cartesian coordinates of 18 anatomical landmarks at the bases and tips of the lingual and labial cusps as well as 12 equidistant semi-landmarks on a manually drawn curve along the anterior loop (Fig. 1b).

Digitization of landmarks and semi-landmarks was performed using TPSdig 2 (Rohlf 2010a). Position, orientation and scaling information from the raw coordinates were standardized by a generalized Procrustes analysis (GPA) using TPSrelw 1.49 (Rohlf 2010b). To combine landmarks and semi-landmarks in the GPA,

Table 2 ^{14}C dates and their calibrated age ranges for 23 *Microtus arvalis* mandibles collected from Orkney

Laboratory number	Sample reference	Site name	Island	^{14}C age BP	95.4% (2s) cal age ranges
OxA 18324	R44	Point of Cott	Westray	4555 ± 40	cal BP 5050–5437
OxA 18782	R37	Point of Cott	Westray	4459 ± 33	cal BP 4967–5288
OxA 18325	R45	Point of Cott	Westray	4451 ± 38	cal BP 4884–5287
OxA-18668	R177	Quanterness	Mainland	4414 ± 27	cal BP 4869–5257
OxA-18784	R179	Quanterness	Mainland	4400 ± 33	cal BP 4861–5213
OxA-18786	R191	Skara Brae	Mainland	4199 ± 33	cal BP 4622–4844
OxA-20309	R84	Skara Brae	Mainland	4145 ± 29	cal BP 4574–4823
OxA-18664	R11	Skara Brae	Mainland	4124 ± 28	cal BP 4529–4815
OxA-18663	R3	Skara Brae	Mainland	3946 ± 27	cal BP 4294–4515
OxA-18787	R194	Skara Brae	Mainland	3939 ± 32	cal BP 4256–4514
OxA-18669	R193	Skara Brae	Mainland	3906 ± 27	cal BP 4249–4419
OxA-18785	R189	Skara Brae	Mainland	3884 ± 31	cal BP 4185–4418
OxA-18666	R23	Holm of Papa Westray	Westray	4089 ± 29	cal BP 4448–4808
OxA-18665	R20	Holm of Papa Westray	Westray	4054 ± 28	cal BP 4435–4784
OxA 18328	R126	Pierowall Quarry	Westray	4000 ± 45	cal BP 4298–4781
OxA 18783	R124	Pierowall Quarry	Westray	3824 ± 34	cal BP 4094–4406
OxA 18327	R99	Pierowall Quarry	Westray	3822 ± 38	cal BP 4092–4406
OxA-18350	R58	Howe	Mainland	1860 ± 28	cal BP 1721–1869
OxA-18351	R59	Howe	Mainland	1849 ± 27	cal BP 1714–1865
OxA-18667	R62	Howe	Mainland	1469 ± 24	cal BP 1308–1396
OxA-20310	RH3	Green Hill, South Walls	Hoy	1100 ± 24	cal BP 958–1060
OxA-20481	RH2	Green Hill, South Walls	Hoy	993 ± 27	cal BP 798–962
OxA 18326	R29	Earl's Bu	Mainland	966 ± 29	cal BP 795–932

the semi-landmarks along the anterior loop curve were constrained to slide along an estimated tangent at each sliding point using the bending energy method (Bookstein 1997).

The overall size parameter for the M_1 is centroid size (square root of the sum of squared distances of landmarks and semi-landmarks from the centroid) and comparisons between populations use a box plot of log-transformed values produced with 'R' v. 2.13.1 (R Development Core Team 2011). The shape variables are the Procrustes coordinates obtained after the GPA, and variation among populations is displayed by principal component analysis (PCA) using MorphoJ 1.05c (Klingenberg 2011).

Cytochrome b sequencing (modern DNA)

A total of 283 specimens of *M. arvalis* were sequenced from 18 localities in Orkney and 50 localities in continental Europe (particularly from around the potential area of human introduction: northern France and Belgium and inland from there). Total genomic DNA was isolated using the DNeasy Tissue Kit (Qiagen, Halden, Germany) and PCR-amplified using previously described primers (Table S4, Supporting information; Jaarola *et al*. 2004). PCR products were purified with QIAquick PCR purification kit (Qiagen) and commercially sequenced using BigDye Terminator Sequencing chemistry (Applied Biosystems, Foster City, CA, USA) with newly designed primers (Table S4, Supporting information) and run on ABI PRISM 3730xl sequencers (Applied Biosystems). Complete *cytb* sequences (1143 bp) were generated (GU190383–GU190665).

Cytochrome b sequencing (ancient DNA)

All aDNA extractions were performed in a laboratory in Durham where no modern molecular biology or post-PCR work is undertaken and where *Microtus* were analysed for the first time. Before use, all materials and work surfaces were wiped with 10% bleach, and the workspace was UV-irradiated overnight. Samples of *Microtus* bone were excised using a scalpel blade and crushed within aluminium foil. The bone powder was incubated overnight on a rotator at 55 °C in 500 μL of extraction buffer (0.5M pH 8.0 EDTA, 0.1M pH 8.0 Tris, 0.05% w/v SDS) with 8 μL of proteinase K (0.3 mg/mL). Digested samples were then extracted using the Qiagen QIAquick PCR purification method, as described in Nichols *et al*. (2007). Final elutions of aDNA were collected in 50 μL of TE buffer following the QIAquick protocol and stored at −20 °C. Negative extraction controls (lacking bone powder) were also performed in parallel in a ratio of approximately 1:10.

Ancient DNA was successfully obtained from 37 specimens (of 190 attempted) from Neolithic to Viking

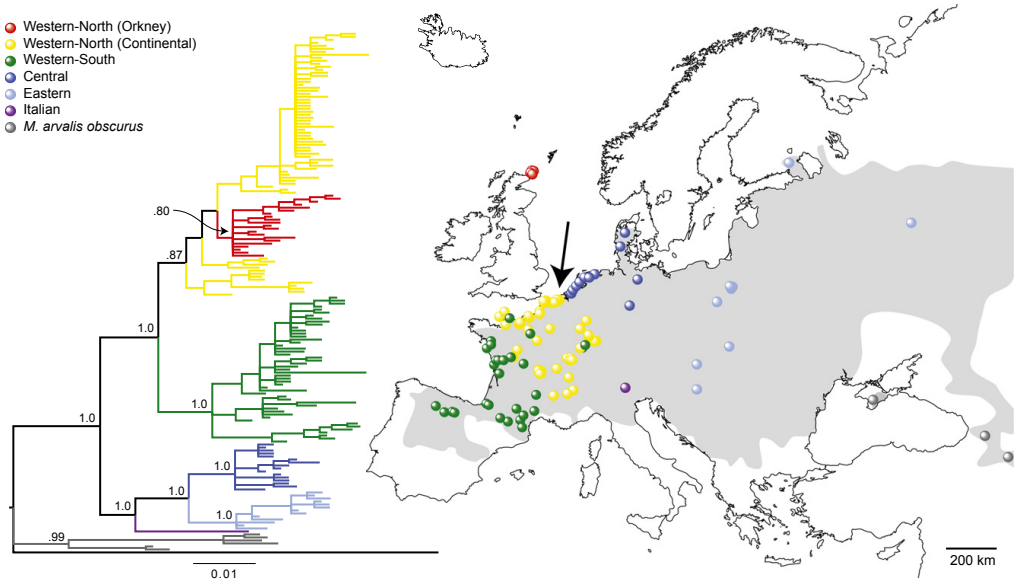


Fig. 2 Bayesian phylogenetic tree depicting all complete cytochrome *b* haplotypes of modern *Microtus arvalis* and localities where each phylogroup is found. Lineages labelled following previous authors (Haynes *et al.* 2003; Tougard *et al.* 2008): Orkney (red) within Western-North (yellow), Western-South (green), Central (dark blue), Eastern (light blue), Italian (purple), *M. a. obscurus* (grey). Posterior probability is presented for major clades only. Outgroup sequence in tree: *Microtus levis*, AY513819 (Jaarola *et al.* 2004). Grey shading: distribution of *M. arvalis* (Shenbrot & Krasnov 2005) where three *obscurus* localities are off the map to the east. Black arrow shows the most likely source locality of the Orkney voles, based on microsatellite data (see Table 1).

Age Orkney and Late Medieval Belgium (Table S3, Fig. S2, Supporting information). Partial *cytb* sequences (1130 bp) were generated through 8–9 overlapping PCRs (Table S4, Supporting information), with the missing *cytb* sequence at the 5′-end. Each PCR used 2 μL of aDNA extract in a 25 μL volume with Hi-Fidelity Platinum *Taq* (Invitrogen, Carlsbad, CA, USA). Purification and sequencing of PCR products followed the protocols used for modern DNA.

As further checks on the aDNA authenticity additional to the negative extraction and PCR controls, two samples were internally replicated from bone powder (R59 and R60), producing identical sequences. All sequences with unique mutations not found in modern Orkney voles were amplified at least twice for confirmation. The regions of overlap between the PCR products were checked and found to be consistent. The assembled *cytb* coding regions for all ancient voles produced a readable protein sequence with no frameshift or nonsense mutations. *M. arvalis* is known to possess a nuclear copy of *cytb* (DeWoody *et al.* 1999) that is easily identifiable due to unique fixed mutations and insertions and was not amplified from any of the aDNA extracts.

Six samples (R44, R58, R125, R131, R189 and R191) were sent to Egham where the entire procedure, including extraction and amplification, was replicated. Two samples (R125 and R131) were sequenced in their entirety and produced identical sequences to those in Durham. Radiocarbon dating and high collagen levels found in 19 samples from Orkney that had been successful for aDNA extraction give corroborative evidence for good biochemical preservation of the samples.

Microsatellites

A total of 492 modern specimens from eight localities in Orkney and 16 localities in continental Europe from around the most probable area of introduction (northern France and Belgium and inland from there; Fig. S1, Supporting information) were typed at 14 nuclear microsatellite loci (Heckel *et al.* 2005; Jaarola *et al.* 2007; Walser & Heckel 2008): CRB5, INb, MAG6, MAG25, MAR003, MAR016, MAR063, MAR076, MAR080, MAR102, MAR113, MM1, MM8 and Moe02. PCR amplifications were performed with the Qiagen Multiplex kit using Gene Amp PCR System 9700 (Applied Biosystems), PCR fragments were separated on an ABI 3100

sequencer (Applied Biosystems), and their lengths determined using GENEMAPPER 3.7 (Applied Biosystems).

For each locus and each population, tests for departure from Hardy–Weinberg equilibrium were performed with ARLEQUIN 3.12 (Excoffier *et al.* 2005), and significance levels were corrected for multiple testing using the sequential Bonferroni–Holm procedure (Rice 1989). There were no significant departures from equilibrium. Mean allelic richness over loci (A_R) was calculated with FSTAT version 2.9.3.2 (Goudet 1995) and the mean heterozygosity (H) over loci and population differentiation (F_{ST}) with Arlequin.

Phylogenetic analysis

For *cytb*, sequence chromatograms were assembled in Sequencher 4.5 (GeneCodes, Ann Arbor, MI, USA) and aligned manually in BioEdit (Hall 1999). Newly obtained modern DNA sequences were analysed together with previously published modern sequences (Martin *et al.* 2000; Haynes *et al.* 2003; Borkowska & Ratkiewicz 2008; Tougard *et al.* 2008). Using the 1143-bp alignment, this yielded a total of 395 sequences distributed among 173 haplotypes. These were derived from 124 localities (106 from continental Europe, 18 from Orkney). *Microtus levis* was used as an outgroup (Martínková & Moravec 2012).

A Bayesian phylogenetic tree of all modern sequences with the complete (1143 bp) *cytb* was estimated in MrBayes 3.1.2 (Ronquist & Huelsenbeck 2003). To avoid long-tree artefact (Marshall 2010), the Γ-distribution shape parameter α was fixed to 1.0736. The value was obtained from the GTR+Γ substitution model selected by Akaike information criterion in MrModeltest 2.3 (Posada & Crandall 1998; Nylander 2004). The two separate Markov chain Monte Carlo (MCMC) analyses each comprised one cold and eleven heated chains. Samples were drawn from the posterior every 1000 steps over 10 million steps. The chain temperature was 0.06, and two chain swaps were attempted every step to optimize mixing. The first 3000 sampled trees were discarded as burn-in, with the resulting average standard deviation of split frequencies equal to 0.007. The 95% credibility interval of the tree length included a maximum-likelihood estimate of the tree length obtained from RAXML 7.2 (Stamatakis 2006).

Median-joining networks were constructed in Network 4.2 (Bandelt *et al.* 1999) with an equal transition/transversion ratio. The analysis was carried out using the 1130-bp alignment because aDNA sequences were included.

Demographic analysis

For *cytb*, nucleotide and haplotype diversities (Nei 1987) and Ramos-Onsins & Rozas's (2002) R_2 were determined using DNASP 5.1 (Rozas *et al.* 2003). The neutrality test statistics Tajima's (1989) *D* and Fu's (1997) F_S were estimated in Arlequin. This software was also used for mismatch distribution analysis (Rogers & Harpending 1992) to detect and date population expansions.

Radiocarbon-dated and modern *M. arvalis cytb* sequences were analysed using the program BEAST 1.5.1 (Drummond & Rambaut 2007), utilizing the 1130-bp alignment because of inclusion of aDNA sequences. For the Orkney data set, the GTR+Γ model of nucleotide substitution was selected using the Akaike information criterion. All BEAST analyses used this model, along with a strict molecular clock. A comparison using Bayes factors selected the constant-size coalescent prior as the most appropriate demographic model. With ancient sequences in the data set, the potential for undetected sequence errors to influence the analyses was also modelled in BEAST (Rambaut *et al.* 2009). All other parameters were co-estimated with the phylogeny, with samples drawn from the posterior every 1000 steps over a total of 10 million steps. The first 1 000 000 steps were discarded as burn-in. Acceptable mixing and convergence to stationarity were checked using the program TRACER 1.4.1 (Rambaut & Drummond 2007).

For modern DNA sequences from continental European populations, we used a demographic model simulating population expansion and used Bayes factors to compare a strict molecular clock, uncorrelated lognormal relaxed clock and uncorrelated exponential relaxed clock (Drummond *et al.* 2006). A model of exponential growth was identified as the best-fitting demographic model using Bayes factors. Samples were drawn from the Markov chain every 10 000 steps over a total of 100 million steps, with the first 30% of sampled trees discarded as burn-in. Estimated effective sample sizes were >200. Mean mutation rate was fixed to that obtained from the tip-dated sequence analysis described above, and each lineage that was expected to exhibit unique demographic history in the target time frame was analysed separately.

For the specific question of time of colonization of Orkney, we used the program IMa (Hey & Nielsen 2007) to estimate the splitting time between the modern Orkney sample (N = 57 for mtDNA, N = 114 for microsatellite loci) and European mainland samples from the potential area of human introduction: northern France, Belgium and nearby areas of Germany (N = 46 for mtDNA, N = 92 for microsatellite loci). Initial runs were performed to estimate parameters, followed by three replicate runs with different random number seeds to check for consistency of results (results from the final run reported). Input data were 1143-bp *cytb* sequences (assuming the HKY mutation model and the mean

mutation rate estimated in BEAST; priors on the range of mutation rate scalars were set one order of magnitude above and below) and 14 microsatellite loci. The inheritance scalar was set at 0.25. Metropolis coupling was implemented using 20 chains and a geometric heating model (term 1: 0.99; term 2: 0.95). The burn-in period was 7 h. Upper bounds were set for prior distributions for theta, migration rate and the splitting time. The final run saved 499 469 trees per locus and ran for 705 h (50 million steps following burn-in). Convergence was tested by ensuring that effective sample sizes exceeded 50 and parameter trend lines were flat.

We analysed the microsatellite data using approximate Bayesian computation (ABC) to estimate demographic parameters of the colonization history of Orkney by *M. arvalis*. Comparisons were made between the continental European and Mainland Orkney populations listed in Table 1 (and mapped in Fig. S1, Supporting information), except that the four continental populations with small sample sizes (15 individuals or fewer) were excluded. Again, we focused on populations from the region of continental Europe that most likely represented a source for the introduction. We specified an ABC model with 11 demographic parameters in which the Orkney population diverged from a continental population after a bottleneck. Moreover, both the Orkney and the continental populations were allowed to pass through independent bottlenecks. The bottleneck of the continental population could occur either before or after the colonization of Orkney. For simplicity, the model was simulated with instantaneous growth after all three bottleneck events. Further details are given in Data S1, Table S5 and Fig. S3 (Supporting information).

Results

Characteristics of the Orkney voles

Among extensive samples of *Microtus arvalis* collected from across their European range, it is clear that the island populations on Orkney (seven samples) and Guernsey (one sample) are highly distinctive compared with their continental counterparts in first lower molar morphology (Fig. 1). This relates to both size (the island voles generally had larger teeth: Fig. 1e) and shape (the Orkney voles were divergent from continental and insular European *M. arvalis* in the principal components scatter plot, with the M_1 displaying a relatively broader anterior loop: Fig. 1c, d).

At the 14 microsatellite loci screened, the mean number of alleles per locus and heterozygosities were systematically lower in the Orkney populations than in northern France, Belgium and nearby areas of Germany (Table 1), the areas of continental Europe from where introduced voles most likely originated. There was no overlap in heterozygosity values and the small overlap in mean number of alleles per locus related to Mainland Orkney populations, which had distinctly higher diversity values than other Orkney Island populations.

Overall F_{ST} between population samples was very high (0.382, $P < 0.0001$) in agreement with previous studies on the species (e.g. Heckel *et al.* 2005). Pairwise comparisons between Orkney populations ranged from $F_{ST} = 0.094$ to 0.682 (mean: 0.412) and between 0.007 and 0.346 for the continental populations (mean: 0.177; Table S6, Supporting information).

Special features of the mitochondrial DNA variation

Together with previously published results, our new data confirm that *M. arvalis* of the same (Western-North) *cytb* lineage found today in France, Belgium, the Netherlands, Germany and Switzerland also occurs on Orkney (Fig. 2 and Table S2, Supporting information). Fig. 2 shows very clearly that the coastline of France and Belgium provides possible maritime access of the Western-North *cytb* lineage to Orkney, consistent with our contention that this is the most likely source region for an introduction of *M. arvalis* there.

In the Bayesian phylogenetic tree (Fig. 2) and the median-joining network (Fig. 3), the Orkney *cytb* sequences form an unsupported monophyletic clade within the Western-North lineage. However, within that context, the sequences are very distinct from those from coastal Belgium/France and other regions in continental Europe, separated by at least four mutations. Thus, despite detailed sampling of the Western-North lineage in general and along the coast of France and Belgium in particular (Fig. 2), we found no *cytb* sequences there that are clearly ancestral to those of the Orkney voles.

Instead, over this coastal region, an area of more than 10 000 km^2 (Table 3), a single mtDNA haplotype (and sequences separated by one mutation from it) is dominant. None of these haplotypes relate closely to the Orkney *M. arvalis* haplotypes (Fig. 3). This 'starburst' of the coastal French and Belgian sequences suggests their recent expansion and dispersal. A significant deviation from neutral expectation for Tajima's *D* and Fu's *F*s, a small Rozas's R_2 and a significant signal in pairwise DNA sequence comparisons in the mismatch distribution are consistent with this (Table 3).

Considering now the Orkney sequences, the same *cytb* lineage and two of the same haplotypes as those in modern specimens were also found in *M. arvalis* dating back to the Neolithic period (Fig. 3; Table S3, Supporting information). Interestingly, the central Orkney

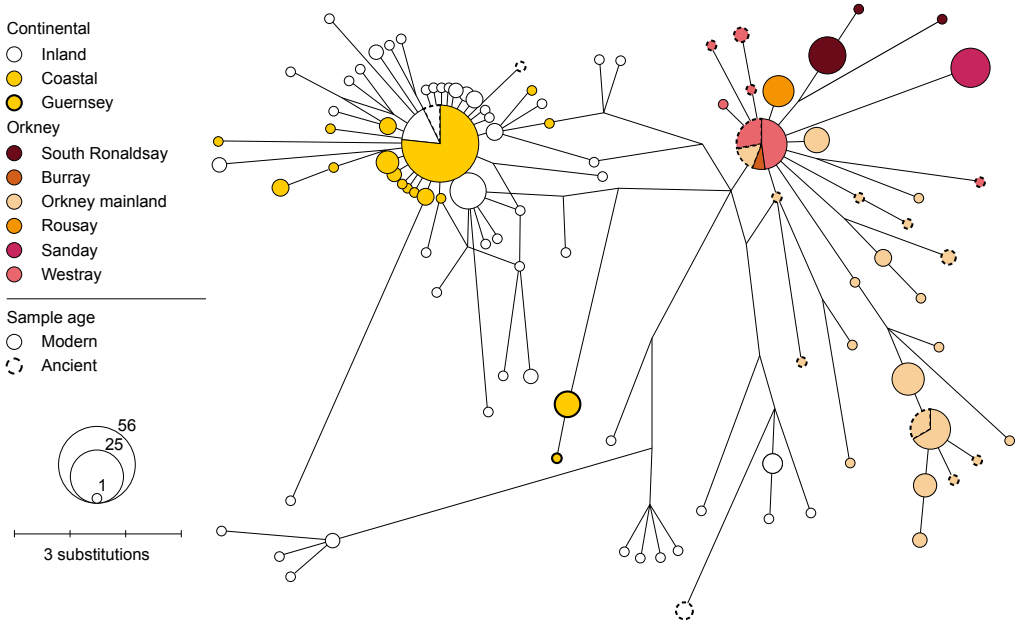


Fig. 3 Median-joining network showing all modern and ancient DNA haplotypes in the Western-North phylogroup of *Microtus arvalis*. The phylogroup is defined in Fig. 2. Node size is proportional to haplotype frequency and edge length to number of substitutions separating the haplotypes. Guernsey is an offshore island also analysed in the geometric morphometric analysis (see Fig. 1).

Table 3 Modern mtDNA comparison of *Microtus arvalis* in Orkney and continental Europe

	Coastal France and Belgium sampled*	Orkney Mainland	All Orkney Islands sampled
Total area (km^2)	10 439	52	73
Number of individuals sequenced	68	42	97
Haplotype diversity (h ± SD)	0.785 ± 0.039	0.864 ± 0.029	0.911 ± 0.108
Nucleotide diversity (π ± SD)	0.00127 ± 0.00017	0.00356 ± 0.00201	0.00416 ± 0.00227
Tajima's D	**−1.986**	−1.102	−1.167
Fu's F_S	**−10.035**	−1.455	−3.014
Rozas's R_2	0.1616	0.0739	0.0587
Mismatch distribution (SSD)	**0.0042**	**0.0112**	**0.0037**

Characteristics of the modern mtDNA sequences of *M. arvalis* from the Orkney archipelago and coastal regions of France and Belgium (where the Orkney *M. arvalis* most likely originated, see text) relating to the population expansion in the two areas.
For neutrality test statistics and the mismatch distribution, significant values ($P < 0.01$) consistent with population expansion are given in bold.
*This included all localities within 100 km of the coast. The area sampled was estimated conservatively as a rectangle with one side given by the distance between the end localities along the coast and the other by the mean distance to the coast of all localities.

haplotype found in archaeological specimens from Mainland Orkney and Westray was detected in modern specimens from Westray and Burray not Mainland Orkney. A second haplotype was found in both archaeological and modern sequences from Mainland Orkney. Both archaeological and modern sequences from Mainland Orkney are represented by many haplotypes in the network (9 and 12, respectively), and there are five archaeological and two modern haplotypes from Westray (Fig. 3).

A possible source area for the introduction of M. arvalis to Orkney

We used ABC on the microsatellite data to compare 12 current populations (both coastal and inland) from northern France, Belgium and nearby areas of Germany as possible source populations for Orkney voles (Table 1 and Fig. S1, Supporting information). Our results confirmed that, if *M. arvalis* were introduced to Orkney from this source region, they most likely originated from coastal Belgium/France (three of the four best-supported populations: Table 1) rather than inland continental populations. The best-supported modern source population is from Stalhille on the Belgian coast (Fig. 2 and Table 1) and geographically the closest point to Orkney within the distribution of the Western-North lineage.

Time of colonization of Orkney

We obtained new ^{14}C dates for 23 specimens from archaeological sites dating from the Late Neolithic period to the Viking Age (Table 2). The estimated dates are consistent with those of the archaeological contexts from which their bones were recovered. They show definitively that *M. arvalis* was present on Orkney at least 5100 BP. Two molecular analyses suggest that they did not arrive much earlier than this. An ABC analysis of microsatellites shows modal point estimates of time of colonization of Orkney that vary between 3000 and 5700 BP (between Stalhille, Belgian coast and three localities on Mainland Orkney). An IMa analysis (Hey & Nielsen 2007) of ancient and modern mtDNA sequences and microsatellites yields an estimated colonization time of 4515 years (90% CI: 2568–9019). Both analyses assume a single generation per year. Multiple generations (for which there is no evidence: Gorman & Reynolds 2008) would imply a more recent post-Neolithic colonization, which does not fit well with our new ^{14}C dates (voles on Orkney at latest 5100 BP) and aDNA sequencing of archaeological specimens (same genetic lineage as the modern specimens).

While the radiocarbon and molecular dating indicate that voles arrived on Orkney about 5000 BP, the time to the most recent common ancestor (tMRCA) for the Orkney *cytb* sequences is much earlier, *c.* 15 400 years (95% CI of 8100–25 400 years), based on Bayesian coalescent analysis incorporating modern and ancient DNA data. Using the same mutation rate ($\mu = 3.27 \times 10^{-7}$ mutations/site/year) as for Orkney sequences, the tMRCA for the sequences from coastal France/Belgium (the putative source area for the Orkney voles) is only 2356 years (95% CI of 723–4941 years). Mismatch analysis (Rogers & Harpending 1992) provides further evidence of the recentness of the derivation of the *cytb* sequences in this potential source area for the Orkney voles, showing an expansion beginning 1860 BP (95% CI: 1201–2665).

Discussion

Cytb mutation rate

Estimates of mutation rates for phylogeographical studies have traditionally been based on fossil calibrations, often founded on very incomplete information. For *Microtus*, Triant and DeWoody (2006) used a fossil calibration point of 1.5 Ma to provide a mutation rate for *cytb* of approximately 8×10^{-8} mutations/site/year. Although this is probably the best-substantiated fossil-based estimate of the *cytb* mutation rate for *Microtus* in the literature, we considered it inadequate for our needs. This is because the fossil calibration point likely has errors (it is not based on precisely dated geological events or fossil transitions) and is geologically very much earlier than the time frame for our study. Instead, we were able to calibrate our rate estimate using the aDNA sequences from radiocarbon-dated specimens as part of a Bayesian coalescent analysis. This method has the advantage of producing an estimate of the *cytb* mutation rate relevant to precisely the population that we were studying and the temporal scale of interest.

The mutation rate we obtained and used here (3.27×10^{-7} mutations/site/year) was about four times higher than that of Triant and DeWoody (2006), but comparable to other aDNA-based estimates of mutation rate for a variety of species (Ho *et al.* 2011). Navascués & Emerson (2009) caution against possible upward bias of such aDNA-based estimates under extreme demographic scenarios, but there is a growing consensus in the literature that aDNA-based estimates are better than traditional fossil-based mutation rates, which are too low for dating divergences and demographic events that have occurred over the last few thousand years (Ho & Larson 2006; de Bruyn *et al.* 2011). With regard to our aDNA-based *cytb* mutation rate for *M. arvalis*, it is extremely close to that obtained for another *Microtus*, which used a very recent geological event as a calibration point (*M. agrestis*; 3.89×10^{-7} mutations/site/year; Herman & Searle 2011). This provides independent support for our aDNA-based estimate, although, as with all molecular dating, interpretations have to be viewed with caution.

Colonization history of Orkney voles

It is striking that there is substantial *cytb* variation in *M. arvalis* in Mainland Orkney and over the whole archipelago (Fig. 3, Table 3). Island populations often

show low genetic diversity (Frankham 1997). This can relate to small population sizes and/or population bottlenecks associated with colonization of islands, particularly by sweepstake dispersal or human introduction. The high *cytb* diversity could indicate that the Orkney population of *M. arvalis* represents an island relict of a previously continuous mainland population, perhaps dating back to before the LGM (Beirne 1952). This would fit with the long tMRCA for the molecular variation on Orkney, potentially dating back to 25 400 BP (within the 95% CI).

However, there are strong arguments against glacial survival of the Orkney vole population. First, the IMa analysis based on microsatellites and *cytb* and the ABC analysis based on microsatellites provide a date of arrival around 5000 BP, considerably more recent than the LGM. Second, all the other species of small mammals on Orkney are most reasonably viewed as human introductions (Yalden 1982), necessitating a special case for *M. arvalis* as a glacial survivor. Third, it is very difficult to make this special case given that *M. arvalis* is not a species currently associated with arctic or even moderately high latitude conditions. Its range extends eastward beyond Lake Baikal and yet barely traverses north of the 60th parallel (Fig. 2; Shenbrot & Krasnov 2005). Orkney was under or near a glacial ice sheet at the LGM (Bowen *et al*. 2002) and *M. arvalis* is not part of the fossil fauna known from Britain from the last glacial period (Yalden 1999; Currant & Jacobi 2001). Fourth, *M. arvalis* is not currently found in Britain (Fig. 2). It is therefore contrary to think that *M. arvalis* should be a glacial relict on Orkney rather than *M. agrestis*, when the latter occurs further north in Eurasia (beyond the 70th parallel) and is distributed throughout Britain, including on many offshore islands, while *M. arvalis* only occurs on Orkney. There have been no land connections between mainland Britain and Orkney after conditions ameliorated following the LGM (Yalden 1982), and hence why *M. agrestis* (and other wide-ranging small mammals in Britain, such as common shrews *Sorex araneus* and bank voles *Myodes glareolus*) failed to colonize Orkney. If *M. arvalis* did not survive on Orkney itself during the last glacial period, the absence of the species in Britain means there are no grounds to suggest sweepstake colonization from there. It is conceivable that there could have been sweepstake colonization of Orkney from Doggerland, the landmass connecting Britain, the Low Countries and Denmark until about 8000 BP (Weninger *et al*. 2008) – but this would require the survival of small mammals on floating mats of vegetation over a substantial marine gap between Doggerland and Orkney.

Thus, human introduction is by far the most likely explanation for the occurrence of *M. arvalis* on Orkney. From the IMa and ABC dates, this introduction at about 5000 BP fits well with the earliest radiocarbon dates for archaeological *M. arvalis* from Neolithic contexts (5100 years old: Table 2) and the beginnings of the Neolithic culture on Orkney (5600 BP: Ritchie 2001; Schulting *et al*. 2010). Voles could have been brought to Orkney by Mesolithic hunter-gatherers, as early as *c*. 9000 BP, but no vole remains have been found in the one excavated Mesolithic site on Orkney (Lee & Woodward 2009), in contrast to their abundance at Neolithic and later sites (Yalden 1999; Thaw *et al*. 2004).

If, as appears most likely, the voles were introduced by Neolithic settlers about 5000 BP, various other implications flow from our molecular data, which are of considerable archaeological interest.

First, the introduction implies long-distance maritime travel by Neolithic people between continental Europe and Orkney, extending on findings from elsewhere (e.g. Broodbank 2006). Our study highlights the Belgian coastline as the most reasonable source of the Orkney voles on the basis of available genetic data. This suggests Neolithic cultural linkages between Belgium and Orkney, of worthwhile focus for future archaeological investigation. *Microtus arvalis* were not introduced successfully into mainland Britain, which is consistent with relatively direct transport to Orkney from the continental source area.

Second, if the introduction occurred about 5000 BP, then, because the tMRCA for the Orkney voles is so long (15 400 years), substantial numbers of female voles must have been introduced to explain the *cytb* variation observed in modern and archaeological Orkney voles. High genetic diversity is already evident in the 16 aDNA sequences dating to 4200 BP or earlier, separated by up to 10 mutations (Fig. 3 and Table S3, Supporting information) and which produce an estimate for the tMRCA (14 780 years; 95% CI of 4681–36 379 years) similar to that of the full ancient and modern data set.

To explain a substantial number of voles arriving accidentally on Orkney implies transport of plentiful grass livestock bedding/fodder in which the vole stowaways could have survived. This in turn may suggest the direct movement of livestock as part of the proposed Neolithic linkage between Belgium and Orkney.

Alternatively, deliberate transport of voles onto Orkney could explain the large numbers introduced (Thaw *et al*. 2004). It is conceivable that voles were taken as food items, pets or for cultural/religious purposes – *M. arvalis* is docile in captivity (Berry 2000), so could theoretically have been 'tamed'. This suggestion of deliberate transportation of small rodents has a precedent: it has been argued that Pacific rats (*Rattus exulans*), now present on islands throughout Oceania,

were intentionally conveyed by Polynesians as a food source (Matisoo-Smith & Robins 2004).

Evolutionary processes affecting voles on Orkney and the continental source area

Compared with continental European *M. arvalis*, those on Orkney and another offshore island (Guernsey) are divergent in terms of tooth morphology, including increased tooth size. For Orkney, this divergence may have occurred over *c.* 5000 years, if introduced during the Neolithic. In addition to having larger teeth, the Orkney and Guernsey *M. arvalis* have a larger body size than continental voles (Gorman & Reynolds 2008). Quick-evolving rodent gigantism has been described previously on islands of the northeast Atlantic (Corbet 1961; Angerbjorn 1986), but not within such a precisely defined time frame. A range of selective factors have been proposed to explain this gigantism, including an absence of small mammalian predators (Lomolino 1985), and a genetic basis for gigantism has been identified in island house mice (Chan *et al.* 2012). Elsewhere, we further explore the dynamics of morphological evolution for the Orkney *M. arvalis* using archaeological specimens (T. Cucchi, R. Barnett, N. Martínková, *et al.* submitted) extending substantially on previous studies (Berry & Rose 1975; Corbet 1986).

Despite their similarity in large tooth and body size, Guernsey and Orkney voles exhibit distinctive mtDNA haplotypes (Fig. 3). It is therefore most reasonable to consider that the Guernsey and Orkney voles attained their large body size independently. It is not clear whether Guernsey was colonized naturally before it became an island, as part of the continental European late glacial/postglacial species expansion (Haynes *et al.* 2003; Heckel *et al.* 2005; Tougard *et al.* 2008), or whether the voles were introduced by people after it became an island (Gorman & Reynolds 2008). However, given that Guernsey voles are likely to come from the same general (northern France/Belgium) source area as the Orkney voles and that they are also different in mtDNA from current northern France/Belgium populations, there would be much interest in further detailed comparison of Orkney, Guernsey and northern France/Belgium voles.

In addition to the operation of selection in the evolution of Orkney voles suggested by morphology, stochastic processes appear to have been important based on microsatellites. The population on the largest island, Mainland Orkney, has retained much of the microsatellite variation found in continental Europe, while all the other Orkney Islands (which are considerably smaller: Fig. 1) show very low levels of microsatellite variation, consistent with founder events and genetic drift. Similar stochastic processes can also explain microsatellite variation among Scottish Island populations of common shrew (White & Searle 2007a).

Our findings with regard to morphology and microsatellites in *M. arvalis* are unsurprising in comparison with previous studies on island small mammals, but the results from our mtDNA analyses are more unexpected. Although the *cytb* sequences from Orkney and the proposed source area for the Orkney colonization both belong to the Western-North lineage of *M. arvalis*, the sequences are remarkably divergent given the time frame for colonization. Also, it might have been expected that (as for the microsatellites) variability would have been lower on Orkney than in continental Europe. In fact, the opposite is the case. Taking either the principal island (Mainland Orkney) or the whole archipelago, mtDNA diversity is higher in Orkney than in coastal France/Belgium (Table 3). Our dating analysis also shows that the mtDNA sequences in coastal France/Belgium have a much more recent derivation than the Orkney sequences.

So, here we are seeing another facet of evolution in association with the colonization of offshore islands, in this case occurring in the mainland population. The presence of derived sequences in coastal France/Belgium suggests a replacement event in *M. arvalis*, with one mtDNA type (the current type) replacing another (the Orkney type), similar to aDNA findings in other species (Barnes *et al.* 2002; Pergams *et al.* 2003; Hofreiter *et al.* 2007). The fact that there is an affiliation between coastal Belgium and Orkney on the basis of microsatellite genotypes argues against a complete population replacement (e.g. by extinction–recolonization) as an explanation for the mtDNA result. Instead, within-population processes of selective sweeps or genetic drift are implicated, and more likely expressed in the mtDNA data, as a single locus with small effective population size than in the microsatellite data. We cannot be sure what environmental factors promoted the replacement. There could, for instance, have been a local, unrecorded disease outbreak. However, it is notable that the replacement occurred over a period when *M. arvalis* populations would have changed dramatically due to human land-use change, and this appears the most likely driver of the replacement. Over several thousand years, sustained forest clearance in continental Europe (Rackham 1998; Cyprien *et al.* 2004) created new agricultural habitats and associated selection pressures that essentially expanded the opportunities for *M. arvalis* as a species that particularly exploits managed grassland (Niethammer & Krapp 1982). In such a habitat, *M. arvalis* populations can undergo massive population expansions and crashes (Delattre *et al.* 1992) that reduce long-term effective population size, promoting genetic change through drift. On Orkney (which saw the rapid

decline of low shrubs and tree species with the arrival of Neolithic farmers: Bunting 1996), *M. arvalis* utilizes a range of open habitats and does not show the same dramatic population fluctuations as seen in parts of continental Europe (Gorman & Reynolds 2008).

Offshore islands as field laboratories

There has been a tendency to view offshore island populations of small mammals (and other organisms with low density and low dispersal) as genetic deviants from the 'norm'. This is because studies of various species have shown results similar to ours for morphology (substantial change) and microsatellites (loss of variation) (Lomolino 1985; Frankham 1997; Boessenkool *et al*. 2007; Millien 2011). These have included detailed studies on small mammals such as wood mouse *Apodemus sylvaticus* (Angerbjorn 1986; Michaux *et al*. 1996), masked shrew *Sorex cinereus* (Stewart & Baker 1992) and common shrew (White & Searle 2007a,b, 2008).

However, as we have demonstrated with our *M. arvalis* mtDNA studies, island populations can also represent genetic 'arks', retaining the ancestral genetic variation, while evolutionary and other processes on the mainland may lead to a loss of that ancestral variation. Islands may have importance therefore in conservation of genetic variation. A further example involving human introduction of a small mammal onto an offshore island is provided by the Eurasian red squirrel *Sciurus vulgaris*. Thus, Irish red squirrels have genetic variants that apparently derive by introduction from Britain, but these are now absent in that source population (Finnegan *et al*. 2008; Searle 2008). For low density and low dispersal organisms such as small mammals, we suggest that genetic surveys of mainland areas should, where available, include populations from neighbouring offshore islands. It is very likely that those island populations will provide a new perspective on the temporal and spatial dynamics of the genetic variation in that region.

The 'ark' concept that we discuss here is of course more general. Populations colonizing new areas will take the genetic and nongenetic characteristics of the source population, and some of those characteristics may subsequently be lost in the source population but retained in the population in the new area. In this way, for instance, the United States is a 'linguistic ark' for various English words that would have been common in the British Isles at the time of settlement of North America by the British, but which have subsequently fallen into disuse in the homeland (e.g. 'fall' meaning 'autumn').

Returning to genetic characteristics of offshore islands, in addition to their potential as genetic 'arks', they also hold potential as field laboratories to study genetic change in the islands themselves. Compared with the classic evolutionary studies on oceanic islands, those based on offshore islands will tend to view events over shorter timescales and thus provide a different perspective on evolutionary processes. Offshore islands are particularly valuable for studying initial stages of diversification, with the opportunity (as in the current study together with T. Cucchi, R. Barnett, N. Martínková, *et al*. submitted) to follow island populations from their foundation to the present day using advanced genetic and morphometric tools as applied to modern and ancient populations of different ages. Extremely accurate dating of ancient populations may be possible (e.g. in archaeological settings). With this short time duration and close proximity to the mainland, there is also a greater chance to find the precise source area for the island colonization, which allows interesting comparison of evolutionary processes on the mainland and island. This brings us back to the value of offshore island populations in interpreting mainland processes. Offshore islands are an underutilized resource for evolutionary analysis, with great potential. In some ways, they represent study systems intermediate between those in a continental setting and those on oceanic islands; they have the simplicity of the oceanic island system yet are clearly relevant to continental situations.

Acknowledgements

We dedicate this article to Michel Pascal, our outstanding co-author, who died on 5 January 2013, and to Anne Brundle, who gave us access to much archaeological material and who died in 2012. We acknowledge receipt of a Marie Curie Intra European Fellowship (to N.M.), support from the Swiss National Science Foundation (projects 31003A-127377, 3100A0-112072 and 3100-126074) to L.E. and G.H., funding from SYNTHESYS2 made available by the European Community Research Infrastructure under FP7 ('Synthesis of Systematic Resources', 226506-CP-CSA-Infra) to S.B., a Wellcome Trust University award to K.M.D. (GR071037) and overarching funding from the Arts and Humanities Research Council (project grant 119396). We thank V. Bretille, A. Frantz, M. Fuster, N. Gould, J. Herman, E. Jones, S. Martínek, R. Marwick, J. Michaux, S. Montuire, J. Pauperio, C. Scott, C. Tougard, B. Walther and N. Wheale for field specimens, T. White for assistance with IMa runs, and A. Ritchie, L. Shepherd and A. Sheridan for archaeological advice. We are grateful to the following for museum and archaeological samples: J. Barrett (MacDonald Institute, University of Cambridge), A. Brundle (Orkney Museum), C. David (Guernsey Museum), A. Ervynck (Flemish Heritage Institute), L. Gordon (Smithsonian Institute), J. Herman (National Museums of Scotland), D. Lee (Orkney College), R. Sabin (British Museum - Natural History, London) and G. Veron (Muséum national d'histoire naturelle, Paris).

References

Angerbjorn A (1986) Gigantism in island populations of wood mice (*Apodemus*) in Europe. *Oikos*, **47**, 47–56.

Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**, 37–48.

Barnes I, Matheus P, Shapiro B, Jensen D, Cooper A (2002) Dynamics of Pleistocene population extinctions in Beringian brown bears. *Science*, **295**, 2267–2270.

Beirne BP (1952) *The Origin and History of the British Fauna*. Methuen, London.

Berry RJ (1996) Small mammal differentiation on islands. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, **351**, 753–764.

Berry RJ (2000) *Orkney Nature*. Poyser, London.

Berry RJ, Rose FEN (1975) Islands and the evolution of *Microtus arvalis* (Microtinae). *Journal of Zoology*, **177**, 395–409.

Boessenkool S, Taylor SS, Tepolt CK, Komdeur J, Jamieson IG (2007) Large mainland populations of South Island robins retain greater genetic diversity than offshore island refuges. *Conservation Genetics*, **8**, 705–714.

Bookstein FL (1997) Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical Image Analysis*, **1**, 225–243.

Borkowska A, Ratkiewicz M (2008) Sex-related spatial structure and effective population size in the common vole, *Microtus arvalis*, as revealed by mtDNA analysis. *Annales Zoologici Fennici*, **45**, 255–262.

Bowen DQ, Phillips FM, McCabe AM, Knutz PC, Sykes GA (2002) New data for the last glacial maximum in Great Britain and Ireland. *Quaternary Science Reviews*, **21**, 89–101.

Broodbank C (2006) The origins and early development of Mediterranean maritime activity. *Journal of Mediterranean Archaeology*, **19**, 199–230.

de Bruyn M, Hoelzel AR, Carvalho GR, Hofreiter M (2011) Faunal histories from Holocene ancient DNA. *Trends in Ecology and Evolution*, **26**, 405–413.

Bunting MJ (1996) The development of heathland in Orkney, Scotland: pollen records from Loch of Knitchen (Rousay) and Loch of Torness (Hoy). *The Holocene*, **6**, 193–212.

Chan YF, Jones FC, McConnell E *et al*. (2012) Parallel selection mapping using artificially selected mice reveals body weight control loci. *Current Biology*, **22**, 794–800.

Corbet GB (1961) Origin of British insular races of small mammals and of Lusitanian fauna. *Nature*, **191**, 1037–1040.

Corbet GB (1986) Temporal and spatial variation of dental pattern in the voles, *Microtus arvalis*, of the Orkney Islands. *Journal of Zoology*, **208**, 395–402.

Currant A, Jacobi R (2001) A formal mammalian biostratigraphy for the Late Pleistocene of Britain. *Quaternary Science Reviews*, **20**, 1707–1716.

Cyprien AL, Visset L, Carcaud N (2004) Evolution of vegetation landscapes during the Holocene in the central and downstream Loire basin (Western France). *Vegetation History and Archaeobotany*, **13**, 181–196.

Delattre P, Giraudoux P, Baudry J *et al*. (1992) Land use patterns and types of common vole (*Microtus arvalis*) population kinetics. *Agriculture, Ecosystems & Environment*, **39**, 153–168.

DeWoody JA, Chesser RK, Baker RJ (1999) A translocated mitochondrial cytochrome *b* pseudogene in voles (Rodentia: *Microtus*). *Journal of Molecular Evolution*, **48**, 380–382.

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology*, **4**, e88.

Ehlers J, Gibbard PL (2004) *Quaternary Glaciations – Extent and Chronology. Part I: Europe*. Elsevier, Amsterdam.

Emerson BC (2002) Evolution on oceanic islands: molecular phylogenetic approaches to understanding pattern and process. *Molecular Ecology*, **11**, 951–966.

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.

Fink S, Fischer MC, Excoffier L, Heckel G (2010) Genomic scans support repetitive continental colonization events during the rapid radiation of voles (Rodentia: *Microtus*): the utility of AFLPs versus mitochondrial and nuclear sequence markers. *Systematic Biology*, **59**, 548–572.

Finnegan LA, Edwards CJ, Rochford JM (2008) Origin of, and conservation units in, the Irish red squirrel (*Sciurus vulgaris*) population. *Conservation Genetics*, **9**, 1099–1109.

Frankham R (1997) Do island populations have less genetic variation than mainland populations? *Heredity*, **78**, 311–327.

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147**, 915–925.

Gorman ML, Reynolds P (2008) Orkney and Guernsey vole. In: *Mammals of the British Isles: Handbook*, 4th edn (eds Harris S, Yalden DW), pp. 107–110. The Mammal Society, Southampton.

Goudet J (1995) FSTAT (Version 1.2): a computer program to calculate F-statistics. *Journal of Heredity*, **86**, 485–486.

Grant PR (ed.) (2008) *Evolution on Islands*. Oxford University Press, Oxford.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acid Symposium Series*, **41**, 95–98.

Haynes S, Jaarola M, Searle JB (2003) Phylogeography of the common vole (*Microtus arvalis*) with particular emphasis on the colonization of the Orkney archipelago. *Molecular Ecology*, **12**, 951–956.

Heckel G, Burri R, Fink S, Desmet JF, Excoffier L (2005) Genetic structure and colonization processes in European populations of the common vole, *Microtus arvalis*. *Evolution*, **59**, 2231–2242.

Hedges REM, Housley RA, Law IA, Perry C, Gowlett JAJ (1987) Radiocarbon dates from the Oxford AMS system: archaeometry datelist 6. *Archaeometry*, **29**, 289–306.

Herman JS, Searle JB (2011) Post-glacial partitioning of mitochondrial genetic variation in the field vole. *Proceedings of the Royal Society. B, Biological Sciences*, **278**, 3601–3607.

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences USA*, **104**, 2785–2790.

Ho SYW, Larson G (2006) Molecular clocks: when times are a-changin'. *Trends in Genetics*, **22**, 79–83.

Ho SYW, Lanfear R, Phillips MJ *et al*. (2011) Bayesian estimation of substitution rates from ancient DNA sequences with low information content. *Systematic Biology*, **60**, 366–375.

Hofreiter M, Munzel S, Conard NJ *et al*. (2007) Sudden replacement of cave bear mitochondrial DNA in the late Pleistocene. *Current Biology*, **17**, R122–R123.

Jaarola M, Martínková N, Gündüz İ *et al*. (2004) Molecular phylogeny of the speciose vole genus *Microtus* (Arvicolinae, Rodentia) inferred from mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, **33**, 647–663.

Jaarola M, Ashford RT, Ratkiewicz M, Brunhoff C, Borkowska A (2007) Isolation and characterization of polymorphic microsatellite loci in the field vole, *Microtus agrestis*, and their cross-utility in the common vole, *Microtus arvalis*. *Molecular Ecology Notes*, **7**, 1029–1031.

Klingenberg CP (2011) MorphoJ: an integrated software package for geometric morphometrics. *Molecular Ecology Resources*, **11**, 353–357.

Lachaise D, Harry M, Solignac M *et al*. (2000) Evolutionary novelties in islands: *Drosophila santomea*, a new melanogaster sister species from Sao Tome. *Proceedings of the Royal Society of London. B, Biological Sciences*, **267**, 1487–1495.

Lee DHJ, Woodward NL (2009) *Links House, Stronsay, Orkney (Phase III)*. Data Structure Report, ORCA.

Lomolino MV (1985) Body size of mammals on islands - the island rule reexamined. *American Naturalist*, **125**, 310–316.

Marshall DC (2010) Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. *Systematic Biology*, **59**, 108–117.

Martin Y, Gerlach G, Schlötterer C, Meyer A (2000) Molecular phylogeny of European muroid rodents based on complete cytochrome *b* sequences. *Molecular Phylogenetics and Evolution*, **16**, 37–47.

Martínková N, Moravec J (2012) Multilocus phylogeny of arvicoline voles (Arvicolini, Rodentia) shows small tree terrace size. *Folia Zoologica*, **61**, 254–267.

Martínková N, McDonald RA, Searle JB (2007) Stoats (*Mustela erminea*) provide evidence of natural overland colonization of Ireland. *Proceedings of the Royal Society. B, Biological Sciences*, **274**, 1387–1393.

Matisoo-Smith E, Robins JH (2004) Origins and dispersals of Pacific peoples: evidence from mtDNA phylogenies of the Pacific rat. *Proceedings of the National Academy of Sciences USA*, **101**, 9167–9172.

Michaux JR, Libois R, Fons R (1996) Différenciation génétique et morphologique du mulot, *Apodemus sylvaticus*, dans le bassin méditerranéen occidental. *Vie Milieu*, **46**, 193–203.

Millien V (2011) Mammals evolve faster on smaller islands. *Evolution*, **65**, 1935–1944.

Morrison ML, Block WM, Jehl JR Jr, Hall LS (1992) Terrestrial vertebrates of the Mono Lake Islands, California. *Great Basin Naturalist*, **52**, 328–334.

Navascués M, Emerson BC (2009) Elevated substitution rate estimates from ancient DNA: model violation and bias of Bayesian methods. *Molecular Ecology*, **18**, 4390–4397.

Nei M (1987) *Molecular Evolutionary Genetics*. Colombia University Press, New York.

Nichols C, Herman J, Gaggiotti OE *et al*. (2007) Genetic isolation of a now extinct population of bottlenose dolphins (*Tursiops truncatus*). *Proceedings of the Royal Society. B, Biological Sciences*, **274**, 1611–1616.

Niethammer J, Krapp F (1982) *Microtus arvalis* (Pallas, 1779) - Feldmaus. In: *Handbuch der Säugetiere Europas 2 (1). Rodentia: II* (eds Niethammer J, Krapp F), pp. 284–318. Akademische Verlagsgesellschaft, Wiesbaden.

Nylander JAA (2004) *MrModeltest v2*. Evolutionary Biology Centre, Uppsala University, Uppsala.

Pergams ORW, Barnes WM, Nyberg D (2003) Rapid change in mouse mitochondrial DNA. *Nature*, **423**, 397.

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. Retrieved from http://www.r-project.org

Rackham O (1998) Implications of historical ecology for conservation. In: *Conservation Science and Action* (ed. Sutherland W), pp. 152–175. Blackwell, Oxford.

Rambaut A, Drummond AJ (2007) *Tracer v1.4*. University of Oxford, Oxford.

Rambaut A, Ho SYW, Drummond AJ, Shapiro B (2009) Accommodating the effect of ancient DNA damage on inferences of demographic histories. *Molecular Biology and Evolution*, **26**, 245–248.

Ramos-Onsins SE, Rozas J (2002) Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution*, **19**, 2092–2100.

Reimer PJ, Baillie MGL, Bard E *et al*. (2004) IntCal04 terrestrial radiocarbon age calibration, 0-26 cal kyr BP. *Radiocarbon*, **46**, 1029–1058.

Rice WR (1989) Analyzing tables of statistical tests. *Evolution*, **43**, 223–225.

Ritchie A (2001) Knap of Howar, Papa Westray. *Discovery and Excavation in Scotland*, **2000**, 124–125.

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, **9**, 552–569.

Rohlf FJ (2010a) *TpsDig 2-Thin Plate Spline Digitizer*. Ecology & Evolution, State University at Stony Brook, New York.

Rohlf FJ (2010b) *TpsRelw 1.49-Thin Plate Spline Relative Warp*. Ecology & Evolution, State University at Stony Brook, New York.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.

Schulting RA, Sheridan A, Crozier R, Murphy E (2010) Revisiting Quanterness: new AMS dates and stable isotope data from an Orcadian chamber tomb. *Proceedings of the Society of Antiquaries of Scotland*, **140**, 1–50.

Searle JB (2008) The genetics of mammalian invasions: a review. *Wildlife Research*, **35**, 185–192.

Shenbrot GI, Krasnov BR (2005) *An Atlas of the Geographic Distribution of the Arvicoline Rodents of the World (Rodentia, Muridae: Arvicolinae)*. Pensoft Publishers, Sofia, Moscow.

Simpson GG (1940) Mammals and land bridges. *Journal of the Washington Academy of Sciences*, **30**, 137–163.

Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Stewart DT, Baker AJ (1992) Genetic differentiation and biogeography of the masked shrew in Atlantic Canada. *Canadian Journal of Zoology*, **70**, 106–114.

Stuiver M, Reimer PJ (1993) Extended ^{14}C database and revised CALIB radiocarbon calibration program. *Radiocarbon*, **35**, 215–230.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Thaw S, Jaarola M, Searle JB, Dobney KM (2004) The origin of the Orkney vole *Microtus arvalis orcadensis*: a proxy for reconstructing human movements. In: *Atlantic Connections and Adaptations* (eds Housley RA, Coles G), pp. 114–119. Oxbow, Oxford.

Tougard C, Renvoisé E, Petitjean A, Quéré J-P (2008) New insight into the colonization processes of common voles: inferences from molecular and fossil evidence. *PLoS ONE*, **3**, e3532.

Triant DA, DeWoody JA (2006) Accelerated molecular evolution in *Microtus* (Rodentia) as assessed via complete mitochondrial genome sequences. *Genetica*, **128**, 95–108.

Wallace AR (1880) *Island Life*. Macmillan, London.

Walser B, Heckel G (2008) Microsatellite markers for the common vole (*Microtus arvalis*) and their cross-species utility. *Conservation Genetics*, **9**, 479–481.

Weninger B, Schulting R, Bradtmöller M *et al.* (2008) The catastrophic final flooding of Doggerland by the Storegga Slide tsunami. *Documenta Praehistorica*, **35**, 1–24.

White TA, Searle JB (2007a) Genetic diversity and population size: island populations of the common shrew, *Sorex araneus*. *Molecular Ecology*, **16**, 2005–2016.

White TA, Searle JB (2007b) Factors explaining increased body size in common shrews (*Sorex araneus*) on Scottish islands. *Journal of Biogeography*, **34**, 356–363.

White TA, Searle JB (2008) The colonization of Scottish islands by the common shrew, *Sorex araneus* (Eulipotyphla: Soricidae). *Biological Journal of the Linnean Society*, **94**, 797–808.

Yalden DW (1982) When did the mammal fauna of the British Isles arrive? *Mammal Review*, **12**, 1–57.

Yalden DW (1999) *The History of British Mammals*. Poyser, London.

K.M.D. and J.B.S. conceived, initiated and coordinated the project. T.C., M.F., G.H., N.M., M.P., Ma.P., J.-P.Q. and J.B.S. organized and collected field specimens, R.B., T.C. and K.M.D. organized and obtained museum and archaeological samples. R.B., T.C. and N.M. conducted laboratory work supported by K.M.D., A.R.H., G.H. and J.B.S. and with major contributions from S.B. and T.H. R.B., T.C., N.M. and R.S. analysed the data supported by K.M.D., L.E., P.O'H., A.R.H., G.H., S.Y.W.H. and J.B.S. J.B.S. led the writing of the text to which all authors contributed.

Data accessibility

The DNA sequences have been deposited in GenBank (GU190383-GU190665). The TreeBASE entry for the phylogenetic tree of *cytb* sequences shown in Fig. 2, the microsatellite genotypes, the input file for the IMa analysis, the modern and ancient *cytb* sequences, and the morphological coordinates collected from 2D images of skulls are all available through DRYAD:

doi:10.5061/dryad.9rf5m.

Supporting information

Additional supporting information may be found in the online version of this article.

Data S1 Material and Methods.

Table S1 List of modern *M. arvalis* samples used for morphometrics, including country and site of origin.

Table S2 List of all modern *M. arvalis* used for *cytb* analysis and collected for this study, including country and site of origin; arranged according to *cytb* haplotype.

Table S3 List of all ancient specimens of *M. arvalis* that successfully provided a *cytb* sequence with details of location collected, calibrated age range (where obtained) and GenBank Accession Number for the sequence.

Table S4 List of primers used for the amplification of cytochrome *b* from *M. arvalis*.

Table S5 Prior distributions of the ABC model parameters (as illustrated in Fig. S3).

Table S6 Pairwise F_{ST} values between population samples analysed with microsatellites (see Table 1 and Fig. S1).

Fig. S1 Map showing distribution of population samples of modern *M. arvalis* used for microsatellite typing, labelled for mtDNA lineage. Population names as listed in Table 1: 1 – Heerenveen, 2 – Dinteloord, 3 – Stalhille, 4 – Veurne, 5 – Pihen lès Guines, 6 – Fressenneville, 7 – Daubeuf, 8 – Thaon, 9 – Ste Marie du Mont, 10 – St Jean du Thomas, 11 – Baie d'Aiguillon, 12 – Aiffres, 13 – Avallon, 14 – Clérmont-Ferrand, 15 – Alflen, 16 – Schiltach, 17 – Loch of Swartmill, 18 – Ness, 19 – Whitemill Bay, 20 – Settiscarth, 21 – Harray Stenness, 22 – St Ola, 23 – Grimness, 24 – Wind Wick.

Fig. S2 Map showing Orkney localities where ancient specimens of *M. arvalis* successfully provided either radiocarbon dates and/or *cytb* sequences. Localities as listed in Tables 2 and S3: 1 - Holm of Papa Westray, Westray, 2 - Point of Cott, Westray, 3 - Pierowall Quarry, Westray, 4 - Quanterness, Mainland, 5 - Earl's Bu, Mainland, 6 - Howe, Mainland, 7 - Skara Brae, Mainland, 8 - Green Hill, South Walls, Hoy.

Fig. S3 Diagram illustrating the ABC model parameters.

Paper 2.3.1

Irwin N. R., Bayerlová M., Missa O., **Martínková N.** 2012. Complex patterns of host switching in New World arenaviruses. *Molecular Ecology* 21: 4137-4150.

MOLECULAR ECOLOGY

Molecular Ecology (2012) **21**, 4137–4150 doi: 10.1111/j.1365-294X.2012.05663.x

Complex patterns of host switching in New World arenaviruses

NANCY R. IRWIN,* MICHAELA BAYERLOVÁ,† OLIVIER MISSA,* and NATÁLIA MARTÍNKOVÁ,†‡

*Department of Biology, University of York, POBOX 373, York, YO10 5DD, UK, †Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic, ‡Institute of Vertebrate Biology, v.v.i., Academy of Sciences of the Czech Republic, Brno, Czech Republic

Abstract

We empirically tested the long-standing hypothesis of codivergence of New World arenaviruses (NWA) with their hosts. We constructed phylogenies for NWA and all known hosts and used them in reconciliation analyses. We also constructed a phylogenetic tree of all Sigmodontinae and Neotominae rodents and tested whether viral–host associations were phylogenetically clustered. We determined host geographical overlap to determine to what extent opportunity to switch hosts was limited by host relatedness or physical proximity. With the exception of viruses from North America, no phylogenetically codivergent pattern between NWA and their hosts was found. We found that different virus clades were clustered differently and that Clade B with members pathogenic to humans was randomly distributed across the rodent phylogeny. Furthermore, viral relatedness within Clade B was significantly explained by the geographic overlap of their hosts' ranges rather than host relatedness, indicating that they are capable of host switching opportunistically. This has important bearings on their potential to become panzootic. Together, these analyses suggest that NWA have not codiverged with their hosts and instead have evolved predominantly via host switching.

Keywords: AxParafit, codivergence, cophylogeny, Phylocom, reconciliation, viral haemorrhagic fever

Received 23 November 2011; revision received 1 May 2012; accepted 2 May 2012

Introduction

Arenaviruses (Arenaviridae) are zoonotic viruses, some of which cause severe diseases in humans, primarily characterized by haemorrhagic fevers and/or meningitis with high mortality rates (Peters *et al.* 1996; Jay *et al.* 2005). Arenaviruses are lipid-enveloped bi-segmented ambisense single-stranded RNA viruses classified into two main groups, separated both serologically and geographically, into the Old World (Lymphocytic choriomeningitis virus) and New World (Tacaribe) complexes (Bowen *et al.* 1997). The Old World arenaviruses (OWA) contain 10 species, two of which cause diseases in humans with high mortality rates (Ogbu *et al.* 2007; Briese *et al.* 2009; de Bellocq *et al.* 2010). The New World arenaviruses (NWA) contain 23 viral species that are subdivided into four well-established lineages; lineages A, B and C occur in Latin America (Bowen *et al.* 1996) and lineage A-Recombinant (A-Rec) in North America (Charrel *et al.* 2001; Archer & Rico-Hesse 2002). Five viruses of lineage B, Junín, Machupo, Guanarito, Sabiá and Chapare, cause severe haemorrhagic fever with a 10–33% case mortality (Gonzalez *et al.* 1996; Delgado *et al.* 2008). Viruses of lineages A, A-Rec and C are not known to be pathogenic to humans, with the exception of Flexal, which caused nonfatal infections in laboratory workers (Peters *et al.* 1996) and Whitewater Arroyo, which was implicated in three fatalities (Enserink 2000).

The natural hosts of arenaviruses are rodents (Kunz 2009), with the exception of the Tacaribe virus discovered

Correspondence: Nancy R. Irwin, Fax: 44 (0)1904 328505; E-mail: nancy.irwin@york.ac.uk

in two *Artibeus* bat species (Downs *et al.* 1963). Infected rodents excrete the viruses through their urine and saliva (Childs & Peters 1993b). Humans get infected either by cuts or by bites or via inhalation or ingestion of excreta particles (Dylla *et al.* 2008). People with regular direct or indirect contact with wildlife, agricultural workers (Childs *et al.* 1995), laboratory workers (Rousseau *et al.* 1997) and in domestic residences (Kilgore *et al.* 1995) are particularly susceptible. There are also rare incidences of human-to-human infection by either close contact (McCormick *et al.* 1987), nosocomial infection (Carey *et al.* 1972; Briese *et al.* 2009) or organ transplantation (Palacios *et al.* 2008). Feral rodents have infected hamsters and guinea pigs for the pet trade (Skinner & Knight 1979) and laboratory mice (Skinner *et al.* 1977; Smith *et al.* 1984), which, in some cases, have gone on to infect humans (Bowen *et al.* 1975; Smith *et al.* 1984; Rousseau *et al.* 1997). Such virulent properties mean that the pathogenic arenaviruses have been classified in the highest category for infectious agents (Category A) and even developed with the intent to be used as a biological weapon (Borio *et al.* 2002).

The arenavirus genome consists of two single-stranded RNA segments designated as the small segment (S) and the large segment (L) (Bowen *et al.* 1997). The S-segment (~3500 bp) encodes the nucleocapsid protein (NP) and glycoprotein precursor (GP). The L segment (~7100 bp) encodes the zinc-binding protein (Z protein) and the RNA-dependent RNA polymerase (L protein). Proteins of each segment are separated by intergenic noncoding regions. The GP undergoes post-translational cleavage to generate envelope proteins G1 and G2 and a signal peptide (Charrel *et al.* 2002). The GP protein is involved in cell entry to the hosts and is therefore critical to the transmission of the virus (Radoshitzky *et al.* 2008). To date, two cell receptors have been identified as targets for cell entry. The first, the alpha-dystroglycan (α-DG) receptor is used by OWA (Cao 1998) as well as Latino and Oliveros viruses from the NWA (Spiropoulou *et al.* 2002). The second is the alpha transferrin receptor 1 protein (TfR1), an iron transport receptor, used by Clade B of the NWA (Radoshitzky *et al.* 2007; Flanagan *et al.* 2008; Abraham *et al.* 2009). These viruses also exploit an independent cell entry pathway, currently not identified (Flanagan *et al.* 2008). It is currently not known which receptors Clade A or A-Rec of NWA use to enter the cell (Spiropoulou *et al.* 2002; Reignier *et al.* 2008).

Rodent species that are considered the primary hosts are found to suffer from a persistent infection of a single virus species and are capable of both vertical and horizontal transmission (Childs & Peters 1993b). The ability to survive with arenavirus infection is interpreted as evidence of a long evolutionary adaptation between the virus and its host (Childs & Peters 1993b; Briese *et al.* 2009; Kunz 2009). Each virus is thought to be associated with a particular host, and therefore, the distribution of the host is likely to determine the distribution of the virus (Charrel & de Lamballerie 2010). The only globally distributed arenavirus is Lymphocytic chorio-meningitis virus mediated by the distribution of its host, the house mouse *Mus musculus* (Salazar-Bravo *et al.* 2002). Rodents from the subfamily Murinae host OWA, whereas in the New World, arenaviruses infect members from the endemic subfamilies Sigmodontinae and Neotominae (Bowen *et al.* 1997). Given that both New and Old World arenaviruses and their respective hosts form distinct evolutionary units, co-evolution dating back to when the rodent groups diverged (20–30 Mya) is often hypothesized (Childs & Peters 1993a; Bowen *et al.* 1997; Mills *et al.* 1997; Salazar-Bravo *et al.* 2002; Gonzalez *et al.* 2007). Yet, there is evidence that some viruses have switched host, in that some viruses are found in more than one host species, for example Bear Canyon virus infects both *Neotoma macrotis* and *Peromyscus californicus* (Fulhorst *et al.* 2002; Cajimat *et al.* 2007), and some host species are known to harbour more than one virus from different clades present in the same region, for example *Zygodontomys brevicauda* in Venezuela has both Pirital virus (Clade A) and pathogenic Guanarito (Clade B) (Weaver *et al.* 2000).

Bowen *et al.* (1997) hypothesized that the evolution of arenaviruses would be best explained by co-evolution (i.e. codivergence) with the occasional host switching, but they lacked a phylogeny of New World rodents to empirically test their hypothesis. Phylogenetic reconciliation analysis has previously been limited to OWA (Hugot *et al.* 2001; Hugot 2003; Coulibaly-N'Golo *et al.* 2011) and a study that included only 12 of the known 23 NWA (Jackson & Charleston 2004). Despite the repeated assertion of a long history of codivergence (Jay *et al.* 2005; Gonzalez *et al.* 2007; Radoshitzky *et al.* 2008), the studies showed no evidence of congruence between the viral and host phylogenies. While Jackson & Charleston (2004) concluded that no cophylogenetic association existed, Hugot *et al.* (2001) concluded that diffuse co-evolution (i.e. host switching on related hosts) occurred on top of an ancient association between arenaviruses and their hosts. In the most recent assessment of the OWA, the authors concluded that multiple host switching and extinction events may have obscured any co-evolutionary signal (Coulibaly-N'Golo *et al.* 2011). With increasing availability of molecular data from both the NWA and rodents, improved host knowledge and advances in methodology, AxParafit (Stamatakis *et al.* 2007), phylogenetic

clustering (Webb *et al.* 2008), phylogenetically constrained Mantel tests (Harmon & Glor 2010), it is now possible to critically reassess the codivergence hypothesis.

In this article, we tested the codivergence hypothesis for New World arenaviruses. We found no evidence that NWA have codiverged with their hosts, but instead that they have evolved via host switching. We found that each NWA clade is differently associated with their hosts and demonstrate either host relatedness or geographic proximity of their hosts has been the drivers of host switching.

Materials and methods

We downloaded all available nucleotide sequences of arenaviruses from GenBank. We used complete coding region sequences of GP, NP, L and Z proteins for phylogenetic analyses. We analysed all available mitochondrial sequences of the complete cytochrome *b* gene of all known hosts as well as all Sigmodontine and Neotomine rodents of the Americas from GenBank. Nucleotide sequences were aligned according to amino acid alignment using global translation alignment in Geneious Pro v4.7 (Biomatters Ltd.) (Drummond *et al.* 2009). For each protein, we identified identical sequences and removed them from the alignment in DAMBE 4.5 (Xia & Xie 2001).

Host–pathogen associations

We assembled all known wild natural hosts of arenaviruses from a literature review (host taxa = 24, associations = 33; Table S1, Supporting information). The hosts for the viruses Chapare and Sabiá are unknown and are therefore not included in any analysis. Likewise, missing host sequence data for *Oecomys paricola* and *Artibeus lituratus* prevented their viral associations being included. The poor taxonomic knowledge of South American rodents may lead to misleading associations and patterns if virus collectors in the field did not collect a sample of the host at the same time. Here, taxonomic revision has changed the taxonomic names of five of the hosts of NWA (see Supporting information). All other associations are firmly established in the literature (Table S1, Supporting information).

Phylogenetic reconstruction

We reconstructed phylogenetic trees for each protein individually using the GTR+Γ+I substitution model selected by all criteria in Modeltest 3.7 (Posada & Crandall 1998). No sequence data were available for the Pinhal virus; therefore, 22 taxa were used in the GP (1626 bp, $n = 62$) and NP (1770 bp, $n = 64$; Table S2, Supporting information) phylogenetic analysis. Tonto Creek, Catarina, Big Brushy Tank and Skinner Tank viruses are also missing data for the L and Z proteins, and these were therefore reconstructed with the remaining 18 NWA taxa (L protein, 6978 bp, $n = 37$; Z protein, 333 bp, $n = 51$; Table S2, Supporting information). Bayesian inference (BI) analysis was run with Markov chain Monte Carlo (MCMC) for two million generations in MrBayes 3.1 (Ronquist & Huelsenbeck 2003), discarding 10% of sampled trees as burn-in. Maximum likelihood (ML) analysis was executed using GTRCAT model in RAxML 7.1.0 (Stamatakis *et al.* 2004). Node support for ML trees was estimated from 10,000 rapid bootstrap replicates. In the NWA analysis, an OWA Mopeia virus was used as the outgroup. Trees were visualized in FigTree v1.2 (Drummond & Rambaut 2007). To estimate the host phylogeny, we constructed a ML phylogenetic tree from 523 of complete mitochondrial cytochrome *b* (1,140 bp) sequences of up to three sequences per taxon for known hosts and all rodents from subfamilies Sigmodontinae and Neotominae (Table S3, Supporting information) using the same methodology as used for viruses with 3,000 rapid bootstrap replicates. This phylogenetic tree includes two hosts from different orders, a mustelid and a bat.

Reconciliation analysis

We tested whether the viruses were distributed nonrandomly with respect to their hosts using two complementary approaches: coassociation analysis and phylogenetic clustering analysis.

Coassociation analysis. New World arenaviruses contain eight viruses that have multiple hosts. Coassociation on mapping the pathogen association onto the host phylogeny (e.g. Treemap) therefore cannot be used, as it is impossible to infer host order (Jackson & Charleston 2004). We therefore tested the coassociation of arenaviruses with their known hosts using ParaFit (Legendre *et al.* 2002) as implemented in the optimized algorithm of AxParafit (Stamatakis *et al.* 2007) used with the CopyCat GUI (Meier-Kolthoff *et al.* 2007). Parafit uses patristic distances from both the host and virus phylogenies to test whether the null hypothesis of independence is significantly rejected, both globally (across the whole tree) and individually (for each viral association). Unlike other cophylogeny methods, it does not infer an evolutionary scenario. The advantages of this method are that it can handle large data sets, viruses with multiple hosts and has been shown to have reliable type-I and type-II errors (Legendre *et al.* 2002). We used a host tree of 21 taxa and tested associations for each

pruned viral protein tree individually (GP viruses = 22, associations = 31; NP viruses = 22, associations = 31; L viruses = 18, associations = 27 and Z viruses = 18, associations = 27). Significance of individual host–virus associations was calculated from 9999 permutations per row of the association matrix. We corrected for multiple comparisons by using a false discovery approach (Benjamini & Hochberg 1995).

We tested the sensitivity of our results to missing data by removing one host species from the analysis and re-running the AxParafit analysis to evaluate whether the patterns of significance changed. We repeated this for all 21 host taxa. The number of associations varied from 28 to 30 depending on which host was removed.

We tested if the global significance of NWA with their hosts was driven by a clade of viruses by removing each viral clade from the coassociation analysis (removing Clade A-Rec; associations = 19, viruses = 13, hosts = 14; removing Clade A, associations = 25, viruses = 15, hosts = 18; removing Clade B; associations = 19, viruses = 14, hosts = 15 or removing Clade C; associations = 29, viruses = 18, hosts = 21).

Clustering analysis. New World arenaviruses phylogenetic host specificity can also be examined using indices of phylogenetic diversity (Poulin *et al.* 2011). Patterns of association at different taxonomic levels can be tested, as the host tree is not restricted to only known host taxa. Here, we constructed a host phylogeny of all available American rodent species in the subfamilies Sigmodontinae and Neotominae and pruned it to one individual per species (n = 269). We also retained the non-rodent hosts, a bat and a mustelid, in the phylogeny (n = 271). We tested whether hosts of arenaviruses were randomly spread, over-dispersed or clustered across the host phylogeny using Phylocom (Webb *et al.* 2008). We did this by calculating the mean phylogenetic distance (MPD) and the nearest phylogenetic taxon distance (MNTD) of the virus hosts and compared these values to those obtained from randomizing host species labels across the host phylogeny. Abundance of viral species found on each host was incorporated into the test, to account for host species that can be infected by more than one virus (either from a single or multiple clades). Two-tailed *P*-values were calculated from 9,999 permutations. We included Pinhal in the analysis as although there is no sequence data for the virus, the host *Calomys tener* is known. We therefore tested the significance of 32 NWA viral associations found in 22 hosts across all rodents of the subfamilies Sigmodontinae and Neotominae. We then tested each NWA clade separately by pruning the host phylogeny to the rodents that occur in the countries where each arenavirus clade is known. Distributions of each rodent species were determined from the Global Mammal Assessment (IUCN 2010) and the Global Biodiversity Information Facility data portal (GBIF 2012) (Table S3, Supporting information).

Geographical overlap between hosts

To evaluate whether host specificity was primarily constrained by host relatedness or whether there was a spatial component, we examined host range overlap of all virus pairs. The distribution of each rodent host species (n = 19) was downloaded from the Global Mammal Assessment (IUCN 2010). Each file was then processed using R (R Development Core Team, 2.12 2010) to create geo-referenced maps of host distributions for each arenavirus lineage. These maps were then used to calculate the amount of geographical overlap between all hosts (in km^2, then log transformed after adding a constant of 1 for all subsequent analyses). A genetic distance matrix was calculated for both viruses (using GP sequences, n = 19) and hosts (n = 19) using Geneious Pro v4.7 (Drummond *et al.* 2009). A number of Mantel tests (simple and partial) were then performed using 29 associations. The simple Mantel tests assessed whether the genetic distance between two viruses was significantly associated with the genetic distance of their respective hosts and/or the geographic overlap existing between their respective hosts. However, because we can expect closely related hosts to overlap more than distantly related hosts, partial Mantel tests were also performed to ascertain whether the above associations were direct or indirect ones. To account for viruses that have more than one host, we calculated the average range overlap between their respective hosts using an arithmetic mean (on the log transformed overlap). Similarly, the average genetic distance between two sets of hosts was calculated using a geometric mean (after adding a constant of 0.01, to account for some hosts being shared by two viruses).

To test for global associations across the whole viral tree, their host and geographic ranges, a Mantel test informed by phylogeny (Lapointe & Garland 2001) was used to reduce the risk of type-I error associated with the classical Mantel test (Harmon & Glor 2010). This permutation procedure swaps closely related species more often than distantly related species across the phylogenetic tree (Lapointe & Garland 2001). However when assessing relationships within each clade of viruses, the classical Mantel test was applied instead as the small sample size (number of viruses in a single clade) would not guarantee a sufficient level of permutation (i.e. the identity of the viruses in the permutated trees would remain too often the original identities) with a Mantel

test informed by phylogeny. To minimize type-I error, the partial Mantel tests (whether phylogeny informed or not) were performed using the permutation procedure on raw data (virus genetic distance) (Legendre 2000). Analyses were completed in R using modified algorithms from Harmon & Glor (2010). To assess the level of significance, 9999 permutations were used.

Results

Phylogeny of New World arenaviruses

All phylogenetic analyses supported four previously recognized evolutionary lineages A, B, C and A-Rec (Fig. S1A–D, Supporting information). The GP phylogeny places Clade A-Rec as sister to Clade B, whereas the NP, Z and L protein place it as sister to A. This well-known incongruence has been shown to be the result of a recombinant progenitor (Charrel *et al*. 2001; Archer & Rico-Hesse 2002). Our larger phylogenetic trees support this conclusion. The rodent phylogenetic tree supported reciprocal monophyly of subfamilies Sigmodontinae and Neotominae, but relationships between genera of Sigmodontinae were poorly resolved (Fig. 1) as has been demonstrated in larger studies using multiple loci (Engel *et al*. 1998; Jansa & Weksler 2004).

Host–viral associations

Coassociation analysis. Global co-evolutionary association of NWA with their known hosts was statistically significant for GP (ParafitGlobal Test Statistic = 51.105, $P = 0.0004$), NP (ParafitGlobal = 27.644, $P = 0.0003$) and L protein (ParafitGlobal = 34.656, $P = 0.0239$), but not significant for Z protein (ParafitGlobal = 22.978, $P = 0.1452$). However, the significance of individual hosts associations varied. In the host–virus associations based on GP sequences, 22 of 31 host–virus associations were not significantly explained by topology (Fig. 2). Three of these associations were found to be on the border of significance ($P = 0.05$) but were not significant after correction for multiple testing. The only significant associations found in NWA were those from Clade A-Rec hosted by woodrats of the genus *Neotoma*. Further analysis of the other proteins showed that the same nine A-Rec viruses were significantly associated with their hosts for the NP, but no cophylogenetic signal was found for any virus with any host for the L and Z proteins (Fig. S2, Supporting information). The results were very similar for analyses based on the GP and NP trees. In subsequent analyses, we chose to use the GP data set as it is known to be associated with host cell entry recognition (Radoshitzky *et al*. 2008) and was

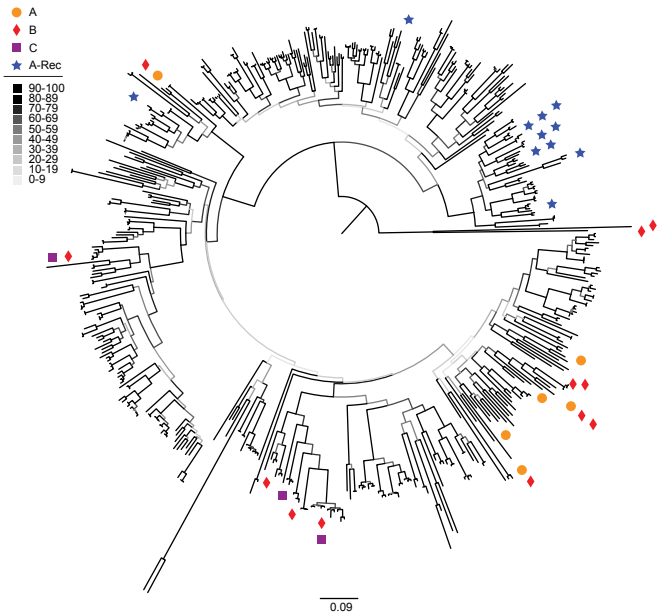


Figure 1 Phylogenetic tree of all Sigmodontinae and Neotominae rodents with hosts of each New World arenavirus indicated. A ML phylogenetic tree of mitochondrial cytochrome *b* of all available species ($n = 523$). Bootstrap support is indicated by grey-scale coding of respective branches. Host species (Table S1, Supporting information) shown for each New World arenaviruses (NWA) clade.

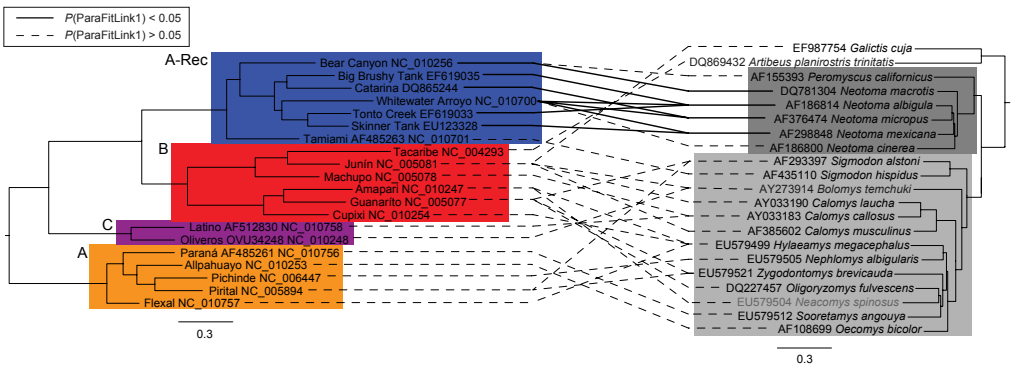


Figure 2 Host–virus associations between New World arenaviruses and their known hosts. 31 associations are drawn on trees pruned to one individual per species from a Bayesian glycoprotein arenavirus tree (n = 22) (Fig. S1A, Supporting information) and known hosts (n = 21) (Fig. 1). When there were no data for a host, we chose the most closely related species to represent the association as indicated by grey type (Table S1, Supporting information). The four New World arenaviruses (NWA) clades are indicated at the base of each clade; the Neotominae rodents (dark grey) and the Sigmodotinae rodents (pale grey). Significance of host–virus associations were calculated from 9,999 permutations, using ParaFit, (Legendre *et al.* 2002). Full lines represent significant cophylogenetic association between specific host-arenavirus pairs, dashed lines represent nonsignificant relationships.

therefore more likely to recover host–virus cophylogeny than the NP.

Testing the sensitivity of the coassociation analysis. Removing 16 of the hosts did not change the pattern of significance in the coassociation matrix. The same nine A-Rec viruses remained significantly associated with their host. However, we found that removing either nonrodent host (the mustelid or the bat) resulted in the Bear Canyon virus becoming significantly coassociated with *Peromyscus californicus*. We found that removing one of the *Neotoma* species that had multiple viral associations (*N. micropus*, *N. mexicana* or *N. albigula*) resulted in all associations previously seen in A-Rec viruses no longer remaining significant. The global statistic for coassociation remained significant after removing viral Clade A, B or C (P = 0.0001, P = 0.0002, P = 0.0001). However, removing Clade A-Rec caused the global association statistic to become nonsignificant (P = 0.41414).

Clustering analysis. New World arenaviruses were randomly distributed across the rodent tree (n = 22, MPD P = 0.137; MNTD P = 0.442). Clade A-Rec (n = 7) viruses were significantly clustered with their hosts for both analyses (MPD P > 0.001; MNTD P = 0.039). Clade C, limited to only 3 viruses, were not significant in either analyses (MPD P > 0.270; MNTD P = 0.162). Clade A (n = 6) were significantly clustered when compared to all available hosts (MPD P = 0.026), but not clustered into particularly closely related taxa (MNTD P = 0.475). Clade B (n = 11), however, were significantly over-dispersed across the phylogenetic tree, for both indices (MPD P = 0.021, MNTD P = 0.047) when the two hosts of different orders were included in the analysis. Removing these two taxa, Clade B was shown to be randomly distributed amongst the rodents (MPD P = 0.164 and MNTD P = 0.964). These results are summarized in Table 1.

Distribution trends for each virus clade

The potential extent of each NWA clade has been mapped by plotting the distribution of their known hosts (Fig. 3). Hosts of Clade A arenaviruses are all distributed in South America (not further south than Uruguay) and part of Central America (not further west than Costa Rica) (Fig. 3). Hosts of Clade B arenaviruses are located in South and Central America (south of Mexico). However, two subclades of Clade B appeared separated by a major barrier to dispersal, the Amazon river. The hosts of subclade B1 (Tacaribe, Junín, Machupo) (Charrel *et al.* 2002) are all located south of the Amazon River except for Tacaribe whose hosts are widespread tropical bat species (*Artibeus lituratus* and *Artibeus jamaicensis*). The hosts of subclade B2 (Amapari, Cupixi and Guanarito) (Charrel *et al.* 2002) are all located north of the Amazon River, except for one host, *Hylaeamys megacephalus*, occurring from Venezuela to southern Brazil. This species, however, is a complex of three lineages; the true host is therefore likely to be an unnamed species with a distribution north of the Amazon river (Miranda *et al.* 2007). Hosts

Table 1. Clustering analysis of New World arenaviruses (NWA)

Virus	No. of viruses	No. of host taxa	No. of nonhosts	MPD	MPD random	MPD *P* value	MNTD	MNTD random	MNTD *P* value
NWA	32	22	249	0.5717	0.5268	0.1372	0.3129	0.2889	0.442
NWA (restricted to rodents)	30	20	249	0.5094	0.5241	0.6302	0.2962	0.0348	0.558
Clade A-Rec	11	7	103	0.2781	0.4596	0.000	0.2496	0.3849	0.039
Clade A	6	6	140	0.3642	0.4635	0.026	0.3540	0.4004	0.475
Clade B	12	11	246	0.6166	0.5033	0.021	0.4378	0.3409	0.047
Clade B (restricted to rodents)	11	9	246	0.4409	0.4909	0.164	0.3561	0.3584	0.964
Clade C	3	3	125	0.3095	0.3713	0.270	0.3432	0.4888	0.162

The phylogenetic tree of all rodents of the Sigmodontinae and Neotominae subfamilies (Fig. 1) was pruned to one sequence per species (n = 269) and included two nonrodent hosts (n = 271). Clustering analysis tested the significance of phylogenetic distance (MPD, Mean Phylogenetic Distance) and the distance of the nearest neighbour (MNTD, Mean Nearest phylogenetic Taxon Distance) between known hosts by comparison with the values generated by 9,999 permutations according to a two-tailed test. Clustering analysis of all NWA across all rodents of these subfamilies was tested. Each virus clade was also tested independently against the phylogeny restricted to their sympatric rodents (Table S3, Supporting information). Two hosts of Clade B virus are found in different orders of mammals; the analysis was therefore also run with and without these taxa for NWA as a whole and for Clade B.

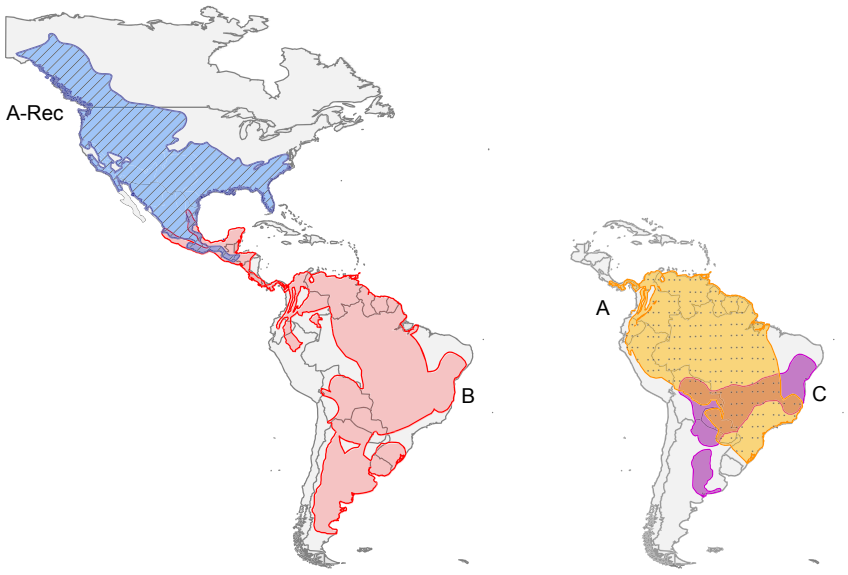


Figure 3 Distribution map of hosts of New World arenavirus Clades. Distributions of hosts (Table S1, Supporting information) were downloaded from the IUCN Global Mammal Assessment Team (IUCN 2010) and processed using R (R Development Core Team, 2.12 2010). Each clade is indicated as follows; figure to the left, Clade A-Rec –stripes (7 host species) Clade B—(9 host species); figure to the right, Clade A—dots (6 host species), Clade C— (3 host species).

of Clade C all occur south of the Amazon River, but, interestingly, the host of Oliveros (*Necromys benefactus*) is restricted to central Argentina and does not overlap with the other Clade C hosts (*Calomys callosus* and *C. tener*). Other hosts for Clade C may therefore exist in North Argentina and Paraguay. Hosts of Clade A-Rec all occur in North America and Mexico, except for *Neotoma mexicana*, which also occurs in Central America as far east as Guatemala (Fig. 3). These maps extend the areas where NWA are predicted to occur, in particular Clade A-Rec's hosts are overlapping with Clade A's hosts in Central America. NWA are likely to be present in Central America, which has been sparsely sampled to date.

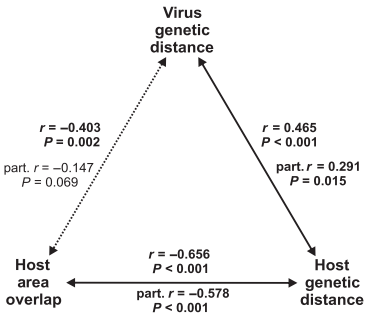


Figure 4 Mantel tests of virus genetic distance to host genetic distance and host area overlap. Simple and partial Mantel tests of each relationship are presented with their corresponding *r* value and significance level. See Table 2 for clade-specific results.

Is viral phylogeny correlated to host phylogeny or to host geographical overlap?

Across all NWA, as the genetic distance between two arenaviruses increased, so did the genetic distance between their respective hosts (partial $r = 0.291$, phylogenetically constrained Mantel $P = 0.015$) (Fig. 4). In turn, as the genetic distance between the hosts of two viruses increased, the spatial overlap existing between them tended to decrease (partial $r = -0.578$, phylog. constr. Mantel $P < 0.001$). The genetic distance between two viruses, however, was not significantly correlated to the amount of spatial overlap existing between the hosts of these viruses (partial $r = -0.147$, phylog. constr. Mantel $P = 0.069$). At the clade level, however, the patterns varied widely and differed markedly from the overall pattern. Clade C had too few viruses to enable analysis to be completed at clade level. Only the viral genetic distance was significantly correlated to the spatial overlap of their respective hosts for Clades A and B (contrary to the overall pattern) (Table 2). For Clade A-Rec, only the correlation between the viruses' genetic distance and the genetic distance of their hosts was significant.

Discussion

We empirically tested whether NWA had codiverged with their hosts using three complementary approaches: reconciliation analysis, phylogenetic clustering and Mantel tests. We found no evidence to support codivergence of NWA with their hosts. Instead, the evidence suggests that host switching has occurred frequently, though to a different extent in the different clades.

Phylogenetic analyses of host-virus associations

In cophylogenetic analysis of NWA confined to known hosts, we found a globally significant association between the viral and host phylogenies with all viral protein trees (GP, NP, L and Z). However, it was only in the larger data sets with more structured trees (NP and GP analysis) that nine individual Clade A-Rec viral associations were significantly associated with their hosts. Testing the sensitivity of the AxParafit analysis showed the results to be generally robust. The ParaFit global statistic only became nonsignificant after removing Clade A-Rec, demonstrating that this clade was driving the global significant association. Furthermore, removing any host that had multiple viral associations with A-Rec viruses caused all other associations to become nonsignificant.

The power to discriminate a significant association (and thus avoid type-II error) in Parafit is associated with completeness of knowledge of all the hosts and viruses in the system (Legendre *et al*. 2002). Nine NWA have been discovered in the past decade (Charrel & de Lamballerie 2010), making it unlikely that all hosts and viruses are currently known. Unsupported phylogenies,

Table 2. Assessment of geographic overlap and host phylogeny on viral phylogeny

Lineage	Virus genetic distance vs. Host genetic distance	Virus genetic distance vs. Host area overlap	Host genetic distance vs. Host area overlap
A	$r = 0.029$	$r = -0.586$*	$r = 0.019$
	part. $r = 0.050$	part. $r = -0.587$*	part. $r = 0.045$
A-Rec	$r = 0.581$*	$r = -0.372$	$r = -0.232$
	part. $r = 0.548$*	part. $r = -0.300$	part. $r = -0.021$
B	$r = 0.680$	$r = -0.757$**	$r = -0.546$**
	part. $r = 0.487$	part. $r = -0.627$**	part. $r = -0.066$

Coefficients of correlation (simple and partial) between the genetic distance separating two arenaviruses, the genetic distance separating their respective hosts and the amount of spatial overlap existing between these hosts. Significant results are indicated by a star, *$P < 0.05$, **$P > 0.001$.

for example rapid radiations, also add to uncertainty in permutation testing procedures. Here, despite the viral phylogeny being well supported, the Sigmodontinae higher taxonomy was difficult to confidently resolve (Jansa & Weksler 2004) and may have contributed to the lack of statistical significance. While the improved algorithms for permutation tests of reconciliation analysis enable multiple hosts and large number of taxa associations to be tested, if one or both of the phylogenies have uncertainty, other statistical approaches are needed to assess the nature of host–virus relationships.

Clustering analysis

To explore viral–host associations further, we examined whether patterns of phylogenetic clustering were significantly different from chance expectations across all the potential Sigmodontinae and Neotominae rodent hosts using different indices of relatedness. We found that NWAs collectively have invaded a wide range of rodent hosts, which were randomly distributed across the rodent phylogeny. However, different patterns were recovered for each Clade when analysis was restricted to rodents within their geographic range. Clade A-Rec was the most tightly constrained clade of NWA viruses, being significantly clustered with their hosts. Clade A was less constrained, being clustered in subsets of the rodent tree but beyond that, not in particularly closely related taxa. Worryingly, Clade B with pathogenic members was the least constrained being randomly distributed across the rodent tree and even able to infect hosts of different mammalian orders.

Geographic proximity analysis

We examined further the relationship between viral phylogeny and host geographical overlap with simple and partial Mantel tests to see whether we could tease out any explanatory factor that differed between clades. Viral phylogeny was explained by host phylogeny only for Clade A-Rec, which was congruent with all other analyses. Host relatedness is therefore a factor that has constrained host switching in Clade A-Rec, with few exceptions. Geographical overlap of hosts was more important than host relatedness in shaping the evolution of both Clades A and B, indicating that physical contact with a new host to infect was more limiting than the need to find a sufficiently related host. Host range is clearly a rather crude proxy for virus distribution, as true viral distribution will be dependant on many factors including host demography and pathogen prevalence (e.g. Mills 1994). This makes our results even more striking, as it was expected a priori that phylogenetic relatedness of hosts should constrain host switching because of host defence and cell entry mechanisms.

Refuting the codivergence hypothesis for NWA

A long-standing codiverged relationship has been asserted for NWA because Clade B viruses infected on the whole Sigmodontinae, while Clade A-Rec viruses primarily infected *Neotoma* species in the subfamily Neotominae (Webb *et al.* 1970; Arata & Gratz 1975; Childs & Peters 1993b; Bowen *et al.* 1997; Mills *et al.* 1997; Mills 1999; Salazar-Bravo *et al.* 2002; Charrel *et al.* 2003; Cajimat *et al.* 2007; Gonzalez *et al.* 2007). The odd viruses such as Bear Canyon or Tamiami that did not follow this pattern were therefore thought to represent rare host switches to a different group of rodents (Cajimat *et al.* 2007). Here, we found host switching to be the predominant pattern of association.

Classically, codivergence implies a pattern of high host specificity where one host is infected by one parasite (Fahrenholz's Rule) (Paterson & Banks 2001). However, a comprehensive literature review showed that one quarter of the NWA infected multiple hosts and one-third of the host species were infected by more than one NWA virus.

Doubts about the validity of the codivergence hypothesis have also been raised from the observation that several subfamilies of rodents sister to the host subfamilies are devoid of arenaviruses (Salazar-Bravo *et al.* 2002). While LCMV (in the Old World) has been recorded in Arvicolinae (Laakkonen *et al.* 2006) and Cricetinae (Skinner & Knight 1979), wild infections in Gerbillinae, Tylomyinae and Deomyinae have never been recorded. This either can be interpreted as lack of field sampling in these groups, extinction of viral lineages (Paterson & Banks 2001), or as evidence that arenaviruses did not coevolve with their hosts. While sampling deficiency is always possible, extinction across three major lineages seems unlikely.

Likewise, topological incongruence between hosts and viruses could be explained by extinction events or lack of sampling (Charleston & Perkins 2006). Although extinctions will have contributed to the NWA system, we believe it has not dominated the pattern of coassociation given that (i) Clade A and B are so widely distributed across the Sigmodontinae and Neotominae tree it would imply widespread extinctions across a large range of taxa; and (ii) physical proximity can already significantly explain the host species that Clades A and B have managed to invade.

Temporal incongruence between the viral and host trees would also support the rejection of the codivergence hypothesis. However, limited heterochronous data for NWA mean that molecular dating is currently

not available. However, substitution rates for OWA (Coulibaly-N'Golo *et al.* 2011) have been estimated at the typical rate for RNA viruses of 10^{-3} to 10^{-4} substitutions per site per year (Domingo & Holland 1997) resulting in dates for the time to the most recent common ancestor (TMRCA) of between 2500 and 7000 ybp (Coulibaly-N'Golo *et al.* 2011). These dates are orders of magnitude younger than the 22 million years postulated by the codivergence hypothesis (Bowen *et al.* 1997). A growing body of RNA virus studies with heterochronous data has also produced speciation events much younger than expected (Duffy *et al.* 2008; Holmes 2008; Ramsden *et al.* 2009; Pagán *et al.* 2010), which has led some to suggest a methodological problem with inferring dates from short time series, especially for the deeper nodes (Worobey *et al.* 2010; Ho *et al.* 2011). Dating inferred from host speciation events based on encapsulated RNA is also orders of magnitude older than those based on molecular clock estimates (Horie *et al.* 2010; Katzourakis 2010; Worobey *et al.* 2010). This paradox can be explained by high lineage turnover in rapidly evolving populations (Holmes 2003, 2008). The molecular clock approach calculates substitution rates based on the genetic variation of current circulating viruses. It does not provide the absolute age of when these viruses split from their sister taxon, but the TMRCA of the viruses sampled (Holmes 2003, 2008). The shallow dates for RNA viruses reflect this. The rapid turnover is therefore likely to remove any signal of ancient history or association between RNA viruses and their hosts.

Each NWA clade has a unique pattern of relationships with their hosts, indicating that each clade has evolved differently. Clade A-Rec is comprised of specialists that infect only a single host or a small number of closely related hosts. The general pattern of cophylogeny for Clade A-Rec viruses with their hosts does not imply codivergence but can be equally explained by host switching (Holmes & Price 1980). In a system dominated by 'preferential' host switching, viruses are more likely to switch successfully on closely related hosts as they are preadapted genetically, physiologically and ecologically to the original host (Charleston & Robertson 2002). These viruses then become adapted to the new host, speciating and obscuring the host-switching signal (Brook & Mclennan 1991). Even the phylogenetically distant host switches seen in Bear Canyon and Tamiami are encompassed, as the concept does not imply that there will be no distant host switching only that it will occur less frequently (Charleston & Robertson 2002). Conversely, the proximity analysis suggested that ecological opportunity to infect new hosts rather than relatedness was the driver for host switching for the viruses of Clades A and B.

The different pattern of association shown for each lineage of NWA implies that each lineage is constrained differently to enter new hosts. This hypothesis is supported by experimental and structural studies, which have demonstrated that cell entry mechanisms for NWA are key to understanding the potential for host spread and disease outbreak (Flanagan *et al.* 2008; Choe *et al.* 2011). For example, currently nonpathogenic viruses of Clade B with only minor changes in the apical domain of the GP used as the cell entry receptor target could emerge as new diseases (Abraham *et al.* 2009, 2010). Pathogenic viruses with the ability to infect a wide range of hosts are of serious concern, but experimental infection has shown that Clade B viruses were unable to infect some nonhost species such as the house mouse and black rat, and that each virus species had a different ability to infect native hosts and humans (Radoshitzky *et al.* 2008; Abraham *et al.* 2010). Switching to a distantly related host may therefore remain a rare phenomenon followed by rapid adaptation to the new host and ultimately resulting in viral speciation.

Owing to the difficulties of screening highly infectious viruses in the wild that are both spatially and temporally patchy in their distribution (Jay *et al.* 2005), sequencing the five interaction motifs of the TfR1 receptor (Abraham *et al.* 2010) of potential hosts may be a targeted approach to predict new hosts. Evidence from the phylogenetic clustering and reconciliation analysis here, as well as laboratory infections (Lukashevich *et al.* 2002; Abraham *et al.* 2010) and examination of the cell receptors (Abraham *et al.* 2009), would suggest that bats, monkeys and other Sigmodontinae and Neotominae rodents, currently not known to be hosts of Clade B, should be screened to understand better the potential for host switching and disease spread. The ability of both OWA and NWA to switch to unrelated hosts including humans makes the development of a general vaccine and/or therapeutic approach (Larson *et al.* 2008; Kotturi *et al.* 2009; Botten *et al.* 2010) all the more important especially as not all viruses may yet be known.

Summary

We found no evidence that NWA have codiverged (associated by descent) with their hosts. Similarly, the pattern of association of Old World arenaviruses (OWA) with their known hosts could not be distinguished from random (Coulibaly-N'Golo *et al.* 2011). However, this led the authors to conclude that the frequent host switching may have obscured an ancient codivergent pattern of association (Coulibaly-N'Golo *et al.* 2011). Here, we would have drawn a similar conclusion for NWA (with the possible exception of

Clade A-Rec), but the addition of phylogenetic clustering analysis demonstrated that each clade of NWA was associated differently with the hosts available to them and that only Clade B was truly randomly distributed. We found that the phylogenetic clustering approach on all the potential hosts was more robust than reconciliation analysis limited to known hosts. Likewise, the Mantel tests suggested that the potential for host switching differed between clades. The diversification of two clades was driven by geographic proximity of their hosts rather than their relatedness, indicating host switching was opportunistic, whereas the diversification of Clade A-Rec was driven by host relatedness. We expect these approaches to be informative for other pathogen–host systems as more molecular data become available for construction of large host phylogenies.

Acknowledgements

NRI was supported by a Daphne Jackson Fellowship sponsored by the Natural Environmental Research Council. Biomatters sponsored NRI with a license to use Geneious Pro. NM was supported by Academy of Sciences of the Czech Republic (grant AV0Z60930519). The data were analysed on the computational cluster of the Institute of Vertebrate Biology and at Bioportal of the University of Oslo. We thank the subeditor Dr. Roman Biek and the three anonymous referees for their comments, which helped improve the paper.

References

Abraham J, Kwong J, Albariño C *et al.* (2009) Host-species transferrin receptor 1 orthologs are cellular receptors for nonpathogenic New World clade B arenaviruses. *PLoS Pathogens*, **5**, e1000358.

Abraham J, Corbett KD, Farzan M, Choe H, Harrison SC (2010) Structural basis for receptor recognition by New World hemorrhagic fever arenaviruses. *Nature Structural & Molecular Biology*, **17**, 438–444.

Arata AA, Gratz NG (1975) The structure of rodent faunas associated with arenaviral infections. *Bulletin of the World Health Organization*, **52**, 621–627.

Archer A, Rico-Hesse R (2002) High genetic divergence and recombination in Arenaviruses from the americas. *Virology*, **304**, 274–281.

de Bellocq JG, Borremans B, Katakweba A *et al.* (2010) Sympatric occurrence of 3 arenaviruses, Tanzania. *Emerging infectious diseases*, **16**, 692–695.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society*, **57**, 289–300.

Borio L, Inglesby T, Peters CJ *et al.* (2002) Hemorrhagic fever viruses as biological weapons: medical and public health management. *JAMA*, **287**, 2391–2405.

Botten J, Whitton JL, Barrowman P *et al.* (2010) A multivalent vaccination strategy for the prevention of Old World arenavirus infection in humans. *Journal of Virology*, **84**, 9947–9956.

Bowen GS, Calisher CH, Winkler WG *et al.* (1975) Laboratory studies of a lymphocytic choriomeningitis virus outbreak in man and laboratory animals. *American Journal of Epidemiology*, **102**, 233–240.

Bowen M, Peters C, Nichol S (1996) The phylogeny of New World (Tacaribe complex) arenaviruses. *Virology*, **219**, 285–290.

Bowen M, Peters C, Nichol S (1997) Phylogenetic analysis of the Arenaviridae: patterns of virus evolution and evidence for cospeciation between arenaviruses and their rodent hosts. *Molecular Phylogenetics and Evolution*, **8**, 301–316.

Briese T, Paweska J, Mcmullan L *et al.* (2009) Genetic detection and characterization of Lujo virus, a new hemorrhagic fever–associated arenavirus from Southern Africa. *PLoS Pathogens*, **4**, e1000455.

Brook DR, Mclennan DA (1991) *Phylogeny, Ecology and Behavior; A Research Program in Comparative Biology*. The University of Chicago Press, Chicago, IL, USA.

Cajimat MNB, Milazzo ML, Hess BD, Rood MP, Fulhorst CF (2007) Principal host relationships and evolutionary history of the North American arenaviruses. *Virology*, **367**, 235–243.

Cao W (1998) Identification of -dystroglycan as a receptor for lymphocytic choriomeningitis virus and lassa fever virus. *Science*, **282**, 2079–2081.

Carey D, Kemp G, White H *et al.* (1972) Lassa fever. Epidemiological aspects of the 1970 epidemic, Jos, Nigeria. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **66**, 402.

Charleston MA, Perkins SL (2006) Traversing the tangle: algorithms and applications for cophylogenetic studies. *Journal of Biomedical Informatics*, **39**, 62–71.

Charleston MA, Robertson DL (2002) Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Systematic Biology*, **51**, 528–535.

Charrel RN, de Lamballerie X (2010) Zoonotic aspects of arenavirus infections. *Veterinary Microbiology*, **140**, 213–220.

Charrel R, de Lamballerie X, Fulhorst C (2001) The whitewater arroyo virus: natural evidence for genetic recombination among Tacaribe serocomplex viruses (family Arenaviridae). *Virology*, **283**, 161–166.

Charrel R, Feldmann H, Fulhorst C *et al.* (2002) Phylogeny of New World arenaviruses based on the complete coding sequences of the small genomic segment identified an evolutionary lineage produced by intrasegmental recombination. *Biochemical and Biophysical Research Communications*, **296**, 1118–1124.

Charrel R, Lemasson J, Garbutt M *et al.* (2003) New insights into the evolutionary relationships between arenaviruses provided by comparative analysis of small and large segment sequences. *Virology*, **317**, 191–196.

Childs J, Peters CJ (1993) Ecology and epidemiology of arenaviruses and their hosts. In: *The Arenaviridae* (ed Salvato MS), pp. 331–384. Plenum press, New York.

Childs JE, Peters CJ (1993b) Ecology and epidemiology of arenaviruses and their hosts. In: *The Arenaviridae* (ed Salvato MS), pp. 331–384. Plenum press, New York.

Childs JE, Mills J, Glass GE (1995) Rodent-borne hemorrhagic fever viruses: a special risk for mammalogists? *Journal of Mammalogy*, **76**, 664–680.

Choe H, Jemielity S, Abraham J, Radoshitzky SR, Farzan M (2011) Transferrin receptor 1 in the zoonosis and pathogenesis of New World hemorrhagic fever arenaviruses. *Current Opinion in Microbiology*, **14**, 476–482.

Coulibaly-N'Golo D, Allali B, Kouassi SK *et al.* (2011) Novel arenavirus sequences in Hylomyscus sp. and Mus (Nannomys) setulosus from Côte d'Ivoire: implications for evolution of arenaviruses in Africa. *PLoS One*, **6**, e20893.

Delgado S, Erickson B, Agudo R *et al.* (2008) Chapare virus, a newly discovered arenavirus isolated from a fatal hemorrhagic fever case in Bolivia. *PLoS Pathogens*, **4**, e1000047.

Domingo E, Holland XX (1997) RNA virus mutations and fitness for survival. *Annual Reviews in Microbiology*, **51**, 151–178.

Downs WG, Anderson CR, Spence L, Aitken THG, Greenhall AH (1963) Tacaribe virus, a new agent isolated from Artibeus bats and mosquitoes in Trinidad, West Indies. *The American Journal of Tropical Medicine and Hygiene*, **12**, 640.

Drummond AJ, Rambaut A (2007) BEAST: bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A (2009) *Geneious v 4.7*. Available at http://www.geneious.com/.

Duffy S, Shackelton LA, Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics*, **9**, 267–276.

Dylla DE, Michele DE, Campbell KP, McCray Jr PB (2008) Basolateral entry and release of New and Old World arenaviruses from human airway Epithelia. *Journal of Virology*, **82**, 6034–6038.

Engel SR, Hogan KM, Taylor JF, Davis SK (1998) Molecular systematics and paleobiogeography of the South American sigmodontine rodents. *Molecular Biology and Evolution*, **15**, 35–49.

Enserink M (2000) Emerging Diseases: new arenavirus blamed for recent deaths in California. *Science*, **289**, 842–843.

Flanagan ML, Oldenburg J, Reignier T *et al.* (2008) New world clade B arenaviruses can use transferrin receptor 1 (TfR1)-dependent and -independent entry pathways, and glycoproteins from human pathogenic strains are associated with the use of TfR1. *Journal of Virology*, **82**, 938–948.

Fulhorst C, Bennett S, Milazzo M *et al.* (2002) Bear Canyon virus: an arenavirus naturally associated with the California mouse (Peromyscus californicus). *Emerging Infectious Diseases*, **8**, 717–721.

GBIF (2012) Biodiversity occurrence data published by The Museum of Vertebrate Zoology, Oklahoma Museum of Natural History, Museum of Southwestern Biology Albuquerque, Michigan State University Museum, Mammal Collection of the University of Kansas Biodiversity Research Centre, Senckenberg Collection Mammalia SMF, Colección de Mamíferos de la Sierra Volcánica Transversal de México (UAM-I), The Field Museum of Natural History, Yale Peabody Museum <data.gbif.org>.

Gonzalez J, Bowen M, Nichol S, Rico-Hesse R (1996) Genetic characterization and phylogeny of Sabia virus, an emergent pathogen in Brazil. *Virology*, **221**, 318–324.

Gonzalez J, Emonet S, de Lamballerie X, Charrel R (2007) Arenaviruses. *Current Topics in Microbiology and Immunology*, **315**, 253–288.

Harmon LJ, Glor RE (2010) Poor statistical performance of the mantel test in phylogenetic comparative analyses. *Evolution*, **64**, 2173–2178.

Ho SYW, Lanfear R, Bromham L *et al.* (2011) Time-dependent rates of molecular evolution. *Molecular Ecology*, **20**, 3087–3101.

Holmes E (2003) Molecular clocks and the puzzle of RNA virus origins. *Journal of Virology*, **77**, 3893–3897.

Holmes EC (2008) Evolutionary history and phylogeography of human viruses. *Annual Reviews in Microbiology*, **62**, 307–328.

Holmes JC, Price PW (1980) Parasite communities: the roles of phylogeny and ecology. *Systematic Zoology*, **29**, 203–213.

Horie M, Honda T, Suzuki Y *et al.* (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature*, **463**, 84–87.

Hugot JP (2003) L'évolution des Arenaviridae et de leurs hôtes muridés résulte-t-elle d'évènements de capture ou de processus de coévolution? Evolution of the Old World Arenaviridae and their rodent hosts: generalized host-transfer or association by descent? *Bulletin de l'Académie Vétérinaire de France* **156**, 37–44.

Hugot JP, Gonzalez JP, Denys C (2001) Evolution of the Old World Arenaviridae and their rodent hosts: generalized host-transfer or association by descent? *Infection Genetics Evolution*, **1**, 13–20.

IUCN (2010) IUCN Red List of Threatened Species. Version 2010. http://www.iucnredlist.org.

Jackson AP, Charleston MA (2004) A cophylogenetic perspective of RNA-virus evolution. *Molecular Biology and Evolution*, **21**, 45–57.

Jansa SA, Weksler M (2004) Phylogeny of muroid rodents: relationships within and among major lineages as determined by IRBP gene sequences. *Molecular Phylogenetics and Evolution*, **31**, 256–276.

Jay M, Glaser C, Fulhorst C (2005) The arenaviruses. *Journal of the American Veterinary Medical Association*, **227**, 904–915.

Katzourakis A (2010) PLoS genetics: endogenous viral elements in animal genomes. *PLoS Genetics*, **6**, e1001191.

Kilgore P, Peters C, Mills J *et al.* (1995) Prospects for the control of Bolivian hemorrhagic fever. *Emerging infectious diseases*, **1**, 97–99.

Kotturi MF, Botten J, Sidney J *et al.* (2009) A multivalent and cross-protective vaccine strategy against arenaviruses associated with human disease. *PLoS Pathogens*, **5**, e1000695.

Kunz S (2009) Receptor binding and cell entry of Old World arenaviruses reveal novel aspects of virus–host interaction. *Virology*, **387**, 245–249.

Laakkonen J, Kallio-Kokko H, Oktem MA *et al.* (2006) Serological survey for viral pathogens in Turkish rodents. *Journal of wildlife diseases*, **42**, 672–676.

Lapointe F-J, Garland T (2001) A generalized permutation model for the analysis of cross-species data. *Journal of Classification*, **18**, 109–127.

Larson R, Dai D, Hosack V, Tan Y, Bolken T (2008) Identification of a broad-spectrum arenavirus entry inhibitor. *Journal of Virology*, **82**, 10768–10775.

Legendre P (2000) Comparison of permutation methods for the partial correlation and partial Mantel tests. *Journal of Statistical Computation and Simulation*, **67**, 37–73.

Legendre P, Desdevises Y, Bazin E (2002) A statistical test for host-parasite coevolution. *Systematic Biologists*, **51**, 217–234.

Lukashevich IS, Djavani M, Rodas JD *et al*. (2002) Hemorrhagic fever occurs after intravenous, but not after intragastric, inoculation of rhesus macaques with lymphocytic choriomeningitis virus. *Journal of Medical Virology*, **67**, 171–186.

McCormick JB, Webb P, Krebs J, Johnson K, Smith ES (1987) A prospective study of the epidemiology and ecology of Lassa fever. *The Journal of Infectious Diseases*, **155**, 437–444.

Meier-Kolthoff JP, Auch A, Huson DH, Goeker M (2007) CopyCat: cophylogenetic analysis tool. *Bioinformatics*, **23**, 898–900.

Mills JN, Ellis BA, Childs JE *et al*. (1994) HPrevalence of infection with Junin virus in rodent populations in the epidemic area of Argentine hemorrhagic fever. The *American Journal of Tropical Medicine and Hygiene*, **51**, 554–562.

Mills JN (1999) The role of rodents in emerging human disease: examples from the hantaviruses and arenaviruses. *Ecologically-based Rodent Management*, 134–160.

Mills J, Bowen M, Nichol S (1997) African arenaviruses-coevolution between virus and murid host? *Belgian Journal of Zoology (Belgium)*, **127**, 19–28.

Miranda G, Andrades-Miranda J, Oliveira L, Langguth A, Mattevi M (2007) Geographic patterns of genetic variation and conservation consequences in three South American rodents. *Biochemical Genetics*, **45**, 839–856.

Ogbu O, Ajuluchukwu E, Uneke CJ (2007) Lassa fever in West African sub-region: an overview. *Journal of vector borne diseases*, **44**, 1–11.

Pagán I, Firth C, Holmes EC (2010) Phylogenetic analysis reveals rapid evolutionary dynamics in the plant RNA virus genus tobamovirus. *Journal of Molecular Evolution*, **71**, 298–307.

Palacios G, Druce J, Du L *et al*. (2008) A new arenavirus in a cluster of fatal transplant-associated diseases. *New England Journal of Medicine*, **358**, 991–998.

Paterson AM, Banks J (2001) Analytical approaches to measuring cospeciation of host and parasites: through a glass, darkly. *International Journal for Parasitology*, **31**, 1012–1022.

Peters CJ, Buchmeier M, Rollin PE, Ksiazek TG (1996) Arenaviruses. In: *Fields Virology* (eds Fields BN, Knipe DM, Howley PM, Chanock RM, Melnick JL, Monath TP, Roizman R and Straus SE), pp. 1521–1552. Lippincott-Raven Publishers, Philadelphia.

Posada D, Crandall K (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Poulin R, Krasnov BR, Mouillot D (2011) Host specificity in phylogenetic and geographic space. *Trends in Parasitology*, **27**, 355–361.

R Development Core Team, 2.12 (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radoshitzky S, Abraham J, Spiropoulou C *et al*. (2007) Transferrin receptor 1 is a cellular receptor for New World haemorrhagic fever arenaviruses. *Nature*, **446**, 92–96.

Radoshitzky S, Kuhn J, Spiropoulou C *et al*. (2008) Receptor determinants of zoonotic transmission of New World hemorrhagic fever arenaviruses. *Proceedings of the National Academy of Sciences*, **105**, 2664.

Ramsden C, Holmes EC, Charleston MA (2009) Hantavirus evolution in relation to Its rodent and insectivore hosts: no evidence for codivergence. *Molecular Biology and Evolution*, **26**, 143–153.

Reignier T, Oldenburg J, Flanagan ML *et al*. (2008) Receptor use by the Whitewater Arroyo virus glycoprotein. *Virology*, **371**, 439–446.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Rousseau MC, Saron MF, Brouqui P, Bourgeade A (1997) Lymphocytic choriomeningitis virus in southern France: four case reports and a review of the literature. *European Journal of Epidemiology*, **13**, 817–823.

Salazar-Bravo J, Ruedas LA, Yates TL (2002) Mammalian reservoirs of arenaviruses. *Current Topics in Microbiology and Immunology*, **262**, 25–63.

Skinner HH, Knight EH (1979) The potential role of Syrian hamsters and other small animals as reservoirs of lymphocytic choriomeningitis virus. *The Journal of Small Animal Practice*, **20**, 145–161.

Skinner HH, Knight EH, Grove R (1977) Murine lymphocytic choriomeningitis: the history of a natural cross-infection from wild to laboratory mice. *Laboratory Animals*, **11**, 219–222.

Smith AL, Paturzo FX, Gardner EP *et al*. (1984) Two epizootics of lymphocytic choriomeningitis virus occurring in laboratory mice despite intensive monitoring programs. *Canadian Journal of Comparative Medicine*, **48**, 335–337.

Spiropoulou CF, Kunz S, Rollin PE, Campbell KP, Oldstone MBA (2002) New World arenavirus clade C, but not clade A and B viruses, utilizes α-dystroglycan as its major receptor. *Journal of Virology*, **76**, 5140–5146.

Stamatakis A, Ludwig T, Meier H (2004) New fast and accurate heuristics for inference of large phylogenetic trees. *Proceedings of 18th IEEE/ACM International Parallel and Distributed Processing Symposium Santa Fe*, NM, *April 26–30*.

Stamatakis A, Auch A, Meier-Kolthoff J, Göker M (2007) AxPcoords & parallel AxParafit: statistical co-phylogenetic analyses on thousands of taxa. *BMC Bioinformatics*, **8**, 405.

Weaver S, Salas R, de Manzione N *et al*. (2000) Guanarito virus (Arenaviridae) isolates from endemic and outlying localities in Venezuela: sequence comparisons among and within strains isolated from Venezuelan hemorrhagic fever patients and rodents. *Virology*, **266**, 189–195.

Webb PA, Johnson KM, Hibbs JB, Kuns ML (1970) Parana, a new Tacaribe complex virus from Paraguay. *Arch Gesamte Virusforsch*, **32**, 379–388.

Webb CO, Ackerly DD, , Kembel SW (2008) Phylocom: software for the analysis of phylogenetic community structure and character evolution. *Bioinformatics*, **24**, 2098–2100.

Worobey M, Telfer P, Souquière S *et al*. (2010) Island biogeography reveals the deep history of SIV. *Science*, **329**, 1487.

Xia X, Xie Z (2001) DAMBE: software package for data analysis in molecular biology and evolution. *Journal of Heredity*, **92**, 371–373.

N.I. is a molecular ecologist with interest in bat systematics, phylogeography and their diseases. M.B. is interested in using evolutionary techniques in areas of medical importance. O.M. is interested in studying biodiversity at all scales (local to global) and in statistical techniques that takes phylogenetic information into account. N.M. works on phylogenetics and evolutionary genetics with interest in infectious diseases in wildlife.

Data accessibility

Alignments for all analyses and the R script used for the Mantel tests have been made available: DRYAD entry doi:10.5061/dryad.231g746f.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Associations of NWA and their hosts. Junior synonyms previously used in the literature are indicated below the current taxonomic name.

Table S2 GenBank accession and strain numbers of NWA sequences used in this study.

Table S3 GenBank accession numbers and rodents species names with their synonyms.

Table S4 Country of occurrence for the American rodent species used in this study and their overlap with each NWA Clade.

Figure S1 Bayesian phylogenetic trees of New World arenaviruses based on complete nucleotide sequences of glycoprotein (A), nucleoprotein (B), L-protein (C) and Z-protein (D).

Figure S2 Reconciliation analysis of NWA based on nucleoprotein (A), L-protein (B), and Z-protein (C) with known hosts.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Paper 2.3.2

Pečnerová P., Moravec J. C., **Martínková N.** 2015. A skull might lie: Modeling ancestral ranges and diet from genes and shape of tree squirrels. *Systematic Biology* 64: 1074-1088.

Syst. Biol. 64(6):1074–1088, 2015

DOI:10.1093/sysbio/syv054
Advance Access publication August 8, 2015

A Skull Might Lie: Modeling Ancestral Ranges and Diet from Genes and Shape of Tree Squirrels

PATRÍCIA PEČNEROVÁ[1,2,3,4,*], JIŘÍ C. MORAVEC[2,5], AND NATÁLIA MARTÍNKOVÁ[2,6]

[1]*Department of Botany and Zoology, Masaryk University, Kotlářská 2, 602 00 Brno, Czech Republic;* [2]*Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, v.v.i., Květná 8, 603 65 Brno, Czech Republic;* [3]*Department of Zoology, Stockholm University, 10691 Stockholm, Sweden;* [4]*Department of Bioinformatics and Genetics, Swedish Museum of Natural History, PO Box 50007, 10405 Stockholm, Sweden;* [5]*Institute of Fundamental Sciences, Massey University, Private Bag 11 222, Palmerston North, New Zealand; and* [6]*Institute of Biostatistics and Analyses, Masaryk University, Kamenice 3, 625 00 Brno, Czech Republic*

[*]*Correspondence to be sent to: Department of Bioinformatics and Genetics, Swedish Museum of Natural History, PO Box 50007, 10405 Stockholm, Sweden; E-mail: patricia.pecnerova@nrm.se*

Received 8 January 2015; reviews returned 23 March 2015; accepted 29 July 2015
Associate Editor: Norm McLeod

Abstract.—Tropical forests of Central and South America represent hotspots of biological diversity. Tree squirrels of the tribe Sciurini are an excellent model system for the study of tropical biodiversity as these squirrels disperse exceptional distances, and after colonizing the tropics of the Central and South America, they have diversified rapidly. Here, we compare signals from DNA sequences with morphological signals using pictures of skulls and computational simulations. Phylogenetic analyses reveal step-wise geographic divergence across the Northern Hemisphere. In Central and South America, tree squirrels form two separate clades, which split from a common ancestor. Simulations of ancestral distributions show western Amazonia as the epicenter of speciation in South America. This finding suggests that wet tropical forests on the foothills of Andes possibly served as refugia of squirrel diversification during Pleistocene climatic oscillations. Comparison of phylogeny and morphology reveals one major discrepancy: *Microsciurus* species are a single clade morphologically but are polyphyletic genetically. Modeling of morphology–diet relationships shows that the only group of species with a direct link between skull shape and diet are the bark-gleaning insectivorous species of *Microsciurus*. This finding suggests that the current designation of *Microsciurus* as a genus is based on convergent ecologically driven changes in morphology. [Ancestral range reconstruction; diet modeling; geometric morphometry; multilocus phylogeny; Sciurini; speciation.]

Speciation in the tropics is a key area of inquiry in the study of biological diversity and the latitudinal gradient in species richness, one of the most intensively discussed patterns in biology (Pianka 1966; Mittelbach et al. 2007). While there is no general explanation for tropical species richness (Mittelbach et al. 2007), studies of diverse model organisms show that an array of mechanisms and processes are responsible for high biodiversity (Haffer 1997; Schluter 2001; Via 2001). In the Neotropics, biodiversity is mostly explained by a combined effect of Pliocene and Miocene orogenic processes and Pleistocene climatic oscillations (Hooghiemstra and van der Hammen 1998; Rull 2008; Turchetto-Zolet et al. 2013). The different effects of orogenic processes and climatic changes have created a mosaic of phylogeographic patterns that can be detected in Neotropical forests today (Turchetto-Zolet et al. 2013). Tree squirrels of the tribe Sciurini represent an excellent model system for studying these processes. Sciurini are a widespread and diverse group with a broad food and habitat niche. These squirrels inhabit much of Eurasia and North America (Thorington et al. 2012), but they are more diverse in the recently colonized tropical forests of Central and South America.

According to molecular dating based on multilocus data (Mercer and Roth 2003), the tribe Sciurini originated in the Northern Hemisphere approximately 23 Ma, when the Sciurini last shared a common ancestor with the Pteromyini. Within the tribe Sciurini, recent phylogenetic analyses show basal divergence between the genus *Tamiasciurus*, which inhabits North America, and the *Sciurus* lineage (including *Microsciurus*, *Rheithrosciurus*, and *Syntheosciurus*). The genus *Sciurus* originated in Eurasia and colonized the Americas after the formation of a land bridge in Beringia (Oshida et al. 2009; Pečnerová and Martínková 2012). This scenario is supported by molecular dating (Mercer and Roth 2003) as well as by independent evidence gathered from the paleobotanical record (Wolfe et al. 1966; Wing 1998), which shows that the Beringian land corridor was forested when the transition occurred between 8.6 and 5.3 Ma (Mercer and Roth 2003). In the Americas, divergence of *Sciurus* species proceeded from north to south, until *Sciurus* colonized the tropical forests of Central America (which is considered to include Mexico throughout the study). After the formation of the Isthmus of Panama, squirrels entered South America approximately 2.8 Ma (Mercer and Roth 2003). In the tropical regions of Central and South America, *Sciurus* squirrels reached high levels of species richness and species of the genera *Microsciurus* and *Syntheosciurus* diverged from *Sciurus* here (Mercer and Roth 2003; Pečnerová and Martínková 2012). The genus *Rheithrosciurus*, which is endemic to Borneo, probably diverged from the genus *Sciurus* early after the origin of *Sciurus* in Eurasia (Pečnerová and Martínková 2012).

Species richness in the tropical forests of Central and South America is also reflected in the Sciurini tree squirrels. Out of the 37 species of the tribe, 22 inhabit tropical regions of Central and South America (Wilson and Reeder 2005). The high Neotropical biodiversity in the phylogenetically youngest group points to rapid

speciation (Roth and Mercer 2008), but very little is known about these species and the driver of their intensive diversification remains unclear.

The taxonomy of Sciurini is based on two morphological studies from the mid-20th century, a qualitative analysis of cranial morphology by Moore (1959) and a paleontological revision with emphasis on osteological and cranial measurements by Black (1963). The tribe Sciurini currently consists of five genera: *Microsciurus*, *Rheithrosciurus*, *Sciurus*, *Syntheosciurus*, and *Tamiasciurus*. However, molecular studies in the last decade show that species of *Microsciurus* and *Syntheosciurus* and possibly also *Rheithrosciurus* are closely related to species of the genus *Sciurus* and cluster into a single lineage (Mercer and Roth 2003; Herron et al. 2004; Steppan et al. 2004). No explanation of the discrepancy between molecular and morphological data has yet been provided.

In this study, we hypothesize that Sciurini differentiated while dispersing to new regions. We address this hypothesis by extending the genetic record of Sciurini and by modeling the possible areas of origin. We also examine the conflicting signals of molecular and morphological data. We consider two hypothetical causes of morphological differentiation and their inconsistency with the molecular phylogeny: allopatric differentiation and ecological adaptation. We test these hypotheses using a combined molecular-morphological approach, accompanied by simulations in a Bayesian framework and predictive modeling.

Materials and Methods

Tissue Samples

The Natural History Museum (NHM) in London provided 34 samples of soft tissue residues retrieved from squirrel bones and the National Museum of Natural History of the Smithsonian Institution in Washington, DC, provided 13 samples of frozen soft tissues (Supplementary Table S1, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). Samples represented 36 species, out of which 34 were species of the tribe Sciurini and two species were used as an outgroup (*Glaucomys volans* and *Pteromys volans*).

DNA Extraction and Amplification

We used DNeasy Blood and Tissue Kit by QIAGEN (Hilden, Germany) to extract total genomic DNA. Due to the characteristic features of museum specimens, including chemical treatment, risk of contamination and DNA damage affecting molecules after the death of an organism (Pääbo et al. 2004; Martínková and Searle 2006), we used a conservative approach to obtain sequences from this material. We accepted results from relatively long PCR products, indicating that the DNA was preserved in sufficient quality and quantity (c.f. Martínková and Searle 2006).

We amplified three mitochondrial loci (12S rRNA, 16S rRNA, and *MT-CYB*) and two nuclear loci (*MYC* and *IRBP*). Sequences of *MT-CYB* for four samples (*Sciurus deppei*, *Sciurus flammifer*, *Sciurus igniventris*, and *Sciurus pyrrhinus*) were obtained from overlapping fragments with newly designed internal primers (Supplementary Table S2, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). Amplifications for sequencing were performed in 50 μL volumes and the reaction mixture contained 2 μL of DNA, 2 mM $MgCl_2$, 0.2 mM dNTPs, 1.0 unit of Platinum *Taq* DNA Polymerase (Invitrogen, Life Technologies, Carlsbad, CA, USA), 1 × PCR buffer, 0.2 μM of each primer, and ddH_2O up to the total volume. Amplification conditions were adjusted for each primer pair (Table 1). The common steps for all amplifications were initial denaturation at 94°C for 3 min, followed by 35–40 cycles of 94°C for 1 min, 49–60°C for 30 s to 1 min and 72°C for 45 s to 3 min, with the final extension at 72°C for 3 min (Table 1). We were able to extract and amplify DNA from 35 samples of 28 species. The amplified products were purified and sequenced by Macrogen Europe (Amsterdam, the Netherlands) with both amplification primers. The sequence data set was enriched with tree squirrel sequences from GenBank. The complete data set for molecular analyses then contained 30 species (Supplementary Table S3, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). We used CodonCode Aligner (CodonCode Corporation, Centerville, MA, USA) to process chromatograms into consensus sequences.

To verify that the generated sequences represent endogenous DNA of Sciurini squirrels, all sequences were checked using BLAST (Altschul et al. 1990). The sequences were the most similar to respective squirrel species expected to be the most closely related, except for three samples (12S rRNA and 16S rRNA sequences of *Sciurus ignitus*, *Sciurus richmondi*, and 12S rRNA sequence of *Sciurus stramineus*) that matched human DNA and these were excluded from further analyses.

Phylogenetic Analyses of the Supermatrix

The DNA sequences were aligned in ClustalW2 (Larkin et al. 2007; Goujon et al. 2010) and the rRNA loci sequences were aligned according to their predicted secondary structure in MAFFT 7.2 with the q-ins-i algorithm (Katoh and Standley 2013). The alignments were manually edited in BioEdit 7.1.3 (Hall 1999).

We used two concepts of phylogenetic inference, supermatrix and supertree. The supermatrix approach requires sequences of all loci and taxa to be concatenated into a single matrix. The concatenated data set was constructed in SequenceMatrix 1.7.8 (Vaidya et al. 2011) and analyzed by maximum-likelihood (ML) method in RAxML 8.1.15 (Stamatakis 2014) and by Bayesian inference (BI) in MrBayes 3.2.1 (Huelsenbeck and Ronquist 2001). Each analysis used a partitioning scheme selected according to the Bayesian Information Criterion in PartitionFinder 1.0 (Lanfear et al. 2012) with linked

TABLE 1. Characteristics of sequence alignments analyzed in this study

Locus	PCR amplification	Alignment		Partitioned substitution model			
		Number of sequences	Length	Gene trees	BI from supermatrix	ML from supermatrix	Species tree
12S rRNA (mt)	$T_a=60$, $t_e=3$	25	867	GTR+Γ	GTR+Γ[1]	GTR+Γ[1]	GTR+Γ[1]
16S rRNA (mt)	$T_a=49$, $t_e=1.5$	18	1077	GTR+Γ	GTR+Γ[1]	GTR+Γ[1]	GTR+Γ[1]
MT-CYB (mt)	$T_a=55$, $t_e=2$	22	1140	GTR+Γ, HKY+I, GTR+Γ	GTR+Γ[1], HKY+I[8], GTR+Γ[9]	GTR+Γ[1], GTR+Γ[4], GTR+Γ[5]	GTR+Γ[2]
D-loop (mt)	n/a	8	1228	GTR+Γ	GTR+Γ[10]	GTR+Γ[6]	GTR+Γ[3]
MYC exon 2 (nuc)	n/a	5	674	JC, HKY (3rd)	JC[6], JC[6], HKY[7]	GTR+Γ[4], GTR+Γ[4], GTR+Γ[2]	HKY[4]
MYC exon 3 (nuc)	$T_a=60$, $t_e=0.75$	14	967	K80, K80 (3rd)	HKY+I[3], HKY+I[4], K80[5]	GTR+Γ[3], GTR+Γ[4], GTR+Γ[3]	GTR+I[5]
IRBP (nuc)	$T_a=58$, $t_e=3$	19	1253	HKY+Γ, HKY, HKY+I	GTR+Γ[2], HKY+I[3], HKY+I[4]	GTR+Γ[2], GTR+Γ[3], GTR+Γ[4]	GTR+I[5]
RAG1 (nuc)	n/a	6	2141	HKY, HKY (3rd)	HKY+I[3], HKY+I[4], GTR+Γ[2]	GTR+Γ[3], GTR+Γ[4], GTR+Γ[2]	GTR+I[5]
Total		117	9347				

Notes: mt, mitochondrial; nuc, nuclear; T_a annealing temperature in °C; t_e, extension time in minutes; n/a, not available; [1–10] partitions with jointly estimated model parameters in the given analysis, model order corresponds to codon positions unless specified as the third codon position.

branches and a greedy algorithm that considered non-coding genes and codon positions of the coding genes. The tested models were: JC, K80, HKY, GTR for individual genes; JC, HKY, GTR for BI of the concatenated data set and predefined RAxML model set for ML tree (see Table 1). Models for individual genes and the concatenated data set could include proportion of invariable sites (*I*) or rate heterogeneity modeled with a Γ distribution model (Γ).

In the ML method implemented in RAxML 8.1.15 (Stamatakis 2014), the concatenated data set was analyzed with six partitions (Table 1). The analysis was conducted with the rapid bootstrapping algorithm, which combines ML search and bootstrapping, and was run with 1000 bootstrap searches. In the BI of the concatenated data set conducted in MrBayes 3.2.1 (Huelsenbeck and Ronquist 2001), a substitution model with 10 partitions was applied (Table 1). Convergence and chain mixing were facilitated by running five Markov Chain Monte Carlo (MCMC) chains for 3 million generations, with trees sampled every 1000th generation. One chain swap was attempted every third generation and the temperature was adjusted to 0.08 after the initial run when the default temperature (0.1) did not converge. Successful neighboring chain swaps between 30% and 70% showed that chain mixing was effective with this setting. A burn-in fraction of 30% was used for all analyses. We verified that the runs converged and reached stationary distribution by observing that after 3 million generations the average standard deviation of split frequencies reached a value below 0.01 (0.0087) and all potential scale reduction factors approached 1.0 as the run converged.

Phylogenetic Analyses of Supertrees

BI of the gene trees was conducted in MrBayes 3.2.1 (Huelsenbeck and Ronquist 2001) with appropriate partitioning schemes and substitution models estimated for each locus separately (Table 1). The BI was conducted with four (for RNA loci) or five (for the other loci) MCMC chains run for 2 million generations, with trees sampled every 1000th generation, a chain swap attempted every third generation of the run and temperature set to 0.1. MCMC convergence was assessed as described above.

In this study, gene trees were analyzed by two methods of supertree reconstruction, SuperTriplets (Ranwez et al. 2010) and a veto method implemented in PhySIC_IST (Scornavacca et al. 2008). The SuperTriplets method uses a polynomial algorithm to search for a median tree (a tree minimizing the sum of distances to the source trees) based on triplets (rooted, binary, three-leaf topologies) (Ranwez et al. 2010). The SuperTriplets analysis was executed with the default settings through the online version of the program available at http://www.supertriplets.univ- montp2.fr/, (last accessed 14 August, 2015). Veto methods show topologies supported by all source (gene) trees and eliminate taxa that produce conflicting topologies. Hence, these methods are more conservative and usually result in less resolved supertrees. We performed a veto analysis in the PhySIC_IST program available online at http://www.atgc-montpellier.fr/physic_ist/, (last accessed 14 August, 2015), with the bootstrap threshold for source clade selection set to default value (0) and the correction threshold used by source tree correction (preprocessing the data to detect anomalies) set to 0.3. Four taxa were omitted from the analysis due to conflicting topologies.

Species Trees Reconstruction

We used two methods that model the coalescent process to infer species trees. SVDquartets (Chifman and Kubatko 2014) uses a score based on decomposition

of a matrix to assess a supported topology for taxa quartets. The matrix consists of frequencies of splits that characterize quartets of investigated taxa based on available sequence data. We used the method embedded in PAUP 4.0a142 (Swofford 2015), utilizing exhaustive quartet sampling (27,405 quartets) and 100 bootstrap replicates for the multispecies coalescent.

BI of species trees in *BEAST 2.2 (Heled and Drummond 2010) jointly estimates gene trees, where the gene trees must be embedded in the species tree, and divergence times and population sizes of present and ancestral taxa using MCMC sampling from the posterior. Substitution model partitioning scheme was estimated without distinguishing codon positions to retain gene partitioning in the species tree reconstruction (Table 1). MCMC was run for 10 million generations sampled every 1000th step with a birth–death speciation tree prior and strict clock. Convergence was checked in Tracer 1.6 (Rambaut et al. 2013), where trace of model likelihood and parameter values oscillated around a plateau value, their posterior densities showed roughly unimodal distribution and estimated sample sizes (ESSs) were >100.

Analysis of Phylogenetic Stability

The impact of missing data was evaluated in Phyloterraces (Sanderson et al. 2011). The algorithm breaks the supertree into subtrees representing taxa sets as they appear in the individual gene trees. The subtrees are further subdivided into triplets that appear in the final, binary tree. The triplets are used to construct all possible parent trees. Given the amount of missing data and its location in the alignment, the parent trees will have identical tree likelihood. For the purpose of this analysis, the BI tree was summarized with resolved all compatible relationships.

Testing Microsciurus Monophyly

Support for *Microsciurus* monophyly was estimated implicitly from the posterior tree sample and explicitly by comparing likelihoods of trees sampled from constrained and negatively constrained analyses.

Following Bergsten et al. (2013) and Suchard et al. (2005), support for a certain topology could be estimated from the posterior tree sample. MCMC samples tree space and the posterior probability of a tree is computed as the frequency of posterior trees conforming to a hypothesis. We explored our posterior tree samples from the BI based on the supermatrix and Bayesian coalescent inference of species trees.

We tested *Microsciurus* monophyly by comparing model likelihoods from two BI analyses: a constraint analysis in which monophyly of *Microsciurus alfari* and *Microsciurus flaviventer* was enforced, and a negative constraint in which monophyly of this pair was disallowed. Analytical parameters were as described above. Model likelihoods were estimated with the harmonic mean estimator and compared with Bayes factors (BFs).

Molecular Dating

Divergence dates were inferred with Bayesian coalescent analysis (Bouckaert et al. 2014) using multiple time constraint priors on internal nodes. The root of Palearctic Sciurini prior was constrained with a uniform distribution between 3 and 36 Ma, assuming that *Sciurus* was present in Europe since late Pliocene, and the earliest known squirrel fossil, *Douglassciurus jeffersoni*, was dated to late Eocene (Thorington et al. 2012). The time constraints on root of Nearctic *Sciurus* between 7.4 and 36 Ma were based on a combination of our results with the fossil record. Our data show that *Sciurus* colonized North America from Asia, but the oldest fossil attributable to *Sciurus* dates back to late Miocene in Nevada (Emry et al. 2005). The colonization therefore must have occurred before the earliest estimated opening of the Bering Strait 7.4 Ma (Marincovich and Gladenkov 1999). Expecting that Sciurini arrived to South America via the Isthmus of Panama, their root prior was set to log-normal distribution with mean of tested prior parameter values equal to 3 Ma (mean $= e^{\mu+\sigma^2/2}$, where $\mu = 1.0186$, $\sigma = 0.4$) (Keigwin 1982). Two time constraints within *Tamiasciurus* are based on their first known fossils. *Tamiasciurus hudsonicus* was first reported from the Irvingtonian (Steele 1998), implemented as exponential prior with mean = 0.5, offset by 0.24 Ma, for the *Tamiasciurus* root. The split between *Tamiasciurus douglasii* and *Tamiasciurus mearnsi* is believed to be complete by the end of the last glaciation (Steele 1999) and was modeled with exponential prior with mean = 0.5, offset by 0.015 Ma. Tree and clock models were tested with BF, comparing marginal likelihoods from analyses with strict clock and log-normal relaxed clock and with birth–death and Yule speciation tree priors. The MCMC algorithm ran for 100 million generations and was sampled every 1000th generation. The substitution model partitioning scheme was identical to that used in the BI from the supermatrix. Convergence was assessed as described above.

Morphological Analyses of Skulls

Morphological data were based on digital photographs of skulls of 525 specimens that represented 25 species, out of which 24 were species of the tribe Sciurini and one species was used as an outgroup (Supplementary Table S5, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). All species were represented by 10–27 specimens, and samples of both sexes and various localities within the species range were included (Supplementary Table S6, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). Ventral and dorsal photographs of the skulls were captured with a digital camera Nikon D90. Two different tripods were used to secure the camera due to data

collection in different museums. As a result, specimens from the NHM in London were photographed from a larger distance. To overcome the influence of different distance, a ruler was used in all photographs and size was recalibrated during scoring landmarks in tpsDig 2.16 (Rohlf 2005). The use of plasticine was not allowed; therefore, for capturing the ventral side, a skull was horizontally aligned by supporting the frontal part of the skull with a ruler, so that the occlusal surface was parallel to the underlying base. The alignment was assessed by eye. For capturing the dorsal side, skulls were aligned by positioning on incisors and auditory bullae. Some specimens were missing incisors (~2%) and there was a slight variability in the preservation of incisors, which might have subtly influenced the angle under which a skull was captured. However, including at least 10 specimens per species should minimize the differences. We also repeated all analyses including only species with at least 20 specimens to verify that the sample size was sufficient and the results were consistent.

Skull shape was captured with two-dimensional landmarks in tpsDig 2.16 (Rohlf 2005). 2D methodology is routinely applied to the study of rodent cranial morphology (Fadda and Corti 2000; Cardini and O'Higgins 2004; Cardini et al. 2005; Macholán et al. 2008) because of its "cost-effectiveness, rapidity of data collection and analytical simplicity" (Cardini 2014). Cardini (2014) studied the influence of missing the third dimension in 2D analyses and concluded that the congruence of 2D and 3D data is modest, but there is a level of inaccuracy when 2D landmarks are applied to a 3D object and these studies require cautious interpretations. For that reason, we refrained from making conclusions regarding species-level differences in morphology and we concentrated on morphological distinction between *Microsciurus*, *Sciurus*, and *Tamiasciurus*. Twelve landmarks were digitized on the dorsal side and 21 on the ventral side (Fig. 1, Table 2). Landmarks were digitized on the left part of the skull to avoid the influence of asymmetry (Cardini and O'Higgins 2004; Macholán et al. 2008). Information on size, location, and orientation of specimens were removed by generalized least-squares Procrustes superimposition (Gower 1975; Rohlf and Slice 1990) in tpsRelw 1.49 (Rohlf 2010). The shape space in geometric morphometry (Kendall's shape space) is a curved surface and has a non-Euclidean geometry. Approximation by the linear tangent space is used (Rohlf 1998), but validity of such an approximation needs to be tested by estimating correlation between the Procrustes distances of Kendall's shape space and Euclidean distances of the linear tangent space. This approximation was evaluated in tpsSmall 1.20 (Rohlf 2003).

Shape change was visualized by the thin-plate spline method (Bookstein 1989), which is based on the deformation of a regular grid by bending energy, as an analogy of energy required for shape change. Bending energy is expressed by variables called principal warps. These indicate the directions of shape change. Other variables, called partial warps, are used to illustrate the proportion of shape change each principal warp accounts for (Bookstein 1989).

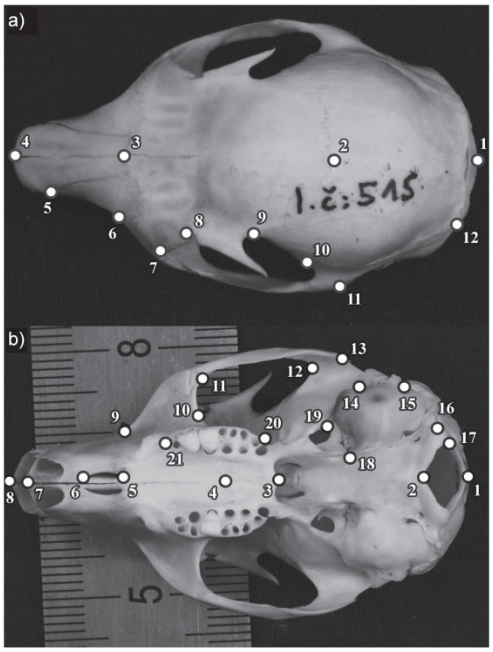


FIGURE 1. Positions of landmarks digitized on photographs of squirrel skulls on their dorsal (a) and ventral (b) side. The skull of *S. vulgaris* was borrowed from the collections of the Institute of Vertebrate Biology of the Academy of Sciences of the Czech Republic.

To reduce the dimensionality of the data, we used relative warp analysis (RWA; Rohlf 1993), which is a version of principal component analysis (Pearson 1901; Hotelling 1933) implemented in tpsRelw. Landmark coordinates were used to perform linear discriminant function analysis (DFA) to determine variables that can be used to divide samples into groups. Multivariate analysis of variance (MANOVA) was performed on Procrustes distances to test the statistical significance of differences between groups, applying Goodall's *F*-test with 999 permutations. Unweighted pair-group method using arithmetic averages (UPGMA) was used as the clustering method to construct trees based on Mahalanobis distances between predicted DFA species averages. MANOVA, DFA, and UPGMA were performed in R (Paradis et al. 2004; Adams and Otarola-Castillo 2013; McFerrin 2013; R Core Team 2013).

The composition of species differed between molecular and morphological analyses, with 19 species being examined with both approaches (Supplementary Table S7, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). The results

TABLE 2. Description of landmark positions on dorsal and ventral sides of squirrel skulls

Side	Number	Description
Dorsal	1	Posterior end of the curvature of the occipital
	2	Intersection of the coronal and sagittal sutures
	3	Intersection of the naso-frontal suture in the midline
	4	Tip of the nasals at their anterior suture
	5	Anterior tip of suture between nasal and premaxilla
	6	Anterior tip of the suture between premaxilla and maxilla
	7	Lateral end of the naso-frontal suture in the dorsal projection
	8	Point on frontal at greatest interorbital constriction
	9	Posterior base of the postorbital process
	10	Rostral end of the zygomatic process of the temporal bone
	11	Most posterior meeting point between jugal and squamosal process of the zygomatic arch
	12	Posterio-lateral end of the interparietal–occipital suture
Ventral	1	Posterior limit of the foramen magnum
	2	Anterior limit of the foramen magnum
	3	Posterior midline suture of palatines
	4	Intersection of the maxillo-palatine suture in the midline
	5	Posterior end of the incisive foramen
	6	Anterior end of the incisive foramen
	7	Rostralmost point of the upper incisor in the midline
	8	Tip of the nasals at their anterior suture
	9	Rostral end of the zygomatic plate
	10	Most anterior point of the orbit (in the ventral view)
	11	Maximal curvature on internal zygomatic bar
	12	Inner extreme curvature point of the zygomatic bar at the squaniosus process
	13	Posterior tip of the zygomatic arch
	14–15	Anterior and posterior tips of the external auditory meatus
	16–17	Lateral tips of the occipital condyle
	18	Lateral end of the basisphenoid–basioccipital suture in the dorsal projection
	19	Posterior extremity of the foramen ovale
	20	Anterior extremity of the toothrow
	21	Posterior extremity of the toothrow

Notes: Landmarks were chosen on the basis of similar morphometric studies on rodents (Fadda and Corti 2000; Cardini and O'Higgins 2004; Cardini et al. 2005; Macholán et al. 2008).

from the molecular genetic data are classified with Roman numerals while the results from the morphological data are classified with Arabic numbers.

Spatial and Ecological Hypotheses Testing

We evaluated Pearson's correlation between genetic and geographic distances with Mantel's test in R (Oksanen et al. 2013; R Core Team 2013), constructing the genetic matrix from phylogenetic distances computed from the chronogram and the geographic matrix from range centroids in the International Union for Conservation of Nature (IUCN) map data set (IUCN 2013) using the Geographic Distance Matrix Generator software (Ersts 2014). To evaluate statistical significance, we used 9999 permutations.

Probability of ancestral states of species ranges was calculated with respect to the topological uncertainty of the phylogeny in SIMMAP 1.5 (Bollback 2006). Species were scored with multistate characters according to their occurrence in Eurasia, North, Central and South America, and ancestral ranges were evaluated from 100 trees randomly sampled from the posterior.

To test for the ancestral area of the South American clade, we discretized range polygons from the IUCN data set for South American species ranges by creating a grid over South America with a cell size equal to three degrees longitude by three degrees latitude. Grid cells were coded as presence/absence data based on overlapping species range polygons. Ancestral area simulations in BayArea 1.0.2 (Landis et al. 2013) were run for 600 million MCMC generations on a pruned chronogram that included only the South American species with truncated geographic distances setting and with 20% burnin. Ancestral ranges were visualized in BayArea-fig (Landis et al. 2013).

To evaluate the relationship between dietary preferences and skull shape, we used a generalized linear model with a probit linkage function with relative warps (RWs) as predictors and binarized dietary preference (Supplementary Table S8, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n) as the response variable. The dietary preferences of Sciurini were estimated on the basis of Thorington et al. (2012). A species' diet was coded as follows: 0—food item not mentioned/not consumed, 1—food item consumed, 2—food item highlighted as a common or preferred diet. We used three sets of models, with dorsal, ventral, and combined RW values. Each binarized food preference was estimated separately and backward stepwise regression using AIC was performed to simplify the

model. Using these models, response variables (dietary components) were predicted and compared with the original values.

To further investigate the relationships between skull shapes and dietary preferences, we constructed a neighbor-joining (NJ) tree from Manhattan distances based on the dietary preference table (Supplementary Table S8, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). The diets of tree squirrels are generally variable, but due to the usage of a matrix of all consumed items, clusters in the NJ tree avoid the problem of a misleading classification into predefined dietary categories. *Sciurus richmondi* was excluded from all analyses of the relationship between diet and morphology as no information about its diet was available.

Results

We obtained 42 sequences, including nine previously unsampled species. With sequences from GenBank, our data set consisted of 117 sequences distributed between eight loci—12S rRNA, 16S rRNA, *MT-CYB*, D-loop, *MYC* exon 2, *MYC* exon 3, *IRBP*, and *RAG1*. Number of sequences per locus ranged between 5 and 25. The concatenated data set consisted of 9347 base pairs (bp). Gene alignments ranged from 674 to 2141 bp (Table 1). The missing data represented 64.27% of the data set.

Phylogenetic Relationships

The ML and BI trees based on the supermatrix yielded almost identical results (Fig. 2) and were consistent with the SuperTriplets and veto trees reconstructed from gene trees and species trees based on the coalescent (Supplementary Fig. S1, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). All trees recognized three lineages. First, there was a split between the genus *Tamiasciurus* (lineage I) and the other taxa. Afterwards, the latter lineage evolved in the direction of *Rheithrosciurus* (lineage II) and the remaining genera (*Microsciurus*, *Sciurus*, and *Syntheosciurus*; lineage III). The likelihood terrace in the tree space with the final tree contained three alternative topologies in total, meaning that the amount of missing data resulted in the inability of likelihood-based tree reconstruction to distinguish unequivocally between these topologies. The consensus tree from all topologies sharing a terrace showed that the position of *Sciurus anomalus* and *Rheithrosciurus macrotis* varied on the alternative tree topologies (Supplementary Fig. S4, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n).

In lineage III, no structure representative of the current taxonomy was observed; instead, species were grouped according to biogeography. After the split from *Rheithrosciurus*, all Palearctic species of *Sciurus* diversified gradually (*Sciurus lis* + *Sciurus vulgaris* and *S. anomalus*), followed by gradual speciation of Nearctic Sciurini (*Sciurus niger*, *Sciurus carolinensis* with undetermined relationship to *Sciurus aberti* + *Sciurus griseus*) and at last the divergence of the Neotropical clade (that included genera *Microsciurus* and *Syntheosciurus*) from the Nearctic ancestor.

Statistical support for the monophyletic origin of Neotropical species was marginally significant in the BI tree, but not significant in the ML tree (taking into account the significance level of 70% for bootstrap in ML and 0.95 for posterior probability in BI throughout

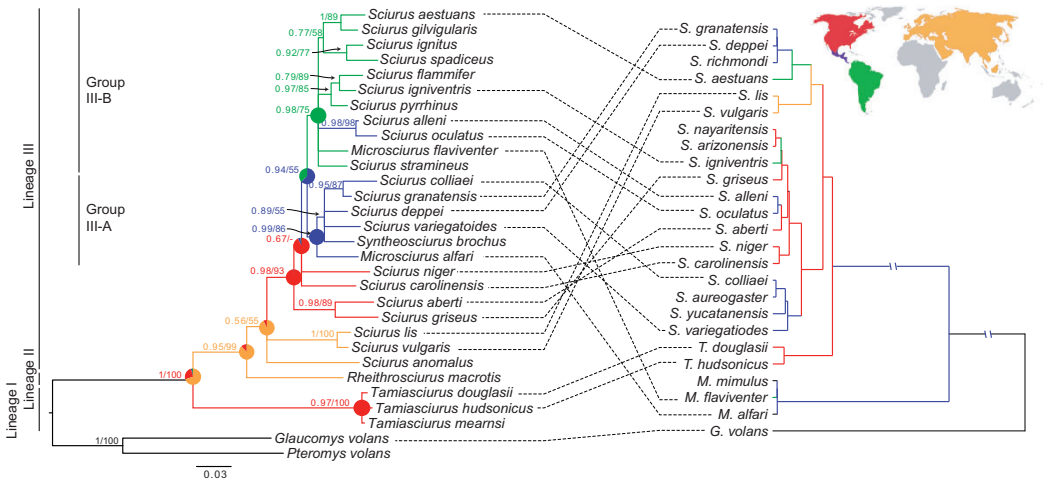


FIGURE 2. Phylogenetic tree of Sciurini tree squirrels calculated from supermatrix DNA sequence data using BI analysis (left) and UPGMA dendrogram of morphological shape (right). Branch labels indicate Bayesian posterior probabilities and bootstrap support from the ML analysis. Pie charts at basal nodes represent probability of ancestral distributions for the node, given the topological uncertainty of the phylogeny. Inset map shows respective continents of origin. Tanglegram depicts disparity between molecular phylogeny and morphological similarity.

the study). Two reciprocally monophyletic groups with significant support were recognized inside the Neotropical clade. The first group (III-A) consisted of Central American taxa with *M. alfari* as the sister branch to other taxa, and an unresolved relationship of *S. deppei*, *Sciurus variegatoides*, *Sciurus colliaei* + *Sciurus granatensis* and *Syntheosciurus brochus*. The second group (III-B) included South American taxa, but relationships of taxa after divergence from the common ancestor were unresolved, in agreement with a rapid diversification. However, three groupings were significantly supported in group III-B. First, *S. pyrrhinus* and *S. flammifer* + *S. igniventris*, second, *Sciurus aestuans* + *Sciurus gilvigularis* and third, *S. ignitus* + *Sciurus spadiceus*.

The SuperTriplets and veto trees constructed from gene trees were similar (Supplementary Fig. S1, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). They recognized the same three clades as retrieved from the supermatrix, but the mutual relationships were not clear. The veto method takes into account only uncontradicted relationships, therefore, several taxa with conflicting topologies in the gene trees were excluded from the analysis (*R. macrotis*, *S. aestuans*, *S. deppei*, and *S. ignitus*). The main difference compared with the supermatrix-based trees (ML and BI) was that the Central American taxa did not form a monophyletic group and almost all relationships inside the Neotropical clade were unresolved. Comparing the SuperTriplets tree to the ML and BI trees, *S. griseus* occurred in a different position than the previously suggested close relationship with *S. aberti* (Supplementary Fig. S1b, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). Two monophyletic groups inside the Neotropical clade were also recognized, but the relationships inside these groups slightly differed (Supplementary Fig. S1b, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n).

In species trees based on the coalescent, the basic tree structure was recognizable, but not supported (Supplementary Fig. S1c,d, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). Both species trees placed North American *S. griseus* within the South American group. The SVDquartets tree (Supplementary Fig. S1c, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n) additionally placed *R. macrotis* at the root of Sciurini.

Phylogenetic–biogeographic discongruency of Sciurus alleni and Sciurus oculatus

Two species (*S. alleni* and *S. oculatus*) inhabiting Central America were placed close to the Neotropical clade. We considered this placement as an analysis artifact, because these two species were represented only by sequences of the 12S rRNA locus that provided no information about the relationships of Neotropical squirrels (Supplementary Fig. S2, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n).

Nonmonophyly of Microsciurus

No trees from the posterior of either the supermatrix BI or species tree BI contained trees where *M. alfari* and *M. flaviventer* were sisters. The probability of this happening by chance alone, if prior distribution is considered, is 5.61×10^{-39} for the BI from a supermatrix posterior sample and 3.15×10^{-62} for the BI of species trees posterior sample. The constraint test of *Microsciurus* monophyly showed very strong support for nonmonophyly of the genus (2 log(BF) = −119.82).

Morphology of Skulls

The RWA showed that the first two RWs captured 62.5% of the variation in the data for the dorsal side and 47.6% for the ventral side of the tree squirrel skulls (Fig. 3; further plots in Supplementary Fig. S2, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). The correlation between Procrustes and Euclidean distances was high (both sides: $r = 1.000$, slope $b = 0.999$, intercept $a = 0.000$), which allowed us to proceed with subsequent analyses. DFA successfully assigned specimens into defined species with an average rate 72.9% (Goodall's $F = 47.083$, $P = 0.001$) for the dorsal side and 76.3% (Goodall's $F = 32.557$, $P = 0.001$) for the ventral side (Supplementary Table S9, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). The most successfully assigned specimens belonged to the outgroup *G. volans* with 100% for both sides, and the least successfully *S. granatensis* for the dorsal side (40.9%). Most misclassifications occurred among closely related species. For example, the genus *Microsciurus* had 100% success in assigning measured skulls into the appropriate genus, but the assignment into particular species of the genus was less successful (58.3–83.3%). Goodall's *F*-test showed that affiliation to a species explained 69.3% of differences in the dorsal shape of the skull ($F = 47.083$; df = 24,500; $P = 0.001$) and 61% of differences in the ventral shape of the skull ($F = 32.557$; df = 24,500; $P = 0.001$).

According to the RWA (Fig. 3) and the cluster analysis (Supplementary Fig. S3, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n), three morphological groups were recognized in the tribe Sciurini. The most distinct skull shape was observed in the genus *Microsciurus*, which formed a separate cluster against all other squirrels. The second group was represented by the genus *Tamiasciurus* and the third group included all *Sciurus* squirrels (Figs. 2 and 3 and Supplementary Fig. S3, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). The *Sciurus* group was divided into three subgroups (Supplementary Fig. S3, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). The first subgroup (3.2) contained squirrels of North America and northern Mexico (except for the Neotropical *S. igniventris*), the second (3.1) consisted of Neotropical squirrels, and the third subgroup (3.3) contained both Neotropical and Palearctic squirrels.

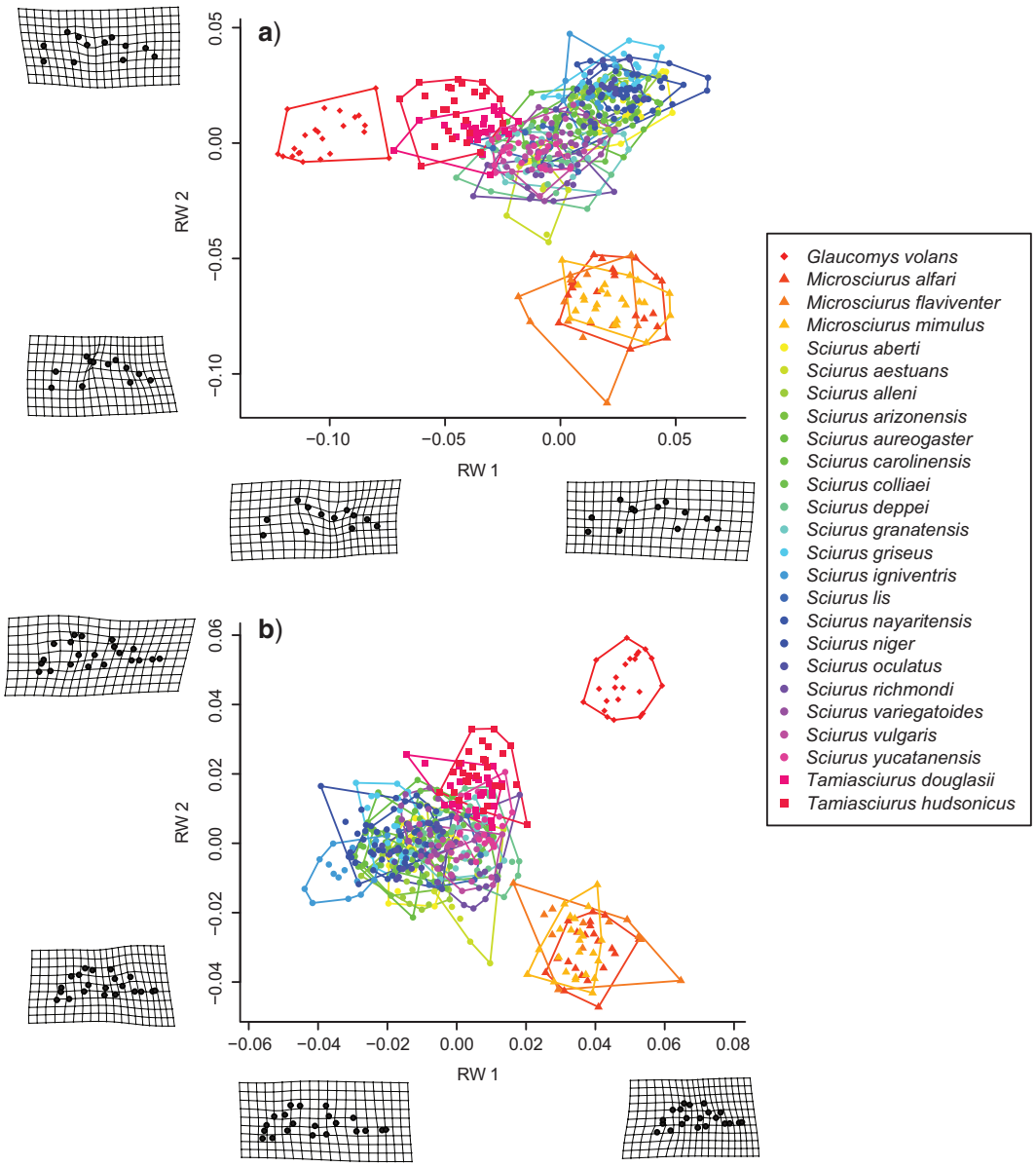


FIGURE 3. Plots of the first two relative warps (RW1, RW2) for dorsal (a) and ventral (b) sides of tree squirrel skulls. The first two RWs accounted for 62.5% and 47.6% of variability in the skull sides, respectively. Deformation grids show the shape change along the first two RW axes.

The most distinct features of the skull of *Microsciurus* in comparison to *Sciurus* and *Tamiasciurus* (except for their smaller and broader shape) were their more convex shape, short zygomatic arches, small orbits, short nasals, and wide interorbital region. In detail, the deformation grids for dorsal RW1 showed how with increasing values the interorbital region gets broader, that is, the interorbital constriction gets narrower, and the width

of a skull gets more uniform. The most negative values (the narrowest interorbital region and relatively broad posterior part of a skull) represented the outgroup taxon *G. volans*. *Tamiasciurus* had an intermediate form and *Microsciurus* and *Sciurus* species had broader interorbital space. The axis for dorsal RW2 showed size change of nasals and zygomatic arches that were shorter and wider in *Microsciurus* (negative values) and longer and narrower for, for example, *S. griseus* (positive values).

The axis for ventral RW1 showed differences in relative length of nasals and the anterior part of a skull that were shorter for more positive values (*Glaucomys* and *Microsciurus*) and longer for more negative values (e.g., *S. igniventris*). The axis for ventral RW2 showed that the relative width of the anterior part of a skull and zygomatic arches became narrower with more positive values (*Glaucomys*) and wider with more negative values (*Microsciurus*).

Spatio-Temporal and Ecological Hypotheses of Molecular and Morphometric Variation

We found a significant correlation between phylogenetic distances of taxa and distances of their species ranges (Mantel's test: $r = 0.3574$, $P = 0.0226$), indicating coupled dispersal and genetic diversification in Sciurini tree squirrels. The probability of occurrence of the ancestor of Sciurini in the Palearctic region is 0.70 and in the Nearctic region is 0.26. We found strong support for a single southward colonization with diversification of lineage III in the Americas (Fig. 2).

To test stability of the ancestral range of South American Sciurini, the number of generations of MCMC sampling was iteratively increased from 100 to 600 million. In each run, Western Amazonia was identified as the ancestral state (Fig. 4). However, in runs with <150 million generations, other putatively ancestral areas showed similar posterior probability. Estimated posterior distributions for parameters for all runs were long tailed, although estimated parameter values were similar. At 600 million generations, ESSs were around 400, posterior distribution was unimodal and trace showed no apparent spike.

In the molecular dating analysis, the log-normal relaxed clock outperformed a strict clock (BF >20), and there was no difference between performance of the tested speciation tree priors (BF <5). A birth–death speciation prior was chosen as a generalization of the Yule process that allows lineage speciation as well as extinction. The molecular dating showed that the diversification of Sciurini in South America occurred 3.7–6.03 Ma (Supplementary Fig. S5, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n).

The dietary preferences of Sciurini were modeled with the probit regression where the final model contained those morphological shape variables expressed as the RWs that provided the lowest AIC values for the model. The evaluated food items showed a strong correlation with several RWs with the exception of feeding on seeds, which did not significantly correlate with any RWs at $\alpha = 0.05$. Although residual variance remained high for multiple models, the misclassification rate was 8% for the

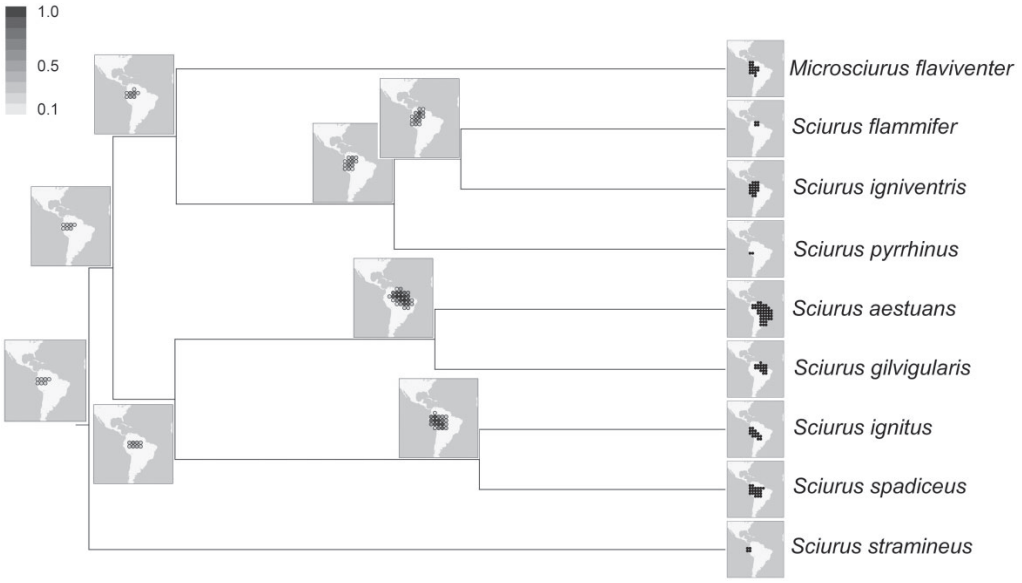


FIGURE 4. Ancestral ranges of South American squirrels reconstructed from the pruned chronogram containing South American species and their current distribution (maps at tips). The putative ancestral area was in the western Amazonia, on the east side of the Andes. Circle shading reflects Bayesian posterior probability that the ancestor occupied the grid cell.

combined RW data set, meaning that at the individual level, 8% of animals could not be correctly predicted to feed on a specific food source based on their skull morphology (Supplementary Table S10, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). The models based on the dorsal and ventral data had a 16% and 13% misclassification rate, respectively. The application of the majority rule over multiple samples per species yielded no improvement.

Our set of models most accurately predicted seeds, fruits, leaves, and bark as dietary preferences, while affinities for flowers and fungi were erroneously predicted in 4 out of 24 species (Supplementary Table S10, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). Dietary preferences of 18 species were predicted accurately with zero or one incorrect prediction.

The NJ tree (Fig. 5) based on the dietary preferences divided taxa according to differences in the composition of their diet. In general, the dietary groups did not reflect association with morphology, as species with similar skull morphology (according to the UPGMA analyses) were distributed in different dietary groups, indicating that the type of food consumed is not mirrored by skull shape. The only exceptions were species that feed on insects, bark, and exudates, which represent the exclusive diet of all *Microsciurus* species (Fig. 5).

Discussion

We showed that the phylogeny of Sciurini is decoupled from morphological differentiation in the Neotropics. Where the phylogeny reflected geographic occurrence of the species, the morphological cluster analyses showed three groups, which correspond with the current taxonomy of the tree squirrels.

Changes in the Skull Shape

Geometric morphometry recognized three morphological groups of Sciurini squirrels (*Microsciurus*, *Sciurus*, and *Tamiasciurus*). While species of *Microsciurus* and *Sciurus* form a clade on a phylogeny, morphologically they represent valid genera. The main characteristic distinguishing the skulls of *Microsciurus* and *Sciurus* is a shorter, broader, and more convex shape. The skulls of *Microsciurus* are characterized by small orbits with wider interorbital space, short zygomatic arches, short palatines and short nasals, more ventral position of foramen magnum. These findings are consistent with morphological description of the genus *Microsciurus* by Allen (1914).

Our results on cranial morphology of *Microsciurus* filled the knowledge gap on the morphology of *Microsciurus* as only data on postcranial and mandibular morphology were previously available. The comparison with other dwarf squirrels with similar morphology and behavior patterns suggests that the changes of skull shape in *Microsciurus* are associated with morphological convergence.

Morphological Convergence

The sharp conflict between molecular and morphological data was in the position of *Microsciurus*. In the molecular analyses, species of the genus *Microsciurus* belonged to different monophyletic groups and were closely related to species of the genus *Sciurus*. However, the morphological data recognized *Microsciurus* species as clearly distinct from *Sciurus* and *Tamiasciurus* species. The skulls of *Microsciurus* are smaller, broader, and more convex, with shorter nasals and wider interorbital regions compared with *Sciurus* and *Tamiasciurus* as already described by Allen (1914). We tested the role of feeding habits on skull morphology, and the probit model categorizing diet based on morphology as well as the comparison of morphological and dietary clusters showed an important influence of bark gleaning and insectivory in explaining skull shape in *Microsciurus*. The overall cranial morphology of *Microsciurus* resembles the cranial morphology of other dwarf squirrels (Neotropical *S. pusillus*, Afrotropical *M. pumilio*, and Oriental *Nannosciurus* and *Exilisciurus* species). The cranial features in dwarf squirrels are accompanied by lengthened limbs and other morphological adaptations for vertical climbing and claw clinging (Thorington and Thorington 1989; Thorington et al. 1997), along with deep mandibular bodies providing support to jaw muscles for a strong incisor bite, needed for bark feeding (Thorington and Darrow 1996; Velhagen and Roth 1997; Casanovas-Vilar and van Dam 2013). The morphological divergence of mandibles in dwarf squirrels is associated with allometry (Roth 1996; Velhagen and Roth 1997). However, each group of dwarf squirrels has a different evolutionary allometric trajectory influenced by, for example, adaptation (Hautier et al. 2009). Morphological similarities of *Microsciurus* with other dwarf squirrels, together with behavioral observations of tearing off bark, claw clinging, and active searching for arthropods in *Microsciurus* species (Youlatos 1999; Thorington et al. 2012) suggest that the morphological differentiation from other Sciurini is related to the reduced body size and adaptations to bark gleaning.

Dietary Preference

Even with relatively simple models, we have been able to achieve high accuracy in predicting dietary preferences (84% for models from combined dorsal and ventral information). However, response was coded as a binary variable that lacked information on proportion of dietary preferences and could not weigh evolutionary pressure on certain skull-shapes. Thus, although generalist squirrels that change diet depending on availability could have different skull shape than

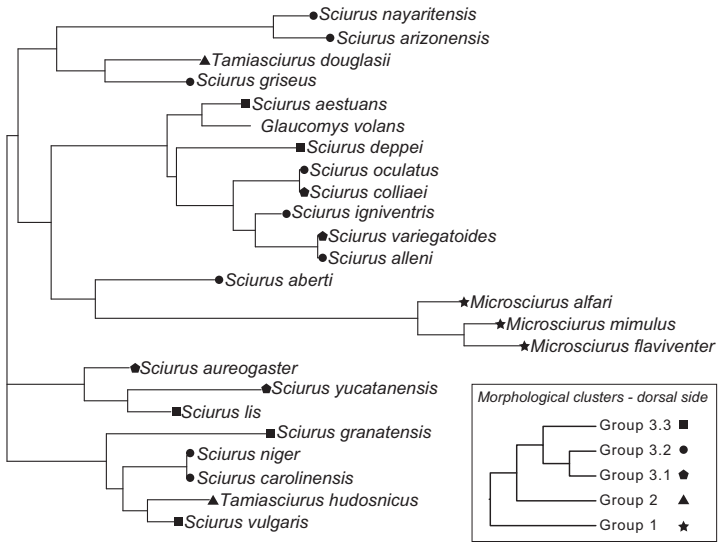


FIGURE 5. NJ tree built from Manhattan distances based on the dietary preference table of Sciurini. Clusters reflect species with similar diet composition. Tip labels indicate affiliation to morphological clusters (inset per Supplementary Fig. S3, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n). In general, dietary groups do not overlap with morphological clusters, except for species of *Microsciurus*.

specialists that continuously eat similar diet, our models could not express this relationship.

For those reasons, we have kept from overinterpretation of the dietary preferences models apart from bark gleaners, as the results for this group were supported from two additional analyses.

Molecular Phylogeny Conflicts with Taxonomy

Evolutionary history estimated from phylogenies supported previous results (Mercer and Roth 2003; Herron et al. 2004; Steppan et al. 2004; Pečnerová and Martínková 2012) suggesting that the genus *Tamiasciurus* (lineage I) is sister to all other taxa clustered in a separate clade (II and III). Compared with earlier research, we obtained more specific results for the monotypic genus *Rheithrosciurus*, which was recognized as a sister lineage (II) to the other taxa of lineage III (*Microsciurus*, *Sciurus*, and *Syntheosciurus*), suggesting an early divergence from a common ancestor with lineage III. Our results confirm that two crossings of Beringia occurred in the evolutionary history of Sciurini: an early crossing that caused divergence of the *Tamiasciurus* lineage and the other two lineages, and later an eastbound crossing after the split of *Rheithrosciurus* ancestors (Pečnerová and Martínková 2012). Additionally, the monophyly of lineage III provides additional support for the need of a taxonomic revision of the tribe Sciurini suggested already by Mercer and Roth (2003).

Tropical Mountain Forests as Epicenters of Squirrel Speciation

The colonization of the Neotropics by Sciurini was accompanied by a split between the Central American (III-A) and the South American clade (III-B) and extensive speciation within these clades. Timing of the diversification of the two clades predates the full formation of the Isthmus of Panama around 3 Ma. Sciurini were probably early colonizers of South America, accidentally transported during shoaling of the Isthmus of Panama. However, this date is dependent on the assumption that *Sciurus* arrived to North America earlier than 7.4 Ma (Supplementary Figure S5, available on Dryad at http://dx.doi.org/10.5061/dryad.v2t5n; Marincovich and Gladenkov 1999; Emry et al. 2005). The ranges of the majority of Neotropical Sciurini species are anchored around the eastern slopes of the Andes, suggesting that this area played a significant role in the evolutionary history of Sciurini. We used a simulation approach to reconstruct the ancestral range of South American tree squirrels and our results revealed an epicenter of tree squirrel speciation in the western Amazonia.

The center of diversification in wet tropical mountain forests demonstrates the importance of this habitat in the process of speciation. Villalobos (2013) studied the biogeography of Mesoamerican squirrels and concluded that a combination of vicariant barriers and dispersal events generated modern levels of Mesoamerican squirrel diversity and proposed that vicariant events

might be explained by ice age cycles. During the Pleistocene ice ages, tropical areas were drier, had less forest and more grassland (Hewitt 2000). According to the refuge hypothesis (Haffer 1997), patches of moist mountain forests served as refugia during the arid periods of ice ages and forests and forest species spread from the refugia in interglacials. However, multiple studies (e.g., Colinvaux et al. 2000; Bush and de Oliveira 2006) showed evidence that the extent of spreading savannas was not of a sufficient magnitude to generate forest refugia. A complex scenario, taking into consideration both approaches has been suggested (Turchetto-Zolet et al. 2013). Our data are consistent with the refuge hypothesis. Moist forest areas of western Amazonia, estimated as the ancestral area of the Sciurini distribution, correspond to the long-recognized Quaternary refugia of Napo and Inambari (Haffer 1969; Cracraft 1985). These refugia are defined by the Andes to the west and they provided a stable forest habitat during ice age cycles (Hewitt 2000). Our findings place tree squirrels of the tribe Sciurini among other Neotropical taxa with a Pleistocene origin of biodiversity, including monkeys (Chiou et al. 2011), birds (Sousa-Neves et al. 2013), butterflies (Garzón-Orduña et al. 2014), and trees (Richardson et al. 2001).

Conclusions

The integrated approach of molecular and morphological data and computational simulations revealed two key findings in the speciation of Sciurini. First, the rapid speciation of tree squirrels in South America was anchored in the area recognized as a Quaternary refugium in the western Amazonia in South America. Mountain forests probably served as areas where squirrels were relegated during the climatic oscillations in the Pleistocene. Second, the ecological diversity of tropical forests, in particular the availability of varied food resources, stimulated ecological adaptation. While molecular analyses considered the genus *Microsciurus* to be polyphyletic, with its species being nested deep inside the clade of the genus *Sciurus*, morphologically, the species of *Microsciurus* form a well-recognized cluster. Geometric morphometrics and analysis of dietary preferences linked the change in skull shape to bark gleaning insectivory, indicating that the present recognition of *Microsciurus* as a genus is based on morphological convergence and not on evolutionary history. By using a set of varied analytical tools, we were able to provide a picture of the processes leading to the rapid speciation in the Neotropics. Our study shows that by studying the evolutionary history of particular model groups we can gain better insight into the complex origin of one of the world's biodiversity hotspots.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.v2t5n.

Acknowledgments

The authors would like to acknowledge Ondřej Mikula for lending a digital camera to obtain photographs for morphological analyses and his valuable advice on geometric morphometrics. They thank Louise Tomsett from the NHM in London and Darrin Lunde from the National Museum of Natural History of the Smithsonian Institution in Washington, DC, for tissue samples and access to museum collections.

Funding

This work was supported by several grants and scholarships: Program rektora, Masaryk University [MUNI/C/0772/2011; Short visit grant, Frontiers of Speciation Research, European Science Foundation [4650]; PROVAZ: PROpojení Vzdělávání A nových přístupů v Zoologicko-ekologickém výzkumu—od teorie k praxi [in the frame of project CZ.1.07/2.4.00/17.0138; Scholarship for support of creative activity, Department of Botany and Zoology, Masaryk University [30144/2012].

References

Adams D.C., Otarola-Castillo E. 2013. geomorph: An R package for the collection and analysis of geometric morphometric shape data. Methods Ecol. Evol. 4:393–399.
Allen J.A. 1914. Review of the genus *Microsciurus*. Bull. Am. Mus. Nat. Hist. 33:145–165.
Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.
Bergsten J., Nilsson A.N., Ronquist F. 2013. Bayesian tests of topology hypotheses with an example from diving beetles. Syst. Biol. 62: 660–673.
Black C.C. 1963. A review of the North American Tertiary Sciuridae. Bull. Mus. Comp. Zool. 130:109–248.
Bollback J.P. 2006. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics 7:88.
Bookstein F.L. 1989. Principal warps—thin-plate splines and the decomposition of deformations. IEEE Trans. PAMI 11:567–585.
Bouckaert R., Heled J., Kühnert D., Vaughan T.G., Wu C.-H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. 2014. BEAST2: A software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 10:e1003537.
Bush M.B., de Oliveira P.E. 2006. The rise and fall of the refugial hypothesis of Amazonian speciation: A paleoecological perspective. Biota Neotrop. 6:1–17.
Cardini A. 2014. Missing the third dimension in geometric morphometrics: How to assess if 2D images really are a good proxy for 3D structures? Hystrix 25:73–81.
Cardini A., Hoffmann R.S., Thorington R.W. Jr. 2005. Morphological evolution in marmots (Rodentia, Sciuridae): Size and shape of the dorsal and lateral surfaces of the cranium. J. Zool. Syst. Evol. Res. 43:258–268.
Cardini A., O'Higgins P. 2004. Patterns of morphological evolution in Marmota (Rodentia, Sciuridae): Geometric morphometrics of the cranium in the context of marmot phylogeny, ecology and conservation. Biol. J. Linn. Soc. 82:385–407.
Casanovas-Vilar I., van Dam J. 2013. Conservatism and adaptability during squirrel radiation: What is mandible shape telling us? PLoS One 8:e61298.
Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30:3317–3324.

Chiou K.L., Pozzi L., Alfaro J.W.L., Di Fiore A. 2011. Pleistocene diversification of living squirrel monkeys (*Saimiri* spp.) inferred from complete mitochondrial genome sequences. Mol. Phylogenet. Evol. 59:736–745.
Colinvaux P.A., De Oliveira P.E., Bush M.B. 2000. Amazonian and neotropical plant communities on glacial time-scales: The failure of the aridity and refuge hypotheses. Quat. Sci. Rev. 19:141–169.
Cracraft J. 1985. Historical biogeography and pattern of differentiation within the South American avifauna: Areas of endemism. Ornithol. Monogr. 36:49–84.
Emry R.J., Korth W.W., Bell M.A. 2005. A tree squirrel (Rodentia, Sciuridae, Sciurini) from the late Miocene (Clarendonian) of Nevada. J. Vertebr. Paleontol. 25:228–235.
Ersts P.J. 2014. Geographic distance matrix generator. American Museum of Natural History, Center for Biodiversity and Conservation. Available from: URL http://biodiversityinformatics.amnh.org/open_source/gdmg, last accessed 14 August, 2015.
Fadda C., Corti M. 2000. Three dimensional morphometric study of the Ethiopian *Myomys-Stenocephalemys* complex (Murinae, Rodentia). Hystrix 10:131–143.
Garzón-Orduña I.J., Benetti-Longhini J.E., Brower A.V.Z. 2014. Timing the diversification of the Amazonian biota: Butterfly divergences are consistent with Pleistocene refugia. J. Biogeogr. 41:1631–1638.
Goujon M., McWilliam H., Li W., Valentin F., Squizzato S., Paern J., Lopez R. 2010. A new bioinformatics analysis tools framework at EMBL–EBI. Nucleic Acids Res. 38:W695–W699.
Gower J.C. 1975. Generalized procrustes analysis. Psychometrika 40:33–51.
Haffer J. 1969. Speciation in Amazonian forest birds. Science 165: 131–137.
Haffer J. 1997. Alternative models of vertebrate speciation in Amazonia: An overview. Biodivers. Conserv. 6:451–476.
Hall T.A. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41:95–98.
Hautier L., Fabre P.-H., Michaux J. 2009. Mandible shape and dwarfism in squirrels (Mammalia, Rodentia): Interaction of allometry and adaptation. Naturwissenschaften 96:725–730.
Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.
Herron M.D., Castoe T.A., Parkinson C.L. 2004. Sciurid phylogeny and the paraphyly of Holarctic ground squirrels (*Spermophilus*). Mol. Phylogenet. Evol. 31:1015–1030.
Hewitt G. 2000. The genetic legacy of the Quaternary ice ages. Nature 405:907–913.
Hooghiemstra H., van der Hammen T. 1998. Neogene and Quaternary development of the neotropical rain forest: The forest refugia hypothesis, and a literature overview. Earth-Sci. Rev. 44:147–183.
Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24:417–441.
Huelsenbeck J.P., Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.
IUCN. 2013. IUCN Red List of Threatened Species. Available from: URL www.iucnredlist.org (downloaded 19 January 2014).
Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30:772–780.
Keigwin L. 1982. Isotopic paleoceanography of the Caribbean and East Pacific: Role of Panama uplift in late Neogene time. Science 217: 350–353.
Landis M.J., Matzke N.J., Moore B.R., Huelsenbeck J.P. 2013. Bayesian analysis of biogeography when the number of areas is large. Syst. Biol. 62:789–804.
Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29:1695–1701.
Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gisson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.
Macholán M., Mikula O., Vohralík V. 2008. Geographic phenetic variation of two eastern-Mediterranean non-commensal mouse species, *Mus macedonicus* and *M. cypriacus* (Rodentia: Muridae) based on traditional and geometric approaches to morphometrics. Zool. Anz. 247:67–80.
Marincovich L. Jr., Gladenkov A.Yu. 1999. Evidence for an early opening of the Bering Strait. Nature 397:149–151.
Martínková N., Searle J.B. 2006. Amplification success rate of DNA from museum skin collections: A case study of stoats from 18 museums. Mol. Ecol. Notes 6:1014–1017.
McFerrin L. 2013. HDMD: Statistical analysis tools for high dimension molecular data (HDMD). R package version 1.2. Available from: URL http://CRAN.R-project.org/package=HDMD, last accessed 14 August, 2015.
Mercer J.M., Roth V.L. 2003. The effects of Cenozoic global change on squirrel phylogeny. Science 299:1568–1572.
Mittelbach G.G., Schemske D.W., Cornell H.V., Allen A.P., Brown J.M., Bush M.B., Harrison S.P., Hurlbert A.H., Knowlton N., Lessios H.A., McCain C.M., McCune A.R., McDade L.A., McPeek M.A., Near T.J., Price T.D., Ricklefs R.E., Roy K., Sax D.F., Schluter D., Sobel J.M., Turelli M 2007. Evolution and the latitudinal diversity gradient: Speciation, extinction and biogeography. Ecol. Lett. 10:315–331.
Moore J.C. 1959. Relationships among the living squirrels of the Sciurinae. Bull. Am. Mus. Nat. Hist. 118:153–206.
Oksanen J., Blanchet F.G., Kindt R., Legendre P., Minchin P.R., O'Hara R.B., Simpson G.L., Solymos P., Stevens M.H.H., Wagner H. 2013. vegan: Community ecology package. R package version 2.0-7. Available from: URL http://CRAN.R-project.org/package=vegan, last accessed 14 August, 2015.
Oshida T., Arslan A., Noda M. 2009. Phylogenetic relationships among the Old World *Sciurus* squirrels. Folia Zool. 58:14–25.
Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.
Pearson K. 1901. On lines and planes of closest fit to systems of points in space. Phil. Mag. 2:559–572.
Pečnerová P., Martínková N. 2012. Evolutionary history of tree squirrels (Rodentia, Sciurini) based on multilocus phylogeny reconstruction. Zool. Scr. 41:211–219.
Pianka E.R. 1966. Latitudinal gradients in species diversity—a review of concepts. Am. Nat. 100:33–46.
Pääbo S., Poinar H., Serre D., Jaenicke-Després V., Hebler J., Rohland N., Kuch M., Krause J., Vigilant L., Hofreiter M. 2004. Genetic analyses from ancient DNA. Annu. Rev. Genet. 38:645–679.
R Core Team. 2013. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from: URL http://www.R-project.org/, last accessed 14 August, 2015.
Rambaut A., Suchard M.A., Xie D., Drummond A.J. 2013. Tracer v1.5. Available from: URL http://beast.bio.ed.ac.uk/Tracer, last accessed 14 August, 2015.
Ranwez V., Criscuolo A., Douzery E.J. 2010. SuperTriplets: A triplet-based supertree approach to phylogenomics. Bioinformatics 26:i115–i123.
Richardson J.E., Pennington R.T., Pennington T.D., Hollingsworth P.M. 2001. Rapid diversification of a species-rich genus of neotropical rain forest trees. Science 293:2242–2245.
Rohlf F.J. 1993. Relative warp analysis and example of its application to mosquito wings. In: Marcus LF, Bello E, Garcia-Valdecasas A, editors. Contribution to morphometrics. Madrid: Museo Nacional de Ciencias Naturales. p. 131–159.
Rohlf F.J. 1998. On applications of geometric morphometrics to studies of ontogeny and phylogeny. Syst. Biol. 47:147–158.
Rohlf F.J. 2003. tpsSmall. New York: Department of Ecology and Evolution, State University of New York at Stony Brook. Available from: URL http://life.bio.sunysb.edu/morph/index.html.
Rohlf F.J. 2005. tpsDig2. New York: Department of Ecology and Evolution, State University of New York at Stony Brook. Available from: URL http://life.bio.sunysb.edu/morph/index.html, last accessed 14 August, 2015.
Rohlf F.J. 2010. tpsRelw. New York: Department of Ecology and Evolution, State University of New York at Stony Brook. Available from: URL http://life.bio.sunysb.edu/morph/index.html, last accessed 14 August, 2015.
Rohlf F.J., Slice D. 1990. Extensions of the procrustes method for the optimal superimposition of landmarks. Syst. Zool. 39:40–59.

Roth V.L. 1996. Cranial integration in the Sciuridae. Amer. Zool. 36: 14–23.
Roth V.L., Mercer J.M. 2008. Differing rates of macroevolutionary diversification in arboreal squirrels. Curr. Sci. 95:857–861.
Rull V. 2008. Speciation timing and neotropical biodiversity: The Tertiary-Quaternary debate in the light of molecular phylogenetic evidence. Mol. Ecol. 17:2722–2729.
Sanderson M.J., McMahon M.M., Steel M. 2011. Terraces in phylogenetic tree space. Science 333:448–450.
Schluter D. 2001. Ecology and the origin of species. Trends Ecol. Evol. 16:372–380.
Scornavacca C., Berry V., Lefort V., Douzery E.J., Ranwez V. 2008. PhySIC_IST: Cleaning source trees to infer more informative supertrees. BMC Bioinformatics 9:413.
Sousa-Neves T., Aleixo A., Sequeira F. 2013. Cryptic patterns of diversification of a widespread Amazonian Woodcreeper species complex (Aves: Dendrocolaptidae) inferred from multilocus phylogenetic analysis: Implications for historical biogeography and taxonomy. Mol. Phylogenet. Evol. 68:410–424.
Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Steele M.A. 1998. *Tamiasciurus hudsonicus*. Mammal. Species 586:1–9.
Steele M.A. 1999. *Tamiasciurus douglasii*. Mammal. Species 630:1–8.
Steppan S.J., Storz B.L., Hoffmann R.S. 2004. Nuclear DNA phylogeny of the squirrels (Mammalia: Rodentia) and the evolution of arboreality from *c-myc* and *RAG*1. Mol. Phylogenet. Evol. 30: 703–719.
Suchard M.A., Weiss R.E., Sinsheimer J.S. 2005. Models for estimating Bayes factors with applications to phylogeny and tests of monophyly. Biometrics 61:665–673.
Swofford D. 2015. PAUP 4.0a142. Available from: URL http://people.sc.fsu.edu/~dswofford/paup_test/ (24 February 2015).
Thorington R.W. Jr., Darrow K. 1996. Jaw muscles of old world squirrels. J. Morphol. 230:145–165.
Thorington R.W. Jr., Darrow K., Betts A.D.K. 1997. Comparative myology of the forelimb of squirrels (Sciuridae). J. Morphol. 234:155–182.
Thorington R.W. Jr., Koprowski J.L., Steele M.A., Whatton J.F. 2012. Squirrels of the World. Baltimore: The Johns Hopkins University Press.
Thorington R.W. Jr., Thorington E.M. 1989. Postcranial proportion of *Microsciurus* and *Sciurillus*, the American pygmy squirrels. In: Redfor KH, Eisenberg JF, editors. Advances in neotropical mammalogy. Gainesville, FL: Sandhill Crane Press, p. 125–136.
Turchetto-Zolet A.C., Pinheiro F., Salgueiro F., Palma-Silva C. 2013. Phylogeographical patterns shed light on evolutionary process in South America. Mol. Ecol. 22:1193–1213.
Vaidya G., Lohman D.J., Meier R. 2011. SequenceMatrix: Concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics 27: 171–180.
Velhagen W.A., Roth V.L. 1997 Scaling of the mandible in squirrels. J. Morphol. 232:107–132.
Via S. 2001. Sympatric speciation in animals: The ugly duckling grows up. Trends Ecol. Evol. 16:381–390.
Villalobos F. 2013. Tree squirrels: A key to understand the historic biogeography of Mesoamerica? Mamm. Biol. 78:258–266.
Wilson D.E., Reeder D.M. 2005. Mammal species of the world. A taxonomic and geographic reference. Baltimore: The Johns Hopkins University Press.
Wing S.L. 1998. Tertiary vegetational history of North America as a context for mammalian evolution. In: Janis C, Jacobs L, Scott K, editors. Evolution of tertiary mammals of North America: Vol. I: Terrestrial carnivores, ungulates, and ungulate-like mammals. Cambridge: Cambridge University Press. p. 37–65.
Wolfe J.A., Hopkins D.M., Leopold E.B. 1966. Tertiary stratigraphy and paleobotany of the Cook Inlet region. Alaska: Discussion of stratigraphic significance of fossil plants from the Chickaloon, Kenai, and Tsadaka formations. U.S. Geological Survey Professional Paper 398-A.
Youlatos D. 1999. Locomotor and postural behavior of *Sciurus igniventris* and *Microsciurus flaviventer* (Rodentia, Sciuridae) in eastern Ecuador. Mammalia 63:405–416.

Paper 2.4.1

Zukal J., Bandouchova H., Bartonicka T., Berkova H., Brack V., Brichta J., Dolinay M., Jaron K. S., Kovacova V., Kovarik M., **Martínková N.**, Ondracek K., Rehak Z., Turner G. G., Pikula J. 2014. White-nose syndrome fungus: a generalist pathogen of hibernating bats. *PLoS ONE* 9: e97224.

OPEN ACCESS Freely available online

PLOS ONE

White-Nose Syndrome Fungus: A Generalist Pathogen of Hibernating Bats

Jan Zukal[1,2], Hana Bandouchova[3], Tomas Bartonicka[2], Hana Berkova[1], Virgil Brack[4], Jiri Brichta[3], Matej Dolinay[2], Kamil S. Jaron[5], Veronika Kovacova[3], Miroslav Kovarik[6], Natália Martínková[1,5], Karel Ondracek[3], Zdenek Rehak[2], Gregory G. Turner[7], Jiri Pikula[3]*

1 Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, Brno, Czech Republic, **2** Department of Botany and Zoology, Masaryk University, Brno, Czech Republic, **3** Department of Ecology and Diseases of Game, Fish and Bees, University of Veterinary and Pharmaceutical Sciences, Brno, Czech Republic, **4** Environmental Solutions & Innovations Inc., Cincinnati, Ohio, United States of America, **5** Institute of Biostatistics and Analysis, Masaryk University, Brno, Czech Republic, **6** Administration of the Moravian Karst Protected Landscape Area, Nature Conservation Agency of the Czech Republic, Blansko, Czech Republic, **7** Pennsylvania Game Commission, Harrisburg, Pennsylvania, United States of America

Abstract

Host traits and phylogeny can determine infection risk by driving pathogen transmission and its ability to infect new hosts. Predicting such risks is critical when designing disease mitigation strategies, and especially as regards wildlife, where intensive management is often advocated or prevented by economic and/or practical reasons. We investigated *Pseudogymnoascus [Geomyces] destructans* infection, the cause of white-nose syndrome (WNS), in relation to chiropteran ecology, behaviour and phylogenetics. While this fungus has caused devastating declines in North American bat populations, there have been no apparent population changes attributable to the disease in Europe. We screened 276 bats of 15 species from hibernacula in the Czech Republic over 2012 and 2013, and provided histopathological evidence for 11 European species positive for WNS. With the exception of *Myotis myotis*, the other ten species are all new reports for WNS in Europe. Of these, *M. emarginatus, Eptesicus nilssonii, Rhinolophus hipposideros, Barbastella barbastellus* and *Plecotus auritus* are new to the list of *P. destructans*-infected bat species. While the infected species are all statistically phylogenetically related, WNS affects bats from two suborders. These are ecologically diverse and adopt a wide range of hibernating strategies. Occurrence of WNS in distantly related bat species with diverse ecology suggests that the pathogen may be a generalist and that all bats hibernating within the distribution range of *P. destructans* may be at risk of infection.

Citation: Zukal J, Bandouchova H, Bartonicka T, Berkova H, Brack V, et al. (2014) White-Nose Syndrome Fungus: A Generalist Pathogen of Hibernating Bats. PLoS ONE 9(5): e97224. doi:10.1371/journal.pone.0097224

Editor: Justin G. Boyles, Southern Illinois University, United States of America

Received January 5, 2014; **Accepted** April 16, 2014; **Published** May 12, 2014

Funding: This study was supported by the Grant Agency of the Czech Republic (Project No. P506/12/1064) and by the National Speleological Society of the USA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: V.B. is currently affiliated with a commercial company Environmental Solutions & Innovations Inc. He provided expertise in bat ecology in this paper. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

* E-mail: pikulaj@vfu.cz

Introduction

Host-pathogen dynamics represent a balance between the pathogen's ability to infect and the host's ability to resist, with an intensive arms race between the two reflected in co-evolutionary adaptations. With host switching, pathogens temporarily escape the arms race. New, naive hosts may show lower resistance and other characteristics favourable to the pathogen. Overlapping distribution of a pathogen and its potential host(s) is key to host switching driven by opportunity [1]. The spread of emerging wildlife pathogens may have economic consequences, even in species indirectly linked to humans [2]. Fungal infections in amphibians and bats that result in population declines [3], for example, can lead to increased agricultural costs where humans chemically compensate for ecosystem services provided by these organisms in terms of insect control.

White-nose syndrome (WNS) is an emerging disease of hibernating bats associated with skin infection by *Pseudogymnoascus [Geomyces] destructans*, a recently recognised fungal pathogen [4–7]. Severe skin damage results in disruption of torpor pattern, premature depletion of fat reserves and mortality in affected bats in North America [8]. High mortality rates at affected localities and the rapid spread of the infection since 2006 continues to threaten bat diversity [9–11].

While WNS has characteristics of an epizootic, gradually expanding through North American hibernacula from its original detection site [12,13], *P. destructans* is pan-European in distribution [14,15]. Aside from seasonality in the appearance of white fungal growth [15], detailed spatio-temporal data for *P. destructans* infection in Europe are lacking. Fortunately, mass mortality has not been observed in European bats to date [14].

WNS can be transmitted either directly through bat-to-bat contact or indirectly through contact with pathogen propagules in the environment [6,16] and the infection's spread is assumed to be both density- and frequency-dependent [10]. Multiple factors, such as hibernation in large assemblages or length of hibernation season, play a role in the epidemiology of this fungal disease and ecological and behavioural characteristics of bat species may affect the risk of infection [3]. Traits such as selection of hibernaculum roost sites with differing microclimatic conditions and solitary

versus gregarious hibernation behaviour may also influence the impact of WNS [10]. Other risk factors associated with hibernating bat mortality in North America include distance from the first WNS-affected site, cluster size, species diversity and composition and type of hibernaculum [13].

Prior to 2012, bat species positive for *P. destructans* in North America included *Myotis austroriparius*, *M. grisescens*, *M. leibii, M. lucifugus, M. septentrionalis, M. sodalis*, *M. velifer, Perimyotis subflavus* and *Eptesicus fuscus* [11,16,17]. Dermatohistopathology has revealed fungal infection with cupping erosions and skin invasion diagnostic for WNS in a *M. myotis* specimen hibernating in the Moravian Karst, Czech Republic [18] and eight species (*M. myotis, M. oxygnathus/M. blythii, M. brandtii, M. daubentonii, M. dasycneme, M. mystacinus, M. nattereri* and *M. bechsteinii*) have been reported positive for the WNS fungus in Europe based on direct microscopy of characteristic *P. destructans* conidia, fungal culture and genetic analysis [14,15,19–22]. Photographic evidence of fungal growth suggests that *M. emarginatus, E. nilssonii* and *Rhinolophus hipposideros* may also prove positive for *P. destructans* [14].

In summary, a total of 17 vespertilionid bat species had been reported positive for the WNS fungus in North America and Europe prior to 2012 and, as the epizootic spreads through North America and surveillance continues in Europe, it is expected that the number of infected species will increase. Hereinafter, the term *P. destructans*-infected or -positive relates to those species for which the fungal pathogen has been confirmed by laboratory methods such as fungal culture and genetic analysis. WNS-positive represents those species where the infection has been diagnosed through characteristic histopathological lesions, such as fungal hyphae densely packed in so-called cupping erosions and/or invasion of the dermis [23].

Little is known about *P. destructans* infection in European bat species less abundant or less commonly observed in hibernacula. Knowledge of pathological effects associated with the WNS fungus in European bat species is even poorer. While it is a commonly held view that European bats are more resistant or resilient than those in North America [17], our monitoring revealed three further species positive for *P. destructans* infection in 2012. Differences in their hibernation behaviour and taxonomy inspired us to examine the ecological and behavioural traits and phylogeny of European and North American species reported positive for *P. destructans* in order to identify any similarities in behaviour and habitat use and to identify any other species that may be at risk.

First, we examined the hypothesis that more bat species are positive for WNS in Europe than currently reported via histopathology, considered as the 'gold standard' for diagnosing WNS [23,24]. Second, we hypothesised that ecology, behaviour and phylogenetic relationships of hibernating bat species influence risk of infection by *P. destructans*. Aside from ecological similarities, those species most often found positive for *P. destructans* and WNS belong to the genus *Myotis*, indicating that phylogenetic relatedness of hosts may facilitate invasion by the fungus. To test this, we compared the ecological and behavioural traits of hibernating bats from Europe and North America. We grouped species with similar behaviour and habitat use and used confirmed positive species to propose possible additional species susceptible to infection. We constructed a phylogeny of vespertilionids and rhinolophids from Europe and North America, to test the hypothesis that infected bats are phylogenetically closely related. Finally, we screened species of unknown infection status in Czech hibernacula to test the validity of our models and predictions.

Here, we provide histopathological evidence of multiple European species positive for WNS. We found that infected bat species are ecologically diverse, utilising a range of hibernating and feeding strategies. Although bat species previously described as being *P. destructans*-positive have been phylogenetically related, the pattern begins to break down with the newly diagnosed taxa; the data presented herein demonstrating that the host range for this fungal pathogen is more diverse than previously realized.

Materials and Methods

Ethics statement

The Czech Academy of Sciences' Ethics Committee has reviewed and approved Animal Use Protocol No. 169/2011 in compliance with Law No. 312/2008 on Protection of Animals against Cruelty, as adopted by the Parliament of the Czech Republic. Bats were monitored for WNS and presence of the causative agent *P. destructans* in the spring of 2012 and 2013 in caves of the Moravian Karst, mines near Mala Moravka in the Jeseniky Mountains, and in the Podyji National Park, all in the Czech Republic. Non-lethal sampling was in compliance with Law No. 114/1992 on Nature and Landscape Protection and was based on permits 01662/MK/2012S/00775/MK/2012, 866/JS/2012 and 00356/KK/2008/AOPK issued by the Nature Conservation Agency of the Czech Republic. Bats were handled so as to minimise stress and duration of sampling procedures between capture and release. Numbered aluminium rings were attached around the forearm for long-term identification prior to release at the site.

Screening bat species in Czech hibernacula for *P. destructans* infection

When screening bats for WNS and *P. destructans* we 1. captured bats emerging from hibernacula at the end of the hibernation season using mist nets, 2. swabbed the wing membrane for fungal culture using the Fungi-Quick transport system (Copan Innovation, Italy), 3. briefly illuminated the bats with a flashlight to detect any visible fungal growth, 4. trans-illuminated the wing membrane using ultraviolet light (UV; wavelength 368 nm) to detect any WNS lesions, 5. photographed wing membranes of each bat under both visible and UV light, 6. took a wing punch biopsy from all WNS-suspected skin lesions (i.e. areas of orange-yellow fluorescence) using a sterile and disposable 4 mm skin biopsy punch (Kruuse, Denmark), 7. used polymerase chain reaction (PCR) to confirm *P. destructans* from fungal cultures or skin swabs using the FLOQSwabs system (CopanFlock Technologies, Italy), and 8. undertook complete histopathological examinations of skin samples.

A total of 276 bats were screened for WNS and *P. destructans* and 123 skin biopsies were taken for histological examination from 15 bat species (Table 1).

Formalin-fixed punch biopsy samples were embedded in paraffin and serial 5 µm tissue sections were prepared and stained for fungi with periodic acid–Schiff stain. Histopathological findings were classified as WNS based on previously described diagnostic criteria [23]. Samples collected to cultivate fungi were transferred onto Petri dishes containing Sabouraud agar, sealed with parafilm, inverted and incubated in the dark at 10 °C. Pure fungal cultures were established from fungal growth developing at 14 days or later. *Pseudogymnoascus destructans* was confirmed through characteristic asymmetrically curved conidia via direct microscopy [5]. Fungal isolates or skin swabs were further identified using PCR and follow-up sequencing of amplicons [25] and real-time PCR [24]. A pure culture of *P. destructans* isolate grown at 10 °C on Sabouraud agar and genetically confirmed (EMBL-Bank accession number: HE588133; [18]) served as a PCR control.

Table 1. Bats examined for white-nose syndrome and *Pseudogymnoascus destructans* infection in Czech hibernacula (Europe).

Species	Screened	Biopsied	Histo+	WNS prevalence (%)	St. error
Myotis myotis[a]	67	56	37	55.22	6.08
Myotis daubentonii[b]	25	13	4	16.00	5.76
Myotis bechsteinii[b]	21	7	2	9.52	6.78
Myotis nattereri[b]	20	8	3	15.00	7.10
Myotis brandtii[b]	17	1	1	5.88	8.23
Myotis alcathoe[b]	8	7	0	0	15.94
Myotis emarginatus[b]	39	7	5	12.82	5.35
Rhinolophus hipposideros[b]	28	5	1	3.57	5.18
Eptesicus nilssonii[b]	4	1	1	25.00	26.89
Plecotus auritus[c]	23	11	5	21.73	8.60
Barbastella barbastellus[c]	17	3	3	17.64	8.24
Plecotus austriacus	3	1	0	0	32.22
Eptesicus serotinus	2	1	0	0	39.61
Pipistrellus pipistrellus	1	1	0	0	48.47
Myotis dasycneme[b]	1	1	1	100	48.47
Total	**276**	**123**	**63**	**22.82**	**2.53**

[a] = species reported positive for WNS fungus prior to 2012, [b] = species recognised as positive in 2012, [c] = bat species recognised as positive in 2013. Screened = numbers of bats captured and examined using UV light trans-illumination to detect WNS lesions, biopsied = numbers of bats biopsied due to WNS-suspected skin lesions viewed under UV light, histo+ = specimens positive for WNS diagnostic features under histopathological examination (i.e. cupping erosions and fungal invasion of dermis), WNS prevalence = percentage of bats positive on histopathology out of the total number screened.
doi:10.1371/journal.pone.0097224.t001

Analysis of European and North American bat ecological traits

The list of European and North American bat species was prepared according to Simmons [26]; those species included in the study being those with complete or partial distribution in continental Europe or North America for which data are available. In total, we reviewed ecological and behavioural variables for 87 species. Of these, 47 were assessed for all variables and were subjected to ecological modelling analysis.

Eleven traits were chosen to describe bat species: 1. Infection status (*P. destructans*-positive/*P. destructans*-negative), 2. Cave or non-cave hibernation, 3. Region (Palearctic/Nearctic distribution), 4. Clustering during hibernation (clustering/non-clustering; i.e. hibernating in groups where multiple individuals touch), 5. Temperature preference (thermophilic/cryophilic; mainly according to Webb *et al.* [27]), 6. Preferred roost type during hibernation (exposed/hidden/both), 7. Size of clusters during hibernation (no clustering/small clusters < 50 bats/medium clusters of 51 to 500 bats/large clusters > 500 bats), 8. Distribution range (very large area/large area of approximately half a continent/moderate size/small area/very small area; according to Horáček *et al.* [28]), 9. Food (dominant insect food group represented by Diptera/Lepidoptera/Coleoptera/generalists), 10. Foraging habitat (open/edge/closed), 11. Body size (small - up to 5 g/medium - 5 to10 g/large - over 10 g).

These traits, which were assessed based on a primary literature review and expert evaluation, were chosen as those most likely to influence susceptibility of bats to WNS, the spread of *P. destructans* infection or survival rate during hibernation [10]. As the disease can inflict long-term damage to affected bats surviving the hibernation season, other medically relevant factors may also influence survival and reproduction in the active season [29]. Categorical variables were coded as $n - 1$ binary, dummy variables, where n is the number of categories. Data for the traits of each bat species are provided in Table S1. Grouping of ecologically similar species was performed via neighbour-joining clustering of squared Euclidean distances using ape in R language [30,31].

Phylogenetic reconstruction of European and North American bats

The phylogeny of bats from Europe and North America was extracted from a maximum likelihood phylogenetic tree using Phylocom version 4.2 [32]. The complete phylogeny of the Vespertilionidae, Miniopteridae and Cistugidae families (from which the tree used here was pruned) was reconstructed from a concatenated DNA sequence matrix of 13 mitochondrial and nuclear genes with 64% of missing data (Table S2). Three *Rhinolophus* species were used as an outgroup. Phylogeny was reconstructed using the partitioning scheme suggested by the greedy algorithm, utilising the Bayesian Information Criterion assessment in PartitionFinder [33]. The tree space was searched using maximum likelihood analysis with automatic majority-rule bootstopping option [34]. By extracting the target species' phylogeny from a comprehensive tree, we were able to obtain a phylogeny that exploits currently available diversity to optimise relationships and branch lengths, and thus mitigate possible analysis artefacts.

Statistical analysis and hypothesis testing of *P. destructans* infection occurrence

We explored the distribution pattern of *P. destructans* infection on a tree based on bat trait variation and molecular phylogeny. Occurrence of *P. destructans* infection represents a presence/absence variable, rather than a continuous trait, meaning that it is suitable for community structure analysis. The phylogenetic signal for explanatory variables and for *P. destructans* infection was calculated in Phylocom using the comstruct function. In order to

assess relatedness of species that share a specific characteristic, mean phylogenetic distance (MPD) and mean nearest phylogenetic taxon distance (MNTD) were compared to the null model, which assumes random dispersal of the trait on the tree. MPD measures the mean branch length between two randomly selected taxa from a sample, and is calculated as the sum of branch lengths to the node representing their most recent common ancestor. MNTD is the mean branch length between a taxon within the sample and its nearest relative. The null model randomised samples across phylogeny in 9,999 replicates. The distribution of the trait on a tree is clustered if values of MPD and MNTD obtained are higher than 95% of values obtained from the null samples standardised by the standard deviation of the null samples. The comparison is expressed as net relatedness index (NRI) and nearest taxon index (NTI) greater than zero [32]. Clustered distribution of a trait or phylogenetic signal means that species that share the trait are more closely related to one another than to a random taxon sampled from the tree.

Species' ecological and behavioural characteristics often show a heritable component such that close relatives have similar traits [35]. Such characteristics might then be adaptive and their evolution further decoupled from the assumption of sample independence needed for general statistical approaches. The evolutionary relationships of traits in our dataset were removed from comparisons by using phylogenetic generalised least squares (PGLS) in the Caper package of R [36]. We used a variance-covariance matrix calculated from the phylogeny with branch lengths transformed according to the Ornstein-Uhlenbeck model in geiger [37]. The PGLS model was developed via a step-down procedure, using the Akaike Information Criterion (AIC) to compare alternative models.

Results

Screening species with unknown infection status in Czech hibernacula

We tested a broad diversity of European hibernating bats for *P. destructans* infection and skin lesions pathognomonic for WNS. Analysis of 123 skin biopsy samples collected in 2012 and 2013 revealed histopathological findings matching criteria used for diagnosis of WNS in 63 bats (22.82% prevalence; Table 1) of 11 species, i.e. *M. myotis*, *M. daubentonii*, *M. bechsteinii*, *M. nattereri*, *M. brandtii*, *M. emarginatus*, *M. dasycneme*, *E. nilssonii*, *R. hipposideros*, *B. barbastellus* and *P. auritus* (Figure 1). With the exception of *M. myotis*, the other ten species are all new reports of WNS in Europe. Of these, *M. emarginatus*, *E. nilssonii*, *R. hipposideros*, *B. barbastellus* and *P. auritus* are new to the list of *P. destructans*-infected bat species. Fungus isolates or skin swabs from histopathologically positive bats were identified as *P. destructans* using PCR.

Risk of *P. destructans* infection in bat species of unknown infection status

Testing the hypothesis of phylogenetic relatedness. *P. destructans*-infected species were clustered together by molecular phylogeny (MPD = 0.212, NRI = 2.913, $p < 0.001$), meaning that pairs of infected species were, on average, more closely related than random species pairs from Europe and North America. When sister species or nearest relatives were considered, however, our results indicated that infection of both, either or neither was random (MNTD = 0.111, NTI = 1.556, $p = 0.06$; Table 2, Figure 2). Nine explanatory variables showed NRI and/or NTI not equal to zero, indicating that phylogenetic comparative methods were needed due to relatedness of taxa with shared traits (Table 2).

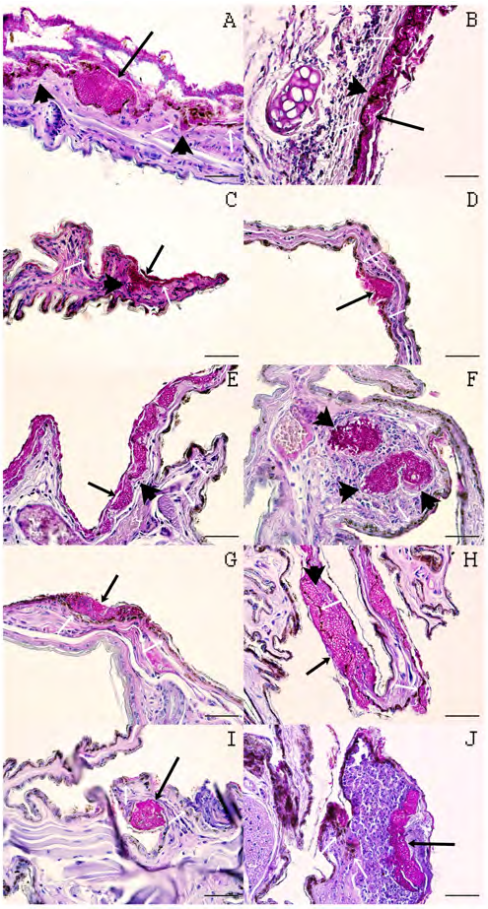


Figure 1. Histopathological skin lesions consistent with white-nose syndrome in ten European bat species. (A) *Myotis emarginatus*, (B) *Eptesicus nilssonii*, (C) *Rhinolophus hipposideros*, (D) *Plecotus auritus*, (E) *Barbastella barbastellus*, (F) *M. dasycneme*, (G) *M. nattereri*, (H) *M. daubentonii*, (I) *M. bechsteinii*, (J) *M. brandtii*. The photographs illustrate i) extensive infection of the wing membrane and cup-shaped epidermal erosions (A, E, H, J; long black arrow); ii) cup-like epidermal erosions in the pinna (B; long black arrow), iii) *Pseudogymnoascus destructans* hyphae obscuring the basement membrane and invading the dermis (A, B, C, E, H; black arrow); iv) a single cupping erosion packed with fungal hyphae in the wing membrane (C, D, G, I; long black arrow); v) colonisation of a hair follicle by *P. destructans*, fungal hyphae present in the associated sebaceous gland and regional connective tissue (F; black arrow); vi) marked signs of inflammation (B, F, J); and vii) a cellular inflammatory crust that sequesters fungal hyphae (A, J). White arrows within each photograph indicate the interface between epidermis and dermis. Periodic acid-Schiff stain; scale bar = 50 µm. *M. myotis* not shown because WNS lesions in this species have already been documented elsewhere [18].
doi:10.1371/journal.pone.0097224.g001

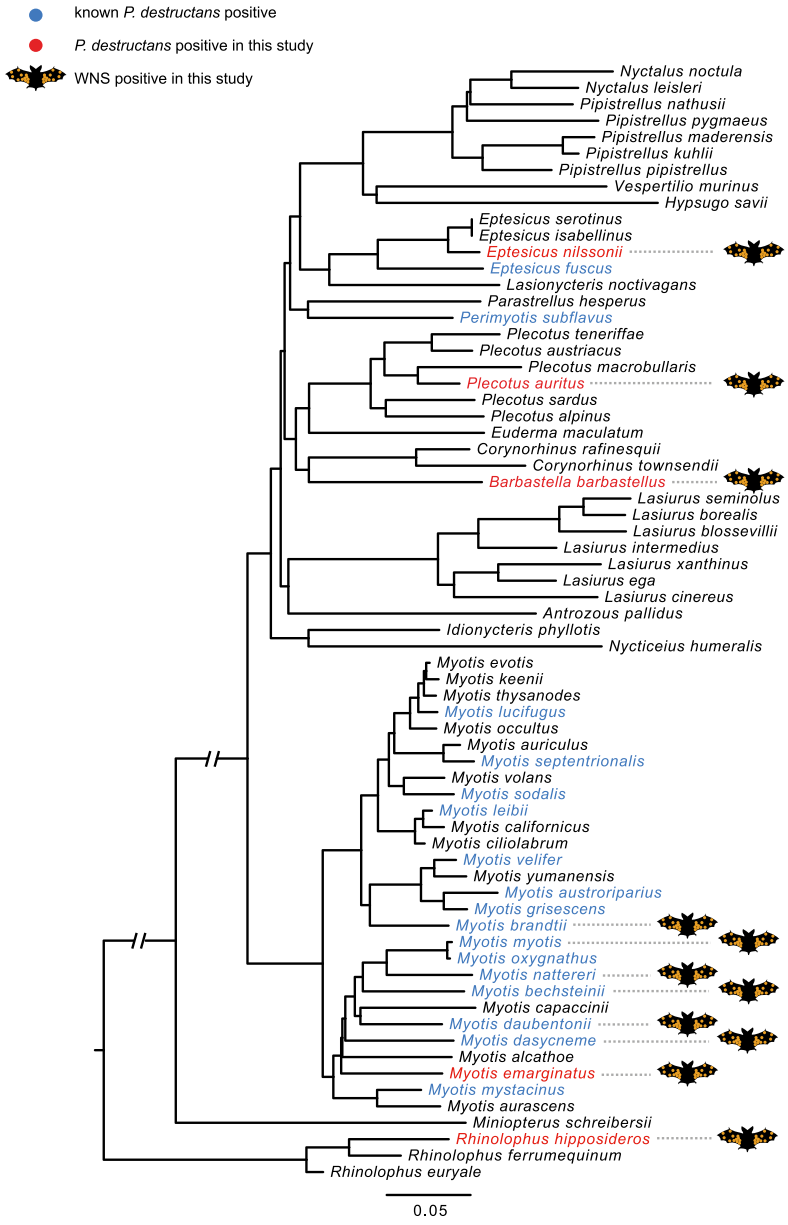


Figure 2. Phylogenetic reconstruction of bats from Europe and North America. The reconstruction was based on a concatenated DNA sequence matrix from 13 loci, purged from a maximum likelihood vespertilionid phylogeny rooted on *Rhinolophus*. Blue = species reported positive for WNS fungus prior to 2012, red = species recognised positive in this study, bat image = bat species diagnosed as WNS positive in this study. doi:10.1371/journal.pone.0097224.g002

Table 2. *Pseudogymnoascus destructans* infection in relation to chiropteran phylogeny and ecological similarity.

Variable	*N*	MPD	NRI	p_{NRI}	MNTD	NTI	p_{NTI}
Explanatory							
CAVE	42	0.270	1.300	0.105	**0.091**	**1.984**	**0.025**
REGION	23	0.332	−1.606	0.944	0.114	1.125	0.135
CLUSTER	27	**0.241**	**2.151**	**0.012**	0.109	1.180	0.124
TEMPERATURE	21	0.318	−1.020	0.841	0.128	0.569	0.292
SHELTERhidden	31	**0.246**	**2.147**	**0.013**	0.121	0.115	0.450
SHELTERexposed	21	0.316	−0.929	0.814	0.119	1.004	0.165
SHELTERboth[a]	5	0.224	1.051	0.106	0.170	0.754	0.229
CSIZEno	20	0.337	−1.639	0.947	0.125	0.736	0.240
CSIZEsmall	7	**0.126**	**3.025**	**0.001**	**0.082**	**2.474**	**0.004**
CSIZEmedium	9	0.280	0.201	0.474	0.162	0.404	0.358
CSIZElarge	11	0.265	0.592	0.306	0.166	0.064	0.488
RANGEverylarge	15	0.333	−1.217	0.875	0.188	−1.328	0.902
RANGElarge	13	0.311	−0.565	0.708	**0.213**	**−1.872**	**0.970**
RANGEmoderate	14	0.246	1.189	0.112	0.133	0.859	0.203
RANGEsmall	5	**0.098**	**2.865**	**0.001**	**0.059**	**2.675**	**0.001**
FOODcolleoptera	5	0.258	0.482	0.330	0.169	0.706	0.249
FOODdiptera	9	0.237	1.093	0.117	0.144	0.902	0.188
FOODgeneralist	18	0.313	−0.763	0.770	0.128	0.772	0.232
FOODlepidoptera	14	0.286	0.117	0.475	0.147	0.355	0.368
FOODother	1	n/a					
HABITATclosed	11	0.280	0.228	0.458	0.142	0.802	0.220
HABITATopen	13	0.314	−0.623	0.724	0.178	−0.652	0.739
HABITATedge	23	0.267	0.855	0.211	**0.093**	**2.321**	**0.007**
BODYsmall	10	0.292	−0.097	0.576	0.142	0.855	0.204
BODYmedium	21	**0.211**	**2.829**	**0.001**	**0.105**	**1.698**	**0.048**
BODYlarge	16	**0.368**	**−2.259**	**0.989**	0.149	0.070	0.475
Response							
Pd+ on phylogeny (Figure 2)	20	**0.212**	**2.914**	**0.001**	0.111	1.555	0.058
Pd+ on 'eco' tree (Figure 3)	20	7.839	0.862	0.201	4.356	0.277	0.384

[a] = species using both types of shelters are also included in the previous categories.

Phylogenetic signal of explanatory variables on a phylogeny of bats from Europe and North America and of *P. destructans* infection on both phylogeny and a neighbour-joining tree based on squared Euclidean distances of ecological and behavioural traits of hibernating bat species. Values in bold indicate significant clustering or over-dispersion of *P. destructans* infection on the tree. Note that all categories of explanatory variables were tested here, but *n* - 1 dummy variables were included in the PGLS model. *N* = number of species scored positive for the given variable, MPD = mean phylogenetic distance, NRI = net relatedness index, MNTD = mean nearest taxon phylogenetic distance, NTI = nearest taxon index.

doi:10.1371/journal.pone.0097224.t002

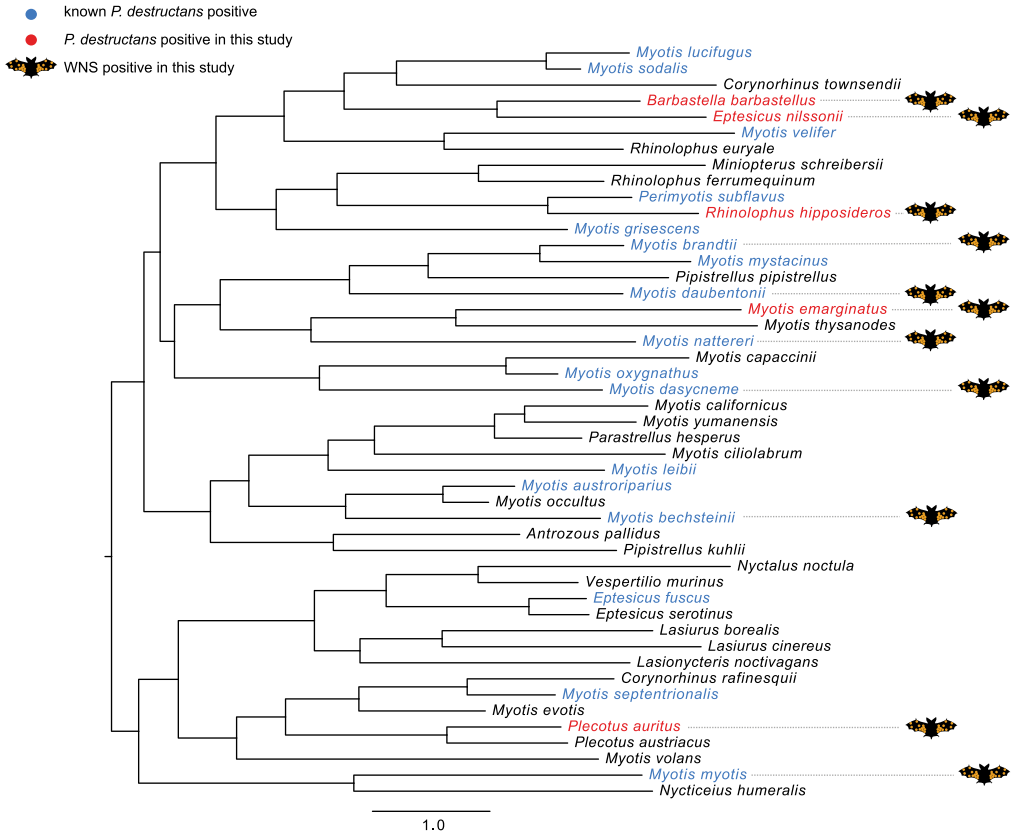


Figure 3. Neighbour-joining tree based on ecological and behavioural traits of bats from Europe and North America (rooted at midpoint). See Figure 2 for a description of the colour scheme.
doi:10.1371/journal.pone.0097224.g003

Testing the hypothesis of ecological similarity. The ecological similarity tree for European and North America bats was constructed using the squared Euclidean distances of the traits dataset (Figure 3). Analysis of *P. destructans*-infected species distribution indicated that infected species were randomly distributed (MPD = 7.839, NRI = 0.862, $p = 0.201$) across the ecological diversity of bats from these two continents. The most ecologically similar species were also infected randomly (MNTD = 4.356, NTI = 0.277, $p = 0.384$; Table 2, Figure 3).

Predicting species at risk from *P. destructans* infection. We explored relationships between ecological traits after removing the effects of bat species relatedness using PGLS. The final model, displaying lowest AIC (AIC = 69.08, *F*-statistic = 8.98, df = 7 and 39, $p < 0.001$, adjusted $R^2 = 0.55$), differed from two more complex models by $\Delta AIC < 3$ (Table S3). The addition of the variables did not markedly alter results of the analysis as reported below, and therefore we used the model with the lowest AIC. It describes the relationship between *P. destructans* infection in bat species and Temperature preference during hibernation ($\beta = -0.207$, SE = 0.085); Roost Shelter during hibernation: Hidden ($\beta = -0.109$, SE = 0.211), Exposed ($\beta = 0.560$, SE = 0.192); Cluster Size during hibernation: Small ($\beta = -0.386$, SE = 0.130), Medium ($\beta = -0.309$, SE = 0.194); Distribution range size: Moderate ($\beta = -0.201$, SE = 0.089); and Feeding habitat: Closed ($\beta = 0.319$, SE = 0.113). Shapiro-Wilks' normality test indicated that model residuals were normally distributed (W = 0.981, $p = 0.63$).

The fitted values of *P. destructans* infection in bats based on the PGLS model showed overlap between the *P. destructans*-positive and -negative bats (Figure 4). Bat species currently recognised as *P. destructans*-negative with highest fitted PGLS values were (in descending order): *Corynorhinus townsendii*, *Lasiurus cinereus*, *Plecotus austriacus*, *Rhinolophus ferrumequinum*, and *Miniopterus schreibersii*.

Discussion

At the end of the hibernation seasons of 2012 and 2013, we screened 276 bats of unknown infection status and biopsied all bats with WNS-suspected skin lesions. Both the number of bats and the number of taxa ($n = 15$) examined make this the most extensive

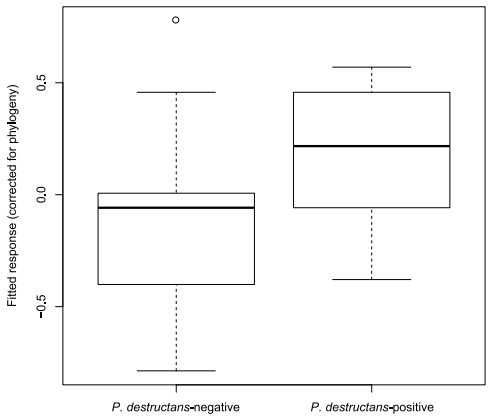


Figure 4. Boxplot of fitted values from the phylogenetic generalised least squares model of *P. destructans* infection. Predictions for infected and non-infected species overlap.
doi:10.1371/journal.pone.0097224.g004

and species-rich study of *P. destructans* infection in Europe to date. Earlier European studies have provided data from bats originally sampled for fungal microscopy, culture and genetic analysis when they exhibited obvious fungal growth during hibernation, numbers of species examined ranging from 1 to 12 and numbers of specimens from 1 to 107 [14,15,19–22].

This study documented five additional bat species as positive for *P. destructans* infection and added ten species from the genera *Myotis*, *Eptesicus*, *Plecotus*, *Barbastella* and *Rhinolophus* to the list of European bat species showing histopathological findings consistent with WNS [23]. Species-specific prevalence of WNS-diagnostic skin lesions ranged from 0 to 100% (4–55% for species with $n > 20$; Table 1). Highest prevalence of WNS lesions was observed in *M. myotis* (after excluding species for which just one specimen was caught, i.e. *M. dasycneme*). Note, however, that with prevalence differing by an order of magnitude, detection of positive specimens may have been biased by small sample size in rarer species and species less frequently visible in hibernacula. With our experimental design (i.e. trapping of winter survivors leaving hibernacula and non-destructive UV fluorescence screening), we were able to confirm that *M. myotis* had highest prevalence of WNS lesions (based on histopathology), with the possible exception of *M. dasycneme*. Our results confirmed WNS skin lesions in 11 bat species, which is in contrast to a previous study that found no invasive growth of *P. destructans* in European bats [38]. The latter study, however, examined a low number of bats and used skin biopsy methods lacking in sensitivity and this may explain the failure to detect lesions diagnostic of WNS.

Bat species with *P. destructans* infection exhibited diverse hibernation behaviours. Among these, *R. hipposideros* was special as it is an exclusively solitary hibernator and hangs free in exposed places with the wing membranes covering the body. It is capable of hibernation at higher temperatures but requires the microclimate stability ensured by using the inner parts of caves [39]. *Rhinolophus hipposideros* is also the most abundant species in winter-monitoring counts, amounting to more than 50% of all bats registered in Czech hibernacula (unpublished data, Czech Bat Conservation Trust). Moreover, as a member of the suborder Pteropodiformes, it is phylogenetically distantly related to all other species infected by *P. destructans* (Figure 2). As documented by the low infection prevalence (3.57%), environmentally mediated indirect and density-dependent transmission probably does not result in higher risk for this rhinolophid species [10]. Similarly, *E. fuscus* and *M. leibii*, both solitary hibernators from North America, were the least impacted species. In comparison, higher declines were observed in large winter colonies of two species that roost solitarily or in small groups, i.e. *P. subflavus* and *M. septentrionalis* [10]. Disease risk, however, was not related to conspecific transmission only. When multiple co-occurring species can host the pathogen, density-dependent transmission can be amplified [40,41].

We confirm here that bat species previously known to be positive for *P. destructans* may later show as WNS-positive based on histopathology. As the lists of bats positive for *P. destructans* or WNS lesions in North American and European species are nearly equal, conclusions drawn from analysis of infection or disease risk should be similar. Traits describing ecological and behavioural characteristics of bats occurring within the known distribution of *P. destructans* indicate that species belonging to additional genera may also be found positive for the infection in the future. Our screening of Czech underground hibernacula, however, demonstrates that the initial, relatively low, number of bat species positive for *P. destructans* infection is more likely the result of sampling bias than a biological phenomenon. Currently, affected species are ecologically diverse, to the point where predictions for infected and non-infected species overlap. We therefore assume that more species will be revealed as WNS-positive with increased sensitivity of detection methods [42].

Phylogenetic representation of *P. destructans* infection indicates that closely related species are most likely to be infected. In terms of field surveys, therefore, some *Myotis* bats are universally likely to be WNS-positive and most effort should be devoted to these species. Interestingly, *M. alcathoe* was free of WNS-positive skin lesions in the present study (but note the low sample size). *Myotis* species typically form clusters during hibernation and such behaviour promotes frequency-dependent transmission of the infection, independent of population size, and may yet drive the species to extinction [9] unless they change their social behaviour, as documented in *M. lucifugus* and *M. sodalis* [10]. In light of our new data, clustering behaviour is not a descriptor common to infected species. Rather, a suite of characteristics, including cluster size, type of shelter during hibernation, temperature at hibernation, as well as size of the distribution range and feeding in closed habitats, play a role in characterising bats with *P. destructans* infection.

Based on the list of species currently known to be affected by WNS or *P. destructans*, it is clear that the fungus is neither species-, genus- nor family-specific. The multi-host occurrence of the pathogen might make the disease less predictable using ecologically- and phylogenetically-based analysis [43]; however, this is likely to change in the future as additional species are revealed as susceptible. Hibernation in contaminated caves and mines under conditions favourable for fungal growth [5,44] appears to be the main risk factor [4,12]. Distribution of *P. destructans* is also correlated with disease in hibernating bats [45]. Importantly, 25 species of insectivorous bats presently hibernate in the United States and Canada, all of which represent possible hosts of the fungal pathogen should the disease spread to their geographic range [46]. This scenario is predicted to happen in most counties with caves in the contiguous United States by the winter of 2105-2106 [47].

The reason bats in North America have been so hard-hit, with millions dying, while bats in Europe apparently cope better with the infection, has not yet been explained. Likewise, the pathogen-

esis of WNS still remains unclear [12]. Behavioural aberrations, physiological disruption and immunosuppression during hibernation are, however, considered key pathomechanisms [4,48,49]. On the other hand, restoration of immune responses in WNS-positive bats early in post-hibernation may result in immune-mediated destruction of infected tissues and death [50].

The fact that our samples were mostly collected from bats emerging from hibernacula at the end of the hibernation season indicates that European bat species can survive *P. destructans* infection and highlights the need for a comparison of European and North American bat population responses to this fungal disease. As all European bat species are strictly protected and any thorough pathological study of *P. destructans* infection would be controversial [18,22], implementation of non-lethal sampling methods is necessary, such as the wing membrane biopsy used in this study [42]. While a detailed comparison of histopathological findings in European and North America bats represents a valid approach to the better understanding of WNS mortality [17,48], it was outside the objectives of this ecological study. We are however, planning a comprehensive study to investigate the extent of WNS wing lesions in hibernating bats from the two continents.

Conclusions

This hypothesis-driven study explored clustering of *P. destructans* infection in relation to chiropteran ecological and behavioural trait variation and phylogeny and supported this with field data. Extension of the surveillance to a broader number of species to test the study's hypotheses identified multiple European species positive for WNS. The increased number of positive bat species resulted in random dispersion of *P. destructans* infection across trait trees and weakened the pattern of phylogenetic clustering of *P. destructans* infection. Distantly related bat species, characterised by diverse life histories, were infected and all hibernating bats may, therefore, be at risk from *P. destructans* infection. Ecological and evolutionary constraints on hibernating bats do not pose a barrier to this generalist fungal pathogen, with WNS occurring in both suborders of bats. Our findings indicate that a wider focus is needed in studying the ecology and epidemiology of this fungal disease of major conservation concern.

Supporting Information

Table S1 Ecological and behavioural traits of European and North American bat species. The dataset includes the most common behavioural or trait value for each bat species.
(XLSX)

Table S2 Accession numbers for phylogeny reconstruction. A total of 13 available mitochondrial and nuclear genes were used for phylogeny reconstruction of bat species from Europe and North America.
(XLSX)

Table S3 Phylogenetic generalized least squares model selection. The model predicting *P. destructans* infection based on ecological and behavioural characteristics of bats was selected with the step-down procedure, where the full model is given on the first line and removed variables are listed subsequently.
(PDF)

Acknowledgments

The authors would like to thank Kevin Roche for his correction of the English text. Comments from Justin Boyles and two anonymous reviewers helped to improve the manuscript.

Author Contributions

Conceived and designed the experiments: JZ H. Bandouchova NM JP. Performed the experiments: H. Bandouchova JP JZ TB H. Berkova JB MD KSJ VK MK KO ZR. Analyzed the data: JZ MD NM KSJ VB GGT. Contributed reagents/materials/analysis tools: JZ H. Bandouchova NM JP GGT. Wrote the paper: JZ H. Bandouchova NM JP. Compiled the ecological and behavioural traits of bats: JZ. Reconstructed the phylogeny of European and North American bats: MD NM. Performed the statistical analysis: JZ KSJ NM. Examined the histopathological findings: H. Bandouchova JP. Collected field data: JZ H. Bandouchova TB H. Berkova JB MD KSJ VK MK KO ZR. Cultured and diagnosed the fungus: JB VK KO. Reviewed compiled data on North American bats: VB GGT. Provided a new method to detect WNS lesions: GGT.

References

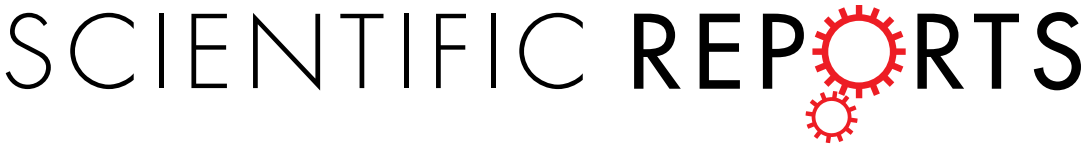
1. Irwin NR, Bayerlová M, Missa O, Martínková N (2012) Complex patterns of host switching in New World arenaviruses. Mol Ecol 21: 4137–4150.
2. Boyles JG, Cryan PM, McCracken GF, Kunz TH (2011) Economic importance of bats in agriculture. Science 332: 41–42.
3. Eskew EA, Todd BD (2013) Parallels in amphibian and bat declines from pathogenic fungi. Emerg Infect Dis doi: http://dx.doi.org/10.3201/eid1903.120707.
4. Blehert DS, Hicks AC, Behr M, Meteyer CU, Berlowski-Zier BM, et al. (2009) Bat white-nose syndrome: An emerging fungal pathogen? Science 323: 227–227.
5. Gargas A, Trest MT, Christensen M, Volk TJ, Blehert DS (2009) *Geomyces destructans* sp nov associated with bat white-nose syndrome. Mycotaxon 108: 147–154.
6. Lorch JM, Meteyer CU, Behr MJ, Boyles JG, Cryan PM, et al. (2011) Experimental infection of bats with *Geomyces destructans* causes white-nose syndrome. Nature 480: 376–378.
7. Minnis AM, Lindner DL (2013) Phylogenetic evaluation of *Geomyces* and allies reveals no close relatives of *Pseudogymnoascus destructans*, comb. nov., in bat hibernacula of eastern North America. Fungal Biol 117: 638–649.
8. Warnecke L, Turner JM, Bollinger TK, Lorch JM, Misra V, et al. (2012) Inoculation of bats with European *Geomyces destructans* supports the novel pathogen hypothesis for the origin of white-nose syndrome. Proc Natl Acad Sci USA 109: 6999–7003.
9. Frick WF, Pollock JF, Hicks AC, Langwig KE, Reynolds DS, et al. (2010) An emerging disease causes regional population collapse of a common North American bat species. Science 329: 679–682.
10. Langwig KE, Frick WF, Bried JT, Hicks AC, Kunz TH (2012) Sociality, density-dependence and microclimates determine the persistence of populations suffering from a novel fungal disease, white-nose syndrome. Ecol Lett 15: 1050–1057.
11. Turner GG, Reeder DM (2009) Update of white-nose syndrome in bats, September 2009. Bat Research News 50(3): 47–53.
12. Blehert DS (2012) Fungal disease and the developing story of bat white-nose syndrome. PLoS Pathog 8: e1002779.
13. Wilder AP, Frick WF, Langwig KE, Kunz TH (2011) Risk factors associated with mortality from white-nose syndrome among hibernating bat colonies. Biol Lett 7: 950–953.
14. Martínková N, Bačkor P, Bartonička T, Blažková P, Červený J, et al. (2010) Increasing incidence of *Geomyces destructans* fungus in bats from the Czech Republic and Slovakia. PLoS One 5: e13853.
15. Puechmaille SJ, Wibbelt G, Korn V, Fuller H, Forget F, et al. (2011) Pan-European distribution of white-nose syndrome fungus (*Geomyces destructans*) not associated with mass mortality. PLoS One 6: e19167.
16. Foley J, Clifford D, Castle K, Cryan P, Ostfeld RS (2010) Investigating and managing the rapid emergence of white-nose syndrome, a novel, fatal, infectious disease of hibernating bats. Conserv Biol 25: 223–231.
17. Puechmaille SJ, Frick WF, Kunz TH, Racey PA, Voigt CC, et al. (2011) White-nose syndrome: is this emerging disease a threat to European bats? Trends Ecol Evol 26: 570–576.
18. Pikula J, Bandouchova H, Novotny L, Meteyer CU, Zukal J, et al. (2012) Histopathology confirms white-nose syndrome in bats in Europe. J Wildl Dis 48: 207–211.
19. Kubátová A, Koukol O, Nováková A (2011) *Geomyces destructans*, phenotypic features of some Czech isolates. Czech Mycol 63: 65–75.
20. Puechmaille SJ, Verdeyroux P, Fuller H, Gouilh MA, Bekaert M, et al. (2010) White-nose syndrome fungus (*Geomyces destructans*) in bat, France. Emerg Infect Dis 16: 290–293.

21. Šimonovičová A, Pangallo D, Chovanová K, Lehotská B (2011) *Geomyces destructans* associated with bat disease WNS detected in Slovakia. *Biologia* 66: 562–564.
22. Wibbelt G, Kurth A, Hellmann D, Weishaar M, Barlow A, et al. (2010) White-nose syndrome fungus (*Geomyces destructans*) in bats, Europe. Emerg Infect Dis 16: 1237–1243.
23. Meteyer CU, Buckles EL, Blehert DS, Hicks AC, Green DE, et al. (2009) Histopathologic criteria to confirm white-nose syndrome in bats. J Vet Diagn Invest 21: 411–414.
24. Muller LK, Lorch JM, Lindner DL, O'Connor M, Gargas A, et al. (2013) Bat white-nose syndrome: a real-time TaqMan polymerase chain reaction test targeting the intergenic spacer region of *Geomyces destructans*. Mycologia 105: 253–259.
25. Lorch JM, Gargas A, Meteyer CU, Berlowski-Zier BM, Green DE, et al. (2010) Rapid polymerase chain reaction diagnosis of white-nose syndrome in bats. J Vet Diagn Invest 22: 224–230.
26. Simmons NB (2005) Chiroptera. In: Wilson DE, Reeder DM, editors. Mammal species of the world: A taxonomic and geographic reference. Smithsonian series in comparative evolutionary biology, 3rd edition. Washington DC: Smithsonian Inst. Press. pp. 312–529.
27. Webb PI, Speakman JR, Racey PA (1996) How hot is a hibernaculum? A review of the temperatures at which bats hibernate. Can J Zool 74: 761–765.
28. Horáček I, Hanák V, Gaisler J (2000) Bats of Palearctic region. In: Woloszyn BW, editor. Proceedings of the VIIIth European bat research symposium: 23–27 August 1999. Krakow, Poland: Institute of Systematics and Evolution of Animals PAS. pp. 11–157.
29. Reichard JD, Kunz TH (2009) White-nose syndrome inflicts lasting injuries to the wings of little brown myotis (*Myotis lucifugus*). Acta Chiropt 11: 457–464.
30. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.
31. R Core Team: R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available: http://www.R-project.org. Accessed 2013 Oct 10.
32. Webb CO, Ackerly DD, Kembel SW (2008) Phylocom: software for the analysis of phylogenetic community structure and character evolution. Bioinformatics 24: 2098–2100.
33. Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29: 1695–1701.
34. Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A (2010) How many bootstrap replicates are necessary? J Comput Biol 17: 337–354.
35. Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125: 1–15.
36. Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, et al. (2013) Caper: Comparative analyses of phylogenetics and evolution in R. R package version 0.5.2. Available: http://CRAN.R-project.org/package = caper. Accessed 2014 Feb 24.
37. Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W (2008) GEIGER: investigating evolutionary radiations. Bioinformatics 24: 129–131.
38. Wibbelt G, Puechmaille SJ, Ohlendorf B, Mühldorfer K, Bosch T, et al. (2013) Skin lesions in European hibernating bats associated with *Geomyces destructans*, the etiologic agent of white-nose syndrome. PLoS ONE 8: e74105.
39. Zukal J, Berková H, Řehák Z (2005) Activity and shelter selection by *Myotis myotis* and *Rhinolophus hipposideros* hibernating in the Kateřinská cave (Czech Republic). Mamm Biol 70: 271–281.
40. Keesing F, Holt RD, Ostfeld RS (2006) Effects of species diversity on disease risk. Ecol Lett 9: 485–498.
41. Ostfeld RS, Keesing F (2012) Effects of host diversity on infectious disease. Annu Rev Ecol Evol Syst 43: 157–182.
42. Turner GG, Meteyer CU, Barton H, Gumbs JF, Reeder DM, et al. (2014) Nonlethal screening of bat-wing skin with the use of ultraviolet fluorescence to detect lesions indicative of white-nose syndrome. J Wildl Dis, doi: 10.7589/2014-03-058.
43. Woolhouse MEJ, Taylor LH, Haydon DT (2001) Population biology of multihost pathogens. Science 292: 1109–1112.
44. Verant ML, Boyles JG, Waldrep W, Wibbelt G, Blehert DS (2012) Temperature-dependent growth of *Geomyces destructans*, the fungus that causes bat white-nose syndrome. PLoS ONE 7: e46280.
45. Lorch JM, Muller LK, Russell RE, O'Connor M, Lindner DL, et al. (2013) Distribution and environmental persistence of the causative agent of white-nose syndrome, *Geomyces destructans*, in bat hibernacula of the Eastern United States. Appl Environ Microbiol 79: 1293–1301.
46. U.S. Geological Survey (2012) White-Nose Syndrome Threatens the Survival of Hibernating Bats in North America. Available: http://www.fort.usgs.gov/WNS/Default.asp. Accessed 2014 Jan 05.
47. Maher SP, Kramer AM, Pulliam JT, Zokan MA, Bowden SE, et al. (2012) Spread of white-nose syndrome on a network regulated by geography and climate. Nat Commun 3: 1306.
48. Cryan PM, Meteyer CU, Boyles JG, Blehert DS (2010) Wing pathology of white-nose syndrome in bats suggests life-threatening disruption of physiology. BMC Biology 8: 135.
49. Cryan PM, Meteyer CU, Blehert DS, Lorch JM, Reeder DM, et al. (2013) Electrolyte depletion in white-nose syndrome bats. J Wildl Dis 49: 398–402.
50. Meteyer CU, Barber D, Mandl JN (2012) Pathology in euthermic bats with white nose syndrome suggests a natural manifestation of immune reconstitution inflammatory syndrome. Virulence 3: 583–588.

Paper 2.4.2

Zukal J., Bandouchova H., Brichta J., Cmokova A., Jaron K. S., Kolarik M., Kovacova V., Kubátová A., Nováková A., Orlov O., Pikula J., Presetnik P., Šuba J., Zahradníková A. Jr., **Martínková N.** 2016. White-nose syndrome without borders: *Pseudogymnoascus destructans* infection confirmed in Asia. *Scientific Reports* 6: 19829.

www.nature.com/scientificreports



OPEN

Received: 08 July 2015
Accepted: 15 December 2015
Published: 29 January 2016

White-nose syndrome without borders: *Pseudogymnoascus destructans* infection tolerated in Europe and Palearctic Asia but not in North America

Jan Zukal[1,2,*], Hana Bandouchova[3,*], Jiri Brichta[3], Adela Cmokova[4], Kamil S. Jaron[1], Miroslav Kolarik[4], Veronika Kovacova[3], Alena Kubátová[5], Alena Nováková[4], Oleg Orlov[6,7], Jiri Pikula[3], Primož Presetnik[8], Jurģis Šuba[9], Alexandra Zahradníková Jr.[10] & Natália Martínková[1,11]

A striking feature of white-nose syndrome, a fungal infection of hibernating bats, is the difference in infection outcome between North America and Europe. Here we show high WNS prevalence both in Europe and on the West Siberian Plain in Asia. Palearctic bat communities tolerate similar fungal loads of *Pseudogymnoascus destructans* infection as their Nearctic counterparts and histopathology indicates equal focal skin tissue invasiveness pathognomonic for WNS lesions. Fungal load positively correlates with disease intensity and it reaches highest values at intermediate latitudes. Prevalence and fungal load dynamics in Palearctic bats remained persistent and high between 2012 and 2014. Dominant haplotypes of five genes are widespread in North America, Europe and Asia, expanding the source region of white-nose syndrome to non-European hibernacula. Our data provides evidence for both endemicity and tolerance to this persistent virulent fungus in the Palearctic, suggesting that host-pathogen interaction equilibrium has been established.

Emerging wildlife infections are a threat to global biodiversity[1], yet the emergence and transmission of such infectious diseases may also be a function of ecosystem quality and biodiversity[2]. Moreover, anthropogenic disturbance and introductions have been highlighted as contributing considerably to such disease outbreaks[3,4]. The majority of emerging wildlife pathogens are of viral origin but fungal infections have also been recognised across a diverse range of taxa, including plants and poikilothermic animals[5]. Relatively few fungal species cause severe diseases in mammals due to their high body temperature and effective immune system[6,7].

Prevalence of pathogens may be high in bat populations[8,9]. Despite this, bats have been reported to survive infections that are lethal to other taxa[10–12]. Bat's capacity to carry pathogens may result from its ability to tolerate damage caused by the agent or the associated immune response[13–15]. Understanding the mechanisms that

[1]Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, Květná 8, 603 65 Brno, Czech Republic. [2]Department of Botany and Zoology, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic. [3]Department of Ecology and Diseases of Game, Fish and Bees, University of Veterinary and Pharmaceutical Sciences Brno, Palackeho 1/3, 612 42 Brno, Czech Republic. [4]Laboratory of Fungal Genetics and Metabolism, Institute of Microbiology, Academy of Sciences of the Czech Republic, Vídeňská 1083, 142 20 Prague 4, Czech Republic. [5]Department of Botany, Faculty of Science, Charles University in Prague, Benátská 2, 128 01 Prague, Czech Republic. [6]Ural State Pedagogical University, Kosmonavtov str. 26, 620017 Yekaterinburg, Russia. [7]Ural State Medical University, Repina str. 3, 620028 Yekaterinburg, Russia. [8]Centre for Cartography of Fauna and Flora, Antoličičeva 1, SI-2204 Miklavž na Dravskem polju, Slovenia. [9]Latvian State Forest Research Institute "Silava", 111 Rigas str., LV-2169 Salaspils, Latvia. [10]Institute of Molecular Physiology and Genetics, Slovak Academy of Sciences, Vlárska 5, 83334 Bratislava, Slovakia. [11]Institute of Biostatistics and Analysis, Masaryk University, Kamenice 3, 625 00 Brno, Czech Republic. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.P. (email: pikulaj@vfu.cz) or N.M. (email: martinkova@ivb.cz)

underlie any trade-off between tolerance and resistance to infectious agents in reservoir hosts is of utmost importance for effective disease control[16].

Most bats in temperate regions enter seasonal hibernation, at which time they reduce their metabolic rate and decrease body temperature until it approaches the ambient temperature of the hibernaculum, which may alter immune response to pathogens[17,18]. During hibernation, bats are vulnerable to infection by the psychrophilic fungus *Pseudogymnoascus destructans* [formerly *Geomyces destructans*][19], which emerged as a novel pathogen in eastern North America in 2006[20]. The *P. destructans* fungus is the causative agent of white-nose syndrome (WNS)[20–23]. Considered an epidemic of major conservation concern in North America, WNS is responsible for an unprecedented decline in bat populations[24–26]. WNS combines some of the worst possible epidemiological characteristics, including a highly virulent pathogen[23,27] with density- and frequency-dependent transmission[28], an environmental reservoir[19], long-term persistence in hibernacula[29,30] and susceptibility of multiple hosts[19,20,31].

The origin of WNS in North America remains unknown and current research focuses on identifying the infectious agent's source and identifying whether the agent is an introduced pathogen. Six main lines of evidence support introduction of WNS into North America: 1) Fungal communities associated with bats and their hibernacula in eastern North America comprise a diverse range of *P. destructans* allies[22,32]; phylogenetic evaluation, however, indicates that none of these is closely related to *P. destructans*[22]. 2) Previous studies have demonstrated clonal dispersal of a single *P. destructans* genotype among WNS-infected bats in the United States[33,34]. 3) Spread of WNS in North America follows a clear invasion front with varying survival rate since first appearance of the disease[24,35–37]. 4) Presence of *P. destructans* has been confirmed in many European countries[38–41] but with no reports of mass mortality[38]. 5) European *P. destructans* isolates are pathogenic for North American bats[23]. Further, virulent skin infections producing focally severe lesions pathognomonic for WNS have been documented in European bats under natural infection conditions[42]. 6) While only one heterothallic fungal mating type has been recorded in North American *P. destructans* populations, two types have been found coexisting in European hibernacula. Effective recombination during sexual reproduction results in genetic variability and may be linked with virulence[43]. These indirect sources of evidence tend to support the introduced pathogen hypothesis, suggesting WNS may have originated outside of North America[22,23,43,44]. Comparative studies between North America and Europe have been proposed to explain the origin and differential manifestation of the fungal infection[4,40,42,45].

Although some authors speculate on a non-European origin[4,25], Europe is believed to be the likely source of WNS[44]. Supporting evidence for long exposure to the WNS fungus based on old European photographs, which appear to show a white cutaneous fungal infection[38,40], lacks specificity for WNS. Another dermatophyte, *Trichophyton redellii*, produces a similar gross appearance in hibernating bats[46] and the photographs might represent either. On the contrary, presence of WNS in multiple and diverse hosts[31] indicates that the WNS fungus could persist across their ranges in the Palearctic.

To date, WNS diagnostic skin lesions have only been reported from the Czech Republic, Europe, with almost half of all species being WNS positive and prevalence based on histopathology reaching 55% in bats emerging from hibernacula in spring[31,42]. Occurrence of WNS in distantly related bat species with diverse ecologies, however, suggests the pathogen is a generalist and that all bats hibernating within the distribution range of *P. destructans* may be at risk of infection[31]. If we assume that *P. destructans* range expansion and its concurrent WNS epidemic occurred unnoticed in Europe in the past, we expect that WNS should be found at any site in Europe with conditions favourable for the pathogen and could extend to non-European hibernacula of the Palearctic region. We tested the following four hypotheses. First, *P. destructans* is present in European and non-European Palearctic hibernacula and, if the fungus is present in a hibernaculum, bats would test positive for WNS under histopathology. Second, an endemic steady-state of *P. destructans* infection in the Palearctic would be reflected in a persistent high prevalence among bats with no associated population declines. Third, given that the differences in population size changes in WNS-positive regions[24,38] are explained as WNS resistance in Europe, Palearctic bats will display a lower pathogen load and smaller size of cupping erosions diagnostic for WNS than Nearctic species. Fourth, in regions with endemic pathogen occurrence, *P. destructans* load will be influenced by geographic location, sampling date and bat community structure.

Results

WNS in the Palearctic. We sampled and examined bats from sites in geographically distant regions (the Czech Republic, Slovenia, Latvia and Russia; $n = 481$) to assess occurrence of WNS in the Palearctic (Table 1, Supplementary Table S1). We found WNS-positive bats from multiple species at all sites, including the West Siberian Plain in Russia. At each site, all of the following criteria categorised the site as WNS positive: suspected fungal growth (Fig. 1), characteristic curved conidia of *P. destructans* observed on an adhesive tape imprint (Fig. 2A), *P. destructans* confirmed by DNA sequencing and qPCR using a species-specific probe, and definitive diagnosis confirmed through histopathology of biopsy punches from wing membrane lesions targeted by UV trans-illumination (Fig. 2B).

Histopathological findings matching WNS diagnostic criteria were present in 100 bats of 13 different species (Table 1). Two species were newly identified as positive for WNS, *Miniopterus schreibersii* and *Rhinolophus euryale*. Species-specific prevalence, using qPCR to detect *P. destructans* and UV trans-illumination to detect lesions indicative of WNS, ranged from 64 to 100% (overall prevalence = 83.1%) and 14 to 100% (overall prevalence = 51.9%), respectively (Table 1).

In order to evaluate *P. destructans* genetic variability, we sequenced six loci resulting in 254 new sequences. Two alleles were found in *TUB2*, *MAT1-1-1*, *MAT1-2-1* and *CAM* loci, four alleles in *TEF1*α and three in *ITS* (Fig. 3). The most frequent alleles from all loci were found in both European and Asian *P. destructans* isolates. In *ITS* and *TEF1*α from Asia, we sampled isolates with one substitution difference from the most frequent allele. The dominant haplotype from the concatenated sequence of four genes (*TUB2*, *CAM*, *TEF1*α, *ITS*) is widespread in North America, Europe and Asia (Fig. 3E). Three divergent concatenated haplotypes were found in the Czech

Region	Bat species	Screened	WNS qPCR+		WNS UV+		WNS histo+	
		n	*n*	qPCR+ (%) ± s.e.	*n*	UV+ (%) ± s.e.	*n*	histo+ (%) ± s.e.
Slovenia	*Miniopterus schreibersii*	21	20	100 ± 7.1	21	47.62 ± 10.9	6	16.67 ± 20.05
	Myotis emarginatus	3	0	NA ± NA	3	100 ± 32.22	3	100 ± 32.22
	Myotis myotis	11	10	100 ± 13.21	11	63.64 ± 12.16	4	0 ± 26.89
	Rhinolophus euryale	14	14	64.29 ± 9.83	14	28.57 ± 9.83	1	100 ± 48.47
Total		**49**	**44**	**88.1 ± 4.88**	**49**	**59.96 ± 7**	**14**	**54.17 ± 13.32**
Czech Republic	*Barbastella barbastellus*	18	17	64.71 ± 11.59	18	50 ± 11.79	4	75 ± 26.89
	Eptesicus nilssonii	3	2	50 ± 39.61	2	0 ± 39.61	1	100 ± 48.47
	Myotis alcathoe	7	7	71.43 ± 17.76	7	14.29 ± 17.76	6	0 ± 20.05
	Myotis bechsteinii	23	23	86.96 ± 6.23	23	39.13 ± 10.18	9	44.44 ± 14.45
	Myotis brandtii	17	16	68.75 ± 8.71	17	23.53 ± 8.24	1	100 ± 48.47
	Myotis dasycneme	1	1	100 ± 48.47	1	100 ± 48.47	1	100 ± 48.47
	Myotis daubentonii	33	30	86.67 ± 4.85	31	51.61 ± 8.98	14	35.71 ± 12.81
	Myotis emarginatus	33	32	96.88 ± 4.56	32	43.75 ± 8.77	5	60 ± 23
	Myotis myotis	156	150	99.33 ± 1.01	108	96.3 ± 1.4	61	73.77 ± 5.63
	Myotis nattereri	22	21	95.24 ± 6.78	22	54.55 ± 10.62	9	33.33 ± 14.45
	Plecotus auritus	23	23	91.3 ± 6.23	22	68.18 ± 9.93	11	54.55 ± 12.16
	Plecotus austriacus	3	3	66.67 ± 32.22	3	33.33 ± 32.22	1	0 ± 48.47
	Rhinolophus hipposideros	35	29	79.31 ± 7.52	35	34.29 ± 8.02	8	50 ± 15.94
Total		**374**	**354**	**81.33 ± 2.07**	**321**	**46.84 ± 2.79**	**131**	**55.91 ± 4.34**
Latvia	*Eptesicus nilssonii*	4	4	25 ± 26.89	4	25 ± 26.89	1	0 ± 48.47
	Myotis brandtii	4	4	50 ± 26.89	4	50 ± 26.89	0	NA ± NA
	Myotis dasycneme	8	8	100 ± 15.94	8	100 ± 15.94	6	50 ± 20.05
	Myotis daubentonii	9	9	100 ± 14.45	9	88.89 ± 14.45	6	66.67 ± 20.05
Total		**25**	**25**	**68.75 ± 9.27**	**25**	**65.97 ± 9.48**	**13**	**38.89 ± 13.52**
Russia	*Eptesicus nilssonii*	5	5	100 ± 23	5	100 ± 23	2	100 ± 39.61
	Myotis brandtii	9	9	100 ± 14.45	9	100 ± 14.45	2	0 ± 39.61
	Myotis dasycneme	17	17	100 ± 8.24	17	100 ± 8.24	12	75 ± 11.27
	Myotis daubentonii	1	1	100 ± 48.47	1	100 ± 48.47	0	NA ± NA
	Plecotus auritus	1	1	100 ± 48.47	1	100 ± 48.47	1	100 ± 48.47
Total		**33**	**33**	**100 ± 4.43**	**33**	**100 ± 4.43**	**17**	**68.75 ± 11.24**
USA	*Myotis lucifugus*	10	0	NA ± NA	10	100 ± 13.21	10	100 ± 13.21
Total		**10**	**0**	NA ± NA	**10**	**100 ± 13.21**	**10**	**100 ± 13.21**

Table 1. Prevalence of white-nose syndrome and *Pseudogymnoascus destructans* infection in bat species from the Holarctic region. Screened = number of bats captured and examined by PCR (to detect the pathogen), UV light trans-illumination or histopathology (to detect WNS lesions). + = percentage of positive bats from the number screened by the method. NA = not available. Samples subjected for histopathology examination were suspect lesions selected under field conditions based on UV trans-illumination and thus prevalence on histopathology is not based on a randomized sample. It rather reflects qualitative information that WNS was confirmed in the species with histopathology.

Republic and one local to Asian Russia. The *P. destructans* mating type proportion was 27 (*MAT1-1-1*) to 19 (*MAT1-2-1*) ($\chi^2 = 1.391$, $P = 0.238$).

Quantitative comparison of WNS on bats. The fungal load on qPCR-positive bats ranged from 0.21 pg to 3.41 μg across the surface of the left wing (Supplementary Fig. S1, see Table 1 for sample sizes). This range included fungal loads reported from Nearctic bat species (Supplementary Fig. S1). Log-transformed fungal load was not lower on Palearctic bats than on Nearctic bats (Wilcoxon $W = 68030$, $P = 1$, $n = 413$ and 247, respectively), even assuming that fungal load on Nearctic bats could vary up to ten-times due to the difference in sample collection method between continents (Wilcoxon $W = 50910$, $P = 0.484$).

In order to account for body size differences between Palearctic bat species, we used log-transformed fungal load per cm^2 of wing area (henceforth referred to as fungal load, unless specified otherwise). Fungal load differed significantly between regions (Kruskal-Wallis $\chi^2 = 94.2$, $P < 0.001$; Fig. 4) and species (Kruskal-Wallis $\chi^2 = 221.2$, $P < 0.001$; Fig. 5), with highest values recorded in the Czech Republic and two euryvalent bat species, *Myotis myotis* and *Myotis nattereri*, respectively. In frequently captured *M. myotis*, fungal load did not change between 2012 and 2014 (ANOVA: $F_{2,146} = 1.209$, $P = 0.301$). Prevalence dynamics of nine species captured in multiple years was statistically equal between years (χ^2 test: $P > 0.05$; Czech Republic). The sampled regions contained multiple species with varying fungal loads (Fig. 5), and some species were present in multiple regions (Table 1). The observed difference of fungal load can be caused by a different set of species or different environmental conditions

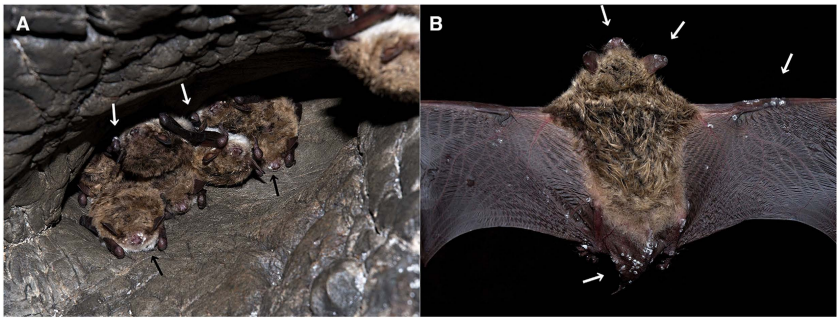


Figure 1. Fungal growth on hibernating bats from Russia. (**a**) A hibernating cluster of pond bats *Myotis dasycneme* in a cave near Yekaterinburg, Russia, in May 2014. Black and white arrows indicate fungal growth on the muzzle and forearm, respectively. (**b**) A pond bat from the same hibernaculum showing visible fungal growths on the uropatagium, pelvic limb toes, plagio- and pro-patagium and the ears and muzzle (white arrows). Photo: Jiri Pikula.

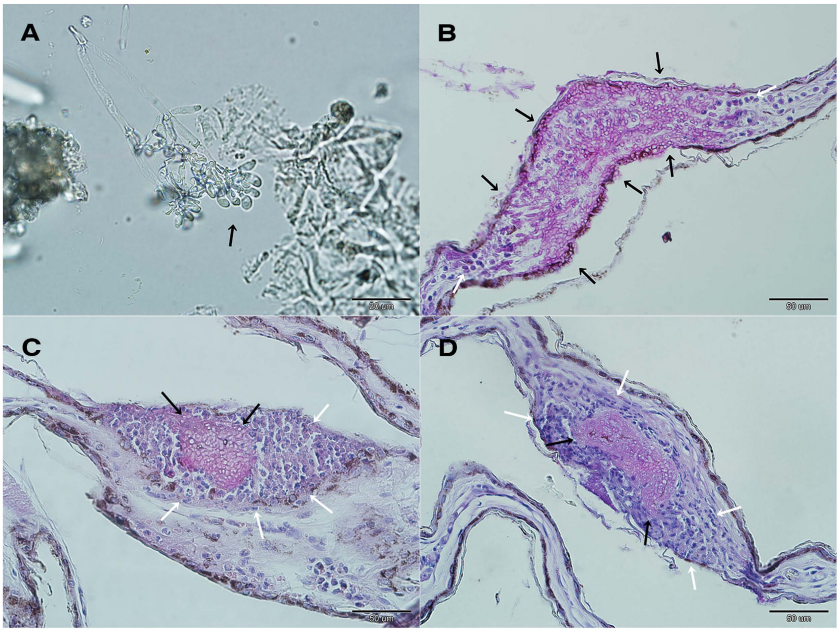


Figure 2. White-nose syndrome on a pond bat *Myotis dasycneme* near Yekaterinburg, Russia, in May 2014. (**A**) Microscopic identification of the characteristic curved conidia of *Pseudogymnoascus destructans* (black arrow). (**B**) Invasive fungal growth penetrating the full-thickness of the wing membrane (black arrows), with several inflammatory cells (neutrophils) situated at both margins of the lesion (white arrows). (**C**) Packed fungal hyphae of cupping erosions (black arrows) sequestered by neutrophils (white arrows). (**D**) Histopathological finding from a WNS-positive Nearctic *Myotis lucifugus*, identical to that found in Palearctic Asia. Skin sections stained with periodic acid-Schiff stain.

in sampled regions. Therefore, we separated the effect of the two variables on fungal load by correcting for the random effect of species and region. This had no impact on significance of the comparison between regions or species. The phylogenetic signal of mean species fungal load was significant (Blomberg's $K = 0.892$, $P = 0.002$), meaning that closely related taxa had more similar mean fungal loads in the Palearctic than expected by chance.

Fungal load was significantly correlated with the number of WNS lesions (Spearman's rank correlation: $r = 0.61$, $P < 0.001$). Phylogenetic generalised least-squares, which accounts for intra-specific variability and

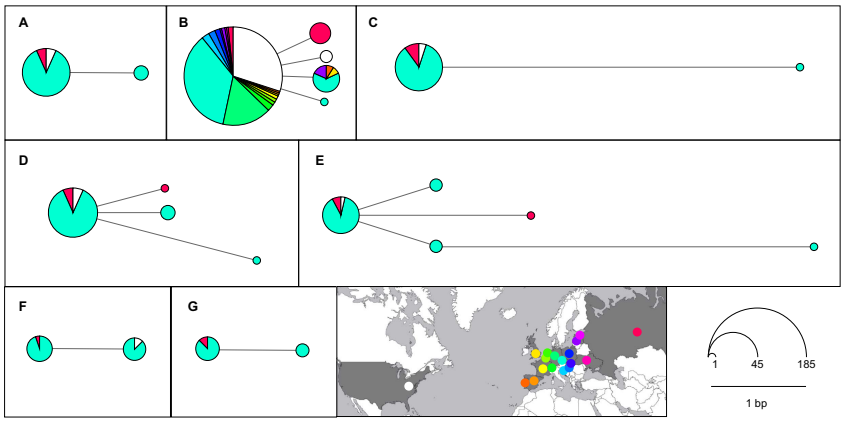


Figure 3. Median-joining networks for *Pseudogymnoascus destructans* DNA sequences. Node centroid distances correspond to the number of substitutions between haplotypes and the node size to the number of isolates sharing the haplotype. Isolate origin is signified through different colours based on the inset map. The map was modified from http://www.amcharts.com/visited_countries/ (last accessed on 9 October, 2015). (**A**) *TUB2* ($n = 46$), (**B**) *ITS* ($n = 203$), (**C**) *CAM* ($n = 47$), (**D**) *TEF1α* ($n = 51$), (**E**) sequences concatenated with *TUB2*, *ITS*, *CAM* and *TEF1α* ($n = 34$); haplotypes pooled according to the available sequence and gaps treated as missing data, (**F**) *MAT1-1-1* ($n = 27$), (**G**) *MAT1-2-1* ($n = 19$).

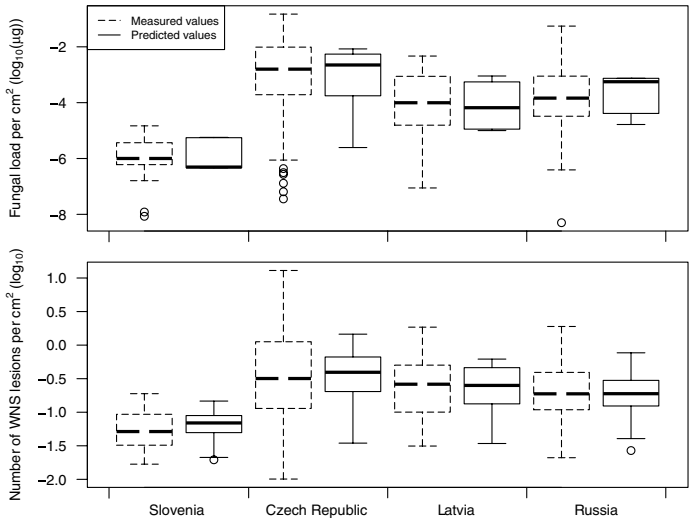


Figure 4. Fungal load and number of WNS lesions in Europe (Czech Republic, Latvia, Slovenia) and in Asia (Russia). Fungal load is quantified as *P. destructans*-specific DNA per cm^2 of wing area (established through quantitative PCR) and number of WNS lesions is counted on the same wing (using UV light trans-illumination).

species phylogenetic relationships, indicated that the number of WNS lesions increased with increasing fungal load across bat diversity in this study (intercept = –0.058, slope = 0.177; Fig. 6). No significant phylogenetic signal was detected in the mean number of WNS lesions by species (Blomberg's $K = 0.511$, $P = 0.073$).

Microscopic WNS cupping erosion width ranged between 28.6 and 397.2 µm (median = 86.34 µm, $n = 105$) and depth between 11.3 and 91.8 µm (median = 31.29 µm), this also being the WNS-lesion size range detectable photographically as individual spots using UV trans-illumination. Size of WNS lesion (log-transformed) did not differ significantly between species (Kruskal-Wallis $\chi^2 = 20.7$, $P = 0.079$, $n = 14$ species across the Holarctic;

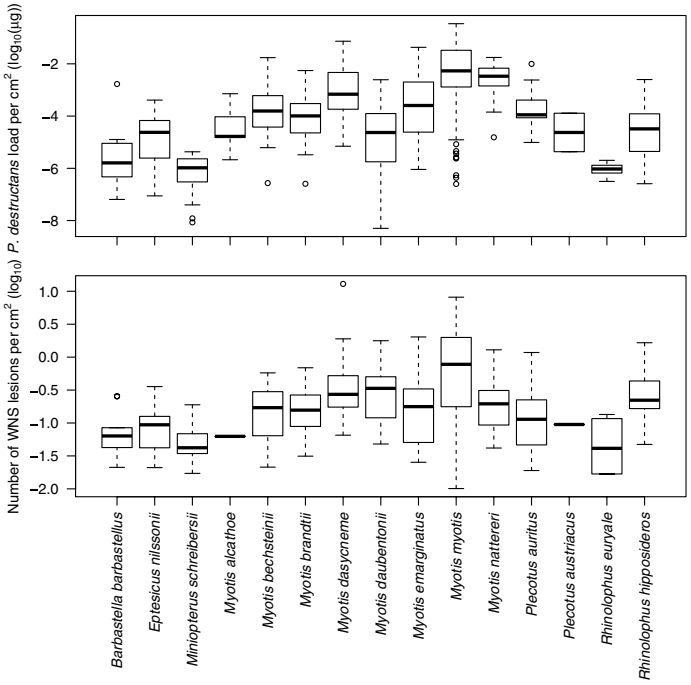


Figure 5. Fungal load and number of WNS lesions on bats from Europe (Czech Republic, Latvia, Slovenia) and Asia (Russia). Fungal load is quantified as *P. destructans*-specific DNA per cm^2 of wing area (established through quantitative PCR) and number of WNS lesions is counted on the same wing (using UV light trans-illumination).

Fig. 7, see Table 1 for sample sizes). Phylogenetic signal for mean species-specific WNS lesion size was also not significant (Blomberg's $K = 0.448$, $P = 0.177$) and there was no significant difference observed between Palearctic and Nearctic bats (phylogenetic ANOVA: $F = 0.43$, $P = 0.292$).

The number of WNS lesions in animals positive over UV ranged from 1 to 805 (median = 13) and differed significantly between Palearctic regions (log-transformed per cm^2; Kruskal-Wallis $\chi^2 = 33.6$, $P < 0.001$; Fig. 4) and species (Kruskal-Wallis $\chi^2 = 76.7$, $P < 0.001$; Fig. 5; correction did not influence significance). The fungal load from UV-negative individuals (median = 3.78×10^{-5} µg) overlapped that from UV-positive individuals (median = 7.46×10^{-4} µg; Fig. 8).

Comparison of AIC values indicated that the best fitting model for fungal load across regions included geographic coordinates (95% confidence interval of parameter estimates: longitude: –0.07–(–0.03), latitude: 0.16–0.36), sampling date (0.002–0.02) and structure of common bat species community found in the respective region (–3.86–(–2.24)) (Supplementary Table S2). On the other hand, disease intensity, quantified as number of WNS lesions, was best modelled using fungal load (0.18–0.28) and sampling date (–0.007–0.0003), though addition or removal of other variables did not change the relationship between WNS-lesion number and fungal load significantly (0.18–0.27). In both cases, species was treated as a random effect. Values predicted from the optimal model for both fungal load and number of WNS lesions were highest in the Czech Republic (Fig. 4). Lowest values were predicted in Slovenia and intermediate values in Latvia and Russia.

Discussion

Here we show WNS infection in Palearctic bat communities within and beyond the borders of Europe, greatly extending the distribution range of *P. destructans* and confirming its generalist nature. Furthermore, we confirm WNS skin lesions in two more bat species, belonging to families Miniopteridae and Rhinolophidae.

Histopathology of skin sections is presently considered the 'gold standard' for diagnosing WNS[47]. The microscopically identified WNS in bat species sampled in Russia was consistent with WNS histopathological criteria[42,47]. These included characteristic cup-shaped epidermal erosions, *P. destructans*-infected hair follicles, associated sebaceous glands and regional connective tissues, and invasive fungal growth throughout the wing-membrane thickness (Fig. 2). Our results demonstrate that *P. destructans* is virulent for Palearctic bats under natural infectious conditions, as previously shown in the Czech Republic[31,42]. Local invasiveness and WNS-lesion severity in wing membranes of bats collected in the Palearctic was comparable with that of Nearctic bats (Figs 2 and 8), with

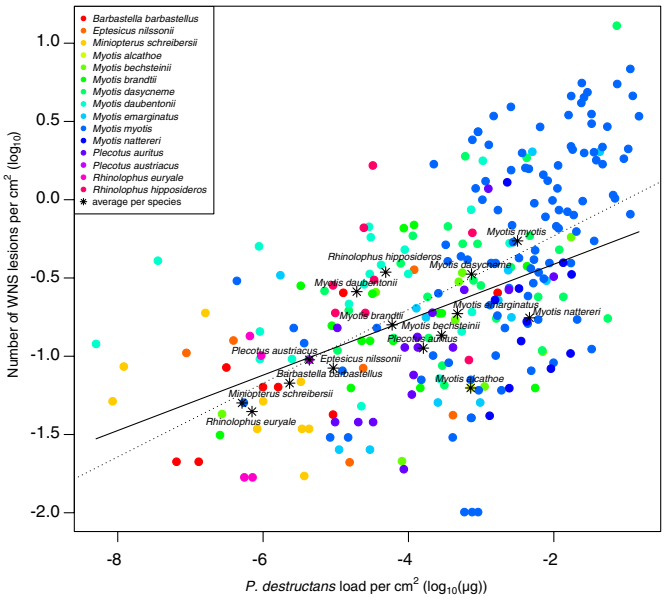


Figure 6. Phylogenetic generalised least-squares for number of WNS lesions, dependent on fungal load and accounting for within-species variation and relatedness. The dotted line represents the linear regression without phylogenetic correction.

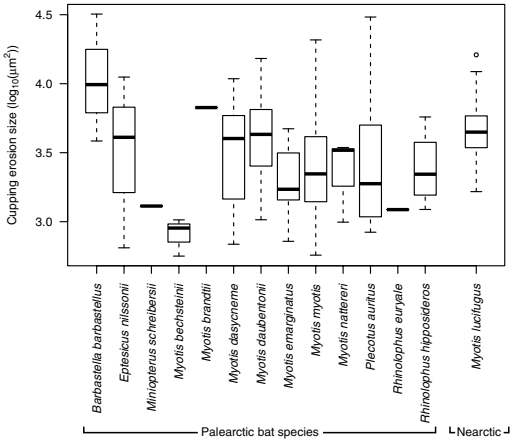


Figure 7. Variation in focal tissue invasiveness among hibernating bat species infected by *Pseudogymnoascus destructans*. A series of 80 periodic acid–Schiff stained sections from each bat's wing membrane biopsy were used to determine mean maximum size of WNS-diagnostic cupping erosions ($\log_{10}(\mu m^2)$). See Table 1 for sample sizes.

diagnostic features ranging from cupping erosions to full-thickness fungal invasion found on all continents. In our experience, detection of WNS-positive bats was greatly enhanced by the use of a UV lamp combined with non-lethal punch biopsies of suspected skin lesions[48]. The effectiveness of this method enabled us to identify seven species as WNS-positive in multiple regions, i.e. *Myotis brandtii*, *Myotis dasycneme*, *Myotis daubentonii*, *Myotis emarginatus*, *M. myotis*, *Eptesicus nilssonii* and *Plecotus auritus* (Table 1). As predicted[31], *M. schreibersii* and *R. euryale* were confirmed positive for skin lesions. These are thermophilic species that only hibernate in

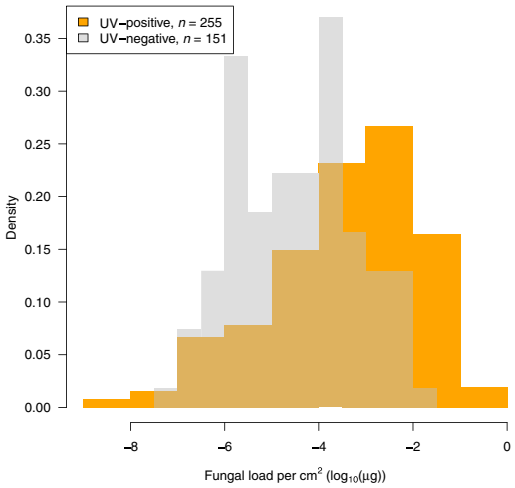


Figure 8. Frequency of *Pseudogymnoascus destructans* fungal load per cm² of wing area. Bats were identified as positive (grey; $n = 255$) and negative (orange; $n = 151$) using UV trans-illumination.

the northern part of their ranges. *Miniopterus schreibersii*, a long-distance migrant, forms large mixed colonies with other cave-dwelling bat species[49] and such characteristics make the species another effective candidate for pathogen dispersal.

Theoretically, WNS spread is limited by conditions prevalent in underground hibernacula being favourable to *P. destructans* and presence of susceptible hibernating bats[50]. *Pseudogymnoascus destructans*, being a generalist pathogen, could infect any bat species hibernating under the right microclimatic conditions and ecological and evolutionary differences in the hibernating bats would not pose a barrier[31]. Distribution ranges of multiple WNS-positive bat species (*E. nilssonii*, *M. dasycneme*, *M. daubentonii*, *M. brandtii*, *P. auritus*) extend across Palearctic Asia and overlap with their sister species or ecological counterparts (e.g. *Myotis petax*, *Myotis sibiricus*). As bats can form multi-species clusters at hibernation sites, there is a good chance that *P. destructans* could switch host species to presently WNS-negative taxa (e.g. *Myotis ikonnikovi*, *M. petax*, *M. sibiricus*). This suggests that WNS may be present throughout the Palearctic where suitable environmental conditions occur and in all bat species that utilise such sites. In fact, a paper published ahead of print during review of this article corroborates our conclusions with one *M. petax* found positive for WNS in North China[51].

Given the tragic impact of WNS on bat biodiversity in North America[24], it is important to identify the source of possible introduction[4]. However, *P. destructans* exhibits notoriously low genetic diversity in both the clonal North American populations[33,34,43] and the sexually reproducing populations of Europe[44], effectively thwarting efforts to pinpoint the source region. Our study further exacerbates the enigma of where the North American *P. destructans* strain originates as the North American haplotype was found in all non-mating gene types in Palearctic Asia, thereby expanding the putative source region (Fig. 3). Additionally, the search for a source region should shift to markers with a faster mutation rate and better resolution for *P. destructans* phylogeography, as the low genetic variability in markers routinely used in fungal population studies indicates a relatively recent origin or expansion of the Palearctic population.

Theory suggests that if a disease is detectable at high prevalence it is probably mild and unlikely to be a major problem[52]. Two methods were used to establish prevalence in this study: qPCR, used to quantify the pathogen on wing skin[53], and UV trans-illumination of the wing membrane, which enables detection of fluorescing lesions associated with WNS skin infection[48]. Both methods provided comparable results of high prevalence, and, together with stable or increasing host population sizes[54], our data strongly suggest endemicity of *P. destructans* within Palearctic bat populations. Persistent high prevalence and pathogen load with absence of population declines in the Palearctic are in sharp contrast to the situation in the Nearctic. In Midwestern United States, high prevalence of *P. destructans* infection was followed by a decrease in number of hibernating bats[55]. Some form of host-pathogen equilibrium, analogous to the amphibian chytrid fungus observed in post-decline frog communities[56], may already have occurred in the Palearctic but not in the Nearctic.

Pathogen load indicates both exposure to the infectious agent and suitability of the host for pathogen replication[57]. *Pseudogymnoascus destructans* fungal load can be evaluated using swab samples from wing membranes[53]. Although swabs do not sample fungus invasion deep in the wing tissue, our results show that fungal load sampled in this way is positively correlated with the number of WNS lesions (Fig. 6). Hence, fungal load from swabs may be assumed to approximate the total load associated with the skin infection. Fungal loads determined across the Palearctic were similar to those observed in Nearctic species (Supplementary Fig. S1). The swabbing technique used in the North American studies[19,58,59], however, does not allow standardisation of *P. destructans* load per cm²

and thus detailed statistical comparisons cannot be made. Nevertheless, fungal loads in Nearctic bats are similar to the observed fungal load of Palearctic bats (Supplementary Fig. S1), even if the former were underestimated ten-fold with the different swabbing technique.

In the Palearctic study area, fungal load differed regionally (Fig. 4) and between species (Fig. 5) with no reported mass decline in bat populations attributable to WNS[38,54]. Fungal load increased with increasing northing and westing, sampling day (load increasing later in the year) and in regions where phylogenetically closely-related hibernating bats predominated (Fig. 4). The driving factor that best modelled the number of WNS lesions observed using UV trans-illumination in the Palearctic was the fungal load. Increasing fungal load positively correlated with disease intensity across species diversity indicates that hyphae are more likely to invade deep tissues and cause WNS lesions with heavy fungal growth on bat wings (Fig. 5). However, with overlap of fungal loads in UV-negative and UV-positive bats, no clear fungal load threshold defines the pathogen pressure where development of WNS lesions starts.

With persistent and similar fungal load (Supplementary Fig. S1) and cupping erosion size (Fig. 7) on bats from the Palearctic and the Nearctic, the difference in population size response is striking. While population sizes dropped dramatically under the WNS epidemic in the Nearctic[24], population size changes remained within normal inter-annual fluctuation levels in the Palearctic[38,54,60]. Our data suggest that, once the hibernacula are contaminated by the fungus, Palearctic bats are exposed to high pathogen pressure and the infection is continuously present to a high level, i.e. the so-called hyperendemic condition. The difference in infection outcome is therefore a function of adaptation to pathogen pressure.

While reduction of pathogen load is a function of host resistance, the ability to limit the harm caused by a given load is a result of tolerance[61,62]. These two alternative and complementary forms of defence may have profound effects on the epidemiology of infectious diseases and on host-pathogen coevolution[62]. Resistance protects the host at the expense of the pathogen and tolerance saves the host from harm without direct negative effects on the infectious agent. Evolution of resistance, therefore, should also reduce the prevalence of the pathogen in host populations. On the other hand, tolerance is expected to have a neutral, or even positive, effect on prevalence of the pathogen. Lack of resistance to WNS infection in hibernating European bats has been demonstrated in 13 species[31,42] that exhibit stable or increasing population sizes[54]. Yet, prevalence of *P. destructans*-positive bats reached 100% in the Palearctic. In light of the above arguments, it would appear that mechanisms promoting tolerance to *P. destructans* infection are in operation in the area studied. Palearctic bat species tolerate comparable fungal loads and WNS-lesion size to their Nearctic counterparts suffering population declines. We hypothesise that the balance between tolerance and resistance mechanisms changes with transition of bats from hibernation to euthermia.

Our results provide evidence of *P. destructans* and pathognomonic WNS skin lesions in hibernating bats sampled from the West Siberian Plain of Russian Asia and indicate endemicity of this virulent fungus in the Palearctic region. Data suggesting bat tolerance imply lowered risk following establishment of equilibrium in the host-pathogen interaction. The extensive spatial distribution of the agent may pose a threat, however, representing a continued source of introduction to other regions with naïve bats not yet exposed to the pathogen. Classical models employing the disease triangle concept suggest that the epidemiological outcome of an infection depends on determinants of the pathogen, host(s) and the environment. Alterations in any of these determinants may trigger shifts in the complex host-pathogen system, as seen here by infection tolerance in hibernating Palearctic bats.

Methods

Material collection. Between 2012 and 2014, we sampled 481 bats (15 species) at 20 sites in Slovenia, the Czech Republic, Latvia and Russia (West Siberian Plain, Asia; Table 1, Fig. 3). Samples were taken as late in the hibernation or as early in the post-hibernation season as possible (February-May) to minimise the impact of disturbance to specific bat species. Following capture, the wings, ears and muzzle were swabbed with a nylon (FLOQ Swabs, Copan Flock Technologies srl, Brescia, Italy) or cotton swab (Plain swab sterile plastic applicator, Copan) in a standardised manner in order to collect fungal biomass from the whole skin area for fungal detection (dorsal side of left wing only for qPCR). The bats were photographed over a 368 nm wavelength UV lamp and wing punch biopsies of suspect tissue collected in 10% formalin for histopathological examination. All bats were then released at the site.

WNS was diagnosed according to current standards[47], i.e. *P. destructans* presence was confirmed with qPCR[53] and selected orange-yellow spots observed over UV were sampled. A series of 80 periodic acid-Schiff stained sections embedded in paraffin were obtained from each wing membrane biopsy. These were then observed under an Olympus BX51 light microscope (Olympus Corporation, Tokyo, Japan). Fungal cell walls were stained magenta under the periodic acid-Schiff stain, which allowed measurement of cupping erosions packed with hyphae. Using cellSense Software tools (Olympus Soft Imaging, GmbH, Münster, Germany), we measured total area (size) of cupping erosions. Trans-illuminated photographs of the left wing membrane stretched over a UV lamp were used to manually enumerate yellow-orange fluorescing pinpoints indicative of WNS lesions[48], using the individual object counting tool of ImageJ[63].

Bats were considered WNS-positive if qPCR confirmed *P. destructans* infection, wings exhibited characteristic UV fluorescence[48] and cupping erosions packed with fungal hyphae or a full thickness fungal invasion of the wing membrane were observed under histopathology[47]. Additional swabs taken from the wings and muzzle after sampling for qPCR were used for cultivation on Sabouraud dextrose agar plates and isolation of the fungus in pure cultures[38]. Representative isolates were deposited at the Culture Collection of Fungi, Charles University in Prague, Czech Republic.

Field work and sampling in the Czech Republic was performed in accordance with Czech Law No. 114/1992 on Nature and Landscape Protection, based on permits 01662/MK/2012S/00775/MK/2012, 866/JS/2012 and 00356/KK/2008/AOPK issued by the Agency for Nature Conservation and Landscape Protection of the Czech

Republic. Experimental procedures were approved by the Ethical Committee of the Academy of Sciences of the Czech Republic (No. 169/2011). Sampling in Latvia was approved by the Nature Conservation Agency (No. 3.15/146/2014-N), in Slovenia by the Ministry of Environment and Spatial Planning of the Slovenian Republic, Slovenian Environment Agency (No. 35601-35/2010-6) and in Russia by the Institute of Plant and Animal Ecology, Ural Division of the Russian Academy of Sciences (No. 16353–2115/325). The authors were authorised to handle free-living bats according to the Czech Certificate of Competency (No. CZ01341; §17, Act No. 246/1992 Coll.) and a permit approved by the Latvian Nature Conservation Agency (No. 05/2014).

DNA isolation. Total DNA was isolated from swabs using the QIAamp DNA Mini Kit (Qiagen, Halden, Germany). Swabs were placed in tissue lysis buffer with proteinase K and incubated at 56 °C for two hours. Lysis buffer was added and the samples incubated for a further 10 min. After this step, we followed the Qiagen Buccal Swab Spin Protocol according to the manufacturer's recommendations.

Quantitative PCR. We performed quantitative PCR[53] (qPCR) using TaqMan® Universal Master Mix II with UNG (Life Technologies, Foster City, CA, USA). To optimise the PCR reaction, we added bovine serum albumin and Platinum® Taq DNA Polymerase (Life Technologies) in final concentrations of 0.05 mg/μl and 0.025 U/μl, respectively. We used both forward and reverse primers (0.3 μM each) and species-specific and genus-specific fluorescently-labelled custom probes were used for quantification of the PCR product. The reaction mix was prepared on ice and three replicates were prepared for each DNA sample. We performed real-time PCR reaction on a LightCycler 480 PCR platform (Life Technologies) with initial inactivation at 50 °C for 2 min and a hot start at 96 °C for 10 min. Nine cycles with a denaturation step at 95 °C for 15 sec and annealing at 62 °C for 1 min were followed by 43 identical cycles with quantification detection. qPCR was finalised by dissociation at 95-60-95 °C for 15 sec each and cooling to 40 °C for 10 min. DNA isolated from CCF3937 culture[38] and water were used as positive and negative controls and as concentration references for each run.

Fungal load on Palearctic bats. We calculated a DNA concentration calibration curve using a CCF3937 dilution series. Exact DNA concentrations (ng μl^{-1}) in the dilution series were determined using Qubit HS fluorometry via the manufacturer's protocol. Effectivity of qPCR was 1.96. We used this dilution series to estimate the relationship between fungal load and qPCR cycle, calculating *P. destructans* DNA concentration in the sample using custom scripts in R[64] with the equation $\log(q_{PdDNA}) = 3.194 - 0.287\ Cp$ ($R^2 = 0.9719$), where q is the DNA concentration and Cp the cycle. Each result was converted to fungal load based on the positive control and overall elution of DNA.

Fungal load on Nearctic bats. We downloaded previously published *P. destructans* loads from natural infections[19,58] or recalculated loads from mean Cp values according to the authors' equation[59]. In one case, we plotted means as individual data points as the authors[19] included species means per site (plus standard error) but fungal loads per individual were not published. The sample sizes for Nearctic bat species sampled within the dates used in this study were: *Corynorhinus rafinesquii* ($n = 2$), *Eptesicus fuscus* ($n = 6$), *Lasiurus borealis* ($n = 2$), *Myotis grisescens* ($n = 11$), *M. leibii* ($n = 1$), *M. lucifugus* ($n = 36$), *M. septentrionalis* ($n = 4$), *M. sodalis* ($n = 13$) and *Perimyotis fuscus* ($n = 172$). None of the North American studies specified sampled area on the bat with sufficient precision to enable standardisation to cm^2 or universal statistical comparison.

Molecular genetic variability of *P. destructans*. To assess genetic variability, we used a set of isolates collected between 2009 and 2014 in the Czech Republic[38,65] and isolates from this study. The isolates were characterised using *ITS* and five other nuclear gene sequences using previously published primers (Supplementary Table S3). Chromatograms were assembled to DNA sequences in Geneious 6 (Biomatters, Auckland, New Zealand) and submitted to the European Nucleotide Archive (LN871244-LN871428). These were checked with blast and *P. destructans* was confirmed at all sites with a sequence identity to previously sequenced strains $\geq$99% in all markers. We then downloaded previously published *P. destructans* sequences of the respective genes from GenBank. Sequence haplotypes were identified based on available nucleotide residues, with gaps and unresolved bases treated as missing data. We estimated relationships between haplotypes using median-joining networks[66].

Individual-based linear mixed models. We modelled intensity of WNS infection using fungal load on the dorsal side of the left wing and the number of WNS lesions on the same wing visualised over UV in R[64]. UV-detectable lesions correspond with cupping erosions diagnostic for WNS in Europe and North America, thereby reflecting disease intensity[47,48]. Our previous experience in the Czech Republic has shown that wing damage tends to be localised and that lesions do not usually merge[31], allowing them to be counted. The yellow-orange fluorescent spots were counted by a researcher with no knowledge of the qPCR and histopathology results for the samples. We scaled variables to 1 cm^2 of wing area[67] and log-transformed them in order to account for body-size differences between species. We included latitude and longitude, sampling day (from the beginning of the calendar year) and scores for the first two principal components (characterising dominant hibernating bat community in each region; Supplementary Fig. S2) as fixed-effect variables and species as a random effect. Microclimate at a specific bat roost might influence fungal load and disease intensity on an individual through a compound effect of optimizing growth conditions for the pathogen and hibernation conditions for the host. We used geographic coordinates as a proxy for possible general site microclimate (mean annual temperature, elevation, humidity) under the assumption that fungal load and number of WNS lesions would increase at higher latitudes and with longitudinal shift from oceanic to more continental climate. Similarly, we expected day of sampling to be reflected in the dependent variables as increased time for fungus propagation might increase its loading. As such,

geographic location, combined with sampling date, can be understood as a proxy for cave microclimate, which would influence growth of the WNS fungus[50,68].

Both relatedness and species assemblage differences at hibernacula in a given region could skew regional fungal load and disease intensity in favour of those areas where severely infested species co-occur. We estimated bat communities present at hibernacula from local surveys. Relatedness of species recorded was assessed using mean phylogenetic distance and mean nearest taxon distance (both with weighted abundance) and UniFrac metric measuring the percentage of phylogeny shared by a given community. We used these metrics for principal components analysis. We used backward stepwise variable selection to develop the model, which was assessed using the Akaike Information Criterion (AIC).

Species-level phylogenetically-informed model. Previously published phylogenetic trees for European bats[31] based on multilocus sequence data were rescaled as an ultrametric tree using penalised likelihood[69,70] with λ and root height = 1. We used R[70,71] to calculate phylogenetic signal in species means and phylogenetic ANOVA[71,72]. We used the phylogenetic generalised least-squares method[72,73] to explain the number of visible WNS lesions on a UV trans-illuminated wing based on *P. destructans* fungal load.

References

1. Daszak, P., Cunningham, A. A. & Hyatt, A. D. Emerging infectious diseases of wildlife - Threats to biodiversity and human health. *Science* **287,** 443–449 (2000).
2. Keesing, F. *et al.* Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature* **468,** 647–652 (2010).
3. Dobson, A. & Foufopoulos, J. Emerging infectious pathogens of wildlife. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356,** 1001–1012 (2001).
4. Puechmaille, S. J. *et al.* White-nose syndrome: is this emerging disease a threat to European bats? *Trends Ecol. Evol.* **26,** 570–576 (2011).
5. Fisher, M. C. *et al.* Emerging fungal threats to animal, plant and ecosystem health. *Nature* **484,** 186–194 (2012).
6. Bergman, A. & Casadevall, A. Mammalian endothermy optimally restricts fungi and metabolic costs. *mBio* **1,** e00212–00210 (2010).
7. Garcia-Solache, M. A. & Casadevall, A. Global warming will bring new fungal diseases for mammals. *mBio* **1,** e00061–00010 (2010).
8. González-González, A. E. *et al. Histoplasma capsulatum* and *Pneumocystis spp*. co-infection in wild bats from Argentina, French Guyana, and Mexico. *BMC Microbiology* **14,** 23 (2014).
9. Johara, M. Y. *et al.* Nipah virus infection in bats (order Chiroptera) in peninsular Malaysia. *Emerg. Infect. Dis.* **7,** 439–441 (2001).
10. Hayman, D. T. S. *et al.* Long-term survival of an urban fruit bat seropositive for Ebola and Lagos bat viruses. *PLoS ONE* **5,** e11978 (2010).
11. Middleton, D. J. *et al.* Experimental Nipah virus infection in Pteropid bats (*Pteropus poliocephalus*). *J. Comp. Pathol.* **136,** 266–272 (2007).
12. Olival, K. J. & Hayman, D. T. S. Filoviruses in bats: Current knowledge and future directions. *Viruses* **6,** 1759–1788 (2014).
13. Calisher, C. H., Childs, J. E., Field, H. E., Holmes, K. V. & Schountz, T. Bats: Important reservoir hosts of emerging viruses. *Clin. Microbiol. Rev.* **19,** 531–545 (2006).
14. Moratelli, R. & Calisher, C. H. Bats and zoonotic viruses: can we confidently link bats with emerging deadly viruses? *Mem. Inst. Oswaldo Cruz.* **110,** 1–22 (2015).
15. O'Shea, T. J. *et al.* Bat flight and zoonotic viruses. *Emerg. Infect. Dis.* **20,** 741–745 (2014).
16. Mandl, J. N. *et al.* Reservoir host immune responses to emerging zoonotic viruses. *Cell* **160,** 20–35 (2015).
17. Meteyer, C. U., Barber, D. & Mandl, J. N. Pathology in euthermic bats with white nose syndrome suggests a natural manifestation of immune reconstitution inflammatory syndrome. *Virulence* **3,** 583–588 (2012).
18. Bouma, H. R., Carey, H. V. & Kroese, F. G. M. Hibernation: the immune system at rest? *J. Leukoc. Biol.* **88,** 619–624 (2010).
19. Langwig, K. E. *et al.* Host and pathogen ecology drive the seasonal dynamics of a fungal disease, white-nose syndrome. *Proc. R. Soc. B Biol. Sci.* **282,** 20142335 (2015).
20. Blehert, D. S. *et al.* Bat white-nose syndrome: An emerging fungal pathogen? *Science* **323,** 227 (2009).
21. Gargas, A., Trest, M. T., Christensen, M., Volk, T. J. & Blehert, D. S. *Geomyces destructans* sp. nov. associated with bat white-nose syndrome. *Mycotaxon* **108,** 147–154 (2009).
22. Minnis, A. M. & Lindner, D. L. Phylogenetic evaluation of *Geomyces* and allies reveals no close relatives of *Pseudogymnoascus destructans*, comb. nov., in bat hibernacula of eastern North America. *Fungal Biol.* **117,** 638–649 (2013).
23. Warnecke, L. *et al.* Inoculation of bats with European *Geomyces destructans* supports the novel pathogen hypothesis for the origin of white-nose syndrome. *Proc. Nat. Acad. Sci. USA* **109,** 6999–7003 (2012).
24. Frick, W. F. *et al.* An emerging disease causes regional population collapse of a common North American bat species. *Science* **329,** 679–682 (2010).
25. Coleman, J. T. H. & Reichard, J. D. Bat white-nose syndrome in 2014: A brief assessment seven years after discovery of a virulent fungal pathogen in North America. *Outlooks in Pest Management* **25,** 374–377 (2014).
26. Turner, G. G., Reeder, D. M. & Coleman, J. T. H. A five-year assessment of mortality and geographic spread of white-nose syndrome in North American bats, with a look at the future. Update of white-nose syndrome in bats. *Bat Research News* **52,** 13–27 (2011).
27. Lorch, J. M. *et al.* Experimental infection of bats with *Geomyces destructans* causes white-nose syndrome. *Nature* **480,** 376–378 (2011).
28. Langwig, K. E. *et al.* Sociality, density-dependence and microclimates determine the persistence of populations suffering from a novel fungal disease, white-nose syndrome. *Ecol. Lett.* **15,** 1050–1057 (2012).
29. Lorch, J. M. *et al.* Distribution and environmental persistence of the causative agent of white-nose syndrome, *Geomyces destructans*, in bat hibernacula of the Eastern United States. *Appl. Environ. Microbiol.* **79,** 1293–1301 (2013).
30. Hoyt, J. R. *et al.* Long-term persistence of *Pseudogymnoascus destructans*, the causative agent of white-nose syndrome, in the absence of bats. *EcoHealth*, doi: 10.1007/s10393-014-0981-4 (2014).
31. Zukal, J. *et al.* White-nose syndrome fungus: A generalist pathogen of hibernating bats. *PLoS ONE* **9,** e97224 (2014).
32. Lorch, J. M. *et al.* A culture-based survey of fungi in soil from bat hibernacula in the eastern United States and its implications for detection of *Geomyces destructans*, the causal agent of bat white-nose syndrome. *Mycologia* **105,** 237–252 (2013).
33. Ren, P. *et al.* Clonal spread of *Geomyces destructans* among bats, midwestern and southern United States. *Emerg. Infect. Dis.* **18,** 883–885 (2012).
34. Rajkumar, S. S. *et al.* Clonal genotype of *Geomyces destructans* among bats with white nose syndrome, New York, USA. *Emerg. Infect. Dis.* **17,** 1273–1276 (2011).
35. Escobar, L. E., Lira-Noriega, A., Medina-Vogel, G. & Townsend Peterson, A. Potential for spread of the white-nose fungus (*Pseudogymnoascus destructans*) in the Americas: use of Maxent and NicheA to assure strict model transference. *Geospat. Health* **9,** 221–229 (2014).

36. Maher, S. P. *et al.* Spread of white-nose syndrome on a network regulated by geography and climate. *Nat. Commun.* **3,** 1306 (2012).
37. Maslo, B. & Fefferman, N. H. A case study of bats and white-nose syndrome demonstrating how to model population viability with evolutionary effects. *Conserv. Biol.* **29,** 1176–1185 (2015).
38. Martínková, N. *et al.* Increasing incidence of *Geomyces destructans* fungus in bats from the Czech Republic and Slovakia. *PLoS ONE* **5,** e13853 (2010).
39. Puechmaille, S. J. *et al.* Pan-European distribution of white-nose syndrome fungus (*Geomyces destructans*) not associated with mass mortality. *PLoS ONE* **6,** e19167 (2011).
40. Wibbelt, G. *et al.* White-nose syndrome fungus (*Geomyces destructans*) in bats, Europe. *Emerg. Infect. Dis.* **16,** 1237–1243 (2010).
41. Pavlinić, I., Đaković, M. & Lojkić, I. *Pseudogymnoascus destructans* in Croatia confirmed. *Eur. J. Wildl. Res.* **61,** 325–328 (2015).
42. Bandouchova, H. *et al. Pseudogymnoascus destructans*: evidence of virulent skin invasion for bats under natural conditions, Europe. *Transbound. Emerg. Dis.* **62,** 1–5 (2015).
43. Palmer, J. M. *et al.* Molecular characterization of a heterothallic mating system in *Pseudogymnoascus destructans*, the fungus causing white-nose syndrome of bats. *G3 (Bethesda)* **4,** 1755–1763 (2014).
44. Leopardi, S., Blake, D. & Puechmaille, S. J. White-nose syndrome fungus introduced from Europe to North America. *Curr. Biol.* **25,** R217–219 (2015).
45. Blehert, D. S. Fungal disease and the developing story of bat white-nose syndrome. *PLoS Pathog.* **8,** e1002779 (2012).
46. Lorch, J. M. *et al.* The fungus *Trichophyton redellii* sp. nov. causes skin infections that resemble white-nose syndrome of hibernating bats. *J. Wildl. Dis.* **51,** 36–47 (2015).
47. Meteyer, C. U. *et al.* Histopathologic criteria to confirm white-nose syndrome in bats. *J. Vet. Diagn. Invest.* **21,** 411–414 (2009).
48. Turner, G. G. *et al.* Nonlethal screening of bat-wing skin with the use of ultraviolet fluorescence to detect lesions indicative of white-nose syndrome. *J. Wildl. Dis.* **50,** 566–573 (2014).
49. Hutterer, R., Ivanova, T., Meyer-Cords, C. & Rodrigues, L. *Bat Migrations in Europe: A Review of Banding Data and Literature*. (Federal Agency for Nature Conservation, 2005).
50. Perry, R. W. A review of factors affecting cave climates for hibernating bats in temperate North America. *Environ. Rev.* **21,** 28–39 (2013).
51. Hoyt, J. R. *et al.* Widespread bat white-nose syndrome fungus, Northeastern China. *Emerg. Infect. Dis.* **22,** doi: 10.3201/eid2201.151314 (2016).
52. McCallum, H. & Dobson, A. Detecting disease and parasite threats to endangered species and ecosystems. *Trends Ecol. Evol.* **10,** 190–194 (1995).
53. Shuey, M. M., Drees, K. P., Lindner, D. L., Keim, P. & Foster, J. T. Highly sensitive quantitative PCR for the detection and differentiation of *Pseudogymnoascus destructans* and other *Pseudogymnoascus* species. *Appl. Environ. Microbiol.* **80,** 1726–1731 (2014).
54. Van der Meij, T. *et al.* Return of the bats? A prototype indicator of trends in European bat populations in underground hibernacula. *Mammalian Biology - Zeitschrift für Säugetierkunde* **80,** 170–177 (2015).
55. Langwig, K. E. *et al.* Invasion dynamics of white-nose syndrome fungus, Midwestern United States, 2012–2014. *Emerg. Infect. Dis.* **21,** doi: 10.3201/eid2106.150123 (2015).
56. Retallick, R. W., McCallum, H. & Speare, R. Endemic infection of the amphibian chytrid fungus in a frog community post-decline. *PLoS Biol.* **2,** e351 (2004).
57. Horrocks, N. P., Matson, K. D. & Tieleman, B. I. Pathogen pressure puts immune defense into perspective. *Integr. Comp. Biol.* **51,** 563–576 (2011).
58. Bernard, R. F., Foster, J. T., Willcox, E. V., Parise, K. L. & McCracken, G. F. Molecular detection of the causative agent of white-nose syndrome on Rafinesque's big-eared bats (*Corynorhinus rafinesquii*) and two species of migratory bats in the southeastern USA. *J. Wildl. Dis.* **51,** 519–522 (2015).
59. Janicki, A. F. *et al.* Efficacy of visual surveys for White-nose syndrome at bat hibernacula. *PLoS ONE* **10,** e0133390, doi: 10.1371/journal.pone.0133390 (2015).
60. Frick, W. F. *et al.* Disease alters macroecological patterns of North American bats. *Glob. Ecol. Biogeogr.* **24,** 741–749, doi: 10.1111/geb.12290 (2015).
61. Medzhitov, R., Schneider, D. S. & Soares, M. P. Disease tolerance as a defense strategy. *Science* **335,** 936–941 (2012).
62. Råberg, L., Graham, A. L. & Read, A. F. Decomposing health: tolerance and resistance to parasites in animals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364,** 37–49 (2009).
63. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Meth.* **9,** 676–682 (2012).
64. R., Core Team. R: A language and environment for statistical computing, <http://www.R-project.org/> (2013).
65. Kubátová, A., Koukol, O. & Nováková, A. *Geomyces destructans*, phenotypic features of some Czech isolates. *Czech Mycol.* **63,** 65–75 (2011).
66. Bandelt, H. J., Forster, P. & Röhl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16,** 37–48 (1999).
67. Norberg, U. M. & Rayner, J. M. V. Ecological morphology and flight in bats (Mammalia, Chiroptera): Wing adaptations, flight performance, foraging strategy and echolocation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **316,** 335–427 (1987).
68. Hallam, T. G. & Federico, P. The panzootic white-nose syndrome: an environmentally constrained disease? *Transbound. Emerg. Dis.* **59,** 269–278 (2012).
69. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20,** 289–290 (2004).
70. Sanderson, M. J. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19,** 101–109 (2002).
71. Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E. & Challenger, W. GEIGER: Investigating evolutionary radiations. *Bioinformatics* **24,** 129–131 (2008).
72. Revell, L. J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3,** 217–223 (2012).
73. Ives, A. R., Midford, P. E. & Garland Jr., T. Within-species measurement error in phylogenetic comparative methods. *Syst. Biol.* **56,** 252–270 (2007).

Acknowledgements

This study was supported through a grant from the Czech Science Foundation (Grant No. P506-12-1064). We are grateful to Tomáš Bartonička, Hana Berková and Masha Orlova for invaluable field assistance, to Matej Dolinay, Jiří C. Moravec, Patrícia Pečnerová and Aneta Reichová for laboratory assistance and to Gregory G. Turner for biopsy samples from Nearctic bats.

Author Contributions

J.Z., J.P. and N.M. designed the study and participated in field research, with help from H.B., J.B., V.K., O.O., P.P. and J.S. H.B., A.C., A.K., A.N., J.P. and A.Z. performed the laboratory analyses. K.S.J., M.K., J.P. and N.M. analysed the data. J.Z., J.P. and N.M. wrote the manuscript, with contributions from all authors.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Zukal, J. *et al.* White-nose syndrome without borders: *Pseudogymnoascus destructans* infection tolerated in Europe and Palearctic Asia but not in North America. *Sci. Rep.* **6**, 19829; doi: 10.1038/srep19829 (2016).

Paper 2.4.3

Pikula J., Amelon S. K., Bandouchova H., Bartonička T., Berkova H., Brichta J., Hooper S., Kokurewicz T., Kolarik M., Köllner B., Kovacova V., Linhart P., Piacek V., Turner G. G., Zukal J., **Martínková N.** 2017. White-nose syndrome pathology grading in Nearctic and Palearctic bats. *PLoS ONE* 12: e0180435.

PLOS ONE

RESEARCH ARTICLE

White-nose syndrome pathology grading in Nearctic and Palearctic bats

Jiri Pikula[1,2]*, Sybill K. Amelon[3], Hana Bandouchova[1], Tomáš Bartonička[4], Hana Berkova[5], Jiri Brichta[1], Sarah Hooper[6], Tomasz Kokurewicz[7], Miroslav Kolarik[8], Bernd Köllner[9], Veronika Kovacova[1], Petr Linhart[1], Vladimir Piacek[1], Gregory G. Turner[10], Jan Zukal[4,5], Natália Martínková[5,11]

1 Department of Ecology and Diseases of Game, Fish and Bees, University of Veterinary and Pharmaceutical Sciences Brno, Brno, Czech Republic, **2** CEITEC—Central European Institute of Technology, University of Veterinary and Pharmaceutical Sciences Brno, Brno, Czech Republic, **3** United States Department of Agriculture Forest Service, Northern Research Station, Columbia, Missouri, United States of America, **4** Department of Botany and Zoology, Masaryk University, Brno, Czech Republic, **5** Institute of Vertebrate Biology, Czech Academy of Sciences, Brno, Czech Republic, **6** Department of Veterinary Pathobiology, University of Missouri, Columbia, Missouri, United States of America, **7** Institute of Biology, Department of Vertebrate Ecology and Palaeontology, Wrocław University of Environmental and Life Sciences, Wrocław, Poland, **8** Laboratory of Fungal Genetics and Metabolism, Institute of Microbiology, Czech Academy of Sciences, Prague, Czech Republic, **9** Institute of Immunology, Friedrich-Loeffler-Institute, Federal Research Institute for Animal Health, Greifswald-Insel Riems, Germany, **10** Pennsylvania Game Commission, Harrisburg, Pennsylvania, United States of America, **11** Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic

* pikulaj@vfu.cz

OPEN ACCESS

Citation: Pikula J, Amelon SK, Bandouchova H, Bartonička T, Berkova H, Brichta J, et al. (2017) White-nose syndrome pathology grading in Nearctic and Palearctic bats. PLoS ONE 12(8): e0180435. https://doi.org/10.1371/journal.pone.0180435

Editor: Sharon Swartz, Brown University, UNITED STATES

Received: December 2, 2016

Accepted: May 26, 2017

Published: August 2, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was supported by the Czech Science Foundation (Grant No. P506-12-1064 and 17-20286S). This research was carried out as part of the CEITEC 2020 project (LQ1601), with further financial support from the Ministry of Education, Youth and Sports of the Czech Republic under National Sustainability Programme II.

Abstract

While white-nose syndrome (WNS) has decimated hibernating bat populations in the Nearctic, species from the Palearctic appear to cope better with the fungal skin infection causing WNS. This has encouraged multiple hypotheses on the mechanisms leading to differential survival of species exposed to the same pathogen. To facilitate intercontinental comparisons, we proposed a novel pathogenesis-based grading scheme consistent with WNS diagnosis histopathology criteria. UV light-guided collection was used to obtain single biopsies from Nearctic and Palearctic bat wing membranes non-lethally. The proposed scheme scores eleven grades associated with WNS on histopathology. Given weights reflective of grade severity, the sum of findings from an individual results in weighted cumulative WNS pathology score. The probability of finding fungal skin colonisation and single, multiple or confluent cupping erosions increased with increase in *Pseudogymnoascus destructans* load. Increasing fungal load mimicked progression of skin infection from epidermal surface colonisation to deep dermal invasion. Similarly, the number of UV-fluorescent lesions increased with increasing weighted cumulative WNS pathology score, demonstrating congruence between WNS-associated tissue damage and extent of UV fluorescence. In a case report, we demonstrated that UV-fluorescence disappears within two weeks of euthermy. Change in fluorescence was coupled with a reduction in weighted cumulative WNS pathology score, whereby both methods lost diagnostic utility. While weighted cumulative WNS pathology scores were greater in the Nearctic than Palearctic, values for Nearctic bats were within the range of those for Palearctic species. Accumulation of wing damage probably influences mortality in affected bats, as demonstrated by a fatal case of *Myotis daubentonii*

Competing interests: The authors have declared that no competing interests exist.

with natural WNS infection and healing in *Myotis myotis*. The proposed semi-quantitative pathology score provided good agreement between experienced raters, showing it to be a powerful and widely applicable tool for defining WNS severity.

Introduction

Wildlife conservation medicine is currently being challenged by a number of infectious and non-infectious diseases [1] that can potentially induce mass mortality events [2]. Recently, the health of temperate-region bats has been compromised by a generalist fungal agent that causes white-nose syndrome, *Pseudogymnoascus destructans* [3–7]. White-nose syndrome (WNS) emerged as a point-source epidemic. Its geographic spread since 2006 has been associated with a major decline in Nearctic bat populations [8–10]. On the other hand, Palearctic bat communities in Europe and Asia appear to tolerate hyperendemic exposure to this virulent pathogen [11].

WNS is characterised by the invasive skin infection caused by *P. destructans* [3,5,12,13]. Extensive damage to flight membranes may alter the torpor pattern of hibernating bats by increasing their arousal frequency and depleting their fat reserves prematurely [14,15]. Disruption of the effective skin's barrier function can, therefore, explain the pathophysiological mechanisms underlying mortality in WNS-affected bats [16–20].

Pathognomonic skin lesions are the only reliable sign of this syndromic disease that are easy to detect using laboratory methods. In combination with identification of the pathogen [4,6], therefore, histopathology is seen as the gold standard for diagnosing WNS qualitatively [13]. While histopathology has indicated equivalent focal skin-tissue invasiveness in multiple bat species naturally infected with *P. destructans* throughout its known geographic range [7,11–13,21], it does not explain the striking difference in infection outcome between Nearctic and Palearctic bats [11]. Likewise, no significant difference in fungal load and cupping erosion size has been found on bats in Europe, Palearctic Asia and North America [11]. In order to better understand WNS progression and severity, quantification of histopathological findings in bats sampled from different regions is needed.

Until recently, bats had to be dead or euthanised for laboratory testing procedures, collection of dermato-histopathological samples from the wings, muzzle and ears and to optimise detection of the disease [13]. What is more, the severity scoring system for WNS currently in use utilises the whole membrane from one wing [14]. Both in Europe and elsewhere, bat species are under strict protection; hence, the need for a non-lethal sampling method is imperative. In response, our team recently validated a new non-lethal technique for identifying and targeting WNS skin lesions for sampling [22]. Trans-illumination of a wing membrane with 366–385 nm ultraviolet (UV) light elicits a distinct orange-yellow fluorescence that corresponds directly with the fungal cupping erosions in histological sections of the respective skin area. The fluorescence emitted from these skin lesions is associated with hyperaccumulation of riboflavin, a secondary fungal metabolite that may also represent a virulence factor leading to skin damage [23]. When not being used in combination with infected wing membrane biopsies, UV transillumination can also be used for non-invasive photographic surveillance of infection intensity.

Based on the need for a standardised non-lethal tool for measuring skin pathology in bats from different regions, the objective of the present study was to establish and validate a novel grading system for defining WNS severity based on single biopsies. Here, we propose a

weighted cumulative WNS pathology score based on UV trans-illumination guided biopsies that allows for semi-quantitative comparison and shows high inter-pathologist reproducibility. We use this grading system to describe and compare histopathological features of *P. destructans* infection in both Nearctic and Palearctic bats previously qualitatively diagnosed with WNS. While non-lethal diagnostic tools offer the opportunity to follow the progression of skin pathology, prior experience with WNS pathology scores may be used to predict the outcome of infection in a diseased bat. With intensive research for treatment of WNS in North America [24], ability to predict patient prognosis becomes imperative. Lacking sufficient sample sizes for statistical model evaluation, case report experience might provide valuable information. We, therefore, used time series data on two bats receiving supportive care in a rehabilitation facility to document the clinical outcome of WNS and to examine the diagnostic utility of the proposed grading system in the early post-hibernation period.

Material and methods

Ethics statement

Collection of bat samples from hibernacula in the Czech Republic complied with Czech Law No. 114/1992 on Nature and Landscape Protection. Collection was based on permits 01662/MK/2012S/00775/MK/2012, 866/JS/2012 and 00356/KK/2008/AOPK issued by the Agency for Nature Conservation and Landscape Protection of the Czech Republic. Approval of all experimental procedures was provided by the Ethical Committee of the Czech Academy of Sciences (No. 169/2011). Sampling at the "Nietopierek" Natura 2000 site (Poland) was approved by the II Local Ethical Commission in Wrocław (No. 45/2015). Sampling in Latvia, Slovenia, Russia and Poland was approved by the Latvian Nature Conservation Agency (No. 3.15/146/2014-N), the Ministry of Environment and Spatial Planning of the Slovenian Republic, the Slovenian Environment Agency (No. 35601-35/2010-6), the Institute of Plant and Animal Ecology—Ural Division of the Russian Academy of Sciences (No. 16353–2115/325) and the Regional Directorate for Environmental Protection in Gorzow Wielkopolski (No. WPN-I-6205.10.2015.AI). Collection of samples from bats in Hannibal, MO complied with Missouri Department of Conservation Scientific Wildlife Collector's Permit (No. 15947/2014) and was performed under a protocol approved by University of Missouri, Animal Care and Use Committee. The authors were authorised to handle wild bats according to the Czech Certificate of Competency (No. CZ01341; §17, Act No. 246/1992 Coll.) and a permit approved by the Latvian Nature Conservation Agency (No. 05/2014).

Surveillance of bats for white-nose syndrome

The origin of Holarctic bat samples has been described previously [7,11,12]. Briefly, bats were sampled at 20 sites in the Czech Republic, Slovenia, Latvia, Poland, Russia and the USA. Additional Nearctic *Myotis septentrionalis* males were sampled in Hannibal, Missouri (39.70 N, 91.36 W; USA) from January to March 2015, two hibernation seasons after the first documentation of *P. destructans* infection.

As described previously [7,11,12], bat wings were swabbed for laboratory examination of *P. destructans* infection using culture and estimation of associated fungal load using quantitative polymerase chain reaction (qPCR), and photographed over a 368 nm UV lamp for later enumeration of lesions indicative of WNS [22]. One WNS-suspect spot from each bat was then sampled under UV guidance. In this way, we collected wing membrane biopsies from 210 specimens of 21 different Nearctic and Palearctic bat species during the late hibernation and early post-hibernation periods of 2012 to 2015 (S1 Table). All bats were handled in such a way as to minimise any impact from disturbance and then quickly released at the capture site. Skin

samples collected with a 4 mm sterile punch (Kruuse, Denmark) were immediately fixed in 10% formalin, then dehydrated in the laboratory and embedded in paraffin. To obtain a representative range of histopathology findings, a series of 80 5 μm sections were prepared from each wing membrane biopsy and stained with periodic acid-Schiff (PAS). The slides were examined with light microscopy with focus on invasive fungal growth and identification of the WNS skin pathology grades described below.

Case reports of WNS progression

In the Czech Republic, single specimens of *M. daubentonii* and *M. myotis* with extensive WNS infection were recognised at hibernacula in the Podyjí National Park and Jeseníky Mountains and sent to the Rescue Centre at the University of Veterinary and Pharmaceutical Sciences Brno (Czech Republic). While healing was being augmented with captive nutritional support, skin disease progression under euthermic conditions was documented using photography and histopathology.

A novel grading system of WNS skin pathology

The semi-quantitative grading system presented here is based on a single bat flight membrane biopsy per animal guided by UV transillumination. Eleven binary encoded grades g, pathognomonic for WNS [13] or associated with the early stages of *P. destructans* skin infection, were selected as index signs for grading and were identified on Periodic acid-Schiff (PAS)-stained histopathological slides. The grades were ordered with increasing invasion severity from presence of the fungus on the wing's surface to replacement of tissue by fungal hyphae throughout the wing's thickness. Skin surface colonisation by the fungus (g_1) was followed by hair follicle infection (g_2), sebaceous gland infection (g_3), single occurrences of cup-like lesions (g_4), multiple and/or confluent cupping erosions (g_5), skin basement membrane breach by the fungus (g_6) and full-thickness invasion (g_7). The bat's immune response to infection was represented by variable inflammatory response manifested as tissue infiltration with neutrophils (g_8) and fungal sequestration by extensive inflammatory response (g_{11}). Findings of skin necrosis (g_9), characterised by loss of identifiable skin structures and skin infarction (g_{10}), were also included among the criteria associated with *P. destructans* skin infection.

In order to evaluate infection severity, we assign a weight w to each grade based on its invasiveness and extensiveness. The weights were selected so that a cumulative score of lower grade findings could not outweigh more severe grades of skin infection. A combination of fungal skin colonisation ($w_1 = 1$) and hair follicle ($w_2 = 2$) and sebaceous gland infection ($w_3 = 2$) represent the least severe finding, with a cumulative score lower than that in samples with single ($w_4 = 6$) or multiple ($w_5 = 12$) cupping erosions. Basement membrane breach ($w_6 = 13$) and full thickness infection ($w_7 = 19$) represent severe disruptions of the deep layers of bat wing membranes. We consider inflammation ($w_8 = 20$), necrosis ($w_9 = 25$) and infarction ($w_{10} = 30$) to be the most severe WNS grades. When the immune response succeeds in fungal sequestration ($w_{11} = -20$ to reflect potential healing), however, recovery from WNS appears more likely.

Inspection of 80 wing membrane sections from each biopsy could result in a weighted cumulative WNS pathology score defined as:

$$\text{histoSum} = \sum_{i=1}^{11} g_i w_i,$$

where $g \in \{0,1\}$, with $g = 0$ meaning absence and $g = 1$ meaning presence of the respective WNS pathology grade. The histoSum values may range from −20 (binary code corresponding

to the listed grades: 00000000001) to 130 (11111111110). As an example, a sample scored 32 if the biopsy contained fungal skin colonisation (+1) together with multiple cupping erosions (+12) and hyphae that breached the basal membrane (+13). Note that a single cupping erosion (+6) is scored automatically when multiple or confluent cupping erosions are present.

Data used in this study is available in S2 Table.

Inter-pathologist reproducibility of the proposed grading system

In order to test whether the above-described index signs for WNS grading were universally recognizable, we photographed 30 randomly selected PAS-stained sections prepared from bat-wing biopsies. In a blind evaluation, five independent pathologists examined and scored the photographs. Clarity of grading in naïve raters was evaluated in an experiment whereby 72 undergraduate students from the University of Veterinary and Pharmaceutical Sciences Brno scored a subset of 10 photographs following 45-minutes training in WNS histopathology. The participating students of veterinary medicine already received education in general pathology during their veterinary study program and volunteered for the task of scoring *P. destructans* skin infection pathology. These volunteers were recruited at a Wildlife Diseases lecture. Assignment to three evaluating groups was performed to limit the time that the volunteers spent reading WNS histopathology.

Estimation of fungal load

Fungal DNA was isolated from swabs of the dorsal side of bat left wing using QIAamp DNA Mini Kit (Qiagen, Halden, Germany). DNA from *P. destructans* was quantified with a dual-probe TaqMan (Life Technologies, Foster City, CA, USA) assay developed by Shuey et al. [25], following the detailed protocol for the qPCR of Zukal et al. [11]. Each sample was run in triplicate and each plate included positive and negative controls. Fungal load was estimated from positive control dilution series calibration curve according to equation log (*P. destructans* DNA concentration) = 3.194–0.287 * cycle, and standardized to total fungal load in nanograms per 1 cm^2 of wing area to correct for species size differences [11].

Statistical analyses

Weighted cumulative WNS grading scores for the Palearctic and Nearctic zones were compared using the Mann-Whitney *U* test. The influence of species-specific histoSum was assessed by filtering the value for species random effect using the *lme4* package in R [26,27].

The relationship between data provided by non-destructive diagnostic methods and initial progression of WNS infection by histopathology was evaluated using logistic regression [27]. Two sets of three logistic regressions were fitted for skin surface colonisation by the fungus, formation of single and multiple cupping erosions. Fungal load (log_{10}(ng)) and number of UV-fluorescent spots on the left wing recalculated to cm^2 of wing membrane [11] were used as independent variables in the respective models. Wing membrane area for *M. septentrionalis* was not available in literature and thus it was calculated from UV photographs using a custom R script and the *splancs* and *jpeg* packages [28,29]. The relationship between histoSum and number of UV fluorescent spots was evaluated using phylogenetic generalised least-squares [30] in the *phytools* package in R [31]. Previously published multilocus phylogeny was used to correct for species relatedness [7].

Inter-rater agreement of pathologists and veterinary students was tested in a paired study using the unweighted Cohen's κ coefficient [32] against scores included in the results below, where confidence was estimated with equations from [33]. The overall evaluation of κ for multiple raters was used according to Fleiss et al.'s modification [34]. The κ coefficient corrects

grading category agreement frequency for an effect where raters achieve consensus by chance. We used custom R scripts along with the *irr* package [35]. Confidence intervals for multi-rater agreement were estimated from 1000 bootstrap replicates.

Results

Prevalence of each WNS pathology grade (Figs 1 and 2) induced by natural *P. destructans* skin infection in Palearctic and Nearctic bat species varied from 0 to 100% in different bat species (Table 1). Signs of fungal skin-surface colonisation not classified as WNS (in the absence of other lesions) were present in all Nearctic bats and in 89% of Palearctic bats. Prevalence of infection within hair follicles and associated glands was highly variable in all hibernating bat species and for species with multiple investigated individuals ranged from 17 to 67% for the hair follicle infection and from 15 to 67% for the sebaceous gland infection. While wing membrane infection progressed to multiple and/or confluent cupping erosions in most bat species, single cupping erosions only were observed in *Miniopterus schreibersii* ($n = 1$), *M. bechsteinii* ($n = 3$) and *Rhinolophus euryale* ($n = 1$). Fungal hyphae regularly breached the epidermal/dermal interface in *M. lucifugus* (100%, $n = 10$) and *M. septentrionalis* (86%, $n = 7$) as well as in Palearctic bats (46–100%). Full thickness invasion of the wing membrane occurred more frequently in Palearctic bats (34% on average) and *M. septentrionalis* (71%) than in *M. lucifugus* (30%). A high percentage of both Palearctic and Nearctic bats (73 and 76% on average, respectively) showed an inflammatory response to fungal invasion, though this mostly resulted in a considerably lower occurrence of fungal sequestration (25 and 22% on average, respectively; Table 1). While *Eptesicus serotinus* ($n = 1$), *M. alcathoe* ($n = 7$), *Nyctalus noctula* ($n = 8$), *Pipistrellus pipistrellus* ($n = 2$), *Pipistrellus pygmaeus* ($n = 2$) and *Plecotus austriacus* ($n = 1$) were also sampled for histopathology, all were confirmed negative for WNS. Among these, four *M. alcathoe* were confirmed positive for *P. destructans* infection using qPCR (S2 Table).

Sum of scores and inter-pathologist agreement in the proposed grading system

While the histoSum score can reach a theoretical maximum of 130, the maximum score in practice was 126 for a fatal case of WNS in the *M. daubentonii* specimen described below (binary code 10011111011). The species with highest average histoSum were *Eptesicus nilssonii* from Asia ($n = 2$), *M. brandtii* ($n = 1$), *Rhinolophus hipposideros* ($n = 4$), *Barbastella barbastellus* ($n = 3$) from Europe and *M. septentrionalis* from North America ($n = 7$; Fig 3). While the mean histoSum for all sampled bats was 39.2 (std. error = 2.74), *E. nilssonii* ($n = 2$), *M. schreibersii* ($n = 2$), *M. bechsteinii* ($n = 7$), *M. dasycneme* ($n = 6$), *M. myotis* ($n = 57$), *M. nattereri* ($n = 8$) and *R. euryale* ($n = 1$) from Europe had a lower average weighted sum of concurrent findings (Fig 3). Nearctic bats had a significantly higher histoSum score (mean ± std. error: 64.9 ± 6.4) than Palearctic bats (35.9 ± 2.9; Mann-Whitney $U = 1688$, $p < 0.001$). Correction for random effect of species had no influence on significance.

Paired scoring agreement corrected for chance between five experienced pathologists and results presented herein ranged from 0.76 to 0.85 (Cohen's κ, $p < 0.001$; 95% confidence intervals (CI): {[0.75, 0.87], [0.8, 0.91], [0.72, 0.85], [0.77, 0.89], [0.7, 0.83]}) with an overall Fleiss' κ_5 of 0.78 ($z = 57.49$, $p < 0.001$, CI: [0.74, 0.82]), signifying good agreement in scoring WNS severity. Inter-rater agreement evaluating each WNS pathology grade separately differed, with basement membrane breach and inflammation being most difficult to score consistently (Table 2). Naïve raters displayed lower inter-rater agreement with the reported scores (Cohen's $\kappa \in [0.5, 0.92]$) and lower within-group agreement (Fleiss' $\kappa \in [0.64, 0.71]$).

Table 1. Prevalence of histopathology severity grades associated with natural skin infection by *Pseudogymnoascus destructans*. Prevalence was calculated as the percentage of positive biopsies at the given grade in the dataset of all bats positive for WNS based on histopathology. All infection grades apply to the wing membrane and, aside from fungal skin surface colonisation in the absence of any other findings, are classified as white-nose syndrome (WNS). Tested = number of individuals examined for histopathology. WNS histo+ = number of individuals confirmed positive for WNS. Std. error (%) = $100\sqrt{p(1-p)/n}$, where n is the overall number of scored grades for the given species and p is the proportion of positively scored grades. The respective severity grades are shown in Figs 1 and 2. Numbers in brackets represent WNS pathology grade weighting—see text for details.

Bat taxon data	WNS histo +	Std. error (%)	Prevalence (%)										
			Fungal skin colonisation (1)	Hair follicle infection (2)	Sebaceous gland infection (2)	Single cupping erosion (6)	Multiple cupping erosions (12)	Basement membrane breach (13)	Full thickness infection (19)	Inflammation (20)	Fungal sequestration (−20)	Necrosis (25)	Infarction (30)
Barbastella barbastellus	3	8.67	100	0	0	100	100	100	66.67	66.67	0	66.67	0
Eptesicus nilssonii	3	7.75	100	66.67	66.67	100	100	100	66.67	100	33.33	66.67	0
Miniopterus schreibersii	1	14.5	100	100	100	100	0	100	0	100	0	100	0
Myotis bechsteinii	3	8.21	66.67	33.33	0	100	0	66.67	0	66.67	0	33.33	0
Myotis brandtii	1	8.67	100	100	100	100	100	100	100	100	100	100	0
Myotis dasycneme	13	4.18	100	30.77	15.38	100	38.46	61.54	23.08	84.62	30.77	38.46	0
Myotis daubentonii	10	4.76	90	30	20	100	70	70	50	50	10	70	10
Myotis emarginatus	6	6.15	100	16.67	16.67	100	83.33	66.67	50	50	16.67	50	0
Myotis myotis	50	2.09	96	46	36	100	46	46	2	34	6	24	0
Myotis nattereri	4	7.02	100	25	25	100	25	50	0	25	0	0	0
Plecotus auritus	7	5.69	100	42.86	28.57	100	71.43	71.43	28.57	71.43	28.57	42.86	0
Rhinolophus euryale	1	14.5	0	0	0	100	0	100	0	100	100	0	0
Rhinolophus hipposideros	4	7.25	100	50	50	100	75	100	50	100	0	75	0
Total/mean in Palearctic bats	106		88.67	41.64	35.25	100	54.56	79.41	33.61	72.95	25.03	51.31	0.77
Std. error in Palearctic bats			2.06	4.75	4.46	0.0	4.85	4.75	3.87	4.85	4.71	0.94	3.29
Myotis lucifugus	10	4.7	100	30	30	100	100	100	30	80	30	40	0
Myotis septentrionalis	7	5.62	100	14.29	14.29	100	100	85.71	71.43	71.43	14.29	71.43	0
Total/mean in Nearctic bats	17		100	22.14	22.14	100	100	92.86	50.71	75.72	22.14	55.72	0
Std. error in Nearctic bats			0.0	10.29	10.29	0.0	0.0	5.71	12.11	10.29	12.11	0.0	10.29

https://doi.org/10.1371/journal.pone.0180435.t001

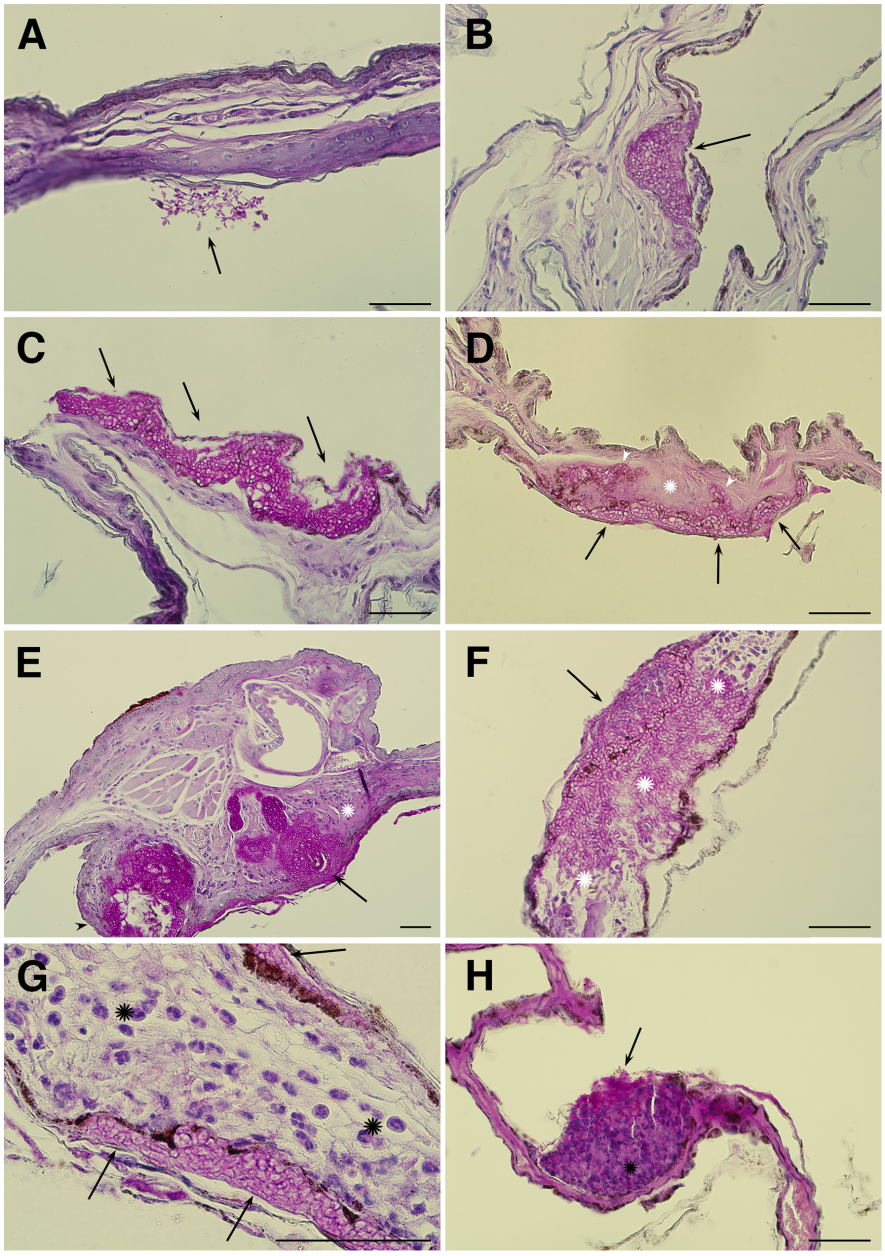


Fig 1. Histopathology grades induced by natural *Pseudogymnoascus destructans* skin infection in Holarctic bats. (A) *Myotis myotis*: fungal skin-surface colonisation with aerial hyphae developing conidia (g_1, black arrow), not classified as

WNS in absence of other findings; (B) *M. myotis*: a single cupping erosion (g_4, black arrow) eroding to the epidermal/dermal interface; (C) *M. myotis*: three confluent cupping erosions (g_5, black arrows); (D) *M. daubentonii*: necrotic wing membrane (witnessed as loss of dermal tissue stainability; g_9, white asterisk) next to multiple cupping erosions packed with *P. destructans* hyphae (g_5, black arrows) that also breached the basement membrane (g_6, white arrowheads); (E) *M. daubentonii*: infection of hair follicle (g_2, black arrow) and associated glands (g_3, black arrowhead). Surface skin colonisation (g_1), multiple cupping erosions (g_5), inflammatory cells (g_8) and necrotic tissue (g_9, white asterisk) are also present in the section; (F) *M. dasycneme*: an outline of a cupping erosion (g_4, black arrow) clearly visible together with full thickness fungal invasion (g_7) replacing the necrotic wing membrane (g_9) and sporadic neutrophils (g_8, white asterisk); (G) *M. lucifugus*: marked inflammatory response (g_8, black asterisks) to fungal invasion of several cupping erosions (g_5, black arrows) on both sides of the wing membrane; (H) *M. dasycneme*: fungal sequestration (g_{11}, black arrow) with neutrophils (g_8, black asterisk) from the wing membrane. Periodic acid-Schiff stain. Scale bar– 50 μm.

https://doi.org/10.1371/journal.pone.0180435.g001

Progression of severe WNS skin infection resulting in extensive skin necrosis and fatality in a *Myotis daubentonii* bat

Progression of severe WNS lesions on the wing membranes of a *M. daubentonii* specimen was investigated at the University of Veterinary and Pharmaceutical Sciences Brno (Czech Republic) Rescue Centre. A time series of UV transillumination and daylight photography images spanning seven days from capture to death indicated that the wing membranes started to show signs of dry necrosis within two days (Fig 4). Flight membrane areas with extensive infection lost tone, elasticity and sheen as they contracted and tore around the WNS lesions in a proximal-to-distal pattern. Histopathology demonstrated wing membrane necrosis associated with confluent cupping-erosion from fungal infection, distended blood vessels with intraluminal neutrophilic infiltration and haemorrhagic infarct consisting of red blood cell and fibrin clots caused by blood flow obstruction (Fig 2).

Healing time series of WNS lesions in a *Myotis myotis* bat

A *M. myotis* bat captured at the end of the hibernation period was diagnosed with pathognomonic WNS skin lesions using UV transillumination (Fig 5A) and histopathology (Fig 1B). The fluorescent yellow-orange WNS lesions disappeared after two weeks at euthermy (Fig 5). Healing progressed until cupping erosion-like structures packed with fungal hyphae were contained within a scab covering the repaired wing membrane (Fig 6). Though this specimen had a histoSum score of 31 (11110001000) on the day of capture, the score had decreased to -20 (00000000001) after 15 days of healing at euthermy.

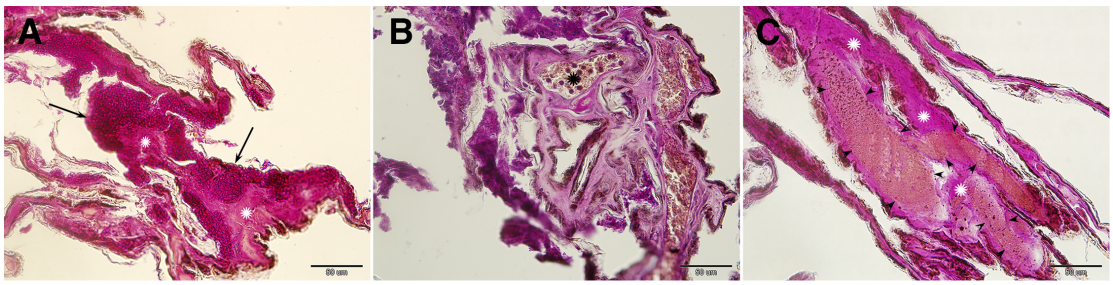


Fig 2. Skin infarction and necrosis associated with progressive white-nose syndrome lesions in *Myotis daubentonii*. Samples for histopathology were collected on Day 7 of the time series documented using UV and daylight photography. Extensive *Pseudogymnoascus destructans* infection of the wing membrane produced confluent cupping erosions (g_5, black arrows) resulting in skin necrosis (g_9, white asterisk), characterised as loss of identifiable skin structures (A). Intraluminal neutrophilic infiltration (g_8, black asterisk) in distended blood vessels was associated with the compromised wing membrane (B). Other skin lesions in this bat included haemorrhagic infarcts (g_{10}, black arrowhead) with stagnant blood and skin necrosis (g_9, white asterisk) (C). Periodic acid-Schiff stain.

https://doi.org/10.1371/journal.pone.0180435.g002

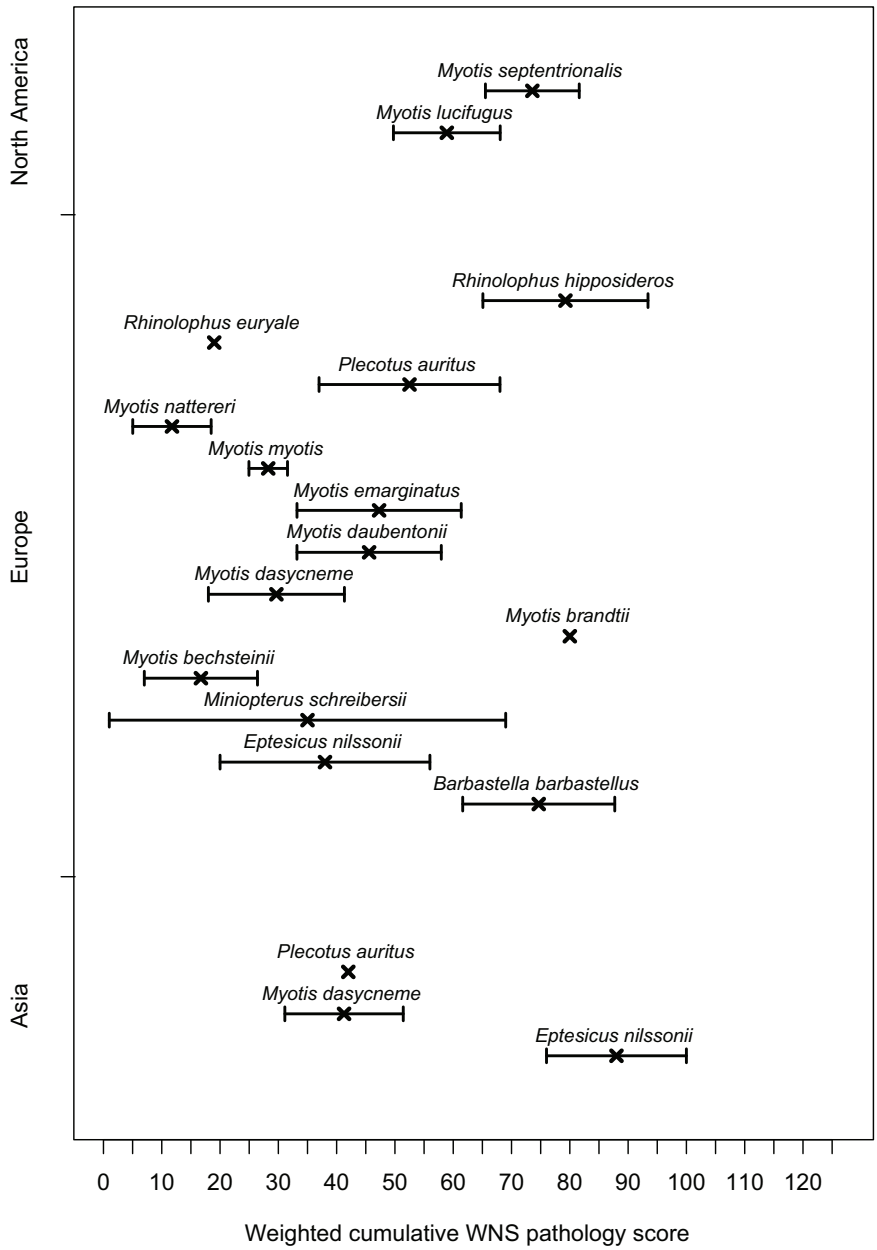


Fig 3. Species-specific weighted cumulative white-nose syndrome pathology score (histoSum). Average sum of weighted qualitative scoring for white-nose syndrome severity grades displayed in Figs 1 and 2 (± std. error). Animals with

histoSum = 1 not classified as positive for WNS on histopathology are included in the figure. Species sampled on multiple continents are presented separately. See Table 1 for sample sizes.

https://doi.org/10.1371/journal.pone.0180435.g003

Functional analysis of WNS pathology scores

Logistic regression (Fig 7A) revealed that the odds of observing surface skin colonization on histopathology increase with increase in fungal load estimated from wing swabs of Palearctic bats (odds ratio: $OR = 2.1$, CI: 1.6–2.8, $p < 0.001$; fungal load was not available for histologically examined Nearctic bats, S2 Table). The OR for observing single cupping erosions dependent on fungal load is lower, but significantly different from 1 ($OR = 2.0$, CI: 1.5–2.6, $p < 0.001$), indicating that single cupping erosions occur in response to roughly one-order-of-magnitude higher fungal loads than that during fungal skin-surface colonisation (Fig 7). In Palearctic bats, wing membrane infection progressed to multiple and/or confluent cupping erosions with fungal loads randomly ($OR = 1.3$, CI: 1–1.6, $p = 0.54$). Investigating both Palearctic and Nearctic bats (S2 Table), a similar pattern was observed in odds of recognizing surface skin colonization ($OR = 10.2$, CI: 4.7–22.2, $p < 0.001$) and single cupping erosions ($OR = 6.6$, CI: 3.5–12.5, $p < 0.001$) on histopathology dependent on the number of UV fluorescent lesions (Fig 7B). The number of UV fluorescent lesions was a better predictor for multiple and/or confluent cupping erosions with $OR = 4.3$ (CI: 2.6–7.2, $p < 0.001$) than fungal load.

Cumulative WNS pathology scores from single 4 mm biopsies increased with the number of UV fluorescent lesions observed on the surface of the whole wing ($\beta_0 = 62.1$, $\beta_1 = 16.41$; Fig 8), indicating that species with more extensive UV fluorescence tend toward multiple concurrent WNS findings on histopathology.

Discussion

Fuelled by no reports of mass die-offs from Europe [36,37], the most erroneous conclusion concerning white-nose syndrome is the belief that *P. destructans* infection causes no harm to Palearctic bats. *Pseudogymnoascus destructans* infection is known to induce complex physiological

Table 2. Inter-rater agreement in scoring white-nose syndrome (WNS) pathology according to Fleiss' κ. Experienced pathologists ($n = 5$) scored 30 photographs of randomly drawn histopathology slides, each group of naïve raters ($n \in \{27, 20, 25\}$) scored a subset of 10 photographs. Standard errors for the given sample sizes were 0.047, 0.017, 0.023 and 0.018, respectively. Negative κ values indicate no inter-rater agreement.

Grade index	WNS pathology grade	Fleiss' κ			
		experienced	naïve group 1	naïve group 2	naïve group 3
1	Fungal skin colonisation	0.671	0.560	0.114	0.014
2	Hair follicle infection	0.931	0.690	0.772	0.850
3	Sebaceous gland infection	0.712	0.074	0.575	0.839
4	Single cupping erosion	0.790	0.521	0.200	0.655
5	Multiple cupping erosions	0.861	0.716	0.445	0.596
6	Basement membrane breach	0.496	0.592	0.237	0.285
7	Full thickness infection	0.704	0.655	0.330	0.809
8	Inflammation	0.469	0.436	0.368	0.392
9	Skin necrosis	0.460	0.293	0.414	0.288
10	Skin infarction	0.851	-0.007	0.415	-0.008
11	Fungal sequestration	0.805	0.755	-0.032	0.159

https://doi.org/10.1371/journal.pone.0180435.t002

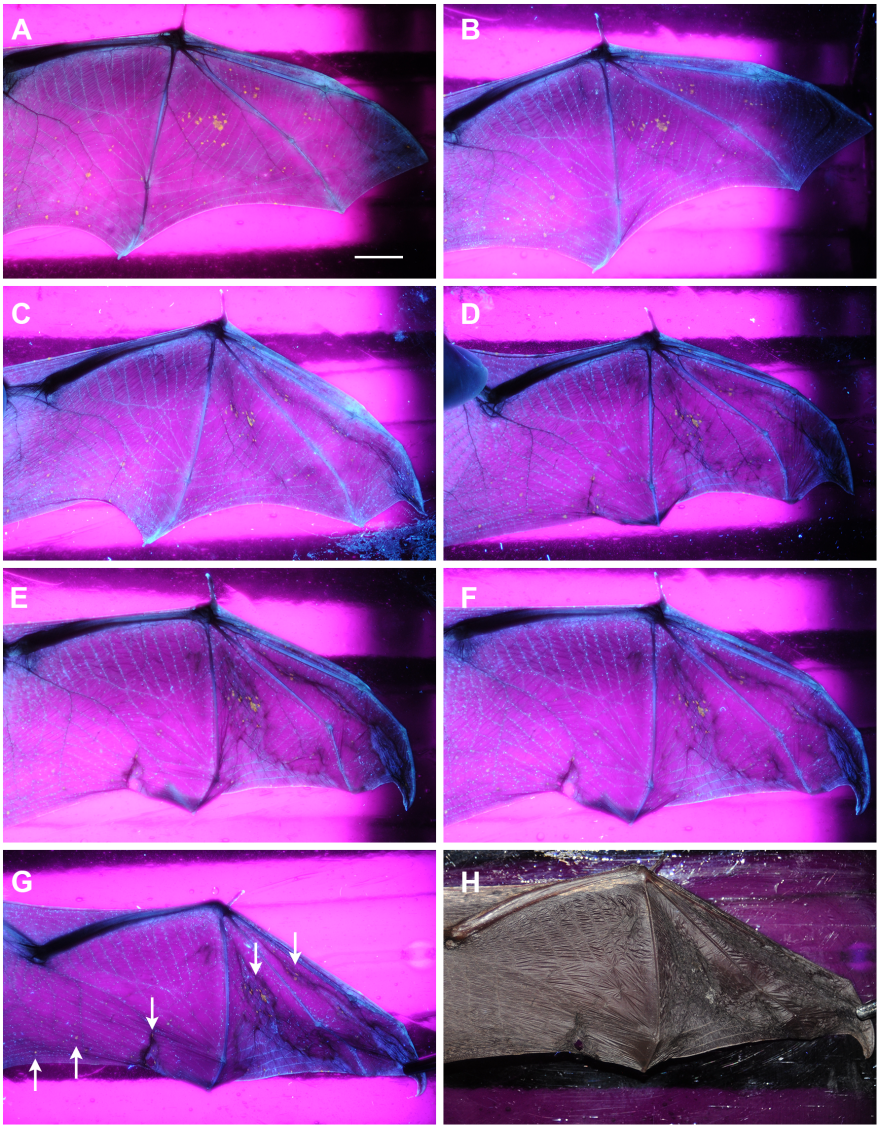


Fig 4. Time series of a *Myotis daubentonii* wing showing progression of white-nose syndrome lesions to fatality. Extensive white-nose syndrome infection was recognised on a *M. daubentonii* bat at a hibernaculum in the Podyjí National Park (Czech Republic). The bat was kept under euthermic conditions at a rescue centre, fed *ad libitum* and supplied with drinking water. The wing was extended over a Wood's lamp at 366 nm wavelength and photographed in a darkroom. (A) = Day 0, (B) = Day 1, (C) = Day 2, (D) = Day 3, (E) = Day 5, (F) = Day 6, (G) = Day 7, white arrows indicate biopsy punch sites (results presented in Fig 5). Day 7 was also documented using daylight photography (H). Scale bar = 1 cm. This time series spanned seven days from capture at the hibernaculum to death in the rescue centre. Wing membrane areas with extensive *Pseudogymnoascus destructans* infection became dry and necrotic within two days of euthermy, whereupon they contracted

and tore around the white-nose syndrome lesions in a proximal-to-distal pattern. The animal displayed loss of skin tone, elasticity and surface sheen, and ceased eating one day prior to death.

https://doi.org/10.1371/journal.pone.0180435.g004

and transcriptional effects in hibernating bats long before the onset of the clinical signs indicative of late-stage WNS [19,38]. Palearctic and Nearctic bats are exposed to similar fungal loads, resulting in equivalent focal skin-tissue invasiveness pathognomonic for WNS lesions [7,11–13,39,40]. It would be reasonable, therefore, to expect both morbidity effects (not always recognisable in the field) and mortality across the distribution range of *P. destructans*. Here, we show the full range of WNS skin pathology in Palearctic and Nearctic bats. Moreover, we document a case of *P. destructans* infection in a European *M. daubentonii* bat progressing to fatality due to extensive skin necrosis and infarction. The differential outcome of WNS in Eurasia and North America appears to be associated with some form of tolerance mechanism [11]. Our data, however, suggests that Palearctic bat tolerance to WNS infection may be limited by severity of wing damage.

In this paper, we presented a novel pathogenesis-based grading system for *P. destructans* skin infection that utilises non-lethal biopsy sampling to examine the WNS lesions resulting from host-pathogen interaction. The severity grades chosen are simple to differentiate on histopathology and show good agreement amongst experienced raters. Identification of inflammation and skin necrosis exhibited lowest agreement and requires careful consideration in multi-rater comparisons. Naïve, minimally trained, raters showed similar agreement with scoring most grades as did experienced raters, but fungal skin colonization, sebaceous gland infection, skin infarction and fungal sequestration were most challenging grades for this group (Table 2). Generally speaking, histopathology enables one to assign nominal diagnostic categories to medical conditions observed in organs and tissues. Grading may provide additional

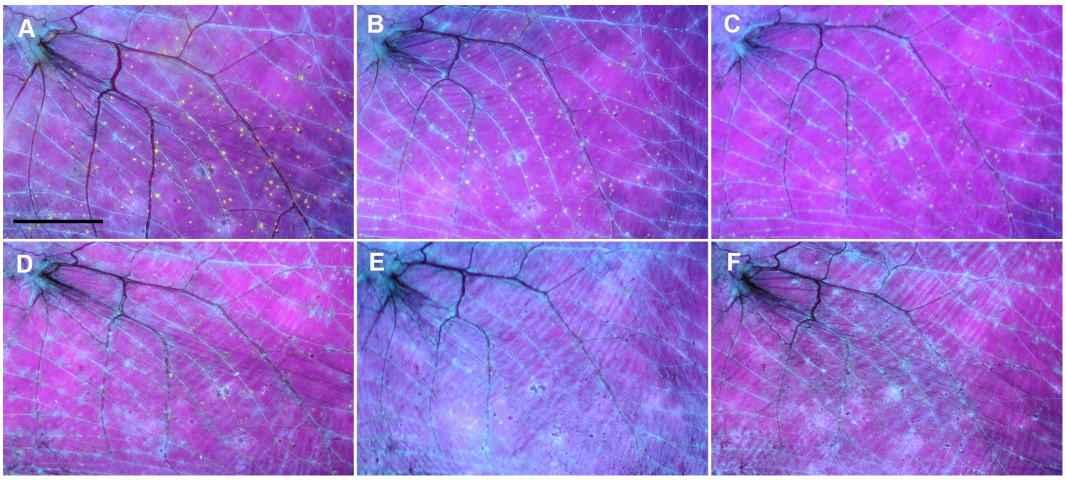


Fig 5. UV transillumination time series of a *Myotis myotis* wing with decrease in fluorescence corresponding to white-nose syndrome lesions over time. The bat was captured at the end of the hibernation period, kept in captivity at euthermy, fed *ad libitum* with cockroaches and mealworms and supplied with drinking water, and released after the white-nose syndrome lesions had healed. The wing was extended over a Wood's lamp at 366 nm wavelength and photographed in a darkroom. **A** = Day 0, **B** = Day 3, **C** = Day 5, **D** = Day 7, **E** = Day 11, **F** = Day 15. The top-left corner matches in each image, wing deformation is due to variable handling of the live animal during sampling. Scale bar = 1 cm.

https://doi.org/10.1371/journal.pone.0180435.g005

Fig 6. Healing of a white-nose syndrome lesion in *Myotis myotis*. Samples for histopathology were collected on Day 15 of the time series described in Fig 5. A cupping erosion-like structure (black arrow) packed with fungal hyphae within a scab covering the healing wing membrane of *M. myotis*. Periodic acid-Schiff stain.

https://doi.org/10.1371/journal.pone.0180435.g006

data allowing one to estimate severity, predict biological behaviour of the disease, predict survival rate and make decisions regarding patient prognosis and treatment [41]. Recent studies linking histopathology WNS severity scores with frequent arousal from hibernation and mortality [14] and altered physiological homeostasis [18–20,42] have used destructive methods for examining bats. While such methods for examining wing membrane damage score WNS severity only at the moment of death or euthanasia [14,18–20], our grading system allows one to follow temporal patterns of disease progression with consecutive multiple biopsies targeted with UV transillumination. UV fluorescence-guided non-lethal biopsy diagnostics of WNS skin showed 95.5% sensitivity and 100% specificity [22]. Collection of additional biopsies from each specimen may further increase diagnostic and grading sensitivity.

Direct comparison between histological WNS severity scoring from the whole wing membrane, which provides a reasonable representation of skin damage associated with *P. destructans* infection [14,20], and our proposed single non-lethal biopsy method is complicated. Instead, we suggest that the number of bats positive for WNS lesions on histopathology ($n = 123$) and the number of taxa ($n = 15$) from the Palearctic and Nearctic regions examined, together with the non-lethal nature of the methodology, make this grading system universally applicable to all species exposed to the infection. The ability to select a biopsy site on the bat's wing displaying the most representative and extensive UV fluorescence [22] enables a reasonable assessment of pathology while reducing the impact of WNS surveillance on the bat.

Previous WNS grading schemes have proposed four or five WNS-positive grades [14,20], with presence, extent and distribution of skin lesions densely packed with fungal hyphae (cupping erosions) used as the main criteria for assigning whole-wing-based scores [14]. As an alternative, we propose a finer, semi-quantitative 11-grade scale. Analogous quantitative measures of *P. destructans* skin infection in this alternative grading system include presence of

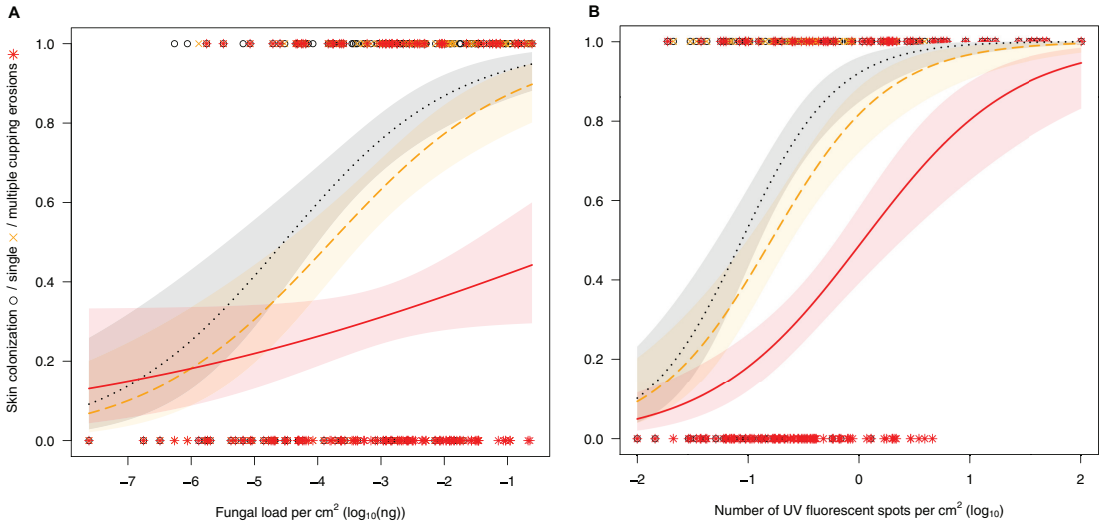


Fig 7. Relationship between data from non-destructive diagnostic methods and selected histopathology severity grades. Logistic regressions of skin surface colonisation by the fungus (black circle, dotted line; coefficients: (A) $\beta_0 = 3.38$, $\beta_1 = 0.74$, (B) $\beta_0 = 2.45$, $\beta_1 = 2.32$), single cupping erosion (orange cross, dashed line; coefficients: (A) $\beta_0 = 2.60$, $\beta_1 = 0.68$, (B) $\beta_0 = 1.49$, $\beta_1 = 1.88$) and multiple cupping erosions (red star, solid line; coefficients: (A) $\beta_0 = -0.09$, $\beta_1 = 0.24$, (B) $\beta_0 = -0.06$, $\beta_1 = 1.44$) dependent on fungal load detected on qPCR in Palearctic bats (A) or number of UV fluorescent spots in Holarctic bats (B). Shaded area represents 95% confidence interval on predicted probability.

https://doi.org/10.1371/journal.pone.0180435.g007

single, multiple and/or confluent cupping erosions within a 4 mm punch biopsy that can be evaluated through image analysis to address invasiveness and size [11], then approximated for the whole wing from UV photographs [22]. Expanding the previous systems towards more severe pathology means that invasive fungal growth penetrating the full-thickness of the wing membrane represents a higher grade of skin infection than cup-like fungal structures eroding to the epidermal/dermal interface only [12,13]. In addition to cupping erosions, the ability of *P. destructans* to invade the dermis and the full thickness of the wing membrane distinguishes WNS from other dermatophytic infections.

As dysregulated immunity may contribute to severe post-emergent pathology [43], we also included bat immune response to infection (inflammation and fungal sequestration) as index criteria in our grading system. Inflammatory responses were frequently associated with fungal invasion in both Palearctic and Nearctic bats, probably due to sampling in the late-hibernation or early post-hibernation periods. However, the ability of bats to contain the infection through sequestration of the fungus with neutrophils and skin debris was only observed in a quarter of bats showing an inflammatory response. Warnecke et al. [20] scored bat tissue replaced by the fungus as necrotic. Unlike their indirect indication, we considered skin necrosis as findings characterised by loss of staining pattern, indistinct outlines of individual bat tissue cells and hypereosinophilia (Figs 1D and 2A). Skin infarction, which has previously been recognised in association with WNS [16], accounted for the most severe wing membrane damage (Fig 2C). In future studies, therefore, we urge that WNS severity be assessed with respect to skin infarction, as wing membrane damage in such cases extended from WNS skin lesions to terminal wing areas affected with blood flow obstruction (cf. Fig 4).

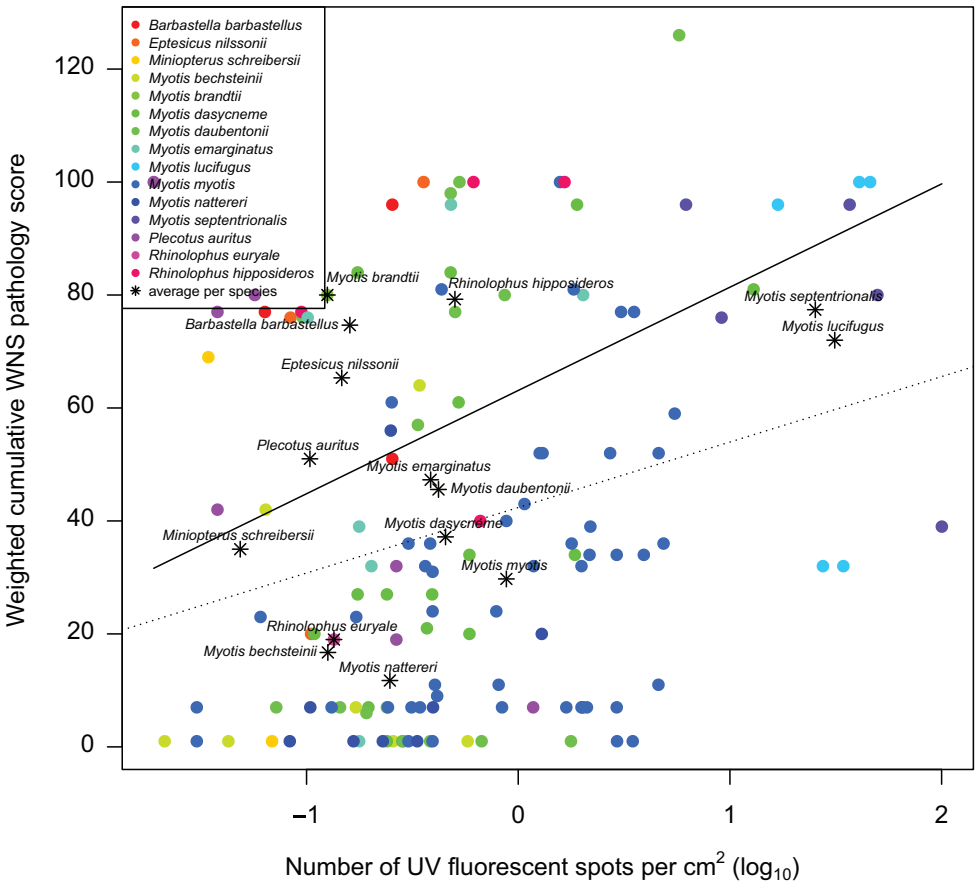


Fig 8. Increase in weighted cumulative white-nose syndrome pathology score with progressing UV fluorescence. Phylogenetic generalised least-squares accounts for phylogeny and intraspecific variability in evaluating the relationship between weighted cumulative white-nose syndrome pathology score dependent on unit number of UV fluorescent lesions ($\beta_0 = 62.1$, $\beta_1 = 16.41$; solid line). Linear regression without phylogenetic correction ($\alpha = 42.38$, $\beta = 11.59$, $F_{1,131} = 9.12$, $p = 0.003$, $r^2 = 0.07$; dotted line).

https://doi.org/10.1371/journal.pone.0180435.g008

While some hibernating bats are able to recover from fungal infection [44,45], extensive skin damage in WNS survivors may alter flight performance, foraging efficiency and metabolism during the post-emergent season [16,43,46,47]. In line with the ecological and evolutionary trade-off concept [48], mounting an immune response against *P. destructans* infection, and the necessity to repair the damaged wing membrane, represent physiological costs that may affect the bat's fitness and survival. To address the series of healing from different grades of skin damage associated with WNS, our histopathology findings in conjunction with the case reports suggest that cupping erosions heal through marked neutrophilic inflammation (Fig 1G), sequestration of the agent from the skin (Fig 1H) and re-epithelialization (Figs 1H and 6). As shown in Fig 6, the outcome of healing is regeneration of the skin, suggesting a return to

full function. On the other hand, tissue necrosis (Figs 1D and 2A) and full-thickness skin infection (Fig 1F) probably need more remodelling and may result in wing membrane scarring. Haemorrhagic infarcts (Fig 2C) induce larger flight membrane defects (Fig 4) and may result in tears that stretch to the wing margin (Fig 4E, 4F, 4G and 4H).

Case reports on infection time-series in *M. daubentonii* and *M. myotis* document the dichotomy of development that may occur in European bats infected with *P. destructans*. Our findings indicate that the severity of invasion and associated wing damage determine the clinical outcome in terms of morbidity and/or mortality. Importantly, the *M. daubentonii* in our case study died during the early post-hibernation period, similar to observations by Meteyer et al. [43]. Sporadic cases of mortality in European bats due to *P. destructans* infection may go undetected, therefore, outside of hibernacula. From a practical diagnostic point of view, it is necessary to bear in mind that WNS-specific UV fluorescence [22] disappears within two to three weeks of euthermy.

Quantified per-unit wing membrane area fungal load correlates with disease intensity measured as the number of WNS lesions detected under UV transillumination [11]. It has been hypothesised that *P. destructans* hyphae are more likely to invade deeper skin layers and produce WNS lesions with heavy fungal growth on bat wings [11]. In fact, our data documents progression of fungal infection to WNS pathognomonic single and multiple cupping erosions in response to one- and several-orders-of-magnitude higher wing membrane fungal loads, respectively, compared with the early stage of infection characterised by skin surface colonisation (cf. Fig 7A). For this chronic skin infection, severity-grading scores were understandably higher in bats with more extensive UV fluorescence (Fig 8). The UV fluorescence associated with WNS is caused by riboflavin, a fungal secondary metabolite, that accumulates in skin lesions and results in cytotoxic damage of adjacent tissues [23] (as reflected by the weighted cumulative WNS pathology score). All three diagnostic methods (i.e. qPCR, UV transillumination and histology), when used in a quantitative or semi-quantitative manner, proved useful for discrimination between colonisation of bats exposed to the pathogen only and those with the invasive skin infection known as WNS.

Comparative findings from Palearctic and Nearctic bats infected with *P. destructans* show differences in prevalence, fungal load and qualitative [11] and semi-quantitative histopathology (this study). To gain greater insight into WNS severity, it is imperative that complex data are collected on infection intensity in each bat. Our proposed biopsy-based grading system for WNS is suitable for this purpose as histopathology reveals degree of invasiveness in terms of skin layers invaded, along with associated damage; qPCR provides data on pathogen load, and image analysis of photographs taken under UV light quantifies total affected wing/body surface area. The ease of diagnostic scoring using our system allows open, effective WNS surveillance across Europe, as well as comparative studies across the Holarctic. The use of uniform criteria and training sessions in WNS histopathology can substantially improve consensus rate in diagnosis of the disease among pathologists and promote its universality. Undertaken alongside UV trans-illumination, which enables targeted biopsy, our WNS severity-grading system provides an attractive option for use with conservation-sensitive bat species.

Supporting information

S1 Table. Species sample sizes. *n*–number of individuals per species undergoing histopathological examination, WNS histo+–number of individuals confirmed positive for white-nose syndrome.
(PDF)

S2 Table. Data on fungal load, ultra-violet fluorescence and histopathology in Holarctic bats. Fungal load and UV fluorescence are given on $\log_{10}$ scale per cm^2 of wing area. Histopathology is reported as a binary code for index signs of grades corresponding to those listed in Methods and Table 2. NA–not available, CZ–Czech Republic, LV–Latvia, PL–Poland, RU–Russia, SI–Slovenia, histoSum–weighted cumulative WNS pathology score.
(TXT)

Acknowledgments

This study was supported by the Czech Science Foundation (Grant No. P506-12-1064 and 17-20286S). This research was carried out as part of the CEITEC 2020 project (LQ1601), with further financial support from the Ministry of Education, Youth and Sports of the Czech Republic under National Sustainability Programme II. We thank the following students of the Veterinary Study Programme at the University of Veterinary and Pharmaceutical Sciences Brno for volunteering to participate in the study: Martina Balážová, Tereza Bartošová, Tereza Bělková, Martina Benešová, Miriam Birošová, Kateřina Bohatá, Denisa Bohušová, Klára Borýsková, Petra Buchníčková, Anna Časová, Ondřej Daněk, Lenka Danielová, Anna Drahoňovská, Marie Dvořáková, Anna Frumarová, Nella Fuchsová, Lenka Horáková, Václav Hůlka, Patricie Janíčková, Martina Jasanská, Miroslava Juranová, Lenka Kapustová, Lenka Klimešová, Magda Kohoutová, Katarína Kopálová, Lucie Košťálová, Tereza Kovářová, Jan Kovol, Barbora Krejčí, Jan Krejčí, Linda Kubaštová, Marie Kubátová, Veronika Kvaková, Petra Laníková, Barbora Löfflerová, Dobromila Malíková, Lenka Malinová, Jana Michalcová, Alena Míčková, Kristína Miklošovičová, Kateřina Musálková, Michaela Petríková, Kateřina Podrábská, Michaela Prokešová, Polina Rapekta, Nelly Reisová, Karel Šlajs, Jiří Slavík, Markéta Sosňáková, Ivana Šťáhlavská, Veronika Stařecká, Natálie Štefaňáková, Václav Štellar, Soňa Struhárová, Monika Šubrtová, Ondřej Táborský, Ivana Timová, Klára Tlačbabová, Anita Turanská, Juraj Turňa, Karel Tvrdoň, Gabriela Vacková, Daniela Valvodová, Martina Vavřincová, Lenka Večerková, Lucie Veselková, Lenka Vojtěchovská, Michaela Vojtková, Zuzana Voláková, Jakub Záleský, Henrieta Zbořilová and Anna Zemanová.

Author Contributions

Conceptualization: Jiri Pikula, Sybill K. Amelon.

Formal analysis: Natália Martínková.

Funding acquisition: Jiri Pikula, Natália Martínková.

Investigation: Jiri Pikula, Hana Bandouchova, Veronika Kovacova, Petr Linhart, Vladimir Piacek, Natália Martínková.

Methodology: Jiri Pikula, Hana Bandouchova, Natália Martínková.

Resources: Jiri Pikula, Sybill K. Amelon, Hana Bandouchova, Tomáš Bartonička, Hana Berkova, Jiri Brichta, Sarah Hooper, Tomasz Kokurewicz, Miroslav Kolarik, Bernd Köllner, Veronika Kovacova, Petr Linhart, Vladimir Piacek, Gregory G. Turner, Jan Zukal, Natália Martínková.

Writing – original draft: Jiri Pikula, Natália Martínková.

Writing – review & editing: Jiri Pikula, Sybill K. Amelon, Hana Bandouchova, Tomáš Bartonička, Jiri Brichta, Sarah Hooper, Tomasz Kokurewicz, Miroslav Kolarik, Veronika Kovacova, Vladimir Piacek, Jan Zukal, Natália Martínková.

References

1. Deem SL, Karesh WB, Weisman W. Putting theory into practice: Wildlife health in conservation. Conserv Biol. 2001; 15: 1224–1233.
2. Fey SB, Siepielski AM, Nusslé S, Cervantes-Yoshida K, Hwan JL, et al. Recent shifts in the occurrence, cause, and magnitude of animal mass mortality events. Proc Natl Acad Sci USA. 2015; 112: 1083–1088. https://doi.org/10.1073/pnas.1414894112 PMID: 25583498
3. Blehert DS, Hicks AC, Behr M, Meteyer CU, Berlowski-Zier BM, et al. Bat white-nose syndrome: An emerging fungal pathogen? Science. 2009; 323: 227. https://doi.org/10.1126/science.1163874 PMID: 18974316
4. Gargas A, Trest MT, Christensen M, Volk TJ, Blehert DS. *Geomyces destructans* sp. nov. associated with bat white-nose syndrome. Mycotaxon. 2009; 108: 147–154.
5. Lorch JM, Meteyer CU, Behr MJ, Boyles JG, Cryan PM, et al. Experimental infection of bats with *Geomyces destructans* causes white-nose syndrome. Nature. 2011; 480: 376–378. https://doi.org/10.1038/nature10590 PMID: 22031324
6. Minnis AM, Lindner DL. Phylogenetic evaluation of *Geomyces* and allies reveals no close relatives of *Pseudogymnoascus destructans*, comb. nov., in bat hibernacula of eastern North America. Fungal Biol. 2013; 117: 638–649. https://doi.org/10.1016/j.funbio.2013.07.001 PMID: 24012303
7. Zukal J, Bandouchova H, Bartonicka T, Berkova H, Brack V, et al. White-nose syndrome fungus: A generalist pathogen of hibernating bats. PLoS ONE. 2014; 9: e97224. https://doi.org/10.1371/journal.pone.0097224 PMID: 24820101
8. Frick WF, Pollock JF, Hicks AC, Langwig KE, Reynolds DS, et al. An emerging disease causes regional population collapse of a common North American bat species. Science. 2010; 329: 679–682. https://doi.org/10.1126/science.1188594 PMID: 20689016
9. Turner GG, Reeder DM, Coleman JTH. A five-year assessment of mortality and geographic spread of white-nose syndrome in North American bats and a look to the future. Bat Research News. 2011; 52: 13–27.
10. Coleman JTH, Reichard JD. Bat white-nose syndrome in 2014: A brief assessment seven years after discovery of a virulent fungal pathogen in North America. Outlooks on Pest Management. 2014; 25: 374–377.
11. Zukal J, Bandouchova H, Brichta J, Cmokova A, Jaron KS, et al. White-nose syndrome without borders: *Pseudogymnoascus destructans* infection tolerated in Europe and Palearctic Asia but not in North America. Sci Reports. 2016; 6: 19829.
12. Bandouchova H, Bartonicka T, Berkova H, Brichta J, Cerny J, et al. *Pseudogymnoascus destructans*: Evidence of virulent skin invasion for bats under natural conditions, Europe. Transbound Emerg Dis. 2015; 62: 1–5. https://doi.org/10.1111/tbed.12282 PMID: 25268034
13. Meteyer CU, Buckles EL, Blehert DS, Hicks AC, Green DE, et al. Histopathologic criteria to confirm white-nose syndrome in bats. J Vet Diagn Invest. 2009; 21: 411–414. https://doi.org/10.1177/104063870902100401 PMID: 19564488
14. Reeder DM, Frank CL, Turner GG, Meteyer CU, Kurta A, et al. Frequent arousal from hibernation linked to severity of infection and mortality in bats with white-nose syndrome. PLoS ONE. 2012; 7: e38920. https://doi.org/10.1371/journal.pone.0038920 PMID: 22745688
15. Warnecke L, Turner JM, Bollinger TK, Lorch JM, Misra V, et al. Inoculation of bats with European *Geomyces destructans* supports the novel pathogen hypothesis for the origin of white-nose syndrome. Proc Natl Acad Sci USA. 2012; 109: 6999–7003. https://doi.org/10.1073/pnas.1200374109 PMID: 22493237
16. Cryan P, Meteyer C, Boyles J, Blehert D. Wing pathology of white-nose syndrome in bats suggests life-threatening disruption of physiology. BMC Biol. 2010; 8: 135. https://doi.org/10.1186/1741-7007-8-135 PMID: 21070683
17. Cryan P, Meteyer C, Boyles J, Blehert D. White-nose syndrome in bats: illuminating the darkness. BMC Biol. 2013; 11: 47. https://doi.org/10.1186/1741-7007-11-47 PMID: 23587401
18. Cryan PM, Meteyer CU, Blehert DS, Lorch JM, Reeder DM, et al. Electrolyte depletion in white-nose syndrome bats. J Wildlife Dis. 2013; 49: 398–402.
19. Verant ML, Carol MU, Speakman JR, Cryan PM, Lorch JM, et al. White-nose syndrome initiates a cascade of physiologic disturbances in the hibernating bat host. BMC Physiol. 2014; 14: 10. https://doi.org/10.1186/s12899-014-0010-4 PMID: 25487871
20. Warnecke L, Turner JM, Bollinger TK, Misra V, Cryan PM, et al. Pathophysiology of white-nose syndrome in bats: A mechanistic model linking wing damage to mortality. Biol Letters. 2013; 9: 20130177.
21. Hoyt JR, Sun K, Parise KL, Lu G, Langwig KE, et al. Widespread bat white-nose syndrome fungus, Northeastern China. Emerg Inf Dis. 2016; 22: 140.

22. Turner GG, Meteyer CU, Barton H, Gumbs JF, Reeder DM, et al. Nonlethal screening of bat-wing skin with the use of ultraviolet fluorescence to detect lesions indicative of white-nose syndrome. J Wildlife Dis. 2014; 50: 566–573.
23. Flieger M, Bandouchova H, Cerny J, Chudíčková M, Kolarik M, et al. Vitamin B_2 as a virulence factor in *Pseudogymnoascus destructans* skin infection. Sci Reports. 2016; 6: 33200.
24. Court MH, Robbins AH, Whitford AM, Beck EV, Tseng FS, et al. Pharmacokinetics of terbinafine in little brown myotis (*Myotis lucifugus*) infected with *Pseudogymnoascus destructans*. Am J Vet Res. 2017; 78: 90–99. https://doi.org/10.2460/ajvr.78.1.90 PMID: 28029293
25. Shuey MM, Drees KP, Lindner DL, Keim P, Foster JT. Highly sensitive quantitative PCR for the detection and differentiation of P*seudogymnoascus destructans* and other *Pseudogymnoascus* species. Appl Environ Microb. 2014; 80: 1726–1731.
26. Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–9. 2015. Available from: https://CRAN.R-project.org/package=lme4.
27. R Core Team. R: A language and environment for statistical computing. 2016. Available from: http://www.R-project.org/.
28. Rowlingson B, Diggle P. splancs: Spatial and space-time point pattern analysis. R package version 2.01–38. 2015. Available from: https://CRAN.R-project.org/package=splancs.
29. Urbanek S. jpeg: Read and write JPEG images. R package version 0.1–8. 2014. Available from: https://CRAN.R-project.org/package=jpeg.
30. Ives AR, Midford PE, Garland T Jr. Within-species variation and measurement error in phylogenetic comparative methods. Syst Biol. 2007; 56: 252–270. https://doi.org/10.1080/10635150701313830 PMID: 17464881
31. Revell LJ. *phytools*: An R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012; 3: 217–223.
32. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960; 20: 37–46.
33. McHugh ML. Interrater reliability: the kappa statistic. Biochem Medica. 2012; 22: 276–282.
34. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971; 76: 378–382.
35. Gamer M, Lemon J, Singh IFP. irr: Various coefficients of interrater reliability and agreement. R package version 0.84. 2012. Available from: https://CRAN.R-project.org/package=irr.
36. Martínková N, Bačkor P, Bartonička T, Blažková P, Červený J, et al. Increasing incidence of *Geomyces destructans* fungus in bats from the Czech Republic and Slovakia. PLoS ONE. 2010; 5: e13853. https://doi.org/10.1371/journal.pone.0013853 PMID: 21079781
37. Puechmaille SJ, Wibbelt G, Korn V, Fuller H, Forget F, et al. Pan-European distribution of white-nose syndrome fungus (*Geomyces destructans*) not associated with mass mortality. PLoS ONE. 2011; 6: e19167. https://doi.org/10.1371/journal.pone.0019167 PMID: 21556356
38. Field KA, Johnson JS, Lilley TM, Reeder SM, Rogers EJ, et al. The white-nose syndrome transcriptome: Activation of anti-fungal host responses in wing tissue of hibernating little brown *Myotis*. PLoS Pathog. 2015; 11: e1005168. https://doi.org/10.1371/journal.ppat.1005168 PMID: 26426272
39. Pikula J, Bandouchova H, Novotný L, Meteyer CU, Zukal J, et al. Histopathology confirms white-nose syndrome in bats in Europe. J Wildlife Dis. 2012; 48: 207–211.
40. Lucan RK, Bandouchova H, Bartonicka T, Pikula J, Zahradnikova A Jr., et al. Ectoparasites may serve as vectors for the white-nose syndrome fungus. Parasit Vectors 2016; 9: 16. https://doi.org/10.1186/s13071-016-1302-2 PMID: 26762515
41. Cross SS, Benes K, Stephenson TJ, Harrison RF. Grading in histopathology. Diagnostic Histopathology. 2011; 17: 263–267.
42. McGuire LP, Turner JM, Warnecke L, McGregor G, Bollinger TK, et al. White-nose syndrome disease severity and a comparison of diagnostic methods. EcoHealth. 2016; 13: 60–71. https://doi.org/10.1007/s10393-016-1107-y PMID: 26957435
43. Meteyer CU, Barber D, Mandl JN. Pathology in euthermic bats with white nose syndrome suggests a natural manifestation of immune reconstitution inflammatory syndrome. Virulence. 2012; 3: 583–588. https://doi.org/10.4161/viru.22330 PMID: 23154286
44. Meteyer CU, Valent M, Kashmer J, Buckles EL, Lorch JM, et al. Recovery of little brown bats (*Myotis lucifugus*) from natural infection with *Geomyces destructans*, white-nose syndrome. J Wildl Dis. 2011; 47: 618–626. https://doi.org/10.7589/0090-3558-47.3.618 PMID: 21719826
45. Fuller NW, Reichard JD, Nabhan ML, Fellows SR, Pepin LC, et al. Free-ranging little brown myotis (*Myotis lucifugus*) heal from wing damage associated with white-nose syndrome. Ecohealth. 2011; 8: 154–162. https://doi.org/10.1007/s10393-011-0705-y PMID: 21922344

46. Reichard JD, Kunz TH. White-nose syndrome inflicts lasting injuries to the wings of little brown myotis (*Myotis lucifugus*). Acta Chiropterol. 2009; 11: 457–464.

47. Voigt CC. Bat flight with bad wings: is flight metabolism affected by damaged wings? J Exp Biol. 2013; 216: 1516–1521. https://doi.org/10.1242/jeb.079509 PMID: 23348945

48. Lochmiller RL, Deerenberg C. Trade-offs in evolutionary immunology: just what is the cost of immunity? Oikos. 2000; 88: 87–98.

Paper 2.5.1

Jaron K. S., Moravec J. C., **Martínková N.** 2014. SigHunt: Horizontal gene transfer finder optimized for eukaryotic genomes. *Bioinformatics* 30: 1081-1086.

BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 8 2014, pages 1081–1086
doi:10.1093/bioinformatics/btt727

Sequence analysis

Advance Access publication December 25, 2013

SigHunt: horizontal gene transfer finder optimized for eukaryotic genomes

Kamil S. Jaron[1,*], Jiří C. Moravec[1] and Natália Martínková[1,2,*]

[1]Institute of Biostatistics and Analyses, Masaryk University and [2]Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, Brno, Czech Republic

Associate Editor: John Hancock

ABSTRACT

Motivation: Genomic islands (GIs) are DNA fragments incorporated into a genome through horizontal gene transfer (also called lateral gene transfer), often with functions novel for a given organism. While methods for their detection are well researched in prokaryotes, the complexity of eukaryotic genomes makes direct utilization of these methods unreliable, and so labour-intensive phylogenetic searches are used instead.

Results: We present a surrogate method that investigates nucleotide base composition of the DNA sequence in a eukaryotic genome and identifies putative GIs. We calculate a genomic signature as a vector of tetranucleotide (*4-mer*) frequencies using a sliding window approach. Extending the neighbourhood of the sliding window, we establish a local kernel density estimate of the *4-mer* frequency. We score the number of *4-mer* frequencies in the sliding window that deviate from the credibility interval of their local genomic density using a newly developed discrete interval accumulative score (DIAS). To further improve the effectiveness of DIAS, we select informative *4-mers* in a range of organisms using the tetranucleotide quality score developed herein. We show that the SigHunt method is computationally efficient and able to detect GIs in eukaryotic genomes that represent non-ameliorated integration. Thus, it is suited to scanning for change in organisms with different DNA composition.

Availability and implementation: Source code and scripts freely available for download at http://www.iba.muni.cz/index-en.php?pg=research–data-analysis-tools–sighunt are implemented in C and R and are platform-independent.

Contact: 376090@mail.muni.cz or martinkova@ivb.cz

Received on August 9, 2013; revised on November 18, 2013; accepted on December 9, 2013

1 INTRODUCTION

Horizontal gene transfer (HGT) occurs when a DNA sequence passes between organisms otherwise than by reproductive descent. It results in a relationship of orthologous sequences that is not tree-like and contains reticulations. Notorious examples include antibiotic resistance plasmids transferred between bacterial strains (Freeman, 1951), pathogenicity islands (Friesen *et al.*, 2006), incorporation of retroviruses (Jern and Coffin, 2008) and artificial HGT in the forms of genetically modified organisms (Wolfenbarger and Phifer, 2000). When a horizontally transferred gene becomes fixed in a population, it is termed a genomic island (GI). Relatively frequent HGT occurs between organisms of similar complexity, such as between prokaryotes, but successful HGT between domains and kingdoms is also known. Incorporation of the alien sequence into a recipient genome must be compatible with survival of the cell; it should not, for example, knock out an essential gene. In eukaryotes, distortion of the open reading frame with HGT is less likely due to the sparseness of coding sequences; yet alien genes face molecular biological limitations relating to metabolism in the recipient organism. When genes are transferred from prokaryotes to eukaryotes, the genetic code difference might hinder correct protein translation. In cases of HGT between eukaryotes, incorrect intron splicing would render the gene product altered and potentially dysfunctional. Nevertheless, successful implementation of HGT in a suitable place within the genome could result in expression of the relevant protein. Proteins encoded in a GI that could be expressed in the recipient organism might provide a novel, highly adaptive function (Casacuberta and Gonzlez, 2013; Schönknecht *et al.*, 2013). To find such a GI is to discover exciting information that is often transformative for the given research field.

HGT detection is well studied in prokaryotes, having started from measuring variability of oligonucleotide frequency along the genome (Karlin and Burge, 1995). Eukaryotic genomes, however, are comparatively heterogeneous in their composition and much more extensive. That situation complicates the HGT search. Therefore, those methods developed for prokaryotes fail either due to their inability to handle the sequence heterogeneity or because their computational requirements skyrocket. Researchers studying HGT in eukaryotes use two types of methods: surrogate and comparative. Surrogate methods use nucleotide base composition of the DNA sequence. They have been applied in the form of chaos game representation clustering based on tetranucleotide composition (Mallet *et al.*, 2010). Comparative methods are computationally intensive because they compute phylogenetic comparisons between large numbers of identified genes or they use local alignment comparisons against a reference sequence database. They require prior annotation of a genome, an extensive database of comparable orthologues and computationally intensive phylogenetic analyses. Despite these limitations, the results from comparative methods are considered most reliable, as they enable identification of HGT and the donor organism in the form of testable hypotheses. Therefore, methods have been developed to reduce the candidate dataset for GIs while balancing the numbers of false positives and false negatives (Boc and Makarenkov, 2011;

*To whom correspondence should be addressed.

Mallet *et al.*, 2010; Podell and Gaasterland, 2007). To date, the effort in eukaryotic genomes has been demonstrated consistently to fail on these criteria due to the genome heterogeneity in eukaryotes (Mallet *et al.*, 2010), thus leading to the need for an expanding reference database for comparative analyses.

We show here, however, that genome complexity may be overcome by carefully examining the pattern of genomic signature along a sequence. We use both the selection of tetranucleotides with greatest interspecific variability in their frequencies and local composition shifts that take into account natural variation of tetranucleotide frequency changes along a chromosomal arm to locate regions that differ and might thus represent recent acquisitions via HGT.

2 ALGORITHM

2.1 Calculation of *4-mer* sliding density

The principle of the SigHunt (genomic signature hunter) method is to use a sliding window approach both to detect the HGT and to take into account DNA sequence composition change along chromosomes. The genomic signature is calculated as a frequency vector of tetranucleotides (*4-mer*) within the window of a genomic sequence. Within each sliding window, frequency of every *4-mer* is calculated as

$$F_m(S) = \frac{C_m(S)}{N_S - 3}$$

where $F_m(S)$ is the frequency of the *4-mer* m within a sequence in the sliding window (S), $C_m(S)$ is the count of m in S and N_S is the length of the sequence in the window S. Only fully resolved sites are counted towards $F_m(S)$. Tetranucleotides that contain an unresolved base are omitted. Unresolved sites are skipped, and the length of the window, but not the length of the sequence, is increased by the given number of nucleotides until it reaches 20% of the window size. This is the maximum extension of the window. There are two reasons for choosing oligonucleotide length. First, *4-mers* provide a vector with a number of dimensions sufficient for complex comparisons. Second, the *4-mers* frequency vector is representative even for relatively short sliding windows, which is of interest in cases when short alien genes could be expected.

Along a chromosomal sequence, DNA base composition changes with functional regions, such as coding and non-coding regions, repetitive elements, telomeres, centromeres or other structural elements that stabilize a chromosomal arm. Differences between repetitive and coding regions in particular are prone to variable frequency of a few *4-mers* that could either distort or inform a signal from alien fragments. To account for this, we develop a novel scoring system and introduce a sliding-density concept. This takes into account genomic regions D that are directly adjacent to the sliding window S, offset at the 5′ and 3′ ends of the sequence. Within the long sliding window of D, we calculate a kernel density estimate for each *4-mer*. The measured *4-mer* frequency in the short sliding window of sequence S is tested for whether or not it is located outside of the credibility interval (CI) of the *4-mer* density in D

$$F_m(S) \in \left(0, \Phi_D^{-1}\left(\frac{\alpha}{2}\right)\right) \cup \left(\Phi_D^{-1}\left(1 - \frac{\alpha}{2}\right), 1\right)$$

where Φ_D is a cumulative distribution function of $F_m(S)$ in D and α is a confidence level. Values found to be outside of the CI are scored in three intervals for $\alpha \in \{0.05, 0.025, 0.01\}$, adding 1, 2 and 3, respectively, to the discrete interval accumulative score (DIAS). To avoid autocorrelation in measuring $F_m(S)$ on D, D is selected in such a way that S and a specified number of its surrounding sliding windows (x) are excluded from the *4-mer* density calculation (eye-of-the-storm approach). Thus, we compare the S sequence to its context but not to its immediate surroundings. To compare the sliding window approach to previous genome-wide signature studies, we also show global *4-mer* density.

DIAS measures how many *4-mers* deviate in their frequency from the local background of the genomic sequence and by how much. While this is more stable along a chromosome than any other compositional measure known to us, it is nevertheless sensitive to local changes.

2.2 Selection of informative *4-mers*

To further improve computational speed of the SigHunt method, informative *4-mers* can be selected to reduce the signal-to-noise ratio. Multiple genomes are used to train the procedure for *4-mer* selection based on their intra- and inter-genomic variability. $F_m(S)$ values for all consecutive windows are used to calculate the *4-mer* density in a given chromosome. All chromosomes are used for the training genomes. Informative *4-mers* are selected as those where means of F_m in organisms are distinctive from the overall estimates and within-genome F_m variance is small. We score the *4-mers* using the tetranucleotide quality score (TES) for each *4-mer*

$$TES_m = \sum_{k=1}^{n} \left(K_k - \tilde{K}\right)^2 - \sum_{k=1}^{n} (A_k + B_k + D_k) + E$$

$$K = \frac{1}{c} \sum_{i=1}^{c} \mu_i$$

$$A = \sum_{i=1}^{c} \frac{1}{c} (\mu_i - K)^2$$

$$B = \sum_{i=1}^{c} \frac{1}{c} (\sigma_i - \overline{\sigma_c})^2$$

$$D = \tilde{\sigma}_c^2$$

$$E = \sum_{k=1}^{n} (K_k - K_e)^2 - (A_e + B_e + D_e)$$

where n is the number of all organisms used for training, c is the number of chromosomes in the given organism, μ_i is mean *4-mer* frequency on the given chromosome $\left[\mu_i = \left(\overline{F_m(S)}\right)_i\right]$, $\tilde{K}$ is the average of all mean frequencies of the given *4-mer* on all tested chromosomes in all organisms within the dataset, σ denotes respective variances and e is the estimated organism intended for the SigHunt search.

Using TES, we are interested to learn the extent to which *4-mer* frequencies vary between organisms with respect to their variability within a genome. To achieve this, K estimates average frequency of a *4-mer* that is found in the given organism. It is calculated as a mean of means to avoid weighting of the value according to the number and size of chromosomes. A sums

squared differences between each typical *4-mer* frequency on a chromosome compared with the whole genome; *B* similarly penalizes *4-mers* that have deviant frequency variance in a chromosome compared with the background genome. By including the *D* component into the equation, we ensure that the frequency variance in an organism is small to facilitate finding and interpreting *4-mer* frequencies outside of the confidence interval of their density. *E* provides a measure that would stress usefulness of the given *4-mer* for discrimination of the home sequence. TES increases where *4-mer* frequencies differ between organisms, they are stable for a given organism and they exhibit little variation within a genome. We demonstrate below that *4-mers* with high TES scores will be informative in recognizing putative HGT. GIs identified with SigHunt should subsequently be verified using comparative methods (Fig. 1).

3 METHODS

The sensitivity and specificity of the SigHunt method was tested using a receiver operating characteristic (ROC) curve on simulated data. We introduced alien sequences into the recipient sequence by randomly selecting and replacing DNA fragments between 10 organisms with complete genomic sequences. Eukaryotes were represented by the fungi *Aspergillus fumigatus* (Nierman *et al.*, 2005), *Encephalitozoon cuniculi* (Katinka *et al.*, 2001) and *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996); the red alga *Cyanidioschyzon merolae* (Matsuzaki *et al.*, 2004); the chromalveolates *Cryptosporidium parvum* (Abrahamsen *et al.*, 2004), *Plasmodium falciparum* (Hall *et al.*, 2002) and *Thalassiosira pseudonana* (Armbrust *et al.*, 2004); and an animal, *Drosophila melanogaster* (Adams *et al.*, 2000). SigHunt's performance in prokaryotes was demonstrated in *Buchnera* sp. (Shigenobu *et al.*, 2000) and *Escherichia coli* (Welch *et al.*, 2002). In each of 500 replicates of the procedure, we replaced three genomic fragments with alien DNA from other organisms. The length of each introduced fragment was 2, 5 and 15 kb in each chromosome, where the origin of each fragment was randomly drawn from the pool of analysed organisms and chromosomes. We scored the sequence according to DIAS for 5 kb windows and 1 kb sliding windows. Those regions with known alien sequences were scored as the only GIs present. We calculated the area under the curve (AUC) from the ROC curve in R (R Development Core Team, 2011; Robin *et al.*, 2011; Sing *et al.*, 2005) and estimated the optimal threshold while maximizing AUC. The same analysis was conducted using INDeGeNIUS, which is a recent surrogate method that uses oligonucleotide frequencies (Shrivastava *et al.*, 2010), and with Alien_Hunter, which uses interpolated variable order motifs of DNA composition (Vernikos and Parkhill, 2006).

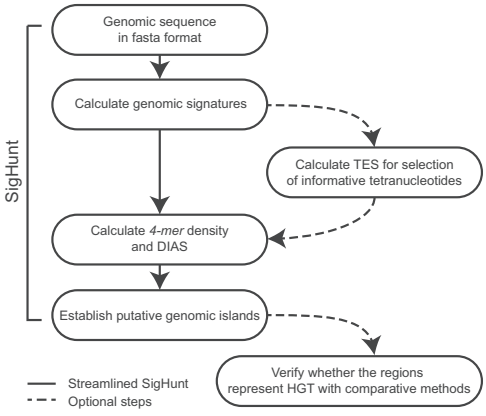


Fig. 1. Flow diagram of GIs analyses using the SigHunt method

3.1 Case studies

To test the SigHunt method on real biological data, we used genomic sequences of *Aspergillus*, *Cryptosporidium* and *Saccharomyces*, as listed above, and added genomic sequences not yet assembled to chromosomes for organisms with known GIs. The latter were the red algae *Galdieria sulphuraria*, wherein the horizontally transferred genes provide multiple environmental adaptations (Schönknecht *et al.*, 2013), and the fungus *Pyrenophora tritici-repentis*, which recently acquired a pathogenicity island (Friesen *et al.*, 2006) and had additional proteins originating from HGT (Sun *et al.*, 2013). We used organisms where GIs had been identified previously and their location was specific in the available genomic sequence (Hall *et al.*, 2005; Huang *et al.*, 2004; Mallet *et al.*, 2010; Schönknecht *et al.*, 2013; Sun *et al.*, 2013). Contigs at least 200 kb were used from these organisms. This enabled us to cross-check SigHunt against previous studies and thereby to demonstrate its utility. The optimal threshold value for the DIAS as tested with ROC on simulated data was 6.04, and we relaxed this value further to account for sequence amelioration. The cut-off value used here was DIAS ≥ 5. Two windows adjacent to the previously identified GI were assessed to compensate for the fact that most GIs were identified as a coding gene sequence and the transferred region could likely include flanking regions (Friesen *et al.*, 2006).

4 RESULTS

We estimated *4-mer* density and its variance in the 10 reference genomes. Comparing the *4-mers* using TES, we selected the 16 most informative *4-mers*. These were used for all subsequent analyses.

4.1 Sensitivity and specificity of SigHunt

SigHunt showed average AUC values equal to 0.77 for global density, 0.72 for sliding density and 0.77 for the eye-of-the-storm approach, meaning that sensitivity and specificity of detecting GIs in simulated data were high (Table 1). We assigned a GI only where it had been artificially introduced, and the remaining home sequence still contained its natural GIs, which were disregarded for the purpose of this test (Table 2). This could have lowered the performance indicators. The analyses of individual chromosomes required 8–60 min to calculate DIAS for all three approaches presented here. INDeGeNIUS and Alien_Hunter performed in a similar way with respect to their ability to correctly score the introduced GIs, and no differences between the methods were significant. INDeGeNIUS showed the highest values of AUC from the tested methods. However, those analyses took 20 min–20 h per chromosome, and *Drosophila* could not be analysed due to extensive memory requirements. The speed of Alien_Hunter was 20–120 min per chromosome.

4.2 HGT detection with SigHunt

We estimated that the studied genomes exhibit regions with deviant genomic signature that could be considered alien. These provide a genome-wide assessment of candidate regions

Table 1. Average area under a ROC curve for SigHunt, INDeGeNIUS and Alien_Hunter analyses on 500 random replacements of three GIs into reference chromosomal sequences from the model organisms

Organism	Global density	Sliding density	Eye of the storm	INDeGeNIUS	Alien_Hunter
Fungi					
Aspergillus	0.75 (0.11)	0.72 (0.11)	0.75 (0.12)	0.84 (0.07)	0.65 (0.17)
Encephalitozoon	0.71 (0.09)	0.70 (0.08)	0.74 (0.10)	0.88 (0.07)	0.88 (0.11)
Saccharomyces	0.81 (0.07)	0.72 (0.07)	0.78 (0.06)	0.90 (0.08)	0.83 (0.12)
Red alga					
Cyanidioschyzon	0.83 (0.08)	0.81 (0.07)	0.86 (0.07)	0.91 (0.07)	0.84 (0.15)
Animal					
Drosophila	0.74 (0.11)	0.68 (0.09)	0.72 (0.09)	n/a	0.75 (0.13)
Chromalveolates					
Cryptosporidium	0.77 (0.07)	0.68 (0.08)	0.74 (0.07)	0.86 (0.07)	0.78 (0.17)
Plasmodium	0.67 (0.12)	0.63 (0.09)	0.70 (0.12)	0.87 (0.07)	0.82 (0.17)
Thalassiosira	0.87 (0.09)	0.81 (0.07)	0.85 (0.08)	0.89 (0.06)	0.78 (0.18)
Prokaryotes					
Buchnera	0.83 (0.06)	0.73 (0.07)	0.81 (0.06)	0.91 (0.06)	0.85 (0.16)
Escherichia	0.76 (0.09)	0.68 (0.09)	0.72 (0.09)	0.85 (0.07)	0.82 (0.16)

Note: Standard deviation is given in parentheses.
n/a, not available.

Table 2. Number of correctly assigned GIs in model organisms from those identified previously as retrieved by SigHunt eye-of-the-storm variant, INDeGeNIUS and Alien_Hunter

Organism	Number of chromosomes or contigs[a]	Sequence length (Mb)	Previously established GIs	SigHunt	INDeGeNIUS	Alien_ Hunter	References
Fungi							
Aspergillus	8	29.4	189	150	189	54	(Mallet *et al*., 2010)
Pyrenophora	19[a]	33.9	17[b]	6	11	0	(Friesen *et al*., 2006; Sun *et al*., 2013)
Saccharomyces	16	12	10[b]	5	2	5	(Hall *et al*., 2005)
Red algae							
Galdieria	17[a]	4.4	79[b,c]	48	15	33	(Schönknecht *et al*., 2013)
Chromalveolates							
Cryptosporidium	8	9.1	30[b]	9	12	11	(Huang *et al*., 2004)

Note: In SigHunt, a selection of 16 *4-mers* identified by TES was used. DIAS ≥ 5 was used as a cut-off value and two windows adjacent to the island borders were considered. Sequence length = size of the analysed genomic sequence.
[a]Number of assessed contigs.
[b]In annotated genes.
[c]GIs found in 13.7 Mb of the genomic sequence, but only the longest contigs were analysed here.

for HGT. We searched for consistency of specific sequences in *Aspergillus*, *Saccharomyces*, *Pyrenophora*, *Galdieria* and *Cryptosporidium* between SigHunt and published results. Compared with the extent of HGT identified in previous studies that used predominantly comparative methods on annotated genes, SigHunt found from 30% (*Cryptosporidium*) to 80% (*Aspergillus*) of previously identified GIs (Table 2). In *Aspergillus*, we were able to find the majority of published GIs as per Mallet *et al*. (2010), which can be expected given that those authors used a surrogate method verified with phylogenetic comparison to localize GIs. We identified 5 of 10 putative GIs recognized in *Saccharomyces* by Hall *et al*. (2005). The missed GIs were protein-coding genes <1.2 kb, for which our chosen window size might not be optimal. In *Pyrenophora*, we searched for 17 genes and 6 were retrieved by SigHunt. HGT in *Galdieria* can be attributed to ~9% of the genomic sequence in the longest scaffolds, which we analysed (Schönknecht *et al*., 2013). SigHunt was able to recognize the genomic signatures of the published GIs as being alien in 61% of the cases. In *Cryptosporidium*, we found 9 of 30 previously identified GIs (Huang *et al*., 2004). As in previous cases with low success, the GIs in *Cryptosporidium* that were missed consisted predominantly of short genes. In these cases, experimentation with window size would be beneficial. INDeGeNIUS found more known GIs in *Aspergillus*, *Pyrenophora* and *Cryptosporidium*, but the analyses took an order of magnitude longer than in SigHunt. SigHunt outperformed Alien_Hunter in recognizing the previously established HGT events in most tested organisms. The exception

was *Cryptosporidium*, where Alien_Hunter found 11 of 30 GIs (Table 2).

5 DISCUSSION

We present here a tool for identifying genomic regions as candidates for HGT assessment in eukaryotes. To our knowledge, this is the first surrogate method primarily optimized for eukaryotic genomes. It detects non-ameliorated HGT in large genomic sequences, it is computationally efficient and its implementation provides step-wise user access to results that enables data exploration and analytical optimization.

We demonstrated good success in using SigHunt to find introduced GIs across kingdoms and GIs in real genomic sequences (particularly in some fungi). Considering from a biological perspective reproduction within this group, HGT might be more common in fungi than in other eukaryotes (Rosewich and Kistler, 2000). With HGT events being relatively common within the group, one could speculate that some of these will be non-ameliorated and thus easily detectable using surrogate methods. The further example of *Galdieria*, within which GIs were plentiful across the genome, seems to corroborate this.

5.1 SigHunt's advantages

The advantage of SigHunt lies in its utilization of informative *4-mers*. Selecting only parts of the genomic signature that are most informative according to TES reduces noise in the data and computational requirements, thereby speeding up the analysis. Computational demands are not negligible in eukaryotic genomics. For example, INDeGeNIUS analysis of the largest chromosome in *Drosophila* would require ≈70 TB of memory, which is beyond the capacity available to many researchers, including our group, and the current version of the program does not allow for changes in memory use. By contrast, SigHunt analysed the same problem using 900 MB of maximum allocated memory. At a given time, SigHunt stores in memory only that chromosome sequence needed to calculate the signatures, or, once the corresponding signatures are calculated and the chromosome sequence deleted from memory, the signatures and densities themselves. Such orderly memory utilization reduces the memory requirements and thus allows computations of even large datasets on regular office computers.

Contrary to other recent methods that use a genomic signature for HGT detection (Elhai *et al*., 2012; Shrivastava *et al*., 2010; but see Mallet *et al*., 2010), SigHunt does not assume a point estimate of the genomic signature. Instead, it uses a density distribution and thus acknowledges the natural variability of DNA composition and distinguishes only those regions that deviate from the broad 'norm' for an organism. We recognize that the density distributions must contain tails even for home signatures. Due to this, DIAS measures accumulations of deviant *4-mer* frequencies rather than their mere occurrence. With sliding density and its eye-of-the-storm variant that avoids autocorrelation, SigHunt is able to find putative GIs in any region of the chromosome. This includes regions rich in repetitive DNA because SigHunt assumes a differing genomic signature typical for a genomic region rather than for the whole chromosome.

SigHunt makes it unnecessary to have knowledge as to the exact position on a chromosome of the examined sequence (Podell and Gaasterland, 2007). It can successfully analyse unassembled genomes, provided that the supercontigs are sufficiently long. It also does not require information about gene locations. By filtering nucleotide positions in a sequence that are not fully resolved, we limit the amount of information while increasing the accuracy. In case of a long eukaryotic genome, a trade-off in favour of accuracy is paramount for SigHunt.

5.2 SigHunt's disadvantages

Unfortunately, SigHunt is not a universal black box solution for all HGT problems. Its very principle denies universality, as it rises and falls on the assumption that there are differences in oligonucleotide frequencies between organisms (Karlin and Burge, 1995). This is not always sufficiently true, as shown by our analyses on manipulated and real data. Some random islands were undifferentiated from the home signature. For others, the GI size in real datasets might have been too small to accurately estimate the *4-mer* frequency density for the DIAS calculation. In addition to reasons of there being similar genomic signatures among the organisms involved, SigHunt is prone to false negatives due to amelioration over time of the compositional bias in the horizontally transferred region compared with the host genome. The false-positive rate might also be increased. In regions with strong selection bias and functional restrictions, home signature might vary locally. The extent to which this is the case remains to be tested.

SigHunt expands the search for GIs across the genome without annotation limitations, yet it is able to guide the comparative search more effectively than do other similar methods. We have shown that SigHunt provides a fair basis of target regions for comparative assessment that consists of true GIs as confirmed by phylogenetic analyses in the published data (Friesen *et al*., 2006; Hall *et al*., 2005; Huang *et al*., 2004; Mallet *et al*., 2010; Schönknecht *et al*., 2013; Sun *et al*., 2013).

5.3 Global density paradox

We claim that the advantage of SigHunt lies in its ability to account for variation of the genomic signature along a chromosomal sequence. Yet, in Table 1, the sensitivity and specificity test shows the highest (albeit not statistically significant) AUC values to be for the global density estimate in six cases. This is probably caused by the fact that we introduced HGT directly between organisms that spanned kingdoms and our islands were thus devoid of any amelioration. In other words, while one would rarely encounter such an event in practice, it is one that is relatively easy to capture by means of genomic signatures. Analysing real biological data would require a more subtle approach. Therefore, both the sliding density and its eye-of-the-storm variant provide room for fine-tuning the method. These are parameterized for window size, sliding window size, sliding-density window size, and eye-of-the-storm size. All these parameters might be optimized to further improve SigHunt for any specific target organism. On the other hand, the global density reached the height of its performance in this study. We nevertheless consider global-density DIAS calculation a useful approach in view

of the fact that it is computationally effective and has low memory demands.

5.4 Usefulness of SigHunt

As a method for investigating genomic signature in complex eukaryotic genomes, SigHunt can provide a rapid analysis tool for ongoing sequencing projects. In particular, it will be sensitive to recent HGT, such as in cases of emergent pathogens that have acquired novel genes. The choice of informative *4-mers* could increase resolution and success for binning of metagenomic DNA fragments (Saeed and Halgamuge, 2009). With the recent finding of bacterial horizontally transferred genes in human tumour cells (Riley *et al*., 2013), SigHunt shows promise to be used in rapid screening of such events in genomic assemblies of specific cell lines. We assume that further research will reveal more fields within which sliding density of DNA composition in eukaryotic genomes and the selection of informative oligonucleotides will prove advantageous.

ACKNOWLEDGEMENT

The bioinformatic analyses were conducted at the MetaCentrum computing facility of Masaryk University.

Funding: Czech Science Foundation (grant number P506/12/1064). The access to the MetaCentrum was funded by the Ministry of Education, Youth, and Sports of the Czech Republic (grant number LM2010005).

Conflict of Interest: none declared.

REFERENCES

Abrahamsen,M.S. *et al.* (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*, **304**, 441–445.

Adams,M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Armbrust,E.V. *et al.* (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science*, **306**, 79–86.

Boc,A. and Makarenkov,V. (2011) Towards an accurate identification of mosaic genes and partial horizontal gene transfers. *Nucleic Acids Res*., **39**, e144.

Casacuberta,E. and Gonzlez,J. (2013) The impact of transposable elements in environmental adaptation. *Mol. Ecol*., **22**, 1503–1517.

Elhai,J. *et al.* (2012) Detection of horizontal transfer of individual genes by anomalous oligomer frequencies. *BMC Genomics*, **13**, 245.

Freeman,V.J. (1951) Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. *J. Bacteriol*., **61**, 675.

Friesen,T.L. *et al*. (2006) Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat. Genet*., **38**, 953–956.

Goffeau,A. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546–567.

Hall,C. *et al*. (2005) Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. *Eukaryot. Cell*, **4**, 1102–1115.

Hall,N. *et al*. (2002) Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature*, **419**, 527–531.

Huang,J. *et al*. (2004) Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. *Genome Biol*., **5**, R88.

Jern,P. and Coffin,J.M. (2008) Effects of retroviruses on host genome function. *Annu. Rev. Genet*., **42**, 709–732.

Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*., **11**, 283–290.

Katinka,M.D. *et al.* (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature*, **414**, 450–453.

Mallet,L. *et al*. (2010) Whole genome evaluation of horizontal transfers in the pathogenic fungus *Aspergillus fumigatus*. *BMC Genomics*, **11**, 171.

Matsuzaki,M. *et al*. (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10d. *Nature*, **428**, 653–657.

Nierman,W.C. *et al.* (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature*, **438**, 1151–1156.

Podell,S. and Gaasterland,T. (2007) DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol*., **8**, R16.

R Development Core Team. (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Riley,D.R. *et al.* (2013) Bacteria-human somatic cell lateral gene transfer is enriched in cancer samples. *PLoS Comput. Biol*., **9**, e1003107.

Robin,X. *et al.* (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.

Rosewich,U.L. and Kistler,H.C. (2000) Role of horizontal gene transfer in the evolution of fungi. *Annu. Rev. Phytopathol*., **38**, 325–363.

Saeed,I. and Halgamuge,S. (2009) The oligonucleotide frequency derived error gradient and its application to the binning of metagenome fragments. *BMC Genomics*, **10**, S10.

Schönknecht,G. *et al.* (2013) Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, **339**, 1207–1210.

Shigenobu,S. *et al*. (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. aps. *Nature*, **407**, 81–86.

Shrivastava,S. *et al.* (2010) INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms. *J. Biosci*., **35**, 351–364.

Sing,T. *et al*. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Sun,B.F. *et al.* (2013) Multiple interkingdom horizontal gene transfers in *Pyrenophora* and closely related species and their contributions to phytopathogenic lifestyles. *PLoS One*, **8**, e60029.

Vernikos,G.S. and Parkhill,J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics*, **22**, 2196–2203.

Welch,R.A. *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 17020–17024.

Wolfenbarger,L.L. and Phifer,P.R. (2000) The ecological risks and benefits of genetically engineered plants. *Science*, **290**, 2088–2093.

Paper 2.5.2

Škrabánek P., **Martínková N.** 2017. Extraction of outliers from imbalanced sets. In: Martínez de Pisón F., Urraca R., Quintián H., Corchado E. (Eds.) *Hybrid Artificial Intelligent Systems. HAIS 2017. Lecture Notes in Computer Science*, vol. 10334. pp. 402-412. Springer, Cham, DOI: 10.1007/978-3-319-59650-1_34.

Extraction of Outliers from Imbalanced Sets

Pavel Škrabánek[1( )] and Natália Martínková[2,3]

[1] Faculty of Electrical Engineering and Informatics, University of Pardubice, Studentská 95, 532 10 Pardubice, Czech Republic
pavel.skrabanek@upce.cz

[2] Institute of Vertebrate Biology, Czech Academy of Sciences, Květná 8, 603 65 Brno, Czech Republic
martinkova@ivb.cz

[3] Institute of Biostatistics and Analyses, Masaryk University, Kamenice 3, 625 00 Brno, Czech Republic

Abstract. In this paper, we presented an outlier detection method, designed for small datasets, such as datasets in animal group behaviour research. The method was aimed at detection of global outliers in unlabelled datasets where inliers form one predominant cluster and the outliers are at distances from the centre of the cluster. Simultaneously, the number of inliers was much higher than the number of outliers. The extraction of exceptional observations (EEO) method was based on the Mahalanobis distance with one tuning parameter. We proposed a visualization method, which allows expert estimation of the tuning parameter value. The method was tested and evaluated on 44 datasets. Excellent results, fully comparable with other methods, were obtained on datasets satisfying the method requirements. For large datasets, the higher computational requirement of this method might be prohibitive. This drawback can be partially suppressed with an alternative distance measure. We proposed to use Euclidean distance in combination with standard deviation normalization as a reliable alternative.

Keywords: Outlier analysis · Distance based method · Global outlier · Single cluster · Mahalanobis distance · Biology

1 Introduction

Data mining reveals new, valuable and non-trivial information in large datasets [14]. It is a process of discovering interesting patterns and knowledge in the data that is not immediately apparent. Various data mining approaches help to specify the patterns in the data mining tasks. Examples include characterization and discrimination, mining of frequent patterns, associations and correlations, classification and regression, clustering analysis, and outlier analysis [10].

The outlier analysis has an important position among data mining approaches. Hawkins specified an intuitive definition of the term *outlier* as: 'Within a given dataset, the outlier is an observation which deviates so much

F.J. Martínez de Pisón et al. (Eds.): HAIS 2017, LNAI 10334, pp. 402–412, 2017.
DOI: 10.1007/978-3-319-59650-1_34

from other observations as to arouse suspicions that it was generated by a different mechanism' [11]. The other observations are usually called *inliers*, *normal data* or *normal observations*. Throughout the text, a predetermined battery of *features* characterizes an observation.

The outlier analysis is used in a wide variety of domains such as the financial industry, quality control, fault diagnosis, intrusion detection, web analytics, and medical diagnosis [2]. The most typical application of the outlier analysis is data cleaning. However, in many applications, outliers are more interesting than inliers. Fraud detection is a classic example, where attention focuses on the outliers, because these more likely represent cases of fraudulent behaviour [12].

The outlier analysis distinguishes three categories of outliers that require specific analytical approaches: *global outliers*, *contextual outliers* (known also as *conditional outliers*), and *collective outliers* [10]. A global outlier is an observation that deviates significantly from the rest of the dataset, whereas a contextual outlier deviates from inliers only with respect to a specific context. The term collective outlier is used for a subset of observations. A subset of observations forms a collective outlier if the subset as a whole deviates significantly from the entire dataset.

To identify outliers, the outlier detection methods create models of normal patterns in the data (so called *data model* or simply *model*), and then compute an *outlier score* of a given observation on the basis of the deviations from the normal patterns [2]. The outlier detection methods utilize *clustering models*, *distance-based models*, *density-based models*, *probabilistic* and *statistical models*, *classification models*, and *information-theoretic models* [2,10].

The selection of the model and outlier score calculation is data-specific and relies on assumptions of information contained in the data. For example, classification models require datasets of labelled observations. Methods based on other models, e.g. statistical models or distance-based models, can be applied to both labelled as well as unlabelled datasets.

The correct choice of the method from the perspective of the data model determines results of the outlier analysis [2]. For example, application of a method based on a statistical model, which expects a uniform distribution of inliers, would be inappropriate for a dataset with the zipf distribution.

In biology, animal group behaviour studies generate specific datasets of observable variables pre-selected in the experimental design [5,18]. Typically, such datasets are unlabelled and may contain numerical as well as categorical data. Given the complexity of animal behaviour, feature space of the observed variables will not be exhaustive on an individual level and determinants of group behaviour will exhibit subtle trends. From amongst the data mining approaches, the outlier analysis provides functionality to identify observations putatively generated by an alternative mechanism, which makes the analysis suitable for application in animal group behaviour research. In order to ensure a simple and reliable recognition of the outliers in such a dataset, we developed an outlier detection method. Our method detects global outliers using a distance-based model. Here, we introduce the method for numerical data.

2 Methods

2.1 Analysis of the Problem

A dataset considered for application with the proposed method contains inliers that form one multidimensional cluster, while the outliers span at a distance from the cluster centre. The outliers may or may not form small clusters. The total number of observations in the dataset range from tens to hundreds of observations. Further, distribution of inliers may significantly vary among various datasets. The data contains no prior knowledge about the outliers, and the information embodied in the outliers is the object of interest. These datasets may include both numerical and categorical data; however, the proposed method is intended for datasets composed of numerical data.

The first step in developing a new outlier detection method is identification of the outlier category. Following the above stated setup, the proposed method detects global outliers. The second step, selection of the model for inliers, delineates the direction of the development process. Herein, we use backward selection to select the proper model. Information-theoretic models are impropriate for the defined datasets, because of the expected type of features. Without prior knowledge about the outliers, the new detection method cannot be based on a classification model. As different datasets may have different distributions of inliers, usage of a probabilistic, density-based or a statistical model is inadvisable. Consequently, the method has to be based on one of the remaining model types; clustering or distance-based models.

Both clustering models and distance-based models represent appropriate choices for the new outlier detection method given the data. Between them, distance-based methods enable a higher granularity of analysis as compared to clustering methods. This property of distance-based methods provides a more refined ability to distinguish between weak and strong outliers in noisy data sets [2]. Hence, the presented method has been developed on a distance-based model.

2.2 Description of the Method

Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n observations $\mathbf{x}$. The i-th observation $\mathbf{x}_i$, where $i \in I$ and $I = \{1, \ldots, n\}$, is a d-dimensional real vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$ of features $x \in \mathcal{F}$, where x_{ik} is the k-th feature of the i-th observation, and $\mathcal{F}$ is a feature space. Let us expect that the majority of the observations, say m, belongs to inliers. The remaining p observations correspond to outliers. A subset of all outliers in the set X will be denoted as O.

The presented method belongs to the group of outlier detection methods based on a distance model. It assumes two attributes in the observations $\mathbf{x} \in X$:

(I) inliers form one predominant cluster and the outliers are at distances from the centre of the cluster,
(II) the number of inliers is much higher than the number of outliers ($m >> p$).

In order to design the separation method, the outlier score had to be properly formulized. For this purpose, an appropriate similarity measure had to be chosen. Similarity of two observations, say $\mathbf{x}_i$, $\mathbf{x}_j \in X$, was assessed using a distance measure. In order to ensure a comparable level of impact for all the features $x \in \mathcal{F}$, the observations should be compared with normalized data or the measure should be unitless and scale-invariant. In our solution, we used Mahalanobis distance [4,7]. This measure is unitless and scale-invariant. For the observations $\mathbf{x}_i, \mathbf{x}_j$, the Mahalanobis distance is defined as

$$d\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j)^{\top}}, \tag{1}$$

where $\mathbf{S}$ is a covariance matrix, and $\top$ symbolizes transposition.

Considering the properties of the datasets expressed via the assumptions (I) and (II), we proposed to formulate the outlier score J for the i-th observation as the sum of distances between the i-th observation and the others, i.e.

$$J_i = \sum_{\forall j \in I} d\left(\mathbf{x}_i, \mathbf{x}_j\right). \tag{2}$$

The distance-based methods usually take into account distances between an evaluated observation and its k nearest neighbours. Nevertheless, the outlier score (2) considers all n distances. An example demonstrates rationalization for the formulation of the score. In Fig. 1, the number of distances is identical regardless of whether an inlier (Fig. 1a) or an outlier (Fig. 1b) is evaluated; however, distributions of their values differ. For the outliers, longer distances appear more frequently than for inliers. This holds for an arbitrarily chosen inlier and outlier, since inliers form a single cluster and $m >> p$.

The specific properties of the dataset lead to the conclusion that the greater the number of nearest neighbours which are included in the analysis, the larger the difference between scores of inliers and outliers. Consequently, inclusion of all observations in the comparison results in higher sensitivity of the method. The associated increase in computational complexity of the method is irrelevant for the expected dataset sizes. For larger datasets, a GPU optimized variant of the method may be developed [3].

Observations evaluated using the outlier score (2) can be easily classified as outliers or inliers using a threshold value t. In our case, the unusual structure of the dataset X inspired the analytical expression of t. Indeed, values of the score for inliers are markedly smaller than for outliers. Considering this fact and the fact that $m >> p$, median of the score's values $\hat{J}$ adequately describes inliers. On the basis of the median and the smallest score values, the range of score values of inliers can be estimated. Thus, the threshold value can be expressed as

$$t = \varepsilon.[\hat{J} - \min_{\forall i \in I} J_i] + \hat{J}, \tag{3}$$

where the parameter ε is used as a tuning parameter. Each observation with a score equal to or greater than the threshold value t is expected to be an outlier.

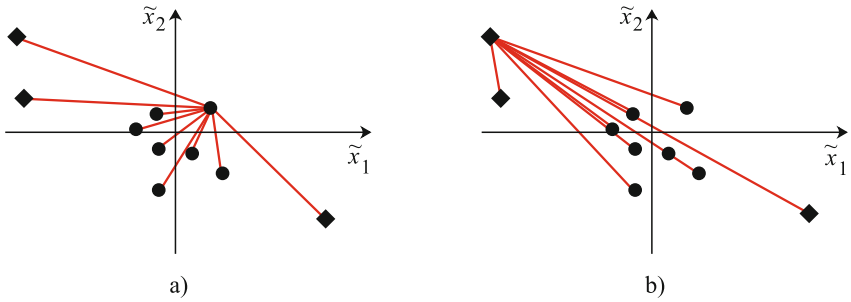


Fig. 1. Demonstration of the idea behind the outlier score by evaluation of: (a) an inlier, and (b) an outlier. In both figure panels, three outliers (diamonds) and seven inliers (circles) are plotted on a two-dimensional centred, rotated and standardized feature space $\tilde{x}_1, \tilde{x}_2 \in \tilde{\mathcal{F}}$. The distance between two observations is symbolized using a red line. (Color figure online)

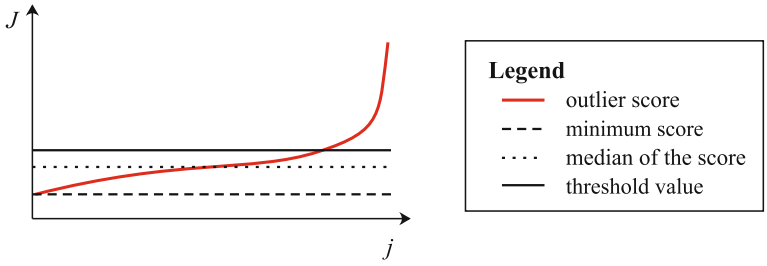


Fig. 2. Visualization of the score J as a function j where j are indexes of the observations sorted according to J in the ascending order. Threshold for outlier classification t is placed at a point with rapid change in score trend

The presented method has one tuning parameter ε, and its setting significantly predetermines the output of the method. We proposed a visualization method in order to estimate the accurate value for ε. The visualization displays a continuous line connecting scores J for $\forall \mathbf{x} \in X$, where the scores are sorted in ascending order. The line is approximately exponential. The initial lag phase with gradual increase in J, includes inliers, and the subsequent exponential phase includes outliers. Using the graph, an expert can estimate the boundary between inliers and outliers and accordingly the threshold value determining ε (Fig. 2).

For datasets satisfying the assumptions, the right placement of the auxiliary line is straightforward. However, the more the dataset X deviates from the ideal, the deeper understanding of the data is necessary for the appropriate placement.

2.3 Algorithmic Expression of the Method

The proposed method can be realized as a function, here presented as a pseudocode (Algorithm 1). The function has two inputs and two outputs. The inputs are the set of observations X and the tuning parameter ε. The outputs are a set of all outliers O and a set of outliers indexes I_o in the original set X.

Algorithm 1. Extraction of Exceptional Observations

1: **function** EEO(X, ε)
Input: Set of n observations $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, constant ε specifying the limit for outliers
Output: Set O of all outliers and set of their indexes I_o in X
2: $J_i \leftarrow \sum_{\forall j \in I} d(\mathbf{x}_i, \mathbf{x}_j), \forall i \in I$ ▷ Evaluation of observations using the criterion
3: $t \leftarrow \varepsilon.[\hat{J} - \min_{\forall i \in I} J_i] + \hat{J}$ ▷ Threshold value for exceptional observations
4: $I_o \leftarrow \{i : i \in I \text{ where } J_i \geq t\}$ ▷ Indexes of exceptional observations
5: $O \leftarrow \{\mathbf{x}_i : i \in I_o\}$ ▷ Set of exceptional observations
6: **return** O, I_o
7: **end function**

2.4 Experimental Evaluation of the Method

We used 44 previously published datasets for evaluation of the proposed method [8]. The datasets originated from areas such as biology, medicine, criminology or astronautics. They contained three types of features: R - real numbers, I - integers, and N - nominal values. The datasets consisted of labelled samples with two classes. All datasets were imbalanced with an imbalance ratio $IR = m/p$, where $IR \in [1.82, 129.44]$. We expected that the minority class represented outliers, while the majority class consisted of inliers.

We adapted three performance measures used in binary classification to evaluate results obtained from our method. Namely, we considered *sensitivity* (Se), *specificity* (Sp), and their geometric mean (G) [8,15]. For the outlier analysis, they can be expressed as

$$Se = \frac{|TO|}{|TO| + |FI|}, \qquad Sp = \frac{|TI|}{|TI| + |FO|}, \qquad G = \sqrt{Se \cdot Sp}, \tag{4}$$

where $|TO|$ is the number of correctly recognized outliers (true outliers), $|FO|$ is the number of inliers labelled as outliers (false outliers), $|TI|$ is the number of correctly recognized inliers (true inliers), $|FI|$ is the number of outliers labelled as inliers (false inliers).

We evaluated our method (extraction of exceptional observations, EEO) for two values of ε. Within the first experiment, we estimated the value of ε from the graph (as per Fig. 2). This value was denoted as $\hat{\varepsilon}$. In the second experiment, we searched for an optimal setting (ε^*) using genetic algorithms [16]. The genetic algorithms used the objective function as $\max G(\varepsilon)$. We applied the MATLAB function `ga`, with no constraints and default settings [1].

3 Results

Due to the presence of nominal variables, EEO could not be applied on datasets 'abalone19' and 'abalone9–18'. Further, it was unsuccessful on datasets 'ecoli-0_vs_1' and 'segment0', in which some features had constant values for all observations. The obtained results are summarized in Table 1. In general, sensitivity of

EEO with manual estimation of $\hat{\varepsilon}$ was lower than for ε^*, established from labelled data with genetic algorithms in lieu of higher specificity. This was accompanied by higher, and thus more conservative, values of $\hat{\varepsilon}$.

To estimate the performance of EEO amongst existing outlier detection methods, we compared our method with Chi et al.'s method with 3 and 5 labels (Chi-3 and Chi-5) [6], Ishibuchi et al.'s method (Ish05) [13], E-Algorithm (E-Alg) [19], Fernández et al.'s method (HFRBCS) [8], and C4.5 decision tree (C4.5) [17]. We adopted the evaluation results published in [8]. The evaluation results using G are summarized for all expected methods, including the EEO with optimal and manual setting of ε, in Table 2.

4 Discussion

The proposed EEO method was tested on 44 datasets previously used for algorithm testing [8]. The datasets differed in the imbalance ratio, in the number of features and their type (Table 1). However, from the viewpoint of EEO testing, many of these datasets did not meet the assumptions that the inliers form one multidimensional cluster (I) and the number of inliers is much higher than the number of inliers (II). This is apparent when displaying the first two principal components of observation scores with their class labels [9]. The dataset 'shuttle-c0-vs-c4' fully met the assumptions (Fig. 3b), and EEO was successful in outlier detection ($G \approx 98$). The datasets 'ecoli-0-1-3-7_vs_2-6', 'shuttle-c2-vs-c4', and 'Wisconsin' similarly showed nearly ideal class assignment with respect to inliers (data not shown). On these datasets, EEO exhibited excellent results according to all three measures (4) both for $\hat{\varepsilon}$ and ε^*. The performance of EEO is fully comparable to all evaluated methods. In fact, EEO provides considerably better separation on 'ecoli-0-1-3-7_vs_2-6' dataset than any other considered method (Table 2).

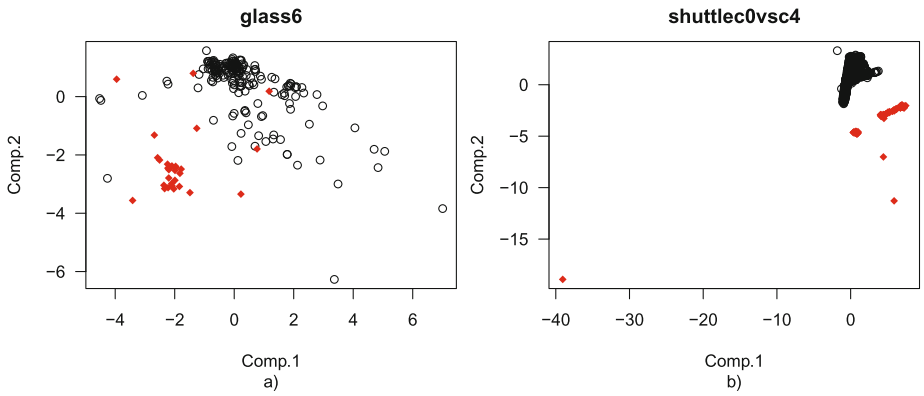


Fig. 3. Example of (a) inappropriate and (b) ideal datasets. The outliers (red diamonds) and inliers (black circles) are plotted in a two-dimensional centred, rotated and standardized feature space using the first two principal components. (Color figure online)

Table 1. Evaluation of EEO on test datasets using sensitivity *Se*, specificity *Sp*, and their geometric mean *G*. In first four columns, basic information about datasets is listed. It includes name, feature type (R - real numbers, I - integers, and N - nominal values), number of observations n, and imbalance ratio *IR*. The remaining columns consist of evaluation results for estimated and suboptimal setting of ε, respectively.

Information about datasets				EEO with $\hat{\varepsilon}$				EEO with ε^*			
Name	R/I/N	n	*IR*	ε	*Se*	*Sp*	*G*	ε	*Se*	*Sp*	*G*
abalone19	7/0/1	4174	129.44	-	-	-	-	-	-	-	-
abalone9-18	7/0/1	472	16.4	-	-	-	-	-	-	-	-
ecoli-0_vs_1	7/0/0	220	1.86	-	-	-	-	-	-	-	-
ecoli-0-1-3-7_vs_2-6	7/0/0	281	39.14	0.60	71.43%	86.13%	78.44%	2.47	71.43%	98.91%	84.05%
ecoli1	7/0/0	336	3.36	0.70	20.78%	83.78%	41.72%	0.04	59.74%	55.98%	57.83%
ecoli2	7/0/0	336	5.46	0.90	9.62%	86.97%	28.92%	0.26	57.69%	64.79%	61.14%
ecoli3	7/0/0	336	8.6	0.90	8.57%	87.04%	27.31%	0.06	65.71%	55.15%	60.20%
ecoli4	7/0/0	336	15.8	0.90	65.00%	90.82%	76.83%	0.75	70.00%	87.34%	78.19%
glass0	9/0/0	214	2.06	1.30	10.00%	68.75%	26.22%	−0.30	51.43%	25.69%	36.35%
glass-0-1-2-3_vs_4-5-6	9/0/0	214	3.2	1.30	52.94%	84.66%	66.95%	0.88	86.27%	83.44%	84.84%
glass-0-1-6_vs_2	9/0/0	192	10.29	1.60	17.65%	77.71%	37.03%	−0.21	88.24%	37.71%	57.69%
glass-0-1-6_vs_5	9/0/0	184	19.44	1.40	22.22%	78.86%	41.86%	0.85	55.56%	69.71%	62.23%
glass1	9/0/0	214	1.82	1.30	19.74%	73.19%	38.01%	−0.22	56.58%	34.78%	44.36%
glass2	9/0/0	214	11.59	1.00	11.76%	85.28%	31.67%	0.12	41.18%	55.84%	47.95%
glass4	9/0/0	214	15.47	1.30	46.15%	77.11%	59.66%	0.59	92.31%	66.67%	78.45%
glass5	9/0/0	214	22.78	1.30	22.22%	75.61%	40.99%	0.88	55.56%	67.80%	61.38%
glass6	9/0/0	214	6.38	1.30	65.52%	82.16%	73.37%	0.87	96.55%	76.76%	86.09%
haberman	0/3/0	306	2.78	0.60	14.81%	95.11%	37.54%	0.13	46.91%	66.22%	55.74%
iris0	4/0/0	150	2	0.80	10.00%	81.00%	28.46%	−0.22	86.00%	30.00%	50.79%
new-thyroid1	4/1/0	215	5.14	0.70	65.71%	83.89%	74.25%	0.31	80.00%	74.44%	77.17%
new-thyroid2	4/1/0	215	5.14	0.70	65.71%	83.89%	74.25%	0.23	82.86%	71.11%	76.76%
page-blocks0	4/6/0	5472	8.79	0.70	85.87%	83.88%	84.87%	0.69	86.05%	83.68%	84.85%
page-blocks-1-3_vs_4	4/6/0	472	15.86	0.70	53.57%	78.60%	64.89%	0.46	71.43%	71.40%	71.41%
pima	8/0/0	768	1.87	0.90	20.15%	88.40%	42.20%	0.10	61.57%	67.20%	64.32%
segment0	19/0/0	2308	6.02	-	-	-	-	-	-	-	-
shuttle-c0-vs-c4	0/9/0	1829	13.87	0.70	100.00%	95.96%	97.96%	0.77	100.00%	97.01%	98.49%
shuttle-c2-vs-c4	0/9/0	129	20.5	0.50	100.00%	88.62%	94.14%	0.97	100.00%	96.75%	98.36%
vehicle0	0/18/0	846	3.25	0.50	16.08%	90.42%	38.13%	0.05	53.27%	60.59%	56.81%
vehicle1	0/18/0	846	2.9	0.40	9.68%	82.83%	28.31%	−0.02	48.85%	46.10%	47.46%
vehicle2	0/18/0	846	2.88	0.40	18.35%	85.83%	39.68%	0.14	31.65%	65.61%	45.57%
vehicle3	0/18/0	846	2.99	0.40	10.85%	83.28%	30.06%	−0.12	70.75%	37.54%	51.54%
vowel0	10/3/0	988	9.98	0.40	48.89%	86.64%	65.08%	0.20	68.89%	72.72%	70.78%
wisconsin	0/9/0	683	1.86	0.50	97.07%	88.51%	92.69%	1.15	93.72%	94.37%	94.05%
yeast-0-5-6-7-9_vs_4	8/0/0	528	9.35	0.50	37.25%	81.34%	55.05%	0.30	54.90%	72.54%	63.11%
yeast1	8/0/0	1484	2.46	0.50	18.41%	78.58%	38.04%	−0.03	48.48%	45.31%	46.87%
yeast-1_vs_7	7/0/0	459	14.3	0.70	40.00%	85.31%	58.42%	0.24	53.33%	66.67%	59.63%
yeast-1-2-8-9_vs_7	8/0/0	947	30.57	0.70	40.00%	83.32%	57.73%	0.39	53.33%	72.30%	62.10%
yeast-1-4-5-8_vs_7	8/0/0	693	22.1	0.70	10.00%	82.50%	28.72%	−0.13	73.33%	40.57%	54.55%
yeast-2_vs_4	8/0/0	514	9.08	0.60	50.98%	84.45%	65.61%	0.18	84.31%	69.76%	76.69%
yeast-2_vs_8	8/0/0	482	23.1	0.60	70.00%	81.82%	75.68%	1.20	65.00%	93.94%	78.14%
yeast3	8/0/0	1484	8.1	0.50	25.15%	80.02%	44.86%	−0.01	69.33%	51.40%	59.69%
yeast4	8/0/0	1484	28.1	0.50	45.10%	80.32%	60.19%	0.16	78.43%	62.11%	69.79%
yeast5	8/0/0	1484	32.73	0.70	34.09%	86.74%	54.38%	0.06	100.00%	55.69%	74.63%
yeast6	8/0/0	1484	41.4	0.70	22.86%	86.34%	44.42%	−0.04	91.43%	47.83%	66.13%

Table 2. Comparison of EEO with other approaches for outlier detection. The geometric mean G of sensitivity and specificity was used as an overall comparison value. Results obtained by EEO are in bold on relevant datasets that meet designed criteria (inliers form a predominant cluster with outliers spanned from it and the number of inliers is greater than the number of outliers)

Dataset	Chi-3	Chi-5	Ish05	E-Alg	HFRBCS	C4.5	EEO	
							$\hat{\varepsilon}$	ε^*
abalone19	62.69%	66.71%	66.09%	0.00%	70.19%	15.58%	-	-
abalone9-18	63.93%	66.47%	65.78%	32.29%	67.56%	53.19%	-	-
ecoli-0_vs_1	92.27%	95.56%	96.70%	95.25%	93.63%	67.95%	-	-
ecoli-0-1-3-7_vs_2-6	71.04%	49.57%	71.31%	73.65%	71.48%	71.21%	**78.44**%	**84.05**%
ecoli1	85.28%	86.05%	85.71%	77.81%	84.18%	76.10%	41.72%	57.83%
ecoli2	88.01%	87.64%	87.00%	70.35%	87.62%	91.60%	28.92%	61.14%
ecoli3	87.58%	91.61%	85.39%	78.54%	90.81%	88.77%	27.31%	60.20%
ecoli4	91.27%	92.11%	86.92%	92.43%	93.02%	81.28%	76.83%	78.19%
glass0	64.06%	63.69%	69.39%	0.00%	76.57%	78.14%	26.22%	36.35%
glass-0-1-2-3_vs_4-5-6	85.83%	85.94%	88.56%	82.09%	88.37%	90.13%	66.95%	84.84%
glass-0-1-6_vs_2	40.84%	56.17%	41.18%	0.00%	58.37%	48.91%	37.03%	57.69%
glass-0-1-6_vs_5	71.48%	75.59%	88.77%	65.14%	77.96%	72.08%	41.86%	62.23%
glass1	64.90%	64.91%	59.29%	0.00%	73.66%	75.11%	38.01%	44.36%
glass2	47.67%	49.24%	43.55%	9.87%	54.84%	33.86%	31.67%	47.95%
glass4	84.96%	81.75%	78.27%	83.38%	70.39%	83.71%	59.66%	78.45%
glass5	81.56%	64.33%	89.96%	50.61%	68.73%	86.70%	40.99%	61.38%
glass6	83.87%	78.13%	86.27%	90.23%	86.95%	83.00%	73.37%	86.09%
haberman	58.91%	60.40%	62.65%	4.94%	57.08%	61.32%	37.54%	55.74%
iris0	100.00%	98.97%	100.00%	100.00%	100.00%	98.97%	28.46%	50.79%
new-thyroid1	87.44%	95.38%	89.02%	88.52%	95.58%	97.98%	74.25%	77.17%
new-thyroid2	89.81%	96.34%	94.21%	88.57%	99.72%	96.51%	74.25%	76.76%
page-blocks0	79.91%	87.25%	32.16%	64.51%	91.40%	94.84%	84.87%	84.85%
page-blocks-1-3_vs_4	91.92%	92.93%	94.53%	94.12%	98.64%	99.55%	64.89%	71.41%
pima	66.80%	66.78%	71.10%	55.01%	68.72%	71.26%	42.20%	64.32%
segment0	94.99%	95.88%	42.47%	95.33%	97.51%	99.26%	-	-
shuttle-c0-vs-c4	99.12%	98.71%	99.16%	98.40%	99.12%	99.97%	**97.96**%	**98.49**%
shuttle-c2-vs-c4	89.99%	78.34%	99.17%	100.00%	97.49%	99.15%	**94.14**%	**98.36**%
vehicle0	86.41%	84.93%	75.94%	39.07%	88.92%	91.10%	38.13%	56.81%
vehicle1	70.92%	71.88%	64.89%	3.09%	71.76%	69.28%	28.31%	47.46%
vehicle2	85.54%	87.19%	67.82%	43.83%	90.61%	94.85%	39.68%	45.57%
vehicle3	69.22%	63.13%	63.12%	0.00%	66.80%	74.34%	30.06%	51.54%
vowel0	98.37%	97.87%	89.03%	89.63%	98.82%	94.74%	65.08%	70.78%
wisconsin	88.91%	43.58%	95.78%	96.01%	88.24%	95.44%	**92.69**%	**94.05**%
yeast-0-5-6-7-9_vs_4	78.91%	75.99%	79.49%	59.99%	73.18%	74.88%	55.05%	63.11%
yeast1	67.69%	69.66%	51.41%	0.00%	71.71%	70.86%	38.04%	46.87%
yeast-1_vs_7	80.05%	63.02%	53.15%	27.55%	70.74%	67.73%	58.42%	59.63%
yeast-1-2-8-9_vs_7	76.12%	69.26%	48.55%	50.00%	69.37%	64.13%	57.73%	62.10%
yeast-1-4-5-8_vs_7	62.40%	58.76%	40.80%	0.00%	62.49%	41.19%	28.72%	54.55%
yeast-2_vs_4	86.80%	86.39%	70.85%	80.92%	89.32%	85.09%	65.61%	76.69%
yeast-2_vs_8	72.75%	78.76%	72.83%	72.83%	72.47%	78.23%	75.68%	78.14%
yeast3	90.13%	89.33%	77.06%	81.99%	90.41%	88.50%	44.86%	59.69%
yeast4	82.99%	83.07%	71.36%	32.16%	82.64%	65.00%	60.19%	69.79%
yeast5	93.41%	93.64%	94.94%	88.17%	94.20%	92.04%	54.38%	74.63%
yeast6	87.50%	87.73%	88.42%	51.72%	84.92%	80.38%	44.42%	66.13%

Good results were obtained also on other datasets, e.g. on 'glass6' (Fig. 3a), 'new-thyroid1', or 'yeast-2_vs_8'. Here, a majority of the inliers were concentrated near the cluster center; however, many inliers (their number was similar to the total number of outliers) were interspersed with the outliers. In such cases, estimation of ε became vague and perfect separation was not possible. Thus, the good EEO results on these datasets were coincidental and the presented method was not suited for them.

While the threshold values t for outlier detection can be directly set from the sorted distance visualization, estimating ε will represent a good practice in data reporting. The ε value defines the position of the outliers relative to the median, providing a data-independent approximation on outlier distribution comparable between studies.

The presented method was based on the Mahalanobis distance (1). While the distance was efficient for the proposed problem, we found the method to be computationally extravagant. Thus, we suggest an alternative approach based on the Euclidean distance for application where computational intensity would be of concern. The Euclidean distance in combination with standard deviation normalization [14] might provide equally good results while its time-complexity would be considerably lower.

5 Conclusion

The outlier analysis has the potential to mine valuable information from a complex dataset, but its sensitivity and specificity is dependent on both suitability of the method and the model, to the data. We designed EEO for the specifics of animal group behaviour observations, where the outliers could reveal alternative mechanisms determining group behaviour. Our testing on varied imbalanced sets demonstrated that the utility of the method is wider. The EEO was able to correctly classify outliers in datasets from engineering, microbiology or medicine. We therefore conclude that global outliers may be detected with EEO based on the threshold estimated from sums of pairwise Mahalanobis distances in datasets across fields that form one predominant multidimensional cluster with outliers distanced from it.

Acknowledgments. The work was supported by the University of Pardubice (PŠ) and the Czech Science Foundation grant number 17-20286S (NM).

References

1. MATLAB: Global optimization toolbox (R2016a) (2016). https://www.mathworks.com/help/gads/index.html
2. Aggarwal, C.C.: Outlier Analysis. Springer, New York (2013)
3. Angiulli, F., Basta, S., Lodi, S., Sartori, C.: GPU strategies for distance-based outlier detection. IEEE Trans. Parallel Distrib. Syst. **27**(11), 3256–3268 (2016)
4. Brereton, R.G.: The Mahalanobis distance and its relationship to principal component scores. J. Chemometr. **29**(3), 143–145 (2015)

5. Broom, D.M., Fraser, A.F.: Domestic Animal Behaviour and Welfare, 4th edn. CABI, Wallingford (2015)
6. Chi, Z., Yan, H., Pham, T.: Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition, vol. 10. World Scientific, Singapore (1996)
7. Deza, M.M., Deza, E.: Encyclopedia of Distances, 3rd edn. Springer, Heidelberg (2014)
8. Fernndez, A., del Jesus, M.J., Herrera, F.: Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. Int. J. Approximate Reasoning **50**(3), 561–577 (2009)
9. Gower, J., Lubbe, S., Roux, N.: Understanding Biplots. Wiley, New York (2010)
10. Han, J., Kamber, M., Pei, J.: Data Mining, 3rd edn. Morgan Kaufmann, San Francisco (2012)
11. Hawkins, D.M.: Identification of Outliers. Springer, Netherlands (1980)
12. Hawkins, S., He, H., Williams, G., Baxter, R.: Outlier detection using replicator neural networks. In: Kambayashi, Y., Winiwarter, W., Arikawa, M. (eds.) DaWaK 2002. LNCS, vol. 2454, pp. 170–180. Springer, Heidelberg (2002). doi:10.1007/3-540-46145-0_17
13. Ishibuchi, H., Yamamoto, T.: Rule weight specification in fuzzy rule-based classification systems. IEEE Trans. Fuzzy Syst. **13**(4), 428–435 (2005)
14. Kantardzic, M.: Data Mining: Concepts, Models, Methods, and Algorithms, 2nd edn. Wiley, Hoboken (2011)
15. Kohl, M.: Performance measures in binary classification. Int. J. Stat. Med. Res. **1**(1), 79–81 (2012)
16. Reeves, C.R., Rowe, J.E.: Genetic Algorithms: Principles and Perspectives: A Guide to GA Theory. Kluwer Academic Publishers, Norwell (2002)
17. Salzberg, S.L.: C4.5: programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Mach. Learn. **16**(3), 235–240 (1994)
18. Ward, A., Webster, M.: Sociality: The Behaviour of Group-Living Animals. Springer International Publishing, Heidelberg (2016)
19. Xu, L., Chow, M.Y., Taylor, L.S.: Using the data mining based fuzzy classification algorithm for power distribution fault cause identification with imbalanced data. In: 2006 IEEE PES Power Systems Conference and Exposition. pp. 1228–1233, October 2006