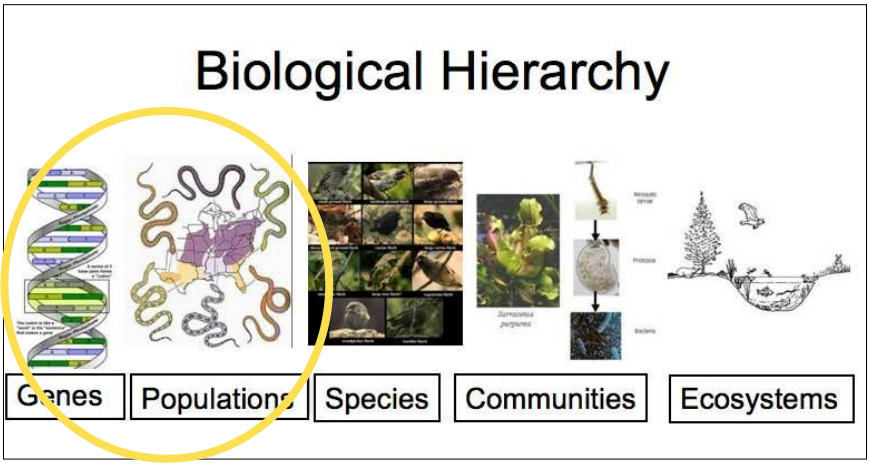
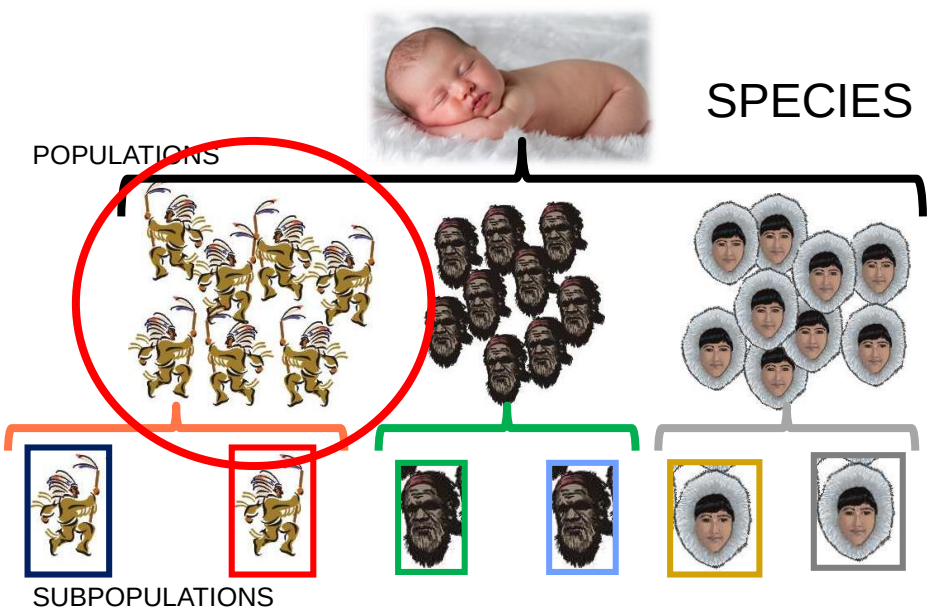
POPULATION GENETICS



I. GENETIC DIVERSITY

13 March 2017



I. GENETIC DIVERSITY – ANALYSIS OF SINGLE POPULATIONS

POPULATION
and problems of definition

- a population is a group of interbreeding indiviuals that exist together in time and space
- to develop the basic concepts of population genetics, we initially consider the **ideal population** = large, random-mating

ALLELE FREQUENCY

- proportion of an allele in comparison to all the others alleles of the same locus (gene) in a population sample
- basic characteristics for genetic diversity (variation) of a population
- population genetics studies genetic diversity and processes that have created it and influence it – i.e. the dynamics of distribution and frequency of alleles (genotypes → phenotypes), i.e. processes shaping **evolution**:

  increase of gen. diversity: mutation and migration
  decrease of gen. diversity: genetic drift (and natural selection)

MUTATIONS

increase genetic diversity
responsible for variation/heterogeneity in populations – essential to **evolution**

1. substitutions (transitions, transversions)

non-coding regions

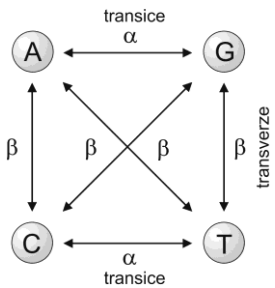
synonymous GTC → GTA
Val → Val

} = silent substitutions

nonsynonymous

missense GTC → TTC
Val → Phe

nonsense AAG → TAG
Lys → *ochre* (stop)



2.
insertion ACGGT → ACAGGT
deletion ACGGT → AGGT

} = indels
→ frameshift mutations

Mutation rate – rate at which number of various types of mutations occur in a given position over time

OBSERVATION

Callimorpha dominula

přástevník hluchavkový

Scarlet tiger moth

OBSERVATION

Callimorpha dominula

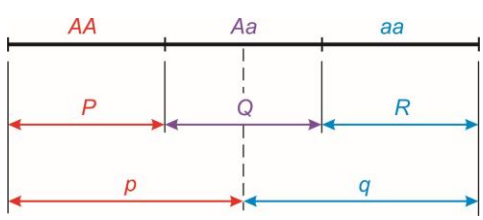
přástevník hluchavkový

Scarlet tiger mo

Table 3.1. Data from a collection of 1612 scarlet tiger moths.

Phenotype	No. of individuals
White spotting	1469
Intermediate	138
Little spotting	5

Genotype and allele frequency



Relative numbers = frequencies: genotype f.: P (G_{AA}), Q (G_{Aa}), R (G_{aa})

allele (gene) f.: p (A), q (a)

$P + Q + R = 1$
$p + q = 1$

Genotype	A_1A_1	A_1A_2	A_2A_2	Total
Number	n_1	n_2	n_3	N
Frequency	$P=n_1/N$	$Q=n_2/N$	$R=n_3/N$	
	$p = (2n_1 + n_2)/2N$		$q = (n_2 + 2n_3)/2N$	

Hardy-Weinberg Equilibrium (HWE)

Ex. Single locus with 2 alleles

Allele	Allele frequency
A	p
a	q

$p + q = 1$
p, q - Allele frequencies known from our samples

Genotype	Expected genotype frequency
AA	p^2
Aa	2pq
aa	q^2

= Hardy-Weinberg equilibrium

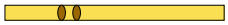
- Observed genotype frequencies (H_o) are known from our samples
- deviation of H_o from HWE conditions ⇒ for example χ^2 test

Expected heterozygosity, (H_e) under HWE
$H_e=1-(p^2+q^2)$ ..... for 1 locus with the allele frequencies *p* and *q*

Assumptions for ideal population in HWE

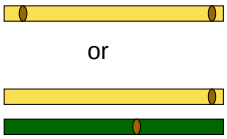
- random-mating
- negligible effect of mutations and migration („closed populations“)
- infinitely large population (negligible effect of random fluctuations in allele frequencies in time – genetic drift) – **in HWE population the allele frequencies are stable = do not change between generations**
- Mendelian inheritance of the analysed loci
- neutral loci – not under selection
- diploid, sexually reproducing organisms with discrete generations

- loci are independent from each other – test for „linkage disequilibrium“



2 loci physically close to each other
(decreased probability of recombination
- **linkage disequilibrium**)

vs.



or

2 loci physically distant
(probability of recombination not influenced
- **linkage equilibrium**)

LINKAGE DISEQUILIBRIUM (LD)

loci in LINKAGE EQUILIBRIUM – segregate independently of each other during meiosis

the most common reason for non-random association among loci (LD) is the **proximity of two loci on a chromosome** (others e.g. small pop. size – gen. drift, immigration, overlapping generations, admixture, etc.)

haplotype diversity – p(AB) ≠ p(A) x p(B)

in presence of LD:

we have **fewer** independent loci for our genetic analysis than anticipated

neutral loci (alleles) linked to selected ones will appear non-neutral

presence of LD **needs to be tested** when analysing data from multiple loci

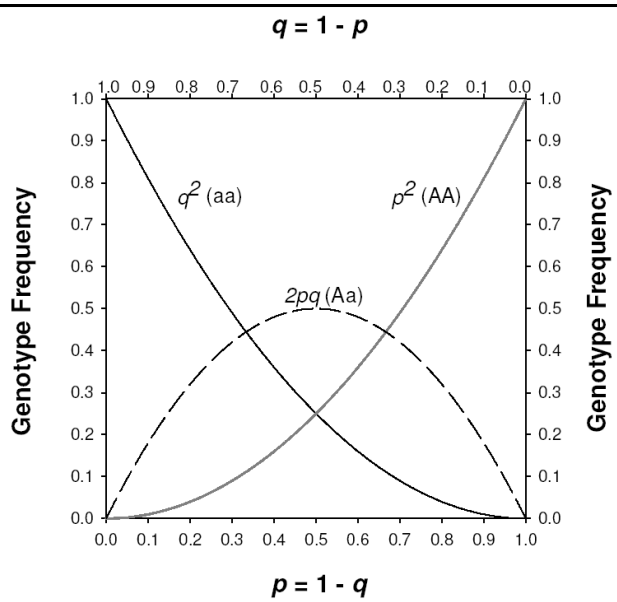


Figure 3.4 The combinations of homozygote and heterozygote frequencies that can be found in populations that are in HWE. Note that the frequency of heterozygotes is at its maximum when $p = q = 0.5$. When the allele frequencies are between 1/3 and 2/3, the genotype with the highest frequency will be the heterozygote.

Example of genetic diversity estimation in a sample of 4 individuals (on 4 loci)

Individual	Locus 1	Locus 2	Locus 3	Locus 4	Average
Ind 1	170/170	223/227	116/116	316/316	
Ind 2	170/172	223/225	112/112	316/316	
Ind 3	172/172	223/225	112/112	316/316	
Ind 4	170/172	223/227	112/112	316/316	
Počet alel	2	3	2	1	**2**
Ho	0,5	1,00	0	0	**0,375**
p	0,5	p = 0,5	0,75	1,00	
q	0,5	q = 0,25 r = 0,25	0,25	0	
He	0,5	0,625	0,375	0	**0,375**

$H_e=1-(p^2+q^2)$

$H_e=1-(p^2+q^2+r^2)$

Proportion of polymorphic loci (polymorphism) = 0,75

Is our population in HWE?

Callimorpha dominula

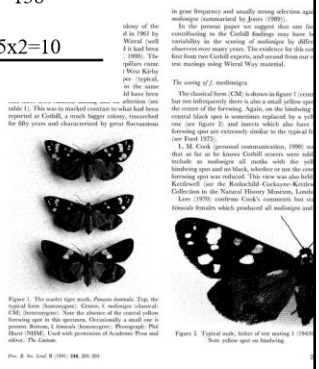
přástevník hluchavkový

Scarlet tiger moth

Is our population in HWE?

Table 3.1. Data from a collection of 1612 scarlet tiger moths.

Phenotype	No. of individuals	Assumed genotype	No. of *A* alleles	No. of *a* alleles
White spotting	1469	*AA*	1469x2=2938	-
Intermediate	138	*Aa*	138	138
Little spotting	5	*aa*	-	5x2=10


d the scarlet tiger moth, *Panaxia*
cies in the scoring of the
nigra

Deviation from HWE

- HWE test – e.g. Genepop software („exact probability tests“) – any case of **significant deviations from HWE** indicates that some of HWE assumptions were not fulfilled → detailed inspection required:
- heterozygote excess
  - negative **assortative mating** (i.e. intentional mating of distinct individuals)
  - used loci are advantageous in heterozygote situation (= balancing **selection** favouring heterozygotes, e.g. MHC genes)
  - **mutation**
  - **migration**
- heterozygote deficit
  - **inbreeding** (all loci are equally affected), assortative mating
  - genetic **structure** in populations
  - **null alleles** (only some loci affected by heterozygote deficit)

Quantifying genetic diversity

Polymorfism (proportion of polymorphic loci) - P

- **polymorphic locus** = with at least two alleles with having frequency of more numerous allele being **less or equal 0.95** (or 0.99)
- e.g. a population sample with four polymorphic loci out of five → P = 0.8

Number of alleles - N_a

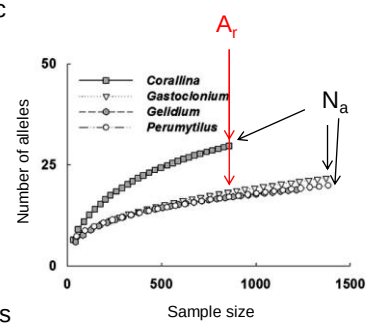
- number of alleles per locus (mean over loci)

Allelic richness - A_r

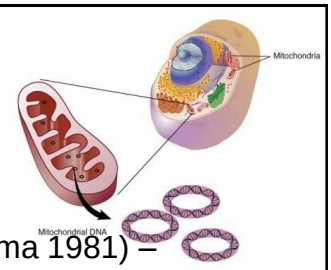
- number of alleles corrected for sample size (rarefaction method e.g. in FSTAT software)

Observed heterozygosity - H_o

- observed frequency of heterozygote genotypes (mean over loci)



HAPLOID DIVERSITY



- genetic diversity for haploid data

HAPLOTYPE DIVERSITY (h; Nei et Tajima 1981) – frequency of different haplotypes

$$H = \frac{N}{N-1}(1 - \sum_i x_i^2)$$

x_i –haplotype frequency of each haplotype in the sample
N – sample size

NUCLEOTIDE DIVERSITY (π; Nei 1987)

– quantifies the mean nucleotide divergence between sequences

– probability that two randomly chosen homologous nucleotides will be identical

$$\pi = \sum_{ij} x_i x_j \pi_{ij}$$

x_i and x_j – respective frequencies of the *i*th and *j*th sequences
π_{ij} – number of nucleotide differences per nucleotide site between the *i*th and *j*th sequences

WHAT INFLUENCES GENETIC DIVERSITY?

- influenced by a multitude of factors
- varies considerably between populations

MOST IMPORTANT DETERMINANTS OF GENETIC DIVERSITY:

- genetic drift
- population bottlenecks
- natural selection
- methods of reproduction

GENETIC DRIFT

population not infinitely large → population not in HWE → increase of influence of CHANCE → allele frequencies vary between generations

in absence of selection, each allele goes to:

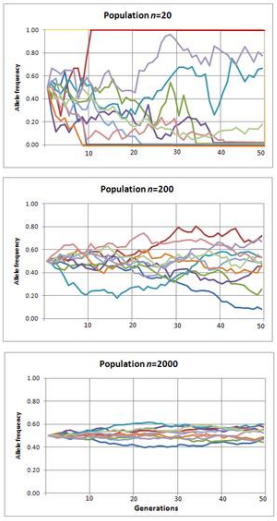
1. fixation
2. extinction

more quickly in smaller populations

→ **DECREASE of genetic diversity**

genetic drift – process causing a population´s allele frequencies to change from one generation to the next as a result of **CHANCE**

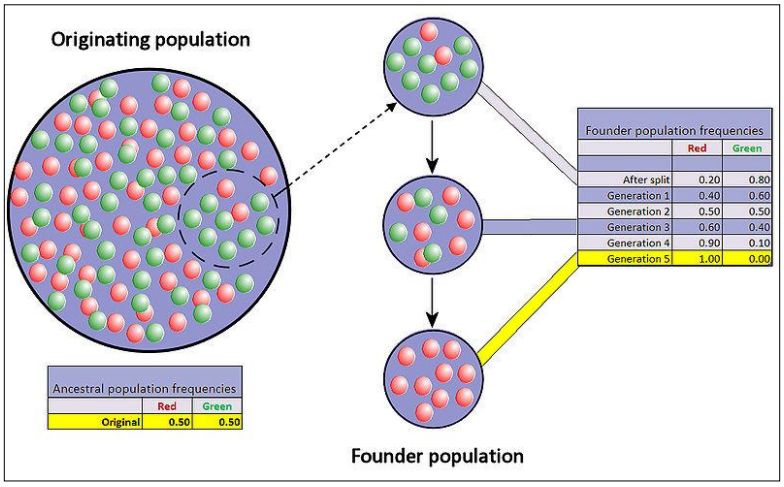
GENETIC DRIFT

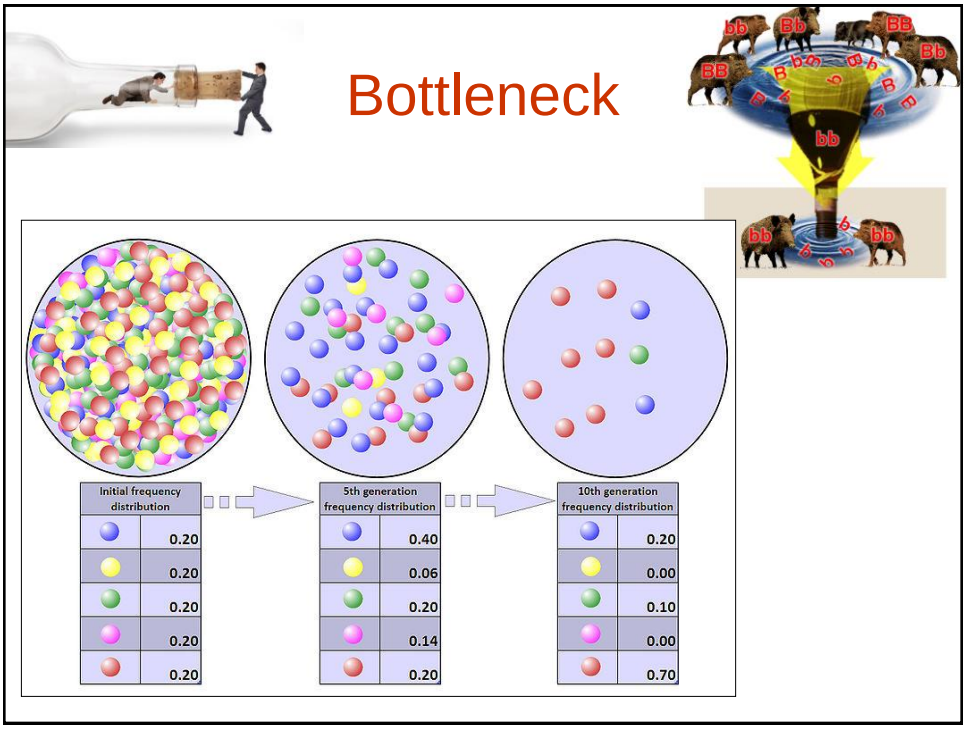


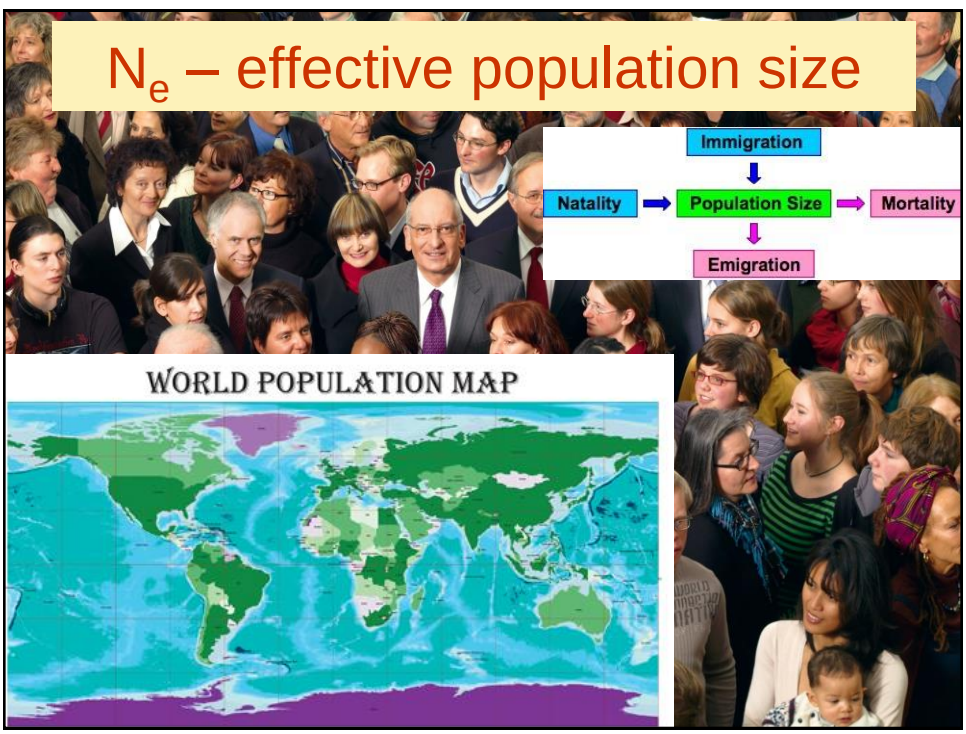
very profound effect of genetic drift in small populations – **founder effect, bottleneck**

inextricable link between genetic drift and population size – **the effective population size**

Founder effect




Bottleneck
bb
Bb
BB
Initial frequency distribution
0.20
0.20
0.20
0.20
0.20
5th generation frequency distribution
0.40
0.06
0.20
0.14
0.20
10th generation frequency distribution
0.20
0.00
0.10
0.00
0.70


N_e – effective population size
Immigration
Natality
Population Size
Mortality
Emigration
WORLD POPULATION MAP

N_e – effective population size

vs. N_c – census population size (may be estimated from N_e – see Luikart *et al.* 2010 *Conserv Genet*)

all else being equal, LARGE pops are MORE LIKELY to survive than small pops

N_e – reflects the rate at which genetic diversity will be lost following genetic drift (this rate is inversely proportional to a population´s N_e)

single-sample estimators of Ne – level of LD due to drift

double sample estimators of Ne – temporal changes in allele frequencies due to genetic drift

OVERVIEW

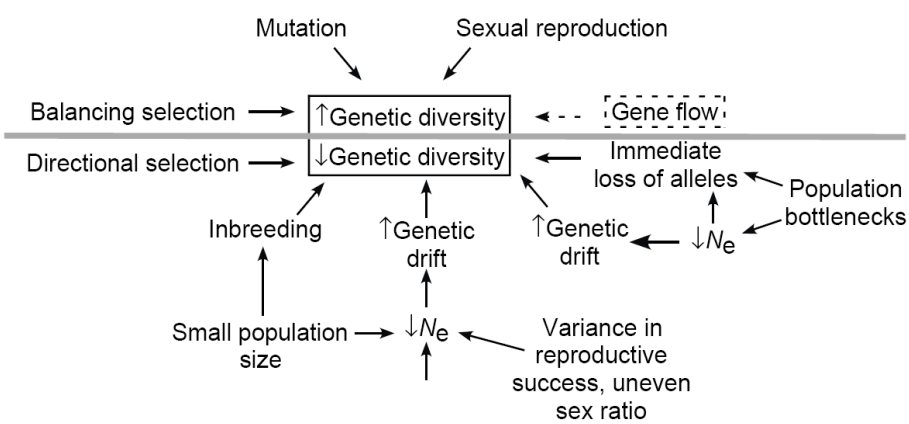


Figure 3.16 An overview of some of the main factors that influence levels of genetic diversity within populations.

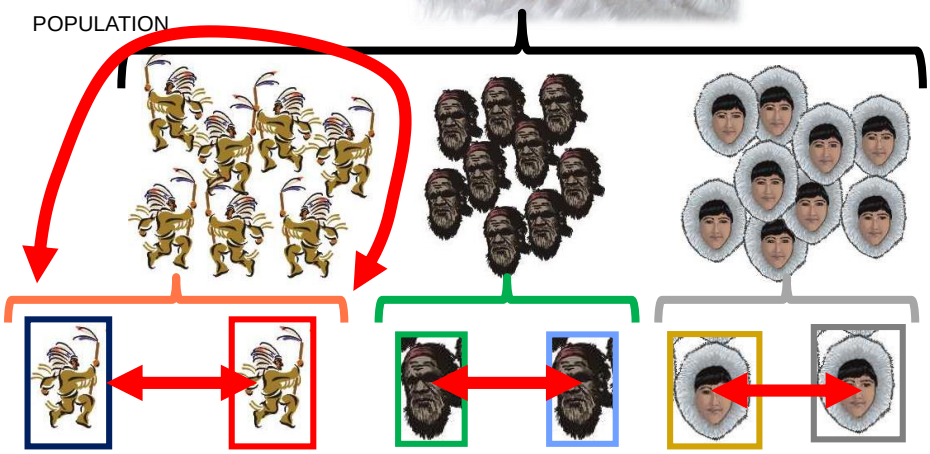
Freeland *et al.* 2011

POPULATION GENETICS

30 March 2017

SPECIES

POPULATION



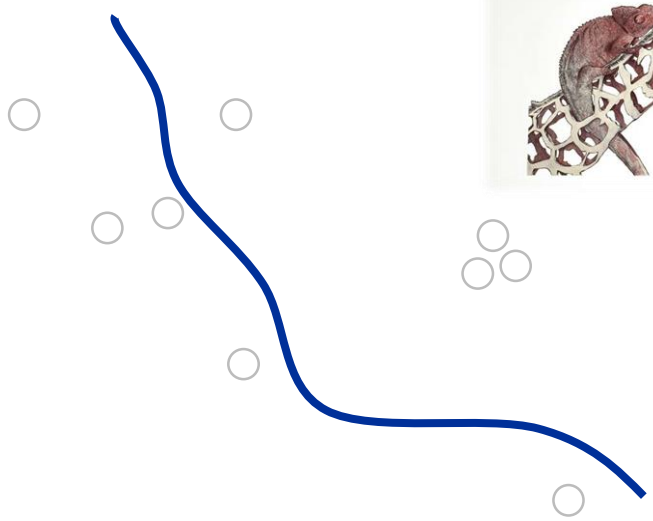
SUBPOPULATIONS

II. INTERPOPULATION VARIATION - GENETIC STRUCTURE OF POPULATION

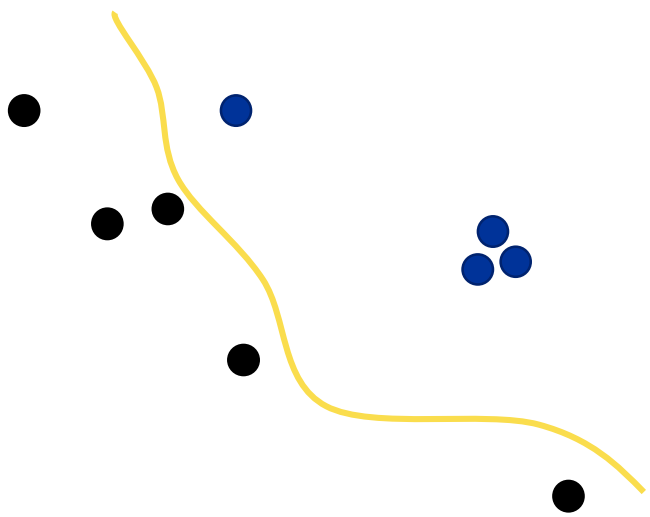
Assumption for population structure analysis:

- **neutral loci** = no effect of selection included
- classical population genetics approach = populations are *(thought to be)* known (e.g. we want to quantify level of genetic differentiation between two localities / ?populations)
- BUT populations are **not usually known** (e.g. due to no obvious spatial heterogeneity over the distribution range) – we want to **reveal any potential population differentiation/structure according to our genetic data**

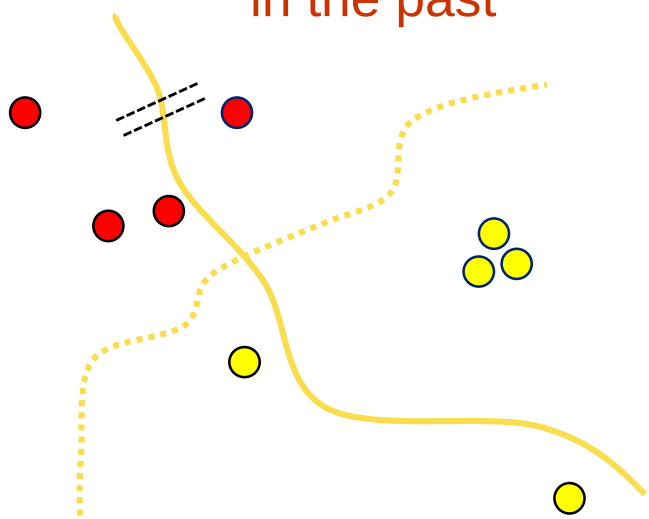
We have sampled animals in nature –
Is it one or several populations???



We are interested in genetic structure of populations



Recently observed genetic structure indicates what happened in the past



Genetic structure – any pattern in the genetic make-up of individuals within a population

AIMS:

- Detection of **any** genetic structure (subdivision) in a population (in my dataset)
- Are there any **differences** between „different“ (in space and time) populations?
- Quantification of such differences = **description of genetic structure in population**
- What factors shape (have shaped) these differences? e.g. **population history**
- Is there any migration/connection between different populations? = detection and quantification of **gene flow**, what influences gene flow (e.g. **spatial heterogeneity**)
- What happens during migration/connection of populations? = **hybridisation**

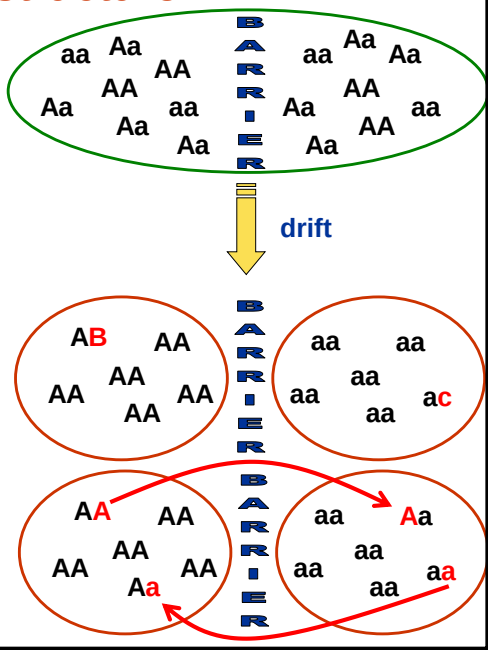
Population genetic structure

neutral markers

- **GENETIC DRIFT**
  - creates subpopulation differentiation
  (changes in allele frequencies – extremely up to fixation of distinct alleles)

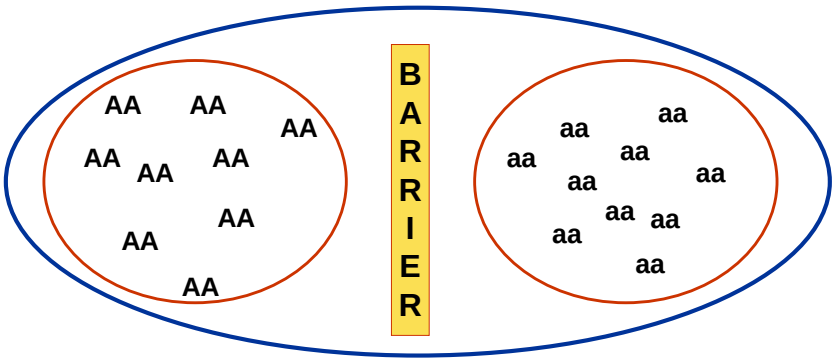
- **MUTATION**
  may increase differentiation (not necessarily – homoplasy)

MIGRATION (GENE FLOW)
- AGAINST subpopulation differentiation



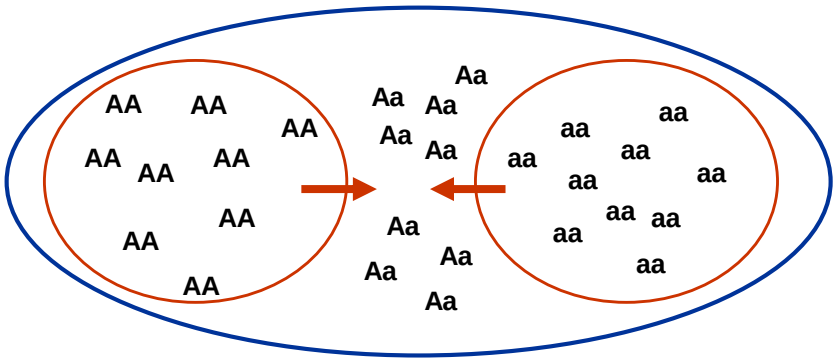
Effect of population structure on heterozygosity

- **Wahlund effect**– first documented by Swedish geneticist Sten Wahlund (1901-1976) in 1928
- two isolated subpopulations with fixed distinct alleles
- both SUBPOPULATIONS are in HWE, but the pooled dataset (the whole POPULATION) shows **deficit of heterozygotes**



Wahlund effect (isolate breaking)

Homozygosity reduction when subpopulations merge



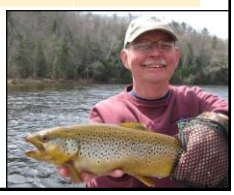
Wahlund, S. (1928) Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas*, 11: 65–106

Wahlund effect – an example

- Bunnersjöarna lake (northern Sweden) – „brown trout“
- one trait with 2 alleles

	170/170	170/172 (= Ho)	172/172	Total	p	2pq (=He)
Přítok	50	0 (0)	0	50	1.000	0.000
Odtok	1	13 (0.26)	36	50	0.150	0.255
Whole lake (expected)	51 (33.1)	13 (0.13) (48.9)	36 (18.1)	100	0.575	0.489

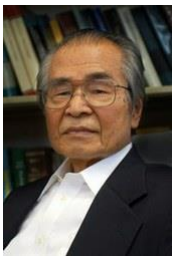
$p^2 = 0.575^2$ $q^2 = 0.425^2$



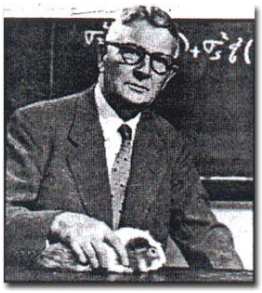
Ryman et al. 1979

Wright´s F-statistics

F_{IS}, F_{ST}, F_{IT}



Masatoshi Nei
***1931**

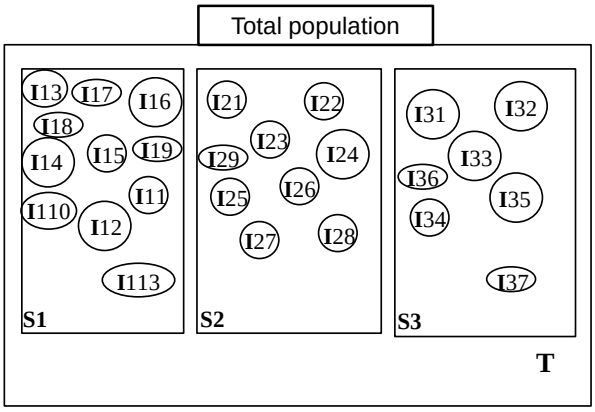


Sewall Wright
1889 - 1988

- Wright (1950), Nei (e.g. 1987)

- **detecting and describing population structure**
- describe heterozygosity (i.e. deviation from HWE) at different levels

Estimate of population structure effect on genetic diversity



- 3 levels (Total, Subpopulation, Individual)
- *x* subpopulations (*x* = 1 to *k*; here *k* = 3)
- each subpopulation has N_x individuals
- AA, AB, BB – genotypes with different symbols
- e.g. I1-13 = 13st individual from the 1st subpopulation

F-statistics and heterozygosity

H_I – averaged observed heterozygosity of an individual in a subpopulation
H_S – expected heterozygosity of an individual in a subpopulation **under HWE**
H_T – expected heterozygosity of an individual over the total population under HWE

$$H_I = \sum_{x=1}^{k} H_x / k$$

H_x = observed heterozygosity in subpopulation x

$$H_S = 1 - \sum_{i=1}^{j} p_{i,x}^2$$

$p_{i,x}^2$ = frequency of i-th allele in subpopulation x

$$\overline{H}_S = \sum_{x=1}^{k} H_S / k$$

averaged expected heterozygosity in subpopulation

$$H_T = 2 p_0 q_0$$

p_o = allele frequency in the total population

- for two alleles at a single locus (Wright 1950)
- more complicated for more alleles (Nei 1987)

F-statistics

$$F_{IS} = \frac{\overline{H}_S - H_I}{\overline{H}_S}$$

Heterozygosity decrease of an individual due to non-random mating in a subpopulation (vs. HWE)

Heterozygosity over all populations

Mean heterozygosity within subpopulations

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T}$$

Influence of division of the total population in subpopulations (i.e. heterozygosity decrease due to Wahlund effect)

$$F_{IT} = \frac{H_T - H_I}{H_T}$$

Total coefficient of inbreeding F_{IT} - measures heterozygosity decrease of an individual in relation to the total population

$$(1-F_{IT}) = (1-F_{ST})(1-F_{IS})$$

Weir & Cockerham (1984) ***f*** ***(~ F_{IS})***, ***θ*** ***(~ F_{ST})***, ***F*** ***(~ F_{IT})***
Correction for sample size and number of subpopulations

Computation of F-statistics

Mean allele A frequency in the whole population

	Subpopulation 1 (N_1=40)				Subpopulation 2 (N_2=20)					
Locus	AA	AB	BB	$p_{1(l)}$	AA	AB	BB	$p_{2(l)}$	$p_{0(l)}$	Note
Loc I	10	20	10	0.5	5	10	5	0.5	0.5	**HWE**
Loc II	16	8	16	0.5	4	4	12	0.3	0.4	**heterozygote deficit**
Loc III	12	28	0	0.65	6	12	2	0.6	0.625	**heterozygote excess**
Loc IV	0	0	40	0.0	20	0	0	1.0	0.5	**alternatively fixed alleles**

Computation of allele frequencies

	Observed heterozygosity		Expected heterozygosity			Wright´s F-statistics		
Locus	$H_{1(l)}$	$H_{2(l)}$	$H_{I(l)}$	$H_{S(l)}$	$H_{T(l)}$	$F_{IS(l)}$	$F_{ST(l)}$	$F_{IT(l)}$
Loc I	0.5	0.5	0.5	0.5	0.5	0.0	0.0	0.0
Loc II	0.2	0.2	0.2	0.46	0.48	0.565	0.042	0.583
Loc III	0.7	0.6	0.65	0.4675	0.46875	-0.39	0.0027	-0.387
Loc IV	0.0	0.0	0.0	0.0	0.5	---	1.0	1.0
Mean						0.058	0.261	0.300

Mean values of F-statistics may hide distinct evolution history of different loci

F-statistics

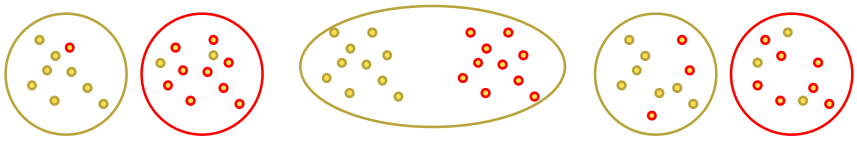
- F_{IS} *decrease of heterozygosity in local subpopulation*
  high values – inbreeding
- F_{IT} summary measure – limited use
- F_{ST} = **subdivision measure** = *limited gene flow between subpopulations* (i.e. existence of a barrier – Wahlund effect)
  - originally developed for estimation of the amount of allelic fixation due to genetic drift (**fixation index**)

Permutation test of Fst significance

1. Real measured populations

2. Merged into a single dataset

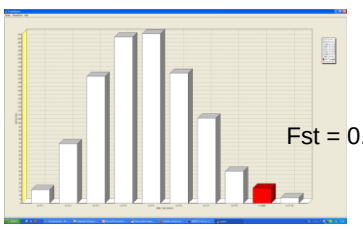
3. 1000 x randomly re-separated populations



Real Fst

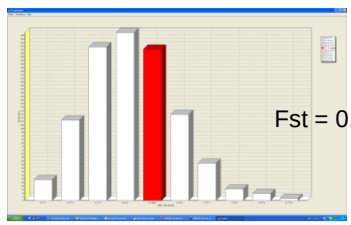
1000 x simulated Fst

TWO DIFFERENT CASES:



Fst = 0.072

0.8 % simulated values higher than real Fst
p = 0.008 (i.e. significant difference)



Fst = 0.0013

35.4 % simulated values higher than real Fst
p = 0.354 (e.g. non-significant difference)

F_{ST} computation – an example

	A/A	A/B (=Ho)	B/B	Total	p	2pq (=He)
Přítok	50	0 (**0**)	0	50	1.000	**0.000**
Odtok	1	13 (**0.26**)	36	50	0.150	**0.255**
Whole lake	51	13 **(0.13)**	36	100	0.575	**0.489**
(expected)	(33.1)	(48.9)	(18.1)			

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T} = \frac{0.489 - 0.128}{0.489} = 0.728$$

As a consequence of gene flow barrier:
Heterozygosity is about 72.8% lower than would be under HWE

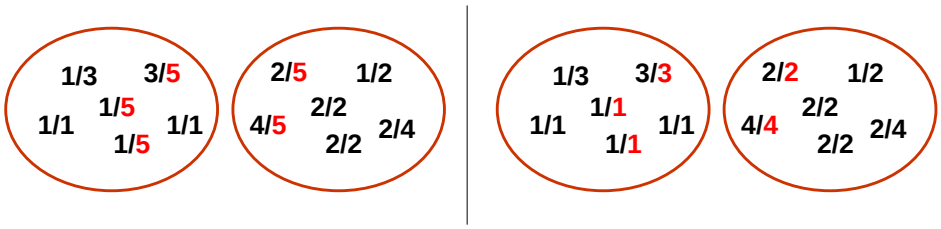
Ryman et al. 1979

F_{ST} analysis – BE AWARE

Global vs. pairwise indices

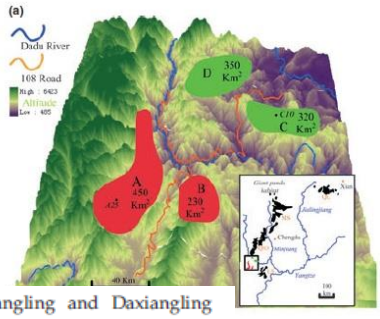
Absolute values depends on heterozygosity level of used loci!!!
(i.e. microsatellite-based F_{ST} cannot be compared to allozyme-based F_{ST})
Demands standardization: $F_{ST}' = F_{ST}/F_{STmax}$ (Hedrick 2005)
– e.g. GenAlEx

In case of null alleles presence: needs to be corrected!
(increased F_{ST} – increase of homozygosity); FreeNA software



Giant Panda

- 192 feces samples→ 136 genotypes→ 53 unique genotypes
- separation by a river (ca 26 ky ago) and by roads (recently)
- even the roads are important barriers, even if less



(Zhu et al., 2011)

Table 3 Pairwise F_{ST} in the Xiaoxiangling and Daxiangling populations

Patch	A	B	C	D
A				
B	0.033*			
C	0.107*	0.062*		
D	0.107*	0.097*	0.037*	

*Significant level after Bonferroni correction ($P < 0.01$).

G_{ST} (Nei 1973)

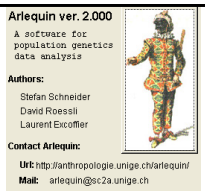
- Analogy of F_{ST} for **haploid (haplodiploid) organisms, mtDNA sequences**
- Takes into account **haplotype (gene) diversity** instead of heterozygosity
- *Haplotype diversity* = probability that any two randomly chosen sequences in a population will be different
- Pracuje tedy jen s frekvencemi alel, ne s procentem heterozygotů

R_{ST}

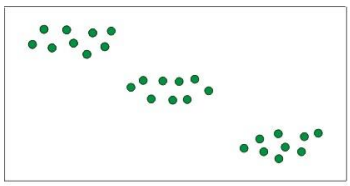
- Analogy of F_{ST}
- Takes into account **the size of alleles** (number of repeats in microsatellite loci)
- Assumption of a known mutation model
  assumption of SMM (*stepwise mutation model*)
- Indicates traces of mutations
  - **$R_{ST}>F_{ST}$ higher effect of mutations**
  - **$R_{ST}=F_{ST}$ higher effect of genetic drift**
- Randomisation tests for R_{ST} significance (Hardy et al. 2003, program SPAGeDi 1.1)

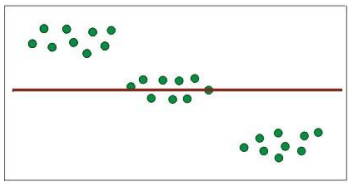
AMOVA

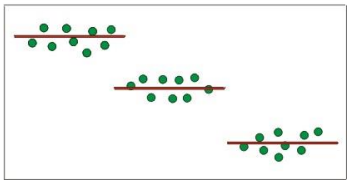
Excoffier et al. 1992



- **A**nalysis of **Mo**lecular **Va**riance
- Analysis of allele frequencies variance *(*before in *Cockerham & Weir 1987,1993)*
- ***Quantifies population differentiation***
- Takes into account difference between alleles – allelic state (mutations)
- Program ARLEQUIN
- Data:
  sequences
  microsatellites (assuming SMM *stepwise mutation model*)



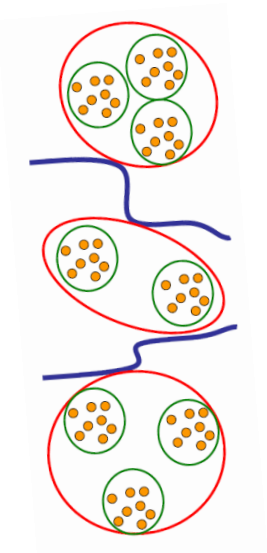




Hierarchical AMOVA

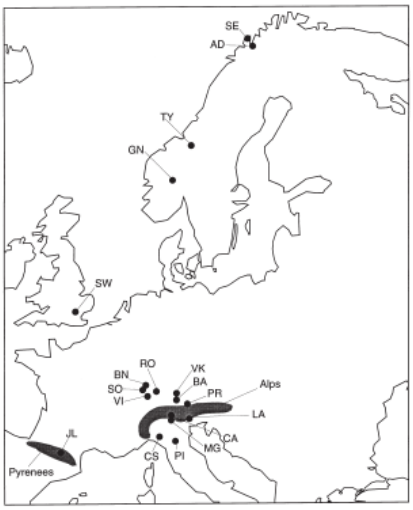
How much variation may be explained by:

- differentiation in big **groups of populations**
- differentiation in **populations** within the groups
- differentiation between **individuals** within the populations



Bombus pascuorum

Widmer & Schmid-Hempel 1999



	F/Φ	d.f.	SSD†	Variance component	% Total variance*
Among populations	F	17	77.71	0.07	4.51*
	Φ	17	5198.20	5.02	8.74*
Among regions	F	4	56.15	0.08	5.16*
	Φ	4	3464.94	4.58	7.49*
Among populations within regions	F	11	24.35	0.02	1.11*
	Φ	11	1773.71	2.16	3.53*
Between north and south of Alps	F	1	38.57	0.11	7.12*
	Φ	1	2622.89	7.25	11.74*
Among populations north and south of the Alps, respectively	F	16	39.14	0.02	1.46*
	Φ	16	2575.31	2.18	3.53*

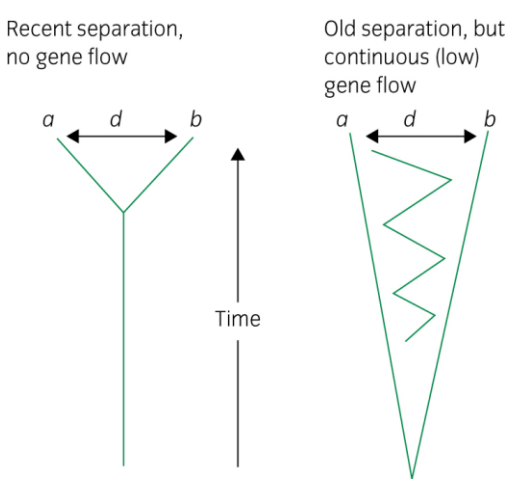
†Sum of squared deviations.
* $P < 0.001$.

Microsatellites, AMOVA
Most explained by the Alps

AMOVA and F-statistics

description of results, not causes ⟶ possible alternative explanations

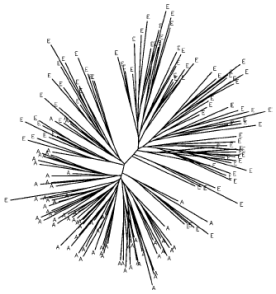
(use of population history analyses – based on coalescency and allele phylogenetics)



Clustering methods

DISTANCE-BASED methods

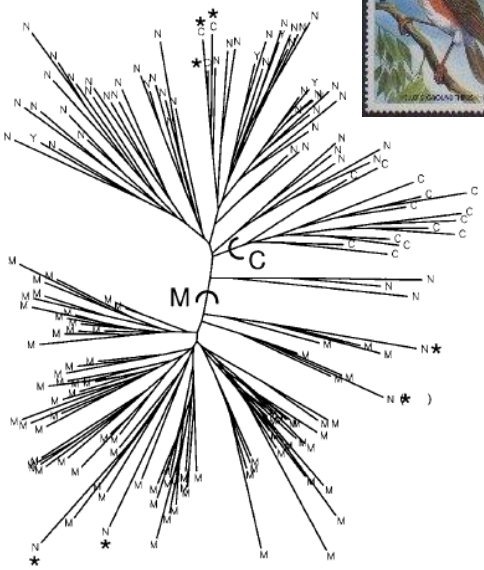
- a tree or a plot is constructed according to a pairwise distance matrix
- clusters then may be defined visually



MODEL-BASED methods

- observations from each cluster are random draws from some parametric model
- inference for the parameters corresponding to each cluster is done jointly with inference for the cluster membership of each individual
- standard statistical methods are used (e.g. maximum-likelihood in Bayeasian methods)

Turdus helleri



- Fragments of humid tropical forest
- Localities Chawia, Ngangao, Mbololo, Yale (Kenya)
- 7 microsatellite loci
- Neighbour-joining
- * wrongly clustered individuals

Clustering method based on microsatellite distances

Factorial correspondence analysis

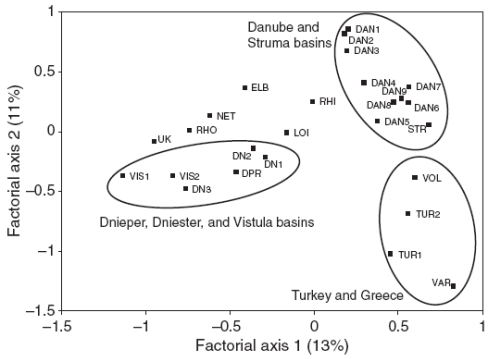


Fig. 2 A two-dimensional plot of the factorial correspondence analysis performed using GENETIX based on 12 microsatellite loci. Three geographical groups are bounded by grey lines.

- each locus as one variable, reduction of number of variables
- **Genetix** – inference about population structure
- individuals vs. populations

STRUCTURE program

Pritchard, Stephens and Donnelly 2000, Genetics

- a model-based Bayesian clustering method
- uses multilocus genotype data (e.g. microsatellites, RFLPs, SNPs; various levels of ploidy)
- MCMC algorithm

- INFERS POPULATION STRUCTURE:
  - presence of population structure
  - assignment of individuals to populations
  - identification of migrants or admixed individuals (parameter Q – individual membership coefficient)

Model implemented in STRUCTURE assumes:

- **K populations/clusters (K may be unknown)**
- each of K populations is characterized by a set of allele frequencies at each locus
- within each of K populations marker loci are at LINKAGE EQUILIBRIUM with each other and in HARDY-WEINBERG EQUILIBRIUM

under these assumptions each allele at each locus in each genotype is an independent draw from the appropriate frequency distribution, and this is completely specified by the probability distribution $P(X|Z,P)$

X – genotypes of the sampled individuals

Z – unknown populations of origin of the individuals

P – unknown allel frequencies in all populations

MODELS in STRUCTURE

ANCESTRY MODELS

- no admixture model
- admixture model
- linkage model
- models with informative priors

ALLELE FREQUENCY MODELS

- independent frequencies model
- correlated frequencies model

Ancestry models:

NO ADMIXTURE MODEL

- each individual is discretely from one of the K populations
- the output reports the posterior probability that individual *i* is from population K
- the prior probability for each population is 1/K

This model is appropriate for studying fully discrete populations and is often more powerful than the admixture model at detecting subtle structure.

Ancestry models:

ADMIXTURE MODEL

- individuals may have mixed ancestry
- each individual has inherited some proportion of its genome from each of the K populations = *Q*
- the output records the posterior mean estimates of these proportions

Recommended as a starting point for most populations.

"It is a reasonably flexible model for dealing with many of the complexities of real populations. Admixture is a common feature of real data, and you probably won't find it if you use the no-admixture model."

Allele frequency models:

INDEPENDENT FREQUENCIES MODEL

- the allele frequencies in each population are independent draws from a distribution that is specified by a parameter λ
- this prior says that we expect allele frequencies in different populations to be reasonably different from each other

Allele frequency models:

CORRELATED FREQUENCIES MODEL

- frequencies in the some populations are likely to be similar (probably due to migration or shared ancestry)
- this prior says that the allele frequencies in different populations may be quite similar between the populations

- better clustering for closely related populations
- but may increase the risk of over-estimating K

- *If one population is quite divergent from the others, the correlated model can sometimes achieve better inference if that population is removed.*

Falush, Stephens and Pritchard 2003, Genetics

MODELS in STRUCTURE

ANCESTRY MODELS

- no admixture model
- admixture model
- linkage model
- models with informative priors

ALLELE FREQUENCY MODELS

- independent frequencies model
- correlated frequencies model

How long to run it

it is not possible to determine suitable run-lengths theoretically

this requires some experimentation on the part of the user

burnin length: how long to run the simulation before collecting data to minimize the effect of the starting configuration

- typically a burnin of 10,000—100,000 is more than adequate

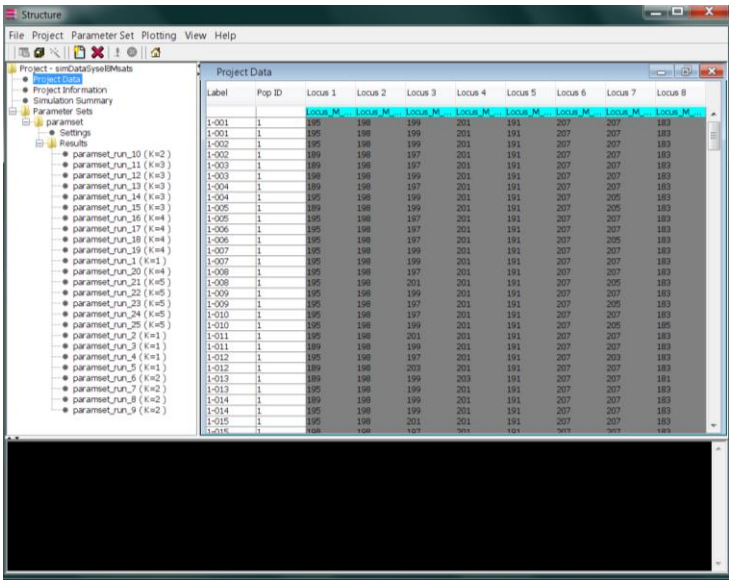
run length: how long to run the simulation after the burnin to get accurate parameter estimates

- several runs at each K, possibly of different lengths, and see whether you get consistent answers
- you can get good estimates of the parameter values (P and Q) with runs of 10,000–100,000 steps, but accurate estimation of Pr(X|K) may require longer runs
- at least 500,000

In practice your run length may be determined by your computer speed and patience as much as anything else.

STRUCTURE program

Pritchard, Stephens et Donnelly 2000, Genetics



Data format: genotypes of an individual in TWO rows

		loc_a	loc_b	loc_c	loc_d	loc_e
George	1	-9	145	66	0	92
George	1	-9	-9	64	0	94
Paula	1	106	142	68	1	92
Paula	1	106	148	64	0	94
Matthew	2	110	145	-9	0	92
Matthew	2	110	148	66	1	-9
Bob	2	108	142	64	1	94
Bob	2	-9	142	-9	0	94
Anja	1	112	142	-9	1	-9
Anja	1	114	142	66	1	94
Peter	1	-9	145	66	0	-9
Peter	1	110	145	-9	1	-9
Carsten	2	108	145	62	0	-9
Carsten	2	110	145	64	1	92

Needs to be specified:
number of individuals, ploidy of the data, number of loci, missing value symbol (integer)

Data format: genotypes of an individual in ONE row

sot_msat_structure - Notepad

File Edit Format View Help

Ssu1	Ssu5	Ssu7	Ssu8	Ssu13	Ssu15	SSU16	IGS-110b		IGS-1	MS45	MS56	MS41														
KV22	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	124	124	115	115
KV23	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	140	140	87	87	124	124	115	115
KV24	1	217	217	149	149	164	166	232	232	180	180	193	193	194	194	137	137	87	87	126	126	115	115
KV25	1	217	217	149	149	164	166	232	232	180	180	193	193	194	194	140	140	87	87	122	122	115	115
KV26	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	122	122	115	115
KV27	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	126	126	115	115
KV28	1	217	217	149	149	164	166	232	232	180	180	193	193	194	194	137	137	87	87	122	124	117	117
KV29	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	120	122	115	115
KV30	1	217	217	149	149	164	166	232	232	180	180	193	193	194	194	137	137	87	87	124	124	115	115
KV31	1	217	217	149	149	164	164	228	232	180	180	193	193	194	194	137	137	87	87	122	122	115	115
KV32	1	217	217	149	149	164	164	235	235	180	180	193	193	194	194	137	137	87	87	122	122	115	115
KV33	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	122	122	113	113
KV34	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	122	124	115	115
KV35	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	124	124	115	115
KV36	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	122	124	113	113
KV37	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	124	124	115	115
KV38	1	217	217	149	149	164	166	235	235	180	180	193	193	194	194	137	137	87	87	122	122	115	115
KV39	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	140	140	87	87	124	124	115	115
KV40	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	122	122	115	115
KV41	1	217	217	149	149	164	166	232	232	180	180	193	193	194	194	137	137	87	87	120	120	113	113
KV42	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	124	124	115	115
KV43	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	124	124	115	115
KV44	1	217	217	149	149	164	166	232	232	180	180	193	193	194	194	137	137	87	87	124	124	115	115
KV45	1	217	217	149	149	164	166	232	232	180	180	193	193	194	194	137	137	87	87	122	122	115	115
KV46	1	217	217	149	149	166	166	232	232	180	180	193	193	194	194	137	137	87	87	122	122	115	115
KV48	1	217	217	149	149	164	166	232	232	180	180	193	193	194	194	137	137	87	87	122	122	115	115
HR1	2	217	217	149	149	166	166	228	232	180	180	190	193	194	225	137	137	87	87	122	122	115	115
HR2	2	217	217	149	149	164	166	232	232	180	180	190	193	194	194	137	137	87	87	122	122	115	115
HR3	2	217	217	149	149	166	166	232	232	180	180	190	193	194	194	146	149	87	87	122	122	117	117
HR4	2	217	217	149	149	166	166	228	232	180	180	190	193	194	228	137	146	87	87	122	122	113	113
HR5	2	217	217	149	149	166	166	228	232	180	180	193	193	194	194	137	146	87	87	122	124	115	115
HR6	2	217	217	149	149	166	166	232	232	180	180	190	193	194	194	146	146	87	87	122	122	115	115
HR7	2	217	217	149	149	164	166	232	232	180	180	191	193	194	225	137	137	87	87	124	124	113	115
HR8	2	217	217	149	149	164	166	228	232	180	180	193	193	194	194	140	148	87	87	124	126	117	117
HR9	2	217	217	149	149	166	166	232	232	180	180	191	193	194	194	137	137	87	87	124	124	115	115
HR10	2	217	217	149	149	166	166	232	232	180	180	191	193	194	194	137	137	87	87	124	124	113	115
HR11	2	217	217	149	149	166	166	228	228	180	180	190	190	194	194	137	146	87	87	120	120	115	115
HR12	2	211	217	149	149	166	166	228	232	180	180	190	191	194	194	137	146	87	87	122	122	115	115
HR13	2	211	217	149	149	166	166	228	232	180	180	190	191	194	194	137	149	87	87	124	124	117	118
HR14	2	211	217	149	149	166	166	232	232	180	180	190	190	194	194	140	140	87	87	124	126	117	117
LET61	3	217	217	149	149	164	166	232	232	180	180	190	193	194	194	137	146	87	87	122	122	113	115
LET62	3	217	217	149	149	166	166	232	232	180	180	190	193	194	194	140	146	87	87	122	122	115	115
LET63	3	217	217	144	149	164	166	232	232	180	180	191	193	194	194	137	140	87	87	122	124	115	115
LET64	3	217	217	144	149	166	166	232	232	180	180	191	193	194	225	137	146	87	87	122	122	115	117
LET65	3	217	217	144	149	166	166	232	232	180	180	190	190	194	194	137	146	87	87	122	122	115	115
LET66	3	217	217	144	144	166	166	232	232	180	180	191	191	194	194	146	146	87	87	122	122	113	115
LET67	3	217	217	144	149	164	166	232	232	180	180	193	193	194	225	137	146	87	87	122	122	113	113
LET68	3	211	217	149	149	164	166	232	232	180	180	190	193	194	194	140	146	87	87	122	122	113	113
LET69	3	217	217	144	149	166	166	232	232	180	180	191	191	194	225	137	146	87	87	122	122	113	115
LET70	3	217	217	149	149	166	166	232	232	180	180	191	193	194	194	137	140	87	87	122	122	113	115
LET71	3	211	217	149	149	166	166	232	232	180	180	193	193	194	194	140	140	87	87	122	122	113	115
LET72	3	217	217	149	149	166	166	232	232	180	180	190	193	194	194	137	137	87	87	122	122	115	115
LET74	3	217	217	144	149	166	166	232	232	180	180	191	193	194	194	146	146	87	87	122	122	115	115
LET75	3	217	217	144	149	166	166	232	232	180	180	190	193	194	194	137	137	87	87	122	122	113	113
LET76	3	217	217	149	149	166	166	232	232	180	180	190	191	194	225	146	146	87	87	122	122	115	115
LET77	3	211	217	144	149	166	166	232	232	180	180	191	193	194	194	137	137	87	87	122	122	115	115
LET78	3	211	217	149	149	164	166	232	232	180	180	193	193	194	194	140	140	87	87	122	122	113	113
LET79	3	217	217	144	149	166	166	232	232	180	180	191	193	194	194	137	140	87	87	122	122	115	117
LET80	3	217	217	144	149	164	166	232	232	180	180	190	191	194	194	140	140	87	87	122	122	113	113
LET81	3	217	217	144	149	166	166	232	232	180	180	190	193	194	194	137	140	87	87	122	122	115	115
LET82	3	211	217	149	149	166	166	232	232	180	180	193	193	194	225	146	146	87	87	126	126	115	115
LET83	3	217	217	144	149	164	164	232	232	180	180	191	193	194	194	137	146	87	87	122	122	113	113

Data format: microsatellites of haploid organisms

project_data - Notepad

File Edit Format View Help

		msat_1	msat_2	msat_3	msat_4	msat_5	msat_6	msat_7	msat_8
1-001	1	240	195	219	225	199	197	191	221
1-002	1	240	205	219	207	201	185	193	227
1-003	1	240	205	219	207	199	237	193	221
1-004	1	240	205	219	207	205	237	197	197
1-005	1	240	195	219	223	199	191	187	219
1-006	1	240	205	219	223	203	191	193	219
1-007	1	240	205	219	207	203	191	197	227
1-008	1	233	195	219	235	199	191	193	227
1-009	1	240	205	219	223	199	237	193	221
1-010	1	240	201	219	209	205	237	197	219
1-011	1	240	205	207	225	201	237	197	213
1-012	1	233	205	219	225	201	237	197	209
1-013	1	240	207	219	223	199	191	193	221
1-014	1	240	205	219	207	203	237	187	227
1-015	1	240	205	207	207	201	191	197	227
1-016	1	240	205	207	223	197	237	193	209
1-017	1	233	201	219	225	201	237	193	197
1-018	1	240	195	219	225	199	185	197	213
1-019	1	240	205	207	225	199	237	193	219
1-020	1	233	195	219	207	201	237	197	197
2-001	2	235	203	215	189	203	191	191	199
2-002	2	235	201	213	189	203	191	191	183
2-003	2	167	203	215	189	203	191	191	183
2-004	2	235	201	215	189	203	205	201	195
2-005	2	167	203	213	189	203	191	191	183
2-006	2	231	197	215	189	203	191	201	183
2-007	2	235	201	213	189	203	191	191	183
2-008	2	167	201	213	189	203	191	201	195
2-009	2	240	201	213	189	203	205	191	195
2-010	2	235	201	213	189	203	191	201	183
2-011	2	235	201	213	189	203	191	191	183
2-012	2	240	201	213	189	203	205	201	183
2-013	2	240	201	213	189	203	191	191	183
2-014	2	235	201	213	187	203	191	201	183
2-015	2	233	201	215	189	203	205	201	183
2-016	2	240	219	215	189	203	205	201	183
2-017	2	167	201	215	189	203	191	201	183
2-018	2	233	201	213	189	203	205	201	183
2-019	2	240	203	215	189	203	191	191	183
2-020	2	240	201	213	189	203	205	201	183
3-001	3	165	213	207	201	197	195	205	203

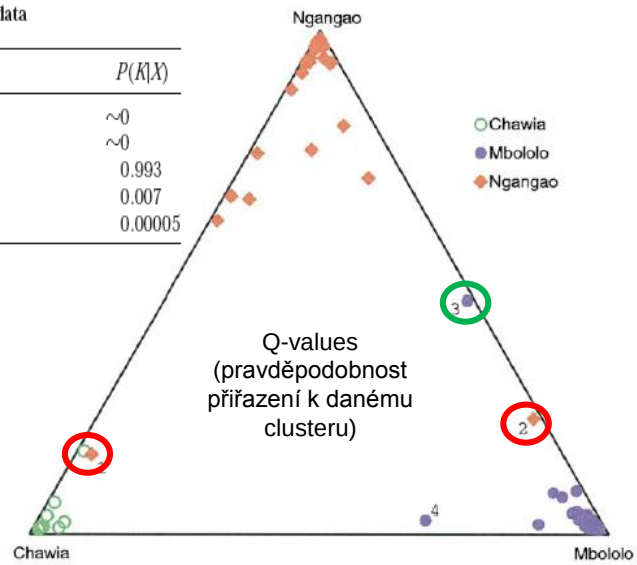
Program STRUCTURE – graphical output

Inferring the value of K, the number of populations, for the *T. helleri* data

K	$\log P(X\|K)$	$P(K\|X)$
1	−3144	~0
2	−2769	~0
3	−2678	0.993
4	−2683	0.007
5	−2688	0.00005

recent migrants

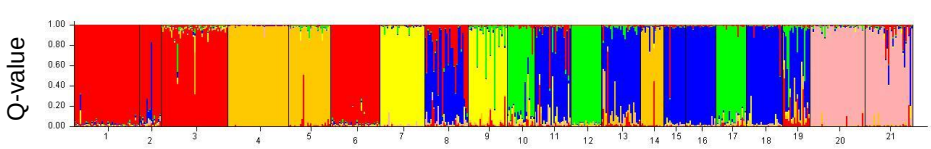
a hybrid?



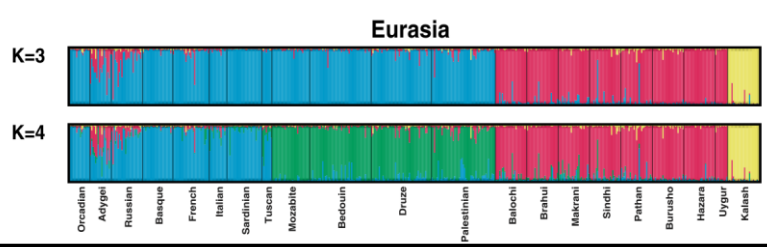
Q-values (pravděpodobnost přiřazení k danému clusteru)

Admixture model – allows assignement of an individual to several clusters

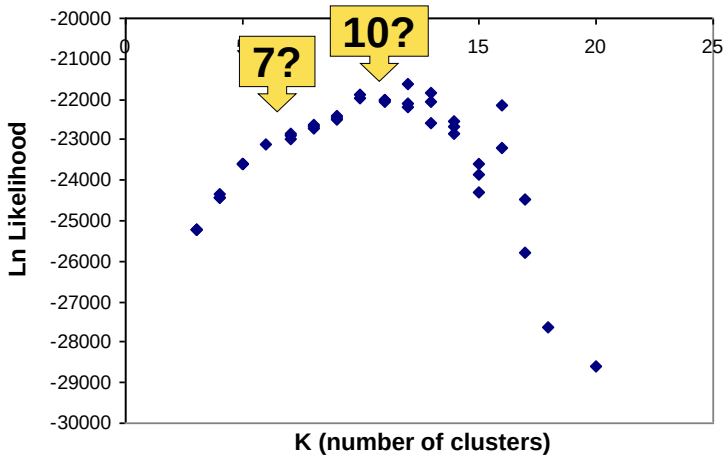
Barplot for K = 7



Genome proportion of each individual assigned to each of K clusters



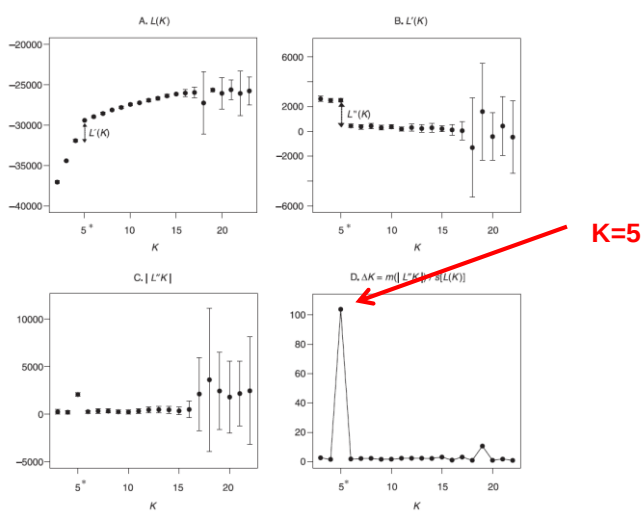
What K is the best???



Molecular Ecology (2005) 14, 2611–2620 doi: 10.1111/j.1365-294X.2005.02553.x

Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study

G. EVANNO, S. REGNAUT and J. GOUDET
Department of Ecology and Evolution, Biology building, University of Lausanne, CH 1015 Lausanne, Switzerland

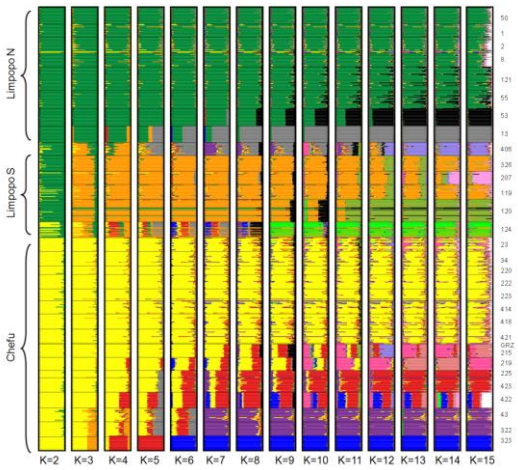


Post-processing of the STRUCTURE outputs

CLUMPAK was designed to aid users in four main objectives:

- Separate distinct solutions obtained from STRUCTURE-like programs.
- Compare and align solutions obtained for different K values.
- Compare results obtained using different models/data subsets/programs.
- Indicate the preferred value of K according to Evanno et al.

Graphical output from STRUCTURE – a serie of barplots with increasing K

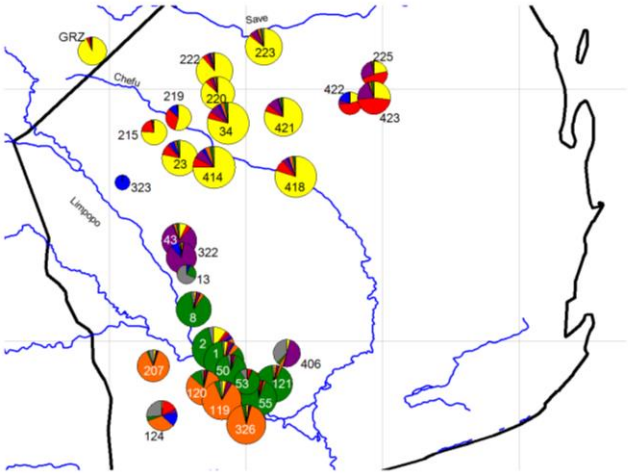


„forced clustering“

Picture of **hierarchical structure between clusters**

Bartáková et al. 2013

- Q-values for whole locality samples (not individuals)



Bartáková et al. 2013