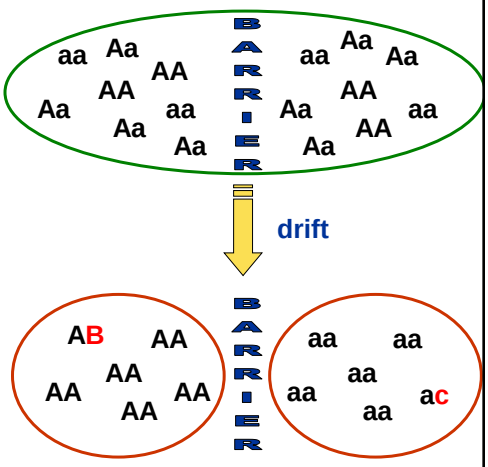
# Genetic structure of populations, drift, mutations

- **Drift**
  → differentiation of populations

  random changes in allele frequencies (may lead to fixation of alternative alleles)

- **Mutations & selection**
  increase differentiation

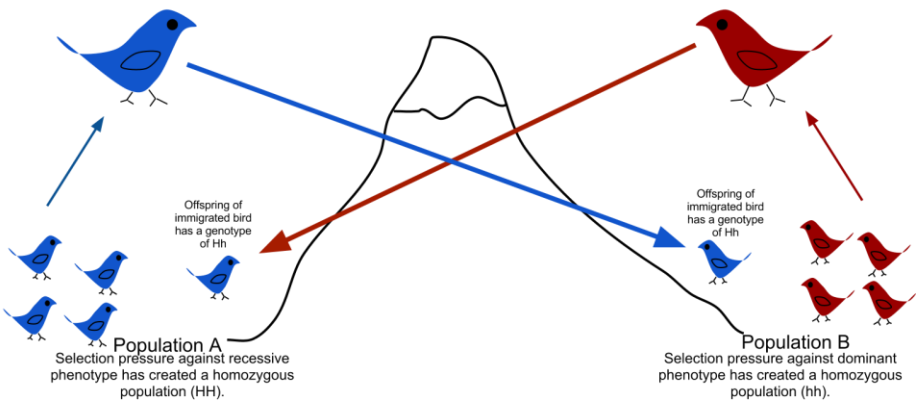



**Gene flow**
- acts against differentiation of subpopulations

# GENE FLOW

= the transfer of genes/alleles from one population to another

⟶ CHANGE OF ALLELIC FREQUENCIES



# MIGRATION versus GENE FLOW

- movement of **individuals** between pops
- immigrants may **not be reproducing** in a new pop! (even a strong migration/dispersal does not mean necessarily any gene flow)
- detectable (with substatntial difficulties) by direct ecological methods

- movement of **alelles (genes)** between pops
- via dispersion of individuals, propagules (gametes – pollen, seeds)
- passive in plants, mostly active in animals

- if strong → **homogenization of allele frequencies** between the pops
- **prevents** pop differentiation, divergence of pos, establishment of pop structure, and ultimately to speciation ---- by mixing the genepools
- **prevents** decrease of abilitiy to survive due to inbreeding

- estimable from genetic data

# Quantifying gene flow

1. Direct methods:
   - observation
   - Capture-Mark-Recapture sampling
   - telemetry

2. Indirect methods – methods of population genetics

   - **we have information about pop structure** (expected subpopulations or estimated from genetic data)

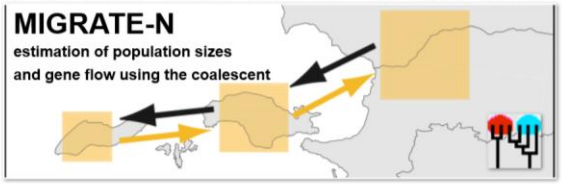
   - based on distribution of genetic variation
   - based on deviations from Hardy-Weinberg equilibrium

   - estimation based on $F_{ST}$
   - model-based methods based on the coalescent theory (eg. MIGRATE software)

Migrate-n Info Download Registration Tutorials FAQ Blog Eventtree Relationship

# MIGRATE-N

estimation of population sizes
and gene flow using the coalescent



Current Version is 3.6 [Fall 2013]
Updated prior system, including a gamma distribution prior. Addition of prior information into the histograms of the PDF outfile. MacOS 106, 10.7, and 10.8 version now have an experimental installer. I switched the copyright license to the MIT opensource license.

Important additional information to help run recent version of migrate:
- Most recent paper on migrate technology: Beerli and Palczewski 2010: Unified Framework to Evaluate Panmixia and Migration Direction Among Multiple Sampling Locations. Genetics (2010) vol. 185 pp. 313--326 (LINK)
- Tutorial on how to compare population genetic models (LINK)
- Opinion about issues with divergence and accuracy of migration rates (LINK)

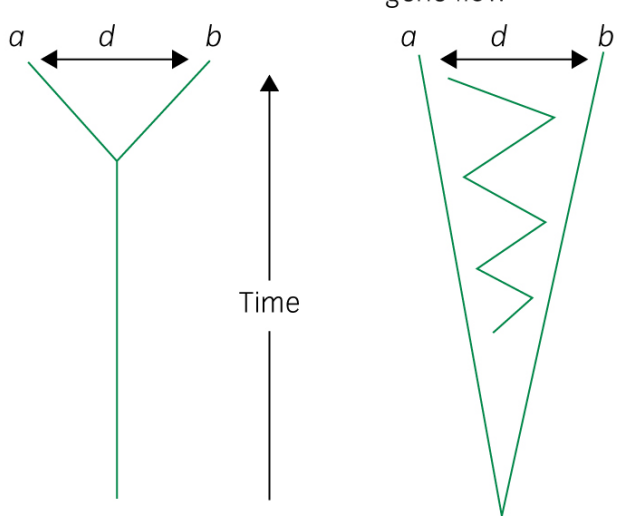
Known problems with current version:
- no known problem (yet)

Migrate estimates effective population sizes and past migration rates between n population assuming a migration matrix model with asymmetric migration rates and different subpopulation sizes.
Migrate uses Bayesian inference (or maximum likelihood) to jointly estimate all parameters. The analysis can be constrained to subsets of migration patterns, such as setting some migration rates between populations to zero or constrain to symmetric rates or average over all migration rates, or use Bayes factors to compare different hypotheses. Migrate can use single-locus or multi-locus data: sequence data using Felsenstein's 84 model with or without site rate variation, single nucleotide polymorphism data, microsatellite data using a stepwise mutation model or a brownian motion mutation model, and electrophoretic data using an 'k+1' allele model.
The output comes in two flavors: PDF and TEXT file. The file can contain:
- Bayesian inference: Estimates of maximum posterior values of parameters and credibility intervals in table form, figures of posterior distribution of parameters. Indications of convergence and effective sample size, presentation of frequency of migration events over time, very approximate skyline plots, genealogy with best likelihood (Printable through ET (eventtree -- http://popgen.sc.fsu.edu/Eventtree.html) or after some editing in Figtree (Andrew Rambaut)

http://popgen.sc.fsu.edu/Migrate/Migrate-n.html

Recent separation, no gene flow

Old separation, but continuous (low) gene flow

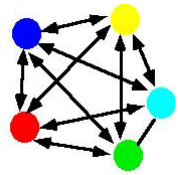


# Models of gene flow

- **island model**

  (Wright 1931)

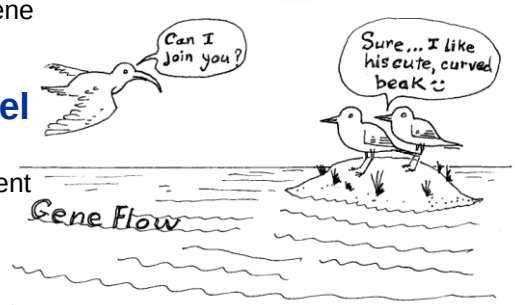
  assume same size of subpops
  assume symetrical flow of genes
  assume equal probability of gene exchange between subpops



- **stepping stone model**

  (Kimura 1953)

  exchange only between adjacent subpops




***$N_e m$* = number of adult, reproducing migrants between subpops per a generation (island model assumed!)**

It is just a rough estimation at a scale of „few“ and „a lot“

- **Private alleles** (Slatkin 1985) – useful for highly polymorphic markers = alleles occuring only in a single subpopulation

  $p(1)$ - frequency of private alleles
  $\ln p(1) = -0{,}505 \ln(N_e m) - 2.44$

- **F statistics** $$F_{ST} = \frac{1}{1+4N_e m}$$ (only for Fst > 0.05-0.10)

<u>Assumptions for using $N_e m$:</u>

- **island model** (= infinite number of subpops, no natural selection, equal size of all subpops, equal probability of migrant exchange between all subpops)
- **migration-drift equilibrium** (= no population expansion, no habitat fragmentation, no population bottleneck)

- *extreme case of a complete genetic isolation:* *Nm = 0, Fst = 1*
- *1 migrant every forth generation:* *Nm = 0.25, Fst = 0.5*
- *1 migrant every second generation:* *Nm = 0.5, Fst = 0.33*
- *1 migrant every generation:* *Nm = 1, Fst = 0.2*
- *2 migrants every generation:* *Nm = 2, Fst = 0.11*

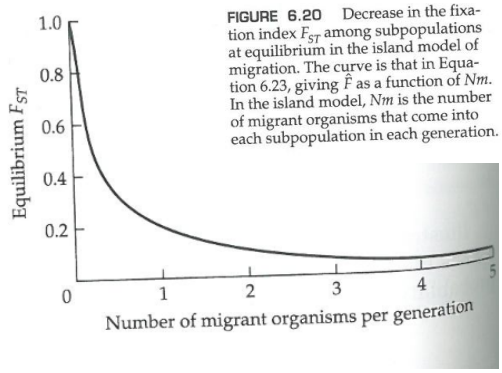



**FIGURE 6.20** Decrease in the fixation index $F_{ST}$ among subpopulations at equilibrium in the island model of migration. The curve is that in Equation 6.23, giving $\bar{F}$ as a function of $Nm$. In the island model, $Nm$ is the number of migrant organisms that come into each subpopulation in each generation.

## but be aware!!!

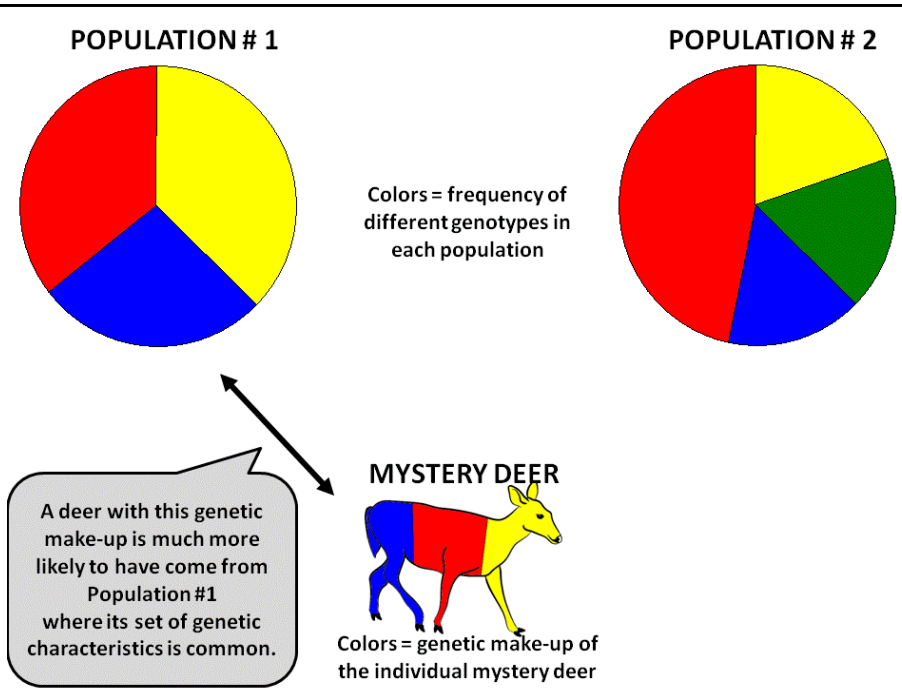
- even in a case of two very very distant populations
- $F_{ST} \rightarrow$ will never be equal to zero, $N_e m \rightarrow$ there had been exchange of individuals in the past
- even pops which have never exchanged any migrants will have never $N_e m$ equals to zero

# Inference of Recent Migration

- **BayesAss: Bayesian Inference of Recent Migration Using Multilocus Genotypes**
- *Reference*: G.A. Wilson and B. Rannala 2003. Bayesian inference of recent migration rates using multilocus genotypes. Genetics 163: 1177-1191.
- http://www.rannala.org/?page_id=245

# Assignment tests

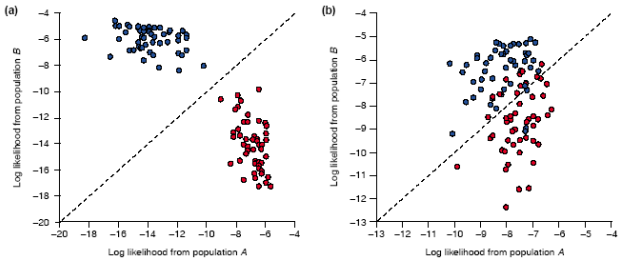
- assign individuals to their most likely population of origin
- done by comparison of individual genotypes to the genetic profiles of various populations
- vs $N_e m$ based indirect methods: **not comparing overall genetic similarities** between pops, but a maximum likelihood method to estimate probabilities that **a given genotypes** arose from alternative pops (Paetkau et al. 1995)
- all pops are assumed to be in HWE and the loci not in LD



# Population assignment tests

- program **GeneClass** (Piry et al. 2004)
- estimates probabilities of a certain genotype being from a certain pre-defined population – identification of recent migrants or samples of unknown origin (fight against poaching)
- may combine data of various genetic markers


# Depends on the level of genetic difference between populations

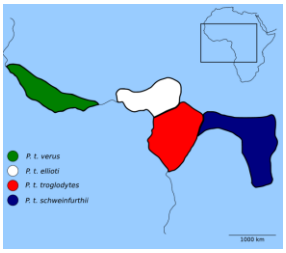



5 microsatellite loci
$F_{st}$ = 0.14
99.9% assigned correctly

5 microsatellite loci
$F_{st}$ = 0.04
90.2% assigned correctly

# Subspecies identification of chimpanzees in Czech ZOOs

- chimpanzees in ZOOs often of unclear origin
- genetic data from natural populations are available (300 msats, Becquet et al. 2007)
- 30 most informative microsatellites – genotypization of all chimpanzees in CZ
- GeneClass: assignment to the subspecies/populations




*Mapua et al. (2011)*

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | 12 | loci | | | 27 | loci | | | 30 loci | | | |
| 2 | | rank | score | rank | score | rank | score | rank | score | rank | score | rank | score |
| 3 | Assigned samp | 1 | % | 2 | % | 1 | % | 2 | % | 1 | % | 2 | % |
| 9 | 77-pop5-60 | Pop1 | 100 | Pop2 | 0 | Pop1 | 100 | Pop2 | 0.001 | Pop1 | 100 | Pop2 | 0.004 |
| 10 | 78-pop5-67 | Pop1 | 100 | Pop4 | 0.001 | Pop2 | 80.65 | Pop1 | 19.35 | Pop2 | 99.76 | Pop1 | 0.239 |
| 11 | Bamia | Pop2 | 57.13 | Pop1 | 42.87 | Pop1 | 100 | Pop2 | 0 | Pop1 | 100 | Pop2 | 0 |
| 12 | Babeta | Pop1 | 95.21 | Pop2 | 4.786 | Pop1 | 98.17 | Pop2 | 1.829 | Pop1 | 64.26 | Pop2 | 33.35 |
| 13 | Bambari | Pop2 | 94.66 | Pop1 | 5.3 | Pop1 | 84.77 | Pop2 | 15.23 | Pop2 | 83.34 | Pop1 | 16.66 |
| 14 | Bonie | Pop1 | 100 | Pop2 | 0 | Pop1 | 100 | Pop2 | 0 | Pop1 | 100 | Pop2 | 0 |
| 15 | Carl | Pop4 | 99.26 | Pop2 | 0.645 | Pop1 | 99.98 | Pop3 | 0.019 | Pop1 | 99.72 | Pop3 | 0.268 |
| 16 | Cindy | Pop4 | 99.98 | Pop1 | 0.022 | Pop3 | 89.59 | Pop4 | 8.614 | Pop4 | 89.06 | Pop3 | 10.19 |
| 17 | Dadula | Pop4 | 99.58 | Pop1 | 0.415 | Pop1 | 67.47 | Pop4 | 32.53 | Pop1 | 92.15 | Pop4 | 7.854 |
| 18 | Dais | Pop1 | 92.04 | Pop2 | 7.957 | Pop1 | 100 | Pop4 | 0 | Pop1 | 100 | Pop4 | 0 |
| 19 | Dingo | Pop1 | 100 | Pop2 | 0.003 | Pop1 | 98.98 | Pop4 | 0.98 | Pop1 | 99.84 | Pop2 | 0.102 |
| 20 | Dorka | Pop1 | 99.34 | Pop2 | 0.399 | Pop2 | 99.48 | Pop4 | 0.46 | Pop2 | 99.67 | Pop1 | 0.326 |
| 21 | Faben | Pop4 | 100 | Pop2 | 0.001 | Pop2 | 95.76 | Pop1 | 4.236 | Pop2 | 98.34 | Pop4 | 0.874 |
| 22 | Gina | Pop2 | 99.26 | Pop1 | 0.736 | Pop2 | 71.24 | Pop1 | 28.77 | Pop1 | 52.48 | Pop2 | 47.53 |
| 23 | Hope | Pop2 | 99.08 | Pop1 | 0.918 | Pop2 | 100 | Pop1 | 0.001 | Pop2 | 100 | Pop1 | 0 |
| 24 | Ingridy | Pop3 | 56.69 | Pop1 | 43.31 | Pop3 | 99.52 | Pop1 | 0.484 | Pop3 | 99.93 | Pop1 | 0.072 |
| 25 | Jakub | Pop1 | 99.99 | Pop2 | 0.015 | Pop1 | 100 | Pop2 | 0 | Pop1 | 100 | Pop3 | 0 |
| 26 | Janis | Pop4 | 99.42 | Pop3 | 0.499 | Pop3 | 95.23 | Pop2 | 4.756 | Pop3 | 86.24 | Pop4 | 9.103 |
| 27 | Jimmy | Pop4 | 99.42 | Pop1 | 0.565 | Pop1 | 100 | Pop2 | 0 | Pop1 | 100 | Pop2 | 0 |
| 28 | Judy | Pop1 | 99.84 | Pop4 | 0.158 | Pop1 | 82.71 | Pop3 | 17.29 | Pop3 | 97.73 | Pop1 | 2.273 |

12lociResults 27lociResults 30lociResults all

Připraven

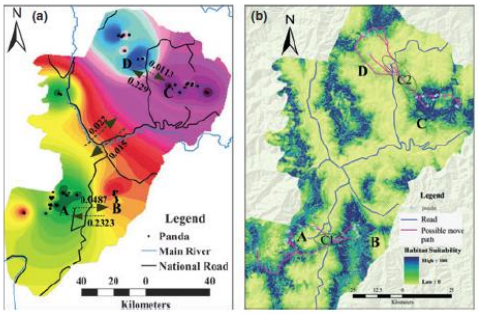
- some individuals are genetically clearly assigned to ESU (Evolutionary Significant Units = subspecies) – Zoo in Liberec, Dvůr Králové
- but also quite a few of hybrids (mainly Ostrava, Brno, etc.)

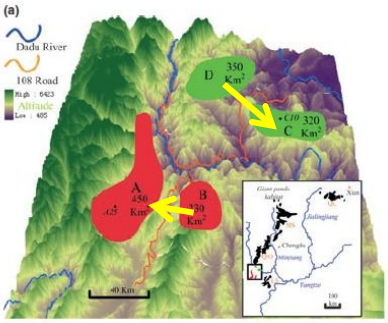
*Mapua et al. (2011)*

Zhu et al. 2011 Mol Ecol

# BayesASS
# GeneClass 2

- Giant panda




-Bayesian estimates of gene flow over few last generations
-identification of two possible first-generation migrants
-recommendations for conservation management – migration corridor construction

# Models of gene flow

- **Island model**

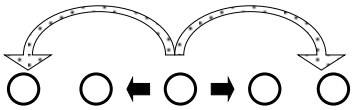
  (Wright 1931)
  assume same size of subpops
  assume symetrical flow of genes
  assume equal probability of gene exchange between subpops

- **Stepping stone model**

  (Kimura 1953)
  exchange only between adjacent subpops

- **Isolation by distance**

  Gene flow rate dicreases with increasing distance between subpops



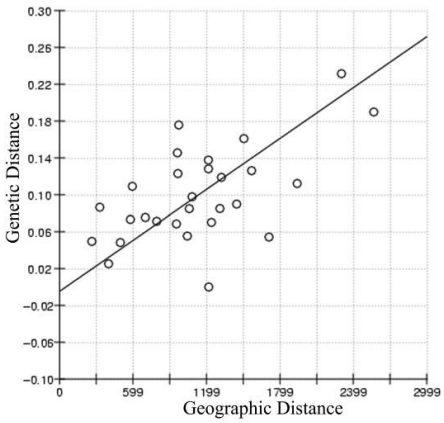
# Isolation by distance (IBD)

= the amount of gene flow between pops is inversely proportional to the geographic distances between them

- Sewall G. Wright (1943)
- regression of log-transformed gene flow estimate (eg. FST) and appropriate log-transformed geographic distances
- significance of correlation tested by **Mantel test** (does not assume independent population pairwise comparisons)
- relevant geographical scale (depends on dispersal abilities)
- migration-drift equilibrium must occur
- IBD (isolation-by-distance) is not
  - in very recently isolated populations
  - in completely isolated populations
  - in case of high amount of migration

# IBD detection

- correlation between matrices of genetic and geographic distances
- Mantel test
- e.g. Genepop

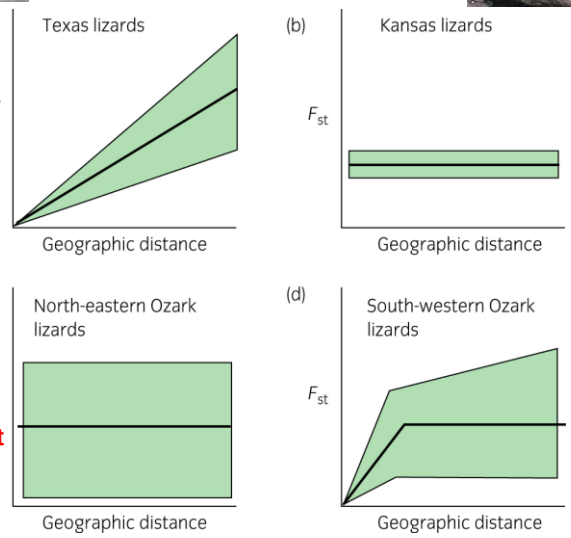


# Isolation by distance

*Crotaphytus collaris*

Hutchinson & Templeton 1999

no barriers for tens of thousands years
**equilibrium between drift and migration**




postglacially no barriers
**influence of migration**

postglacially fragmentation
**incluence of drift**

postglacially increasing fragmentation
**influence of drift**
at big scales
**equilibrium**
at small scales

Wellenreuther et al. 2010

# Population assignments

## Classical problems of population genetics

- Populations are defined, individuals *a priori* assigned to populations, we are interested in population characteristics (F-statistics) → i.e. pop genetic diversity and structure
- Populations are defined, but we want to assign individuals of unknown origin to them
- Cryptic population structure = nothing is known at the beginning → we want to estimate clusters (i.e. natural homogeneous populations) and assign the individuals to the clusters (**population assignments**)

## A. Direct methods

- morphological variation (geographical races)
- leg-bands or similar markers (ex. over one million *Ficedula hypoleuca* have been ringed in UK and Sweden – only six recaptured on wintering grounds in Africa
- satellite telemetry – expensive, not useful for small animals

## B. Biogeochemical approaches

- ratios of stable isotopes of naturally occurring elements (C, H, N, Sr) vary across the landscape
- determined by the relative frequency of C3 and C4 plants, climate, and bedrock

(1) geographical structure of isotopic ratio distributions

(2) knowledge about where animals incorporate isotopes

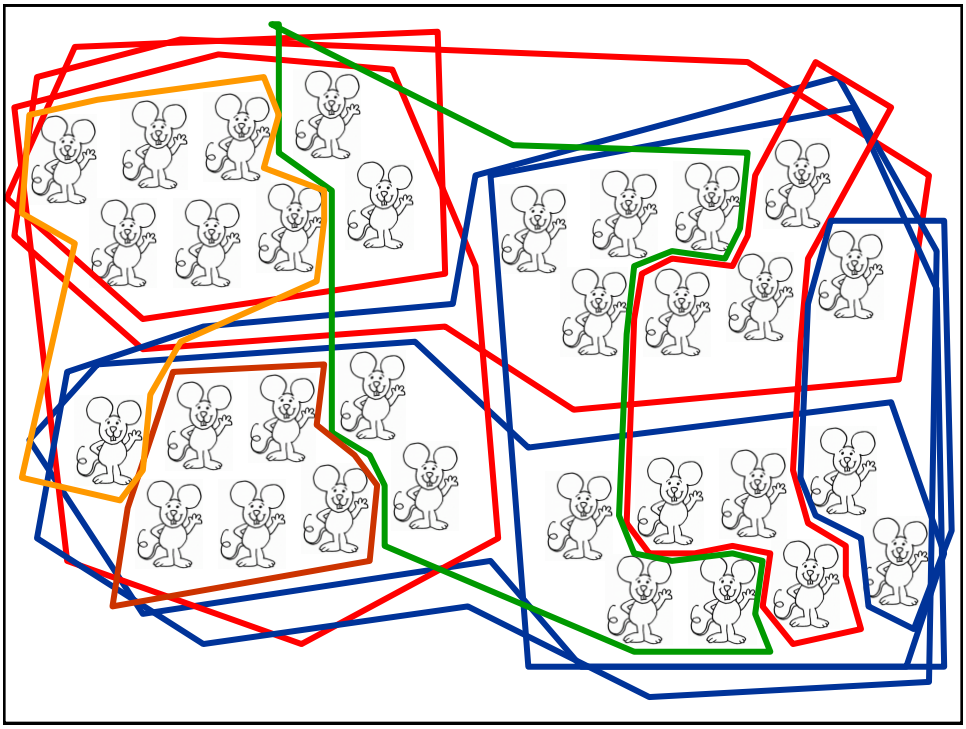
(3) tissue samples from individuals at different parts of their annual cycle

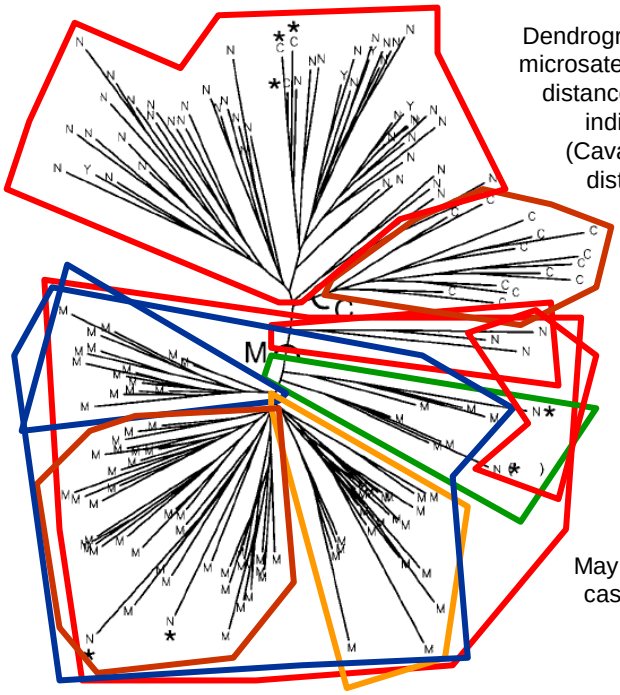
# C. Genetic approaches

- « few birds have rings, but everybody has genotype »
- genetic data about population structure
- problems: (1) low genetic differentiation between pops (intense dispersal), (2) low differentiation in temperate zone – recent postglacial colonization
- Solutions: (1) use more genetic markers, (2) study of parasite DNA (e.g. avian malaria) – parasites have quicker evolution, are more differentiated

# Individual-based assignments

- cryptic population structure
- unknown number of clusters
- **level of an individual**
- identify clusters and assign individuals to them simultaneously
- we have individual genotypes (sometimes also geographical coordinates)
- Data: msats (other codominant loci, *SINE*), AFLP





Dendrogram based on microsatellite genotype distances between individuals (Cavali-Sforza distances)

May be biased in case of too few markers

# LANDSCAPE GENETICS

- approach combining population genetics, spatial statistics (GIS) and landscape ecology
- aiming to quantify the **influence of landscape features and environmental variables on the distribution of allele frequencies** among populations

  = to understand the relationship between habitats and gene flow
- „landscape“ – the area that the organism of interest is utilizing (ie. number of various habitats of varying suitability)

- homogeneous vs. heterogeneous landscape ???
- homogeneous: panmictic population
- homogeneous, but larger than the dispersal distance of an individual: IBD
- heterogeneous (ie. various habitats): gene flow in not equal throughout the landscape

## Bayesian spatial clustering

### Spatially explicit analyses = spatial genetics = landscape genetics

- based on Bayesian clustering approach (of STRUCTURE type) – **individual-based models**
- for modelling is added information of both genetic data and geographical coordinates
- e.g. programs BAPS, TESS, Geneland (the „best“ number of clusters – K – is estimated automatically)

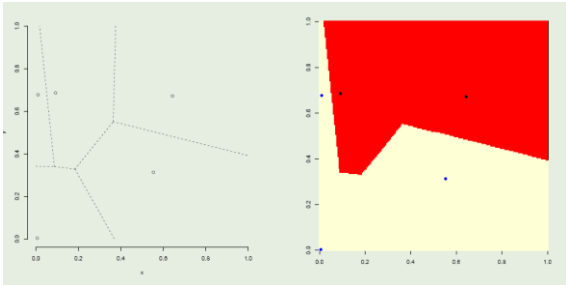
## Spatial models use **Voronoi diagrams**

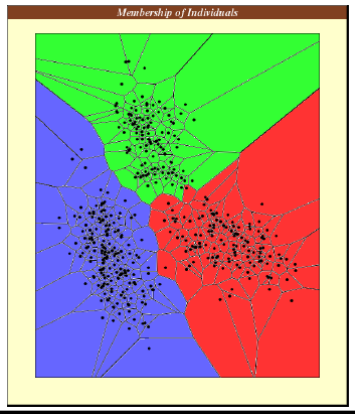
G. F. Voronoi (1868-1908)

Voronoi polygons, Dirichlet tessellation

- type of decomposition of metric space defined by distances to a given discrete set of objects in space, e.g. a discrete set of points

- separation of plane according to a given set of points *M*
- Voronoi diagram is a separation of plane in such a way that each point *b* from *M* is provided by an area *V(b)* whose all points are closer to the point *b* than to any other point of *M*

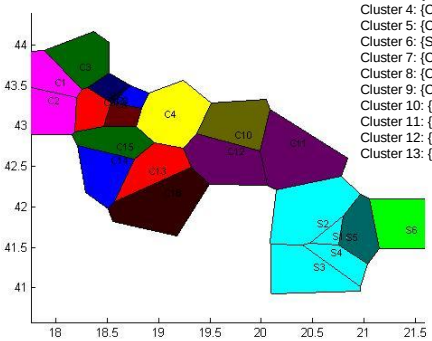


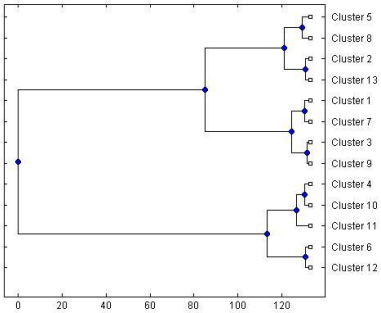



http://is.muni.cz/th/143320/fi_b_a2/animace/voroneho_diagram.html
http://ivankuckir.blogspot.cz/2011/03/voroneho-diagram-v-as3.html

The example of very fragmented populations: the best model in BAPS for Central and Southern *Dinaromys* populations (spatial clustering of groups of individuals): K=13 (i.e. evidence of very high structuration)

Best Partition:
Cluster 1: {C9, C13}
Cluster 2: {S6}
Cluster 3: {C8, C14}
Cluster 4: {C4}
Cluster 5: {C1, C2}
Cluster 6: {S1, S2, S3, S4}
Cluster 7: {C6}
Cluster 8: {C3, C15}
Cluster 9: {C5, C7}
Cluster 10: {C10}
Cluster 11: {C11, C12}
Cluster 12: {S5}
Cluster 13: {C16}




# program BAPS

software for Bayesian Analysis of genetic Population Structure

http://www.helsinki.fi/bsg/software/

# Geneland homepage

home papers applications courses events contact

## Overview

Geneland is a computer program for statistical analysis of population genetics data. Its main goal is to detect population structure in form of systematic variation of allele frequency that can be detected from departure from Hardy-Weinberg and linkage equilibrium. Geneland requires individual multilocus genetic data that are optionally geo-referenced. It implements several models that can make use of both geographic and genetic informations to estimate the number of populations in a dataset and delineate their spatial organisation.

Important areas of application include landscape genetics, conservation genetics, human genetics, anthropology and epidemiology.

Geneland can handle all common types of co-dominant or dominant markers (microsatellites, SNPs, AFLP, sequence data).

Since version 4.0.0, the program can also process phenotypic data and therefore any combination of genetic, phenotypic and geographic information.

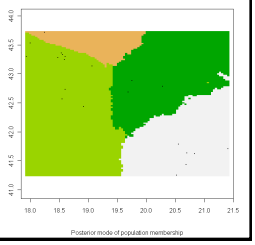
The program is released as an add-on to the free statistical program R and is currently available for Linux, Mac-OS and Windows. It includes **a fully clickable user interface** requiring no particular knowledge of R.

## 2 Models

Three types of quantities are involved:

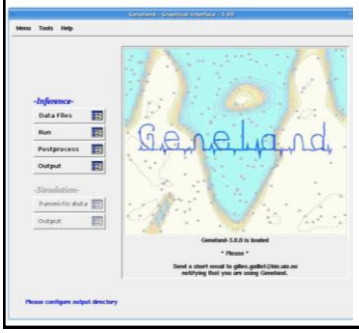
- the (usually unknown) number of populations $K$
- the parameters (or hidden variable) coding for population membership (of individuals and pixels)
- the parameters of the genetic model conditionally on the the number of populations and on population memberships.

They are modelled separately. $K$ is assumed to follow a uniform distribution between 0 and an upper bound $K_{max}$ prescribed by the user. The genetic and the spatial model are specified conditionally on $K$. This is described below.




# GENELAND

Population genetic and morphometric data analysis using R and the Geneland program



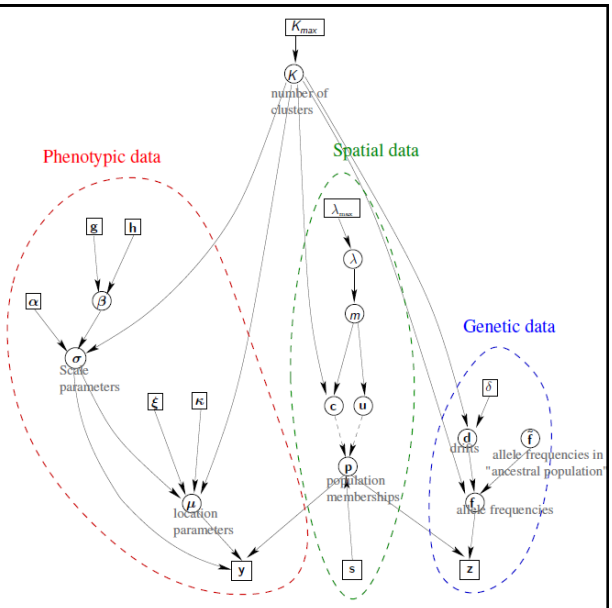
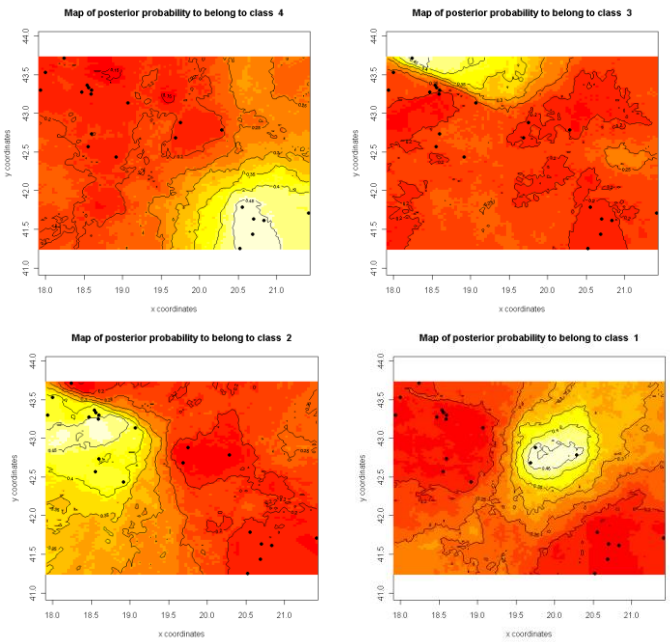



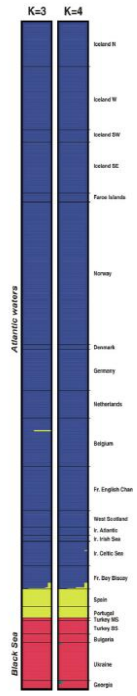

Figure 5: Graph of the global model. Continuous black lines represent stochastic dependencies, dashed black lines represent deterministic dependencies. Boxes enclose data or fixed hyper-parameters, circles enclose inferred parameters. Bold symbols refer to vector parameters. The red, green and blue dashed lines enclose parameters relative to the phenotypic, geographic and genetic parts of the model respectively. The parameters of interest to biologists are the number of clusters $K$, the vector $\mathbf{p}$ which encode the cluster memberships, and possibly allele frequencies $\mathbf{f}$, mean phenotypic values $\boldsymbol{\mu}$, phenotypic variance $\sigma^2$ which quantify the genetic and phenotypic divergence between and within clusters. Other parameters can be viewed mostly as nuisance parameters.

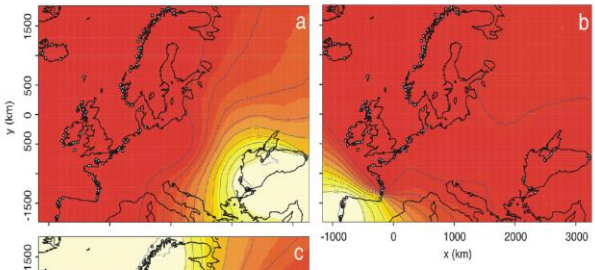
R platform

Posterior probability maps

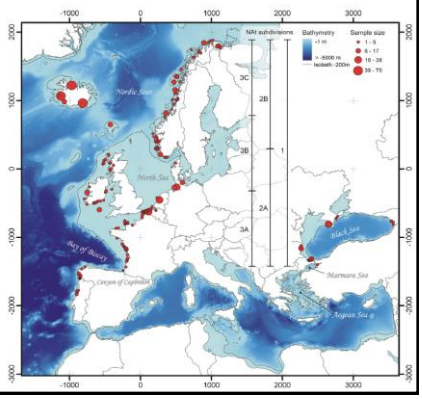



Spatial population genetics *Fontaine et al. 2007*




*Phocoena phocoena*

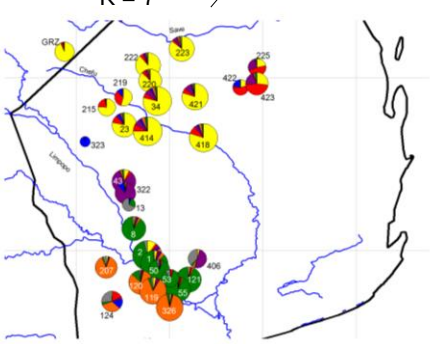


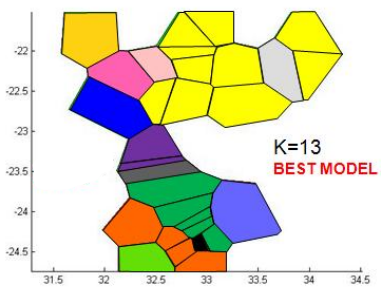
## Comparison of features of various „individual-based assignment“ programs

| | Structure | Partition | BAPS | Geneland |
|---|---|---|---|---|
| Estimate *K* | ● | ● | ● | ● |
| Spatial | ● | ● | ● | ● |
| Admixture | ● | ● | ● | ● |
| Inbreeding | ● | ● | ● | ● |
| Linked loci | ● | ● | ● | ● |
| Corr. freq. | ● | ● | ● | ● |
| Co-dom. markers | ● | ● | ● | ● |
| Null alleles | ● | ● | ● | ● |

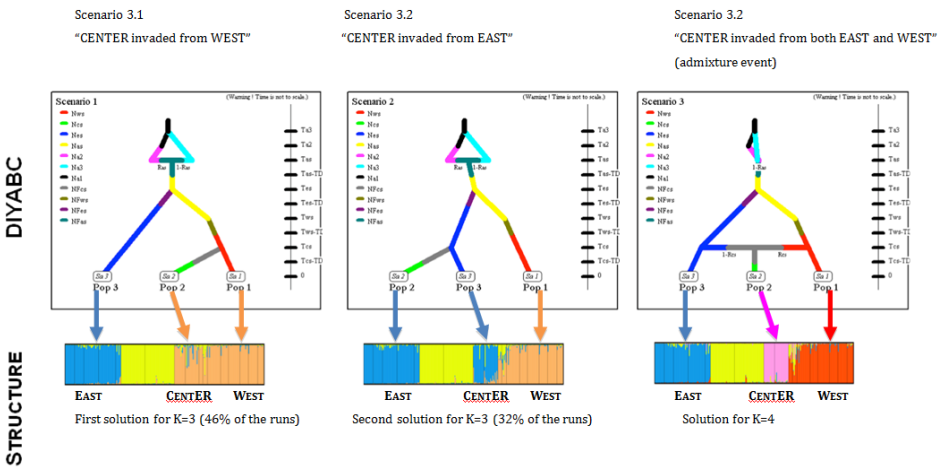
## STRUCTURE vs. BAPs

K = 7






Robust support of the population structure




# Population structure - summary

| | Connected populations (gene flow) | Isolated populations (no gene flow) |
|---|---|---|
| $N_e$ | ↑ | ↓ |
| Genetic drift | ↓ | ↑ |
| Genetic diversity | ↑ | ↓ |
| Population differentiation | ↓ | ↑ |