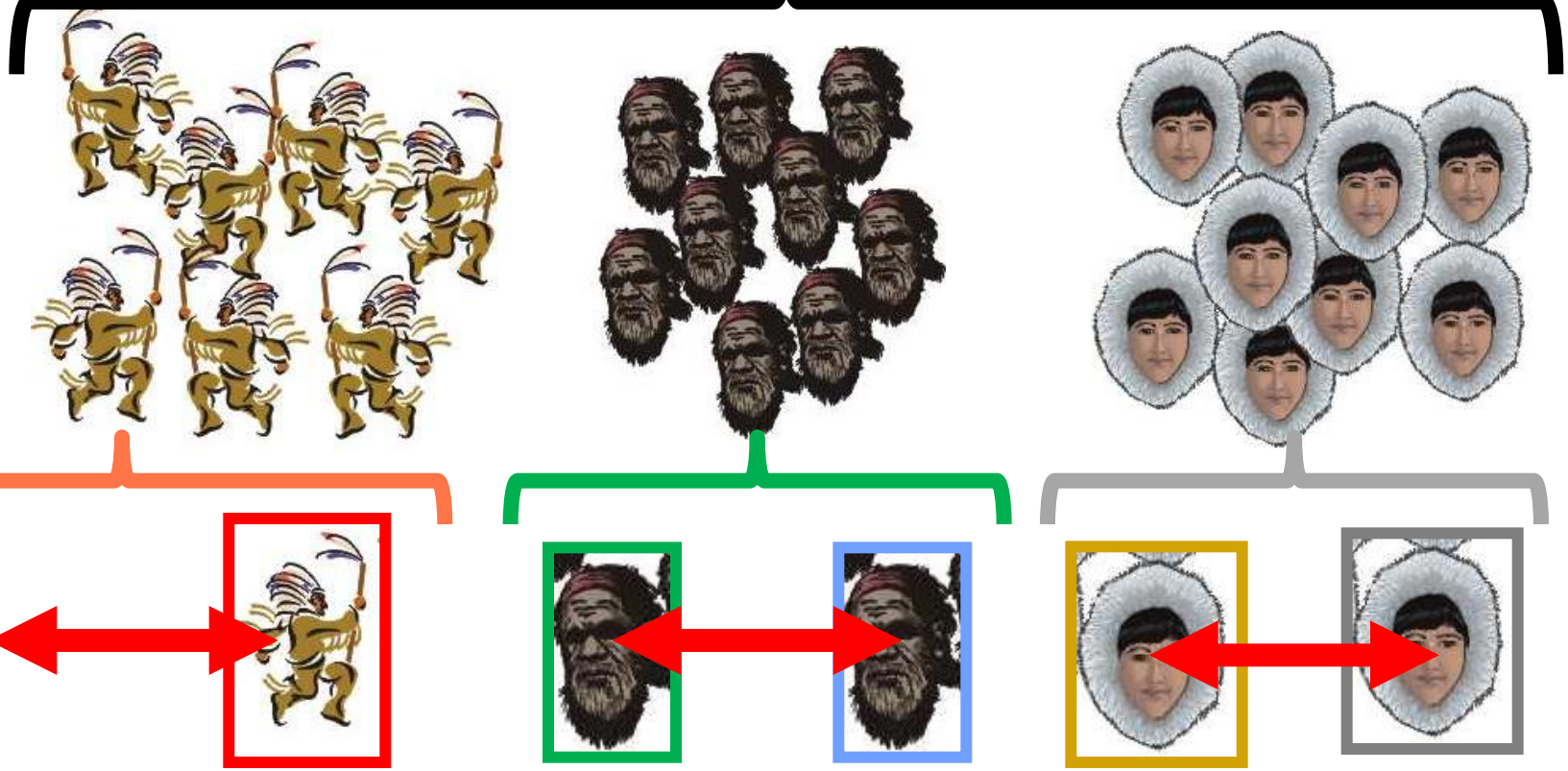
POPULATION GENETICS

SPECIES

POPULATION



SUBPOPULATIONS

II. GENETIC STRUCTURE OF POPULATION

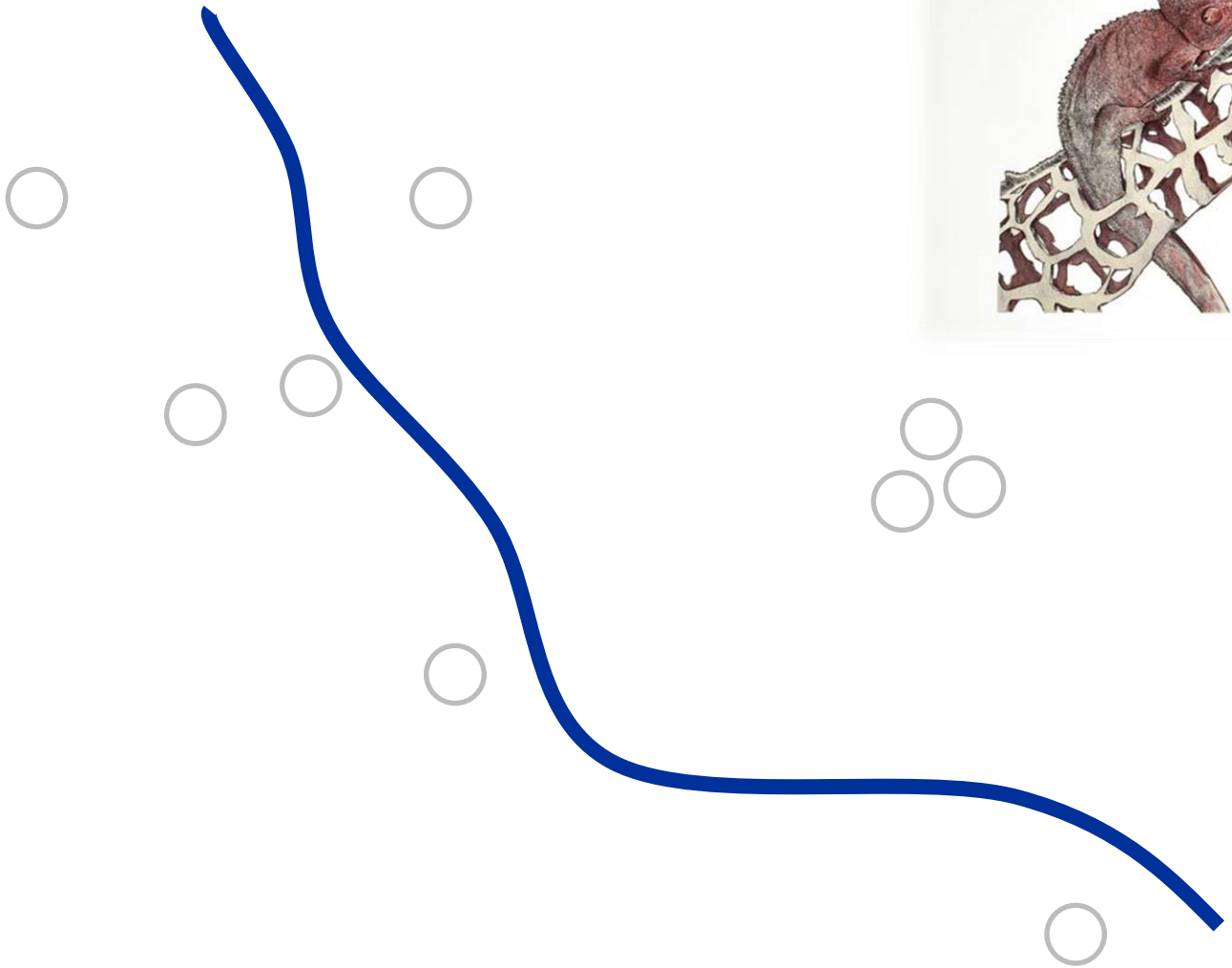
Assumption for population structure analysis:

- **neutral loci** = no effect of natural selection included

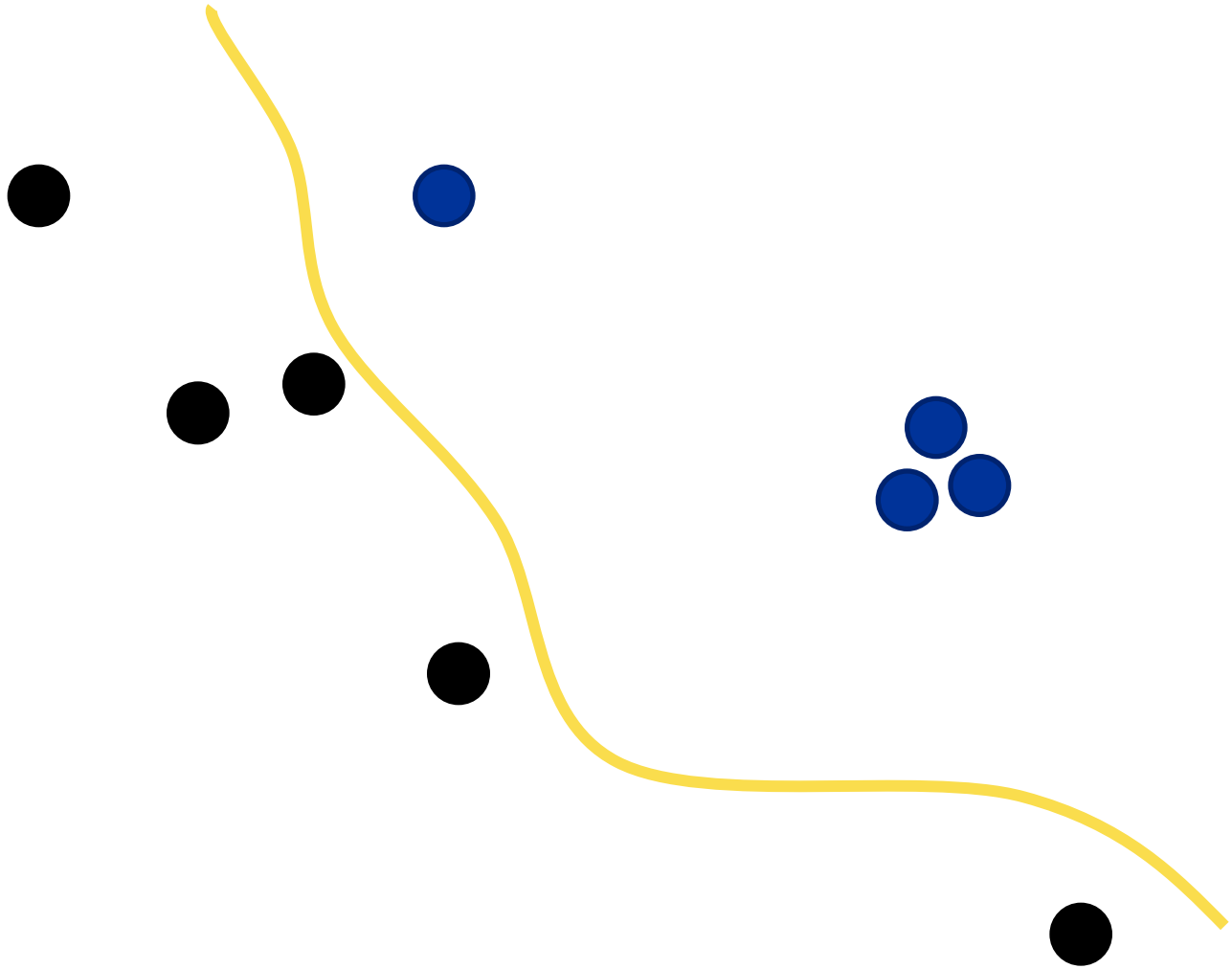
- classical population genetics approach = populations are ***a priori*** *(thought to be)* known (e.g. we want to quantify level of genetic differentiation between two localities / ?populations)

- 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 -> non-*a priori* methods**

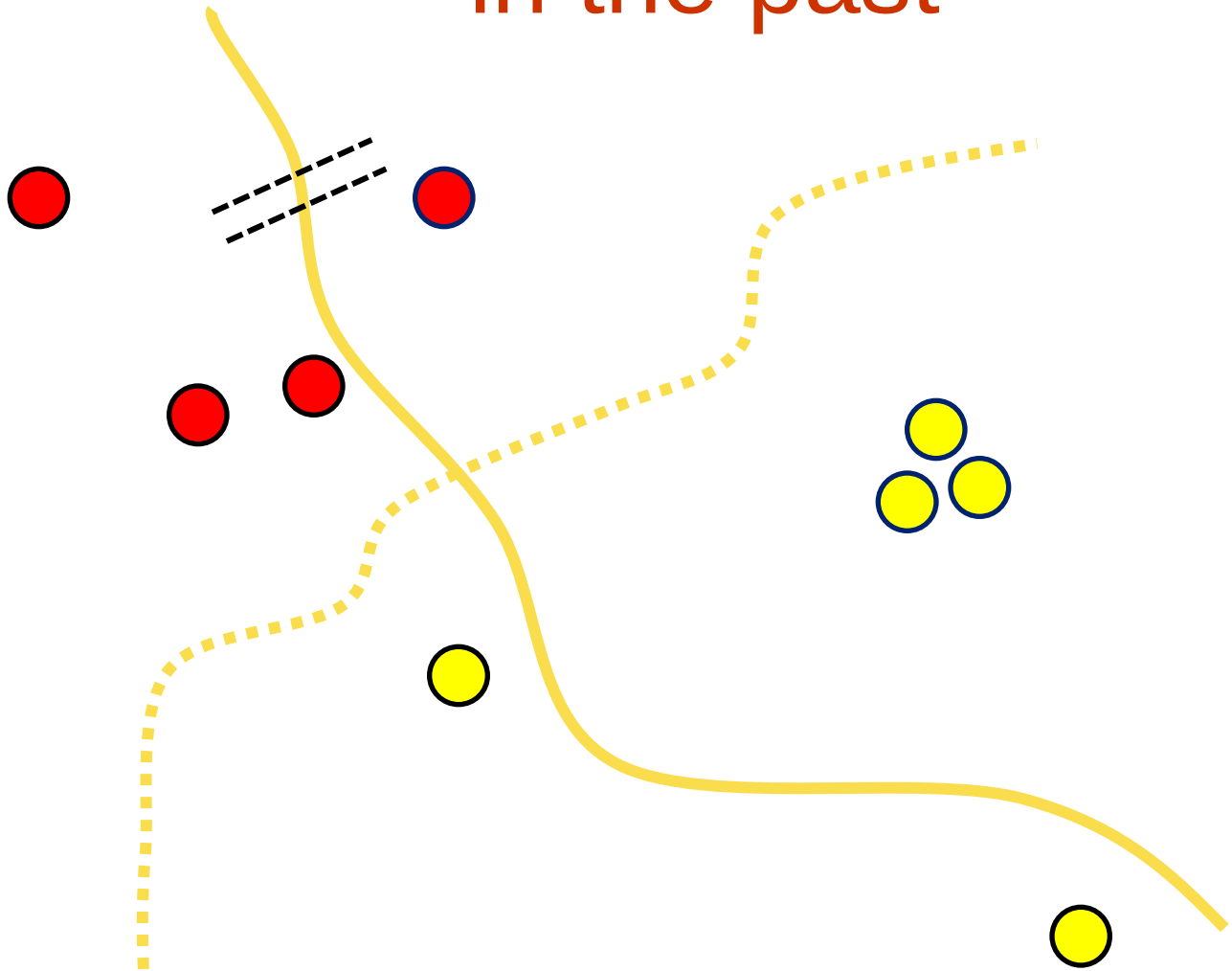
We have sampled animals in nature – Is it one or several populations???



We are interested in genetic structure of populations



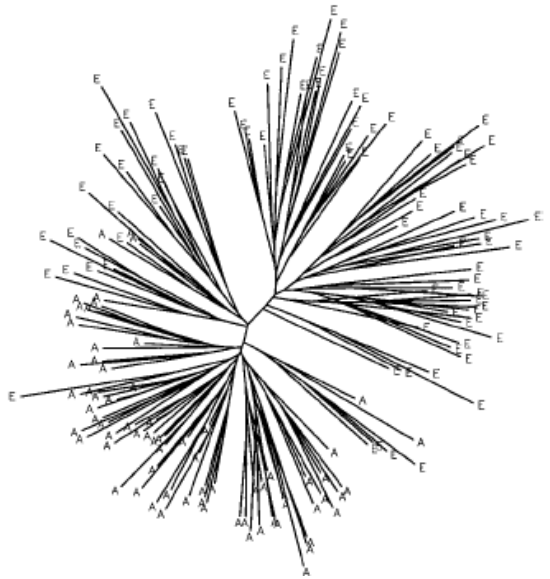
Recently observed genetic structure indicates what happened in the past



Clustering methods for „non-*a priori*“ identification of populations

DISTANCE-BASED methods

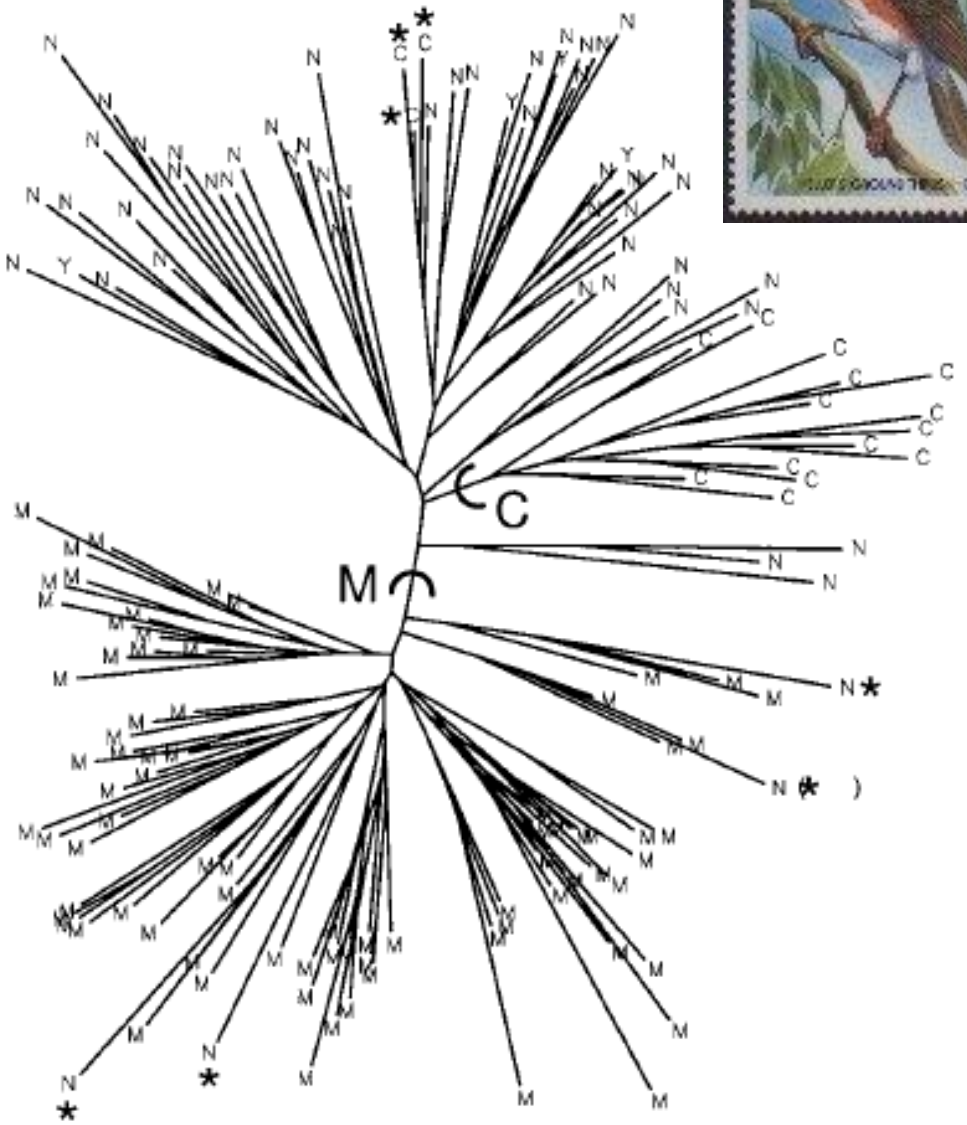
- a tree or a plot is constructed according to a pairwise distance matrix
- clusters then may be defined visually



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



Clustering method based on microsatellite distances

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

Frequency correspondence analysis

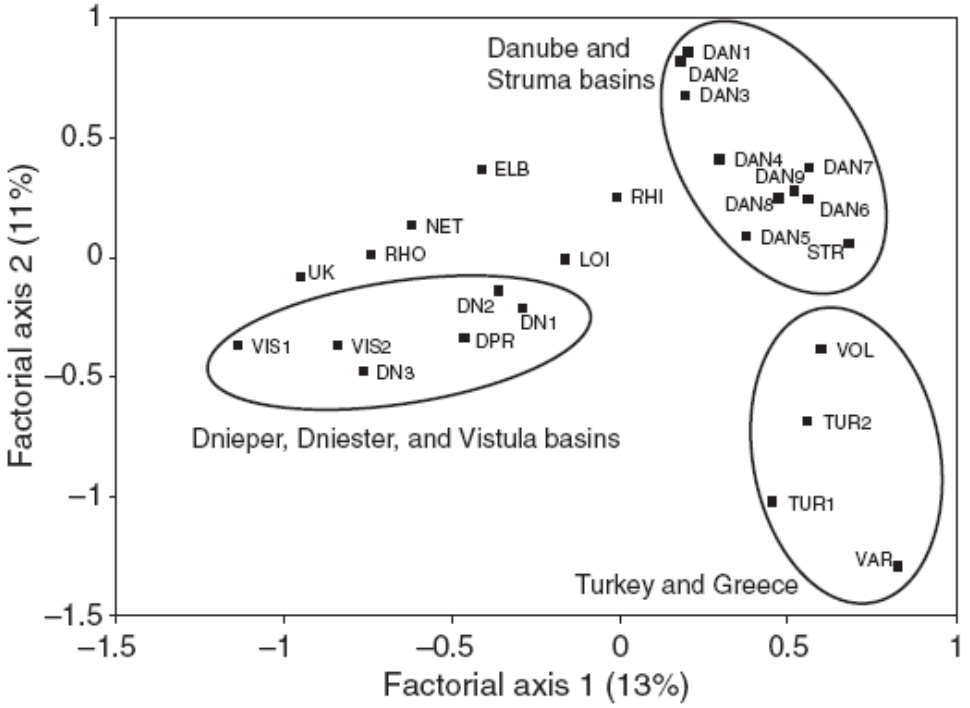


Fig. 2 A two-dimensional plot of the factorial correspondence analysis performed using GENETIX based on 12 microsatellite loci. Three geographical groups are bounded by grey lines.

PCA
Eigensoft – Patterson et al. 2006
adegenet – Jombart 2008

- each locus as one variable, reduction of number of variables
- **Genetix** – Belkhir et al. 1999 – 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)

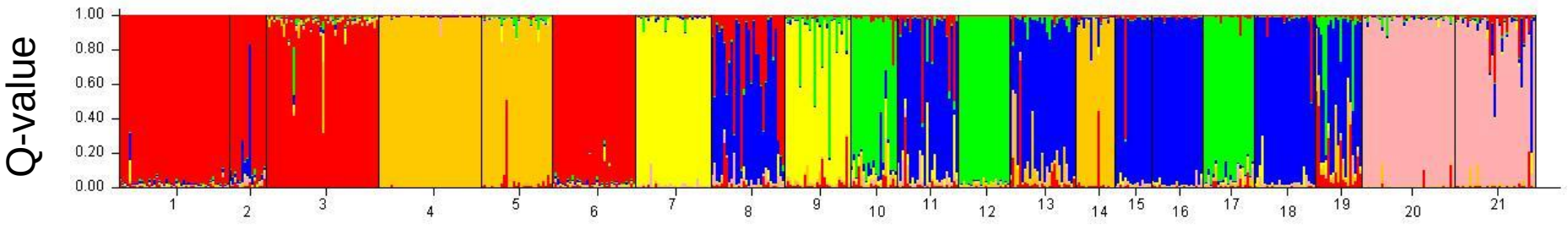
Model implemented in STRUCTURE assumes:

- **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
- i.e. the model tries to explain/correct deviation from HWE and LD by introducing the population genetic structure

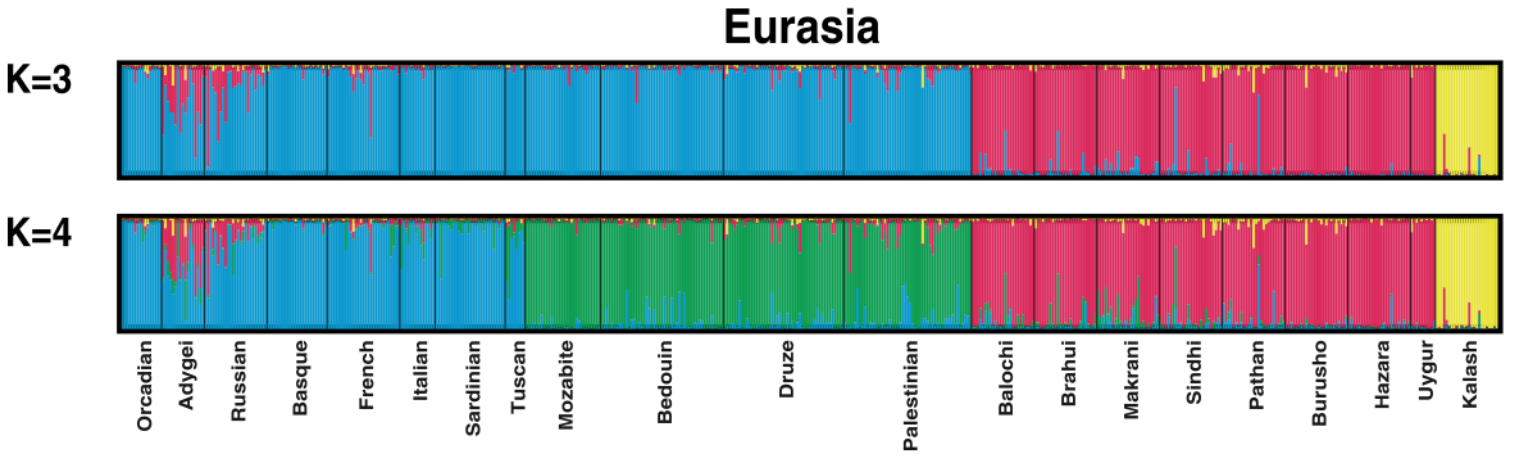
- Unknown number of populations characterised by distinct allele frequencies → **the number of populations (clusters – K) and the allele frequencies** to be estimated
- The individuals are assigned to the clusters simultaneously

Admixture model – allows assignement of an individual to several clusters

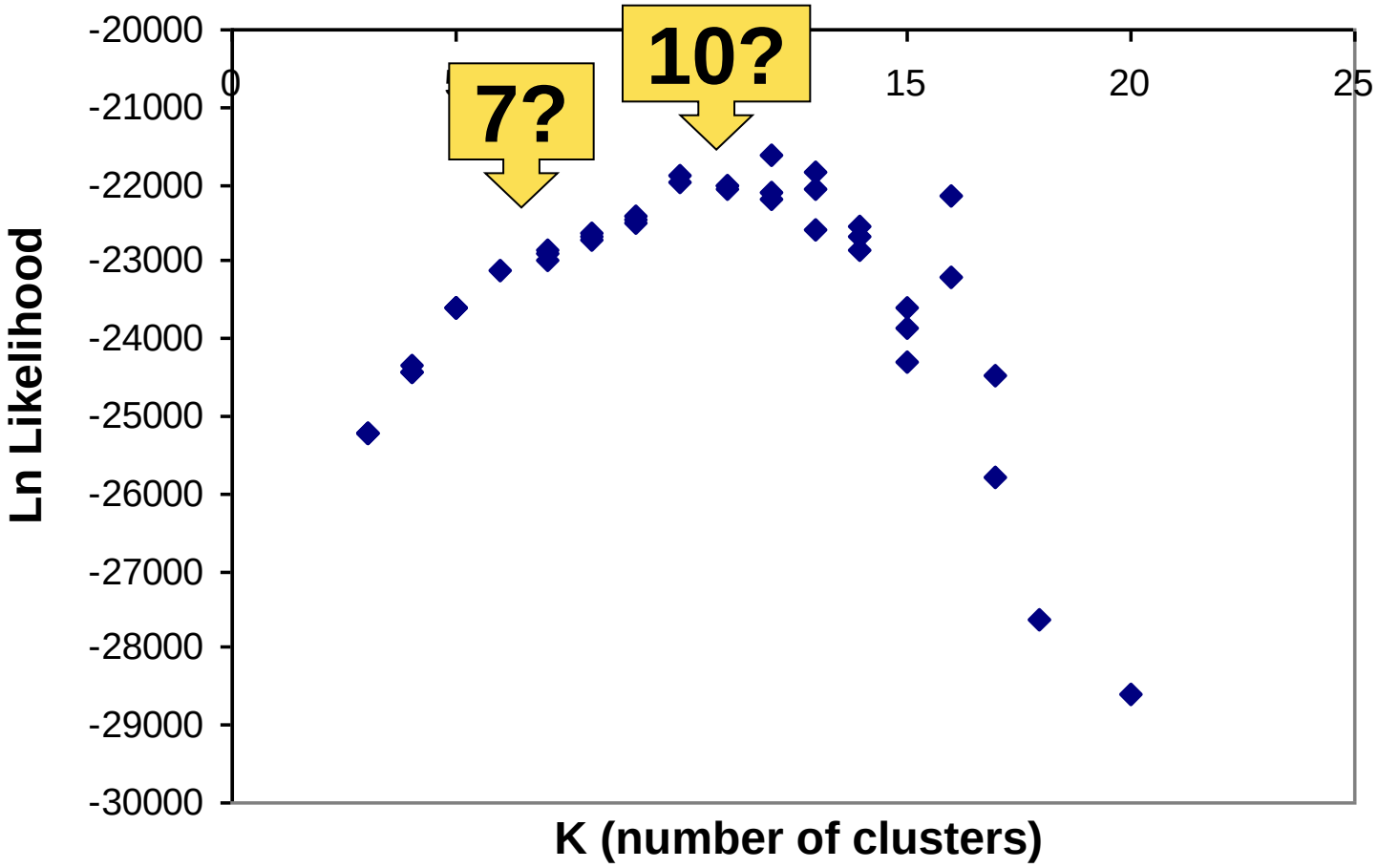
Barplot for K = 7



Genome proportion of each individual assigned to each of K clusters



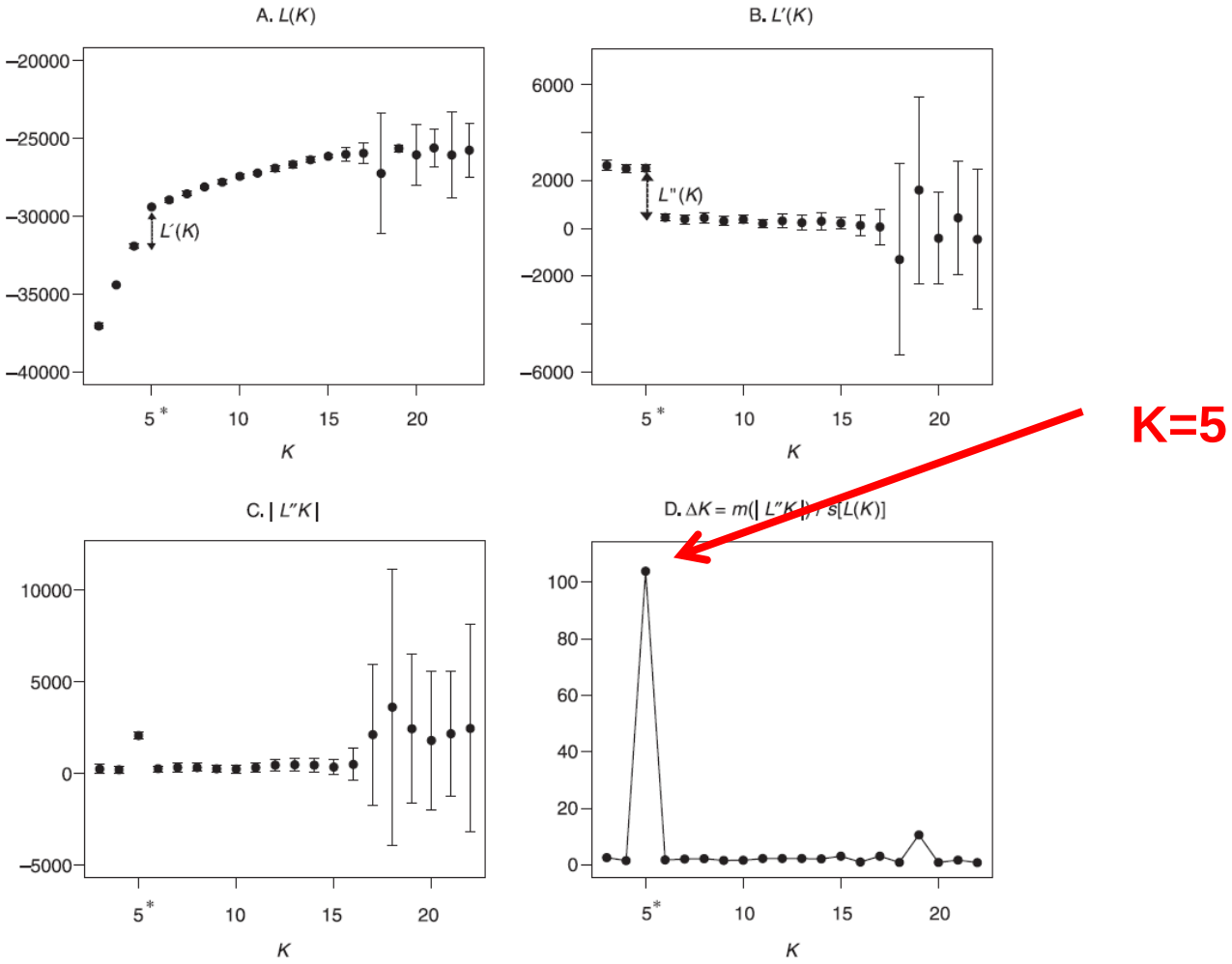
What K is the best???



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

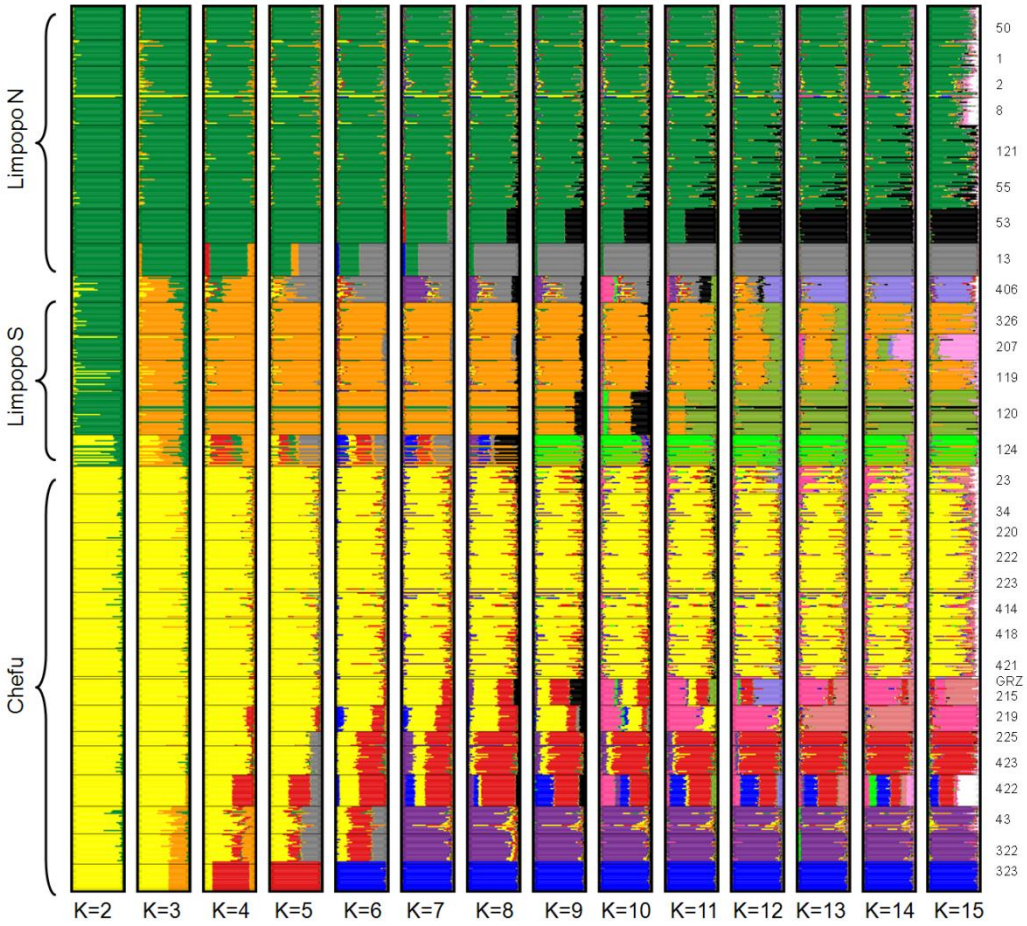


Post-processing of the STRUCTURE outputs

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

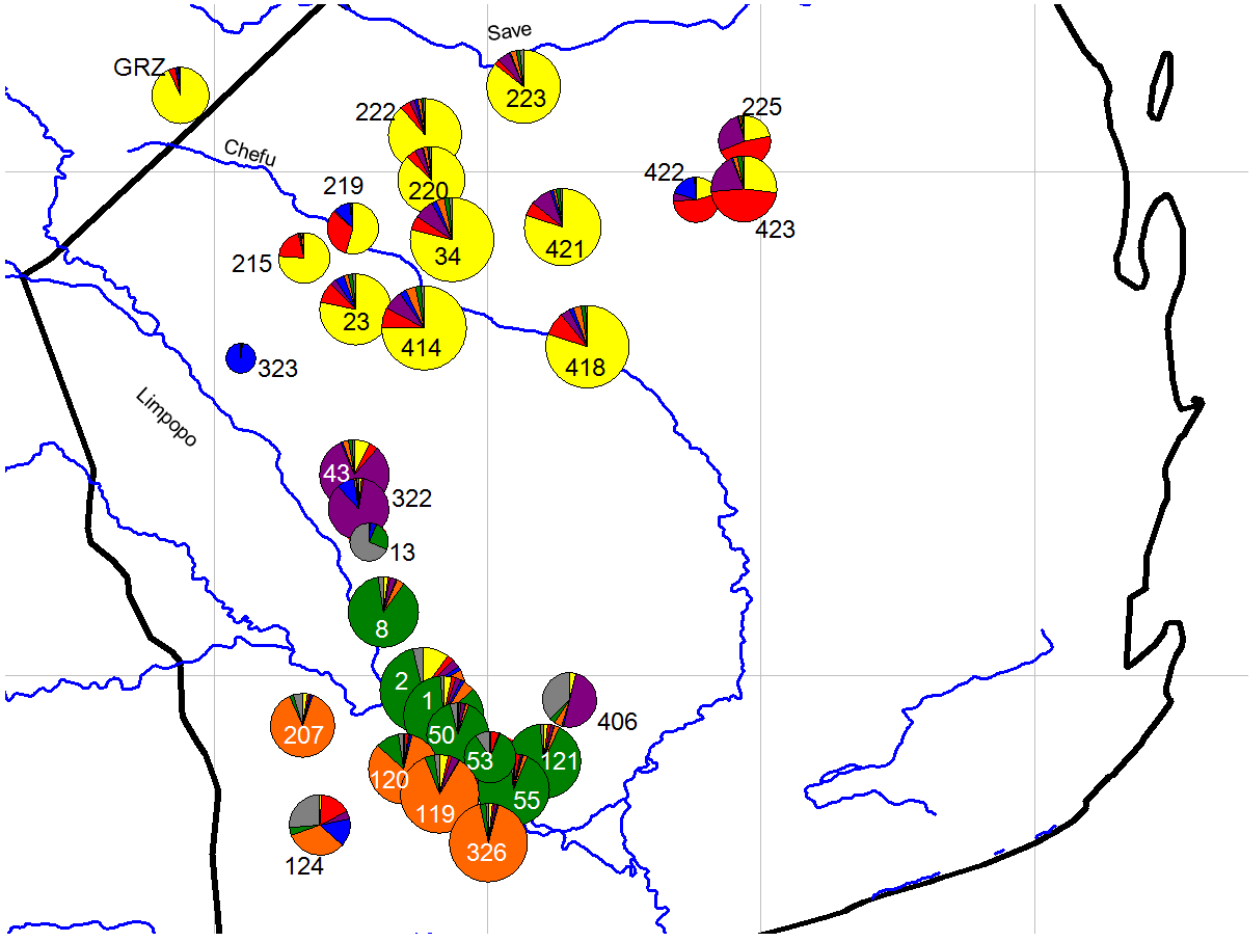


Picture of **hierarchical structure between clusters**

Bartáková et al. 2013

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

K = 7

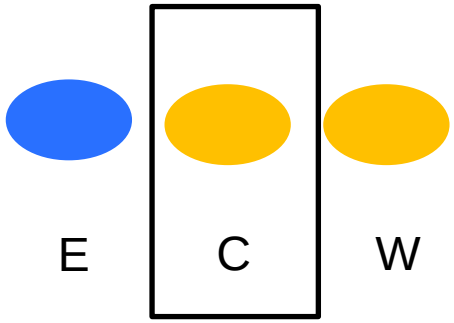


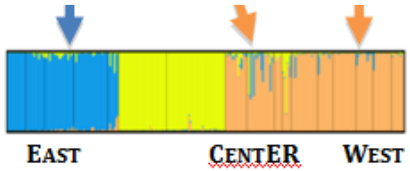
Bartáková et al. 2013

! introgression vs. ancestral polymorphism

STRUCTURE

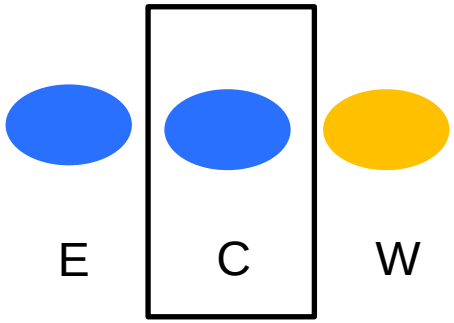
K=3

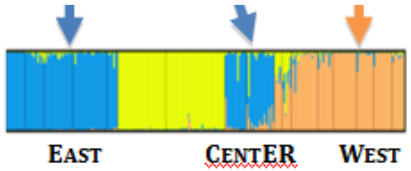




First solution for K=3 (46% of the runs)

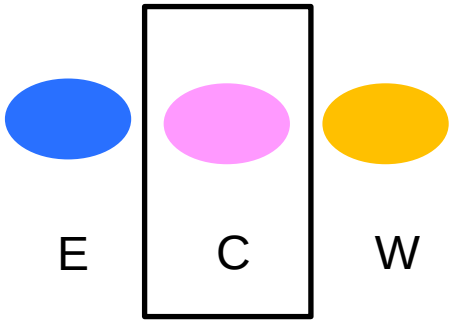
K=3

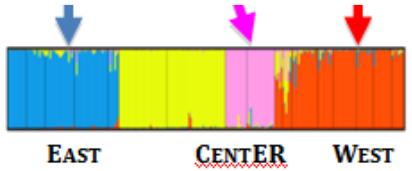




Second solution for K=3 (32% of the runs)

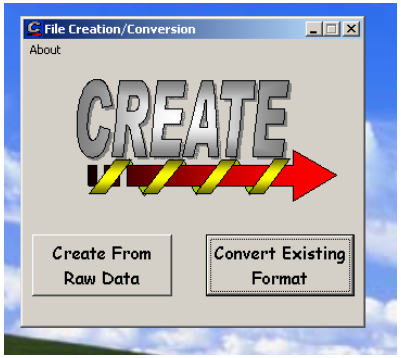
K=4





Solution for K=4

A whole bunch of population genetics software (with specific input data formats)



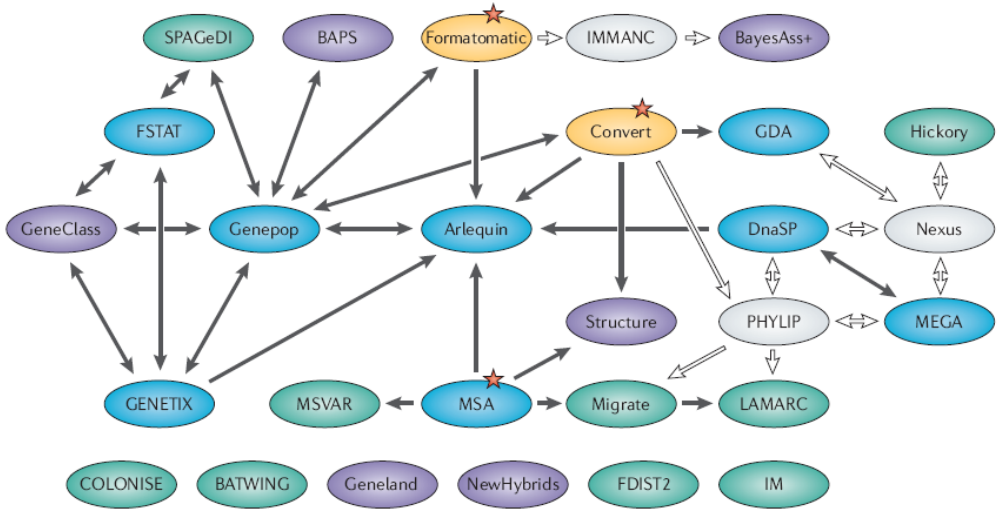
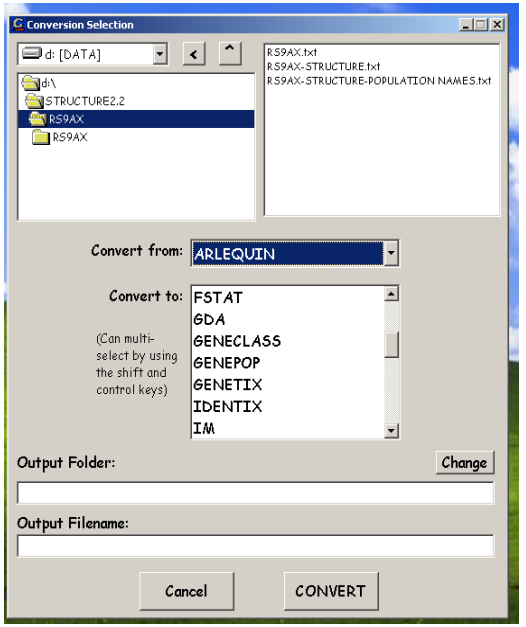


Figure 1 | **Flow chart of possible data exchange between different population genetics programs.** Although many programs have their own input-file specification, data files can still be exchanged between most programs (black arrows), avoiding tedious reformatting processes. The red stars are recommended starting points to format an initial data set. Blue ellipses represent multi-purpose packages, whereas individual-centred programs are shown in violet. The two conversion programs are shown in yellow. Specialized programs are shown in green, and light grey ellipses represent programs that are not reviewed here, but the data formats of which are used by other programs allowing indirect data exchange (white arrows). The data files associated with the programs listed on the bottom row cannot be exchanged directly with the other programs.



CREATE is software for the creation of new and conversion of existing data input files for 64 genetic data analysis software programs

Computer programs for population genetics data analysis: a survival guide

Laurent Excoffier and Gerald Heckel

Acoording to purpose of our population genetic analysis

Table 5 | **List of computer programs suited for a given analysis and genetic marker**

	Multi-allelic markers*	STR	Dominant markers (AFLP)	SNP	DNA sequences
Descriptive statistics	Arlequin, FSTAT, GDA, Genepop, GENETIX, MSA, SPAGeDi, Hickory		SPAGeDi		Arlequin, DnaSP, MEGA
Linkage disequilibrium	Arlequin, FSTAT, GDA, Genepop, GENETIX, Structure				
Analysis of population subdivision	Arlequin, FSTAT, GDA, Genepop, GENETIX, MSA, SPAGeDi, Hickory, Structure, BAPS, Geneland	Arlequin, FSTAT, GDA, Genepop, MSA, SPAGeDi	Hickory		Arlequin, DnaSP, MEGA
Detection of new immigrants: known populations	BayesAss+, GeneClass				
Detection of new immigrants: inferred populations	BAPS, NewHybrids, Structure, Geneland	BATWING, IM, LAMARC, MSVAR			
Demographic expansion or decline		BATWING, IM, LAMARC, Migrate, MSVAR		BATWING, LAMARC, Migrate	Arlequin, DnaSP, IM, LAMARC, Migrate
Population size	Migrate	BATWING, IM		BATWING, LAMARC, Migrate	IM, LAMARC, Migrate
Divergence time	Arlequin, FSTAT, GDA, Genepop, GENETIX	BATWING, IM, LAMARC, Migrate, MSVAR		BATWING, LAMARC, Migrate	DnaSP, IM, LAMARC, Migrate
Migration rates	Arlequin, FSTAT, Genepop, BayesAss+, COLONISE, Migrate	BATWING, IM, LAMARC, Migrate, MSVAR		BATWING, LAMARC, Migrate	DnaSP, IM, LAMARC, Migrate
Neutrality tests	Arlequin, FDIST2				Arlequin, DnaSP, MEGA
Spatially explicit analyses	SPAGeDi, Geneland, COLONISE				

*By multi-allelic markers, we mean loci for which no specific mutation model is assumed, or for which mutations can be neglected. In the latter case, computations are based on allele frequencies only. Otherwise, specific mutation models are assumed. [illegible] DNA sequence, STR and SNP allele frequencies, as well as nucleotide [illegible] packages to estimate descriptive statistics and linkage disequilibrium, and to detect new immigrants. AFLP, amplified fragment length polymorphism; STR, short tandem repeat.

allele frequency

allele frequency + mutation model

Software for analysis of intra-population genetic variation (genetic diversity)

- Conversion of input data formats:
  - GenAlEx (http://biology-assets.anu.edu.au/GenAlEx/Download.html)
  - CREATE (https://bcrc.bio.umass.edu/pedigreesoftware/node/2)

- GenAlEx – Ho, He, HWE
- Genepop – LD, HWE
- FSTAT – allelic richness

popgen softwares:
https://courses.washington.edu/popgen/Software.htm

GenAlEx

Genetic Analysis in Excel

6.5

Professor Rod Peakall
Evolution, Ecology and Genetics
Research School of Biology
The Australian National University, Canberra ACT 0200, Australia.

Professor Peter Smouse
Department of Ecology, Evolution and Natural Resources
School of Environmental and Biological Sciences
Rutgers University, New Brunswick NJ 08901-8551, USA.

Peakall R. and Smouse P.E. (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. Bioinformatics 28, 2537-2539. Peakall R. and Smouse P.E. (2006) GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol. Ecol. Notes 6, 288-295.

Australian National University

Proudly supported by The Australian National University

http://biology.anu.edu.au/GenAlEx/

Logo Design by GreenIdeasCreative.com

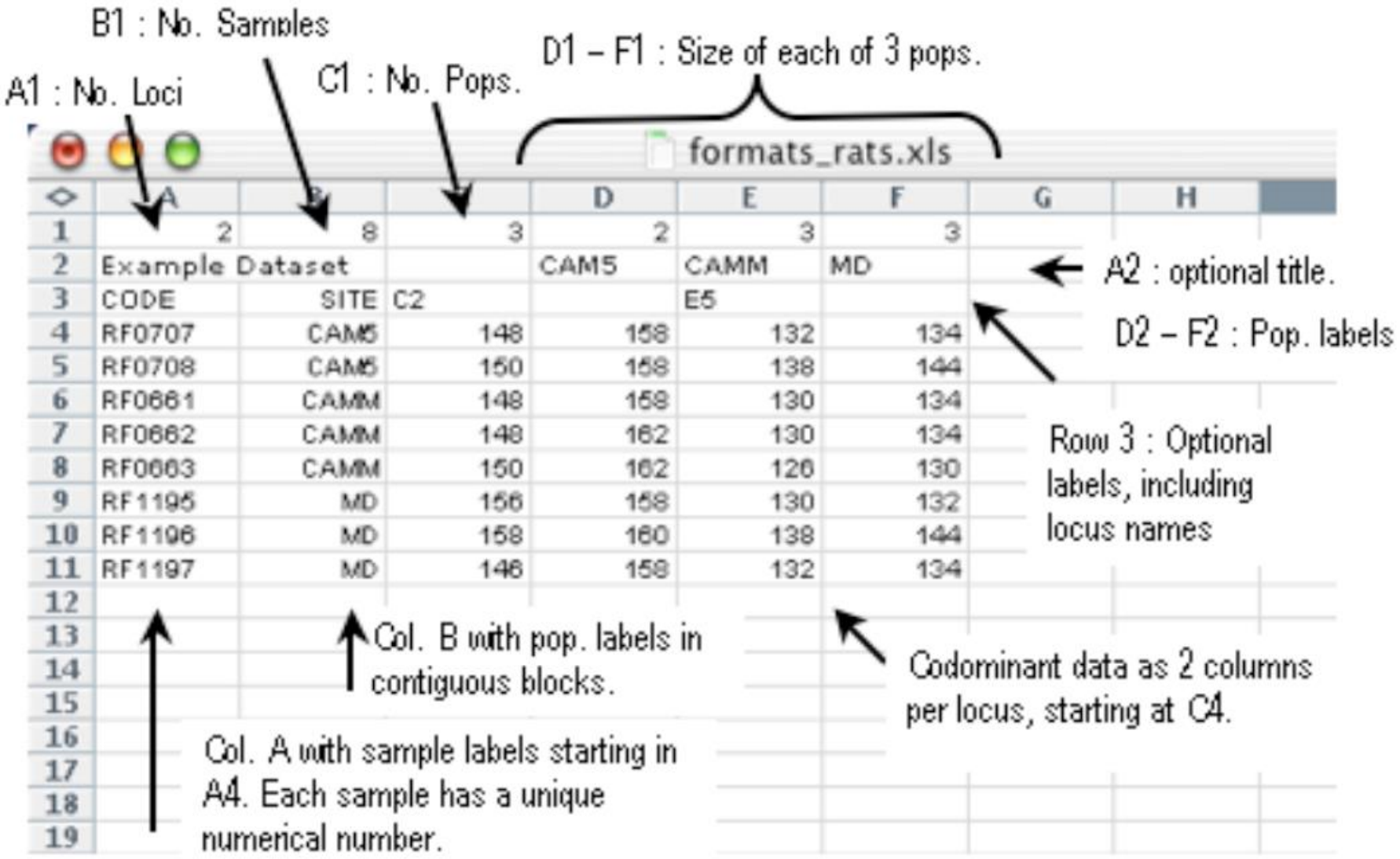
GenAlEx - Genetic Analysis in Excel (Peakall and Smouse 2006, 2012) is designed as a user-friendly package with an intuitive and consistent interface that allows users to analyse a wide range of population genetic data within a software environment with which most users will have some familiarity (MS Excel).

Example of codominant microsatellite data, with genotypes by fragment size.

Formats.xls

	A	B	C	D	E	F	G
1	2	8	2	4	4		
2	Codominant data - fragment size			EC	TT		
3	Sample no.	Pop	CA2		GA8		
4	HE001	EC	294	298	274	274	
5	HE002	EC	292	300	256	258	
6	HE003	EC	296	298	258	258	
7	HE004	EC	298	300	258	258	
8	HE010	TT	298	298	256	256	
9	HE011	TT	292	296	256	260	
10	HE012	TT	296	296	254	256	
11	HE013	TT	292	296	214	248	
12							

Diploid codominant markers (microsatellites)



	A	B	C	D	E	F	G	H
1	2	8	3	2	3	3		
2	Example Dataset			CAM5	CAMM	MD		
3	CODE	SITE	C2		E5			
4	RF0707	CAM5	148	158	132	134		
5	RF0708	CAM5	150	158	138	144		
6	RF0661	CAMM	148	158	130	134		
7	RF0662	CAMM	148	162	130	134		
8	RF0663	CAMM	150	162	126	130		
9	RF1195	MD	156	158	130	132		
10	RF1196	MD	158	160	138	144		
11	RF1197	MD	146	158	132	134		
12								
13								
14								
15								
16								
17								
18								
19								