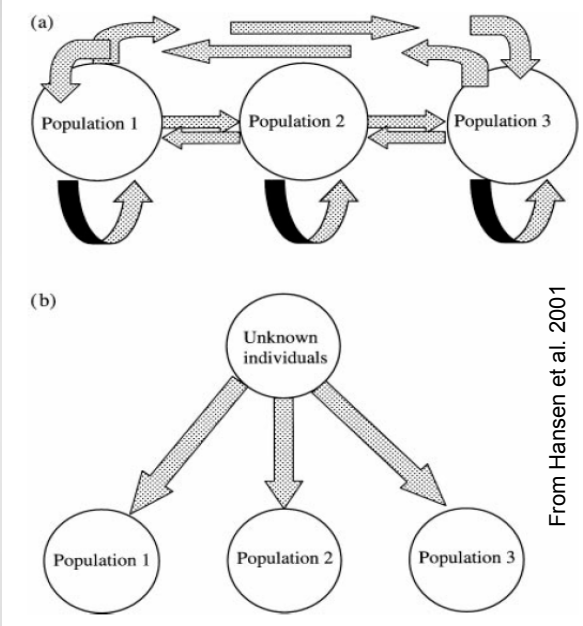
Assignment and clustering algorithms for individual multilocus genotypes



Mathieu Gautier, Raphael Leblois
Centre de Biologie pour la Gestion des Populations , CBGP
INRA, Montpellier

Master B2E, December 2012

Ex1 : Provenance de défenses d'éléphants saisies?

Conservation Biology, Volume 22, No. 4, 1065–1071

DOI: 10.1111/j.1523-1739.2008.01012.x

Contributed Paper

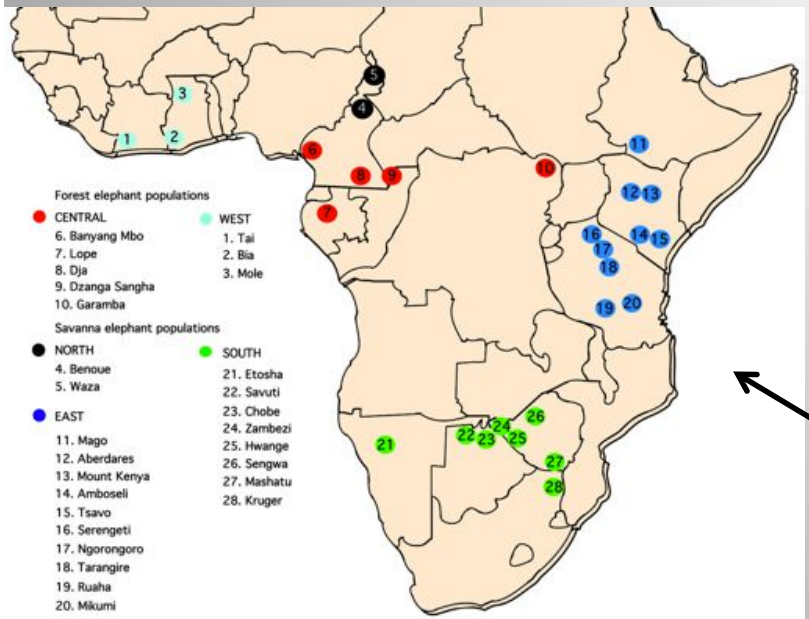
Combating the Illegal Trade in African Elephant Ivory with DNA Forensics

SAMUEL K. WASSER,*‡‡ WILLIAM JOSEPH CLARK,† OFIR DRORI,‡ EMILY STEPHEN KISAMO,§ CELIA MAILAND,* BENEZETH MUTAYOBA,** AND MATTHEW STEPHENS††

*Center for Conservation Biology, Department of Biology, University of Washington, Seattle, WA 98195-1800, U.S.A.
†Interpol Working Group on Wildlife Crime and Department of Law Enforcement, Israel Nature and Parks Authority, Jerusalem 95463, Israel
‡Last Great Ape Organization, Vallee Nlongkak, Younde, Cameroon
§Lusaka Agreement Task Force, P.O. Box 3533, Nairobi, Kenya
**Faculty of Veterinary Medicine, Sokoine University of Agriculture, Morogoro, Tanzania
††Department of Human Genetics and Statistics, University of Chicago, Chicago, IL 60637, U.S.A.

Ex1 : Provenance de défenses d'éléphants saisies?

(a) Map of Africa showing the collection sites divided into five regions: West Africa (cyan), Central forest (red), and Central (black), South (green) and East (blue) savanna. (b) Estimated locations of elephant tissue and faecal samples



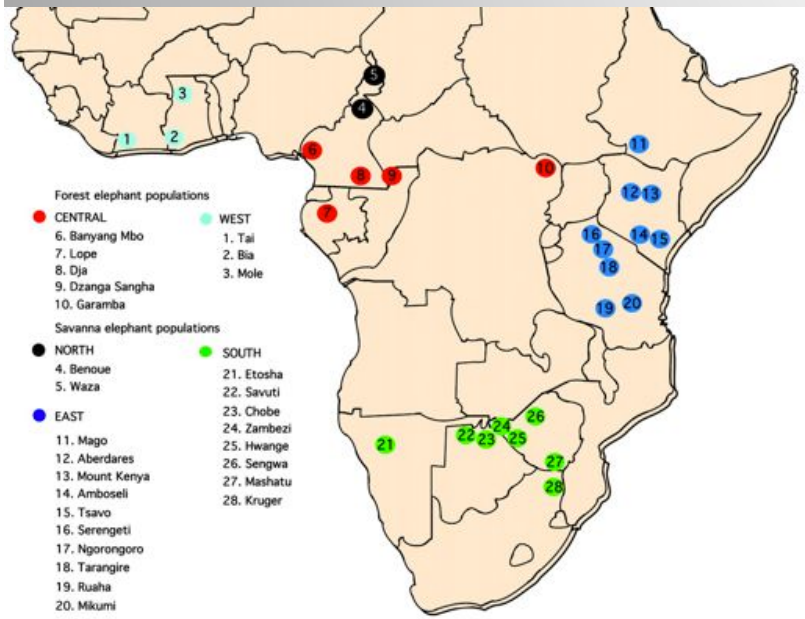
37 tusks from a large seizure in Singapore

Where are they coming from?

Known populations, genetically characterized

Ex1 : D'ou viennent ces défenses d'éléphants?

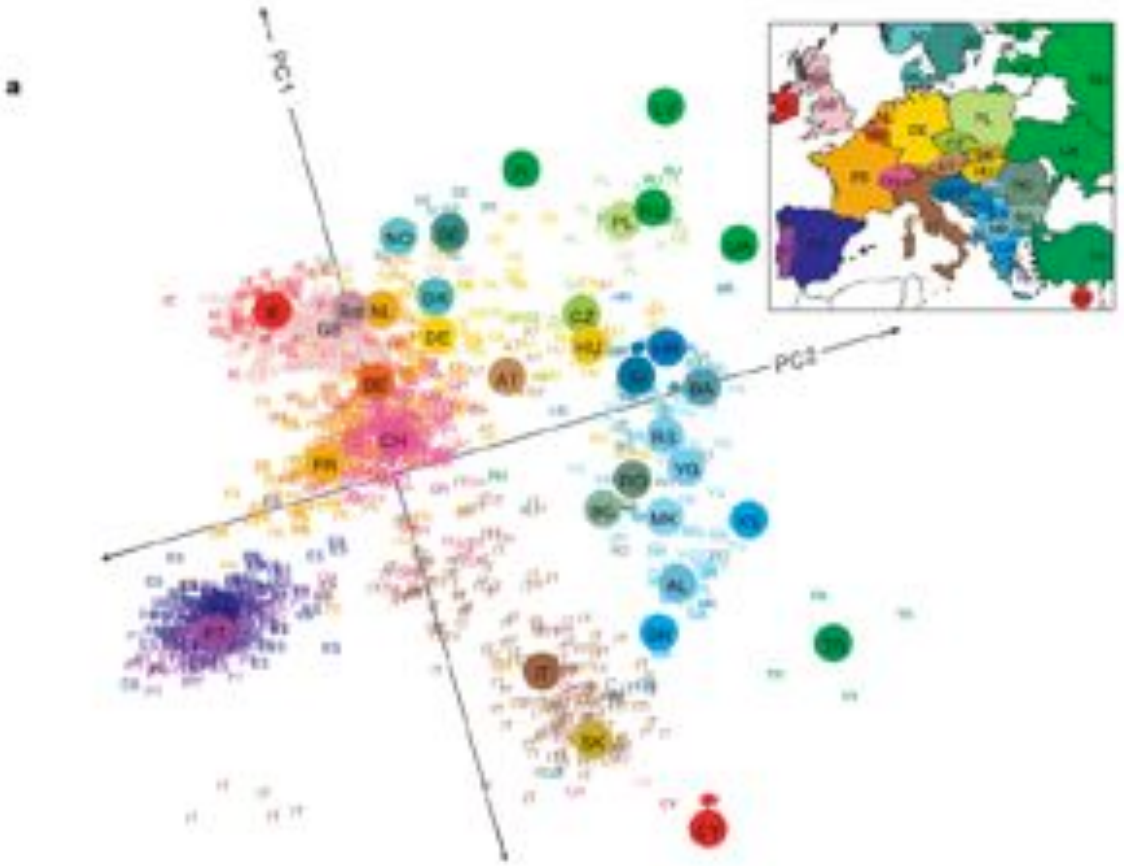
Estimated locations of the tusks :

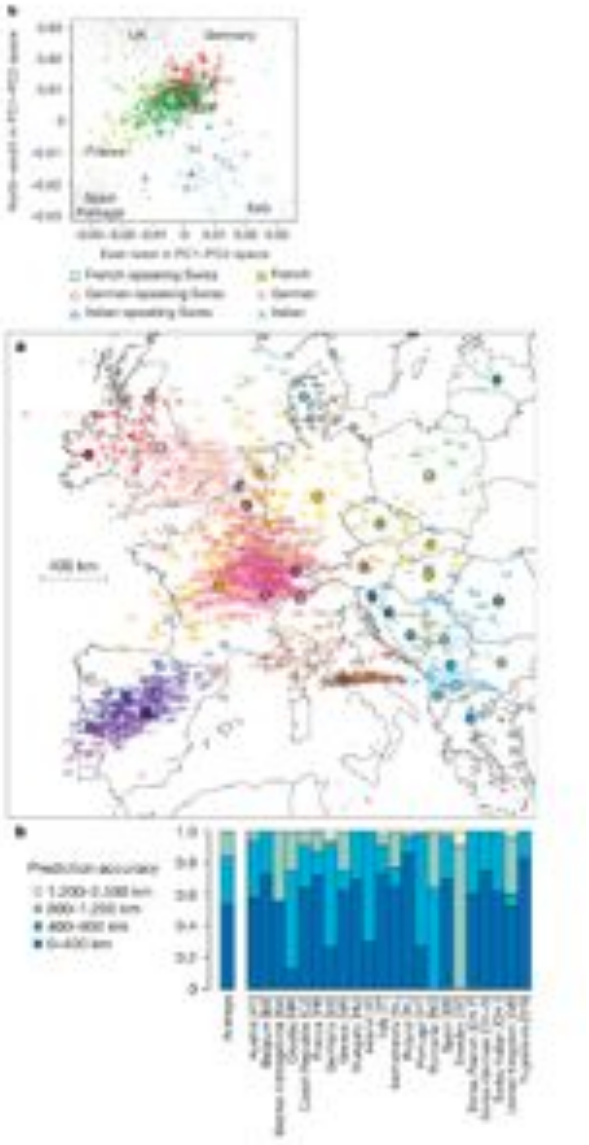


Biological questions : genetic diversity description PCA

LETTERS

Genes mirror geography within Europe

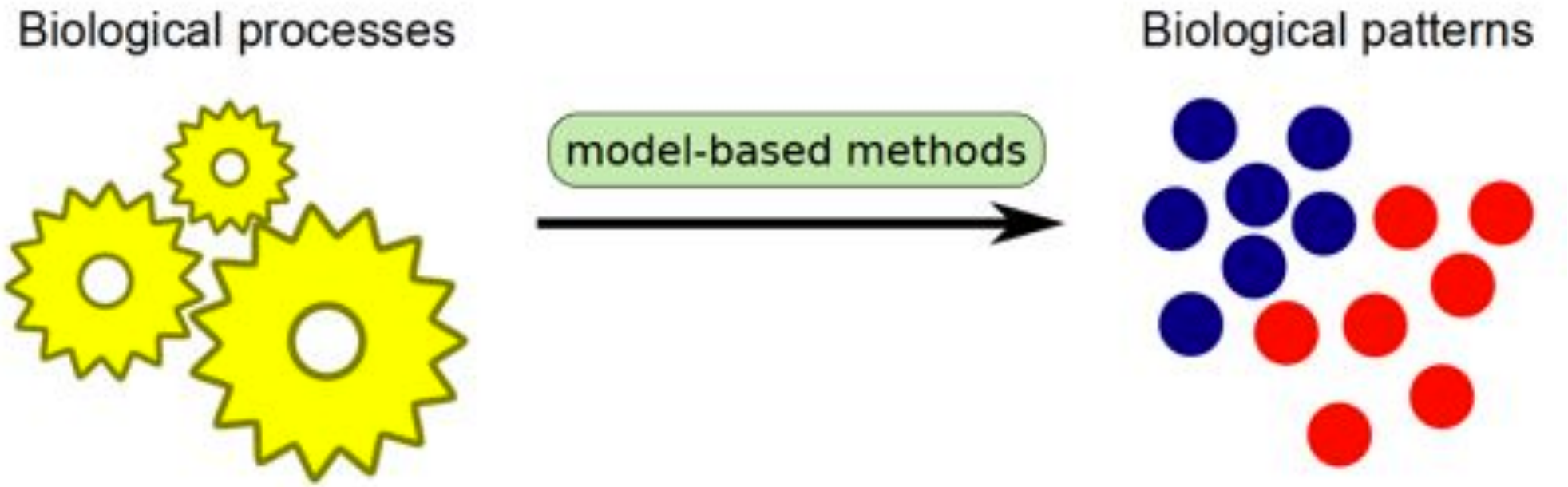




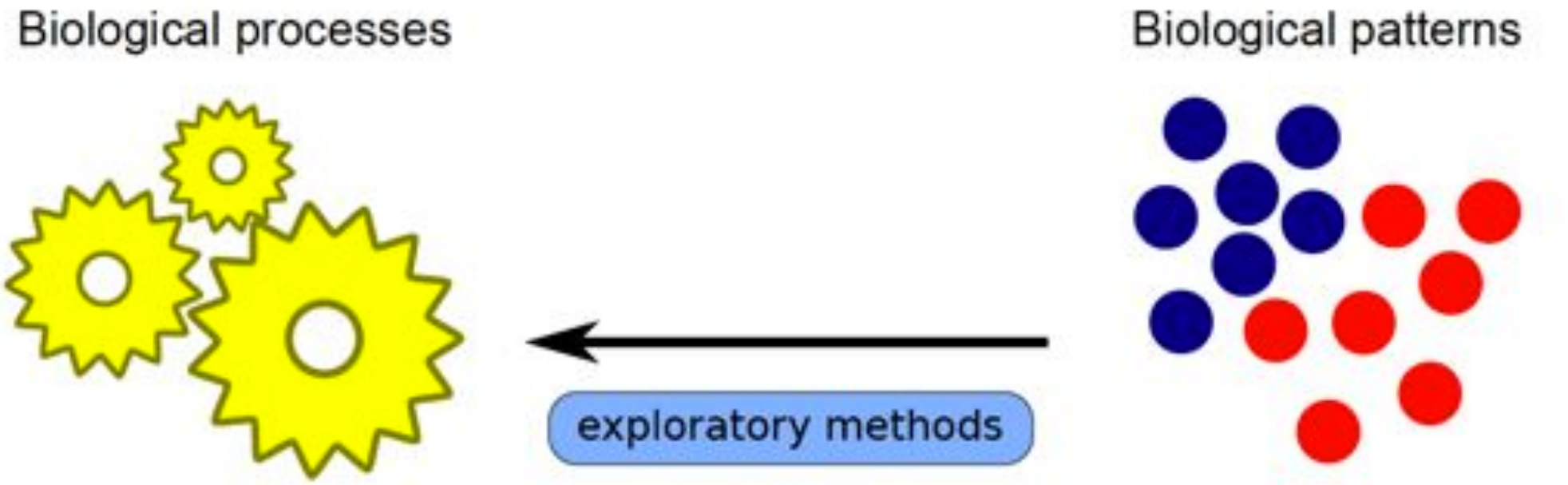
Biological questions

- Geographic origin of some individuals with unknown origin
- Population delimitation, spatial or not
- Migrant detection / inference of recent migration rates
- Analysis of genetic introgression / hybridization

Model-based vs. exploratory approaches

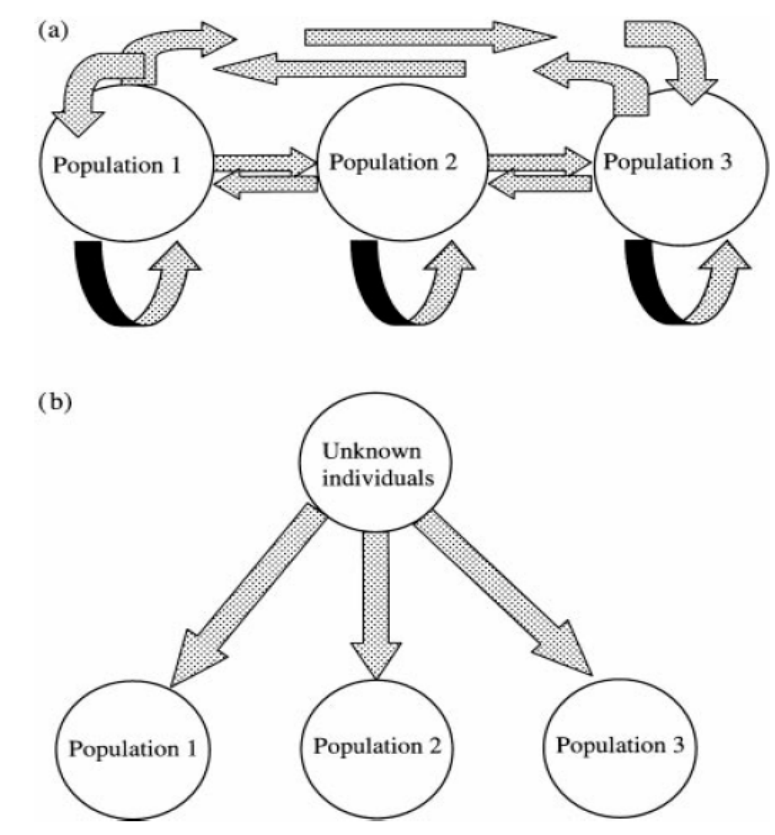


Model-based vs. exploratory approaches



Classification vs. Clustering (model-based approaches)

What is a priori known about sampled population and individuals ?



Assignment: some focal individuals, of unknown origin, are assigned to a priori defined populations or groups

Software : GENECLASS2

Clustering : unknown a priori populations or groups, clusters are build from the genetic data

Software : STRUCTURE, GENELAND, …

Assignment principle

Definition : Assign individuals of unknown origin to a priori known populations (i.e. genetically characterized), using their multilocus genotypes

Main assumptions :

1- known populations and large genetic samples from each pop

2- In each population : - Hardy-Weinberg equilibrium
- linkage equilibrium

Ex : Paetkau et al. 1995, Rannala & Mountain 1997

First algorithm : Paetkau et al. 1995

Hardy Weinberg + linkage equilibrium ➠ allows likelihood computation using the probability that a given multilocus genotype came from a given population

For a single locus, the likelihood *L* of a genotype occurrence in a population is proportional to its expected genotype frequencies under HW given the allelic frequencies in the population :

p_{ijk} : frequency of allele *k* at locus *j* in pop *i*

$$L \approx 2 * p_{ijk} * p_{ijk'} \quad \textbf{if heterozygote } kk'$$

$$\text{or } L \approx p_{ijk}^2 \quad \textbf{if homozygote } kk$$

Independent loci ➠ the multilocus likelihood is the product of the likelihood at each locus

First algorithm : Paetkau et al. 1995

3 steps of the algorithm:

1- Computation of allelic frequencies in each population

2- Computation of the likelihood of the membership of each focal individual to each population

3- Assignment of the focal individuals to the population for which they have the highest likelihood of membership (Maximum likelihood)

Supplementary assumption : allelic frequencies inferred from the genotypes sampled in each population are close to the true values

First algorithm : Paetkau et al. 1995

Supplementary assumption : allelic frequencies inferred from the genotypes sampled in each population are close to the true values

Potential problem:

one allele, present in the genotype of a focal individual, is not present in a population ➠ null likelihood because $p_{ijk}=0$

However this allele may be rare and may not have been sampled just by chance (small sample bias)

2 ad-hoc solutions :

- Always put a low frequency to potentially unsampled alleles (arbitrary or 1/(gene sample size))
- Always add the focal individual genotype to each population for population allelic frequency computations

Paetkau et al. 1995

Example : Brown Bears population structure

Paetkau et al. 1998, Molecular Ecology

Fig. 1 Study areas (black). Fifteen individual samples were obtained from southeast coastal Alaska (l–z). Glaciers and icefields are shown in grey. According to Kurtén (1973) the Kuskokwim, Alaska Range, and Kluane samples are *Ursus arctos horribilis* whereas the ABC, Izembek and southeast coastal areas fall within the range of *U.a. dalli*.

Paetkau et al. 1995

Example : Brown Bears population structure

Paetkau et al. 1998, Molecular Ecology

Study area	2N	H_O	H_E
Admiralty	60	0.646	0.628
Baranof	18	0.493	0.496
Chichagof	52		
Kluane	100	0.788‡	0.761‡
Alaska Rge.	56	0.759	0.779
Kuskokwim	110	0.700‡	0.682‡
Izembek	28	0.536	0.532
Kodiak	68	0.298‡	0.265‡
Coast (l–z)	30	0.617	0.757

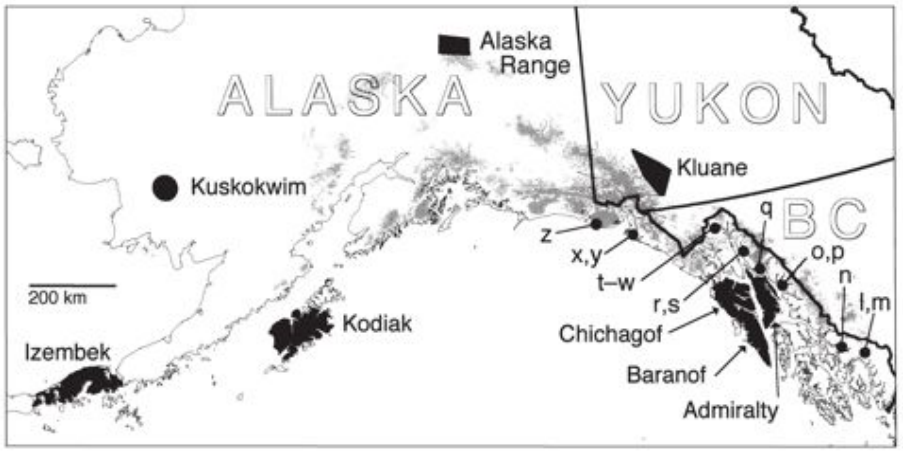


Fig. 1 Study areas (black). Fifteen individual samples were obtained from southeast coastal Alaska (l–z). Glaciers and icefields are shown in grey. According to Kurtén (1973) the Kuskokwim, Alaska Range, and Kluane samples are *Ursus arctos horribilis* whereas the ABC, Izembek and southeast coastal areas fall within the range of *U.a. dalli*.

Paetkau et al. 1995

Example : Brown Bears population structure

Paetkau et al. 1998, Molecular Ecology

assignment results :

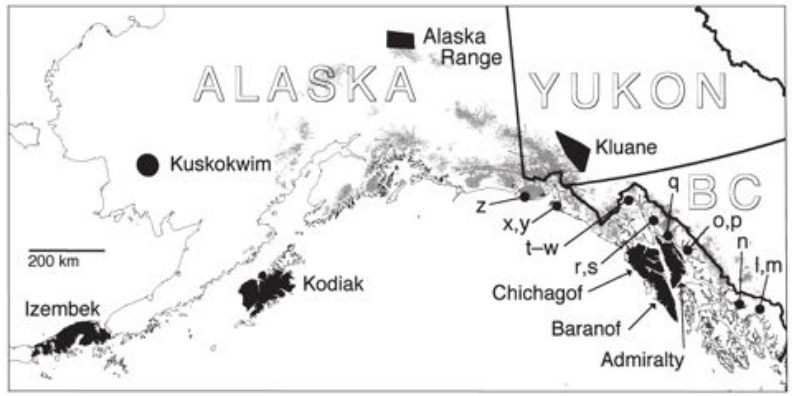


Fig. 1 Study areas (black). Fifteen individual samples were obtained from southeast coastal Alaska (l–z). Glaciers and icefields are shown in grey. According to Kurtén (1973) the Kuskokwim, Alaska Range, and Kluane samples are *Ursus arctos horribilis* whereas the ABC, Izembek and southeast coastal areas fall within the range of *U.a. dalli*.

Source population / *N*	Population to which individuals were assigned						
	Adm.	B-C	Klu.	Ala.	Kus.	Ize.	Kod.
Admiralty / 30	**29**	1					
B-C / 35		**35**					
Kluane / 50		1	**45**	4			
Alaska Rge. / 28	1		1	**24**	1	1	
Kuskokwim / 55			1	1	**47**	6	
Izembek / 14					2	**12**	
Kodiak / 34							**34**
Coast (l–z) / 15	1	2	7	5			

Paetkau et al. 1995

Example : Brown Bears population structure

Paetkau et al. 1998, Molecular Ecology

1 – migration between adjacent population

2 – >7km-wide rivers act as strong barriers

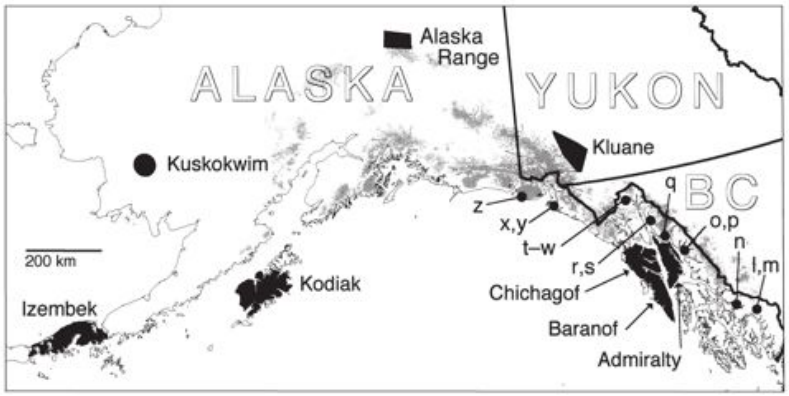


Fig. 1 Study areas (black). Fifteen individual samples were obtained from southeast coastal Alaska (l–z). Glaciers and icefields are shown in grey. According to Kurtén (1973) the Kuskokwim, Alaska Range, and Kluane samples are *Ursus arctos horribilis* whereas the ABC, Izembek and southeast coastal areas fall within the range of *U.a. dalli*.

Source population / *N*	Population to which individuals were assigned						
	Adm.	B-C	Klu.	Ala.	Kus.	Ize.	Kod.
Admiralty / 30	**29**	1					
B-C / 35		**35**					
Kluane / 50		1	**45**	4			
Alaska Rge. / 28	1		1	**24**	1	1	
Kuskokwim / 55			1	1	**47**	6	
Izembek / 14					2	**12**	
Kodiak / 34							**34**
Coast (l–z) / 15	1	2	7	5			

interpretated in terms of migration…

Second algorithm : Cornuet et al. 1999

This method does not assume HW nor linkage equilibrium, it is strictly based on individual genetic distances

Distances = Cavalli-Sforza & Edwards chord distance, shared allele distance or $(\delta\mu)^2$ especially designed for microsatellites

Focal individuals are assigned to the "closest" population, i.e. the population showing the shortest distance to the focal individual

The main potential problem of both algorithms

Those algorithms always assign individuals to the population showing the largest "score" (highest likelihood or shortest distance)

However, the set of sampled populations may not contain the true population of origin of the focal individual

➠ need for a measure of the confidence of each assignment

The exclusion method of Cornuet et al. 1999

Principle: Confidence measure based on the estimation by simulation of the distribution of the assignment score (for all possible genotypes) for membership in a population

Computing the assignment score for all possible genotypes is too computationally intensive ➠ Monte Carlo simulations

The exclusion method of Cornuet et al. 1999

Principle: Confidence measure based on the estimation by simulation of the distribution of the assignment score (for all possible genotypes) for membership in a population

Simulation method of Cornuet et al. 1999 :

1. Simulate a large number of genotypes (e.g. 1000) from the (estimated) allelic frequencies in the population

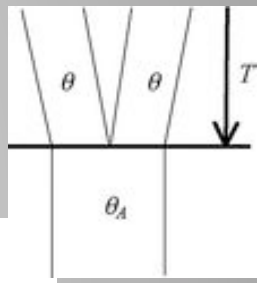
2. Compute the assignment score for each of those simulated genotypes ➠ "null" distribution

3. Compute the probability of observing the focal individual score under the null distribution

The exclusion method of Cornuet et al. 1999

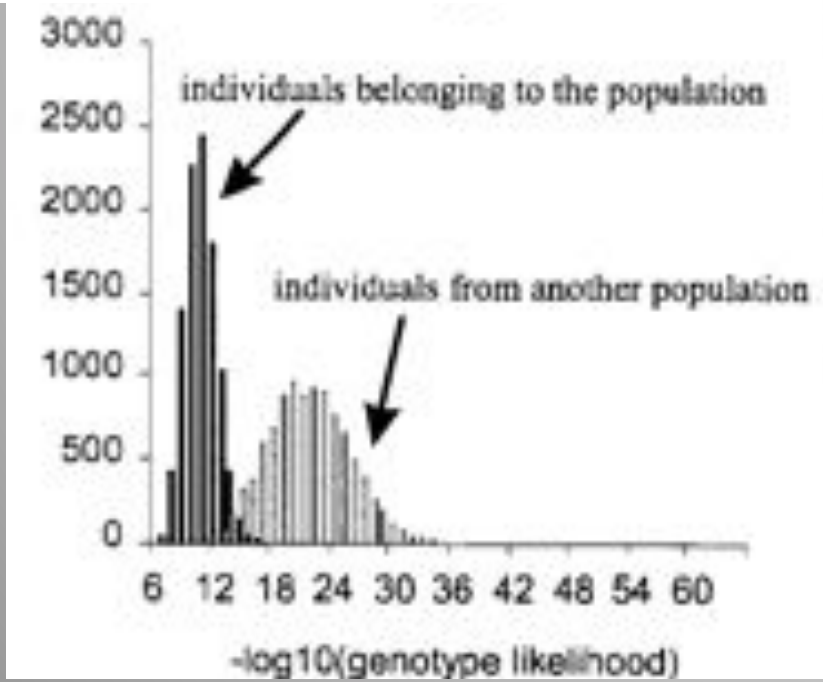
Principle: Simulation of the null distribution of the assignment score for membership in a population

The proportion of the distribution with assignment scores lower than the score of the focal individual gives a measure of the probability that the focal individual is effectively a member of the tested population

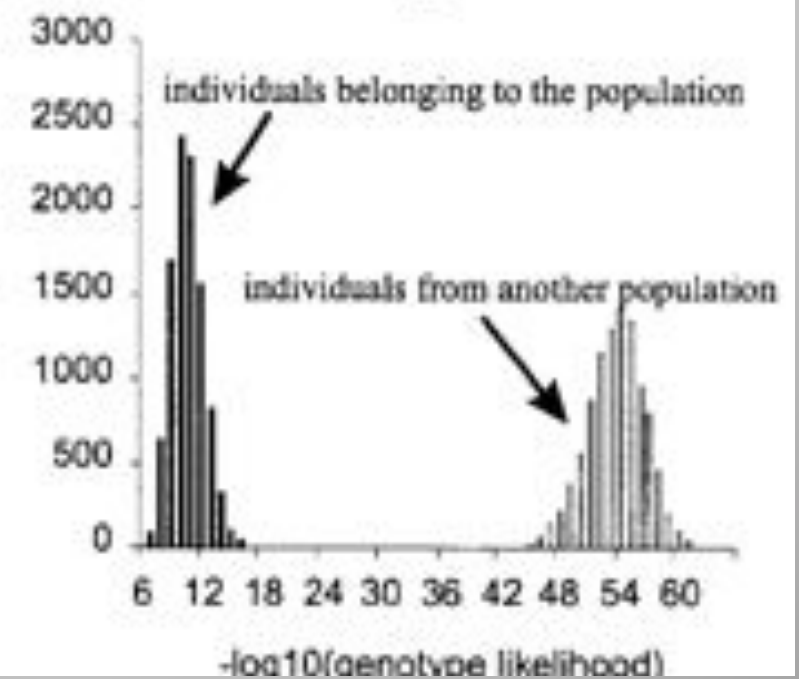


simulation test in 2 diverging populations :

Divergence = 20 generations



Divergence = 200 generations



Comparison of different algorithms (Cornuet et al. 1999)

Simulation test under a model of divergence of the effects of:

- Mutational model
- Sample sizes
- Locus number
- differentiation (i.e. divergence time)

on the proportion of well classified individuals
with the methods of

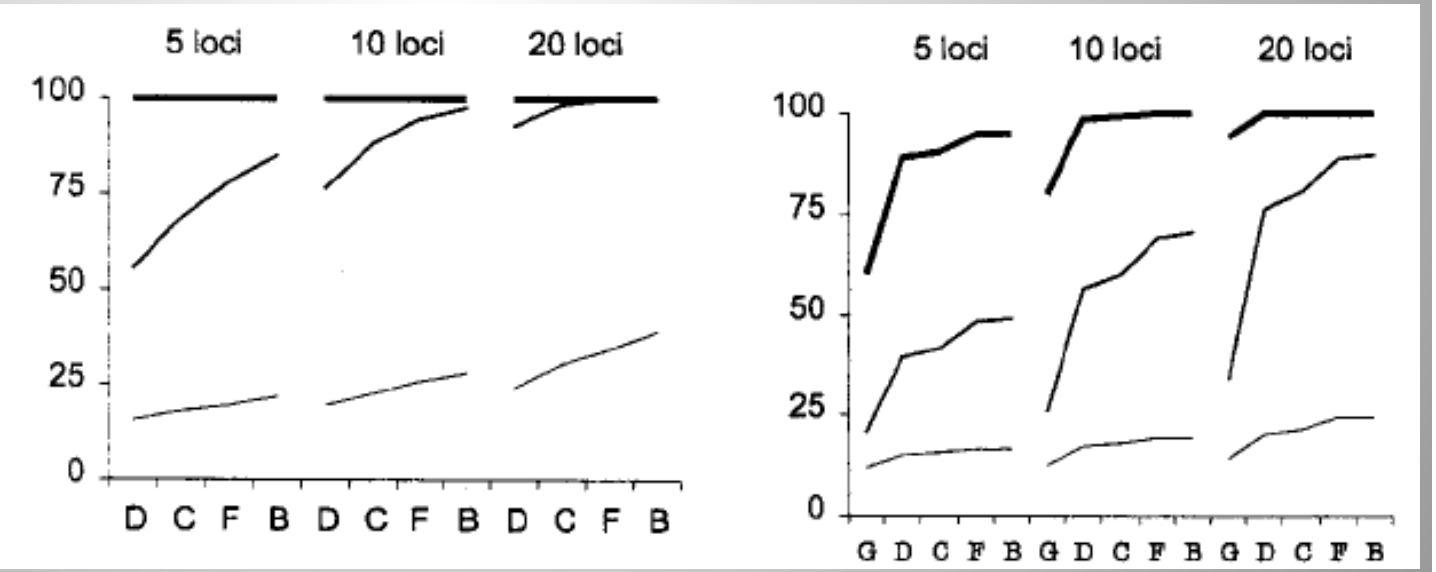
Paetkau et al. 1995 (F), Rannala & Mountain 1997 (B, highly similar to F),

and the distance method of Cornuet et al. 1999 with shared allele distance (D), Cavalli-Sforza a Edwards distance (C) and $(\delta\mu)^2$ (G only for SMM)

Comparison of different algorithms (Cornuet et al. 1999)

- Mutational model :
  Infinite number of Allele Model (IAM, no homoplasy ➠ most informative model) vs. Stepwise Mutation Model (SMM, for microsatellites)
- Differentiation (Fst, directly linked to divergence time Div T)
- Locus number

Div T	Fst
2000	0.3
200	0.08
20	0.01

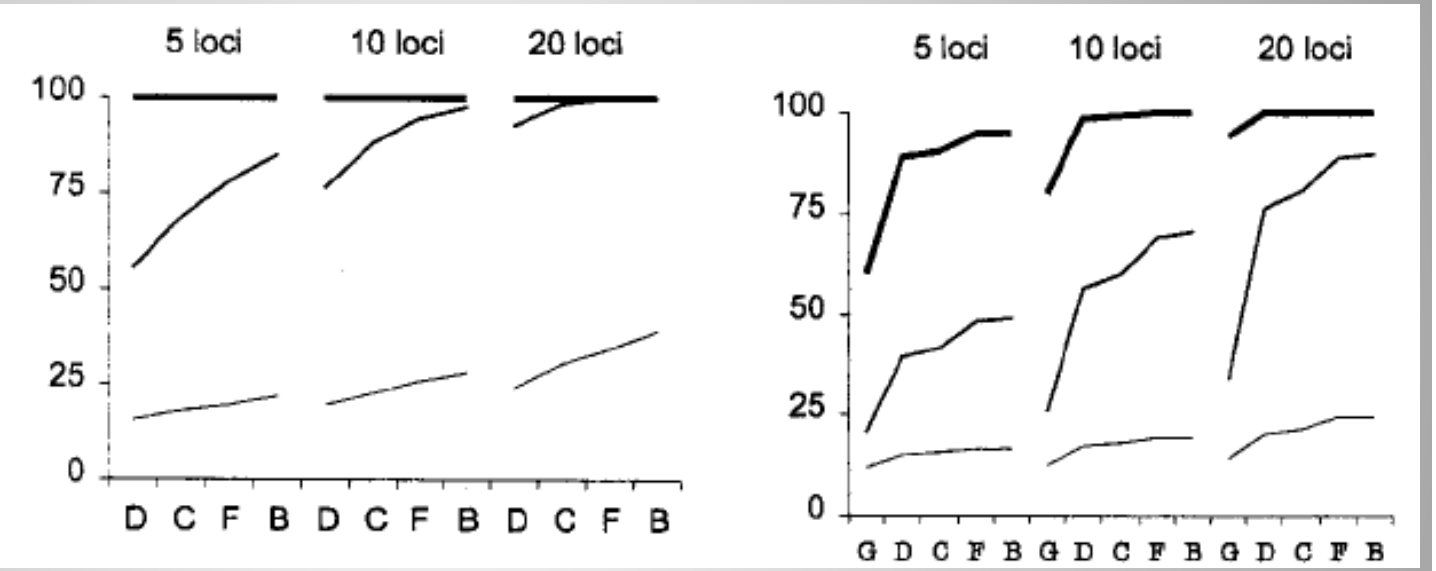


IAM SMM

Comparison of different algorithms (Cornuet et al. 1999)

- IAM vs. SMM, differentiation level, locus number

Div T	Fst
2000	0.3
200	0.08
20	0.01



IAM SMM

- strong effect of the mutation processes, better under IAM than SMM
- B > F > chord distance > shared alleles distance > $(\delta\mu)^2$ distance
- better for larger differentiation and larger number of loci

➠ no surprise

From individual assignments to the inference of migration rates

- Cornuet et al. (1999) is a good example for comparison of methods using simulations but no consideration of migration (pure divergence model)

Most models in population genetics ($F_{statistics}$, diffusion, coalescent) assume demographic equilibrium (mutation – drift - migration)

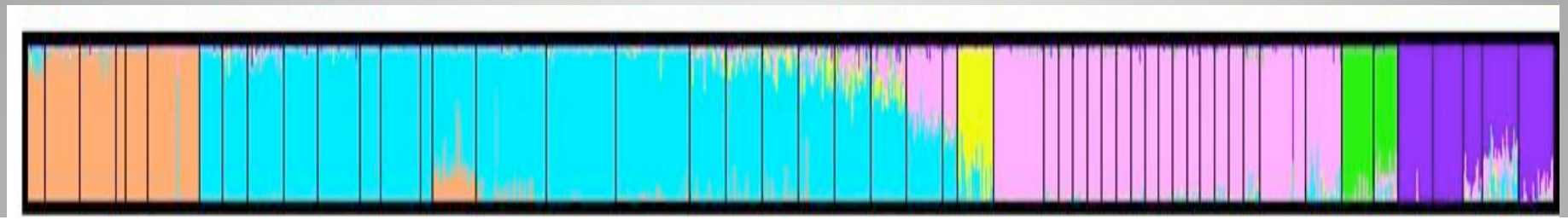
➠ Integrative over long time periods (with few exceptions e.g. IBD)

➠ recent migration events are hardly detectable with such methods

By contrast, no demographic equilibrium assumptions for assignment methods

➠ allows to study recent migration processes

non-spatialized clustering : the STRUCTURE software



Inference of Population Structure Using Multilocus Genotype Data

Jonathan K. Pritchard, Matthew Stephens and Peter Donnelly

Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom

Falush, Stephens, and Pritchard (2003, 2007)

Hubisz, Falush, Stephens and Pritchard (2009)

STRUCTURE Objectives

Grouping individuals into homogeneous genetic clusters using their multilocus genotypes only,

and jointly inferring allele frequencies in those clusters

Also :

- Inferring the level of introgression/hybridization of each individuals
- Inferring the origin of a particular locus (i.e. a part of a chromosome)
- Inferring the most likely number of cluster *K* in a data set

STRUCTURE
principle and assumptions

Same assumptions than for assignment methods:

Hardy-Weinberg equilibrium in each cluster
linkage equilibrium between loci

“Our main modeling assumptions are Hardy-Weinberg equilibrium within populations and complete linkage equilibrium between loci within populations”

“Loosely speaking, the idea here is that the model accounts for the presence of HWD or LD by introducing population structure and attempts to find populations groupings that (as far as possible) are not in disequilibrium”

STRUCTURE: The model

the data = *X* = individual multilocus genotypes (genetic sample)

	locus 1		**locus *i***		**locus l**	
	j=1	**j=2**	**j=1**	**j=2**	**j=1**	**j=2**
$X =$	$(x_1^{(1,1)}$	$x_1^{(1,2)})$	$(x_l^{(1,1)}$	$x_l^{(1,2)})$	$(x_L^{(1,2)}$	$x_L^{(1,2)})$
	...	...	...	...	...	...
	$(x_1^{(i,1)}$	$x_1^{(i,2)})$	$(x_l^{(i,1)}$	$x_l^{(i,2)})$	$(x_L^{(i,1)}$	$x_L^{(i,2)})$
	...	...	...	...	...	...
	$(x_1^{(N,1)}$	$x_1^{(N,2)})$	$(x_l^{(N,1)}$	$x_l^{(N,2)})$	$(x_L^{(N,1)}$	$x_L^{(N,2)})$

X is (N×2I) where N is the number of individuals and I the number of loci

STRUCTURE: The model

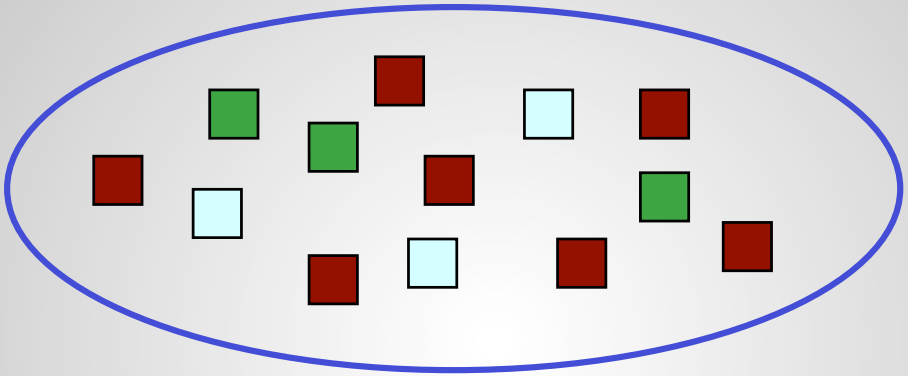
the data = *X* = individual multilocus genotypes (genetic sample)

microsatellite data set example

	phi011		phi015		phi029		phi031		phi062	
1	215	215	82	98	150	150	223	223	164	164
2	218	218	82	102	158	158	187	227	164	164
3	218	218	86	98	150	150	187	227	164	164
4	215	215	86	98	154	154	187	191	164	164
5	218	218	-9	-9	154	158	191	223	164	164
6	215	215	86	86	158	158	227	227	164	164

STRUCTURE: Modeling cluster of origin

Model 1



MODEL 1= (“basic”) model : 'without admixture'

Assumption :

each individual come from a unique ancestral population (cluster)

i.e., all his genes come from a unique cluster among the K possible clusters

STRUCTURE: Model 1

Z = cluster membership of each individual

$$z = \begin{bmatrix} z^{(1)} \\ \cdots \\ z^{(i)} \\ \cdots \\ z^{(N)} \end{bmatrix}$$

Z is a vector of length N

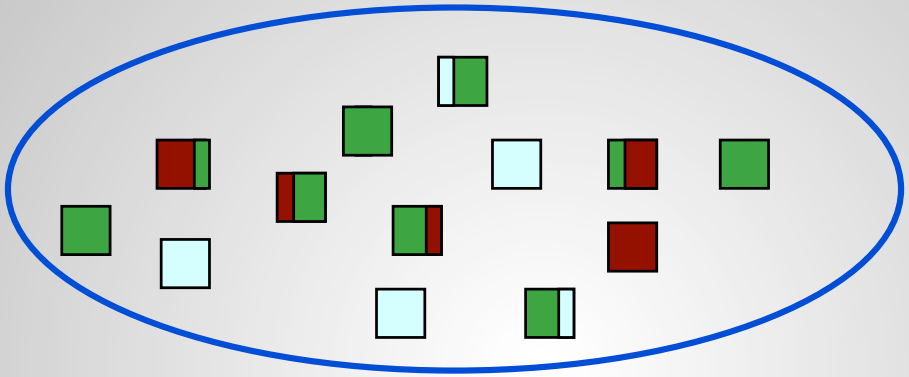
- if individual *i* is a member of cluster *k* then
  $z^{(i)} = k$

$P(z^{(i)} = k)$ is the probability that individual *i* is a member of cluster *k*

$$z_{(N \times 1)}$$

STRUCTURE: Modeling cluster of origin

Model 2



MODEL 2 = model with 'admixture' (most commonly used)

Assumption:

the different genes of an individual may come from different clusters due to recent introgression / hybridization / migration events.

Inference is then done on the proportion of genes *Q* that comes from the *K* different clusters

STRUCTURE Model 2

Z = cluster membership of each locus allele copy from each individual => Z is a matrix of dimension *N x 2L* (Rq: if haploids: *N x L*)

$$Z = \begin{bmatrix} (z_1^{(1,1)} & z_1^{(1,2)}) & \cdots & (z_l^{(1,1)} & z_l^{(1,2)}) & \cdots & (z_L^{(1,2)} & z_L^{(1,2)}) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ (z_1^{(i,1)} & z_1^{(i,2)}) & \cdots & (z_l^{(i,1)} & z_l^{(i,2)}) & \cdots & (z_L^{(i,1)} & z_L^{(i,2)}) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ (z_1^{(N,1)} & z_1^{(N,2)}) & \cdots & (z_l^{(N,1)} & z_l^{(N,2)}) & \cdots & (z_L^{(N,1)} & z_L^{(N,2)}) \end{bmatrix}$$

$Z_{(N \times 2L)}$

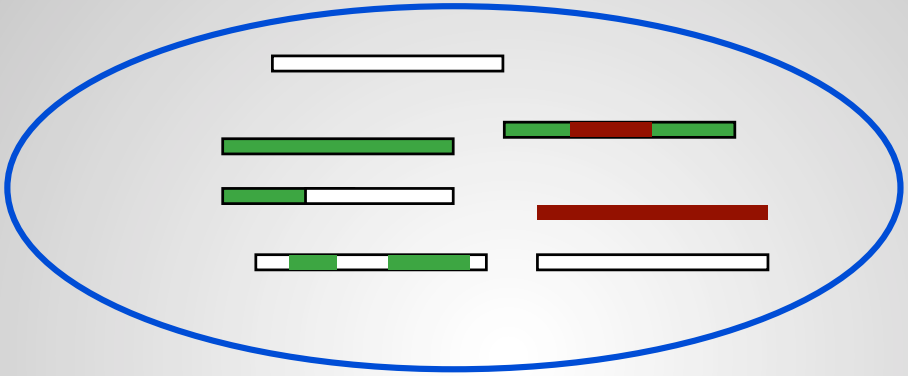
P($z_n^{(i,l)}$ = *k*) is the probability that the allele copy carried by the individual *n* at locus *i* on his/her chromosome *l* *(eg l=1 if paternal and l=2 if maternal)* originates from cluster *k*

Q = proportion of individual genome from each cluster => Q is a matrix of dimension N x K

$$Q = \begin{pmatrix} q_1^1 & \cdots & q_1^K \\ \cdots & \cdots & \cdots \\ q_N^1 & \cdots & q_N^K \end{pmatrix}$$

STRUCTURE: Extending model 2 to account for LD *(i.e. dependancy of loci along chromosomes)*

MODEL 3



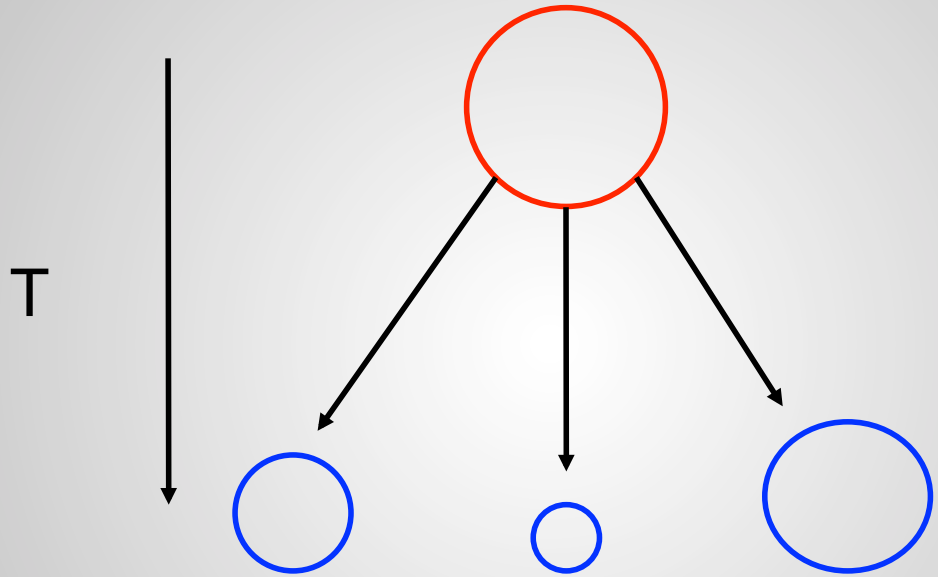
MODEL 3: the 'linkage' model ***(explicit recombination on chromosomes)***

generalization of the admixture model with higher probabilities of coming from the same cluster for different loci with low level of recombination

i.e. different "chunks" on each chromosomes may come from different clusters => $P(z_n^{(i,l)} = k \mid z_n^{(i-1,l)} = k)$ depends on r

STRUCTURE: Extending model 2 to account for correlation of cluster All. Freq.

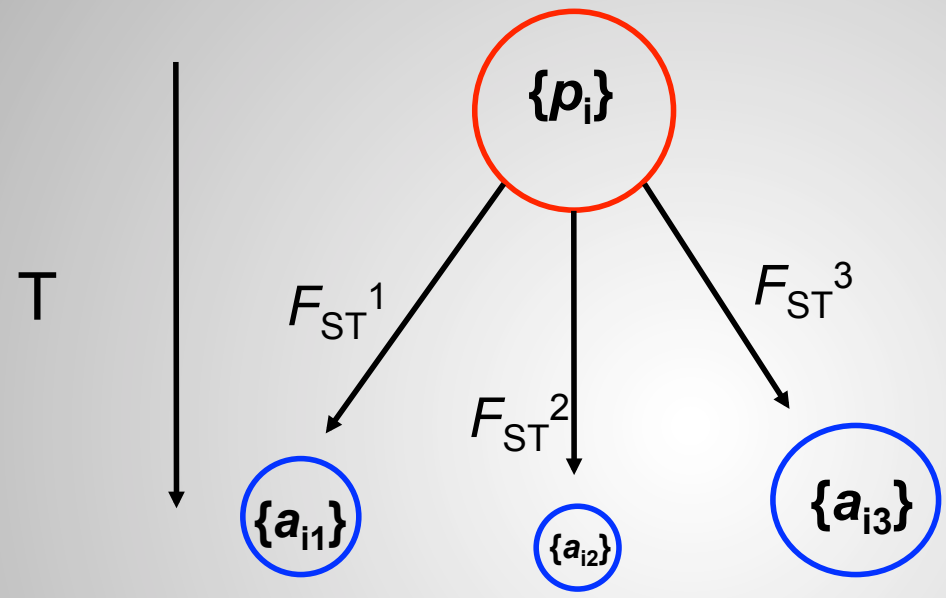
MODEL 4



MODEL 4 = Modeling correlation among allele freq. across clusters: the F-model

Assuming cluster represent ancestral population, allele frequencies across cluster might be correlated due to demographic history relating those populations.

MODEL 4



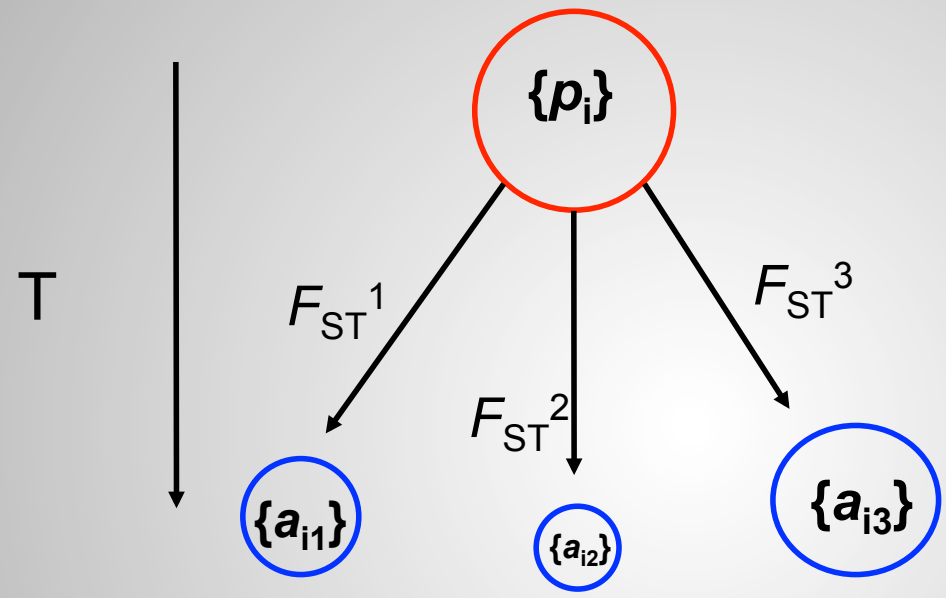
4. Modeling correlation among allele freq. across clusters: the F-model

{p_i} : Vector of allele frequencies at locus i in the pop ancestral to all the cluster pops (star-shaped phylogeny);

{a_{ik}} : Vector of allele frequencies at locus i in cluster pop k

{F_{ST}^k} : amount of differentiation (div. time≈t/2N) between the actual cluster and the population ancestral to all clusters

MODEL 4

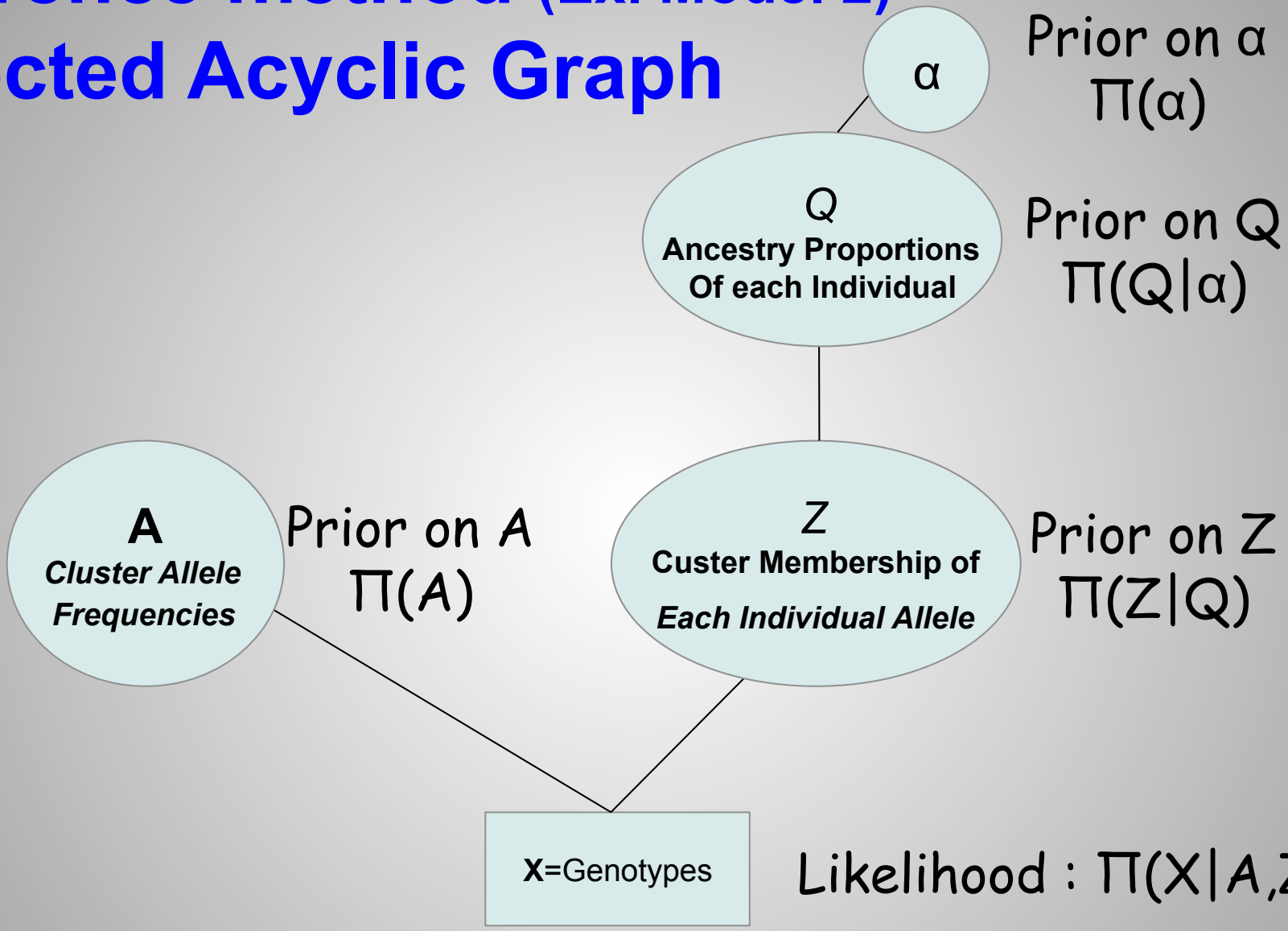


4. Modeling correlation among allele freq. across clusters: the F-model

Statistical modeling:

! Interpretation as a pure drift model of divergence (works rather well providing divergence time are not too high F_{ST}<0.4)

Inference method (Ex. Model 2)
Directed Acyclic Graph



$$\pi(X,Q,Z,A) = \pi(X \mid Q,Z,A)\pi(Z \mid Q)\pi(Q \mid \alpha)\pi(A)\pi(\alpha)$$

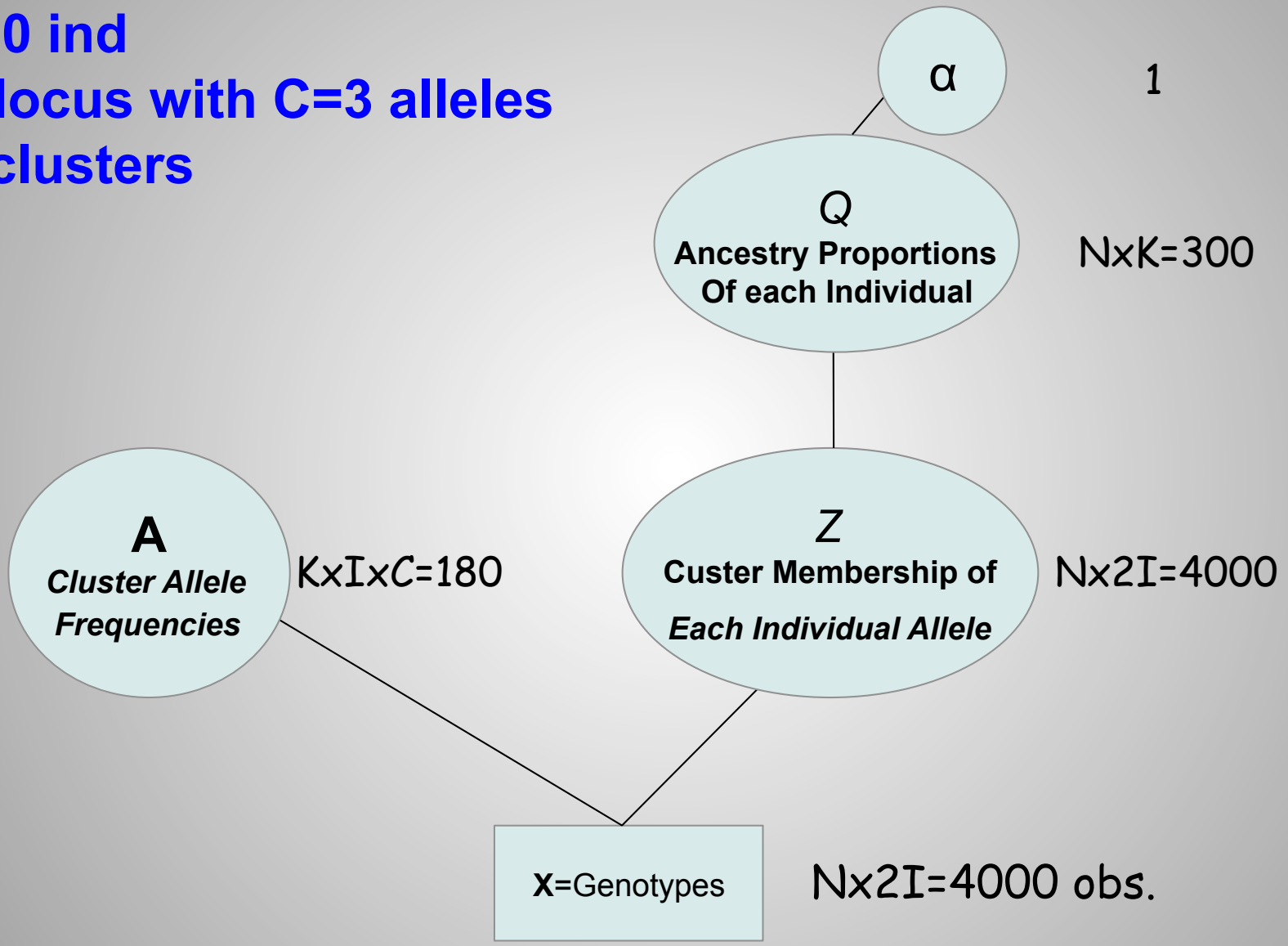
Number of Parameters

Ex.

N=100 ind

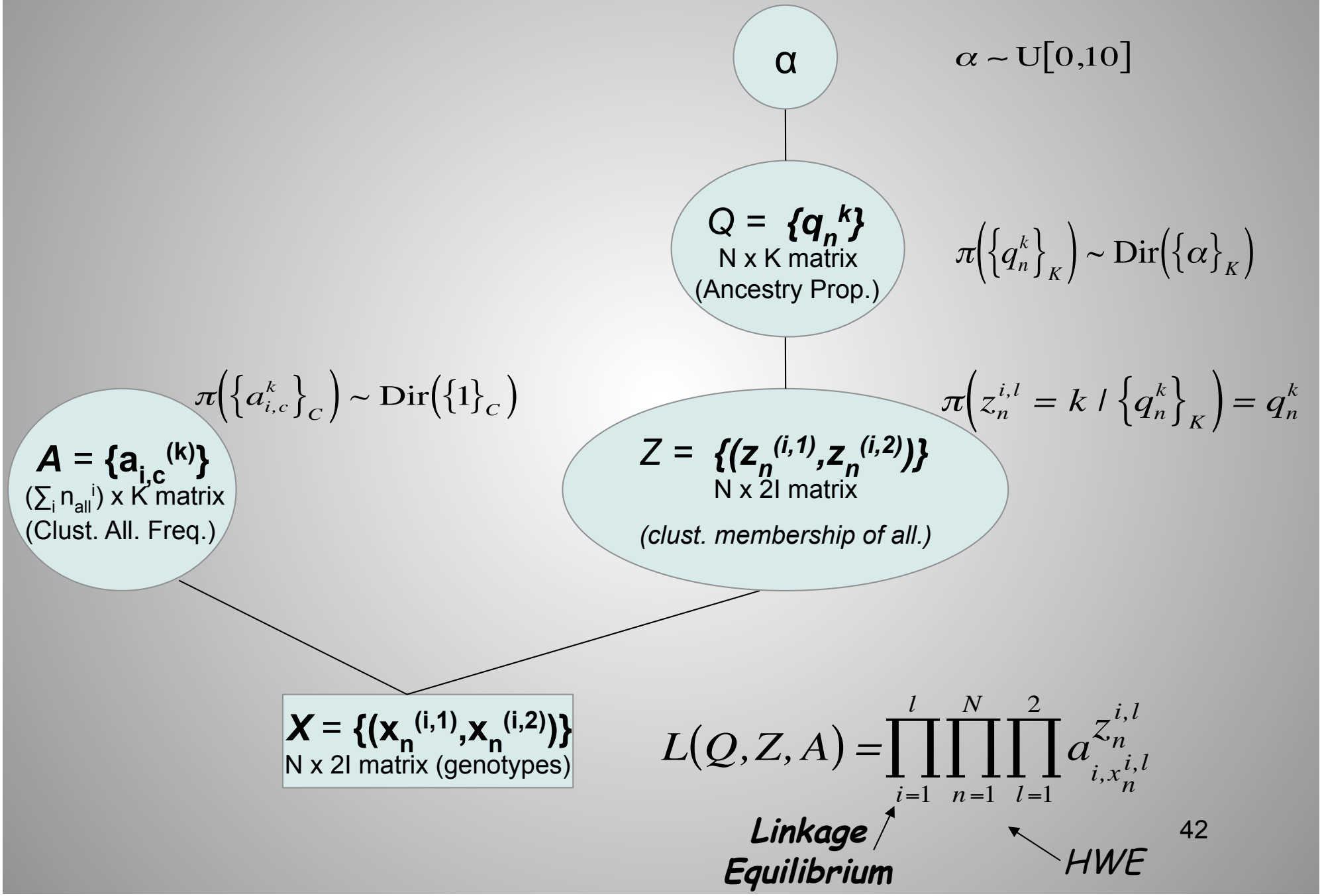
I=20 locus with C=3 alleles

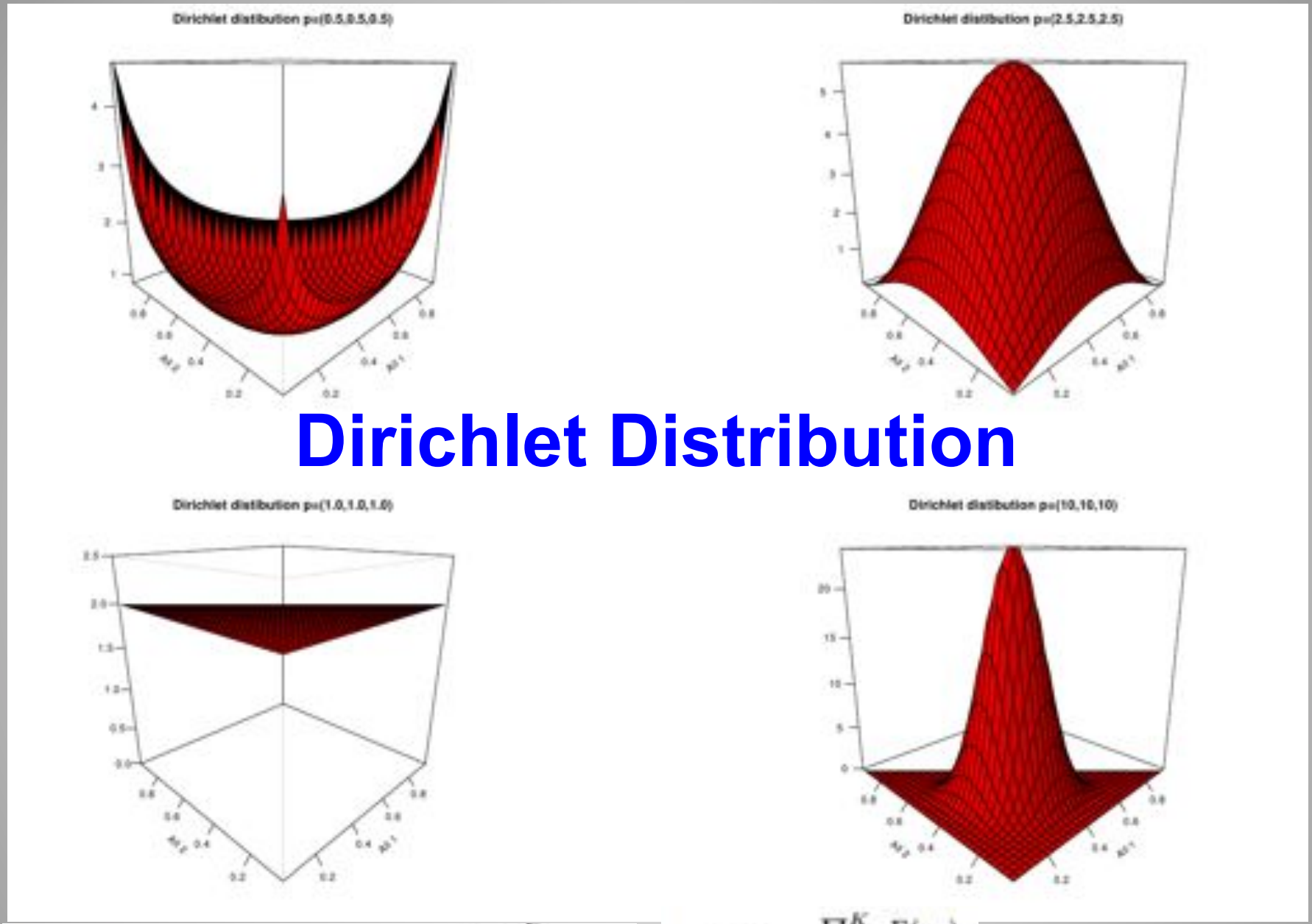
K=3 clusters



Ex. : 4,000 + 180 + 300 + 1 = 4,481 parameters !!!

Model 2 Specification





Dirichlet Distribution

$$f(x_1,\ldots,x_{K-1};\alpha_1,\ldots,\alpha_K) = \frac{1}{\mathrm{B}(\alpha)}\prod_{i=1}^{K} x_i^{\alpha_i-1}$$

$$\mathrm{B}(\alpha) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)},$$

Estimation of Parameters
MCMC

Step 1: Initialize all parameter values. For instance:

- Cluster Allele Freq.:

  $(a_{i,c}^{(k)}(0)) = 1/C$ (C Nbr of alleles)

- Ind. Allele membership:

  $z_n^{(i,l)}(0) = (1,1)$ *or* $(2,K)$ *or* $(1,K)$ *… at random*

- Ind. Ancestry Proportion*:*

  $q_n^k\,(0) = (1/K\ ,\ 1/K,\ldots,1/K)$

Step 2: Iterate from t=1 to t=T times sampling from full conditional distributions for each parameters in turn to obtain samples from the posterior distributions of the parameter of interest

MCMC algorithm: Step 2a)

2a) Vector of cluster allele of frequencies at marker *i* given others parameters values

=> I step (one per locus)

=> $\{a_{i,c}^{(k)}(t)\}_C \mid X, Z(t-1)$

$$\left\{a_{i,c}^{k}\right\}_C / \left\{x_n^{i,l}\right\}_N, \left\{\left(z_n^{i,1}, z_n^{i,2}\right)\right\}_N \sim \mathrm{Dir}\left(\left\{1 + n_{k,c}\right\}_C\right) \quad \left(n_{k,c} := \#\, x_n^{i,l} = c \text{ and } z_n^{i,l} = k\right)$$

2a) Exemple: Update Freq. Of Marker 1 in cluster 1

=>Marker 1 has three alleles: 110-113-114

=>At iteration t-1, given Z:

– 50 alleles 110, 28 alleles 113 and 12 alleles 114 among the observed ones (X) originate from cluster 1

=>$\{a_{1,"110"}^{(1)}, a_{1,"113"}^{(1)}, a_{1,"114"}^{(1)}\}(t) \sim \mathrm{Dir}(\{51,29,13\})$

MCMC algorithm: Step 2b, 2c, and 2d

2b) Update cluster membership of each of the two alleles from each individual in turn ($z_n^{(i,l)}(t) \mid X, Q(t), A(t)$)

=> 2NxI steps (2 per individual and per locus)

$$P\left(z_n^{i,l} = k \mid ..\right) = \frac{q_k^i P\left(x_n^{i,l} \mid \left\{a_i^k\right\}\right)}{\sum_{k'=1}^{k'=K} q_{k'}^i P\left(x_n^{i,l} \mid \left\{a_i^{k'}\right\}\right)}$$

2c) Update Ancestry proportion vector of each individual in turn ($\{q_n^i(t)\}_K \mid X, Z(t-1), \alpha(t-1)$)

=> N steps (one per individual)

$$\left\{q_n^k\right\}_K \mid \left\{\left(z_n^{i,1}, z_n^{i,2}\right)\right\}_N, \alpha \sim \mathrm{Dir}\left(\left\{\alpha + m_{n,k}\right\}_K\right) \quad \left(m_{n,k} := \#\, z_n^{i,l} = k\right)$$

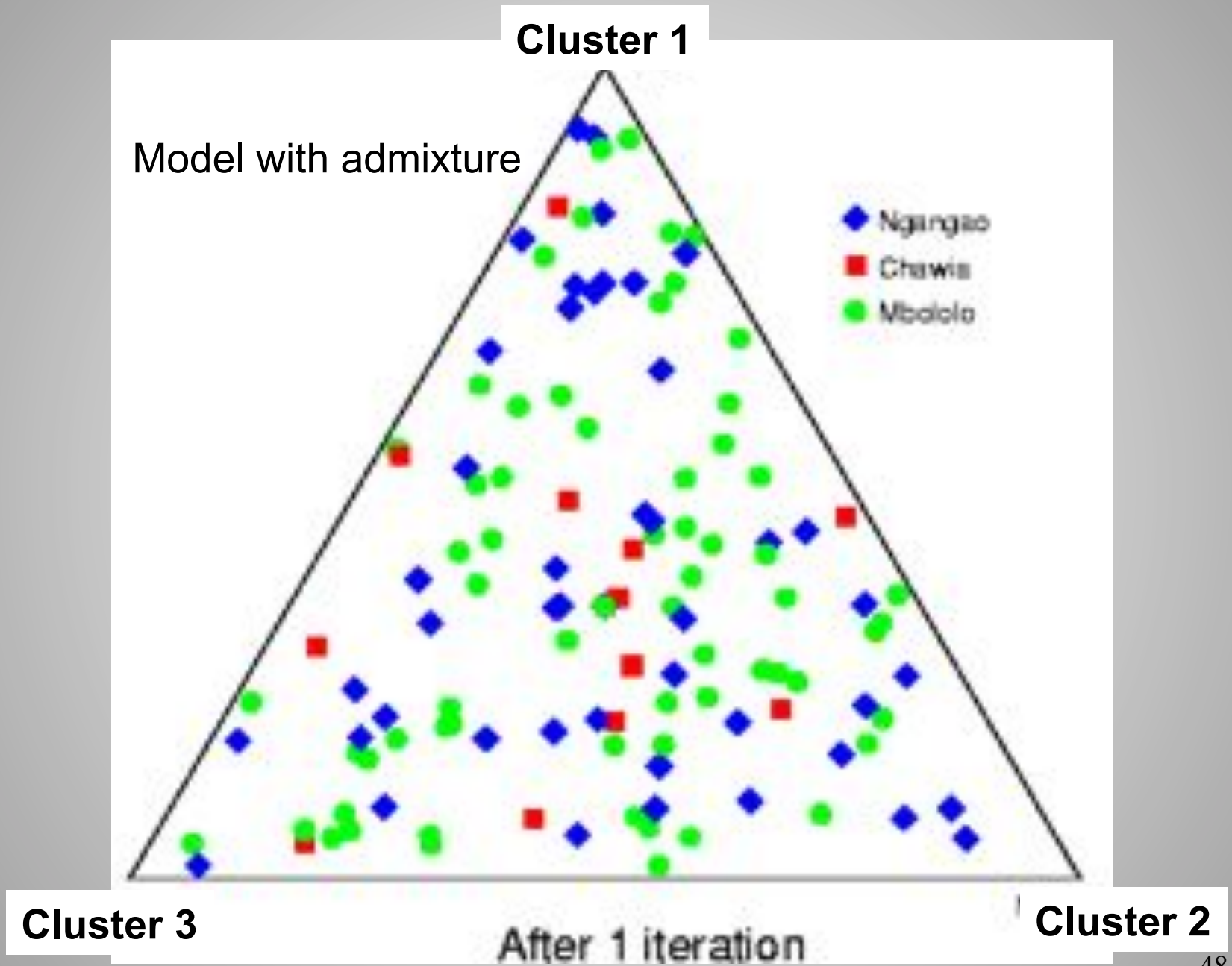
2d) Update parameters α (α(t)|Q(t),α(t-1)):

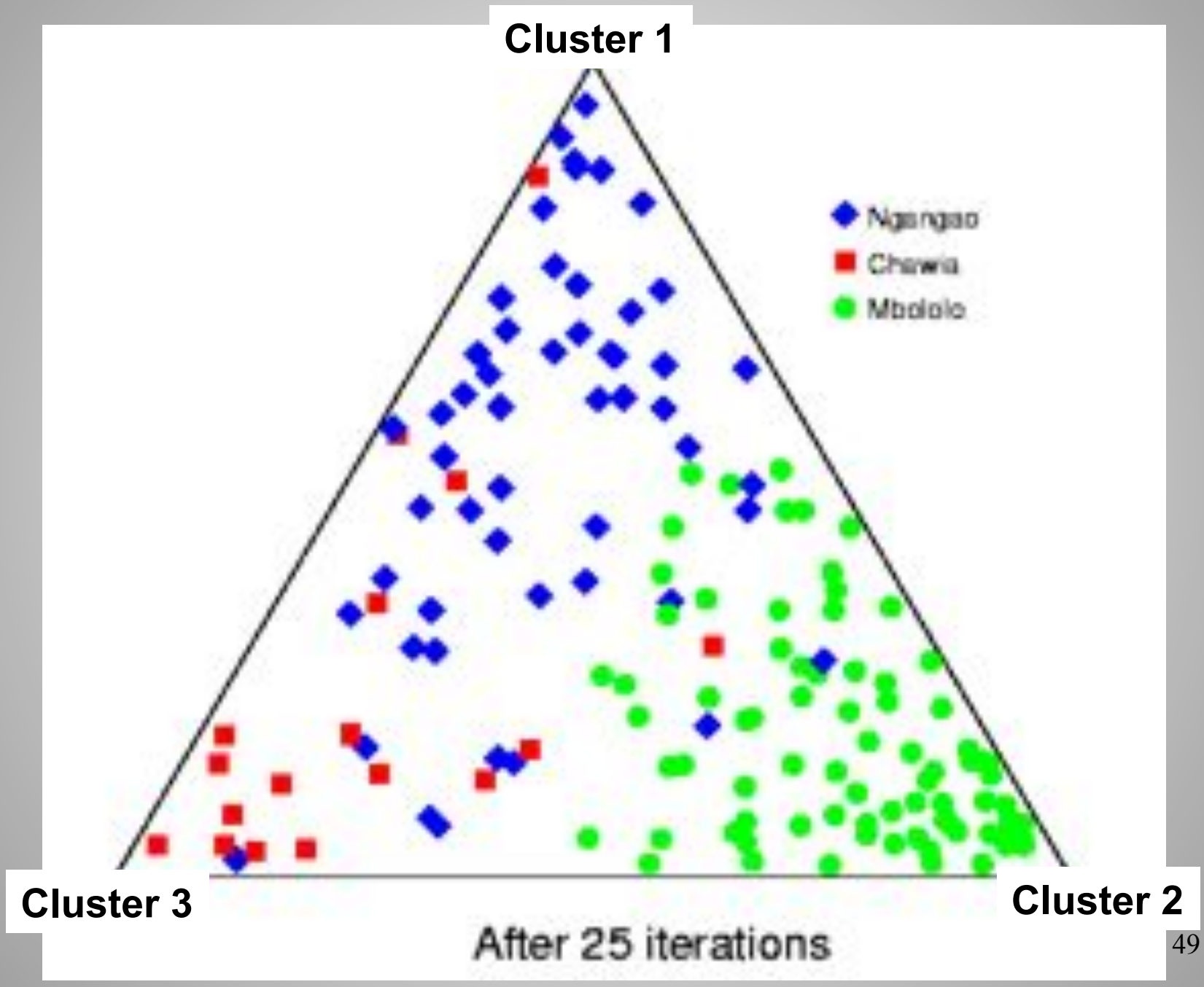
=>not of usual form

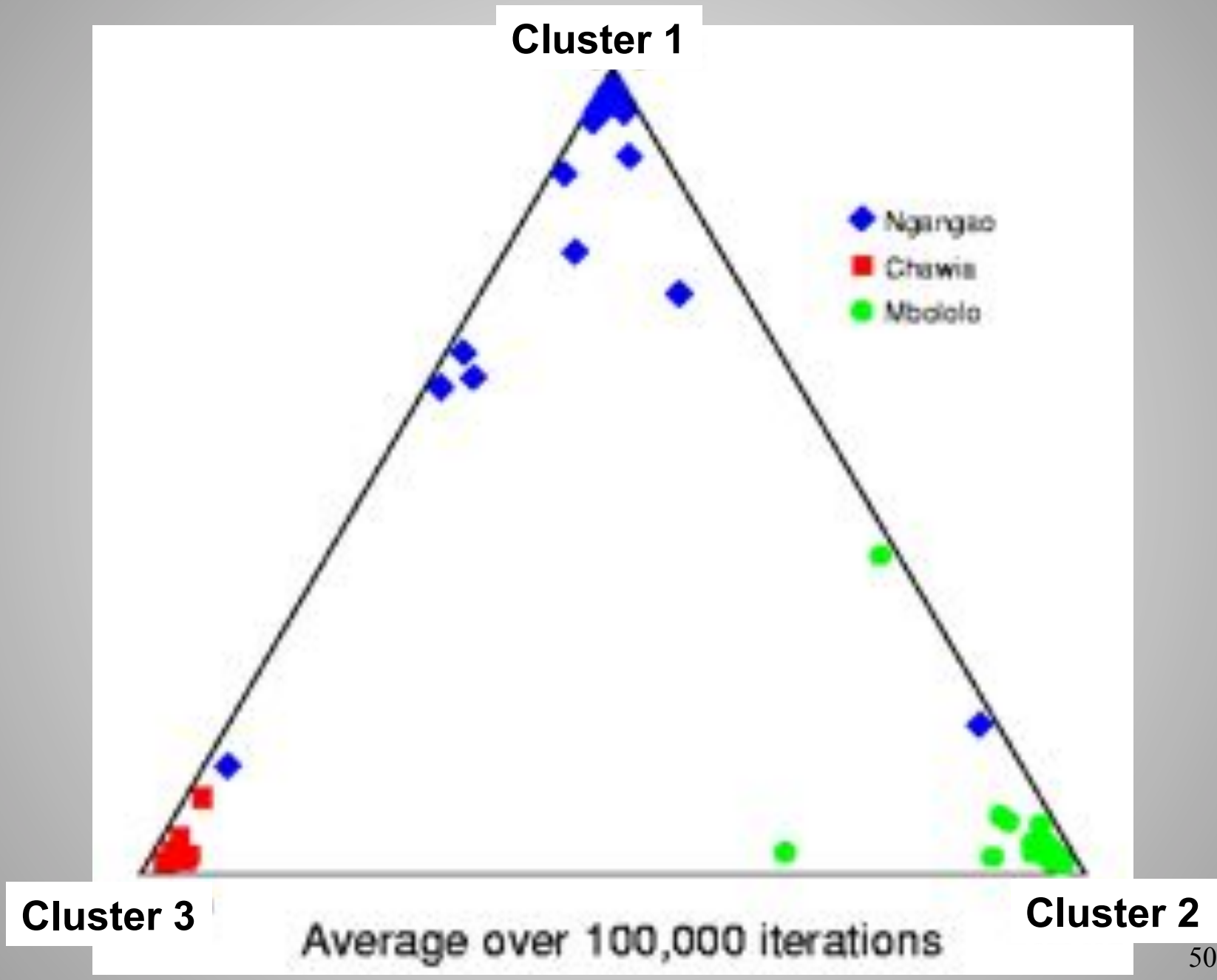
=> 1 step (propose a value and accept/refuse it according to MH rule...)

Example: Taita Thrush data

- three main sampling locations in Kenya
- low migration rates (radio-tagging study))
- 155 individuals, genotyped at 7 microsatellite loci


Cluster 1
Model with admixture
Ngangao
Chawia
Mbalala
Cluster 3
Cluster 2
After 1 iteration


Cluster 1
Ngarngao
Chirwia
Mbololo
Cluster 3
Cluster 2
After 25 iterations


Cluster 1
Ngangao
Chawia
Mbololo
Cluster 3
Cluster 2
Average over 100,000 iterations

Example on highly structured populations

OPEN ACCESS Freely available online

PLoS GENETICS

Genetic Structure of Chimpanzee Populations

Celine Becquet[1], Nick Patterson[2], Anne C. Stone[3], Molly Przeworski[1*], David Reich[2,4*]

1 Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America, 2 Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America, 3 School of Human Evolution and Social Change, Arizona State University, Tempe, Arizona, United States of America, 4 Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America

Little is known about the history and population structure of our closest living relatives, the chimpanzees, in part because of an extremely poor fossil record. To address this, we report the largest genetic study of the chimpanzees to date, examining 310 microsatellites in 84 common chimpanzees and bonobos. We infer three common chimpanzee populations, which correspond to the previously defined labels of "western," "central," and "eastern," and find little evidence of gene flow between them. There is tentative evidence for structure within western chimpanzees, but we do not detect distinct additional populations. The data also provide historical insights, demonstrating that the western chimpanzee population diverged first, and that the eastern and central populations are more closely related in time.

Example on highly structured populations

OPEN ACCESS Freely available online

PLoS GENETICS

Genetic Structure of Chimpanzee Populations

Celine Becquet[1], Nick Patterson[2], Anne C. Stone[3], Molly Przeworski[1*], David Reich[2,4*]

Table 3. Genetic Differentiation among Populations

Location	Eastern	Central	Bonobo
Western	0.31 (0.32)	0.25 (0.29)	0.68 (0.68)
Eastern	—	0.05 (0.09)	0.57 (0.54)
Central	—	—	0.51 (0.49)

Example on highly structured populations

OPEN ACCESS Freely available online

PLoS GENETICS

Genetic Structure of Chimpanzee Populations

Celine Becquet[1], Nick Patterson[2], Anne C. Stone[3], Molly Przeworski[1*], David Reich[2,4*]

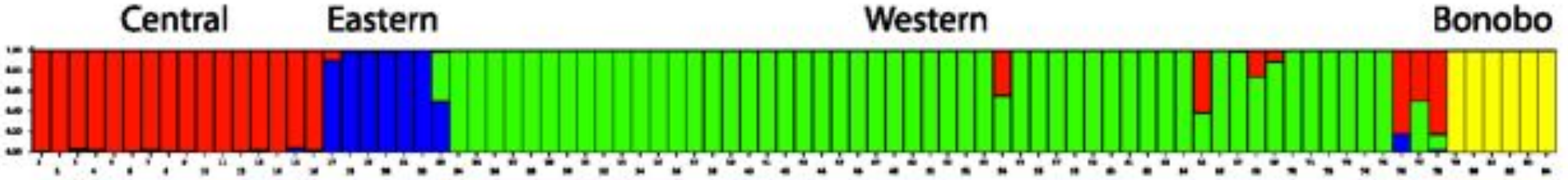


Figure 1. STRUCTURE Analysis, Blinded to Population Labels, Recapitulates the Reported Population Structure of the Chimpanzees

Individuals 76–78 are reported hybrids. Only two individuals with a >5% proportion of ancestry in more than one inferred cluster are wild born: number 54 and number 17. Red, central; blue, eastern; green, western; yellow, bonobo.

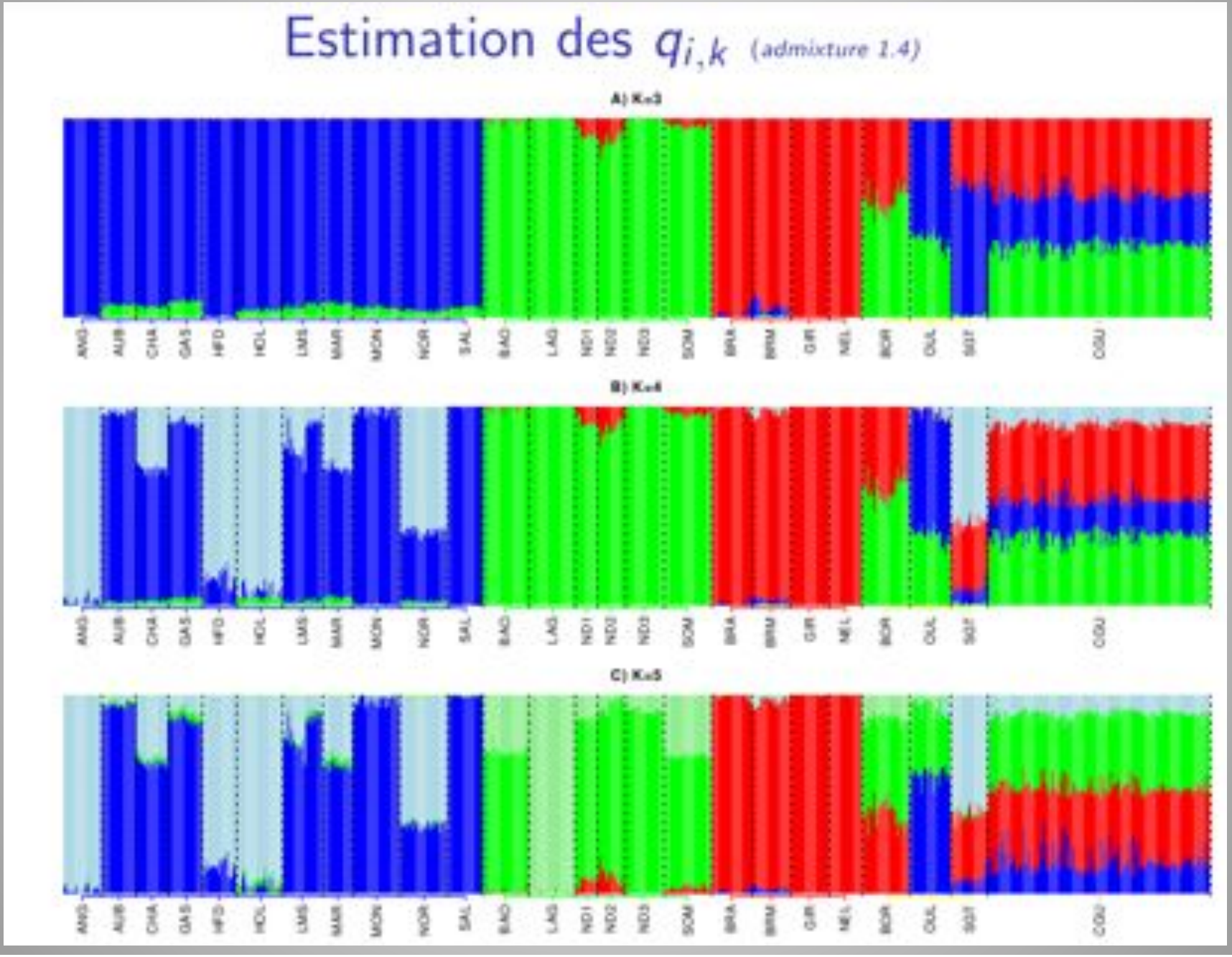
doi:10.1371/journal.pgen.0030066.g001

Very clear structure,
few migration/hybridization events detected

Example on admixed populations

- Taurins Européens (origine ibérique) avec les premiers colons espagnols (1493) et très récemment (races françaises)
- Taurins Africains (même voie que le commerce triangulaire) entre les 16èmeet 18èmesiècle
- Zébus originaires d'Inde (ou d'Afrique ?) introduits en Amérique (Sud et US) au 19èmesiècle

Example on admixed populations



Inference of the number of clusters *K*

STRUCTURE do not infer the number of cluster using MCMC,

K should be inferred afterwards from many MCMC runs with different *K* values by choosing the runs with the higher posterior probabilities of the data :

Assumed value of *K*	Posterior probability of *K*
1	~0
2	~0
3	0.993
4	0.007
5	0.00005

Taita Thrush data

Inference of the number of clusters *K*

STRUCTURE do not infer the number of cluster using MCMC,

Assumed value of *K*	Posterior probability of *K*
1	~0
2	~0
3	0.993
4	0.007
5	0.00005

Taita Thrush data

problem : statistical theory state that the likelihood should always increase between models when the number of degrees of freedom increases

the likelihood should increase with *K* …

there may be a convergence problem with this data set?

Inference of the number of clusters *K*

Hopefully, sometimes it is much better :

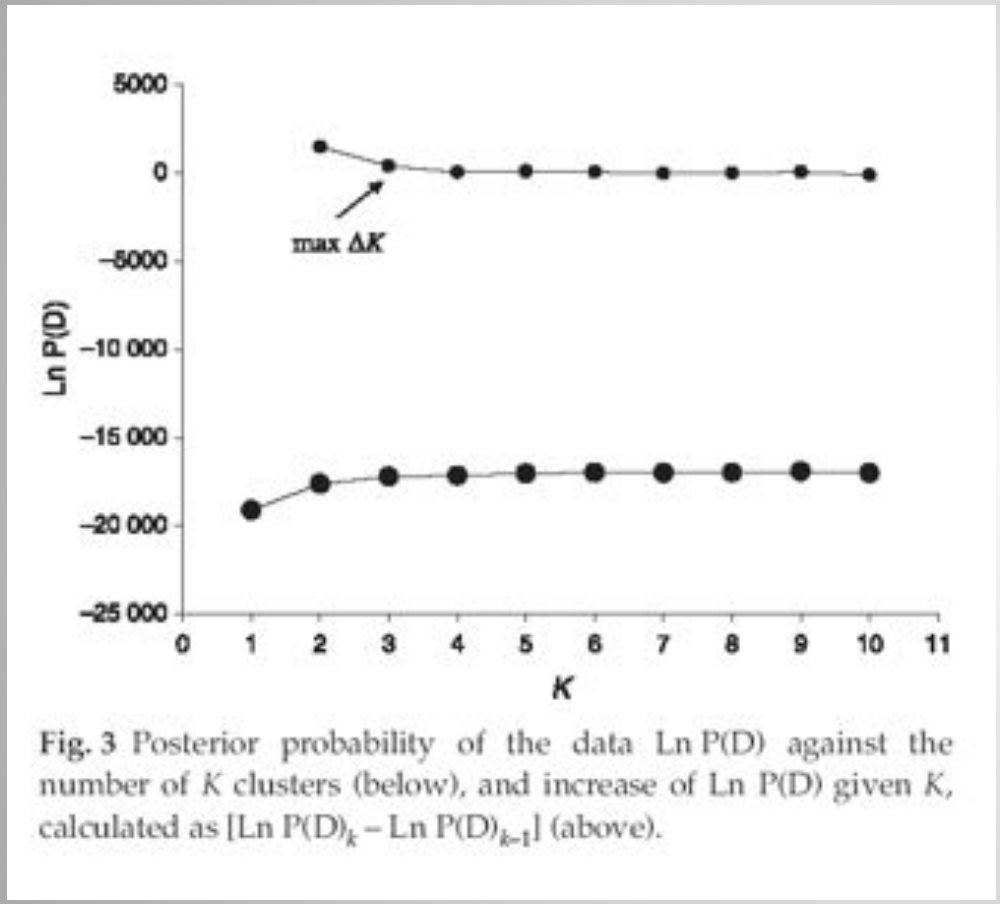


Fig. 3 Posterior probability of the data Ln P(D) against the number of *K* clusters (below), and increase of Ln P(D) given *K*, calculated as [Ln $P(D)_k$ – Ln $P(D)_{k-1}$] (above).

Scottish feral cat

the variation in likelihood between different K values can also be used (ΔK)

Inference of the number of clusters *K*

STRUCTURE do not infer the number of cluster using MCMC, and what K exactly represents is not clear, especially in cases of hierarchical "barriers"/groups

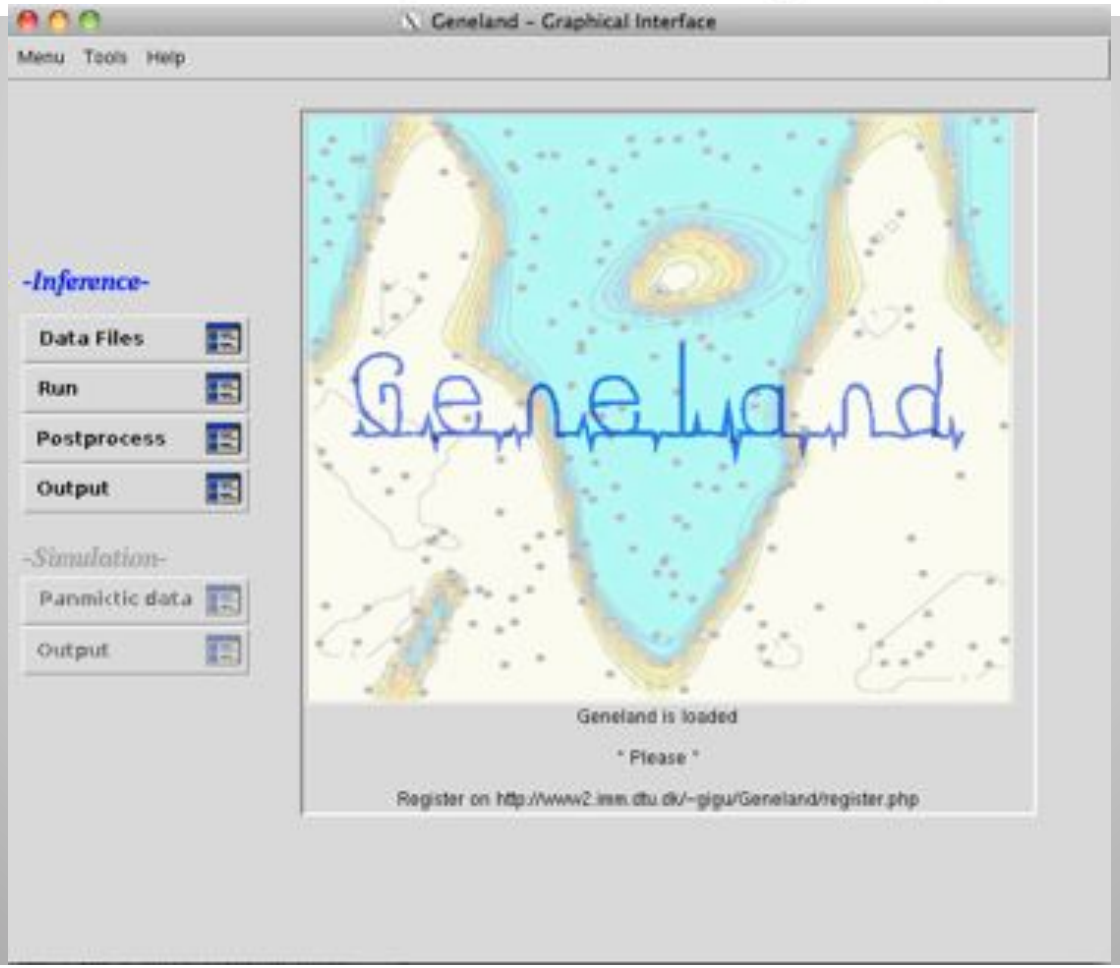
It is usually better to analyze different values of *K*, and conclude from all of them instead of focusing on the "best" *K* value.

Spatial clustering: the GENELAND software

DOI: 10.1534/genetics.104.033803

A Spatial Statistical Model for Landscape Genetics

Gilles Guillot,*,[1] Arnaud Estoup,[†] Frédéric Mortier[‡] and Jean François Cosson[§]



Spatial clustering: the GENELAND software

Aim : spatial delimitation of genetically homogeneous clusters from individual multilocus genotypes with spatial coordinates

= locate genetic discontinuities in space

and also :

- Infer the number of cluster on the sampled area (integrated in the MCMC, but not more meaningful than for STRUCTURE)
- Assign individuals to the different clusters (migrant detection)

GENELAND spatial population model

Set of spatialized panmictic populations

Each cluster (one panmictic population) is a formed by a set of polygons which contains individuals belonging to this cluster :

it is called the colored Voronoi tessellation ➠ 1 pop is 1 color

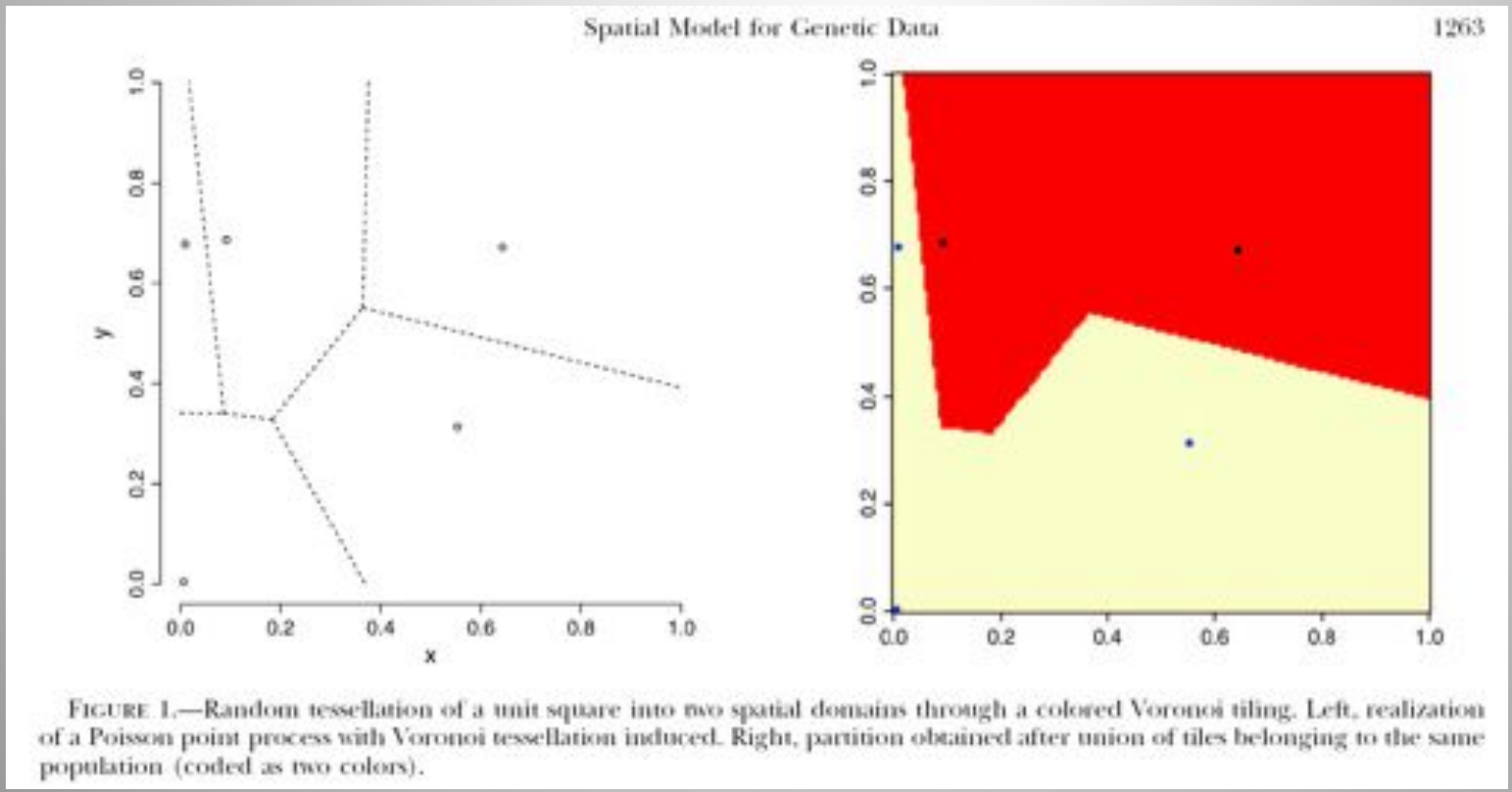
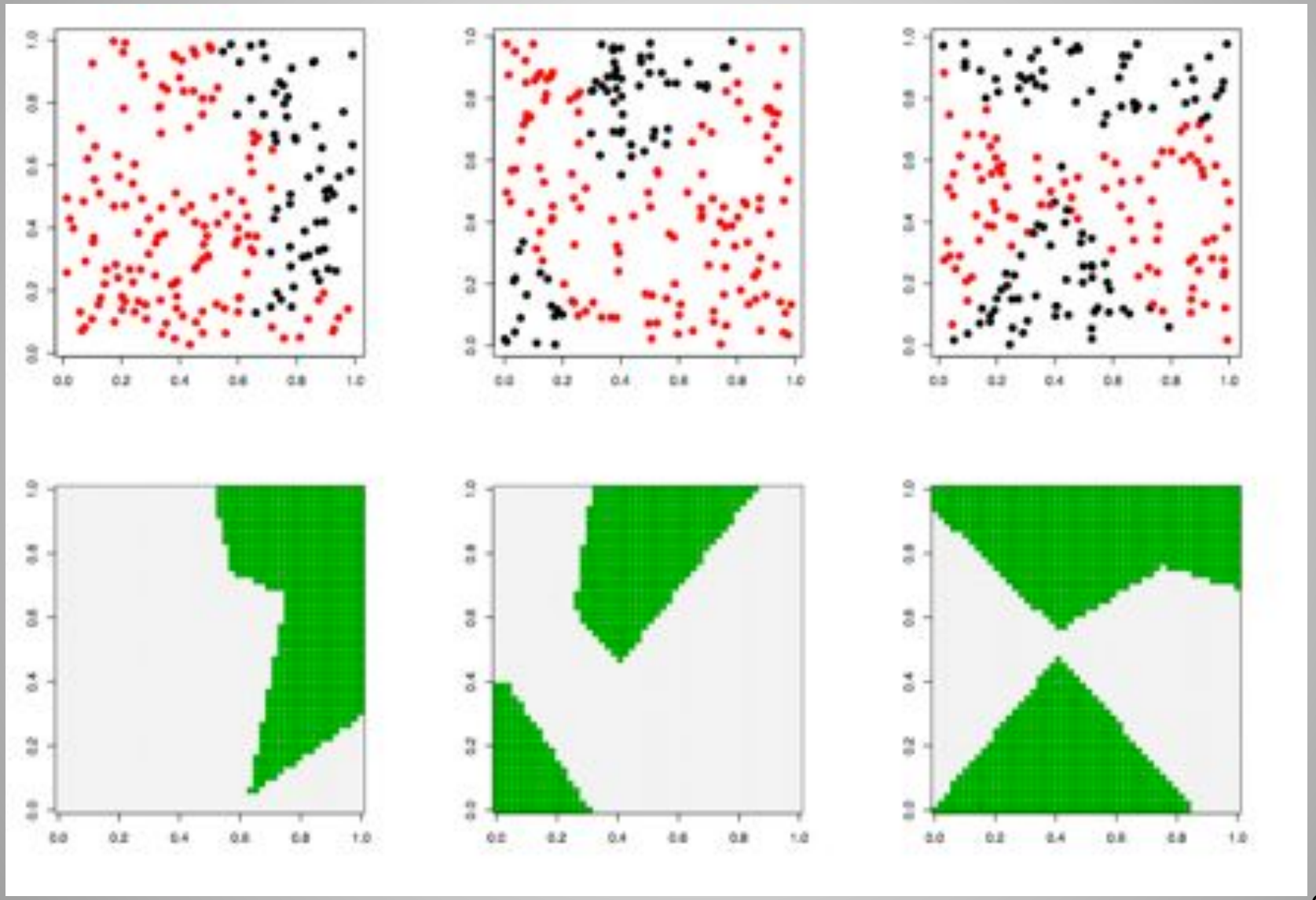


FIGURE 1.—Random tessellation of a unit square into two spatial domains through a colored Voronoi tiling. Left, realization of a Poisson point process with Voronoi tessellation induced. Right, partition obtained after union of tiles belonging to the same population (coded as two colors).

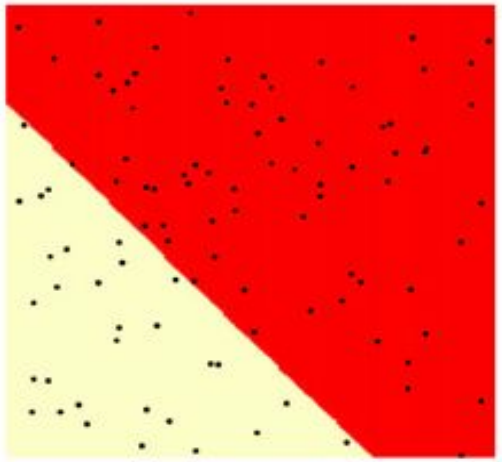
GENELAND spatial population model



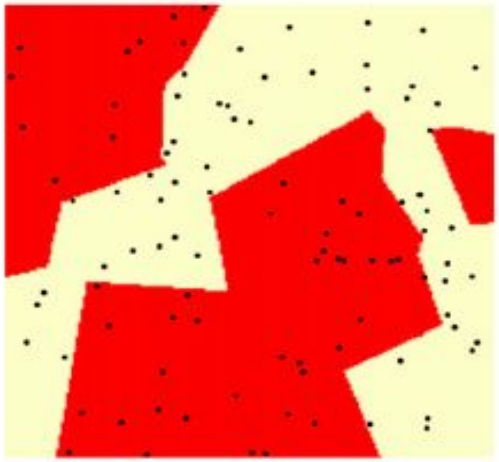
GENELAND spatial population model

Set of spatialized panmictic populations

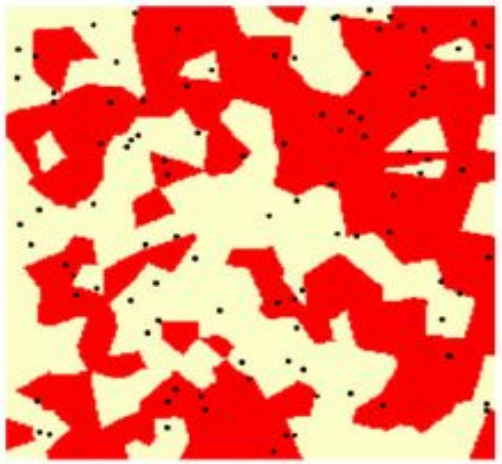
example of different Tessellation outputs for different spatial correlations



$m = 2$



$m = 20$



$m = 200$

FIGURE 6.—Examples of simulated spatial organization of 100 individuals (black dots) into two populations (coded as two colors) with various levels of spatial dependence. This level is controlled by parameter m (number of Voronoi tiles). The nuclei of the tiles are not depicted for clarity.

The spatial correlation is modeled through the parameter *m = max number of disjointed polygons that form a cluster*

small *m* ➠ more spatial correlation,
large m ➠ less spatial correlation
because p(2 ind ∈ single cluster) increase with m

! not really linked to IBD !

GENELAND method

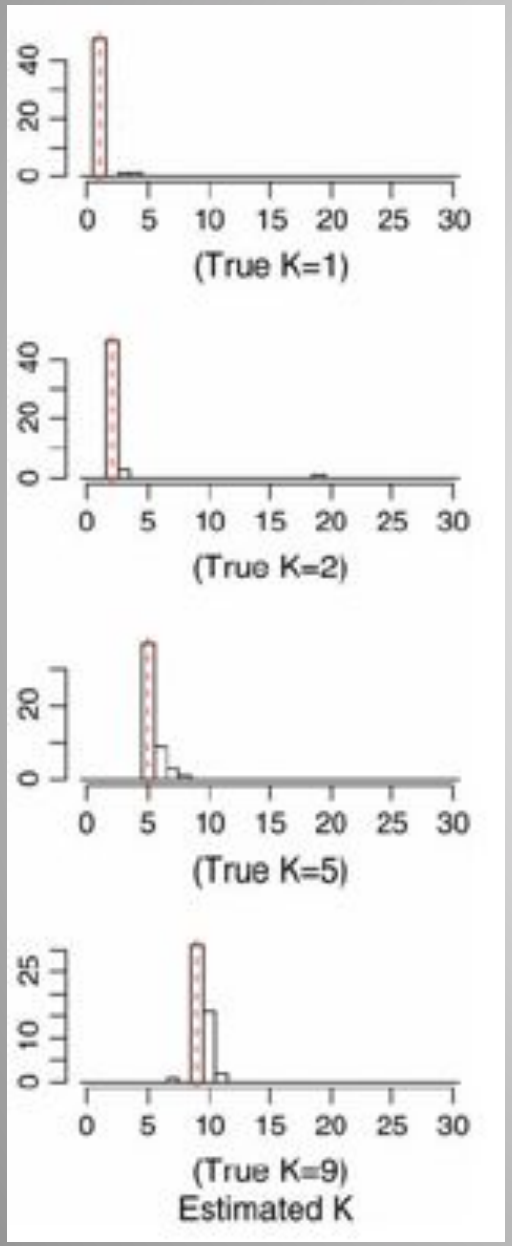
the principle of the method is very close to STRUCTURE method with additional parameters for the spatial arrangement of the different cluster

The main assumptions are :

- the colored Tessellation
- Hardy-Weinberg equilibrium in each cluster
- linkage equilibrium between loci in each cluster

Contrary to STRUCTURE, the MCMC algorithm implemented in GENELAND also include the parameter *K*, the number of clusters.

GENELAND : simulation test
Inference of K



Contrary to STRUCTURE, the MCMC algorithm implemented in GENELAND also include the parameter *K*, the number of clusters.

Simulation test of the inference of *K*

GENELAND : simulation test
Individual assignment

TABLE 1

Average false classification rates (in percentage) for all simulated data sets and subsamples with various levels of genetic and spatial structure

Structure		Spatial		Nonspatial	
Genetic	Spacial	F-model	D-model	F-model	D-model
		Results with 10 loci			
All	All	1.8	2.6	3.8	3.3
$F_{ST} < 0.04$	All	7.8	14.2	15	13.5
$F_{ST} < 0.06$	All	4.7	7.6	9	8.5
$F_{ST} > 0.11$	All	0.3	0.3	0.2	0.2
All	$m < 12$	2.3	1.9	11.4	6
All	$m < 25$	1.7	1.8	6.8	4.4
All	$m > 80$	2.2	3	2.8	3
$F_{ST} < 0.06$	$m < 25$	2.7	5.3	11.8	9.5
$F_{ST} < 0.04$	$m < 12$	3.5	1	24	16.7
		Results with 3 loci			
All	All	11.3	12.5	17.5	17.5

The level of genetic and spatial structure increases with F_{ST} and decreases with m, respectively. Results are shown from 1000 simulated data sets of 100 individuals in two populations, with $L = J_{l=1,\ldots L} = 10$ and $L = 3$, $J_{l=1,\ldots L} = 10$.

Geneland

Structure

GENELAND makes less assignment errors than STRUCTURE, especially when there is a strong spatial structure (small *m*) and a weak differentiation (low F_{ST})

GENELAND : simulation test spatial cluster delimitation

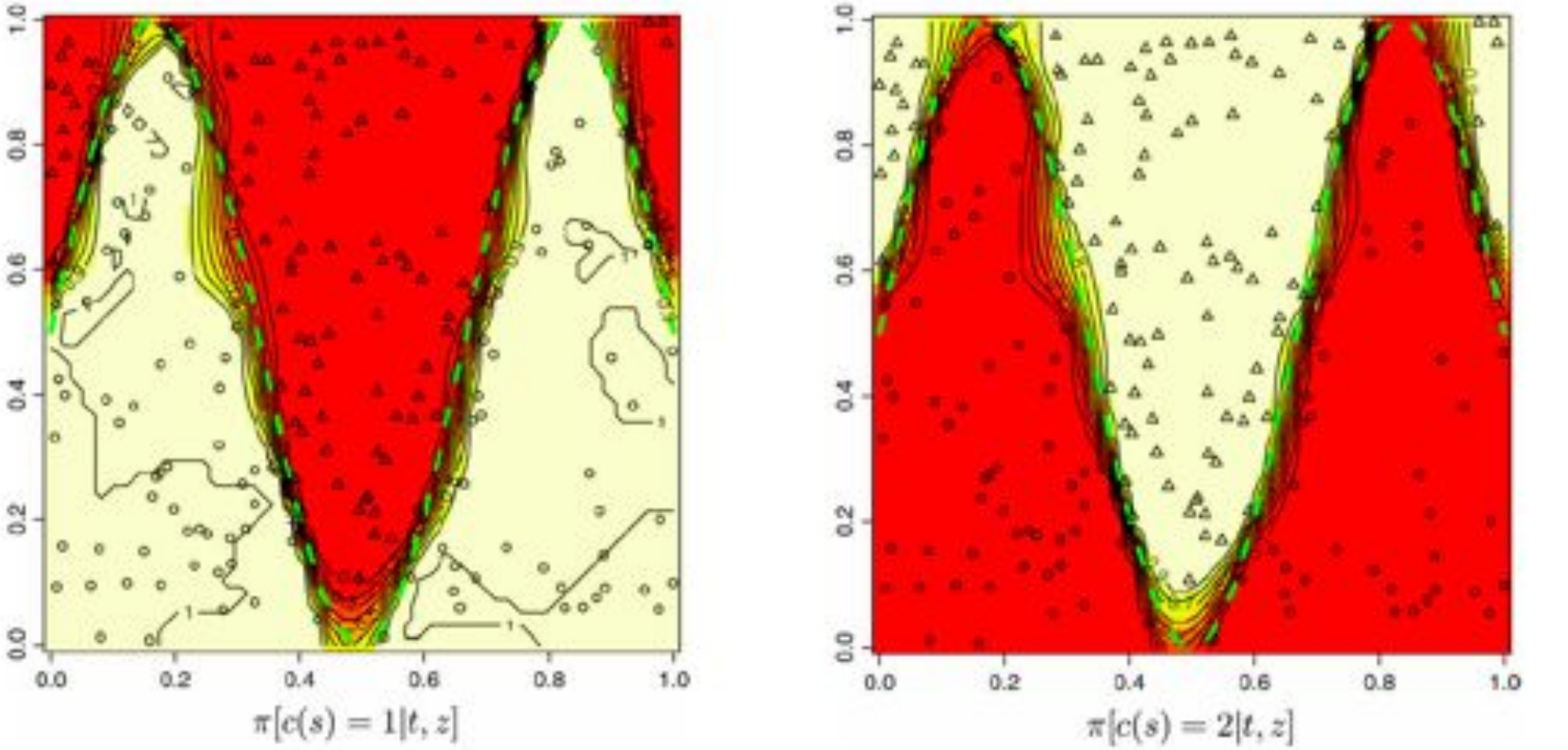


FIGURE 7.—Maps of posterior probabilities, simulated data set A. The dashed green line depicts the true sine-shaped line of discontinuity. $F_{ST} = 0.16$, $L = J_{t=1,\ldots,L} = 10$.

Very good spatial delimitation of genetic clusters with F_{ST}=0.16

GENELAND : simulation test spatial cluster delimitation

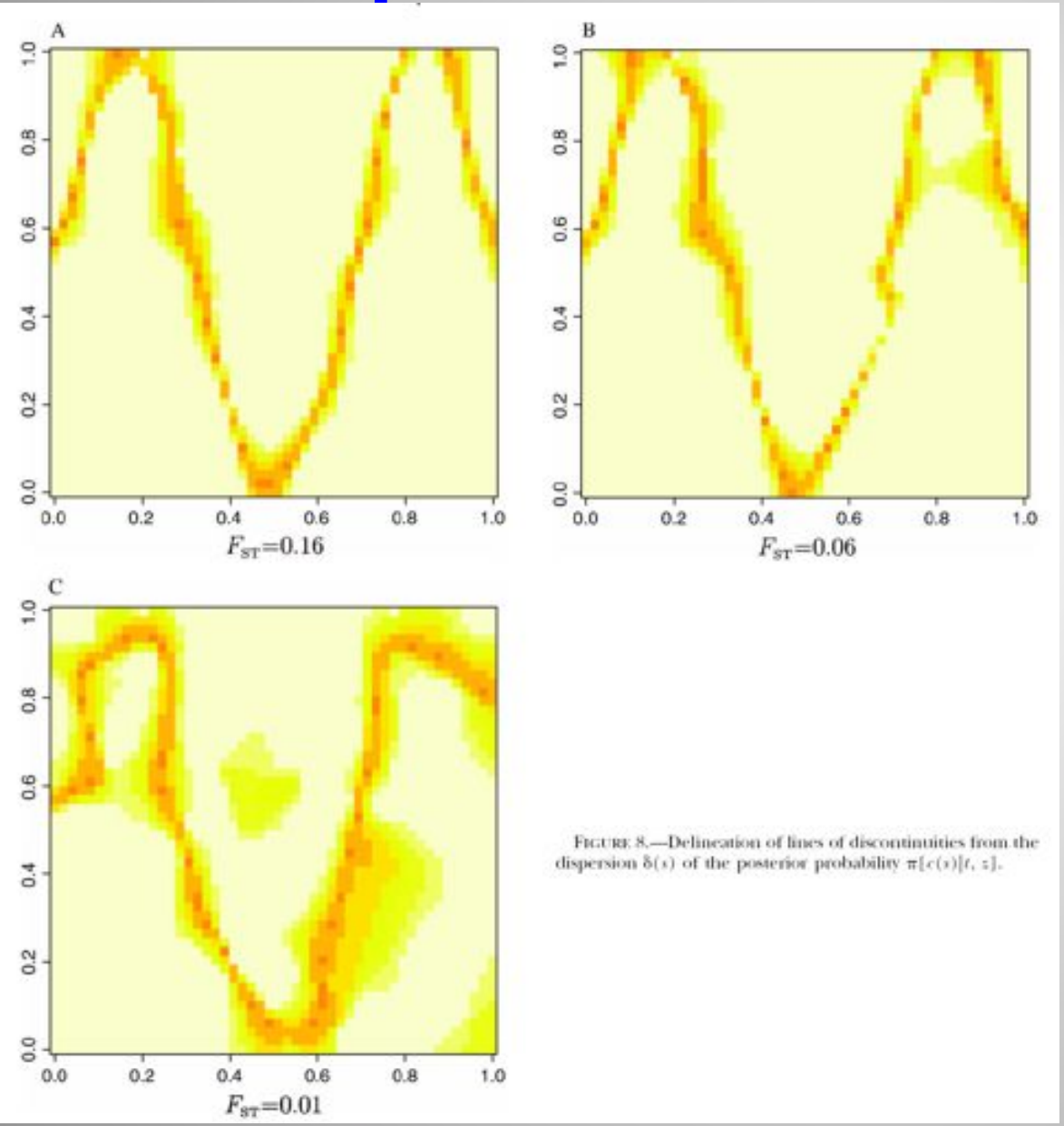


FIGURE 8.—Delineation of lines of discontinuities from the dispersion $\delta(s)$ of the posterior probability $\pi[c(s)|t, z]$.

less and less precision when genetic differentiation decreases

GENELAND : simulation test immigrant detection

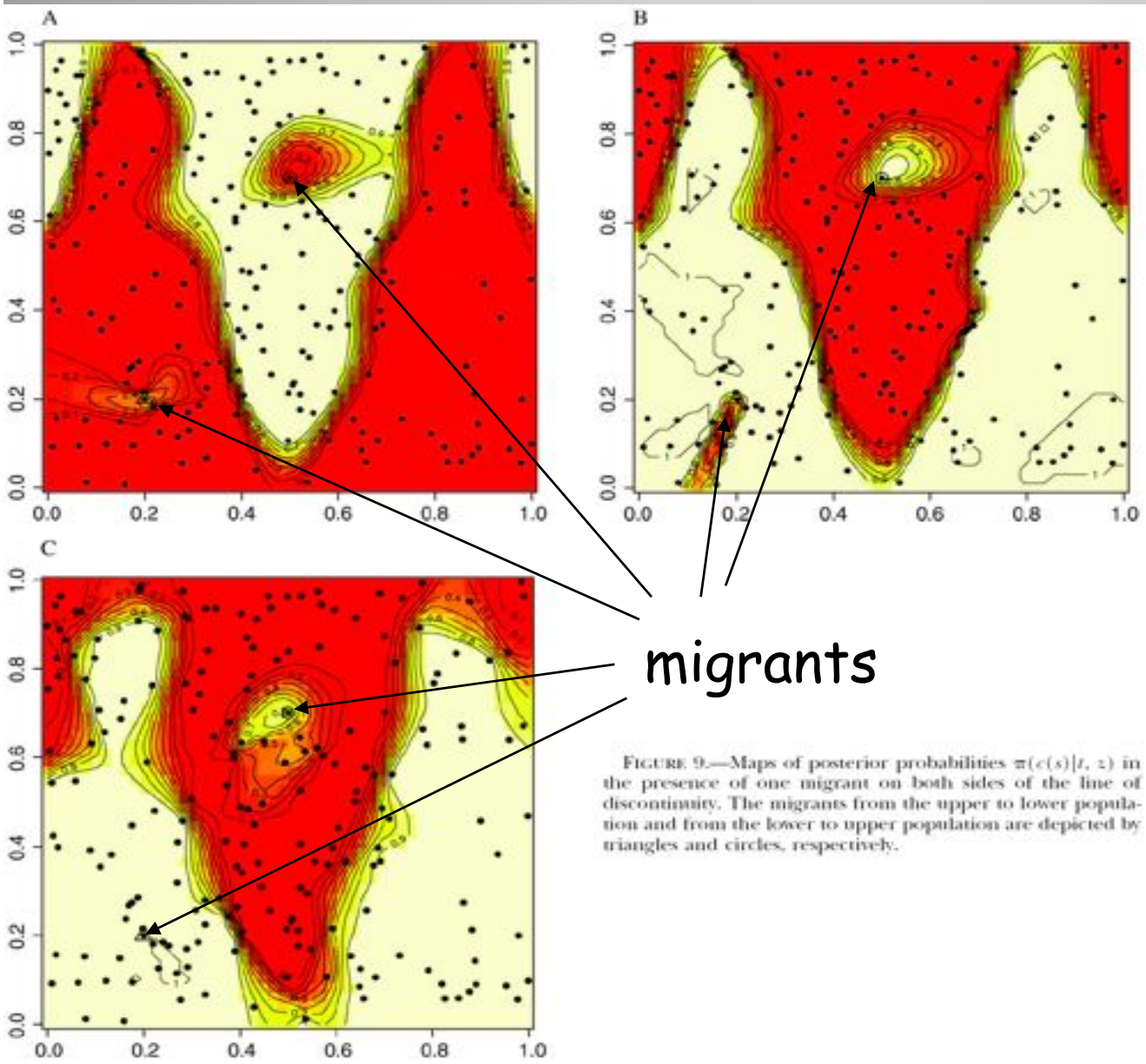


FIGURE 9.—Maps of posterior probabilities $\pi(c(s)|t, z)$ in the presence of one migrant on both sides of the line of discontinuity. The migrants from the upper to lower population and from the lower to upper population are depicted by triangles and circles, respectively.

good detection

Migrants do not strongly affect the spatial delimitation of the clusters

Migrants are more easily detected if they are isolated rather than surrounded by residents (especially for small *m*)

GENELAND : test on a real data set

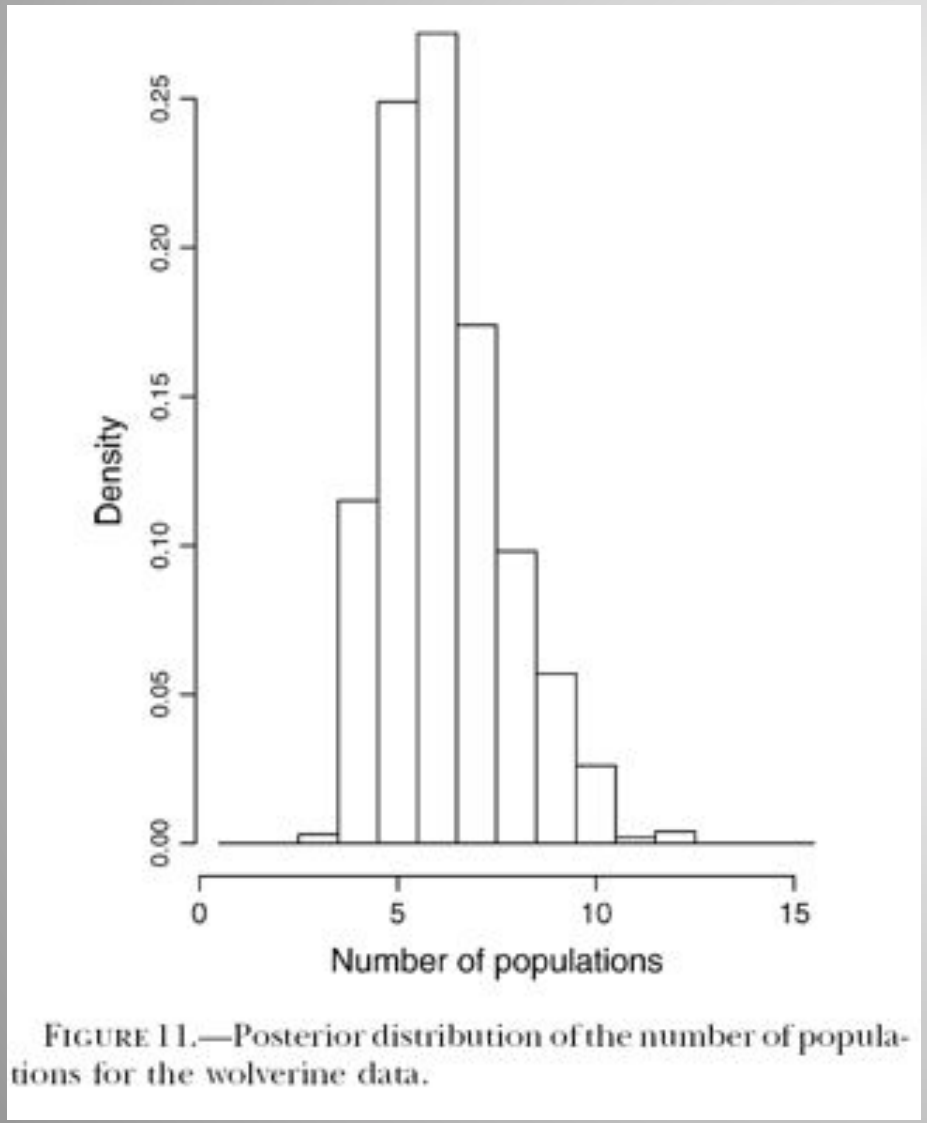
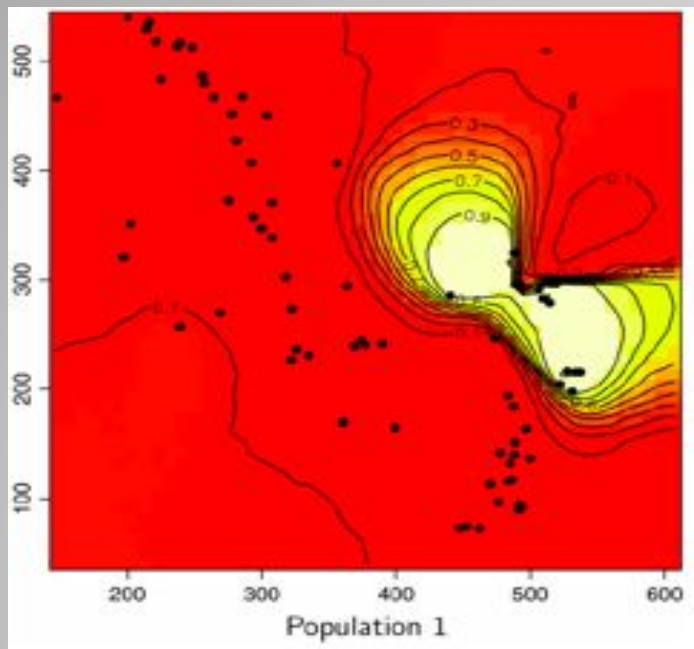
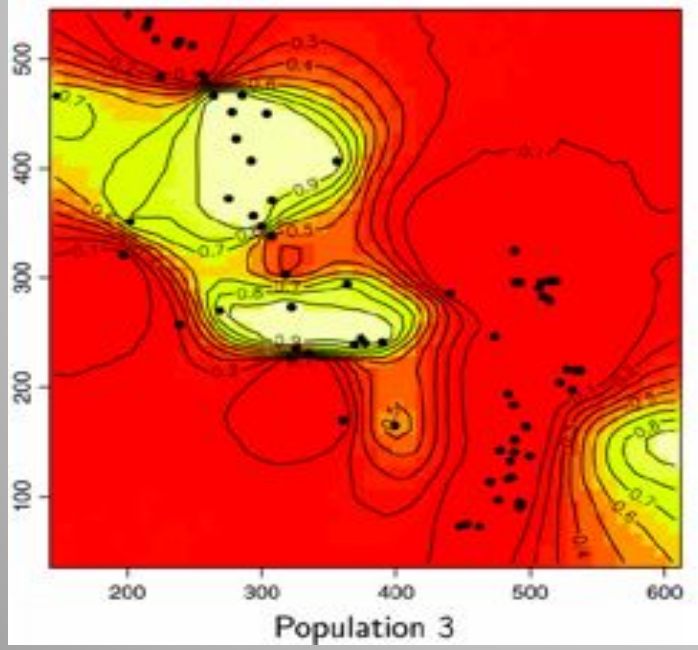
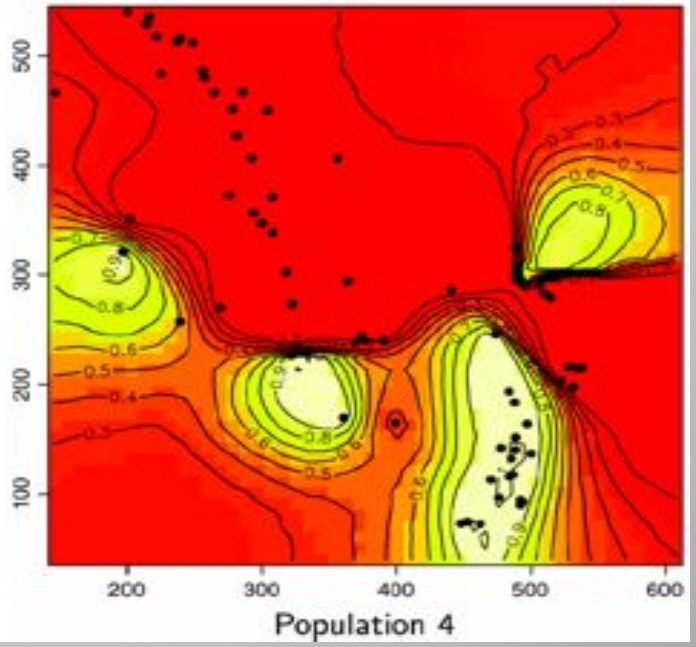


FIGURE 11.—Posterior distribution of the number of populations for the wolverine data.

GENELAND : test on a real data set







GENELAND : test on a real data set

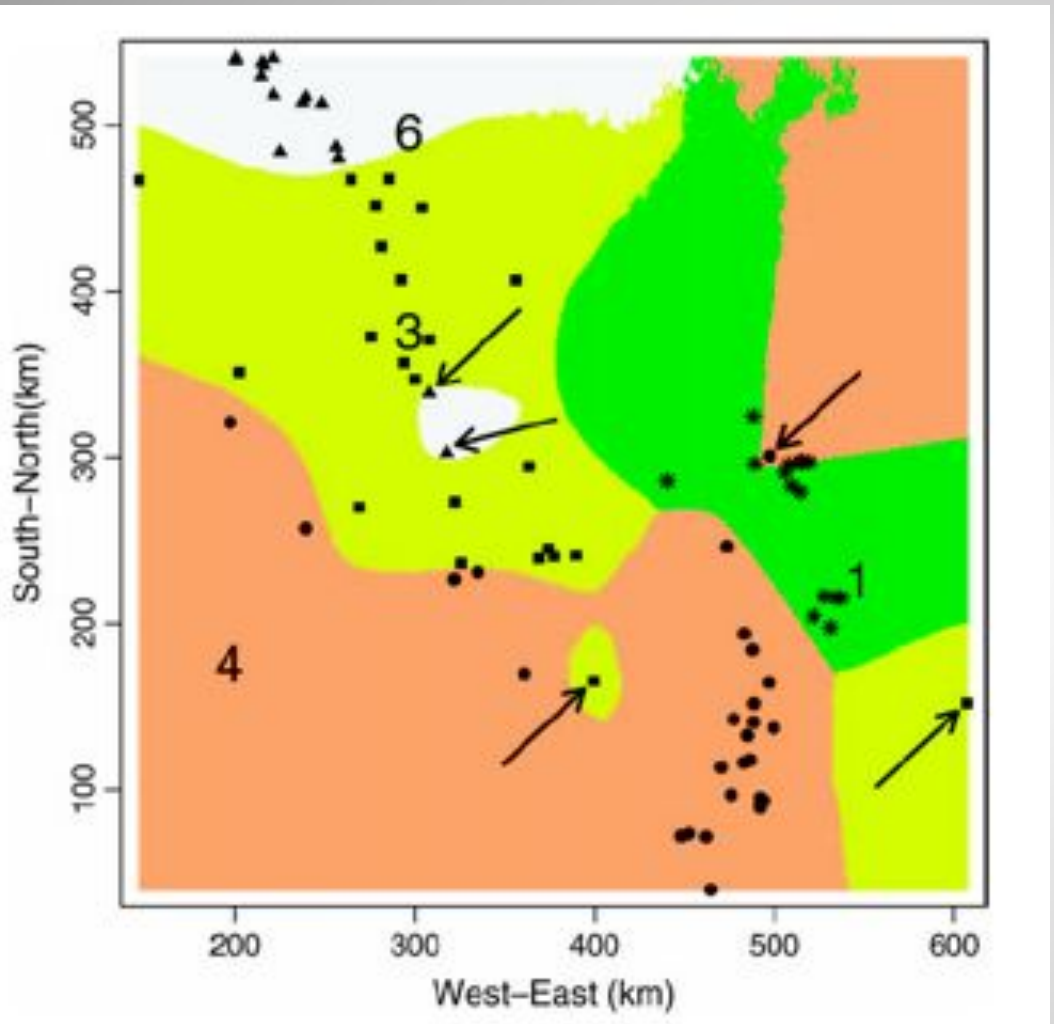


FIGURE 13.—Map of the mode of the posterior probability to belong to each class for the wolverine data. Large character numbers indicate population labels. Arrows indicate putative migrants.

spatial delimitation of 6 genetic clusters
detection of 5 migrants

GENELAND : test on a real data set

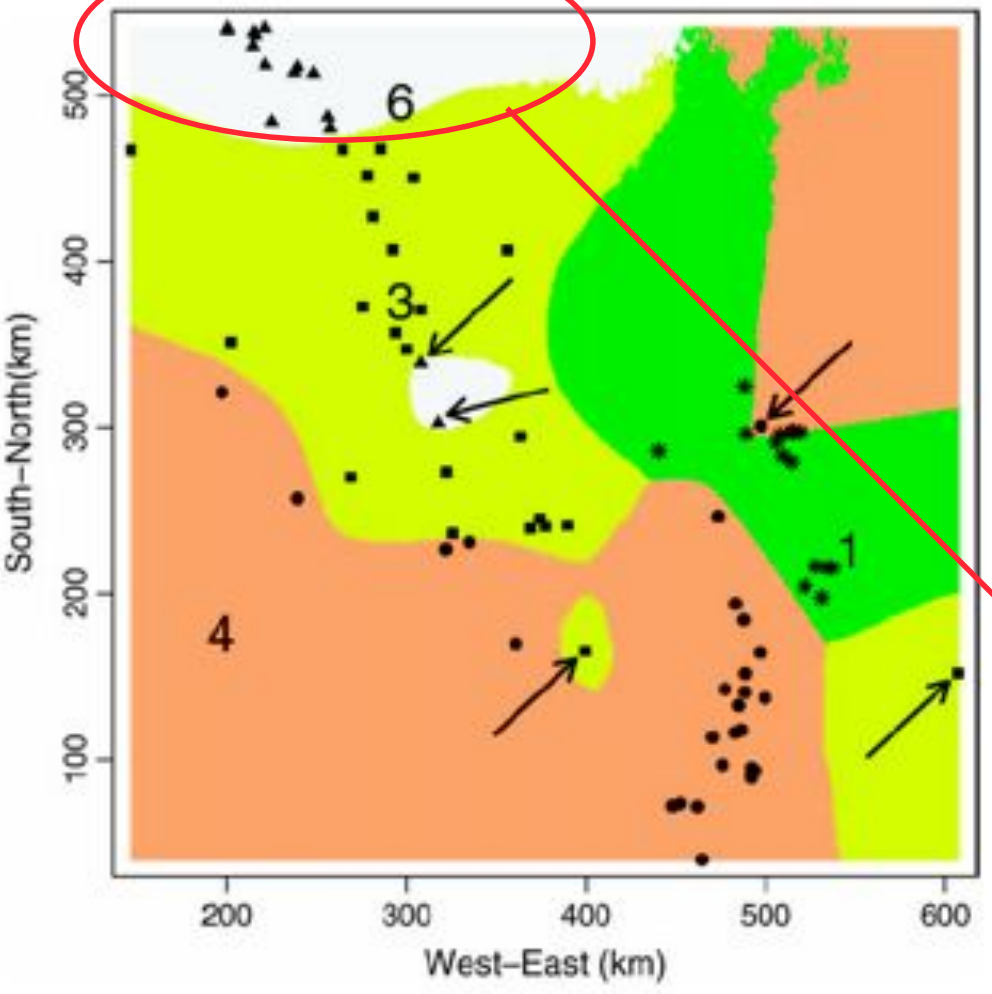


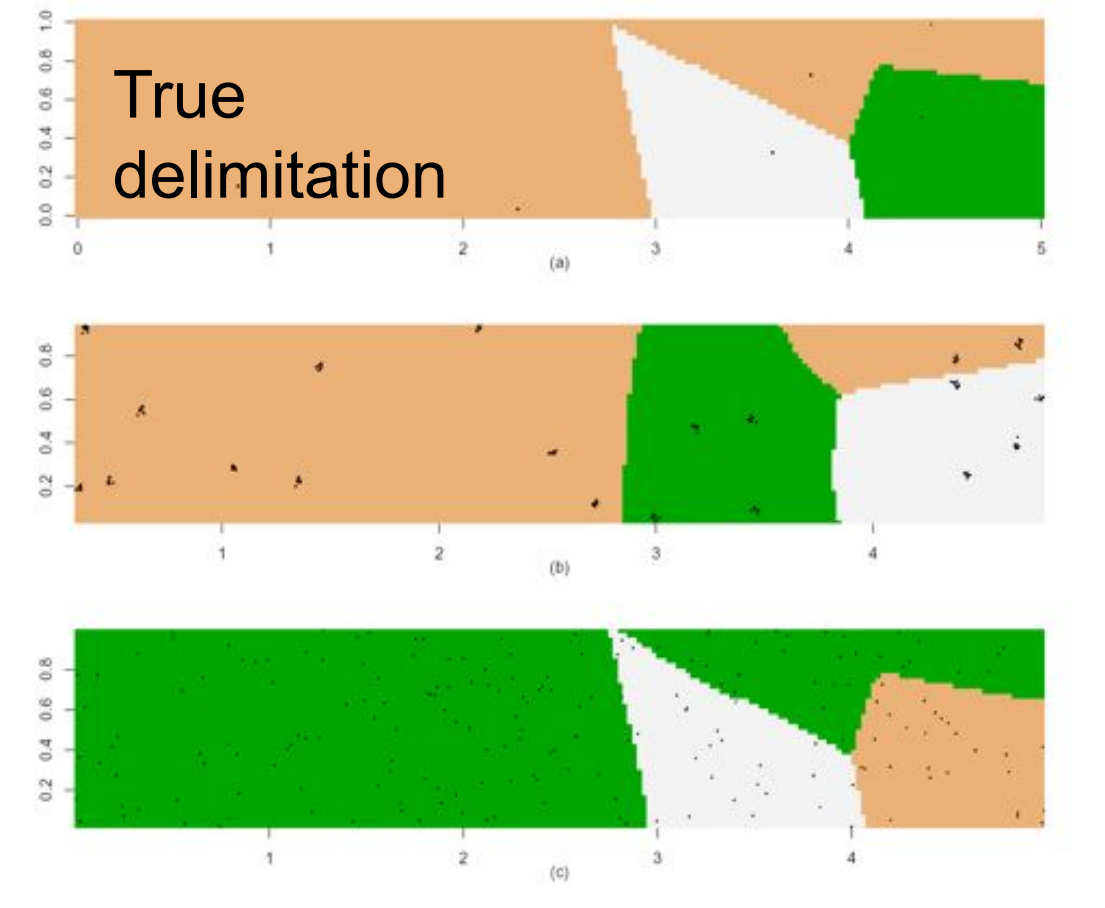
FIGURE 13.—Map of the mode of the posterior probability to belong to each class for the wolverine data. Large character numbers indicate population labels. Arrows indicate putative migrants.

This cluster was not detected with other methods : GENECLASS, STRUCTURE

Better performance or bias of the spatial method?

GENELAND : simulation tests of potential problems

What happens when samples are aggregated in space ?



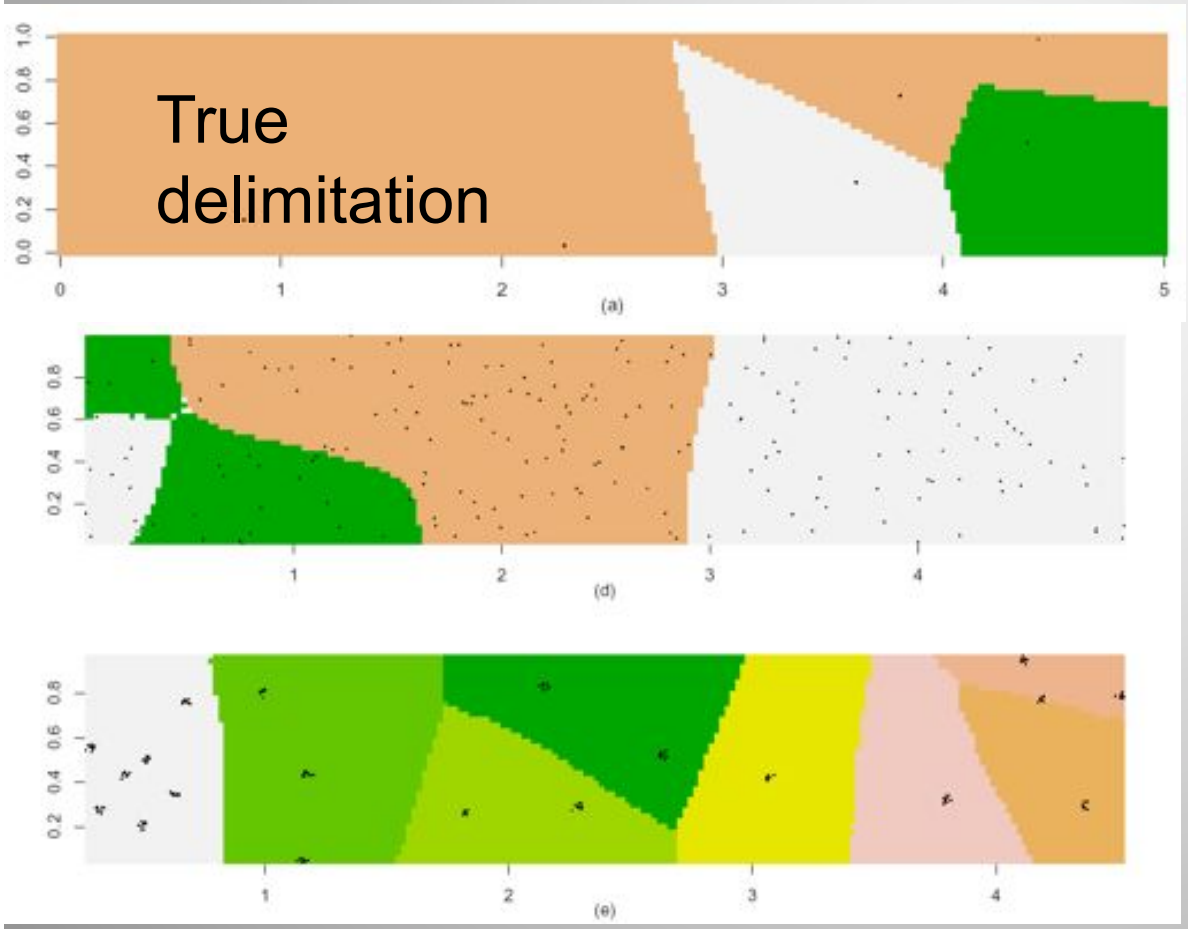
Results are intuitive:

Spatial cluster delimitation is precise when there are sampled individuals around them.

➠ better to sample homogeneously around the potential barriers

GENELAND : simulation tests of potential problems

What happens when there is Isolation By Distance ?

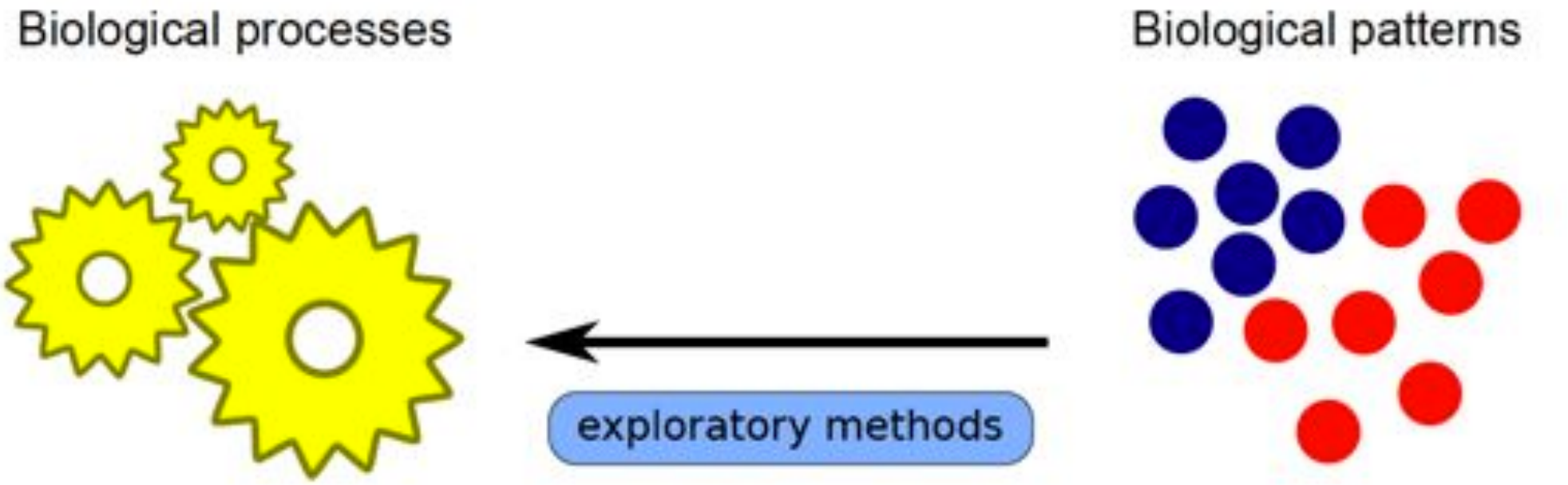


Results are also intuitive:

Spatial cluster delimitation is not working for strong IBD and is worth when samples are aggregated

➠ need for a new version designed for IBD

quick example of an exploratory method : PCA

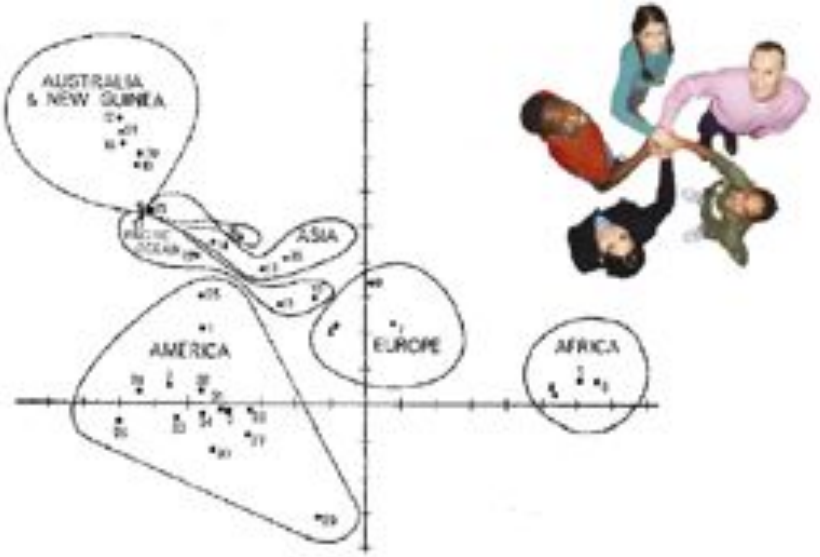


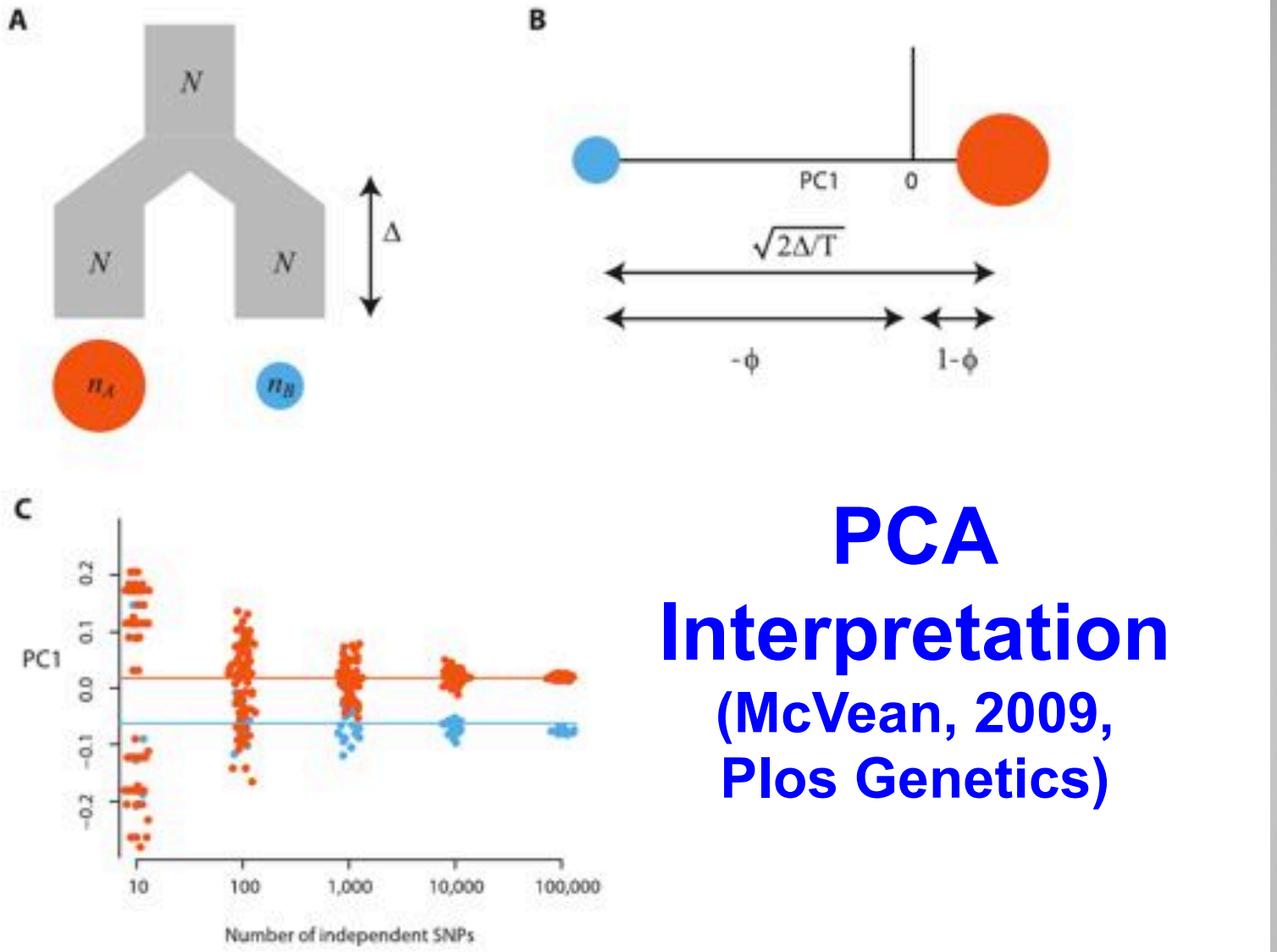
quick example of an exploratory method : PCA

Analyse en Composantes Principales

Principe Général

- Réduction de la dimension des jeux de données en préservant le plus de variabilité possible
- Premières applications en génétique des populations par Cavalli-Sforza (1966)





PCA Interpretation

(McVean, 2009, Plos Genetics)

Figure 2. Principal component analysis of two populations. (A) Consider a sample of n_A individuals from population A (indicated by the red circle) and n_B from population B (indicated by the blue circle), where the two populations have the same effective population size of N and are both derived from a single ancestral population, also of size N, with the split happening a time Δ in the past. (B) The expected locations of these two sets of samples on the first PC is defined by the time since divergence (the Euclidean distance between the samples is $\sqrt{2\Delta/T}$) (see text for definitions) and the relative sample size from the populations, with the larger sample lying closer to the origin. Defining $\phi = n_A/(n_A + n_B)$, the relative location of the two populations on the first PC are $1-\phi$ for samples from population A and $-\phi$ for samples from population B (note that the sign is arbitrary). (C) To investigate the effect of finite genome size simulations were carried out for the model shown in part A with 80 genomes sampled from population A, 20 from population B and a split time of 0.02 N_e generations ($F_{ST} = 0.01$) and between 10 and 10^5 SNPs. Lines indicate the analytical expectation. A jitter has been added to the x-axis for clarity. Note that the separation of samples with 10 SNPs does not correlate with population and simply reflects random clustering arising from the small numbers of SNPs.
doi:10.1371/journal.pgen.1000686.g002

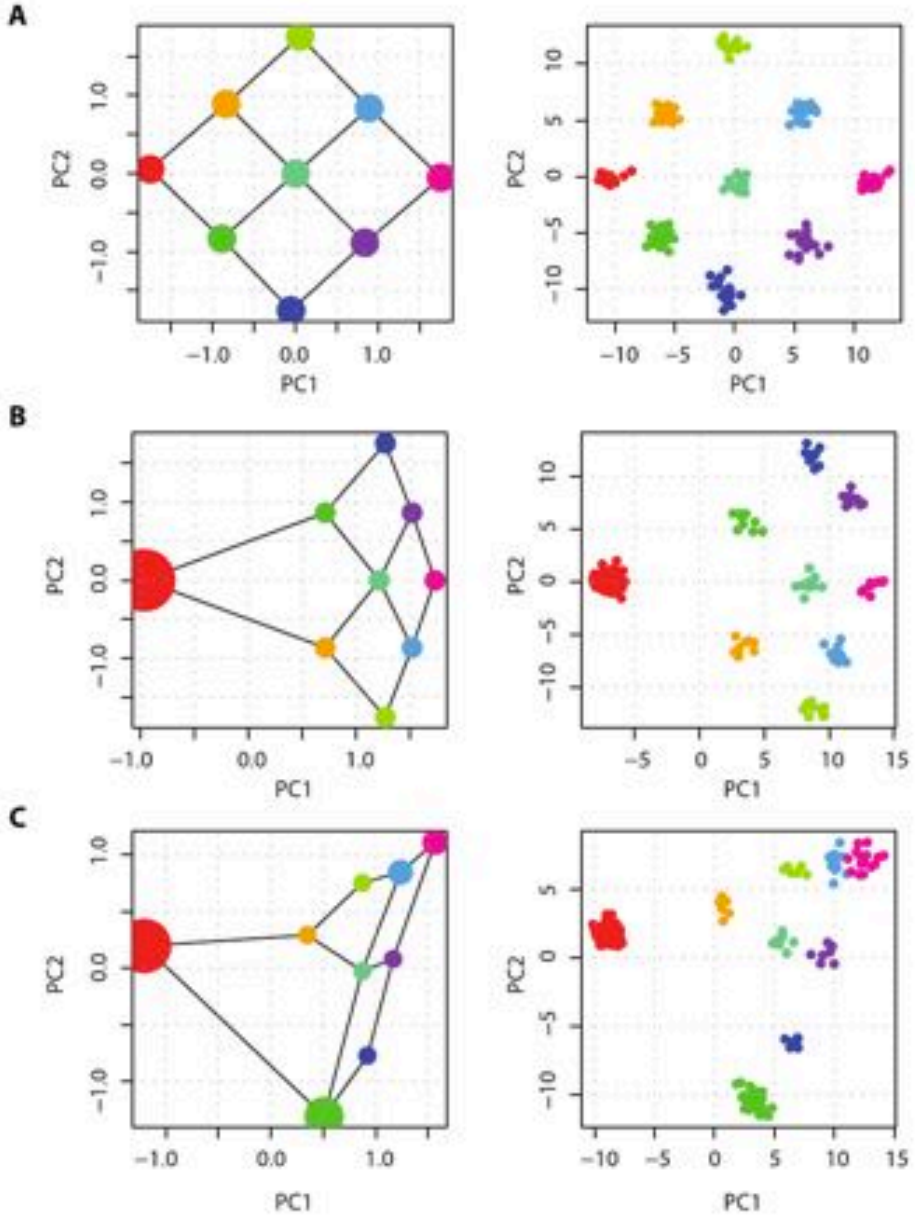
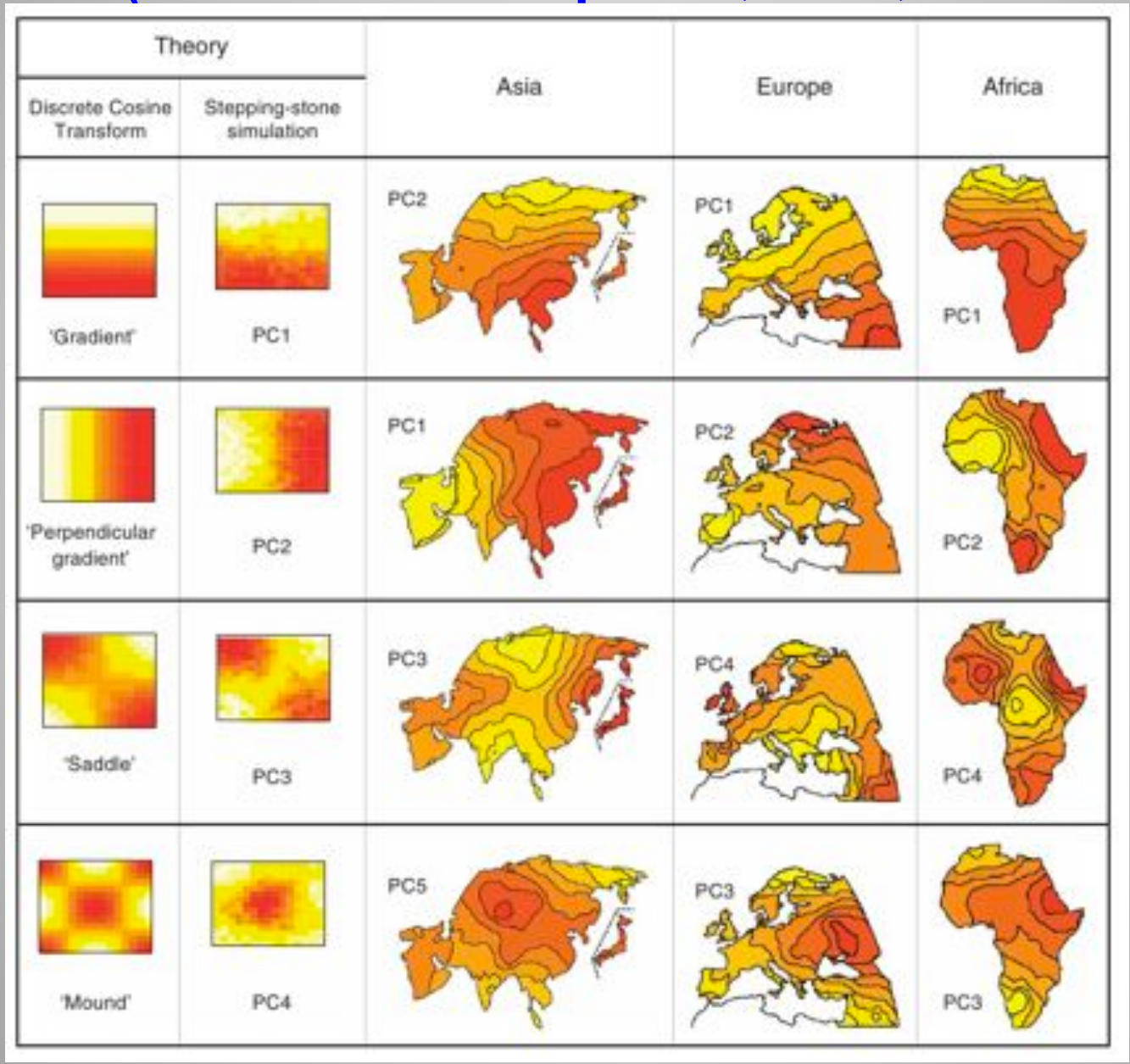


Figure 3. The effect of uneven sampling on PCA projection. PCA projection of samples taken from a set of nine populations arranged in a lattice, each of which exchanges migrants at rate M per N_e generations with each adjoining neighbour, leads to a recovery of the migration-space if samples are of equal size (A), or a distortion of migration-space if populations are not equally represented (B,C). In each part the left-hand panel shows the analytical solution (the area of each point represents the relative sample size) with migration routes illustrated while the right-hand panel shows the result of a simulation with a total sample size of 180 and 10,000 independent SNP loci. All examples are for $M = 2$.
doi:10.1371/journal.pgen.1000686.g003

PCA Interpretation (McVean, 2009, Plos Genetics)

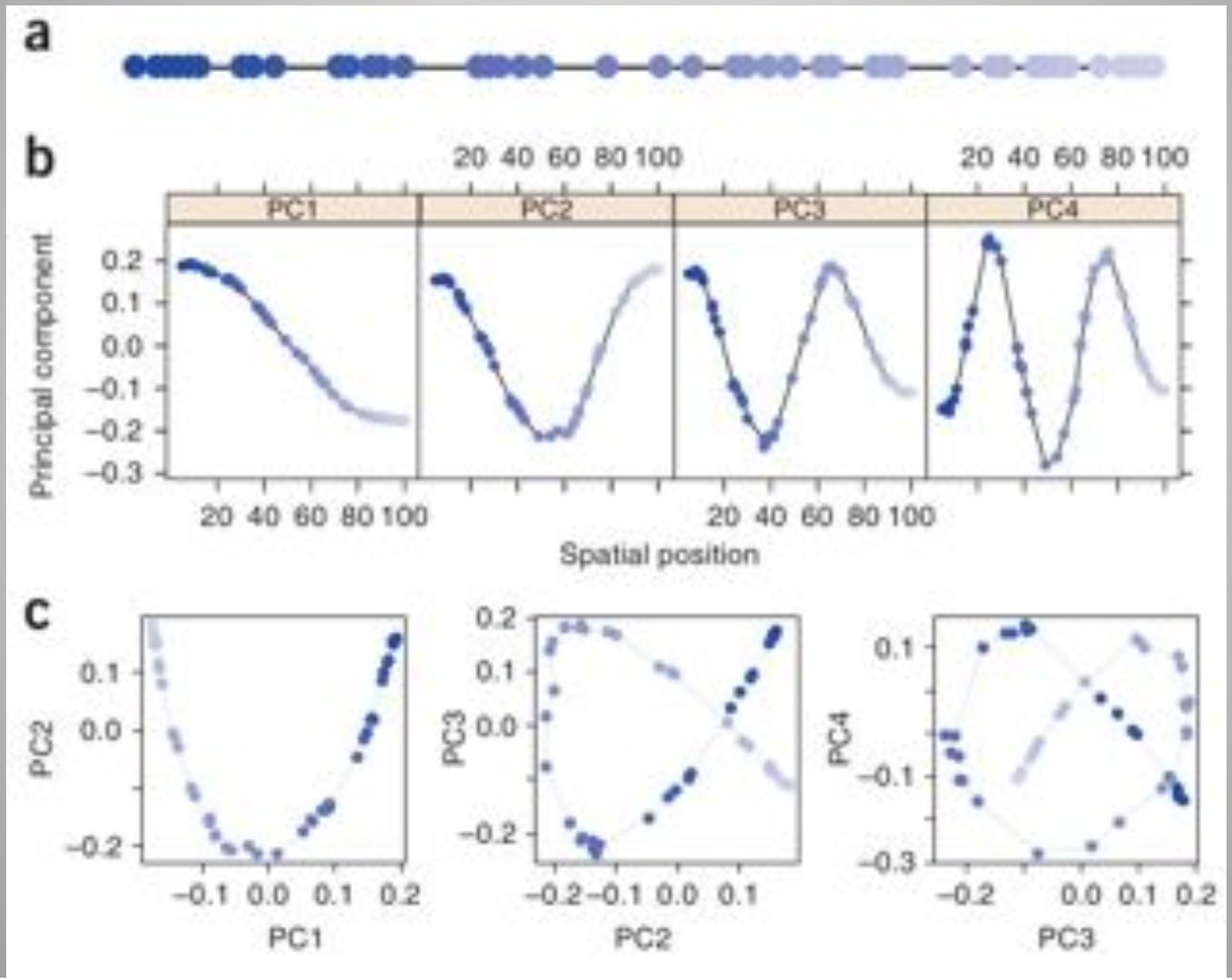
PCA Interpretation

(Novembre & Stephens, 2008, Nat Genet)



PCA Interpretation

(Novembre & Stephens, 2008, Nat Genet)



PCA : conclusion

One of the numerous methods used to highlight genetic structure

Advantages of PCA analyses :

- Well known statistical properties
- Very efficient for large data sets
- Numerous extensions (e.g. sPCA, DAPC : Jombart & Co)

Conclusions: Assignment, Clustering and PCA

Limits of above approaches :

- **Assignation**: some prior information about the populations is needed (equivalent to supervised clustering)
- **PCA and Unsupervised Clustering**: only representation of the genetic diversity (that might be strongly affected by the sample characteristics)

These approaches do not provide information about the (historical) events that resulted in the observed genetic structure.
They might at best help in defining compteing demographic scenarios

Conclusions: Assignment, Clustering and PCA

Limits of above approaches :

- **Assignation**: prior information needed
- **PCA and Unsupervised Clustering**: only representative

Do not provide information about the (historical) events that resulted in the observed genetic structure

e.g.: when using STRUCTURE, extreme cautions is needed when interpreting clusters as ancestral populations

➔ different demographic scenarios might result in same PCA (or Unsup. Clustering) results

Conclusions: Assignment, Clustering and PCA

Do not provide information about the (historical) events that resulted in the observed genetic structure

Other inference methods are needed to infer demographic history:

- Estimating parameters and comparing scenarios:
  - ✓Likelihood based inferences (cf. Raphael Leblois)
  - ✓ABC approach, e.g. DIY-ABC (cf. Arnaud Estoup)

- Estimating (or comparing) trees: e.g. Phylip (Felsesntein & co), Treemix (Pickrell, Pritchard, 2012), kim_tree (Gautier, Vitalis, 2012)

Pour le TD de Lundi…..

Lire rapidement le document

Lire "R pour les débutants" Emmanuel Paradis
(http://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf)

on ne présentera pas le logiciel STRUCTURE et son interface mais n'hésitez pas a poser des questions sur son utilisation si vous en avez…