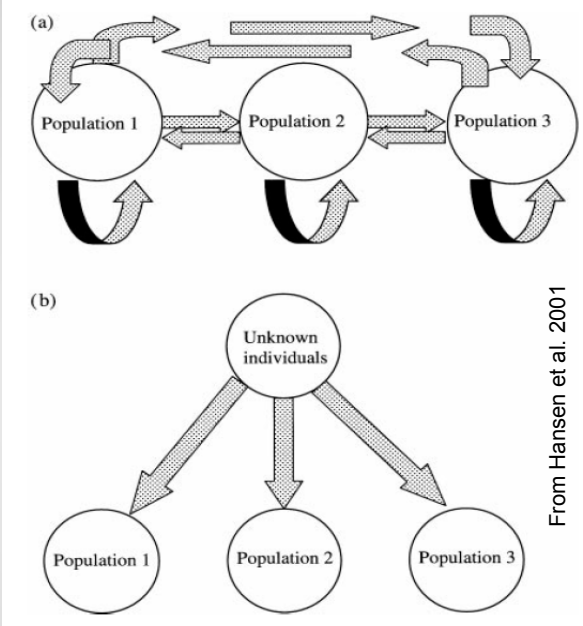
Assignment and clustering algorithms for individual multilocus genotypes



Raphael Leblois
Centre de Biologie et de Gestion des Populations , CBGP
INRA, Montpellier

Master MEME, March 2011

Assignment and Clustering from individual multilocus genotypes

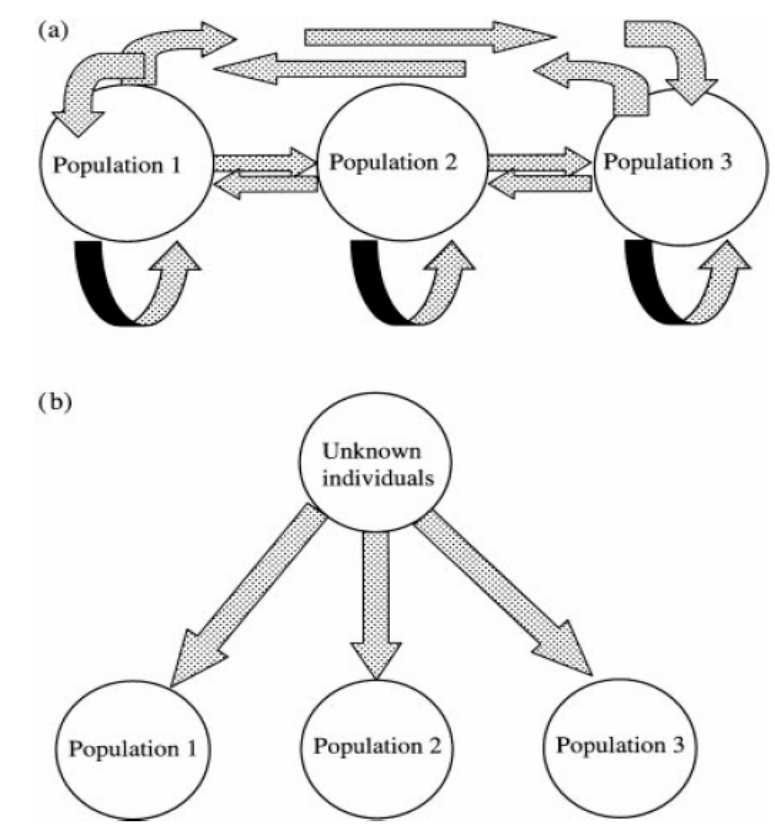
1. Introduction to genetic assignment and clustering methods
2. Few assignment algorithms
3. Inference of migration rates using assignment methods
4. Non-spatialized clustering : STRUCTURE
5. Spatialized clustering : GENELAND

Biological questions

- What the geographic origin or the population of origin of a focal individual
- Population delimitation
- Migrant detection / inference of recent migration rates
- Analysis of genetic introgression / hybridization

Classification vs. Clustering

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, Cornuet et al. 1999

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*

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

or $L \approx p_{ijk}^2$ 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

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 and $(\delta\mu)^2$ 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

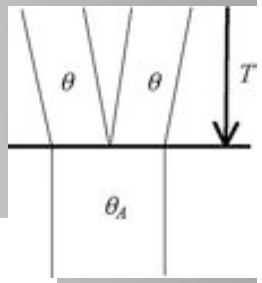
2. Compute the assignment score for each of those simulated genotypes ➠ "null" distribution

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

The exclusion method of Cornuet et al. 1999

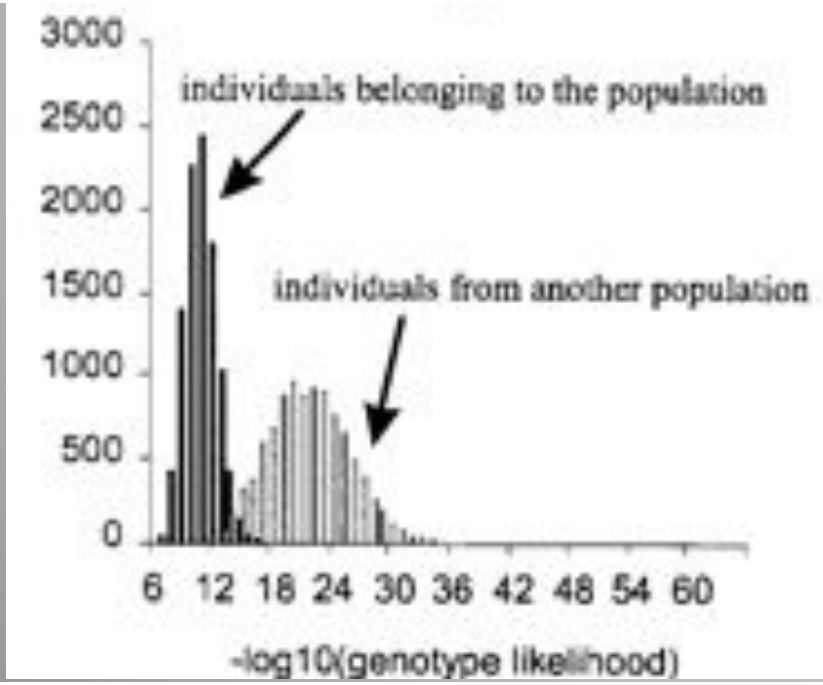
Principle: Simulation of the null distribution of the assignment score for membership in a population

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

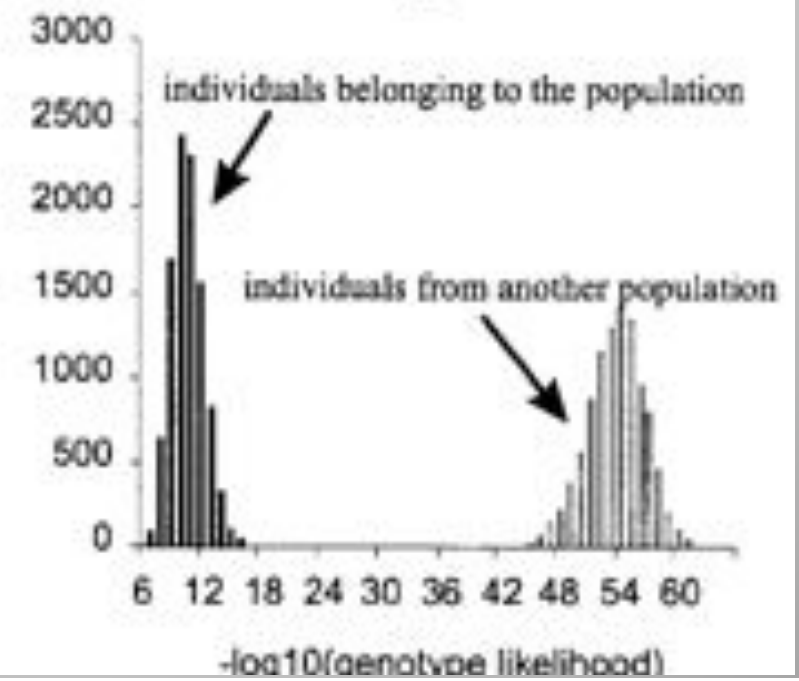


simulation test in 2 diverging populations :

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

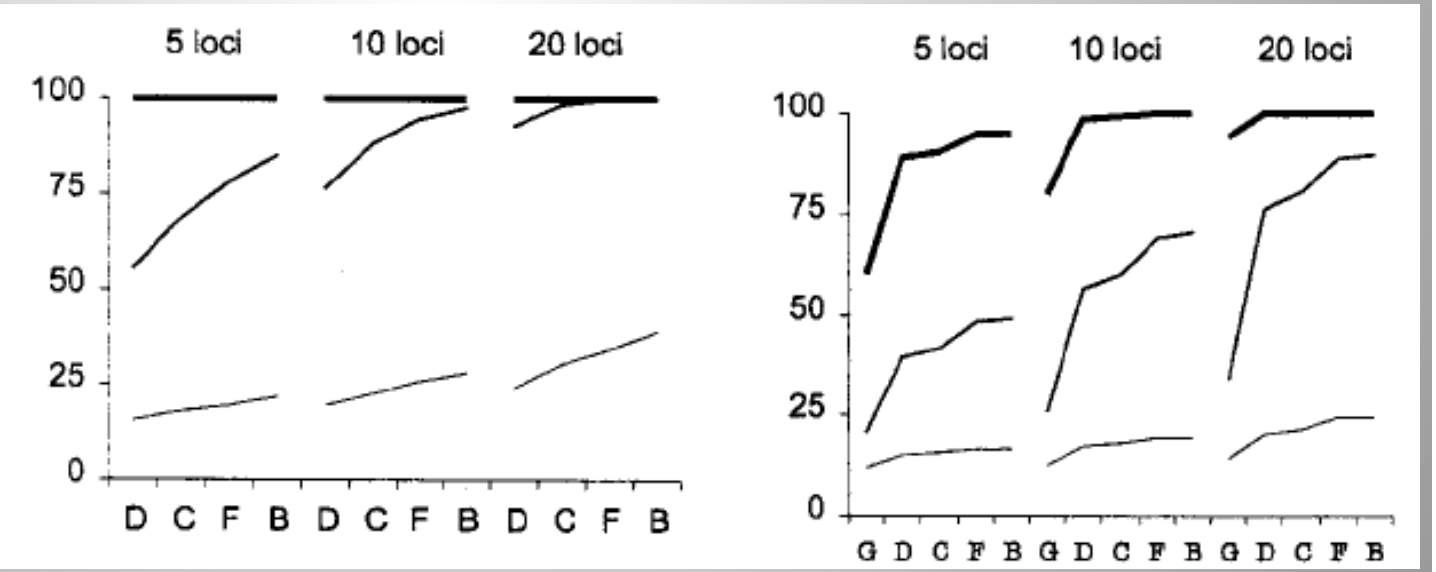
Paetkau et al. 1995 (F), Rannala & Mountain 1997 (B, highly similar to F),

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

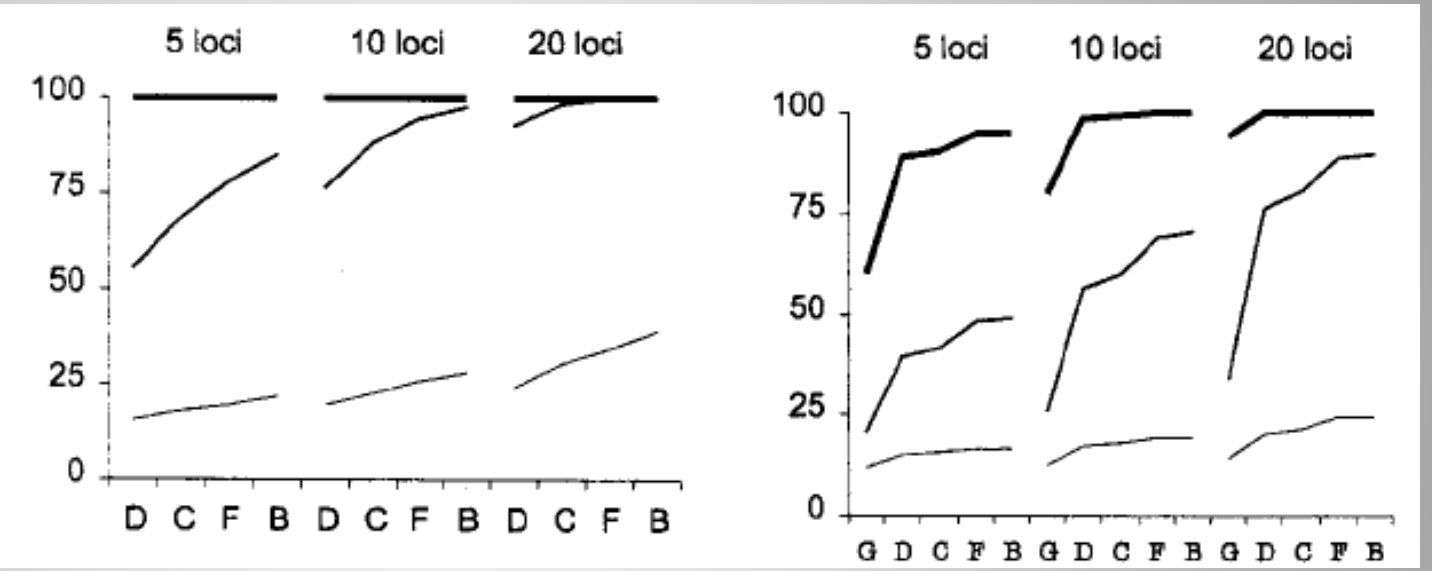


IAM SMM

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



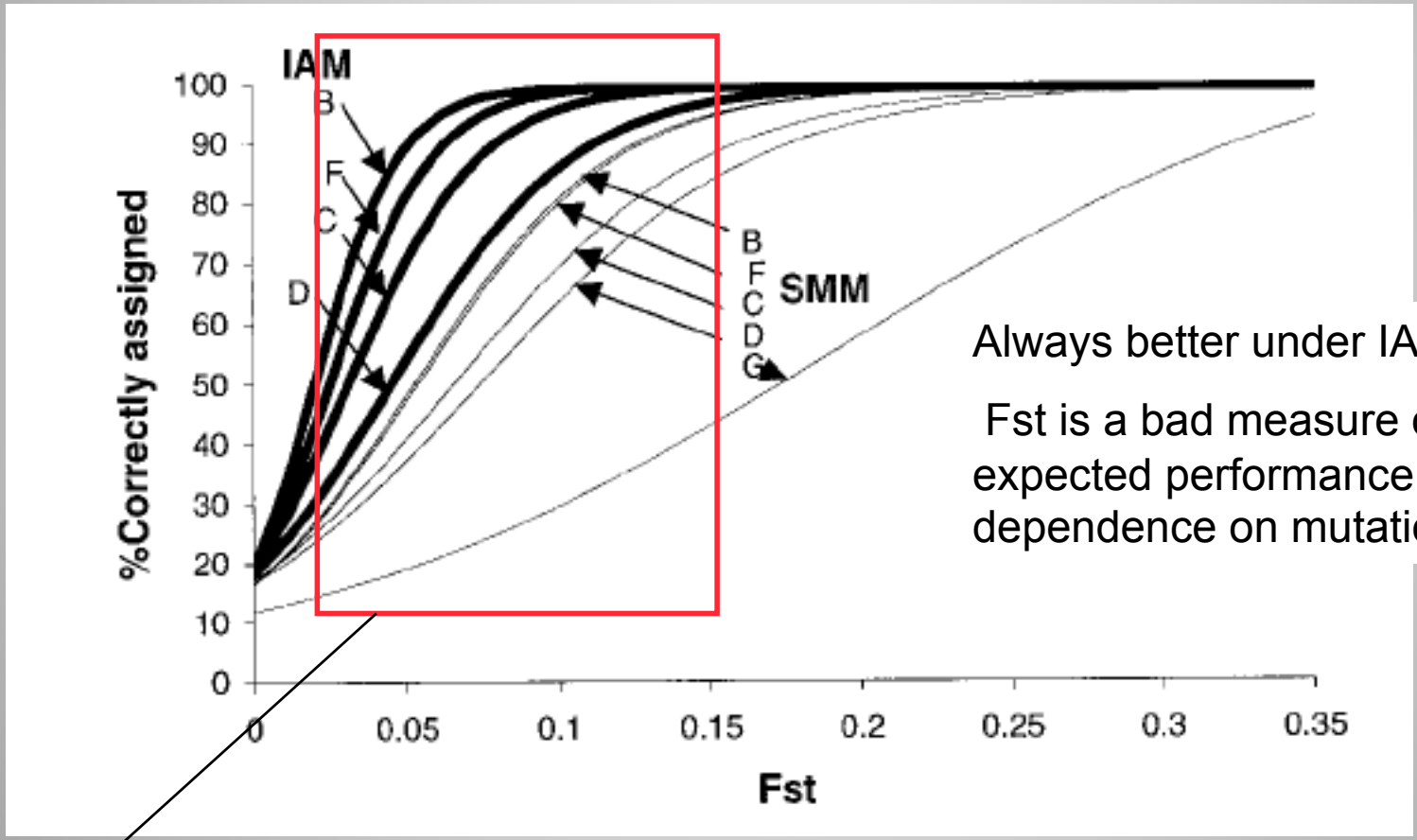
IAM SMM

- 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

Comparison of different algorithms (Cornuet et al. 1999)

Differentiation (Fst, directly linked to divergence time)



Always better under IAM than SMM

Fst is a bad measure of the expected performance (strong dependence on mutation processes)

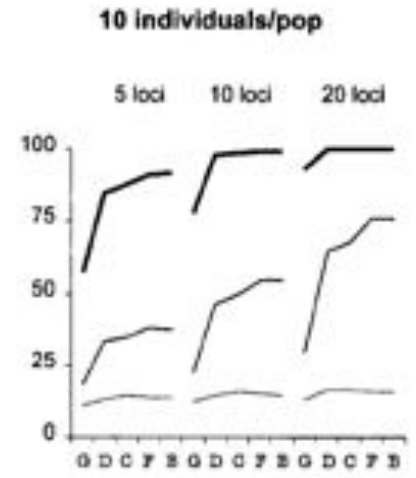
choice of the method is important

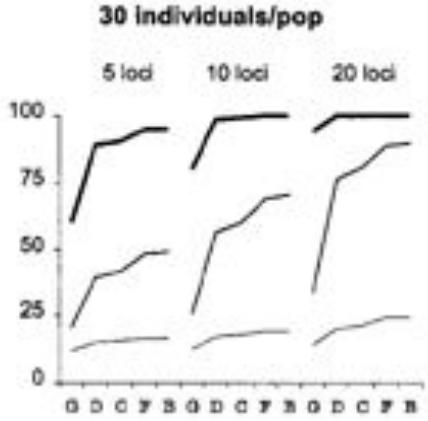
Comparison of different algorithms (Cornuet et al. 1999)

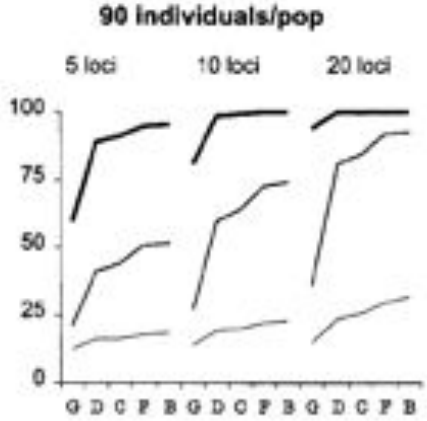
Sample size per population, locus number, differentiation

➠ weak influence of the sample size compared to the other factors

Td	Fst
2000	0.3
200	0.08
20	0.01







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

From individual assignments to the inference of migration rates

H_0 : the focal individual was born in the population where it has been sampled

Principle:

1. Compute one by one assignment scores for all individuals to their population of sampling, removing its genotype from the population

2. Compute the exclusion probability for all individuals to their sampling population

3. Detect as a immigrants all individuals for which the exclusion probability is larger than an arbitrary threshold α (e.g. 0.95)

From individual assignments to the inference of migration rates

H_0 : the focal individual was born in the population where it has been sampled

Principle:

1. Compute one by one assignment scores for all individuals to their population of sampling, removing its genotype from the population
2. Compute the exclusion probability for all individuals to their sampling population
3. Detect as a immigrants all individuals for which the exclusion probability is larger than an arbitrary threshold α (e.g. 0.95)

Paetkau et al. (2004) : Test of the same methods than in Cornuet et al. (1999) but for the detection of F0 migrants

From individual assignments to the inference of migration rates

the most important part is the exclusion probability computation :

➠ to know if an individual that is excluded from its sampling population is really a recent immigrant or if it is just mis-assigned by chances (i.e. its genotype is rare in the population)

type I error = probability of detecting a resident as a immigrant
Power = 1 - type II error = probability that an immigrant is detected as immigrant

From individual assignments to the inference of migration rates

Main limitation of Cornuet et al. (1999) exclusion approach is that the loci are considered as independent (no linkage disequilibrium) whereas an immigrant individuals corresponds to the migration of a complete haplotype

➠ strong linkage disequilibrium

Paetkau et al. (2004) designed a new exclusion algorithm by simulating multilocus genotype on the 10 last generations instead of independent loci

➠ Simulating gamete haplotypes from randomly chosen pairs of parents haplotypes

From individual assignments to the inference of migration rates

different possible exclusion criterion :

- the likelihood directly as in Cornuet et al. (1999)

  ➠ better when some population were not sampled (ghost pops)

- likelihood ratio L_home/L_max as in Paetkau et al. (2004)

  ➠ better when all populations were sampled

From individual assignments to the inference of migration rates

Simulation test in Paetkau et al. (2004) : test the resident/immigrant status of each individuals in an island model of migration

strong effect of the haplotypic vs. allelic simulation methods

Cornuet et al. 1999

Paetkau et al. 2004 is much better

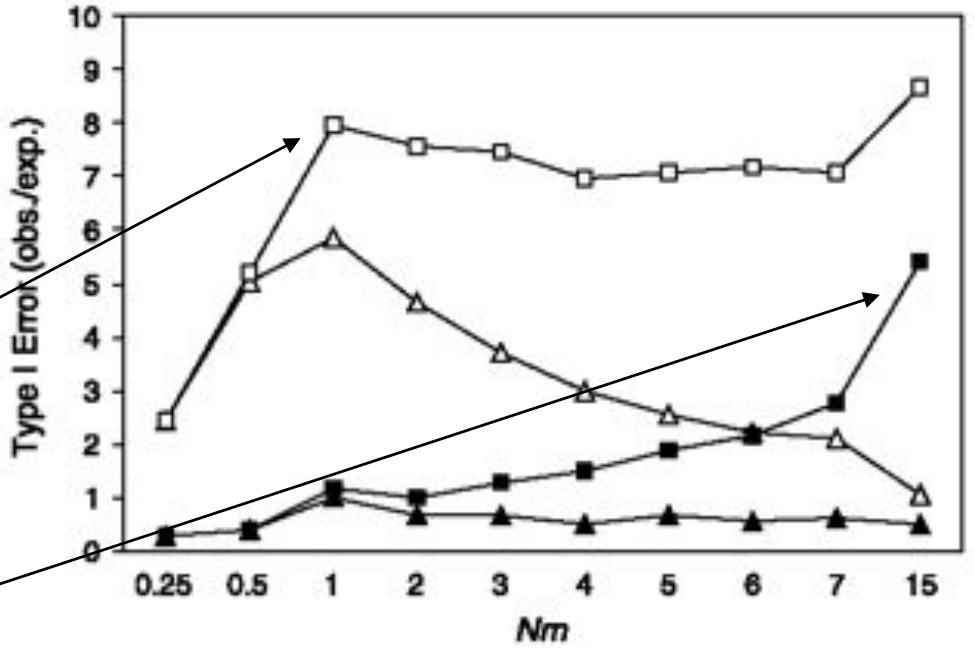


Fig. 1 Four different methods for drawing genotypes from the sample data to produce distributions of the test statistic A: drawing gametes (filled symbols) vs. alleles (open symbols) and analysing genotypes in sets of 10 times (squares) or 1 times (triangles) the size of the original data set. $N = s = 50$, $\mu = 0.005$, $l = 10$, $\alpha = 0.002$.

From individual assignments to the inference of migration rates

Simulation test in Paetkau et al. (2004) : effect of sample size in each population

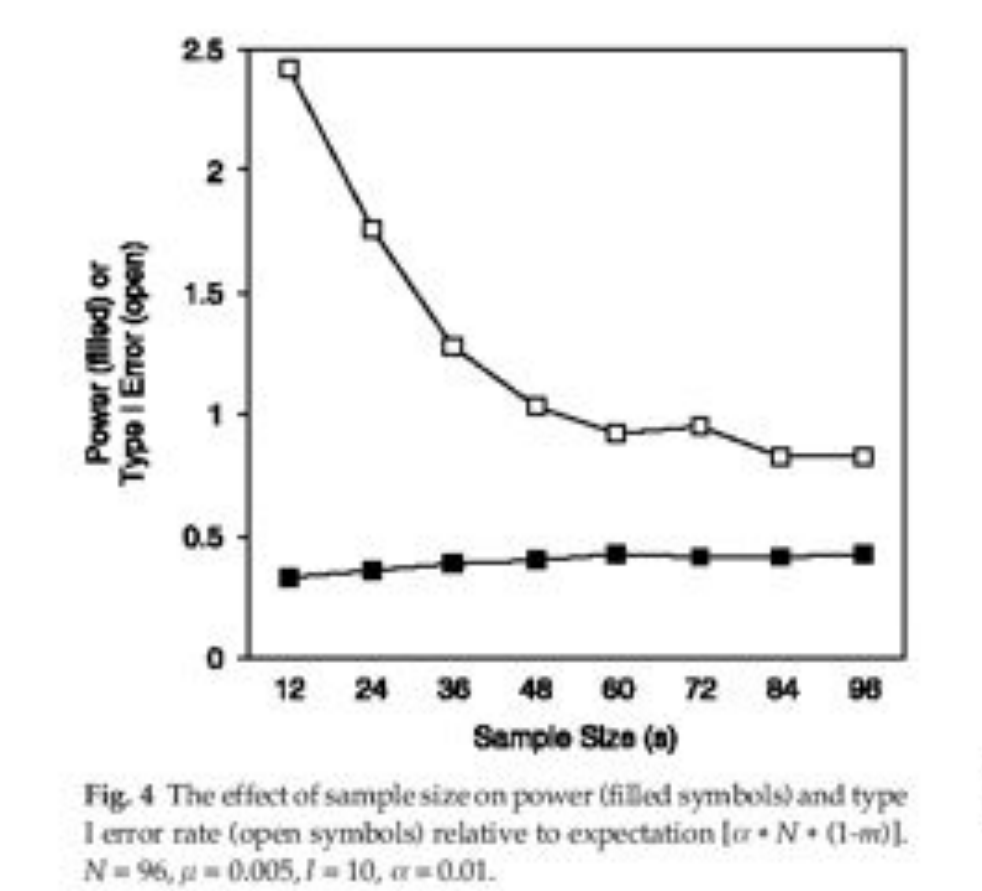


Fig. 4 The effect of sample size on power (filled symbols) and type I error rate (open symbols) relative to expectation [$\alpha * N * (1-m)$]. $N = 96$, $\mu = 0.005$, $I = 10$, $\alpha = 0.01$.

Strong effect of the sample size on the type I error, none on the power of the method

important because sample size usually do not have much effect when > 30 in population genetics

From individual assignments to the inference of migration rates

Simulation test in Paetkau et al. (2004) : how to predict the power of immigrant detection on a data set

D_{LR} = mean genotype likelihood ratio

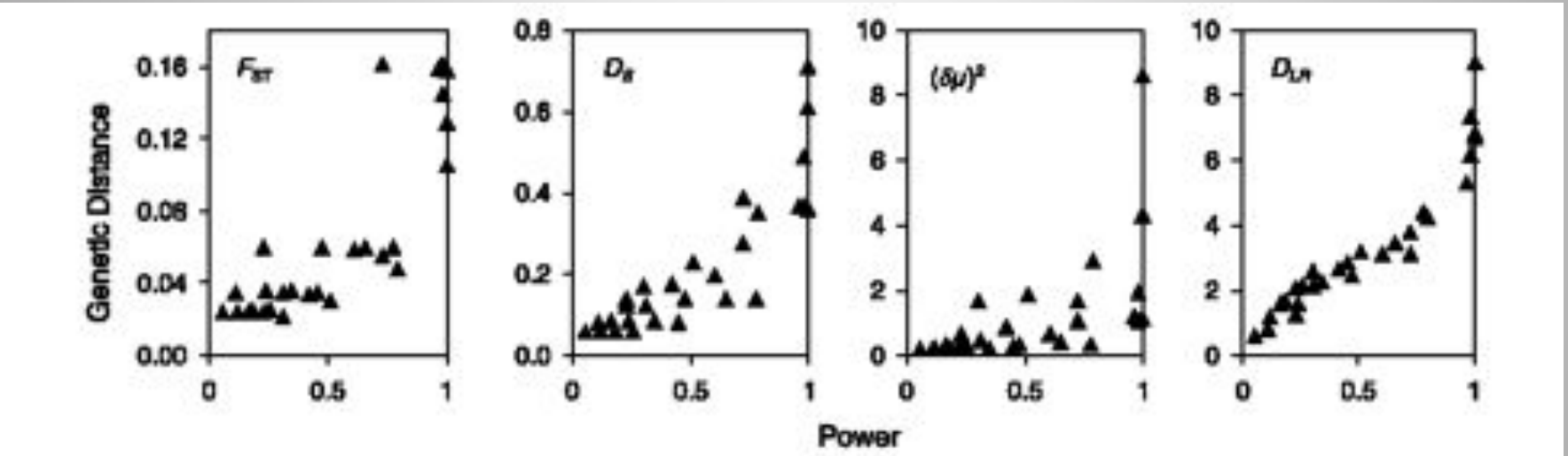


Fig. 7 Performance of four population genetic statistics in predicting power to identify *F0* immigrants. The 28 data sets involved different combinations of *I* (5–20), μ (0.0025–0.02) and *Nm* (1, 3, 5 or 7).

$\delta\mu^2$ < *Fst* < *shared allele distance Ds* **<< D_{LR} (Paetkau *et al.* 1997)**

Assignment and migration : an example on human populations

Rannala & Mountain (1997) : detecting immigrant individuals or individuals having immigrant parents

Comparison of the power of the approach for highly differentiated and moderately population

Australian and New Guinean (F_{ST}=0.056)

Japanese and Senegalese (F_{ST}=0.232)

12 individuals from each "population", RFLP markers

Assignment and migration : an example on human populations

Rannala & Mountain (1997) : detecting immigrant individuals or individuals having immigrant parents, grand-parents, etc…

$$\Lambda = \frac{p(\text{ind } i \text{ is born where he was sampled })}{p(\text{ind } i \text{ is an immigrant})}$$

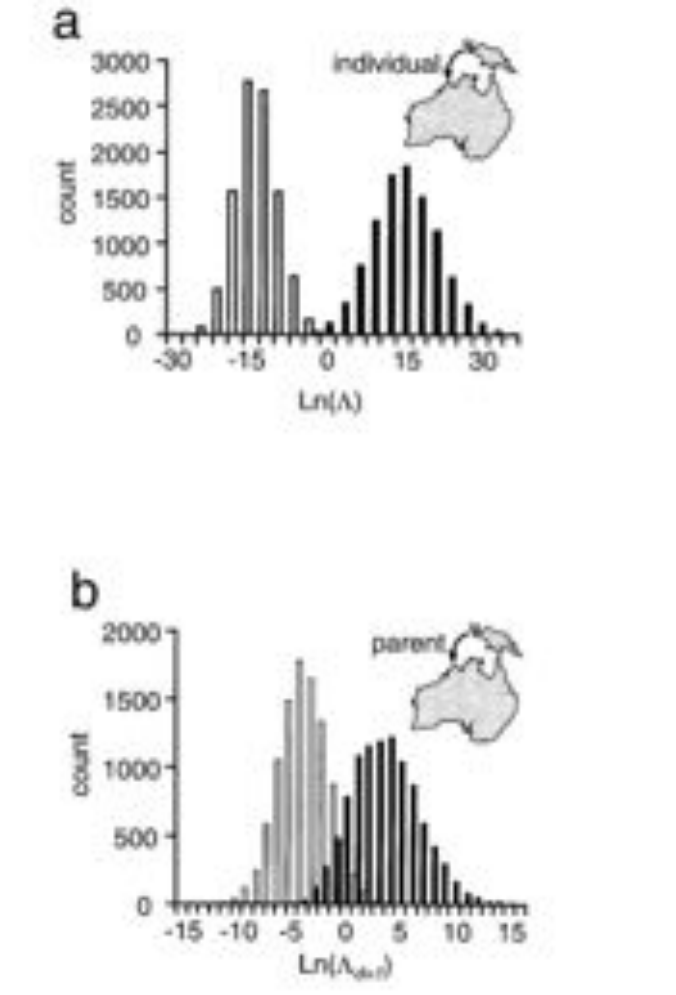
$$\Lambda_d = \frac{p(\text{all parents of ind } i \text{ , } d \text{ generation ago, were born where } i \text{ was sampled })}{p(\text{at least one parent of ind } i \text{ was an immigrant } d \text{ generation ago})}$$

ln Λ > 0 : the individual is a resident
ln Λ < 0 : the individual is an immigrant

ln Λ = -2.3 : the individual has 10 times more chance of being a immigrant than a resident

Assignment and migration : an example on human populations

Rannala & Mountain (1997) : detecting immigrant individuals or individuals having immigrant parents, grand-parents, etc…



power of the methods to detect:

- a immigrant individual
- an individual with one immigrant parent

from New Guinea to Australia

FIG. 2. Histograms indicating the power of the immigration tests for two cases. (*a*) The hypothesis that an Australian individual is an immigrant ($d = 0$) from New Guinea is considered. The shaded columns represent the distribution of ln Λ generated given the alleles observed for the Australian sample, while the unshaded columns represent the distribution of ln Λ generated given the alleles observed for the New Guinean sample. (*b*) The hypothesis that one parent of an Australian individual was an immigrant ($d = 1$) from New Guinea is considered. The shaded columns represent the distribution of ln Λ generated given the alleles observed for the Australian sample, while the unshaded columns represent the distribution of ln Λ generated given the alleles observed for the Australian and New Guinean samples and assuming that the individual received one allele at each locus from each population.

Assignment and migration : an example on human populations

Rannala & Mountain (1997) : detecting immigrant individuals or individuals having immigrant parents, grand-parents, etc…

4 individuals show signals of immigration :

3 Australian from New Guinea, 1 Japanese from Senegal (!)

Table 2. Power of the posterior probability ratio test to detect immigrant ancestry: Four individuals with posterior probability ratios indicating possible immigration ($\alpha < 0.05$)

Individual	Potential source	No. of markers	Value	Hypothetical immigrant ancestor: Individual ($d = 0$)	Parent ($d = 1$)	Grandparent ($d = 2$)	Great-grandparent ($d = 3$)
AUS1	NGN	76	ln A	−2.76	−2.89	−1.65	−0.89
			α	0.000	0.009	0.022	0.037
			Power	1.000	0.821	0.347	0.197
AUS2	NGN	73	ln A	4.48	0.87	−0.37	−0.11
			α	0.032	0.179	0.244	0.288
			Power	1.000	0.828	0.332	0.136
AUS3	NGN	82	ln A	5.23	−0.50	−0.90	−0.56
			α	0.032	0.049	0.064	0.092
			Power	1.000	0.862	0.375	0.149
JPN1	SEN	69	ln A	17.80	1.52	−1.26	−1.10
			α	0.021	0.014	0.029	0.045
			Power	1.000	0.999	0.771	0.431

Twelve individuals from each of four populations were included. Australians (AUS) were considered as possible immigrants, or descendants of immigrants, from New Guinea (NGN), and *vice versa*. Japanese (JPN) were considered as possible immigrants, or descendants of immigrants, from the Senegalese (SEN) population, and *vice versa*. Values of ln A or ln A_d are given in the first row for each individual. Values in the second row are significance levels (α values) approximated using the Monte Carlo approach (1,000 iterations per test). Values in the third row are the power of the test for this individual ($\alpha < 0.05$).

Assignment and migration : an example on human populations

Rannala & Mountain (1997) : detecting immigrant individuals or individuals having immigrant parents, grand-parents, etc…

ln Λ corresponding to α=0.05

ln Λ for individual AUS 1

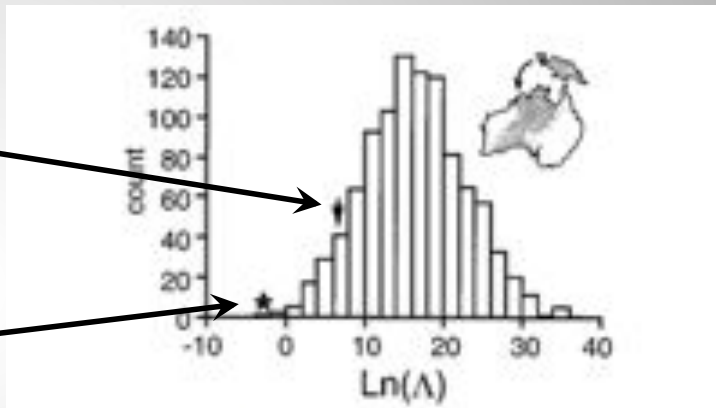


FIG. 1. Illustration of Monte Carlo method for examining significance of test statistic ln Λ for comparison of Australian (sample) and New Guinean (potential source) populations. Histogram of 1,000 values of the ln-probability difference generated by simulating genotypes given the allele counts observed for the Australian (sample) population. A total of 72 markers, for which the individual Australian 1 has been typed, were used to generate the distribution. See Eqs. 23–25. The critical region for the test statistic ($\alpha = 0.05$) is that portion of the distribution to the left of the arrow. The posterior probability ratio ($\ln \Lambda = -2.76$) for the individual Australian 1 is indicated by an asterisk.

Table 2. Power of the posterior probability ratio test to detect immigrant ancestry: indicating possible immigration ($\alpha < 0.05$)

Individual	Potential source	No. of markers	Value	Hypothetical immigrant ancestor: Individual ($d = 0$)	Parent ($d = 1$)	Grandparent ($d = 2$)	Great-grandparent ($d = 3$)
AUS1	NGN	76	ln Λ	−2.76	−2.89	−1.65	−0.89
			α	0.000	0.009	0.022	0.037
			Power	1.000	0.821	0.347	0.197

Assignment and migration : an example on human populations

Rannala & Mountain (1997) : detecting immigrant individuals or individuals having immigrant parents, grand-parents, etc…

4 individuals show signals of immigration :

3 Australian from New Guinea, 1 Japanese from Senegal (!)

Table 2. Power of the posterior probability ratio test to detect immigrant ancestry: Four individuals with posterior probability ratios indicating possible immigration ($\alpha < 0.05$)

Individual	Potential source	No. of markers	Value	Hypothetical immigrant ancestor: Individual ($d = 0$)	Parent ($d = 1$)	Grandparent ($d = 2$)	Great-grandparent ($d = 3$)
AUS1	NGN	76	ln Λ	−2.76	−2.89	−1.65	−0.89
			α	0.000	0.009	0.022	0.037
			Power	1.000	0.821	0.347	0.197

AUS 1 is probably a direct immigrant, or a descendant of an immigrant

Assignment and migration : an example on human populations

Rannala & Mountain (1997) : detecting immigrant individuals or individuals having immigrant parents, grand-parents, etc…

4 individuals show signals of immigration :

3 Australian from New Guinea, 1 Japanese from Senegal (!)

Table 2. Power of the posterior probability ratio test to detect immigrant ancestry: Four individuals with posterior probability ratios indicating possible immigration ($\alpha < 0.05$)

Individual	Potential source	No. of markers	Value	Hypothetical immigrant ancestor: Individual ($d = 0$)	Parent ($d = 1$)	Grandparent ($d = 2$)	Great-grandparent ($d = 3$)
AUS1	NGN	76	ln Λ	−2.76	−2.89	−1.65	−0.89
			α	0.000	0.009	0.022	0.037
			Power	1.000	0.821	0.347	0.197
AUS2	NGN	73	ln Λ	4.48	0.87	−0.37	−0.11
			α	0.032	0.179	0.244	0.288
			Power	1.000	0.828	0.332	0.136
AUS3	NGN	82	ln Λ	5.23	−0.50	−0.90	−0.56
			α	0.032	0.049	0.064	0.092
			Power	1.000	0.862	0.375	0.149

AUS 1 is probably a direct immigrant (relatively good confidence)

AUS 2 may be a descendant of a immigrant 2 or 3 generations ago

AUS 3 may be a descendant of a immigrant 1, 2 or 3 generations ago

much less confidence for AUS 2 and 3 than for AUS 1

Assignment and migration : an example on human populations

Rannala & Mountain (1997) : detecting immigrant individuals or individuals having immigrant parents, grand-parents, etc…

4 individuals show signals of immigration :

3 Australian from New Guinea, 1 Japanese from Senegal (!)

Table 2. Power of the posterior probability ratio test to detect immigrant ancestry: Four individuals with posterior probability ratios indicating possible immigration ($\alpha < 0.05$)

Individual	Potential source	No. of markers	Value	Hypothetical immigrant ancestor: Individual ($d = 0$)	Parent ($d = 1$)	Grandparent ($d = 2$)	Great-grandparent ($d = 3$)
JPN1	SEN	69	ln A	17.80	1.52	−1.26	−1.10
			α	0.021	0.014	0.029	0.045
			Power	1.000	0.999	0.771	0.431

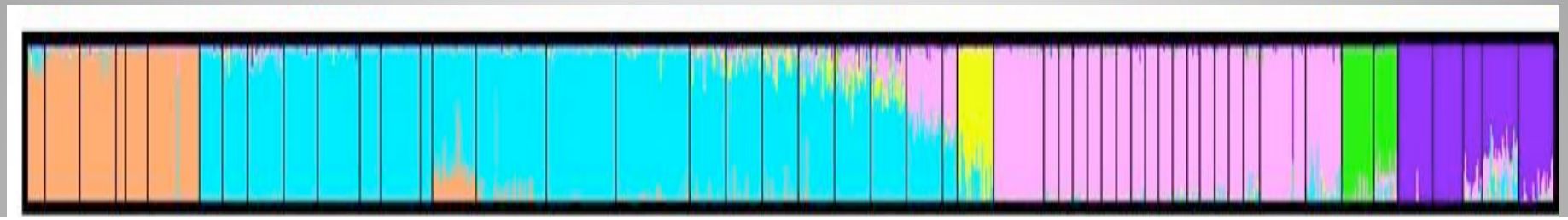
Twelve individuals from each of four populations were included. Australians (AUS) were considered as possible immigrants, or descendants of immigrants, from New Guinea (NGN), and *vice versa*. Japanese (JPN) were considered as possible immigrants, or descendants of immigrants, from the Senegalese (SEN) population, and *vice versa*. Values of ln A or ln A_d are given in the first row for each individual. Values in the second row are significance levels (α values) approximated using the Monte Carlo approach (1,000 iterations per test). Values in the third row are the power of the test for this individual ($\alpha < 0.05$).

JPN 1 may be a descendant of a immigrant 2 generations ago

But Paetkau et al. (2004) showed that Rannala & Mountain method was to confident in detecting immigrants !

because of "bad" Monte Carlo simulation of the criterion distribution (simulation of allelic vs. haplotypic migration)

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 allelic 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

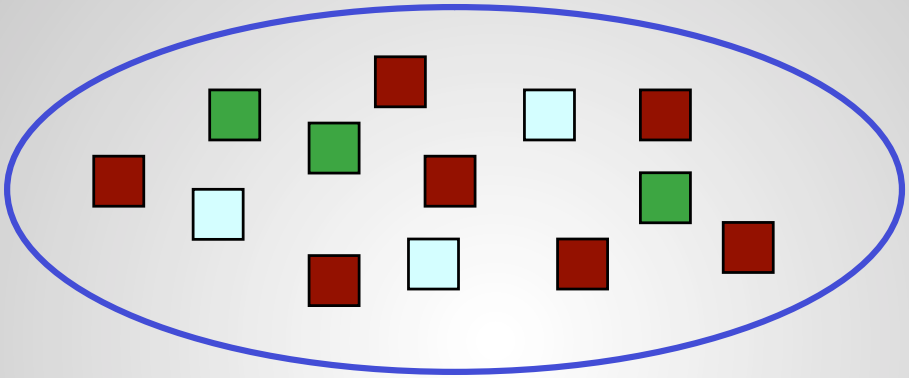
Same assumptions than for assignment methods:

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
4 different models



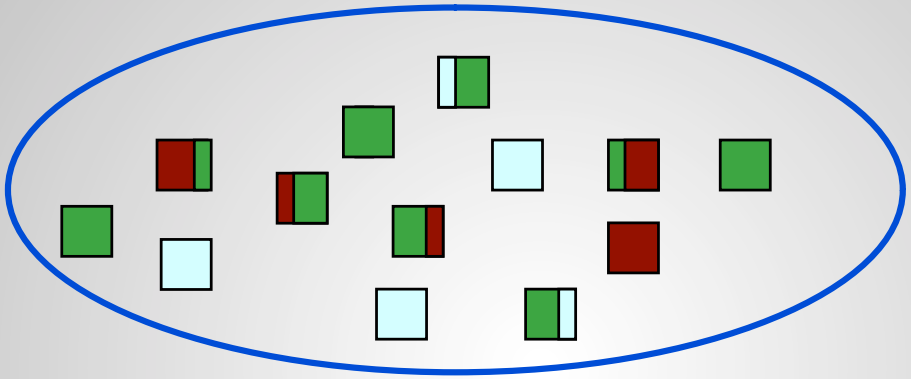
1. **the basic model without admixture**

   Assumption :

   each individual come from a unique

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

STRUCTURE
4 different models



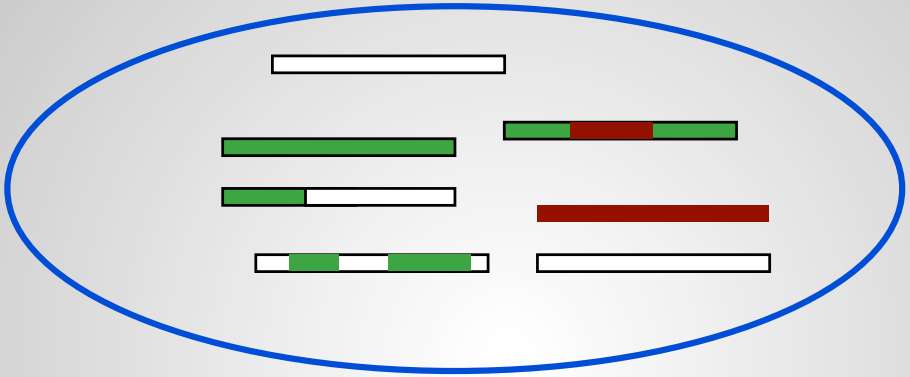
2. the model with admixture (most commonly used)

Assumption:

the different genes of an individual may come from different cluster 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
4 different models

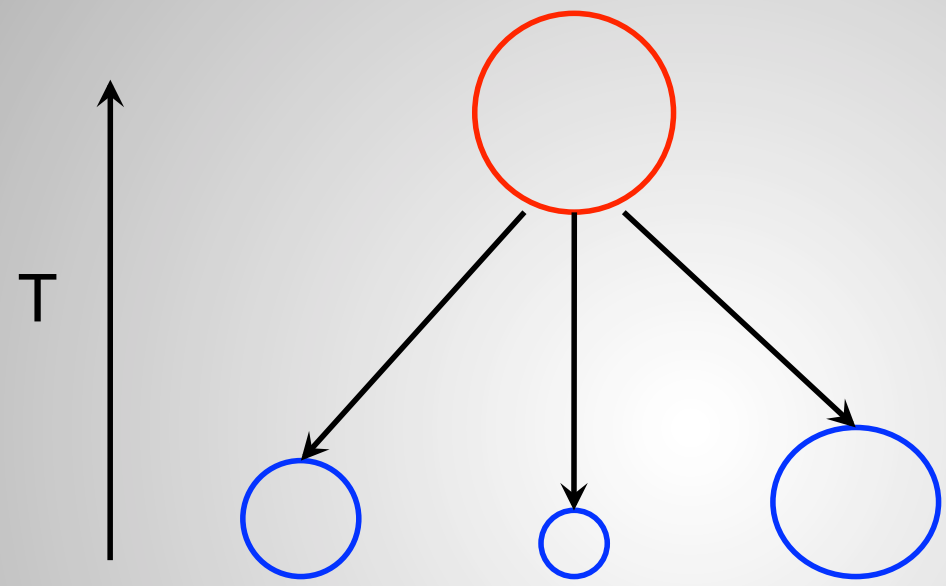


3. **the linkage model (explicit recombination on chromosomes)**

   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

STRUCTURE
4 different models

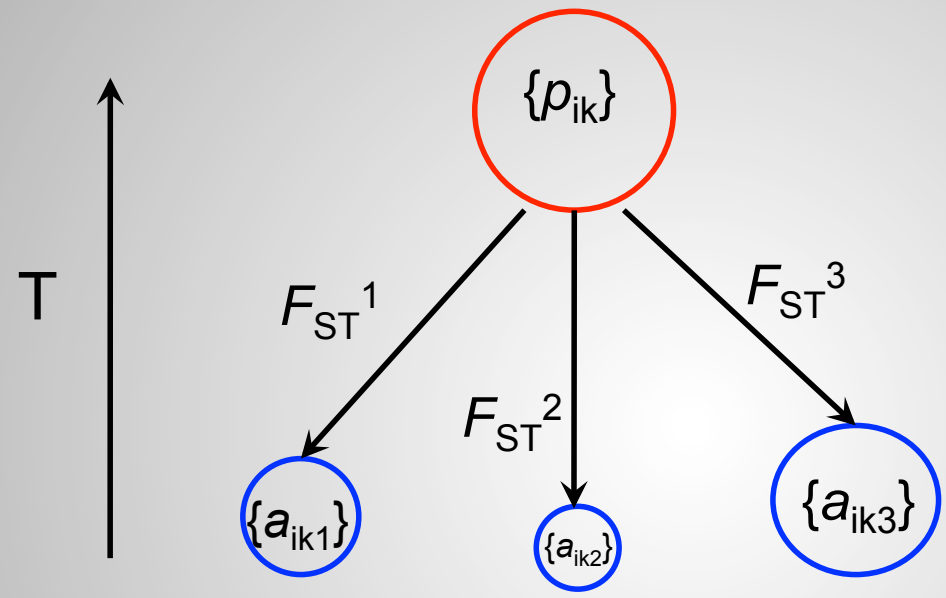


4. the F-model of ancestry (the correlated allele frequency model)

instead of considering independent allele frequencies in each cluster, the dependence between allele frequencies in the different cluster are modeled using a pure drift model for the ancestry of the different clusters

It can be use with the different models described above

STRUCTURE
4 different models

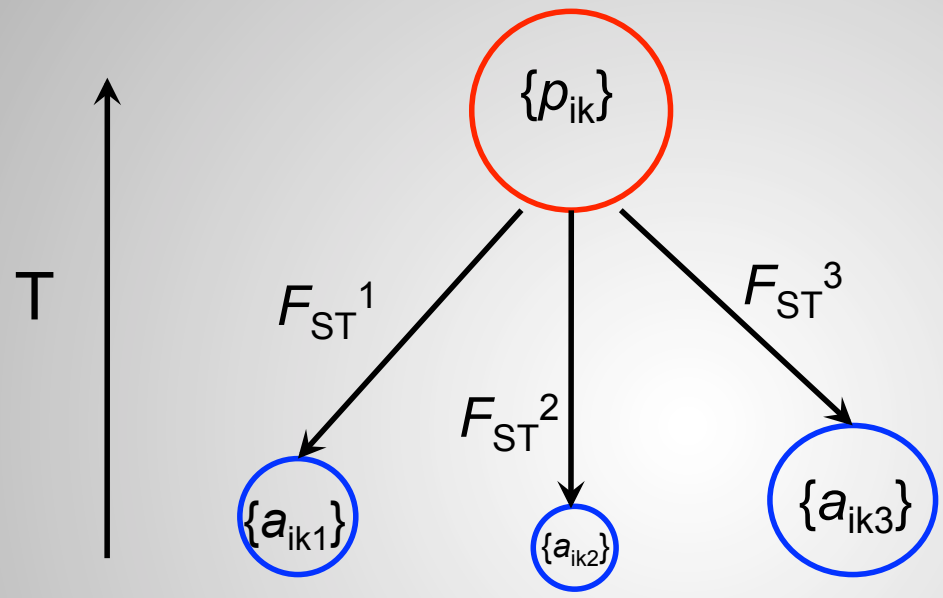


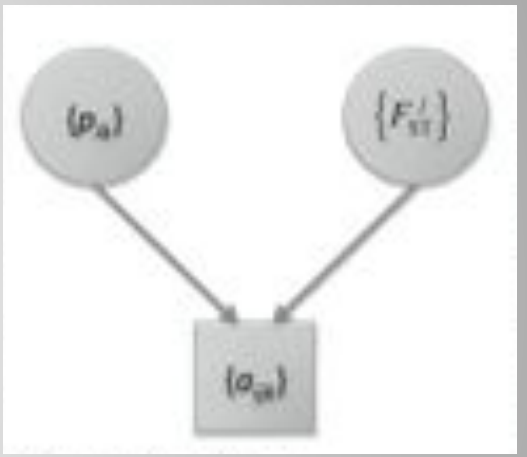
4. the F-model of ancestry (the correlated allele frequency model)

{p_{ik}} : allele frequencies in the ancestral pop;
{a_{ikj}} : allele frequencies in the actual populations

{F_{ST}^j} : differentiation level between the actual and the ancestral population
= measure of the level of drift acting on the derived populations

STRUCTURE
4 different models



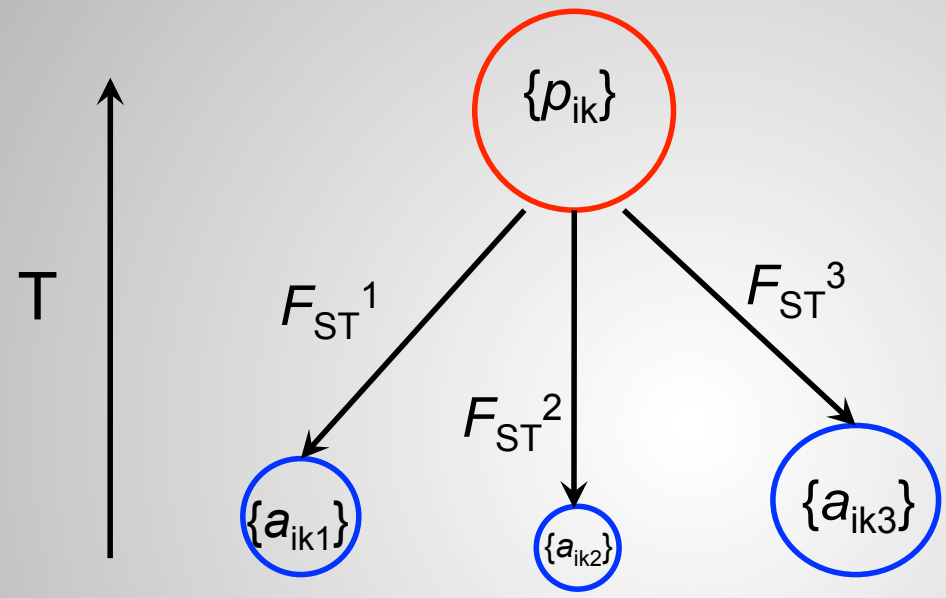


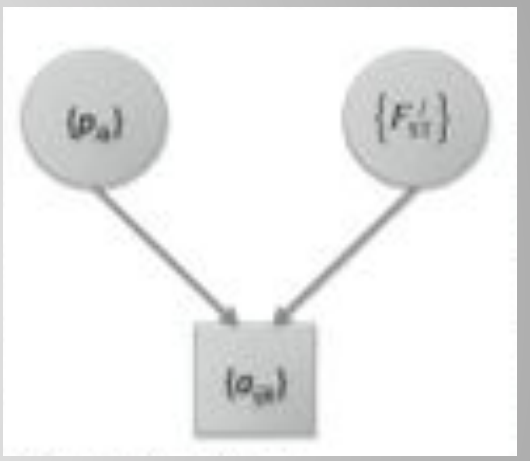
4. the F-model of ancestry (the correlated allele frequency model)

$\{p_{ik}\}$: allele frequencies in the ancestral pop;
$\{a_{ikj}\}$: allele frequencies in the actual populations

$\{F_{ST}^{j}\}$: differentiation level between the actual and the ancestral population
= measure of the level of drift acting on the derived populations

STRUCTURE
4 different models





4. the F-model of ancestry (the correlated allele frequency model)

This model is considering drift only but not migration (there is an equivalent model for allelic frequency correlation under an island model but not implemented in STRUCTURE)

It must thus be used on biological data that do not strongly deviate from this assumption, otherwise it is risky!

STRUCTURE inference method

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

	loc = 1			**l = *l***			**l = L (locus number)**	
	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×2L)

STRUCTURE inference method

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

microsatellite data set example

	phi011		phi015		phi029		phi031		phi062	
1	212	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 inference method

the data = *X* = individual multilocus genotypes

unknown variables:
Z = cluster membership of each individual

For the model without admixture, *Z* is a vector

- 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*

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

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

STRUCTURE inference method

the data = *X* = individual multilocus genotypes

unknown variables:
Z = cluster membership of each individual or each individual locus

For the model with admixture or the linkage model, *Z* is a matrix

$P(z^{(i,l)} = k)$ is the probability that locus (or chromosome part) *l* of individual *i* is a member of cluster *k*

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

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

STRUCTURE inference method

the data = X = individual multilocus genotypes

unknown variables:
Z = cluster membership of each individual or each individual locus
P = allele frequencies in each cluster

	loc = 1			l = *l*			l = *L*	
	j=1	j=2		j=1	j=2		j=1	j=2

$$\mathbf{P} = \begin{bmatrix} (p_{111} & p_{112}) & \cdots & (p_{1l1} & p_{1l2}) & \cdots & (p_{1L1} & p_{1L2}) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ (p_{k11} & p_{k12}) & \cdots & (p_{kl1} & p_{kl2}) & \cdots & (p_{kL1} & p_{kL2}) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ (p_{K11} & p_{K12}) & \cdots & (p_{Kl1} & p_{Kl2}) & \cdots & (p_{KL1} & p_{KL2}) \end{bmatrix}$$

$$\mathbf{P}_{(K \times 2L)}$$

STRUCTURE inference method

the data = *X* = individual multilocus genotypes

unknown variables:
Z = cluster membership of each individual or each individual locus
P = allele frequencies in each cluster

For the model with correlated allele frequencies, there are two additional variables :

P' = vector of allele frequencies in the ancestral populations
F = vector of the *K* F_{ST} values between the ancestral and the derived clusters

STRUCTURE inference method

the data = *X* = individual multilocus genotypes

unknown variables:
Z = cluster membership of each individual or each individual locus
P = allele frequencies in each cluster

the idea (i.e. simplified algorithm) is that assuming Hardy-Weinberg and linkage equilibrium, the likelihood of the sample for a given partition is proportional to :

$$p(X \mid Z,P) = \prod_{\text{ind } i} \prod_{\text{locus } l} 2 \cdot \mathrm{p}_{\mathrm{z(i,1,l),i,1,l}} \cdot \mathrm{p}_{\mathrm{z(i,2,l),i,2,l}}$$

impossible to explore all partitions ➠ Markov chain Monte Carlo simulation

STRUCTURE inference method

For a fixed value of the number of clusters *K* ,the probability that individual i is a member of cluster *k* can be expressed as (Bayes rules) :

$$p(Z_i = k | X_i, P) = \frac{p(X_i | Z_i = k, P) \cdot \ p(Z_i = k)}{\sum_{j=pops} p(X_i | Z_i = j, P) \cdot \ p(Z_i = j)}$$

where p(Z_i=*k*) is the prior probability of membership of individual *i* (equals 1/*K* for all *I* and *k*)

STRUCTURE inference method

an estimator for allele frequencies in each pop is :

$$\hat{p}_{jlk} = \frac{\text{number of genes of type } j \text{ in pop } k}{\text{total number of genes in pop } k}$$

in STRUCTURE, a Dirichlet distribution is used for allele frequencies

STRUCTURE inference method

MCMC algorithm : inference of cluster membership of all individuals = partition of the sample into *K* clusters (fixed *K* value)

the main steps of the MCMC :

step 1 : Allele frequencies for each cluster are inferred from individual genotypes assigned to the each cluster at the previous step

step 2 : individuals are assigned to clusters using the allele frequencies computed previously

STRUCTURE inference method

MCMC algorithm : inference of cluster membership of all individuals = partition of the sample into *K* clusters (fixed *K* value)

the main steps of the MCMC :

step 1 : Allele frequencies for each cluster are inferred from individual genotypes assigned to the each cluster at the previous step

step 2 : individuals are assigned to clusters using the allele frequencies computed previously

if those steps are repeated a large number of times, the partition of individuals/loci will converge towards its stationary distribution

STRUCTURE inference method

MCMC algorithm : inference of cluster membership of all individuals = partition of the sample into *K* clusters (fixed *K* value)

the main steps of the MCMC :

Initialization: place individuals at random on all clusters $p(Z_i=k) = 1/K$
then :

Repeat
m=1,2,..M
times

1. draw $P(m)$ from $p(P|X, Z(m\text{-}1))$

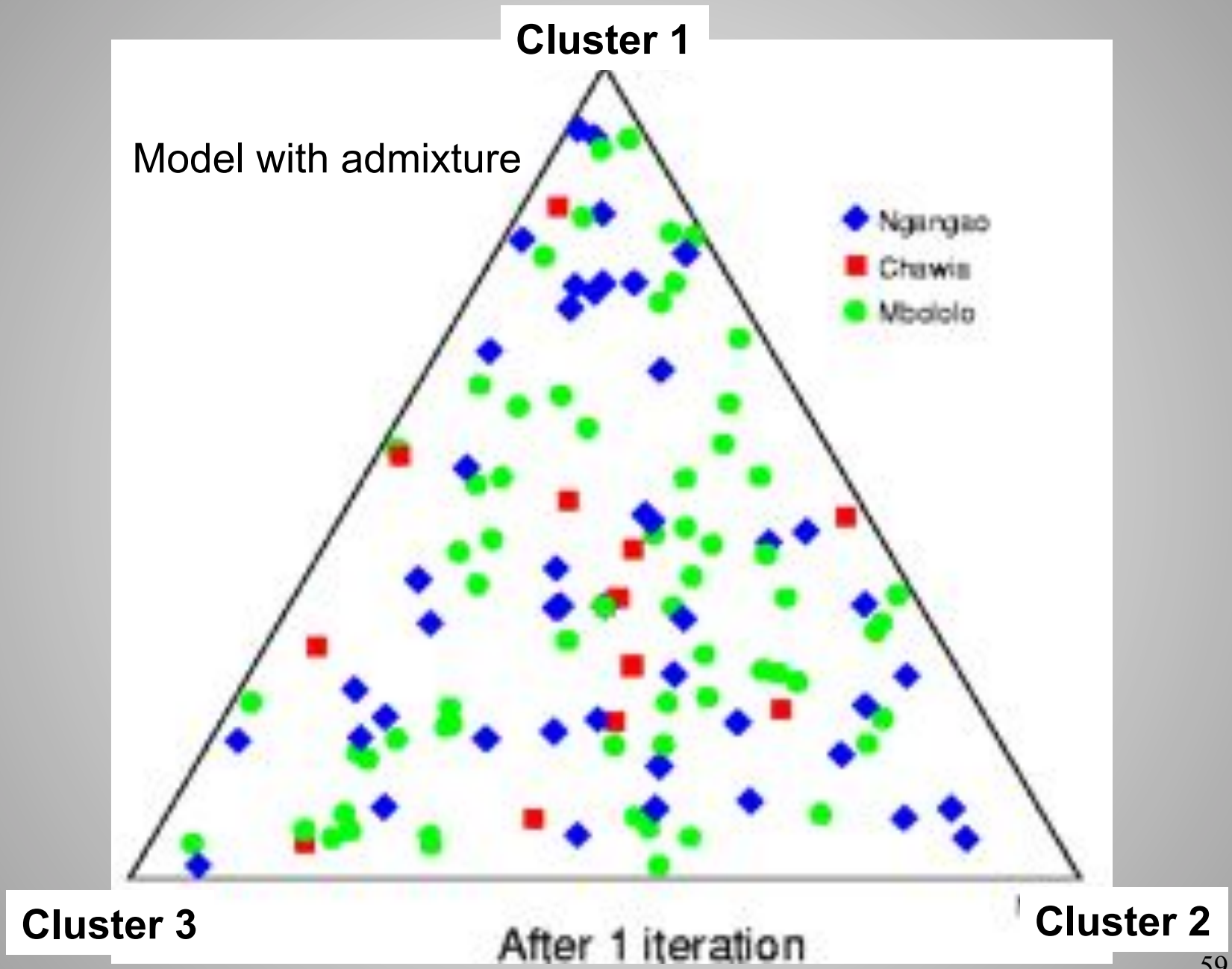
2. draw $Z(m)$ from $p(Z|X, P(m))$

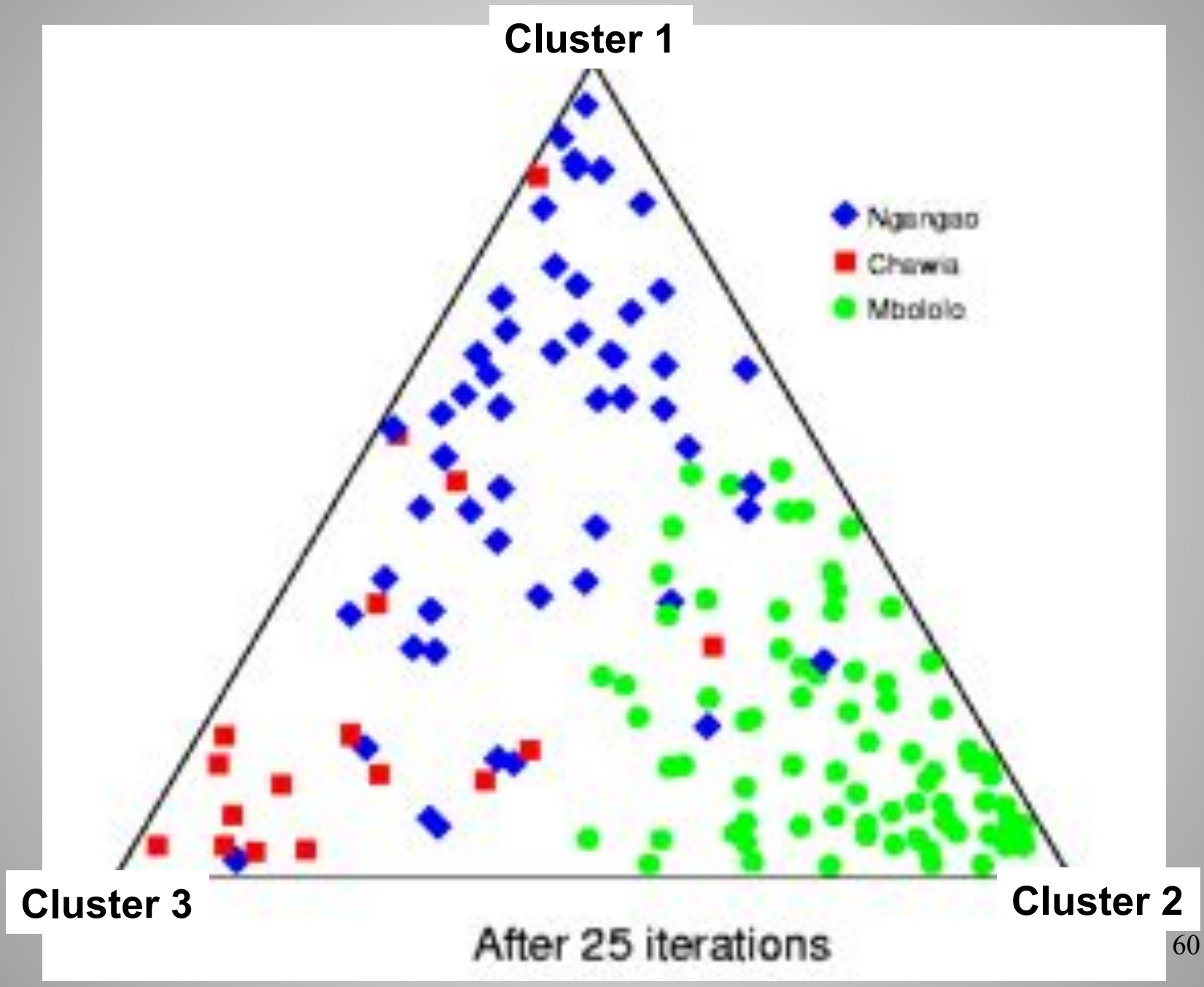
For large *M*, *P* and *Z* will converge towards their stationary distributions

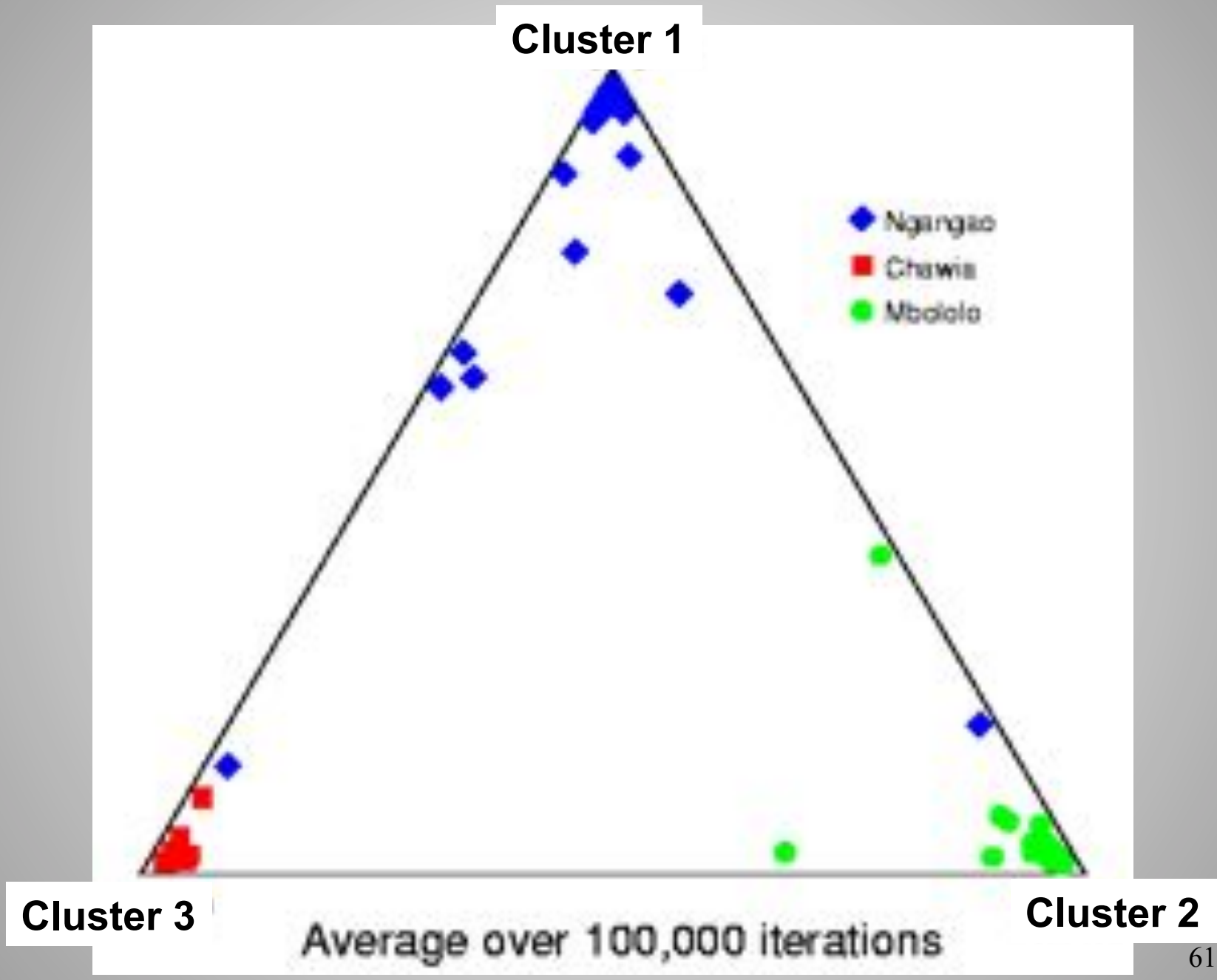
Example: Taita Thrush data

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

*Data courtesy of Dr Peter Galbusera


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


Cluster 1
Ngangao
Chawia
Mbololo
Cluster 3
Cluster 2
After 25 iterations


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

Inferred ancestry of individuals:

Proportion of individuals' genotypes, originating from each K populations.

	Label	(%Miss)	Pop:	Inferred clusters
1	CH01()	(0)	1 :	0.403 0.544 0.053
2	CH02()	(0)	1 :	0.877 0.072 0.051
3	CH03()	(0)	1 :	0.808 0.030 0.162
4	CH04()	(0)	1 :	0.136 0.010 0.854
5	CH04()	(0)	1 :	0.956 0.023 0.021
6	CH06()	(0)	1 :	0.941 0.026 0.033
7	CH07()	(0)	1 :	0.648 0.106 0.246
8	CH09()	(0)	1 :	0.775 0.038 0.187
9	CH10()	(0)	1 :	0.892 0.034 0.074
10	CH11()	(0)	1 :	0.617 0.039 0.344
11	CH14()	(0)	1 :	0.678 0.142 0.181
12	CH14()	(0)	1 :	0.766 0.036 0.198
13	CH16()	(0)	1 :	0.554 0.235 0.210
14	CH17()	(0)	1 :	0.870 0.042 0.088
15	CH18()	(0)	1 :	0.809 0.078 0.113
16	CH19()	(0)	1 :	0.808 0.059 0.133
17	CH20()	(4)	1 :	0.341 0.017 0.641
18	CH1()	(0)	1 :	0.575 0.356 0.069
19	CH2()	(0)	1 :	0.125 0.015 0.860
20	CH3()	(4)	1 :	0.794 0.015 0.190
21	CH4()	(0)	1 :	0.850 0.017 0.133

Estimated Allele Frequencies in each population

First column gives estimated ancestral frequencies

Locus 1 :

2 alleles

0.0% missing data

1	(0.681)	0.691	0.579	0.582
2	(0.319)	0.309	0.421	0.418

Locus 2 :

2 alleles

0.3% missing data

1	(0.694)	0.698	0.434	0.796
2	(0.306)	0.302	0.566	0.204

Locus 3 :

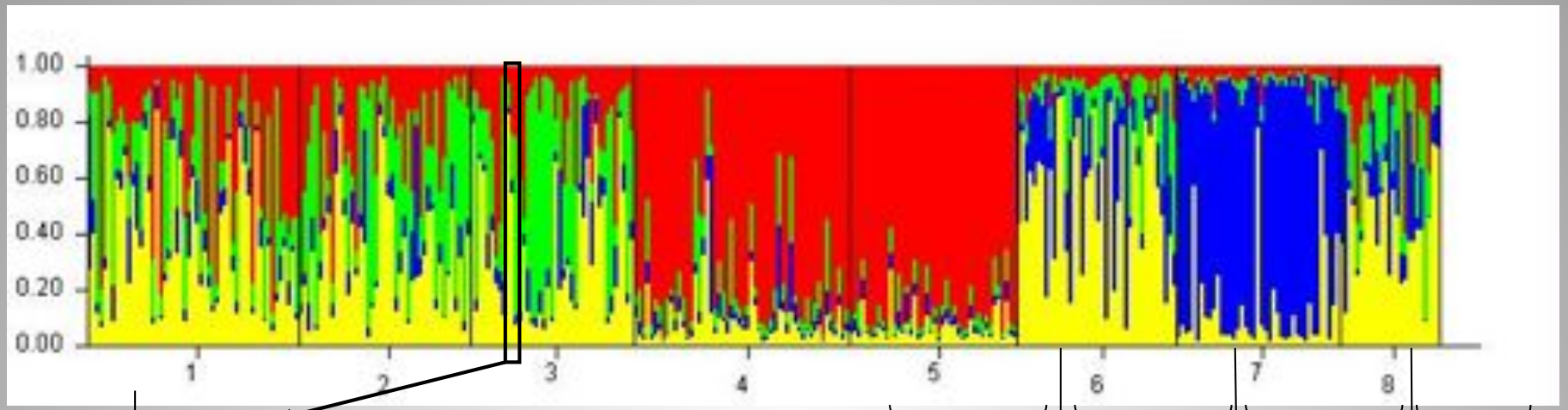
2 alleles

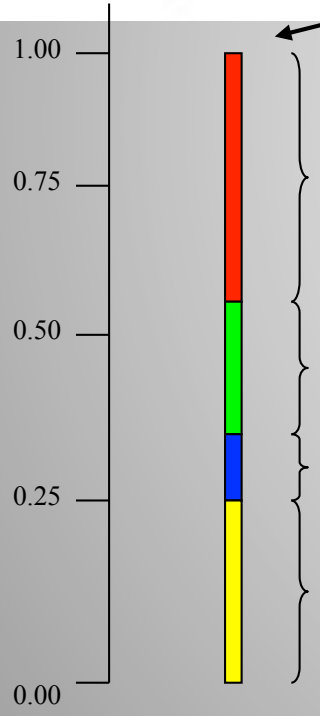
2.1% missing data

1	(0.434)	0.433	0.297	0.510
2	(0.566)	0.567	0.703	0.490

Cluster 1 Cluster 2 Cluster 1

STRUCTURE typical plots





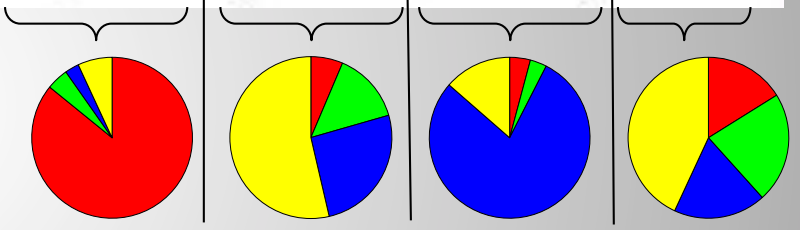
Individual
74

***Proportion of "#74"'s genotype belonging to cluster* K=1**

K=2

K=3

K=4



Pies:

Structure prints out a summary table of the average proportions of membership of each <u>pre-defined population</u> in each of the K clusters, that can be plotted in pies.

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*]

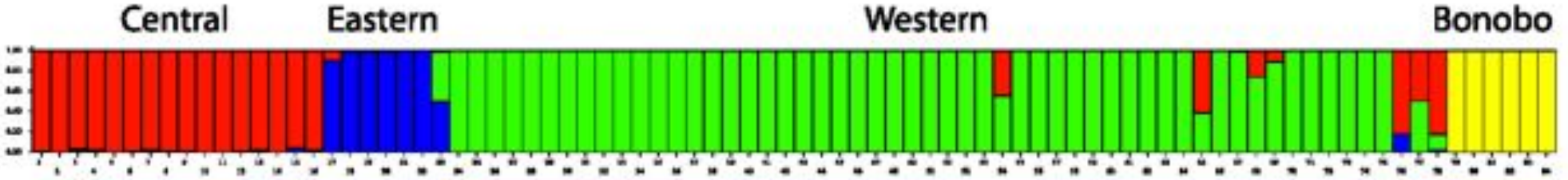


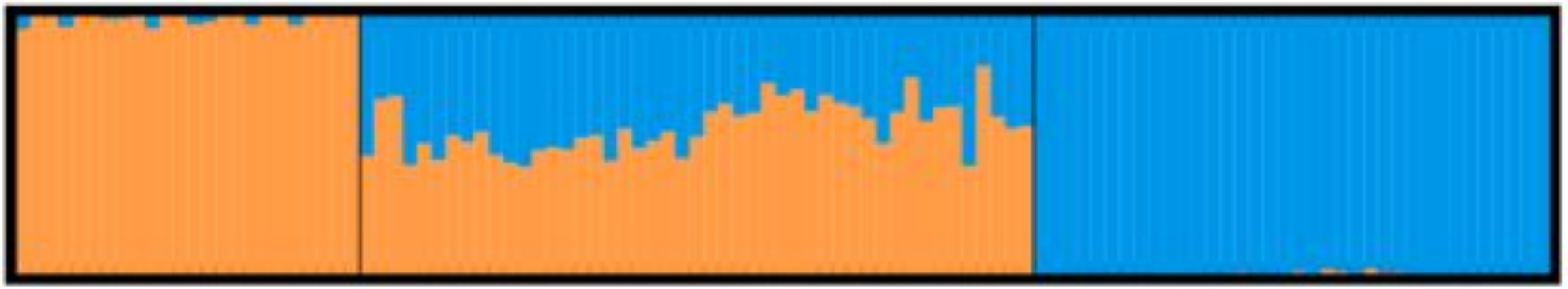
Figure 1. STRUCTURE Analysis, Blinded to Population Labels, Recapitulates the Reported Population Structure of the Chimpanzees

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



ZEBU FULANI (N=30)

BORGOU (N=47)

SOMBA (N=32)

clear admixture pattern

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 :

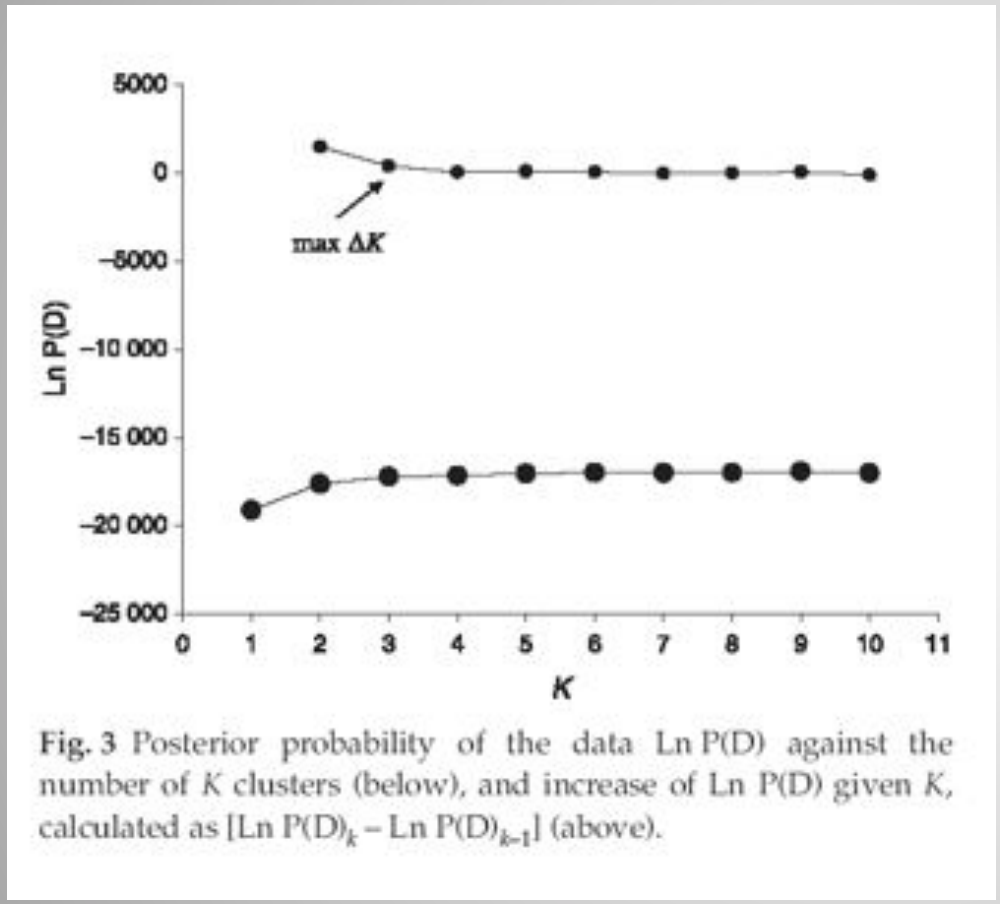


Fig. 3 Posterior probability of the data Ln P(D) against the number of *K* clusters (below), and increase of Ln P(D) given *K*, calculated as [Ln $P(D)_k$ – Ln $P(D)_{k-1}$] (above).

Scottish feral cat

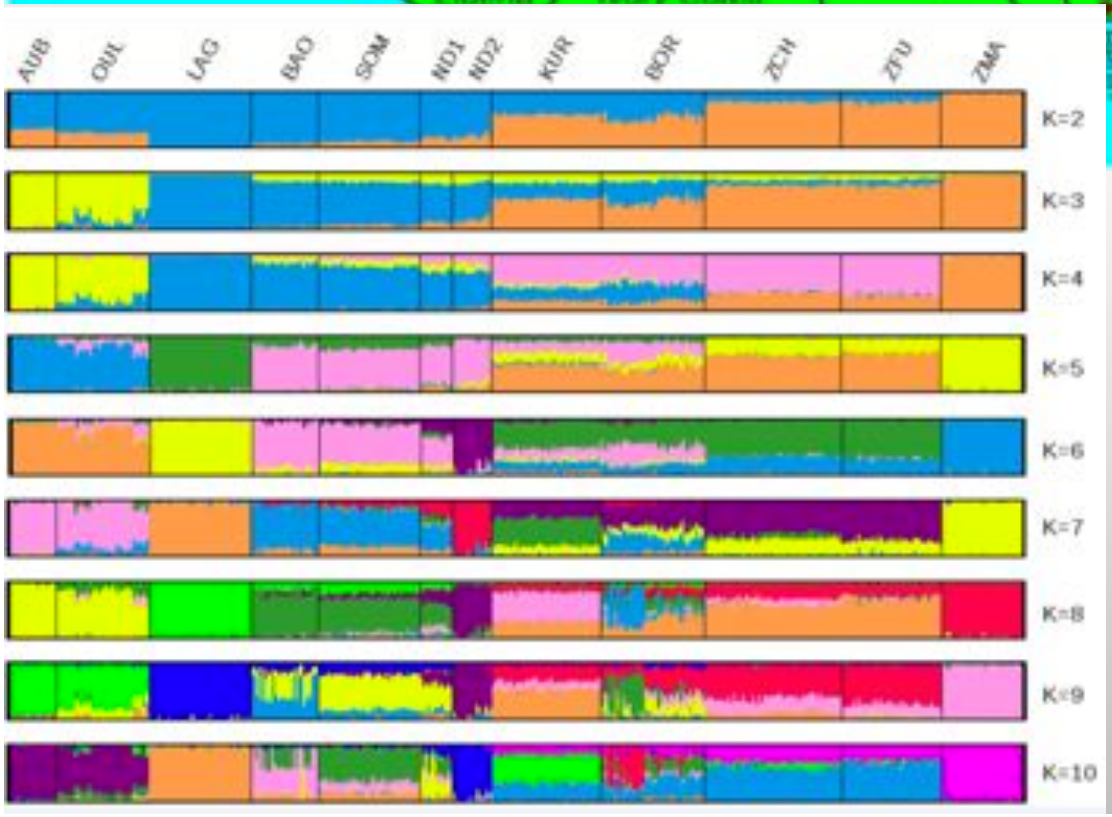
the variation in likelihood between different K values can also be used (ΔK)

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.

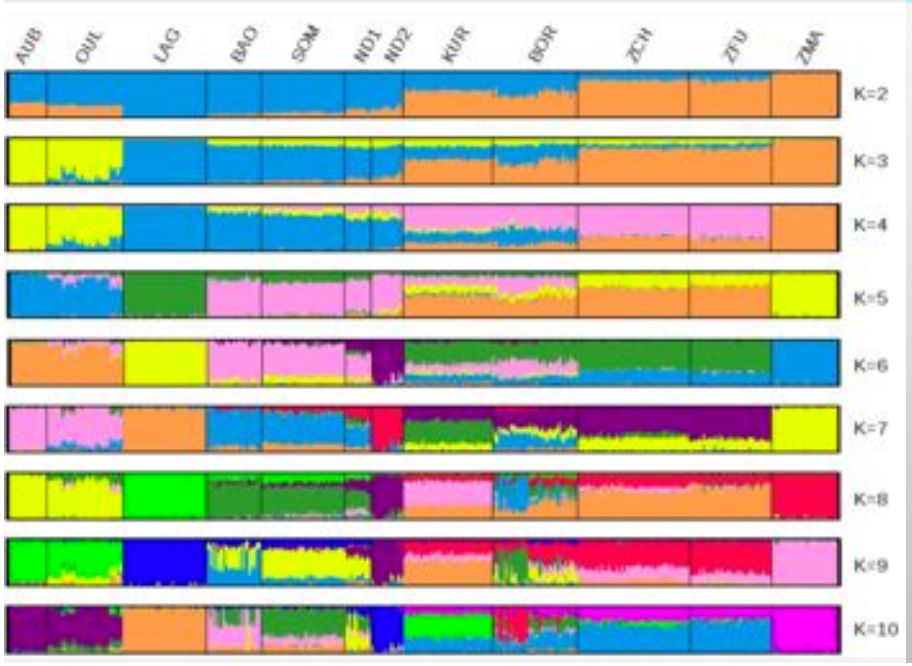
Inference of the number of clusters *K*

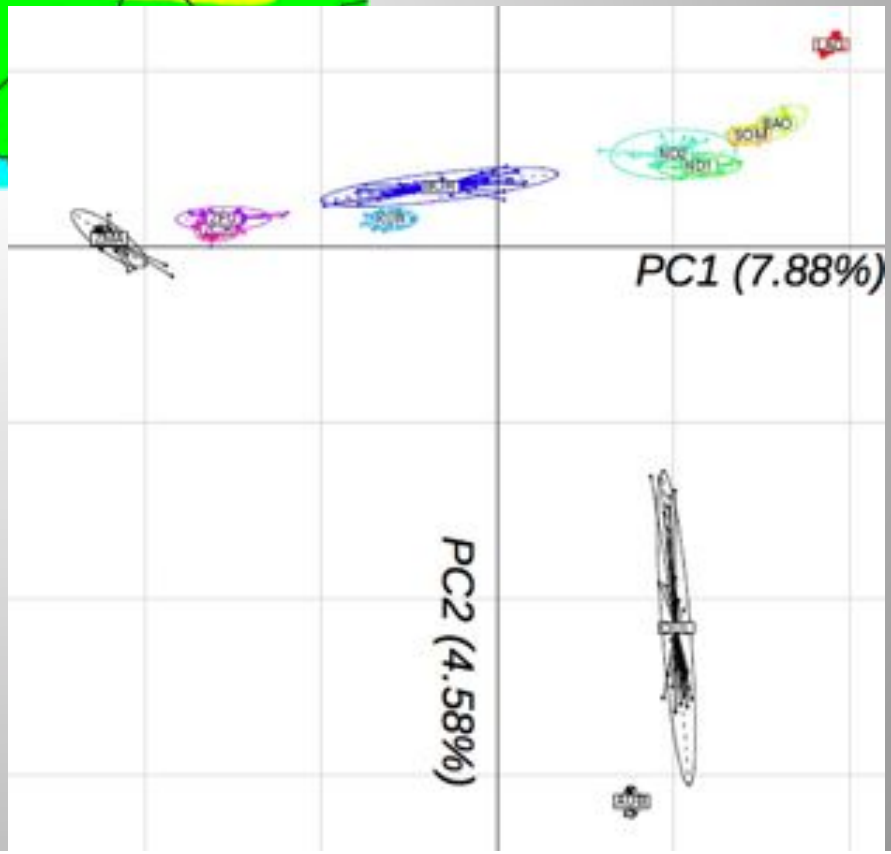


It is usually better to analyze different values of *K*, and conclude from all of them instead of focusing on the "best" *K* value

Inference of the number of clusters *K*

STRUCTURE may thus be considered as a representative population genetic tool (like PCA) rather than an inference method strictly



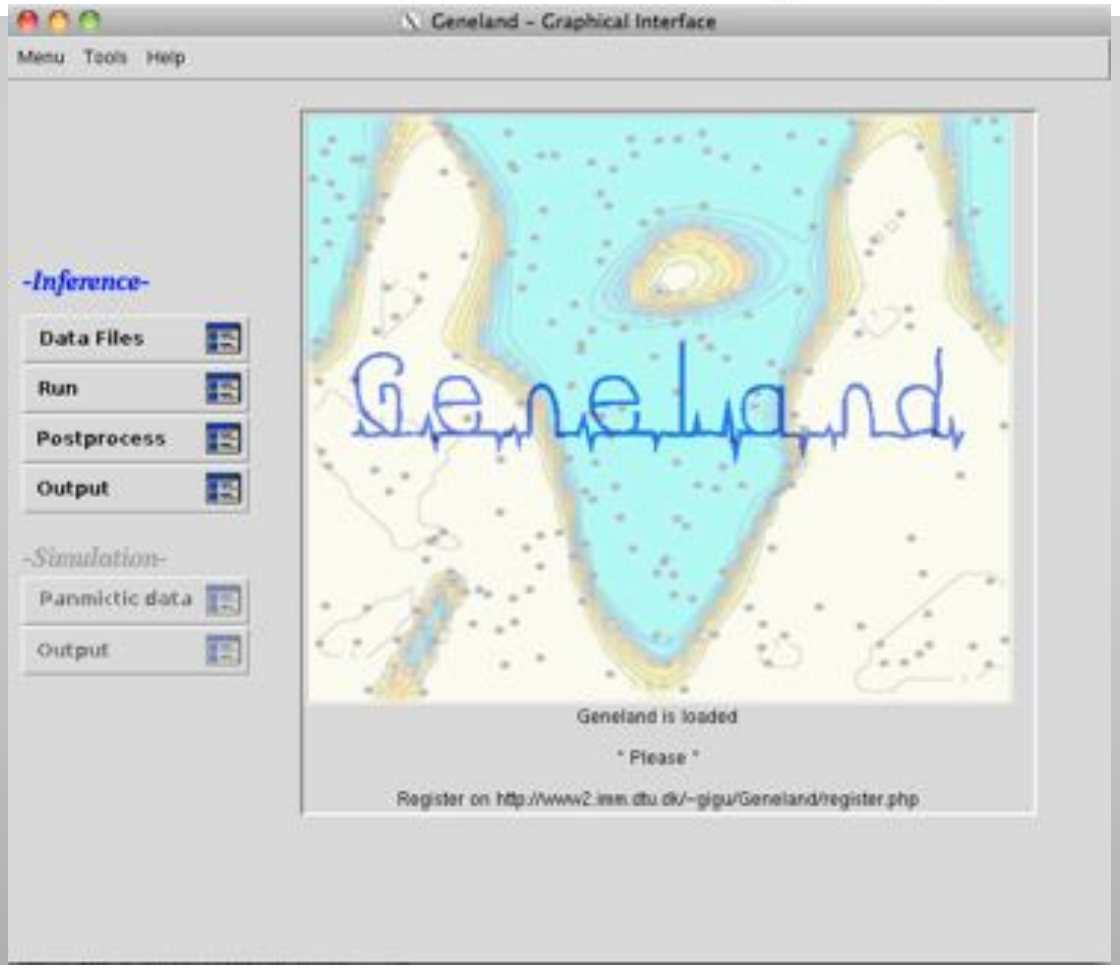


Spatial clustering: the GENELAND software

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

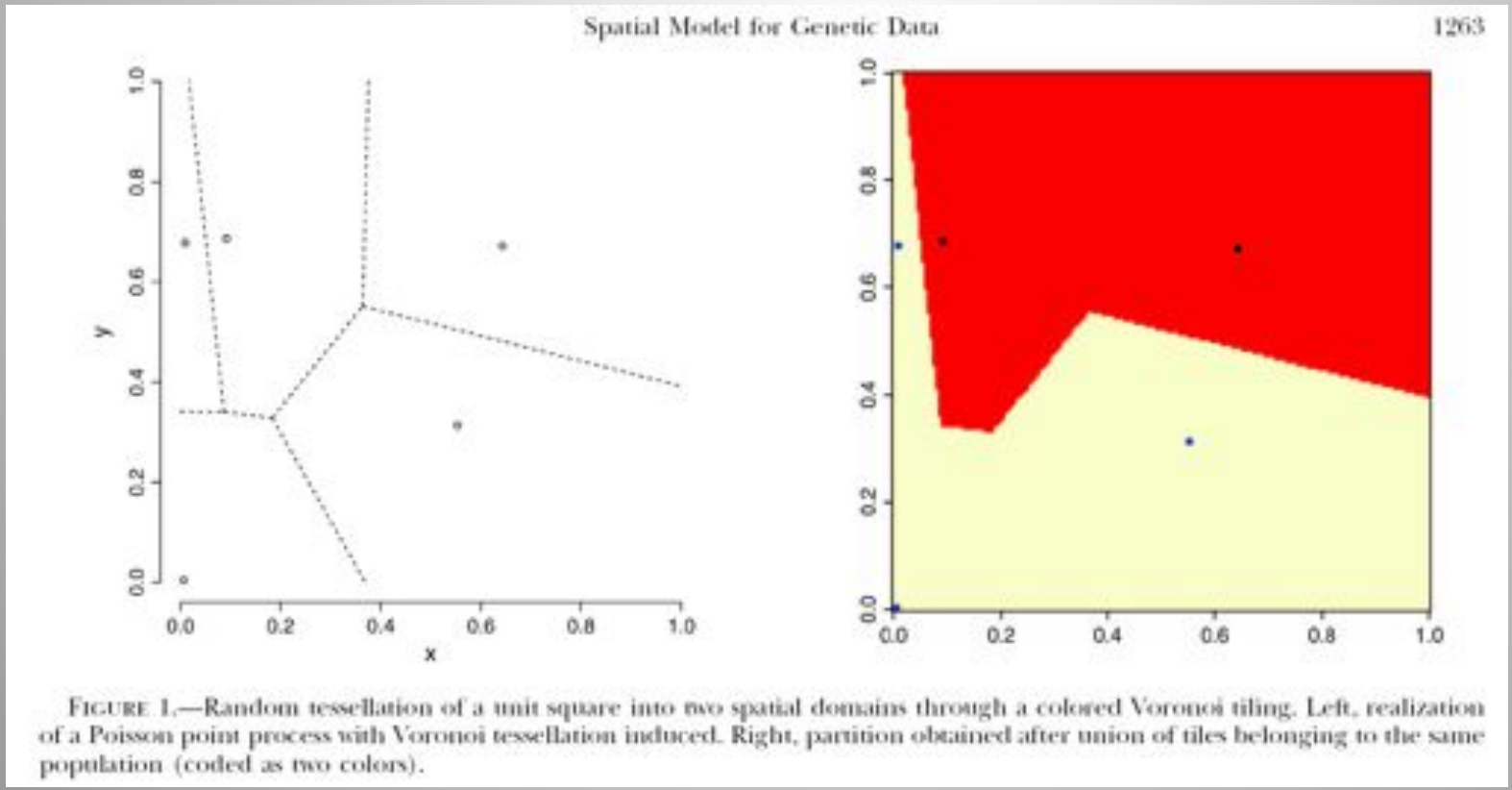
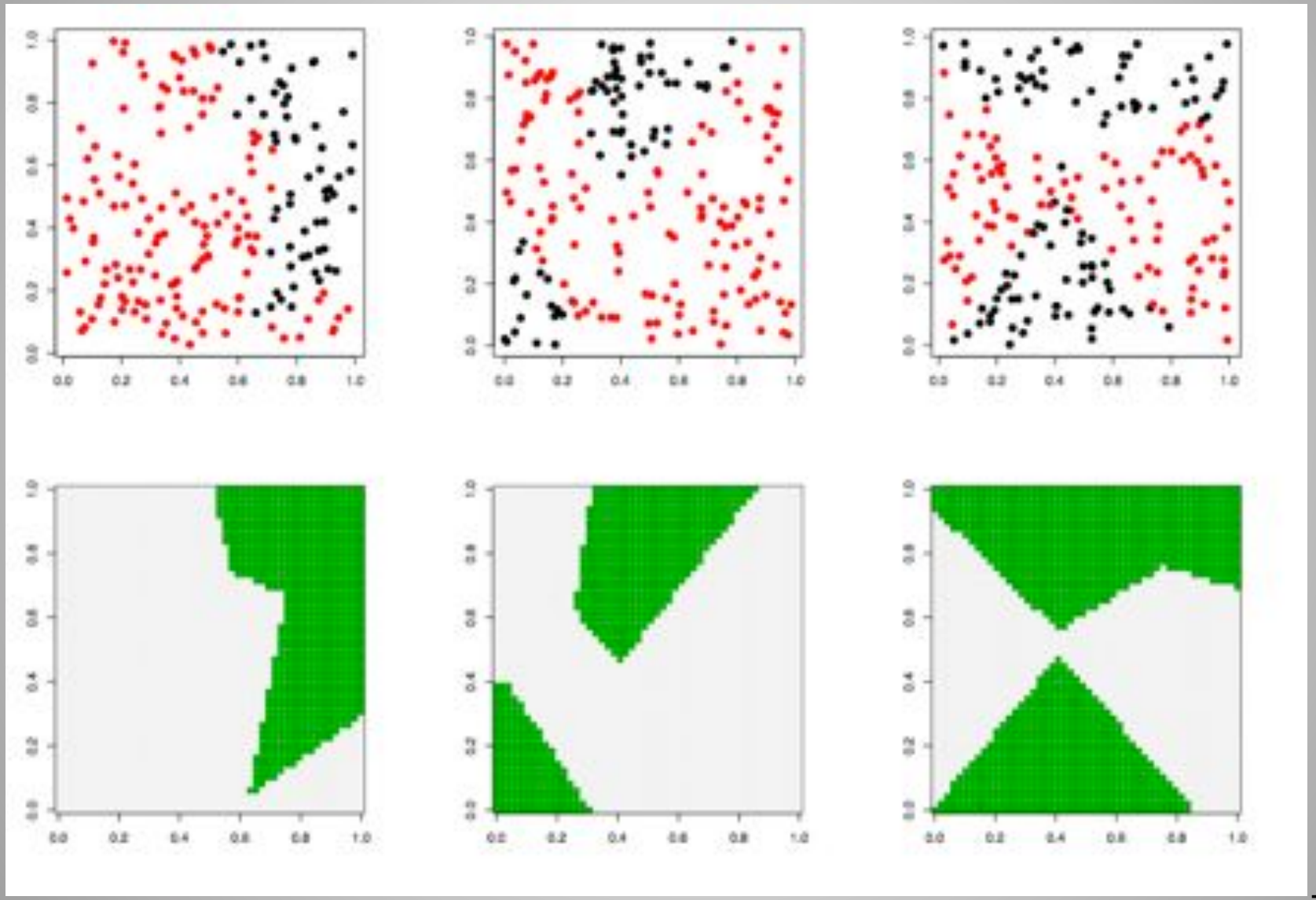


FIGURE 1.—Random tessellation of a unit square into two spatial domains through a colored Voronoi tiling. Left, realization of a Poisson point process with Voronoi tessellation induced. Right, partition obtained after union of tiles belonging to the same population (coded as two colors).

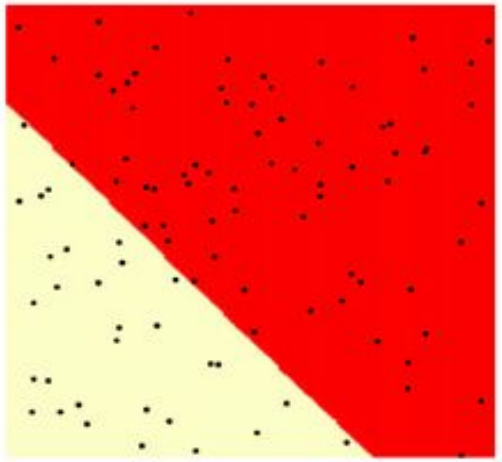
GENELAND spatial population model



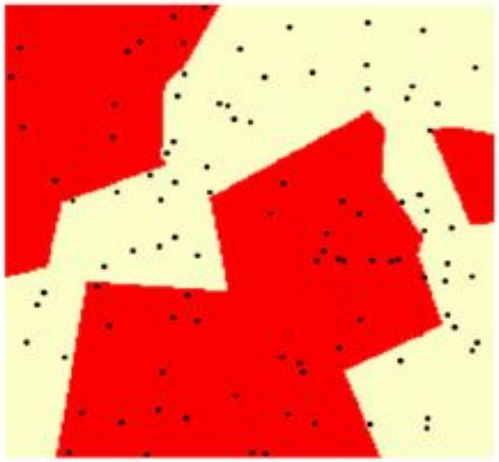
GENELAND spatial population model

Set of spatialized panmictic populations

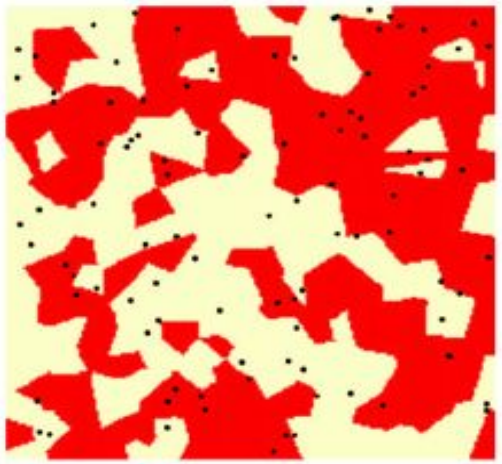
example of different Tessellation outputs for different spatial correlations



$m = 2$



$m = 20$



$m = 200$

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

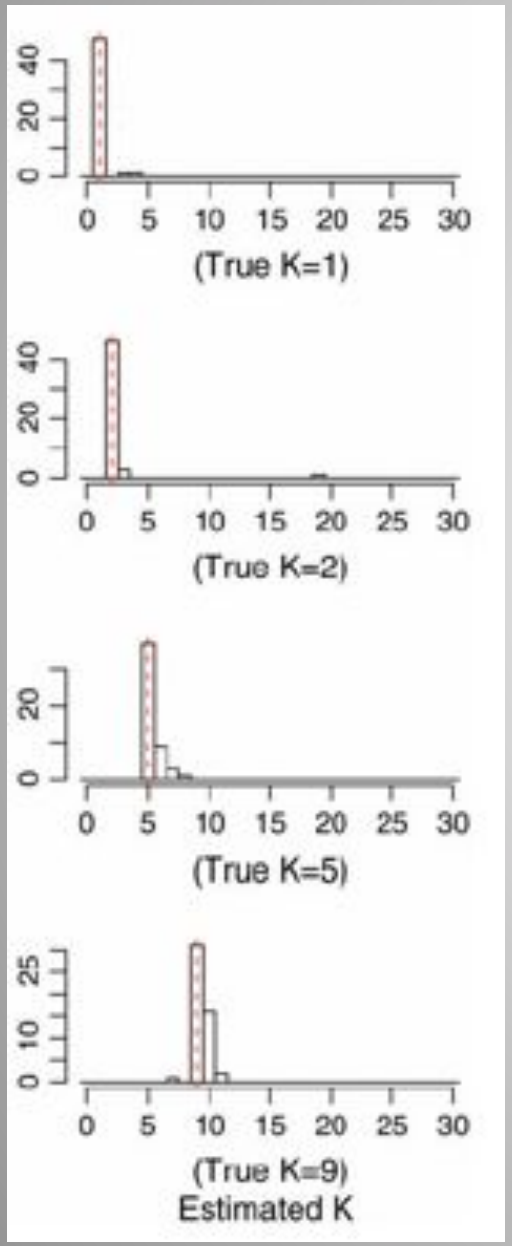
the principle of the method is very close to STRUCTURE method with additional parameters for the spatial arrangement of the different cluster

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

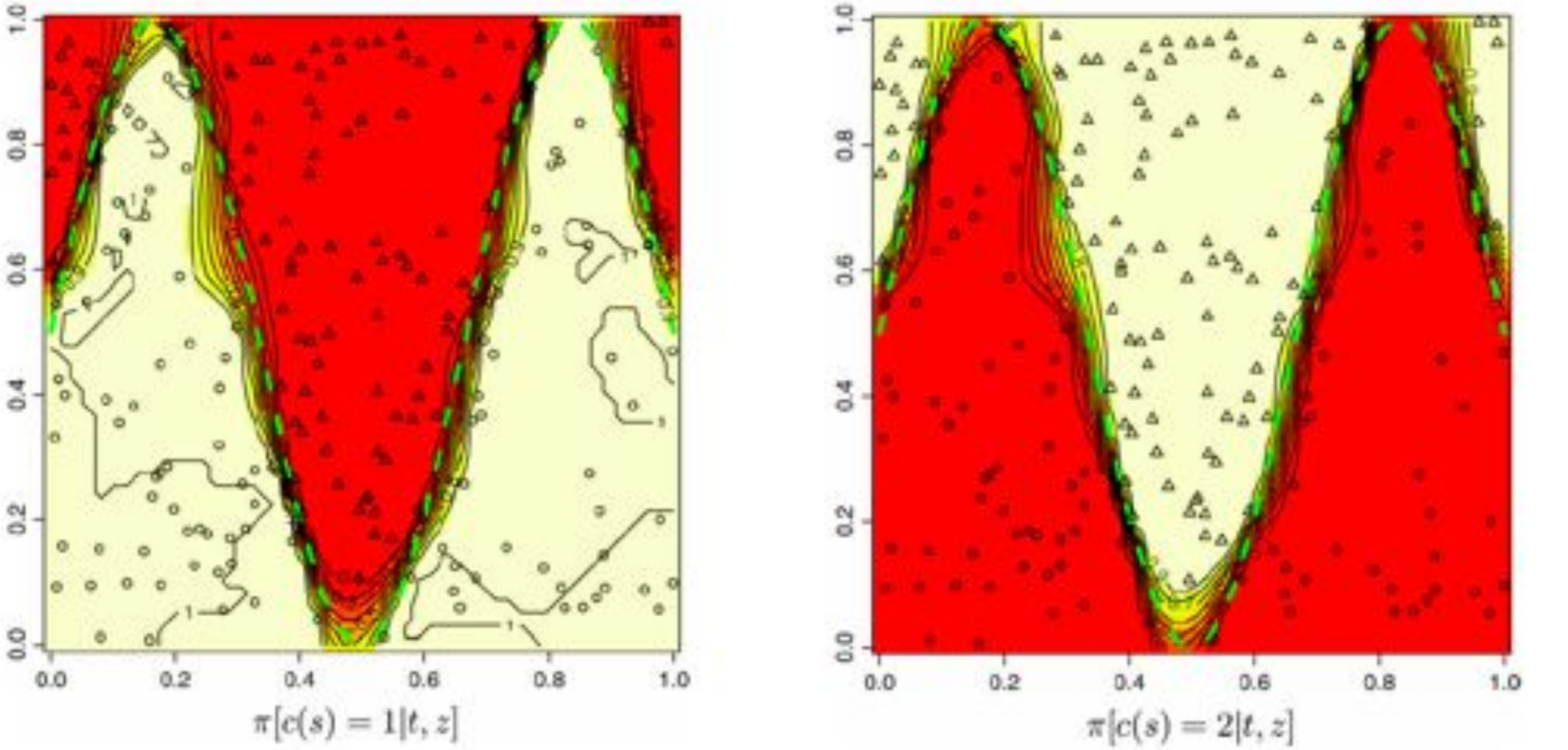


FIGURE 7.—Maps of posterior probabilities, simulated data set A. The dashed green line depicts the true sine-shaped line of discontinuity. $F_{ST} = 0.16$, $L = J_{t=1,\ldots,L} = 10$.

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

GENELAND : simulation test spatial cluster delimitation

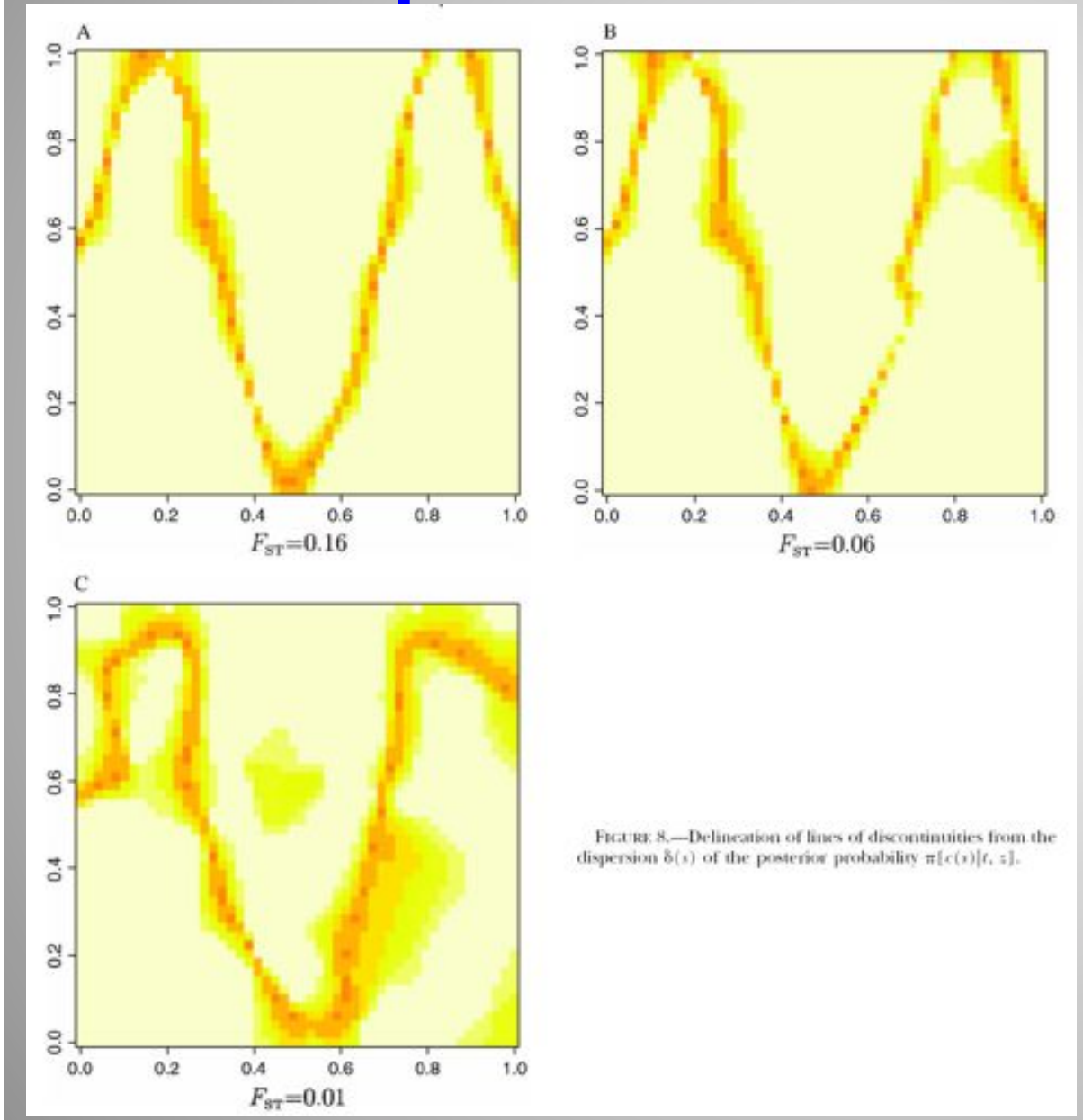


FIGURE 8.—Delineation of lines of discontinuities from the dispersion $\delta(s)$ of the posterior probability $\pi[c(s)|t, z]$.

less and less precision when genetic differentiation decreases

GENELAND : simulation test immigrant detection

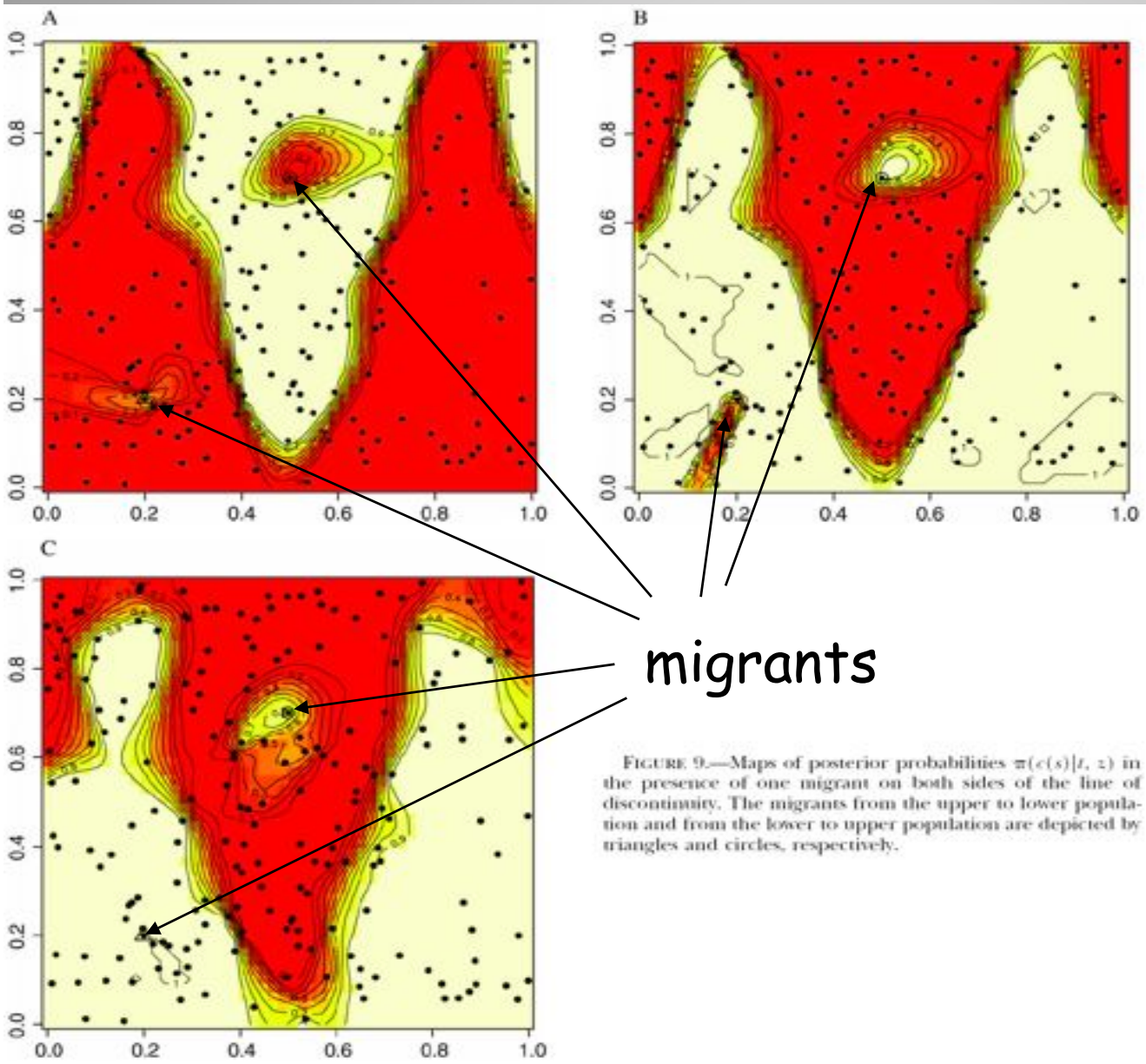


FIGURE 9.—Maps of posterior probabilities $\pi(c(s)|t, z)$ in the presence of one migrant on both sides of the line of discontinuity. The migrants from the upper to lower population and from the lower to upper population are depicted by triangles and circles, respectively.

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

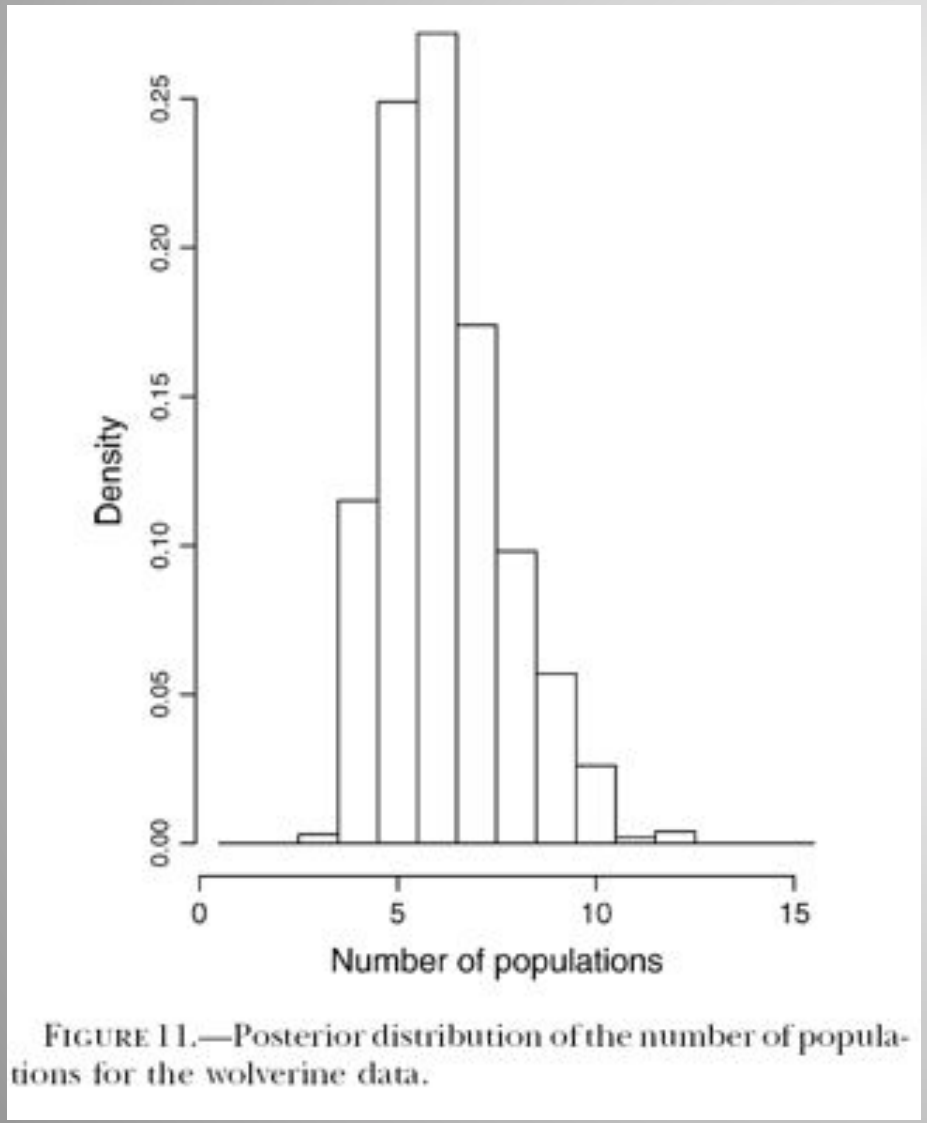
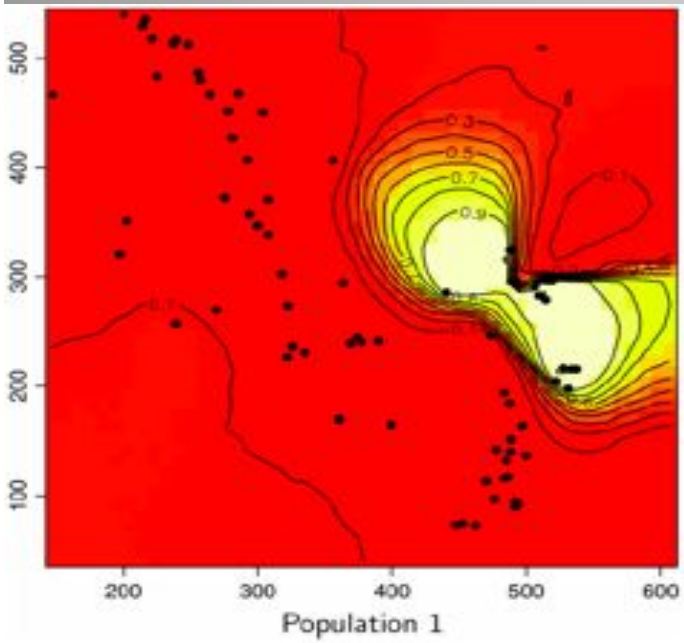
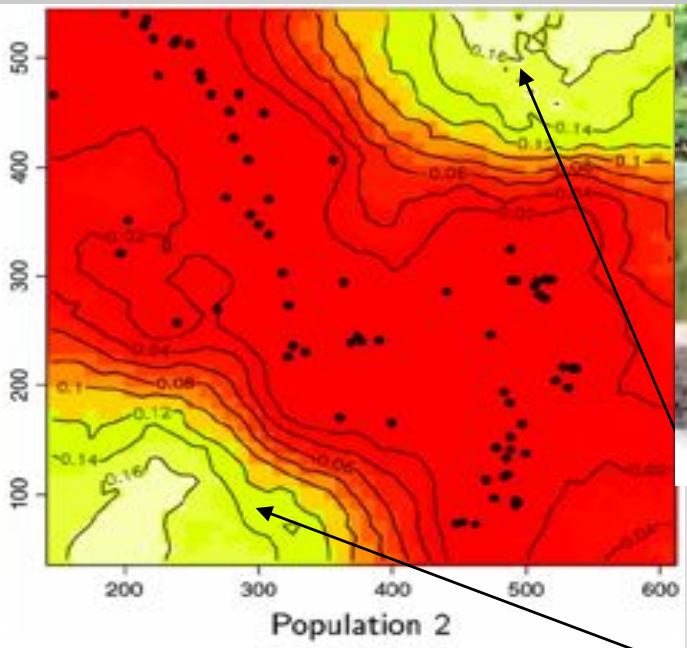
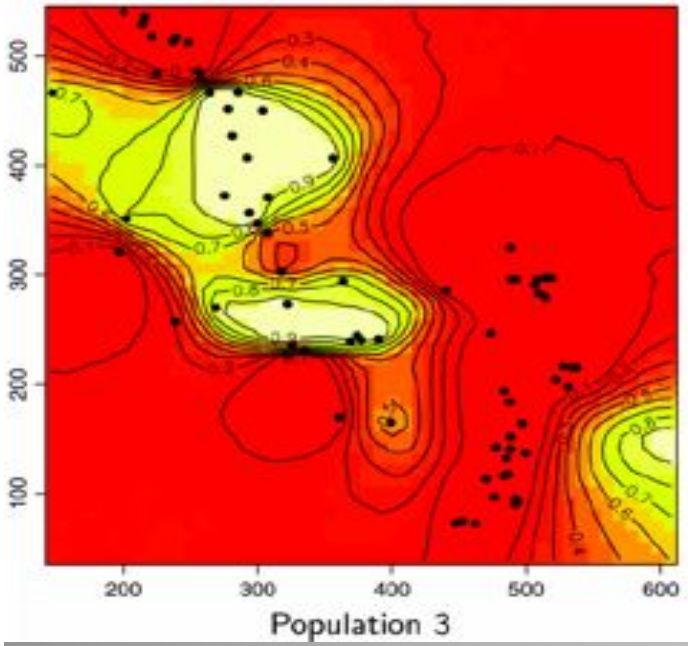


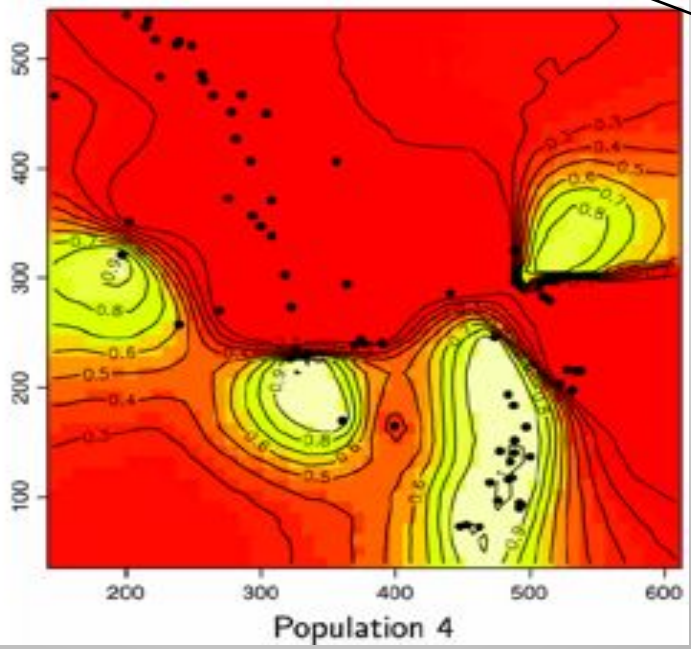
FIGURE 11.—Posterior distribution of the number of populations for the wolverine data.

GENELAND : test on a real data set









Ghost population : does not contain any individual !

GENELAND : test on a real data set

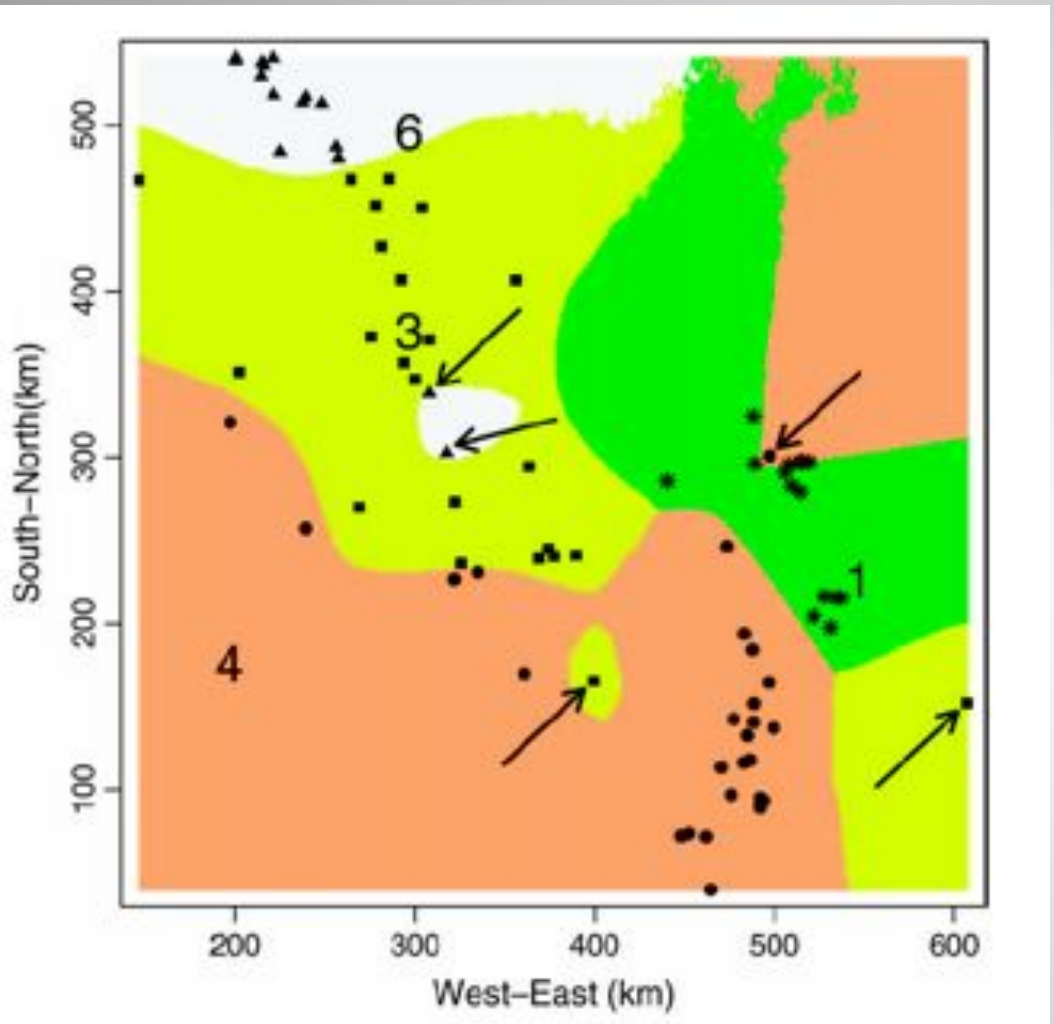


FIGURE 13.—Map of the mode of the posterior probability to belong to each class for the wolverine data. Large character numbers indicate population labels. Arrows indicate putative migrants.

spatial delimitation of 6 genetic clusters
detection of 5 migrants

GENELAND : test on a real data set

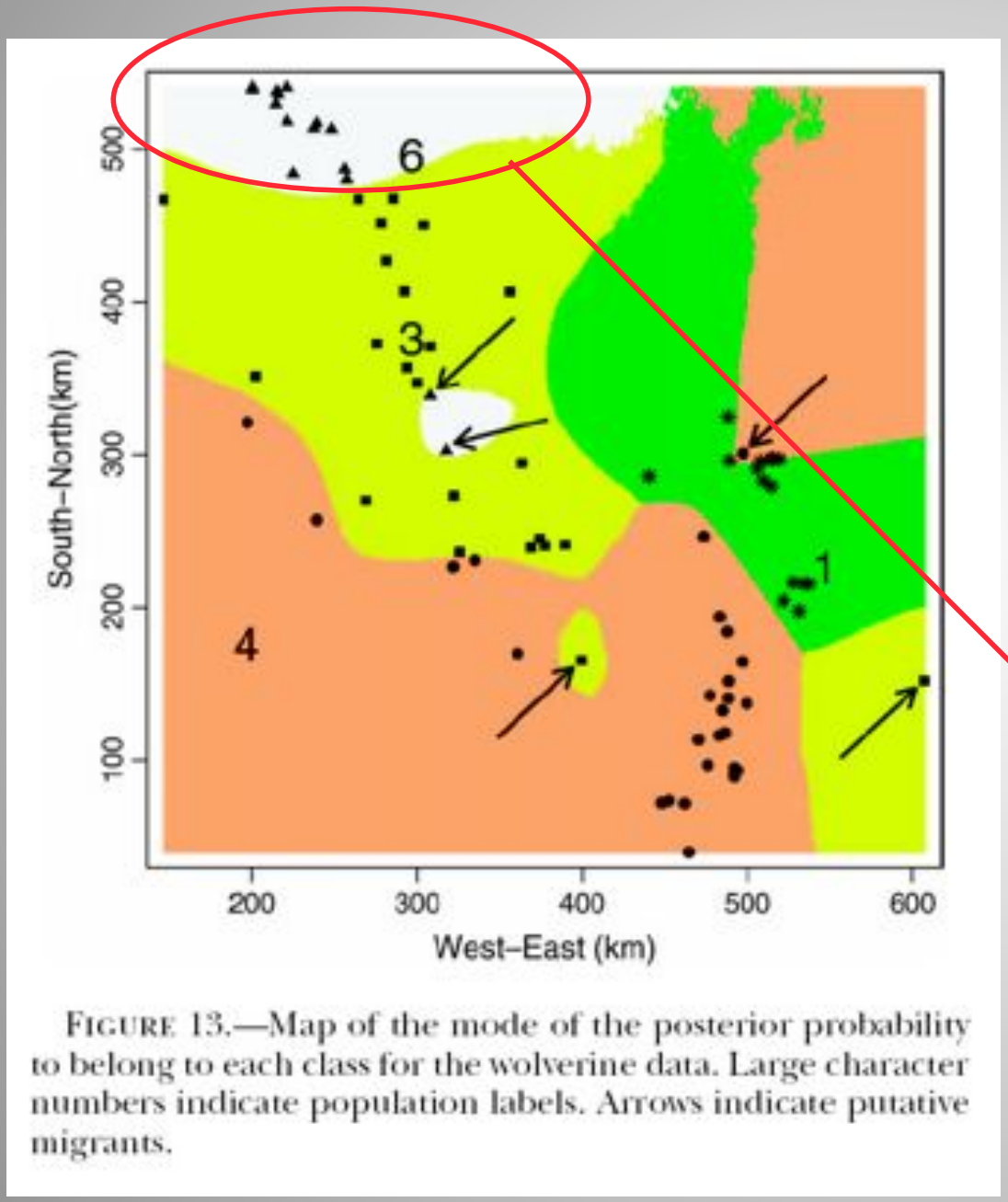


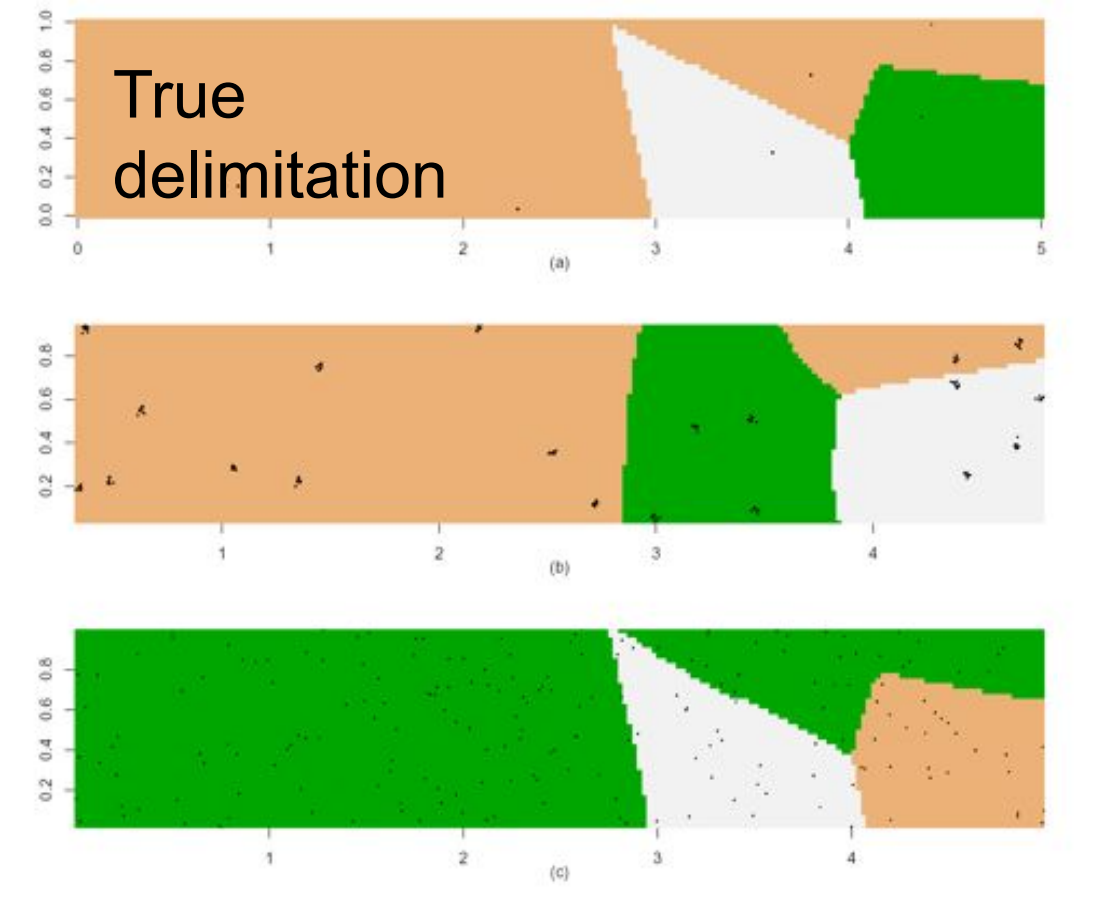
FIGURE 13.—Map of the mode of the posterior probability to belong to each class for the wolverine data. Large character numbers indicate population labels. Arrows indicate putative migrants.

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 ?



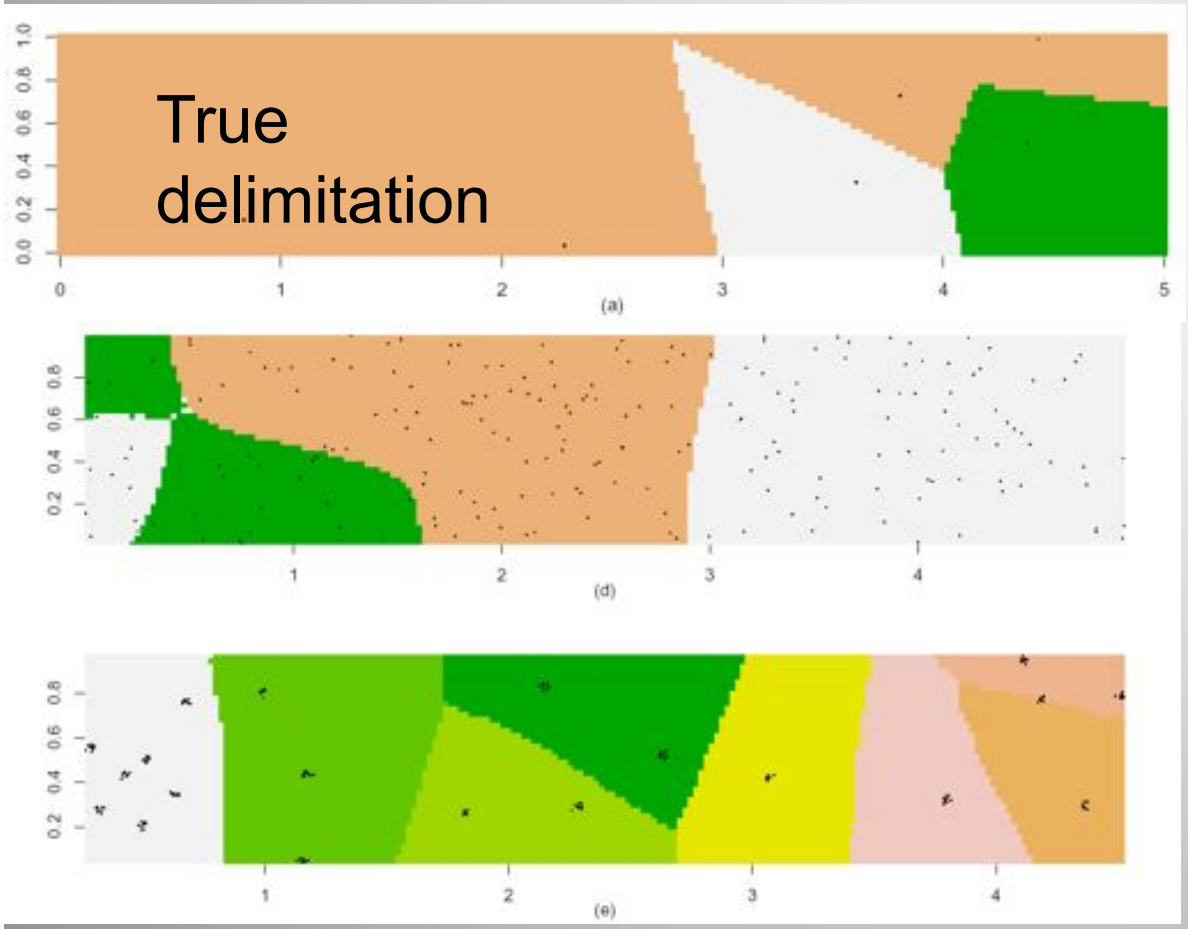
Results are intuitive:

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