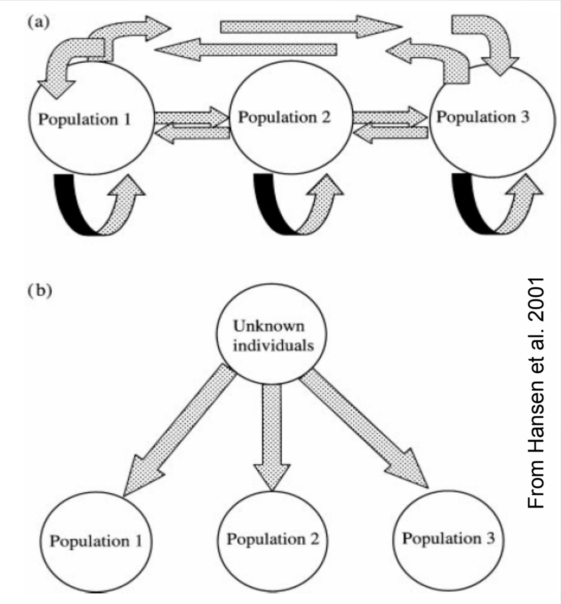
Clustering algorithms for individual multilocus genotypes

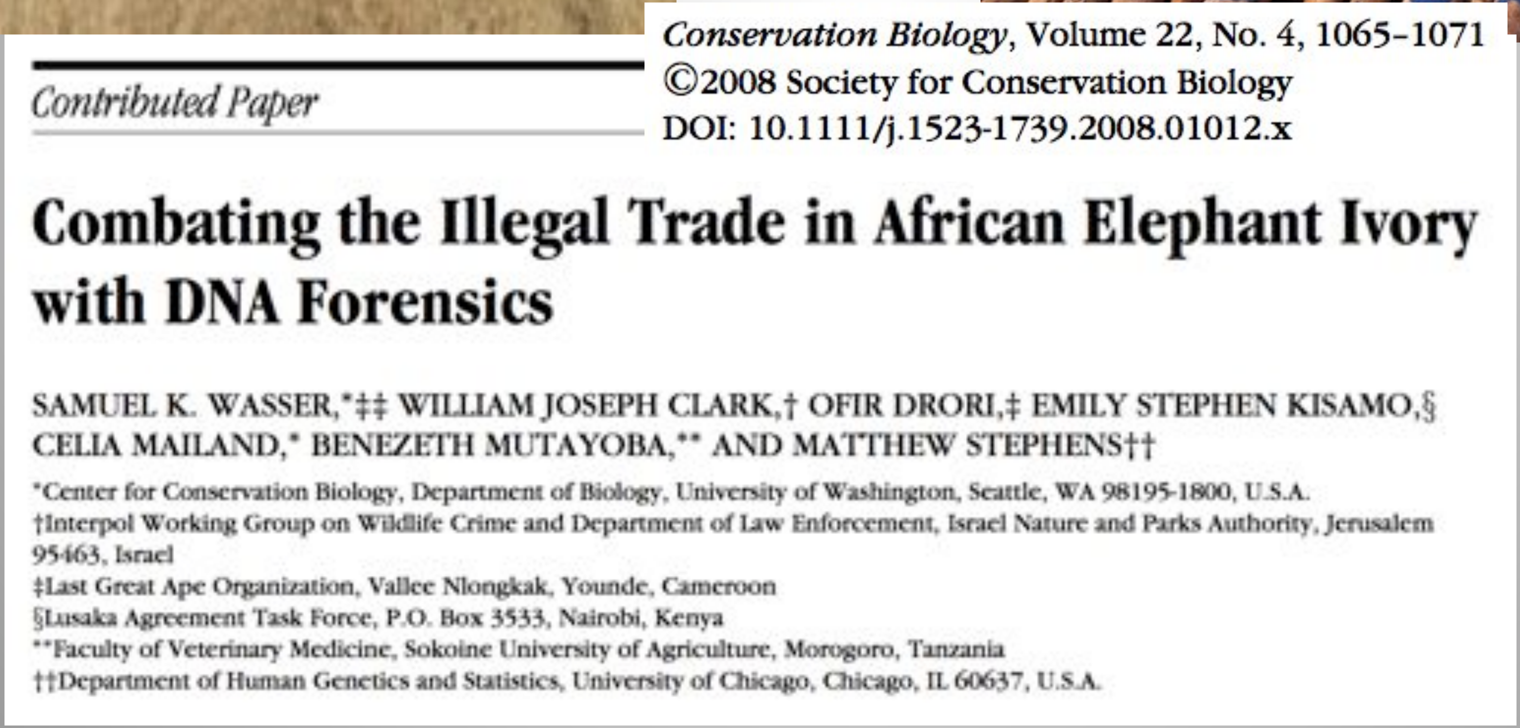


From Hansen et al. 2001

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

Master B2E, December 2014

Ex1 : Provenance de défenses d'éléphants saisies?



Contributed Paper

Conservation Biology, Volume 22, No. 4, 1065–1071
©2008 Society for Conservation Biology
DOI: 10.1111/j.1523-1739.2008.01012.x

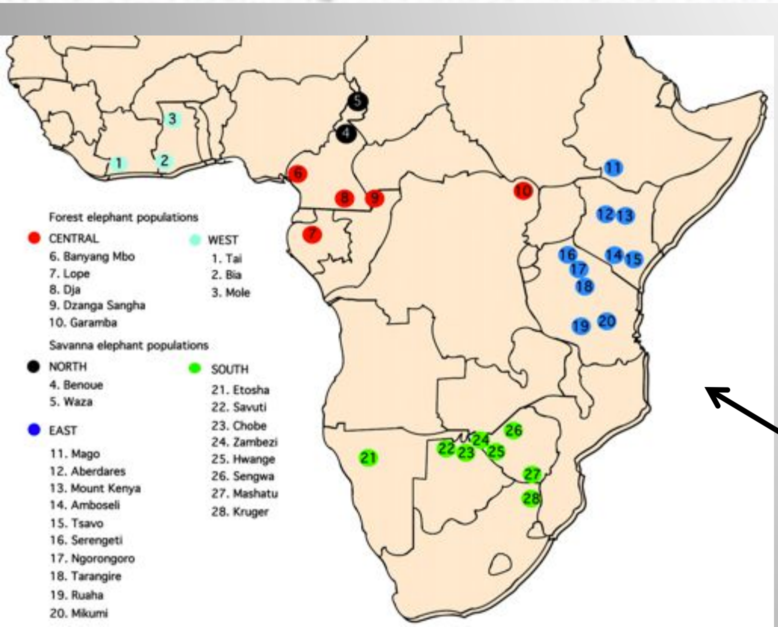
Combating the Illegal Trade in African Elephant Ivory with DNA Forensics

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



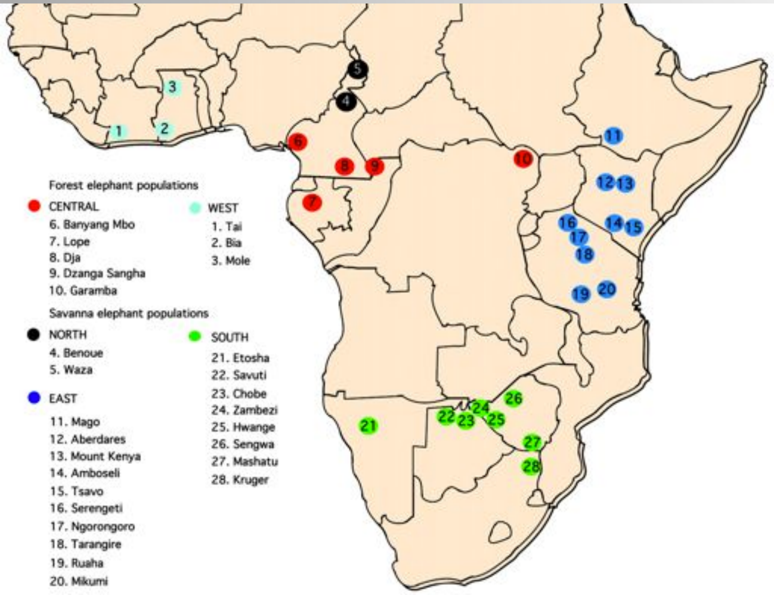
37 tusks from a large seizure in Singapore

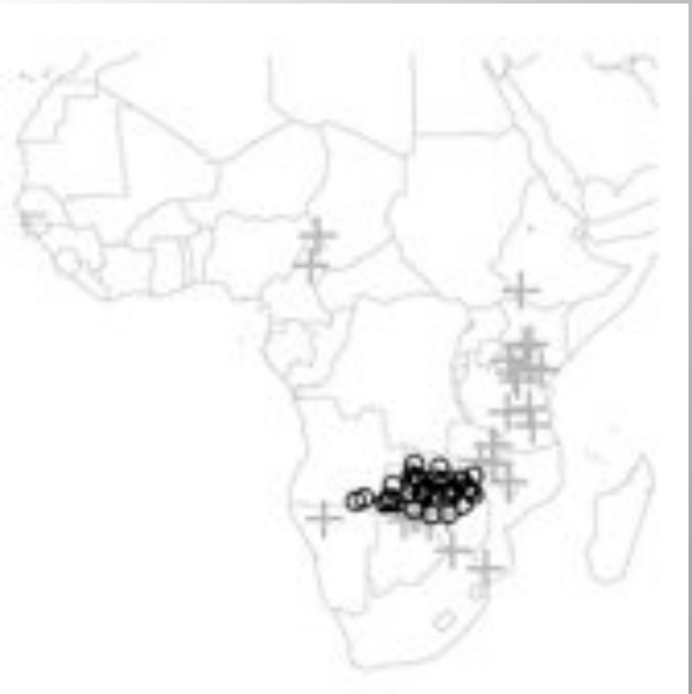
Where are they coming from?

Known populations, genetically characterized

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

Estimated locations of the tusks :

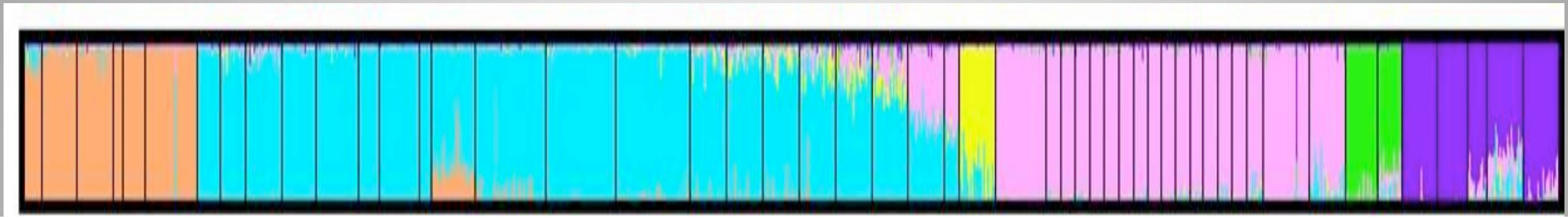


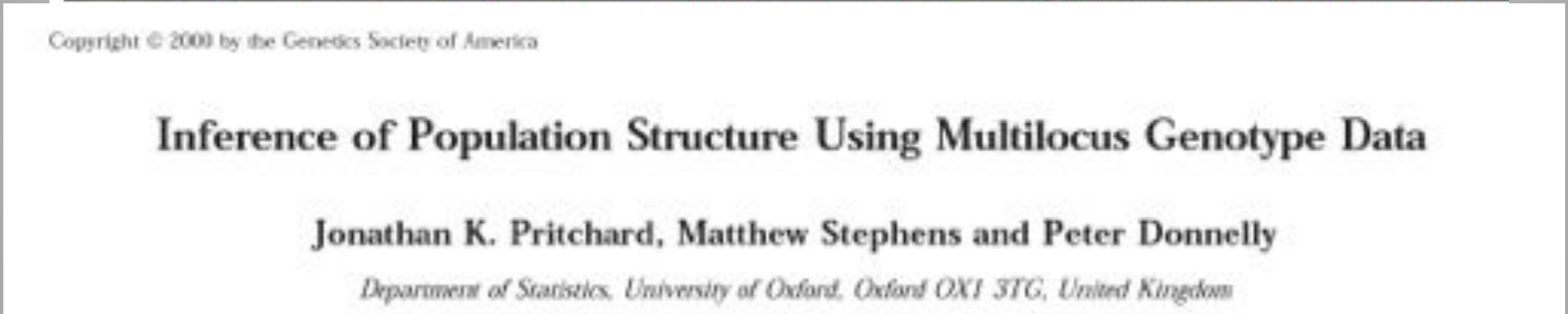


Biological questions

- 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

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

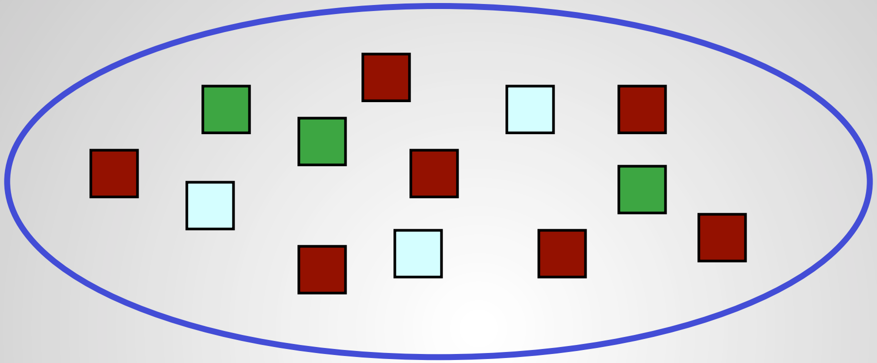
Same assumptions than for assignment methods:

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: Modeling cluster of origin
Model 1



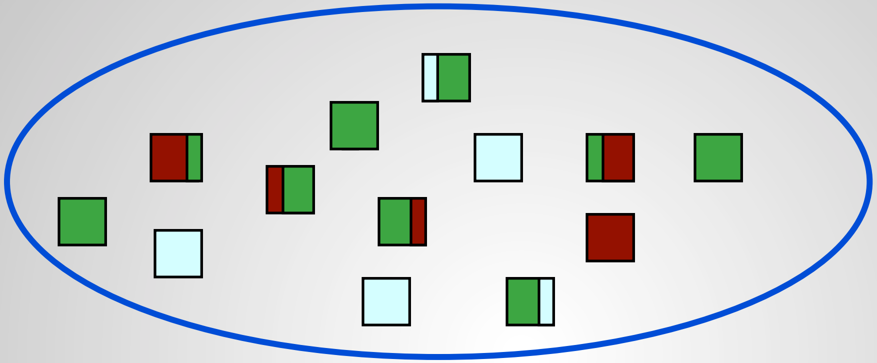
MODEL 1= ("basic") model : 'without admixture'

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: Modeling cluster of origin
Model 2



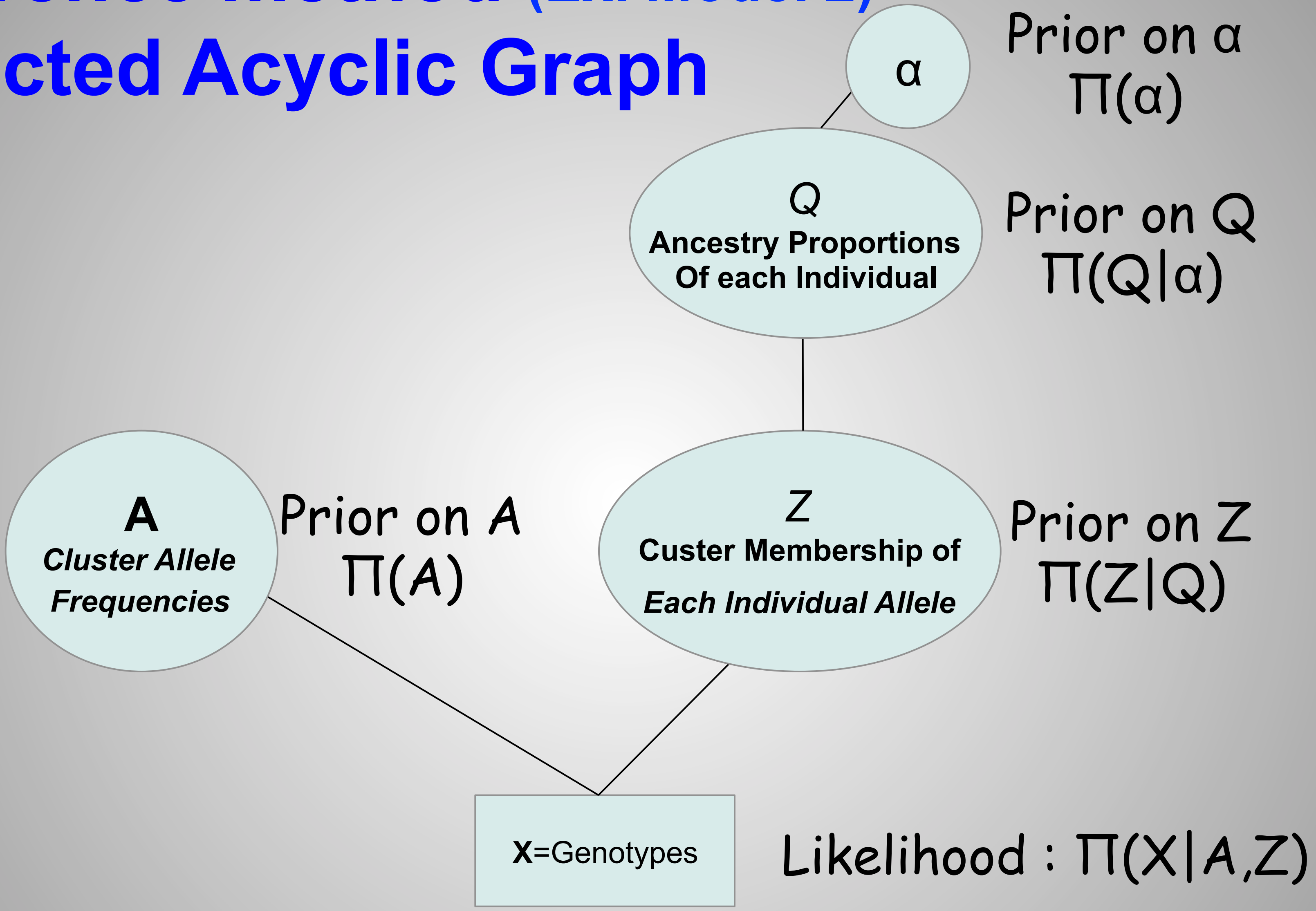
MODEL 2 = model with 'admixture' (most commonly used)

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

Inference method (Ex. Model 2)
Directed Acyclic Graph



α

Prior on α
Π(α)

Q
Ancestry Proportions Of each Individual

Prior on Q
Π(Q|α)

A
Cluster Allele Frequencies

Prior on *A*
Π(*A*)

Z
Custer Membership of
Each Individual Allele

Prior on *Z*
Π(*Z*|Q)

X=Genotypes

Likelihood : Π(X|*A*,*Z*)

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

Inference method (Ex. Model 2)

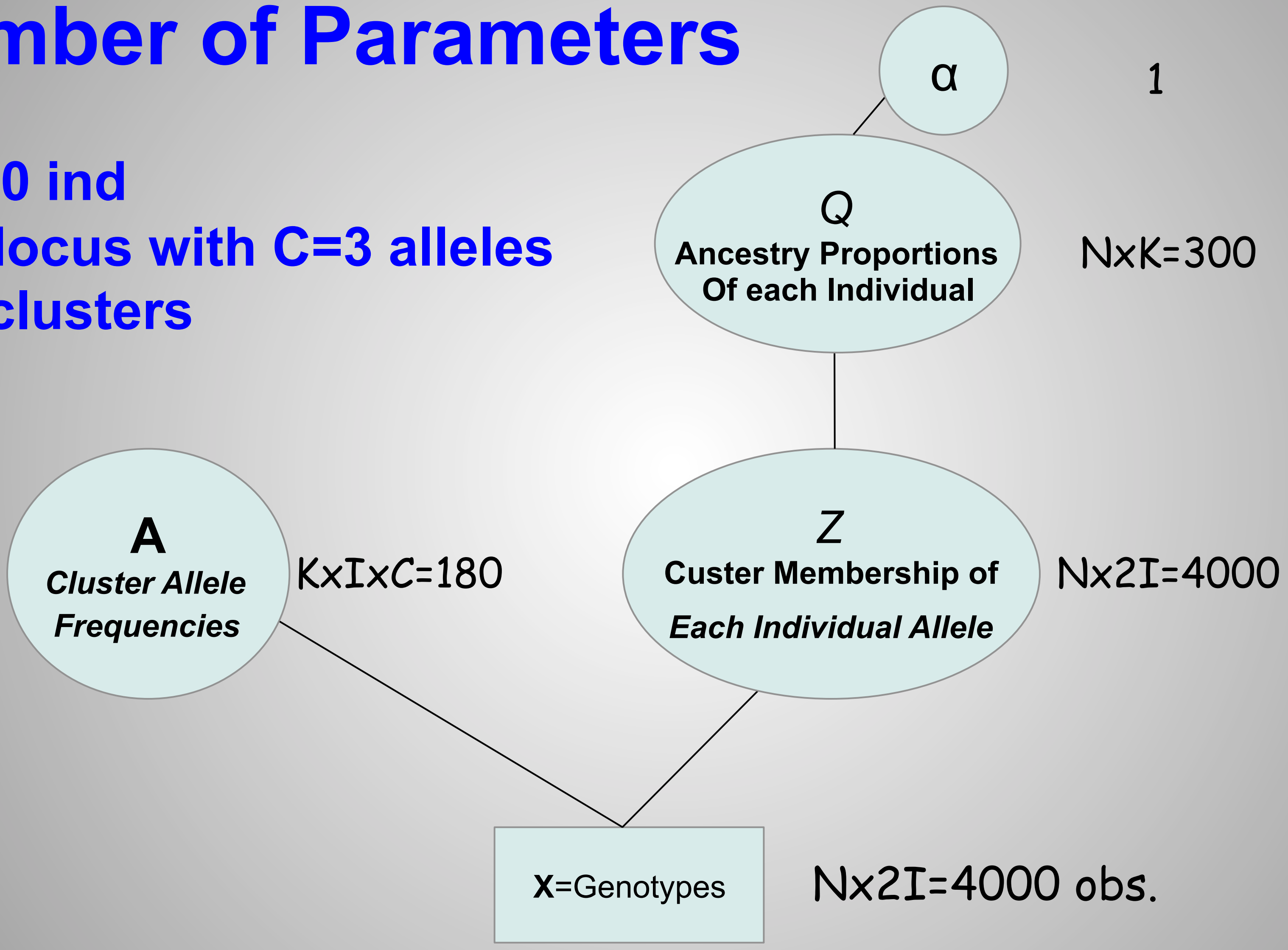
Number of Parameters

Ex.

N=100 ind

I=20 locus with C=3 alleles

K=3 clusters



α

1

Q

Ancestry Proportions Of each Individual

NxK=300

A

Cluster Allele Frequencies

KxIxC=180

Z

Custer Membership of *Each Individual Allele*

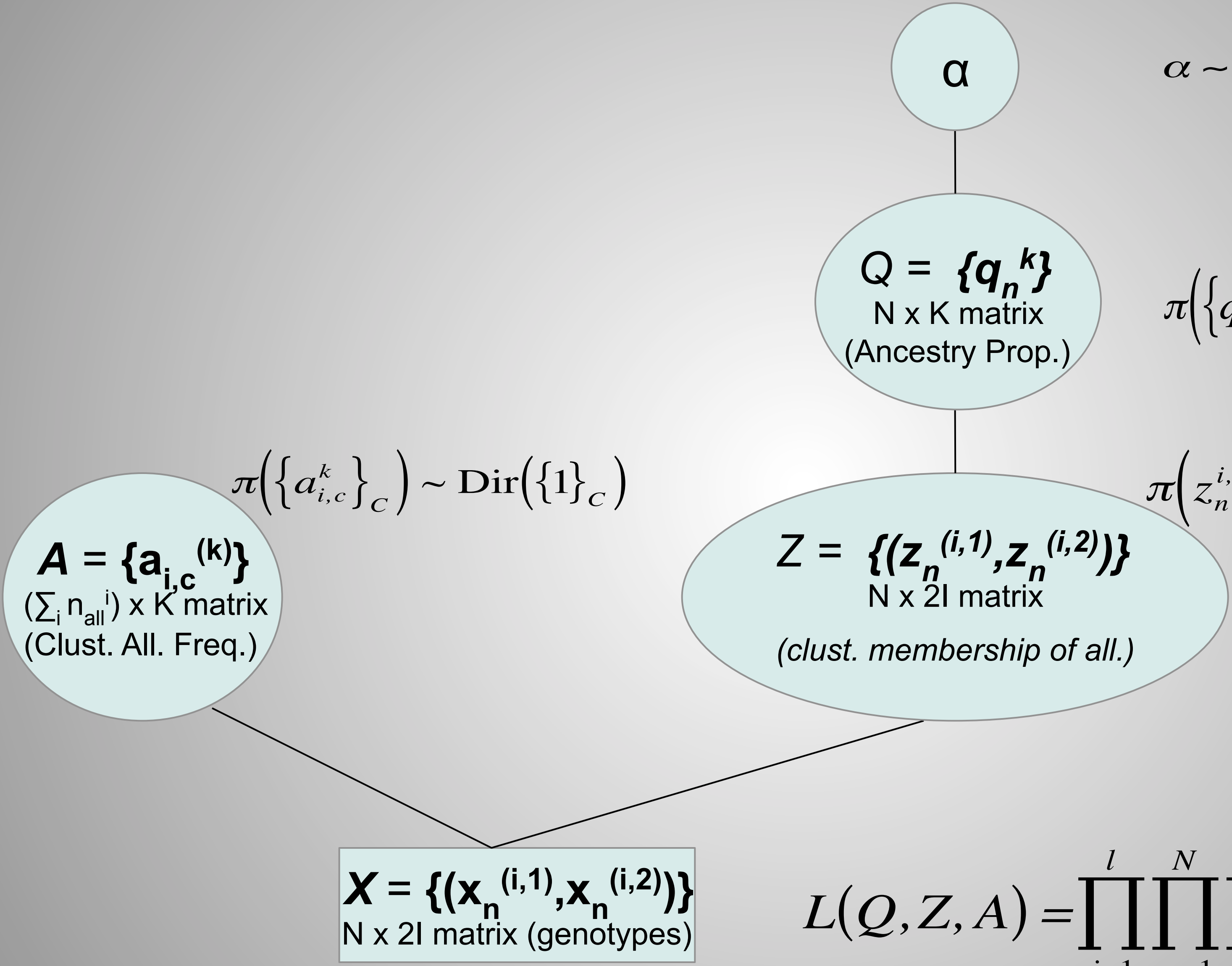
Nx2I=4000

X=Genotypes

Nx2I=4000 obs.

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

Model 2 Specification



α

$\alpha \sim \mathrm{U}[0,10]$

$Q = \{q_n^k\}$
N x K matrix
(Ancestry Prop.)

$\pi\left(\left\{q_n^k\right\}_K\right) \sim \mathrm{Dir}\left(\left\{\alpha\right\}_K\right)$

$\pi\left(\left\{a_{i,c}^k\right\}_C\right) \sim \mathrm{Dir}\left(\left\{1\right\}_C\right)$

$\pi\left(z_n^{i,l} = k \,/\, \left\{q_n^k\right\}_K\right) = q_n^k$

$A = \{a_{i,c}^{(k)}\}$
$(\sum_i n_{all}^i)$ x K matrix
(Clust. All. Freq.)

$Z = \{(z_n^{(i,1)}, z_n^{(i,2)})\}$
N x 2l matrix

(clust. membership of all.)

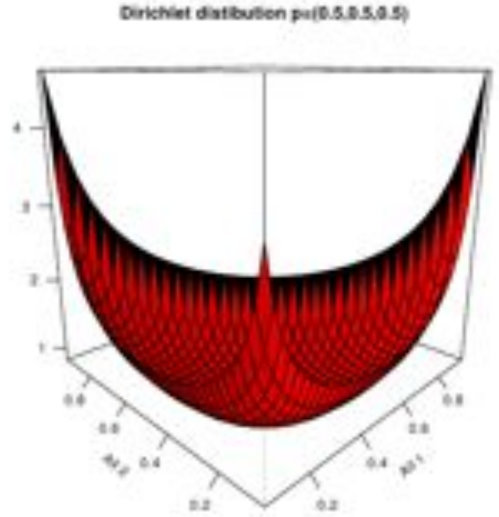
$X = \{(x_n^{(i,1)}, x_n^{(i,2)})\}$
N x 2l matrix (genotypes)

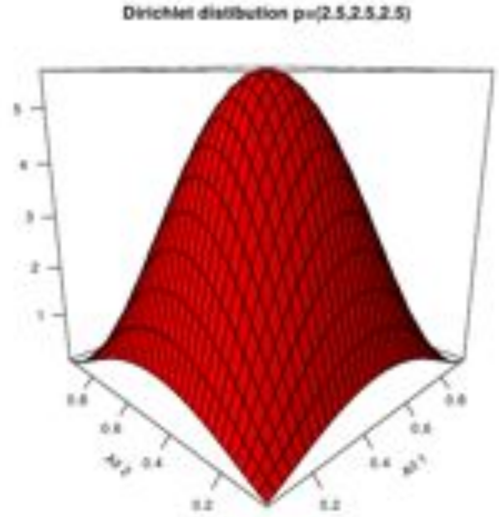
$$L(Q,Z,A) = \prod_{i=1}^{l}\prod_{n=1}^{N}\prod_{l=1}^{2} a_{i,x_n^{i,l}}^{z_n^{i,l}}$$

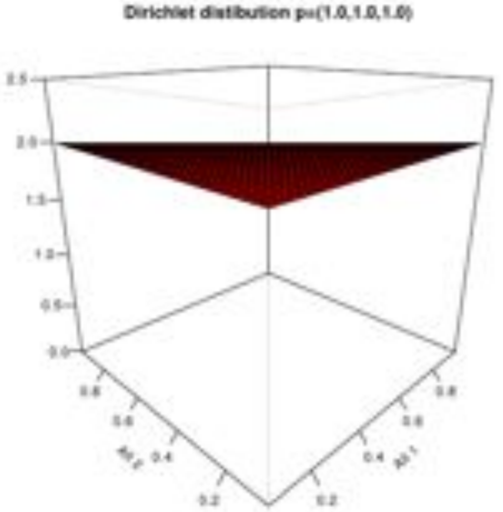
Linkage Equilibrium

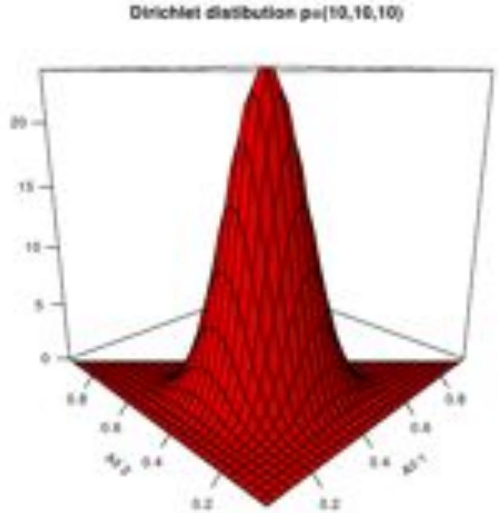
HWE

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)

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

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

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

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

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)

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

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)

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

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

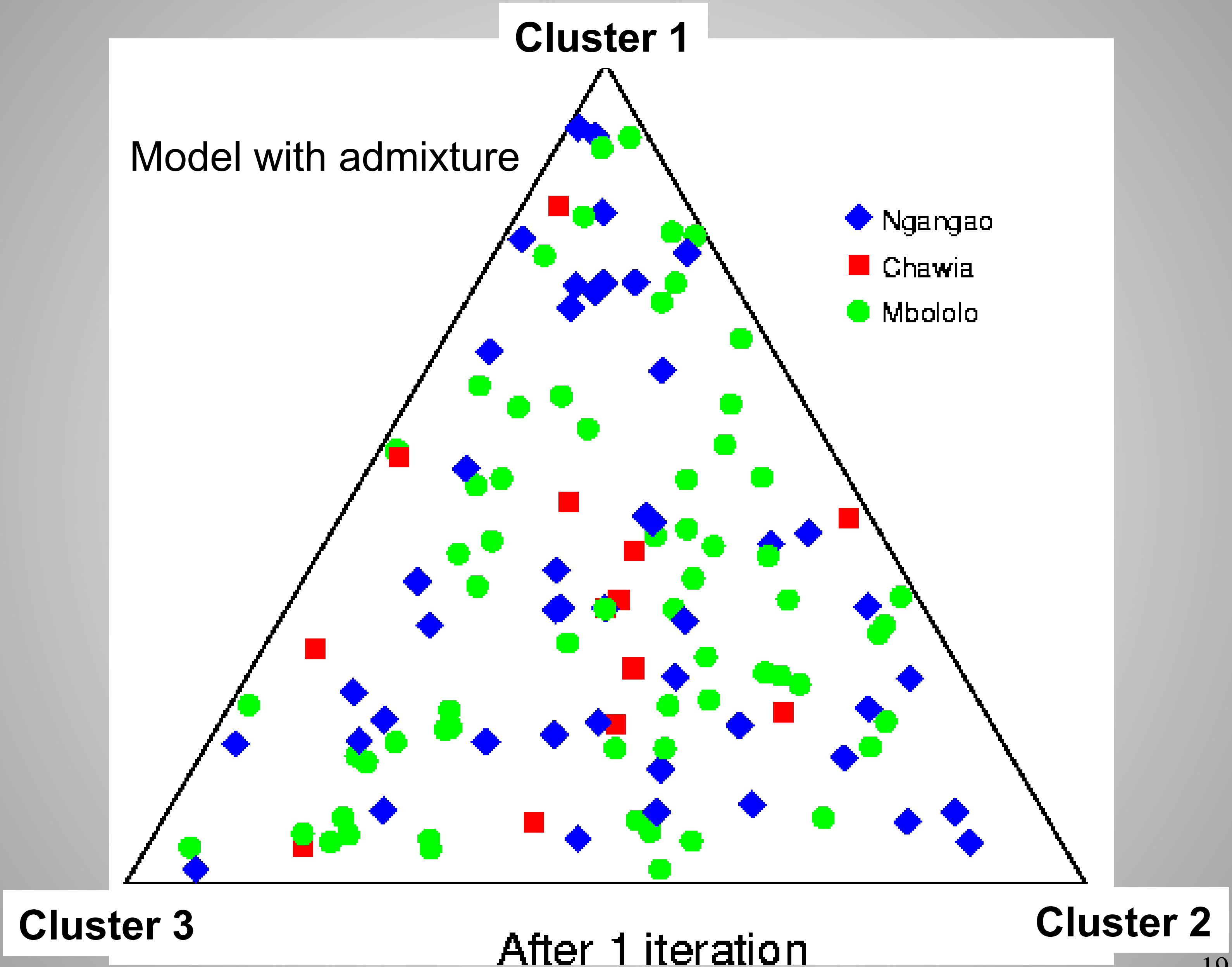
=>not of usual form

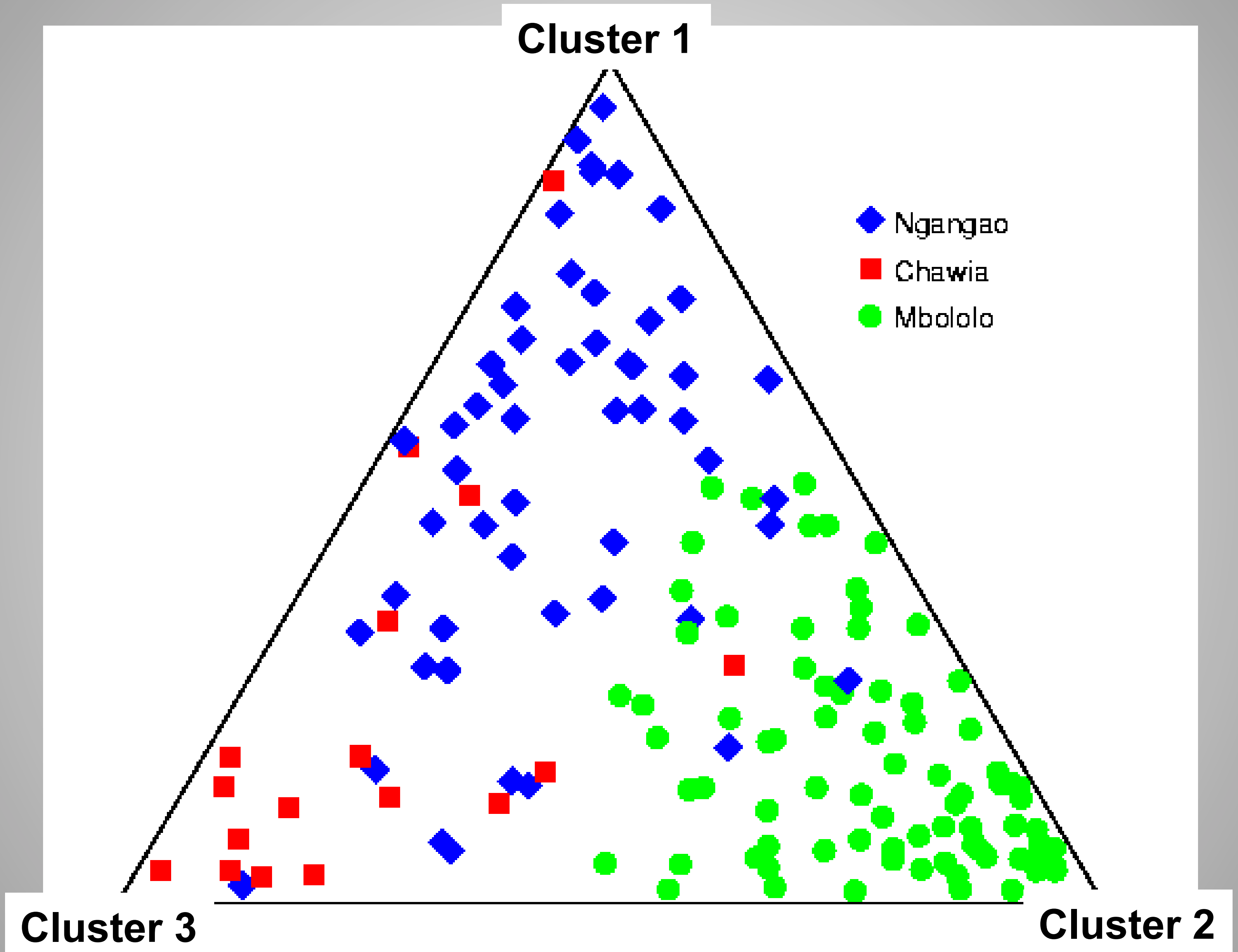
=> 1 step (propose a value and accept/refuse it according to MH rule...)

Example: Taita Thrush data

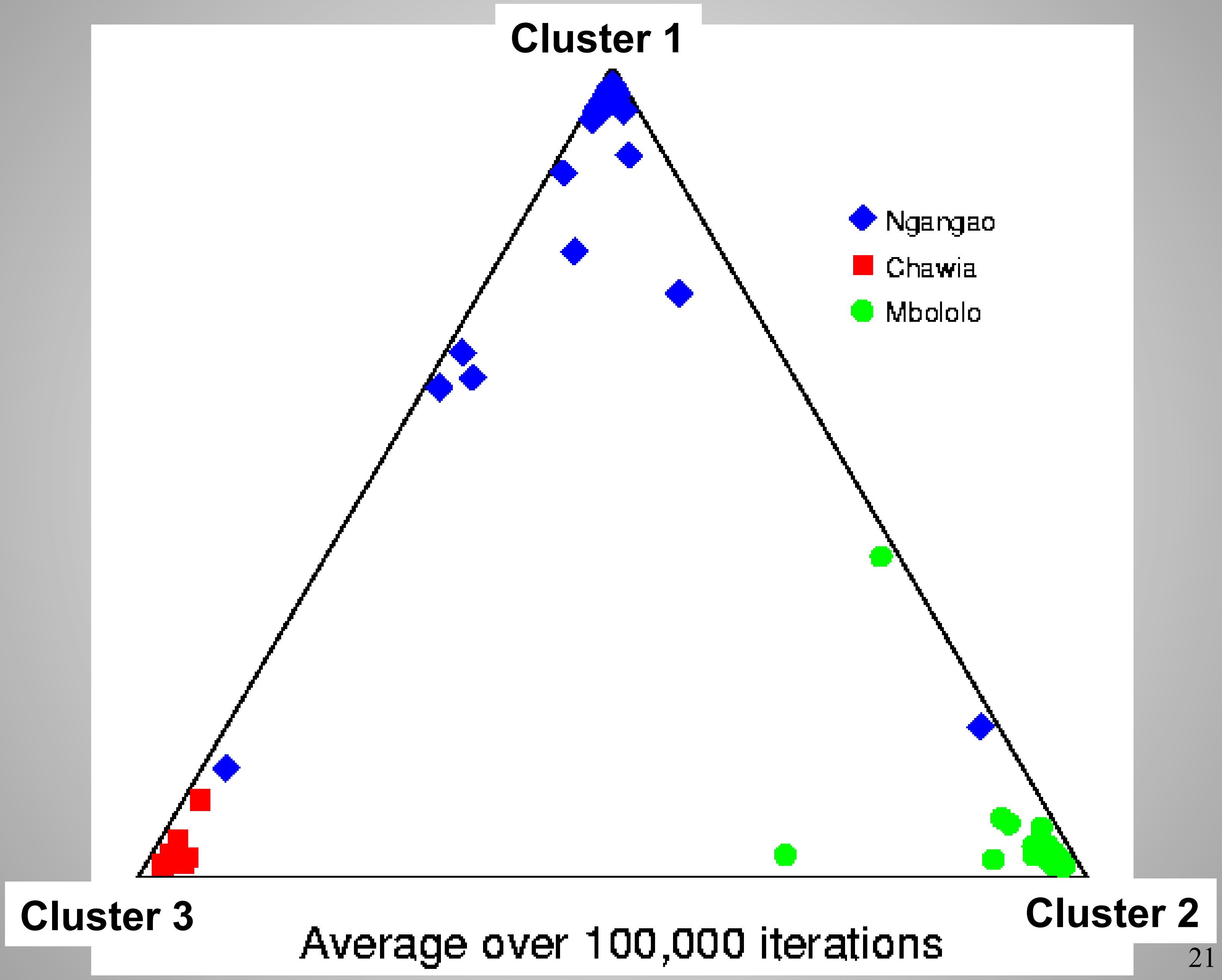
- three main sampling locations in Kenya
- low migration rates (radio-tagging study))
- 155 individuals, genotyped at 7 microsatellite loci

*Data courtesy of Dr Peter Galbusera


Cluster 1
Model with admixture
◆ Ngangao
■ Chawia
● Mbololo
Cluster 3
Cluster 2
After 1 iteration



After 25 iterations



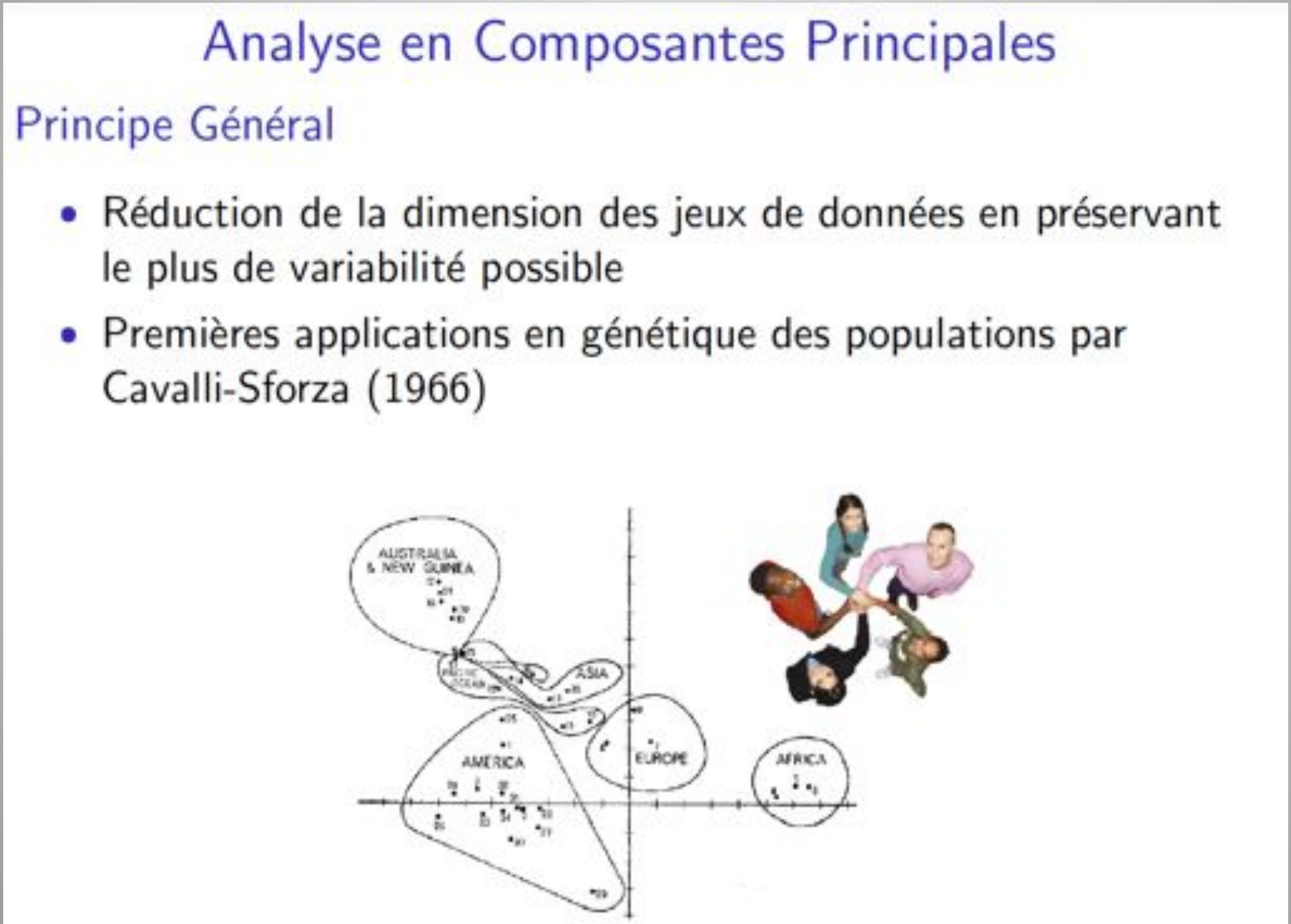
Average over 100,000 iterations

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.

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)

Une approche courante (Patterson *et al.*, 2006, Plos Gen.)

Normalisation des données

- Soit $p_j = \frac{1}{2I} \sum_{i=1}^{I} x_{ij}$ la fréquence de l'allèle de réference pour le SNP j dans l'ensemble des individus
- Soit $\mathbf{M} = \{m_{i,j}\}$ la matrice $\mathbf{X}$ normalisée avec $m_{i,j} = \frac{x_{i,j} - 2p_j}{\sqrt{p_j(1-p_j)}}$
- $\Rightarrow$ poids similaire à tous les SNPs ($\sim$ variance similaire sur chaque colonne)
- Rq : Les données manquantes pour le SNP j peuvent être mises à $2p_j$

ACP (en théorie)

- Décomposition spectrale : $\mathbf{S} = \frac{1}{J-1}\mathbf{M}'\mathbf{M} = \sum_{k=1}^{r} \lambda_k \mathbf{a_k} \mathbf{a_k'}$

  $\lambda_k = k^{ième}$ valeurs propres ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$) et $\mathbf{a_k}$ = vecteur propre associé (avec $\mathbf{a_k'}\mathbf{a_k} = 1$)
- $I \ll J \Rightarrow r = I - 1$ valeurs propres non nulles (Rq : si doublons parfaits $r < I - 1$)

  Si K populations, $K - 1$ axes correspondants à chaque population ("significatifs"), et $I - K$ axes correspondant à du bruit intra-pop
- Coordonnées principales : $\mathbf{Z} = \mathbf{MA}$ (où $\mathbf{A} = \{\mathbf{a_k}\}$)

ACP en pratique (1)

Rappels

- Soit $\mathbf{X}$ une matrice $n \times p$ telle que $\mathbf{X'X}$ soit de rang $r < p < n$ $\Rightarrow r$ val. prop. non nulles et $p - r$ val. prop. nulles
- Alors $\mathbf{XX'}$ a les mêmes r val. prop. non nulles et $n - r$ val. prop. nulles $(\mathbf{XX'c_l} = \eta_l \mathbf{c_l} \Rightarrow (\mathbf{X'X})\, \mathbf{X'c_l} = \eta_l\, (\mathbf{X'c_l}))$
- Si $\mathbf{c_l} = l^{ième}$ vect. prop. de $\mathbf{X'X}$ alors $\mathbf{d_l} = \frac{1}{\sqrt{\eta_l}} \mathbf{X'c_l}$ est le $l^{ième}$ vect. prop. de $\mathbf{XX'}$ (soumis à $\mathbf{d_l'd_l} = 1$)

Utilisation du principe de dualité

- $\mathbf{S} = \frac{1}{J-1}\mathbf{M'M}$ de dim JxJ ($t_c \propto J^2 I$) alors que $\mathbf{T} = \frac{1}{I-1}\mathbf{MM'}$ de dim IxI : ($t_c \propto I^2 J$)
- Décomposition spectrale de : $\mathbf{T} = \sum_{k=1}^{r} \tau_k \mathbf{b_k b_k'}$
- $\lambda_k = \frac{I-1}{J-1}\tau_k$ (*smartpca donne* : $\lambda_k^{(std)} = (I-1)\lambda_k / \sum_{k=1}^{r} \lambda_k$)
  $(\frac{1}{I-1}\mathbf{MM'b_k} = \tau_k \mathbf{b_k} \Rightarrow \mathbf{M'M}\,(\mathbf{M'b_k}) = (I-1)\tau_k(\mathbf{M'b_k}) \Rightarrow \frac{1}{J-1}\mathbf{M'M}\,(\mathbf{M'b_k}) = \frac{I-1}{J-1}\tau_k(\mathbf{M'b_k}))$
- $\mathbf{a_k} = \frac{1}{\sqrt{(I-1)\tau_k}}\mathbf{M'b_k}$
  $(\mathbf{a_k} = \alpha \mathbf{M'b_k}$ avec $1 = \mathbf{a_k'a_k} = \alpha^2 \mathbf{b_k'MM'b_k} = \alpha^2 \tau_k (I-1)\mathbf{b_k'b_k} = \alpha^2\tau_k(I-1)$ d'où $\alpha = \frac{1}{\sqrt{(I-1)\tau_k}})$

ACP en pratique (2)

Calcul des PCs

- $\mathbf{Z} = \mathbf{MA}$ et $\mathrm{var}(\{z_k\}) = \tau_k$
- Il est courant de représenter des PCs "standardisés" $\tilde{z}_{i,k} = \frac{z_{i,k}}{\sqrt{(I-1)\tau_k}}$

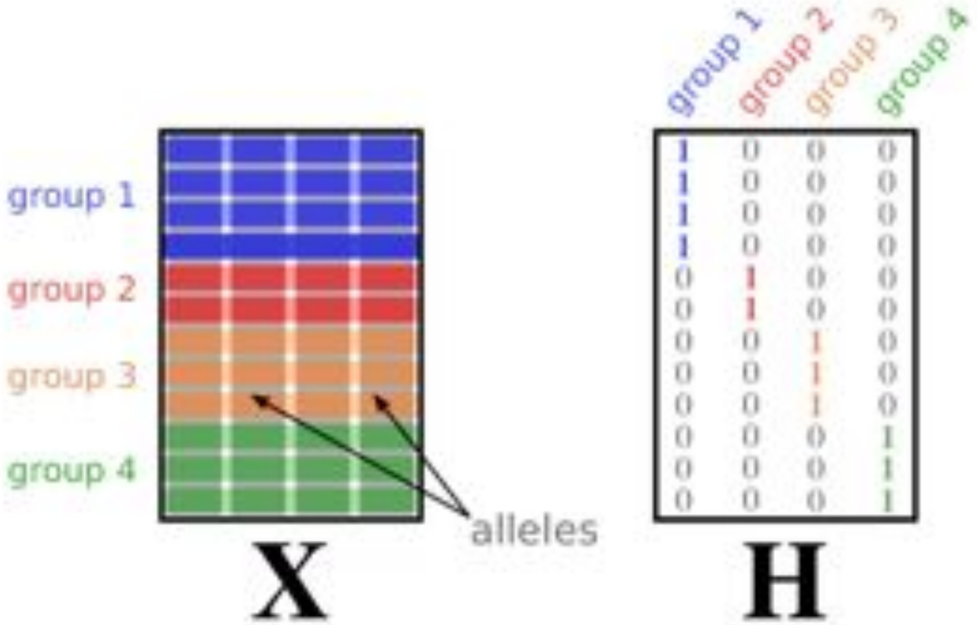
Décomposition en Valeurs Singulières

- $\mathbf{M} = \mathbf{ULA}'$
  - $\mathbf{U}$ = vecteur propres de $\mathbf{MM}'$ (dim=$I \times I - 1$)
  - $\mathbf{A}$ = vecteur propres de $\mathbf{M}'\mathbf{M}$ (dim=$J \times I - 1$)
  - $\mathbf{L}$ = matrice diagonale $l_{i,i} = \sqrt{\tau_k}$ (dim=$I - 1 \times I - 1$)
- $\mathbf{Z} = \mathbf{MA} = \mathbf{ULA}'\mathbf{A} = \mathbf{UL}$ ou $\mathbf{U} = \mathbf{ZL}^{-1}$

Applications/Outils

- Avec R : Utilisation des fonctions *eigen* et *svd* du package *base*
- Packages *ade4* et *adegenet* qui offrent une grande versatilité
- Programme *smartpca* (suite *eigenstrat*, Patterson et al.)

Partitionnement du jeu de données



$$\mathbf{H} = \begin{array}{cccc}
\text{group 1} & \text{group 2} & \text{group 3} & \text{group 4} \\
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1
\end{array}$$

Décomposition : $\mathbf{X} = \mathbf{XP} + (\mathbf{I} - \mathbf{P})\mathbf{X}$

- Projecteur $\mathbf{P} = \mathbf{H}\left(\mathbf{H}^{\mathsf{T}}\mathbf{D}\mathbf{H}\right)^{-1}\mathbf{H}^{\mathsf{T}}\mathbf{D}$ (e.g. D=matrice diagonale de rang n telle que $d_{ii} = \frac{1}{n}$)
- $\mathbf{XP} \Longleftrightarrow$ matrice $\mathbf{X}$ où chaque valeur est remplacée par la moyenne intra-groupe (sur chaque colonne). (e.g. $x_{ij} = \frac{1}{g(i)} \sum_{k=1}^{g(i)} x_{kj}$)

Analyses Multivariées sur les différentes partitions

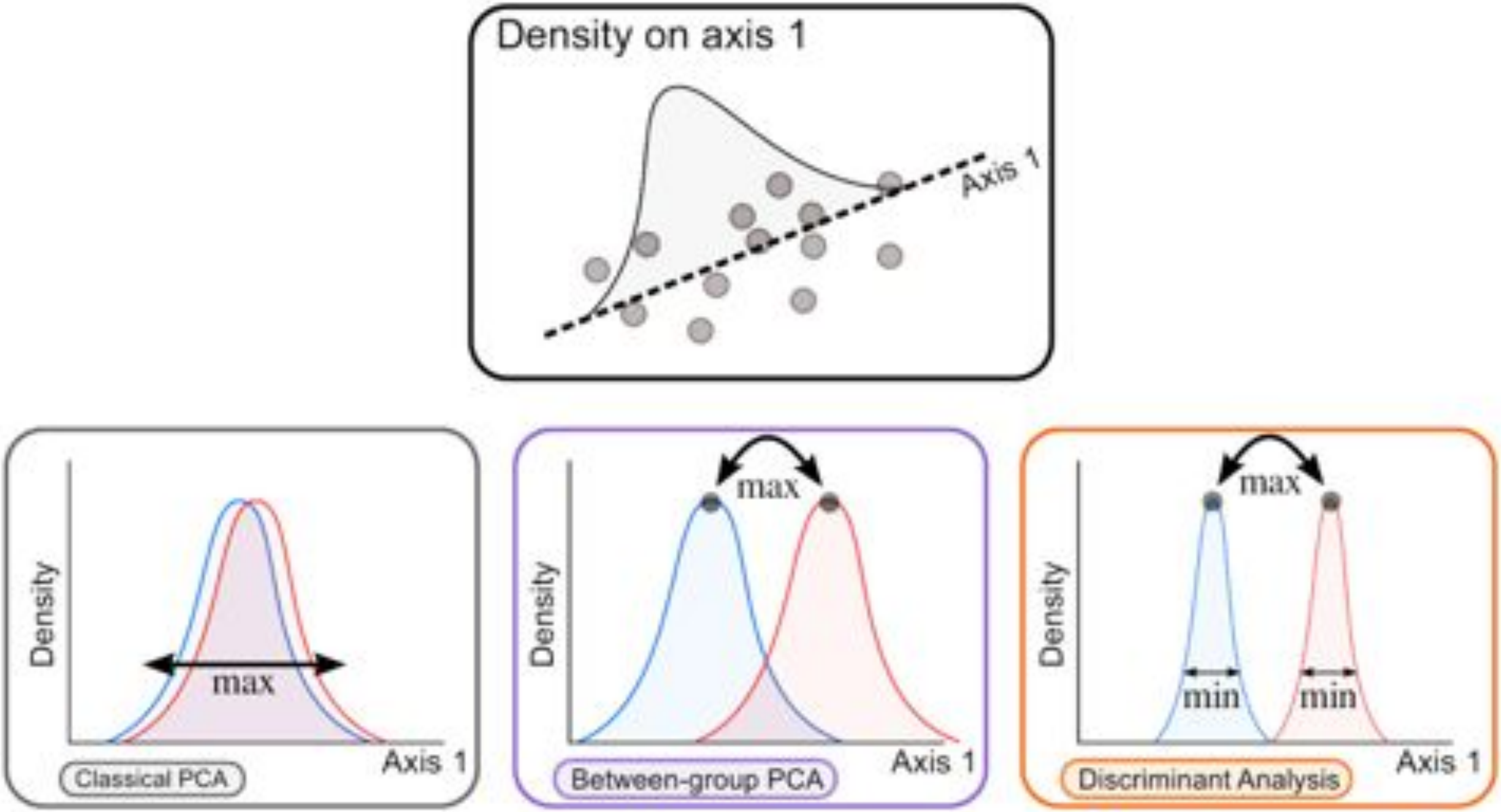
Partionnement de la variance (modèle MANOVA)

$$
\begin{aligned}
\mathrm{V}(\mathbf{X}) &= B(\mathbf{X}) + W(\mathbf{X}) \\
&= \mathrm{V}(\mathbf{XP}) + \mathrm{V}((\mathbf{I}-\mathbf{P})\mathbf{X}) \\
\mathrm{tr}\left(\mathbf{X}^{\mathsf{T}}\mathbf{DX}\right) &= \left[\mathrm{tr}\left(\mathbf{X}^{\mathsf{T}}\mathbf{P}^{\mathsf{T}}\mathbf{DPX}\right)\right] + \left[\mathrm{tr}\left(\mathbf{X}^{\mathsf{T}}(\mathbf{I}-\mathbf{P})^{\mathsf{T}}\mathbf{D}(\mathbf{I}-\mathbf{P})\mathbf{X}\right)\right]
\end{aligned}
$$

Analyses selon contraintes d'optimisation sur les PCs ($\mathbf{Xu}$)

- ACP classique : maximise variabilité globale ($\mathrm{V}(\mathbf{Xu})$)
- ACP between-group : maximise variabilité inter–groupe ($\mathrm{V}(\mathbf{B}(\mathbf{Xu}))$)
- Analyse discriminante (DA) : maximise variabilité inter-groupe ($\mathrm{V}(\mathbf{B}(\mathbf{Xu}))$) et minimise conjointement variabilité intra–groupe ($\mathrm{V}(\mathbf{W}(\mathbf{Xu}))$)
  (⇔ optimise la séparation des individus selon leur groupe d'origine).

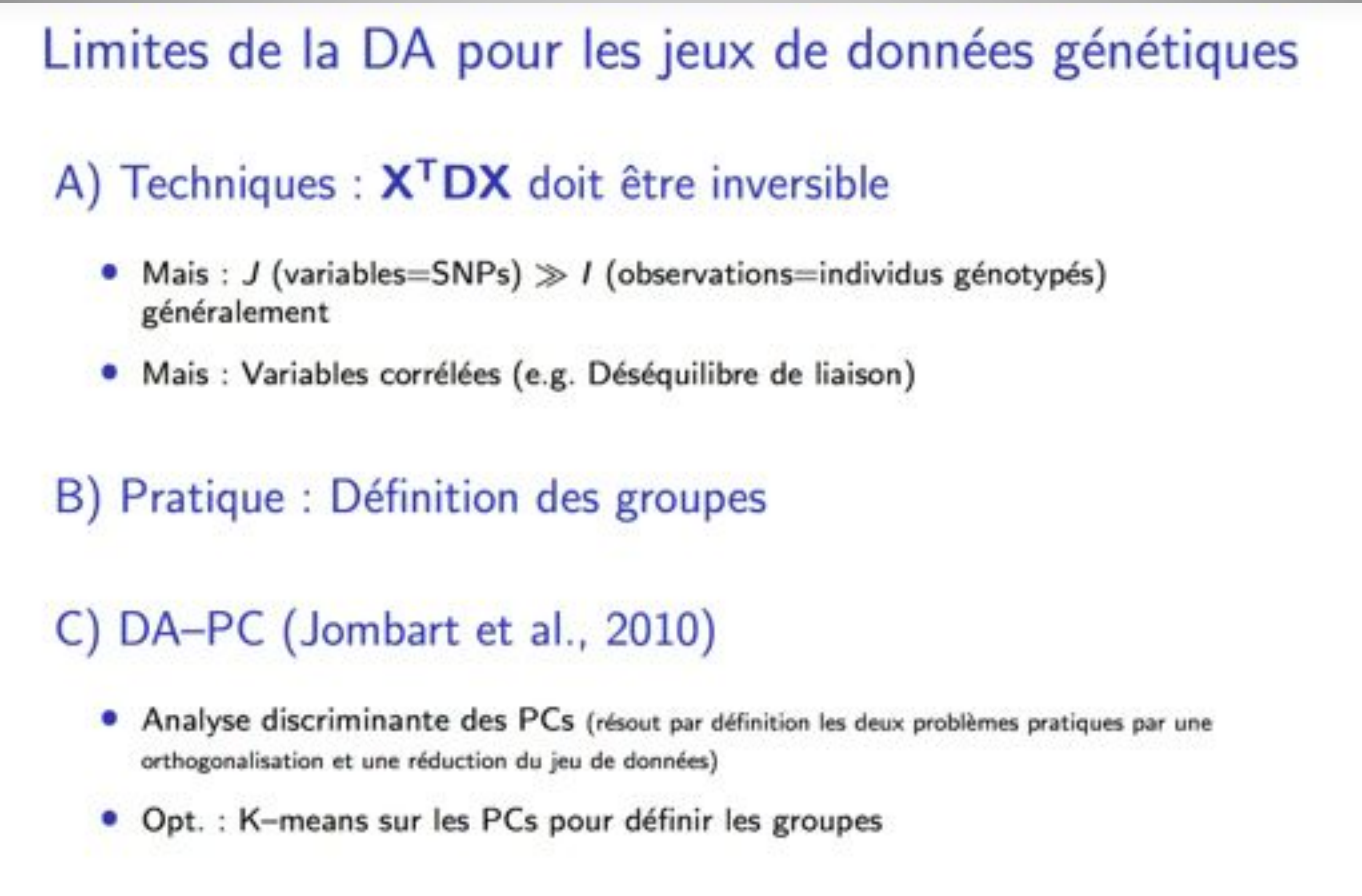
Illustration



Max. total diversity

Max. diversity between groups

Max. separation of groups



Limites de la DA pour les jeux de données génétiques

A) Techniques : $\mathbf{X}^T\mathbf{D}\mathbf{X}$ doit être inversible

- Mais : J (variables=SNPs) $\gg$ I (observations=individus génotypés) généralement
- Mais : Variables corrélées (e.g. Déséquilibre de liaison)

B) Pratique : Définition des groupes

C) DA–PC (Jombart et al., 2010)

- Analyse discriminante des PCs (résout par définition les deux problèmes pratiques par une orthogonalisation et une réduction du jeu de données)
- Opt. : K–means sur les PCs pour définir les groupes

Rappels : K–means

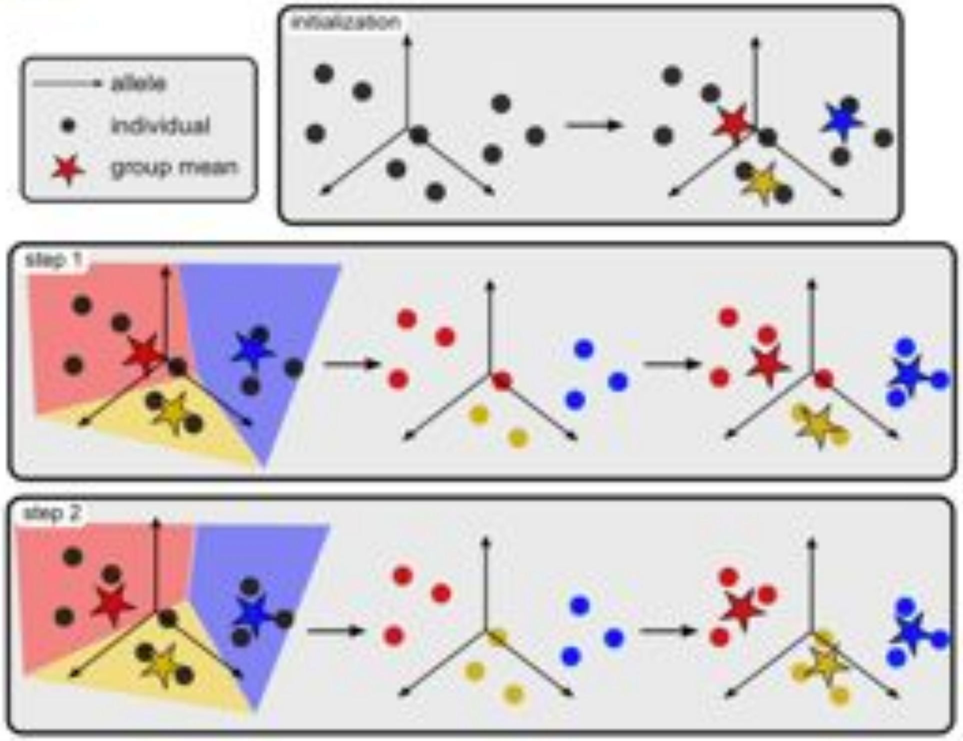
Principe

Pour un nombre K de groupes donné, assigner chaque individu à un groupe en minimisant la variance intra–groupe des PCs ($W(\mathbf{Z})$)

Valeur de K optimale ?

- K optimal = celui qui minimise le BIC$= n\log(W(\mathbf{X})) + K\log(n)$
- Support jugé fort si $\Delta(BIC) > 6$

Algorithme



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)

PCA Interpretation

(McVean, 2009, Plos Genetics)

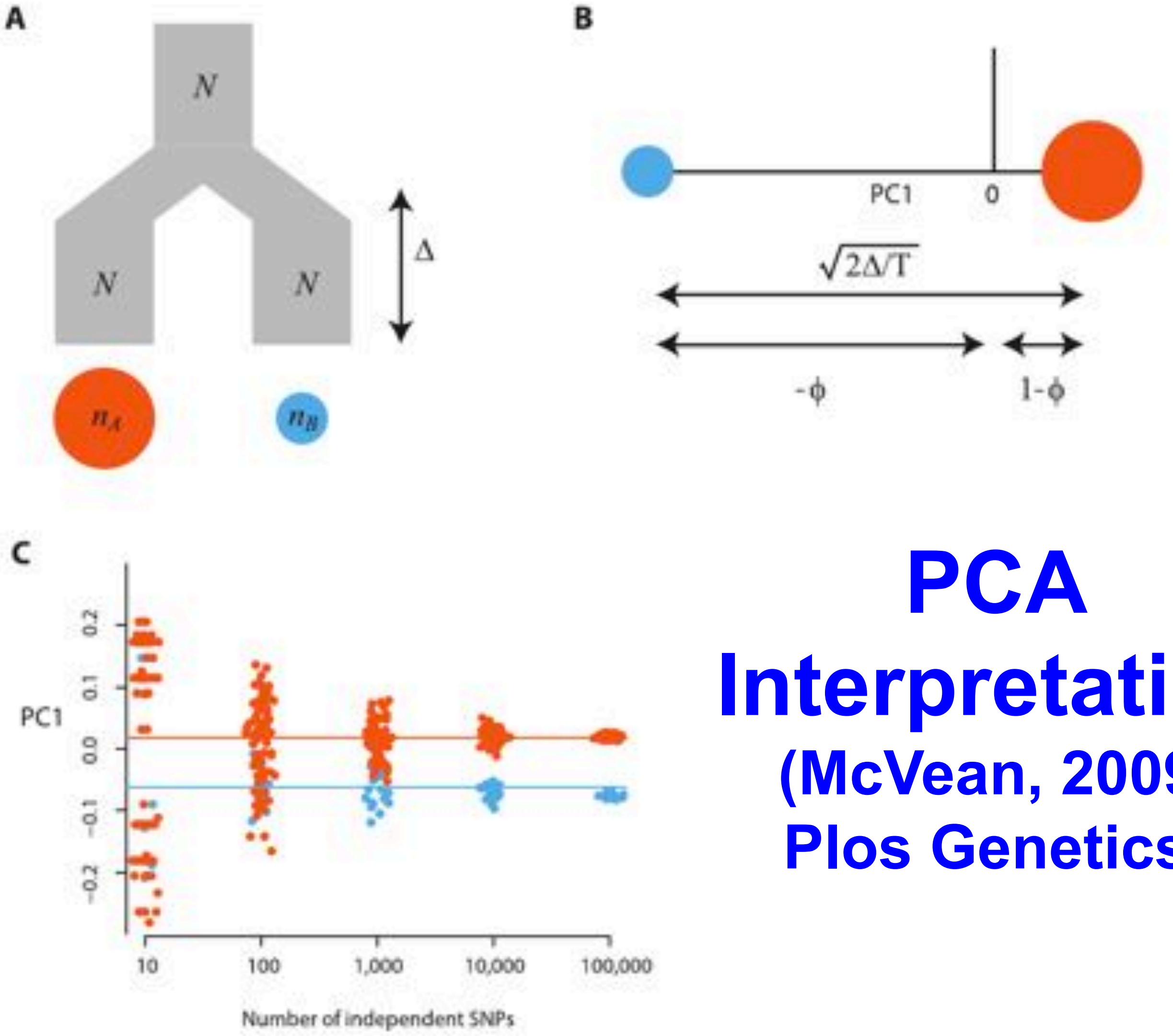


Figure 2. Principal component analysis of two populations. (A) Consider a sample of n_A individuals from population A (indicated by the red circle) and n_B from population B (indicated by the blue circle), where the two populations have the same effective population size of N and are both derived from 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

PCA Interpretation

(McVean, 2009, Plos Genetics)

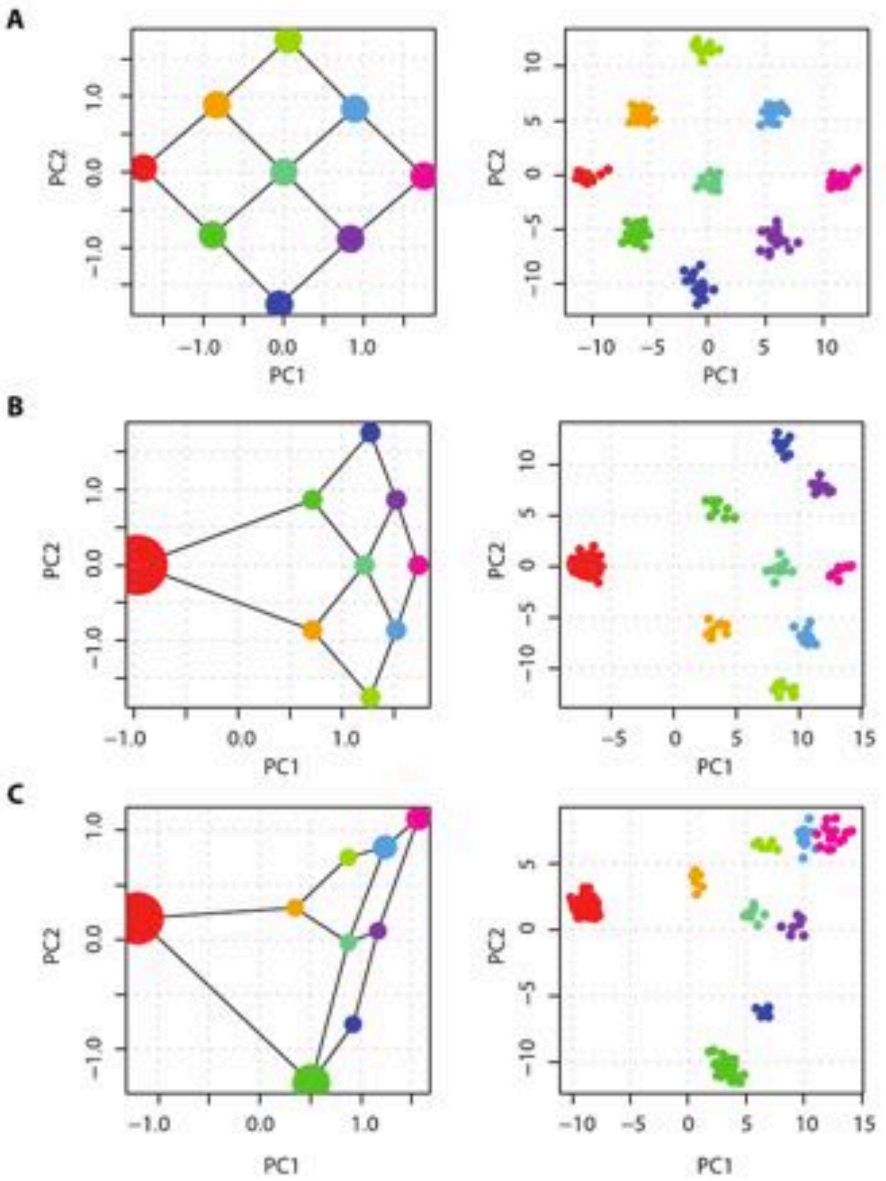
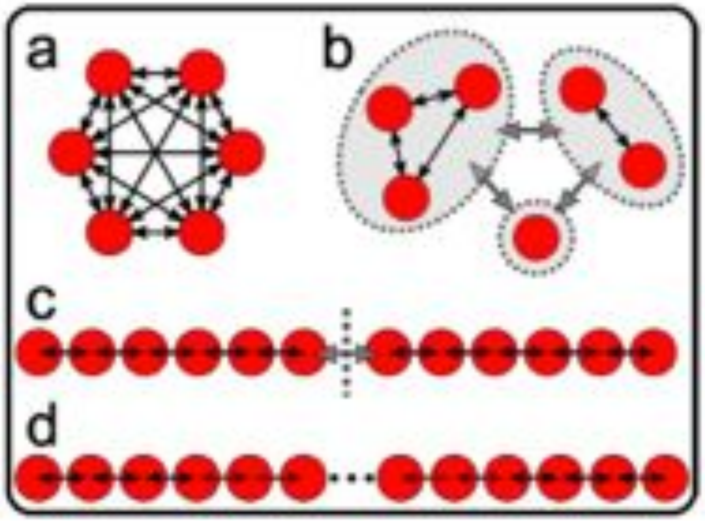


Figure 3. The effect of uneven sampling on PCA projection. PCA projection of samples taken from a set of nine populations arranged in a lattice, each of which exchanges migrants at rate M per N_e generations with each adjoining neighbour, leads to a recovery of the migration-space if samples are of equal 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

Illustration sur données simulées (Jombart et al., 2010)



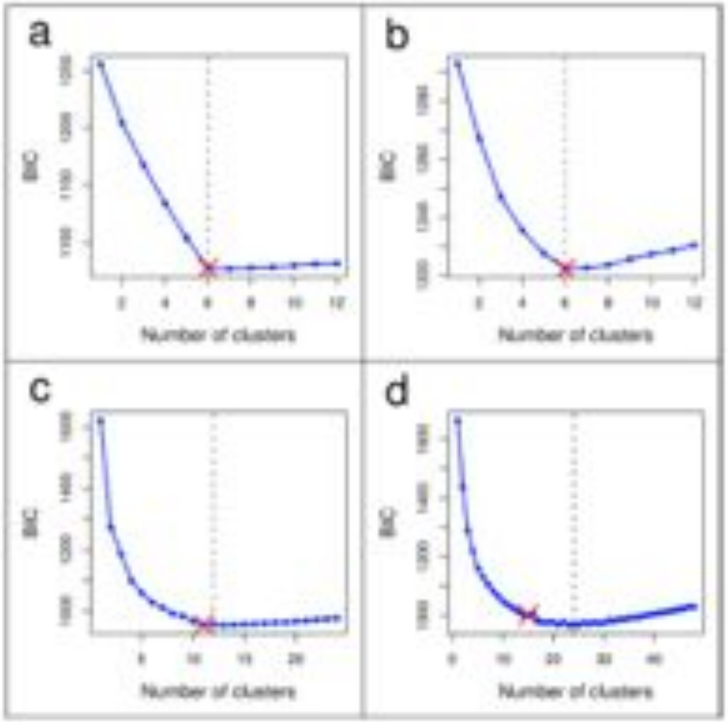
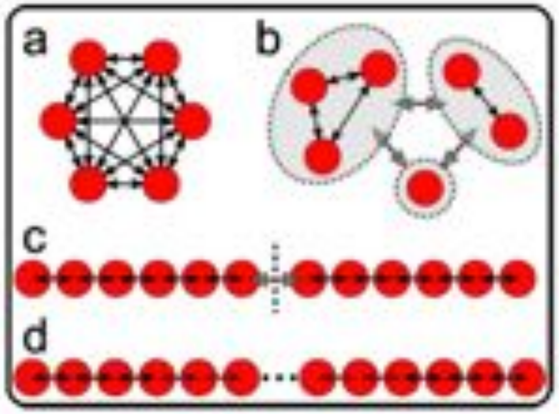
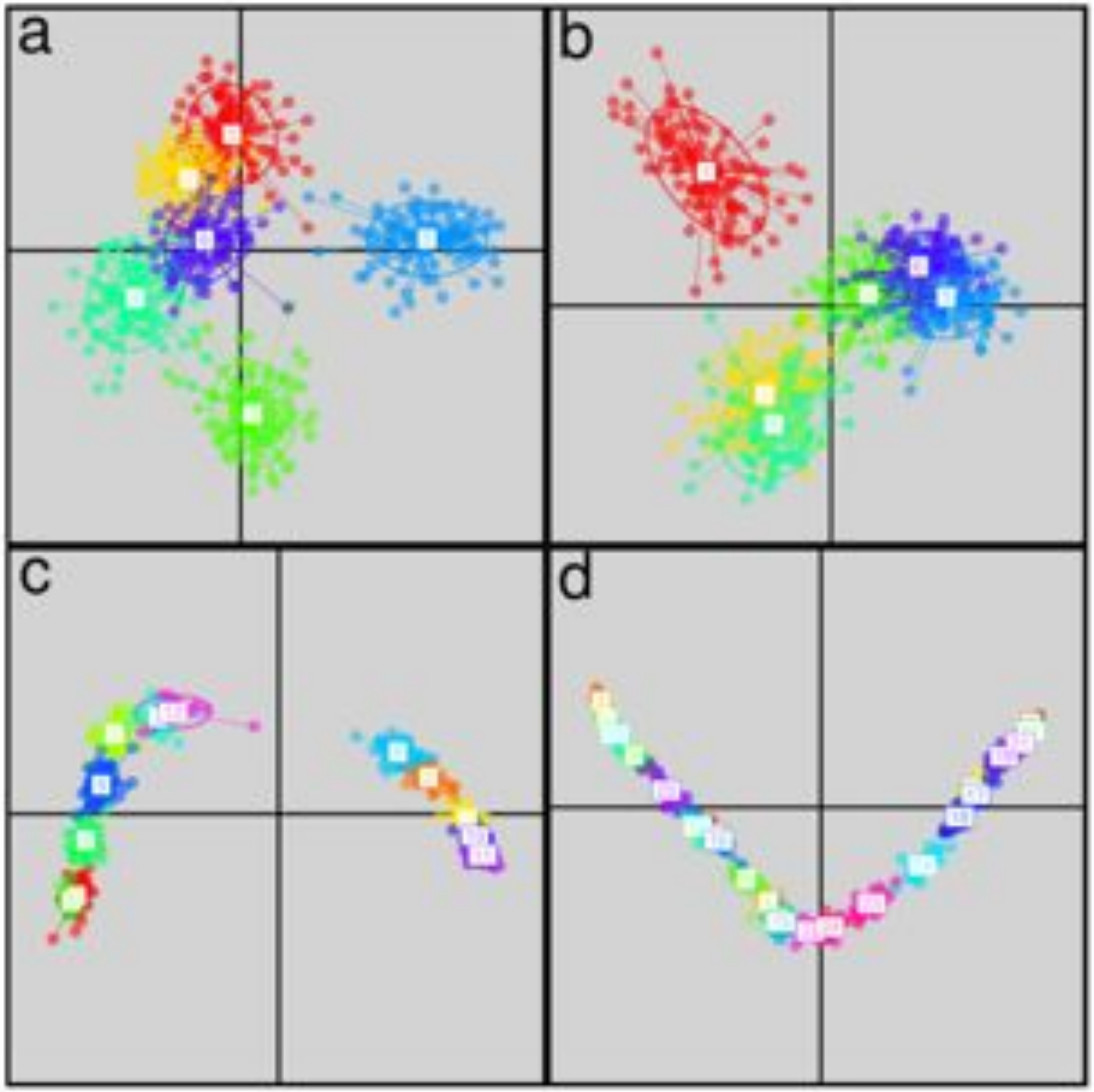
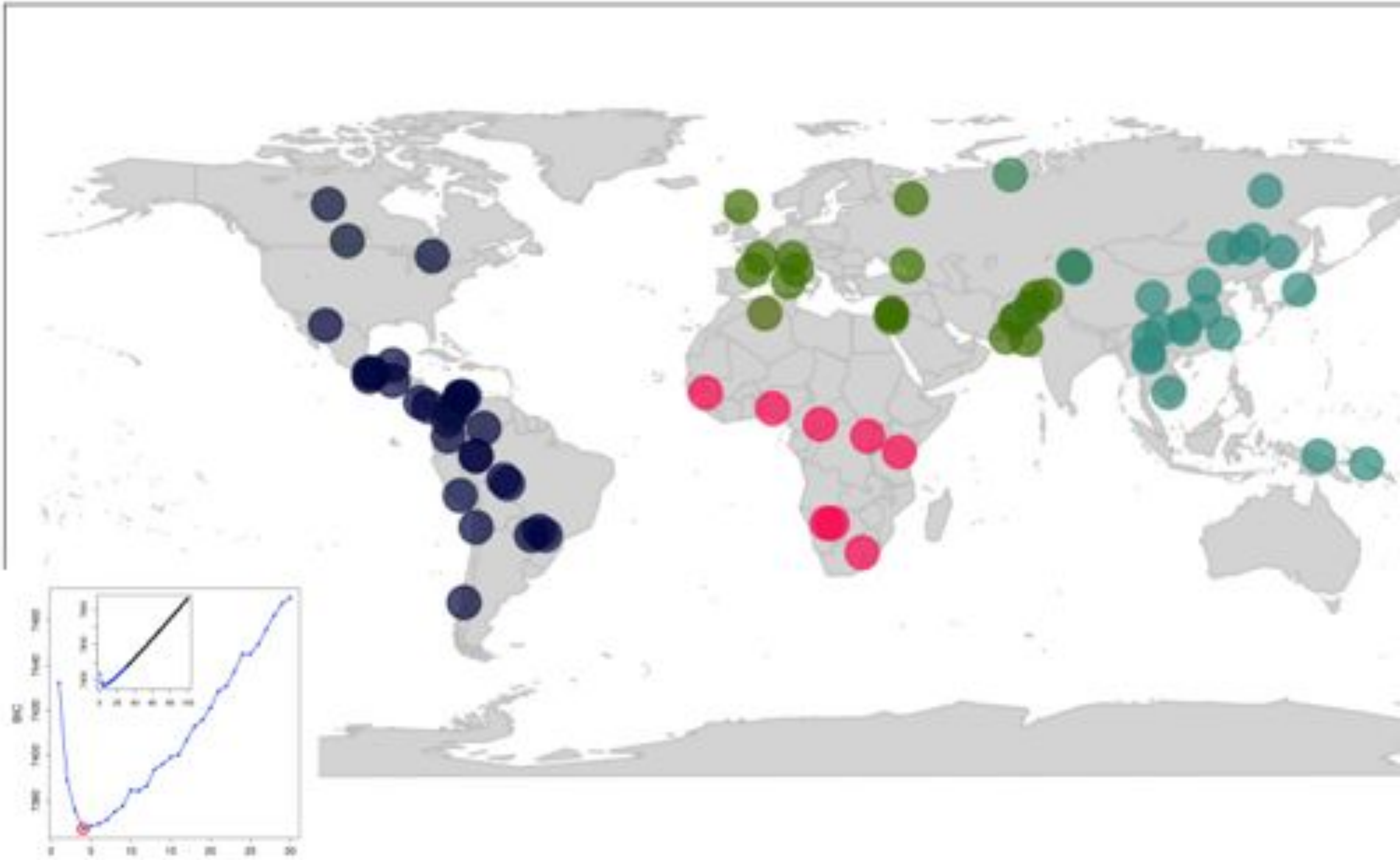


Illustration sur données simulées (Jombart et al., 2010)

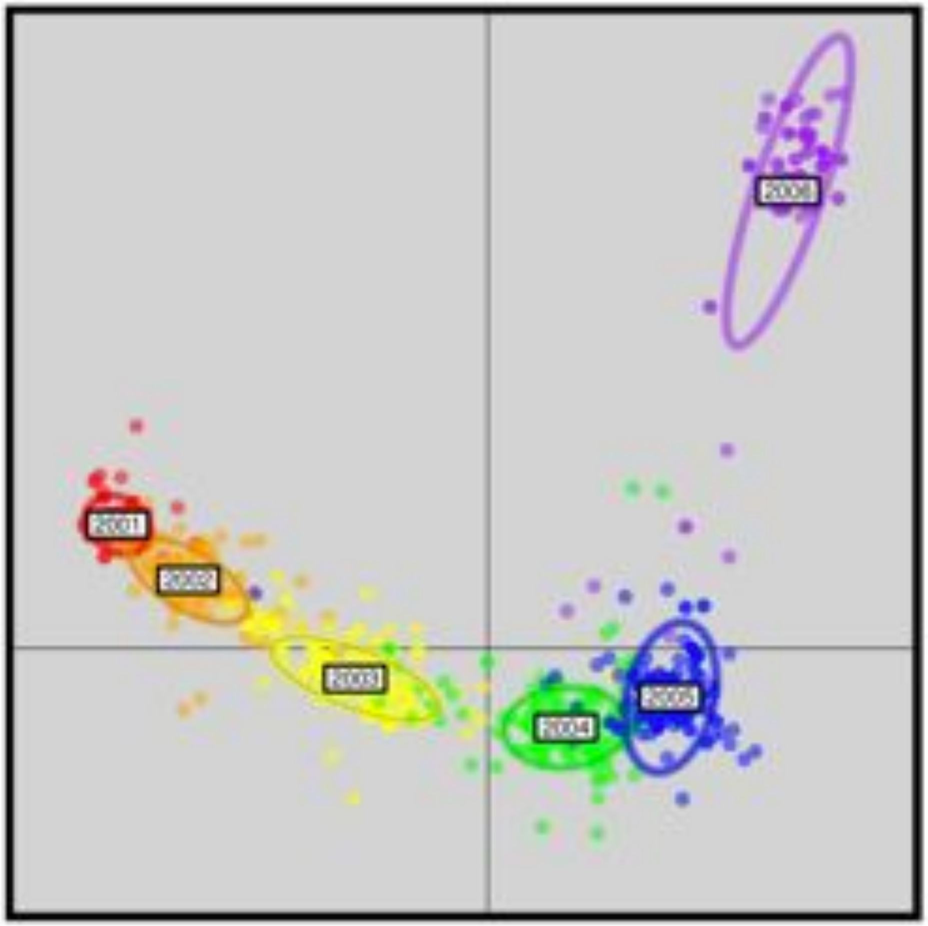


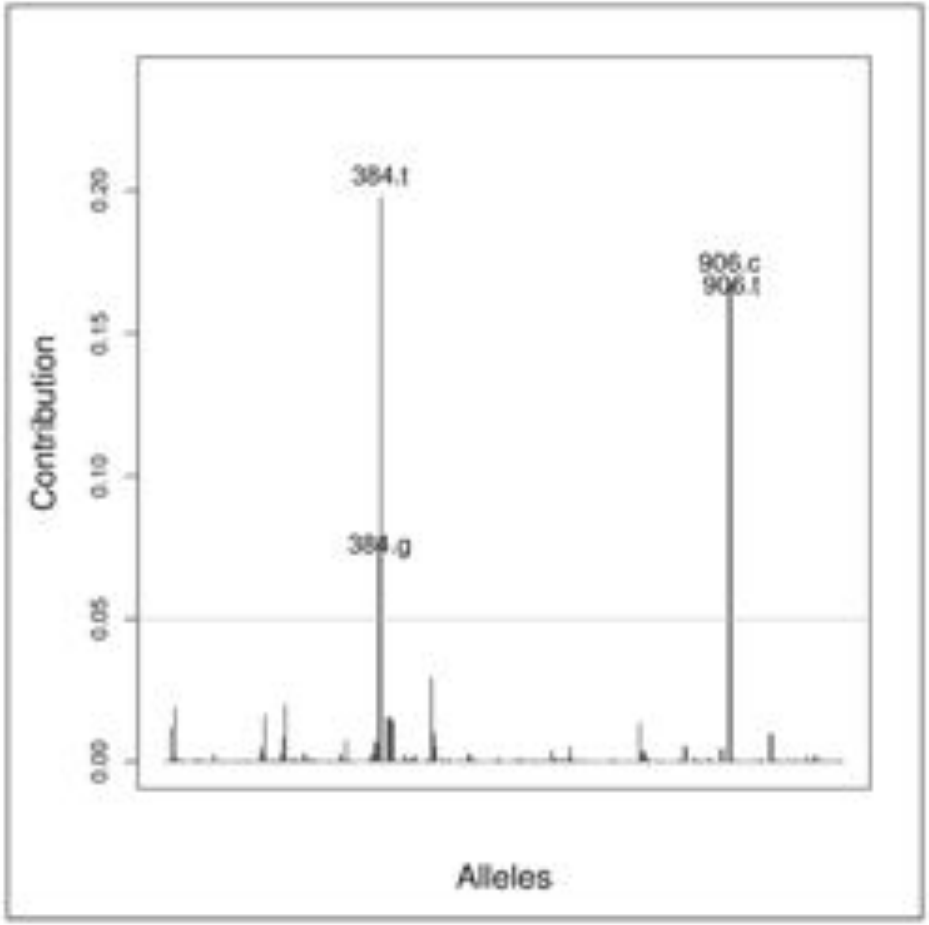


Données humaines : 1,350 ind. (79 pops.) sur 678 µsat (8,170 alleles) (Jombart et al., 20



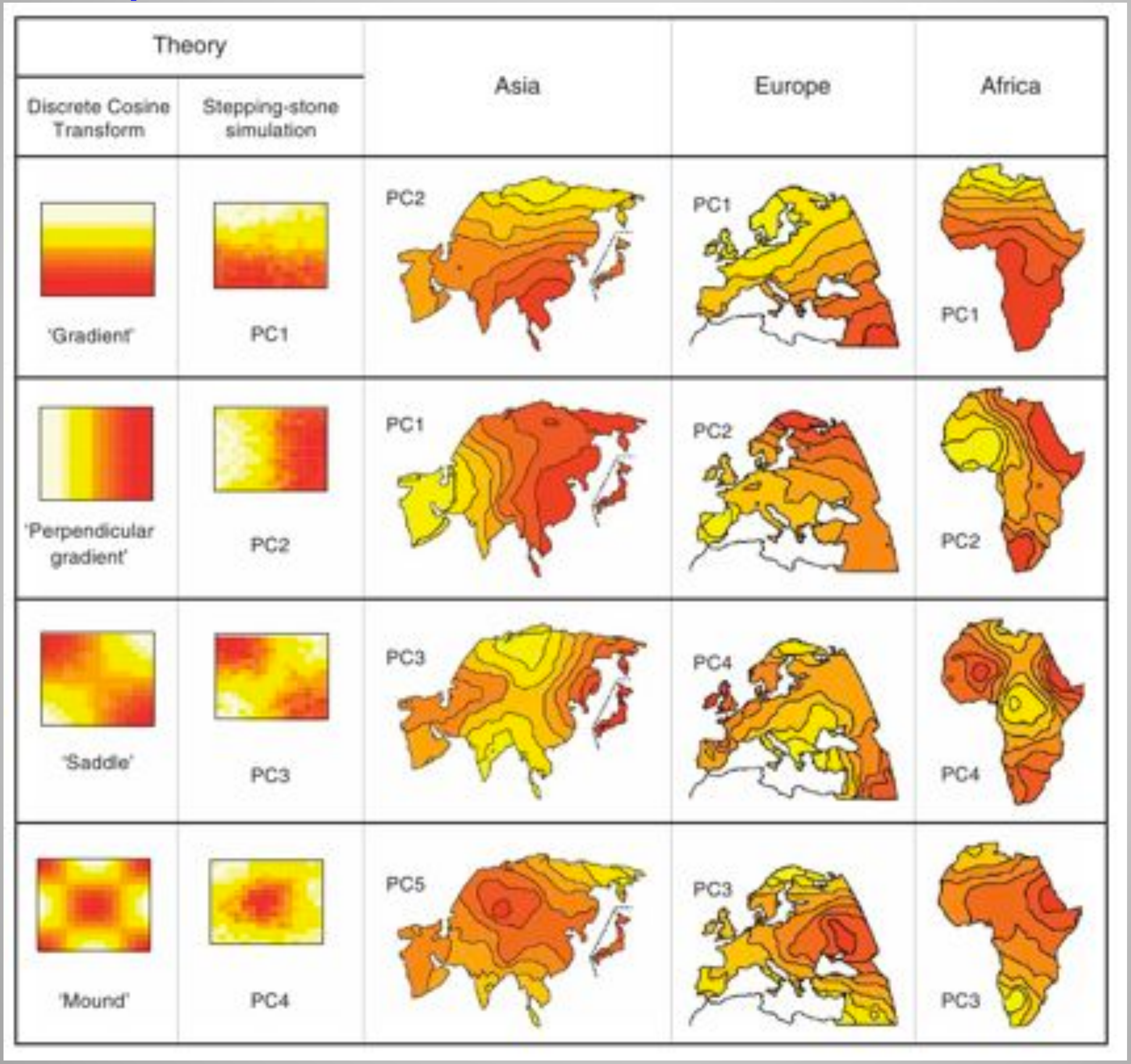
Virus de la grippe : 1,903 souches (6 hivers) séquencées ⇒ 125 SNPs (334 alleles) (Jombart et al., 2010)



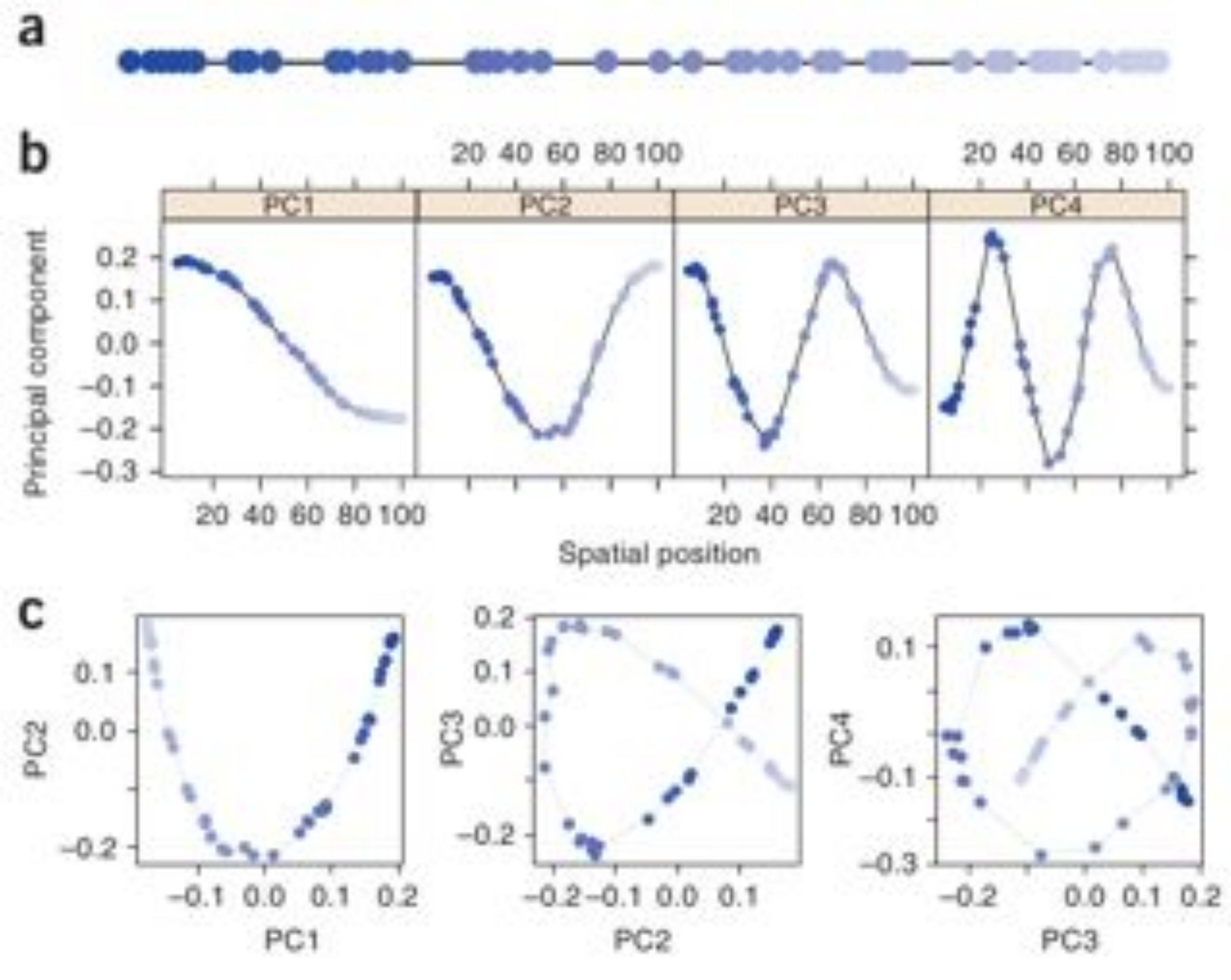


PCA Interpretation

(Novembre & Stephens, 2008, Nat Genet)



PCA Interpretation
(Novembre & Stephens, 2008, Nat Genet)


a
b
PC1
PC2
PC3
PC4
Principal component
Spatial position
c
PC2
PC1
PC3
PC4