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

From Hansen et al. 2001
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

First algorithm: Paetkau et al. 1995

Hardy Weinberg + linkage equilibrium \(\Rightarrow\) 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:

\[
p_{ijk} : \text{frequency of allele } k \text{ at locus } j \text{ in pop } i
\]

\[
L \approx 2p_{ijk}p_{ijk'} \quad \text{if heterozygote } kk'
\]

or \(L \approx p_{ijk}^2\) \quad \text{if homozygote } kk

Independent loci \(\Rightarrow\) 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 $1/(\text{gene sample size})$)
• 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

<table>
<thead>
<tr>
<th>Div T</th>
<th>Fst</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>0.3</td>
</tr>
<tr>
<td>200</td>
<td>0.08</td>
</tr>
<tr>
<td>20</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Comparison of different algorithms (Cornuet et al. 1999)

- IAM vs. SMM, differentiation level, locus number

<table>
<thead>
<tr>
<th>Div T</th>
<th>Fst</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>0.3</td>
</tr>
<tr>
<td>200</td>
<td>0.08</td>
</tr>
<tr>
<td>20</td>
<td>0.01</td>
</tr>
</tbody>
</table>

- strong effect of the mutation processes, better under IAM than SMM
- B > F > chord distance > shared alleles distance > (δμ)^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

<table>
<thead>
<tr>
<th>Td</th>
<th>Fst</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>0.3</td>
</tr>
<tr>
<td>200</td>
<td>0.08</td>
</tr>
<tr>
<td>20</td>
<td>0.01</td>
</tr>
</tbody>
</table>
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 ($F_{\text{statistics}}$, 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 immigrants all individuals for which the exclusion probability is larger than an arbitrary threshold $\alpha$ (e.g. 0.95)
H₀ : 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 immigrants all individuals for which the exclusion probability is larger than an arbitrary threshold \( \alpha \) (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
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 $L_{\text{home}}/L_{\text{max}}$ 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 individual 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
From individual assignments to the inference of migration rates

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

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

Fig. 4 The effect of sample size on power (filled symbols) and type I error rate (open symbols) relative to expectation \( [\alpha \times N \times (1-m)] \). \( N = 96, \mu = 0.005, I = 10, \alpha = 0.01 \).
Simulation test in Paetkau et al. (2004) : how to predict the power of immigrant detection on a data set

$D_{LR} = \text{mean genotype likelihood ratio}$

$\delta \mu^2 < Fst < \text{shared allele distance } Ds << D_{LR}$ (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…

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

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

\ln \Lambda > 0: \text{ the individual is a resident}
\ln \Lambda < 0: \text{ the individual is an immigrant}
\ln \Lambda = -2.3: \text{ the individual has 10 times more chance of being an 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
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 ($\alpha < 0.05$)

<table>
<thead>
<tr>
<th>Individual</th>
<th>Potential source</th>
<th>No. of markers</th>
<th>Value</th>
<th>Hypothetical immigrant ancestor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Individual ($d = 0$) Parent ($d = 1$) Grandparent ($d = 2$) Great-grandparent ($d = 3$)</td>
</tr>
<tr>
<td>AUS1</td>
<td>NGN</td>
<td>76</td>
<td>$\ln\Lambda$</td>
<td>2.76</td>
</tr>
<tr>
<td>AUS2</td>
<td>NGN</td>
<td>73</td>
<td>$\ln\Lambda$</td>
<td>4.48</td>
</tr>
<tr>
<td>AUS3</td>
<td>NGN</td>
<td>82</td>
<td>$\ln\Lambda$</td>
<td>5.23</td>
</tr>
<tr>
<td>JPN1</td>
<td>SEN</td>
<td>69</td>
<td>$\ln\Lambda$</td>
<td>17.80</td>
</tr>
</tbody>
</table>

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 $\ln\Lambda$ or $\ln\Lambda_\alpha$ are given in the first row for each individual. Values in the second row are significance levels ($\alpha$ 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…

\[ \ln \Lambda \] corresponding to \( \alpha = 0.05 \)

\[ \ln \Lambda \] for individual AUS 1

---

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

<table>
<thead>
<tr>
<th>Individual</th>
<th>Potential source</th>
<th>No. of markers</th>
<th>Value</th>
<th>Individual (( d = 0 ))</th>
<th>Parent (( d = 1 ))</th>
<th>Grandparent (( d = 2 ))</th>
<th>Great-grandparent (( d = 3 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUS1</td>
<td>NGN</td>
<td>76</td>
<td>( \ln \Lambda )</td>
<td>-2.76</td>
<td>-2.89</td>
<td>-1.65</td>
<td>-0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \alpha )</td>
<td>0.000</td>
<td>0.009</td>
<td>0.022</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Power</td>
<td>1.000</td>
<td>0.821</td>
<td>0.347</td>
<td>0.197</td>
</tr>
</tbody>
</table>
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 (!)

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

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 ($\alpha < 0.05$)

<table>
<thead>
<tr>
<th>Individual</th>
<th>Potential source</th>
<th>No. of markers</th>
<th>Value</th>
<th>Hypothetical immigrant ancestor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Individual ($d = 0$)</td>
</tr>
<tr>
<td>JPN1</td>
<td>SEN</td>
<td>69</td>
<td>ln $\alpha$</td>
<td>1.780</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\alpha$</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Power</td>
<td>1.000</td>
</tr>
</tbody>
</table>

JPN 1 may be a descendant of an 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 individual
- 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”
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
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
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
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
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
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
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)

\[
X = \begin{bmatrix}
(x_{1}^{(1,1)} & x_{1}^{(1,2)}) & \ldots & (x_{l}^{(1,1)} & x_{l}^{(1,2)}) & \ldots & (x_{L}^{(1,1)} & x_{L}^{(1,2)}) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
(x_{1}^{(i,1)} & x_{1}^{(i,2)}) & \ldots & (x_{l}^{(i,1)} & x_{l}^{(i,2)}) & \ldots & (x_{L}^{(i,1)} & x_{L}^{(i,2)}) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
(x_{1}^{(N,1)} & x_{1}^{(N,2)}) & \ldots & (x_{l}^{(N,1)} & x_{l}^{(N,2)}) & \ldots & (x_{L}^{(N,1)} & x_{L}^{(N,2)})
\end{bmatrix}
\]

\( X \) is \((N \times 2L)\)
**STRUCTURE inference method**

The data = \( X \) = individual multilocus genotypes (genetic sample)

**Microsatellite data set example**

<table>
<thead>
<tr>
<th></th>
<th>phi011</th>
<th>phi015</th>
<th>phi029</th>
<th>phi031</th>
<th>phi062</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>212</td>
<td>215</td>
<td>82</td>
<td>98</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>150</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>223</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>223</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
</tr>
<tr>
<td>2</td>
<td>218</td>
<td>218</td>
<td>82</td>
<td>102</td>
<td>158</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>158</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>187</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>227</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
</tr>
<tr>
<td>3</td>
<td>218</td>
<td>218</td>
<td>86</td>
<td>98</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>150</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>187</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>227</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
</tr>
<tr>
<td>4</td>
<td>215</td>
<td>215</td>
<td>86</td>
<td>98</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>154</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>187</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>191</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
</tr>
<tr>
<td>5</td>
<td>218</td>
<td>218</td>
<td>-9</td>
<td>-9</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>158</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>191</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>223</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
</tr>
<tr>
<td>6</td>
<td>215</td>
<td>215</td>
<td>86</td>
<td>86</td>
<td>158</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>158</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>227</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>227</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>164</td>
</tr>
</tbody>
</table>
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 $k$
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 $k$

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

$Z_{(N \times 2L)}$
The data = $X =$ individual multilocus genotypes.

Unknown variables:
- $Z =$ cluster membership of each individual or each individual locus
- $P =$ allele frequencies in each cluster

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

\[
p(X \mid Z, P) = \prod_{\text{ind } i} \prod_{\text{locus } l} 2 \cdot p_{z(i,1,l),i,1,l} \cdot p_{z(i,2,l),i,2,l}
\]

impossible to explore all partitions \( \Rightarrow \) Markov chain Monte Carlo simulation
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):

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

where $p(Z_i = k)$ is the prior probability of membership of individual $i$ (equals $1/K$ for all $I$ and $k$)
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 $m=1,2,\ldots,M$ times

1. draw $P(m)$ from $p(P|X, Z(m-1))$

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*
Model with admixture
**Inferred ancestry of individuals:**

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

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

**Estimated Allele Frequencies in each population**

First column gives estimated ancestral frequencies

**Locus 1:**
2 alleles
0.0% missing data

<table>
<thead>
<tr>
<th>Locus 1</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele</td>
<td>0.681</td>
<td>0.319</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.691</td>
<td>0.309</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>0.579</td>
<td>0.421</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.582</td>
<td>0.418</td>
</tr>
</tbody>
</table>

**Locus 2:**
2 alleles
0.3% missing data

<table>
<thead>
<tr>
<th>Locus 2</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele</td>
<td>0.694</td>
<td>0.306</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.698</td>
<td>0.302</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>0.434</td>
<td>0.566</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.796</td>
<td>0.204</td>
</tr>
</tbody>
</table>

**Locus 3:**
2 alleles
2.1% missing data

<table>
<thead>
<tr>
<th>Locus 3</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele</td>
<td>0.434</td>
<td>0.566</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.433</td>
<td>0.570</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>0.297</td>
<td>0.703</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.510</td>
<td>0.490</td>
</tr>
</tbody>
</table>

Cluster 1  Cluster 2  Cluster 1
STRUCTURE typical plots

"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 pre-defined population in each of the K clusters, that can be plotted in pies.
Example on highly structured populations

Genetic Structure of Chimpanzee Populations

Celine Becquet¹, Nick Patterson², Anne C. Stone³, Molly Przeworski¹*, David Reich²,⁴*

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

Genetic Structure of Chimpanzee Populations

Celine Becquet¹, Nick Patterson², Anne C. Stone³, Molly Przeworski¹*, David Reich²,4*

Table 3. Genetic Differentiation among Populations

<table>
<thead>
<tr>
<th>Location</th>
<th>Eastern</th>
<th>Central</th>
<th>Bonobo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Western</td>
<td>0.31 (0.32)</td>
<td>0.25 (0.29)</td>
<td>0.68 (0.68)</td>
</tr>
<tr>
<td>Eastern</td>
<td>—</td>
<td>0.05 (0.09)</td>
<td>0.57 (0.54)</td>
</tr>
<tr>
<td>Central</td>
<td>—</td>
<td>—</td>
<td>0.51 (0.49)</td>
</tr>
</tbody>
</table>
Example on highly structured populations

Genetic Structure of Chimpanzee Populations

Celine Becquet¹, Nick Patterson², Anne C. Stone³, Molly Przeworski¹*, David Reich²,4*

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 clusters 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:

<table>
<thead>
<tr>
<th>Assumed value of $K$</th>
<th>Posterior probability of $K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>~0</td>
</tr>
<tr>
<td>2</td>
<td>~0</td>
</tr>
<tr>
<td>3</td>
<td>0.993</td>
</tr>
<tr>
<td>4</td>
<td>0.007</td>
</tr>
<tr>
<td>5</td>
<td>0.00005</td>
</tr>
</tbody>
</table>

Taita Thrush data
Inference of the number of clusters $K$

STRUCTURE do not infer the number of cluster using MCMC,

<table>
<thead>
<tr>
<th>Assumed value of $K$</th>
<th>Posterior probability of $K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>~0</td>
</tr>
<tr>
<td>2</td>
<td>~0</td>
</tr>
<tr>
<td>3</td>
<td>0.993</td>
</tr>
<tr>
<td>4</td>
<td>0.007</td>
</tr>
<tr>
<td>5</td>
<td>0.00005</td>
</tr>
</tbody>
</table>

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:

the variation in likelihood between different $K$ values can also be used ($\Delta K$)
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.
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 speaking.
Spatial clustering: the GENELAND software
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 $\Rightarrow$ 1 pop is 1 color
GENELAND spatial population model
GENELAND spatial population model
Set of spatialized panmictic populations
example of different Tessellation outputs for different spatial correlations

The spatial correlation is modeled through the parameter $m = \text{max number of disjointed polygons that form a cluster}$.

small $m \Rightarrow$ more spatial correlation, large $m \Rightarrow$ less spatial correlation because $p(\text{2 ind } \subseteq \text{ single cluster})$ 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.
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 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

Very good spatial delimitation of genetic clusters with $F_{ST}=0.16$
GENELAND: simulation test spatial cluster delimitation

less and less precision when genetic differentiation decreases
GENELAND: simulation test
immigrant detection

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

spatial delimitation of 6 genetic clusters
detection of 5 migrants

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.
GENELAND : test on a real data set

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