Inferring demographic histories from genomic polymorphism data

Jean-Michel Marin

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck (IMAG)
Institut de Biologie Computationnelle (IBC)
LabEx NUMEV

Thanks

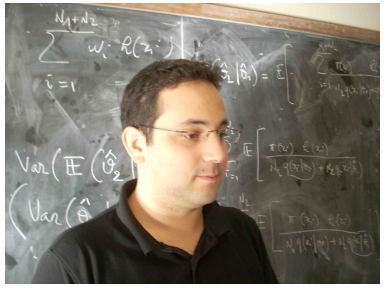




Mark Beaumont, Jean-Marie Cornuet, Arnaud Estoup, Mathieu Gautier, Coralie Merle







Raphaël Leblois, Pierre Pudlo, Christian Robert, François Rousset, Mohammed Sedki

Introduction

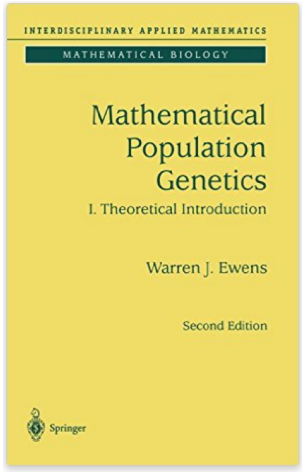
Population genetics

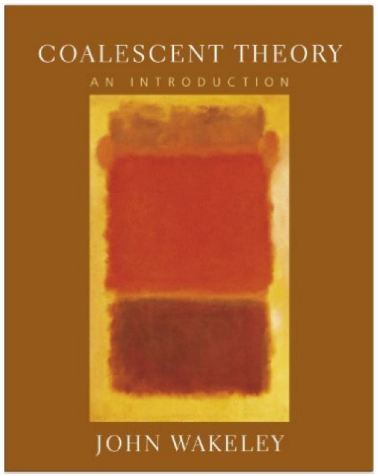
Population genetics is concerned with the causes and effects of genetic variation.

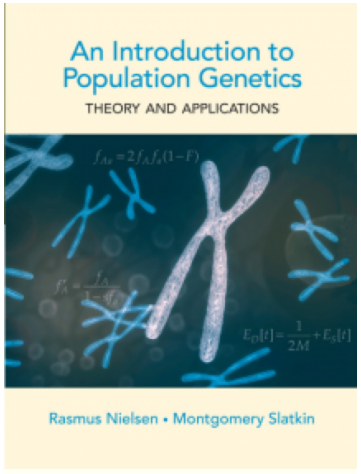
- classical population genetics theory: allele and haplotype frequencies
  describe the genotypes, estimate the allele frequencies, determine their distributions within and between populations;

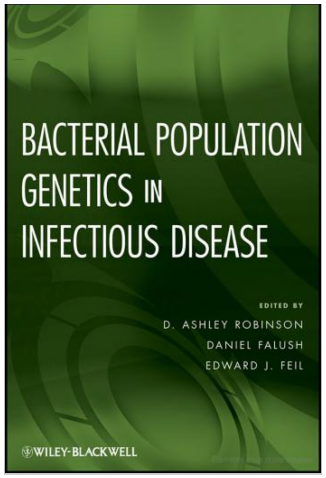
- modern population genetics theory: coalescent theory
  predict and understand the evolution of gene frequencies in populations as a result of various factors.

One of the main developments in population genetics modeling is the use of coalescent methods.









The goal is to recover some elements of populations history. To analyse the structure of genetic data, these methods use the gene trees.

The formulation of a model is constrained by an evolutionary scenario that mimics the historical and demographic reality.

Such a scenario summarizes the evolutionary history of populations by a sequence of demographic events from an ancestral population.

Our datasets are composed of genetic informations coming from several locus, **more and more locus**...

There are several options to model the relationship between these different loci: common genealogy, partially shared genealogies and recombination, or independent genealogies.

In this talk, we assume that the loci are independent.

We focus on a class of probabilistic models that includes inter-populational events such as divergence, admixture and migration.

In this talk, we consider neutral models (Kimura (1968, 1983)): no selection effect.

The observed polymorphisms are the result of genetic mutations on the genealogy of individuals.

With these models, we can answer questions of biological interest:

- estimate divergence times, quantify reductions or increases in effective population sizes, infer migration rates...
  parameter estimation problems

- determine from which ancestral sources comes a population, describe the invasion routes...
  model choice questions

Human populations example

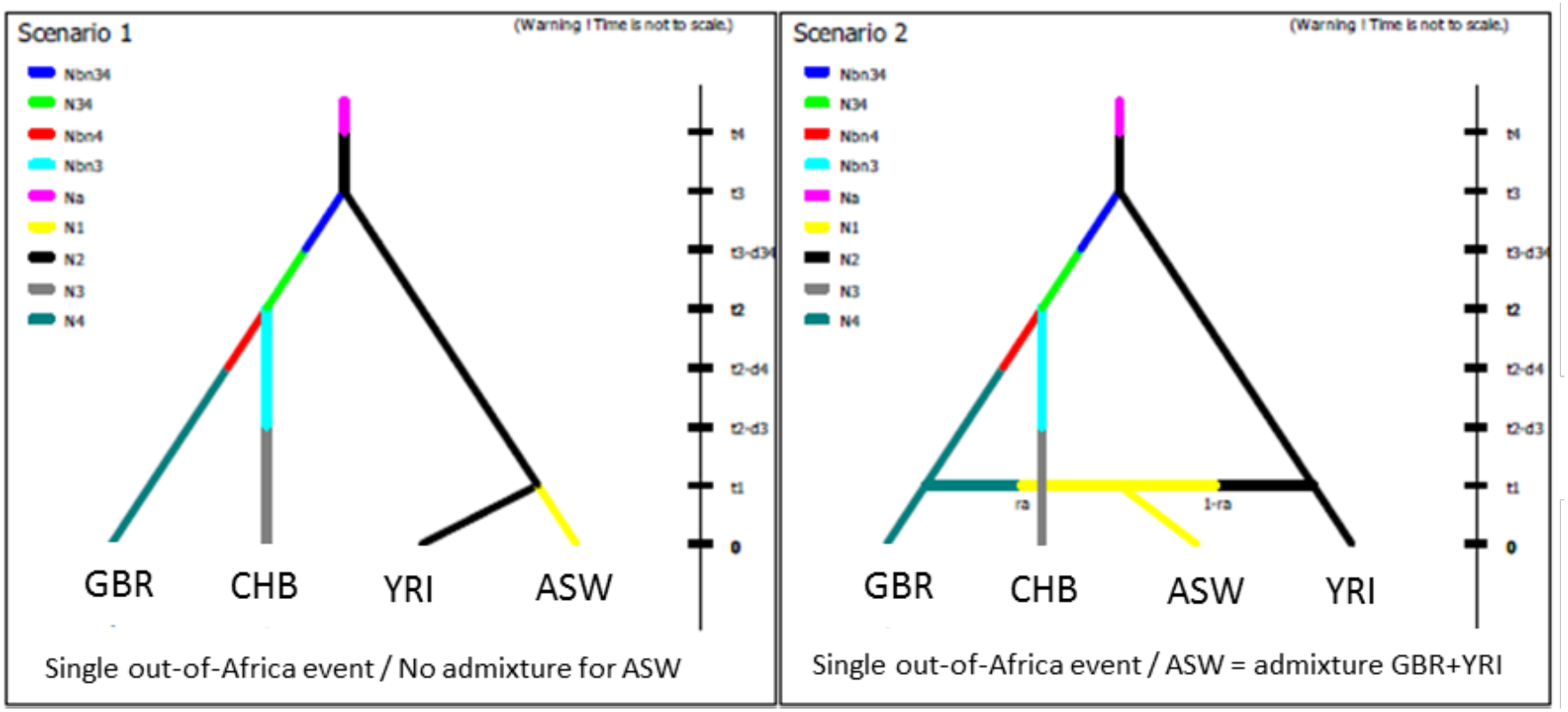
50,000 SNP markers genotyped in four Human populations: Yoruba (Africa), Han (East Asia), British (Europe) and American individuals of African Ancestry; 30 individuals per population.

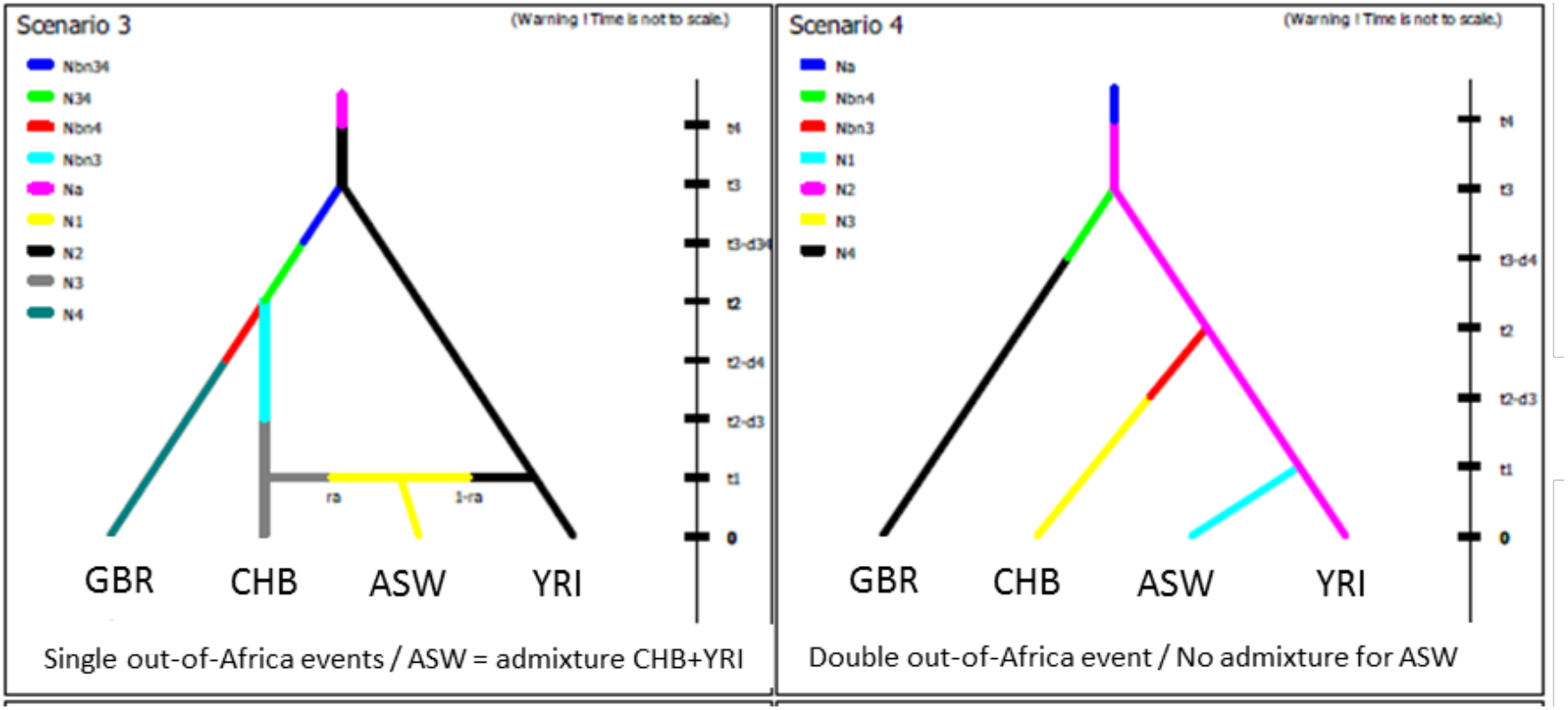
The dataset: a vector of length $6,000,000$.

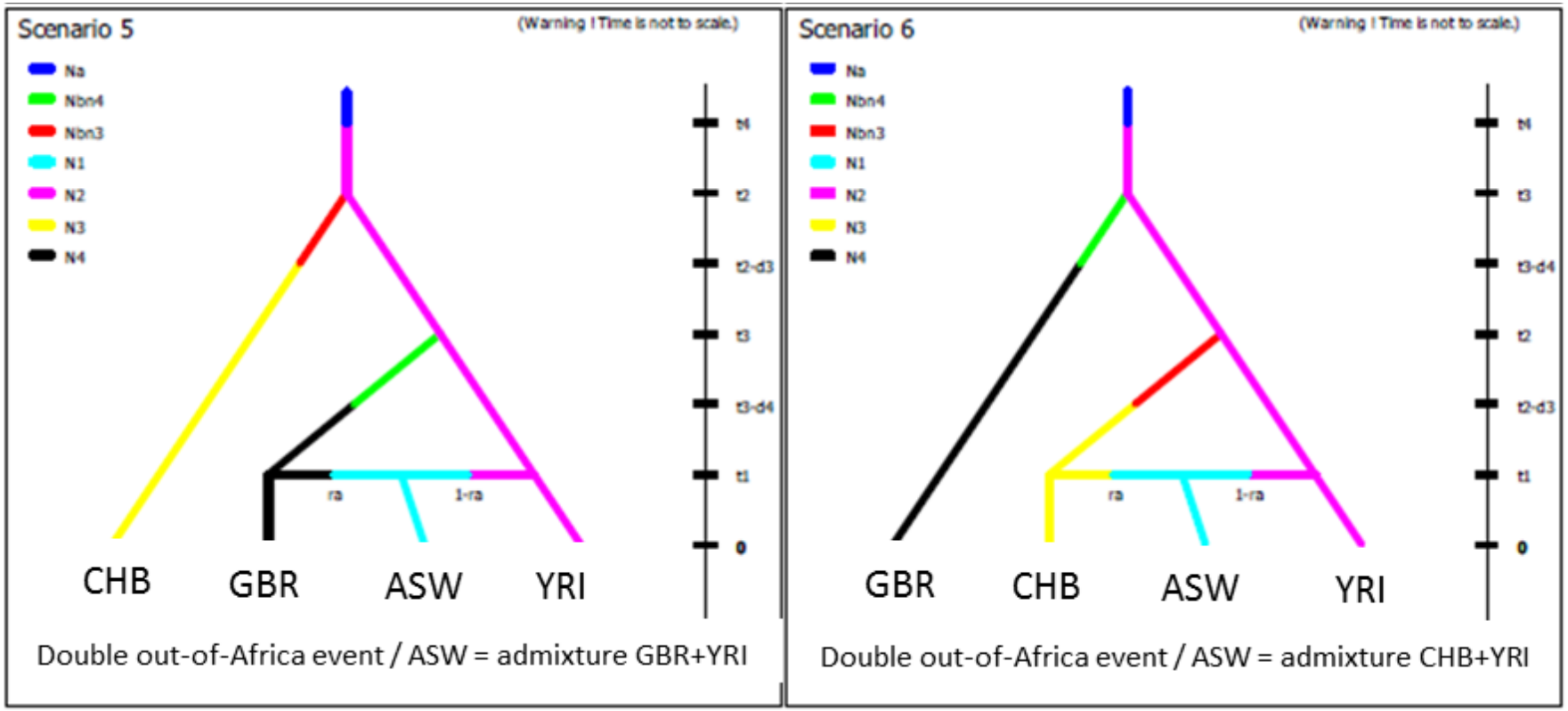
We compared six scenarios of evolution which differ from each other by one ancient and one recent historical events:

A) a single out-of-Africa colonization event giving an ancestral out-of-Africa versus two independent out-of-Africa colonization events;

B) the possibility of a recent genetic admixture of Americans of African origin with their African ancestors and individuals of European or East Asia origins.


Scenario 1
(Warning ! Time is not to scale)
Nbn34
N34
Nbn4
Nbn3
Na
N1
N2
N3
N4
tN
t3
t3-d34
t2
t2-d4
t2-d3
t1
0
GBR
CHB
YRI
ASW
Single out-of-Africa event / No admixture for ASW
Scenario 2
(Warning ! Time is not to scale)
ra
1-ra
GBR
CHB
ASW
YRI
Single out-of-Africa event / ASW = admixture GBR+YRI


Scenario 3
(Warning ! Time is not to scale.)
Nbn34
N34
Nbn4
Nbn3
Na
N1
N2
N3
N4
ra
1-ra
t4
t3
t3-d34
t2
t2-d4
t2-d3
t1
0
GBR
CHB
ASW
YRI
Single out-of-Africa events / ASW = admixture CHB+YRI
Scenario 4
(Warning ! Time is not to scale.)
Na
Nbn4
Nbn3
N1
N2
N3
N4
t4
t3
t3-d4
t2
t2-d3
t1
0
GBR
CHB
ASW
YRI
Double out-of-Africa event / No admixture for ASW


Scenario 5
(Warning ! Time is not to scale.)
Na
Nbn4
Nbn3
N1
N2
N3
N4
t4
t2
t2-d3
t3
t3-d4
t1
0
ra
1-ra
CHB
GBR
ASW
YRI
Double out-of-Africa event / ASW = admixture GBR+YRI
Scenario 6
(Warning ! Time is not to scale.)
Na
Nbn4
Nbn3
N1
N2
N3
N4
t4
t3
t3-d4
t2
t2-d3
t1
0
ra
1-ra
GBR
CHB
ASW
YRI
Double out-of-Africa event / ASW = admixture CHB+YRI

Outline

Data

Sample genealogies

Mutation process

Inferential difficulties

Simulation techniques as inferential tools

Data

The dataset consists of different samples.

Each sample corresponds to a population.

We consider D populations, $Pop1, \ldots, PopD$, the sample size of population $Popi$ is denoted by n_i.

We assimilate a diploid individual to two haploid individuals.

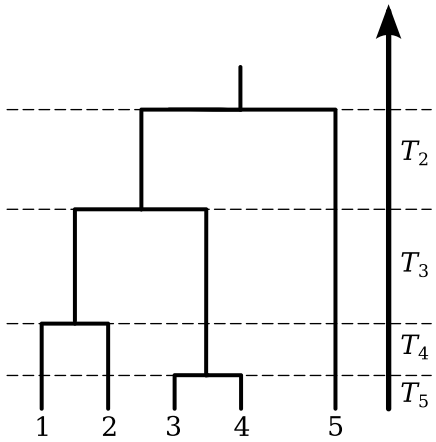
For each individual, **we consider a large number of locus on the genome**.

For these loci, the DNA sequence can vary for an individual to another due to mutations: genetic polymorphism.

Different types of loci: **microsatellite**, **SNP** (Single Nucleotide Polymorphism) or **a sequence of DNA**.

Sample genealogies

Coalescent theory (Kingman (1982), Tajima, Tavaré...)



The Kingman coalescent model describes the genealogy of a sample of genes back to the Most Recent Common Ancestor (MRCA) of the sample.

The genealogy of a sample is represented by a dendrogram.

Ancestral lineages are generated until the MRCA.

A coalescent event occurs when the lineages of two individuals merge at a node of the dendrogram.

The genealogy of a sample of k individuals is composed of $k-1$ coalescent events.

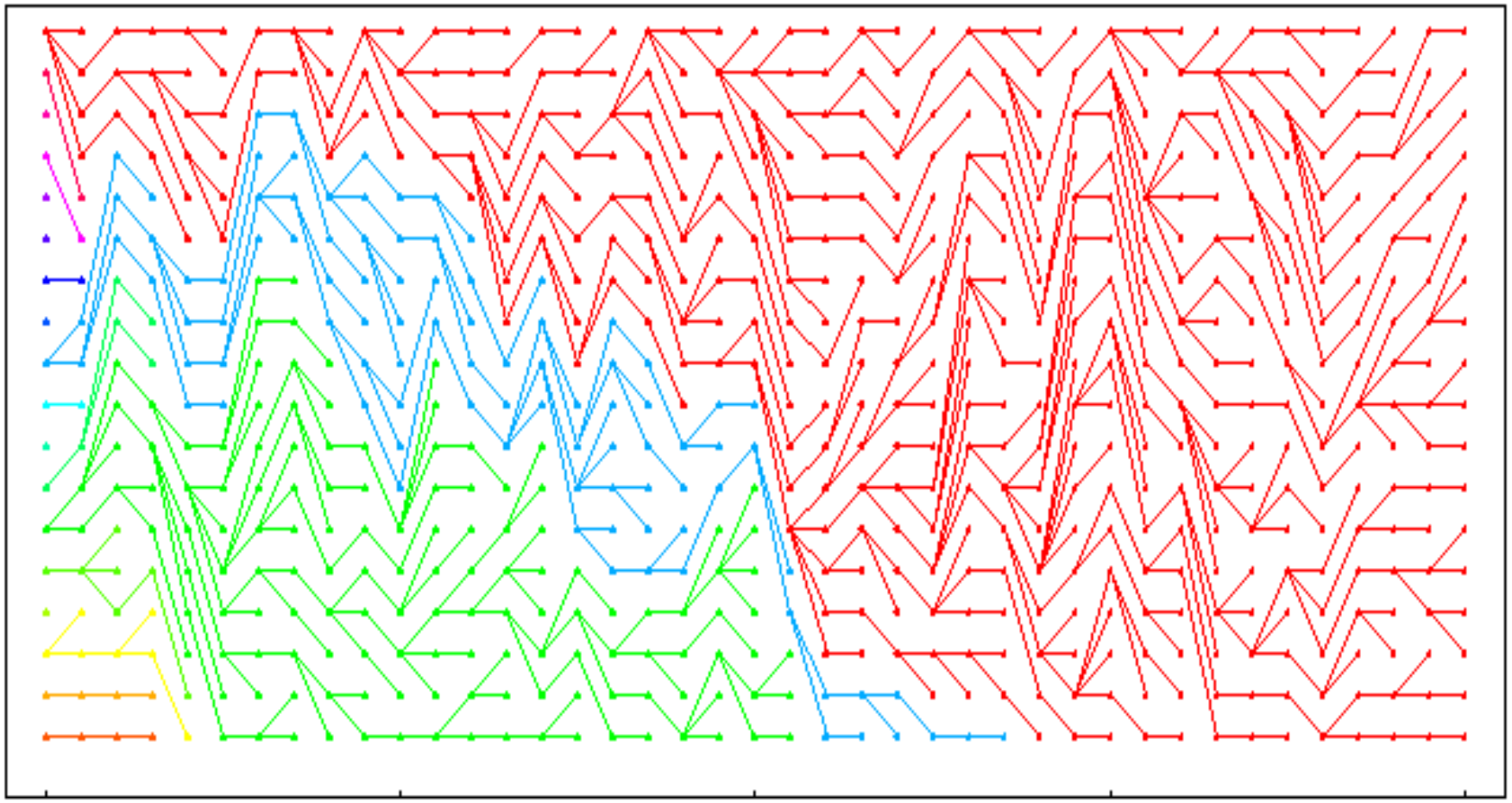
Let $T_k, \ldots, T_2$ be the durations between successive coalescent events.

The genealogy probability distribution of k individuals is characterized by the choice of the lineages at each coalescent event and the distribution of durations between coalescent events $T_k, \ldots, T_2$.

Let Ne be the effective population size.

For the Kingman coalescent, the durations between coalescent events $T_k, \ldots, T_2$ are independent and T_k is distributed according to an exponential distribution with parameter $k\big(k-1\big)/(2Ne)$.

Asymptotic approximation of the Wright-Fisher model



while $k \geq 2$ **do**

1) Generate a duration T_k from an exponential distribution with parameter $\frac{k\left(k-1\right)}{2Ne}$

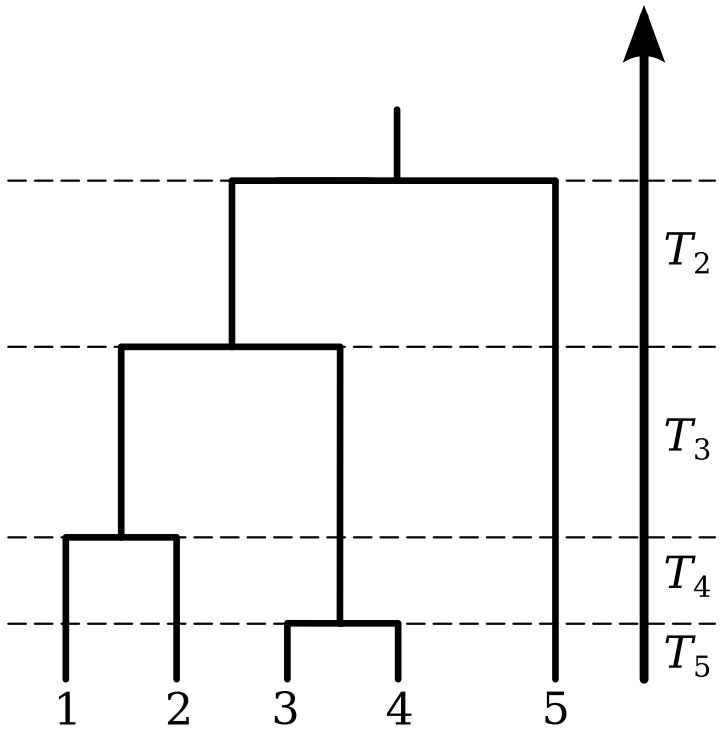
2) Add T_k to the lengths of the k lineages

3) Choose at random (uniformly) two lineages and merge them to create a node of the dendrogram

4) $k \leftarrow k - 1$

end while

Five individuals from a closed population at equilibrium



Several structured populations

We consider an evolutionary scenario described by inter-populational events.

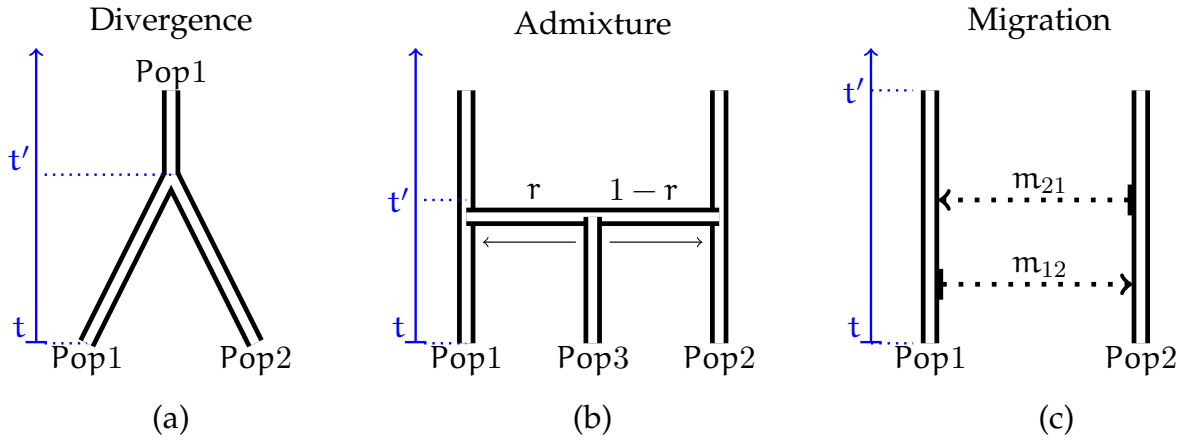
We combine these events with the Kingman coalescent which describes the intra-populational genealogies.

Three inter-populational events:

- **Divergence** the fusion of two populations back in time.
- **Admixture** the split of a population in two parts.
- **Migration** move of lineages from one population to another over a fixed period.

At a divergence event, back in time, the lineages of the two populations are merged to constitute a new population.

At an admixture event, back in time, the admixture sample $Pop3$ is split at random in two parts: a lineage of $Pop3$ is associated to $Pop1$ with probability r and to $Pop2$ with probability $1-r$ where r is the admixture rate.



a coalescent process within each pipeline

Mutation process

Position of mutations on the tree

The mutation rate per diploid individual is denoted by μ.

Conditional on the genealogy, the mutations are distributed according to a Poisson point process with intensity $\mu/2$.

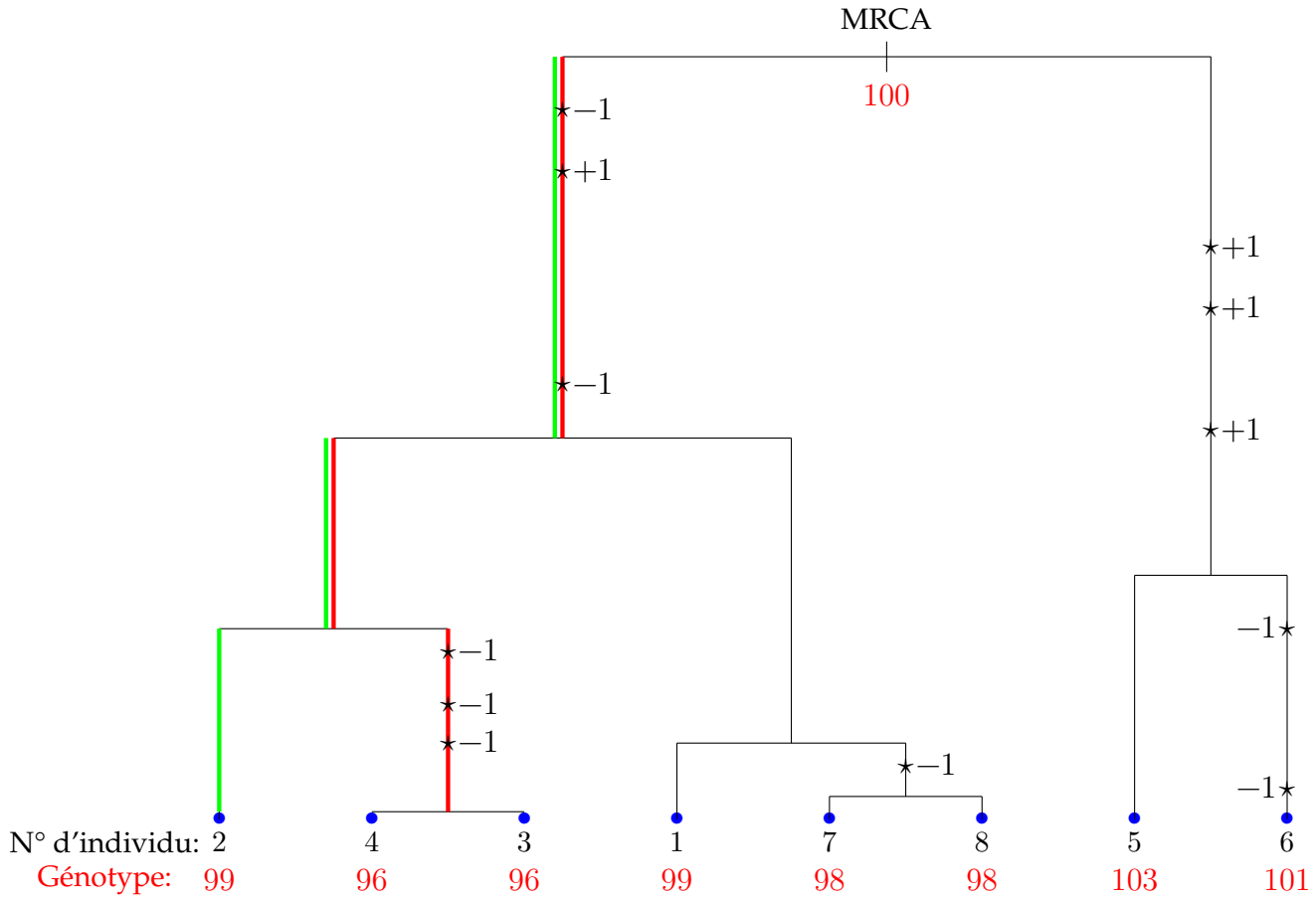
On a branch of length t, the number of mutations N is distributed according to a Poisson distribution with parameter $\mu t/2$ and the N mutations are uniformly distributed on the branch.

For microsatellite data, two mutation models: SMM (Stepwise Mutation Model) and GSM (Generalized Stepwize Mutation Model), symmetric random walks.

SNP loci have low mutation rates, we consider that polymorphism at such loci results from a single mutation.

For SNP data, a single mutation event is put at random on one branch of the genealogy, the branch being chosen with a probability proportional to its length.

To generate the genotypes of a sample at a given locus, we just have to modify the genotype of the MRCA along the genealogy.


MRCA
100
⋆−1
⋆+1
⋆−1
⋆+1
⋆+1
⋆+1
⋆−1
⋆−1
⋆−1
⋆−1
−1⋆
−1⋆
N° d'individu: 2 4 3 1 7 8 5 6
Génotype: 99 96 96 99 98 98 103 101

Inferential difficulties

Each model is characterized by a set of parameters $\boldsymbol{\theta}$ historical (divergence times, admixture times, ...), demographic (effective population sizes, admixture rates, migration rates, ...) and genetic (mutation rate, ...).

The goal is to estimate these parameters from a polymorphism dataset $\mathbf{x}$ observed at the present time.

Difficulty: we cannot calculate $f_{\boldsymbol{\theta}}(\mathbf{x})$.

Let $f_{\boldsymbol{\theta}}(\mathcal{G})$ denote the probability density of the genealogy of genes $\mathrm{d}\mathcal{G}$.

Let $f_{\boldsymbol{\theta}}(\mathcal{M}|\mathcal{G})$ denote the probability density of the mutation process $\mathcal{M}$ given the genealogy $\mathcal{G}$.

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{i \in \{locus\}} \int_{\mathcal{M}_i \to \mathbf{x}_i} f_{\boldsymbol{\theta}}\left(\mathcal{M}_i \middle| \mathcal{G}_i\right) f_{\boldsymbol{\theta}}\left(\mathcal{G}_i\right) \mathrm{d}\mathcal{G}_i \, \mathrm{d}\mathcal{M}_i,$$

where $\mathbf{x}_i$ is the data at locus i and $\mathcal{M}_i \to \mathbf{x}_i$ is the set of genotypes on the dendrogram compatible with $\mathbf{x}_i$.

That is a very high-dimensional integral with discrete (such as the genotype) and continuous (such as the length of branches) parts.

Despite the simplicity of the Kingman coalescent and the mutation processes, we cannot expect any simplification in the calculation of the likelihood.

Simulation techniques as inferential tools

Two strategies

- avoid to approximate the likelihood, simulate from the model and compare simulated datasets to the observed one
  Approximate Bayesian Computation methods

- approximate the likelihood using advanced Monte Carlo methods
  Importance Sampling methods