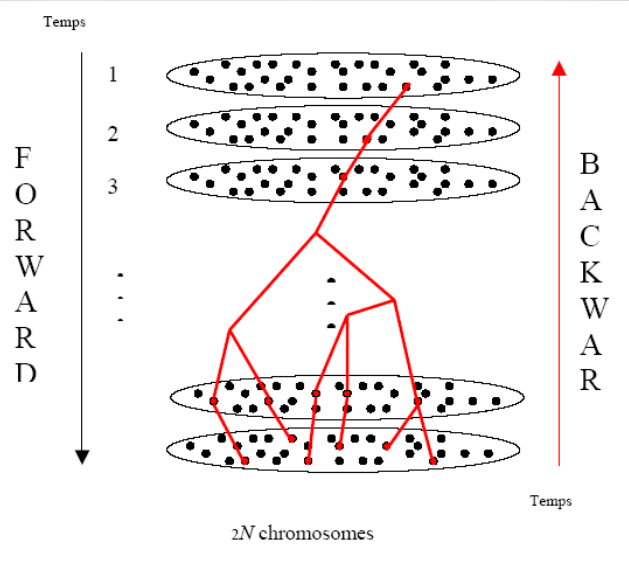
Likelihood-based demographic inference using the coalescent

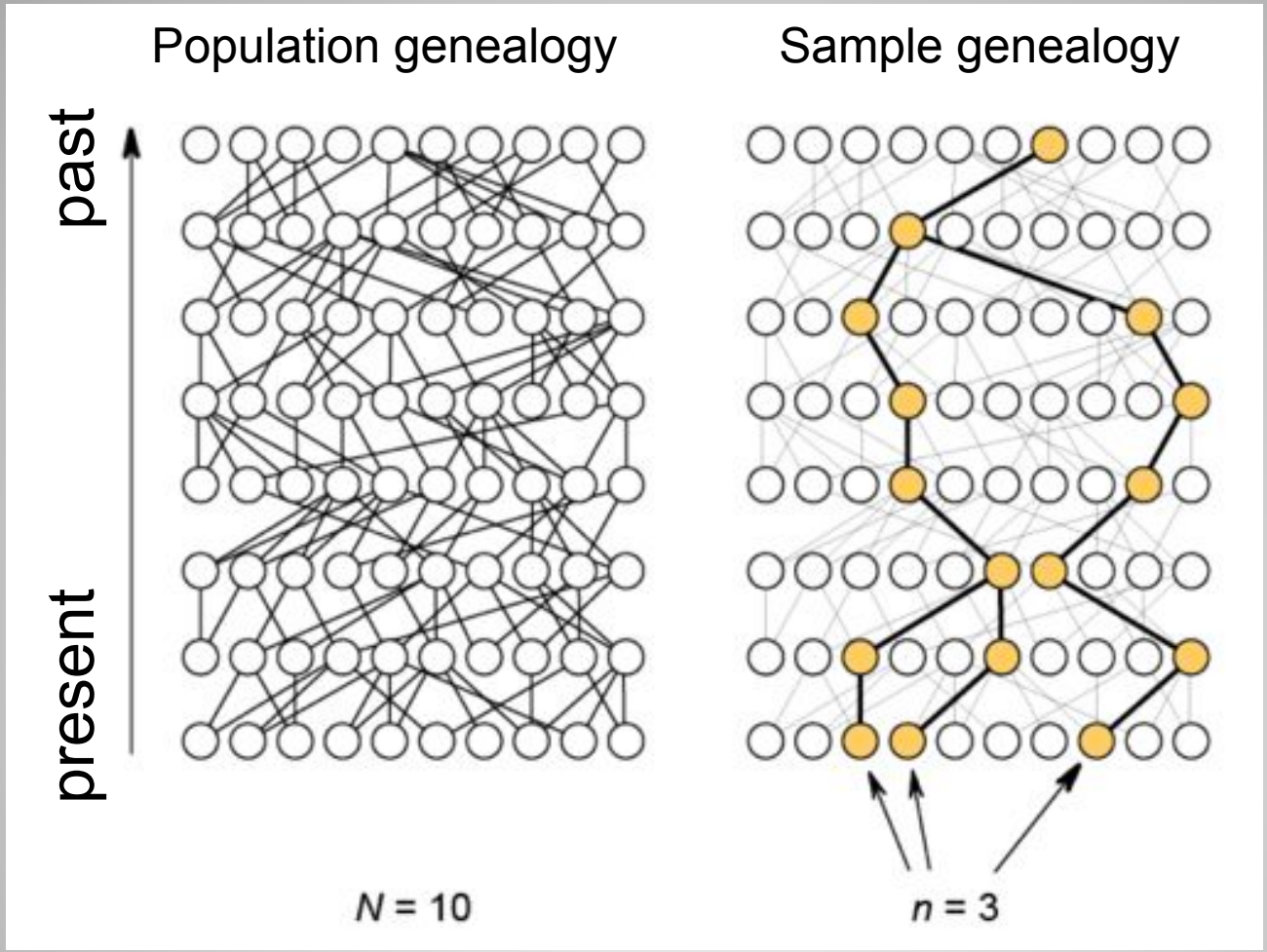


Raphael Leblois
Centre de Biologie et de Gestion des Populations , CBGP
INRA, Montpellier

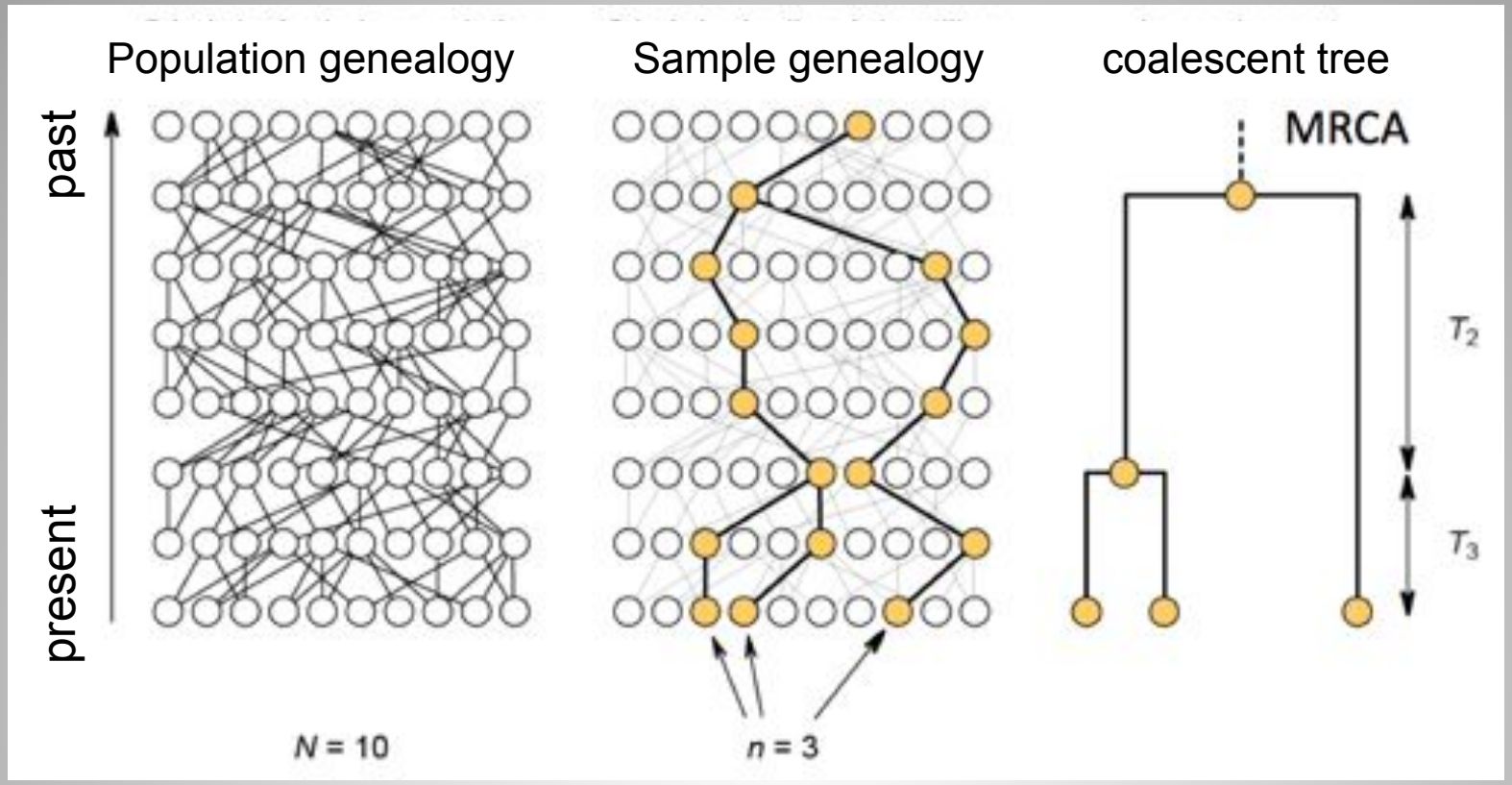
Master MEME, March 2011

Likelihood-based demographic inference using the coalescent

1. Reminder : main coalescence principles
2. Simulating coalescent trees and polymorphism data
3. Likelihood-based inferences
4. Maximum likelihood and Isolation by Distance



In the coalescent theory, we look at the genealogy of a sample of genes going backward in time until the most recent common ancestor (MRCA)



→ **a new approach in population genetics :**

✓ Classical approach

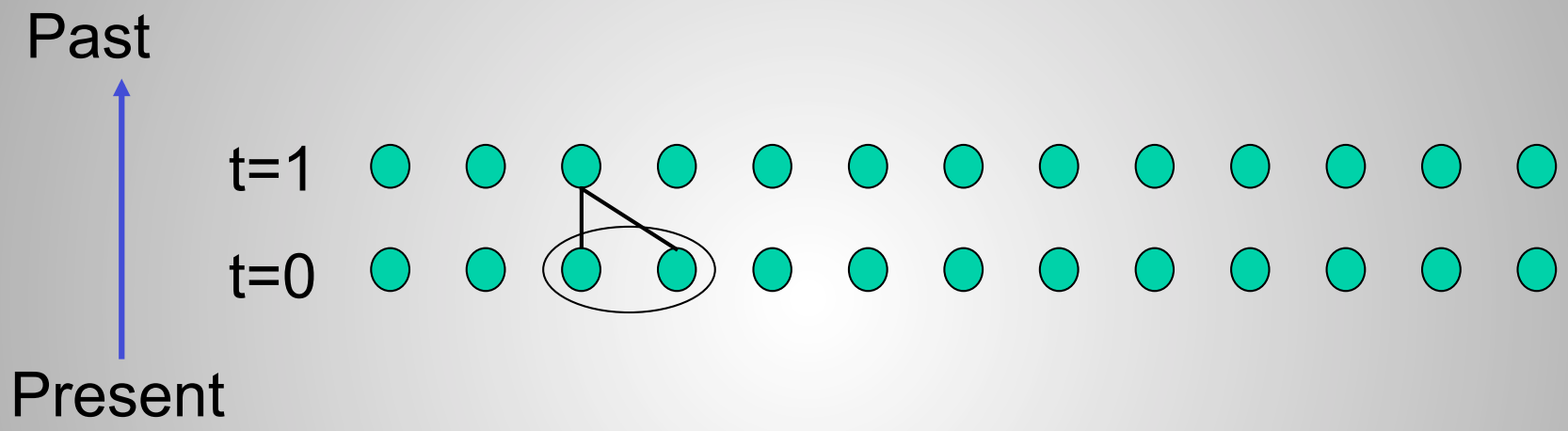
- Population
- Gene frequencies
- Forward in time

✓ Coalescent approach

- Sample
- Gene Genealogies
- backward in time

Coalescence of 2 genes in one generation

in a haploid population of size *N*

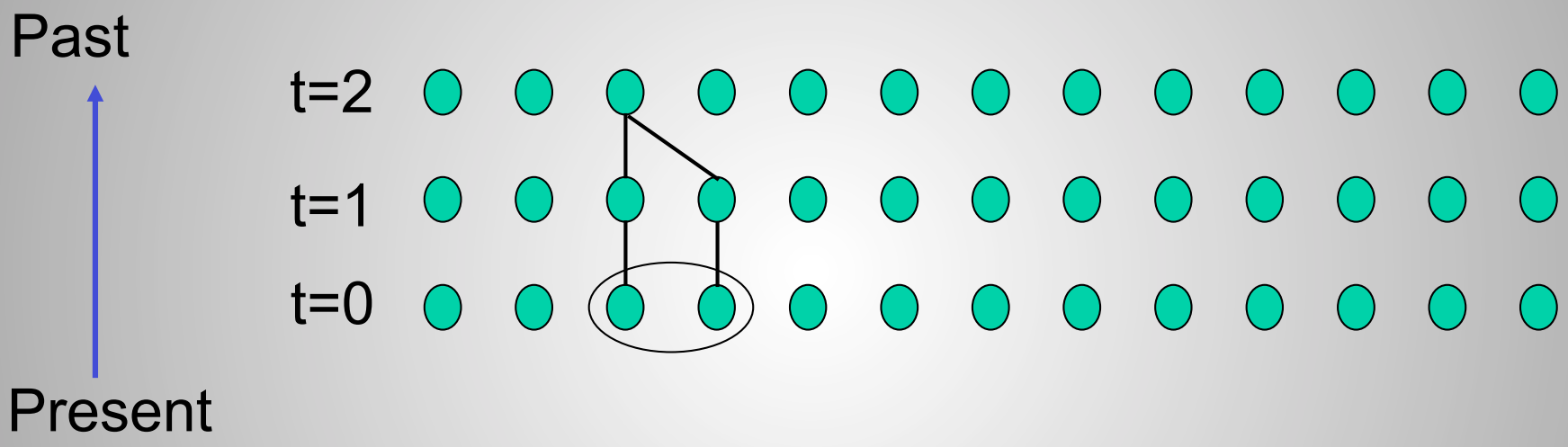


Probability of coalescence of 2 genes in one generation
= probability that the two genes have a common parental gene

$$P(T_2 = 1) = \frac{1}{N}$$

Coalescence of 2 genes in 2 generations

in a haploid population of size *N*

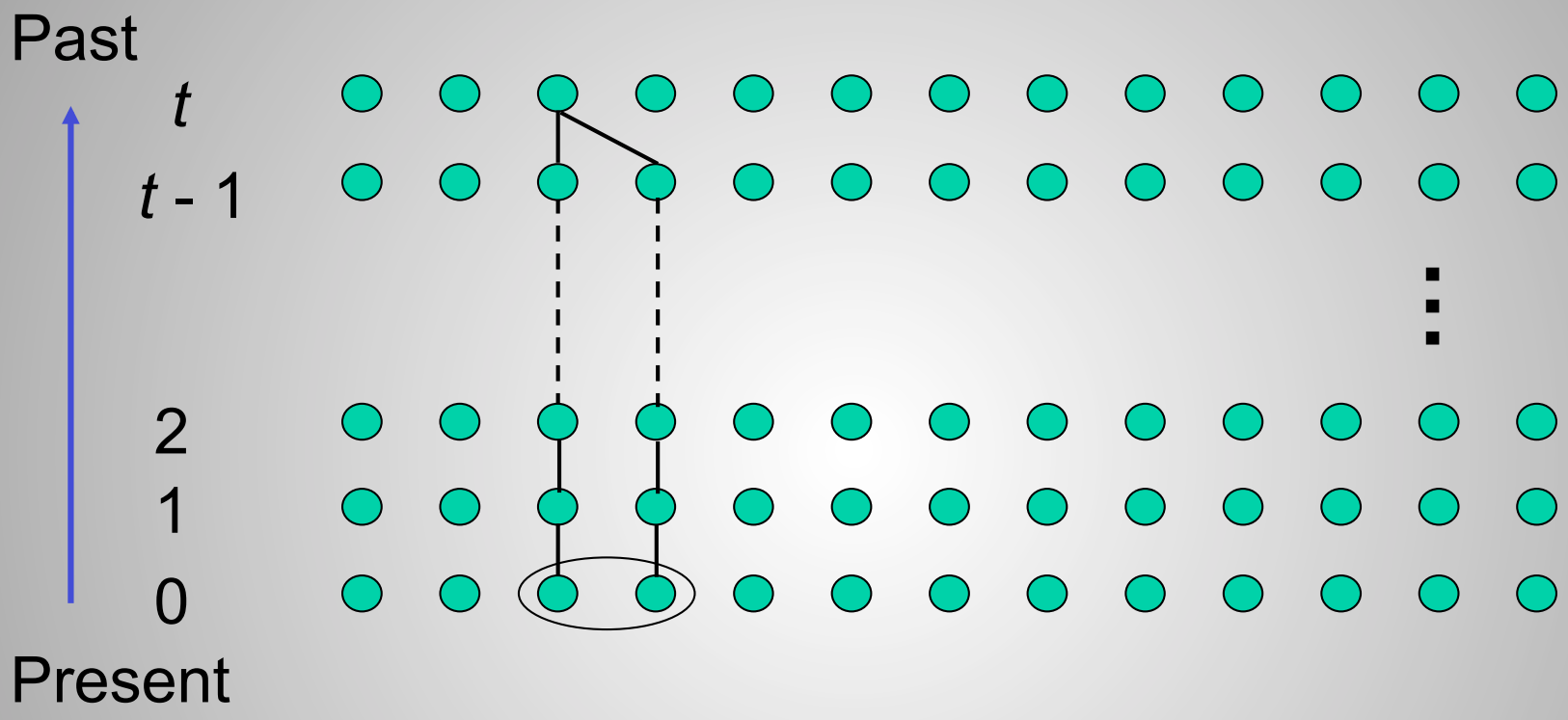


(Prob. that the 2 gene do not coalesce at *t*=1)
*(Prob. that the 2 gene coalesce at t=2)

$$P(T_2 = 2) = \left(1 - \frac{1}{N}\right)\frac{1}{N}$$

Coalescence of two genes in *t* generations

in a haploid population of size *N*



(Prob. that the 2 gene do not coalesce in the first *t*-1 generations)
*(Prob. that the 2 gene coalesce at *t*)

$$P(T_2 = t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N}$$

Coalescence of two genes in *t* generations

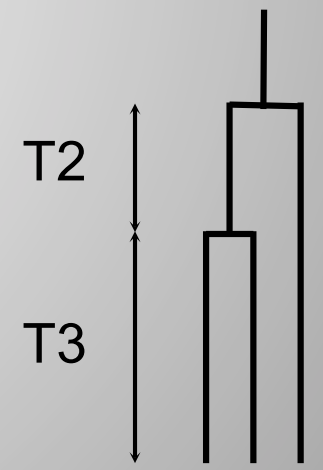
in a haploid population of size *N*

for $x << 1 \quad (1-x)^t \approx e^{-xt}$

The discret geometric distribution can be approximated by an continuous exponential distribution for large *N*

$$P(T_2 = t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N} \approx \frac{1}{N} e^{-Nt}$$

Coalescence times follow an exponential ditribution of rate *N* (it is also its expectation)



Coalescence of *j* genes in *t* generations
in a haploid population of size *N*

Assumption: no multiple coalescence for large *N*

$\binom{j}{2}$ = *j**(*j* - 1)/2 gene pairs can coalesce with probability 1/*N*

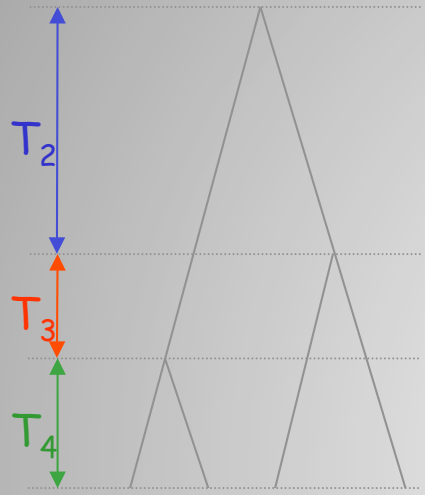
$$\Pr(\text{two genes among } j \text{ coalesce in one generation}) = \frac{j(j-1)}{2N}$$

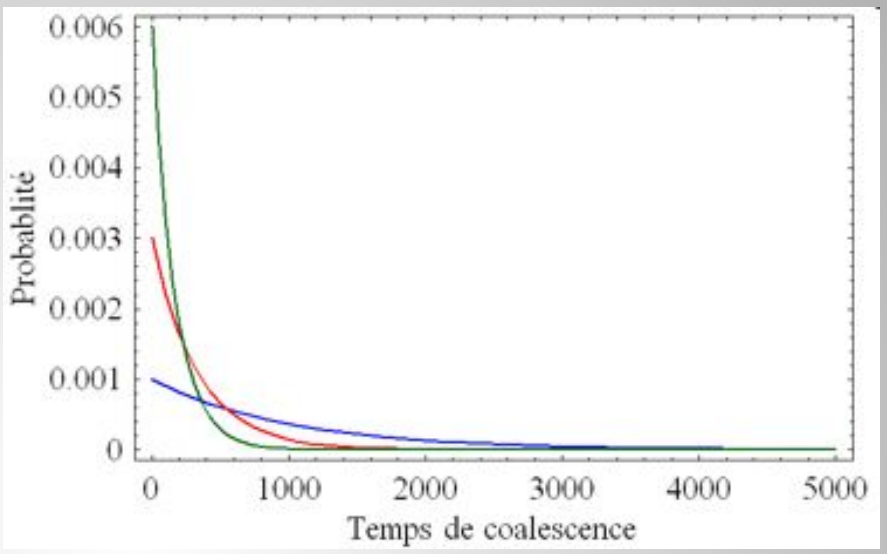
coalescence times for a sample of *j* genes/lineages follow a geometric distribution with parameter *j**(*j*-1)/2*N*, and can be approximated by an exponential distribution with expectation 2*N* / (*j** (*j* -1))

$$\Pr(T_j = t) = (1 - \frac{j(j-1)}{2N})^{t-1}(\frac{j(j-1)}{2N}) \approx \frac{j(j-1)}{2N} e^{-\frac{j(j-1)}{2N}t}$$

Coalescence of *j* genes in *t* generations

in a haploid population of size *N*





$$E(T_j) = \frac{2N}{j(j-1)}$$

the larger the sample size or lineage number is, the larger the expected coalescence times are

$$\mathrm{var}(T_j) = \frac{4N^2}{j^2(j-1)^2}$$

Coalescence times have high variance : two independent loci could show very different coalescence times, and thus very different coalescent trees (genealogies)

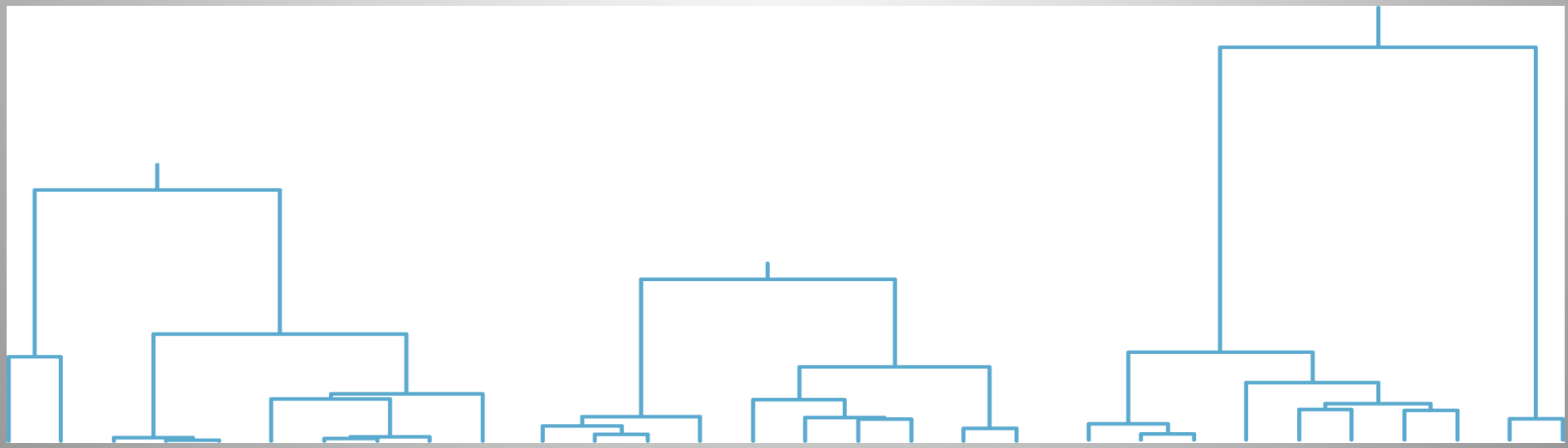
Coalescence en *t* générations de *j* lignées

$$E(T_j) = \frac{2N}{j(j-1)}$$

the larger the sample size or lineage number is, the larger the expected coalescence times are

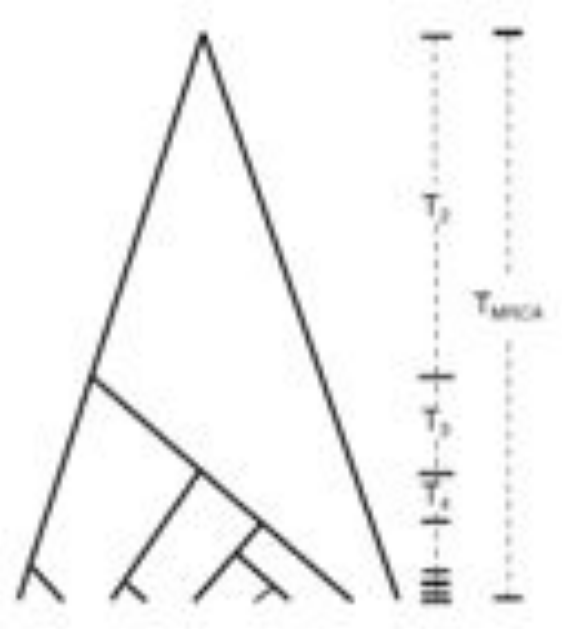
$$\mathrm{var}(T_j) = \frac{4N^2}{j^2(j-1)^2}$$

Coalescence times have high variance : two independent loci could show very different coalescence times, and thus very different coalescent trees (genealogies)



TMRCA : length of the coalescent trees

TMRCA = Time to the Most Recent Common Ancestor
= Time of the last node (coalescence) of the tree
= Tree length



$$E[TMRCA] = \sum_{i=2}^{j} E[T_i] = \sum_{i=2}^{j} \frac{2N}{i(i-1)}$$

$$= 2N \times \sum_{i=2}^{j} \left(\frac{1}{i-1} - \frac{1}{i}\right)$$

$$= 2N\left(1 - \frac{1}{j}\right)$$

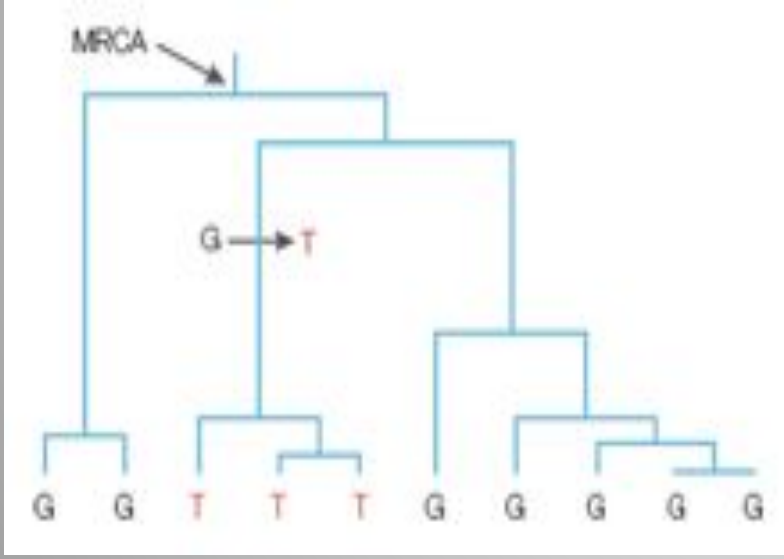
✓ TMRCA expectation tends to 2*N* for large samples

✓ TMRCA of a relatively small sample is close the TMRCA of the whole population

coalescent trees and mutations

Under neutrality assumption, mutations are independent of the genealogy, because genealogical process strictly depends on demographic parameters

First, genealogies are build given the demographic parameters considered (e.g. *N*),



Then mutation are added a posteriori on each branch of the genealogy, from MRCA to the leaves

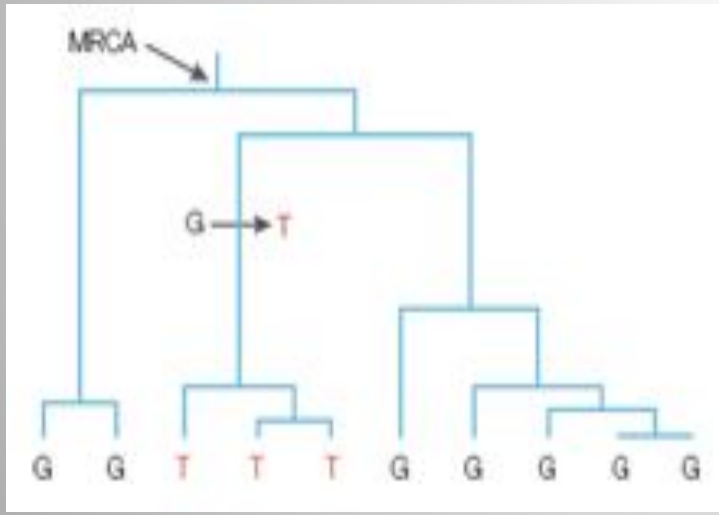
We thus obtain polymorphism data under the demographic and mutational model considered

coalescent trees and mutations

The number of mutations on each branch is a function of the mutation rate of the genetic marker (μ) and the branch length (t).

μ = mean number of mutation per locus per generation.

e.g. 5.10^{-4} for microsatellites, 10^{-7} per nucleotide for DNA sequences



For a branch of length t, the number of mutation thus follows a binomial distribution with parameters (μ,t).

Often approximated by a Poisson distribution with parameter (μ*t).

$$\Pr(k \text{ mut} | t) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

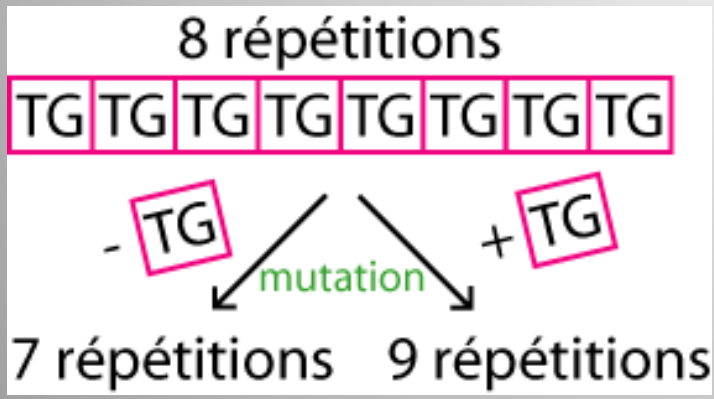
Arbre de coalescence et mutations

Different mutational models for the different genetic markers, e.g. :

✓ For DNA sequences : mutation matrix for different nucleotide transition rates (Pr[A➔T], Pr[A➔C], Pr[T➔G], etc…)

✓ For SNPs : Pr(Ancestral ➔Derived)=μ, Pr(Anc. ➔Der.)=0

✓ For microsatellites :stepwise models



SMM (Stepwise mutation model): each mutation add or remove a motif to the parental allele

Main advantages of the coalescent

- The coalescent is a powerful probabilistic model for gene genealogies

The genealogy of a population genetic sample, and more generally its evolutionary history, is often unknown and cannot be repeated
⇒ the coalescent allows to take this unknown history into account

- The coalescent often simplifies the analyses of stochastic population genetic models and their interpretation

Genetic data polymorphism largely reflects the underlying genealogy
⇒ the coalescent greatly facilitate the analysis of the observed genetic variability and the understanding of evolutionary processes that shaped the observed genetic polymorphism.

Main advantages of the coalescent

- The coalescent allows extremely efficient simulations of the expected genetic variability under various demo-genetic models (sample vs. entire population)

- The coalescent allows the development of powerful methods for the inference of populational evolutionary parameters (genetic, demographic, reproductive,…), some of those methods uses all the information contained in the genetic data (likelihood-based methods)

Trees and polymorphism data simulation

- reminder :
  - ✓ For neutral markers, the number of offspring is independent of the genetic types of their parents

    ➔ Demographic processes are thus independent of mutational processes

  - ✓ Simulation of polymorphism data can thus be done in two steps :

(1) Tree simulation : topology and branch length

(2) Addition of mutations on the tree

Coalescent tree simulation

Two main methods :

- **Coalescent continuous approximations**

  very fast but approximations only valid for large population sizes, weak mutation and migration rates and "simple" demographic models

- **Generation by generation**

  Ok for all demograophic and mutational models but relatively slow

RAPIDITY :

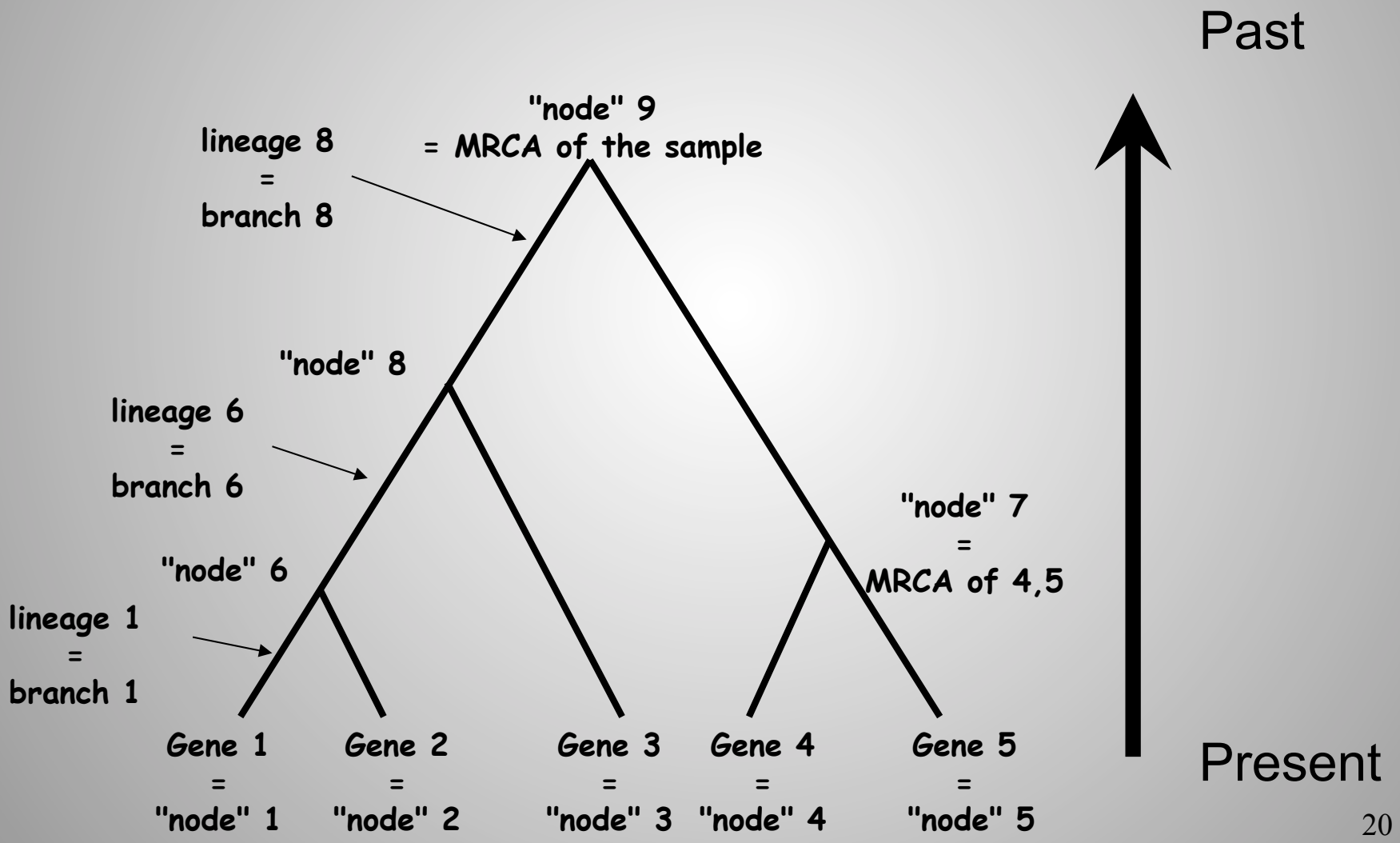
Continuous approximations > Generation by generation

FLEXIBILITY :

Generation by generation > Continuous approximations

Coalescent tree simulation

- Tree representation



Coalescent tree simulation
Generation par generation

- very simple and exact (without any approximations):
  - ✓ Go backward in time generation by generation
  - ✓ At each generation, we stochastically draw potential events affecting the genealogy

    e.g. coalescence, migration, recombinaison
  - ✓ Stop at the most recent common ancestor of all sampled genes = MRCA

Coalescent tree simulation
Generation par generation

- Toy example :
  - ✓ 4 gene sample
  - ✓ single neutral locus
  - ✓ panmictic haploid population of size N=10

Coalescent tree simulation
Generation par generation

- Example : 4 genes, neutral, 1 pop *N*=10

nodes / lineages numbering	1	2	3	4
random number between 1 and *N* for each lineage				
lineage starting generation	0	0	0	0

① ② ③ ④

Gn=0

Coalescent tree simulation
Generation par generation

- Example : 4 genes, neutral, 1 pop N=10

nodes / lineages numbering	1	2	3	4
random number between 1 and N for each lineage				
lineage starting generation	0	0	0	0

Prob for a coalescence in j lineages in one generation

$=j(j-1)/2N$

= probability of drawing 2 identical integers in j uniform drawings between 1 and N

Gn=0

Coalescent tree simulation
Generation par generation

- Example : 4 genes, neutral, 1 pop N=10

  Probability of a coalescence in j lineages in one generation

  $=j(j-1)/2N$

  = probability of drawing 2 identical integers in j uniform drawings between 1 and N

  in other terms, we randomly and uniformly draw a parent for each gene/lineage among the N potential parents (stable population size)

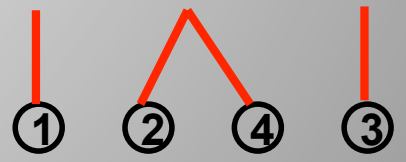
  Genes/lineages sharing the same parent coalesce

Coalescent tree simulation
Generation par generation

- Example : 4 genes, neutral, 1 pop *N*=10

nodes / lineages numbering	1	2	3	4
random number between 1 and *N* for each lineage	2	6	5	6
lineage starting generation	0	0	0	0

Coalescence at generation 1 of nodes/lineages 3 and 4



Gn=1

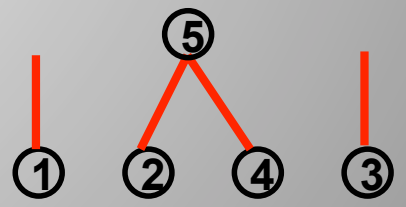
Coalescent tree simulation
Generation par generation

- Example : 4 genes, neutral, 1 pop *N*=10

nodes / lineages numbering	1	3	5
random number between 1 and *N* for each lineage	2	5	6
lineage starting generation	0	0	1

Coalescence at generation 1 of nodes/lineages 3 and 4

new node 5



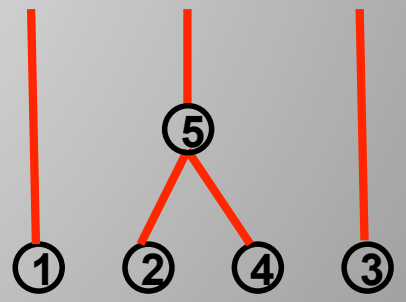
Gn=1

Coalescent tree simulation
Generation par generation

- Example : 4 genes, neutral, 1 pop *N*=10

nodes / lineages numbering	1	3	5
random number between 1 and *N* for each lineage	3	1	7
lineage starting generation	0	0	1

nothing happened at generation 2



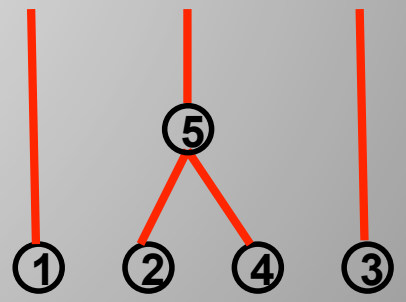
Gn=2

Coalescent tree simulation
Generation par generation

- Example : 4 genes, neutral, 1 pop *N*=10

nodes / lineages numbering	1	3	5
random number between 1 and *N* for each lineage	7	4	8
lineage starting generation	0	0	1

nothing happened at generation 3



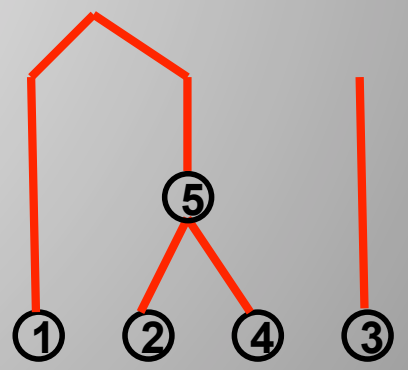
Gn=3

Coalescent tree simulation
Generation par generation

- Example : 4 genes, neutral, 1 pop *N*=10

nodes / lineages numbering	1	3	5
random number between 1 and *N* for each lineage	5	2	5
lineage starting generation	0	0	1

Coalescence at generation 4 of nodes/lineages 1 and 5



Gn=4

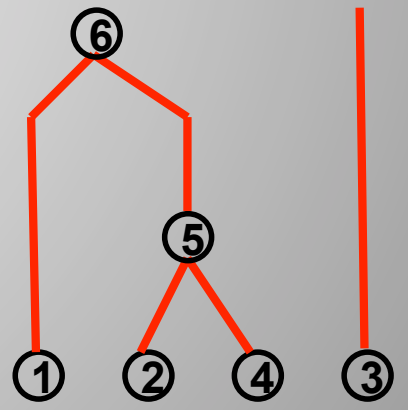
Coalescent tree simulation
Generation par generation

- Example : 4 genes, neutral, 1 pop *N*=10

nodes / lineages numbering	3	6
random number between 1 and *N* for each lineage	2	5
lineage starting generation	0	5

Coalescence at generation 4 of nodes/lineages 1 and 5

new node 6



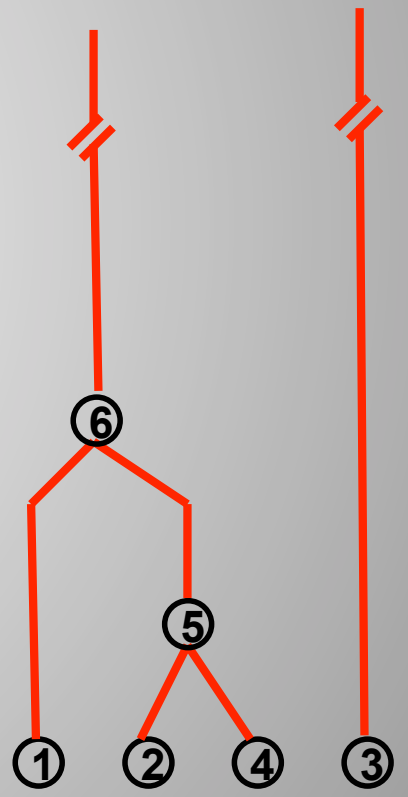
Gn=4

Coalescent tree simulation
Generation par generation

- Example : 4 genes, neutral, 1 pop *N*=10

nodes / lineages numbering	3	6
random number between 1 and *N* for each lineage	3	9
lineage starting generation	0	5

nothing at generation 5,6,…



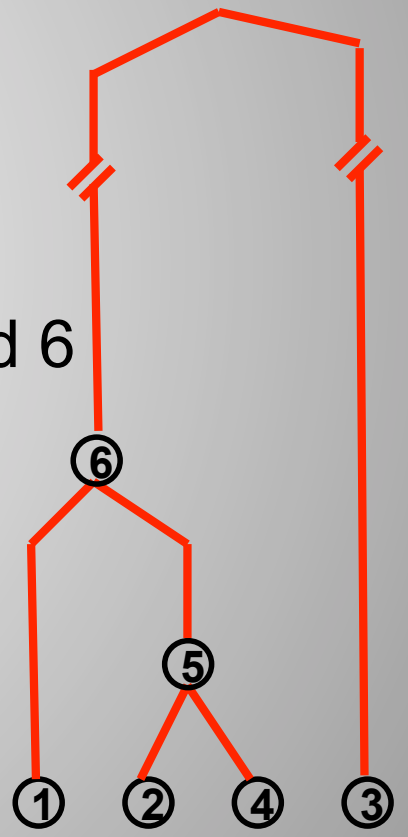
Gn=5

Coalescent tree simulation
Generation par generation

- Example : 4 genes, neutral, 1 pop *N*=10

nodes / lineages numbering	3	6
random number between 1 and *N* for each lineage	7	7
lineage starting generation	0	5

Coalescence at generation 20 of the two last lineages 3 and 6



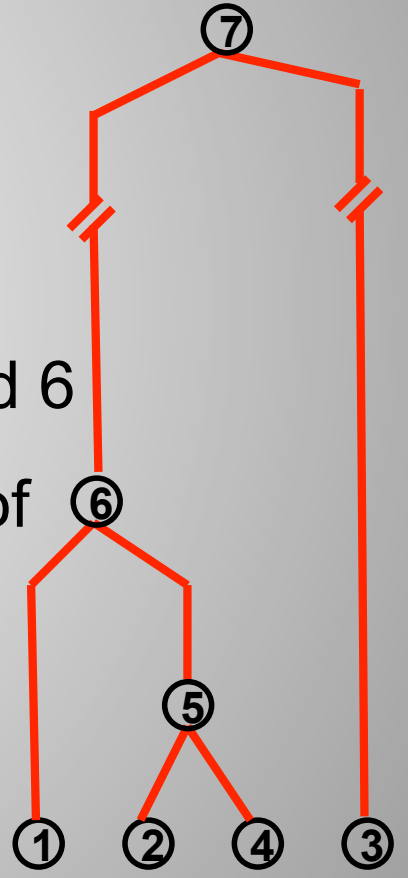
Gn=20

Coalescent tree simulation
Generation par generation

- Example : 4 genes, neutral, 1 pop *N*=10

nodes / lineages numbering	3	6
random number between 1 and *N* for each lineage	7	7
lineage starting generation	0	5

Coalescence at generation 20 of the two last lineages 3 and 6 new node 7 = MRCA of the sample



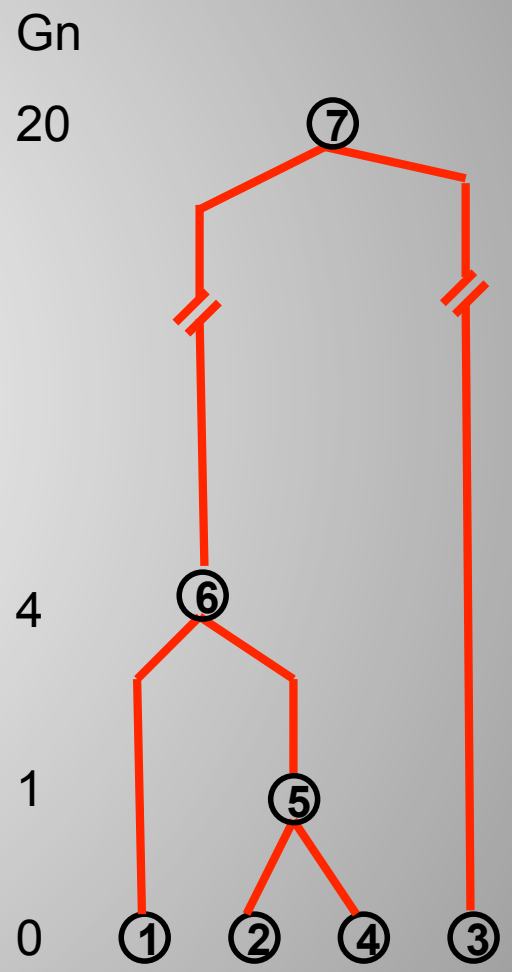
Gn=20

Coalescent tree simulation
Generation par generation

The coalescence tree (topology and branch lengths) is build.

It is a stochastic process, so if we build many trees, they will all be different but share some common properties.

To get polymorphism data, we need to add mutations on the tree…



Coalescent tree simulation
Hudson continuous approximations

- Principle: 2 successive steps

  (1) The topology of the tree is build by randomly coalescing lineages

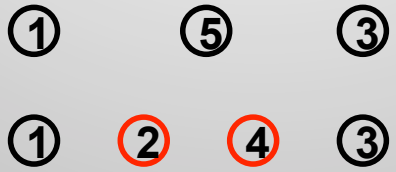
  (2) Branch length are simulated using expected coalescence times between two coalescence events

Coalescent tree simulation
Hudson continuous approximations

- Example : 4 genes, neutral, 1 pop *N*=10

(1) The topology of the tree is build by randomly coalescing lineages

1st coalescence = random draw of 2 lineages among the 4
➔ lineages 2 and 4 coalesce to give lineage 5

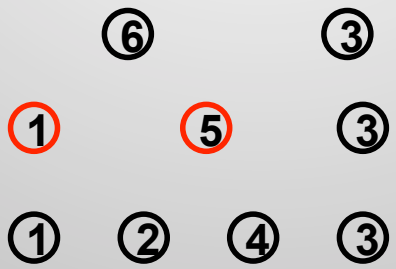


Coalescent tree simulation
Hudson continuous approximations

- Example : 4 genes, neutral, 1 pop *N*=10

(1) The topology of the tree is build by randomly coalescing lineages

2^d coalescence = random draw of 2 lineages among the 3 lineages left ➔ lineages 1 and 5 coalesce to give lineage 6

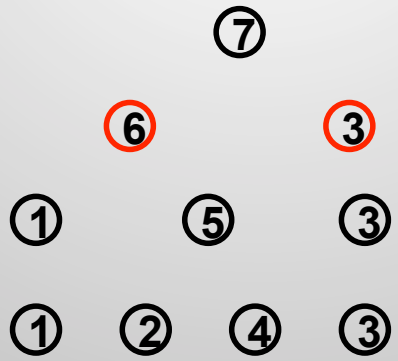


Coalescent tree simulation
Hudson continuous approximations

- Example : 4 genes, neutral, 1 pop *N*=10

(1) The topology of the tree is build by randomly coalescing lineages

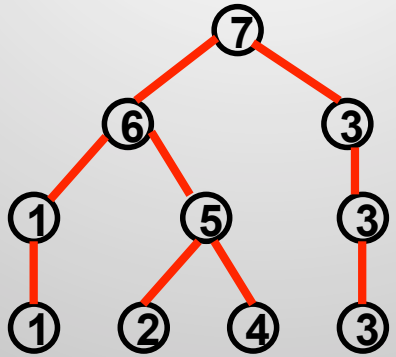
3[d] and last coalescence = the last 2 lineages 6 and 3 coalesce to give lineage 7, the MRCA



Coalescent tree simulation
Hudson continuous approximations

- Example : 4 genes, neutral, 1 pop *N*=10

  (1) Topology is build

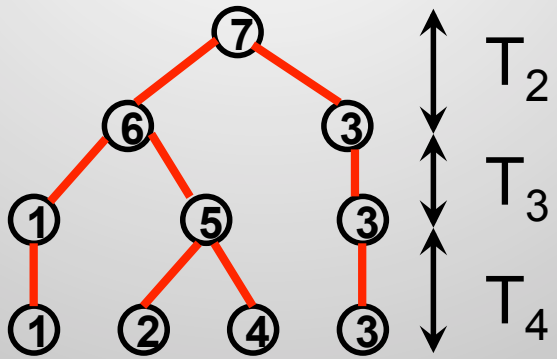


Coalescent tree simulation
Hudson continuous approximations

- Example : 4 genes, neutral, 1 pop *N*=10

(2) Branch length simulation

there are 3 branch lengths to simulate T_4, T_3, T_2



Coalescent tree simulation
Hudson continuous approximations

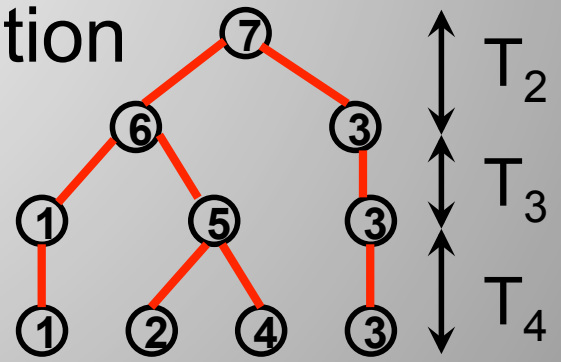
- Example : 4 genes, neutral, 1 pop *N*=10

3 branch lengths to simulate T4, T3, T2

$$\Pr(T_j = k) = \frac{j(j-1)}{2N} e^{\frac{-j(j-1)}{2N}k}$$

T_4 drawn from an exponential distribution with parameter (expectation)

j(*j*-1) /2*N*=4*3/2*10



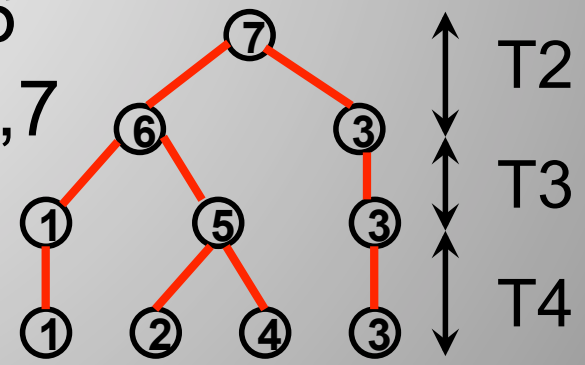
(algorithms to draw exponential deviates are availables)

Coalescent tree simulation
Hudson continuous approximations

- Example : 4 genes, neutral, 1 pop *N*=10

3 branch lengths to simulate T4, T3, T2

Ex:
T4 drawn from exp. (j(j-1) /2N=4*3/2*10) ➔ 1,2
T3 drawn from exp. (3*2/2*10) ➔ 2,6
T2 drawn from exp. (2*1/2*10) ➔ 15,7

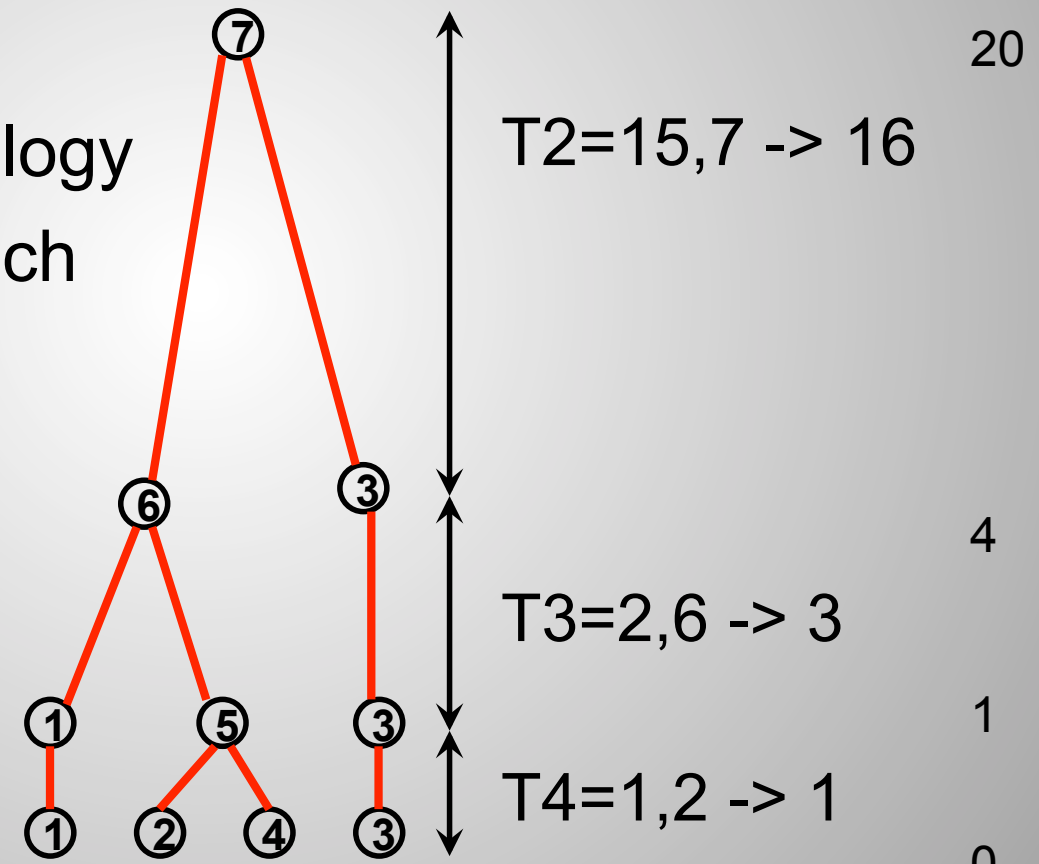


Coalescent tree simulation
Hudson continuous approximations

Example : 4 genes, neutral, 1 pop *N*=10

We then have the topology and branch length, which correspond to the total coalescent tree

Coalescence times distributions must be known under the demographic model considered!



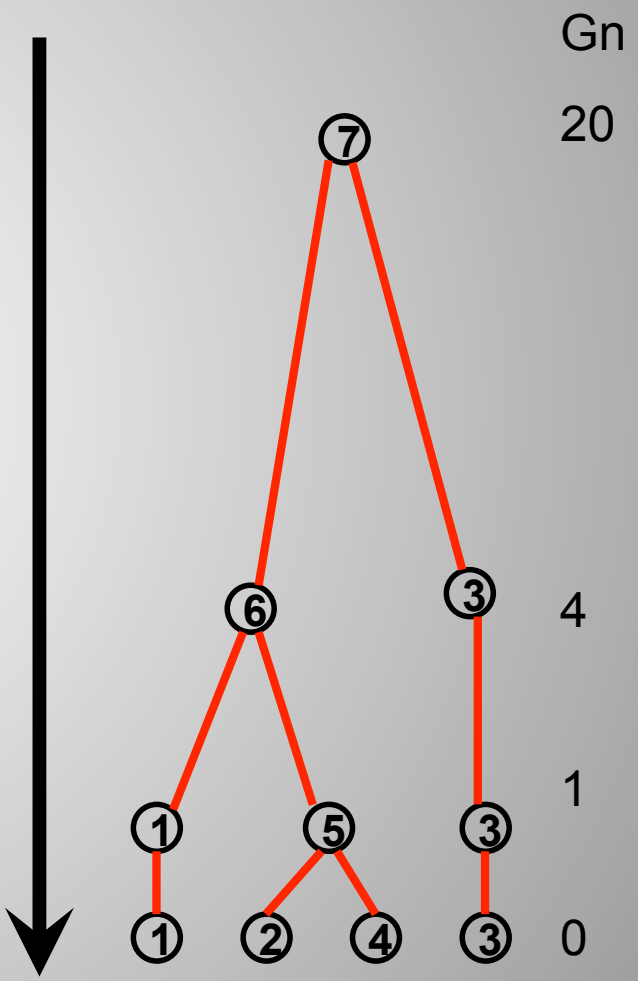
Polymorphism data simulation starting from a coalescent tree

General principle (reminder) :

Mutations are distributed on the different branches from the MRAC to the leaves as a function of the mutation rate μ

Each mutation induce a change in the allelic/nucleotidic state of the descending node

This genetic state change is made according to the mutational model considered, which may reflect real mutational processes of some genetic markers

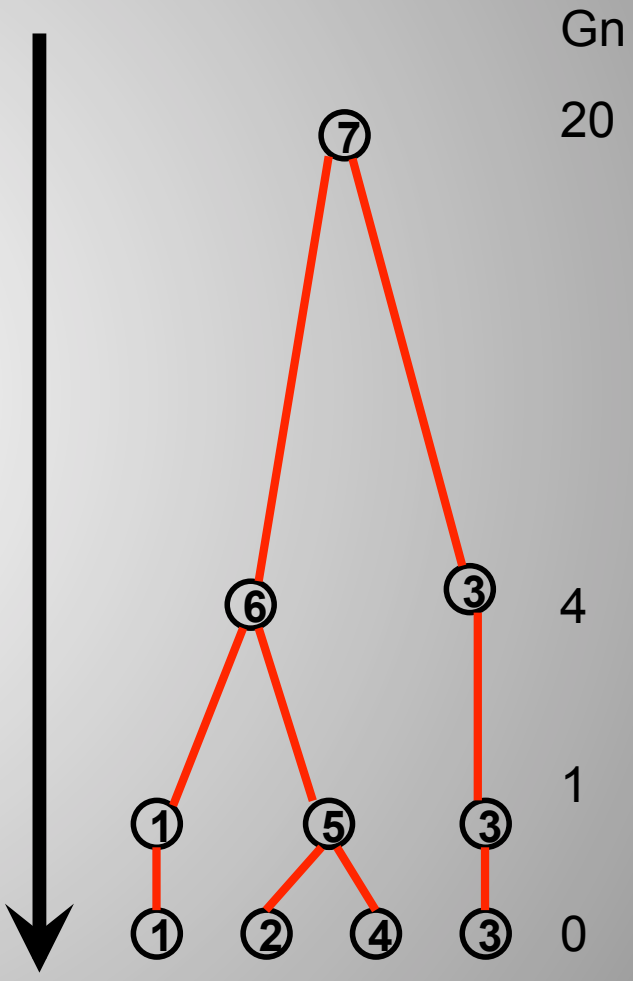


Polymorphism data simulation starting from a coalescent tree

On a branch of length t, the number of mutation follow a binomial with parameters (μ,t)

Approximated by a Poisson distribution with parameter (μ*t)

$$\Pr(k \text{ mut} | t) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

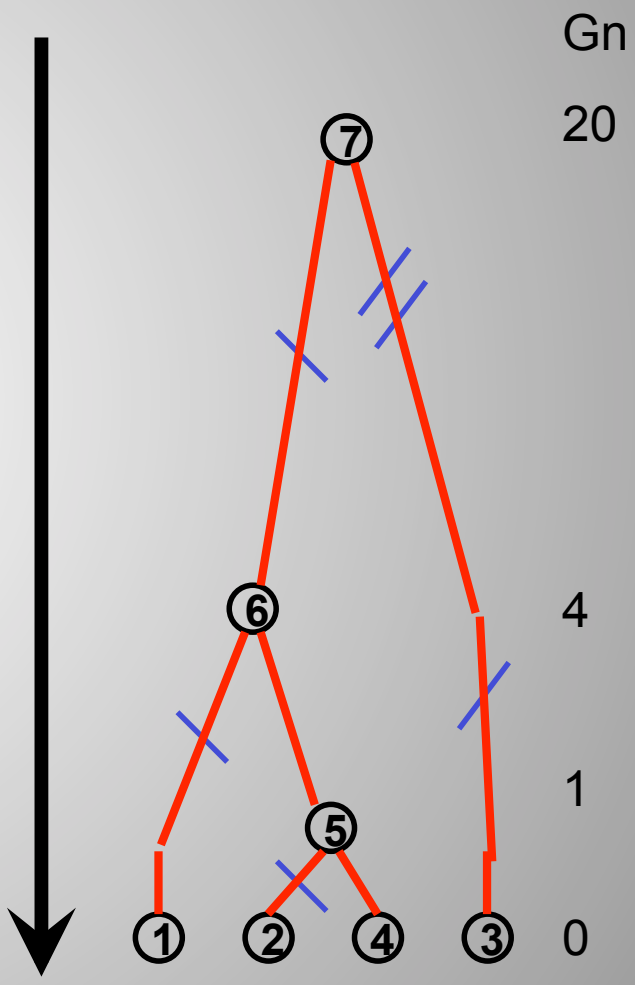


Polymorphism data simulation starting from a coalescent tree

Example for microsatellites under a SMM : gain or loss of a motif (repeat) for each mutation

addition of mutation numbers on each branch following the Poisson distribution

$$\Pr(k \text{ mut} | t) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$



Polymorphism data simulation starting from a coalescent tree

Example for microsatellites under a SMM : gain or loss (p=0.5) of a motif (repeat) for each mutation

Choice of the MRCA type (random): 20

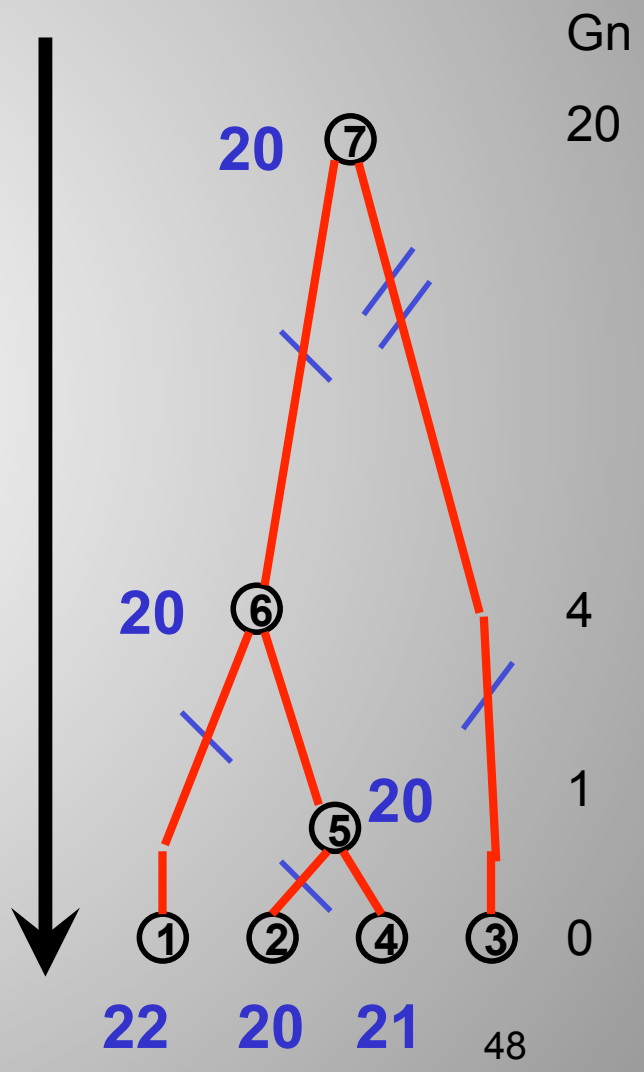
node 7 to 6 : one time ±1 ➔ 21

node 6 to 1 : one time ±1 ➔ 22

node 6 to 5 : 0 time ±1 ➔ 21

node 5 to 2 : one time ±1 ➔ 20

node 5 to 4 : 0 time ±1 ➔ 21

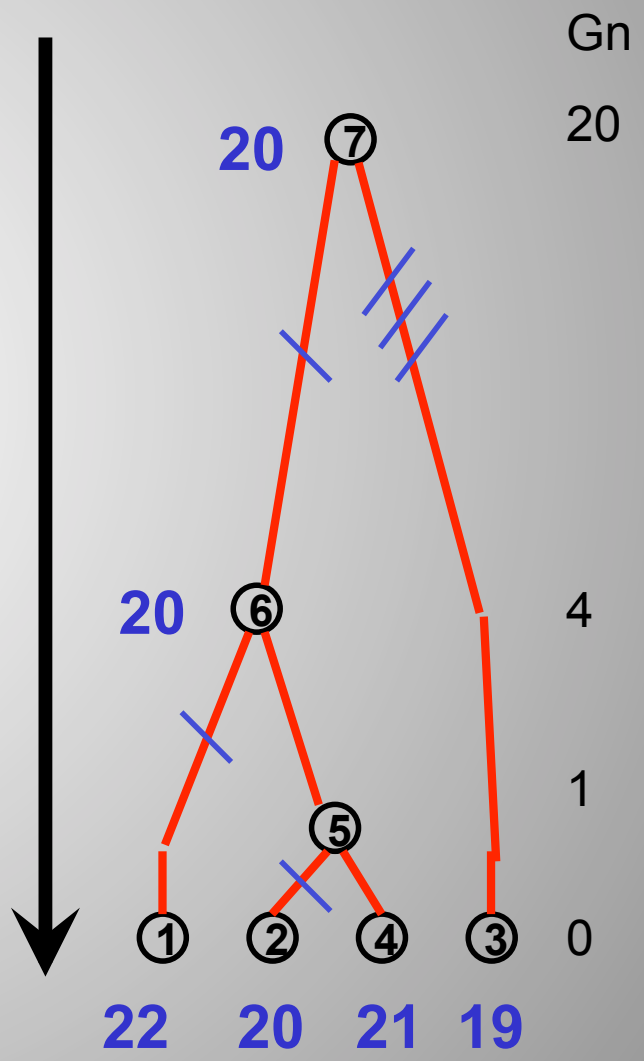


Polymorphism data simulation starting from a coalescent tree

Example for microsatellites under a SMM : gain or loss (p=0.5) of a motif (repeat) for each mutation

node 7 to 3 : 3 times ±1 ➔ 19

A polymorphism sample of 4 genes is obtained with allelic states 19, 20, 21, 22



Polymorphism data simulation starting from a coalescent tree

Example on DNA sequence markers (5 bp).

Choice of the ancestral sequence (ATTGC)

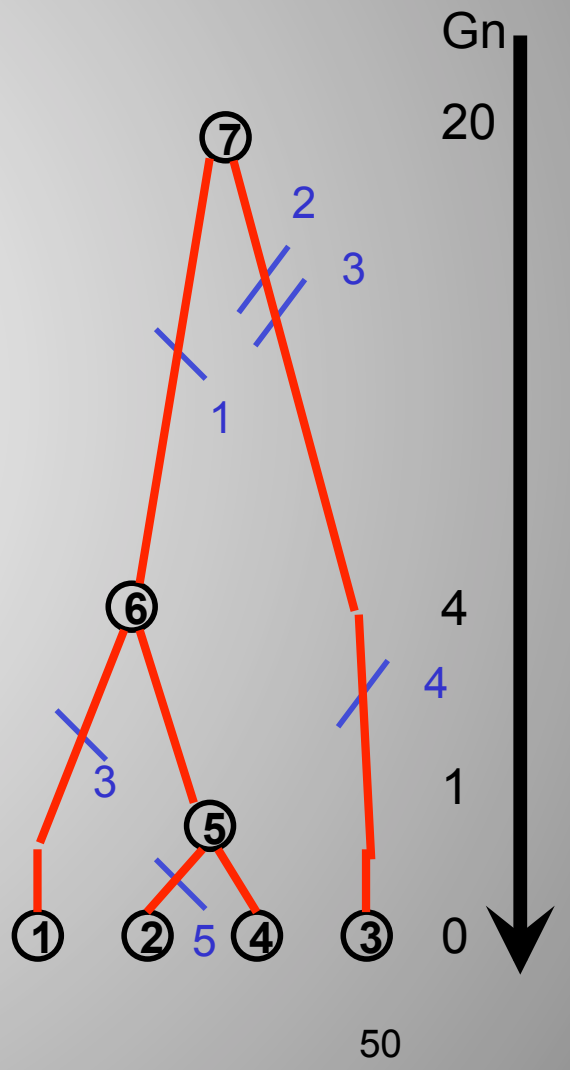
independent mutation on each site

7 to 6 : 1 mut on site 1 ➔ TTTGC

6 to 1 : 1 mut on site 3 ➔ TTAGC

5 to 2 : 1 mut on site 5 ➔ TTTGG

7 to 3 : 1 mut on each site 2,3,4 ➔ AAACC



Polymorphism data simulation starting from a coalescent tree

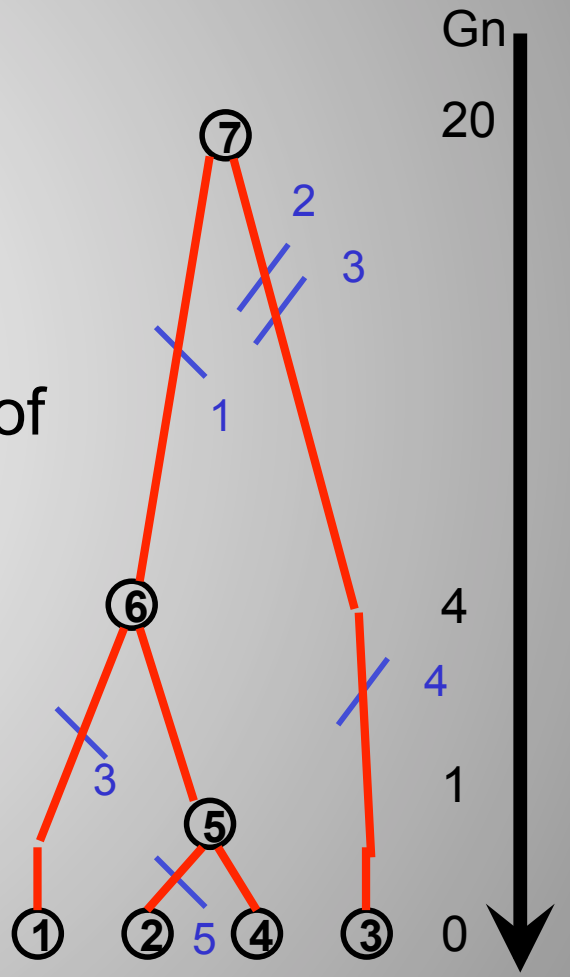
Example on DNA sequence markers (5 bp).

Choice of the ancestral sequence (ATTGC)

independent mutation on each site

The polymorphism sample is then composed of 4 different sequences :

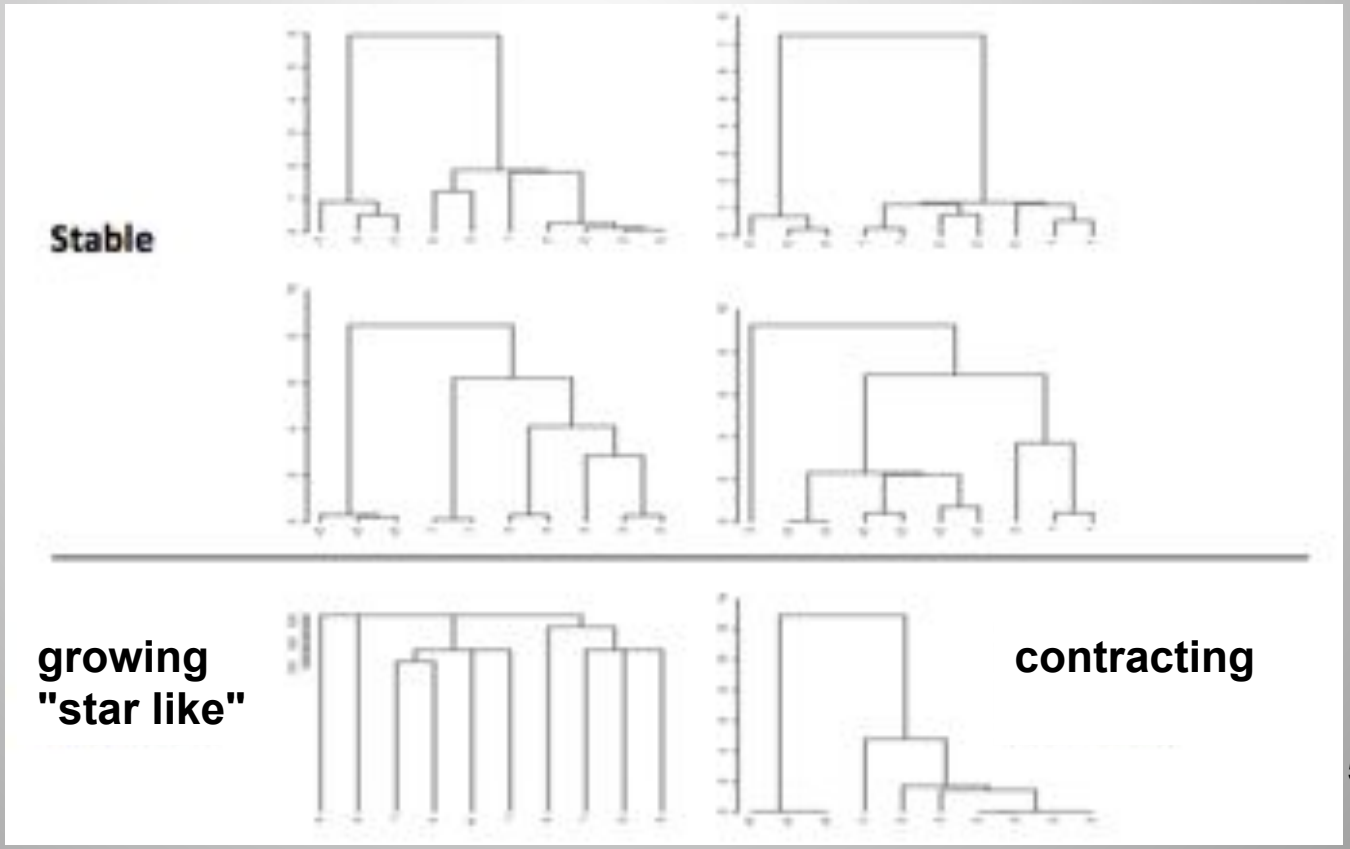
TTAGC,TTTGG,TTTGC,AAACC



what can we do with those coalescent trees and genetic data simulation?

- Exploratory approaches : to study the effects of various parameters on the shape of coalescent trees and on the distribution of polymorphism in a sample

Ex: past demography effects



what can we do with those coalescent trees and genetic data simulation?

- Exploratory approach : demographic effects
  - growing population size (e.g. invasion of a new habitat)

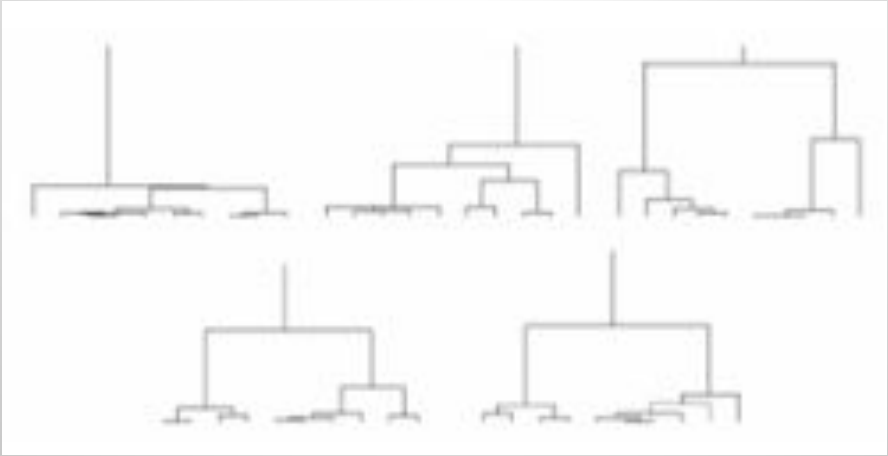
  There are more ancient coalescences (small *N*) than recent coalescences (large *N*), coalescent trees thus have longer terminal branches

A population size growth induces an excess of low frequency alleles (rare alleles)

what can we do with those coalescent trees and genetic data simulation?

- Exploratory approach : demographic effects
  - population size contraction (e.g. threatened species)

  There are more recent coalescences (small *N*) than ancient coalescences (large *N*), coalescent trees thus have shorter terminal branches



A contraction induces a deficit of low frequency alleles

what can we do with those coalescent trees and genetic data simulation?

- **Exploratory approach :** to study the effects of various parameters on the shape of coalescent trees, on the distribution of polymorphism in a sample and on various sumary statistics computed on a genetic sample (e.g. H_e, F_{ST},…)
- **Simulation tests** : to create simulated data sets to test the precision and robustness of genetic data analysis methods
- **Inferential approach** : to estimate populational evolutionary parameters (pop sizes, dispersal, demographic history) from polymorphism data

Demographic inference under the coalescent

- Inferential approaches are based on the modeling of population genetic processes. Each population genetic model is characterized by a set of demographic and genetic parameters *P*
- The aim is to infer those parameters from a polymorphism data set (genetic sample)
- The genetic sample is then considered as the realization ("output") of a stochastic process defined by the demo-genetic model

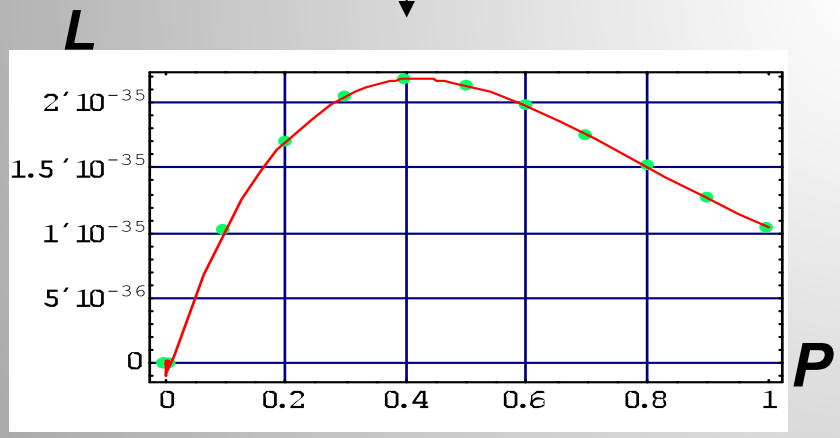
Demographic inference under the coalescent

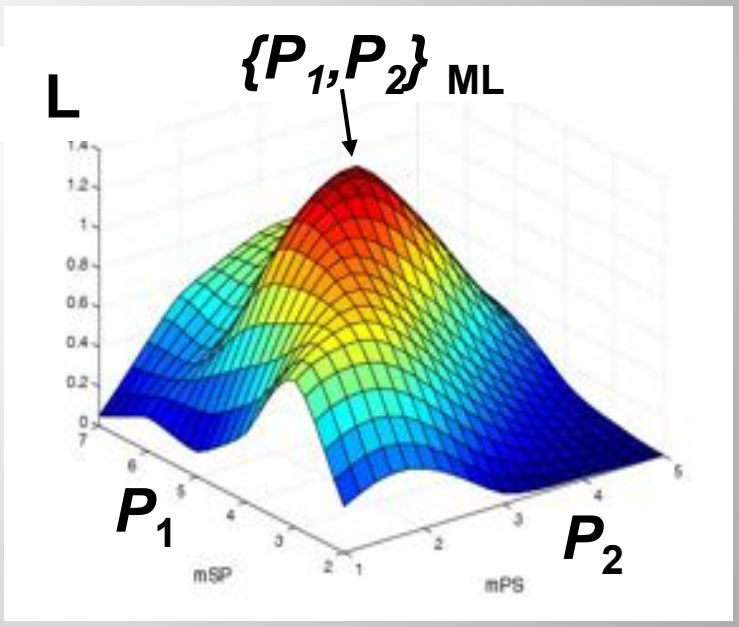
- First, compute or estimate the probability Pr(D |P^*) of observing the data D given some parameter values P^*, it is the likelihood : L(P^* | D)=Pr(D | P^*)
- Second, find the set of parameter values that maximize this probability of observing the data (maximum likelihood method)

Demographic inference under the coalescent

- Maximum likelihood method

P_{ML} = maximum likelihood estimate





!! many parameters → large parameter space to explore !!

Demographic inference under the coalescent

- Problem : Most of the time, the likelihood Pr($D|P$) of a genetic sample cannot be computed directly because there is no explicite mathematical expression
- However, the probability Pr($D|P,G_i$) of observing the data D given a specific genealogy G_i and the parameter values P can be computed.
- then we take the sum of all genealogy-specific likelihoods on the whole genealogical space, weighted by the probability of the genealogy given the parameters :

$$L(P|D) = \int_G \Pr(D|G;P)\Pr(G|P)dG$$

Demographic inference under the coalescent

- The likelihood can be written as the sum of $\Pr(D|P,G_i)$ over the genealogical space (all possible genealogies) :

$$L(P|D) = \int_G \Pr(D|G;P)\Pr(G|P)dG$$

mutational parameters

Coalescent theory demographic parameters

- Genealogies are nuisance parameters (or missing data), they are important for the computation of the likelihood but there is no interest in estimating them

very different from the phylogenetic approaches

Demographic inference under the coalescent

$$L(P|D) = \int_G \Pr(D|G;P)\Pr(G|P)dG$$

Sum over all possible genealogies

⇒ usually untractable !!!

Monte Carlo simulations are used : a large number *K* of genealogies are simulated according to Pr(*G*|*P*) and the mean over those simulations is taken as the expectation of Pr(*D*|*G*;*P*) :

$$L(P|D) = E_{pr(G|P)}\left[\Pr(D|G;P)\right] \approx \frac{1}{K}\sum_{k=1}^{K}\Pr(D|G_k;P)$$

simulation of many genealogies is necessary to get a good estimation of the likelihood

Demographic inference under the coalescent

$$L(P|D) = E_{pr(G|P)}\left[\Pr(D|G;P)\right] \approx \frac{1}{K}\sum_{k=1}^{K}\Pr(D|G_k;P)$$

Monte Carlo simulations are often not very efficient because there are too many genealogies giving extremely low probabilities of observing the data, more efficient algorithms are used to explore the genealogical space and focuss on genealogies well suppported by the data.

Demographic inference under the coalescent

More efficient algorithms :

- IS : Importance Sampling
- MCMC : Monte Carlo Markov chains associated with Metropolis-Hastings algorithm

allows better exploration of the genealogies proportionnaly to their probability of explaining the data P(D|P;G).

Demographic inference under the coalescent
the approach of Felsenstein et al. (MCMC)

- Probability of a genealogy given the parameters of the demographic model $\Pr(G_i|P)$ can be computed from the continuous time approximations (cf. Hudson approximations to construct coalescent trees)
- then the probability of the data given a genealogy and mutational parameters $\Pr(D|G_i,P)$ can be easily computed from the mutation model parameters, the mutation rate and the Poison distribution of mutations.
- From this, an efficient algorithm to explore the genealogical and the parameter spaces should allows the inference of the likelihood over the spaces.

Demographic inference under the coalescent
the approach of Felsenstein et al. (MCMC)

- Probability of a genealogy given the parameters of the demographic model (N, or $\{N_i, m_{ij}\}$ if structured populations)

example for a unique panmictic population

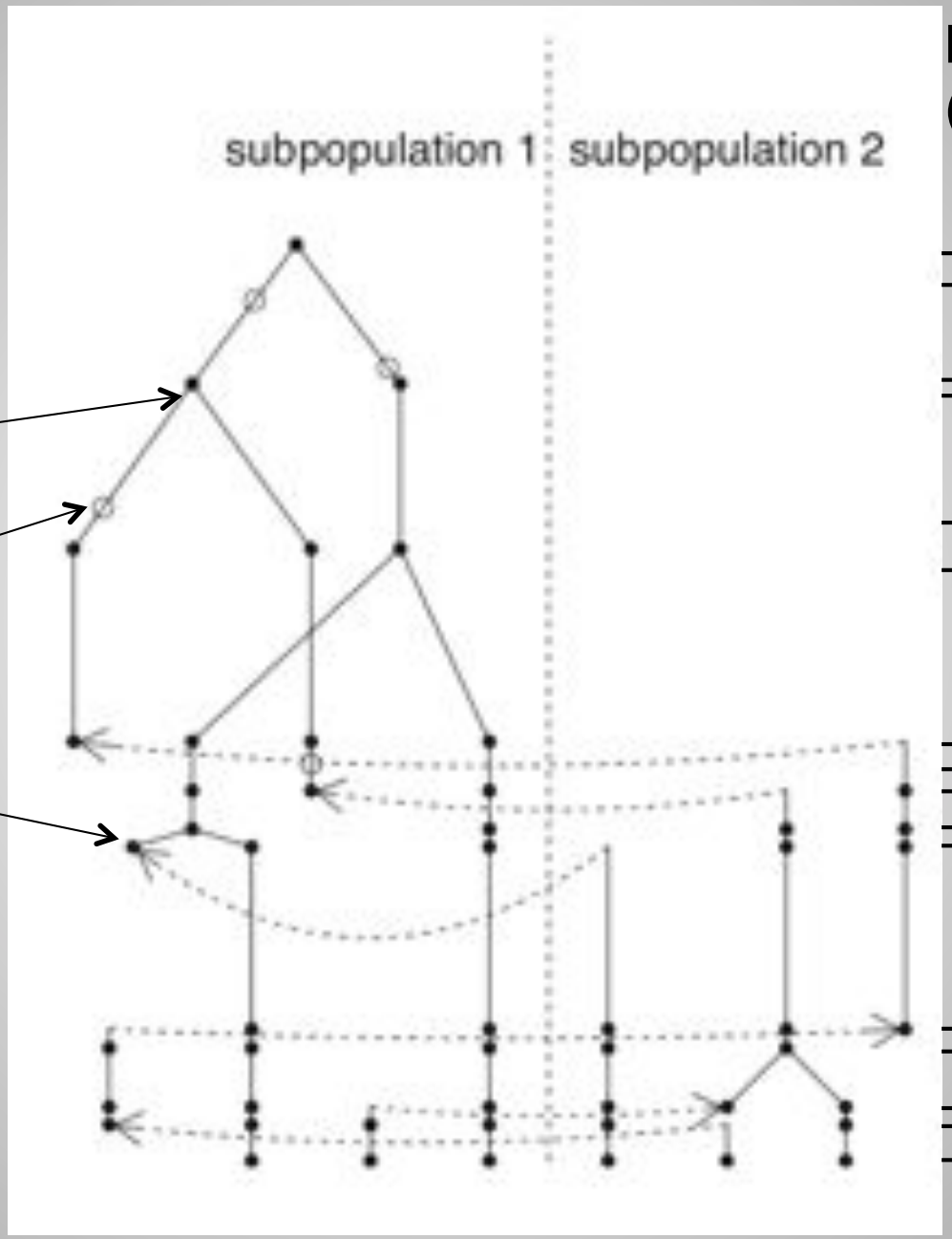
$$\Pr(G|P) = \prod_{\tau=1}^{TMRCA} \left(\frac{j_\tau(j_\tau - 1)}{4N} e^{-\frac{j_\tau(j_\tau - 1)}{4N} \mathrm{k}_\tau} \right)$$

Product over all demographic events (coalescence or migration) affecting the genealogy

lineage number before the event

Time interval between this event and the previous one

Time intervals between events (coa, mut, mig)



coa

mut

mig

Demographic inference under the coalescent
the approach of Felsenstein et al. (MCMC)

- Probability of a genealogy given the parameters of the demographic model

$$\Pr(G|P) = \prod_{\tau=1}^{T} \left(\frac{j_\tau (j_\tau - 1)}{4N} \mathrm{e}^{\frac{j_\tau (j_\tau - 1)}{4N} \mathrm{k}_\tau} \right)$$

- Probability of the sample given the genealogy and mutational parameters (mutation rate μ, M_{mut} mutation matrix)

$$\Pr(D|G) = \prod_{b=1}^{B} \left((M_{mut})^{i_b} \frac{(\mu L_b)^{i_\mathrm{b}}}{i_\mathrm{b}!} \mathrm{e}^{\mu L_b} \right)$$

Product over all tree branches

mutation number on branch *b*

Poisson probability of getting i_b mutations on a time interval L_b

length of branch *b*

Demographic inference under the coalescent
the approach of Felsenstein et al. (MCMC)

- Probability of a genealogy given the parameters of the demographic model

$$\Pr(G|P) = \prod_{\tau=1}^{T} \left(\frac{j_\tau (j_\tau - 1)}{4N} \mathrm{e}^{\frac{j_\tau (j_\tau - 1)}{4N} \mathrm{k}_\tau} \right)$$

- Probability of the sample given the genealogy and mutational parameters

$$\Pr(D|G) = \prod_{b=1}^{B} \left((P_{mut})^{i_b} \frac{(\mu L_b)^{i_b}}{i_b!} \mathrm{e}^{\mu L_b} \right)$$

- by definition

$$L(P|D) \approx \frac{1}{K} \sum_{k=1}^{K} \Pr(D|G_k; P) \approx \frac{1}{K} \sum_{k=1}^{K} \Pr(D|G_k) \Pr(G_k|P)$$

Demographic inference under the coalescent
the approach of Felsenstein et al. (MCMC)

It is a very complexe problem because of the large genealogical and parameter spaces to explore

more parameters ➠ more complexe genealogies

Models with more parameters will need more computation times or more efficient algorithms to explore the 2 spaces

➔ better to always try to consider simple but robust models

Metropolis-Hastings algorithm for the parameter space

(1) start from a point (vector of parameter values, Θ)

(2) propose a change in the parameter space
Θ' from the proposal distribution $q(\Theta \rightarrow \Theta')$

(3) accept the change with probability

$$h = \min\left(1, \frac{L(\Theta';D)}{L(\Theta;D)} \frac{P(\Theta')}{P(\Theta)} \frac{q(\Theta' \rightarrow \Theta)}{q(\Theta \rightarrow \Theta')}\right)$$

(4) go back to (1)

This algorithm ensure that the parameter space is explored proportionnaly to the likelihood

Metropolis-Hastings algorithm : an efficient exploration of the space

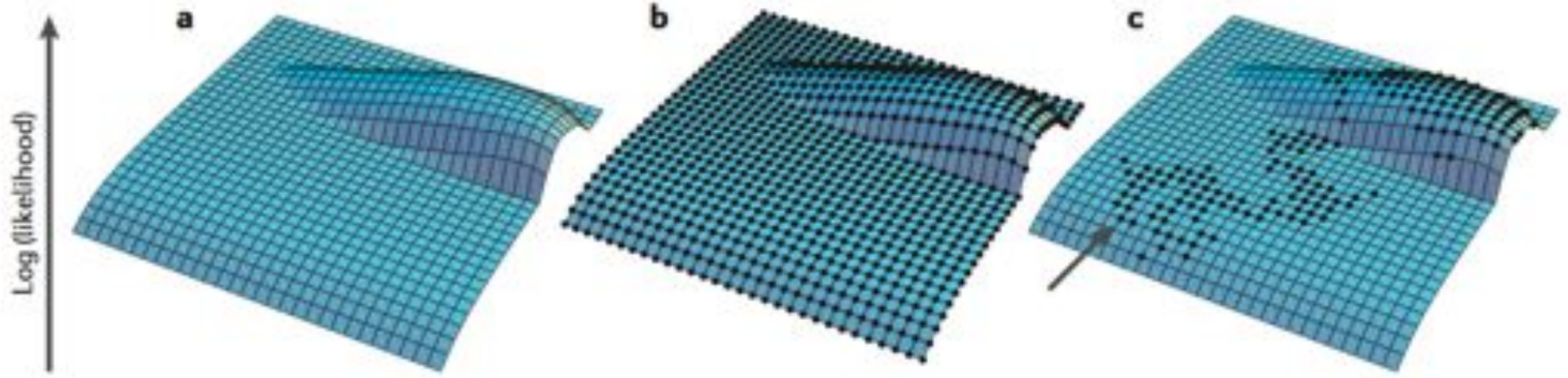
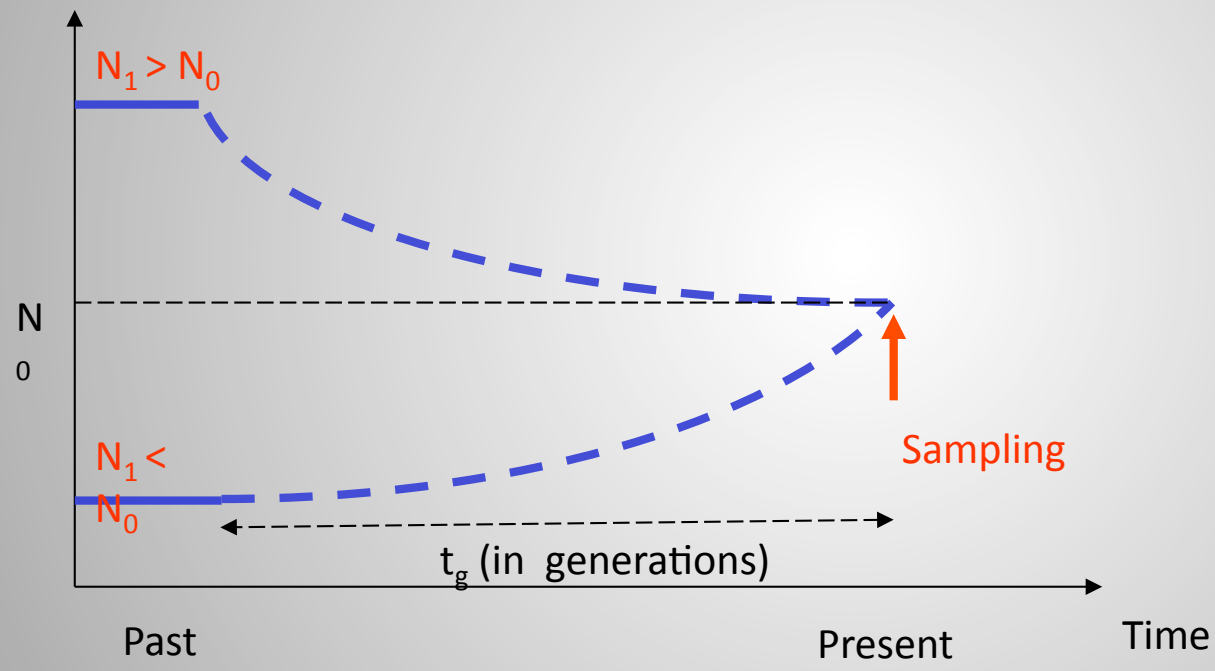


Image courtesy of Peter Beerli, Florida State University, USA.

One example : MsVar (Beaumont 1999)

✓ **Demographic model :** one population with variable size

Population contraction or expansion



3+1 parameters N_0, N_1 et t_g (+ μ) to be estimated using a MCMC Metropolis-Hastings algorithm

One example : MsVar (Beaumont 1999)

- **Monte Carlo Markov chains simulation using the Metropolis-Hastings algorithm (MCMC)**
  - ✓ To explore the genealogy space
  - ✓ and the parameter space

One example : MsVar (Beaumont 1999)

- **Monte Carlo Markov chains simulation (MCMC)**

  ✓ To explore the genealogies, we then build a new genealogy by a "partial deletion-reconstruction" algorithm from the current one :

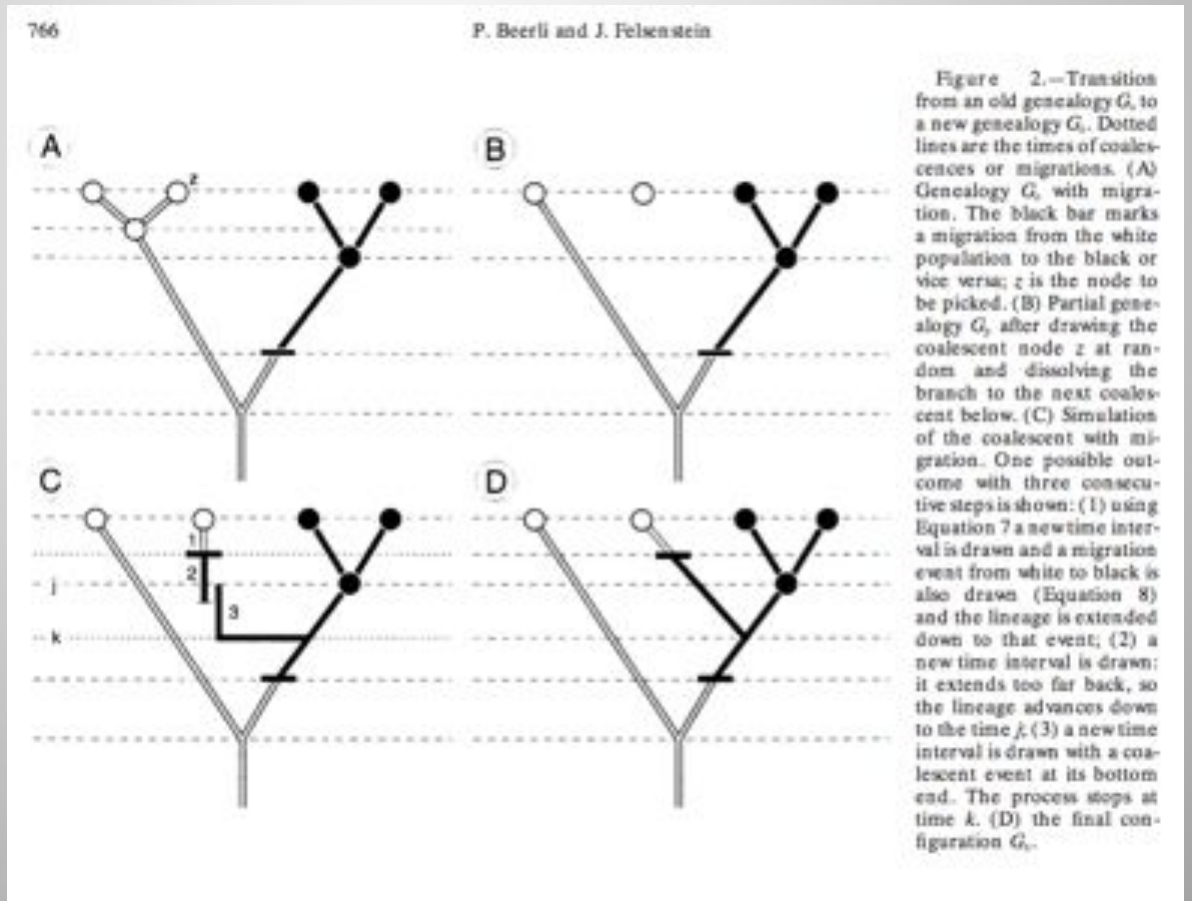
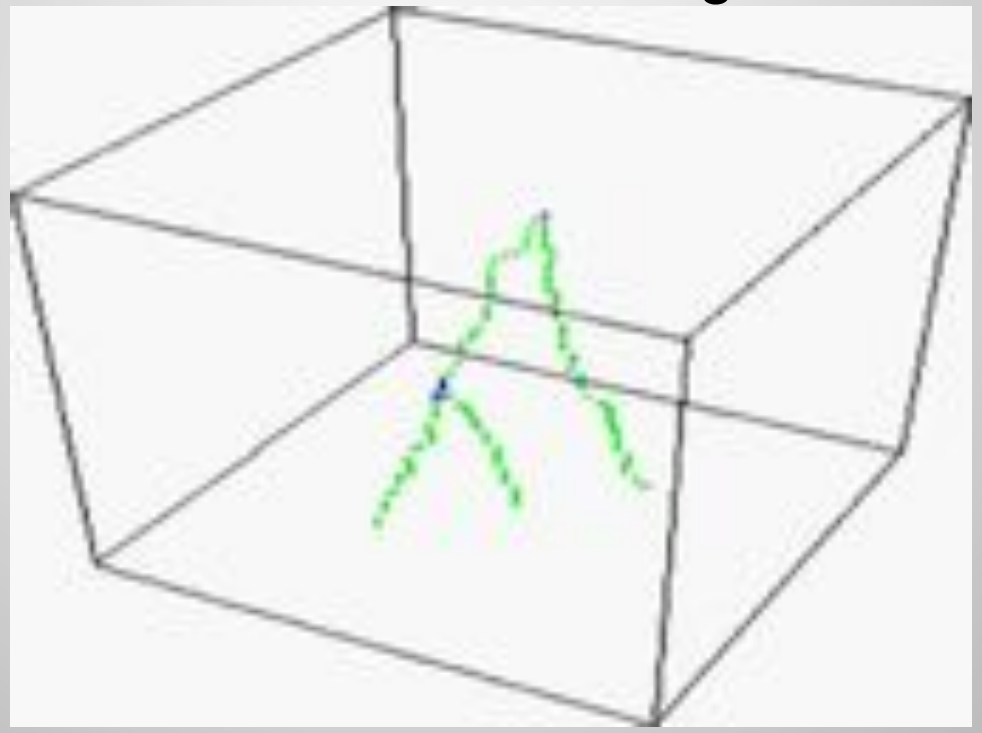


Figure 2.—Transition from an old genealogy G_i to a new genealogy G_{i-1}. Dotted lines are the times of coalescences or migrations. (A) Genealogy G_i with migration. The black bar marks a migration from the white population to the black or vice versa; z is the node to be picked. (B) Partial genealogy G_j after drawing the coalescent node z at random and dissolving the branch to the next coalescent below. (C) Simulation of the coalescent with migration. One possible outcome with three consecutive steps is shown: (1) using Equation 7 a new time interval is drawn and a migration event from white to black is also drawn (Equation 8) and the lineage is extended down to that event; (2) a new time interval is drawn: it extends too far back, so the lineage advances down to the time j; (3) a new time interval is drawn with a coalescent event at its bottom end. The process stops at time k. (D) the final configuration G_{i-1}.

One example : MsVar (Beaumont 1999)

- **Monte Carlo Markov chains simulation (MCMC)**

  ✓ To explore the genealogies, we then build a new genealogy by a "partial deletion-reconstruction" algorithm from the current one :



potential problem : Trees are correlated…

One example : MsVar (Beaumont 1999)

- **Monte Carlo Markov chains simulation (MCMC)**

  ✓ To explore the genealogies, a new genealogy is build by a "partial deletion-reconstruction" algorithm from the current genealogy

  ✓ in parallel, the parameter space will be explored by modifying parameter values in the MCMC

  ➔ at each step of the MCMC:

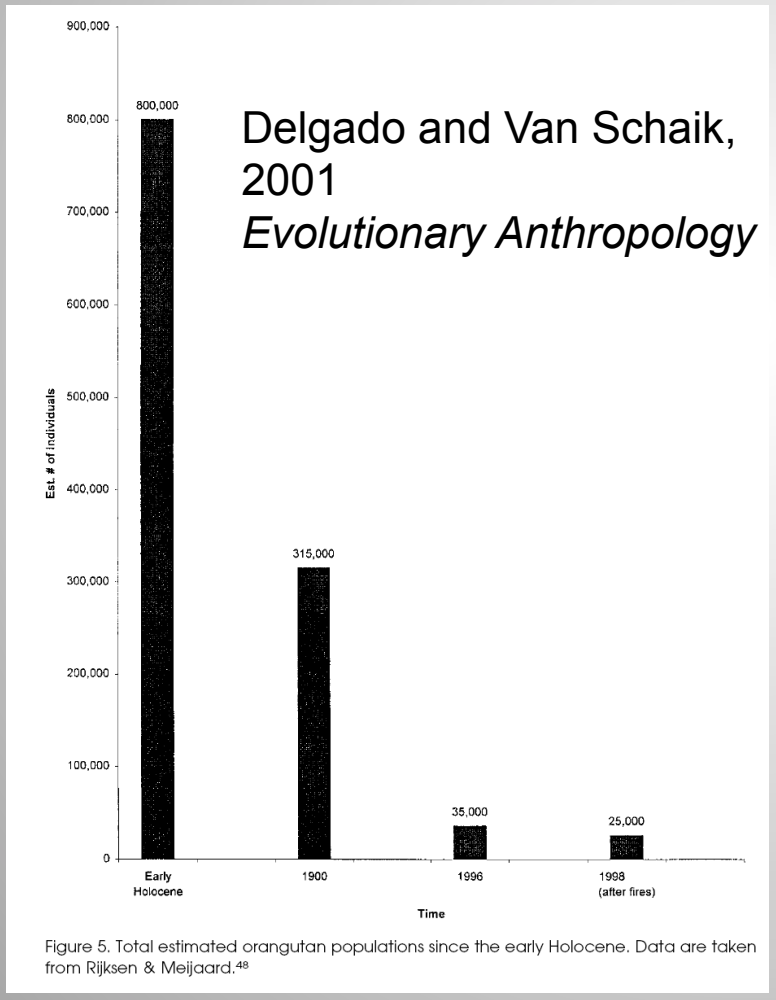
  either the genealogy is modified,

  or a parameter value is modified

Results presented by Renaud Vitalis

Un exemple d'application de la méthode MsVar

- Les Orangs-Outans et la déforestation



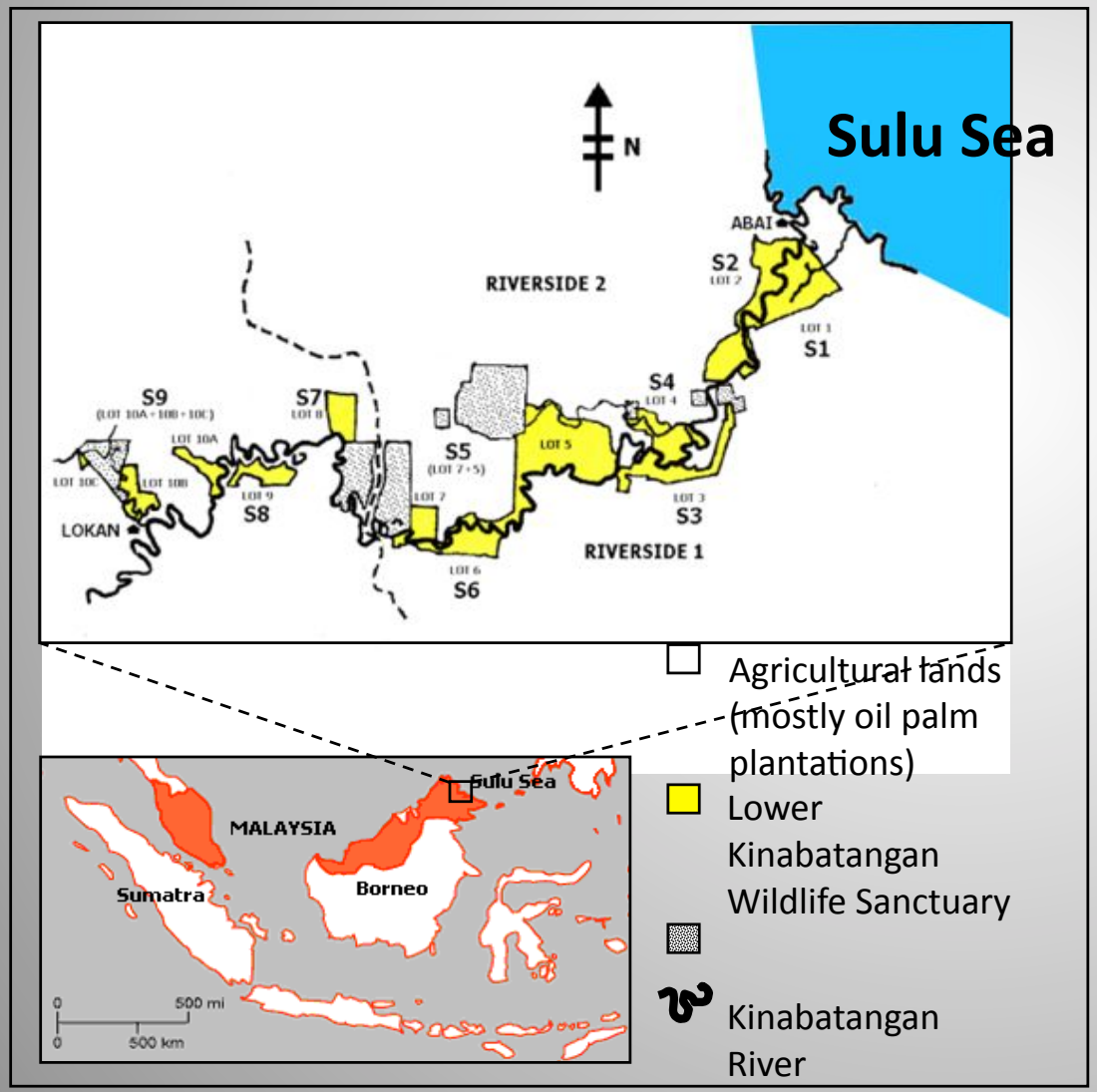
Delgado and Van Schaik, 2001
Evolutionary Anthropology

Figure 5. Total estimated orangutan populations since the early Holocene. Data are taken from Rijksen & Meijaard.[48]

Quelle est la cause de la baisse de taille de population?
La génétique peut elle nous aider?

Un exemple d'application de la méthode MsVar

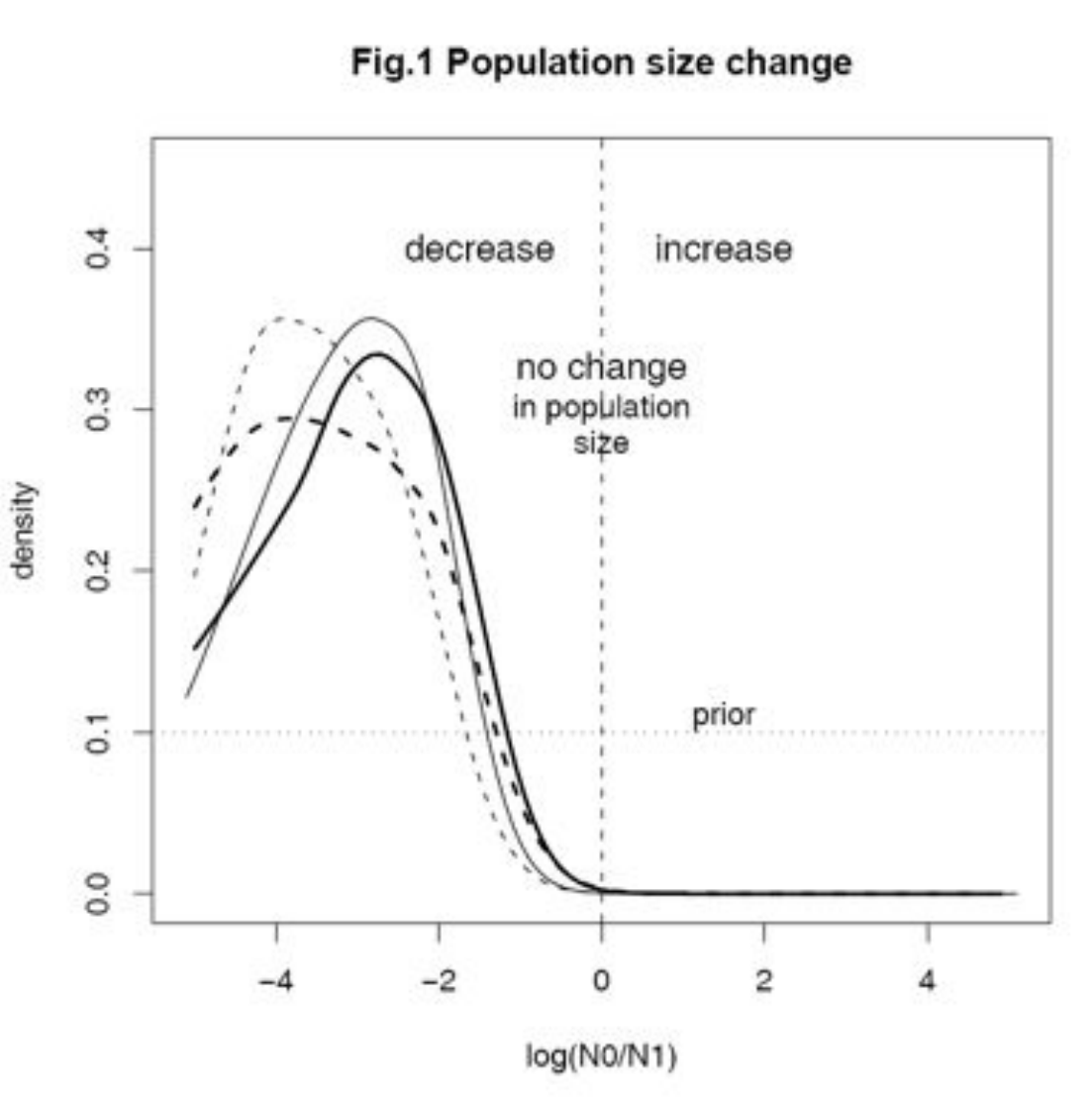
- Les Orangs-Outans et la déforestation : les données



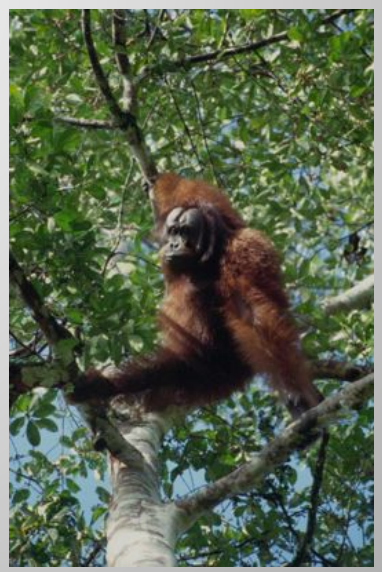
200 individus
14 locus microsatellites

Un exemple d'application de la méthode MsVar

- Les Orangs-Outans et la déforestation :



MsVar détecte bien un réduction de taille de population



Un exemple d'application de la méthode MsVar

- Les Orangs-Outans et la déforestation :

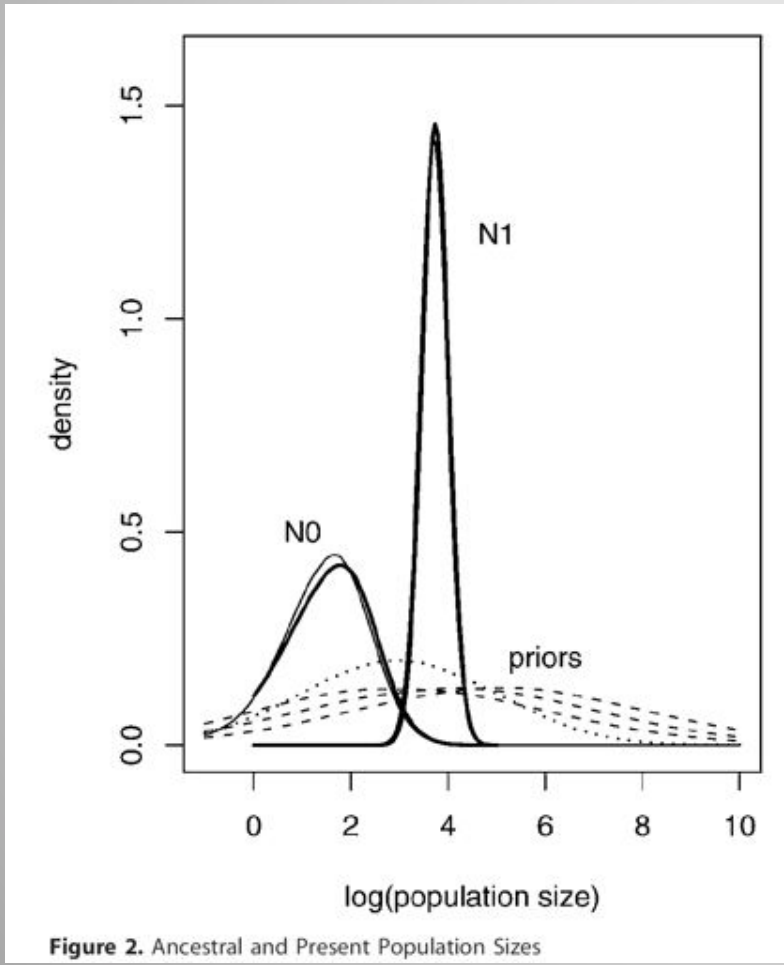
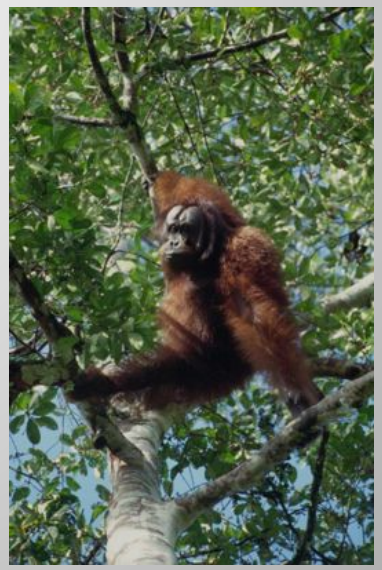


Figure 2. Ancestral and Present Population Sizes

MsVar détecte bien un réduction de taille de population



Un exemple d'application de la méthode MsVar

- Les Orangs-Outans et la déforestation :

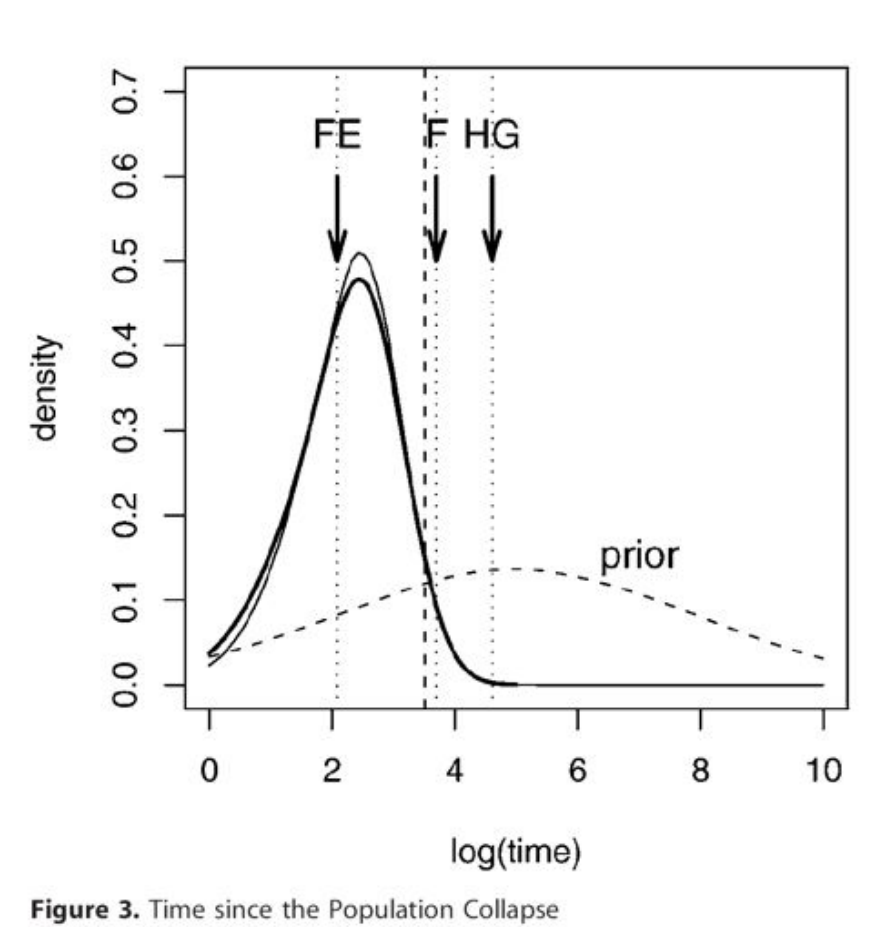


Figure 3. Time since the Population Collapse

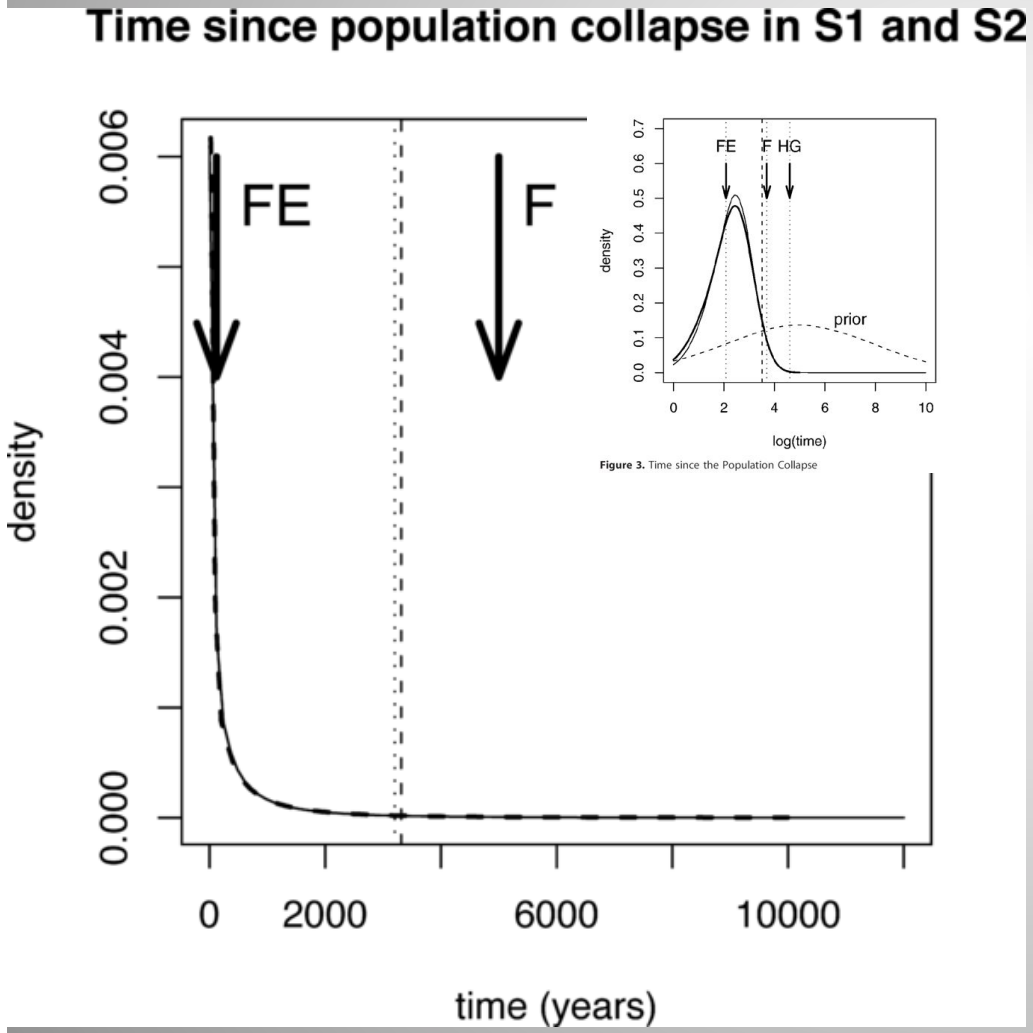
MsVar détecte bien un réduction de taille de population

et permet d'obtenir une datation

FE : Forest exploitation
F: Farmers
HG: Hunter-gatherers

Un exemple d'application de la méthode MsVar

- Les Orangs-Outans et la déforestation :



MsVar détecte bien un réduction de taille de population

et permet d'obtenir une datation: l'exploitation de la forêt semble être la cause…

FE : Forest exploitation
F: Farmers
HG: Hunter-gatherers

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- Probability of a sample *D* given mutational and demographic parameters of the model considered can be computed using the probabilities of transition between the different events affecting the genealogy (with mutations), i.e. the different ancestral states H_k .

a genealogy = genealogical history of the sample can be divided into *m* successive events/states H_k (coalescences, mutations, migrations)

$$G_i=\{H_k;\ 0 > k > -m\} = \{ H_0, H_{-1}, .. , H_{-m+1}, H_{-m},\}$$

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- Probability of a sample *D* given mutational and demographic parameters of the model considered can be computed using the probabilities of transition between the different ancestral states H_k.

$$G_i=\{H_k;\ 0 > k > -m\} = \{ H_0, H_{-1}, \ldots, H_{-m+1}, H_{-m},\}$$

the probability of a given state H_k can be expressed as the probability of all possible ancestral states H_{k-1} multiplied by their associated transition probability possible $\Pr(H_k|H_{k-1})$

$$p(H_k) = \sum_{\{H_{k-1}\}} p(H_k \mid H_{k-1}) p(H_{k-1}).$$

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- the principle of Griffiths et al. importance sampling approach :

  = the recurrence between ancestral samples

$$p(H_k) = \sum_{\{H_{k-1}\}} p(H_k \mid H_{k-1}) p(H_{k-1}).$$

exploring all possible ancestral sample configurations is usually impossible,

➠ Monte Carlo simulations are used to explore a given number *K* of possible genealogies by building genealogies backward in time from the initial sample configuration H_0 to the MRCA (Absorbing Markov chains with absorbing state being the MRCA)

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- the principle of Griffiths et al. importance sampling approach :
  = the recurrence between ancestral samples $p(H_k) = \sum_{\{H_{k-1}\}} p(H_k \mid H_{k-1}) p(H_{k-1}).$
  ➠ Monte Carlo simulations on *Z* possible genealogies build backward in time from $D=H_0$ to the MRCA

$$p(D = H_0) \approx E_Z[p(H_{z,0}|H_{z,-1})\,p(H_{z,-1}|H_{z,-2})\ldots p(H_{z,-m+1}|H_{z,-m})\,p(H_{z,-m})]$$

the probability of the data for a given genealogy *z* is the product of all transition probabilities between ancestral states from H_0 to H_{-m} = the MRCA

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- the principle of Griffiths et al. importance sampling approach :

  = the recurrence between ancestral samples $p(H_k) = \sum_{\{H_{k-1}\}} p(H_k \mid H_{k-1}) p(H_{k-1}).$

  ➠ Monte Carlo simulations

$$p(D = H_0) \approx E_Z[p(H_{z,0}|H_{z,-1})\,p(H_{z,-1}|H_{z,-2})\ldots p(H_{z,-m+1}|H_{z,-m})\,p(H_{z,-m})]$$

This is the approach of Griffiths & Tavaré 1984, implemented in GeneTree for DNA sequence

Genealogies / coalescent trees are explored according to the forward transition probabilities Pr(H_k|H_{k-1})

The approach is working but relatively inefficiently because it uses forward transition probabilities to build genealogies backward

➠ too many tree simulation needed

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- the new importance sampling of Delorio & Griffiths 2004 :

  = the recurrence between ancestral samples $p(H_k) = \sum_{\{H_{k-1}\}} p(H_k \mid H_{k-1}) p(H_{k-1}).$

  ➠ it is much better to try to simulate from the backward transition probabilities Pr(H_{k-1}|H_k). Those probabilities are unknown but they may be approximated

$$p(H_k) = \sum_{\{H_{k-1}\}} \frac{p(H_k \mid H_{k-1})}{\hat{p}(H_{k-1} \mid H_k)} p(H_{k-1}) \hat{p}(H_{k-1} \mid H_k)$$

$$= \mathbb{E}_{\hat{p}} \sum_{\{H_{k-1}\}} \left[\frac{p(H_k \mid H_{k-1})}{\hat{p}(H_{k-1} \mid H_k)} \mid H_k \right].$$

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- the new importance sampling of Delorio & Griffiths 2004 :

  = the recurrence between ancestral samples $p(H_k) = \sum_{\{H_{k-1}\}} p(H_k \mid H_{k-1}) p(H_{k-1}).$

  ➠ it is much better to try to simulate from the backward transition probabilities Pr($H_{k-1}|H_k$). Those probabilities are unknown but they may be approximated

importance sampling weights
= correction for simulating according to $\hat{p}(H_{k-1} \mid H_k)$

$$p(H_k) = \sum_{\{H_{k-1}\}} \frac{p(H_k \mid H_{k-1})}{\hat{p}(H_{k-1} \mid H_k)} p(H_{k-1}) \hat{p}(H_{k-1} \mid H_k)$$

$$= \mathbb{E}_{\hat{p}} \sum_{\{H_{k-1}\}} \left[\frac{p(H_k \mid H_{k-1})}{\hat{p}(H_{k-1} \mid H_k)} \mid H_k \right].$$

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- the new importance sampling of Delorio & Griffiths 2004 :

  = the recurrence between ancestral samples $p(H_k) = \sum_{\{H_{k-1}\}} p(H_k \mid H_{k-1}) p(H_{k-1}).$

  ➠ it is much better to simulate genealogies from approximated backward transition probabilities $\hat{p}(H_{k-1} \mid H_k)$

$$p(H_0) = \mathbb{E}_{\hat{p}} \left[\frac{p(H_0 \mid H_{-1})}{\hat{p}(H_{-1} \mid H_0)} \cdots \frac{p(H_{-m+1} \mid H_{-m})}{\hat{p}(H_{-m} \mid H_{-m+1})} \times p(H_{-m}) \right],$$

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- the new importance sampling of Delorio & Griffiths 2004 :

  = the recurrence between ancestral samples $p(H_k) = \sum_{\{H_{k-1}\}} p(H_k \mid H_{k-1}) p(H_{k-1}).$

  ➠ it is much better to simulate genealogies from approximated backward transition probabilities $\hat{p}(H_{k-1} \mid H_k)$

$$p(H_0) = \mathbb{E}_{\hat{p}}\left[\frac{p(H_0 \mid H_{-1})}{\hat{p}(H_{-1} \mid H_0)} \cdots \frac{p(H_{-m+1} \mid H_{-m})}{\hat{p}(H_{-m} \mid H_{-m+1})} \times p(H_{-m})\right],$$

the probability of the data for a given genealogy *z* is the product of all transition importance weights $w_{IS}(H_k, H_{k-1})$ between ancestral states from H_0 to H_{-m} = the MRCA

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- the new importance sampling of Delorio & Griffiths 2004 :

  = the recurrence between ancestral samples $p(H_k) = \sum_{\{H_{k-1}\}} p(H_k \mid H_{k-1}) p(H_{k-1}).$

  ➠ it is much better to simulate genealogies from approximated backward transition probabilities $\hat{p}(H_{k-1} \mid H_k)$

$$p(H_0) = \mathbb{E}_{\hat{p}} \left[\frac{p(H_0 \mid H_{-1})}{\hat{p}(H_{-1} \mid H_0)} \cdots \frac{p(H_{-m+1} \mid H_{-m})}{\hat{p}(H_{-m} \mid H_{-m+1})} \times p(H_{-m}) \right],$$

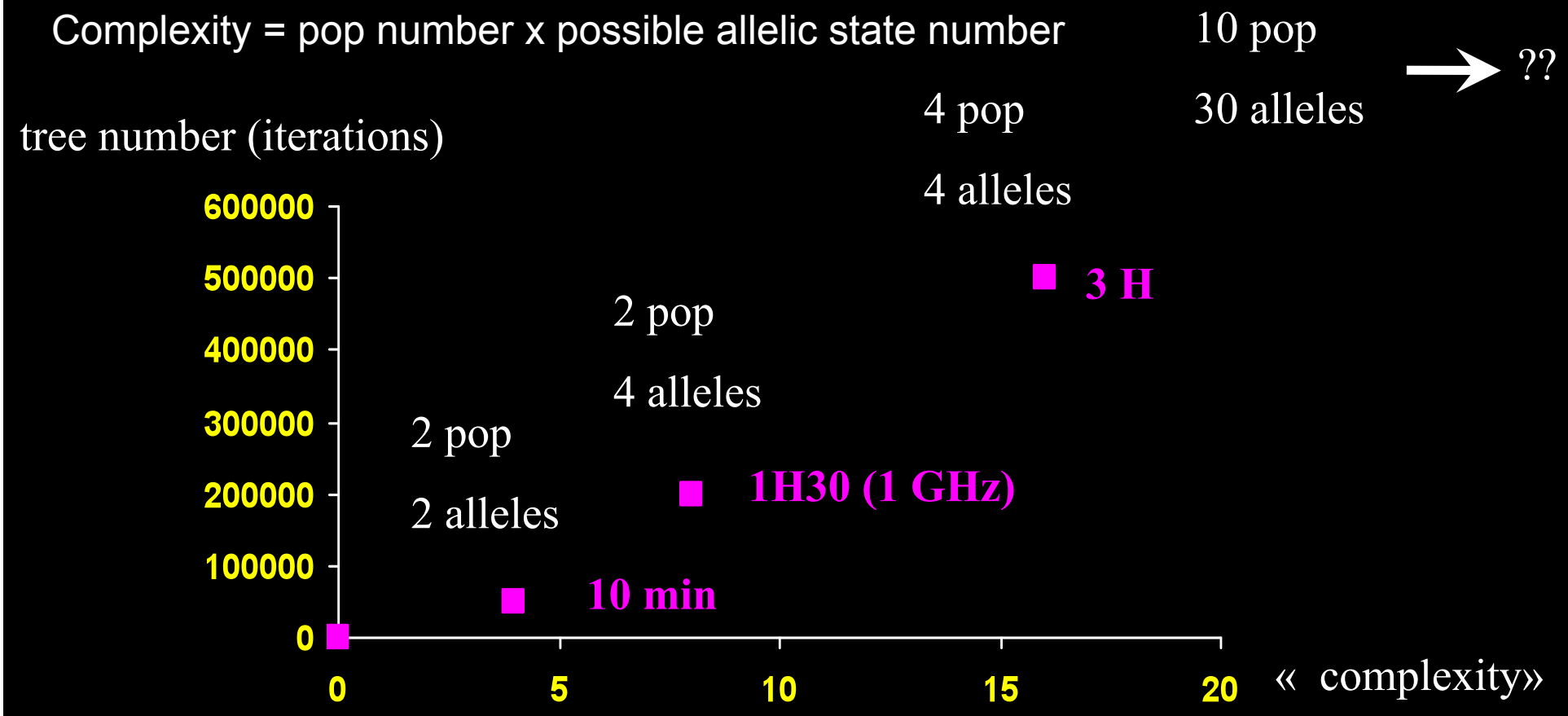
The approach of Delorio & Griffiths 2004 is much more efficient,

1,000 times less trees to explore!

Computation time and model complexity with Griffiths & Tavaré (1984) algorithm

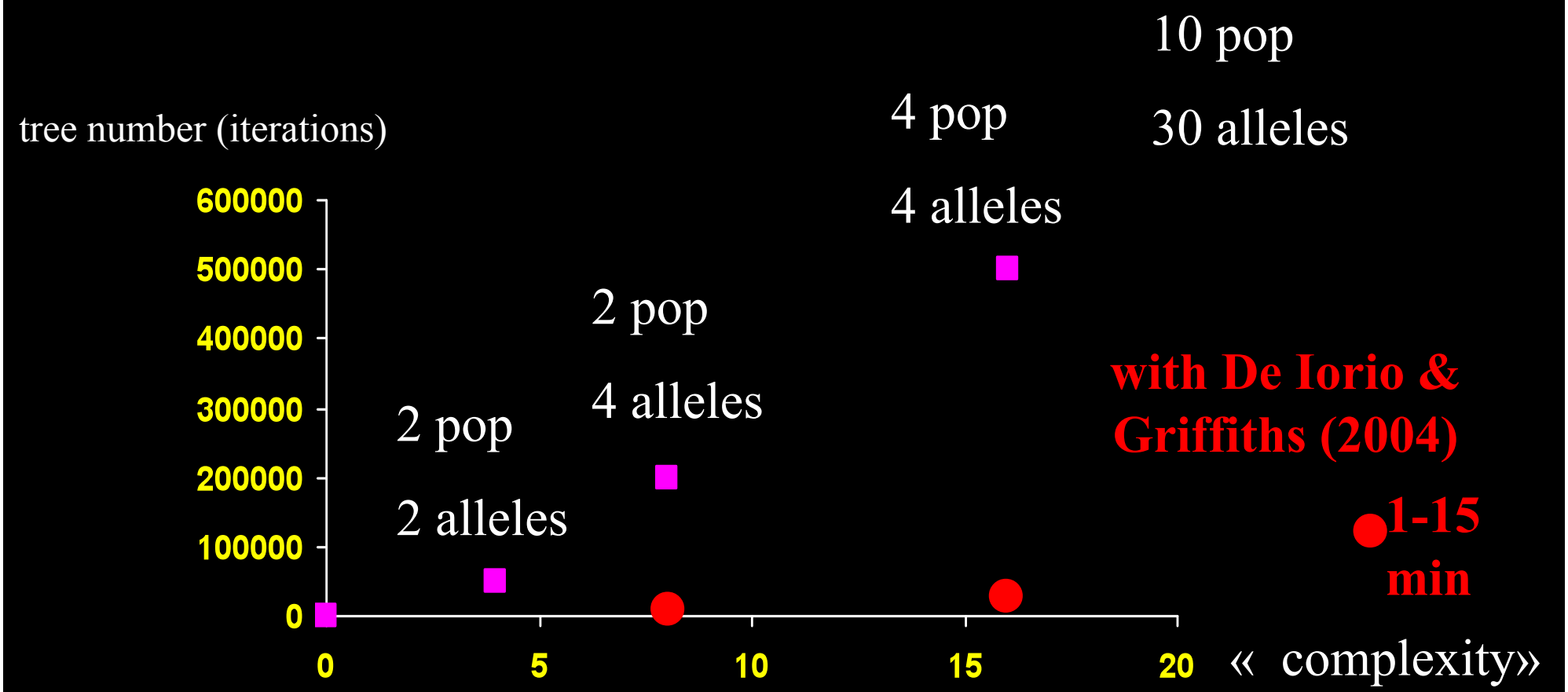
Number of genealogies (= iterations) and time to correctly infer the likelihood of a sample at a single parameter point (one vector Θ of parameter values)

Complexity = pop number x possible allelic state number



Too slow to be practically used or inferences

Computation time and model complexity with DeIorio & Griffiths (2004) algorithm



much more efficient, practically usable for inferences

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- the new importance sampling of Delorio & Griffiths 2004 :

possible genealogies (= coalescent trees) with mutations are build backward in time event by event (i.e. H_k, each time the sample configuration changes) until the MRCA is found.

Those coalescent tree simulations (absorbing Markov chains) are used to explore the genealogy space

The importance sampling fonction $\hat{p}(H_{k-1} \mid H_k)$ is used to more efficiently explore the genealogical space (i.e. more likely genealogies, with the more likely events)

the parameter space is explored independently

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- the new importance sampling of Delorio & Griffiths 2004 :

$$p(H_k) = \sum_{H_{k-1}} w_{\hat{p}}(H_k, H_{k-1}) \times \hat{p}(H_{k-1} \mid H_k) \times p(H_{k-1})$$

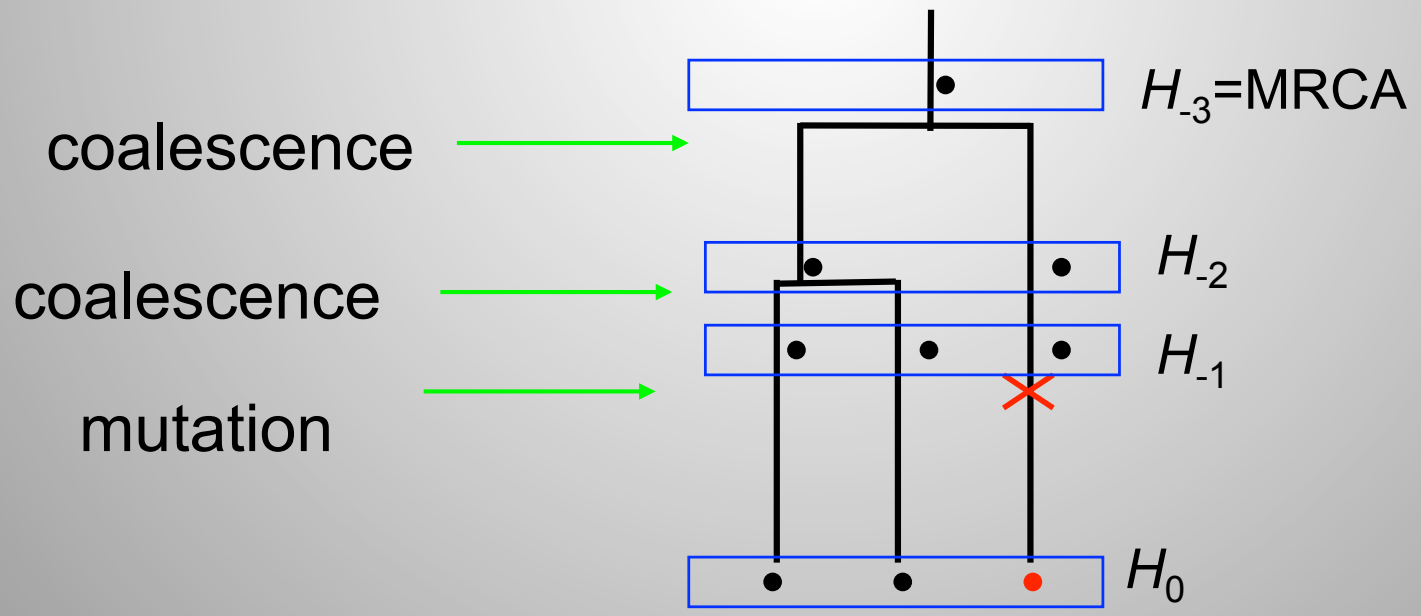
$$p(H_0) = E_{IS}\left[\prod_{k=0}^{k=-m+1=MRCA} w_{IS}(H_k, H_{k-1}) \times p(H_{-m}) \right]$$

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- the recurrence :

$$p(H_k) = \sum_{H_{k-1}} w_{IS}(H_k, H_{k-1}) \bullet \hat{p}(H_{k-1} \mid H_k) \bullet p(H_{k-1})$$

$$p(H_0) = E_{IS}\left[\prod_{k=0}^{k=-m+1=MRCA} w_{IS}(H_k, H_{k-1}) \bullet p(H_{-m}) \right]$$

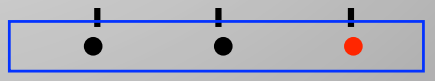


Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- Coalescent tree building

1. Start with the sample configuration H_0
2. Draw randomly an event among all possible events (=**coa** ou **mig** ou **mut**) from the IS transition probabilities

   → new ancestral configuration H_{k-1}
3. compute and store the IS transition weight $w_{IS}(H_{k-1},H_k)$
4. Go back to 2 until the MRCA is found



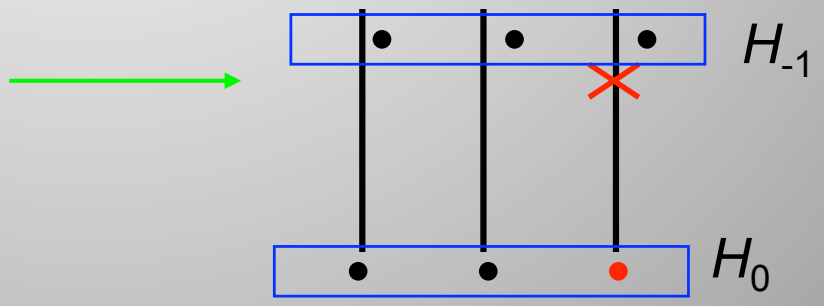
Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- Coalescent tree building

1. Start with the sample configuration H_0
2. Draw randomly an event among all possible events (=**coa** ou **mig** ou **mut**) from the IS transition probabilities

   → new ancestral configuration H_{k-1}
3. compute and store the IS transition weight $w_{IS}(H_{k-1},H_k)$
4. Go back to 2 until the MRCA is found

mutation

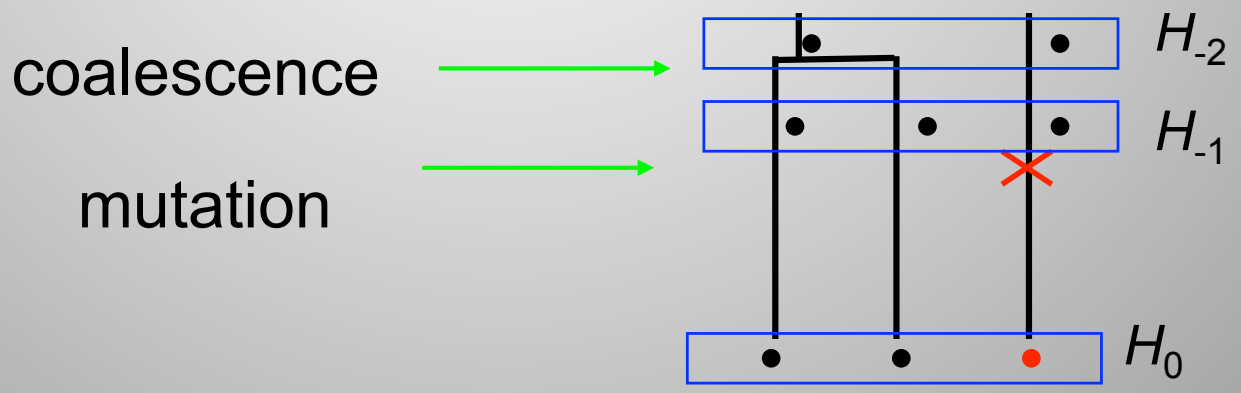


Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- Coalescent tree building

1. Start with the sample configuration H_0
2. Draw randomly an event among all possible events (=**coa** ou **mig** ou **mut**) from the IS transition probabilities

   → new ancestral configuration H_{k-1}
3. compute and store the IS transition weight $w_{IS}(H_{k-1},H_k)$
4. Go back to 2 until the MRCA is found

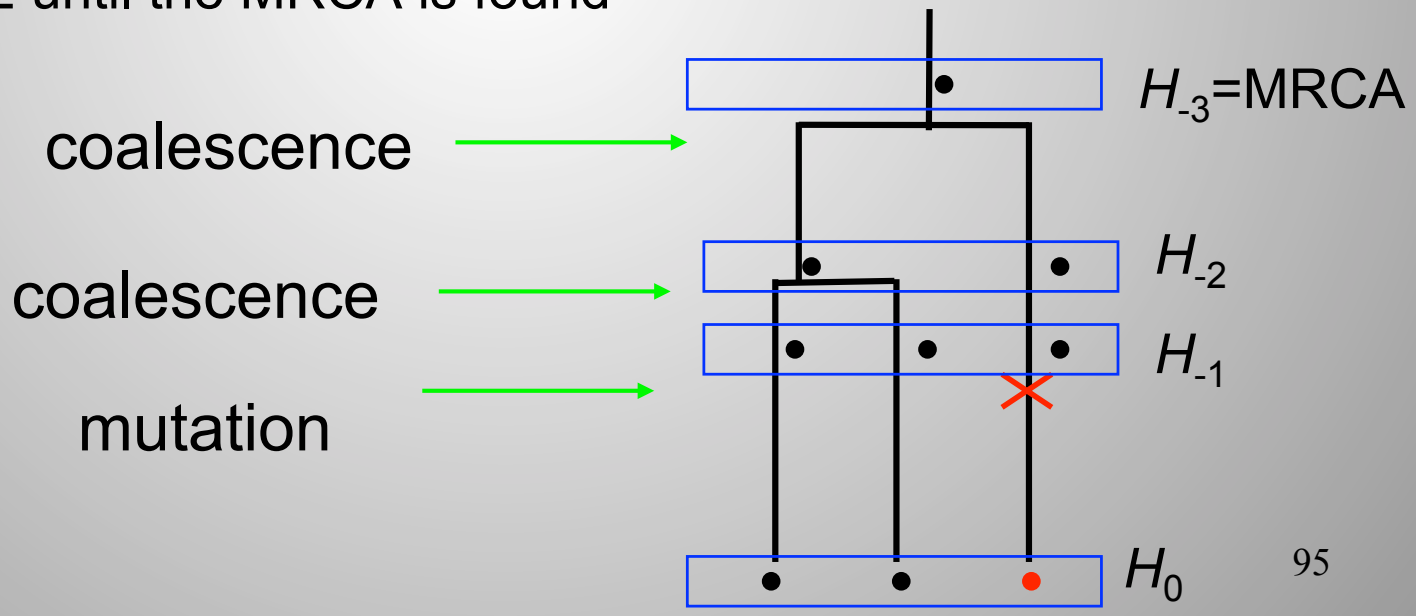


Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- Coalescent tree building

1. Start with the sample configuration H_0
2. Draw randomly an event among all possible events (=**coa** ou **mig** ou **mut**) from the IS transition probabilities

   → new ancestral configuration H_{k-1}
3. compute and store the IS transition weight $w_{IS}(H_{k-1},H_k)$
4. Go back to 2 until the MRCA is found



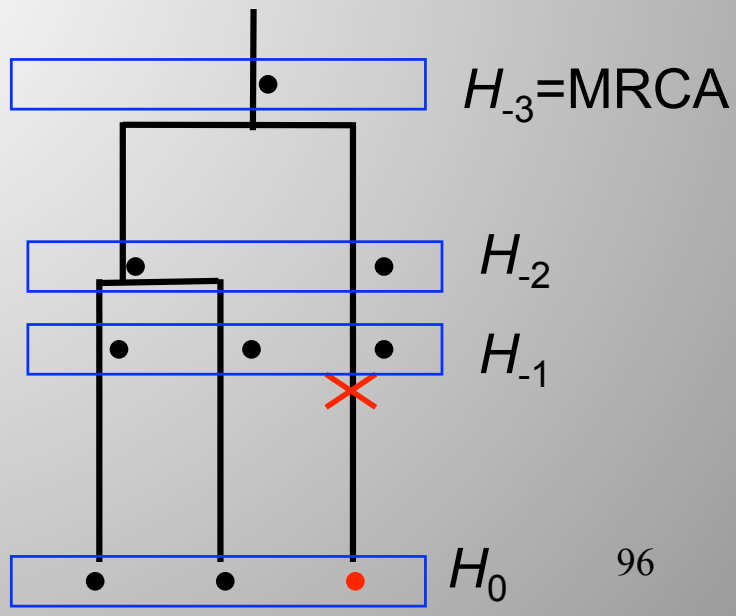
Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- Coalescent tree building

1. Start with the sample configuration H_0
2. Draw randomly an event among all possible events (=**coa** ou **mig** ou **mut**) from the IS transition probabilities
   → new ancestral configuration H_{k-1}
3. compute and store the IS transition weight $w_{IS}(H_{k-1},H_k)$
4. Go back to 2 until the MRCA is found

probabiity of the MRCA = probability of the allelic state of the MRCA in the stationnary distribution of the mutation model

for most model it is equal to 1/K,
with K the number of possible allelic state



Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- Probability of the sample for a given coalescent tree:

All transition weight $w_{IS}(H_{k-1},H_k)$ were computed and stored

$$p(H_0 \mid G_z) = \prod_{k=0}^{k=-m+1=MRCA} w_{IS}(H_k, H_{k-1}) \times p(H_{-m})$$

$$p(H_0 \mid G_z) = \prod_{k=0}^{k=-m+1=MRCA} w_{IS}(H_k, H_{k-1}) \times 1/K$$

Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

- Probability of the sample using Monte Carlo integration over a large number (Z) of coalescent trees:

$$p(H_0 \mid G_z) = \prod_{k=0}^{k=-m+1=MRCA} w_{IS}(H_k, H_{k-1}) \times 1/K$$

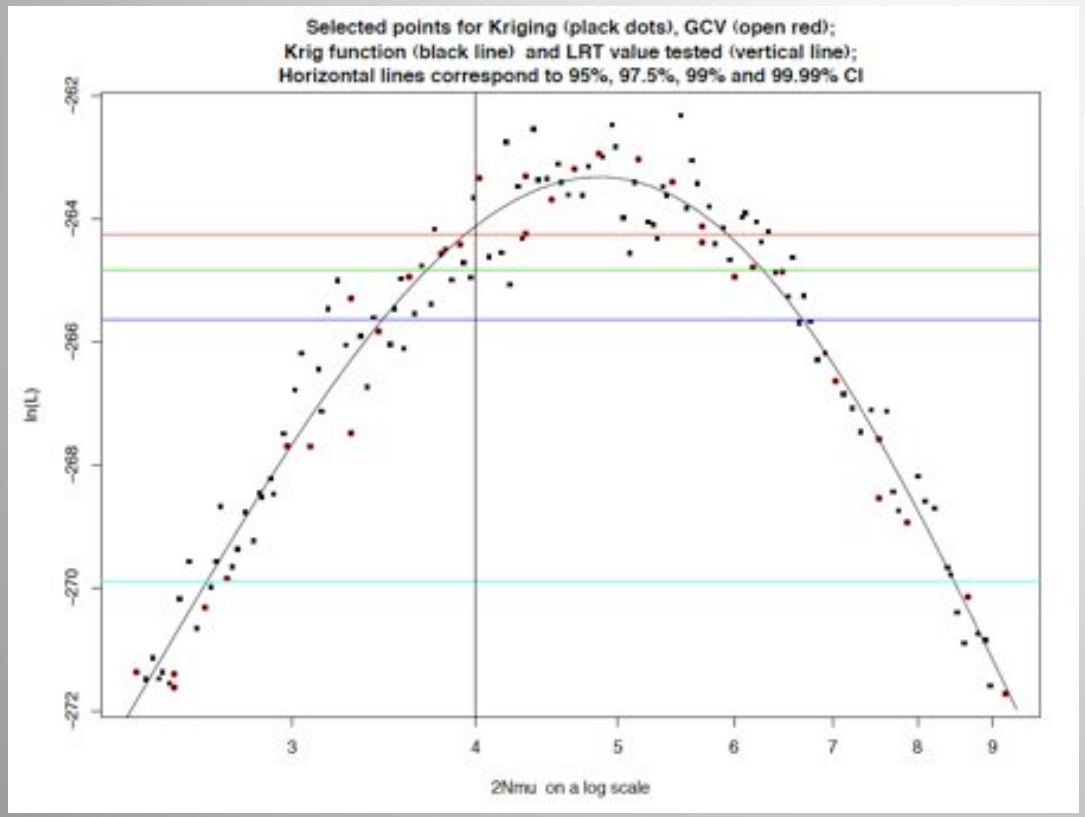
$$p(H_0) = E_{IS}\left[p(H_0 \mid G_z) \right]$$

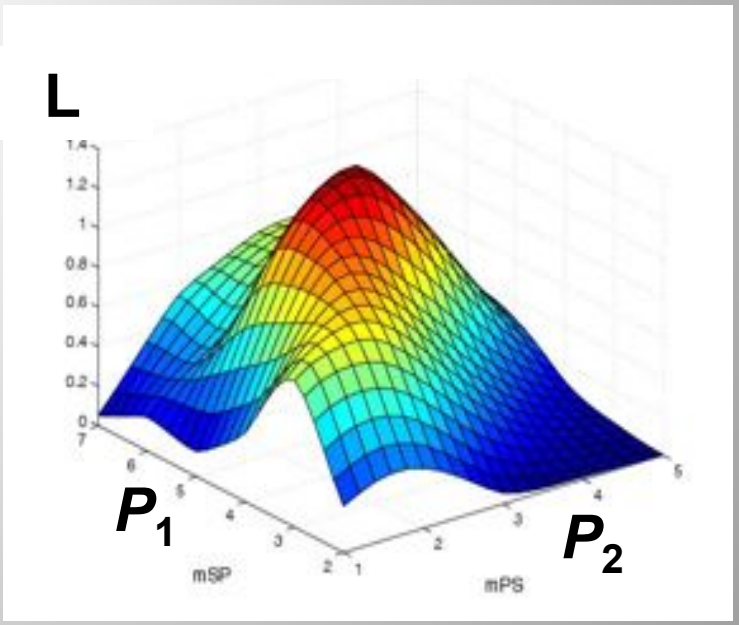
$$p(H_0) \approx \frac{1}{Z} \sum_z p(H_0 \mid G_z)$$

$$L(P|D) = \frac{1}{K} \sum_{k=1}^{K} \Pr(D|G_k; P)$$

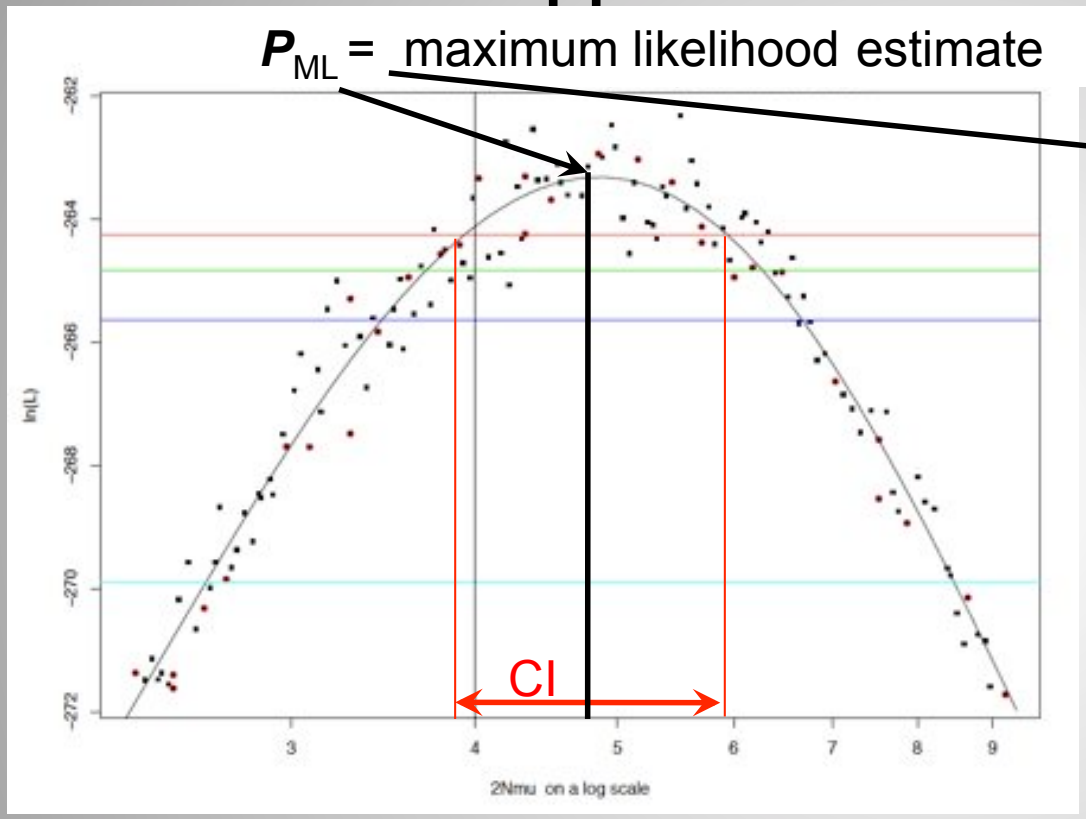
Demographic inference under the coalescent
the approach of Griffiths et al. (IS)

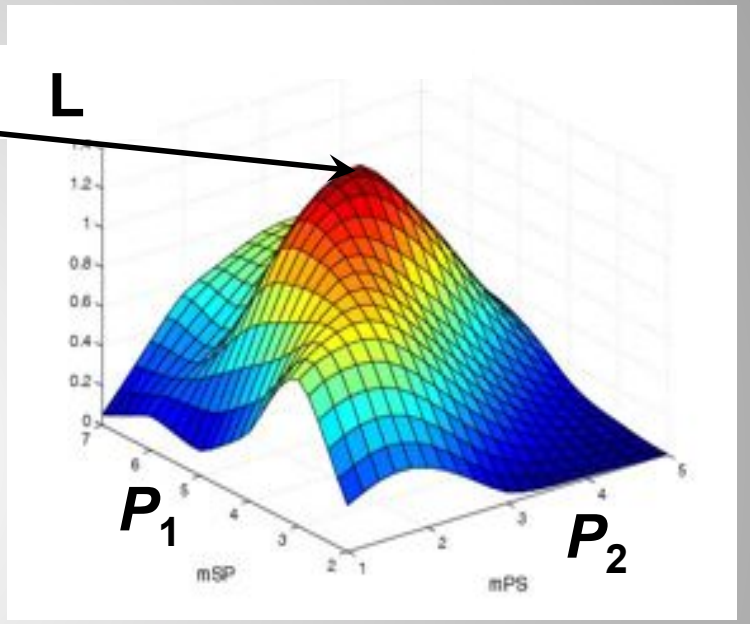
- the likelihood of the sample L($P|D$)=$p(H_0)$ is computed for many points (random or on a grid) over the parameter space and the likelihood surface is interpolated using Kriging





Demographic inference under the coalescent
the approach of Griffiths et al. (IS)





- ML point estimate and Confidence intervals are determined from this interpolated likelihood surface

➠ no convergence required!

IBD and coalescence based maximum likelihood inferences

In theory, Maximum Likelihood methods (ML) should be more powerful than moment based methods (F_{ST}) because :

- Use all the information present in the genetic data
- Powerful maximum likelihood statistical framework
- Possible to make inference on parameters other than $D\sigma^2$
  - ✓ Migration rates (Nm)
  - ✓ Shape of the distribution
  - ✓ Total population size
  - ✓ Mutation rate

IBD and maximum likelihood inference

Two main approaches for maximum likelihood demographic inference under the coalescent :

1. "MCMC" : Felsenstein et al. (e.g. MsVar, IM, LAMARCK, MIGRATE)

   High computation times → difficult to test

3. "IS" : Griffiths et al. (GENETREE, MIGRAINE)

   Much less developed than MCMC, simpler models

   Only recently : 1D and 2D IBD, One pop with variable size ("MsVar like" model)

IBD and maximum likelihood inference

Griffiths et al. (IS, software MIGRAINE)
IBD 1D, recent development for 2D IBD (in prep)

Likelihood and Approximate Likelihood Analyses of Genetic Structure in a Linear Habitat: Performance and Robustness to Model Mis-Specification

François Rousset and Raphaël Leblois†*

*Université, Montpellier 2, CNRS, Institut des Sciences de l'Évolution, France; and †Unité Origine, Structure et Évolution de la Biodiversité, Museum National d'Histoire Naturelle, Paris, France

Mol. Biol. Evol. 24(12):2730–2745. 2007

Demic model of IBD on a circle or on a line with absorbing boundaries

IS much faster than MCMC (10x + easy parallel computing)

Number of parameters reduced by consideration of homogeneous IBD model

IBD and ML inference

1- First results under **stepping stone migration** (i.e. no middle/long distance migrants):

very good precision and robustness on *Nm* inference :

Rel biais =[0.04-0.12] and Rel RMSE=[0.15-0.5]

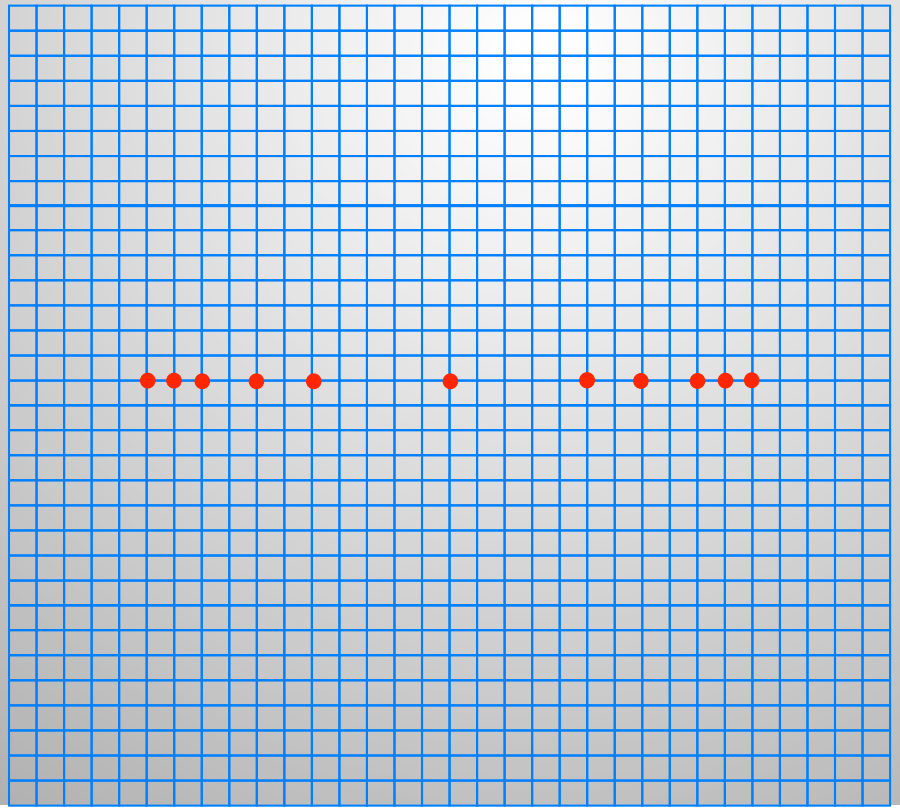
relatively good precision for $N\mu$

Rel biais =[0.04-0.40] and Rel RMSE=[0.25-0.8])

IBD and ML inference

1- First results under **stepping stone migration** (i.e. no middle/long distance migrants):

$N\mu$ slightly influenced by the total number of sub-populations considered in the analysis ("Ghost populations")



IBD and ML inference

2- **geometric dispersal** $\frac{m}{2}(1-g)g^{|k-1|}$ (i.e. with middle/long distance migrants):

Large *g* ➠ more long distance, large $D\sigma^2$ $\sigma^2 = m(1+g)/(1-g)^2.$

$D\sigma^2$ and *Nm* inferences much more precise and robust than for *g*

large *m* and *g* (i.e. more migrants, at larger distances)

➔ more influence of the **ghost/unsampled pops** and of the **mutation process**

Stronger effect for $N\mu$ and *g* than *Nm*,

not much effect for $D\sigma^2$ (compensation of different bias)

IBD and ML inference

2- **geometric dispersal** $\frac{m}{2}(1-g)g^{|k-1|}$ (i.e. with middle/long distance migrants):

Large g ➠ more long distance, large $D\sigma^2$ $\sigma^2 = m(1+g)/(1-g)^2.$

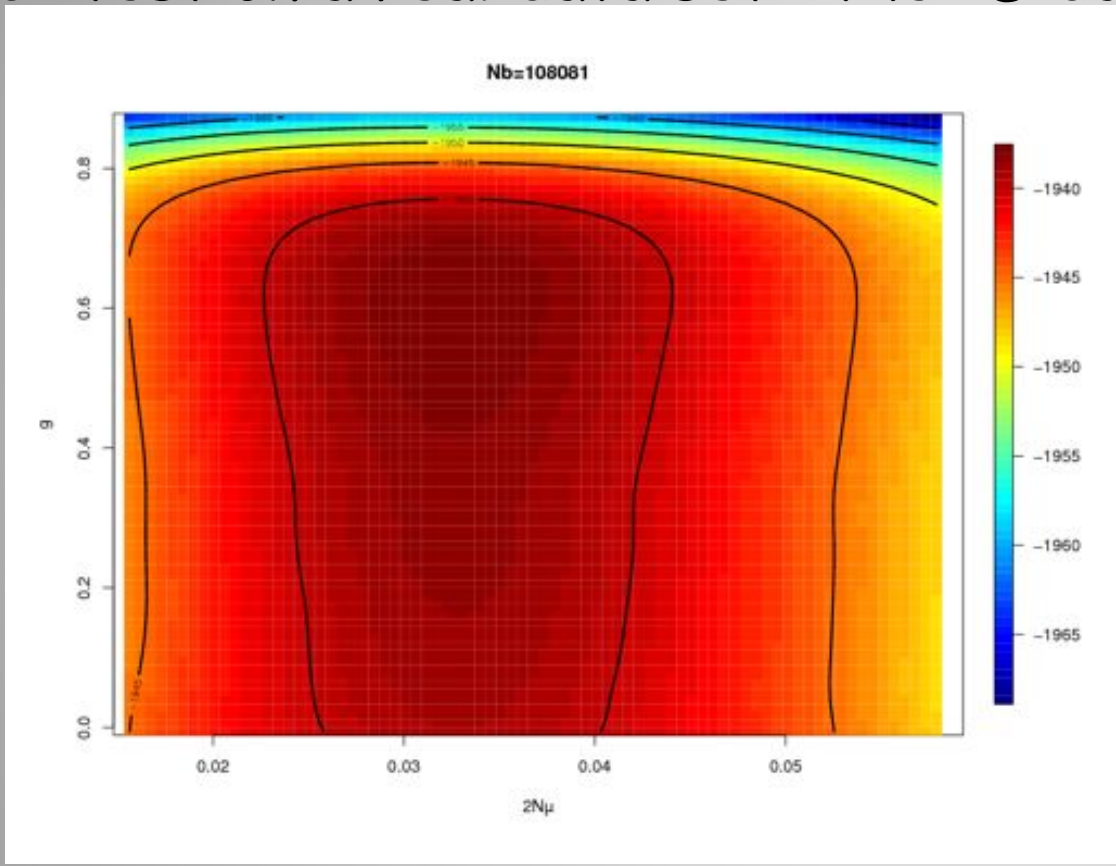
$D\sigma^2$ and Nm inferences much more precise and robust than for g

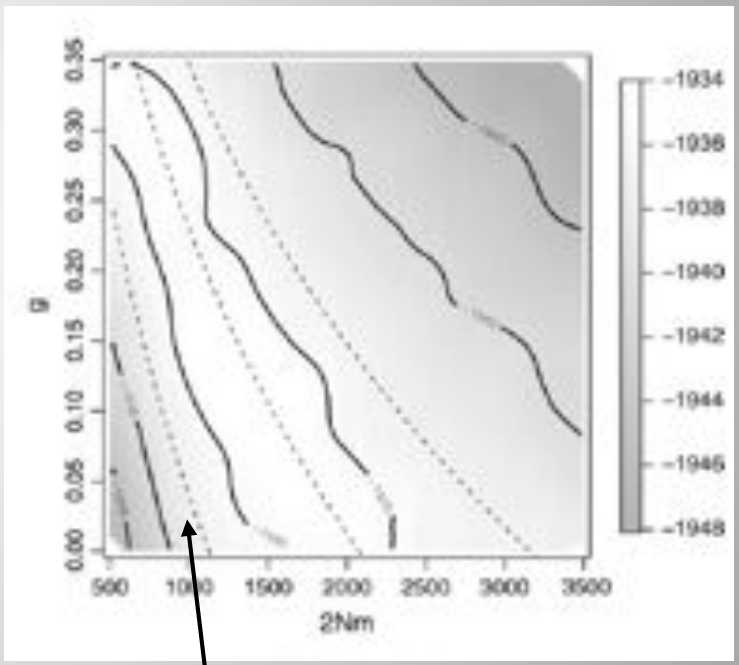
ML more accurate than moment based regression method when analyzed under the good model (i.e. nb of sub-pops and mutation processes well specified)

Hopefully the results are also very accurate for most cases with misspecifications

IBD and ML inference

3- test on a real data set : the 1D damselflies data set



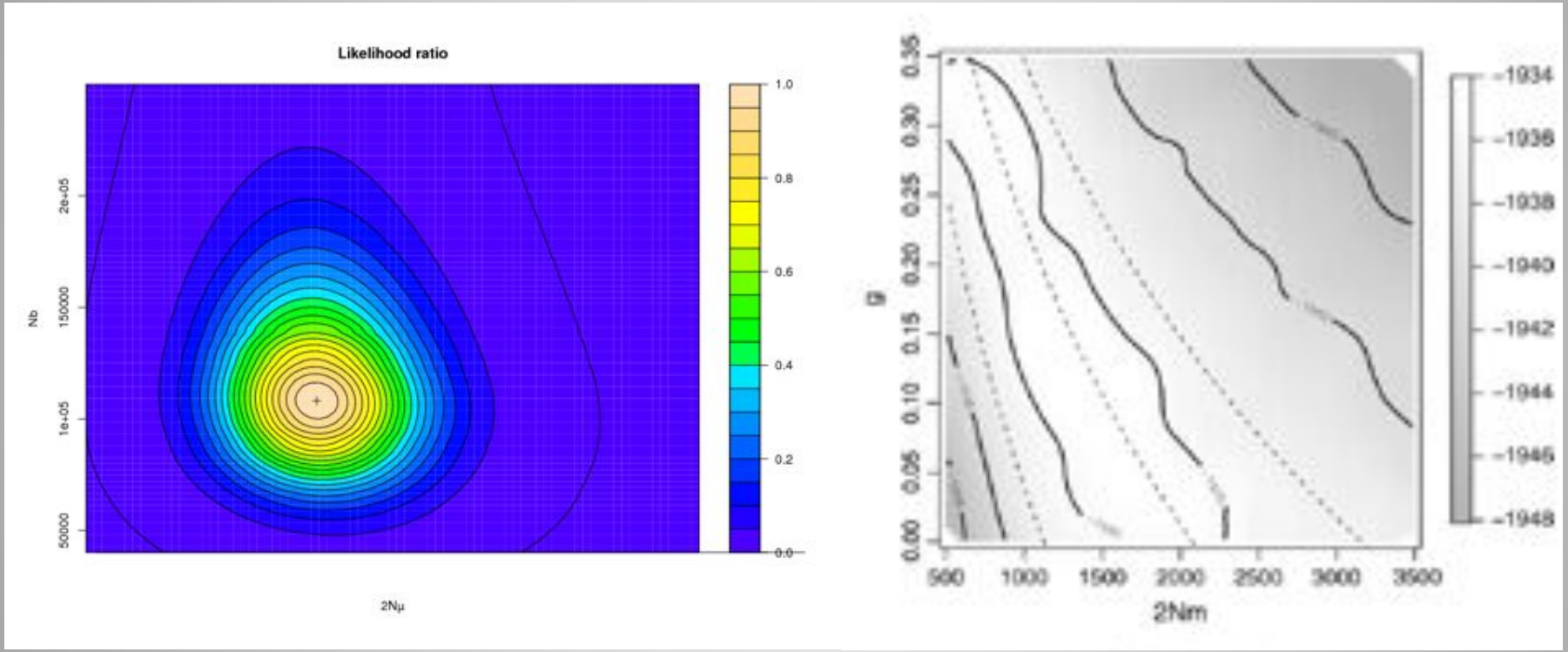


Not much information on *g*, because of a strong correlation with *Nm*

Lines of equal $4D\sigma^2$ values

IBD and IS inference (MIGRAINE)

3- test on a real data set : the 1D damselflies data set



$Nb = 4D\sigma^2$ -> 108 [50-220] $2N\mu$ = 0.04 [0.02-0.07]

More information about *Nb* than *Nm* and *g* separetely

More information about *Nm* than *g*.

6. IBD and IS inference (MIGRAINE)

4 - Comparison with demographic estimates and the moment based regression method on the damselflies example

	Inference of $D\sigma^2$		
	Direct (demo)	Indirect (Regression)	Indirect (MIGRAINE)
Site 1 (1D)	277	222 [66-392]	108 [50-220]

"Effective" demographic estimates are probably overestimated (not corrected for temporal variations in density)

CI obtained by the regression method overlaps widely with the one given by MLE.

6. IBD and IS inference (MIGRAINE)

4 - Comparison with demographic estimates and the moment based regression method on the damselflies example

	Inference of $D\sigma^2$		
	Direct (demo)	Indirect (Regression)	Indirect (MIGRAINE)
Site 1 (1D)	277	222 [66-392]	108 [50-220]

Both genetic methods may estimate (with different small-sample biases) the same effective $D\sigma^2$ and the demographic estimate may be slightly overestimated.

Further comparisons necessary to demonstrate systematic differences of this magnitude.

6. IBD and IS inference (MIGRAINE)

4 - Comparison with demographic estimates and the moment based regression method on the damselflies example

	Inference of $D\sigma^2$		
	Direct (demo)	Indirect (Regression)	Indirect (MIGRAINE)
Site 1 (1D)	277	222 [66-392]	108 [50-220]

Other possible explanations for the observed differences:

- Shape of the dispersal distribution (i.e. not geometric in reality)
- Influence of past demographic processes/fluctuations
- Mutation processes, edge effects, number of sub-populations, binning (but showed only moderate effects on simulations)

ML and IBD : Conclusions

+ Very good performances, even when the model is mis-specified

- Very slow for large network of populations (>100)

- some problems for large migration rates, long distance migration, and small population sizes (due to the coalescent approximations)

➠ impossible to model continuous populations (ABC methods??)

➠ geographic data binning needed to deal with continuous samples

- need to test robustness to past demographic fluctuations

- not much information in "classical" samples on the shape of the dispersal distribution (i.e. inference of param. others than $D\sigma^2$ and Nm)

+ may be used for other developments (e.g. IBD between habitats, landscape genetics)

6. IBD and maximum likelihood inference??

Two main approaches for maximum likelihood demographic inference under the coalescent :

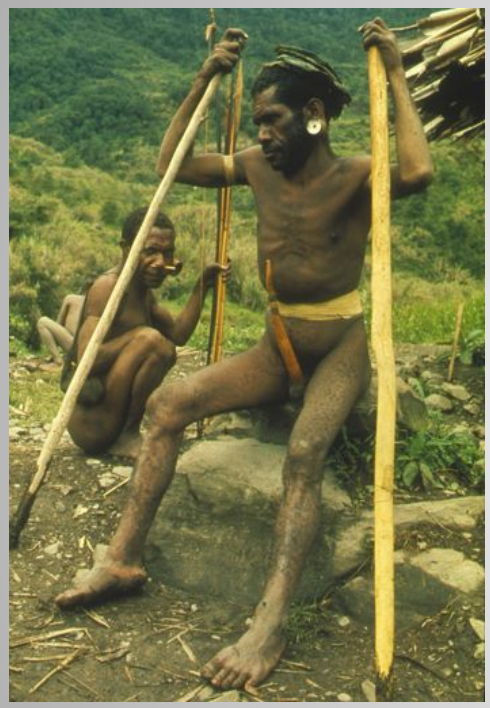
1. "MCMC" : Felsenstein et al. (MIGRATE)
   - High computation times → difficult to test
   - Bad results (simulations and comparison with demography) under IBD but MIGRATE is not especially designed for IBD

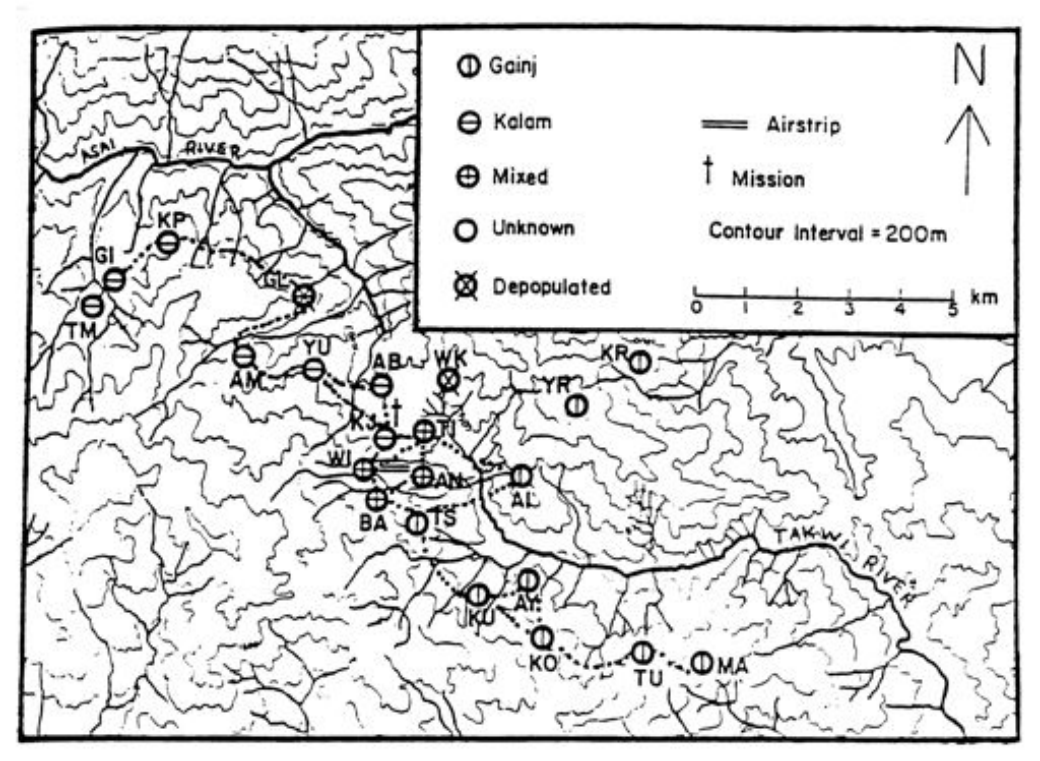
3. "IS" : Griffiths et al. (MIGRAINE)
   - Much less developed than MCMC, simpler models
   - Only recently : Linear IBD (i.e. one dimension)

Preliminary results

First test of MIGRATE : comparison with demographic data

Human data : villages of New Guinea

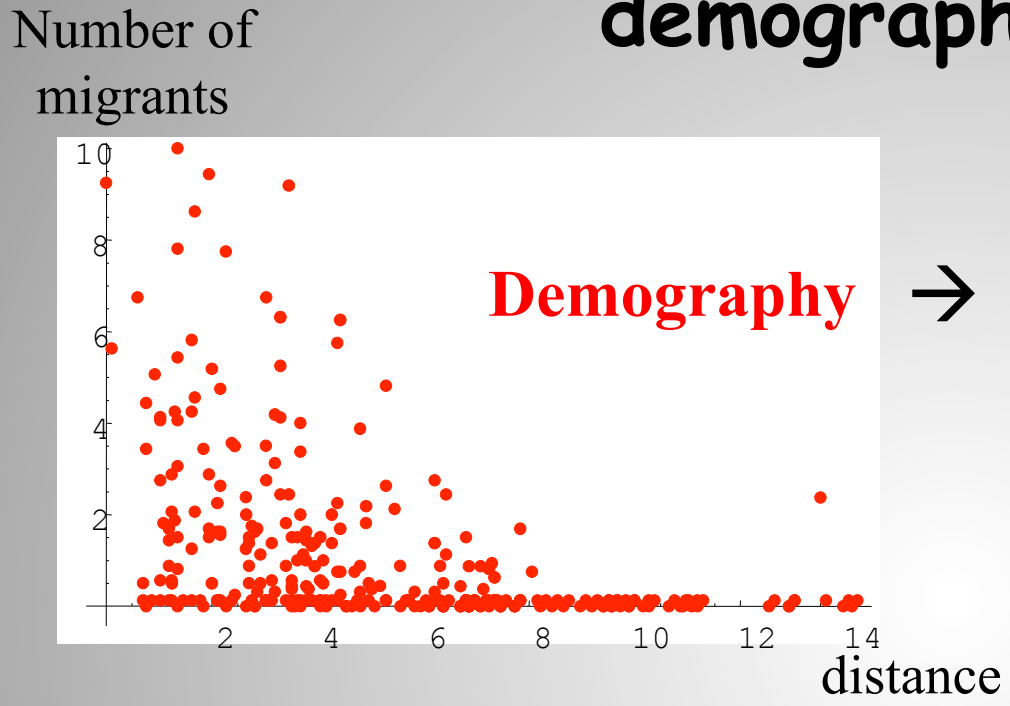




Limited dispersal : few kilometers per generation

Demographic data : Wood et al. *Am. Nat.* 1985
Genetic data (allozymes) : Long et al. *Am. J. Phys. Anth.*1986

First test of MIGRATE : comparison with demographic data



→ Inferred σ^2 :
1.9 km²/generation

moment based regression method

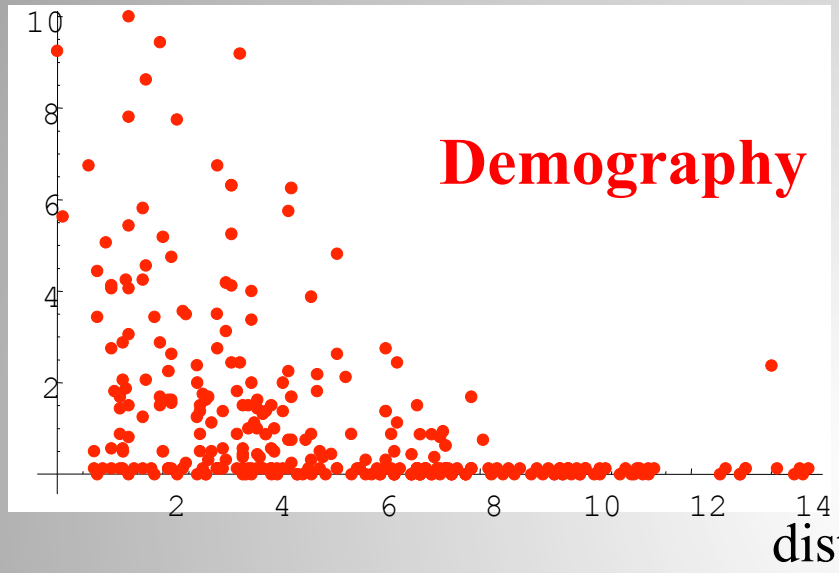
(a_r)

→ inference of σ^2 :

1.4 km²/generation

First test of MIGRATE : comparison with demographic data

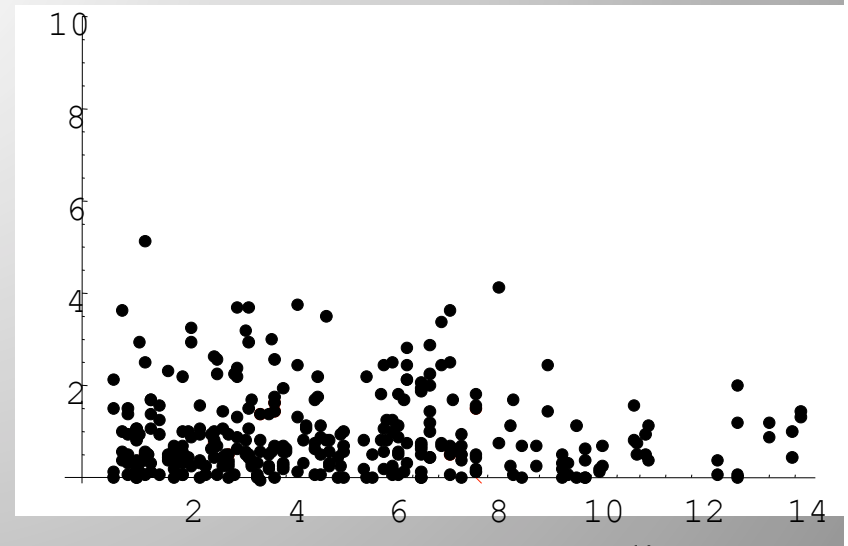
Number of migrants



Demography

→ Inferred σ^2 :
1.9 km²/generation

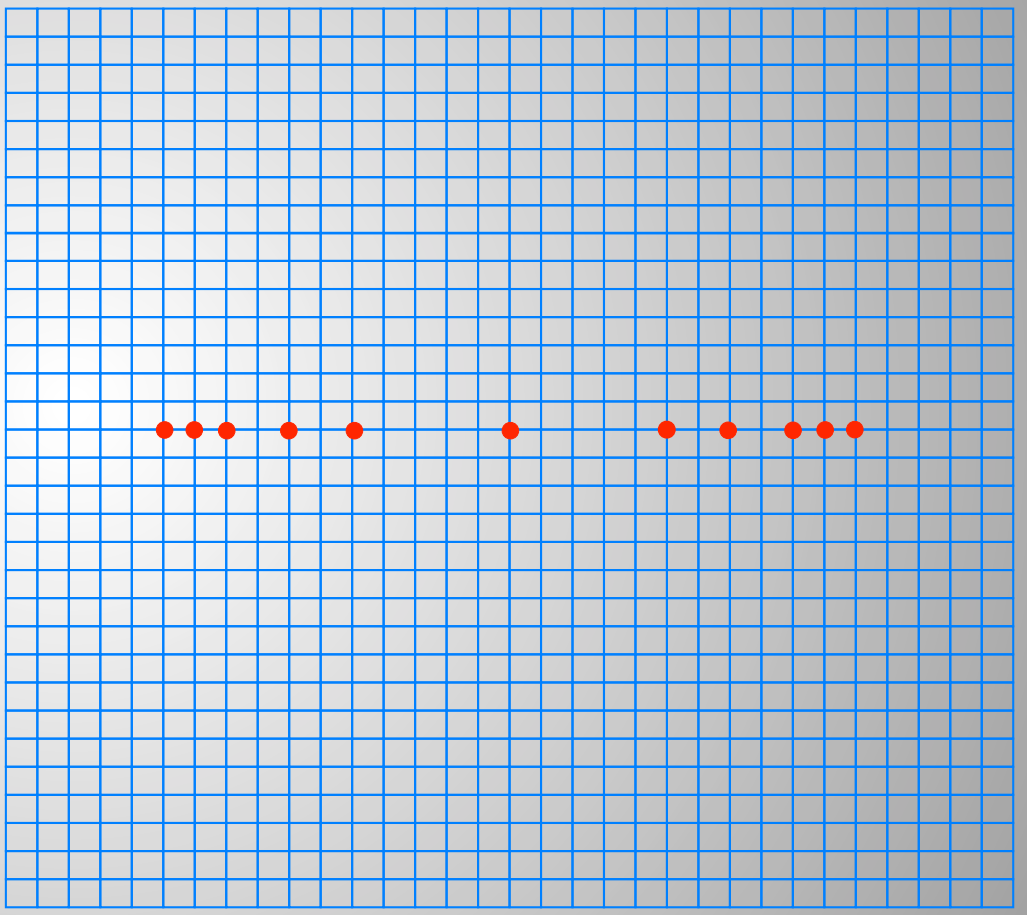
Number of migrants



MIGRATE
Over estimation of σ^2 :
16.3 km²/generation

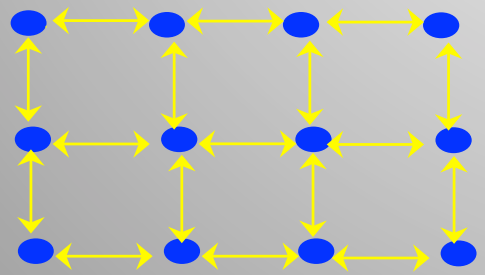
Complementary tests using simulations

11 samples (•) of 20 individuals evolving on a lattice of **40 000** (200x200) sub-populations



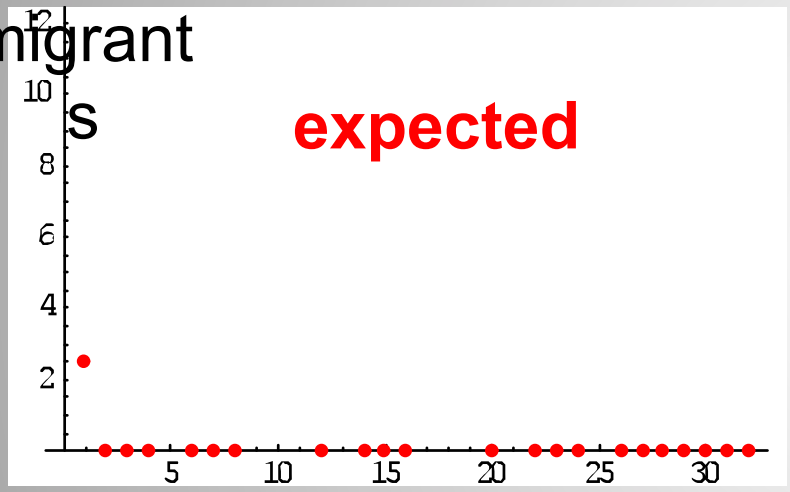
5 loci
KAM 10 alleles
Mutation rate of 5.10^{-4}

stepping stone migration



Inference of the number of migrants by MIGRATE

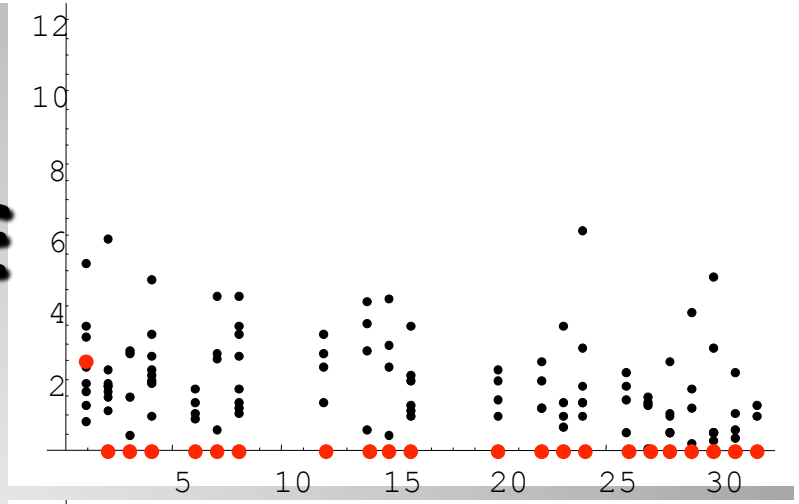
Number of migrants

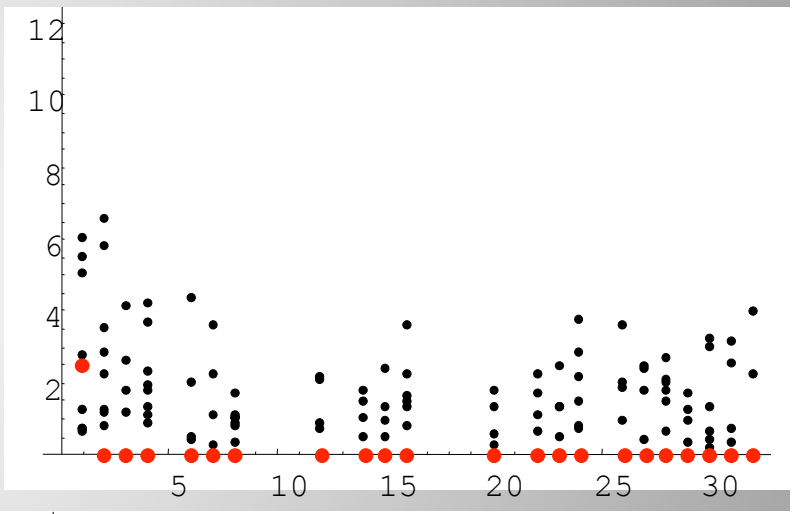


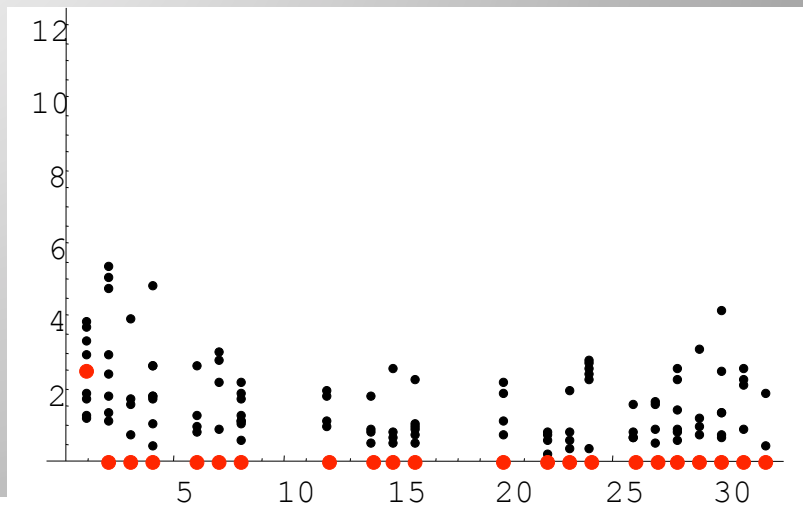
expected

distance

Obtained →







Over-estimation at large distances

Possible explanations…(1)

Inherent Bias of the method?

Yes

Observed on simulations by Beerli et Felsenstein (2001) :
Expected bias when low number of migrants

Possible explanations…(2)

Wrong mutation model?

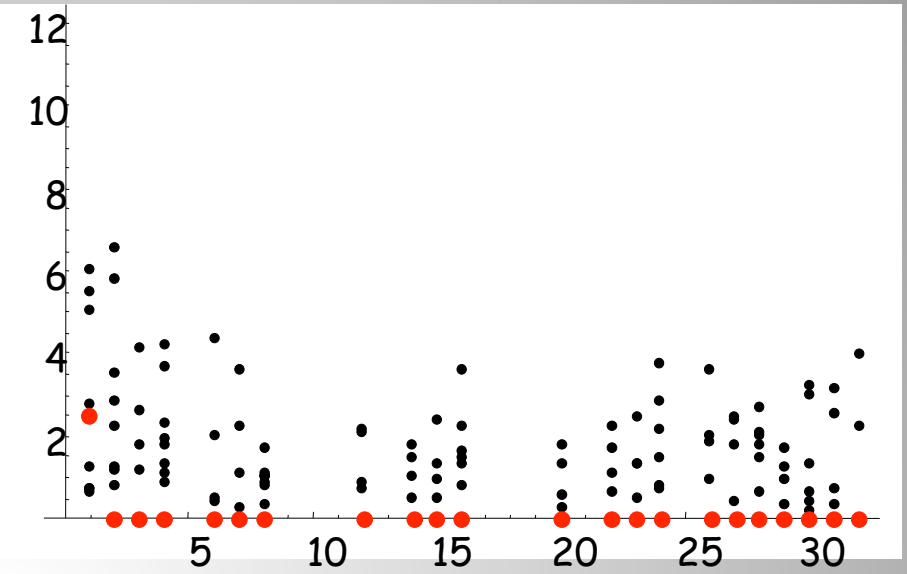
Number of populations

Slow convergence of MCMC?

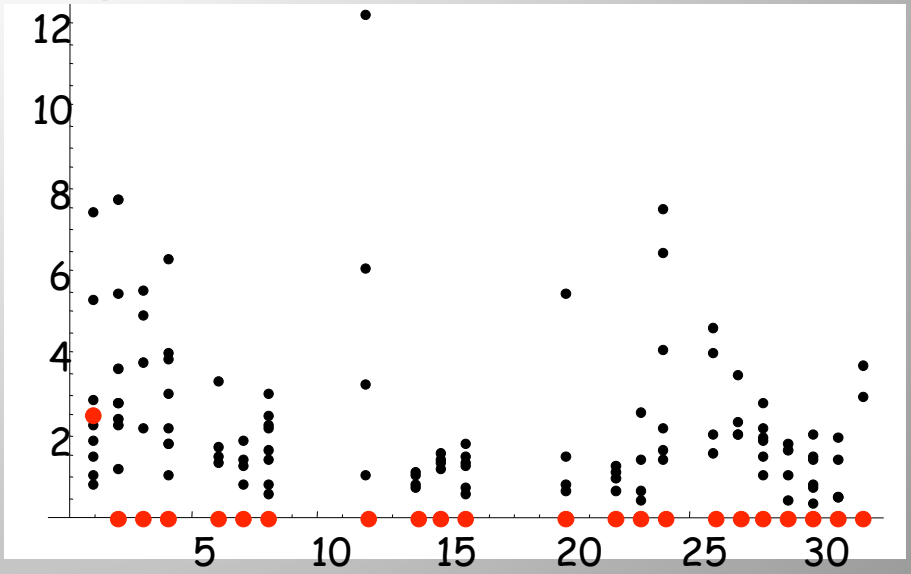
Inherent bias of the method?

No major effects

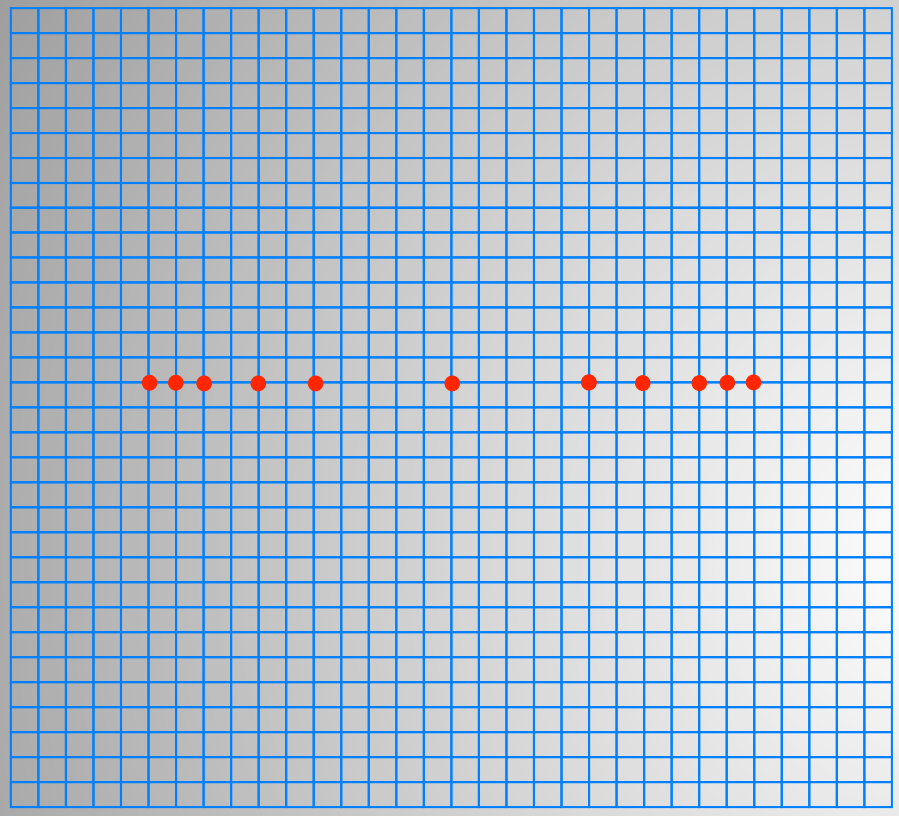
Small run (6 days at 1GHz)



Long run (3 weeks at 1Ghz)



Possible explanations…(3)



	MIGRATE	SIMULATIONS
Total number of sub-populations vs number of sampled sub-populations?	11	40000

Not easy to solve in practice

Many possible explanations…

Inherent bias to method?	Yes
Slow convergence of MCMC?	No major effects
Total nb of sub-populations VS nb of sampled sub-populations?	? – not easy to solve in practice
Too many parameters to infer?	Expected to have an important effect

Very slow → difficult to test

Bad precision under IBD