Estimating the evolution history of species from genome sequences

Simon Boitard, Olivier Mazet, Bertrand Servin

Estimating the evolution history of species from genome sequences

Simon Boitard[1], Bertrand Servin[1], Olivier Mazet[2]

1 : INRA, Génétique Physiologie et Systèmes d'Elevage (GenPhySE), Toulouse
2 : INSA, Institut de Mathématiques de Toulouse

CIMI, Semestre Mathématiques et Informatique pour les sciences du vivant
20 septembre 2017

Population genetics

- Inference of evolutionary history over short time scale, typically **within species** (in contrast to phylogeny).
- Typical questions (humans): "out of Africa" hypothesis, Neandertal introgression within modern humans ...
- Typical questions (breeding species): domestication process, impact of intensive selection since the 50's ...

Outline

1 Data and objectives

2 Single population model

3 Estimation of population size from single locus data

4 Estimation of population size from whole-genome sequences

Outline

1 Data and objectives

2 Single population model

3 Estimation of population size from single locus data

4 Estimation of population size from whole-genome sequences

Genetic data

- n DNA sequences of length L from the same species.
- Mostly similar, but at some positions different **alleles** exist (genetic polymorphism).

A-A-C-G-**G**-G-T-A-**T**-C-G- ....

A-A-C-G-**G**-G-T-A-**A**-C-G- ....

A-A-C-G-**C**-G-T-A-**T**-C-G- ....

Single Nucleotide Polymorphism (SNP)

- Only one nucleotide is changed.
- Very common on the genome.
- Result from a single mutation event during evolution
  $\rightarrow$ only two distinct alleles, one **ancestral** (denoted 0) and one **derived** (denoted 1).

A-A-C-G-**G**-G-T-A-**T**-C-G- ....

A-A-C-G-**G**-G-T-A-**A**-C-G- ....

A-A-C-G-**C**-G-T-A-**T**-C-G- ....

Single Nucleotide Polymorphism (SNP)

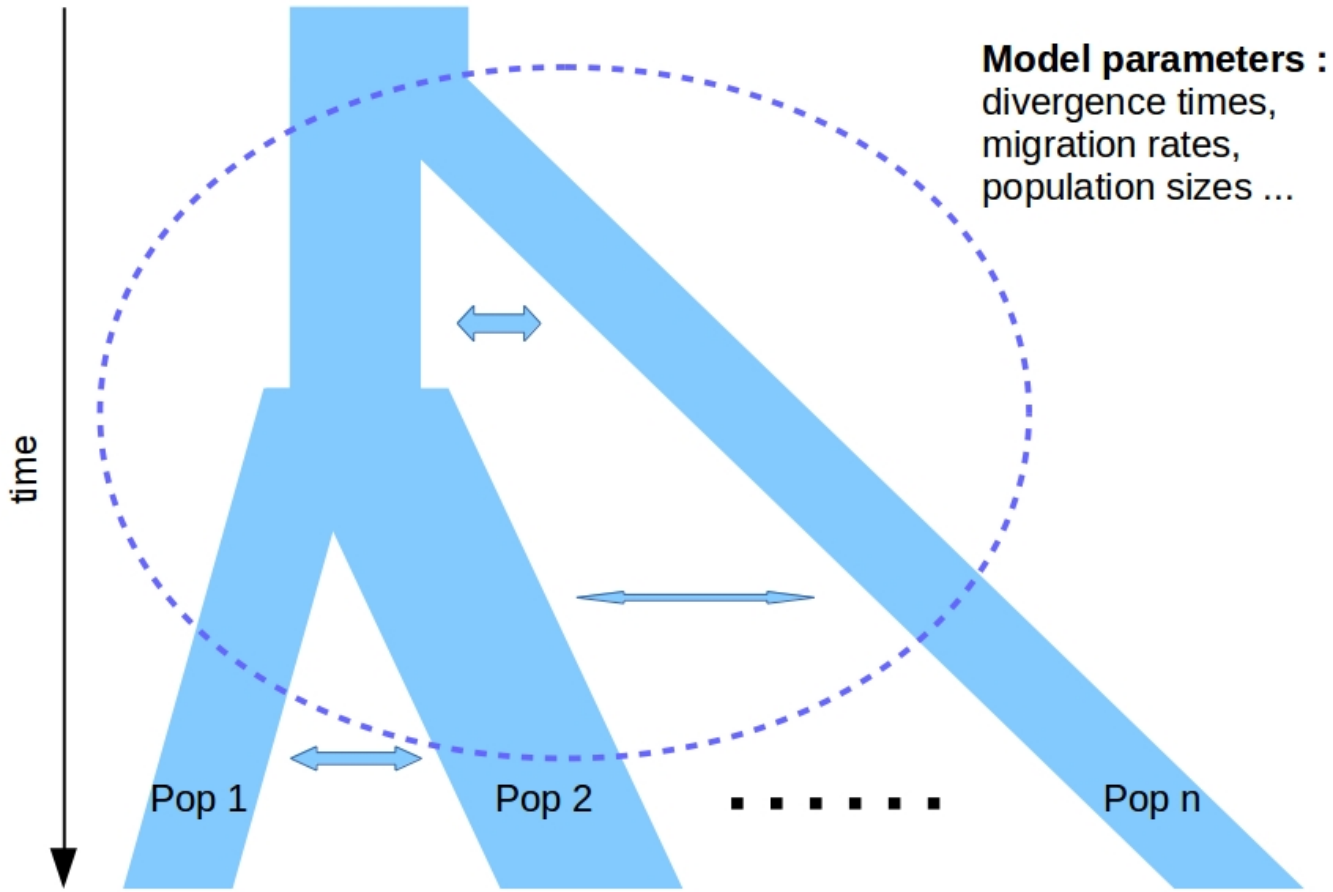
- Only one nucleotide is changed.
- Very common on the genome.
- Result from a single mutation event during evolution
  $\rightarrow$ only two distinct alleles, one **ancestral** (denoted 0) and one **derived** (denoted 1).

0-0-0-0-**1**-0-0-0-**0**-0-0- ....

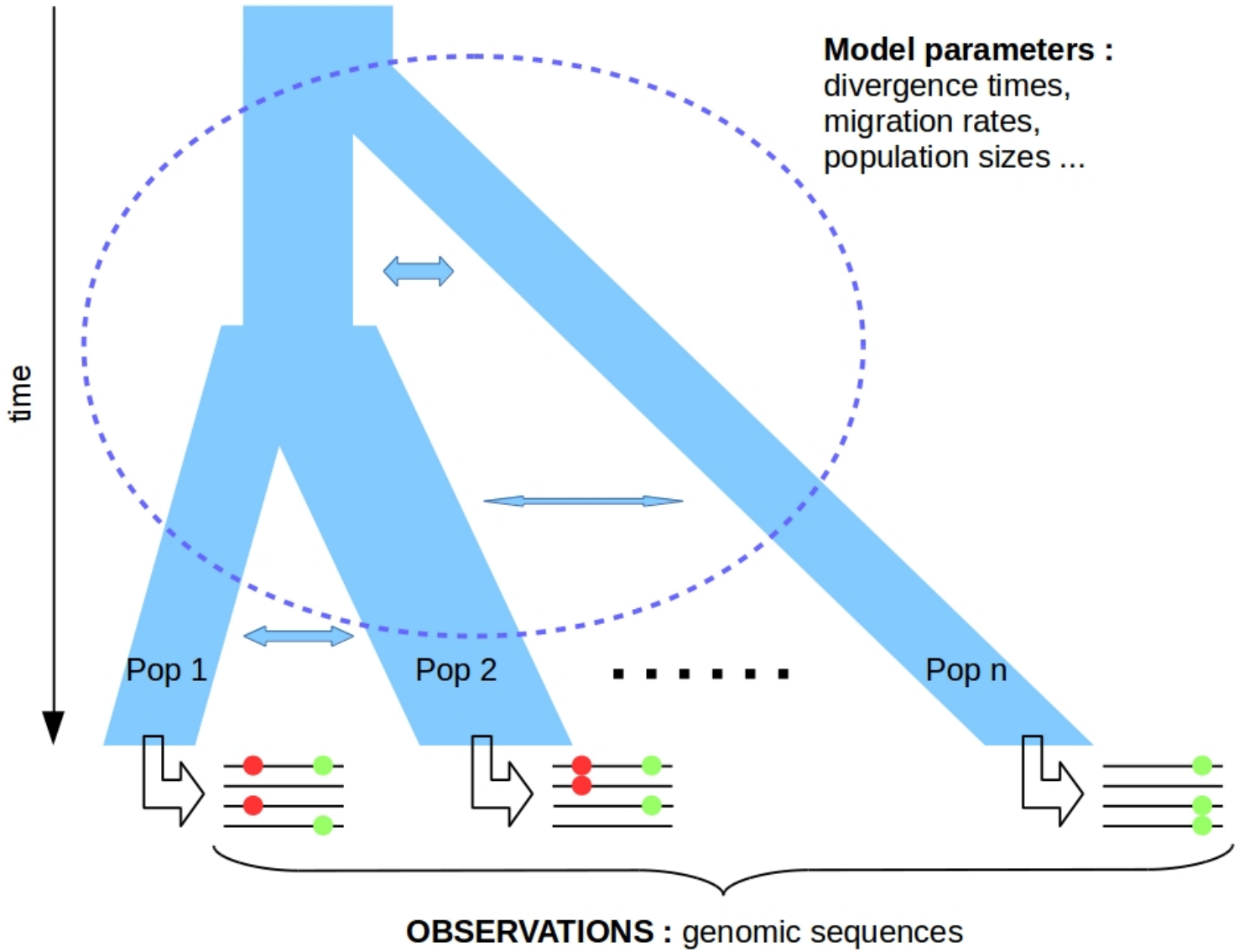
0-0-0-0-**1**-0-0-0-**1**-0-0- ....

0-0-0-0-**0**-0-0-0-**0**-0-0- ....

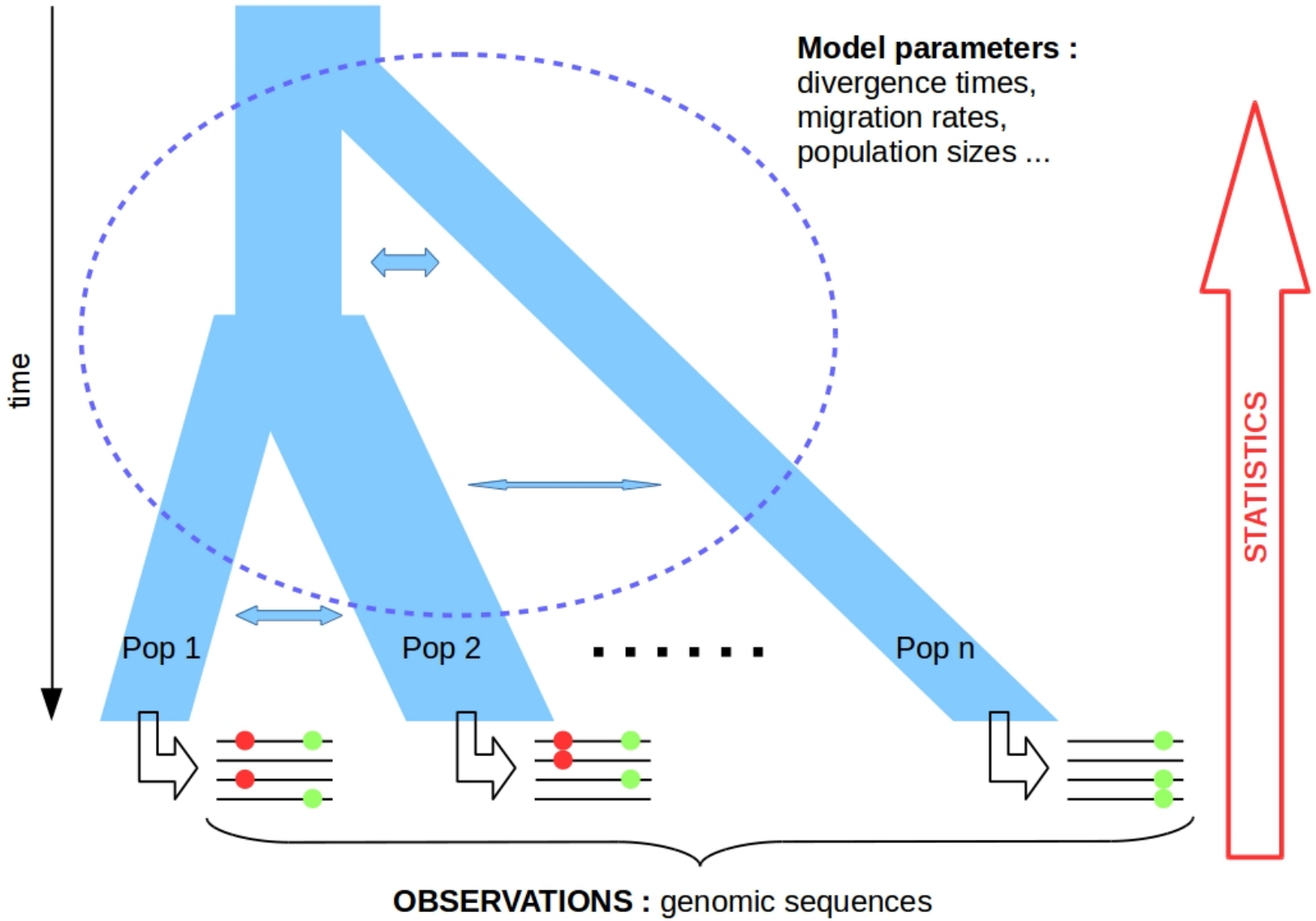
General evolution model



General evolution model



General evolution model



Outline

1 Data and objectives

2 Single population model

3 Estimation of population size from single locus data

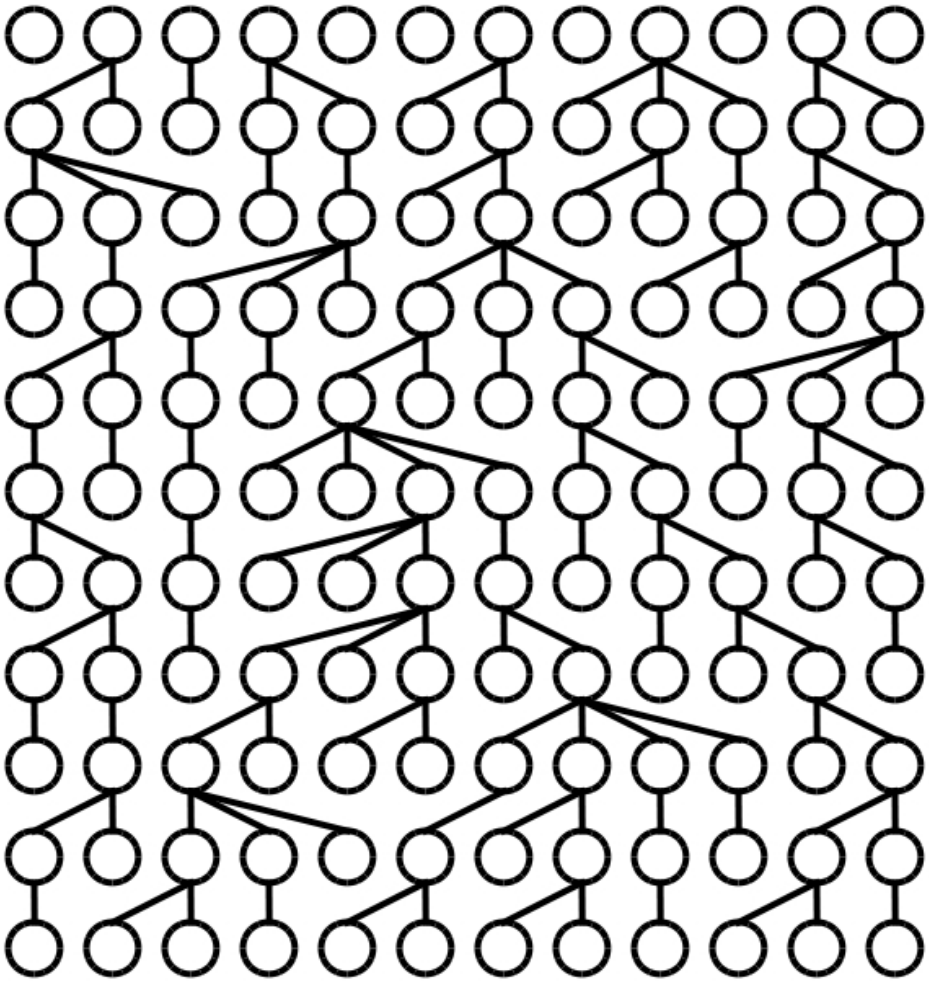
4 Estimation of population size from whole-genome sequences

the Wright-Fisher process

the Wright-Fisher process

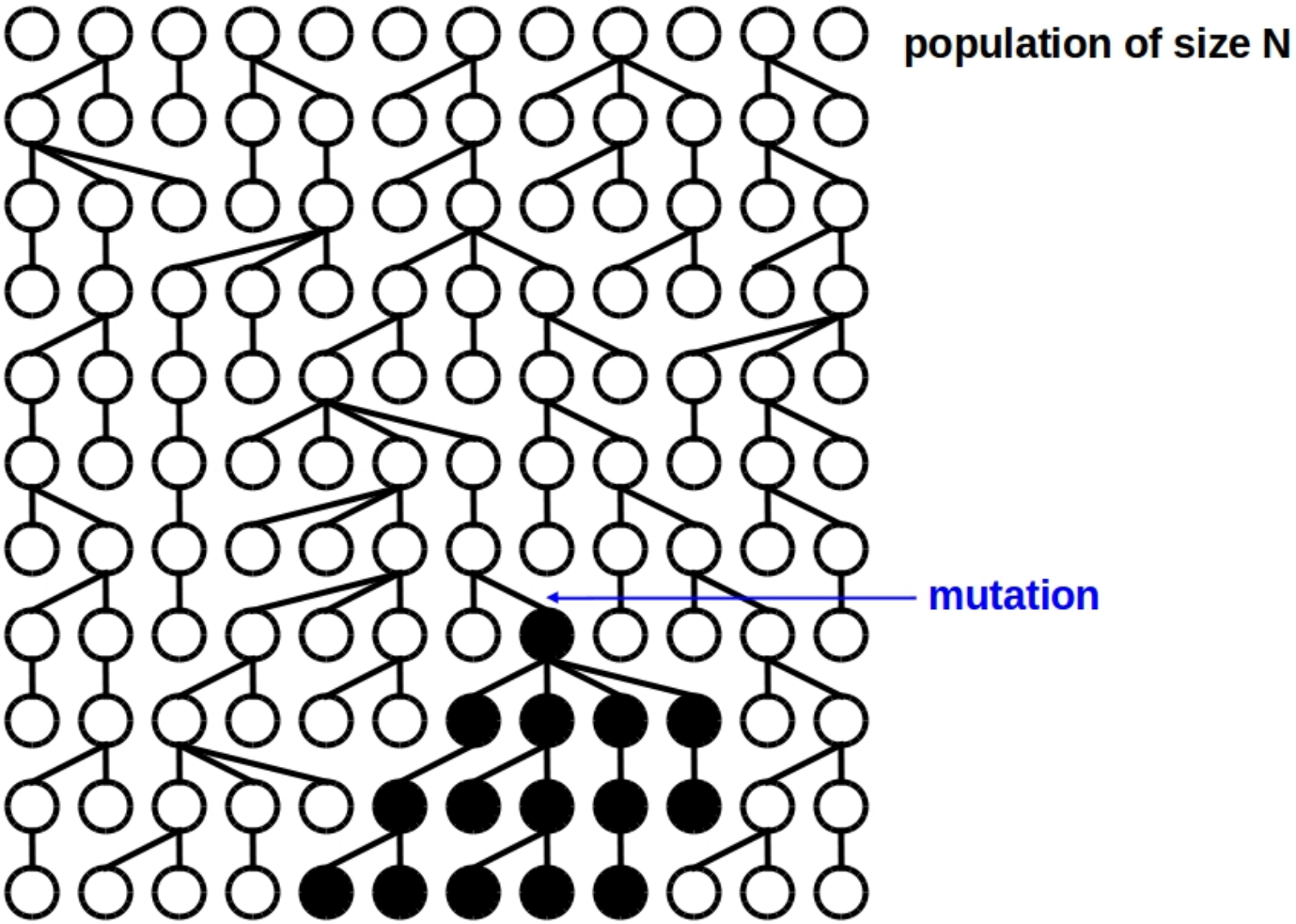
population of size N

the Wright-Fisher process

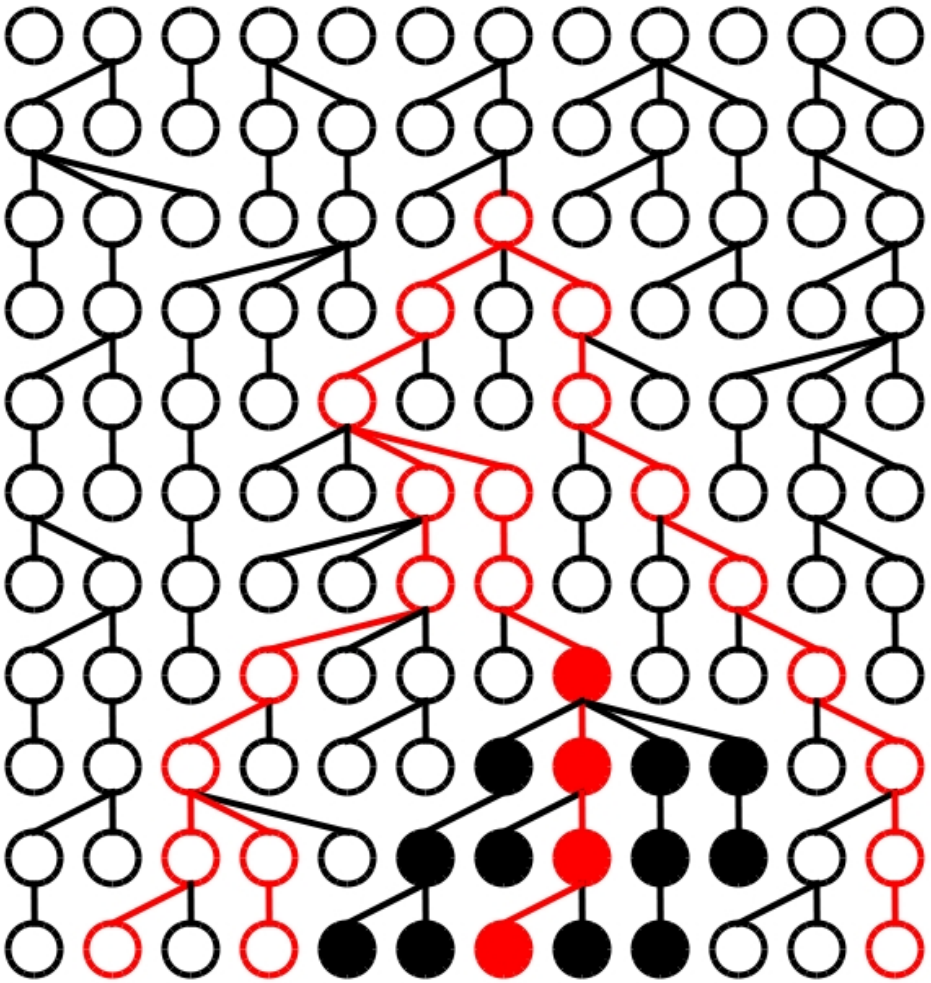


population of size N

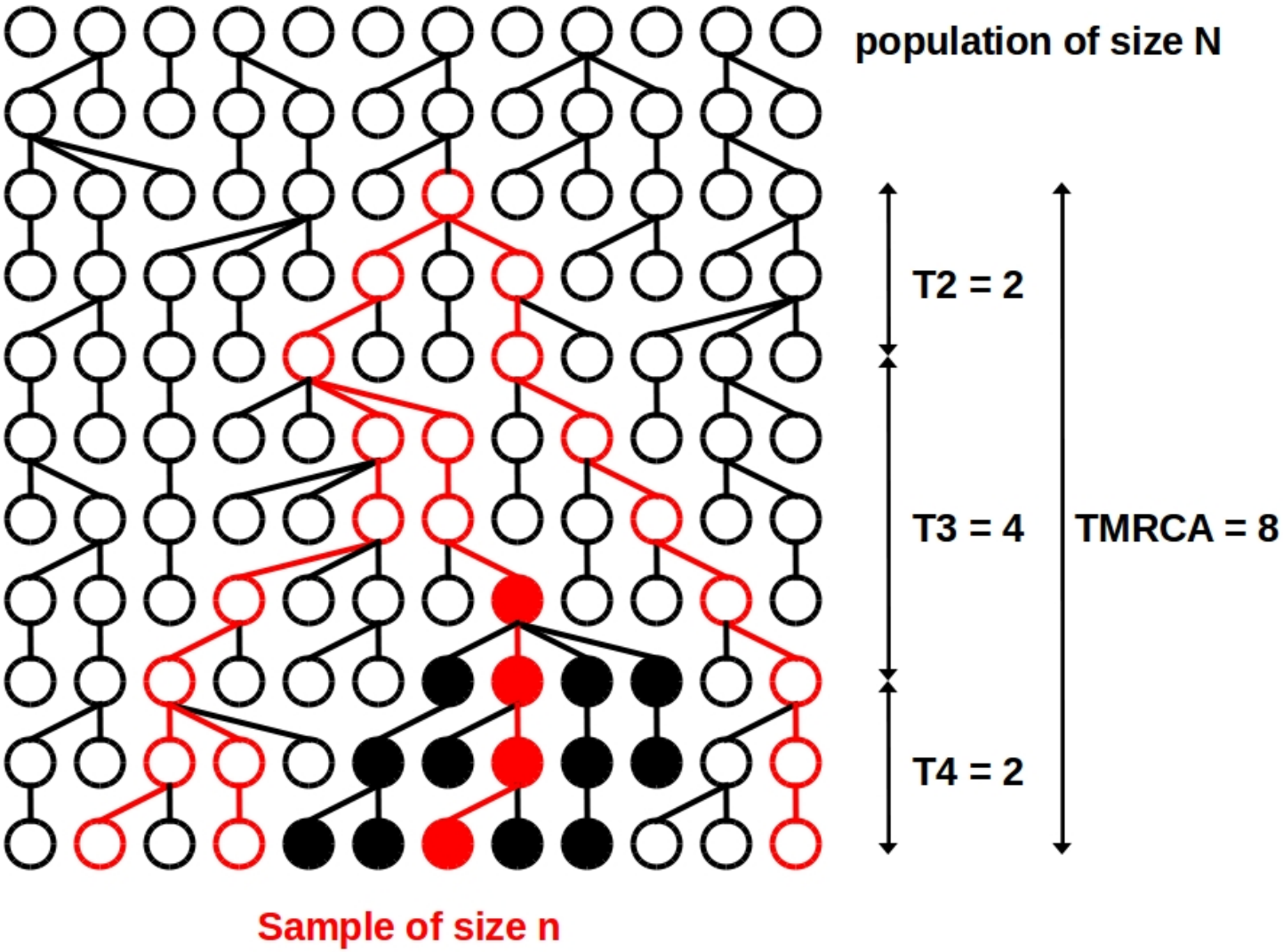
the Wright-Fisher process



The coalescent process



The coalescent process



Important properties

- At each generation, **probability** that **no coalescence** occurs is

$$q^N(n) = \prod_{i=1}^{n-1}(1 - \frac{i}{N}) = 1 - \frac{n(n-1)}{2N} + O(\frac{1}{N^2})$$

- Coalescence time T_k^N $(2 \leq k \leq n)$ has geometric distribution

$$\mathbb{P}(T_k^N > t) = (q^N(k))^t$$

- All lineages coalesce at the same rate.
- Number of mutations on a branch of length t is Binomial $\mathcal{B}(t, \mu)$, μ mutation rate per meiosis and per nucleotide (bioliogically known).

Kingman's coalescent (1982)

- $N \to +\infty$, rescaled time $\tau = \frac{t}{N}$
- Coalescence time T_k^N tends to T_k, with exponential distribution

$$\mathbb{P}(T_k > \tau) = e^{-\frac{k(k-1)}{2}\tau}$$

- **Coalescence** events imply **one single pair of lineages**.
- Number of mutations on a branch of length τ is Poisson $\mathcal{P}(\frac{\theta}{2}\tau)$, with $\theta = 2N\mu$ (population scaled mutation rate).

Advantages of the coalescent approach

- Very efficient way to **simulate genetic data**.
- Easily extended to **more complex models** (variable population sizes, structured populations . . .).
- Used to **express the likelihood** of observed genetic data.
- Provides **conceptual framework** to understand the influence of some evolutionary parameters on observed genetic data.

Outline

1 Data and objectives

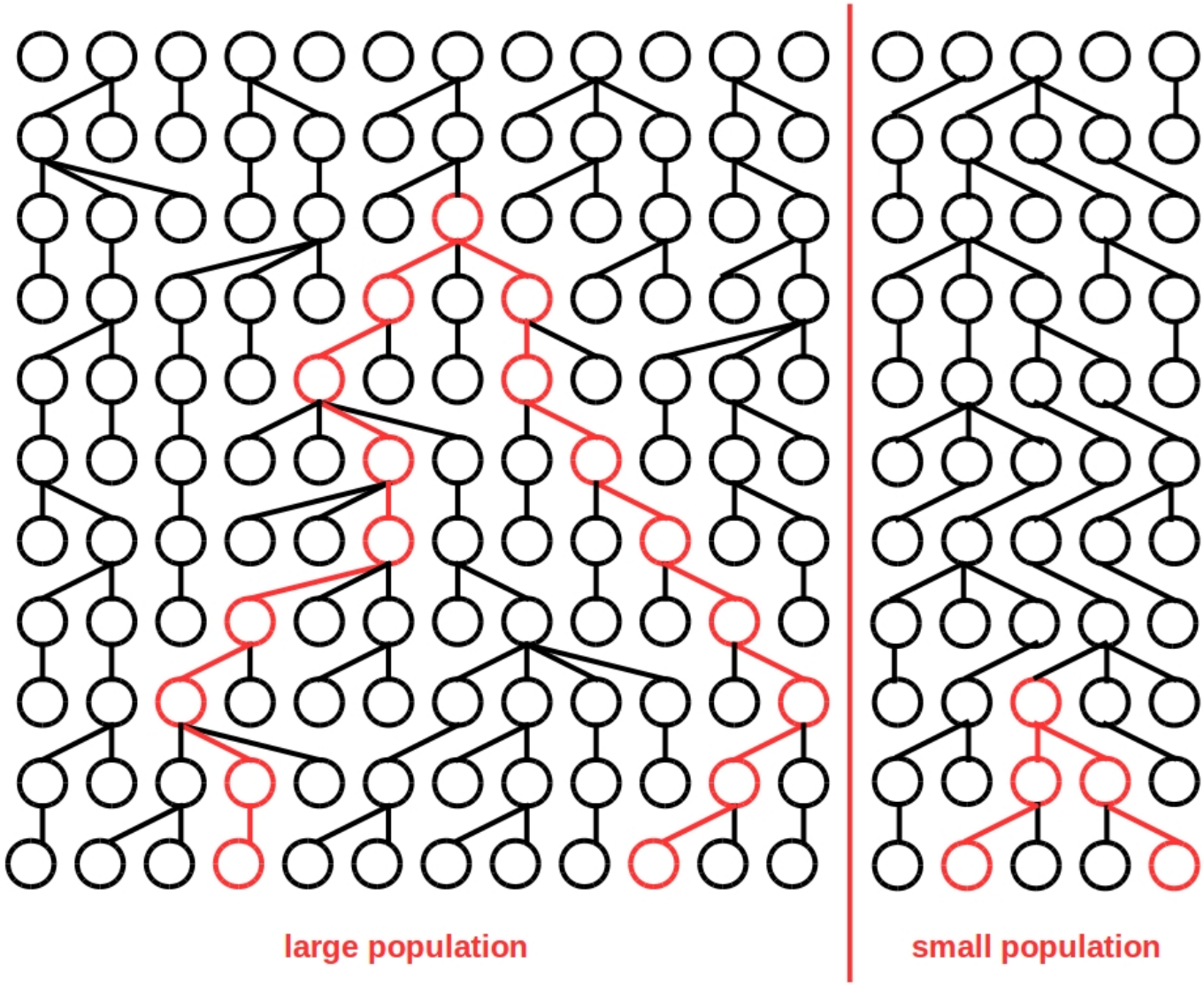
2 Single population model

3 Estimation of population size from single locus data

4 Estimation of population size from whole-genome sequences

Constant population size : intuition

larger $N \rightarrow$ longer coalescence times $\rightarrow$ more mutations.



Constant population size : the Watterson estimator

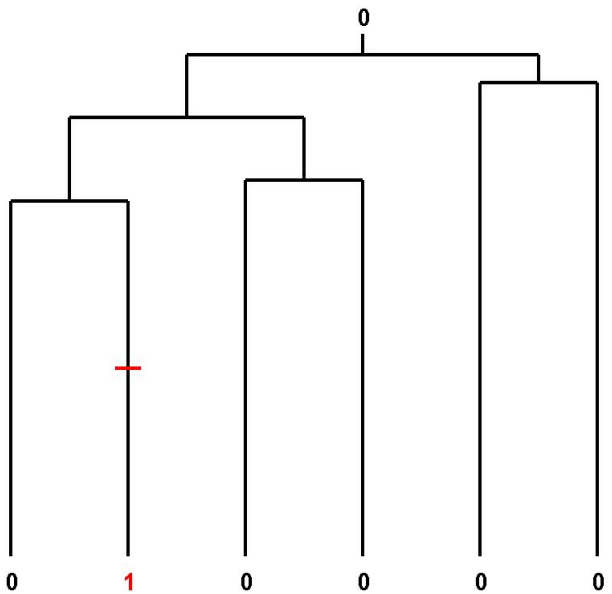
- S_n number of polymorphic sites in a sample of n DNA sequences of length L.

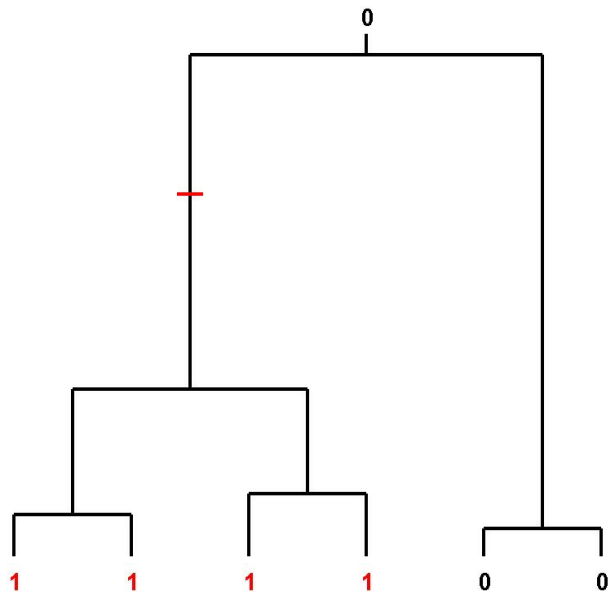
$$\theta_W = \frac{1}{L} S_n \left(\sum_{k=1}^{n-1} \frac{1}{k}\right)^{-1}$$

- $\mathbb{E}[\theta_W] = \theta = 2N\mu$.
- As μ is known, this provides an **unbiased estimator of** N.
- $\mathrm{Var}(\theta_W) = (\theta^2 \sum_{k=1}^{n-1} \frac{1}{k^2} + \frac{\theta}{L} \sum_{k=1}^{n-1} \frac{1}{k})(\sum_{k=1}^{n-1} \frac{1}{k})^{-2}$

Variable population size : intuition

- Population **expansion** $\rightarrow$ larger coalescence times in the recent past $\rightarrow$ higher proportion of **derived alleles at low frequency**.
- Population **decline** $\rightarrow$ larger coalescence times in the distant past $\rightarrow$ higher proportion of **derived alleles at intermediate frequency**.





Variable population size : estimation approaches

- **Likelihood**:

$$\mathbb{P}(\mathcal{D} \mid N()) = \sum_{\mathcal{T}} \mathbb{P}(\mathcal{D} \mid \mathcal{T}) \mathbb{P}(\mathcal{T} \mid N())$$

  $\mathcal{D}$ observed sequences, $N()$ population size history, $\mathcal{T}$ coalescence tree.

- **No analytical expression**

$$\mathbb{P}(\mathcal{D} \mid N()) \approx \sum_{i=1}^{I} \mathbb{P}(\mathcal{D} \mid \mathcal{T}_i)$$

  $\mathcal{T}_i$ simulated from $\mathbb{P}(\mathcal{T} \mid N())$.

- **High dimension** of $\mathcal{T}$
  $\rightarrow$ explore using **Markov Chain Monte Carlo** (MCMC) or **Importance Sampling** (IS) algorithms (Beaumont, 1999; Drummond et Rambaut, 2007; Hobolt et al, 2008).

Outline

1 Data and objectives

2 Single population model

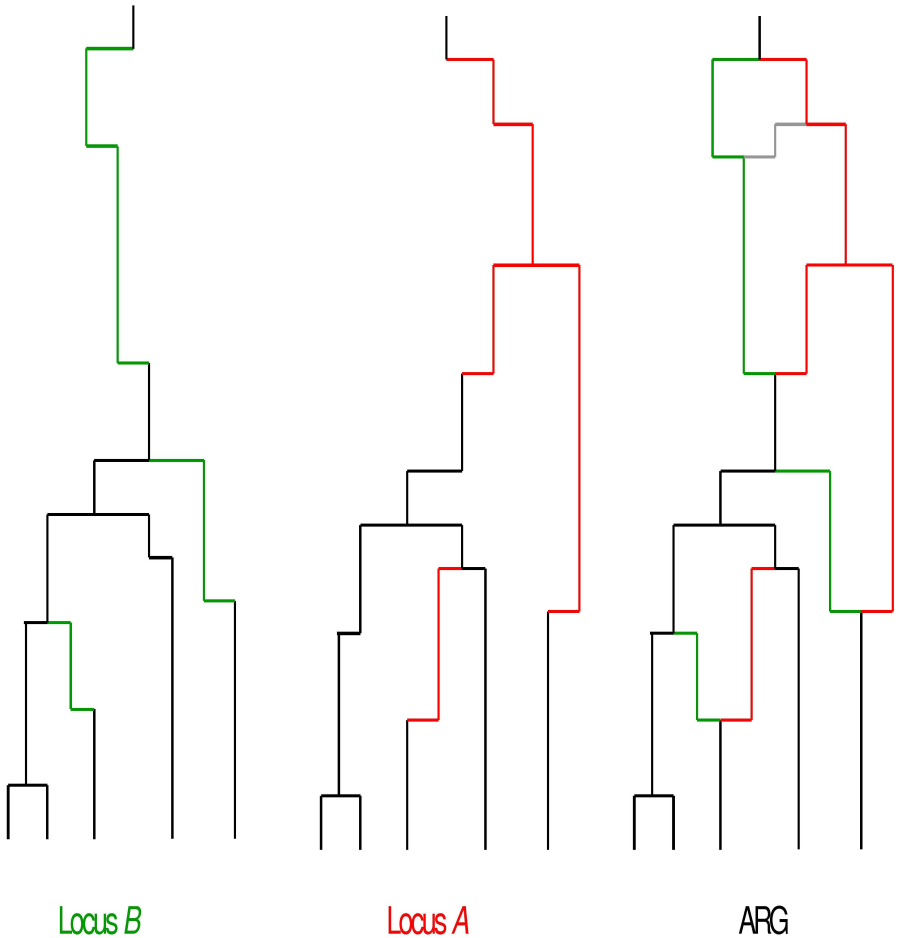
3 Estimation of population size from single locus data

4 Estimation of population size from whole-genome sequences

Recombination

- In diploid species (e.g. humans), recombination during meiosis $\rightarrow$ each gamete is a **mixture of two sequences**, one inherited from the mother and one from the father.
- Negligible at short distance (single locus), but important when studying whole-genome sequences.
- The **genealogy** of n DNA sequences becomes a **graph**.

The Ancestral Recombination Graph (ARG)



Consequences for inference

- Coalescence trees at two distinct loci are neither similar nor independent.
- **Complex correlation structure**.
- Dimension of the space of genealogies explodes, **previous** MCMC or IS **approaches no longer possible**.
- Open and active research area.

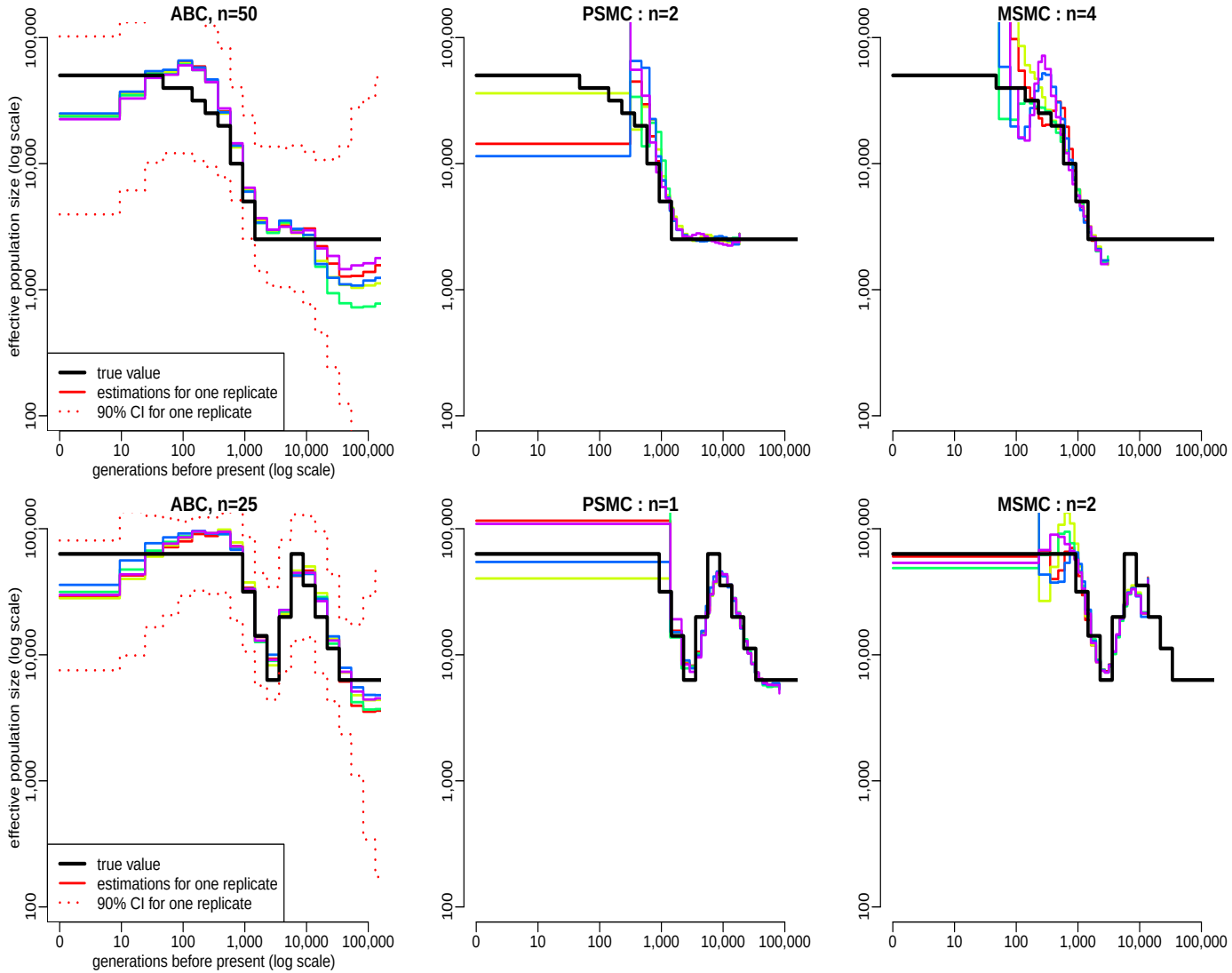
The Approximate Bayesian Computation (ABC) approach

- First proposed by Beaumont (2002), allows estimating parameters θ of a model when likelihood cannot be evaluated.
- Approximate the posterior $\mathbb{P}(\theta|\mathcal{D})$ by the posterior $\mathbb{P}(\theta|\mathcal{S})$, for a set $\mathcal{S}$ of (meaningfull!) **summary statistics**.
- Estimate $\mathbb{P}(\theta|\mathcal{S})$ using **intensive simulations**:
  1. Compute $\mathcal{S} = f(\mathcal{D})$
  2. For i from 1 to I:
     1. Sample parameter θ_i from a prior distribution.
     2. Simulate dataset $\mathcal{D}_i$ from the model with parameter θ_i.
     3. Compute $\mathcal{S}_i = f(\mathcal{D}_i)$.
     4. Select the simulation if $dist(\mathcal{S}_i, \mathcal{S}) < \epsilon$.
  3. Estimate the posterior distribution of θ from the empirical distribution of selected θ_i values.

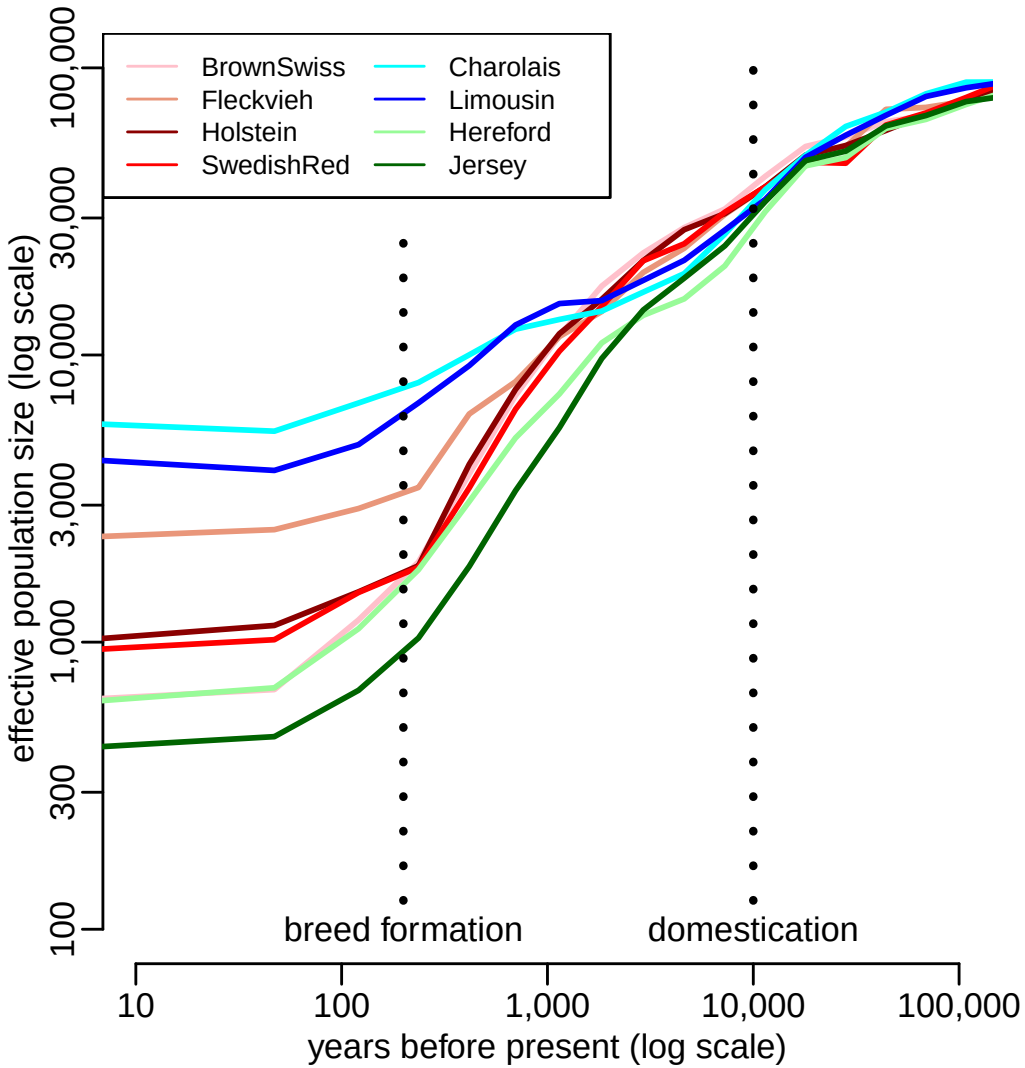
Application to population sizes (Boitard *et al*, 2016)

- **Panmictic population** with population size history $N()$.
- **Large sample of whole-genome sequences** from this population.
- Approximates $\mathbb{P}(N()|\mathcal{S})$ using ABC.
- Set of ≈ 50 summary statistics, describing (among others) the **distribution of allele frequencies**.

Simulation results



Analysis of cattle genomes



- **Common trajectory before domestication**.
- Continuous decline since domestication.
- Ranking of recent sizes consistent with current knowledge of these breeds.

Influence of population structure

- In real life, **populations not isolated**.
- Relationship between populations **affect population size estimations**.
  $\rightarrow$ **Identifiability issue:** can we distinguish population size changes and population structure from genetic data?
- Mazet *et al* (2016): **not from two genomes**, because population size change models can reproduce every possible distribution of T_2.
- Important conclusion, because one popular estimation method (Li and Durbin, 2011) is based on this distribution.
- Distinction would be in theory possible from the joint distribution of (T_2, T_3) (Grusea *et al*, in prep.).

Conclusions

- **Genetic data informative** about species history.
- **Population genetics:** a very active field of research, interface between biology and applied mathematics.
- Contributes to answer **fundamental questions** about human (and other species) history.
- Many **theoretical and computational issues** to be solved.
- **Massive amount of data** to be analyzed.

Our research group in Toulouse

- INRA: Bertrand Servin, Simon Boitard
- INSA/IMT: Olivier Mazet, Simona Grusea, Willy Rodriguez, Didier Pinchon
- EDB : Lounès Chikhi