Three Lectures about :
“Evolutionary Processes and Patterns of Biodiversity”

Lecture 1/3 : Genealogy of one or many genes

Amaury Lambert

amaury_upmc

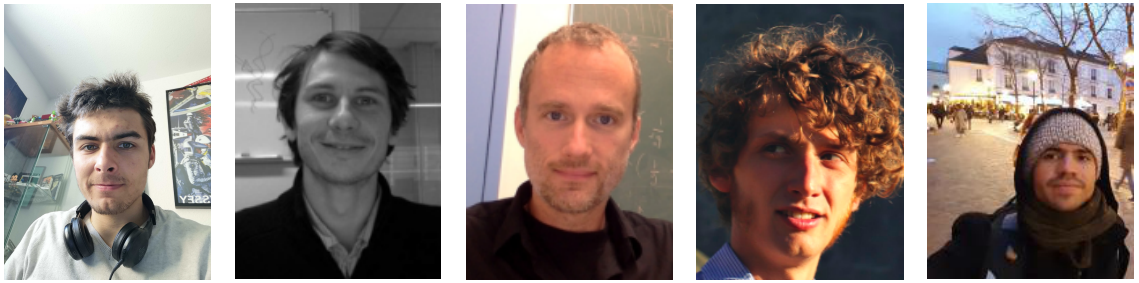
IICD & Probability and Society Initiative Joint Seminar Series
Columbia University
October 9, 2020

SMILE : An interdisciplinary group in Paris

Below : SMILE members in May 2020

Jasmine 12/01 | François 15/01 | Guillaume T. 25/02 | Pete 29/03 | Emmanuel 18/04

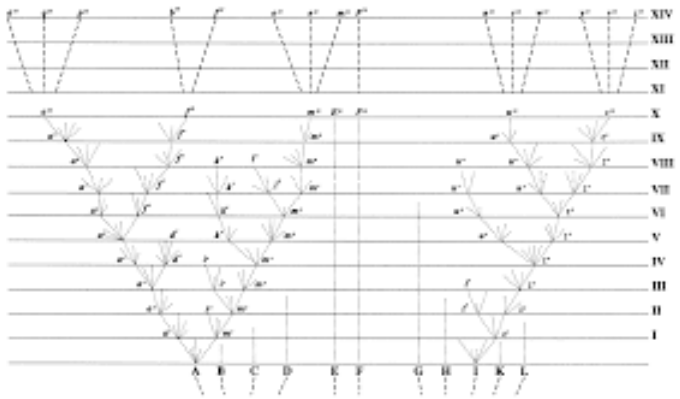
Laura 21/05 | Jean-Jil 24/05 | Felix 05/09 | Rob 16/09 | Julie 21/09 | Elise 10/10



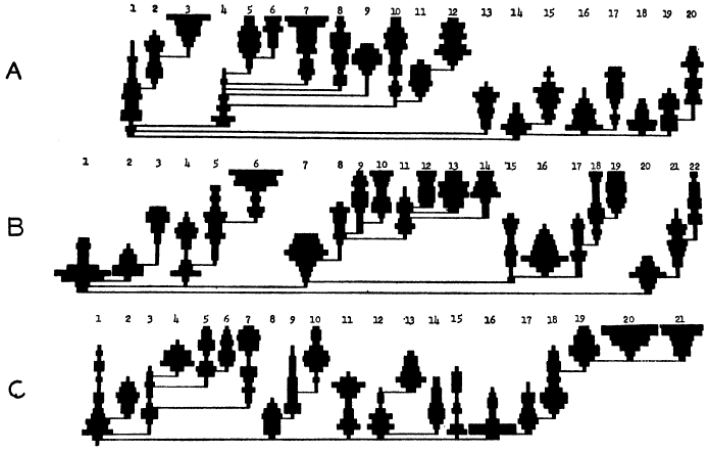
Léo 07/11 | Guillaume A. 09/12 | Amaury 16/12 | Philibert 30/12 | Alejandro

SMILE = **S**tochastic **M**odels for the **I**nference of **L**ife **E**volution

Evolution as a generic process



Darwin (1859)



Raup et al (1973)

- Darwin, Wallace, Lamarck : Evolution is a generic process
- 1920-30's First models of micro-evolution ('population genetics')
  - 1924–34 Haldane "A mathematical theory of natural and artif selection"
  - 1931 Wright "Evolution in Mendelian populations"
  - 1937 Fisher "The wave of advance of advantageous genes"
- 1960-70's First models of macro-evolution
  - 1925 Yule "A mathematical theory of evolution, based on..."
  - 1967 Cavalli-Sforza & Edwards "Phylogenetic analysis : models and estimation procedures"
  - 1973 Farris "A probability model for inferring evolutionary trees"
  - 1973 Raup, Gould, Schopf & Simberloff "Stochastic models of phylogeny and the evolution of diversity"
  - 1985 Felsenstein "Phylogenies and the comparative method"

Outline

Introduction : genetic diversity, relatedness, genealogy

- Sample n individuals from a population and sequence their DNA.

Introduction : genetic diversity, relatedness, genealogy

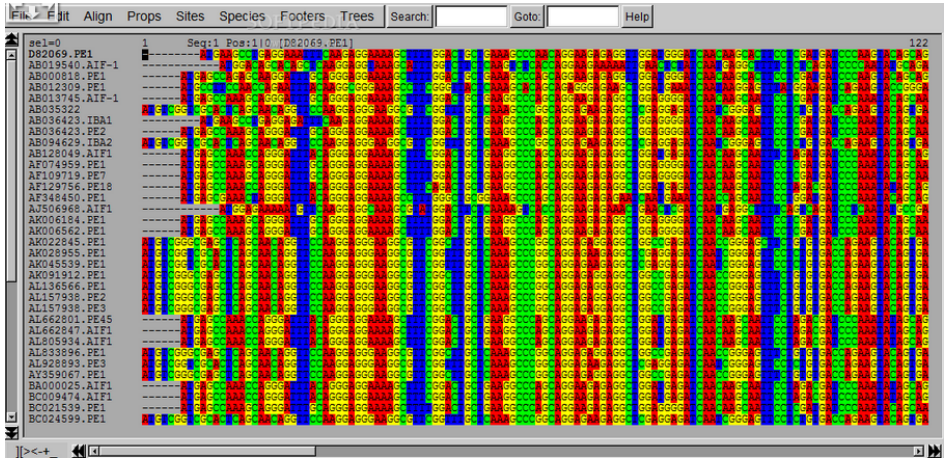
- Sample *n* individuals from a population and sequence their DNA.

- **Q. What diversity do you expect to observe in this sample of DNA sequences ?**

Introduction : genetic diversity, relatedness, genealogy

- Sample *n* individuals from a population and sequence their DNA.

- **Q. What diversity do you expect to observe in this sample of DNA sequences ?**



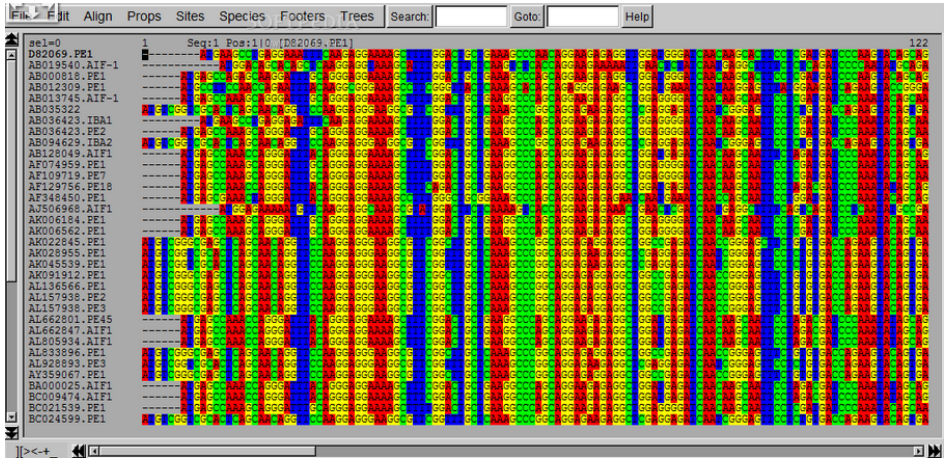
A multiple sequence alignment

Introduction : genetic diversity, relatedness, genealogy

- Sample *n* individuals from a population and sequence their DNA.

- **Q. What diversity do you expect to observe in this sample of DNA sequences ?**

- Mutations : mutation rate ↑, diversity ↑



A multiple sequence alignment

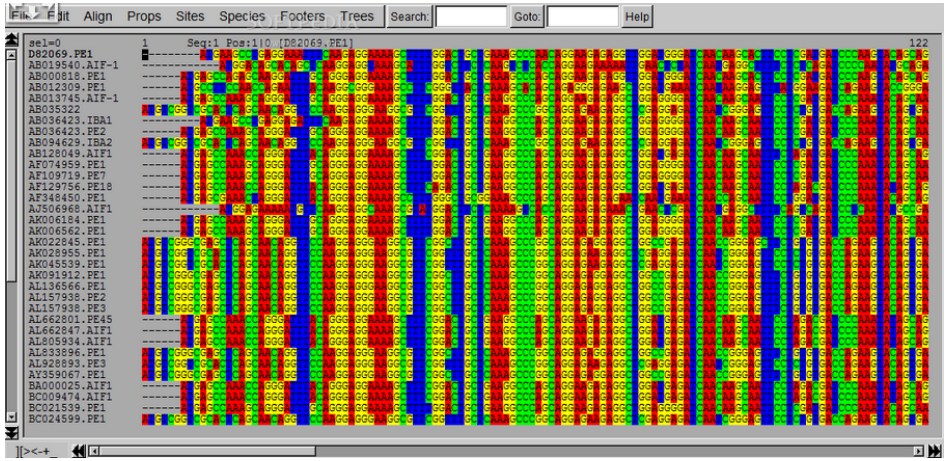
Introduction : genetic diversity, relatedness, genealogy

- Sample *n* individuals from a population and sequence their DNA.

- **Q. What diversity do you expect to observe in this sample of DNA sequences ?**

- Mutations : mutation rate ↑, diversity ↑

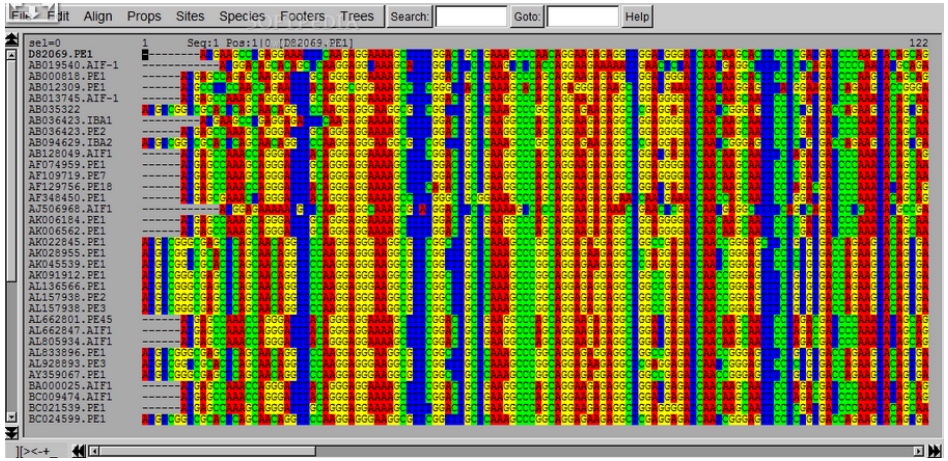
- Genealogy : relatedness ↑, diversity ↓



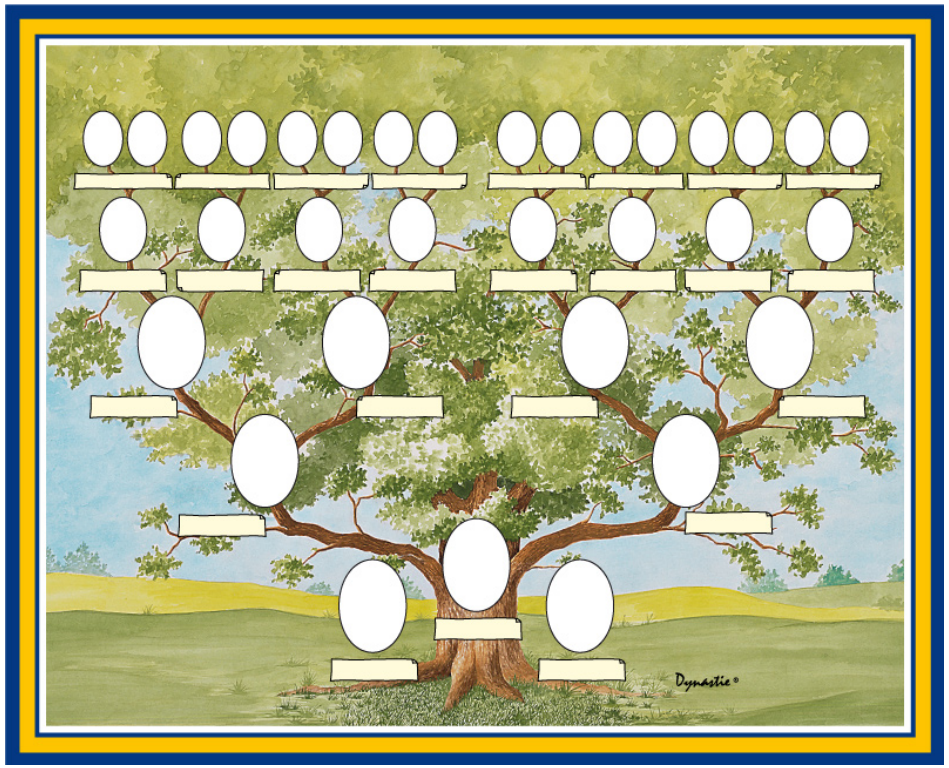
A multiple sequence alignment

Introduction : genetic diversity, relatedness, genealogy

- Sample *n* individuals from a population and sequence their DNA.
- **Q. What diversity do you expect to observe in this sample of DNA sequences ?**
- Mutations : mutation rate ↑, diversity ↑
- Genealogy : relatedness ↑, diversity ↓



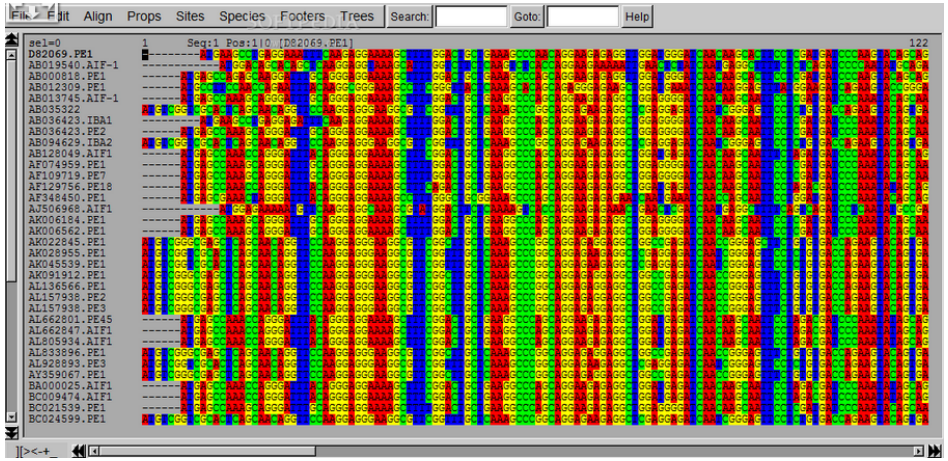
A multiple sequence alignment



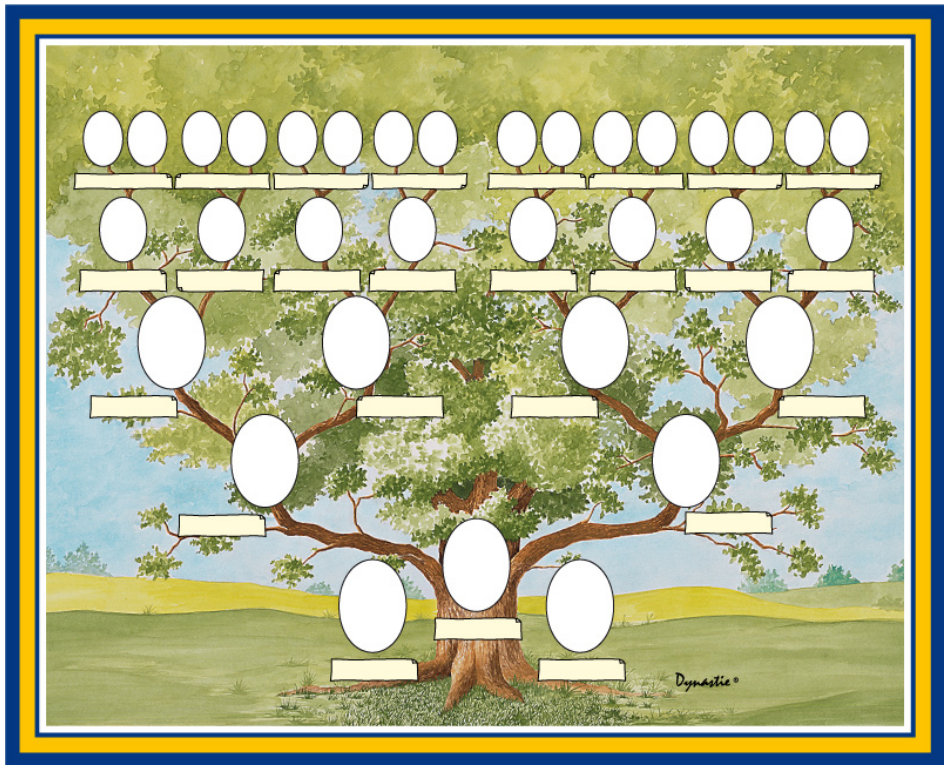
A pedigree = 2 parents ($\neq$ genealogy = 1 parent)

Introduction : genetic diversity, relatedness, genealogy

- Sample *n* individuals from a population and sequence their DNA.
- **Q. What diversity do you expect to observe in this sample of DNA sequences ?**
- Mutations : mutation rate ↑, diversity ↑
- Genealogy : relatedness ↑, diversity ↓
- Each locus (gene, site) is inherited from one single parent : simple gene genealogy



A multiple sequence alignment

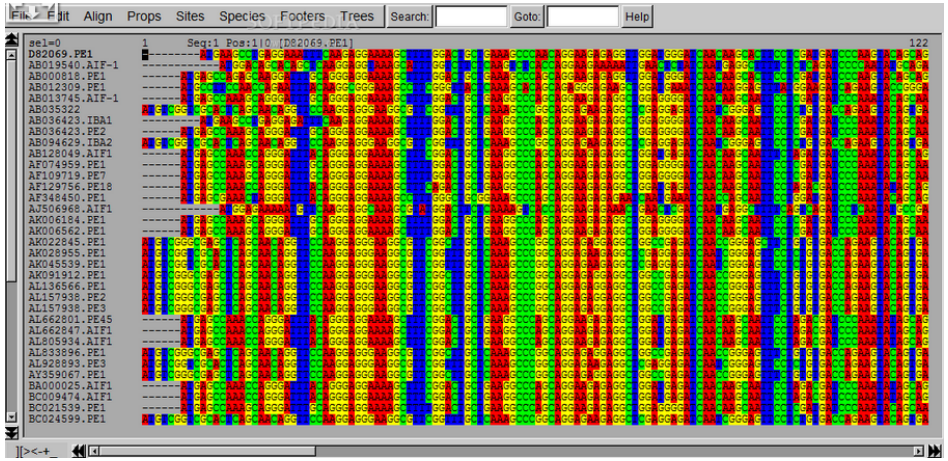


A pedigree = 2 parents ($\neq$ genealogy = 1 parent)

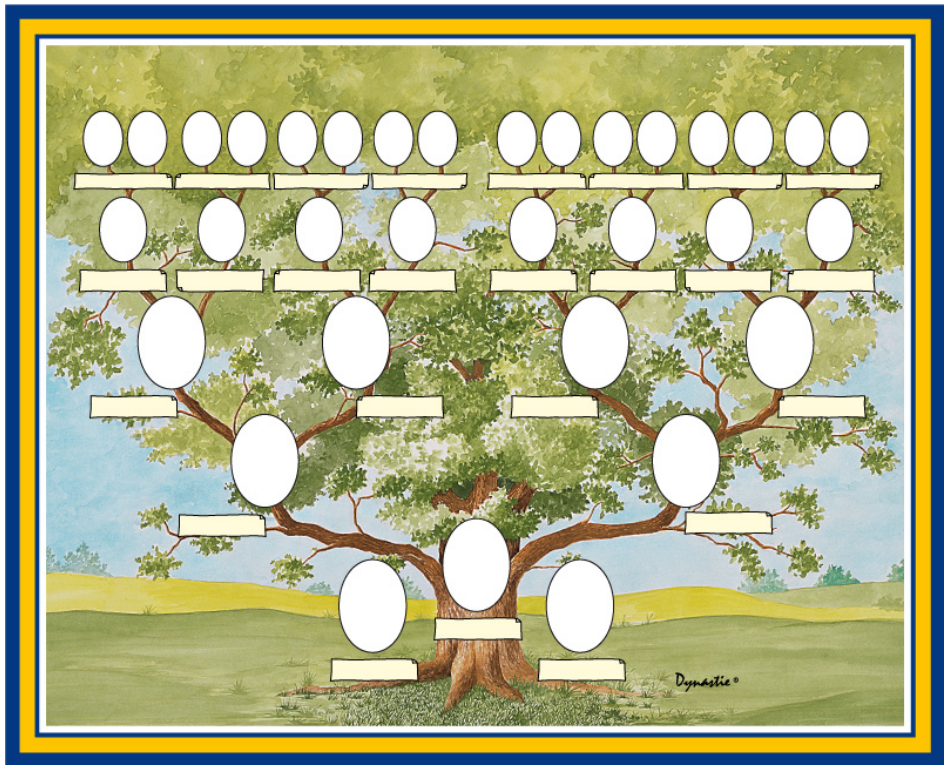
Introduction : genetic diversity, relatedness, genealogy

- Sample *n* individuals from a population and sequence their DNA.
- **Q. What diversity do you expect to observe in this sample of DNA sequences ?**
- Mutations : mutation rate ↑, diversity ↑
- Genealogy : relatedness ↑, diversity ↓
- Each locus (gene, site) is inherited from one single parent : simple gene genealogy
- Half of our genes come from our mother and the other half from our father

  ⇒ Different loci have different genealogies



A multiple sequence alignment



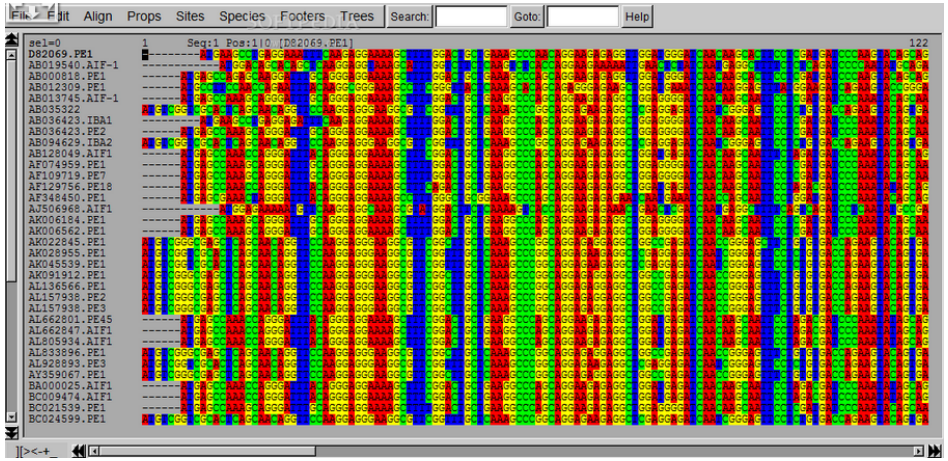
A pedigree = 2 parents (≠ genealogy = 1 parent)

Introduction : genetic diversity, relatedness, genealogy

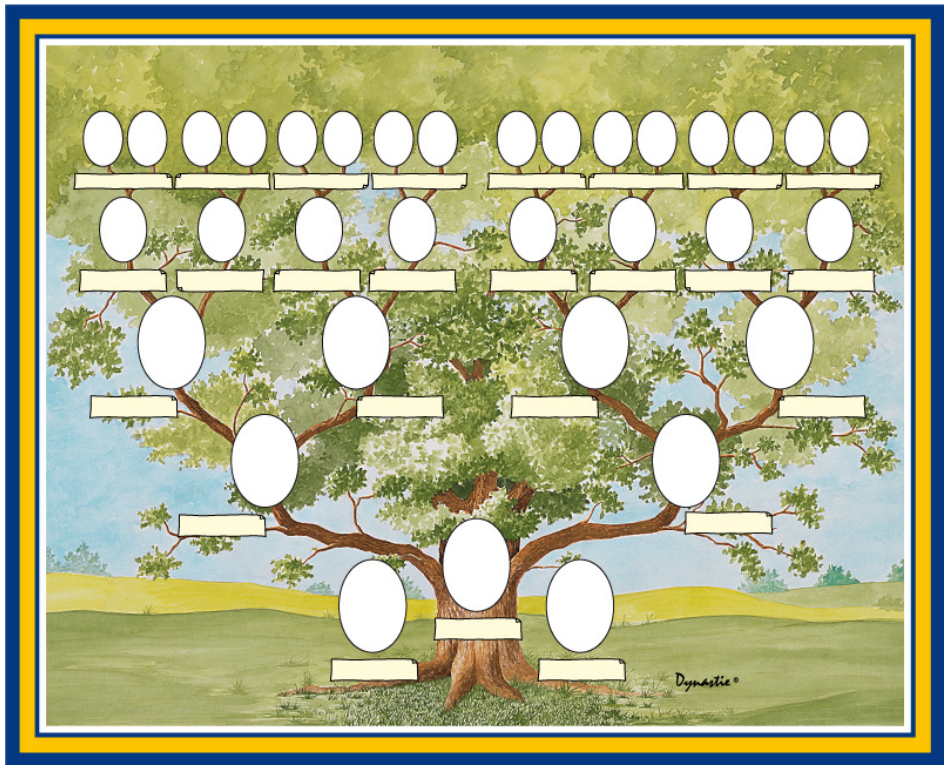
- Sample *n* individuals from a population and sequence their DNA.
- **Q. What diversity do you expect to observe in this sample of DNA sequences ?**
- Mutations : mutation rate ↑, diversity ↑
- Genealogy : relatedness ↑, diversity ↓
- Each locus (gene, site) is inherited from one single parent : simple gene genealogy
- Half of our genes come from our mother and the other half from our father

  ⇒ Different loci have different genealogies

  - In diploid species, each chromosome is in two copies in each cell



A multiple sequence alignment

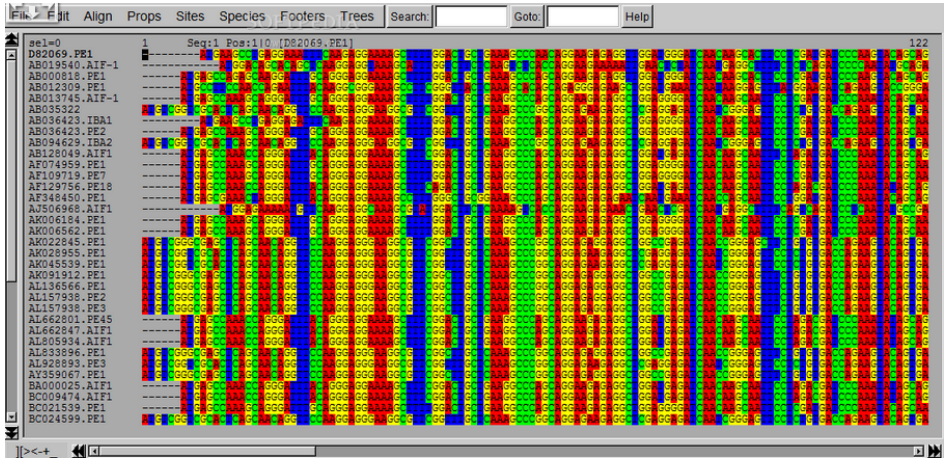


A pedigree = 2 parents (≠ genealogy = 1 parent)

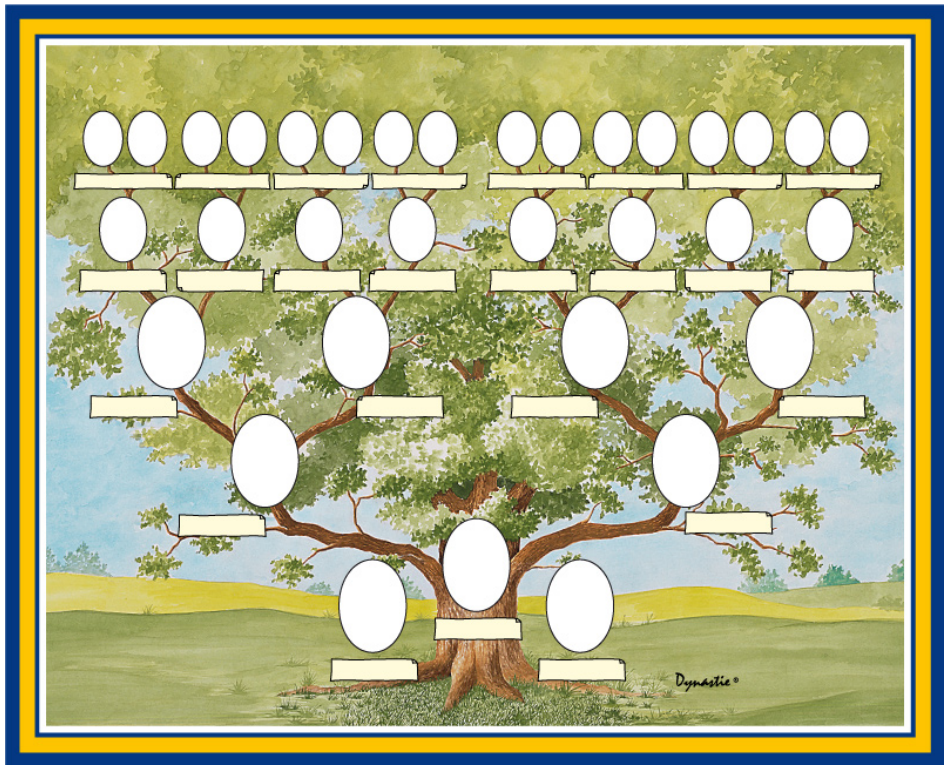
Introduction : genetic diversity, relatedness, genealogy

- Sample *n* individuals from a population and sequence their DNA.
- **Q. What diversity do you expect to observe in this sample of DNA sequences ?**
- Mutations : mutation rate ↑, diversity ↑
- Genealogy : relatedness ↑, diversity ↓
- Each locus (gene, site) is inherited from one single parent : simple gene genealogy
- Half of our genes come from our mother and the other half from our father

  ⇒ Different loci have different genealogies
  - In diploid species, each chromosome is in two copies in each cell
  - Each parent contributes one copy of each chromosome : the 2 copies have ≠ ancestries



A multiple sequence alignment

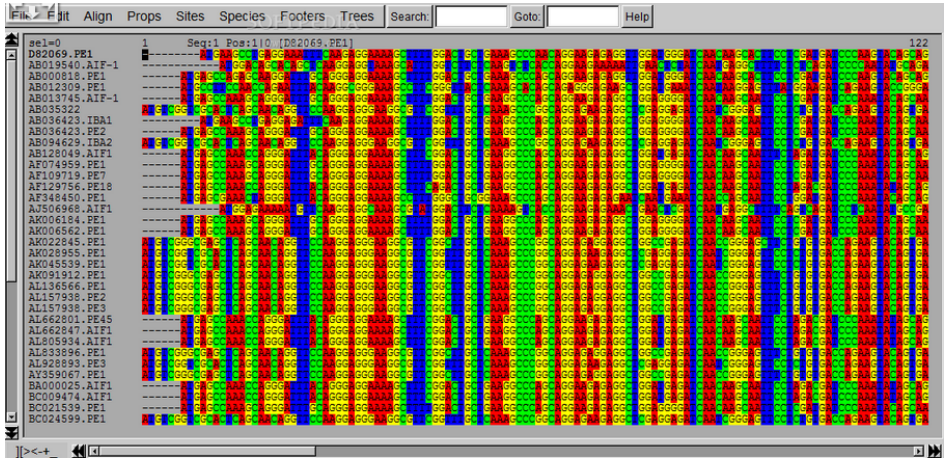


A pedigree = 2 parents (≠ genealogy = 1 parent)

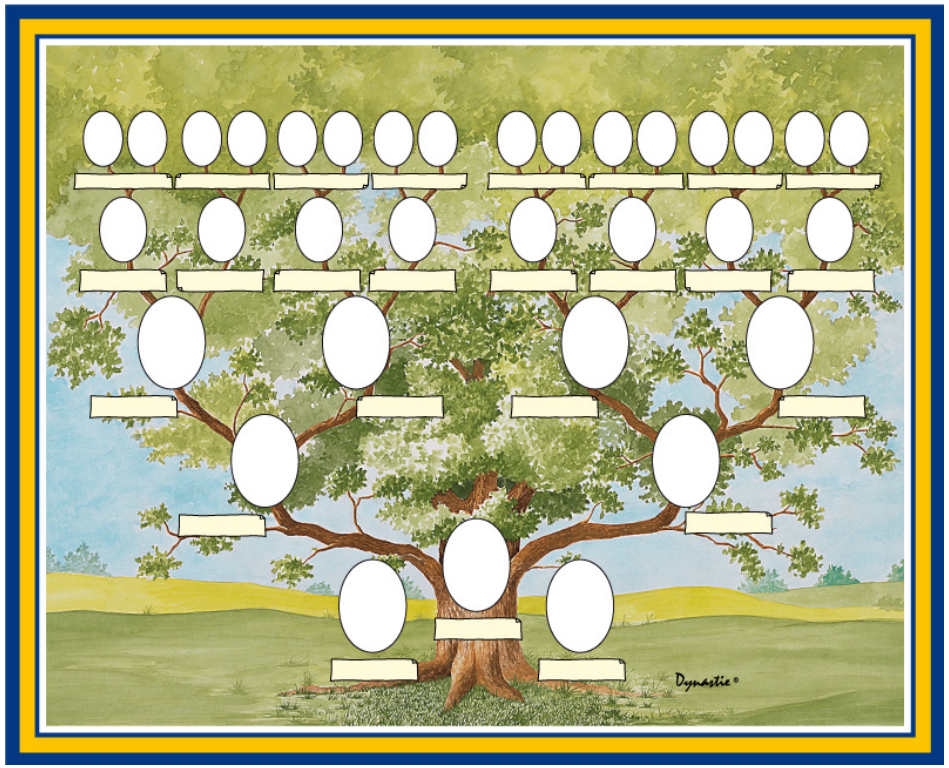
Introduction : genetic diversity, relatedness, genealogy

- Sample *n* individuals from a population and sequence their DNA.
- **Q. What diversity do you expect to observe in this sample of DNA sequences ?**
- Mutations : mutation rate ↑, diversity ↑
- Genealogy : relatedness ↑, diversity ↓
- Each locus (gene, site) is inherited from one single parent : simple gene genealogy
- Half of our genes come from our mother and the other half from our father

  ⇒ Different loci have different genealogies
  - In diploid species, each chromosome is in two copies in each cell
  - Each parent contributes one copy of each chromosome : the 2 copies have ≠ ancestries
  - Even loci located on same chr don't have the same genealogy ⟵ recombination



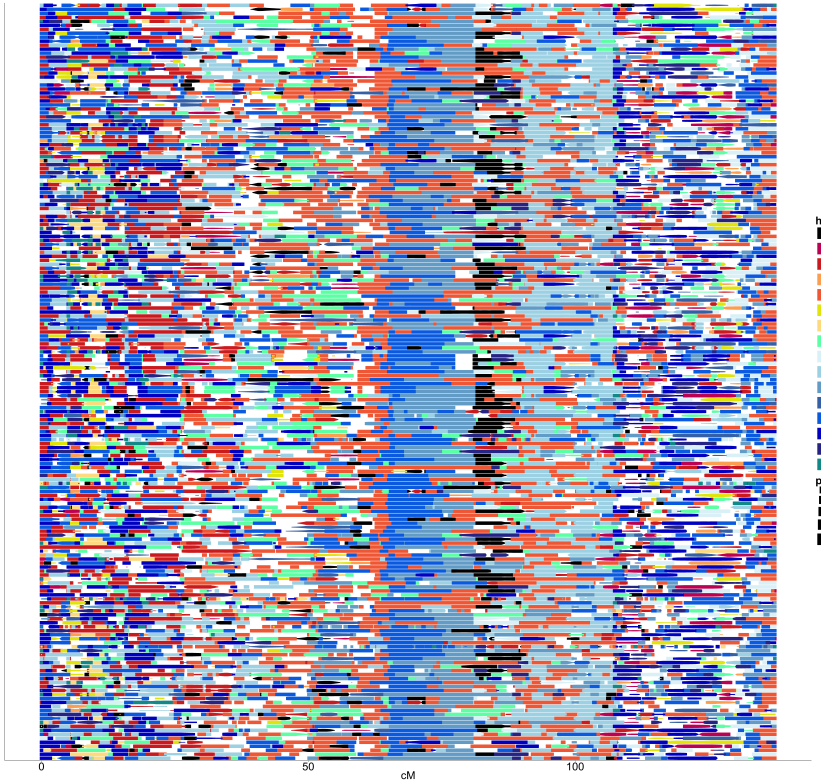
A multiple sequence alignment



A pedigree = 2 parents (≠ genealogy = 1 parent)

Outline

- Models for the genealogy of one gene, coalescent theory
- Patterns of genetic diversity at one locus, relation to population size
- Models coupling genealogies of several genes
- Applications
  - **Q1.** How does genome-wide diversity inform us on the past demography ?
  - **Q2.** If the genome of each ancestor was painted in a different color, how would the mosaic of colors in the pop look like in the long run ?



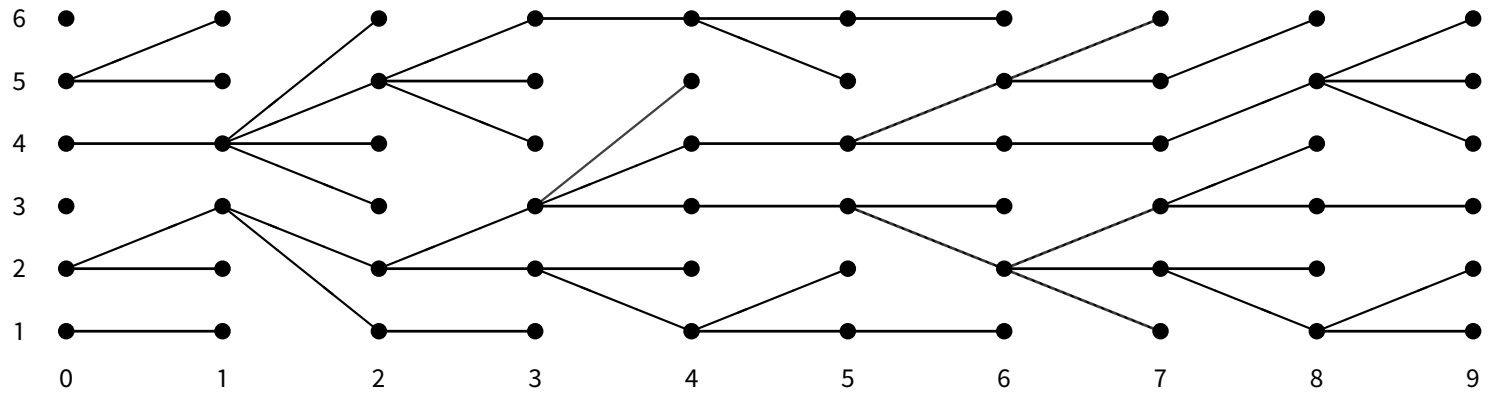
Experimental evolution with C. elegans
16 colors, $n = 300$, $N = 10^4$
Teotónio, Estes, Phillips & Baer (2017)

Outline

Neutral models of population genetics

Wright, Fisher, Cannings...

- The size of the population is constant, fixed equal to $N \gg 1$.
- Individual i of generation t has $\nu_i^{(t)}$ children $\in$ generation $t+1$



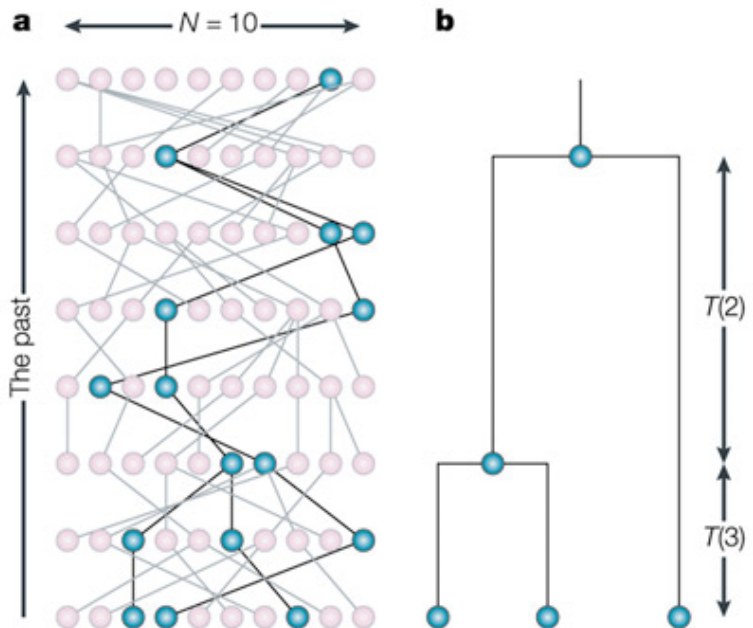
- **Cannings model(s)** : The vectors $(\nu_1^{(t)}, \nu_2^{(t)}, \ldots, \nu_N^{(t)})_{t\in\mathbb{Z}}$ are independent copies of a vector $(\nu_1, \nu_2, \ldots, \nu_N)$ such that
  - $\sum_{i=1}^N \nu_i = N$
  - The law of $(\nu_1, \nu_2, \ldots, \nu_N)$ is invariant by permutation (exchangeable)
- **Wright–Fisher model** : $(\nu_1, \ldots, \nu_N)$ is multinomial with parameters $(N; 1/N, \ldots, 1/N)$ $\Leftrightarrow$ Each ind in generation $t+1$ **picks her parent uniformly and indep'ly** in generation t

The coalescent for two individuals

- Sample 2 individuals uniformly at random and follow their ancestors backwards in time
- Let $T_N(2)$ be the number of generations counted backwards until the two lineages find their **most recent common ancestor (MRCA)**
- Then $T_N(2)$ is geometric with success probability $c_N := \mathbb{P}(\text{ 2 random ind are sisters })$.
- In the Wright–Fisher model, $c_N = 1/N$, otherwise

$$c_N = \mathbb{E}\left(\sum_{i=1}^{N} \frac{\nu_i(\nu_i - 1)}{N(N-1)}\right) = \frac{\mathbb{E}\left(\nu_1(\nu_1 - 1)\right)}{N-1}$$

- If $c_N \to 0$ as $N \to \infty$, then $T_N(2) = O(1/c_N)$ and $c_N T_N(2) \to T(2) \sim \mathcal{E}(1)$.
- Wright-Fisher : $T_N(2) = O(N)$ and $T_N(2)/N \to \mathcal{E}(1)$.



The coalescent for n individuals

Kingman, Griffiths, Möhle...

- Sample n individuals uniformly at random and follow their ancestors backwards in time
- Recall $c_N = \mathbb{P}(\text{ 2 random ind are sisters })$ and set $d_N := \mathbb{P}(\text{ 3 random ind are sisters })$, so that $d_N = 1/N^2$ in the WF model.

Theorem (Möhle's lemma)

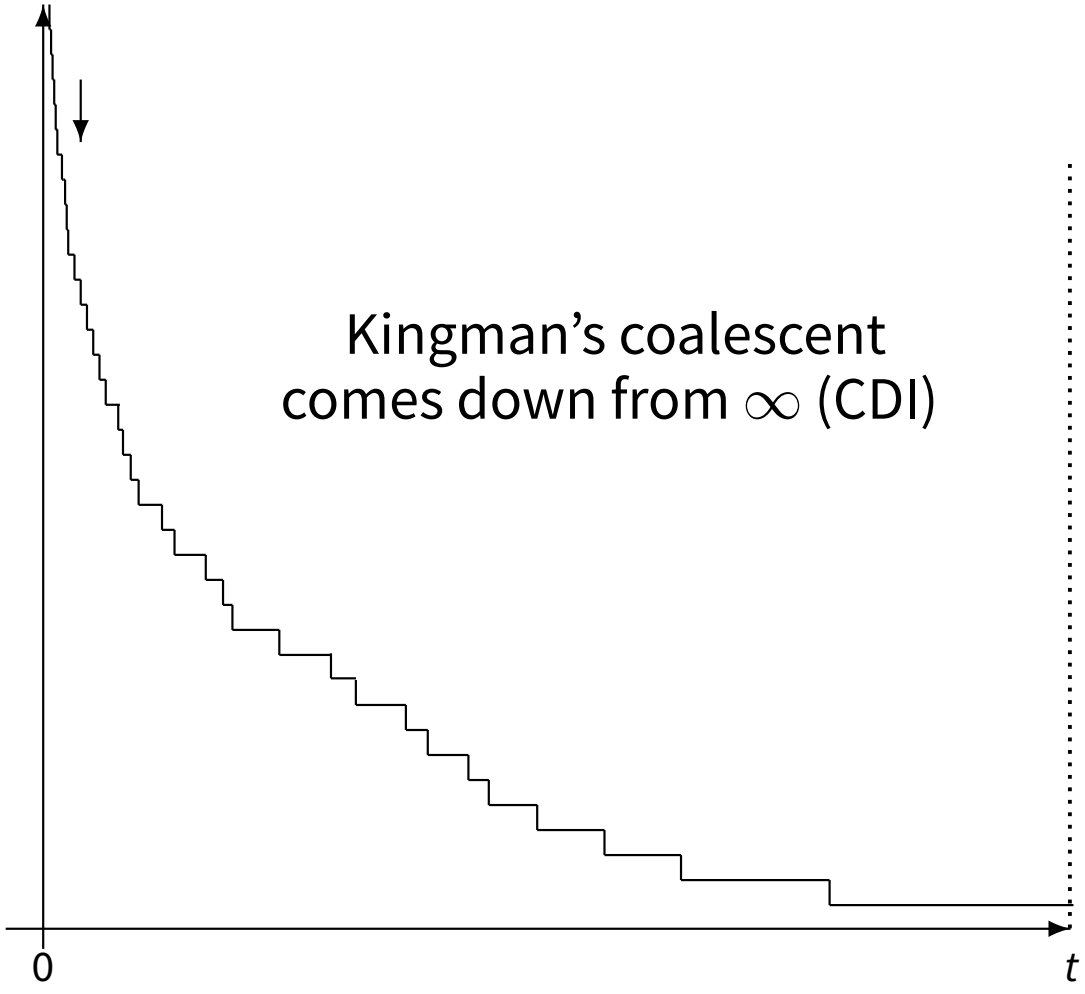
As $N \to \infty$, under the assumption that $d_N/c_N \to 0$, the genealogy of the sample t/c_N units of time ago, converges to **Kingman's coalescent** *:*

1. *The waiting time $T(k)$ from k to $k-1$ lineages is exponential with parameter $\binom{k}{2}$*
2. *The next coalescing pair is chosen uniformly at random.*

- $\Leftrightarrow$ "Each pair of lineages coalesces independently at rate 1"...
- ...Or at rate $1/x(t)$, if pop size $= Nx(Nt)$
- No multiple mergers. Shorter edge lengths close to present. Sampling-consistent.
- The genealogy of n also has length $= O(N)$.

Note : The coalescent at time t is represented by the partition of $\{1, \ldots, n\}$ induced by the relation $i \sim_t j$ if i and j have found their common ancestor t time units ago

Large sample limit $1 \ll n \ll N$



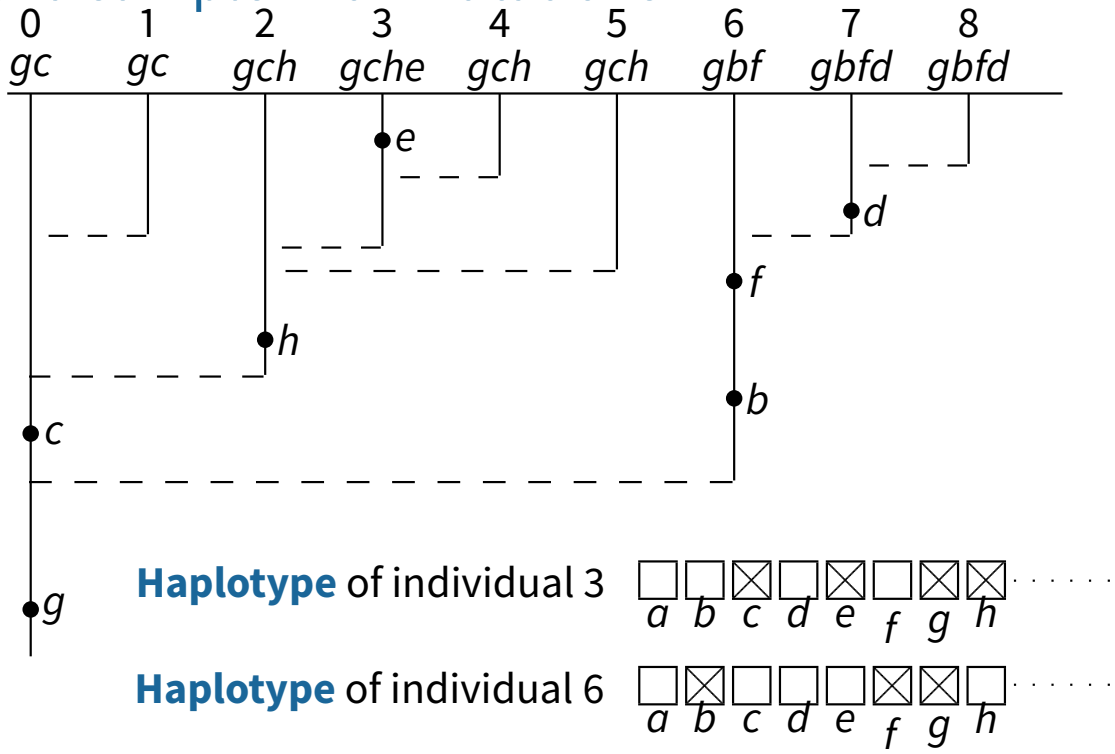
- The process counting the number of lineages in Kingman's coalescent is a pure-death process going from k to $k-1$ at rate $\binom{k}{2}$
- The sojourn time T_k in state k has expectation $\mathbb{E}(T_k) = \binom{k}{2}^{-1}$ so

$$\mathbb{E}\left(\sum_{k\geq 2} T_k\right) = \sum_{k\geq 2} \mathbb{E}(T_k) < \infty,$$

  so $\sum_{k\geq 2} T_k < \infty$ with probability 1.
- There is a unique entrance law $\mathbb{P}_\infty =:$ standard coalescent.

Outline

Genealogy of a sample with mutations



- Mutations are visible if present in $k \in \{1, \ldots, n-1\} =:$ **polymorphic/segregating site**
  $\Rightarrow$ Visible mutations occur within the genealogical tree, with length $= O(N)$
- Goldilocks zone for **proba u_N of gene-wide mutation** at birth :
  - If $Nu_N \ll 1$, no segregating site in sample
  - If $Nu_N \gg 1$, infinitely many segregating sites in sample
  - If $Nu_N = O(1)$, finite number of mutations (Poisson cond on tree length)
- If $Nu_N = O(1)$ and sequence long enough : mutations all occur at different sites
  $=$ infinitely-many-site model

  $\Rightarrow$ Each mutation gives rise to a new haplotype $=$ infinitely-many-allele model

Assumptions and notation

- Population size N, **mutation proba** $u_N \sim \theta/2N$ (recall Goldilocks zone : $Nu_N = O(1)$)

- As $N \to \infty$, convergence to **Kingman's coalescent** with **Poissonian marks rate** $\theta/2$

- $S_n := \#$ polymorphic **sites** $= \sum_k S_n(k)$, where
- $S_n(k) := \#$ polymorphic **sites carried by** k ind (in sample of n)
  $=:$ Site Frequency Spectrum (SFS), $1 \leq k \leq n-1$.

  Note : Conditional on total tree length L_n, S_n is Poisson with parameter $\theta L_n/2$.

- $A_n := \#$ distinct **haplotypes** $= \sum_k A_n(k)$, where
- $A_n(k) := \#$ **haplotypes carried by** k ind (in sample of n)
  $=:$ Allele Frequency Spectrum (AFS), $1 \leq k \leq n$.

  Note : Haplotypes induce the so-called allelic partition of the sample, so

$$\sum_k kA_n(k) = n$$

Law of the allelic partition : Ewens' Sampling Formula

Time reversal argument :

Coalescent (pairwise rate 1) w deaths (rate $\theta/2$)

$\Leftrightarrow$ Birth process (rate 1) with immigration (rate θ)

$\Leftrightarrow$ Chinese restaurant process
$=$ when $(k+1)$-st customer enters the dining room,

- She sits next to customer i with proba $1/(k+\theta)$,
- Or she sits at an empty table with proba $\theta/(k+\theta)$.

Theorem (Ewens 1972)

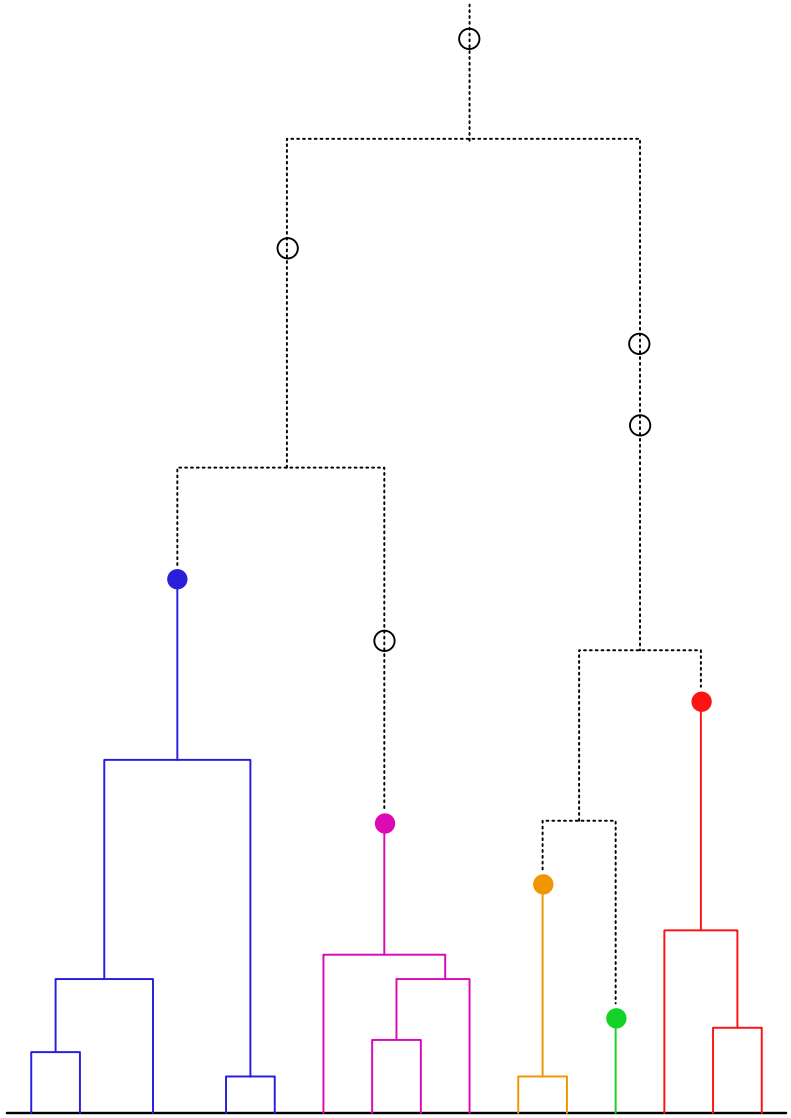
For any vector $(a_1, \ldots, a_n)$ *st* $\sum_{k=1}^n k a_k = n$,

$$\mathbb{P}(A_n(1) = a_1, \ldots, A_n(n) = a_n) = c_{\theta,n} \prod_{k=1}^n \frac{\left(\frac{\theta}{k}\right)^{a_k}}{a_k!}$$

where $c_{\theta,n} := n! \,/\, [\theta(\theta+1)\cdots(\theta+n-1)]$.

$\Leftrightarrow (A_n(1), \ldots, A_n(n)) \overset{(d)}{=} (Y_1, \ldots, Y_n \mid \sum_{k=1}^n k Y_k = n)$:

- Y_k*'s are independent*
- Y_k *is a* **Poisson r.v. with parameter** θ/k.



$n-$coalescent with Poissonian mutations,
each sampled haplotype has its own color

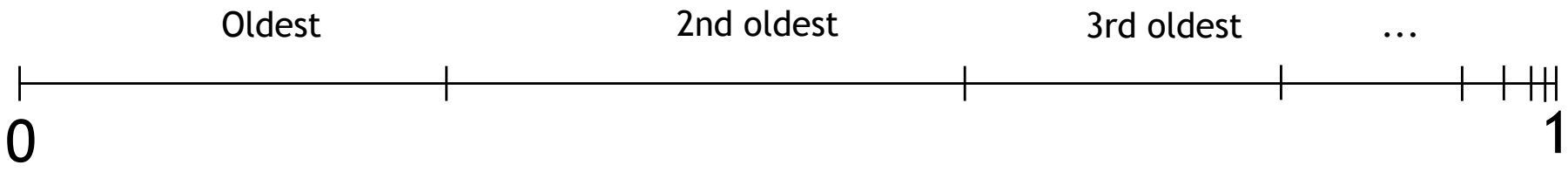
Large sample limit $1 \ll n \ll N$

Ewens, Donnelly & Tavaré...

- As $n \to \infty$,
$$S_n \sim \theta \ln(n) \quad \text{and} \quad A_n \sim \theta \ln(n),$$
with convergence rate $\sqrt{\ln(n)}$.

- Small families (fixed k).
$$\lim_{n\to\infty} A_n(k) \overset{(d)}{=} Y_k,$$
where Y_k denotes a Poisson r.v. with parameter θ/k.

- Large families.
Set $X_n(i) :=$ size of i-th oldest family.
As $n \to \infty$, $\left(n^{-1}X_n(k)\right)_{k\geq 1}$ converges (fdd) to the GEM vector $(P_k)_{k\geq 1}$ defined as
$$P_k := Z_k \prod_{i=1}^{k-1}(1 - Z_i),$$
where the (Z_i) are i.i.d. with density $\theta(1-z)^{\theta-1}$ (Beta $(1,\theta)$).



Estimating population size

- Pop size $\uparrow$, relatedness $\downarrow$, diversity $\uparrow$

Estimating population size

- Pop size $\uparrow$, relatedness $\downarrow$, diversity $\uparrow$

- If mutation proba u_N known, then any estimator of $\theta = 2Nu_N$ yields an estimate of N. For example θ can be estimated by $S_n / \ln(n)$ (Watterson 1975)

Estimating population size

- Pop size $\uparrow$, relatedness $\downarrow$, diversity $\uparrow$

- If mutation proba u_N known, then any estimator of $\theta = 2Nu_N$ yields an estimate of N.
  For example θ can be estimated by $S_n / \ln(n)$ (Watterson 1975)

- Genealogy (predicts)(can be inferred from) genetic diversity.
  For $n = 2$, $\mathbb{P}(\text{ identity }) := \text{Homozygosity } h = (1 - u_N)^{2T_N(2)} \approx \exp(-\theta T(2)))$

Estimating population size

- Pop size $\uparrow$, relatedness $\downarrow$, diversity $\uparrow$

- If mutation proba u_N known, then any estimator of $\theta = 2Nu_N$ yields an estimate of N.
  For example θ can be estimated by $S_n / \ln(n)$ (Watterson 1975)

- Genealogy (predicts)(can be inferred from) genetic diversity.
  For $n = 2$, $\mathbb{P}(\text{ identity }) := \text{Homozygosity } h = (1 - u_N)^{2T_N(2)} \approx \exp(-\theta T(2)))$

- Assume N constant. If the **value of** $T_N(2)$ can be estimated from diversity,
  then $1/T_N(2) =$ **estimator of** N

Estimating population size

- Pop size $\uparrow$, relatedness $\downarrow$, diversity $\uparrow$

- If mutation proba u_N known, then any estimator of $\theta = 2Nu_N$ yields an estimate of N.
  For example θ can be estimated by $S_n / \ln(n)$ (Watterson 1975)

- Genealogy (predicts)(can be inferred from) genetic diversity.
  For $n = 2$, $\mathbb{P}(\text{ identity }) := \text{Homozygosity } h = (1 - u_N)^{2T_N(2)} \approx \exp(-\theta T(2)))$

- Assume N constant. If the **value of** $T_N(2)$ can be estimated from diversity,
  then $1/T_N(2) =$ **estimator of** N

- If pop size not constant, $N(t) = Nx(Nt)$, then

$$\mathbb{P}(T_N(2)/N > t) \longrightarrow \exp\left\{-\int_0^t \frac{ds}{x(s)}\right\}$$

Estimating population size

- Pop size $\uparrow$, relatedness $\downarrow$, diversity $\uparrow$

- If mutation proba u_N known, then any estimator of $\theta = 2Nu_N$ yields an estimate of N. For example θ can be estimated by $S_n / \ln(n)$ (Watterson 1975)

- Genealogy (predicts)(can be inferred from) genetic diversity.
  For $n = 2$, $\mathbb{P}(\text{ identity }) := \text{Homozygosity } h = (1 - u_N)^{2T_N(2)} \approx \exp(-\theta T(2)))$

- Assume N constant. If the **value of** $T_N(2)$ can be estimated from diversity, then $1/T_N(2) =$ **estimator of** N

- If pop size not constant, $N(t) = Nx(Nt)$, then

$$\mathbb{P}(T_N(2)/N > t) \longrightarrow \exp\left\{ - \int_0^t \frac{ds}{x(s)} \right\}$$

- Recall that different genes have different genealogies — in theory, if the **distribution of** $T_N(2)$ could be estimated from diversity at many 'independent' loci, the **variations of** N **through time** could be inferred!!

Estimating population size

- Pop size $\uparrow$, relatedness $\downarrow$, diversity $\uparrow$

- If mutation proba u_N known, then any estimator of $\theta = 2Nu_N$ yields an estimate of N. For example θ can be estimated by $S_n / \ln(n)$ (Watterson 1975)

- Genealogy (predicts)(can be inferred from) genetic diversity. For $n = 2$, $\mathbb{P}(\text{ identity }) := \text{Homozygosity } h = (1 - u_N)^{2T_N(2)} \approx \exp(-\theta T(2)))$

- Assume N constant. If the **value of** $T_N(2)$ can be estimated from diversity, then $1/T_N(2) =$ **estimator of** N

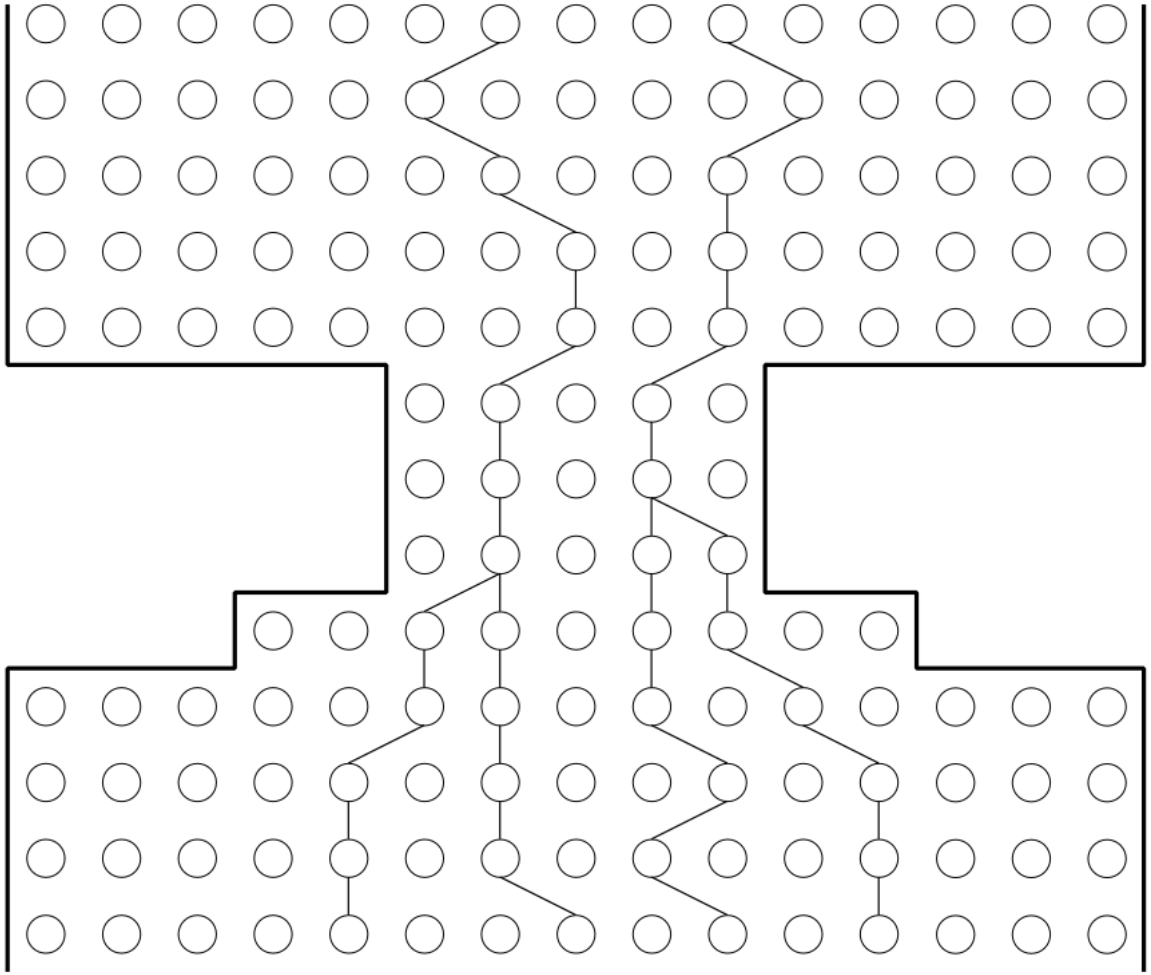
- If pop size not constant, $N(t) = Nx(Nt)$, then

$$\mathbb{P}(T_N(2)/N > t) \longrightarrow \exp\left\{-\int_0^t \frac{ds}{x(s)}\right\}$$

- Recall that different genes have different genealogies — in theory, if the **distribution of** $T_N(2)$ could be estimated from diversity at many 'independent' loci, the **variations of** N **through time** could be inferred!!

- Requires understanding how genealogies of different genes are coupled...

The example of a bottleneck



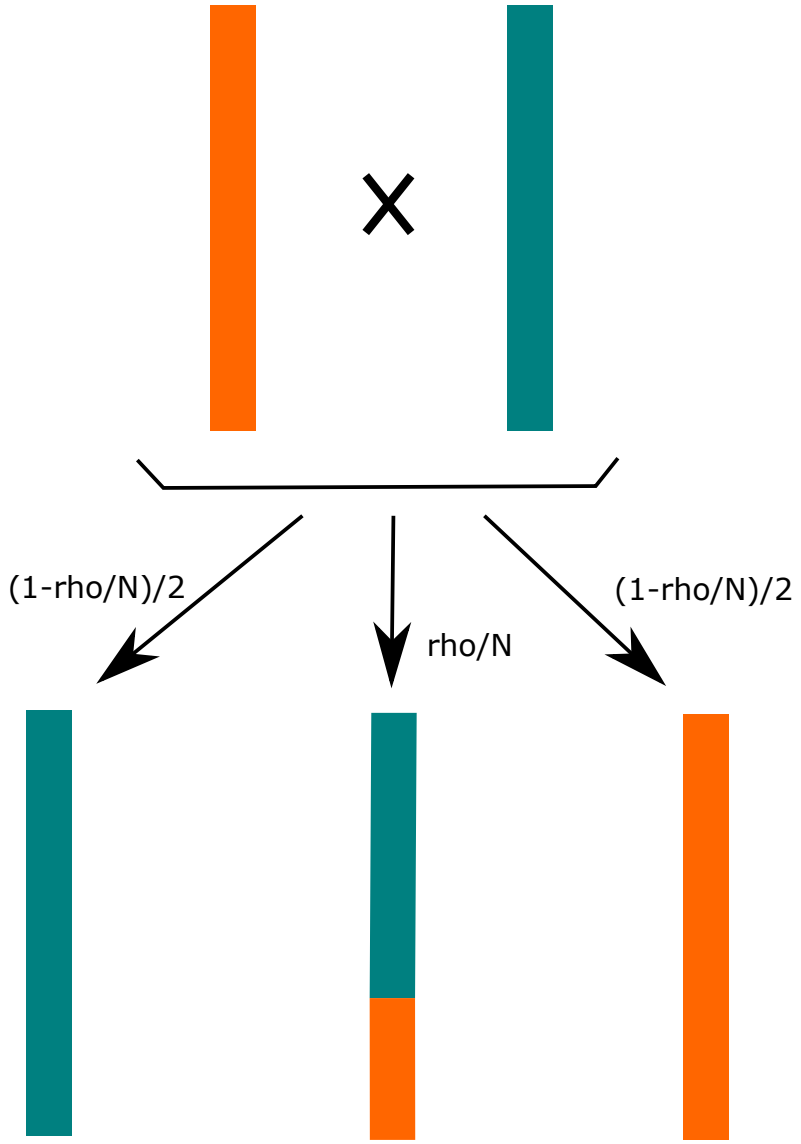
A provisional reduction in population size, or bottleneck

Densities of coalescence times peak at a bottleneck time

Outline

Wright–Fisher model with recombination

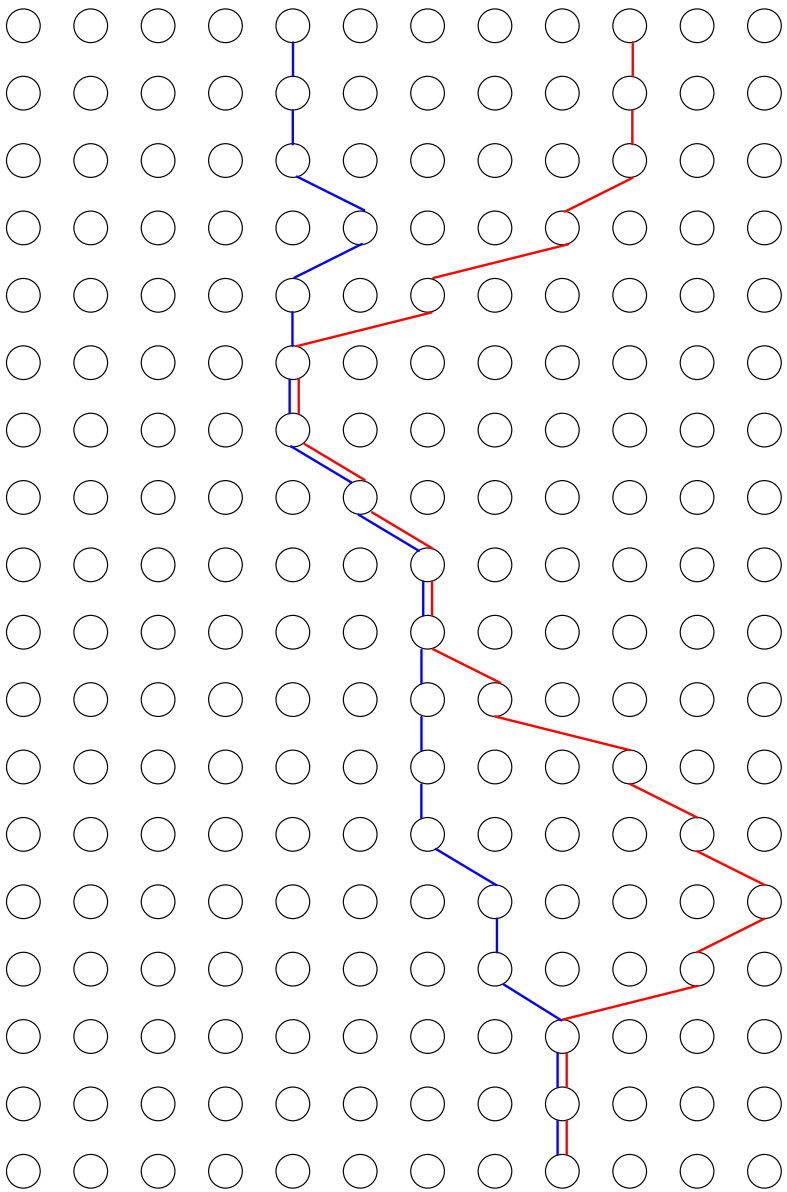
- Constant pop size N, but now : **TWO parents per ind**
- Each ind carries one chromosome $=$ interval $[0, 1]$
- At each generation, each individual chooses her two parents uniformly at random
- The two **parental chromosomes recombine with probability** ρ/N
- ...as a single, uniformly distributed cross-over
- Otherwise, only one of the two chromosomes is passed on.



Ancestral Recombination Graph : 2 sites, $n = 1$

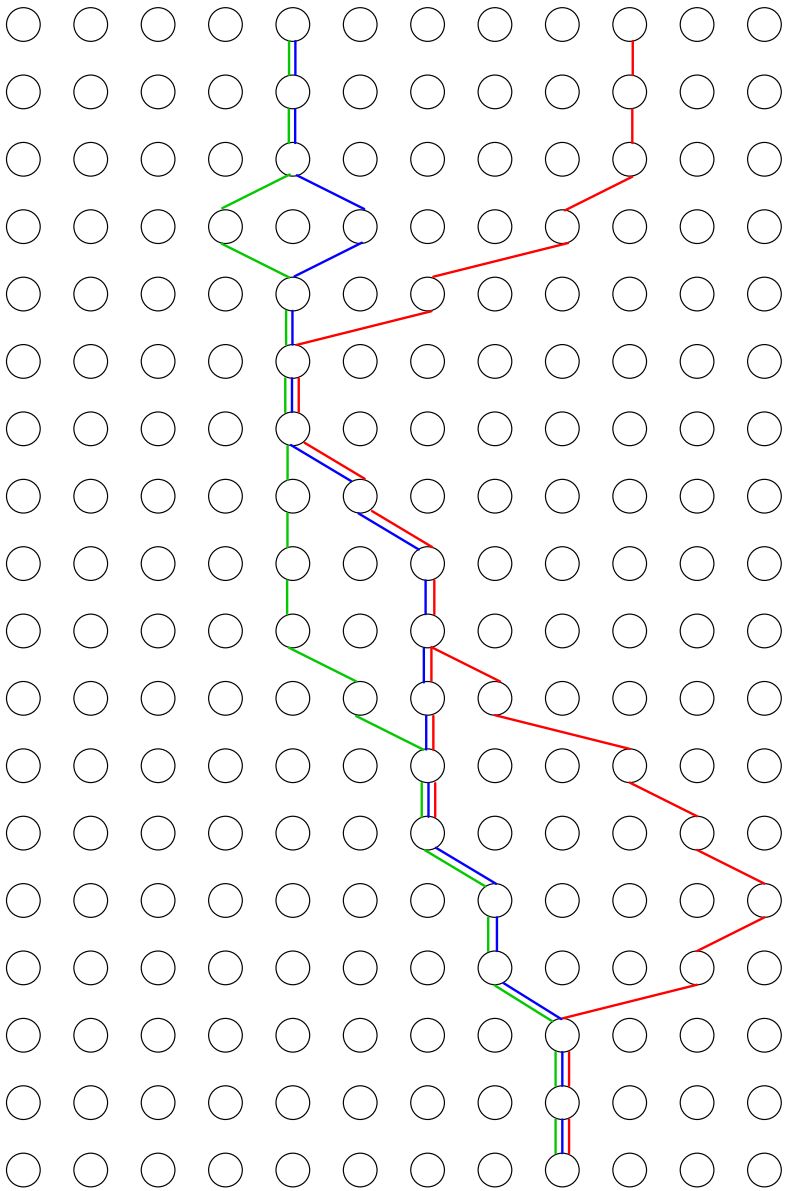
Grifiths & Marjoram, Wiuf & Hein, Jenkins & Song...

- **Sample $n = 1$ individual**
- Consider two sites x and y at distance ℓ and follow their ancestry as time goes backward
- At each generation, the common line of descent $\{x, y\}$ **splits with probability** $\rho\ell/N$
- At each generation, the singleton lines $\{x\}$ and $\{y\}$ coalesce with probability $1/N$
- As $N \to \infty$, the time-rescaled ARG **splits at rate** $\rho\ell$ **and merges at rate** 1.



The Ancestral Recombination Graph : 3 sites, $n = 1$

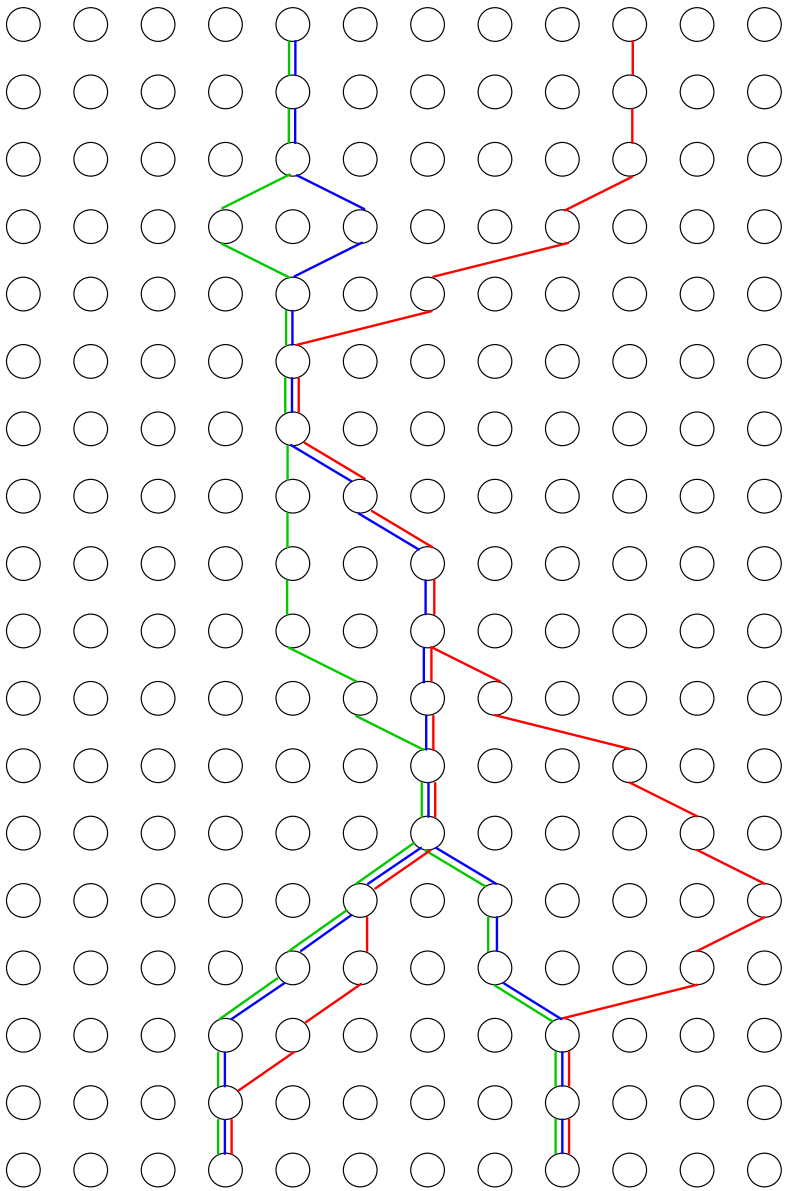
- Sample $n = 1$ individual
- Consider three sites $\{x, y, z\}$ at distances ℓ_1 and ℓ_2
- In the limit $N \to \infty$, the block $\{x, y, z\}$
  - splits into $\{x, y\}$ and $\{z\}$ at rate $\rho\ell_2$
  - splits into $\{x\}$ and $\{y, z\}$ at rate $\rho\ell_1$
- Block $\{x, y\}$..., block $\{y, z\}$..., block $\{x, z\}$...
- Each pair of lines coalesces at rate 1.



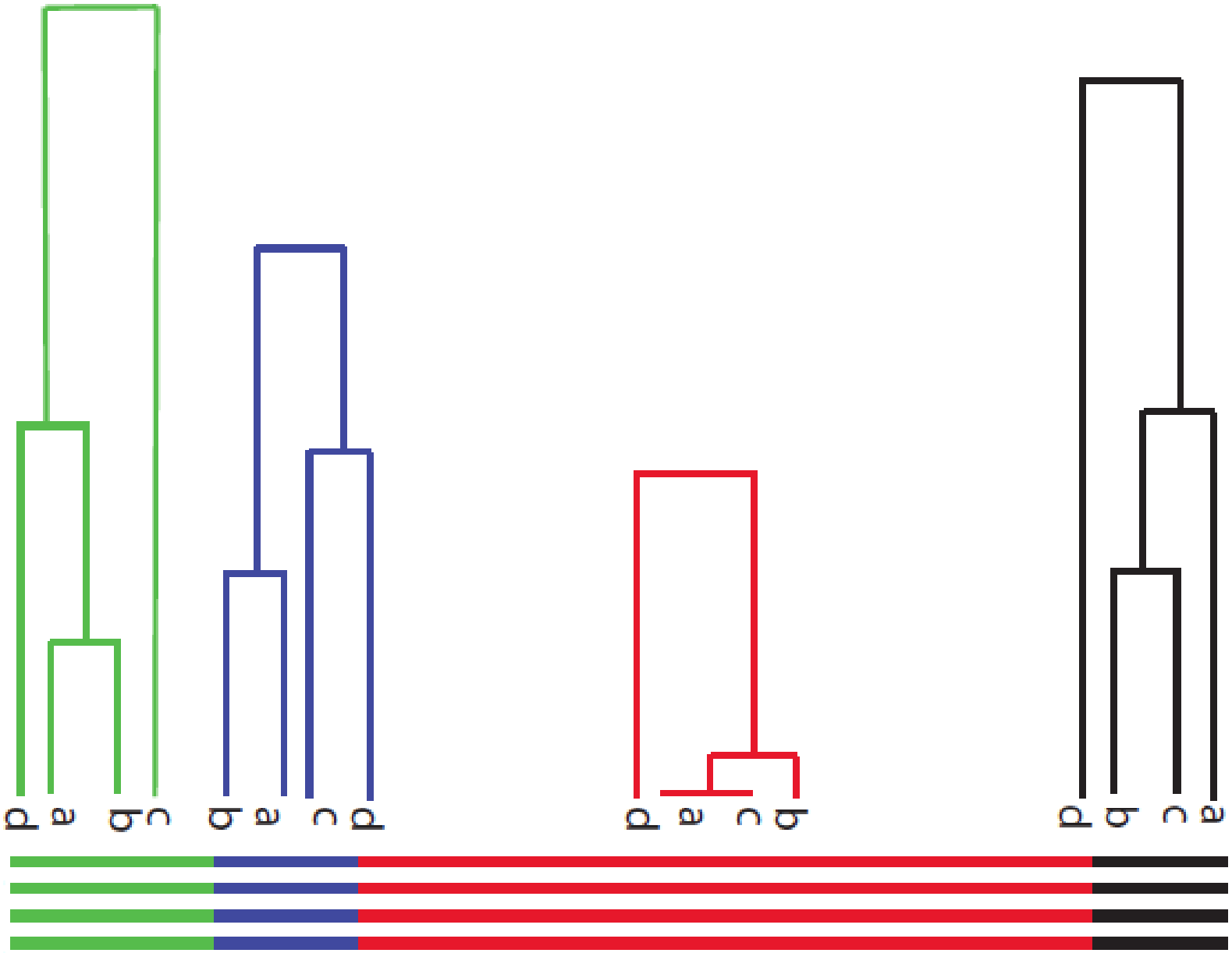
Note : When $n = 1$, the ARG on k loci can be generated by a Markov process valued in the partitions of $\{1, \ldots, k\}$.

The Ancestral Recombination Graph : 3 loci, $n = 2$

- Now sample $n = 2$ individuals
- Now same color-lines can additionally coalesce
- Observe that green and blue loci have the same time to MRCA, $\neq$ red locus.
- Moving along the chromosome, we see a **sequence of trees** ($n = 2$: a sequence of cherries)...
- **IBD segment** ("identical by descent") := maximal connected segment of sites sharing the same genealogy.



Tree sequence



picture by Guillaume Achaz

Shallow trees are carried by a longer IBD segment

Outline

Sequentially Markovian Coalescent (SMC)

McVean & Cardin, Li & Durbin, Schiffels & Durbin...

- ARG : complex dependencies betw gene trees

Sequentially Markovian Coalescent (SMC)

McVean & Cardin, Li & Durbin, Schiffels & Durbin...

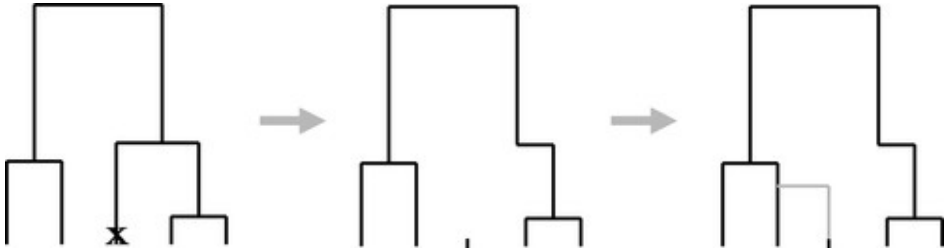
- ARG : complex dependencies betw gene trees

- SMC $:=$ Markovian approximation to the ARG,
  as we move along chromosome

Sequentially Markovian Coalescent (SMC)

McVean & Cardin, Li & Durbin, Schiffels & Durbin...

- ARG : complex dependencies betw gene trees
- SMC := Markovian approximation to the ARG, as we move along chromosome
  - Starting from gene tree with length L...

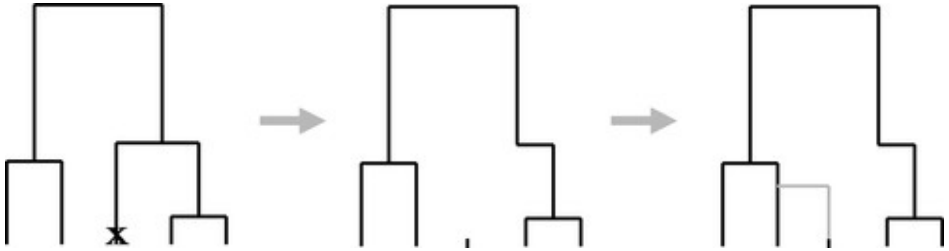


McVean & Cardin (2005)

Sequentially Markovian Coalescent (SMC)

McVean & Cardin, Li & Durbin, Schiffels & Durbin...

- ARG : complex dependencies betw gene trees
- SMC $:=$ Markovian approximation to the ARG, as we move along chromosome
  - Starting from gene tree with length L...
  - Wait an exponential 'distance' with param ρL...

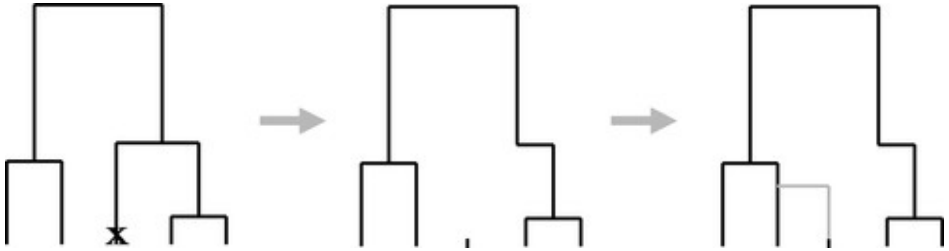


McVean & Cardin (2005)

Sequentially Markovian Coalescent (SMC)

McVean & Cardin, Li & Durbin, Schiffels & Durbin...

- ARG : complex dependencies betw gene trees
- SMC := Markovian approximation to the ARG, as we move along chromosome
  - Starting from gene tree with length L...
  - Wait an exponential 'distance' with param ρL...
  - Detach lineage at a uniform point and regraft it coalescent-like...

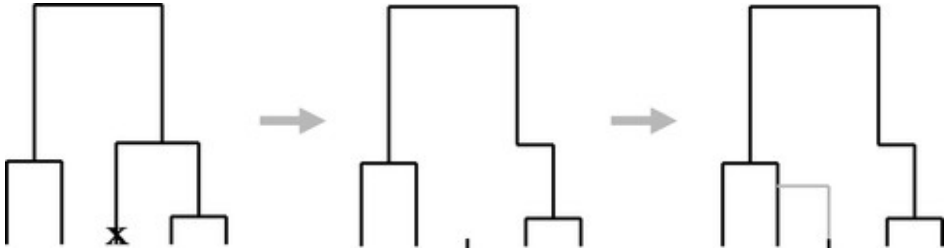


McVean & Cardin (2005)

Sequentially Markovian Coalescent (SMC)

McVean & Cardin, Li & Durbin, Schiffels & Durbin...

- ARG : complex dependencies betw gene trees
- SMC := Markovian approximation to the ARG, as we move along chromosome
  - Starting from gene tree with length L...
  - Wait an exponential 'distance' with param ρL...
  - Detach lineage at a uniform point and regraft it coalescent-like...
- Li & Durbin (*Nature* 2011) : the tour de force of inferring the (pop size) history of human pop by sequencing one individual = one diploid genome = TWO sequences

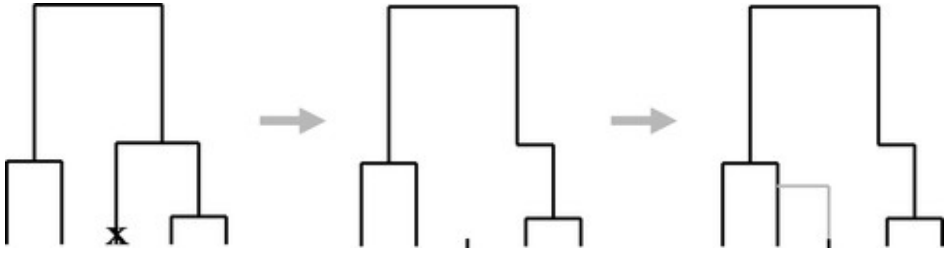


McVean & Cardin (2005)

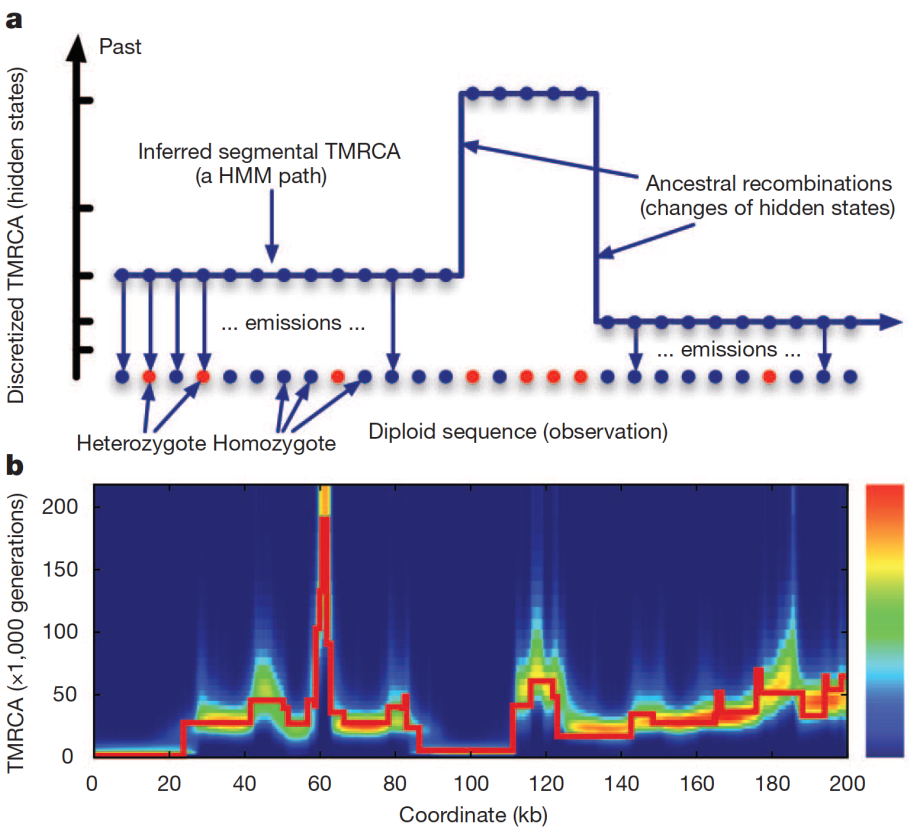
Sequentially Markovian Coalescent (SMC)

McVean & Cardin, Li & Durbin, Schiffels & Durbin...

- ARG : complex dependencies betw gene trees
- SMC := Markovian approximation to the ARG, as we move along chromosome
  - Starting from gene tree with length L...
  - Wait an exponential 'distance' with param ρL...
  - Detach lineage at a uniform point and regraft it coalescent-like...
- Li & Durbin (*Nature* 2011) : the tour de force of inferring the (pop size) history of human pop by sequencing one individual = one diploid genome = TWO sequences
- PSMC= Pairwise SMC:= infer past variations of population size from one diploid genome by HMM, where hidden state = T_{MRCA}



McVean & Cardin (2005)

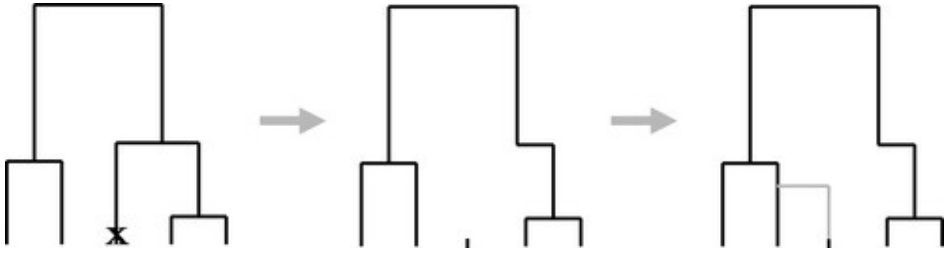


Li & Durbin (2011)

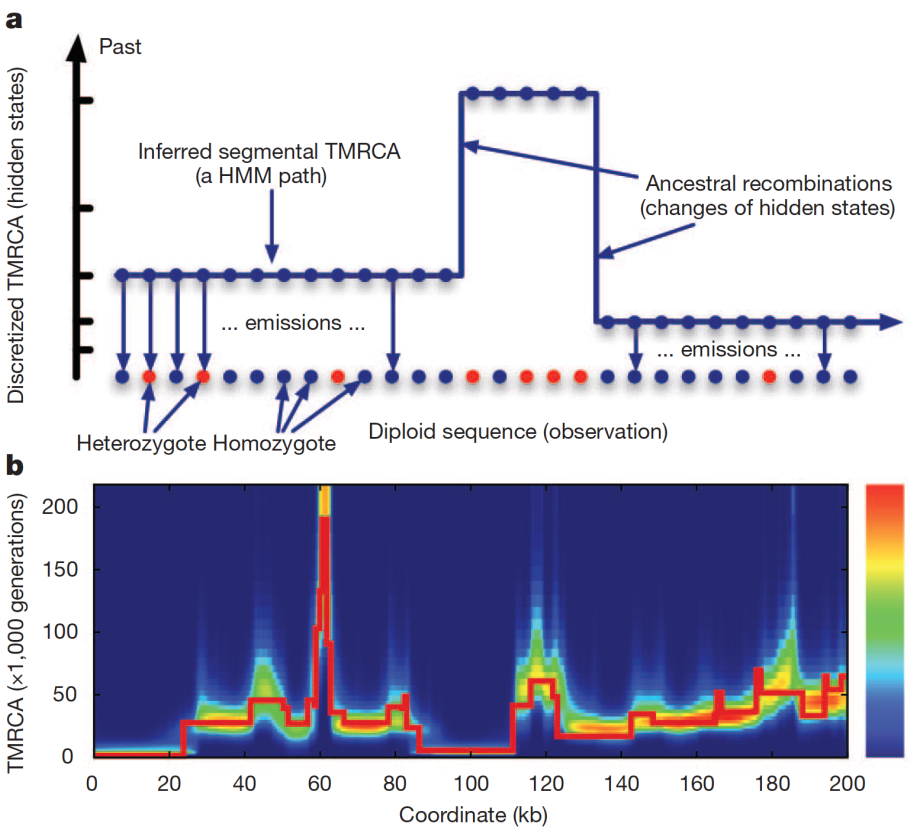
Sequentially Markovian Coalescent (SMC)

McVean & Cardin, Li & Durbin, Schiffels & Durbin...

- ARG : complex dependencies betw gene trees
- SMC := Markovian approximation to the ARG, as we move along chromosome
  - Starting from gene tree with length L...
  - Wait an exponential 'distance' with param ρL...
  - Detach lineage at a uniform point and regraft it coalescent-like...
- Li & Durbin (*Nature* 2011) : the tour de force of inferring the (pop size) history of human pop by sequencing one individual = one diploid genome = TWO sequences
- PSMC= Pairwise SMC:= infer past variations of population size from one diploid genome by HMM, where hidden state = T_{MRCA}
  - Shallow tree = Long segment, low density of heterozygous sites



McVean & Cardin (2005)

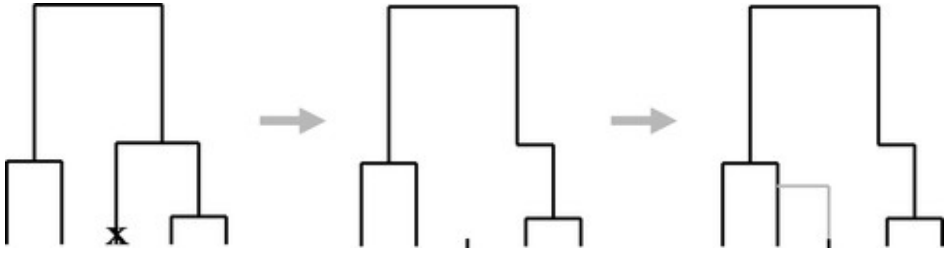


Li & Durbin (2011)

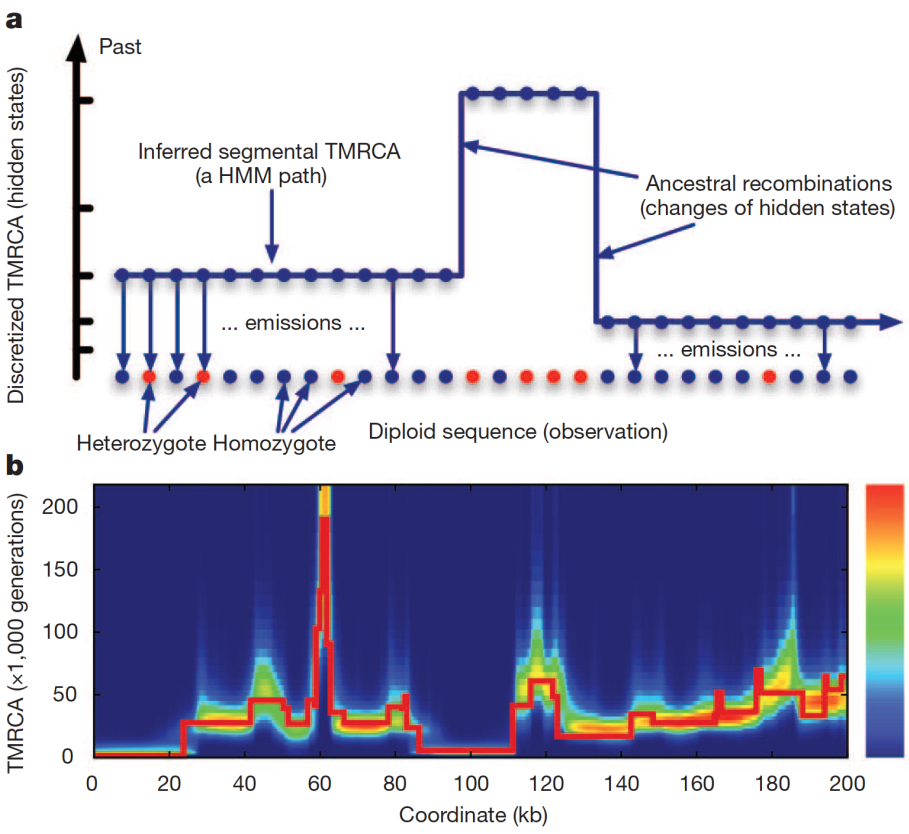
Sequentially Markovian Coalescent (SMC)

McVean & Cardin, Li & Durbin, Schiffels & Durbin...

- ARG : complex dependencies betw gene trees
- SMC := Markovian approximation to the ARG, as we move along chromosome
  - Starting from gene tree with length L...
  - Wait an exponential 'distance' with param ρL...
  - Detach lineage at a uniform point and regraft it coalescent-like...
- Li & Durbin (*Nature* 2011) : the tour de force of inferring the (pop size) history of human pop by sequencing one individual = one diploid genome = TWO sequences
- PSMC= Pairwise SMC:= infer past variations of population size from one diploid genome by HMM, where hidden state = T_{MRCA}
  - Shallow tree = Long segment, low density of heterozygous sites
  - Deep tree = Short segment, high density of heterozygous sites



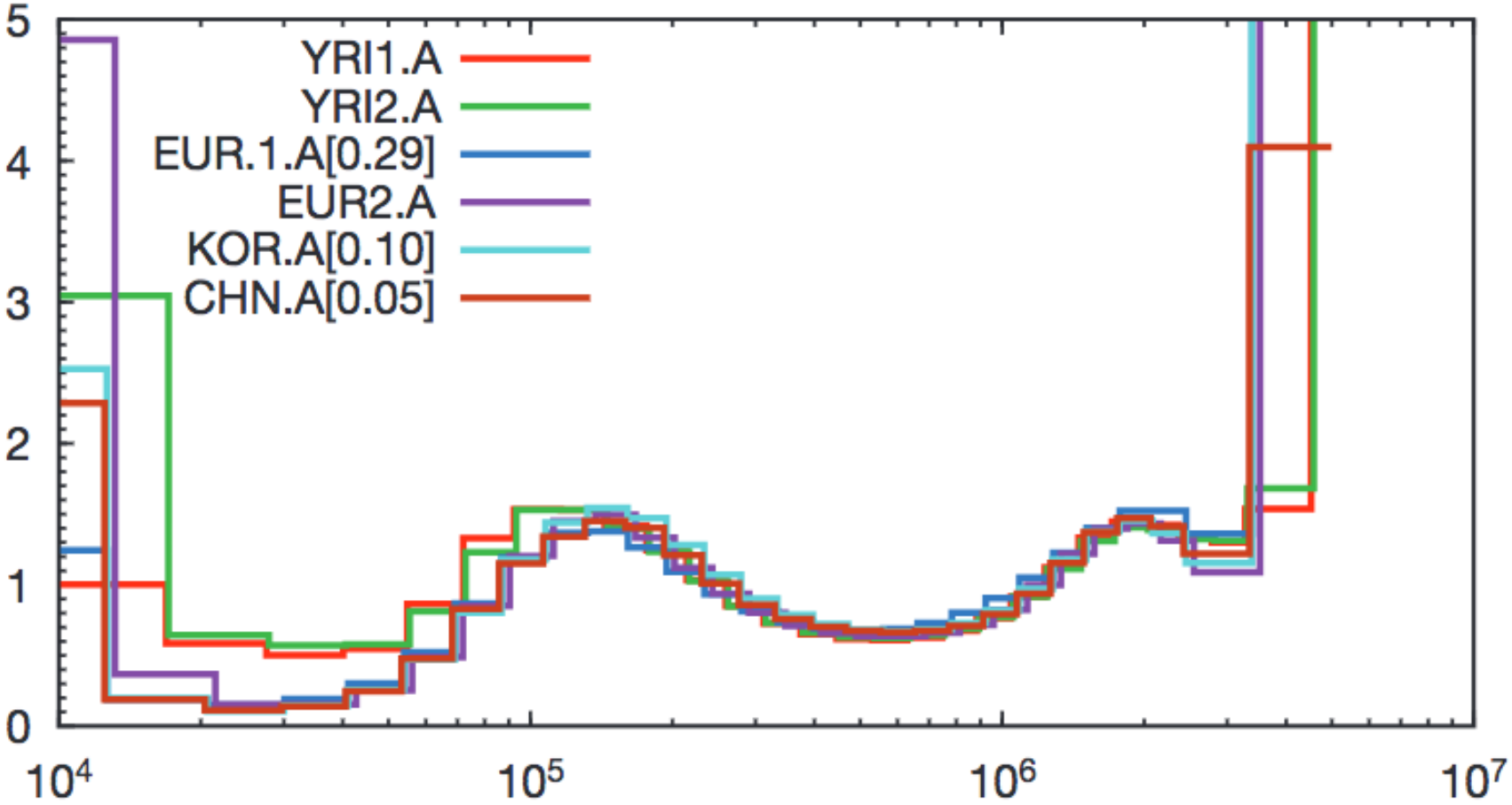
McVean & Cardin (2005)



Li & Durbin (2011)

Inference of demographic history of human ancestry by PSMC

Generation time = 25 years, mutation proba $u = 2.5 \times 10^{-8}$ per generation per bp



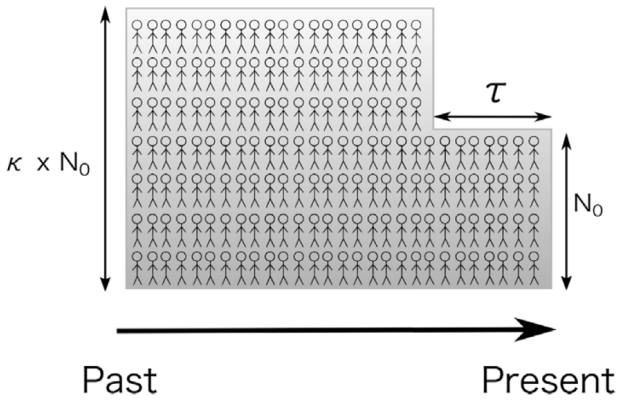
Li & Durbin (2011)

- Severe bottleneck 10–60 kyr ago
- Differentiation of genetically modern humans starting as early as 100–120 kyr ago
- Elevated pop size betw 60 and 250 kyr ago, possible artefact due to pop substructure (involving small, separately evolving isolated pops).

A quantitative assessment of extinction risk

Kerdoncuff, Lambert & Achaz, "Testing for population decline using maximal linkage disequilibrium blocks" *TPB* 2020

- Goal : detect decline $\kappa N_0 \rightarrow N_0, \tau N_0$ generations ago

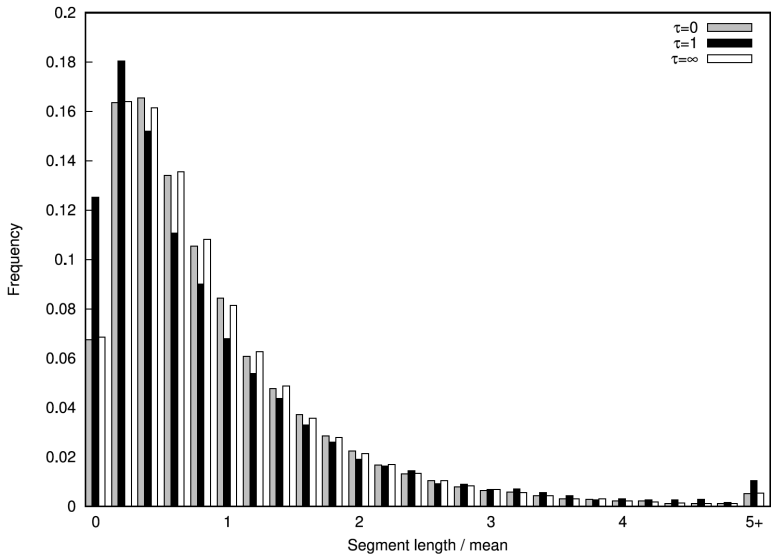


- Maximum LD block = maximal segment with mutations compatible with a single tree

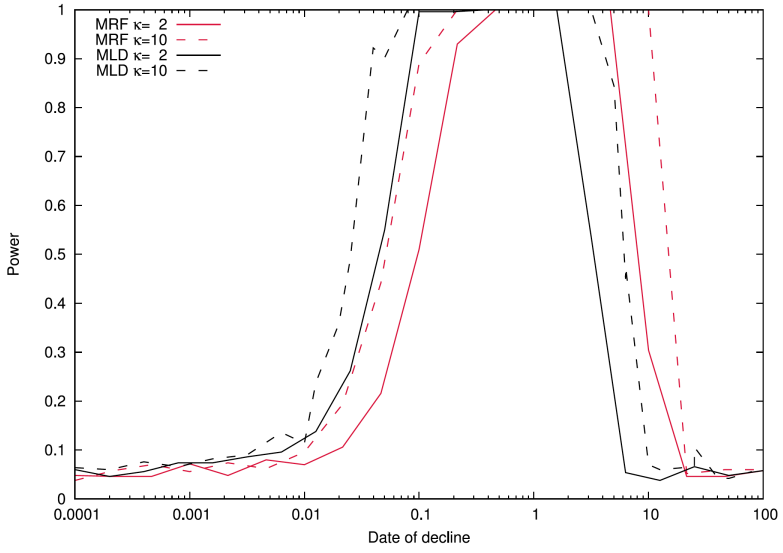
- Normalized distribution of MLD block lengths insensitive to pop size, sensitive to pop size variations

- Example. Power $> 50\%$ for $n = 10$ and a smooth, recent decline ($\kappa = 2, \tau = 0.05$).

- Requires good quality sequences. Sensitive to population structure.



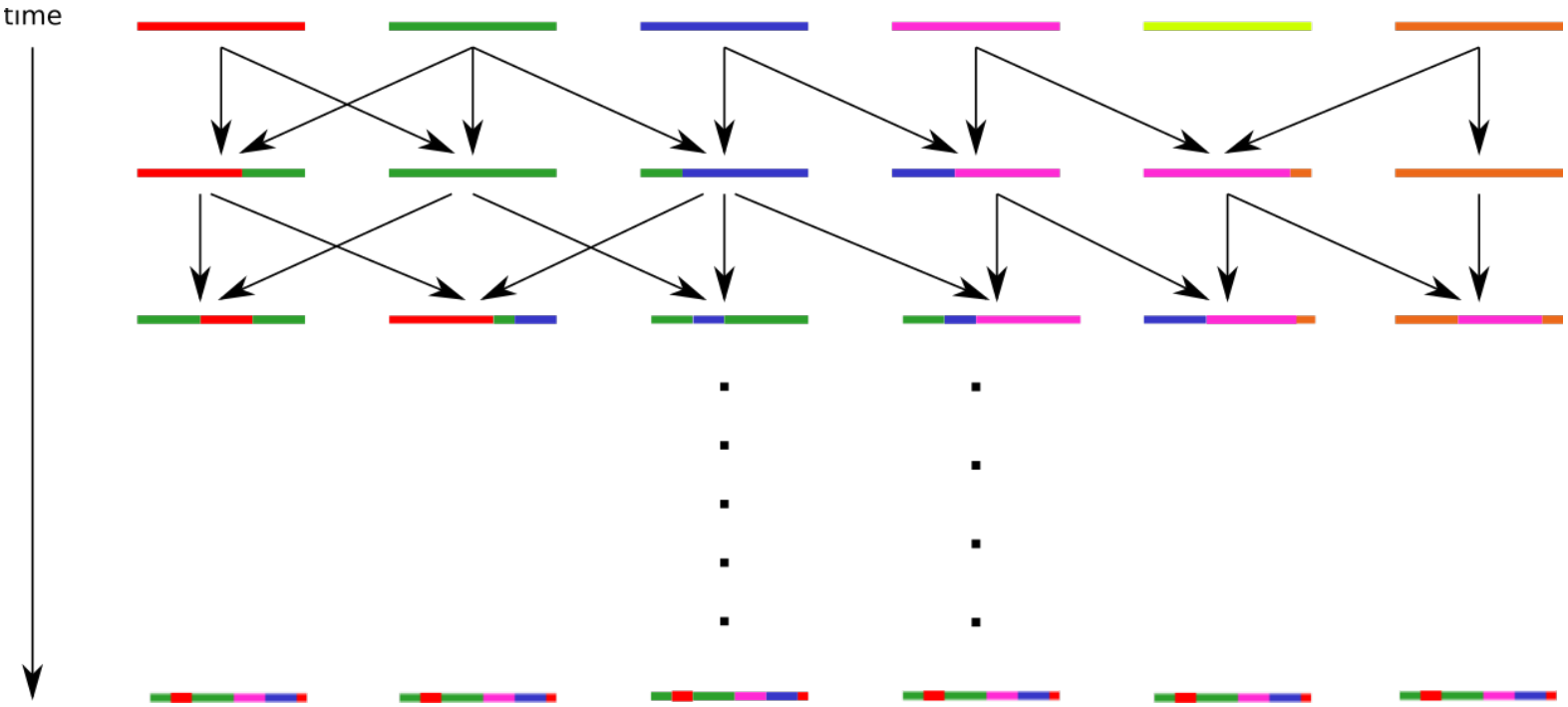
$n = 10, \kappa = 3$



$n = 10, \kappa = 2$ or 10

Chromosome painting

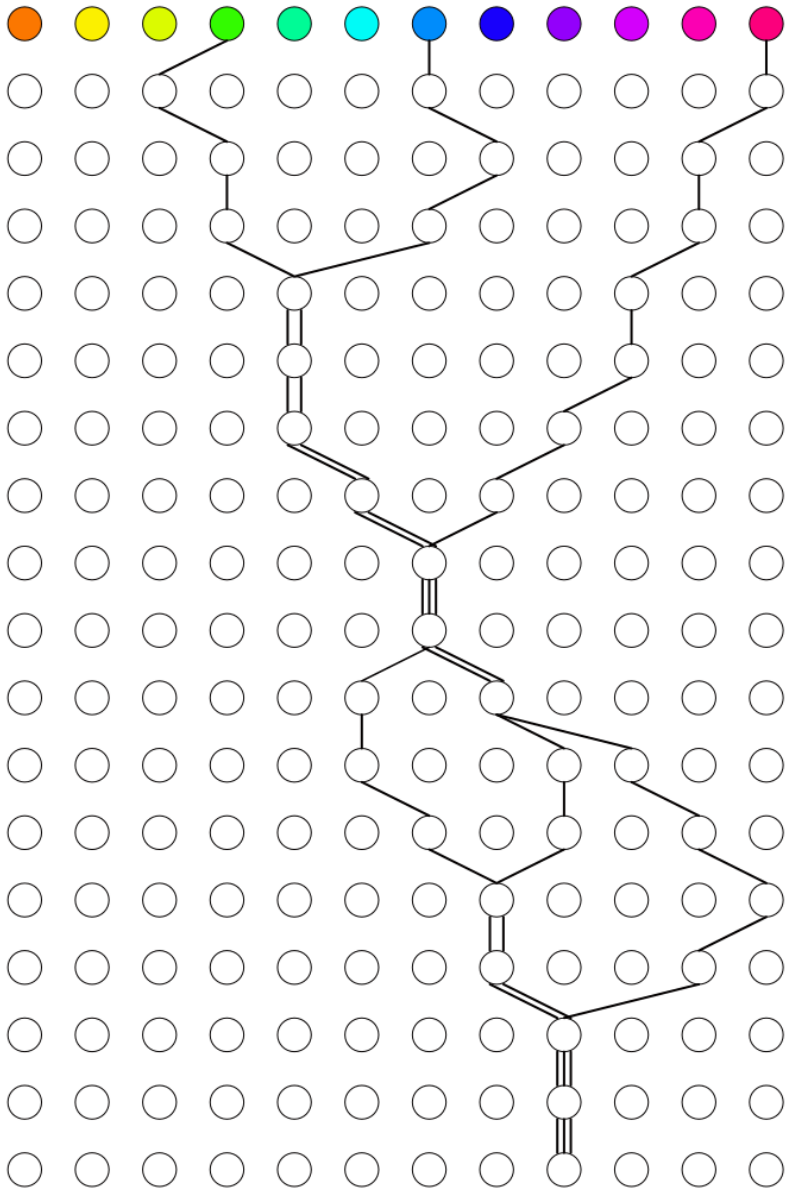
- Recall Wright–Fisher model with pop size N, recombination prob ρ/N (here $\rho = R$)
- Chromosome $=$ interval $[0, R]$
- Start with N ind and paint each of these N initial sequences with a different color.
- After some fixed amount of time, pick one individual at random : how does the mosaic of colors on this chromosome look like ?
- When time is sufficiently large, all individuals carry the same fixed chromosome : How does the fixed mosaic look like ?



picture by Verónica Miró Pina

The Ancestral Recombination Graph – cont'd

- Recall that when $n = 1$, the ARG can be described by a partition of $[0, R]$, induced by the relation of common ancestry
- Initial state : coarse partition
- The fixed mosaic is given by the stationary distribution of this partitioning process

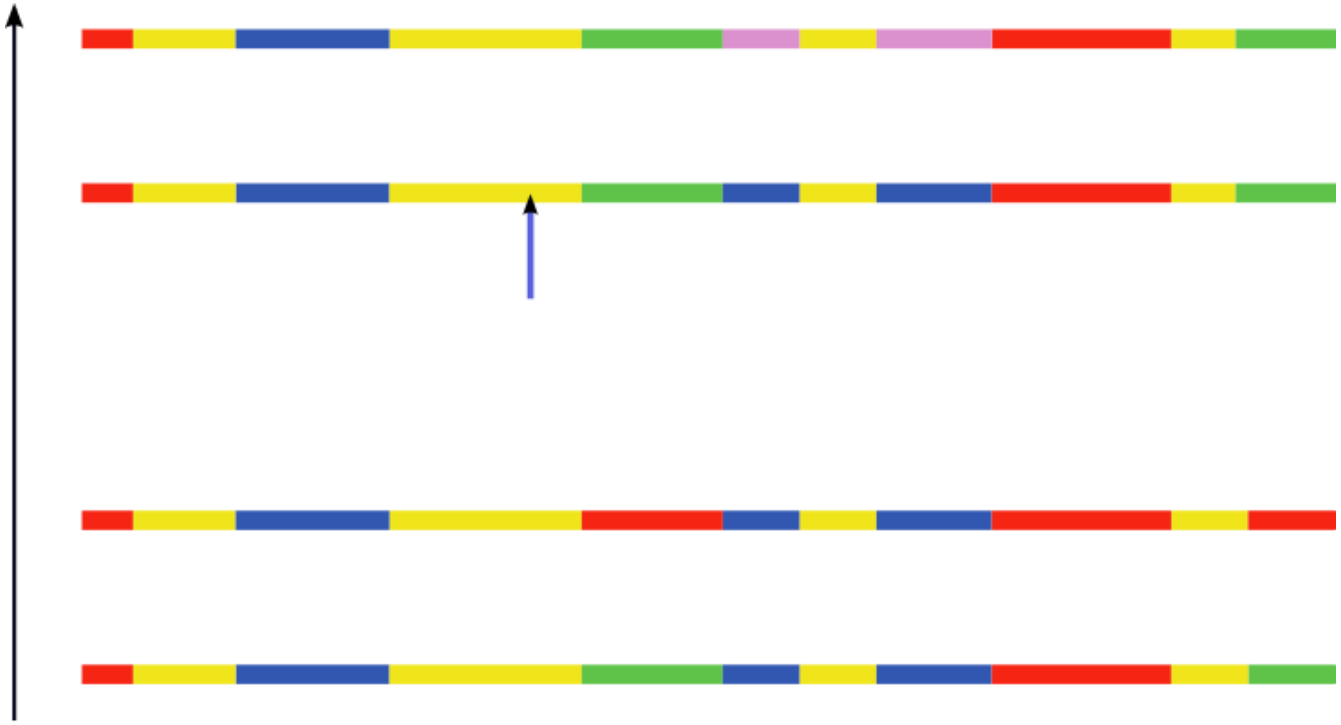


The partitioning process

Esser, Probst & Baake, Lambert, Miró Pina & Schertzer

Recall recombinations fall at rate 1 per unit time, per unit length.

- Each cluster (here, blue) independently splits into two at rate equal to its diameter at a point uniformly distributed in its convex hull
- Each pair of clusters (here, red and green) independently coalesces at rate 1



Zooming out logarithmically on the fixed mosaic

Lambert, Miró Pina & Schertzer "Chromosome painting : how recombination mixes ancestral colors" *Ann Appl Prob* (2020)

- Recall $0 \sim x$, if x carries same color as left extremity of chromosome (say, red)
- Define length of cluster containing 0

$$L_R := \int_0^R \mathbb{1}_{0\sim x}\, dx$$

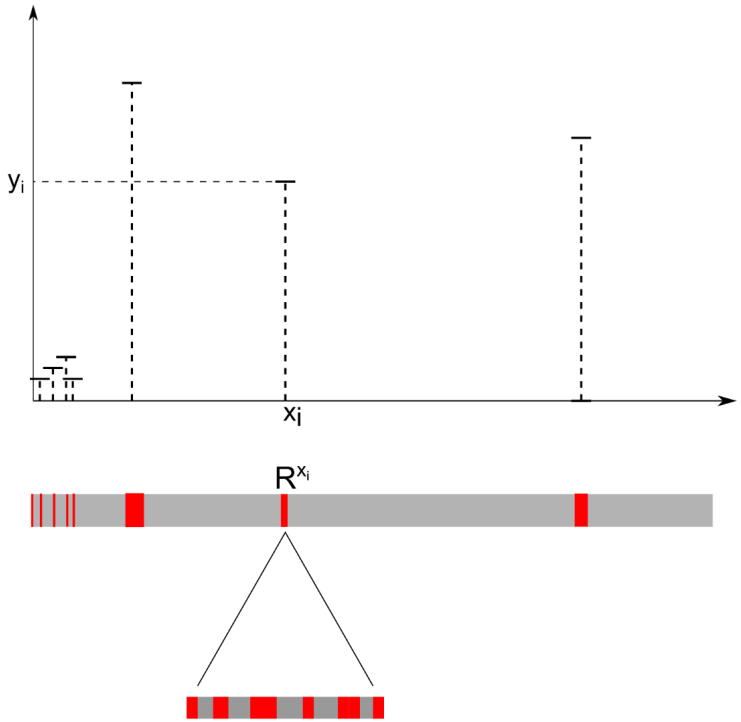
and for $0 \leq a \leq b \leq 1$,

$$\vartheta_R([a,b]) := \frac{1}{\log(R)} \int_{R^a}^{R^b} \mathbb{1}_{0\sim x}\, dx$$

Theorem (L., Miró Pina & Schertzer 2020)

As $R \to \infty$,

- $L_R / \log(R) \to \mathcal{E}(1)$
- $\vartheta_R \to \sum_i y_i \delta_{x_i}$ *where (x_i, y_i) are the atoms of a PPP with intensity $x^{-2} e^{-y/x} dx\, dy$.*



In the logarithmic scale, the segments IBD with 0 are distributed according to the scale-invariant PPP (intensity $x^{-1}dx$) and the length of segment at R^x is exponential with mean $x \log(R)$.

Insert shows complex geometry of these segments at finer scale not described in the Theorem.

Number of ancestors contributing to today's genomes

Theorem (L., Miró Pina & Schertzer 2020)

Let $\epsilon > 0$ *and let* $M_\epsilon(R) =$ *number of clusters in* $[0, R]$ *with length larger than* $\epsilon \ln(R)$. *Then*

$$\lim_{\epsilon \to 0} \lim_{R \to \infty} \frac{\ln(R)}{R} M_\epsilon(R) = 1 \qquad \text{in probability.}$$

Conjecture (Wiuf and Hein 1997)

There exists a constant $c \approx 1.38$ such that

$$\lim_{R \to \infty} \frac{\ln(R)}{R} M(R) = c \qquad \text{(in law, a.s.?)}$$

Collaborators

Guillaume ACHAZ (SMILE) . . . . . 

Élise KERDONCUFF (SMILE) . . . . . 

Verónica MIRÓ PINA (SMILE) . . . . . 

Emmanuel SCHERTZER (SMILE) . . . . . 

Outline

Books

Durrett, R. (2008) *Probability models for DNA sequence evolution.*

Wakeley, J. (2009) *Coalescent theory.*

Etheridge, A. (2011) *Some mathematical models from population genetics — École d'été de probabilités de Saint-Flour XXXIX-2009.*

Ewens, W.J. (2012) *Mathematical population genetics.*

Monographs

Lambert, “Population dynamics and random genealogies”, *Stoch Models* (2008)

Berestycki, “Recent progress in coalescent theory”, *Ensaios matem* (2009)

Lambert, “Probabilistic models for the (sub)tree(s) of life”, *Braz J Probab Stat* (2017)

Lambert, “Random ultrametric trees and applications”, *ESAIM P& S* (2018)

Papers

Kingman, “The coalescent”, *Stoch Proc Appl* (1982)

Donnelly & Tavaré “The ages of alleles and a coalescent”, *Adv Appl Prob* (1986)

Griffiths & Marjoram, “An ancestral recombination graph”, *IMA volume : Progress in Population Genetics and Human Evolution* (1997)

Wiuf & Hein, “On the number of ancestor to a DNA sequence”, *Genetics* (1997)

Möhle, “Robustness results for the coalescent” *J Appl Prob* (1998)

McVean & Cardin, “Approximating the coalescent with recombination” *Philos Transac RS* (2005)

Jenkins & Song, “An asymptotic sampling formula for the coalescent with recombination”, *Ann Appl Probab* (2010)

Li & Durbin, “Inference of human population history from individual whole-genome sequences”, *Nature* (2011)

Liu & Fu, “Exploring population size changes using SNP frequency spectra”, *Nature Genetics* (2015)

Régnier*, Achaz*, Lambert, Cowie, Bouchet & Fontaine, “Mass extinction in poorly known taxa”, *PNAS* (2015)

Esser, Probst & Baake, “Partitioning, duality, and linkage disequilibria in the Moran model with recombination”, *J Math Biol* (2016)

Lapierre, Lambert & Achaz, “Accuracy of demographic inferences from Site Frequency Spectrum : The case of the Yoruba population” *Genetics* (2017)

Kerdoncuff, Lambert & Achaz, “Testing for population decline using maximal linkage disequilibrium blocks” *Theoret Popul Biol* (2020)

Lambert, Miró Pina & Schertzer, “Chromosome painting : how recombination mixes ancestral colors” *Ann Appl Prob* (2020)

Peer Community In Evolutionary Biology

PCI Evol Biol

FREE and transparent **preprint** and postprint **recommendations** in Evolutionary Biology

WHY?

The current publication system is too expensive, particularly given that it relies heavily on the unpaid work of scientists (as authors, editors and reviewers), and is generally not transparent.

Preprints are free but are not peer-reviewed

WHAT?

A community of researchers evaluate by **peer review** and **recommend preprints** deposited in open archives

HOW?

1. YOU DEPOSIT YOUR PREPRINT IN AN OPEN ARCHIVE & OBTAIN A DOI

2. YOU REQUEST A RECOMMENDATION FOR YOUR PREPRINT FROM PCI EVOL BIOL

3. A RECOMMENDER CHOOSES TO HANDLE YOUR PREPRINT & OBTAINS ≥ 2 PEER-REVIEWS

AFTER ≥ 1 ROUND OF REVIEW/CORRECTION, YOUR PREPRINT IS

NOT RECOMMENDED BUT YOU HAVE REVIEWS TO IMPROVE IT

RECOMMENDED, PUBLICISED AND CAN NOW BE CITED AS PEER REVIEWED

YOU ARE FREE TO SUBMIT YOUR PREPRINT TO A CLASSIC JOURNAL

https://evolbiol.peercommunityin.org