On modern problems and methods for data analysis in human genomics

Vladimir Shchur, Richard Durbin

CSP2015, 8th September 2015, Moscow

Talk overview

- What is our data?
- Population models: from basics to coalescent.
- Ancestral Recombination Graph: data structure for handling recombinations.
  - How recombinations affect data?
  - How to make inference taking in account recombinations?
- Li and Stephens model: a different probabilistic model of genomic data.

Basic conceptions

- A DNA molecule is a sequence of four nucleotides: cytosine (C), guanine (G), adenine (A) and thymine (T). Genomic data is a text over a four-letter alphabet.
- A *genome* is the genetic material of an organism consisting of DNA (or RNA for some viruses). It includes genes and non-coding regions and packed and organised into *chromosomes*, each of which is a long DNA molecule.
- Human genome is *diploid*: it contains two sets of chromosomes, one coming from each parent. Genetic material from one parent is called a *haplotype*.

Basic conceptions

- A DNA molecule is a sequence of four nucleotides: cytosine (C), guanine (G), adenine (A) and thymine (T). Genomic data is a text over a four-letter alphabet.
- A *genome* is the genetic material of an organism consisting of DNA (or RNA for some viruses). It includes genes and non-coding regions and packed and organised into *chromosomes*, each of which is a long DNA molecule.
- Human genome is *diploid*: it contains two sets of chromosomes, one coming from each parent. Genetic material from one parent is called a *haplotype*.

Basic conceptions

- A DNA molecule is a sequence of four nucleotides: cytosine (C), guanine (G), adenine (A) and thymine (T). Genomic data is a text over a four-letter alphabet.
- A *genome* is the genetic material of an organism consisting of DNA (or RNA for some viruses). It includes genes and non-coding regions and packed and organised into *chromosomes*, each of which is a long DNA molecule.
- Human genome is *diploid*: it contains two sets of chromosomes, one coming from each parent. Genetic material from one parent is called a *haplotype*.

Some numbers

- Human genome length $\approx$ 3Gb (Giga-basepairs).
- There are $\approx$ 3 million differences between two typical human haplotypes, e.g. maternal and paternal versions in one person.
- Most of these are shared with other people, caused by mutations in the distant past, 10s or 100s of thousands of years ago.
- Each one of us receives approximately 35 new mutations in our genome from our parents, 10^{-8} per bp per generation (though this estimate varies a lot!).

Some numbers

- Human genome length $\approx$ 3Gb (Giga-basepairs).
- There are $\approx$ 3 million differences between two typical human haplotypes, e.g. maternal and paternal versions in one person.
- Most of these are shared with other people, caused by mutations in the distant past, 10s or 100s of thousands of years ago.
- Each one of us receives approximately 80 new mutations in our genome from our parents, 10^{-8} per bp per generation (though this estimate varies a lot!).
- Almost 100 millions of Single Nucleotide Polymorphisms (SNP) are validated according to dbSNP.

Some numbers

- Human genome length $\approx$ 3Gb (Giga-basepairs).
- There are $\approx$ 3 million differences between two typical human haplotypes, e.g. maternal and paternal versions in one person.
- Most of these are shared with other people, caused by mutations in the distant past, 10s or 100s of thousands of years ago.
- Each one of us receives approximately 80 new mutations in our genome from our parents, 10^{-8} per bp per generation (though this estimate varies a lot!).
- Almost 100 millions of Single Nucleotide Polymorphisms (SNP) are validated according to *dbSNP*.

Some numbers

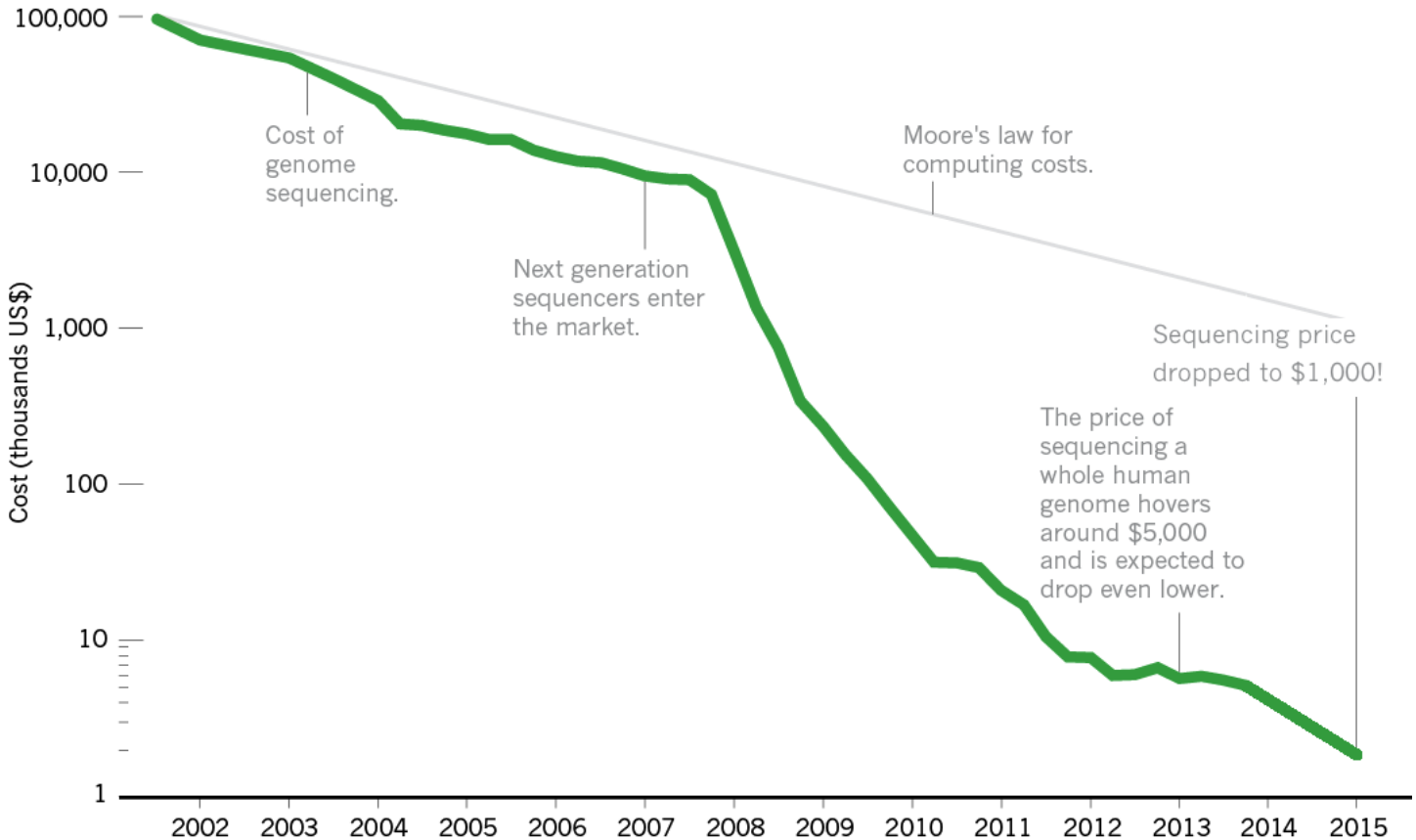
- Human genome length $\approx$ 3Gb (Giga-basepairs).
- There are $\approx$ 3 million differences between two typical human haplotypes, e.g. maternal and paternal versions in one person.
- Most of these are shared with other people, caused by mutations in the distant past, 10s or 100s of thousands of years ago.
- Each one of us receives approximately 80 new mutations in our genome from our parents, 10^{-8} per bp per generation (though this estimate varies a lot!).
- Almost 100 millions of Single Nucleotide Polymorphisms (SNP) are validated according to *dbSNP*.

Some numbers

- Human genome length $\approx$ 3Gb (Giga-basepairs).
- There are $\approx$ 3 million differences between two typical human haplotypes, e.g. maternal and paternal versions in one person.
- Most of these are shared with other people, caused by mutations in the distant past, 10s or 100s of thousands of years ago.
- Each one of us receives approximately 80 new mutations in our genome from our parents, 10^{-8} per bp per generation (though this estimate varies a lot!).
- Almost 100 millions of Single Nucleotide Polymorphisms (SNP) are validated according to *dbSNP*.

Sequencing costs

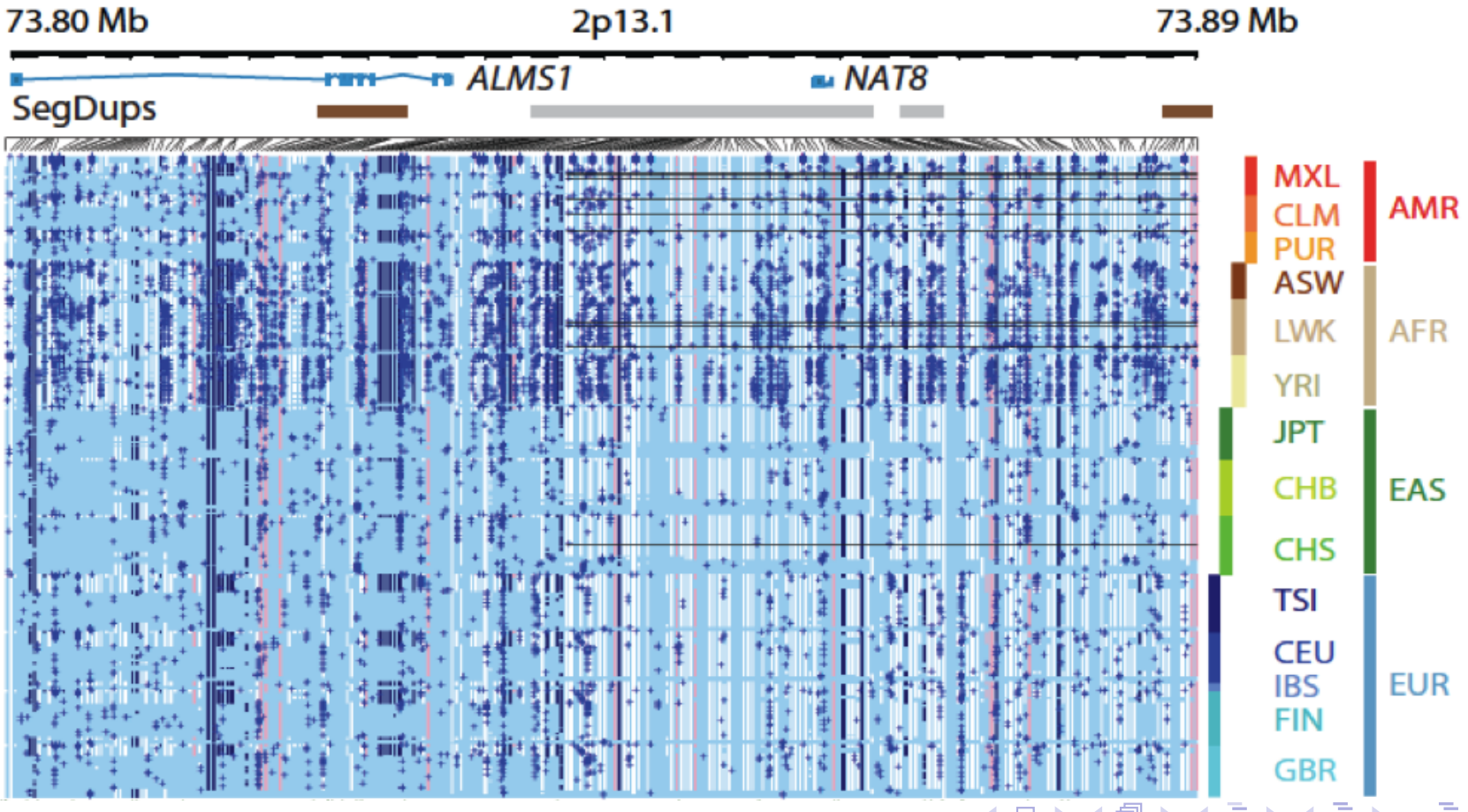
Sequencing price reduced dramatically which allow to create huge genomic data bases.



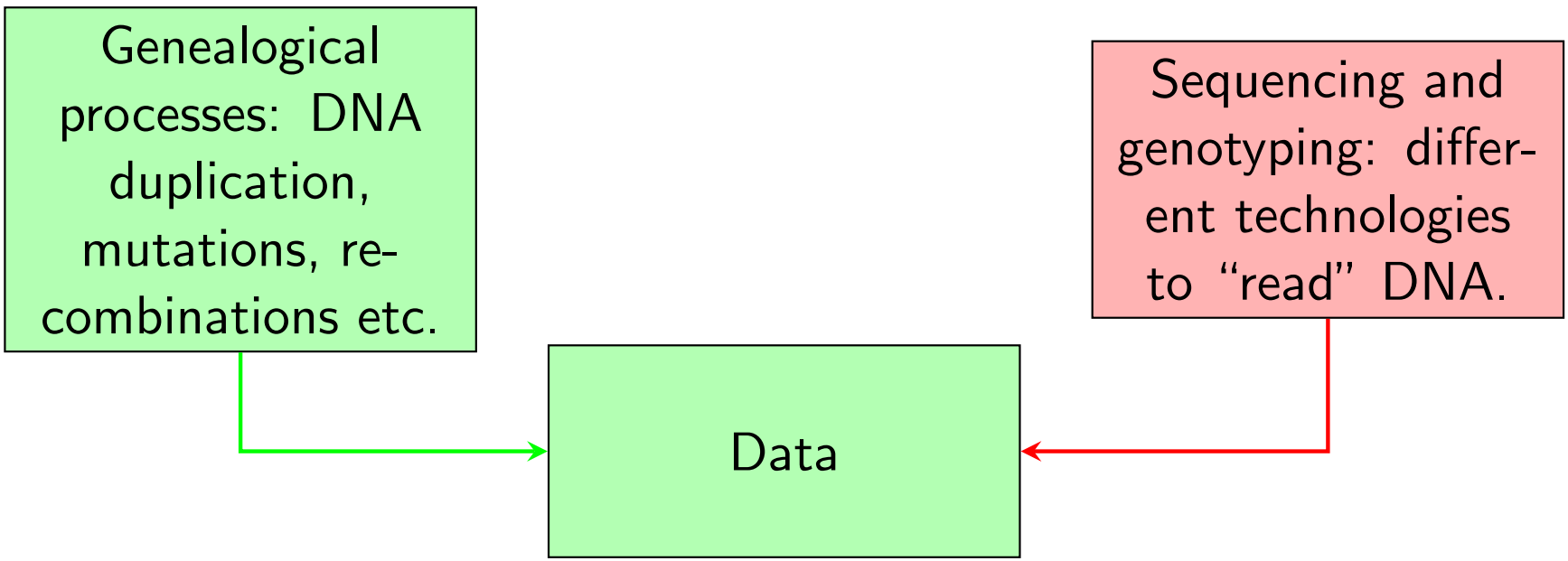
Erika Check Hayden, Technology: The $1,000 genome. Nature, 2014

1000 Genome project

1000 Genome Project is one of the biggest genomic data sets. Currently (phase 3) it contains 2504 human individuals with 88 millions variant sites.



What shapes the data?



Sequencing

Sequencing technologies also influence on the data. Sequencing produce a pool of *reads* - short strands of DNA, current standard is 70-150 base pairs long, but we can get up to 10,000bp. Each position can be covered by several reads (this number is random though). This poses the following problems:

- De-novo assembly.
- Read alignment against reference genome.
- Variant calling: are there enough reads that support the variant?
- Phasing: if SNPs carries variants G and A and SNP2 with C and T, there are two possible genomes which could explain the data.

Sequencing

Sequencing technologies also influence on the data. Sequencing produce a pool of *reads* - short strands of DNA, current standard is 70-150 base pairs long, but we can get up to 10,000bp. Each position can be covered by several reads (this number is random though). This poses the following problems:

- De-novo assembly.
- Read alignment against reference genome.
- Variant calling: are there enough reads that support the variant?
- Phasing: if SNP_1 carries variants G and A and SNP_2 with C and T, there are two possible genomes which can underly the data:

$$\begin{array}{c|c} \ldots G \ldots A \ldots & \ldots G \ldots T \ldots \\ \ldots C \ldots T \ldots & \ldots C \ldots A \ldots \end{array}$$

Sequencing

Sequencing technologies also influence on the data. Sequencing produce a pool of *reads* - short strands of DNA, current standard is 70-150 base pairs long, but we can get up to 10,000bp. Each position can be covered by several reads (this number is random though). This poses the following problems:

- De-novo assembly.
- Read alignment against reference genome.
- Variant calling: are there enough reads that support the variant?
- Phasing: if SNP_1 carries variants G and A and SNP_2 with C and T, there are two possible genomes which can underly the data:

$$\begin{array}{c|c} \ldots G \ldots A \ldots & \ldots G \ldots T \ldots \\ \ldots C \ldots T \ldots & \ldots C \ldots A \ldots \end{array}$$

Sequencing

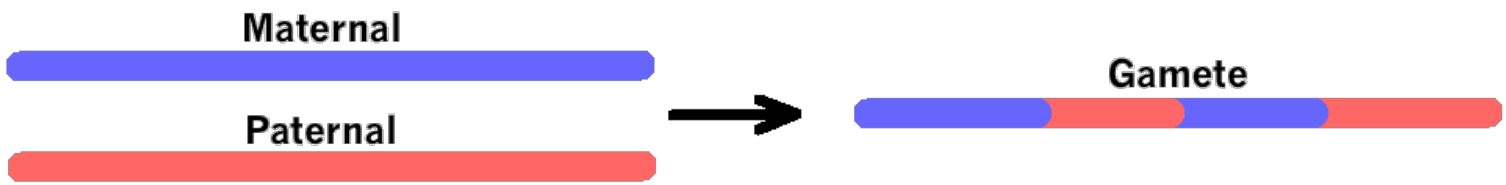
Sequencing technologies also influence on the data. Sequencing produce a pool of *reads* - short strands of DNA, current standard is 70-150 base pairs long, but we can get up to 10,000bp. Each position can be covered by several reads (this number is random though). This poses the following problems:

- De-novo assembly.
- Read alignment against reference genome.
- Variant calling: are there enough reads that support the variant?
- Phasing: if SNP_1 carries variants G and A and SNP_2 with C and T, there are two possible genomes which can underly the data:

$$\begin{array}{c|c} \ldots \mathrm{G} \ldots \mathrm{A} \ldots & \ldots \mathrm{G} \ldots \mathrm{T} \ldots \\ \ldots \mathrm{C} \ldots \mathrm{T} \ldots & \ldots \mathrm{C} \ldots \mathrm{A} \ldots \end{array}$$

Genealogical processes

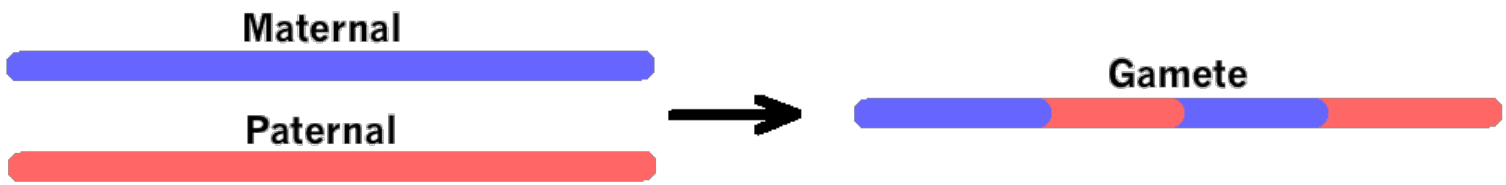
- All the life reproduction is based on cell division. Genetic material is duplicated during this process.
- Errors can occur during duplication. It can be a single nucleotide polymorphism (SNP), insertions, deletions and some other.
- Human gametes (reproduction cells) contain only one set of chromosomes which is a mosaic of parental two sets of chromosomes, which is created by *recombinations*.



Problem: Estimate mutation and recombination rates.

Genealogical processes

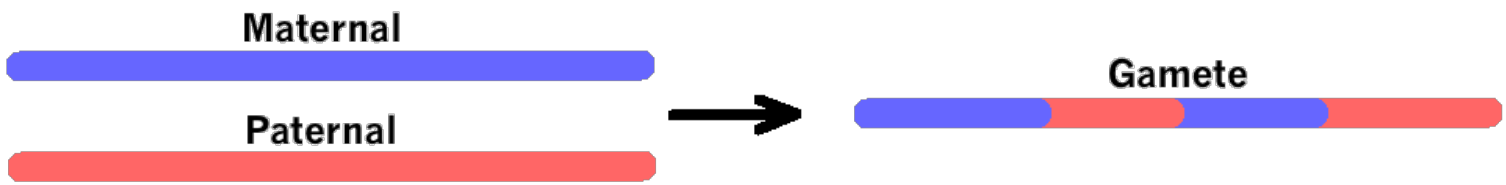
- All the life reproduction is based on cell division. Genetic material is duplicated during this process.
- Errors can occur during duplication. It can be a single nucleotide polymorphism (SNP), insertions, deletions and some other.
- Human gametes (reproduction cells) contain only one set of chromosomes which is a mosaic of parental two sets of chromosomes, which is created by *recombinations*.



Problem: Estimate mutation and recombination rates.

Genealogical processes

- All the life reproduction is based on cell division. Genetic material is duplicated during this process.
- Errors can occur during duplication. It can be a single nucleotide polymorphism (SNP), insertions, deletions and some other.
- Human gametes (reproduction cells) contain only one set of chromosomes which is a mosaic of parental two sets of chromosomes, which is created by *recombinations*.



Problem: Estimate mutation and recombination rates.

Basic models

Problem: model population genomes.

- The most evident population models work forward in time introducing birth (together with the choice of parent) and death of individuals.
- Wright-Fisher model and Moran model are the classical examples.
- The important parameter which affects the shape of genealogy is the *effective population size*: the number of breeding individuals in an idealised population.
- These models have more theoretical than computational interest.

Basic models

Problem: model population genomes.

- The most evident population models work forward in time introducing birth (together with the choice of parent) and death of individuals.
- Wright-Fisher model and Moran model are the classical examples.
- The important parameter which affects the shape of genealogy is the *effective population size*: the number of breeding individuals in an idealised population.
- These models have more theoretical than computational interest.

Basic models

Problem: model population genomes.

- The most evident population models work forward in time introducing birth (together with the choice of parent) and death of individuals.
- Wright-Fisher model and Moran model are the classical examples.
- The important parameter which affects the shape of genealogy is the *effective population size*: the number of breeding individuals in an idealised population.
- These models have more theoretical than computational interest.

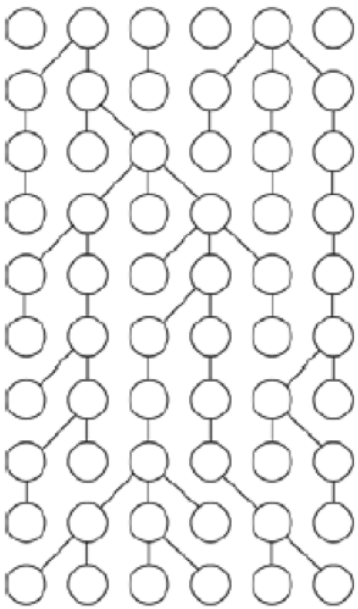
Basic models

Problem: model population genomes.

- The most evident population models work forward in time introducing birth (together with the choice of parent) and death of individuals.
- Wright-Fisher model and Moran model are the classical examples.
- The important parameter which affects the shape of genealogy is the *effective population size*: the number of breeding individuals in an idealised population.
- These models have more theoretical than computational interest.

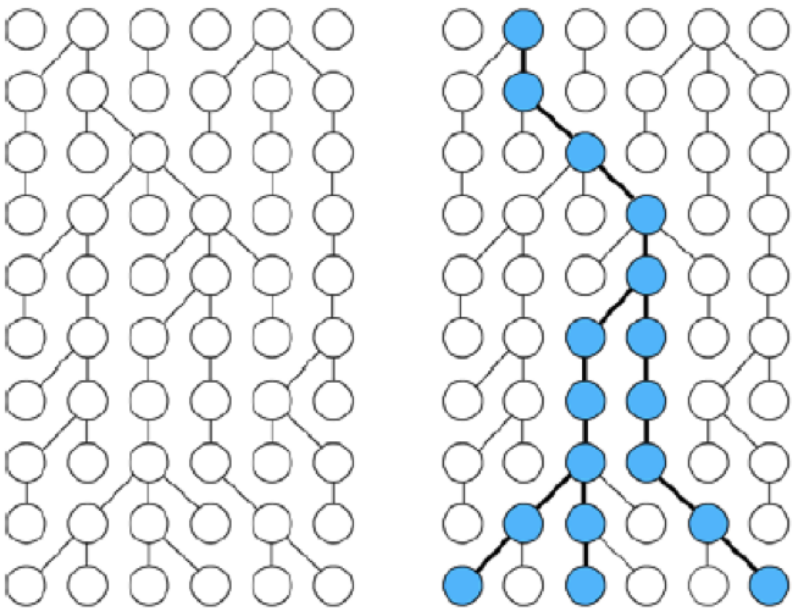
Relationship between forwards in time (Wright-Fisher) and backwards in time (Coalescent) models

- In the absence of recombinations, a genealogy of genome samples is a tree. The internal nodes of the tree corresponds to the *most recent common ancestors* of two lineages.
- Coalescent approach models genealogies backward in time, which is computationally efficient.



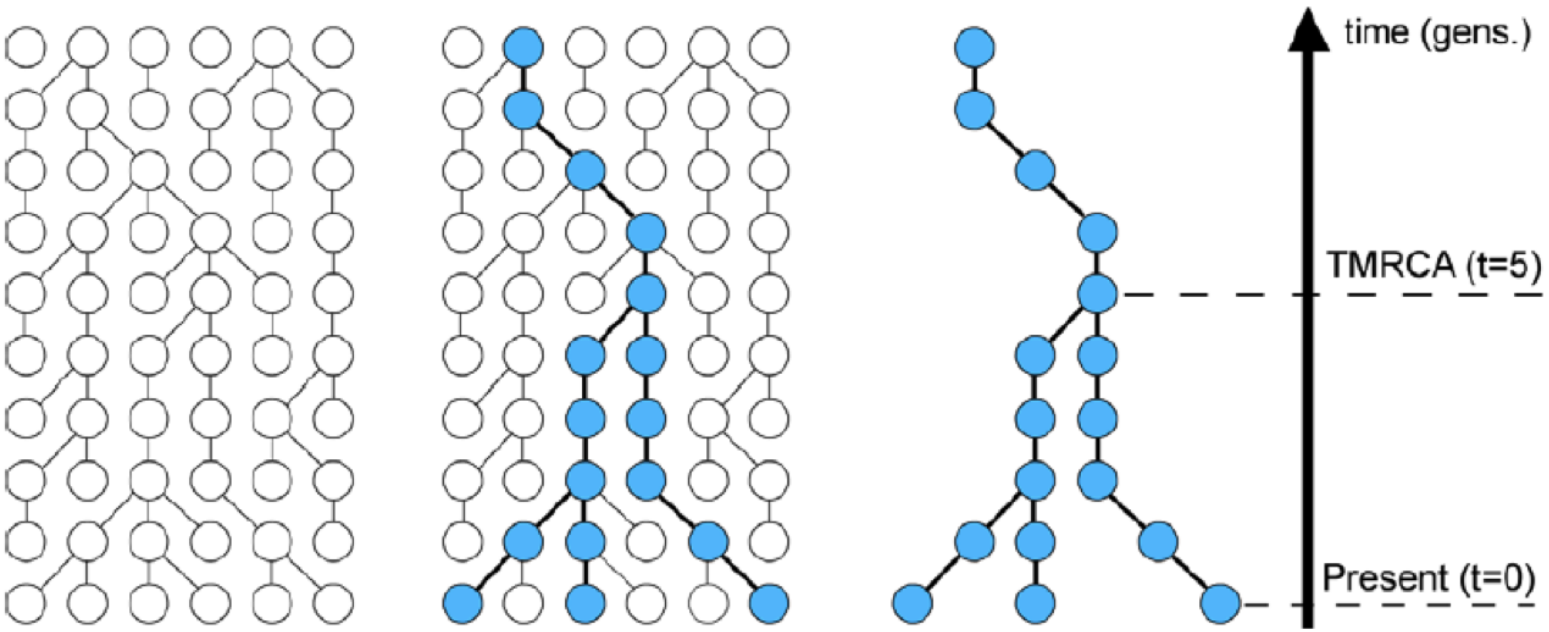
Relationship between forwards in time (Wright-Fisher) and backwards in time (Coalescent) models

- In the absence of recombinations, a genealogy of genome samples is a tree. The internal nodes of the tree corresponds to the *most recent common ancestors* of two lineages.
- Coalescent approach models genealogies backward in time, which is computationally efficient.



Relationship between forwards in time (Wright-Fisher) and backwards in time (Coalescent) models

- In the absence of recombinations, a genealogy of genome samples is a tree. The internal nodes of the tree corresponds to the *most recent common ancestors* of two lineages.
- Coalescent approach models genealogies backward in time, which is computationally efficient.



Coalescent model

Problem: to infer population history.

- Under constant population size, the times between successive coalescent events are distributed exponentially with the parameter $\binom{k}{2}$, where k is the number of lineages at the corresponding interval.
- The distribution of counts of allele frequencies j is $1/j$.
- The deviations from this law can be used to detect variation in effective population size and different population histories (isolation, migration, bottlenecks etc.). Tajima's D statistic is the classical measure reflecting this property.

Coalescent model

Problem: to infer population history.

- Under constant population size, the times between successive coalescent events are distributed exponentially with the parameter $\binom{k}{2}$, where k is the number of lineages at the corresponding interval.
- The distribution of counts of allele frequencies j is $1/j$.
- The deviations from this law can be used to detect variation in effective population size and different population histories (isolation, migration, bottlenecks etc.). Tajima's D statistic is the classical measure reflecting this property.

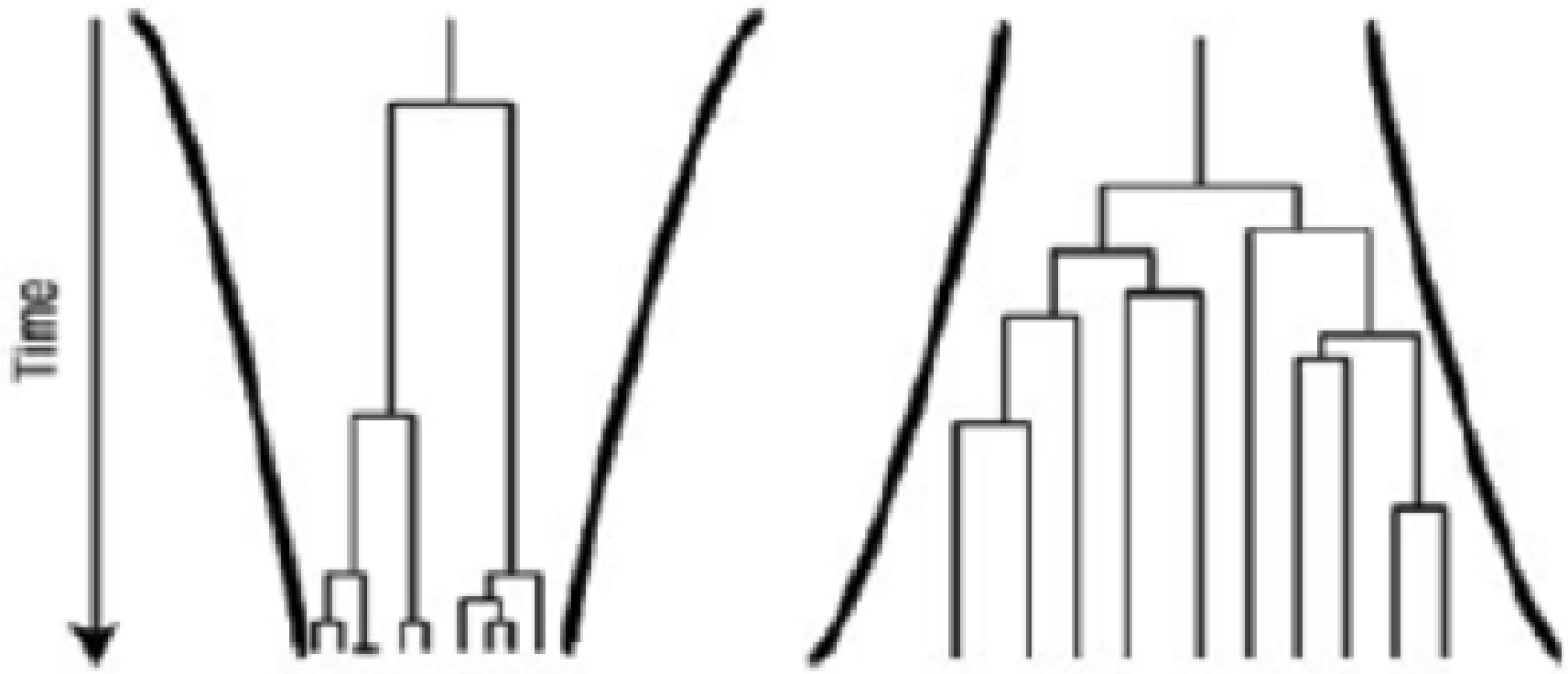
Coalescent model

Problem: to infer population history.

- Under constant population size, the times between successive coalescent events are distributed exponentially with the parameter $\binom{k}{2}$, where k is the number of lineages at the corresponding interval.
- The distribution of counts of allele frequencies j is $1/j$.
- The deviations from this law can be used to detect variation in effective population size and different population histories (isolation, migration, bottlenecks etc.). Tajima's D statistic is the classical measure reflecting this property.

Coalescent model

Here are two examples of decreasing and increasing effective population size. In the first scenario the number of singletons is relatively small, though in the second singletons will be overrepresented.

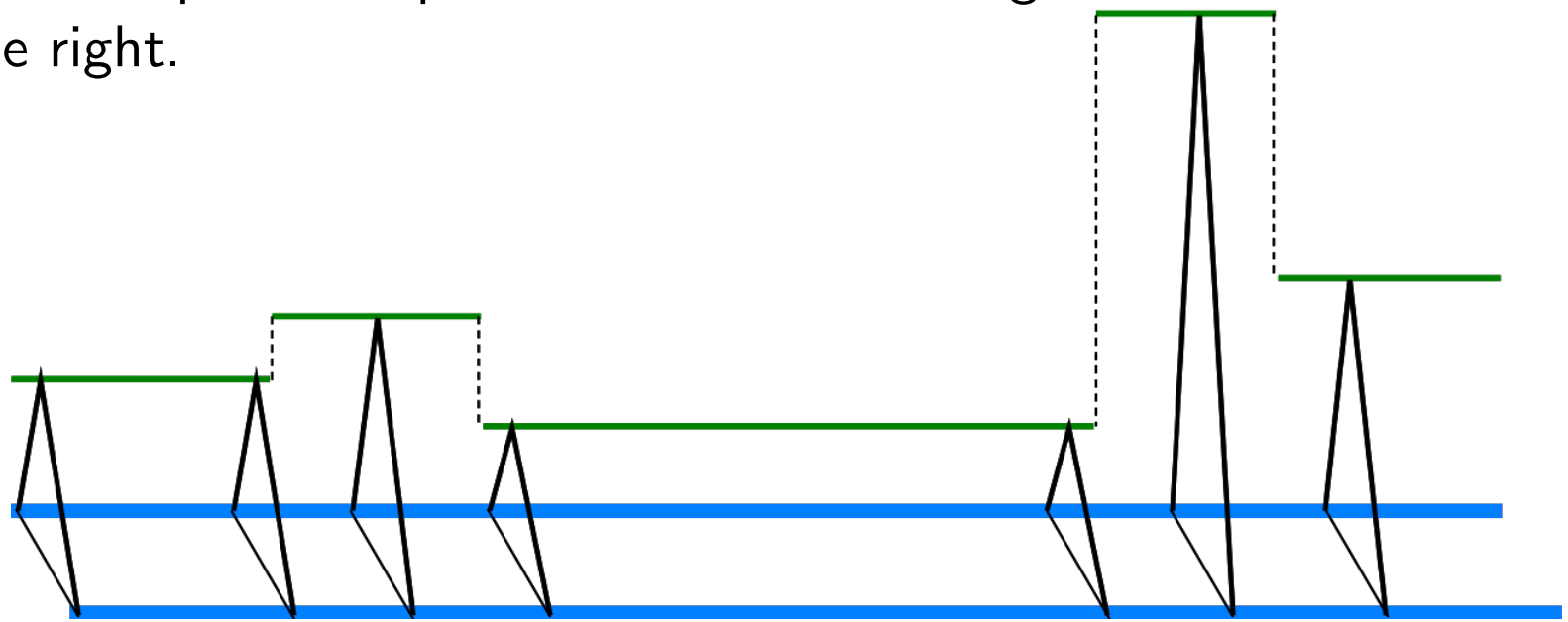


Recombinations

- If points on the genome are very close, e.g. adjacent, they share the same tree.
- If points are very far, their trees are sampled from the coalescent independently.

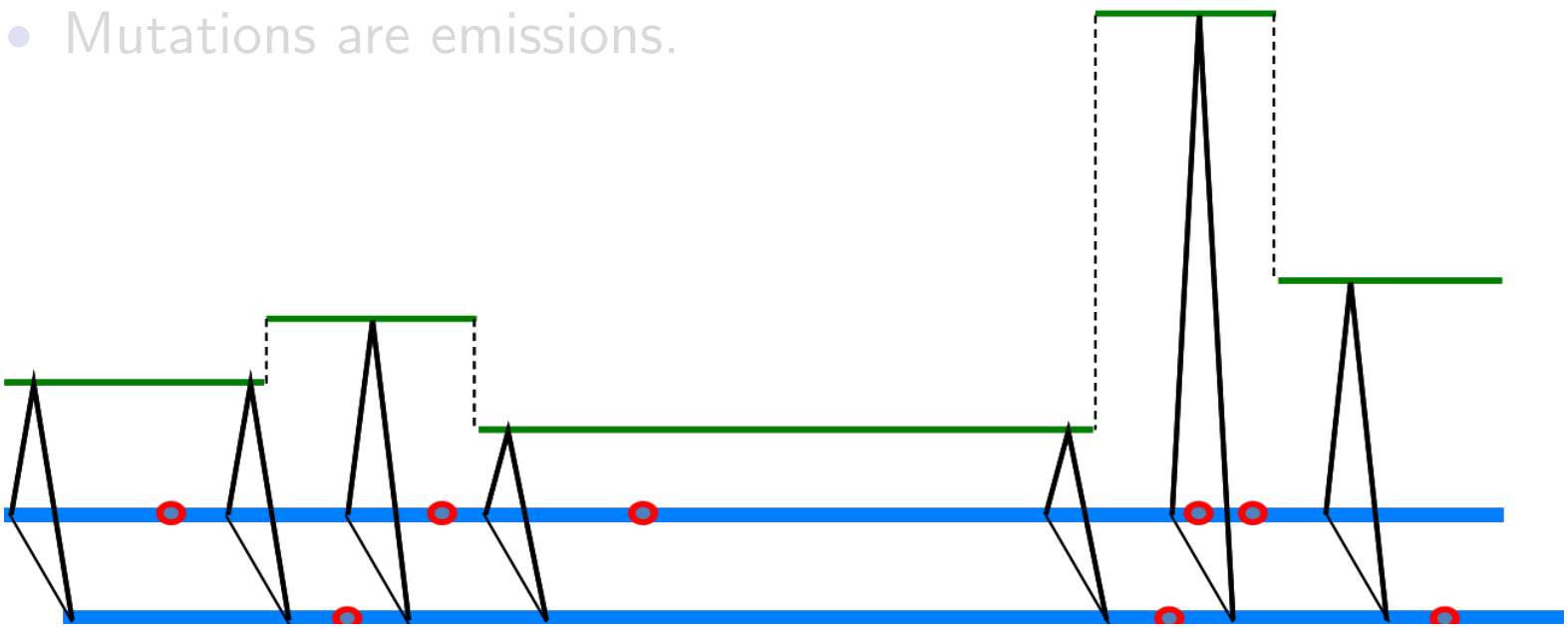
Problem: What happens in between?

A recombination in the ancestor of a modern sequence made it out of two separate sequences, one contributing to the left and one to the right.



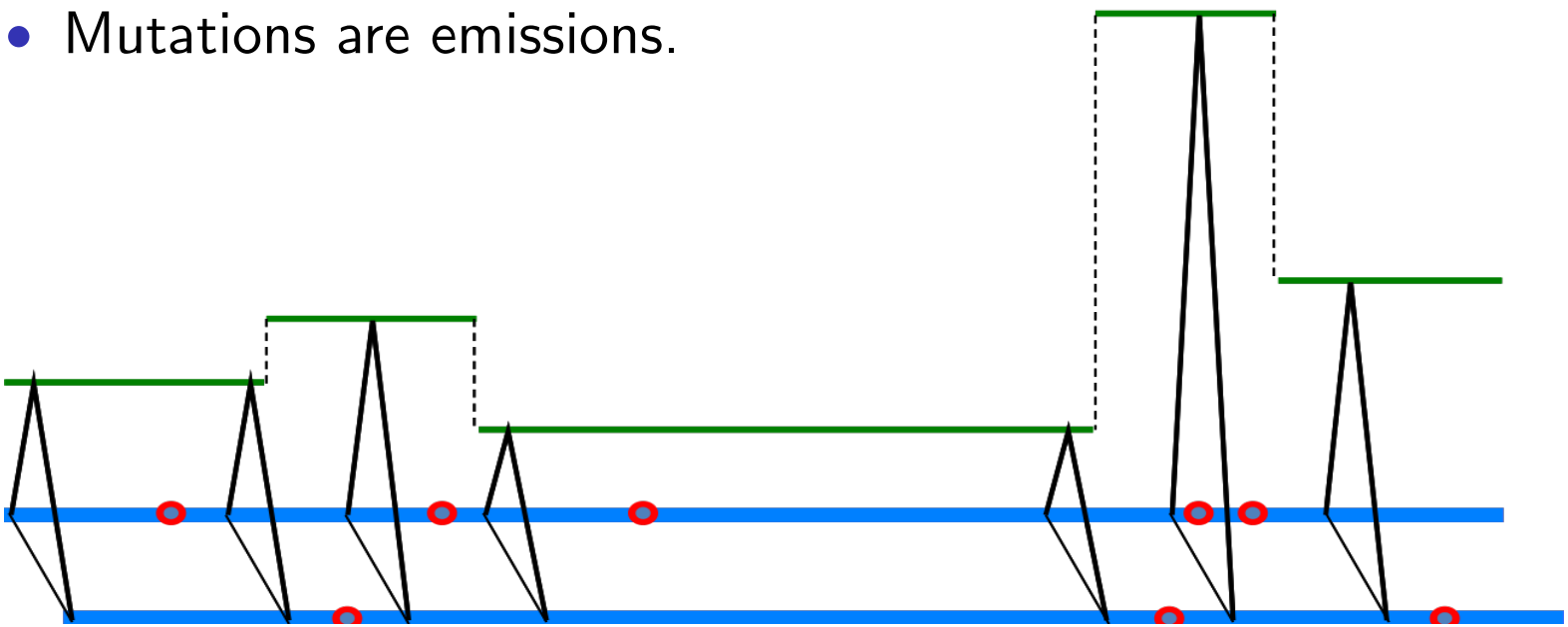
Pairwise Sequentially Markovian Approximation to the Coalescent (PSMC)

- For two haplotypes, the tree is very simple. Recombinations change its height.
- Local trees are states of a Hidden Markov Model
- Recombinations are transitions.
- Mutations are emissions.



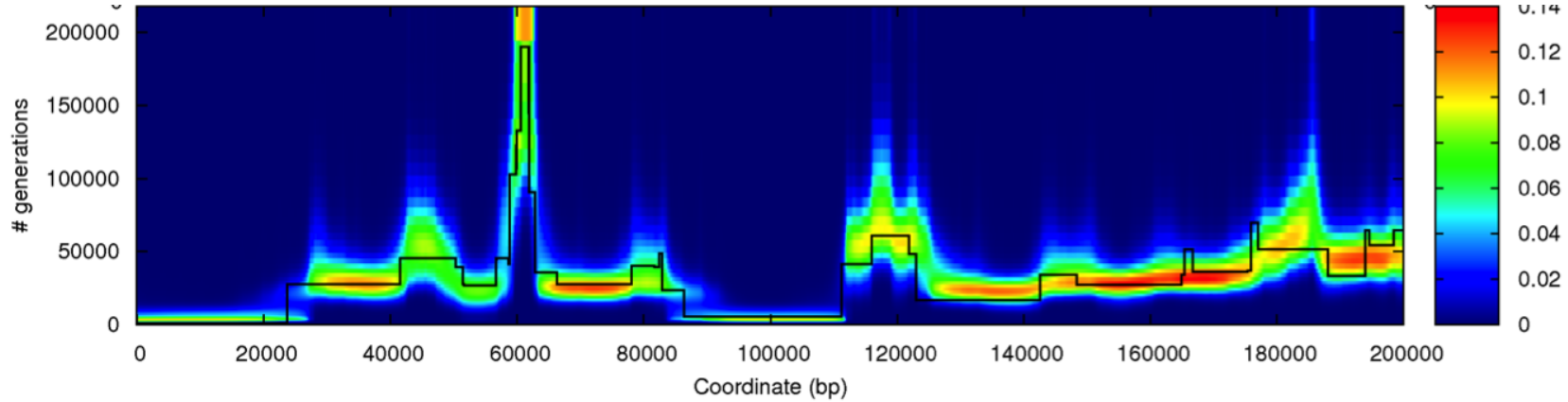
Pairwise Sequentially Markovian Approximation to the Coalescent (PSMC)

- For two haplotypes, the tree is very simple. Recombinations change its height.
- Local trees are states of a Hidden Markov Model
- Recombinations are transitions.
- Mutations are emissions.



PSMC on simulated data

PSMC reconstructs individual history. It fits effective population size and few other parameters.
Two haplotypes were simulated.

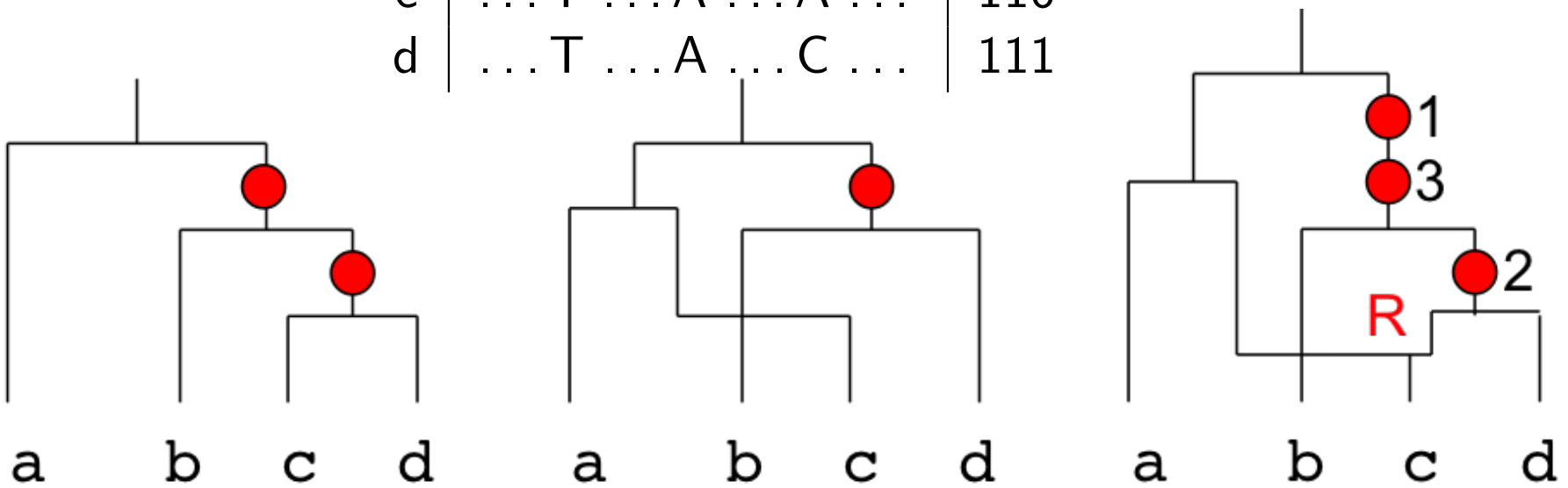


Ancestral Recombination Graph ($\mathcal{ARG}$)

The Ancestral Recombination Graph describes the way that individual sequences in a population are related.

- At a locus, sequences are related by a tree.
- Ancestral recombinations change the tree as one moves along the chromosome.

a	...C...G...A...	000
b	...T...G...C...	101
c	...T...A...A...	110
d	...T...A...C...	111



Coalescent with recombination

Problem: infer relations along the genome at many scales simultaneously.

- $\mathcal{ARG}$ is a structure (data type).
- The probability distribution over $\mathcal{ARG}$s that arises when recombination is added to the standard (Wright-Fisher) model is called the Coalescent with Recombination.
- 2 possible events going backwards in time
  - Coalescence: which merges two sequences.
    For i sequences, rate is $i(i-1)/2N$.
  - Recombination: which splits a sequence into two.
    For i sequences, rate is $iL\rho$

Coalescent with recombination

Problem: infer relations along the genome at many scales simultaneously.

- $\mathcal{ARG}$ is a structure (data type).
- The probability distribution over $\mathcal{ARG}$s that arises when recombination is added to the standard (Wright-Fisher) model is called the Coalescent with Recombination.
- 2 possible events going backwards in time
  - Coalescence: which merges two sequences.
    For i sequences, rate is $i(i-1)/2N$.
  - Recombination: which splits a sequence into two.
    For i sequences, rate is $iL\rho$

Coalescent with recombination

Problem: infer relations along the genome at many scales simultaneously.

- $\mathcal{ARG}$ is a structure (data type).
- The probability distribution over $\mathcal{ARG}$s that arises when recombination is added to the standard (Wright-Fisher) model is called the Coalescent with Recombination.
- 2 possible events going backwards in time
  - Coalescence: which merges two sequences.
    For i sequences, rate is $i(i-1)/2N$.
  - Recombination: which splits a sequence into two.
    For i sequences, rate is $iL\rho$

Approximating Coalescent with Recombination

- Statistical inference under Coalescent with Recombination is very limited due to the complexity of the state space.
- Sequentially Markovian Coalescent approximation (McVean and Cardin, 2005) considers $\mathcal{ARG}$ as a sequence of local trees with a "prune-and-regraft" operation on them. The transitions between trees are Markovian.
- It turns to be a very good approximation with a much more tractable state space.

Approximating Coalescent with Recombination

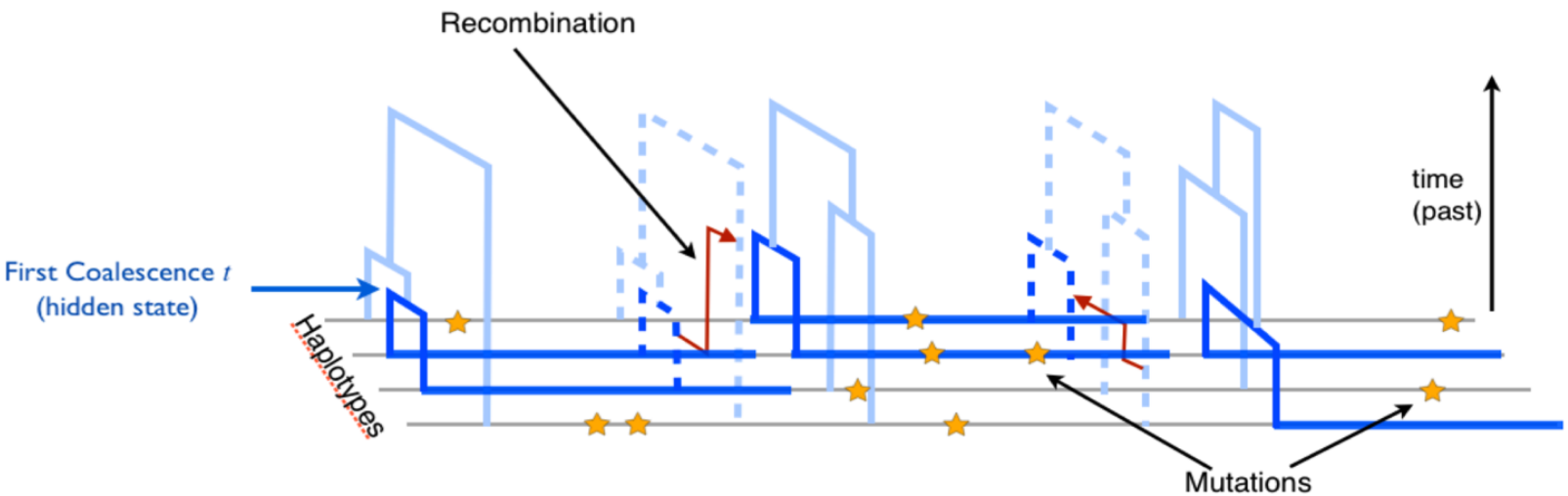
- Statistical inference under Coalescent with Recombination is very limited due to the complexity of the state space.
- Sequentially Markovian Coalescent approximation (McVean and Cardin, 2005) considers $\mathcal{ARG}$ as a sequence of local trees with a "prune-and-regraft" operation on them. The transitions between trees are Markovian.
- It turns to be a very good approximation with a much more tractable state space.

Approximating Coalescent with Recombination

- Statistical inference under Coalescent with Recombination is very limited due to the complexity of the state space.
- Sequentially Markovian Coalescent approximation (McVean and Cardin, 2005) considers $\mathcal{ARG}$ as a sequence of local trees with a "prune-and-regraft" operation on them. The transitions between trees are Markovian.
- It turns to be a very good approximation with a much more tractable state space.

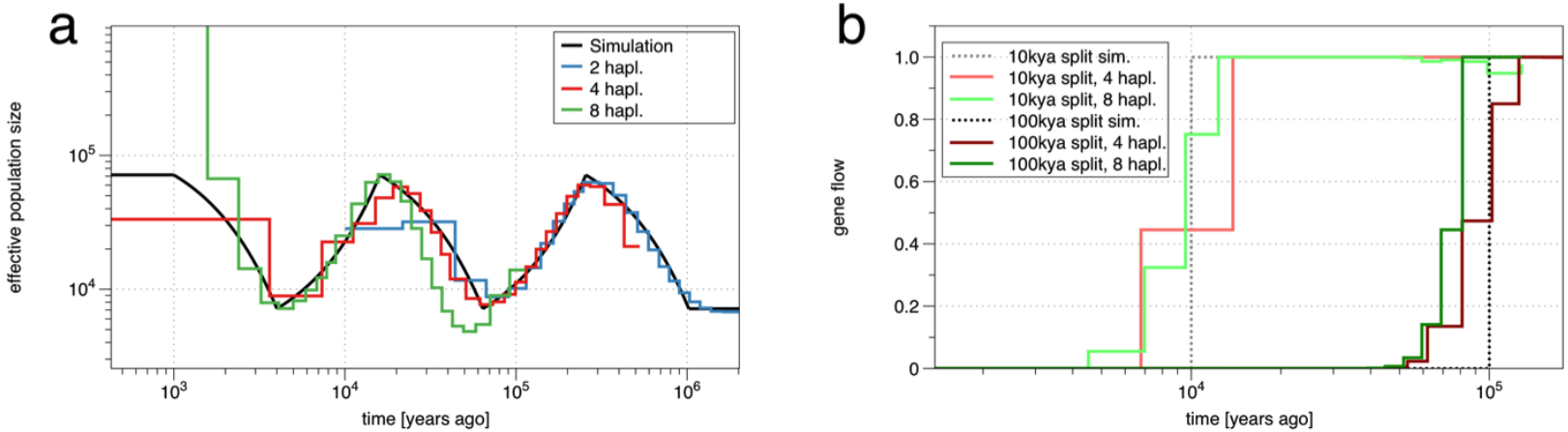
MSMC

- MSMC is the method introduced by S. Schiffels and R. Durbin.
- For a given set with n haplotypes it finds the first coalescence for each segment which allows to infer both population size and separation history.
- It is approximately Markov with the state space $O(n^2 T)$



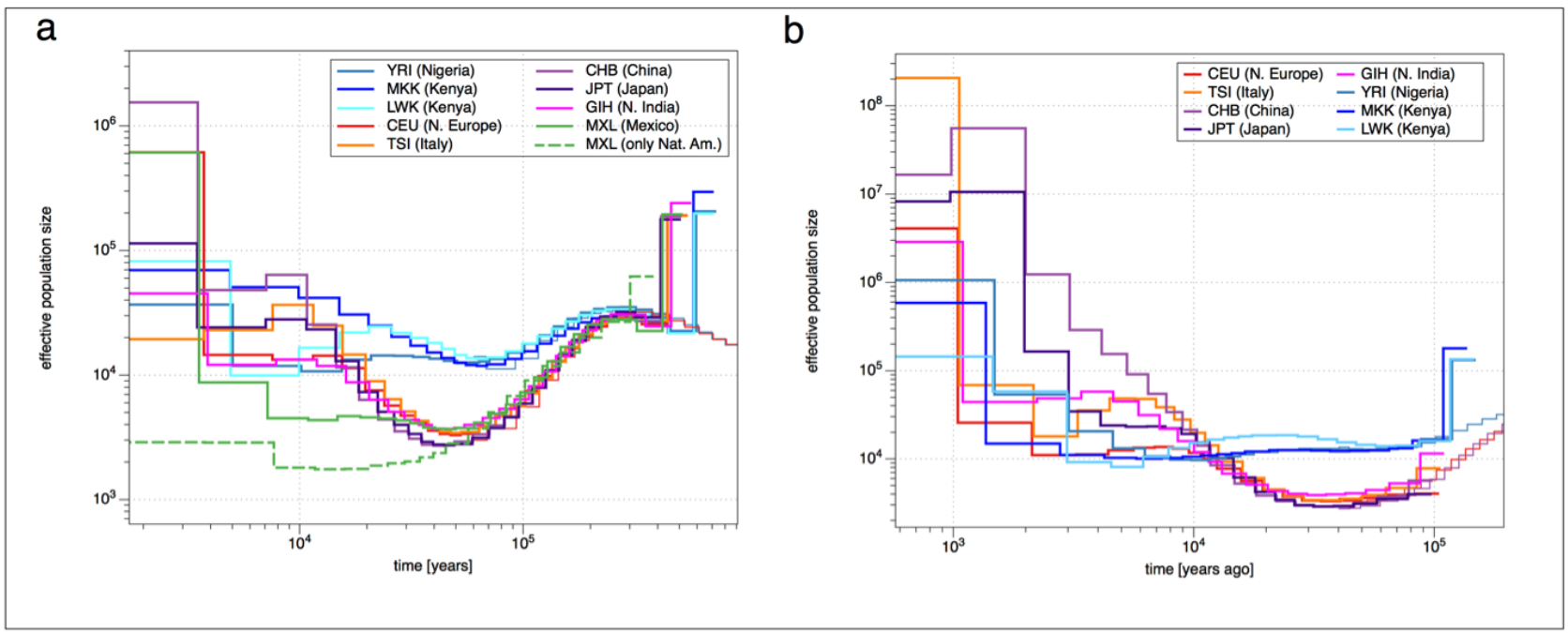
MSMC on simulated data

- Separation is inferred by the ratio of coalescences.
- Here are results for simulated data.



MSMC on real data

The analysis of real data: out of Africa bottleneck and recent fast population growth are supported clearly.



$\mathcal{ARG}$ inference

- Even under SMC approximation the inference of $\mathcal{ARG}$ is computationally hard problem.
- Discretised SMC is an approximation of SMC introduced by Matthew D. Rasmussen, Melissa J. Hubisz, Ilan Gronau, Adam Siepel in 2014.
- To generate an $\mathcal{ARG}$ they use the following strategy implemented in the software called *ARGweaver*. Given an $\mathcal{ARG}$ G^n for sequences $h_1, \ldots, h_n$, sample an extended $\mathcal{ARG}$ $G^{n+1} \supset G^n$ for the same sequences and a new one h_{n+1}.

$\mathcal{ARG}$ inference

- Even under SMC approximation the inference of $\mathcal{ARG}$ is computationally hard problem.
- Discretised SMC is an approximation of SMC introduced by Matthew D. Rasmussen, Melissa J. Hubisz, Ilan Gronau, Adam Siepel in 2014.
- To generate an $\mathcal{ARG}$ they use the following strategy implemented in the software called *ARGweaver*. Given an $\mathcal{ARG}$ G^n for sequences $h_1, \ldots, h_n$, sample an extended $\mathcal{ARG}$ $G^{n+1} \supset G^n$ for the same sequences and a new one h_{n+1}.

$\mathcal{ARG}$ inference

- Even under SMC approximation the inference of $\mathcal{ARG}$ is computationally hard problem.
- Discretised SMC is an approximation of SMC introduced by Matthew D. Rasmussen, Melissa J. Hubisz, Ilan Gronau, Adam Siepel in 2014.
- To generate an $\mathcal{ARG}$ they use the following strategy implemented in the software called *ARGweaver*. Given an $\mathcal{ARG}$ G^n for sequences $h_1, \ldots, h_n$, sample an extended $\mathcal{ARG}$ $G^{n+1} \supset G^n$ for the same sequences and a new one h_{n+1}.

ARGweaver sampling strategies

- One starts with a single sequence and trivial $\mathcal{ARG}$ G^1 and subsequently adds sequences one by one.
- Gibbs sampling: given G^n, remove one sequence and add it again on the resulting G^{n-1}. Internal nodes are poorly mixed by this strategy.
- To overcome this limitation, they allow to rethread subtrees.
- Even with this can only look at ≈ 25 sampled genomes. Mice

Problem: how to analyse $\mathcal{ARG}$? What can be inferred?

ARGweaver sampling strategies

- One starts with a single sequence and trivial $\mathcal{ARG}$ G^1 and subsequently adds sequences one by one.
- Gibbs sampling: given G^n, remove one sequence and add it again on the resulting G^{n-1}. Internal nodes are poorly mixed by this strategy.
- To overcome this limitation, they allow to rethread subtrees.
- Even with this can only look at ≈ 25 samples genome wide.

Problem: how to analyse $\mathcal{ARG}$? What can be inferred?

ARGweaver sampling strategies

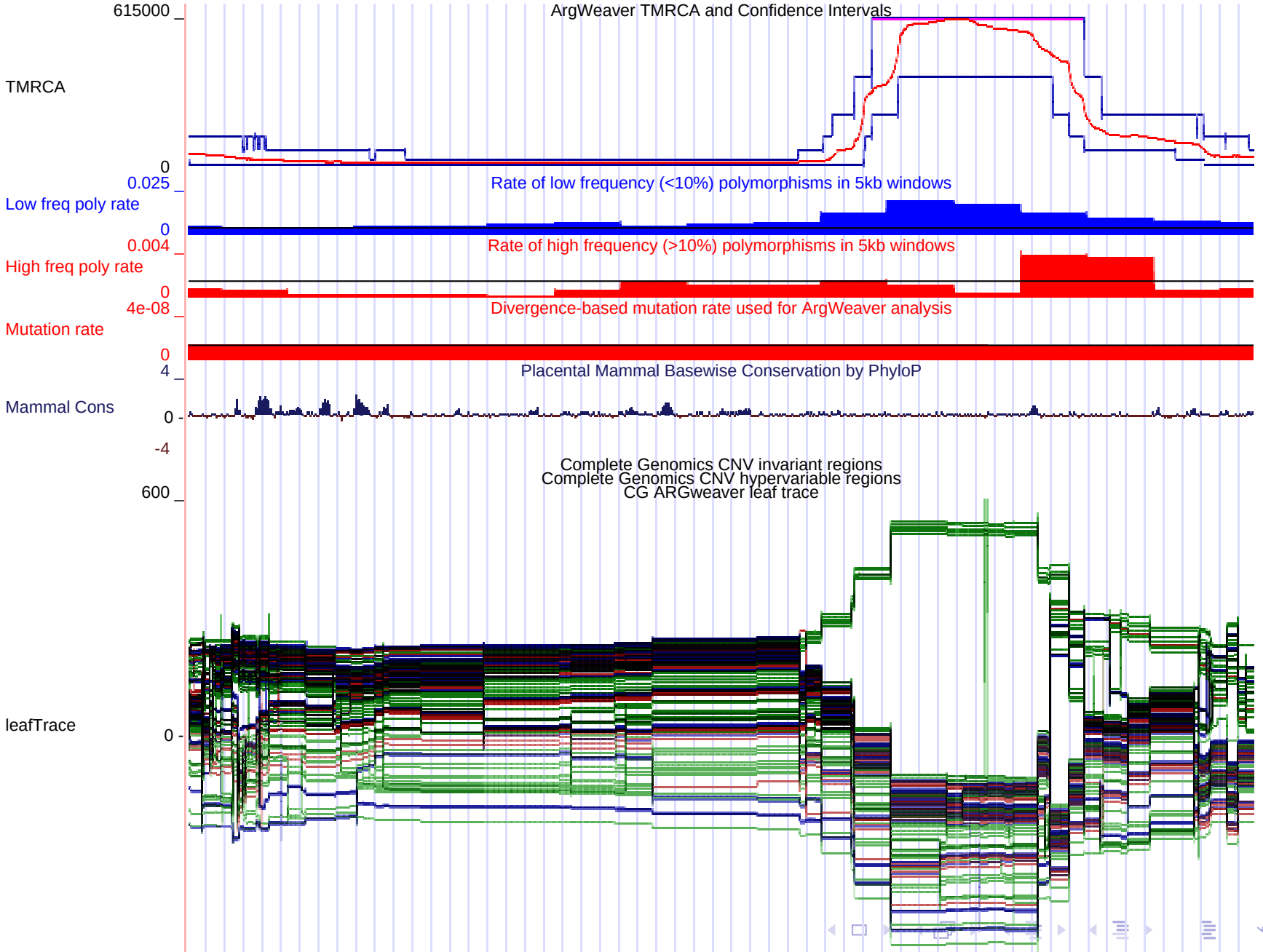
- One starts with a single sequence and trivial $\mathcal{ARG}$ G^1 and subsequently adds sequences one by one.
- Gibbs sampling: given G^n, remove one sequence and add it again on the resulting G^{n-1}. Internal nodes are poorly mixed by this strategy.
- To overcome this limitation, they allow to rethread subtrees.
- Even with this can only look at ≈ 25 samples genome wide.

Problem: how to analyse $\mathcal{ARG}$? What can be inferred?

ARGweaver sampling strategies

- One starts with a single sequence and trivial $\mathcal{ARG}$ G^1 and subsequently adds sequences one by one.
- Gibbs sampling: given G^n, remove one sequence and add it again on the resulting G^{n-1}. Internal nodes are poorly mixed by this strategy.
- To overcome this limitation, they allow to rethread subtrees.
- Even with this can only look at ≈ 25 samples genome wide.

Problem: how to analyse $\mathcal{ARG}$? What can be inferred?


ArgWeaver TMRCA and Confidence Intervals
615000
TMRCA
0
0.025
Rate of low frequency (<10%) polymorphisms in 5kb windows
Low freq poly rate
0
0.004
Rate of high frequency (>10%) polymorphisms in 5kb windows
High freq poly rate
0
4e-08
Divergence-based mutation rate used for ArgWeaver analysis
Mutation rate
0
4
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
0
-4
Complete Genomics CNV invariant regions
Complete Genomics CNV hypervariable regions
CG ARGweaver leaf trace
600
leafTrace
0

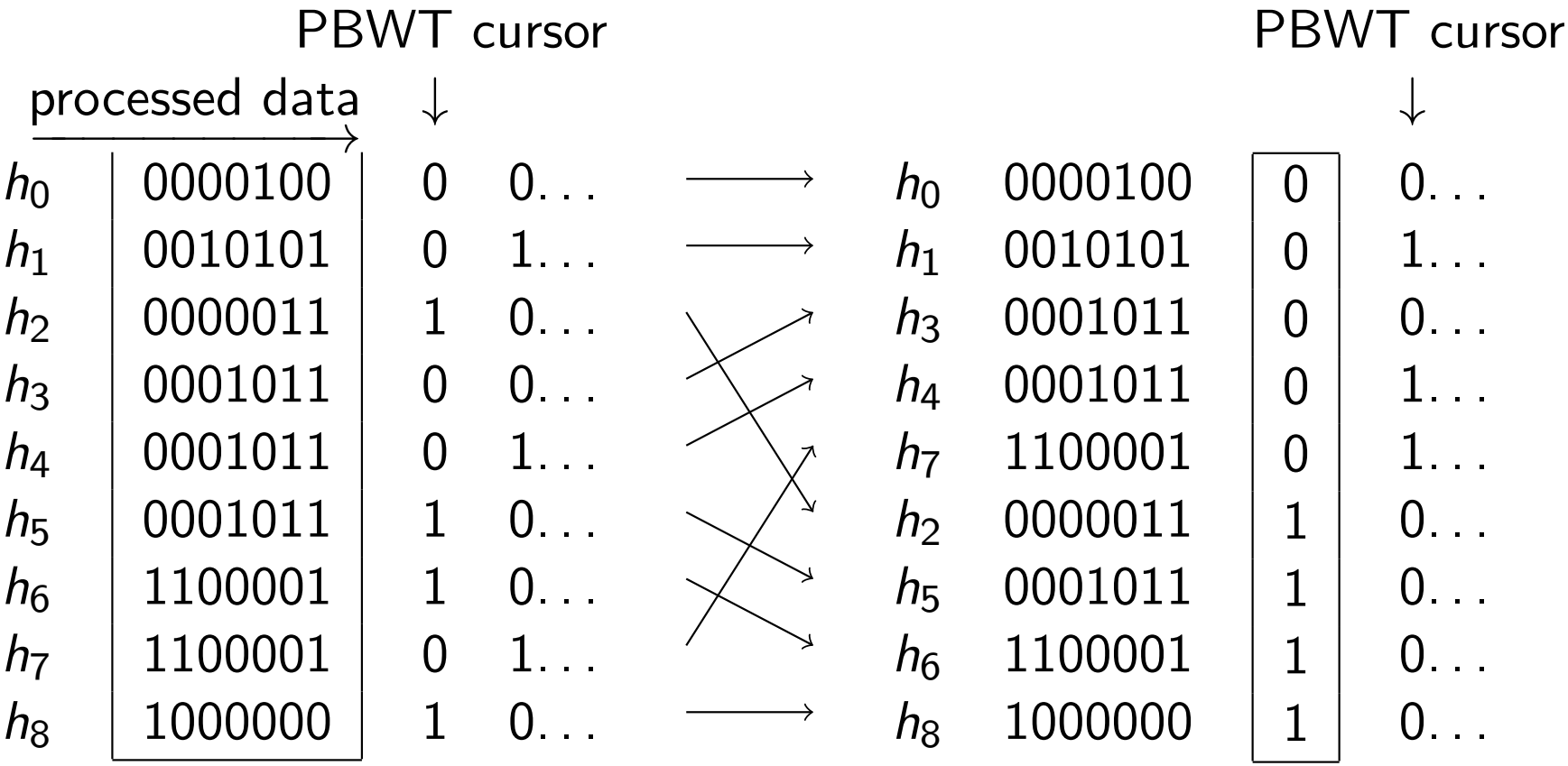
Data structures for $\mathcal{ARG}$

- The $\mathcal{ARG}$ inference is a computationally hard problem.
- Need data structure which allow fast operations on $\mathcal{ARG}$.
- Positional Burrow-Wheeler transform, or shortly $\mathcal{PBWT}$ (R. Durbin, Bioinformatics, 2014) inspires the solution.
- $\mathcal{PBWT}$ is a new data structure for haplotype matching: lexicographical reverse prefix order at every position of genome.

Data structures for $\mathcal{ARG}$

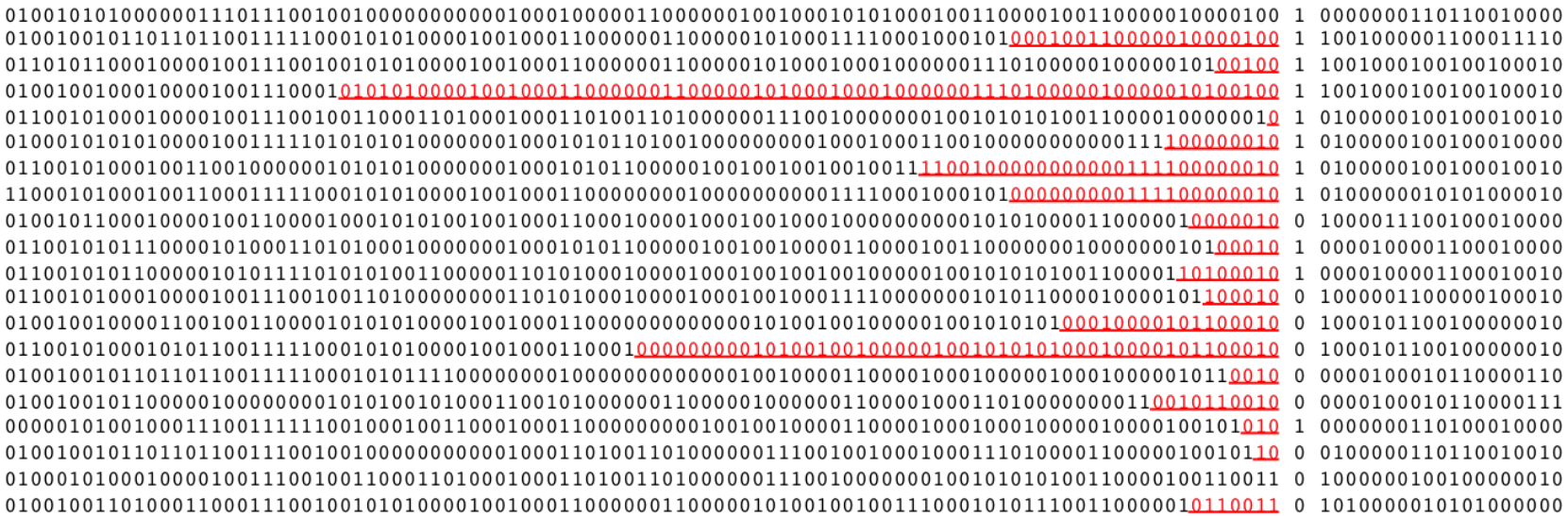
- The $\mathcal{ARG}$ inference is a computationally hard problem.
- Need data structure which allow fast operations on $\mathcal{ARG}$.
- Positional Burrow-Wheeler transform, or shortly $\mathcal{PBWT}$ (R. Durbin, Bioinformatics, 2014) inspires the solution.
- $\mathcal{PBWT}$ is a new data structure for haplotype matching: lexicographical reverse prefix order at every position of genome.

$\mathcal{PBWT}$: example



Positional Burrows Wheeler Transform

The algorithm moves left-to-right, while re-sorting haplotypes at every position following a certain rule.



Matches are adjacent in the sort order.

Data compression

Simulate 100k sequences of length 20Mbp with $\mathcal{ARG}$ simulator MaCS (Chen et al. 2009)

- 370,264 sites (one per 54bp): 37GB raw output
- gzip compresses to 1.02GB (35x compression)
- PBWT compresses to 7.7MB (4800x compresssion)

Real data: 1000 Genomes phase1 chromosome 1

- 2184 chromosomes, 3,007,196 sites
- PBWT 51,186,641
- gzip 302,883,517
- factor 5.9

Data compression

Simulate 100k sequences of length 20Mbp with $\mathcal{ARG}$ simulator MaCS (Chen et al. 2009)

- 370,264 sites (one per 54bp): 37GB raw output
- gzip compresses to 1.02GB (35x compression)
- PBWT compresses to 7.7MB (4800x compresssion)

Real data: 1000 Genomes phase1 chromosome 1

- 2184 chromosomes, 3,007,196 sites
- PBWT 51,186,641
- gzip 302,883,517
- factor 5.9

Tree consistent $\mathcal{PBWT}$

- Tree consistent $\mathcal{PBWT}$ is an evolution of the $\mathcal{PBWT}$ data structure.
- Currently it infers tree topologies.
- Without recombinations, the order converges and the correct underlying genealogy topology.
- Tree is encoded by distances between it leaves (linear memory).
- Basic strategy:
  - Assume infinite site model: at most one mutation at site.
  - If a column inconsistent with a tree, divide it to maximal consistent branches.
  - Rebuild the tree with minimal pruning and regrafting operations.

Tree consistent $\mathcal{PBWT}$

- Tree consistent $\mathcal{PBWT}$ is an evolution of the $\mathcal{PBWT}$ data structure.
- Currently it infers tree topologies.
- Without recombinations, the order converges and the correct underlying genealogy topology.
- Tree is encoded by distances between it leaves (linear memory).
- Basic strategy:
  - Assume infinite site model: at most one mutation at site.
  - If a column inconsistent with a tree, divide it in maximal consistent branches.
  - Rebuild the tree with minimal prune-and-regraft operations.
- All operations can be performed with either linear or quadratic complexity in the number of tree leaves (same for the data structure).

Tree consistent $\mathcal{PBWT}$

- Tree consistent $\mathcal{PBWT}$ is an evolution of the $\mathcal{PBWT}$ data structure.
- Currently it infers tree topologies.
- Without recombinations, the order converges and the correct underlying genealogy topology.
- Tree is encoded by distances between it leaves (linear memory).
- Basic strategy:
  - Assume infinite site model: at most one mutation at site.
  - If a column inconsistent with a tree, divide it in maximal consistent branches.
  - Rebuild the tree with minimal prune-and-regraft operations.
- All operations can be performed with either linear or sublinear complexity in the number of tree leaves using the data structure.

Tree consistent $\mathcal{PBWT}$

- Tree consistent $\mathcal{PBWT}$ is an evolution of the $\mathcal{PBWT}$ data structure.
- Currently it infers tree topologies.
- Without recombinations, the order converges and the correct underlying genealogy topology.
- Tree is encoded by distances between it leaves (linear memory).
- Basic strategy:
  - Assume infinite site model: at most one mutation at site.
  - If a column inconsistent with a tree, divide it in maximal consistent branches.
  - Rebuild the tree with minimal prune-and-regraft operations.
- All operations can be performed with either linear or sublinear complexity in the number of tree leaves using the data structure.

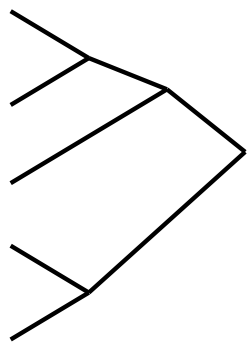
Tree consistent $\mathcal{PBWT}$

- Tree consistent $\mathcal{PBWT}$ is an evolution of the $\mathcal{PBWT}$ data structure.
- Currently it infers tree topologies.
- Without recombinations, the order converges and the correct underlying genealogy topology.
- Tree is encoded by distances between it leaves (linear memory).
- Basic strategy:
  - Assume infinite site model: at most one mutation at site.
  - If a column inconsistent with a tree, divide it in maximal consistent branches.
  - Rebuild the tree with minimal prune-and-regraft operations.
- All operations can be performed with either linear or sublinear complexity in the number of tree leaves using the data structure.

$tc\mathcal{PBWT}$: perfect phylogeny example and run-time

Perfect phylogeny: $tc\mathcal{PBWT}$ will find the correct topology in linear time.

	B_1	B_2	B_3	B_4	B_5	encoded tree
h_0	1	0	1	0	0	-
h_1	1	0	1	1	0	1
h_2	0	0	1	0	0	2
h_3	0	1	0	0	1	3
h_4	0	1	0	0	0	1

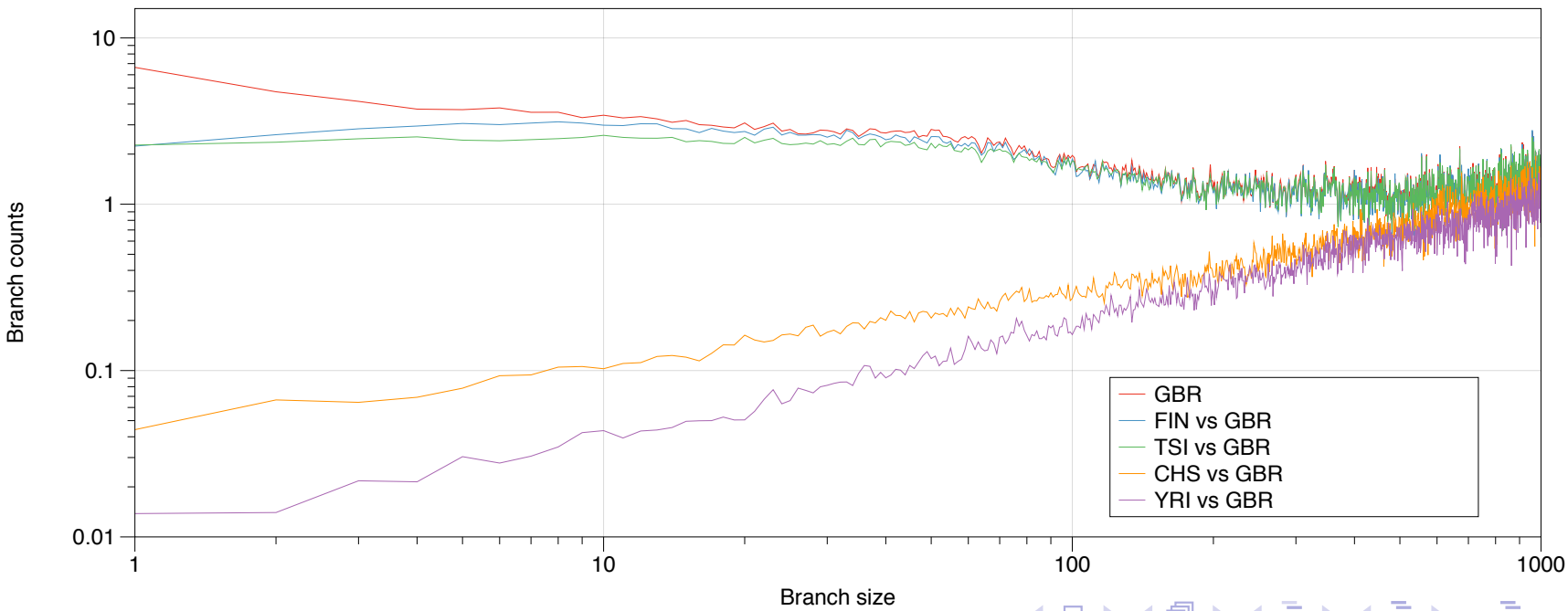


To generate an $\mathcal{ARG}$ for chromosome 20 of 1000 Genome Project, phase 3 (5006 sequences, $\approx$ 860 thousands SNPs) it took $\approx$10 minutes.

$tc\mathcal{PBWT}$: conclusions

- $tc\mathcal{PBWT}$ reduces state space of $\mathcal{ARG}$s.
- It provides a new framework for scalable $\mathcal{ARG}$ generation.
- Tree are encoded and updated in terms of distance which encourages to search for algorithms for inference of times of events.

Population analysis based on an $\mathcal{ARG}$ inferred by $tc\mathcal{PBWT}$.



Li and Stephens model

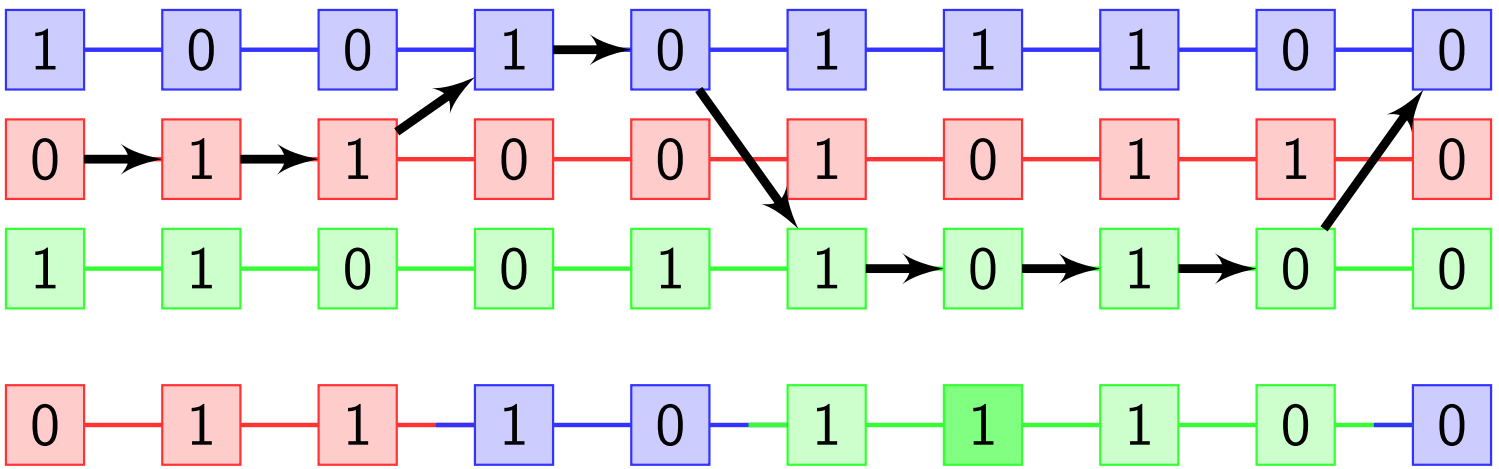
- As we have just seen, inference under coalescent is a challenging problems. Li and Stephens model (or copying model) is a Hidden Markov Model-based approach which allows to infer full likelihood straightforwardly.
- Given a reference panel $h_1, \ldots, h_n$ of haplotypes, a new haplotype h_{n+1} is generated:
  - Select randomly a haplotype.
  - Copy the first locus with a small probability of error.
  - With high probability the next locus is copied from the same haplotype.
  - Alternatively, the haplotype for copying is re-chosen, uniformly and independently.

Li and Stephens model

- As we have just seen, inference under coalescent is a challenging problems. Li and Stephens model (or copying model) is a Hidden Markov Model-based approach which allows to infer full likelihood straightforwardly.
- Given a reference panel $h_1, \ldots, h_n$ of haplotypes, a new haplotype h_{n+1} is generated:
  - Select randomly a haplotype.
  - Copy the first locus with a small probability of error.
  - With high probability the next locus is copied from the same haplotype.
  - Alternatively, the haplotype for copying is re-chosen, uniformly and independently.

Li and Stephens model

- Mutations are incorporated in the model as copying errors.
- Recombinations are presented as reselecting haplotypes.
- Li and Stephens model is a Hidden Markov chain with transition probabilities corresponding to reselecting haplotypes and with emissions corresponding to copying or miscopying alleles.
- Computation of maximum likelihood of a set of haplotypes is straightforward under this model.



Summary

- Large data sets appear in genomic nowadays.
- Statistical inference is challenging due to the huge state-space of underlying models and the data size.
- New data structures and algorithms are needed to make data processing scalable and precise in the same time.