Roadmap of the problems of modern genomics

Vladimir Shchur

16th May 2018, Moscow

Talk overview

- What is a genome?
- Sequencing: the way we observe the data.
- Population genomics: studying natural processes which creates our data and inferring history from the genomes.
  - Forwards in time models.
  - Backwards in time models.
  - Methods for population history inference.
- Medical and other applications: Genome Wide Association Study (GWAS), cancer genomics and metagenomics.

What is a genome?

- 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*.

What is a genome?

- 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*.

What is a genome?

- 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 30 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 11.4 million of Single Nucleotide Polymorphisms (SNP) are validated according to dbSNP (Database 2012).

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 114 millions of Single Nucleotide Polymorphisms (SNP) are validated according to *dbSNP* (October 2017).

Some numbers

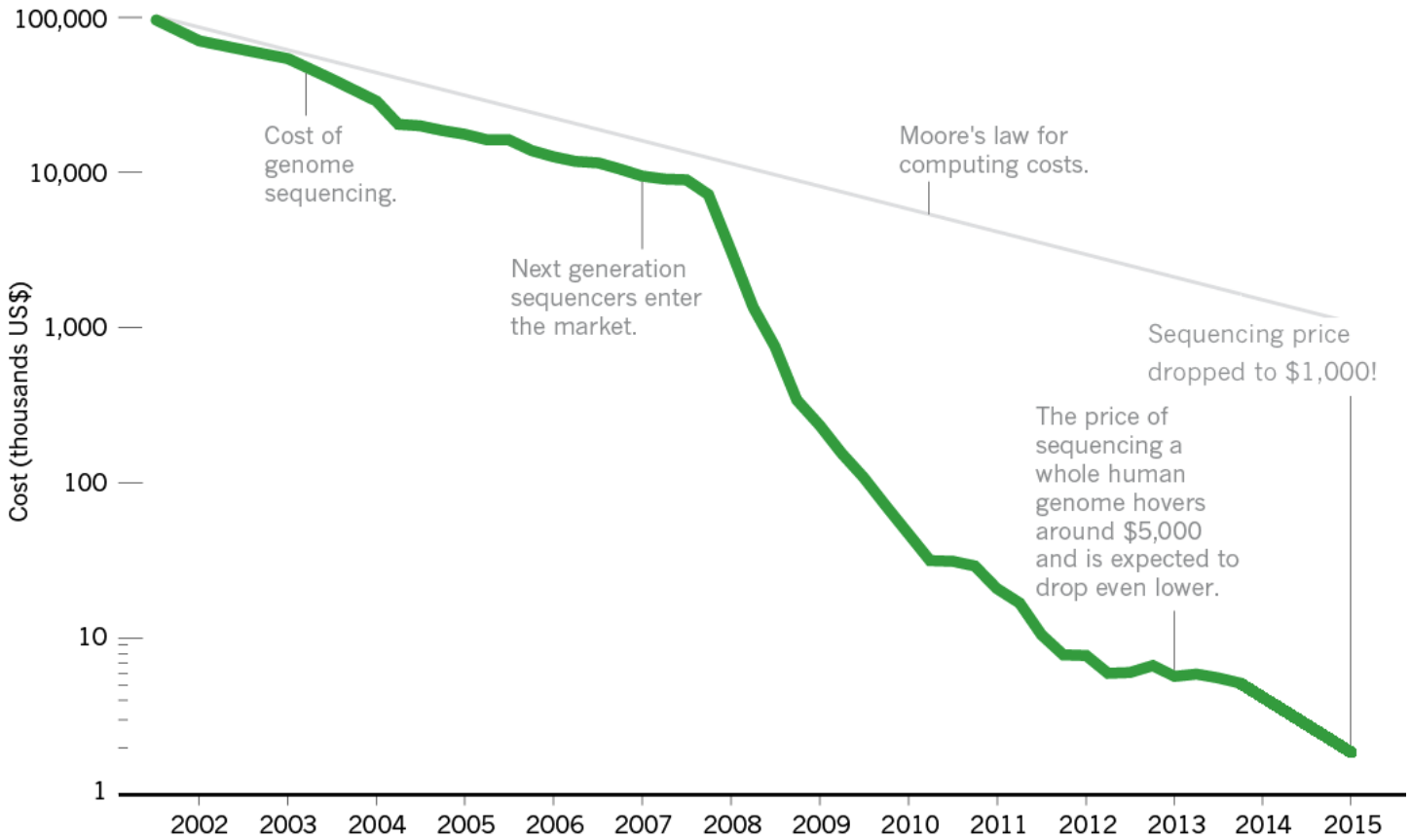
- Human genome length $\approx$ 3Gb (Giga-basepairs).
- There are $\approx$ 3 million differences between two typical human haplotypes, e.g. maternal and paternal versions in one person.
- Most of these are shared with other people, caused by mutations in the distant past, 10s or 100s of thousands of 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 114 millions of Single Nucleotide Polymorphisms (SNP) are validated according to *dbSNP* (October 2017).

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 114 millions of Single Nucleotide Polymorphisms (SNP) are validated according to *dbSNP* (October 2017).

Sequencing costs

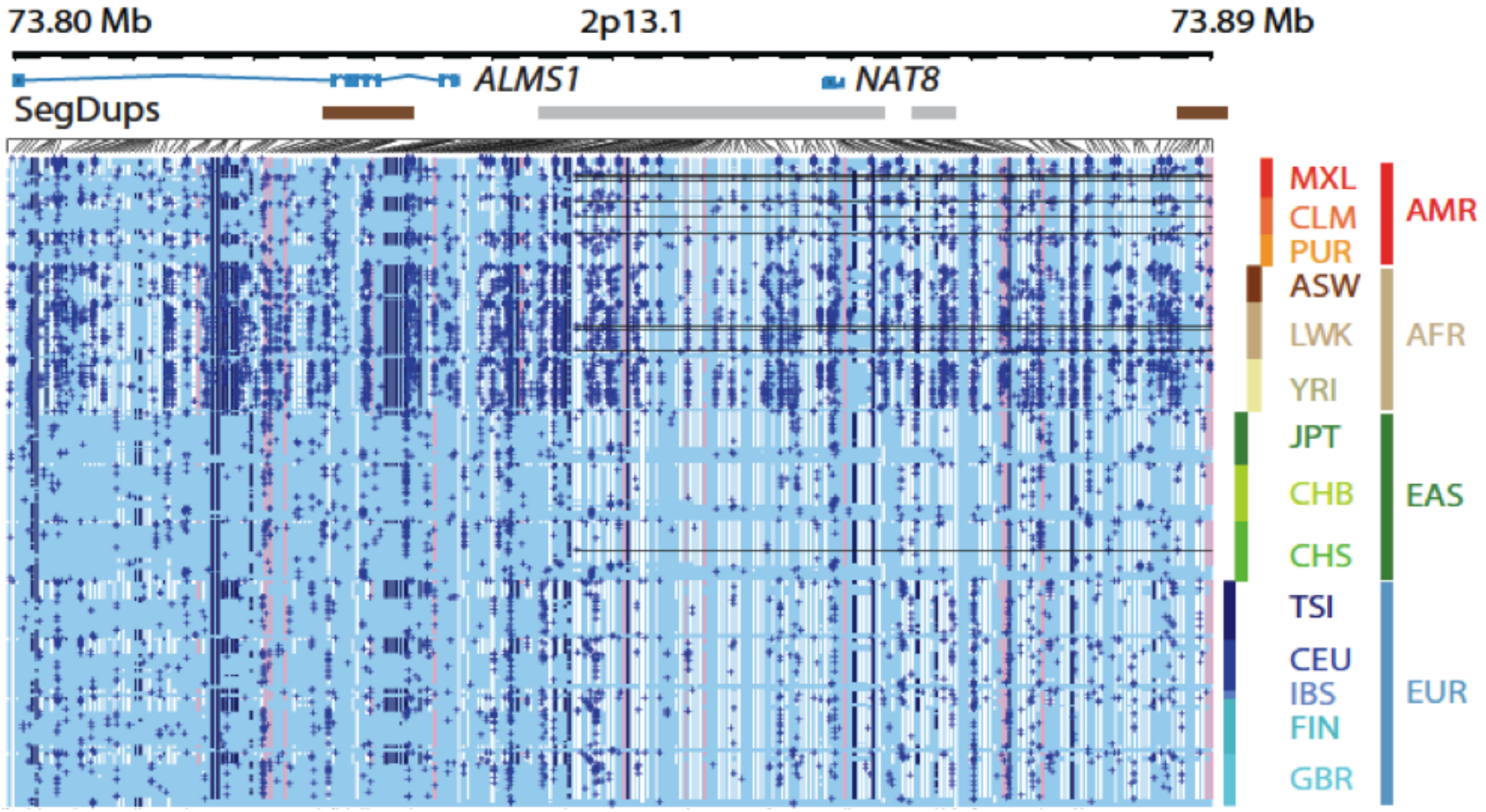
Sequencing price reduced dramatically which allow to create huge genomic data bases.



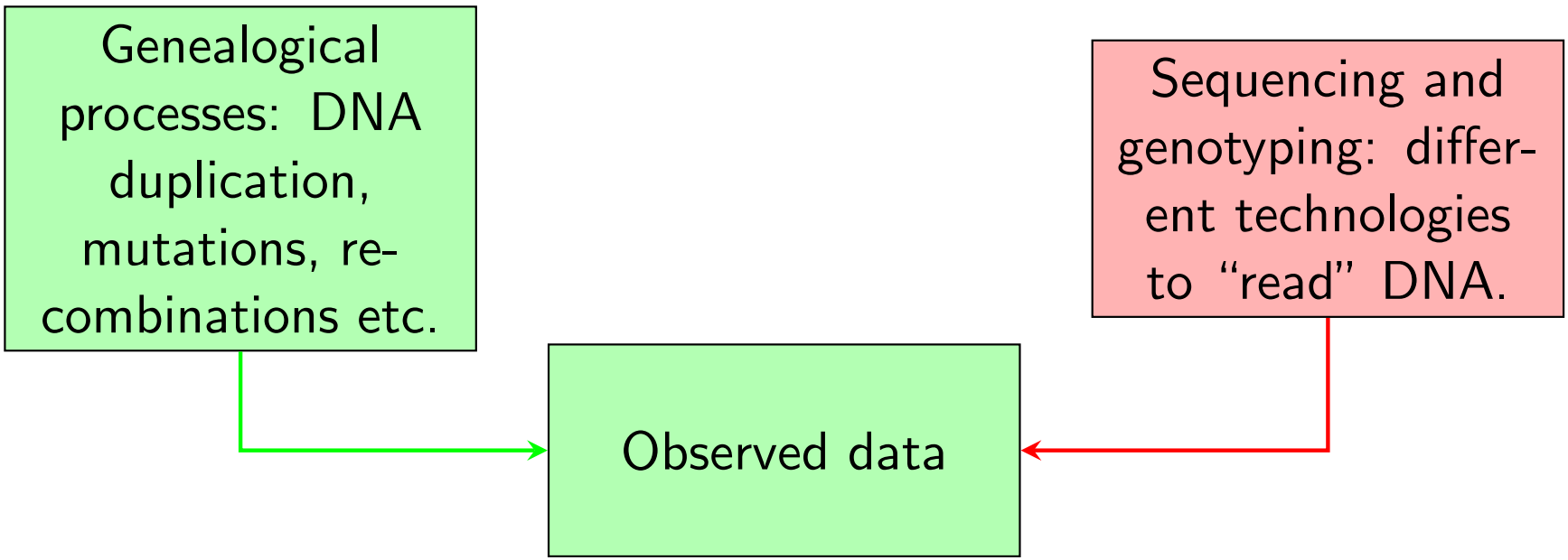
Erika Check Hayden, Technology: The $1,000 genome. Nature, 2014

1000 Genome projects

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).

- 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:

$$\begin{array}{cc} G & C \\ A & T \end{array} \quad \text{or} \quad \begin{array}{cc} G & T \\ A & C \end{array}$$

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).

- 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:

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

- Imputation: the way to treat missing 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).

- 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:

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

- Imputation: the way to treat missing 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).

- 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:

...G...A...	...G...T...
...C...T...	...C...A...

- Imputation: the way to treat missing 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).

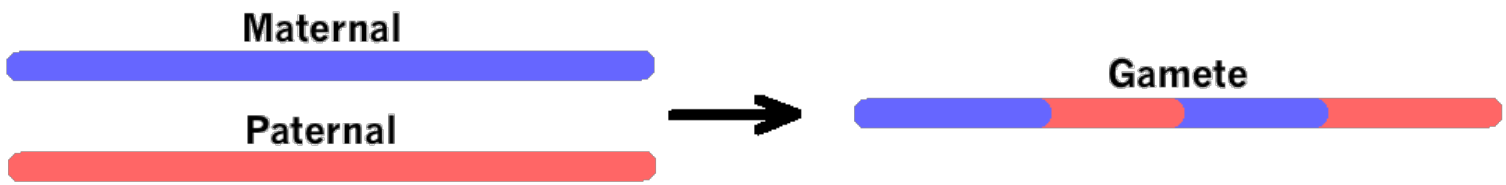
- De-novo assembly.
- Read alignment against reference genome.
- Variant calling: are there enough reads that support the variant?
- Phasing: if SNP_1 carries variants G and A and SNP_2 with C and T, there are two possible genomes which can underly the data:

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

- Imputation: the way to treat missing data.

Genealogical processes: from molecular level ...

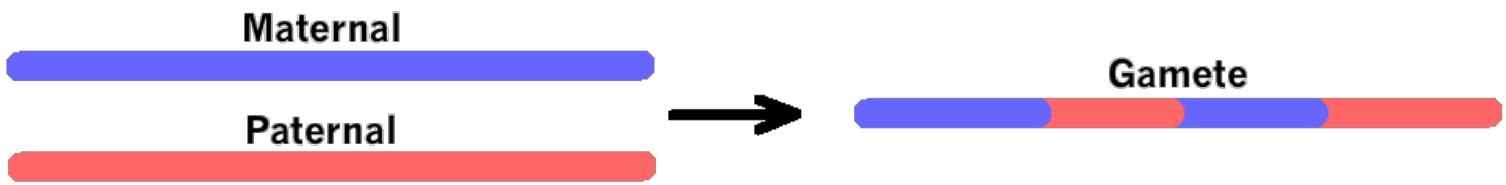
- All the life reproduction is based on cell division. Genetic material is duplicated during this process.
- Errors can occur during duplication. It can be a single nucleotide polymorphism (SNP), insertions, deletions and some other.
- Human gametes (reproduction cells) contain only one set of chromosomes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



Problem: Estimate mutation and recombination rates.

Genealogical processes: from molecular level ...

- 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: from molecular level ...

- 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: ... to population study

Problem: what is the relation between population history and 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.

Genealogical processes: ... to population study

Problem: what is the relation between population history and 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.

Genealogical processes: ... to population study

Problem: what is the relation between population history and 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.

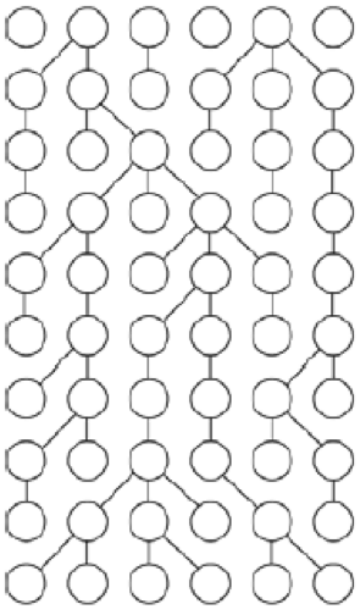
Genealogical processes: ... to population study

Problem: what is the relation between population history and 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.

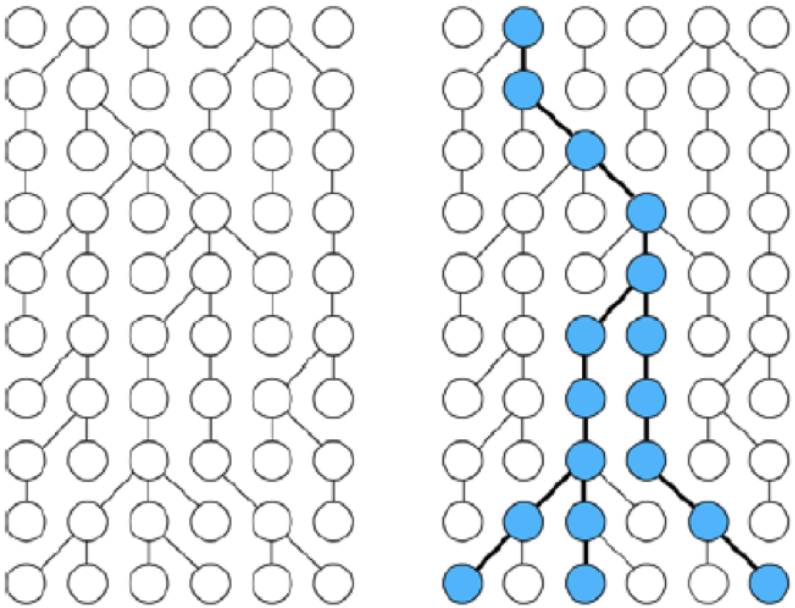
Relationship between forwards in time (Wright-Fisher) and backwards in time (Coalescent) models

- 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.



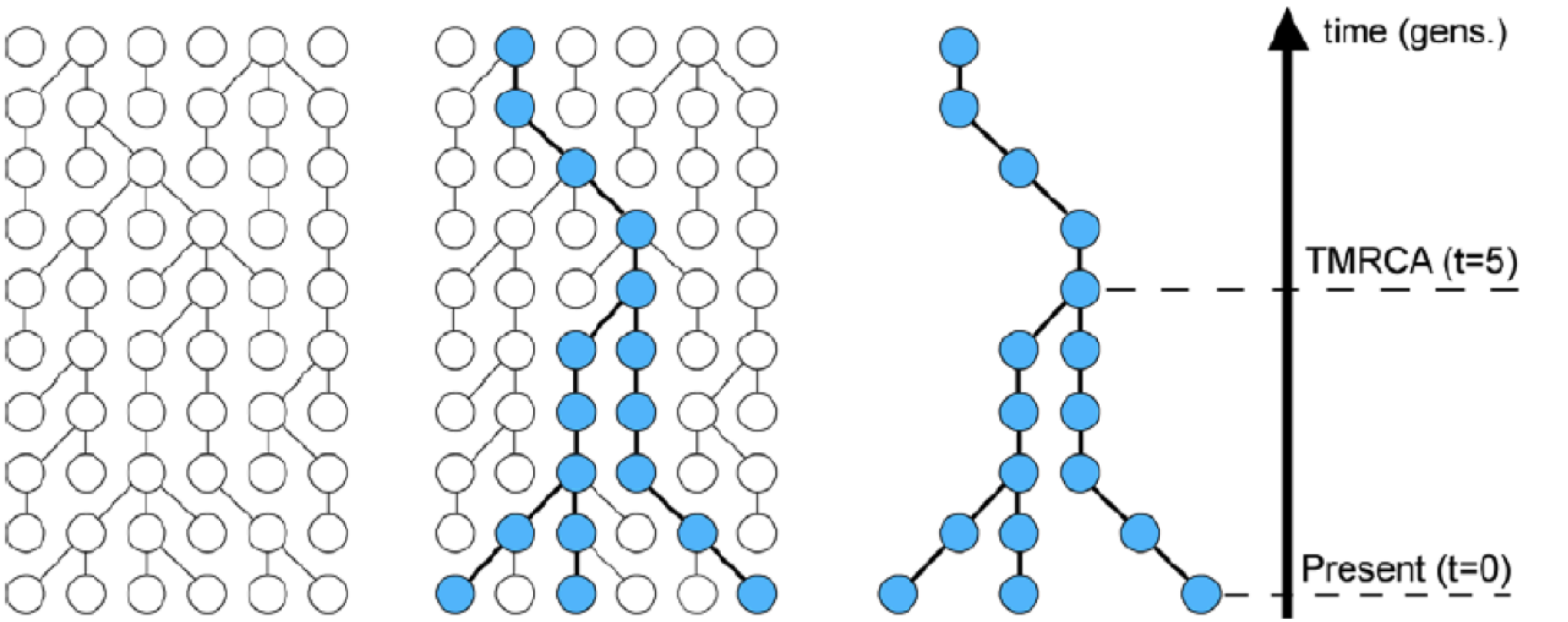
Relationship between forwards in time (Wright-Fisher) and backwards in time (Coalescent) models

- 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: inferring population history and structure from genomes (modern and ancient).

- Coalescent model is a limiting distribution which is consistent with forwards in time models for large effective population size.
- Lineages coalesce according to a Poisson process with parameter proportional to the scaled effective population size.
- If effective population size is constant, 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 one-site theory test.

Coalescent model

Problem: inferring population history and structure from genomes (modern and ancient).

- Coalescent model is a limiting distribution which is consistent with forwards in time models for large effective population size.
- Lineages coalesce according to a Poisson process with parameter proportional to the scaled effective population size.
- If effective population size is constant, 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: inferring population history and structure from genomes (modern and ancient).

- Coalescent model is a limiting distribution which is consistent with forwards in time models for large effective population size.
- Lineages coalesce according to a Poisson process with parameter proportional to the scaled effective population size.
- If effective population size is constant, 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.

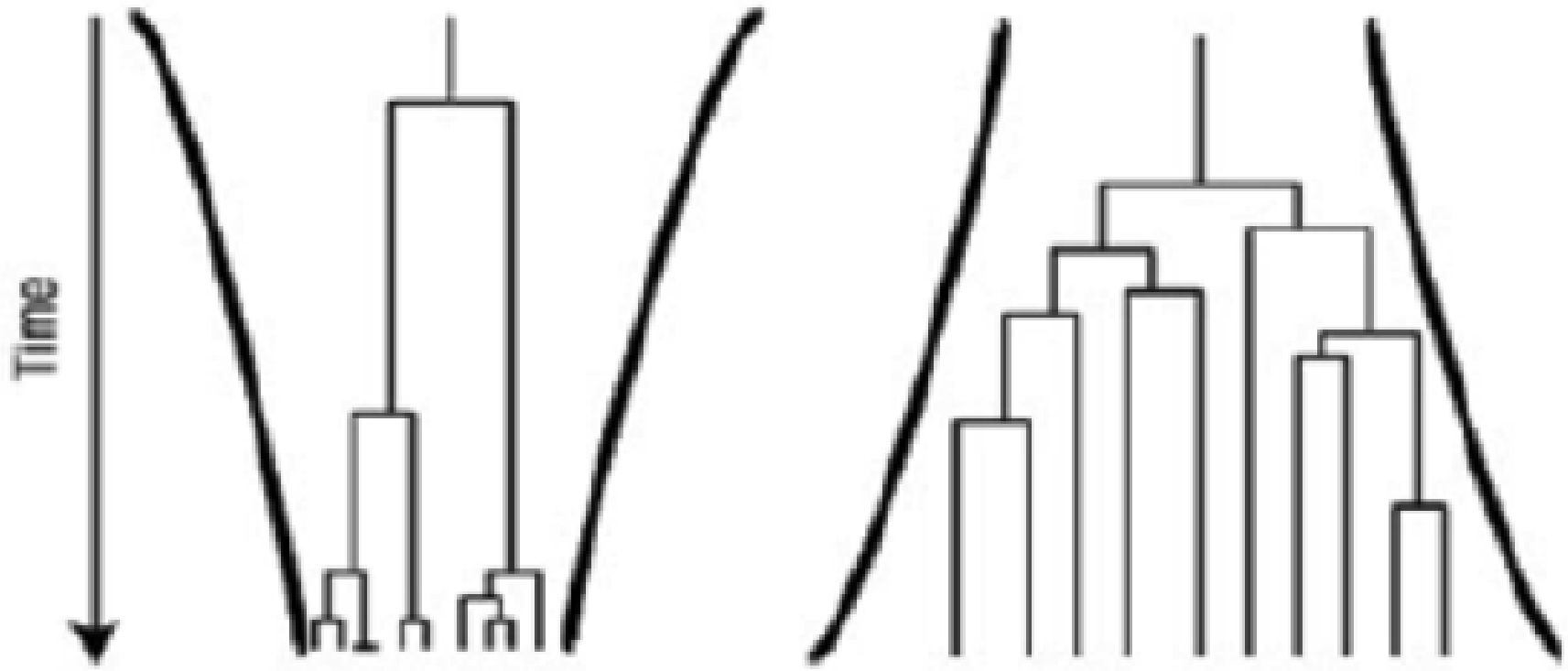
Coalescent model

Problem: inferring population history and structure from genomes (modern and ancient).

- Coalescent model is a limiting distribution which is consistent with forwards in time models for large effective population size.
- Lineages coalesce according to a Poisson process with parameter proportional to the scaled effective population size.
- If effective population size is constant, 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.



Population structure of recent generations

- Coalescent model works fine for large time scales because it is a limiting distribution.
- It does not work for very recent past.
- For example, characterising the number of siblings or cousins in a sample requires working with Wright-Fisher model [Shchur, Nielsen, under reviewing].
- For a random mating population with effective population size $2N$, in a sample of size K, the number of individuals without siblings of cousins (in the same sample) is approximately

$$e^{-(2^{2p-1})K/N}.$$

With a fast growth of data sets, it is important to correct our studies (e.g. GWAS) for relatives.

Example from criminalistics

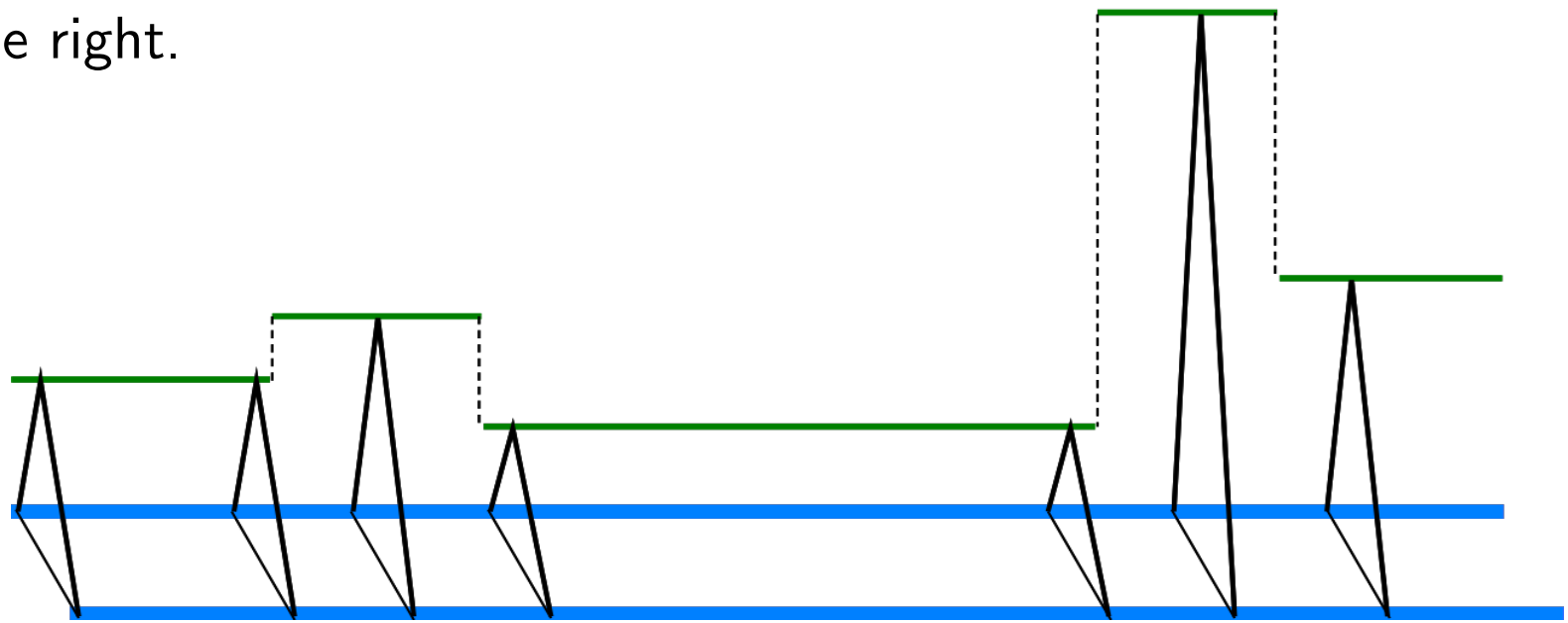
- On April 24, 2018, the Sacramento County Sheriff's Department arrested J. DeAngelo (72 y.o.) known as the Golden State Killer who committed more than 150 crimes from 1974 through 1986.
- Law enforcement uploaded the Golden State Killer's DNA profile to personal genomics website GEDmatch.
- The website identified a dozen of distant relatives of the Golden State Killer. The investigators then traced the family tree to the main suspect.
- Our manuscript was used in a blogpost by Prof. Graham Coop (UC Davis) to analyse how lucky the genetic investigation was.

Adding 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.

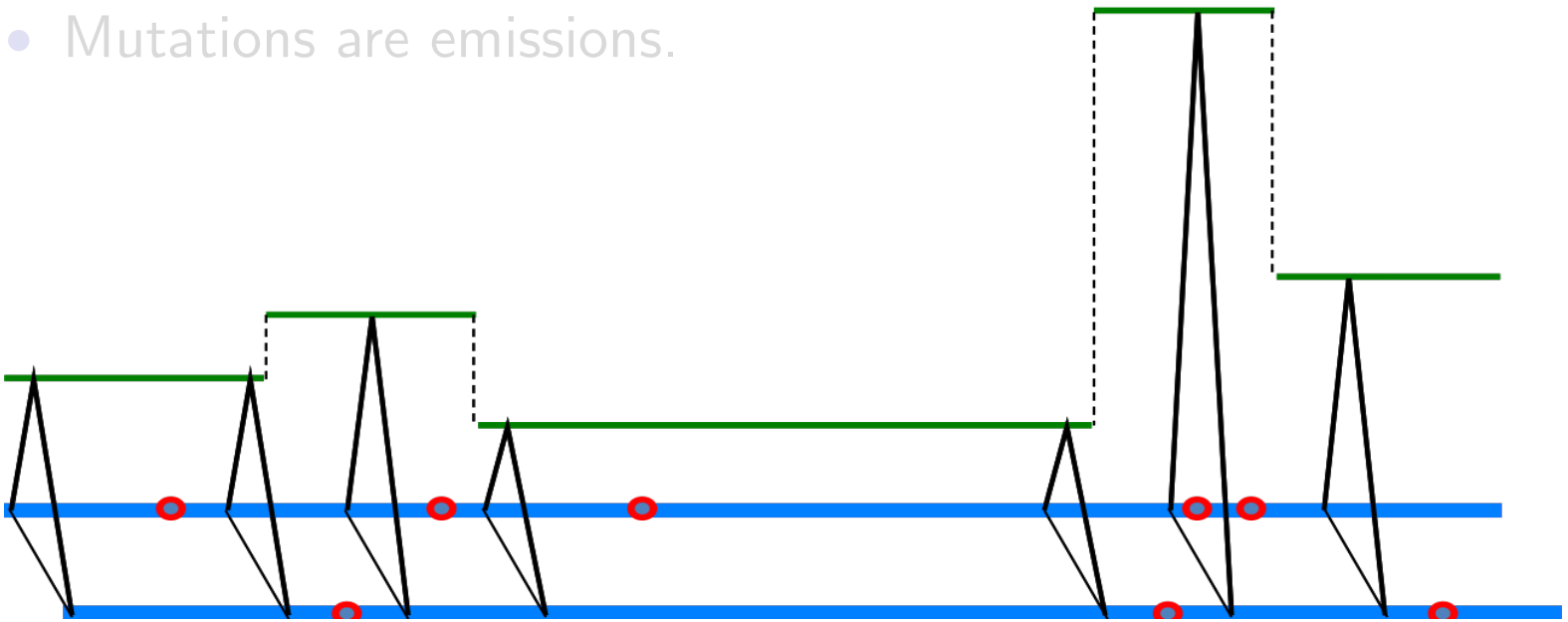
Problem: What happens in between?

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.



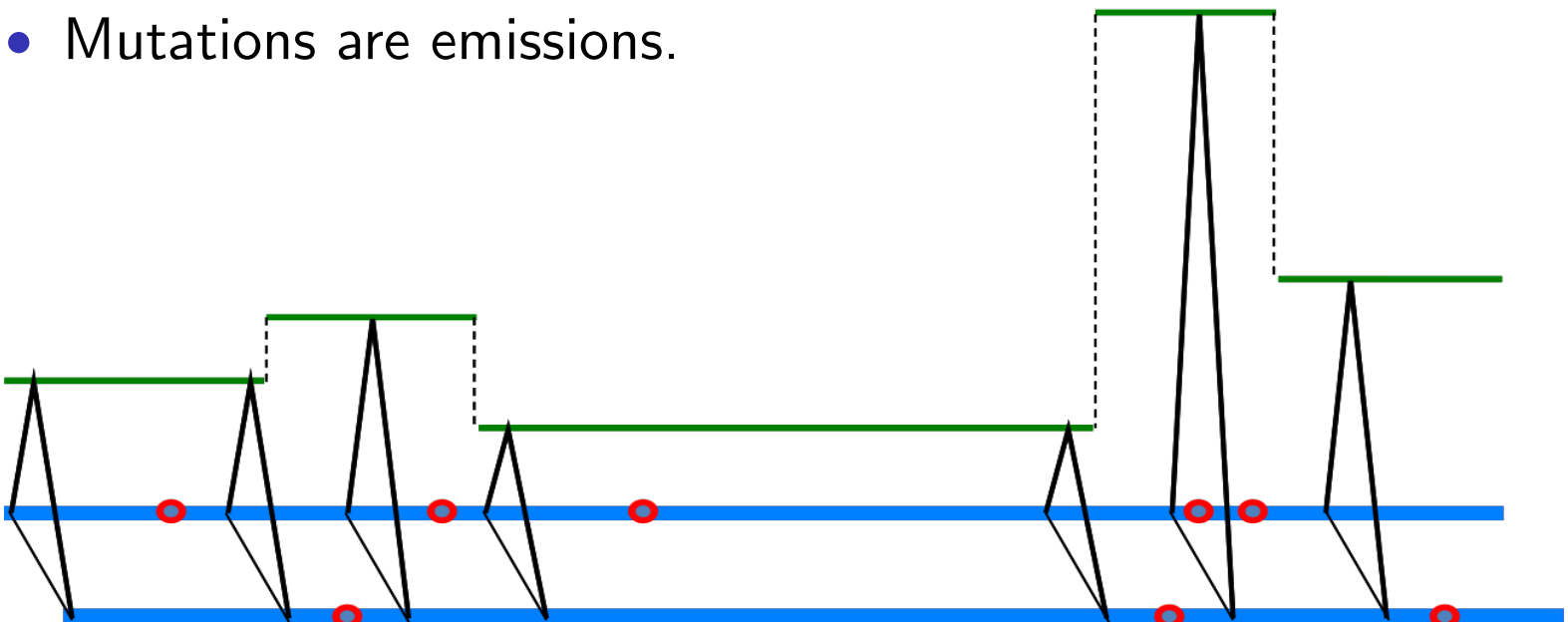
Pairwise Sequentially Markovian Approximation to the Coalescent (PSMC)

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



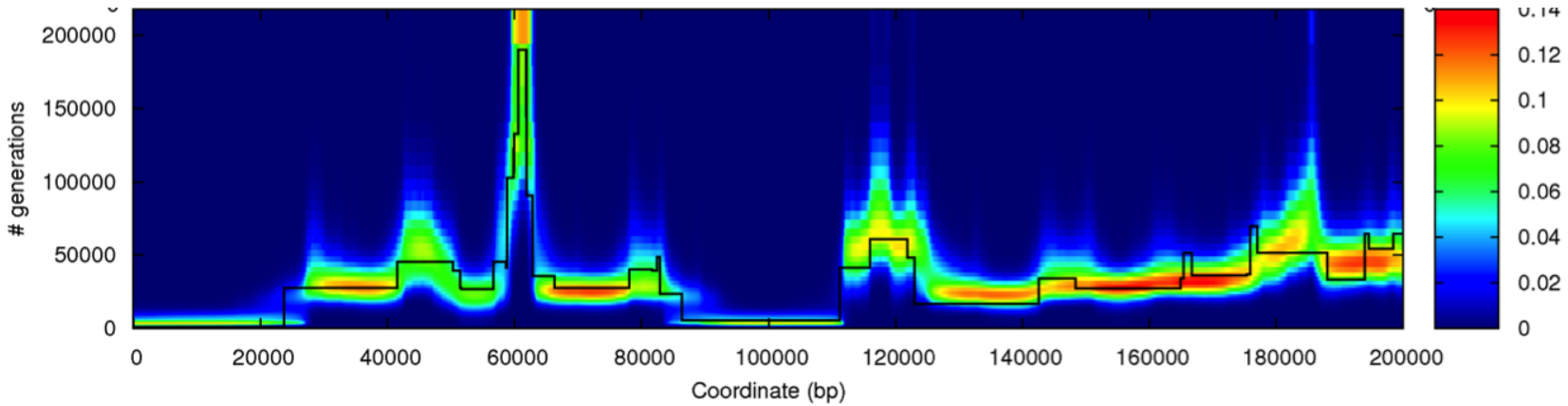
Pairwise Sequentially Markovian Approximation to the Coalescent (PSMC)

- For two haplotypes, the tree is very simple. Recombinations change its height.
- Local trees are states of a Hidden Markov Model (H. Li, R. Durbin)
- 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.



PSMC for low quality data

- To call SNPs confidently, we need expensive high-coverage sequencing.
- Genotype likelihoods allow to work with low quality data, incorporate uncertainty for many different types of errors.
- We are developing a new version of PSMC with emissions based on genotype likelihoods (with Thorfinn Sand, University of Copenhagen).
- We are also developing a method to infer population rates and split time from PSMC on two different genomes by fitting allele frequency spectrum.

PSMC for low quality data

- To call SNPs confidently, we need expensive high-coverage sequencing.
- Genotype likelihoods allow to work with low quality data, incorporate uncertainty for many different types of errors.
- We are developing a new version of PSMC with emissions based on genotype likelihoods (with Thorfinn Sand, University of Copenhagen).
- We are also developing a method to infer migration rates and split time from PSMC on two different genomes by fitting allele frequency spectrum.

PSMC for low quality data

- To call SNPs confidently, we need expensive high-coverage sequencing.
- Genotype likelihoods allow to work with low quality data, incorporate uncertainty for many different types of errors.
- We are developing a new version of PSMC with emissions based on genotype likelihoods (with Thorfinn Sand, University of Copenhagen).
- We are also developing a method to infer migration rates and split time from PSMC on two different genomes by fitting allele frequency spectrum.

PSMC for low quality data

- To call SNPs confidently, we need expensive high-coverage sequencing.
- Genotype likelihoods allow to work with low quality data, incorporate uncertainty for many different types of errors.
- We are developing a new version of PSMC with emissions based on genotype likelihoods (with Thorfinn Sand, University of Copenhagen).
- We are also developing a method to infer migration rates and split time from PSMC on two different genomes by fitting allele frequency spectrum.

Genome Wide Association Study (GWAS)

- Study design: typically it is a comparison of a control group against a case group.
- Mendelian traits are relatively easy to discover (e.g. lactase persistance).
- Some traits (e.g. height or schizophrenia) can be caused by both genetic and environmental factors.
- Epistasis: non-linear interaction between alleles in the same or different genes and even non-coding regions.

Genome Wide Association Study (GWAS)

- Study design: typically it is a comparison of a control group against a case group.
- Mendelian traits are relatively easy to discover (e.g. lactase persistance).
- Some traits (e.g. height or schizophrenia) can be caused by both genetic and environmental factors.
- Epistasis: non-linear interaction between alleles in the same or different genes and even non-coding regions.
- Joe Pickrell's idea (genome.org): sequence an average quality but huge amount of individuals.

Genome Wide Association Study (GWAS)

- Study design: typically it is a comparison of a control group against a case group.
- Mendelian traits are relatively easy to discover (e.g. lactase persistance).
- Some traits (e.g. height or schizophrenia) can be caused by both genetic and environmental factors.
- Epistasis: non-linear interaction between alleles in the same or different genes and even non-coding regions.
- Joe Pickrell's idea (gencove.com): sequence at very low quality, but huge amount of individuals.

Genome Wide Association Study (GWAS)

- Study design: typically it is a comparison of a control group against a case group.
- Mendelian traits are relatively easy to discover (e.g. lactase persistance).
- Some traits (e.g. height or schizophrenia) can be caused by both genetic and environmental factors.
- Epistasis: non-linear interaction between alleles in the same or different genes and even non-coding regions.
- Joe Pickrell's idea (gencove.com): sequence at very low quality, but huge amount of individuals.

Genome Wide Association Study (GWAS)

- Study design: typically it is a comparison of a control group against a case group.
- Mendelian traits are relatively easy to discover (e.g. lactase persistance).
- Some traits (e.g. height or schizophrenia) can be caused by both genetic and environmental factors.
- Epistasis: non-linear interaction between alleles in the same or different genes and even non-coding regions.
- Joe Pickrell's idea (gencove.com): sequence at very low quality, but huge amount of individuals.

Cancer genomics

- Cancer genomics is our main hope to treat cancer efficiently.
- This vast field includes population genomics (study of tumour evolution) and GWAS (identification of driver mutations).
- This knowledge can help in diagnostics (blood test for specific genetic markers), treatment planning (tumours with different mutations have different response to treatments).
- One of the most promising treatments is the viral therapy.

Cancer genomics

- Cancer genomics is our main hope to treat cancer efficiently.
- This vast field includes population genomics (study of tumour evolution) and GWAS (identification of driver mutations).
- This knowledge can help in diagnostics (blood test for specific genetic markers), treatment planning (tumours with different mutations have different response to treatments).
- One of the most promising treatments is the viral therapy.

Cancer genomics

- Cancer genomics is our main hope to treat cancer efficiently.
- This vast field includes population genomics (study of tumour evolution) and GWAS (identification of driver mutations).
- This knowledge can help in diagnostics (blood test for specific genetic markers), treatment planning (tumours with different mutations have different response to treatments).
- One of the most promising treatments is the viral therapy.

Cancer genomics

- Cancer genomics is our main hope to treat cancer efficiently.
- This vast field includes population genomics (study of tumour evolution) and GWAS (identification of driver mutations).
- This knowledge can help in diagnostics (blood test for specific genetic markers), treatment planning (tumours with different mutations have different response to treatments).
- One of the most promising treatments is the viral therapy.

Metagenomics

- Metagenomics studies genomic material from environmental samples: gut microbes, soils etc.
- Machine learning can be used to identify new species of viruses.
- Gives valuable insight in the organisation of different biomes, can detect presence of different species.
- Other applications: medicine, agriculture, recycling

Metagenomics

- Metagenomics studies genomic material from environmental samples: gut microbes, soils etc.
- Machine learning can be used to identify new species of viruses.
- Gives valuable insight in the organisation of different biomes, can detect presence of different species.
- Other applications: medicine, agriculture, recycling.

Metagenomics

- Metagenomics studies genomic material from environmental samples: gut microbes, soils etc.
- Machine learning can be used to identify new species of viruses.
- Gives valuable insight in the organisation of different biomes, can detect presence of different species.
- Other applications: medicine, agriculture, recycling.

Metagenomics

- Metagenomics studies genomic material from environmental samples: gut microbes, soils etc.
- Machine learning can be used to identify new species of viruses.
- Gives valuable insight in the organisation of different biomes, can detect presence of different species.
- Other applications: medicine, agriculture, recycling.

Summary

- Genomics is a multidisciplinary science which includes biology, mathematics, statistics and computer science.
- It is one of the most dynamic fields of knowledge, and in the 21 century it might have an impact as strong as physics, chemistry, and computational technology had in the previous century.

Perspectives in Russia

- Russia is a multinational country with a unique history and uneven population. Genetic profile of the country would be of high interest from cultural, economical and health aspects.
- Russia is a source of highly valuable archeological discoveries (Denisova man at Altai, mammoths in North-East Siberia). Study of ancient genomes allows to solve many mysteries.