Mathematical models in population genomics and their applications

Vladimir Shchur

NATIONAL RESEARCH
UNIVERSITY

November 20, 2020

Genomics

Genomics is a **multidisciplinary** and **data-driven** science including following fields:

- Biology (genetics)
- Mathematics
- Computer science
- Archaeology, medicine, ecology, ...

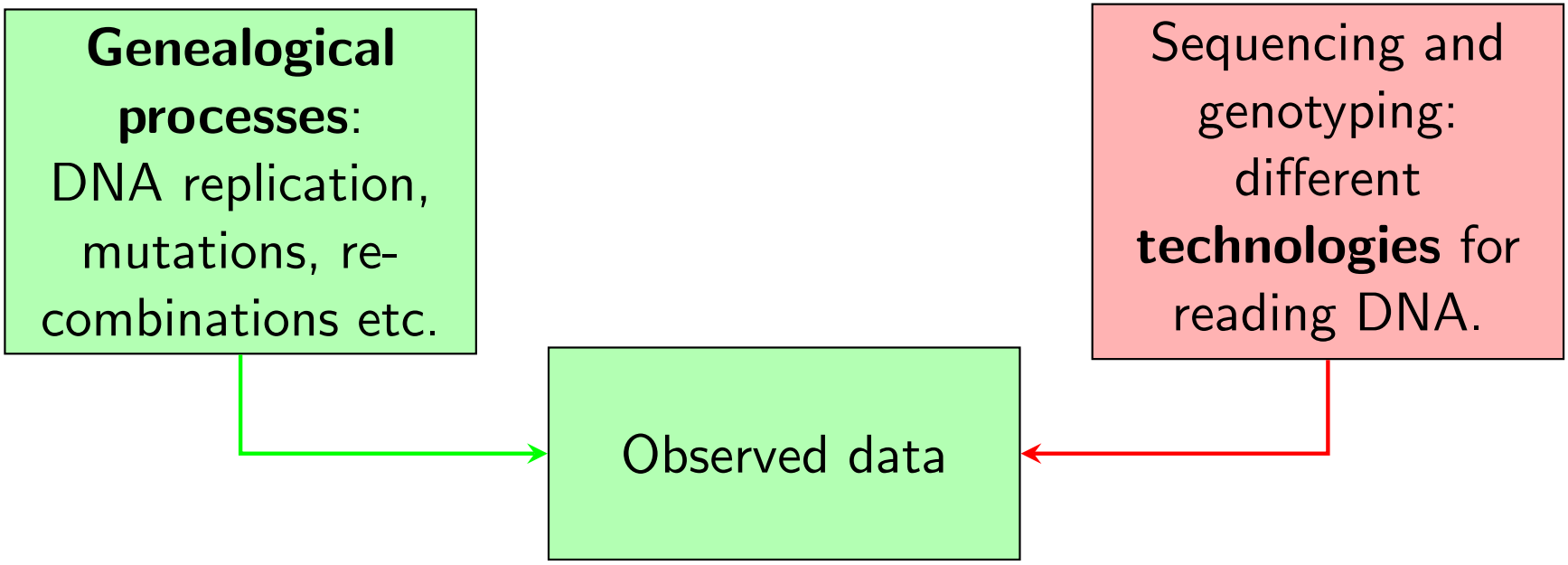
Genomics

Genomics is a **multidisciplinary** and **data-driven** science including following fields:

- Biology (genetics)
- Mathematics
- Computer science
- Archaeology, medicine, ecology, ...

Today we talk about problems of *population genomics*.

What shapes the data?



Wright-Fisher model

Time measured in generations. Constant population size N.

Number of siblings and p-th cousins

p-th cousins are individuals sharing ancestor $p+1$ generations ago.

Theorem

The expected proportion of individuals with at least one p-th cousin in a sample of K individuals is approximately $1 - e^{-2^{2p-1}K/N}$ *in a diploid dioecious Wright-Fisher population of size N.*[1]

[1] Shchur, Nielsen (2018) On the number of siblings and p-th cousins in a large population sample. Math. Biol.

Number of siblings and p-th cousins

p-th cousins are individuals sharing ancestor $p+1$ generations ago.

Theorem

The expected proportion of individuals with at least one p-th cousin in a sample of K individuals is approximately $1 - e^{-2^{2p-1}K/N}$ *in a diploid dioecious Wright-Fisher population of size N.*[1]

Golden State Killer committed more than 150 crimes in 1974-1986. J. DeAngelo was arrested in April 2018.

[1] Shchur, Nielsen (2018) On the number of siblings and p-th cousins in a large population sample. Math. Biol.

Number of siblings and p-th cousins

p-th cousins are individuals sharing ancestor $p+1$ generations ago.

Theorem

The expected proportion of individuals with at least one p-th cousin in a sample of K individuals is approximately $1 - e^{-2^{2p-1}K/N}$ *in a diploid dioecious Wright-Fisher population of size N.*[1]

Golden State Killer committed more than 150 crimes in 1974-1986. J. DeAngelo was arrested in April 2018.

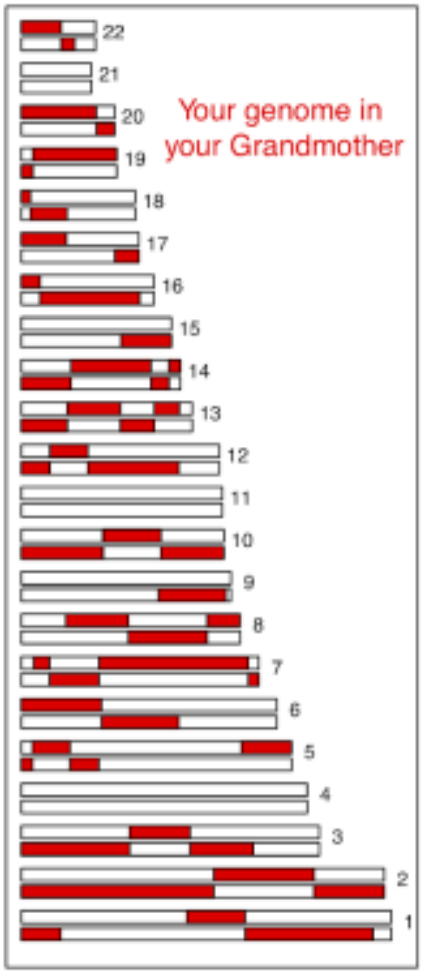
How lucky was the genetic investigation in the Golden State Killer case?

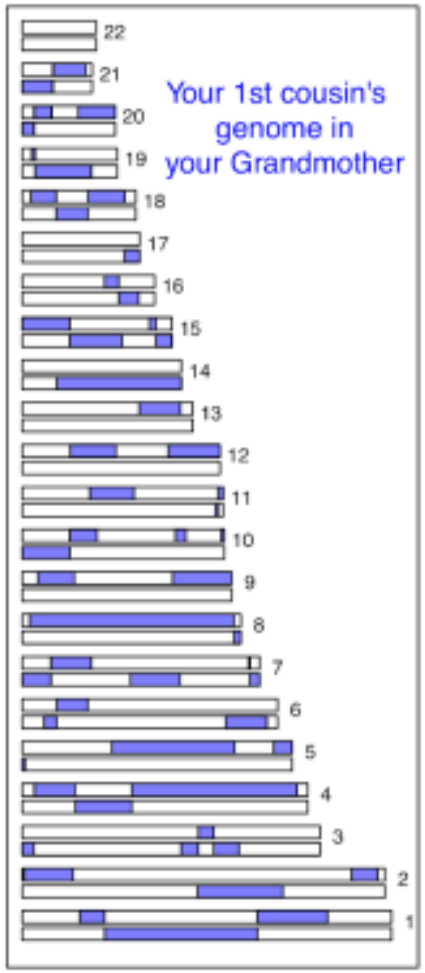
Posted on May 7, 2018

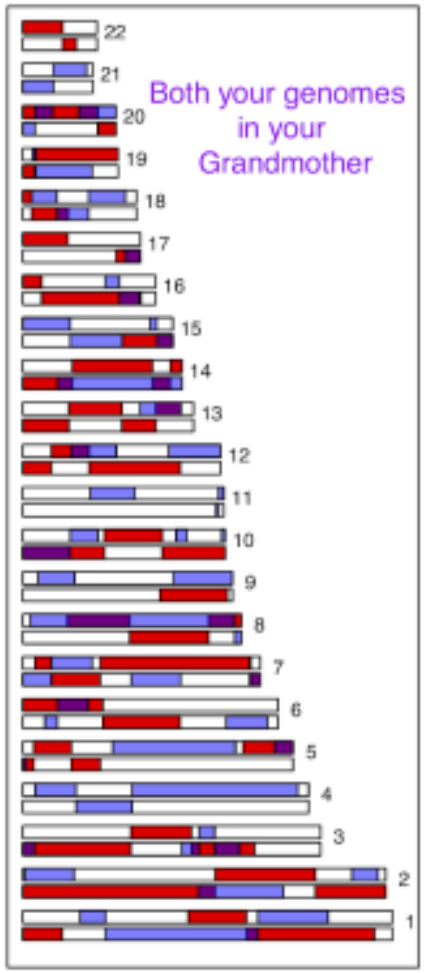
Last week, police arrested Joseph DeAngelo as a suspect in case of the Golden State Killer, an infamous serial murderer and rapist whose case has been open for over forty years. The arrest is huge news in and of itself, but for people interested in the social uses of genetic data, the way in which DeAngelo was identified—using genetic genealogy & genetic data

[1] Shchur, Nielsen (2018) On the number of siblings and p-th cousins in a large population sample. Math. Biol.

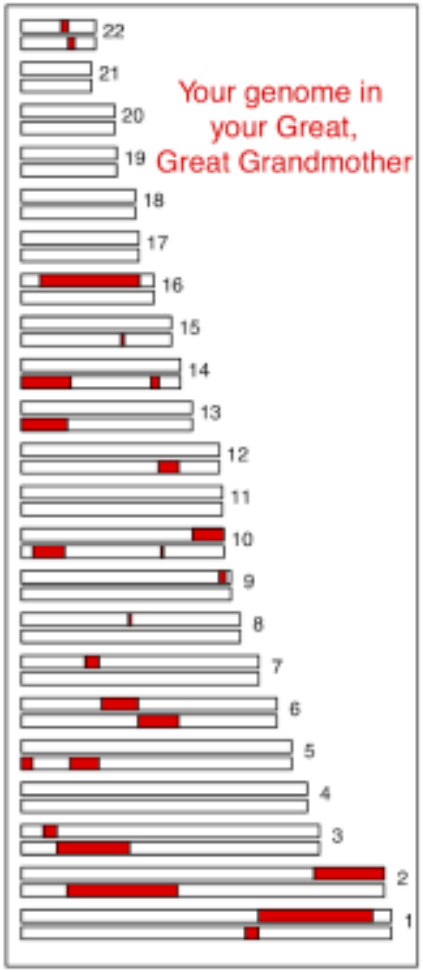
Genealogical and genetic relatives

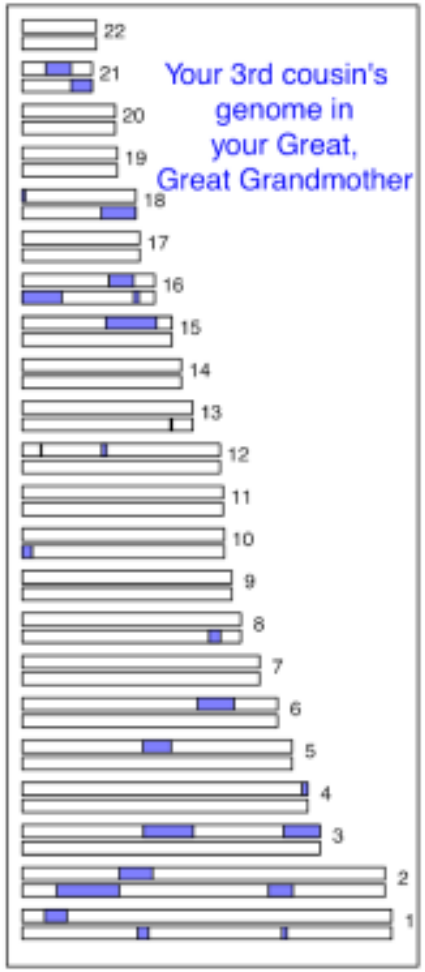


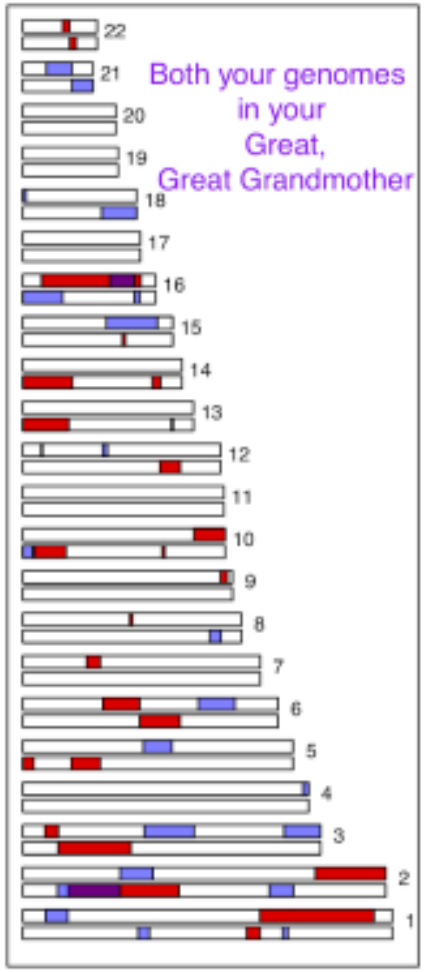




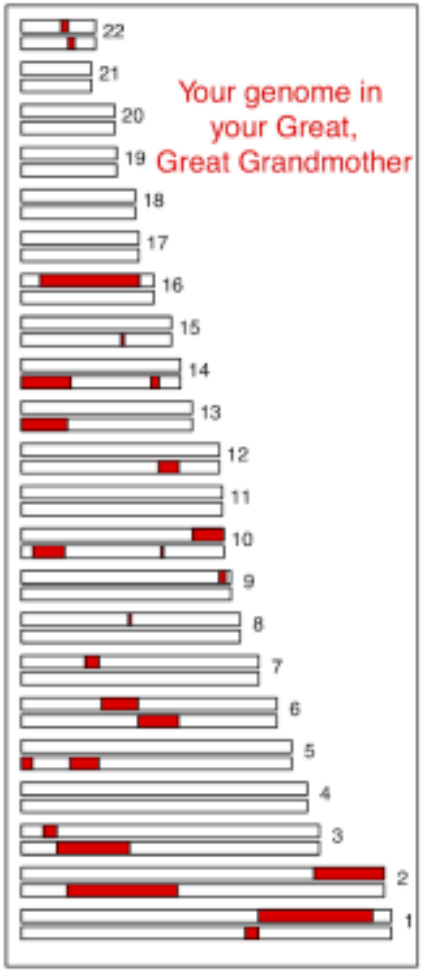
Genealogical and genetic relatives

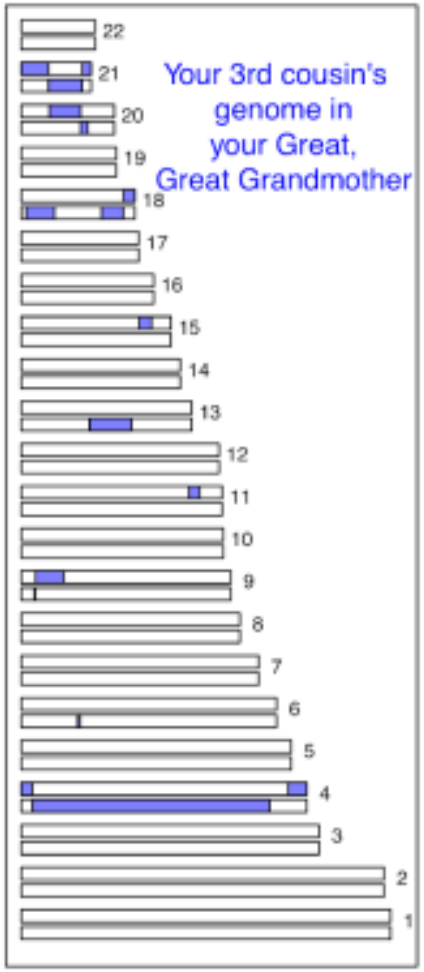


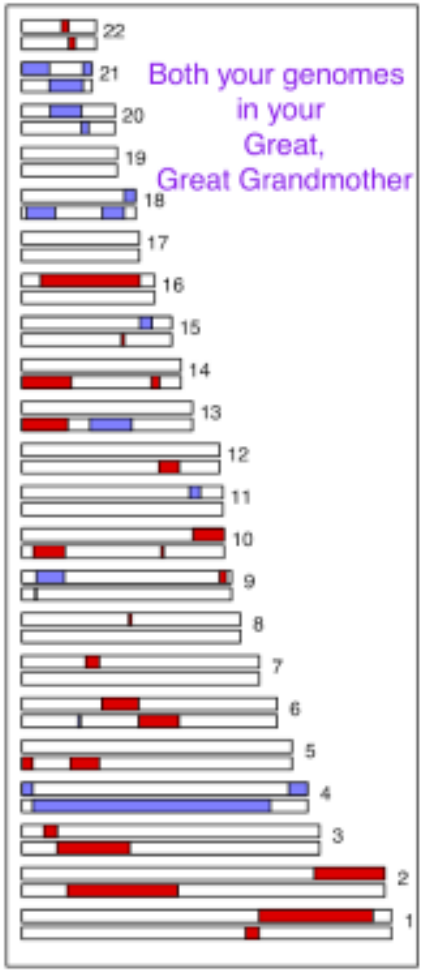




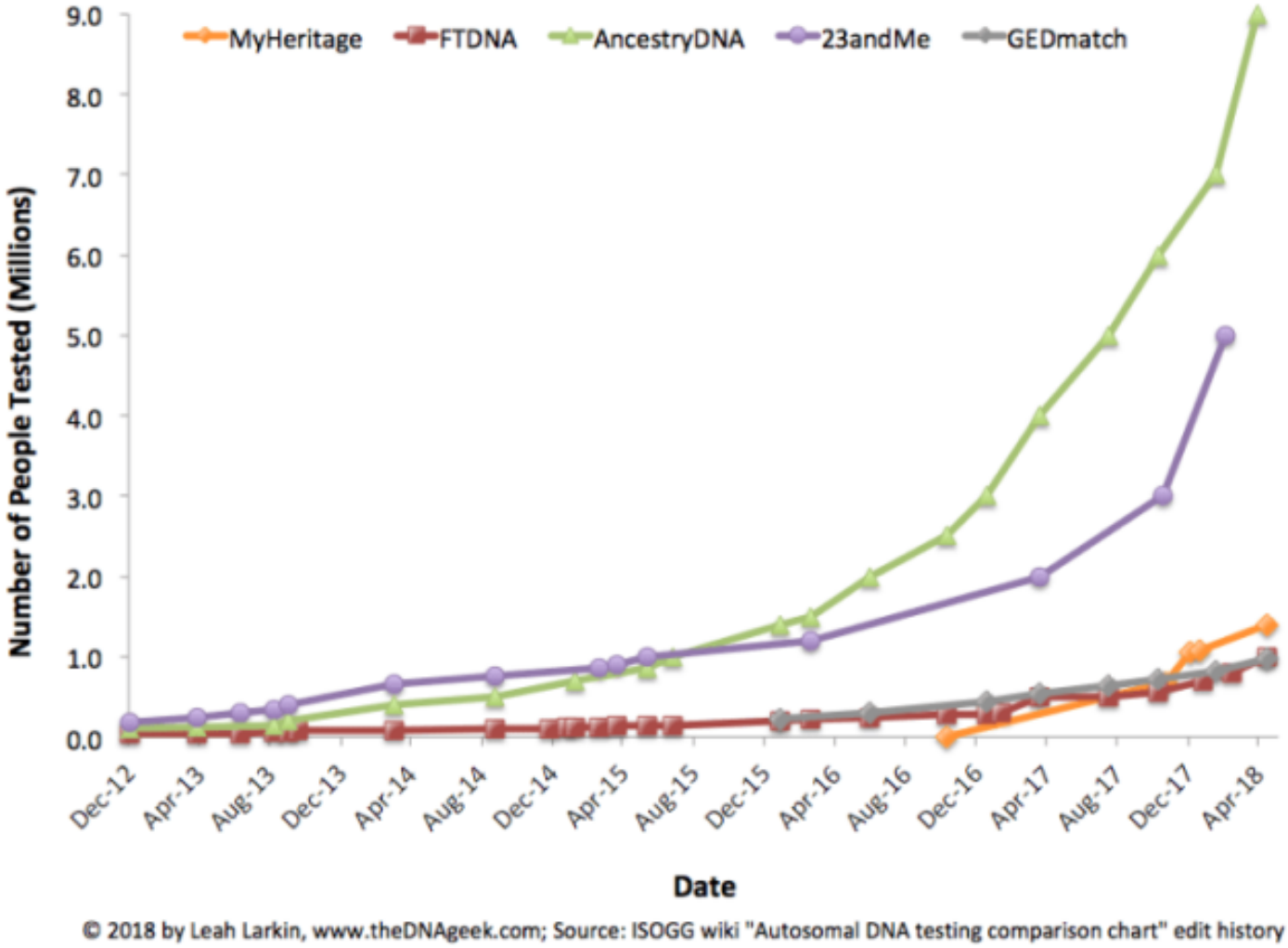
Genealogical and genetic relatives





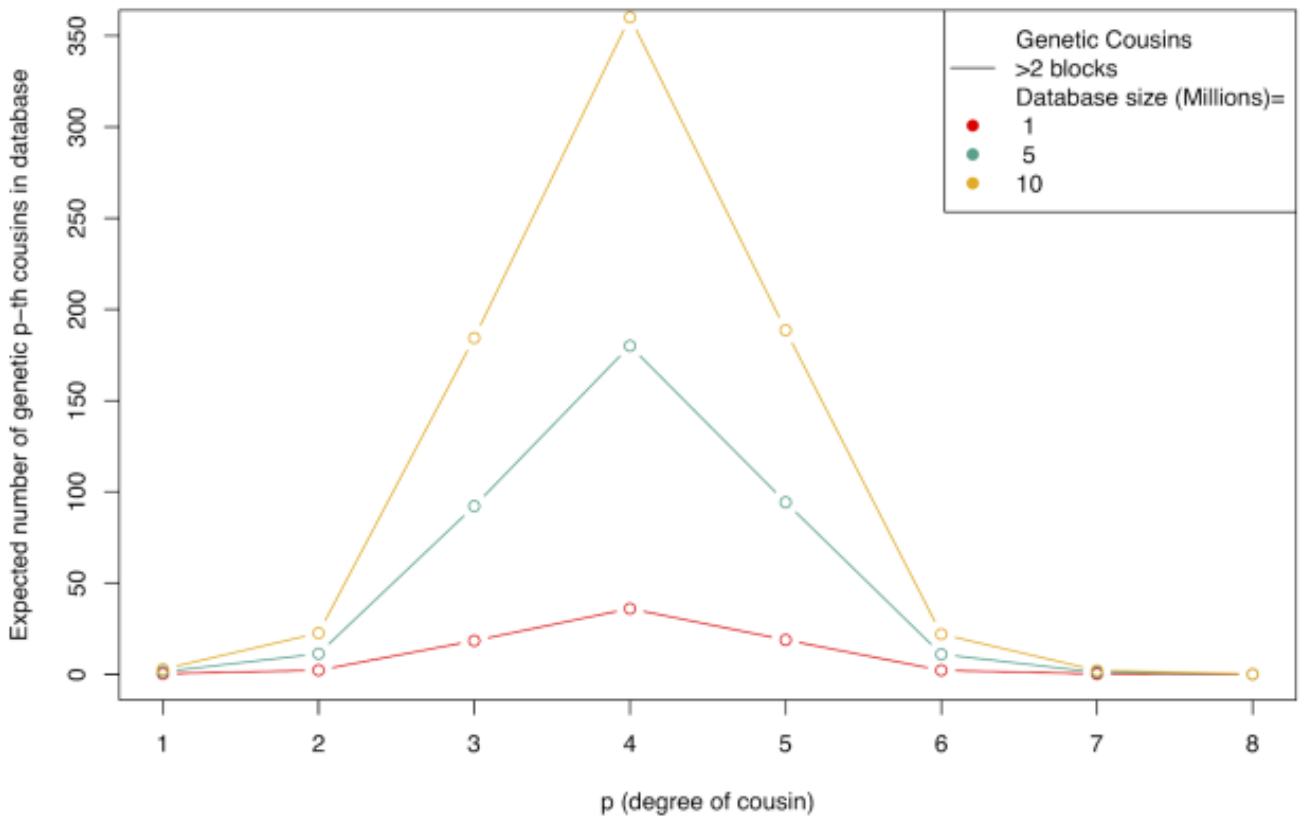


Genetic databases growth



Number of genetic cousins

Police found between 10 to 20 genetic p-th cousins of Golden State Killer. Then they use civil records to narrow the search to a single extended family.



What is a genome?

- A *DNA molecule* is a sequence of four nucleotides: cytosine (C), guanine (G), adenine (A) and thymine (T).
- 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).
- 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).
- 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 $\approx$ 3.2Gb (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 60 new mutations in our genome from our parents.

Some numbers

- Human genome length $\approx$ 3.2Gb (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.

Some numbers

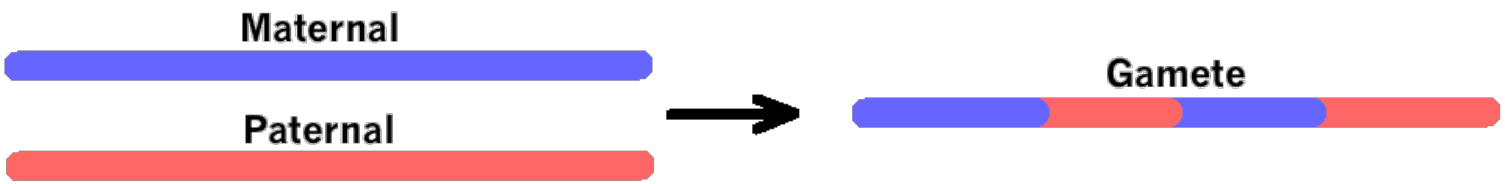
- Human genome length $\approx$ 3.2Gb (Giga-basepairs).
- There are $\approx$ 3 million differences between two typical human haplotypes, e.g. maternal and paternal versions in one person.
- Most of these are shared with other people, caused by mutations in the distant past, 10s or 100s of thousands of years ago.
- Each one of us receives approximately 80 new mutations in our genome from our parents.

Some numbers

- Human genome length $\approx$ 3.2Gb (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.

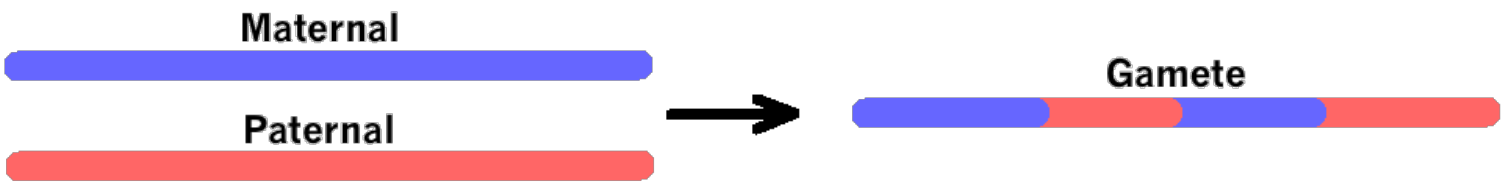
DNA replication

- All the life reproduction is based on cell division. Genetic material is duplicated during this process.
- Errors can occur during duplication: single nucleotide polymorphisms (SNPs), insertions, deletions etc.
- Human gametes contain only one set of chromosomes which is a mosaic of parental two sets of chromosomes, which is created by *recombinations*.



DNA replication

- All the life reproduction is based on cell division. Genetic material is duplicated during this process.
- Errors can occur during duplication: single nucleotide polymorphisms (SNPs), insertions, deletions etc.
- Human gametes contain only one set of chromosomes which is a mosaic of parental two sets of chromosomes, which is created by *recombinations*.

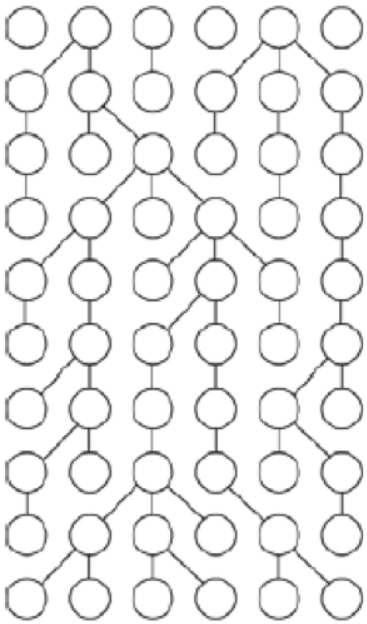


DNA replication

- All the life reproduction is based on cell division. Genetic material is duplicated during this process.
- Errors can occur during duplication: single nucleotide polymorphisms (SNPs), insertions, deletions etc.
- Human gametes contain only one set of chromosomes which is a mosaic of parental two sets of chromosomes, which is created by *recombinations*.

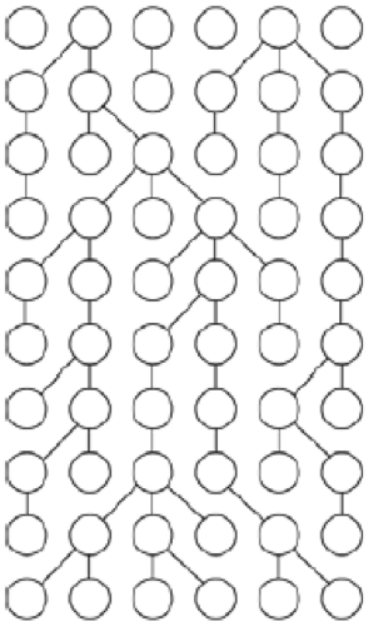
Coalescent model

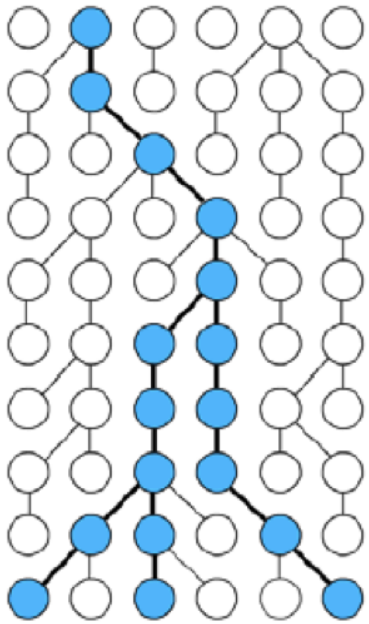
- Wright-Fisher model.



Coalescent model

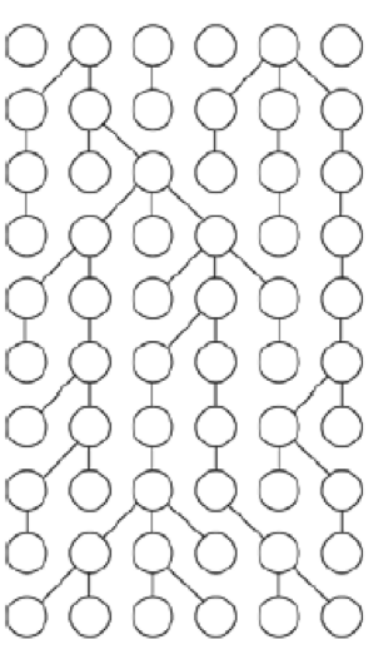
- Wright-Fisher model.
- Sample genealogy.

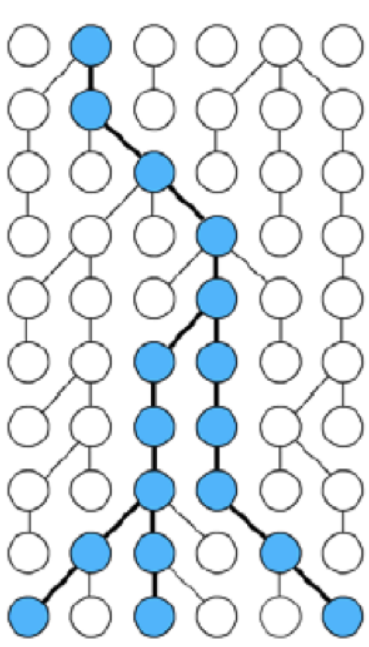


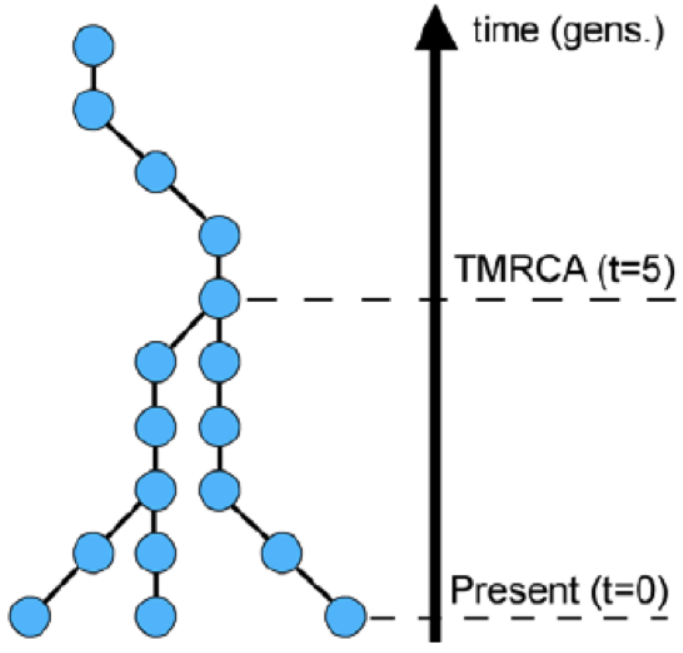


Coalescent model

- Wright-Fisher model.
- Sample genealogy.
- The internal nodes of the tree corresponds to the *most recent common ancestors* of two *lineages*.







Coalescent model

limiting distribution of WF model

Backward continuous time. Sample size $K << N$.

Coalescent with recombination

Continuous Markov process.

Sequentially Markov Coalescent (SMC)

Markov process along the genome.

Inference under SMC

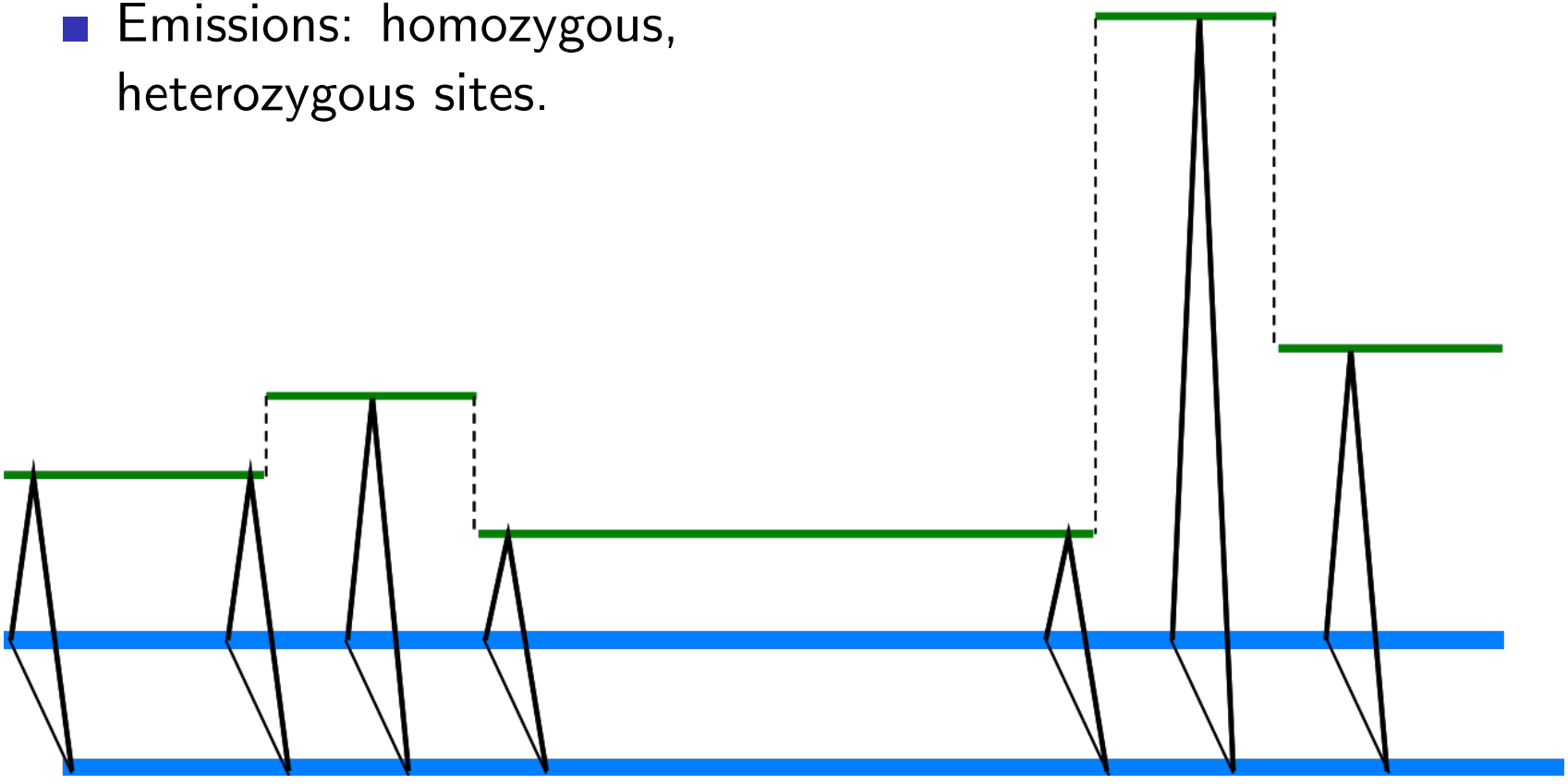
- Coalescent Hidden Markov Model.
  - PSMC method [1]
- MCMC algorithms.
  - BEAST2 package [2]

[1] Li and Durbin (2011) Inference of human population history from individual whole-genome sequences. Nature

[2] Bouckaert, R. et al. (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol.

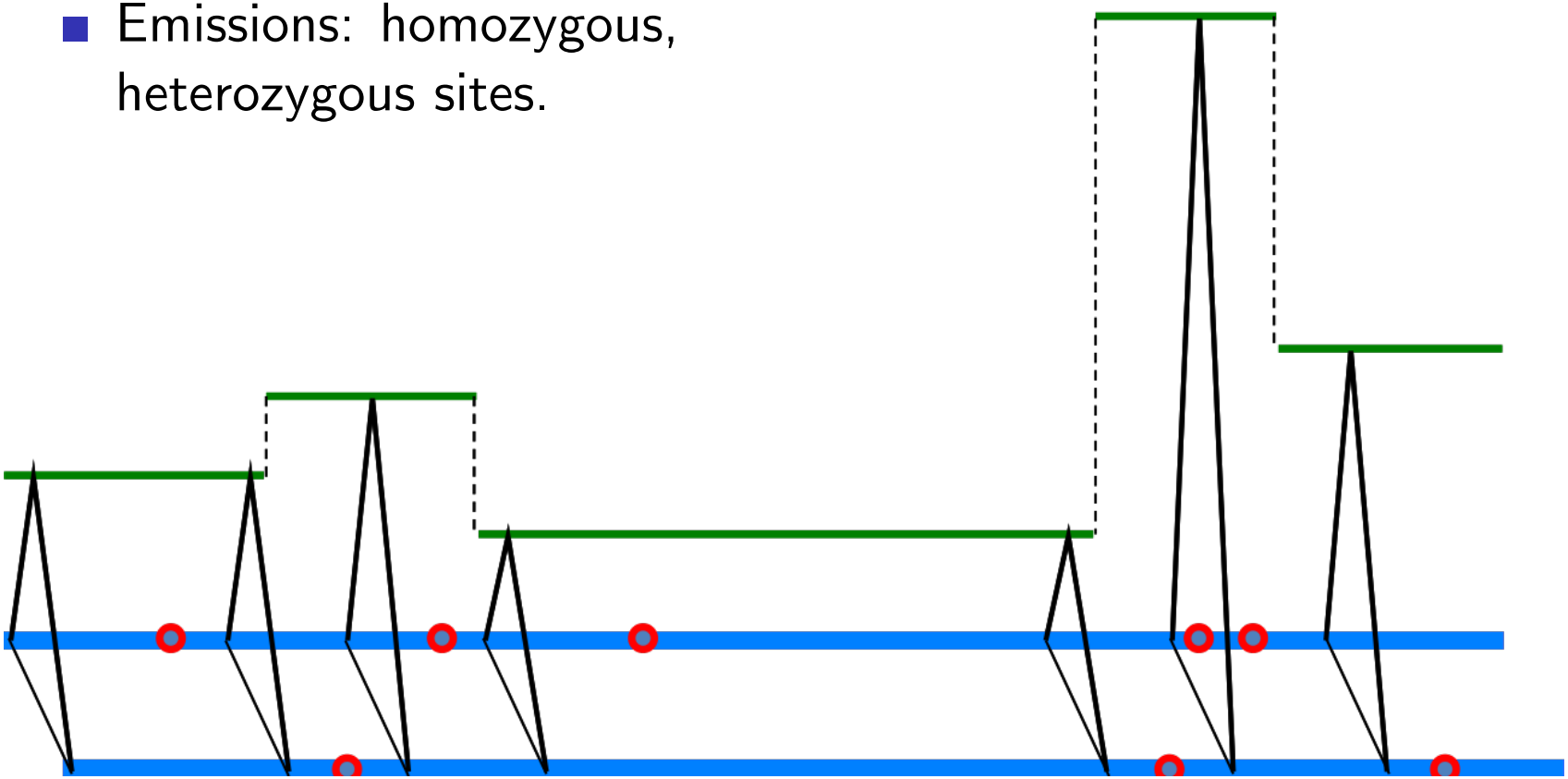
PSMC

- Hidden states: times to the MRCA.
- Emissions: homozygous, heterozygous sites.

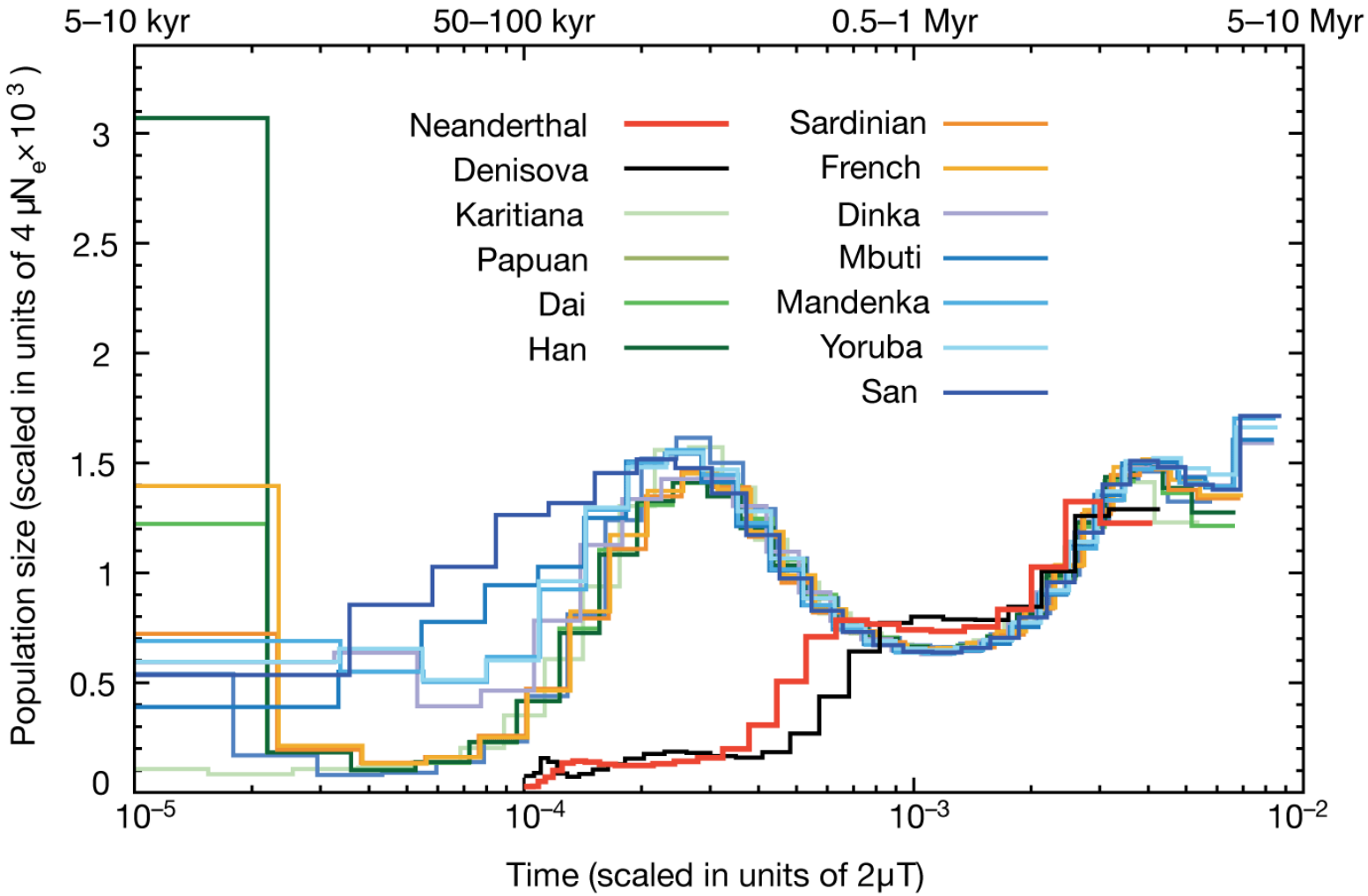


PSMC

- Hidden states: times to the MRCA.
- Emissions: homozygous, heterozygous sites.

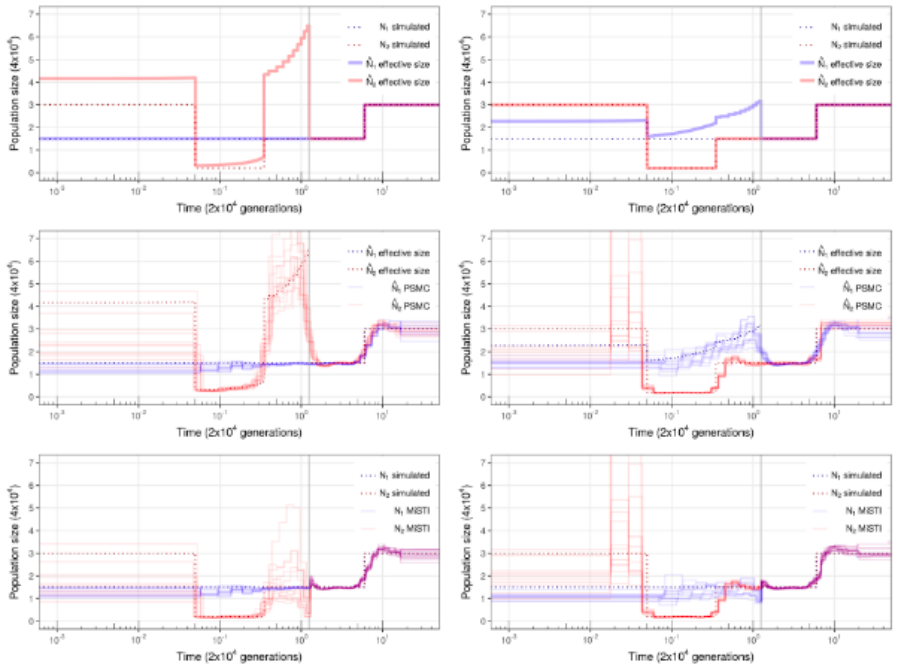


Human population size



Inference under SMC: challenges

- Population structure (migration, non-random mating etc) might affect the inference.
- Most methods rely on variant calls.
- We develop three PSMC-inspired methods to infer migration and population split times.



Covid-19 pandemic

- 211 SARS-CoV-2 sequences (March 11-April 23).
- Including 52 sequences from Vreden Institute of Traumatology.
- Some samples have known travel data.

Covid-19 pandemic

- 211 SARS-CoV-2 sequences (March 11-April 23).
- Including 52 sequences from Vreden Institute of Traumatology.
- Some samples have known travel data.

With: SkolTech, Smorodintsev Research Institute of Influenza, Institute for Information Transmission Problems, Vreden Russian Research Institute of Traumatology and Orthopaedics

Covid-19 pandemic

- 211 SARS-CoV-2 sequences (March 11-April 23).
- Including 52 sequences from Vreden Institute of Traumatology.
- Some samples have known travel data.

With: SkolTech, Smorodintsev Research Institute of Influenza, Institute for Information Transmission Problems, Vreden Russian Research Institute of Traumatology and Orthopaedics

Genomic epidemiology of the early stages of SARS-CoV-2 outbreak in Russia

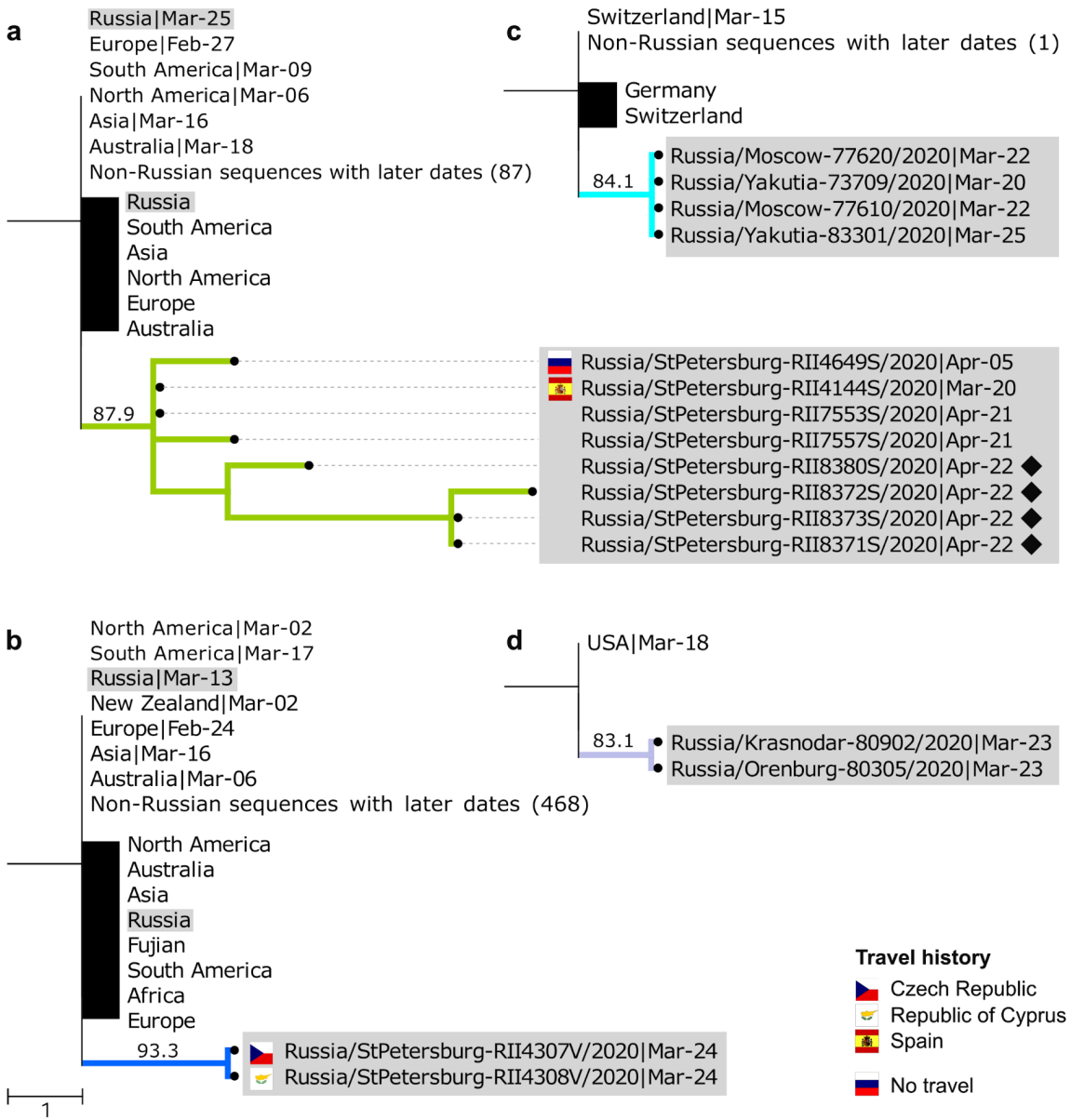
Andrey B Komissarov, Ksenia R Safina, Sofya K Garushyants, Artem V Fadeev, Mariia V Sergeeva, Anna A Ivanova, Daria M Danilenko, Dmitry Lioznov, Olga V Shneider, Nikita Shvyrev, Vadim Spirin, Dmitry Glyzin, Vladimir Shchur, Georgii A Bazykin

doi: https://doi.org/10.1101/2020.07.14.20150979

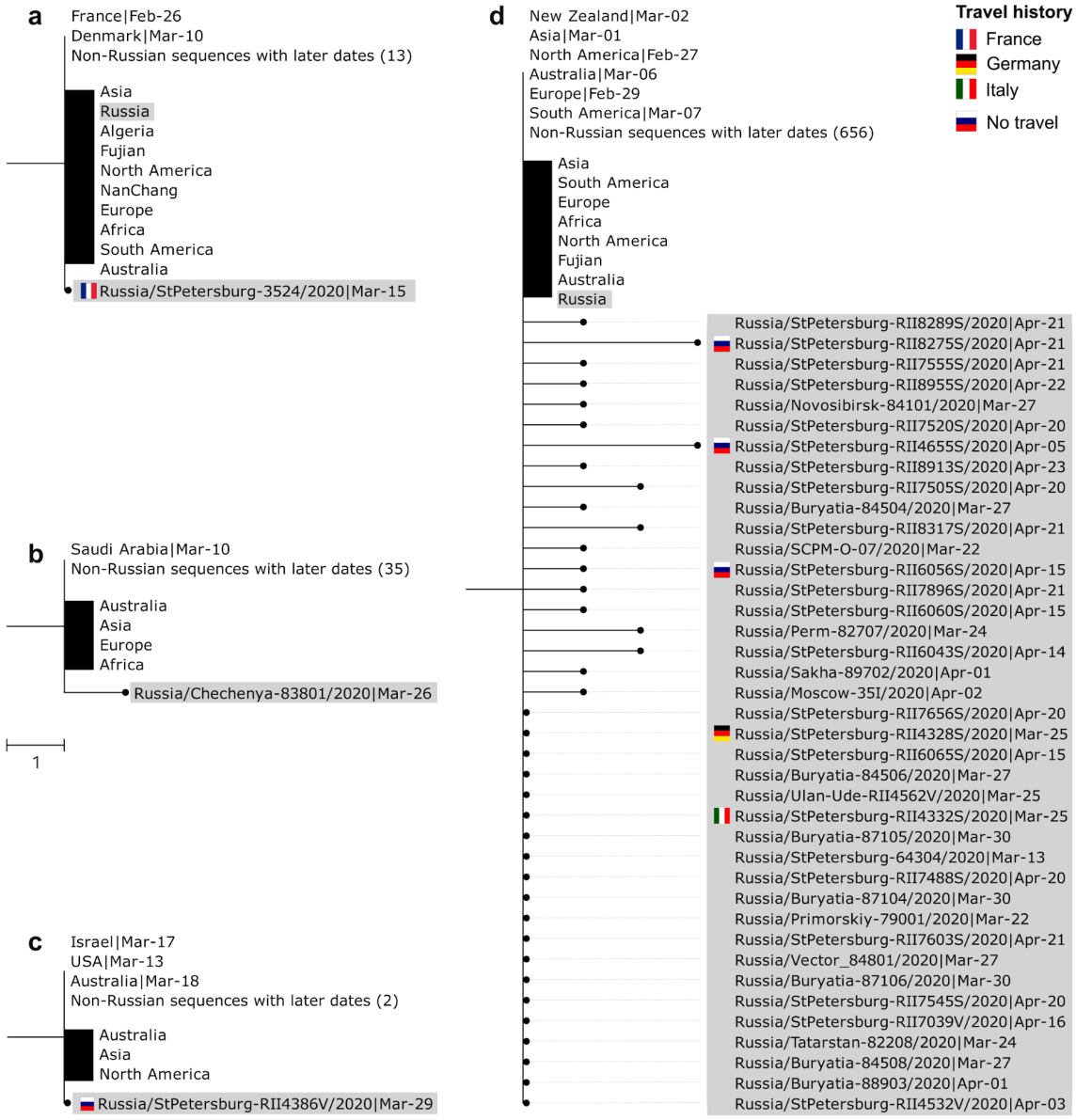
Epidemiology in Russia

- We combine ML-tree (Russian sequences and 19623 world-wide sequences from GISAID) with travel data.
- We did not find any direct imports from China.
- Expected number of introductions for our dataset is 67 separate events.

Epidemiology in Russia

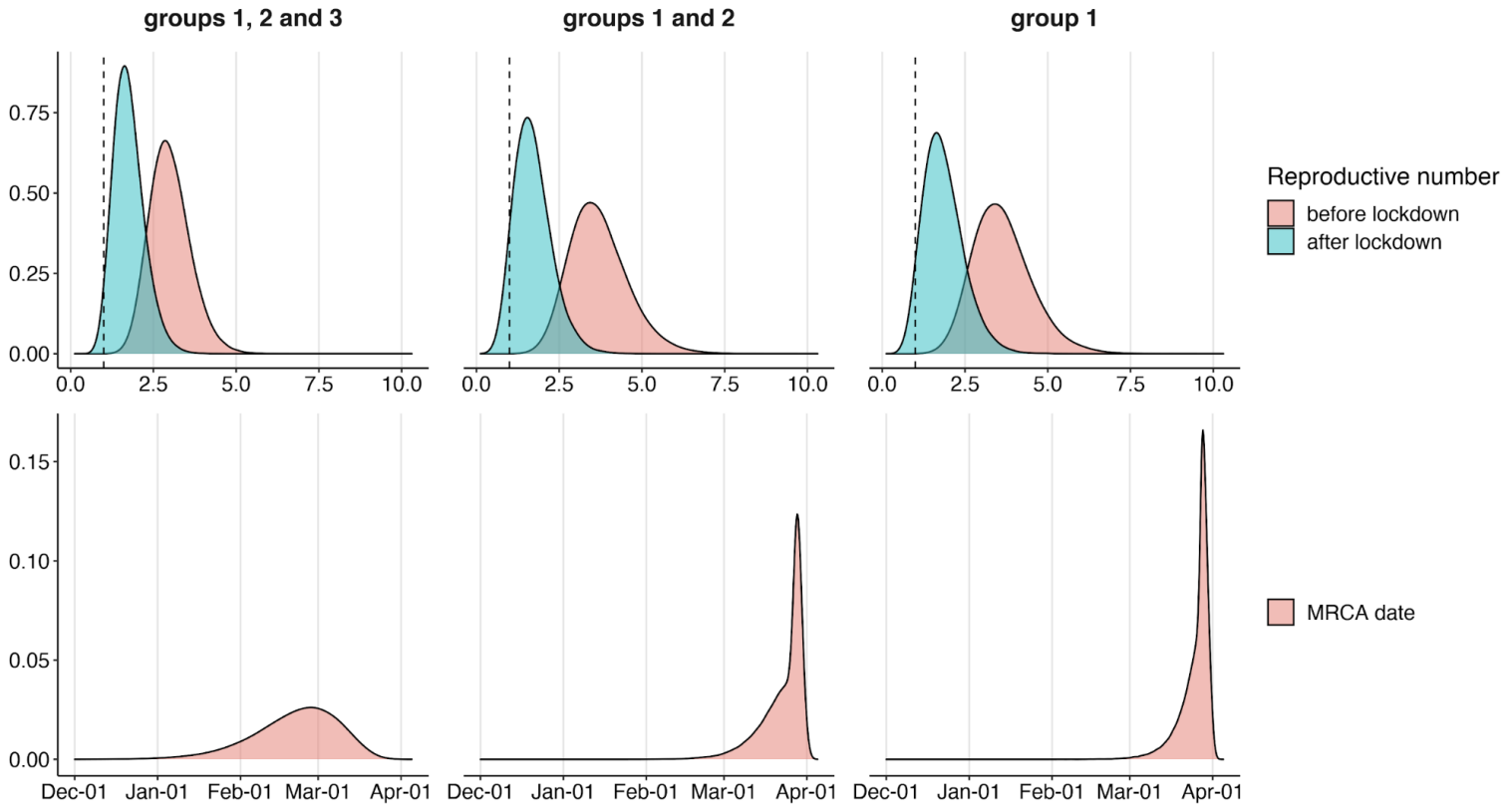


Epidemiology in Russia

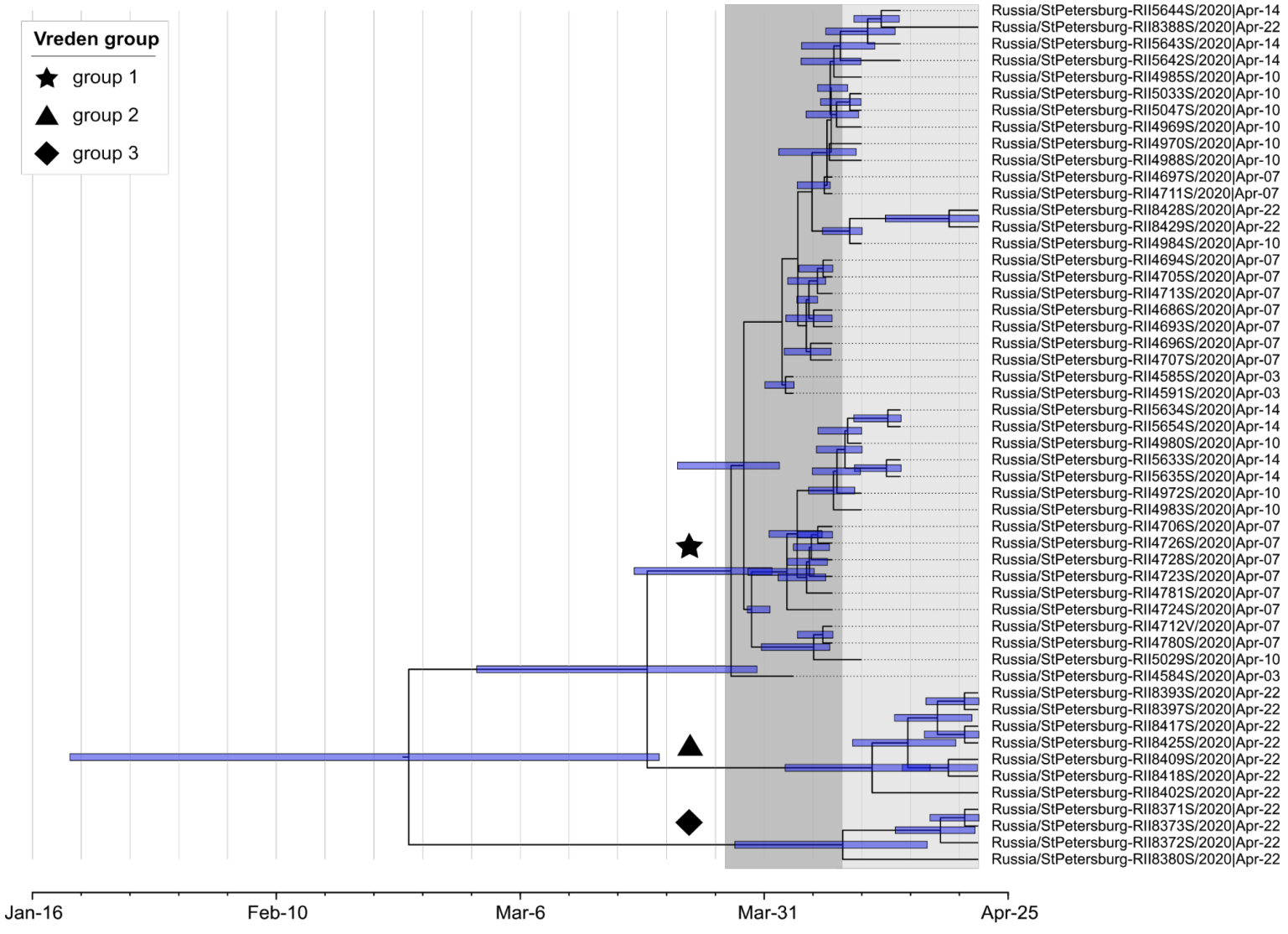


Vreden hospital outbreak

- At least three separates introductions.
- Slower spread after quarantine.



Vreden hospital outbreak



International laboratory of statistical and computational genomics

- Job opportunity for PostDoc/Researcher position.
- PhD or candidate degree in mathematics, computer science, physics or similar.
- Please contact me *vshchur@hse.ru*.