

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ のQで

Mathematical models in population genomics and their applications

Vladimir Shchur



NATIONAL RESEARCH UNIVERSITY

November 20, 2020



▲□▶ ▲□▶ ▲□▶ ▲□▶ ■ ●の00

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

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



▲□▶ ▲□▶ ▲□▶ ▲□▶ ■ ●の00

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?



▲□▶ ▲□▶ ▲□▶ ▲□▶ ■ ●の00





▲□▶ ▲□▶ ▲ □▶ ▲ □▶ □ のへぐ

Time measured in generations. Constant population size N.



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

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



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.

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



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

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

Genealogical and genetic relatives







1-2 1 1-1 1 1-1

Genealogical and genetic relatives







1-8 1 1-1 [] P-1

Genealogical and genetic relatives







1-8 1 1-1 [] P-1

Genetic databases growth



© 2018 by Leah Larkin, www.theDNAgeek.com; Source: ISOGG wiki "Autosomal DNA testing comparison chart" edit history

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへぐ

1-8 1 1-1 [] P-1



(日) (四) (日) (日) (日)

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.



p (degree of cousin)



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*.



▲□▶ ▲□▶ ▲□▶ ▲□▶ ■ ●の00

- 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*.



・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・

- 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*.



• Human genome length \approx 3.2Gb (Giga-basepairs).

- There are ≈ 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.



- Human genome length \approx 3.2Gb (Giga-basepairs).
- There are ≈ 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.



▲□▶ ▲□▶ ▲□▶ ▲□▶ ■ ●の00

- Human genome length \approx 3.2Gb (Giga-basepairs).
- There are ≈ 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.



・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・
・

- Human genome length \approx 3.2Gb (Giga-basepairs).
- There are ≈ 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.



- 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*.





- 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*.





- 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*.





▲□▶ ▲□▶ ▲ □▶ ▲ □▶ □ のへぐ

• 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.





limiting distribution of WF model



▲□▶ ▲□▶ ▲ □▶ ▲ □▶ □ のへぐ

Backward continuous time. Sample size $K \ll N$.

Coalescent with recombination



Continuous Markov process.





Markov process along the genome.





Coalescent Hidden Markov Model.

- PSMC method ¹
- MCMC algorithms.
 - BEAST2 package ²

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





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







▲□▶ ▲□▶ ▲三▶ ▲三▶ 三三 のへで

- 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.





▲□▶ ▲□▶ ▲□▶ ▲□▶ ■ ●の00

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



- 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



- 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



- 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



э



Epidemiology in Russia



a 	France (Phi-26, Domnark (Ma-10) Non-Russian sequences with later dates (13) Acia Russian Algeria Algeria Fujian America Algeria Algeria Fujian America Algeria	d	New Zealand Mar-02 Asia Mar-01 North AmericalTeb-27 Australia Mar-06 Europe Feb-29 South America Mar-05 Non-Russian sequences Asia South America Europe Africa North America Fujian Australia	Travel history II Franco Germany II Italy with later dates (656) No fravel
				Russis/SPetersburg-RUIT2095/2020/Apr-21 Russis/SPetersburg-RUIT2955/2020/Apr-21 Russis/SPetersburg-RUIT2955/2020/Apr-22 Russis/SPetersburg-RUIT5955/2020/Apr-22 Russis/SPetersburg-RUIT5955/2020/Apr-20 Russis/SPetersburg-RUIT5955/2020/Apr-20 Russis/SPetersburg-RUIT5955/2020/Apr-20 Russis/SPetersburg-RUIT5955/2020/Apr-20 Russis/SPetersburg-RUIT5955/2020/Apr-20 Russis/SPetersburg-RUIT5955/2020/Apr-20 Russis/SPetersburg-RUIT5955/2020/Apr-20
b	Saud Arabia Mar-10 Non-Russian sequences with later dates (35) Australia Asia Europe Africa • Russia/Checkenya-83801/2020 Mar-26			Russia/SCPH-0-07/2020 [Mm-22 Russia/SCPH-0-07/2020 [Am-23 Russia/SCPHerraburg-R100565/2020] (Am-15 Russia/SCPHerraburg-R100565/2020] (Am-16 Russia/SCPHerraburg-R100545/2020] (Am-26 Russia/SCPHerraburg-R100545/2020] (Am-26 Russia/SCPHerraburg-R103265/2020] (Am-27 Russia/SCPHerraburg-R103265/2020] (Am-27 Russia/SCPHerraburg-R103265/2020] (Am-27 Russia/SCPHerraburg-R103265/2020] (Am-27 Russia/SCPHerraburg-R103265/2020] (Am-27 Russia/SCPHerraburg-R103265/2020] (Am-27 Russia/SCPHerraburg-R103265/2020] (Am-27 Russia/SCPHerraburg-R103265/2020] (Am-27 Russia/SCPHerraburg-R103265/2020] (Am-27 Russia/SCPHerraburg-R103265/2020] (Am-26 Russia/SCPHerraburg-R103265/2020] (Am-26 Russia/SCPHerraburg-R103265/2020] (Am-26 Russia/SCPHerraburg-R10326) (Am-26) (Am-26
с 	Israel Mar-17 USA/Mar-13 Actralia/Mar-18 Actralia/Mar-18 Actralia/Mar-18 Action Action Action Action North America Factoria/Stefeterburg-RI/4386V/2020/Mar-29			Russia/SPeternburg-4524/2020/Har-13 Russia/SPeternburg-41714855/2020/Har-20 Russia/SPeternburg-41714855/2020/Har-20 Russia/Peternburg-4171055/2020/Har-22 Russia/Veternburg-417102/SPETERDS/ Russia/Veternburg-417102/SPETERDS/ Russia/Veternburg-417102/SPETERDS/ Russia/Referenburg-417102/SPETERDS/ Russia/Referenburg-4171029/Veternburg- Russia/Referenburg-4171029/Veternburg- Russia/Referenburg-4171029/Veternburg- Russia/Referenburg-4171029/Veternburg- Russia/Referenburg-4171029/Veternburg- Russia/Referenburg-4509/2020/Har-27 Russia/Refurg-4509/2020/Har-21 Russia/Refurg-418-509/2020/Har-21

▲ 臣 ▶ 臣 • • • • • •

Vreden hospital outbreak



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



Vreden hospital outbreak



Russia/StPetersburg-RII5644S/2020IApr-14 Russia/StPetersburg-RII8388S/2020IApr-22 Russia/StPetersburg-RII5643S/2020IApr-14 Russia/StPetersburg-RII5642S/2020|Apr-14 Russia/StPetersburg-RII4985S/2020 Apr-10 Russia/StPetersburg-RII5033S/2020 Apr-10 Russia/StPetersburg-RII5047S/2020IApr-10 Russia/StPetersburg-RII4969S/2020IApr-10 Russia/StPetersburg-RII4970S/2020|Apr-10 Russia/StPetersburg-RII4988S/2020|Apr-10 Russia/StPetersburg-RII4697S/2020 Apr-07 Russia/StPetersburg-RII4711S/2020[Apr-07 Russia/StPetersburg-RII8428S/2020IApr-22 Russia/StPetersburg-RII8429S/2020[Apr-22 Russia/StPetersburg-RII4984S/2020 Apr-10 Russia/StPetersburg-RII4694S/2020 Apr-07 Russia/StPetersburg-RII4705S/2020IApr-07 Russia/StPetersburg-RII4713S/2020IApr-07 Russia/StPetersburg-RII4686S/2020[Apr-07 Russia/StPetersburg-RII4693S/2020 Apr-07 Russia/StPetersburg-RII4696S/2020 Apr-07 Russia/StPetersburg-RII4707S/2020 Apr-07 Russia/StPetersburg-RII4585S/2020IApr-03 Russia/StPetersburg-RII4591S/2020[Apr-03 Russia/StPetersburg-RII5634S/2020|Apr-14 Russia/StPetersburg-RII5654S/2020IApr-14 Russia/StPetersburg-RII4980S/2020IApr-10 Russia/StPetersburg-RII5633S/2020IApr-14 Russia/StPetersburg-RII5635S/2020IApr-14 Russia/StPetersburg-RII4972S/2020|Apr-10 Russia/StPetersburg-RII4983S/2020 Apr-10 Russia/StPetersburg-RII4706S/2020IApr-07 Russia/StPetersburg-RII4726S/2020IApr-07 Russia/StPetersburg-RII4728S/2020IApr-07 Russia/StPetersburg-RII4723S/2020|Apr-07 Russia/StPetersburg-RII4781S/2020 Apr-07 Russia/StPetersburg-RII4724S/2020|Apr-07 Russia/StPetersburg-RII4712V/2020IApr-07 Russia/StPetersburg-RII4780S/2020IApr-07 Russia/StPetersburg-RII5029S/2020[Apr-10 Russia/StPetersburg-RII4584S/2020 Apr-03 Russia/StPetersburg-RII8393S/2020|Apr-22 Russia/StPetersburg-RII8397S/2020IApr-22 Russia/StPetersburg-RII8417S/2020IApr-22 Russia/StPetersburg-RII8425S/2020[Apr-22 Russia/StPetersburg-RII8409S/2020|Apr-22 Russia/StPetersburg-RII8418S/2020 Apr-22 Russia/StPetersburg-RII8402S/2020 Apr-22 Russia/StPetersburg-RII8371S/2020IApr-22 Russia/StPetersburg-RII8373S/2020IApr-22 Russia/StPetersburg-RII8372S/2020|Apr-22 Russia/StPetersburg-RII8380S/2020|Apr-22

1-1 1-1 1-1

▲□▶ ▲□▶ ▲三▶ ▲三▶ 三三 のへで



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