NATIONAL RESEARCH UNIVERSITY

International laboratory of statistical and computational genomics,
AI Institute, Faculty of Computer Science
HSE University

Human population genomics: from present times into deep past

Популяционная геномика человека: от настоящего до глубокого прошлого

Vladimir Shchur

Dynamics in Siberia 2026, Novosibirsk

Геномика - междисциплинарная наука, основанная на экспериментальных данных.

генетика

компьютерные науки

математика

Геномика - междисциплинарная наука, основанная на экспериментальных данных.

генетика **компьютерные науки** **математика**

а также

медицина **археология** **экология**

криминалистика **эпидемиология**

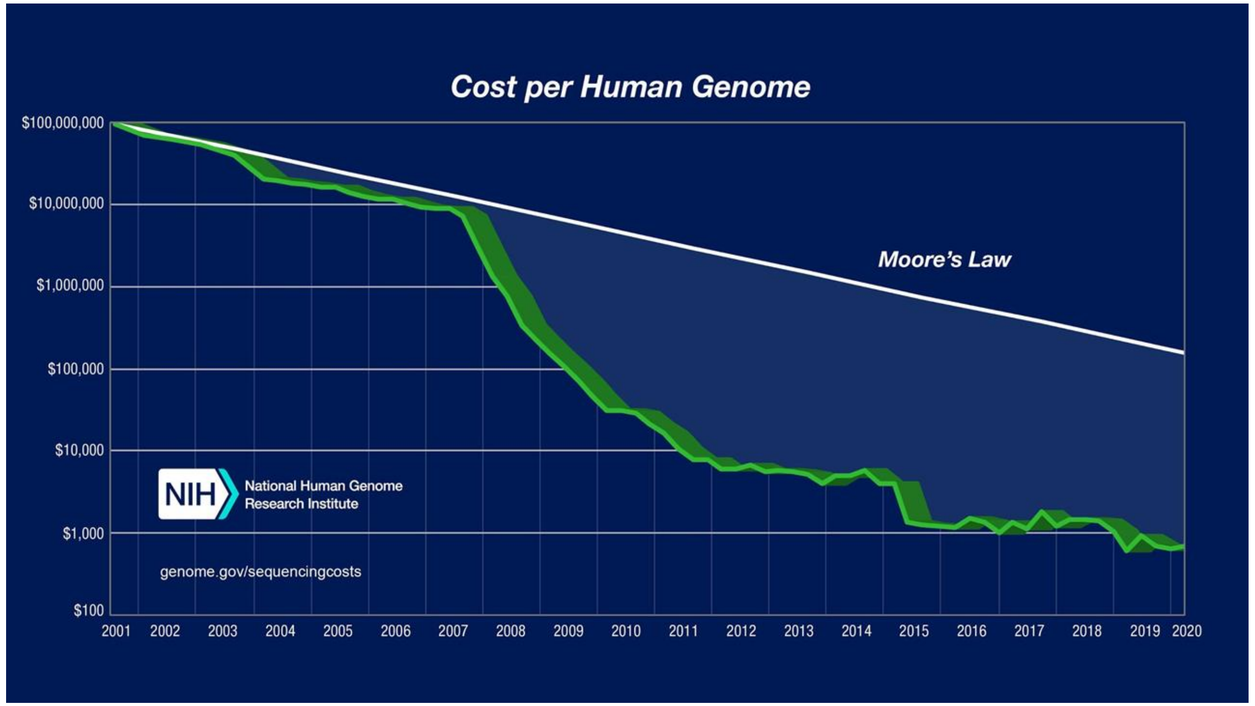
Популяционная геномика

Мотивация: взрывной рост генетических данных, доступных для анализа, дает возможность изучать историю различных популяций и эволюционные механизмы.

Цель: изучение популяционной и эволюционной истории из генетических данных.

Основные задачи:

- Когда разделялись популяции?
- Как менялся размер популяций?
- Как происходила миграция?
- На какие участки генома действовал естественный отбор?
- Какова структура (неоднородность) современных популяций?



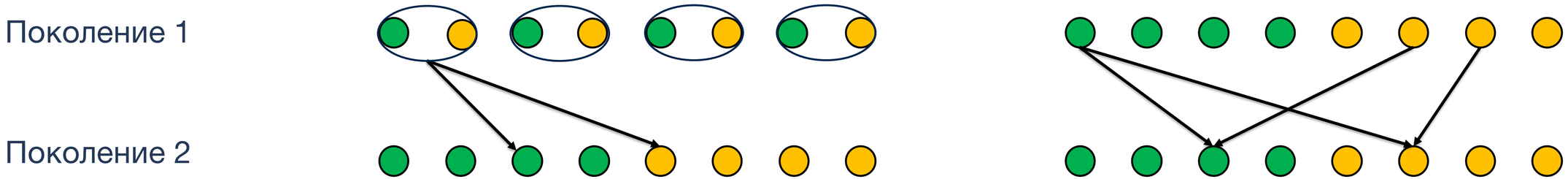
- 2003 - первый геном человека $300M
- 2015 - геном человека стоит $1K

О числе *p*-сестер в большой выборке из популяции

Рассмотрим диплоидную популяцию Райта-Фишера:

- поколения не пересекаются,
- число мужских и женских особей постоянно и равно N.

Rasmus Nielsen
(UC Berkeley)



Моногамная модель

Немоногамная модель

Shchur V., Nielsen R. “On the number of siblings and p-th cousins in a large population sample”, Journal of Mathematical Biology 77(5) (2018), pp. 1279-1298

О числе *p*-сестер в большой выборке из популяции

***p*-сестрами** назовем особей, у которых есть пара общих предков *p* поколений назад.
***p*-полусестрами** назовем особей, у которых есть один общий предок *p* поколений назад.

Например, *2*-сестры - *это двоюродные* сестры, то есть особи, имеющие общих бабушку и дедушку.

Рассмотрим выборку *S* размера *K* из моногамной популяции Райта-Фишера. Пусть U_2 - случайная величина, равная числу особей в выборке *S* без *2*-сестер. Тогда математическое ожидание U_2 равно

$$\mathbb{E}(U_2) = K \frac{\sum_{m=1}^{K} S(K,m) \binom{N}{m} m! N (N-1)(N-2)^{2m-2}}{\sum_{m=1}^{K} S(K,m) \binom{N}{m} m! N^{2m}}$$

Shchur V., Nielsen R. “On the number of siblings and p-th cousins in a large population sample”, Journal of Mathematical Biology 77(5) (2018), pp. 1279–1298

О числе *p*-сестер в большой выборке из популяции

Теорема. Пусть *K* – число особей в случайной выборке *S*. Пусть U_p - случайная величина, равная числу особей в выборке *S* без *p*-сестер, также попавших в выборку *S*, в моногамной модели Райта-Фишера. V_p – аналогичная случайная величина для *p*-полусестер в немоногамной модели. Пусть также $K/N=\alpha+o(1)$ при $N\to\infty$. Тогда

$$\lim_{N\to\infty}\frac{EU_p}{K}=e^{-(2^{2p-2})\alpha}$$

и

$$\lim_{N\to\infty}\frac{EV_p}{K}=e^{-(2^{2p-1})\alpha}\ .$$

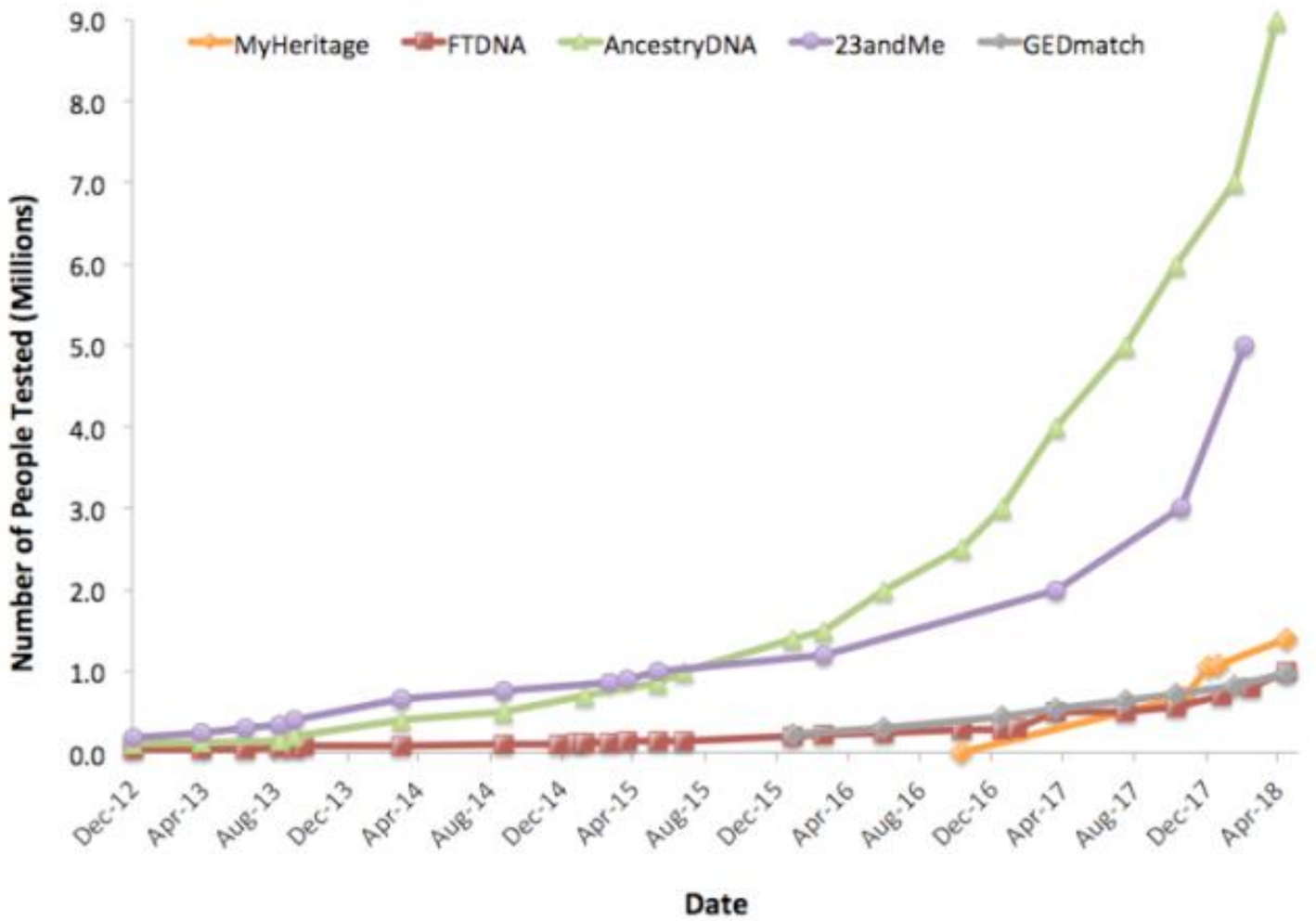
Также в работе получены явные выражения для распределений U_1 и V_1 и математических ожиданий U_p и V_p.

Shchur V., Nielsen R. “On the number of siblings and p-th cousins in a large population sample”, Journal of Mathematical Biology 77(5) (2018), pp. 1279–1298

Убийца из Золотого Штата

- Golden State Killer: более 150 преступлений с 1974 по 1986.
- Джозеф Деанджело арестован в апреле 2018.

Генетические базы данных



Генетические и генеалогические родственники

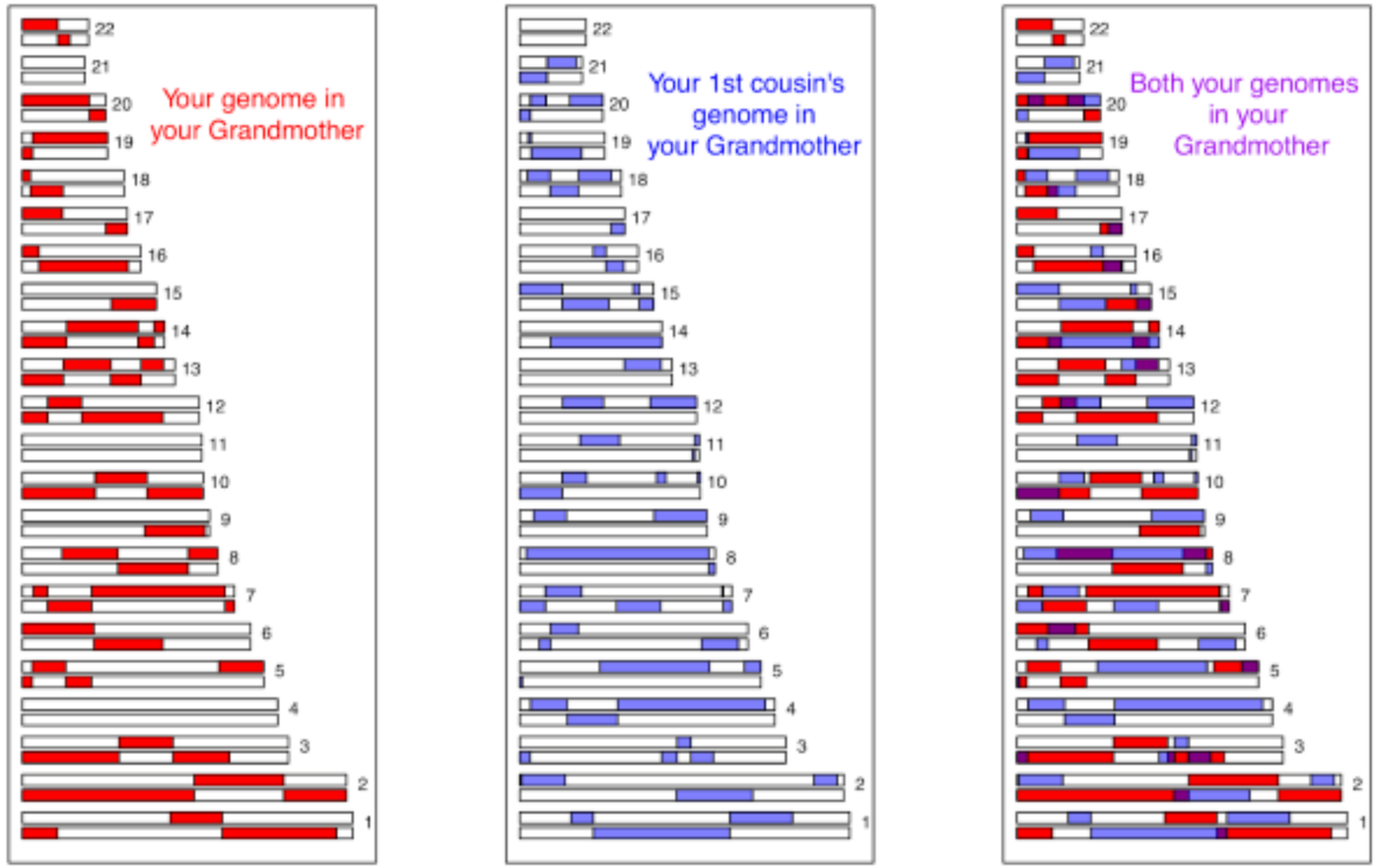


Рисунок: Edge, Coop (2019) How lucky was the genetic investigation in the Golden State Killer case? biorxiv

Генетические и генеалогические родственники

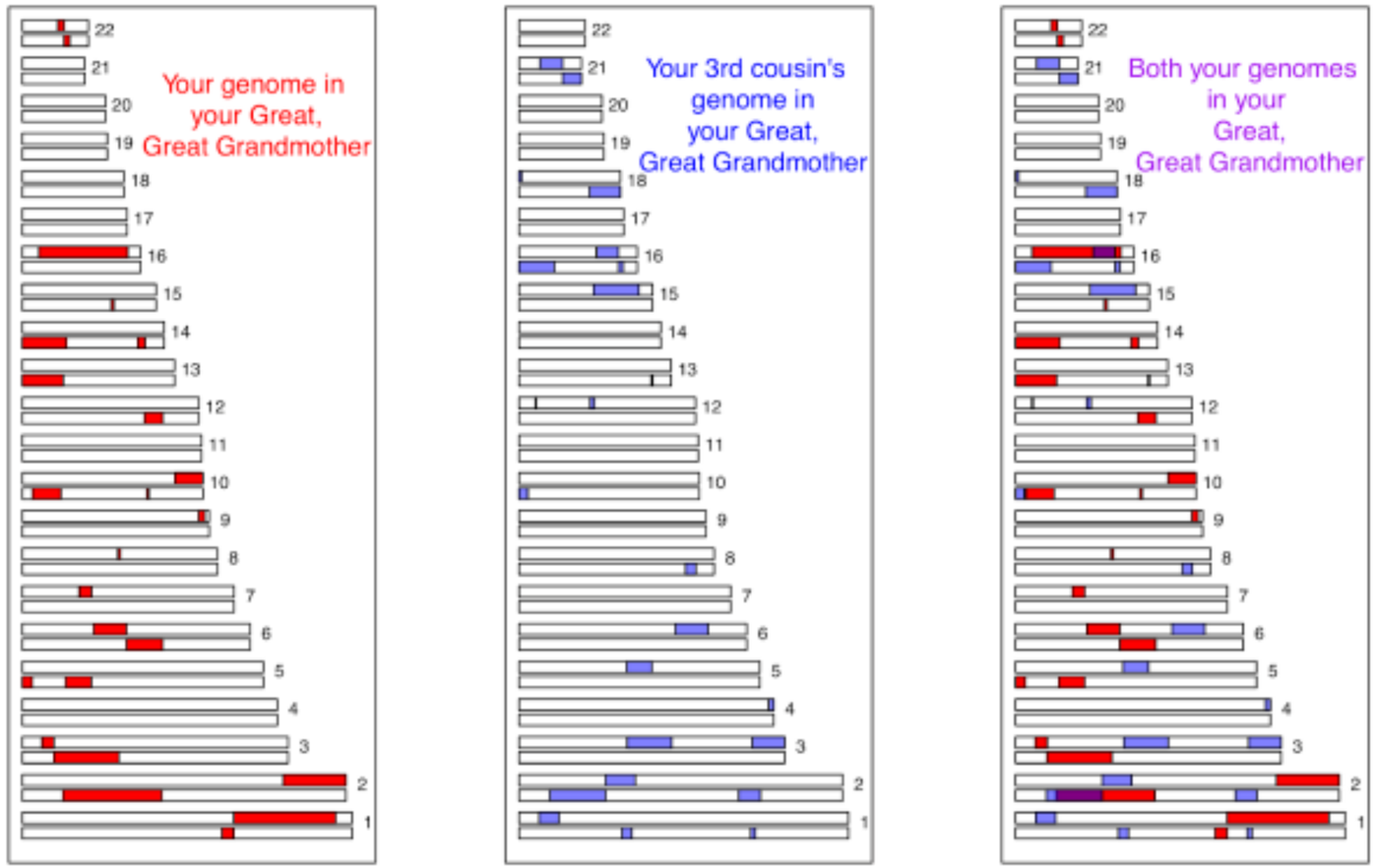


Рисунок: Edge, Coop (2019) How lucky was the genetic investigation in the Golden State Killer case? biorxiv

Генетические и генеалогические родственники

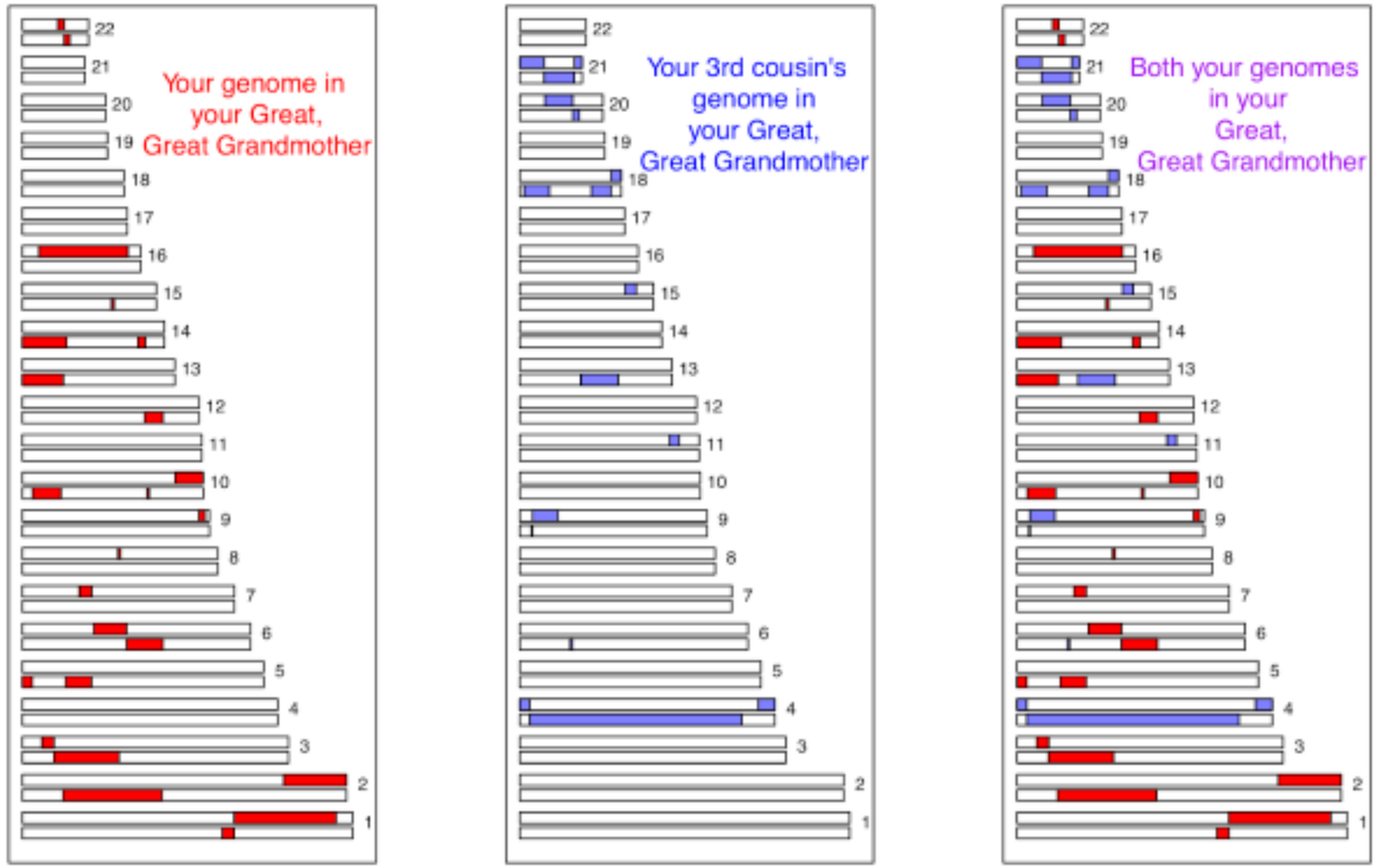
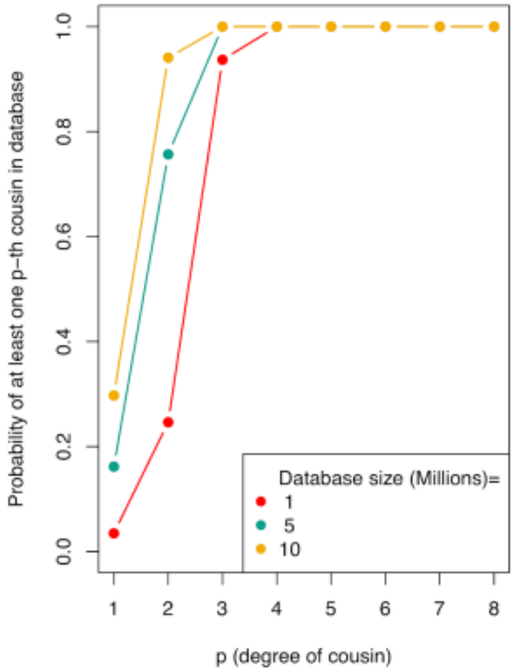
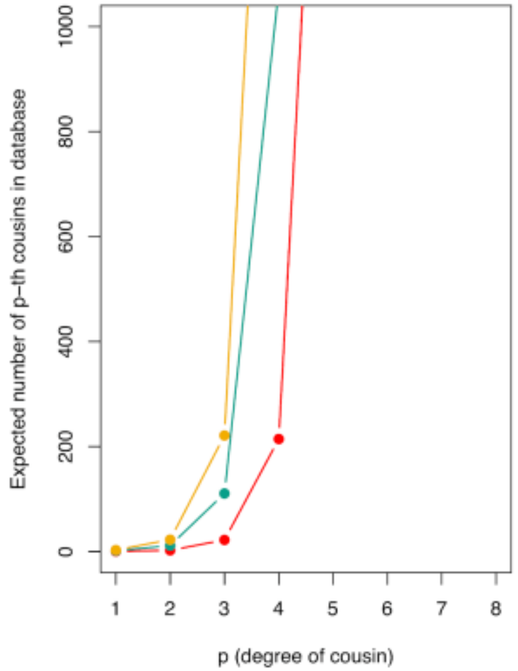


Figure: Edge, Coop (2019) How lucky was the genetic investigation in the Golden State Killer case? biorxiv





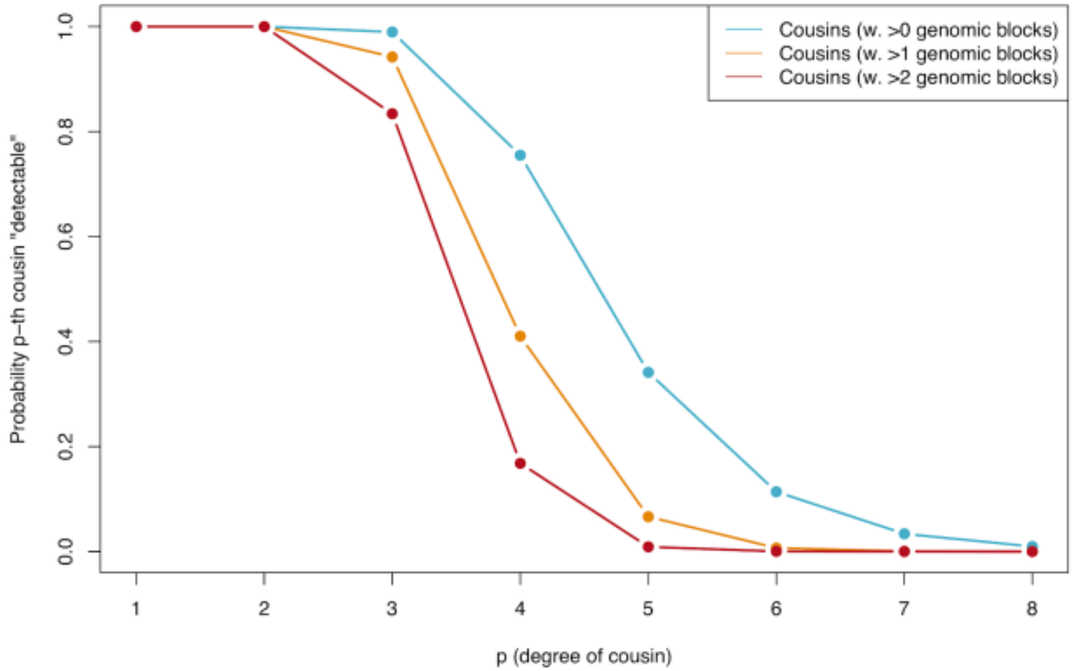
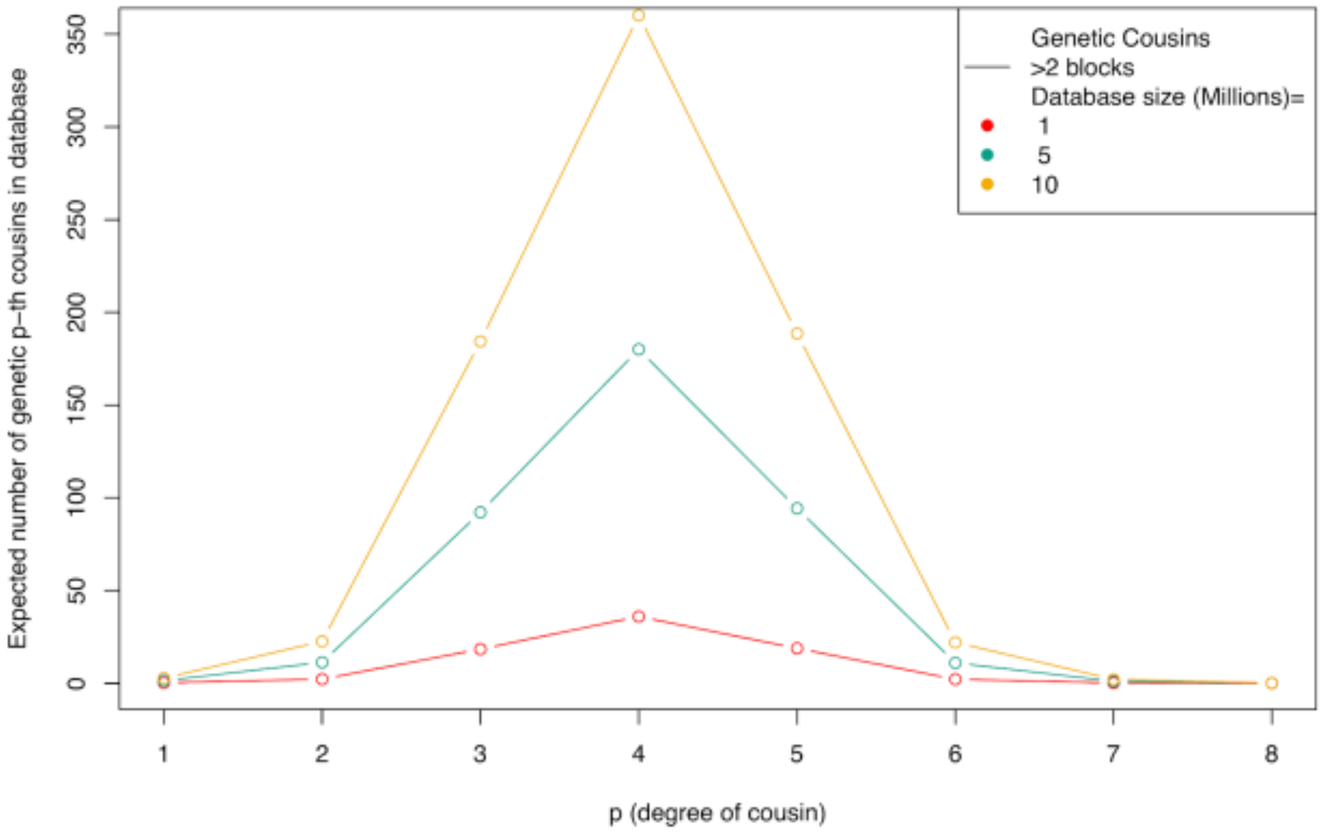


Figure: Edge, Coop (2019) How lucky was the genetic investigation in the Golden State Killer case? biorxiv

Вероятность обнаружить в большой генетической базе вашего генетического родственника близка к 1.

И даже не одного!



How to distinguish individuals from closely related populations?

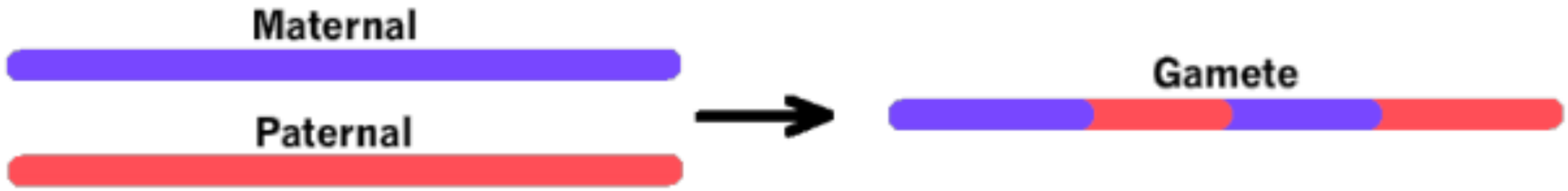
Aleksei Shmelev
(HSE University)

Alexander Rakitko
(Genotek)

with Genotek

Воспроизведение жизни

- Жизнь воспроизводится копированием последовательностей ДНК
- Человеческий геном состоит из двух наборов хромосом (гаплотипов)
- Каждый гаплотип передаётся от одного родителя
- Гаплотип является мозаикой исходного диплоидного генома



Воспроизведение жизни

От поколения к поколению, в мозаике становится всё больше фрагментов.

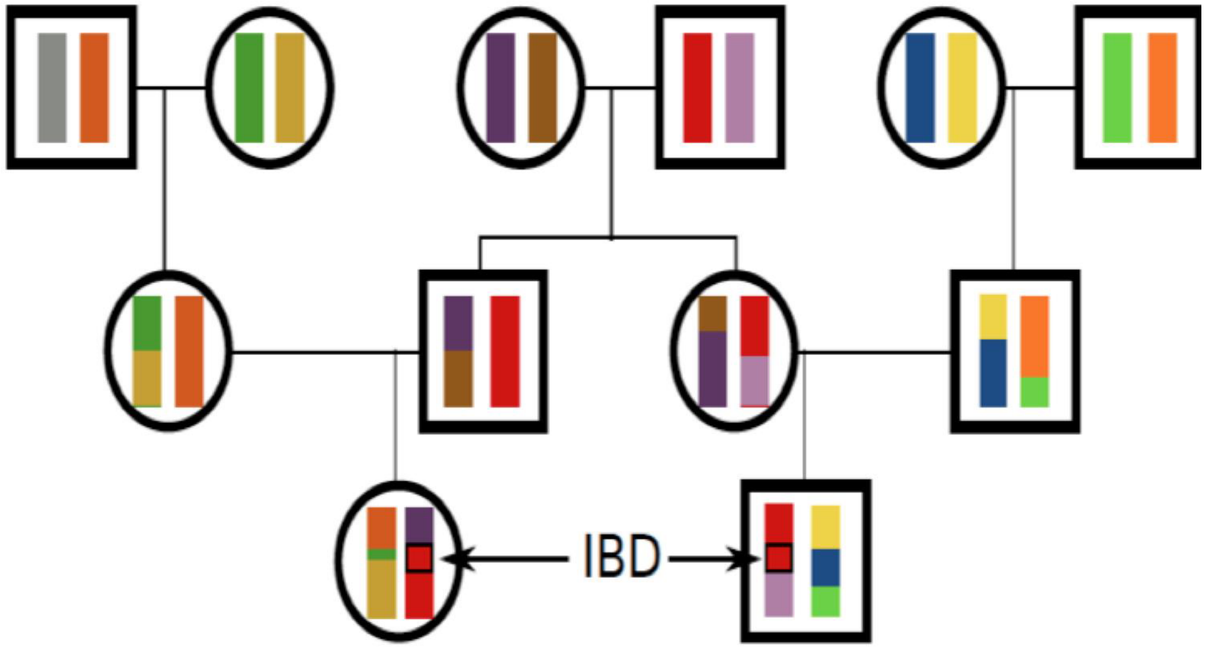


Figure: Graham Coop. Notes on Population Genetics

Предковый граф рекомбинаций

- В основе генома – сложная структура данных.
- Разные участки гаплотипов происходят от разных предков.

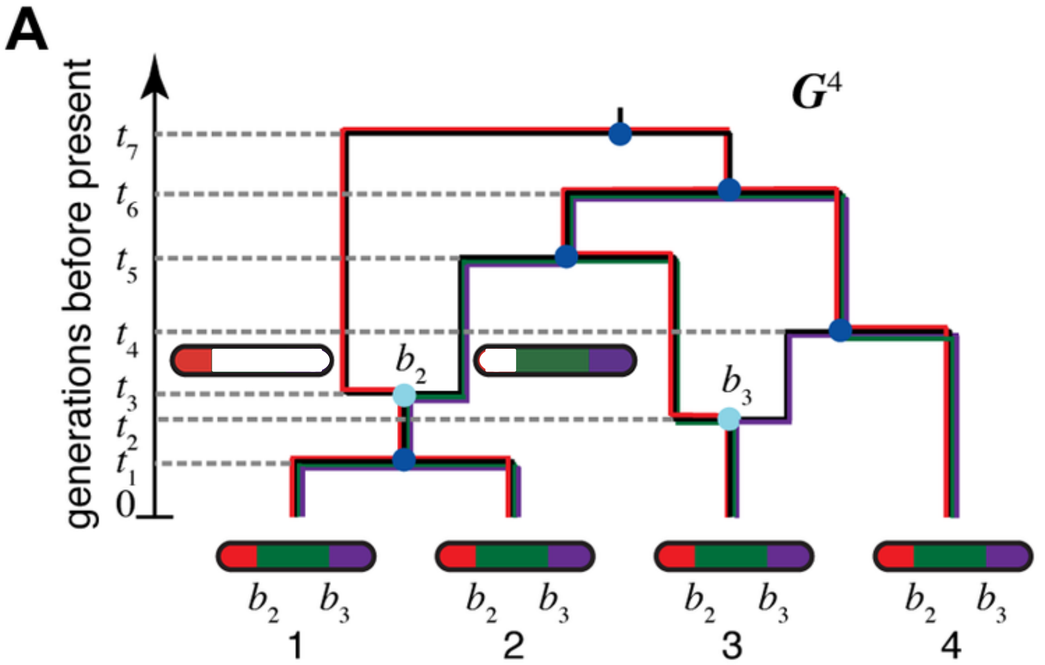


Рисунок: Rasmussen et al. (2014) Genome-Wide Inference of
Ancestral Recombination Graphs. PLoS Genetics

Предковый граф рекомбинаций

- В основе генома – сложная структура данных.
- Разные участки гаплотипов происходят от разных предков.

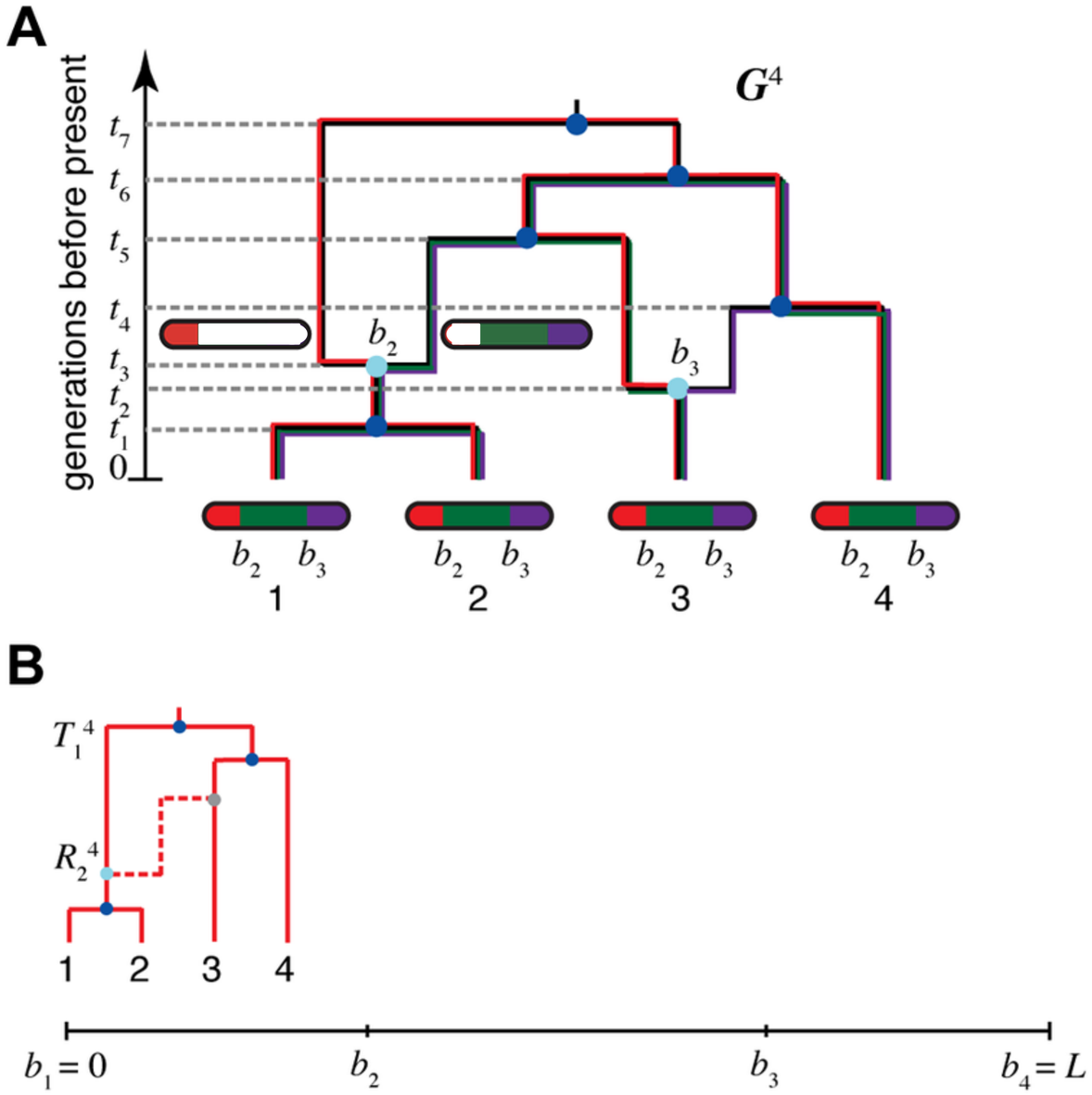


Рисунок: Rasmussen et al. (2014) Genome-Wide Inference of
Ancestral Recombination Graphs. PLoS Genetics

Предковый граф рекомбинаций

- В основе генома – сложная структура данных.
- Разные участки гаплотипов происходят от разных предков.

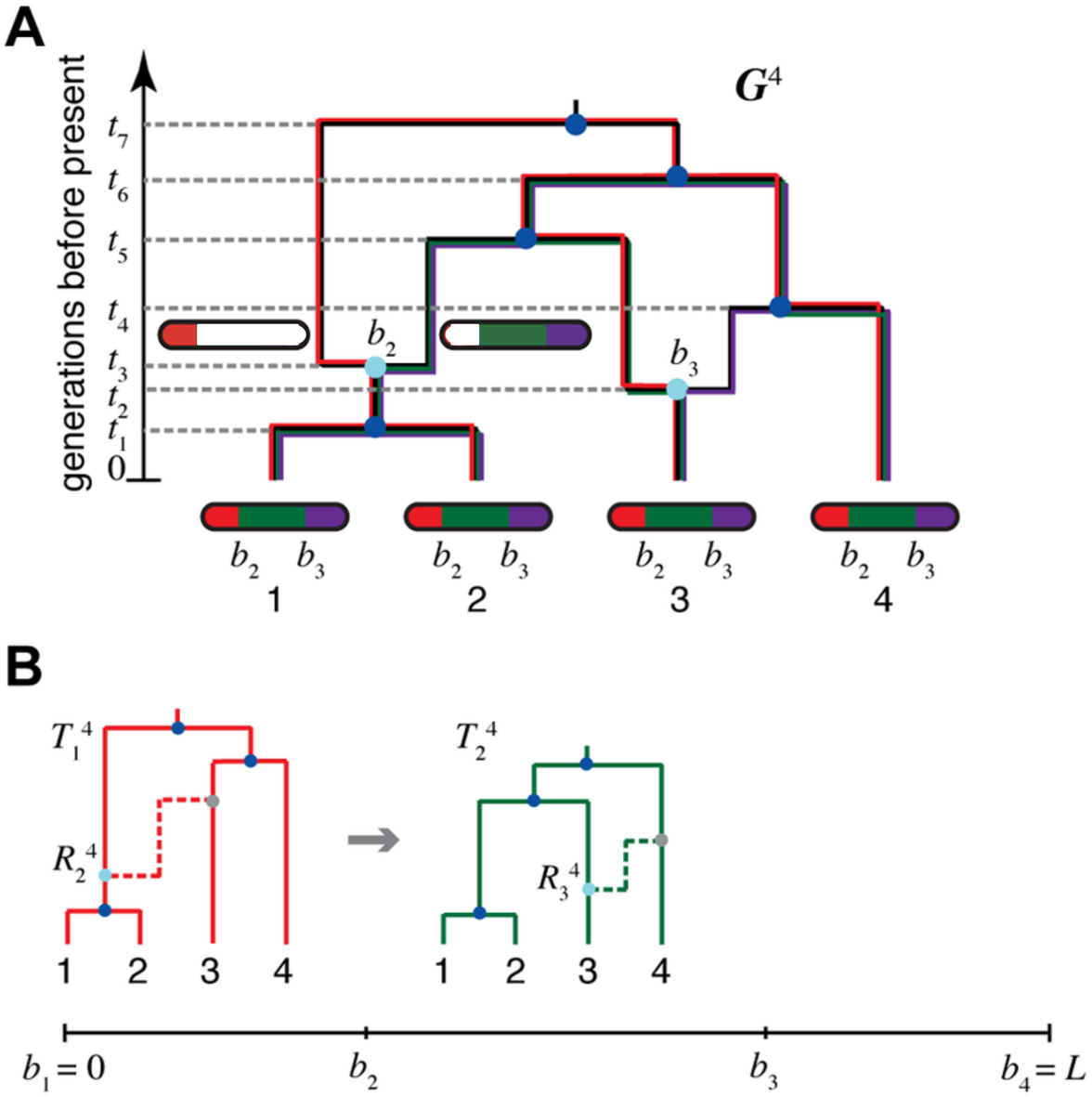


Рисунок: Rasmussen et al. (2014) Genome-Wide Inference of
Ancestral Recombination Graphs. PLoS Genetics

Предковый граф рекомбинаций

- В основе генома – сложная структура данных.
- Разные участки гаплотипов происходят от разных предков.

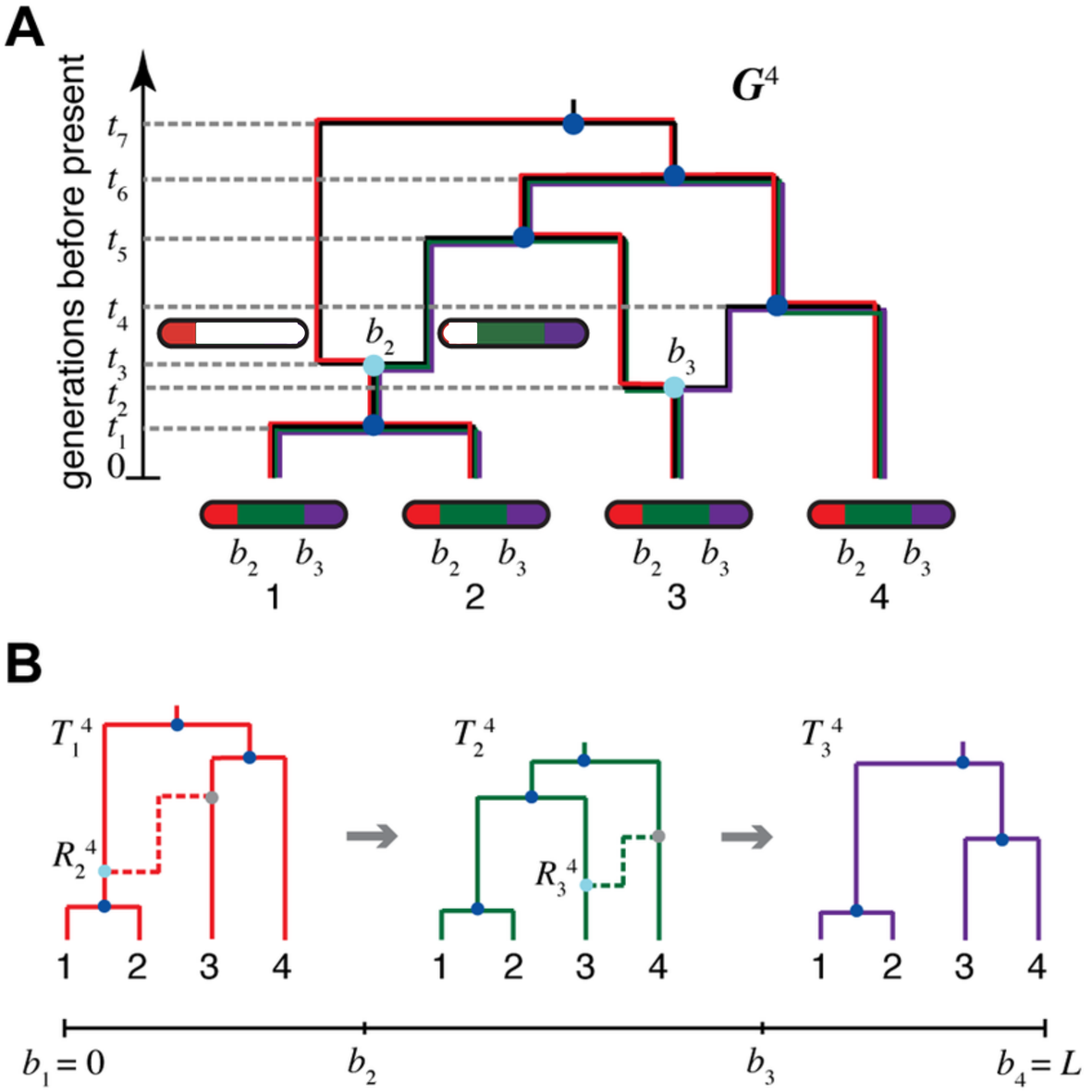


Рисунок: Rasmussen et al. (2014) Genome-Wide Inference of
Ancestral Recombination Graphs. PLoS Genetics

IBD graph and Graph Neural Networks (GNN)

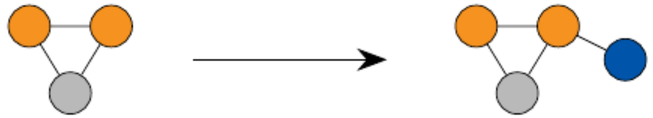
Assumptions
- C populations
- Each individual is of “pure” ancestry

IBD graph G
- Vertices are individuals
- Edges – shared IBD segments
- Weights – total length of IBD segments

“Semi-inductive” learning
- Reference graph G_r – frozen
- Add a single individual to the reference G_r to make prediction.

No node feature and unlabeled nodes

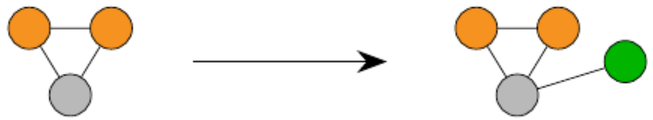
Individual A case



Initial dataset (training data)

Initial dataset with individual A

Individual B case



Initial dataset (training data)

Initial dataset with individual B

We need to use such feature generation algorithm that saves features of training nodes but can be generalized to new graph structure!

+

Somehow deal with nodes with no classes.

Graph based features

Neighbor count: $n_{i,c} = \sum_{j \in N_i} 1\{L_j = c\},$

Average edge weight: $\bar{w}_{i,c} = \frac{1}{n_{i,c}} \sum_{\substack{j \in N_i \\ L_j = c}} W(i,j),$

Maximum edge weight: $w_{i,c}^{\max} = \max_{\substack{j \in N_i \\ L_j = c}} W(i,j),$

Standard deviation of edge weights: $\sigma_{i,c} = \sqrt{\frac{1}{n_{i,c}} \sum_{\substack{j \in N_i \\ L_j = c}} \Big(W(i,j) - \bar{w}_{i,c}\Big)^2},$

Average number of IBD segments: $\bar{\mathrm{IBD}}_{i,c} = \frac{1}{n_{i,c}} \sum_{\substack{j \in N_i \\ L_j = c}} \mathrm{IBD}(i,j).$

$$\mathbf{x}_i = \Big(n_{i,1}, \ldots, n_{i,C},\ \bar{w}_{i,1}, \ldots, \bar{w}_{i,C},\ \sigma_{i,1}, \ldots, \sigma_{i,C},\ w_{i,1}^{\max}, \ldots, w_{i,C}^{\max},\ \bar{\mathrm{IBD}}_{i,1}, \ldots, \bar{\mathrm{IBD}}_{i,C}\Big)^\top$$

One hot features

For known labels:

$$x_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$$

For unknown labels:

$$x_i = \left(\frac{1}{C}, \frac{1}{C}, \ldots, \frac{1}{C}\right)^\top$$

East Slavs dataset

% unlabeled nodes	GNN (graph-based)	GNN (one-hot)	MLP	LR	Community Detection
0%	0.6146 ± 0.02	0.6165 ± 0.02	0.6125 ± 0.02	0.6059 ± 0.02	0.5307 ± 0.01
1%	0.6285 ± 0.01	0.6255 ± 0.01			
5%	0.6500 ± 0.02	0.6467 ± 0.02			
25%	0.6844 ± 0.02	0.6672 ± 0.01		*NA*	
50%	0.6962 ± 0.02	0.6765 ± 0.01			
100%	$\mathbf{0.7135 \pm 0.01}$	0.6825 ± 0.01			

And now diving into the past

Neanderthals

- We know about them from archeological evidence.
- The first recognized fossil discovered in 1856 in Germany.
- Neanderthals lived from Atlantics to Altai.

- Are they our ancestors?
- Or is it a completely independent human branch which we've never met?

Figures: CNN; ResearchGate; Natural History museum.

Neanderthals

Until 2010 there were only controversial hypothesis.
Genomics solves this mystery.

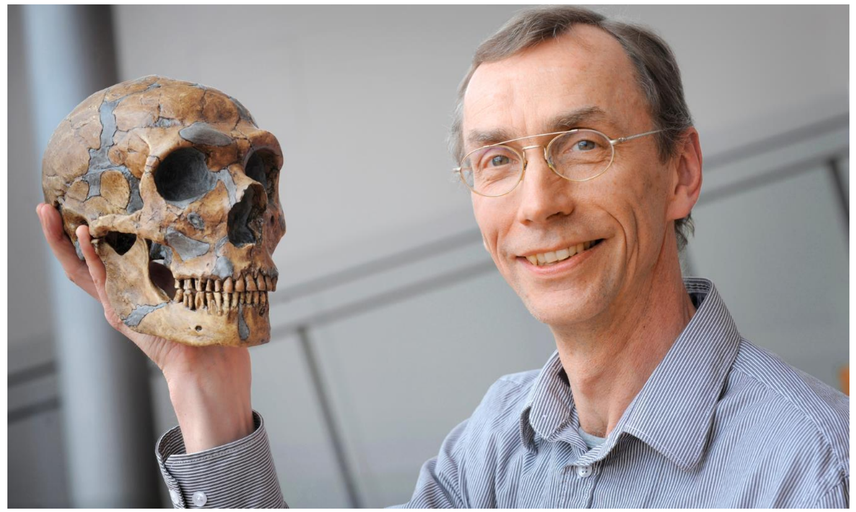
- Ancestors of Neanderthal and anatomically modern human diverges 550-850 kya.
- They met again around 55kya.
- All humans of non-African origin have on average 1-3% of Neanderthal genome.

kya = thousand years ago

Genomes of archaic humans

- Mid 80-s: first attempts to sequence ancient DNA (Cell).
- 2010 – draft of the first Neanderthal genome (Science).
- 2012 – Denisovan genome (Science)
- 2013 – the whole Altai Neanderthal genome (Nature)
- 2017 – the whole Vindija (cave in Chroatia) Neanderthal genome (Science)
- 2018 – the genome of a daughter of a Neanderthal mother and Denisovan father (Nature)

Denisova cave, Altai



Svante Pääbo
Nobel Prize in medicine 2022

How to infer genome segments (tracts) of archaic origin?

Method development

Anna Ilina
(HSE University)

Léo Planche
(University Paris Saclay)

Biological interpretation

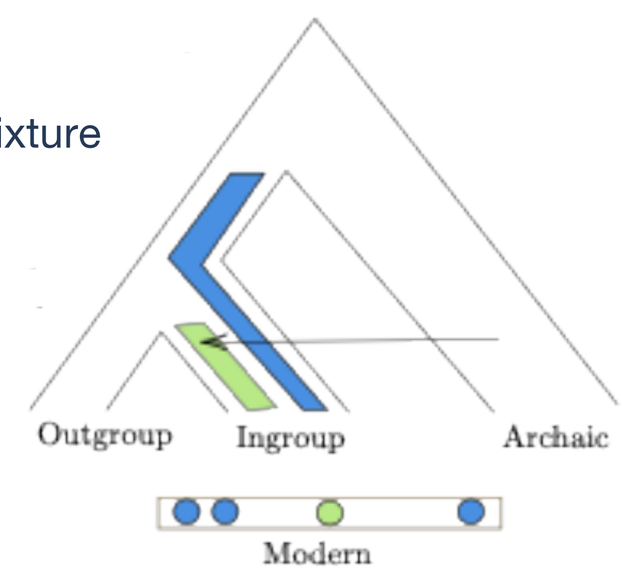
Emilia Huerta-Sanchez
(Brown University and
Trinity College Dublin)

Linda Ongaro
(Trinity College Dublin)

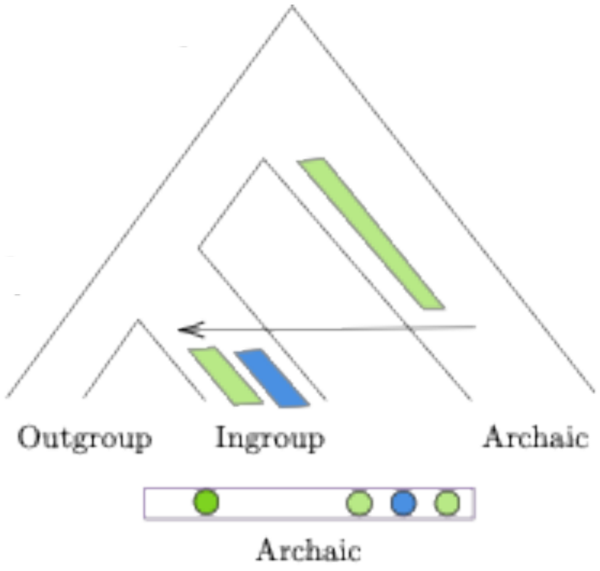
DAIseg: method for inferring archaic tracts in modern humans

Hidden Markov Model
- States: ancestries
- Emissions: bins of 1000bp summarized as tuples of the number of private variants relatively to each reference population (outgroups or archaic)

- Three versions available:
  - ✓ single archaic introgression,
  - ✓ multiple archaic introgressions,
  - ✓ joint inference of archaic and modern admixture
- Multiple outgroups
- Archaic genomes can be used as reference



Emission (3, 1)

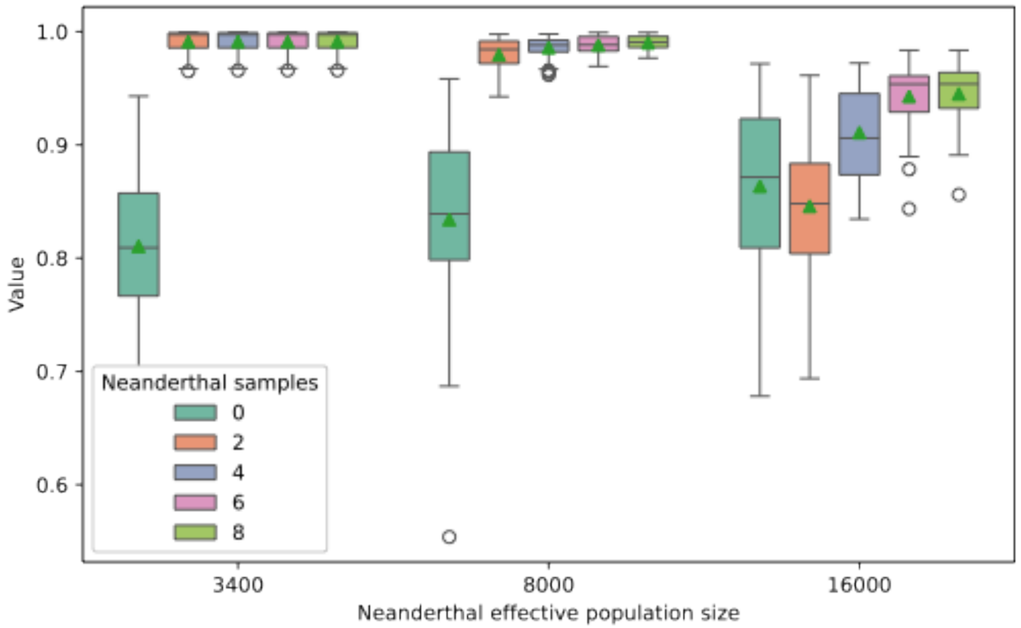


(1, 3)

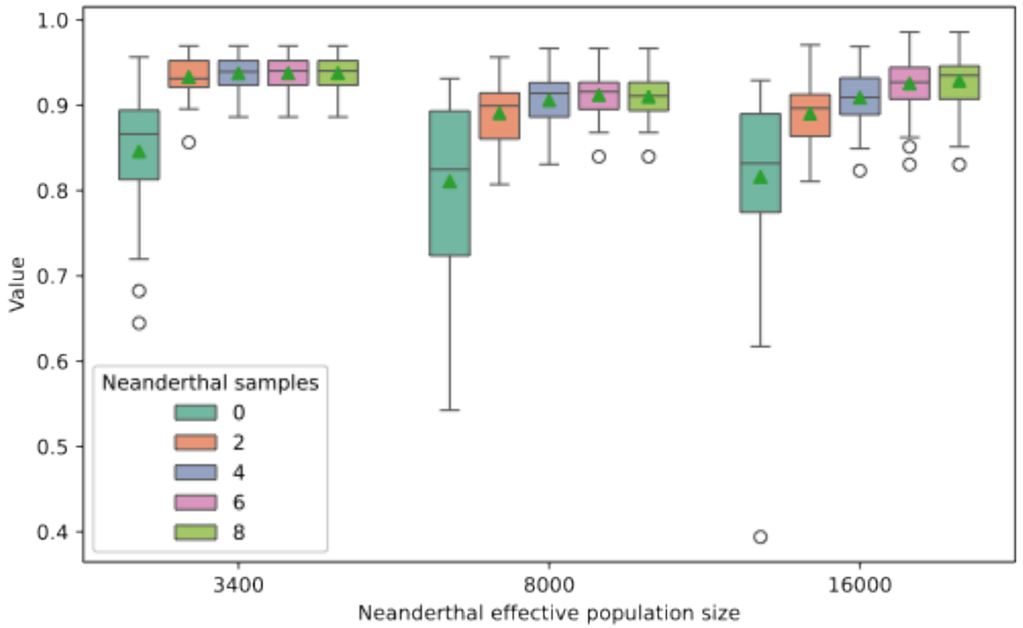
DAIseg: detecting archaically introgressed segments
also in Russian дай [dai] – give *(me that segment)*

DAIseg: precision and recall on simulated data (single archaic introgression)

Precision



Recall

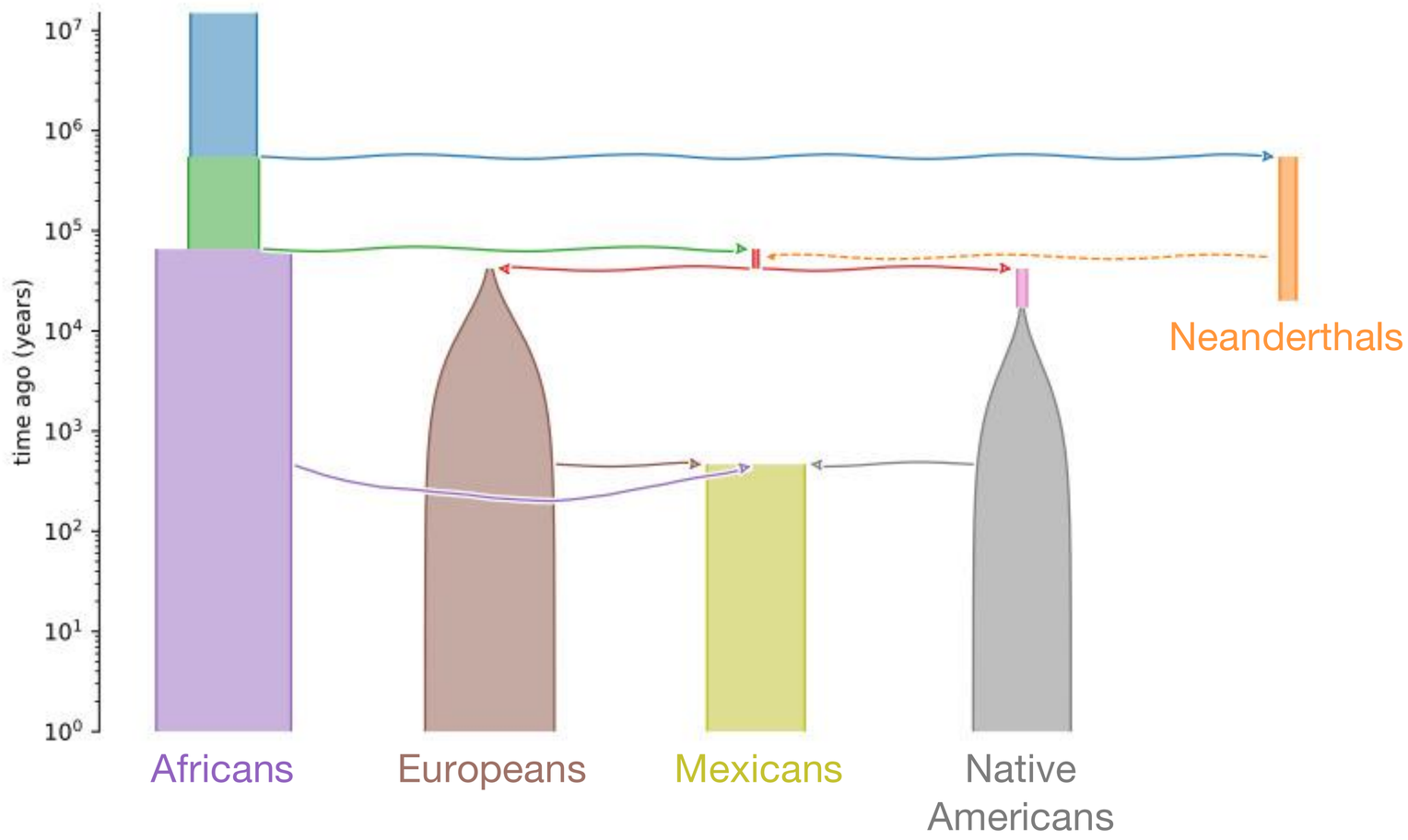


Takeaway

Including archaic samples in the reference panel increases both precision and recall.

Archaic tracts in recently admixed populations

Inference of both the archaic tracts and their modern source population



Hidden states - ancestries

European
Native American
African
} Modern

Euro-Neanderthal
NA-Neanderthal
} Archaic

Emissions – bins of 1000bp

tuples `(i, j, k, l)` with counts of private variants relatively to

`i` – Europeans
`j` – Native Americans
`k` – Africans
`l` – Neanderthals

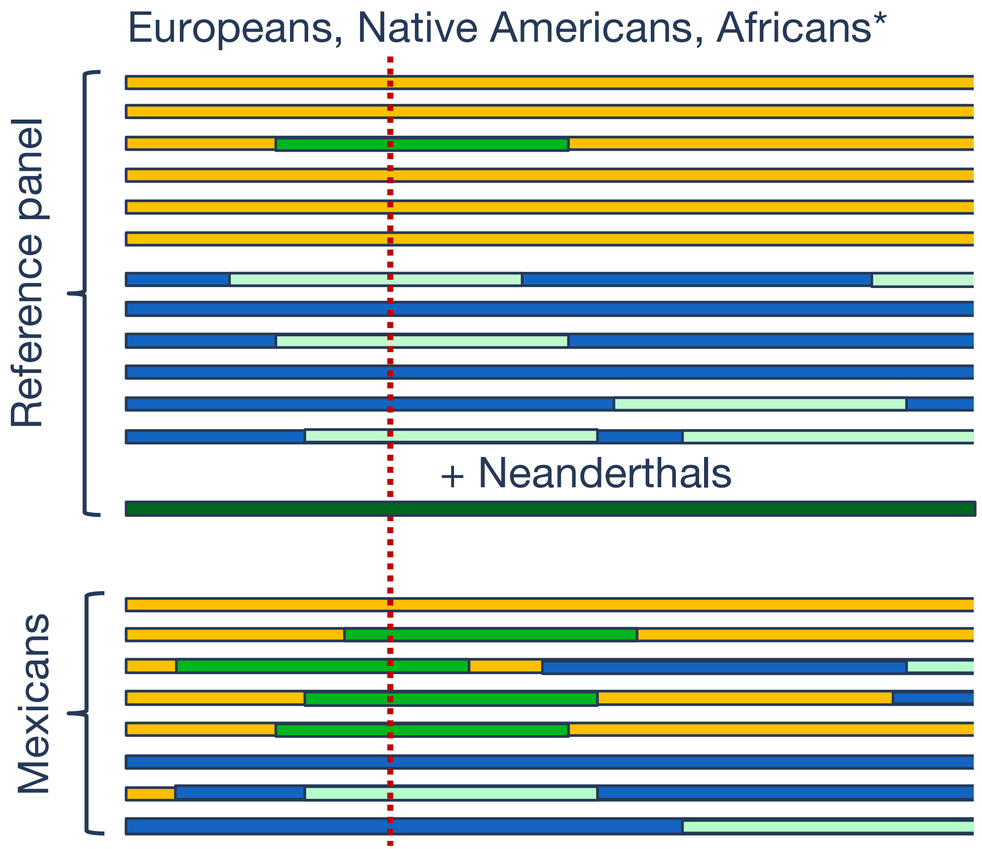
Precision and recall on simulated data

Simulations were performed with *msprime* using parameters recommended by *stdpopsim* community.

Ancestry	Precision (sd)	Recall (sd)
Euro Neanderthal	0.987 (0.015)	0.937 (0.04)
Native American Neanderthal	0.988 (0.014)	0.931 (0.068)

https://tskit.dev/software/msprime.html
https://popsim-consortium.github.io/

Selection on archaic component in recently admixed population

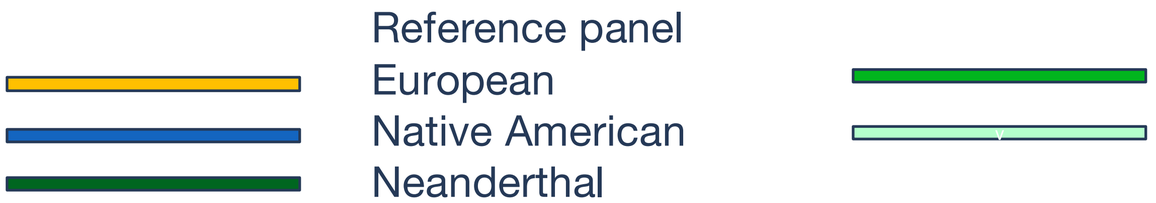


Toy example
Consider the locus marked by the red dashed line

ND frequency in the reference panel *0.67*
Euro-ND + NA-ND frequency in Mexicans *0.63*

ND frequency in Europeans *0.17*
Euro-ND frequency in the admixed population *0.8*

No signal based on total ND ancestry at the locus
But there is a Euro-ND specific ancestry signal for **post-admixture positive selection**



*Africans are not shown on the Figure for clarity

Selection on Euro-Neanderthal ancestry in Mexicans

We discovered selection signal in 22 genomic regions (protein-coding genes and long non-coding RNAs) using 1000 Genome Project data.

Positive selection

Chr 15: *TCF12* (transcriptional regulation; craniosynostosis) and *UNC13C* (associated with Autism Spectrum Disorder)

Chr 5: *IL17B (*Interleukin 17B, T-cell cytokine; leiomyoma and psoriasis).

Chr 2: *PSMD14* (cystic fibrosis and Machado-Joseph disease, a rare neurological disorder)

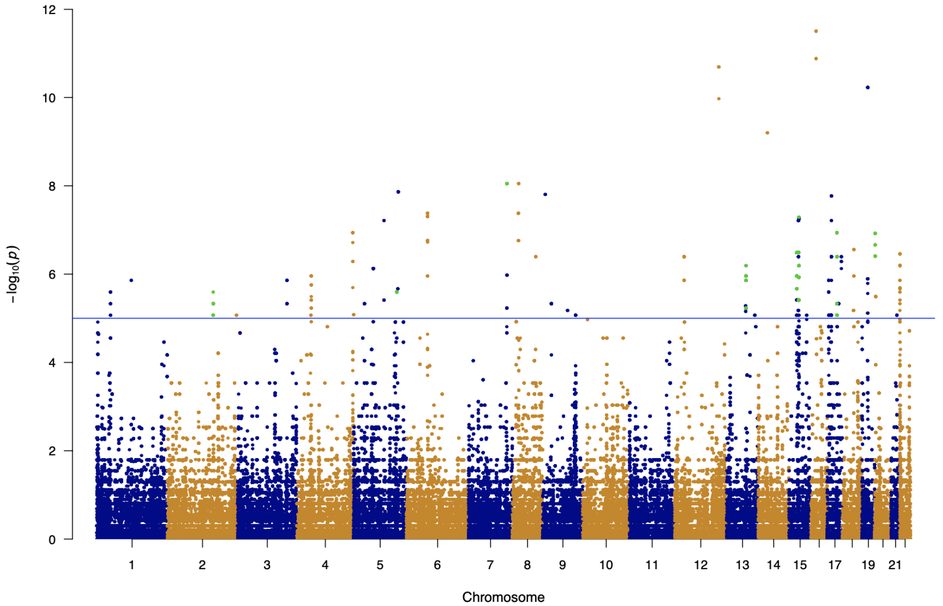
Chr 13: *PIBF1* (pregnancy maintenance).

Negative selection

Chr 3: *NEK10* (environmental stress) and *SLC4A7* (ion transport).

Chr 4: *TLR6* (immune system activation) - appears to be under purifying selection.

Chr 2: *POLR2D* (Type 2 Diabetes Mellitus).



What can we learn about Neandertals from modern humans?

Anna Ilina
(HSE University)

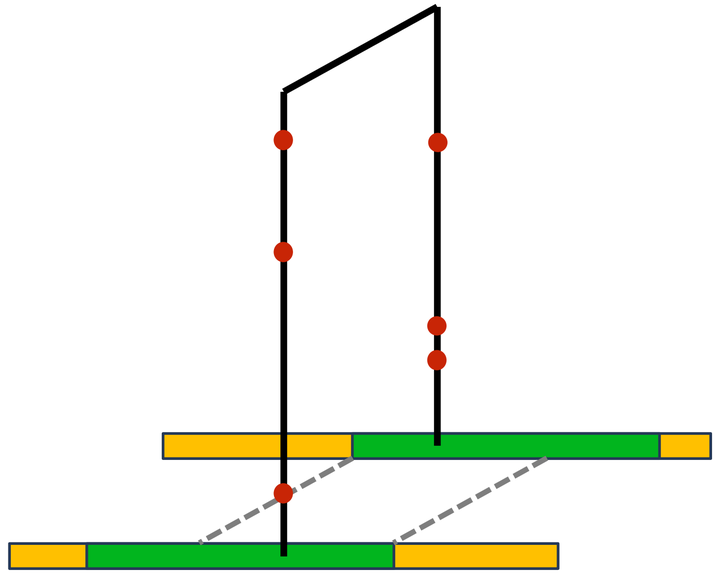
Egor Lappo
(Stanford University)

Demographic inference for Neanderthals

Approach

For each pairwise intersection of archaic segments, one can write down the likelihood for the divergence between the two sequences that depends on the demographic parameters of the Neanderthal population and admixture fraction. We assume infinitely-many-sites mutation model and piecewise-constant population size.

Composite likelihood for all pairwise intersections can be used to infer Neanderthal population size at admixture.

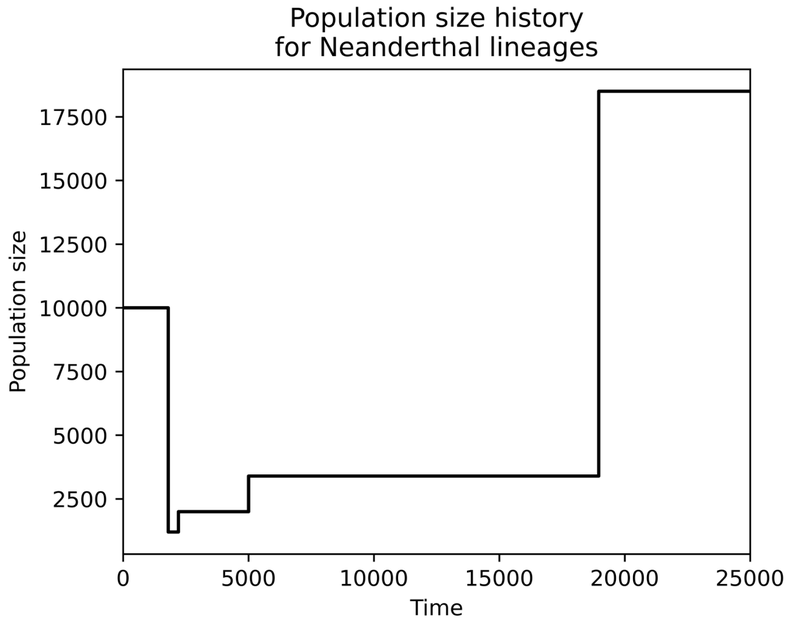


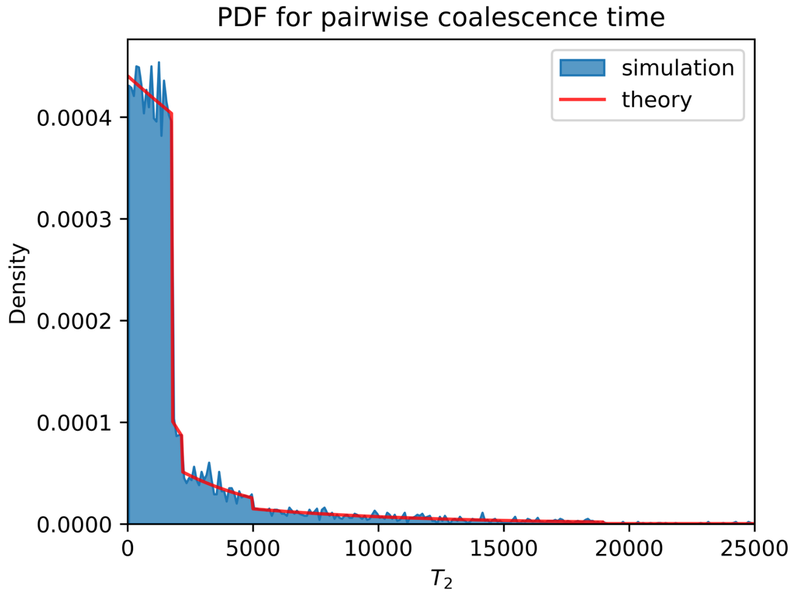
$$P_{\text{Pois}}(K = k \mid N(t), p_{\text{admix}}, \mu) = \sum_{m=1}^{M} \left(\frac{p_{\text{admix}}^{1-\delta_{m1}}}{P_{\text{arch}}} \prod_{i=1}^{m-1} e^{-(t_i - t_{i-1})/2N_i} \times \int_{t_{m-1}}^{t_m} \frac{(2\mu\tau)^k e^{-2\mu\tau}}{2N_m k!} e^{-(\tau - t_{m-1})/2N_m} \, d\tau \right),$$

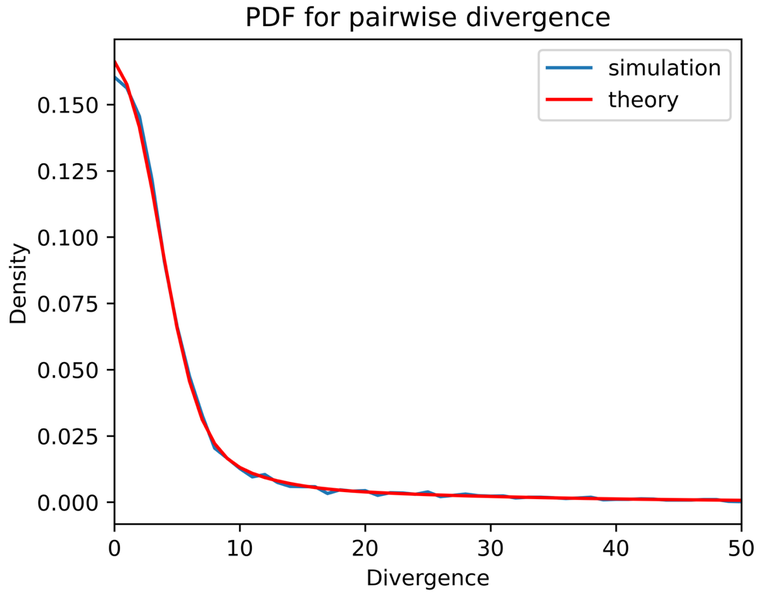
$$P_{\text{arch}} = \left(1 - e^{-t_1/2N_1}\right) + p_{\text{admix}} e^{-t_1/2N_1}$$

Demographic inference for Neanderthals

Theoretical results. Analytical expressions for distributions of coalescence time and divergence between two intersecting archaic segments.

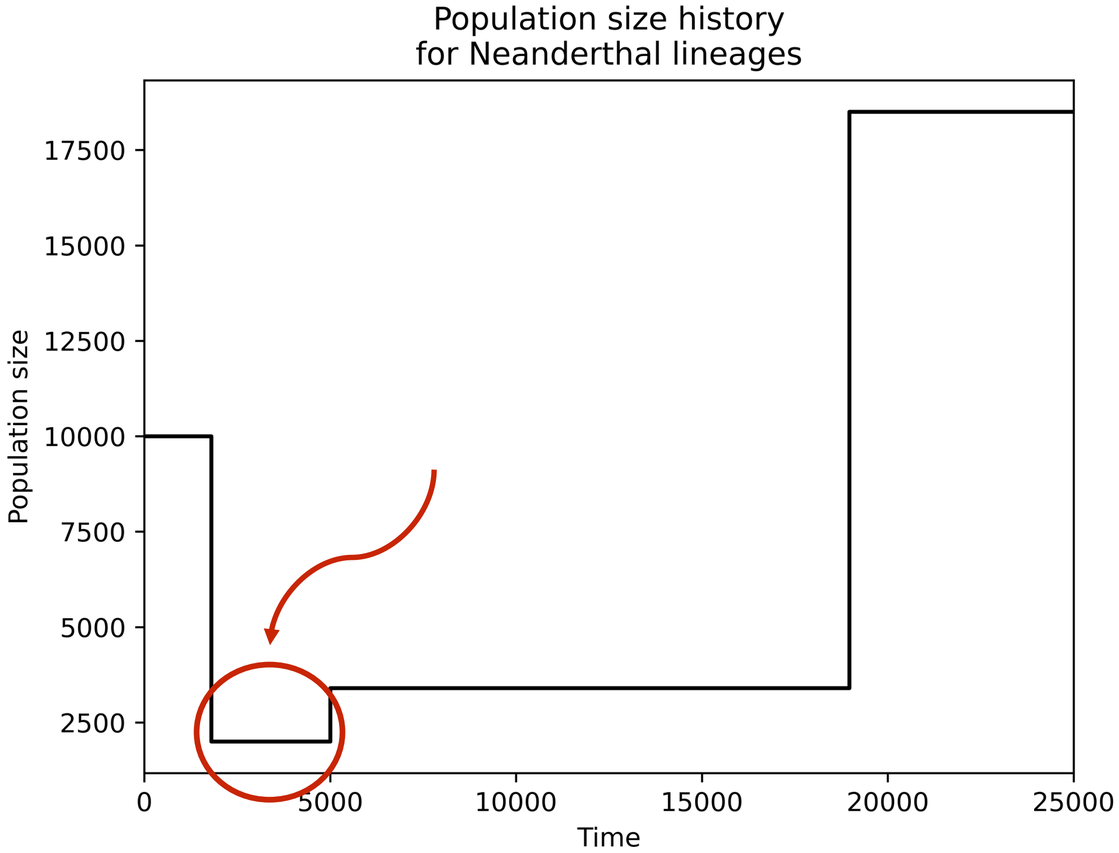


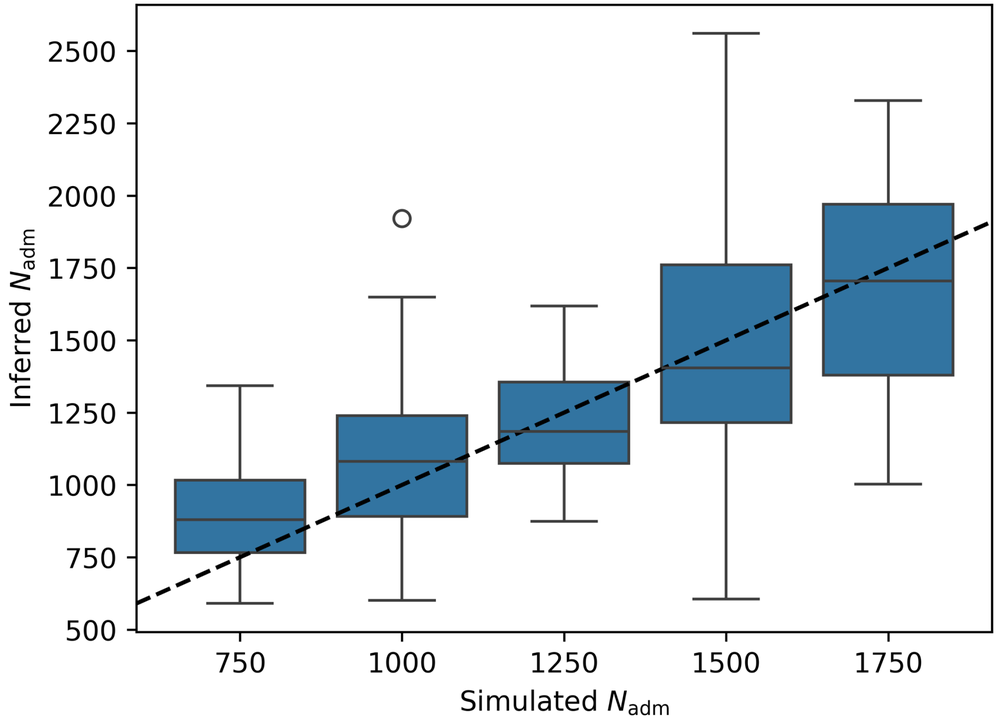




Demographic inference for Neanderthals

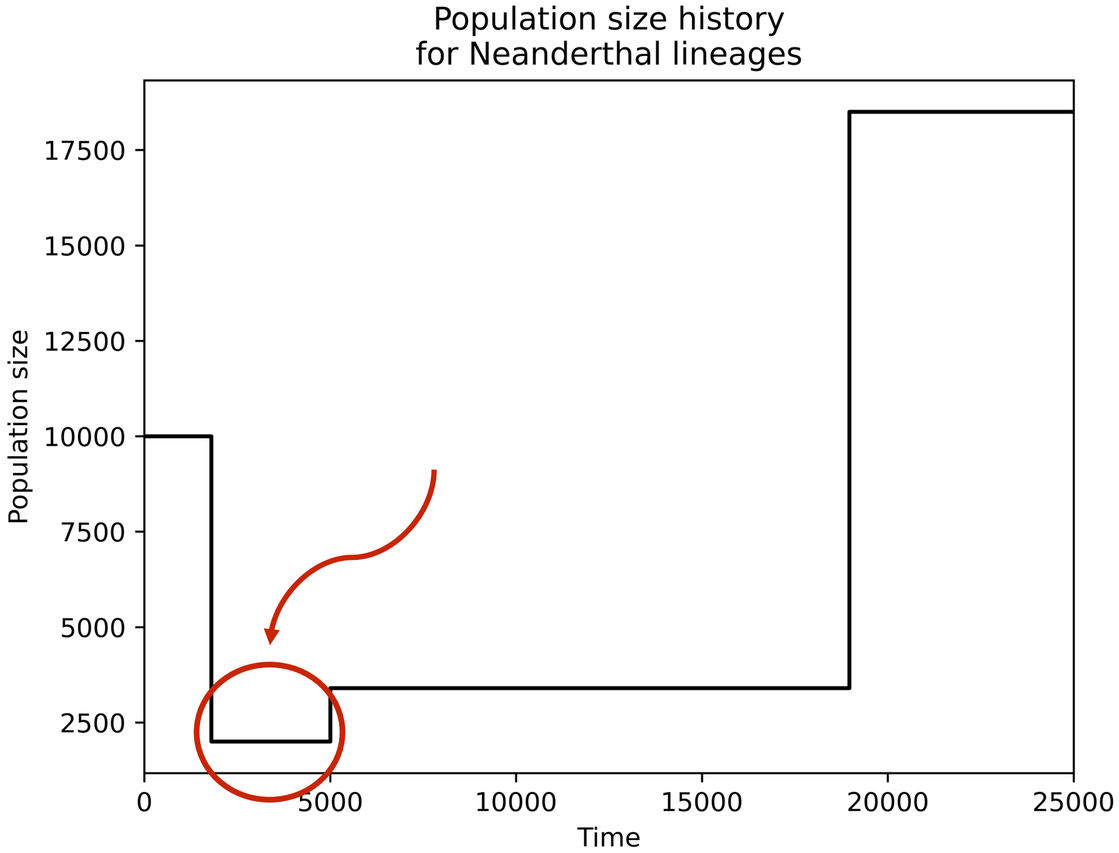
Inference on simulations. Assume that Neanderthals went through a bottleneck prior to admixture. Numerical optimization of composite likelihood recovers bottleneck population size in simulations.





Demographic inference for Neanderthals

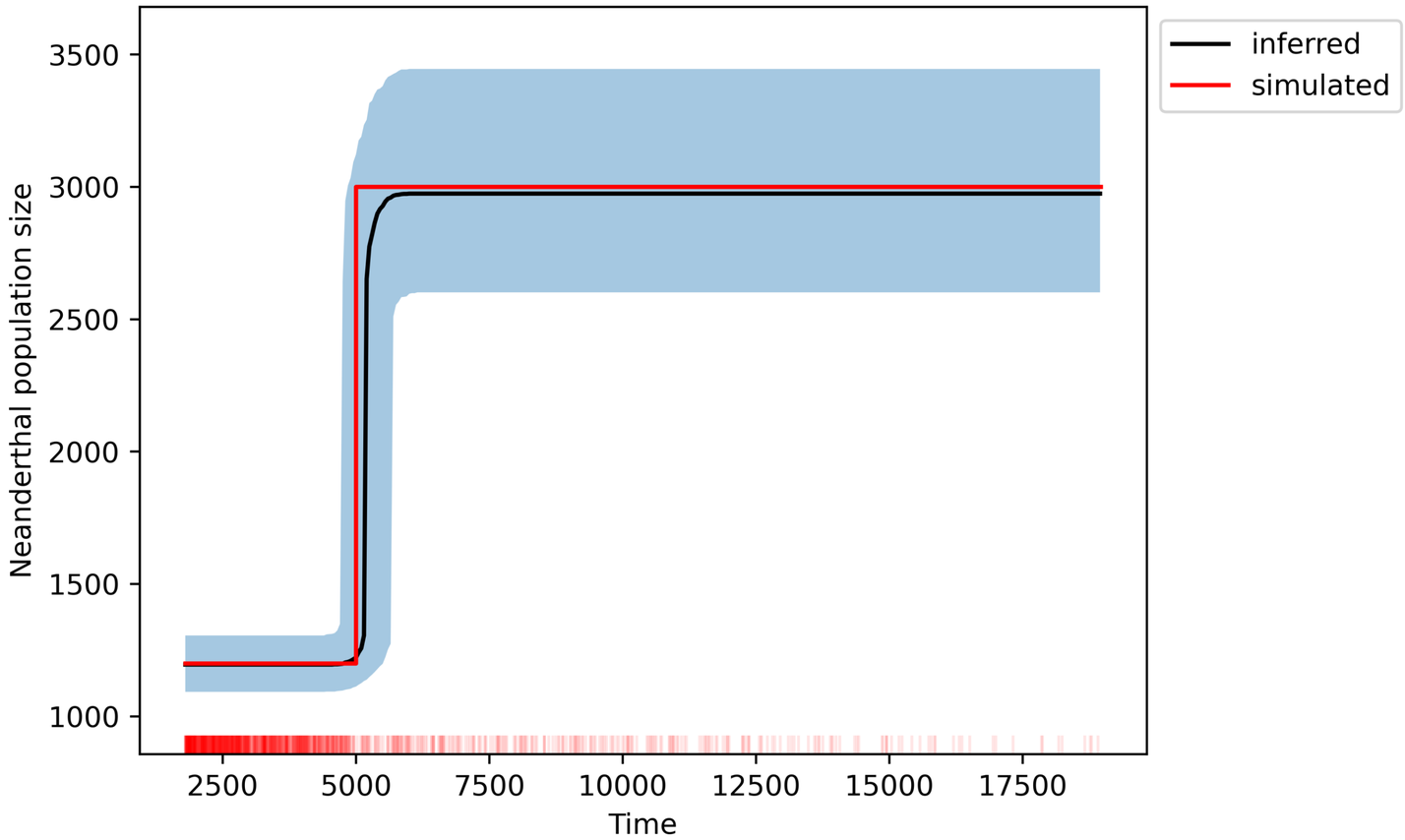
Inference on multiple 1000 Genomes populations shows different bottleneck strengths for different populations.



Population	Inferred N_{adm}
IBS	2670.30 [2650.62, 2690.84]
TSI	2696.91 [2681.59, 2718.64]
GBR	2669.23 [2650.30, 2689.27]
FIN	3013.27 [2988.59, 3033.88]
CHB	3638.00 [3617.10, 3666.16]
CHS	3609.15 [3588.72, 3629.66]

Demographic inference for Neanderthals

PRELIMINARY. Inference of more complicated demographic scenarios with an MCMC approach allows us to also date the timing of the bottleneck.



We are hiring

Мы ищем математиков на позиции пост-доков (научных сотрудников) и аспирантов

- Палеогеномика
- Геномная эпидемиология
- Эволюция прокариот и Crispr-Cas системы

vshchur@hse.ru

@VLSHCHUR

Знаний биологии не требуется!

NATIONAL RESEARCH
UNIVERSITY