

Генетическая генеалогия

Александр Ракитъко



What do we do?

Personal DNA tests

Microarray genotyping

- complex traits
- ancestry composition
- carrier screening

Clinical genetics

Next Generation Sequencing

- hereditary diseases
- NIPT

Research studies

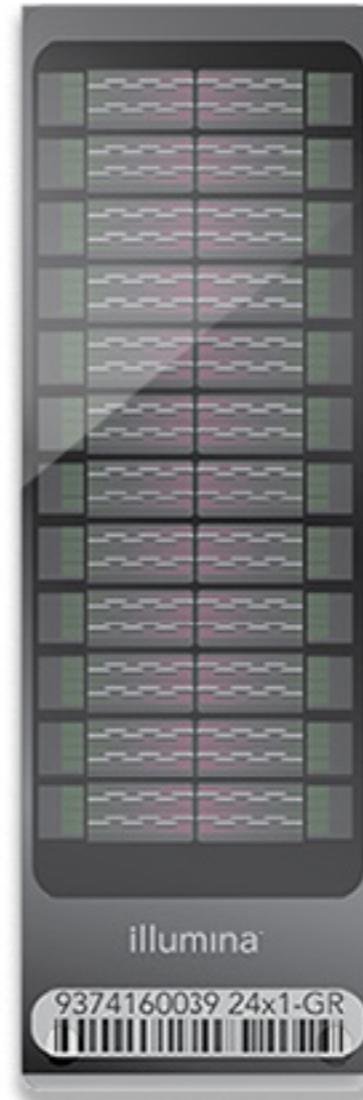
- DNA sequencing
- RNA-Seq
- Metagenome
- Methylome



MyHeritage ▶DNA



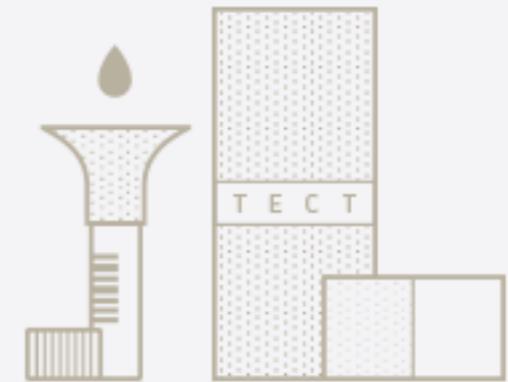
Genotek



Illumina Global Screening Array

Как проходит ДНК тестирование?

1. Сбор биоматериала



2. Микрочиповое генотипирование или секвенирование (NGS)



3. Результаты в личном кабинете
(генетика + опросник)

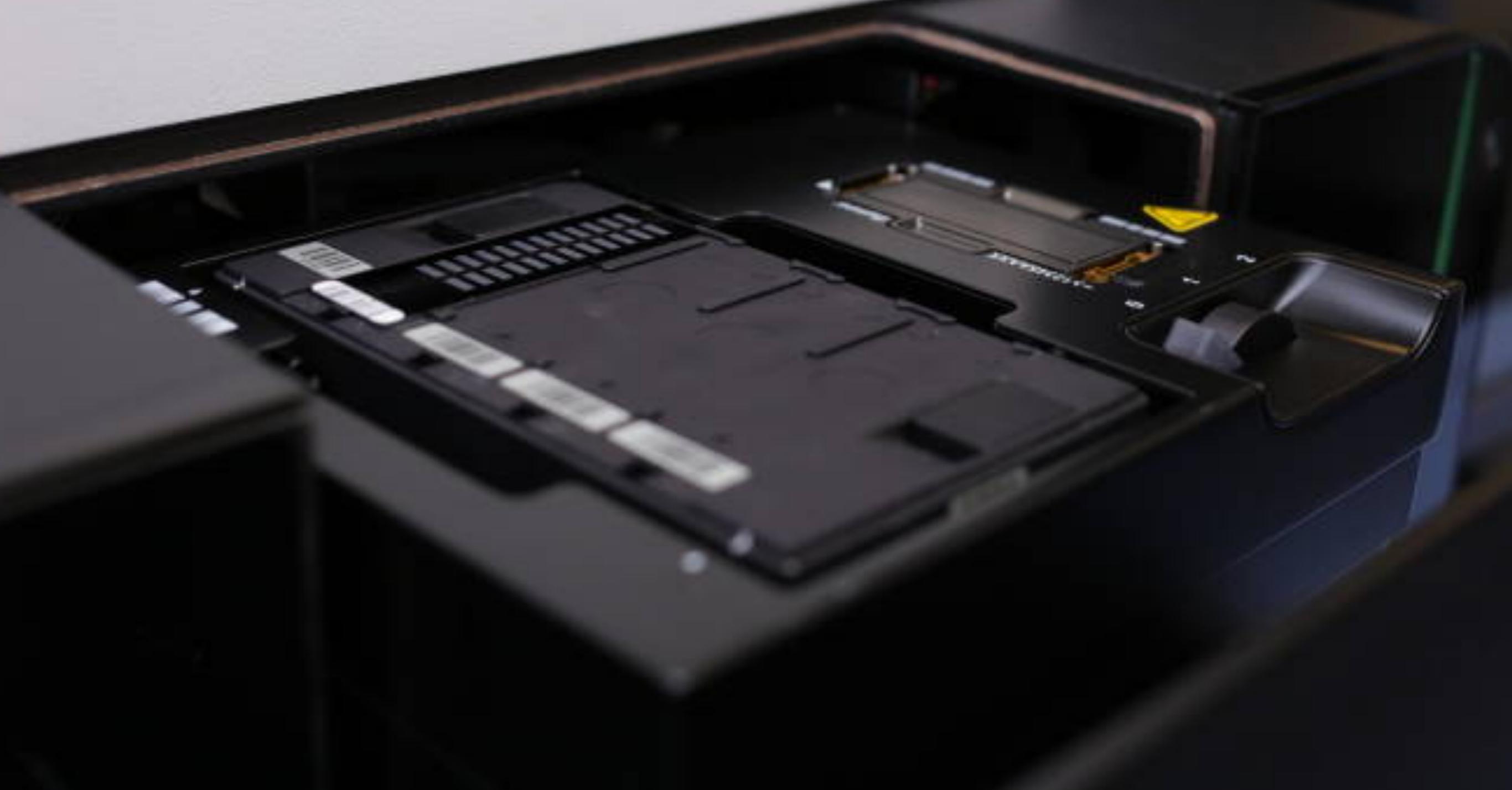


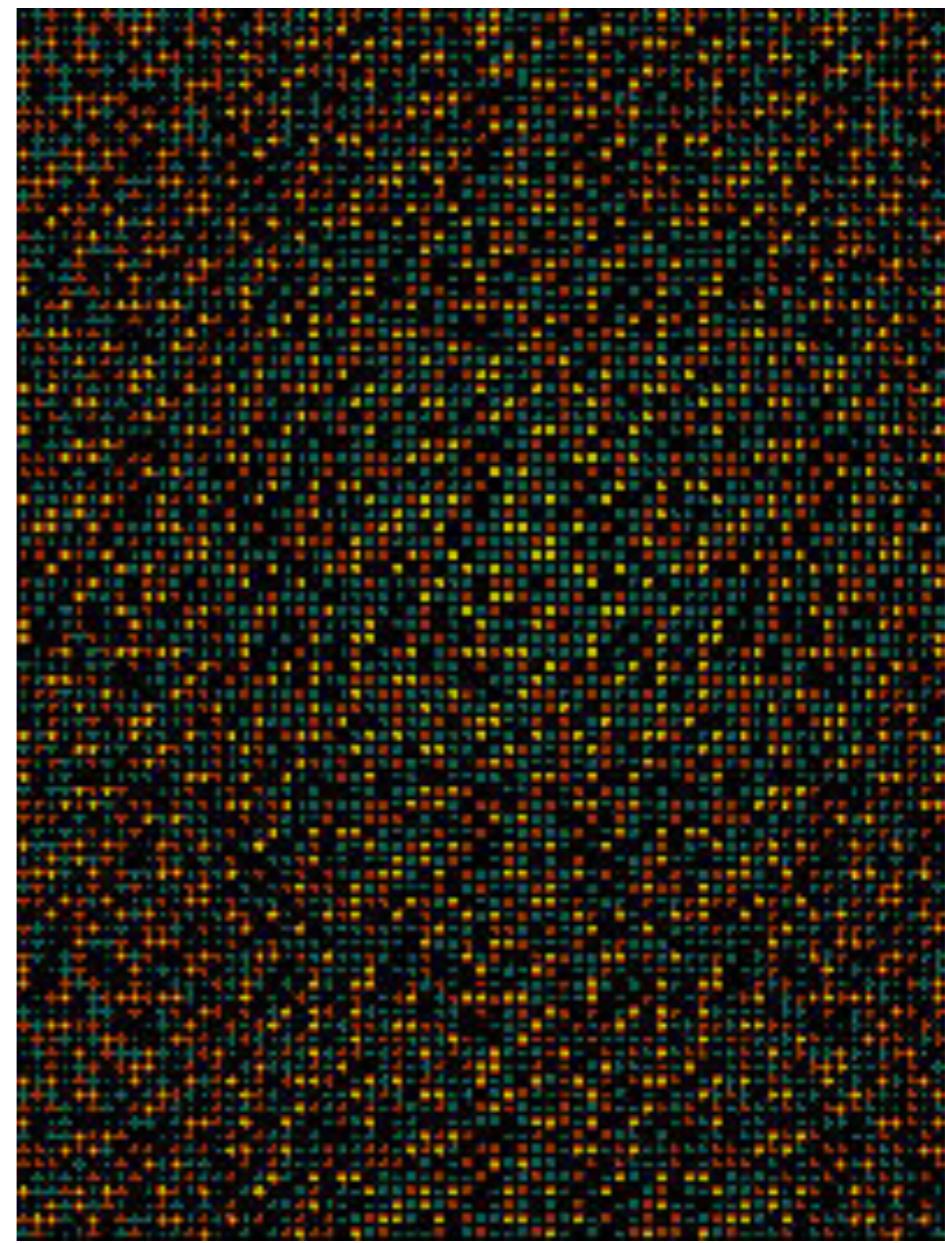
HiScan

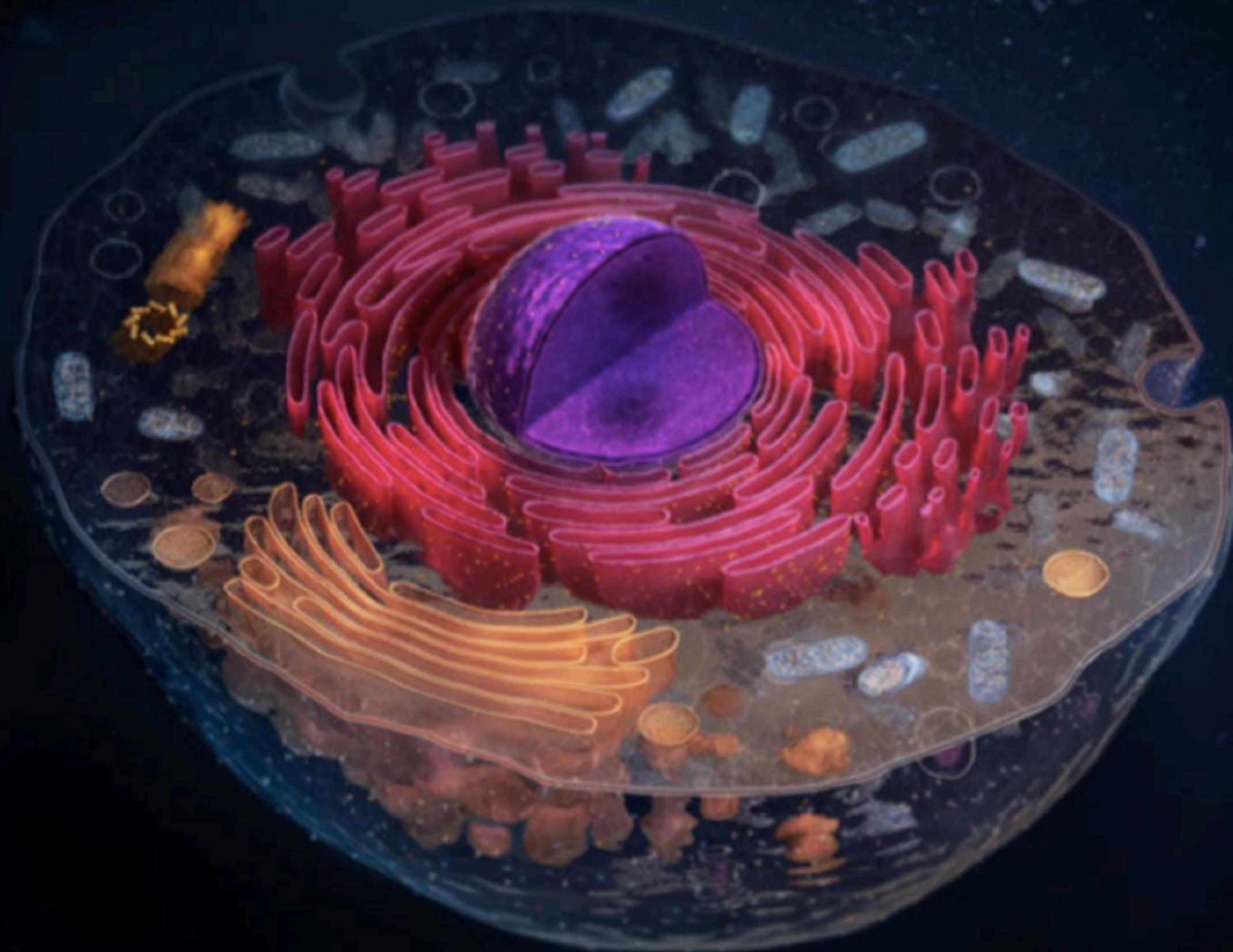
illumina[®]



illumina







Расшифровка генома в лаборатории



ЭТНИЧЕСКИЙ СОСТАВ

50 % ваших предков — армяне, азербайджанцы, персы и турки

50%

Армяне, азербайджанцы, персы и турки

30%

Белорусы, русские, украинцы

20%

Ашkenазские евреи



ПОИСК РОДСТВЕННИКОВ

Мы нашли 25 родственников

4 новых за предыдущий месяц

25 всего



РИСКИ ЗАБОЛЕВАНИЙ

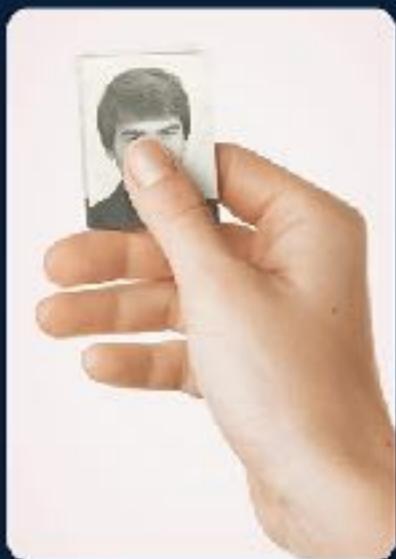
У вас 1 повышенный риск критического заболевания

2 высоких риска в группе «болезни кожи и подкожной клетчатки»

12 рекомендаций по профилактике заболеваний



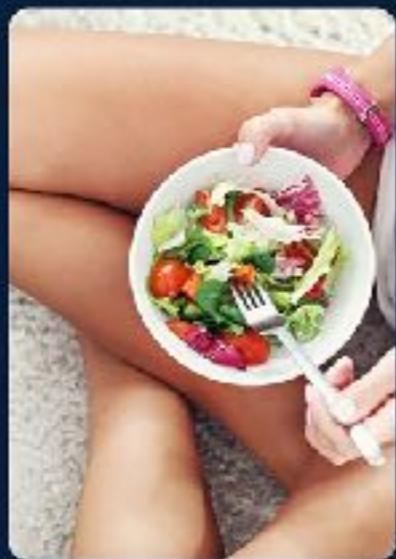
Генетический паспорт – это 7 ДНК-тестов в 1 + консультация с врачом



Происхождение



Риски заболеваний



Питание



Спорт



Эффективность лекарств



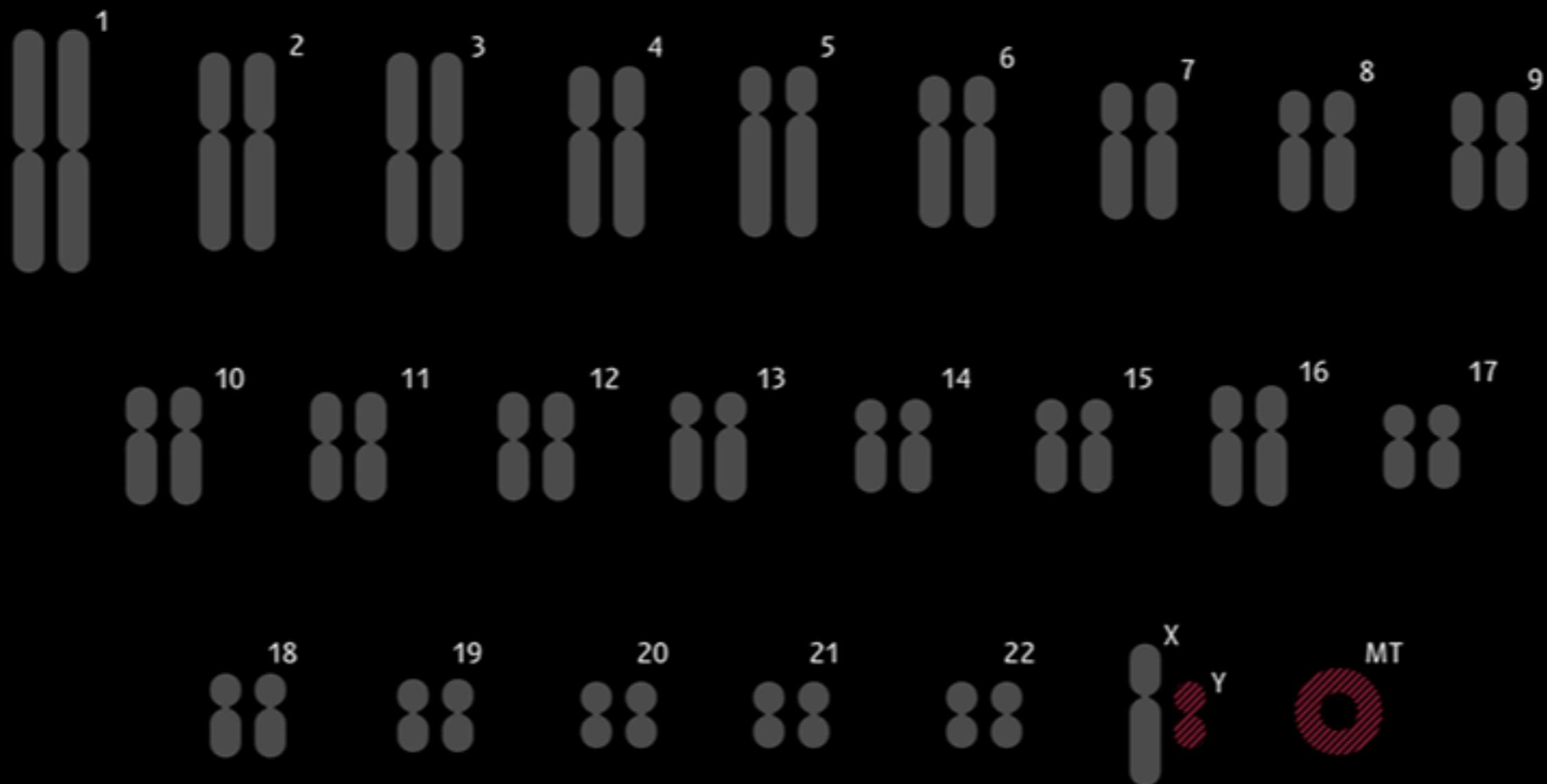
Планирование детей



Способности и характер



Консультация с врачом



77 000
маркеров этнического
происхождения (доступно
для мужчин и женщин)

1500
маркеров происхождения
по отцовской линии
(доступно только для
мужчин)

600
маркеров происхождения
по материнской линии
(доступно для мужчин и
женщин)

Что такое генетический вариант?



... AAT**A**CAGTGC...
... AATGCAGTGC...



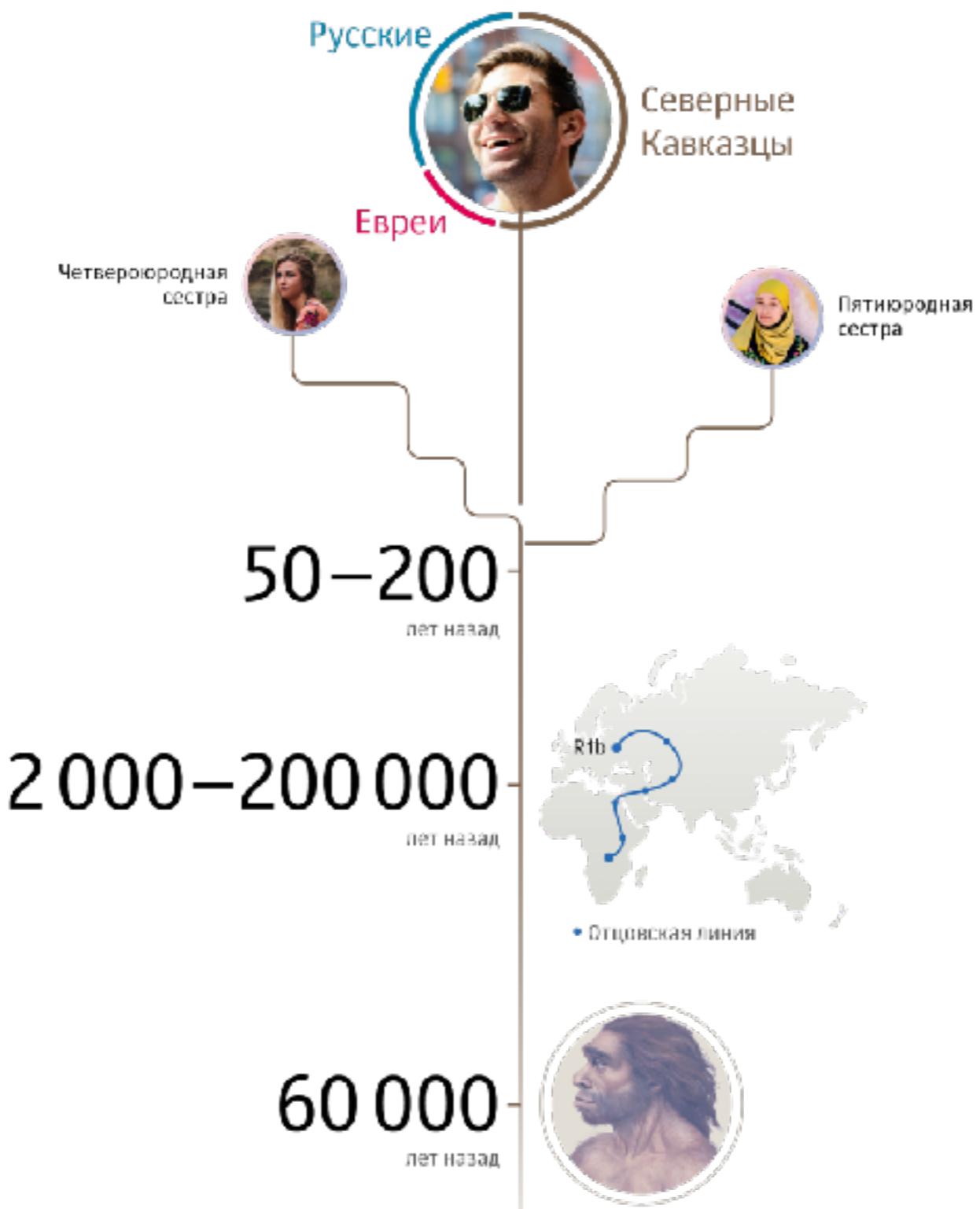
... AATGCAGTGC...
... AATGCAGT**CC**...



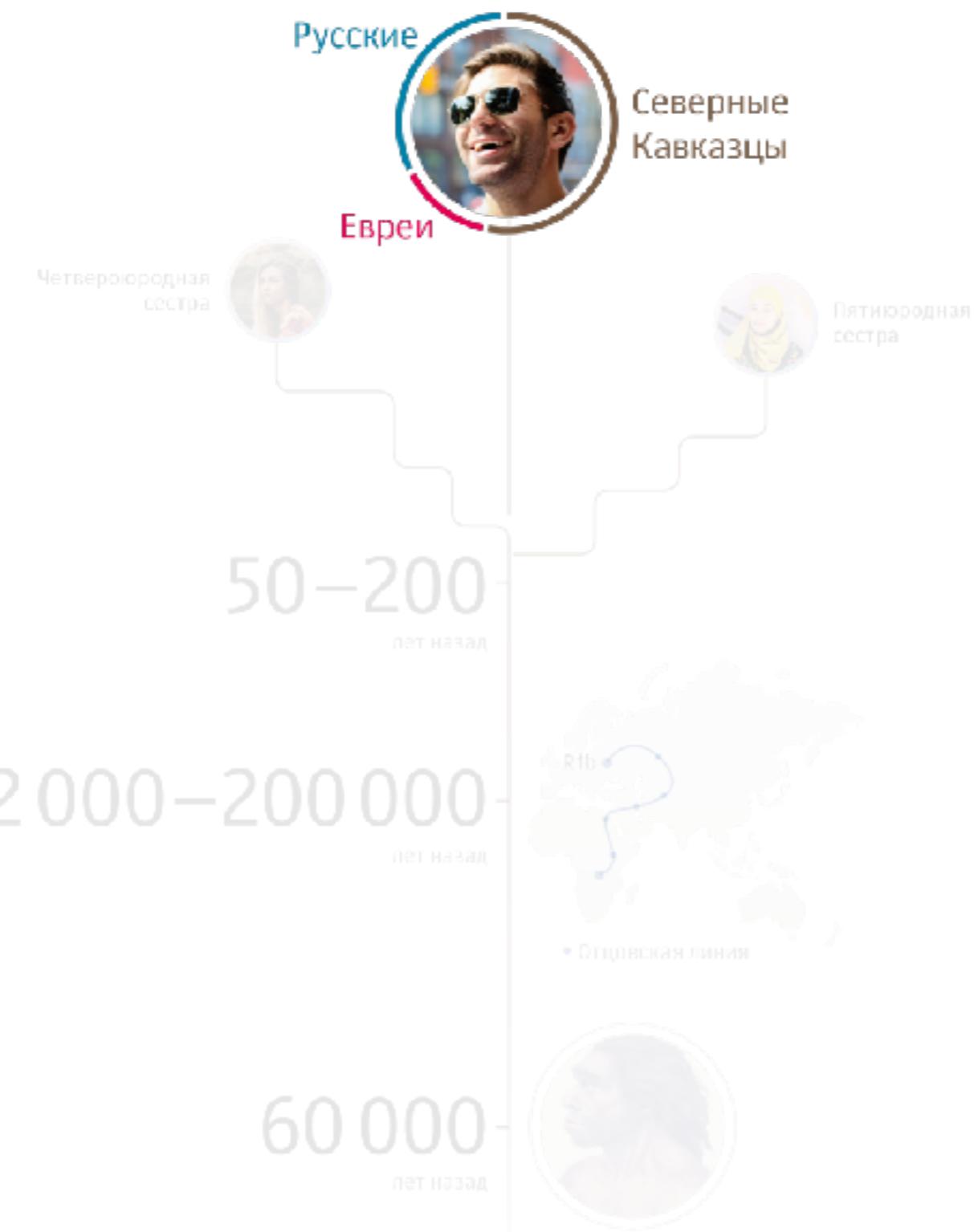
... AATGCAGTGC...
... AAT**A**CAGTGC...

rs3131972	1	752721	AG
rs114525117	1	759036	GG
rs79373928	1	801536	TT
rs116452738	1	834830	GG
rs72631887	1	835092	TT
rs4970383	1	838555	AA
rs28678693	1	838665	TT
rs4970382	1	840753	CC
rs4475691	1	846808	CT
rs72631889	1	851390	GG
rs7537756	1	854250	AG
rs13302982	1	861808	GG
rs376747791	1	863130	AA
rs2880024	1	866893	CC
rs13302914	1	868404	TT
rs76723341	1	872952	CC
rs148327885	1	878331	CC
rs143853699	1	879911	GG
rs2272757	1	881627	AA
rs67274836	1	884767	GG
rs3748597	1	888659	CC
rs3828049	1	889238	GG
rs77608078	1	891277	CC
rs13303010	1	894573	AA
rs13303229	1	897564	CC
rs3935066	1	900730	AA
rs6669800	1	903321	AA
rs35241590	1	904752	TT
rs28561399	1	910473	GG

Происхождение



Этнический состав

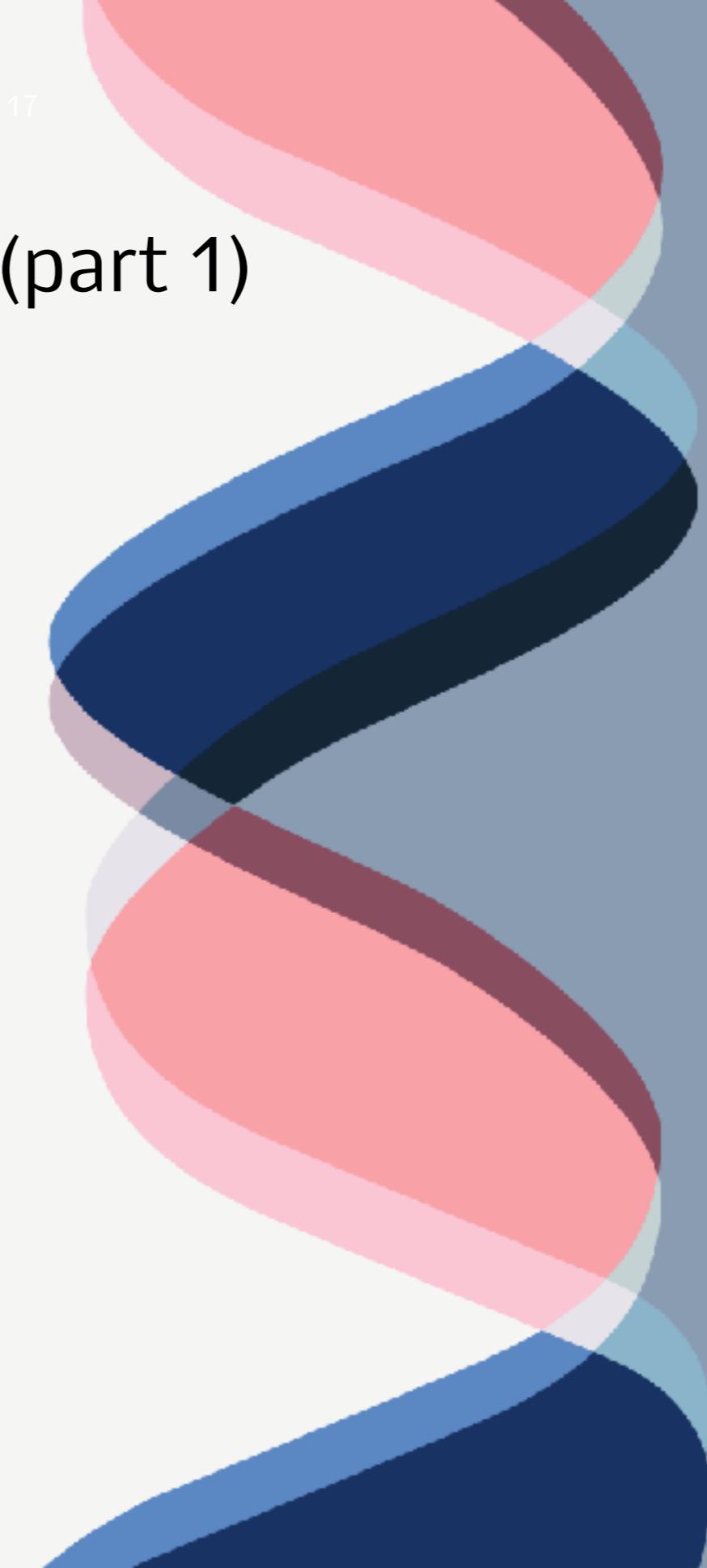




Ancestry decomposition (part 1)

Create database

1. Find datasets





Estonian Biocentre

Free data

For academic research, please find here human DNA sequence and genotype data from recent papers.

Complete high coverage human genomes and Y chromosome sequence data has been generated by Complete Genomics

SNP data has been generated with different Illumina genotyping arrays since 2008. Data has been lifted to Build37 and rsNumbers come from hg19 b131 (see the log below). The data is in PLINK binary PED (bed, bim, fam) format. Please cite the papers and contact mailt@ebc.ee for any questions.

IMPORTANT NOTE

Current Biology
Report

Shifts in the Genetic Landscape of the Western Eurasian Steppe Associated with the Beginning and End of the Scythian Dominance

Mart Arde^{1,2,3}, Ildi Saag¹, Kristjan Kyn¹, Jüri Kola¹, Atri G. Pathak⁴, Francesco Amato⁵, Koenraad Peeters⁶, Marta Zvelebil⁷, Lauri Kaas⁸, Kristiina Tambur⁹, Jüri Kola¹⁰, Katalin Kere¹¹, Ann Ebbé¹², Ute Winkel¹³, Birte Brinkmann¹⁴, Daniel Gruhier¹⁵, Paul Chikofsky¹⁶, Kyrre Gallo¹⁷, Oktayda Emeyen¹⁸, Anatoliy Abov¹⁹, Human Pollio²⁰, Ermalifajlija²¹, Enesay Kimhi²², Andriana Talić²³, Arman Salamyan²⁴, Märt Küllust²⁵

Current Biology
Report

The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East

Léa Lévy^{1,2,3}, Marcell Llamas¹, Uri Vayn¹, Martin Malas¹, Heidi Voll¹, María A. Rojo⁴, Iván G. Shirokova⁵, Valeri I. Khartanovich⁵, Darya R. Miltadjanova⁵, Klimov Klyazminovich⁵, Oksana L. Sloboda⁵, Amy Solid¹, Tuuli Roosberg¹, Ulri Park^{1,2}, Leor Serag, Erez Melamed¹, Sri Ravid¹, Francesco Marzilli¹, Maiko Ikeda¹, Reeva Meij¹, Eugenia U. Koncan¹, Emiko Hyonosuke Ono¹, David Reich^{1,2,3}, Ryan G. Thomas^{1,2,3}, Alena Novak¹, Leonidas Vlachos^{1,2,3}, Michaela Röhl¹, Max Mengel¹, and Antonio Salas¹

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

HOME | A
Search

David Reich Lab

Harvard Medical School

Downloadable genotypes of present-day and ancient DNA data (compiled from published papers)

On this page you can download a merged dataset consisting of genotypes for thousands of ancient and present-day individuals at up to 1.23 million positions in the genome (in hg19 coordinates).

All data released here:

- (a) have already been published (some by our group and some by other groups - see full list of references below),
- (b) have permissions appropriate for fully public data release,
- (c) are typed at a set of 1,233,013 sites in the genome (or 597,573 sites for present-day individuals genotyped on the Affymetrix Human Origins array). Typing is typically pseudo-haploid for ancient samples, when coverage is too low for full genotyping.

There are two datasets:

"1240K" : Ancient and present-day individuals (from either shotgun sequencing or ancient DNA)

"1240K+HO": Data from the above set merged with present-day individuals typed on the Human Origins array

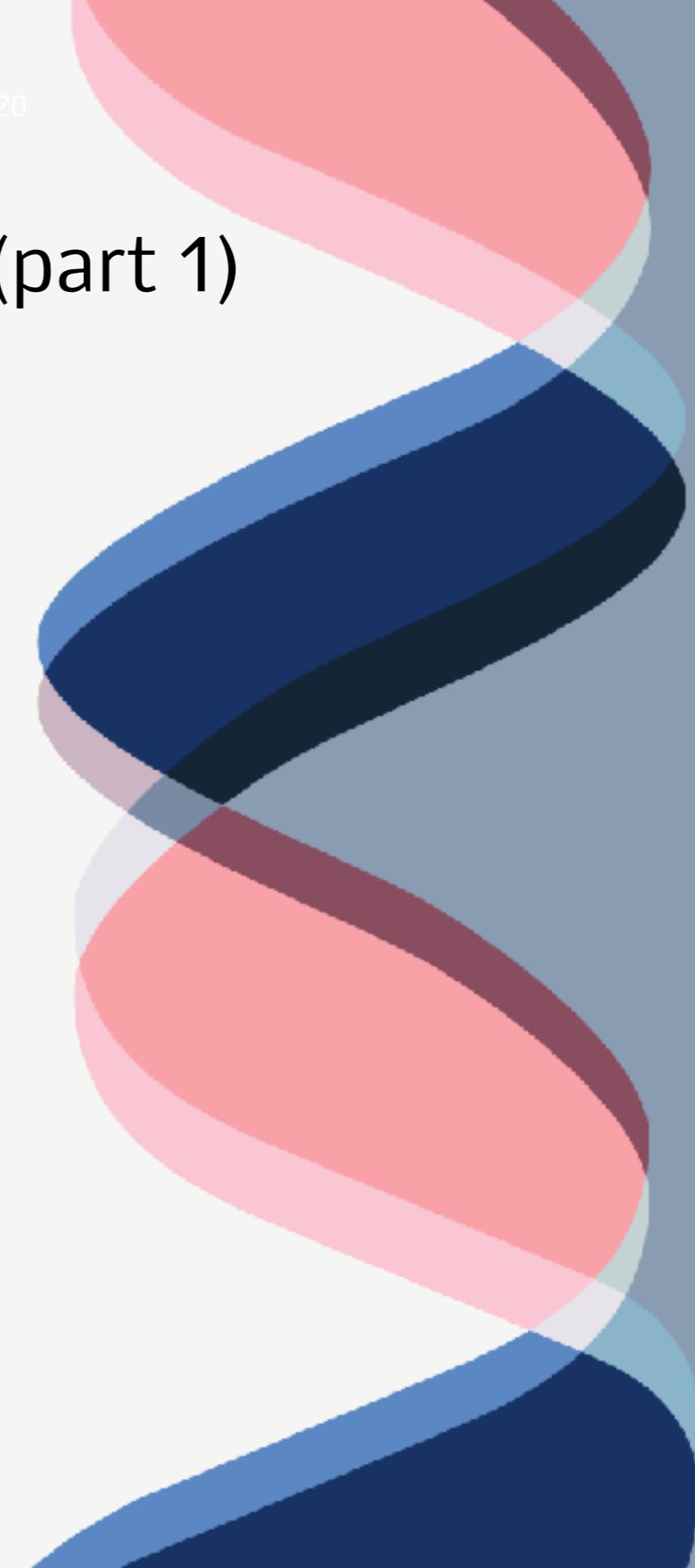
Each dataset consists of four files, in [eigenstrat](#) format. For details, please see:
[eigensoft](#):

- .anno: Rich meta-information for each individual.
- .ind: Three columns: Individual ID, sex determination, and group label.
- .snp: Information on each analyzed SNP position (SNP id, physical/genetic

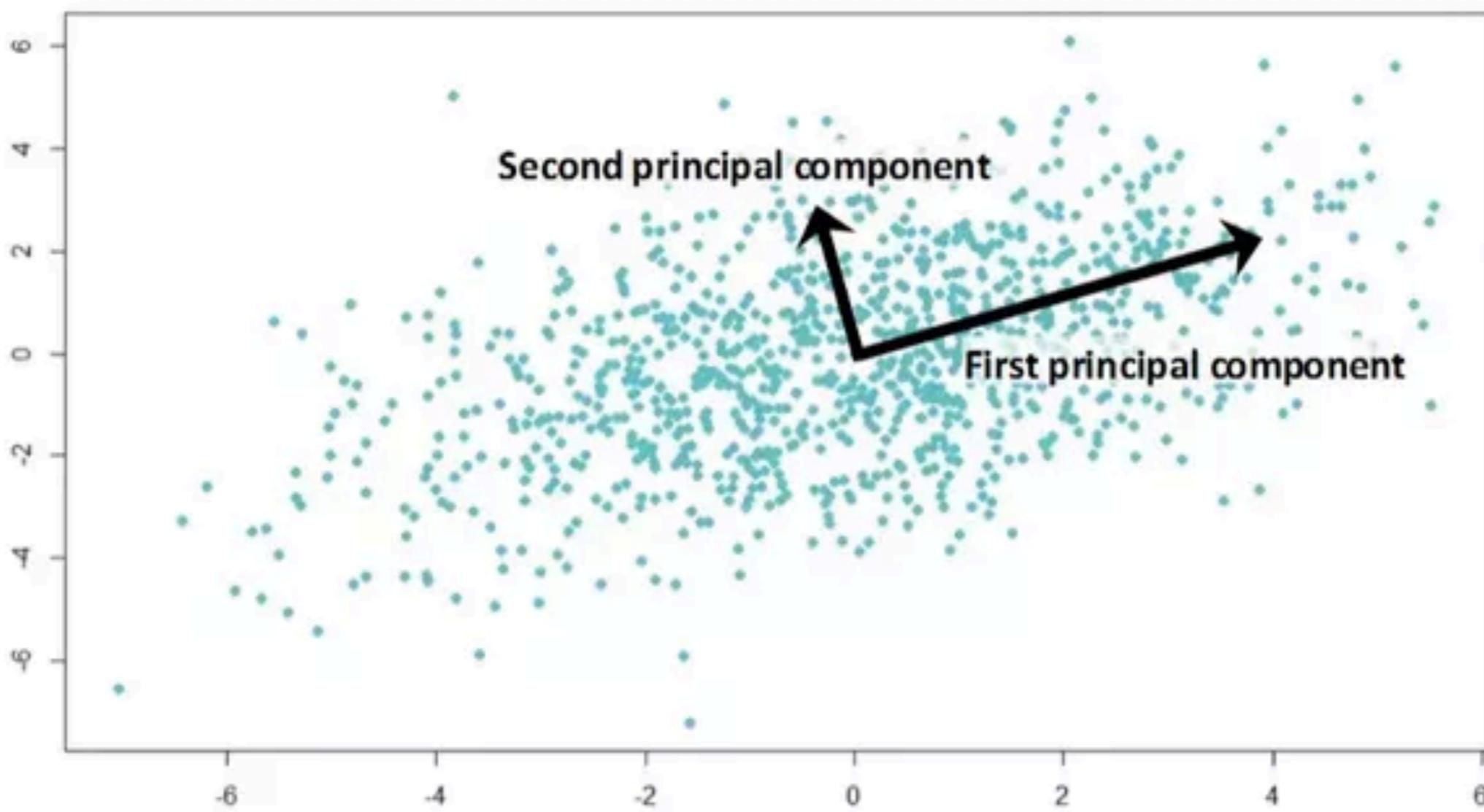
Ancestry decomposition (part 1)

Create database

1. Find datasets
2. Impute and phase data
3. Visualize data (PCA, tSNE, UMAP)



Метод главных компонент (PCA)



*.ped

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

*.bed

Contains binary version of the SNP info of the *.ped file.
(not in a format readable for humans)

*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

Covariate file

FID	IID	C1	C2	C3
1	1	0.00812835	0.00606235	-0.000871105
2	2	-0.0600943	0.0318994	-0.0827743
3	3	-0.0431903	0.00133068	-0.000276131

Legend			
FID	Family ID	rs{x}	Alleles per subject per SNP
IID	Individual ID	Chr	Chromosome
PID	Paternal ID	SNP	SNP name
MID	Maternal ID	GD	Genetic distance (morgans)
Sex	Sex of subject	BPP	Base-pair position (bp units)
P	Phenotype	C{x}	Covariates (e.g., Multidimensional Scaling (MDS) components)

Whole genome association analysis toolset

[Introduction](#) | [Basics](#) | [Download](#) | [Reference](#) | [Formats](#) | [Data management](#) | [Summary stats](#) | [Filters](#) | [Stratification](#) | [ES/AED](#) | [Association](#) | [Family-based](#) | [Permutation](#) | [LD calculations](#) | [Haplotypes](#) | [Conditional tests](#) | [Proxy association](#) | [Imputation](#) | [Dosage data](#)

[Meta-analysis](#) | [Result annotation](#) | [Clumping](#) | [Gene Report](#) | [Epistasis](#) | [Rare CNVs](#) | [Common CNVs](#) | [R-plug-ins](#) | [SNP annotation](#) | [Simulator](#) | [Profiles](#) | [ID helper](#) | [Resources](#) | [Flow chart](#) | [MISC.](#) | [FAQ](#) | [gPLINK](#)

1. introduction

2. Basic Information

- [Using PLINK](#)
- [Reporting problems](#)
- [What's new?](#)
- [PDF documentation](#)

3. Download and general notes

- [Stable download](#)
- [Development code](#)
- [General notes](#)
- [MS-DOS notes](#)
- [Unix/Linux notes](#)
- [Compilation](#)
- [Using the command line](#)
- [Viewing output files](#)
- [Version history](#)

4. Command reference table

- [List of options](#)
- [List of output files](#)
- [Under development](#)

5. Basic usage/data formats

- [Running PLINK](#)
- [PED files](#)
- [MAP files](#)
- [Transposed filesets](#)
- [Long-format filesets](#)
- [Binary PED files](#)
- [Alternate phenotypes](#)
- [Covariate files](#)
- [Cluster files](#)
- [Set files](#)

A PLINK tutorial

In this tutorial, we will consider using PLINK to analyse example data: randomly selected genotypes (approximately 80,000 autosomal SNPs) from the 89 Asian HapMap individuals. A phenotype has been simulated based on the genotype at one SNP. In this tutorial, we will walk through using PLINK to work with the data, using a range of features: data management, summary statistics, population stratification and basic association analysis.

NOTE These data do not, of course, represent a realistic study design or a realistic disease model. The point of this exercise is simply to get used to running PLINK.

89 HapMap samples and 80K random SNPs

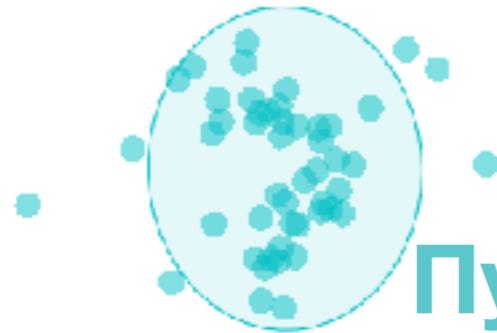
The first step is to obtain a working copy of PLINK and of the example data files.

1. Make sure you have PLINK installed on your machine (see [these instructions](#)).
2. Download the example data archive file which contains the genotypes, map files and two extra phenotype files, described below (zipped, approximately 2.3M)
3. Create a new folder/directory on your machine, and unzip the file you downloaded (called `hapmap1.zip`) into this folder.

HINT! If you are a Windows user who is unsure how to do this, follow this link

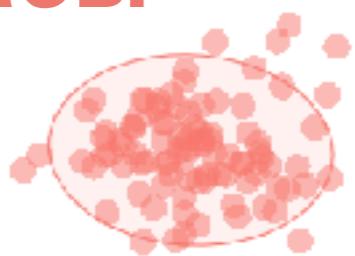
Two phenotypes were generated: a quantitative trait and a disease trait (affection status, coded 1=unaffected, 2=affected), based on a median split of the quantitative trait. The quantitative trait was generated as a function of three simple components:

- A random component
- Chinese versus Japanese identity

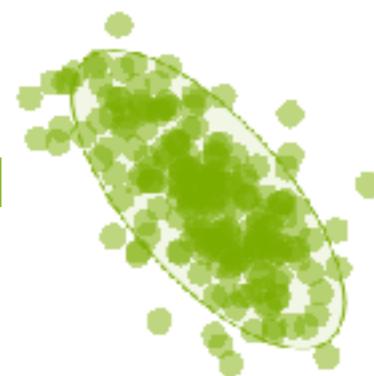


Пуштуны

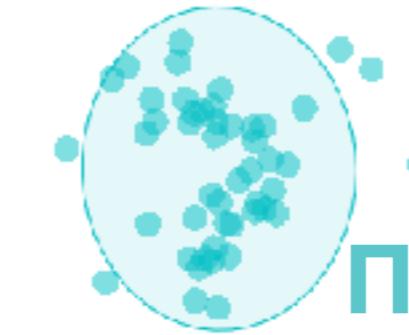
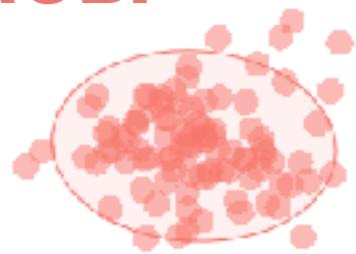
Арабы



Прибалты



Арабы

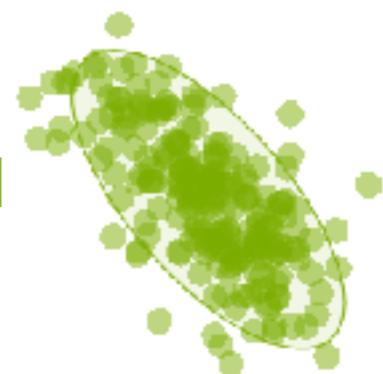


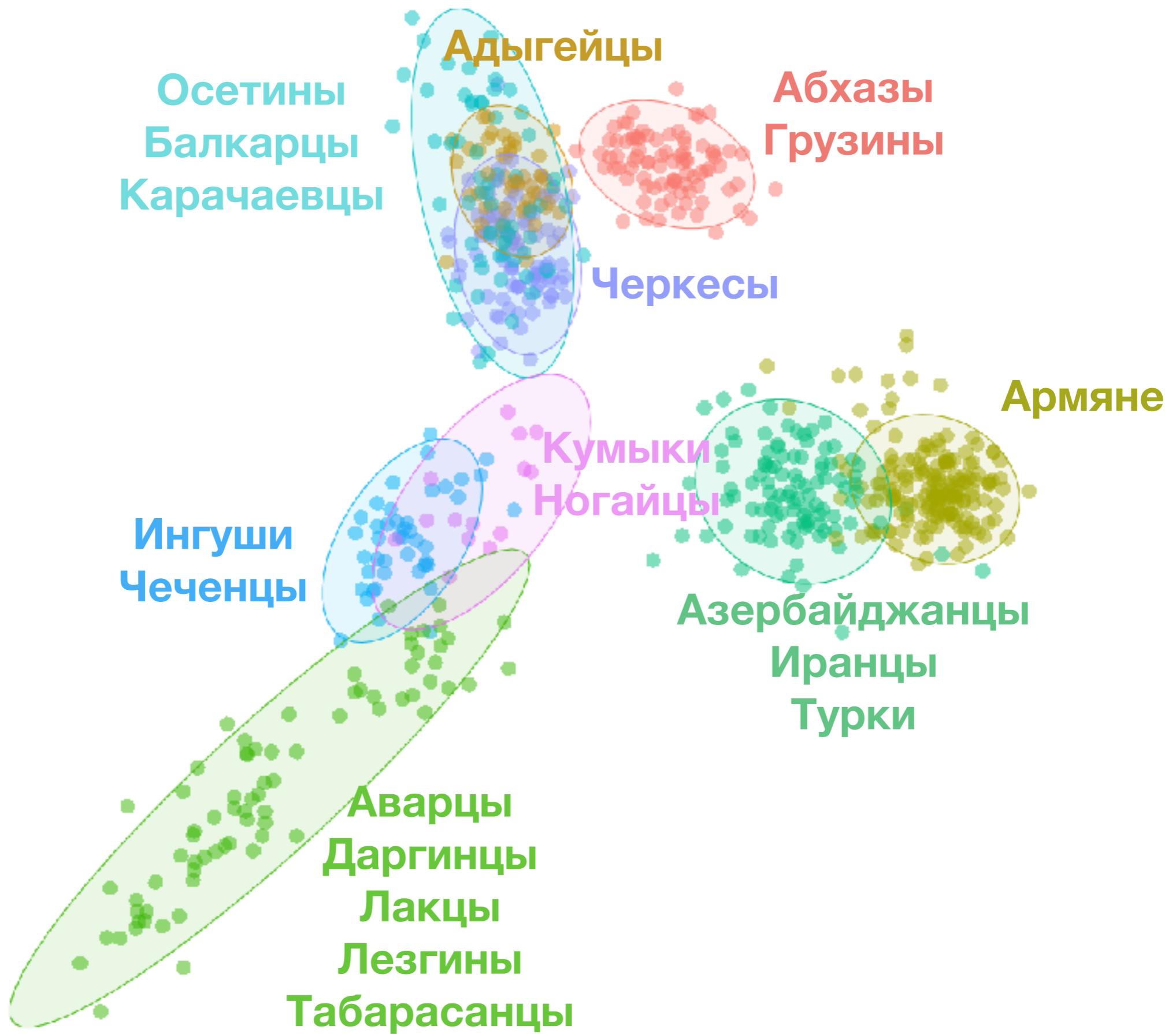
Пуштуны

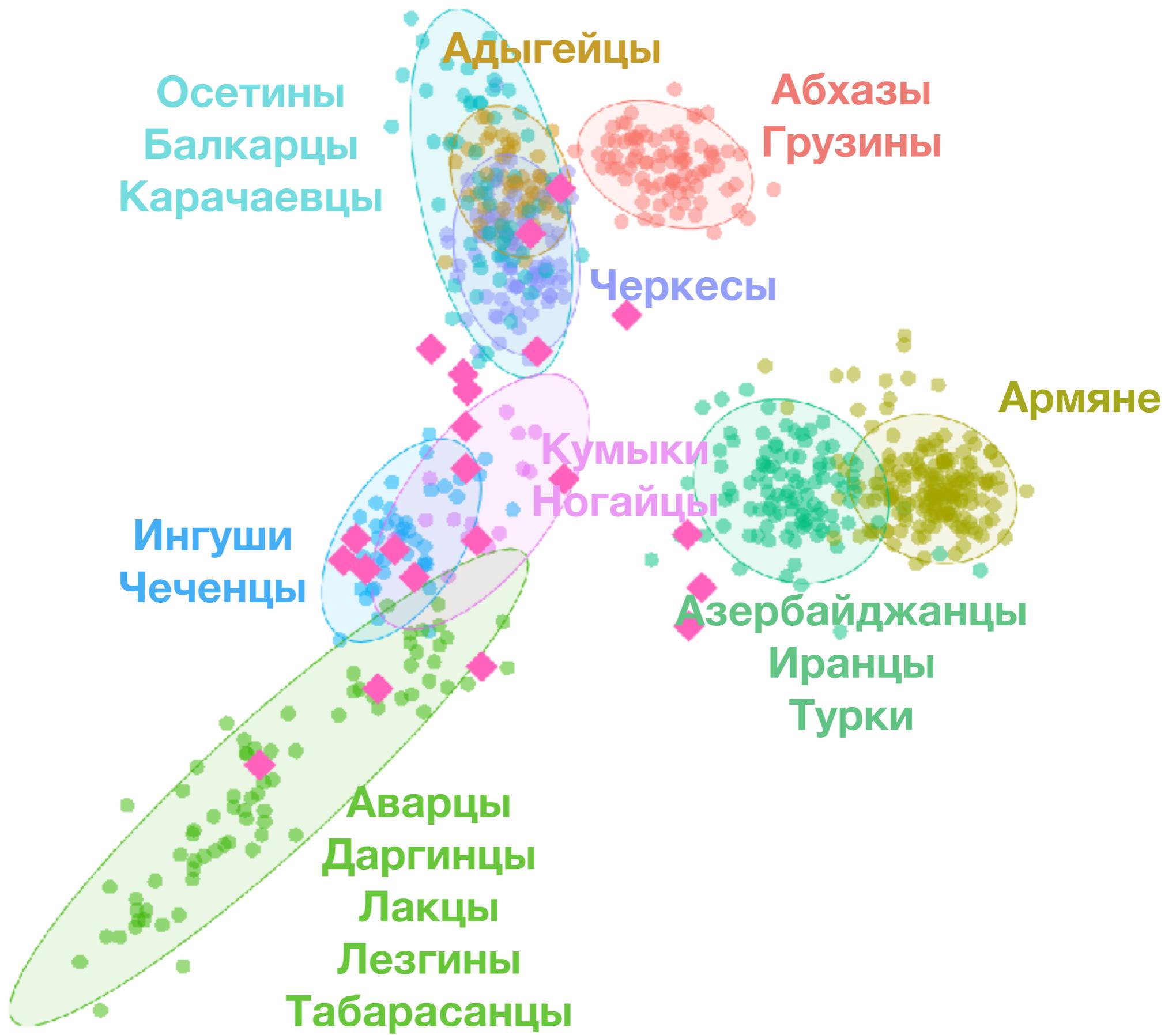
Участники



Прибалты







Курашева Екатерина

Доциев Арслан

Захарова Алёна

Юнаев Моисей

Дидаров Сослан

Гуева Мадина

Давыдов Давид

Алборова Залина

Пареулидзе Ибрагим

Хамидов Муса

Гапаев Ала

Калиматов Адам

Боков Руслан

Албаков Магомед

Хутиев Адам

Амиров Мухтар

Асретов Джалил

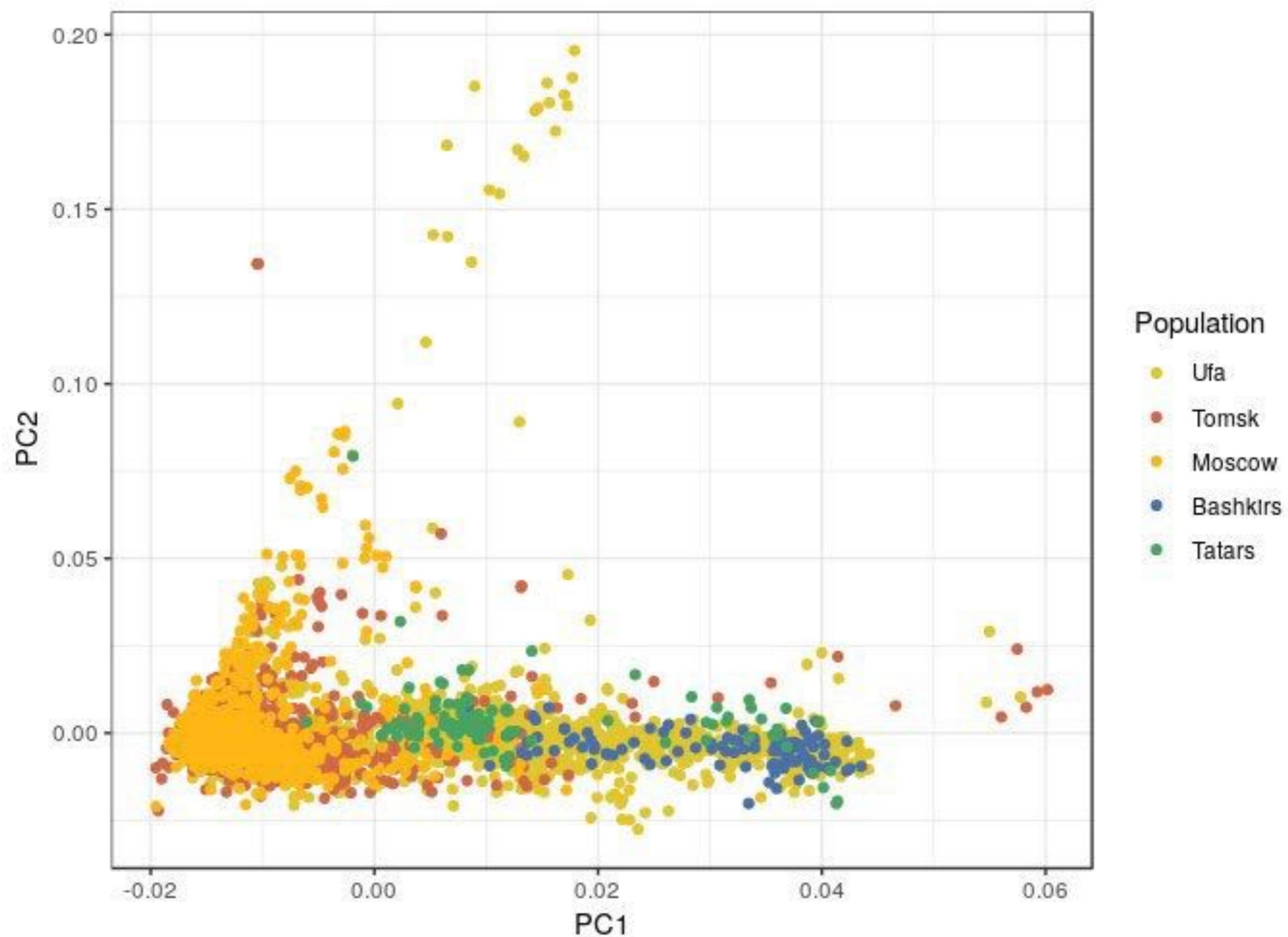
Шехсаидов Карим

Фарахманд Париса

Исрафилов Исрафил

Зинковская Дарья

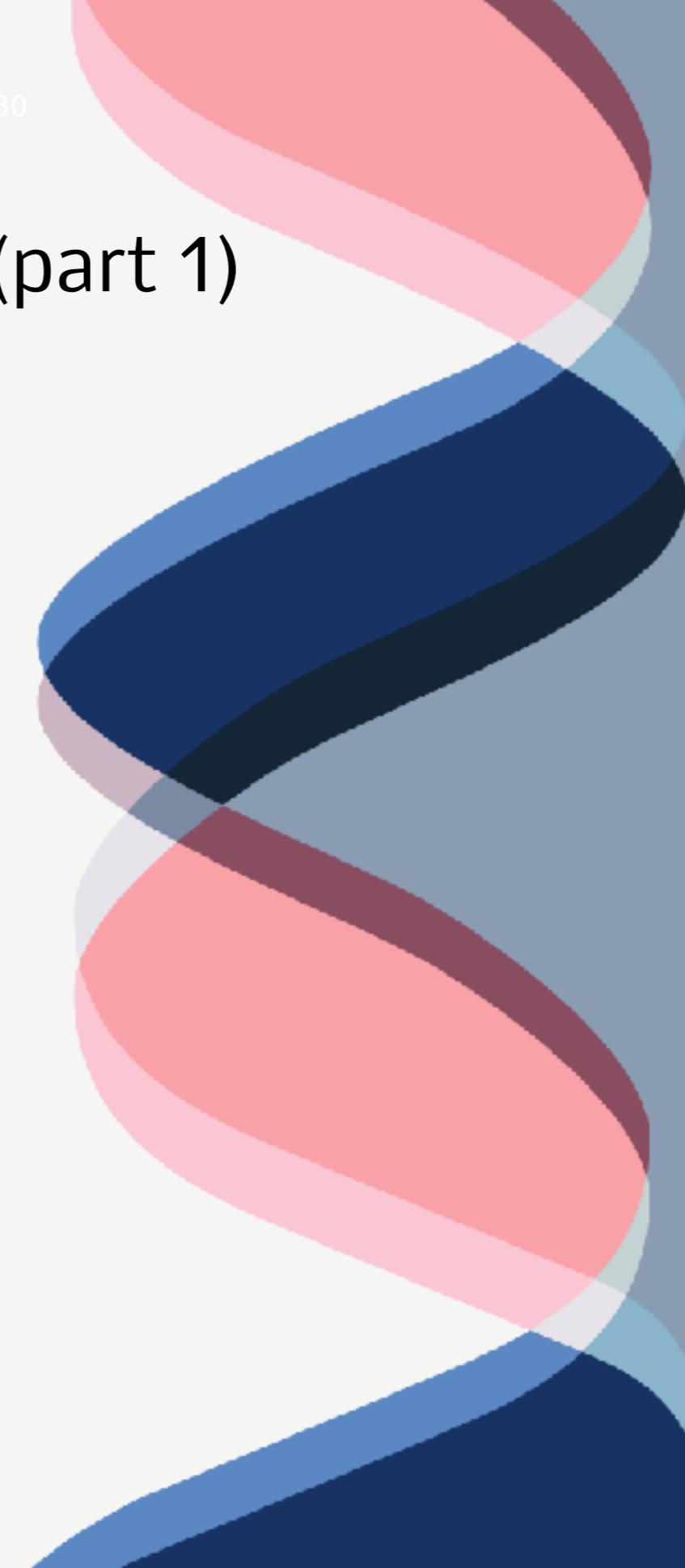
Орлова Нина

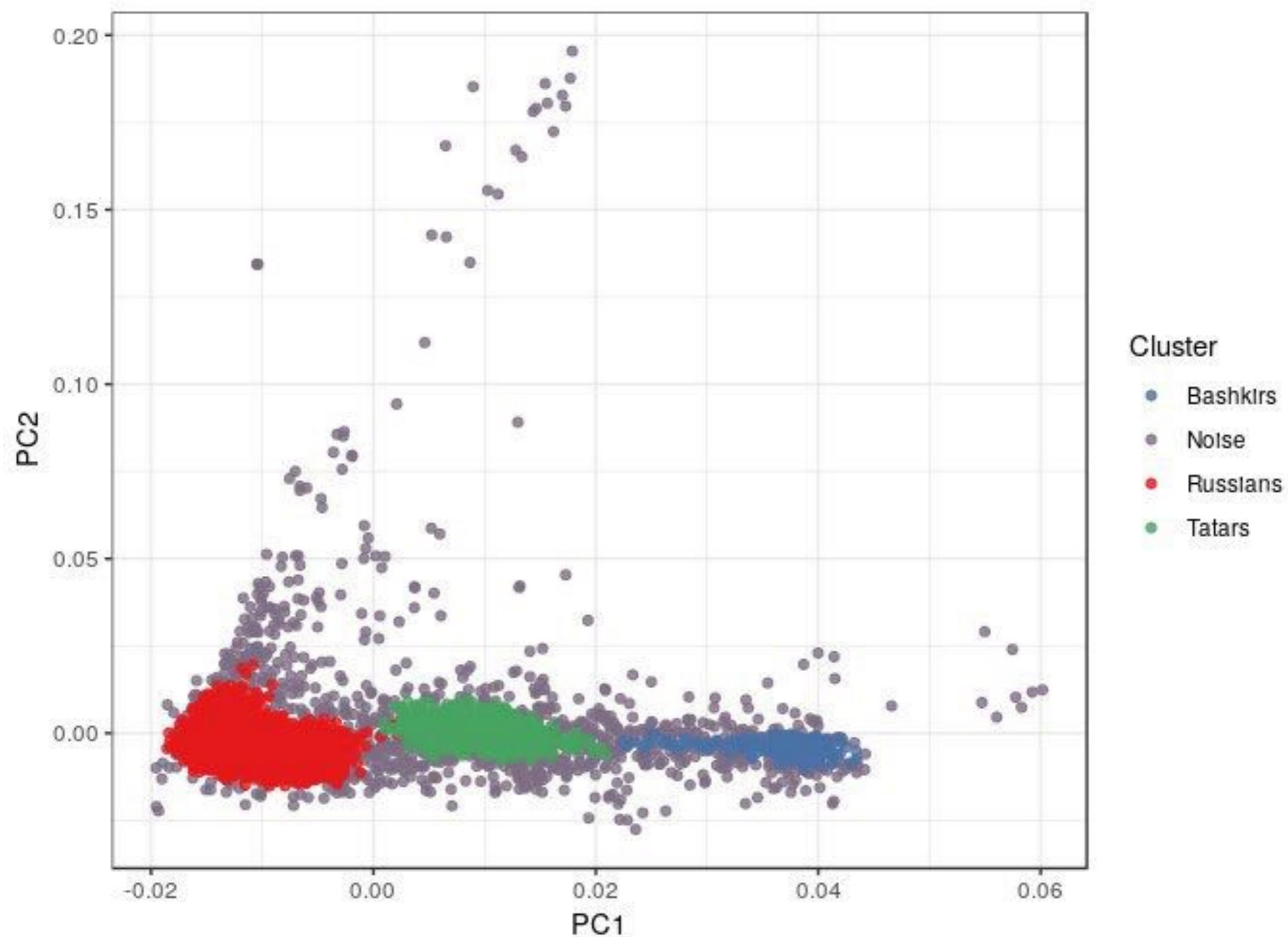


Ancestry decomposition (part 1)

Create database

1. Find datasets
2. Impute and phase data
3. Visualize data (PCA, tSNE, UMAP)
4. Filter out admixed individual, outliers, form clusters (HDBSCAN)

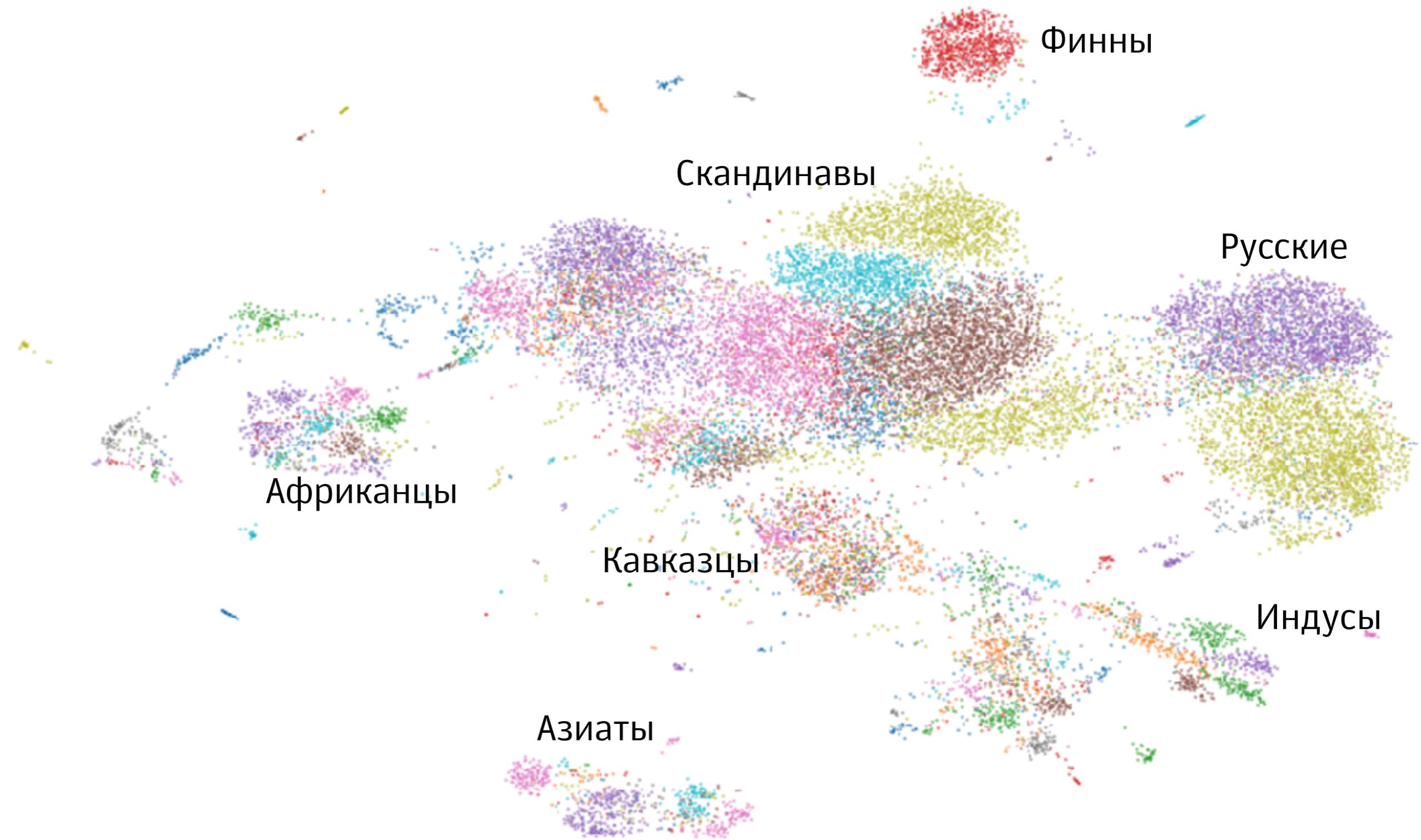




Генетическая карта народов России



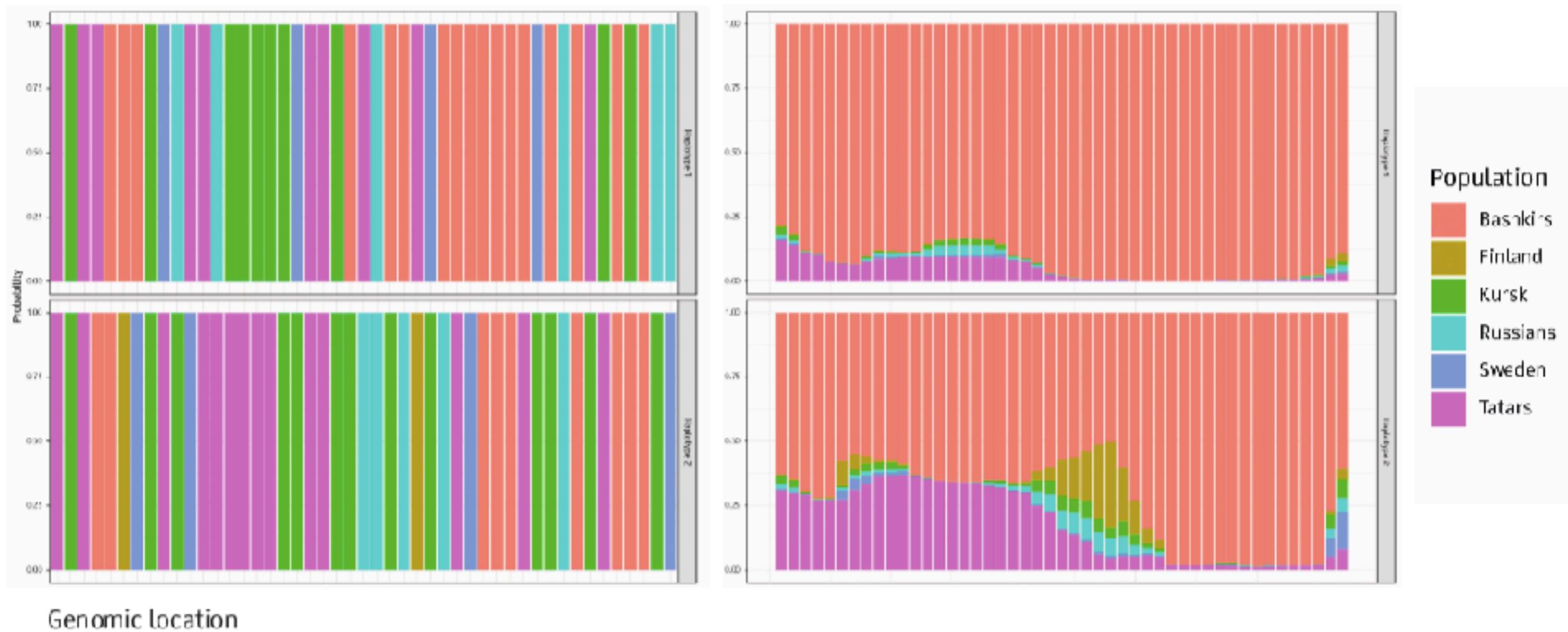
UMAP of 255 populations



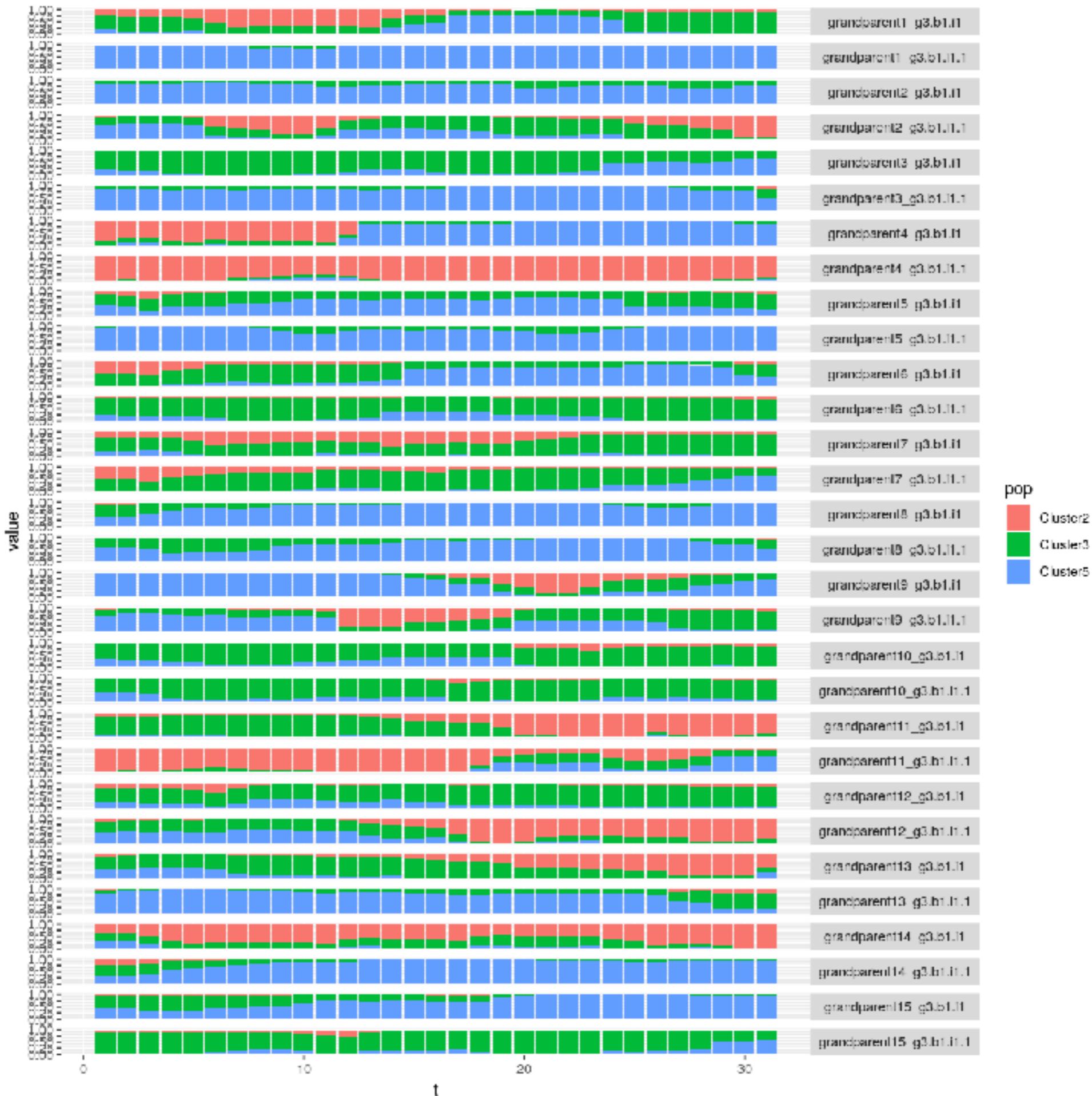
Ancestry decomposition (part 2) Local ancestry inference

1. Classify 100-500 SNP windows
(SVM)
2. Error correction (HMM)

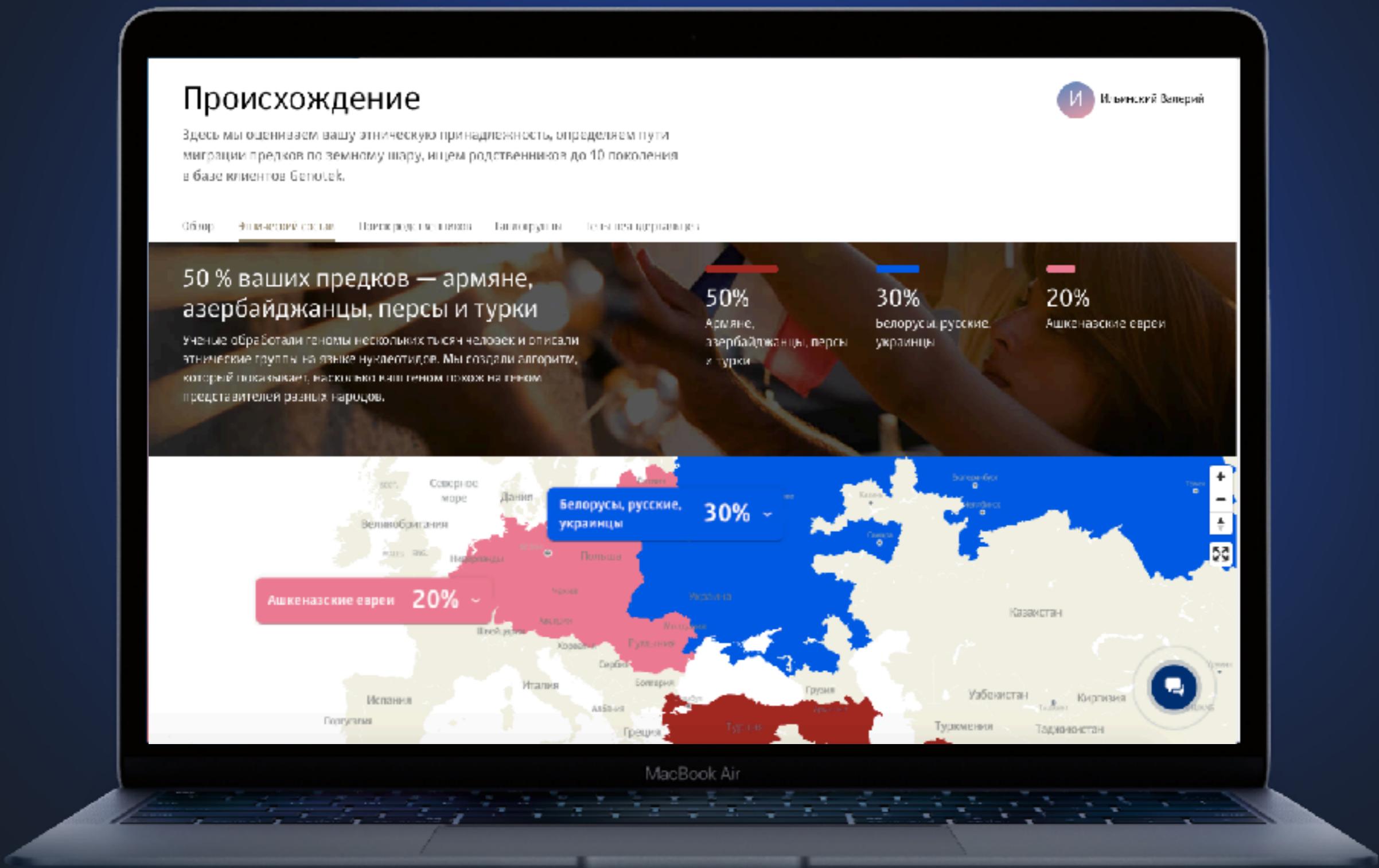
Ancestry proportions of the individual with Bashkir ethnic origin estimated with SVM (left) and HMM (right)



For Bashkirs the precision is over 95% and the recall is about 70%.



Результаты этнического состава



Population adjustment for Polygenic Risk Scores

Regarding stratification, most PRS methods do not explicitly address recent admixture, and none consider recently admixed individuals' unique local mosaics of ancestry; thus, further methodological development is needed.

Martin, Alicia R., et al. "Clinical use of current polygenic risk scores may exacerbate health disparities." *Nature genetics* 51.4 (2019): 584.

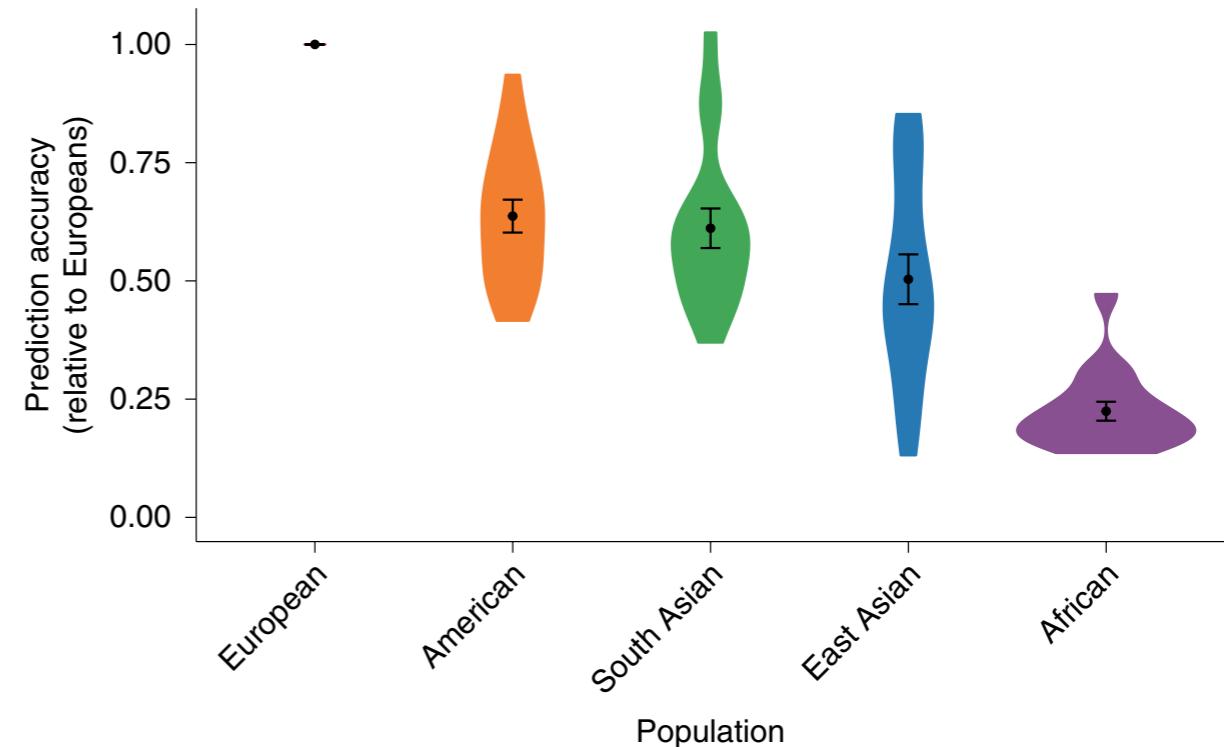


Fig. 3 | Prediction accuracy relative to European-ancestry individuals across 17 quantitative traits and 5 continental populations in the UKBB. All phenotypes shown here are quantitative anthropometric and blood-panel traits, as described in Supplementary Table 6, which includes discovery-cohort sample sizes. Prediction target individuals do not overlap with the discovery cohort and are unrelated; sample sizes are shown in Supplementary Table 7. Violin plots show distributions of relative prediction accuracies, points show mean values, and error bars show s.e.m. values. Prediction R^2 for each trait and population are shown in Supplementary Fig. 12.

Витамины и минералы

ДЕФИЦИТ ВИТАМИНА D

В Российской Федерации до 84% людей страдают от недостаточности и дефицита витамина D. Только каждый третий ребенок до 3 лет жизни имеет достаточный уровень витамина D.

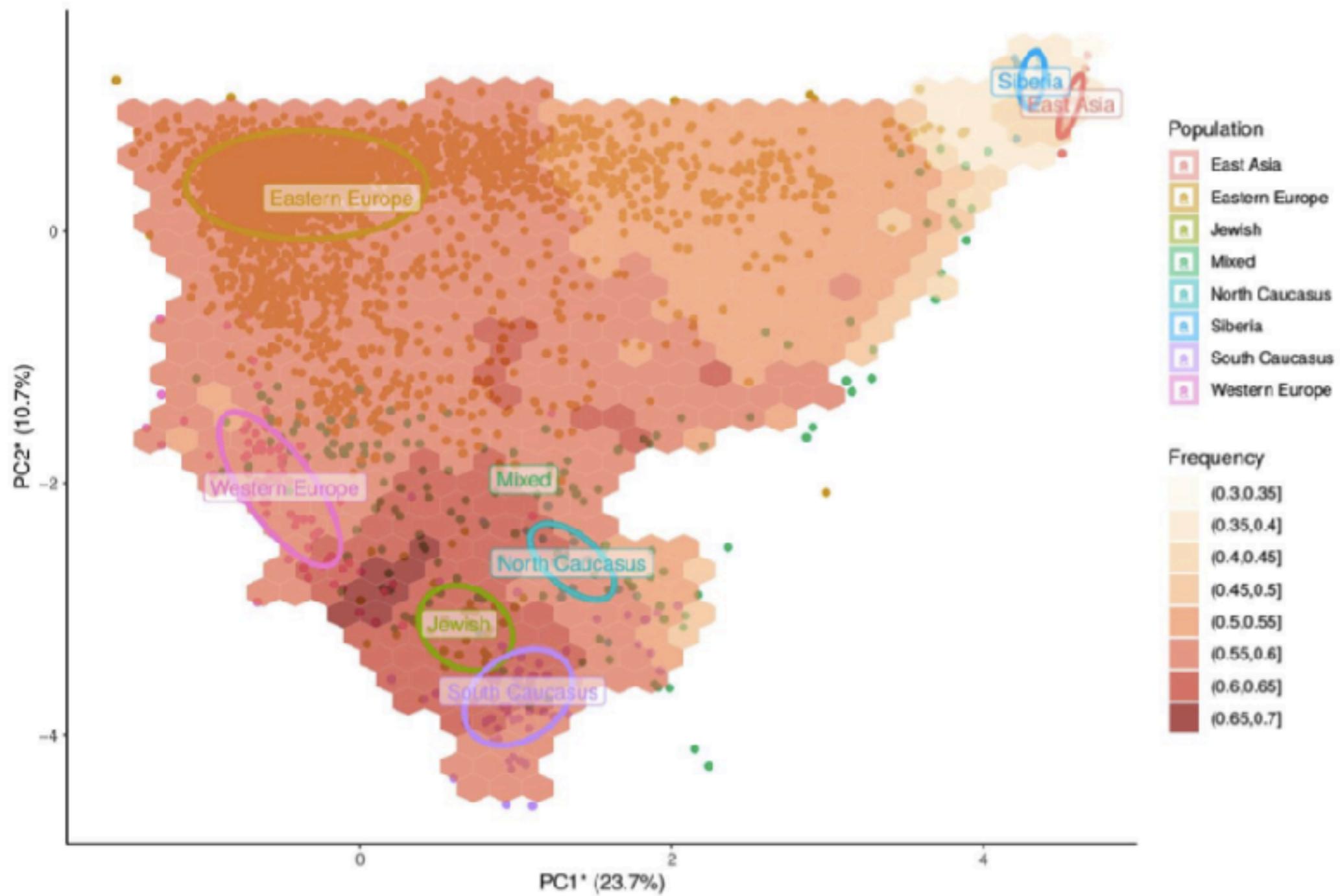
ДЕФИЦИТ ВИТАМИНА В12

Распространенность дефицита витамина B12 по разным оценкам достигает 1.5 – 15%, повышаясь с возрастом. Учитывая повышенный интерес к вегетарианскому типу питания последние годы, распространенность дефицита B12 может увеличиваться.

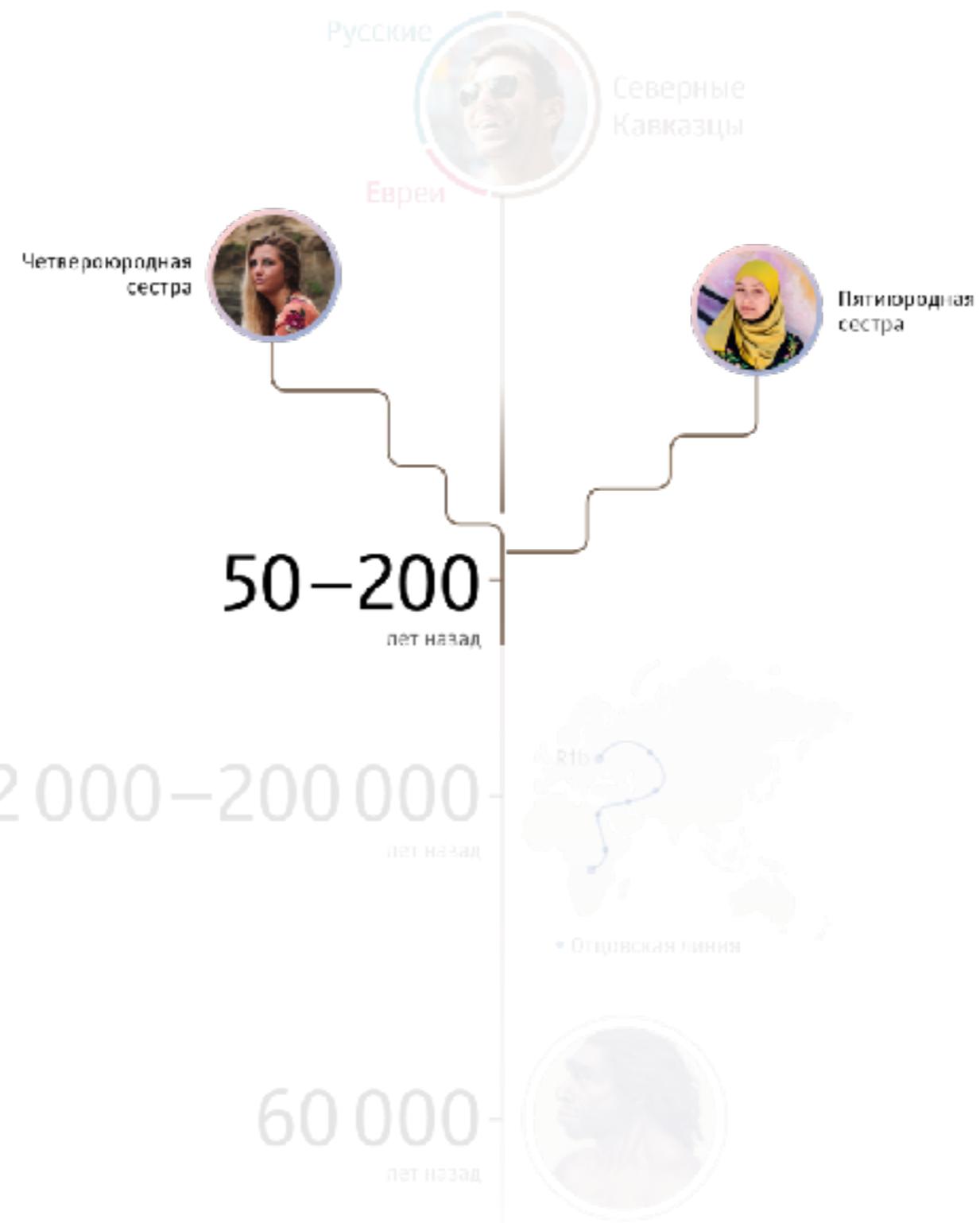
ДЕФИЦИТ ЖЕЛЕЗА

Около трети населения планеты страдает от железодефицита. Дефицит железа может приводить к тяжелым последствиям в виде анемии, которой страдают 25% людей во всем мире.

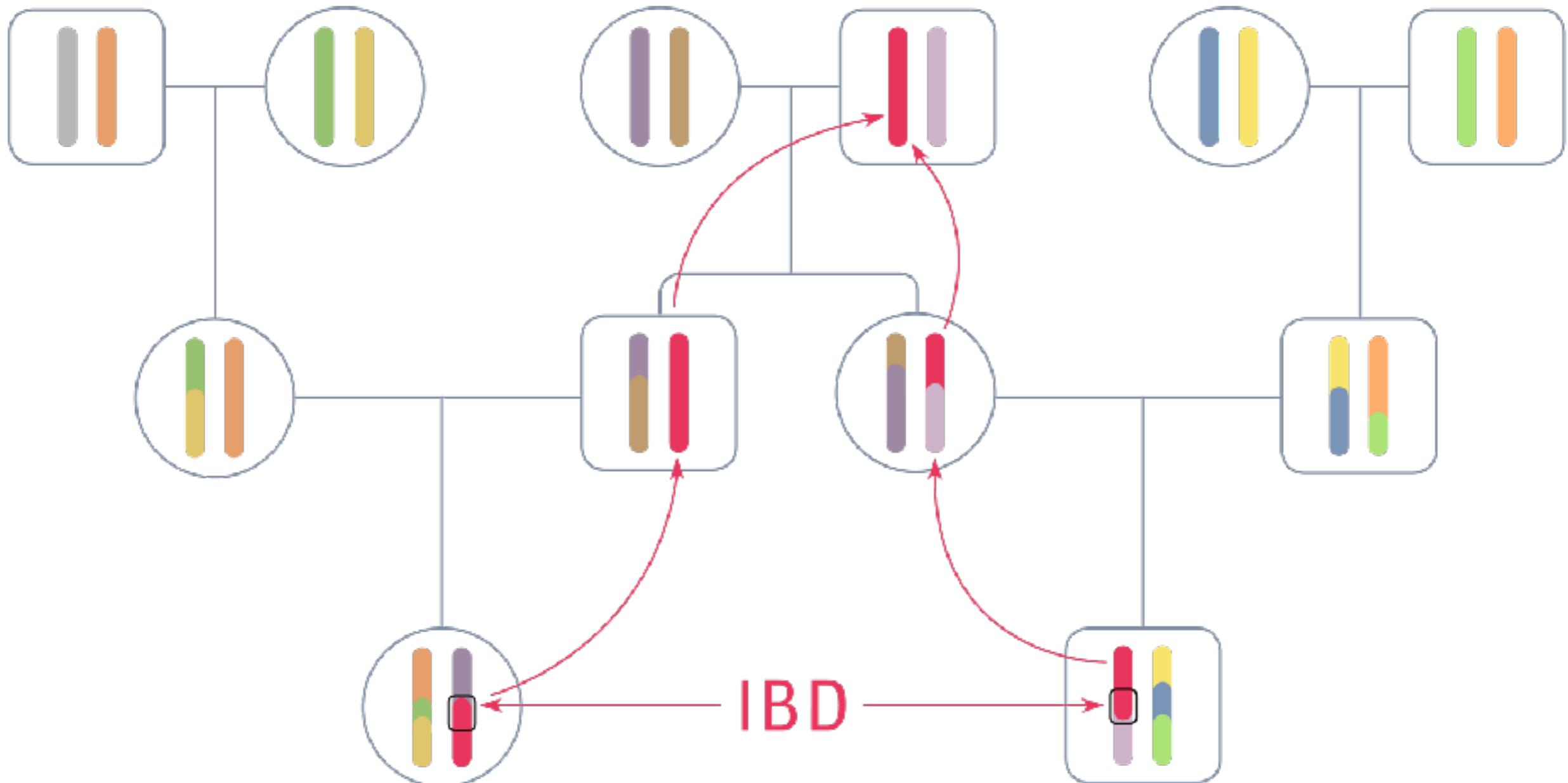
Распространенность гаплотипа GC1S

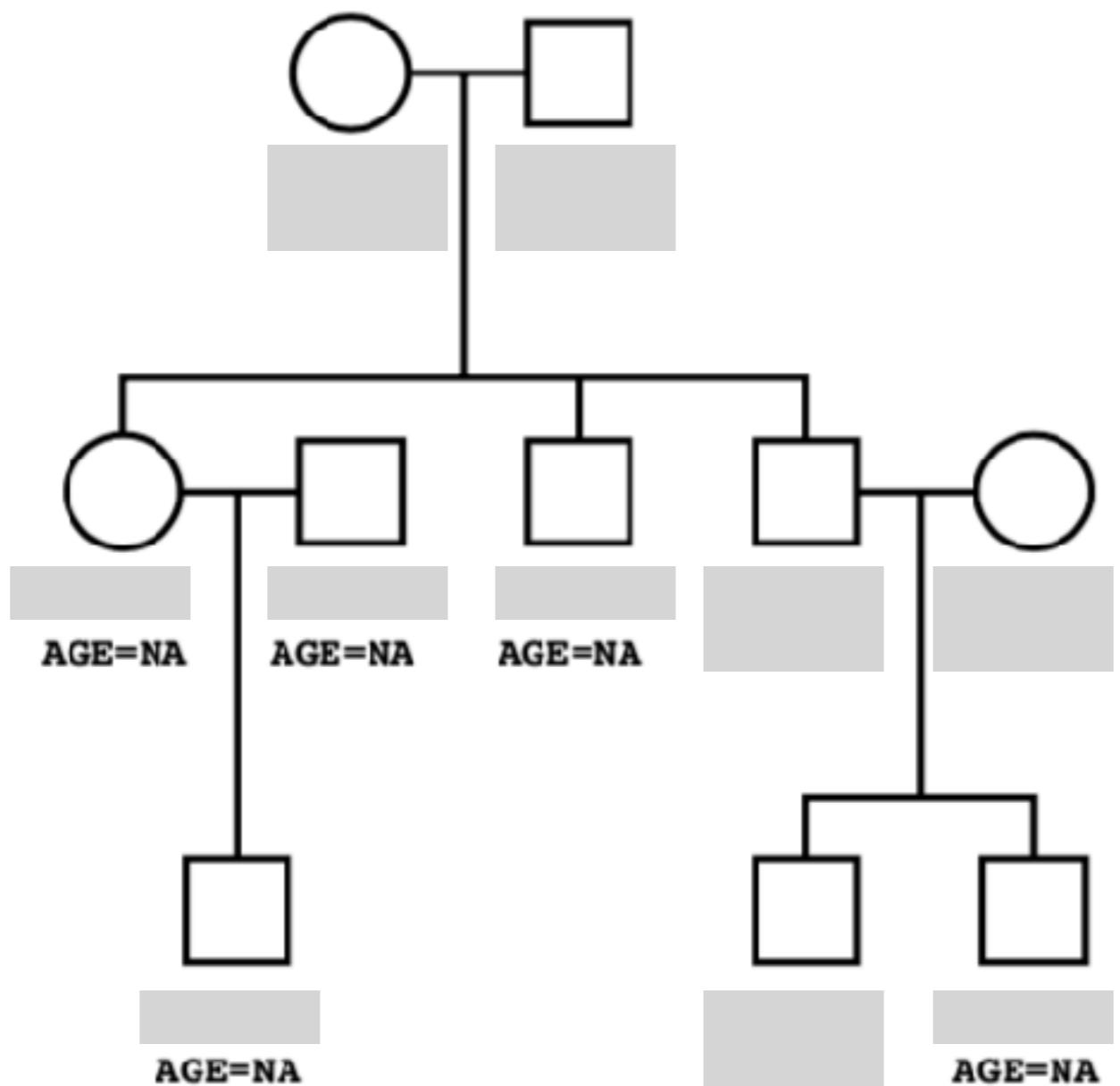


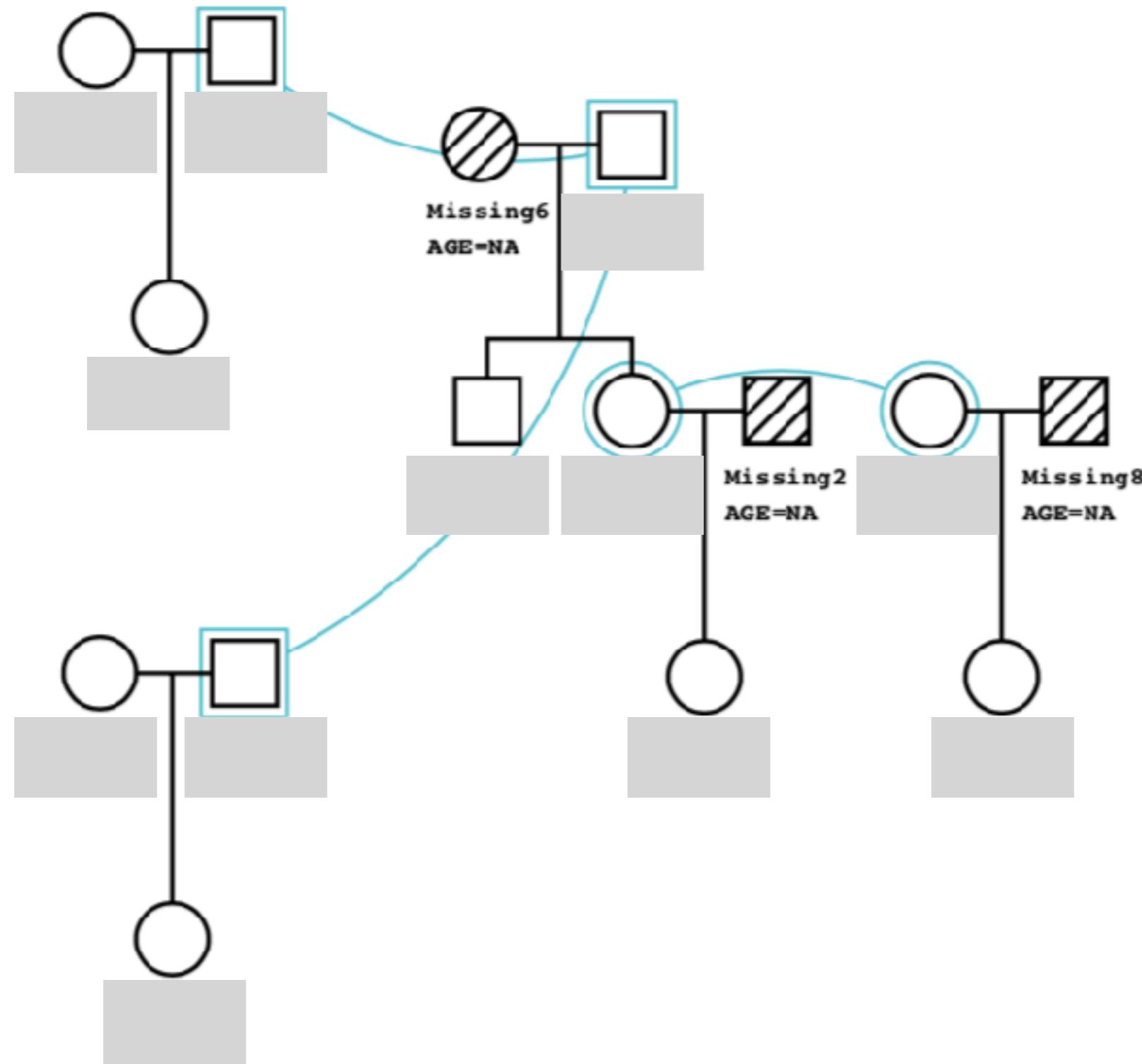
Поиск родственников



Сегменты от общего предка

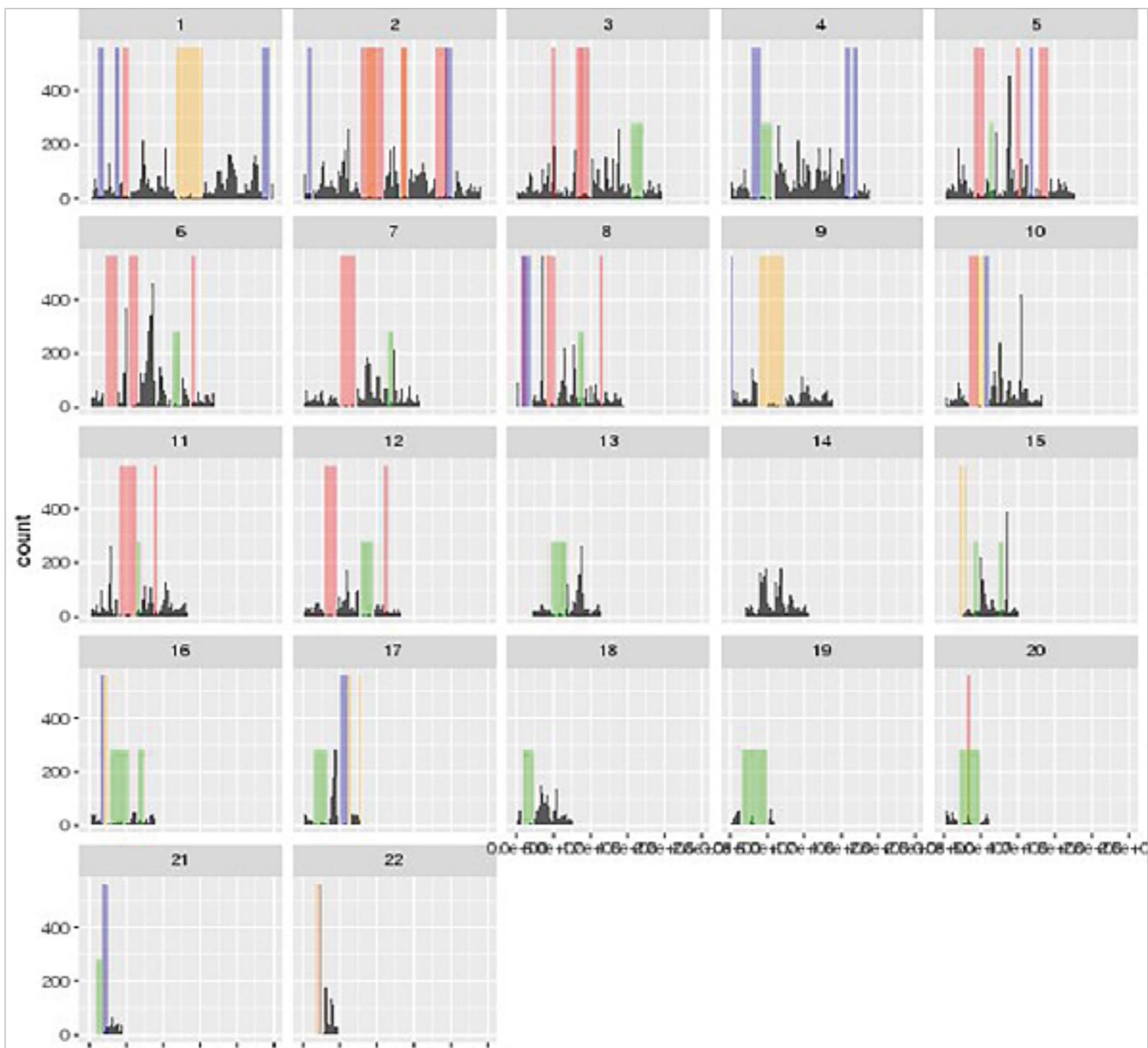






Common regions to be masked

After exteneded external masks + ersa



Результаты поиска родственников

The screenshot displays the Genotek software interface on a MacBook Air laptop screen. The main window is titled "Генеалогия" (Genealogy). On the left, a sidebar menu lists various features: Домашняя страница (Home page), Консультация с врачом (Consultation with a doctor), Проверка (Check), Проверка наследственных заболеваний (Hereditary disease check), Здоровье (Health), Риски заболеваний (Disease risks), Питание (Nutrition), Спорт (Sports), Способности и характер (Abilities and character), Планирование детей (Planning children), Риски зачатия (Fertilization risks), Риски беременности (Pregnancy risks), Генетический риск для последующих поколений (Genetic risk for future generations), and Неонатальный скрининг (Neonatal screening). A red banner at the bottom of the sidebar indicates "Акция: скидка 20% на 23%", and "Исходные данные ДНК-тестов" (Initial DNA test data) is listed. The main content area shows a search bar with "Название заболевания, генотипа" and a magnifying glass icon. Below it, a message states: "Здесь мы определяем вашу этническую принадлежность, определяем пути миграции предков по земному шару, ищем родственников до 10 поколений в базе клиентов БиоНек". Navigation tabs include "Обзор" (Overview), "Этнический состав" (Ethnic composition), "Поиск родственников" (Search for relatives) (which is selected), "Генетические группы" (Genetic groups), and "Наследование" (Inheritance). The main section displays "У вас 4 близких родственника" (You have 4 close relatives) and instructions to "Найдите родственников вплоть до 10 поколений. Общайтесь и делитесь контактами." Below this, three relatives are listed:

Фото	Имя	Возраст	Место жительства	Генетическая группа	Действие	
	Константин Константинович Константиновский	64 года	г. Константинополь	Родственник в 3-8 поколениях по материнской линии	Генетическая карта: 2atlh2at	Получить сообщение
	Ф. А.	64 года		Мальчик	Генетическая карта: 2atlh2at	Получить сообщение
	И.	64 года		Родственник в 3-8 поколениях по материнской линии	Генетическая карта: 2atlh2at	Получить сообщение

Distant relatives prediction

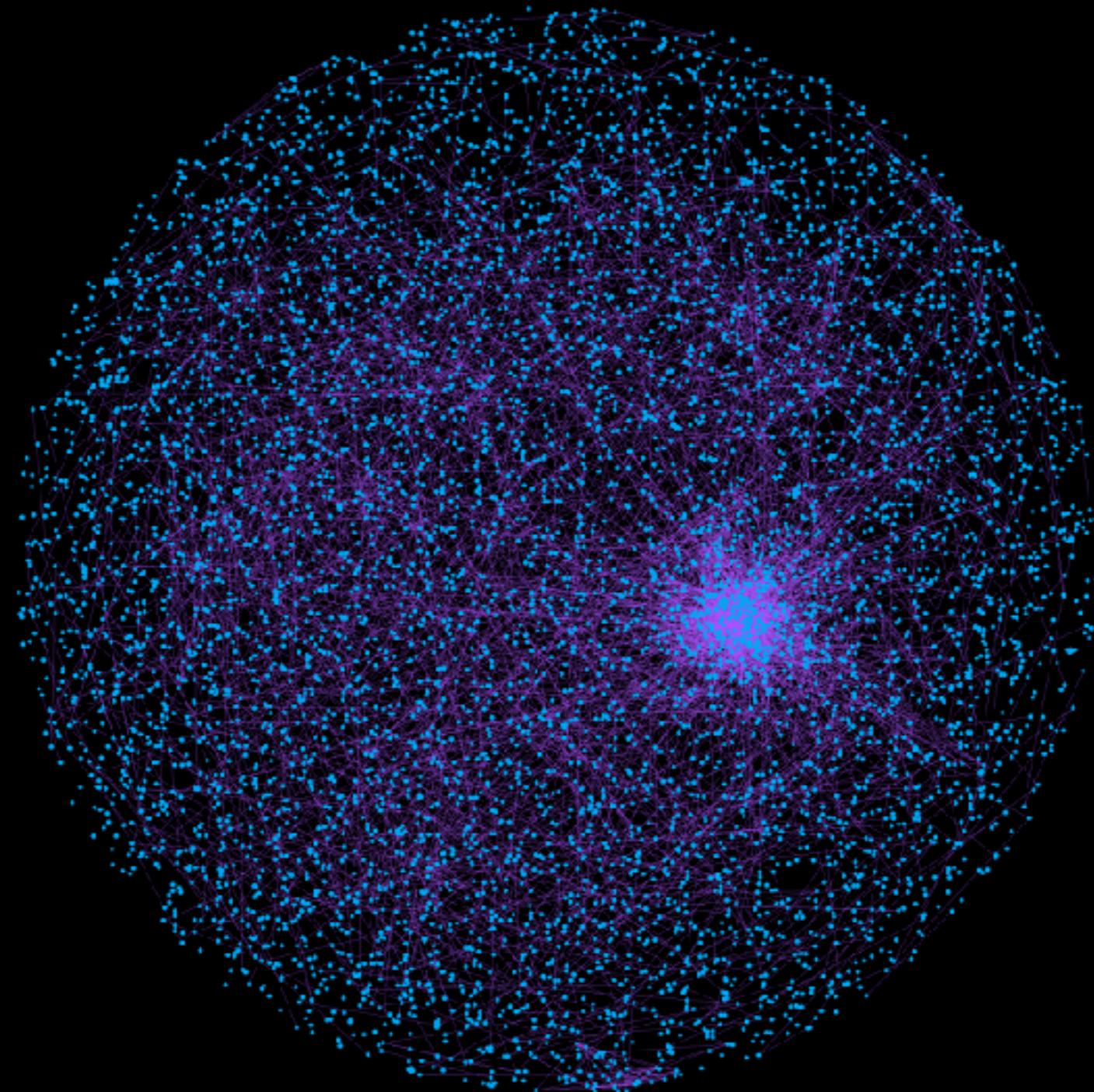
Degree	% of correct degree	% of correct degree ± 1
1	100 %	100 %
2	100 %	100 %
3	97 %	100 %
4	65 %	100 %
5	62 %	100 %
6	48 %	86 %
7	27 %	63 %

Distant relatives

1. Impute and phase data
2. Find IBD segments (GERMLINE)
3. Predict the degree of the relationship (ERSA)
4. Build the family tree (involving Age, mtDNA and X,Y chromosomes) (PRIMUS)
5. Updated distant degrees (PADRE)
6. Detect lineage



Сеть родственников



Ашkenазские евреи

Происхождение

Согласно гипотезе популяция ашkenазских евреев образовалась в результате смешения евреев-выходцев с Ближнего Востока и евреев, населявших Европу. В ходе истории их численность сокращалась с относительной периодичностью. Это явление называется «бутылочным горлышком» (рис. справа).

Около 700 лет назад резкое сокращение численности ашkenазских евреев привело к падению генетического разнообразия. Вероятно это связано с вспышкой чумы в Европе.

Современные ашkenазские евреи –
потомки лишь 350 людей, живших
600–800 лет назад



Наследственные заболевания

Наследственные заболевания

Прохождение популяции ашкеназских евреев через «бутылочное горлышко» несколько сотен лет назад, а также вступление в брак преимущественно с представителями своего народа привело к снижению генетического разнообразия ашкеназских евреев. Многие генетические маркеры присутствуют у них значительно чаще, чем у других народов, и приводят к наследственным заболеваниям.

ген
HEXA

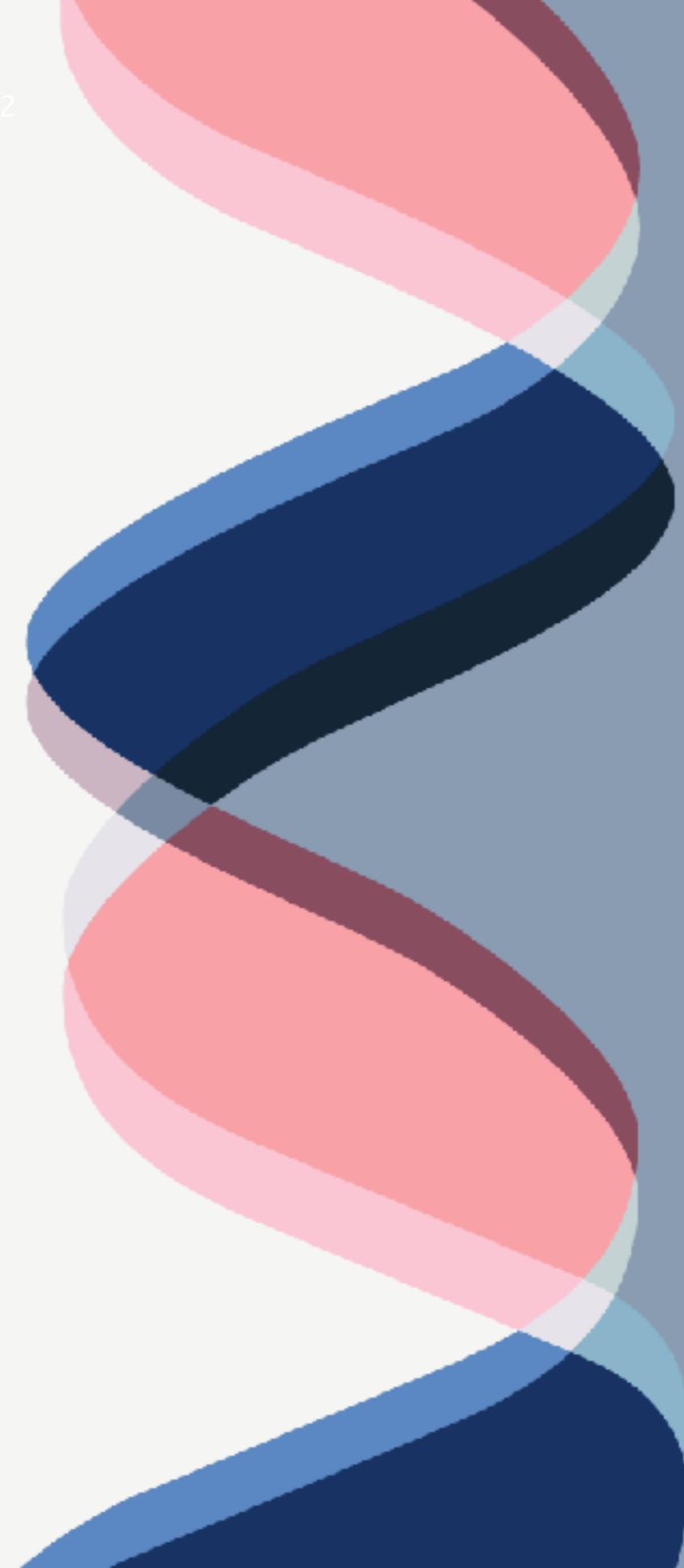
1 из 26 ашкеназских евреев является носителем болезни Тея-Сакса

Болезнь Тея-Сакса

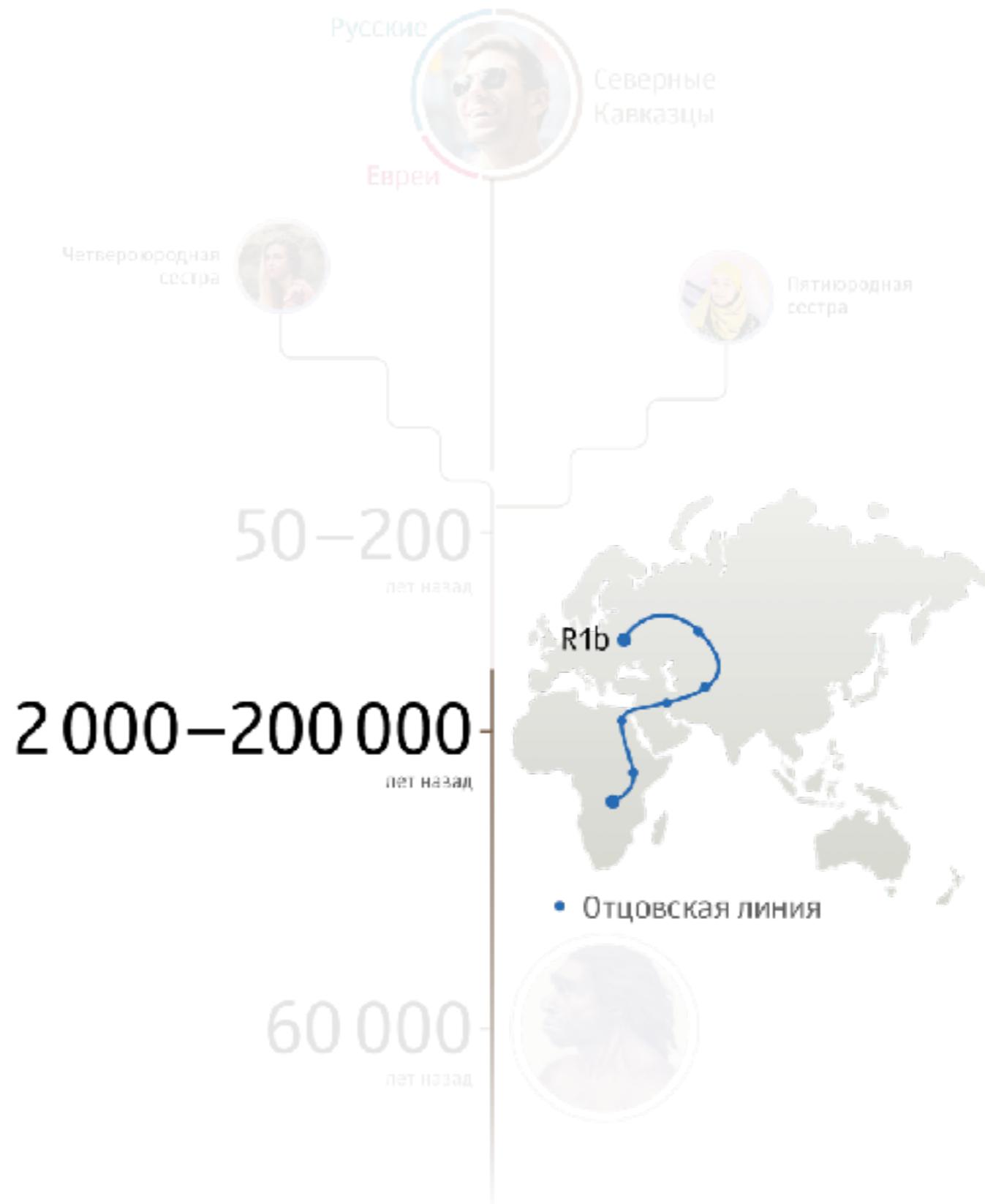
Дети с этим заболеванием рождаются без каких-либо симптомов и развиваются normally первые месяцы жизни. В возрасте около полугода происходит ухудшение умственного и физического развития. Дети теряют зрение, слух, возникают судороги, затрудняется глотание. Позже происходит атрофия мышц и паралич. Летальный исход наступает в возрасте до 4 лет.

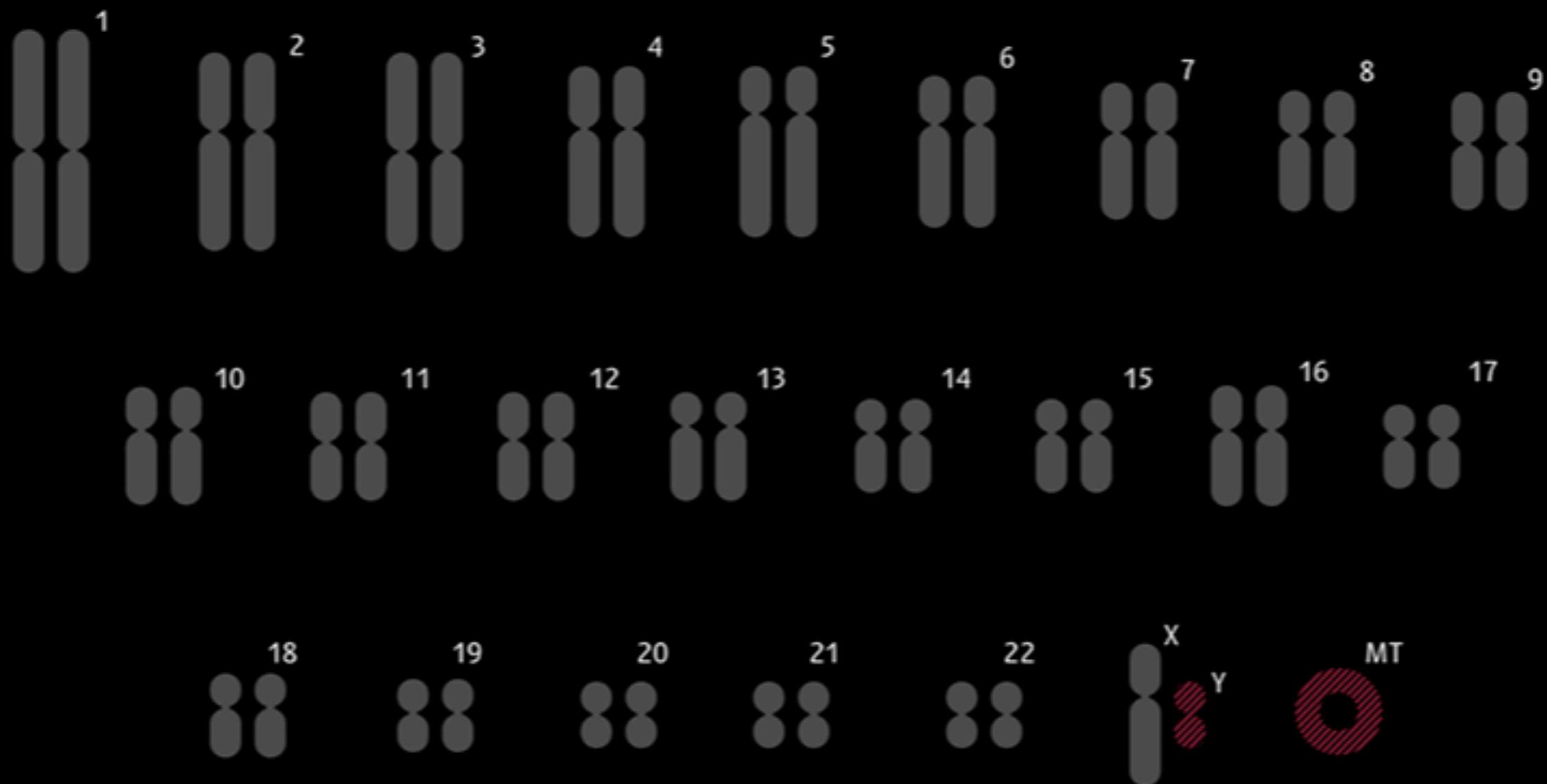
Haplogroups

1. Detect variants in Y chromosome and mtDNA
2. Find the optimal path in haplotree



Гаплогруппы

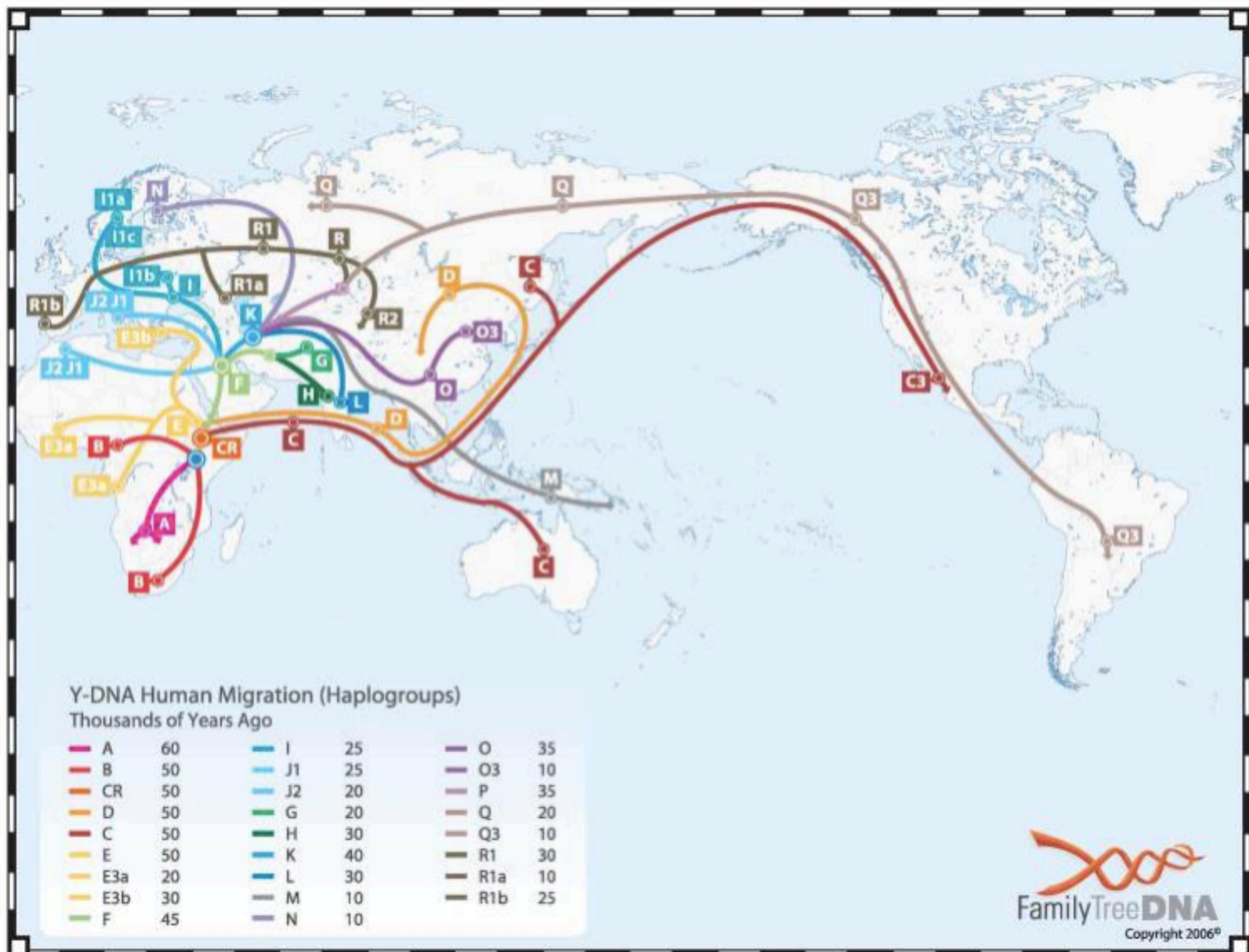




77 000
маркеров этнического
происхождения (доступно
для мужчин и женщин)

1500
маркеров происхождения
по отцовской линии
(доступно только для
мужчин)

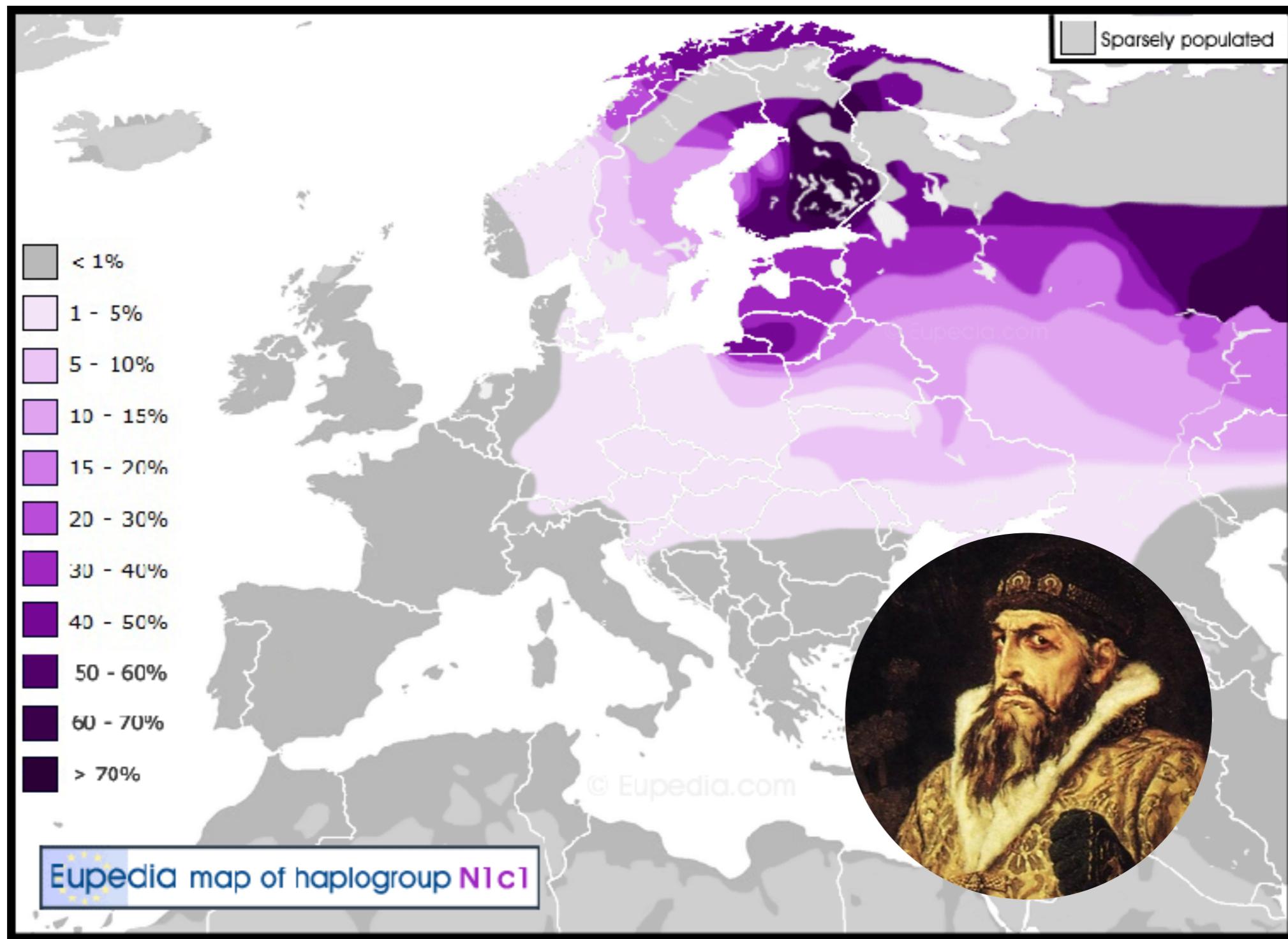
600
маркеров происхождения
по материнской линии
(доступно для мужчин и
женщин)



yDNA Path to N-L591



Распространенность гаплогруппы N1c1



Home A0-T A1 A1b BT CT CF F GHIJK HIJK IJK K K2 K-M2335 NO N N-Z4762 N-L729 N-Z1956 N-TA-
N-L1025 N-Y5580 N-BY158

N-L591 Y5581 * Y5583 * FGC76068 +5 SNPs formed 2200 ybp, TMRCA 2000 ybp info

N-L591*

N-Y24022 Y24023 * Y24022 * Y24024 formed 2000 ybp, TMRCA 1550 ybp info

 └ id:YF06727 LTU [LT-VL]

 └ id:YF03662 LTU [LT-PN]

N-Y5582 Y5576 * Y5577 * Y5582 +2 SNPs formed 2000 ybp, TMRCA 1650 ybp info

 └ id:YF66302 i

 └ id:YF01718

N-BY32561 BY32559/Y39460 * Y42620/BY32561 formed 2000 ybp, TMRCA 1500 ybp info

N-BY32561*

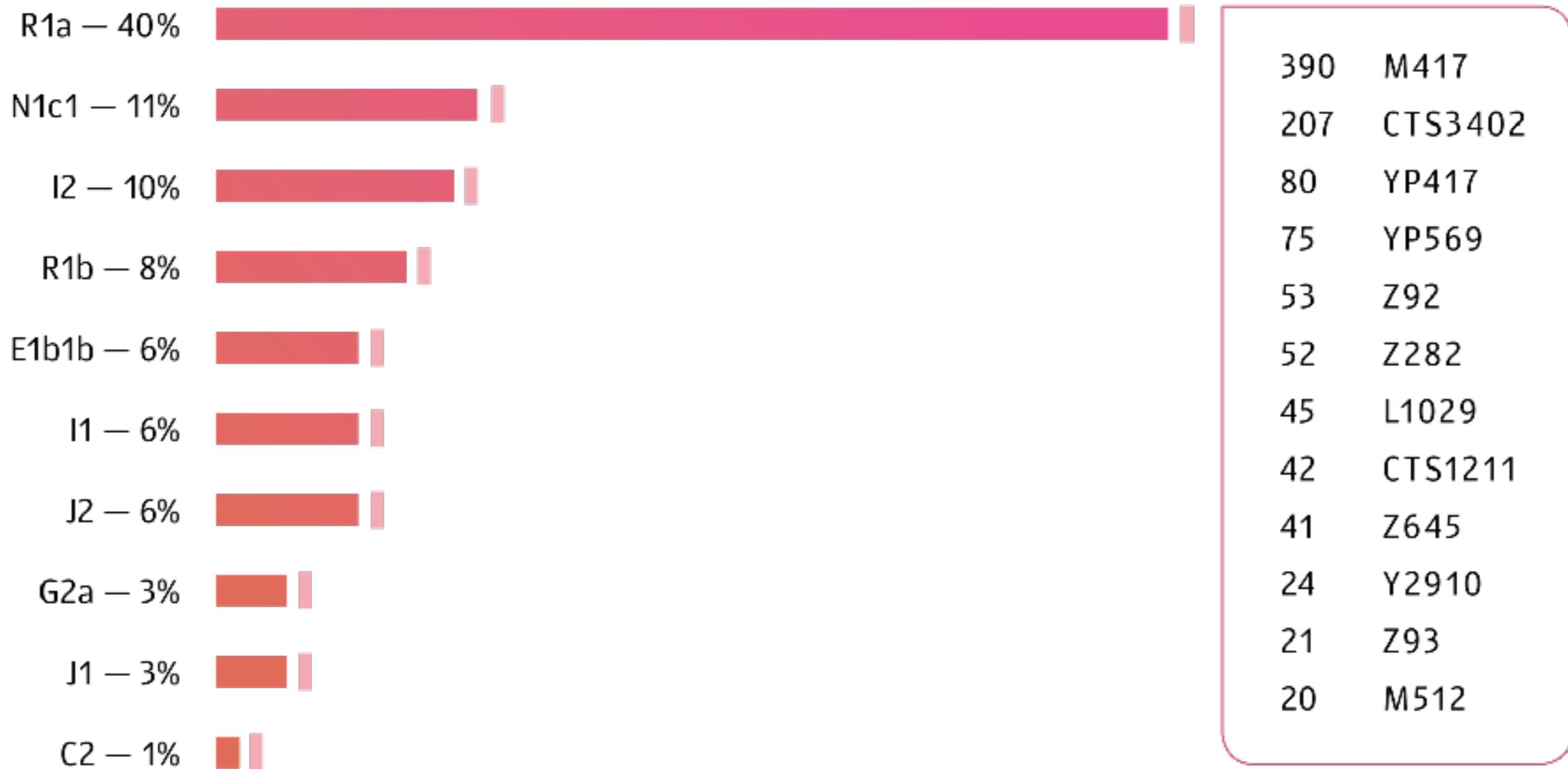
 └ id:YF12441 BLR [BY-VI]

N-Y40148 BY32560/Y40148 * BY32562/Y43192 formed 1500 ybp, TMRCA 1400 ybp info

 └ id:YF66873

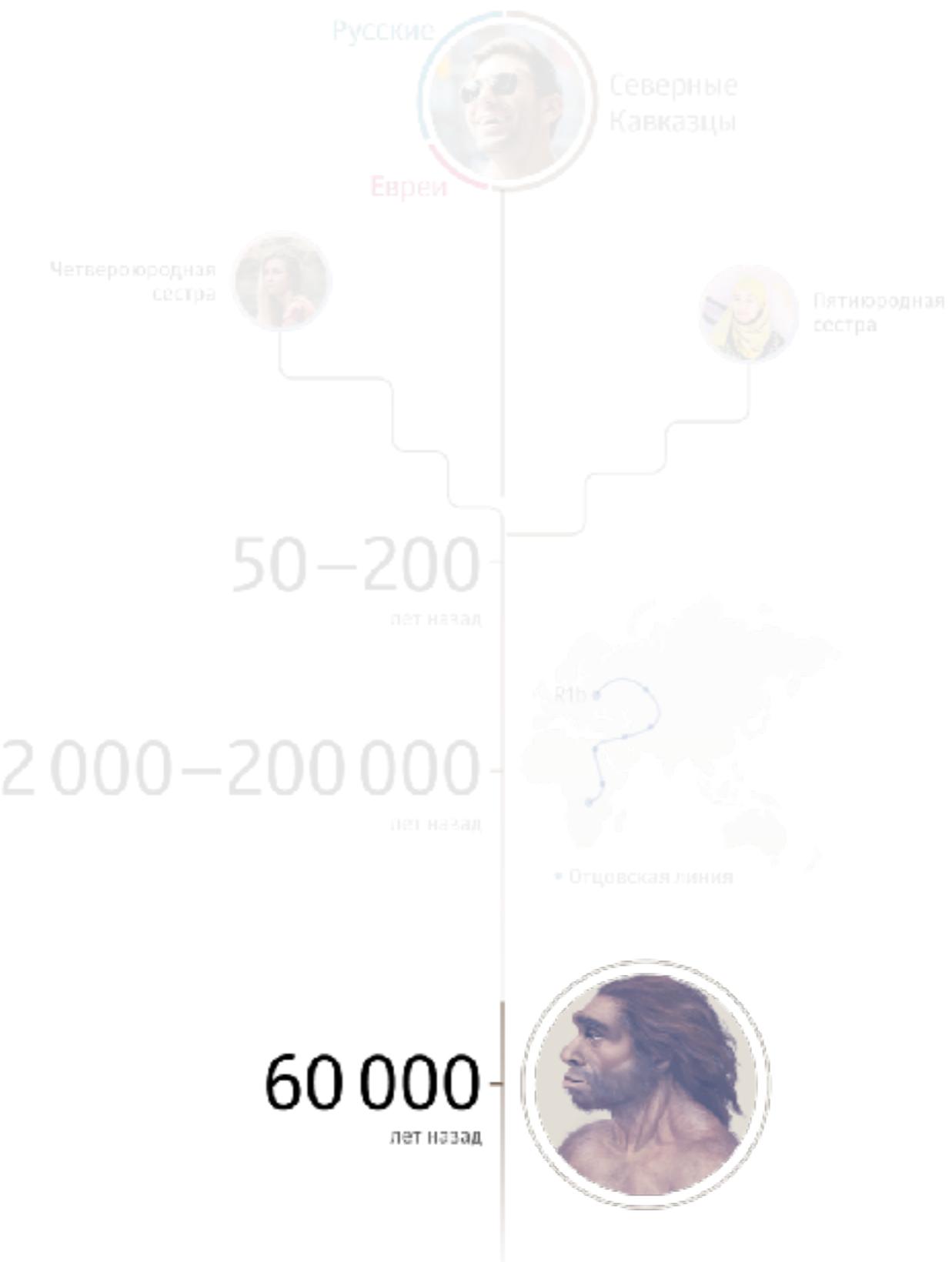
 └ id:YF02399

Y chromosome haplogroups

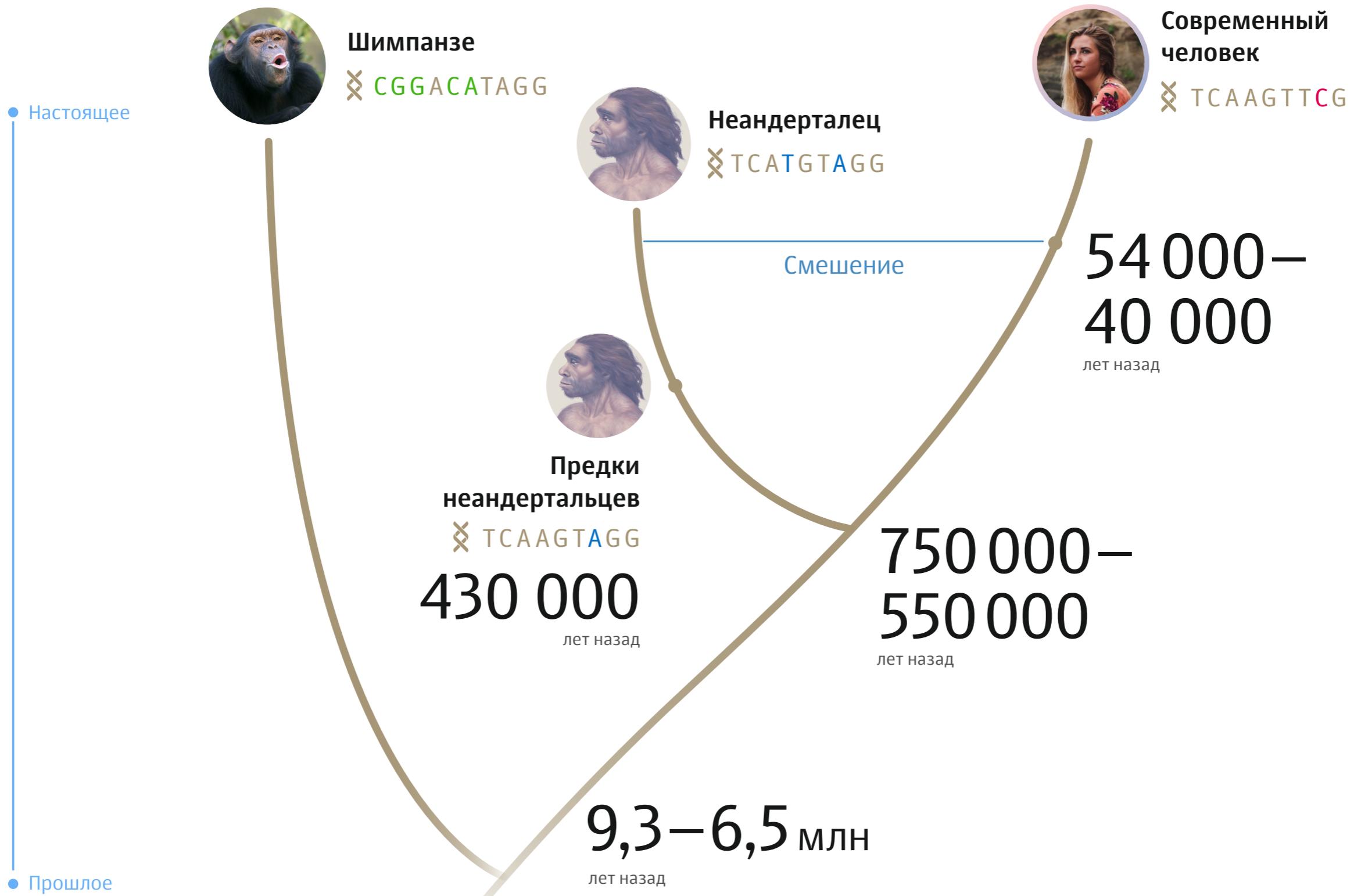


We studied the frequency distribution of mtDNA and Y-chromosome haplogroups. For example, the most frequent haplogroups for Y-chromosome were R1a (40%), I2 (10%), N1c1 (11%). M417 and CTS3402 mutations dominate among R1a subclades. More than 700 unique mtDNA subclades were detected.

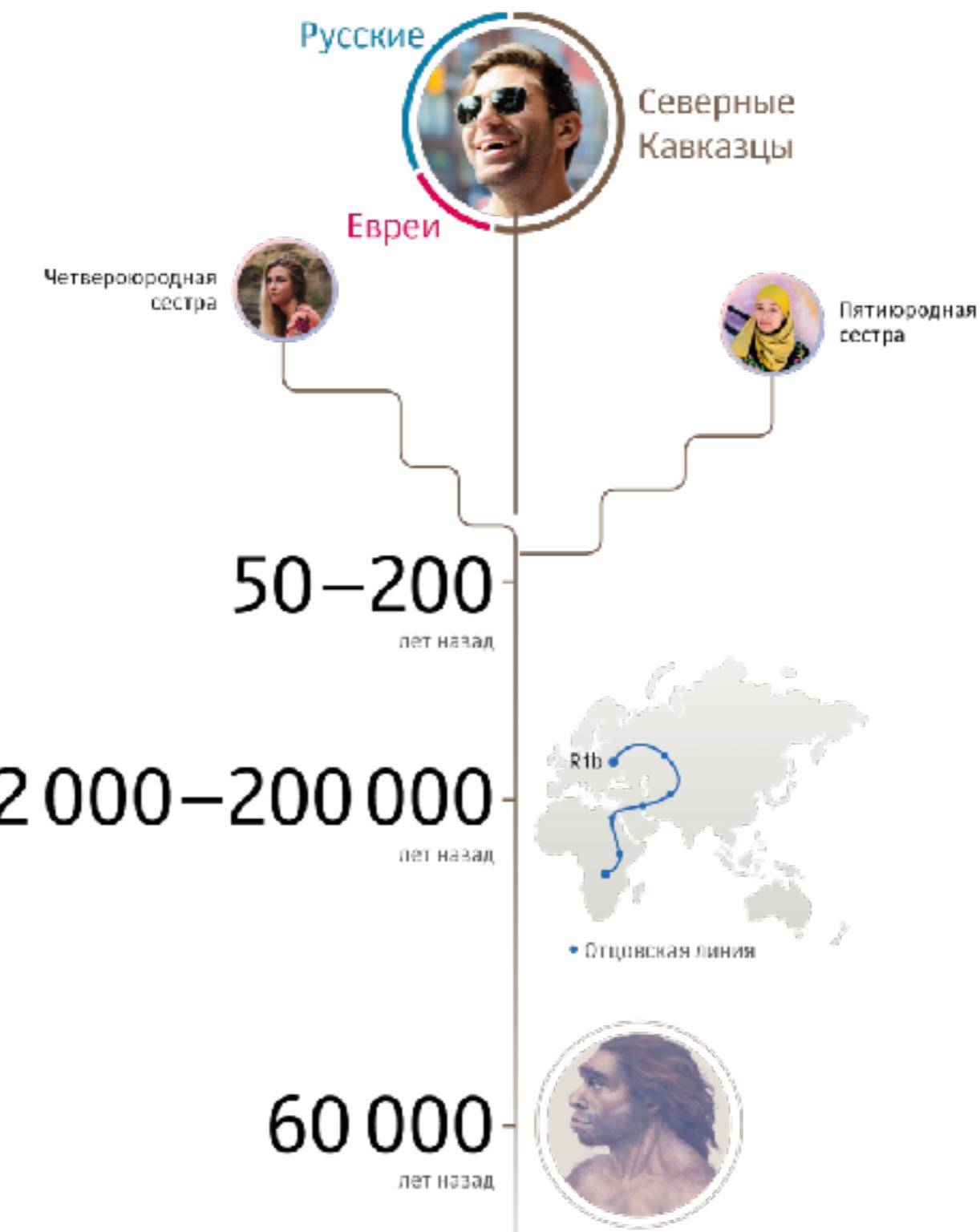
Гены неандертальца



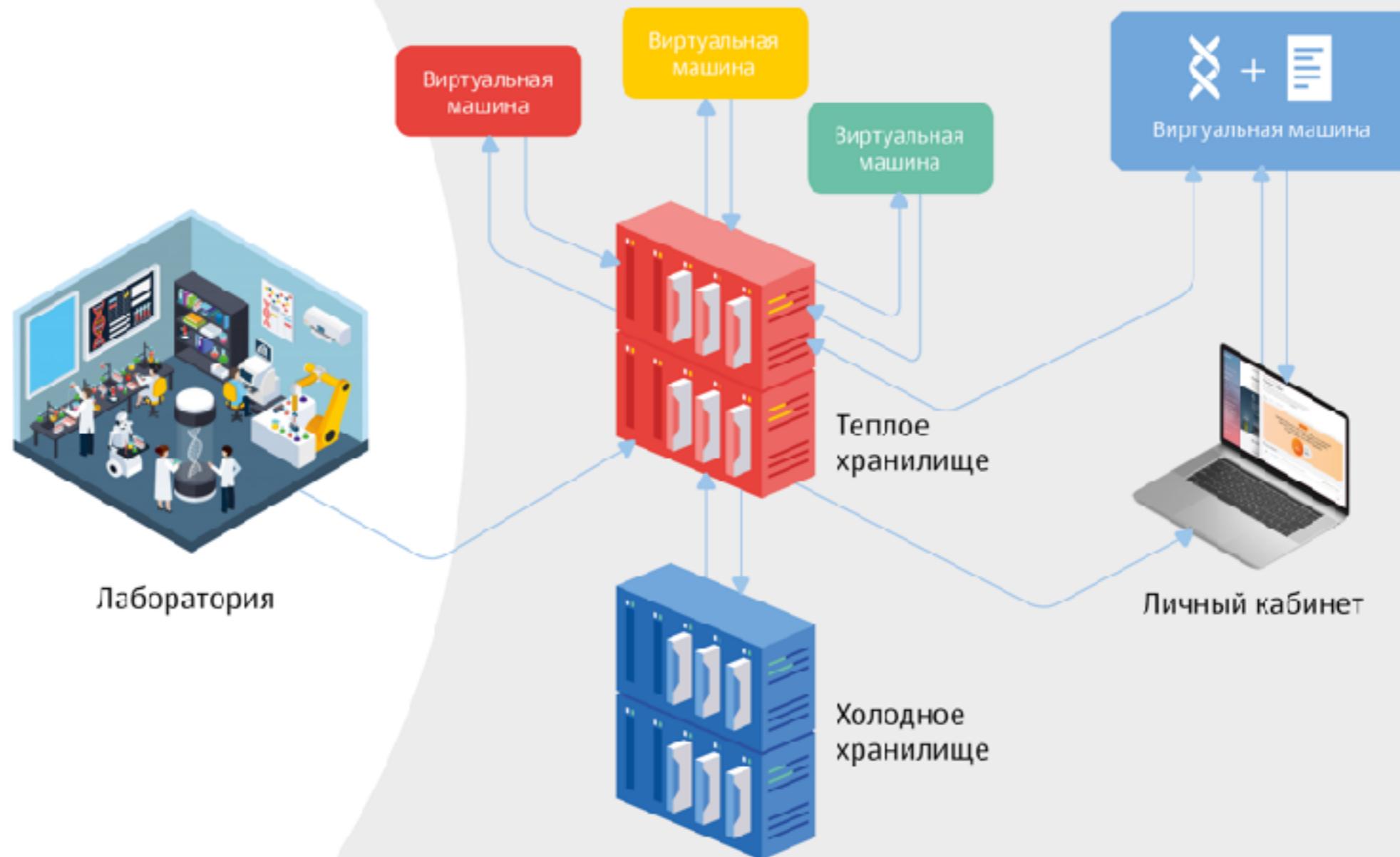
Гены неандертальца



Этнический состав



Существовавшая инфраструктура



Размерность данных

100 000 людей

До 80 000 000
генетических
вариантов для
каждого
человека

0/0	0/1	0/0	0/1	0/0					

- Данные хранятся в файлах
- Для работы с данными используются инструменты: bcftools, plink

Типовые запросы к генетическим

Запрос 1
1 мутация

у одного
человека

«Если ли у человека мутация
Лейдена, повышающая риск
тромбоза в 4.5 раза?»

Запрос 2
1000 мутаций

у одного
человека

«Посчитать риск
диабета 2-го типа»

Запрос 3
650 000

мутаций у 10

«Проверить родство,
построить семейное
дерево»

ClickHouse и генетические данные

		Ind1	Ind2	...	Ind100
Chr1	Pos1	0/1	0/0	...	0/0
Chr1	Pos2	0/0	1/1	...	0/0
...
Chr22	Pos100	0	1/1	1/1	0/0



Chr1	Pos1	Ind1	0/1
Chr1	Pos2	Ind1	0/0
...
Chr22	Pos1000	Ind1	1/1
Chr1	Pos1	Ind2	0/0
Chr1	Pos2	Ind2	1/1
...
Chr22	Pos1000	Ind2	1/1
Chr1	Pos1	Ind3	0/0
Chr1	Pos2	Ind3	0/0
...
Chr22	Pos1000	Ind3	0/0

query_10_8M	224m51.103s	186m46.171s	66m24.761s	70m30.792s	91mb.248s	//m/.648s
	300GB				25GB	

Бенчмарк сравнения по типовым запросам в секундах

В качестве модельного датасета использовались
данные проекта 1000 Genomes (80М мутаций x 2.5К
людей)

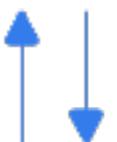


Новая инфраструктура



Virtual Private Cloud (VPC)

Yandex
Functions



Managed
Databases



ПРИМЕР

Конвертация генетических данных клиента

	РЕШЕНИЕ 1 Хранить готовые файлы	РЕШЕНИЕ 2 Конвертировать	РЕШЕНИЕ 3 Конвертировать в формат
ТЕХНОЛОГИИ	S3	S3 + СЕРВЕР	CLICKHOUSE + YANDEX FUNCTION
Скорость	Быстро	Медленно	Быстро
Обслуживание сервера	Не требуется	Требуется	Не требуется
Дополнительные ресурсы на хранение	Требуется	Не требуется	Не требуется

Genome assembly using quantum and quantum-inspired annealing

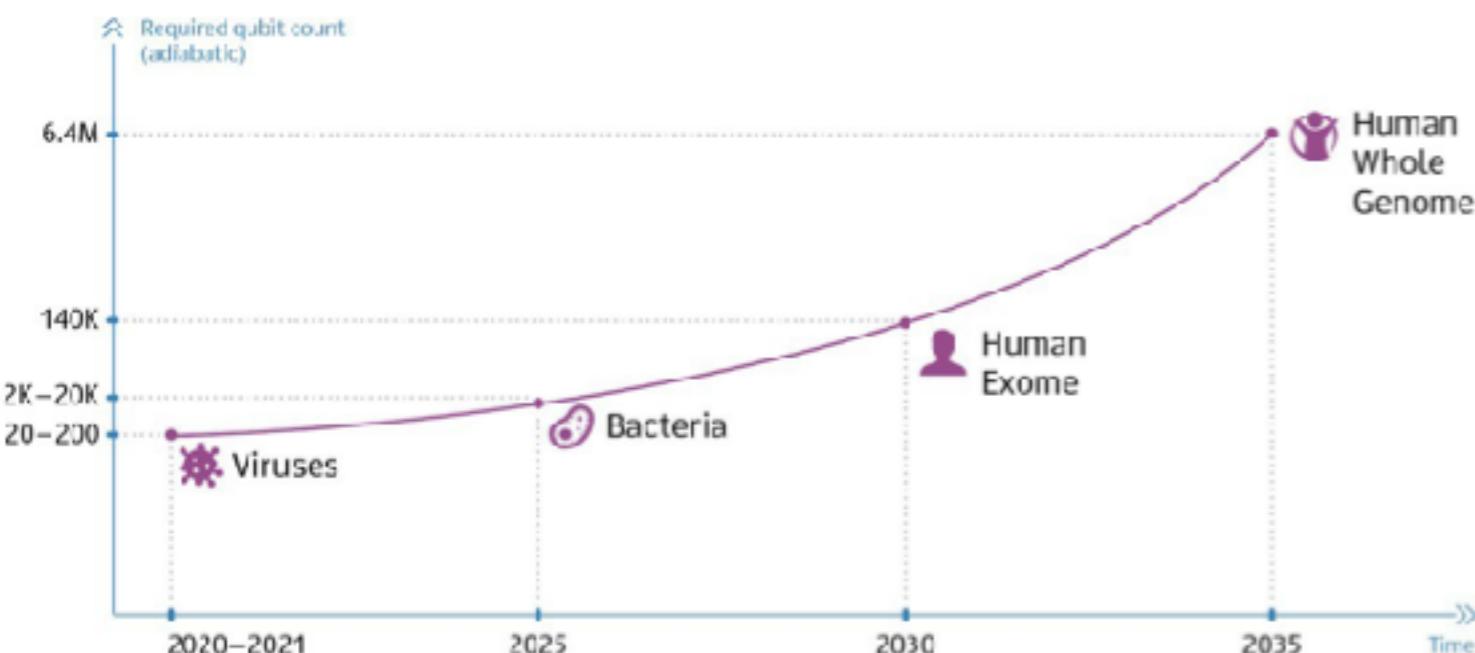
A.S. Boev,¹ A.S. Rakitko,² S.R. Usmanov,¹ A.N. Kobzeva,¹
I.V. Popov,² V.V. Ilinsky,² E.O. Kiktenko,¹ and A.K. Fedorov¹

¹Russian Quantum Center, Skolkovo, Moscow 143025, Russia

²Genotek ltd., Moscow 105120, Russia

(Dated: April 23, 2020)

Recent advances in DNA sequencing open prospects to make whole-genome analysis rapid and reliable, which is promising for various applications including personalized medicine. However, existing techniques for *de novo* genome assembly, which is used for the analysis of genomic rearrangements, chromosome phasing, and reconstructing genomes without a reference, require solving tasks of high computational complexity. Here we demonstrate a method for solving genome assembly tasks with the use of quantum and quantum-inspired optimization algorithms. Within this method, we present experimental results on genome assembly using quantum annealers both for simulated data and the ϕX 174 bacteriophage. Our results pave a way for a significant increase in the efficiency of solving bioinformatics problems with the use of quantum computing technologies and, in particular, quantum annealing might be an effective method. We expect that the new generation of quantum annealing devices would outperform existing techniques for *de novo* genome assembly. To the best of our knowledge, this is the first experimental study of *de novo* genome assembly problems both for real and synthetic data on quantum annealing devices and quantum-inspired algorithms.



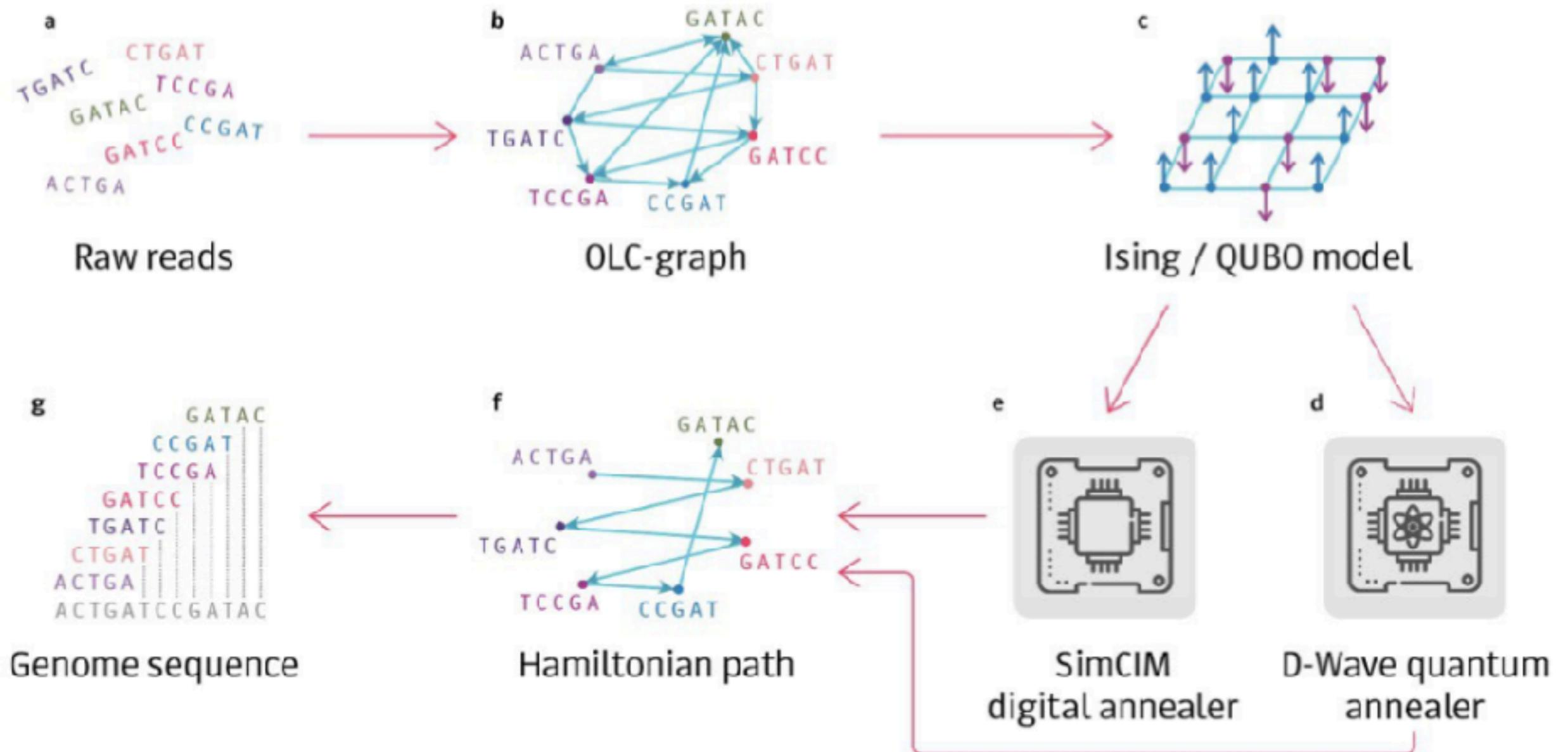


Figure 1. Solving the *de novo* genome assembly problem using quantum annealers and quantum-inspired algorithms: a) raw reads; b) raw reads are transformed to the overlap-layout-consensus (OLC) graph; c) finding the Hamiltonian path for the OLC graph is reduced to the Ising problem; d) and e) the Ising problem in QUBO form can be solved using quantum annealers (D-Wave) and quantum-inspired algorithms (SimCIM); in f) the output is the Hamiltonian path; in g) the genome sequence is obtained as the solution.

Thank you for your attention!



<https://twitter.com/AlexRakitko>

rakitko@genotek.ru



+7 495 215-15-14
www.genotek.ru
info@genotek.ru

Наставнический переулок,
д. 17, стр. 1, корп. 15
Москва, 105120, Россия