Генетическая генеалогия

Александр Ракитько

Genotek

My background

- Studied theory of probability at MSU, Russia
- Developed personal DNA tests at Genotek as team-lead of biostatisticians
- Work as CPO at Genotek

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

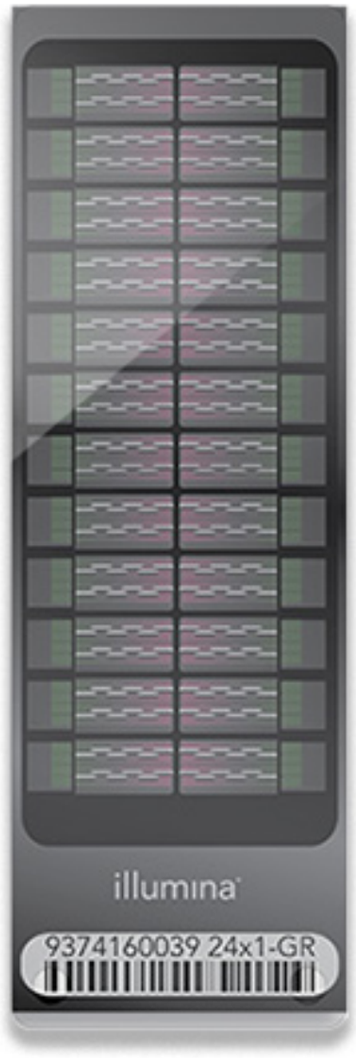
- DNA sequencing
- RNA-Seq
- Metagenome
- Methylome

23andMe®

MyHeritage DNA

FamilyTreeDNA

Genotek

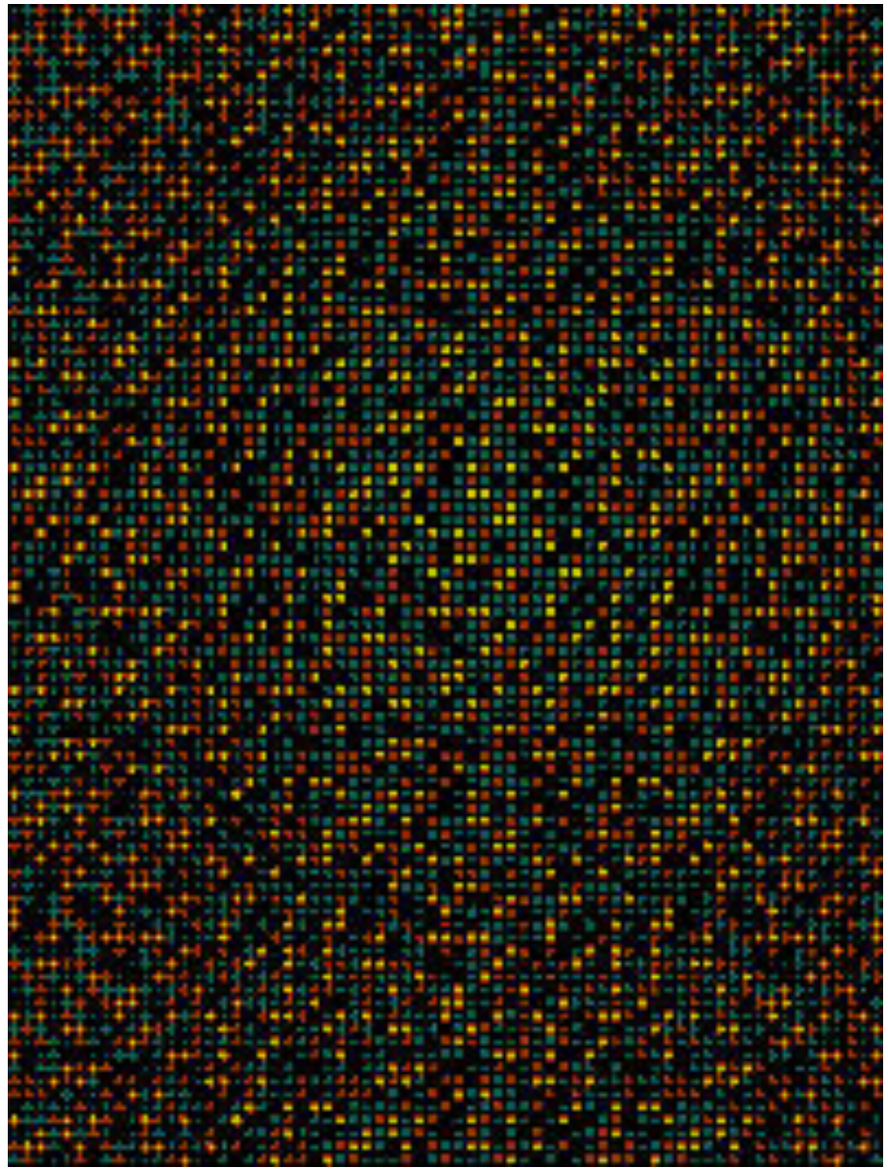


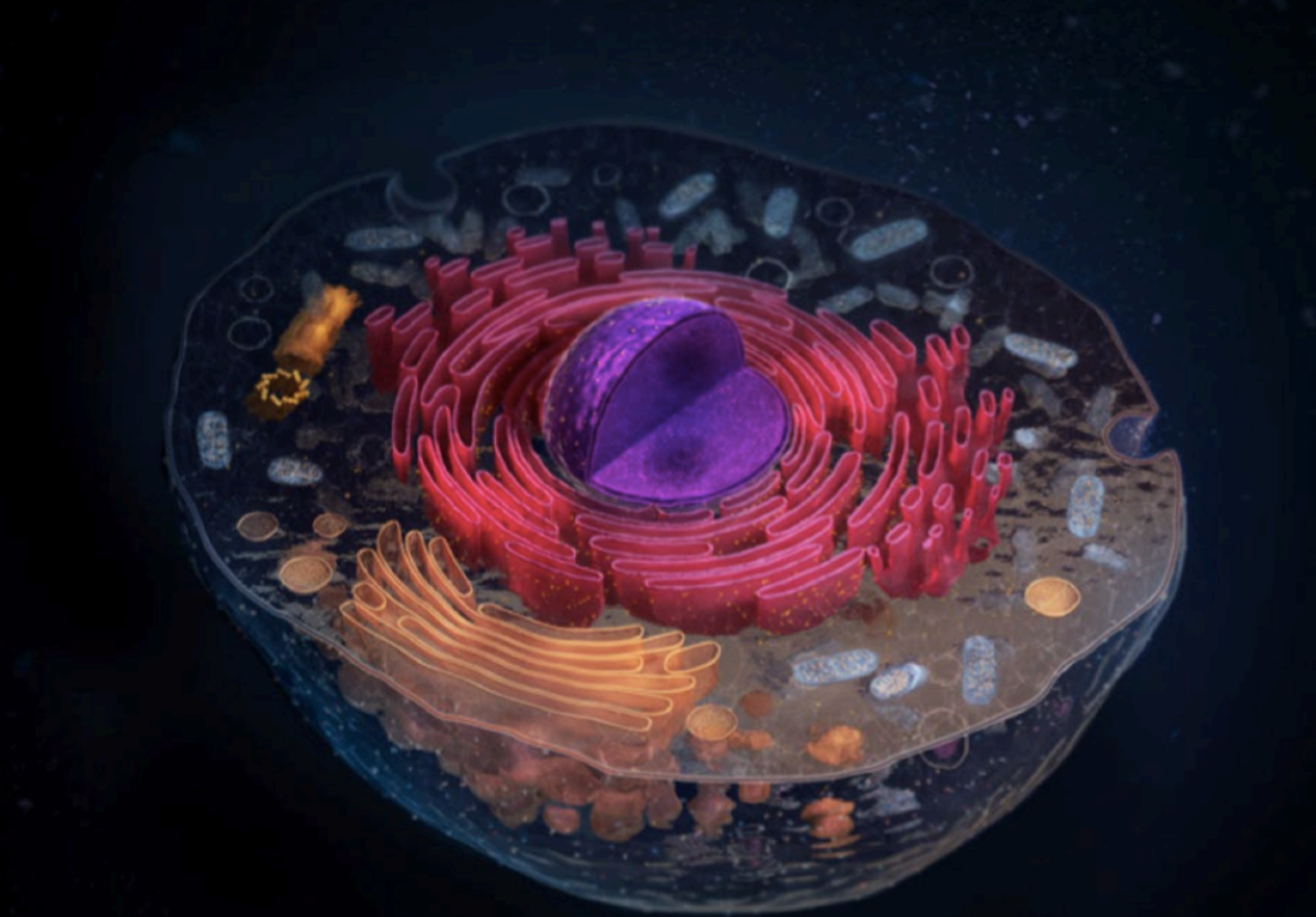
Illumina Global Screening Array

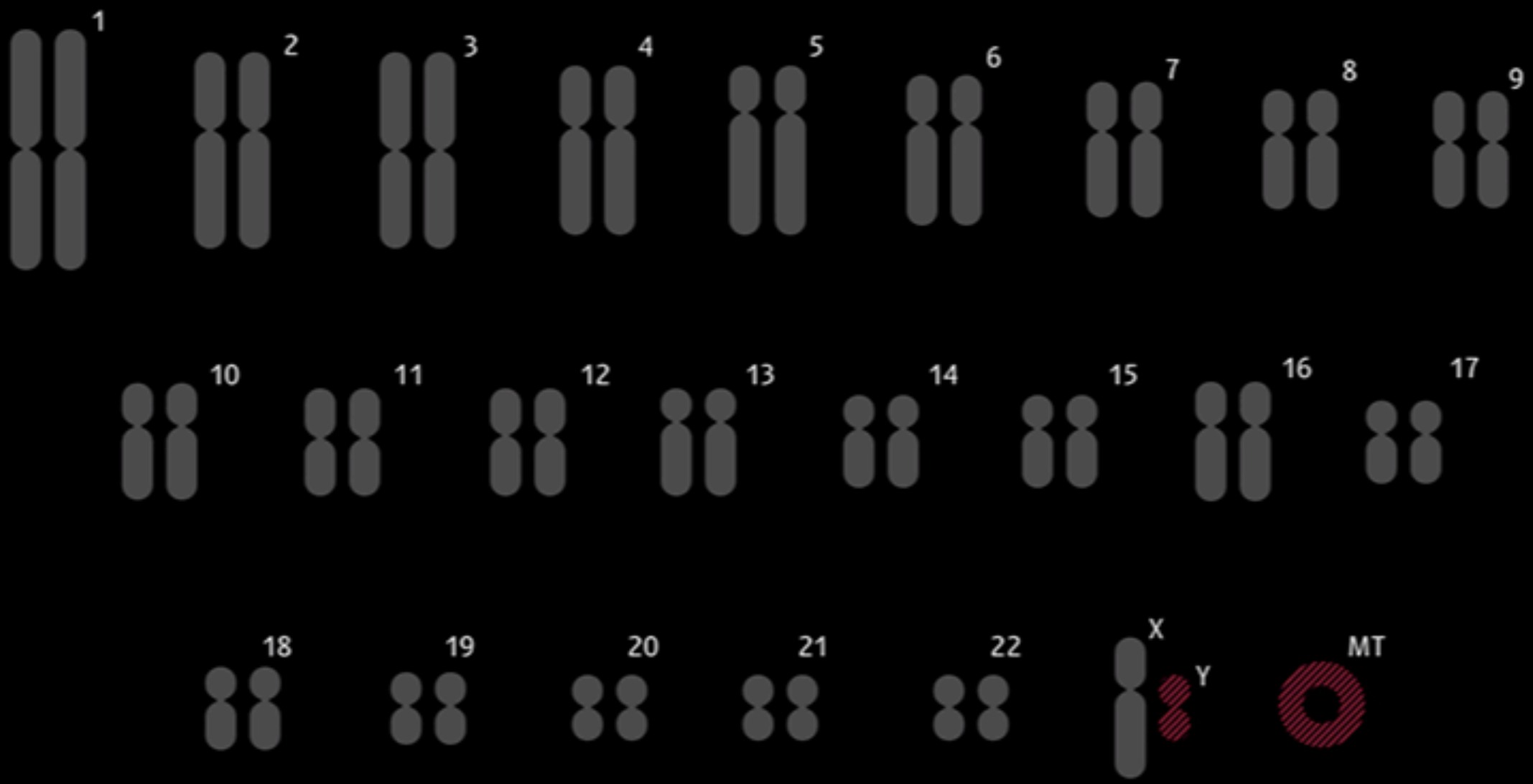

HiScan
illumina


illumina


0114848012





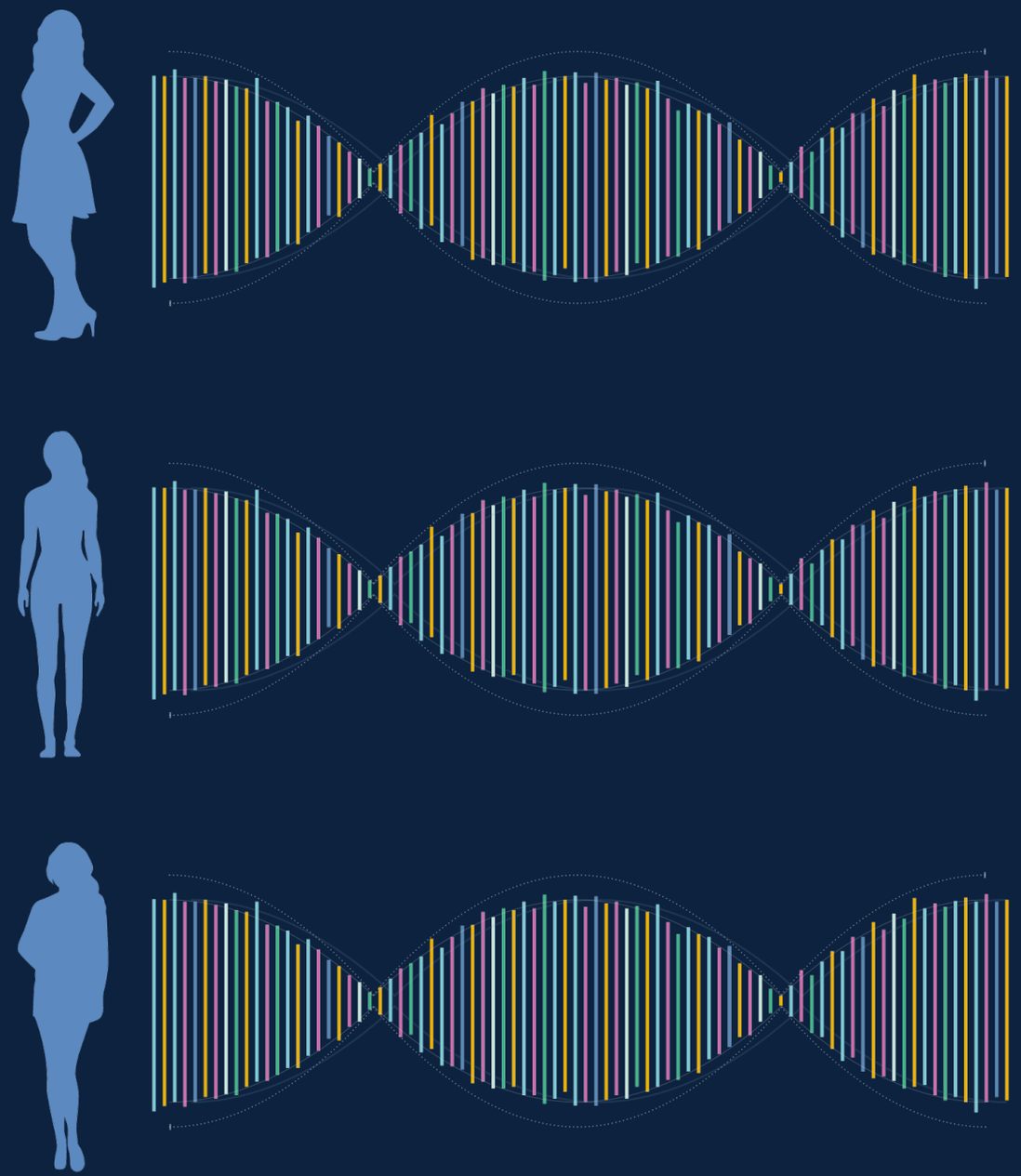


77 000
маркеров этнического происхождения (доступно для мужчин и женщин)

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

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

Что такое генетический полиморфизм?



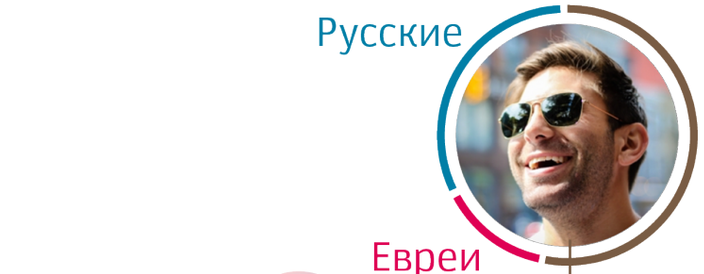
...AATGCAATGCGA...

...AATGCAGTGCGA...

...AATGCACTGCGA...

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




Четвероюродная сестра


Пятиюродная сестра

50–200
лет назад

2 000–200 000
лет назад

• Отцовская линия

60 000
лет назад

Этнический состав

Русские

Северные Кавказцы

Евреи

Четвероюродная сестра

Пятиюродная сестра

50–200 лет назад

2 000–200 000 лет назад

Отцовская линия

60 000 лет назад

Ancestry decomposition (part 1)
Create database

1. Find datasets

Estonian Biocentre

Free data

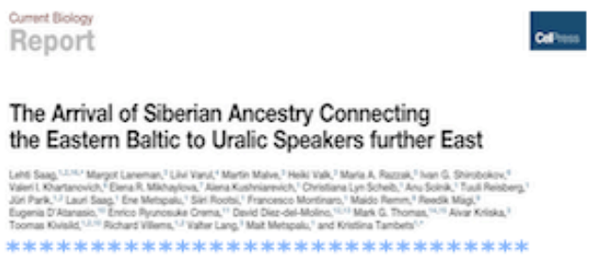
For academic research, please find here human DNA sequence and genotype data from recent papers.
Complete high coverage human genomes and Y chromosome sequence data has been generated by Complete Genomics
SNP data has been generated with different Illumina genotyping arrays since 2008. Data has been 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 mait@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

paper: || Data


Current Biology
Report

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

paper: || Data

bioRxiv

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 sequ
"1240K+HO": Data from the above set merged with present-day individuals t
```

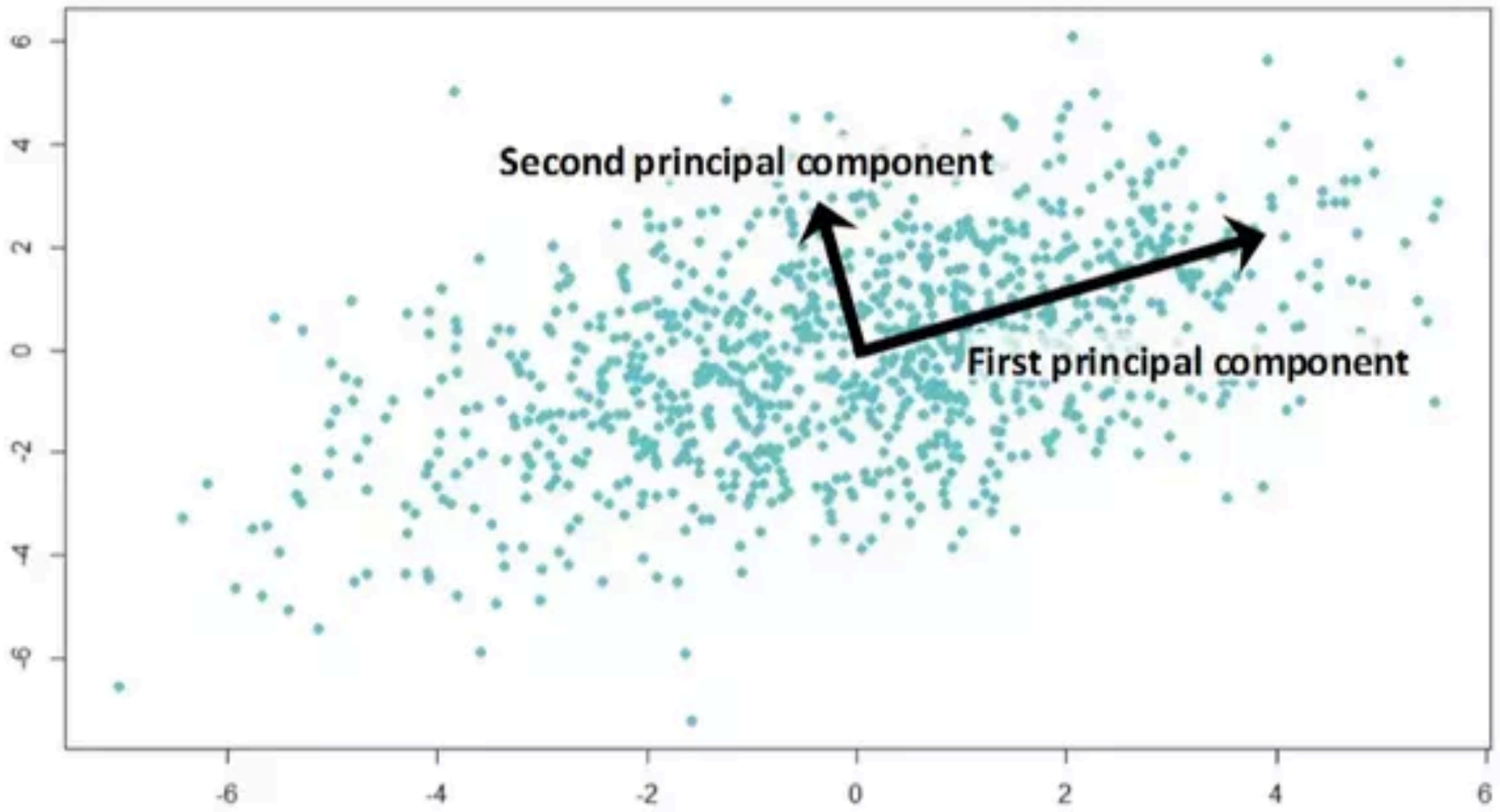
Each dataset consists of four files, in eigenstrat format. For details, please see: eigensoft:

```
.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/geneti
```

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)

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.

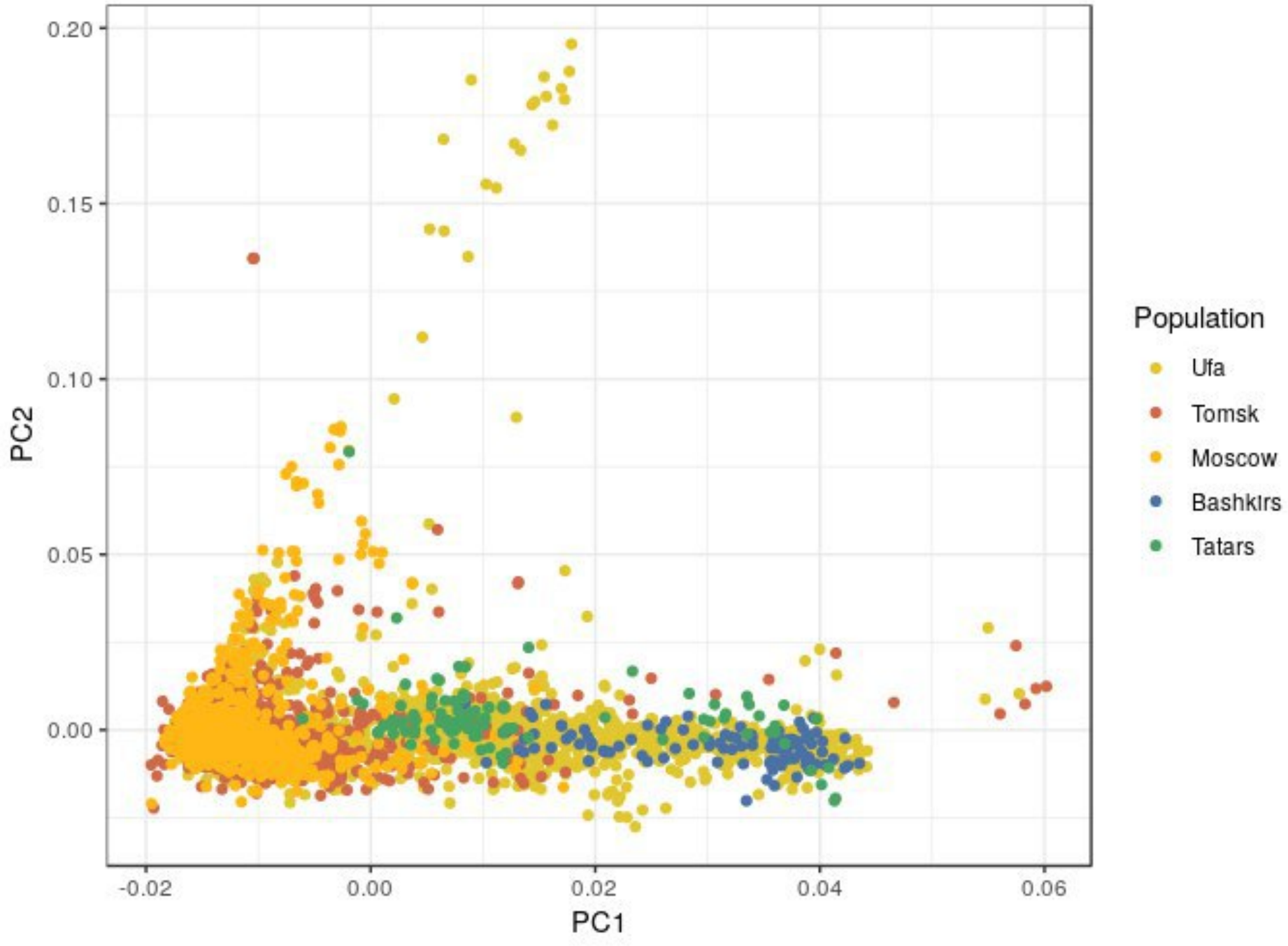
1. Make sure you have PLINK installed on your machine (see these instructions).
2. Download the example data archive file which contains the genotypes, map files and two extra phenotype files, described below (zipped, approximately 2.8M)
3. Create a new folder/directory on your machine, and unzip the 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

---

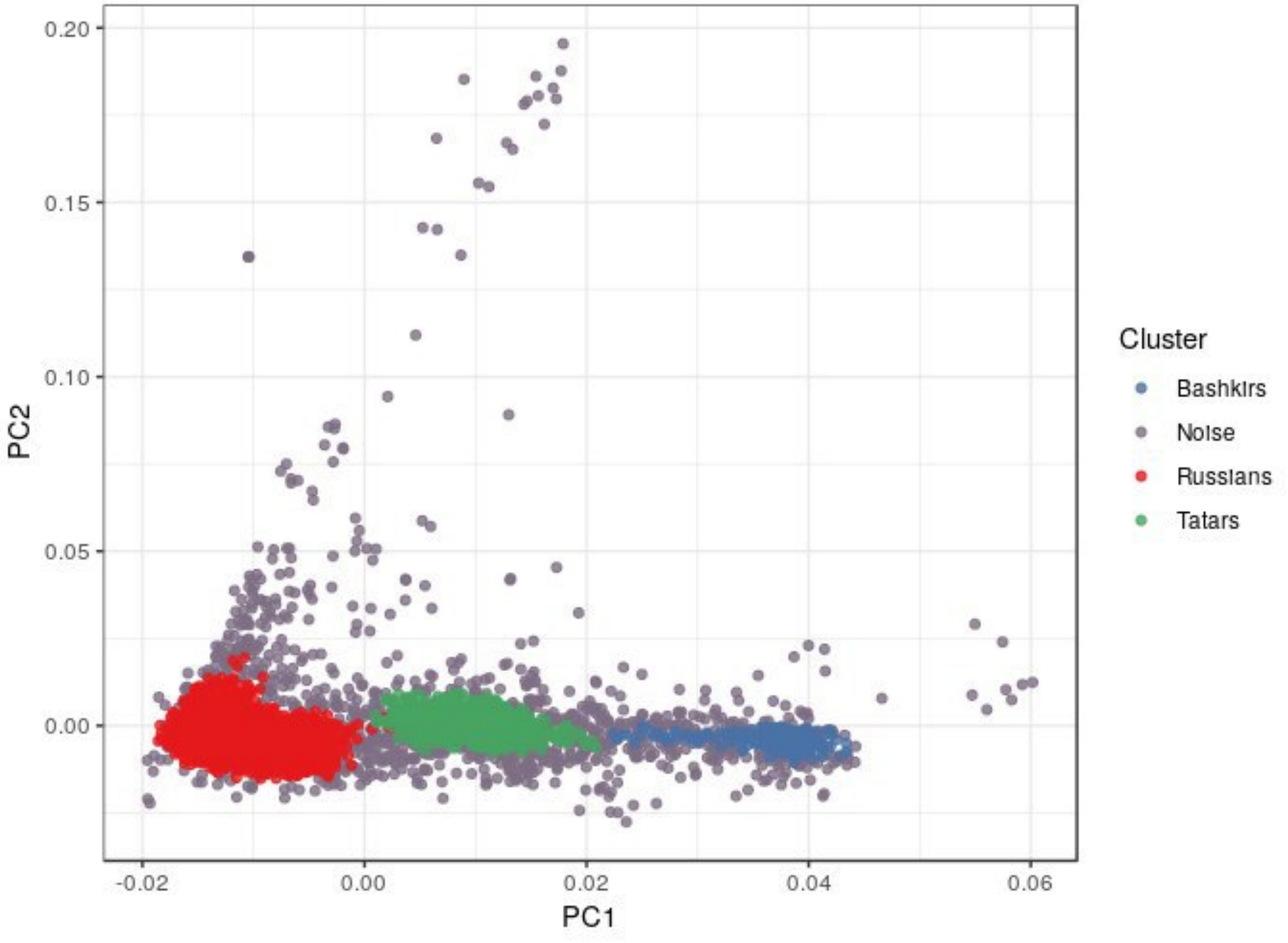
Two phenotypes were generated: a quantitative triat and a disease trait (affection status, coded 1=unaffected, 2=affected), based on a median split of the quantitative trait. The quantitative trait was generated as a function of three simple components:

- A random component
- Chinese versus Japanese identity

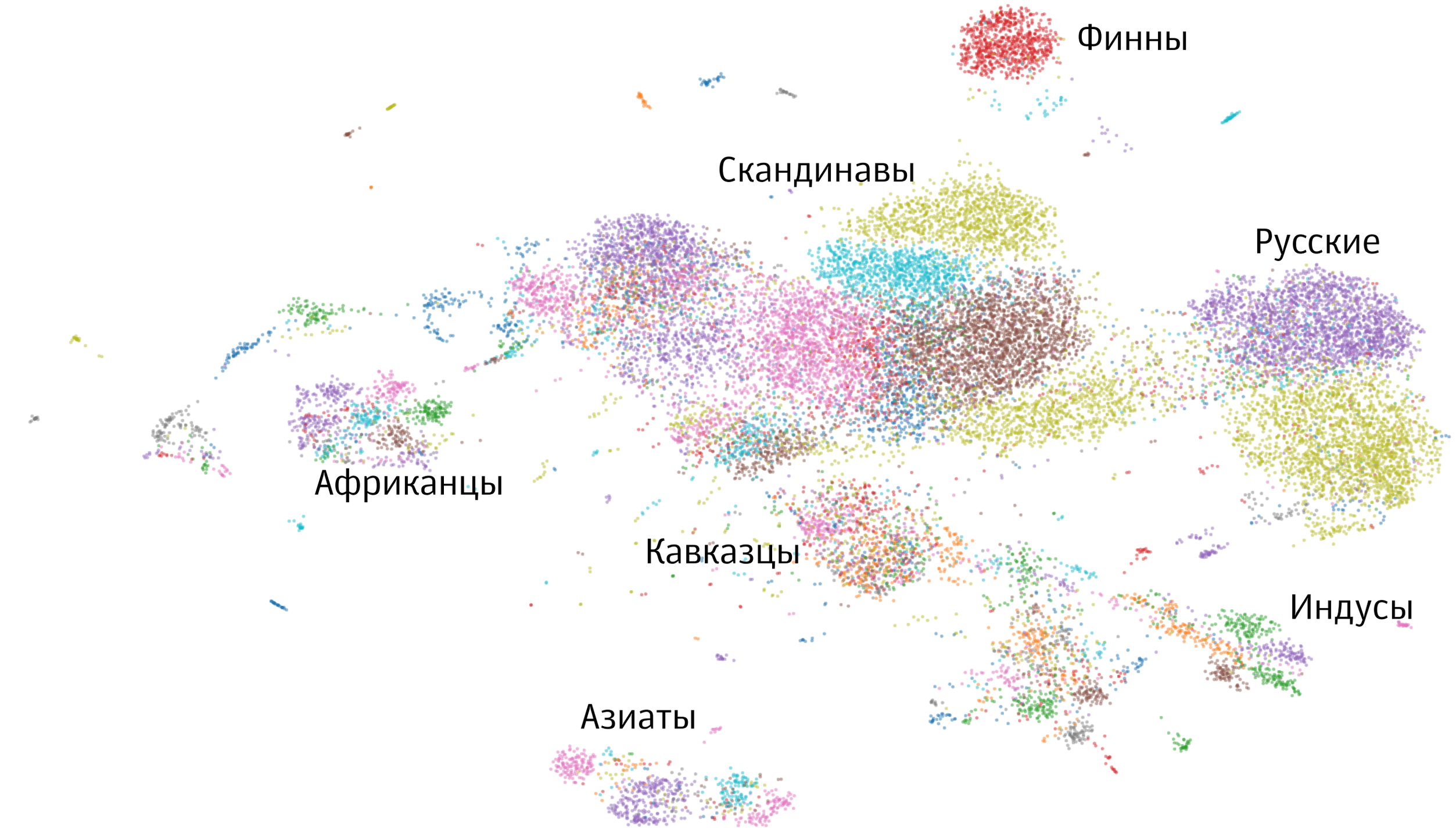

0.20
0.15
0.10
0.05
0.00
PC2
-0.02
0.00
0.02
0.04
0.06
PC1
Population
Ufa
Tomsk
Moscow
Bashkirs
Tatars

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)


0.20
0.15
0.10
0.05
0.00
PC2
-0.02
0.00
0.02
0.04
0.06
PC1
Cluster
Bashkirs
Noise
Russians
Tatars

UMAP of 255 populations

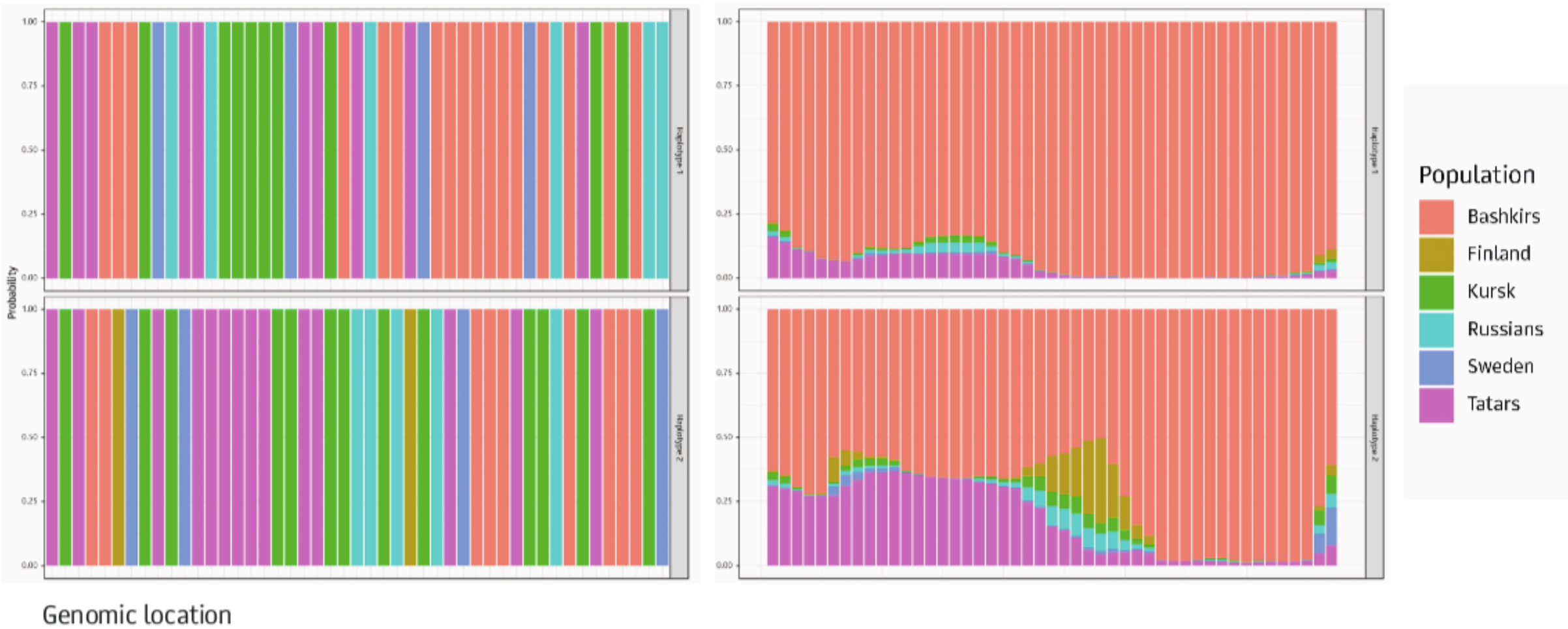


Ancestry decomposition (part 2)
Local ancestry inference

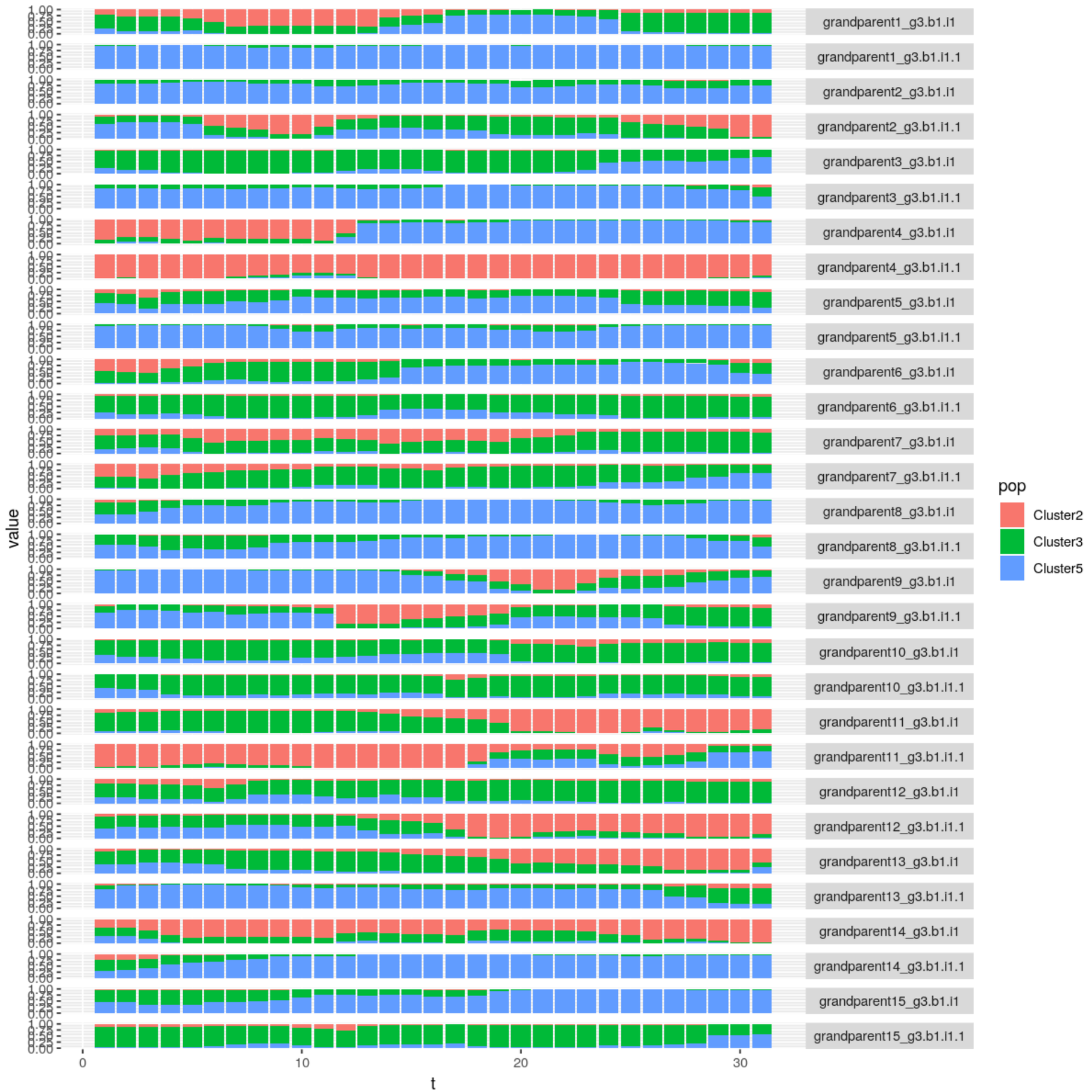
1. Classify 100-500 SNP windows (SVM)

2. Error correction (HMM)

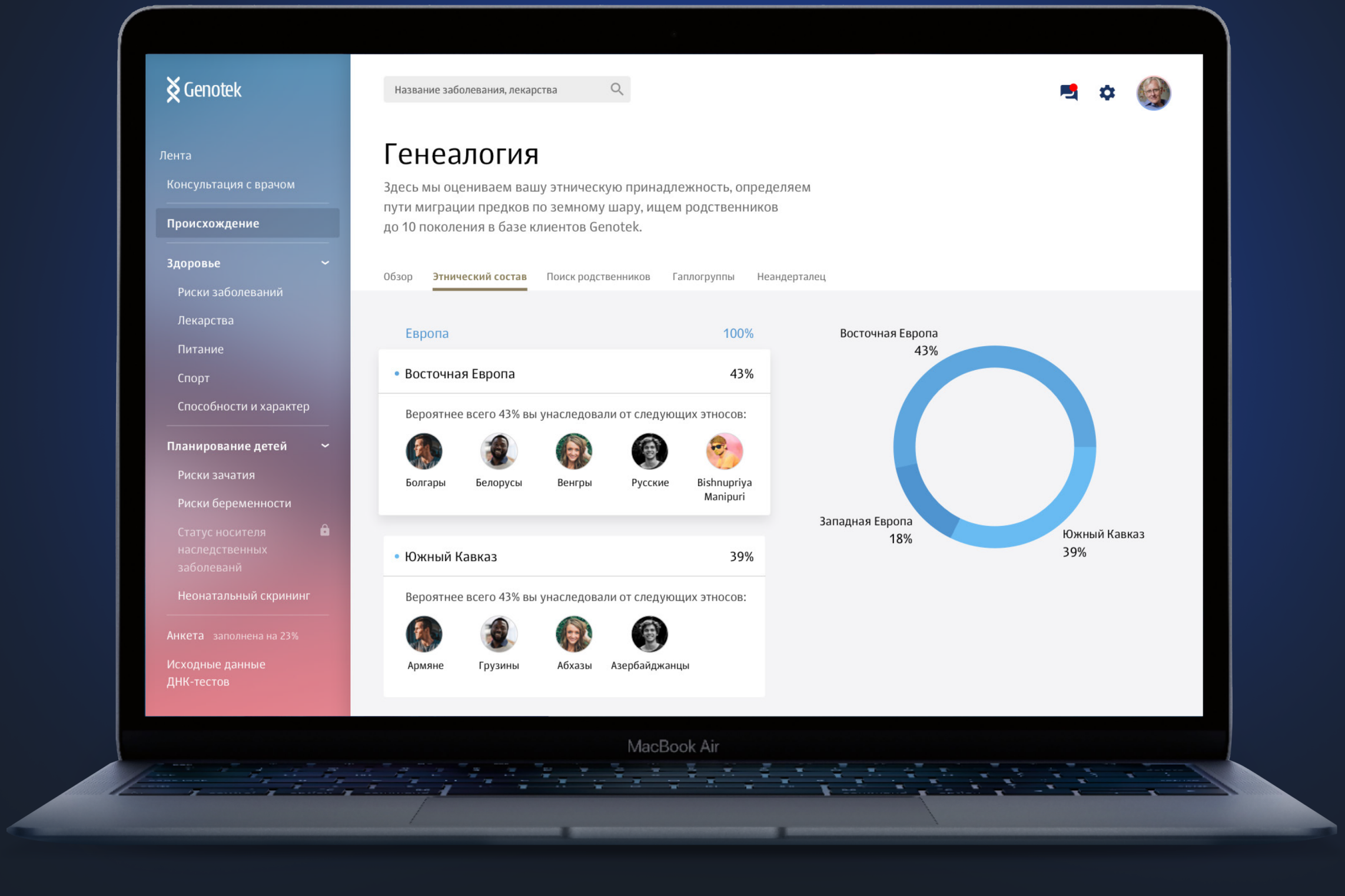
Ancestry proportions of the individual with Bashkir ethnic origin estimated with SVM (left) and HMM (right)



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


grandparent1_g3.b1.i1
grandparent1_g3.b1.i1.1
grandparent2_g3.b1.i1
grandparent2_g3.b1.i1.1
grandparent3_g3.b1.i1
grandparent3_g3.b1.i1.1
grandparent4_g3.b1.i1
grandparent4_g3.b1.i1.1
grandparent5_g3.b1.i1
grandparent5_g3.b1.i1.1
grandparent6_g3.b1.i1
grandparent6_g3.b1.i1.1
grandparent7_g3.b1.i1
grandparent7_g3.b1.i1.1
grandparent8_g3.b1.i1
grandparent8_g3.b1.i1.1
grandparent9_g3.b1.i1
grandparent9_g3.b1.i1.1
grandparent10_g3.b1.i1
grandparent10_g3.b1.i1.1
grandparent11_g3.b1.i1
grandparent11_g3.b1.i1.1
grandparent12_g3.b1.i1
grandparent12_g3.b1.i1.1
grandparent13_g3.b1.i1
grandparent13_g3.b1.i1.1
grandparent14_g3.b1.i1
grandparent14_g3.b1.i1.1
grandparent15_g3.b1.i1
grandparent15_g3.b1.i1.1
value
0
10
20
30
t
pop
Cluster2
Cluster3
Cluster5

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.

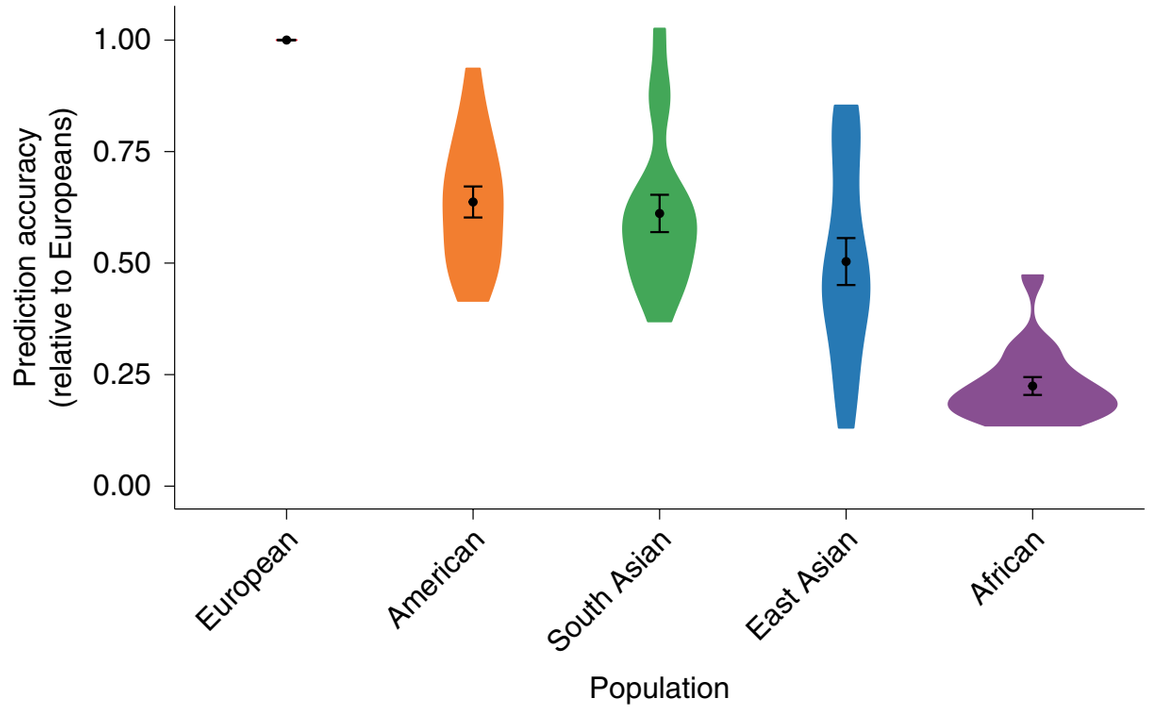


Fig. 3 | Prediction accuracy relative to European-ancestry individuals across 17 quantitative traits and 5 continental populations in the UKBB. All phenotypes shown here are quantitative anthropometric and blood-panel traits, as described in Supplementary Table 6, which includes discovery-cohort sample 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.

Поиск родственников

Русские

Северные Кавказцы

Евреи

Четвероюродная сестра

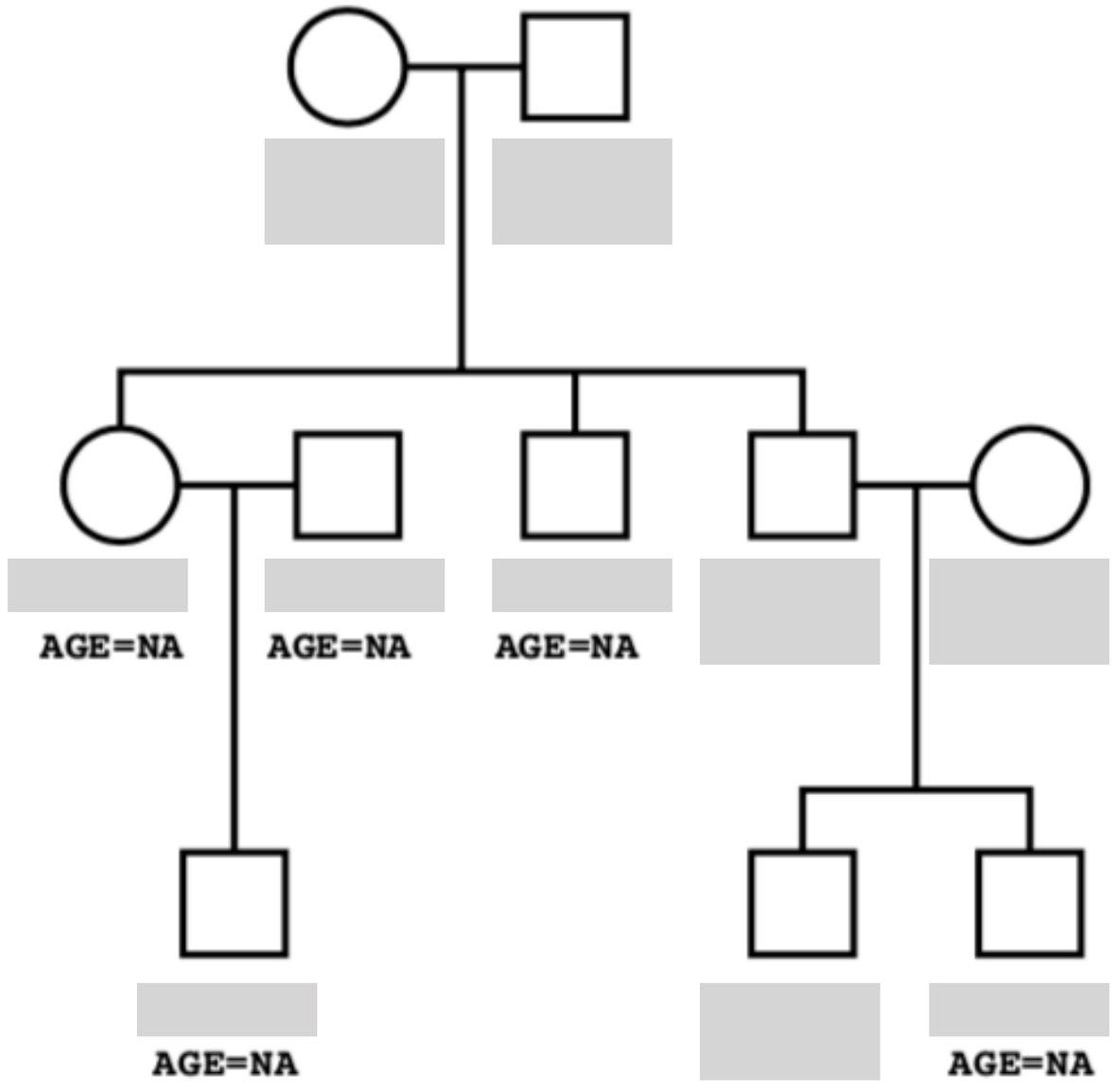
Пятиюродная сестра

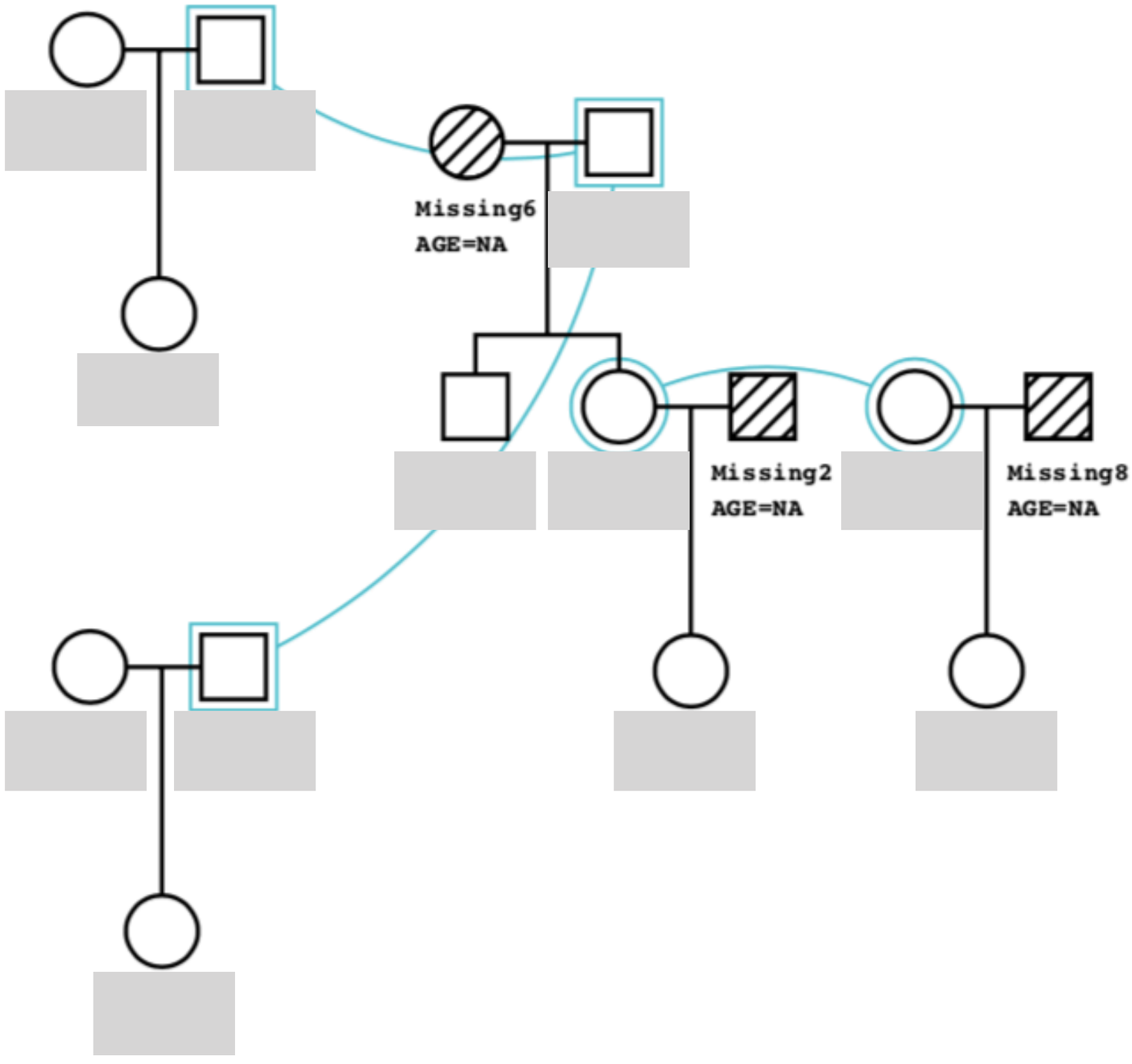
50–200
лет назад

2 000–200 000
лет назад

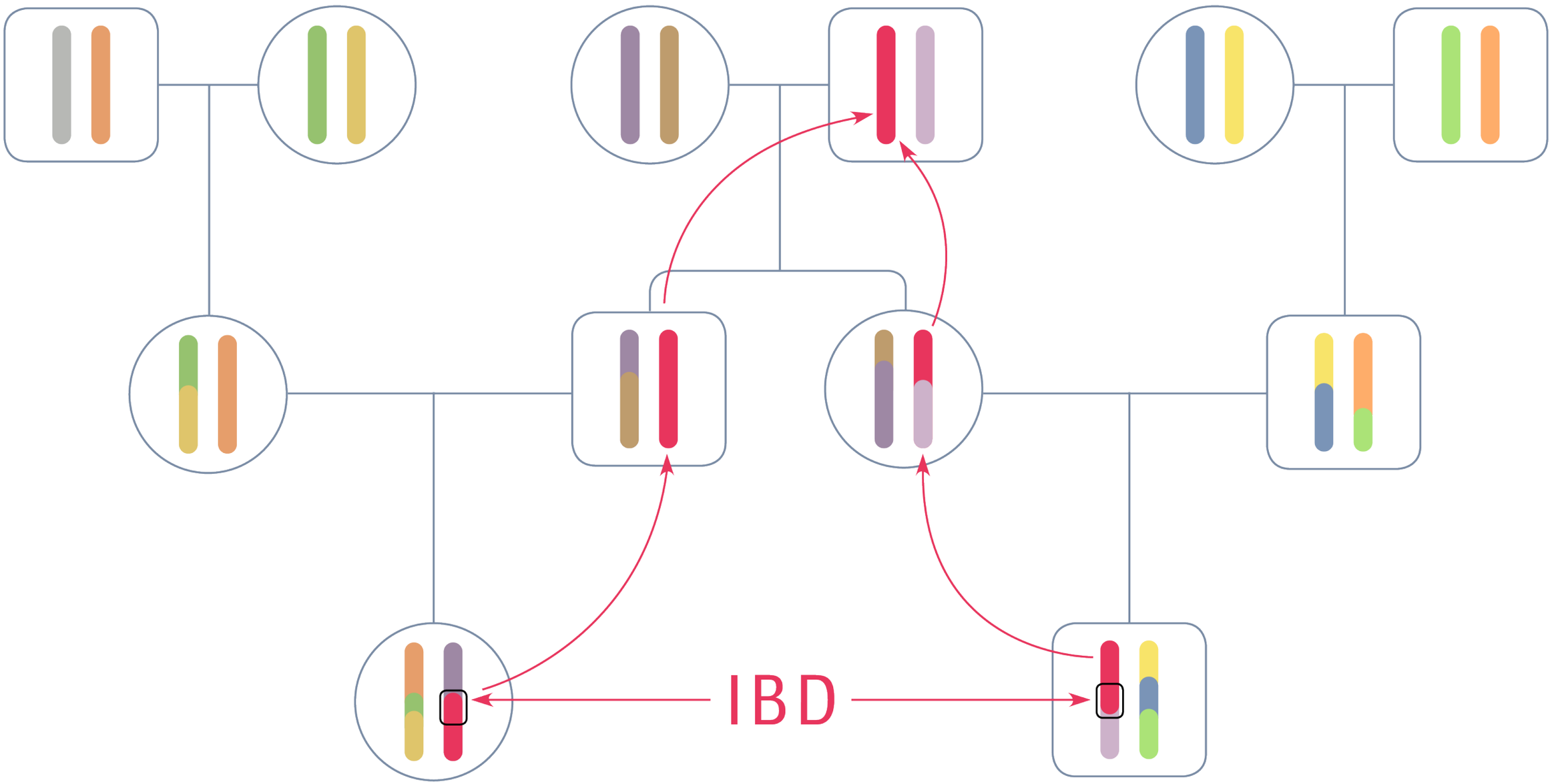
Отцовская линия

60 000
лет назад


AGE=NA
AGE=NA
AGE=NA
AGE=NA
AGE=NA

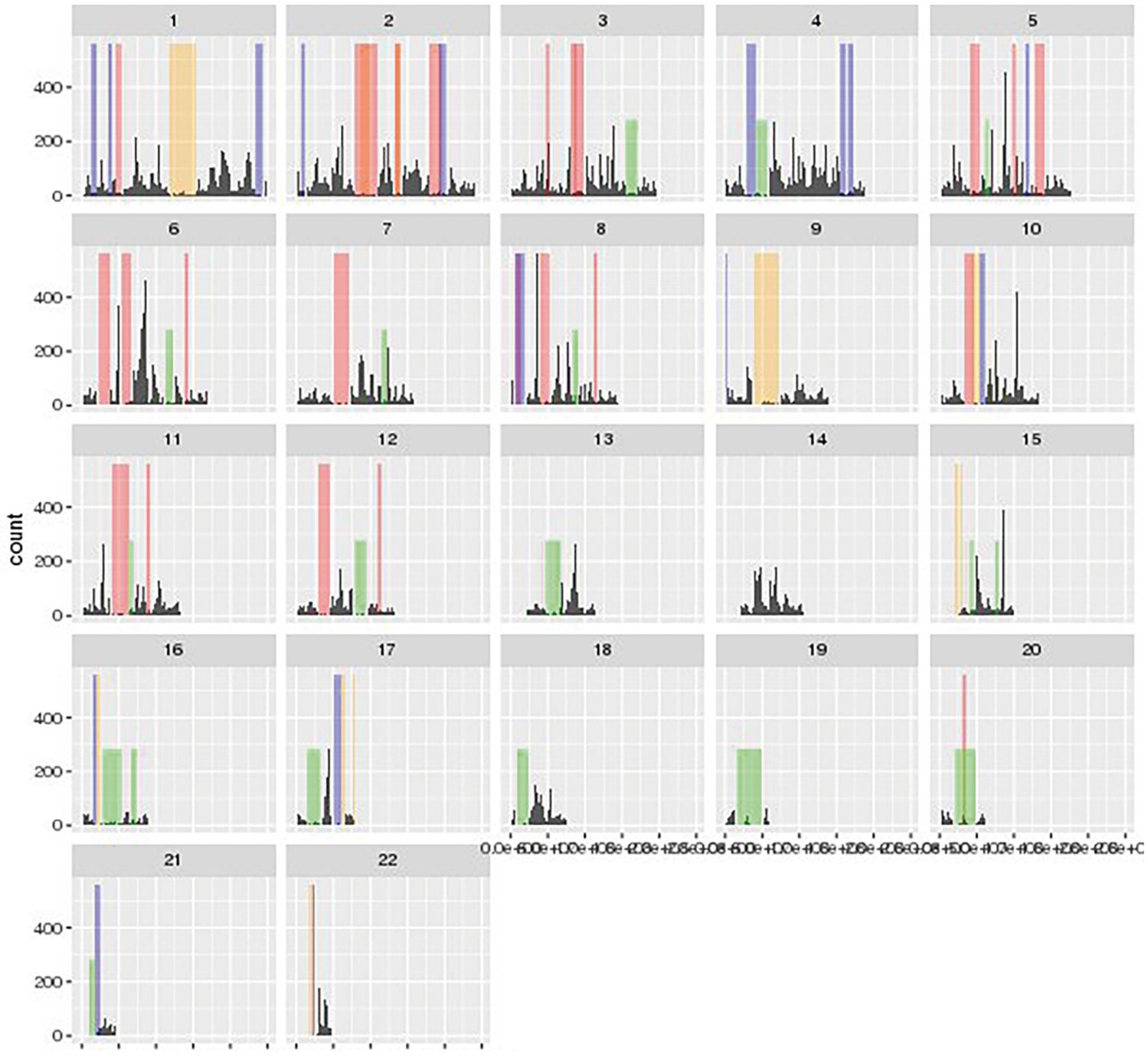

Missing6
AGE=NA
Missing2
AGE=NA
Missing8
AGE=NA

Сегменты от общего предка

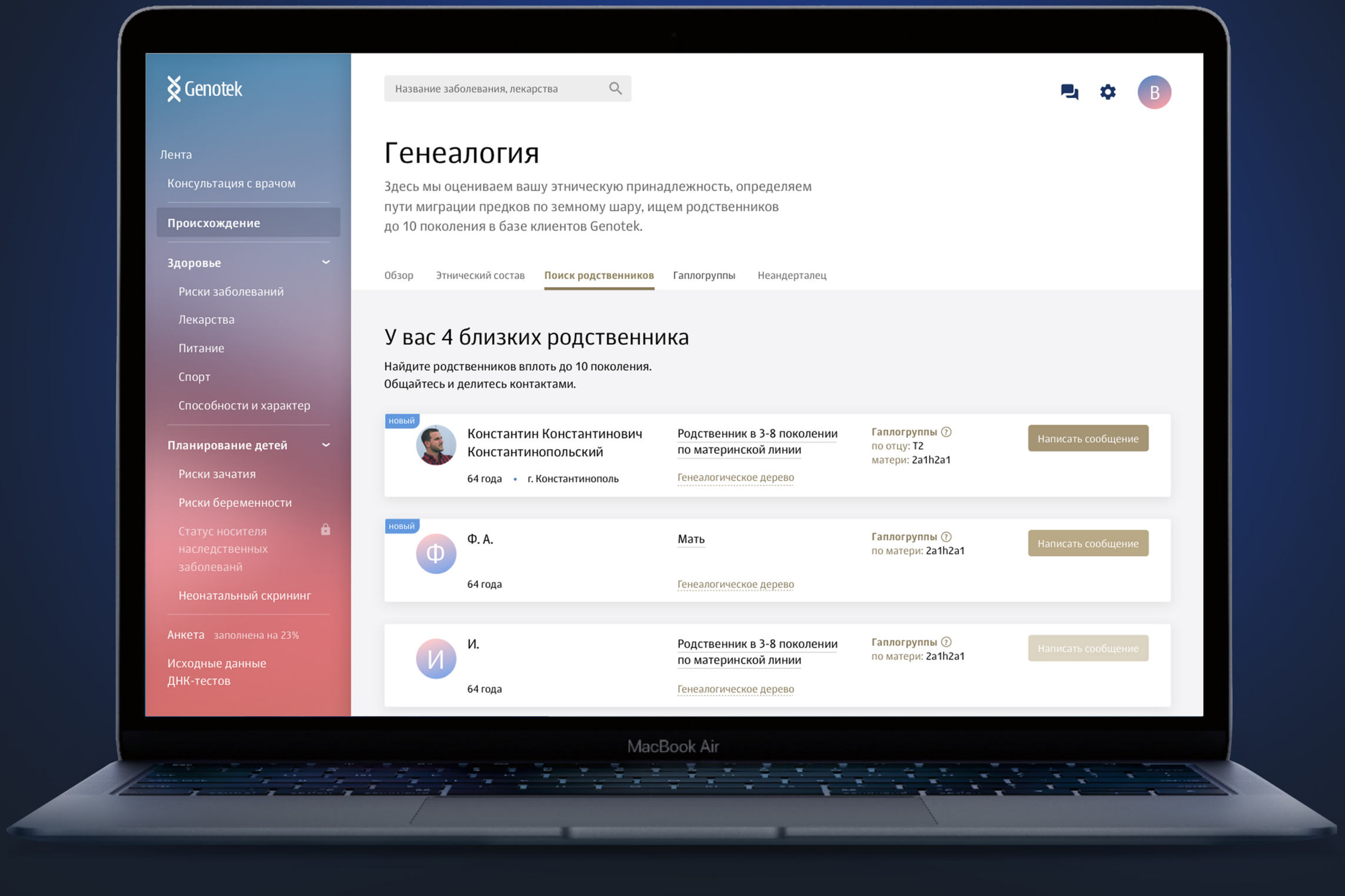


Common regions to be masked

After exteneded external masks +ersa



Результаты поиска родственников

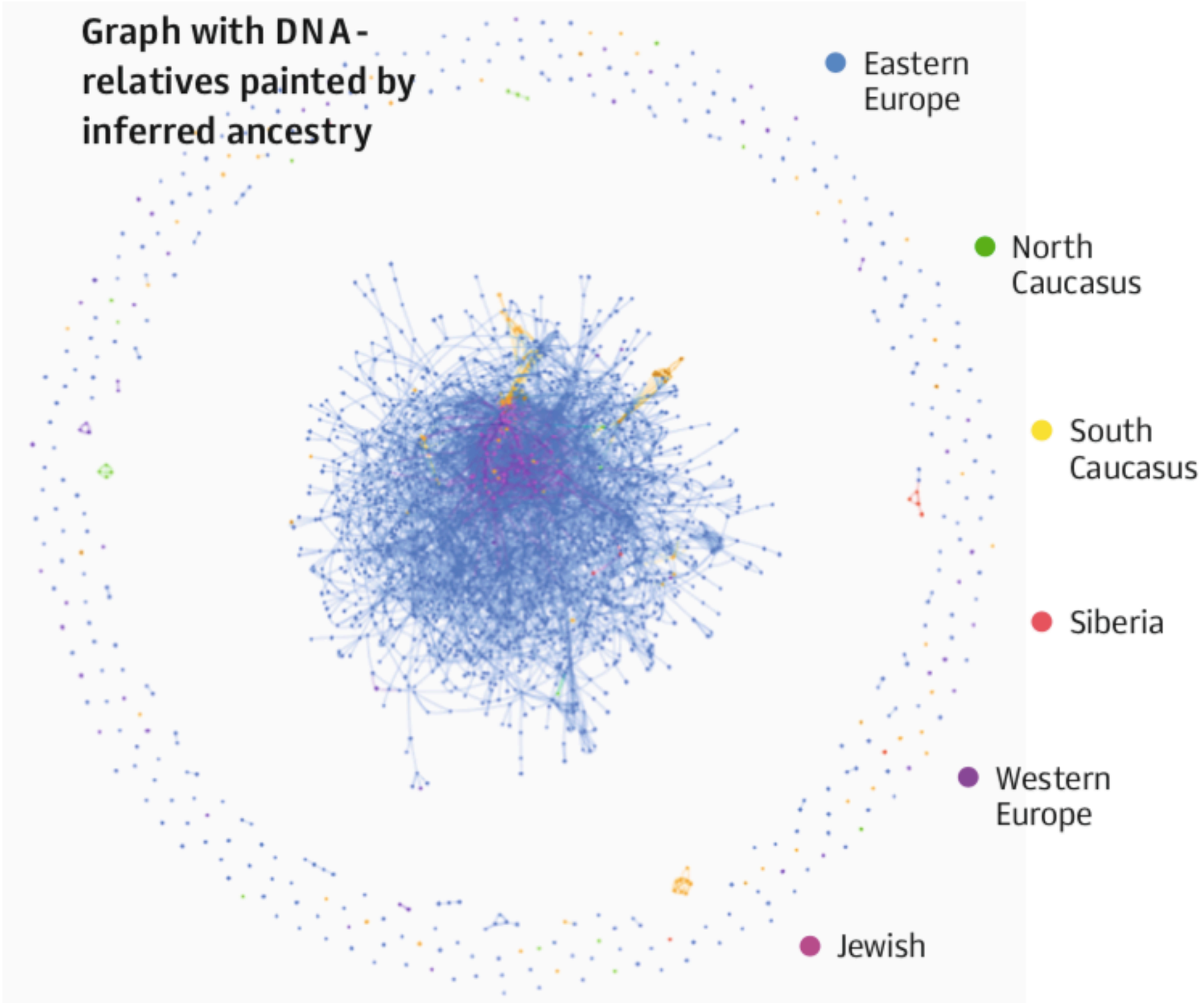


Distant relatives prediction

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


Graph with DNA-relatives painted by inferred ancestry
Eastern Europe
North Caucasus
South Caucasus
Siberia
Western Europe
Jewish

Ашкеназские евреи

Происхождение

Согласно гипотезе популяция ашкеназских 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 ашкеназских евреев привело к падению генетического разнообразия. Вероятно это связано с вспышкой чумы в Европе.

Современные ашкеназские 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 ашкиназских евреев. Многие генетические маркеры присутствуют у них значительно чаще, чем у других народов, и приводят к наследственным заболеваниям.

ген
HEXA
1 из 26 ашкиназских евреев является носителем болезни Тея-Сакса

Болезнь Тея-Сакса
Дети с этим заболеванием рождаются без каких-либо симптомов и развиваются нормально 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

Четвероюродная
сестра

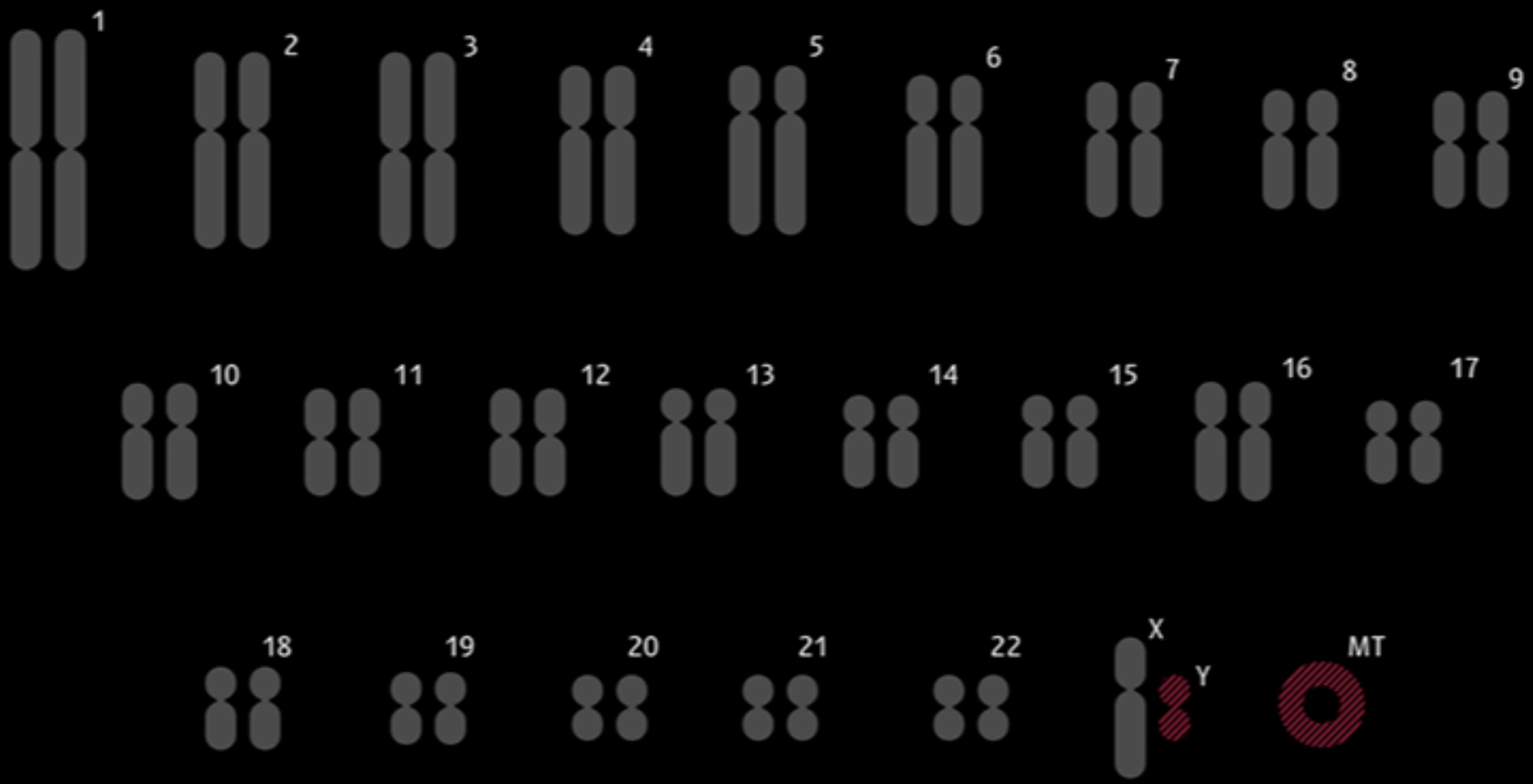
Пятиюродная
сестра

50–200
лет назад

2 000–200 000
лет назад

60 000
лет назад

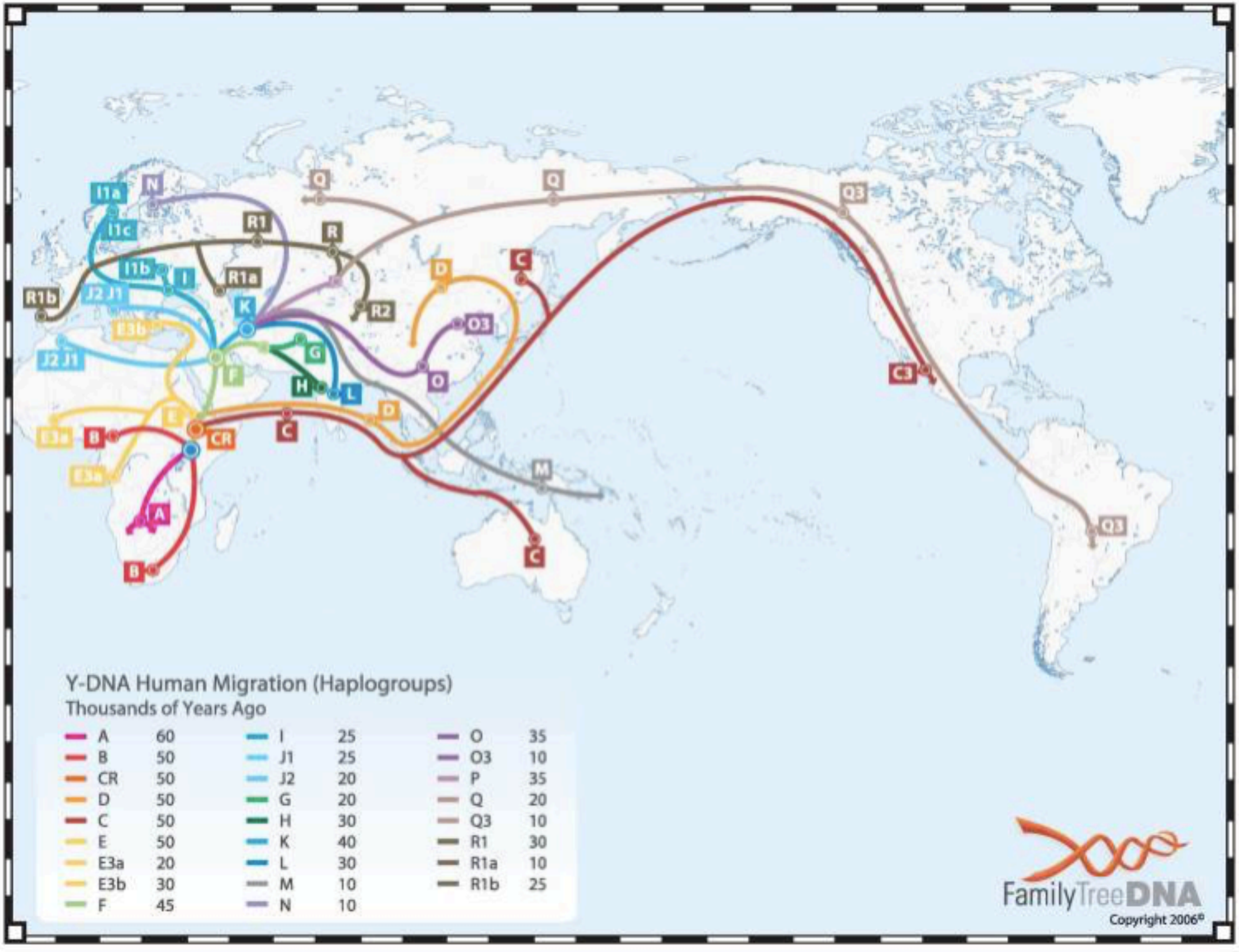
• Отцовская линия



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

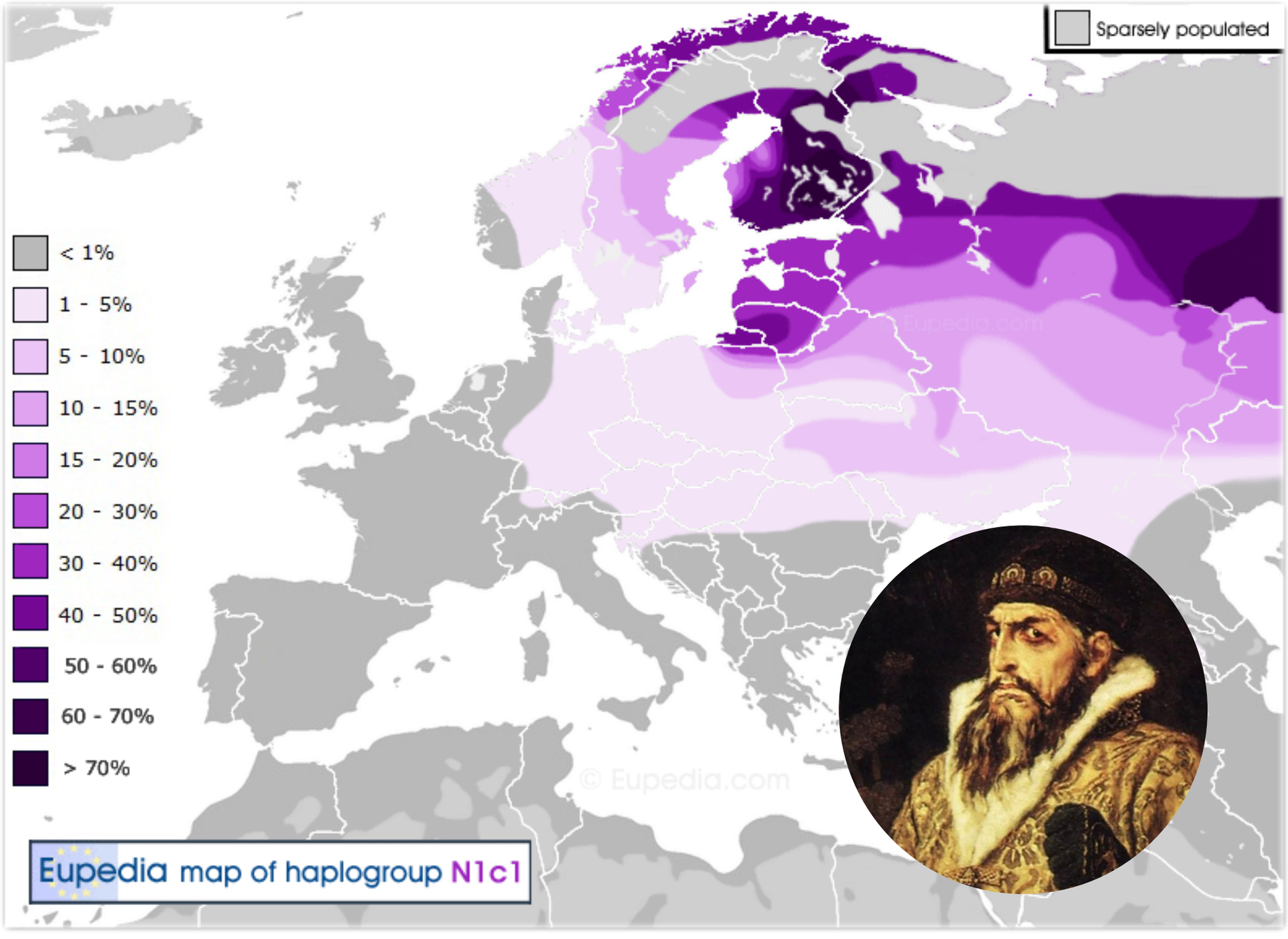
1500
маркеров происхождения по отцовской линии (доступно только для мужчин)

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


I1a
N
I1c
R1
Q
Q
Q3
R
I1b
I
R1a
R1b
J2 J1
K
R2
D
C
O3
E3b
J2 J1
G
F
H
L
O
C3
E
D
E3a
B
CR
C
M
E3a
A
B
C
Q3
Y-DNA Human Migration (Haplogroups)
Thousands of Years Ago
A 60
B 50
CR 50
D 50
C 50
E 50
E3a 20
E3b 30
F 45
I 25
J1 25
J2 20
G 20
H 30
K 40
L 30
M 10
N 10
O 35
O3 10
P 35
Q 20
Q3 10
R1 30
R1a 10
R1b 25
FamilyTreeDNA
Copyright 2006©


yDNA Path to N-L591
N-L1026
N-M2005
N-F1419
N-L591
N-CTS10760
N-M178
N-TAT
N-L729
N-F1206
N-M231
NO-M214
K-F549
K-M526
K-M9
IJK-L15
HIJK-PF3494
GHIJK-F1329
CF-P143
CT-M168
A-PR2921
A-V168
A-V221
BT-M42
Modern
Medieval
Roman
Iron
Bronze
Neolithic
Mesolithic
Paleolithic

Распространенность гаплогруппы N1c1

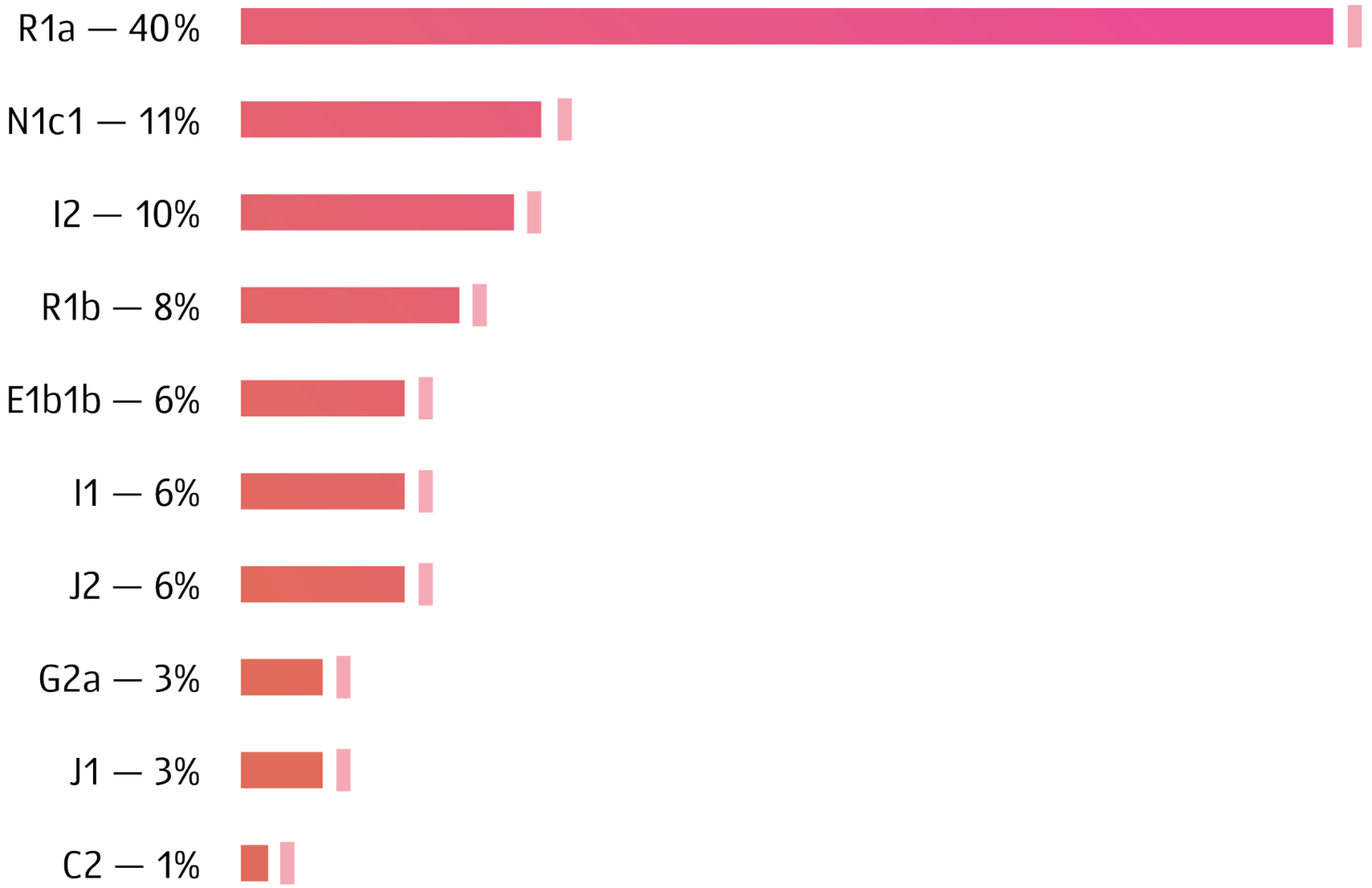


YFull YTree v8.01.00 Посмотреть: scientific new chart new

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

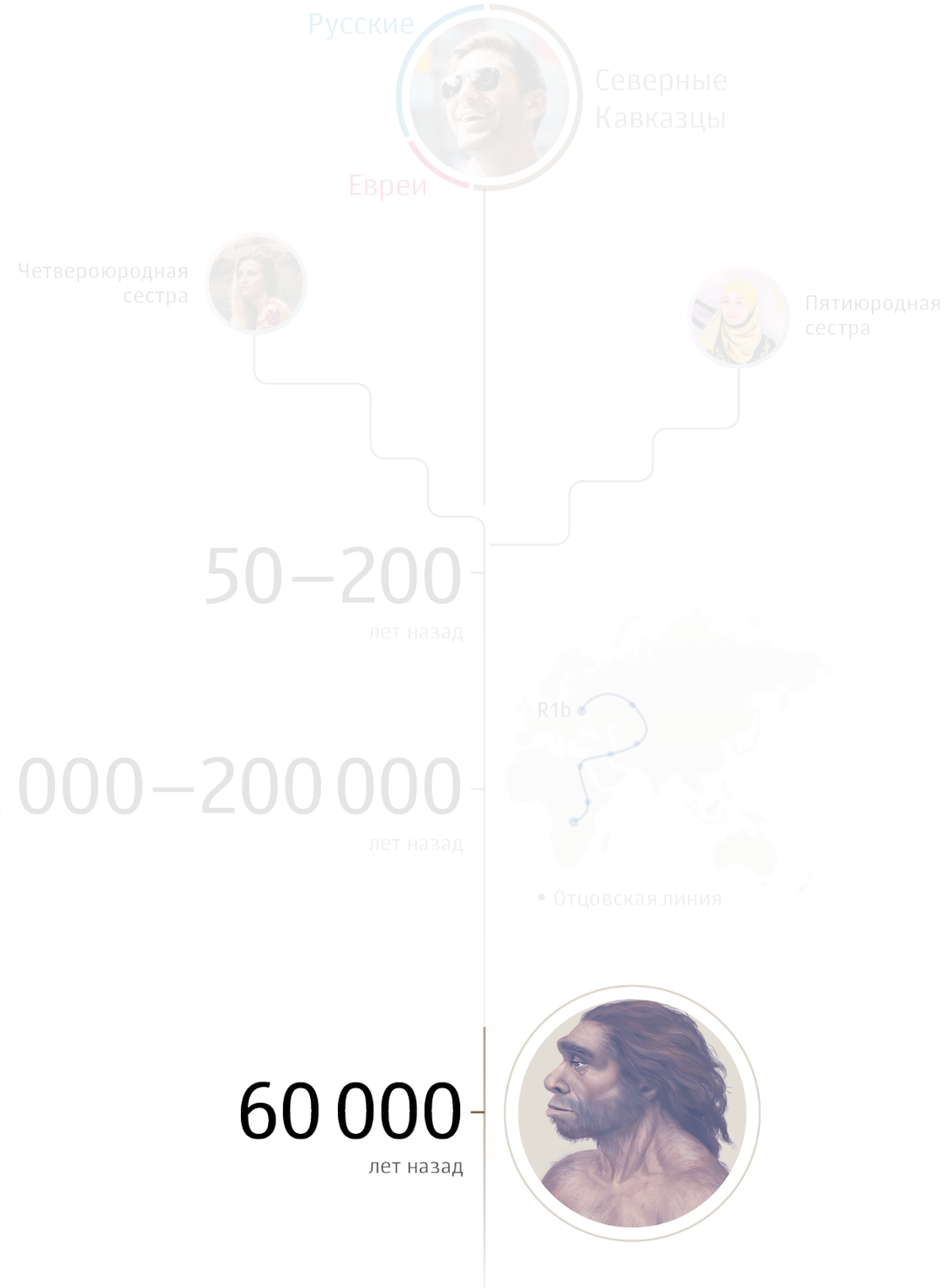
Y chromosome haplogroups



390	M417
207	CTS3402
80	YP417
75	YP569
53	Z92
52	Z282
45	L1029
42	CTS1211
41	Z645
24	Y2910
21	Z93
20	M512

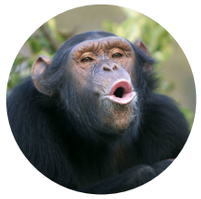
We studied the frequency distribution of mtDNA and Y-chromosome haplogroups. For example, the most frequent haplogroups for Y-chromosome were R1a (40%), I2 (10%), N1c1 (11%). M417 and CTS3402 mutations dominate among R1a subclades. More than 700 unique mtDNA subclades were detected.

Гены неандертальца



Гены неандертальца

Настоящее



Шимпанзе

CGGACATAGG

Неандерталец

TCATGTAGG

Современный человек

TCAAGTTCG

Смешение

54 000–40 000 лет назад

Предки неандертальцев

TCAAGTAGG

430 000 лет назад

750 000–550 000 лет назад

9,3–6,5 млн лет назад

Прошлое

Этнический состав

Русские

Северные Кавказцы

Евреи

Четвероюродная сестра

Пятиюродная сестра

50–200 лет назад

• Отцовская линия

2 000–200 000 лет назад

60 000 лет назад

Thank you for your attention!

https://t.me/alexrakitko
rakitko@gmail.com

Genotek

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

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