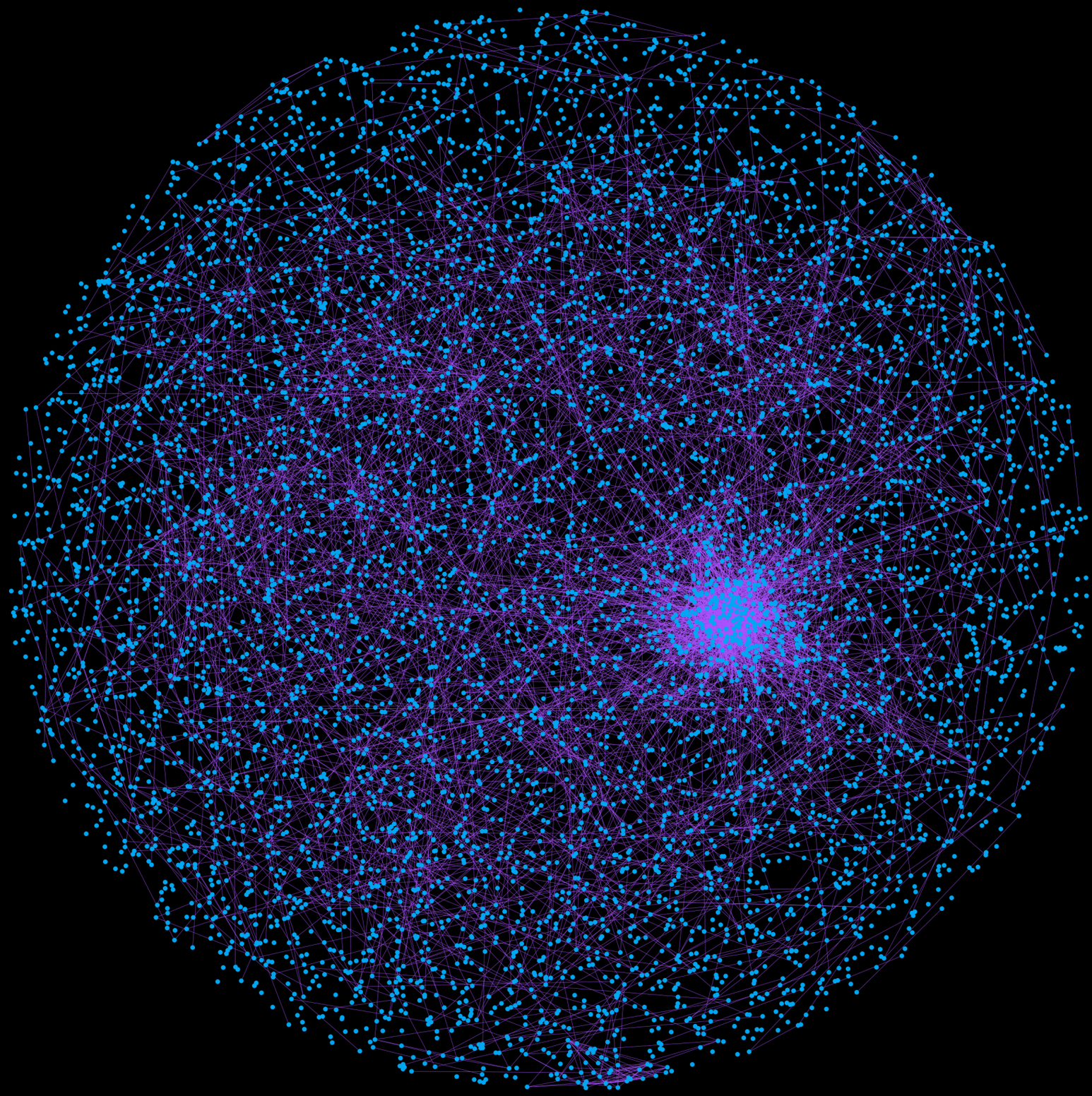
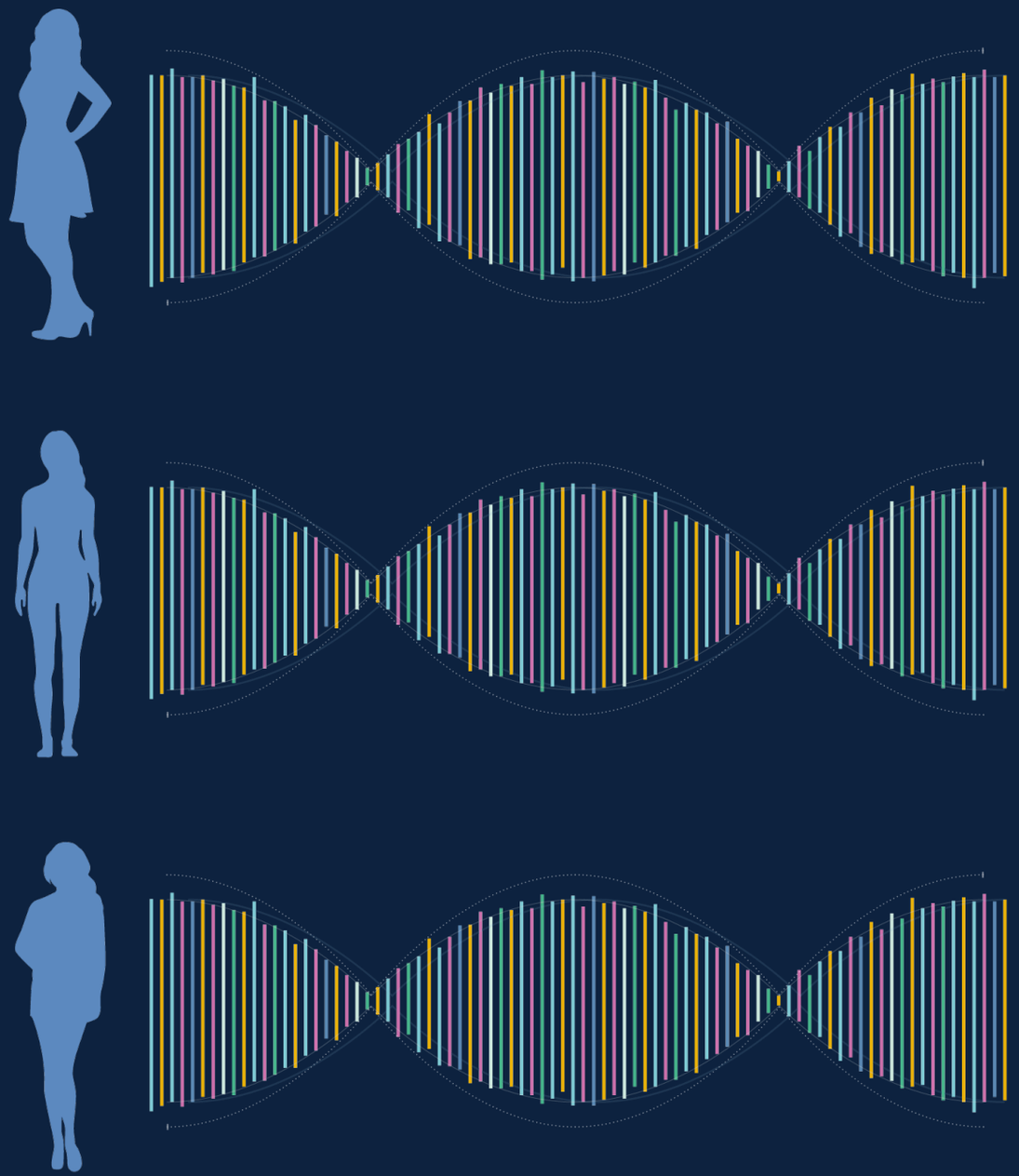
Machine Learning in population genetics

Alexander Rakitko, MSU

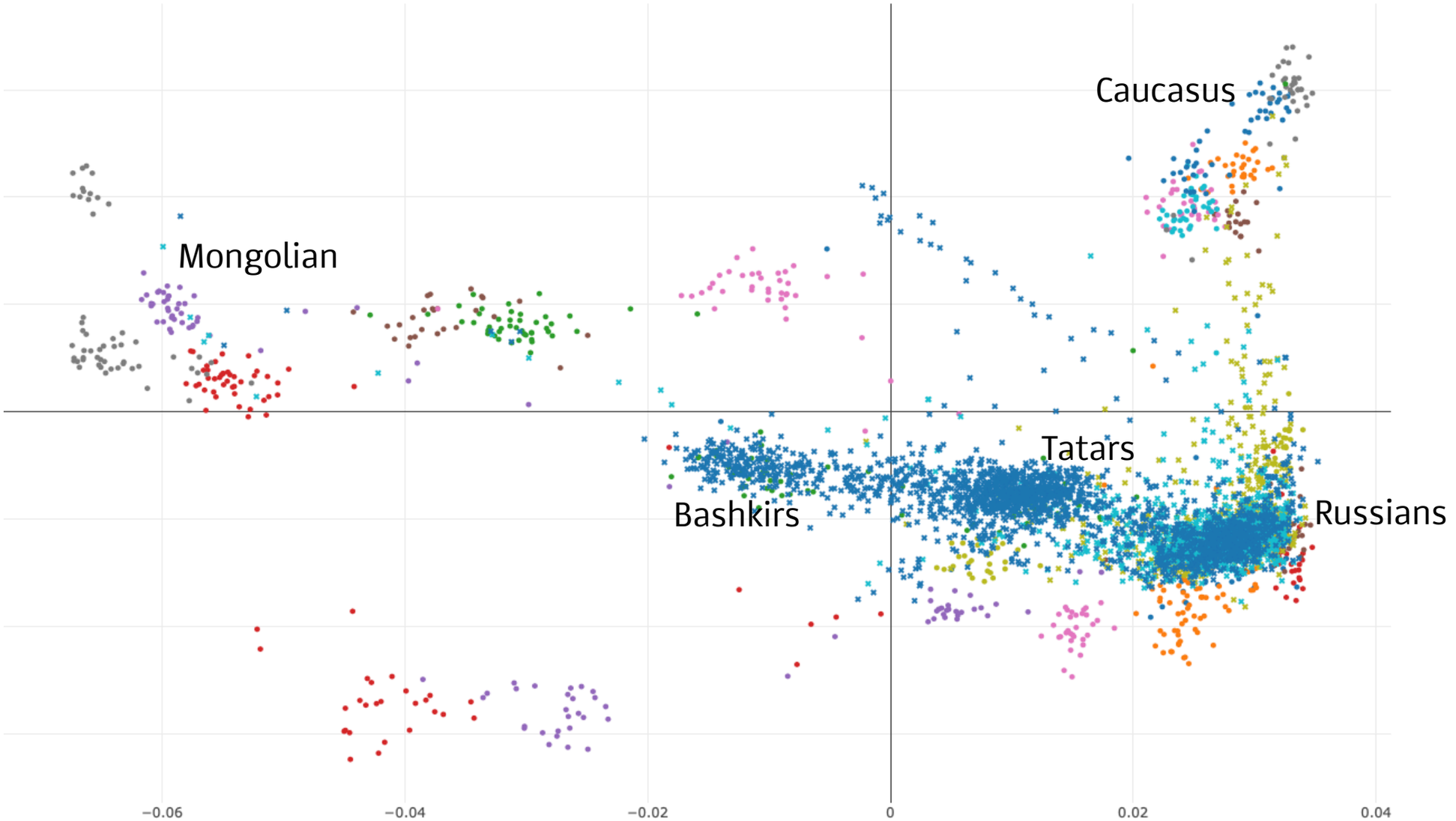
Network of genetic relatives



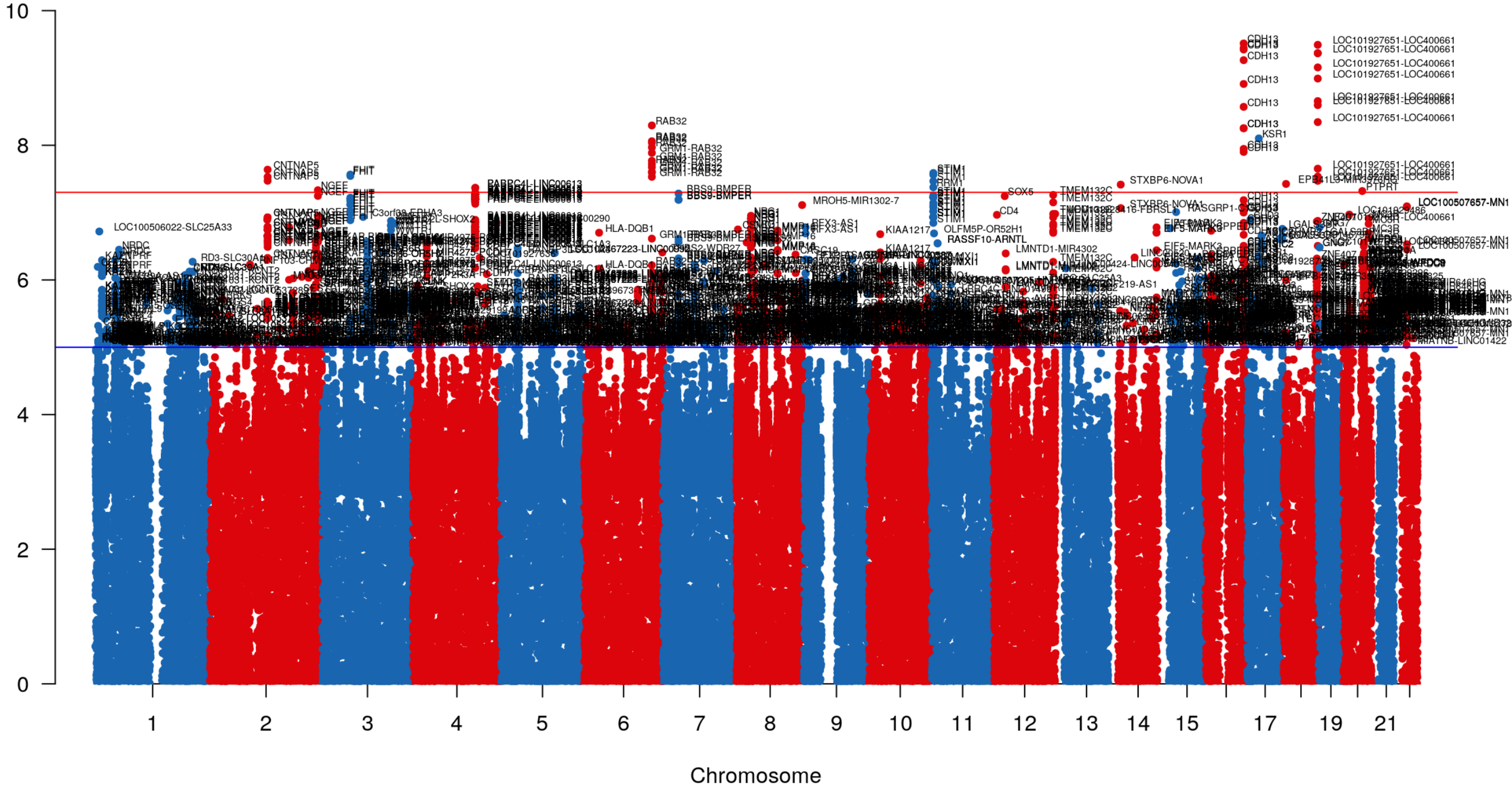
We work with SNPs

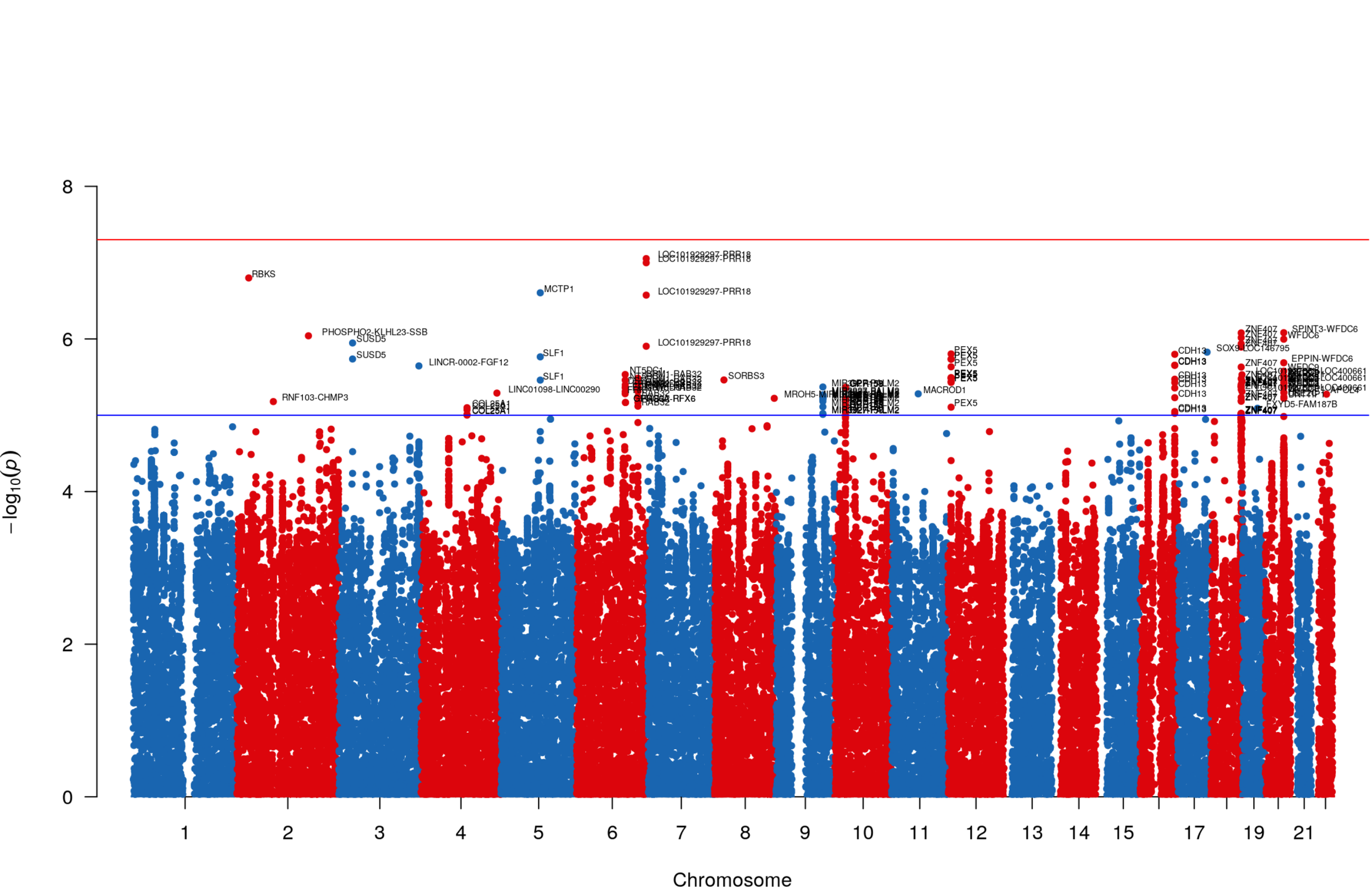


GWAS of schizophrenia with Russian National Consortium for Psychiatric Genetics (RNCPG)

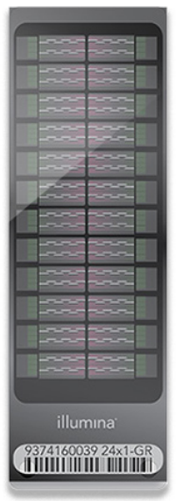


Triska, Petr, et al. "Between Lake Baikal and the Baltic Sea: genomic history of the gateway to Europe." BMC genetics 18.1 (2017): 110.

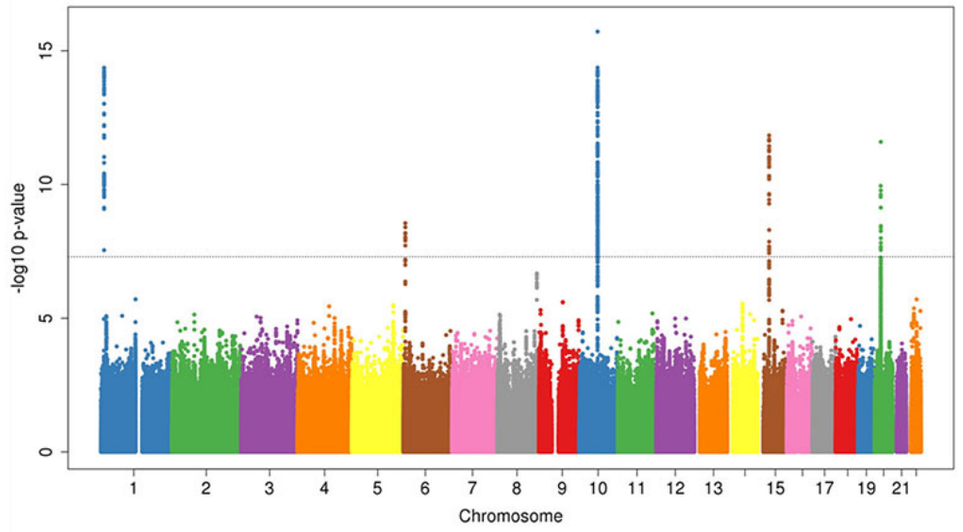

10
8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
Chromosome
RAB32
CNTNAP5
FHIT
NGEE
BBS9-BMPER
MROH5-MIR1302-7
HLA-DQB1
LOC100506022-SLC25A33
KIAA1217
OLFM5P-OR52H1
RASSF10-ARNTL
SOX5
CD4
TMEM132C
STXBP6-NOVA1
LMNTD1-MIR4302
CDH13
KSR1
LOC101927651-LOC400661
EPB41L3-MIR3922HG
PTPRT
LOC100507657-MN1


8
6
4
2
0
$-\log_{10}(p)$
RBKS
PHOSPHO2-KLHL23-SSB
SUSD5
SUSD5
LINCR-0002-FGF12
RNF103-CHMP3
MCTP1
SLF1
SLF1
LINC01098-LINC00290
COL25A1
COL25A1
LOC101929297-PRR18
LOC101929297-PRR18
LOC101929297-PRR18
SORBS3
GPR82-RFX6
MACROD1
PEX5
PEX5
PEX5
PEX5
PEX5
CDH13
CDH13
SOX9-LOC146795
ZNF407
SPINT3-WFDC6
WFDC6
EPPIN-WFDC6
FXYD5-FAM187B
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
19
21
Chromosome

Polygenic Risk Scores



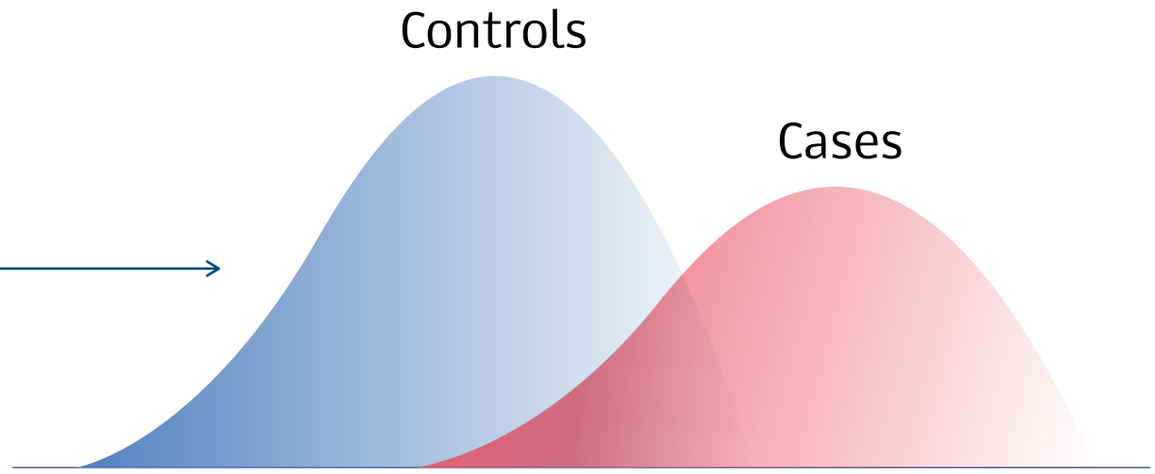
Microarray Genotyping



GWAS analysis

$$PRS = \beta_1 X_1 + \ldots + \beta_n X_n$$

Genome-Wide Polygenic Risk Score



PRS for coronary artery disease (CAD)

Polygenic Risk Score

The PRS previously described in Khera et al. 2018 was used as a predictive model. The PRS was based on analysis of 6,630,150 SNPs.

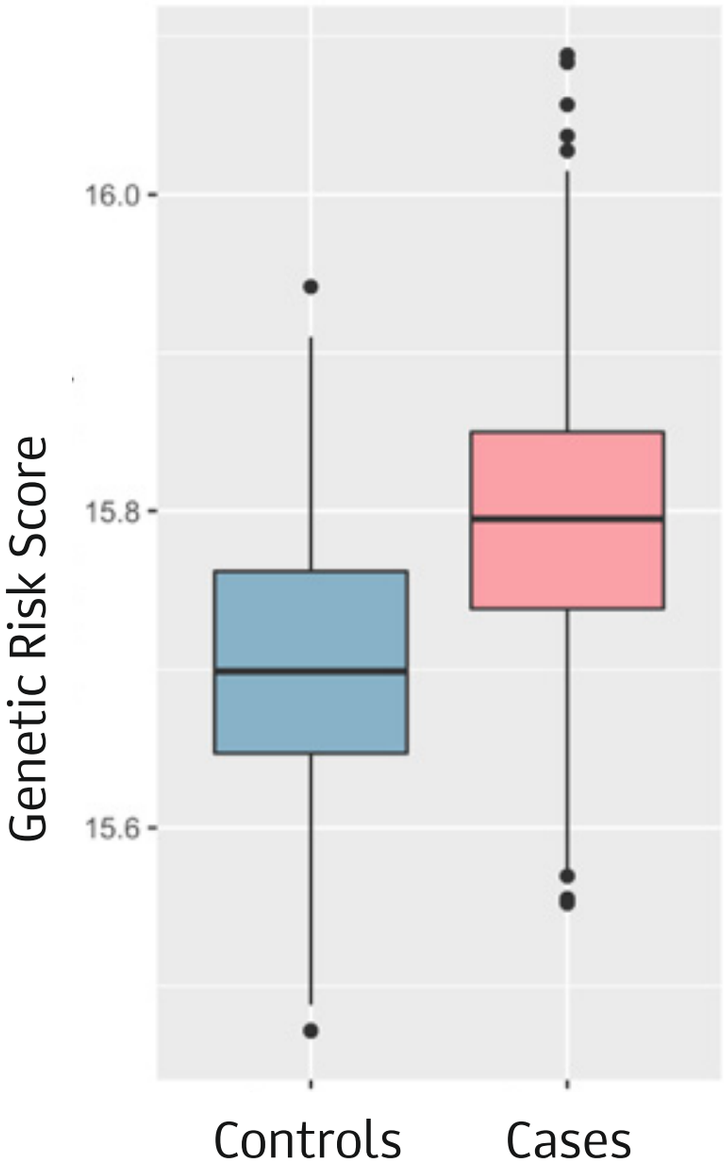
Test dataset

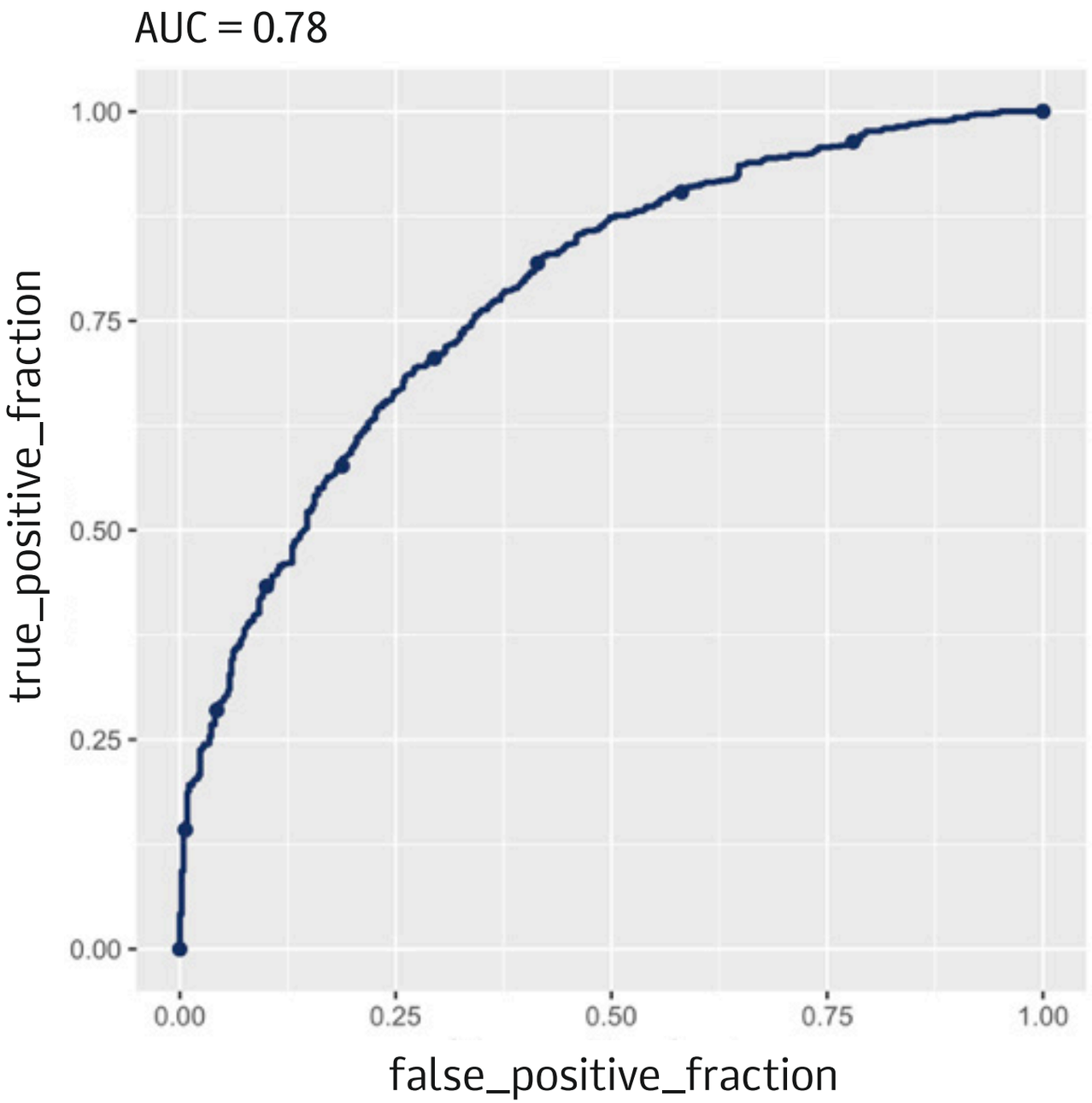
A PennCath study subset was used as a test sample (Reilly et al. 2011). The total sample size was 1,401 individuals, including 933 cases with CAD and 468 controls. The genotyping array contains 500,000 SNPs. Age, sex, triglyceride, LDL and HDL levels were available for every individual in addition to genetic data.

Khera, Amit V., et al. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations." Nature genetics 50.9 (2018): 1219.

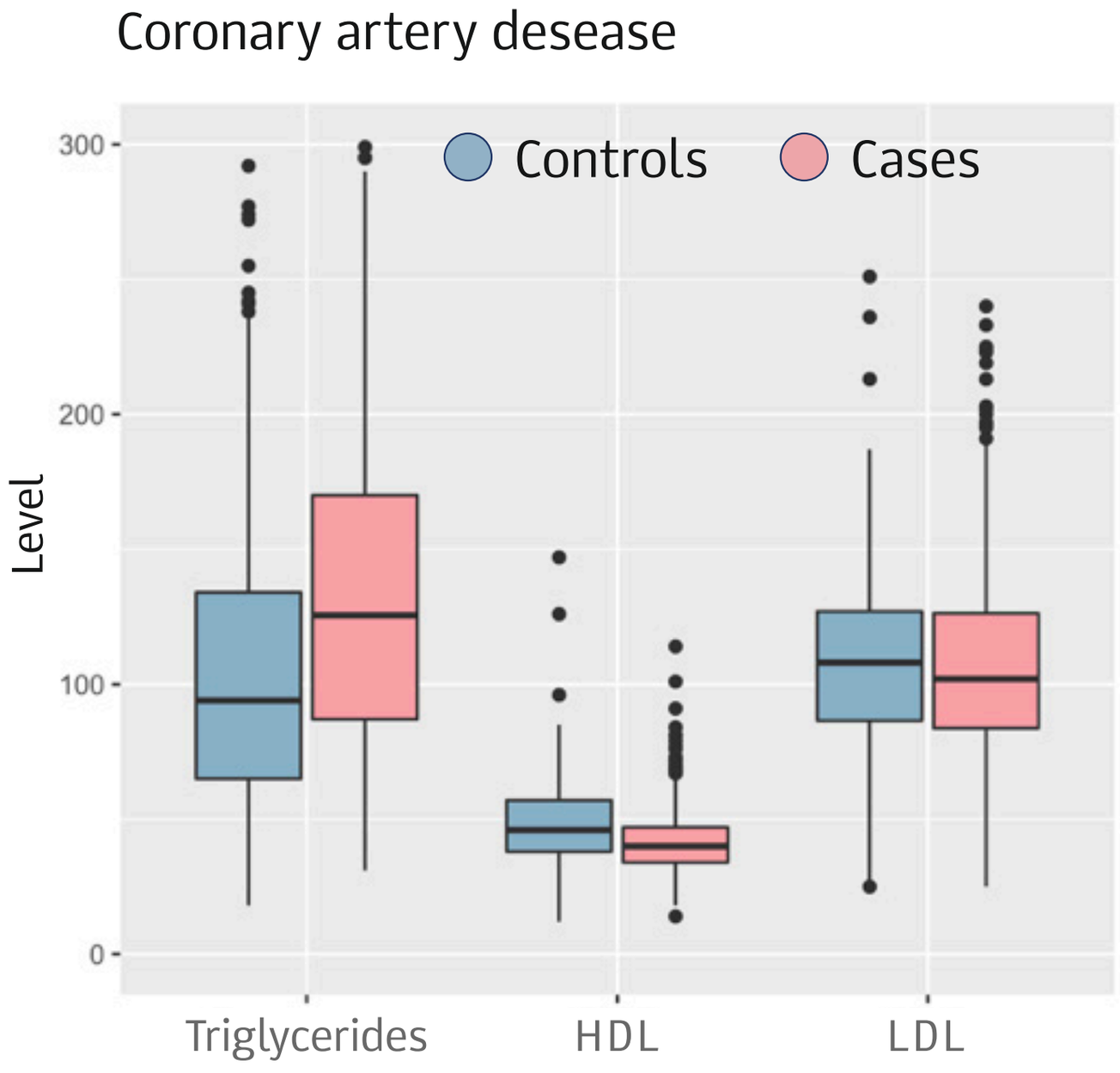
Reilly, Muredach P., et al. "Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies." The Lancet 377.9763 (2011): 383-392.

Validation of Polygenic Risk Score





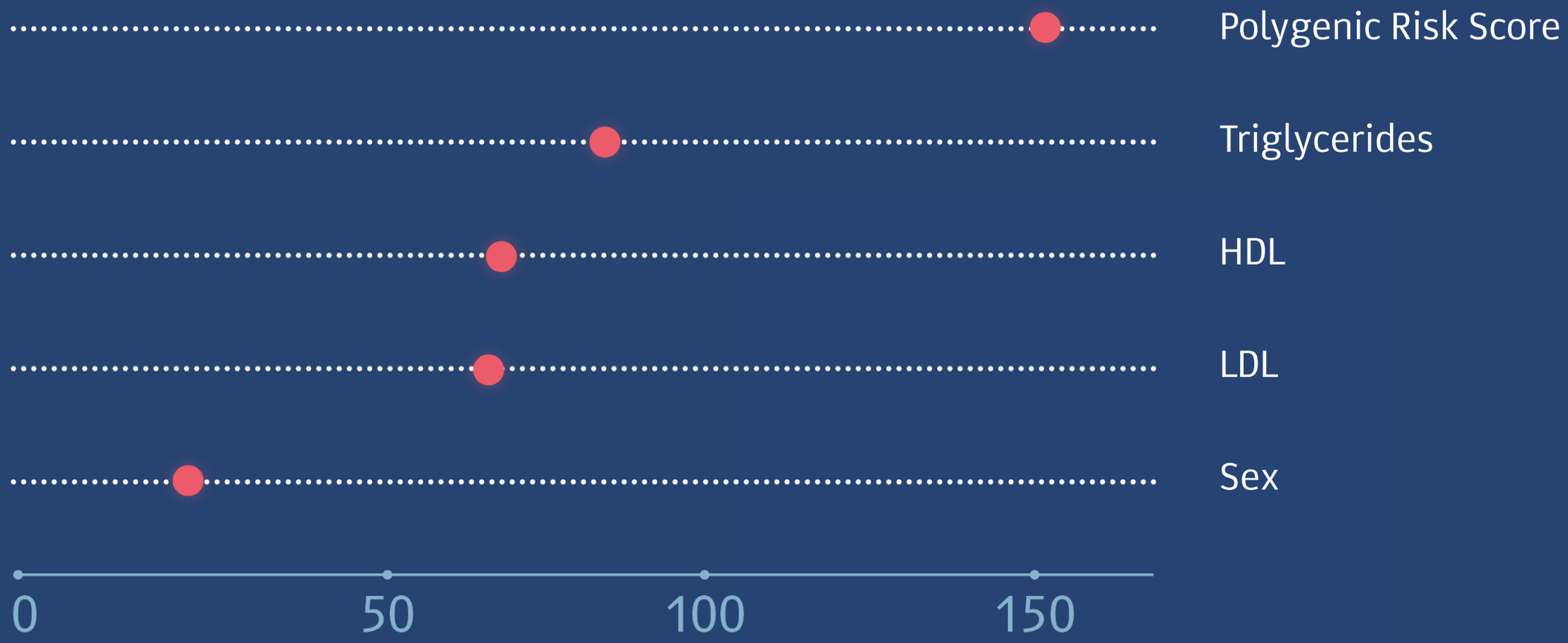
Non-genetic risk factors



Comparison of predictive models

Model	Non-genetic factors	Genetic + non-genetic factors
Logistic regression	69,6 % (3%)	75,3 % (3,5%)
Random forests	65,3% (3%)	74,5 % (2%)
Gradient boosting	67,2 % (2%)	**75,8 %** (3%)

Feature importance (RF model)



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.

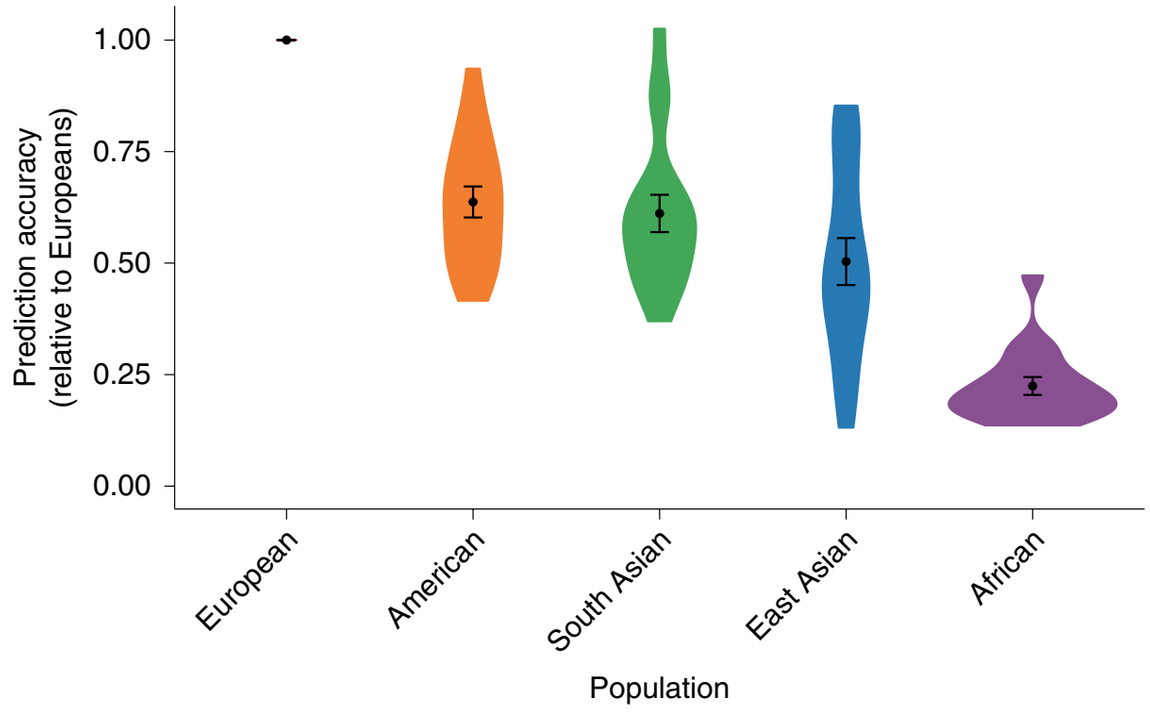
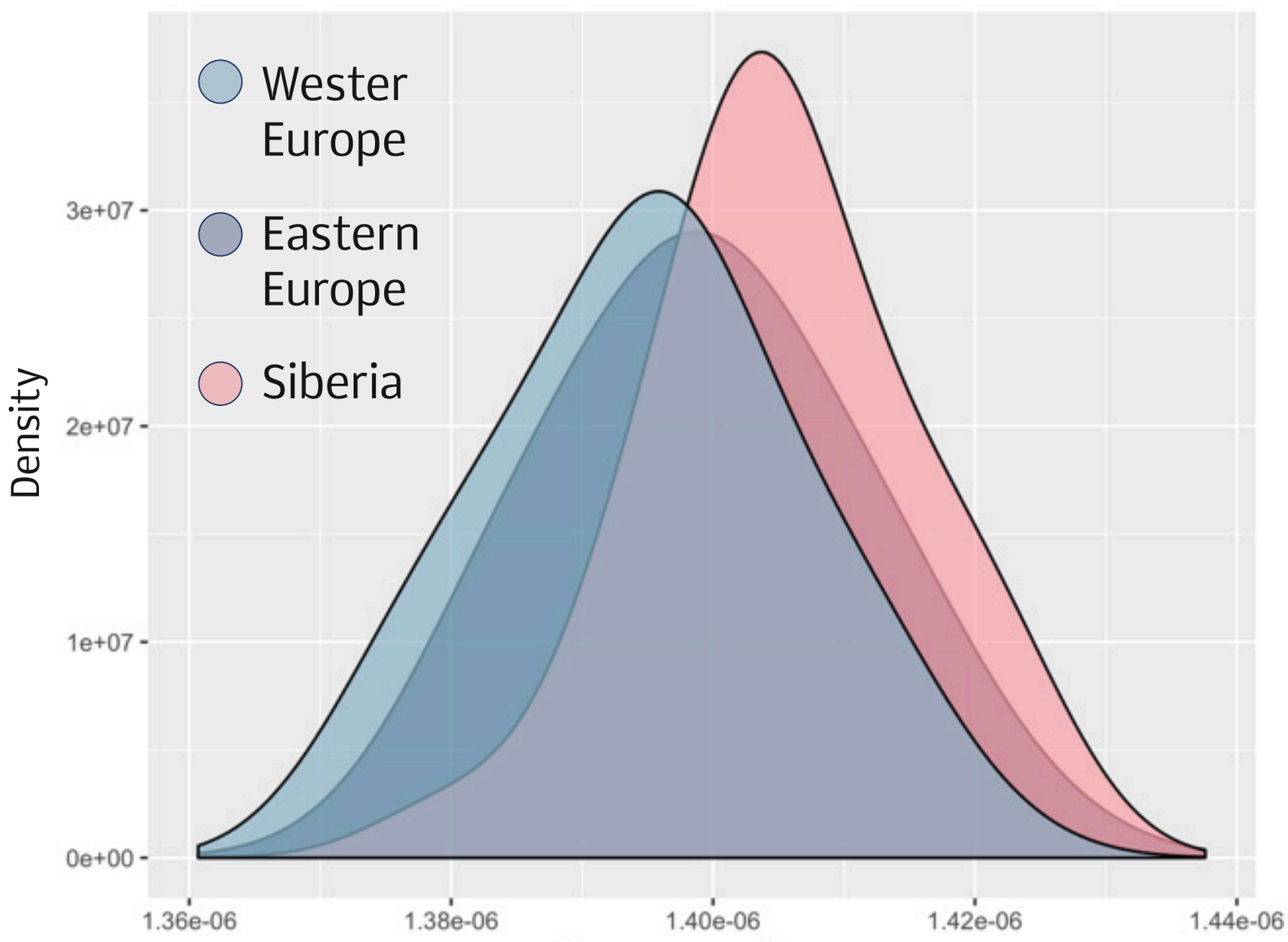
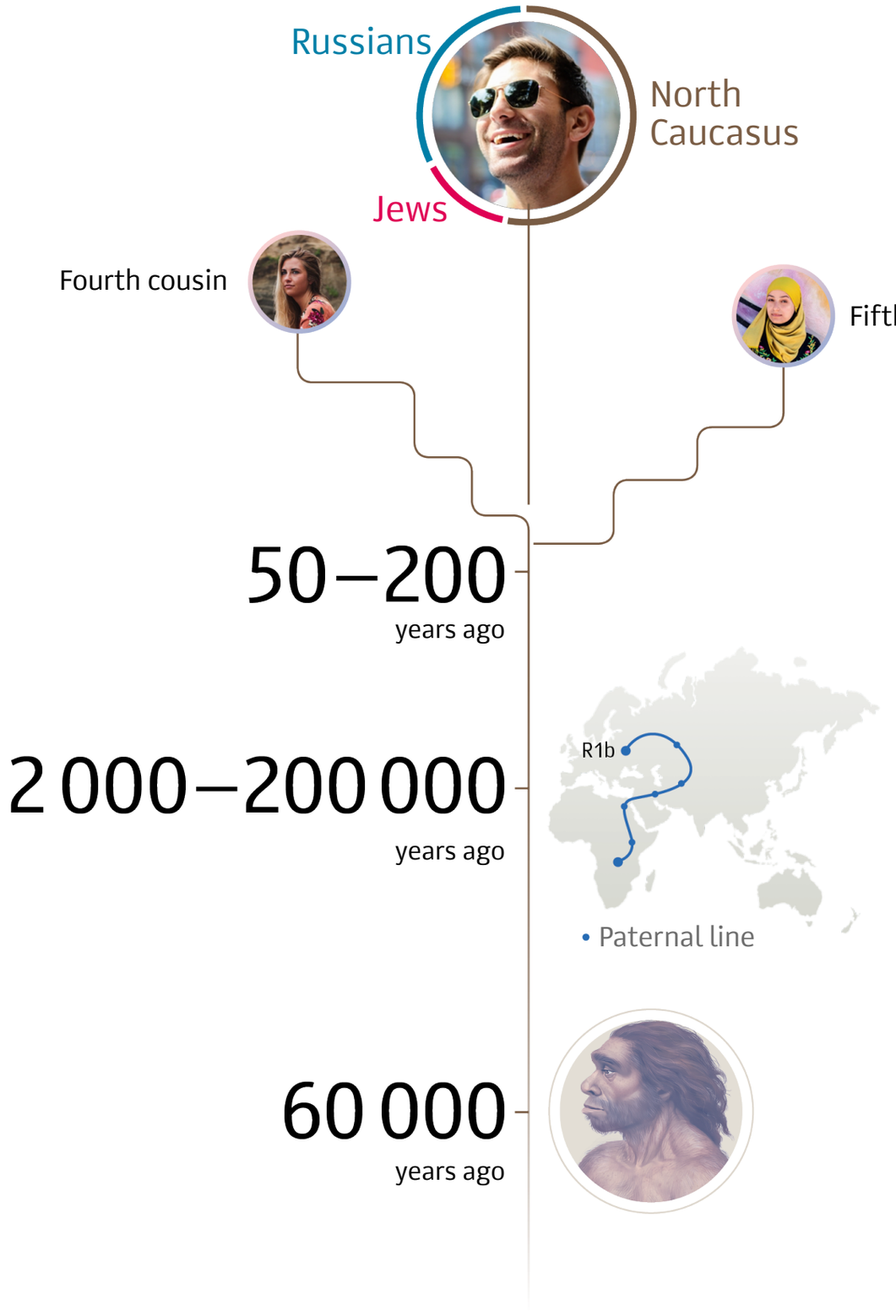


Fig. 3 | Prediction accuracy relative to European-ancestry individuals across 17 quantitative traits and 5 continental populations in the UKBB. All phenotypes shown here are quantitative anthropometric and blood-panel traits, as described in Supplementary Table 6, which includes discovery-cohort sample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

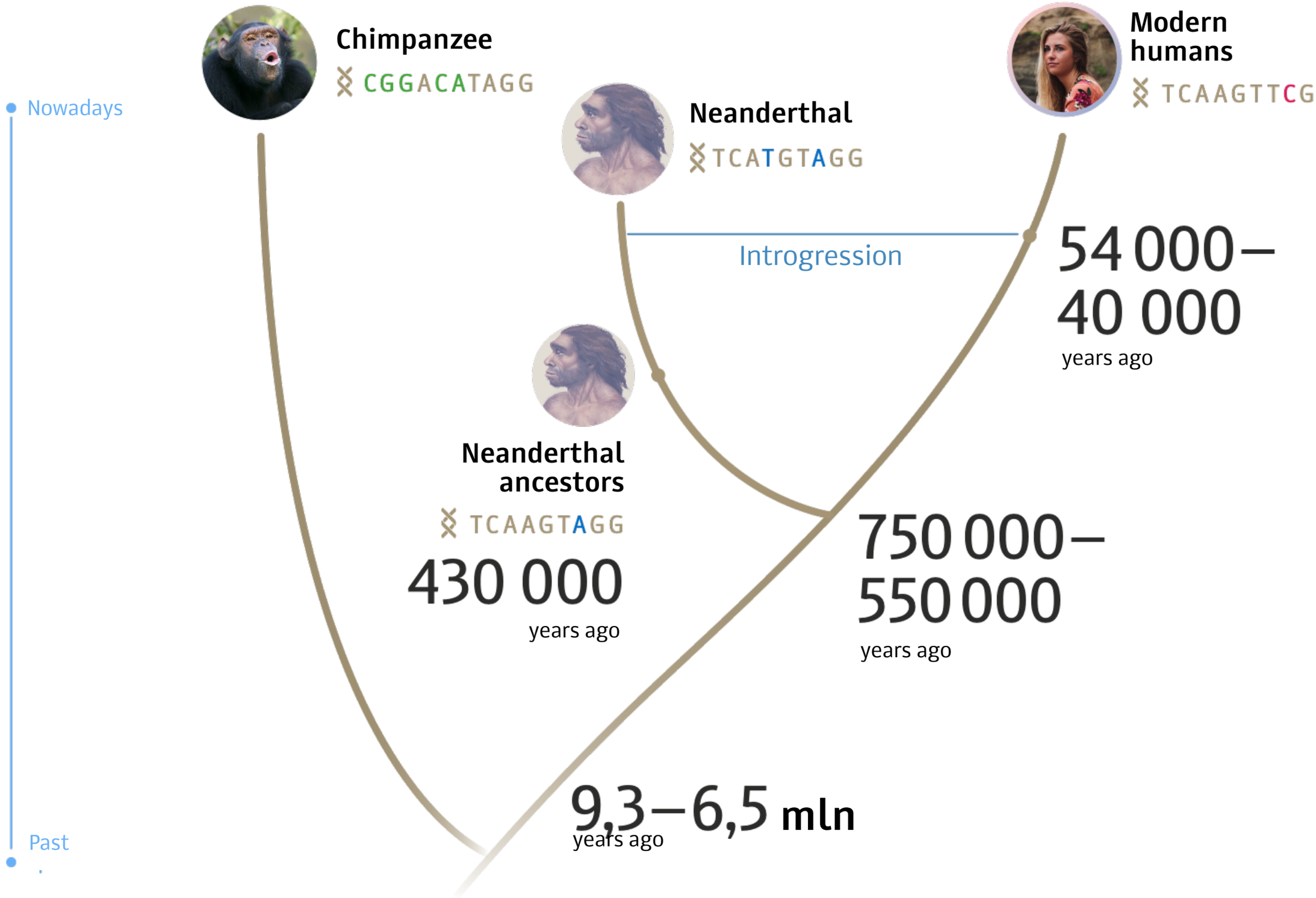
Population differences



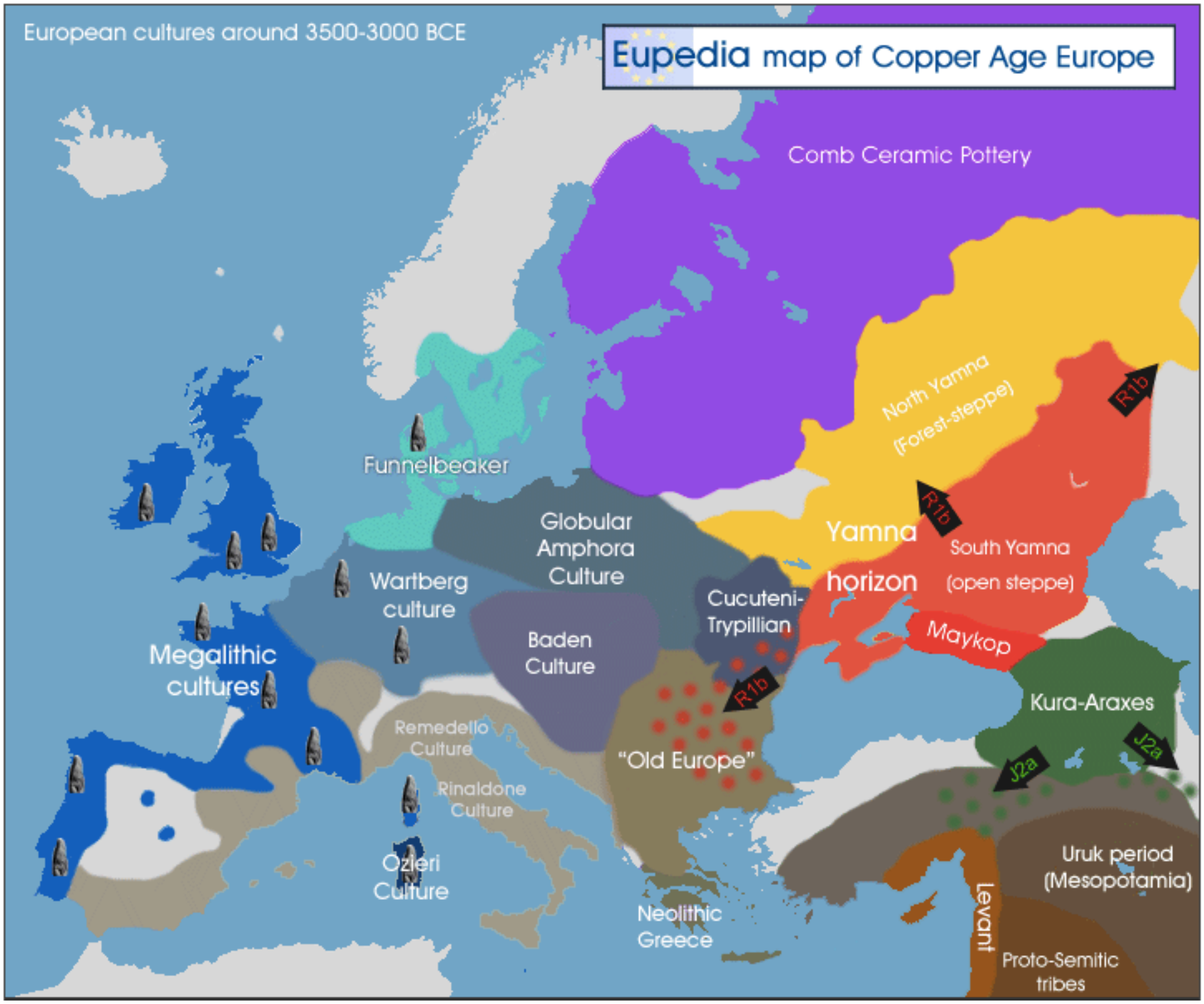
The goal of hackathon: discover your ancestry



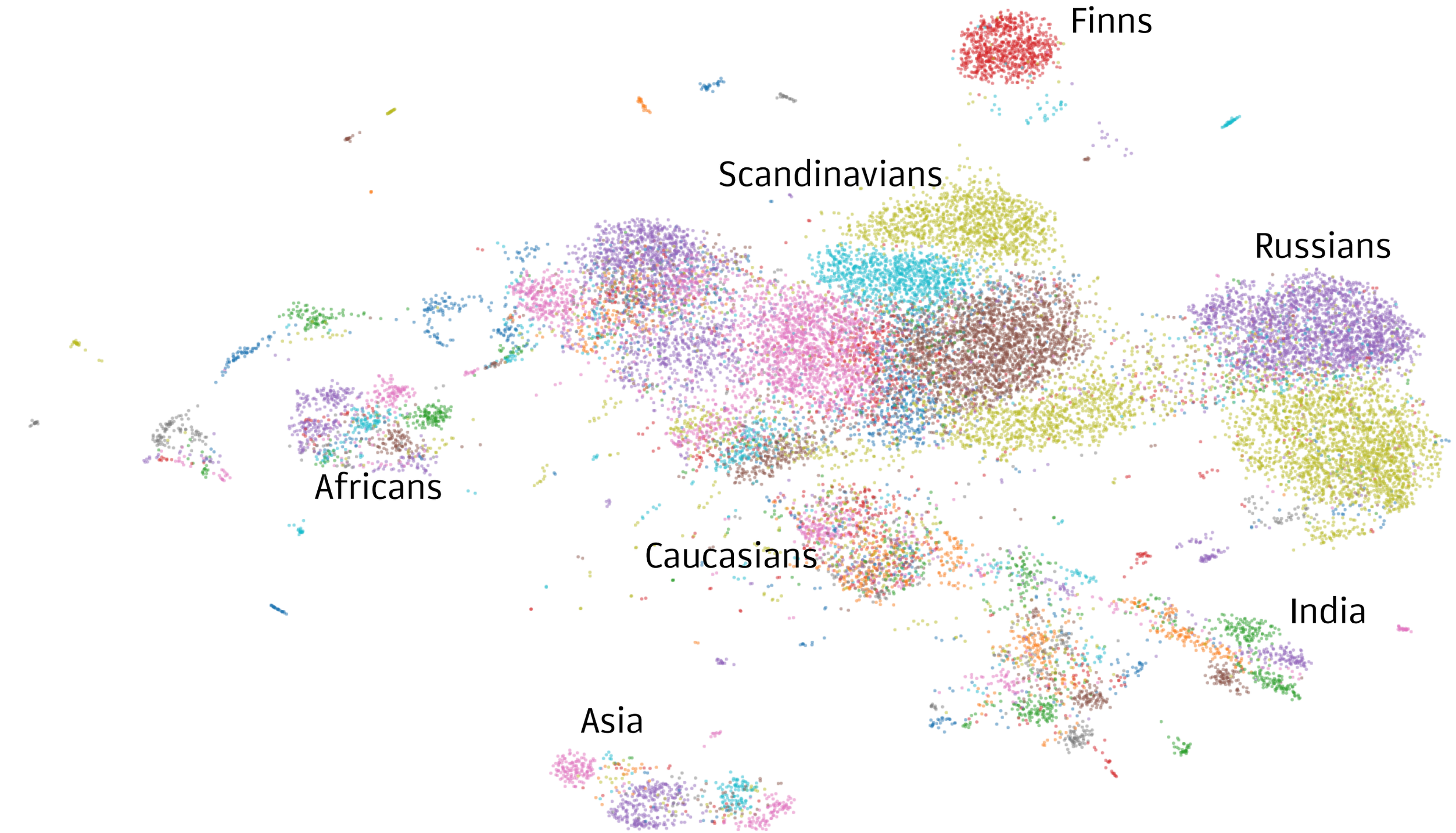
Find population with the highest Neandethal introgression



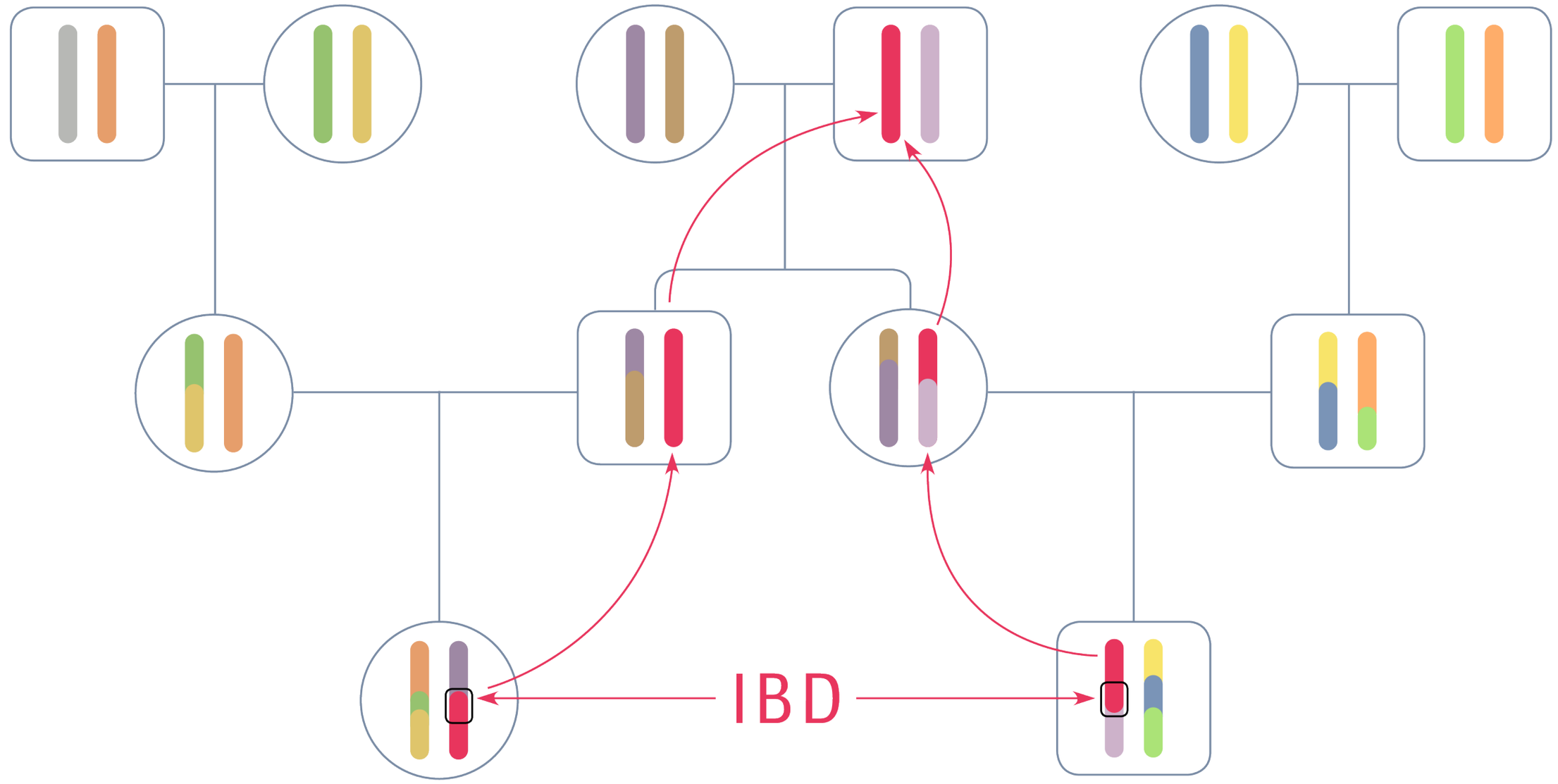
Proportion of Yamna culture in modern Polish genomes



Plot the genetic map of Europe



Paint chromosomes into modern populations



Hackathon summary

Algorithms

- Neanderthal genes (phylogenetic metrics)
- Ancient cultures (Max Likelihood Est)
- Genetic map of Europe (PCA, tSNE, UMAP)
- Find populations (k-means, HDBSCAN)
- Paint chromosomes (SVM / k-NN / NN / RF + HMM)

Dataset

- 2K+ ancient genomes
- 5K+ modern genomes

Languages

- R
- Python
- Bash