# The entangled biobank: On the topology of high-dimensional human genetic data

Alexandr Diaz-Papkovich

Quantitative Life Sciences, McGill University

Montréal, Québec, Canada

August 2023

A thesis submitted to McGill University in partial

fulfillment of the requirements of the degree of

Doctor of Philosophy

# Contents

**Abstract** i

**Résumé** ii

**Acknowledgements** iii

**Contribution to original knowledge** v

**Contributions of authors** vi

**List of abbreviations** viii

**List of figures** xii

**List of tables** xiii

## I Introduction and literature review 1

**1 Overview** 3

1.1 Genetic diversity . . . 4

1.1.1 Population structure . . . 6

1.1.2 Genome-wide association studies and polygenic scores . . . 9

1.2 Biobanks . . . 11

1.3 Exploratory data analysis . . . 12

1.3.1 Dimensionality reduction . . . 14

1.3.2 Topological data analysis . . . 15

1.3.3 t-distributed stochastic neighbour embedding . . . 16

1.3.4 Uniform manifold approximation and projection . . . 17

1.3.5 Clustering . . . 19

1.4 Rationale and objective of research . . . 25

## II Original contributions to knowledge 27

**2 UMAP reveals cryptic population structure and phenotype heterogeneity** 29

2.0 Preface . . . 29

2.1 Abstract . . . 30

2.2 Author summary . . . 31

2.3 Introduction . . . 31

2.4 Results . . . 36
2.4.1 Fine-scale visualization of the 1KGP dataset . . . 36
2.4.2 Admixed individuals fall along a genetic continuum . . . 38
2.4.3 Regional patterns in the Hispanic subpopulation . . . 38
2.4.4 Population structure in the UKBB reflects local and global genetic variation 42
2.4.5 Identifying patterns in phenotype variation related to genetic population structure . . . 46
2.4.6 Comparing t-SNE and UMAP . . . 49
2.5 Discussion . . . 50
2.5.1 Caveats . . . 51
2.6 Materials and methods . . . 53
2.6.1 Ethics statement . . . 57
2.7 Acknowledgments . . . 57
2.8 Supporting information . . . 58
References . . . 123

**3 A review of UMAP in population genetics 126**
3.0 Preface . . . 126
3.1 Abstract . . . 127
3.2 Introduction . . . 127
3.3 Visualizing genomic cohorts . . . 130
3.4 Supporting analyses: What do I do with a UMAP projection? . . . 135
3.5 Discussion . . . 136
3.6 Conclusion . . . 140
3.7 Materials and methods . . . 140
References . . . 142

**4 Topological stratification of continuous genetic variation 145**
4.0 Preface . . . 145
4.1 Abstract . . . 146
4.2 Introduction . . . 147
4.3 Methods . . . 150
4.4 Results . . . 153
4.4.1 Clustering captures population structure from sample design . . . 153
4.4.2 Correlates between populations and sociodemographic, phenotypic, and environmental variables . . . 157
4.4.3 Phenotype smoothing and modelling . . . 160
4.4.4 Evaluating transferability of polygenic scores . . . 163
4.4.5 Quality control for complex multi-ethnic cohorts . . . 168
4.5 Discussion . . . 169

4.5.1 Applications . . . 172
4.5.2 Considerations . . . 174
4.6 Acknowledgements . . . 176
4.7 Materials and Methods . . . 176
4.8 Supporting information . . . 178
4.9 Supplementary figures and tables . . . 181
References . . . 200

**III Summary 205**

**5 Discussion 207**
5.1 The value of visualization . . . 207
5.2 Implications for biomedical and epidemiological research . . . 209
5.3 Clustering in population genetics . . . 212
5.4 Limitations . . . 215
5.5 Future directions . . . 216

**6 Conclusion 218**

**Master reference list 219**

**A Re-use permissions 225**

# Abstract

Advances in genomics have led to the rise of the biobank. Now containing the genetic data of millions of individuals, biobanks are rich repositories, used regularly and fuelling scientific discovery. The human genome spans approximately three billion base pairs, making the study of large-scale genetic data one of very high dimensions. The sheer scale of the problem is challenging, as is the complexity of human genetic diversity. With new methods in topological data analysis, we are able to reduce the dimensionality of these massive data sets while preserving significant amounts of information using a methodology that is tractable on the biobank-scale.

In this research, we apply uniform manifold approximation and projection (UMAP), a form of non-linear dimensionality reduction, and HDBSCAN, a density-based clustering algorithm, to several biobanks of human genetic data. We use UMAP and HDBSCAN to study population structure, the phenomenon in which genetic variation is non-random and correlated with factors like geography, demographic history, migration, and social structure. We develop a methodology in which we can visualize genetic data, identify structure in genetic variation ranging from a handful to hundreds of thousands of individuals. We uncover subtle relationships between genetics, history, geography, and phenotype distributions.

# Résumé

Avec les progrès de la génomique, les biobanques sont en plein essor. Contenant les données génétiques de millions d'individus, les biobanques sont utilisées régulièrement et permettent de nombreuses découvertes scientifiques. Le génome humain s'étend sur environ trois milliards de paires de bases, ce qui fait de l'étude des données génétiques à grande échelle une tâche dans un espace de grande dimension. La taille des données et la richesse de la diversité génétique humaine compliquent l'analyse des données. L'objectif de cette thèse est de développer des méthodes basées sur la topologie pour transformer ces données en une version en plus basse dimension qui préserve le plus d'information possible.

Dans le cadre de cette recherche, nous appliquons «uniform manifold approximation and projection» (UMAP), une forme de la réduction de la dimension non linéaire, et HDBSCAN, un algorithme de regroupement basé sur la densité des données, à plusieurs biobanques de données génétiques humaines. Nous utilisons UMAP et HDBSCAN pour étudier la structure de la population, le phénomène dans lequel la variation génétique n'est pas aléatoire et est corrélée à des facteurs tels que la géographie, l'histoire démographique, la migration et la structure sociale. Nous développons une méthodologie qui nous permet de visualiser les données génétiques, d'identifier la structure de la variation génétique allant d'une poignée à des centaines de milliers d'individus. Nous découvrons des relations subtiles entre la génétique, l'histoire, la géographie et la distribution des phénotypes.

# Acknowledgements

This thesis would not have been possible without my many communities, nor without the contributions of those who donated their data to further our understanding of genetics.

I would like to thank Simon Gravel for his guidance and mentorship and his balance of energetic curiosity with scientific scepticism, and the members of my supervisory committee, Nada Jabado and Ryan Hernandez. I would also like to thank the QLS program, particularly Celia Greenwood and Alex DeGuise for all of their hard work.

Thank you to my QLS cohort for grinding it out with me: Myriah Haggard, Matt D'Iorio, Jeffrey Hyacinthe, Yixiao Zheng, Sara Zapata-Marin, and Selin Jessa; to Noor Al-Sharif and Jake Vogel, for making Montreal home; to my Thomson House Trivia crew: Matt, Barbara, Hector, Gerardo, JT, Val, and Rozzy, for being champions; to my lab, especially my desk-mate Chief Ben-Eghan for cracking me up; to the floræ and faunæ of Parc La Fontaine, for raising my spirits.

Thank you to my friends in my other homes. Those in Ottawa: Dylan, Leo, Julie, Ashley, Shane, Isaac, Amy, thanks for always hosting me when I was in town. To those in KW: Jake, Kayla, Warren, Kaitlin, Jon, Steph, Derek, Christina, Dave, Rebecca, Travis, Cynthia, Tyler, and Aine, thanks for entertaining my rants on Slack—may your broods grow belligerent and numerous.

Thank you to my loved ones: to Melissa, for your love and support and for exploring the forests and rivers with me; to my brother, Andrew, and my mother, Marina, for carrying me when I stumbled; and lastly to my father, Andrés, who started this marathon with me so long ago—the finish line is in sight, and I can hear you cheering.

To my family; and,

To the memory of Andrés Carlos Díaz Bravo.

# Contribution to original knowledge

This body of work presents a novel unified methodology for dimensionality reduction, visualization, and clustering in population genetics with possible extensions to genomics at large. We explore applications of UMAP, a nonlinear dimensionality-reduction method, and HDBSCAN($\hat{\epsilon}$), a density-based clustering algorithm. These form the basis of our topological data analysis. The approach is tractable, easy-to-implement, and fits in the paradigm of exploratory-confirmatory analysis of biobank data, particularly for large and complex cohorts.

In Chapter 2 we apply UMAP to population genetic data for the first time. We establish that it efficiently reveals fine-scale population structure in biobanks and use it in data visualization in 2- and 3-dimensions. These visualizations correlate with a number of phenotypic, geographic, and socio-demographic measures, as well as demographic histories, and often appear as clusters. As UMAP became popular in the field, in Chapter 3 we review its use in population genetics and provide insights on its optimal use and potential downstream applications.

Finally, in Chapter 4 we address algorithmic clustering of UMAP results. We apply HDBSCAN($\hat{\epsilon}$) to UMAP data in $3$ to $5$ dimensions and use it to stratify biobank data, particularly large biobanks with unbalanced population sizes. We discover clusters of structure in genetic data and leverage them to study demographic histories, polygenic score transferability, and phenotype distributions. We also use clusters for quality control and to evaluate potentially influential alleles in polygenic scores.

We have provided online repositories of code using publicly available genotype data.

# Contributions of authors

## Chapter 2

- **Alex Diaz-Papkovich**: Conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – original draft, writing – review & editing
- **Luke Anderson-Trocmé**: Visualization
- **Chief Ben-Eghan**: Data curation
- **Simon Gravel**: Supervision, writing – review & editing

Acknowledgements: Audrey Grant, Ryan Hernandez, Jose Sergio Hleap, Mark Lathrop, Dominic Nelson, Markus Munter, Stephen Sawcer, Melissa Spear, and Dara Torgerson for useful discussions about science, programming, and data access; David Poznik, Liz Babalola, and Adam Auton for discussing findings in the 1KGP; Selin Jessa for introducing us to UMAP

## Chapter 3

- **Alex Diaz-Papkovich**: Conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – original draft, writing – review & editing
- **Luke Anderson-Trocmé**: Visualization, data curation
- **Simon Gravel**: Supervision, writing – review & editing

## Chapter 4

- **Alex Diaz-Papkovich**: Conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – original draft, writing – review & editing
- **Shadi Zabad**: Software, data curation
- **Chief Ben-Eghan**: Data curation
- **Luke Anderson-Trocmé**: Visualization, data curation
- **Georgette Femerling**: Data curation
- **Vikram Nathan**: Methods testing
- **Jenisha Patel**: Methods testing
- **Simon Gravel**: Supervision, writing – review & editing

Acknowledgements: Claude Bhérer, Melissa Spear, and Paul Verdu for scientific discussion.

# List of abbreviations

### Biobanks, technical terms, and organizations

| Abbreviation | Definition |
|---|---|
| 1000GP, 1KGP | 1000 Genomes Project |
| ARG | Ancestral recombination graph |
| CAG | CARTaGENE |
| COB | Country of birth |
| EB | Ethnic background |
| EDA | Exploratory data analysis |
| FEV1 | Forced expiratory volume in 1 second |
| FVC | Forced vital capacity |
| GBMI | Global Biobank Meta-analysis Initiative |
| GWAS | Genome-wide association study |
| HRS | Health and Retirement Study |
| IBD | Identity-by-descent |
| IBD | Isolation-by-distance |
| IBS | Identity-by-state |
| LD | Linkage disequilibrium |
| MAF | Minor allele frequency |
| MDS | Multidimensional scaling |
| NHS | National Health Service |
| PCA | Principal component analysis |
| PGS | Polygenic score |
| PRS | Polygenic risk score |
| SNP | Single nucleotide polymorphism |
| SNV | Single nucleotide variant |
| SVD | Singular value decomposition |
| TDA | Topological data analysis |
| TMRCA | Time to most recent common ancestor |
| T-SNE | t-distributed stochastic neighbour embedding |
| UKB, UKBB | UK biobank |
| UMAP | Uniform manifold approximation and projection |

### The 1000 Genomes Project populations

| Abbreviation | Population name |
|---|---|
| ACB | African Caribbean in Barbados |
| ASW | African Ancestry in SW USA |
| BEB | Bengali in Bangladesh |
| CDX | Chinese Dai in Xishuangbanna, China |
| CEU | Utah residents with Northern/Western European ancestry |
| CHB | Han Chinese in Beijing, China |
| CHS | Han Chinese South |
| CLM | Colombian in Medellín, Colombia |
| ESN | Esan in Nigeria |
| FIN | Finnish in Finland |
| GBR | British From England and Scotland |
| GWD | Gambian in Western Division – Mandinka |
| GIH | Gujarati Indians in Houston, Texas, USA |
| IBS | Iberian Populations in Spain |
| ITU | Indian Telugu in the UK |
| JPT | Japanese in Tokyo, Japan |
| KHV | Kinh in Ho Chi Minh City, Vietnam |
| LWK | Luhya in Webuye, Kenya |
| MSL | Mende in Sierra Leone |
| MXL | Mexican Ancestry in Los Angeles, CA, USA |
| PEL | Peruvian in Lima, Peru |
| PJL | Punjabi in Lahore, Pakistan |
| PUR | Puerto Rican in Puerto Rico |
| STU | Sri Lankan Tamil in the UK |
| TSI | Toscani in Italy |
| YRI | Yoruba in Ibadan, Nigeria |

# List of figures

1.1 Contrasting Fisher's paradigm with Tukey's paradigm in biology . . . . . 13
1.2 Density clustering . . . . . 22

2.1 Four methods of dimension reduction of 1KGP genotype data . . . . . 35
2.2 Applying UMAP to subsets of data can reveal deep population structure . . . . . 40
2.3 The UKBB coloured by self-reported ethnic background . . . . . 41
2.4 UMAP captures relationships between population structure and geography . . . . . 44
2.5 Maps coloured by 3D UMAP projections . . . . . 45
2.6 UMAP captures relationships between population structure and phenotype heterogeneity. . . . . 47
2s1 Montage of $t$-sne and UMAP on up to 9 PCs of 1KGP data . . . . . 58
2s2 Montage of $t$-sne and UMAP on 10 to 50 PCs of 1KGP data . . . . . 59
2s3 Montage of UMAP on progressively more PCs of 1KGP data . . . . . 60
2s4 UMAP on PCs 100 to 3350 of 1KGP data . . . . . 61
2s5 Number of neighbours and families forming disjoint clusters . . . . . 62
2s6 UMAP on HRS data coloured by ethnicity . . . . . 63
2s7 UMAP on HRS data coloured by admixture . . . . . 64
2s8 UMAP on HRS data coloured by birth region . . . . . 65
2s9 UMAP on HRS data with 1KGP data overlaid . . . . . 66
2s10 Pairwise plots of PCs of Hispanic HRS data . . . . . 67
2s11 UMAP on Hispanic HRS data coloured by admixture . . . . . 68
2s12 UMAP on Hispanic HRS data coloured by birth region . . . . . 69
2s13 UMAP on Asian UKBB data coloured by self-identified ethnicity . . . . . 70
2s14 UMAP on UKBB data with some countries of birth identified . . . . . 71
2s15 UMAP on UKBB data coloured by distance from London . . . . . 72
2s16 Montage of UMAP on top 40 PCs of UKBB data coloured by ethnicity . . . . . 73
2s17 Montage of UMAP on top 40 PCs of UKBB data coloured by northing . . . . . 74
2s18 Montage of UMAP on top 40 PCs of UKBB data coloured by easting . . . . . 75
2s19 Map of Asia coloured by 3D UMAP coordinates of UKBB data . . . . . 76
2s20 Map of Caribbean coloured by 3D UMAP coordinates of UKBB data . . . . . 77
2s21 Map of Europe coloured by 3D UMAP coordinates of UKBB data . . . . . 78
2s22 $t$-sne on UKBB data coloured by self-identified ethinicity . . . . . 79
2s23 UMAP on UKBB data coloured by basophil count (female) . . . . . 80
2s24 UMAP on UKBB data coloured by basophil count (male) . . . . . 81
2s25 UMAP on UKBB data coloured by eosinphil count (female) . . . . . 82
2s26 UMAP on UKBB data coloured by eosinphil count (male) . . . . . 83
2s27 UMAP on UKBB data coloured by FEV1 (female) . . . . . 84
2s28 UMAP on UKBB data coloured by FEV1 (male) . . . . . 85
2s29 UMAP on UKBB data coloured by height (female) . . . . . 86

2s30 UMAP on UKBB data coloured by height (male) . . . 87
2s31 UMAP on UKBB data coloured by leukocyte count (female) . . . 88
2s32 UMAP on UKBB data coloured by leukocyte count (male) . . . 89
2s33 UMAP on UKBB data coloured by neutrophil count (female) . . . 90
2s34 UMAP on UKBB data coloured by neutrophil count (male) . . . 91
2s35 Box plots of height in the UKBB by self-identified ethnicity (female) . . . 92
2s36 Box plots of height in the UKBB by self-identified ethnicity (male) . . . 93
2s37 Box plots of FEV1 in the UKBB by self-identified ethnicity (female) . . . 94
2s38 Box plots of FEV1 in the UKBB by self-identified ethnicity (male) . . . 95
2s39 Subset (left) of UKBB UMAP projection coloured by height, FEV1, and self-identified ethnicity . . . 96
2s40 Subset (top) of UKBB UMAP projection coloured by height, FEV1, and self-identified ethnicity . . . 97
2s41 East Asian individuals from UKBB UMAP projection selected for FEV1 investigation . . . 98
2s42 Ridge plots of East Asian individuals from UKBB UMAP projection selected for FEV1 investigation . . . 99
2s43 Comparison of $t$-sne error by initialization on UKBB data . . . 100
2s44 Comparing visualizations of $t$-sne and UMAP of UKBB data by initialization . . . 101
2s45 PCs 1 and 2 of the UKBB coloured by height (female) . . . 102
2s46 PCs 1 and 2 of the UKBB coloured by FEV1 (female) . . . 102
2s47 $t$-sne projection of UKBB data coloured by height (female) . . . 103
2s48 $t$-sne projection of UKBB data coloured by FEV1 (female) . . . 104
2s49 Zoomed in views of UMAP projection of UKBB data, coloured by self-identified ethnicity . . . 105
2s50 Comparing visualizations of $t$-sne and UMAP of 1KGP data by initialization . . . 106
2s51 Comparing visualizations of $t$-sne and UMAP of HRS data by initialization . . . 107
2s52 Comparison of $t$-sne error by initialization on 1KGP data . . . 108
2s53 Comparison of $t$-sne error by initialization on HRS data . . . 109
2s54 Box plots of basophil count in the UKBB by self-identified ethnicity (female) . . . 110
2s55 Box plots of basophil count in the UKBB by self-identified ethnicity (male) . . . 111
2s56 Box plots of eosinophil count in the UKBB by self-identified ethnicity (female) . . . 112
2s57 Box plots of eosinophil count in the UKBB by self-identified ethnicity (male) . . . 113
2s58 Box plots of leukocyte count in the UKBB by self-identified ethnicity (female) . . . 114
2s59 Box plots of leukocyte count in the UKBB by self-identified ethnicity (male) . . . 115
2s60 Box plots of neutrophil count in the UKBB by self-identified ethnicity (female) . . . 116
2s61 Box plots of neutrophil count in the UKBB by self-identified ethnicity (male) . . . 117
2s62 UMAP projection of combined HRS and 1KGP data . . . 118
2s63 Alternate colouring of 2s7 . . . 119
2s64 An alternate colouring of 2s11 . . . 120

2s65 Admixture plot of Hispanic individuals in the HRS . . . 121

3.1 UMAP with and without HLA regions filtered . . . 131
3.2 PCA compared to UMAP of the 1KGP . . . 132
3.3 PCA compared to UMAP of the UKB . . . 133
3.4 UMAP of gnomAD and Biobank Japan . . . 134
3.5 UMAP parametrization changes the connectivity of points . . . 139

4.1 Overview of visualization and clustering pipeline . . . 150
4.2 Clusters generated from 1KGP genotype data reflect its population sampling . . . 154
4.3 Clusters capture structure in populations with overlapping admixture proportions in the 1KGP . . . 155
4.4 Clusters of population structure in the UKB . . . 159
4.5 Smoothed phenotypic measures across multiple parametrizations of clustering . . . 162
4.6 MSE of cluster-based phenotype estimation . . . 166
4.7 PGS accuracy by $F_{ST}$ . . . 167
4.8 Clustering identifies data collection errors in CaG . . . 168
4s1 Clustering the UKB with basic HDBSCAN . . . 181
4s2 Word clouds generated from four clusters in the UKB . . . 182
4s3 Admixture proportions for 5 populations . . . 183
4s4 Proportion of daily smokers . . . 184
4s5 Distributions of FEV1 by cluster . . . 185
4s6 Distributions of neutrophil count by cluster . . . 186
4s7 Cluster 22 from Figure 4.4a highlighted coloured in on a plot of PC1 and PC2. . . . 187
4s8 Regression line of PGS vs MAF of rs4420638 . . . 188
4s9 Regression line of PGS vs MAF of rs7412 . . . 189
4s10 Measuring the number of individuals not clustered . . . 190
4s11 Alternative clustering of the UKB . . . 191

5.1 The relationship between target and study populations . . . 210
5.2 The treachery of clustering . . . 213

A1 Copyright permission for Chapter 2 . . . 225
A2 Copyright permission for Chapter 3 . . . 226
A3 Copyright permission for Chapter 4 . . . 227

# List of tables

2s1 Variance explained by PCs in the 1KGP . . . . . 122

4s1 Regression summary of PGS against MAF of rs4420638 . . . . . 188
4s2 Regression summary of PGS vs MAF of rs7412 . . . . . 189
4s3 Names and abbreviations of 1KGP populations. . . . . 192
4s4 Cluster assignments for each 1KGP population . . . . . 193
4s5 Composition of each cluster broken down by 1KGP population . . . . . 194
4s6 Possible values for ethnic background in the UKB . . . . . 195
4s7 Frequency of country of birth by cluster . . . . . 196
4s8 Frequency of selected EB by cluster . . . . . 197
4s9 Comparing phenotype models by EB (1) . . . . . 198
4s10 Comparing phenotype models by EB (2) . . . . . 199

# Part I

# Introduction and literature review

***Fry:*** *Here you are in the year 3000 or so, yet you just sit around like it's the boring time I came from.*
***Professor Farnsworth:*** *Boring?! Wasn't that the period when they cracked the human genome, and boy bands roamed the earth?*
*—Futurama, S03E11 (2001)*

# Chapter 1: Overview

In 2001, the first analyses of the draft human genome were published in sister papers in *Nature* and *Science*. The Human Genome Project had been budgeted US$3 billion in 1990; by 2020 the cost of sequencing a human genome had dropped to US$1,000, and a relative paucity of data has given way to abundance[1]. Massive biobanks are becoming commonplace.

Every genome carries both the stories of its ancestors and the basic programming of its bearer's physiology. By identifying patterns across many genomes and their associated data, we can infer their histories and study distributions of biomedical traits. The complexities of human history and society, to say nothing of the complexities of biology itself, ensure that this is a non-trivial task.

With each genome spanning 3 billion base pairs, any mathematical investigation is high-dimensional. Within these data, there are systematic patterns in the distributions of genetic variants: population structure. To study structure, it is helpful to reduce the dimensionality of the data. Uniform manifold approximation and projection (UMAP) is a recently-developed method of dimensionality reduction that is rooted in topology. This thesis explores topological data analysis of population genetics data through UMAP and the density clustering algorithm HDBSCAN($\hat{\epsilon}$). We will study how the high-dimensional topology of population genetic data reflects the demographic histories of the individuals we have sequenced, informs us of their biomedical measures, and how to explore this new trove of data.

The thesis is organized into three parts. The first part contains this chapter and provides a literature review and description of relevant methods; the second, consisting of three chapters,

makes up the original contributions to the field; and the third consists of a general discussion and the final conclusions.

In the literature review, we will describe the origins of human genetic diversity and population structure. We will discuss the methods used to study population structure, the challenges it poses to biomedical studies, and the types of data sets used in the field—biobanks. Having motivated the study of population structure in biobanks, we will outline a methodology of exploratory data analysis using dimensionality reduction and clustering.

The three chapters of original contributions have been previously published as manuscripts. In Chapter 2 we apply UMAP to human genetic data for the first time. We use genotype data from three biobanks, generate visualizations and observe patterns in relatedness, demographic histories, geographic distribution, phenotype distributions, and other phenomena. In Chapter 3 we review the applications of UMAP in other human genetic datasets, such as different biobanks or other types of genetic data. Finally, in Chapter 4 we formalize a methodology to use UMAP in higher dimensions ($n \geq 3$), extract clusters algorithmically, and apply these abstractions to a variety of problems common in biobank research.

## 1.1 Genetic diversity

The human genome is organized across 23 pairs of chromosomes—22 pairs of autosomes and one pair of sex chromosomes—with some DNA present in mitochondria (mtDNA). It is diploid with one set of chromosomes coming from each parent via their gametes; these chromosomes are cre-

ated through the process of meiotic recombination, in which the chromosomes of grandparents are aligned, cross over, and recombine. Along with mutation, recombination generates diversity. Approximately $99.9\%$ of DNA shared between humans is identical, with genetic variants (alleles) arising through mutations. Single nucleotide polymorphisms (SNPs) are relatively common variants, usually defined as having a frequency above $1\%$.

Variants that lie along the same chromosome and are separated through recombination are co-inherited and are linked. The block of allelic states along a DNA molecule is referred to as a haplotype, and when the same variants exist between two individuals, they are said to be identical by state (IBS). If the shared variant is inherited from a common ancestor without recombination, they are also said to be identical by descent (IBD); alleles that are IBS are typically IBD, with rare exceptions. Alleles that are physically closer are more likely to be inherited together, and those that appear together more often than expected at random are said to be in linkage disequilibrium (LD). Combining two haplotypes gives a diploid genotype, and assuming free recombination, the theoretical maximum number of possible unique haplotypes is $2^L$, where $L$ is the number of biallelic SNPs.

Recombination is not uniformly random. DNA that does not lie in the pseudoautosomal regions (PAR1 and PAR2) of the Y chromosome, as well as mtDNA, does not recombine[2]. Recombination rates also vary within chromosomes with certain regions known to be hotspots[3]. Germline mutations may result from copying errors during replication or from spontaneous errors from DNA's instability or external factors like UV radiation. Whole genome sequencing pedigree-

based studies estimate the overall mutation rate at about $10^{-8}$ per base pair per generation, though this rate may vary depending on the mechanistic source of the mutation[4].

### 1.1.1 Population structure

Population structure, also called stratification, is the systematic difference in allele frequencies between groups. The distribution of genetic variation is not fully random—genetic differentiation in subpopulations is nearly ubiquitous across organisms. Allele frequencies are influenced by factors like natural selection favouring certain genotypes, founder effects, genetic drift, etc.[5]. The mating range of an individual is usually much more constrained than the range of whole species; this leads to isolation by distance (IBD) where there is local structure because of random drift[6]. Positive assortative mating, in which individuals choose mates with similar phenotypes, can increase homozygosity[5]. Structure may also arise from population bottlenecks or geographic isolation[7] or admixture between groups that had been separated[8].

Population structure in humans is universal, complex, and multifactorial, resulting from all of the aforementioned causes (as well as phenomena like language[9] and culture[10]), and presents as both continuous and discrete[11]. Inferring deep population structure is challenging because of a lack of data and fossil records and the complexity of the required models. Most present-day population structure is relatively recent, having arisen within the last $100,000$ years; prior structure may be best explained by repeated divergence, isolation, and merger between weakly-differentiated *Homo sapiens* populations, though some combination of multiple populations and

archaic introgression is also possible[12].

### Hardy-Weinberg principle

The Hardy-Weinberg model was outlined in 1908. It is used to model genotype frequencies, given several simplifying assumptions[13]. If we have a diploid genome with alleles $A_1$ and $A_2$ that occur at respective frequencies $p$ and $q$, they follow the binomial distribution and are expected to naturally occur at rates:

$$A_1 \times A_1 = p^2$$
$$A_2 \times A_2 = q^2$$
$$A_1 \times A_2 = 2pq$$

We can test for excess departure from Hardy-Weinberg equilibrium with a $\chi^2$ statistic, which can indicate technical errors, non-random mating, or population stratification.

### Fixation index

We expect the level of heterozygosity ($\pi$) to be lower in the presence of population structure. The fixation index ($F_{ST}$) measures population structure; values close to $0$ indicate no structure, while values close to $1$ indicate fully structured populations. Several different estimators exist, depending on the context[14]. The sample size-independent definition for two populations is:

$$F_{ST} = \frac{Var(\pi_i)}{\pi(1-\pi)}$$

where $Var(\pi)$ is the variance in allele frequencies between two populations.

$F_{ST}$ is used as a measure of differentiation between populations. Though it is sometimes described as a pairwise distance between populations, it is not a true mathematical distance as it does not satisfy the triangle inequality[15].

### Wright-Fisher model

The Wright-Fisher model is a forward simulation of random genetic drift[16]. It assumes a fixed population size of $N$ and discrete non-overlapping generations, where individuals from generation $k+1$ draw their alleles at random from generation $k$. It is a simple foundational approach that can be modified to allow for forces like selection, mutation, etc. The alleles follow a binomial distribution with parameters $2N$ and $p$, where $p$ is the proportion of individuals with an allele in the previous generation.

### Coalescent

The trajectory of $N$ loci can be modelled backwards using the coalescent approach[16]. For large $N$, the probability that two branches coalesce at time $t$ is exponentially distributed with mean $N$. To incorporate recombination, the problem is formulated as an Ancestral Recombination Graph (ARG) where nodes in the tree are events (either common ancestors or recombinations). This approach scaled poorly, leading to the development of the sequentially Markov coalescent (SMC) approximation where observed data are dependent on underlying sequence data and are modelled

as a hidden Markov Model (HMM). There have been considerable improvements in computational approaches for coalescent models and it is an active area of research[17–19].

### 1.1.2 Genome-wide association studies and polygenic scores

One of the greatest promises of the genomic age is to reveal the genetic architecture of phenotypes and disease risk. Two common and related methodologies are the genome-wide association study (GWAS) and the polygenic score (PGS), also called the polygenic risk score (PRS) when used in the context of disease risk.

The most basic GWAS is a linear regression model. We assume that for some phenotype $y$, the variant $x$ has an additive effect that can be modelled linearly. Then, we can study the association of each variant across the genome, hence the name GWAS. The basic outline is:

$$y_i = \beta_0 + x_j\beta_j + \epsilon_i \text{ , } \epsilon_i \sim N(0, \sigma^2)$$

where $x_j$ is the vector of genotype values at SNP $j$ and $\beta_j$ is its effect size. GWAS commonly include terms for genetic relatedness between samples as well as covariates related to genetic ancestry and environment. The model can be written as, e.g.[20]:

$$\begin{aligned} Y &= W\alpha + X_j\beta_j + g + \epsilon \\ g &\sim N(0, \sigma_A^2\psi) \\ \epsilon &\sim N(0, \sigma_\epsilon^2 I) \end{aligned}$$

Here, $Y$ is a vector of the phenotypes of individuals, $W$ is the matrix of covariates with associated effect sizes $\alpha$, $X_j$ is a vector of genotypes at locus $j$ with effect size $\beta_j$, $g$ is the random polygenic effect of other SNPs with $\sigma_A^2$ measuring the phenotype's additive genetic variance and $\psi$ the genetic relatedness matrix, and $\epsilon$ is the regression model residual. To account for multiple testing, the threshold for statistical significance for $\beta_j$ is usually set to $10^{-8}$. A variety of methods have been developed to estimate $\beta$ in different frameworks, e.g. via linear mixed-models[21], incorporating rare variants[22], using ridge regression[23], etc.

Given a set of estimates $\hat{\beta}$ for each SNP (or for each significant SNP), one could construct an estimate of a phenotype as the sum of the estimated effects. For an individual $i$ and for SNPs $j = 1, \ldots, p$ we estimate the PGS $\hat{y}_i = \sum_j^p x_{ij}\hat{\beta}_j$. Importantly, PGS have been noted to transfer poorly to populations that are more differentiated from the populations used to estimate the values of $\hat{\beta}$[24].

Despite significant advances in data collection and methodology, population structure is a persistent confounder in GWAS and PGS. If a variant is more common in one population than another, and a measure differs systematically between them for non-genetic reasons (e.g. environmental exposure, cultural differences in diet, etc.), a GWAS would return a spurious correlation[25]. These systematic differences can accrue in downstream applications like PGS[26].

The effects of structure can be pernicious. In the 2010s several studies identified an increase in height along a south-to-north cline in Europe and attributed this to natural selection; two studies in 2019 found that these analyses were likely confounded by population structure[27, 28]. Even

in populations considered to be relatively homogeneous, structure has been found to bias PGS estimates[29]. With applications like PGS becoming more common, understanding population structure in biobanks is critical[30].

## 1.2 Biobanks

Biobanks are repositories of biological samples and data; we are concerned with those containing genetic data from humans. Large-scale biobanks are now common, with many having data from hundreds of thousands of participants. The Global Biobank Meta-analysis Initiative (GBMI) lists $23$ biobanks with genetic and phenotypic data from over $2.2$ million individuals in total[31]. These biobanks are commonly used for studying genetic ancestry, demographic history, and biomedical research such as GWAS and PGS.

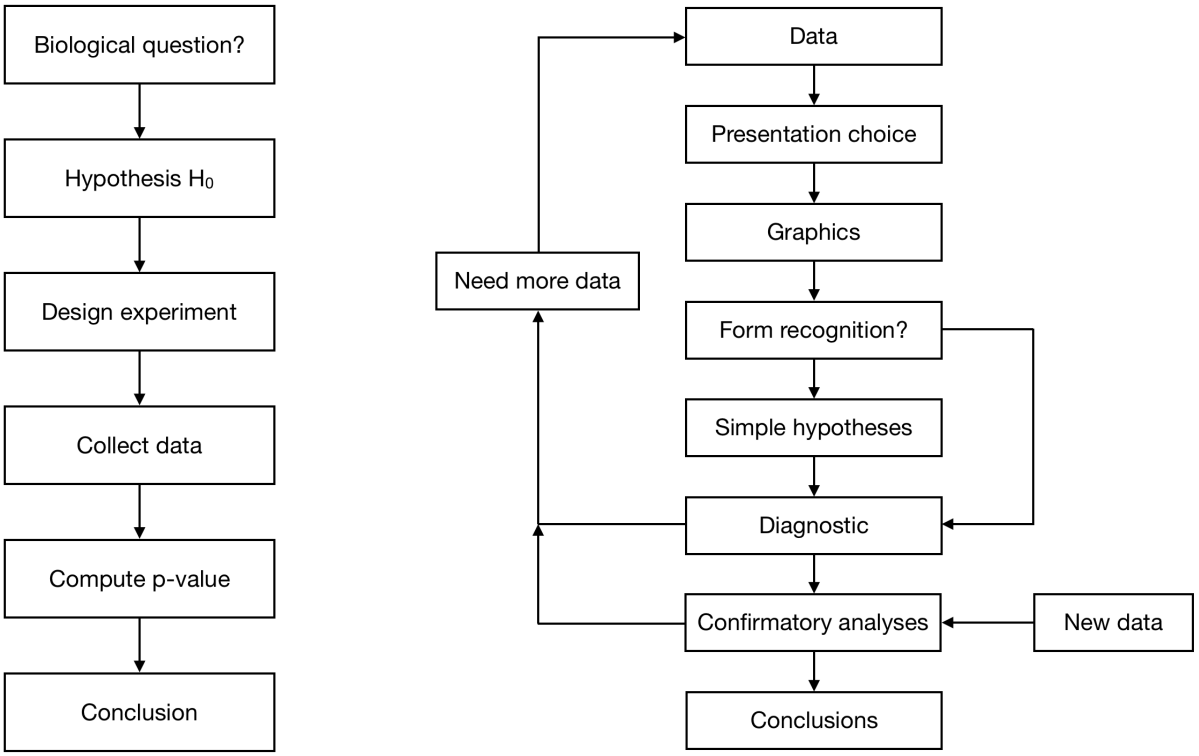
The sampling methodologies of biobanks vary widely—two common approaches include those that are designed to study genetic diversity, and those that provide genetic data for existing health databases. For example, the 1000 Genomes Project (1KGP) was designed to represent a diverse collection of populations from around the world[32]. It contains $3,450$ genotypes sampled from $26$ populations with $104$ to $183$ individuals in each group. We study its population structure throughout this thesis, finding clusters of structure in Chapter 2 using UMAP, exploring the concept of connectivity in Chapter 3, and automatically extracting clusters in Chapter 4 using HDBSCAN($\hat{\epsilon}$). As the data are publicly available, there are repositories of our code and data provided for each chapter.

In contrast, public health biobanks are usually sampled from some combination of political jurisdiction and existing health care services. The UK biobank (UKB)[33] and CARTaGENE (CaG)[34], for example, recruit volunteers who are registered with their respective health care administrations within certain geographical regions and political boundaries. Such administrative records are convenient sampling frames, since they tend to have contact information, associated data such as electronic health records, and established epidemiological methodologies. The wealth of health data, particularly longitudinal data, makes these biobanks invaluable for health care research. Their population structure, however, is extraordinarily complex and not known in advance. Even limiting a biobank to one city requires in-depth knowledge of its regional history, its migrations, its cultures, its physical and social environments, etc. Without a means to understand population structure and how it relates to the aforementioned, biobanks can transform from a panacea into a witch's brew of statistical biases. In Chapters 2 and 4, we study the relationships between population structure and several variables in biobanks.

## 1.3 Exploratory data analysis

Traditional statistical analyses of biological data follow the recommendations of R.A. Fisher in the 1930s[35]. They have a linear structure: an investigator begins with a biological question, forms a hypothesis and an associated null ($H_0$), designs an experiment, collects data, tests $H_0$ with a p-value, and formulates a conclusion. Beginning in the 1970s, statistician John W. Tukey proposed the alternative framework of exploratory data analysis (EDA)[36, 37]. This approach is iterative

and instead begins with the data: we visualize it, understand it, and use it to inform what sort of analysis to use in an accompanying confirmatory data analysis. The two contrasting approaches are schematized in Figure 1.1.




**(a)** Fisher's paradigm. **(b)** Tukey's paradigm.

**Figure 1.1: Contrasting Fisher's paradigm with Tukey's paradigm in biology.** Fisher's paradigm (left) takes a sequential approach to data analysis, beginning with a well-defined question and strong assumptions. Tukey's paradigm (right) is iterative, beginning with the data, emphasizing exploratory analysis through visualization, and complemented by confirmatory analyses that are robust and do not rely on complex assumptions. Figure adapted from [35].

Writing in 1980, Tukey emphasized that science neither begins with a tidy question nor ends with a tidy answer[38]. This is especially true in modern biology. Statistical questions from the 1930s typically had a few parameters $p$ with a manageable sample size $N$ (where $N > p$), and the

people posing questions were involved in data collection. Today we sit at the opposite extreme; it is not unusual for data to have $p \gg N$ with the two values differing by orders of magnitude. When studying a biobank, we may have several thousand individuals and several hundred thousand genetic markers. Often, the people investigating data have not collected it. These factors make Tukey's paradigm much better suited to our analytical needs[35].

### 1.3.1 Dimensionality reduction

In Figure 1.1b, the iterative process includes presentation choices, graphics, and form recognition. This provokes a natural question: what approaches ought we to use here? With genomic data comes the "curse of dimensionality": though we have many dimensions to our data, the signal is sparse and many methods are computationally intractable. This motivates dimensionality reduction—we wish to reduce our data to a relatively low number of dimensions, ideally preserving important characteristics of the data. Given a satisfactory representation of the data set, we can visualize it.

#### Principal component analysis

Principal components analysis (PCA) is a non-parametric linear transformation that projects data onto a series of orthogonal axes based on a linear combination of the original data. The axes are generated and ordered according to the eigenvalues of the covariance matrix of the data, and the ratio of each axis' corresponding eigenvalue to the sum of all eigenvalues represents the proportion of variance explained by that axis. We can think of PCA as fitting an ellipsoid around the data in

high dimensions and the axes of that ellipsoid are the principal components. By only selecting the largest axes—corresponding to the most explained variance — we can reduce the dimensionality of our data while preserving significant explanatory value. We can also interpret our dimensionally reduced data in terms of how much of the overall variance it explains. Principal components are calculated through eigendecomposition of the covariance matrix. Detailed examinations of PCA in the context of population genetics can be found in [11] and [39].

PCA has seen wide application in population genetics. The top PCs often reflect isolation-by-distance and are used for visualization (e.g. within Europe[40]). However, using them for visualization requires selecting which components to examine and is limited to $2$ or $3$ dimensions; if there is signal beyond the first few PCs, it may go unnoticed. We expand on this in Chapter 2.

They are also used to correct for population structure in genome-wide association studies (GWAS) by their inclusion as covariates in models[25]. There are varying rules-of-thumb on how many PCs to include in a model, such as using the top $10$, looking for an "elbow" in the scree plot, or testing for significance in the Tracy-Widom distribution. We explore the impact of PC adjustment for phenotypes in biobanks in Chapter 4.

### 1.3.2 Topological data analysis

We are often interested in understanding the large-scale structure of our data, e.g., identifying different cell types or related individuals. Though we have some definitions of distances, we are interested in notions of *similarity*, *nearness*, and *connectivity*. Topology provides the mathematical

machinery for ideas rooted in qualitative geometry[41]. Topological data analysis (TDA) is a set of statistical methods that uses ideas of shape and connectivity to study data[42]. We will focus on manifold learning, nonlinear dimensionality reduction, and density clustering.

TDA assumes that we observe a sample $X_1, \ldots, X_n \sim P$ with $P$ supported on some set $\text{supp}(P) = \mathcal{X} \subseteq \mathbb{R}^d$. In the simplest case of manifold learning, we suppose that $P$ is actually supported on some set $S$ with dimension $r$, where $r < d$ and $S$ is a smooth and compact manifold, and we may estimate $S$. PCA is a special case of linear manifold learning where data are assumed to lie on or near an affine subspace[42]. In cases where there is local nonlinear structure (such as clustering), nonlinear methods of manifold learning are more useful[43].

### 1.3.3 t-distributed stochastic neighbour embedding

$t$-distributed stochastic neighbour embedding ($t$-SNE) is a method of manifold learning used for visualization that was developed in 2008[44]. By then, several methods existed to approximate the local structure of manifolds, but they suffered from the "crowding problem"—in an attempt to preserve local distances between points, many of them are crunched together, eliminating the gaps between clusters. $t$-SNE addressed this by introducing a repulsion force between points, modelling pairwise distances between points $i$ and $j$ as a $t$-distributed random variable with $1$ degree of freedom (equivalent to a Cauchy distribution). The distances are modelled as probabilities:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}$$

This choice was ad-hoc and was later found to work because it optimized structure at the local scale (i.e. within clusters) as well as causing points to repel each other (i.e. causing clusters to separate)[45]. This repulsion allowed $t$-SNE and related methods to preserve topology[42]. Because $t$-SNE could only reduce data to $2$ or $3$ dimensions, it was not recommended as a general purpose dimensionality reduction algorithm[44]. It saw considerable use in visualization in single-cell genomics[46], but its application in population genetics was limited (e.g. [47]). We provide details on $t$-SNE's performance in population genetics in Chapter 2.

### 1.3.4 Uniform manifold approximation and projection

Uniform manifold approximation and projection (UMAP) is a general purpose dimensionality reduction method rooted in algebraic topology and Riemannian geometry that was introduced in 2018[48]. The motivation underlying UMAP is to explicitly represent the high-dimensional topology of data in low dimensional space. We will briefly outline UMAP; details on the topology and theoretical justifications are available in [48], with a more intuitive explanation available in online documentation[49].

We assume our data $X = \{X_1, \ldots, X_n\}$ lay on some manifold and are uniformly distributed. For this assumption to hold, each point $X_i$ has its own custom distance, defined as the normalized distance to its $k^{\text{th}}$ nearest neighbour; thus, each $X_i$ has its own metric space, and is the centre of a unit ball that extends to the $k^{\text{th}}$ nearest neighbour. If we represent this as a graph, each $X_i$ is a point with edges to its $k$ neighbours, where the distances represent the edge weights. We can

represent this as a simplicial complex—a set of points and edges—where each point is a $0$-simplex and each edge is a $1$-simplex. This simplicial complex forms an open cover of the underlying topological space. As the edge weights are between $0$ and $1$, we may also interpret the values as the belongingness to an open set in a cover rather than a binary "yes" or "no" value—a fuzzy topological cover. To harmonize the respective edge weights $a, b$ from points $X_a$ to $X_b$ (since each point has its own local metric), UMAP defines the combined weight as $a + b - a \times b$, interpreted as the probability that an edge weight between $X_a$ and $X_b$ exists. The final high-dimensional product is a fuzzy simplicial complex, which can be represented as a weighted graph, and is a fuzzy topological representation of the data.

For the low-dimensional representation, we carry out the same process of building a fuzzy topological representation. Rather than using a locally-varying metric, we assume that our data will lay on a low-dimensional Euclidean space, and we specify a minimum distance we wish to have between our points in this space. The algorithm then minimizes the cross-entropy function between the high- and low-dimensional representations. If $E$ is the set of all possible $1$-simplices, $w_h(e)$ is the weight of edge $e$ in the high-dimensional space and $w_l(e)$ the low-dimensional space, we minimize:

$$\sum_{e \in E} w_h(e) \log \left( \frac{w_h(e)}{w_l(e)} \right) + (1 - w_h(e)) \log \left( \frac{1 - w_h(e)}{1 - w_l(e)} \right)$$

UMAP allows for reduction to an arbitrary number of dimensions. The value of $k$ defines the scale of the topology we wish to approximate, with lower values being more local and finer-scale

and higher values approximating broader manifold structure. Each chapter of this thesis discusses the uses and parametrizations of UMAP in population genetics: briefly, lower values of $k$ approximate closer relationships, e.g., at a structure as fine-scale as families; higher values of minimum distance facilitate visualization, while lower values facilitate algorithmic cluster detection. Points that appear near each other in low-dimensional space are closely related to one another. UMAP often generates visual clusters—while the clusters themselves can be interpreted, the distances between them are generally not meaningful. Importantly, UMAP is a fast and scalable algorithm, capable of handling millions of data points.

UMAP is the core method of this thesis. In Chapter 2, we use UMAP in population genetics for the first time, exploring its potential applications thoroughly and compare it to PCA and t-SNE. In Chapter 3, we review its uses in the field and discuss different data inputs and parametrizations. In Chapter 4, we introduce the use of UMAP for topological stratification of complex biobank data by using it to pre-process data for clustering rather than visualization.

### 1.3.5 Clustering

Broadly, clustering is a class of methods that puts similar data points into the same cluster while keeping dissimilar data points in different clusters[50]. Clustering in population genetics can be separated into model-based methods, which are common in global ancestry estimation, and distance-based methods, which measure pairwise distances between individuals. Though we use the former in some analyses, Chapter 4 focuses on the latter, specifically density clustering.

It can be useful to model populations as $K$ discrete demes, e.g. when modelling admixture, studying population splits and merges, or testing the transferability of GWAS and PGS. Clustering is a natural approach to this problem. A variety of clustering methods have been used in population genetics. These methods often require specifying $K$; though multiple methods have been proposed to estimate $K$ (e.g. [51, 52]), it does not have a "correct" value as genetic data does not fall into natural discrete groups, and it is more useful to consider how well a value of $K$ helps to explain the question being posed[53].

### Model-based clustering

Model-based clustering assumes that observations are randomly drawn from some parametric model. The first such method used in population genetics was STRUCTURE, in 2000, which assumed that each cluster was defined by some allele frequency; having observed the genotypes $X$, it infers the populations of origin $Z$ and allele frequencies $P$ through Bayesian modelling of $Pr(Z, P|X)$[54]. Each genome is then modelled as a mixture of some $K$ source populations—often presented in literature as a stacked bar graph where each individual is one bar with split ancestry proportions—and termed "global ancestry estimation"[55]. The idea has since evolved into many other methods (e.g. ADMIXTURE[55], FRAPPE[56], sparse non-negative matrix factorization[57], archetypal analysis[58]) and is an active area of research to handle more complex model assumptions and larger data sets.

### K-means clustering

One of the most common distance-based algorithms is $K$-means clustering. Developed in the 1970s, it works by dividing $M$ points in $N$ dimensions into $K$ groups by minimizing the within-cluster sum of squares[59]. $K$-means clustering has been used to define populations by, e.g., using the coordinates of individuals in PCA space and comparing to some known reference population within a cluster. $K$-means implicitly assumes that data points form a Gaussian distribution around some centroid[60]. This does not describe genetic data well—data points are either partitioned arbitrarily, stripping clusters of interpretation, or the distances from the centroids have sharp cut-off points. In the latter case many individuals, particularly from admixed populations or uncommon genetic ancestries, will never be placed into a cluster, especially if there is no reference population to compare them to[61].

### Density clustering

Density clustering finds sets of data with a high density of points; a formal outline can be found in Wasserman's review of TDA[42]. Briefly, assume we observe a sample $X_1, \ldots, X_n$ from some distribution $P$ with density $p$ where $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$. For any $\lambda \geq 0$, we define the upper level set $L_\lambda = \{x : p(x) > \lambda\}$. The density clusters at level $\lambda$ are denoted by $\mathcal{C}_\lambda$ and are the connected components of $L_\lambda$. The set of all density clusters is:

$$\mathcal{C} = \bigcup_{\lambda \geq 0} \mathcal{C}_\lambda$$

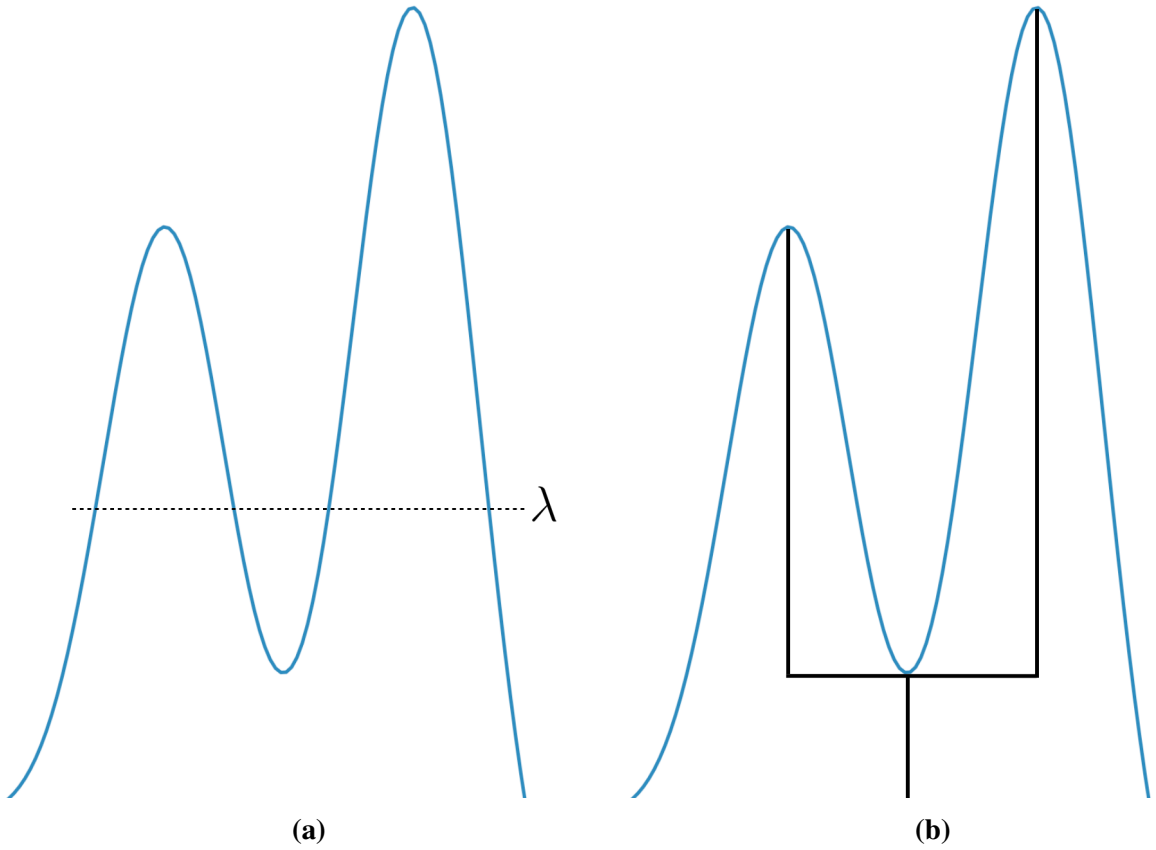



**Figure 1.2: Density clustering.** A density function $p$. **(a)** Density clusters corresponding to $L_\lambda = \{x : p(x) > \lambda\}$. **(b)** The density tree.

By varying the threshold $\lambda$, level sets $L_\lambda$ become nested within each other; that is, if $A, B \in \mathcal{C}$ then $A \subset B$ or $B \subset A$ or $A \cap B = \emptyset$. Thus the set $\mathcal{C}$ can be represented as a tree, and each branch represents a cluster. This approach is useful if we know there is structure in our data but do not know how many clusters exist—if we can find an appropriate threshold for density, we can discover the structure. Unlike other clustering methods, this approach does not necessarily require specifying a number of clusters $K$ in advance—though some algorithms do incorporate $K$, we do not consider them in this thesis.

The first density clustering algorithm was Density Based Spatial Clustering of Applications with Noise (DBSCAN); developed in 1996, it became the basis of many other algorithms[62, 63]. DBSCAN formalized the intuition that clusters were spaces that were dense with data, while areas that were relatively sparse were noise. For each point in a cluster, the neighbourhood of a given radius $\epsilon$ must contains at least $m$ points. Any points that meet this criterion, or fall within the neighbourhood, are assigned to clusters, while the rest are deemed noise. However, this requires a global density $\epsilon$ threshold—clusters of variable densities are not considered.

**HDBSCAN($\hat{\epsilon}$)**

A hierarchical version (HDBSCAN) was developed in 2013, capable of handling variable densities and only requiring the minimum number of points $k$ in a cluster as a parameter[64]. HDBSCAN uses single-linkage clustering, an agglomerative clustering algorithm that iteratively takes the closest points in space and connects them. Since this is sensitive to noise (e.g. a single point may act as a "bridge" between two clusters), HDBSCAN transforms the distance between points. The core distance $d_c$ of a point $x$ is the distance to the $k^{\text{th}}$ nearest neighbour of $x$, and the minimum reachability distance between two points is

$$d_{MRD}(x_p, x_q) = \max\{d_c(x_p), d_c(x_q), d(x_p, x_q)\}$$

The distance $d_{MRD}$ helps to separate density (i.e. clusters) from noise; it pushes points apart by at least their core distance, keeping points in dense areas close to each other and pushing sparse points

farther from each other. It builds a minimum spanning tree from $d_{MRD}$, which is the hierarchy of candidate clusters.

HDBSCAN defines the density $\lambda$ at a local level: $\lambda = \frac{1}{\epsilon}$, where $\epsilon$ is the distance from a point to its $k^{\text{th}}$ neighbour. To create a set of non-overlapping clusters from the hierarchy, it uses the excess of mass method, which maximizes stability across all clusters (explained in full detail in [60] and [65]). Essentially, it traverses the minimum spanning tree from the root onwards and records the value of $\lambda$ whenever a cluster forms ($\lambda_{born}$) or loses a point ($\lambda_p$). When a point leaves a cluster $C$, we record the stability $(\lambda_p - \lambda_{born})$, and we sum this up as each point leaves each cluster, i.e. $\sum_{p \in C} \lambda_p - \lambda_{born}$. The final clustering is the one that maximizes stability across all clusters and points—dense regions with lots of points are considered clusters.

In certain cases the data may contain both large populations with very high density and smaller populations with relatively low density. If $k$ is too high, HDBSCAN will label small populations as noise; if $k$ is too small, HDBSCAN will create numerous micro-clusters with many points in between labelled as noise. This led to the development of HDBSCAN($\hat{\epsilon}$), which assumes that the data set contains some very large and dense clusters as well as small, less dense clusters[65]. It imposes a threshold on $\epsilon$, i.e. clusters with $\lambda_{born} \leq 1/\hat{\epsilon}$ have their stability set to $0$. In practice, this allows us to have a small value of $k$, meaning we can identify small and large clusters of structure.

This latter method is key to Chapter 4, where we pre-process biobank data with UMAP and apply HDBSCAN($\hat{\epsilon}$) to extract clusters. The sizes of populations in many biobanks vary widely—

it is not unusual for a biobank to have a few ancestrally-related populations making up, e.g., $90\%$ of the cohort with the remaining $10\%$ consisting of several small populations whose genetic ancestries differ from the majority and from each other because of immigration history, admixture, highly-structured demographic history (as in the case of islands), etc. The UKB, for example, has $488,377$ genotyped individuals; well over $90\%$ claim European ancestry, but the remaining groups still have thousands of individuals from a diverse set of backgrounds.

## 1.4 Rationale and objective of research

The biobank space is rapidly growing with large and diverse cohorts being announced and released on a regular basis. Methods like GWAS and PGS are now standard and studying genetic ancestry is common both in scientific circles and in consumer-facing products. We require new tractable methodologies that can bolster our understanding of population structure in these data sets, which continue to grow in size and complexity. We also must understand the interplay between genetics, phenotypes, and the environment.

Thus the main objectives of this thesis are:

- To apply topological data analysis to human genetic data
- To use nonlinear dimensionality reduction and density clustering to characterize population structure in a variety of biobanks, focusing on using UMAP and HDBSCAN($\hat{\epsilon}$)
- To provide a framework for a new methodology for population genetics and demonstrate its

utility to the field

- To relate population structure to other data in biobanks, such as, phenotypic information, geographic coordinates, population labels, demographic history, genetic ancestry, and other variables

- To study the implications of our results to GWAS and PGS

# Part II

# Original contributions to knowledge

*No catalog of techniques can convey a willingness to look for what can be seen, whether or not anticipated. Yet this is at the heart of exploratory data analysis. The graph paper—and transparencies—are there, not as a technique, but rather as a recognition that the picture-examining eye is the best finder we have of the wholly unanticipated.*
*—John W. Tukey (1980)*

# Chapter 2

## 2.0 Preface

In Chapter 2, we apply UMAP to population genetic data for the first time. Until this time, dimensionality reduction in population genetics was largely limited to PCA, with the occasional foray into methods like t-SNE. We provide an in-depth analysis and comparison of PCA, t-SNE, and UMAP on genotype data from three biobanks: the 1KGP, the HRS, and the UKB.

We explore a variety of visualization methods and illustrate the relative strengths of UMAP as well as its limitations compared to other methods. We use UMAP to reduce our data to 2 dimensions and uncover fine-scale population structure in each of our data sets and colour it with sociodemographic data, geographic coordinates, phenotype distributions, admixture estimates, and other variables to reveal intricate patterns. We use UMAP to reduce our data to 3 dimensions and translate this from $(x, y, z)$ coordinates to $(R, G, B)$ values to show how to use topological data analysis to reveal spatial gradients in population structure.

This manuscript became the basis of several UMAP analyses by other researchers in a wide variety of contexts. It is now standard for new biobanks to publish a UMAP plot of their population structure. This manuscript was released as a preprint on *BioRxiv* in 2018 and published in *PLoS Genetics* in 2019.

# UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts

Alex Diaz-Papkovich[1,2], Luke Anderson-Trocmé[2,3], Chief Ben-Eghan[2,3], Simon Gravel[2,3,*]

[1]Quantitative Life Sciences, McGill University, Montreal, Québec, Canada

[2]McGill University and Genome Quebec Innovation Centre, Montreal, Québec, Canada

[3]Department of Human Genetics, McGill University, Montreal, Québec, Canada

* Corresponding author: simon.gravel@mcgill.ca

Published in *PLoS Genetics* 15.11 (2019): e1008432.

## 2.1 Abstract

Human populations feature both discrete and continuous patterns of variation. Current analysis approaches struggle to jointly identify these patterns because of modelling assumptions, mathematical constraints, or numerical challenges. Here we apply uniform manifold approximation and projection (UMAP), a non-linear dimension reduction tool, to three well-studied genotype datasets and discover overlooked subpopulations within the American Hispanic population, fine-scale relationships between geography, genotypes, and phenotypes in the UK population, and cryptic structure in the Thousand Genomes Project data. This approach is well-suited to the influx of large and diverse data and opens new lines of inquiry in population-scale datasets.

## 2.2 Author summary

The demographic history of human populations features varying geographic and social barriers to mating. Over time, these barriers have led to varying levels of genetic relatedness among individuals. This population structure is informative about human history, and can have a significant impact on studies of medical genetics. Because population structure depends on myriad demographic, ecological, and social forces, a priori visualization is useful to identify subtle patterns of population structure. We use a dimension reduction method—UMAP—to visualize population structure in three genomic datasets and find previously unobserved patterns, revealing fine-scale population structure and illustrating differences between groups in traits such as white blood cell count, height, and FEV1, a measure of lung function. Using UMAP is computationally efficient and can identify fine-scale population structure in large population datasets. We find it particularly useful to reveal phenotypic variation among genetically related populations, and recommend it is a complement to principal component analysis in primary data visualization.

## 2.3 Introduction

Questions in medicine, anthropology, and related fields hinge on interpreting the deluge of genomic data provided by modern high-throughput sequencing technologies. Because genomic datasets are high-dimensional, their interpretation requires statistical methods that can comprehensively condense information in a manner that is understandable to researchers and minimizes the amount

of data that is sacrificed. Both model-based and model-agnostic approaches to summarize data have played important roles in shaping our understanding of the evolution of our species (e.g., [1–5]).

Here we will focus on nonparametric approaches to visualize relatedness patterns among individuals within populations. If we consider unphased single nucleotide polymorphism (SNP) data, an individual genome can be represented as a sequence of integers corresponding to the number of copies of the alleles carried by the individual at each of the $L$ SNPs for which genotypes are available, with $L$ ranging from hundreds of thousands to hundreds of millions. Since each individual is represented as an $L$-dimensional vector, dimension reduction methods are needed to visualize the data.

Principal component analysis (PCA) is often the first dimensional reduction tool used for genomic data. It identifies and ranks directions in genotype space that explain most-to-least variance among individuals. Positions of individuals along directions of highest variance can then be used to summarize individual genotypes. PCA coordinates have natural genealogical interpretations in terms of expected times to a most recent common ancestor (TMRCA) [6], and are used empirically to reveal admixture [7], continuous isolation-by-distance [8, 9], as well as technical artefacts. PCA coordinates are particularly well-suited to correct for population structure in GWAS[4].

The amount of information encoded in the highest-variance PCs increases slowly with sample size, so researchers typically examine multiple two-dimensional projections to lower-variance PCs to explore data. In this process, finer features of the data may be hidden by the projections or

hard to interpret. To display finer features of the data in a two dimensional figure, we can use non-linear transformations that emphasize the local structure of the data. A popular method for such visualization is t-distributed stochastic neighbour embedding (t-SNE)[10]. t-SNE has been used before to visualize SNPs[11]. Using data from the 1000 Genomes Project (1KGP)[12], it groups individuals corresponding roughly to their continent of origin, with smaller ethnic subgroups visible within the larger continental clusters[13]. However, t-SNE struggles with very large datasets, when a large number of locally optimal configurations make convergence to a globally satisfying solution difficult.

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique designed to model and preserve the high-dimensional topology of data points in the low-dimensional space[14]. With genotype data, UMAP creates a neighbourhood around each individual's genetic coordinates and identifies a pre-selected number of neighbours to build high-dimensional manifolds. The end result is a patchwork of low-dimensional representations of neighbourhoods that groups genetically similar individuals together on a local scale while better preserving long-range topological connections to more distantly related individuals. The method has been successfully applied to single-cell RNA sequencing datasets [15].

Non-linear dimension reduction methods tend to be computationally intensive. A common practice to reduce this burden is to first apply PCA to data, and perform dimensional reduction on data projected to leading principal components (PCs). In addition to being computationally advantageous, this discards noise that can confound non-linear approaches: population structure arising

from $n$ isolated randomly-mating demes can be described by the leading $n-1$ PCs, with the following PCs describing stochastic variation in relatedness [4]. Selecting the leading PCs therefore has potential to extract meaningful population structure while filtering out stochastic noise. We explore different strategies to pre-process the data and investigate discrete and continuous population structure patterns present in large datasets of human genotypes: the 1KGP, the Health and Retirement Study (HRS)[16], and the UK BioBank (UKBB)[17], and compare UMAP's performance to t-SNE.

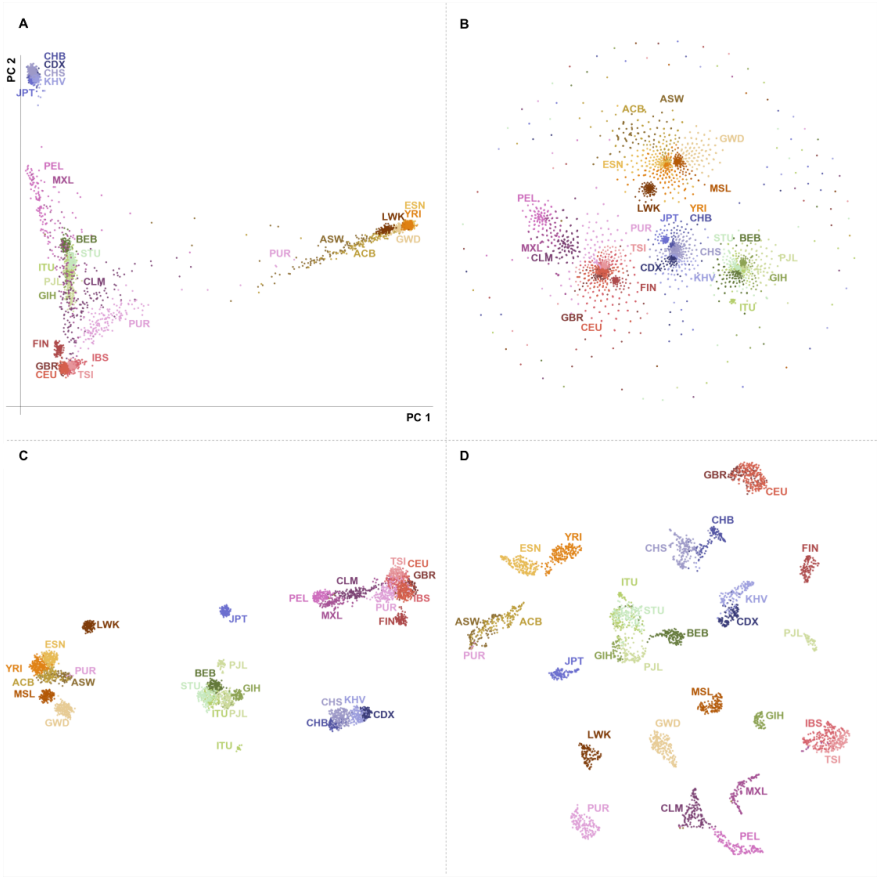



**Figure 2.1: Four methods of dimension reduction of 1KGP genotype data with population labels.** **(A)** PCA maps individuals in a triangle with vertices corresponding to African, Asian, and European continental ancestry. Discarding lower-variance PCs leads to overlap of populations with no close affinity, such as Central and South American populations with South Asians. **(B)** t-SNE forms groups corresponding to continents, with some overlap between European and Central and South American people. Smaller subgroups are visible within continental clusters. The cloud of peripheral points results from the method's poor convergence. **(C)** UMAP forms distinct clusters related to continent with clearly defined subgroups. Japanese, Finnish, Luhya, and some Punjabi and Telugu populations form separate clusters consistent with their population history[12]. **(D)** UMAP on the first 15 principal components forms fine-scale clusters for individual populations. Groups closely related by ancestry or geography, such as African Caribbean/African American, Spanish/Italian, and Kinh/Dai populations cluster together. Results using t-SNE on principal components are presented in Figure 2s1. Axes in UMAP and t-SNE are arbitrary. Since the algorithms prioritize local distances, long distances between clusters are not meaningful.

## 2.4 Results

### 2.4.1 Fine-scale visualization of the 1KGP dataset

The 1KGP contains genotype data of 3,450 individuals from 26 relatively distinct labeled populations[12]. Figure 2.1 shows visualizations using PCA, t-SNE, UMAP, and UMAP with PCA pre-processing. Using UMAP and t-SNE on the genotype data presents clusters that are roughly grouped by continent, with UMAP showing a clear hierarchy of population and continental clusters, whereas t-SNE fails to assign many individuals to population clusters. Using either method on the top principal components leads to distinct population clusters and less defined continental structure. Adding more components results in progressively finer clusters until approximately 20 populations appear using 15 components; adding further components converges to results similar to using the entire genotype data (see Figures 2s1, 2s2, and 2s3). To investigate the population information contained in low-variance PCs, we performed UMAP on data projected onto PCs 100 to 3450 (i.e., without information about the leading 99 PCs). Figure 2s4 shows that population structure is still clearly visible.

Focusing on UMAP with the leading 15 principal components (Figure 2.1D), several population clusters reflect shared ancestries. British individuals from England and Scotland form a cluster mixed with those from Utah who claim Northern and Western European ancestry. Toscani and Iberian individuals form a cluster reflecting their Mediterranean heritage. African Americans in the Southwest US, African Caribbean individuals in Barbados, and some Puerto Ricans also form a cluster. Three East Asian clusters appear: one is largely Han and Southern Han individuals,

another is comprised of the Chinese Dai in southern China and the Kinh from Vietnam, and the third is the Japanese population. Other clusters are comprised of Colombians and Peruvians, the Esan and Yoruba populations of Nigeria, and several South Asian populations.

Within population clusters, family members were projected near each other within broader population groups. When UMAP was parameterized to use only 5 nearest neighbours, however, families often formed distinct clusters (Figure 2s5): Using few neighbours to build a manifold emphasizes closer relatedness.

A few individuals cluster with populations different from their label: some Mexican individuals cluster with Spanish and Italian populations; some Puerto Rican individuals cluster with African American and Caribbean populations; and one Gambian individual clusters with the Mende of Sierra Leone. Two populations form multiple clusters: Gujarati Indians in Houston, Texas and Punjabi people in Lahore, Pakistan. This clustering effect is robust to the number of components considered (Figure 2s2). Differentiation in the 1KGP Gujarati population has been previously identified through a PCA restricted to the Gujarati [18]. Following a preprint version of the present article, 23andme released a statement [19] arguing that one of the two clusters could be traced, via 23andMe participant recontact, to individuals from a group in Western India with shared ancestry and patronym[D. Poznik, 23andMe, personal communication][19].

### 2.4.2 Admixed individuals fall along a genetic continuum

The 1KGP sampled individuals from relatively distinct populations, so the data are more likely to form clusters. Most medical cohorts, however, comprise larger numbers of individuals sampled across extended geographical areas. The HRS contains genotype data of 12,454 Americans from a variety of backgrounds. Using UMAP on the first 10 principal components, we demonstrate projections that present a collection of sub-populations and a continuum of genetic variation.

The HRS forms several large clusters, reflecting both ethnicity (Figure 2s6) and admixture proportions (Figure 2s7). Gradients in admixture proportion are visible within the predominantly Hispanic cluster, but not within the predominantly Black cluster, perhaps because the variance in ancestry proportions is greater among Hispanics. The "White Not Hispanic" (WNH) group forms several interconnected clusters, and these do not correspond to broad geographical areas (Figure 2s8). By generating the PC axes and UMAP embedding for the HRS data in Figure 2s6, and projecting the 1KGP data onto it, we reveal substructure within the Hispanic cluster, groupings of Finnish individuals within the WNH groups, as well as Italian and Spanish individuals grouping near the White Hispanic population (Figure 2s9). One group of WNH individuals regularly appears at the periphery of the main cluster and does not cluster with any 1KGP populations.

### 2.4.3 Regional patterns in the Hispanic subpopulation

Applying UMAP to self-identified Hispanic individuals in the HRS reveals clear groupings related to birth region in Figure 2.2A. The highlighted cluster consists almost entirely of individuals

born in the Mountain Region of the United States. This cluster is not apparent when looking at a grid of pairwise plots of the first 8 principal components, provided in Figure 2s10, as the signal is distributed along PCs 3, 4, and 6. Even though continental admixture patterns do correlate with UMAP position (Figure 2s11), these do not explain the Mountain Region cluster. Individuals from 1KGP populations do not appear in the cluster when projected to the UMAP embedding. The cluster possibly comprises the Hispano/Nuevomexicano population of the Southwest US, who have been present in the Mountain Region area long before the more recent immigrants from Latin America, and whose ancestry is expected to reflect both distinct Native ancestry and population-specific drift relative to other Hispanic populations. Such a cluster has been previously identified in AncestryDNA data using network-based clustering on identity-by-descent connections[20]; a recent preprint discusses the Mountain Region Hispanics with a more detailed historical description[21].


**Figure 2.2: Applying UMAP to subsets of data can reveal deep population structure. (A)** UMAP on the top 7 principal components of the self-identified Hispanic population of the HRS reveals a cluster. Colouring the points by birthplace shows they were born almost entirely in the Mountain region (in green) of the United States (New Mexico, Arizona, Colorado, Utah, Nevada, Wyoming, Idaho, and Montana). When populations from the 1KGP are projected onto the UMAP embedding they do not map to the cluster. Six 1KGP populations are presented: CLM, Colombian in Medellin, Colombia; IBS, Iberian in Spain; MXL, Mexican in Los Angeles, California; PEL, Peruvian; PUR, Puerto Rican; TSI, Toscani in Italy. Figures 2s11 and 2s12 present the same projection of individuals from the HRS coloured by estimated admixture proportions census region of birth, respectively. **(B)** UMAP on the top 8 principal components of the self-identified Asian populations of the UKBB creates clusters. Indian individuals born in Kenya (in purple) form one such cluster. A version coloured by self-identified ethnicity is presented in Figure 2s13.

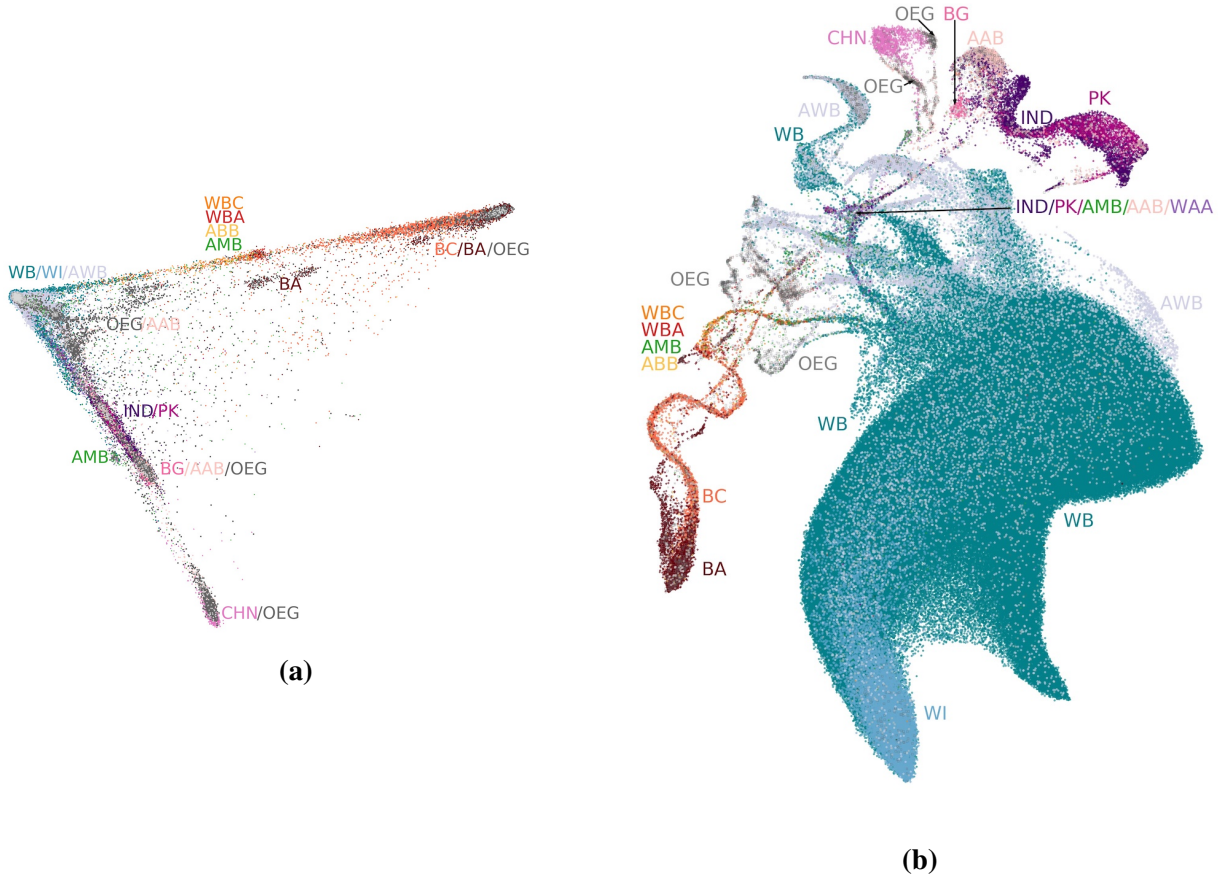



**Figure 2.3: The UKBB coloured by self-reported ethnic background.** **(a)** The first two principal components, showing the usual triangle with vertices corresponding to African, Asian, and European ancestries, and intermediate values indicating admixture or lack of relationship to the vertex populations. **(b)** UMAP on the first 10 principal components. The cluster of White British and White Irish individuals is greatly expanded, with the Irish forming a distinct sub cluster mixed with the White British population. South Asian and East Asian individuals form their separate clusters, as do individuals of African or Caribbean backgrounds. Population clusters are connected by "trails" comprised of large proportions of individuals with mixed backgrounds. BA, Black African; BC, Black Caribbean; BG, Bangladeshi; CHN, Chinese; IND, Indian; PK, Pakistani; WB, White British; WI, White Irish; WBC, White and Black Caribbean; WBA, White and Black African; WAA, White and Asian; AAB, Any other Asian Background; ABB, Any other Black Background; AWB, Any other White Background; AMB, Any other Mixed Background; OEG, Other ethnic group.

### 2.4.4 Population structure in the UKBB reflects local and global genetic variation

The UKBB contains data on 488,377 individuals including genotypes, phenotypic measures and self-identified ethnic backgrounds. Figure 2.3 compares UMAP to PCA applied to the UKBB. As expected, PCA captures major axes of variation emphasizing continental ancestry, whereas UMAP reveals finer structure. UMAP on the top 10 principal components reveals continuous and discrete population structure (Figure 2.3B): the patchwork of local topologies identifies multiple sub-populations, as well as continuous structure within populations and admixture gradients between populations. The result is a succinct illustration of the complex structure and population relationships in a large and multi-ethnic dataset.

The largest cluster in Figure 2.3B consists of the White British and Irish populations. The Irish population forms a sub-cluster, but many individuals are also scattered throughout the British-identifying population. Individuals identifying as Black African and Black Caribbean partially overlap, but admixed individuals form distinct trails leading to Asian and European clusters. Chinese individuals form a cluster, within a broader East Asian population; Indian, Pakistani, and Bangladeshi populations form a closely bound cluster as well. The East Asian and South Asian populations each have large clusters of individuals who identify as having an "other Asian background" or belonging to an "other ethnic group". The patchwork of genetic neighbourhoods is connected by trails of admixed individuals, which converge at a nexus of individuals with a variety of ethnicities. Many claim mixed ancestry, and there are clusters of individuals who belong to an "other ethnic group". Using data on countries of birth, we identified many finer groups in Fig-

ure 2s14, and confirmed they appeared in intuitive areas with, e.g., Japanese and Filipino clusters being projected near Chinese clusters.

Figure 2.4 presents the UMAP projection from Figure 2.3B coloured instead by geographical coordinates from the Ordnance Survey National Grid (OSGB1936), with distances defined as a north or east position relative to the Isles of Scilly. Geographic clusters form in the large White British grouping, reflecting the relationship between genetic and geographic distance, as has been observed in Europe and British-wide data[8, 22]. Figure 2.3B shows that the admixed individuals have UMAP coordinates next to White British individuals residing in the South East of the UK, where London is located (see also Figure 2s15, where individuals are coloured by distance from London). This likely reflects the high migration levels to the city and surrounding area: the UMAP projection attempted to preserve both the genetic similarity among admixed individuals and the relatedness with White British individuals in cosmopolitan areas.

The detailed shape of extended clusters is not stable as we vary the number of PCs included, but the patterns mentioned above are preserved. Figures 2s16, 2s17, and 2s18 show UMAP plots using the top $40$ PCs from the UKBB.

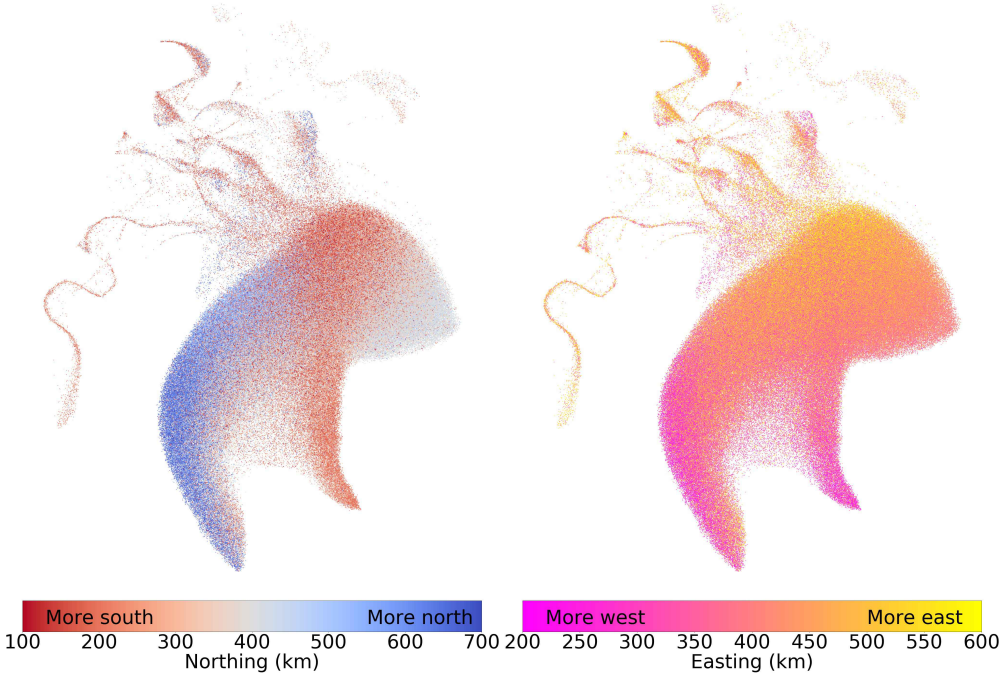
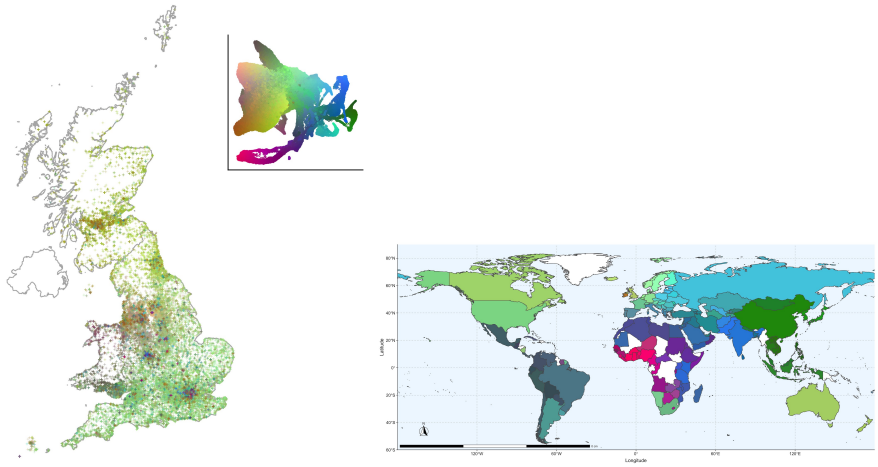



**Figure 2.4: UMAP captures relationships between population structure and geography.** Each individual is coloured by their geographical coordinates of residence. Coordinates follow the UKBB's OSGB1936 geographic grid system and represent distance from the Isles of Scilly, which lie southwest of Great Britain. The left image colours individuals by their north-south ("northing") coordinates, and the right image colours them by their east-west ("easting") coordinates. Adding more components creates finer clusters (Figures 2s17 and 2s18). Northing values were truncated between 100km and 700km, and easting values were truncated between 200km and 600km.



**(a)** **(b)**

**Figure 2.5: Maps coloured by 3D UMAP projections of the top 20 principal components of the UKBB.** Each individual is assigned a 3D RGB vector based on 3D UMAP coordinates (a flattened projection is in the top right of panel A). Individuals who are closer to each other in the projection will be closer in colour in the maps. More details on colouring, as well as randomization of points to protect participant privacy, are available in the materials and methods. **(a)** Each point is an individual placed based on where they live. Patterns in genetic similarity are visible in Scotland, South England, North and South Wales, the East and West Midlands, and major urban centres. **(b)** Geographic distribution of UMAP coordinates. Using the country of birth of individuals in the UKBB, we colour countries by the closeness in 3D UMAP space of those born there. Broad patterns of similarity appear in East Asia, South Asia, North African and the Middle East, West Africa, and South America. Differences between neighbouring countries can reflect both ancient population structure and recent differences in migration history. Evidence of migrations related to colonialism are visible with, e.g., European ancestry in South Africa and South Asian ancestry in Kenya and Tanzania. Because of the large number of White British individuals born abroad, to avoid skewing the colour scale they were not included unless they were born in the UK, Europe, Australia, Canada, or the United States, where UKBB participants already tended to have European ancestry. Zoomed maps of East Asia, the Caribbean, and Europe are available in Figures 2s19, 2s20, and 2s21, respectively.

As an alternate visualization of geography and genetic diversity, we performed a 3D UMAP projection and converted the UMAP coordinates into RGB values, allowing us to plot individuals on a map of Great Britain, emphasizing both spatial gradients of genetic relatedness and increased diversity in urban centers (Figure 2.5A). The geographical patterns outside major urban centers are similar to those reported in [22] using the haplotype-based CHROMOPAINTER on British individuals whose grandparents lived nearby. Using data about country of birth, we performed a similar analysis of a world map in Figure 2.5B, revealing subtle regional variation around the world.

Similarly to UMAP, t-SNE applied to the UKBB data both displays diversity within the "White British" population and identifies clusters among other groups. However, it has three drawbacks: it is much slower, requiring 2.26 hours for its first thousand iterations alone on 10 principal components against UMAP's 14 minutes; it fails to find a global optimum, which results in a scattering of individuals and groups that are not stable across independent runs; and it does not identify continuity between different continental groups resulting from admixture (Figure 2s22).

### 2.4.5 Identifying patterns in phenotype variation related to genetic population structure

Covariates such as height and leukocyte count (Figure 2.6) and autoimmune and asthma-related measures (Figures 2s23 to 2s34) correlate strongly with both discrete and continuous population structure. Several populations in Figure 2.6, including South Asian, East Asian, African, and others have noticeably lower-than-average heights. More subtle patterns are also visible: the area

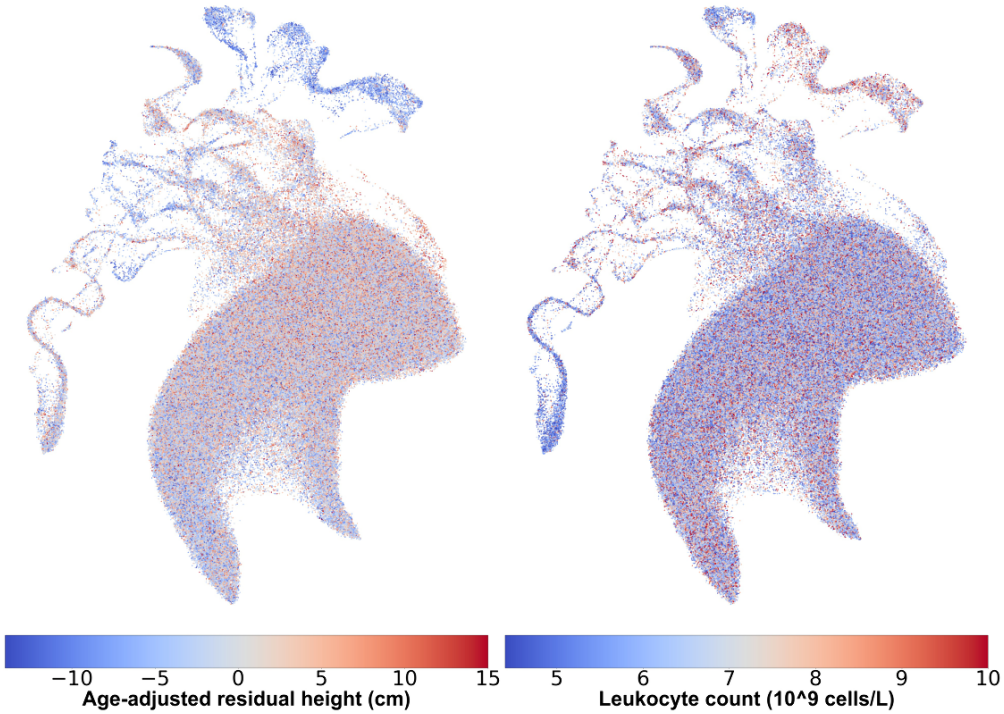



**Figure 2.6: UMAP captures relationships between population structure and phenotype heterogeneity.** Females from the UMAP projection in Figure 2.3B, coloured by age-adjusted difference from mean population height (left) and leukocyte counts (right). Individuals with missing data were excluded. To protect participant privacy, data in these images has been randomized as explained in the materials and methods section.

of the projection in Figure 2.3B with the cluster of White Irish people appears more blue than the main body of White British individuals. To quantify and statistically test these qualitative observations, we performed an unpaired two sample t-test of self-identified White Irish and White British individuals and found British males taller on average by 0.846cm (p-value $2.10 \times 10^{-23}$) and British females by 0.763cm (p-value $3.65 \times 10^{-23}$) (see Figures 2s35 and 2s36 for boxplots). Height differences between Irish and British populations have been previously observed but the direction of the difference is not consistent[23, 24].

In clinical settings, baseline Forced Expiratory Volume in 1 second (FEV1) is determined via equations that include ethnicity or race[25], but studies in several populations have shown that there is considerable variation based on ancestry, even within self-defined ethnicity[26]. Figures 2s27 and 2s28 show strong correlations with genetic clustering: certain populations—South Asian, African, and Caribbean—have considerably lower measurements on average (see Figures 2s37 and 2s38 for boxplots and p-values).

Notably, there appears to be a juncture in the admixed population, highlighted in Figure 2s39, where the distribution of FEV1 changes. This roughly corresponds to the transition from Black African/Caribbean individuals to those who identified having mixed backgrounds. Boxplots and statistical testing suggest that relative to White British populations, FEV1 values are significantly lower for Black African and Black Caribbean populations, but not for White and Black Caribbean and White and Black African populations (Figures 2s37 and 2s38).

Figure 2s40 further suggests a difference in FEV1 between those who self-identified as Chi-

nese and a nearby cluster enriched in individuals born in Japan; to our knowledge there have not been studies into differences in FEV1 between these populations. To focus on individuals of Asian ancestry (rather than, e.g., individuals born in Japan but who have European ancestry), we first selected the individuals whose UMAP coordinates were near the Chinese cluster. We then focused on individuals born in Japan, Malaysia, and the Philippines as well as the self-identified Chinese population. These four groups are mutually exclusive and are shown in Figure 2s41. After adjusting for age, age$^2$, height, and sex, an unpaired two-sample t-test shows those born in Japan have a higher mean FEV1 than Chinese individuals by 0.224 (p-value $2.787 \times 10^{-15}$). By sex, there is a difference of 0.213 (p-value $5.4 \times 10^{-13}$) among females and 0.317 (p-value $5.1 \times 10^{-4}$) among males, though there are considerably fewer males in the sample (distributions presented in Figure 2s42). For comparison, the adjusted difference between self-identified African and British individuals in the UKBB is 0.762 (p-value $2.2 \times 10^{-16}$).

### 2.4.6 Comparing t-SNE and UMAP

Identifying the best dimension reduction technique is challenging, both because the "best" representation depends on context, and because convergence issues may mean that a good theoretical model for dimensional reduction might perform poorly because of challenges in numerical optimization. To assess whether the relatively poor performance of t-SNE could be due to convergence rather than a flawed model, we used UMAP to preprocess the UKBB data and provide a starting point to a standard t-SNE implementation. This led to representations that were objectively bet-

ter (according to the t-SNE metric) than the default t-SNE implementation (Figure 2s43). Yet, these representations were much less detailed than the UMAP embedding provided as a starting point (Figure 2s44). Given these results, we recommend UMAP over t-SNE for large and diverse genomic datasets.

## 2.5 Discussion

Methods such as UMAP and t-SNE focus on preserving local distances to reveal fine-scale structure in populations, and in the process may preserve aspects of global structure as well. In contrast, PCA preserves long range distances but hides finer-scale details. Hierarchical clustering of networks has also successfully detected fine-scale population structure using identity-by-descent similarity by attempting to preserve relations between global networks and smaller local ones (e.g., [20]). We speculate that the addition of weak constraints favouring the preservation of longer distances in UMAP-like approaches has the potential of preserving the desirable local properties while encouraging more intuitive positioning of clusters on a global scale.

UMAP comprehensively illustrates genotypic information at fine scales and within the context of global population structure. It is easy to use and fast: given PCA data and a desktop computer, UMAP can be performed in 15 to 25 minutes on a sample of hundreds of thousands of individuals over tens of dimensions. It excels with larger datasets containing individuals with admixed backgrounds, which present discrete and continuous population structure.

Using UMAP reveals clusters that would have been difficult to identify via pairwise PCA plots

or Admixture analysis, such as the geographically restricted cluster within the Hispanic population of the HRS, or the splits within the Gujarati and Punjabi population samples in the 1KGP. More importantly, UMAP helps reveal patterns of covariation between geography, phenotypes and genotypes. Traits such as height showed continuous variation across admixture edges and geographic gradients, as expected from genetically controlled complex traits, and others, such as leukocyte counts or FEV1, showed sharper boundaries and non-linear behavior consistent with the existence of strong regional environmental influences.

We found that pre-processing the data with PCA allowed for time savings, but identifying an optimal number of PCs to use is challenging. Groupings on ethnicity formed slowly as PCs were added until reaching a stable number around 10 to 15 PCs. Geographical patterns in the UKBB continued to appear even up to 40 components, as visible in Figures 2s17 and 2s18.

### 2.5.1 Caveats

In contrast to PCA, UMAP has more adjustable parameters. Changing the PC cutoff, minimum distance, and number of neighbours can change characteristics of the visualizations. Using a minimal number of neighbours (e.g. 5 rather than the default 15) can result in the formation of disjoint clusters comprised of related family members (Figure 2s5), and using a low minimum distance (e.g. 0.001 rather than the default 0.1) can result in clusters becoming more compact, losing visual detail. We used a minimum distance of 0.5 and 15 neighbours; however, default suggested parameters in UMAP generally perform well across datasets.

In the absence of clear theoretical rationales, we suggest to use as many PCs as are available and computationally feasible, even though we sometimes found that a lower number of PCs led to a simpler shape that facilitated discussion (e.g. Figure 2.3B). Overall, we recommend reporting on a range of parameter values and following up on observations with statistical testing.

Like most non-linear methods, UMAP lacks direct interpretability. It emphasizes local distances over global distances; while points that are very close in UMAP space are likely close in the original data, points that are distant in UMAP space are not necessarily very different in the original data. Disconnected clusters may also change their positions relative to other clusters over the course of multiple projections, as in Figure 2s2. For these reasons, UMAP coordinates should not be used as GWAS covariates or for quantifying distances between populations. UMAP is sensitive to sample sizes and spends more visual space on populations with larger sample sizes. This is useful to identify significant patterns in a cohort, but it makes comparing visualization across cohorts difficult and may appear to exaggerate the genetic variation within the most sampled populations, such as the White British population in the UKBB. We did not assign meaning to wiggles in UMAP figures, which occurred consistently in the UKBB but may be an artefact of the dimensional reduction strategy rather than a meaningful feature of the data. Hand-waving interpretations of pretty plots have a history of getting population geneticists in trouble (as pointed out, e.g., in [27]): visualization is not a replacement for statistical testing.

With these caveats in mind, a priori data visualization plays a central role in quality control, hypothesis generation, and confounder identification for a wide range of genomic applications.

Non-linear approaches, despite their limitations, become increasingly useful as the size of datasets increases. UMAP, in particular, reveals a wide range of features that would not be apparent using linear maps. Given its ease of use, broad applicability, and low computational cost, we propose that UMAP should become a default companion to PCA and other population structure visualization and inference methods in large genomic cohorts.

## 2.6 Materials and methods

We used genotype data from 12,454 individuals from the Health and Retirement Study (HRS), genotyped on the Illumina Human Omni 2.5M platform[16]. Principal components were computed in PLINK v1.90b5.2 64-bit[28] using variants with a minor allele frequency greater than 0.05, Hardy-Weinberg p-value of more than $1 \times 10^{-6}$, and genotype missing rate of less than $0.1$, and sample with genotype missing rate of less than $0.1$. We used the principal components of genotype data from 488,377 individuals in the UK BioBank (UKBB) as computed by the cohort [17]. We used genotype data from 3,450 individuals from the 1KGP project using Affy 6.0 genotyping[12].

The HRS contains genotype data of 12,454 American individuals across all 50 states who have provided racial identity (10,434 White, 1,652 Black, 368 Other) as well as whether they identify as Hispanic (1,203 total) and, if so, whether they identify as Mexican-American (705 total)[16]. We crossed these three variables to form a composite self-reported ethnicity resulting in 10 categories (e.g. White Hispanic Mexican-American), and considered birth regions based on the 10 census regions and divisions used by the US Census Bureau. Admixture proportions for each individual

were estimated in [29] by assuming ancestral African, Asian, and European populations using RFMIX [30]. We have scaled each of the three proportions to values between 0 and 255 (with 100% corresponding to 255), to colour individual points by their estimated admixture represented by RGB where red, green, and blue respectively correspond to African, European, and Asian/Native American ancestry. To project 1KGP data on HRS embeddings, as in Figures 2.2A and 2s9, we created the PC axes and UMAP embedding for the HRS data and then projected the 1KGP data onto them.

The UKBB provides genotype data on 488,377 individuals along with self-identified ethnic background in a hierarchical tree-structured dictionary. Participants provided ethnic background on two occasions. We used the initial ethnicity after finding minimal differences between the two. The dataset is majority White (88.3% British, 2.6% Irish, 3.4% other), with large populations identifying as Black (1.6% either African, Caribbean, or other), Asian (1.9% either Indian, Pakistani, Bangladeshi, or other), Chinese (0.3%), an other ethnic group (0.8%), mixed ethnicity (0.6%), or an unavailable response (0.5%).

Scripts for all tests and plotting functions can be found on `https://github.com/diazale/gt-dimred` with a command line script for UMAP available at `https://github.com/diazale/gt-dimred/scripts/general_umap_script.py`. A demo version using freely available 1KGP data is available at `https://github.com/diazale/1KGP_dimred`. PCA and standard t-SNE were done with Scikit-learn[31]. UMAP was performed using a Python implementation[14]. Statistical testing was done in SciPy[32], StatsModels[33], and R[34]. Vi-

sualizations were created with Matplotlib[35] and ggplot2[36], and maps were made with Natural Earth.

Both UMAP and t-SNE feature a number of adjustable parameters. Among the parameters that we varied, the number of PCs used in pre-processing of the data has the largest effect for both methods (see Figures 2s1 and 2s2). With UMAP, there are other parameters, such as the learning rate and the distance metric; these were left to the default values.

We tested different choices for perplexity in t-SNE. The default value of 30 provided comparable performance to other parameter choices. Similarly, we tested different parameter choices for UMAP, with the clearest results generated by specifying 15 nearest neighbours (the default value) and a "minimum distance" between points in low dimensions of 0.5. UMAP developers described "sensible" values for nearest neighbours as between 5 and 50 and minimum distance between 0.5 and 0.001. Tuning these parameters will not change qualitative results much but may make patterns easier to identify. Increasing the number of neighbours will increase the computational load, and a smaller minimum distance can break the connectivity between clusters, though the same individuals will continue to group together.

UMAP and t-SNE projections were carried out on an iMac with a 3.5GhZ Intel Core i7 processor, 32 GB 1600 MHz DDR3 of RAM, and an NVIDIA GeForce GTX 775M 2048 MB graphics card.

Colours for maps in Figures 2.5A, 2.5B, 2s19, 2s20, and 2s21 were determined by projecting data to 3D and using each 3D coordinate as an RGB coordinate. For the world map, countries

were determined using the country of birth variable, with a country's colour being determined by the mean $x$, $y$, and $z$ values of all individuals born in that country. Because many self-identified White British individuals were born abroad, including them everywhere would skew the colour scheme; they were included only if they were born in the UK, Europe, Australia, Canada, or the United States. This approach to colouring is sensitive to sample sizes as UMAP will give more space to larger populations.

To reduce the potential risks for re-identification from results in this publication, data has been randomly permuted so that the population characteristics are preserved but individual-level data is not presented directly in the figures. We rounded each attribute to an attribute-specific number of bins, and then permuted the data in the following way: For each point (i.e. each individual) in UMAP visualizations, and each attribute, we identified the 9 nearest neighbouring points, and copied the attribute from a randomly selected neighbour (thus allowing for the possibility of one value being printed twice). Because this process is done independently for each visualization, a given point shown on the figure will copy values from different randomly selected individuals. Projections coloured by participants' spatial coordinates have random noise added (normally distributed about 0 with a standard deviation of 50km) before binning to the nearest 50km. Projections coloured by participants' distance from London have random noise added (normally distributed about 0 with a standard deviation of 5km) before binning to the nearest km. For each point in Figure 2.5A we identified the nearest 50 neighbouring individuals and copied the colour value from a randomly selected neighbour.

### 2.6.1 Ethics statement

HRS data was accessed under IRB Study No. A11-E91-13B - The apportionment of genetic diversity within the United States. UKBB data was accessed under accession number 6728.

## 2.7 Acknowledgments

We thank all participants in the HRS, UKBB, and 1KGP for providing their genetic data as well as the teams who generated and assembled the datasets. We also thank Audrey Grant, Ryan Hernandez, Jose Sergio Hleap, Mark Lathrop, Dominic Nelson, Markus Munter, Stephen Sawcer, Melissa Spear, and Dara Torgerson for useful discussions about science, programming, and data access; David Poznik, Liz Babalola, and Adam Auton from 23andMe for discussing findings in the 1KGP; and Selin Jessa for introducing us to UMAP.

## 2.8 Supporting information

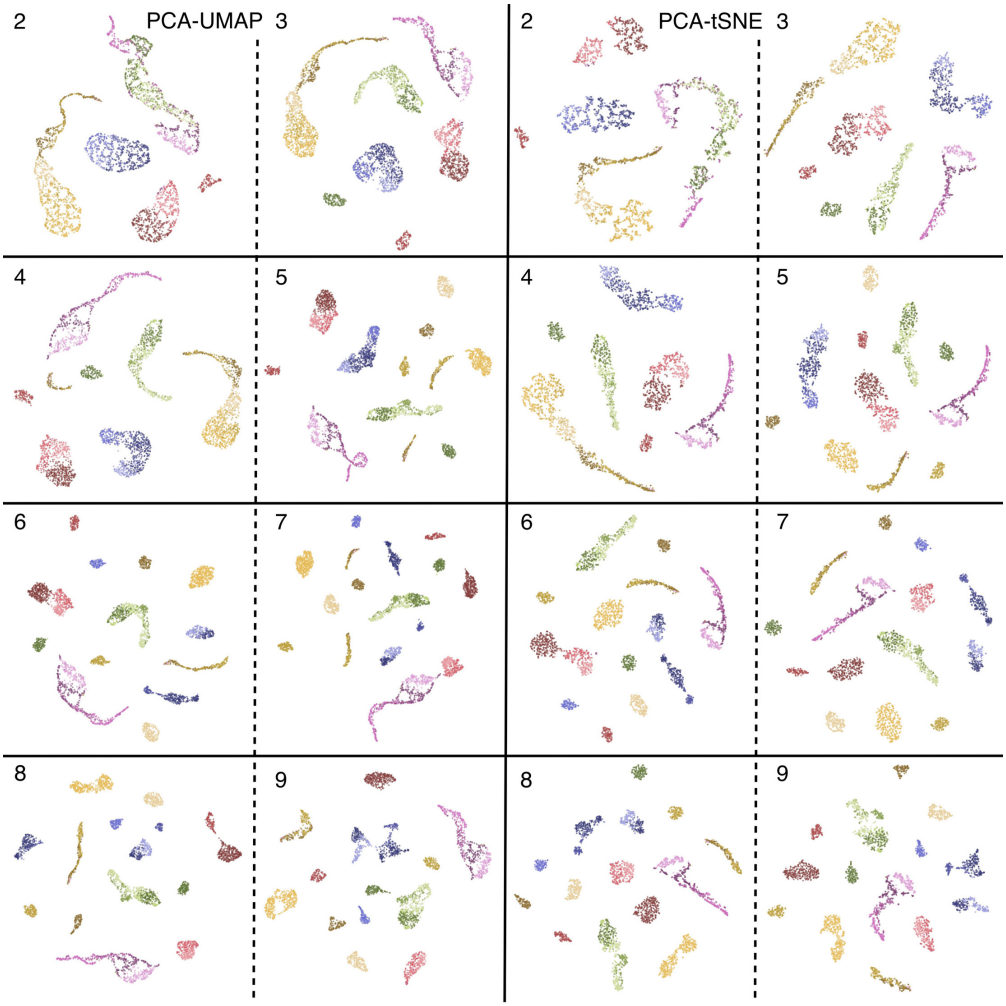



**Figure 2s1: Montage of $t$-sne and UMAP on up to 9 PCs of 1KGP data.** UMAP (left two columns) and $t$-sne (right two columns) applied to the top principal components of the 1KGP labelled by the number of components used. Adding more components results in progressively finer population clusters using both methods.

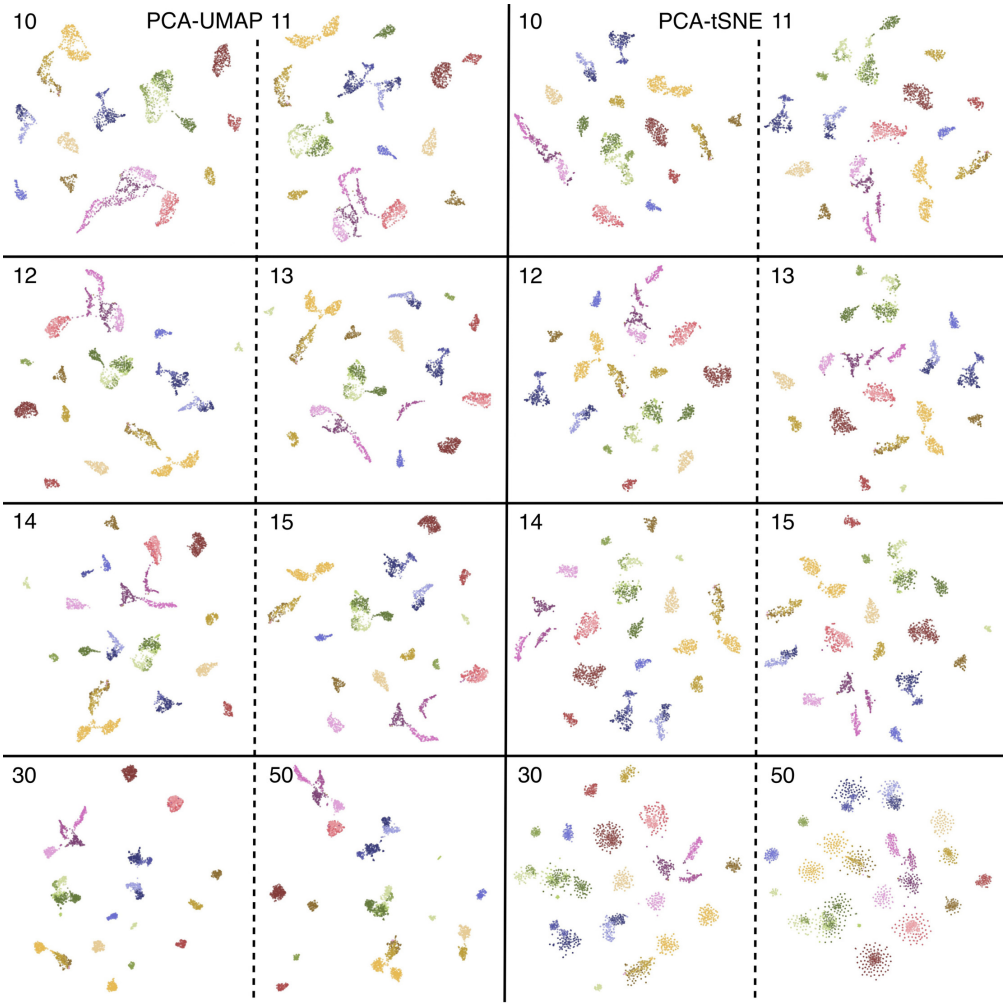



**Figure 2s2: Montage of $t$-sne and UMAP on 10 to 50 PCs of 1KGP data.** UMAP (left two columns) and $t$-sne (right two columns) applied to the top principal components of the 1KGP labelled by the number of components used. Results are similar until approximately 11 components, where $t$-sne breaks apart clusters of South Asian (in green) and Central and South American populations (in pink) while UMAP preserves them. At approximately 30 components populations begin to drift together with UMAP and disperse with $t$-sne.

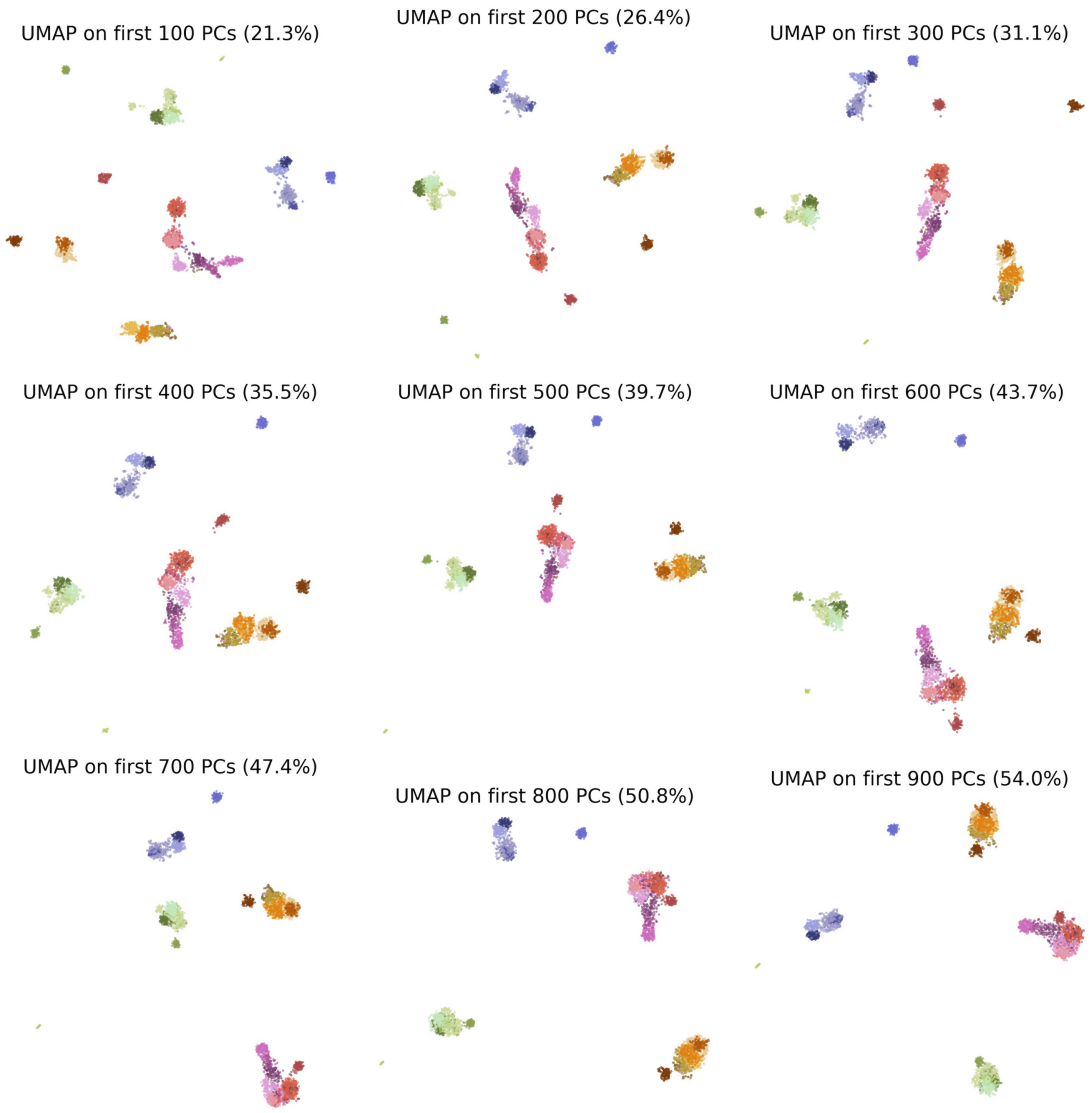



**Figure 2s3: Montage of UMAP on progressively more PCs of 1KGP data.** UMAP applied to the first few hundred principal components of the 1KGP data with the amount of variance explained in parentheses. As more components are added, the figure begins to resemble that of UMAP carried out on the full genotype dataset.

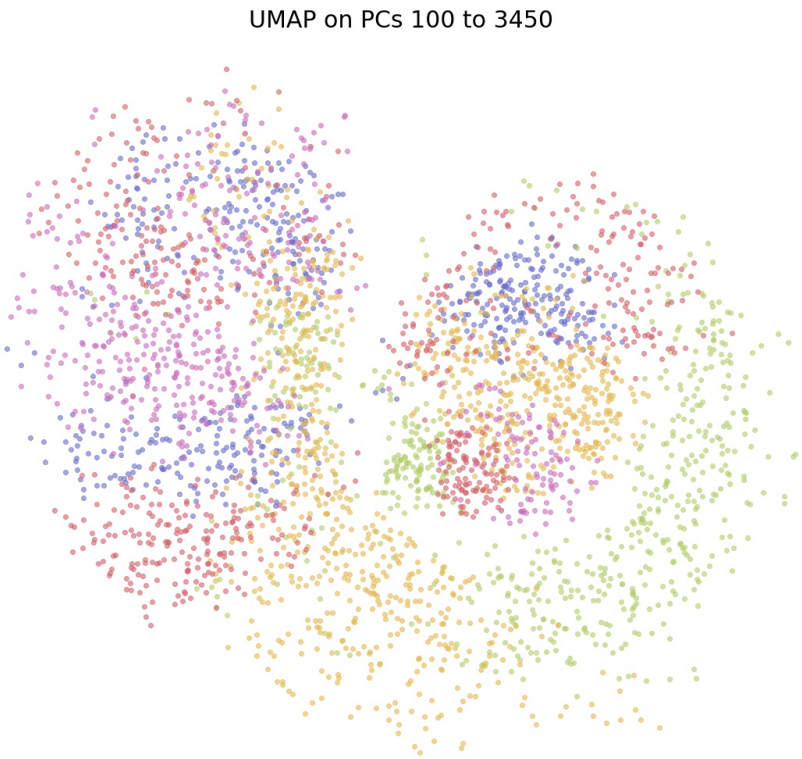



**Figure 2s4: UMAP on PCs 100 to 3350 of 1KGP data.** UMAP applied the last 3350 principal components of the 1KGP, which explain 78.7% of the variation. The colour scheme is the same as in 2.1.

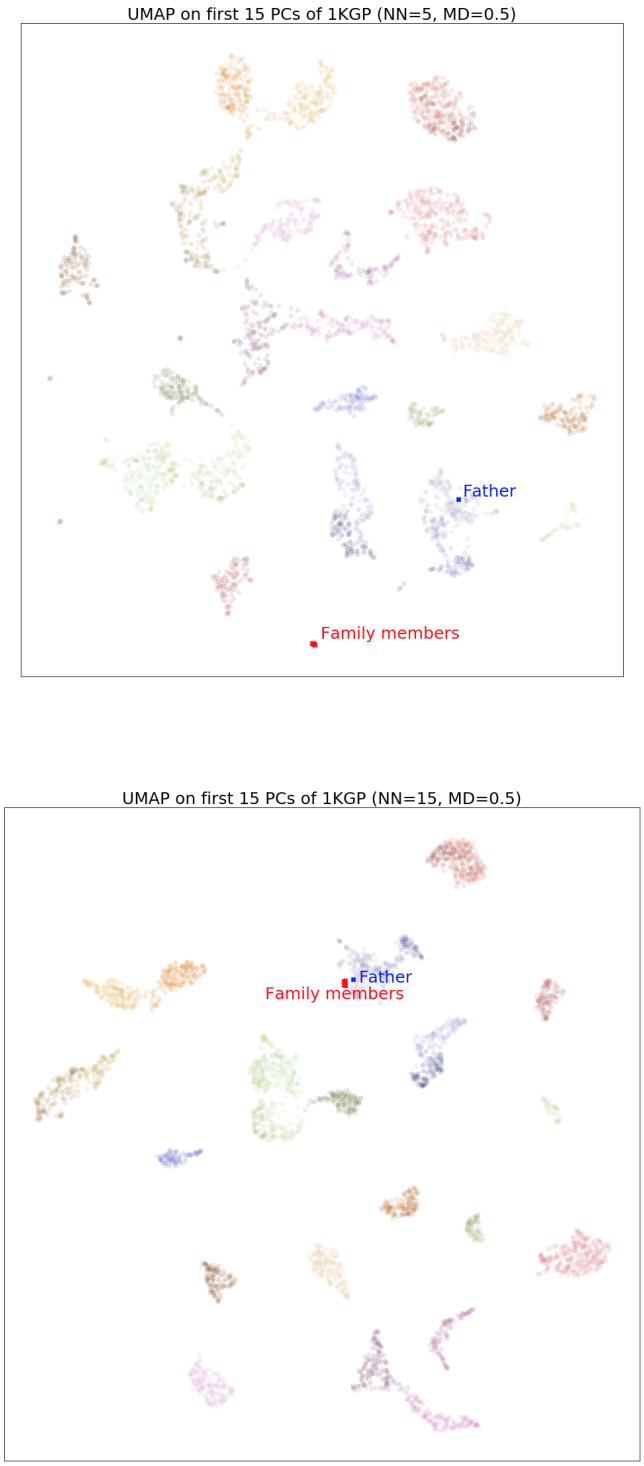



**Figure 2s5: Number of neighbours and families forming disjoint clusters.** UMAP applied to the first 15 principal components of the 1KGP, with the number of neighbours set to 5 (top) and 15 (bottom). Six members of one Southern Han Chinese family are highlighted: HG00656 (grandfather), HG00657 (grandmother), HG00658 (uncle, mother's brother), HG00701 (mother), HG00702 (father), HG00703 (child). When using UMAP with five neighbours, the father (in blue) is projected to the cluster of the Southern Han Chinese population while the rest of the family members (in red) form their own disjoint cluster. Using 15 neighbours, the family still clusters together, but as part of the Southern Han Chinese population rather than a separate cluster.

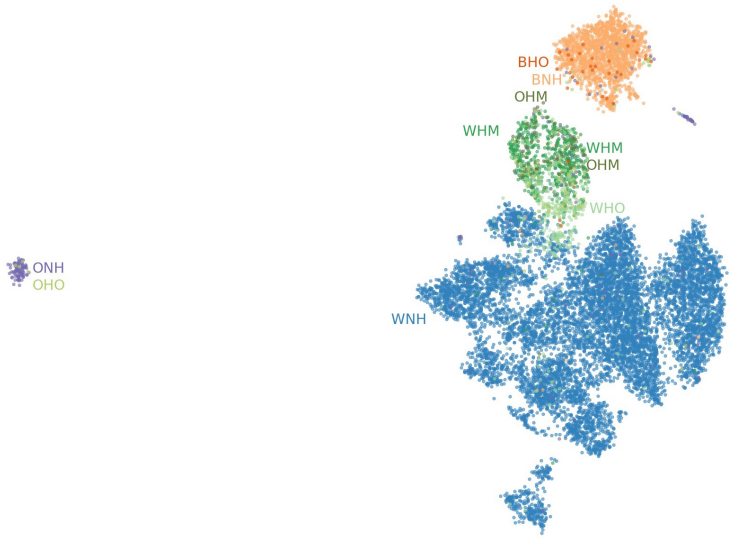



**Figure 2s6: UMAP on HRS data coloured by ethnicity.** UMAP applied to the first 10 principal components of HRS data. Points coloured by self-identified race, Hispanic status, and Mexican-American status. The cluster on the left is mostly people who identify as neither Black nor White and were born outside the contiguous United States or in the Pacific census region. Clustering with the 1KGP data places them with Asian-identified populations. BNH, Black (not Hispanic); BHO, Black (Hispanic, Other); WNH, White (not Hispanic); WHM, White (Hispanic, Mexican-American); WHO, White Hispanic (Other); ONH, Other (not Hispanic); OHM, Other (Hispanic, Mexican-American); OHO, Other (Hispanic, Other).

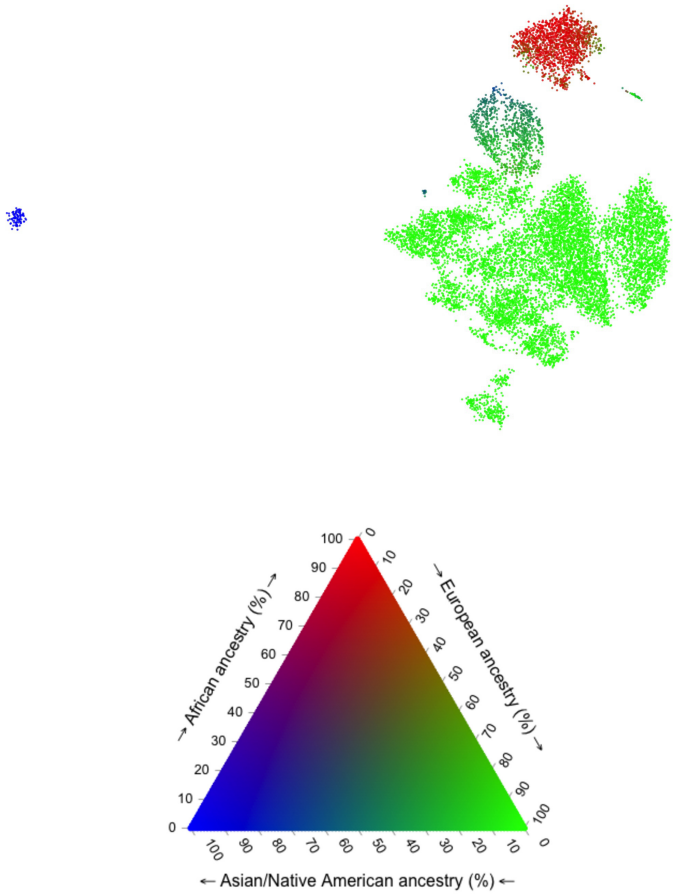



**Figure 2s7: UMAP on HRS data coloured by admixture.** UMAP on the first 10 principal components of HRS data. colouring individuals by estimated admixture from three ancestral populations reveals considerable diversity in the Hispanic population. This projection coloured by self-identified race and Hispanic status is presented in 2s6. Admixture proportions for each individual were estimated in (Baharian 2016) by assuming ancestral African, Asian, and European populations using RFMIX. We have scaled each of the three proportions to values between 0 and 255 (with 100% corresponding to 255), to colour individual points by their estimated admixture represented by RGB where red, green, and blue respectively correspond to African, European, and Asian/Native American ancestry. An alternate colouring is provided in 2s63.

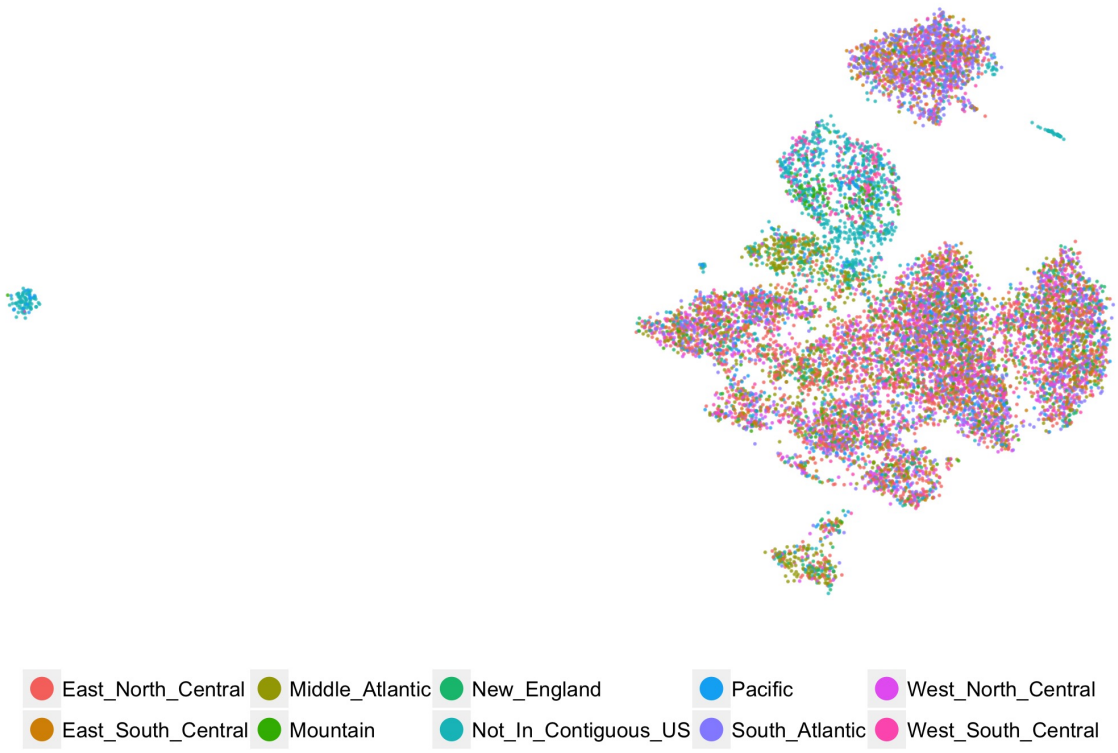



**Figure 2s8: UMAP on HRS data coloured by birth region.** UMAP on the top 10 principal components of the HRS dataset, coloured by Census Bureau birth region. Each colour represents one of the 10 birth regions. There is no obvious pattern in the clusters of majority "White Not Hispanic" individuals.

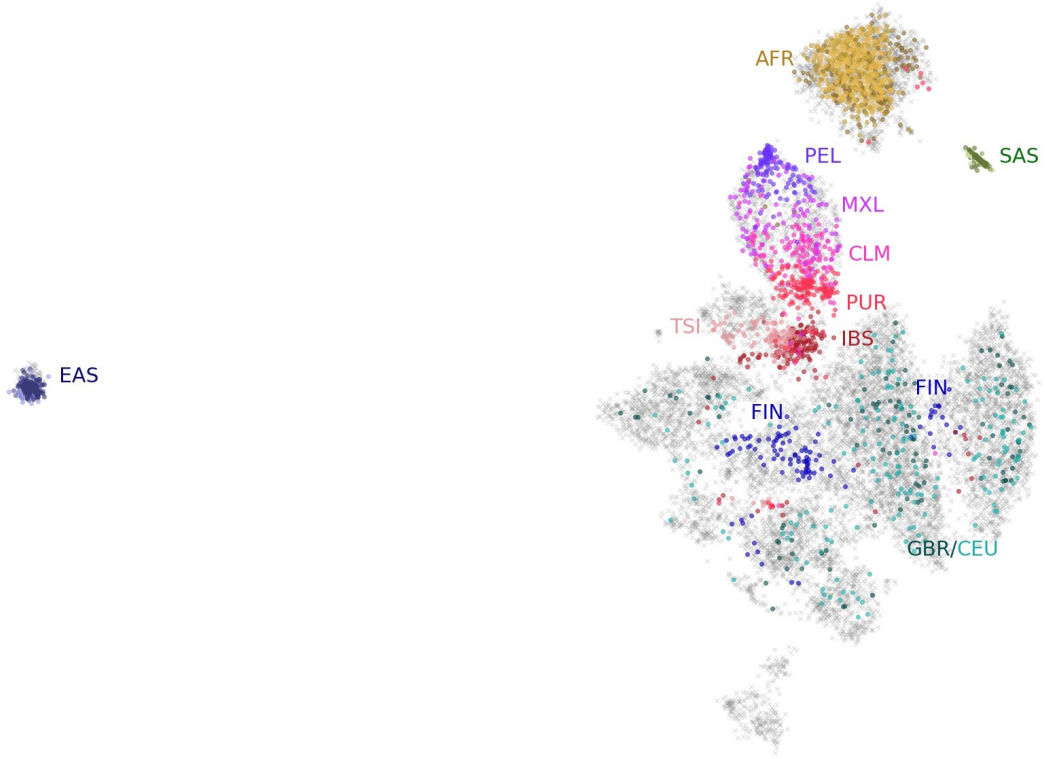



**Figure 2s9: UMAP on HRS data with 1KGP data overlaid.** UMAP on the top 10 principal components of the HRS data, with 1KGP data projected onto the embedding. Individuals from the HRS are grey. British (GBR) and other European (CEU) individuals are scattered throughout the "White Not Hispanic" clusters. Finns (FIN) form clear groupings. Spanish (IBS) and Italian (TSI) individuals cluster near the Hispanic grouping. There are sub-groups in the Hispanic cluster formed of Puerto Ricans (PUR), Colombians (CLM), Mexicans (MXL), and Peruvians (PEL). Populations with African ancestry (AFR) appear with Black individuals. East Asian (EAS) populations comprising Chinese, Kinh, and Japanese individuals cluster together with what appears in 2s7 as a population of mostly Asian ancestry. South Asian (SAS) populations with Indian, Pakistani, and Sri Lankan ancestry cluster in a separate area. One "White Not Hispanic" cluster at the bottom does not cluster with any 1KGP populations.

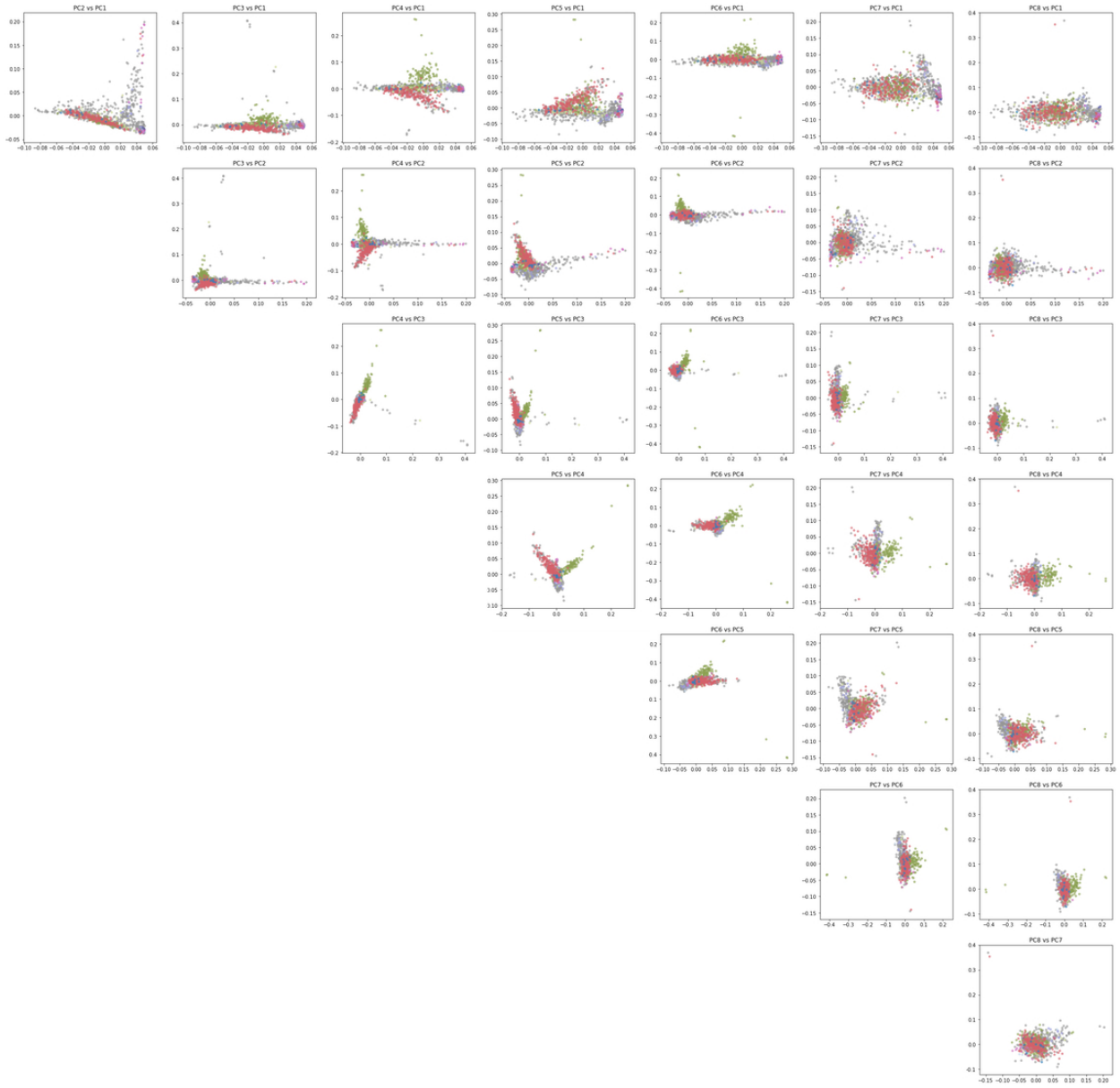



**Figure 2s10: Pairwise plots of PCs of Hispanic HRS data.** Pairwise plots of the first 8 principal components of the Hispanic subset of the HRS. Those born in the Mountain region are coloured green.

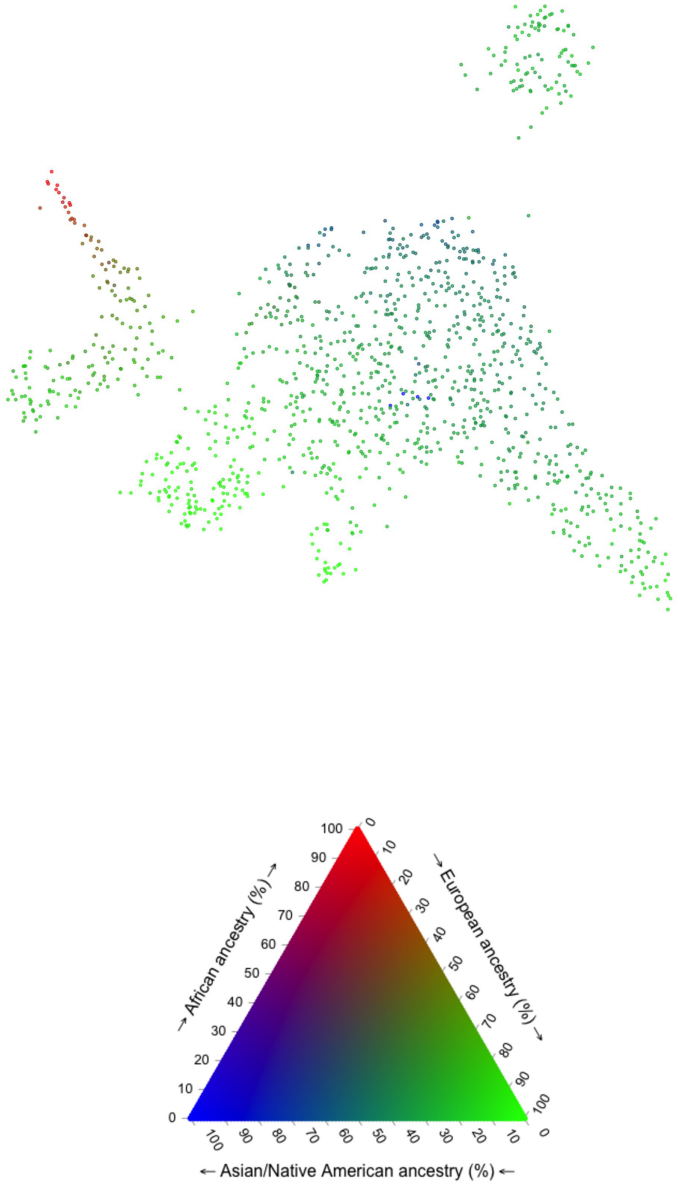



**Figure 2s11: UMAP on Hispanic HRS data coloured by admixture.** UMAP of the first 7 principal components of the Hispanic population of the HRS, coloured by estimated admixture proportions. Admixture proportions for each individual were estimated in (Baharian, 2016) by assuming ancestral African, Asian, and European populations using RFMIX. We have scaled each of the three proportions to values between 0 and 255 (with 100% corresponding to 255), to colour individual points by their estimated admixture represented by RGB where red, green, and blue respectively correspond to African, European, and Asian/Native American ancestry. An alternate colouring is provided in 2s64.


**Figure 2s12: UMAP on Hispanic HRS data coloured by birth region.** UMAP of the first 7 principal components of the Hispanic population of the HRS, coloured region of birth.

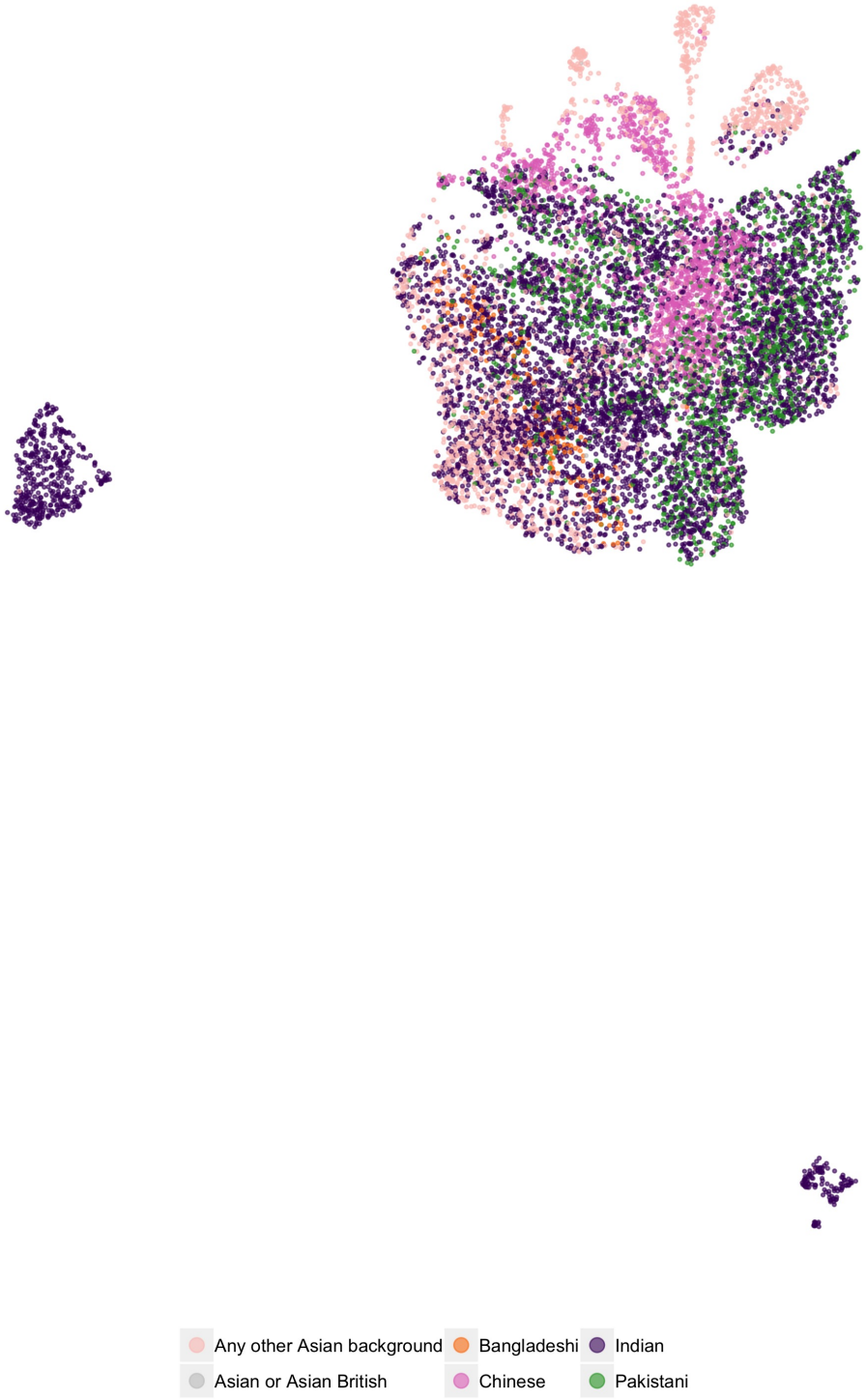



**Figure 2s13: UMAP on Asian UKBB data coloured by self-identified ethnicity.** UMAP of the first 8 principal components of the Asian population in the UKBB coloured by self-identified ethnicity. This is an alternate colouring of 2.2B.

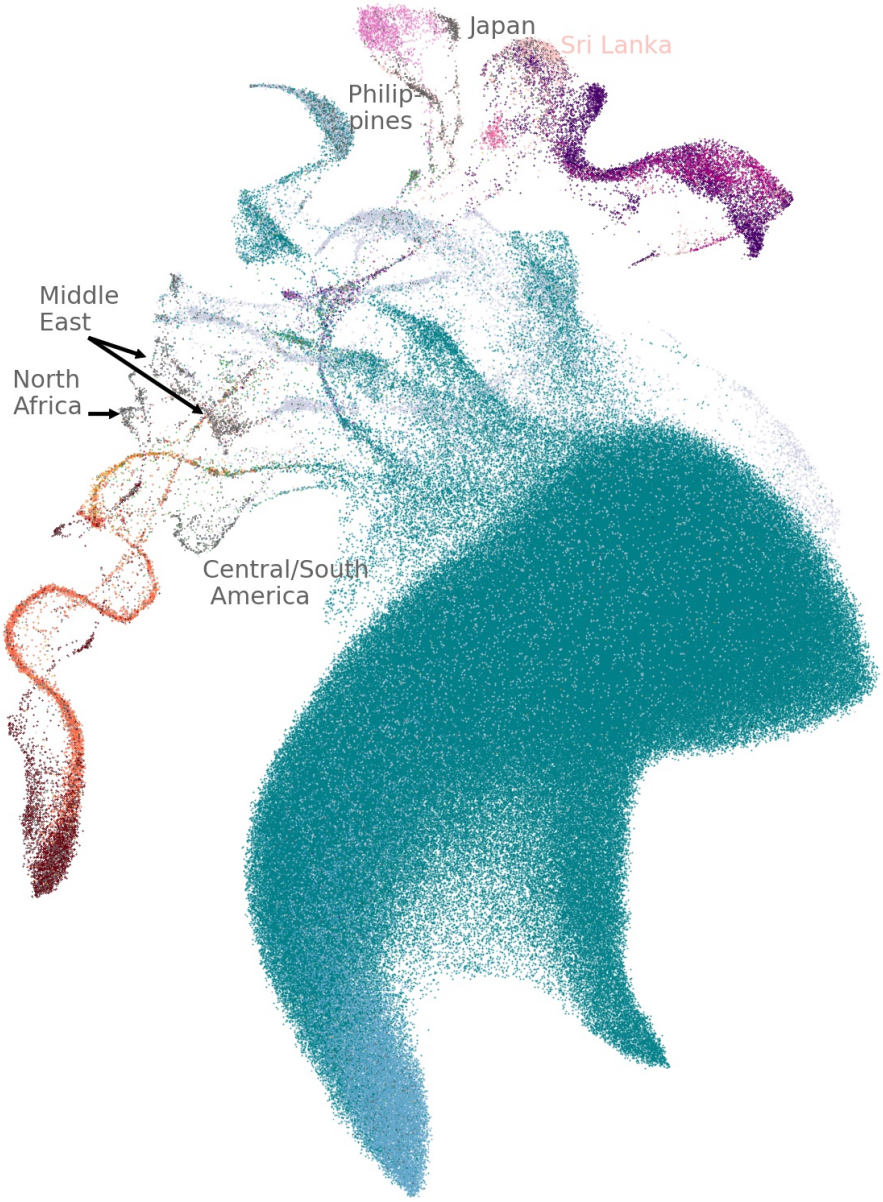



**Figure 2s14: UMAP on UKBB data with some countries of birth identified.** Using country of birth data, some of the larger unidentified groups from 2.3B were identified as being born mostly in Japan, the Philippines, North Africa, the Middle East, and Central and South America. The large cluster of "Any other Asian Background" were mostly born in Sri Lanka.

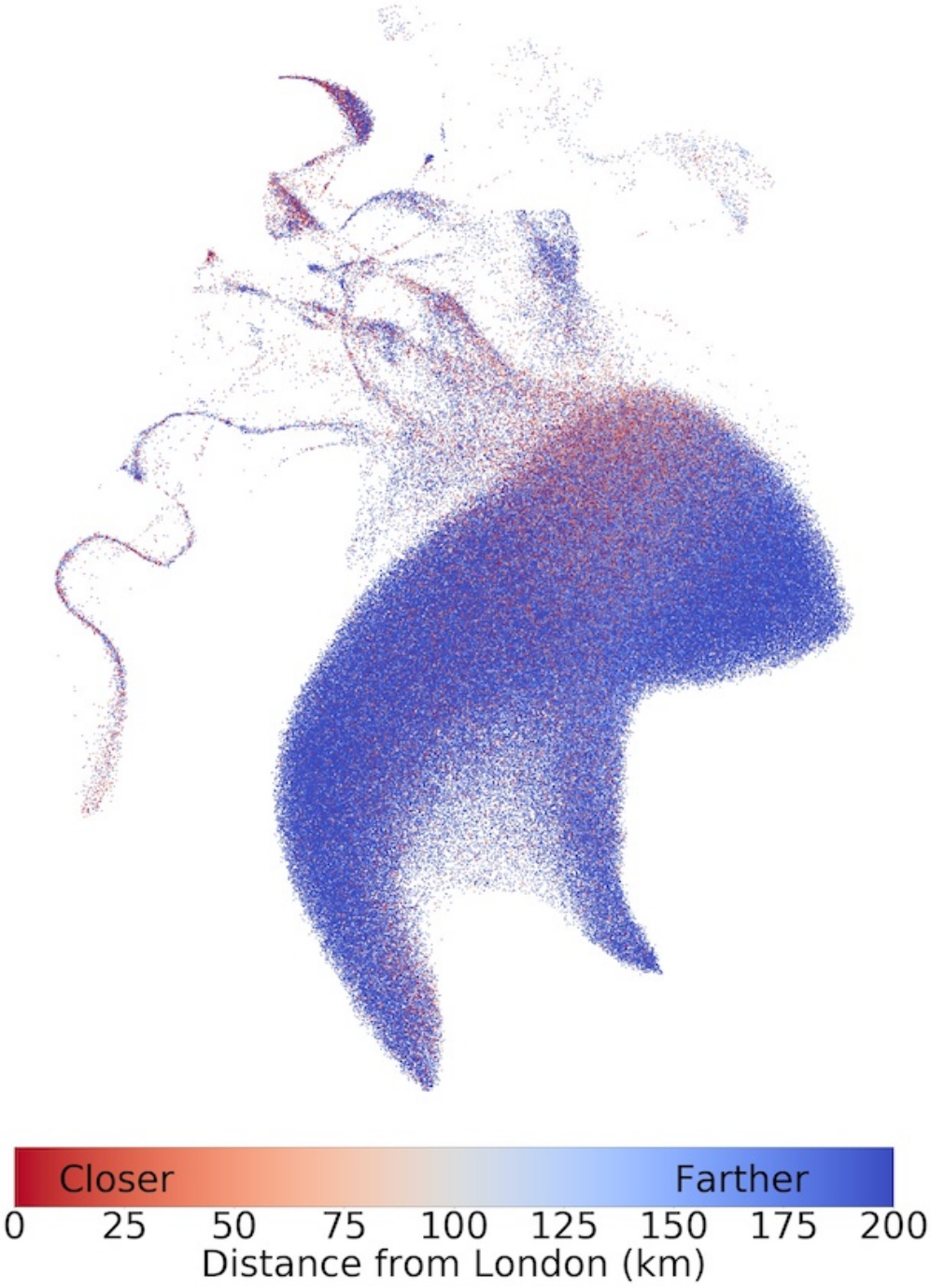



**Figure 2s15: UMAP on UKBB data coloured by distance from London.** UMAP on UKBB data, coloured by distance from London, with red representing those living closer to London and blue representing those living farther from London. A 200km radius extends roughly to Cardiff, and a 100km radius extends roughly to cities such as Leicester and Bath, and contains cities such as Oxford, Cambridge, and Peterborough. Data has been randomized as explained in the materials and methods section.

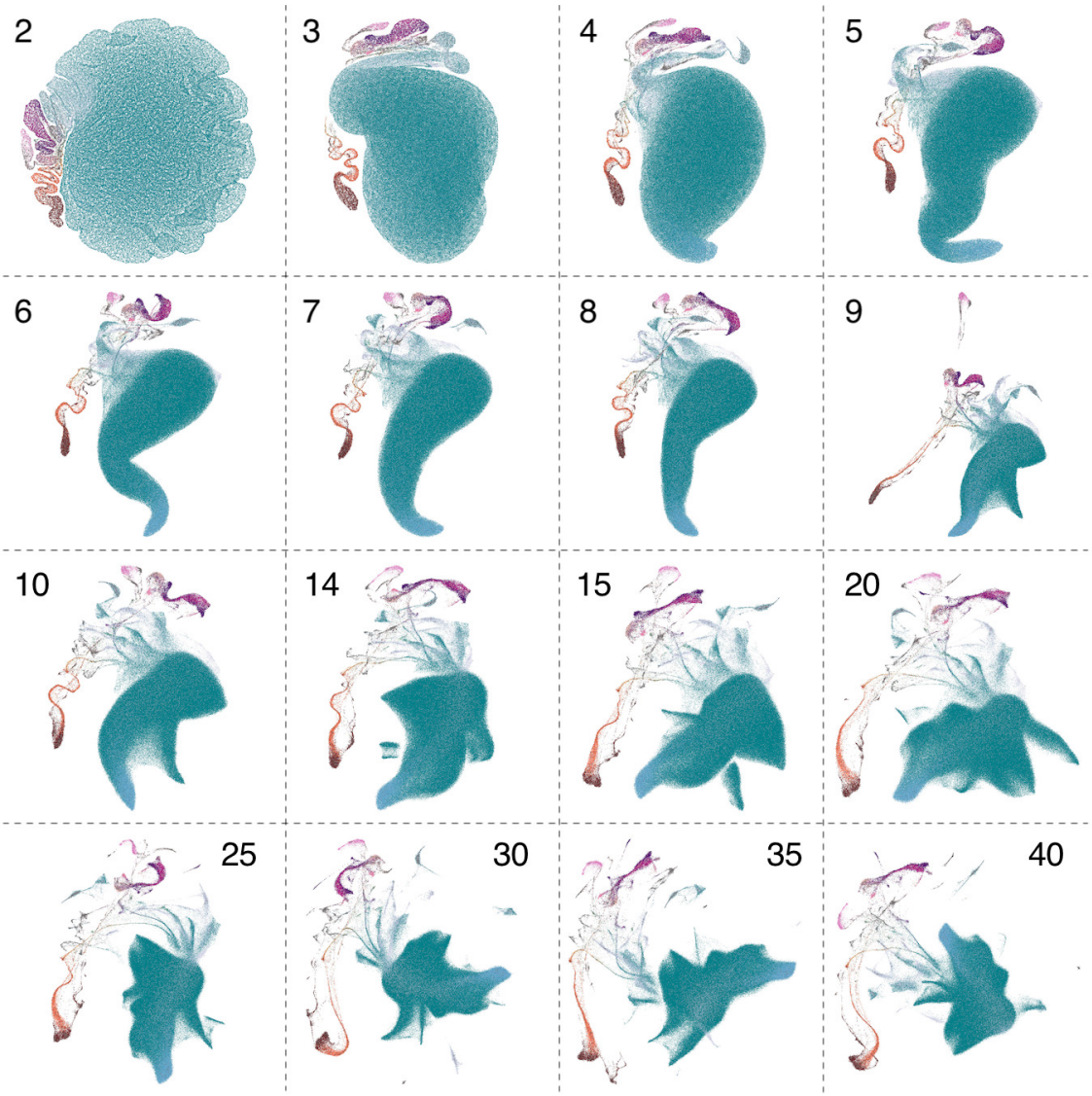



**Figure 2s16: Montage of UMAP on top 40 PCs of UKBB data coloured by ethnicity.** UMAP on UKBB data, coloured by self-identified ethnic background. Images are labelled by the number of components included.

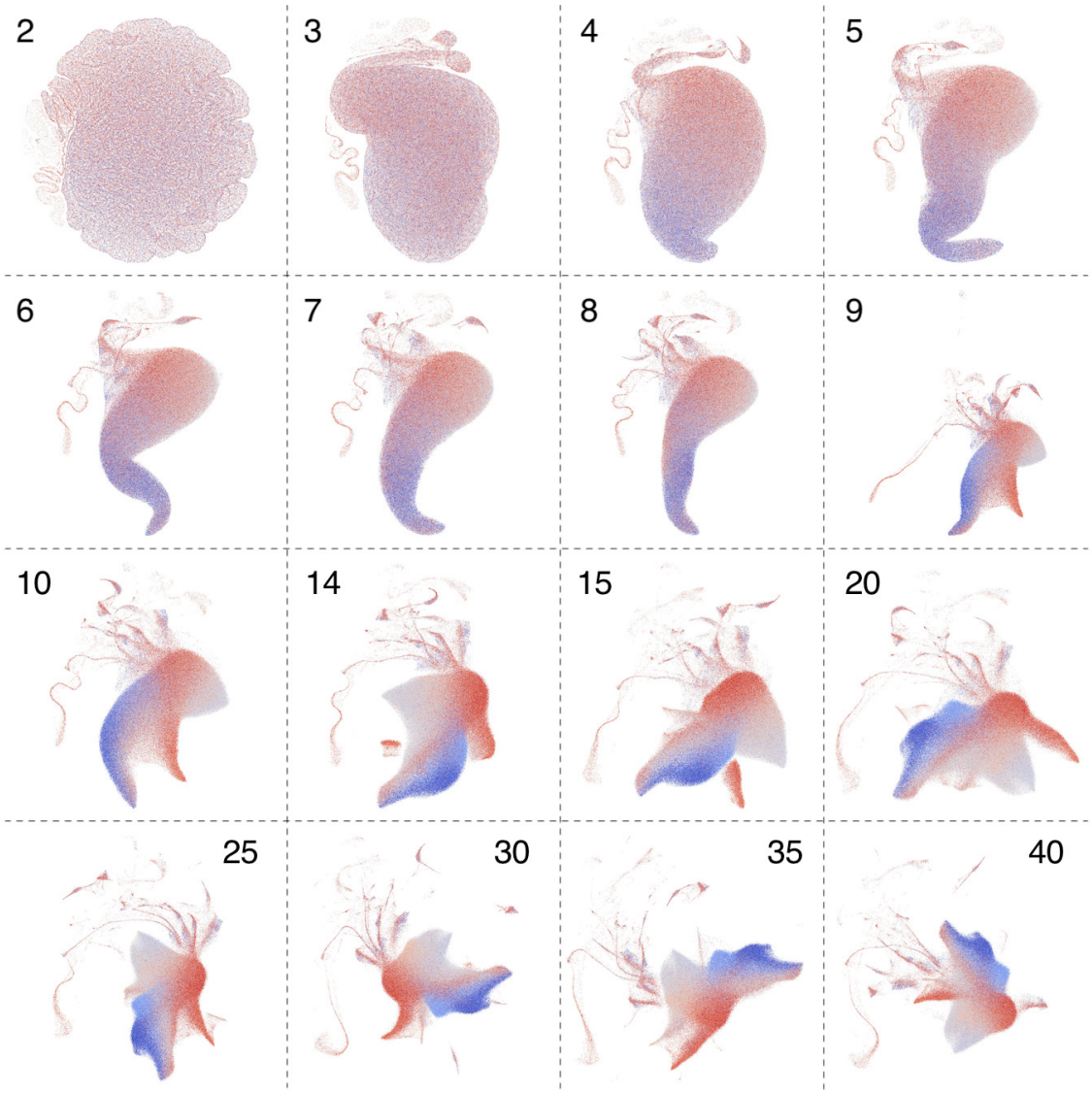



**Figure 2s17: Montage of UMAP on top 40 PCs of UKBB data coloured by northing.** UMAP on UKBB data, coloured by northing values, with more blue representing more northern coordinates and more red representing more southern coordinates. Images are labelled by the number of components included. Data has been randomized as explained in the materials and methods section.

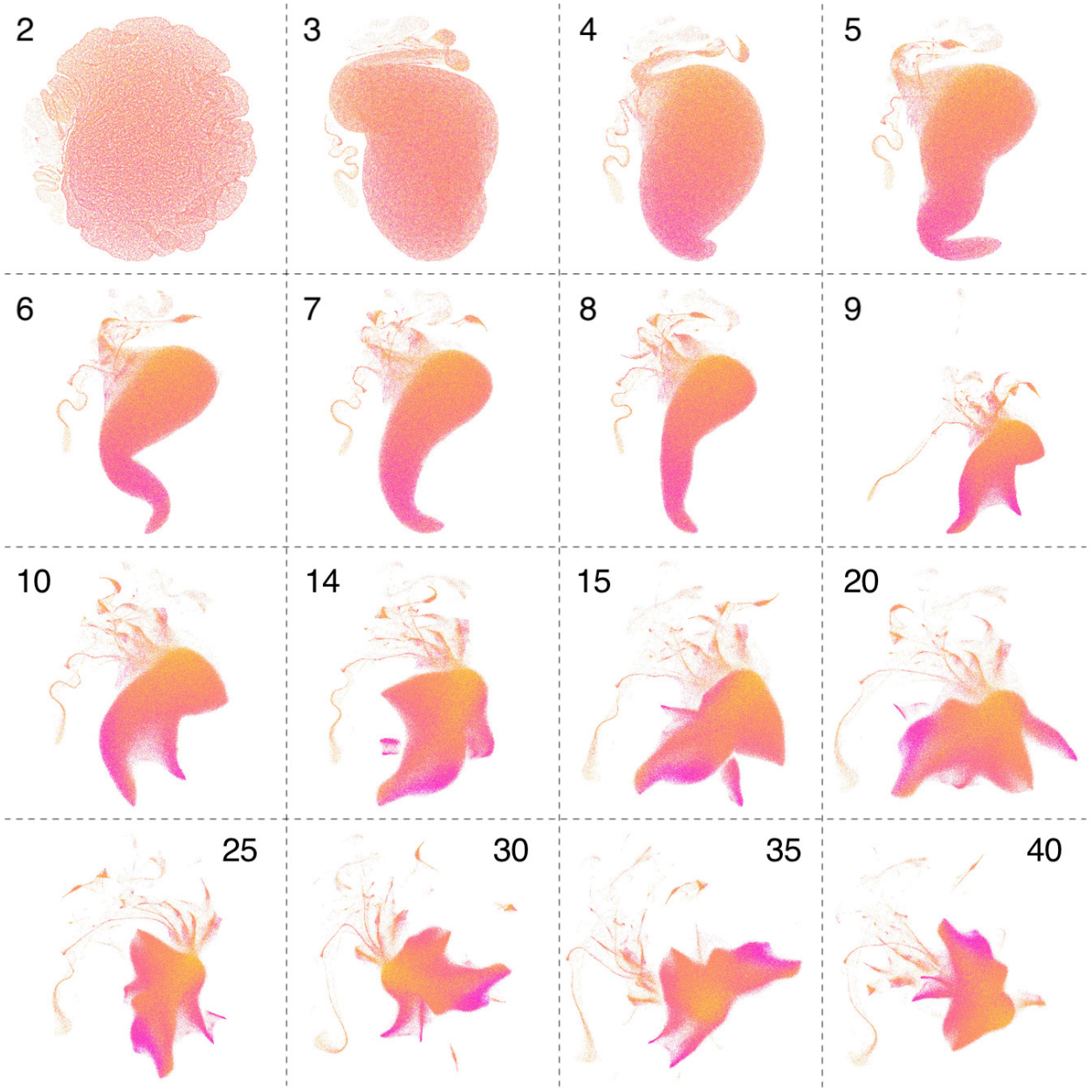



**Figure 2s18: Montage of UMAP on top 40 PCs of UKBB data coloured by easting.** UMAP on UKBB data, coloured by easting values, with more yellow representing more eastern coordinates and more pink representing more western coordinates. Images are labelled by the number of components included. Data has been randomized as explained in the materials and methods section.

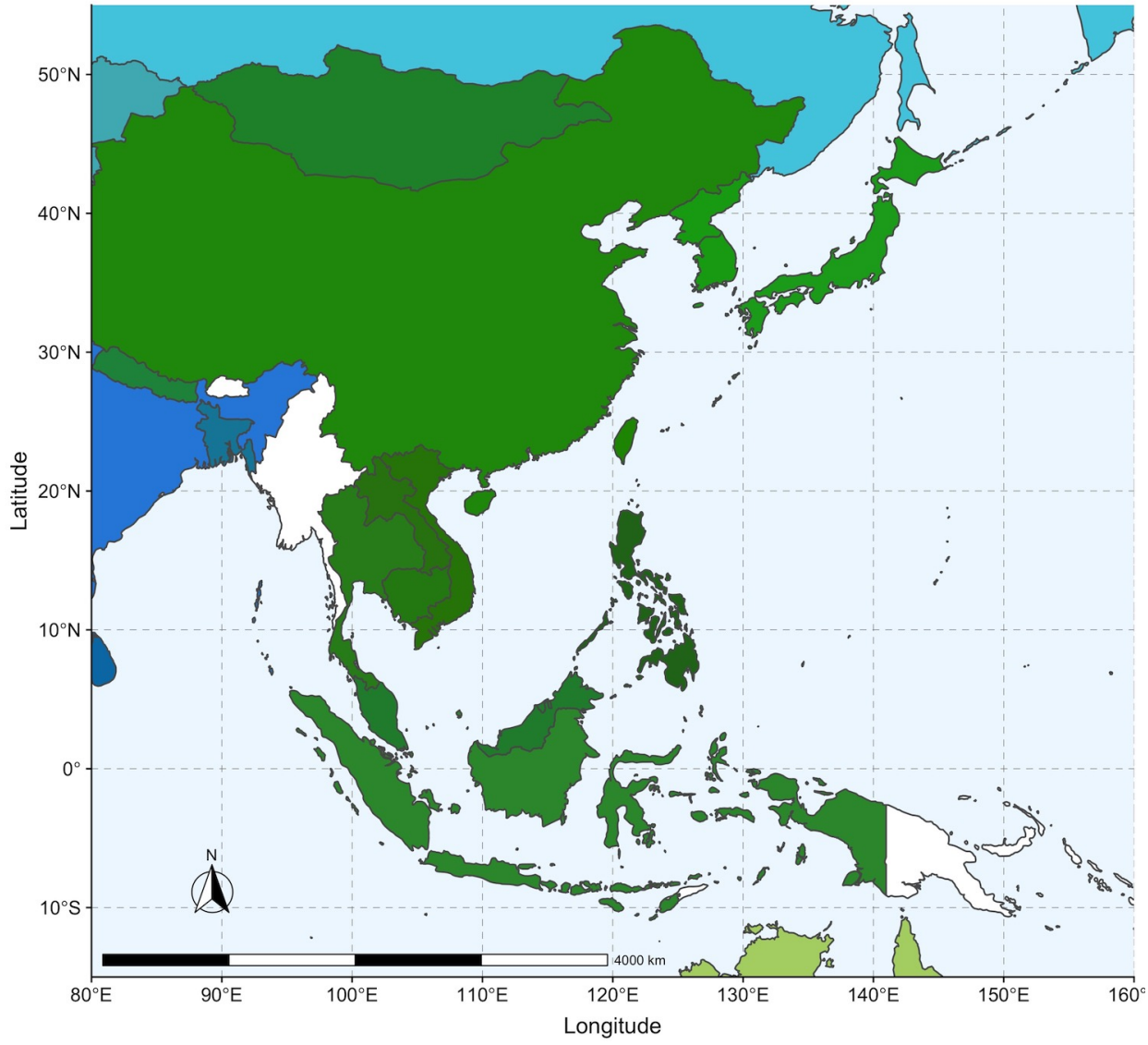



**Figure 2s19: Map of Asia coloured by 3D UMAP coordinates of UKBB data.** Fig 5b, zoomed in on Asia. Geographic distribution of UMAP coordinates. Using the country of birth of individuals in the UKBB, we colour countries by the closeness in 3D UMAP space of those born there. Broad patterns of similarity appear in East Asia, South Asia, North African and the Middle East, West Africa, and South America. Differences between neighbouring countries can reflect both ancient population structure and recent differences in migration history. Evidence of migrations related to colonialism are visible with, e.g., European ancestry in South Africa and South Asian ancestry in Kenya and Tanzania. Because of the large number of White British individuals born abroad, to avoid skewing the colour scale they were not included unless they were born in the UK, Europe, Australia, Canada, or the United States, where UKBB participants already tended to have European ancestry.

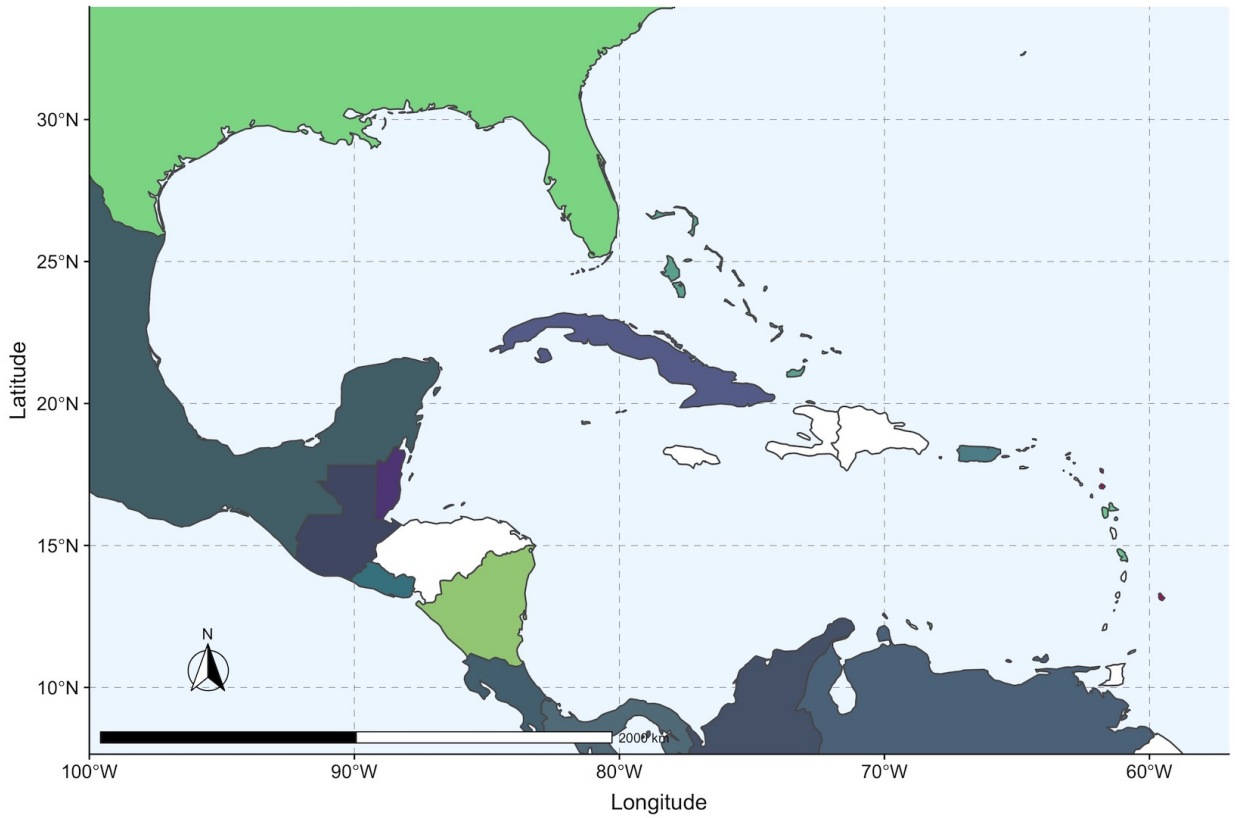



**Figure 2s20: Map of Caribbean coloured by 3D UMAP coordinates of UKBB data.** Fig 5b, zoomed in on the Caribbean. Geographic distribution of UMAP coordinates. Using the country of birth of individuals in the UKBB, we colour countries by the closeness in 3D UMAP space of those born there. Broad patterns of similarity appear in East Asia, South Asia, North African and the Middle East, West Africa, and South America. Differences between neighbouring countries can reflect both ancient population structure and recent differences in migration history. Evidence of migrations related to colonialism are visible with, e.g., European ancestry in South Africa and South Asian ancestry in Kenya and Tanzania. Because of the large number of White British individuals born abroad, to avoid skewing the colour scale they were not included unless they were born in the UK, Europe, Australia, Canada, or the United States, where UKBB participants already tended to have European ancestry.

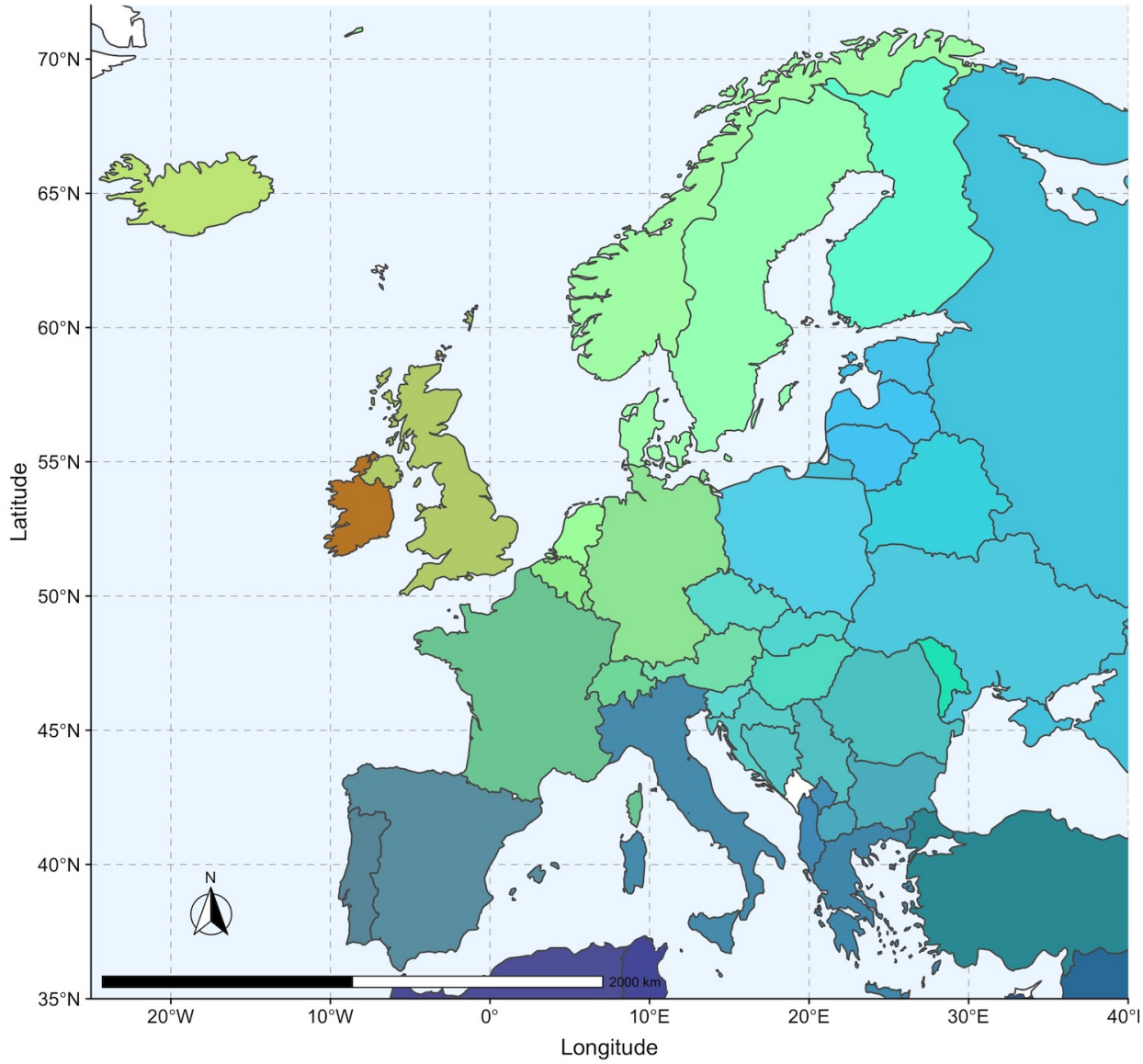



**Figure 2s21: Map of Europe coloured by 3D UMAP coordinates of UKBB data.** Fig 5b, zoomed in on Europe. Geographic distribution of UMAP coordinates. Using the country of birth of individuals in the UKBB, we colour countries by the closeness in 3D UMAP space of those born there. Broad patterns of similarity appear in East Asia, South Asia, North African and the Middle East, West Africa, and South America. Differences between neighbouring countries can reflect both ancient population structure and recent differences in migration history. Evidence of migrations related to colonialism are visible with, e.g., European ancestry in South Africa and South Asian ancestry in Kenya and Tanzania. Because of the large number of White British individuals born abroad, to avoid skewing the colour scale they were not included unless they were born in the UK, Europe, Australia, Canada, or the United States, where UKBB participants already tended to have European ancestry.

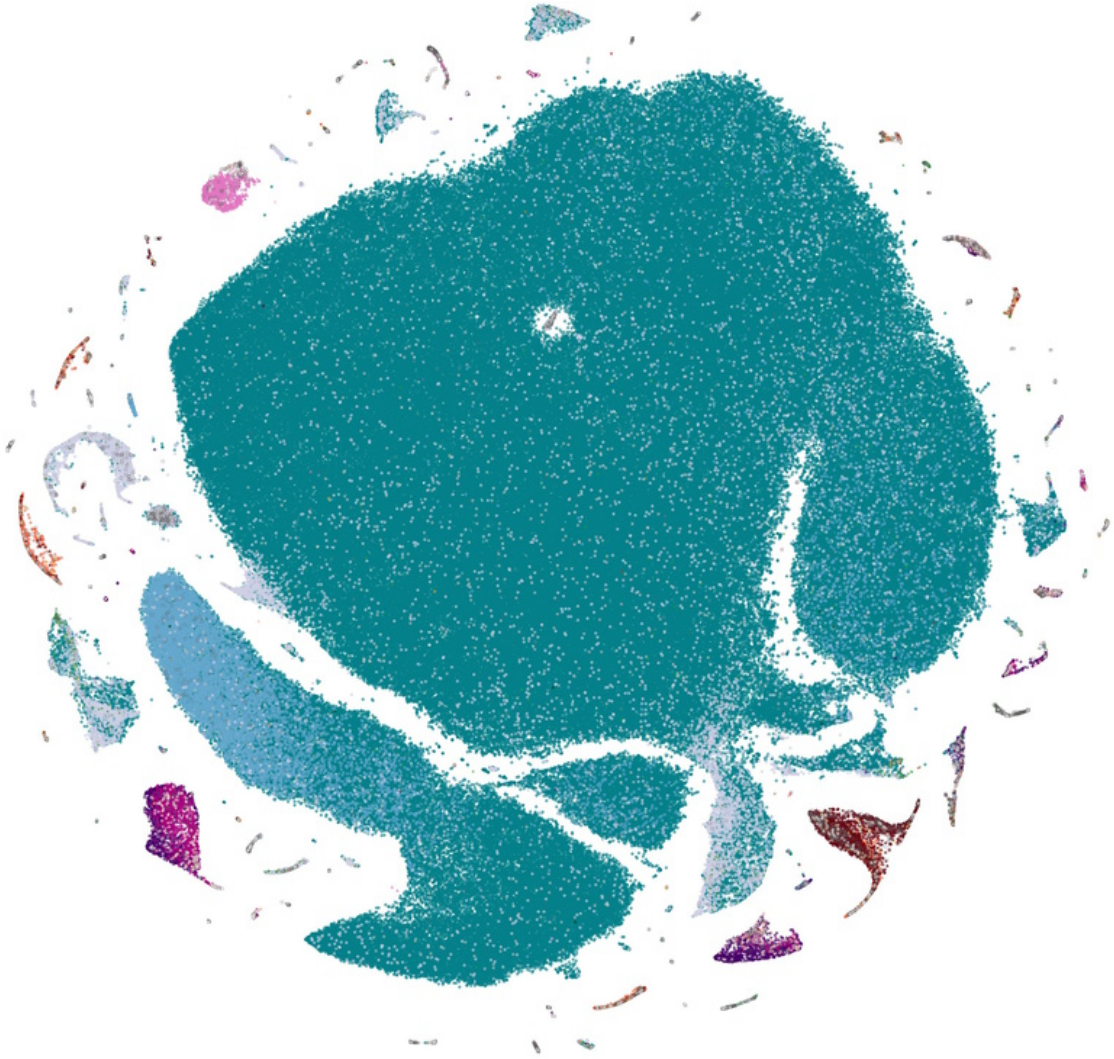


**Figure 2s22: $t$-sne on UKBB data coloured by self-identified ethinicity.** $t$-sne applied to the top 10 principal components of the UKBB, coloured by ethnic background. The unbalanced populations resulted in many individuals and populations being orphaned along the periphery of the main cluster.

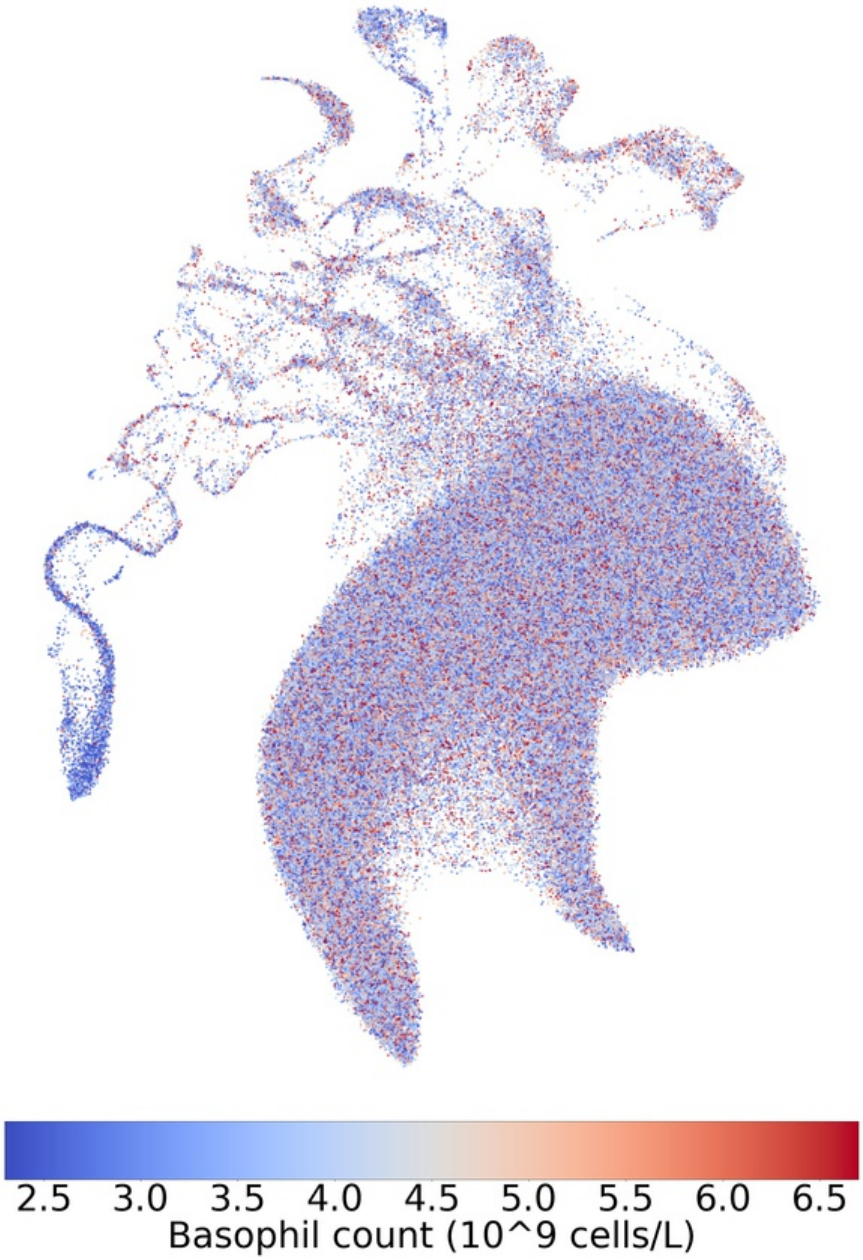



**Figure 2s23: UMAP on UKBB data coloured by basophil count (female).** UMAP on the top 10 principal components of the UKBB coloured by basophil count (female). Data has been randomized as explained in the materials and methods section.

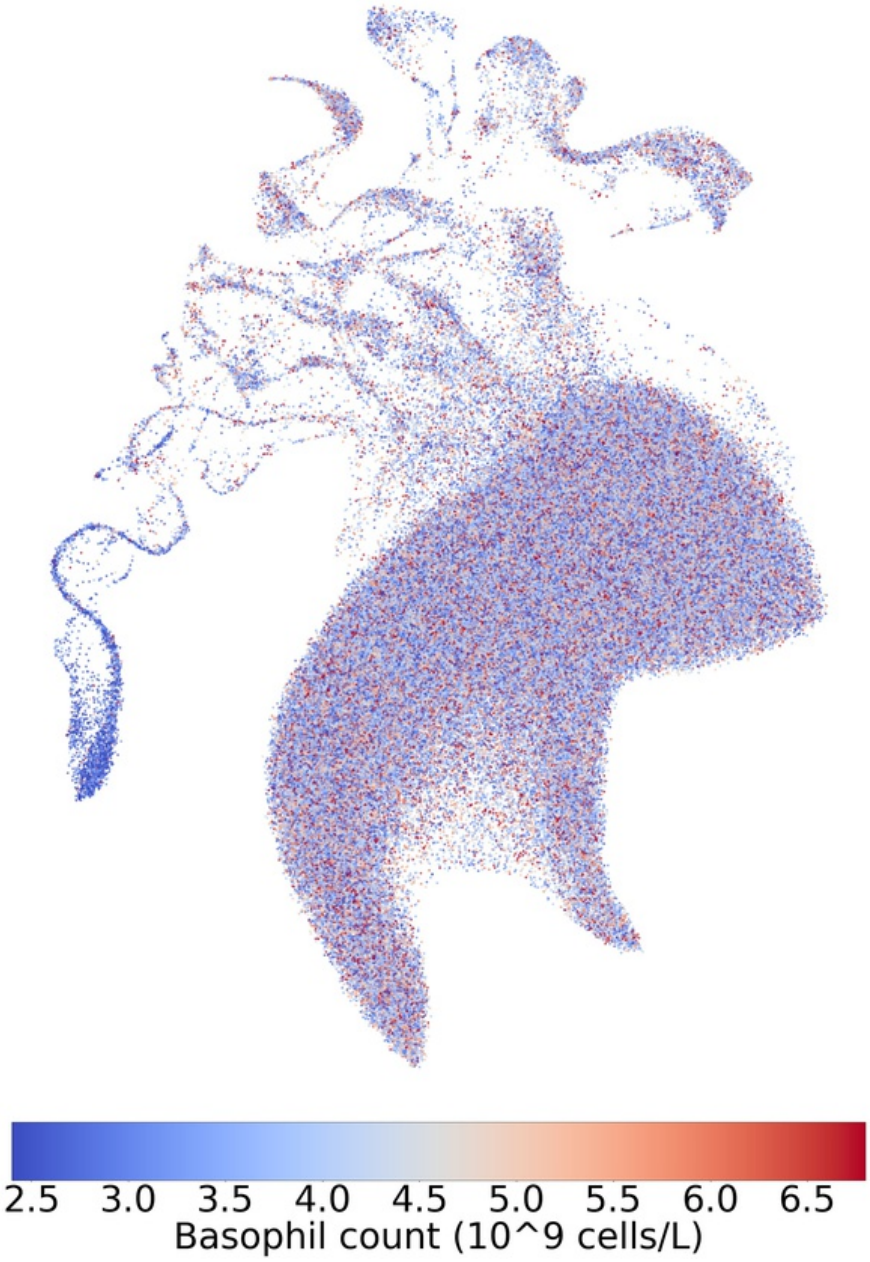



**Figure 2s24: UMAP on UKBB data coloured by basophil count (male).** UMAP on the top 10 principal components of the UKBB coloured by basophil count (male). Data has been randomized as explained in the materials and methods section.

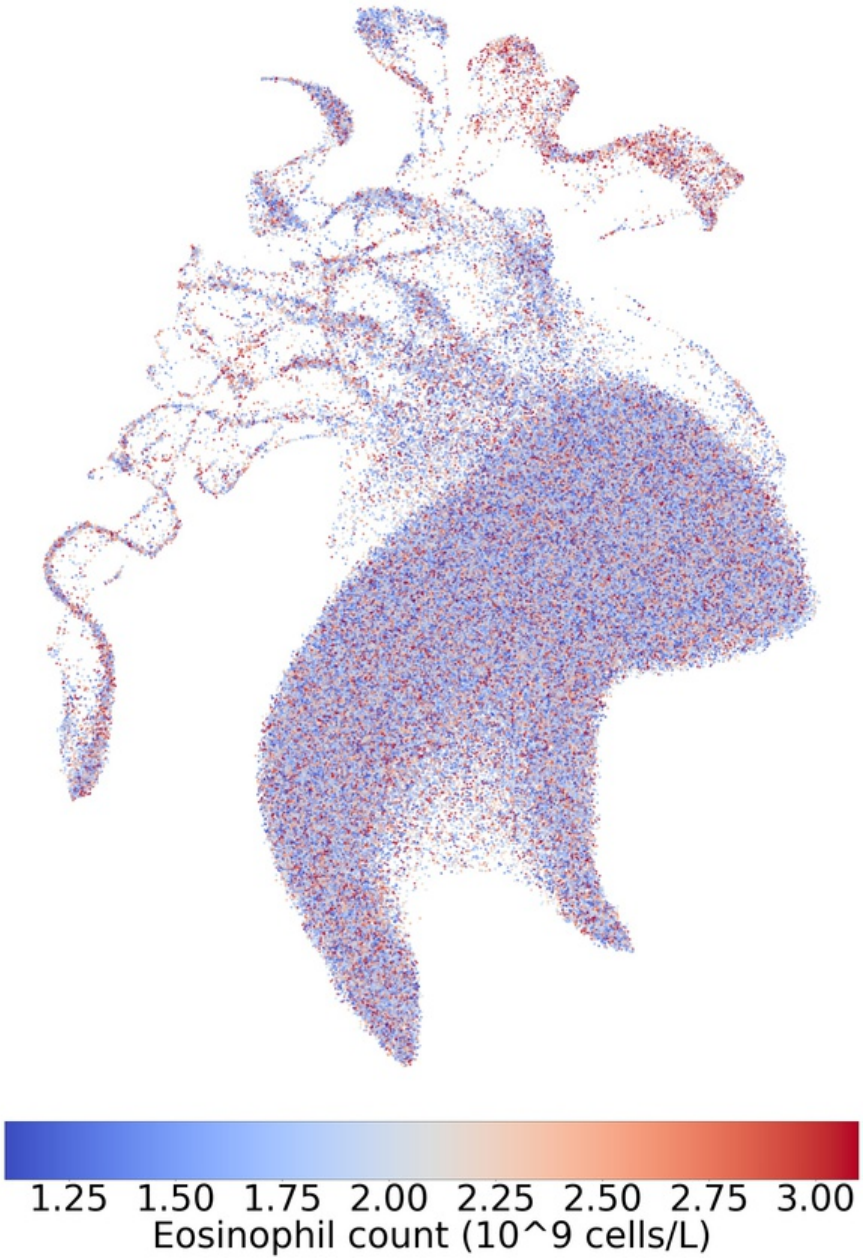



**Figure 2s25: UMAP on UKBB data coloured by eosinphil count (female).** UMAP on the top 10 principal components of the UKBB coloured by eosinophil count (female). Data has been randomized as explained in the materials and methods section.

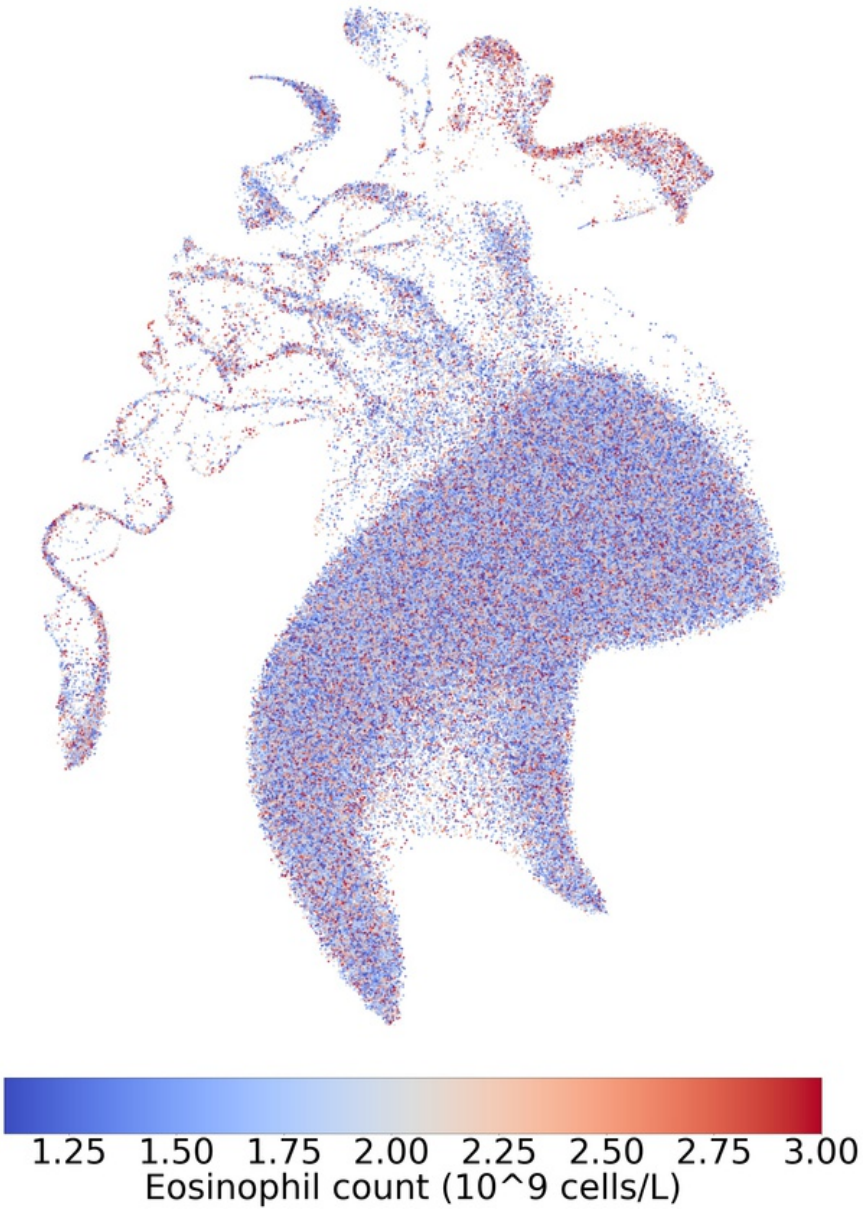



**Figure 2s26: UMAP on UKBB data coloured by eosinphil count (male).** UMAP on the top 10 principal components of the UKBB coloured by eosinophil count (male). Data has been randomized as explained in the materials and methods section.

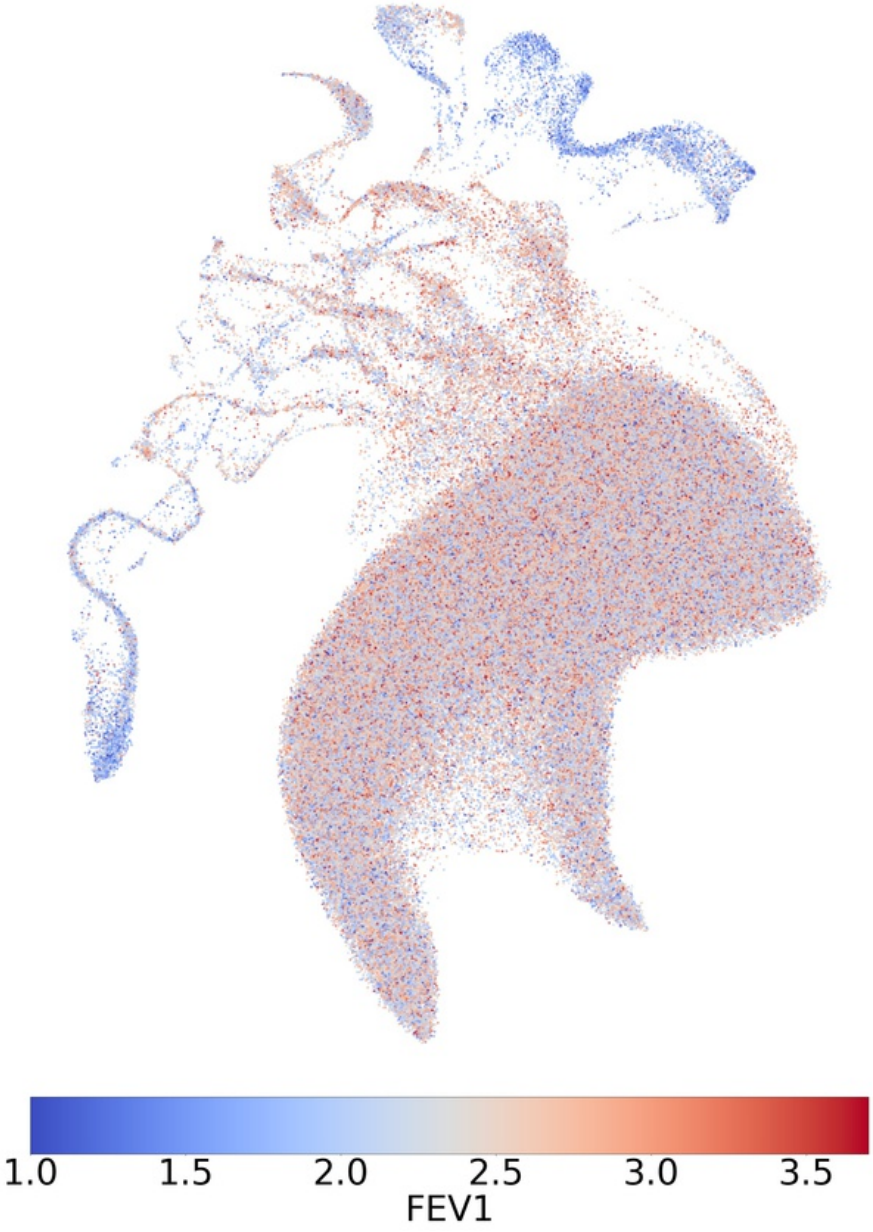



**Figure 2s27: UMAP on UKBB data coloured by FEV1 (female).** UMAP on the top 10 principal components of the UKBB coloured by FEV1 (female). Data has been randomized as explained in the materials and methods section.


**Figure 2s28: UMAP on UKBB data coloured by FEV1 (male).** UMAP on the top 10 principal components of the UKBB coloured by FEV1 (male). Data has been randomized as explained in the materials and methods section.

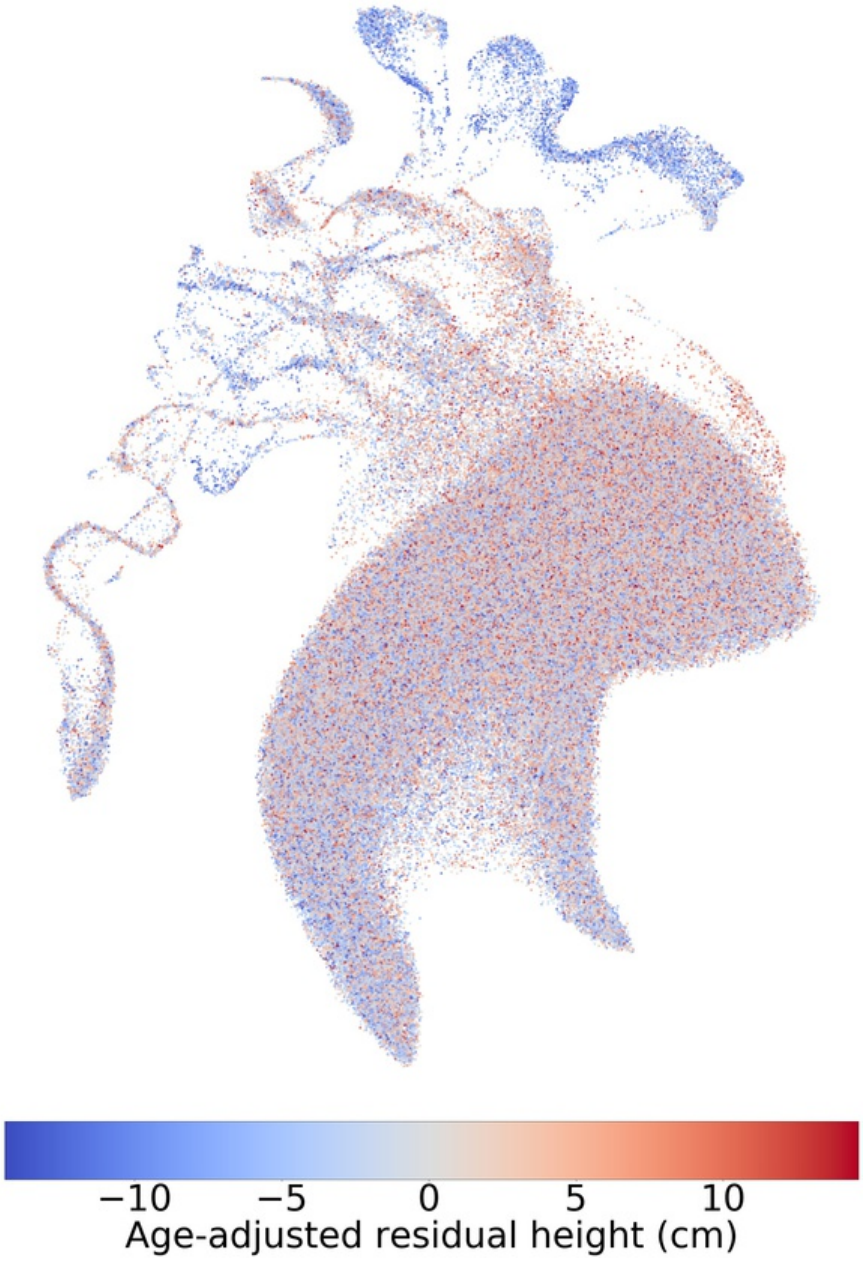



**Figure 2s29: UMAP on UKBB data coloured by height (female).** UMAP on the top 10 principal components of the UKBB coloured by height (female). Data has been randomized as explained in the materials and methods section.

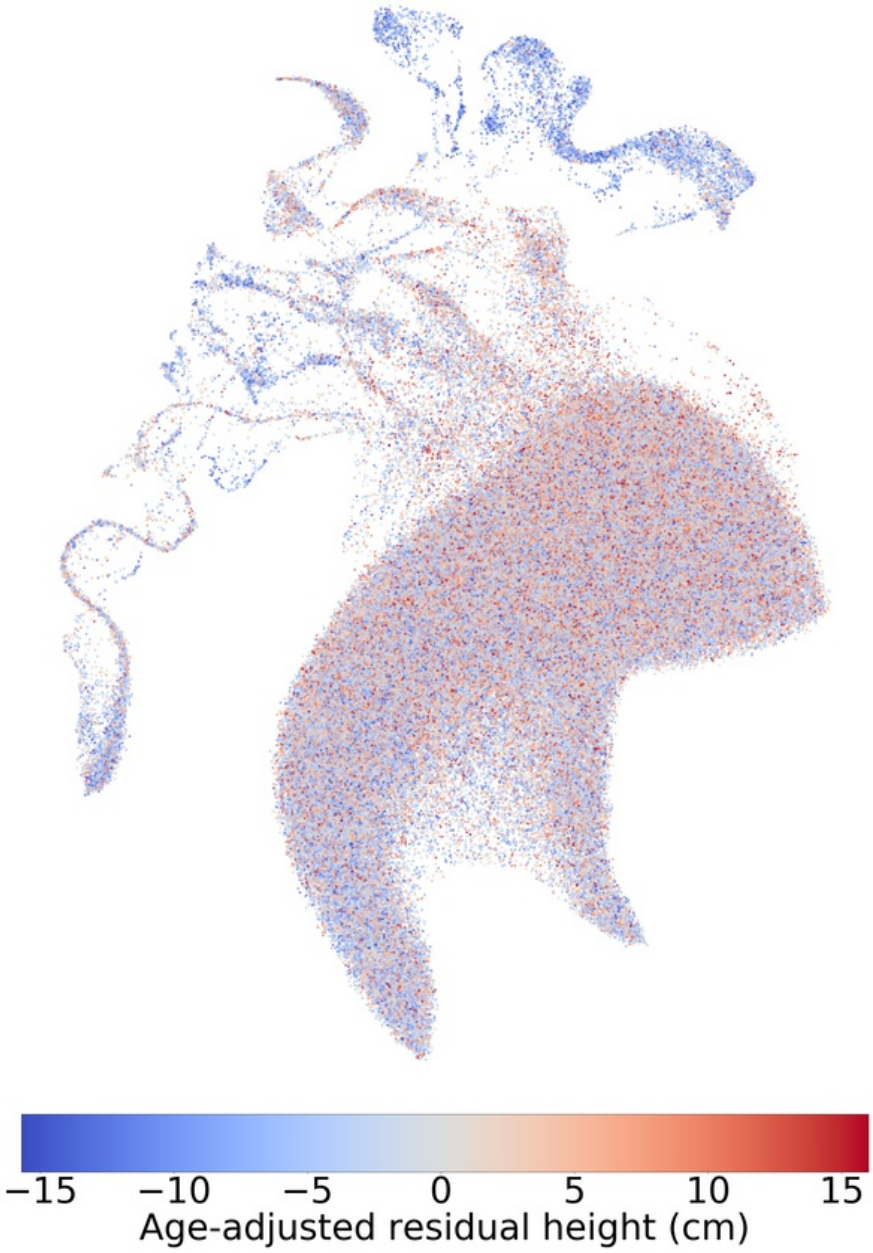



**Figure 2s30: UMAP on UKBB data coloured by height (male).** UMAP on the top 10 principal components of the UKBB coloured by height (male). Data has been randomized as explained in the materials and methods section.

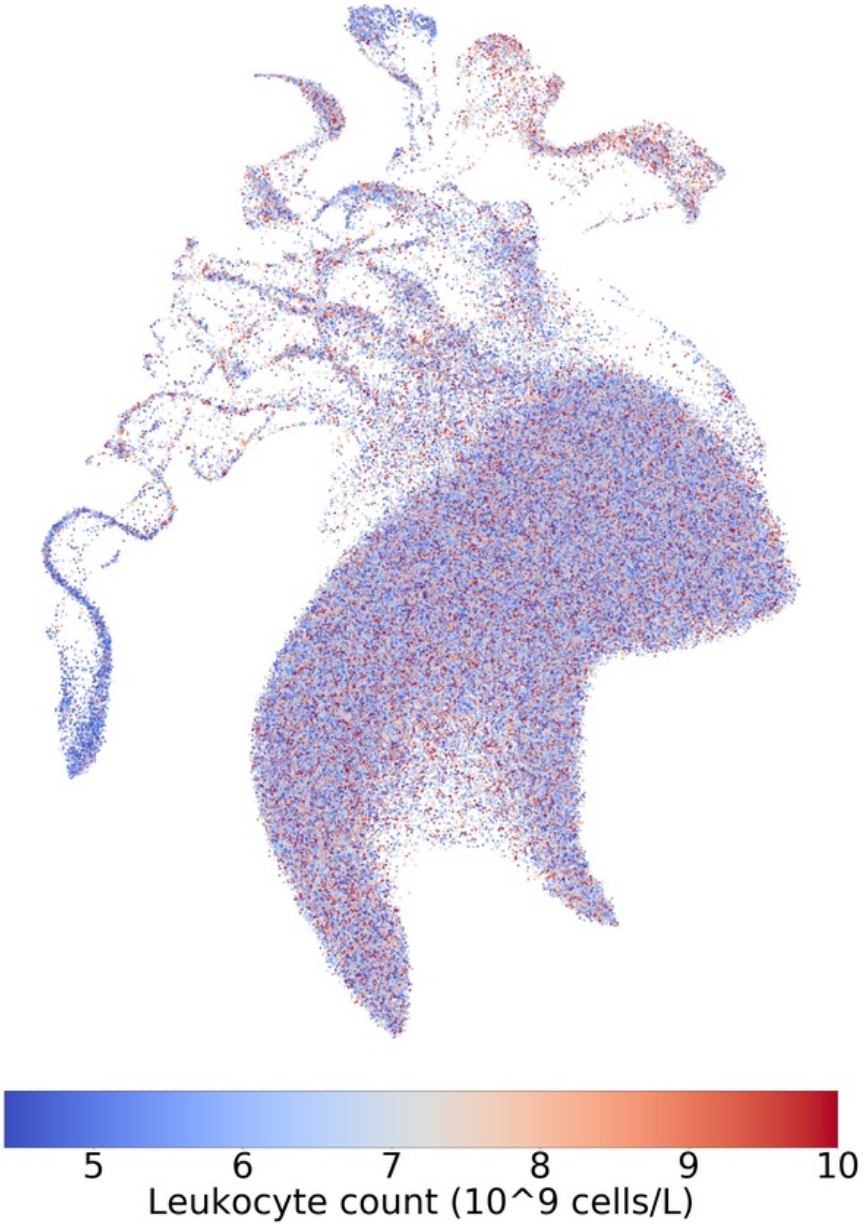



**Figure 2s31: UMAP on UKBB data coloured by leukocyte count (female).** UMAP on the top 10 principal components of the UKBB coloured by leukocyte count (female). Data has been randomized as explained in the materials and methods section.

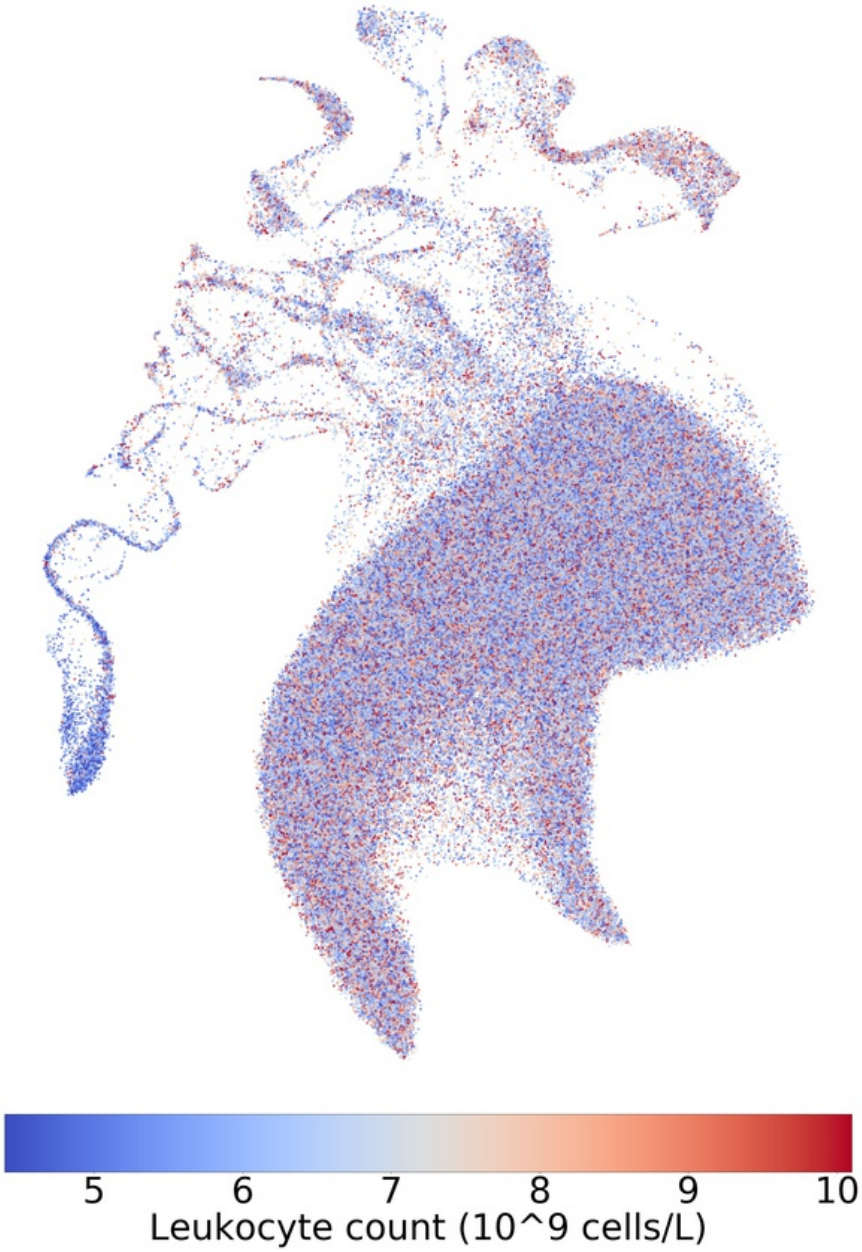



**Figure 2s32: UMAP on UKBB data coloured by leukocyte count (male).** UMAP on the top 10 principal components of the UKBB coloured by leukocyte count (male). Data has been randomized as explained in the materials and methods section.

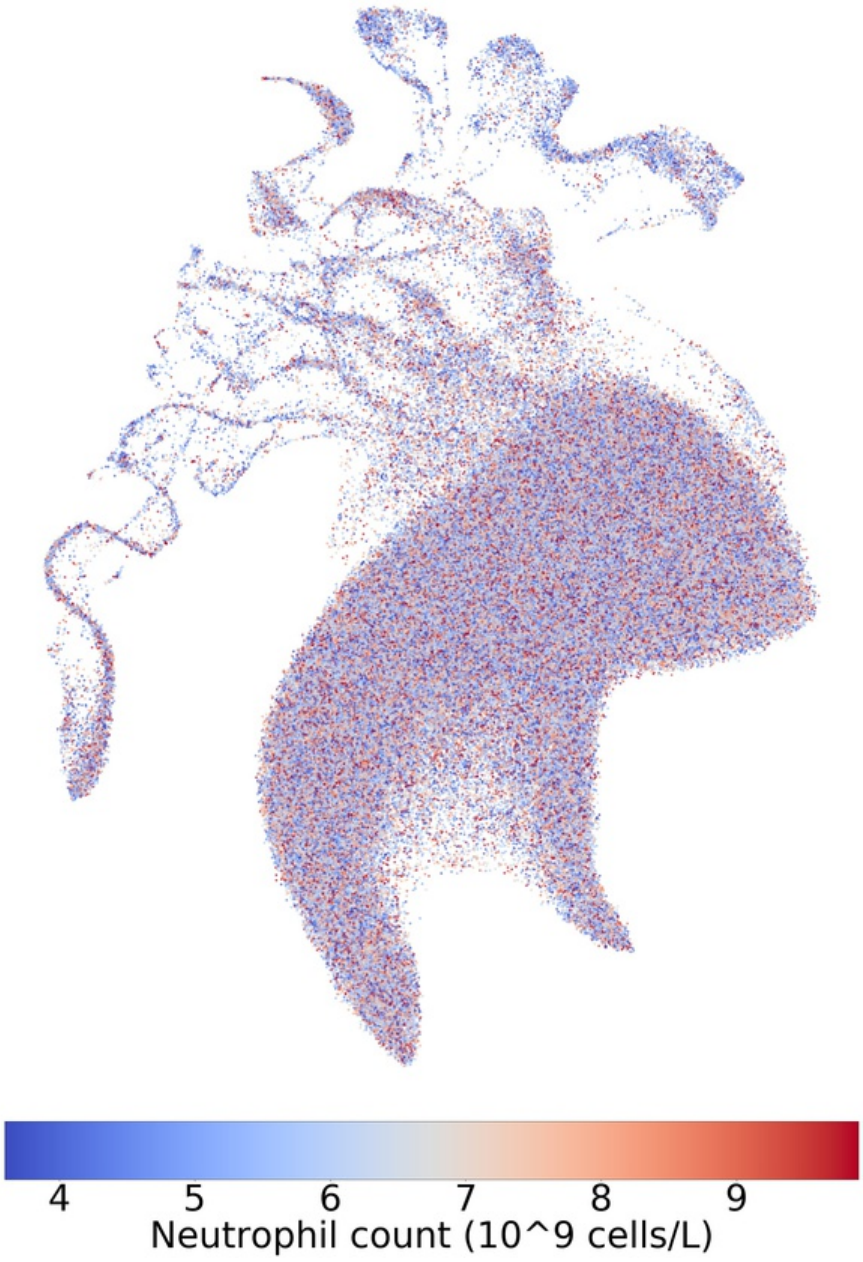



**Figure 2s33: UMAP on UKBB data coloured by neutrophil count (female).** UMAP on the top 10 principal components of the UKBB coloured by neutrophil count (female). Data has been randomized as explained in the materials and methods section.


**Figure 2s34: UMAP on UKBB data coloured by neutrophil count (male).** UMAP on the top 10 principal components of the UKBB coloured by neutrophil count (male). Data has been randomized as explained in the materials and methods section.

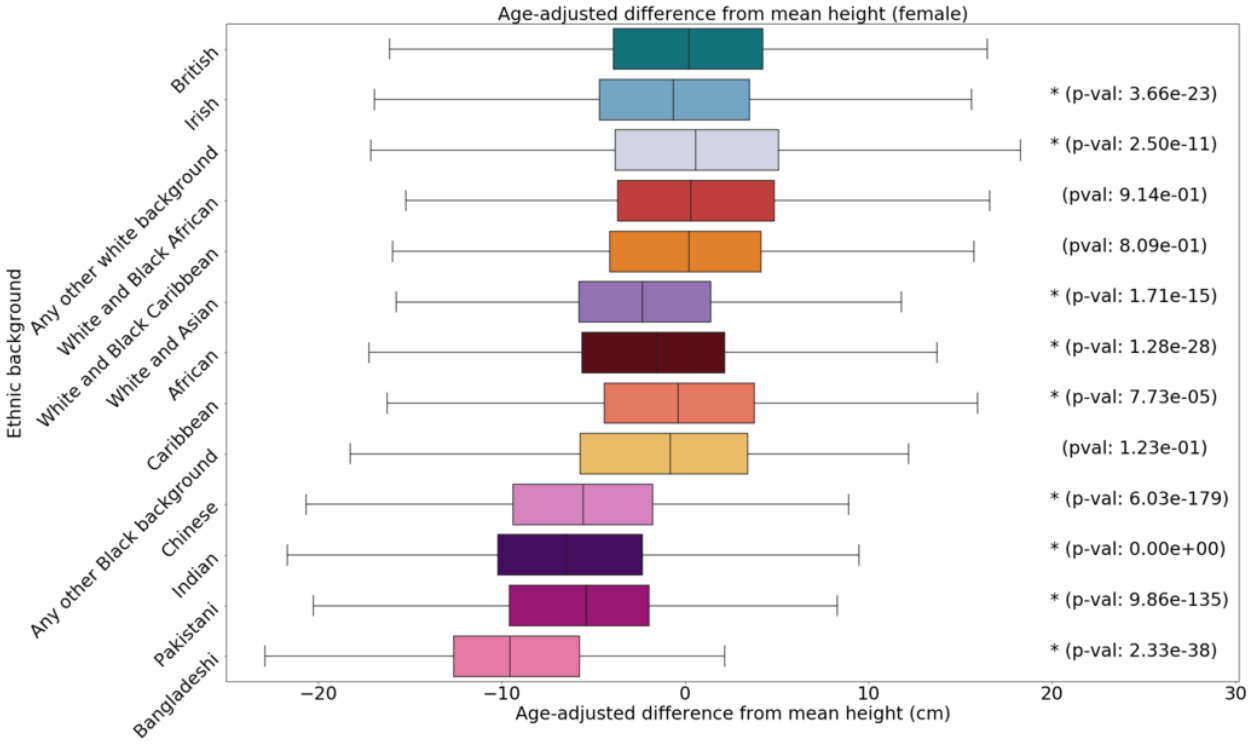



**Figure 2s35: Box plots of height in the UKBB by self-identified ethnicity (female).** Height by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

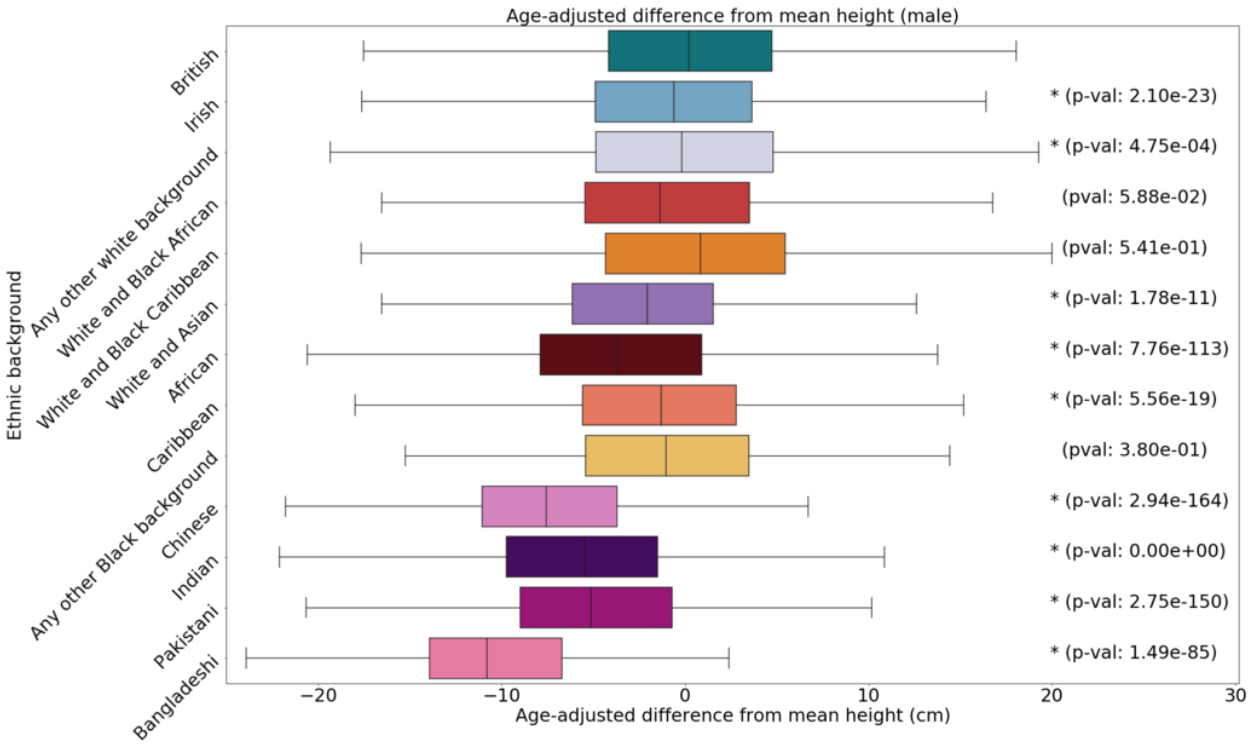



**Figure 2s36: Box plots of height in the UKBB by self-identified ethnicity (male).** Height by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

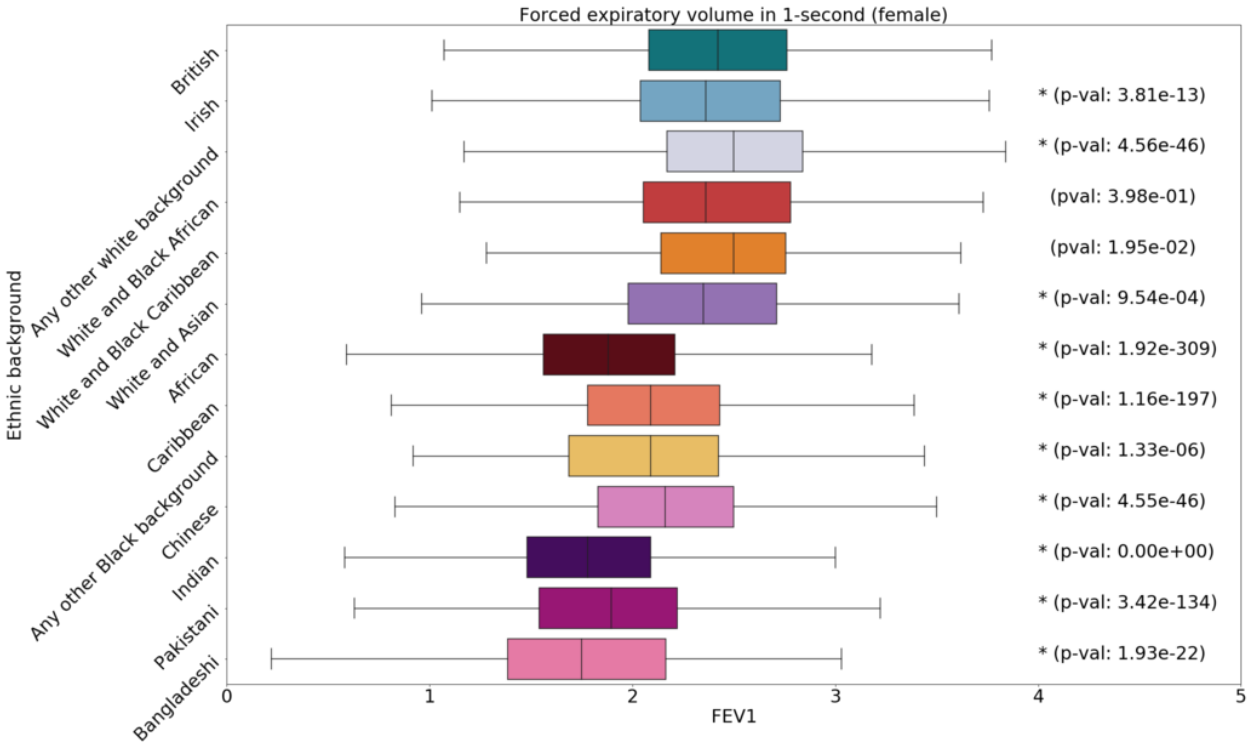



**Figure 2s37: Box plots of FEV1 in the UKBB by self-identified ethnicity (female).** FEV1 by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

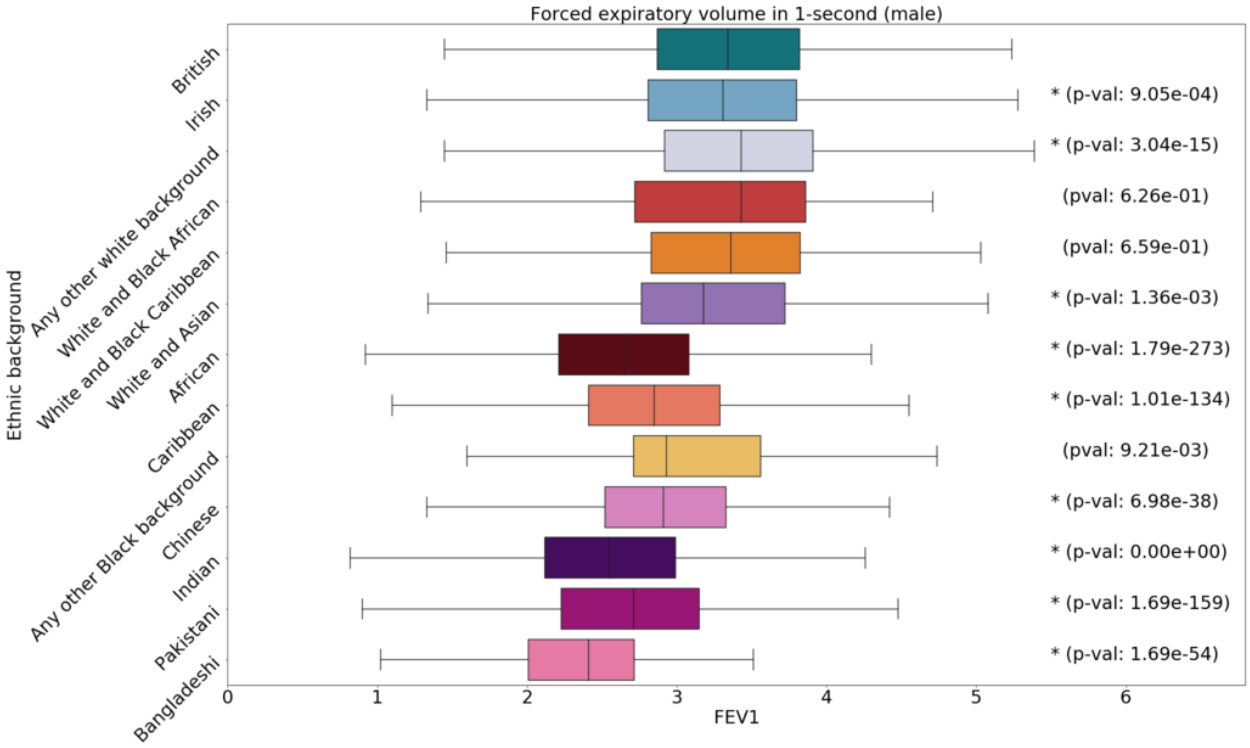



**Figure 2s38: Box plots of FEV1 in the UKBB by self-identified ethnicity (male).** FEV1 by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

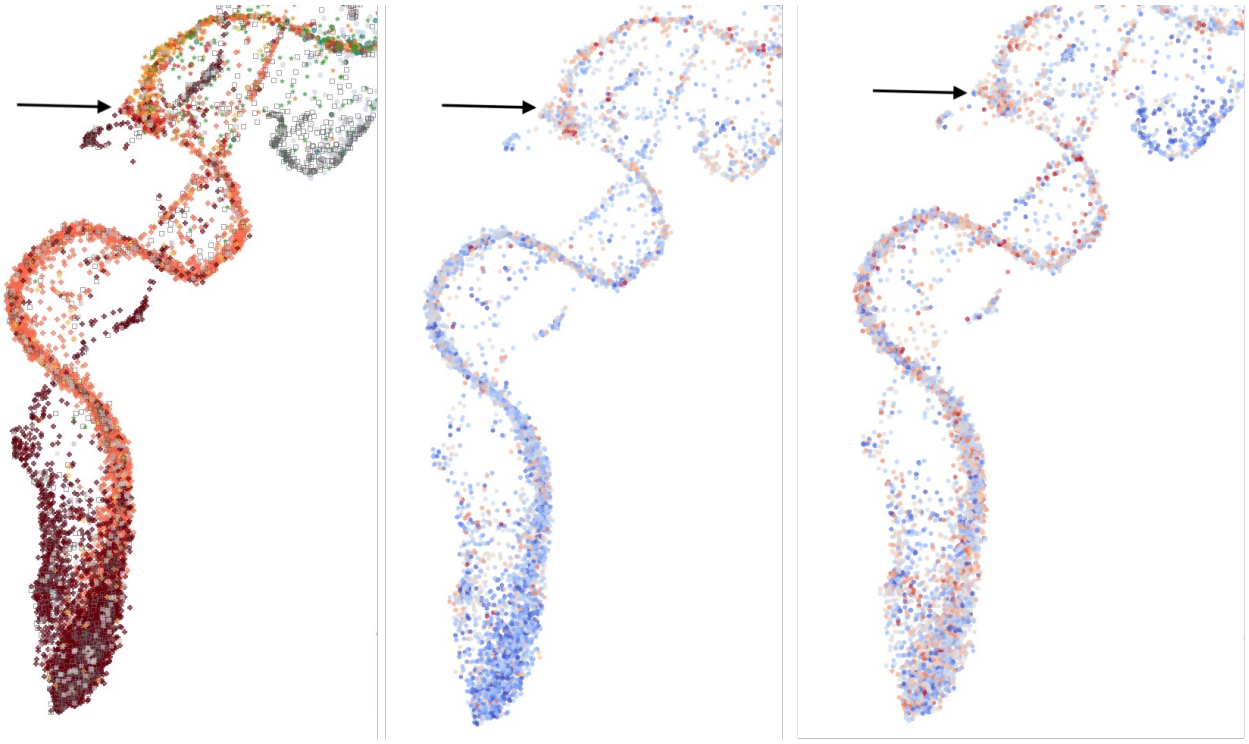


**Figure 2s39: Subset (left) of UKBB UMAP projection coloured by height, FEV1, and self-identified ethnicity.** Individuals of Black African, Black Caribbean, and mixed backgrounds (primarily White and Black Caribbean/African) coloured by self-identified ethnic background (left, from 2.3B), FEV1 (middle), and age-adjusted height (right). An arrow points to an area where the FEV1 distribution appears to change, corresponding to where the clusters contain more people with self-identified mixed backgrounds.

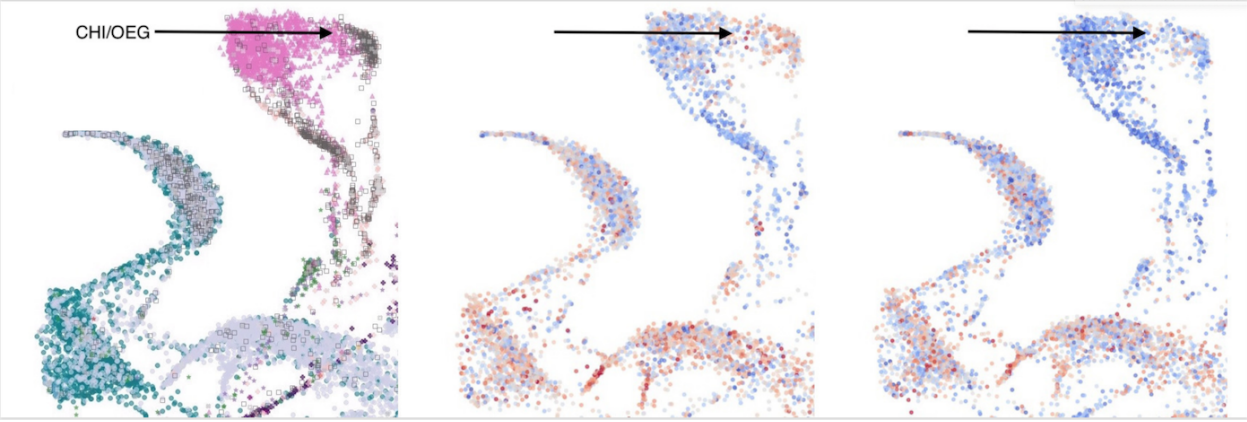



**Figure 2s40: Subset (top) of UKBB UMAP projection coloured by height, FEV1, and self-identified ethnicity.** Zoomed in section of 2.3B, focused on individuals with Chinese (CHI), White British (GBR), any other white background, or any other ethnic group (OEG) coloured by ethnicity (left), FEV1 (middle), and age-adjusted height (right). The OEG cluster next to the Chinese cluster appears redder on the middle panel, suggesting higher levels of FEV1.

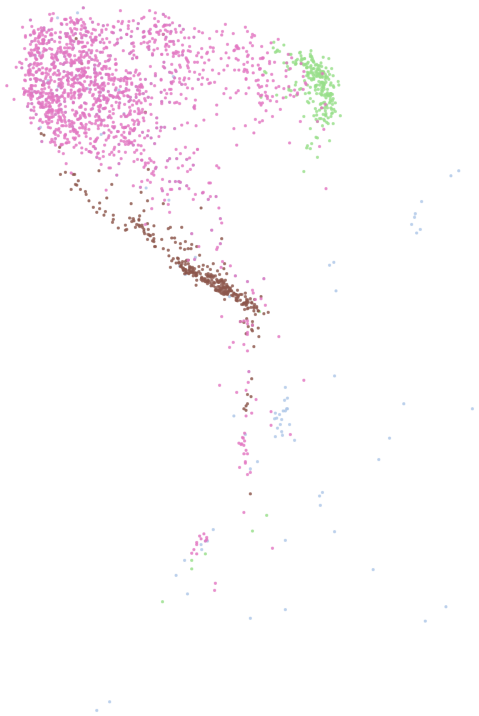


**Figure 2s41: East Asian individuals from UKBB UMAP projection selected for FEV1 investigation.** Individuals from the zoomed in section in 2s40 used in statistical testing, coloured the same as in 2s42. Brown, blue, and green represent those born in the Philippines, Malaysia, and Japan; pink represents those who self-identify as Chinese. The Chinese individuals were those who self-identified their ethnic background as Chinese, and the remaining populations were determined based on country of birth; the categorizations are mutually exclusive.

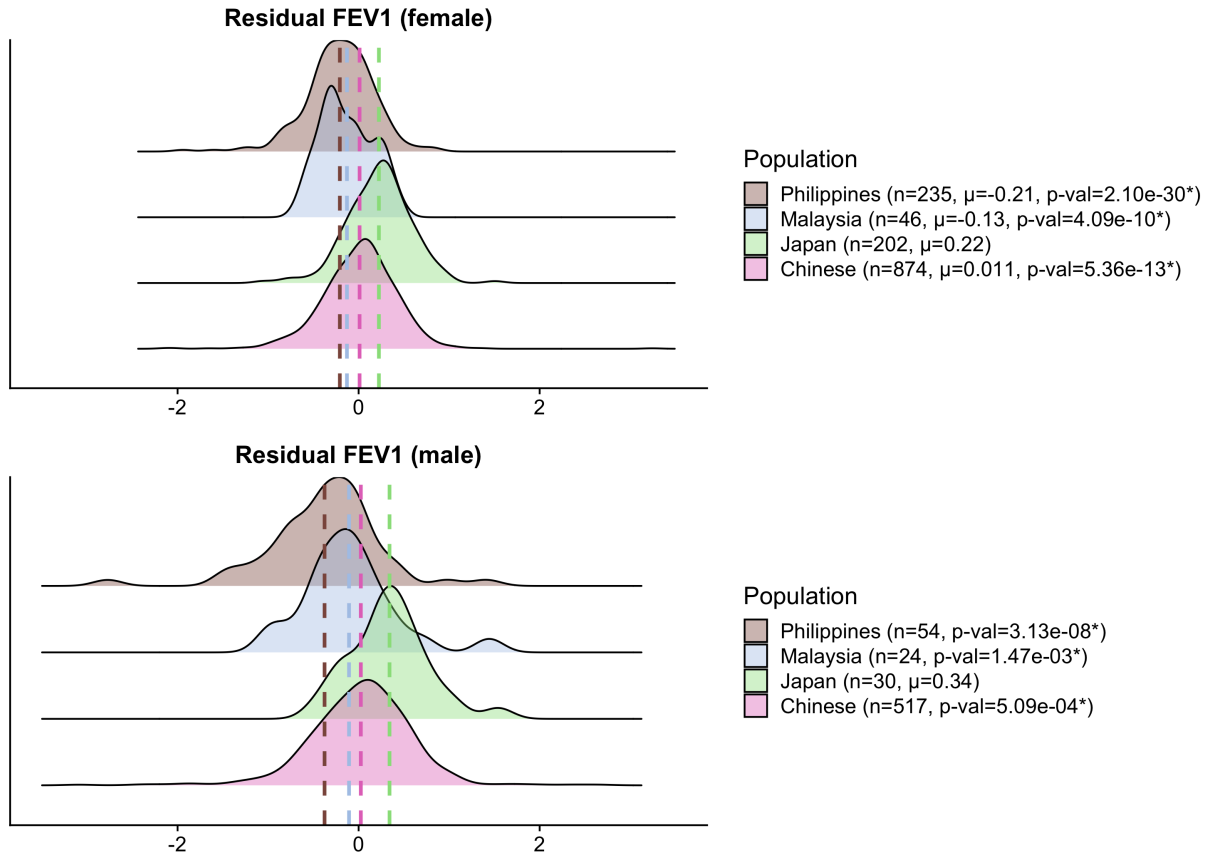



**Figure 2s42: Ridge plots of East Asian individuals from UKBB UMAP projection selected for FEV1 investigation.** Plots of the distributions of residual FEV1 by sex for East Asian populations, after adjusting for height, age, age$^2$, and sex through linear regression. Individuals were limited to those in the "Chinese/Other Ethnic Group" cluster from 2s40. The Chinese individuals were those who self-identified their ethnic background as Chinese, and the remaining populations were determined based on country of birth; the categorizations are mutually exclusive. Asterisks indicate significant difference from the Japanese population, using Welch's unpaired t-test with a Bonferroni correction for 3 groups. The dashed lines are the means of the distributions, and Japanese populations have consistently higher means.

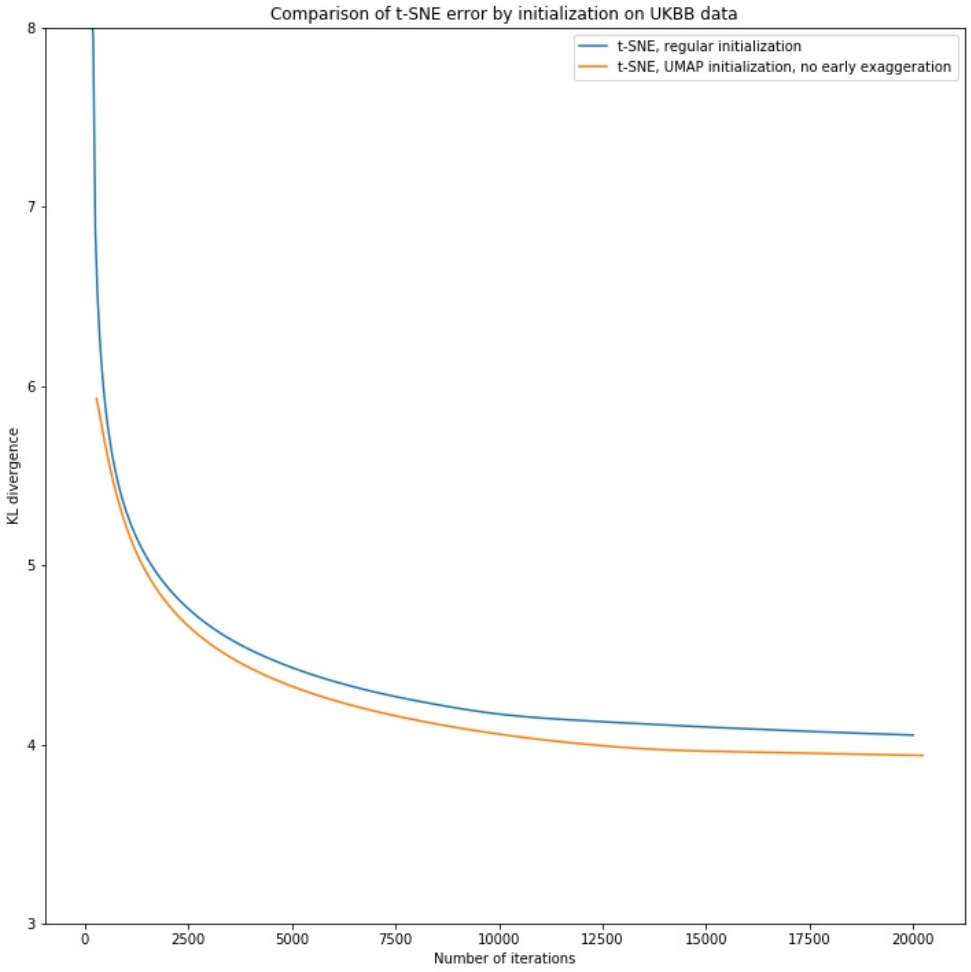



**Figure 2s43: Comparison of $t$-sne error by initialization on UKBB data.** Comparing the error terms of standard $t$-sne versus $t$-sne initialized with a UMAP embedding and no early exaggeration. Done on the UKBB dataset with 20000 iterations. The UMAP-initialized graph has been shifted by 230 iterations to approximate the 230 epochs UMAP uses for large datasets ($n > 10,000$).

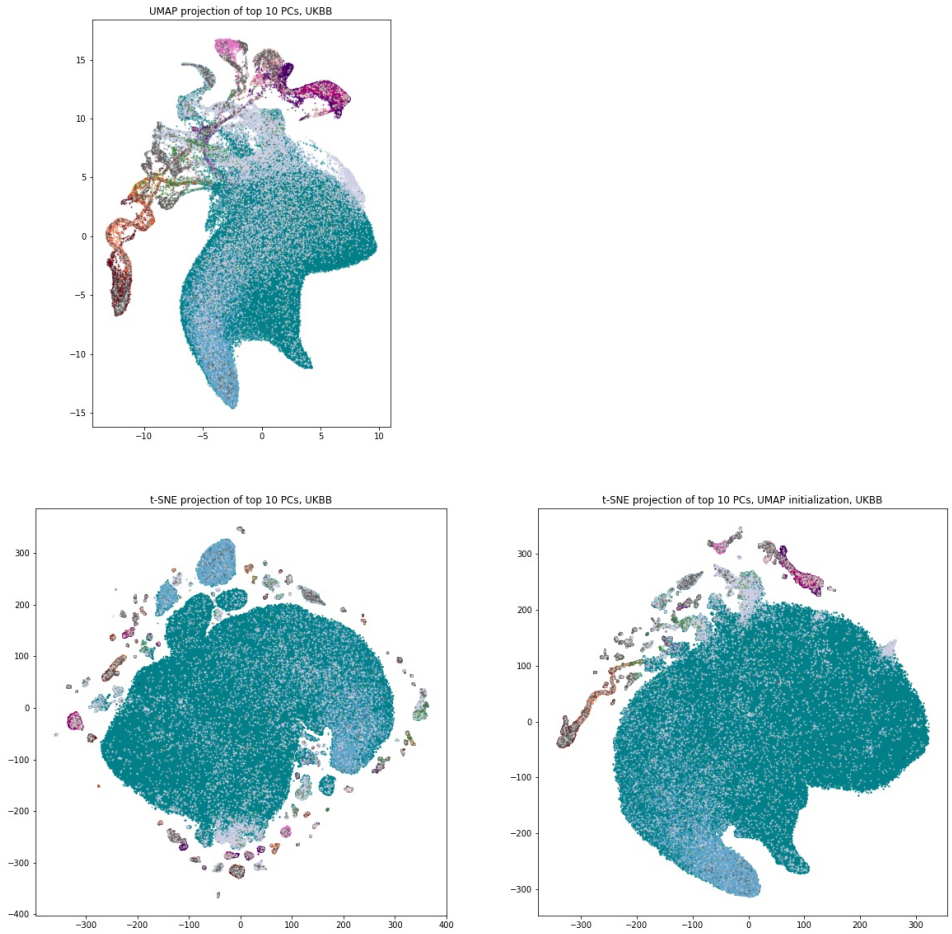



**Figure 2s44: Comparing visualizations of $t$-sne and UMAP of UKBB data by initialization.** Comparing the visualizations of UMAP, standard $t$-sne, and $t$-sne initialized with a UMAP projection, on the top 10 principal components of the UKBB. $t$-sne used 20000 iterations.

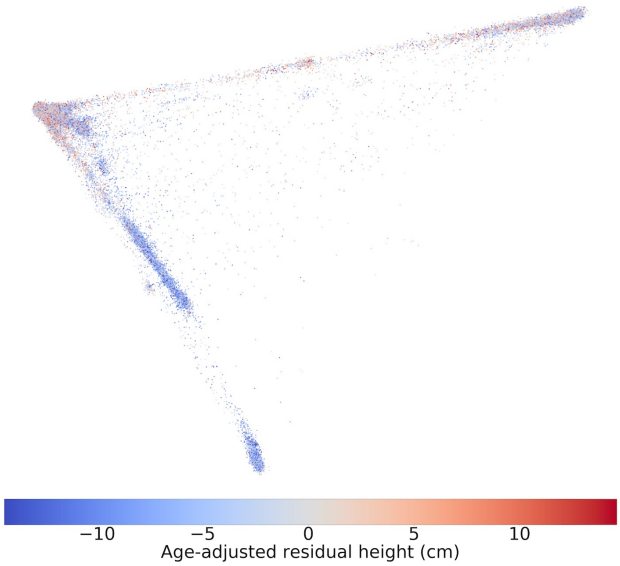



**Figure 2s45: PCs 1 and 2 of the UKBB coloured by height (female).** Principal components 1 and 2 from the UKBB, coloured by age-adjusted residual height (female). Data has been randomized as explained in the materials and methods section.

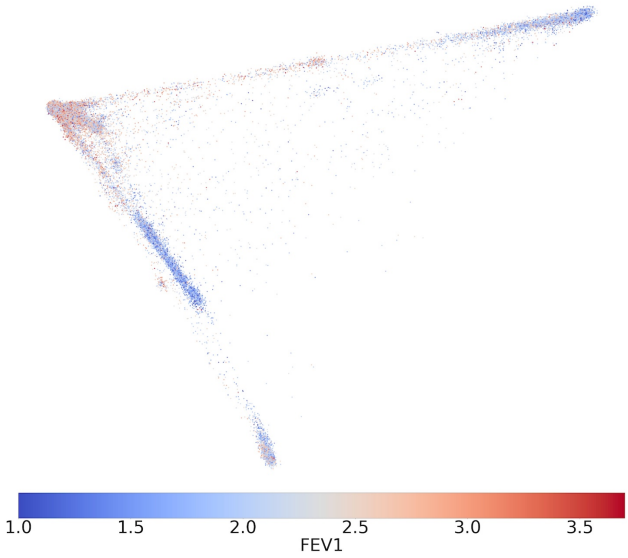



**Figure 2s46: PCs 1 and 2 of the UKBB coloured by FEV1 (female).** Principal components 1 and 2 from the UKBB, coloured by FEV1 (female). Data has been randomized as explained in the materials and methods section.

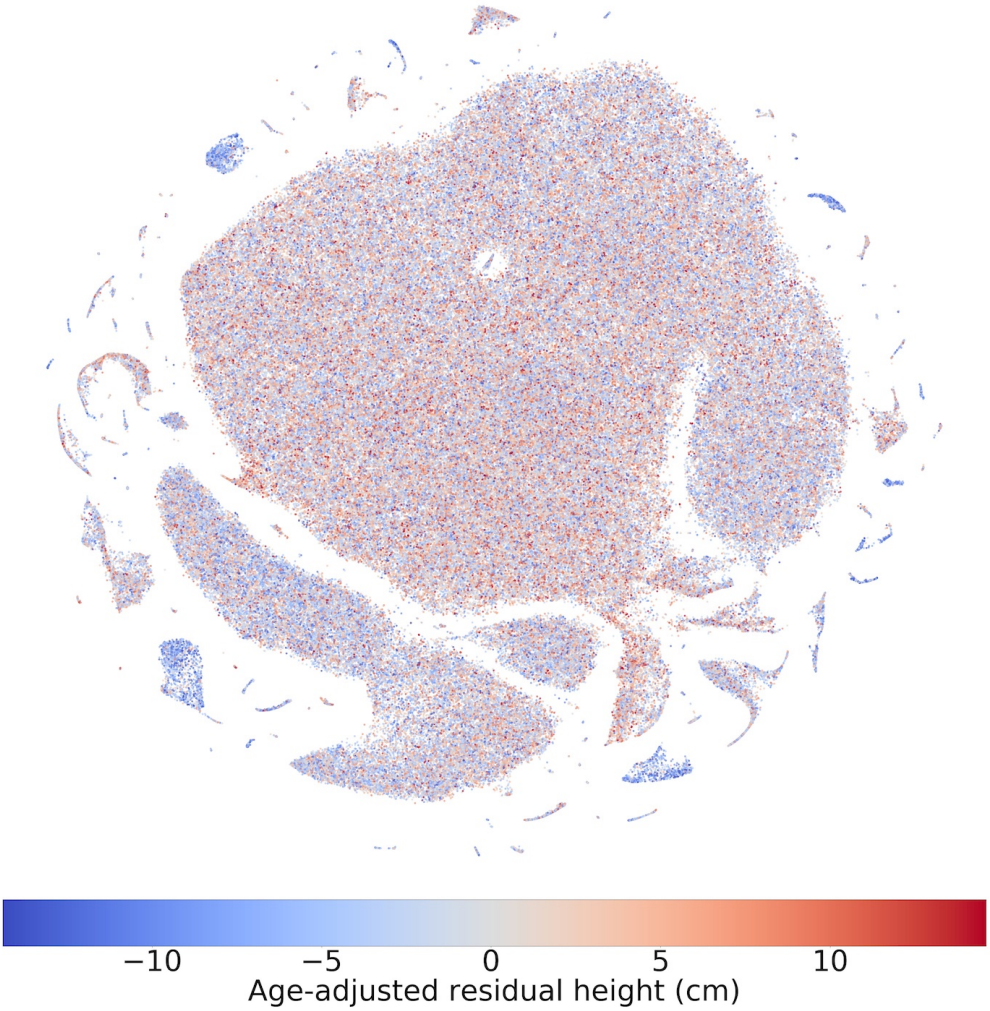



**Figure 2s47: $t$-sne projection of UKBB data coloured by height (female).** $t$-sne on the first 10 principal components from the UKBB, coloured by age-adjusted residual height (female). Data has been randomized as explained in the materials and methods section.

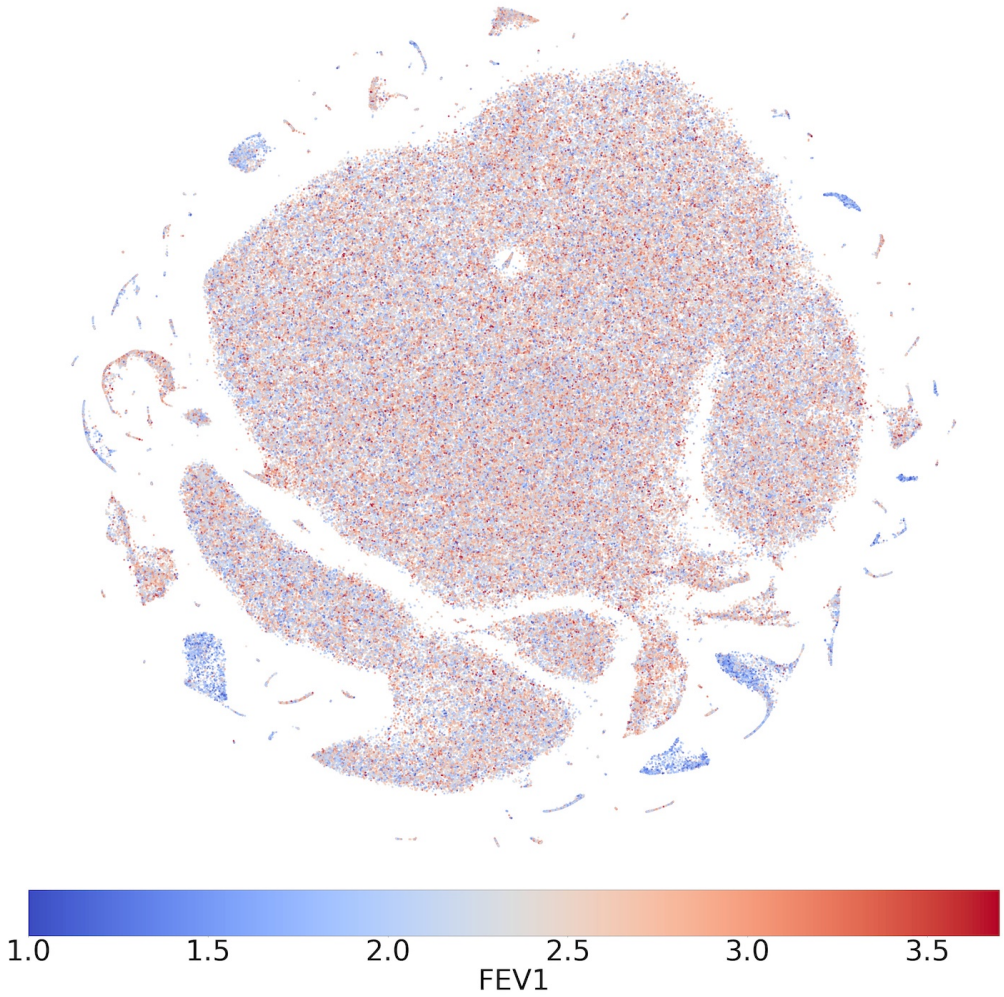



**Figure 2s48: $t$-sne projection of UKBB data coloured by FEV1 (female).** $t$-sne on the first 10 principal components from the UKBB, coloured by FEV1 (female). Data has been randomized as explained in the materials and methods section.

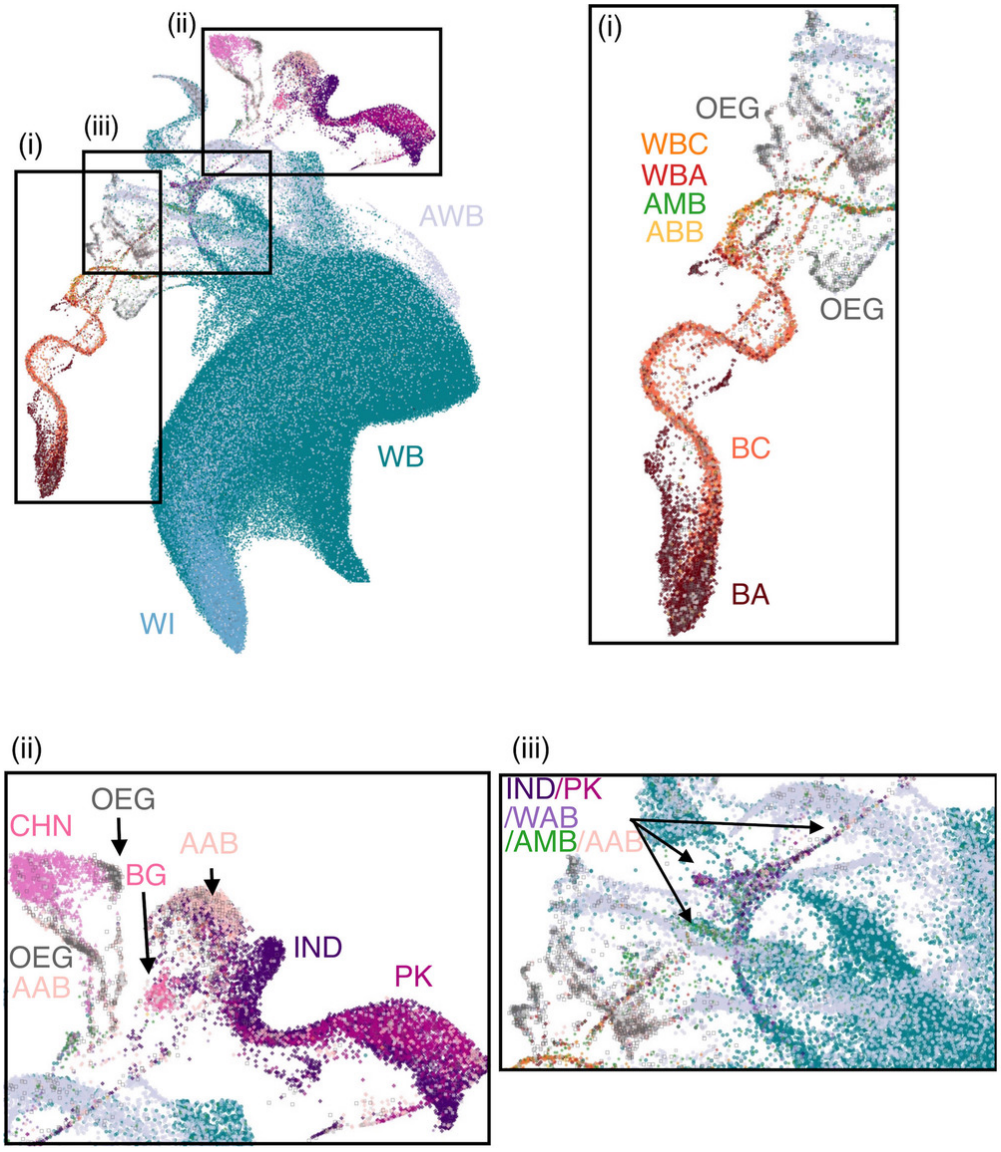



**Figure 2s49: Zoomed in views of UMAP projection of UKBB data, coloured by self-identified ethnicity.** Zoomed in areas of 2.3B. Sections (i) and (ii) respectively focus on the African and Asian superpopulations, and section (iii) focuses on an area with individuals from many ethnic backgrounds. Noticeable clusters of unidentified ethnic backgrounds appear and are labelled "OEG" ("Other Ethnic Group").

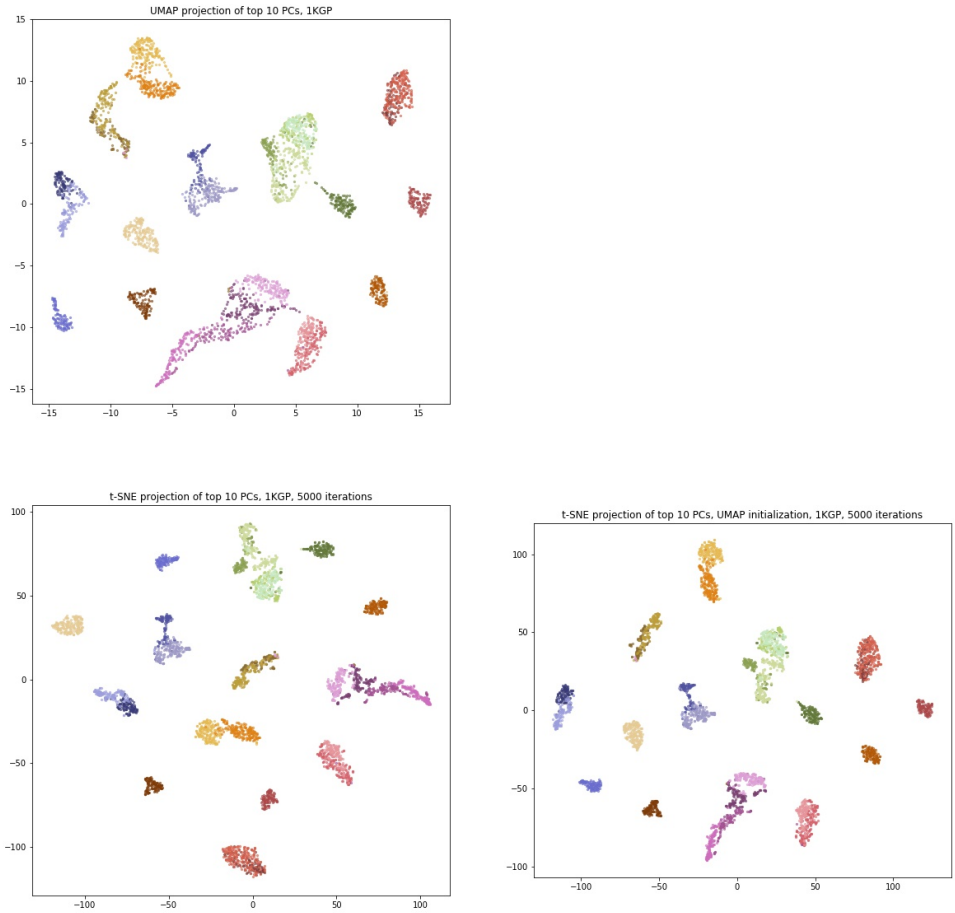



**Figure 2s50: Comparing visualizations of $t$-sne and UMAP of 1KGP data by initialization.** Comparing the visualizations of UMAP, standard $t$-sne, and $t$-sne initialized with a UMAP projection, on the top 10 principal components of the 1KGP. $t$-sne used 5000 iterations. Initializing $t$-sne with UMAP breaks the continuous structure of the projection and instead forms many small clusters.

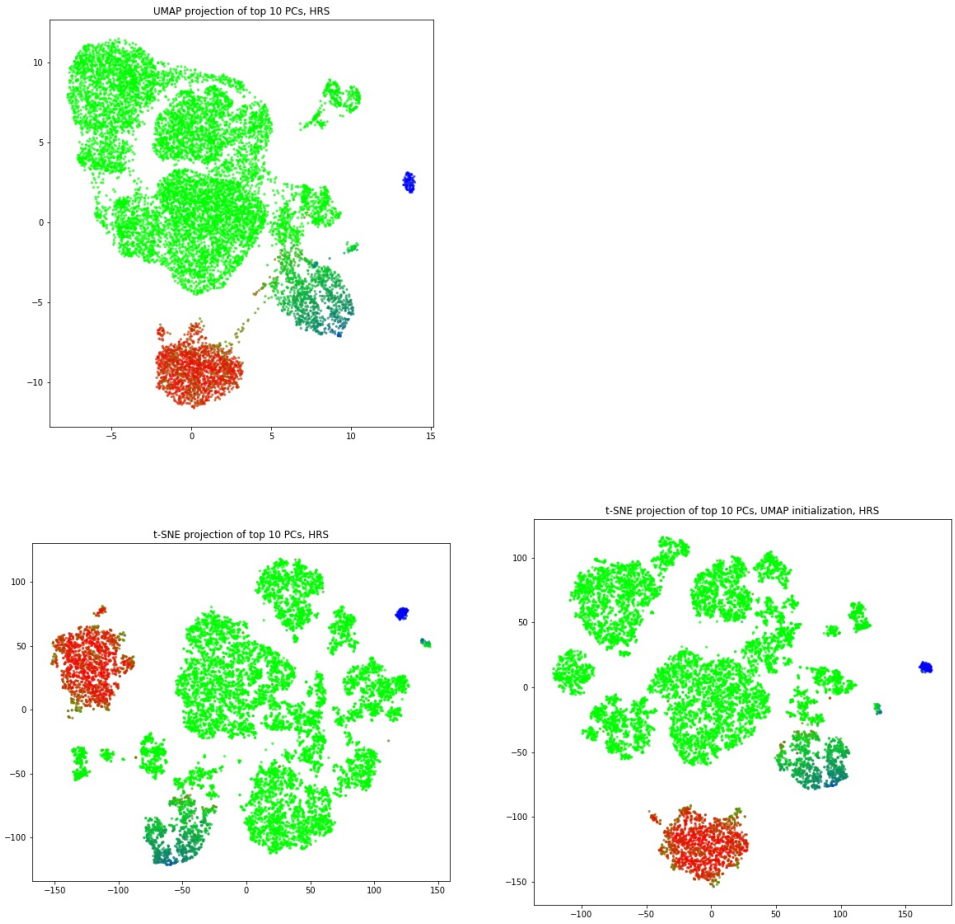



**Figure 2s51: Comparing visualizations of $t$-sne and UMAP of HRS data by initialization.** Comparing the visualizations of UMAP, standard $t$-sne, and $t$-sne initialized with a UMAP projection, on the top 10 principal components of the HRS. $t$-sne used 5000 iterations.

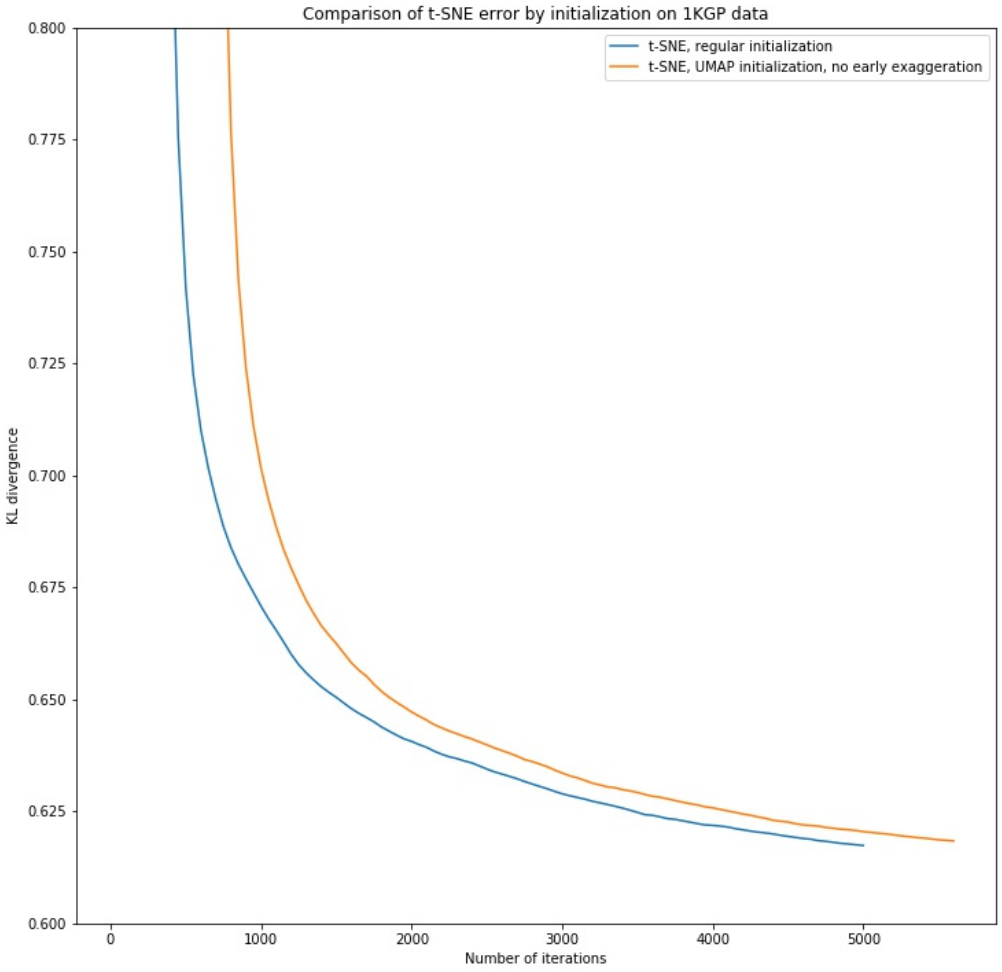



**Figure 2s52: Comparison of $t$-sne error by initialization on 1KGP data.** Comparing the error terms of standard $t$-sne versus $t$-sne initialized with a UMAP embedding and no early exaggeration. Done on the 1KGP dataset with 5000 iterations. The UMAP-initialized graph has been shifted by 600 iterations to approximate the 600 epochs UMAP uses for small datasets ($n <= 10,000$).

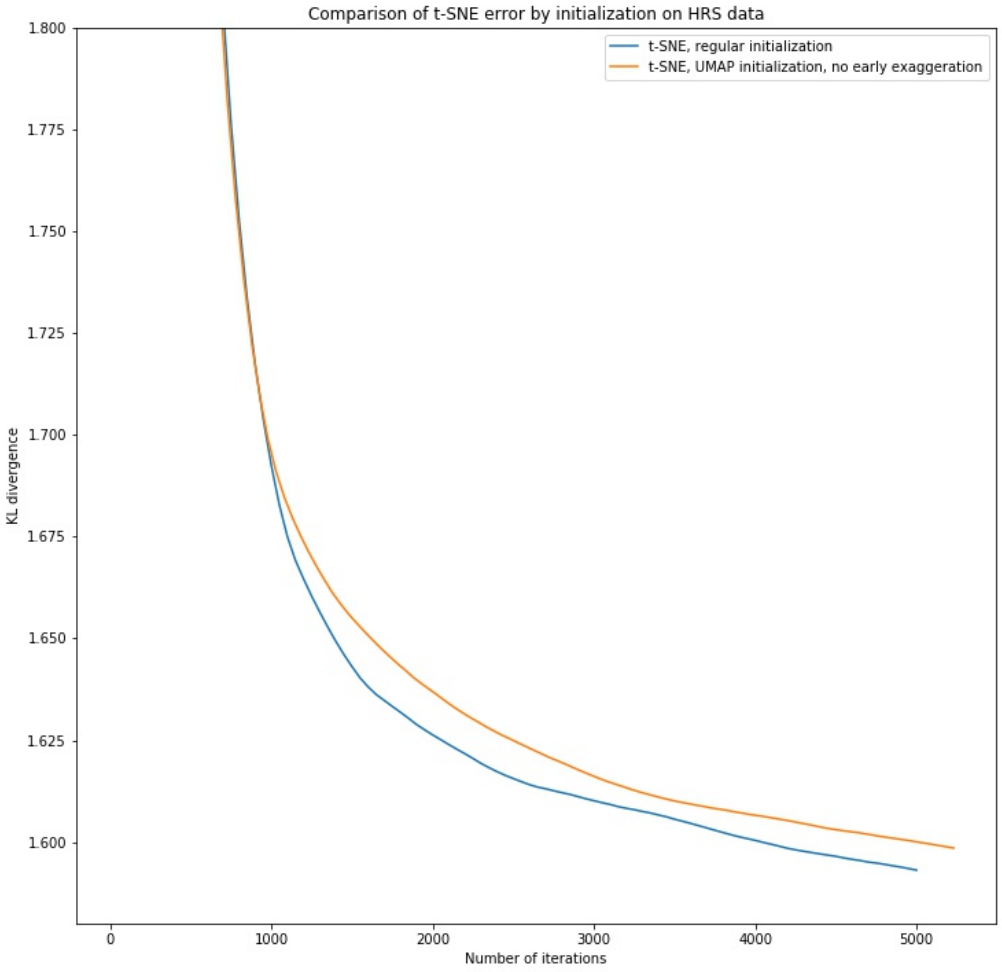



**Figure 2s53: Comparison of $t$-sne error by initialization on HRS data.** Comparing the error terms of standard $t$-sne versus $t$-sne initialized with a UMAP embedding and no early exaggeration. Done on the HRS dataset with 5000 iterations. The UMAP-initialized graph has been shifted by 230 iterations to approximate the 230 epochs UMAP uses for large datasets ($n > 10,000$).

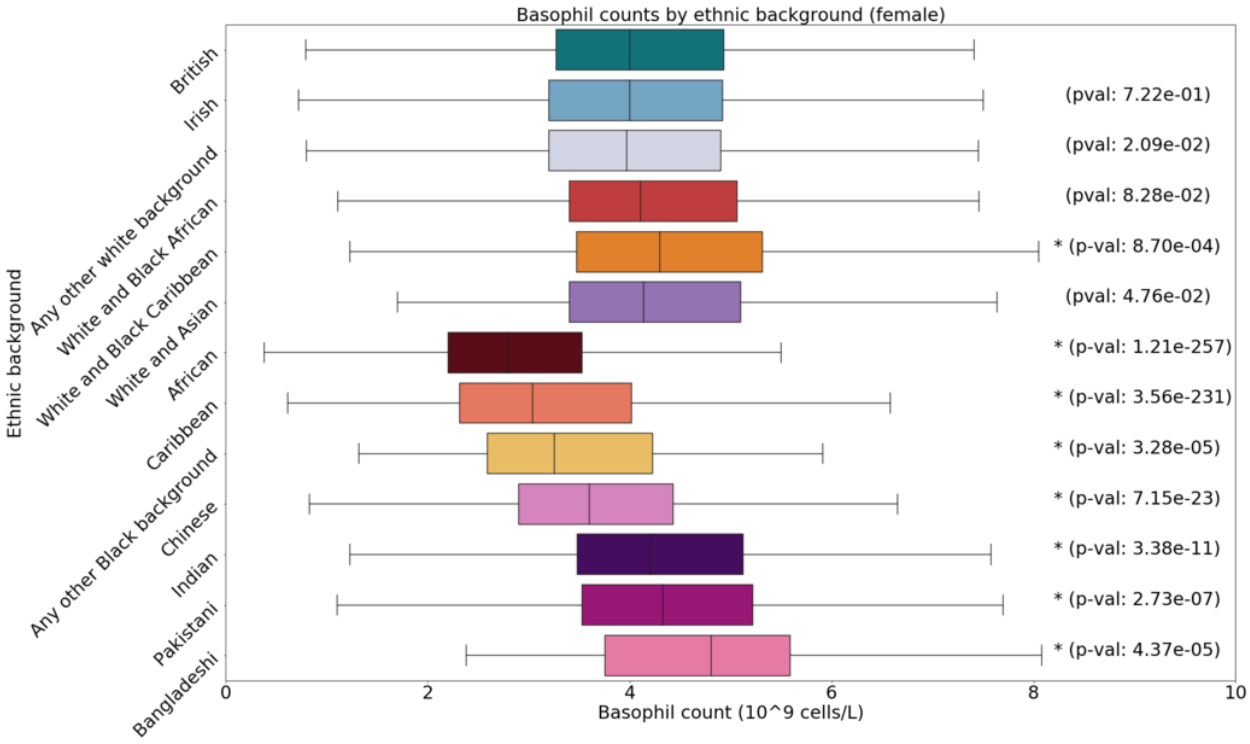



**Figure 2s54: Box plots of basophil count in the UKBB by self-identified ethnicity (female).** Basophil counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

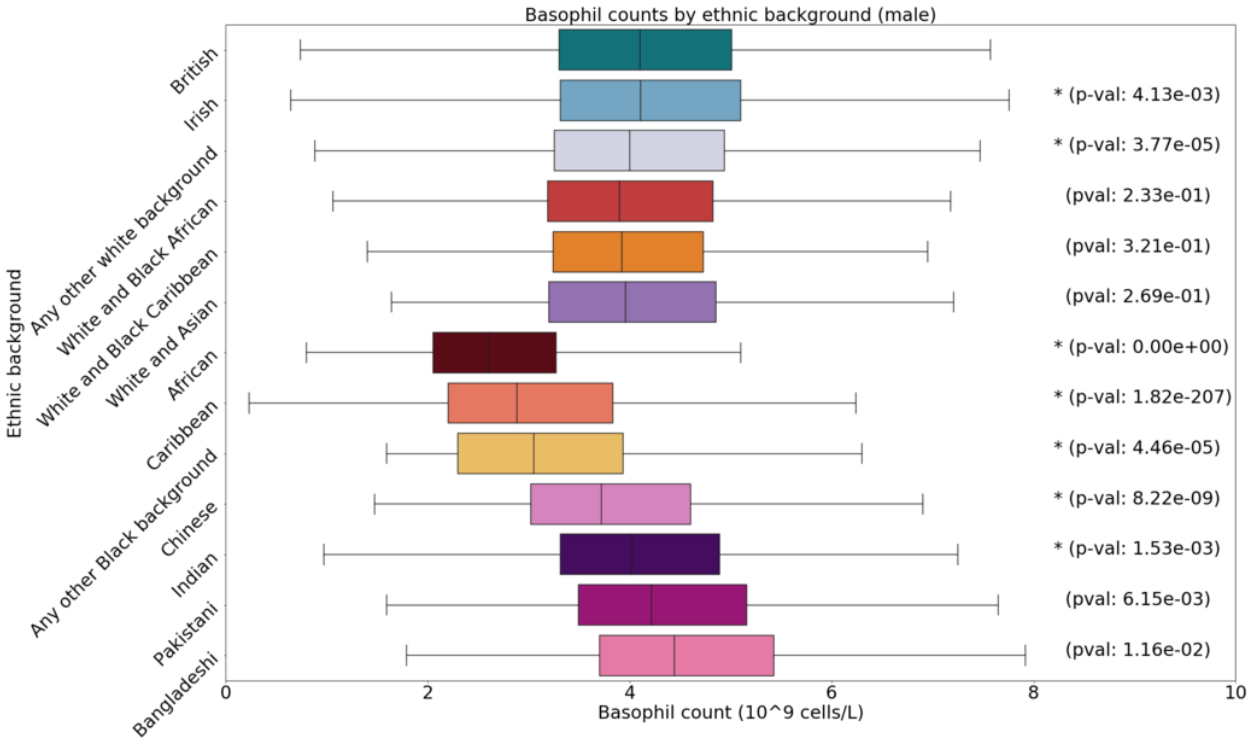



**Figure 2s55: Box plots of basophil count in the UKBB by self-identified ethnicity (male).** Basophil counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

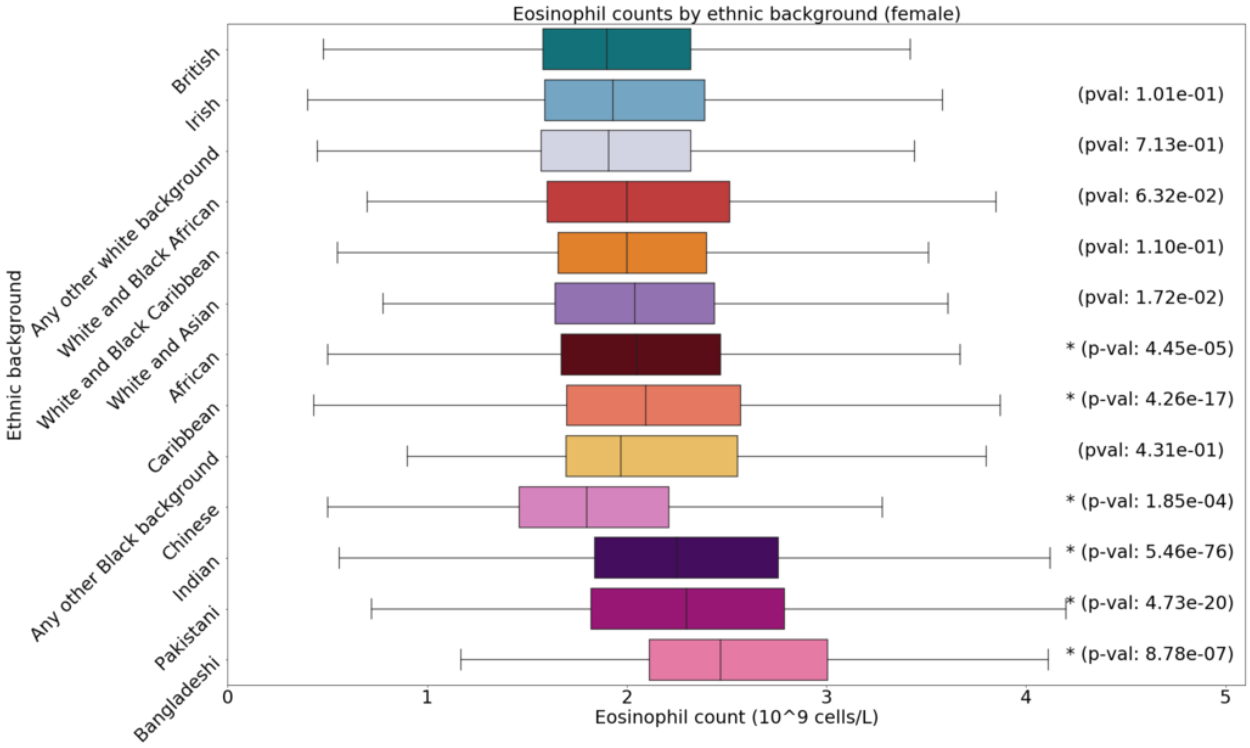



**Figure 2s56: Box plots of eosinophil count in the UKBB by self-identified ethnicity (female).** Eeosinophil counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

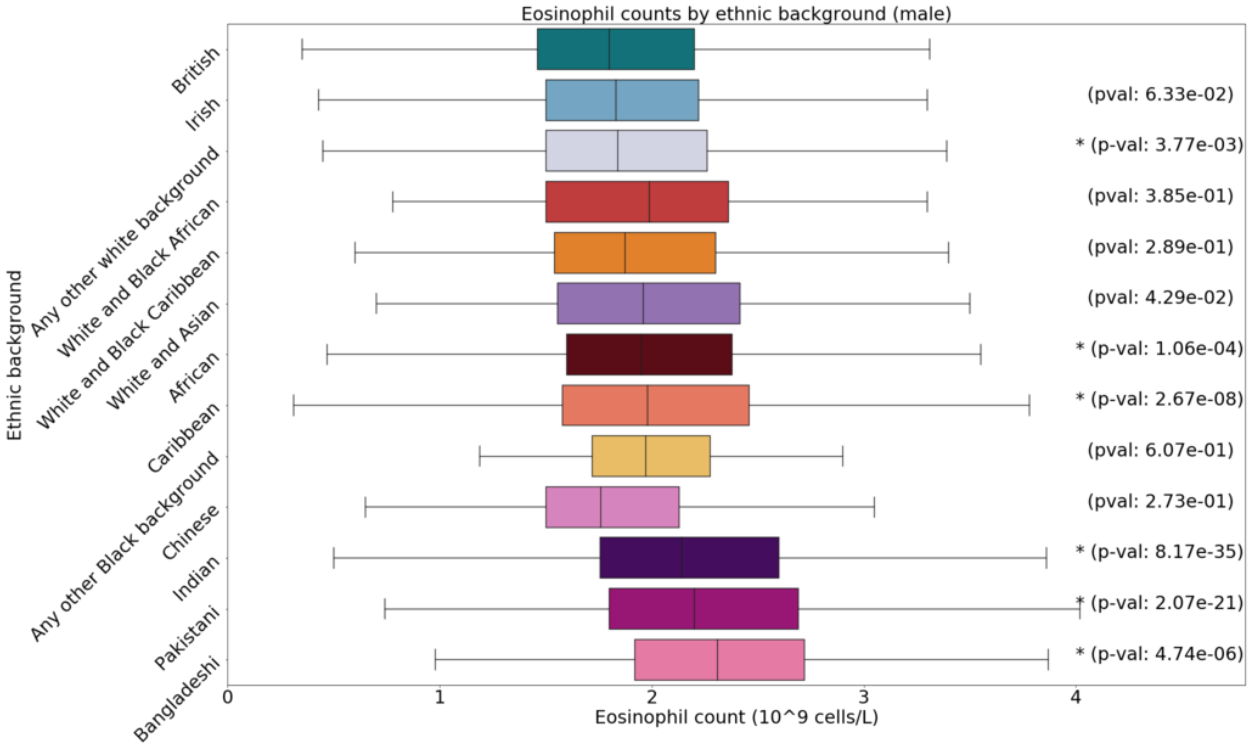



**Figure 2s57: Box plots of eosinophil count in the UKBB by self-identified ethnicity (male).** Eosinophil counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

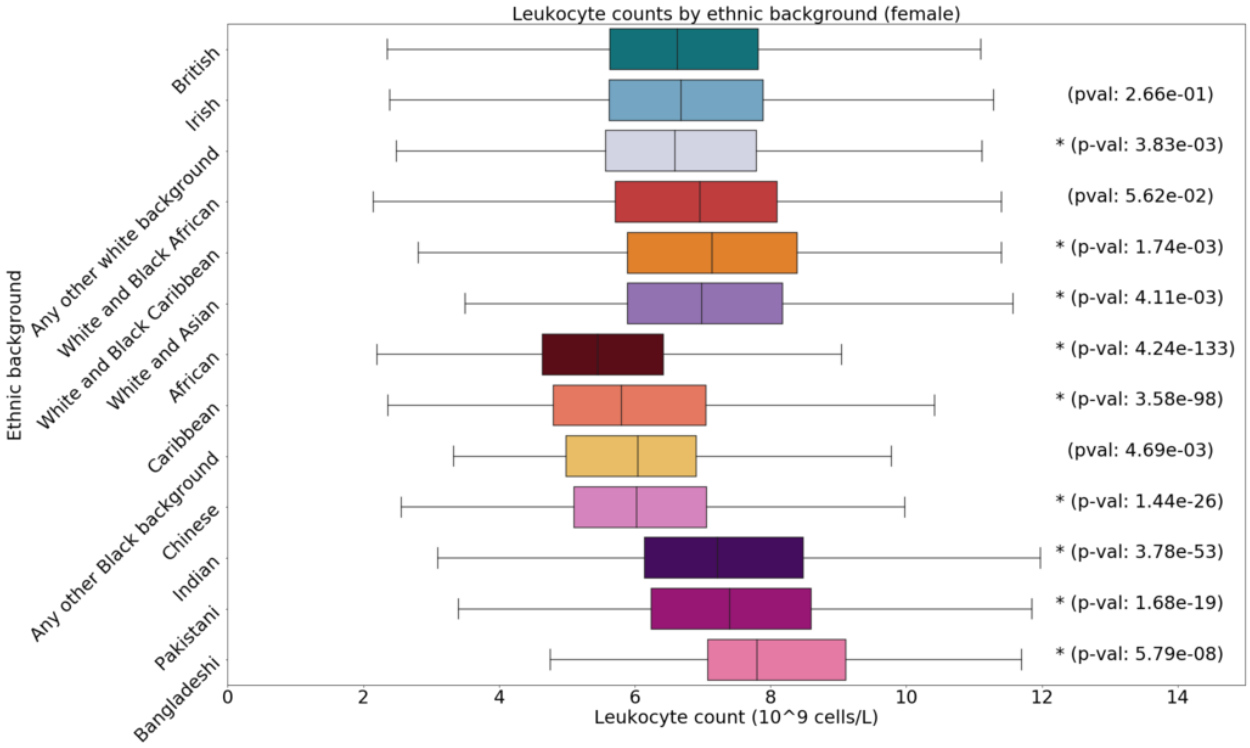



**Figure 2s58: Box plots of leukocyte count in the UKBB by self-identified ethnicity (female).** Leukocyte counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

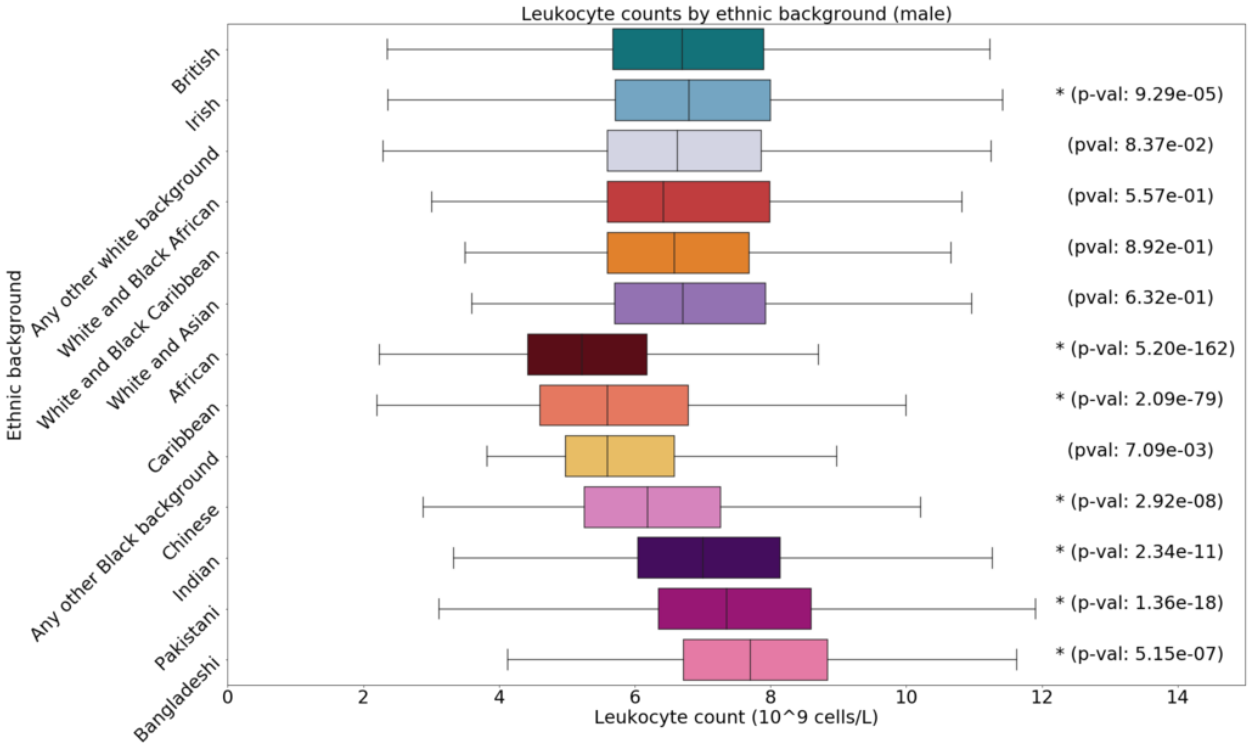



**Figure 2s59: Box plots of leukocyte count in the UKBB by self-identified ethnicity (male).** Leukocyte counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

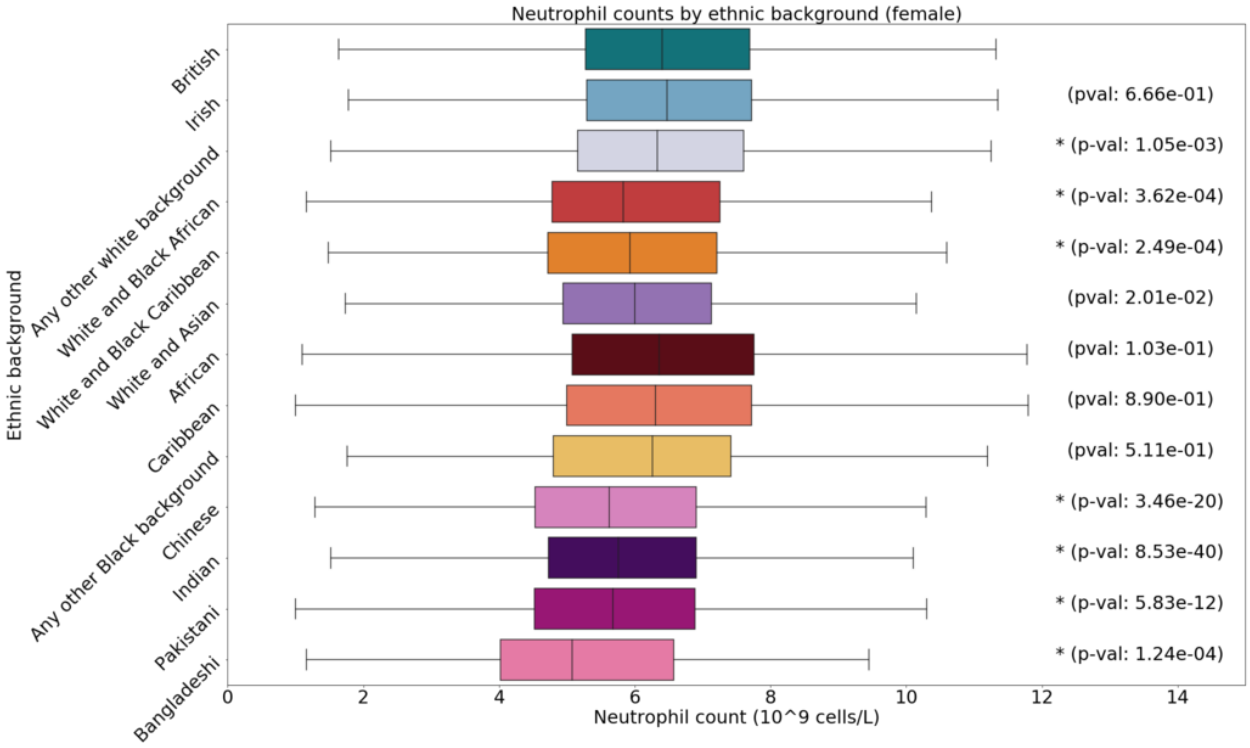



**Figure 2s60: Box plots of neutrophil count in the UKBB by self-identified ethnicity (female).** Neutrophil counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

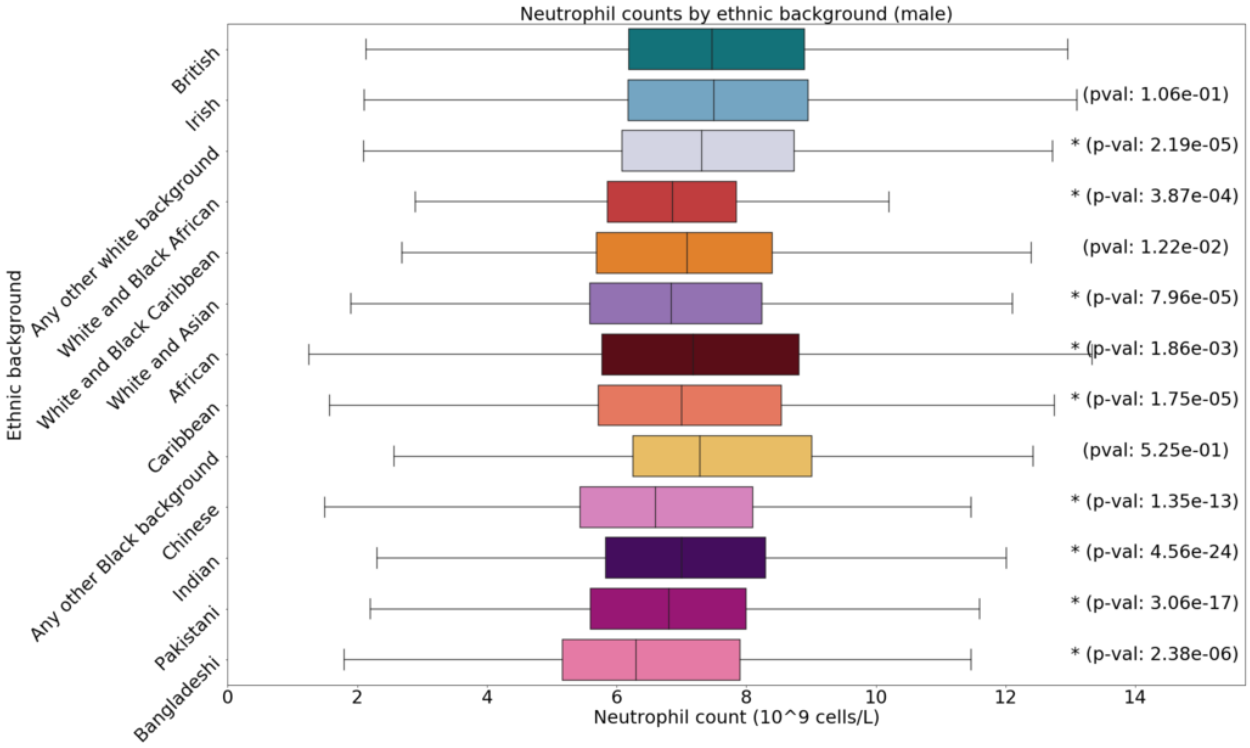



**Figure 2s61: Box plots of neutrophil count in the UKBB by self-identified ethnicity (male).** Neutrophil counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

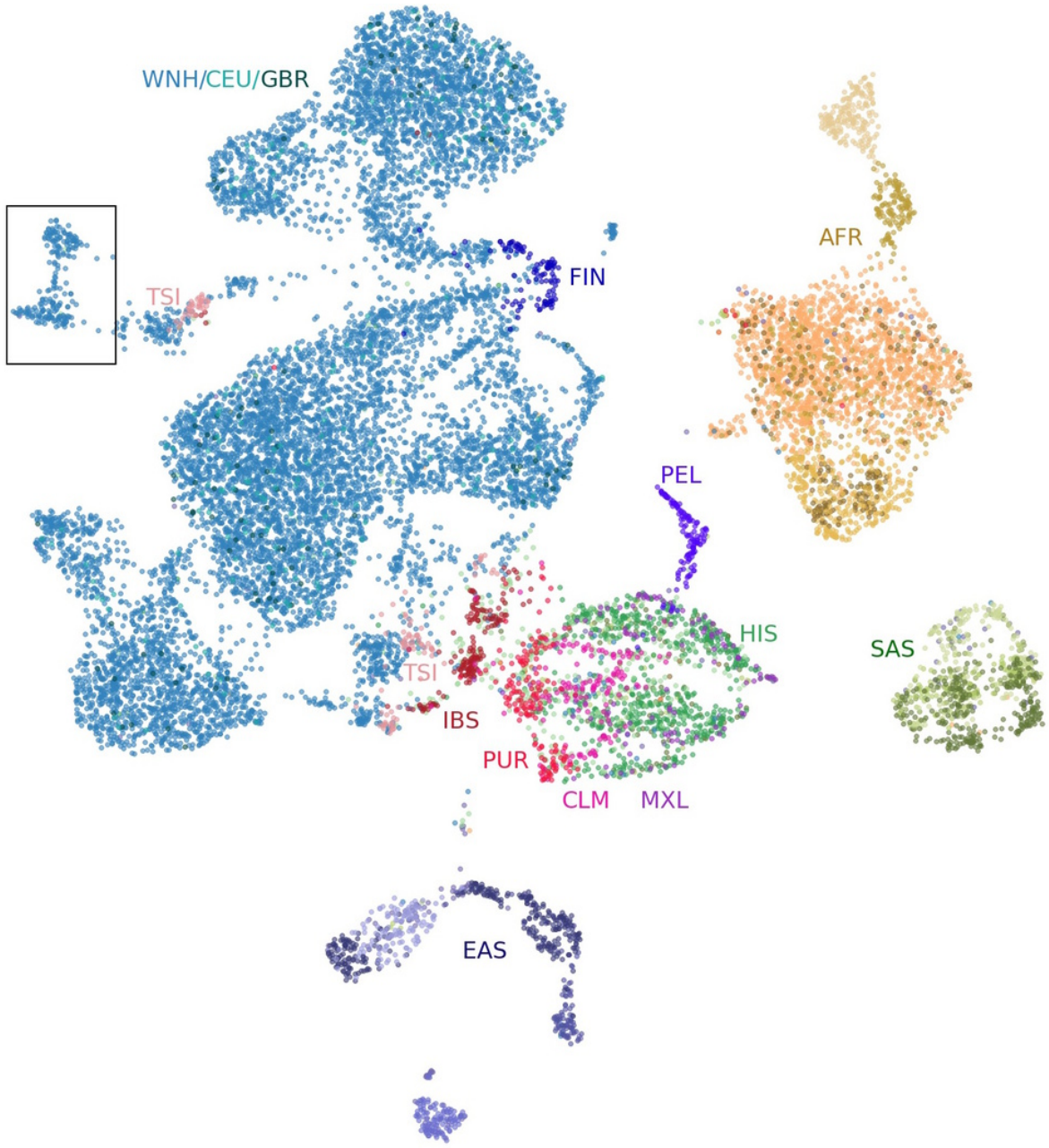



**Figure 2s62: UMAP projection of combined HRS and 1KGP data.** UMAP projection of the top 10 principal components of the combined HRS and 1KGP datasets. One cluster (in the box) does not group with any of the 1KGP populations. A cluster of Finnish (FIN) individuals consistently appears in the "White Not Hispanic" (WNH) group. Groups of Central and South American populations from the 1KGP (CLM, Colombian; MXL, Mexican; PEL, Peruvian; PUR, Puerto Rican) form nearby or within the HRS Hispanic cluster (HIS). Iberian individuals (IBS) cluster near the Hispanic population. Toscani individuals (TSI) form some small clusters and sometimes appear near the Iberian and Hispanic populations. Individuals with British/Scottish (GBR) or Northern/Western European ancestry (CEU) are scattered throughout the WNH clusters. Individuals with African ancestry from the 1KGP group with Black Americans from the HRS (AFR). Similar population groupings occur with South Asian (SAS) and East Asian (EAS) individuals.

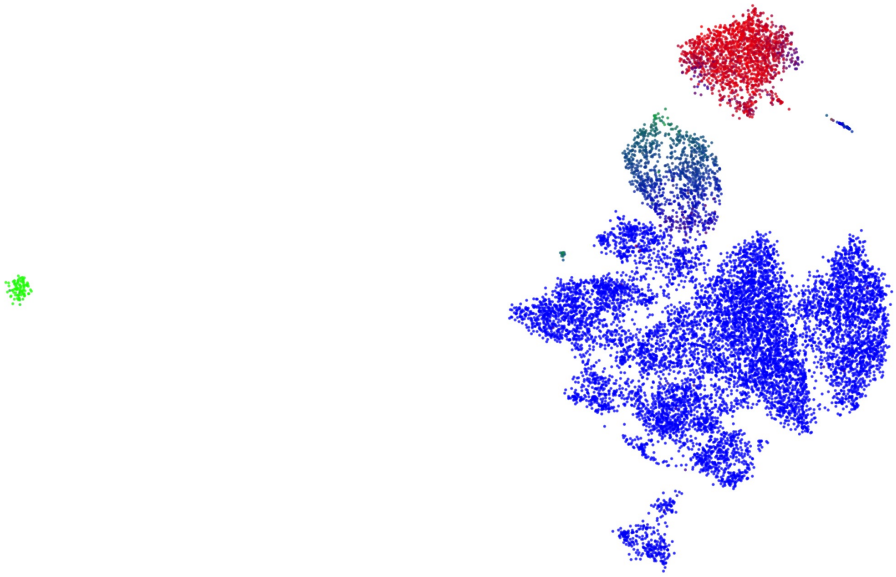


**Figure 2s63: Alternate colouring of 2s7.** An alternate colouring of 2s7. Here red, green, and blue correspond to African, Asian/Native American, and European ancestry, respectively.

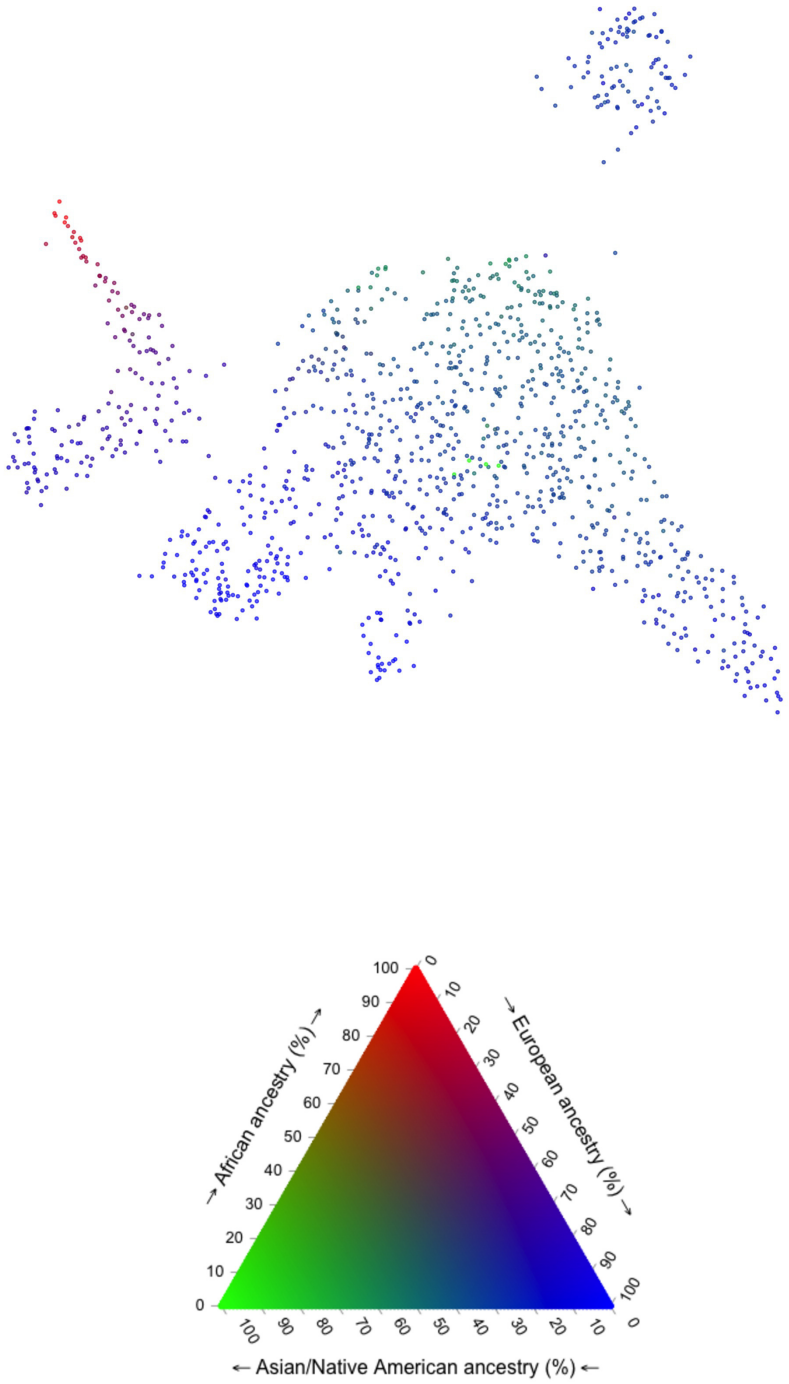



**Figure 2s64: Alternate colouring of 2s11.** An alternate colouring of 2s11. Here red, green, and blue correspond to African, Asian/Native American, and European ancestry, respectively.

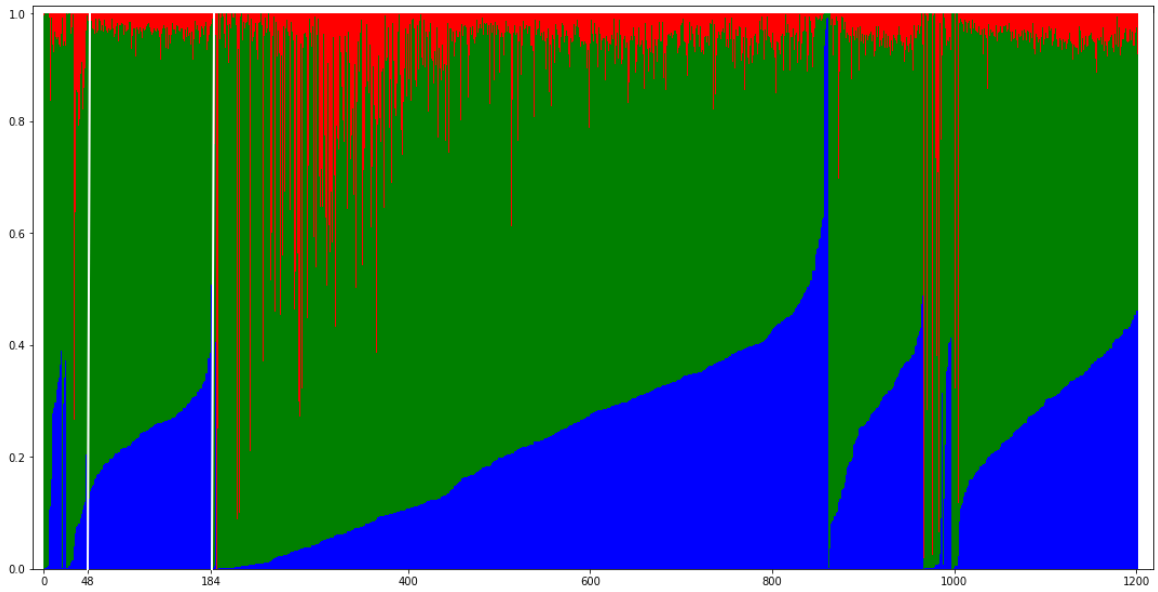



**Figure 2s65: Admixture plot of Hispanic individuals in the HRS.** Admixture plot of Hispanic individuals in the HRS. Individuals born in the Mountain census region fall between the white lines (indices 48 to 184).

| Number of PCs | Variance explained | Number of PCs | Variance explained |
|---|---|---|---|
| 2 | 13.6% | 1000 | 57.0% |
| 3 | 14.7% | 1100 | 60.0% |
| 4 | 15.5% | 1200 | 62.8% |
| 5 | 15.7% | 1300 | 65.4% |
| 6 | 15.8% | 1400 | 68.0% |
| 7 | 15.9% | 1500 | 70.5% |
| 8 | 16.0% | 1600 | 73.0% |
| 9 | 16.1% | 1700 | 75.3% |
| 10 | 16.2% | 1800 | 77.5% |
| 11 | 16.2% | 1900 | 79.7% |
| 12 | 16.3% | 2000 | 81.8% |
| 13 | 16.4% | 2100 | 83.8% |
| 14 | 16.5% | 2200 | 85.8% |
| 15 | 16.5% | 2300 | 87.6% |
| 30 | 17.5% | 2400 | 89.4% |
| 50 | 18.6% | 2500 | 91.1% |
| 100 | 21.3% | 2600 | 92.7% |
| 200 | 26.4% | 2700 | 94.2% |
| 300 | 31.1% | 2800 | 95.3% |
| 400 | 35.5% | 2900 | 96.3% |
| 500 | 39.7% | 3000 | 97.1% |
| 600 | 43.7% | 3100 | 97.8% |
| 700 | 47.4% | 3200 | 98.5% |
| 800 | 50.8% | 3300 | 99.2% |
| 900 | 54.0% | 3400 | 99.7% |

**Table 2s1:** Variance explained in the 1KGP data by the number of principal components used.

## References


[1] Daniel John Lawson et al. "Inference of population structure using dense haplotype data". In: *PLoS genetics* 8.1 (2012), e1002453.

[2] John Novembre and Benjamin M Peter. "Recent advances in the study of fine-scale population structure in humans". In: *Current opinion in genetics & development* 41 (2016), pp. 98–105.

[3] Jeffrey P Spence et al. "Inference of population history using coalescent HMMs: review and outlook". In: *Current opinion in genetics & development* 53 (2018), pp. 70–76.

[4] Nick Patterson, Alkes L Price, and David Reich. "Population Structure and Eigenanalysis". In: *PLOS Genetics* 2 (Dec. 2006), pp. 1–20.

[5] Garrett Hellenthal et al. "A Genetic Atlas of Human Admixture History". In: *Science* 343.6172 (2014), pp. 747–751.

[6] Gil McVean. "A genealogical interpretation of principal components analysis". In: *PLoS genetics* 5.10 (2009), e1000686.

[7] Abra Brisbin et al. "PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations". In: *Human biology* 84.4 (2012), p. 343.

[8] John Novembre et al. "Genes mirror geography within Europe". In: *Nature* 456 (2008), pp. 98–101.

[9] Matthew R Nelson et al. "The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research". In: *The American Journal of Human Genetics* 83.3 (2008), pp. 347–358.

[10] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605.

[11] Alexander Platzer. "Visualization of SNPs with t-SNE". In: *PloS one* 8.2 (2013), e56883.

[12] 1000 Genomes Project Consortium. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), p. 68.

[13] Wentian Li et al. "Application of t-SNE to human genetic data". In: *Journal of Bioinformatics and Computational Biology* 15.04 (2017). PMID: 28718343, p. 1750017.

[14] L. McInnes and J. Healy. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[15] Etienne Becht et al. "Dimensionality reduction for visualizing single-cell data using UMAP". In: *Nature biotechnology* 37.1 (2019), pp. 38–44.

[16] F Thomas Juster and Richard Suzman. "An overview of the Health and Retirement Study". In: *Journal of Human Resources* (1995), S7–S56.

[17] Cathie Sudlow et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLoS medicine* 12.3 (2015), e1001779.

[18] David Reich et al. "Reconstructing Indian population history". In: *Nature* 461.7263 (2009), pp. 489–494.

[19] 23andMe. *23andMe Tests New Ancestry Breakdown in Central and South Asia*. 2019.

[20] Eunjung Han et al. "Clustering of 770,000 genomes reveals post-colonial population structure of North America". In: *Nature communications* 8.1 (2017), p. 14238.

[21] I.King Jordan, Lavanya Rishishwar, and Andrew B Conley. "Cryptic Native American ancestry recapitulates population-specific migration and settlement of the continental United States". In: *bioRxiv* (2018).

[22] Stephen Leslie et al. "The fine-scale genetic structure of the British population". In: *Nature* 519.7543 (2015), p. 309.

[23] Matthew R Robinson et al. "Population genetic differentiation of height and body mass index across Europe". In: *Nature genetics* 47.11 (2015), p. 1357.

[24] AU Komlos. *Stature, living standards, and economic development: Essays in anthropometric history*. 1994.

[25] Philip H Quanjer et al. *Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations*. 2012.

[26] Victor E Ortega and Rajesh Kumar. "The effect of ancestry and genetic variation on lung function predictions: what is "normal" lung function in diverse human populations?" In: *Current allergy and asthma reports* 15 (2015), pp. 1–11.

[27] John Novembre and Matthew Stephens. "Interpreting principal component analyses of spatial population genetic variation". In: *Nature genetics* 40.5 (2008), p. 646.

[28] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American Journal of Human Genetics* 81.3 (2007), pp. 559–575.

[29] Soheil Baharian et al. "The great migration and African-American genomic diversity". In: *PLoS genetics* 12.5 (2016), e1006059.

[30] Brian K Maples et al. "RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference." In: *Am J Hum Genet* 93.2 (2013), pp. 278–288.

[31] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[32] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. [Online; accessed 2018-02-02]. 2001.

[33] Skipper Seabold and Josef Perktold. "Statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[34] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013.

[35] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing In Science & Engineering* 9.3 (2007), pp. 90–95.

[36] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

# Chapter 3

## 3.0 Preface

In the two years following the publication of a preprint of Chapter 2, UMAP gained widespread adoption in population genetics, being applied across many biobanks and to different types of genetic data, such as structural variants and ancient DNA. It had also been applied to animal data to study introgression, conservation genetics, and disease vectors. There was a growing discussion regarding the interpretations of UMAP results and best practices for genetic data.

In this chapter, we review the applications of UMAP. We discuss the impacts of parametrizations on visualizations, the impacts of data filtering steps for LD and the human leukocyte antigen (HLA) region, and updates to the functionality of the Python implementation. We also discuss the use of UMAP in the context of exploratory data analysis.

This chapter was originally published in the *Journal of Human Genetics* in 2020.

## A review of UMAP in population genetics

Alex Diaz-Papkovich[1,2], Luke Anderson-Trocmé[2], Simon Gravel[2,*]

[1]Quantitative Life Sciences, McGill University, Montreal, Québec, Canada

[2]Department of Human Genetics, McGill University, Montreal, Québec, Canada

* Corresponding author: simon.gravel@mcgill.ca

Published in the *Journal of Human Genetics* 66.1 (2021): 85-91.

## 3.1 Abstract

Uniform manifold approximation and projection (UMAP) has been rapidly adopted by the population genetics community to study population structure. It has become common in visualizing the ancestral composition of human genetic datasets, as well as searching for unique clusters of data, and for identifying geographic patterns. Here we give an overview of applications of UMAP in population genetics, provide recommendations for best practices, and offer insights on optimal uses for the technique.

## 3.2 Introduction

One of the primary challenges of genomic data analysis is high dimensionality. The human genome has over three billion base pairs, and many biobanks contain hundreds of thousands of individuals and above. Relationships among individuals are relevant for historical studies as well as for

studies that seek to identify genetic roots of diseases. These relationships can be influenced by demography, sampling strategies, and technical variation. A first step in many genomic analyses is dimensionality reduction to visualize the data to identify relevant relatedness patterns.

One of the most common methods of dimensionality reduction is principal component analysis (PCA). PCA identifies directions, in the high-dimensional space, along which data is most variable. The projection of genomic data along these directions provides a low-dimensional representation that captures as much variance as possible. Because PCA projection is a linear operation, it has a relatively straightforward interpretation in terms of demographic events (i.e, distances between populations can be interpreted in terms of times to the most recent common ancestors) [1]. It is also well-suited to the correction of population structure in genome-wide association studies (GWAS)[2], and is therefore widely used.

Dimensionality reduction requires tradeoffs. Because PCA projection identifies directions of maximal variance in the data and ignores variation along other directions, it tends to obscure finer scale patterns of population structure. Many nonlinear neighbour graph-based dimension reduction algorithms, such as t-SNE[3], have been developed over the years to overcome this limitation. Here we focus on uniform manifold approximation and projection (UMAP)[4], a method developed in 2018 that has seen widespread use across fields (e.g. single-cell genomics[5]).

Rather than trying to preserve large-scale structure, UMAP seeks to preserve local neighbourhoods in a dataset. For each individual in a genetic dataset, UMAP identifies a pre-set number of nearest neighbours and represents distances to these neighbours as a weighted graph where the

nearest neighbours are weighted more heavily. The goal is then to find a low-dimensional representation of the data that preserves these neighbourhoods as much as possible. By focusing on preserving neighborhood topology rather than absolute distances, UMAP allows for data-dense regions to be "stretched out" in the representation. This can have the benefit of reducing overcrowding of the low-dimensional representation, but comes at the cost of a more challenging interpretation of distances. This is an important distinction relative to algorithms such as PHATE [6] that allow nonlinear transformations of the data while seeking to preserve meaningful distances.

A consequence of the focus on topology is that the meaning of distances in the reduced space is difficult to interpret. Even though most nonlinear dimension reduction methods allow for some stretching of distances to improve visualization of local structure, UMAP can be thought of as particularly permissive, as it does not penalize uniform stretching. Because of this, UMAP representations can also contain arbitrarily small distances between points. Though such small distances might be a faithful representation of the original data topology, they are not ideal for visualization. UMAP allows for specification of a minimum distance between nearest neighbours in low-dimensional space: higher values are useful for visualization, but values near or equal to zero can be used for downstream analyses, such as clustering.

In the context of genetic data, UMAP finds the nearest genetic neighbours for each individual and creates low-dimensional representations that group more closely-related individuals together, and partially preserves longer-range relatedness through intermediary individuals. When used in visualizations, UMAP embeddings uncover many subtle features of data, such as distinct demo-

graphic histories and covariation between genetics, geography, and phenotypes[7]. Figure 3.2 compares visualizations of PCA to UMAP using genotype data from the Thousand Genomes Project (1000GP)[8]. PCA flattens the third dimension, obscuring the distinction between South Asian and Central/South American population clusters, whereas UMAP places them in more clearly visible clusters. UMAP has become widely used to study population structure in humans and other species, in conjunction with existing methods. Here we will describe the current state of the use of UMAP in population genetics.

## 3.3 Visualizing genomic cohorts

The most straightforward and common use of UMAP is for visualization. This has proven useful for data composed of relatively homogeneous populations as well as those with considerable diversity in ancestries. UMAP will dedicate more visual space to larger populations within a cohort, and consequently can illustrate the ancestral composition of a cohort in the context of its population structure as well as the size of the data. Often these data are combined with reference panels such as the 1000GP or the Human Genome Diversity Project (HGDP)[9]. As with PCA, researchers can either perform the dimensionality reduction jointly or project one dataset onto UMAP embeddings of reference data. In most surveyed literature, data are restricted to common variants with a minor allele frequency (MAF) greater than some threshold, e.g. $0.01$. This has the benefit of increasing computational speed and reduces possible confounding by false positive variants. Given sufficient power and high quality data, however, UMAP can be run on unfiltered data.

Data cleaning, including LD thinning, is important when performing UMAP. Certain regions, such as the human leukocyte antigen (HLA) region in the genome, can unduly influence clustering and visualization results — whereas the influence of HLA might be only observed in a higher-order PC, UMAP can identify the clustering of haplotypes at a single, densely typed locus and represent carriers of that haplotype as a distinct cluster (figure 3.1). LD thinning addresses this issue. Thus careful data preparation is necessary for UMAP, and researchers should resist the tendency to assign a demographic explanation to each cluster without careful analysis.

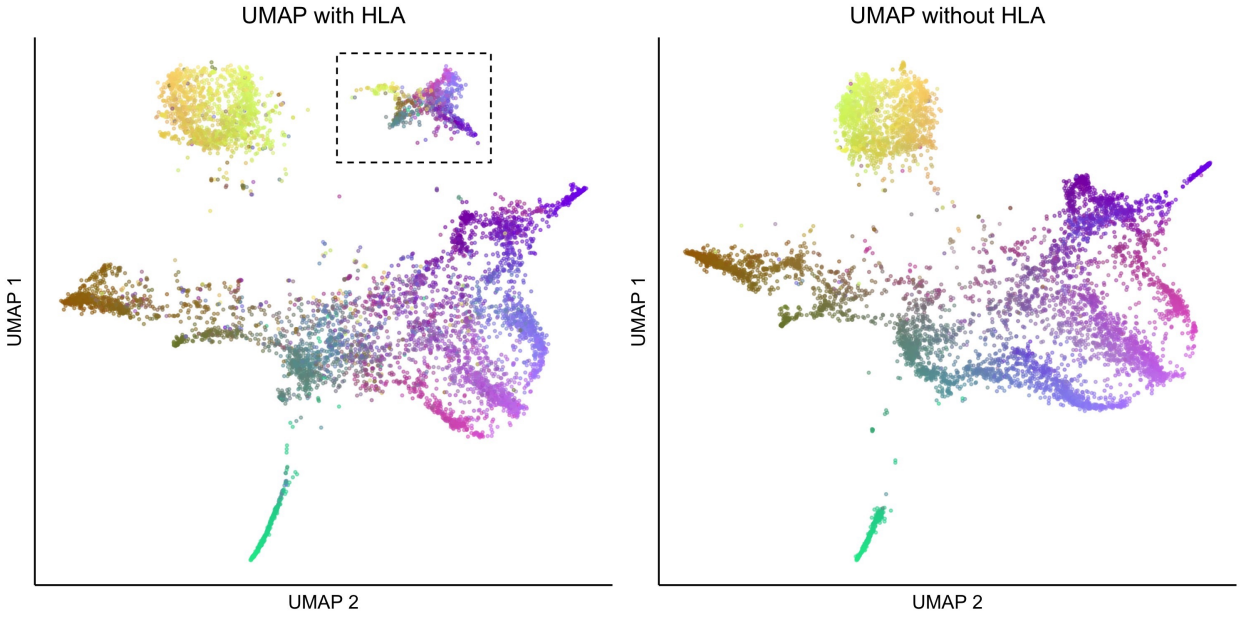
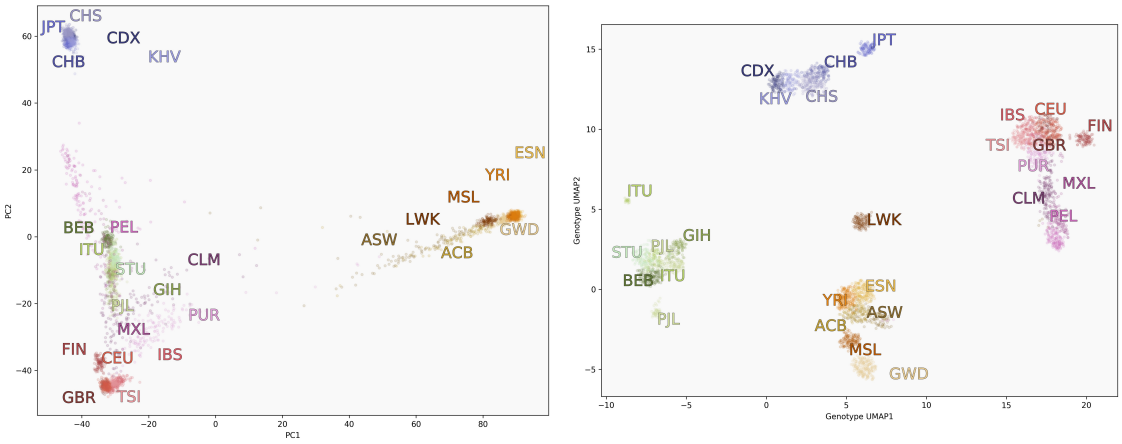



**Figure 3.1: UMAP with (left) and without (right) HLA regions used on the Genizon database.** The cluster in the dotted lines disappears when filtering for HLA and linkage disequilibrium.

Comparing PCA and UMAP on the 1000GP and UKB datasets shows how the sampling scheme influences UMAP representation. PCA for both datasets presents aspects of genetic variation related to the out-of-Africa expansion, forming a triangle shape with African, East Asian,

and European populations at the vertices and admixed populations falling between (Figures 3.2 and 3.3). Since continental ancestry is expected to be the largest source of differences in population structure, PCA will put these populations far apart, and this is useful as a sanity check. In the 1000GP, which sampled individuals from geographically or culturally distinct groups, UMAP forms clusters corresponding to the different groups. By contrast, the UKB performed population-based sampling, and UMAP captures individuals with different levels of admixture from different ancestries. UMAP identifies admixture "bridges" between the different clusters and arguably provides a more detailed representation of the relationships among study participants.




**(a)** **(b)**

**Figure 3.2: Visualizations of data from the 1000GP.** The first two principal components (left) versus a two-dimensional UMAP embedding (right).

Since its strength is in revealing fine-scale population structure, UMAP is well-suited to data with a high number of significant PCs, and can also extract population structure signal from the

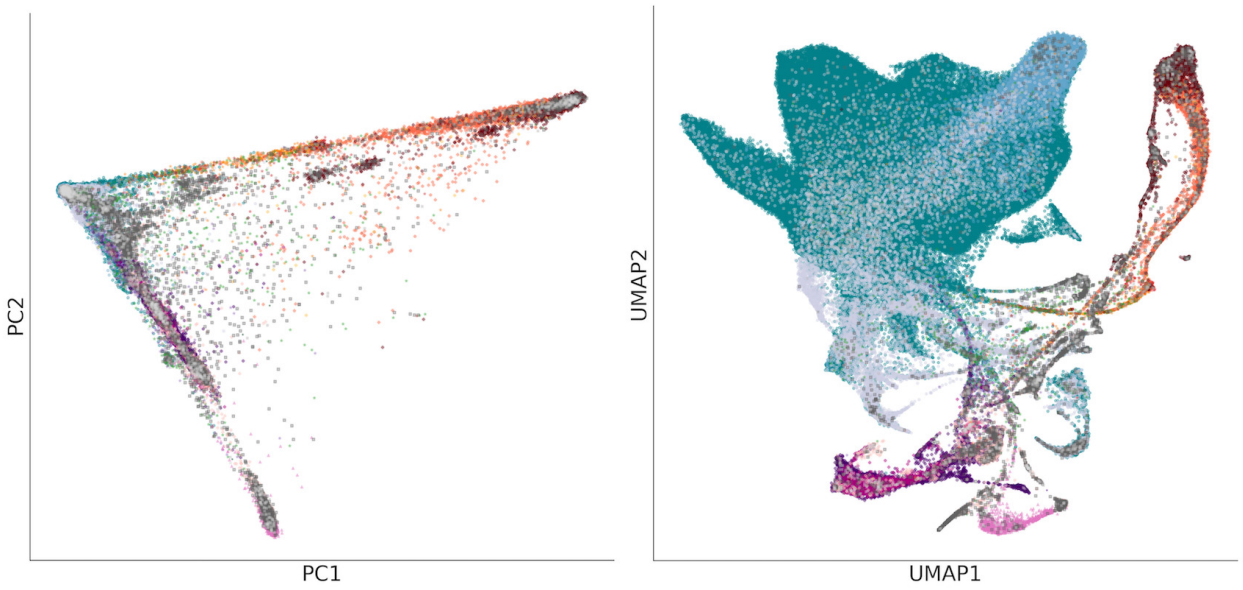
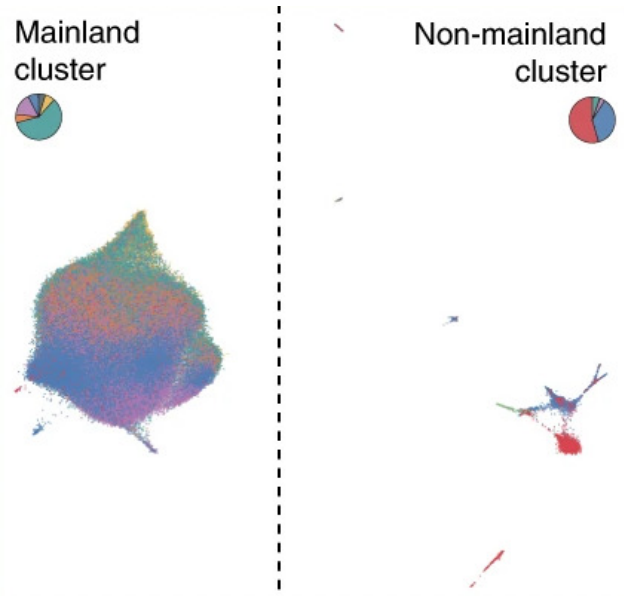



**Figure 3.3: PCA (left) and UMAP (right) projections of the UKB data, coloured by self-identified ethnic background.** Unlike PCA, UMAP focuses on preserving local relationships and emphasizes fine-scale patterns in data. Groups in the UMAP projection are less compressed showing, for example, the relative size of the British and Irish populations in the UKB, alongside populations of other ancestries, while simultaneously showing the population structure between and within groups.

collection of high-order PCs. [7]. Figures 3.4a and 3.4b visualize, respectively, the Genome Aggregation Database (gnomAD v3) from the Broad Institute[10] and Biobank Japan (BBJ)[11, 12], each of which contains over $100,000$ individuals. When applied to ethnically diverse groups such as the UKB, Bio*Me*[13] and the Million Veterans Program (MVP) [14], UMAP tends to highlight groups with different international migration and admixture histories. In relatively more homogeneous populations such as BBJ, it highlights clusters related to geographic features such as island populations.


**(a) gnomADv3 data visualized using UMAP.**




**(b) BBJ data visualized using UMAP.**

**Figure 3.4: The Genome Aggregation Database (gnomAD, top) and Biobank Japan (BBJ, bottom) visualized using UMAP.** UMAP illustrates the ancestral diversity of gnomAD, showing many the relationships between populations on continental and subcontinental levels. For the relatively more homogeneous BBJ data, it splits data geographically into the large mainland cluster (consisting of Hokkaido, Tohoku, Kanto-Koshinetsu, Chubu-Hokuriku, Kinki, and Kyushu regions), and smaller non-mainland clusters. The gnomAD image is reproduced from [10], and the BBJ image is reproduced from [12].

UMAP has also been successfully used with ancient DNA samples combined with modern and contemporary populations to identify shared population structure[15], as well as animal populations to study spatial introgression in mussels[16], genetic bottlenecks in the white rhino population[17], and the geographic origin of disease-carrying mosquitoes[18, 19].

In all these applications, data points were colored using categorical variables such as geographic origin or self-reported ancestry to help with interpretation. We have also found it informative to colour visualizations by continuous variables such as geographical coordinates, phenotype values, or global admixture proportions as in [7], [20], and [21].

## 3.4 Supporting analyses: What do I do with a UMAP projection?

Within Tukey's paradigm of exploratory data analysis, visualization with UMAP can be one of the first steps to the interrogation of complex data[22]. UMAP is useful for identifying clusters in genetic data when the number of clusters is not known in advance[23], and when there are a high number of significant PCs[7]. One straightforward approach is to run UMAP again on a cluster itself to examine subcontinental population structure, as in the National Geographic Genographic Project[20]. One may run UMAP on several types of genetic data; this was the case with Almarri et al's study of structural variants, where they found population stratification in all classes of genetic variants, with Oceanian populations consistently forming their own clusters[24]. In Spear et al., we identified several clusters of Hispanic/Latinx populations using UMAP on the top PCs, despite these groups having overlapping proportions of continental ancestry proportions, and further

studied the Mexican-American population to identify temporal and demographic patterns in their admixture histories[21]. In each case, these projections were combined with traditional statistical approaches such as $F_{ST}$, ADMIXTURE[25], or fineSTRUCTURE[26].

One promising application is the use of clusters as covariates in GWAS and polygenic scores (PGS). Fine-scale population structure continues to confound studies of polygenic traits whether in studies of ancestrally diverse or relatively homogeneous populations (e.g. [27–29]), making it an important area of research. Sakaue et al. used UMAP to identify substructure within the Japanese population, separating it into a mainland population and Hokkaido-Ainu with surrounding islands, reflecting known demographic history in Japan[12]. They identified systematic shifts in PGS for multiple traits across UMAP clusters.

The capacity of UMAP to identify haplotype structure was used by Yamamoto et al to visualize mitochondrial DNA (mtDNA). Though UMAP correctly identified sub-haplogroup clusters of mitchondrial DNA, it did not identify parent clusters as readily as PCA or phylogenetic analysis, and is not particularly advantageous for single-locus analysis[30].

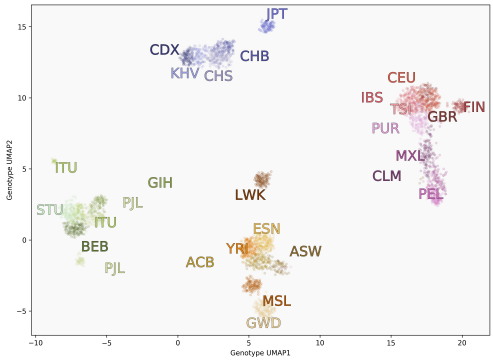
## 3.5 Discussion

UMAP is now used regularly to visualize the ancestral composition of cohorts as well as to examine fine-scale population structure and subtle patterns in biobanks of all compositions. In this sense, UMAP — and dimensionality reduction at large — is to data what a microscope is to biological samples: an effective tool to scientifically examine a subject and provoke deeper investigation. In

both cases, calibration is an important factor, as is understanding the tool's limitations. The main parameters to calibrate in UMAP are the number of nearest neighbours (NN) and the minimum distance (MD). Studies varied in their parameter selection, but generally chose NN close to $15$; setting $NN < 10$ can result in disjoint clusters made up of closely-related individuals, such as families. The minimum distance was usually $0.1 < MD < 0.5$; values of $MD$ close to $0$ create very tight clusters, which can be appropriate for downstream process such as cluster analysis but less pleasing visually. We recommend running multiple parametrizations and to combine UMAP plots with PCA plots and methods like fineSTRUCTURE[26], ADMIXTURE[25], or traditional statistics such as $F_{ST}$ to make inferences. As with PCA and other dimensionality reduction methods, genetically defined clusters represent some degree of shared ancestry. While genetic clusters correlate with variables like self-identified ethnicity or race, they are distinct concepts and not interchangeable[31].

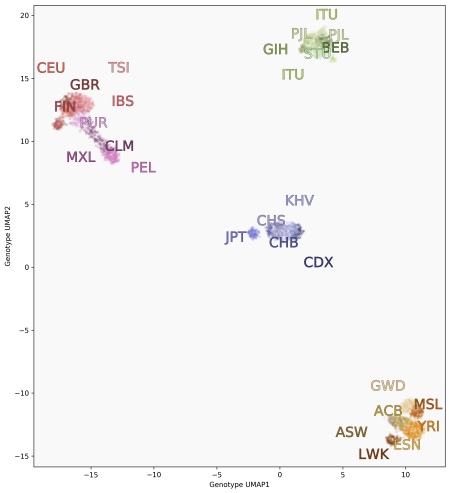
The reference implementation of UMAP is regularly updated with new features[32]. A recent update enabled visualization of the simplicial complex underlying the algorithm, which can highlight how input data and parameterization impact the formation and placement of clusters relative to one another. We demonstrate this using genotype data from the 1000GP in Figure 3.5. Increasing the value of $NN$ increases the size of the complex (at a higher computational cost), but clusters that are completely disjoint from the rest of the data when $NN = 15$ become connected as $NN$ is increased to $200$. In Figures 3.5a and 3.5b the simplicial complexes of South Asian and East Asian populations do not connect to other populations; that is, for these continental clusters, every

individual's $15$ closest genetic neighbours fall within the cluster. In Figures 3.5c and 3.5d, where $NN = 200$, all continental populations become connected. Some populations, such as the Luhya (LWK) and Japanese (JPT), become more closely connected to their continental groups, and the embedding with $NN = 200$ places their subclusters closer to their respective continental populations. These visualizations also clarify that since UMAP preserves these topological connections, the positions of connected clusters may be flipped or rotated relative to each other when carrying out multiple runs with identical parameters.

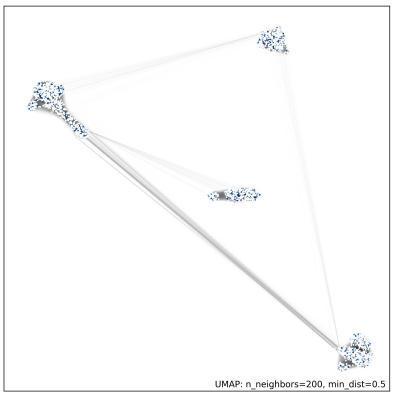



**(a)** UMAP with 15 neighbours.


**(b)** Connectivity map of 15 neighbours.




**(c)** UMAP with 200 neighbours.




**(d)** Connectivity map of 200 neighbours.

**Figure 3.5: UMAP projection of the same genotype data from the 1000GP comparing parametrization with a small (top) and large (bottom) number of nearest neighbours.** Left images are coloured by population; right images are the same points but with the simplicial complex drawn. When adding more neighbours, subclusters become less separated, as with the LWK population, for example. Looking at the connectivity maps, we see new connections between continental groups, such as the Central/South American clusters and East Asian clusters. Darker lines indicate that individuals are closer to each other in genotype space.

## 3.6 Conclusion

With its effective performance and widespread use in under two years, UMAP shows considerable promise as part of the toolbox of a population geneticist, especially in the case of large cohorts. Beyond its capacity to visualize data, it holds promise for downstream methods such as clustering, correction for fine-scale population structure in GWAS and PGS, and identifying unique demographic histories. We anticipate that UMAP and/or related methods of dimensionality reduction will continue to find applications in the field, bolstering our exploration and understanding of human genomic data and the study of complex polygenic traits.

## 3.7 Materials and methods

All code used to process 1000GP data and generate images is available at `https://github.com/diazale/umap_review`. We used genotype data from $3,450$ individuals from the 1000GP using Affy 6.0 genotyping[8]. Genotype data from the 1000GP is available at `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/` and `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/`. The Genizon cohort is comprised of $7,843$ genotyped individuals from Quebec. The genotype data from this cohort was compiled from 4 different chips (HumanHap375, HumanHap550, Illumina1M and Human610-Quad). The missing data from the merging of these datasets was imputed using the Michigan Imputation Server. The UKB provides genotype data and principal components

on $488,377$ individuals. Visualizations were done with matplotlib[33] and PCA was done using sklearn[34].

## References

[1] Gil McVean. "A genealogical interpretation of principal components analysis". In: *PLoS genetics* 5.10 (2009), e1000686.

[2] Nick Patterson, Alkes L Price, and David Reich. "Population structure and eigenanalysis". In: *PLoS genetics* 2.12 (2006), e190.

[3] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605.

[4] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020.

[5] Etienne Becht et al. "Dimensionality reduction for visualizing single-cell data using UMAP". In: *Nature biotechnology* 37.1 (2019), pp. 38–44.

[6] Kevin R Moon et al. "Visualizing structure and transitions in high-dimensional biological data". In: *Nature biotechnology* 37.12 (2019), pp. 1482–1492.

[7] Alex Diaz-Papkovich et al. "UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts". In: *PLoS genetics* 15.11 (2019), e1008432.

[8] The 1000 Genomes Project Consortium. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74.

[9] Howard M Cann et al. "A human genome diversity cell line panel". In: *Science* 296.5566 (2002), pp. 261–262.

[10] Konrad J. Karczewski et al. "The mutational constraint spectrum quantified from variation in 141,456 humans". In: *Nature* 581.7809 (2020), pp. 434–443.

[11] Akiko Nagai et al. "Overview of the BioBank Japan Project: study design and profile". In: *Journal of epidemiology* 27.Supplement_III (2017), S2–S8.

[12] Saori Sakaue et al. "Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction". In: *Nature Communications* 11.1 (2020), p. 1569.

[13] Gillian M. Belbin et al. "Towards a fine-scale population health monitoring system". In: *bioRxiv* (2019), p. 780668.

[14] Haley Hunter-Zinck et al. "Genotyping Array Design and Data Quality Control in the Million Veteran Program". In: *The American Journal of Human Genetics* 106.4 (2020), pp. 535–548.

[15] Ashot Margaryan et al. "Population genomics of the Viking world". In: *bioRxiv* (2019), p. 703405.

[16] Alexis Simon et al. "Local introgression at two spatial scales in mosaic hybrid zones of mussels". In: *bioRxiv* (2019), p. 818559.

[17] Fátima Sánchez-Barreiro et al. "Historical population declines prompted significant genomic erosion in the northern and southern white rhinoceros (Ceratotherium simum)". In: *bioRxiv* (2020), p. 2020.05.10.086686.

[18] The Anopheles gambiae 1000 Genomes Consortium et al. "Genome variation and population structure among 1,142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii". In: *bioRxiv* (2020), p. 864314.

[19] Thomas L. Schmidt et al. "Population genomics of two invasive mosquitoes (Aedes aegypti and Aedes albopictus) from the Indo-Pacific". In: *bioRxiv* (2020), p. 2020.03.15.993055.

[20] Chengzhen L. Dai et al. "Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis". In: *The American Journal of Human Genetics* 106.3 (2020), pp. 371–388.

[21] Melissa L Spear et al. "Recent shifts in the genomic ancestry of Mexican Americans may alter the genetic architecture of biomedical traits". In: *Elife* 9 (2020), e56029.

[22] Susan Holmes and Wolfgang Huber. *Modern Statistics for Modern Biology*. Cambridge University Press, 2019. Chap. Introduction.

[23] Gerry Tonkin-Hill et al. "Fast hierarchical Bayesian analysis of population structure". In: *Nucleic Acids Research* 47.11 (2019), pp. 5539–5549.

[24] Mohamed A. Almarri et al. "Population Structure, Stratification and Introgression of Human Structural Variation". In: *bioRxiv* (2020), p. 746172.

[25] David H Alexander, John Novembre, and Kenneth Lange. "Fast model-based estimation of ancestry in unrelated individuals". In: *Genome research* 19.9 (2009), pp. 1655–1664.

[26] Daniel John Lawson et al. "Inference of population structure using dense haplotype data". In: *PLoS genetics* 8.1 (2012), e1002453.

[27] Sini Kerminen et al. "Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland". In: *The American Journal of Human Genetics* 104.6 (2019), pp. 1169–1181.

[28] Jeremy J Berg et al. "Reduced signal for polygenic adaptation of height in UK Biobank". In: *eLife* 8 (2019), e39725.

[29] Mashaal Sohail et al. "Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies". In: *eLife* 8 (2019), e39702.

[30] Kenichi Yamamoto et al. "Genetic and phenotypic landscape of the mitochondrial genome in the Japanese population". In: *Communications Biology* 3.1 (2020), pp. 1–11.

[31] Iain Mathieson and Aylwyn Scally. "What is ancestry?" In: *PLOS Genetics* 16.3 (2020), e1008624.

[32] Leland McInnes et al. “UMAP: Uniform Manifold Approximation and Projection”. In: *The Journal of Open Source Software* 3.29 (2018), p. 861.

[33] J. D. Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing In Science & Engineering* 9.3 (2007), pp. 90–95.

[34] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

# Chapter 4

## 4.0 Preface

One common use of UMAP in population genetics is to identify clusters and to treat them as populations for downstream analysis. However, there is no effective way to algorithmically extract clusters from UMAP plots. Though it generates clusters visually, it is a dimensionality reduction algorithm and not a clustering algorithm. Centroid- or archetype-based approaches fail to capture many individuals, rely on arbitrary definitions of population groups, or require a pre-specified number of populations. Researchers often resorted to hand-delineating clusters, which is limited to 2D projections and not scalable in the presence of many populations.

We apply HDBSCAN($\hat{\epsilon}$), a hierarchical density-based clustering algorithm to UMAP data. This approach can use UMAP embeddings of arbitrary dimensions—importantly allowing us to work in 3 or more dimensions. Running on the order of seconds for massive biobanks, it creates topological clusters that reflect the demgraphic histories of populations. We apply the algorithm to three biobanks (the 1KGP, UKB, and CaG cohorts) and demonstrate its effectiveness at capturing population structure, usefulness in analysis of biobank data, potential downstream applications (e.g. for PGS transferability), and its use as a quality control tool.

This manuscript was released as a preprint on *bioRxiv* in 2023.

## Topological stratification of continuous genetic variation in large biobanks

Alex Diaz-Papkovich[1,2], Shadi Zabad[3], Chief Ben-Eghan[2], Luke Anderson-Trocmé[2], Georgette Femerling[2], Vikram Nathan[3], Jenisha Patel[3,4], Simon Gravel[2,*]

[1]Quantitative Life Sciences, McGill University, Montreal, Québec, Canada

[2]Department of Human Genetics, McGill University, Montreal, Québec, Canada

[3]School of Computer Science, McGill University, Montreal, Québec, Canada

[4]Department of Bioengineering, McGill University, Montreal, Québec, Canada

* Corresponding author: simon.gravel@mcgill.ca

Released as a preprint on *bioRxiv* in 2023.

## 4.1 Abstract

Biobanks now contain genetic data from millions of individuals. Dimensionality reduction, visualization and stratification are standard when exploring data at these scales; while efficient and tractable methods exist for the first two, stratification remains challenging because of uncertainty about sources of population structure. In practice, stratification is commonly performed by drawing shapes around dimensionally reduced data or assuming populations have a "type" genome. We propose a method of stratifying data with topological analysis that is fast, easy to implement, and integrates with existing pipelines. The approach is robust to the presence of sub-populations of varying sizes and wide ranges of population structure patterns. We demonstrate its effectiveness

on genotypes from three biobanks and illustrate how topological genetic strata can help us understand structure within biobanks, evaluate distributions of genotypic and phenotypic data, examine polygenic score transferability, identify potential influential alleles, and perform quality control.

## 4.2 Introduction

Following improvements in genomic technologies, large-scale biobanks have become commonplace. The Global Biobank Meta-analysis Initiative (GBMI), for example, lists 23 biobanks with genetic data and health records from over 2.2 million individuals[1]. The growth in sample sizes has led to increased potential for scientific findings; methods like genome-wide association studies (GWAS) and polygenic scores (PGS) have gained widespread popularity, and accordingly thousands of genetic loci have been implicated across numerous phenotypes. Though the growth of biobanks has fuelled discovery, population structure—the phenomenon in which allele frequencies systematically differ between populations—remains a persistent confounder in GWAS and PGS (e.g. [2, 3]). Many methods in population genetics seek to describe and account for population structure, but the complexity of human history, along with factors like biobank recruitment strategies, preclude model-based approaches from effectively capturing the many determinants of observed genetic variation.

Dimensionality reduction and visualization are common in examining both discrete and continuous aspects of genetic variation (e.g. [4, 5]). Within the framework of exploratory-confirmatory data analysis, visualization of complex data enables pattern-recognition and the generation and

testing of hypotheses[6]. Visualization alone, though, cannot be used for analysis, and data stratification is often necessary. Algorithmic genetic stratification or clustering is often based on principal component analysis (PCA), sometimes using reference panels or assuming a "type" genome—e.g., all points within a certain radius in PCA space are classified as "European". In recently admixed populations (i.e., populations who derive ancestry from "source" populations who had been in relative isolation), grouping based on inferred admixture proportions is also common, often with the use of a reference panel as a proxy for the source populations. These approaches may not work for populations with no reference panel, or with complex admixture histories, or small sample sizes[7]. Other approaches cluster based on shared identity-by-descent (IBD) segments or recent genetic relatedness (e.g. [8, 9]). These approaches typically capture finer scale population structure, but are analytically and computationally demanding. Self-declared variables like race and ethnicity are also sometimes used for genetic stratification but are imperfect indicators of genetic ancestry and are no longer recommend as proxies for it[10, 11].

Despite the demand, there is not an effective, fast, and tractable method for stratifying biobank data based on patterns of genetic structure. In practice, researchers often manually group participants into discrete ad hoc "clusters" that they perceive in low-dimensional visualizations, which they use as strata in downstream analyses regarding, e.g., heterogeneity in ancestry and allele frequencies[12], environmental exposures[4], or assessing the performance of PGS[2, 13]. There are many drawbacks to such ad hoc approaches. For example, in cosmopolitan cohorts, there are many subgroups with distinct ancestral histories, leading researchers to manually distinguish between a

"majority" cluster and an "everybody else" cluster—often to be discarded due to its heterogeneity[14, 15].

We propose topological data analysis as an alternative approach. Rather than fitting individuals to a pre-defined notion of a population, a topological approach describes the network of neighbourhoods between data points—here, this would be the network of genetic similarity between individuals. It is especially well-suited to describe collections of points in high-dimensional space with smooth distributions but with no clear centre or "archetype". We assume that structure in high-dimensional genetic data can be represented topologically, and can be locally approximated and reconstructed in a low-dimensional space. After reconstructing data in the low-dimensional space, we identify dense clusters of data—i.e., the genetic strata. This approach is unsupervised, requiring neither a number of clusters nor a reference panel, and thus fits naturally with population genetic data, which is sparse and contains numerous sub-populations of unknown and varying sizes, often without *a priori* definitions.

We demonstrate the effectiveness of this approach on three biobanks, showing that we can consistently and effectively identify and characterize sources of population structure in each cohort, as well as relate many key variables to this structure. We identify subtle population structure quickly, simultaneously identifying structured groups as small as $100$ individuals and as large as $400,000$ within the same cohort. We show that this provides important insight into the relationships between genetic structure, environmental and sociodemographic variables, and phenotype distributions. We use stratification to identify populations for which PCA adjustment fails within a biobank (often

admixed populations) and populations for which PGS transferability is poor (often, but not always, populations diverged from the training population). Finally, we illustrate how to use topological modelling as a quality control tool, a critical if less glamorous aspect of the fast-growing biobank space.

In summary, we argue that topological modelling, which describes data in terms of local neighbourhoods in a high-dimensional space, is a powerful alternative to ancestry-based modelling for the description of genetic variation in complex cohorts.

## 4.3 Methods

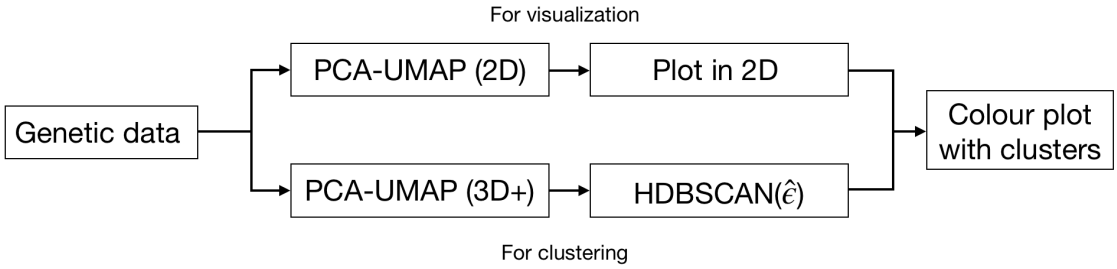



**Figure 4.1: Overview of visualization and clustering pipeline.** We use the same genotype data to generate visualizations as well as cluster labels and combine them within one figure. For visualization, we reduce data to 2D and optimize for visual clarity. For clustering, we use higher UMAP dimensions to maximize information and minimize distortions before HDBSCAN($\hat{\epsilon}$) processing. These clusters are then used to colour the 2D plot, where each point is an individual and the cluster is represented by colour. Genotype data is pre-processed with PCA.

Our method works on structured genotype data, represented by a matrix of allele counts for each individual and genetic variant. To reduce computational burden, we perform analyses using leading principal component (PCs) on the raw genotype data. This approach has the additional ben-

efit that many genetic cohorts have PC coordinates pre-computed as part of standard quality control pipelines. We use uniform manifold approximation and projection (UMAP)[16], a dimensionality reduction method, and a clustering algorithm, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), using an implementation by Malzer and Baum called HDBSCAN($\hat{\epsilon}$)[17]. Both methods are unsupervised.

UMAP is designed to preserve the topology of high-dimensional data by assuming the data lie on a manifold and then approximating the manifold on a local level[16]. The algorithm requires three parameter inputs: the target number of dimensions, the number of nearest neighbours (used to define the size of high-dimensional neighbourhoods to approximate), and the minimum distance between points in the low-dimensional space. We have previously explored its use for visualization in $2$ and $3$ dimensions[4]. Figure 4.1 highlights the two distinct roles played by UMAP in this work, each requiring distinct parameters:

1. For visualization, reducing data to $2$ dimensions and using a relatively high minimum distance ($0.3$ to $0.5$), to facilitate human perception and understanding

2. For clustering, reducing data to $3$ or more dimensions and using a very low minimum distance (near or equal to $0$) to facilitate algorithmic identification of dense clumps of data.

After reducing genetic data to $3$ or more dimensions with UMAP in step 2, we use HDBSCAN($\hat{\epsilon}$) to extract clusters. HDBSCAN($\hat{\epsilon}$) is a hierarchical density-based clustering algorithm based on predecessors HDBSCAN and DBSCAN*[17]. It is motivated by situations where we expect data to

be in a sparsely populated space with relatively dense clusters throughout. The number of clusters is not known, and the sizes of the clusters are assumed to vary. This describes biobank data particularly well, since it is expected to contain population structure at many different scales, and it is usually difficult to specify in advance a useful number of subgroups to consider. The parameter $\hat{\epsilon}$ allows clusters to have widely varying sizes; we provide more details on parameters in the Supporting Information (SI).

We use UMAP-assisted density-based clustering on data from three biobanks: the Thousand Genomes Project (1KGP), the UK biobank (UKB), and CARTaGENE (CaG). The 1KGP data consists of the genotypes of $3,450$ individuals sampled from $26$ populations from around the world; the populations were decided in advance and their sample sizes are similar, ranging from $104$ to $183$ samples[18]. The UKB is a cohort of $488,377$ individuals from the United Kingdom (UK) with genotypic, phenotypic, and sociodemographic data. UKB participants were recruited by inviting 9 million individuals registered with the National Health Service (NHS) who lived near a testing centre[19]. CaG is a cohort of residents of the Canadian province of Québec, with genotype data for $29,337$ participants who were recruited using registration data from the Régie de l'assurance maladie du Québec (RAMQ), the provincial health authority, from four metropolitian areas in the province[20]. Unlike the 1KGP, CaG and the UKB do not have *a priori* populations defined, though they collected information about ethnicity, country of birth, and residential geographic distribution.

## 4.4 Results

### 4.4.1 Clustering captures population structure from sample design

The 1KGP's relatively balanced global sample design makes it useful for testing algorithms to identify population structure. We have previously shown that UMAP results in clear visual clusters from 1KGP data in two dimensions[4]. Figure 4.2 shows a UMAP representation of the 1KGP. Figure 4.2a shows the data without population labels (to mimic data with unknown populations), Figure 4.2b shows the data with corresponding population labels from the 1KGP, and Figure 4.2c shows the data with cluster labels generated by HDBSCAN($\hat{\epsilon}$) run on a $5D$ UMAP.

The major source of genetic structure in 1KGP data is its sampling scheme, which selected individuals from geographically diverse populations. The clusters formed by UMAP and extracted by HDBSCAN($\hat{\epsilon}$) largely reflect this sampling strategy, with some exceptions noted below. Figure 4.2d shows that there is strong agreement between population label and cluster label, with full breakdowns provided in Tables 4s4 and 4s5. These results are comparable to a supervised neural network approach to predict sampled population label (e.g. Figure 3 in [21]), though our approach is unsupervised and runs on the order of seconds.

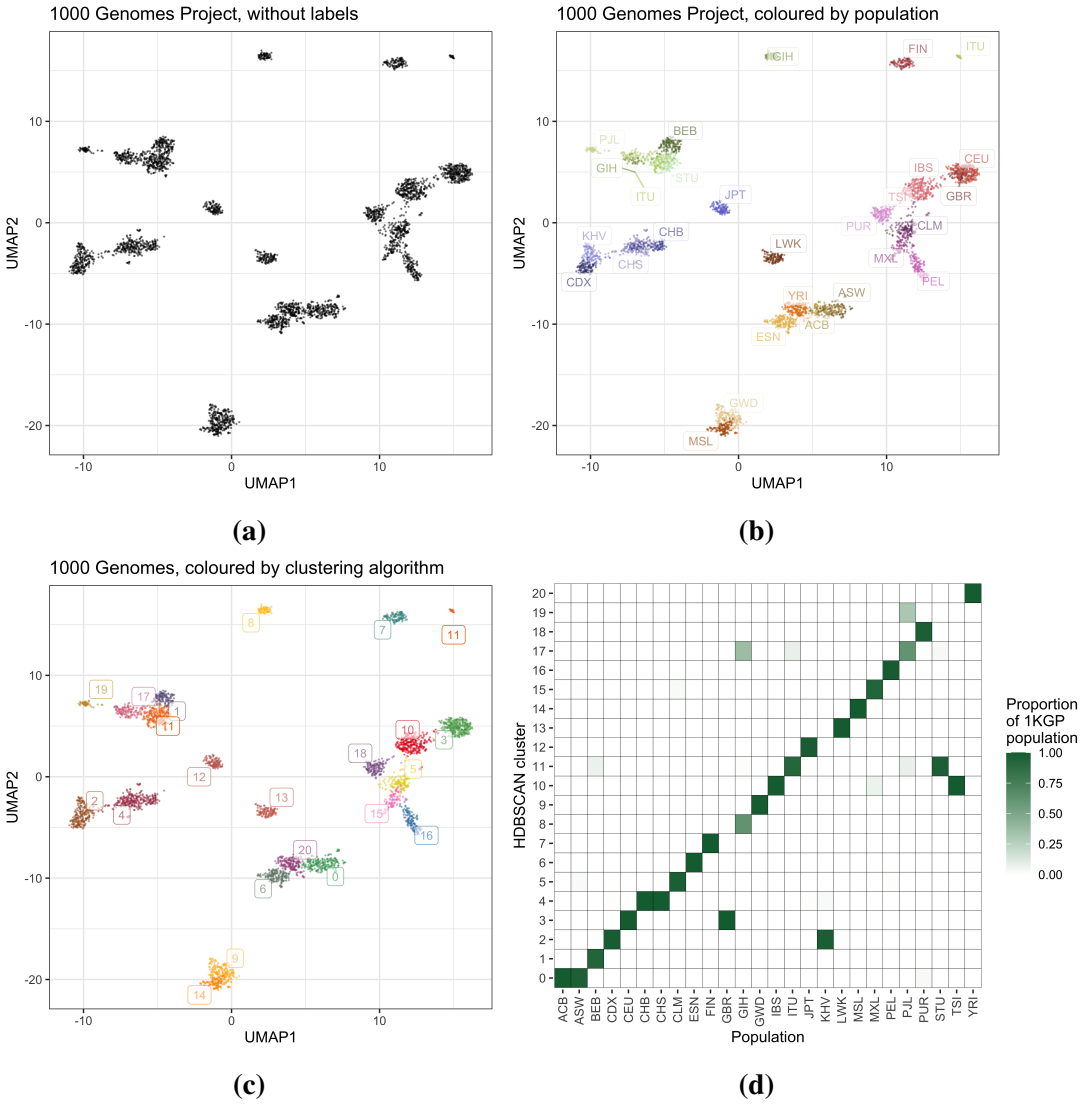



**Figure 4.2: Clusters generated from 1KGP genotype data reflect its population sampling.** (A) UMAP embedding of data without labels. (B) UMAP embedding of data, coloured by population label. (C) UMAP embedding of data, coloured by clusters derived from HDBSCAN($\hat{\epsilon}$) applied to a 5D UMAP embedding. (D) Proportions of each 1KGP population contained within a given cluster. Most populations fall almost entirely within a single cluster, with a few splitting into multiple clusters. Population labels are provided in Table 4s3.

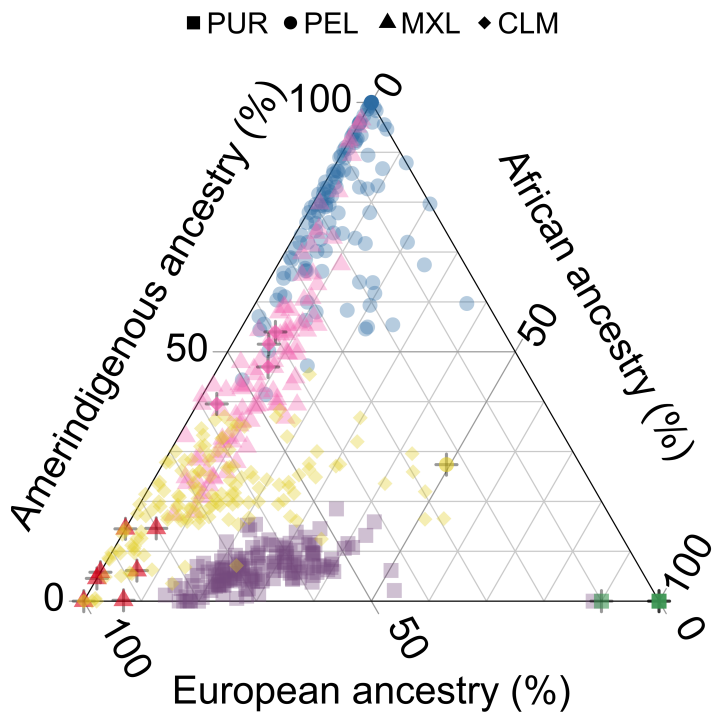



**Figure 4.3: Clusters capture structure in populations with overlapping admixture proportions in the 1KGP.** A ternary plot of the PUR, PEL, MXL, and CLM populations from the 1KGP with axes corresponding to global ancestry proportions estimated using ADMIXTURE ($K = 3$). Shapes indicate 1KGP label, colours indicate cluster label and match Figure 4.2c; bolded points with a + symbol overlaid indicate individuals who are not members of the modal cluster of their 1KGP population (full results given in Tables 4s4 and 4s5). Many individuals from the populations have similar admixture proportions; UMAP-HDBSCAN($\hat{\epsilon}$) clusters reflect structure from the sample populations, while clusters based on inferred admixture proportions would not.

One benefit of the unsupervised approach is that we do not require *a priori* assumptions about the origins of structure, making it possible to capture meaningful clusters despite considerable within-cluster heterogeneity, including in admixed populations. The Central and South American clusters largely match their 1KGP labels despite overlapping distributions in ADMIXTURE-estimated continental ancestry within each group (Figure 4.3). Some populations are clustered together: GBR and CEU (British From England and Scotland; and Utah residents with Northern/Western European ancestry), CDX and KHV (Chinese Dai in Xishuangbanna, China; and Kinh in Ho Chi Minh City, Vietnam), IBS and TSI (Iberian Populations in Spain; and Toscani in Italy), ACB and ASW (African Caribbean in Barbados; and African Ancestry in SW USA). While these groups differ in their sampling and history, supervised learning methods also struggle in distinguishing most of these pairs (Figure 3A in [21]). The CDX and KHV (Cluster 2 in Figure 4.2b) populations are present at opposite ends of one continuous cloud of points. In other words, two groups belonging to one cluster does not mean that the groups are indistinguishable. Rather, it means that HDBSCAN($\hat{\epsilon}$) could find a relatively continuous path in genetic space linking individuals sampled in one group to individuals sampled in the other.

Some South Asian populations are split into different clusters, possibly from stronger patterns of relatedness within those groups[4, 22]. We note the ITU (Indian Telugu in the UK) population is visibly split into two groups in 2D, while clustering carried out in 5D groups them together (Cluster 11). While some clusters will tend to persist across many parametrizations of UMAP and HDBSCAN($\hat{\epsilon}$), others based on more subtle patterns or in populations with more continuous vari-

ation will be less stable—though discrete groupings can help us understand data, the delineations are always, to a degree, arbitrary.

Some South Asian populations are split into different clusters, possibly from stronger patterns of relatedness within those groups[4, 22]. We note the ITU (Indian Telugu in the UK) population is visibly split into two groups in 2D, while clustering carried out in 5D groups them together (Cluster 11). While some clusters will tend to persist across many parametrizations of UMAP and HDBSCAN($\hat{\epsilon}$), others based on more subtle patterns or in populations with more continuous variation will be less stable—though discrete groupings can help us understand data, the delineations are always, to a degree, arbitrary.

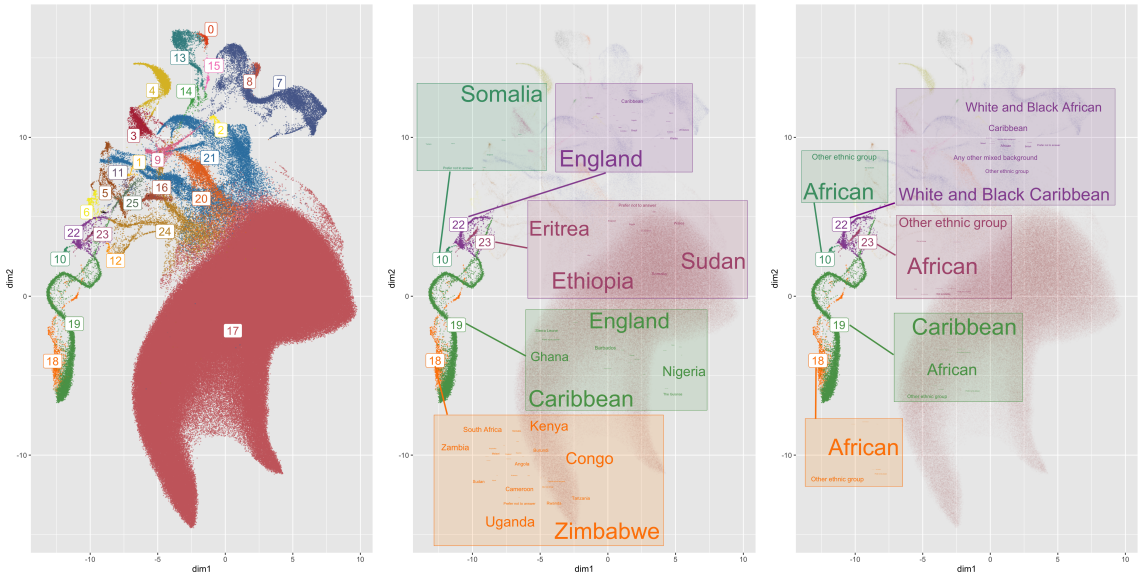
### 4.4.2 Correlates between populations and sociodemographic, phenotypic, and environmental variables

The UK biobank (UKB) contains $488,377$ genotypes from volunteers with an array of demographic, phenotypic, and biomedical data, with individuals' ages ranging from $40$ to $69$. The demographic data collected for the UKB include Country of Birth (COB) and Ethnic Background (EB), which is selected from a nested set of pre-determined options (see Table 4s6). Participants first select their "ethnic group" from a list (e.g. "White"; "Black or Black British"), which determines the list of possible "ethnic background" (e.g. "British"; "Caribbean"). The most common countries of birth in the data set are England, Scotland, Wales, and the Republic of Ireland, comprising $77.8\%$, $8.0\%$, $4.4\%$, and $1.0\%$, respectively. For EB, $88.3\%$ of participants selected "White

British", with an additional $5.8\%$ selecting "White Irish" or "Any other white background". Here we primarily focus on the $28,814$ individuals with other backgrounds.

The biobank is one of the most-used resources for genetic analyses. Despite its multi-ethnic composition, many studies discard non-European samples, sometimes citing concerns related to confounding from population structure[14]. The population structure has been deeply explored, though typically focused on British or European individuals[23–25]. Because its sub-populations are numerous, geographically/ancestrally diverse, and of widely varying sizes, clustering the UKB data is challenging, requiring overly broad categorization (e.g. a small number of continental populations [12, 26]) and/or significant computational resources. The original implementation of HDBSCAN, without the $\hat{\epsilon}$ parameter, discards much of the UKB data as noise and splits populations into hundreds of microclusters that are not interpretable (see Fig 4s1).

Figure 4.4a shows $26$ clusters generated by HDBSCAN($\hat{\epsilon}$), placing $99.99\%$ of individuals in clusters. We generated word clouds for COB and EB, shown in Figures 4.4b and 4.4c, which allow us to illustrate sources of structure without having to impose a label to groups which may be heterogeneous. Individuals in Cluster 10, for example, are mostly born in Somalia ($84\%$), while those in Cluster 23 are mostly born in East Africa (Ethiopia, Sudan, Eritrea; $33\%$, $29\%$, $25\%$, respectively). Those in Cluster 18 are mostly born in sub-Saharan Africa, and $77\%$ chose "African" as their EB, while $19\%$ chose "Other ethnic group". Figure 4s2 presents word clouds for another subset of data. Individuals in Cluster 0 are mostly born in Japan and South Korea ($84\%$ and $9\%$, respectively), and those in Cluster 15 are mostly born in Nepal ($80\%$). In contrast, individuals




(a) (b) (c)

**Figure 4.4: An example of clusters of population structure in the UKB.** The clusters reflect a mixture of demographic history within the UK, the geographic origins of recent immigrants, the colonial history of the British Empire, and ongoing admixture. (a) Left: A 2D UMAP of UKB genotypes coloured by HDBSCAN($\hat{\epsilon}$). This parametrization generated 26 clusters. (b) Middle: Five clusters are highlighted with word clouds for the most common countries of birth within the cluster. (c) Right: The same five clusters are highlighted with word clouds for the most common EB within the cluster. Admixture proportions for clusters are presented in Figure 4s3. Detailed breakdowns of EB and country of birth are presented in Tables 4s7 and 4s8.

in Cluster 13 are born in a variety of East/Southeast Asian jurisdictions; the most common EB was "Chinese" (70%), followed by "Other ethnic group" (16%) and "Any other Asian background" (11%). Tables 4s7 and 4s8 provide breakdowns for clusters.

Clusters 14 and 22 both capture structure resulting from recent admixture following immigration and colonial history, with 49% and 66% of their respective populations being born in England (see also Figure 4s3). No single EB represents a majority in either cluster; the most common EB in Cluster 14 is "Any other mixed background" (29%), while for Cluster 22 it is "Mixed, White

and Black Caribbean" ($39\%$).

Notably, significant proportions of majority-African-born clusters identify as "Other ethnic group"—a respective $24\%$, $19\%$, and $37\%$ in Clusters 10, 18, and 23. This suggests that filtering individuals by EB alone for further analyses could limit sample sizes by discarding relevant data. Cluster 18 captures individuals born in sub-Saharan Africa, while Cluster 19 consists of individuals born in the Caribbean ($31\%$), England ($28\%$), as well as Nigeria ($14\%$) and Ghana ($12\%$). These regions are historically linked to the UK; between the years 1641 and 1808, an estimated $325,311$ Africans from the Bight of Benin, between the coasts of modern-day Ghana and Nigeria, were enslaved by British ships and sent to the British Caribbean[27, 28].

Despite the complexity of the UKB, topological clustering identifies population structure that is interpretable from historical or demographic perspectives and includes all or almost all individuals. Such structure is difficult to infer from a single label such as geography or ethnicity; once it is characterized, it can clarify the genetic structure of the cohort.

### 4.4.3 Phenotype smoothing and modelling

Epidemiological research often focuses on observed differences between groups—for example, finding the mean of a phenotype or sociodemographic measure and comparing between populations. Clustering is one method to define groups based on shared demographic history. However, clustering data featuring continuous variation patterns can be sensitive to input parameters, may not reflect true boundaries, and risks encouraging the perception that clusters are more distinct than

they really are[29]. As an illustration of how topological approaches can help interpret data beyond specific clustering choices, we use HDBSCAN($\hat{\epsilon}$) to define a simple regularization method, defined in Algorithm 1, that allows us to incorporate alternative parameterizations. We use this smoothing method to examine phenotypic and sociodemographic data with respect to population structure and identify outstanding patterns.

---

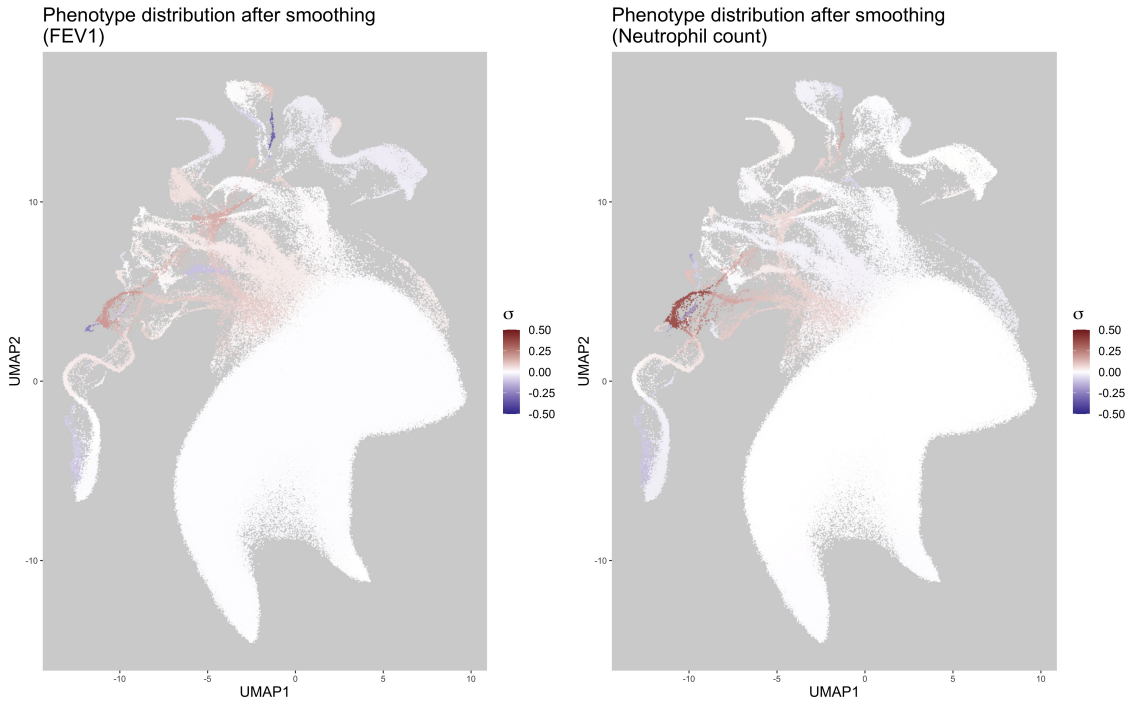
**Algorithm 1** We create a regularized value for each measure by taking the mean of cluster means for each individual. Given a set of parameters $P$ for the clustering algorithm, each parametrization $p$ will result in a set of clusters $C_p$. We use varying cluster assignments across parametrizations to smooth a measured quantity (e.g. phenotype) $m$ for individual $i$.

---

Given a set of parametrizations $P$, each with a set of clusters $C_p$, for some measure of interest $m$, we calculate the regularized value $\mu_i$ for each individual $i$.
**for** $p$ in $P$ **do**
  **for** $c$ in $C_p$ **do**
    For each individual $i$ in $C_p$, set the mean value $\mu_{p,i} := \sum_i m_i / \mid C_p \mid$
  **end for**
**end for**
Set $\mu_i := \frac{\sum_{p \in P} \mu_{p,i}}{|P|}$

---

In Figure 4.5, we visualize the smoothing across $604$ parametrizations of UMAP-HDBSCAN($\hat{\epsilon}$) FEV1 and neutrophil count. Despite regressing out the effects of the top $40$ principal components, there remains structure in the distribution of the residuals. For example, the average residual value is noticeably higher in individuals who fall in Cluster 22 in Figure 4.4a. This cluster is composed mostly of individuals with admixed African/European backgrounds, and although they are intermediate to African and European ancestry populations in PCA space (Figure 4s7), their phenotype distributions are not intermediate to clusters of primarily European- and African-ancestry individuals (Figure 4s5, Figure 4s6).




**(a)** **(b)**

**Figure 4.5: Smoothed phenotypic measures across multiple parametrizations of clustering.** A 2D UMAP coloured by phenotype value after having removed the top $40$ PCs and averaged by cluster, run over $604$ parametrizations of the clustering pipeline. The colour scale runs from $-0.5\sigma$ to $0.5\sigma$, for the standard deviation $\sigma$ of each phenotype after regressing the linear effects of the top $40$ PCs. We observe that the distributions of phenotypes among some groups are not centred about $0$ even after PC adjustment. (a) Left: FEV1. (b) Right: Neutrophil count.

To test if smoothed cluster estimates have explanatory power for these admixed individuals, we carried out an $80-20$ split and compared simple linear models for phenotype prediction using the top $40$ PCs versus using the smoothed estimates made from residuals after removing the effects of the top $40$ PCs. We compared the models for populations that selected "Mixed" as their EB in the UKB questionnaire and found that for individuals who selected "White and Black Caribbean" ($n = 573$) or "White and Black African" ($n = 389$), the smoothed cluster estimates indeed outperformed

the PCA model, with an improved mean squared error across several phenotypes (see Figure 4.6; full table of MSE values in Tables 4s9 and 4s10).

Analysis based on topological components can help to visualize the impact of covariate adjustment in the context of population structure and to identify residual heterogeneity in phenotype distributions and environmental data (e.g. smoking rates in Figure 4s4).

### 4.4.4 Evaluating transferability of polygenic scores

Most investigations of PGS transferability are done at a population-level using large-scale geographical groups (e.g. "African", "European", "Asian"). However, these broad populations themselves exhibit population structure[30]. Instead, we use our $26$ cluster labels from Figure 4.4a, and compared the transferability of PGS across them.

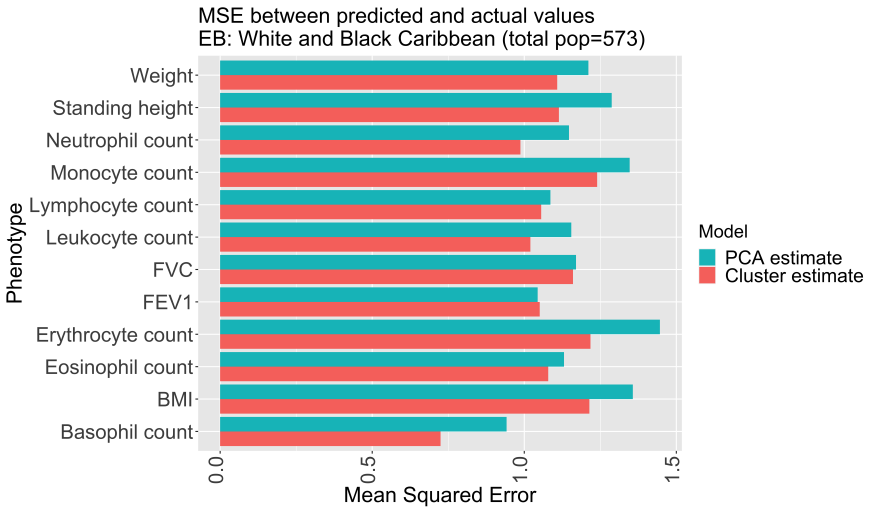
Using UKB data, we estimated effect sizes of SNPs using VIPRS[31]. As a training population, we used individuals who selected "White British" as their EB to mimic the well-documented overrepresentation of European-ancestry individuals in GWAS. We estimated phenotypes for individuals and calculated the values of the fixation index ($F_{ST}$) between the clusters. In Figure 4.7, we plot the PGS accuracy for two phenotypes—standing height and low-density lipoprotein cholesterol (LDL)—against the $F_{ST}$ for each cluster relative to Cluster 17, a cluster with over $400,000$ individuals and with significant overlap with the training population ($> 95\%$ selected "White British" as their EB). We observe for height (Figure 4.7a) that as the $F_{ST}$ between populations grows, the predictive value of the PGS decreases; such a decrease is expected, due to factors like population-

specific causal variants, gene-by-environment interaction, differences in allele frequencies, and linkage disequilibrium between assayed SNPs and causal variants[32].

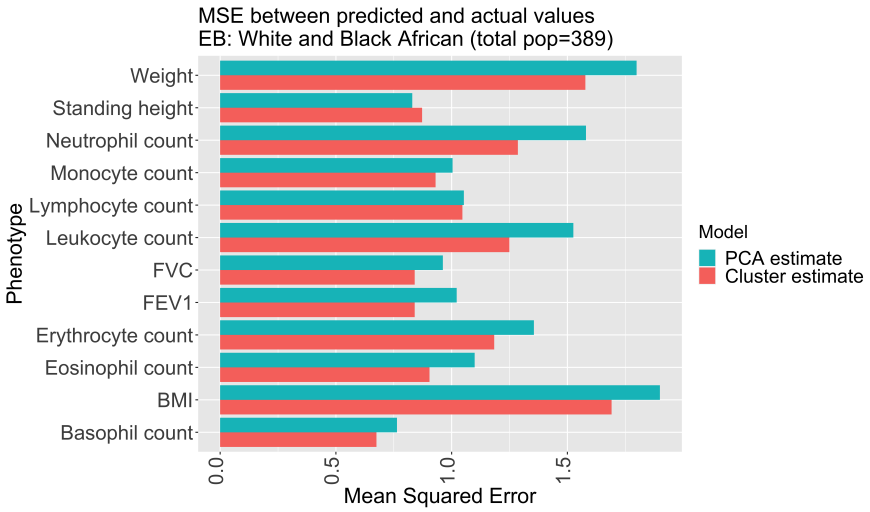
However, we see no such relationship for LDL (Figure 4.7b). Cluster 18, composed mostly of individuals born in sub-Saharan Africa and of whom $77\%$ selected the EB "Black African", has one of the best PGS predictions despite its large $F_{ST}$ from the training population. This may be because there are a few variants with large effect sizes; in contrast to height, LDL has been noted for its relatively low polygenicity[7]. Since $F_{ST}$ compares genome-wide variation, the accuracy of a PGS constructed from relatively few variants with strong effects is not expected to correlate as strongly with $F_{ST}$.

To test if the frequencies of certain alleles impacted the PGS estimates, we modelled the $R^2$ from the VIPRS estimates for each cluster against minor allele frequencies (MAF) of the top $100$ SNPs and found the two strongest results were for $rs4420638$ and $rs7412$ (Tables 4s1 and 4s2; Figures 4s8 and 4s9, respectively). Both have their highest frequencies in Cluster 18 and both markers are in the apolipoprotein E (APOE) gene cluster; $rs7412$ had the largest overall effect size ($\hat{\beta} = -0.1812$), while $rs4420638$ had the second largest effect size in the opposite direction ($\hat{\beta} = 0.02813$). The $rs7412$ allele has been linked to LDL[33] and was found to explain significant variation in LDL in African Americans[34]. The $rs4420638$ allele was associated with LDL even in the presence of the $rs7412$ allele in a study of Sardinian, Norwegian, and Finnish individuals[35]; it was also found to affect LDL in studies with of children in Germany[36] and China[37].

The relationship between PGS accuracy and fine-scale population structure is complex and will vary by phenotype. It is not immediately obvious whether a PGS will transfer when there is a large degree of differentiation between the estimand and training populations. However, an approach like UMAP-HDBSCAN($\hat{\epsilon}$) can provide a detailed picture of the likely performance of a PGS in various genetic subgroups.

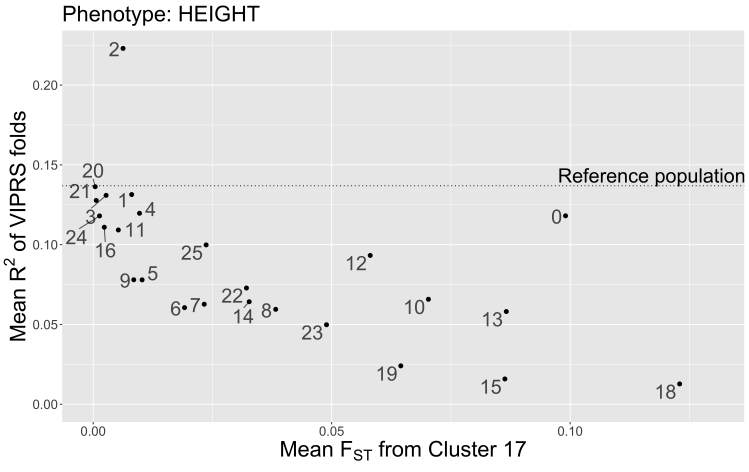



**(a)**




**(b)**

**Figure 4.6: Cluster-based estimation can improve phenotype models.** To test the explanatory value of smoothed cluster estimates generated from Algorithm 1, we carried out an $80-20$ split on the UKB data and compared phenotype prediction using the top $40$ principal components versus estimates generated from the residual structure, presented in Figure 4.5.




**(a)**


**(b)**

**Figure 4.7: PGS accuracy by $F_{ST}$ for standing height and LDL.** A plot of the mean $R^2$ of a PGS against the difference in $F_{ST}$ from the White British in the UKB. We use clusters extracted using HDBSCAN($\hat{\epsilon}$). There is a negative linear relationship between $F_{ST}$ from the largest cluster and PGS accuracy. (a) Top: A PGS of height shows a strong decay between $R^2$ and $F_{ST}$, as expected. (b) Bottom: A PGS of LDL-cholesterol has an unclear relationship between $R^2$ and $F_{ST}$. Cluster $18$ has the largest $F_{ST}$ but one of the highest $R^2$ values; the cluster also has the highest frequency of the $rs7412$ and $rs4420638$ alleles.

### 4.4.5 Quality control for complex multi-ethnic cohorts

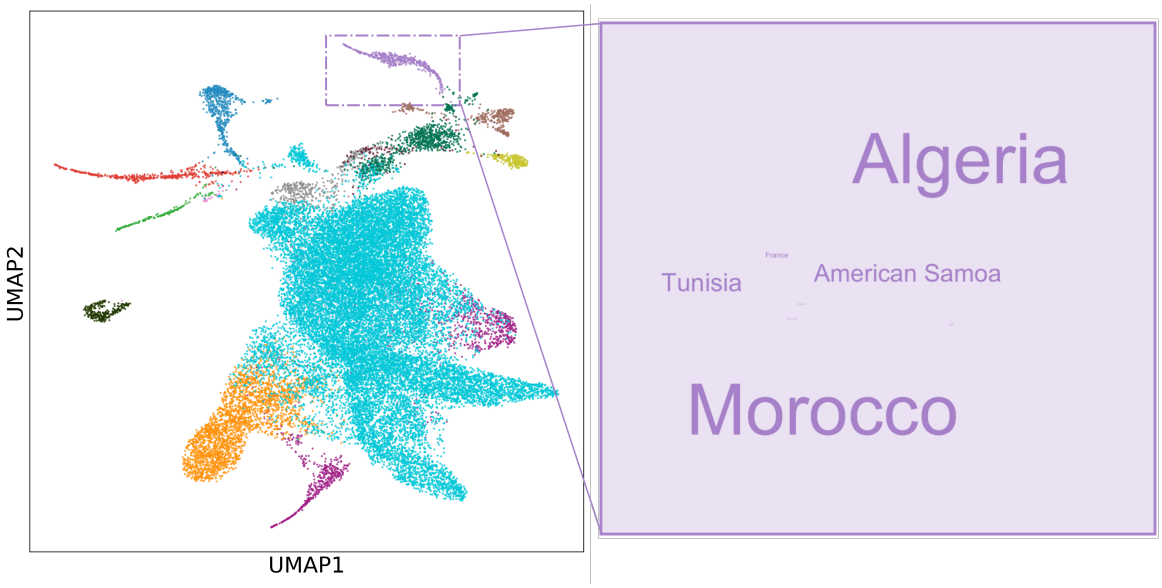



**Figure 4.8: Clustering can identify data collection errors.** A 2D UMAP of CARTaGENE data coloured by clusters extracted using HDBSCAN($\hat{\epsilon}$). The highlighted cluster was found to have most of its individuals born in North Africa. A word cloud shows that a significant minority of individuals were born in American Samoa, which was found to be a coding error.

Generally the fine-scale structure of biobank data is not known in advance. The structure of under-represented groups in particular, such as minority populations or those with complex histories of recent migration and admixture, can also be intricate and poorly understood, at least by geneticists. Individuals with uncommon combinations of ancestral, geographic, and ethnic descriptors are present in all biobanks. These combinations can be real and represent the completely different nature of genetic ancestry and ethnicity; they may also represent clerical errors[38]. Distinguishing the two is especially relevant when biobanks are used as sample frames for deeper sequencing or for follow-up studies, and when variables like country of birth and ethnicity are

used as selection criteria. Using HDBSCAN($\hat{\epsilon}$) to explore the relationship between clusters membership and auxiliary variables can detect data collection errors before sample selection is carried out, preventing serious methodology problems or unnecessary exclusion of individuals.

CARTaGENE is a biobank of residents from Quebec, Canada, that has recently genotyped $29,337$ individuals[20]. We were interested in identifying populations of North African descent for further study. In Figure 4.8, we identified a cluster of $446$ people born largely in North Africa with $51$ individuals ($11.4\%$) recorded as being born in American Samoa, an American island territory in the South Pacific Ocean with fewer than $50,000$ inhabitants. After researching possible historical explanations (e.g. migration between American Samoa and North Africa), we traced the result to a coding error from different country codes used over the course of data collection; the actual birth country was corrected to Algeria. The same coding error was found in other clusters, affecting $266$ individuals born in $43$ countries. While this error was easy to discover using $HDBSCAN(\hat{\epsilon})$, it is not obvious whether or how it would have been identified otherwise given that it affected less than $1\%$ of the cohort. Efficient data exploration, aided by visualization and clustering, remains one of our best tools to combat the dual evils of bookkeeping errors and batch effects.

## 4.5 Discussion

We present UMAP-HDBSCAN($\hat{\epsilon}$), a new approach to capturing population structure that approximates the topology of high-dimensional genetic data and detects dense clusters in a low-dimensional space. This approach does not assume a specific number of populations, nor does

it assume that there is a type genome for a given population. Our method requires neither reference panels nor *a priori* definitions of populations, but can use auxiliary data such as population labels, country of birth, ethnicity, geographic coordinates, etc., to characterize the clusters *a posteriori* and learn about their history or origins. With tools like PCA and UMAP already common in population genetics[39], it integrates easily with existing analysis pipelines. Given UMAP data from the UKB—a matrix of dimension $(488,377 \times 5)$—HDBSCAN($\hat{\epsilon}$) takes under $60$ seconds to execute on a single core, making it tractable for large-scale data. Being robust to the presence of many populations of widely varying sizes, it is a powerful and flexible method and is well-suited to modern biobanks.

Stratification is important in data exploration and analysis, and many stratification strategies have been proposed. Using self-identification is common, and can be appropriate if the outcome of interest is tied to identity. It is however an inconsistent measure of genetic ancestry, and is limited to identities that are available in questionnaire data. Its use in genetic screening has led to missed carriers in at-risk populations[11], and a report from the US National Academies of Sciences, Engineering, and Medicine has recommended against using such variables as proxies for genetic variation[10].

The most commonly used metrics for fine-scale genetic community identification are based on recent relatedness. One such approach is identity-by-descent (IBD; see e.g. [8, 9, 40, 41]), which has been used for downstream clustering (e.g. characterizing demographic history in [42] and identifying selection within populations in [43]). An IBD-based approach in ATLAS, for example,

recently identified associations between genetic clusters and genetic, clinical, and environmental data[44]. The ability of IBD clustering to identify fine-scale structure can be due to two effects. First, it focuses on recent relatedness between individuals, which may be helpful in identifying recent demographic effects. Second, because it is by nature pairwise, it encourages the use of clustering methods that focus less on archetypes and more on genetic neighbourhoods, i.e., on more topological approaches.

The topological approach presented here only uses overall genetic similarity— which reflects both recent and background relatedness—to capture population structure. Since it bypasses the need to perform phasing and IBD calling, it requires fewer analytical tools and computational resources. Because IBD clustering is demanding, and because researchers are often interested in identifying clusters they see in PCA or UMAP space (which may not relate to IBD clusters) researchers commonly rely on hand-selected delineations of dimensionally-reduced data (e.g. [2, 12]). The approach we propose is faster, less arbitrary, identifies structure at a finer scale, and takes advantage of the higher-dimensional nature of the data to identify structure more consistently.

A recent publication on polygenic scoring by Ding et al[7] suggested moving entirely away from stratification based on genetic clusters. Instead, they argued in favour of individual-level measures. They cite three issues with clusters: (i) clustering algorithms fail to capture populations without reference panels, such as those that are relatively small or recently admixed; (ii) clusters ignore inter-individual variation; and (iii) clustering results change based on algorithms and reference panels. We believe that these criticisms are valid for the type of stratification they considered:

Ding et al clustered UKB data based on proximity to an archetype in PCA-space—if an individual fell within a certain distance of one of nine pre-defined population centroids, they were considered a member of a cluster; otherwise, their ancestry was considered unknown.

We believe that the first two objections can be resolved by topological approaches. In the UKB, $91\%$ of participants were placed into clusters in [7]. In contrast, across $604$ parametrizations, the median percentage of individuals placed in a cluster was $99.99\%$ (Figure 4s10), with the three worst-performing runs of UMAP-HDBSCAN($\hat{\epsilon}$) respectively assigning $99.11\%$, $99.69\%$, and $99.86\%$ of individuals in the UKB to a cluster. The clusters reflect groups that have shared genetic and geographic histories, including for relatively small and recently admixed groups which were often excluded based on prior approaches[7, 15]. We achieved similar results with CaG and 1KGP data, suggesting that our approach is robust to the idiosyncratic composition of a biobank.

### 4.5.1 Applications

Understanding the population structure of a biobank is a necessary precursor to many analyses. In the 1KGP, the source of its structure is largely the sampling scheme, which is reflected in Figure 4.2—to ensure diversity in the data, the populations were deliberately sampled from multiple locations around the world with similar sample sizes. The sources of population structure of the UKB, on the other hand, reflect a complex history of migrations at different geographic and time scales, including isolation by distance within the UK and recent immigration and admixture between populations from regions of the former British Empire.

We chose the clustering in Figure 4.4a based on its suitability for examining PGS transferability and examining allele frequency versus PGS accuracy. Alternative parameterizations highlight structure at different scales. In Figure 4s11, for example, we see the large cluster of mostly European-born individuals is split into three smaller clusters. The structure of a typical biobank is more similar to the UKB than the 1KGP, as the recruitment methodology is often based on residence within a jurisdiction. Examples include municipal (ATLAS in Los Angeles[44], Bio*Me* in New York City[45]), regional (CARTaGENE in Quebec[20]) and national (Million Veterans Project (MVP),[46], CANPATH[47]) biobanks. Leveraging these diverse cohorts can improve variant discovery[48, 49].

Though population labels like ethnicity can be useful, individuals may identify as "Other" or "Unknown", leading to incomplete data. In the MVP, missing data were imputed using a support vector machine trained on race/ethnicity data to harmonize genetic data with labels for an ethnicity-specific GWAS[50]. A similar supervised approach with random forests was used by gnomAD[51]. Rather than assigning ethnicities to individuals, we constructed clusters from genetic data and investigated the distributions of auxiliary variables within clusters, including missing values. We found word clouds to be well-suited for describing data without imposing a reductive label.

The goal of genetic stratification is in no way to replace self-declared variables in contexts where they are relevant. In fact, genetic stratification revealed interesting trends in self-declared variables. For example, in Cluster 17 of Figure 4.4a, $97.6\%$ of individuals were born in Britain and Ireland and $99.5\%$ chose an ethnic group label; in contrast, $18.9\%$ of those in Cluster 18 (mostly

born in sub-Saharan Africa) and $36.5\%$ in Cluster 23 (mostly born in the Horn of Africa) chose "Other", highlighting differential completeness of questionnaire data. As mentioned above, UKB strata with "mixed" ethnic backgrounds as their mode featured multiple ethnic background labels, likely reflecting both the fact that (genetically) admixed individuals may have a diversity of ethnic backgrounds, and the fact that individuals with both mixed genetic and cultural heritage may have to choose among potentially inadequate labels (see, e.g., discussion in [15]). The presence or absence of a label in data collection can critically influence how people identify: Canadian demographers noted that between the 2011 National Household Survey and the 2016 Census, there was a $53.6\%$ drop in people who identified as "Jewish"—this result was traced to the list of ethnicities not presenting the label as an example in 2016[52].

### 4.5.2 Considerations

Unlike archetype-based methods, HDBSCAN($\hat{\epsilon}$) identifies groups that can be created by linking nearby individuals—it is possible to have a very long chain containing many individuals who are each closely related to those near them within a cluster but not to those at the distant end. In this way, admixed populations can form a single cluster even though individuals within the cluster can differ as much as individuals from the different ancestral "source" populations. In a sense, HDBSCAN($\hat{\epsilon}$) identifies groups of individuals whose distribution in genetic space suggests a common sampling or demographic history, rather than genetic similarity. For this reason, topological stratification may be less conducive to reification of clusters and the notion that population labels

reflect a true underlying "type". However, given the weaponization of population genetics research in the past[53], it is worth emphasizing limitations common to all clustering approaches.

No single label is an individual's "true" ancestry, race, or ethnicity, as these are complex, multifactorial population descriptors[15, 54]. Thus clustering does not have a well-defined ground truth [55], and clusters are most useful as "helpful constructs that support clarification"[56]. With real genetic data, there is no "correct" number of populations[57] and discrete groupings provide a flattened view of a high-dimensional landscape[15, 29]. The clusters generated are sensitive to the input samples, since the demographic composition of a biobank will impact the clustering, and they are also affected by the parameters at the filtering, dimensionality reduction, and clustering steps. This is a reflection of the data, as genetic data are not composed of "natural types". These clusters can be useful in understanding how genetics relates to health and the environment, but it is worth repeating that variation in phenotypes across genetic clusters does not imply a genetic cause, as differences in environment or systemic discrimination are also expected to produce such variation[58]. Each identified cluster also likely features considerable genetic heterogeneity. The UK biobank clusters of majority sub-Saharan-born individuals, for example, encompass considerable genetic substructure[59]. Different choices of metrics for clustering (i.e., genetic relatedness vs. IBD) can emphasize different types of structure. There are no true clusters.

Ultimately, however, many useful analyses require some definition of "populations". For example, an allele frequency can only be calculated and reported within a population. Data exploration and quality control often require investigating relevant subsets of the data to decide whether they

likely reflect technical artefacts or meaningful subgroups. To date there has not been a method of stratification that is tractable, easy to implement, robust to the presence of many populations of many sizes, and that captures all or almost all individuals with complex population histories. We believe our topological approach satisfies these important needs. Looking forward, we expect that topological approaches underlying UMAP and HDBSCAN($\hat{\epsilon}$) also present a promising avenue to move towards a more continuous description of genetic variation in complex cohorts.

## 4.6 Acknowledgements

We are grateful to the participants in each biobank who provided their genetic data. We thank the CARTaGENE team for troubleshooting data with us, and C. Bhérer, M. L. Spear, and P. Verdu for scientific discussion.

**Funding:** This research was also supported by the Canadian Institute for Health Research (CIHR) project grant 437576, Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2017-04816, the Canada Research Chair program, and the Canada Foundation for Innovation.

## 4.7 Materials and Methods

Our code is available at `https://github.com/diazale/topstrat`. We have provided command line tools to run Python implementations of UMAP and HDBSCAN($\hat{\epsilon}$).

We used three datasets for this analysis: the 1000 Genomes project (1KGP), the UK biobank

(UKB), and CARTaGENE (CaG). For the 1KGP we used $3,450$ genotypes using Affy 6.0 genotyping[18]. We generated the principal components using a Python script and have made the top PCs available in the repository to demonstrate the code. We used the genotype and population label files:

- `ALL.wgs.nhgri_coriell_affy_6.20140825.genotypes_has_ped.vcf.gz`
- `affy_samples.20141118.panel 20131219.populations.tsv`

available at `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/`.

For the UKB, we limited our analyses to the $488,377$ individuals with genotype data. We used the UKB's top 40 pre-computed PCs (Data-Field 22009), blood cell counts (Data-Fields 30000, 30010, 30120, 30130, 30140, 30150, 30160), lung function measures (Data-Fields 3062, 3063), age (Data-Field 21003), sex (Data-Field 31), standing height (Data-Field 50), weight (Data-Field 21002), BMI (Data-Field 21001), smoking status (Data-Field 20116), country of birth (Data-Fields 1647, 20115), and ethnic group/background (Data-Field 21000). Ethnic group/background is a hierarchical item in which participants are prompted to select from a pre-populated list of options for Ethnic Group (e.g. "White") and, if available, a secondary option for Ethnic Background (e.g. "British"). Phenotypes used in analyses were normalized with respect to variables $sex$, $age$, and $age^2$. Access to the UKB can be granted at `https://www.ukbiobank.ac.uk/scientists-3/genetic-data/`.

For CARTaGENE, we used $29,337$ individuals with genotype data. We generated the PCs using PLINK[60] after filtering for linkage disequilibrium and HLA (chromosome 6, $25000000$–$33500000$ ). The options used were:

- `indep-pairwise 1000 50 0.1` (PLINK2)
- `maf 0.05`
- `mind 0.1`
- `geno 0.1`
- `hwe 1e-6`.

We used the Python implementations of UMAP[16] (0.3.6) and HDBSCAN (0.8.24), integrating the updates from Malzer and Baum[17]. To calculate PGS, we used VIPRS[31].

## 4.8 Supporting information

For visualization, we reduce our data to 2D via UMAP and set a relatively high minimum distance ($MD$; usually between $0.3$ and $0.5$); this enables us to view fine-scale patterns of structure. We find satisfactory results with the number of neighbours ($NN$) varying from $15$ to $50$; higher values will require more computational resources, but they increase the connectivity between points in the data, as discussed in [39]. For clustering, we set a low value of minimum distance (equal to or close to $0$) and reduce the number of dimensions to at least $3$—in our analyses, we used $3$, $4$, and

$5$ dimensions. The low minimum distance encourages dimensionally-reduced data to form dense clusters, while keeping the dimensionality at $\geq 3$ preserves the complexities of data that can be lost because of artificial tearing in the drop from $3$ to $2$ dimensions. The number of neighbours will vary depending on what is a reasonable expectation for the data. For the 1KGP data, which consists of geographically diverse samples of roughly similar size, $50$ neighbours capture the structure well. For biobank data, it is common for structure to arise from a handful of individuals; we found $10$ to $25$ neighbours to work best. Lower neighbourhood values (e.g. $NN = 5$) will create smaller clusters, but can also highlight highly-localized structure within larger populations. $2D$ visualizations can give intuition as to the presence and sizes of clusters. If pre-processing the data with PCA, more PCs tend to reveal finer-scale structure (see e.g. the relationship with geographical coordinates in Figures S17 and S18 in [4]). For the 1KGP clusters in Figure 4.2 we used the top $16$ PCs; for the UKB in Figure 4.4a and CaG in Figure 4.8 we used the top $25$.

In parametrizing HDBSCAN($\hat{\epsilon}$), the parameter $\hat{\epsilon}$ defines a threshold at which clusters are merged or split. We find values of $\hat{\epsilon}$ ranging from $0.3$ to $0.5$ to be effective at ensuring all or almost all individuals are clustered while still identifying fine-scale structure. The minimum number of points ($MP$) should not be significantly higher than the number of neighbours used in the associated UMAP. If $MP$ is high and $NN$ is low, it can result in a large number of points being classified as noise since the UMAP data will tend to form small clusters; e.g. a UMAP parametrized with $NN = 10$ and HDBSCAN($\hat{\epsilon}$) with $MP = 100$ may return poor results.

Changing parameters will result in different clusters being generated. Given the low computa-

tional costs of UMAP and HDBSCAN($\hat{\epsilon}$), we recommend running a grid search for visualization and exploratory analysis. Clusters can then be characterized using auxiliary data, such as country of birth, geographical location, population label, self-identification, etc. We selected the clustering for the UKB for its suitability for comparing PGS results by $F_{ST}$ from the training population. For CaG, we selected one of the clustering runs that generated a cluster of individuals with North African ancestry.

We calculated pairwise $F_{ST}$ for UKB clusters using PLINK[60]. We calculated admixture proportions using ADMIXTURE 1.3.0[61]. For computational reasons, for the UKB we calculated admixture proportions on individuals not falling into Cluster 17 (the largest cluster, containing around $400,000$ individuals) in Figure 4.4a.

Visualizations and statistical analyses were done in R (3.5.3)[62]. We used ggplot2[63] for graphics and ggwordcloud for word clouds, and stargazer[64] to generate tables.

For phenotype smoothing, we removed the effects of the top $40$ PCs using linear regression, working with the residuals. For phenotype $p$ and individuals $i = 1 \dots I$, we use the model:

$$y_{p,i} = \beta_{p,0} + \sum_{j=1}^{40} \beta_{p,j} PC_{j,i} + \epsilon_{p,i}, \epsilon_{p,i} \sim N(0, \sigma_p^2)$$

We visualize the data in Figure 4.5 with the values $e_{p,i} = y_{p,i} - (\hat{\beta_{p,0}} + \sum_{j=1}^{40} \hat{\beta_{p,j}} PC_{j,i})$

## 4.9 Supplementary figures and tables

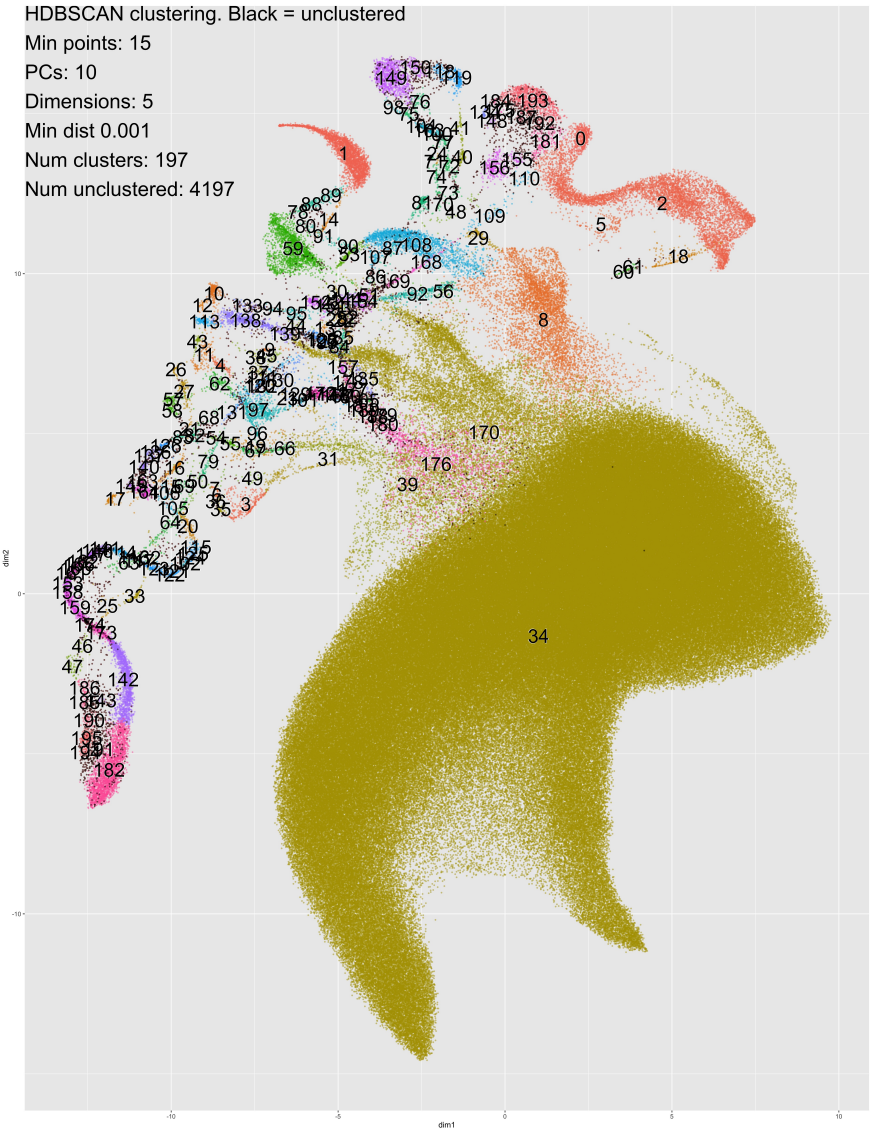
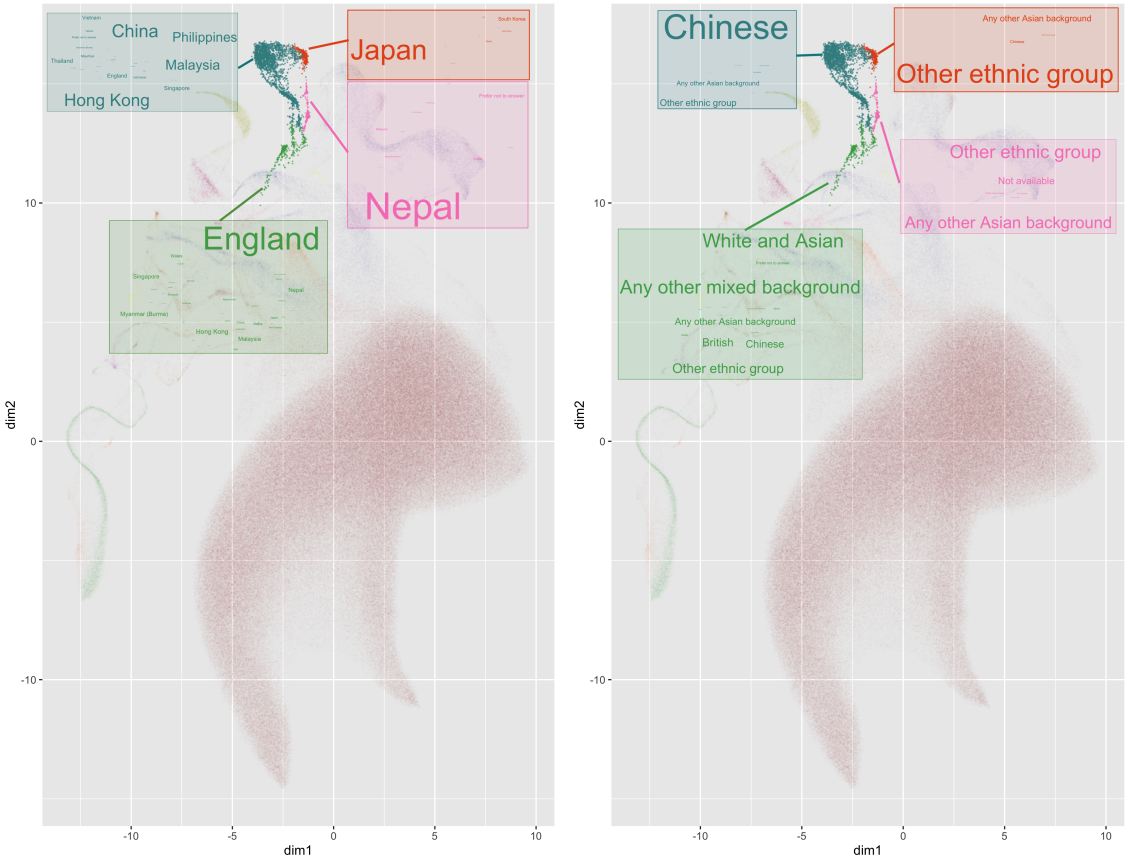



**Figure 4s1:** An example of a clustering of the UKB data using HDBSCAN rather than HDBSCAN($\hat{\epsilon}$). The algorithm fails to adequately cluster many of the sub-populations, categorizing $4,197$ individuals as noise and generated $197$ micro-clusters.




**(a)** **(b)**

**Figure 4s2:** Word clouds generated from four clusters in the UKB from Figure 4.4. (a) Left: Word clouds of the most common countries of birth within each cluster. Most individuals in the orange cluster (Cluster 0) were born in Japan, and most in the pink cluster (Cluster 15) were born in Nepal. (b) Right: Word clouds for the most common EB. The most common in the blue cluster (Cluster 13) was "Chinese", while those in the green cluster (Cluster 14) select a variety, including "White British", "Chinese", "Mixed", or "Other". Detailed breakdowns are available in Tables 4s7 and 4s8.

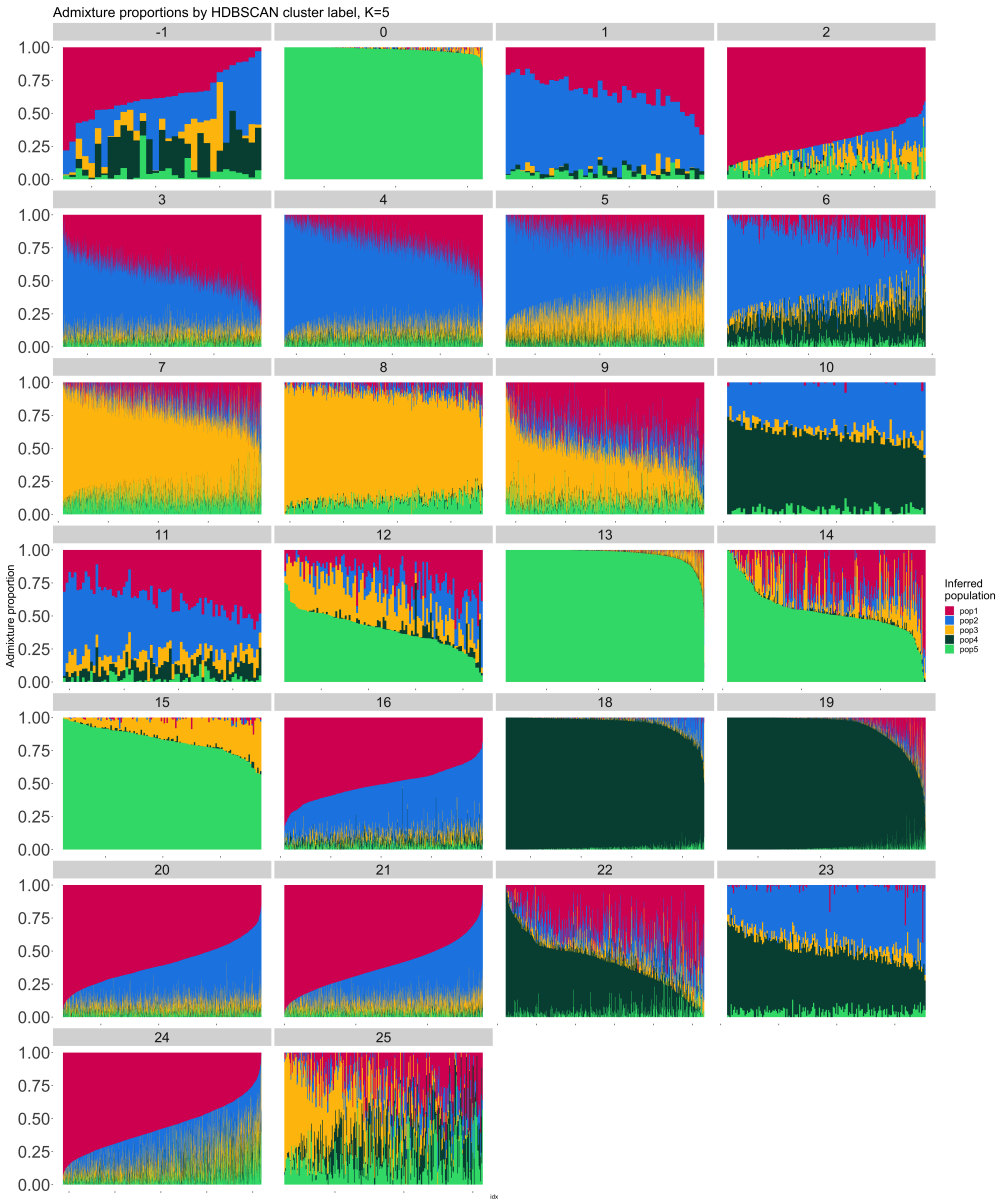



**Figure 4s3:** Admixture proportions for $K = 5$ populations on each of the clusters in Figure 4.4. Cluster 17 ($n > 400,000$) was excluded for computational reasons. Individuals not assigned to a cluster are labelled as $-1$.

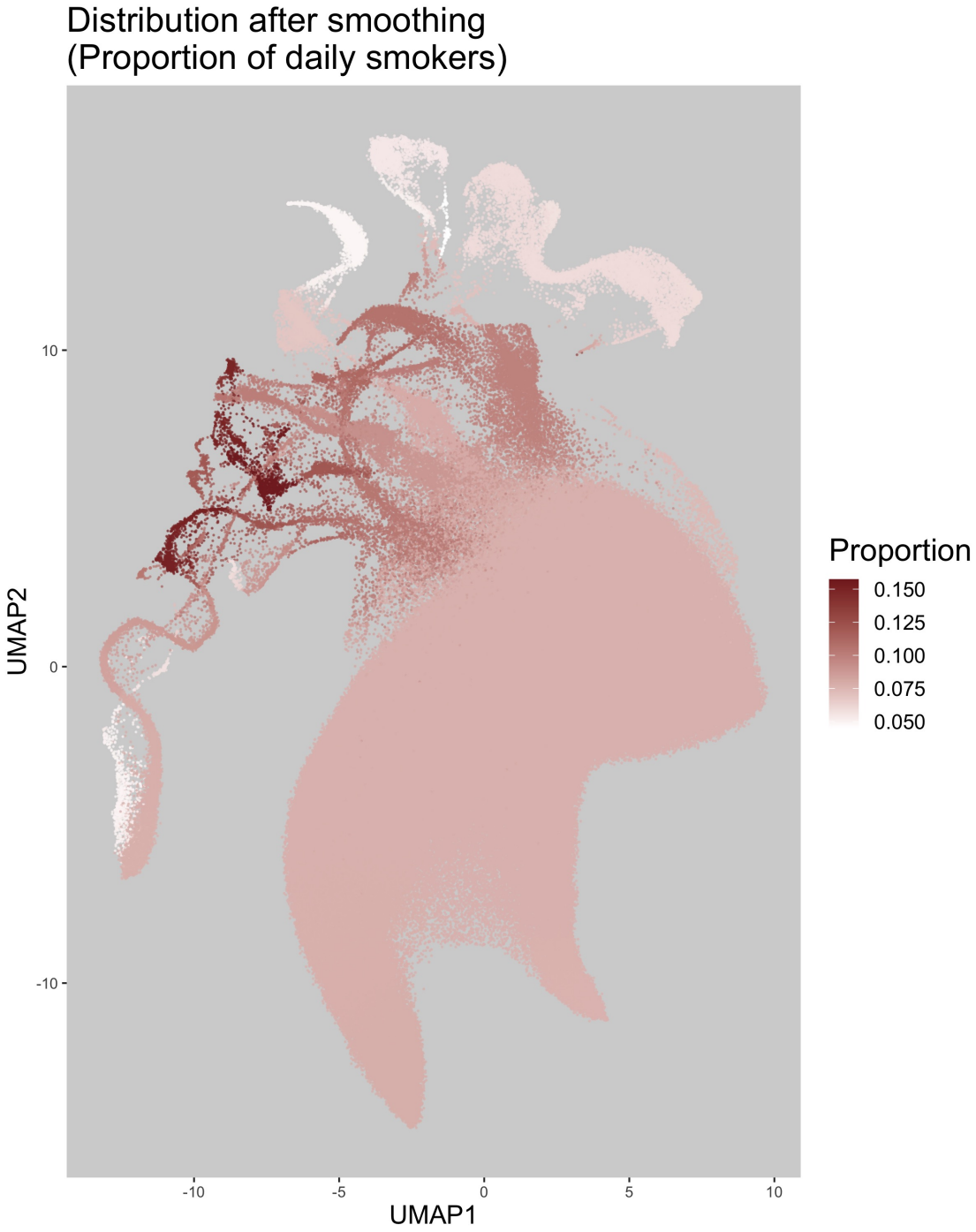
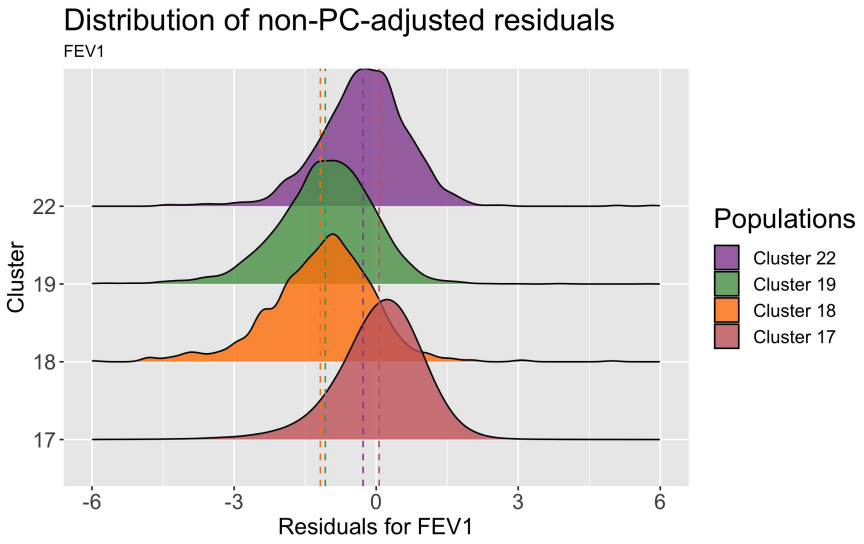
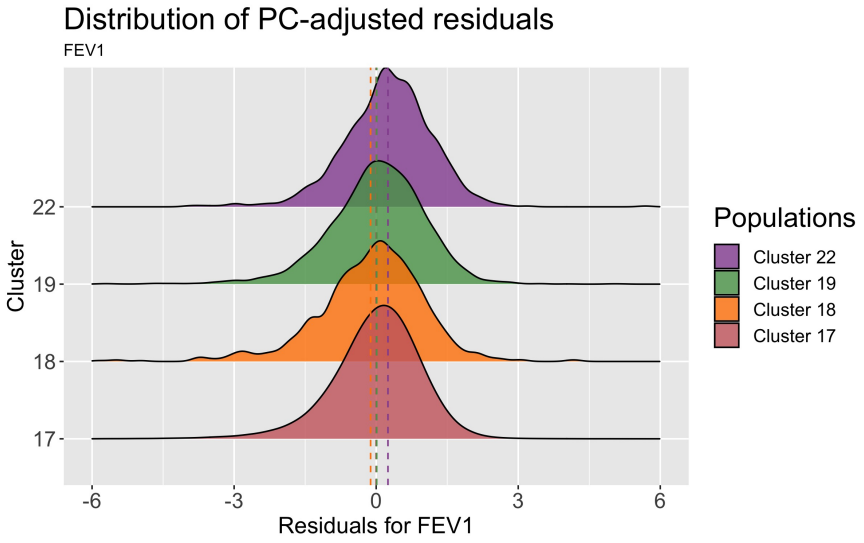



**Figure 4s4:** Proportion of daily smokers, smoothed using Algorithm 1.

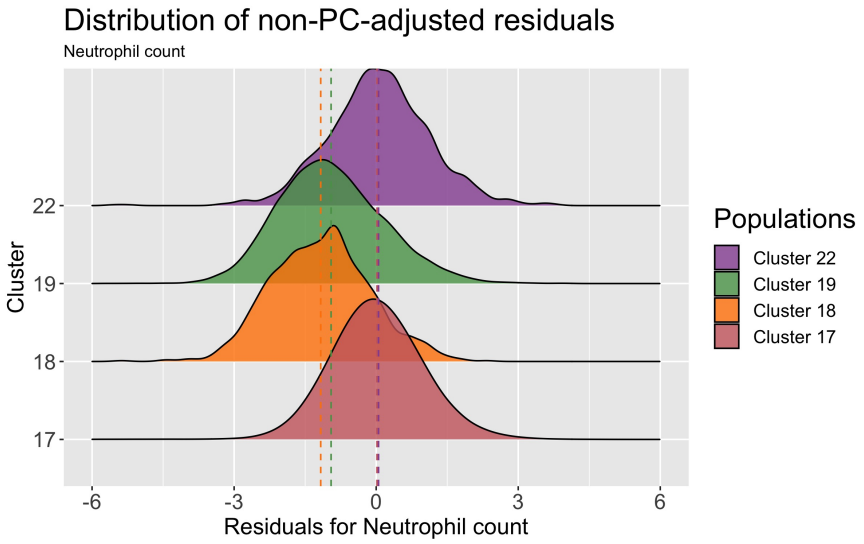



**(a)**

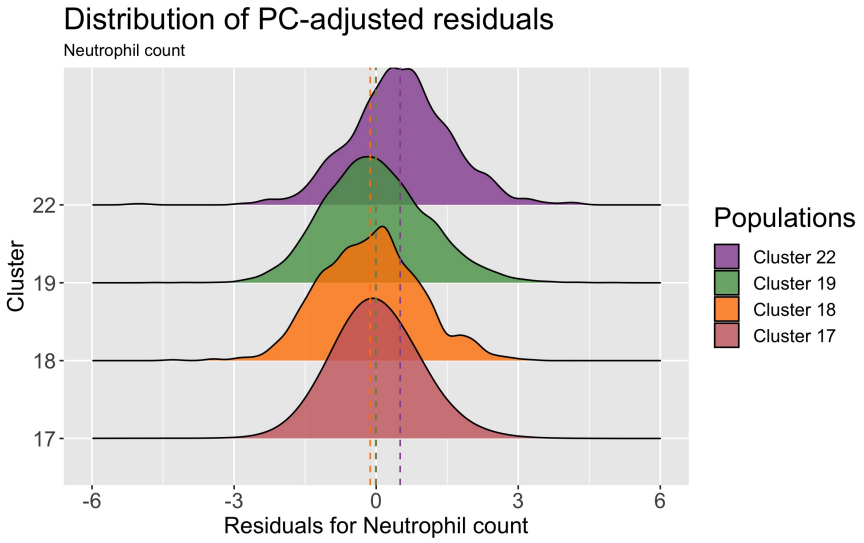



**(b)**

**Figure 4s5:** Distributions of FEV1 adjusted for age and sex stratified by cluster. Vertical dotted lines represent the mean of the distribution. Cluster labels and colours match those in Figure 4.4a. Cluster 17 is mostly European-born individuals, Cluster 18 is mostly sub-Saharan African born individuals, Cluster 19 is mostly individuals born in England, the Caribbean, Ghana, and Nigeria, and Cluster 22 is mostly individuals born in England who chose the EB "White and Black Caribbean" or "White and Black African". (a) Top: Distribution of FEV1 by cluster without adjusting for population structure. (b) Bottom: Distribution of FEV1 by cluster after having adjusted for the top 40 principal components.




**(a)**




**(b)**

**Figure 4s6:** Distributions of neutrophil count adjusted for age and sex stratified by cluster. Vertical dotted lines represent the mean of the distribution. Cluster labels and colours match those in Figure 4.4a. Cluster 17 is mostly European-born individuals, Cluster 18 is mostly sub-Saharan African born individuals, Cluster 19 is mostly individuals born in England, the Caribbean, Ghana, and Nigeria, and Cluster 22 is mostly individuals born in England who chose the EB "White and Black Caribbean" or "White and Black African". (a) Top: Distribution of neutrophil count by cluster without adjusting for population structure. (b) Bottom: Distribution of neutrophil count by cluster after having adjusted for the top 40 principal components.

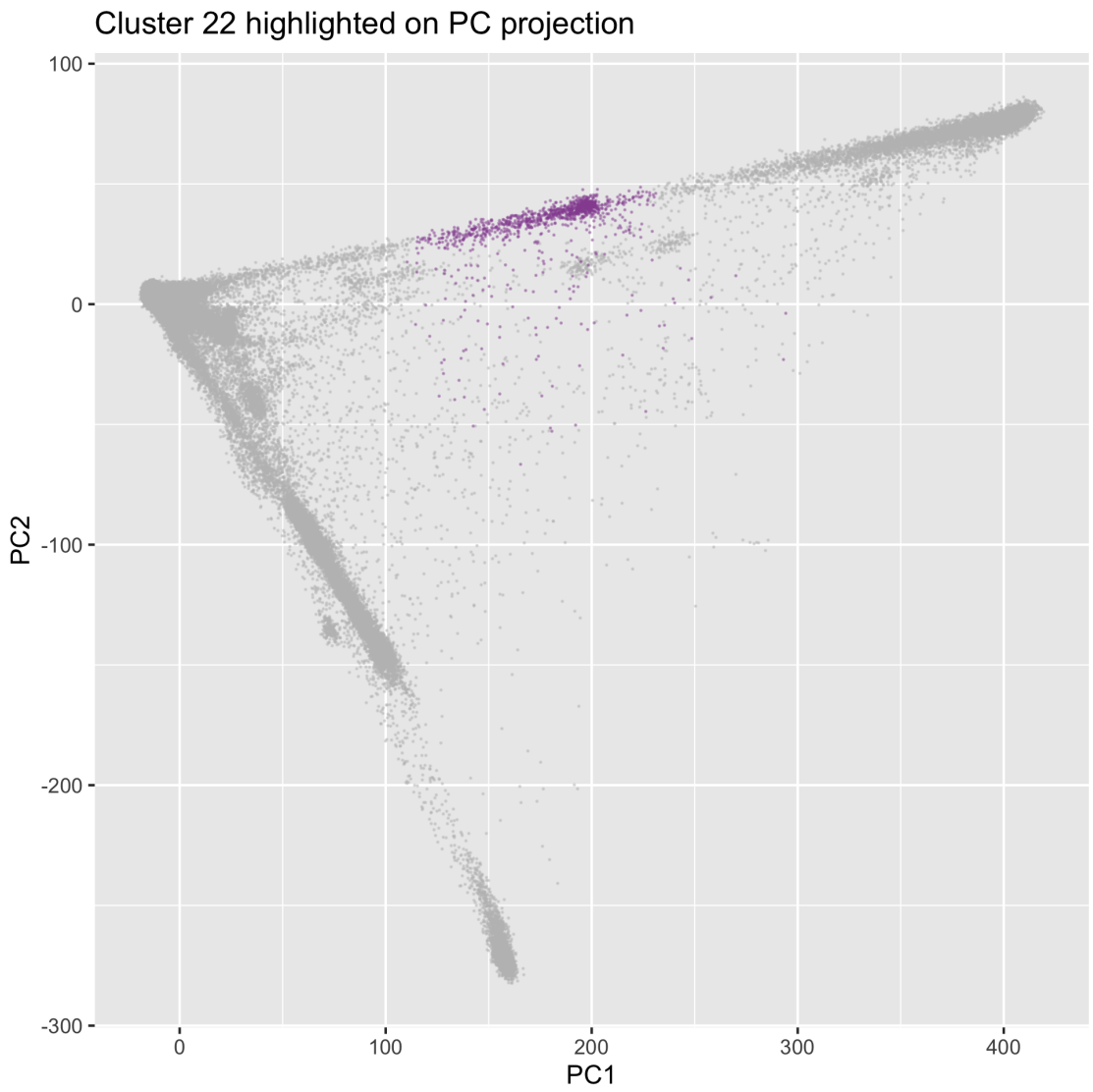



**Figure 4s7:** Cluster 22 from Figure 4.4a highlighted coloured in on a plot of PC1 and PC2.

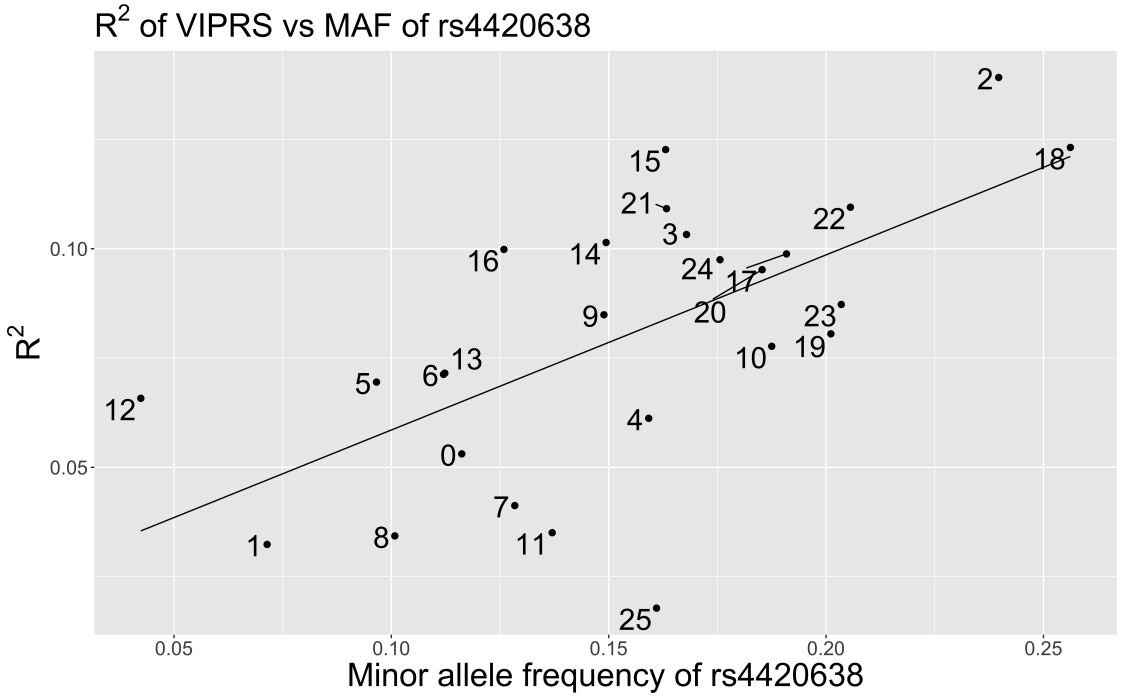



**Figure 4s8:** Regression line of the $R^2$ of a PGS generated by VIPRS versus the minor allele frequency $rs4420638$, labelled by clusters from Figure 4.4a. The regression summary is presented in Table 4s1.

| | *Dependent variable:* |
|---|---|
| | R2 |
| MAF | 0.401*** |
| | (0.100) |
| Constant | 0.018 |
| | (0.016) |
| Observations | 26 |
| $R^2$ | 0.402 |
| Adjusted $R^2$ | 0.378 |
| Residual Std. Error | 0.025 (df = 24) |
| F Statistic | 16.163*** (df = 1; 24) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

**Table 4s1:** Linear regression model between minor allele frequency (MAF) of $rs4420638$ within each cluster from Figure 4.4a and the $R^2$ of a PGS for LDL generated by VIPRS using the clusters from Figure 4.4a. The plot of the regression is present in Figure 4s8.

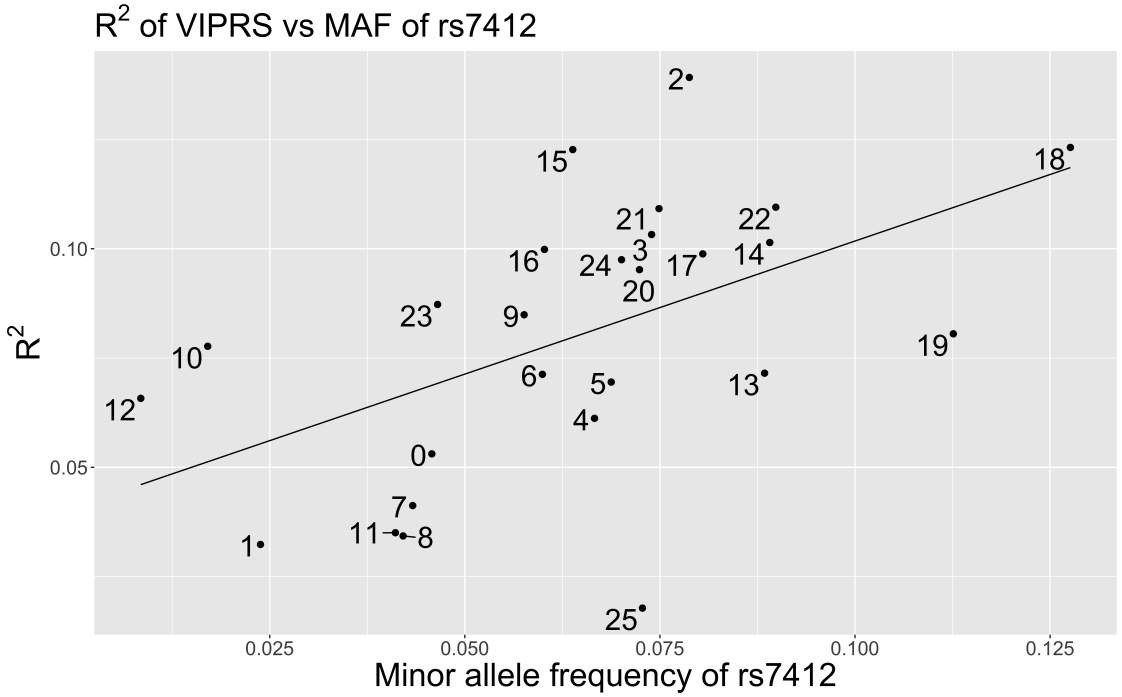
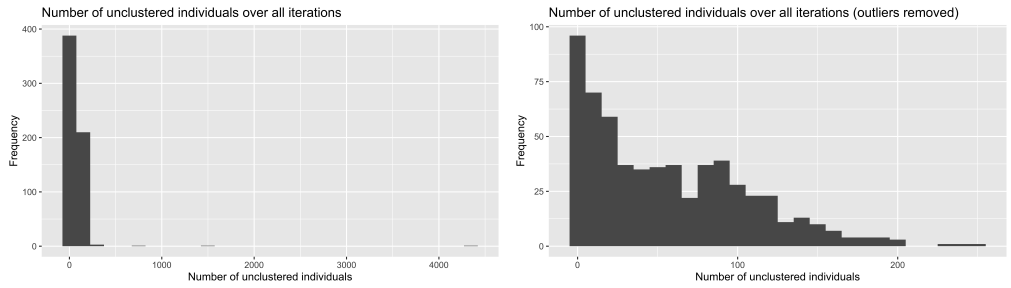



**Figure 4s9:** Regression line of the $R^2$ of a PGS generated by VIPRS versus the minor allele frequency $rs7412$, labelled by clusters from Figure 4.4a. The regression summary is presented in Table 4s2.

| | *Dependent variable:* |
|---|---|
| | R2 |
| MAF | 0.609*** |
| | (0.202) |
| Constant | 0.041*** |
| | (0.014) |
| Observations | 26 |
| $R^2$ | 0.275 |
| Adjusted $R^2$ | 0.245 |
| Residual Std. Error | 0.027 (df = 24) |
| F Statistic | 9.126*** (df = 1; 24) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

**Table 4s2:** Linear regression model between minor allele frequency (MAF) of $rs7412$ within each cluster from Figure 4.4a and the $R^2$ of a PGS for LDL generated by VIPRS using the clusters from Figure 4.4a. The plot of the regression is present in Figure 4s9.




**(a)** **(b)**

**Figure 4s10:** For each of the $604$ runs of UMAP-HDBSCAN($\hat{\epsilon}$) on the UKB, we count the number of individuals not assigned to a cluster. (a) Top: Across all $604$ runs. (b) Bottom: To improve the scale of the figure, we remove $3$ outlier runs in which $684$, $1,535$, and $4,346$ individuals were not assigned to a cluster.

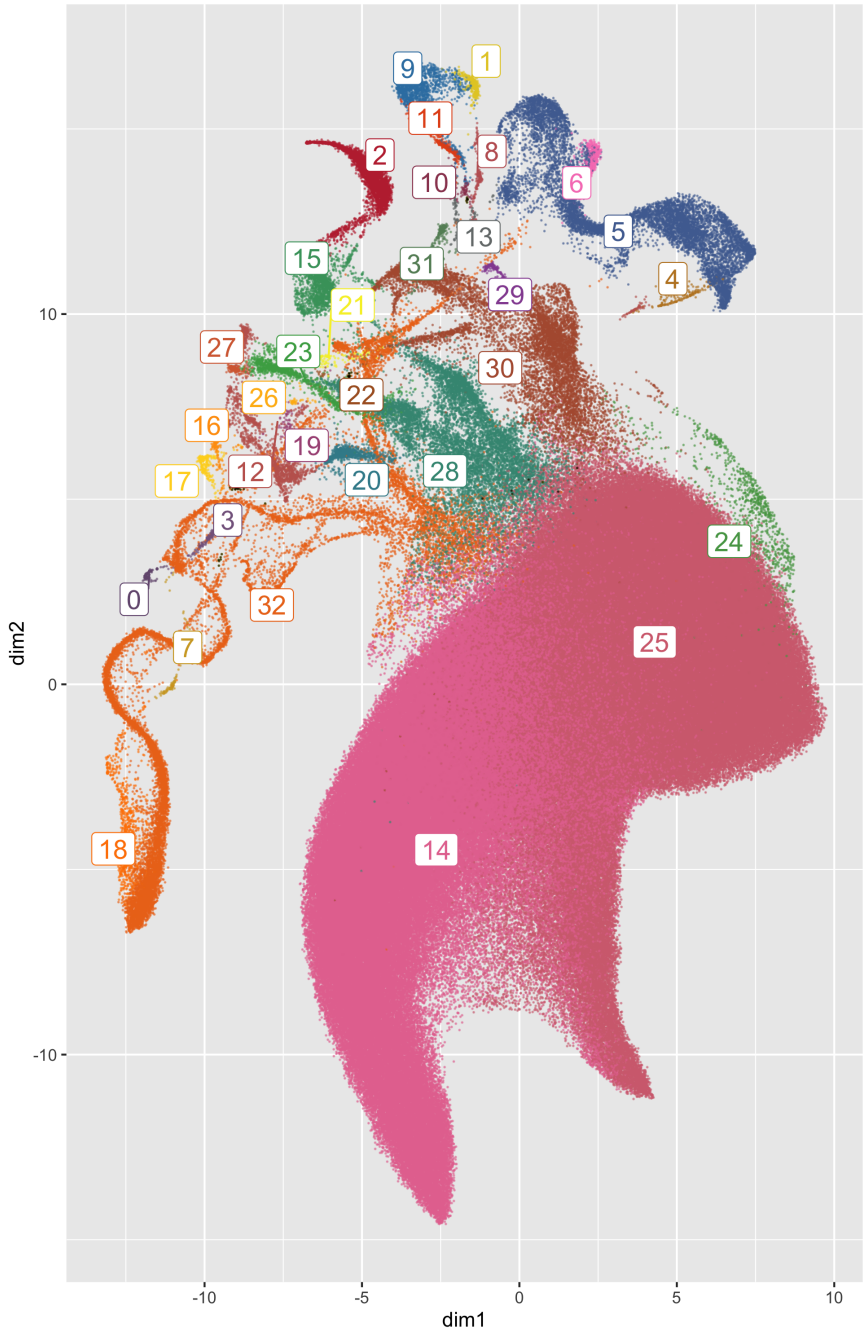



**Figure 4s11:** An alternative clustering of UKB data. Compared to Figure 4.4a, the largest cluster (Cluster 17 in that figure) has been split into three smaller clusters (Clusters 14, 24, 25 in this figure). Other clusters have been split or merged, while some remain the same between runs.

| Abbreviation | Population name |
|---|---|
| ACB | African Caribbean in Barbados |
| ASW | African Ancestry in SW USA |
| BEB | Bengali in Bangladesh |
| CDX | Chinese Dai in Xishuangbanna, China |
| CEU | Utah residents with Northern/Western European ancestry |
| CHB | Han Chinese in Beijing, China |
| CHS | Han Chinese South |
| CLM | Colombian in Medellín, Colombia |
| ESN | Esan in Nigeria |
| FIN | Finnish in Finland |
| GBR | British From England and Scotland |
| GWD | Gambian in Western Division – Mandinka |
| GIH | Gujarati Indians in Houston, Texas, USA |
| IBS | Iberian Populations in Spain |
| ITU | Indian Telugu in the UK |
| JPT | Japanese in Tokyo, Japan |
| KHV | Kinh in Ho Chi Minh City, Vietnam |
| LWK | Luhya in Webuye, Kenya |
| MSL | Mende in Sierra Leone |
| MXL | Mexican Ancestry in Los Angeles, CA, USA |
| PEL | Peruvian in Lima, Peru |
| PJL | Punjabi in Lahore, Pakistan |
| PUR | Puerto Rican in Puerto Rico |
| STU | Sri Lankan Tamil in the UK |
| TSI | Toscani in Italy |
| YRI | Yoruba in Ibadan, Nigeria |

**Table 4s3:** Names and abbreviations of 1KGP populations.

| 1KGP population | Cluster label | 1KGP in cluster | Total in 1KGP | Proportion in cluster |
|---|---|---|---|---|
| ACB | 0 | 122 | 122 | 1.0000000 |
| ASW | 0 | 103 | 107 | 0.9626168 |
| ASW | 5 | 3 | 107 | 0.0280374 |
| ASW | 15 | 1 | 107 | 0.0093458 |
| BEB | 1 | 133 | 143 | 0.9300699 |
| BEB | 11 | 10 | 143 | 0.0699301 |
| CDX | 2 | 104 | 105 | 0.9904762 |
| CDX | 4 | 1 | 105 | 0.0095238 |
| CEU | 3 | 183 | 183 | 1.0000000 |
| CHB | 4 | 105 | 105 | 1.0000000 |
| CHS | 4 | 171 | 171 | 1.0000000 |
| CLM | 5 | 142 | 146 | 0.9726027 |
| CLM | 15 | 4 | 146 | 0.0273973 |
| ESN | 6 | 172 | 172 | 1.0000000 |
| FIN | 7 | 104 | 104 | 1.0000000 |
| GBR | 3 | 105 | 105 | 1.0000000 |
| GIH | 8 | 69 | 111 | 0.6216216 |
| GIH | 17 | 41 | 111 | 0.3693694 |
| GIH | 11 | 1 | 111 | 0.0090090 |
| GWD | 9 | 179 | 180 | 0.9944444 |
| GWD | 14 | 1 | 180 | 0.0055556 |
| IBS | 10 | 162 | 162 | 1.0000000 |
| ITU | 11 | 109 | 118 | 0.9237288 |
| ITU | 17 | 9 | 118 | 0.0762712 |
| JPT | 12 | 104 | 105 | 0.9904762 |
| JPT | 4 | 1 | 105 | 0.0095238 |
| KHV | 2 | 118 | 121 | 0.9752066 |
| KHV | 4 | 3 | 121 | 0.0247934 |
| LWK | 13 | 110 | 110 | 1.0000000 |
| MSL | 14 | 122 | 122 | 1.0000000 |
| MXL | 15 | 97 | 104 | 0.9326923 |
| MXL | 10 | 7 | 104 | 0.0673077 |
| PEL | 16 | 128 | 129 | 0.9922481 |
| PEL | 5 | 1 | 129 | 0.0077519 |
| PJL | 17 | 95 | 155 | 0.6129032 |
| PJL | 19 | 48 | 155 | 0.3096774 |
| PJL | 11 | 12 | 155 | 0.0774194 |
| PUR | 18 | 145 | 149 | 0.9731544 |
| PUR | 0 | 4 | 149 | 0.0268456 |
| STU | 11 | 124 | 128 | 0.9687500 |
| STU | 17 | 4 | 128 | 0.0312500 |
| TSI | 10 | 111 | 111 | 1.0000000 |
| YRI | 20 | 181 | 182 | 0.9945055 |
| YRI | 6 | 1 | 182 | 0.0054945 |

**Table 4s4:** Cluster assignments for each 1KGP population, showing how many individuals from each population ended up in each cluster.

| Cluster | 1KGP population | 1KGP population in cluster | Proportion |
|---|---|---|---|
| 0 | ACB | 122 | 0.5327511 |
| 0 | ASW | 103 | 0.4497817 |
| 0 | PUR | 4 | 0.0174672 |
| 1 | BEB | 133 | 1.0000000 |
| 2 | CDX | 104 | 0.4684685 |
| 2 | KHV | 118 | 0.5315315 |
| 3 | CEU | 183 | 0.6354167 |
| 3 | GBR | 105 | 0.3645833 |
| 4 | CDX | 1 | 0.0035587 |
| 4 | CHB | 105 | 0.3736655 |
| 4 | CHS | 171 | 0.6085409 |
| 4 | JPT | 1 | 0.0035587 |
| 4 | KHV | 3 | 0.0106762 |
| 5 | ASW | 3 | 0.0205479 |
| 5 | CLM | 142 | 0.9726027 |
| 5 | PEL | 1 | 0.0068493 |
| 6 | ESN | 172 | 0.9942197 |
| 6 | YRI | 1 | 0.0057803 |
| 7 | FIN | 104 | 1.0000000 |
| 8 | GIH | 69 | 1.0000000 |
| 9 | GWD | 179 | 1.0000000 |
| 10 | IBS | 162 | 0.5785714 |
| 10 | MXL | 7 | 0.0250000 |
| 10 | TSI | 111 | 0.3964286 |
| 11 | BEB | 10 | 0.0390625 |
| 11 | GIH | 1 | 0.0039062 |
| 11 | ITU | 109 | 0.4257812 |
| 11 | PJL | 12 | 0.0468750 |
| 11 | STU | 124 | 0.4843750 |
| 12 | JPT | 104 | 1.0000000 |
| 13 | LWK | 110 | 1.0000000 |
| 14 | GWD | 1 | 0.0081301 |
| 14 | MSL | 122 | 0.9918699 |
| 15 | ASW | 1 | 0.0098039 |
| 15 | CLM | 4 | 0.0392157 |
| 15 | MXL | 97 | 0.9509804 |
| 16 | PEL | 128 | 1.0000000 |
| 17 | GIH | 41 | 0.2751678 |
| 17 | ITU | 9 | 0.0604027 |
| 17 | PJL | 95 | 0.6375839 |
| 17 | STU | 4 | 0.0268456 |
| 18 | PUR | 145 | 1.0000000 |
| 19 | PJL | 48 | 1.0000000 |
| 20 | YRI | 181 | 1.0000000 |

**Table 4s5:** Composition of each cluster broken down by 1KGP population.

| Ethnic group | Ethnic background |
|---|---|
| White | British |
| White | Irish |
| White | Any other white background |
| Mixed | White and Black Caribbean |
| Mixed | White and Black African |
| Mixed | White and Asian |
| Mixed | Any other mixed background |
| Asian or Asian British | Indian |
| Asian or Asian British | Pakistani |
| Asian or Asian British | Bangladeshi |
| Asian or Asian British | Any other Asian background |
| Black or Black British | Caribbean |
| Black or Black British | African |
| Black or Black British | Any other Black background |
| Chinese | |
| Other ethnic group | |
| Do not know | |
| Prefer not to answer | |

**Table 4s6:** Possible values for ethnic background in the UKB (Data-Field 21000). Participants are first asked "What is your ethnic group?" and then asked "What is your ethnic background?" For "Chinese", there is no second question. Participants may also select "Prefer not to answer" for the second question; it is possible to have ethnic background recorded as ethnic group (e.g. just "White" or "Mixed". Excluding "Do not know", "Prefer not to answer", and "Not available", there were 20 unique values of ethnic background.

| Cluster | COB | Count | Proportion |
|---|---|---|---|
| n/a | England | 18 | 0.51 |
| n/a | Morocco | 3 | 0.09 |
| n/a | Sudan | 3 | 0.09 |
| n/a | Libya | 2 | 0.06 |
| n/a | Wales | 2 | 0.06 |
| 0 | Japan | 242 | 0.84 |
| 0 | South Korea | 26 | 0.09 |
| 1 | Italy | 35 | 0.83 |
| 1 | England | 6 | 0.14 |
| 2 | Finland | 136 | 0.92 |
| 3 | England | 1707 | 0.82 |
| 4 | England | 2418 | 0.76 |
| 4 | Scotland | 181 | 0.06 |
| 4 | USA | 170 | 0.05 |
| 5 | Iran | 502 | 0.31 |
| 5 | Iraq | 303 | 0.19 |
| 5 | England | 169 | 0.10 |
| 5 | Cyprus | 163 | 0.10 |
| 5 | Turkey | 135 | 0.08 |
| 6 | Egypt | 72 | 0.22 |
| 6 | Algeria | 70 | 0.21 |
| 6 | Morocco | 66 | 0.20 |
| 6 | Libya | 37 | 0.11 |
| 7 | India | 3019 | 0.33 |
| 7 | Pakistan | 1344 | 0.15 |
| 7 | Kenya | 1067 | 0.12 |
| 7 | England | 743 | 0.08 |
| 7 | Sri Lanka | 644 | 0.07 |
| 8 | India | 140 | 0.33 |
| 8 | Kenya | 124 | 0.29 |
| 8 | Uganda | 81 | 0.19 |
| 8 | England | 31 | 0.07 |
| 8 | Tanzania | 24 | 0.06 |
| 9 | England | 553 | 0.60 |
| 9 | India | 190 | 0.21 |
| 10 | Somalia | 76 | 0.84 |
| 10 | Prefer not to answer | 7 | 0.08 |
| 11 | England | 50 | 0.58 |
| 11 | Wales | 10 | 0.12 |
| 11 | France | 7 | 0.08 |
| 11 | Egypt | 5 | 0.06 |
| 12 | Peru | 35 | 0.29 |
| 12 | Ecuador | 24 | 0.20 |
| 12 | Mexico | 21 | 0.17 |
| 12 | Bolivia | 13 | 0.11 |
| 12 | Colombia | 13 | 0.11 |
| 13 | Hong Kong | 459 | 0.22 |
| 13 | China | 373 | 0.18 |
| 13 | Philippines | 321 | 0.16 |
| 13 | Malaysia | 314 | 0.15 |
| 14 | England | 194 | 0.49 |
| 14 | Myanmar (Burma) | 24 | 0.06 |
| 14 | Hong Kong | 23 | 0.06 |
| 15 | Nepal | 123 | 0.80 |
| 15 | Prefer not to answer | 11 | 0.07 |
| 16 | Spain | 330 | 0.39 |
| 16 | Portugal | 282 | 0.33 |
| 16 | England | 56 | 0.07 |
| 17 | England | 355844 | 0.82 |
| 17 | Scotland | 37490 | 0.09 |
| 18 | Zimbabwe | 258 | 0.26 |
| 18 | Congo | 144 | 0.14 |
| 18 | Uganda | 126 | 0.13 |
| 18 | Kenya | 111 | 0.11 |
| 18 | Zambia | 56 | 0.06 |
| 18 | South Africa | 53 | 0.05 |
| 19 | Caribbean | 2268 | 0.31 |
| 19 | England | 2077 | 0.28 |
| 19 | Nigeria | 1017 | 0.14 |
| 19 | Ghana | 867 | 0.12 |
| 20 | England | 3528 | 0.87 |
| 21 | England | 8338 | 0.54 |
| 21 | Germany | 970 | 0.06 |
| 21 | Scotland | 938 | 0.06 |
| 22 | England | 697 | 0.66 |
| 22 | Caribbean | 79 | 0.08 |
| 23 | Ethiopia | 57 | 0.33 |
| 23 | Sudan | 50 | 0.29 |
| 23 | Eritrea | 44 | 0.25 |
| 24 | England | 3178 | 0.62 |
| 25 | England | 74 | 0.21 |
| 25 | South Africa | 69 | 0.20 |
| 25 | Mauritius | 63 | 0.18 |
| 25 | Caribbean | 36 | 0.10 |

**Table 4s7:** Frequency of country of birth by cluster for Figure 4.4a. Proportion refers to the proportion within the cluster. Categories with proportion below $0.05$ are not listed.

| Cluster | EB | Count | Proportion |
|---|---|---|---|
| n/a | Mixed, White and Black African | 10 | 0.29 |
| n/a | Mixed, Any other mixed background | 7 | 0.20 |
| n/a | Other ethnic group, Other ethnic group | 6 | 0.17 |
| n/a | White, British | 6 | 0.17 |
| n/a | White, Any other white background | 3 | 0.09 |
| n/a | Black or Black British, African | 2 | 0.06 |
| 0 | Other ethnic group, Other ethnic group | 220 | 0.76 |
| 0 | Asian or Asian British, Any other Asian background | 54 | 0.19 |
| 1 | White, Any other white background | 39 | 0.93 |
| 2 | White, Any other white background | 145 | 0.98 |
| 3 | White, British | 1585 | 0.76 |
| 3 | White, Any other white background | 407 | 0.20 |
| 4 | White, British | 1880 | 0.59 |
| 4 | White, Any other white background | 993 | 0.31 |
| 4 | Other ethnic group, Other ethnic group | 239 | 0.08 |
| 5 | Other ethnic group, Other ethnic group | 751 | 0.46 |
| 5 | White, Any other white background | 435 | 0.27 |
| 5 | Asian or Asian British, Any other Asian background | 229 | 0.14 |
| 5 | White, British | 103 | 0.06 |
| 6 | Other ethnic group, Other ethnic group | 223 | 0.68 |
| 6 | White, Any other white background | 47 | 0.14 |
| 6 | Mixed, White and Black African | 18 | 0.05 |
| 7 | Asian or Asian British, Indian | 5177 | 0.57 |
| 7 | Asian or Asian British, Pakistani | 1726 | 0.19 |
| 7 | Asian or Asian British, Any other Asian background | 1049 | 0.12 |
| 7 | Other ethnic group, Other ethnic group | 498 | 0.06 |
| 8 | Asian or Asian British, Indian | 419 | 0.99 |
| 9 | Mixed, White and Asian | 432 | 0.47 |
| 9 | White, British | 141 | 0.15 |
| 9 | Asian or Asian British, Indian | 84 | 0.09 |
| 9 | Mixed, Any other mixed background | 76 | 0.08 |
| 9 | Other ethnic group, Other ethnic group | 53 | 0.06 |
| 10 | Black or Black British, African | 67 | 0.74 |
| 10 | Other ethnic group, Other ethnic group | 22 | 0.24 |
| 11 | Mixed, Any other mixed background | 30 | 0.35 |
| 11 | White, British | 20 | 0.23 |
| 11 | White, Any other white background | 13 | 0.15 |
| 11 | Mixed, White and Asian | 9 | 0.10 |
| 11 | Other ethnic group, Other ethnic group | 7 | 0.08 |
| 12 | Other ethnic group, Other ethnic group | 90 | 0.74 |
| 12 | Mixed, Any other mixed background | 15 | 0.12 |
| 12 | White, Any other white background | 11 | 0.09 |
| 13 | Chinese, Chinese | 1454 | 0.70 |
| 13 | Other ethnic group, Other ethnic group | 323 | 0.16 |
| 13 | Asian or Asian British, Any other Asian background | 232 | 0.11 |
| 14 | Mixed, Any other mixed background | 114 | 0.29 |
| 14 | Mixed, White and Asian | 93 | 0.23 |
| 14 | Other ethnic group, Other ethnic group | 61 | 0.15 |
| 14 | Asian or Asian British, Any other Asian background | 44 | 0.11 |
| 14 | Chinese, Chinese | 31 | 0.08 |
| 14 | White, British | 31 | 0.08 |
| 15 | Asian or Asian British, Any other Asian background | 63 | 0.41 |
| 15 | Other ethnic group, Other ethnic group | 63 | 0.41 |
| 15 | Not Available | 22 | 0.14 |
| 16 | White, Any other white background | 753 | 0.89 |
| 16 | White, British | 53 | 0.06 |
| 17 | White, British | 412206 | 0.95 |
| 18 | Black or Black British, African | 773 | 0.77 |
| 18 | Other ethnic group, Other ethnic group | 190 | 0.19 |
| 19 | Black or Black British, Caribbean | 4143 | 0.56 |
| 19 | Black or Black British, African | 2225 | 0.30 |
| 19 | Other ethnic group, Other ethnic group | 602 | 0.08 |
| 20 | White, British | 3778 | 0.93 |
| 20 | White, Any other white background | 221 | 0.05 |
| 21 | White, British | 7848 | 0.51 |
| 21 | White, Any other white background | 7020 | 0.45 |
| 22 | Mixed, White and Black Caribbean | 408 | 0.39 |
| 22 | Mixed, White and Black African | 254 | 0.24 |
| 22 | Mixed, Any other mixed background | 111 | 0.11 |
| 22 | Black or Black British, Caribbean | 106 | 0.10 |
| 22 | Other ethnic group, Other ethnic group | 69 | 0.07 |
| 23 | Black or Black British, African | 95 | 0.54 |
| 23 | Other ethnic group, Other ethnic group | 64 | 0.37 |
| 24 | White, British | 3416 | 0.67 |
| 24 | White, Any other white background | 646 | 0.13 |
| 24 | Other ethnic group, Other ethnic group | 285 | 0.06 |
| 25 | Other ethnic group, Other ethnic group | 100 | 0.29 |
| 25 | Mixed, Any other mixed background | 83 | 0.24 |
| 25 | Mixed, White and Black African | 24 | 0.07 |
| 25 | Black or Black British, Caribbean | 23 | 0.07 |

**Table 4s8:** Frequency of selected EB by cluster for Figure 4.4a. Proportions refer to the proportion within the cluster. Categories with proportions below $0.05$ are not listed.

| Phenotype | Model | Caribbean | Indian | African | Any other Asian background | Not available | White | White and Black African | White and Black Caribbean | Bangladeshi |
|---|---|---|---|---|---|---|---|---|---|---|
| FVC | PCA | 0.886 (n=78623) | 1.021 (n=134) | 1.601 (n=320) | 1.215 (n=293) | 0.83 (n=2876) | 1.329 (n=330) | 0.969 (n=2365) | 0.844 (n=197) | 1.081 (n=737) |
| FVC | CLS | 0.891 (n=78623) | 1.029 (n=134) | 1.972 (n=320) | 1.209 (n=293) | 0.874 (n=2876) | 1.32 (n=330) | 0.969 (n=2365) | 0.844 (n=197) | 1.316 (n=737) |
| FEV1 | PCA | 0.913 (n=78623) | 1.092 (n=134) | 1.612 (n=320) | 1.035 (n=293) | 0.854 (n=2876) | 1.21 (n=330) | 1.056 (n=2365) | 0.809 (n=197) | 0.981 (n=737) |
| FEV1 | CLS | 0.917 (n=78623) | 1.085 (n=134) | 1.896 (n=320) | 0.997 (n=293) | 0.884 (n=2876) | 1.173 (n=330) | 1.058 (n=2365) | 0.791 (n=197) | 1.213 (n=737) |
| Standing height | PCA | 0.951 (n=85995) | 1.03 (n=144) | 0.98 (n=350) | 0.831 (n=310) | 0.945 (n=3139) | 0.957 (n=348) | 0.916 (n=2606) | 0.913 (n=214) | 0.93 (n=809) |
| Standing height | CLS | 0.965 (n=85995) | 0.969 (n=144) | 1.059 (n=350) | 0.806 (n=310) | 1.054 (n=3139) | 0.873 (n=348) | 0.922 (n=2606) | 0.972 (n=214) | 1.099 (n=809) |
| BMI | PCA | 0.986 (n=85902) | 0.903 (n=144) | 1.181 (n=347) | 0.683 (n=310) | 0.956 (n=3136) | 1 (n=348) | 0.964 (n=2604) | 1.156 (n=214) | 1.067 (n=805) |
| BMI | CLS | 0.989 (n=85902) | 0.823 (n=144) | 1.18 (n=347) | 0.666 (n=310) | 0.966 (n=3136) | 0.976 (n=348) | 0.964 (n=2604) | 1.197 (n=214) | 1.2 (n=805) |
| Weight | PCA | 0.983 (n=85934) | 1.046 (n=144) | 1.129 (n=347) | 0.752 (n=310) | 0.967 (n=3136) | 0.929 (n=348) | 0.97 (n=2605) | 1.125 (n=214) | 1.122 (n=807) |
| Weight | CLS | 0.984 (n=85934) | 0.943 (n=144) | 1.132 (n=347) | 0.728 (n=310) | 0.977 (n=3136) | 0.9 (n=348) | 0.972 (n=2605) | 1.138 (n=214) | 1.347 (n=807) |
| Leukocyte count | PCA | 0.991 (n=83649) | 0.874 (n=142) | 1.176 (n=333) | 1.13 (n=302) | 0.909 (n=3050) | 0.843 (n=346) | 1.064 (n=2544) | 0.959 (n=210) | 1.021 (n=798) |
| Leukocyte count | CLS | 0.992 (n=83649) | 0.827 (n=142) | 1.124 (n=333) | 1.041 (n=302) | 0.917 (n=3050) | 0.836 (n=346) | 1.063 (n=2544) | 0.871 (n=210) | 1.087 (n=798) |
| Erythrocyte count | PCA | 0.971 (n=83652) | 0.844 (n=142) | 0.969 (n=333) | 1.525 (n=302) | 1.011 (n=3050) | 1.483 (n=346) | 1.004 (n=2544) | 1.113 (n=210) | 1.344 (n=798) |
| Erythrocyte count | CLS | 0.975 (n=83652) | 0.803 (n=142) | 0.98 (n=333) | 1.505 (n=302) | 1.025 (n=3050) | 1.433 (n=346) | 1.004 (n=2544) | 0.957 (n=210) | 1.381 (n=798) |
| Lymphocyte count | PCA | 0.982 (n=83505) | 0.874 (n=142) | 1.073 (n=333) | 1.061 (n=302) | 0.908 (n=3046) | 0.895 (n=345) | 1.042 (n=2539) | 0.763 (n=209) | 0.958 (n=794) |
| Lymphocyte count | CLS | 0.983 (n=83505) | 0.831 (n=142) | 1.081 (n=333) | 0.959 (n=302) | 0.914 (n=3046) | 0.866 (n=345) | 1.039 (n=2539) | 0.778 (n=209) | 0.963 (n=794) |
| Monocyte count | PCA | 0.995 (n=83505) | 1.808 (n=142) | 0.908 (n=333) | 0.908 (n=302) | 0.902 (n=3046) | 1.144 (n=345) | 1.049 (n=2539) | 1.752 (n=209) | 0.895 (n=794) |
| Monocyte count | CLS | 0.995 (n=83505) | 1.848 (n=142) | 0.883 (n=333) | 0.851 (n=302) | 0.905 (n=3046) | 1.137 (n=345) | 1.048 (n=2539) | 1.747 (n=209) | 0.932 (n=794) |
| Neutrophil count | PCA | 0.984 (n=83505) | 0.974 (n=142) | 1.19 (n=333) | 1.188 (n=302) | 0.909 (n=3046) | 0.939 (n=345) | 1.058 (n=2539) | 1.224 (n=209) | 1.102 (n=794) |
| Neutrophil count | CLS | 0.985 (n=83505) | 0.922 (n=142) | 1.178 (n=333) | 1.11 (n=302) | 0.916 (n=3046) | 0.919 (n=345) | 1.053 (n=2539) | 1.127 (n=209) | 1.192 (n=794) |
| Eosinophil count | PCA | 0.982 (n=83505) | 1.249 (n=142) | 0.997 (n=333) | 1.155 (n=302) | 0.892 (n=3046) | 1.992 (n=345) | 0.999 (n=2539) | 1.206 (n=209) | 1.136 (n=794) |
| Eosinophil count | CLS | 0.982 (n=83505) | 1.101 (n=142) | 1.03 (n=333) | 1.111 (n=302) | 0.888 (n=3046) | 1.988 (n=345) | 0.996 (n=2539) | 1.091 (n=209) | 1.126 (n=794) |
| Basophil count | PCA | 0.997 (n=83505) | 0.657 (n=142) | 0.736 (n=333) | 0.957 (n=302) | 0.757 (n=3046) | 1.209 (n=345) | 1.103 (n=2539) | 1.185 (n=209) | 1.122 (n=794) |
| Basophil count | CLS | 0.998 (n=83505) | 0.618 (n=142) | 0.722 (n=333) | 0.956 (n=302) | 0.754 (n=3046) | 1.206 (n=345) | 1.097 (n=2539) | 1.114 (n=209) | 1.101 (n=794) |

**Table 4s9:** Comparing two phenotype models split by EB. One model (PCA) uses the top 40 PCs to estimate phenotypes, while the other (CLS) uses a cluster-smoothed phenotype estimate from Algorithm 1 in addition to the top 40 PCs.

| Phenotype | Model | Caribbean | Indian | African | Any other Asian background | Not available | White | White and Black African | White and Black Caribbean | Bangladeshi |
|---|---|---|---|---|---|---|---|---|---|---|
| FVC | PCA | 1.364 (n=788) | 1.127 (n=1010) | 1.548 (n=589) | 1.106 (n=329) | 1.792 (n=76) | 1.576 (n=91) | 0.962 (n=84) | 1.17 (n=97) | |
| FVC | CLS | 1.365 (n=788) | 1.149 (n=1010) | 1.573 (n=589) | 1.214 (n=329) | 1.943 (n=76) | 1.411 (n=91) | 0.841 (n=84) | 1.16 (n=97) | |
| FEV1 | PCA | 1.071 (n=788) | 1.007 (n=1010) | 1.232 (n=589) | 1.127 (n=329) | 1.419 (n=76) | 1.92 (n=91) | 1.022 (n=84) | 1.044 (n=97) | |
| FEV1 | CLS | 1.085 (n=788) | 1.034 (n=1010) | 1.262 (n=589) | 1.243 (n=329) | 1.917 (n=76) | 1.687 (n=91) | 0.84 (n=84) | 1.051 (n=97) | |
| Standing height | PCA | 1.012 (n=868) | 0.992 (n=1073) | 0.987 (n=642) | 0.942 (n=358) | 1.171 (n=88) | 1.404 (n=103) | 0.83 (n=88) | 1.288 (n=111) | 0.884 (n=52) |
| Standing height | CLS | 1.023 (n=868) | 1.011 (n=1073) | 0.993 (n=642) | 0.943 (n=358) | 1.17 (n=88) | 1.308 (n=103) | 0.872 (n=88) | 1.113 (n=111) | 0.864 (n=52) |
| BMI | PCA | 1.23 (n=867) | 0.898 (n=1071) | 1.057 (n=641) | 0.842 (n=358) | 1.164 (n=86) | 1.212 (n=103) | 1.9 (n=87) | 1.357 (n=110) | 1.258 (n=52) |
| BMI | CLS | 1.204 (n=867) | 0.889 (n=1071) | 1.055 (n=641) | 0.819 (n=358) | 1.061 (n=86) | 1.208 (n=103) | 1.691 (n=87) | 1.213 (n=110) | 0.775 (n=52) |
| Weight | PCA | 1.241 (n=871) | 0.916 (n=1094) | 1.081 (n=642) | 0.828 (n=360) | 1.053 (n=86) | 1.264 (n=103) | 1.799 (n=87) | 1.211 (n=110) | 1.262 (n=52) |
| Weight | CLS | 1.234 (n=871) | 0.922 (n=1094) | 1.108 (n=642) | 0.852 (n=360) | 0.988 (n=86) | 1.233 (n=103) | 1.579 (n=87) | 1.109 (n=110) | 0.823 (n=52) |
| Leukocyte count | PCA | 1.14 (n=828) | 0.892 (n=1061) | 0.926 (n=636) | 0.993 (n=354) | 1.136 (n=84) | 1.262 (n=103) | 1.526 (n=89) | 1.155 (n=107) | 0.893 (n=50) |
| Leukocyte count | CLS | 1.148 (n=828) | 0.884 (n=1061) | 0.914 (n=636) | 0.962 (n=354) | 1.205 (n=84) | 0.919 (n=103) | 1.25 (n=89) | 1.02 (n=107) | 0.596 (n=50) |
| Erythrocyte count | PCA | 1.826 (n=828) | 1.301 (n=1061) | 1.504 (n=636) | 0.952 (n=354) | 1.026 (n=84) | 0.868 (n=103) | 1.355 (n=89) | 1.446 (n=107) | 2.186 (n=50) |
| Erythrocyte count | CLS | 1.81 (n=828) | 1.306 (n=1061) | 1.507 (n=636) | 0.937 (n=354) | 0.955 (n=84) | 0.881 (n=103) | 1.185 (n=89) | 1.217 (n=107) | 1.156 (n=50) |
| Lymphocyte count | PCA | 1.014 (n=826) | 1.138 (n=1060) | 0.93 (n=636) | 1.176 (n=354) | 0.893 (n=84) | 1.048 (n=103) | 1.053 (n=89) | 1.086 (n=107) | 1.356 (n=50) |
| Lymphocyte count | CLS | 1.004 (n=826) | 1.135 (n=1060) | 0.914 (n=636) | 1.168 (n=354) | 0.801 (n=84) | 1.016 (n=103) | 1.041 (n=89) | 1.055 (n=107) | 0.549 (n=50) |
| Monocyte count | PCA | 1.008 (n=826) | 1.069 (n=1060) | 0.944 (n=636) | 1.391 (n=354) | 0.986 (n=84) | 0.984 (n=103) | 1.004 (n=89) | 1.347 (n=107) | 1.817 (n=50) |
| Monocyte count | CLS | 0.993 (n=826) | 1.063 (n=1060) | 0.929 (n=636) | 1.352 (n=354) | 0.922 (n=84) | 0.834 (n=103) | 0.93 (n=89) | 1.24 (n=107) | 1.242 (n=50) |
| Neutrophil count | PCA | 1.223 (n=826) | 0.879 (n=1060) | 1.058 (n=636) | 0.992 (n=354) | 1.179 (n=84) | 1.177 (n=103) | 1.58 (n=89) | 1.147 (n=107) | 0.83 (n=50) |
| Neutrophil count | CLS | 1.232 (n=826) | 0.878 (n=1060) | 1.027 (n=636) | 0.95 (n=354) | 1.27 (n=84) | 0.836 (n=103) | 1.287 (n=89) | 0.988 (n=107) | 0.678 (n=50) |
| Eosinophil count | PCA | 1.277 (n=826) | 1.529 (n=1060) | 1.561 (n=636) | 1.616 (n=354) | 1.355 (n=84) | 0.838 (n=103) | 1.099 (n=89) | 1.131 (n=107) | 2.652 (n=50) |
| Eosinophil count | CLS | 1.291 (n=826) | 1.514 (n=1060) | 1.543 (n=636) | 1.646 (n=354) | 1.412 (n=84) | 0.729 (n=103) | 0.904 (n=89) | 1.078 (n=107) | 2.28 (n=50) |
| Basophil count | PCA | 0.954 (n=826) | 0.766 (n=1060) | 1.607 (n=636) | 0.532 (n=354) | 0.945 (n=84) | 2.296 (n=103) | 0.764 (n=89) | 0.942 (n=107) | 0.436 (n=50) |
| Basophil count | CLS | 0.943 (n=826) | 0.767 (n=1060) | 1.622 (n=636) | 0.505 (n=354) | 0.955 (n=84) | 0.67 (n=103) | 0.675 (n=89) | 0.724 (n=107) | 0.399 (n=50) |

**Table 4s10:** Comparing two phenotype models split by EB. One model (PCA) uses the top 40 PCs to estimate phenotypes, while the other (CLS) uses a cluster-smoothed phenotype estimate from Algorithm 1 in addition to the top 40 PCs.

## References

[1] Wei Zhou et al. “Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease”. In: *Cell Genomics* 2.10 (2022), p. 100192.

[2] Saori Sakaue et al. “Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction”. In: *Nature Communications* 11.1 (2020), p. 1569.

[3] Arslan A Zaidi and Iain Mathieson. “Demographic history mediates the effect of stratification on polygenic scores”. In: *eLife* 9 (2020), e61548.

[4] Alex Diaz-Papkovich et al. “UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts”. In: *PLoS genetics* 15.11 (2019), e1008432.

[5] C J Battey, Gabrielle C Coffing, and Andrew D Kern. “Visualizing population structure with variational autoencoders”. In: *G3* 11.1 (2021), jkaa036.

[6] Susan Holmes and Wolfgang Huber. *Modern Statistics for Modern Biology*. Cambridge University Press, 2019. Chap. Introduction.

[7] Yi Ding et al. “Polygenic scoring accuracy varies across the genetic ancestry continuum”. In: *Nature* (2023), pp. 1–8.

[8] William A Freyman et al. “Fast and Robust Identity-by-Descent Inference with the Templated Positional Burrows–Wheeler Transform”. In: *Molecular Biology and Evolution* 38.5 (2021), pp. 2131–2151.

[9] Ruhollah Shemirani et al. “Rapid detection of identity-by-descent tracts for mega-scale datasets”. In: *Nature Communications* 12.1 (2021), p. 3546.

[10] Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*. Washington, D.C.: National Academies Press, 2023.

[11] Kristjan E. Kaseniit et al. “Genetic ancestry analysis on >93,000 individuals undergoing expanded carrier screening reveals limitations of ethnicity-based medical guidelines”. In: *Genetics in Medicine* 22.10 (2020), pp. 1694–1702.

[12] Bjarni V. Halldorsson et al. “The sequences of 150,119 genomes in the UK Biobank”. In: *Nature* 607.7920 (2022), pp. 732–740.

[13] Alicia R. Martin et al. “Clinical use of current polygenic risk scores may exacerbate health disparities”. In: *Nature Genetics* 51.4 (2019), pp. 584–591.

[14] Chief Ben-Eghan et al. “Don’t ignore genetic data from minority populations”. In: *Nature* 585.7824 (2020), pp. 184–186.

[15] Daphne O. Martschenko et al. “Including multiracial individuals is crucial for race, ethnicity and ancestry frameworks in genetics and genomics”. In: *Nature Genetics* 55.6 (2023), pp. 895–900.

[16] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020.

[17] Claudia Malzer and Marcus Baum. "A Hybrid Approach To Hierarchical Density-based Cluster Selection". In: *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. 2020, pp. 223–228.

[18] The 1000 Genomes Project Consortium. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74.

[19] Cathie Sudlow et al. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age". In: *PLOS Medicine* 12.3 (2015), e1001779.

[20] Philip Awadalla et al. "Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics". In: *International Journal of Epidemiology* 42.5 (2013), pp. 1285–1299.

[21] Adriana Romero et al. "Diet Networks: Thin Parameters for Fat Genomics". In: *arXiv* (2017).

[22] David Reich et al. "Reconstructing Indian population history". In: *Nature* 461.7263 (2009), pp. 489–494.

[23] Abdel Abdellaoui et al. "Genetic correlates of social stratification in Great Britain". In: *Nature Human Behaviour* 3.12 (2019), pp. 1332–1342.

[24] Edmund Gilbert, Ashwini Shanmugam, and Gianpiero L. Cavalleri. "Revealing the recent demographic history of Europe via haplotype sharing in the UK Biobank". In: *Proceedings of the National Academy of Sciences* 119.25 (2022), e2119281119.

[25] Alec M. Chiu et al. "Inferring population structure in biobank-scale genomic data". In: *The American Journal of Human Genetics* 109.4 (2022), pp. 727–737.

[26] Shashwat Deepali Nagar et al. "Socioeconomic deprivation and genetic ancestry interact to modify type 2 diabetes ethnic disparities in the United Kingdom". In: *eClinicalMedicine* 37 (2021), p. 100960.

[27] Slave Voyages: The Trans-Atlantic Slave Trade Database. *Trans-Atlantic Slave Trade - Estimates*. `http://www.slavevoyages.org/estimates/BeZD1wTh`. 2023.

[28] Cesar Fortes-Lima and Paul Verdu. "Anthropological genetics perspectives on the transatlantic slave trade". In: *Human Molecular Genetics* 30.R1 (2021), R79–R87.

[29] Anna C. F. Lewis et al. "Getting genetic ancestry right for science and society". In: *Science* 376.6590 (2022), pp. 250–252.

[30] Abram B. Kamiza et al. "Transferability of genetic risk scores in African populations". In: *Nature Medicine* 28.6 (2022), pp. 1163–1166.

[31] Shadi Zabad, Simon Gravel, and Yue Li. "Fast and accurate Bayesian polygenic risk modeling with variational inference". In: *The American Journal of Human Genetics* (2023).

[32] Ying Wang et al. "Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations". In: *Nature Communications* 11.1 (2020), p. 3865.

[33] Anna M. Bennet et al. "Pleiotropy in the Presence of Allelic Heterogeneity: Alternative Genetic Models for the Influence of APOE on Serum LDL, CSF Amyloid-$\beta$ 42, and Dementia". In: *Journal of Alzheimer's Disease* 22.1 (2010), pp. 129–134.

[34] Laura J. Rasmussen-Torvik et al. "High Density GWAS for LDL Cholesterol in African Americans Using Electronic Medical Records Reveals a Strong Protective Variant in APOE". In: *Clinical and Translational Science* 5.5 (2012), pp. 394–399.

[35] Serena Sanna et al. "Fine Mapping of Five Loci Associated with Low-Density Lipoprotein Cholesterol Detects Variants That Double the Explained Heritability". In: *PLOS Genetics* 7.7 (2011), e1002198.

[36] Clara Breitling et al. "Genetic Contribution of Variants near SORT1 and APOE on LDL Cholesterol Independent of Obesity in Children". In: *PLOS ONE* 10.9 (2015), e0138064.

[37] Hui Wang et al. "Associations of genetic variants of lysophosphatidylcholine metabolic enzymes with levels of serum lipids". In: *Pediatric Research* 91.6 (2022), pp. 1595–1599.

[38] A. K. MacLeod et al. "Some principles and practices of genetic biobanking studies". In: *European Respiratory Journal* 33.2 (2009), pp. 419–425.

[39] Alex Diaz-Papkovich, Luke Anderson-Trocmé, and Simon Gravel. "A review of UMAP in population genetics". In: *Journal of Human Genetics* 66.1 (2021), pp. 85–91.

[40] Yu Qian, Brian L. Browning, and Sharon R. Browning. "Efficient clustering of identity-by-descent between multiple individuals". In: *Bioinformatics* 30.7 (2014), pp. 915–922.

[41] Daniel John Lawson et al. "Inference of Population Structure using Dense Haplotype Data". In: *PLOS Genetics* 8.1 (2012), e1002453.

[42] Eunjung Han et al. "Clustering of 770,000 genomes reveals post-colonial population structure of North America". In: *Nature Communications* 8.1 (2017), p. 14238.

[43] Juba Nait Saada et al. "Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations". In: *Nature Communications* 11.1 (2020), p. 6130.

[44] Christa Caggiano et al. "Disease risk and healthcare utilization among ancestrally diverse groups in the Los Angeles region". In: *Nature Medicine* 29.7 (2023), pp. 1845–1856.

[45] Gillian M. Belbin et al. "Toward a fine-scale population health monitoring system". In: *Cell* 184.8 (Apr. 2021), 2068–2083.e11.

[46] Haley Hunter-Zinck et al. “Genotyping Array Design and Data Quality Control in the Million Veteran Program”. In: *The American Journal of Human Genetics* 106.4 (2020), pp. 535–548.

[47] Trevor J. B. Dummer et al. “The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for research on chronic disease prevention”. In: *CMAJ* 190.23 (2018), E710–E717.

[48] Genevieve L. Wojcik et al. “Genetic analyses of diverse populations improves discovery for complex traits”. In: *Nature* 570.7762 (2019), pp. 514–518.

[49] Meng Lin et al. “Admixed Populations Improve Power for Variant Discovery and Portability in Genome-Wide Association Studies”. In: *Frontiers in Genetics* 12 (2021).

[50] Huaying Fang et al. “Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies”. In: *The American Journal of Human Genetics* 105.4 (2019), pp. 763–772.

[51] Konrad J. Karczewski et al. “The mutational constraint spectrum quantified from variation in 141,456 humans”. In: *Nature* 581.7809 (2020), pp. 434–443.

[52] Trevor Smith and Scott McLeish. *Technical report on changes in response related to the census ethnic origin question: Focus on Jewish origins, 2016 Census integrated with 2011 National Household Survey*. 2019.

[53] Jedidiah Carlson et al. “Counter the weaponization of genetics research by extremists”. In: *Nature* 610.7932 (2022), pp. 444–447.

[54] Wendy D. Roth. “The multiple dimensions of race”. In: *Ethnic and Racial Studies* 39.8 (2016), pp. 1310–1338.

[55] Shai Ben-David. “Clustering - What Both Theoreticians and Practitioners are Doing Wrong”. In: *arXiv* (2018).

[56] Christian Hennig. “What are the true clusters?” In: *arXiv:1502.02555 [stat]* (2015).

[57] Daniel J. Lawson, Lucy van Dorp, and Daniel Falush. “A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots”. In: *Nature Communications* 9.1 (2018), p. 3258.

[58] Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. “Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms”. In: *New England Journal of Medicine* 383.9 (2020), pp. 874–882.

[59] Ananyo Choudhury et al. “High-depth African genomes inform human migration and health”. In: *Nature* 586.7831 (2020), pp. 741–748.

[60] Shaun Purcell et al. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses”. In: *The American Journal of Human Genetics* 81.3 (2007), pp. 559–575.

[61] David H. Alexander, John Novembre, and Kenneth Lange. “Fast model-based estimation of ancestry in unrelated individuals”. In: *Genome Research* 19.9 (2009), pp. 1655–1664.

[62] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018.

[63] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[64] Marek Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia, 2018.

# Part III

# Summary

*A map **is not** the territory it represents, but, if correct, it has a **similar structure** to the territory, which accounts for its usefulness.*
*—Alfred Korzybski (1933)*

# Chapter 5: Discussion

We have presented a method of nonlinear dimensionality reduction that is compatible with the scale and composition of modern biobanks. With UMAP and HDBSCAN($\hat{\epsilon}$) we have uncovered a wide variety of patterns of fine-scale population structure in every biobank studied. Since the publication of Chapter 2 , UMAP has become a standard method for visualization and exploratory data analysis in population genetics. Given the demand for cluster extraction noted in Chapters 3 and 4, it is possible that HDBSCAN($\hat{\epsilon}$)—or some other density clustering method—could see similar adoption.

The timing of the development UMAP coincides fortuitously with the growth of biobanks and the genomic revolution. Writing in his review of TDA in 2018, Wasserman asked, "is it possible to derive low dimensional embedding methods that explicitly preserve topological features of the data? This is an interesting open question."[42] This review was published in the same year that UMAP, an explicitly topological method, was released. Topological analysis and density clustering seem especially well-suited to the task of understanding population structure in diverse biobanks.

## 5.1 The value of visualization

One may ask: why visualize data at all? Are these figures simply decorations for manuscripts? Though analyses, inferences, and models can shed light on the biological world, there is an inherent value to being able to literally see your data and understand its structure. In Chapter 3 we

describe the relationship between dimensionality reduction and biological data as analogous to that of microscopes and biological samples. This analogy has been made before with respect to mathematics in general[66]. Seeing our data can compel us to study it more deeply and to crystallize theoretical concepts. In population genetics, one of the most cited papers is "Genes mirror geography within Europe"[40], which visualizes the relationship between PCA and isolation-by-distance. Although isolation-by-distance had been characterized almost a century earlier, the figure of the top principal components superimposed over Europe was an elegant illustration of the geographical distribution of human genetic variation.

With UMAP, we are able to map complex genetic structure down to 2 or 3 dimensions for visualization, examine the fine-scale structure that a method like PCA compresses, while preserving interesting signal in the data. Using auxiliary data—geographic data, population labels, phenotypic information, environmental variables, etc.—we can visually scan for patterns and generate hypotheses. In Chapter 2 we present a variety of visualization methods, including colouring points by admixture levels to uncover gradients in ancestry and converting 3D UMAP coordinates to RGB colour levels for colouring maps. One common method used in single-cell studies is to colour UMAP plots by gene expression levels (e.g. [67] Figure 4h). A similar approach could be used in exploring allele distributions by clusters to see whether they are relatively more common in different populations—one such method, called the Allele Dispersion Score, was proposed by Correard et al. in 2022[68]. Though UMAP figures are now standard in genetic studies, most are limited to a simple 2D scatterplot coloured by some population label—in every biobank, there is

certainly value in exploring deeper and using other variables.

There is also value in going beyond static 2D figures. Software libraries such as plotly[69] can generate interactive figures, are available for Python and R, and are straightforward to implement. Interactive exploration can assess how individual points fit within larger projections, rapidly identifying patterns or using them for diagnostics. Working in 3D is much easier when the figures are interactive, and we have, e.g., explored the relationships between individuals and populations in the UKB. Finally, as noted in Chapter 4, clustering works well in dimensions above 2—while it can be tempting to delineate clusters from 2D figures alone, the lower the dimensionality of our representation, the less faithful it is to the data. Despite UMAP's now-widespread use, its ability to work in 3 or more dimensions is, in my view, underappreciated.

## 5.2 Implications for biomedical and epidemiological research

In inferential statistics, we require some definition of a "population". To establish the scope of study we define a target population (the group to whom the research applies) and the study population (the group which is covered by the study and is, ideally, drawn from the target population as illustrated in Figure 5.1a)[70]. Given the results in Chapter 4, density clustering with UMAP could potentially be used to define populations for study. It has several advantages over methods based on, e.g., $K$-means clustering: it clusters all or almost all individuals, including those with recent admixture or from uncommon ancestries; it is does not assume a number of populations $K$; it does not require reference panels; its assumptions of the structure of genetic data are more real-

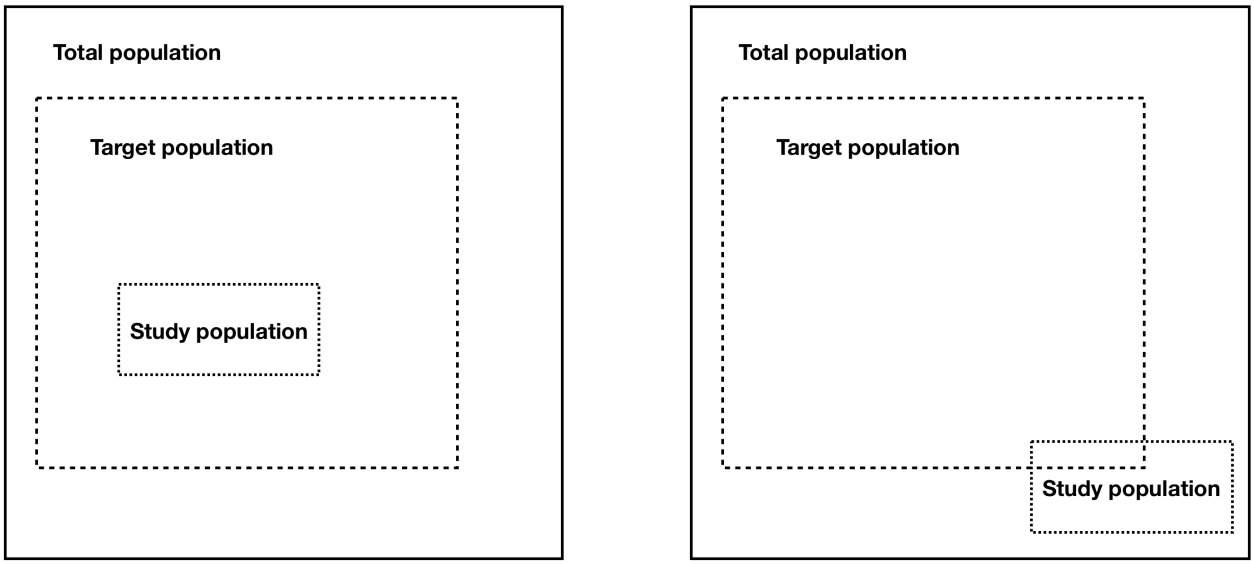



**Figure 5.1: The study population should match the target population.** In **(a)**, the study population is a subset of the target population, so inferences drawn from it will apply to the target population. In **(b)**, the study population is not fully a subset of the target population, so its inferences will be less reliable.

istic. It also has advantages over variables like self-identification or country of birth that, though sometimes used as proxies for population structure, are not actually genetic.

Unlike methods that try to, e.g., harmonize labels with genetic data, the density clustering approach is label-agnostic. This is beneficial when there is differential response in identification, as we saw in Chapter 4 with many individuals from non-majority populations (in this case, non-European individuals) simply identifying as "other". There is currently a significant bias in genomic studies, with a tendency to discard data from non-European populations, often citing concerns over low sample sizes or issues with population structure[71]. Approximately $86\%$ of genomic studies by 2021 were based on individuals of European ancestry, an increase from $81\%$

in 2016[30, 72]. Using UMAP and density clustering presents an avenue to boost sample sizes and include a more diverse array of populations in studies by not depending on clustering that is, essentially, based on population labels.

Biobanks are as complex as they are rich. Across visualizations in Chapters 2 and 4, we see patterns. Some appear quickly at the population level, as in Figure 2.6 where we see population-level differences in height and blood cell count; others require iterative smoothing, such as Figure 4.5 where we see residual structure in phenotypes after removing the effects of the top $40$ PCs. These patterns also appear in behavioural measures, such as smoking in Figure 4s4. Figures 2s17 and 2s18 show how geographic structure is closely tied to genetic structure. Using variables like admixture proportions, country of birth, immigration history, environmental data, etc., we can see the difficulty of disentangling genetic and environmental effects—structure and potential confounding are omnipresent.

Biobanks are generally not random samples of a population and suffer from selection bias[73]; there are genetic biases in enrolment itself[74, 75]. Though the target population for a study may be, e.g., all European-ancestry individuals, if there is significant bias in the composition of a biobank, the study population will not be aligned with the target population and the inference may not be reliable. As such, though we have a prospective methodology to define populations, we must remember that our data are not necessarily representative of the population at large.

If we do not understand the gene-environment interplay, we may also incorrectly credit genetic architecture with a phenotype or disease that is actually caused by environmental factors. In a

well-known paper in the 1980s[76], epidemiologist Geoffrey Rose posed a thought experiment: what if everyone smoked 20 cigarettes per day? Taking a purely genetic approach, we would (wrongly) deduce that lung cancer is a genetic disease. Similarly, there are non-genetic factors, such as pollution or discriminatory policies, that may disproportionately or exclusively impact certain groups. Such factors may not be measured at all within biobanks. It is therefore crucial to understand as much as possible the broader context of a population's environment.

## 5.3 Clustering in population genetics

As we have seen throughout this work (particularly Chapter 4), clusters are useful abstractions. The structure they find is "real" in the sense that it is a discrete measure of relatedness between individuals, with the caveat that the structure identified is conditioned on the data and methods. When we require a genetic definition of a population, such as to calculate a MAF or train/test a PGS or to explore biobank data efficiently, the approach of using UMAP and HDBSCAN($\hat{\epsilon}$) shows great promise.

Unlike problems like classification, clustering does not have a well-defined ground truth, and even the most basic definition, such as "similar points are grouped together and dissimilar points are grouped separately", can be self-contradictory[50]. In a genetic cluster, individuals may be closely related to their immediate neighbours, but not to those in a far end of a cluster; they are grouped by similarity in one sense, but not separated by dissimilarity. Though various quality metrics exist, even untrained humans excel at assessing 2D clusterings[77]—there is an element of

"I know it when I see it", and since clustering is a data-driven computational method, it carries an air of objectivity.


**Figure 5.2: A cluster is only an abstraction, albeit a useful one.** A cluster can represent a population, but it is not a population itself.

However, it is erroneous to describe clusters as "objective" measures of human population structure, both for the aforementioned reasons, and also because clustering changes based on many subjective inputs: algorithms, hyperparameters, subsets of data, etc.[78]. This myth of objectivity in machine learning has perpetuated systemic biases because it spares methods from critical examination[79]. The cost of epistemological ignorance here varies, but at its most extreme, it has led to justification of racism and violence[80–82]. Dimensionality reduction, visualization, and clustering are powerful tools, and we must understand not only their strengths and limitations, but their uses and misuses in the population genetics.

Population genetic clusters are a discrete categorical measure of genetic ancestry, which is a continuous and multidimensional variable. Clusters that are clearly separated in one sample will ultimately fall into a continuum among larger and more diverse samples[83]. We have seen through this thesis that clusters can split, merge, reappear, or disappear; in the same vein, in microscopy objects may appear or disappear depending on the depth and focus and type of microscope used. It would be folly to look at, e.g., a static 2D slice of a single cell and its sub-cellular components at one single time point and to conclude that this image defines the "real" cell. It is, however, a useful representation of the structure of the cell—provided we keep in mind that everything we see is conditioned on the sample (e.g. the cell type and state, the health of the organism, the environment in which the cell is viewed, etc).

Data sets do not spring forth naturally, nor do they examine themselves; they are sampled, collected, processed, analysed, and are products of societies and institutions situated within their own contexts and with their own biases[84]. Population labels are an excellent example, since we often use them as shorthand for clusters and populations. In the UKB, there are 20 unique values for "ethnic background" (see Table 4s6) and they are selected in a nested manner (first you select from a list of "ethnic groups", then from a second list of "ethnic backgrounds"). It is not possible to identify as, for example, "Chinese British" or "Mixed Black and Asian", though there are likely individuals who would select those labels. The term "Asian" itself also has multiple meanings depending on context, including: a cultural label for South Asians (in British English), a cultural label for East Asians (North American English), or a label of geographic origin. These details shape how

we perceive populations in biobanks, how we talk about them, and—eventually—the perceptions of those who consume our research. With genetic data, we are cartographers of unknown territory.

## 5.4 Limitations

Dimensionality reduction necessarily sacrifices information. To help us navigate the physical world, we use two-dimensional Euclidean maps of a three-dimensional spherical planet. This transformation introduces artificial tearing and distortion into world maps—PCA, $t$-SNE, UMAP, and density clustering are no different, although with maps there is a more intuitive understanding of how we distort information. After transforming data, we may see patterns or clusters that are artefactual, arising from unidentified batch effects or the distribution of genetic variants particular to a biobank. The relationships between visual objects may be misleading. Specific combinations of parameters may generate unusual results because of something particular to our data. Patterns may also form from noise. It is crucial to carry out multiple runs of dimensionality reduction, to examine any results with a sceptical eye, and to test hypotheses with confirmatory analyses.

UMAP works within a particular paradigm: it represents topology by approximating the local structure of the data. By patching together many local topological representations, it achieves a semblance of the global structure of data; however, the distances themselves do not have meaningful interpretations. PCA, which explicitly models global variation in data, has been interpreted in connection to TMRCA and F-statistics[11, 85]. Several other nonlinear dimensionality reduction methods aim to balance global and local structure (e.g. PHATE[86], POPVAE[87]). Since UMAP

captures topology and not geometry, it approximates the shapes of data, but not how the shapes are positioned relative to one another. Consider three nested hollow spheres in 3D: each sphere forms one connected shape, but a 2D UMAP would represent them as three disconnected clusters[88]. This correctly identifies the three connected components but not their relationships to each other. Thus it is important to use UMAP in combination with other methods, providing snapshots of our data from many angles and resolutions.

## 5.5 Future directions

Topological data analysis has proven a useful addition to the population geneticist's toolbox. The work done in this thesis has been on genotypic data, and usually using the top principal components for computational reasons. Since the publication of the manuscripts, multi-threaded versions of UMAP and HDBSCAN($\hat{\epsilon}$) have been released. One avenue for future research is to work directly with genotype data that has not been processed with PCA. We explored this briefly in Chapter 2 using 1KGP data; the improvement in compute time with multi-threading should make deeper investigation more feasible.

There are possible connections between genetic topology and IBD. As noted in Chapter 4, UMAP and HDBSCAN($\hat{\epsilon}$) extract population structure that appears similar in scale to IBD studies of diverse biobanks; the two methods may capture similar information via different routes. Clusters extracted by our approach likely contain substructure themselves, and an approach that breaks clusters down further may reveal finer details.

Studying different subsets of populations or genetic data may be illuminating. We have noted studies in Chapter 3 that use, e.g., structural variants rather than genotype data. Sex-bias in admixture has been noted in other studies (e.g. [89–91]), whereas all of the analyses presented here have been on autosomal data and combining the sexes. Each of these approaches seems likely to uncover interesting patterns.

Within the biomedical realm, the confounding of GWAS and transferability of PGS are perennial areas of research. We presented a method of visualization in Chapter 4 that illustrates how PCA adjustment affects populations across an entire biobank, with an additional analysis that studied the transferability of PGS. These approaches could also extend to studying the interplay between environment, genetics, and biomedical traits.

# Chapter 6: Conclusion

This thesis has introduced a new methodology for high-dimensional population genetic data, applying UMAP and HDBSCAN($\hat{\epsilon}$) to several biobanks. In Chapter 2 we applied UMAP to population genetic data for the first time, using it for visualization and revealing fine-scale population structure on data sets of up to half a million individuals. We illustrated its potential for exploratory analyses with examples of many applications; it is now a standard method in the field. In Chapter 3, we review the uses of UMAP in population genetics, discuss the impacts of data filtering and parametrization, and provide guidance on best practices. Finally, in Chapter 4, we provide a method to reliably extract clusters in UMAP data. We use these methods to stratify biobank data, characterize population structure, carry out quality control, study the impacts of PCA correction in GWAS and PGS and the transferability of PGS between different populations. In each chapter, we expound upon the relationships between UMAP data, clusters, and auxiliary data such as population labels, distributions of phenotypes, the geographic distribution of genetic variation, and demographic history. In each case, we have also made our code and data publicly available.

In conclusion, this thesis has established a basis for deep exploration of biobank data using topological analysis. It is a methodology that is fast, tractable, creates rich visualizations that provoke investigation, and has potential for many follow-up studies.

# Master reference list


[1] Richard A. Gibbs. “The Human Genome Project changed everything”. In: *Nature Reviews Genetics* 21.10 (2020), pp. 575–576.

[2] Mark Jobling, Matthew Hurles, and Chris Tyler-Smith. *Human Evolutionary Genetics: Origins, Peoples & Disease*. 2013.

[3] Nicolas Altemose et al. “A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis”. In: *eLife* 6 (2017).

[4] Laure Ségurel, Minyoung J. Wyman, and Molly Przeworski. “Determinants of Mutation Rate Variation in the Human Germline”. In: *Annual Review of Genomics and Human Genetics* 15.1 (2014), pp. 47–70.

[5] Daniel L Hartl and Andrew G Clark. *Principles of population genetics*. Vol. 116. Sinauer associates Sunderland, MA, 2007. Chap. 6.

[6] Sewall Wright. “Isolation by Distance”. In: *Genetics* 28.2 (1943), pp. 114–138.

[7] JF Crow and M Kimura. *An Introduction to Population Genetics Theory*. Harper & Row, NY, 1970. Chap. 3.

[8] JF Crow and M Kimura. *An Introduction to Population Genetics Theory*. Harper & Row, NY, 1970. Chap. 9.

[9] G Barbujani and R R Sokal. “Zones of sharp genetic change in Europe are also linguistic boundaries.” In: *Proceedings of the National Academy of Sciences* 87.5 (1990), pp. 1816–1819.

[10] Michael C. Campbell and Sarah A. Tishkoff. “The Evolution of Human Genetic and Phenotypic Variation in Africa”. In: *Current Biology* 20.4 (2010), R166–R173.

[11] Benjamin M. Peter. “A geometric relationship of F2, F3 and F4-statistics with principal component analysis”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1852 (2022), p. 20200413.

[12] Aaron P. Ragsdale et al. “A weakly structured stem for human origins in Africa”. In: *Nature* 617.7962 (2023), pp. 755–763.

[13] Daniel L Hartl and Andrew G Clark. *Principles of population genetics*. Vol. 116. Sinauer associates Sunderland, MA, 2007. Chap. 2.

[14] Gaurav Bhatia et al. “Estimating and interpreting FST: The impact of rare variants”. In: *Genome Research* 23.9 (2013), pp. 1514–1521.

[15] Ilana M. Arbisser and Noah A. Rosenberg. “FST and the triangle inequality for biallelic markers”. In: *Theoretical Population Biology*. Fifty years of Theoretical Population Biology 133 (2020), pp. 117–129.

[16] Daniel L Hartl and Andrew G Clark. *Principles of population genetics*. Vol. 116. Sinauer associates Sunderland, MA, 2007. Chap. 3.

[17] Jerome Kelleher et al. “Inferring whole-genome histories in large population datasets”. In: *Nature genetics* 51.9 (2019), pp. 1330–1338.

[18] Débora YC Brandt et al. “Evaluation of methods for estimating coalescence times using ancestral recombination graphs”. In: *Genetics* 221.1 (2022), iyac044.

[19] Brian C Zhang et al. “Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits”. In: *Nature Genetics* (2023), pp. 1–9.

[20] Emil Uffelmann et al. “Genome-wide association studies”. In: *Nature Reviews Methods Primers* 1.1 (2021), pp. 1–21.

[21] Po-Ru Loh et al. “Mixed-model association for biobank-scale datasets”. In: *Nature Genetics* 50.7 (2018), pp. 906–908.

[22] Wei Zhou et al. “SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests”. In: *Nature Genetics* 54.10 (2022), pp. 1466–1469.

[23] Joelle Mbatchou et al. “Computationally efficient whole-genome regression for quantitative and binary traits”. In: *Nature Genetics* 53.7 (2021), pp. 1097–1103.

[24] Ying Wang et al. “Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations”. In: *Nature Communications* 11.1 (2020), p. 3865.

[25] Alkes L. Price et al. “Principal components analysis corrects for stratification in genome-wide association studies”. In: *Nature Genetics* 38.8 (2006), p. 904.

[26] Arslan A Zaidi and Iain Mathieson. “Demographic history mediates the effect of stratification on polygenic scores”. In: *eLife* 9 (2020), e61548.

[27] Jeremy J Berg et al. “Reduced signal for polygenic adaptation of height in UK Biobank”. In: *eLife* 8 (2019), e39725.

[28] Mashaal Sohail et al. “Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies”. In: *eLife* 8 (2019), e39702.

[29] Sini Kerminen et al. “Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland”. In: *The American Journal of Human Genetics* 104.6 (2019), pp. 1169–1181.

[30] Jonathan Michael Kaplan and Stephanie M. Fullerton. “Polygenic risk, population structure and ongoing difficulties with race in human genetics”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1852 (2022), p. 20200427.

[31] Wei Zhou et al. “Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease”. In: *Cell Genomics* 2.10 (2022), p. 100192.

[32] The 1000 Genomes Project Consortium. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), pp. 68–74.

[33] Cathie Sudlow et al. “UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age”. In: *PLOS Medicine* 12.3 (2015), e1001779.

[34] Philip Awadalla et al. “Cohort profile of the CARTaGENE study: Quebec’s population-based biobank for public health and personalized genomics”. In: *International Journal of Epidemiology* 42.5 (2013), pp. 1285–1299.

[35] Susan Holmes and Wolfgang Huber. *Modern Statistics for Modern Biology*. Cambridge University Press, 2019. Chap. Introduction.

[36] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley Publishing Company, 1977.

[37] David C. Hoaglin. “John W. Tukey and Data Analysis”. In: *Statistical Science* 18.3 (2003), pp. 311–318.

[38] John W. Tukey. “We Need Both Exploratory and Confirmatory”. In: *The American Statistician* 34.1 (1980), pp. 23–25.

[39] Gil McVean. “A Genealogical Interpretation of Principal Components Analysis”. In: *PLOS Genetics* 5.10 (2009), e1000686.

[40] John Novembre et al. “Genes mirror geography within Europe”. In: *Nature* 456 (2008), pp. 98–101.

[41] Gunnar Carlsson. “Topology and data”. In: *Bulletin of the American Mathematical Society* 46.2 (2009), pp. 255–308.

[42] Larry Wasserman. “Topological Data Analysis”. In: *Annual Review of Statistics and Its Application* 5.1 (2018), pp. 501–532.

[44] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605.

[45] Miguel Á. Carreira-Perpiñan. “The elastic embedding algorithm for dimensionality reduction”. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML 10. 2010, pp. 167–174.

[47] Wentian Li et al. “Application of t-SNE to human genetic data”. In: *Journal of Bioinformatics and Computational Biology* 15.04 (2017), p. 1750017.

[48] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020.

[49] *How UMAP Works — umap 0.5 documentation*.

[50] Shai Ben-David. “Clustering - What Both Theoreticians and Practitioners are Doing Wrong”. In: *arXiv* (2018).

[51] G. Evanno, S. Regnaut, and J. Goudet. “Detecting the number of clusters of individuals using the software structure: a simulation study”. In: *Molecular Ecology* 14.8 (2005), pp. 2611–2620.

[52] Robert Verity and Richard A. Nichols. “Estimating the Number of Subpopulations (K) in Structured Populations”. In: *Genetics* 203.4 (2016), pp. 1827–1839.

[53] Daniel J. Lawson, Lucy van Dorp, and Daniel Falush. “A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots”. In: *Nature Communications* 9.1 (2018), p. 3258.

[54] Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. “Inference of Population Structure Using Multilocus Genotype Data”. In: *Genetics* 155.2 (2000), pp. 945–959.

[55] David H. Alexander, John Novembre, and Kenneth Lange. “Fast model-based estimation of ancestry in unrelated individuals”. In: *Genome Research* 19.9 (2009), pp. 1655–1664.

[56] Hua Tang et al. “Estimation of individual admixture: Analytical and study design considerations”. In: *Genetic Epidemiology* 28.4 (2005), pp. 289–301.

[57] Eric Frichot et al. “Fast and Efficient Estimation of Individual Ancestry Coefficients”. In: *Genetics* 196.4 (2014), pp. 973–983.

[58] Julia Gimbernat-Mayol et al. “Archetypal Analysis for population genetics”. In: *PLOS Computational Biology* 18.8 (2022), e1010301.

[59] J. A. Hartigan and M. A. Wong. “Algorithm AS 136: A K-Means Clustering Algorithm”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), pp. 100–108.

[60] Leland McInnes and John Healy. “Accelerated Hierarchical Density Clustering”. In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 2017, pp. 33–42.

[61] Yi Ding et al. “Polygenic scoring accuracy varies across the genetic ancestry continuum”. In: *Nature* (2023), pp. 1–8.

[62] Martin Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. Vol. 96. 34. 1996, pp. 226–231.

[63] Kamran Khan et al. “DBSCAN: Past, present and future”. In: *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*. 2014, pp. 232–238.

[64] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. “Density-Based Clustering Based on Hierarchical Density Estimates”. In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Jian Pei et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 160–172.

[65] Claudia Malzer and Marcus Baum. “A Hybrid Approach To Hierarchical Density-based Cluster Selection”. In: *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. 2020, pp. 223–228.

[66] Joel E. Cohen. “Mathematics Is Biology’s Next Microscope, Only Better; Biology Is Mathematics’ Next Physics, Only Better”. In: *PLOS Biology* 2.12 (2004), e439.

[67] Selin Jessa et al. “Stalled developmental programs at the root of pediatric brain tumors”. In: *Nature Genetics* 51.12 (2019), pp. 1702–1713.

[68] Solenne Correard, Laura Arbour, and Wyeth W. Wasserman. *Allele Dispersion Score: Quantifying the range of allele frequencies across populations, based on UMAP*. 2022.

[69] Plotly Technologies Inc. *Collaborative data science*. Montreal, QC, 2015.

[70] Sarah Franklin and Charlene Walker, eds. *Survey methods and practices*. Statistics Canada, 2003.

[71] Chief Ben-Eghan et al. “Don’t ignore genetic data from minority populations”. In: *Nature* 585.7824 (2020), pp. 184–186.

[72] Segun Fatumo et al. “A roadmap to increase diversity in genomic studies”. In: *Nature Medicine* 28.2 (2022), pp. 243–250.

[73] Jonathan Yinhao Huang. “Representativeness Is Not Representative: Addressing Major Inferential Threats in the UK Biobank and Other Big Data Repositories”. In: *Epidemiology* 32.2 (2021), p. 189.

[74] Nicola Pirastu et al. “Genetic analyses identify widespread sex-differential participation bias”. In: *Nature Genetics* 53.5 (2021), pp. 663–671.

[75] Stefania Benonisdottir and Augustine Kong. “Studying the genetics of participation using footprints left on the ascertained genotypes”. In: *Nature Genetics* (2023), pp. 1–8.

[77] Joshua M Lewis, Margareta Ackerman, and Virginia R de Sa. “Human Cluster Evaluation and Formal Quality Measures: A Comparative Study”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 34 (2012), pp. 1870–1875.

[78] David S. Watson. “On the Philosophy of Unsupervised Learning”. In: *Philosophy & Technology* 36.2 (2023), p. 28.

[79] Timnit Gebru. “Race and Gender: Data-Driven Claims about Race and Gender Perpetuate the Negative Biases of the Day”. In: *The Oxford Handbook of Ethics of AI*. Ed. by Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford University Press, 2020. Chap. 13, pp. 253–269.

[80] Aaron Panofsky and Joan Donovan. “Genetic ancestry testing among white nationalists: From identity repair to citizen science”. In: *Social Studies of Science* 49.5 (2019), pp. 653–681.

[81] Aaron Panofsky, Kushan Dasgupta, and Nicole Iturriaga. “How White nationalists mobilize genetics: From genetic ancestry and human biodiversity to counterscience and metapolitics”. In: *American Journal of Physical Anthropology* 175.2 (2021), pp. 387–398.

[83] Anna C. F. Lewis et al. “Getting genetic ancestry right for science and society”. In: *Science* 376.6590 (2022), pp. 250–252.

[84] Catherine D’Ignazio and Lauren F. Klein. *Data Feminism*. Cambridge, Massachusetts: The MIT Press, 2020. Chap. 3.

[85] Gil McVean. “A genealogical interpretation of principal components analysis”. In: *PLoS genetics* 5.10 (2009), e1000686.

[86] Kevin R Moon et al. “Visualizing structure and transitions in high-dimensional biological data”. In: *Nature biotechnology* 37.12 (2019), pp. 1482–1492.

[87] C J Battey, Gabrielle C Coffing, and Andrew D Kern. “Visualizing population structure with variational autoencoders”. In: *G3* 11.1 (2021), jkaa036.

[88] Moritz Herrmann et al. *Enhancing cluster analysis via topological manifold learning*. 2022.

[89] Linda Ongaro et al. “Evaluating the Impact of Sex-Biased Genetic Admixture in the Americas through the Analysis of Haplotype Data”. In: *Genes* 12.10 (2021), p. 1580.

[90] Katharine L Korunes et al. “Sex-biased admixture and assortative mating shape genetic variation and influence demographic inference in admixed Cabo Verdeans”. In: *G3* 12.10 (2022).

[91] Beatriz Marcheco-Teruel et al. “Cuba: Exploring the History of Admixture and the Genetic Basis of Pigmentation Using Autosomal and Uniparental Markers”. In: *PLOS Genetics* 10.7 (2014). Publisher: Public Library of Science, e1004488.

# Re-use permissions

**Citation:** Diaz-Papkovich A, Anderson-Trocmé L, Ben-Eghan C, Gravel S (2019) UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLoS Genet 15(11): e1008432. https://doi.org/10.1371/journal.pgen.1008432

**Editor:** Sarah A. Tishkoff, University of Pennsylvania, UNITED STATES

**Received:** July 15, 2019; **Accepted:** September 17, 2019; **Published:** November 1, 2019



**Figure A1:** Copyright permission for Chapter 2.

A review of UMAP in population genetics

**Author:** Alex Diaz-Papkovich et al

**Publication:** Journal of Human Genetics

**Publisher:** Springer Nature

**Date:** Oct 14, 2020

*Copyright © 2020, The Author(s), under exclusive licence to The Japan Society of Human Genetics*

**Author Request**

If you are the author of this content (or his/her designated agent) please read the following. If you are not the author of this content, please click the Back button and select no to the question "Are you the Author of this Springer Nature content?".
Ownership of copyright in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights:
To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).
The author and any academic institution, where they work, at the time may reproduce the contribution for the purpose of course teaching.
To reuse figures or tables created by the Author and contained in the Contribution in oral presentations and other works created by them.
To post a copy of the contribution as accepted for publication after peer review (in locked Word processing file, of a PDF version thereof) on the Author's own web site, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the contribution on the publisher's website.
Authors wishing to use the published version of their article for promotional use or on a web site must request in the normal way.

If you require further assistance please read Springer Nature's online author reuse guidelines.

For full paper portion: Authors of original research papers published by Springer Nature are encouraged to submit the author's version of the accepted, peer-reviewed manuscript to their relevant funding body's archive, for release six months after publication. In addition, authors are encouraged to archive their version of the manuscript in their institution's repositories (as well as their personal Web sites), also six months after original publication.

v1.0

BACK CLOSE WINDOW

**Figure A2:** Copyright permission for Chapter 3.


New Results

Follow this preprint

**Topological stratification of continuous genetic variation in large biobanks**

Alex Diaz-Papkovich, Shadi Zabad, Chief Ben-Eghan, Luke Anderson-Trocmé, Georgette Femerling, Vikram Nathan, Jenisha Patel, Simon Gravel

**doi:** https://doi.org/10.1101/2023.07.06.548007



**Abstract**

Biobanks now contain genetic data from millions of individuals. Dimension-ality reduction, visualization and stratification are standard when exploring data at these scales; while efficient and tractable methods exist for the first two, stratification remains challenging because of uncertainty about sources of population structure. In practice, stratification is commonly performed by drawing shapes around dimensionally reduced data or assuming populations have a “type” genome. We propose a method of stratifying data with topo-logical analysis that is fast, easy to implement, and integrates with existing pipelines. The approach is robust to the presence of sub-populations of vary-ing sizes and wide ranges of population structure patterns. We demonstrate its effectiveness on genotypes from three biobanks and illustrate how topolog-ical genetic strata can help us understand structure within biobanks, evaluate distributions of genotypic and phenotypic data, examine polygenic score trans-ferability, identify potential influential alleles, and perform quality control.

**Competing Interest Statement**

The authors have declared no competing interest.



**Figure A3:** Copyright permission for Chapter 4.