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,
[2]Department of Human Genetics, McGill University, Montreal
[3]School of Computer Science, McGill University, Montreal
[4]Department of Bioengineering, McGill University, Montreal

*To whom correspondence should be addressed; E-mail: simon.gravel@mcgill.ca.

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 trans-

ferability, identify potential influential alleles, and perform quality control.

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 (typically of European ancestry) 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 K 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.

Methods

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 benefit 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

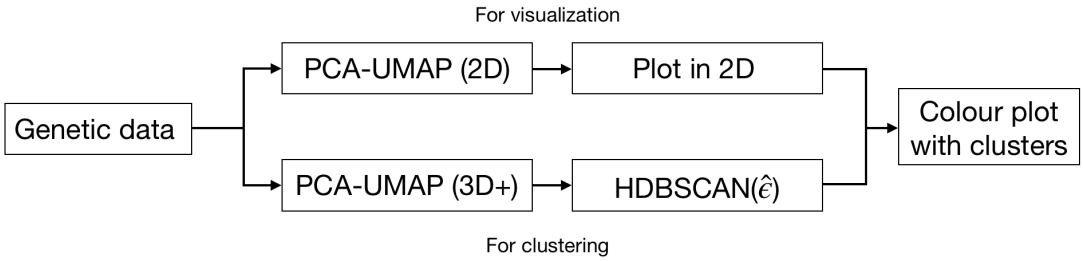


Figure 1: **Overview of visualization and clustering pipeline**. We use the same genotype data to generate visualizations as well as cluster labels and combine them within one figure. For visualization, we reduce data to 2D and optimize for visual clarity. For clustering, we use higher UMAP dimensions 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.

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 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.

Findings

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

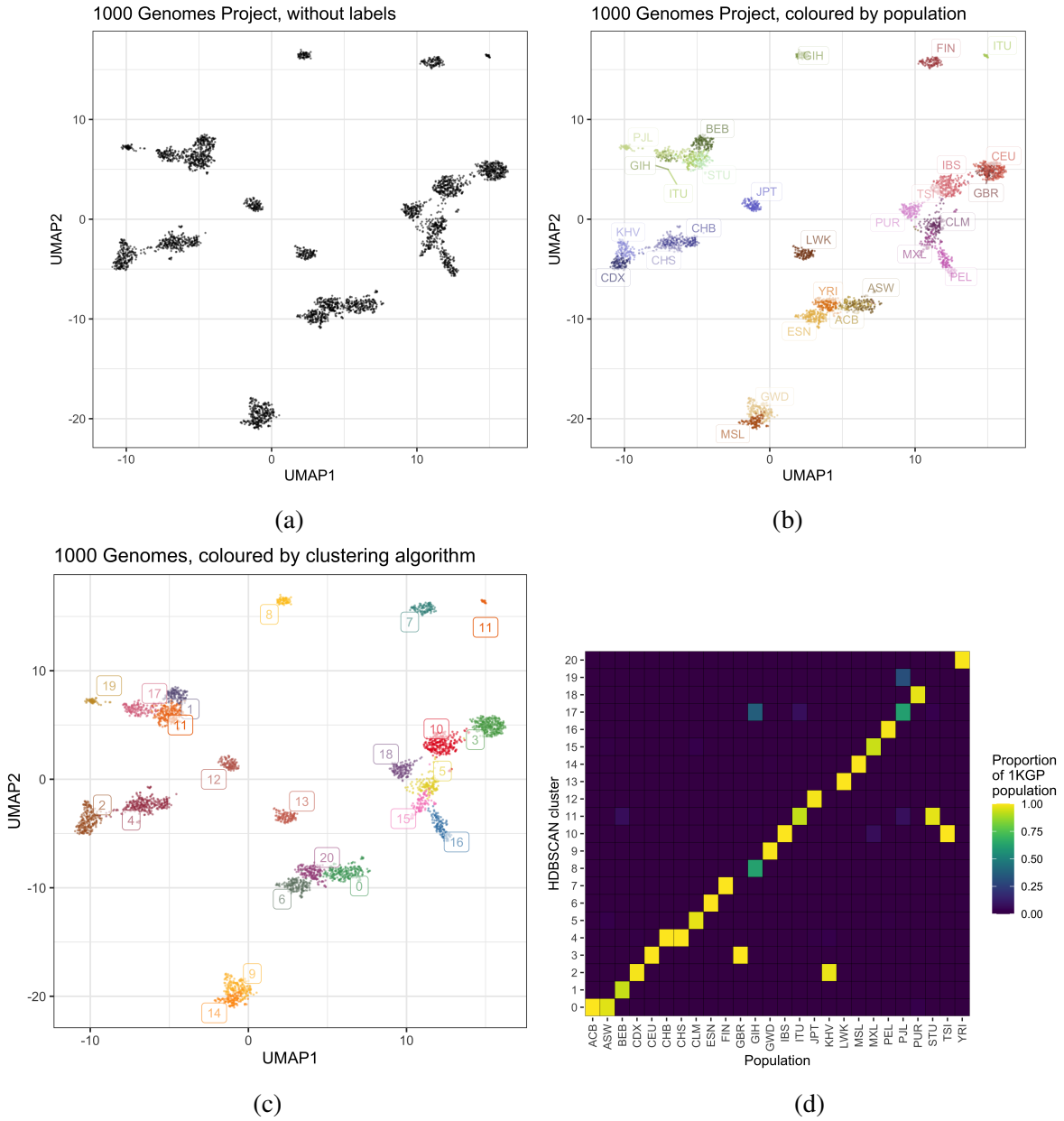


Figure 2: **Clusters generated from 1KGP genotype data reflect its population sampling.** (A) UMAP embedding of data without labels. (B) UMAP embedding of data, coloured by population label. (C) UMAP embedding of data, coloured by clusters derived from HDBSCAN($\hat{\epsilon}$) applied to a 5D UMAP embedding. (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 S3.

clusters from 1KGP data in two dimensions[4]. Figure 2 shows a UMAP representation of the 1KGP. Figure 2a shows the data without population labels (to mimic data with unknown populations), Figure 2b shows the data with corresponding population labels from the 1KGP, and Figure 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 2d shows that there is strong agreement between population label and cluster label, with full breakdowns provided in Tables S4 and S5. 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.

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 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 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

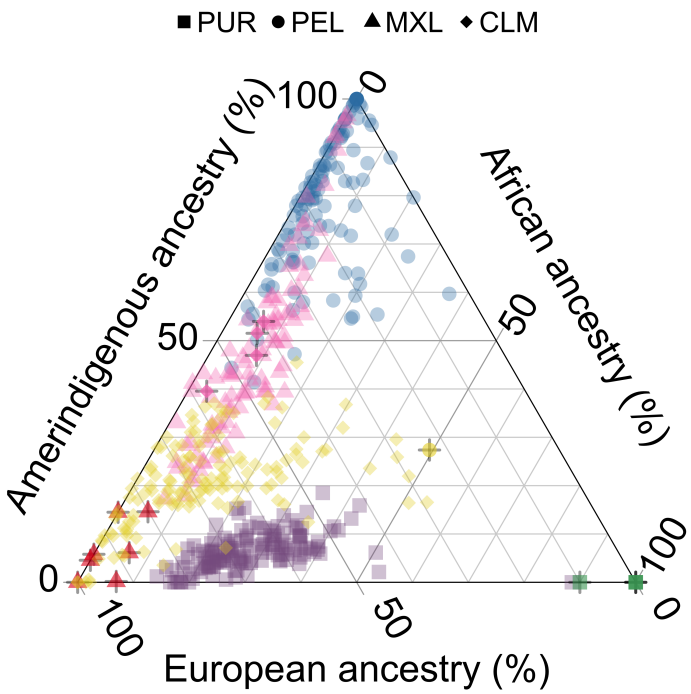


Figure 3: **Clusters capture structure in populations with overlapping admixture proportions in the 1KGP.** A ternary plot of the PUR, PEL, MXL, and CLM populations from the 1KGP with axes corresponding to global ancestry proportions estimated using ADMIXTURE ($K = 3$). Shapes indicate 1KGP label, colours indicate cluster label and match Figure 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 S4 and S5). 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.

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)

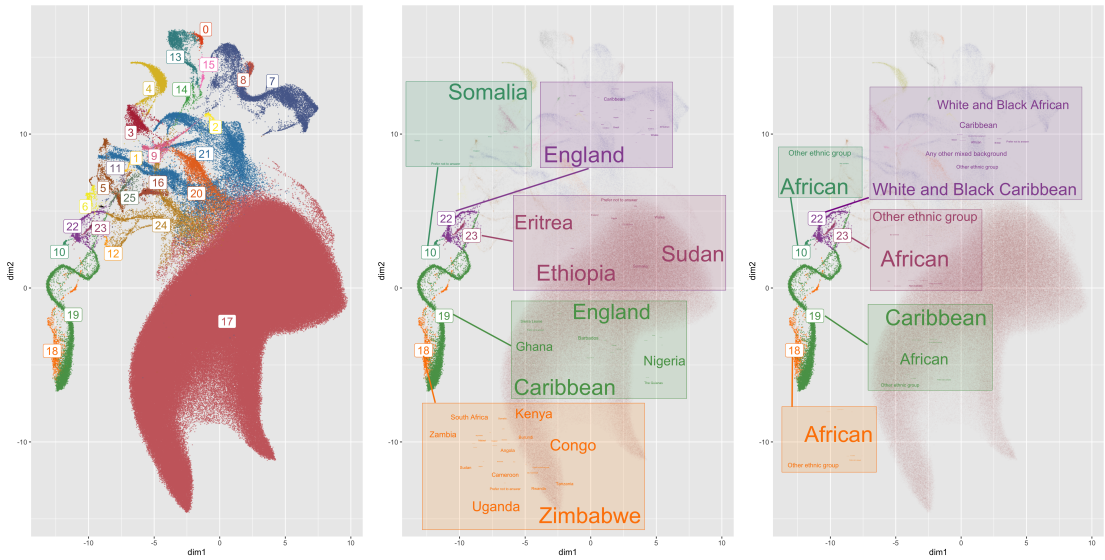
population is visibly split into two groups in 2D, while clustering carried out in 5D groups them together (Cluster 11). While some clusters will tend to persist across many parametrizations of UMAP and HDBSCAN($\hat{\epsilon}$), others based on more subtle patterns or in populations with more continuous variation will be less stable—though discrete groupings can help us understand data, the delineations are always, to a degree, arbitrary.

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 pre-determined hierarchical list. Participants first select their "ethnic group" from a list (e.g. "White"), which determines the list of "ethnic background" options they may select (e.g. "British"). There are 23 values of EB and 203 values for COB. 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 also been deeply explored, though typically focused on British or European individuals[23, 24, 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 S1).



(a) (b) (c)

Figure 4: **An example of clusters of population structure in the UKB.** The clusters reflect a mixture of native demographic history of the UK, the geographic origins of immigrants, the colonial history of the British Empire, and 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 S3. Detailed breakdowns of EB and country of birth are presented in Tables S6 and S7.

Figure 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 4b and 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 S2 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 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 S6 and S7 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 S3). 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.

Phenotype smoothing and modelling

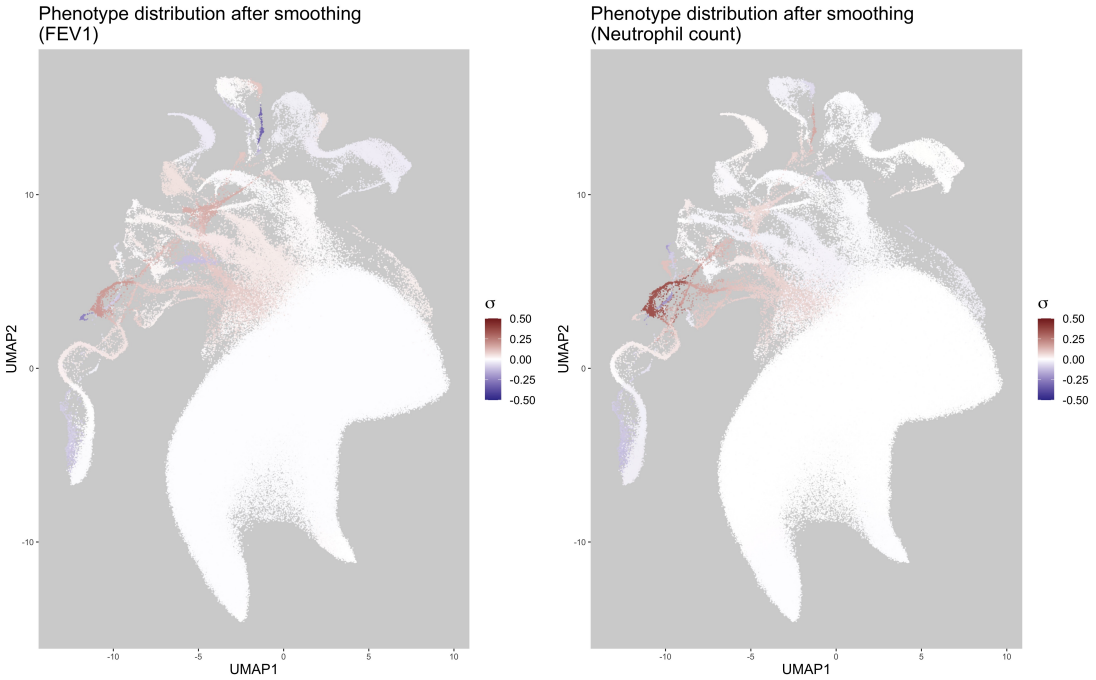
Epidemiological research often focuses on observed differences between groups—for example, finding the mean of a phenotype or sociodemographic measure and comparing between populations. Clustering on genetic data gives us one method to define groups based on shared

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 any number of 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 μ_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 5, we visualize the smoothing across 604 parametrizations of UMAP-HDBSCAN($\hat{\epsilon}$) for two phenotypes: FEV1 and neutrophil count. Despite having regressed out the effects of the top 40 principal components, there remains structure in the distribution of the residuals, visible at the scale of 0.5σ where σ is the standard deviation of the phenotype across the UKB. For example, the average residual value is noticeably higher in individuals who fall in Cluster 22 in Figure 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 S7), their phenotype distributions are not intermediate to clusters of

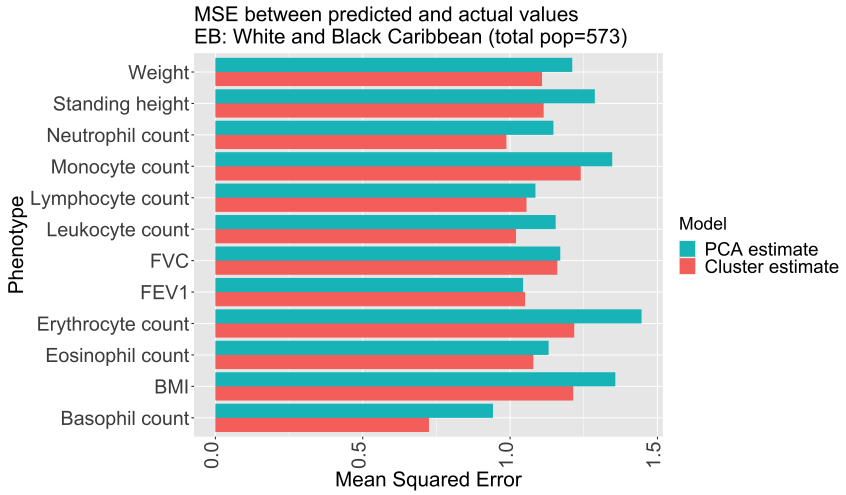


(a) (b)

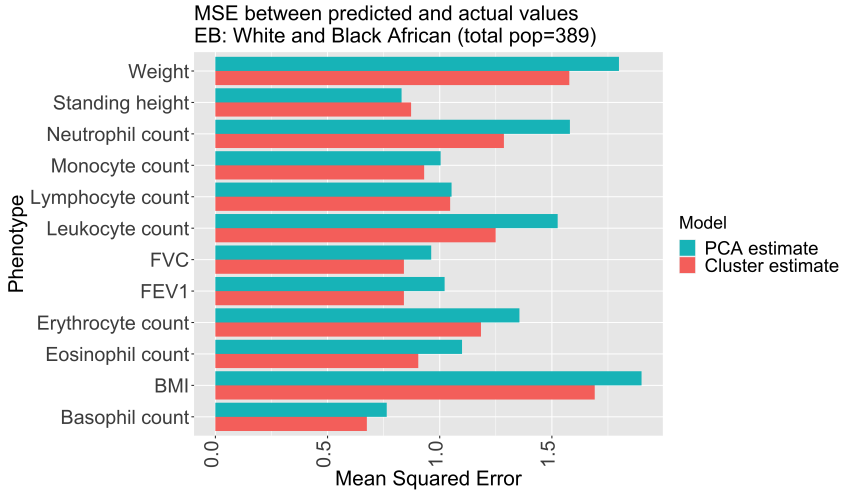
Figure 5: **Smoothed phenotypic measures across multiple parametrizations of clustering.** A 2D UMAP coloured by phenotype value after having removed the top 40 PCs and averaged by cluster, run over 604 parametrizations of the clustering pipeline. The colour scale runs from -0.5σ to 0.5σ, for the standard deviation σ 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.

primarily European- and African-ancestry individuals (Figure S5, Figure S6).

To test if the smoothed cluster estimates have explanatory power for these admixed individuals, we carried out an $80 - 20$ split on UKB data 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 6; full table of MSE values in Tables S8



(a)



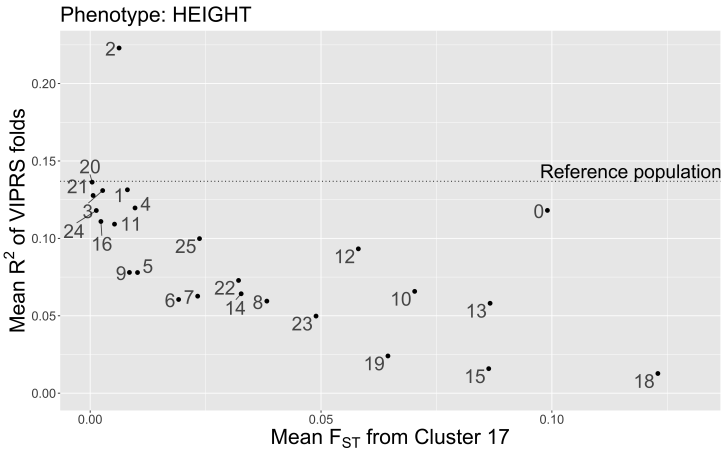
(b)

Figure 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 5.

and S9).

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 S4).

Evaluating transferability of polygenic scores



(a)

(b)

Figure 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.

Most investigations of PGS transferability are done at a population-level using a handful of large-scale geographical groups (e.g. "African", "European", "Asian"). However, these broad

populations can themselves exhibit population structure[30]. Instead, we use our 26 cluster labels from Figure 4a, and compared the transferability of PGS across them for several phenotypes.

Using UKB data, we calculated estimated effect sizes of SNPs using Variational Inference for Polygenic Risk Scores (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 studies. We then estimated phenotypes for individuals in the UKB, and calculated the values of the fixation index (F_{ST}) between the clusters. In Figure 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 7a) that as the F_{ST} between populations grows, the predictive value of the PGS tends to decrease; such a decrease is expected, due to potential 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 7b), and a regression model had no explanatory value. 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 a trait like 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 estimates of PGS, we modelled the

R^2 from the VIPRS estimates for each cluster against minor allele frequencies (MAF) of the top 100 SNPs in separate simple linear regressions and found the two strongest results were for $rs4420638$ and $rs7412$ (shown in Tables S1 and S2 with plots of regression in Figures S8 and S9, respectively). Both SNPs have their highest frequencies in Cluster 18 and both markers are in the apolipoprotein E (APOE) gene cluster; $rs7412$ had the largest effect size overall estimated by VIPRS ($\hat{\beta} = -0.1812$), 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 a large portion of variation in LDL in African Americans[34]. The $rs4420638$ allele was found to be significantly 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 successfully when there is a large degree of differentiation between the estimand population and the training population. However, an approach like UMAP-HDBSCAN($\hat{\epsilon}$) can provide a more detailed picture of the likely performance of a PGS in various genetic subgroups.

Quality control for complex multi-ethnic cohorts

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 processes like sample selection are carried out, preventing serious methodology problems or unnecessary exclusion of individuals.

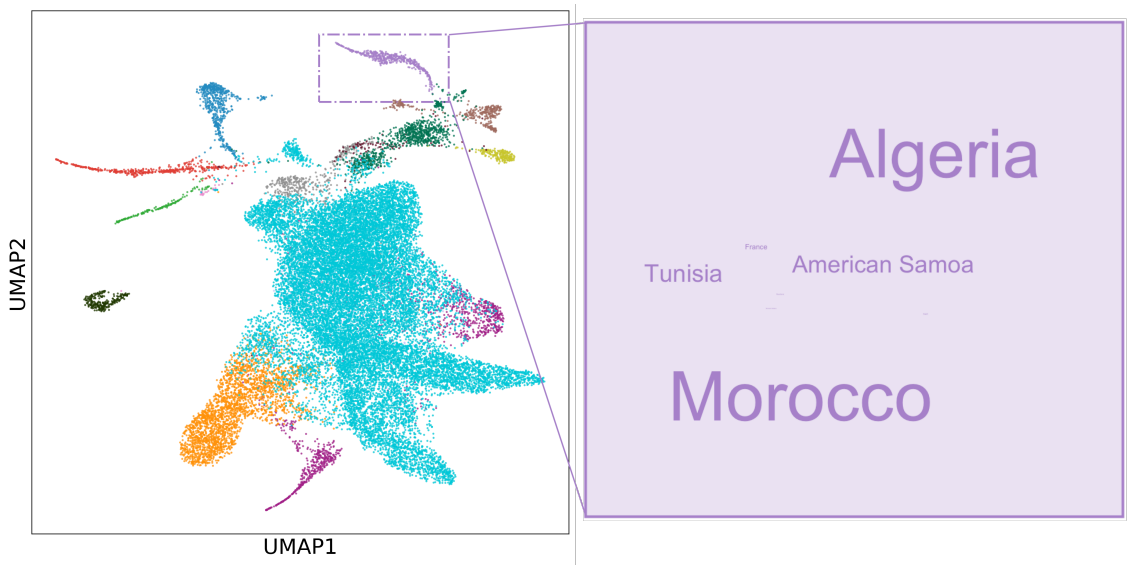


Figure 8: **Clustering can identify data collection errors.** A 2D UMAP coloured by clusters extracted using HDBSCAN($\hat{\epsilon}$). The highlighted cluster was found to have most of its individuals born in North Africa. A word cloud shows that a significant minority of individuals were born in American Samoa, which was found to be a coding error.

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 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.

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 K 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 an appropriate choice 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 Na-

tional 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. [12, 2]). The approach we propose is preferable to hand-delineation as it 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. Though some papers have used versions of HDBSCAN to process genetic data [26, 45], they used low-dimensional projections of data in PCA space and with versions of HDBSCAN less robust to differences in population sizes, leading to much less complete clustering with many individuals being left out. We believe that the approach presented here is a much more robust and easily implementable strategy to stratify

data.

A recent publication on polygenic scoring by Ding et al[7] suggested moving entirely away from stratification based on genetically inferred ancestry (GIA) clusters. Instead, they argued in favour of individual-level measures. They cite three issues with GIA 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 S10), 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 extracted are sensible reflections of the populations in the UKB, representing groups that have shared genetic and geographic histories, including for relatively small and recently admixed groups which were often excluded based on prior GIA approaches [7, 15]. We achieved similar results with CaG and 1KGP data, suggesting that our approach is robust to the idiosyncratic composition of a given biobank.

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 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 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 S11, for example, we see the large cluster of mostly European-born individuals is split into three smaller clusters, recapitulating the finding that there is considerable structure even in restricting analysis to, e.g., the White British population[3]. 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[46]), regional (CARTaGENE in Quebec[20]) and national (Million Veterans Project (MVP),[47], CANPATH[48]) biobanks. Leveraging these diverse cohorts can improve variant discovery[49, 50].

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[51]. A similar supervised approach with random forests was used by gnomAD[45]. 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 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 the self-declared 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].

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 K[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.

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.

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 file `ALL.wgs.nhgri_coriell_affy_6.20140825.genotypes_has_ped.vcf.gz` and population labels `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].

References

[1] W. Zhou, *et al.*, Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease, *Cell Genomics* **2**, 100192 (2022).

[2] S. Sakaue, *et al.*, Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction, *Nature Communications* **11**, 1569 (2020).

[3] A. A. Zaidi, I. Mathieson, Demographic history mediates the effect of stratification on polygenic scores, *eLife* **9**, e61548 (2020).

[4] A. Diaz-Papkovich, L. Anderson-Trocmé, C. Ben-Eghan, S. Gravel, UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts, *PLoS Genetics* **15** (2019).

[5] C. J. Battey, G. C. Coffing, A. D. Kern, Visualizing population structure with variational autoencoders, *G3 Genes|Genomes|Genetics* **11**, jkaa036 (2021).

[6] S. H. Holmes, W. Huber, *Modern statistics for modern biology* (Cambridge University Press, 2018).

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

[8] W. A. Freyman, *et al.*, Fast and Robust Identity-by-Descent Inference with the Templated Positional Burrows–Wheeler Transform, *Molecular Biology and Evolution* **38**, 2131 (2021).

[9] R. Shemirani, *et al.*, Rapid detection of identity-by-descent tracts for mega-scale datasets, *Nature Communications* **12**, 3546 (2021).

[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* (National Academies Press, Washington, D.C., 2023).

[11] K. E. Kaseniit, I. S. Haque, J. D. Goldberg, L. P. Shulman, D. Muzzey, Genetic ancestry analysis on >93,000 individuals undergoing expanded carrier screening reveals limitations of ethnicity-based medical guidelines, *Genetics in Medicine* **22**, 1694 (2020).

[12] B. V. Halldorsson, *et al.*, The sequences of 150,119 genomes in the UK Biobank, *Nature* **607**, 732 (2022).

[13] A. R. Martin, *et al.*, Clinical use of current polygenic risk scores may exacerbate health disparities, *Nature Genetics* **51**, 584 (2019).

[14] C. Ben-Eghan, *et al.*, Don't ignore genetic data from minority populations, *Nature* **585**, 184 (2020).

[15] D. O. Martschenko, H. Wand, J. L. Young, G. L. Wojcik, Including multiracial individuals is crucial for race, ethnicity and ancestry frameworks in genetics and genomics, *Nature Genetics* **55**, 895 (2023).

[16] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2020).

[17] C. Malzer, M. Baum, *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, *Nature* **526**, 68 (2015).

[19] C. 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, *PLOS Medicine* **12**, e1001779 (2015).

[20] P. Awadalla, *et al.*, Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics, *International Journal of Epidemiology* **42**, 1285 (2013).

[21] A. Romero, *et al.*, Diet Networks: Thin Parameters for Fat Genomics, *arXiv:1611.09340 [cs, stat]* (2017).

[22] D. Reich, K. Thangaraj, N. Patterson, A. L. Price, L. Singh, Reconstructing Indian population history, *Nature* **461**, 489 (2009).

[23] A. Abdellaoui, *et al.*, Genetic correlates of social stratification in Great Britain, *Nature Human Behaviour* **3**, 1332 (2019).

[24] E. Gilbert, A. Shanmugam, G. L. Cavalleri, Revealing the recent demographic history of Europe via haplotype sharing in the UK Biobank, *Proceedings of the National Academy of Sciences* **119**, e2119281119 (2022).

[25] A. M. Chiu, E. K. Molloy, Z. Tan, A. Talwalkar, S. Sankararaman, Inferring population structure in biobank-scale genomic data, *The American Journal of Human Genetics* **109**, 727 (2022).

[26] S. D. Nagar, A. M. Nápoles, I. K. Jordan, L. Mariño-Ramírez, Socioeconomic deprivation and genetic ancestry interact to modify type 2 diabetes ethnic disparities in the United Kingdom, *eClinicalMedicine* **37**, 100960 (2021).

[27] Slave Voyages: The Trans-Atlantic Slave Trade Database, Trans-atlantic slave trade - estimates (2023). http://www.slavevoyages.org/estimates/BeZD1wTh.

[28] C. Fortes-Lima, P. Verdu, Anthropological genetics perspectives on the transatlantic slave trade, *Human Molecular Genetics* **30**, R79 (2021).

[29] A. C. F. Lewis, *et al.*, Getting genetic ancestry right for science and society, *Science* **376**, 250 (2022).

[30] A. B. Kamiza, *et al.*, Transferability of genetic risk scores in African populations, *Nature Medicine* **28**, 1163 (2022).

[31] S. Zabad, S. Gravel, Y. Li, Fast and accurate Bayesian polygenic risk modeling with variational inference, *The American Journal of Human Genetics* (2023).

[32] Y. Wang, *et al.*, Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations, *Nature Communications* **11**, 3865 (2020).

[33] A. M. Bennet, *et al.*, Pleiotropy in the Presence of Allelic Heterogeneity: Alternative Genetic Models for the Influence of APOE on Serum LDL, CSF Amyloid-β 42, and Dementia, *Journal of Alzheimer's Disease* **22**, 129 (2010).

[34] L. 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, *Clinical and Translational Science* **5**, 394 (2012).

[35] S. Sanna, *et al.*, Fine Mapping of Five Loci Associated with Low-Density Lipoprotein Cholesterol Detects Variants That Double the Explained Heritability, *PLOS Genetics* **7**, e1002198 (2011).

[36] C. Breitling, *et al.*, Genetic Contribution of Variants near SORT1 and APOE on LDL Cholesterol Independent of Obesity in Children, *PLOS ONE* **10**, e0138064 (2015).

[37] H. Wang, *et al.*, Associations of genetic variants of lysophosphatidylcholine metabolic enzymes with levels of serum lipids, *Pediatric Research* **91**, 1595 (2022).

[38] A. K. MacLeod, *et al.*, Some principles and practices of genetic biobanking studies, *European Respiratory Journal* **33**, 419 (2009).

[39] A. Diaz-Papkovich, L. Anderson-Trocmé, S. Gravel, A review of UMAP in population genetics, *Journal of Human Genetics* **66**, 85 (2021).

[40] Y. Qian, B. L. Browning, S. R. Browning, Efficient clustering of identity-by-descent between multiple individuals, *Bioinformatics* **30**, 915 (2014).

[41] D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of Population Structure using Dense Haplotype Data, *PLOS Genetics* **8**, e1002453 (2012).

[42] E. Han, *et al.*, Clustering of 770,000 genomes reveals post-colonial population structure of North America, *Nature Communications* **8**, 14238 (2017).

[43] J. Nait Saada, *et al.*, Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations, *Nature Communications* **11**, 6130 (2020).

[44] C. Caggiano, *et al.*, Health care utilization of fine-scale identity by descent clusters in a Los Angeles biobank (2022).

[45] K. J. Karczewski, *et al.*, The mutational constraint spectrum quantified from variation in 141,456 humans, *Nature* **581**, 434 (2020).

[46] G. M. Belbin, *et al.*, Toward a fine-scale population health monitoring system, *Cell* **184**, 2068 (2021).

[47] H. Hunter-Zinck, *et al.*, Genotyping Array Design and Data Quality Control in the Million Veteran Program, *The American Journal of Human Genetics* **106**, 535 (2020).

[48] T. J. B. Dummer, *et al.*, The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for research on chronic disease prevention, *CMAJ* **190**, E710 (2018).

[49] G. L. Wojcik, *et al.*, The PAGE Study: How Genetic Diversity Improves Our Understanding of the Architecture of Complex Traits, *preprint*, Genetics (2017).

[50] M. Lin, D. S. Park, N. A. Zaitlen, B. M. Henn, C. R. Gignoux, Admixed Populations Improve Power for Variant Discovery and Portability in Genome-Wide Association Studies, *Frontiers in Genetics* **12** (2021).

[51] H. Fang, *et al.*, Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies, *The American Journal of Human Genetics* **105**, 763 (2019).

[52] T. Smith, S. 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] J. Carlson, B. M. Henn, D. R. Al-Hindi, S. Ramachandran, Counter the weaponization of genetics research by extremists, *Nature* **610**, 444 (2022).

[54] W. D. Roth, The multiple dimensions of race, *Ethnic and Racial Studies* **39**, 1310 (2016).

[55] S. Ben-David, Clustering - What Both Theoreticians and Practitioners are Doing Wrong, *arXiv:1805.08838 [cs, stat]* (2018).

[56] C. Hennig, What are the true clusters?, *arXiv:1502.02555 [stat]* (2015).

[57] D. J. Lawson, L. van Dorp, D. Falush, A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots, *Nature Communications* **9**, 3258 (2018).

[58] D. A. Vyas, L. G. Eisenstein, D. S. Jones, Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms, *New England Journal of Medicine* **383**, 874 (2020).

[59] A. Choudhury, *et al.*, High-depth African genomes inform human migration and health, *Nature* **586**, 741 (2020).

[60] S. Purcell, *et al.*, PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses, *The American Journal of Human Genetics* **81**, 559 (2007).

[61] D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals, *Genome Research* **19**, 1655 (2009).

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

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

[64] M. Hlavac, stargazer: Well-Formatted regression and summary statistics tables (2018).

Supporting Information (SI)

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 ≥ 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 2 we used the top 16 PCs; for the UKB in Figure 4a and CaG in Figure 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 computational 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 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 $i = 1 \ldots I$, 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 5 with the values $e_{p,i} = y_{p,i} - (\hat{\beta_{p,0}} + \sum_{j=1}^{40} \hat{\beta_{p,j}} PC_{j,i})$

SI Figures

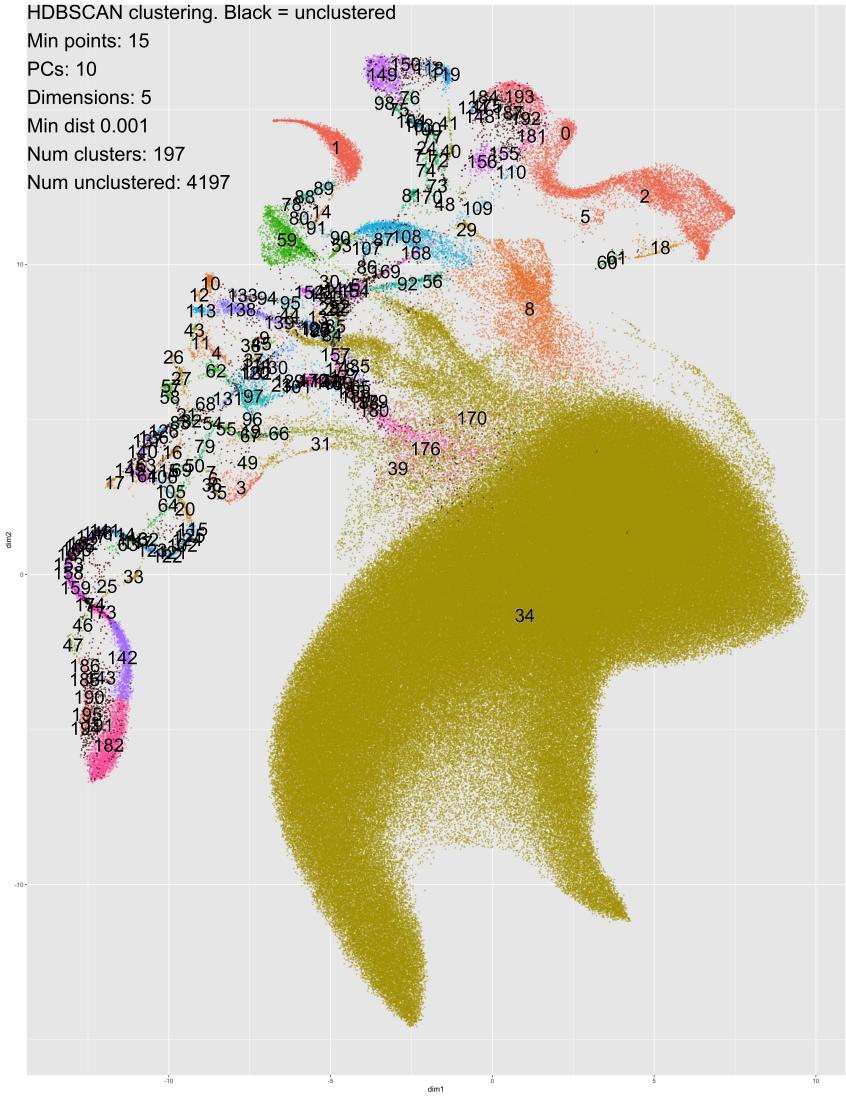
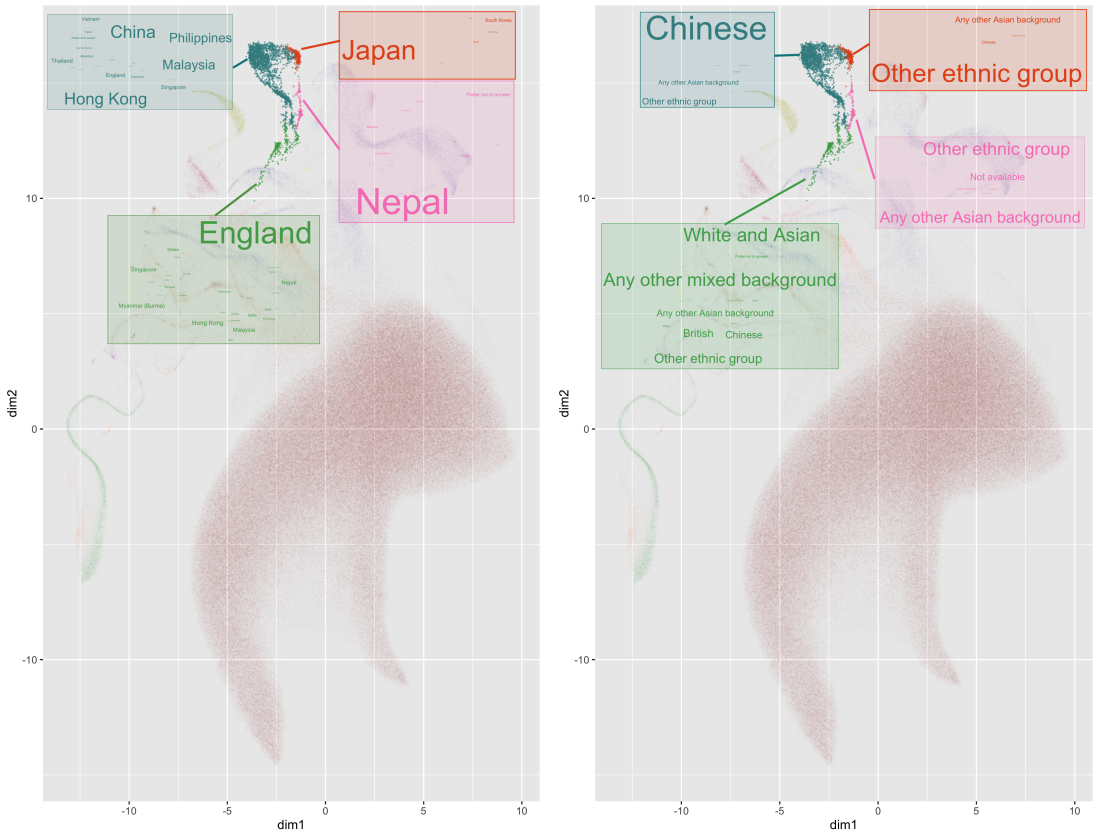


Figure S1: An example of a clustering of the UKB data using HDBSCAN rather than HDBSCAN($\hat{\epsilon}$). The algorithm fails to adequately cluster many of the sub-populations, categorizing $4,197$ individuals as noise and generated 197 micro-clusters.



(a) (b)

Figure S2: Word clouds generated from four clusters in the UKB from Figure 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 S6 and S7.

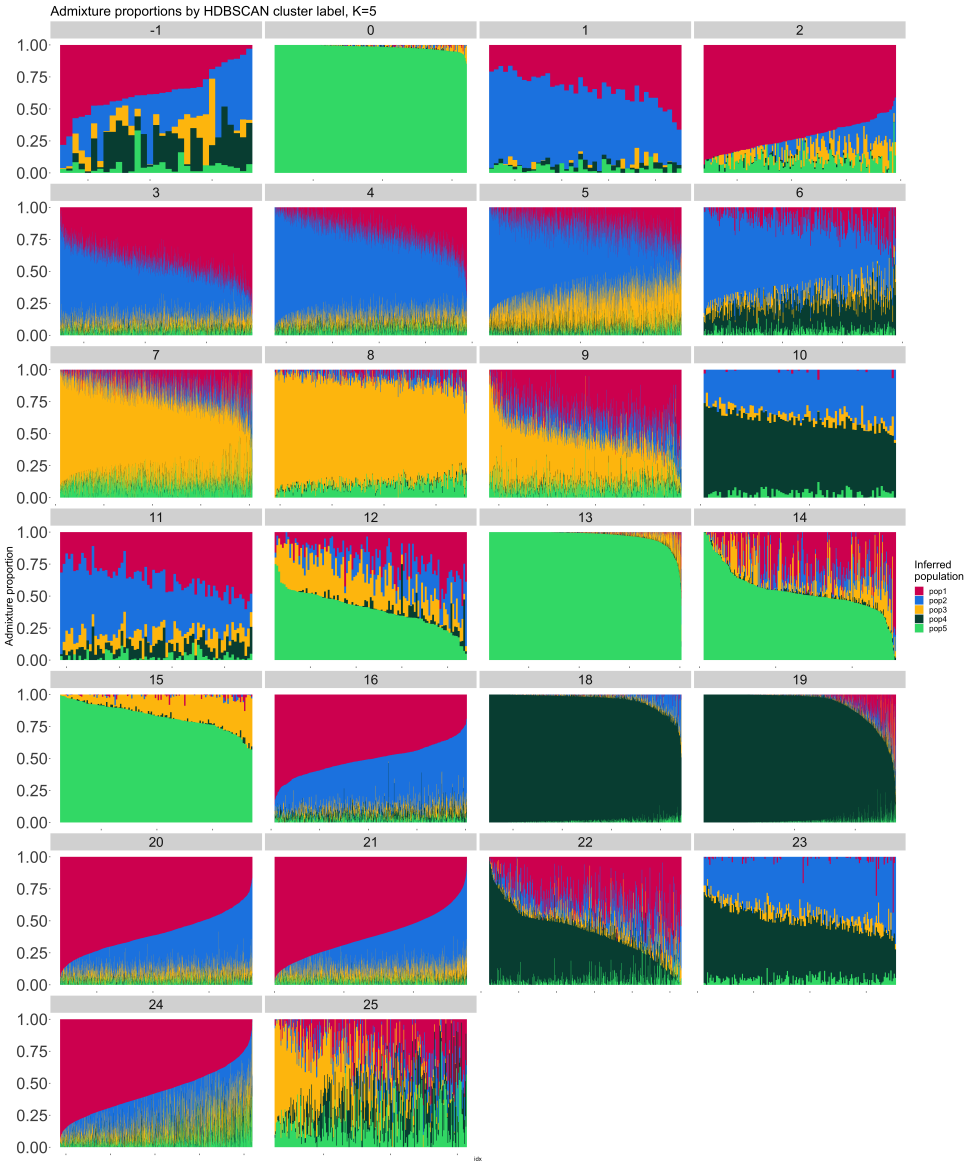


Figure S3: Admixture proportions for $K = 5$ populations on each of the clusters in Figure 4. Cluster 17 ($n > 400,000$) was excluded for computational reasons. Individuals not assigned to a cluster are labelled as -1.

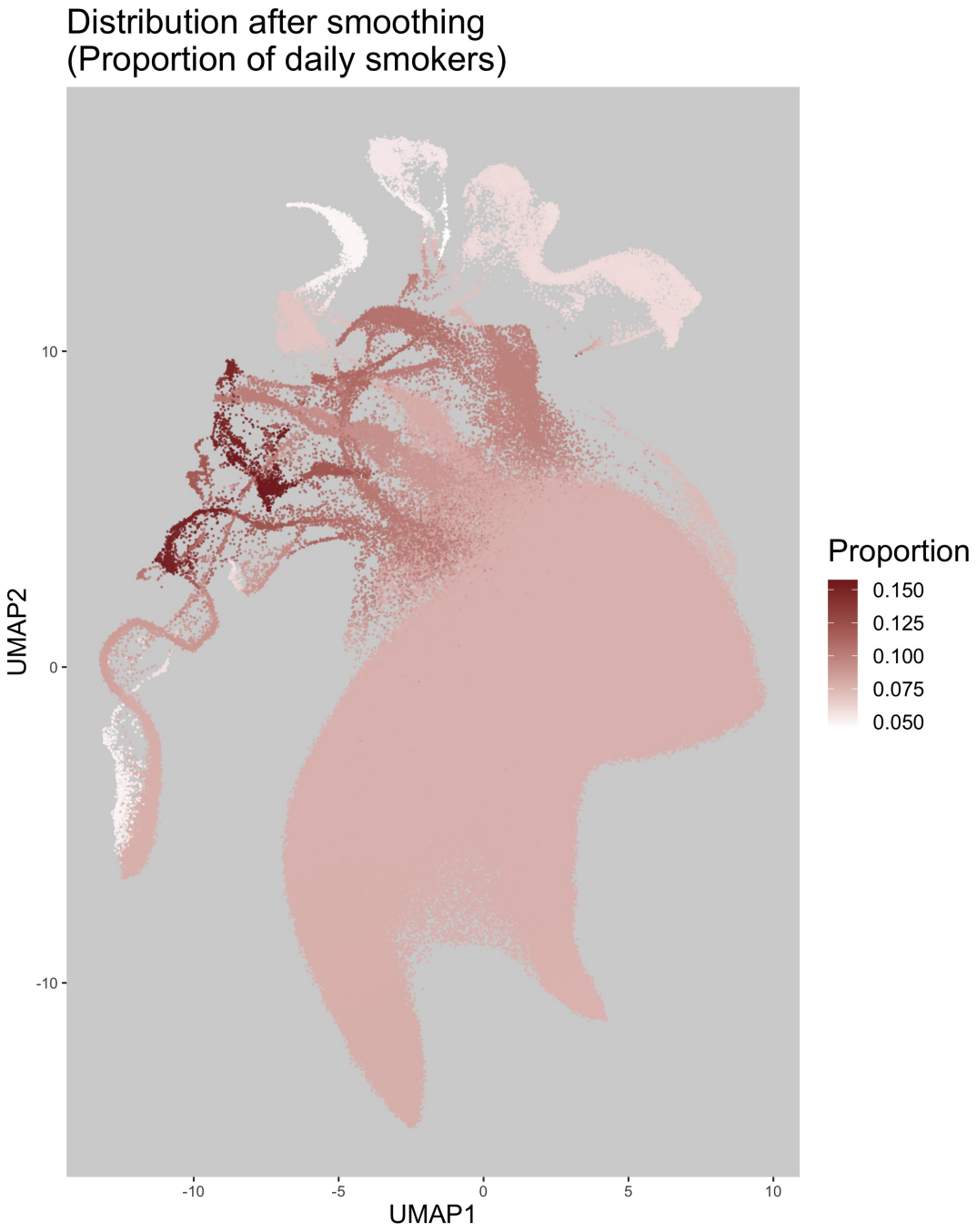
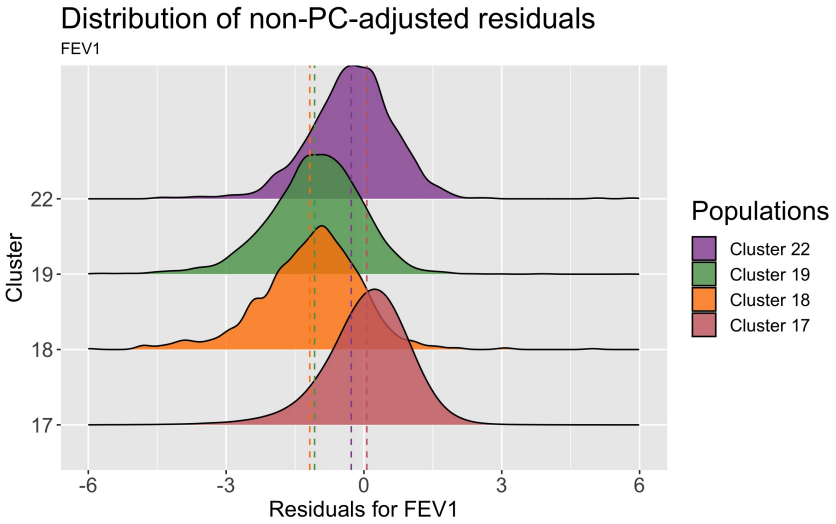
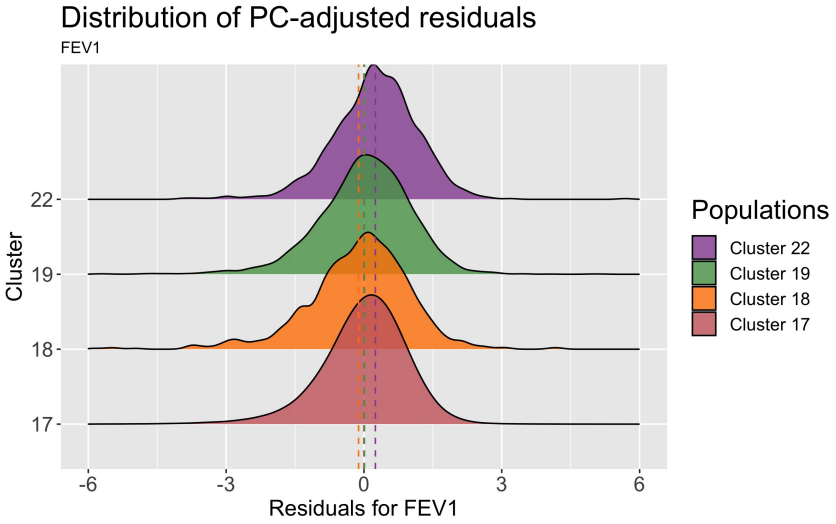


Figure S4: Proportion of daily smokers, smoothed using Algorithm 1.

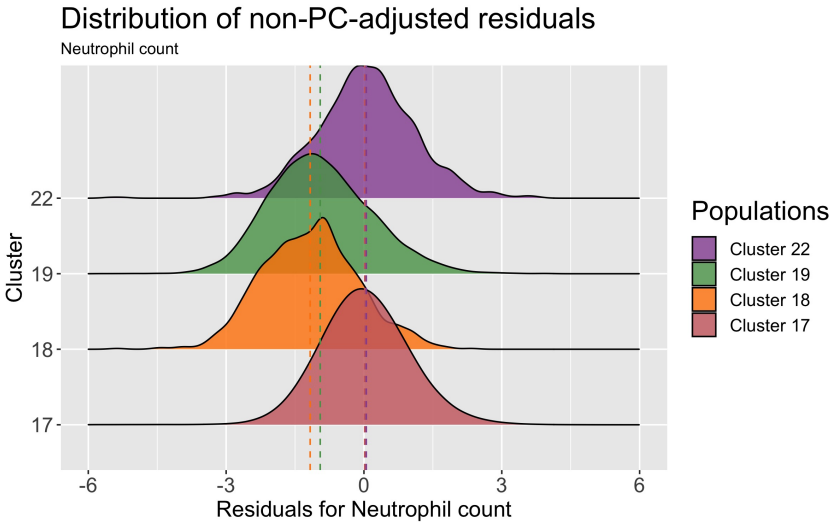


(a)

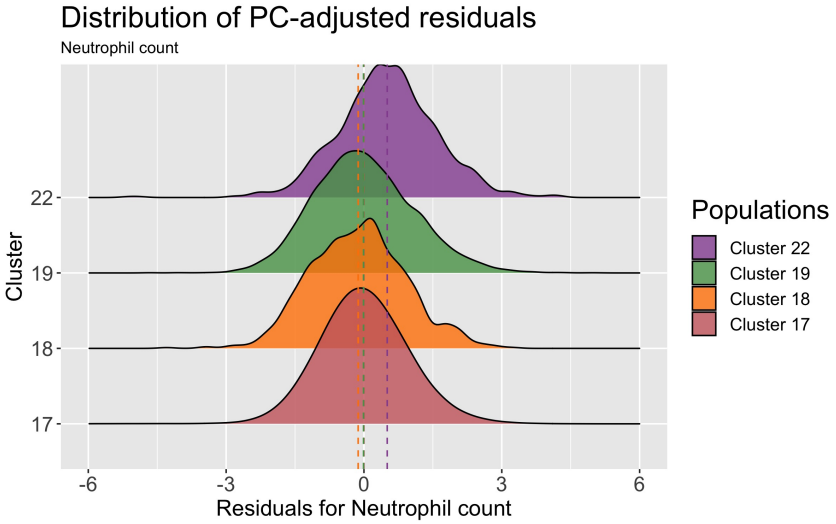


(b)

Figure S5: 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 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 S6: 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 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.

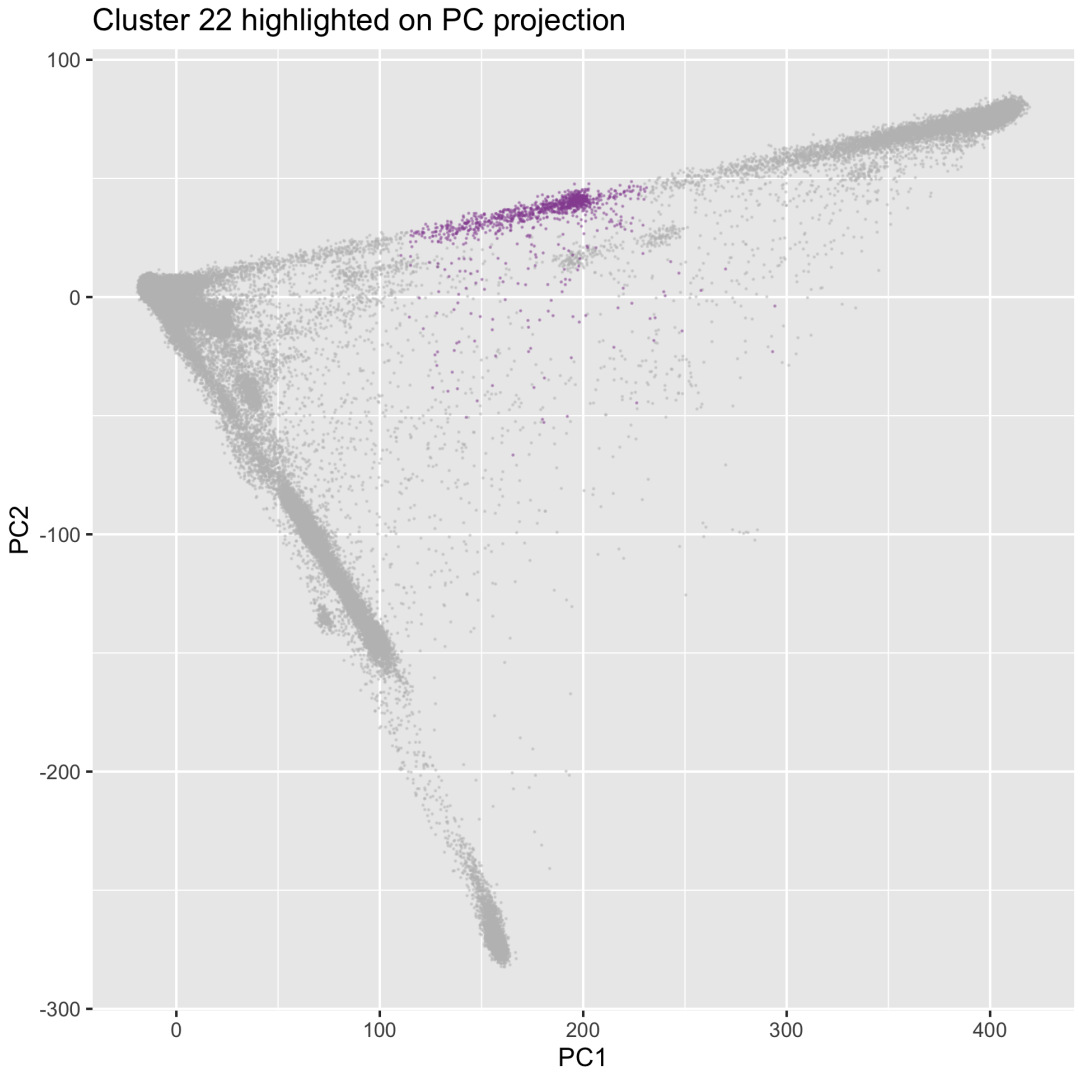


Figure S7: Cluster 22 from Figure 4a highlighted coloured in on a plot of PC1 and PC2.

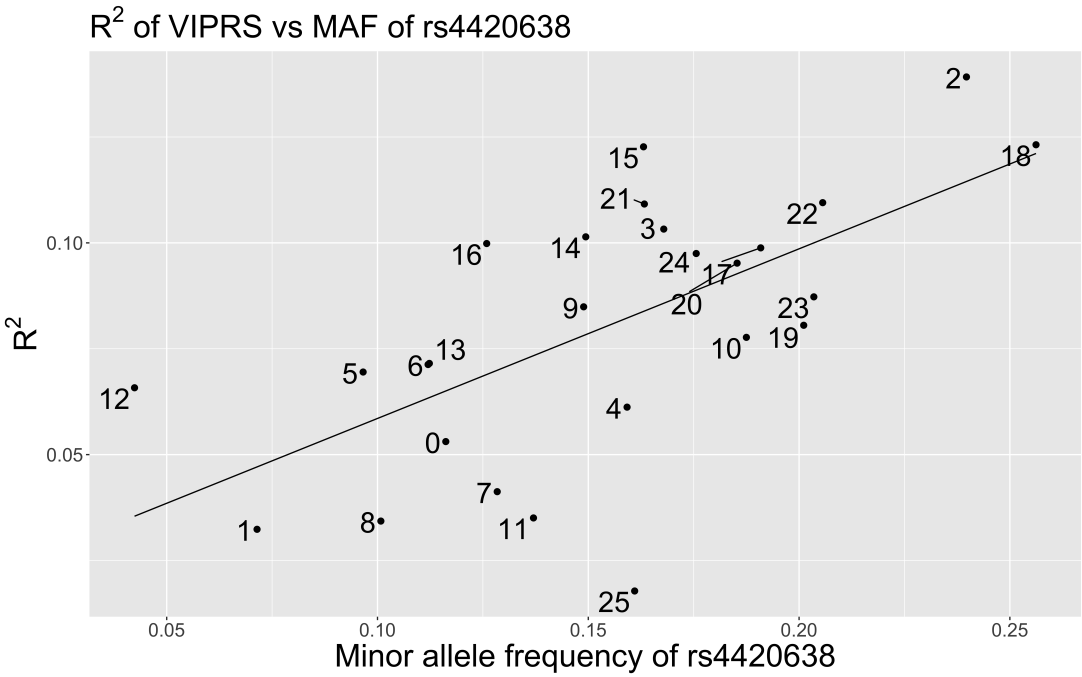


Figure S8: Regression line of the R^2 of a PGS generated by VIPRS versus the minor allele frequency $rs4420638$, labelled by clusters from Figure 4a. The regression summary is presented in Table S1.

	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 S1: Linear regression model between minor allele frequency (MAF) of $rs4420638$ within each cluster from Figure 4a and the R^2 of a PGS for LDL generated by VIPRS using the clusters from Figure 4a. The plot of the regression is present in Figure S8.

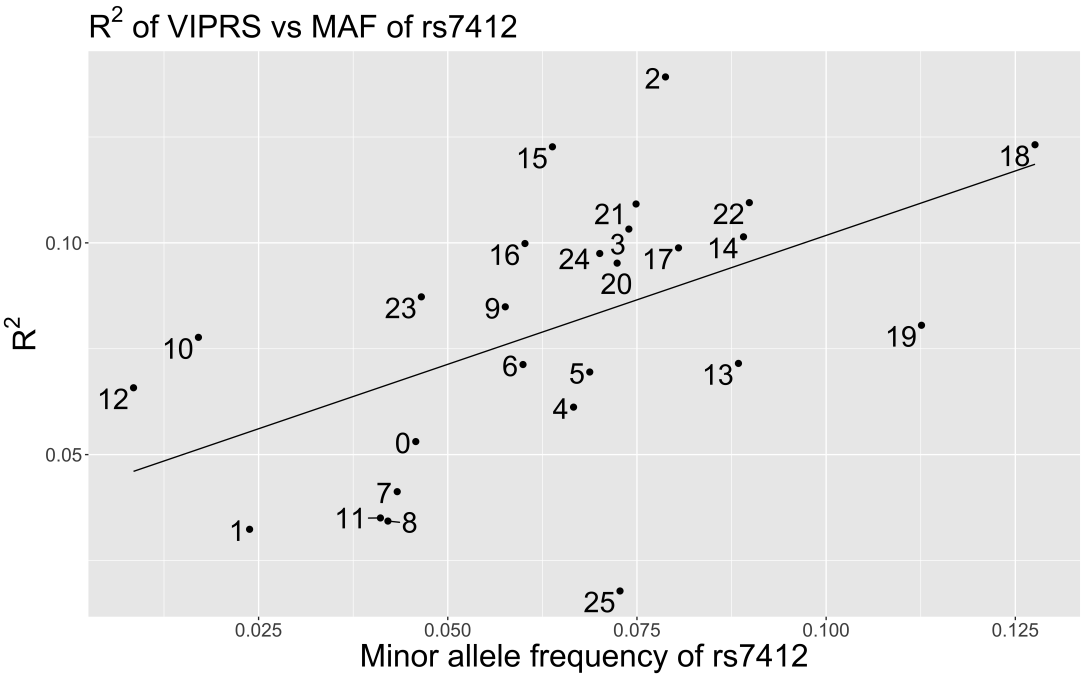
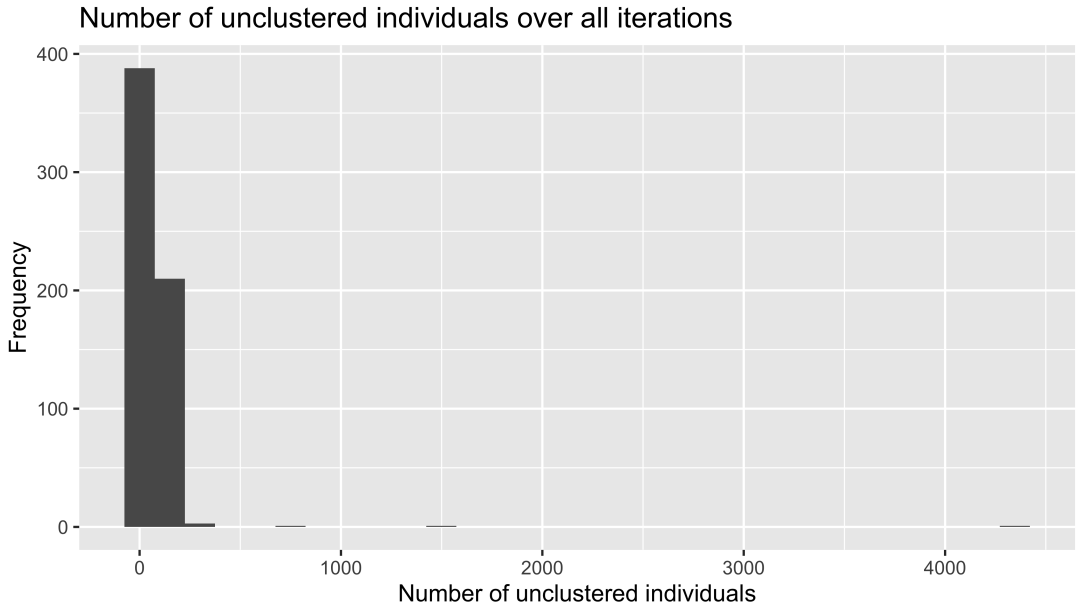


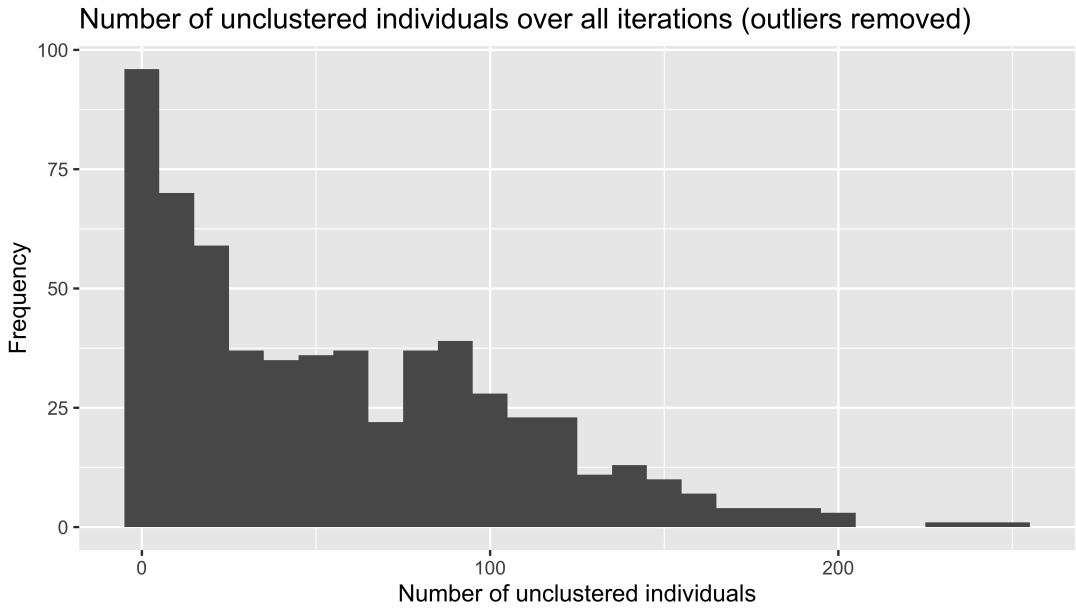
Figure S9: Regression line of the R^2 of a PGS generated by VIPRS versus the minor allele frequency $rs7412$, labelled by clusters from Figure 4a. The regression summary is presented in Table S2.

	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 S2: Linear regression model between minor allele frequency (MAF) of $rs7412$ within each cluster from Figure 4a and the R^2 of a PGS for LDL generated by VIPRS using the clusters from Figure 4a. The plot of the regression is present in Figure S9.



(a)



(b)

Figure S10: 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.

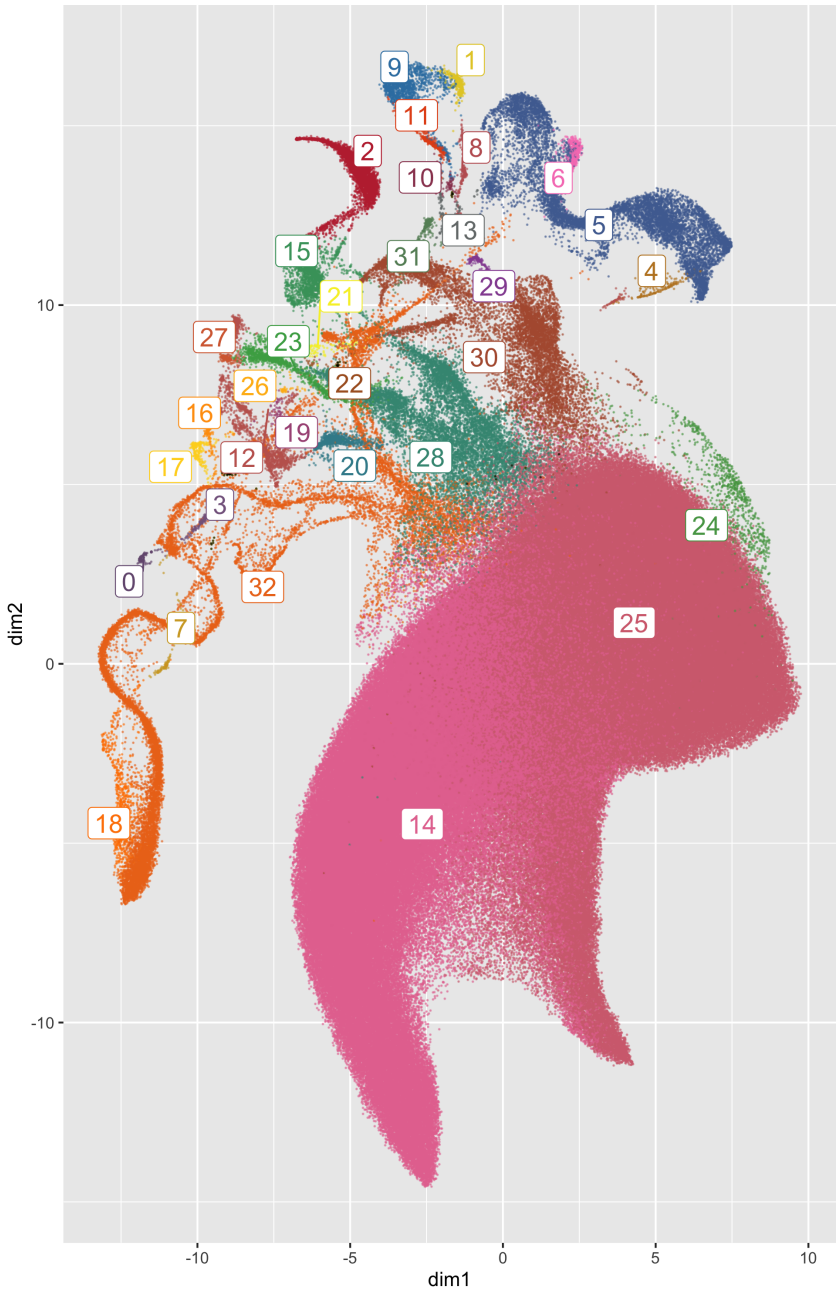


Figure S11: An alternative clustering of UKB data. Compared to Figure 4a, the largest cluster (Cluster 17 in that figure) has been split into three smaller clusters (Clusters 14, 24, 25 in this figure). Other clusters have been split or merged, while some remain the same between runs.

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 S3: 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 S4: 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 S5: Composition of each cluster broken down by 1KGP population.

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

Table S6: Frequency of country of birth by cluster for Figure 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 S7: Frequency of selected EB by cluster for Figure 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 S8: 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 S9: 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.