Title

A framework for research into continental ancestry groups of the UK Biobank

Authors

Andrei-Emil Constantinescu[1,2,3], Ruth E. Mitchell[1,2], Jie Zheng[1,2], Caroline J. Bull[1,2,3], Nicholas J. Timpson[1,2], Borko Amulic[4], Emma E. Vincent[1,2,3], David A. Hughes[1,2]*

1 MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom.

2 Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom.

3 School of Translational Health Sciences, University of Bristol, Bristol, United Kingdom.

4 School of Cellular and Molecular Medicine, University of Bristol, Bristol, United Kingdom.

*Corresponding author: Dr. David A. Hughes (d.a.hughes@bristol.ac.uk)

Keywords

Ancestry, UK Biobank, Population structure

Declarations

Ethics approval and consent to participate

UK Biobank received ethical approval from the NHS National Research Ethics Service North West (11/NW/0382; 16/NW/0274) and was conducted in accordance with the Declaration of Helsinki. All participants provided written informed consent before enrolment in the study.

Consent for publication

All authors consented to the publication of this work.

Availability of data and material

Genetic data from UK Biobank were made available as part of project code 15825. Analytical code is available on GitHub at https://github.com/andrewcon/popgen-biobank.

Competing interests

None to declare.

Funding

AC acknowledges funding from a Medical Research Council PhD studentship (MR/N013794/1). NJT and REM acknowledge funding from the Medical Research Council (MC_UU_00011/1). NJT is the PI of the Avon Longitudinal Study of Parents and Children (Medical Research Council & Wellcome Trust 217065/Z/19/Z) and is supported by the University of Bristol NIHR Biomedical Research Centre (BRC-1215-2001). EEV, CJB, NJT and DH acknowledge funding from the Wellcome Trust (202802/Z/16/Z). EEV, CJB and

NJT also acknowledge funding by the CRUK Integrative Cancer Epidemiology Programme (C18281/A29019). EEV and CJB are supported by Diabetes UK (17/0005587) and the World Cancer Research Fund (WCRF UK), as part of the World Cancer Research Fund International grant program (IIG_2019_2009). JZ is supported by the Academy of Medical Sciences (AMS) Springboard Award, the Wellcome Trust, the Government Department of Business, Energy and Industrial Strategy (BEIS), the British Heart Foundation and Diabetes UK (SBF006\1117). JZ is funded by the Vice-Chancellor Fellowship from the University of Bristol and is supported by Shanghai Thousand Talents Program. BA acknowledges funding from the Medical Research Council (MR/R02149x/1). The funders of the study had no role in the study design, data collection, data analysis, data interpretation or writing of the report.

Authors' contributions

AC, DH, and REM conceived the idea for the paper. AC and DH conducted the analysis. All authors contributed to the interpretation of the findings. AC and DH wrote the manuscript. All authors critically revised the paper for intellectual content and approved the final version of the manuscript.

Acknowledgements

We are grateful to the UK Biobank study and its participants. This research has been conducted using the UK Biobank resource under Application 15825.

Authors' information (optional)

MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom.

AC, REM, JZ, CJB, NJT, EEV and DH

Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom.

AC, REM, JZ, CJB, NJT, EEV and DH

School of Translational Health Sciences, University of Bristol, Bristol, United Kingdom.

AC, CJB and EEV

School of Cellular and Molecular Medicine, University of Bristol, Bristol, United Kingdom.

BA

Abstract

Background

The UK Biobank is a large prospective cohort, based in the United Kingdom, that has deep phenotypic and genomic data on roughly a half a million individuals. Included in this resource are data on approximately 78,000 individuals with “non-white British ancestry”. Whilst most epidemiology studies have focused predominantly on populations of European ancestry, there is an opportunity to contribute to the study of health and disease for a broader segment of the population by making use of the UK Biobank’s “non-white British ancestry” samples. Here we present an empirical description of the continental ancestry and population structure among the individuals in this UK Biobank subset.

Results

Reference populations from the 1000 Genomes Project for Africa, Europe, East Asia, and South Asia were used to estimate ancestry for each individual. Those with at least 80% ancestry in one of these four continental ancestry groups were taken forward (N=62,484). Principal component and K-means clustering analyses were used to identify and characterize population structure within each ancestry group. Of the approximately 78,000 individuals in the UK Biobank that are of “non-white British” ancestry, 50,685, 6,653, 2,782, and 2,364

individuals were associated to the European, African, South Asian, and East Asian continental ancestry groups, respectively. Each continental ancestry group exhibits prominent population structure that is consistent with self-reported country of birth data and geography.

Conclusions

Methods outlined here provide an avenue to leverage UK Biobank's deeply phenotyped data allowing researchers to maximise its potential in the study of health and disease in individuals of non-white British ancestry.

Introduction

As the research community strives to understand the genetic architecture of disease [1], it has increasingly realized the necessity of inclusion and diversity – of ethnically, ancestrally, environmentally, and geographically diverse populations [2–5]. Not simply to enhance knowledge about health and disease, but to insure health equity. Epidemiological studies, including genome-wide associations studies (GWAS), have been overwhelmingly conducted in European populations [2]. However, funding efforts and studies including the Human Heredity and Health in Africa (H3Africa) Initiative [6], the Population Architecture using Genomics and Epidemiology (PAGE) Consortium [7], Trans-Omics for Precision Medicine Consortium [8], Hispanic Community Health Study / Study of Latinos (SOL) [9], and the All of Us Research Program [10] are making concerted efforts to include and increase the number of under-represented populations in genomic epidemiology studies.

The UK Biobank project (UKBB) has phenotypic and genomic data from a prospective cohort of approximately 500,000 individuals from across the United Kingdom [11,12]. It has become an outstanding resource for studies of health and disease, and genetic diversity within the United Kingdom. Whilst it is made up of around 430,000 "white British

ancestry" individuals, as defined by UKBB, it also contains a wealth of diversity from other self-described ethnicities (~78,000). This is a resource that should be utilized to help expand inclusion and diversity in epidemiological studies.

The Pan-UK Biobank, or the Pan-ancestry genetic analysis of the UKBB, has leveraged the diversity present in UKBB and is freely providing GWAS summary statistics for over seven thousand phenotypes in six continental ancestry groups (https://pan.ukbb.broadinstitute.org). The genetic "ancestry" groups identified by Pan-UK Biobank and within our study refer to groups of individuals with a shared genetic ancestry and demographic history. Studies and public resources like Pan-UK Biobank are vital to the goal of increasing under-represented populations and the larger goal of describing and understanding the genetic architecture of phenotypic traits and disease. However, the limited information on intra-population structure and non-specific use of covariates in Pan-UK Biobank GWAS models may influence association effect estimates. A description of the continental diversity and population structure present in the UKBB will aid future study design, methodological choice(s) and ultimately improve our understanding of how genotype influences phenotype.

Here, we describe an approach to define continental ancestry groups and provide a description of the structure and population differentiation within them. We define "ancestry" here as genetic ancestry or the complex inheritance of one's genetic material, but in practice we will be using methodologies that use genetic similarity to identify groups of individuals with high (genetic) affinity or likeness [13]. The aim is to identify relatively homogenous groups of individuals that approach populations consistent with a Hardy-Weinberg model and are resultantly more appropriate for many of the assumptions built into many of the methods

used in genomic epidemiology studies [14,15]. We leverage public data from the 1000 Genomes Project (1KG) [16] to provide reference populations from four, therein described, superpopulations or (sub)-continental ancestry groups (CAGs) – namely, Africa (AFR), Europe (EUR), South Asia (SAS), and East Asia (EAS). We note that we will refer to the groupings or clusters of individuals derived by this work, not as populations, but as groups or clusters of individuals. Further, the groups and clusters identified here are used as discrete units, but ancestry does not have decisive boundaries and is a continuum [17–20]. The use of discrete units is an analytical simplification. Finally, the overarching purpose of our study is to provide a description of the population structure present in the UKBB as an aid to future research investigating the health of individuals from diverse ancestries.

Results

Estimations of continental ancestry

Each of the 78,296 UKBB "non-white British" were included in a supervised ADMITXTURE analysis to estimate a proportion of ancestry to each of African (AFR), European (EUR), South Asian (SAS), and East Asian (EAS) continental ancestry groups (**Figure 1**). The proportion of continental ancestry is further illustrated, for each individual, within the context of UKBB population structure on principal components (PC) one and two as provided by the UKBB (**Figure 2**). AFR ancestry (**Figure 2A**) runs largely parallel with PC1, the major axis of variation. EUR ancestry runs at a roughly 135-degree angle (**Figure 2B**) along PC1 and PC2, while SAS (**Figure 2C**) and EAS (**Figure 2D**) ancestry run, largely, along PC2. Of the approximately 78,000 UKBB samples included in the ADMIXTURE analysis 50,685, 6,653, 2,782, and 2,364 individuals had 80% or more of their ancestry attributed to the EUR, AFR, SAS, and EAS continental super-populations, respectively. These individuals were carried forward into further analyses of population structure within

these continental ancestry groups (CAGs). The 80% threshold was chosen to allow some error in the broader continental classification whilst also placing a limit on the complex structure and admixture evaluated in these subsets. A total of 15,812 "non-white British" UKBB study participants were not included in any of the four CAGs, given the methods and cut-offs used here.

Population structure within continental regions.

To evaluate the level of population structure among the UKBB CAGs, we first re-estimated principal components for each, while also projecting individuals from 1KG populations from each super-population respectively, onto the newly derived PCs (**Figure 3**, **Supplementary Table 1**). For each there is considerable overlap between UKBB individuals and 1KG populations, providing some context for the diversity that is present within the UKBB. In the AFR continental ancestry group principal component one distinguishes West African from East African 1KG populations, while PC3 distinguishes among populations of West Africa (**Figure 3A**). In the EUR continental ancestry group, the PCs and 1KG populations illustrate a strong North-South axis along PC2, with a similar but less distinctive trend on PC1 (**Figure 3B**). In the SAS continental ancestry group, there is a South-North trend along PC1, but no remarkable pattern can be attributed to the PCs (**Figure 3C**). The 1KG sample populations in the EAS ancestry group appear to indicate a North-South axis along PC1, and a West to East axis along PC2 (**Figure 3D**).

K-means clustering of PCs

Given that many population genetics and epidemiological analyses, such as genome-wide association studies, depend on limited population structure, a common desire is to have a relatively homogeneous population sample for these analyses. As such, we used an

unsupervised algorithm to identify groups of individuals that approach Hardy-Weinberg population assumptions. To do so we performed a K-means analysis on the top PCs (see Methods, **Supplementary Figure 1**), from each CAG, to identify 'K' subclusters or groups within each. An optimum number of K-clusters was determined by a silhouette analysis (see Methods, **Supplementary Figure 2**). For each CAG, using only the UKBB participants, we identified seven, two, four, and three K-clusters of individuals for AFR, EUR, SAS, and EAS, respectively (**Supplementary Figure 3**). However, for the EUR CAG we choose the second-best K-cluster (K=6) for the remaining analyses to improve our ability to investigate the utility of this analytical method to discriminate population structure (**Figure 4**).

Country of birth

To evaluate the informativeness of these K-clusters we mapped each individuals' country of birth and United Nations (UN) geographic regions onto the PCs (**Figure 5** and **Supplementary Figure 4-5**). These figures further illustrate the diversity and structure present in the sample. Each CAG presents an observable degree of population structure, and region of birth (ROB) data illustrates non-specific associations between CAGs and ROB (Figure 5). For example, a large number of individuals have an East African ROB but are estimated to have more than 80% of their ancestry from South Asia (Figure 5 C and G). Nevertheless, ROB data illustrates structure across principal components for each CAG. Yet to ascertain if there is a correlation among the K-clusters identified above and the self-reported place of birth we performed a correspondence analysis for each CAG. The analyses indicate a correlation between K-means clusters and the UN regions for each continent: AFR (Dim1 53.29%, Dim2 41.88%), EUR (Dim1 58.25%, Dim2 28.67%), SAS (Dim1 80.00%, Dim2 18.2%), EAS (Dim1 92.11%, Dim2 7.89%) (**Figure 6A**). When UN regions for a smaller geographical region were substituted, namely country of birth (COB; Supplementary

Figures 6-9), an attenuated but correlated structure remained: AFR (Dim1 28.32%, Dim2 25.02%), EUR (Dim1 40.43%, Dim2 31.89%), SAS (Dim1 61.60%, Dim2 25.31%), EAS (Dim1 50.49%, Dim2 49.51%) (**Figure 6B**, **Supplementary Figure 10**).

Population differentiation

An evaluation of the degree of population differentiation within each CAG was performed by estimating Fst, or the fixation index between each pair of K-cluster groups and 1KG populations. All single-nucleotide polymorphisms (SNPs) that were included in each CAG's principal component analysis were used here. An average, minimum, and maximum estimate was used to summarize the distribution of estimates between pairs (**Figure 7**). Relative to the population differentiation observed in the 1KG sample populations we observed, on average, a small degree a population differentiation among AFR and EUR K-means clusters, and larger average estimates among SAS and EAS groups. Among the UKBB samples average Fst estimates indicate that the EAS CAG has the largest amount of population differentiation with an average Fst of 0.0133. This is followed by SAS with an average estimate of 0.0092, EUR with 0.0037, and finally AFR with the smallest average estimate of 0.003. However, we note that these estimates were derived from SNPs with a European ascertainment bias and as such they may not coincide with analyses using an unbiased set of genetic variants.

Discussion

Here we present an analytical pipeline to identify individual participants of the UKBB study with diverse and under-represented ancestries to be used in genomic epidemiology studies. Whilst cohort studies centred in diverse geographic locations are essential for elucidating the effect of environment and genotype on disease, the diversity present in deeply

phenotyped studies such as the UKBB should be utilized where possible. This study presents a description of some of the diversity present in the UKBB. Further, the methods presented here provide an approach to identify subsets of individuals to help broaden, inform, and improve the relevance of genetic epidemiological studies and their findings for those of, in this specific instance, a non-white British ancestry (**Figure 8**).

Throughout the paper, when we speak of ancestry, we are referring to "genetic ancestry", or individuals who share a demographic history [13,21,22]. They would, at the population level, share a history of mutation, genetic drift, recombination, migration, natural selection, environment, and culture (niche construction). As a product, they would have different genetic variants, allele frequencies, and patterns of linkage disequilibrium across their genomes [23–25].

The need to perform analyses, like association studies, separately in unique ancestral populations is largely born from the need to avoid correlations between phenotype and genetic ancestry, or differences in allele frequencies among populations [13,26]. For example, if a disease (or environmentally influenced trait) is more frequent in ancestral population 'A' than it is in 'B' and your association analysis pools these ancestral populations together you may erroneously identify any allele that is more frequent in population 'A' as a genetic variant associated with the disease. To avoid these confounding issues, analyses are commonly limited to relatively homogenous populations.

In genome-wide association studies, the aim is to derive accurate unbiased effect estimates for a genetic variant on a trait. However, the task becomes increasingly challenging, as variation in genetic ancestry comes with different allele frequencies, genetic backgrounds and environments [27]. Methods such as the inclusion of relatedness matrixes and principal components [28–31] are used to account for cryptic relatedness and undetected, fine-scale population stratification. In addition, they are also used to account for correlations

between phenotype and genetic ancestry [32,33]. However, are the inclusion of relatedness matrixes or principal components enough to control the structure present in the CAGs presented here? Or would smaller (K-means clusters) more homogenous populations be better suited to epidemiological analyses, like GWAS?

The problems introduced by population stratification persist even in populations like the "white British" subset of the UKBB, where individual genetic variants and polygenic scores for individual traits can retain correlations with geography, even after correcting for population structure [34,35]. Moreover, when sampling populations across Europe – where genetic ancestry does mirror geography [36,37] – and meta-analysing independently run GWASs [38], effect estimates appear to retain a bias introduced by population structure [39,40]. These fine scale issues exemplify some of the reasons for performing separate epidemiological analysis, like GWAS, for populations with deeper population differentiations, i.e. unique ancestries, demographic histories, and environments.

The complications of population stratification and opportunities for improving health outcomes for more people, even at the continental level, are precisely why a description of the structure within each continental ancestry group was provided here. Namely, the structure present within a CAG, as identified here, may also be too great to be properly accounted for with common methodologies and may thus need to be resolved into smaller more homogenous groups. At the very least, careful consideration is warranted when interpreting results where CAGs are used - because structure matters [41]. The unsupervised clustering performed within each CAG is not a perfect solution for identifying true "populations" – an exercise that may in fact be an impractical goal – but it is a method to identify groups of individuals with a more similar, homogeneous ancestry. Other techniques like uniform manifold approximation and projection [42] or more explicit leveraging of self-described ethnicity could help improve the identification of homogenous groups. Self-described

ethnicity is not a synonym for genetic ancestry though, as it is a sociocultural construct. It would however help inform cultural, social, and other environmental influences – important aspects of a "population" - on phenotypes and disease [22].

In summary, we assigned individuals to continental ancestry groups (**Figure 1 and 2**); illustrated the structure present among individuals within each CAG (**Figure 3**), identified unsupervised clusters or groups of individuals within each (**Figure 4**) and demonstrated that those clusters have an affinity to regions and countries of birth – i.e. the K-means clusters are consistent with geographic structure and isolation by distance models [43,44] (**Figure 5**). Notably, each CAG presents extensive structure, inconsistent with a randomly mating population, but rather with the sampling of unique, geographically distant populations. In particular, East Asian, South Asian, and African CAGs have isolated, or discontinuous groups of individuals in the UKBB sample, exemplified in the K-means clustering analysis (Figure 4) [19,20]. For example, groups K1 and K3 in the EAS CAG (**Figure 4D**) epitomizes this discontinuous structure as they correspond to individuals born on the islands of Philippines and Japan, respectively (**Figure 5**, **Supplementary Figure 8**).

The methods employed here do have several limitations: First, a single 1KG population was used to represent each of four continental ancestry groups evaluated – Africa, Europe, South Asia, and East Asia. One population is a poor proxy for all of the variation present in any one (sub)-continent. However, as the 1KG project does not have optimal population coverage, including more or all the 1KG populations of a CAG would still poorly represent all the variation present in a (sub)-continent and would complicate the assignment of individuals to a single ancestry group. Second, our analysis was limited to four (sub-)continental ancestry groups, to the exclusion of the Americas (AMR, a 1KG superpopulation). Populations from the Americas often have a large and varying amount of recent admixture from various European and African populations [25,45–49]. As such,

316 including an AMR population in the ADMIXTURE analysis, as a reference population, could
317 confound the genetic ancestries being estimated. However, whilst we limit this study to a
318 few, broad, well characterized ancestry groups the approach presented her can be generalised
319 to other, specific ancestries.

320 Third, the UKBB Axiom array used to genotype all UKBB participants was designed
321 to optimize imputation of a European population while also including genetic variants
322 previously associated with disease and other phenotypic traits derived from studies primarily
323 conducted in European populations [11,12]. As a product, the genomic data used here will
324 have an ascertainment bias [50] that would influence imputation accuracy (although no
325 imputation data was used here), allele frequency distributions, estimates of linkage
326 disequilibrium and diversity and divergence within and among populations. Each of these
327 may influence estimations of population differentiation, principal component estimates and
328 the inferences made from them [51,52]. Specific study designs [53,54] have been made to
329 remove ascertainment bias in genotype arrays so that unbiased inferences could be made for a
330 wider range of genetic ancestries, but this was not available here.

331 Fourth, the principal components illustrated and used in the unsupervised K-means
332 clustering analyses were derived from the UKBB participants only and resultantly represents
333 the diversity (point three) and genetic ancestry found in that data set. The inclusion or use of
334 other public data sets with more numerous sample populations, that better represent regional,
335 or continental diversity will provide alternative patterns of structure. Fifth, we are limited by
336 the reference population used in the analyses. Whilst the 1KG data set shall remain an
337 essential reference panel for broad analyses like those conducted here, researchers with
338 specific continental or geographically specific research questions could strengthen and refine
339 the observations made here by including other geographically specific data sets. Finally, the
340 unsupervised K-means clustering analysis is dependent upon the number of PCs included in

it. Here the number of PCs chosen did have an element of subjectivity (Supplementary Figure 1). Whilst analytical methods are available to select a number of informative PCs [55], we did not implement such methods here. Given that the K-means algorithm weights each PC equally, we sought to limit the PCs included to only those with the largest proportions of variance explained and not necessarily all that are analytically estimated to be informative.

Conclusions

The approach presented here demonstrates a method to leverage the deeply phenotyped and widely used UKBB data set to help improve the inclusion and equity of epidemiological studies for under-represented populations. Careful considerations must be given to the diversity present within continental ancestry groups. However, given the thousands of individuals present in the genetic ancestry groups identified here, the UKBB data set shall prove insightful for studies of health and disease in populations beyond the British Isles. While the methods presented here do not describe a perfect solution to identify populations, we hope that they provide an avenue to leverage the diverse data available in UKBB and a methodological platform to improve and build upon.

Methods

Description of working environment

All analyses were performed in a Linux environment supported by the University of Bristol's Advanced Computing Research Centre (ACRC) using the following publicly available software packages: Plink v1.9 and v2.0 [56,57], ADMIXTURE v1.3.0 [58,59], and EIGENSOFT v8.0.0 [29,30]. In addition, bespoke scripts, analyses, and figures were run and

364 generated in the R environment using version 3.6.2 on the ACRC computer clusters and
365 version 4.0.2 (Taking Off Again) on local computers [60].
366
367 *UK Biobank data*
368 This research has been conducted using the UKBB Resource under Application
369 Number 15825, from which directly genotyped SNP data (N=784,256 SNPs) were made
370 available. It includes data for a total of 78,296 individuals identified by UKBB as "non-white
371 British" participants – our analyses were restricted to this subset. In addition to genotypic
372 data, we also acquired several variables of interest (self-described ancestry, country of birth)
373 data for this subset of individuals. 365 exclusions were made when filtering those with sex
374 chromosome mismatch and/or aneuploidy, and outliers with high genetic heterozygosity and
375 missing rates [61].
376
377 *1000 Genomes data*
378 Genetic data (v5a.20130502) from phase three of the 1KG, which includes data from
379 5 continental, or 1KG described super-populations [Europe (EUR), East Asia (EAS), South
380 Asia (SAS), Africa (AFR), and the Americas (AMR)], were used to provide reference
381 populations for admixture analyses and population structure inferences ([62]
382 http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/). Our analyses did not include populations from
383 the AMR superpopulation. This is to maintain a simplified analysis that avoided the
384 complicating factors of the potentially recent admixture events that occurred in the Americas.
385 Included in our analyses are five populations from 1KG super-population label: (AFR), also
386 known as the continental Africa ancestry group (1) Yoruba in Ibadan, Nigeria (YRI); (2)
387 Luhya in Webuye, Kenya (LWK); (3) Gambian in Western Division, The Gambia -
388 Mandinka (GWD); (4) Mende in Sierra Leone (MSL) and (5) Esan in Nigeria (ESN). Five

populations from the super-population label EUR or the continental Europe ancestry group: (1) Utah residents with Northern and Western European ancestry (CEU); (2) Toscani in Italia (TSI); (3) British in England and Scotland (GBR); (4) Finnish in Finland (FIN) and (5) Iberian populations in Spain (IBS). Five populations from the super-population label SAS or the continental South Asian ancestry group: (1) Gujarati Indian in Houston, Texas (GIH); (2) Punjabi in Lahore, Pakistan (PJL); (3) Bengali in Bangladesh (BEB); (4) Sri Lankan Tamil in the UK (STU) and (5) Indian Telugu in the UK (ITU). Finally, five populations from the super-population label EAS or the continental East Asian ancestry group: (1) Han Chinese in Beijing, China (CHB); (2) Japanese in Tokyo, Japan (JPT); (3) Han Chinese South (CHS); (4) Chinese Dai in Xishuangbanna, China (CDX) and (5) Kinh in Ho Chi Minh City, Vietnam (KHV).

Merging UK Biobank and 1000 Genomes

The directly genotyped data from UKBB was used to identify SNPs with the same SNP identifier (RefSNP ID) present in the 1KG data set. A total of 718,711 SNPs were identified with the same ID and extracted from both data sets using PLINK v2.0. The two datasets were then merged using the -bmerge function in PLINK v2.0. After removing problematic SNPs (e.g. multi-allelic, duplicate) in the merge step, a total of 718,487 SNPs remained.

Linkage disequilibrium pruning

Prior to ancestry estimation the merged dataset was reduced to a set of independent SNPs based on linkage disequilibrium (LD) estimates using the PLINK v2.0 function and parameters "--indep-pairwise 50 10 0.025", indicating an r^2 threshold of 0.025, a window size of 50 kilobases and a window step size of 10 kilobases. In addition, 24 previously identified

genomic regions with extensive linkage disequilibrium were also excluded [63,64]. LD estimates in this analysis were limited to unrelated individuals from the 1KG YRI population sample. A total of 30,320 SNPs remained following LD pruning.

Estimating African, European, South Asian, and East Asian ancestry

Four 1KG populations were included as reference populations in a supervised Admixture (v1.3.0) analysis. They were (1) British in England and Scotland (GBR), of the European ancestry (EUR) superpopulation, (2) Yoruba in Ibadan, Nigeria (YRI), of the African ancestry (AFR) superpopulation, (3) Indian Telugu in the UK (ITU), of the South Asian ancestry (SAS) superpopulation, and (4) Han Chinese South (CHS), of the East Asian ancestry (EAS) superpopulation. These singular population samples were chosen to broadly represent each of their four respective continental (superpopulation) ancestry groups, with an average population differentiation (Fst, or fixation index) value of 0.1055 amongst them, as estimated by ADMIXTURE. The supervised ADMIXTURE analysis provides, for each UKBB sample, a proportion of ancestry for each of the four reference populations. Those individuals with at least 80% of their ancestry attributed to one continental ancestry group, or 1KG defined superpopulation, were carried forward into further analyses.

Derivation of continental principal components

Unrelated individuals in each CAG, including both 1KG and UKBB samples with >=80% ancestry to that CAG were identified (using all 718,487 SNPs in the overlapping data set, and the plink (v1.9) function --rel-cutoff and a minor allele frequency (MAF) filter of 0.05 (--maf 0.05)). Then for each CAG and using all (1KG + UKBB) unrelated individuals assigned to the CAG, a list of approximately 40 thousand LD independent SNPs were identified (using the plink (v2.0) function --indep-pairwise 50 10 0.025 (--indep-pairwise 50

10 0.02 for AFR and --indep-pairwise 50 10 0.05 for SAS) along with a MAF filter of 0.01, and the exclusion of the 24 previously identified genomic regions with extensive linkage disequilibrium [63,64]). New plink files including only the LD independent SNPs identified in step two were subsequently generated. smartrel from the EIGENSOFT (https://github.com/DReichLab/EIG) package was used to generate a new list of related individual pairs, along with our script "greedy_unrelated_selection.R" to identify a list of related individuals to exclude from principal component derivation [29,30]. An exception this step was made for the European CAG as its sample-size was prohibitively large to run smartrel, instead the list of unrelated individuals generated from step one was used. Finally, smartpca of the EIGENSOFT package was used to estimate principal components (PC), using only unrelated UKBB samples. Related and 1KG samples were subsequently projected upon these PCs by smartpca. Sample outliers were excluded from the PC analysis by smartpca with the following parameters: using 10 PCs to identify outliers (numoutlierevec), at six standard deviations from the mean (outliersigmathresh), and with 5 outlier removal iterations (numoutlieriter). **Supplementary Table 1** provides numbers for each of these steps, for each CAG. The EUR CAG was treated uniquely due to its larger sample-size. Smartpca was run twice as described above, once with "fastmode=NO" and then with "fastmode=YES". The former provided estimates of the eigenvalues but not the eigenvectors, while the latter provided eigenvectors but not eigenvalues.

K-means clustering of principal components

For each CAG, we estimated the variance explained by each principal component (PC) by dividing the eigenvalue of each PC by the sum of all eigenvalues. To identify the number of top PCs we generated a scree plot, using the variance explained estimates, and identified the elbow or valley in each plot (**Supplementary Figure 1**, **Supplementary Table**

2). The top PCs, and the top PCs only, were then used in an unsupervised K-means clustering analysis (k set from 2 to 20; using the function "kmeans()" from the R stats package) to identify clusters of UKBB individuals that maximize between cluster sums of squares and minimize within cluster sums of squares. An optimum number of clusters (k) was identified by silhouette analysis using the function "pamk()" from the fpc R package (**Supplementary Figure 2**) [65]. These analyses are implemented in our function "DetermineK()" found in this study's GitHub repository.

Correspondence analysis

Each UKBB study participants' country of birth information was placed into United Nations defined geographic regions (**Supplementary Table 3**). To determine if the K-means population clusters have any relationship with an individual's country of birth or country of birth UN-region we performed correspondence analyses (CAs) using the function "ca()" from the R package "ca", for each continental ancestry group [36]. In addition, a chi-square test was performed on the contingency table used in the correspondence analysis. Any UN-region or country of birth with fewer than 10 observations was excluded. Individuals for which country of birth information was not available were also excluded.

Population differentiation among K-means population clusters

For each CAG, we took the best K-means population clusters, as defined by the silhouette analysis, and re-ran smartpca. However, on this run we had smartpca provide for us only an estimation of the average fixation index (Fst) for each pair of populations in the data set, including 1KG populations and UKBB K-means clusters. This was done with the inclusion of the paramaters "fstonly" and "phylipoutname" [55], the latter of which provides a distance matrix of mean Fst values between populations. Estimations of Fst, which range

from 0 to 1, provide a measure of population differentiation among populations. In brief, these describe the proportion of total variation at a SNP that is explained by variation between populations. For any SNP a value of 0 would indicate that minimal variation is attributable to variation between populations. A value of 1 would indicate a fixed difference i.e., the two populations are both invariable but for alternative alleles.

List of abbreviations

1KG = 1000 Genomes Project

ACRC = Advanced Computing Research Centre

AFR = African

AMR = Americas

BEB = Bengali in Bangladesh

CA = correspondence analysis

CAG = continental ancestry group

CDX = Chinese Dai in Xishuangbanna, China

CEU = Utah residents with Northern and Western European ancestry

CHB = Han Chinese in Beijing, China

CHS = Han Chinese South

COB = country of birth

EAS = East Asian

ESN = Esan in Nigeria

EUR = European

FIN = Finnish in Finland

Fst = fixation index

GBR = British in England and Scotland

GIH = Gujarati Indian in Houston, Texas

514 GWAS = Genome-wide association study

515 GWD = Gambian in Western Division, The Gambia - Mandinka

516 IBS = Iberian populations in Spain

517 ITU = Indian Telugu in the UK

518 JPT = Japanese in Tokyo, Japan

519 KHV = Kinh in Ho Chi Minh City, Vietnam

520 LD = linkage disequilibrium

521 LWK = Luhya in Webuye, Kenya

522 MAF = minor allele frequency

523 MSL = Mende in Sierra Leone

524 PC = principal component

525 PJL = Punjabi in Lahore, Pakistan

526 ROB = region of birth

527 SAS = South Asian

528 SNP = single-nucleotide polymorphism

529 STU = Sri Lankan Tamil in the UK

530 TSI = Toscani in Italia

531 UKBB = UK Biobank

532 UN = United Nations

533 YRI = Yoruba in Ibadan, Nigeria

534

535 References

536 [1] Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic

537 architecture: the shape of the genetic contribution to human traits and disease. Nat Rev

538 Genet 2017 192 2017;19:110–24. https://doi.org/10.1038/nrg.2017.101.

[2] Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell 2019;177:26–31. https://doi.org/10.1016/J.CELL.2019.02.048.

[3] Bentley AR, Callier SL, Rotimi CN. Evaluating the promise of inclusion of African ancestry populations in genomics. Npj Genomic Med 2020 51 2020;5:1–9. https://doi.org/10.1038/s41525-019-0111-x.

[4] Cooke Bailey JN, Bush WS, Crawford DC. Editorial: The Importance of Diversity in Precision Medicine Research. Front Genet 2020;0:875. https://doi.org/10.3389/FGENE.2020.00875.

[5] Green ED, Gunter C, Biesecker LG, Di Francesco V, Easter CL, Feingold EA, et al. Strategic vision for improving human health at The Forefront of Genomics. Nat 2020 5867831 2020;586:683–92. https://doi.org/10.1038/s41586-020-2817-4.

[6] Consortium TH. Enabling the genomic revolution in Africa: H3Africa is developing capacity for health-related genomics research in Africa. Science 2014;344:1346. https://doi.org/10.1126/SCIENCE.1251546.

[7] Matise TC, Study for the P, Ambite JL, Study for the P, Buyske S, Study for the P, et al. The Next PAGE in Understanding Complex Traits: Design for the Analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. Am J Epidemiol 2011;174:849–59. https://doi.org/10.1093/AJE/KWR160.

[8] Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nat 2021 5907845 2021;590:290–9. https://doi.org/10.1038/s41586-021-03205-y.

[9] Gallo LC, Penedo FJ, Carnethon M, Isasi C, Sotres-Alvarez D, Malcarne VL, et al. The Hispanic Community Health Study/Study of Latinos Sociocultural Ancillary Study: Sample, Design, and Procedures. Ethn Dis 2014;24:77.

[10] Investigators TA of URP. The “All of Us” Research Program.

Https://DoiOrg/101056/NEJMsr1809937 2019;381:668–76. https://doi.org/10.1056/NEJMSR1809937.

[11] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, 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 Med 2015;12:e1001779–e1001779. https://doi.org/10.1371/journal.pmed.1001779.

[12] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562:203–9. https://doi.org/10.1038/s41586-018-0579-z.

[13] Mathieson I, Scally A. What is ancestry? PLOS Genet 2020;16:e1008624. https://doi.org/10.1371/JOURNAL.PGEN.1008624.

[14] Rodriguez S, Gaunt TR, Day INM. Hardy-Weinberg Equilibrium Testing of Biological Ascertainment for Mendelian Randomization Studies. Am J Epidemiol 2009;169:505–14. https://doi.org/10.1093/AJE/KWN359.

[15] Graffelman J, Weir BS. On the testing of Hardy-Weinberg proportions and equality of allele frequencies in males and females at biallelic genetic markers. Genet Epidemiol 2018;42:34–48. https://doi.org/10.1002/GEPI.22079.

[16] Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061–73. https://doi.org/10.1038/nature09534.

[17] Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. Science (80-) 2002;298:2381–5. https://doi.org/10.1126/SCIENCE.1078311/SUPPL_FILE/ROSENBERG.SOM.PDF.PDF.

[18] Berezovskiĭ ND, Giria VN. [Estimation of combining ability of specialized types of

the big white breed]. Tsitol Genet 1991;25:56–60.

[19] Serre D, Pääbo S. Evidence for Gradients of Human Genetic Diversity Within and Among Continents. Genome Res 2004;14:1679. https://doi.org/10.1101/GR.2529604.

[20] Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. PLoS Genet 2005;1:e70. https://doi.org/10.1371/JOURNAL.PGEN.0010070.

[21] Birney E, Inouye M, Raff J, Rutherford A, Scally A. The language of race, ethnicity, and ancestry in human genetic research n.d.

[22] Peterson RE, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, et al. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. Cell 2019;179:589–603. https://doi.org/10.1016/J.CELL.2019.08.051.

[23] Przeworski M, Wall JD. Why is there so little intragenic linkage disequilibrium in humans? Genet Res 2001;77:143–51. https://doi.org/10.1017/S0016672301004967.

[24] Ptak SE, Voelpel K, Przeworski M. Insights into recombination from patterns of linkage disequilibrium in humans. Genetics 2004;167:387. https://doi.org/10.1534/GENETICS.167.1.387.

[25] Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature 2015;526:68–74. https://doi.org/10.1038/nature15393.

[26] Lander ES, Schork NJ. Genetic dissection of complex traits. Science (80-) 1994;265:2037–48. https://doi.org/10.1126/SCIENCE.8091226.

[27] Vilhjálmsson BJ, Nordborg M. The nature of confounding in genome-wide association studies. Nat Rev Genet 2012 141 2012;14:1–2. https://doi.org/10.1038/nrg3382.

614 [28] Loh P-R, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for
615 biobank-scale datasets. Nat Genet 2018 507 2018;50:906–8.
616 https://doi.org/10.1038/s41588-018-0144-6.
617 [29] Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet
618 2006;2:2074–93. https://doi.org/10.1371/journal.pgen.0020190.
619 [30] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal
620 components analysis corrects for stratification in genome-wide association studies. Nat
621 Genet 2006;38:904–9. https://doi.org/10.1038/ng1847.
622 [31] Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et
623 al. Efficient Bayesian mixed-model analysis increases association power in large
624 cohorts. Nat Genet 2015;47:284–90. https://doi.org/10.1038/ng.3190.
625 [32] Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population
626 stratification in genome-wide association studies. Nat Rev Genet 2010;11:459.
627 https://doi.org/10.1038/NRG2813.
628 [33] Zaidi AA, Mathieson I. Demographic history mediates the effect of stratification on
629 polygenic scores. Elife 2020;9:1–30. https://doi.org/10.7554/ELIFE.61548.
630 [34] Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, Budu-Aggrey A, et al.
631 Apparent latent structure within the UK Biobank sample has implications for
632 epidemiological analysis. Nat Commun 2019;10. https://doi.org/10.1038/S41467-018-
633 08219-1.
634 [35] Abdellaoui A, Hugh-Jones D, Yengo L, Kemper KE, Nivard MG, Veul L, et al.
635 Genetic correlates of social stratification in Great Britain. Nat Hum Behav 2019 312
636 2019;3:1332–42. https://doi.org/10.1038/s41562-019-0757-5.
637 [36] Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, et al. Correlation
638 between Genetic and Geographic Structure in Europe. Curr Biol 2008;18:1241–8.

https://doi.org/10.1016/J.CUB.2008.07.049.

[37] Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature 2008;456:98. https://doi.org/10.1038/NATURE07331.

[38] Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 2014;46:1173. https://doi.org/10.1038/NG.3097.

[39] Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK biobank. Elife 2019;8. https://doi.org/10.7554/eLife.39725.

[40] Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife 2019;8. https://doi.org/10.7554/eLife.39702.

[41] Barton N, Hermisson J, Nordborg M. Why structure matters. Elife 2019;8. https://doi.org/10.7554/ELIFE.45380.

[42] Diaz-Papkovich A, Anderson-Trocmé L, Ben-Eghan C, Gravel S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLoS Genet 2019;15. https://doi.org/10.1371/journal.pgen.1008432.

[43] Morton NE. Isolation by Distance. Genetics 1943;28:114. https://doi.org/10.1016/B978-0-12-374984-0.00820-2.

[44] Slatkin M. ISOLATION BY DISTANCE IN EQUILIBRIUM AND NON-EQUILIBRIUM POPULATIONS. Evolution 1993;47:264–79. https://doi.org/10.1111/J.1558-5646.1993.TB01215.X.

[45] Kidd JM, Gravel S, Byrnes J, Moreno-Estrada A, Musharoff S, Bryc K, et al. Population Genetic Inference from Personal Genome Data: Impact of Ancestry and

664 Admixture on Human Genomic Variation. Am J Hum Genet 2012;91:660.
665 https://doi.org/10.1016/J.AJHG.2012.08.025.
666 [46] Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P,
667 et al. Genomic Insights into the Ancestry and Demographic History of South America.
668 PLoS Genet 2015;11. https://doi.org/10.1371/JOURNAL.PGEN.1005602.
669 [47] Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, et
670 al. Reconstructing the Population Genetic History of the Caribbean. PLoS Genet
671 2013;9. https://doi.org/10.1371/JOURNAL.PGEN.1003925.
672 [48] Ongaro L, Scliar MO, Flores R, Raveane A, Marnetto D, Sarno S, et al. The Genomic
673 Impact of European Colonization of the Americas. Curr Biol 2019;29:3974-3986.e4.
674 https://doi.org/10.1016/J.CUB.2019.09.076/ATTACHMENT/8D05D549-D774-
675 4CBA-9BE7-94D3B60AD79D/MMC3.XLSX.
676 [49] Montinaro F, Busby GBJ, Pascali VL, Myers S, Hellenthal G, Capelli C. Unravelling
677 the hidden ancestry of American admixed populations. Nat Commun 2015;6.
678 https://doi.org/10.1038/NCOMMS7596.
679 [50] Geibel J, Reimer C, Weigend S, Weigend A, Pook T, Simianer H. How array design
680 creates SNP ascertainment bias. PLoS One 2021;16:e0245178–e0245178.
681 https://doi.org/10.1371/journal.pone.0245178.
682 [51] Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: Why
683 it is important, and how to correct it. BioEssays 2013;35:780–6.
684 https://doi.org/10.1002/bies.201300014.
685 [52] Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment biases in SNP chips affect
686 measures of population divergence. Mol Biol Evol 2010;27:2534–47.
687 https://doi.org/10.1093/molbev/msq148.
688 [53] Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient

Admixture in Human History. Genetics 2012;192:1065. https://doi.org/10.1534/GENETICS.112.145037.

[54] Lu Y, Patterson N, Zhan Y, Mallick S, Reich D. Technical design document for a SNP array that is optimized for population genetics n.d.

[55] Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature 2009;461:489–94. https://doi.org/10.1038/nature08365.

[56] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75. https://doi.org/10.1086/519795.

[57] Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 2015;4. https://doi.org/10.1186/s13742-015-0047-8.

[58] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19:1655–64. https://doi.org/10.1101/gr.094052.109.

[59] Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 2011;12:246. https://doi.org/10.1186/1471-2105-12-246.

[60] Core R Team. R: A Language and Environment for Statistical Computing. R Found Stat Comput 2019;2:https://www.R--project.org. http://www.r-project.org (accessed March 2, 2021).

[61] Mitchell RE, Hemani G, Dudding T, Corbin L, Harrison S, Paternoster L. UK Biobank Genetic Data: MRC-IEU Quality Control, version 2, 18/01/2019 n.d.

[62] Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature

2015;526:75–81. https://doi.org/10.1038/nature15394.

[63] Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna K V., et al. Long-Range LD Can Confound Genome Scans in Admixed Populations. Am J Hum Genet 2008;83:132. https://doi.org/10.1016/J.AJHG.2008.06.005.

[64] Weale ME. Quality Control for Genome-Wide Association Studies. In: Barnes MR, Breen G, editors. Genet. Var. Methods Protoc., Humana Press, New York, NY; 2010, p. 31.

[65] Batool F, Hennig C. Clustering with the Average Silhouette Width. Comput Stat Data Anal 2021;158:107190. https://doi.org/10.1016/J.CSDA.2021.107190.

Figure Legends

Figure 1 Ancestry estimates for the UKBB non-white British subset: Estimates of ancestry proportions for each UKBB participant previously labeled as non-white British individuals by UKBB. Ancestry was derived from a supervised ADMIXTURE analysis using four 1000 Genomes reference populations - Yoruba in Ibadan, Nigeria for (AFR) Africa, British in England, and Scotland for (EUR) Europe, Indian Telugu in the UK for (SAS) South Asia, and Han Chinese South for (EAS) East Asia.

Figure 2 Ancestry proportions on UKBB PCs: Continental (A) African, (B) European, (C) South Asian, and (D) East Asian ancestry proportions placed on principal components one and two, as supplied by the UK Biobank.

738 **Figure 3 UKBB continental PCs with 1000 Genomes populations:** Principal components
739 one through four for each CAG. UKBB samples are colored in gray, while the 1KG sub-
740 populations for each CAG are plotted in other colors, as indicated by each legend. The
741 proportion of variation explained by each PC is indicated on each axis.
742
743 **Figure 4 UKBB continental PCs with K-means clusters:** Principal components one
744 through four for each CAG with each individual colored by its assigned K-means population
745 cluster, as indicated by each legend. The proportion of variation explained by each PC is
746 indicated on each axis.
747
748 **Figure 5 Principal components for CAG with geographic regions of birth:** Principal
749 components one and two for each CAG, with (A-D) individuals colored by their region of
750 birth (A-D), and with (E-H) the PC center also colored by region of birth. PC centers were
751 estimated as the average PC1 and PC2 values for all individuals of that ROB. Regions of
752 birth are denoted in the figure legend, and the proportion of variation explained by each PC is
753 indicated on each axis.
754
755 **Figure 6 Correspondence analysis:** Correspondence plots between (A) K-means population
756 clusters (colored circles) and regions of birth (grey squares), and (B) K-means population
757 clusters (colored circles) and country of birth (grey squares) (B). The x and y axes are the
758 first and second dimension of each correspondence analysis, respectively, with the proportion
759 of variance explained indicated in the parentheses of each axis.
760
761 **Figure 7 Fst estimates:** The minimum, mean and maximum fixation index values for each
762 CAG in the 1KG project and the UK Biobank dataset. Fst values in the 1KG project are

763 between the sub-populations of each super-population, while UK Biobank estimates are
764 derived between K-means population cluster of each CAG.
765
766 **Figure 8 Graph outlining the possible effects of geographic structure in population**
767 **genetics:** Suppose one might want to use Mendelian randomization to study the relationship
768 between neutrophil count and severe malaria caused by P. *Falciparum* – a disease largely
769 absent in European environments. Using summary statistics from a neutrophil count GWAS
770 derived from individuals with European ancestry (Box 1A) may affect estimates due to
771 geographic structure (Ancestry + Demography + Environment). This can be overcome by
772 running a GWAS in people of African ancestry (Box 1B).
773
774 **Supplementary Figure 1 Continental ancestry PCA Scree plots:** Legend: Scree plots
775 illustrating the proportion of variation explained by each of the top 20 PCs, in each UKBB
776 continental ancestry principal component analyses. The Scree plots were used to identify the
777 number of top PCs to carry forward into the K-means clustering analysis. The continental
778 ancestry supergroups are Africa (AFR), Europe (EUR), South Asia (SAS), and East Asia
779 (EAS). The number of PCs selected as top PCs are AFR = 4, SAS = 5, EAS = 4, EUR = 5.
780 The horizontal line in each plot denotes where 10% variance explained is in each plot to aid
781 in inter-CAG comparisons.
782
783 **Supplementary Figure 2 K-means k selection with silhouette analysis:** Selection of an
784 optimum number of k clusters in the K-means analysis of the top PCs, by silhouette analysis.
785 A silhouette plot for each UKBB continental supergroup (AFR) African, (EUR) European,
786 (SAS) South Asian, and (EAS) East Asian is provided. The x-axis indicates the number of k
787 clusters evaluated, and the y-axis provides an estimate of the average silhouette width

(ASW). ASW is an estimation of cluster quality, or intra- and inter- cluster distances derived from a partitioning around medoids (PAM). The optimum number of k clusters in each UKBB continental supergroup were identified as AFR = 7, EUR = 2, SAS = 4, and EAS = 3.

Supplementary Figure 3 UKBB continental ancestry group PCs with K-means clusters: UK Biobank continental ancestry group PCs with K-means population clusters color coded: AFR (A), EUR (B), SAS (C), EAS (D).

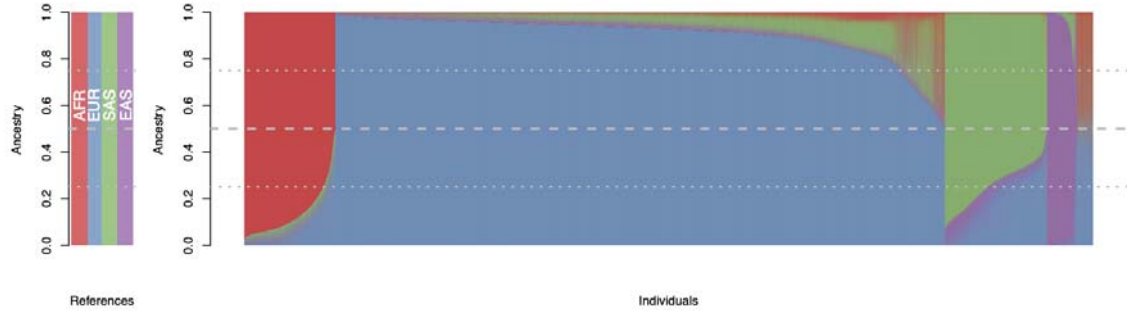
Supplementary Figure 4 Population structure by UN defined geographic region: UK Biobank continental ancestry group PCs colored by the region of birth color coded: AFR (A-C), EUR (D-F), SAS (G-I), EAS (J-L).

Supplementary Figure 5 Population structure centers, as defined by UN geographic region: UK Biobank continental ancestry group PCs 1-4 with region of birth centers (averaged across all individuals from each ROB) colour coded: AFR (A-C), EUR (D-F), SAS (G-I), EAS (J-L).

Supplementary Figure 6 Population structure by country of birth in AFR by region: UK Biobank continental ancestry group PCs 1-4 for the AFR CAG divided by UN regions of birth: Eastern (A), Central (B), Western (C), Northern/Southern (D). Samples are color coded by their country of birth.

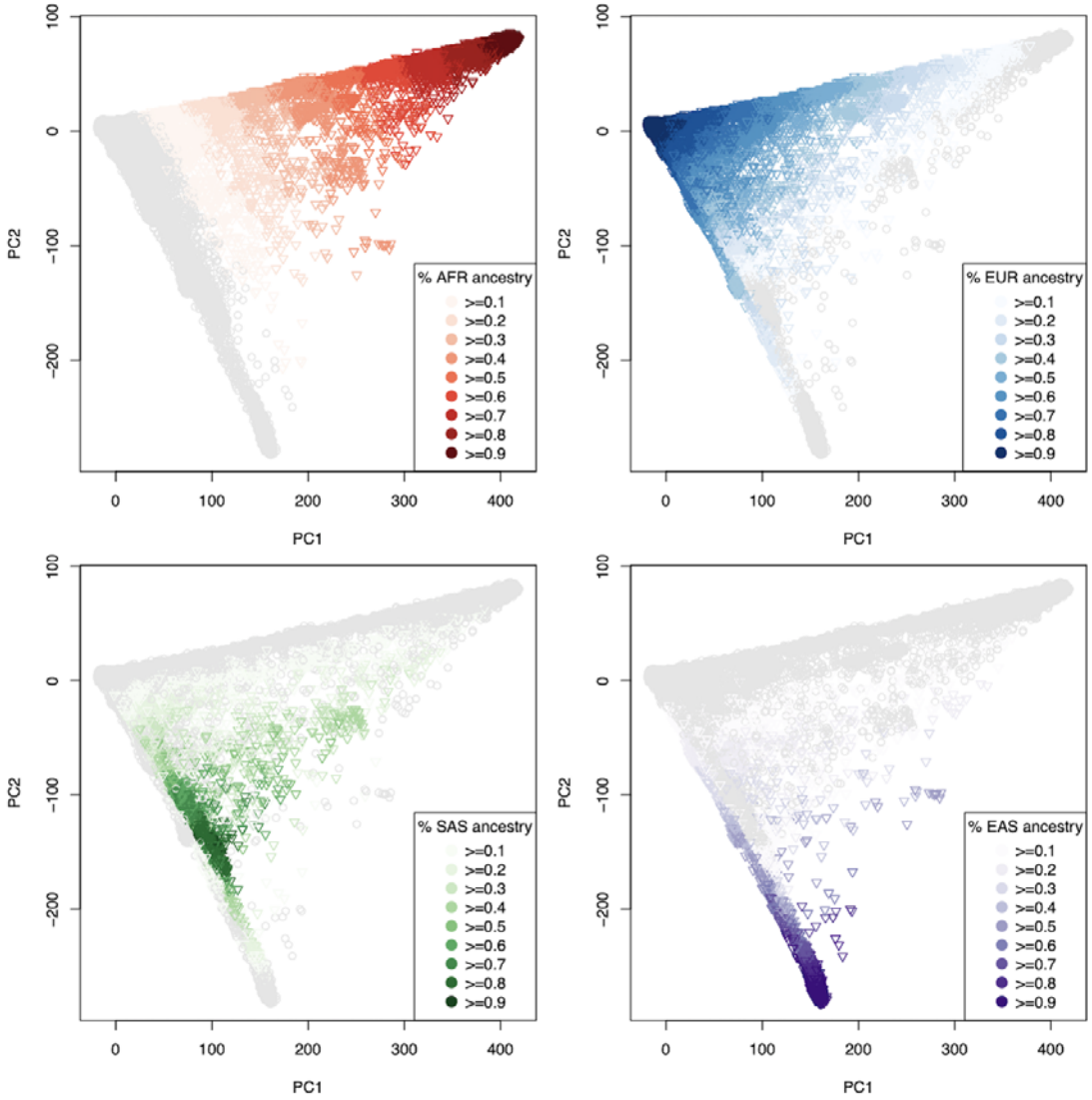
Supplementary Figure 7 Population structure by country of birth in EUR by region: UK Biobank continental ancestry group PCs 1-4 for the EUR CAG divided by UN regions of

812 birth: Northern (A), Eastern (B), Southern (C), Western (D). Samples are color coded by their
813 country of birth.
814
815 **Supplementary Figure 8 Population structure by country of birth in SAS and EAS:** UK
816 Biobank continental ancestry group PCs 1-4 for the SAS and EAS CAGs divided by UN
817 regions of birth: Southern Asia (A), East Asia and South-eastern Asia (B). Samples are color
818 coded by their country of birth.
819
820 **Supplementary Figure 9 Population structure centers by country of birth:** UK Biobank
821 continental ancestry group centers colored by the country of birth: AFR (A-C), EUR (D-F),
822 SAS (G-I), EAS (J-L).
823
824 **Supplementary Figure 10 Population structure centers by country of birth:** UK Biobank
825 continental ancestry group centers in the correspondence analysis, colored by K-means
826 population clusters, overlapping with country of birth data in grey: AFR (A-C), EUR (D-F),
827 SAS (G-I), EAS (J-L).
828 Figure 1 Ancestry estimates for the UKBB non-white British subset



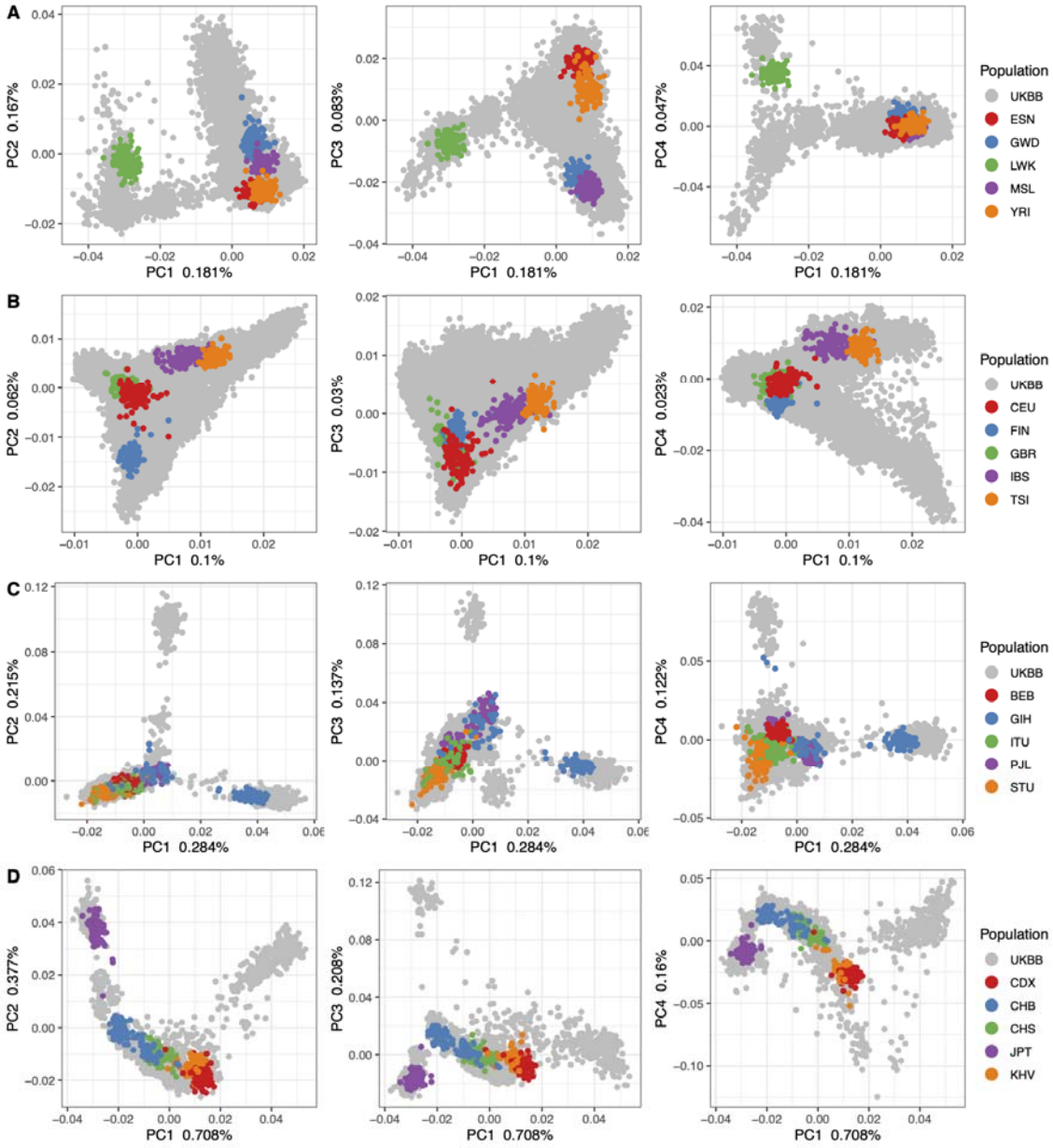
829
830

831 Figure 2: Ancestry proportions on UKBB PCs

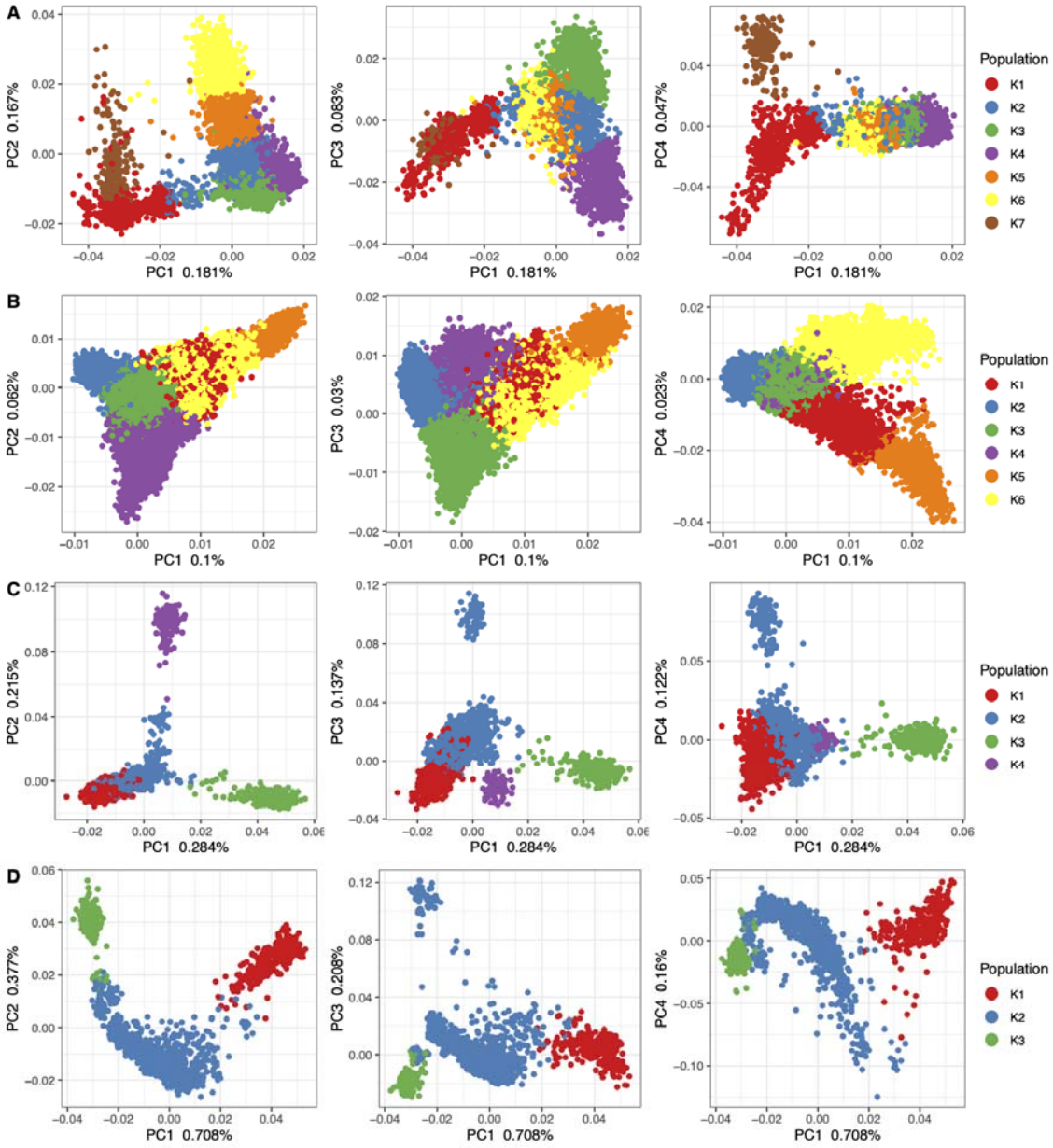


832
833

Figure 3: UKBB continental PCs with 1000 Genomes populations



836 Figure 4: UKBB continental PCs with K-means clusters



837
838

Figure 5: Principal components for CAG with geographic regions of birth.

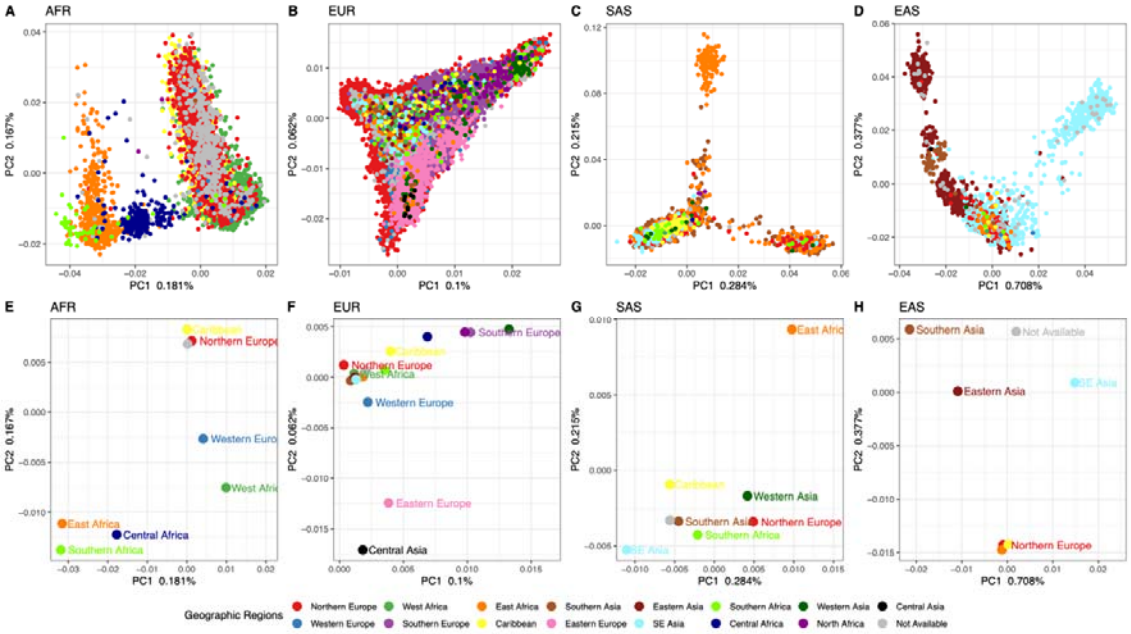


Figure 6: Correspondence analysis

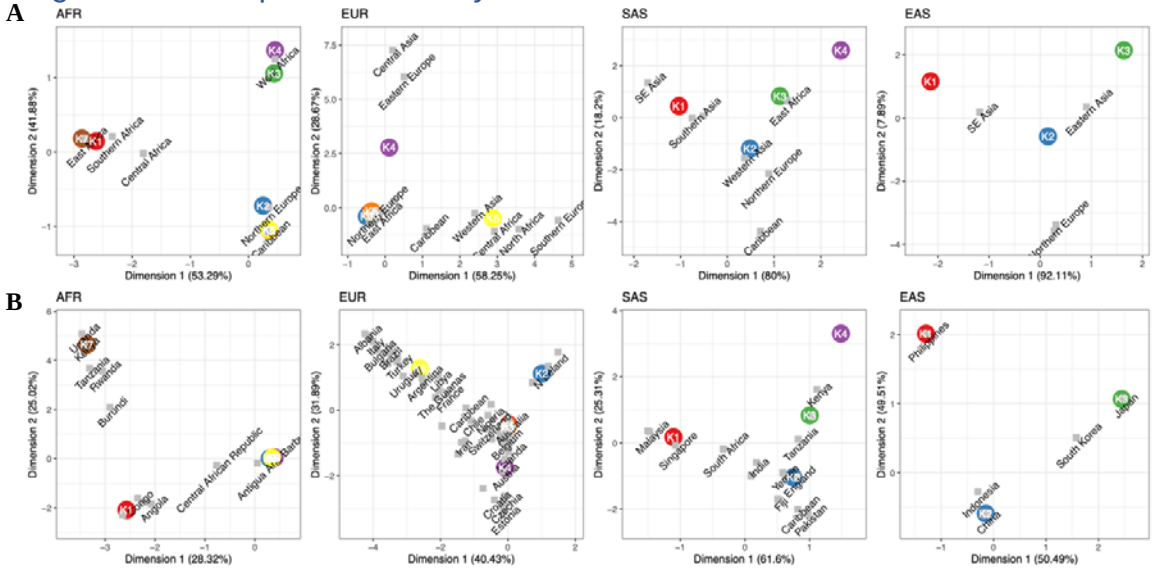


Figure 7: Fst estimates

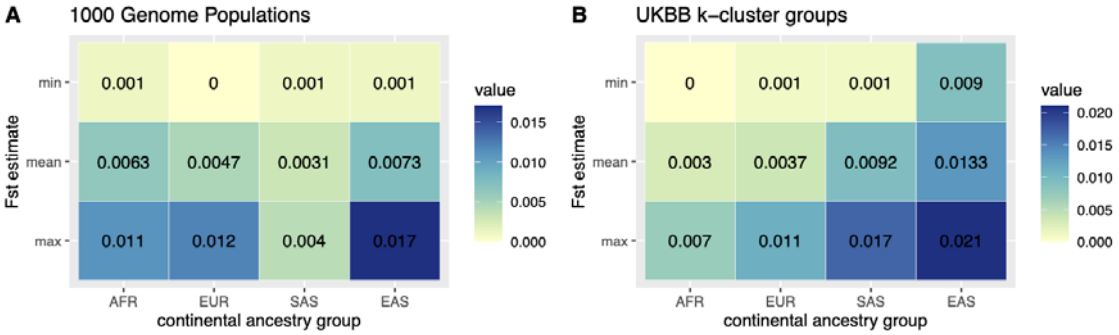


Figure 8: Graph outlining the possible effects of geographic structure in population genetics