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

Alex Diaz-Papkovich[a,b], **Luke Anderson-Trocmé**[b,c], **and Simon Gravel**[b,c,1]

[a]Department of Quantitative Life Sciences, McGill University, Montreal, QC, H3A 0G1 Canada; [b]McGill University and Genome Quebec Innovation Centre, Montreal, QC, H3A 0G1, Canada; [c]Department of Human Genetics, McGill University, Montreal, QC, H3A 0G1, Canada. [1]To whom correspondence should be addressed. E-mail: simon.gravel@mcgill.ca

Background: Population structure in genetic data depends on complex demographic processes including geographic isolation, genetic drift, migration, and admixture. Together with technical artifacts, population structure is a prominent confounder of genomic studies. Identifying such patterns is therefore central to the genomic enterprise. Whereas many methods can identify specific types of population structure, few are able to provide simple representations of genomic diversity across a range of scales. We investigate an approach to dimension reduction and visualization of genomic data that combines principal components analysis (PCA) with uniform manifold approximation and projection (UMAP).

Results: We demonstrate using genotype data from the 1000 Genomes Project, the Health and Retirement Study, and the UK Biobank that projections using PCA-UMAP effectively cluster individuals who are genetically closely related while placing them in a global continuum of genetic variation. These projections reveal non-trivial population groupings, reflect ethnicity and geography on fine-scale levels, and uncover patterns in the distributions of a variety of phenotypes.

Conclusion: These projections succinctly illustrate population structure in large cohorts and capture relationships on local and global scales, establishing PCA-UMAP as a general-purpose approach to exploratory analysis in genomics.

genomics | genetics | population structure | ethnicity | ancestry | machine learning | data visualization | dimension reduction

Background. Questions in medicine, anthropology, and related fields hinge on interpreting the deluge of genomic data provided by modern high-throughput sequencing technologies. Because genomic datasets are high-dimensional, their interpretation requires statistical methods that can comprehensively condense information in a manner that is understandable to researchers and minimizes the amount of data that is sacrificed. Both model-based and model-agnostic approaches to summarize data have played important roles in shaping our understanding of the evolution of our species (**?**).

Here we will focus on nonparametric approaches to visualize relatedness patterns among individuals within populations. If we consider unphased single nucleotide polymorphism (SNP) data, an individual genome can be represented as a sequence of integers corresponding to the number of derived alleles carried by the individual at each of the L SNPs for which genotypes are available, with L typically larger than $100,000$. Since each individual is represented as an L-dimensional vector, dimension reduction methods are needed to visualize the data.

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

As sample sizes increase, the amount of information encoded in the lower-variance principal components increases, and researchers typically examine multiple two-dimensional projections to get a sense of the data. While many features of the data can be identified in this manner, other features may be hidden by the projections or hard to interpret.

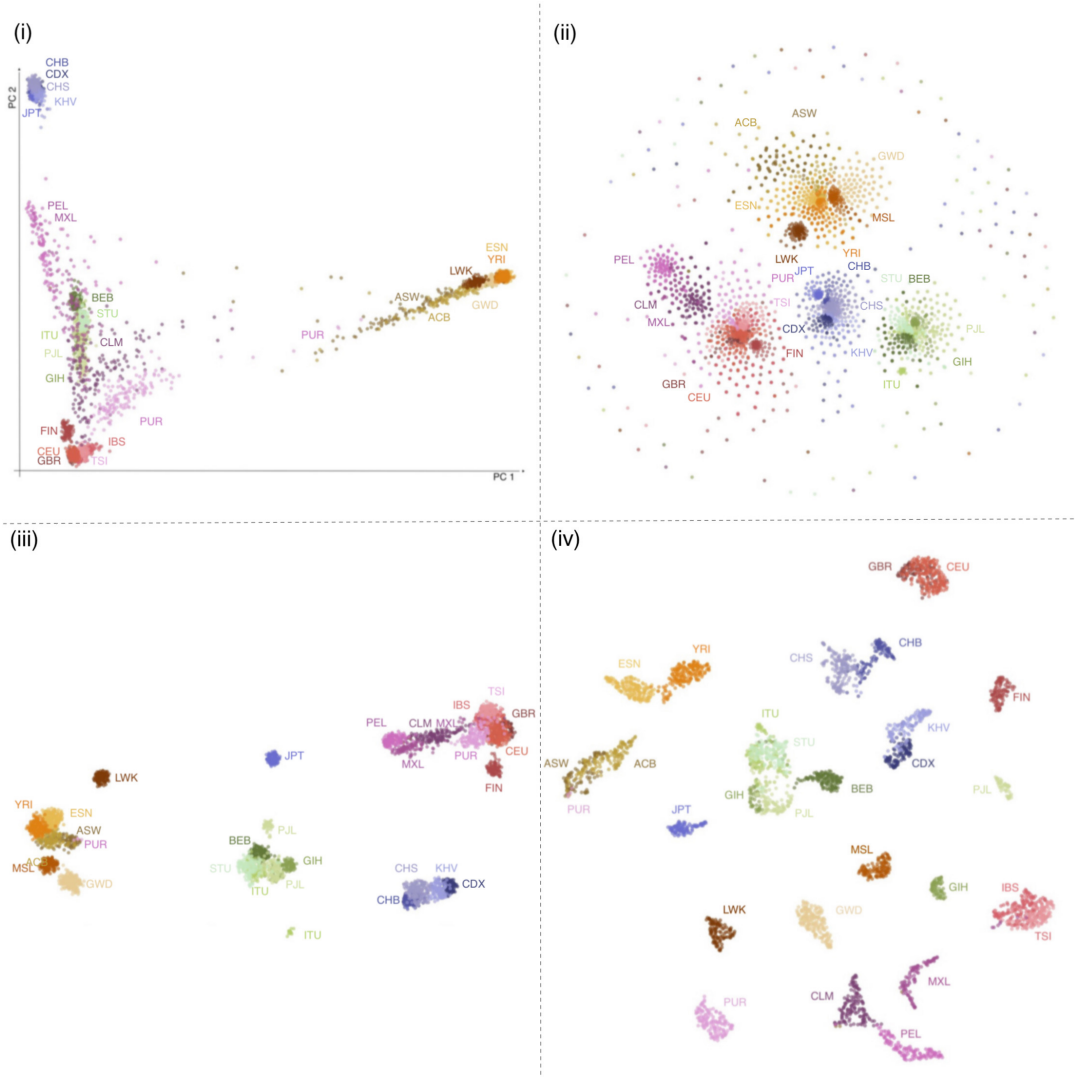


Fig. 1. Four methods of dimension reduction of 1KGP genotype data with population labels (i) PCA maps individuals in a triangle with vertices corresponding to African, Asian/Native American, and European continental ancestry. Discarding lower-variance PCs leads to overlap of populations with no close affinity, such as Central and South American populations with South Asians. (ii) t-SNE forms groups corresponding to continents, with some overlap between European and Central and South American people. Smaller subgroups are visible within continental clusters. The cloud of peripheral points results from the method's poor convergence. (iii) UMAP forms distinct clusters related to continent with clearly defined subgroups. Japanese, Finnish, Luhya, and some Punjabi and Telugu populations form separate clusters consistent with their population history(**?**). (iv) UMAP on the first 15 principal components forms fine-scale clusters for individual populations. Groups closely related by ancestry or geography, such as African Caribbean/African American, Spanish/Italian, and Kinh/Dai populations cluster together. Results using t-SNE on principal components are presented in figure **??**. Axes in UMAP and t-SNE are arbitrary. ACB, African Caribbean in Barbados; ASW, African Ancestry in Southwest US; BEB, Bengali; CDX, Chinese Dai; CEU, Utah residents with Northern/Western European ancestry; CHB, Han Chinese; CHS, Southern Han Chinese; CLM, Colombian in Medellin, Colombia; ESN, Esan in Nigeria; FIN, Finnish; GBR, British in England and Scotland; GWD, Gambian; GTH, Gujarati; IBS, Iberian in Spain; ITU, Indian Telugu in the UK; JPT, Japanese; KHV, Kinh in Vietnam; LWK, Luhya in Kenya; MSL, Mende in Sierra Leone; MXL, Mexican in Los Angeles, California; PEL, Peruvian; PJL, Punjabi in Lahore, Pakistan; PUR, Puerto Rican; STU, Sri Lankan Tamil in the UK; TSI, Tuscani in Italy; YRI, Yoruba in Nigeria

optimal configurations make convergence to a globally satisfying solution difficult.

Uniform Manifold Approximation and Projection (UMAP) is a new dimension reduction technique grounded in Riemannian geometry, algebraic topology, and category theory, and designed to model and preserve the high-dimensional topology of data points in the low-dimensional space (**?**). The assumption behind UMAP is that data are uniformly distributed on local manifolds in high-dimensional space, which can be approximated as fuzzy sets that are patched together to form a

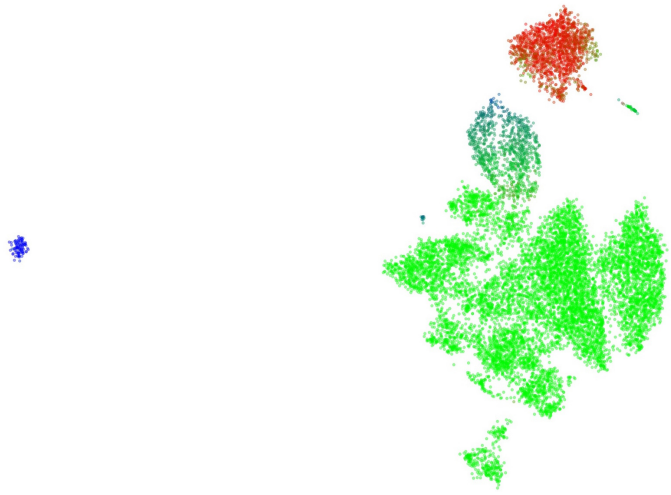


Fig. 2. ADP: Using a new image here UMAP on the first 10 principal components of HRS data. Coloring individuals by estimated admixture from three ancestral populations reveals considerable diversity in the Hispanic population. This projection colored by self-identified race and Hispanic status is presented in figure **??**. Individuals with intermediate admixture ratios sometimes fall between clusters of Black, White, and Hispanic individuals (alternate projection shown in figure **??**).

topological representation. One can then construct a low-dimensional topological representation that minimizes the differences between the two representations. With genotype data, UMAP creates a neighbourhood around each individual's genetic coordinates and identifies a pre-selected number of neighbours to build high-dimensional manifolds. The end result is a low-dimension representation that groups genetically similar individuals together on a local scale while preserving long-range topological connections to more distantly related individuals.

A common practice in dimensional reduction is to first apply PCA to reduce the number of dimensions before performing nonlinear dimensional reduction. In addition to being computationally advantageous, this discards statistical noise that can confound nonlinear approaches: Population structure arising from n isolated randomly-mating demes can be described by the leading $n-1$ PCs, with the following PCs describing stochastic variation in relatedness (**?**). Selecting the leading PCs therefore has potential to extract meaningful population structure while filtering out stochastic noise.

We explore different strategies to pre-process the data and investigate discrete and continuous population structure patterns present in large datasets of human genotypes: the 1KGP, the Health and Retirement Study (HRS)(**?**), and the UK BioBank (UKBB)(**?**).

Results.

Fine-scale visualization of the 1KGP dataset. The 1KGP contains genotype data of 3,450 individuals from 26 relatively distinct labeled populations(**?**). Figure **??** shows visualizations using PCA, t-SNE, UMAP, and PCA-UMAP (that is, UMAP with PCA pre-processing). Using UMAP and t-SNE on the genotype data presents clusters that are roughly grouped by continent, with UMAP showing a clear hierarchy of population and continental clusters, whereas t-SNE fails to assign many individuals to population clusters. Using either on the top principal components leads to more distinct population clusters and less defined continental structure (see figure **??** for PCA-tSNE). Adding more components results in progressively finer clusters until approximately 20 populations appear using 15 components; further components gradually approach results similar to using the entire genotype data (see figures **??** and **??**).

Fig. 3. The top 7 principal components of the Hispanic population of the HRS projected by UMAP, colored by region of birth. The highlighted region consists almost entirely of individuals who were born in the Mountain region of the United States. This region contains New Mexico, Arizona, Colorado, Utah, Nevada, Wyoming, Idaho, and Montana. Figure **??** presents the same figure colored according to estimated continental ancestry proportion. Figure **??** shows that populations from the 1KGP do not map to the cluster when projected onto the UMAP embedding.

Focusing on PCA-UMAP with 15 principal components (figure **??** (iv)), we also find several population clusters that reflect shared ancestries. British individuals from England and Scotland form a cluster mixed with those from Utah who claim Northern and Western European ancestry. Toscani and Iberian individuals form a group reflecting their Mediterranean heritage. African Americans in the Southwest US, African Caribbean individuals in Barbados, and some Puerto Ricans also form a cluster. The East Asian super-population forms three sub-populations split by geography: one is largely Han and Southern Han individuals, another is comprised of the Chinese Dai in southern China and the Kinh from Vietnam, and the third is the Japanese population. Looser geographical groupings include Colombians and Peruvians, and the Esan and Yoruba populations of Nigeria; both groupings appear as connected sub-clusters. The South Asian super-population also forms a loose grouping.

Only a few individuals cluster differently than the majority of individual bearing the same population label: a few Mexican individuals cluster with Spanish and Italian individuals, and a few Puerto Ricans cluster with the African Americans and African Caribbeans, likely resulting from ancestry proportions that differ from the majority. One Gambian-identified individual is present in a cluster that is otherwise entirely Mende people from Sierra Leone. Only two populations form multiple clusters: Gujarati Indians in Houston, Texas and Punjabi people in Lahore, Pakistan. This clustering is robust to, e.g., the choice of the number of PCs considered (see figure **??**).

Finally, contrasting UMAP and t-SNE, we find that UMAP preserves more of the global structure of the data than t-SNE, and is more robust to choices of data pre-processing (figure **??**).

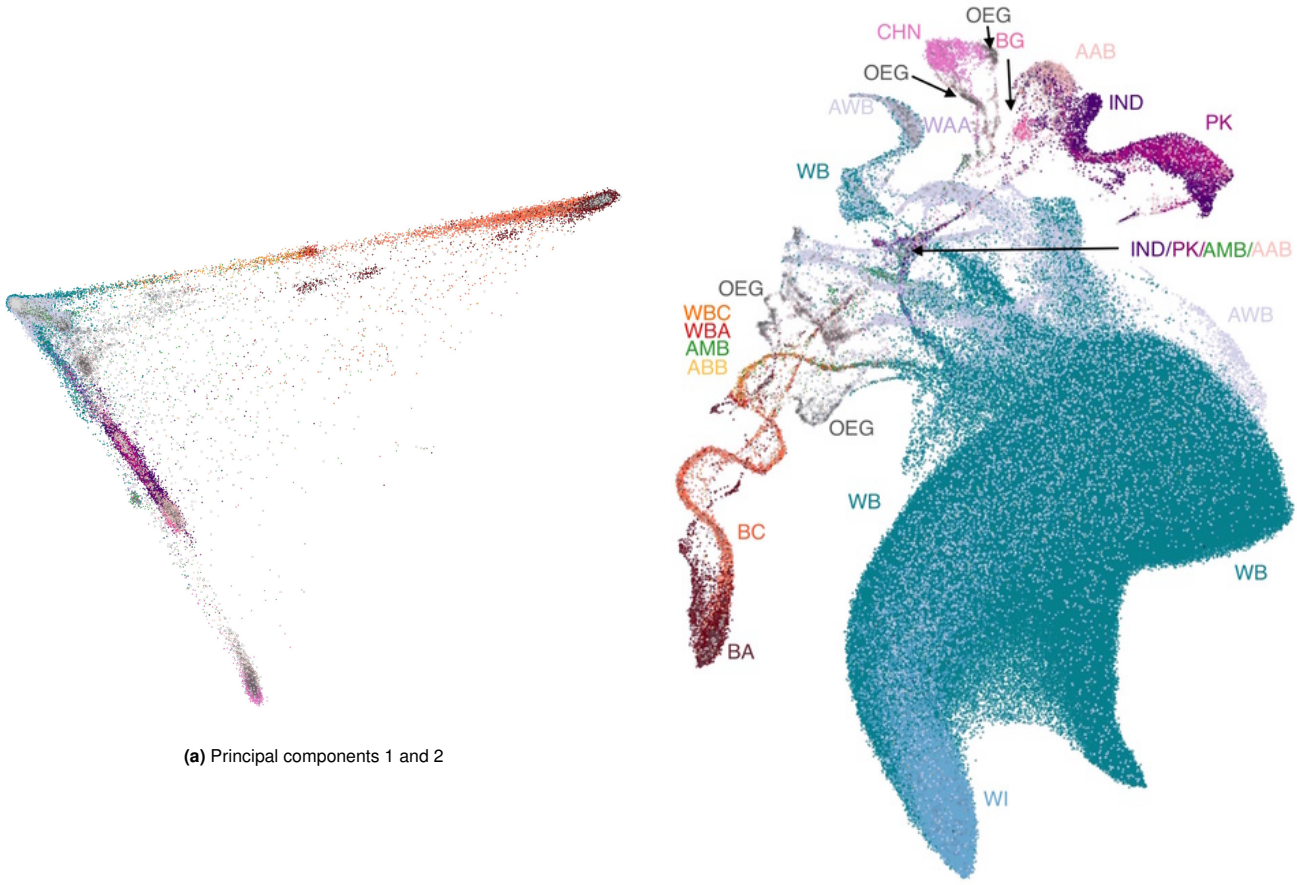
The genetic continuum of admixed populations. The 1KGP sampled individuals from relatively distinct population groups across the world, which makes the data particularly easy to cluster. Most medical cohorts comprise larger numbers of individuals sampled across extended geographical areas.

For example, the HRS contains genotype data of 12,454 American individuals across all 50 states who have provided racial identity (10,434 White, 1,652 Black, 368 Other) as well as whether they identify as Hispanic (1,203 total) and, if so, whether they identify as Mexican-American (705 total)(**?**). We crossed these three variables to form a composite self-reported ethnicity resulting in 10 categories (e.g. White Hispanic Mexican-American), and considered birth regions based on the 10 census regions and divisions used by the US Census Bureau. Admixture proportions for each individual were estimated in (**?**) by assuming ancestral African, Asian/Native American, and European populations using RFMIX (**?**). We have scaled these three proportions to values between 0 and 255, to color individual points by their estimated admixture represented by RGB where red, green, and blue respectively correspond to African, European, and Asian/Native American ancestry. Using the first 10 principal components and UMAP, we demonstrate projections that present a collection of sub-populations and a continuum of genetic variation.

ADP: Talking about using HRS with 1KGP here. Changed "four" clusters to "several" The HRS forms several large groupings and clusters, reflecting both ethnicity (figure **??**) and admixture proportions (figure **??**). Gradients in admixture proportion are clearly visible within the predominantly Hispanic cluster, but not within the predominantly Black cluster, perhaps because the variance in ancestry proportions is greater among Hispanics. The "White Not Hispanic" (WNH) group forms several interconnected clusters, and these do not correspond to broad geographical areas (figure **??**). The clarity of the interconnected clusters varies by parameterization, but they consistently form a large, roughly connected group.

To investigate possible ancestries related to populations in the 1KGP we took two approaches. In the first we generated PC axes and a UMAP embedding using UMAP and 1KGP data together (figure **??**). In the second we used the PC axes and UMAP embedding generated in figure **??** and projected 1KGP data onto it (figure **??**). In the latter case, we again see a gradient in the Hispanic cluster, though this time with ethnicity rather than admixture. In both cases we see groupings of Finnish individuals within the WNH groups, as well as Italian and Spanish individuals grouping near the White Hispanic population. One group of WNH individuals regularly appears at the periphery of the main cluster and does not cluster with any 1KGP populations.

Regional patterns in the American Hispanic population. In contrast to the WNH individuals, applying PCA-UMAP to self-identified Hispanic individuals reveals clear groupings related to birth region. One separate cluster, highlighted in figure **??**, consists almost entirely of individuals born in the Mountain Region of the United States. This cluster is not apparent when looking at a grid of pairwise plots of the first 8 principal components, provided in figure **??**, as the signal is distributed along PCs 3, 4, and 6. Even though continental admixture patterns do correlate with UMAP position (figure **??**), these do not explain the Mountain Region cluster. Individuals from 1KGP populations do not appear in the cluster when projected to the UMAP embedding (figure **??**).ADP: Added the 1KGP note here The cluster possibly comprises the Hispano population of the Southwest US, who have been present in the Mountain Region area long before the more recent immigrants from Latin America, and whose ancestry is expected to reflect both distinct Native ancestry and population-specific drift relative to other Hispanic populations. A recent preprint discusses the Mountain Region Hispanics and provides a more detailed historical description (**?**).



(a) Principal components 1 and 2

(b) UMAP on first 10 principal components

Fig. 4. The UKBB projected onto two dimensions, colored by self-reported ethnic background. (a) The first two principal components, showing the usual triangle with vertices corresponding to African, Asian/Native American, and European ancestries, and intermediate values indicating admixture or lack of relationship to the vertex populations. (b) UMAP on the first 10 principal components. The cluster of White British and White Irish individuals is greatly expanded, with the Irish forming a distinct sub cluster mixed with the White British population. South Asian and East Asian individuals form their separate clusters, as do individuals of African or Caribbean backgrounds. Population clusters are connected by "trails" comprised of large proportions of individuals with mixed backgrounds. BA, Black African; BC, Black Caribbean; BG, Bangladeshi; CHN, Chinese; IND, Indian; PK, Pakistani; WB, White British; WI, White Irish; WBC, White and Black Caribbean; WBA, White and Black African; WAA, White and Asian; AAB, Any other Asian Background; ABB, Any other Black Background; AWB, Any other White Background; AMB, Any other Mixed Background; OEG, Other ethnic group.

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

UMAP on the top 10 principal components reveals both continuous and discrete population structure (figure **??**): The patchwork of local topologies identifies continuous structure within the British population as well as admixture gradients despite the very unbalanced population sizes. The result is a comprehensive portrait of genetic variation capturing population relationships not visible using other methods, succinctly illustrating the complex structure of large and multi-ethnic datasets.

The largest body in the figure consists of the White British and Irish populations. The Irish population concentrates in a portion of this group, but many individuals are also scattered throughout the British-identified population. Individuals identifying as Black African and Black Caribbean partially overlap, but admixed individuals form distinct trails leading to Asian and European

clusters. Chinese individuals form a cluster, within what is suspected to be a broader East Asian super-population; Indian, Pakistani, and Bangladeshi populations form a closely bound group as well. The East and South Asian super-populations each have large clusters of individuals who identify as having an "other Asian background" or belonging to an "other ethnic group". The patchwork of genetic neighbourhoods is connected by trails of admixed individuals. These trails come together in a nexus of individuals with a variety of ethnicities; many claim mixed ancestry, and there are clear groups of individuals who belong to an "other ethnic group". Although their ethnicities are unknown to us, given their proximity to African, South Asian, and White individuals, possible candidates for these groups are North African, Middle Eastern, and West Asia backgrounds. Additionally, there are many individuals whose ethnicity is White but neither British nor Irish (AWB) forming clusters distinct from the British and Irish cluster.

Figure **??** presents the projection in figure **??** colored in by geographical coordinates according to the Ordnance Survey National Grid (OSGB1936) as distances north and east of the Isles of Scilly. UMAP coordinates within the "White British" cluster broadly map to geographic coordinates, as has been observed in Europe-wide data(**?**). Most admixture lines connect to the South East corner of this cluster, corresponding to the position of the city of London and reflecting its high migrant population.

The detailed shape of extended clusters is not stable as we vary the number of PCs included. Figure **??** shows a UMAP plot using the top 20 PCs from the UKBB. The shape of the "White British Cluster" is notably different, and we observe finer patterns of geographic variation, yet the qualitative observations made above are maintained. As an alternate visualization of diversity's correlation with geography, we performed a 3D UMAP projection and converted the normalized UMAP values into RGB values, allowing us to plot individuals on a map of Great Britain, emphasizing both spatial gradients of genetic relatedness and increased diversity in urban centers (figure **??**). The patterns in rural areas observed are similar to those reported in (**?**) using the haplotype-based CHROMOPAINTER on British individuals whose grandparents lived nearby.

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

Patterns in phenotype related to population structure. Our involvement with the UK biobank data is through a project on autoimmune disease and asthma. More than in geographic coordinates, we are interested in whether genetic population structure correlates with phenotypes and covariates of interest.

Covariates such as height (figure **??**) and autoimmune and asthma-related measures (figures **??** to **??**) correlate strongly with both discrete and continuous population structure. Several populations in figure **??**, including South Asian, East Asian, African, and several unidentified ethnic groups have noticeably lower-than-average heights. More subtle patterns are also visible: the area of the projection in figures **??** with the cluster of White Irish people appears more blue than the main body of White British individuals; an unpaired two sample t test of self-identified White Irish and White British individuals reveals statistically significant differences in age-adjusted mean height between the populations, with British males being taller on average by 0.846cm (p-value 2.10×10^{-23}) and British females by 0.763cm (p-value 3.65×10^{-23}) (see figures **??** and **??** for boxplots). Height differences between Irish and British populations have been previously observed but the direction of

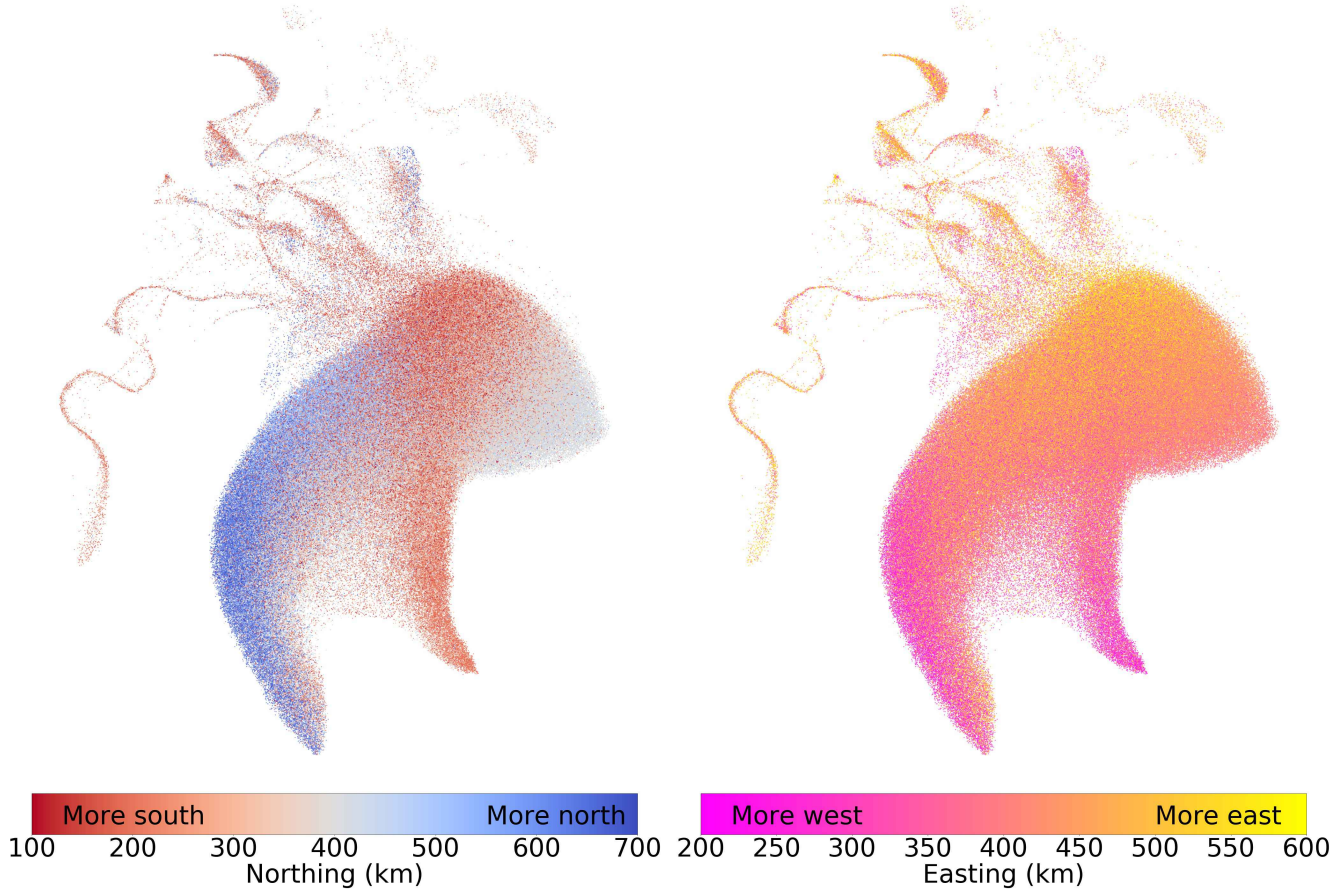


Fig. 5. The UKBB projected onto two dimensions using PCA-UMAP with each individual colored by their geographical coordinates of residence. Coordinates follow the UKBB's OSGB1936 geographic grid system and represent distance from the Isles of Scilly, which lie southwest of Great Britain. The left image colors individuals by their north-south ("northing") coordinates, and the right image colors them by their east-west ("easting") coordinates. Adding more components creates finer clusters. (figures **??** and **??**). Individuals with missing geographic data are not shown. To prevent outlying individuals from washing out the colour scheme, northing values were truncated between 100km and 700km, and easting values were truncated between 200km and 600km. To protect participant privacy, data has been randomized as explained in the materials and methods section.

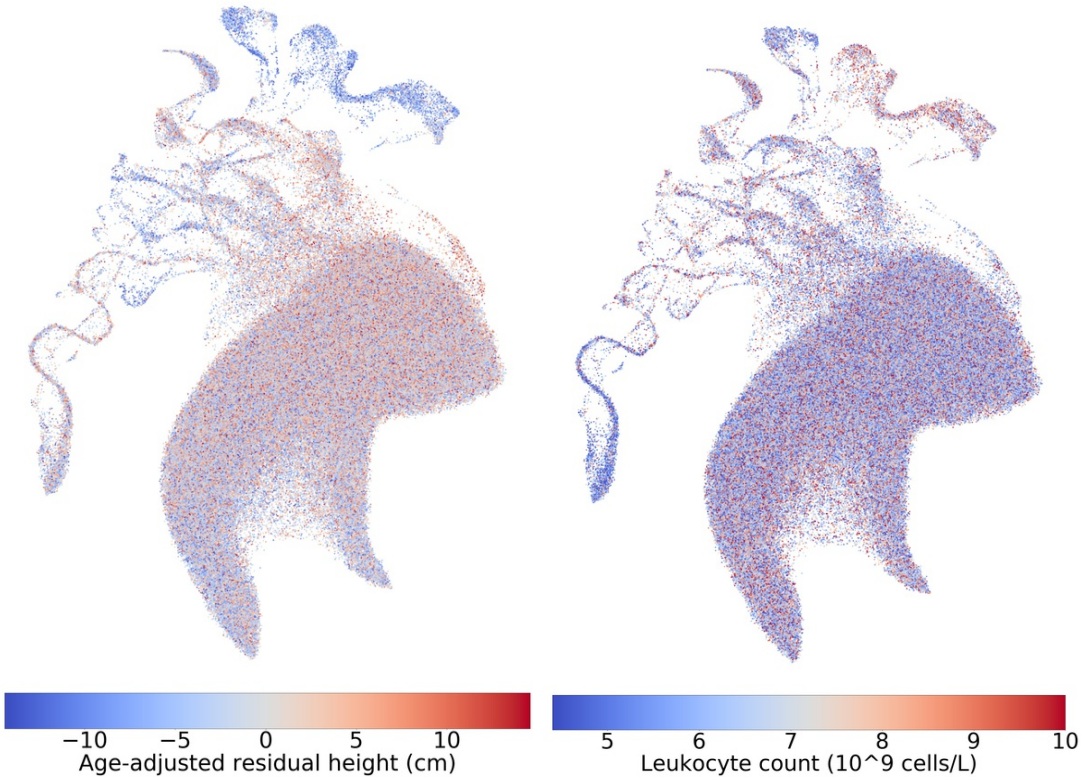


Fig. 6. The UKBB projected onto two dimensions using PCA-UMAP (as in **??**), with females colored by age-adjusted difference from mean population height (left) and leukocyte counts (right). To protect participant privacy, data has been randomized as explained in the materials and methods section.

the difference is not consistent(**? ?**).

Forced expiratory volume in 1 second (FEV1) (Figure **??** also shows strong correlations with certain populations — South Asian, African, and Caribbean — having considerably lower measurements on average (see figures **??** and **??** for boxplots and p-values). Notably, there appears to be a juncture in the admixture continuum, highlighted in figure **??**, where the distribution of FEV1 changes. This roughly corresponds to the transition from Black African/Caribbean individuals to those who identified having mixed backgrounds. Boxplots and statistical testing suggest that relative to White British populations FEV1 values are significantly lower for Black African and Black Caribbean populations, but not for White and Black Caribbean and White and Black African populations. Unidentified populations highlighted in figure **??** suggest that one ethnic group close to the Chinese may have higher than average FEV1 values compared to the relatively low values of the Chinese themselves; while another close to European and British populations has lower values relative to the population mean. These results merit further investigation and underscore the exploratory value of PCA-UMAP — these populations are largely unidentified and there is no straightforward way to separate them within the data otherwise.

Discussion. Understanding population structure is important to identify confounders in medical genomics and studies of anthropology and human evolution. PCA of genomic data reflects genealogical and geographic data, but visualization in large datasets still requires scanning through a large number of pairwise plots. UMAP condenses these components and comprehensively illustrates information — phenotypic, geographic, and ancestral — contained within genotypes on fine-scale levels and within the context of a global population structure. In large datasets where the number

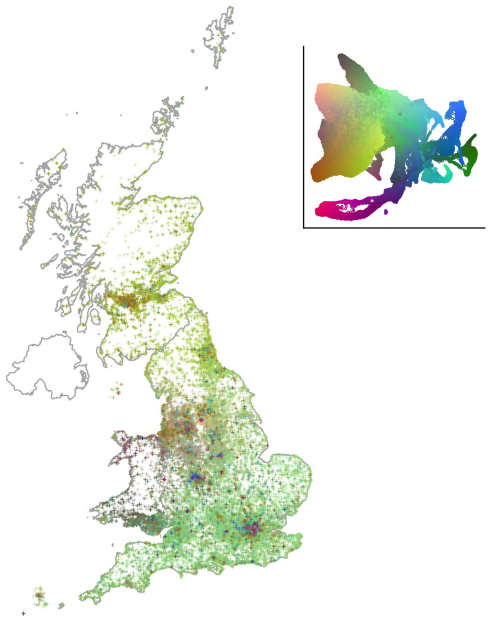


Fig. 7. A map of Great Britain colored by a 3D UMAP projection. Each individual is assigned a 3D RGB vector based on 3D UMAP coordinates (inset): Individuals who are closer to each other in the projection will appear to be closer in color. Patterns in genetic similarity are visible in Scotland, South England, the East and West Midlands, and major urban centres. A flattened 2D view of of the 3D projection used for coloring is presented in the top right. To protect participant privacy, data has been randomized as explained in the materials and methods.

of significant PCs is large, the resulting representation has important advantages over PCA alone and provides a superior visualization to t-SNE.

Examinations of clustering in the three datasets provided many intriguing clusters that would otherwise have been difficult to identify. In particular, several areas from figure **??**, highlighted in figure **??**, show multiple unidentified groups related to each of the East Asian and South Asian super-populations, as well as to either or both of African or admixed populations. Additionally, the Hispanic population of the HRS contains a geographically-restricted cluster that could not have been identified from pairwise examinations of principal components. The 1KGP — frequently used in medical and population studies — contained splits in the Gujarati and Punjabi population samples that were not visible PCA or Admixture analysis alone (although a split among Gujarati is arguably visible in the Admixture analysis with K=12 in (**?**)).

Application to the UKBB underscores the strength of PCA-UMAP in large cohorts. We see clear, fine relationships between genotype and phenotype and geography, and this is presented in a visualization that accounts for natural genetic clustering. Figures **??**, **??**, and **??** demonstrate phenotypic variation within and across clusters, with phenotypes such as height showing continuous variation across admixture edges, as expected from genetically controlled traits, and others, such as leukocyte counts or FEV1, showing sharper boundaries, as expected from environmentally determined traits.

Importantly, using UMAP is straightforward and fast. Most of the plots presented in this article were generated directly from the PCA data using UMAP with default parameters, except that we set the a "minimum distance" parameter to 0.5 which made fine features on UMAP more visible (results with default parameter 0.1 provided qualitatively similar results). Given PCA data and a desktop computer, UMAP can be performed in 15 to 25 minutes on a sample of hundreds of thousands of individuals over tens of dimensions.

There are downsides to using nonlinear approaches to visualize the data. Both UMAP and t-SNE are sensitive to sample size, and spend more visual real estate for populations with larger sample sizes compared to PCA. This is useful to identify significant patterns in a cohort, but it makes comparing visualization across cohorts difficult. Nonlinearity also complicates the interpretation of results. Distances in UMAP or tSNE space should not be used as a proxy for genetic distance. We did not assign meaning to wiggles in UMAP figures, which occurred consistently in the UKBB but may be an artifact of the dimensional reduction strategy rather than a meaningful feature of the data. Hand-waving interpretations of pretty plots has a history of getting population geneticists in trouble (as pointed out, e.g., in (**?**)): visualization is not a replacement for statistical testing.

Conclusion. With these caveats in mind, a priori data visualization plays a central role in quality control, hypothesis generation, and confounder identification for a wide range of genomic applications. Nonlinear approaches, despite their limitations, become increasingly useful as the size of datasets increases: We have shown that UMAP, in particular, reveals a wide range of features that would not be apparent using linear maps. Given its ease of use, breadth of results, and low computational cost, we propose that UMAP should become a default companion to PCA in large genomic cohorts.

Materials and Methods

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

Scripts for all tests and plotting functions can be found on https://github.com/diazale/gt-dimred. A demo version using freely available 1KGP data is available at https://github.com/diazale/1KGP_dimred. PCA and standard t-SNE were done with Scikit-learn(**?**). UMAP was performed using a Python implementation(**?**). Statistical testing was done in SciPy(**?**) and StatsModels(**?**).

Both UMAP and t-SNE feature a number of adjustable parameters. Among the parameters that we varied, the number of PCs used in pre-processing of the data has the largest effect for both methods (see figures **??** and **??**).

We tested different choices for perplexity in t-SNE. The default value of 30 provided comparable performance to other parameter choices. Similarly, we tested different parameter choices for UMAP, with the clearest results generated by specifying 15 nearest neighbours (the default value) and a "minimum distance" between points in low dimensions of 0.5. UMAP developers described "sensible" values for nearest neighbours as between 5 and 50 and minimum distance between 0.5 and 0.001.

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

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

For each point in figure **??** we identified the nearest 50 neighbouring individuals and copied the colour value from a randomly selected neighbour.

Declarations.

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

Consent for publication. Not applicable.

Availability of data and material. All data is publicly available to researchers.

Competing interests. We have no competing interests to declare.

Funding. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program and CIHR grant MOP-136855.

Authors' contributions. A.D.P. and S.G. designed the research. A.D.P. carried out dimension reduction and analysis. S.G. provided analysis. A.D.P and S.G. wrote the paper. L.A.T. provided the UK map visualization and analysis.

Acknowledgements. We thank all participants in the HRS, UKBB, and 1KGP for providing their genetic data as well as the teams who generated and assembled the dataset. We also thank Chief Ben-Eghan, Jose Sergio Hleap, Mark Lathrop, Dominic Nelson, Markus Munter, Stephen Sawcer, and Audrey Grant for useful discussions about science, programming, and data access, and Selin Jessa for introducing us to UMAP. This research has been conducted using the UK Biobank Resource under Application Number 6728.

1. Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS genetics* 8(1):e1002453.
2. McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS genetics* 5(10):e1000686.
3. Brisbin A, et al. (2012) PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human biology* 84(4):343.
4. Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456:98–101.
5. Nelson MR, et al. (2008) The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics* 83(3):347–358.
6. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLOS Genetics* 2(12):1–20.
7. 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526(7571):68.
8. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov):2579–2605.
9. Platzer A (2013) Visualization of SNPs with t-SNE. *PloS one* 8(2):e56883.
10. Li W, Cerise JE, Yang Y, Han H (2017) Application of t-SNE to human genetic data. *Journal of Bioinformatics and Computational Biology* 15(04):1750017. PMID: 28718343.
11. McInnes L, Healy J (2018) UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
12. Juster FT, Suzman R (1995) An overview of the Health and Retirement Study. *Journal of Human Resources* pp. S7–S56.
13. Sudlow C, et al. (2015) 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(3):e1001779.
14. Baharian S, et al. (2016) The great migration and African-American genomic diversity. *PLoS genetics* 12(5):e1006059.
15. Maples BK, Gravel S, Kenny EE, Bustamante CD (2013) RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* 93(2):278–288.
16. Jordan I, Rishishwar L, Conley AB (2018) Cryptic Native American ancestry recapitulates population-specific migration and settlement of the continental United States. *bioRxiv.*
17. Leslie S, et al. (2015) The fine-scale genetic structure of the British population. *Nature* 519(7543):309.
18. Robinson MR, et al. (2015) Population genetic differentiation of height and body mass index across Europe. *Nature genetics* 47(11):1357.
19. Komlos A (1994) *Stature, living standards, and economic development: Essays in anthropometric history*. (University of Chicago Press).
20. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature genetics* 40(5):646.
21. Purcell S, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 81(3):559–575.
22. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.
23. Jones E, Oliphant T, Peterson P, , et al. (2001–) SciPy: Open source scientific tools for Python. [Online; accessed 2018-02-02].
24. Seabold S, Perktold J (2010) Statsmodels: Econometric and statistical modeling with python in *9th Python in Science Conference*.

Supporting Information (SI).

SI Figures.

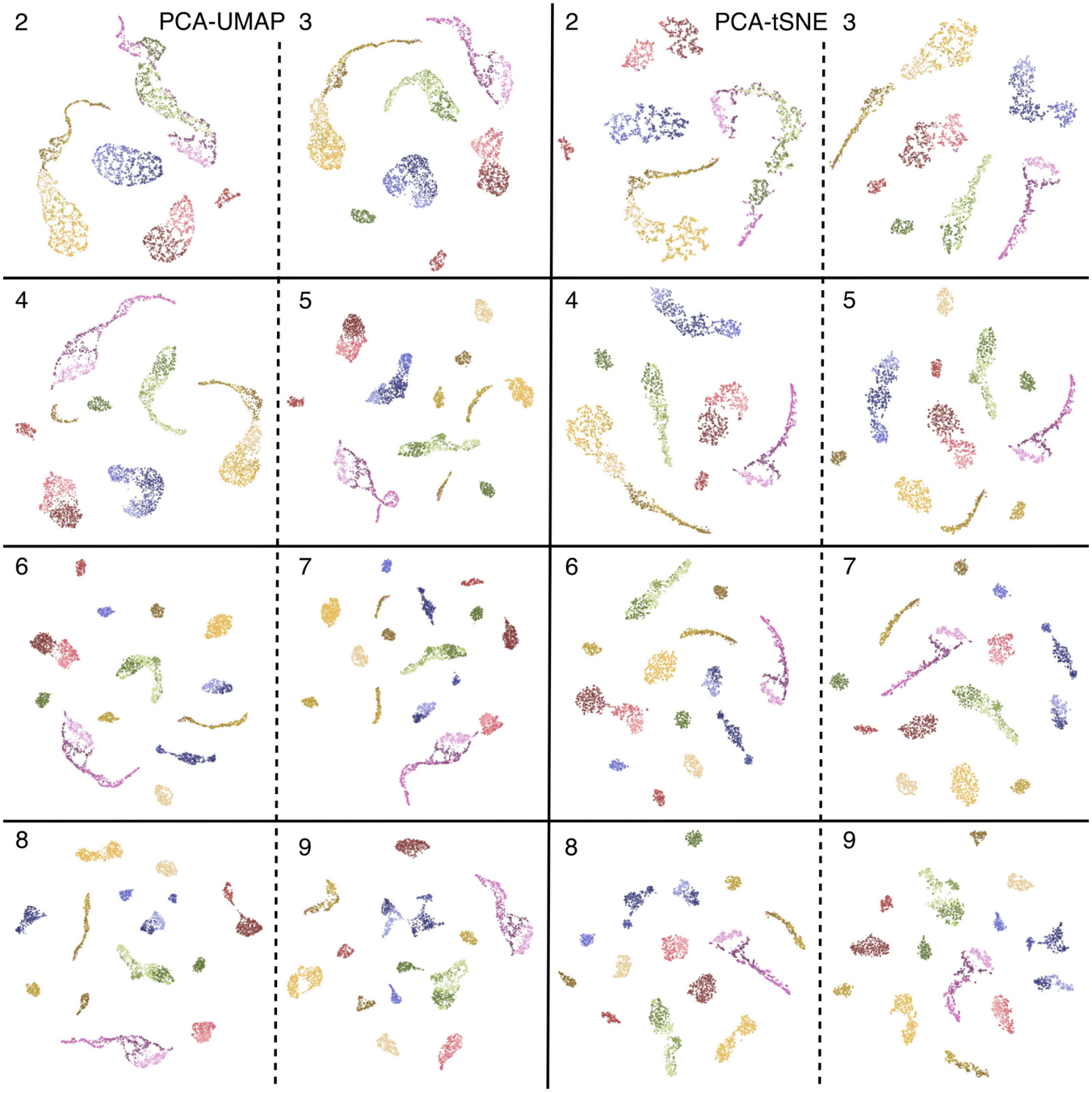


Fig. S1. UMAP (left two columns) and t-SNE (right two columns) applied to the top principal components of the 1KGP labelled by the number of components used. Adding more components results in progressively finer population clusters using both methods.

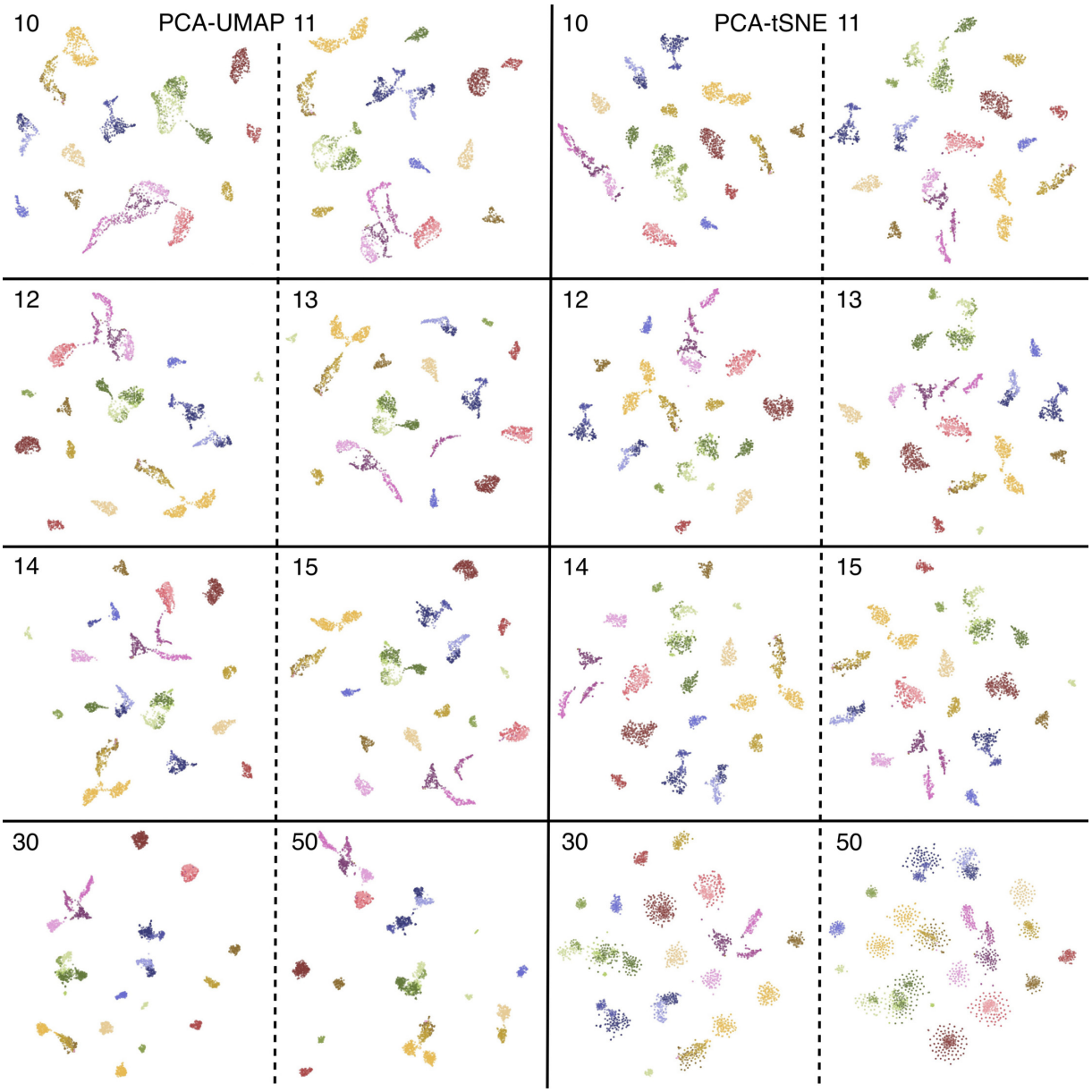


Fig. S2. UMAP (left two columns) and t-SNE (right two columns) applied to the top principal components of the 1KGP labelled by the number of components used. Results are similar until approximately 11 components, where t-SNE breaks apart clusters of South Asian (in green) and Central and South American populations (in pink) while UMAP preserves them. At approximately 30 components populations begin to drift together with UMAP and disperse with t-SNE.

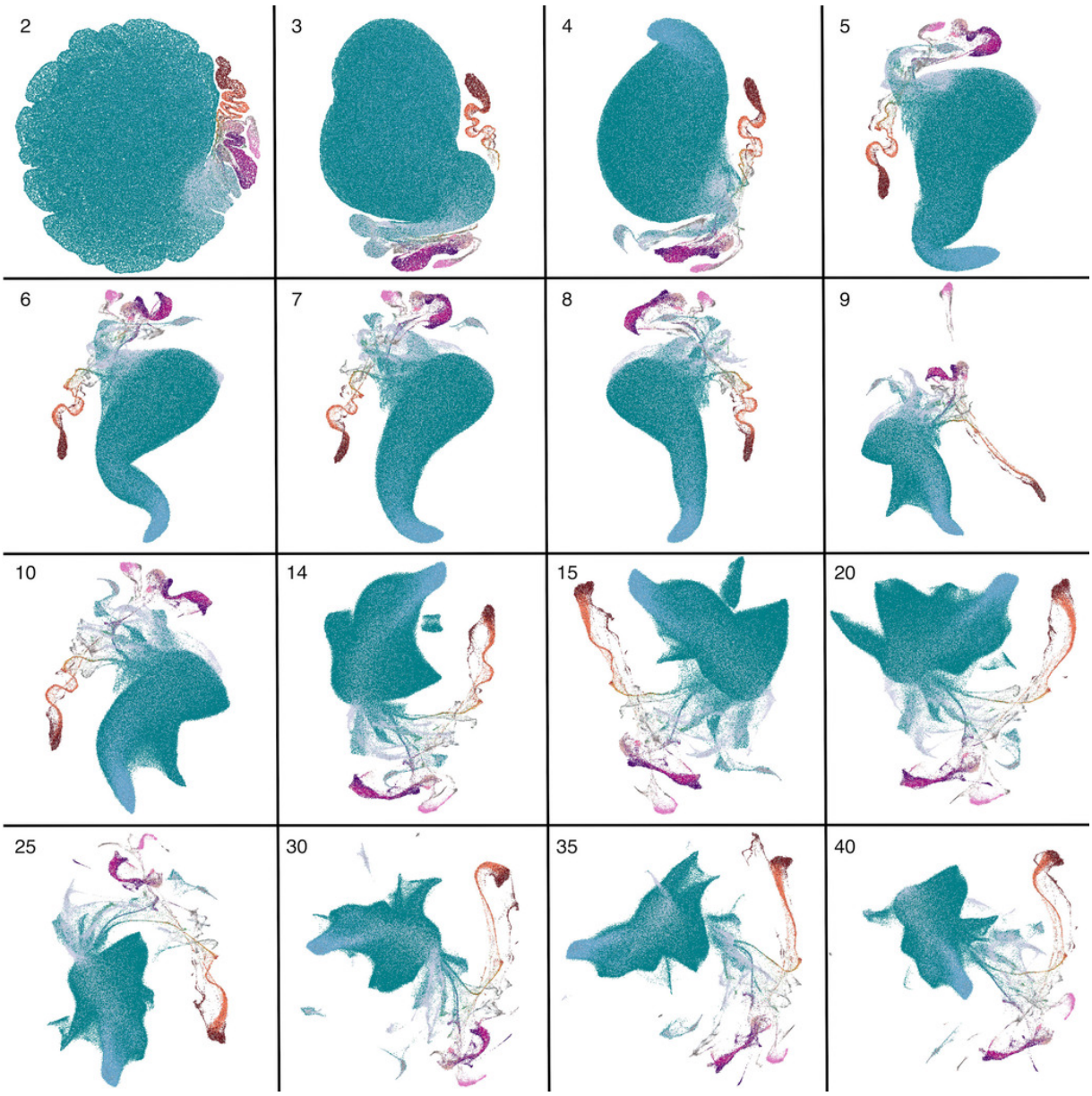


Fig. S3. PCA-UMAP on UKBB data, colored by self-identified ethnic background. Images are labelled by the number of components included.

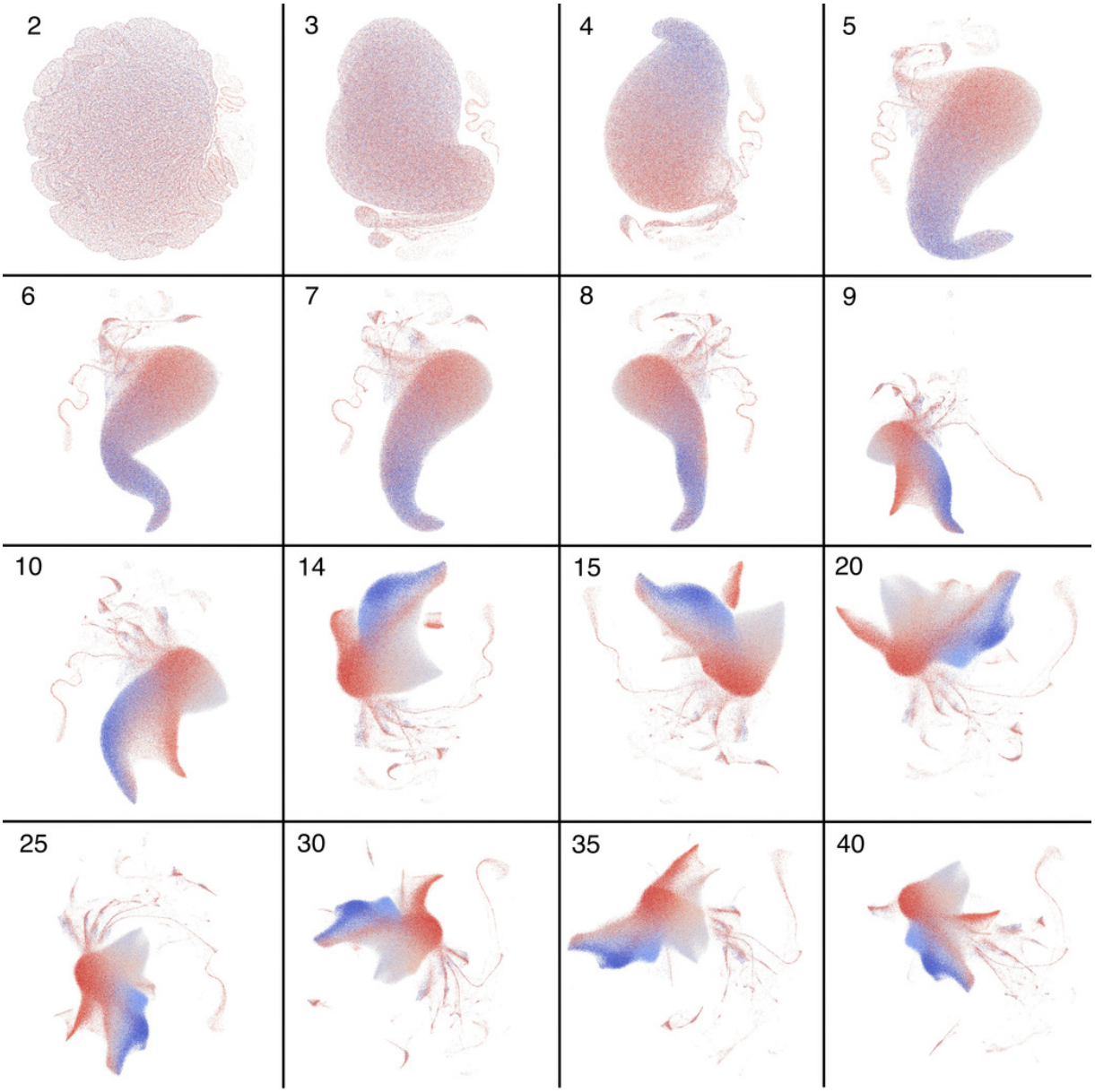


Fig. S4. PCA-UMAP on UKBB data, colored by northing values, with more blue representing more northern coordinates and more red representing more southern coordinates. Images are labelled by the number of components included. Data has been randomized as explained in the materials and methods section.

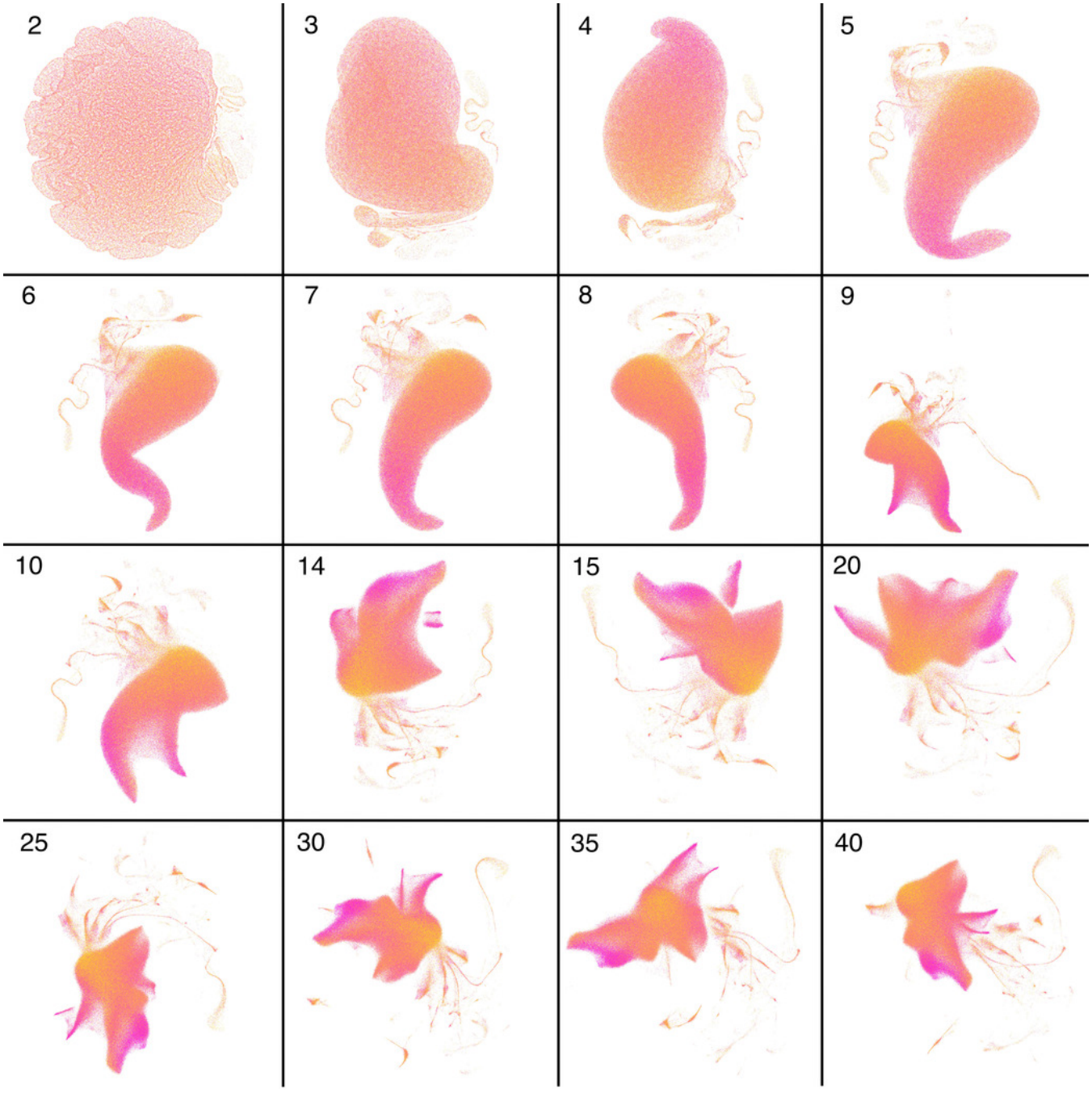


Fig. S5. PCA-UMAP on UKBB data, colored by easting values, with more yellow representing more eastern coordinates and more pink representing more western coordinates. Images are labelled by the number of components included. Data has been randomized as explained in the materials and methods section.

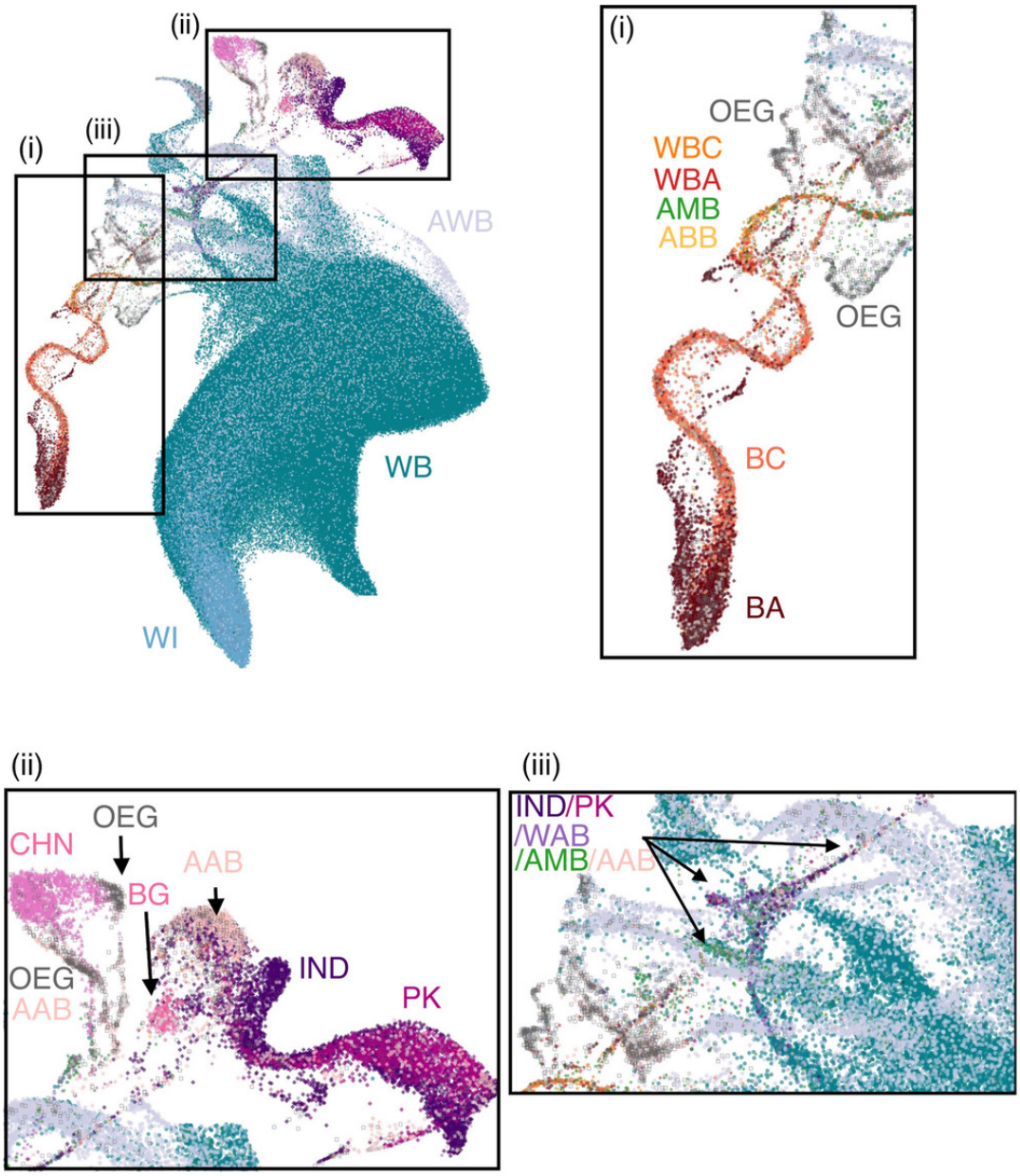


Fig. S6. Zoomed in areas of figure **??**. Sections (i) and (ii) respectively focus on the African and Asian superpopulations, and section (iii) focuses on an area with individuals from many ethnic backgrounds. Noticeable clusters of unidentified ethnic backgrounds appear and are labelled "OEG" "(Other Ethnic Group)".

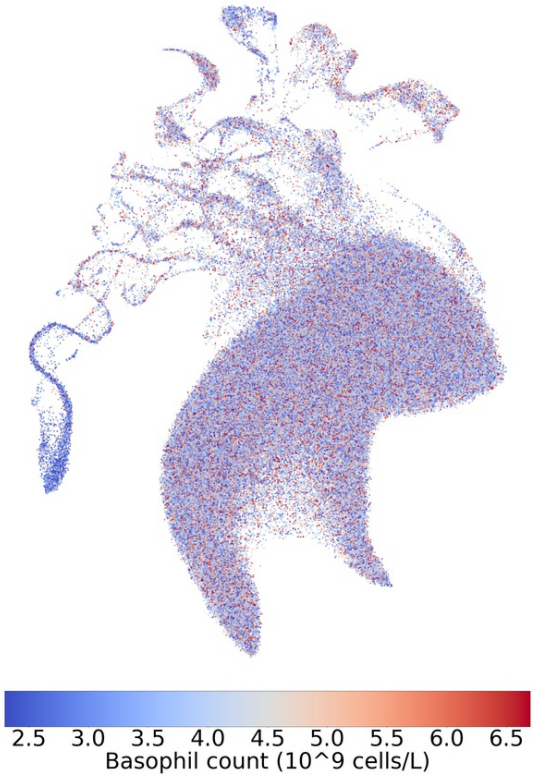


Fig. S7. PCA-UMAP on the top 10 principal components of the UKBB colored by basophil count (female). Data has been randomized as explained in the materials and methods section.

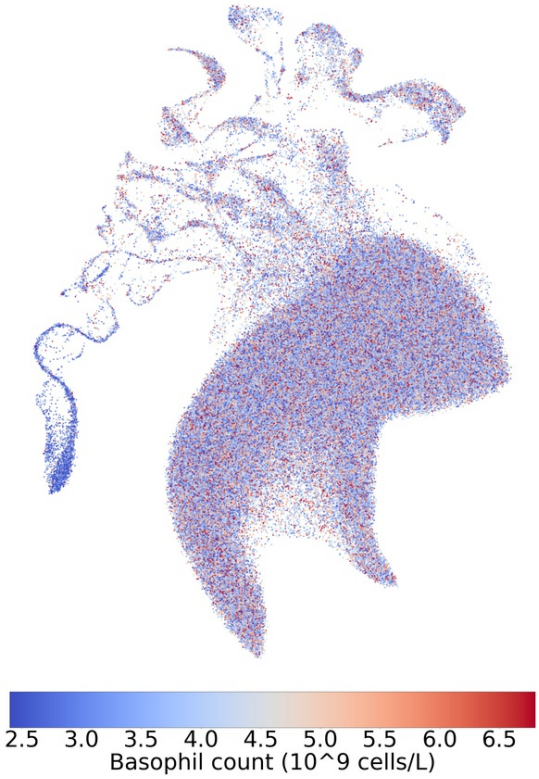


Fig. S8. PCA-UMAP on the top 10 principal components of the UKBB colored by basophil count (male). Data has been randomized as explained in the materials and methods section.

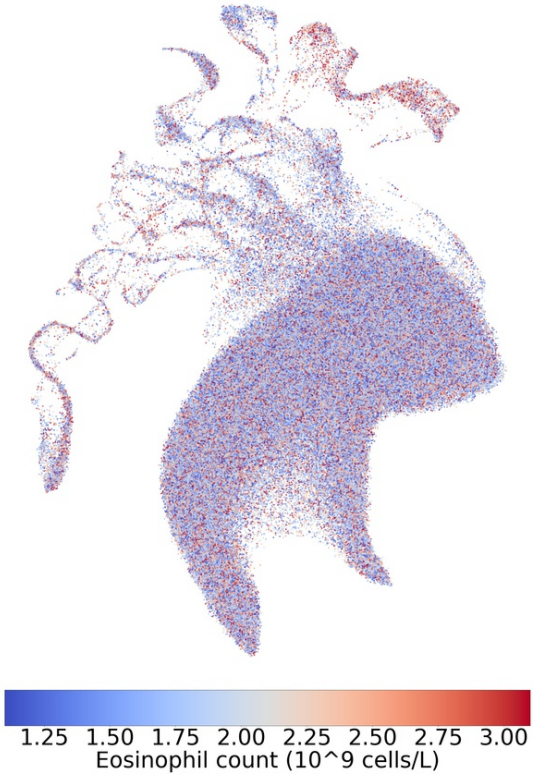


Fig. S9. PCA-UMAP on the top 10 principal components of the UKBB colored by eosinophil count (female). Data has been randomized as explained in the materials and methods section.

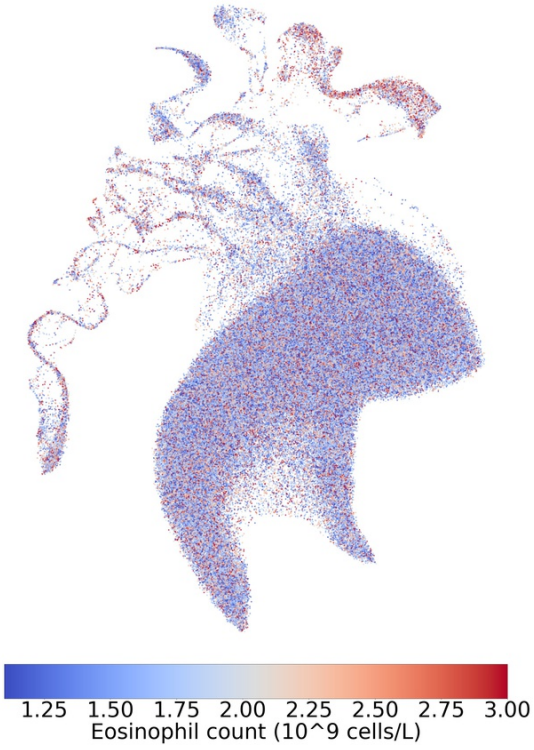


Fig. S10. PCA-UMAP on the top 10 principal components of the UKBB colored by eosinophil count (male). Data has been randomized as explained in the materials and methods section.

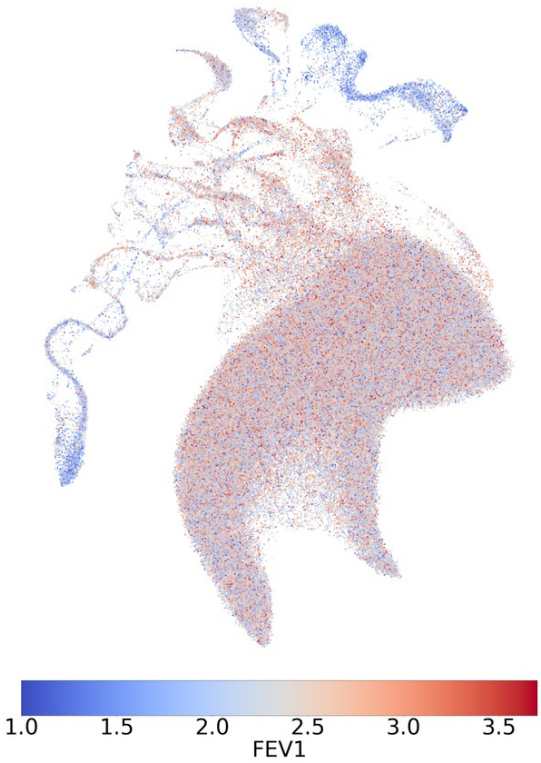


Fig. S11. PCA-UMAP on the top 10 principal components of the UKBB colored by FEV1 (female). Data has been randomized as explained in the materials and methods section.

Fig. S12. PCA-UMAP on the top 10 principal components of the UKBB colored by FEV1 (male). Data has been randomized as explained in the materials and methods section.

Fig. S13. PCA-UMAP on the top 10 principal components of the UKBB colored by height (female). Data has been randomized as explained in the materials and methods section.

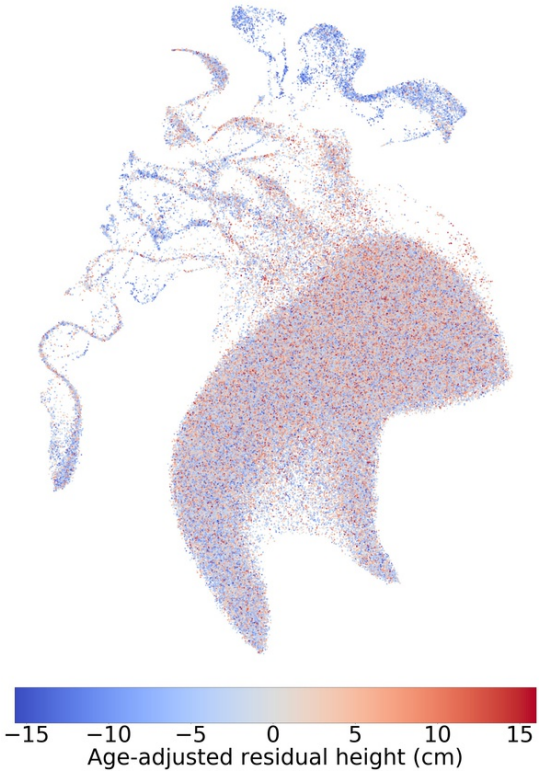


Fig. S14. PCA-UMAP on the top 10 principal components of the UKBB colored by height (male). Data has been randomized as explained in the materials and methods section.

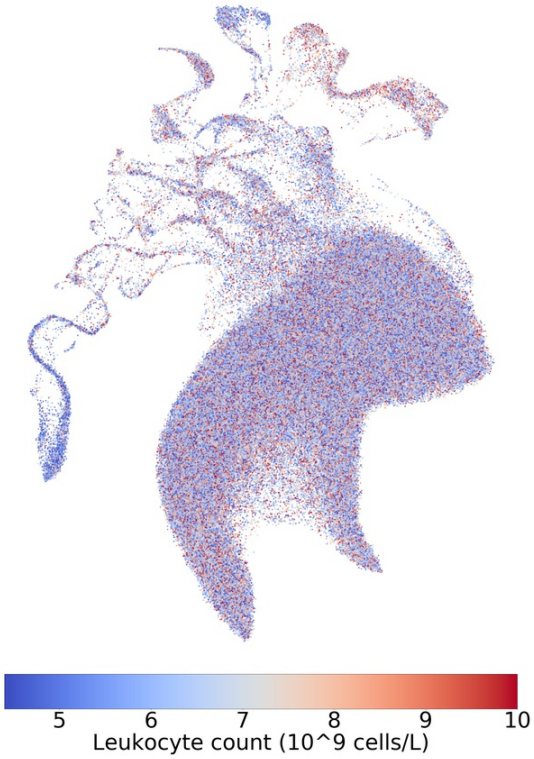


Fig. S15. PCA-UMAP on the top 10 principal components of the UKBB colored by leukocyte count (female). Data has been randomized as explained in the materials and methods section.

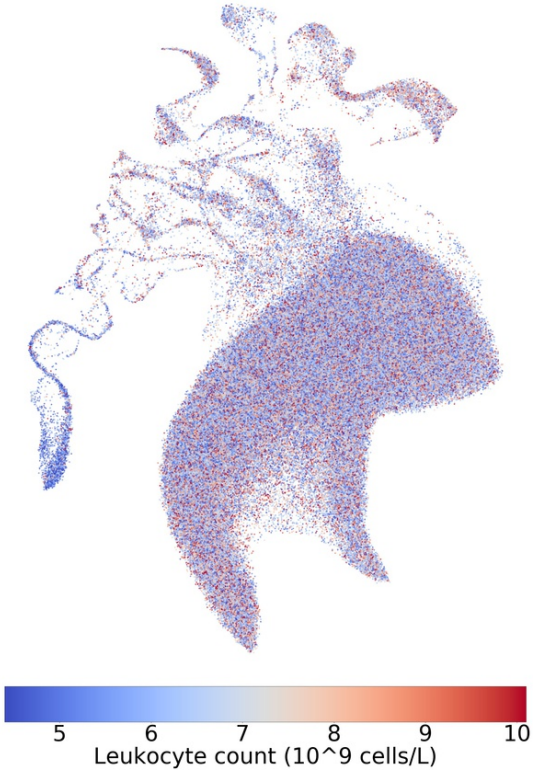


Fig. S16. PCA-UMAP on the top 10 principal components of the UKBB colored by leukocyte count (male). Data has been randomized as explained in the materials and methods section.

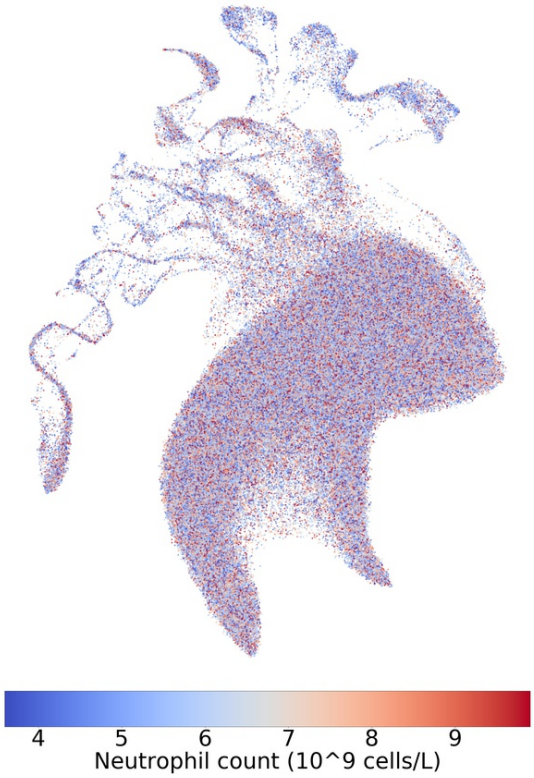


Fig. S17. PCA-UMAP on the top 10 principal components of the UKBB colored by neutrophil count (female). Data has been randomized as explained in the materials and methods section.

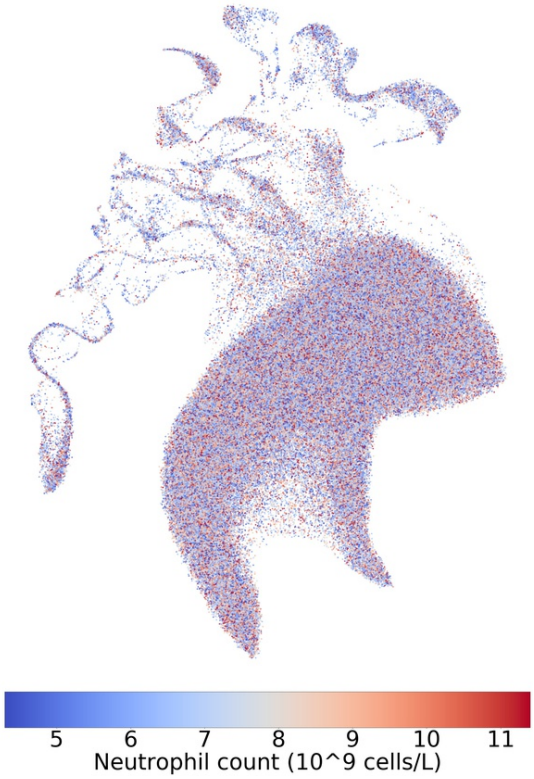


Fig. S18. PCA-UMAP on the top 10 principal components of the UKBB colored by neutrophil count (male). Data has been randomized as explained in the materials and methods section.

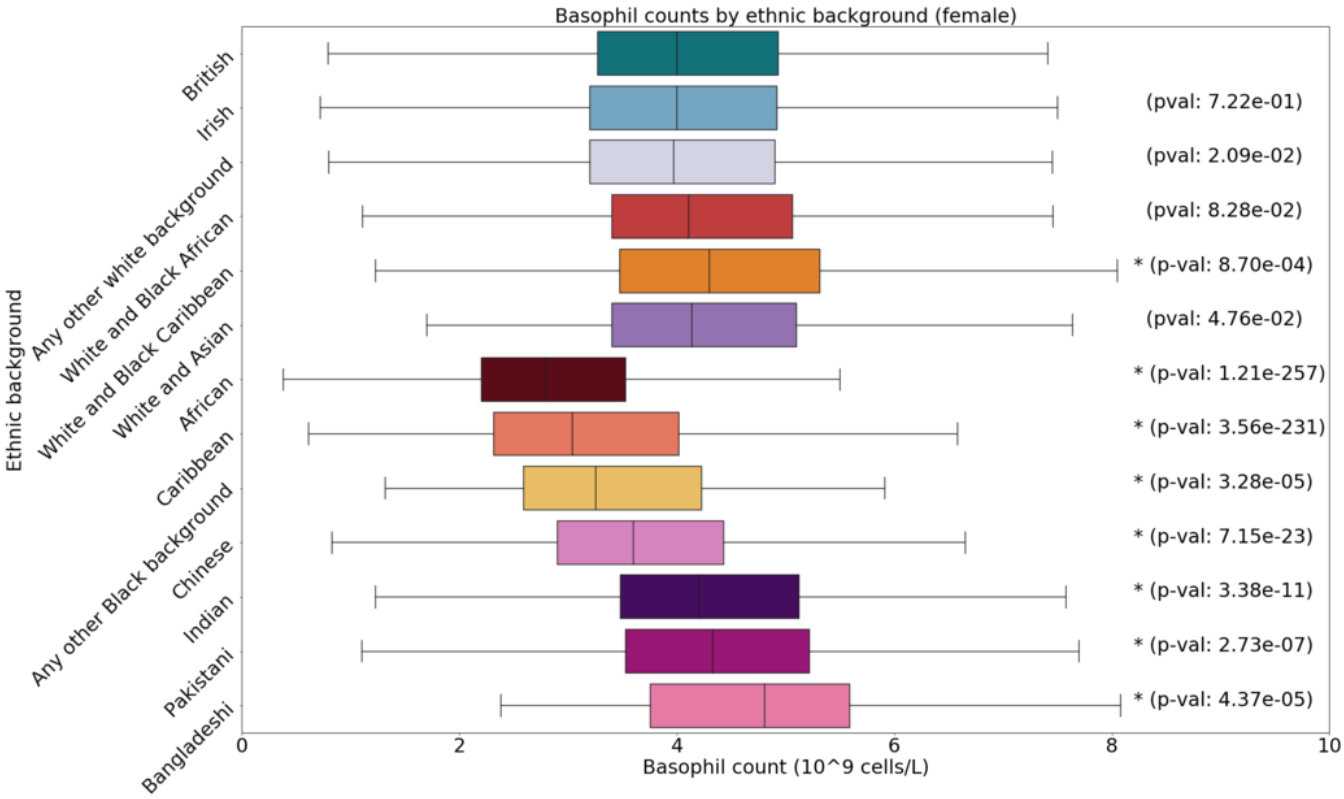


Fig. S19. Boxplot of basophil counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

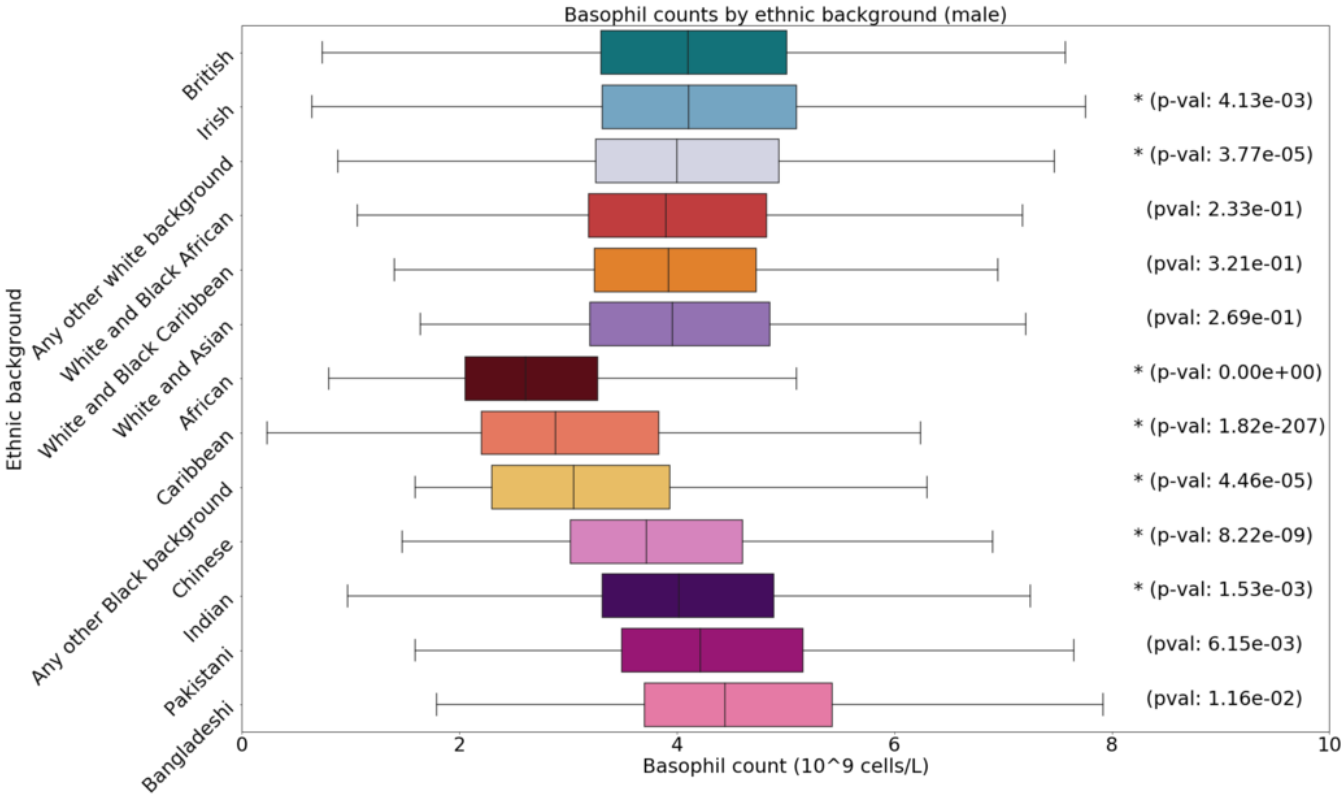


Fig. S20. Boxplot of basophil counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

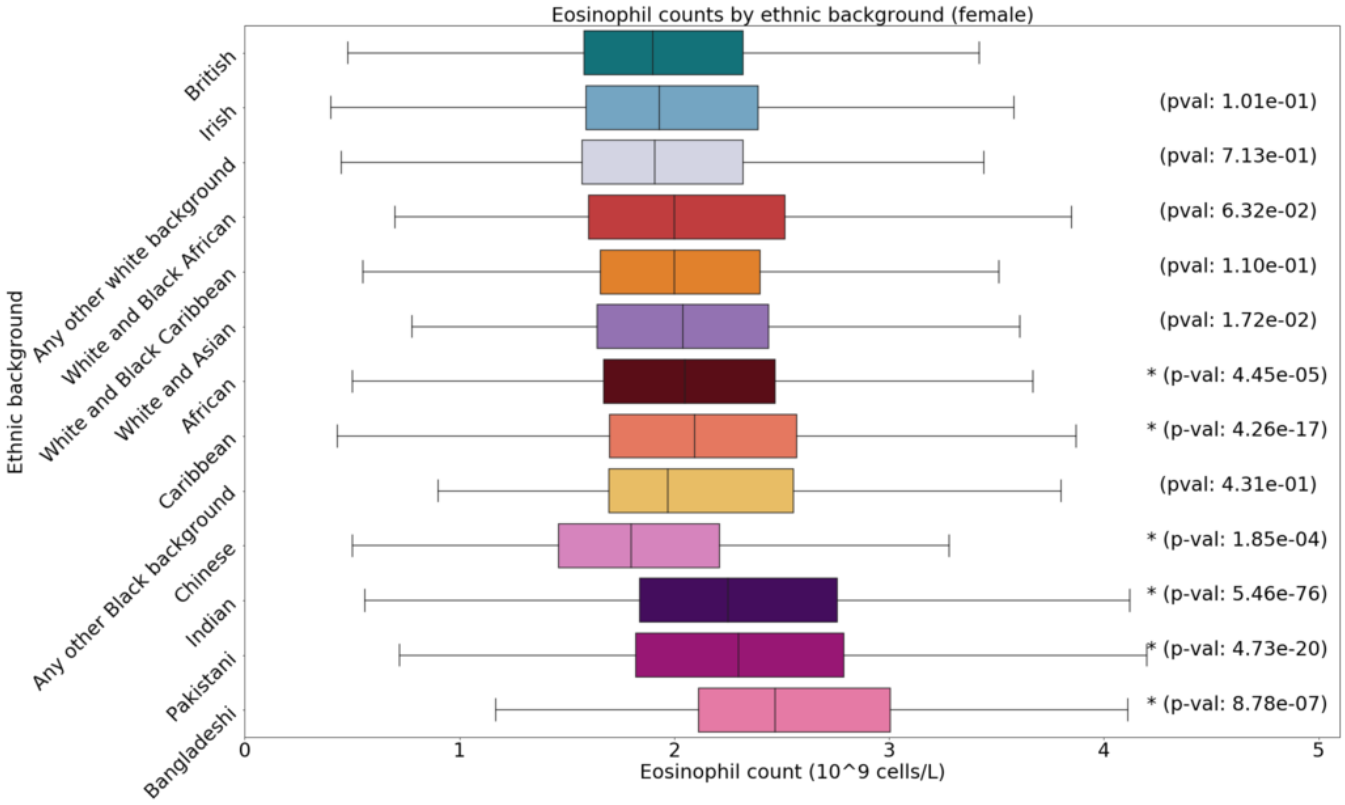


Fig. S21. Boxplot of eosinophil counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

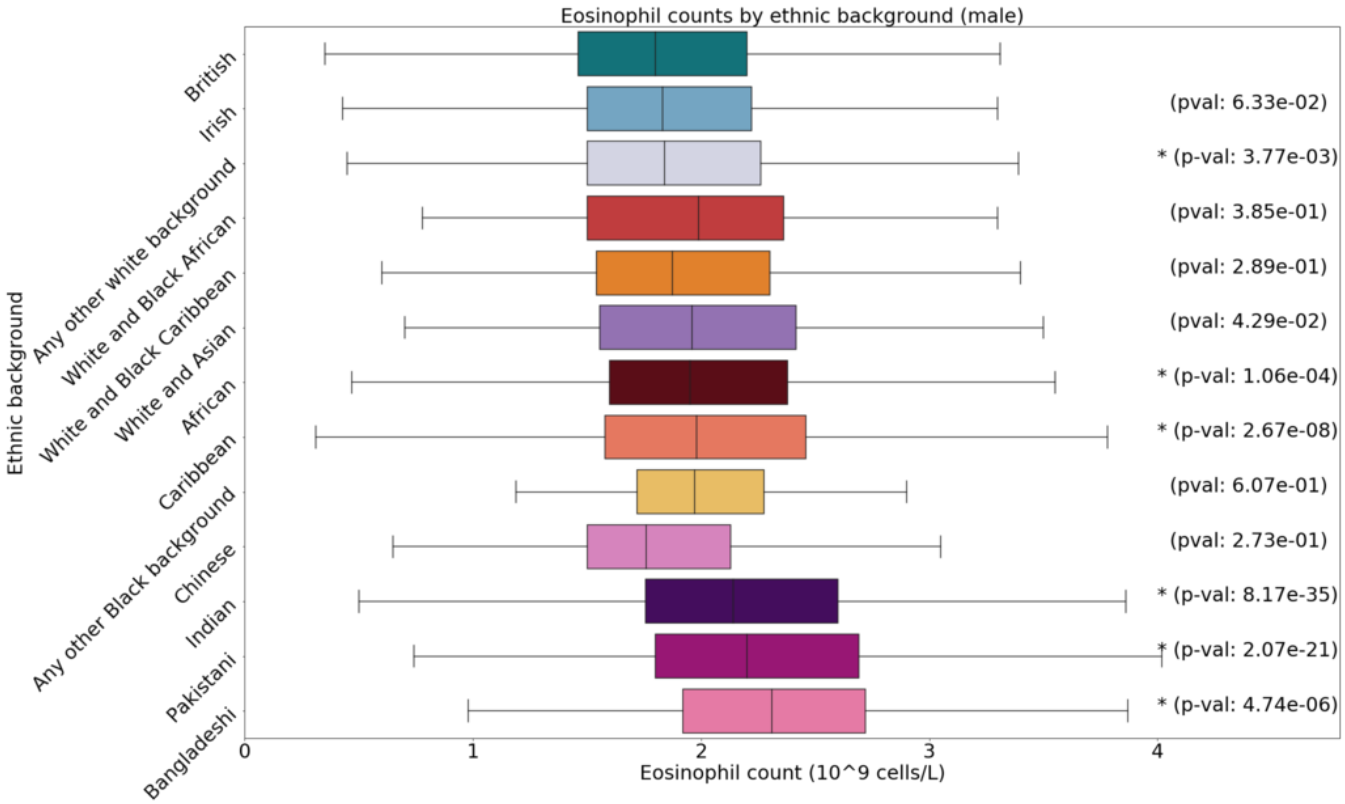


Fig. S22. Boxplot of eosinophil counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

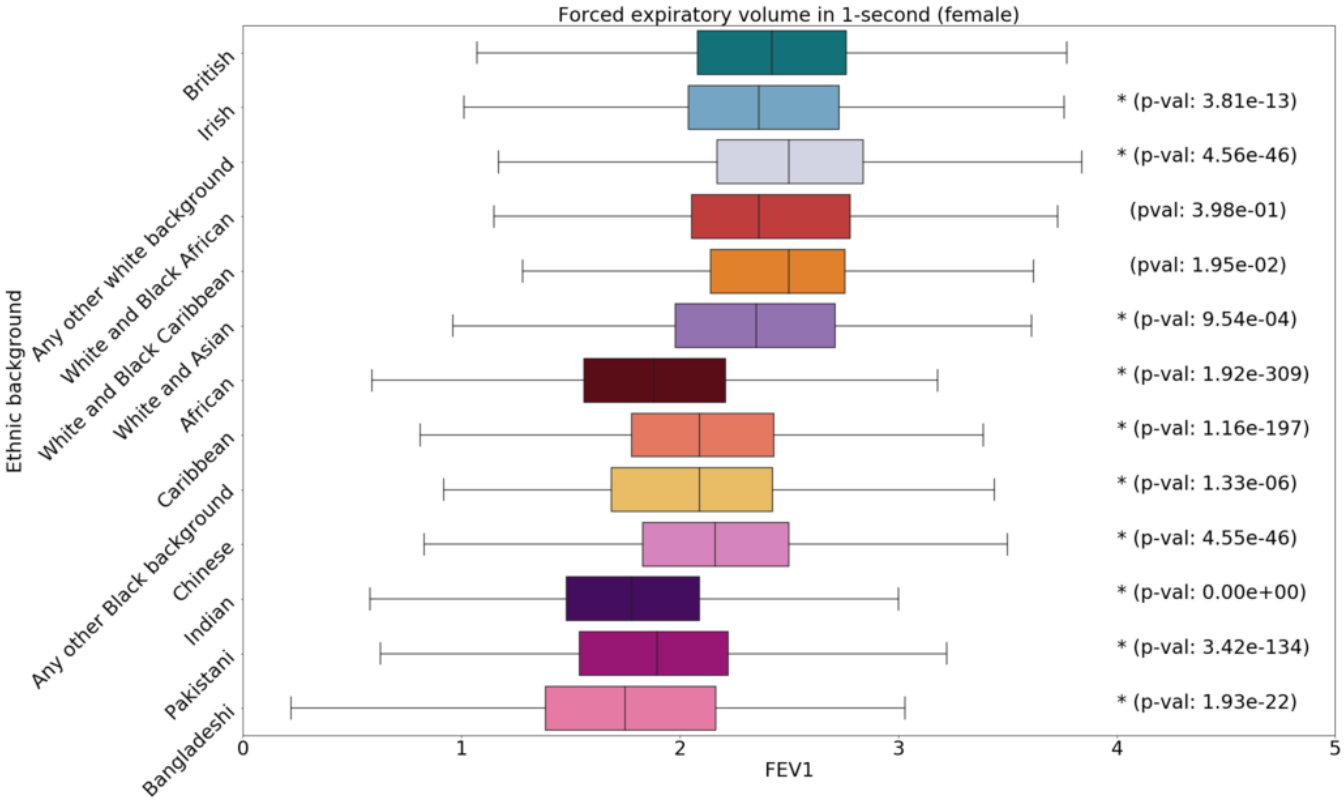


Fig. S23. Boxplot of FEV1 by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

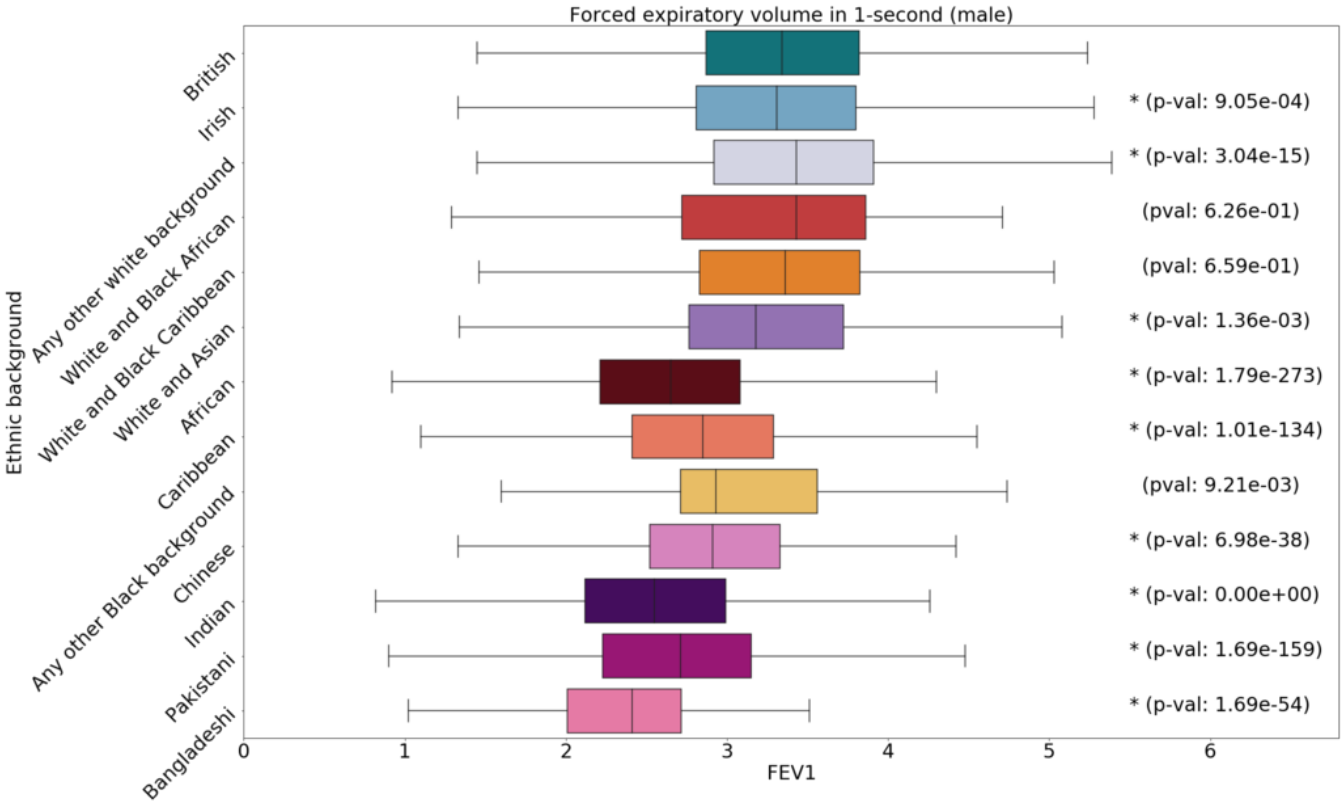


Fig. S24. Boxplot of FEV1 by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

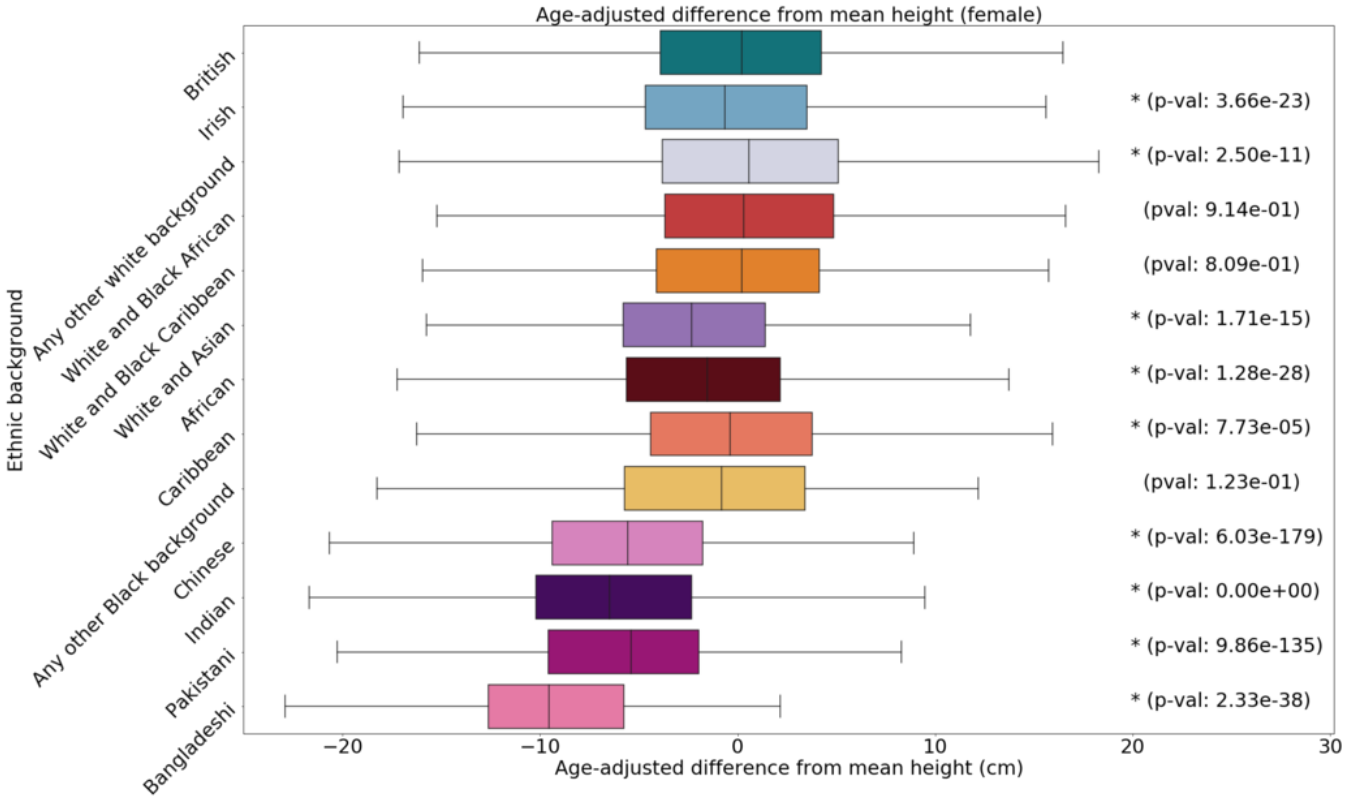


Fig. S25. Boxplot of height by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

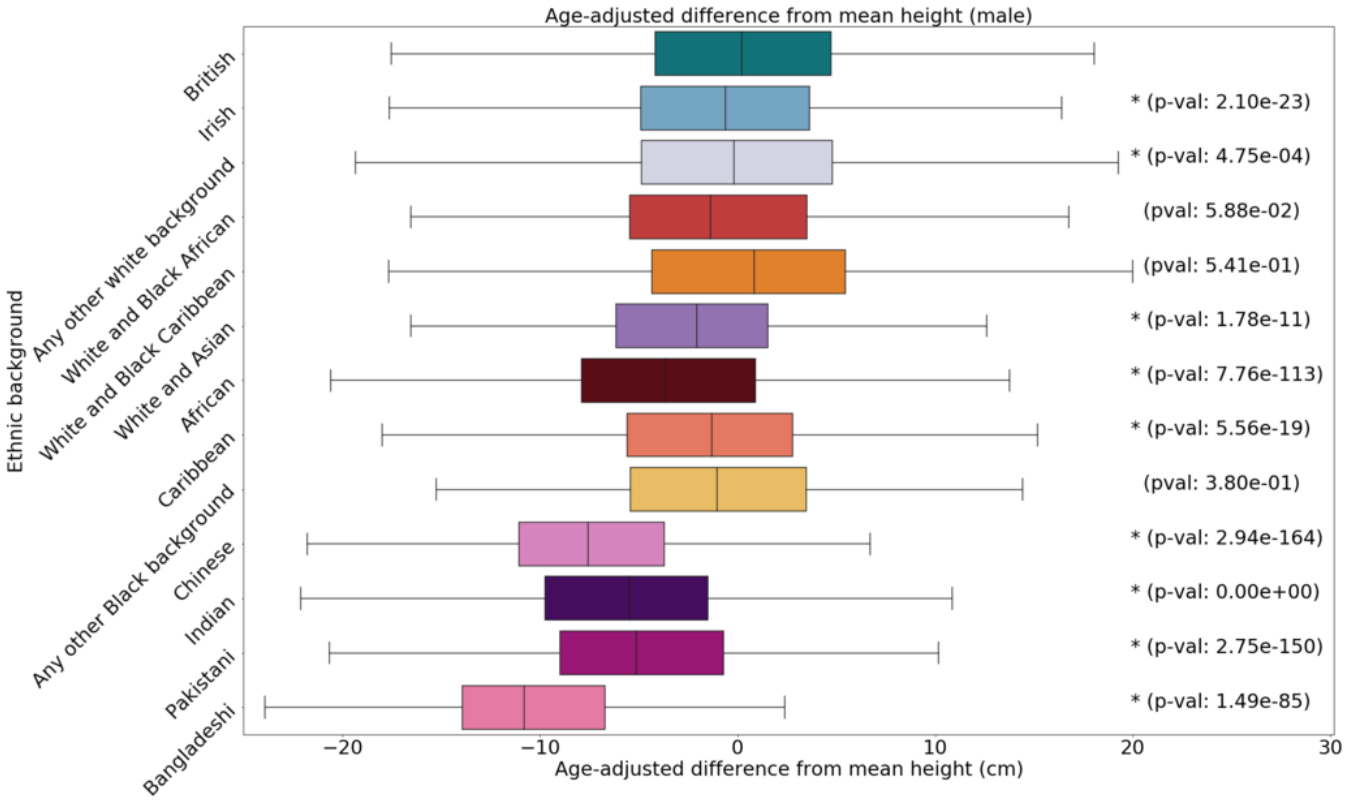


Fig. S26. Boxplot of height by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

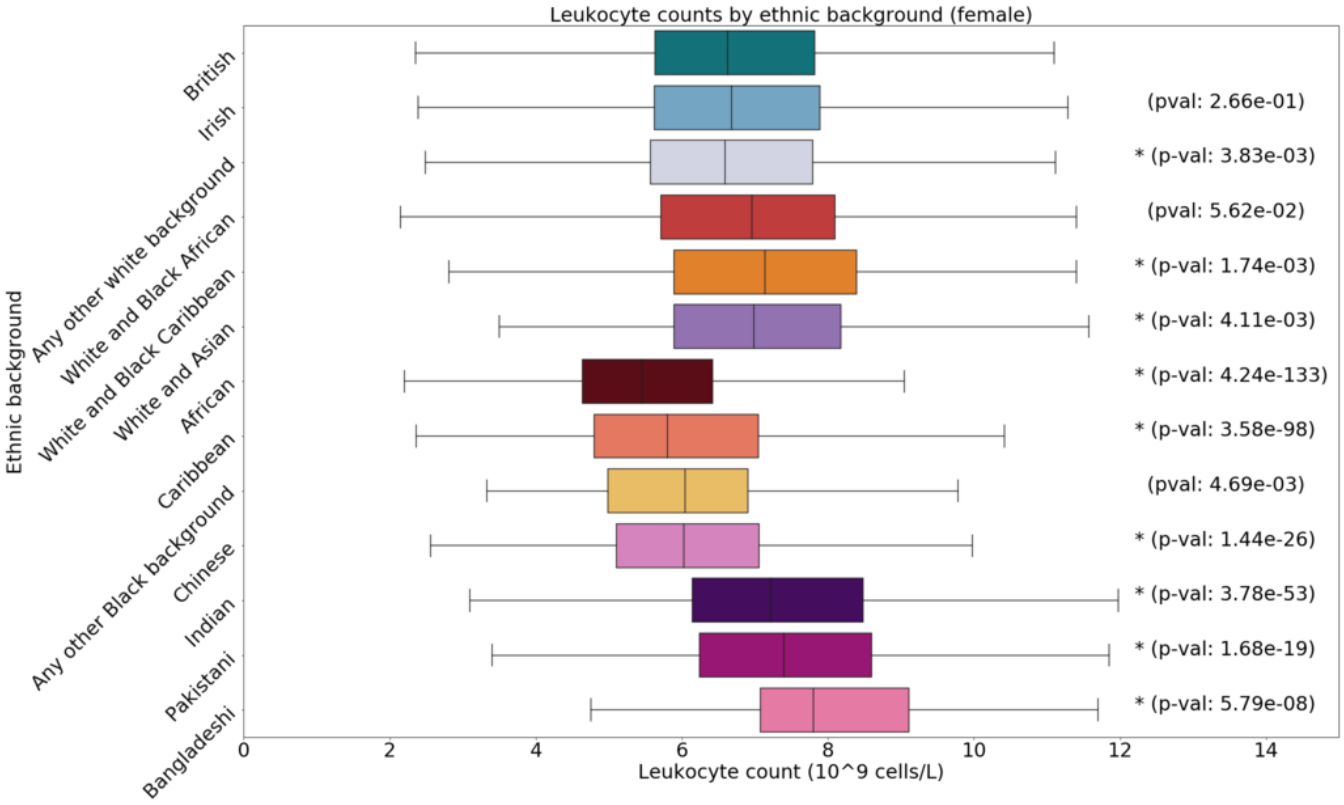


Fig. S27. Boxplot of leukocyte counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

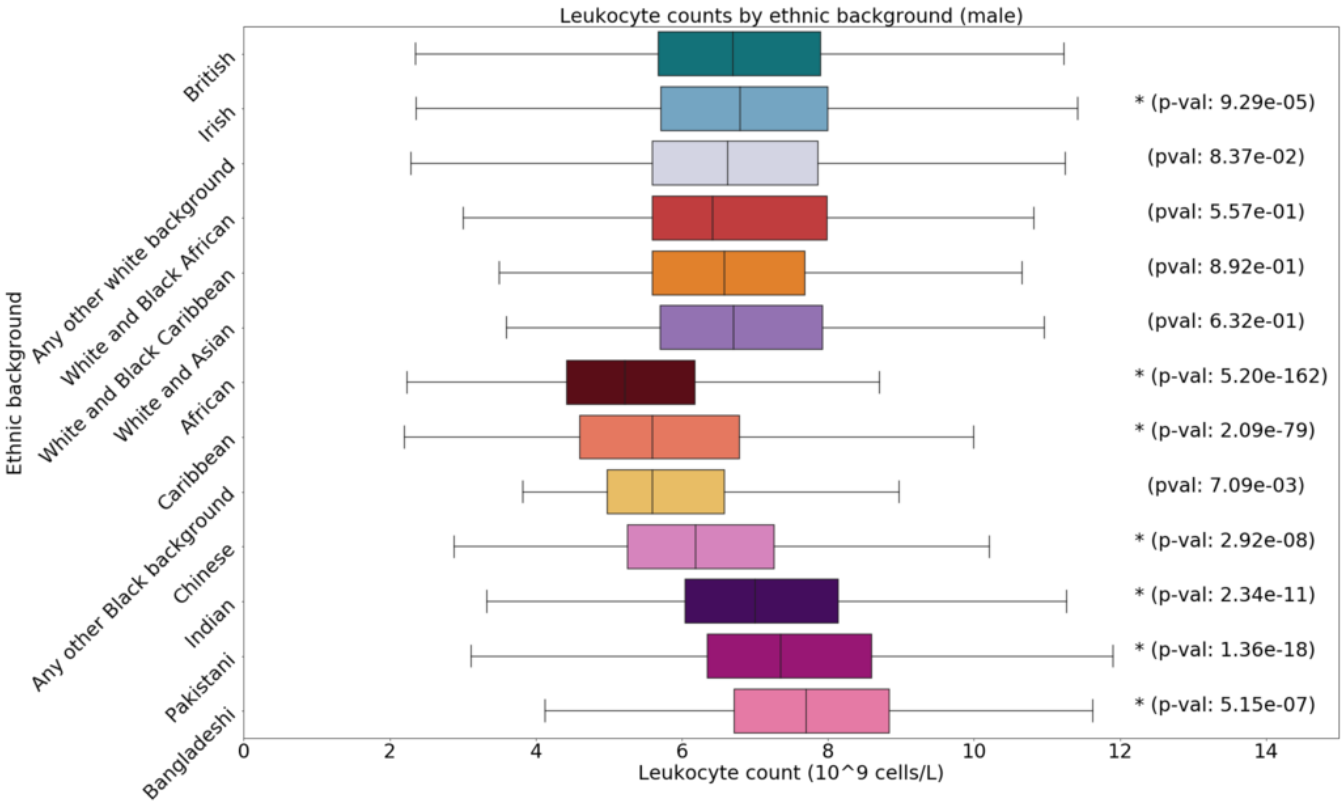


Fig. S28. Boxplot of leukocyte counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

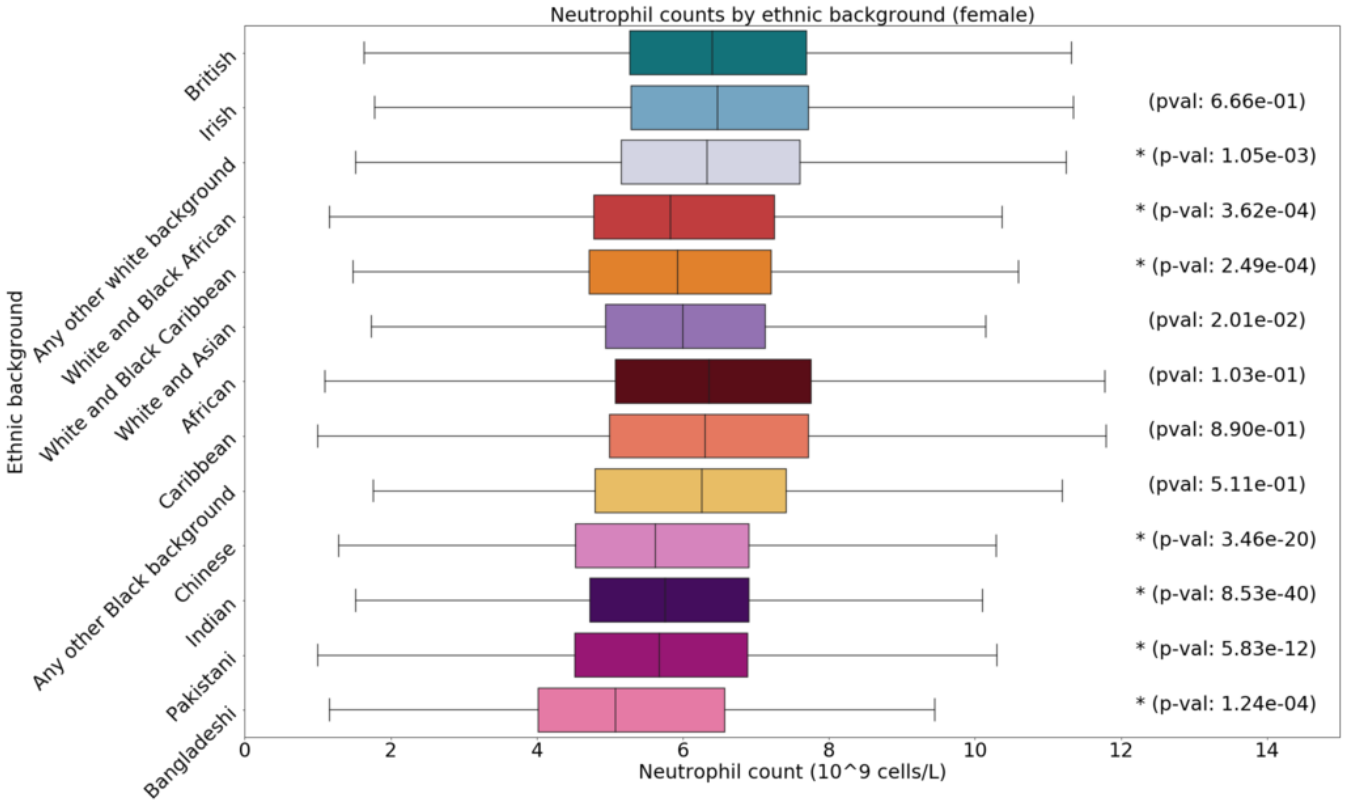


Fig. S29. Boxplot of neutrophil counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

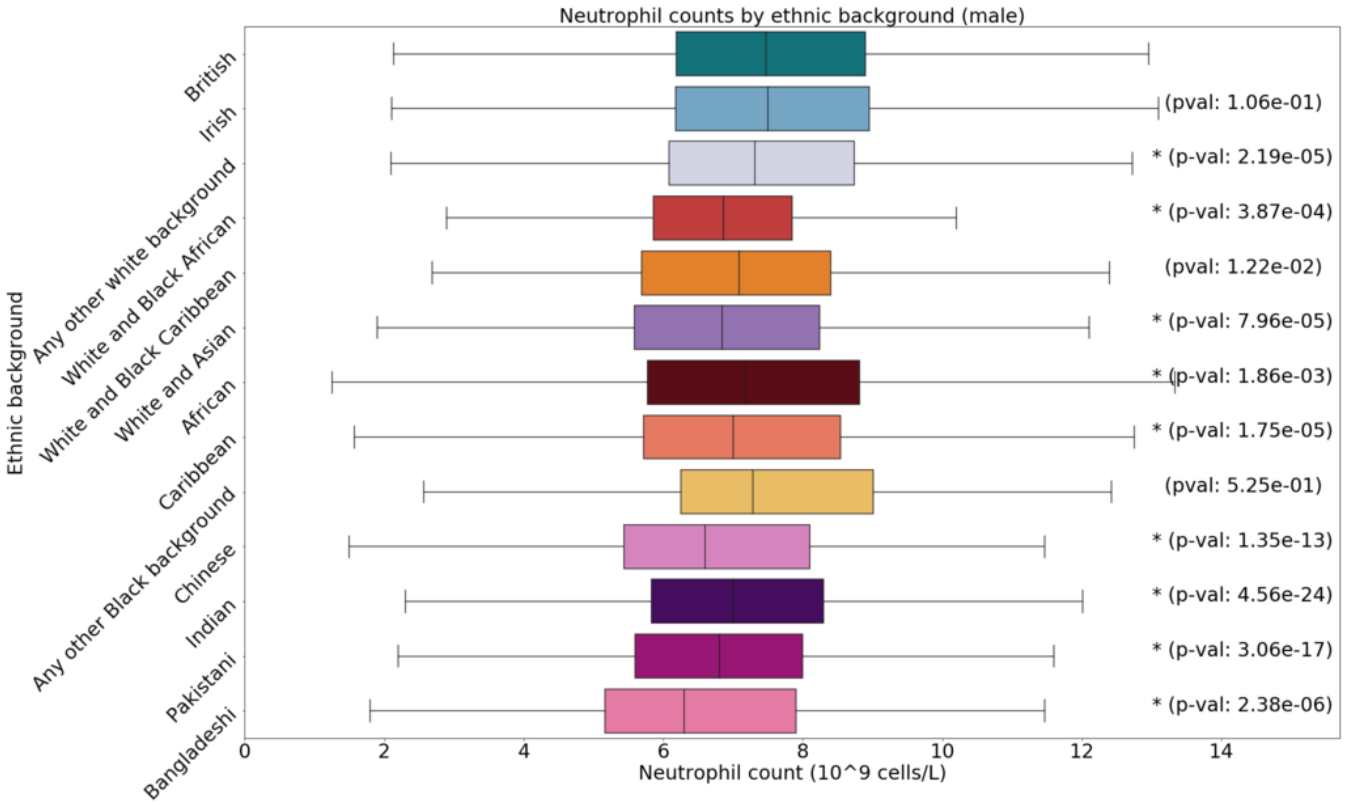


Fig. S30. Boxplot of neutrophil counts by sex and ethnic group, annotated with p-values. Asterisks indicate significant difference from the White British group with a Bonferroni correction for 12 groups.

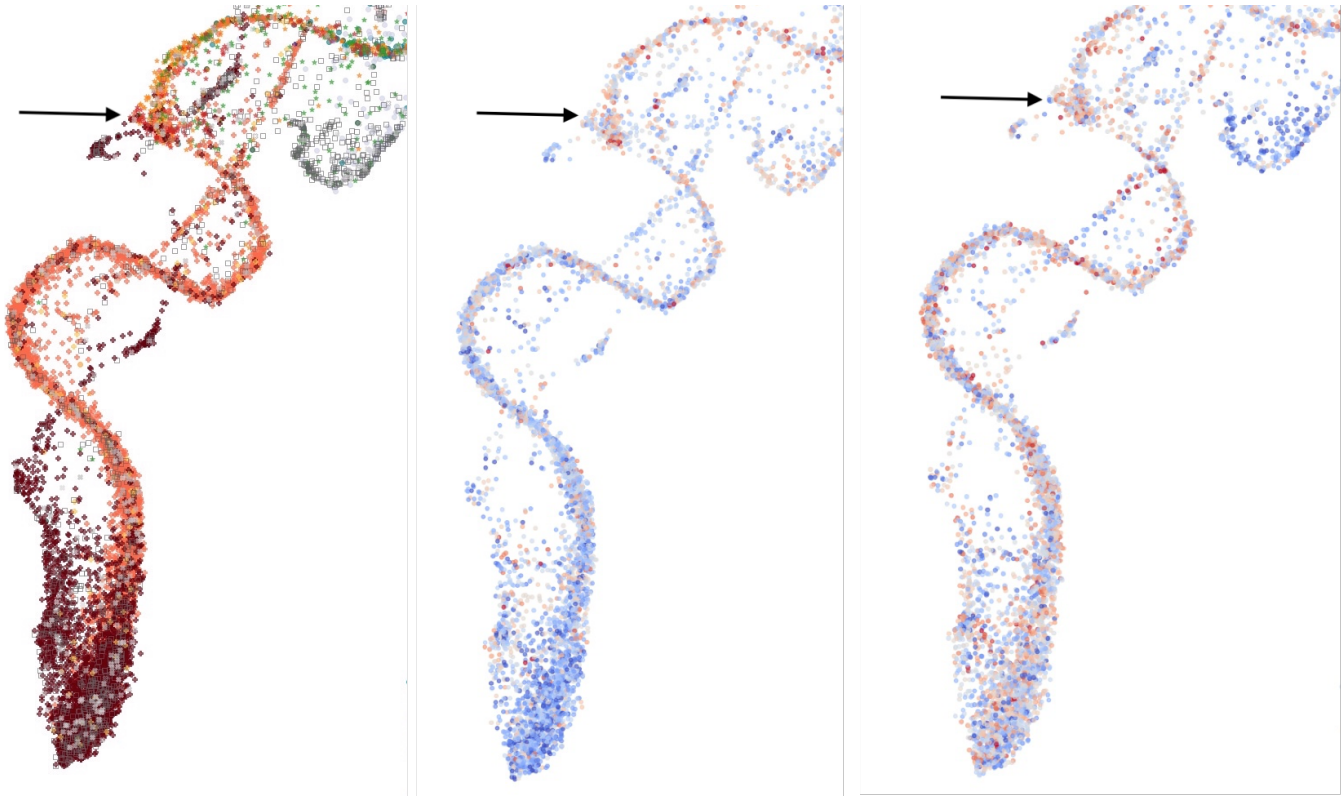


Fig. S31. Individuals of Black African, Black Caribbean, and mixed backgrounds (primarily White and Black Caribbean/African) colored by self-identified ethnic background (left, from figure **??**), FEV1 (middle), and age-adjusted height (right). An arrow points to an area where the FEV1 distribution appears to change, corresponding to where the clusters contain more people with self-identified mixed backgrounds.

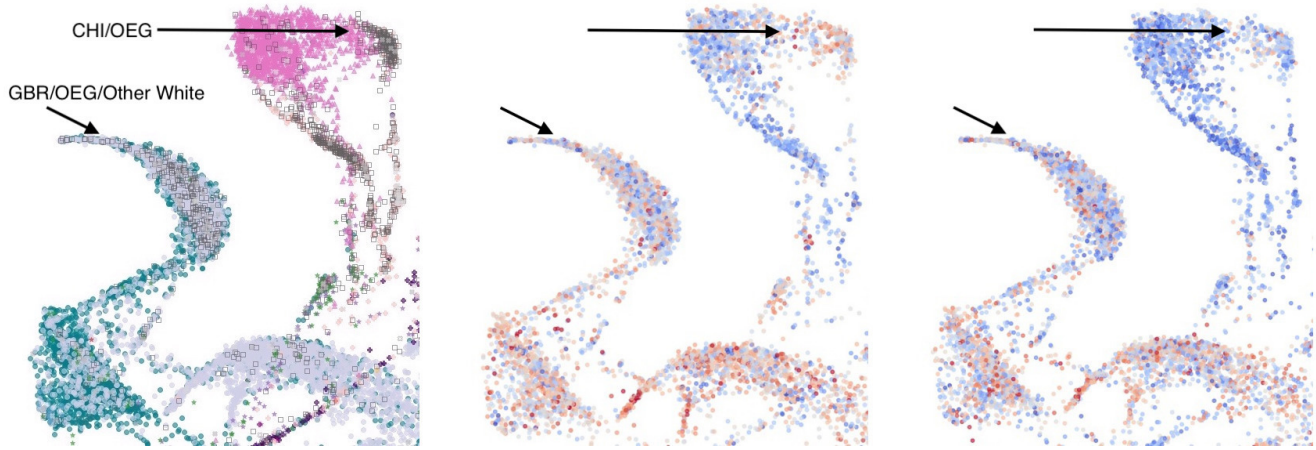


Fig. S32. Zoomed in section of figure **??** focused on individuals with Chinese (CHI), White British (GBR), any other white background, or any other ethnic group (OEG) colored by ethnicity (left), FEV1 (middle), and age-adjusted height (right). The OEG cluster next to the Chinese cluster is colored differently, suggesting this population may have different FEV1 characteristics. A cluster of OEG/other white individuals is more blue, suggesting they may have lower than average FEV1 values relative to the rest of the British or white population.

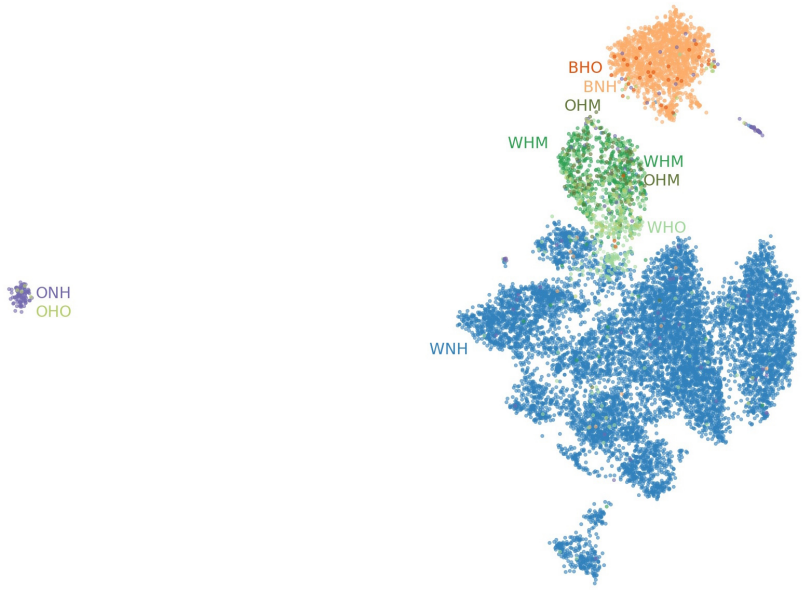


Fig. S33. UMAP applied to the first 10 principal components of HRS data. Points colored by self-identified race, Hispanic status, and Mexican-American status. The cluster on the left is mostly people who identify as neither Black nor White and were born outside the contiguous United States or in the Pacific census region. Clustering with the 1KGP data places them with Asian-identified populations. BNH, Black (not Hispanic); BHO, Black (Hispanic, Other); WNH, White (not Hispanic); WHM, White (Hispanic, Mexican-American); WHO, White Hispanic (Other); ONH, Other (not Hispanic); OHM, Other (Hispanic, Mexican-American); OHO, Other (Hispanic, Other).

Fig. S34. UMAP on the top 10 principal components of the HRS dataset, colored by Census Bureau birth region. There is no obvious pattern in the clusters of majority "White Not Hispanic" individuals.

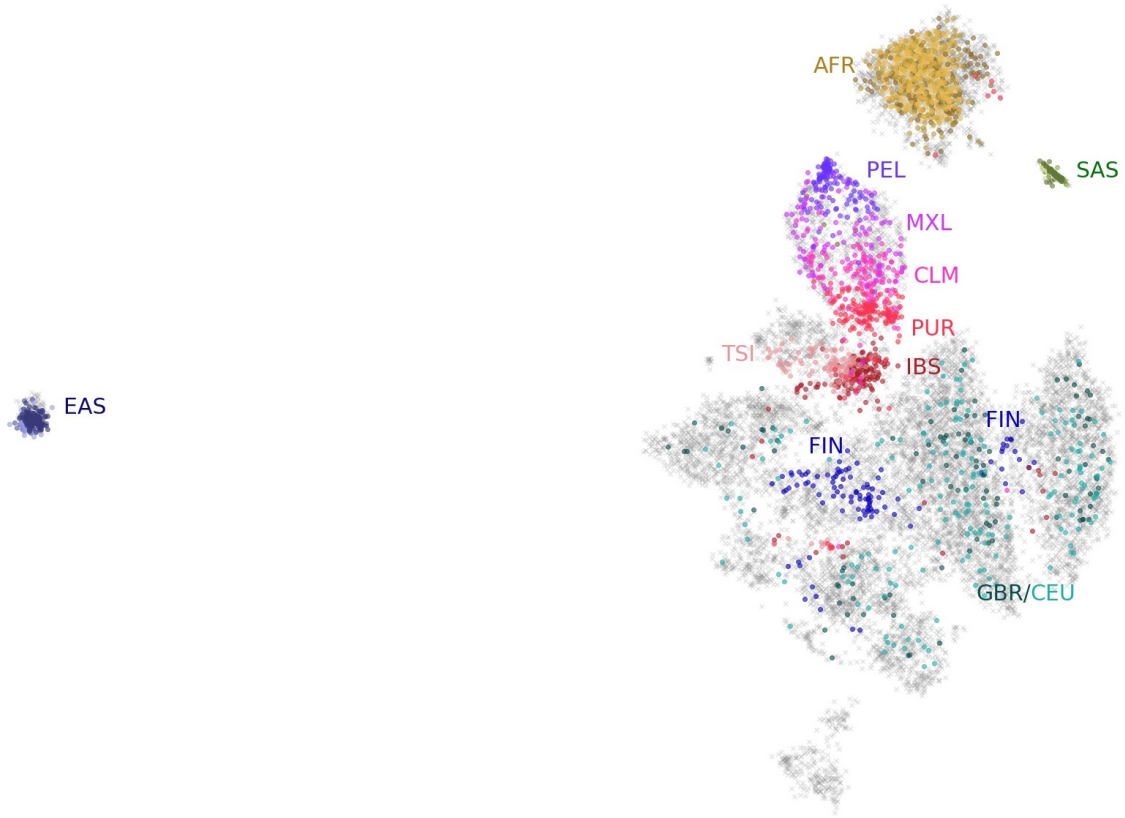


Fig. S35. ADP: This is also new UMAP on the top 10 principal components of the HRS data, with 1KGP data projected onto the embedding. Individuals from the HRS are coloured grey and largely transparent. British (GBR) and other European (CEU) individuals are scattered throughout the "White Not Hispanic" clusters. Finns (FIN) form clear groupings. Spanish (IBS) and Italian (TSI) individuals cluster near the Hispanic grouping. There is a clear gradient in the Hispanic cluster formed of Puerto Ricans (PUR), Colombians (CLM), Mexicans (MXL), and Peruvians (PEL). Populations with African ancestry (AFR) appear with Black individuals. East Asian (EAS) populations comprising Chinese, Kinh, and Japanese individuals cluster together with what appears in figure **??** as a population of mostly Asian ancestry. South Asian (SAS) populations with Indian, Pakistani, and Sri Lankan ancestry cluster in a separate area. One "White Not Hispanic" cluster at the bottom does not cluster with any 1KGP populations.

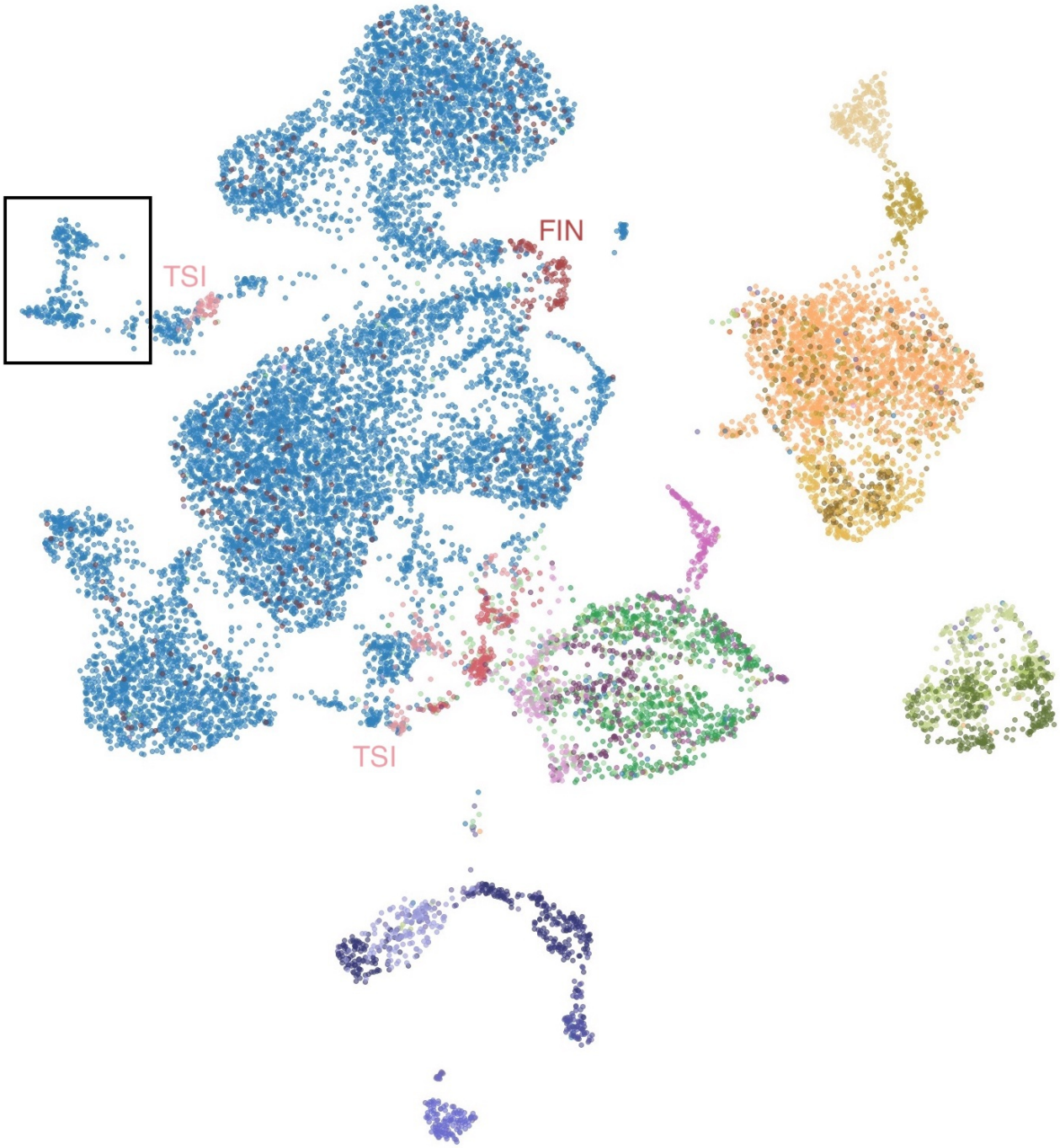


Fig. S36. UMAP on the top 10 principal components of the combined HRS and 1KGP datasets. The colors match those used in the separate 1KGP and HRS plots. The large blue clusters are "White Not Hispanic" individuals. Scattered brown dots within these clusters are Utah residents with Northern/Western European ancestry, and British in England and Scotland ("CEU" and "GBR", respectively, in the 1KGP datasets). A cluster of Finnish individuals consistently shows up and is labelled "FIN". Toscani individuals regularly form clusters and are labelled "TSI". However, their locations relative to the other clusters of "White Not Hispanic" are not consistent, sometimes appearing next to clusters of Spanish individuals and other times not. One cluster of "White Not Hispanic" individuals is highlighted in a box — unlike the other clusters, CEU and GBR individuals are not projected here.

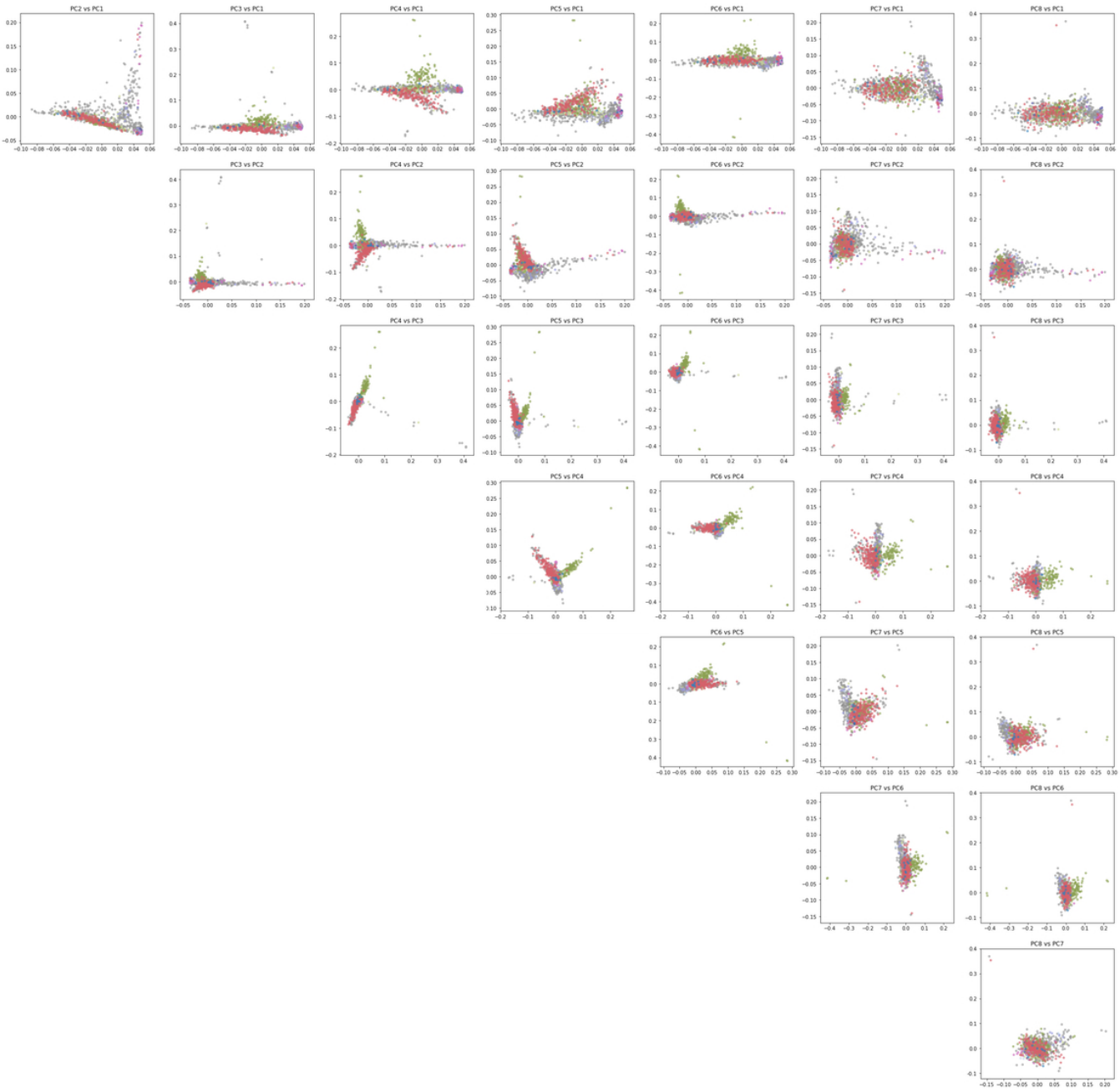


Fig. S37. Pairwise plots of the first 8 principal components of the Hispanic subset of the HRS. Those born in the Mountain region are colored green.

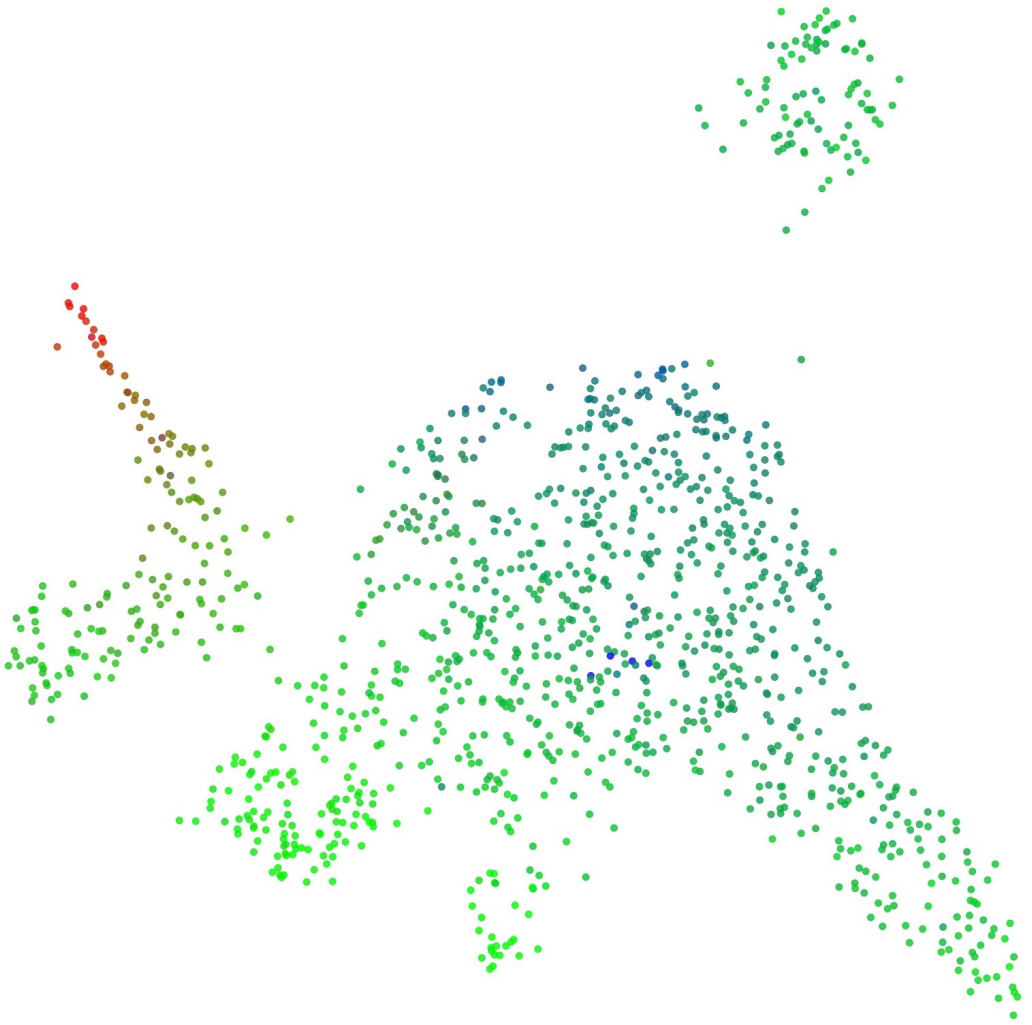


Fig. S38. UMAP of the first 7 principal components of the Hispanic population of the HRS, colored by estimated admixture proportions.

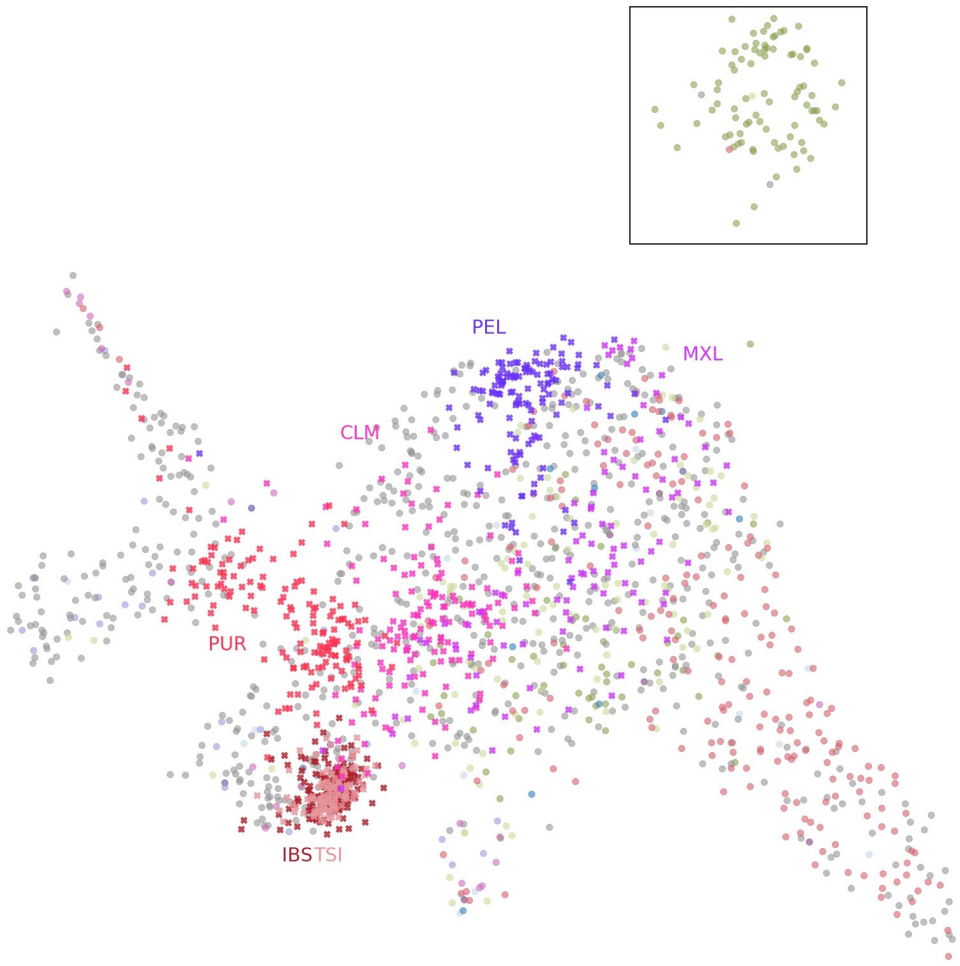


Fig. S39. UMAP of the first 7 principal components of the Hispanic population of the HRS, colored by birth region, with selected 1KGP populations projected onto the UMAP embedding. The highlighted area consists almost entirely of individuals born in the Mountain region of the census (New Mexico, Arizona, Colorado, Utah, Nevada, Wyoming, Idaho, and Montana). The six populations do not fall onto the highlighted cluster. CLM, Colombian in Medellin, Colombia; IBS, Iberian in Spain; MXL, Mexican in Los Angeles, California; PEL, Peruvian; PUR, Puerto Rican; TSI, Tuscani in Italy.

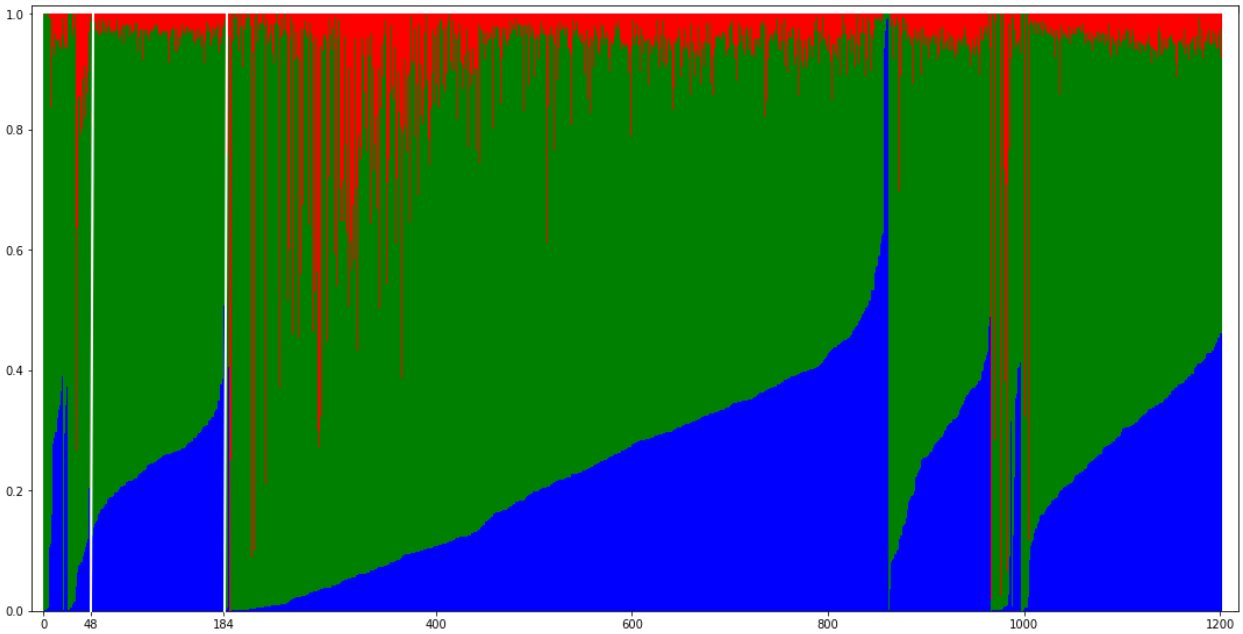


Fig. S40. Admixture plot of Hispanic individuals in the HRS. Those born in the Mountain census region fall between the white lines (indices 48 to 184)

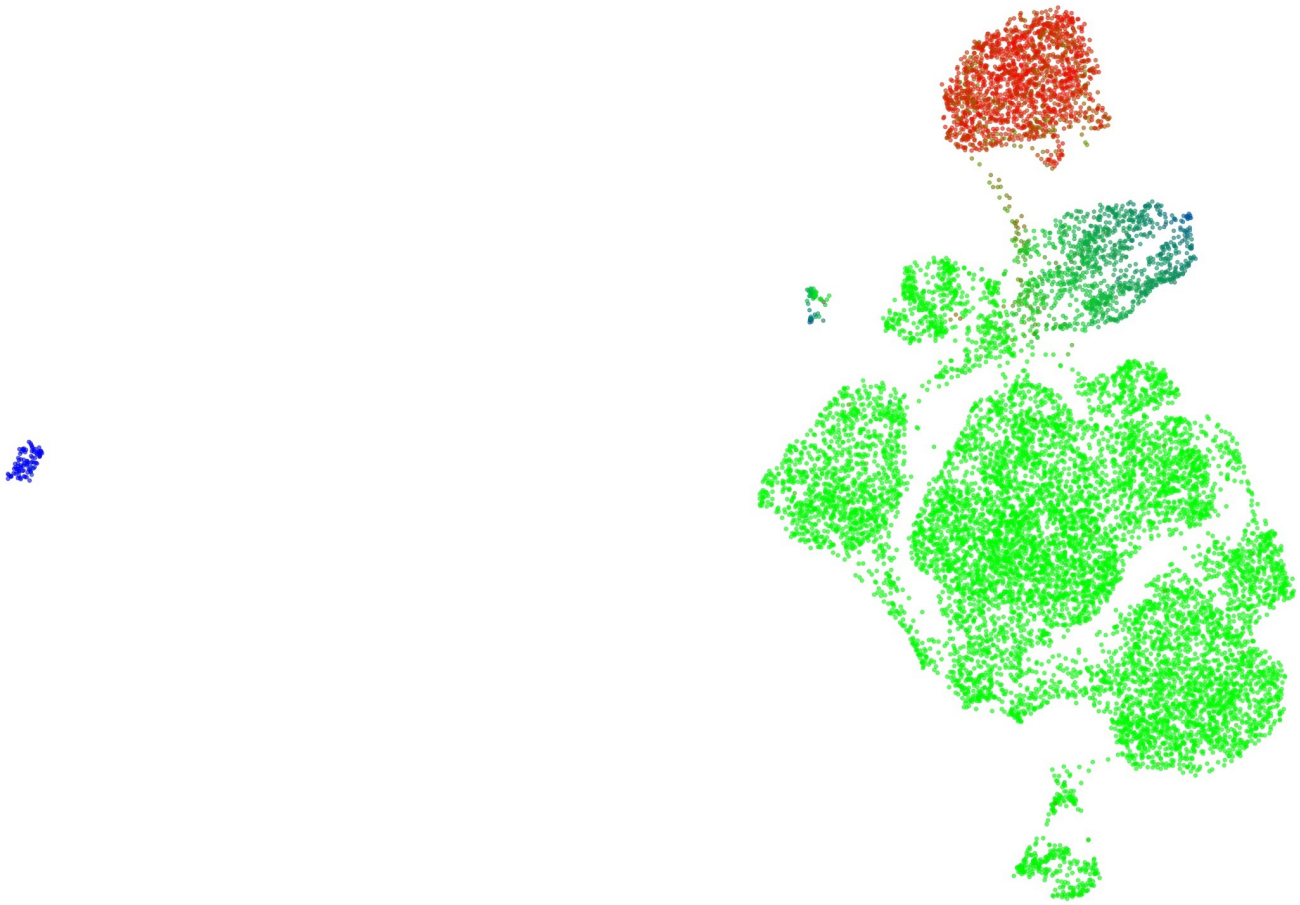


Fig. S41. UMAP on the first 10 principal components of HRS data, coloring individuals by estimated admixture from three ancestral populations. Alternate projection to demonstrate how individuals may fall between clusters.

Fig. S42. t-SNE applied to the top 10 principal components of the UKBB, colored by ethnic background. The unbalanced populations resulted in many individuals and populations being orphaned along the periphery of the main cluster.