Supplemental materials

Data harmonization

Genotype data

Genotype data was processed as described in (Chen et al. 2024). Briefly, reads were mapped using BWA-MEM, then filtered using the GATK Best Practices pipeline (DePristo et al. 2011), and gVCFs were generated using GATK HaplotypeCaller (Poplin et al. 2018). Joint calling was performed using the Hail combiner (Hail Team 2021) and converted to a VariantDataset (VDS), which was then densified into a dense MatrixTable used for analysis. These datasets are released on Google Cloud Platform, Amazon Web Services, and Microsoft Azure, and can be found on the Downloads page of the gnomAD browser (https://gnomad.broadinstitute.org/downloads#v3-hgdp-1kg).

Meta-data

Where possible, we combined meta-data from the 1000 Genomes Project and HGDP by combining the “super population” data from the 1000 Genomes project (1000 Genomes Project Consortium et al. 2015) and region information from HGDP (Bergström et al. 2020). We created a harmonized combined label with 3-letter codes for all groups, which we refer to as geographical/genetic regions throughout the text. Where a region was only clearly contained in HGDP, we used the HGDP information to define a 3-letter code. The CENTRAL_SOUTH_ASIA code contained within HGDP is more geographically expansive than the SAS label contained in the 1000 Genomes Project, so we expanded the 3 letter code to be CSA, as shown in **Table S1**.

After combining region data, we used principal components analysis (PCA) to identify ancestry outliers within regions. We identified outliers as described in **Table S2** and provide final sample counts in **Table S3**.

QC Meta-data Summaries

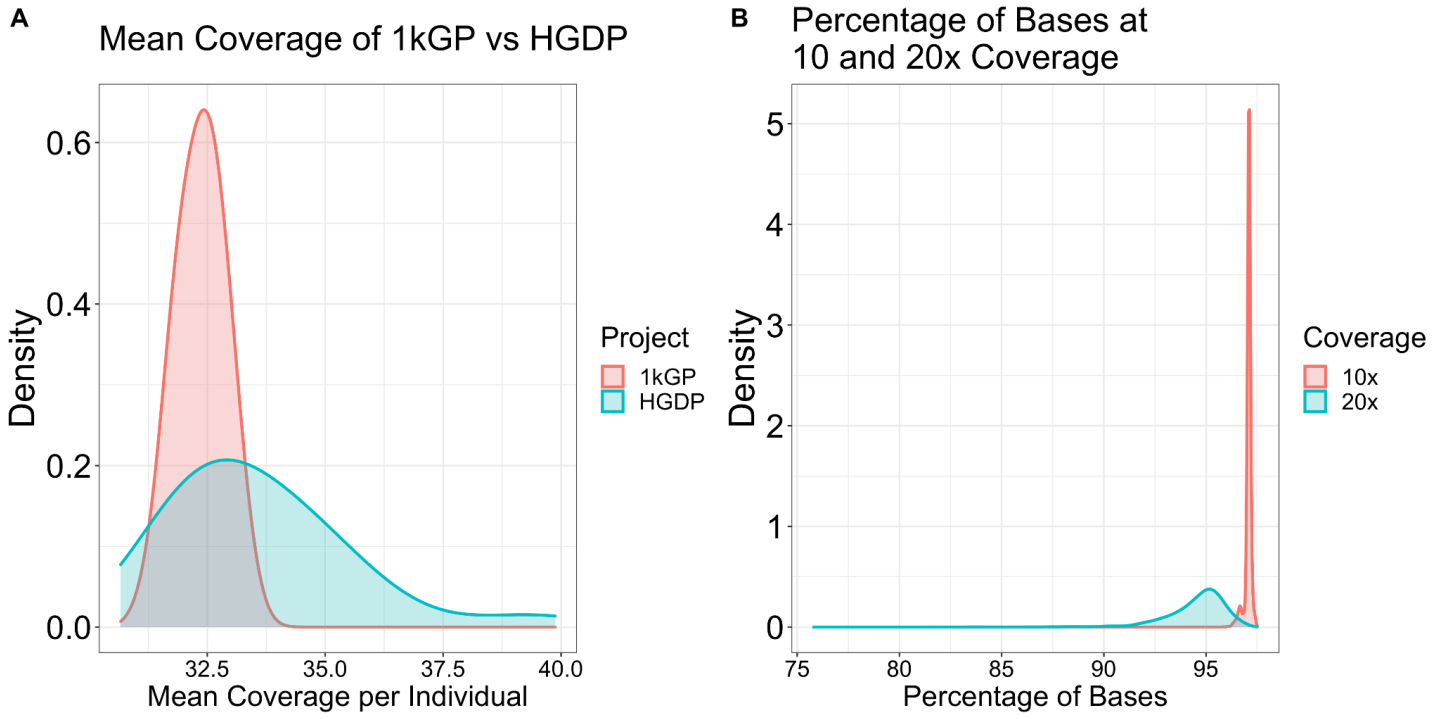

Figure S1 | Coverage across the 1kGP and HGDP.
A) Coverage in both datasets is uniformly above 30×, with an average of 33× coverage across the harmonized dataset. The coverage of the HGDP genomes is more variable than in 1kGP, as expected based on a variety of technical differences such as multiple sequencing batches, PCR+ vs PCR-free, and older cell lines in HGDP compared to 1kGP. The differences in project coverages also impacts the distribution of coverage statistics by Geographical region given their tally by project (**Table S4**). The overall coverage distributions by population are shown in **Figure S2**. B) Over 95% of bases are covered over 10×, and over 90% of bases are covered over 20× in HGDP+1kGP.

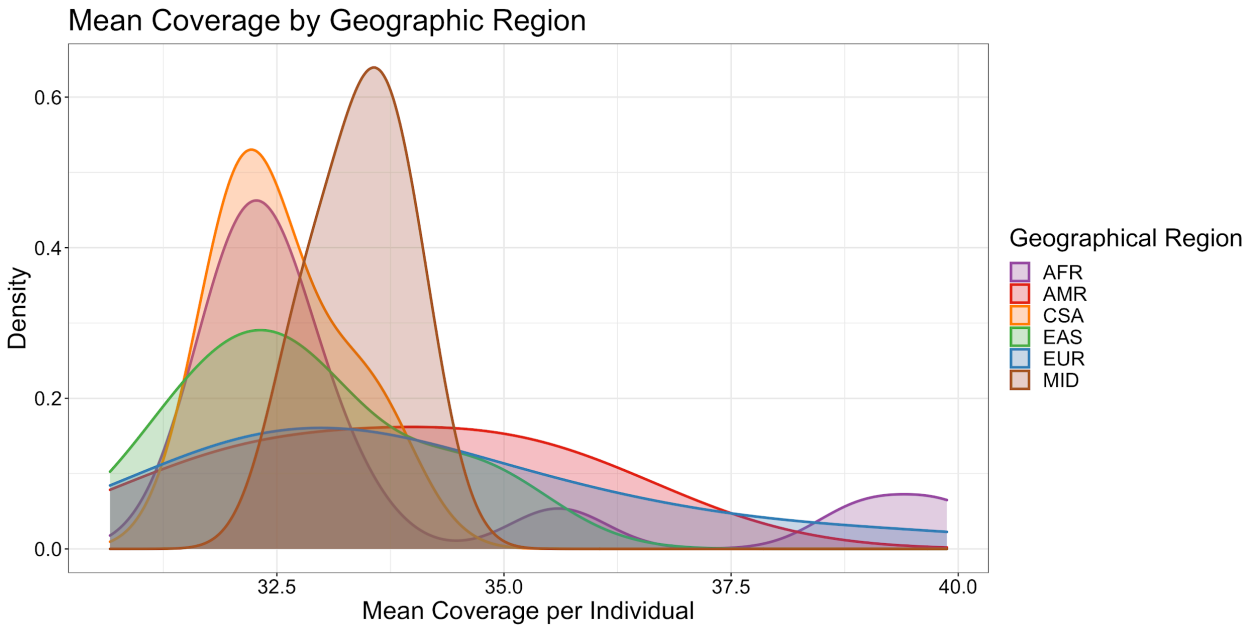

Figure S2 | Coverage across 1kGP and HGDP by population.
Regional abbreviations are as described in **Table S1**. OCE is excluded from this plot as it is represented by only two populations. Mean coverage across the different regions is 33× with coverage consistently above 30× for all regions.

Structural variants (SVs)

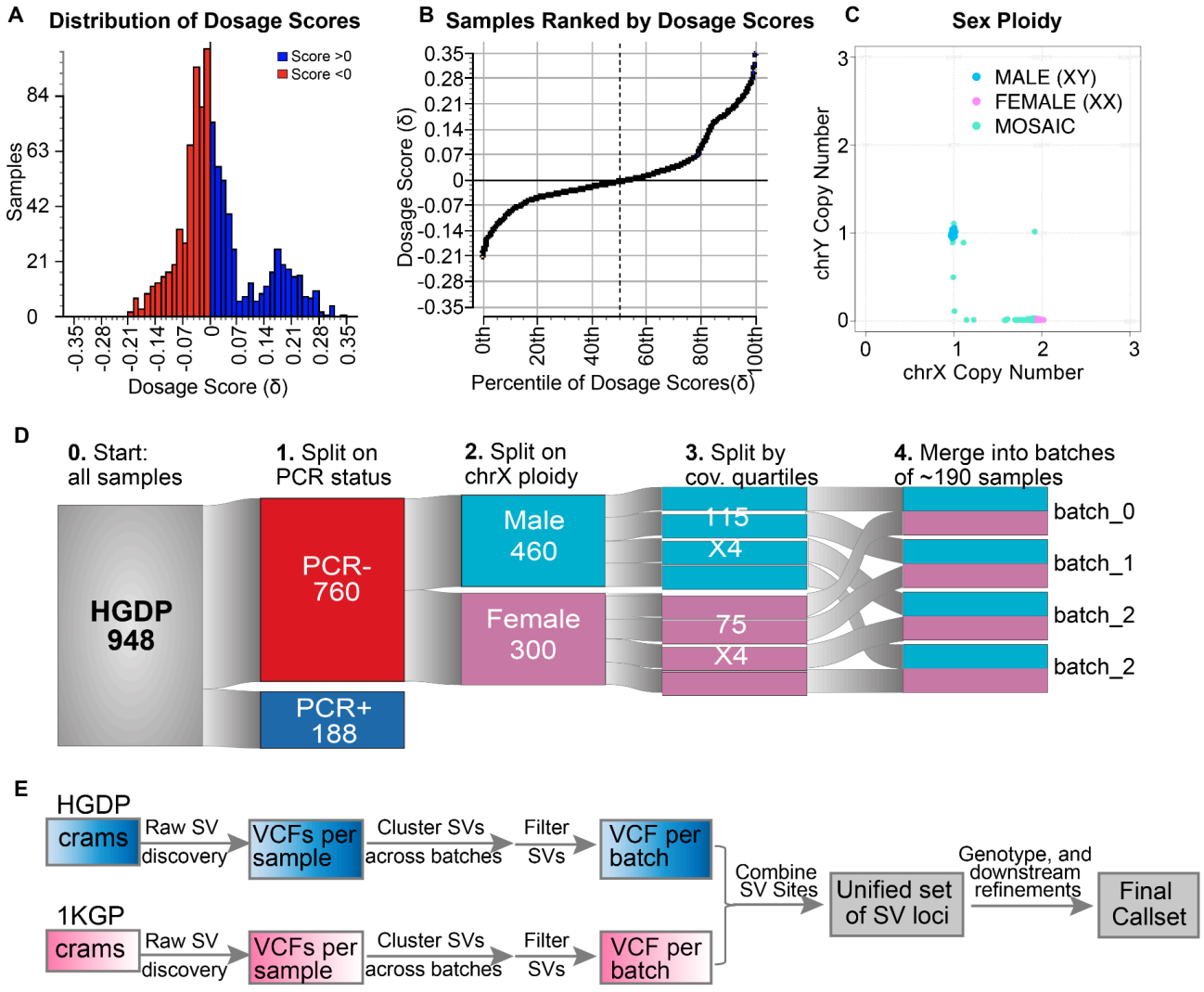

Figure S3 | Dosage and sex ploidy of HGDP samples and batching strategy.
A) Distribution of dosage scores across HGDP samples. We used the previously developed whole genome dosage model (Collins et al 2020) to quantify non-uniform distribution of sequencing coverage.The dosage scores corresponded predominantly to PCR-amplified (PCR+) and PCR-free (PCR-) library protocols. B) Samples ranked by dosage score. C) Distribution of Chr X copy number across HGDP samples. D) Batching strategy for SV calling. HGDP samples were first split by their PCR status and Chr X ploidy. PCR- samples were then ranked by their sequencing depth from low to high, and split into four sub batches of equivalent sizes. Male and female batches with matched coverage quantiles are combined to form the final batches. E) Workflow of SV discovery from the HGDP and 1KGP genomes. The HGDP and 1KGP samples have been processed separately through the first steps of GATK-SV (Collins et al. 2020), including raw SV discovery, batching SVs across each batch and initial filtering of SVs using the “FilterBatch” method in GATK-SV (Collins et al. 2020). The filtered SVs were then merged across HGDP and 1KGP to form a non-redundant set of SV loci, systematically genotype across both HGDP and 1KGP samples, and processed through downstream steps of GATK-SV ((Collins et al. 2020), see https://github.com/broadinstitute/gatk-sv for details).

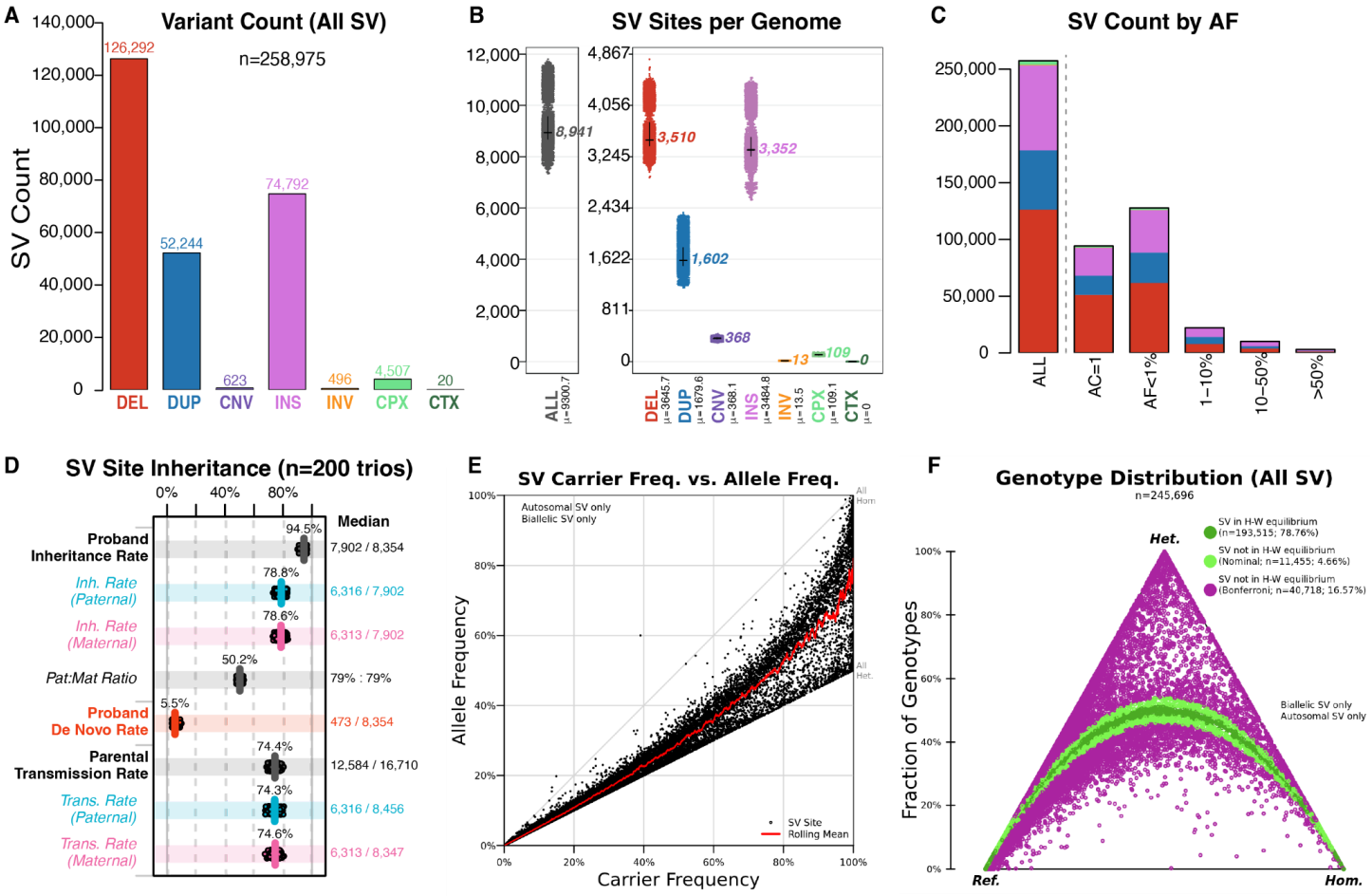

Figure S4 | SV callset and quality evaluation results.
A) Count of SV sites across 4,151 HGDP and 1kGP samples by variant type. B) Count of SVs per genome by variant type. Median counts of SVs per genome by SV type are annotated in the figure, and mean SV counts are annotated on x-axis. C) Count of SV sites by allele frequency. D) Inheritance of SVs calculated in 100 pather-mother-child trio families. Proband Inheritance Rate - proportion of SVs in children's genome that were inherited from either parents; Paternal Inheritance Rate - proportion of SVs in children's genome that were shared by paternal genome; Maternal Inheritance Rate - proportion of SVs in children's genome that were shared by maternal genome; Parental Transmission Rate - proportion of SVs in parents' genome that were transmitted into children's genome; Trans. Rate (Paternal) - proportion of SVs in paternal genome that were transmitted into children's genome; Trans. Rate (Maternal) - proportion of SVs in maternal genome that were transmitted into children's genome. E) Correlation of allele frequencies. F) Hardy-Weinberg Equilibrium distribution of SVs across all samples. Each point is a single biallelic autosomal SV projected onto HWE ternary axes corresponding to its ratio of homozygous reference (0/0), heterozygous (0/1), and homozygous alternate (1/1) genotypes across all samples in the indicated population. The distance of a point to a vertex indicates the fraction of samples with that genotype. Deviation from HWE was assessed using a chi-square goodness-of-fit test with one degree of freedom, and points are colored based on their p-value. Green points are SVs within bounds defined for HWE based on the number of sites documented in each population, and purple points are SVs outside of these p-value bounds. The proportion of SVs corresponding to each p-value cutoff is provided at the right of each panel. Plots were generated using the "HardyWeinberg" package in R (Team 2013).

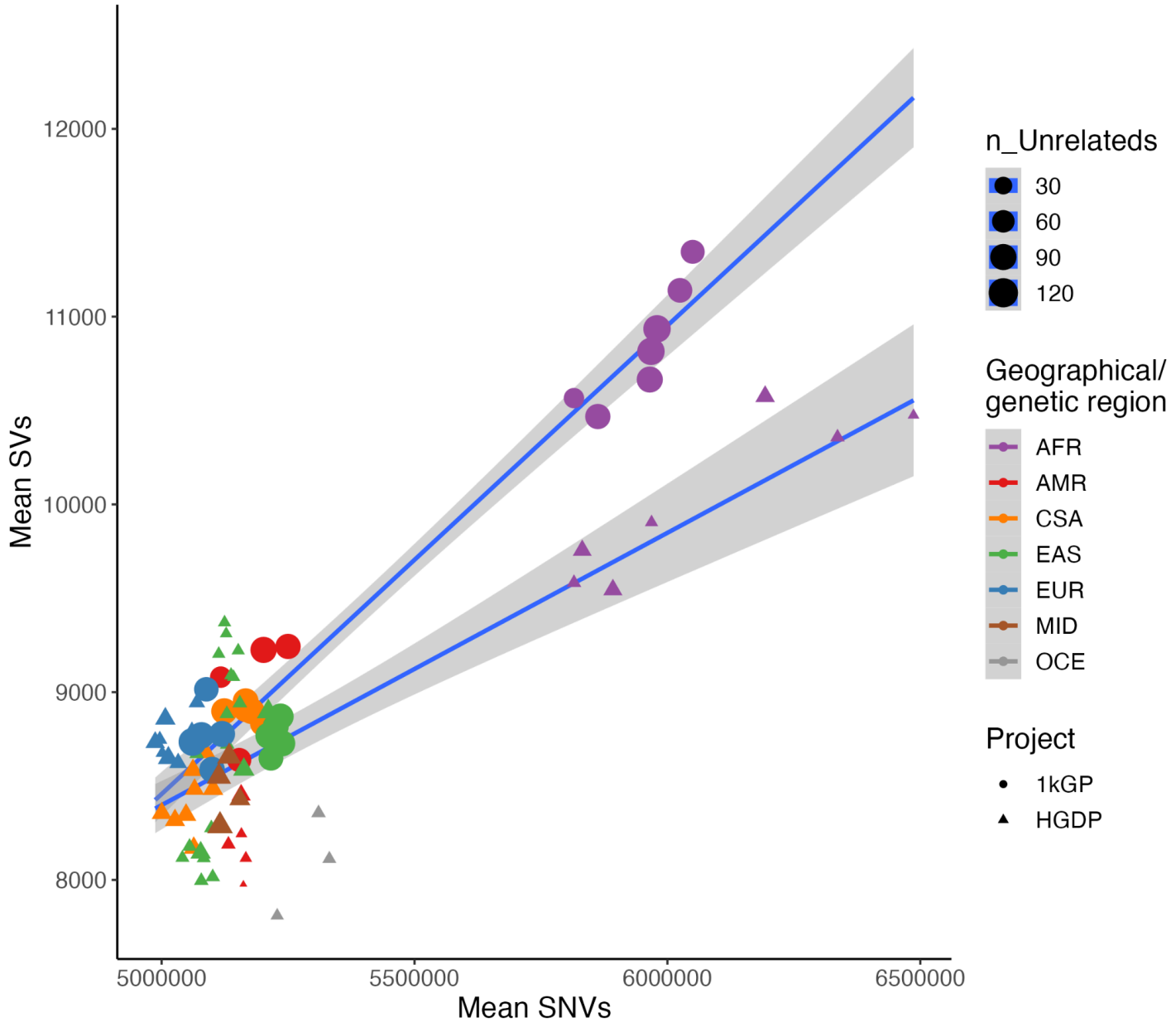

Figure S5 | Mean count of SVs versus SNVs by project, region, and number of individuals. Top line shows a fitted regression line to the 1000 Genomes Project points, and bottom line is fitted to HGDP points. A larger number of SVs are present in the 1000 Genomes Project data, which was explored more fully in **Figure S6**.

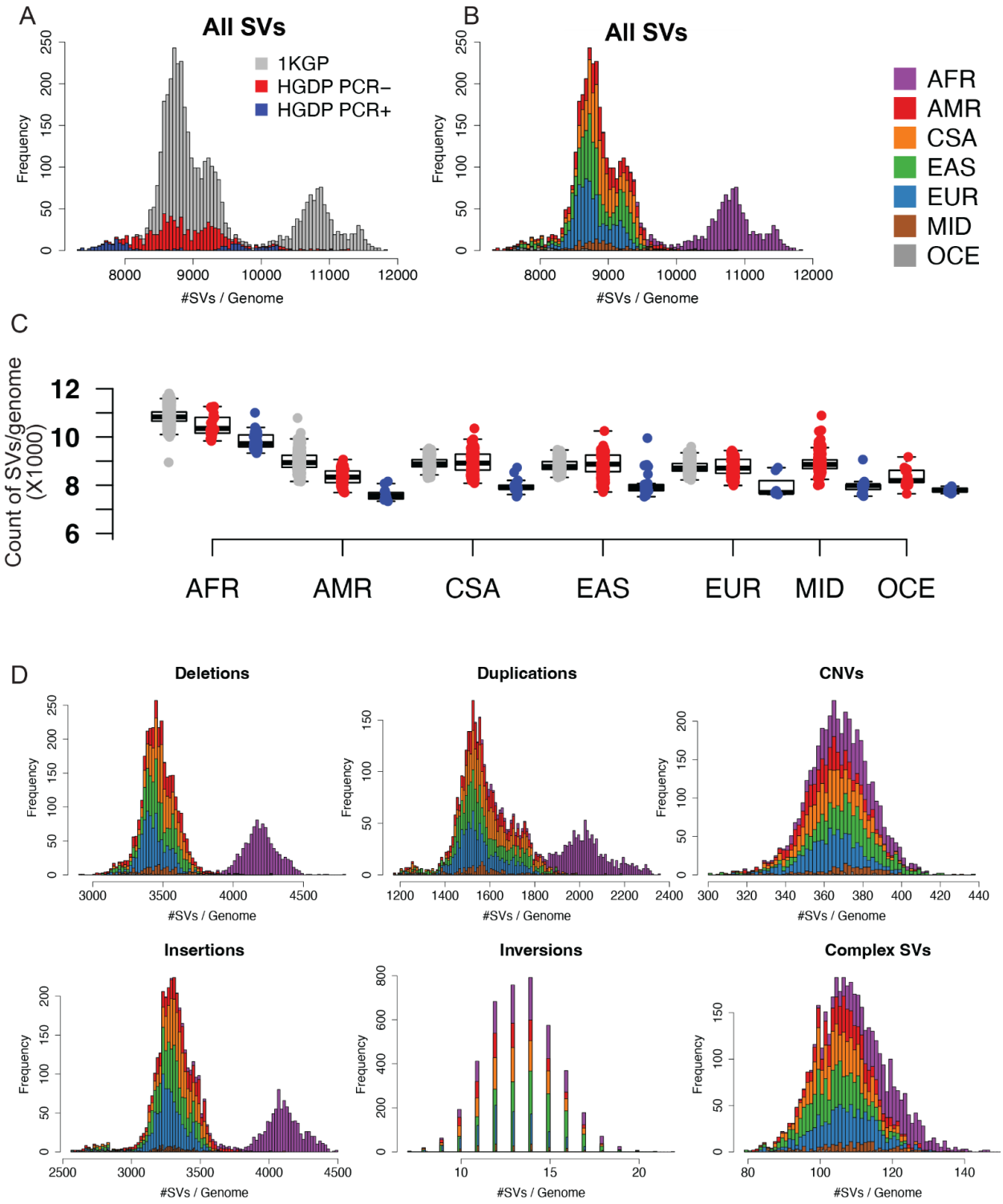

Figure S6 | SV breakdown in count by class across HGDP and 1kGP (HGSV).
Per genome SV counts by study and PCR status (A,C), and population (B). Per genome SV counts are also broken down by SV type, including deletions, duplications, multi-allelic CNVs, insertions, inversions, and complex SVs in D).

Population genetic comparisons

The breakdown of ancestry and population structure by ADMIXTURE is similar to that identified in global PCA, with K=2 highlighting structure in the AFR, K=3 highlighting structure in the EAS, K=4 highlighting structure in the EUR and CSA, K=5 highlighting structure in the AMR, K=6 highlighting structure in the OCE, K=7 highlighting structure in the MID, and subsequent values of K highlighting structure within meta-data labels (**Figure S7**).

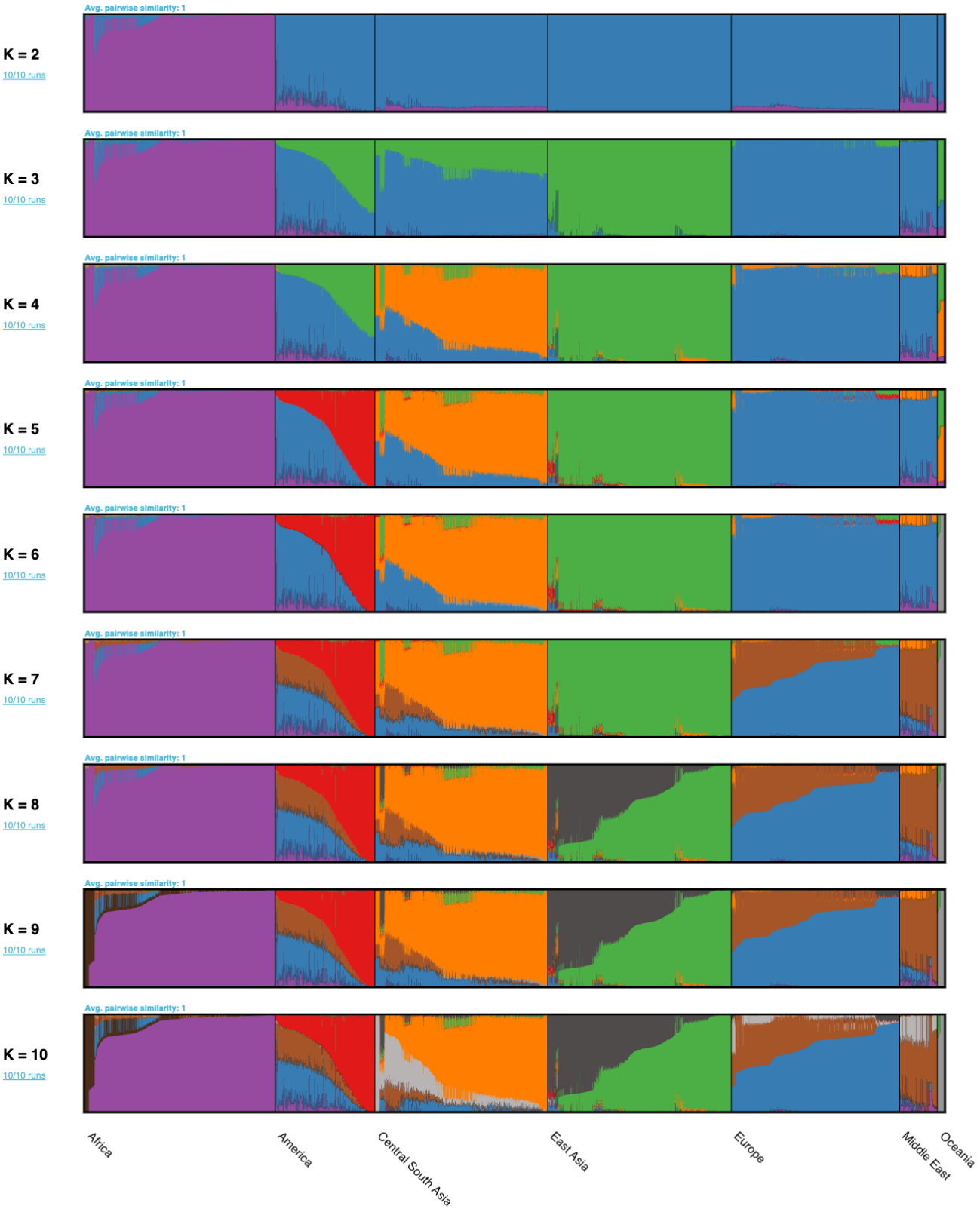

Figure S7 | ADMIXTURE analysis of the HGDP and 1kGP resource.
We ran ADMIXTURE with values of K=2 through K=10 across populations and harmonized geographical/genetic regions. Each row of bar plots shows the breakdown of regional substructure as K increases, where K is the number of genetic ancestry components fit in that run. For example, when K=2, AFR separates from the rest of the populations as the most distinct population due to high levels of genetic diversity.

When K=3 EUR separates from the rest, and so on. We chose the best fit value of K to be K=6 based on a reduction in the rate of change of 5-fold cross validation error as shown in **Figure S8**.

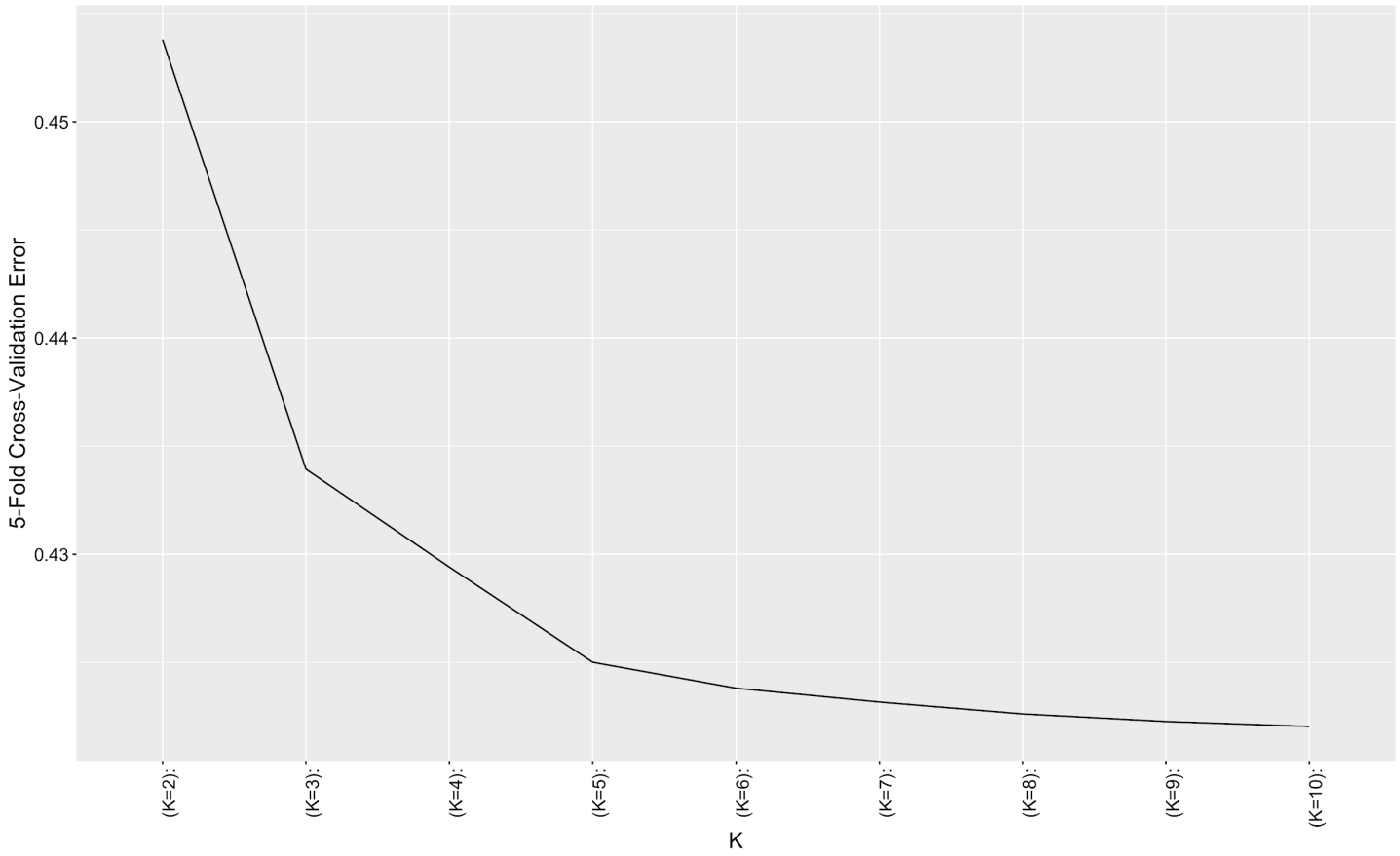

Figure S8 | 5-fold cross-validation error across ADMIXTURE runs.
We selected K=6 as the point at which cross-validation error leveled out. As described in the ADMIXTURE manual, the cross-validation error enables users to identify the value of K for which the model has best predictive accuracy, as determined by “holding out” data points. It partitions observed genotypes into 5 roughly equally sized folds, masks genotypes for each fold, then predicts the genotypes.

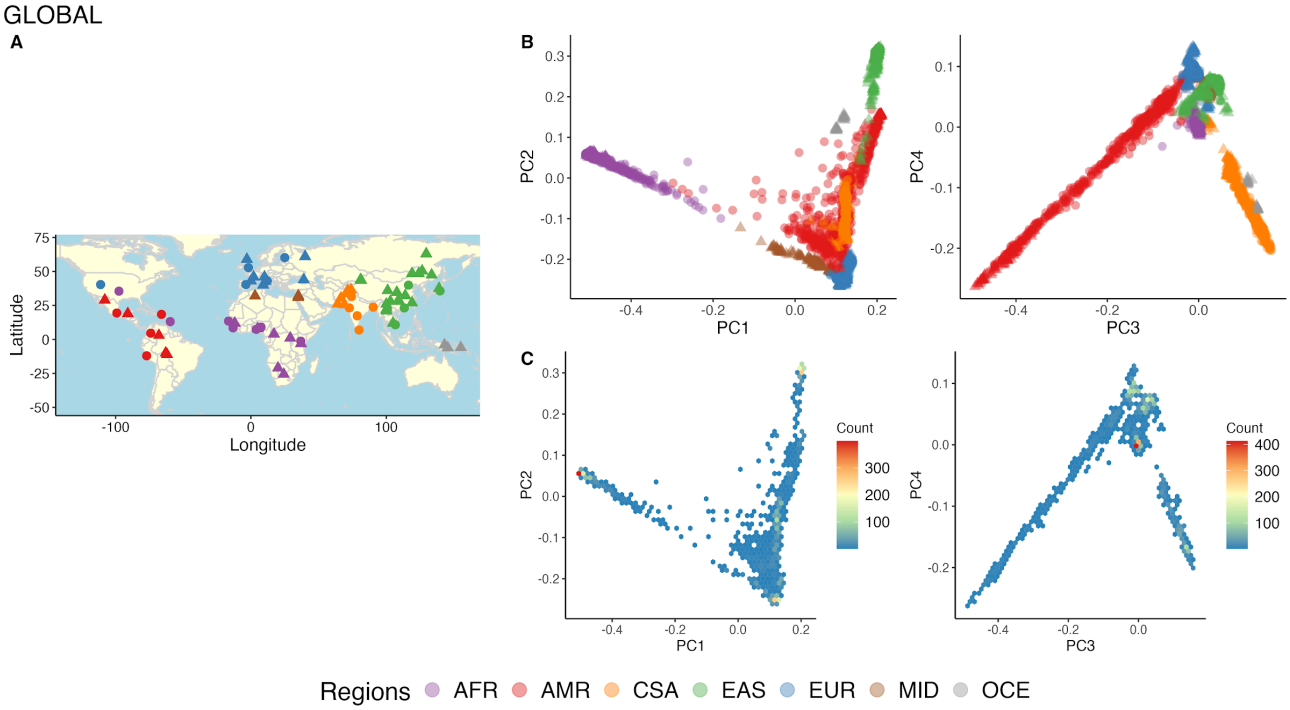

Figure S9 | PCA biplots and densities globally.
A) Map shows where all samples in analyses are from. B) PCA biplots of PCs 1-4. PCA outliers were removed prior to this analysis. Filled circles indicate populations in the 1000 Genomes Project, while filled triangles indicate populations in HGDP. Population codes are as in **Table S1**. C) Density plot of PCA biplots of PCs 1-4.

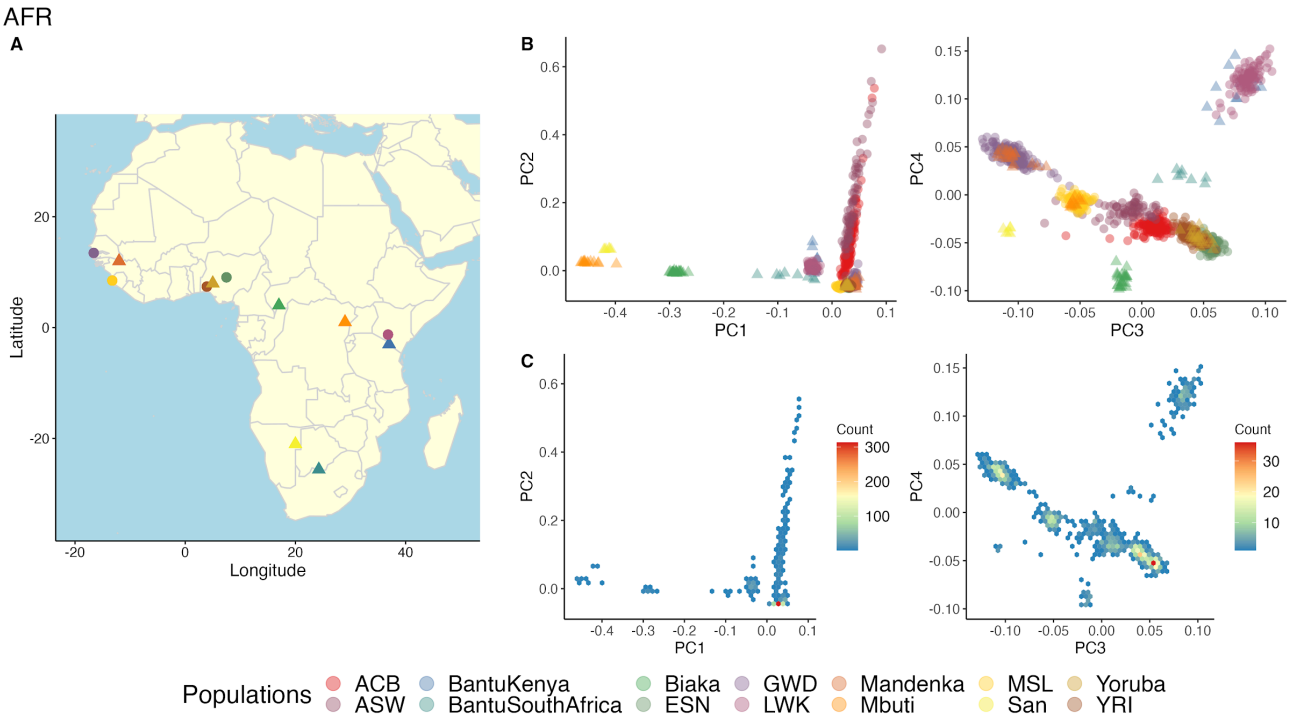

Figure S10 | Subcontinental PCA in AFR populations.
A) Map shows where all AFR samples in analyses are from. B) PCA biplots of PCs 1-4. PCA outliers were removed prior to this analysis. Filled circles indicate populations in the 1000 Genomes Project, while filled triangles indicate populations in HGDP. Population codes are as in **Table S1**. C) Density plot of PCA biplots of PCs 1-4.

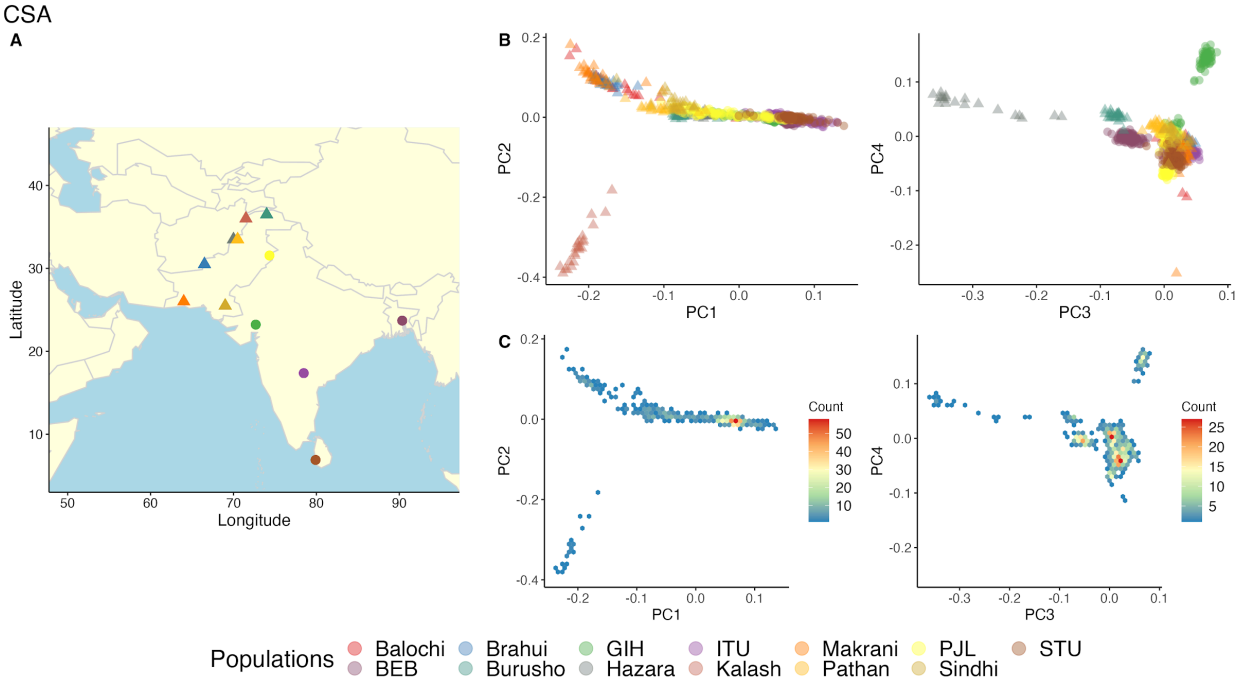

Figure S11 | Subcontinental PCA in CSA populations.
A) Map shows where all CSA samples in analyses are from. B) PCA biplots of PCs 1-4. PCA outliers were removed prior to this analysis. Filled circles indicate populations in the 1000 Genomes Project, while filled triangles indicate populations in HGDP. Population codes are as in **Table S1**. C) Density plot of PCA biplots of PCs 1-4.

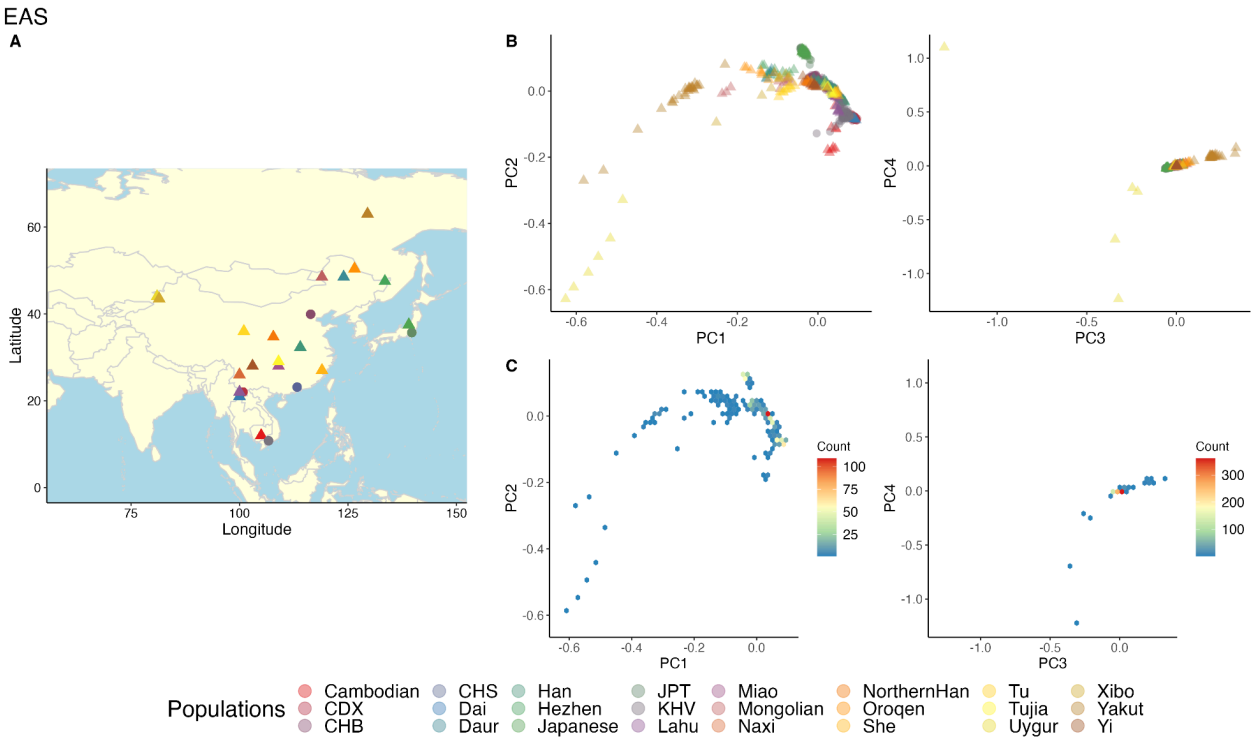

Figure S12 | Subcontinental PCA in EAS populations.
A) Map shows where all EAS samples in analyses are from. B) PCA biplots of PCs 1-4. PCA outliers were removed prior to this analysis. Filled circles indicate populations in the 1000 Genomes Project, while filled triangles indicate populations in HGDP. Population codes are as in **Table S1**. C) Density plot of PCA biplots of PCs 1-4.

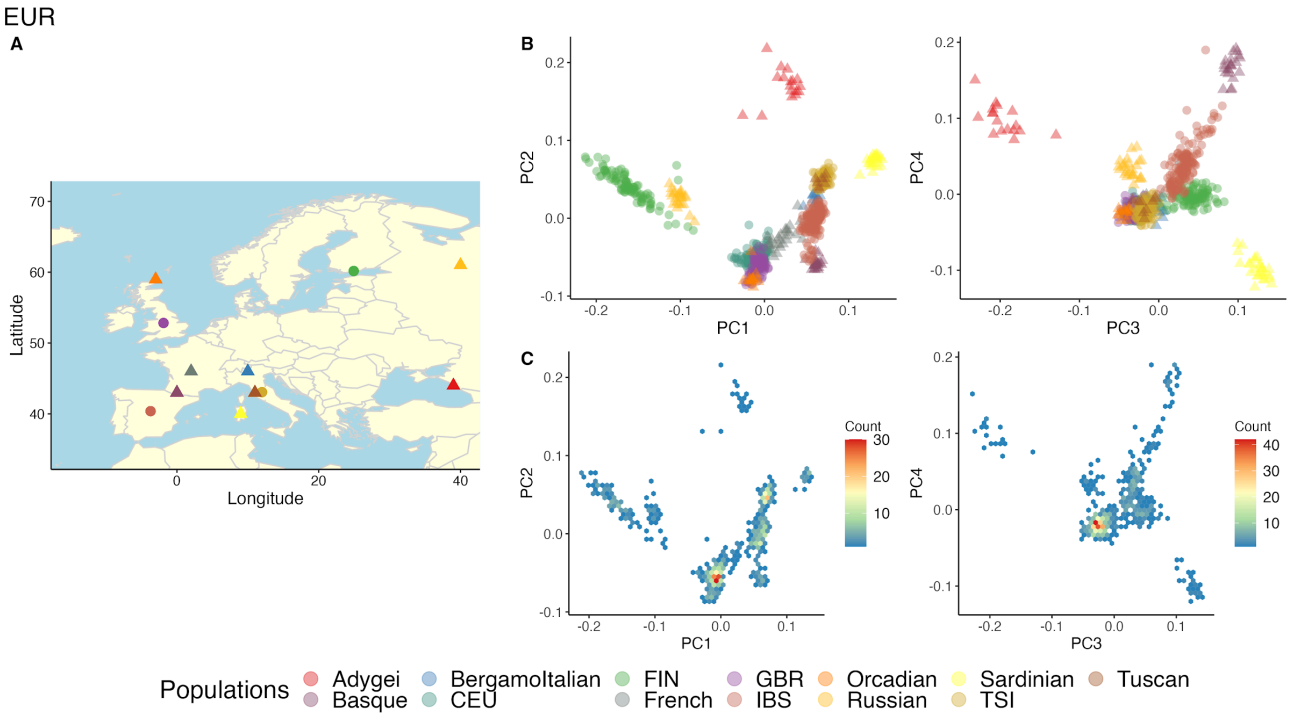

Figure S13 | Subcontinental PCA in EUR populations.
A) Map shows where all EUR samples in analyses are from. B) PCA biplots of PCs 1-4. PCA outliers were removed prior to this analysis. Filled circles indicate populations in the 1000 Genomes Project, while filled triangles indicate populations in HGDP. Population codes are as in **Table S1**. C) Density plot of PCA biplots of PCs 1-4.

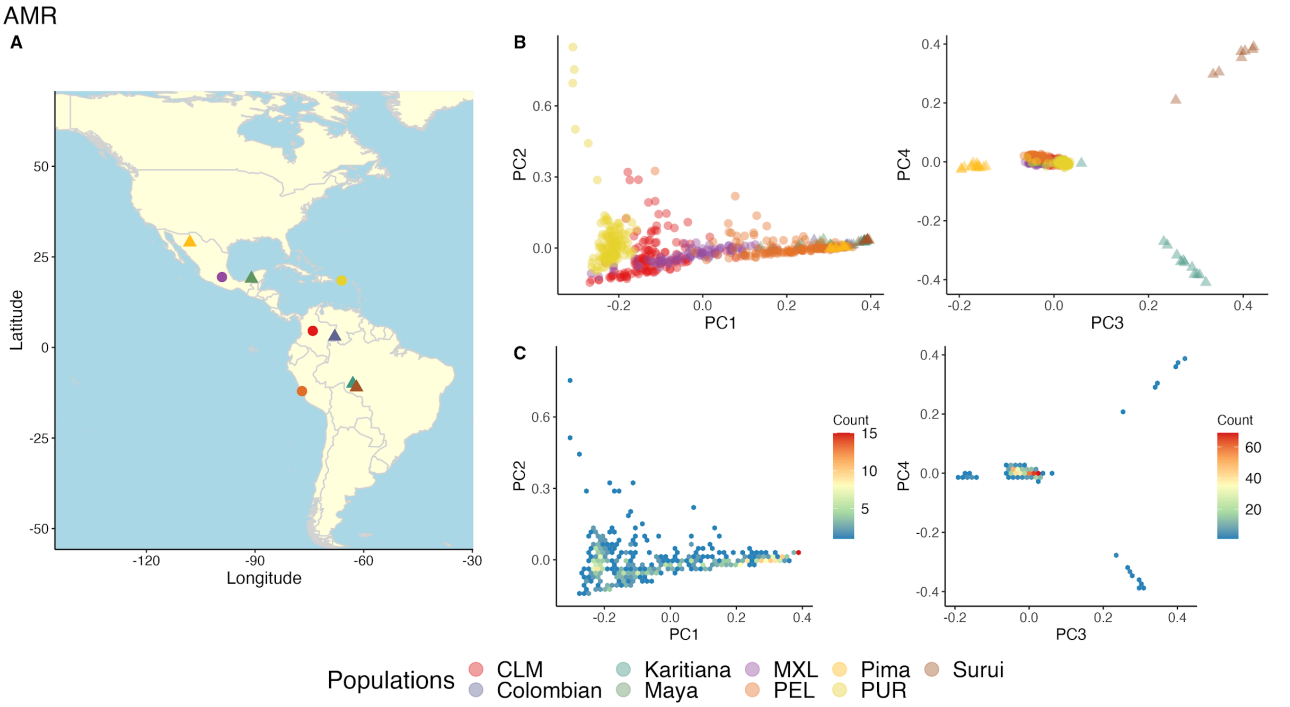

Figure S14 | Subcontinental PCA in AMR populations.
A) Map shows where all AMR samples in analyses are from. B) PCA biplots of PCs 1-4. PCA outliers were removed prior to this analysis. Filled circles indicate populations in the 1000 Genomes Project, while filled triangles indicate populations in HGDP. Population codes are as in **Table S1**. C) Density plot of PCA biplots of PCs 1-4.

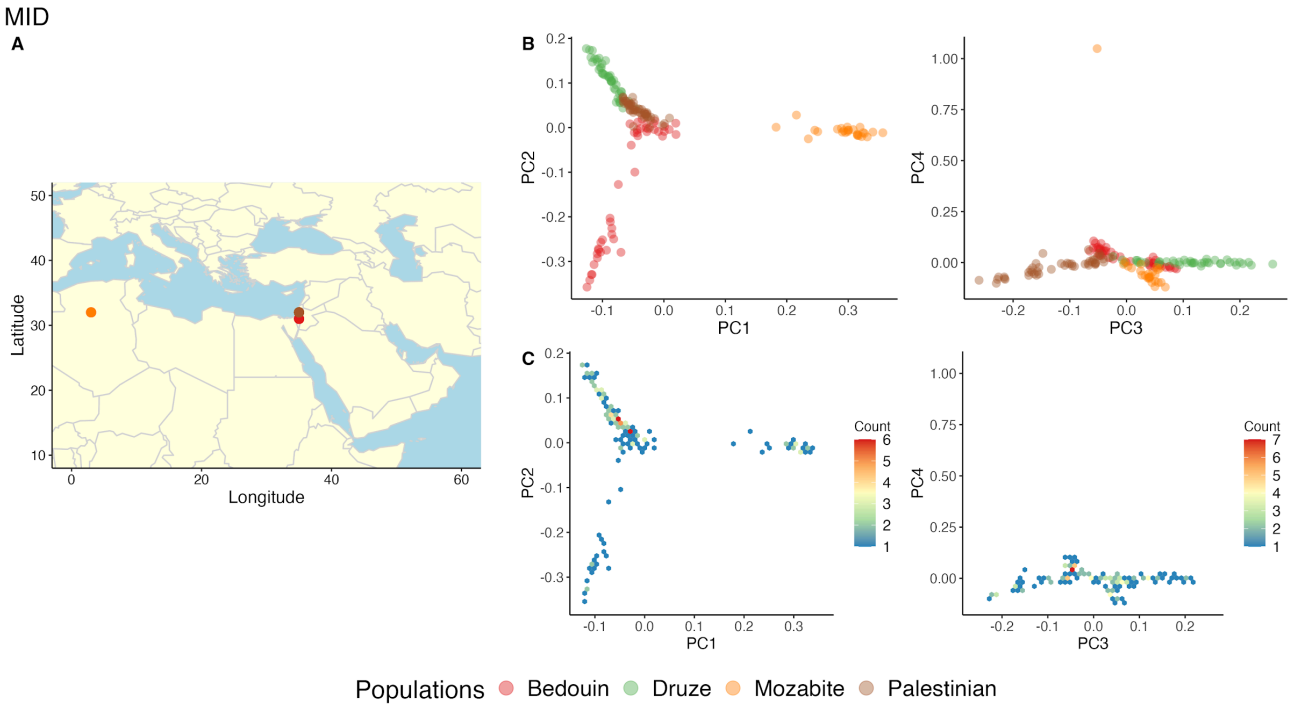

Figure S15 | Subcontinental PCA in MID populations.
A) Map shows where all MID samples in analyses are from. Palestinian and Druze have the same geographical coordinates. B) PCA biplots of PCs 1-4. PCA outliers were removed prior to this analysis. All MID populations are from HGDP. Population codes are as in **Table S1**. C) Density plot of PCA biplots of PCs 1-4.

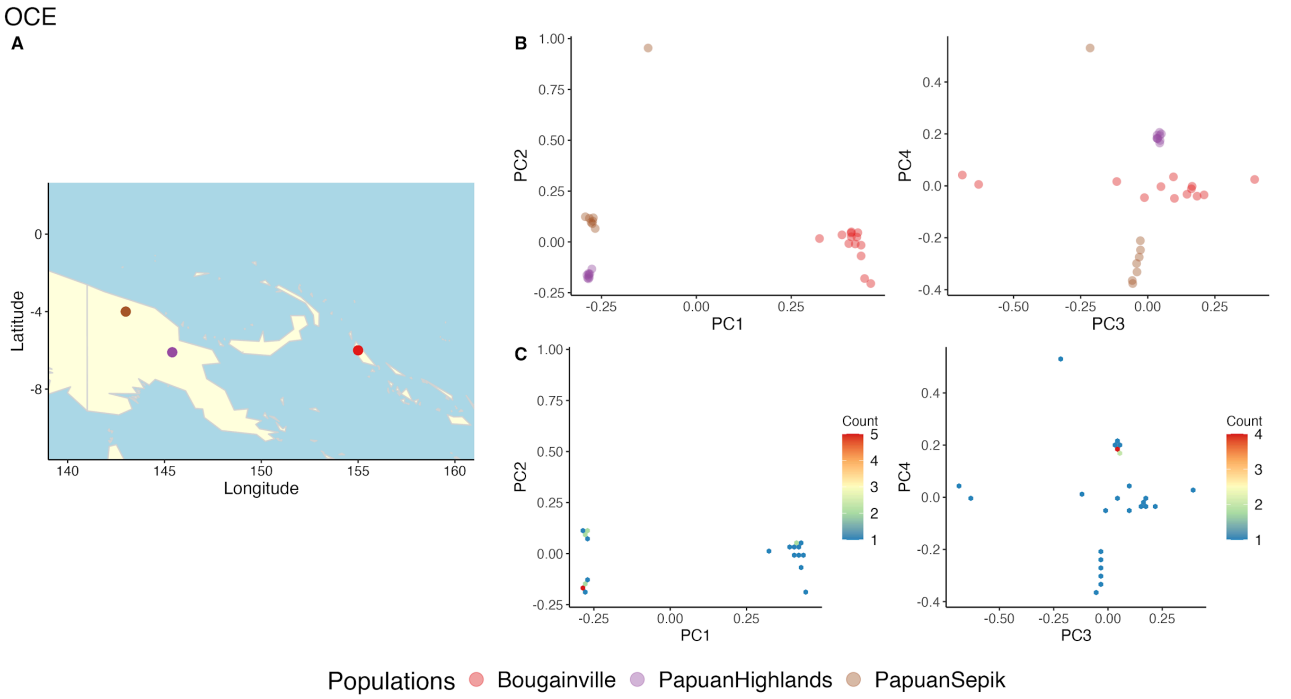

Figure S16 | Subcontinental PCA in OCE populations.
A) Map shows where all OCE samples in analyses are from. B) PCA biplots of PCs 1-4. PCA outliers were removed prior to this analysis. All OCE populations are from HGDP. Population codes are as in **Table S1**. C) Density plot of PCA biplots of PCs 1-4.

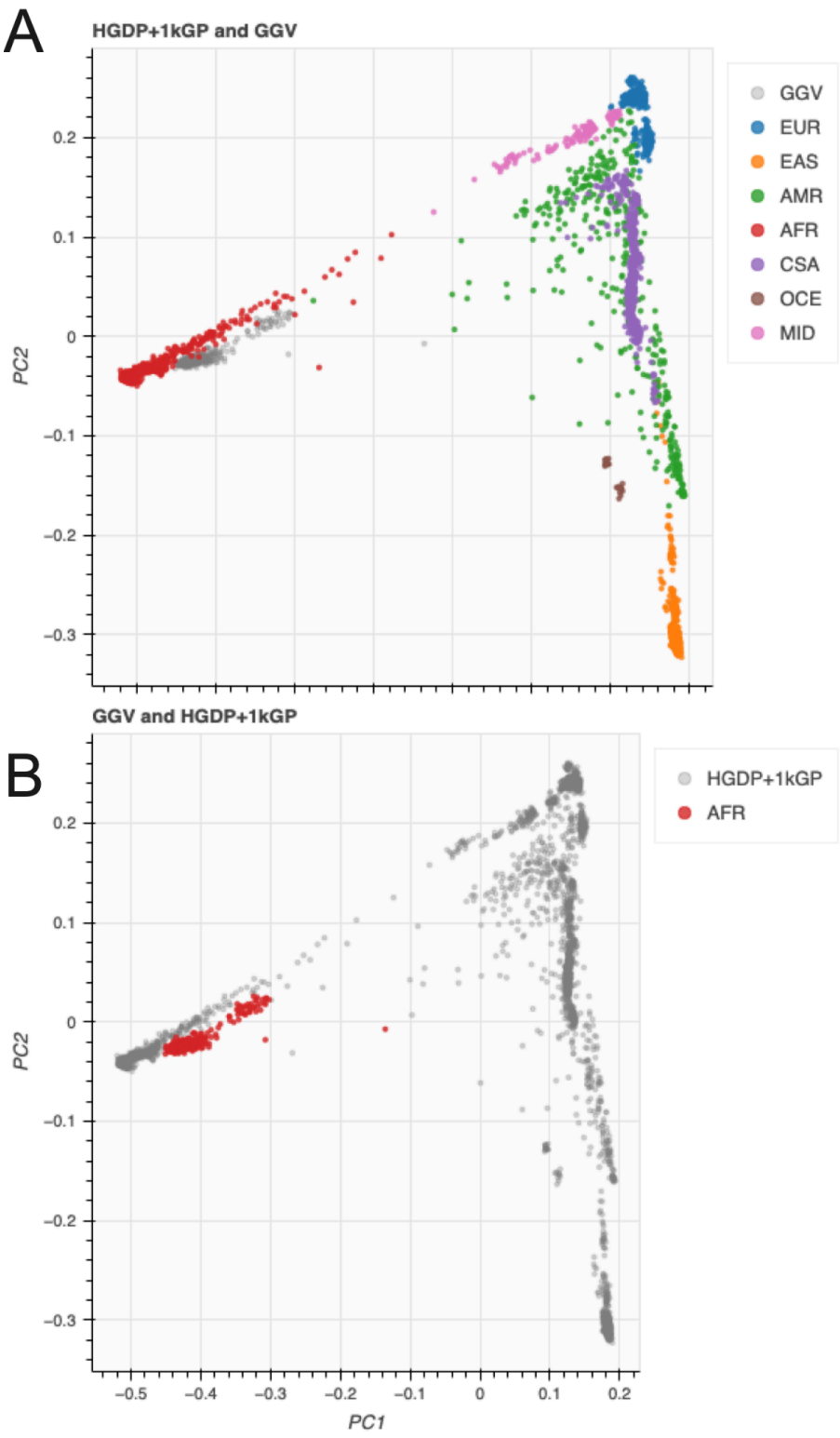

Figure S17 | HGDP+1kGP ancestry labels applied to the Gambian Genome Variation (GGV) Project. A) PCs 1 and 2 of all HGDP+1kGP samples with GGV projected into the same PC space, with each reference population colored and the GGV samples shown in grey. B) The same PCs with the reference data shown in grey and the GGV samples showing the assigned ancestry–all AFR.

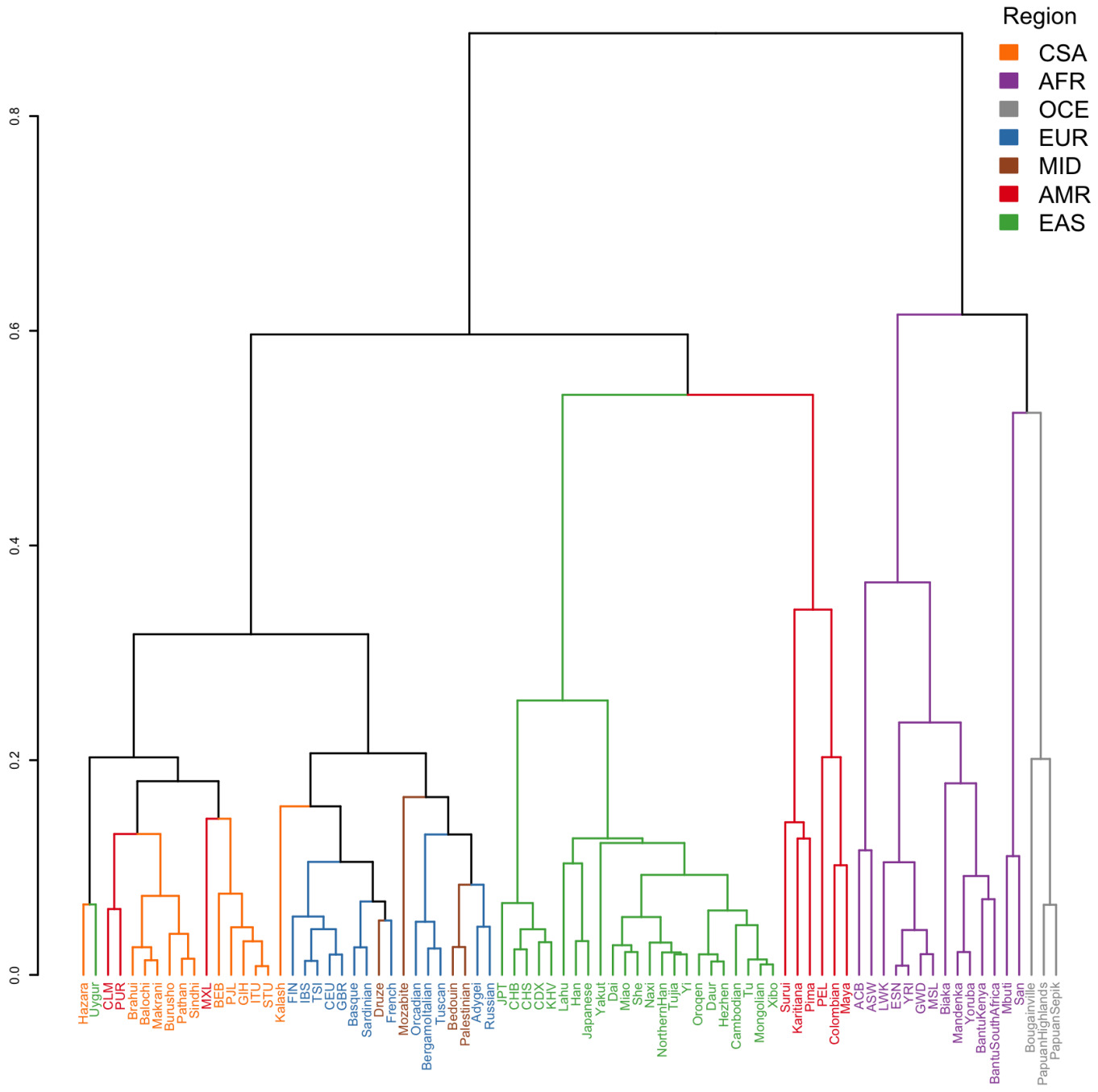

Figure S18 | Dendrogram of the pairwise F_{ST} heatmap between populations colored by geographical/genetic regions.
Populations largely cluster by region with a few exceptions. MID and three AMR populations for example are interspersed among other regions.

Quality control

Our sample QC procedure was mostly the same as in gnomAD, but differed slightly. Specifically, because whole populations were removed from gnomad 'fail_' filters, we did not filter on the basis of these, which were used in gnomAD v3.1. The clearest example of filters that failed was the fail_n_snp_residual filter, as shown in **Figure S19**.

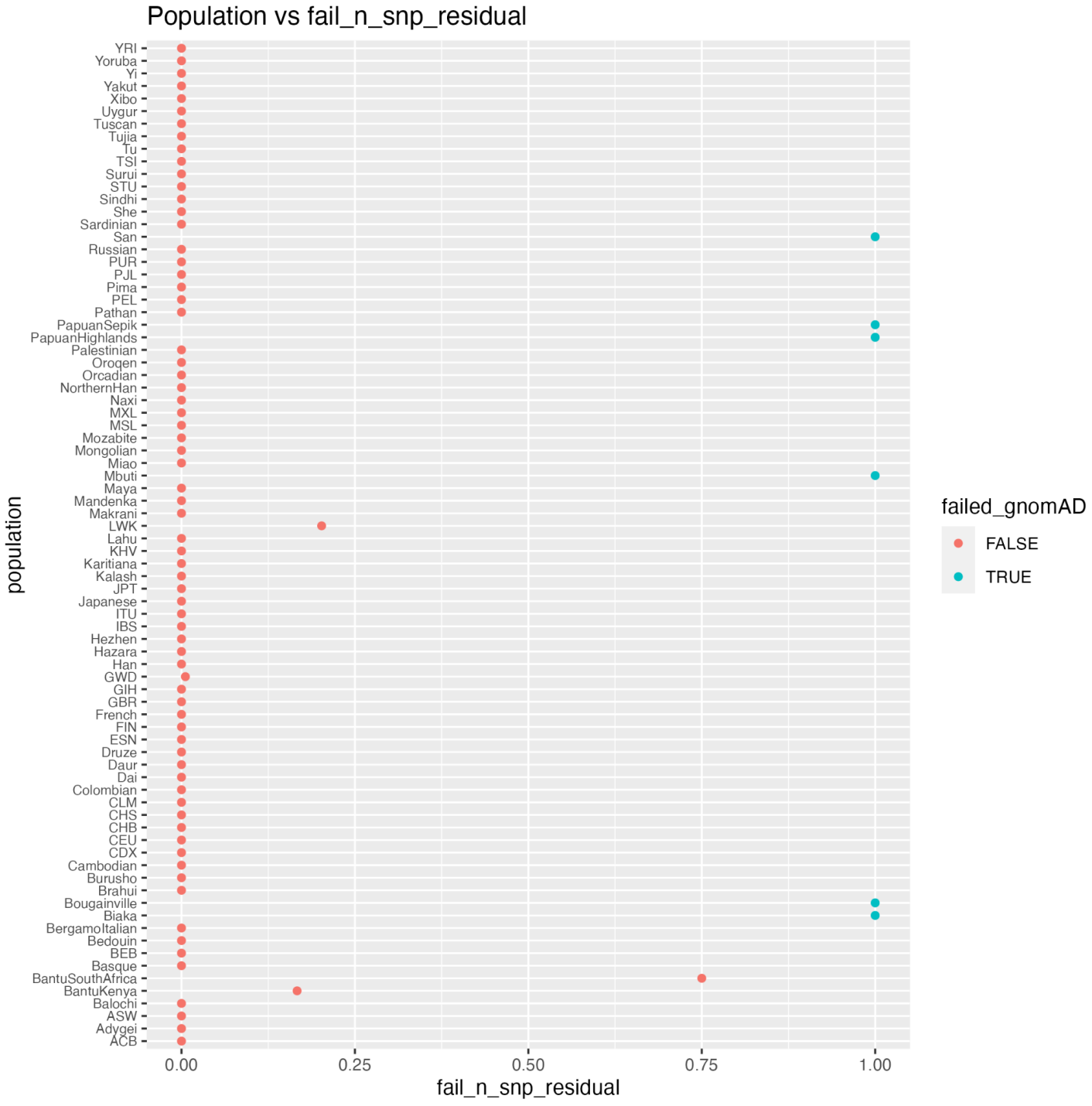

Figure S19 | Example of a filter that was included in gnomAD v3.1 but excluded from this project. The "fail_n_snp_residual" filter, which regresses out principal components from the number of SNPs in an effort to identify technical outliers, would have excluded whole continental groups and populations in this resource because these groups are distinct from the majority of individuals in gnomAD.

Analysis tutorials

To show examples of how to use the individual-level data in a cloud-computing environment, we have created a series of tutorials in iPython notebooks that make use of Hail. These tutorials show how to merge datasets, apply sample and variant QC, run ancestry analysis via PCA and visualization, generate summary statistics of genomes by population, compute and plot population divergence statistics via F_{ST} and f_2 statistics, and intersect external datasets with this dataset and infer ancestry information using project meta-data. The organization of these notebooks is outlined in **Figure 6**.

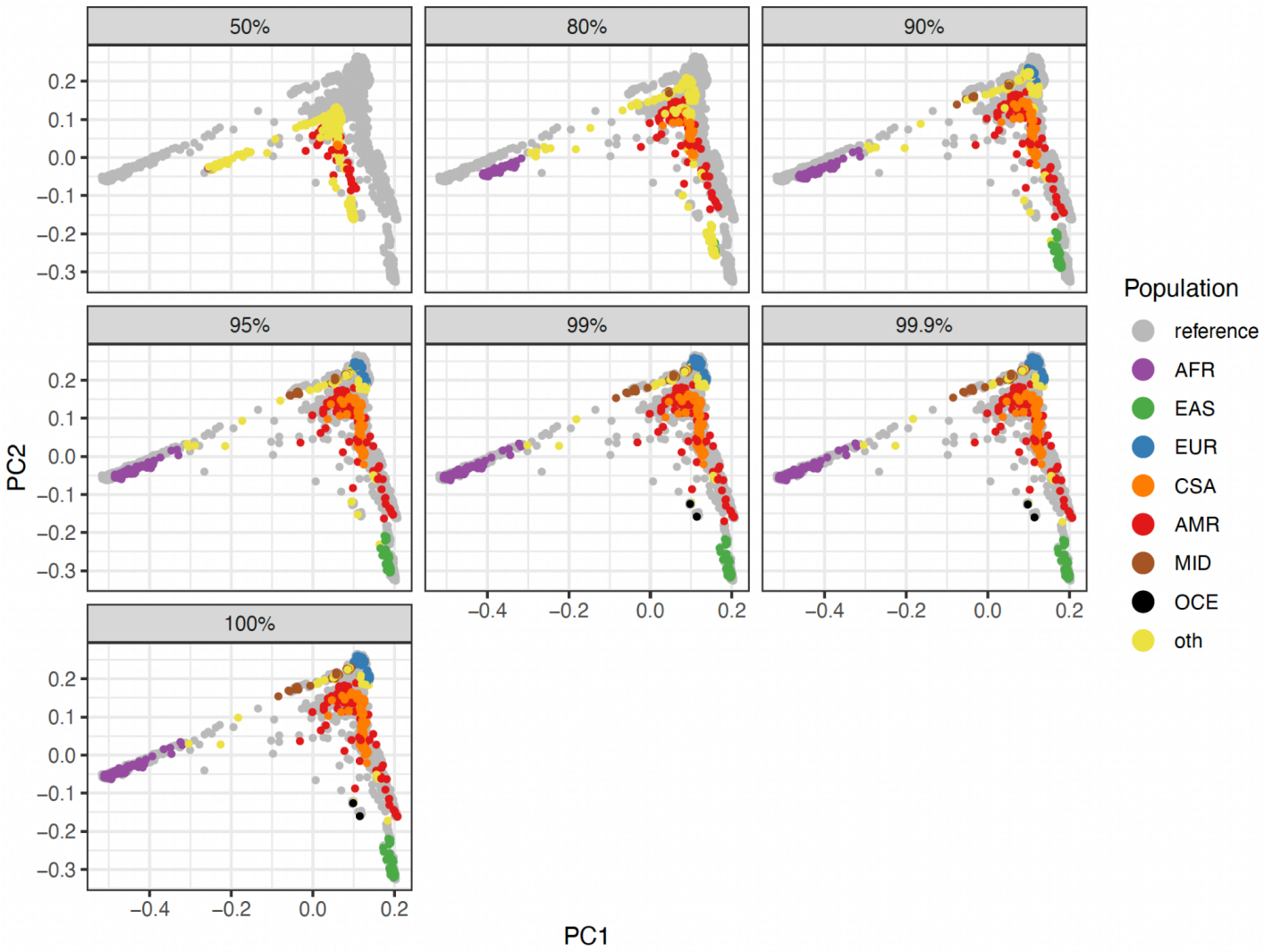

Figure S20 | PCA shrinkage analysis to determine acceptable levels of missingness before ancestry resolution becomes too low to accurately assign population labels. We started with a set of SNPs that were used in other PCA (e.g. **Figure 2**), which had undergone minor allele frequency filtering, missingness filtering, and LD pruning. We randomly selected 80% of samples (N=2,720) to train the random forest with corresponding meta-data labels as usual and held out 20% of samples as a test dataset (N=680). After filtering out monomorphic sites from the training dataset once samples were divided, we

retained 200,403 variants which were used to train the random forest. We randomly downsampled SNPs in the test dataset to include 50%, 80%, 90%, 95%, 99%, 99.9%, and 100% of SNPs in the training dataset. These plots show the corresponding projected PCs in the test dataset, showing the extent to which shrinkage affects analyses. **Table S13** shows rates of unclassified individuals by SNP missingness in the test dataset.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, et al. 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**. https://science.sciencemag.org/content/367/6484/eaay5012/tab-pdf.

Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, Alföldi J, Watts NA, Vittal C, Gauthier LD, et al. 2024. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**: 92–100.

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581**: 444–451.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Hail Team. 2021. *Hail*. https://zenodo.org/record/4504325.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al. 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178. https://www.biorxiv.org/content/10.1101/201178v3 (Accessed July 6, 2020).

Team RC. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing. *(No Title)*. https://cir.nii.ac.jp/crid/1370857669939307264.