Supplementary Information

Supplementary Methods

Quality control, relatedness and genetic ancestry inference in the full TOPMed sample ……..… 4

Genetic ancestry inference in the TOPMed subset of self-identified Black/African American ….. 7

Heritability estimation methods in GCTA ……... 8

Transformation of observed heritability on the liability scale ……..… 8

Comparison of genomic relatedness estimation methods ……..... 9

Mean and variance of log enrichment ratio ……..…. 10

Supplementary Table and Figures

Supplementary Table 1. List of TOPMed studies in Freeze 9 with coronary artery disease (CAD) status available ……... 14

Supplementary Figure 1. The first ten principal components (PC) of 2,403 unrelated individuals from the 1000 Genomes Project …..……... 15

Supplementary Figure 2. Projection of 48,729 TOPMed samples onto the first two principal components (PC1, PC2) of the 2,403 unrelated individuals from the 1000 Genomes Project … 16

Supplementary Figure 3. Ancestral fractions from ADMIXTURE with K = 5 populations ...……. 17

Supplementary Figure 4. Fraction of European ancestry as computed by ADMIXTURE against the pairwise orthogonalized Gnanadesikan-Kettenrin robust Mahalanobis distances from the European continental cluster in the principal component analysis ..……... 18

Supplementary Figure 5. Scree plot of the singular values of the first 20 principal components (PCs) in 22,443 TOPMed samples of European genetic ancestry ...……..... 19

Supplementary Figure 6. Variant loadings for each of the first 20 PCs in 22,443 TOPMed samples of European genetic ancestry ...……..… 20

Supplementary Figure 7. PC scores of 22,443 TOPMed samples of European genetic ancestry ……...…….. 21

Supplementary Figure 8. Fraction of African ancestry as computed by ADMIXTURE against the pairwise orthogonalized Gnanadesikan-Kettenrin robust Mahalanobis distances from the African continental cluster in the principal components analysis 22

Supplementary Figure 9. Scree plot of the singular values of the first 20 principal components in 9,516 TOPMed samples of African genetic ancestry 23

Supplementary Figure 10. Variant loadings for each of the first 20 PCs in 9,516 TOPMed samples of African genetic ancestry 24

Supplementary Figure 11. PC scores of 9,516 TOPMed samples of African genetic ancestry .. 25

Supplementary Figure 12. Contribution of each LD score-MAF bin to the observed heritability h^2 of CAD in the European genetic ancestry sample (GRMs estimated by ratio of averages method and contributions with REML AI algorithm) 26

Supplementary Figure 13. Proportion of observed heritability in each LD score-MAF bin against the proportion of SNVs in that bin 27

Supplementary Figure 14. Contribution of each LD score-MAF bins (16 bins) to the observed heritability h^2 of CAD in the European genetic ancestry sample (GRMs estimated by average of ratios method and contributions with REML EM algorithm) 28

Supplementary Figure 15. Genomic relatedness matrices (GRMs) estimated by the average of ratios method 29

Supplementary Figure 16. Genomic relatedness matrices (GRMs) estimated by the ratio of averages method 30

Supplementary Figure 17. Comparison of genomic relatedness estimation methods in GCTA (ratio of averages versus average of ratios) for variants with high LD scores (above the median) and minor allele frequency (MAF) $\leq$ 0.1% 31

Supplementary Figure 18. Contribution of each LD score-MAF bin to the observed heritability h^2 of CAD in the European genetic ancestry sample (GRMs estimated by average of ratios method and contributions with REML AI algorithm) 32

Supplementary Figure 19. Contribution of each LD score-MAF bin to the observed heritability h^2 of CAD in the European genetic ancestry sample using Tcheandjieu's LD score-MAF binning and adding ultra-rare SNVs (28 GRMs estimated by ratio of averages method and contributions with REML EM algorithm) 33

Supplementary Figure 20. Contribution of each LD score-MAF bin to the observed heritability of CAD in African ancestry (GRMs estimated by ratio of averages method and contributions with REML EM algorithm) 34

Supplementary Figure 21. Sankey diagram showing the distribution of shared SNVs in the European and African genetic ancestry samples across the MAF bins 35

Supplementary Figure 22. Proportion of observed heritability in each LD score-MAF-Impact bin against the proportion of SNVs in that bin (functional impact predicted by SnpEff) 36

Supplementary Figure 23. Distribution of the number of SNVs overlapping with the snATAC-seq peaks 37

Supplementary Figure 24. Proportion of observed heritability in each LD score-MAF-Peak bin against the proportion of SNVs in that bin for all 13 snATAC-seq cell types 38

Supplementary Figure 25. Absolute and relative contribution per variant of each LD score-MAF-Peak bin to the global CAD heritability estimate for the 13 snATAC-seq cell types 39

Supplementary Figure 26. Contribution of each LD score-MAF-aPC-Conservation bin to the global CAD heritability estimate, along with log enriched conservation ratio of each LD score-MAF bin 43

Supplementary Figure 27. Contribution of each LD score-MAF-aPC-Protein-Function bin to the global CAD heritability estimate, along with log enriched protein-function ratio of each LD score-MAF bin 44

Supplementary Figure 28. Proportion of each LD score-MAF-Functionality bin to the global CAD heritability estimate for aPCs at Phred score = 10 or 20 45

Supplementary Figure 29. CAD heritability log enrichment ratio of High over Low functionality SNVs in each LD score-MAF bin for aPCs at Phred score = 10 or 20 46

Brief description of TOPMed studies included 48

TOPMed study-specific acknowledgements 53

NHLBI TOPMed Consortium banner authors 57

TOPMed Atherosclerosis Working Group members 59

Supplementary References 61

Quality control, relatedness and genetic ancestry inference in the full TOPMed sample

First, variants were removed if they were: 1) indels or multiallelic; 2) share the same base pair position; 3) FILTER != PASS in the vcf files; 4) minor allele count (MAC) < 5. We then excluded 3,630 samples with ambiguous coronary artery disease (CAD) status, 23 samples with missing call rate > 5% and removed an additional 158,502 variants with missing call rate > 1%. This left us with 60,744 samples and 84,586,686 biallelic autosomal single nucleotide variants (SNVs).

Second, we identified and excluded related pair of samples. The TOPMed Data Coordinating Center (DCC) carried out a relatedness analysis to identify all pairs of samples related at degree four and above (kinship coefficient ≥ $2^{-11/2}$) using PC-AiR and PC-Relate [1,2]. The analysis was performed using ~630,000 passing SNVs from the minDP10 genotype dataset with minor allele frequency (MAF) > 1%, missing call rate < 1%, and pruned to have low linkage disequilibrium (LD r^2 < 0.1) with each other. We found that 1,073 out of 1,101 samples from the Amish study are related at degree four and above. Consequently, we decided to exclude the whole study from subsequent analyses. Then, using an in-house greedy algorithm based on R package igraph, we further removed one member of each pair of related samples. Our algorithm took into account age-at-onset to prioritize selection of younger cases and older controls. This step removed 12,015 samples related at degree four and above with another sample. After removal of SNVs with missing call rate > 1% and/or MAC < 5, our dataset was comprised of 48,729 samples and 72,247,814 biallelic autosomal SNVs.

Third, we classified samples according to their genetic ancestry as determined by a principal component analysis (PCA) and by ADMIXTURE 1.3 [3]. We first performed a PCA by using the pipeline proposed by Privé et al [4] and implemented in the R package bigsnpr. We iteratively identified a set of 2,403 unrelated individuals from the 1000 Genomes Project [5] using the function snp_plinkKINGQC (pairs of individuals with kinship coefficient ≥ $2^{-9/2}$ were excluded), and by removing outliers identified using a statistic based on K nearest neighbors and by visual inspection of its distribution. Principal component (PC) scores and loadings of these 2,403 individuals were estimated using a truncated singular value decomposition (SVD) of the scaled genotype matrix implemented in function bed_autoSVD. This function performs LD clumping with (default) r^2 > 0.2 and (default) MAC ≥ 10, and automatically detects and removes regions with long-range LD by computing robust Mahalanobis distances of the PC loadings. We

repeated this PCA procedure by varying the number of PCs and determined, after removal of variant and sample outliers, that ten PCs were sufficient to classify the 1000 Genomes individuals into their respective superpopulation/continental ancestry groups (African, American, East Asian, South Asian and European) (Supplementary Fig. 1). The PC scores and loadings from this final model were then used to project our 48,729 TOPMed samples into the PC space (Supplementary Fig. 2) using an optimized implementation of the Online Augmentation, Decomposition and Procustes (OADP) transformation. Participants who self-identified as Black/African American (BLK) or Hispanic/Latino (HSP) do not cluster tightly compared to East Asian (EAS) or White/European (WHT) (when reporting results involving the use of race, ethnicity and/or genetic ancestry, we followed recommendations made by the TOPMed program [6]). To detect samples that differ too much from any superpopulation/continental ancestry groups, we computed the pairwise orthogonalized Gnanadesikan-Kettenrin robust Mahalanobis distances on PC scores as implemented in function dist_ogk in the R package bigutilsr.

To infer genetic ancestry proportions in our 48,729 TOPMed samples, we utilized as a reference set the 2,403 unrelated 1000 Genomes samples in ADMIXTURE. We used the same set of pruned SNVs as for the PCA above. To identify the number *K* of ancestral populations to include in the admixture model, five-fold cross-validation (--cv option) was performed by varying *K* from 1 to 10 in unsupervised mode. We determined that the optimal number of ancestral populations was *K* = 5. The allele frequencies (*P*) estimated from the training sample set of 2,403 samples for *K* = 5 were then used to project the 48,729 TOPMed samples (Supplementary Fig. 3) and estimate their ancestral fractions (*Q*). ADMIXTURE confirmed that most TOPMed participants who self-identify as EAS or WHT are not genetically admixed at a continental level (Supplementary Fig. 3). On the contrary, participants who self-identified as BLK, HSP or SAS (South Asian) show high levels of admixture.

We plotted the fraction of European ancestry inferred by ADMIXTURE against the pairwise orthogonalized Gnanadesikan-Kettenrin robust Mahalanobis distances on PC scores of five European populations from the 1000 Genomes project. In the 1000 Genomes 2,403 unrelated participants, we clearly observed a strong non-linear relationship between the fraction of European ancestry and the Mahalanobis distance (Supplementary Fig. 4, left panel). The Finnish subpopulation (FIN) clusters apart from the other four European subpopulations with up to 10% of non-European genetic ancestry, which is supported by other lines of genetic evidence pointing to a non-negligible Asian contribution to the gene pool of Finns [7]. Upon inspection, a

nearly homogeneous TOPMed sample of European genetic ancestry (Supplementary Fig. 4, right panel) was identified by selecting participants with an ADMIXTURE European ancestry fraction > 98% and a Mahalanobis distance < 100. This step reduces our sample to 23,046 participants. In this smaller sample, we excluded SNVs with missing call rate > 1%, MAC < 5, Hardy-Weinberg test p-value < 10^{-6}, case-control missingness test p-value < 0.05: only 28,545,707 biallelic autosomal SNVs remained after this quality control (QC) step.

Fourth, we computed the inbreeding coefficient of the remaining 23,046 TOPMed participants by running the --het option in plink 2.0. The set of pruned SNVs in approximate LD were obtained by running --indep-pairwise 50kb 1 0.1. An in-house R script then iteratively removes outliers from the boxplot statistics computed by R: as a result, 603 additional samples were removed. Applying the same QCs as before on the SNVs, our final TOPMed sample consists of 22,443 samples (4,949 cases and 17,494 controls) with genotype data for 28,051,806 biallelic autosomal SNVs.

To identify possible residual stratification in this final sample, an unsupervised PCA was performed using once again the R package bigsnpr. We set to 20 the number of PCs to be computed and chose the default options of function bed_autoSVD. After a few iterations of removing long-range LD regions, a total of 5,047,983 (pruned) SNVs were kept in the PCA. Pairwise robust Mahalanobis distances on PC scores were again calculated using function dist_ogk in R package bigutilsr. Supplementary Fig. 5 displays the singular values of the first 20 principal components (PCs). The SNV loadings for each of these 20 PCs are shown in Supplementary Fig. 6. Some long-range linkage disequilibrium (LD) regions (peaks) were visible for PC16 to PC20. On the contrary, PC1 to PC15 seemed to capture residual stratification not influenced by any of these long-range LD regions. Based on Supplementary Figs 5 and 6, we included PC1 to PC15 as fixed effects in our heritability analyses, and excluded PC16 to PC20 as they might capture LD structure rather than population structure (Privé et al [4], and references cited therein). In Supplementary Fig. 7, PC scores of all 22,443 TOPMed participants were plotted and colored by their Mahalanobis distance, and confirmed that the final sample of 22,443 participants did not include any genetic outlier relative to the European genetic ancestry.

Genetic ancestry inference in the TOPMed subset of self-identified Black/African American

We repeated the same quality control steps as above in the TOPMed subset of self-identified Black/African American (BLK) participants. The fraction of African ancestry inferred by ADMIXTURE against the Mahalanobis distances on PC scores of seven African populations from the 1000 Genomes project is plotted in Supplementary Fig. 8. We observed the same non-linear relationship between the fraction of African ancestry and the Mahalanobis distance (Supplementary Fig. 8, left panel). To reach a sufficient sample size for further heritability analysis, we applied a more relaxed threshold with respect to the inferred African ancestry in the TOPMed sample: a fairly homogeneous TOPMed sample of African genetic ancestry (Supplementary Fig. 8, right panel) was restricted to participants with an ADMIXTURE African ancestry fraction > 75% and a Mahalanobis distance < 1000. These selection criteria excluded participants with inferred East African ancestry, represented by the Luhya in Webuye, Kenya (LWK) individuals in the 1000 Genomes. Many studies showed that African American ancestors mostly originated from the western part of Africa [8,9], hence resulting in very few exclusions of TOPMed participants. Unfortunately, due to high admixture, only 9,816 participants remained at this point. In this smaller sample, after applying the same QC steps to SNVs, and after exclusion of 299 additional individuals with high inbreeding coefficients, the sample consisted of 9,517 TOPMed participants of inferred African genetic ancestry.

As we did previously, an unsupervised PCA was performed to detect residual stratification. Again, we set the number of PCs to 20 and ran the default options of function bed_autoSVD. After a few iterations of removing long-range LD regions, a total of 4,748,912 (pruned) SNVs were kept in the PCA. Mahalanobis distances on PC scores were again calculated, and one participant was considered as an outlier and hence excluded. Supplementary Fig. 9 displays the singular values of the first 20 PCs: only one PC seemed to capture residual stratification (elbow method), and this was confirmed by the plot of SNV loadings for each of these 20 PCs (Supplementary Fig. 10). Long-range LD peaks were clearly visible for all PCs except PC1. This might be surprising that, compared to the inferred European sample, residual stratification is minimal in the inferred African sample. However, the African sample size is ~2.4 times lower than the European one, which could lead to potentially less observed genetic diversity. Also, European American cohorts have revealed subcontinental stratification, along with underappreciated admixture at the subcontinental level [10]. Consequently, in all subsequent

heritability analyses, we only included PC1 as a fixed effect (including the first 15 PCs did not modify the estimations; results not shown). The PC scores plot of all 9,516 TOPMed participants did not display any outlier relative to the inferred African genetic ancestry cluster (Supplementary Fig. 11). Our final dataset is comprised of 9,516 (1,733 cases and 7,783 controls) participants and 35,738,556 biallelic autosomal SNVs, ~27% more variants than in the European selected sample.

Heritability estimation methods in GCTA

In GCTA, heritability (variance components) is estimated by including all genomic relatedness matrices (GRMs) into a linear mixed effects model. Different REML (restricted maximum likelihood) estimation algorithms are available in GCTA: 0 for Average Information (AI, default method), 1 for Fisher-Scoring (FS) and 2 for Expectation-Maximization (EM). All heritability estimates were produced under the (default) restricted mode, that is, all variance components were not allowed to be negative. Because REML algorithms 0 or 1 stopped under the restricted mode for some analyses, we decided to report results generated by the REML algorithm 2 only. Each time, we set the maximum number of iterations at 10,000 (GCTA maximum allowed). If algorithm 2 (EM) did not output an heritability estimate after 10,000 iterations, the last estimated variance components were input as prior variance estimates using the --reml-priors option.

Transformation of observed heritability on the liability scale

Observed heritability (h^2_{obs}) was transformed to heritability on the liability scale (h^2_{liab}) using the following equation [11]:

$$h^2_{liab} = h^2_{obs}\frac{K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)}$$

where *K* is the population prevalence, *P* is the proportion of cases in the sample, and *z* is the density value at threshold *t* of the standard normal distribution *Z* for which $\mathrm{Prob}(Z > t) = K$. The variance of h^2_{liab} was computed using

$$\mathrm{Var}(h^2_{liab}) = \mathrm{Var}(h^2_{obs})\left[\frac{K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)}\right]^2$$

Comparison of genomic relatedness estimation methods

Two different genomic relatedness matrix (GRM) estimation methods are available in GCTA: i) average of ratios (AoR, default method) and ii) ratio of averages (RoA). Let x_{ij} denote the number of minor alleles (0, 1 or 2) for SNV *i* in individual *j*, *N* the number of SNVs and p_i the sample minor allele frequency (MAF). In the average of ratios (AoR) method:

$$\mathrm{GRM}_{\mathrm{AoR}} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}$$

In the ratio of averages (RoA) method:

$$\mathrm{GRM}_{\mathrm{RoA}} = \frac{\sum_{i=1}^{N}(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2\sum_{i=1}^{N}p_i(1-p_i)}$$

As mentioned in Wainschtein et al. [12], the AoR estimation method could be biased, especially with very rare variants. We investigated the effect that non-European genetic ancestry estimate and inbreeding coefficient might have on the diagonal elements of the GRMs in all eight LD score-MAF bins in both estimation methods. These diagonal elements are expected to be close to 1 with low levels of relatedness, admixture and inbreeding. Regarding the AoR method, Supplementary Fig. 15 shows that diagonal elements in ultra-rare and rare variant bins (MAF ≤ 1%) tend to be larger for samples with higher fraction of non-European genetic ancestry, while no clear relationship exists with inbreeding coefficient (except maybe for common SNVs with high LD score, but the values are tightly clustered around 1). This effect seems less marked when GRMs are estimated with the RoA method (Supplementary Fig. 16). In both estimation methods, diagonal elements are the largest in the ultra-rare (MAF ≤ 0.1%) SNV bin with high LD score. We compared the diagonal and off-diagonal elements (expected value close to 0) of this LD score-MAF bin between the two methods and observed that the RoA method produces larger values than the AoR method (Supplementary Fig. 17).

Choice of the GRM estimation method and REML algorithm could impact heritability estimate. For example, using the AoR method and REML AI algorithm, the observed heritability is h^2_{obs} = 28.5% (SE = 11.9%), higher than the heritability estimate reported in the main paper (Fig. 1,

Supplementary Data 1). The largest contribution (Supplementary Data 4, Supplementary Fig. 18) still comes from ultra-rare SNVs (MAF ≤ 0.1%) with low LD score: ~62% of the total observed heritability is attributable to this bin (0.176/0.285). In contrast, common SNVs (10% < MAF ≤ 50%) with high LD score only contributes ~11% of the total observed heritability (0.032/0.285). When using the (less biased) RoA method, all the non-zero variance contributions decreased except for common SNVs (10% < MAF ≤ 50%) with high LD score (Supplementary Fig. 12), resulting in a total observed heritability of h^2_{obs} = 20.0% (SE = 10.3%). The largest contribution (Supplementary Data 2) still comes from ultra-rare SNVs (MAF ≤ 0.1%) with low LD score: ~62% of the total observed heritability is attributable to this bin (0.125/0.200). In comparison, common SNVs (10% < MAF ≤ 50%) with high LD score now contributes ~18% of the total observed heritability (0.035/0.200).

Current GRM estimation methods are prone to bias in presence of population structure, and this bias is exacerbated especially for rare variants in high LD. To accurately estimate kinship between samples, the AoR method must satisfy more restrictive conditions than the RoA method, although both methods are biased when loci are not in linkage equilibrium [13]. At the time of writing, the default GRM estimation method in GCTA is the AoR method, which consistently inflated our results compared to the RoA method. More robust estimators have recently been proposed to compute GRMs, although bias remains when loci are in LD [13,14]. Finally, we recall that REML AI (algorithm 0) or REML FS (algorithm 1) did not always converge under the restricted mode of GCTA, so we opted to run REML EM (algorithm 2) for consistency purposes. However, the REML EM method tended to report more non-null variance estimates in some LD score-MAF bins compared to algorithms 0 or 1, hence inflating the resulting total heritability estimate.

Mean and variance of log enrichment ratio

We derive approximate expressions for the mean and variance of the logarithm of the enrichment ratio in each MAF-LD bin using as an example the constrained versus non-constrained bins. First, we define variables X_{in} and X_{out}, representing the proportion of heritability explained per SNV by constrained, and by non-constrained bins, respectively:

$$X_{in} = \frac{V(G)_{in}/V_p}{N_{in}}, \quad X_{out} = \frac{V(G)_{out}/V_p}{N_{out}},$$

where $V(G)_{in}/V_p$ and $V(G)_{out}/V_p$ are obtained directly from GCTA output. Here, $V(G)$ is an estimate of heritability explained by the corresponding MAF-LD bin, and V_p is an estimate of the total heritability for the phenotype on the observed scale. N_{in} and N_{out} are extracted from the GRM estimate of each MAF-LD bin. Note that the same methodology was applied when we estimated the SnpEff impact ratio of protein-altering over non-protein-altering SNVs, and the SNVs inside snATAC-seq peaks over SNVs outside peaks .

The (natural) logarithm of the enrichment ratio is then expressed as a function of X_{in} and X_{out}:

$$\mathrm{g}(X_{in}, X_{out}) = \ln(X_{in}/X_{out}) = \ln(X_{in}) - \ln(X_{out}) \quad (1)$$

Unless $X_{in} = 0$ or $X_{out} = 0$, the function g could be approximated by using a Taylor series expansion around the mean $(\mu_{in}, \mu_{out}) \equiv (\mathrm{E}(X_{in}), \mathrm{E}(X_{out}))$.

In general, for two random variables X_1 and X_2, one can write

$$g(X_1, X_2) = g(\mu_1, \mu_2) + (X_1 - \mu_1)\, g'_{X_1}(\mu_1, \mu_2) + (X_2 - \mu_2)\, g'_{X_2}(\mu_1, \mu_2) + R \quad (2)$$

where g'_{X_1} and g'_{X_2} are the first derivative of g with respect to X_1 and X_2, respectively, evaluated at (μ_1, μ_2), and R designates higher order terms in the expansion.

A first order approximation of the mean of $g(X_1, X_2)$ is obtained by dropping the higher order terms R, and by taking the expectation on both sides of Equation (2). The second and third terms in the right-hand side of Equation (2) now become 0, since $\mathrm{E}(X_1) = \mu_1$ and $\mathrm{E}(X_2) = \mu_2$, by definition. Hence, we obtain

$$\mathrm{E}\big(g(X_1, X_2)\big) \approx \mathrm{E}\big(g(\mu_1, \mu_2)\big) = g(\mu_1, \mu_2) \quad (3)$$

A first order approximation of the variance of $g(X_1, X_2)$ is obtained along the same lines. Using Equations (2) and (3),

$$\mathrm{Var}\big(g(X_1, X_2)\big) = \mathrm{E}\left[\Big(g(X_1, X_2) - \mathrm{E}\big(g(X_1, X_2)\big)\Big)^2\right]$$

$$\approx \mathrm{E}\left[\left(g(\mu_1,\mu_2) + (X_1-\mu_1)\, g'_{X_1}(\mu_1,\mu_2) + (X_2-\mu_2)\, g'_{X_2}(\mu_1,\mu_2) - \mathrm{E}\big(g(X_1,X_2)\big)\right)^2\right]$$

$$= \mathrm{E}\left[\left((X_1-\mu_1)\, g'_{X_1}(\mu_1,\mu_2) + (X_2-\mu_2)\, g'_{X_2}(\mu_1,\mu_2)\right)^2\right]$$

$$= \mathrm{E}(X_1-\mu_1)^2\left[g'_{X_1}(\mu_1,\mu_2)\right]^2 + \mathrm{E}(X_2-\mu_2)^2\left[g'_{X_2}(\mu_1,\mu_2)\right]^2$$
$$+2\,\mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)]\, g'_{X_1}(\mu_1,\mu_2)\, g'_{X_2}(\mu_1,\mu_2)$$

$$= \mathrm{Var}(X_1)\left[g'_{X_1}(\mu_1,\mu_2)\right]^2 + \mathrm{Var}(X_2)\left[g'_{X_2}(\mu_1,\mu_2)\right]^2$$
$$+2\,\mathrm{Cov}(X_1,X_2)\, g'_{X_1}(\mu_1,\mu_2)\, g'_{X_2}(\mu_1,\mu_2) \qquad (4)$$

Now, using g as defined in Equation (1), we have

$$\mathrm{E}\big(g(X_{in},X_{out})\big) \approx g(\mu_{in},\mu_{out}) = \ln(\mu_{in}/\mu_{out}) = \ln(\mu_{in}) - \ln(\mu_{out}) \qquad (5)$$

Straightforward calculus gives $g'_{X_1}(\mu_1,\mu_2) = \frac{1}{\mu_1}$ and $g'_{X_2}(\mu_1,\mu_2) = -\frac{1}{\mu_2}$. Substituting these values into Equation (4), we obtain

$$\mathrm{Var}\big(g(X_{\mathrm{in}},X_{\mathrm{out}})\big) \approx \frac{\mathrm{Var}(X_{\mathrm{in}})}{\mu_{in}^2} + \frac{\mathrm{Var}(X_{\mathrm{out}})}{\mu_{out}^2} - 2\frac{\mathrm{Cov}(X_{\mathrm{in}},X_{\mathrm{out}})}{\mu_{\mathrm{in}}\,\mu_{\mathrm{out}}} \qquad (6)$$

To compute the mean and variance from real data, we must provide an estimate for all terms in Equations (5) and (6). Assuming that all SNVs contribute uniformly and independently in each MAF-LD bin, $\mathrm{E}(X_{in}) = \mu_{in}$ and $\mathrm{E}(X_{out}) = \mu_{out}$ could be estimated using

$$\hat{X}_{in} = \frac{\hat{V}(G)_{in}/\hat{V}_P}{N_{in}} \text{ and } \hat{X}_{out} = \frac{\hat{V}(G)_{out}/\hat{V}_P}{N_{out}}$$

directly from the GCTA output, and by invoking the law of large numbers applied to a proportion (N_{in} and N_{out} are large enough in all MAF-LD bins). Estimated variances of $\hat{V}(G)_{in}$ and $\hat{V}(G)_{out}$, along with $\mathrm{Cov}(\hat{V}(G)_{in},\hat{V}(G)_{out})$, are also outputted by GCTA, and can be substituted into $\mathrm{Var}(\hat{X}_{in})$, $\mathrm{Var}(\hat{X}_{out})$ and $\mathrm{Cov}(\hat{X}_{in},\hat{X}_{out})$ to obtain the variance estimate of Equation (6).

When estimating the average log enrichment ratio in each MAF-LD bin across all 13 snATAC-seq cell types, we assumed that all cell types were statistically independent and simply applied the following formulas to estimate the mean and variance:

$$\mathrm{E}\left(\sum_{\text{all}} \frac{\ln(X_{in}/X_{out})}{13}\right) = \frac{1}{13}\sum_{\text{all}} \mathrm{E}(\ln(X_{in}/X_{out})) \approx \frac{1}{13}\sum_{\text{all}} \ln(\hat{X}_{in}/\hat{X}_{out})$$

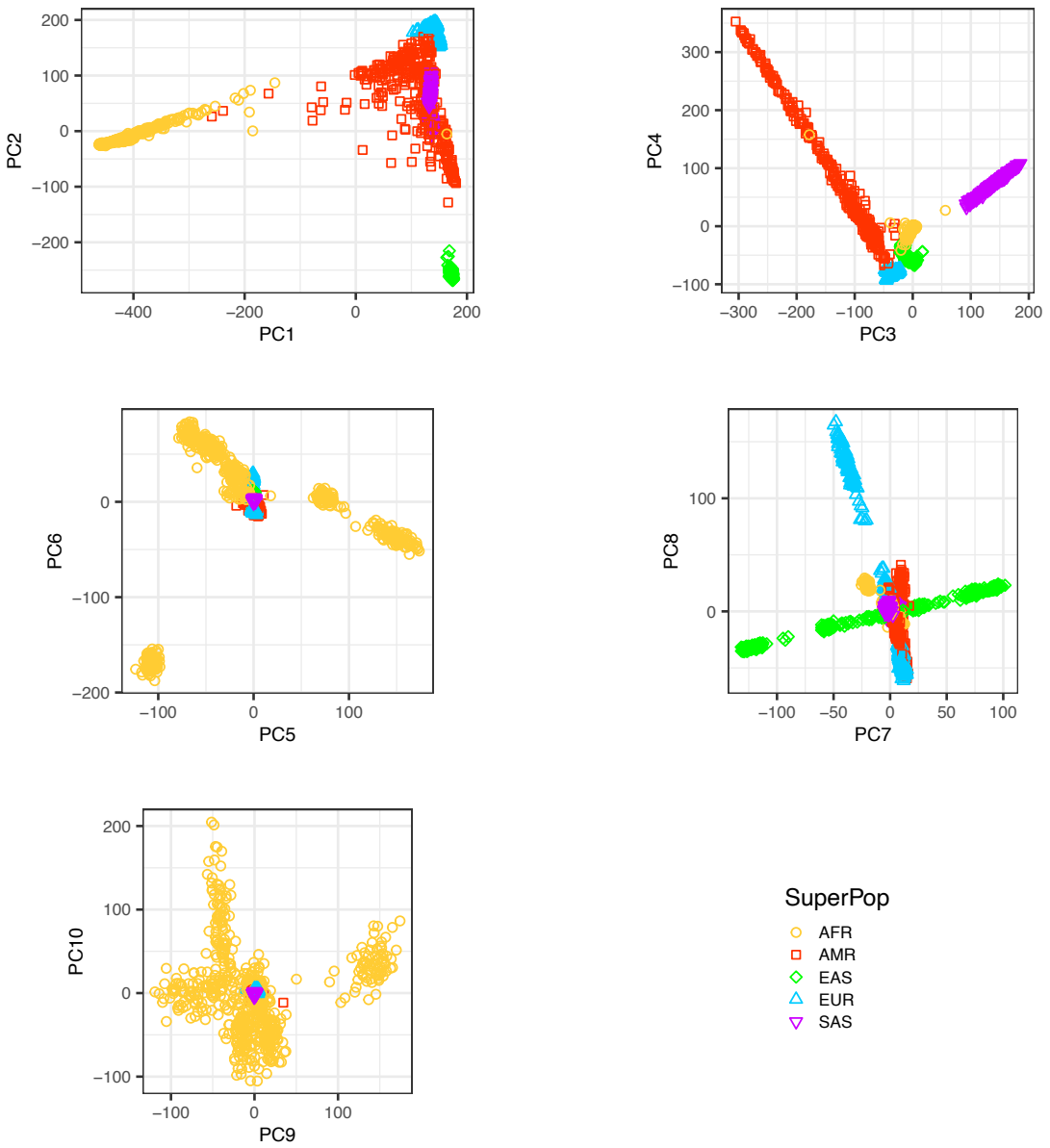
$$\mathrm{Var}\left(\sum_{\text{all}} \frac{\ln(X_{in}/X_{out})}{13}\right) = \frac{1}{13^2}\sum_{\text{all}} \mathrm{Var}(\ln(X_{in}/X_{out}))$$

$$\approx \frac{1}{13^2}\sum_{\text{all}} \left[\frac{\mathrm{Var}(\hat{X}_{in})}{\hat{X}_{in}^2} + \frac{\mathrm{Var}(\hat{X}_{out})}{\hat{X}_{out}^2} - 2\frac{\mathrm{Cov}(\hat{X}_{in}, \hat{X}_{out})}{\hat{X}_{in}\,\hat{X}_{out}}\right]$$

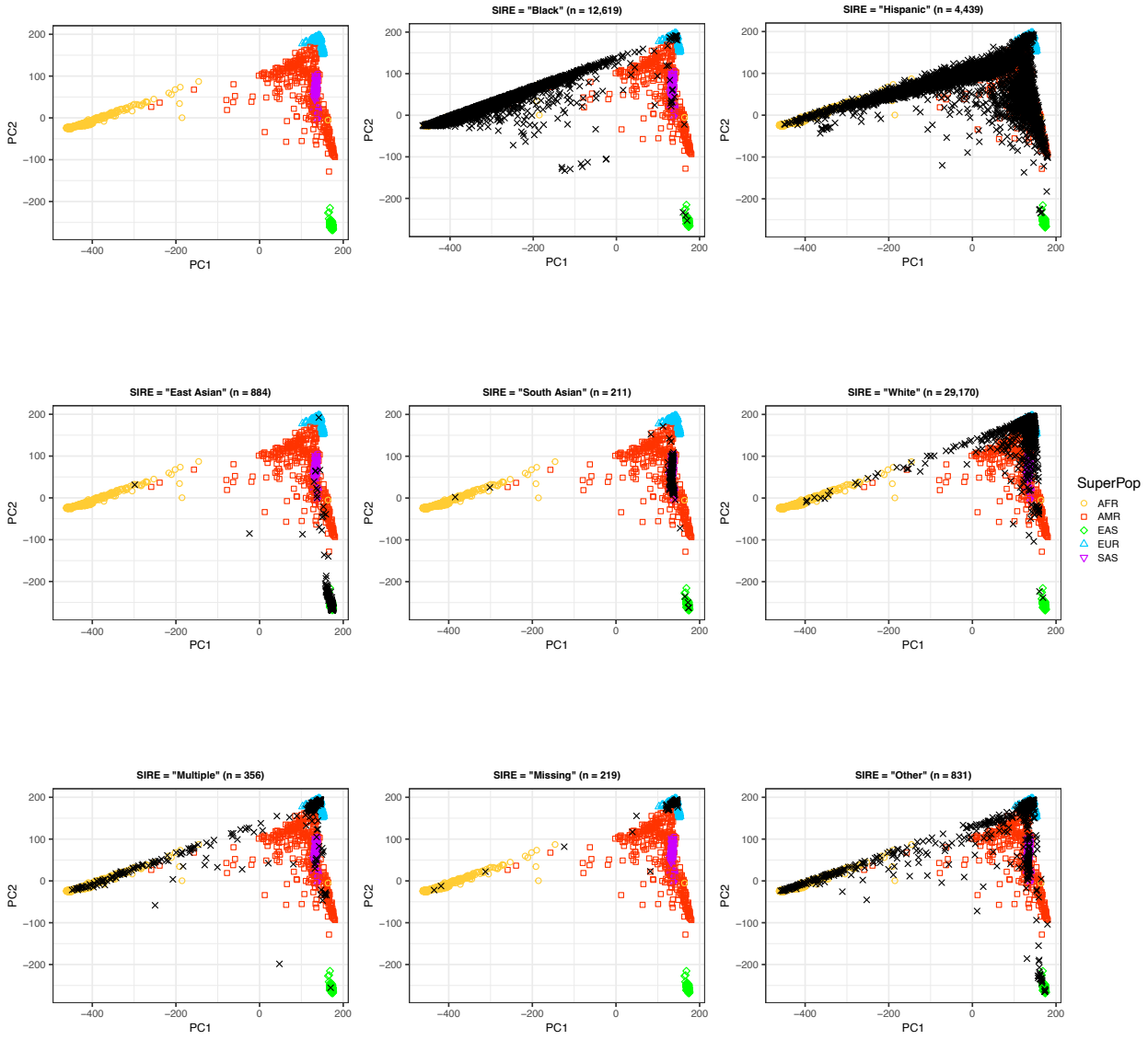
Supplementary Table 1. List of TOPMed studies in Freeze 9 with coronary artery disease (CAD) status available.

TOPMed project	dbGaP TOPMed study accession	Study abbreviation	Sample size	dbGaP parent study accession
Amish	phs000956	Amish	1,101	
AFGen, VTE	phs001211	ARIC	8,125	phs000280
BioMe	phs001644	BioMe	12,010	phs000925
CARDIA	phs001612	CARDIA	3,073	phs000285
CHS, VTE	phs001368	CHS	3,542	phs000287
COPD	phs000951	COPDGene	10,177	phs000179
AA_CAC	phs001412	DHS	392	phs001012
AFGen, FHS	phs000974	FHS	4,176	phs000007
AA_CAC, GeneSTAR	phs001218	GeneSTAR	1,585	phs001074
AA_CAC, HyperGEN_GENOA	phs001345	GENOA	1,253	phs001238
JHS	phs000964	JHS	3,344	phs000286
AA_CAC, MESA	phs001416	MESA	4,595	phs000209
WHI	phs001237	WHI	11,024	phs000200
		Total =	64,397	

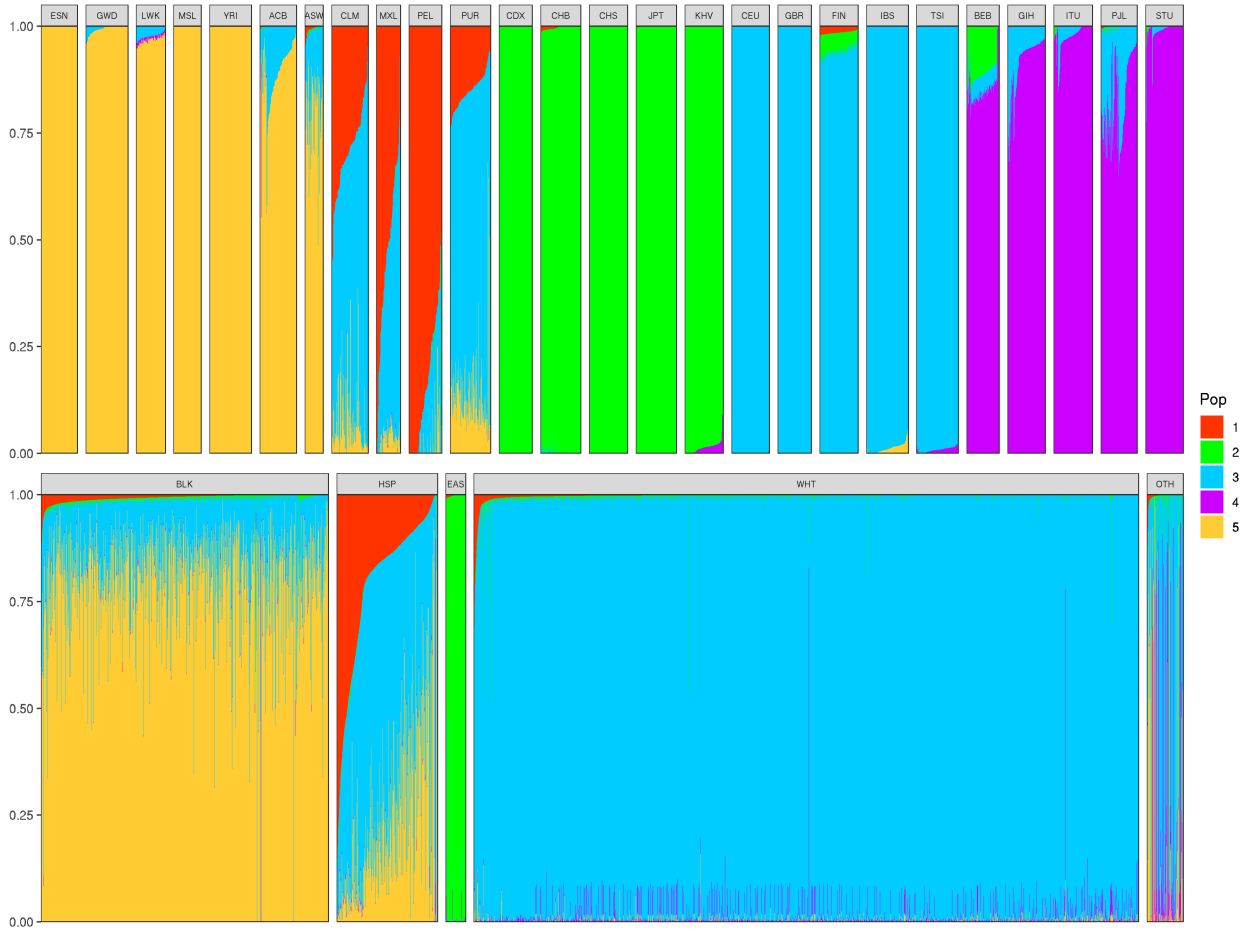
AA_CAC: African American Coronary Artery Calcification project; AFGen: Identification of Common Genetic Variants for Atrial Fibrillation and PR Interval - Atrial Fibrillation Genetics Consortium; Amish: Genetics of Cardiometabolic Health in the Amish; ARIC: Atherosclerosis Risk in Communities; BioMe: Mount Sinai BioMe Biobank; CARDIA: Coronary Artery Risk Development in Young Adults; CHS: Cardiovascular Health Study; COPD: Genetic Epidemiology of COPD; DHS: Diabetes Heart Study; FHS: Framingham Heart Study; GeneSTAR: Genetic Studies of Atherosclerosis Risk; GENOA: Genetic Epidemiology Network of Arteriopathy; HyperGEN_GENOA: Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arteriopathy; JHS: Jackson Heart Study; MESA: Multi-Ethnic Study of Atherosclerosis; VTE: Venous Thromboembolism project; WHI: Women's Health Initiative.
More details at https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-methods-freeze-9



Supplementary Figure 1. The first ten principal components (PC1 to PC10) of 2,403 unrelated individuals from the 1000 Genomes Project. Each sample is colored according to its classification into one of the five superpopulation/continental ancestry groups (SuperPop) as reported in the 1000 Genomes Project (AFR = African, AMR = American, EAS = East Asian, EUR = European, SAS = South Asian).



Supplementary Figure 2. Projection of 48,729 TOPMed samples onto the first two principal components (PC1, PC2) of the 2,403 unrelated individuals from the 1000 Genomes Project. The top left panel represents the 1000 Genomes 2,403 individuals, while other panels represent each TOPMed sample projected values of PC2 versus PC1 (black cross) by self-identified race/ethnicity (SIRE) categories. SIRE categories are Black/African American, Hispanic/Latino, East Asian, South Asian, White/European, Multiple, Missing/Unknown, and Other.



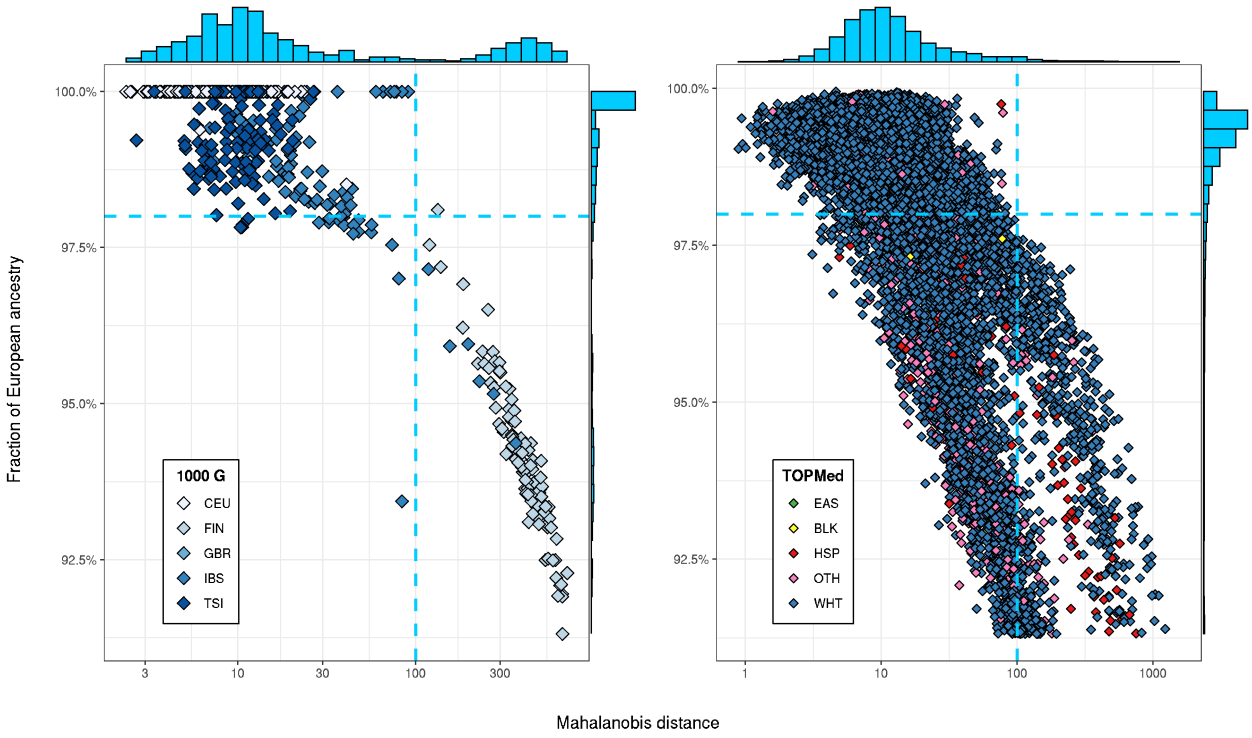
Supplementary Figure 3. Ancestral fractions from the ADMIXTURE software program with *K* = 5 populations. Each thin bar represents a participant colored with their corresponding ancestral fractions from each of the five populations (Pop).

Top panel: 2,403 unrelated samples from 26 populations in the 1000 Genomes Project.

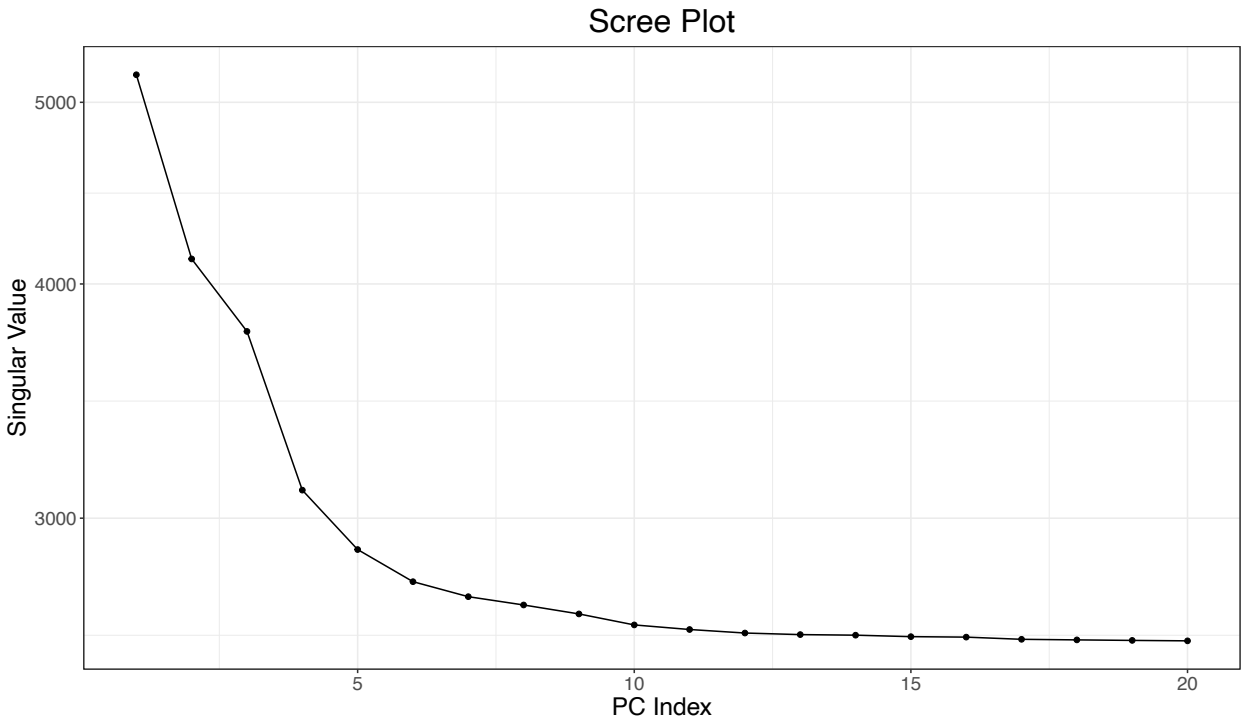
ESN = Esan in Nigeria; GWD = Gambian in Western Division, Mandinka; LWK = Luhya in Webuye, Kenya; MSL = Mende in Sierra Leone; YRI = Yoruba in Ibadan, Nigeria; ACB = African Caribbean in Barbados; ASW = African-American in Southwest USA; CLM = Colombians in Medellin, Colombia; MXL = Mexican-American in Los Angeles, USA; PEL = Peruvians in Lima, Peru; PUR = Puerto Ricans in Puerto Rico; CDX = Chinese Dai in Xishuangbanna, China; CHB = Han Chinese in Beijing, China; CHS = Southern Han Chinese; JPT = Japanese in Tokyo, Japan; KHV = Kinh in Ho Chi Minh City, Vietnam; CEU = Utah residents with Northern and Western European ancestry; GBR = British in England and Scotland; FIN = Finnish in Finland; IBS = Iberian populations in Spain; TSI = Toscani in Italia; BEB = Bengali in Bangladesh; GIH = Gujarati Indians in Houston, USA; ITU = Indian Telugu in United Kingdom; PJL = Punjabi in Lahore, Pakistan; Tamil = Sri Lankan Tamil in United Kingdom.

Bottom panel: 48,729 unrelated TOPMed samples displayed with respect to their self-identified race/ethnicity (SIRE).

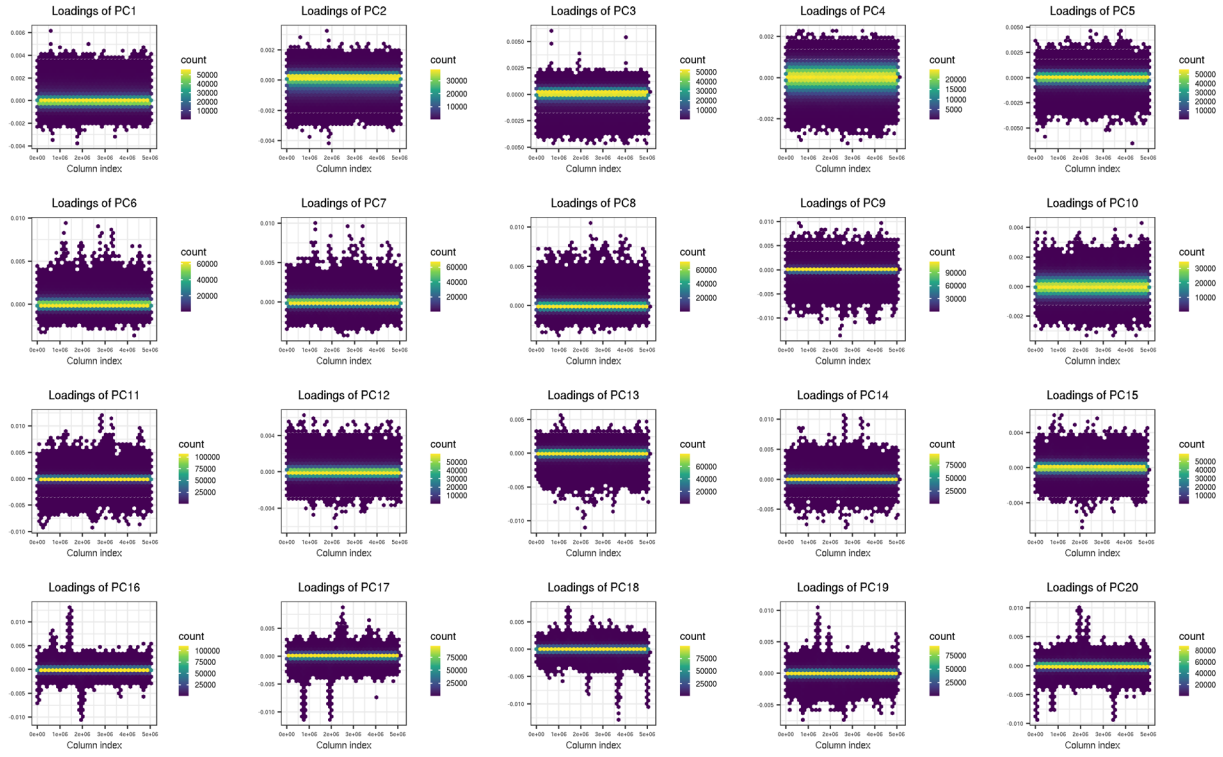
BLK = Black/African American; HSP = Hispanic/Latino; EAS = East Asian; WHT = White/European; OTH = South Asian, Multiple, Missing/Unknown, and Other.



Supplementary Figure 4. Fraction of European ancestry as computed by ADMIXTURE against the pairwise orthogonalized Gnanadesikan-Kettenrin robust Mahalanobis distances from the European continental cluster in the principal components analysis. In both panels, blue dotted lines indicate a fraction of European ancestry = 98% (horizontal), and Mahalanobis distance = 100 (vertical), respectively. X-axes are displayed in log scale.
Left panel: 2,403 unrelated samples from five European populations in the 1000 Genomes Project (1000 G).
CEU = Utah residents with Northern and Western European ancestry; FIN = Finnish in Finland; GBR = British in England and Scotland; IBS = Iberian populations in Spain; TSI = Toscani in Italia.
Right panel: 48,729 unrelated samples from the TOPMed project colored according to their self-identified race/ethnicity (SIRE).
BLK = Black/African American; EAS = East Asian; HSP = Hispanic/Latino; OTH = South Asian, Multiple, Missing/Unknown, and Other; WHT = White/European.
.

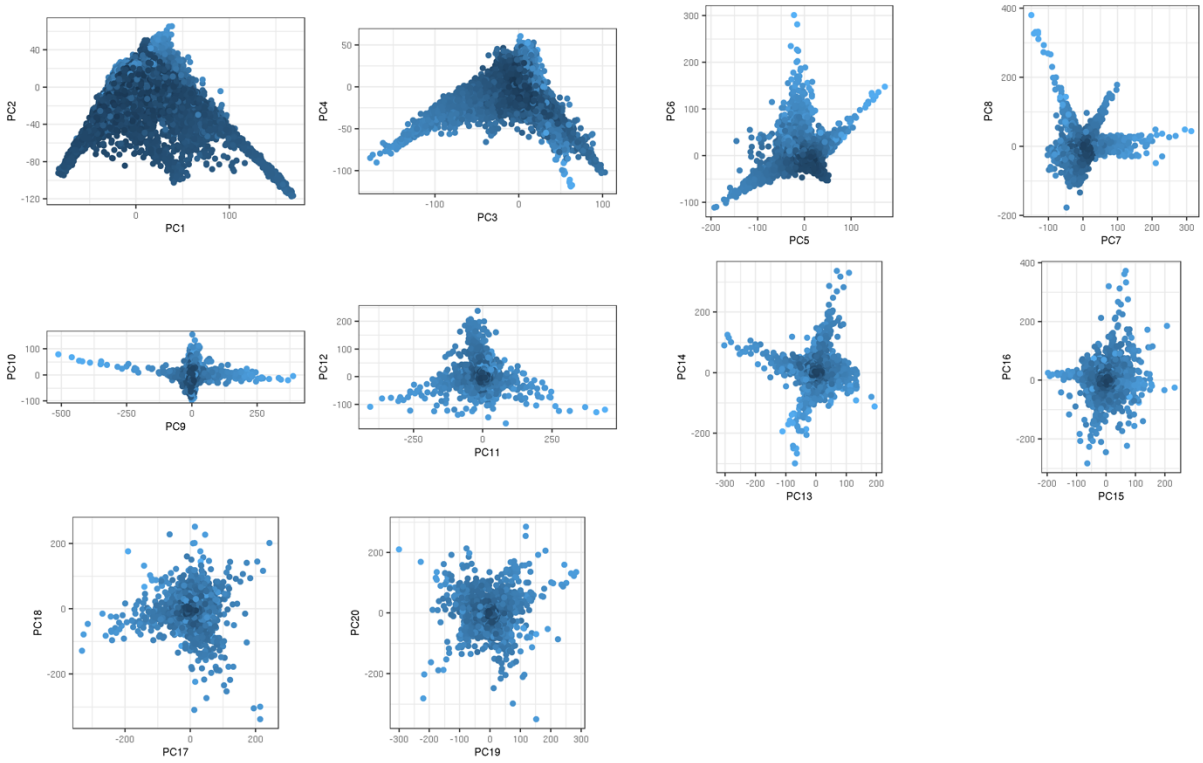


Supplementary Figure 5. Scree plot of the singular values of the first 20 principal components (PC) in 22,443 TOPMed samples of European genetic ancestry.

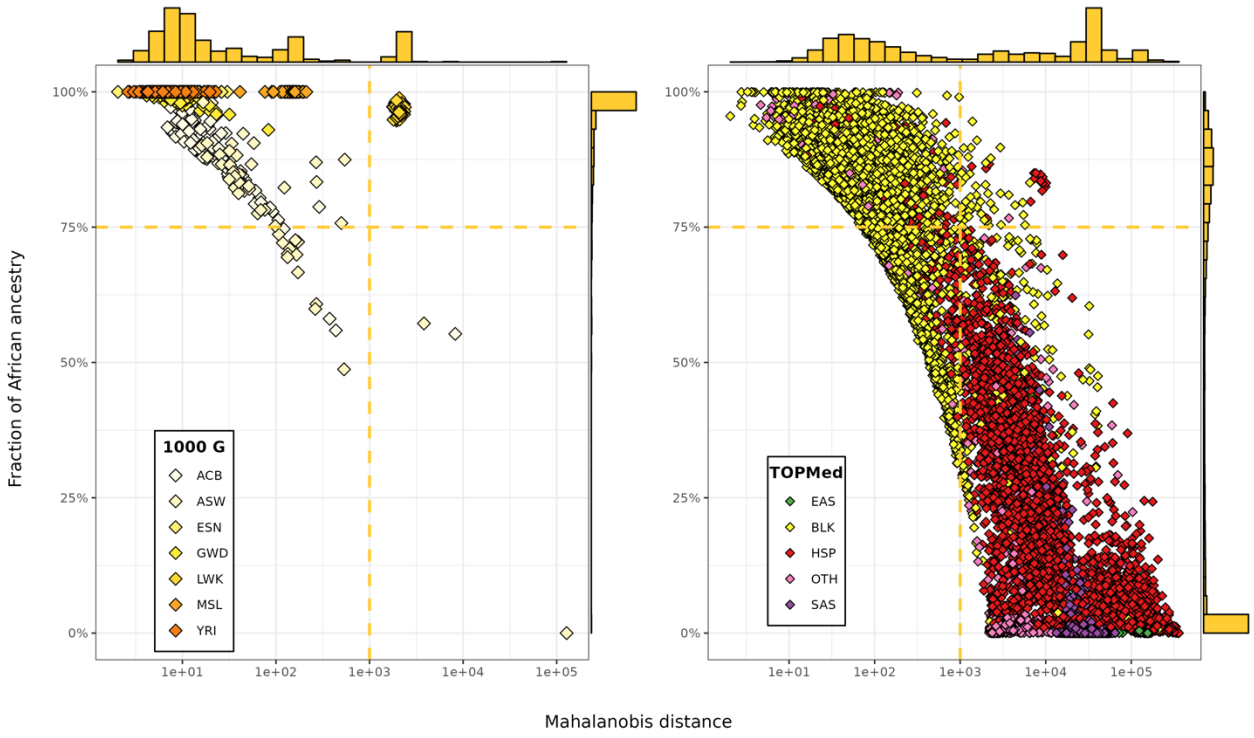


Supplementary Figure 6. SNV loadings for each of the first 20 PCs (PC1 to PC20) in 22,443 TOPMed samples of European genetic ancestry. The 5,047,983 SNVs are ordered by chromosome and position on the x-axis. Points are hex-binned and colored relative to their count.

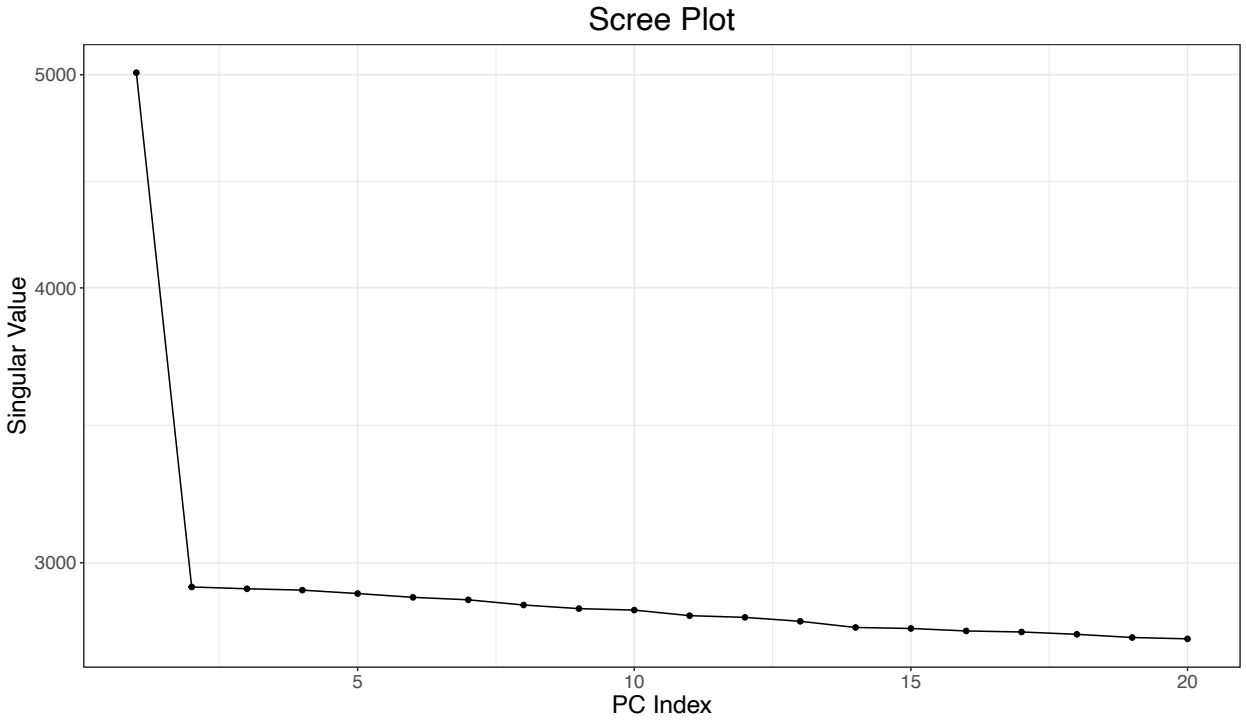
SNV, single nucleotide variant.



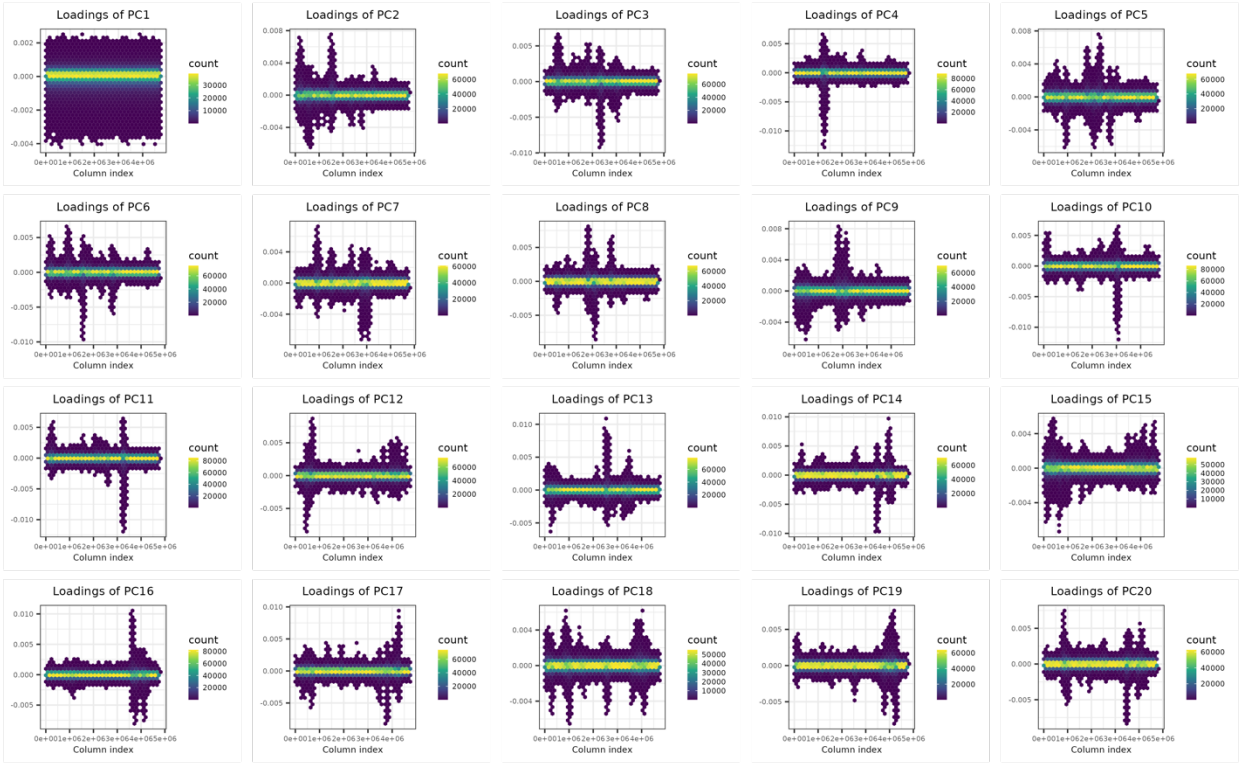
Supplementary Figure 7. Principal component scores (PC1 to PC20) of 22,443 TOPMed samples of European genetic ancestry. A lighter blue color represents a greater Mahalanobis distance.



Supplementary Figure 8. Fraction of African ancestry as computed by ADMIXTURE against the pairwise orthogonalized Gnanadesikan-Kettenrin robust Mahalanobis distances from the African continental cluster in the principal components analysis.
In both panels, yellow dotted lines indicate a fraction of African ancestry = 75% (horizontal), and Mahalanobis distance = 1000 (vertical), respectively. X-axes are displayed in log scale.
Left panel: 2,403 unrelated samples from seven African populations in the 1000 Genomes Project (1000 G).
ACB = African Caribbean in Barbados; ASW = African-American in Southwest USA; ESN = Esan in Nigeria; GWD = Gambian in Western Division, Mandinka; LWK = Luhya in Webuye, Kenya; MSL = Mende in Sierra Leone; YRI = Yoruba in Ibadan, Nigeria.
Right panel: 26,176 unrelated samples from the TOPMed project colored according to their self-identified race/ethnicity (SIRE). For clarity, self-identified White/European (WHT) participants are not shown in the plot.
BLK = Black/African American; EAS = East Asian; HSP = Hispanic/Latino; OTH = South Asian, Multiple, Missing/Unknown, and Other.

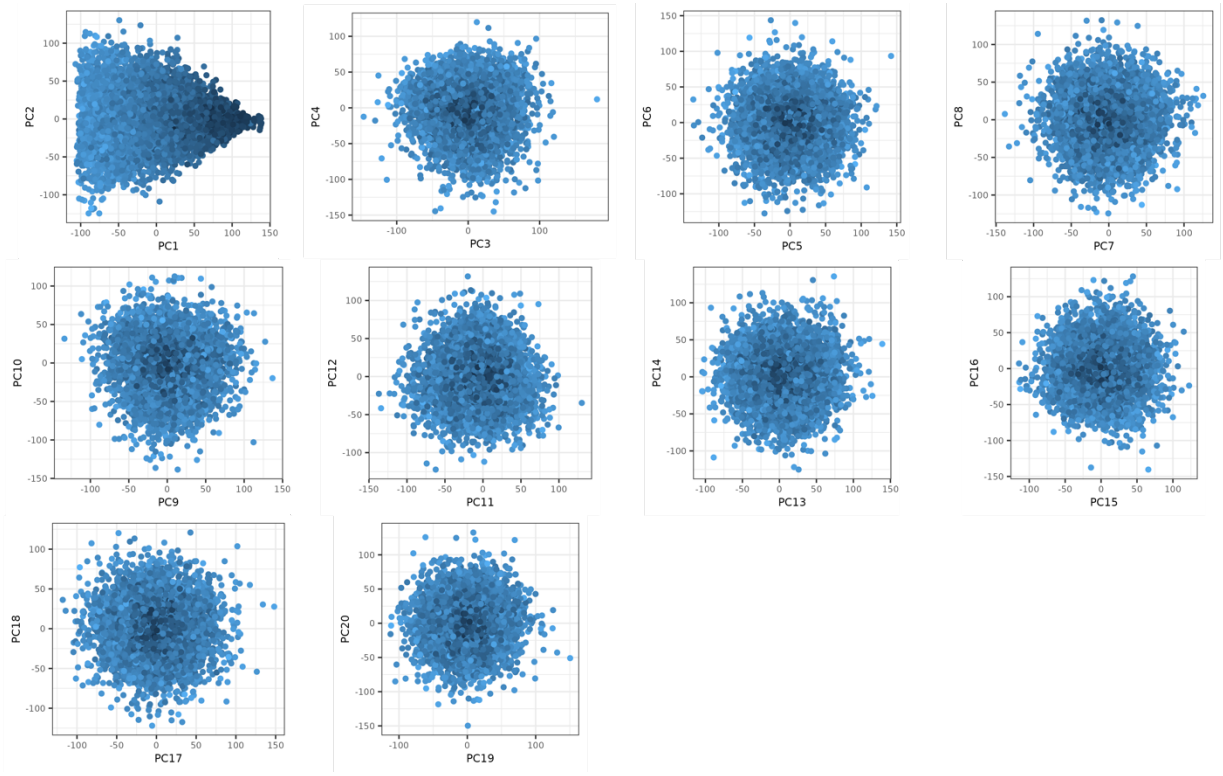


Supplementary Figure 9. Scree plot of the singular values of the first 20 principal components (PC) in 9,516 TOPMed samples of African genetic ancestry.

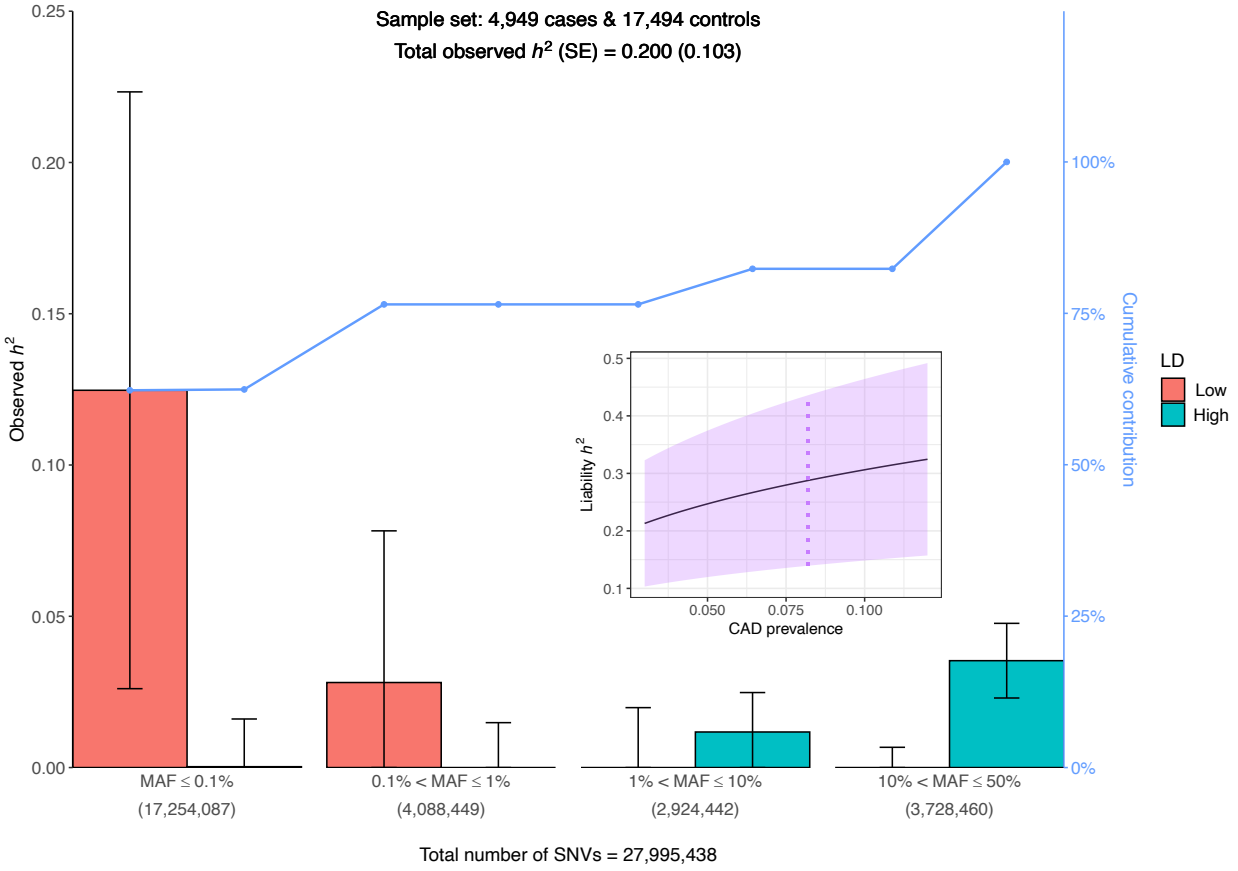


Supplementary Figure 10. SNV loadings for each of the first 20 PCs (PC1 to PC20) in 9,516 TOPMed samples of African genetic ancestry. The 4,748,912 SNVs are ordered by chromosome and position on the x-axis. Points are hex-binned and colored relative to their count.

SNV, single nucleotide variant.

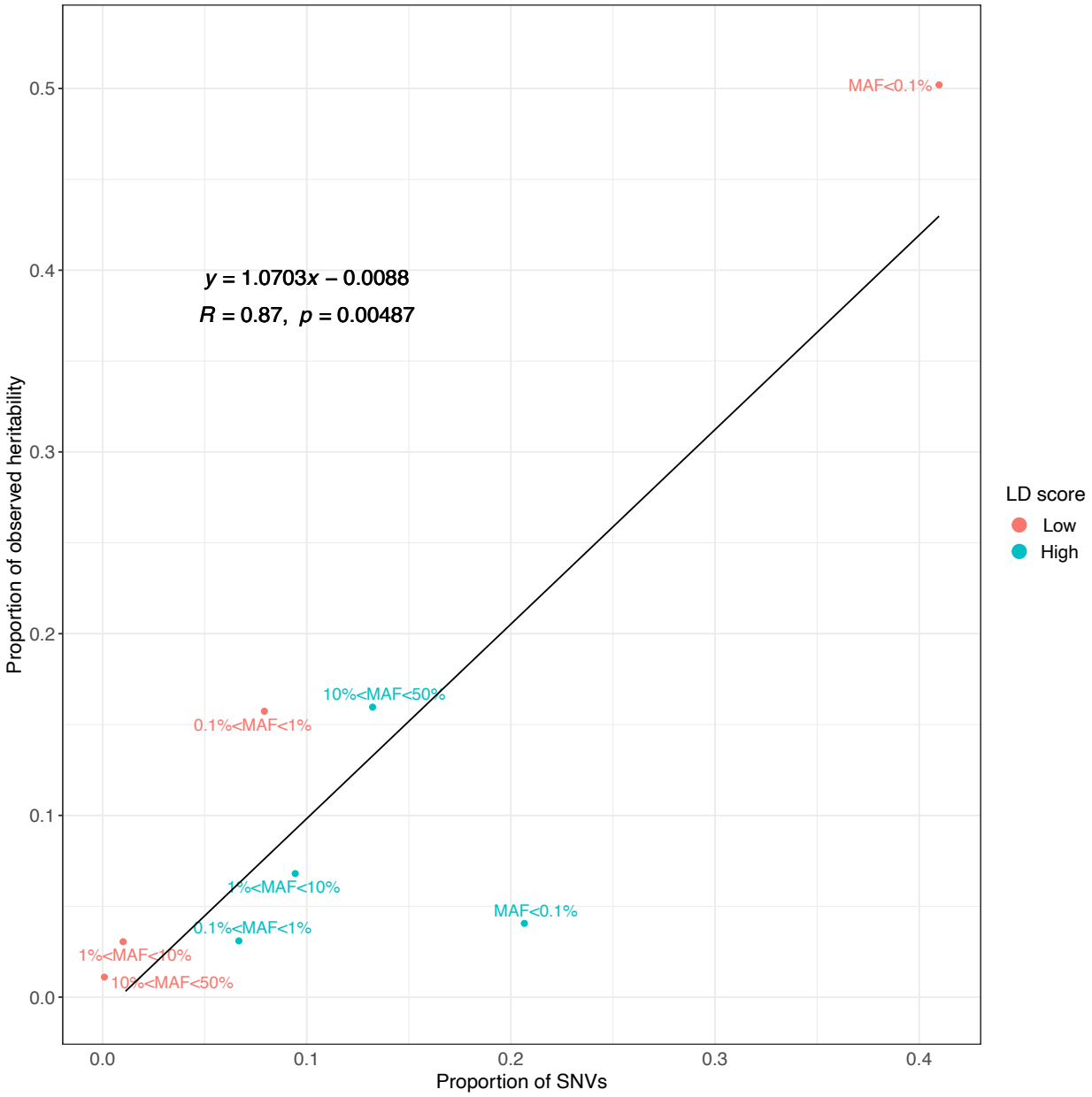


Supplementary Figure 11. Principal component scores (PC1 to PC20) of 9,516 TOPMed samples of African genetic ancestry. A lighter blue color represents a greater Mahalanobis distance.



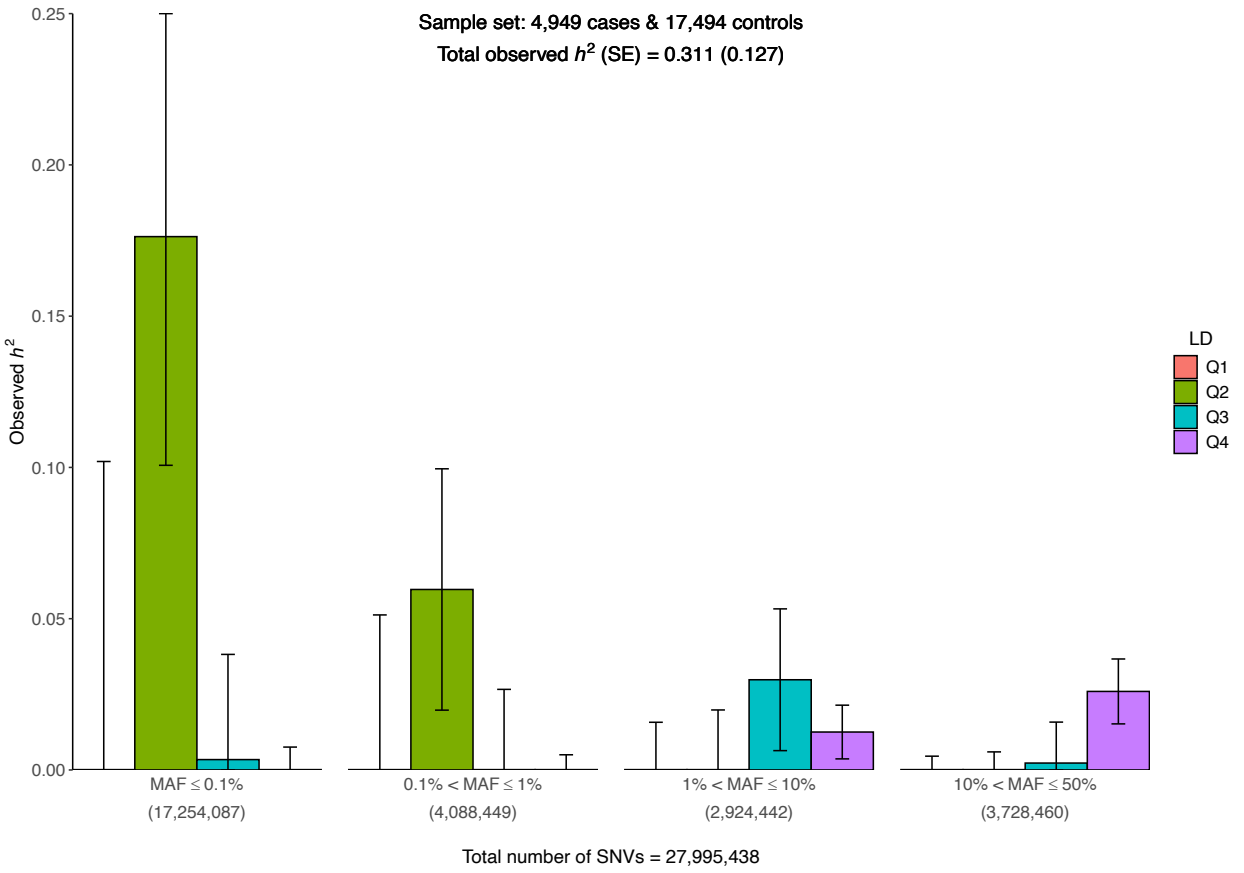
Supplementary Figure 12. Contribution of each LD score-MAF bin to the observed heritability h^2 of CAD in the European genetic ancestry sample. Error bars represent ± one SE from each contribution point estimate. SEs are calculated by GCTA and are proportional to the effective number of independent variants in each bin and inversely proportional to the total sample size (4,949 cases + 17,494 controls). The number of SNVs in each of the four MAF bins is indicated in parentheses. Low (High) category in the legend represents SNVs with LD scores below (above) the median, respectively. The broken line (in blue) displays the cumulative contribution (in %) of each LD score-MAF bin to the observed heritability estimate. Inset represents CAD heritability (estimate ± SE) on the liability scale for CAD prevalence ranging from 3% to 12% in the population (violet shaded area). The vertical dotted line (in violet) indicates the heritability estimate for a population prevalence of 8.2% in White/European ancestry[15]. The GRMs are estimated by the ratio of averages (RoA) method and contributions to h^2 are estimated with the REML AI algorithm.

CAD, coronary artery disease; GRM, genomic relatedness matrix; LD, linkage disequilibrium; MAF, minor allele frequency; SE, standard error; SNV, single nucleotide variant.

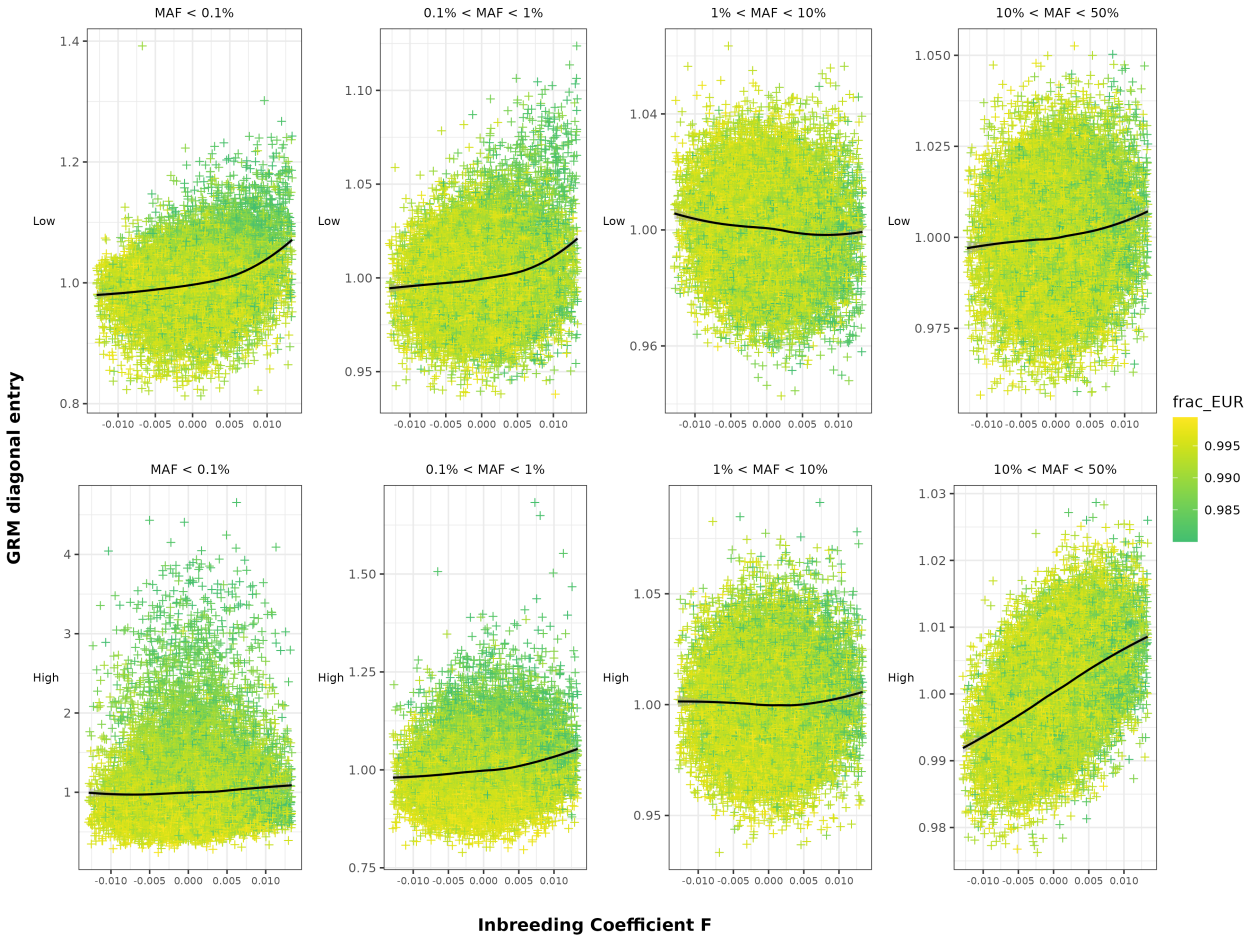


Supplementary Figure 13. Proportion of observed heritability in each LD score-MAF bin against the proportion of SNVs in that bin (number of SNVs in the bin divided by the total number of SNVs). The black line shows the regression line, whose equation is displayed in the upper left corner (n = 8). R designates the Pearson correlation coefficient, while p is the p-value associated to the two-sided test of null correlation.

LD, linkage disequilibrium; MAF, minor allele frequency; SNV, single nucleotide variant.

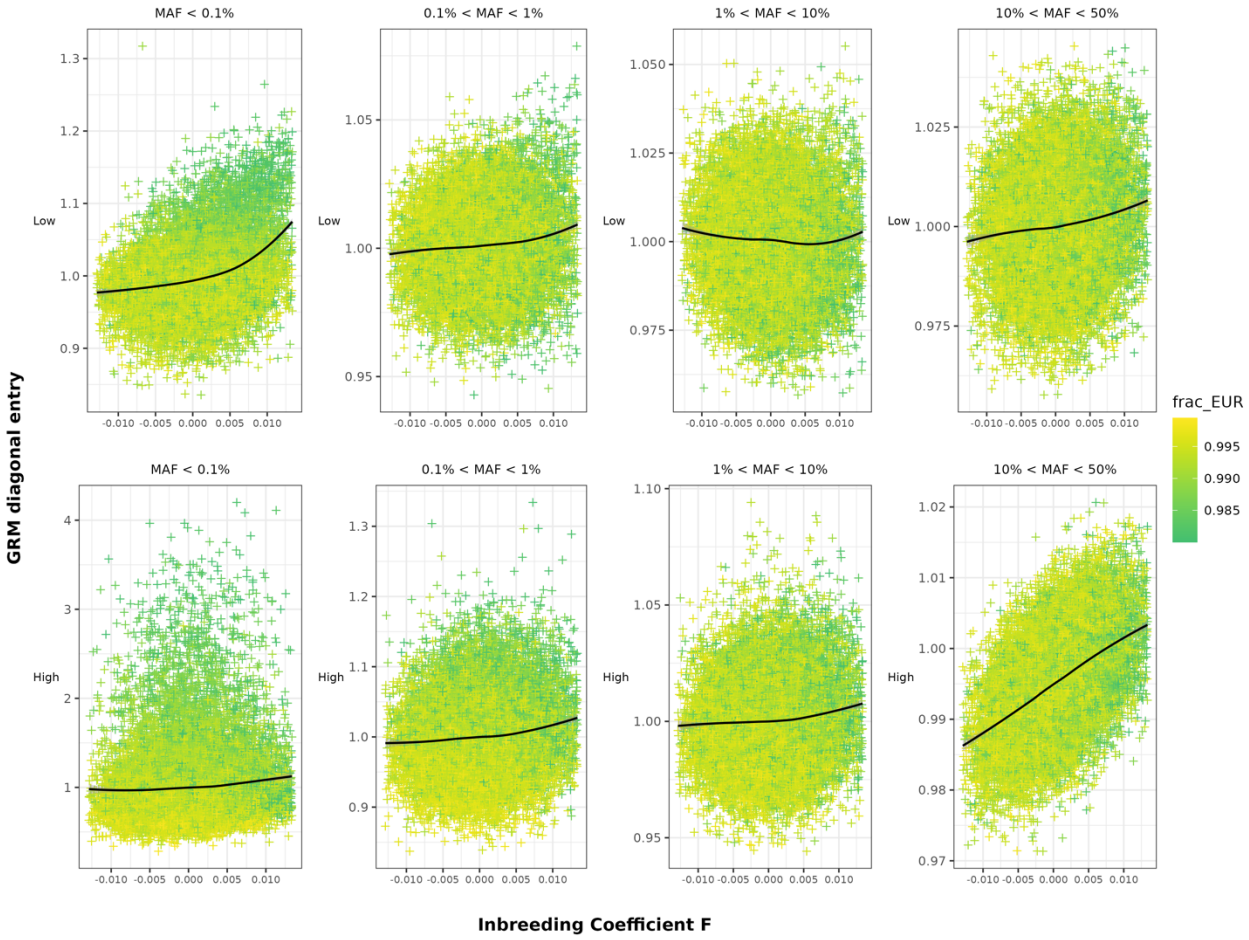


Supplementary Figure 14. Contribution of each LD score-MAF bin to the observed heritability h^2 of CAD in the European genetic ancestry sample. Error bars represent ± one SE from each contribution point estimate. SEs are calculated by GCTA and are proportional to the effective number of independent variants in each bin and inversely proportional to the total sample size (4,949 cases + 17,494 controls). The number of SNVs in each of the four MAF bins is indicated in parentheses. The first, second, third and fourth quartiles of LD scores are denoted Q1, Q2, Q3, and Q4, respectively. The GRMs are estimated by the average of ratios (AoR) method and contributions to h^2 are estimated with the REML EM algorithm.
CAD, coronary artery disease; GRM, genomic relatedness matrix; LD, linkage disequilibrium; MAF, minor allele frequency; SE, standard error; SNV, single nucleotide variant.



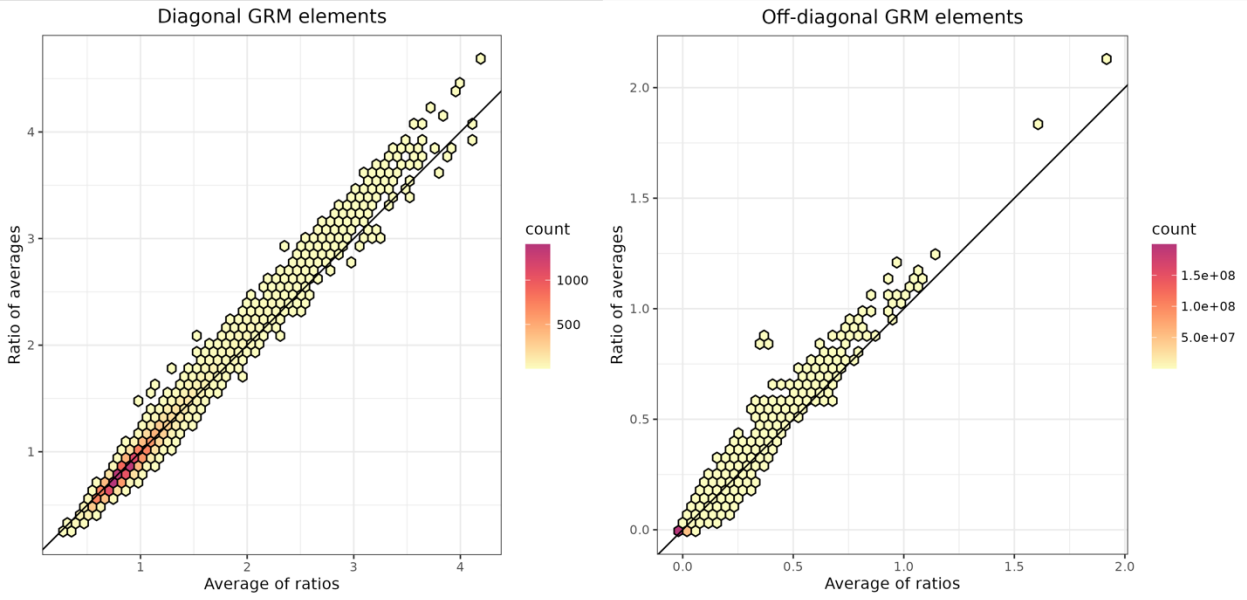
Supplementary Figure 15. GRMs estimated by average of ratios (AoR) method. For each sample, the diagonal value of each GRM matrix (*n* = 22,443) is plotted against the sample's inbreeding coefficient *F*. The eight GRMs are displayed by MAF (columns) and LD scores (rows). Top (bottom) row displays GRMs computed from SNVs with LD scores below (above) the median, respectively. Black lines show LOESS regression fits. Color coding represents the fraction of European ancestry for each sample (frac_EUR varies between 0.98 and 1) as estimated by ADMIXTURE.

GRM, genomic relatedness matrix; LD, linkage disequilibrium; LOESS, locally estimated scatterplot smoothing; MAF, minor allele frequency; SNV, single nucleotide variant.



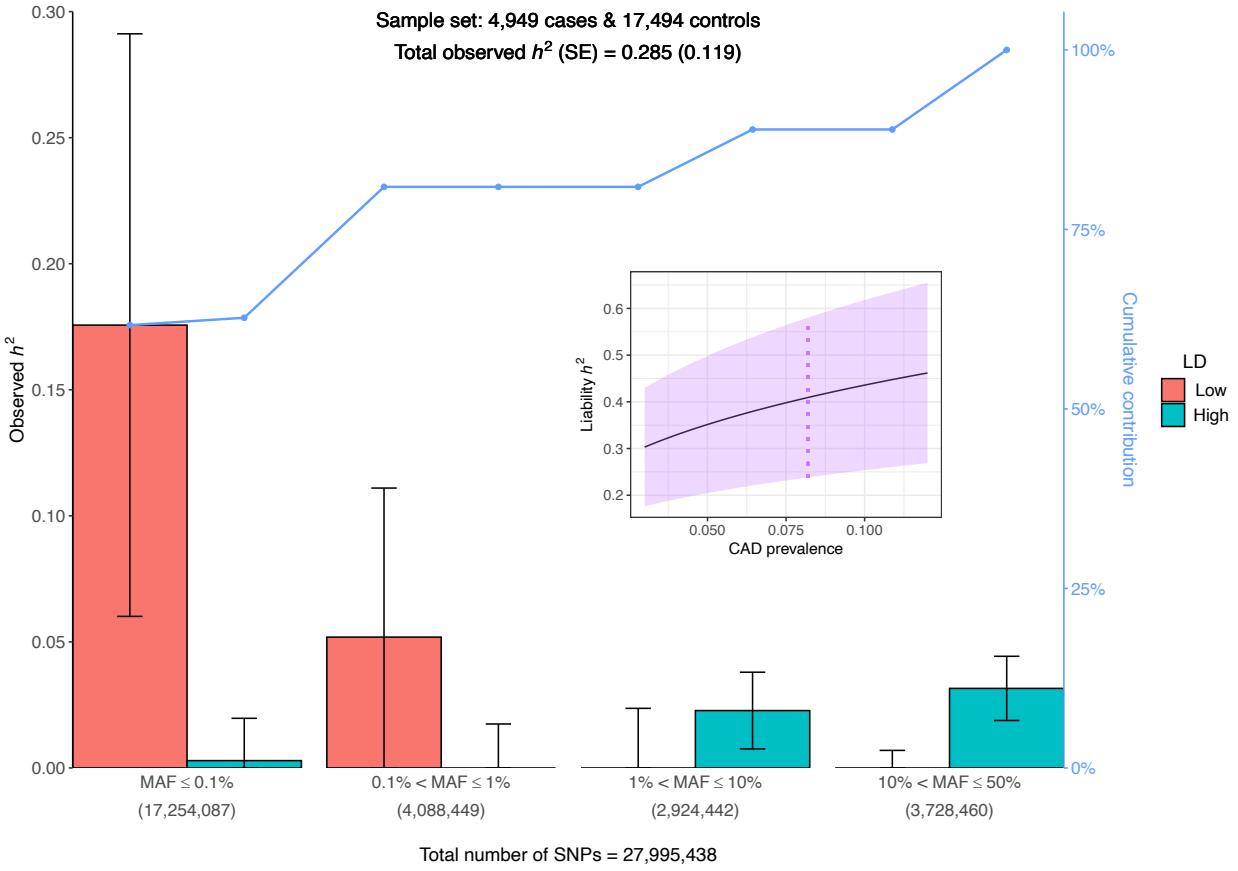
Supplementary Figure 16. GRMs estimated by the ratio of averages (RoA) method. For each sample, the diagonal value of each GRM matrix (*n* = 22,443) is plotted against the sample's inbreeding coefficient *F*. The eight GRMs are displayed by MAF (columns) and LD scores (rows). Top (bottom) row displays GRMs computed from SNVs with LD scores below (above) the median, respectively. Black lines show LOESS regression fits. Color coding represents the fraction of European ancestry for each sample (frac_EUR varies between 0.98 and 1) as estimated by ADMIXTURE.

GRM, genomic relatedness matrix; LD, linkage disequilibrium; LOESS, locally estimated scatterplot smoothing; MAF, minor allele frequency; SNV, single nucleotide variant.



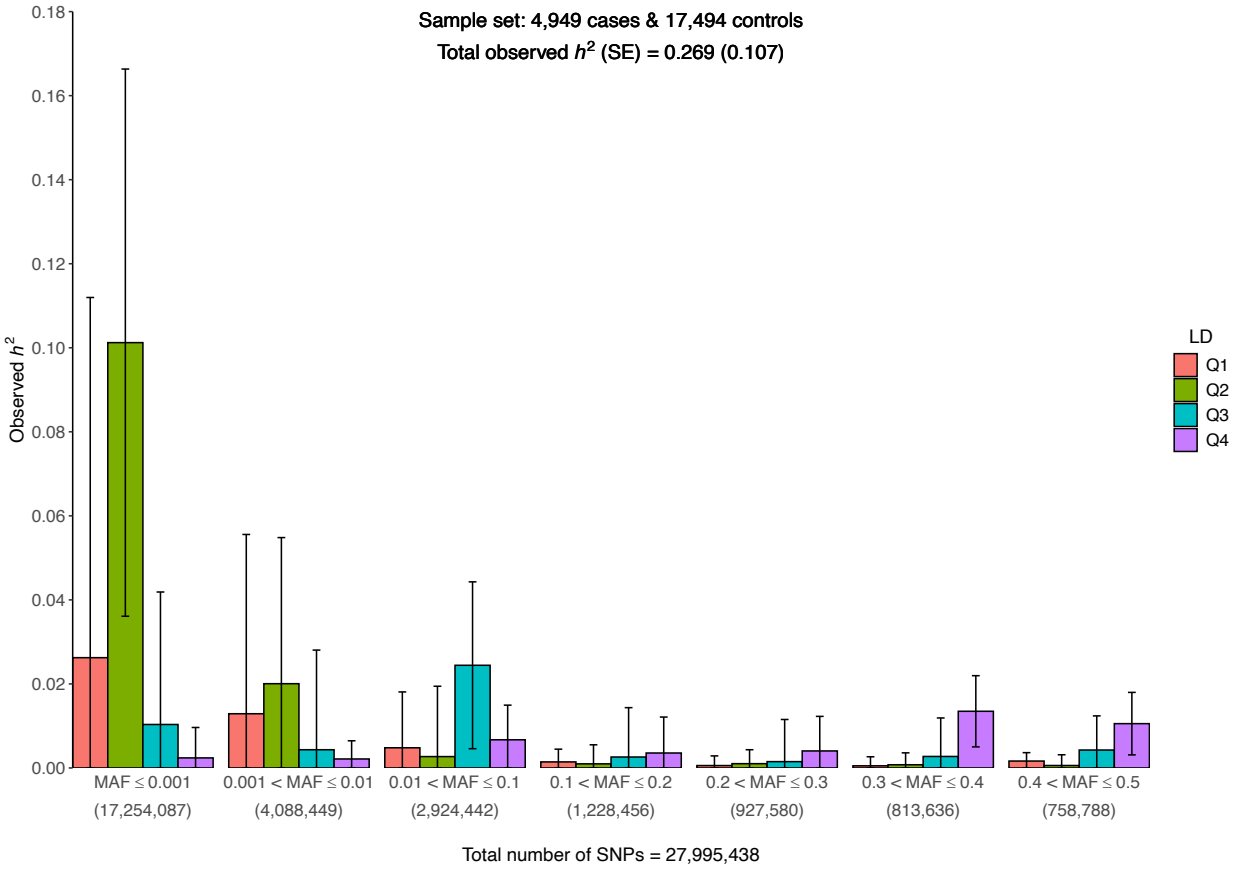
Supplementary Figure 17. Comparison of GRM estimation methods ratio of averages (RoA) versus average of ratios (AoR) for SNVs with high LD scores (above the median) and MAF ≤ 0.1%. The black line indicates a straight line of slope 1 and intercept 0. The RoA method show larger values compared to the AoR method for both the diagonal elements (left panel, *n* = 22,443) and off-diagonal elements (right panel, *n* = 22,443×22,442/2 = 251,832,903). Points are hex-binned and colored relative to their count.

GRM, genomic relatedness matrix; LD, linkage disequilibrium; MAF, minor allele frequency; SNV, single nucleotide variant.



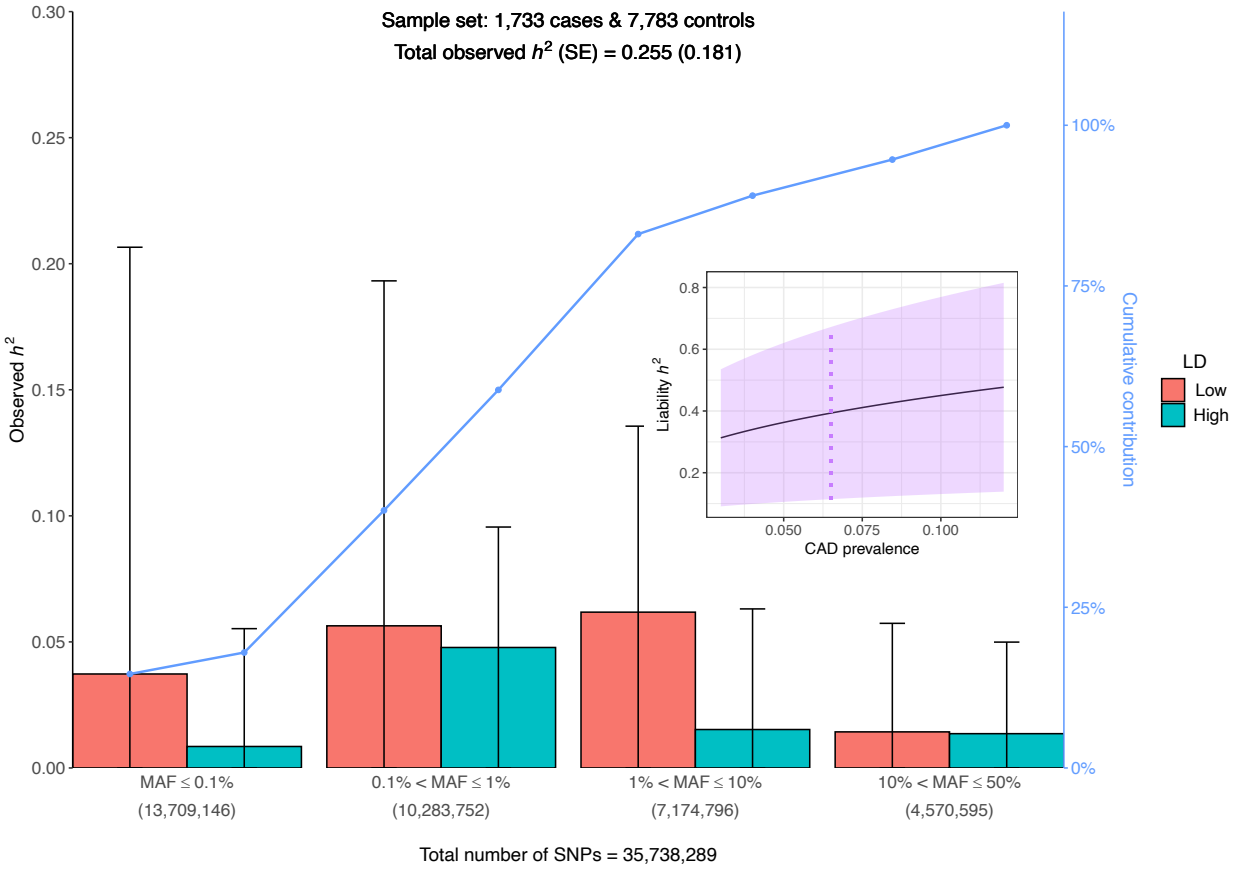
Supplementary Figure 18. Contribution of each LD score-MAF bin to the observed heritability h^2 of CAD in the European genetic ancestry sample. Error bars represent ± one SE from each contribution point estimate. SEs are calculated by GCTA and are proportional to the effective number of independent variants in each bin and inversely proportional to the total sample size (4,949 cases + 17,494 controls). The number of SNVs in each of the four MAF bins is indicated in parentheses. Low (High) category in the legend represents SNVs with LD scores below (above) the median, respectively. The broken line (in blue) displays the cumulative contribution (in %) of each LD score-MAF bin to the observed heritability estimate. Inset represents CAD heritability (estimate ± SE) on the liability scale for CAD prevalence ranging from 3% to 12% in the population (violet shaded area). The vertical dotted line (in violet) indicates the heritability estimate for a population prevalence of 8.2% in White/European ancestry[15]. The GRMs are estimated by the average of ratios (AoR) method and contributions to h^2 are estimated with the REML AI algorithm.

CAD, coronary artery disease; GRM, genomic relatedness matrix; LD, linkage disequilibrium; MAF, minor allele frequency; SE, standard error; SNV, single nucleotide variant.



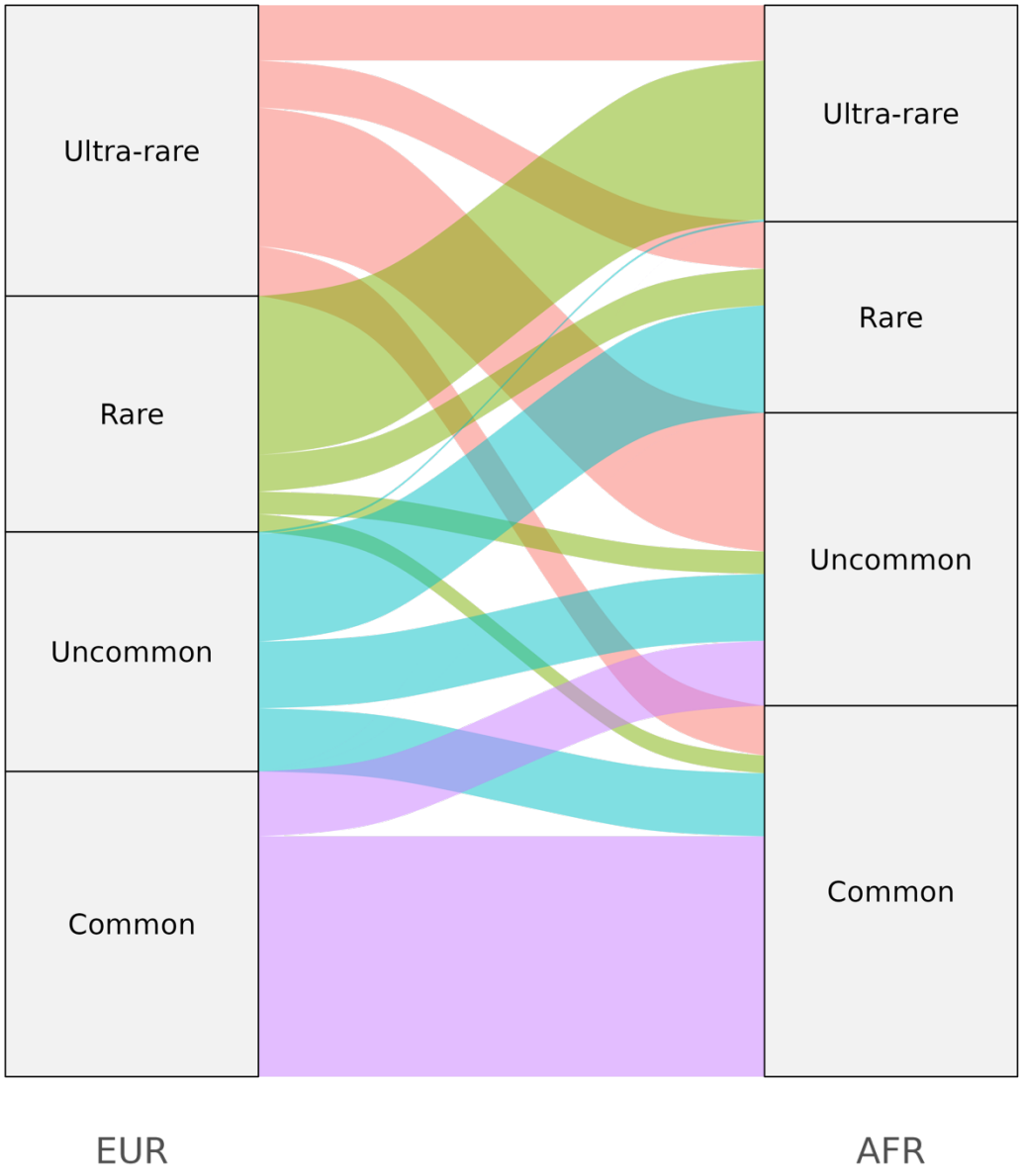
Supplementary Figure 19. Contribution of each LD score-MAF bin to the observed heritability h^2 of CAD in the European genetic ancestry sample using Tcheandjieu's et al.[15] LD score-MAF binning and adding ultra-rare SNVs. Error bars represent ± one SE from each contribution point estimate. SEs are calculated by GCTA and are proportional to the effective number of independent variants in each bin and inversely proportional to the total sample size (4,949 cases + 17,494 controls). The number of SNVs in each of the seven MAF bins is indicated in parentheses. The first, second, third and fourth quartiles of LD scores are denoted Q1, Q2, Q3, and Q4, respectively. The 28 GRMs are estimated by the ratio of averages (RoA) method and contributions to h^2 by the REML EM algorithm.

CAD, coronary artery disease; GRM, genomic relatedness matrix; LD, linkage disequilibrium; MAF, minor allele frequency; SE, standard error; SNV, single nucleotide variant.

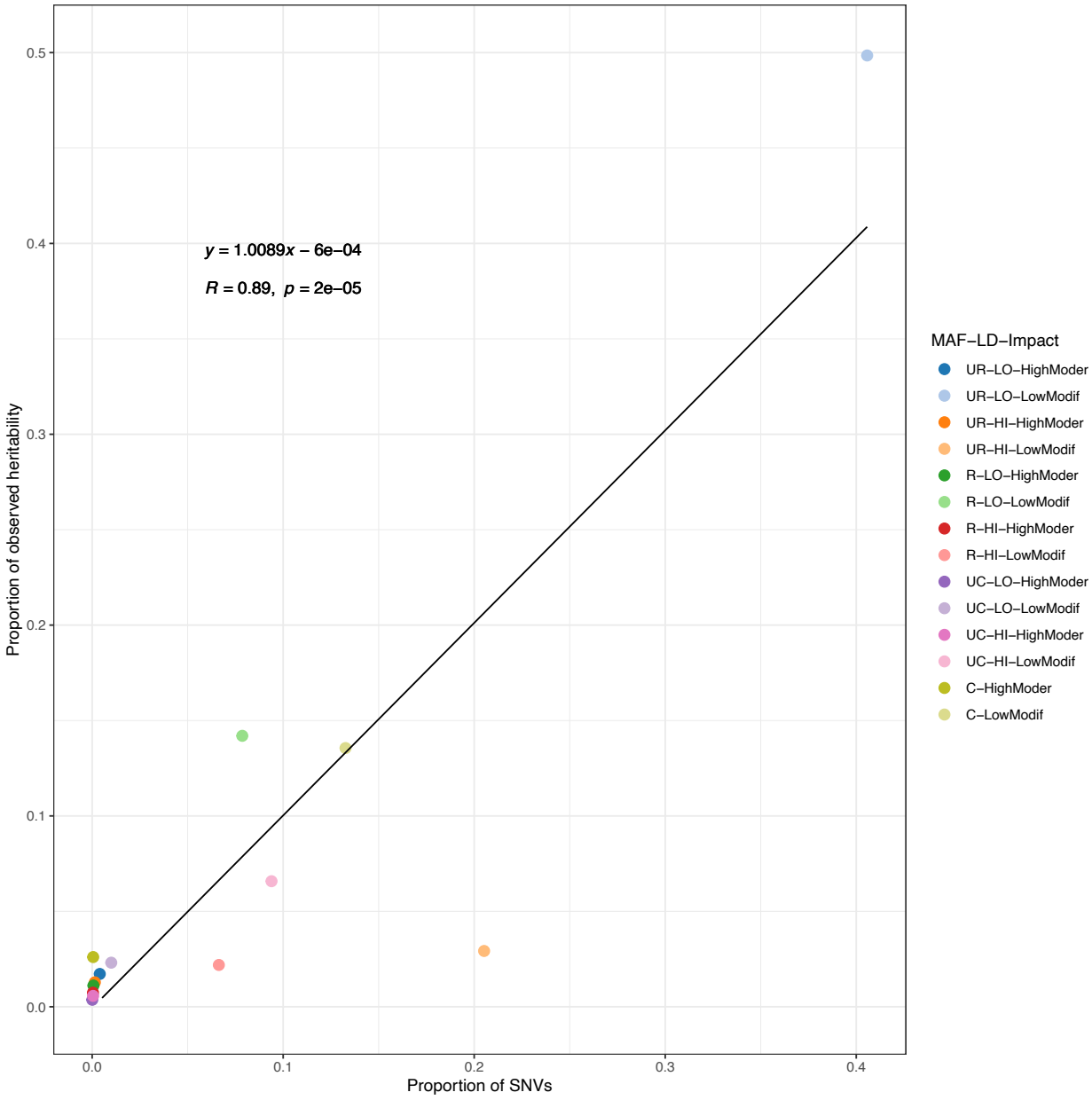


Supplementary Figure 20. Contribution of each LD score-MAF bin to the observed heritability h^2 of CAD in African ancestry. Error bars represent ± one SE from each contribution point estimate. SEs are calculated by GCTA and are proportional to the effective number of independent variants in each bin and inversely proportional to the total sample size (1,733 cases + 7,783 controls). The number of SNVs in each of the four MAF bins is indicated in parentheses. Low (High) category in the legend represents SNVs with LD scores below (above) the median, respectively. The broken line (in blue) displays the cumulative contribution (in %) of each LD score-MAF bin to the observed heritability estimate. Inset represents CAD heritability (estimate ± SE) on the liability scale for CAD prevalence ranging from 3% to 12% in the population (violet shaded area). The vertical dotted line (in violet) indicates the heritability estimate for a population prevalence of 6.5% in the U.S. Black population[15]. The GRMs are estimated by the ratio of averages (RoA) method and contributions to h^2 are estimated by the REML EM algorithm.

CAD, coronary artery disease; LD, linkage disequilibrium; MAF, minor allele frequency; SE, standard error; SNV, single nucleotide variant.

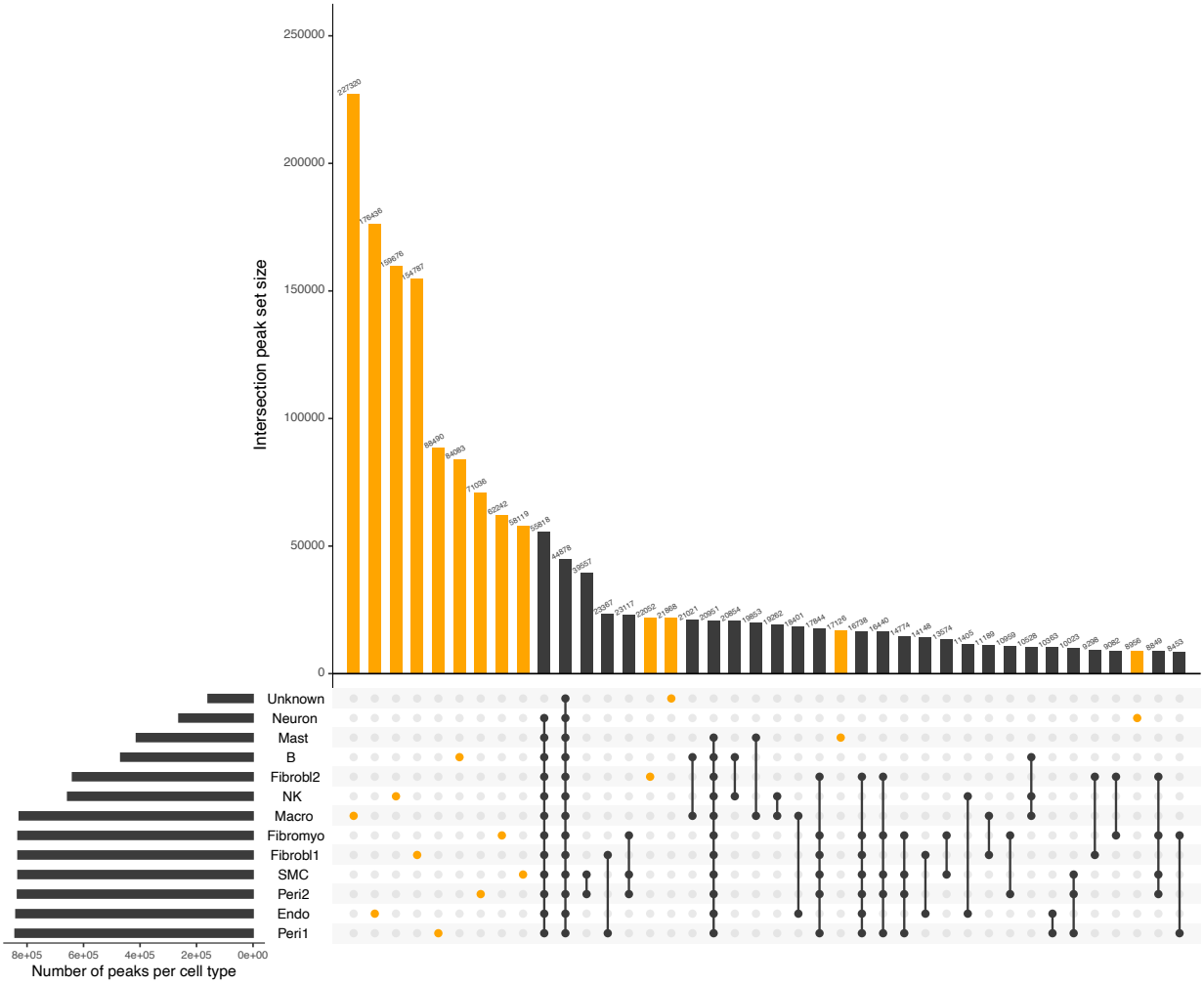


Supplementary Figure 21. Sankey diagram showing the distribution of shared SNVs (*n* = 12,889,748) in the European (EUR) and African (AFR) genetic ancestry samples across the MAF bins. The largest overlap is found in the common variant bin.
MAF, minor allele frequency; SNV, single nucleotide variant.

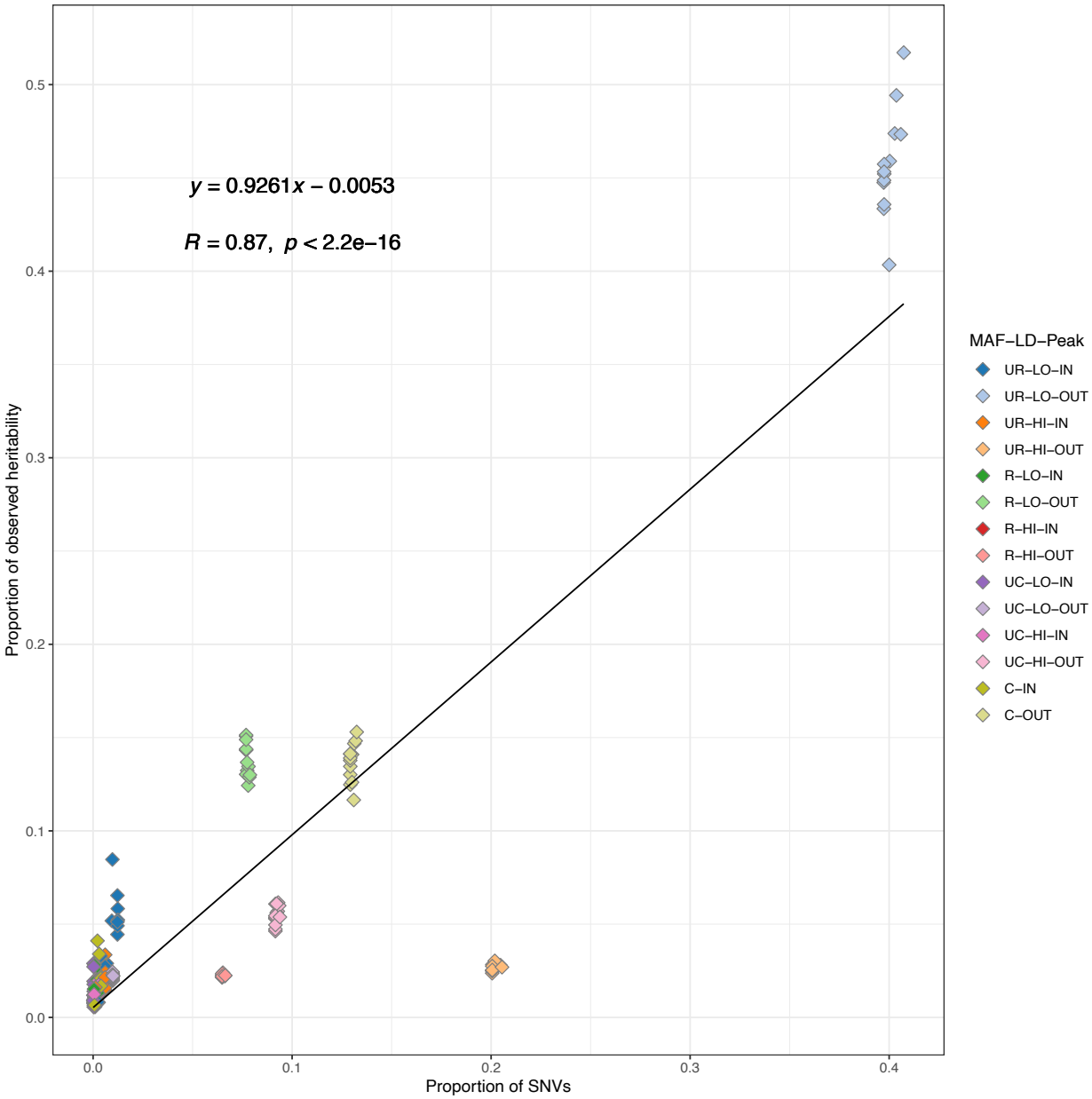


Supplementary Figure 22. Proportion of observed heritability in each LD score-MAF-Impact against the proportion of SNVs in that bin (number of SNVs in the bin divided by the total number of SNVs). Functional impact was predicted by SnpEff. Each label in the legend represents a combination of: i) MAF (UR: ultra-rare (MAF ≤ 0.1%), R: rare (0.1% < MAF ≤ 1%), UC: uncommon (1% < MAF ≤ 10%), C: common (10% < MAF ≤ 50%)); ii) LD score (LO: Low, HI: High); and iii) Impact (HighModer: protein-altering variants, LowModif: non-protein-altering variants). The black line shows the regression line, whose equation is displayed in the upper left corner (n = 14). R designates the Pearson correlation coefficient, while p is the p-value associated to the two-sided test of null correlation.

LD, linkage disequilibrium; MAF, minor allele frequency; SNV, single nucleotide variant.

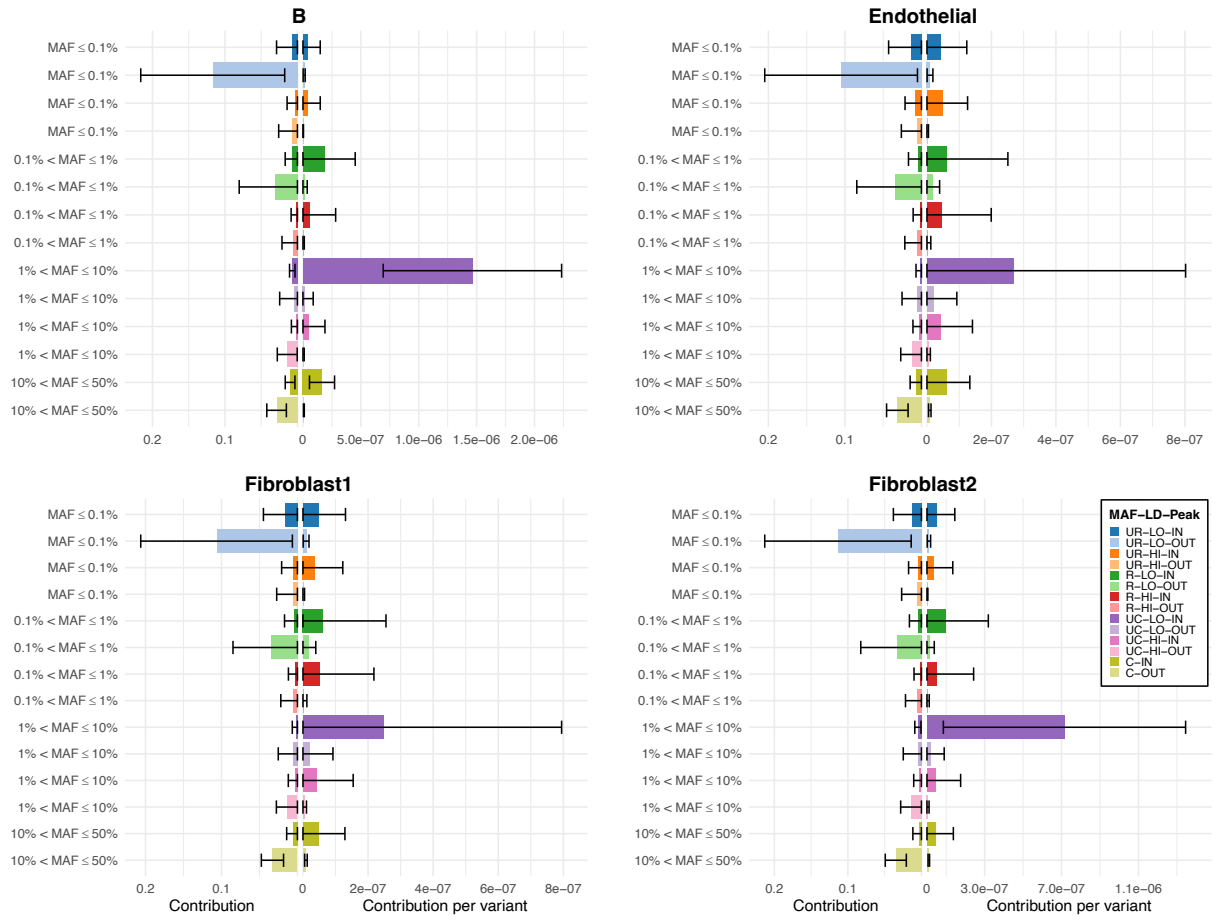


Supplementary Figure 23. Distribution of the number of SNVs (N = 2,553,042) overlapping with the snATAC-seq peaks reported in Turner et al.[16]. The intersection set sizes across all 13 cell types are ordered from the largest to the smallest (not all $2^{13} - 1$ sets are shown). Sets of SNVs that are unique to only one cell type are highlighted in orange.
Endo, endothelial cells; Fibrobl, fibroblasts; Fibromyo, fibromyocytes; Macro, macrophages, NK, natural killer cells; Peri, Pericytes, SMC, smooth muscle cells; snATAC-seq, single-nucleus assay for transposase accessible chromatin with sequencing; SNV, Single Nucleotide Variant.



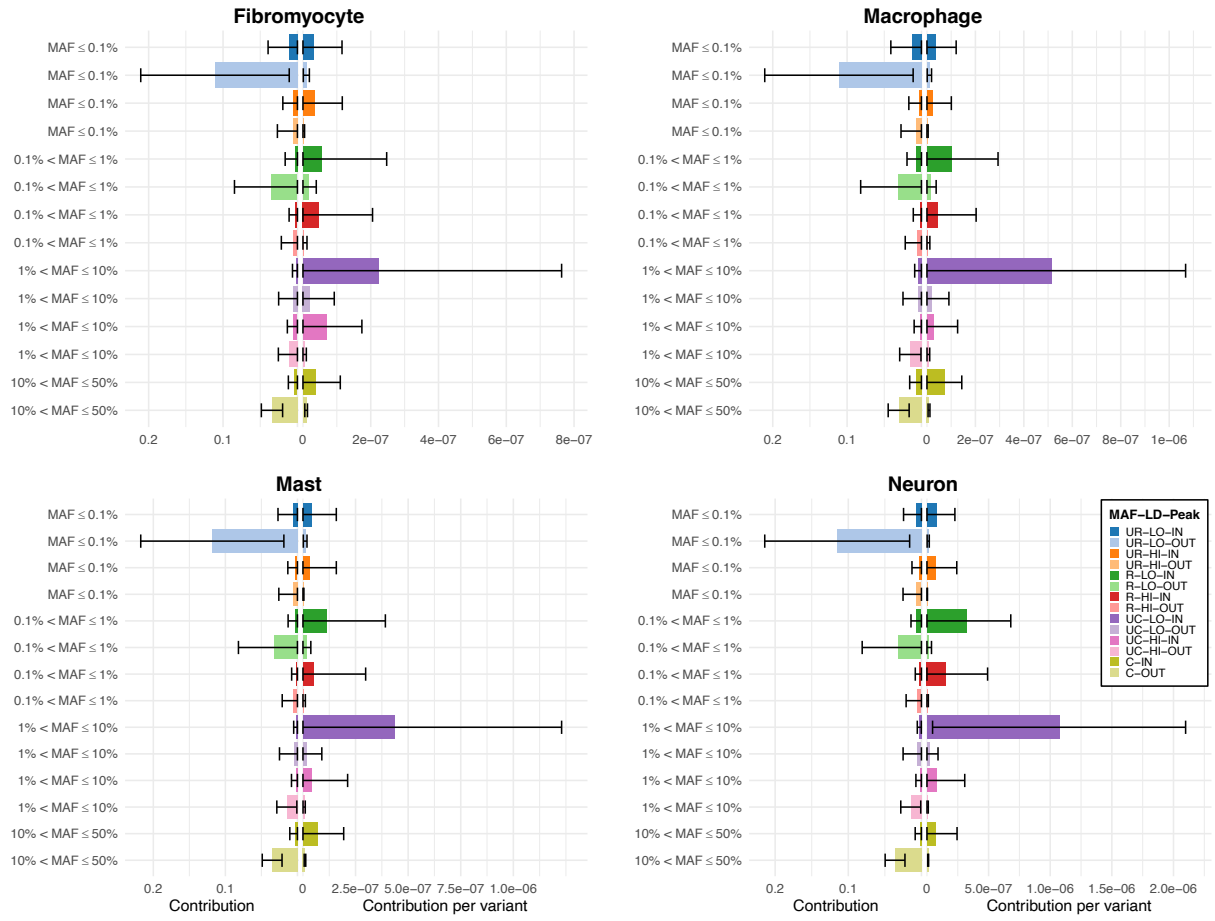
Supplementary Figure 24. Proportion of observed heritability in each LD score-MAF-Peak bin against the proportion of SNVs in that bin (number of SNVs in the bin divided by the total number of SNVs) for all 13 snATAC-seq cell types. Each label in the legend represents a combination of: i) MAF (UR: ultra-rare (MAF ≤ 0.1%), R: rare (0.1% < MAF ≤ 1%), UC: uncommon (1% < MAF ≤ 10%), C: common (10% < MAF ≤ 50%)); ii) LD score (LO: low, HI: high); and iii) Peak (IN: inside, OUT: outside). Each colored point is repeated 13 times. The black line shows the regression line, whose equation is displayed in the upper left corner (n = 14×13 = 182). R designates the Pearson correlation coefficient, while p is the p-value associated to the two-sided test of null correlation.

LD, linkage disequilibrium; MAF, minor allele frequency; SNV, single nucleotide variant.

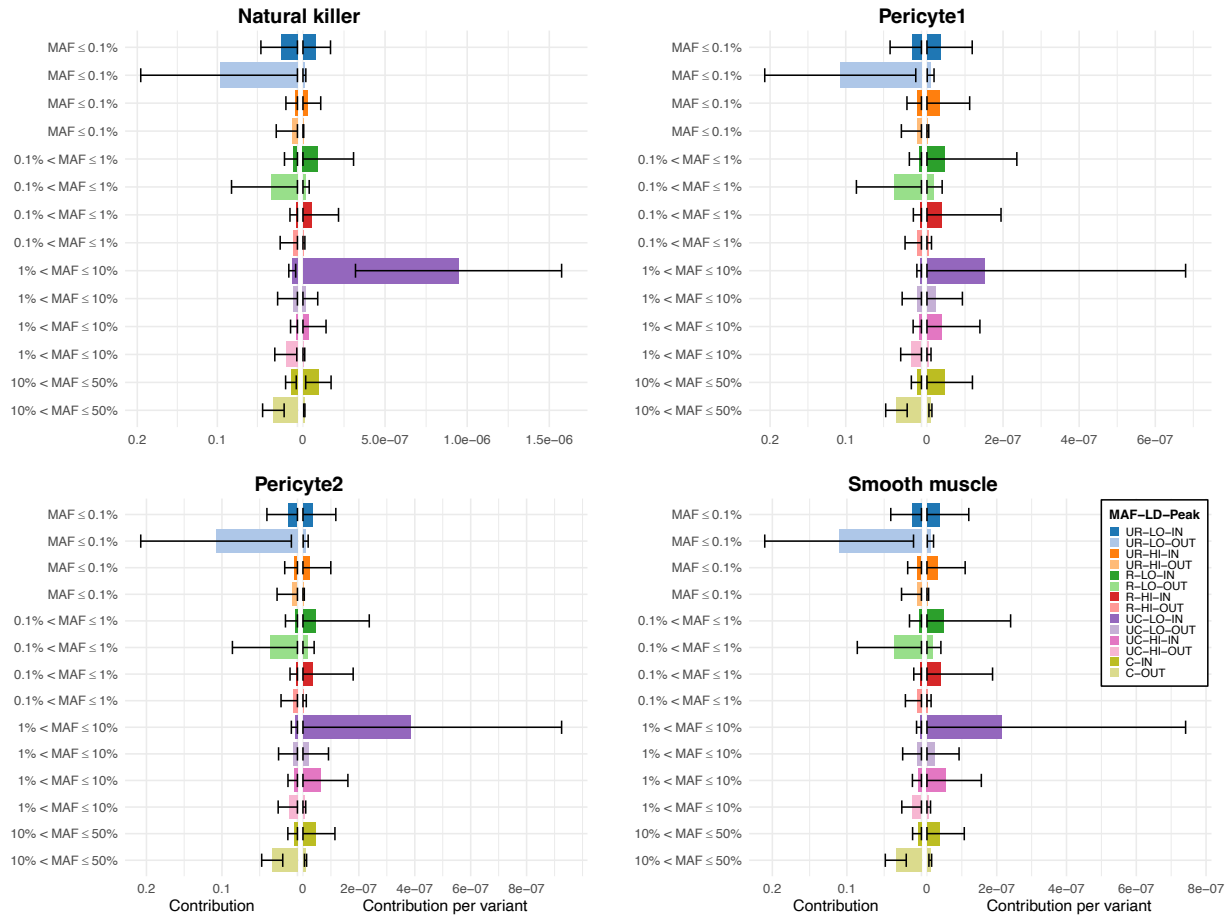


Supplementary Figure 25. Absolute (left) and relative (right) contribution per variant of each MAF-LD score-Peak bin to the CAD heritability estimate for each cell type. Each label in the legend represents a combination of: i) MAF (UR: ultra-rare (MAF ≤ 0.1%), R: rare (0.1% < MAF ≤ 1%), UC: uncommon (1% < MAF ≤ 10%), C: common (10% < MAF ≤ 50%)); ii) LD score (LO: low, HI: high); and iii) Peak (IN: inside, OUT: outside). Error bars represent ± one SE from each contribution point estimate. Absolute SEs (left) are calculated by GCTA and are proportional to the effective number of independent variants in each bin and inversely proportional to the total sample size (4,949 cases + 17,494 controls). Relative SEs (right) are obtained by dividing the corresponding absolute SEs by the square root of the number of variants.

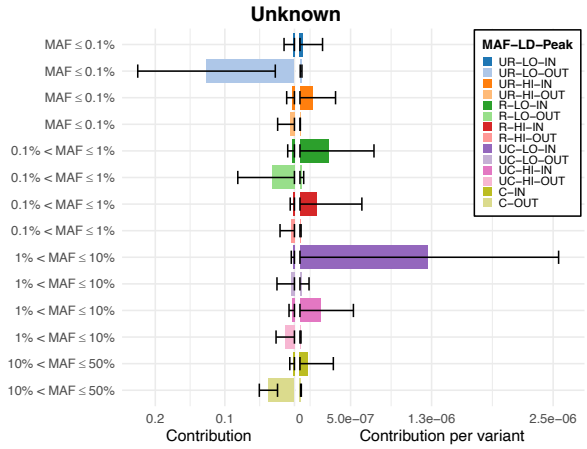
CAD, coronary artery disease; LD, linkage disequilibrium; MAF, minor allele frequency; SE, standard error; SNV, single nucleotide variant.



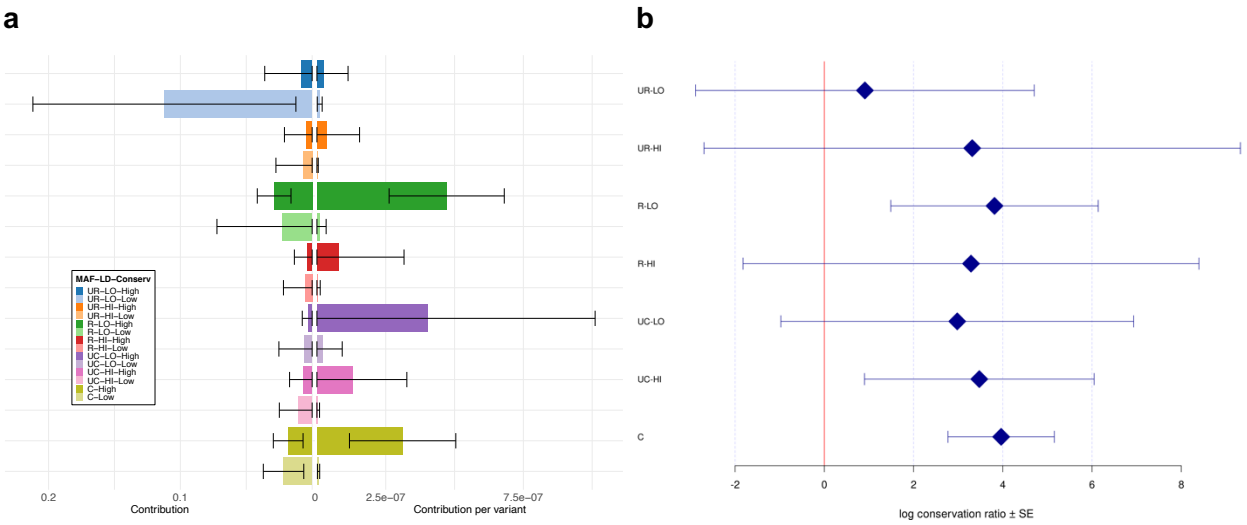
Supplementary Figure 25 (continued)



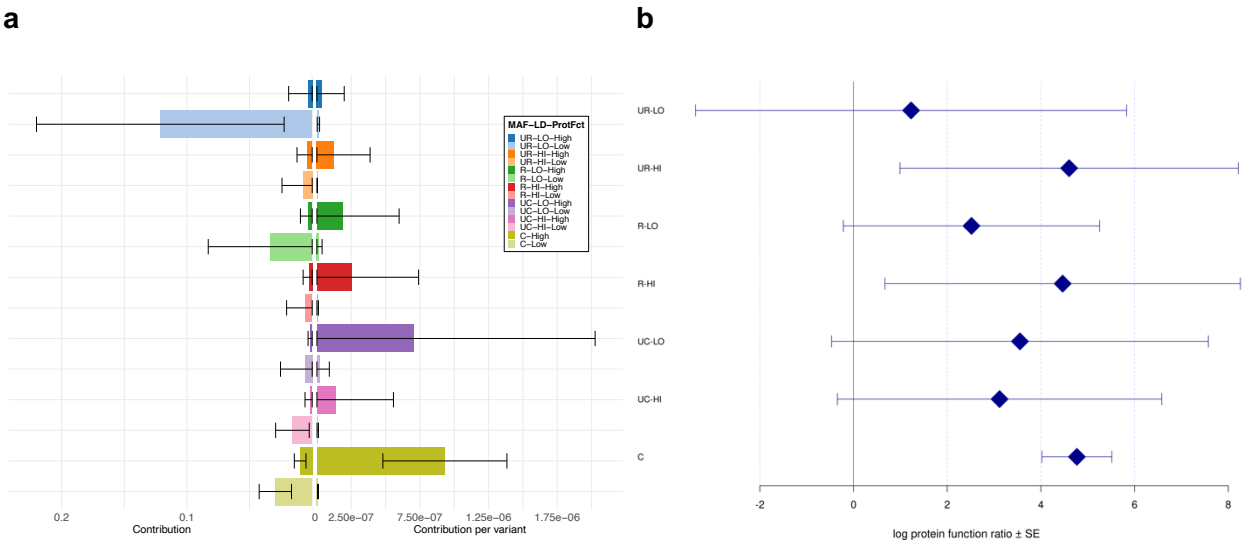
Supplementary Figure 25 (continued)



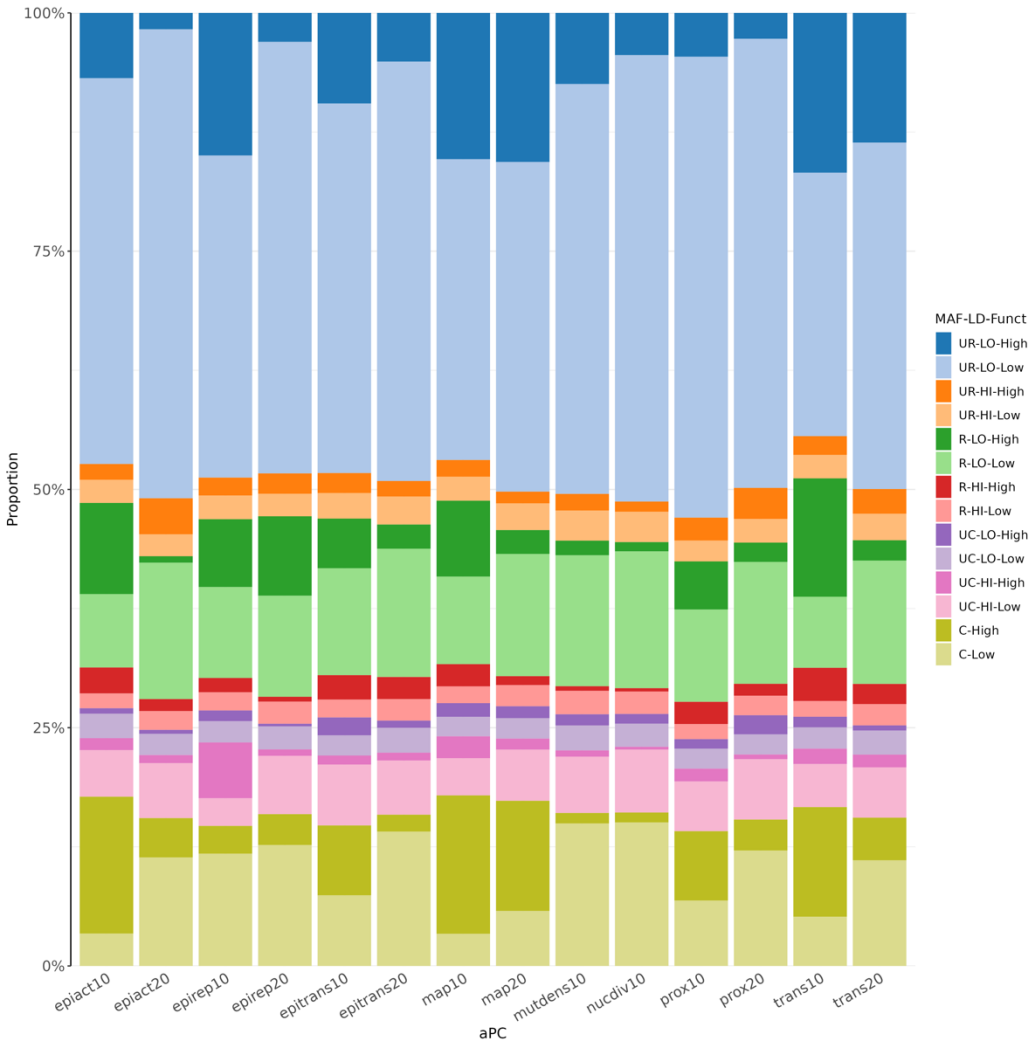
Supplementary Figure 25 (continued)



Supplementary Figure 26. Contribution of each LD score-MAF-aPC-Conservation bin to the global CAD heritability estimate, along with log enriched conservation ratio of each LD score-MAF bin. a, Absolute (left) and relative (right) contribution per variant of each LD score-MAF-aPC-Conservation bin to the global CAD heritability estimate. Each label in the legend represents a combination of: i) MAF (UR: ultra-rare (MAF ≤ 0.1%), R: rare (0.1% < MAF ≤ 1%), UC: uncommon (1% < MAF ≤ 10%), C: common (10% < MAF ≤ 50%)); ii) LD score (LO: low, HI: high); and iii) Conservation functionality (High, Low). Error bars represent ± one SE from each contribution point estimate. Absolute SEs (left) are calculated by GCTA and are proportional to the effective number of independent variants in each bin and inversely proportional to the total sample size (4,949 cases + 17,494 controls). Relative SEs (right) are obtained by dividing the corresponding absolute SEs by the square root of the number of variants. **b**, Log conservation ratio of conserved over non-conserved variants in each LD score-MAF bin. Each label on the y-axis is defined as in **a**. Error bars represent ± one SE from each log conservation ratio estimate. SEs are calculated from GCTA's output of the covariance matrix of contribution estimates to heritability in each bin and their corresponding number of SNVs (see section "Mean and variance of log enrichment ratio" for derivation details).
Conserv, aPC-Conservation; LD, linkage disequilibrium; MAF, minor allele frequency; SE, standard error; SNV, single nucleotide variant.

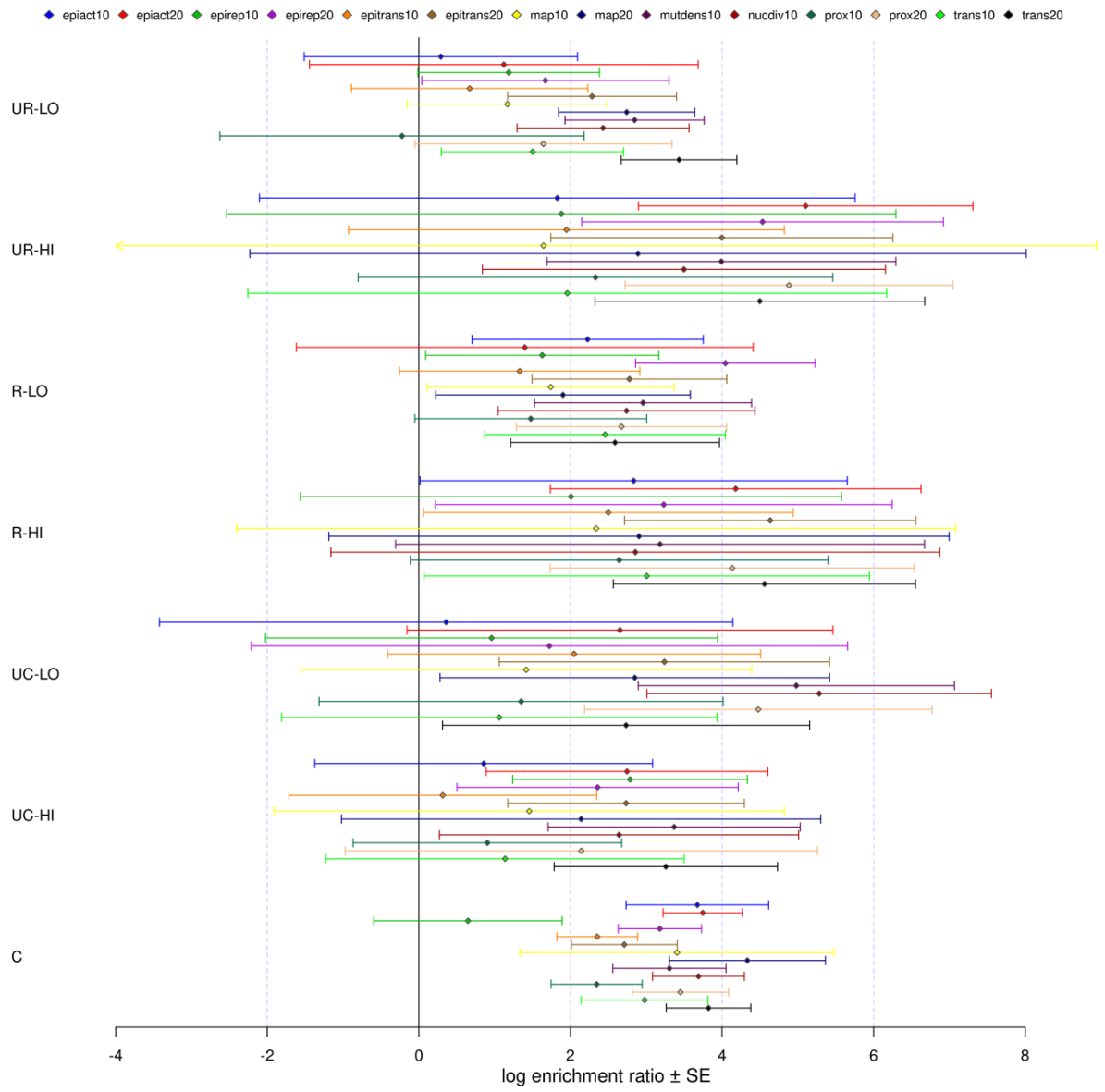


Supplementary Figure 27. Contribution of each LD score-MAF-aPC-Protein-Function bin to the global CAD heritability estimate, along with log enriched protein-function ratio of each LD score-MAF bin. a, Absolute (left) and relative (right) contribution per variant of each LD score-MAF-aPC-Protein-Function bin to the global CAD heritability estimate. Each label in the legend represents a combination of: i) MAF (UR: ultra-rare (MAF $\leq$ 0.1%), R: rare (0.1% < MAF $\leq$ 1%), UC: uncommon (1% < MAF $\leq$ 10%), C: common (10% < MAF $\leq$ 50%)); ii) LD score (LO: low, HI: high); and iii) Protein-Function functionality (High, Low). Error bars represent $\pm$ one SE from each contribution point estimate. Absolute SEs (left) are calculated by GCTA and are proportional to the effective number of independent variants in each bin and inversely proportional to the total sample size (4,949 cases + 17,494 controls). Relative SEs (right) are obtained by dividing the corresponding absolute SEs by the square root of the number of variants. **b**, Log protein function ratio of high over low protein-function functionality variants in each LD score-MAF bin. Each label on the y-axis is defined as in **a**. Error bars represent $\pm$ one SE from each log protein function ratio estimate. SEs are calculated from GCTA's output of the covariance matrix of contribution estimates to heritability in each bin and their corresponding number of SNVs (see section "Mean and variance of log enrichment ratio" for derivation details). ProtFct, aPC-Protein-Function; LD, linkage disequilibrium; MAF, minor allele frequency; SE, standard error; SNV, single nucleotide variant.



Supplementary Figure 28. Proportion of each LD score-MAF-Functionality bin to the global CAD heritability estimate for aPCs at Phred = 10 or 20. Each label in the legend represents a combination of: i) MAF (UR: ultra-rare (MAF ≤ 0.1%), R: rare (0.1% < MAF ≤ 1%), UC: uncommon (1% < MAF ≤ 10%), C: common (10% < MAF ≤ 50%)); ii) LD score (LO: low, HI: high); and iii) Functionality (Low, High).

CAD, coronary artery disease; epiact10, aPC-Epigenetics-Active (Phred = 10); epiact20, aPC-Epigenetics-Active (Phred = 20); epirep10, aPC-Epigenetics-Repressed (Phred = 10); epirep20, aPC-Epigenetics-Repressed (Phred = 20); epitrans10, aPC-Epigenetics-Transcription (Phred = 10); epitrans20, aPC-Epigenetics-Transcription (Phred = 20); Funct, functionality; LD, linkage disequilibrium; MAF, minor allele frequency; map10, aPC-Mappability (Phred = 10); map20, aPC-Mappability (Phred = 20); mutdens10, aPC-Mutation-Density (Phred = 10); nucdiv10, aPC-Local-Nucleotide-Diversity (Phred = 10); prox10, aPC-Proximity-To-TSS-TES (Phred = 10); prox20, aPC-Proximity-To-TSS-TES (Phred = 20); SNV, single nucleotide variant; trans10, aPC-Transcription-Factor (Phred = 10); trans20, aPC-Transcription-Factor (Phred = 20).



Supplementary Figure 29. Log enrichment ratio of high over low functionality variants in each LD score-MAF bin for aPCs at Phred = 10 or 20. Each label on the y-axis represents a combination of: i) MAF (UR: ultra-rare (MAF ≤ 0.1%), R: rare (0.1% < MAF ≤ 1%), UC: uncommon (1% < MAF ≤ 10%), C: common (10% < MAF ≤ 50%)); and ii) LD score (LO: low, HI: high). Error bars show ± one SE from each log enrichment ratio estimate. SEs are calculated from GCTA's output of the covariance matrix of contribution estimates to heritability in each bin and their corresponding number of SNVs (see section "Mean and variance of log enrichment ratio" for derivation details).

epiact10, aPC-Epigenetics-Active (Phred = 10); epiact20, aPC-Epigenetics-Active (Phred = 20); epirep10, aPC-Epigenetics-Repressed (Phred = 10); epirep20, aPC-Epigenetics-Repressed (Phred = 20); epitrans10, aPC-Epigenetics-Transcription (Phred = 10); epitrans20, aPC-Epigenetics-Transcription (Phred = 20); LD, linkage disequilibrium; MAF, minor allele frequency; map10, aPC-Mappability (Phred =

10); map20, aPC-Mappability (Phred = 20); mutdens10, aPC-Mutation-Density (Phred = 10); nucdiv10, aPC-Local-Nucleotide-Diversity (Phred = 10); prox10, aPC-Proximity-To-TSS-TES (Phred = 10); prox20, aPC-Proximity-To-TSS-TES (Phred = 20); trans10, aPC-Transcription-Factor (Phred = 10); trans20, aPC-Transcription-Factor (Phred = 20).

Brief description of TOPMed studies included in this paper

Genetics of Cardiometabolic Health in the Amish (Amish)

The Amish Complex Disease Research Program includes a set of large community-based studies focused largely on cardiometabolic health carried out in the Old Order Amish (OOA) community of Lancaster, Pennsylvania. The OOA population of Lancaster County, PA immigrated to the Colonies from Western Europe in the early 1700's. There are now 43,000 OOA individuals in the Lancaster area, nearly all of whom can trace their ancestry back 12-14 generations to approximately 500 founders. Investigators at the University of Maryland School of Medicine have been studying the genetic determinants of cardiometabolic health in this population since 1993. Subjects included in TOPMed were from the Amish HAPI Heart Study and the Amish Longevity Study. Individuals aged 20 years and older were eligible to participate in the HAPI Heart Study along with their age-eligible family members. Amish Longevity Study subjects included long-lived probands aged 90 years or older, their offspring, and spouses of their offspring.

Atherosclerosis Risk in Communities (ARIC)

The ARIC study is a prospective longitudinal investigation of the development of atherosclerosis and its clinical sequelae in which 15,792 individuals aged 45 to 64 years were enrolled at baseline. At the inception of the study in 1986-1989, the participants were selected by probability sampling from four communities in the United States: Forsyth County, North Carolina; Jackson, Mississippi (African-Americans only); the northwestern suburbs of Minneapolis, Minnesota; and Washington County, Maryland. Four examinations have been carried out at three-year intervals (exam 1, 1987-1989; exam 2, 1990-1992; exam 3, 1993-1995; exam 4, 1996-1998), and subjects are contacted annually to update their medical histories between examinations. A second component of the study involves community surveillance of morbidity and mortality by abstracting hospital records and death certificates and investigating deaths that take place outside of hospitals. The inclusion criteria for selection of ARIC study participants in TOPMed were full consent or consent for cardiovascular disease-specific research, sufficient DNA for sequencing, and unrestricted use of DNA. Individuals were selected from the Venous Thromboembolism project (VTE) and Identification of Common Genetic Variants for Atrial Fibrillation and PR Interval - Atrial Fibrillation Genetics Consortium (AF Gen).

BioMe Biobank at Mount Sinai (BioMe)

The Charles Bronfman Institute of Personalized Medicine Bio*Me* Biobank, founded in September 2007, is an ongoing, broadly-consented electronic health record (EHR)-linked clinical care biobank that enrolls participants non-selectively from the Mount Sinai Medical Center patient population. Bio*Me* currently comprises >50,000 participants from diverse ancestries, characterized by a broad spectrum of longitudinal biomedical traits. Participants enroll through an opt-in process and consent to be followed throughout their clinical care (past, present, and future) in real-time, allowing us to integrate their genomic information with their

EHRs for discovery research and clinical care implementation. BioMe participants consent for recall, based on their genotype and/or phenotype, permitting in-depth follow-up and functional studies for selected participants at any time. Phenotypic and genomic data are stored in a secure database and made available to investigators, contingent on approval by the Bio*Me* Governing Board. For the TOPMed Program, only adult (>18 yrs) participants, who had been diagnosed with chronic obstructive pulmonary disease (COPD) and/or coronary artery disease (CAD) (cases) and/or atrial fibrillation, and who did not have a diagnosis of COPD and CAD (controls), were selected for sequencing.

Coronary Artery Risk Development in Young Adults (CARDIA)

CARDIA is a study examining the etiology and natural history of cardiovascular disease beginning in young adulthood. In 1985-1986, a cohort of 5,115 healthy black and white men and women, aged 18-30 years, were selected to have approximately the same number of people in subgroups of age (18-24 and 25-30), sex, race, and education (high school or less, and more than high school) within each of four US Field Centers. These same participants were asked to participate in follow-up examinations during 1987-1988 (Year 2), 1990-1991 (Year 5), 1992-1993 (Year 7), 1995-1996 (Year 10), 2000-2001 (Year 15), 2005-2006 (Year 20), 2010-2011 (Year 25), 2015-2016 (Year 30), and 2020-2022 (Year 35). In addition to the follow-up examinations, participants are contacted regularly for the ascertainment of information on out-patient procedures and hospitalizations experienced between contacts. In TOPMed, black and white adults from four communities aged 18-30 during the baseline examination in 1985-1986 were included and various clinical data, such as incident CAD, were collected during follow-up examinations.

Cardiovascular Health Study (CHS)

CHS is an NHLBI-funded observational study of risk factors for cardiovascular disease in adults 65 years of age or older conducted across four US field centers. The original cohort of 5,201 persons was recruited in 1989-1990 from random samples of the Medicare eligibility lists. An additional 687 participants, nearly all African Americans, were enrolled in 1992-1993, for a total sample of 5,888. Starting in 1989, and continuing through 1999, participants underwent annual extensive clinical examinations. Follow-up for events remains ongoing through the present. For TOPMed, CHS participants were initially included if they had appropriate consent, available DNA, and met any of the following criteria fully described here: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001368.v3.p2

Genetic Epidemiology of COPD (COPDGene)

The COPDGene study established a racially diverse cohort that is sufficiently large and appropriately designed for genome-wide association analysis of chronic obstructive pulmonary disease (COPD). Study recruitment began in February 2008 at 21 clinical centers throughout the United States. A total of 10,720 subjects were recruited representing the full range of pulmonary function, including non-smoking controls, current smokers, and former smokers. This

baseline cohort is being used for cross-sectional analysis, with long-term longitudinal follow-up visits after five years and after ten years. The primary focus of the study is to identify the genetic risk factors that determine susceptibility for COPD and COPD-related phenotypes. Detailed inclusion/exclusion criteria can be found at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000951.v5.p5

Diabetes Heart Study (DHS)

The Diabetes Heart Study (DHS) is a family-based study enriched for type 2 diabetes (T2D). The cohort included 1,443 European American and African American participants from 564 families with multiple cases of type 2 diabetes. The cohort was recruited between 1998 and 2006. Participants were extensively phenotyped for measures of subclinical cardiovascular disease (CVD) and other known CVD risk factors. Primary outcomes were quantified burden of vascular calcified plaque in the coronary artery, carotid artery, and abdominal aorta all determined from non-contrast computed tomography scans. Singletons with T2D and siblings concordant for T2D with T2D developing after the age of 35 years and treated with insulin and/or oral agents and confirmed by measurement of blood glucose and glycosylated hemoglobin at recruitment were included, while participants with historical evidence of ketoacidosis were excluded. DHS participants with coronary artery calcification (CAC) were selected for whole genome sequencing in TOPMed, prioritizing inclusion of families.

Framingham Heart Study (FHS)

The Framingham Heart Study (FHS) is a prospective longitudinal investigation of the development of atherosclerosis and its clinical sequelae. Study participants were recruited at three time periods. The study was initiated in 1948-50 with the recruitment of 5,209 individuals ages 28-62 (including some spouse pairs, parent-offspring pairs and siblings) for the purpose of investigating the multiple factors involved in the development of cardiovascular disease. This group, known as the Original Cohort, has been examined every two years with a total of thirty-two examinations to date. In 1971-1975, offspring of the Original Cohort and the offspring spouses were recruited to examine among other goals the familial components of cardiovascular disease and its risk factors. In 2002-2005, the third generation (children of the Offspring and grandchildren of the Original Cohort) was recruited. The Offspring Cohort totaled 5,124 and the Third Generation totaled 4,095 at recruitment and have been examined every 4 to 8 years. The Offspring Cohort now has 9 examinations completed and the Third Generation has 2 examinations completed. Additionally, there are two minority cohorts totaling ~900 participants that have been followed since the mid-1990s. DNA samples have been collected and immortalized since the mid-1990s and are available on ~8,000 study participants in 1,037 families. The inclusion criteria in TOPMed for selection of participants (~4,100 subjects) for the FHS study were full informed consent and sufficient DNA for sequencing.

Genetic Study of Atherosclerosis Risk (GeneSTAR)

GeneSTAR began in 1982 as the Johns Hopkins Sibling and Family Heart Study, a prospective longitudinal family-based study conducted originally in healthy adult siblings of people with documented early onset coronary disease under 60 years of age. Commencing in 2003, the siblings, their offspring, and the coparent of the offspring participated in a 2 week trial of aspirin 81 mg/day with pre and post ex vivo platelet function assessed using multiple agonists in whole blood and platelet rich plasma. Extensive additional cardiovascular testing and risk assessment was done at baseline and serially. Follow-up was carried out to determine incident cardiovascular disease, stroke, peripheral arterial disease, diabetes, cancer, and related comorbidities, from 5 to 30 years after study entry. The goal of several additional phenotyping and interventional substudies has been to discover and amplify understanding of the mechanisms of atherogenic vascular diseases and attendant comorbidities. Detailed inclusion/exclusion criteria can be found at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001218.v3.p1

Genetic Epidemiology Network of Arteriopathy (GENOA)

GENOA is one of four networks in the NHLBI Family-Blood Pressure Program (FBPP). GENOA's long-term objective is to elucidate the genetics of target organ complications of hypertension, including both atherosclerotic and arteriolosclerotic complications involving the heart, brain, kidneys, and peripheral arteries. The longitudinal GENOA Study recruited European-American and African-American sibships with at least 2 individuals with clinically diagnosed essential hypertension before age 60 years. All other members of the sibship were invited to participate regardless of their hypertension status. Participants were diagnosed with hypertension if they had either 1) a previous clinical diagnosis of hypertension by a physician with current anti-hypertensive treatment, or 2) an average systolic blood pressure = 140 mm Hg or diastolic blood pressure = 90 mm Hg based on the second and third readings at the time of their clinic visit. Only participants of the African-American Cohort were sequenced through TOPMed. More detailed inclusion/exclusion criteria in the GENOA TOPMed substudy can be found at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001345.v3.p1

Jackson Heart Study (JHS)

JHS is a large, community-based, observational study whose 5,306 participants were recruited from among the non-institutionalized African-American adults from urban and rural areas of the three counties (Hinds, Madison, and Rankin) that make up the Jackson, Mississippi metropolitan statistical area (MSA). JHS participants were chosen randomly from the Accudata America commercial listing, which provides householder name, address, zip code, phone number (if available), age group in decades, and family components. A structured volunteer sample was also included in which demographic cells for recruitment were designed to mirror the eligible population. Enrollment was opened to volunteers who met census-derived age, sex, and socioeconomic status (SES) eligibility criteria for the Jackson MSA. In addition, a family component was included in the JHS. The sampling frame for the family study was a participant in any one of the ARIC, random, or volunteer samples whose family size met eligibility requirements. Eligibility included having at least two full siblings and four first degree relatives

(parents, siblings, children over the age of 21) who lived in the Jackson MSA and who were willing to participate in the study. Recruitment was limited to persons 35-84 years old except in the family cohort, where those 21 years old and above were eligible. DNA samples of all JHS participants who provided consent that allows sharing of data through dbGaP, and who had adequate DNA samples available, were included in the TOPMed project.

Multi-Ethnic Study of Atherosclerosis (MESA)

MESA is a study of the characteristics of subclinical cardiovascular disease (CVD) and the risk factors that predict progression to clinically overt CVD or progression of the subclinical disease. MESA researchers study a diverse, population-based sample of 6,814 asymptomatic men and women aged 45-84. Participants were recruited from six field centers across the United States: Wake Forest University, Columbia University, Johns Hopkins University, University of Minnesota, Northwestern University and University of California - Los Angeles. They are being followed for identification and characterization of CVD events, including acute myocardial infarction and other forms of coronary heart disease (CHD), stroke, and congestive heart failure; for CVD interventions; and for mortality. The first examination took place over two years, from July 2000 - July 2002. It was followed by five examination periods that were 17-20 months in length. Participants have been contacted every 9 to 12 months throughout the study to assess clinical morbidity and mortality. The MESA Family study also provided DNA samples for TOPMed via the study/designation of "AA CAC" (African-American Coronary Artery Calcium consortium study). The general goal of the MESA Family study is to locate and identify genes contributing to the genetic risk for CVD, by looking at the early changes of atherosclerosis within families (mainly siblings). 2,128 individuals from 594 families, yielding 3,026 sib-pairs divided between African-Americans and Hispanic-Americans, were recruited by utilizing the existing framework of MESA. A detailed MESA selection algorithm for TOPMed is described at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001416.v3.p1

Women's Health Initiative (WHI)

WHI is a long-term national health study that has focused on strategies for preventing heart disease, breast and colorectal cancer, and osteoporotic fractures in postmenopausal women. The original WHI study included 161,808 postmenopausal women enrolled between 1993 and 1998. All women enrolled in the WHI were between 50 and 79 years old and were postmenopausal at the time of enrollment. In addition, eligibility criteria for the clinical trial (CT) and observational study (OS) included ability and willingness to provide written informed consent and an intention to reside in the area for at least 3 years after enrollment. In TOPMed, ~11,100 women were sequenced: approximately 1,100 cases of venous thromboembolism (VTE), 4,000 cases of ischemic stroke, 900 cases of hemorrhagic stroke, and 5,100 controls. The inclusion criteria for cases were consent status allowing for data sharing through dbGaP, and incidence case of stroke or VTE after enrollment in WHI. Inclusion criteria for controls were consent status allowing for data sharing through dbGaP, and no history of stroke or VTE.

TOPMed study-specific acknowledgements

Genetics of Cardiometabolic Health in the Amish (Amish)

The Amish studies upon which these data are based were supported by NIH grants R01 AG18728, U01 HL072515, R01 HL088119, R01 HL121007, and P30 DK072488. See publication: PMID: 18440328. The TOPMed component of the Amish Research Program was supported by NIH grants R01HL121007, U01HL072515 and R01AG18728.

Atherosclerosis Risk in Communities (ARIC)

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions.

BioMe Biobank at Mount Sinai (BioMe)

The Mount Sinai BioMe Biobank has been supported by The Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI (U01HG00638001; U01HG007417; X01HL134588). We thank all participants in the Mount Sinai Biobank. We also thank all our recruiters who have assisted and continue to assist in data collection and management and are grateful for the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Coronary Artery Risk Development in Young Adults (CARDIA)

The Coronary Artery Risk Development in Young Adults Study (CARDIA) is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the University of Alabama at Birmingham (HHSN268201800005I & HHSN268201800007I), Northwestern University (HHSN268201800003I), University of Minnesota (HHSN268201800006I), and Kaiser Foundation Research Institute (HHSN268201800004I). CARDIA was also partially supported by the Intramural Research Program of the National Institute on Aging (NIA) and an intra-agency agreement between NIA and NHLBI (AG0005). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The Coronary Artery Risk Development in Young Adults Study (CARDIA) is supported by contracts HHSN268201300025C, HHSN268201300026C, HHSN268201300027C,

HHSN268201300028C, HHSN268201300029C, and HHSN268200900041C from the National Heart, Lung, and Blood Institute (NHLBI).

Cardiovascular Health Study (CHS)

Cardiovascular Health Study: This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006 and grants R01HL105756, U01HL080295 and U01HL130114 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Genetic Epidemiology of COPD (COPDGene)

The COPDGene project described was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. COPDGene is also supported by the COPD Foundation through contributions made to an Industry Advisory Board that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer, and Sunovion. A full listing of COPDGene investigators can be found at: http://www.copdgene.org/directory

Diabetes Heart Study (DHS)

Diabetes Heart Study: The investigators acknowledge the cooperation of our Diabetes Heart Study (DHS) and AA-DHS participants. This work was supported by R01 HL92301, R01 HL67348, R01 NS058700, R01 AR48797, R01 DK071891, R01 AG058921, the General Clinical Research Center of the Wake Forest University School of Medicine (M01 RR07122, F32 HL085989), the American Diabetes Association, and a pilot grant from the Claude Pepper Older Americans Independence Center of Wake Forest University Health Sciences (P60 AG10484).

Framingham Heart Study (FHS)

The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195, HHSN268201500001I and 75N92019D00031 from the National Heart, Lung and Blood Institute and grant supplement R01 HL092577-06S1 for this research. We also acknowledge the dedication of the FHS study participants without whom this research would not be possible. Dr. Vasan is supported in part by the Evans Medical Foundation and the Jay and Louis Coffman Endowment from the Department of Medicine, Boston University School of Medicine.

Genetic Study of Atherosclerosis Risk (GeneSTAR)

GeneSTAR was supported by grants from the National Institutes of Health/National Heart, Lung, and Blood Institute (HL112064, U01 HL72518, HL087698, HL49762, HL59684, HL58625, HL071025, HL092165, HL099747, K23HL105897), National Institutes of Health/National Institute of Nursing Research (NR0224103, NR008153), National Institutes of Health/National Institute of Neurological Disorders and Stroke (NS062059), by a grant from the National Institutes of Health/National Center for Research Resources (M01-RR000052) to the Johns Hopkins General Clinical Research Center, and by grants from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health to the Johns Hopkins Institute for Clinical & Translational Research (UL1 RR 025005, UL1TR001079).

Genetic Epidemiology Network of Arteriopathy (GENOA)

Support for GENOA was provided by the National Heart, Lung and Blood Institute (HL054457, HL054464, HL054481, HL119443, and HL087660) of the National Institutes of Health. We would like the thank the Mayo Clinic Genotyping Core, the DNA Sequencing and Gene Analysis Center at the University of Washington, and the Broad Institute for their genotyping and sequencing services. We would also like to thank the GENOA participants. Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for the TOPMed GENOA Study was performed at the Northwest Genomics Center at the University of Washington for the HyperGen/GENOA project (PI: Donna Arnett) and the Broad Institute of MIT and Harvard for the AACAC project (PI: Kent Taylor). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

Jackson Heart Study (JHS)

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I, and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staff and participants of the JHS.

Multi-Ethnic Study of Atherosclerosis (MESA)

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)" (phs001416.v3.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1), and TOPMed MESA Multi-Omics (HHSN2682015000031/HSN26800004). The MESA projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for the Multi-Ethnic Study of Atherosclerosis (MESA) projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1TR001881, DK063491, and R01HL105756. The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutes can be found at http://www.mesa-nhlbi.org. This study was also supported in part by the NHLBI contracts R01HL151855, R01HL146860, and R01HL163262.

Women's Health Initiative (WHI)

The Women's Health Initiative (WHI) program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005. This manuscript was prepared in collaboration with investigators of the WHI, and has been reviewed and/or approved by the Women's Health Initiative (WHI). WHI investigators are listed at https://www-whi-org.s3.us-west-2.amazonaws.com/wp-content/uploads/WHI-Investigator-Short-List.pdf.

NHLBI TOPMed Consortium banner authors

Name	Institution
Anugu, Pramod	University of Mississippi
Auer, Paul	Medical College of Wisconsin
Barwick, Lucas	The Emmes Corporation
Becker, Diane	Johns Hopkins University
Becker, Lewis	Johns Hopkins University
Bis, Joshua	University of Washington
Carty, Cara	Washington State University
Castaldi, Peter	Brigham & Women's Hospital
Chaffin, Mark	Broad Institute
Chang, Yi-Cheng	National Taiwan University
Choi, Seung Hoan	Broad Institute
Chuang, Lee-Ming	National Taiwan University
Chung, Ren-Hua	National Health Research Institute Taiwan
Crandall, Carolyn	University of California, Los Angeles
David, Sean	University of Chicago
de las Fuentes, Lisa	Washington University in St Louis
Deka, Ranjan	University of Cincinnati
DeMeo, Dawn	Brigham & Women's Hospital
Duan, Qing	University of North Carolina
Eaton, Charles	Brown University
Ekunwe, Lynette	University of Mississippi
El Boueiz, Adel	Harvard University
Franceschini, Nora	University of North Carolina
Gao, Shanshan	University of Colorado at Denver
Gao, Yan	University of Mississippi
Gass, Margery	Fred Hutchinson Cancer Research Center
Ghosh, Auyon	Brigham & Women's Hospital
Grine, Daniel	University of Colorado at Denver
Hall, Michael	University of Mississippi
Hersh, Craig	Brigham & Women's Hospital
Hobbs, Brian	Brigham & Women's Hospital
Hsiung, Chao (Agnes)	National Health Research Institute Taiwan
Hung, Yi-Jen	Tri-Service General Hospital National Defense Medical Center
Huston, Haley	Blood Works Northwest
Hwu, Chii Min	Taichung Veterans General Hospital Taiwan
Jackson, Rebecca	Oklahoma State University Medical Center
Johnsen, Jill	University of Washington
Lange, Christoph	Harvard School of Public Health
Lange, Ethan	University of Colorado at Denver
LeBoff, Meryl	Brigham & Women's Hospital
Lee, Wen-Jane	Taichung Veterans General Hospital Taiwan

Li, Yun	University of North Carolina
Liu, Simin	Brown University
Liu, Yu	Stanford University
Manson, JoAnn	Brigham & Women's Hospital
Martin, Lisa	George Washington University
Mathai, Susan	University of Colorado at Denver
Mei, Hao	University of Mississippi
Naik, Rakhi	Johns Hopkins University
Naseri, Take	Ministry of Health, Government of Samoa
Neltner, Bonnie	University of Colorado at Denver
Ochs-Balcom, Heather	University at Buffalo
Paik, David T.	Stanford University
Parker, Cora	RTI International
Perez, Marco	Stanford University
Peters, Ulrike	Fred Hutchinson Cancer Research Center
Phillips, Lawrence S	Emory University
Powers Becker, Julia	University of Colorado at Denver
Psaty, Bruce	University of Washington
Reupena, Muagututi'a Sefuiva	Lutia I Puava Ae Mapu I Fagalele
Roselli, Carolina	Broad Institute
Russell, Pamela	University of Colorado at Denver
Sabino, Ester Cerdeira	Universidade de Sao Paulo
Sandow, Kevin	Lundquist Institute
Schwander, Karen	Washington University in St Louis
Sciurba, Frank	University of Pittsburgh
Silver, Brian	UMass Memorial Medical Center
Smoller, Sylvia	Albert Einstein College of Medicine
Snively, Beverly	Wake Forest Baptist Health
Storm, Garrett	University of Colorado at Denver
Sung, Yun Ju	Washington University in St Louis
Tang, Hua	Stanford University
Taub, Margaret	Johns Hopkins University
Tinker, Lesley	Fred Hutchinson Cancer Research Center
Tirschwell, David	University of Washington
Tiwari, Hemant	University of Alabama
Vaidya, Dhananjay	Johns Hopkins University
Walker, Tarik	University of Colorado at Denver
Wallace, Robert	University of Iowa
Walts, Avram	University of Colorado at Denver
Weng, Lu-Chen	Massachusetts General Hospital
Yang, Ivana	University of Colorado at Denver
Zhao, Snow Xueyan	National Jewish Health

TOPMed Atherosclerosis Working Group members

Name	Institution
Assimes, Tim	Stanford University
Beame, David	University of Washington
Becker, Diane	Johns Hopkins University
Becker, Lewis	Johns Hopkins University
Bielak, Larry	University of Michigan
Bierig, Trevor	Harvard School of Public Health
Bis, Joshua	University of Washington
Bowers, Michael	University of Washington
Cade, Brian	Brigham & Women's Hospital
Carty, Cara	Washington State University
Chan, Kei Hang Katie	Brown University
Civelek, Mete	University of Virginia
Clarke, Shoa	Stanford University
Conomos, Matthew	University of Washington
Cosentino, Rhea	University of Maryland
Damrauer, Scott	University of Pennsylvania
de Vries, Paul	University of Texas Health at Houston
Divers, Jasmin	Wake Forest Baptist Health
Do, Ron	Icahn School of Medicine at Mount Sinai
Floyd, James	University of Washington
Franceschini, Nora	University of North Carolina
Gagliano Taliun, Sarah	University of Montreal
Hajek, Cassie	Lundquist Institute
Hamik, Anne	Stony Brook Medicine
Hasbani, Natalie	University of Texas Health at Houston
Herrington, David	Wake Forest Baptist Health
Hixson, James	University of Texas Health at Houston
Hou, Lifang	Northwestern University
Jain, Deepti	University of Washington
Kathiresan, Sekar	Broad Institute
Keely, Addison	University of Washington
Keramati, Ali	Johns Hopkins University
Khera, Amit	Broad Institute
Kizer, Jorge	University of California, San Francisco
Klarin, Derek	Broad Institute
Kooperberg, Charles	Fred Hutchinson Cancer Research Center
Koyama, Satoshi	Broad Institute
Kral, Brian	Johns Hopkins University
Laurie, Cecelia	University of Washington
Lebo, Matthew	Brigham & Women's Hospital

Lin, Xihong	Harvard School of Public Health
Liu, Qing	Brown University
Liu, Simin	Brown University
Liu, Yongmei	Duke University
Lu, Yingchang	Vanderbilt University
Lutz, Sharon	Harvard Medical School, Harvard Pilgrim Health Care
Malhotra, Rajeev	Harvard Medical School
Manson, JoAnn	Brigham & Women's Hospital
Martin, Lisa	George Washington University
McHugh, Caitlin	University of Washington
Mikulla, Julie	National Heart, Lung, and Blood Institute, National Institutes of Health
Miller, Clint	University of Virginia
Mitchell, Braxton D.	University of Maryland
Mitchell, Gary F.	Cardiovascular Engineering Inc
Musani, Solomon	University of Mississippi
O'Donnell, Christopher	National Heart, Lung, and Blood Institute, National Institutes of Health
Peyser, Patricia	University of Michigan
Pontes, Rosalice P.	Massachusetts General Hospital
Post, Wendy	Johns Hopkins University
Psaty, Bruce	University of Washington
Purnell, Jennifer Anne	University of Washington
Qi, Qibin	Albert Einstein College of Medicine
Ramachandran, Vasan	Boston University
Regan, Elizabeth	National Jewish Health
Reiner, Alex	Fred Hutchinson Cancer Research Center
Rocheleau, Ghislain	Icahn School of Medicine at Mount Sinai
Rotter, Jerome	Lundquist Institute
Ryan, Kathleen	University of Maryland
Salimi, Shabnam	University of Maryland
Shabani, Mahsima	Johns Hopkins University
Sheu, Wayne Hui-Heng	Taichung Veterans General Hospital Taiwan
Stilp, Adrienne M.	University of Washington
Sullivan, Randi	Wake Forest Baptist Health
Taylor, Kent D.	Lundquist Institute
Terry, James Gregory	Vanderbilt University
Thorington, Daune	Lundquist Institute
Vargas, Jose	Johns Hopkins University
Verbanck, Marie	Icahn School of Medicine at Mount Sinai
Wehr, Kate	University of Washington
Wong, Quenna	University of Washington
Xu, Huichun	University of Maryland

Yanek, Lisa	Johns Hopkins University
Young, Kendra	University of Colorado at Denver

Supplementary references

1. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genet. Epidemiol.* **39**, 276–293 (2015).
2. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free Estimation of Recent Genetic Relatedness. *The American Journal of Human Genetics* **98**, 127–148 (2016).
3. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
4. Privé, F., Luu, K., Blum, M. G. B., McGrath, J. J. & Vilhjálmsson, B. J. Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* **36**, 4449–4457 (2020).
5. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
6. Khan, A. T. *et al.* Recommendations on the use and reporting of race, ethnicity, and ancestry in genetic research: Experiences from the NHLBI TOPMed program. *Cell Genomics* **2**, 100155 (2022).
7. Kittles, R. A. *et al.* Dual Origins of Finns Revealed by Y Chromosome Haplotype Variation. *The American Journal of Human Genetics* **62**, 1171–1179 (1998).
8. Tishkoff, S. A. *et al.* The Genetic Structure and History of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
9. Gouveia, M. H. *et al.* Origins, Admixture Dynamics, and Homogenization of the African Gene Pool in the Americas. *Molecular Biology and Evolution* **37**, 1647–1656 (2020).
10. Gouveia, M. H. *et al.* Unappreciated subcontinental admixture in Europeans and European Americans and implications for genetic epidemiology studies. *Nat Commun* **14**, 6802 (2023).
11. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating Missing Heritability for Disease from Genome-wide Association Studies. *The American Journal of Human Genetics* **88**, 294–305 (2011).
12. Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet* **54**, 263–273 (2022).

13. Ochoa, A. & Storey, J. D. Estimating FST and kinship for arbitrary population structures. *PLoS Genet* **17**, e1009241 (2021).
14. Jiang, W., Zhang, X., Li, S., Song, S. & Zhao, H. An unbiased kinship estimation method for genetic data analysis. *BMC Bioinformatics* **23**, 525 (2022).
15. Tcheandjieu, C. *et al.* Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat Med* **28**, 1679–1692 (2022).
16. Turner, A. W. *et al.* Single-nucleus chromatin accessibility profiling highlights regulatory mechanisms of coronary artery disease risk. *Nat Genet* **54**, 804–816 (2022).