The American Journal of Human Genetics

Supplemental Data

Improved Ancestry Estimation for both Genotyping and Sequencing Data using Projection Procrustes Analysis and Genotype Imputation

Chaolong Wang, Xiaowei Zhan, Liming Liang, Gonçalo R Abecasis, and Xihong Lin

Supplementary Figures

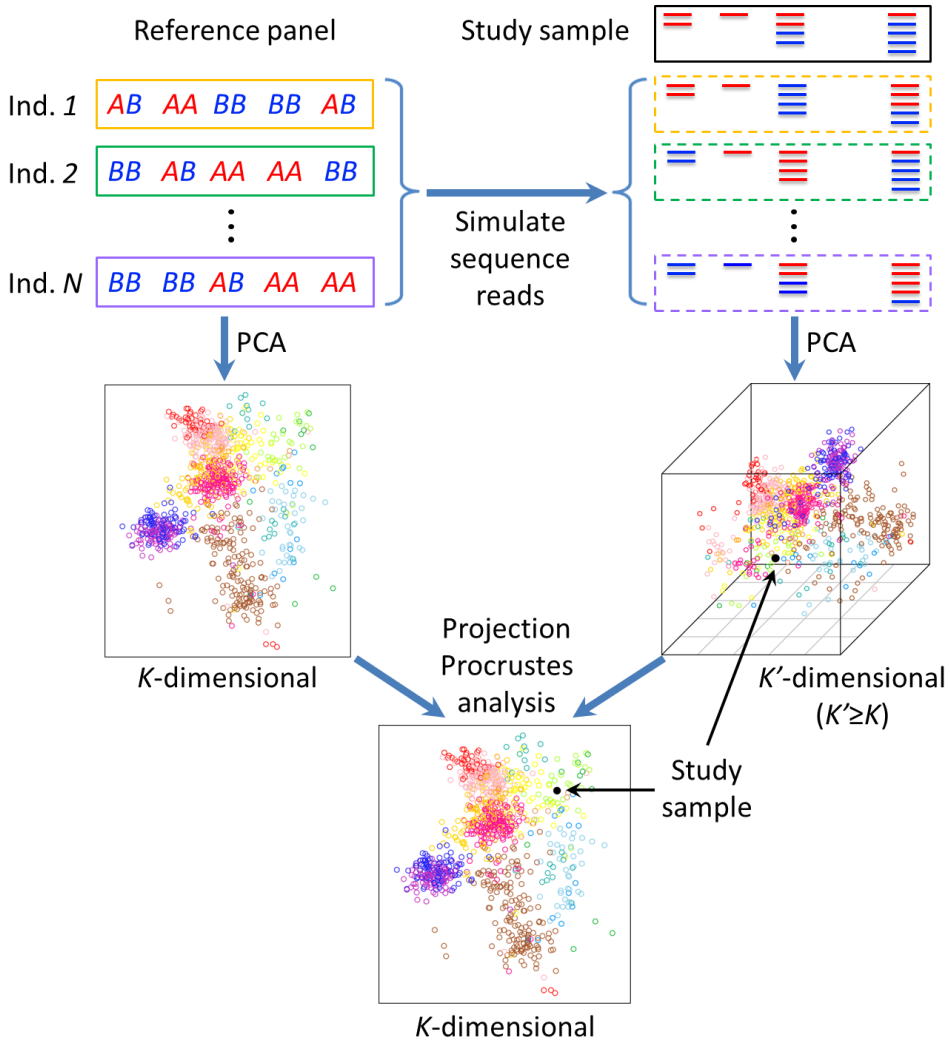

Figure S1. Graphical illustration of the algorithm for estimating an individual's genetic ancestry using sequence reads. This algorithm was implemented in the *laser* program as part of the LASER 2.0 package.

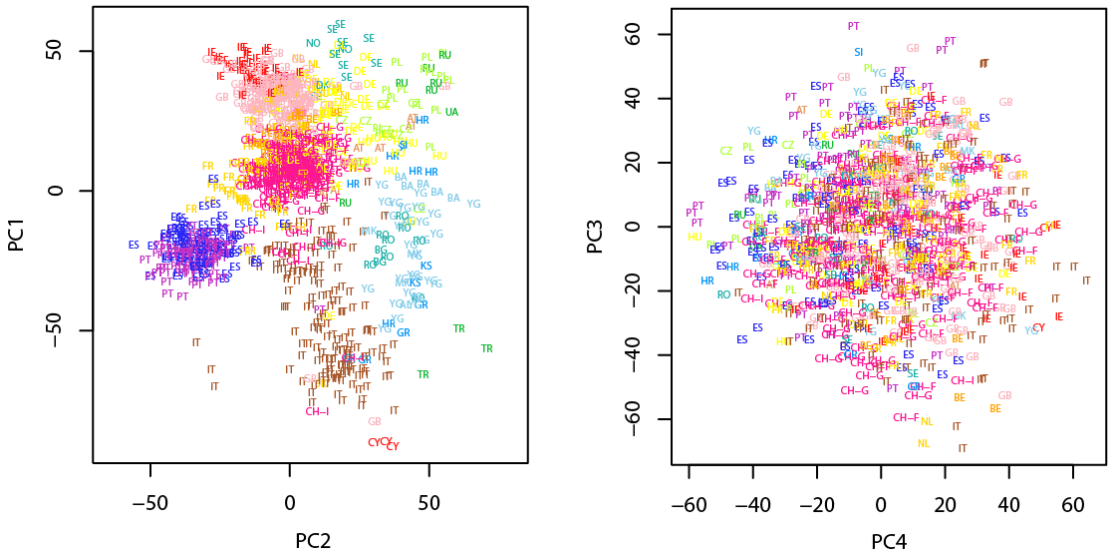

Figure S2. PCA on genotypes across 318,682 SNPs the original POPRES data. Analyses were based on 1,000 POPRES individuals who were randomly selected as the reference panel. Colors and labels follow **Figure 2**.

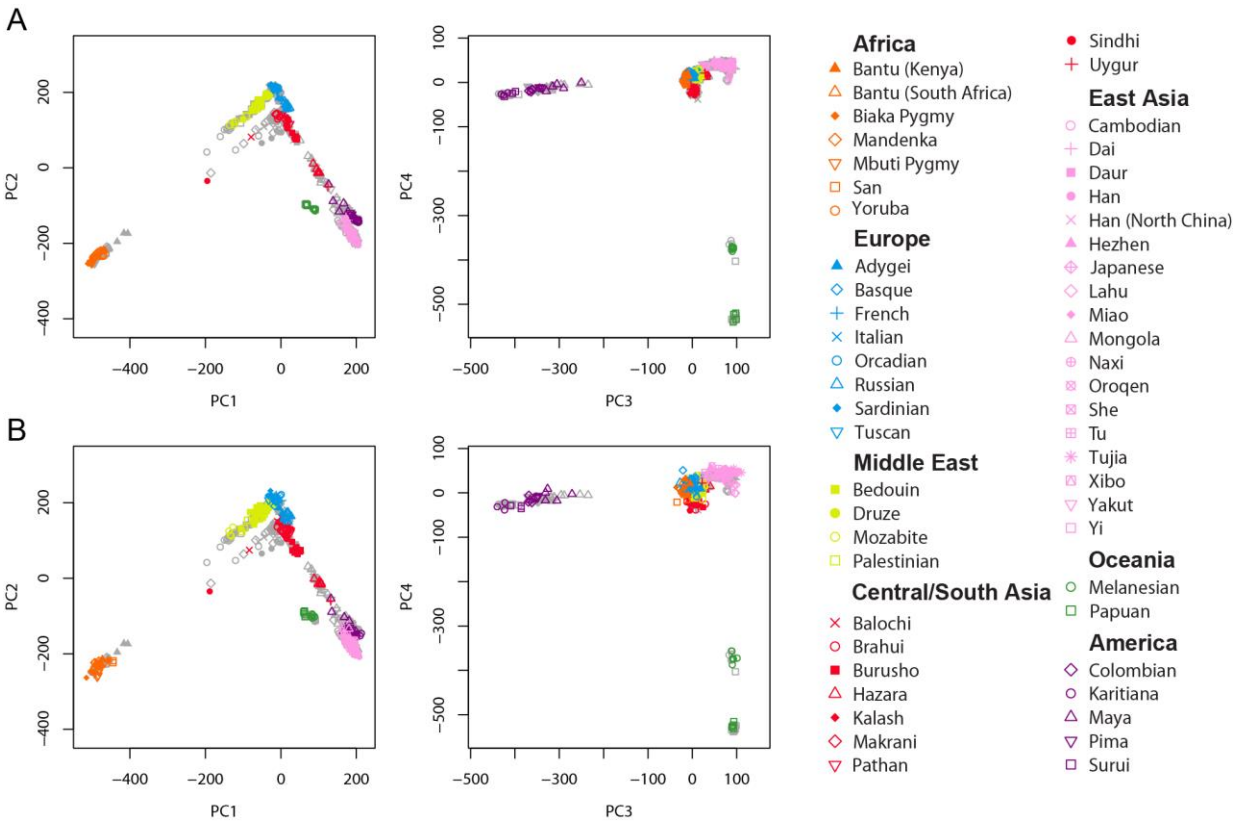

Figure S3. Estimation of worldwide ancestry using exomechip genotypes. Grey points represent 700 reference individuals randomly selected from the HGDP data. Colored points represent the remaining 238 HGDP individuals who were used as the test set. (A) Ancestry estimates based on all 632,958 SNPs in the HGDP data ($K = K' = 4$). (B) Ancestry estimates based on 12,580 SNPs shared by the exomechip and the HGDP data ($K = K' = 4$). The Procrustes similarity of a four-dimensional comparison between A and B is $t_0 = 0.9985$.

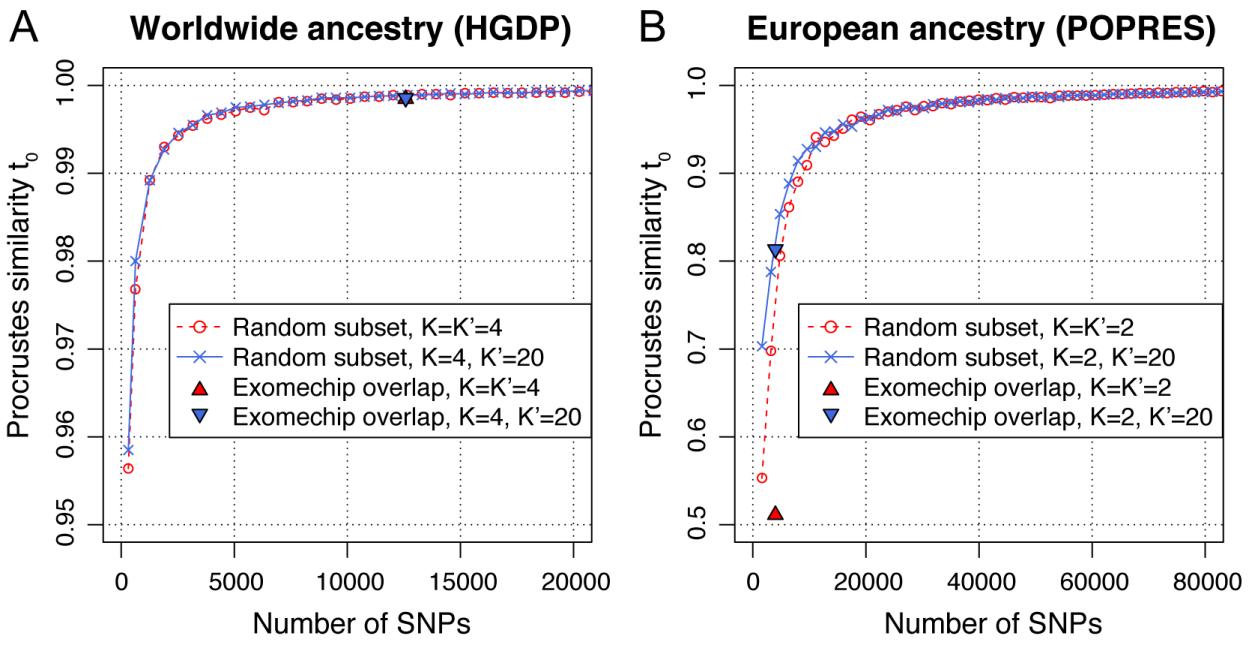

Figure S4. Estimation of worldwide continental ancestry and fine-scale European ancestry using genotypes across random subsets of SNPs in the HGDP data set and the POPRES data set, respectively. The x-axis indicates the number of genotyped SNPs in each random subset. The y-axis indicates the Procrustes similarity score when comparing estimates based on random subsets of SNPs to coordinates estimated based on all genotyped SNPs (*i.e.*, **Figure S3A** for the HGDP comparisons and **Figure 2A** for the POPRES comparisons). (A) Performance of *trace* in estimating worldwide continental ancestry based on the HGDP data set (238 test individuals and 700 reference individuals). The number of SNPs overlapped between the exomechip and the HGDP data is 12,580. (B) Performance of *trace* in estimating European ancestry based on the POPRES data set (385 test individuals and 1000 reference individuals). The number of SNPs overlapped between the exomechip and the POPRES data is 3,983.

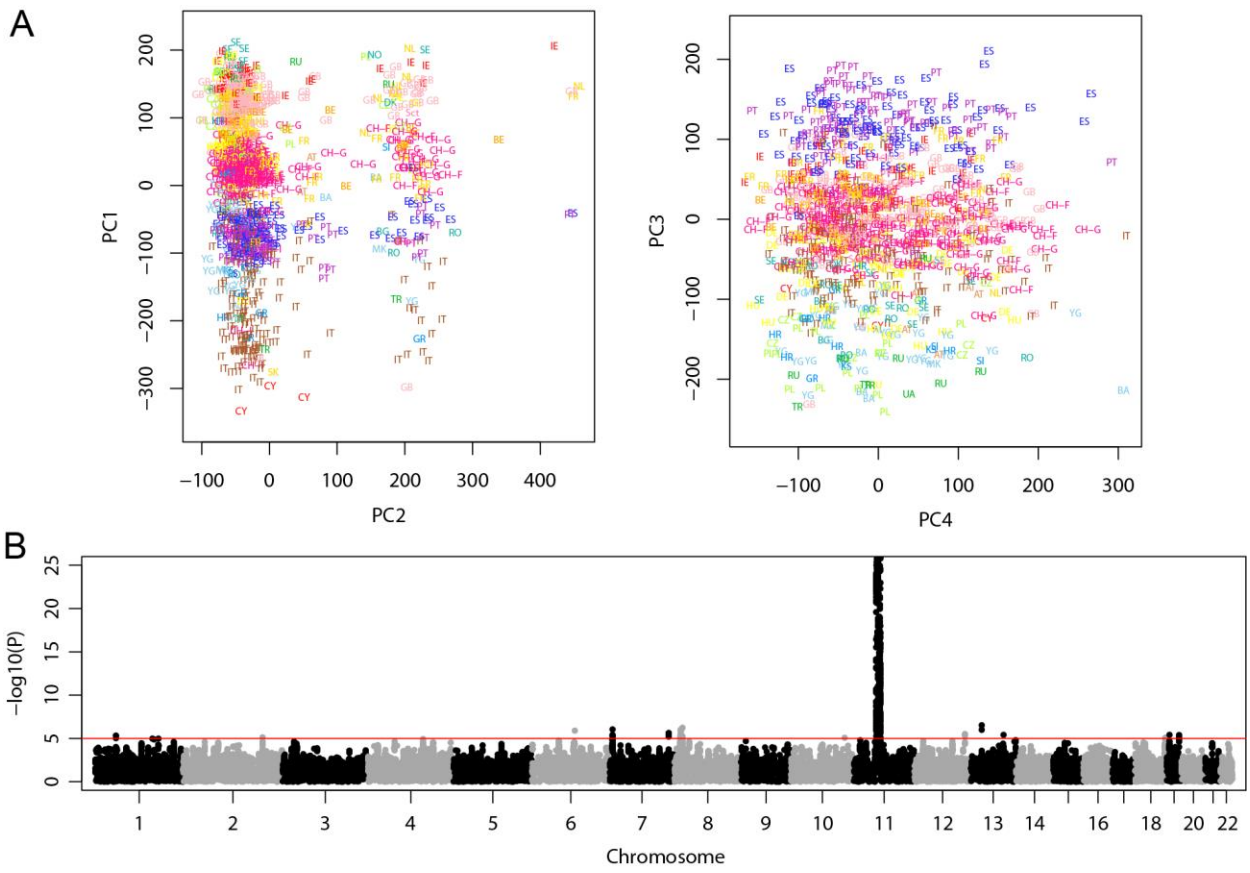

Figure S5. Genotype imputation introduces unexpected patterns in PCA. (A) PCA on the imputed POPRES reference panel (1000 individuals and 4,172,127 SNPs). Colors and labels follow **Figure 2**. (B) P-values for single SNP association tests with PC2. We find that the unexpected patterns along PC2 are driven by a cluster of 9,178 SNPs ($P < 10^{-5}$) around the centromere of chromosome 11(between positions 46,743,232 and 57,208,915). Most of these SNPs (8,936 out of 9,178) are imputed SNPs.

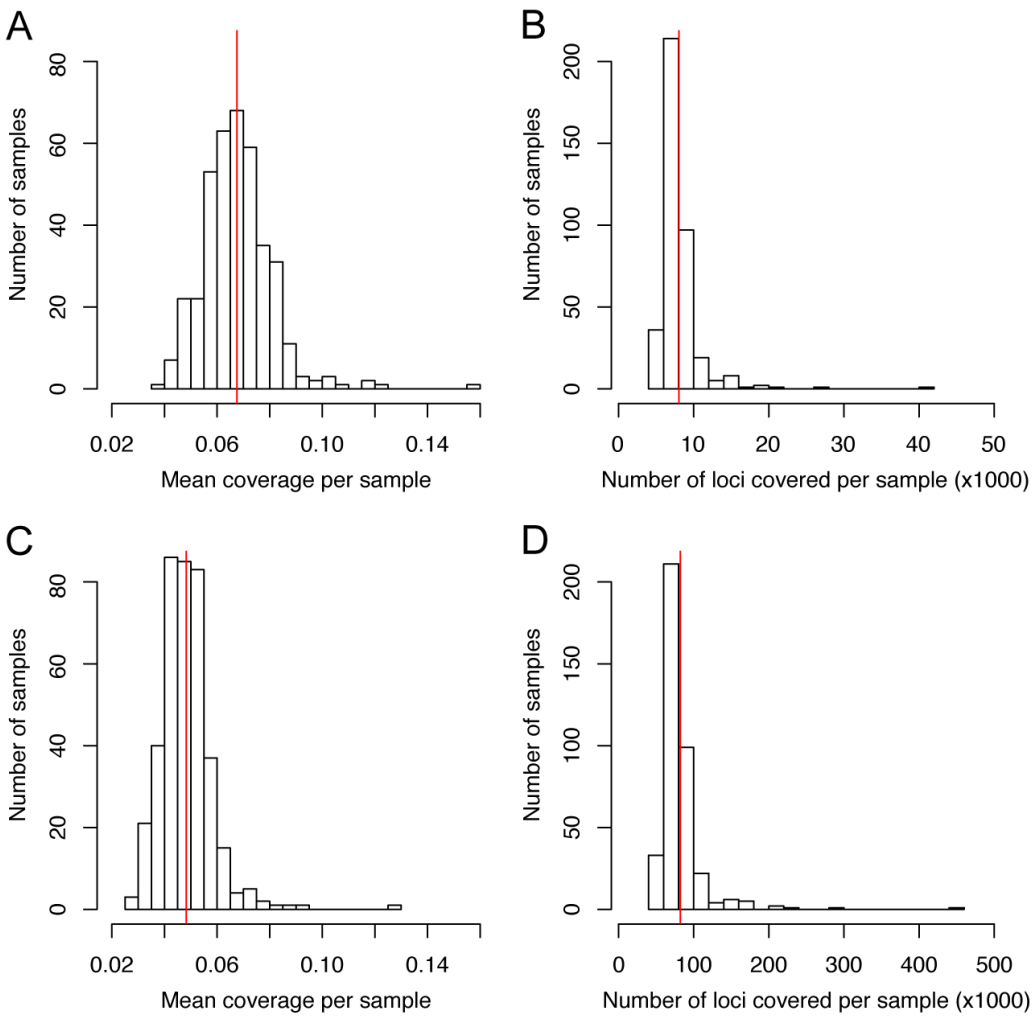

Figure S6. Coverage distribution for the simulated off-target sequence data. The red line in each plot indicates the mean value of the distribution. (A) Coverage per sample averaged across 318,682 SNPs in the original POPRES data. The mean coverage is 0.068X. (B) Number of loci covered by > 0 reads among 318,682 SNPs in the original POPRES data. The average number is 8,040 SNPs per sample. (C) Coverage per sample averaged across 4,172,127 SNPs in the imputed POPRES data. The mean coverage is 0.048X. (D) Number of loci covered by >0 reads among 4,172,127 SNPs in the imputed POPRES data. The average number is 82,152 SNPs per sample.

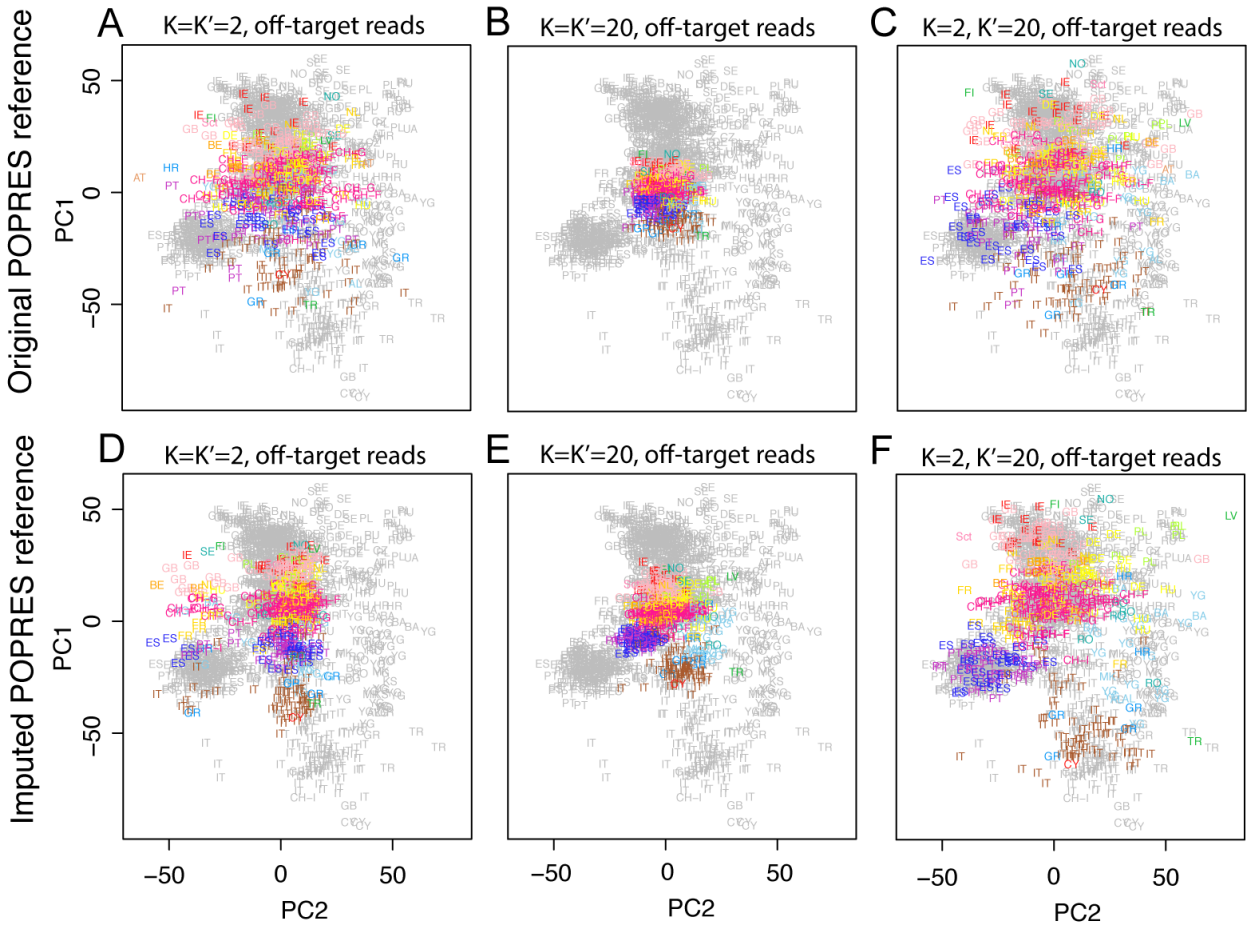

Figure S7. Estimation of European ancestry using low-coverage sequence reads. Grey symbols represent 1,000 reference individuals randomly selected from the POPRES data. Colored symbols represent the remaining 385 POPRES individuals, whose sequence reads were simulated with coverage set to ~5% of the off-target coverage produced in exome sequencing experiments (**Figure S6**). Colors and labels follow **Figure 2**. The Procrustes similarity score t_0 was calculated by comparing top 2 PCs in each panel to the estimates based on genome-wide SNPs shown in **Figure 2A**. (A-C) Estimates using the original POPRES reference panel. (A) $K = K' = 2$, $t_0 = 0.6565$. (B) $K = K' = 20$, $t_0 = 0.7677$. (C) $K = K' = 20$, $t_0 = 0.7841$. (D-F) Estimates using the imputed POPRES reference panel. (D) $K = K' = 2$, $t_0 = 0.6412$. (E) $K = K' = 20$, $t_0 = 0.9375$. (F) $K = 2$, $K' = 20$, $t_0 = 0.9388$.

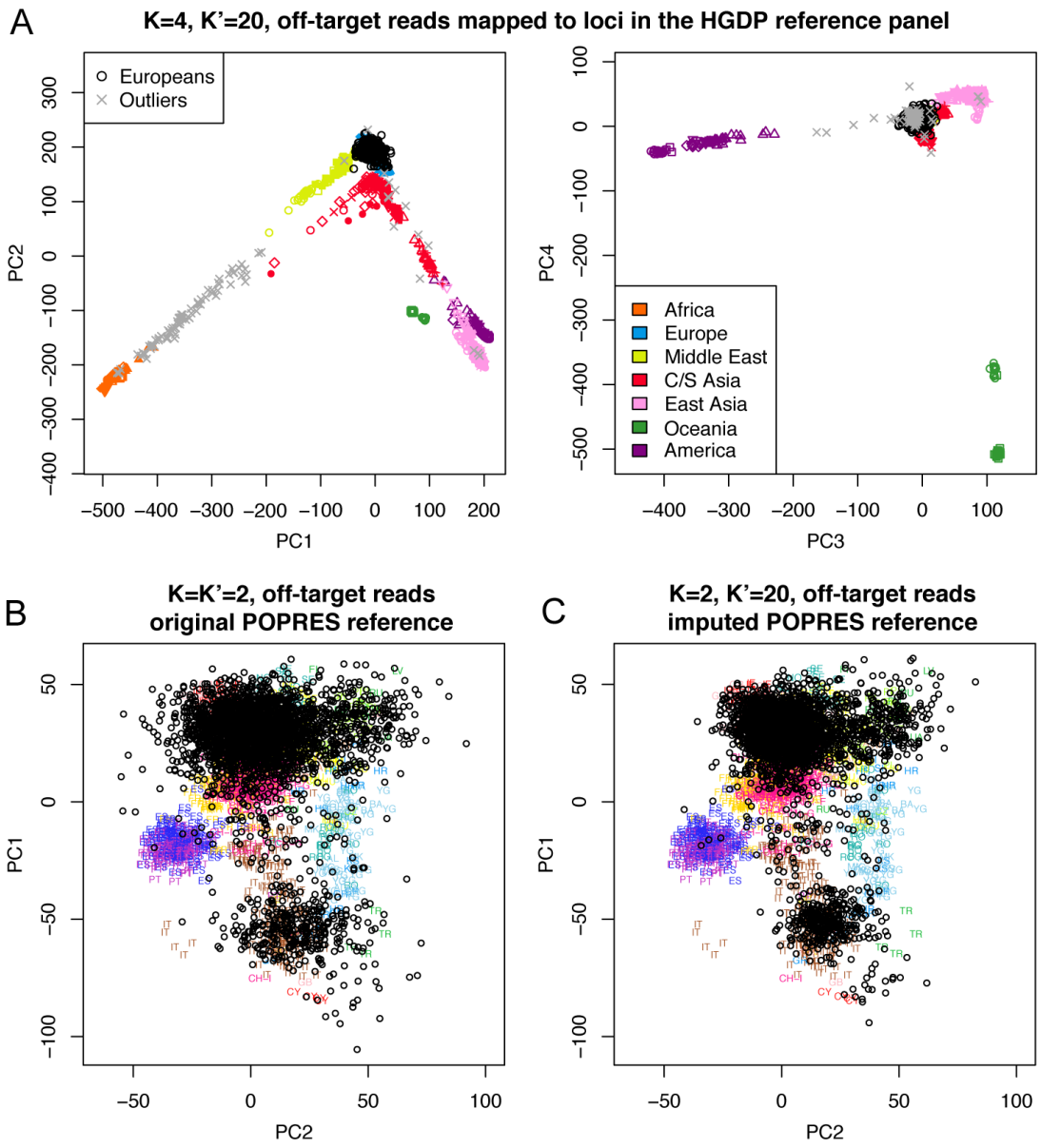

Figure S8. Ancestry analyses of the AMD targeted sequencing data. (A) Worldwide ancestry using the HGDP reference panel. HGDP reference individuals are represented by colored symbols, following the legend of **Figure S3**. We identified 3068 samples with European ancestry (black circles) and 90 non-European outliers (grey crosses). (B) Fine-scale ancestry of 3068 European samples using the original POPRES reference panel with $K = K' = 2$. (C) Fine-scale ancestry of 3068 European samples (black circles) using the imputed POPRES reference panel with $K = 2$ and $K' = 20$. The POPRES reference individuals are represented by colored symbols following **Figure 2.**

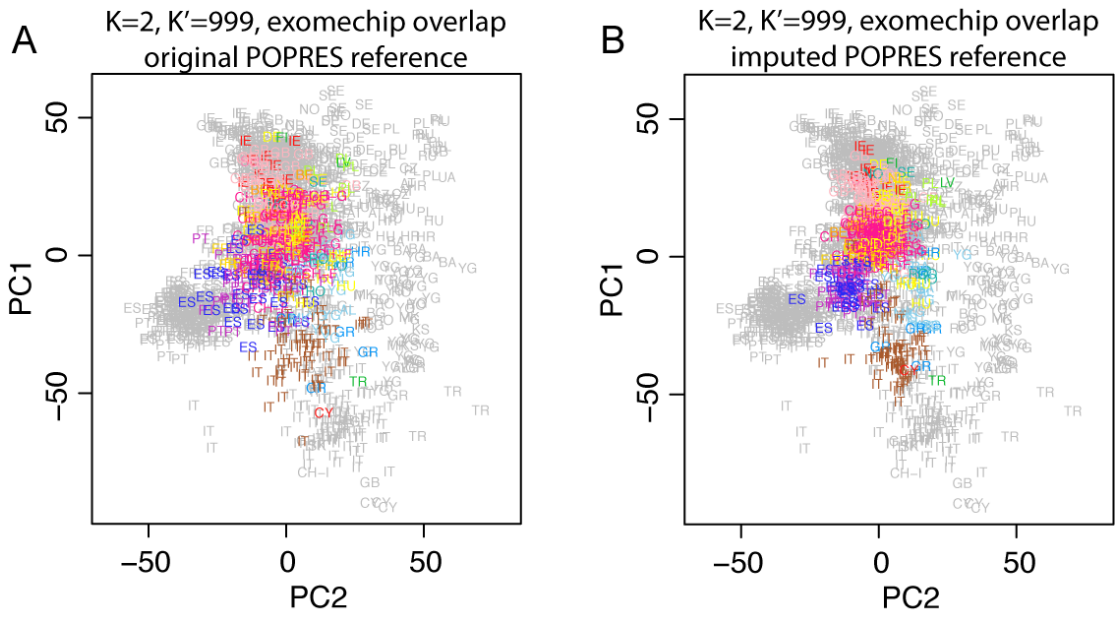

Figure S9. Illustration of the overfitting problem when using projection from an ultra-high-dimensional space. Grey symbols represent 1,000 reference individuals randomly selected from the POPRES data. Colored symbols represent the remaining 385 POPRES test individuals. Colors and population labels follow **Figure 2**. (A) Ancestry estimates based on 3,983 exomechip SNPs shared with the original POPRES data when $K = 2$ and $K' = 999$. The Procrustes similarity is $t_0 = 0.8598$ when compared to estimates based on genome-wide SNPs shown in **Figure 2A**. (B) Ancestry estimates based on 19,123 exomechip SNPs shared with the imputed POPRES data when $K = 2$ and $K' = 999$. The Procrustes similarity is $t_0 = 0.9323$ when compared to estimates in **Figure 2A**. In both panels, the test samples tend to shrink toward the center of the reference map, even though the sample-specific Procrustes similarity scores t are close to 1 for all test samples. This observation indicates that when K' is large (close to the reference sample size N), projection Procrustes analysis might over fit the data in searching for the optimal transformations between two sets of coordinates for the reference individuals, resulting in poor prediction of the coordinates for the test sample.

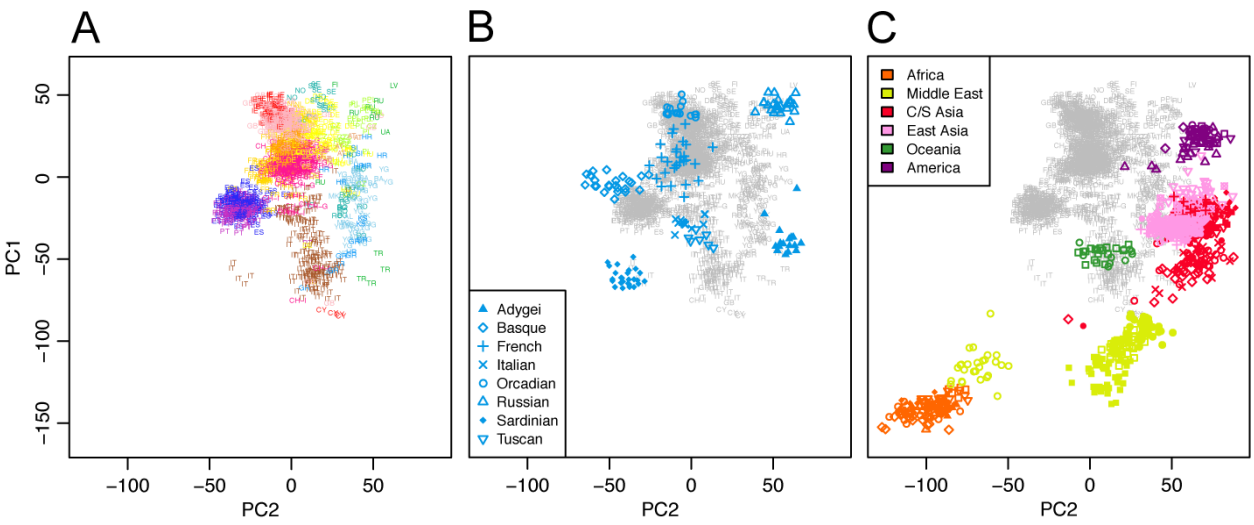

Figure S10. Ancestry analyses for worldwide samples in a European reference ancestry space. (A) European reference ancestry space constructed by the top 2 PCs of PCA on the POPRES data set (1,385 individuals and 318,682 SNPs). Colors and labels for the POPRES individuals follow **Figure 2**. (B) HGDP Europeans in the POPRES reference ancestry space. (C) HGDP non-Europeans in the POPRES reference ancestry space. In panels B and C, grey symbols represent the POPRES reference individuals, and colored symbols represent the HGDP individuals. Colors and symbols for the HGDP samples follow **Figure S3**. We placed the HGDP individuals on the POPRES PCA map using *trace* based on 82,692 shared SNPs between the HGDP and POPRES data sets ($K = 2$, $K' = 20$).

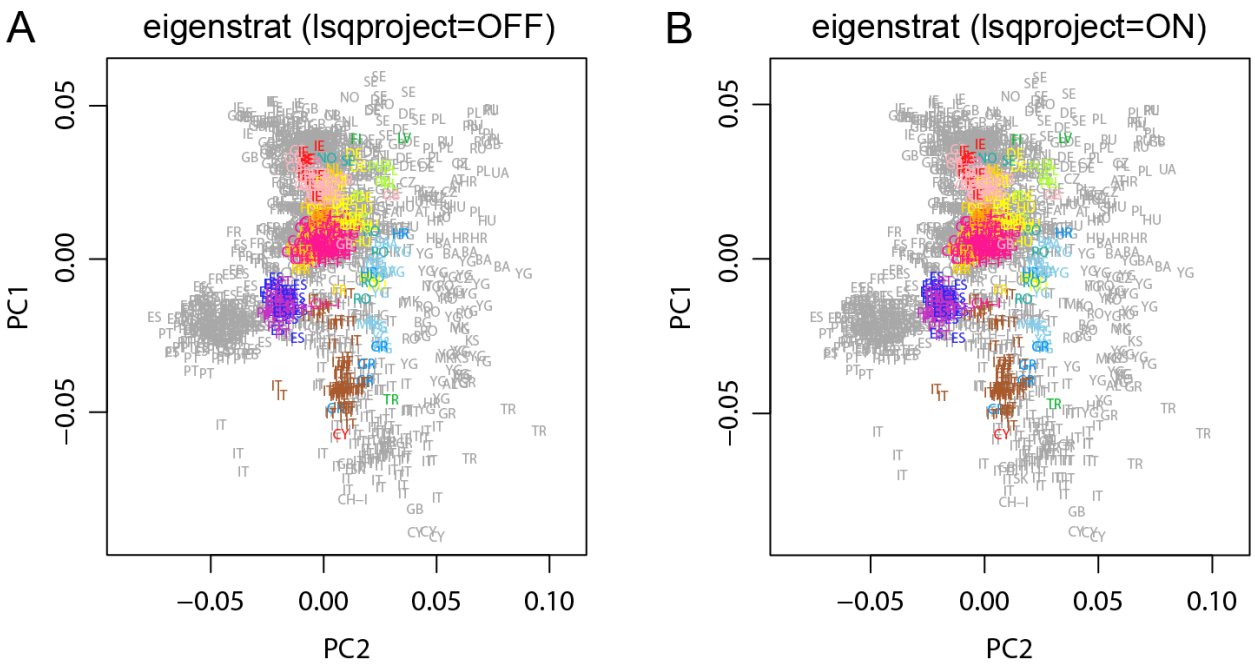

Figure S11. Projecting European samples onto a reference PCA map using EIGENSTRAT. Grey symbols represent 1,000 reference individuals randomly selected from the POPRES data. Colored symbols represent the remaining 385 POPRES test individuals. Colors and labels follow **Figure 2**. The analyses were based on 318,682 genotyped SNPs. (A) Projection based on the PC loadings of all SNPs. (B) Projection based on the least squares projection (*lsqproject*) option in EIGENSTRAT. Projected coordinates in panels A and B are highly similar because there is little missing data in the POPRES genotypes. The *lsqproject* option becomes identical to projection based on the PC loadings when there are no missing data. The scale of PCs in this figure is different from those produced by LASER 2.0 because of different normalization procedures in EIGENSTRAT and in LASER 2.0 when performing PCA.

Supplementary Table

Study sample size	$\boldsymbol{n = 1,000}$		$\boldsymbol{n = 10,000}$		$\boldsymbol{n = 100,000}$	
Method	**Time**	**Memory**	**Time**	**Memory**	**Time**	**Memory**
PCA (based on EVD)	0.5	13	85	3,536	-	-
PCA (based on SVD)	3	315	186	4,088	1,718	31,146
trace (N=200, K=K'=2)	1	15	12	15	148	15
trace (N=200, K=2, K'=20)	1	15	13	15	117	15
trace (N=400, K=K'=2)	8	28	53	28	368	28
trace (N=400, K=2, K'=20)	5	28	51	28	430	28
trace (N=800, K=K'=2)	32	68	286	68	3,557	68
trace (N=800, K=2, K'=20)	31	68	338	68	3,747	68

Table S1. Computational time (in minutes) and memory usage (in megabytes) for PCA and *trace* when applied to genotype data. We simulated diploid SNP data with $L = 10{,}000$ loci and no missing data. *N* is the number of individuals in the reference panel. We computed PCA using two approaches: (1) performing eigenvalue decomposition (EVD) on a $n \times n$ genetic relationship matrix; and (2) performing singular value decomposition (SVD) on a $n \times L$ standardized genotypic matrix. Both approaches for computing PCA were implemented in the LASER 2.0 software package. When $n = 100{,}000$, the EVD-based PCA was aborted because the $n \times n$ genetic relationship matrix is too large for decomposition. The evaluation was based on a 2.3 GHz CPU in a Unix computing cluster.