iScience, Volume *26*

Supplemental information

Federated generalized linear mixed models for collaborative genome-wide association studies

Wentao Li, Han Chen, Xiaoqian Jiang, and Arif Harmanci

Supplementary Figures

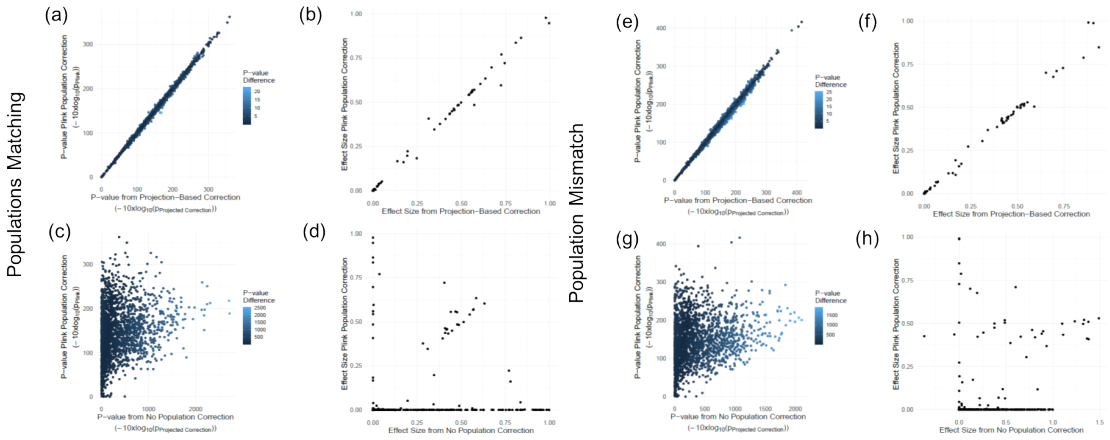

Figure S1: Comparison of projection-based population stratification with PCA-based stratification among 100 simulated GWAS studies (Related to Fig 2). (a) Scatter plot shows the p-value ($-10 \times \log_{10}(p-value)$) estimates with plink2 PCA-based population correction (y-axis) versus p-value estimates from population correction using projection-based estimation of population covariates. (b) Scatter plot shows the effect size estimates with plink2's PCA-based population correction (y-axis) versus effect size estimates from population correction using projection-based estimation of population covariates. (c) Scatter plot shows the p-value estimates with plink2's PCA-based population correction (y-axis) versus p-value estimates without population correction. (d) Scatter plot shows the effect size estimates with plink2's PCA-based population correction (y-axis) versus effect size estimates without population correction. (e) Scatter plot shows the p-value ($-10 \times log_{10}(p-value)$) estimates with plink2's PCA-based population correction (y-axis) versus p-value estimates from population correction using projection-based estimation of population covariates where the reference population is not matching to the populations of GWAS individuals. (f) Scatter plot shows the effect size estimates with plink2's PCA-based population correction (y-axis) versus effect size estimates from population correction using projection-based estimation of population covariates for non-matching reference population. (g) Scatter plot shows the p-value estimates with plink2's PCA-based population correction (y-axis) versus p-value estimates without population correction for non-matching reference population. (h) Scatter plot shows the effect size estimates with plink2's PCA-based population correction (y-axis) versus effect size estimates without population correction for non-matching reference population.

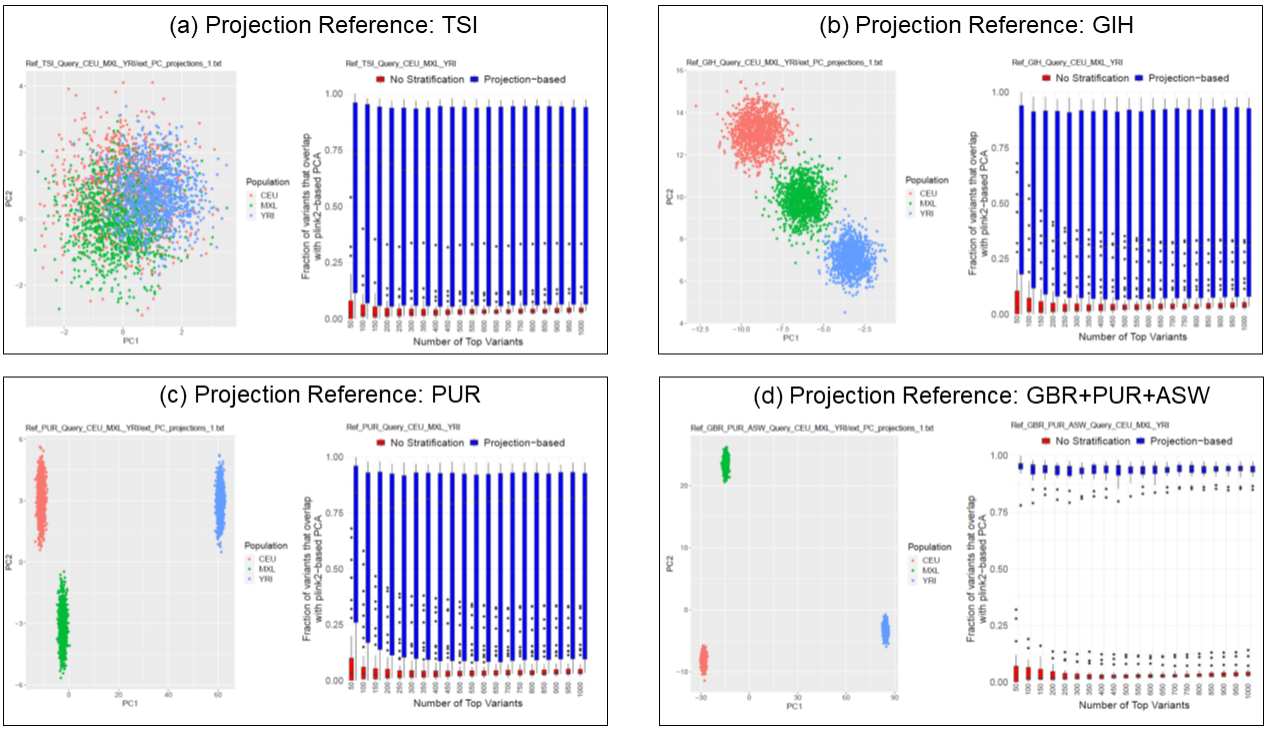

Figure S2: The impact of reference panel selection for projected PC calculations using simulated study cohort that consists of CEU, MXL, YRI populations from The 1000 Genomes Project (Related to Fig 2). (a) The scatter plot of PC1, PC2 (left) shows each subject colored by their populations when the reference is selected as a European panel (TSI). The p-value concordance of top associated variants (right) shows the fraction of variants that are matching (y-axis) with increasing number of variants (x-axis). Concordance is calculated by comparing the plink2 runs using the covariants calculated using a full-PCA vs the covariants calculating from the projection. (b) The scatter plot of subjects (left) and p-value concordance (right) when the reference panel is GIH (Gujarati Indians in Houston). (c) Scatter plot and p-value concordance when the reference panel is PUR (Puerto Ricans living in Puerto Rico). (d) Scatter plot and p-value concordance when the reference panel consists of GBR, PUR, ASW. Note that this panel is not exactly matching to the simulated study cohort subjects.

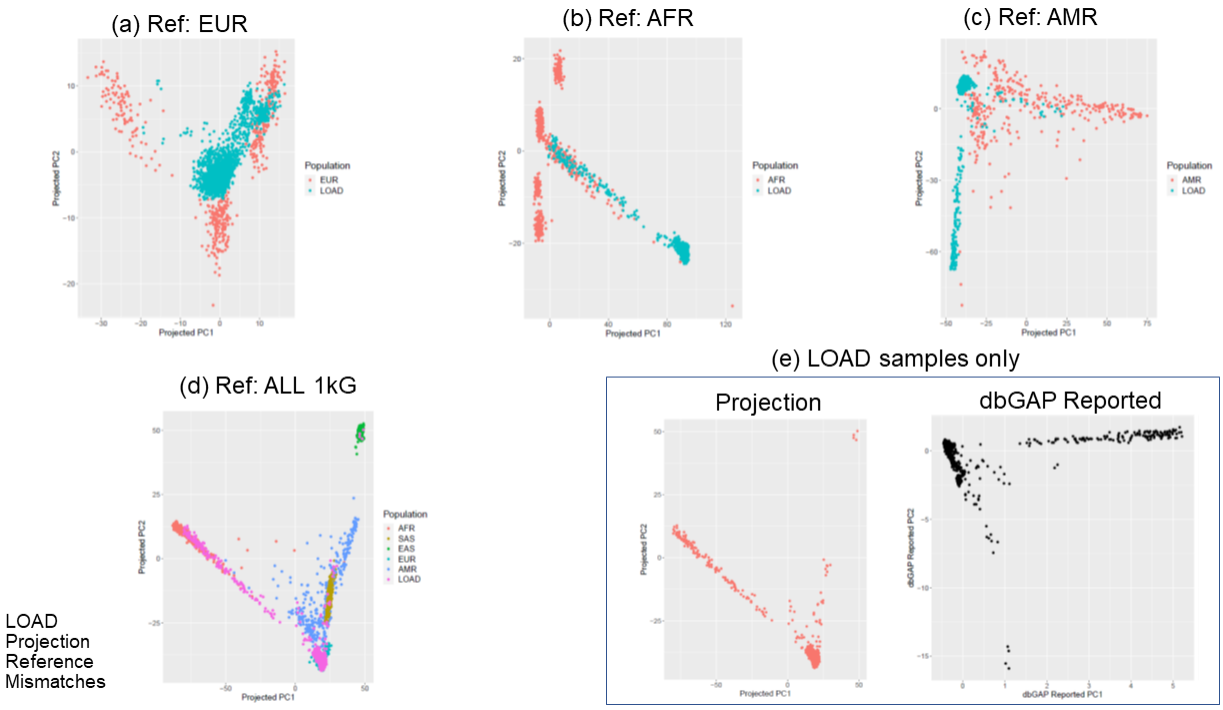

Figure S3: Scatter plots of the dbGAP GWAS cohort subjects with different reference panel for calculating projected population covariates (Related to Fig 2). (a) Scatter plot shows the projected PC1-vs-PC2 for the reference panel individuals (magenta dots) and the study participants (cyan dots) when the reference panel consists of EUR (European) super-population. (b) Scatter plot shows the PC1-vs-PC2 when the reference panel consists of AFR super-population (African). (c) Scatter plot shows the PC1-vs-PC2 when the reference panel consists of AMR (American) super-population. (d) Scatter plot shows the distribution of reference and dbGAP subjects on projected PC1-vs-PC2 when all 1000 Genomes populations are used. (e) The comparison of the projected PCs (left) and the dbGAP reported PCs (right) for the study subjects only. Both scatter plots reflect the triangular shape of the European-African-Asian triangular shape of the genetic ancestry in the embedded PC space.

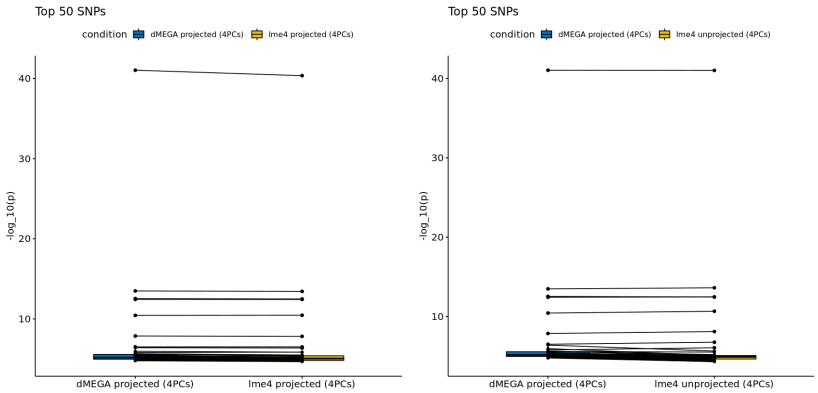

Figure S4: Comparison of significance levels of top 50 SNPs (Related to Fig 4). (Left) Paired boxplot of comparison †; (Right) Paired boxplot of comparison ∗

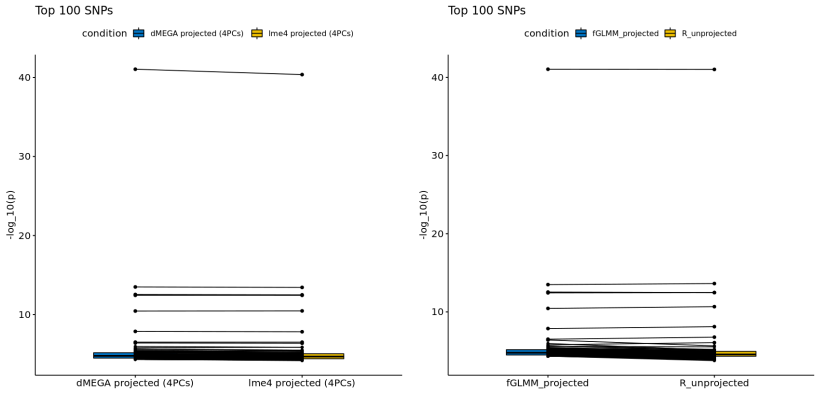

Figure S5: Comparison of significance levels of top 100 SNPs (Related to Fig 4). (Left) Paired boxplot of comparison †; (Right) Paired boxplot of comparison ∗

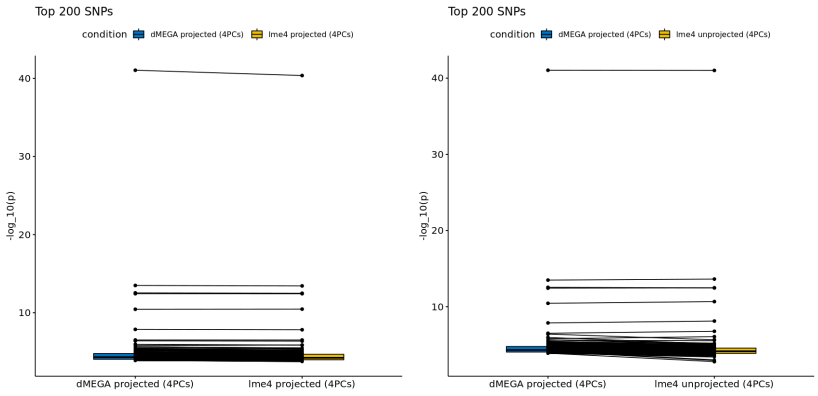

Figure S6: Comparison of significance levels of top 200 SNPs (Related to Fig 4). (Left) Paired boxplot of comparison †; (Right) Paired boxplot of comparison ∗

Supplementary Tables

Table S1: Summary of successful runs with cGLMM (Related to Section *Comparison with cGLMM*). Out of the 20 runs of cGLMM algorithm, table reports the number of iterations and the minimum squared error (MSE) loss value for the 7 successful runs.

Run NO.	Number of iterations to converge	MSE Loss
#7	7	0.5475
#9	7	0.5475
#12	7	0.5475
#16	7	0.5475
#18	3	0.4433
#19	7	0.5475
#20	5	0.3672

Table S2: Summary of successful runs with dMEGA. Out of the 20 runs of 10 random SNPs with dMEGA algorithm (Related to Section *Comparison with cGLMM*). Table reports the mean and standard deviation of P-values and Coefficients of 10 SNPs.

SNP	**# Run**	**P-value (Mean)**	**P-value (Std. Dev.)**	**Coef. (Mean)**	**Coef. (Std. Dev.)**
rs8008645	20/20	0.0079	0.0005	0.1720	0.0010
rs7030500	20/20	0.0004	0.0000	0.2313	0.0022
rs4770728	20/20	0.0005	0.0000	0.3824	0.0011
rs4562717	20/20	0.1158	0.0041	-0.1254	0.0012
rs6019464	20/20	0.0363	0.0011	0.1650	0.0012
rs10503305	20/20	0.0026	0.0003	0.1888	0.0020
rs2488736	20/20	0.0181	0.0017	-0.2955	0.0038
rs11209333	20/20	0.0311	0.0016	0.1864	0.0015
rs9556476	20/20	0.0050	0.0004	-0.2551	0.0024
rs6469733	20/20	0.0019	0.0003	0.2048	0.0030