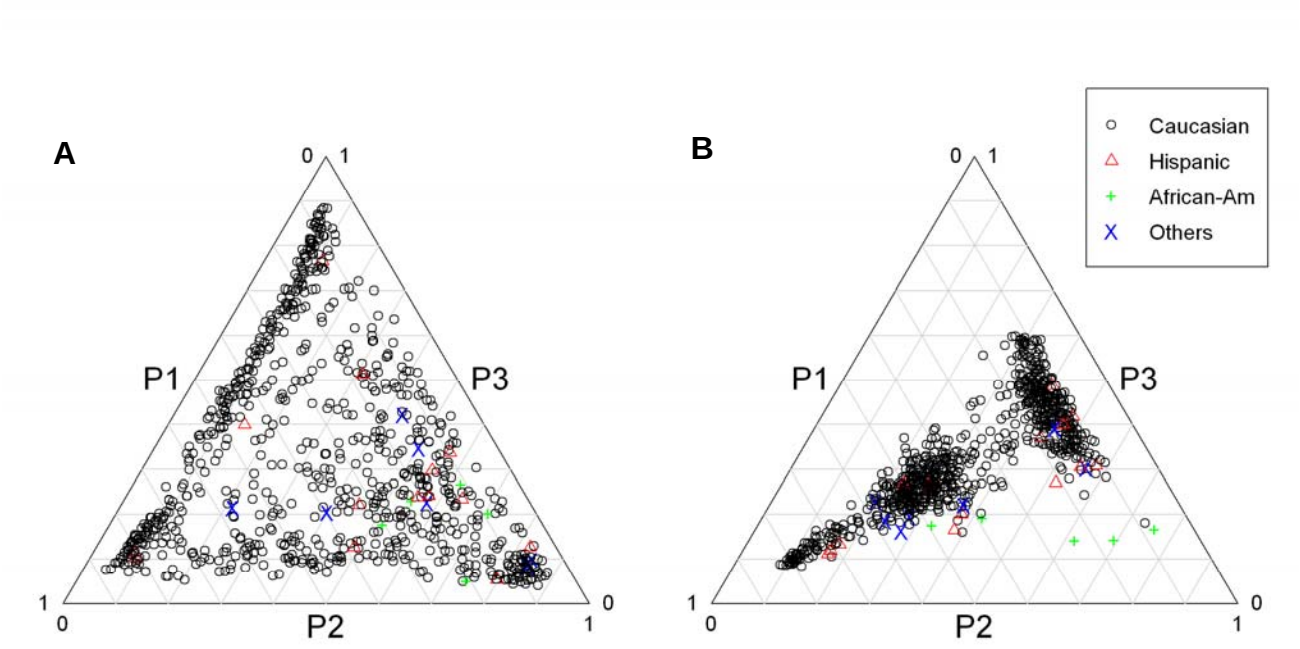
Supplemental Materials:

1. *STRUCTURE analysis (real data 1):* Assuming that there are three underlying sub-populations, we first applied STRUCTURE to the real data 1 using all 81 ancestry informative SNPs. We then randomly removed 4 SNPs that are in high LD with other SNPs (leaving one SNP with each LD block), and re-applied STRUCTURE to the reduced data set. Supplemental Figure 1 summarizes the analysis.

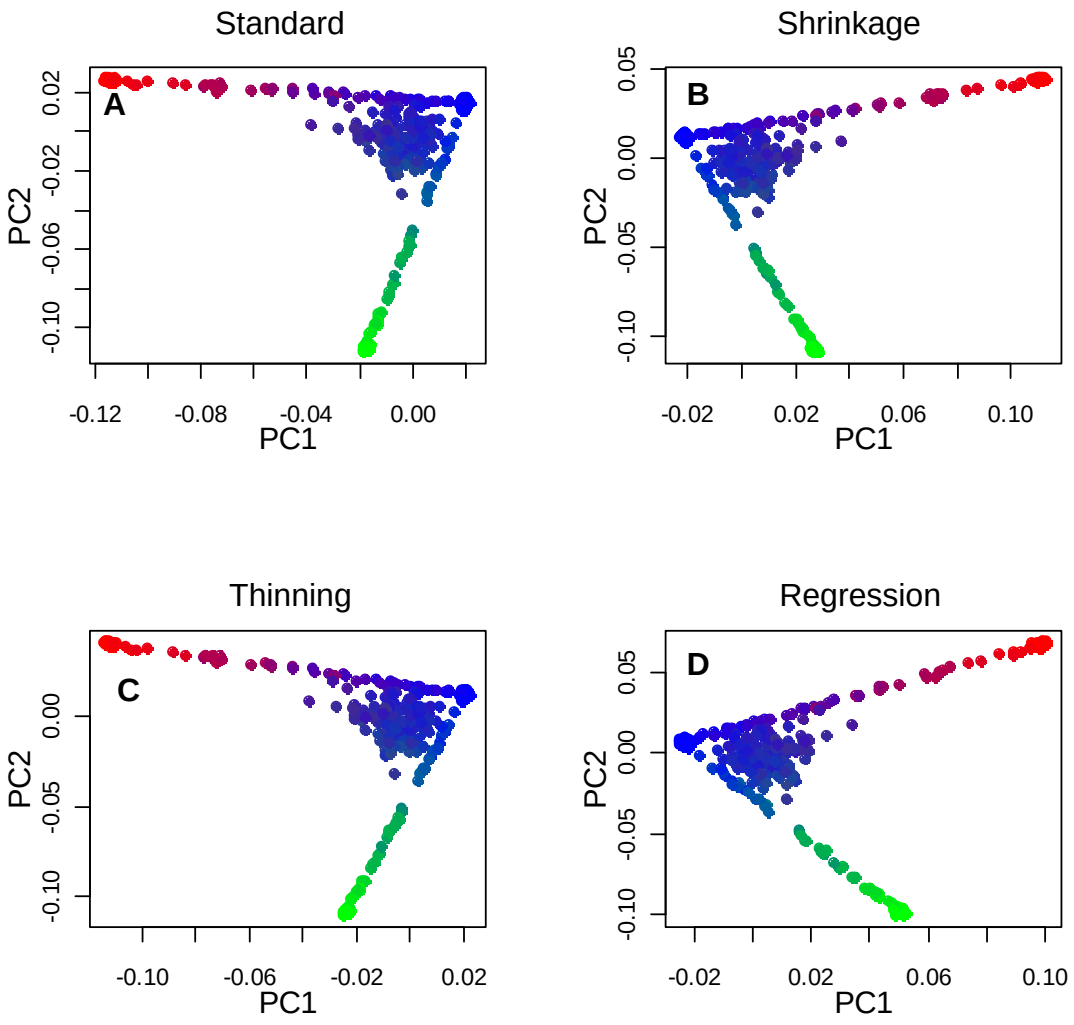


Supplemental Figure 1: STRUCTURE analysis on real data 1. The left panel is based on the data using all 81 ancestry informative SNPs, while the right panel is from the data with only 77 nearly independent SNPs.

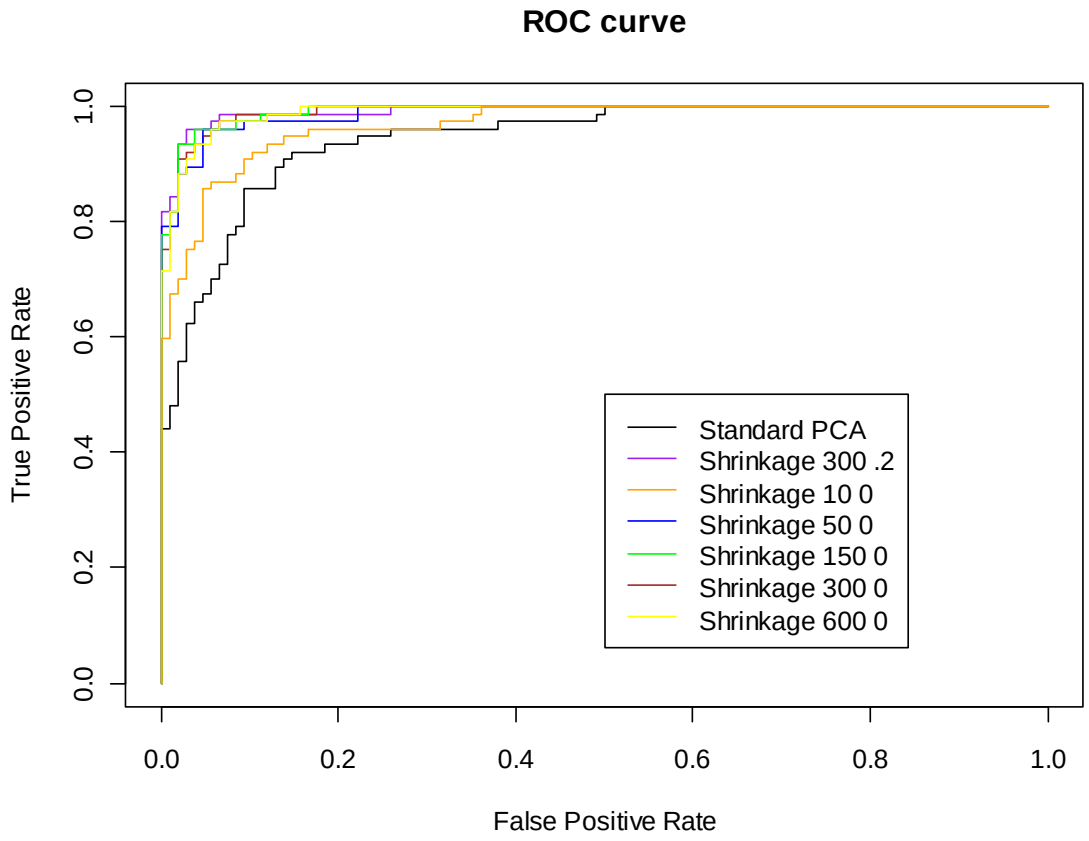
2. *Simulation 3 (substantially stratified GWAS data based on Hapmap Samples):* With *HapSample* software, we first simulated 450 CEU samples, 50 YRI samples, and 50 JP+CH samples respectively using the SNPs on the Affymetrix 100K array [Wright, et al. 2007]. *HapSample* generates data by resampling from existing phased Hapmap datasets and therefore preserves the observed local LD structure in Hapmap samples. We then generated additional 225 individuals with mixed genomes from the three populations, using our modified code from *HapSample*. Specifically, we generated 50 admixture samples of CEU and YRI, 50 admixture samples of CEU and JP+CH, and the reminder125 are admixture samples of the three populations. That is, for the *i*th admixture sample, we have

$$(p_{i1},\quad p_{i2},\quad p_{i3}) = \begin{cases} (u_i,\quad 1-u_i,\quad 0), & i \le 50 \\ (u_i,\quad 0,\quad 1-u_i), & 51 \le i \le 100 \\ (g_{i1},\quad g_{i2},\quad 1-g_{i1}-g_{i2}), & 101 \le i \le 225 \end{cases}$$

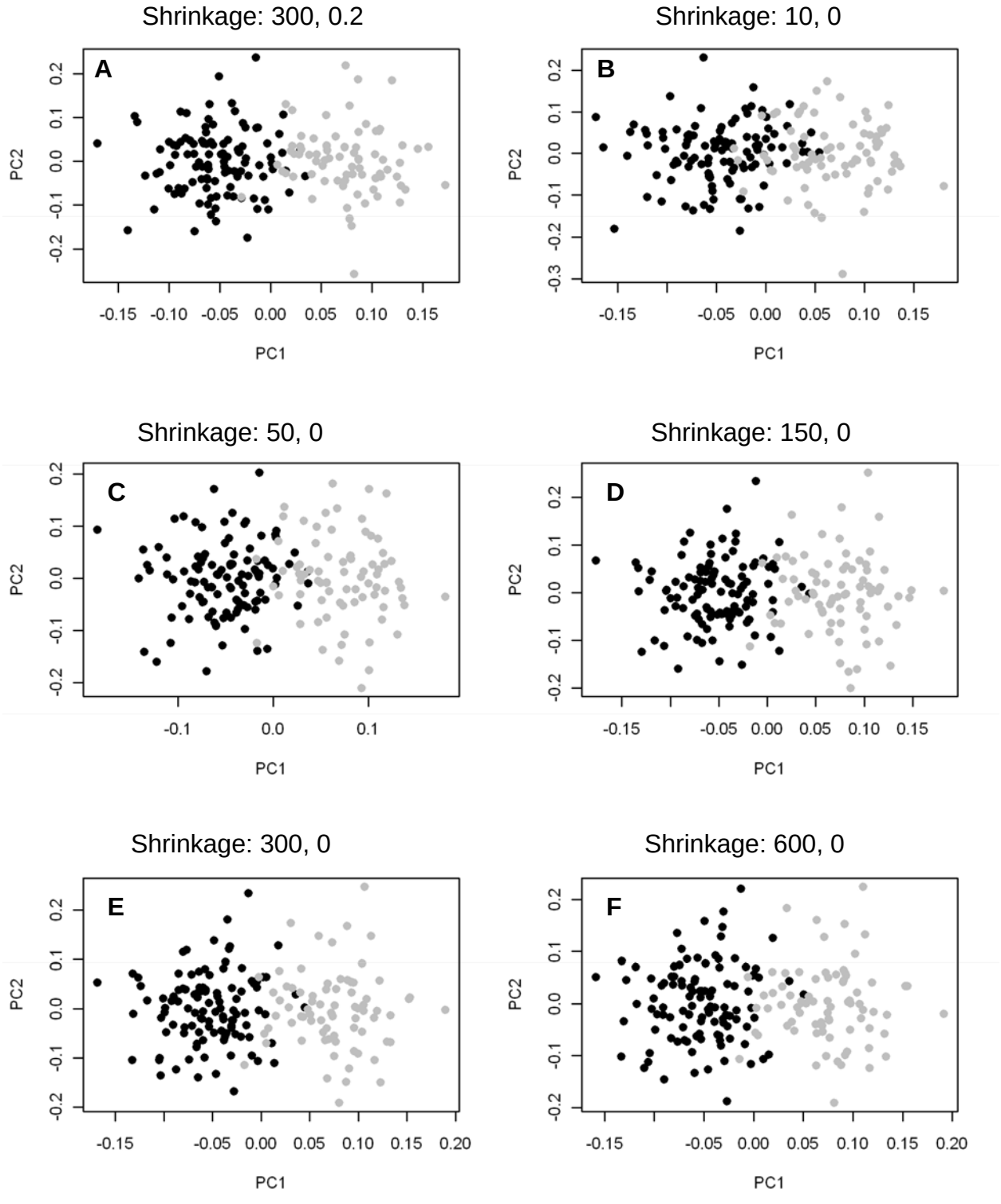
where $p_{i1}, p_{i2,} p_{i3}$ are the corresponding CEU, YRI and JP+CH genome proportions, respectively, $u_i \sim Unif(0,1)$ and $(g_{i1}, g_{i2}) \sim Dirichlet(70,15,15)$. The final simulated data has 775 samples and 109,723 SNPs. Supplemental Figure 2 presents the scatter plots of the top two PCs derived from the four PCA methods. Clearly, for substantially stratified populations where a large number of SNPs are available, all four methods perform equally well.



Supplemental Figure 2: Results from simulation 3. Scatter plots of the top two PCs of 4 different methods with different colors for CEU (blue), YRI (red) and JP+CH (green). Admixture samples are plotted in colors according to their genomic proportions of the three populations.



Supplemental Figure 3: ROC curves of shrinkage PCA on real data 2 with varying parameters in w_i. The two numbers after "Shrinkage" refer the number of markers and the correlation threshold, respectively.



Supplemental Figure 4: Scatter plots of Shrinkage PCA on real data 2 with varying parameters

in w_i . The two numbers after “Shrinkage” refer to the number of markers and the correlation threshold, respectively.