HGEN Module 3 Project (data 4-6)

Ling Chen

I think Data 4 are European samples, because it has fewer variants and does not have obvious population struction. Data 5 and Data 6 are exactly the same in terms of people's ancestry, but the case and control lables are shuffled to make Data 5 have cases in one population and controls in another population and make Data 6 to have matched population substructre for cases and controls. Data 5 and Data 6 should be a mixture of European samples and African samples.

1. Allele Frequency estimates

- Estimate the allele frequency for each variant

- Plot the distribution (histogram) of the allele frequency

- Discuss your thoughts of the allele frequency pattern regarding genetic association studies.

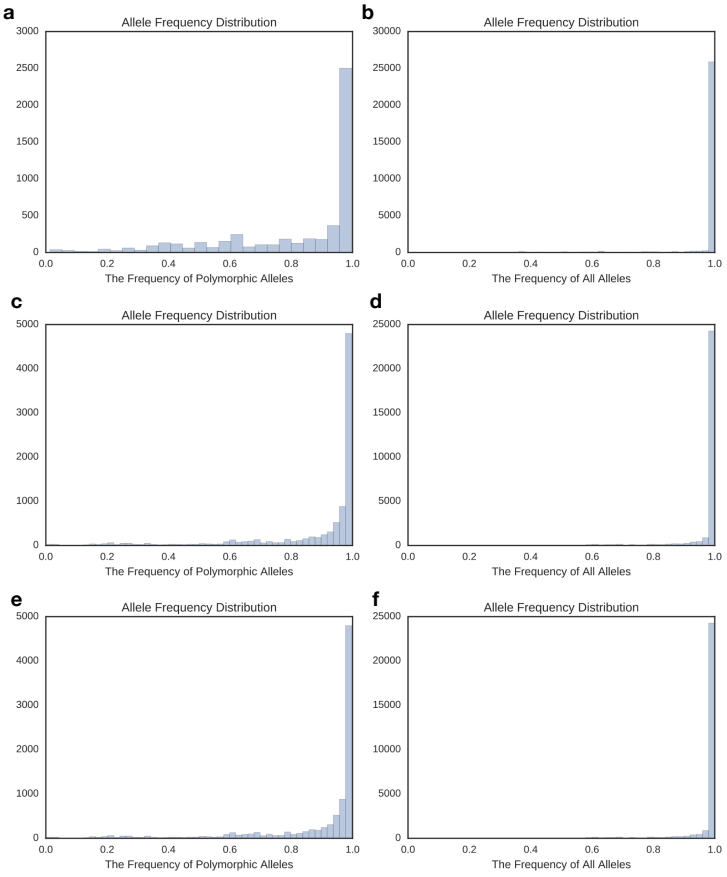


Figure 1. Allele frequency distribution of dataset 4-6. (a) The distribution of frequency of polymorphic alleles in data 4. (b) The distribution of frequency of all alleles in data 4. (c) The distribution of frequency of polymorphic alleles in data 5. (d) The distribution of frequency of all alleles in data 5. (e) The distribution of frequency of polymorphic alleles in data 6. (f) The distribution of frequency of all alleles in data 6.

I calculated the allele frequency of each variant in data 4-6 and plot the distribution for the allele frequencies for each data set.

Data 4-6 contains 150 individuals and 28746 SNPs. Those SNPs are supposed to be common SNPs used in GWAS studies. However, as we saw in the allele frequency distribution of polymorphic alleles in Figure 1a, 1c, 1e, the y axis is much smaller than that in the allele frequency distribution of polymorphic alleles in Figure 1b, 1d, 1f, suggesting that many of the common SNPs are monomorphic in the selected 150 individuals. Data 4 has much fewer polymorphic alleles than dataset 5 and 6, suggesting it may come from a younger population compared to the data 5 and data 6. Data 5 and 6 seems to have the same allele frequency distribution.

Although the SNPs in data 4-6 are supposed to be a collection of common SNPs in population, many SNPs in data 4-6 have very low minor allele frequency and low allele frequency SNPs have low power to be detected as risk associated SNPs in GWAS studies. With sample size of 150, it would be only possible to detect common SNPs with relative large odds ratio (e.g. >2) at a reasonable power (e.g. 80%). For these datasets, it is unlikely to detect any risk associated rare SNPs due to (i) data is from genotyping, not from sequencing (ii) small sample size.

2. Hardy-Weinberg Equilibrium

- Calculate the p value of HWE for each variant

- Plot the distribution of the p values and QQ plot of the p values (log scale)

- Discuss the HWE patterns observed in the datasets.

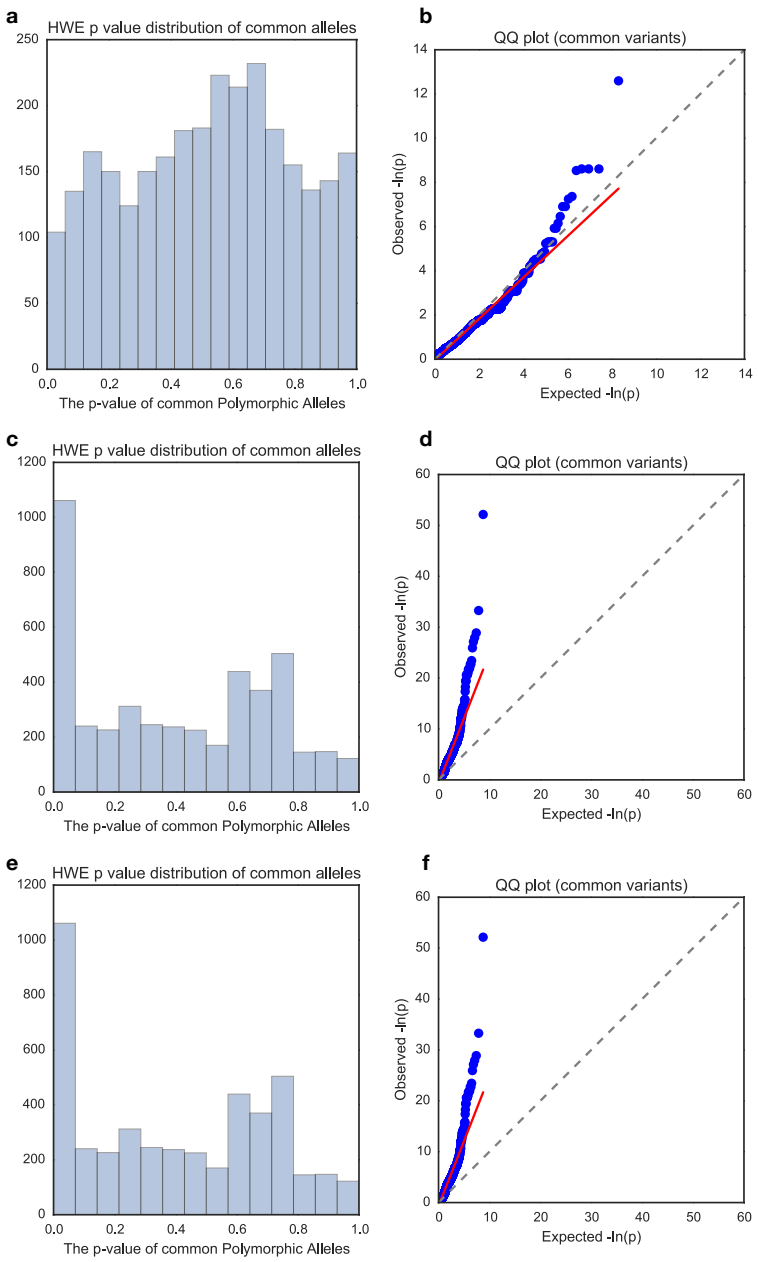


Figure 2. Hardy-Weinberg P value distribution of common polymorphic alleles (minor allele frequency > 2%) in data 4-6. (a) The Hardy-Weinberg P value distribution of data 4. (b) QQ plot of the Hardy-Weinberg P values of data 4. (c) The Hardy-Weinberg P value distribution of data 5. (d) QQ plot of the Hardy-Weinberg P values of data 5. (e) The Hardy-Weinberg P value distribution of data 6. (f) QQ plot of the Hardy-Weinberg P values of data 6.

I calculated whether allele frequencies of a SNP is significantly different from what is expected under Hardy-Weinberg equilibrium using chi-squared goodness of fit with 1 degree of freedom for every SNP in data 4-6.

Most SNPs in data 4 are in Hardy-Weinberg equilibrium (Figure 2a, 2b). Only a few SNPs have lower p-value than expected. It is probably due to (i) they are

disease associated alleles. (ii) genotyping error. In data 5 and 6, the HWE p value distribution seems to be the same (Figure 2c, 2d, 2e and 2f). It is systematically skewed towards low p values. This indicates there are subpopulation structures in these two datasets.

3. Linkage Disequilibrium

- Calculate the pairwise LD among all pairs of the variants (D, D' and R2). Store the LD in an M by M table for each of the D, D' and R2. You need to use EM algorithm to estimate the haplotype frequencies.

- Plot the LD in the M by M matrix for each of the LD metrics, using the same color scheme so that you can contrast different LD measurements.

- Investigate how LD (D, D' and R2) patterns are influenced by the allele frequency, and discuss the implication for genetic association studies.

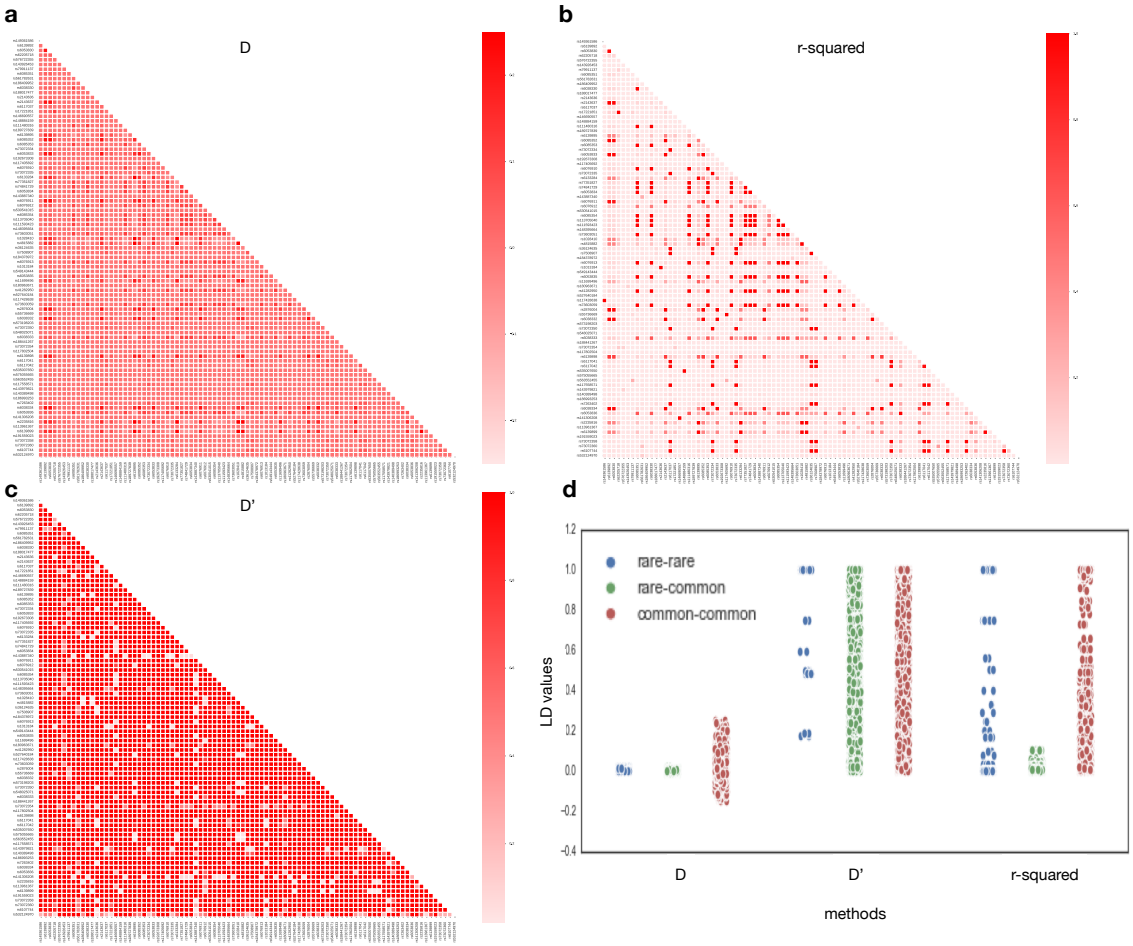


Figure 3. LD values of polymorphic alleles in first 500 SNPs in Data 4 based on different LD metrics (D, D', r^2). (a) D value of pairs of polymorphic alleles in first 500 SNPs in Data 4. (b) r^2 value of pairs of polymorphic alleles in first 500 SNPs in Data 4. (c) D' value of pairs of polymorphic alleles in first 500 SNPs in Data 4. (d) Comparison of LD values between rare-rare snp pair, rare-common snp pair, common-common snp pair.

I first estimate the haplotype frequencies through EM algorithm. Then I calculated the LD values usging D, D' and r^2. Here I included the LD heatmaps for polymorphic alleles in first 500 SNPs in Data 4 in Figure 3a-c. The LD heatmaps for Data 5 and Data 6 are in Supplementary Figure 1.

Compared to D' and r^2, D values can be both negative and positive and have lower abosolute value. D' is essentially a measure of recombination. If there is no recombination, then even the allele frequency is very low, D' can still be high. r^2 value computes the correlation of the frequency of two alleles.

To investigate how LD (D, D' and R2) patterns are influenced by the allele frequency, I sampled 100 low frequency alleles ('rare', allele frequency 0.02) from data 5. Note that the definition I used here is not common standard because there are very few true rare variants in these dataset and I want to separate the allele frequencies of the two classes away. From Figure 3d, I observed: (i) For D, the rare-rare pairs and rare-common pairs have lower D absolute values compared to the common-common pairs. (ii) For D', there is no obvious difference of D' values between rare-rare pairs, rare-common pairs and common-common pairs. (iii) For r^2, the rare-common pairs have lower R2 values compared to the rare-rare pairs and the common-common pairs. In summary, the D' seems to not be greatly affected by the allele frequencies. However, both D and R2 are largely dependent on the allele frequencies. Genetic association studies often use genotyping arrays that have pre-selected SNPs and focus on common SNPs. The untyped SNPs are then inferred from tagSNPs. Because rare variants have low R2 values with tagSNPs, causal rare variants are very hard to detect in the genetic associtation studies using common SNP genotyping arrays.

4. Principal component analysis

- Code the genotypes using an additive model (i.e. use 0, 1 and 2 to code 0/0, 0/1 and 1/1) and carry out PCA. Calculate the proportion of variance accounted for by the 1st, the 2nd, and the 3rd PC.

- Plot the first two PCs of all individuals on an X-Y plot. You need to label the cases in one color and controls in another color.

- Repeat 4.2 for PC1 vs. PC3 and PC2 vs. PC3. Which plot gives clear patterns about the pop substructure?

- Discuss whether population substructure is a confounder or not in each of the datasets. For those datasets in which pop substructure is not a confounder discuss whether you want to use PCA to account for population substructure or not.

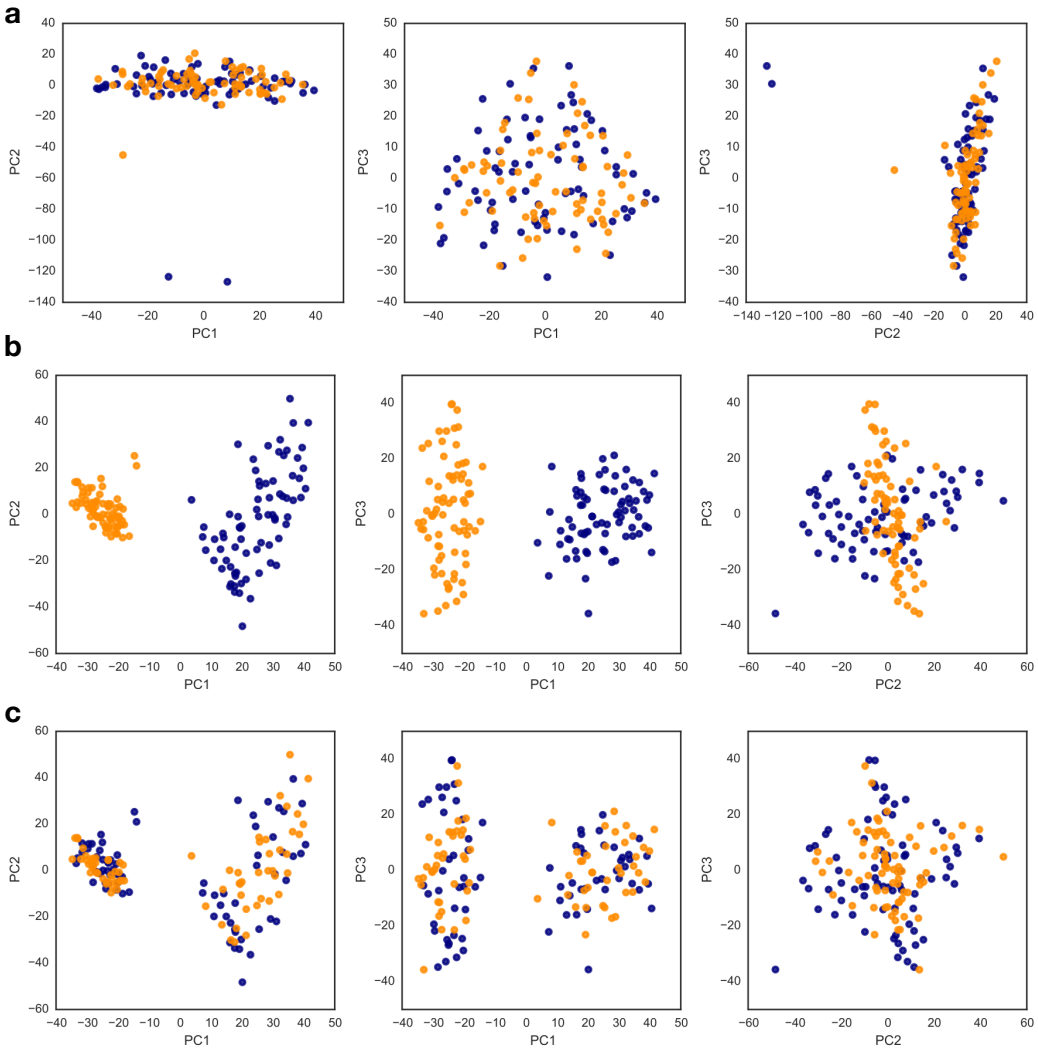


Figure 4. PCA analysis of population substructure in Data 4-6. (a) PCA analysis of Data 4. The three plots are the PC1 against PC2, PC1 against PC3 and PC2 against PC3 from left to right. (b) PCA analysis of Data 5. (c) PCA analysis of Data 6. Orange are cases and blue are controls

In data 4, there is no population structure. Therefore, the population substructure is not a confounder for data 4.

In data 5, there are two subpopulations, one only contains the cases and one only contains controls. Therefore, the population substructure is a confounder for data 5. To control for the population substructure in 5, I would include the first several PC components in the regression model that is used in the genetic association studies.

In data 6, there are two subpopulations. However, the cases and controls are equally distributed in two subpopulations. Therefore, the population substructure is a confounder for data 6. I would still include the first several PC components in the regression model that is used in the genetic association studies to increase the power.

5. For datasets that you suspect have population substructure based on the plots in 4), develop a Gibbs sampler to infer the memberships of all samples. You may decide the number of substructures based on your visual inspection of the PCA plots. Output the membership of each individual with the marginal probability of being in the assigned substructure based on your Gibbs sampler results.

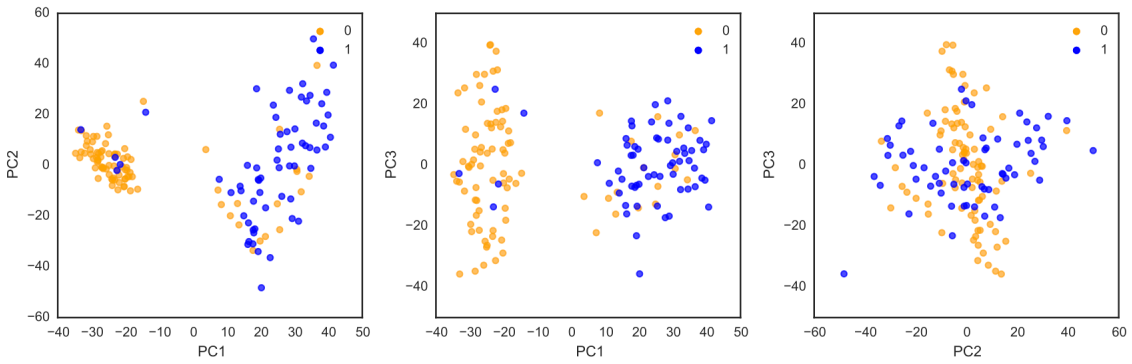
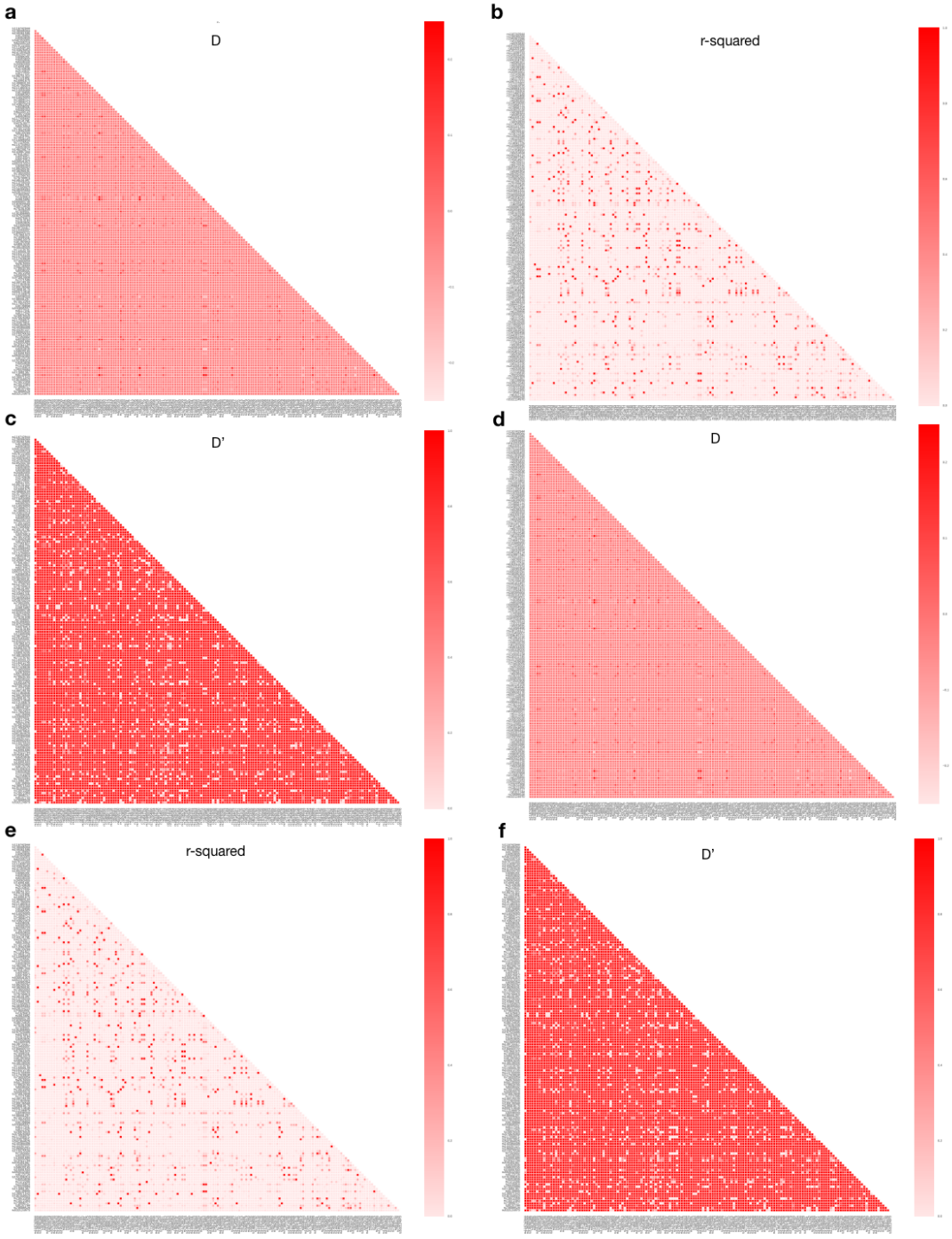


Figure 5. The comparison of gibbs sampling method and the PCA anlysis of Data 5. The color codes the membership of the gibbs sampling results.

I implemented the gibbs sampling methods and run it for common alleles in first 1000 SNPs of Data 5. Then I plot the subpopulation membership assignment from stationary distribution obtained from gibbs sampling method with the PCA plot of Data 5. The results of gibbs sampling mostly agree with the PCA analysis (Figure 5).

Supplementary Figures:



Supplementary Figure 1. LD values of polymorphic alleles in first 500 SNPs in Data 5 and Data 6 based on different LD metrics (D, D', r^2). (a) D value of pairs of polymorphic alleles in first 500 SNPs in Data 5. (b) r^2 value of pairs of polymorphic alleles in first 500 SNPs in Data 5. (c) D' value of pairs of polymorphic alleles in first 500 SNPs in Data 5. (e) D value of pairs of polymorphic alleles in first 500 SNPs in Data 6. (f) r^2 value of pairs of polymorphic alleles in first 500 SNPs in Data 6. (g) D' value of pairs of polymorphic alleles in first 500 SNPs in Data 6.