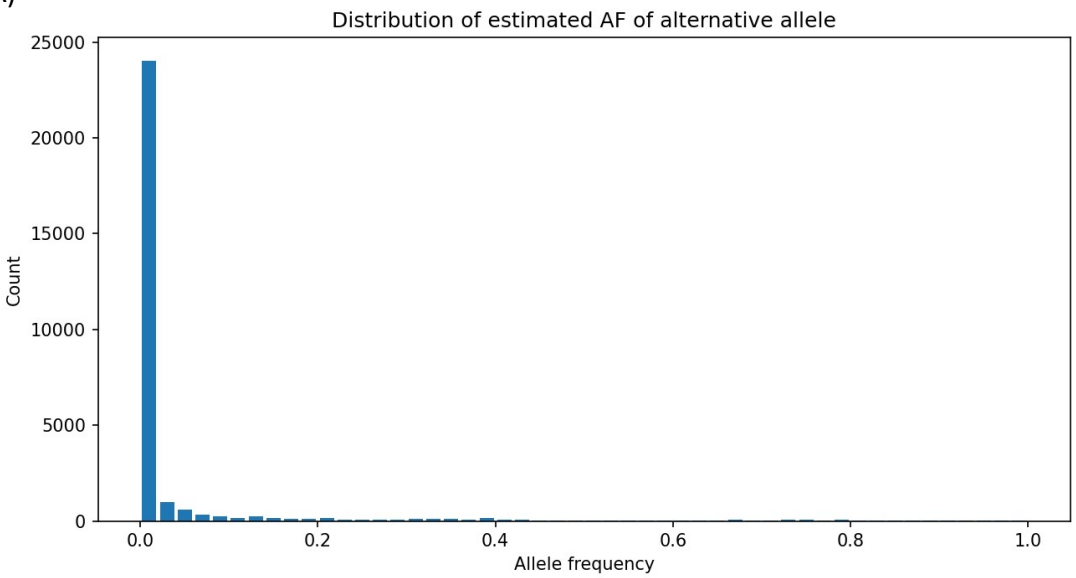
1. Allele frequency estimates

Majority of the variants have very low allele frequencies as showed in figure 1A. If we only plot rare SNPs with minor allele frequency (MAF) less than 0.01 (figure 1B), we can observe that most SNPs have MAF of zero.

Since GWAS does not have high power to detect phenotype association with rare SNPs, I think it is not appropriate to apply GWAS on this dataset. A much larger sample size could be helpful, but using a different analysis method might be better.

(A)



(B)

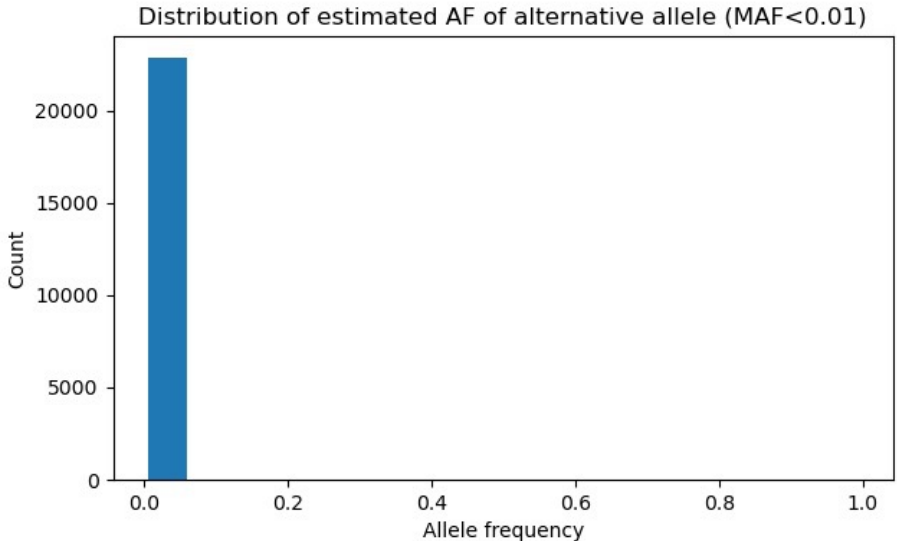


Figure 1. Allele frequency distribution

2. Hardy-Weinberg Equilibrium

There are a lot of SNPs with p values lower than 0.05 (figure 2A) even after removal of variants with MAF<0.05. These SNPs are not in Hardy-Weinberg equilibrium and indicates there might be substructures in the population. QQ plot in figure 2B also shows inflation of the data, as observed p values depart from expected p values almost from beginning of the curve.

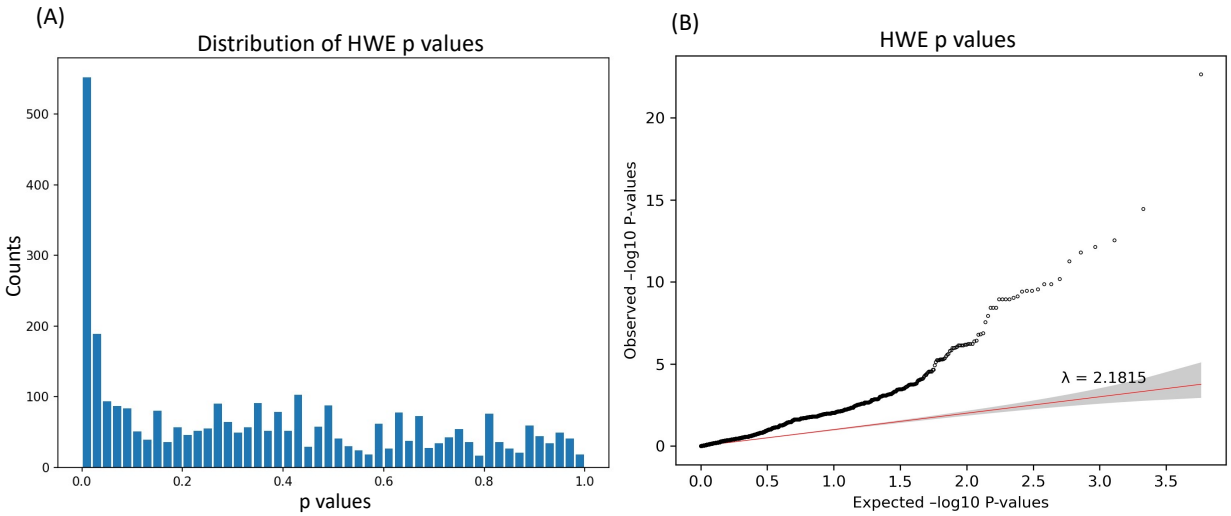


Figure 2

3. Linkage disequilibrium

Below equations are used to calculate D, D' and r^2:

$D(AB) = P(AB) - P(A)P(B)$

$D'(AB) = D(AB) / D_{max}(AB)$

- If D>0: $D_{max} = \min [P(A)P(b), P(a)P(B)]$
- If D<0: $D_{max} = \min [P(A)P(B), P(a)P(b)]$

$$r^2 = \frac{D^2(AB)}{P(A)P(a)P(B)P(b)}$$

Value and range of D depend on allele frequencies, so it is difficult to compare D to D' and r^2 directly.

D' is standardized version of D and ranges from -1 to 1, but often treated as 0 to 1 since signs of D and D' are arbitrary. Generally, D' of zero means no linkage and D of 1 indicates high linkage. Correlation coefficient r^2 is another way to standardize D. It also ranges from 0 to 1 with zero means no linkage and 1 means high linkage.

Figure 3A and 3E show that values of D' is high when r^2 is high but not vice versa. D' will be 1 if only 2 or 3 haplotypes are observed, therefore we get extreme D' values even when r^2 are low. In this dataset, this issue is likely caused by small allele frequencies. Figure 3F proves there are still a lot of SNPs with low allele frequencies after filtering by MAF>=0.05.

Figure 3B-3D are heatmaps of LD D, D' and r^2 between first 100 SNPs. Graph of r^2 has the clearest demonstration of LD blocks (red triangles) among selected variants.

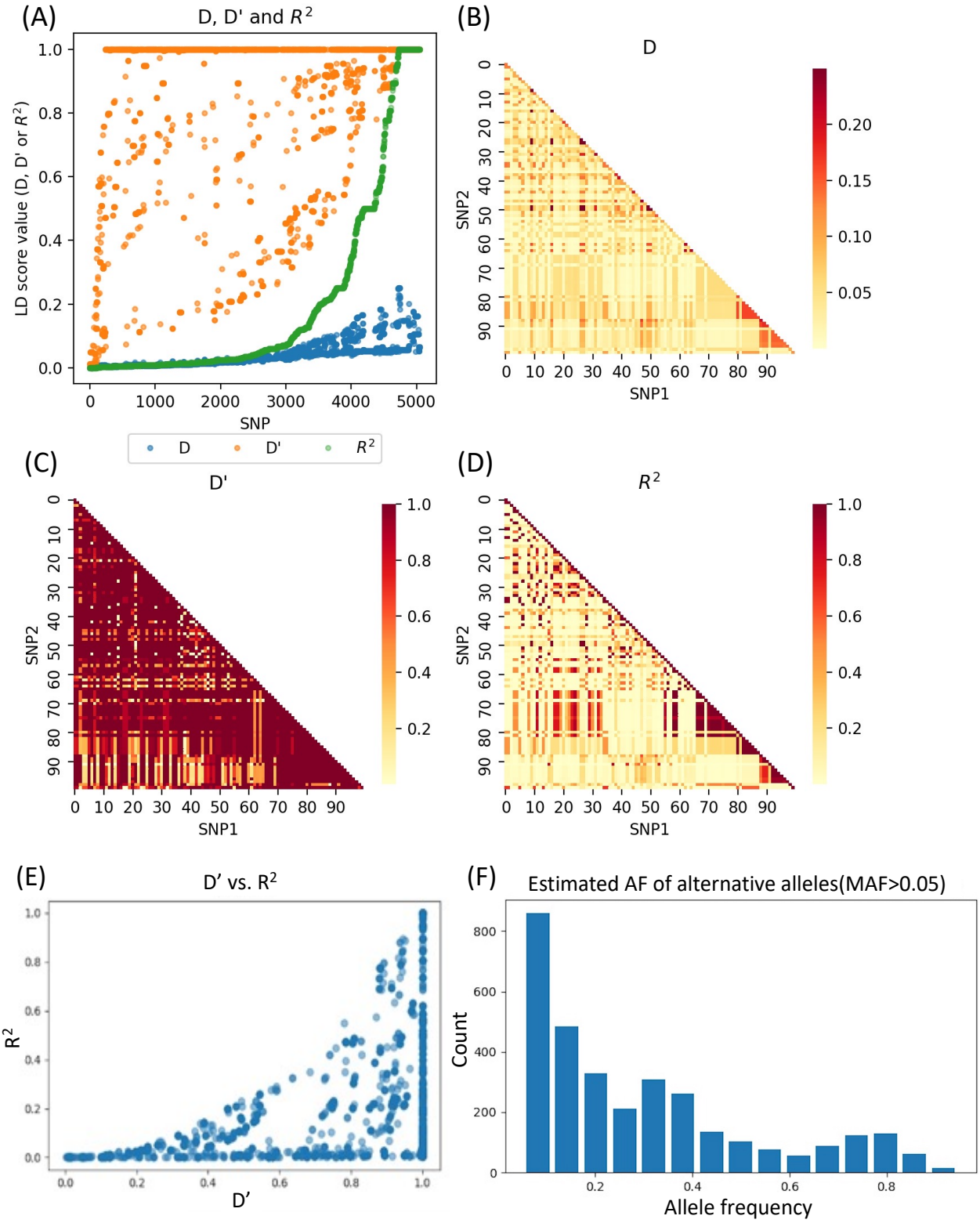

(A)
D, D' and R^2
LD score value (D, D' or R^2)
SNP
D
D'
R^2
(B)
D
SNP2
SNP1
(C)
D'
SNP2
SNP1
(D)
R^2
SNP2
SNP1
(E)
D' vs. R^2
R^2
D'
(F)
Estimated AF of alternative alleles(MAF>0.05)
Count
Allele frequency

Figure 3

4. Principal component analysis

As showed in figure 4A, the first PC captures the most variance of the data, the second PC captures the second to most variance, and so on. Figure 4B (PC1 vs. PC2) and 4C (PC1 vs. PC3) both separate substructures of the population well, most like due to difference of PC1 values among individuals. PC2 vs. PC3 (figure 4D) does not distinguish pattens clearly.

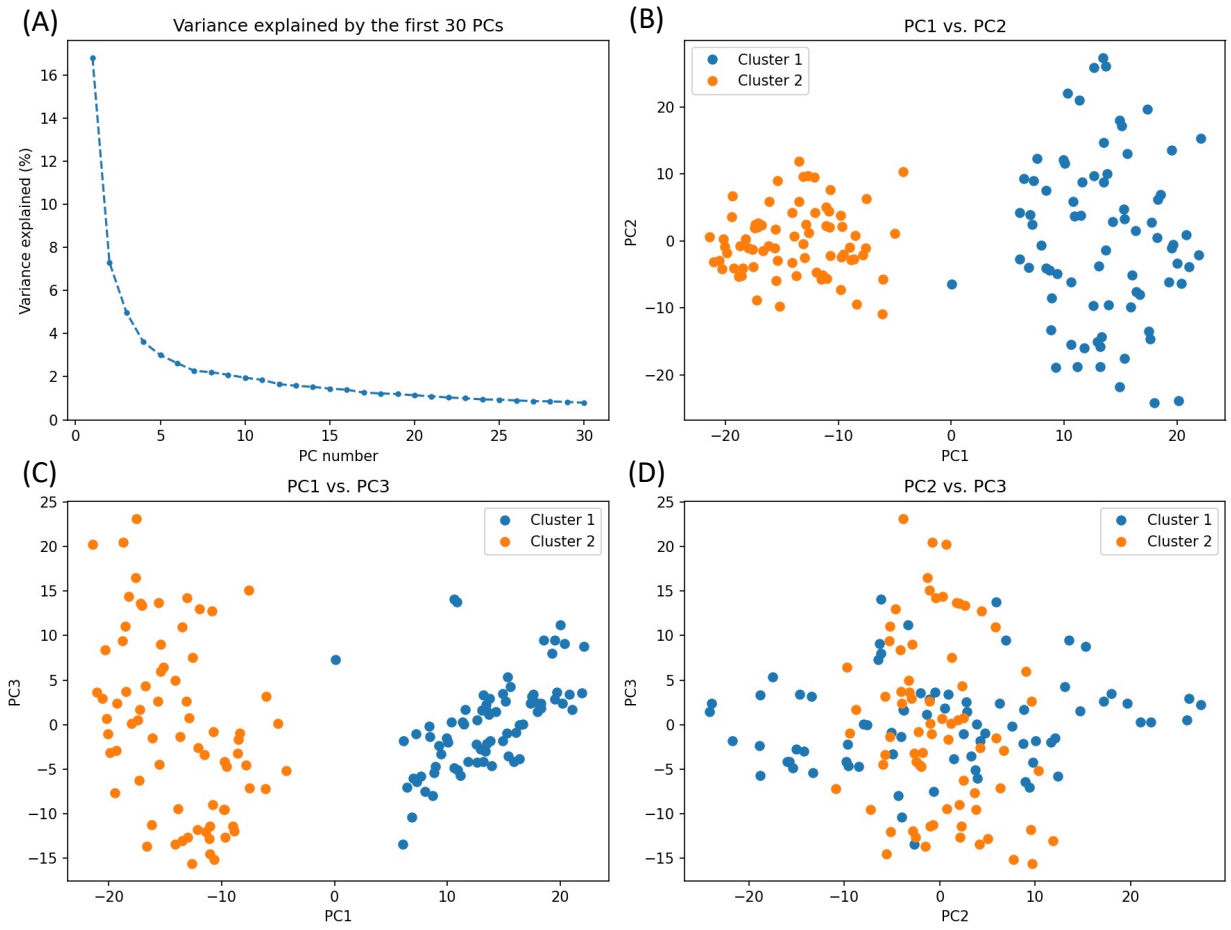


Figure 4