Summary: Principal components analysis corrects for stratification in genome-wide association studies

Liu Huihang
SA18017026
QQ: 184050473
MAIL: huihang@mail.ustc.edu.cn

1. Theoretical Summary

1.1 Problem

Biostatisticians study the associations between diseases and genotypes to reveal the genes related to the occurrence, development and treatment of diseases. This is an intuitive approach, but problems arised in the transcontinental research is very important.

Population stratification (allele frequency differences between cases and controls) owing to systematic ancestry differences can cause spurious associations in GWAS.

For example, assume that using chopsticks is a disease and frequency of a ramdon allele A is higher among Chinese than other countried, then we may come to the conclusion that allele A is highly corelated with the disease. But we know the allele A is chosen randomly and is nothing to do with the phenotype.

1.2 Solution

The paper gives a useful method to correct stratification as described following.

1. Apply PCA to genotype data to infer continuous axes of genetic variation;
2. Adjust genotypes and sphenotypes by amounts attributable to ancestry along each axis;
3. Compute association statistics using ancestry-adjusted genotypes and phenotypes.

1.3 Advantages

Most importantly, this method can correct spurious associations due to systematic ancestry differences. At the same time, PC scores provide the most useful description of within-continent genetic variation. What's more, the results are insensitive to the number of axes inferred. In addition, the approach is computationally tractable on a genome-wide scale.

2. Simulation Summary

In the following, I will review the simulated experiment introduced in the paper and try to reproduce the results by myself.

2.1 Settings

Data is generated at 100000 random SNPs for 500 cases and 500 controls, with 60% of the cases and 40% of the controls sampled from population 1 and the remaining cases and controls sampled from population 2. Then, allele frequencies for population 1 and population 2 were generated using the Balding-Nichols model with $F_{ST} = 0.01$.

They simulated three categories of candidate SNPs to compare the effectiveness of different stratification correction methods. The first category contains random SNPs with no association to disease. The second category contains differentiated SNPs with no association. (This category is discussed in ”problem”, it is different between stratifications.) The third category contains causal SNPs which cause the disease actually. These categories have different parameters when using Balding-Nichols model.

2.2 Correct stratification and check association

This is a very simple approch, as we described above.

Firstly, they use PCA to get only top 1 orthogonal score. Then, checking the association using 1st score of PCA.

2.3 Reuslts

The simulation results clearly support the author’s argument.

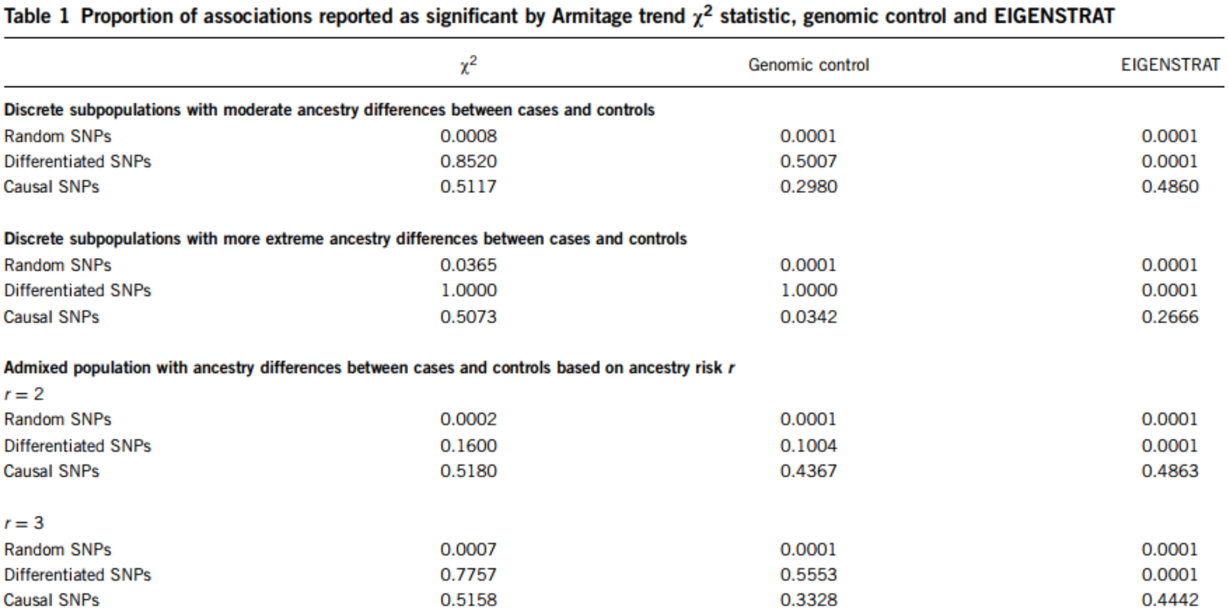

Table 1 Proportion of associations reported as significant by Armitage trend χ^2 statistic, genomic control and EIGENSTRAT

	χ^2	Genomic control	EIGENSTRAT
Discrete subpopulations with moderate ancestry differences between cases and controls			
Random SNPs	0.0008	0.0001	0.0001
Differentiated SNPs	0.8520	0.5007	0.0001
Causal SNPs	0.5117	0.2980	0.4860
Discrete subpopulations with more extreme ancestry differences between cases and controls			
Random SNPs	0.0365	0.0001	0.0001
Differentiated SNPs	1.0000	1.0000	0.0001
Causal SNPs	0.5073	0.0342	0.2666
Admixed population with ancestry differences between cases and controls based on ancestry risk *r*			
$r = 2$			
Random SNPs	0.0002	0.0001	0.0001
Differentiated SNPs	0.1600	0.1004	0.0001
Causal SNPs	0.5180	0.4367	0.4863
$r = 3$			
Random SNPs	0.0007	0.0001	0.0001
Differentiated SNPs	0.7757	0.5553	0.0001
Causal SNPs	0.5158	0.3328	0.4442

The table above contains three three types of stratification, they have different kinds of ancestry differences between cases and controls.

The numbers reported the proportion of candidate SNPs in each category at which each method reports a significant association with $P < 0.0001$ with many independent simulations and a large number of SNPs.

It shows that the proposed method EIGENSTRAT can distinguish and correct spurious associations better than using χ^2 test directly and better than genomic control method.

2.4 My Attempt

The theoretical method seems very easy, but problems occurres when I try to reproduce the result.

But the simulation is more complicate than throry. I didn't get a appropriate results.