**Abstract**

**Introduction**

Recently, three research groups simultaneously proposed a new kind of -like approaches in detecting loci under selection in structured populations (Chen *et al.*, 2016; Duforet-Frebourg *et al.*, 2016; Kevin J. Galinsky *et al.*, 2016). Chen *et al* (2016) further found a simple-marker regression form of the method, and because of its similarity to genome-wide association studies (GWAS), we proposed the name EigenGWAS. The EigenGWAS has also been adopted statistical analysis packages, such as pcadpt (Luu, 2017) and sommer (Covarrubias-Pazaran, 2016).

Compared with the conventional , which asks a cutting-clear definition for subpopulation but probably unavailable especially for continuous sampling for the population of question, the newly EigenGWAS can bypass this cumbersome task. These methods found their applications in detecting loci under selection, such as in human genetics (Kevin J. Galinsky *et al.*, 2016; Chaput *et al.*, 2014; Shen *et al.*, 2016; Liu *et al.*, 2018; Lee *et al.*, 2017), ecology (Kim *et al.*, 2017; Bosse *et al.*, 2017; Armstrong *et al.*, 2018), breeding population (Ma *et al.*, 2016; Liu *et al.*, 2017; Zhao *et al.*, 2018), and other applications – especially various re-sequencing populations that are embarking.

We have observed and analyzed more data and gained further understanding of the method, in this study we are going to help further refine the application of the new method as below:

1. Reconcile the three papers aforementioned (Chen *et al.*, 2016; Duforet-Frebourg *et al.*, 2016; Kevin J Galinsky *et al.*, 2016), in which EigenGWAS will be a reasonably elementary form of the three methods.
2. For the one-degree-of-freedom chi-square test that is included in testing the statistical significance of the loci under selection, non-centrality parameter of the test statistic is defined. Consequently, statistical power of the method can be conducted.
3. Technical adjustment for population structure is verified by choosing between eigenvalue, which will be revealed having a mixture distribution and consequently not always suitable for the correction of population structure, and genomic inflation factor (Devlin and Roeder, 1999).
4. Caveats for incompletely sampled populations, especially for the population under strong selection, is discussed, and empirical evidence to characterize those populations is also presented.

For the brevity of the text, we will use EigenGWAS to represent the -like approaches.

**Materials and methods**

**Characterizing the sample**

Given the genotypic matrix , an genotypic matrix in which is the sample size and the number of markers. By setting the current population as the reference population, we have the additive genomic relation matrix , in which represents the standardized form of ; in particular between individual and , . This matrix has been widely used in genetic analysis, such as genetic prediction (VanRaden, 2008).

Two useful statistics can be inferred from , the off-diagonal elements of the matrix. In theory, if it is a randomly sampled population, , and upon how inbred the samples are, the lower bound of is . So, we can have an estimate of the samples, .

In addition, the variation of , in which is the squared Pearson’s correlation for a pair of markers—the most commonly used metric for the linkage disequilibrium (LD) of a pair of markers (Devlin and Risch, 1995). We can further define (Chen, 2014). It is obviously that is between 1, every marker is in full LD between each other, and , no LD between any pair of markers. It has been well-known in evolution that the when there is hitch-hicking effect the LD will be increased for the markers within the genomic region under selection (Maynard and Haigh, 1974).

So, the mean and variance of can help elucidate some basic characters of the sample of study. It is expected for inbred populations, such as Arabidopsis (Alonso-Blanco *et al.*, 2016), , while for random mating population, .

**The non-centrality parameter for EigenGWAS**

After eigenvalue decomposition analysis of , we have two matrices , which is an matrix, and the diagonal matrix for the ordered eigenvalues. A linear model, EigenGWAS, can be constructed as

(Eq 1)

in which is the eigenvector/column, the grand mean, the SNP/column, and the residual. We are interested in the regression coefficient . In Appendix, we can derive the genetic interpretation of .

Under the null distribution of no association, . When the locus is associated with , the NCP is approximately . and , and are the numbers of samples that have positive and negative values in .

There are many kinds of , we only introduce two forms here that is closed related to EigenGWAS. and . The connection between the NCP and can be, after some algebra, established

Given expression above, we can learn that

* EigenGWAS only see “two subgroups” that at the two sides of the point zero.
* the power is proportional to sample size , and maximized when .
* NCP is zero if .

If we are interested in construct a statistical hypotheses test for whether a locus is under selection, it is better to control the background information, such as genetic drift. Because the conventional statistic contains both genetic drift and selection, we are aim to control genetic drift.

**Eigenvalue and genomic inflation factor**

As the distribution of the test statistic is known, it is possible to coin the test statistic further, for example to test the selection by controlling genetic drift. For a locus of interest, is an overall measurement for genetic differentiation, and in order to control for genetic drift, the adjustment such as genomic inflation factors can be used.

Originally, the three papers all used the eigenvalue that is associated with to adjust the test statistic, so the de facto test statistic is (Chen *et al.*, 2016; Duforet-Frebourg *et al.*, 2016; Kevin J Galinsky *et al.*, 2016)

Of note, in Duforet-Frebourg et al, it was a scan statistic and its distribution was not that well characterized. Later on, Luu et al (Luu, 2017) adopted a test statistic and has been updated it to

in which is the genomic inflation factor (Devlin and Roeder, 1999). In practice, some authors also adopted in data analysis (Bosse *et al.*, 2017).

As observed by us (Table 1 in our paper), we found the eigenvalue and the genomic inflation factor were very similar and often highly correlated along many scanned eigenvectors, and we adopted eigenvalues as a way to control for genetic drift (Chen *et al.*, 2016). It is very interesting to see these subtle difference. Because we know both eigenvalue and genomic inflation factor can reflect population structure, what is the consequence and implication for the respective application in data analysis. In particular, whether it will influence the statistical power because of the choice of the technical adjustment for population structure.

It is known that there is connection between and eigenvalue, in particular for human population the largest eigenvalue often reflects genetic drift for the population geographically distributed (Patterson *et al.*, 2006; McVean, 2009). We shows that an eigenvalue is a mixture distribution, and at least can be written as two components

in which is the proportion of the loci under selection the genetic differentiation of that is and the rest of the loci under genetic differentiation . .

It could be anticipated that when is a small number, often small for such as human population, the median of the observed reflects because the signals soak of selection signals are ranked at the high end of the observed . So, adjust the test statistic with will help control genetic drift. However, if is small and the selection is moderate, we are not expected to see differs from much.

Upon the genetic architecture of the population, these two adjustment of the test statistic differ dramatically. The interpretation of , upon the populations, may find its interpretation of not. For human population, the top eigenvalue often indicate geographic isolation of the samples. A very great example is great tit samples. For great tit samples, the first eigenvalue reflects the genetic isolation of the samples by the ocean, and the second eigenvalue the separation between two Netherland locations (Bosse *et al.*, 2017). Need another scan for PC2.

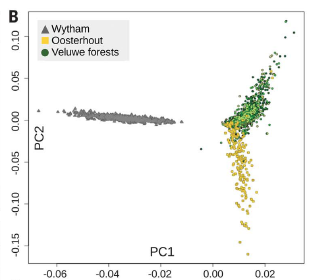


Table 1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Species | Population |  |  | [maf>0.1] |  |  | Source |
| Human | HapMap |  |  |  |  |  |  |
|  | POPRES | 2,466 | 2,450 | 615,494 | 119,407 |  |  |
|  | UK Biobank |  |  |  |  |  |  |
| Great tit | Great tit | 3015 | 5740.75 | 484370 | 6212.54 |  |  |
| Dog |  | 1792 | 1447.35 | 164,164 | 47.54 |  |  |
| Plant | Maize HapMap3 | 879 | 902 | 335848 | 104.12 |  |  |
|  | Arabidopsis 294 | 294 | 293 | 156,744 | 394.73 |  |  |
|  | Arabidopsis 1135 |  |  |  |  |  |  |
|  | 3K rice | 357 [which 357] | 348 | 2,702,622 | 8 |  |  |

For eigenvalues , but it is not necessary the case for the genomic inflation factors for example maize HapMap3.

**Simulation**

Validation of the NCP

We validated our NCP by the simulation below. We simulated a locus with allele frequency of 0.4 in subpopulation 1 and 0.6 in subpopulation 2, and the subpopulations had equal sample size of 500. According to the NCP was 82.51, and we replicated the sample 500 times. Then in a quantile-quantile plot, we compared the observed chisq test statistic with the one sampled from a chisq distribution with NCP of 82.51.

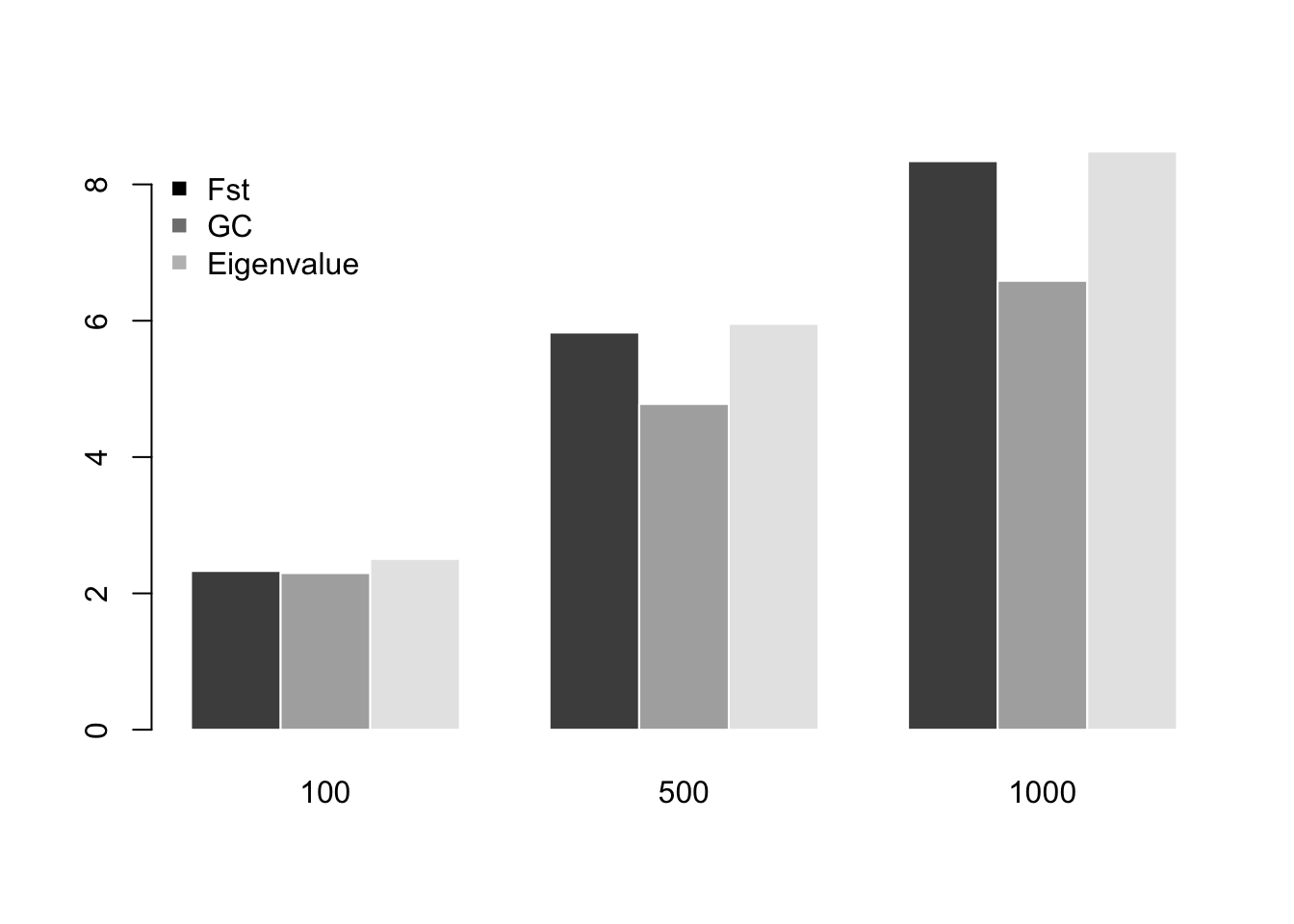
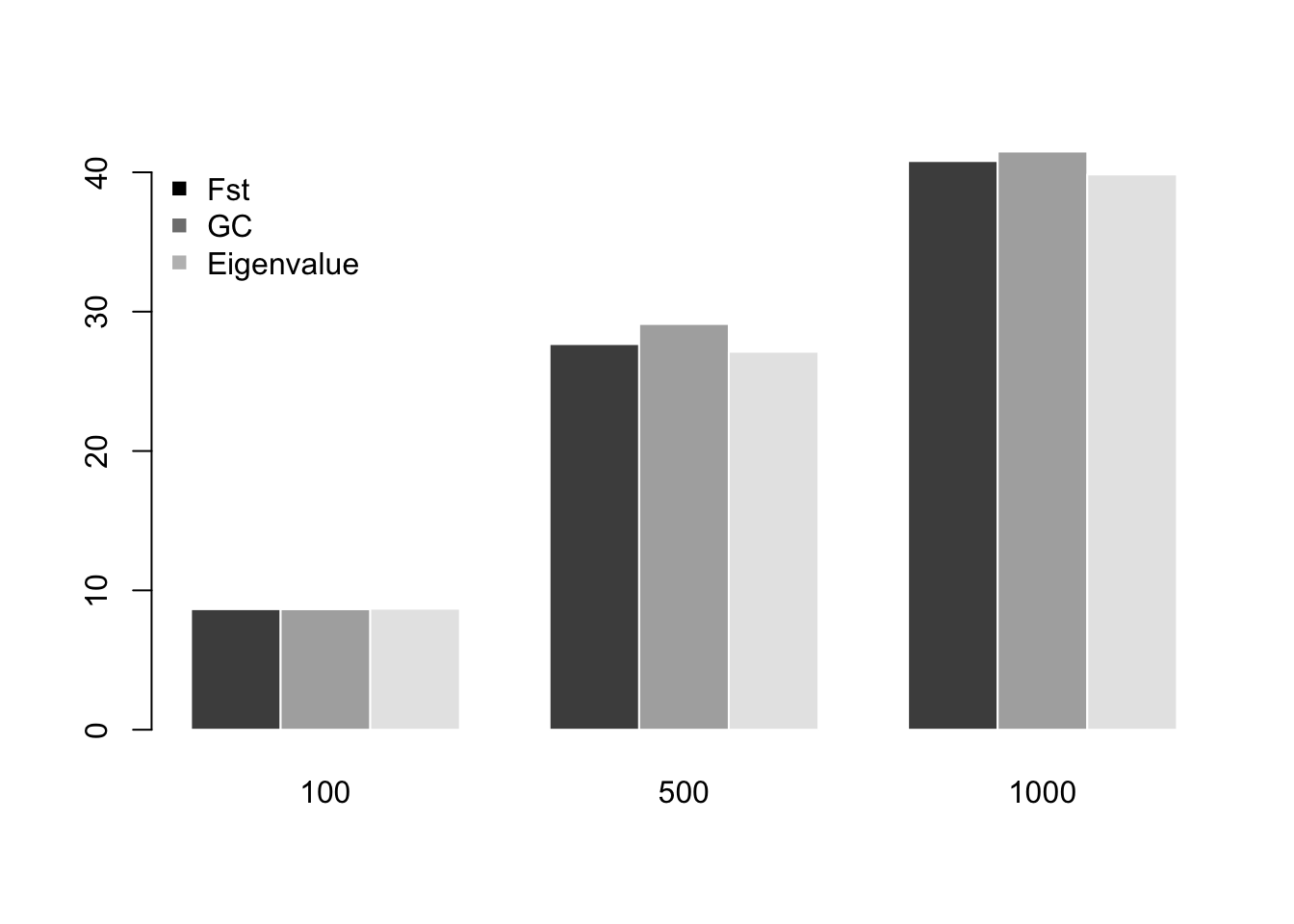
Alternatively, we changed the sample sizes to 300 vs 700, and the observed fitted the expectation well too.



The mixture distribution of eigenvalue

We show that the distribution of the eigenvalue can be a mixture distribution. We simulated population 1 the eigenvalue was dominated by a single , which mimicked genetic drift. The allele frequencies of the two subpopulations were sampled from , in which is the allele frequency of their common ancestor, and ; in the second simulation, the loci were partitioned into two sets, the first set was driven by genetic drift sampled from , and the second set from .

For the simulated samples, we conducted EigenGWAS. Their , , and as illustrated below



**Real data analysis**

UKBiobank

**Discussion**

**Acknowledgements**

This study was supported by the National Natural Science Foundation of China (31771392 to G-B C, and 31871707 to H-M X) and Zhejiang Provincial People’s Hospital Research Foundation (ZRY2018A004 to G-B C). The funders played no role in experimental data and data analysis.

**Reference**

Alonso-Blanco,C. *et al.* (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, **166**, 481–491.

Armstrong,C. *et al.* (2018) Genomic associations with bill length and disease reveal drift and selection across island bird populations. *Evol. Lett.*, **2**, 22–36.

Bosse,M. *et al.* (2017) Recent natural selection causes adaptive evolution of an avian polygenic trait. *Science*, **358**, 365–368.

Chaput,J.-P. *et al.* (2014) Findings from the Quebec Family Study on the Etiology of Obesity: Genetics and Environmental Highlights. *Curr. Obes. Rep.*, **3**, 54–66.

Chen,G.-B. *et al.* (2016) EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. *Heredity*, **117**, 51–61.

Chen,G.-B. (2014) Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. *Front. Genet.*, **5**, 107.

Covarrubias-Pazaran,G. (2016) Genome-Assisted prediction of quantitative traits using the R package sommer. *PLoS ONE*, **11**, e0156744.

Devlin,B. and Risch,N. (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, **29**, 311–22.

Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Duforet-Frebourg,N. *et al.* (2016) Detecting genomic signatures of natural selection with principal component analysis: Application to the 1000 genomes data. *Mol. Biol. Evol.*, **33**, 1082–1093.

Galinsky,K.J. *et al.* (2016) Fast principal components analysis reveals independent evolution of ADH1B gene in Europe and East Asia. *Am. J. Hum. Genet.*, **98**, 456–472.

Galinsky,K.J. *et al.* (2016) Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure. *Am. J. Hum. Genet.*, **99**, 1130–1139.

Kim,K.-W. *et al.* (2017) A sex-linked supergene controls sperm morphology and swimming speed in a songbird. *Nat. Ecol. Evol.*

Lee,S.H. *et al.* (2017) Using information of relatives in genomic prediction to apply effective stratified medicine. *Sci. Rep.*, **7**, 42091.

Liu,S. *et al.* (2018) Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell*, **175**, 347–359.

Liu,Z. *et al.* (2017) Comparison of Genetic Diversity between Chinese and American Soybean (Glycine max (L.)) Accessions Revealed by High-Density SNPs. *Front. Plant Sci.*, **8**, 2014.

Luu,K. (2017) pcadapt : an R package to perform genome scans for selection based on principal component analysis. *Mol. Ecol. Resour.*, **17**, 67–77.

Ma,X. *et al.* (2016) Genome-Wide Association Study for Plant Height and Grain Yield in Rice under Contrasting Moisture Regimes. *Front. Plant Sci.*, **7**, 1801.

Maynard,J. and Haigh,J. (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.*, **23**, 23–35.

McVean,G. (2009) A genealogical interpretation of principal components analysis. *PLoS Genet.*, **5**, e1000686.

Patterson,N. *et al.* (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.

Shen,Q.K. *et al.* (2016) Was ADH1B under Selection in European Populations? *Am. J. Hum. Genet.*, **99**, 1217–1219.

VanRaden,P.M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci.*, **91**, 4414–4423.

Weir,B.S. and Cockerham,C.C. (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, **38**, 1358–1370.

Zhao,Q. *et al.* (2018) Identifying Genetic Differences between Dongxiang Blue-shelled and White Leghorn Chickens using Sequencing Data. *G3*, **8**, 469–476.

**NCP for the additive model**

**For random mating populations**

For linear model ,

In which the standardized eigenvector of interest, codes for the additive effect, the regression coefficien, and the residual of the model. The regression coefficient can be expressed as

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | | | Marginal probability | |
|  |  |  |  |  |
| Subpop 1 |  |  |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Subpop 2 |  |  |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |  |  | |

Under the condition that and , and the sampling variance further shrinks to zero, the NCP becomes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | | | Marginal probability |
|  |  |  |  |  |
| Subpop 1 |  |  |  |  |  |
|  |  |  |
| Subpop 2 |  |  |  |  |  |
|  |  |  |
|  |  |  |  |  |  |

we have

The sampling variance of is , so the z-score statistic is

and , with ncp of .

**For inbred lines**

For linear model ,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | | Marginal probability |
|  |  |  |  |  |
| Subpop 1 |  |  |  |  |
|  |  |
| Subpop 2 |  |  |  |  |
|  |  |
|  |  |  |  |  |

then the NCP approximates to .

So, the NCP for inbred population is the half of that of a random mating population.

**NCP for dominance model**

For linear model ,

In which the standardized eigenvector of interest, codes for the dominance effect, the regression coefficient for the dominance effect, and the residual of the model. The regression coefficient can be expressed as

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | | | Marginal probability |
|  |  |  |  |  |
| Pop 1 |  |  |  |  |  |
|  |  |  |
| Pop 2 |  |  |  |  |  |
|  |  |  |
|  |  |  |  |  |  |

We have ,

in which .

The sampling variance of is , so the z-score statistic is

and , with ncp of . In comparison, for the additive model, the ncp approximates to .