Principle Component Analysis and Mixed Linear Model on cross-population genetic correlations

Zhang Leyi[1]

[1]MSc in BiRC, Aarhus University, Aarhus, Denmark

2023-01

Abstract

Genome-wide Association Studies(GWAS) can be used for genetic analysis on the performance of phenotypes in multiple populations. While in an admixed population data, due to population stratification, the accuracy of the result would be limited.

Principle Component Analysis is a method in dimension reduction and visualization of sample overlap due to the common origin or ancestry is considered evidence of identity. However, PCA is only a visualization method and cannot directly provide statistical conclusions about genetic correlation. If the individuals are highly influenced by genetic relatedness, there would be a lot of false-positive results.

By using the mixed linear model analysis approach, the contribution of genetic relatedness to phenotypes in multiethnic genotype data can be accurately explored. This approach can be used to identify population-specific genetic variants associated with complex traits.

In this analysis, I tested the limitation of GWAS and showed the contributions of the joint analysis with PCA and MLM.

Keywords: GWAS, PCA, MLM, population stratification, genetic relatedness

1 Introduction

Genome-wide Association Studies(GWAS) can be used for genetic analysis on the performance of phenotypes in multiple populations. This refers to a research method that conducts GWAS in multiple populations and integrates the results of GWAS in multiple populations. It can be used to find trait-associated loci that co-occur in multiple populations, Zeggini and Ioannidis (2009). However, GWAS also has some drawbacks, such as data standardization, correction of population stratification, etc.

Human phenotypes are affected by many factors, including genetic factors, environmental factors, and biological factors. In data with population stratification, if we only perform GWAS, there will be biased results. In population structure analysis, Principle Component Analysis(PCA) can help determine the distribution of the population, including population density and the distribution of population. However, PCA analysis has obvious deficiencies, such as PCA can only find the most important variables, but not all variables. This means that if some variables that have a small impact on the population structure are not found, PCA cannot fully reflect the characteristics of the population structure. In addition, PCA can only explain the influence of environmental aspects on the performance of phenotypes, but not the influence of genetic factors (genetic relatedness). In the analysis of cross-population datasets, it is widely concerned that false-positive genetic signals come from inflated test statistics of population stratification, Elhaik (n.d.).

In this experiment, the data are Genomes with using short-read DNA sequencing data **1000 Genome project(1000G)** of 2504 people from 26 countries, and their race and gender are also accurately marked, Sudmant et al. (2015). However, these data do not include their traits, so here I use the LDAK model to simulate Phenotypes for these 2504 individuals.

Firstly, I analyzed the data with PCA. The data were reduced to a small number of dimensions called Principal Components (PCs); each described a reduction in the proportion of genomic variation. The genotypes were then projected onto the space spanned by the PC axis, and in this visualization sample overlap due to the common origin or ancestry is considered evidence of identity. The most attractive property of PCA is the inter-cluster The distance is said to reflect the genetic and geographic distance between them, Harris, Brunsdon, and Charlton (2011)

However, excessive reliance on PCA analysis often leads to unreliable results. PCA may have a biasing effect in genetic studies [6]. I used MLM to further analyze the results of the PCA analysis. MLM can effectively control the multivariate relationship of gene expression and can better reflect the complexity of gene expression.

Overall, my research aims to highlight the limitations of GWAS and the benefits of using both PCA and MLM to analyze on population structure, which may lead to a better understanding of the genetic and demographic factors influencing population structure.

2 Methods and Results

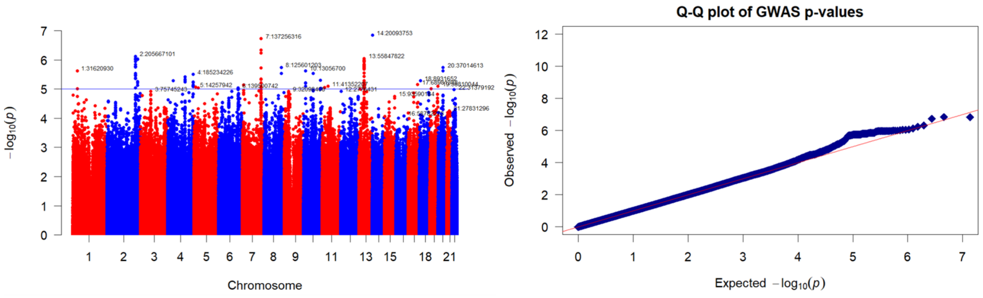
2.1 Quality Control

Plink is a software tool for gene association analysis, which can be used to process a large amount of genotype data and perform gene association analysis. Plink is widely used in human genomics research and can handle many common genotype data file formats, Purcell et al. (2007).

The original data is in .bed, .bim, .fam format. I used Plink software to integrate it and perform Quality Control on the original data. MAF is set at 0.05, Marees et al. (2018). For 12123491 variants obtained from .bim, 5259242 exceeding the threshold are deleted.

2.2 Genome-wide Association Studies

I use simulated phenotypes to run for GWAS analysis, and the phenotypes were simulated by LDAK software. I used GEMMA software to obtain the kinship matrix and perform association analysis. GEMMA (Genome-wide Efficient Mixed-Model Association) is a software tool for gene association analysis. GEMMA uses the MLM association analysis method, which can simultaneously consider the influence of multiple gene loci, and has high precision and complexity, Zhou and Stephens (2012). Then I drew the Manhattan plot and the QQ plot for the obtained association data, as shown in Figure.1.



(a) Manhattan plot of 1000G

(b) QQ plot of 1000G

Figure 1: **GWAS results of 1000G data**

(a)Manhattan plot shows the significance of each variant associated with the phenotype. Each point represents a single nucleotide polymorphism (SNP), and SNPs are ordered on the x-axis according to their genomic position. The y-axis indicates their association strength, measured as P-values transformed by $-\log_{10}$. The blue line marks the genome-wide significance threshold for $P < 5 \times 10^{-8}$. (b)QQ plot shows the distribution of expected P values versus observed P values. The P value is the result corrected by $-\log_{10}$ transformation. Deviations from expectations under the null hypothesis (red line) indicate a true causal effect or insufficiently corrected population stratification.

However, a large number of SNPs in the Manhattan plot were of high statistical significance(above the blue line), suggesting the possibility of population stratification. The p-values of the QQ plot were relatively evenly distributed, which was inconsistent with the results of the Manhattan plot, which reflects the possibility of false positives brought about by population stratification. Therefore, we need to consider the existence of population stratification further.

2.3 Principal Component Analysis

Principal Component Analysis(PCA) is a commonly used method for dimension reduction. Its purpose is that, by finding the principal components of the data, and transform the original high-dimensional data into low-dimensional data, with preserving the total variance of the original data as much as possible.

When using PCA for data dimension reduction, we usually do the following steps:

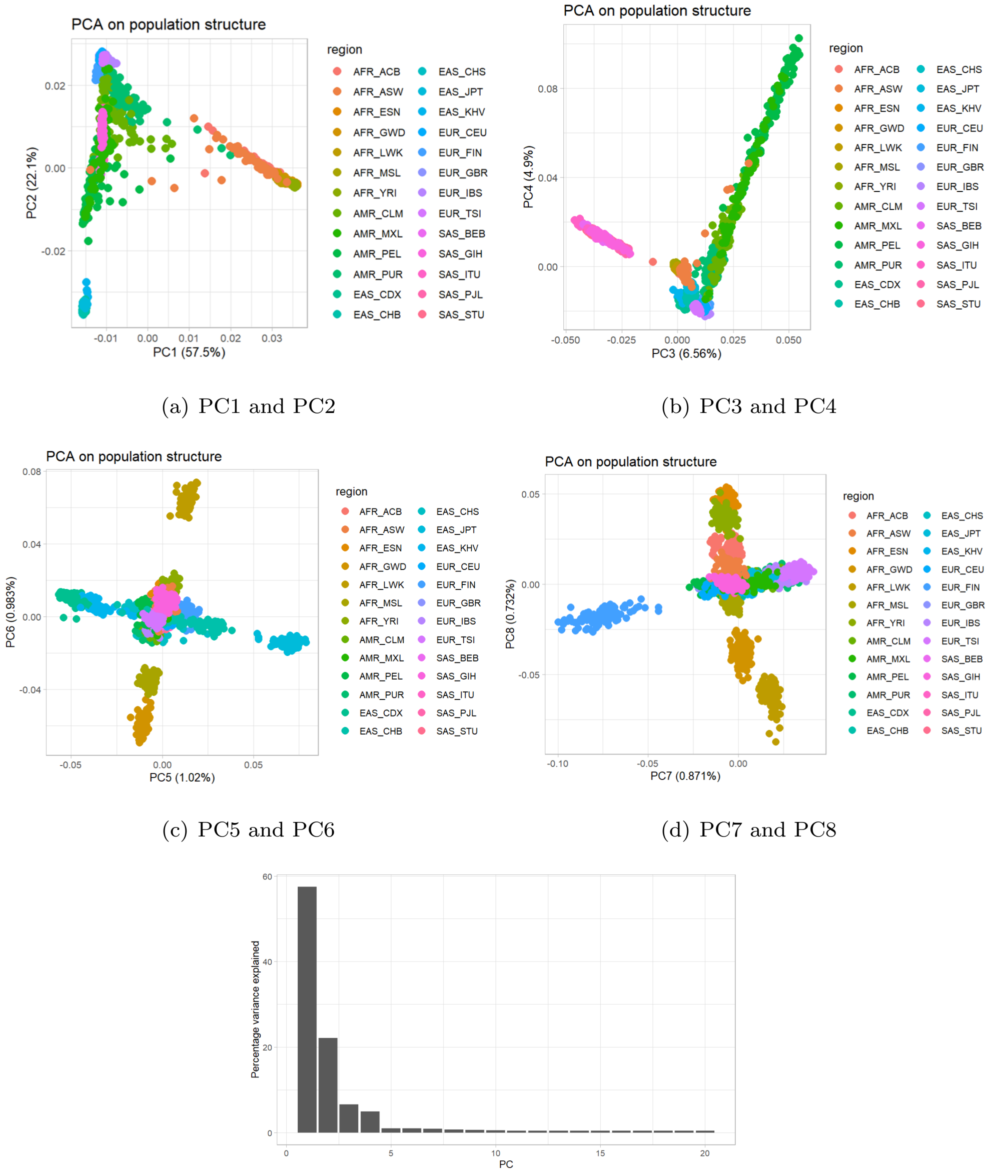
1. Centralize the data, that is, subtract its mean value from each dimension, so that the mean value of the data is 0.
$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)},$$
$$X_{cent} = X - \mu,$$
where μ is the mean vector of all samples, and X_{cent} is the centred data matrix.
2. Calculate the covariance matrix. For an nm data matrix X, its covariance matrix is:
$$C = \frac{1}{m-1} X_{cent}^{T} X_{cent}$$
3. Compute the eigenvalues and eigenvectors of the covariance matrix. The eigenvalue represents the variance of the data in the direction of the eigenvector, which is the vector corresponding to the eigenvalue.
$$Cv = \lambda v,$$
where C is the covariance matrix, v is the eigenvector, and λ is the eigenvalue.
4. Sort the eigenvalues in descending order. Select the eigenvectors corresponding to the top k largest eigenvalues, and denote the matrix composed of these k eigenvectors as Vk. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$
5. Project the data matrix X onto the matrix Vk to obtain a new low-dimensional data matrix Y.
$$Y = XV_k,$$
where Y is a low-dimensional data matrix, X is the original high-dimensional data matrix, and V_k is a matrix composed of eigenvectors corresponding to the first k largest eigenvalues.

When studying the genetic structure of a population, PCA can be used to visualize genetic data, which helps us better understand genetic differences among populations. For a set of genetic data including population, use PCA to reduce the data dimension, and then use point plot to visualize PC1 and PC2. In such a plot, the populations will be distributed along the direction of the principal components. If there are obvious genetic differences and population stratification among the populations, there will be obvious grouping in the point plot.

2.4 Finding population stratification by PCA

Population stratification refers to the group differences caused by certain factors in the population, which may include the genetic characteristics of the population, lifestyle, environment, etc.

In this analysis, the population information of the data was annotated. In the point plot of PCA, Figure.2, the population data are clearly stratified. The points of the same population group are roughly concentrated, which shows that the genetic differences among populations can lead to the emergence of population stratification. However, this plot also shows that there are obvious differences within some populations, e.g. AMR_CLM has a large difference on PC1. This indicates that the amount of genetic information of some groups in this population is quite different from that of residents in the same population, and there is a subpopulation with large differences in genetic characteristics. AMR_CLM stands for Colombian, and Colombia is a country of immigrants, and there are widely distributed people of mixed race, whites, Arabs, Africans, indigenous peoples, etc. in the country, which proves the above point.



(a) PC1 and PC2

(b) PC3 and PC4

(c) PC5 and PC6

(d) PC7 and PC8

(e) Percentages of variance explained by each PC

Figure 2: **PCA of population structure.**

8 PCs with the highest scores of the 1000 Genomes project. Percentages of variance explained by each PC are shown in (e)

Then I used a boxplot to draw the phenotypes of seven populations, Figure.3. It can be seen that the differences of phenotypes within the populations are also quite significant, and there is an obvious level of variation, which means there are differences caused by genetic relatedness within the population. However, PCA is only a visualization method and cannot directly provide statistical conclusions about genetic correlation. I used MLM combined with genetic association analysis for genetic correlation analysis. I mainly used heritability to estimate the proportion of genetic factors affecting traits.

$$h^2 = \frac{C_{genetic}}{C_{total}},$$

where h^2 is the estimated value of heritability, $C_{genetic}$ is the effect of genetic factors on the phenotypes, and C_{total} is the total variation of the trait.

The total variation of traits is not only affected by genetic factors, but also by environmental factors.

$$C_{total} = C_{genetic} + C_{environmental}$$

Therefore, the formula of heritability can be written as:

$$h^2 = \frac{C_{genetic}}{C_{genetic} + C_{environmental}}$$

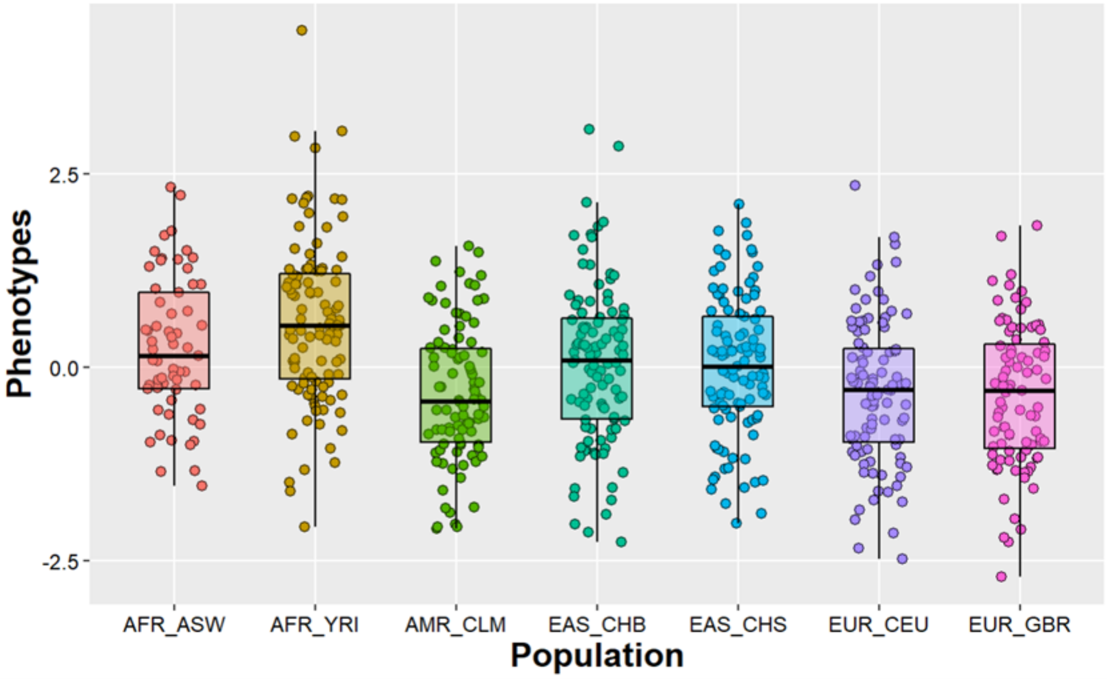


Figure 3: **Boxplot of genetic and environmental factors.**

Seven populations are extracted from the 26 populations of the 1000 Genome Project. Each box represents a population, and the middle black line is the median. AFR_YRI presents a normal distribution, AFR_ASW and AMR_CLM are positive skew, and the other four are negative skew.

2.5 Kinship Matrix

Kinship Matrix is a tool commonly used in genetics research, which can be used to represent the genetic similarity among samples. It is a matrix where each element represents the genetic similarity between two samples.

I used GEMMA to calculate the Kinship Matrix. GEMMA is a model for solving genetic correlation(kinship matrix). I used GEMMA software to calculate Kinship Matrix. GEMMA is a software tool for solving genetic correlation (kinship) matrix. GEMMA utilizes mixed linear models to calculate associations between genes. The advantage of GEMMA is that it can effectively deal with the impact of population structure and genetic relatedness, thereby improving the statistical effectiveness of GWAS, Zhou and Stephens (2014).

After calculating the kinship matrix among 2504 individuals, I randomly selected 50 individuals and built a subset, as shown in Figure.4. Each value of the matrix represents the genetic similarity between two samples, the diagonal (lower left to upper right) is all 1, and is symmetrical along the diagonal axis. And the lighter the colour, the higher the genetic similarity. It can be seen from the figure that the genetic similarities between individuals #669, #771, #777, and #837 are relatively high, and the similarity between #771 and #777 is higher than that between #771 and #669. Individuals #669, #771, #777 are from EAS_KHV population, and 837 is from EAS_CDX. This shows that in the same population, the genetic similarity is generally higher than that of different populations, but there are also factors such as geographical proximity and population migration that lead to the higher similarity between some populations, which further proves that the genetic differences between races lead to Emergence of population stratification. However, there are some differences in the genetic similarity between individuals belonging to the same EAS_CDX ethnic group, which also shows that to a certain extent, there will be some populations within the same ethnic group whose genetic similarity is higher than other populations of the same ethnic group.

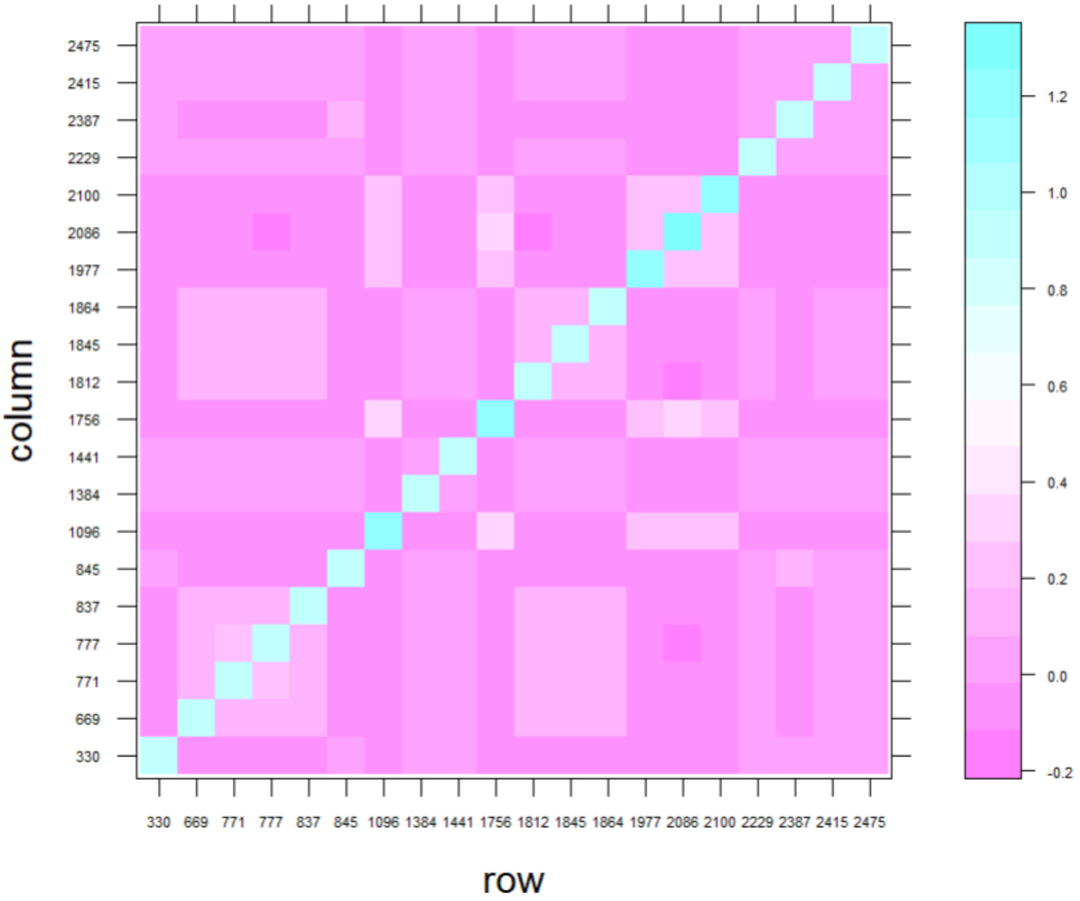


Figure 4: **Heatmap of kinship matrix.**

The kinship matrix by GEMMA provides insight into the relationship between each genotype in the 1000 Genome Project population. The picture is a matrix composed of 50 people randomly selected from 2504 people, which is represented by the heatmap. The genetic relatedness of genotypes is represented by different colours.

I then further analyzed genetic variation among individuals. I use restricted maximum likelihood (REML) to explore the relationship between genotypes and traits. REML is a method for estimating the parameters of fixed and random effects, Molenberghs and Verbeke (2000).

$$L(b) = (2\pi)^{-\frac{p}{2}}\,|R|^{-\frac{1}{2}}|X'X|^{-1}|y - Xb|^2,$$

where $L(b)$ represents the likelihood value of the parameter b, b is the parameter vector of the fixed effect and random effect to be estimated, p is the number of parameters of the fixed effect, R is the parameter vector of the random effect Covariance matrix, X is the design matrix of explanatory variables, and y is a vector of explained variables.

I used LDAK software, with kinship matrix, to conduct REML analysis on all population data, and got $Heritability = 0.533367$ for all data. It uses the LDAK (Linear Mixed Model Association Kernel) model, which uses genomic data to automatically detect genetic correlations and evaluate the results. It has the best performance among single-parameter models, Speed, Holmes, and Balding (2020). After estimating the heritability contributed by a kinship matrix using REML, I then computed the results using BLUP to estimate the effect sizes of the predictor variables used to compute the kinship matrix. BLUP uses a linear model to estimate the effect value of a single individual and its formula is:

$$\hat{y} = X\hat{b} + Z\hat{u},$$

where $\hat{y}$ is an estimate of the effect size for a single individual; X is the design matrix of explanatory variables; $\hat{b}$ is the estimated value of the fixed effect; Z is the design matrix of random effects; $\hat{u}$ is the estimated value of the random effect.

I used the $h2$ density plot to draw the effects obtained by BLUP to get the distribution of heritability, Figure.5. The $h2$ density plot shows the shape of a normal distribution, and the peak is high and the width is narrow, indicating that the feature has a certain heritability, but the effect is relatively low.

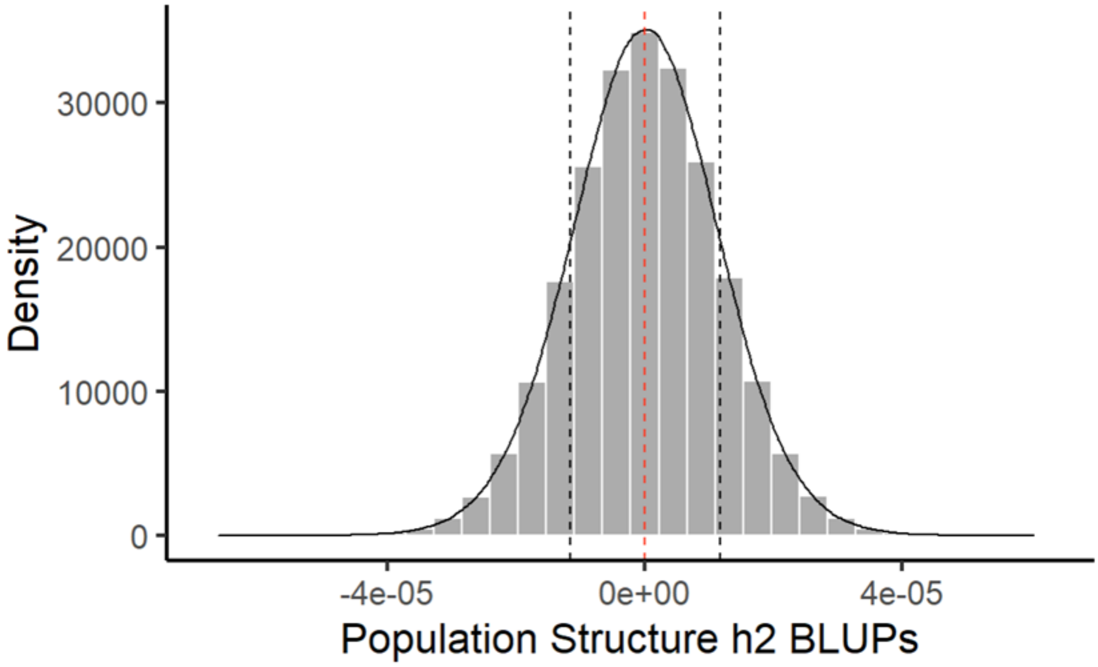


Figure 5: **Distribution of h2.**

The red dashed line is the mean ± one standard deviation in black dashed lines

2.6 Mixed Linear Model

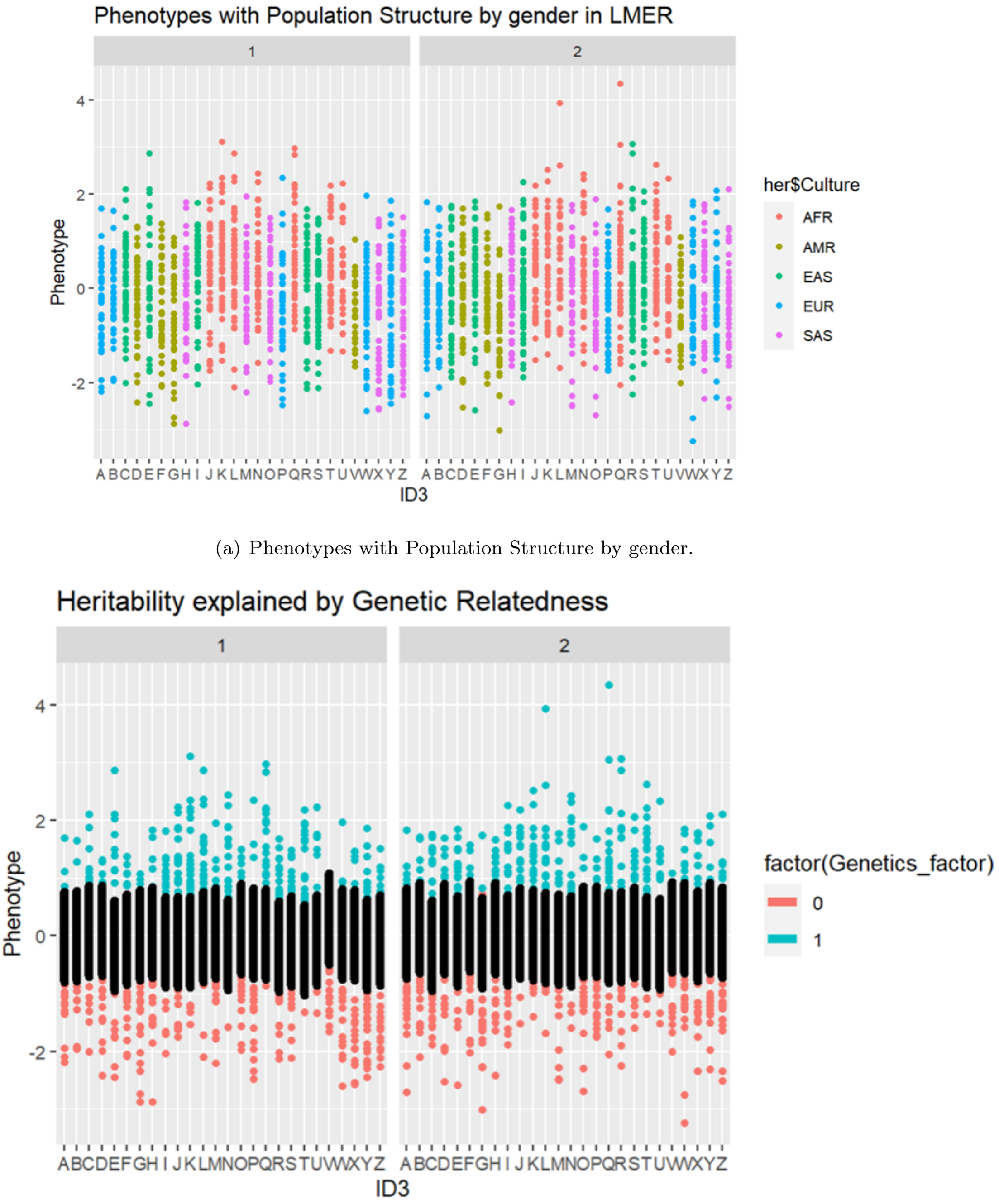
Mixed linear models are a method for estimating the genotype contribution to a trait (heritability). When analyzing multi-ethnic genotype data, mixed linear models can be used to consider the influence of ethnicity, Bauer (2003).

The mixed linear model analysis method is usually as follows:

1. Modeling: Build a mixed linear model for the multi-ethnic genotype data, which included genetic relatedness as a random factor.
2. Calculating heritability: The heritability of each ethnic group is estimated using a mixed linear model.
3. Comparing heritability differences between ethnic groups: Compare the heritability differences among ethnic groups to explore the contribution of ethnic groups to phenotypes.

By using the mixed linear model analysis approach, the contribution of genetic relatedness to phenotypes in multiethnic genotype data can be accurately explored, Blood, Cabral, Heeren, and Cheng (2010). The advantage of using MLM is that multiple genetic factors associated with associated traits can be considered simultaneously, and a kinship matrix can be used to control for genetic correlations. This allows for more accurate estimation of effect sizes and more efficient genomic association study, Korte et al. (2012).

I used Linear Mixed-Effects Regression (LMER) to fit a mixed linear model of fixed effects and random effects, with population and gender as fixed effects and genetic relatedness as random effects. I marked the distribution of phenotypes of each population, as shown in Figure.6. Figure(a) is calculated before using the mixed linear model, and Figure(b) marks with the proportion of heritability explained by Genetic Relatedness, where the positive relatedness is blue and the negative relatedness is red.



(a) Phenotypes with Population Structure by gender.

(b) Heritability explained by Genetic Relatedness.

Figure 6: **Random effects of 1000 Genome Project by LMER.**

In the x-axis, the 26 letters represent each population, see the Table.1 below. The y-axis represents the phenotypes, black is a fixed effect, the genetic(random) effect is calculated by LMER, blue is the positive genetic effect, red is the negative genetic effect.

A	B	C	D	E	
EUR_GBR	EUR_FIN	EAS_CHS	AMR_PUR	EAS_CDX	
F	G	H	I	J	
AMR_CLM	AMR_PEL	SAS_PJL	EAS_KHV	AFR_ACB	
K	L	M	N	O	
AFR_GWD	AFR_ESN	SAS_BEB	AFR_MSL	SAS_STU	
P	Q	R	S	T	
EUR_CEU	AFR_YRI	EAS_CHB	EAS_JPT	AFR_LWK	
U	V	W	X	Y	Z
AFR_ASW	AMR_MXL	EUR_TSI	SAS_GIH	EUR_IBS	SAS_ITU

Table 1: **The list of population name.**

3 Discussion

This analysis shows the disadvantages of only using GWAS in the analysis of Admixed Population and the advantages of using PCA and MLM to analyze the population structure. When different research participants are included in the genomics study, it is common to get false positive results simply by using GWAS, because of the potential population structure related to the phenotype in the sample, which may confound the genetic association test, Peterson et al. (2019).

When there exists population stratification, PCA can be applied to correct it. PC can reveal population structure when estimating data points based on the genetic markers of each individual. GWAS samples can be projected onto the PC axis for group comparison with known reference populations, Peterson et al. (2017).

By drawing the kinship matrix heatmap, I found that within the same population group, there would also be differences in genetic similarity within some populations. Failure to take into account intra-ethnic genetic differences when performing gene association analysis or PCA can lead to several errors:

1. False-positive association: Due to the influence of genetic differences within the population group, genetic association analysis may result in a false-positive association, that is, there is an association in some populations, but there is no association in the entire sample set.
2. False-negative associations: Due to the influence of genetic differences within population groups, genetic association analysis may lead to false-negative associations, that is, there is no association in some populations, but there is an association in the entire sample set.
3. Wrong association conclusions: Due to the influence of genetic differences within population groups, genetic association analysis may draw wrong association conclusions, that is, there is an association in some populations, but there is no association in the entire sample set.

Conjoint analysis with a mixed linear model approach enables the inclusion of all participants, regardless of their population. However, basic admixture models may not fully control population structure in different populations, and there are environmental factors in the association of phenotypes with ancestry, which could be discussed in the future. I used a mixed linear model to calculate heritability, fit values of fixed effects and random effects, and perform genetic association analysis efficiently. This approach can be used to identify population-specific genetic variants associated with complex traits.

Conducting research on the limitations of genome-wide association studies and the benefits of using PCA and MLM to analyze population structure can contribute to our understanding of the genetic and demographic factors that shape populations and may have important implications for fields such as population genetics and evolutionary biology.

4 Acknowledgement

I highly appreciate Bjarni Vilhjálmsson and Doug Speed for their excellent teaching during this whole semester.

I gratefully acknowledge Doug Speed for providing the association study software: LDAK.

References

Bauer, D. J. (2003, 6). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*, 135-167. DOI: 10.3102/10769986028002135

Blood, E. A., Cabral, H., Heeren, T., & Cheng, D. M. (2010, 12). Performance of mixed effects models in the analysis of mediated longitudinal data. *BMC Medical Research Methodology*, *10*, 16. DOI: 10.1186/1471-2288-10-16

Elhaik, E. (n.d.). Why most principal component analyses (pca) in population genetic studies are wrong. Retrieved from https://doi.org/10.1101/2021.04.11.439381 DOI: 10.1101/2021.04.11.439381

Harris, P., Brunsdon, C., & Charlton, M. (2011, 10). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, *25*, 1717-1736. DOI: 10.1080/13658816.2011.554838

Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., & Nordborg, M. (2012, 9). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, *44*, 1066-1071. DOI: 10.1038/ng.2376

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018, 6). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*. DOI: 10.1002/mpr.1608

Molenberghs, G., & Verbeke, G. (2000). *Linear mixed models for longitudinal data.* Springer New York. DOI: 10.1007/978-1-4419-0300-6

Peterson, R. E., Edwards, A. C., Bacanu, S.-A., Dick, D. M., Kendler, K. S., & Webb, B. T. (2017, 8). The utility of empirically assigning ancestry groups in cross-population genetic studies of addiction. *The American Journal on Addictions*, *26*, 494-501. DOI: 10.1111/ajad.12586

Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C. Y., Popejoy, A. B., Periyasamy, S., … Duncan, L. E. (2019, 10). *Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations* (Vol. 179). Cell Press. DOI: 10.1016/j.cell.2019.08.051

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., … Sham, P. C. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559-575. DOI: 10.1086/519795

Speed, D., Holmes, J., & Balding, D. J. (2020, 4). Evaluating and improving heritability models using summary statistics. *Nature Genetics*, *52*, 458-462. DOI: 10.1038/s41588-020-0600-y

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., … Korbel, J. O. (2015, 10). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*, 75-81. DOI: 10.1038/nature15394

Zeggini, E., & Ioannidis, J. P. (2009, 2). Meta-analysis in genome-wide association studies. *Pharmacogenomics*, *10*, 191-201. DOI: 10.2217/14622416.10.2.191

Zhou, X., & Stephens, M. (2012, 7). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*, 821-824. DOI: 10.1038/ng.2310

Zhou, X., & Stephens, M. (2014, 4). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, *11*, 407-409. DOI: 10.1038/nmeth.2848