Lecture 3: GWAS in Samples with Structure & Introduction to the REGENIE Software

Instructors: Joelle Mbatchou and Loic Yengo

Summer Institute in Statistical Genetics 2023

Introduction

- Genetic association studies are widely used for the identification of genes that influence complex traits.
- To date, hundreds of thousands of individuals have been included in genome-wide association studies (GWAS) for the mapping of both dichotomous and quantitative traits.
- Large-scale genomic studies often have high-dimensional data consisting of
  - Tens of thousands of individuals
  - Genotypes data on a million (or more!) SNPs for all individuals in the study
  - Many phenotypes of interest such as Height, BMI, HDL cholesterol, blood pressure, diabetes, etc.

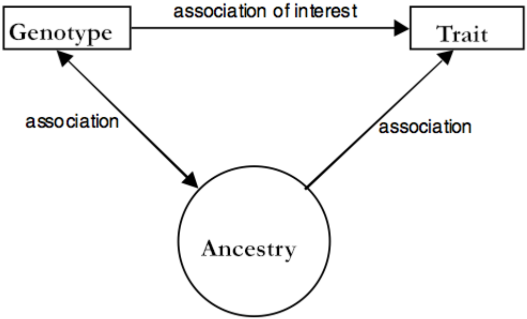
Introduction

- The vast majority of these studies have been conducted in populations of European ancestry
- Non-European populations have largely been underrepresented in genetic studies, despite often bearing a disproportionately high burden for some diseases.
- Recent genetic studies have investigated more diverse populations.

Confounding due to Ancestry

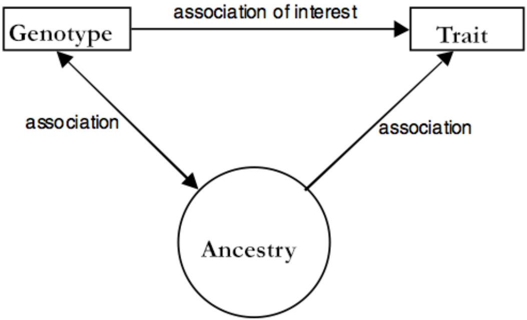
- The observations in association studies can be confounded by population structure
  - **Population structure**: the presence of subgroups in the population with ancestry differences
- Neglecting or not accounting for ancestry differences among sample individuals can lead to **false positive** or **spurious associations!**
- This is a serious concern for all genetic association studies.

Confounding due to Ancestry



In statistics, a **confounding variable** is an extraneous variable in a statistical model that correlates with both the dependent variable and the independent variable.

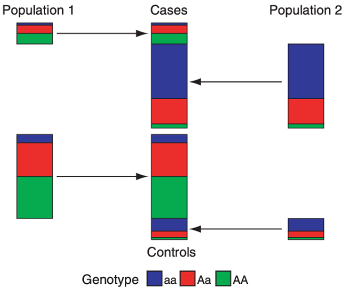
Confounding due to Ancestry



- Ethnic groups (and subgroups) often share distinct dietary habits and other lifestyle characteristics that leads to many traits of interest being correlated with ancestry and/or ethnicity.

Spurious Association

- Association test aims to compare of allele frequency between cases and controls.
- Consider a sample from 2 populations:
  - No differences in allele frequencies between cases/controls **within each population**
  - Large differences in allele frequencies **between populations**
  - Population 2 is overrepresented among cases in the sample. $\Longrightarrow$ spurious association between disease and genetic marker



Marchini et al., *Nature Genetics*, 2004

Genomic Control

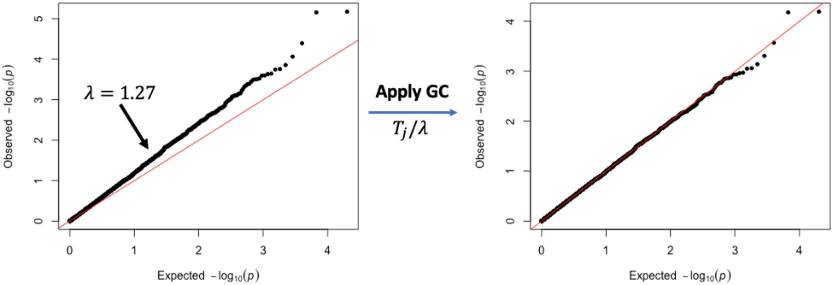
- Devlin and Roeder (1999) proposed correcting for substructure via a method called "genomic control."
- If there is no population structure, then **at unlinked variants** the test statistic $T \sim \chi^2_1$.
- If there is population structure, the statistic will deviate from a χ^2_1 distribution by an approximate constant factor $T \sim \lambda \chi^2_1$ which is estimated as

$$\lambda = \frac{median(T)}{median(\chi^2_1)} = \frac{median(T)}{.456}$$

- It is then applied to the test statistic values at all markers:

$$\tilde{T}_j = \frac{T_j}{\lambda}$$

Genomic Control



LD Score Regression

- In practice, λ is computed using all variants
- Polygenicity can cause $\lambda > 1$
  - Hard to separate confounding from polygenicity when $\lambda > 1$
- LD score regression separates these by regressing "LD scores" L_j on the test statistics

$$E[T_j] = Nh_g^2/M \cdot L_j + Na + 1$$

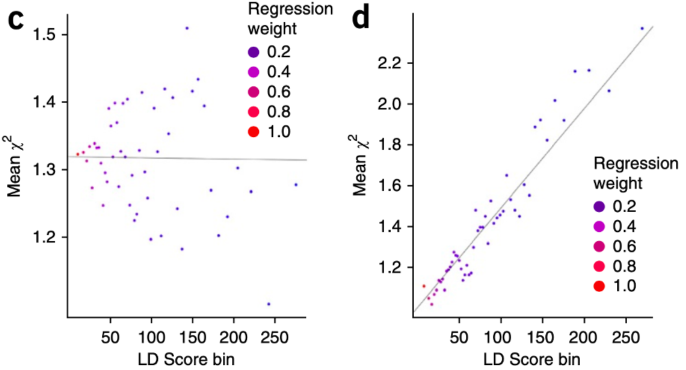
Slope → captures polygenicity
Intecept → captures confounding

LD Score Regression

Population structure & no heritability

$\lambda = 1.32$

No Population structure & with heritability



Bulik-Sullivan, *Nature Genetics*, 2015

Correcting for Population Structure with PCA

- Principal Components Analysis (PCA) is the most widely used approach for identifying and adjusting for ancestry differences among sample individuals
- Consider the genetic relationship matrix $\hat{\boldsymbol{\Psi}}$ discussed in the previous lecture with components $\hat{\psi}_{ij}$ for each pair of individuals as:

$$\hat{\psi}_{ij} = \frac{1}{M} \sum_{l=1}^{M} \frac{(G_{il} - 2\hat{p}_l)(X_{jl} - 2\hat{p}_l)}{\hat{p}_l(1 - \hat{p}_l)}$$

  where $G_{il} = \{0, 1, 2\}$ is the genotype value and $\hat{p}_l$ is a corresponding allele frequency estimate at marker l

Correcting for Population Structure with PCA

- Price et al. (2006) proposed correcting for structure in genetic association studies by applying PCA to $\hat{\boldsymbol{\Psi}}$.
- They developed a method called EIGENSTRAT for association testing in structured populations where the top principal components (highest eigenvalues) are used as covariates in a linear regression model to correct for sample structure.

$$Y = \beta_0 + \beta_1 G + \beta_2 PC_1 + \beta_3 PC_2 + \beta_4 PC_3 + \cdots + \epsilon$$

$$H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0$$

Samples with Population Structure and Relatedness

- Relatedness (family structure or cryptic relatedness) in the sample can lead to spurious association in genetic association studies
- The EIGENSTRAT method was developed for unrelated samples with population structure
  - In the presence of relatedness, PCs may not fully capture this finer-scale structure
- Many genetic studies include relatedness & modeling it directly can lead to improvements in statistical power

Association Testing in Samples with Population Structure and Relatedness

- Linear mixed models (LMMs) have been demonstrated to be a flexible approach for association testing in structured samples. Consider the following model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G_s}\gamma + \mathbf{g} + \boldsymbol{\epsilon}$$

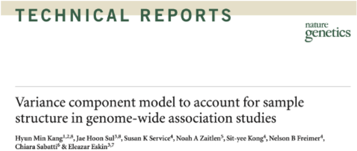
- **Fixed effects:**
  - **X** is a $n \times (k+1)$ matrix of covariates that includes an intercept
  - $\boldsymbol{\beta}$ is the $(k+1)$-length vector of covariate effects
  - γ is the (scalar) association parameter of interest, measuring the effect of genotype on phenotype

Linear Mixed Models for Genetic Association

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{G_s}\gamma + \mathbf{g} + \epsilon$$

- **Random effects:**
  - $\mathbf{g}$ is a n-length vector of polygenic effects with $\mathbf{g} \sim N(\mathbf{0}, \sigma_g^2 \mathbf{\Psi})$
    - σ_g^2 represents additive genetic variance and $\mathbf{\Psi}$ is a $n \times n$ matrix of pairwise measures of genetic relatedness (e.g. kinship matrix, GRM)
    - $\mathbf{g}$ should capture correlation between individuals due to genetic relatedness
  - ϵ is a n-length vector with $\epsilon \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I})$
    - σ_e^2 represents variance due to non-genetic effects assumed to be acting independently on individuals

LMM methods for Quantitative Traits

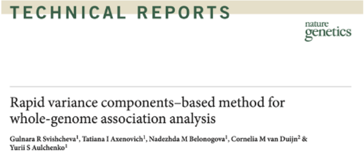


TECHNICAL REPORTS

nature genetics

Variance component model to account for sample structure in genome-wide association studies

Hyun Min Kang[1,2,8], Jae Hoon Sul[3,8], Susan K Service[4], Noah A Zaitlen[5], Sit-yee Kong[4], Nelson B Freimer[4], Chiara Sabatti[6] & Eleazar Eskin[3,7]

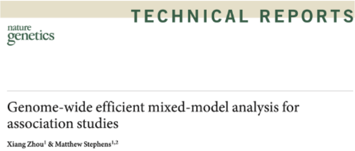


TECHNICAL REPORTS

nature genetics

Rapid variance components–based method for whole-genome association analysis

Gulnara R Svishcheva[1], Tatiana I Axenovich[1], Nadezhda M Belonogova[1], Cornelia M van Duijn[2] & Yurii S Aulchenko[1]

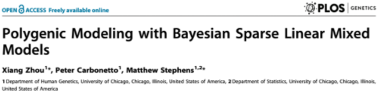


TECHNICAL REPORTS

nature genetics

Genome-wide efficient mixed-model analysis for association studies

Xiang Zhou[1] & Matthew Stephens[1,2]



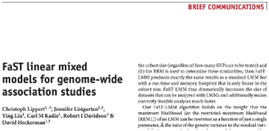
OPEN ACCESS Freely available online

PLOS | GENETICS

Polygenic Modeling with Bayesian Sparse Linear Mixed Models

Xiang Zhou[1*], Peter Carbonetto[1], Matthew Stephens[1,2*]

1 Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America, 2 Department of Statistics, University of Chicago, Chicago, Illinois, United States of America



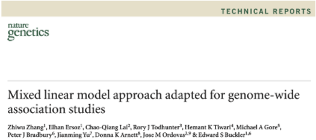
BRIEF COMMUNICATIONS

FaST linear mixed models for genome-wide association studies

Christoph Lippert[1–3], Jennifer Listgarten[1,3], Ying Liu[1], Carl M Kadie[1], Robert I Davidson[1] & David Heckerman[1,3]

the cohort size (regardless of how many SNPs are to be tested) and (ii) the RRM is used to determine these similarities, then FaST-LMM produces exactly the same results as a standard LMM but with a run time and memory footprint that is only linear in the cohort size. FaST-LMM thus dramatically increases the size of datasets that can be analyzed with LMMs and additionally makes currently feasible analyses much faster.

Our FaST-LMM algorithm builds on the insight that the maximum likelihood (or the restricted maximum likelihood (REML)) of an LMM can be rewritten as a function of just a single parameter, δ, the ratio of the genetic variance to the residual vari-



TECHNICAL REPORTS

nature genetics

Mixed linear model approach adapted for genome-wide association studies

Zhiwu Zhang[1], Elhan Ersoz[1], Chao-Qiang Lai[2], Rory J Todhunter[3], Hemant K Tiwari[4], Michael A Gore[5], Peter J Bradbury[6], Jianming Yu[7], Donna K Arnett[8], Jose M Ordovas[2,9] & Edward S Buckler[1,6]

LMMs: Two Step Procedure

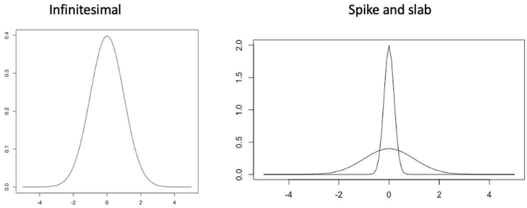
- Many LMM methods use a two-step procedure for GWAS
- Step 1 considers a null model without the tested SNP of interest (i.e. $\gamma = 0$)

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{g} + \epsilon$$

  - Obtain parameter estimates to get predictions for the polygenic effects **g**
  - Same for all variants tested so only performed once which reduces the computational burden

LMMs: Two Step Procedure

- Many methods differ mainly in Step 1 approach
  - Model used for the additive polygenic random effect term



  - Algorithm used to obtain parameter estimates
    - Parameter estimates are obtained using various approaches (e.g. maximum likelihood, restricted maximum likelihood [REML],...)

LMMs: Two Step Procedure

- In Step 2, association testing of SNP and phenotype ($H_0 : \gamma = 0$) is performed based on the model including the tested SNP

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{G_s}\gamma + \mathbf{g} + \boldsymbol{\epsilon}$$

- A score test is performed using the null parameter estimates obtained from Step 1.
- Use Leave-One-Chromosome-Out (LOCO) scheme in Step 1 so polygenic term doesn't capture effects on tested chromosome (i.e. proximal contamination)

$$\mathbf{g} \sim N(\mathbf{0}, \sigma_g^2 \boldsymbol{\Psi}_{-chr(G_s)})$$

LMMs on biobank scale data

- Largest biobanks have gathered data on 100,000s of individuals (e.g. UK Biobank at $N = 500,000$ individuals)
- Many LMM methods involved computationally expensive operations due to the $N \times N$ GRM

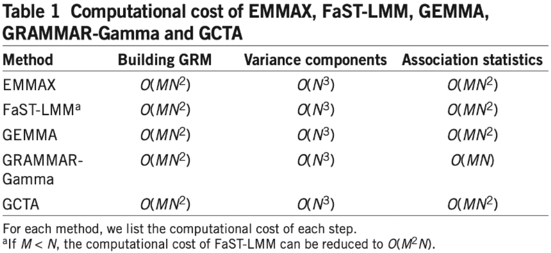


Table 1 Computational cost of EMMAX, FaST-LMM, GEMMA, GRAMMAR-Gamma and GCTA

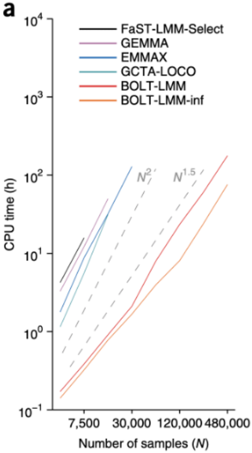
Method	Building GRM	Variance components	Association statistics
EMMAX	$O(MN^2)$	$O(N^3)$	$O(MN^2)$
FaST-LMM[a]	$O(MN^2)$	$O(N^3)$	$O(MN^2)$
GEMMA	$O(MN^2)$	$O(N^3)$	$O(MN^2)$
GRAMMAR-Gamma	$O(MN^2)$	$O(N^3)$	$O(MN)$
GCTA	$O(MN^2)$	$O(N^3)$	$O(MN^2)$

For each method, we list the computational cost of each step.
[a] If $M < N$, the computational cost of FaST-LMM can be reduced to $O(M^2N)$.

Yang et al., *Nature Genetics* 2014

LMMs on biobank scale data

- Loh et al. (2015) proposed BOLT-LMM which used very efficient algorithms (Variational Bayes) to reduce scaling to $\sim O(MN^{1.5})$ for Step 1 and could be applied to biobank-scale data
- Jiang et al. (2019) proposed fastGWA which made use of a sparse GRM leading to further improvements for Step 1 $\sim O(MN)$



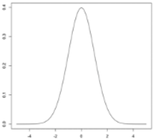
Loh et al., Nature Genetics 2015

LMMs & Whole Genome Regression

- LMMs are closely related to whole genome regression

$$Y = W\beta + \underset{\substack{\big\uparrow \\ N(0,\sigma_g^2\,\boldsymbol{\Psi}) \\ \boldsymbol{\Psi} = \boldsymbol{G}\boldsymbol{G}^{\boldsymbol{T}}/\boldsymbol{M}}}{g} + \epsilon \quad \Longleftrightarrow \quad Y = W\beta + \sum_{l=1}^{M} G_l \underset{\substack{\big\uparrow \\ N(0,\sigma_g^2/M)}}{\theta_l} + \epsilon$$

GRM using M variants



LMMs & Whole Genome Regression

- LMMs are closely related to whole genome regression

1 parameter

$$Y = W\beta + g + \epsilon$$

$$\uparrow$$

$$N(0, \sigma_g^2\, \boldsymbol{\Psi})$$

$$\boldsymbol{\Psi = GG^T/M}$$

GRM using M variants

$$\Longleftrightarrow$$

M parameters

$$Y = W\beta + \sum_{l=1}^{M} G_l \boldsymbol{\theta_l} + \epsilon$$

$$\uparrow$$

$$N(0, \sigma_g^2/M)$$

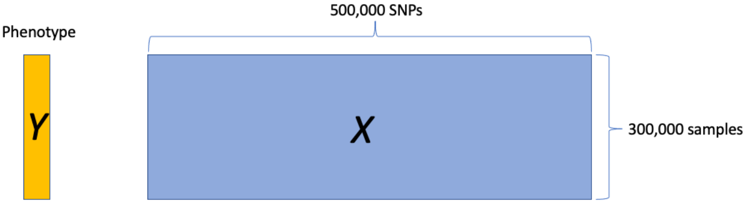
REGENIE: Whole Genome Regression

- Step 1: computationally efficient whole genome regression

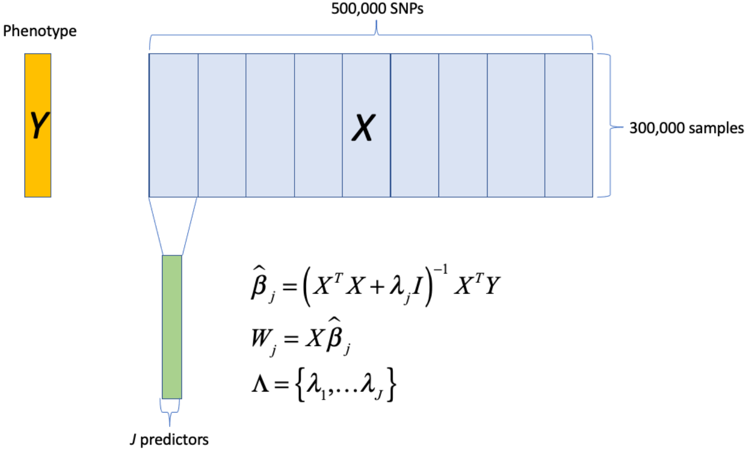
$$\mathbf{Y} = \mathbf{X}\beta + \sum_{l=1}^{M} G_l \theta_l + \epsilon$$

- M is usually $\sim$ 500,000 SNPs across the genome
- REGENIE splits genetic data into blocks and runs local regressions in each block to obtain local genetic scores

REGENIE: Whole Genome Regression



REGENIE: Whole Genome Regression

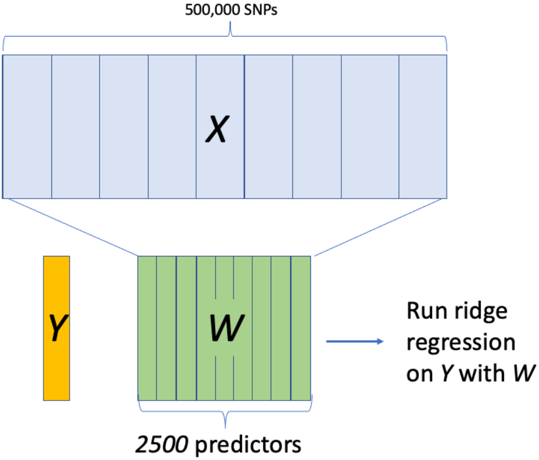


$$\hat{\beta}_j = \left(X^T X + \lambda_j I\right)^{-1} X^T Y$$

$$W_j = X\hat{\beta}_j$$

$$\Lambda = \left\{\lambda_1, \ldots \lambda_J\right\}$$

REGENIE: Whole Genome Regression



500,000 SNPs

X

Y

W

2500 predictors

Run ridge regression on Y with W

REGENIE: Whole Genome Regression

- Step 1: computationally efficient whole genome regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{l=1}^{M} G_l \theta_l + \boldsymbol{\epsilon}$$

- Divide into two levels of regressions
  - Reads genetic data in blocks and within each block fits ridge regression (penalized linear regression)
  - Fit another round of ridge regression on all the block predictors
- Polygenic predictions ($\sum_{l=1}^{M} G_l \hat{\theta}_l$) capture population structure, relatedness as well as polygenicity using LOCO

REGENIE: Whole Genome Regression

- Step 2: test the association parameter γ under the null hypothesis of $H_0 : \gamma = 0$.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + G_s\gamma + \sum_{l=1}^{M} G_l\hat{\theta}_l + \epsilon$$

- Test on millions of genetic variants (array/imputed/exome)
- Also works on binary traits where logistic regression is used instead of linear regression

`https://rgcgithub.github.io/regenie/`

Summary

- Population structure is an important source of confounding in GWAS
- Genomic inflation λ_{GC} and LD score regression can be used to detect its presence
- Adding PCs as covariates can control for population stratification but not relatedness
- Mixed models can directly capture genetic relatedness and help improve statistical power as well as avoid inflated type 1 error

References

- Devlin, B. & Roeder, K. Genomic Control for Association Studies. *Biometrics* **55**, 997-1004 (1999).
- Bulik-Sullivan, B.K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291-295 (2015).
- Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904-909 (2006).
- Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M. & Price, A.L. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* **46**, 100-106 (2014).

References

- Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284-290 (2015).
- Jiang, L. et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics* **51**, 1749-1755 (2019).
- Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50**, 1335-1341 (2018).
- Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics* **53**, 1097-1103 (2021).