Module 17: Computational Pipeline for WGS Data

GWAS in Samples with Structure

Timothy A. Thornton

University of Washington, TOPMed Data Coordinating Center

Summer Institute in Statistical Genetics
July 2019

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
  - Phenotype or Trait values of interest such as Height, BMI, HDL cholesterol, blood pressure, diabetes, etc.

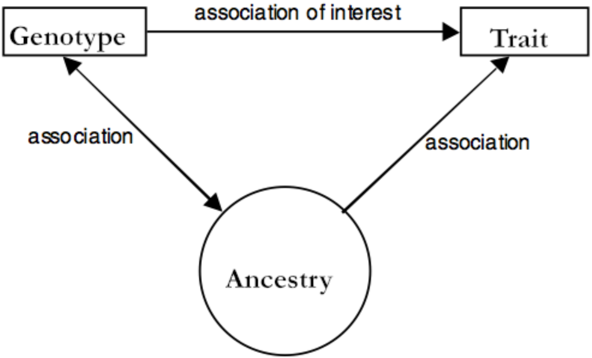
Introduction

- 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 in Genetic Association Studies

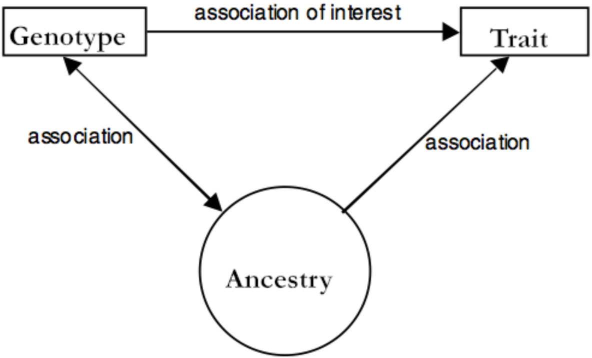
- The observations in association studies can be confounded by population structure
  - **Population structure**: the presence of subgroups in the population with ancestry differences
- Neglecting or not accounting for ancestry differences among sample individuals can lead to **false positive** or **spurious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

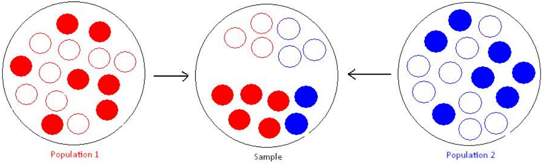
Confounding due to Ancestry



- Ethnicgroups (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

- Case/Control association test
  - Comparison of allele frequency between cases and controls.
- Consider a sample from 2 populations:

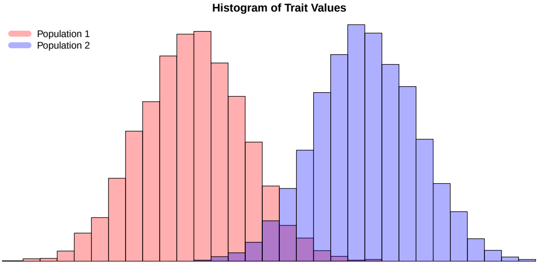


  - Red population overrepresented among cases in the sample.
  - Genetic markers that are not influencing the disease but with significant differences in allele frequencies between the populations

    $\Longrightarrow$ spurious association between disease and genetic marker

Spurious Association

- Quantitative trait association test
  - Test for association between genotype and trait value
- Consider sampling from 2 populations:



  - Blue population has higher trait values.
  - Different allele frequency in each population
    $\Longrightarrow$ spurious association between trait and genetic marker if one population is overrepresented in the sample

Genotype and Phenotype Data

- Suppose the data for the genetic association study include genotype and phenotype on a sample of n individuals
- Let $\mathbf{Y} = (Y_1, \ldots Y_n)^T$ denote the $n \times 1$ vector of phenotype data, where Y_i is the quantitative trait value for the ith individual.
- Consider testing SNP s in a genome-screen for association with the phenotype, where $\mathbf{G_s} = (G_1^s, \ldots G_n^s)^T$ is $n \times 1$ vector of the genotypes, where $G_i^s = 0, 1,$ or 2, according to whether individual i has, respectively, 0, 1 or 2 copies of the reference allele at SNP s.

Correcting for Population Structure with PCA

- As previously discussed, Principal Components Analysis (PCA) is a popular approach for identifying population structure from genetic data
- PCA is also widely used to adjust for ancestry difference among sampled individuals in a GWAS.
- Consider the genetic relationship matrix $\hat{\boldsymbol{\Psi}}$ discussed in the previous lecture with components $\hat{\psi}_{ij}$:

$$\hat{\psi}_{ij} = \frac{1}{M} \sum_{s=1}^{M} \frac{(G_i^s - 2\hat{p}_s)(G_j^s - 2\hat{p}_s)}{\hat{p}_s(1 - \hat{p}_s)}$$

where $\hat{p}_s$ is an allele frequency estimate for the type 1 allele at marker s

Correcting for Population Structure with PCA

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

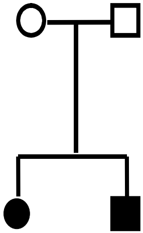
$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{G_s} + \beta_2 \mathbf{PC}_1 + \beta_3 \mathbf{PC}_2 + \beta_4 \mathbf{PC}_3 + \cdots + \boldsymbol{\epsilon}$$

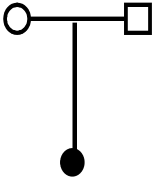
- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

Samples with Population Structure and Relatedness

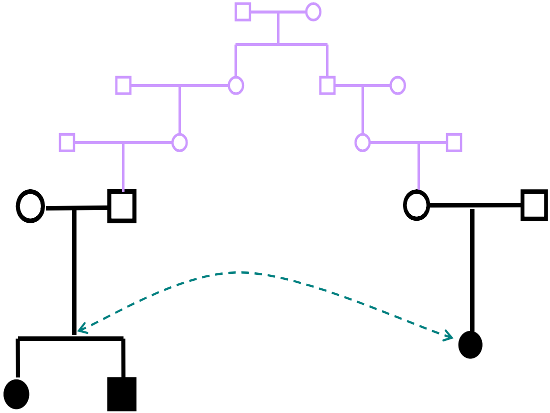
- The EIGENSTRAT methods was developed for unrelated samples with population structure
- Methods may not be valid in samples with related individuals (known and/or unknown)
- Many genetic studies have samples with related individuals
- Cryptic and/or misspecified relatedness among the sample individuals can also lead invalid association results

Incomplete Genealogy





Incomplete Genealogy



Association Testing in samples with Population Structure and Relatedness

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

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

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

Linear Mixed Models for Genetic Association

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

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

LMMs For Cryptic Structure

- The matrix $\mathbf{\Psi}$ will be generally be unknown when there is population structure (ancestry differences) and/or cryptic relatedness among sample individuals.
- Kang et al. [Nat Genet, 2010] proposed the EMMAX linear mixed model association method that is based on an empirical genetic relatedness matrix (GRM) $\hat{\mathbf{\Psi}}$ calculated using SNPs from across the genome. The $(i,j)th$ entry of the matrix is estimated by

$$\hat{\mathbf{\Psi}}_{ij} = \frac{1}{S}\sum_{s=1}^{S}\frac{(G_i^s - 2\hat{p}_s)(G_j^s - 2\hat{p}_s)}{2\hat{p}_s(1-\hat{p}_s)}$$

where $\hat{p}_s$ is the sample average allele frequency. S will generally need to be quite large, e.g., larger than 100,000, to capture fine-scale structure.

EMMAX

- For genetic association testing, the EMMAX mixed-model approach first considers the following **null model** without including any of the SNPs as fixed effects:

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\epsilon} \tag{1}$$

- The variance components, σ_g^2 and σ_e^2, are then estimated using either a maximum likelihood or restricted maximum likelihood (REML), with $\mathbf{Cov}(\mathbf{Y})$ set to $\sigma_g^2\hat{\boldsymbol{\Psi}} + \sigma_e^2\mathbf{I}$ in the likelihood with fixed $\hat{\boldsymbol{\Psi}}$

EMMAX

- Association testing of SNP s and phenotype is then based on the model

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{G^s}\gamma + \mathbf{g} + \boldsymbol{\epsilon}$$

- The EMMAX association statistic is the score statistic for testing the null hypothesis of $\gamma = 0$ using a generalized regression with $Var(\mathbf{Y}) = \boldsymbol{\Sigma}$ evaluated at $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}_g^2 \hat{\boldsymbol{\Psi}} + \hat{\sigma}_e^2 \mathbf{I}$
- EMMAX calculates $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ only once from model (1) to reduce computational burden.

GEMMA

- Zhou and Stephens [2012, Nat Genet] developed a computationally efficient mixed-model approach named GEMMA
- GEMMA is very similar to EMMAX and is essentially based on the same linear mixed-model as EMMAX

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{G^s}\gamma + \mathbf{g} + \boldsymbol{\epsilon}$$

- However, the GEMMA method is an "exact" method that obtains maximum likelihood estimates of variance components $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ for each SNP s being tested for association.

Zhou and Stephens (2012) "Genome-wide efficient mixed-model analysis for association studies" Nature Genetics 44

Other LMM approaches for Quantitative Traits

- A number of similar linear mixed-effects methods have recently been proposed for association testing with quantitative triatwhen there is cryptic structure: Zhang at al. [2010, Nat Genet], Lippert et al. [2011, Nat Methods], Zhou & Stephens [2012, Nat Genet], and Svishcheva [2012, Nat, Genet], and others.

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

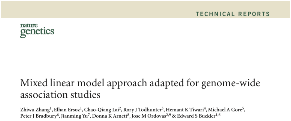
Gulnara R Svishcheva[1], Tatiana I Axenovich[1], Nadezhda M Belonogova[1], Cornelia M van Duijn[2] & Yurii S Aulchenko[1]

TECHNICAL REPORTS

nature genetics

Genome-wide efficient mixed-model analysis for association studies

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

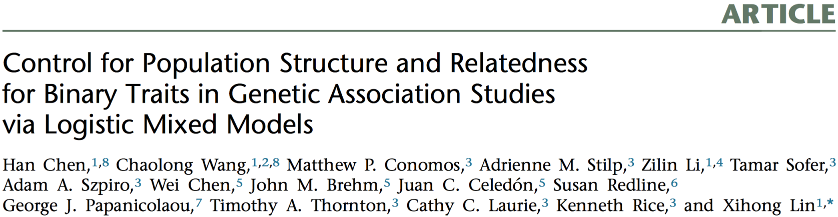

TECHNICAL REPORTS

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]

GMMAT: Logistic Mixed Model for Dichotomous Phenotypes



ARTICLE

Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models

Han Chen,[1,8] Chaolong Wang,[1,2,8] Matthew P. Conomos,[3] Adrienne M. Stilp,[3] Zilin Li,[1,4] Tamar Sofer,[3] Adam A. Szpiro,[3] Wei Chen,[5] John M. Brehm,[5] Juan C. Celedón,[5] Susan Redline,[6] George J. Papanicolaou,[7] Timothy A. Thornton,[3] Cathy C. Laurie,[3] Kenneth Rice,[3] and Xihong Lin[1,*]

The GMMAT can be used to conduct a logistic mixed model regression analysis of binary traits for GWAS with population structure and relatedness.

GMMAT: Logistic Mixed Model for Dichotomous Phenotypes

- Let π_i be the probability that individual i is affected with the disease. The GMMAT logistic mixed model is:

$$log\left(\frac{\pi_i}{1-\pi_i}\middle|\mathbf{W},\mathbf{G_s}\right)$$

$$=\mathbf{W}\boldsymbol{\beta}+\mathbf{G_s}\gamma+\mathbf{g}$$

- **Random effect:**
  - **g** is a length n random vector of polygenic effects with $\mathbf{g}\sim N(\mathbf{0},\tau_g\boldsymbol{\Psi})$
  - τ_g is the variance component parameter for polygenic effects
- GMMAT tests the association parameter γ under the null hypothesis of $H_0:\gamma=0$.

References

- Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, Szpiro AA, Chen W, Brehm JM, Celedn JC, Redline S, Papanicolaou GJ, Thornton T, Laurie CC, Rice K, Lin X (2016) Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *American Journal of Human Genetics.* **98**,653-666.
- Devlin B, Roeder K (1999). Genomic control for association studies. *Biometrics* **55**, 997-1004.
- Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.Y., Freimer, N. B., Sabatti, C. Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348-354.

References

- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904-909.
- Zhou, X., Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**,821-824.