

# Lecture 6: GWAS in Samples with Structure

Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2015

## 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.

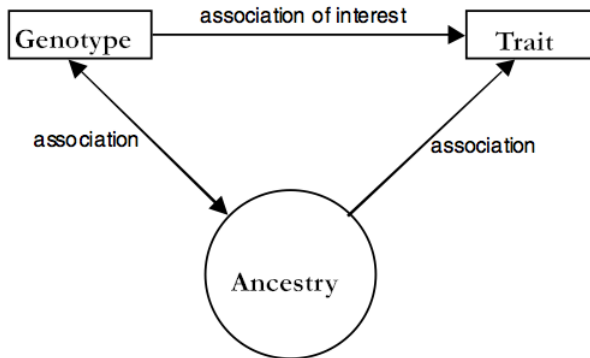
## Introduction

- ▶ 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.

## Case-Control Association Testing

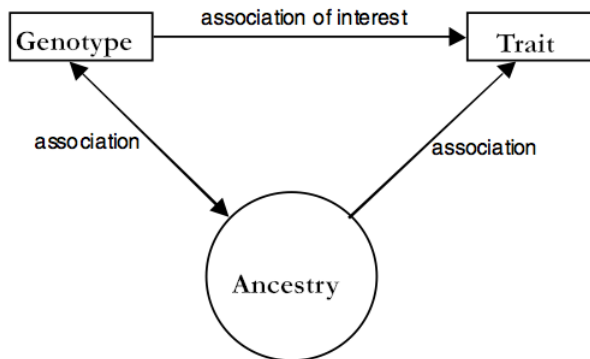
- ▶ The observations in association studies can be confounded by population structure
  - ▶ **Population structure**: the presence of subgroups in the population with ancestry differences
- ▶ Neglecting or not accounting for ancestry differences among sample individuals can lead to **false positive** 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.

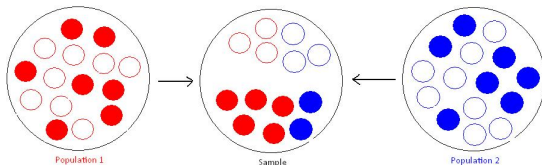
## Confounding due to Ancestry



- ▶ 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

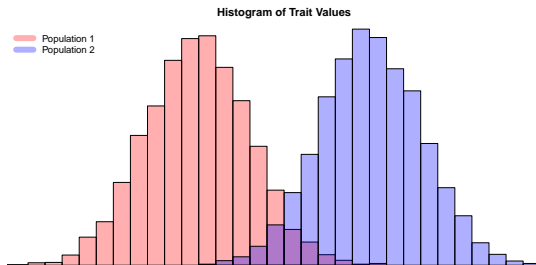
- ▶ Case/Control association test
  - ▶ Comparison of allele frequency between cases and controls.
- ▶ Consider a sample from 2 populations:



- ▶ **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  
 $\implies$  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  
 $\implies$  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  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$  denote the  $n \times 1$  vector of phenotype data, where  $Y_i$  is the quantitative trait value for the  $i$ th individual.
- ▶ Consider testing SNP  $s$  in a genome-screen for association with the phenotype, where  $\mathbf{G}_s = (G_1^s, \dots, G_n^s)^T$  is  $n \times 1$  vector of the genotypes, where  $G_i^s = 0, 1, \text{ or } 2$ , according to whether individual  $i$  has, respectively, 0, 1 or 2 copies of the reference allele at SNP  $s$ .

## Genomic Control

- ▶ Devlin and Roeder (1999) proposed correcting for substructure via a method called "genomic control."
- ▶ For each marker  $s$ , the Armitage trend statistic is calculated

$$A_{r_s} = Nr_{G_s Y}^2$$

where  $r_{G_s Y}^2$  is the squared correlation between the genotype variable  $\mathbf{G}_s$  for marker  $s$  and the phenotype variable  $\mathbf{Y}$ .

- ▶ If there is no population structure, the distribution of  $A_{r_s}$  will approximately follow a  $\chi^2$  distribution with 1 degree of freedom.
- ▶ If there is population structure, the statistic will deviate from a  $\chi^2_1$  distribution due to an inflated variance.

## Genomic Control

- ▶ Use  $\lambda = \frac{\text{median}(A_{r_1}, \dots, A_{r_s}, \dots, A_{r_M})}{.456}$  as a correction factor for cryptic structure, where .456 is the median of a  $\chi^2_1$  distribution.
- ▶ The uniform inflation factor  $\lambda$  is then applied to the Armitage trend statistic values

$$\tilde{A}_{r_s} = \frac{A_{r_s}}{\lambda}$$

- ▶  $\tilde{A}_{r_s}$  will approximately follow a  $\chi^2$  distribution with 1 degree of freedom.

## Correcting for Population Structure with PCA

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

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

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  $\hat{\Psi}$ .
- ▶ 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 multi-linear regression model to correct for sample structure.

$$Y = \beta_0 + \beta_1 X + \beta_2 PC_1 + \beta_3 PC_2 + \beta_4 PC_3 + \dots + \epsilon$$

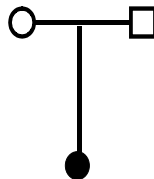
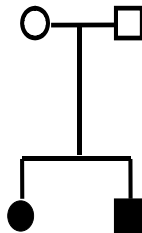
- ▶  $H_0 : \beta_1 = 0$  vs.  $H_a : \beta_1 \neq 0$

## Samples with Population Structure and Relatedness

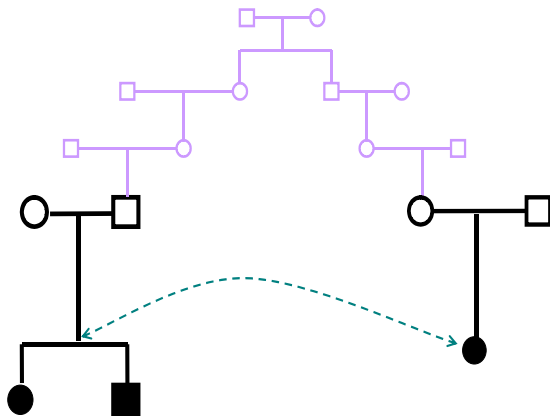
- ▶ The EIGENSTRAT methods was developed for unrelated samples with population structure
- ▶ Methods may not be valid in samples with related individuals (known and/or unknown)
- ▶ Many genetic studies have samples with related individuals

## Incomplete Genealogy

- ▶ Cryptic and/or misspecified relatedness among the sample individuals can also lead to spurious association in genetic association studies



## Incomplete Genealogy





## 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:

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

- ▶ **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
  - ▶  $\boldsymbol{\gamma}$  is the (scalar) association parameter of interest, measuring the effect of genotype on phenotype

# Linear Mixed Models for Genetic Association

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

► **Random effects:**

- $\mathbf{g}$  is a length  $n$  random vector of polygenic effects with  $\mathbf{g} \sim N(\mathbf{0}, \sigma_g^2 \boldsymbol{\Psi})$
- $\sigma_g^2$  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  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I})$
- $\sigma_e^2$  represents non-genetic variance due to non-genetic effects assumed to be acting independently on individuals

## LMMs For Cryptic Structure

- ▶ The matrix  $\Psi$  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)  $\hat{\Psi}$  calculated using SNPs from across the genome. The  $(i, j)$ th entry of the matrix is estimated by

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

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 model without including any of the SNPs as fixed effects:

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\epsilon} \quad (1)$$

- ▶ The variance components,  $\sigma_g^2$  and  $\sigma_e^2$ , 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  $\hat{\boldsymbol{\Psi}}$
- ▶ Association testing of SNP  $s$  and phenotype is then based on the model

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

- ▶ The EMMAX association statistic is the score statistic for testing the null hypothesis of  $\boldsymbol{\gamma} = \mathbf{0}$  using a generalized regression with  $\text{Var}(\mathbf{Y}) = \boldsymbol{\Sigma}$  evaluated at  $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}_g^2 \hat{\boldsymbol{\Psi}} + \hat{\sigma}_e^2 \mathbf{I}$

## 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

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

- ▶ 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

- ▶ A number of similar linear mixed-effects methods have recently been proposed when there is cryptic structure: Zhang et 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

Hyeon Min Kang<sup>1,2</sup>, Ju Hoon Seol<sup>1,2</sup>, Susan K Service<sup>1</sup>, Noah A Zaitlin<sup>1</sup>, Si-yeon Kang<sup>1</sup>, Nathan E Freidman<sup>1</sup>, Oksun Suhart<sup>1</sup> & Matthew Eskin<sup>1,2</sup>

### TECHNICAL REPORTS

nature  
genetics

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

Gadara B Svishcheva<sup>1</sup>, Tatiana I Antonovich<sup>1</sup>, Nadezhda M Belongovaya<sup>1</sup>, Corinna M van Duijn<sup>1</sup> & Yuri S Izhmanskii<sup>1</sup>

### TECHNICAL REPORTS

nature  
genetics

Genome-wide efficient mixed-model analysis for association studies

Xiang Zhou<sup>1</sup> & Matthew Stephens<sup>1,2</sup>

### TECHNICAL REPORTS

nature  
genetics

Mixed linear model approach adapted for genome-wide association studies

Zhen Zhang<sup>1</sup>, Ehsan Emami<sup>1</sup>, Chao-Qiang Lu<sup>1</sup>, Rory J Tibshirani<sup>1</sup>, Hossein K Tizabi<sup>1</sup>, Michael A Gore<sup>1</sup>, Peter J Bradbury<sup>1</sup>, Jianming Yu<sup>1</sup>, Dennis K Aronoff<sup>1</sup>, Jose M Ordovas<sup>1,2</sup> & Edward S Raskin<sup>1,4</sup>

## ROADTRIPS for Dichotomous Phenotypes

- ▶ Similar to LMMs, the ROADTRIPS approach of Thornton and McPeck (2010) also incorporates an empirical covariance matrix  $\hat{\Psi}$ .
- ▶ ROADTRIPS was developed for valid association testing in case-control samples with partially or completely unknown population and pedigree structure
- ▶ ROADTRIPS extensions, to samples with structure, have been developed for a number of association tests including Pearson  $\chi^2$  test and the Armitage trend test

## References

- ▶ 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. *Nat Genet* **42**, 348-354.
- ▶ Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-909.



## References

- ▶ Thornton, T., McPeck, M.S. (2010). ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. *Am. J. Hum. Genet.* **86**, 172-184.
- ▶ Zhou, X., Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**,821-824.