Population Structure in GWAS

H3ABioNet 2018 Genotyping Chip Data Analysis and GWAS lecture series

Ananyo.Choudhury@wits.ac.za
Section I
Population structure

- Association testing
- Stratification in Association
- Basic idea of confounders
- Genetic associations
- Population structure
- How population structure affects GWAS?
**Association testing**

**Fertilizer increases higher vitamin C content in fruits?**

- **Site A**
  - Fertilizer Applied
  - Measure vitamin C content
  - Compare - estimate statistical significance of difference
  - Fertilizer application is associated to Higher Vitamin C production or not

- **Site B**
  - No Fertilizer
  - Measure vitamin C content
**Stratification** - where association is observed due to systemic differences in the groups rather than differences in the outcome variable.

Significant statistic difference in Vitamin C content: ✔

Not due to fertilizer application, but due to stratification: ✗
Confounding: Factors other than the major variable that could effect the outcome.
Example of controlling structure with covariates
Example of controlling structure with covariates

No Control
A: Simple correlation
   \[ r = 0.49, p < .001 \]

Imperfect Control
B: Controlling for subjective heat
   Partial \( r = 0.33, p < .001 \)

Control
C: Controlling for recorded temperature
   Partial \( r = -0.02, p = .81 \)
Results from a case-control genome-wide association study investigating genetic variants associated with heart disease.

Image credit: Genome Research Limited
Continuous trait

A linear regression model is defined as

\[ y = x\beta_1 + \beta_0 + \varepsilon \]

Data:
- \( y \): a continuous trait
- \( x \): SNP genotype at a given locus

Parameters:
- \( \beta_1 \): regression coefficient, represents the strength of association between \( x \) and \( y \)
- \( \beta_0 \): intercept term (is 0 or ignored)
- \( \varepsilon \): noise or the part of \( y \) that is not explained by \( x \) (e.g., environmental effect)

Assumptions:
- The individuals in the study are not related
- The phenotype \( y \) has a normal distribution
GWAS for the chopstick gene

High chopstick skills

Low chopstick skills

Successful Use of Selected Hand Instrument !!
GWAS for the chopstick gene

High chopstick skills

Low chopstick skills

Ancestral composition HCS

Ancestral composition LCS

Allele frequency of variant - HLA-A in European (EU) and East Asians (EA)

Successful Use of Selected Hand Instrument!!

Molecular Psychiatry (2000) 5, 11-13

NEWS & VIEWS

Beware the chopsticks gene

The nature of confounding in genome-wide association studies

Bjarne J. Vihjálmsson1,2 and Magnus Nordborg3,4

VOLUME 14 | JANUARY 2013
Population stratification/structure is the presence of multiple subpopulations (e.g., individuals with different ethnic background) in a study.

Subpopulations in addition to differing in allele frequencies might also differ in disease rate, trait variable, cultural practices, diet etc.

If both allele frequencies and trait variables differ between subpopulations, PS can lead to false positive associations and/or mask true associations.
There could be a number of other confounders such as age, sex, and BMI which could affect GWAS inference.
Effect of PS on GWAS

• Campbell et al. 2005, conducted a GWAS for height based on a set of European American individuals.

• They found a SNP in the gene LCT strongly to be associated with height (P < 10^{-6})

• They also noted this SNP to show high allele frequency variation among European populations.

• Grouping individuals on the basis of European ancestry greatly reduced the apparent association that was due to population stratification.
Section II
Types of Population Structure and their origin

- Types of Structure
- Human history for the last few hundred millennia
- Cryptic Relatedness
Types of PS

- Structure due to inclusion of usually geographically distinct populations
  - NW Europeans and SE Europeans

- Structure due to variation in ancestral contributions (differential admixture) from genetically distinct populations.
  - Gila river community

<table>
<thead>
<tr>
<th>Indian Heritage</th>
<th>$Gm^{3;5;13;14}$ %</th>
<th>% Diabetes age adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>69%</td>
<td>18.5%</td>
</tr>
<tr>
<td>4</td>
<td>45%</td>
<td>28.6%</td>
</tr>
<tr>
<td>8</td>
<td>.01%</td>
<td>39.2%</td>
</tr>
</tbody>
</table>

Adapted from Knowler, 1988
Peopling of the world

Major human migrations across the world inferred through analyses of genomic data. (Nielsen et al. 2017)
Change in human diversity with migration. The number of different colors (representing amount of genetic variation) reduces as explorers move into newer territories.

(From: https://blogs.plos.org/)
Lactase Persistence

http://www.ucl.ac.uk/mace-lab/resources/glad
Schlebusch C and Jakobsson M, 2018
Map of major pre-farming population stratification across the African continent, showing non-Africans

Migration of herders and farmers since Holocene
Ethnolinguistic composition of present day Africa
Cryptic relatedness

- Cryptic relatedness refers to the idea of presence of relatives in a set of ostensibly unrelated individuals in a case-control association study.

- An essential assumption for a GWAS is the independence of subject genomes. Cryptic relatedness violates this assumption and could, thereby, confound the inference of an association study.

- Factors that give rise to cryptic relatedness
  - Assortative mating
  - Effective population sizes / Recent bottleneck
  - Sampling biases
Section III. How to detect PS and CR?

Pre association
- PCA based approach
- IBD based approach

Post association
- Genomic control
- QQ plots

Other approaches
Principal Component Analysis

Genotype information

Distance:
- P0, P1: 4
- P0, P2: 3
- P1, P2: 5
Genotype information:

P0: AA AC AT AA
P1: TT AA TT AA
P2: AA AC TT TT
P3: AT CC TT AA

Distance:

<table>
<thead>
<tr>
<th></th>
<th>P0</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>P0</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>P1</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>P2</td>
<td>0</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P0: (0,0,0)  P1: (4,0,0)  P2: (0,3,0)  P3=(2,0.33,2.21)
PCA

Uses a method known as *eigendecomposition* – in which it takes distance matrix and produces:

- **Eigenvalues**: $\lambda_i$ is relative importance of dimension $i$
- **Eigenvectors**: $v_i$ coordinates of each individual in the $i$-th dimension.
- Preserve relative distance between individuals.
- Number of dimensions/components are reduced.
- Components independent of each other.
- Ordered by importance.
Example of PCA

PCA coloured by population, Global

1000 Genomes, 2012
To what extent can population structure be detected using PCA?

A geographical map of Europe arises naturally as an efficient two-dimensional summary of genetic variation in Europeans

Novembre et al., 2008
PCA is easy

Running PCA

plink --bfile mydata --pca --out mypca

It is advisable to perform a LD based pruning before running the PCA.

Creating the file mypca.eigenvec and mypca.eigenval, containing the eigenvector and values respectively.

http://www.bioinf.wits.ac.za/software/genesis/
Assessing Structure Due to Relatedness

- Presence of duplicate and related individuals in the dataset may introduce bias and cause genotypes in families to be over-represented.

- To identify duplicate and related individuals, a metric (identity by descent, IBD) is calculated for each pair of individuals based on the average proportion of alleles shared in common at genotyped SNPs (excluding the sex chromosomes)
• The degree of **recent shared ancestry for a pair of individuals** (identity by descent, IBD) can be estimated using genome-wide IBS data using Plink. (IBD shown as pi_hat in plink)

    CALCULATING IBD
    plink -bfile example--genome --out example

• The expectation is that:
  
  IBD = 1 for duplicates or monozygotic twins
  IBD = 0.5 for first-degree relatives,
  IBD = 0.25 for second-degree relatives
  IBD = 0.125 for third-degree relatives

• The IBS method works best when only **independent SNPs** are included in the analysis. Independent SNP set for IBS calculation is generally prepared by removing regions of extended LD and pruning the remaining regions so that no pair of SNPs within a given window (say, 50kb) is correlated.
Genomic control

• In the presence of population structure, the chi-squared statistic $\chi^2$ is inflated by a constant inflation factor $\lambda$,

• Lambda is defined as the empirical median of $L$ unrelated statistics divided by the expected median under the null distribution.

• In theory, $\lambda$ should be equal to 1 in a homogeneous population. So a **value greater than one** implies **population structure**.

• PLINK estimates this value as GIF, while running association test. GIF greater than **1.05** often indicates the presence of a structure.

Other factors could also cause GIF to have higher values (Clayton et al. 2005)
Q-Q plots

- Quantile-quantile (Q-Q) plots could detect the existence of population structure.

- The Q-Q plot is constructed as a scatter plot of the observed ranked P-values from the largest to smallest against the theoretical values under the null hypothesis of no association.

- If the statistics come from null distribution, the plot should go along the diagonal linearly. Large/Early deviation from the diagonal indicates population structure.
Other methods

- Robust methods for population stratification in genome wide association studies
- Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness
- Using Network Methodology to Infer Population Substructure
Section IV.
Correcting for Population Structure and Relatedness

- PCA based
- GIF based correction
- Removing related individuals
- LMMs
- PCA and Kinship matrix as covariates
- Pros and cons of various approaches
Using PCA to remove outliers

Draw a boundary containing the core of cases and control.

Exclude individuals outside the boundary

Hulur et al. 2017
Using PCA to remove outliers

Which ellipse is better? How many to exclude?

http://www.bioinf.wits.ac.za/software/poputils/

Is the correction enough?
Using PCA to remove outliers

How to tackle more complex structure?

Huge drop in sample size and power

Behr et al. 2017
PCA as covariates

• The top $n$ principal components, selected on the basis of distribution of the values are incorporated into the regression testing model as covariates. Depending on the dataset - $n$ usually ranges from 3 to 10.

• Let $Z_{1j}$, $Z_{2j}$, ... denote the eigenvectors for $j$th person

$$g(E(Y|X)) = a + bX + c_1Z_1 + c_2Z_2 + ...$$
Removing related individuals

The expectation is that:
- IBD = 1 for duplicates or monozygotic twins
- IBD = 0.5 for first-degree relatives,
- IBD = 0.25 for second-degree relatives
- IBD = 0.125 for third-degree relatives

Genotyping error, LD and population structure cause variation around these theoretical values and it is typical to remove one individual from each pair with an IBD > 0.1875 (halfway between the expected IBD for third- and second-degree relatives).

In some cases this may lead to loss of lot many individuals.
- There might still be more distant relatedness which is not addressed at higher cutoffs but could lead to spurious associations.

**CALCULATING IBD**

plink -bfile example --genome --out example
Genomic control based correction

The GC based correction implemented in PLINK can be accessed using --gc flag in addition to the --adjust flag, while running association.

```
plink --bfile --assoc --gc --adjust --out results_with_gc
```

\[
\chi^2_{fair} = \chi^2_{biased} / \lambda
\]

Plink provides GC adjusted P-Values

- Implementation is difficult for non-additive model
- The correction applies to all the variants, however, some SNPs exhibit more differences in their allele frequencies than others; thus, the uniform adjustment is inappropriate and leads to a loss of power.
- Threshold for applying GC is empirical.
Linear mixed models are an extension of simple linear models to allow both fixed and random effects, and are particularly used when there is non independence in the data, such as arises from a hierarchical structure.
LMM can address both PS and Relatedness

\[ Y_i = \mu + \sum_{j=1}^{3} \beta_j PC_{ji} + \alpha_i + Line_i + \varepsilon_i \]

- Grand Mean
- Marker effect
- Random effects: account for familial relatedness
- Fixed effects: account for population structure
- Observed SNP alleles of \( i \)th individual
- Random error term

- \( (Line_1, \ldots, Line_n) \sim \text{MVN}(0, 2K\sigma_G^2) \)
- \( K = \text{kinship matrix} \)
- \( \varepsilon_i \sim \text{i.i.d. N}(0, \sigma_E^2) \)

Yu et al. (2006)
While these tools are primarily targeted for quantitative traits, these can also be applied to analyze binary traits, by treating them as quantitative traits. However, using these for unbalanced case-control might provide false associations.

Special modification for case-control using LMM such as liability-threshold mixed linear model (LTMLM) are also available.

<table>
<thead>
<tr>
<th>Method</th>
<th>Requires $O(MN^2)$ time</th>
<th>Avoids proximal contamination</th>
<th>Models non-infinitesimal genetic architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMMAX (ref. 3)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FaST-LMM (ref. 5)</td>
<td>X^d</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>FaST-LMM-Select (refs. 9,11,15)</td>
<td>X^d</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>GEMMA (ref. 6)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRAMMAR-Gamma (ref. 10)</td>
<td>X^d</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>GCTA-LOCO (ref. 12)</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>BOLT-LMM</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

BOLT-LMM only recommended for sample sizes >5000.
Which correction to use?

Structured vs. Homogeneous

Independent vs. Related

Populations Structure

Cryptic relatedness

PCA based correction

Mixed model + PCA based correction

Genomic Control

Mixed model

Meta-analysis

DREAM !

Family based ??
• Population structure often confounds inference and needs to be addressed for a successful GWAS.

• There are a variety of computational approaches to enable meaningful association studies to be conducted in a dataset with considerable structure and/or relatedness.

• Given the diversity of African populations and the history of migration and admixture, observing population structure in samples from most geographies is unsurprising. Do not be happy with P-values unless you have seen the Q-Q plot!

• Different approaches for correction are better suited for different scenarios. Choose your approach judiciously.

• Most of the real life analysis (especially those based on Imputed data) are computationally intensive and require nuanced interpretation.
- Population structure often confounds inference and needs to be addressed for a successful GWAS.

- There are a variety of computational approaches to enable meaningful association studies to be conducted in a dataset with considerable structure and/or relatedness.

- Given the diversity of African populations and the history of migration and admixture, observing population structure in samples from most geographies is unsurprising. Do not be happy with P-values unless you have seen the Q-Q plot!

- Different approaches for correction are better suited for different scenarios. Choose your approach judiciously.

- Most of the real life analysis (especially those based on Imputed data) are computationally intensive and require nuanced interpretation.