Population Substructure and Control Selection in Genome-wide Association Studies

Kai Yu, Ph.D.
Division of Cancer Epidemiology and Genetics, NCI
Acknowledgements

CGEMS & DCEG
Gilles Thomas
Zhaoming Wang
Stephen Chanock
Sholom Wacholder
Qizhai Li
Robert Hoover
Kevin Jacobs
Meredith Yeager
Joseph Fraumeni
Daniela Gerhard
Xiang Deng
Nick Orr
Robert Welch
Nilanjan Chatterjee
Richard Hayes
Margaret Tucker
Marianne Rivera-Silva

HSPH
David Hunter
Peter Kraft
David Cox
Sue Hankinson

CeRePP, France
Olivier Cussenot
Geraldine Cancel-Tassin
Antoine Valeri

ACS
Michael Thun
Heather Feigelson
Eugenia Calle

NPHI, Finland
Jarmo Virtamo

Wash. U., St Louis
Gerald Andriole

biöwulf
AT THE NIH
Background

- **Genome wide association studies (GWAS) based on case-control design**
  - Compare genotype frequency at each genetic markers (SNP)
- **Population stratification (PS)**
  - Genotype frequency differences at a given SNP between cases and controls due to ancestry differences (confounding by ethnicity).
PS example: $LCT$ and height (Campbell et al., 2005)

<table>
<thead>
<tr>
<th></th>
<th>Matching on four grandparents</th>
<th>Ancestry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Four US-born</td>
</tr>
<tr>
<td>P-value</td>
<td>$3.6 \times 10^{-7}$</td>
<td>0.098</td>
</tr>
</tbody>
</table>

Note: after adjustment for the three classes, the P-value is 0.0074
More on PS

- PS can occur in a case-control study conducted in a non-homogeneous population
  - Due to disease risk heterogeneity across (hidden) subpopulations
  - Due to sampling bias that results into ancestry background difference between cases and controls
Motivation

• Longstanding debate on the impact of PS on well-designed genetic studies
• The temptation to use external controls to save costs (using controls from another study, using shared controls)
Focus of This Talk

Using empirical data from CGMES

• Evaluate the impact of PS in GWAS conducted in *European Americans* with different sample selection strategies
  – Nested case-control design
  – The use of external controls

• How to effectively correct for PS
Identifying Genetic Markers for Prostate & Breast Cancer

Genome-Wide Analysis
Public Health Problem
Prostate (1 in 8 Men)
Breast (1 in 9 Women)
Analyze Long-Term Studies
NCI PLCO Study
Nurses’ Health Study (NHS)

Initial Study
Follow-up #1
Follow-up #2
Establish Loci

Fine Mapping
Functional Studies
Validate Plausible Variants
Possible Clinical Testing

http://cgems.cancer.gov
Material for Analysis

- **PLCO (Prostate, Lung, Colorectal and Ovarian cancer screening trial)**
  - Men from a randomized trial for cancer prevention
  - Removing subjects with European admixture coefficient <90%
  - 1,171 prostate cancer cases
  - 1,094 controls

- **NHS (Nurses’ Health Study)**
  - Women from a prospective cohort study on nurses
  - Removing subjects with European admixture coefficient <90%
  - 1,140 breast cancer cases
  - 1,138 controls

- # testing autosomal SNP: 450K
  - >5% minor allele frequency in PLCO and in NHS
  - <5% missing rate in PLCO and in NHS
Null markers are useful
Because of the availability of many null SNPs in GWAS

– Monitor extent of PS
  • Q-Q plot, inflation factor
– Estimate the population ancestry and correct for PS (at the cost of power)
  • PCA: capture correlation between genotypes to identify axes with large genetic variation
  • STRUCTURE: Attempts to interpret the correlation between genotypes in terms of admixture among a defined number of ancestral populations
241,238 genomic control SNPs for over-dispersion evaluation

475,118 testing SNPs

The first set of 12,898 PCA SNPs for population substructure inference

The second set of 7,017 PCA SNPs for population substructure inference

Low interlocus LD
Using PCA to study population substructure

Summarize the information measured on $N$ structure inference SNPs and represents study participants in a lower dimensional space so that the Euclidean distance between two subjects represents their genetic difference.
An Illustration for PCA
PCA of joint sample of HapMap and NHS

PCA_HapMap(208)_NHS(2287)
PCA in CGEMS PLCO and NHS GWAS
Principal component comparisons (P-values) between cases and controls based on the Wilcoxon rank-sum test

<table>
<thead>
<tr>
<th></th>
<th>PLCOca-PLCOco</th>
<th>PLCOca-NHSco</th>
<th>NHSca-NHSco</th>
<th>NHSca-PLCOco</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC #1</td>
<td>0.294</td>
<td>4.5 x 10^{-8}</td>
<td>0.664</td>
<td>4.3 x 10^{-6}</td>
</tr>
<tr>
<td>PC #2</td>
<td>0.871</td>
<td>2.2 x 10^{-7}</td>
<td>0.289</td>
<td>6.9 x 10^{-12}</td>
</tr>
<tr>
<td>PC #3</td>
<td>0.340</td>
<td>0.282</td>
<td>0.036</td>
<td>4.0 x 10^{-3}</td>
</tr>
<tr>
<td>PC #4</td>
<td>0.588</td>
<td>1.2 x 10^{-4}</td>
<td>0.015</td>
<td>0.191</td>
</tr>
<tr>
<td>PC #5</td>
<td>0.490</td>
<td>0.385</td>
<td>0.943</td>
<td>0.157</td>
</tr>
</tbody>
</table>
Observations I

• Similar population sub-structure patterns in GWAS conducted in PLCO and NHS
  – The exchange of controls may be feasible

• Demonstrable genetic background difference between the two GWAS, partially due to
  – Difference in geographic locations of the two source populations
Inflation factor (IF)

Let $T_i$, $i = 1, \ldots, M$, be the association test statistics for all testing markers, the inflation factor $\lambda$ can be defined as

$$\lambda = \frac{\text{median}\{T_i, \ i = 1, \ldots, M\}}{\text{median of } T \text{ under the null}}.$$

For example, if using 1 df Chi-squared test,

$$\lambda = \frac{\text{median}\{T_i, \ i = 1, \ldots, M\}}{0.455}.$$

If using 2 df Chi-squared test,

$$\lambda = \frac{\text{median}\{T_i, \ i = 1, \ldots, M\}}{1.386}.$$
Q-Q Plot for the test without PC adjustment

PLCOca-PLCOcn

IF = 1.025

NHSca-NHScn

IF = 1.005

PLCOca-NHScn

IF = 1.090

NHSca-PLCOcn

IF = 1.062
PC selection strategies for the correction of PS

$$\log \frac{p}{1-p} = \alpha + u_1 \beta_1 + u_2 \beta_2 + g \gamma$$

- Select a fixed number of PCs (e.g., top 10 PCs)
- Select PCs with “large” genetic variations (e.g., PCs with Tracy-Widom test P-value < 0.05)
- Select PCs correlated with the outcome
A Algorithm to Select PCs for PS correction

Order the top \( L \) PCs according to their Wilcoxon rank-sum statistics, define them as \( u_1, u_2, \ldots, u_L \). First, to evaluate whether to include \( u_1 \) for the PS adjustment.

1. Obtain the inflation factor \( \lambda_1 \) based on

\[
\log \frac{p}{1-p} = \alpha + u_1 \beta_1 + g \gamma.
\]

2. Random permute \( u_1 \) to \( u_1^{(b)}, b = 1, \ldots, B \). Based on \( u_1^{(b)} \), obtain \( \lambda_1^{(b)} \) from

\[
\log \frac{p}{1-p} = \alpha + u_1^{(b)} \beta_1 + g \gamma. \lambda_1^{(b)}, b = 1, \ldots, B.
\]

3. Estimate the empirical P-value as \( p = \# \{ \lambda_1 > \lambda_1^{(b)}, b = 1, \ldots, B \} / B \).

4. Include \( u_1 \) if \( p \) is small (say < 0.05), not include \( u_1 \) otherwise.
Algorithm (cont.)

Suppose we have chosen \( u_2 \) and \( u_3 \), and try to decide whether to include \( u_5 \), do

1. Obtain the inflation factor \( \lambda_5 \) based on

\[
\log \frac{p}{1-p} = \alpha + u_2 \beta_2 + u_3 \beta_3 + u_5 \beta_5 + g \gamma.
\]

2. Random permute \( u_5 \) \( B \) times to have \( u_5^{(b)} \), \( b = 1, \ldots, B \). Based on \( u_5^{(b)} \), obtain \( \lambda_5^{(b)} \) from

\[
\log \frac{p}{1-p} = \alpha + u_2 \beta_2 + u_3 \beta_3 + u_5^{(b)} \beta_5 + g \gamma. \quad b = 1, \ldots, B.
\]

3. Estimate the empirical P-value as \( p = \# \{ \lambda_5 > \lambda_5^{(b)} \}, \quad b = 1, \ldots, B \}/B \)

4. Include \( u_5 \) if \( p \) is small (say <0.05), not include \( u_5 \) otherwise.
## PCs selected

<table>
<thead>
<tr>
<th>PC #</th>
<th>$PLCOca$-</th>
<th>$PLCOca$-</th>
<th>$NHSca$-</th>
<th>$NHSca$-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$PLCOco$</td>
<td>$NHSco$</td>
<td>$NHSco$</td>
<td>$PLCOco$</td>
</tr>
<tr>
<td>PC #1</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>PC #2</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>PC #3</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>PC #4</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PC #5</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Over-dispersion factor for association tests with adjustment for various numbers of PCs

<table>
<thead>
<tr>
<th>PCs chosen for the adjustment</th>
<th>PLOCca-PLCOco</th>
<th>PLCOca-NHSco</th>
<th>NHSc-</th>
<th>NHSc-PLCOco</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 PC</td>
<td>1.025</td>
<td>1.090</td>
<td>1.005</td>
<td>1.062</td>
</tr>
<tr>
<td>1st PC</td>
<td>1.020</td>
<td>1.055</td>
<td>1.006</td>
<td>1.040</td>
</tr>
<tr>
<td>1-2 PCs</td>
<td>1.022</td>
<td>1.040</td>
<td>1.004</td>
<td>1.013</td>
</tr>
<tr>
<td>1-3 PCs</td>
<td>1.021</td>
<td>1.040</td>
<td>1.005</td>
<td>1.006</td>
</tr>
<tr>
<td>1-4 PCs</td>
<td>1.021</td>
<td>1.032</td>
<td>1.005</td>
<td>1.007</td>
</tr>
<tr>
<td>1-5 PCs</td>
<td>1.023</td>
<td>1.032</td>
<td>1.006</td>
<td>1.008</td>
</tr>
<tr>
<td>1-10 PCs</td>
<td>1.025</td>
<td>1.036</td>
<td>1.008</td>
<td>1.010</td>
</tr>
<tr>
<td>Selected PCs</td>
<td>1.020</td>
<td>1.032</td>
<td>1.003</td>
<td>1.006</td>
</tr>
</tbody>
</table>
Q-Q Plot for the test with and without PC adjustment

unadjusted

PLCOca-PLCOcn

Log10(Observed P-value) vs. log10(Expected P-value)

IF = 1.025

adjusted

PLCOca-PLCOcn

Log10(Observed P-value) vs. log10(Expected P-value)

IF = 1.020

PLCOca-NHScn

Log10(Observed P-value) vs. log10(Expected P-value)

IF = 1.032

PLCOca-NHScn

Log10(Observed P-value) vs. log10(Expected P-value)

IF = 1.090
Q-Q Plot for the test with and without PC adjustment

unadjusted

NHSca-NHScn

log10(Expected P-value)

log10(Observed P-value)

IF = 1.005

NHSca-PLCOcn

adjusted

NHSca-NHScn

log10(Observed P-value)

log10(Expected P-value)

IF = 1.003

NHSca-PLCOcn

log10(Observed P-value)

log10(Expected P-value)

IF = 1.006
Discussions

- We observed population heterogeneity exists within the European American population
- PS does not appear to be a problem in well-design studies
- Problem of PS is more extensive when external controls are used, but the adjustment of PCs can help
- We used empirical data for European Americans, what about other populations, such as African Americans?
- More issues to be considered when using “external controls”, such as,
  - Power issue
  - Covariate measurement harmonization
Discrepancy in SNP selection before and after PC adjustment
(selecting top 5% ranked SNPs)

PLCO cases vs. PLCO controls

PLCO cases vs. NHS control
Rank shuffling in PLCOca-PLCOcn

![Graph showing rank distribution](image)
Rank shuffling in PLCOca-NHScn

Rank Distribution

0-1 1-2 2-3 3-4 4-5 5-6 6-7 7-8 8-9 9-10 >10
0 30 60 0 30 60 0 30 60 0 30 60 0 30 60
PS-I example: *LCT* and height

<table>
<thead>
<tr>
<th></th>
<th>Matching on four grandparents</th>
<th>Ancestry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Four US-born</td>
</tr>
<tr>
<td>P-value</td>
<td>3.6 x 10^{-7}</td>
<td>0.098</td>
</tr>
</tbody>
</table>

Note: after adjustment for the three classes, the P-value is 0.0074
Campbell et al. (NG, 2005)
Sample selection and PS-II

Assuming common disease risk, any sampling bias that leads to ancestral difference will cause PS-II.

- **Nested case-control design**
  - the source population (cohort) is well defined
  - Minimal systematic bias in case control collection

- **Standard case-control design**
  - source population is not well defined
  - Control participation rate difference across subpopulations

- **External controls (shared controls, freezer controls)**
  - Cases and controls could represent different populations
Check of loadings (r2<0.004)
Check of loadings ($r^2<0.01$)