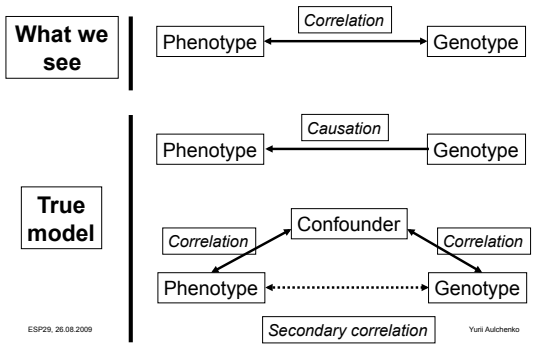
GWA in presence of population stratification

Yurii Aulchenko
Erasmus MC Rotterdam
26.08.2009

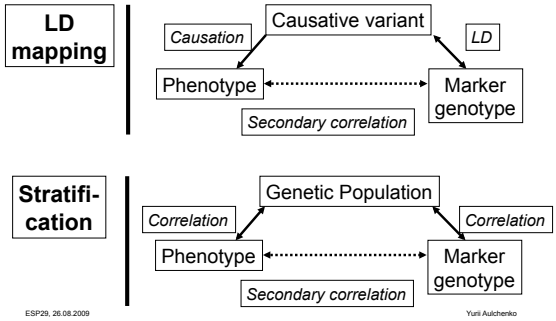
Outline

- Confounding and stratification in GWA studies
- Genomic Control and Structured Association
- PCA correction (EIGENSTRAT)
- Quality Control (QC) of genetic data

Reasons for genetic association



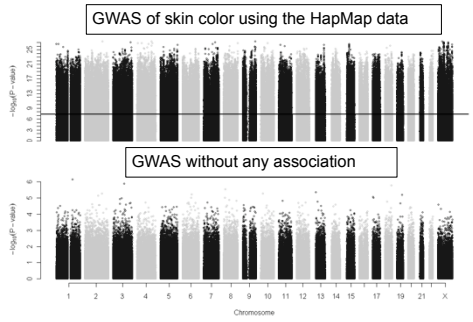
Confounding in genetic studies



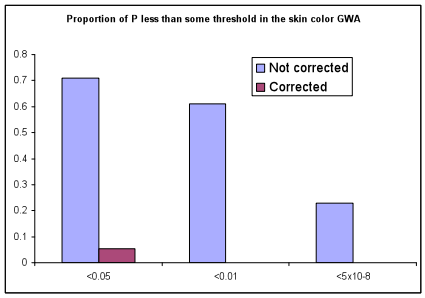
Stratification

- Some factor is a confounder for genotypes and disease prevalence
  - Dark skin is more prevalent in people Africans than in Europeans. The genotypic frequencies are also different between two populations.
  - A study of skin color, which would mix Africans and Europeans is likely to generate multiple false positives
- Other causes of genetic stratification are “cryptic” relations or systematic pedigree structure presented in a sample

Skin color scan



Consequences of stratification

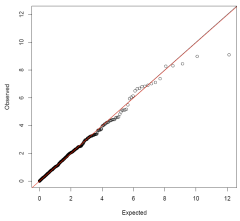


Outline

- Confounding and stratification in GWA studies
- **Genomic Control and Structured Association**
- PCA correction (EIGENSTRAT)
- Quality Control (QC) of genetic data

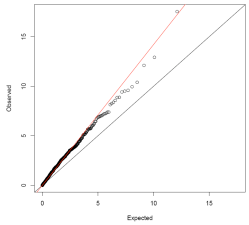
Distribution of the test statistics under the null hypothesis

- 200 random SNPs
- In Linkage Equilibrium
- Not related to the disease
- No stratification
- The distribution of the test statistics for association is χ^2_1



Idea of the genomic control

- There is stratification
- **Assumption: stratification acts in the same manner across all loci**
- This leads to uniform inflation of the test statistics
- The distribution of the test statistics is $\lambda \cdot \chi^2_1$ ($\lambda \geq 1$)



Genomic control

- Consider a test distributed as χ^2_1 under the null (e.g. trend test)
- Select N (>200) independent SNPs and compute the vector of test statistics $\{T^2_1, T^2_2, T^2_3, \ldots, T^2_{N-1}, T^2_N\}$
- Estimate λ as
  - Median$\{T^2_1, T^2_2, T^2_3, \ldots, T^2_{N-1}, T^2_N\}$ /0.456
  - Slope of regression of observed onto expected
- The GC-corrected test statistics
  - $T^2/\lambda \sim \chi^2_1$
- In practice, all (or large proportion of) GW test are used to estimate λ

λ is dependent on sample size

- λ is related to non-centrality parameter, thus it grows with sample size. Therefore λ should be estimated per certain sample size. This is especially important if
  - SNP call rate is very different between SNPs
  - When reporting the results
- For QT analysis, $\lambda_n = 1 + (\lambda_{nref} - 1)\, n/n_{ref}$
  where n_{ref} is the reference sample size
- For case/control design

$$\lambda_{n_j,m_j} = 1 + \left(\lambda_{n_{ref},m_{ref}} - 1\right)\left(\frac{1}{n_{ref}} + \frac{1}{m_{ref}}\right) \Big/ \left(\frac{1}{n_j} + \frac{1}{m_j}\right)$$

where n & m refer to size of samples of cases and controls

When GC does not work (well)?

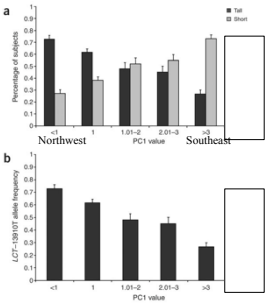
When stratification is large (say, $\lambda_{1000} > 1.1$) other, more powerful methods are to be used

GC assumes that stratification acts in the same manner across all loci

- This is not true for loci differentiated between population e.g. because of selection
- Such loci will still be falsely detected after GC correction

Example: association of stature to LCT

- SNP in the lactase (LCT) gene was strongly associated with height ($P<10^{-6}$)
- GC λ was 1.0
- The LCT SNP is selected and differentiated between European populations
- Little evidence left after applying structured association



•*Campbell et al, Nat Genet 2005*

Structured association (SA)

- Identify genetic populations (strata)
- Mantel-Haenszel test for structured association
- Basically, components of the score test (association score and its' variance) are computed in each strata separately. These could be added up and give single test
- Apply GC to correct for residual inflation ($1 < \lambda < 1.1$)
- Problems with SA
  - Strata not always known or easy to identify
  - Is not powerful when there is a strong case/control mismatch

Outline

- Confounding and stratification in GWA studies
- Genomic Control and Structured Association
- **PCA correction (EIGENSTRAT)**
- Quality Control (QC) of genetic data

Idea of Multidimensional Scaling

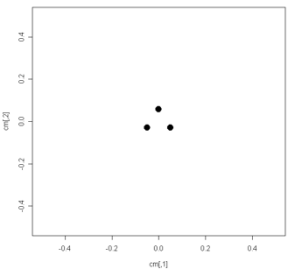
- Study of *N* subjects
- *NxN* matrix of pair-wise distances (0 = the same subject, 1 = very different)
- Multi-Dimensional (MD) scaling takes this matrix
  - Returns coordinates for *N* points in a MD-space
  - The vectors are called "Principal Axes of Variation" (or Principal Components)
  - The distance between the points in this MD-space are as close as possible to the distances observed in the original *NxN* matrix
- Classical MDS is also known as Principal Components Analysis

Example CMDS

- Distance matrix

	ID1	ID2	ID3
ID1	0	0.1	0.1
ID2	0.1	0	0.1
ID3	0.1	0.1	0

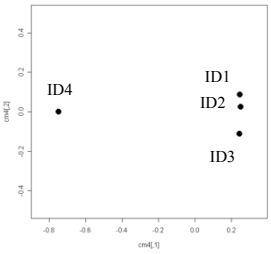
- Results of CMDS:
- PC1 PC2
- ID1 0.00 0.29
- ID2 -0.25 -0.14
- ID3 0.25 -0.14



Example CMDS

- Distance matrix

	ID1	ID2	ID3	ID4
ID1	0	0.1	15	1.00
ID2	0.1	0	0.20	1.00
ID3	0.15	0.20	0	1.00
ID4	1.00	1.00	1.00	0



- Results of CMDS:
- PC1 PC2
- ID1 0.25 0.02
- ID2 0.25 0.09
- ID3 0.25 -0.11
- ID4 -0.75 0.00

Relationship matrix from genomic data

2 x Kinship between people *i* and *j* is the expected proportion of genome shared identical by descent

Distance matrix: 0.5 - kinship

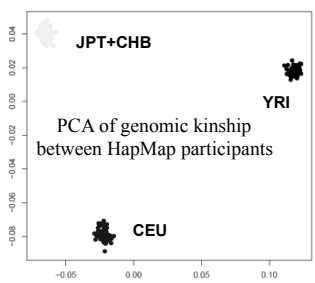
Genomic estimate of kinship between *i* and *j* is computed with

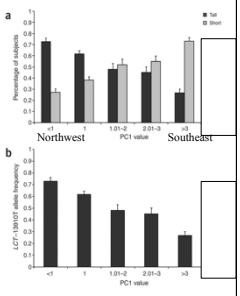
$$f_{ij} = \frac{1}{n}\sum_{k=1}^{n}\frac{(g_{ik}-p_k)(g_{jk}-p_k)}{p_k(1-p_k)}$$

g_{ik} is the genotype (0, 0.5, 1) of the *i*-th person at *k*-th SNP

p_k is the frequency of "1" allele

PCA of genomic kinship



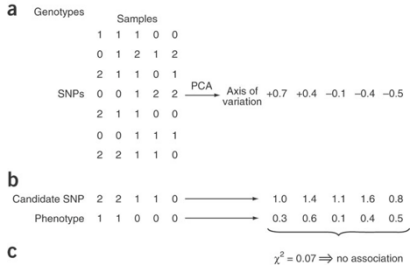


Idea of EIGENSTRAT method

- Quantify genetic origin of study participants with a number (3 to 10) principal axes of variation returned from CMDS analysis of genomic kinship matrix
- In analysis of association, adjust both phenotypes and genotypes for these principal axes of variation
- Apply GC to correct for residual inflation ($1 < \lambda < 1.1$)
- EIGENSTRAT can also (at least partly) correct for differences between genotyping cohorts

 •Price et al, Nat Genet 2006

EIGENSTRAT method



 •Price et al, Nat Genet 2006

Using genomic kinship PC as covariates

- One can use PCs for adjustment using liner model

Trait ~ m + SNP + PC_1 + PC_2 + …

- A problem of using PCs: what PCs to include? How many?

Summary

- If homogeneous group is studied
  - Detect (hopefully few) genetic outliers
  - Remove them from analysis
  - Apply GC to correct for residual stratification
  - Verify findings with EIGENSTRAT
- If multiple strata are expected by design
  - Identify genetic strata
  - Cross-validate with external information
  - If case/control matching is good, apply SA
  - Else, apply EIGENSTRAT analysis
- If strata are not known/difficult to identify
  - Apply EIGENSTRAT / PC adjustment
  - Apply methods using genomic kinship matrix as a whole