GWA in presence of population stratification

Yurii Aulchenko

Erasmus MC Rotterdam

29.08.2007

Outline

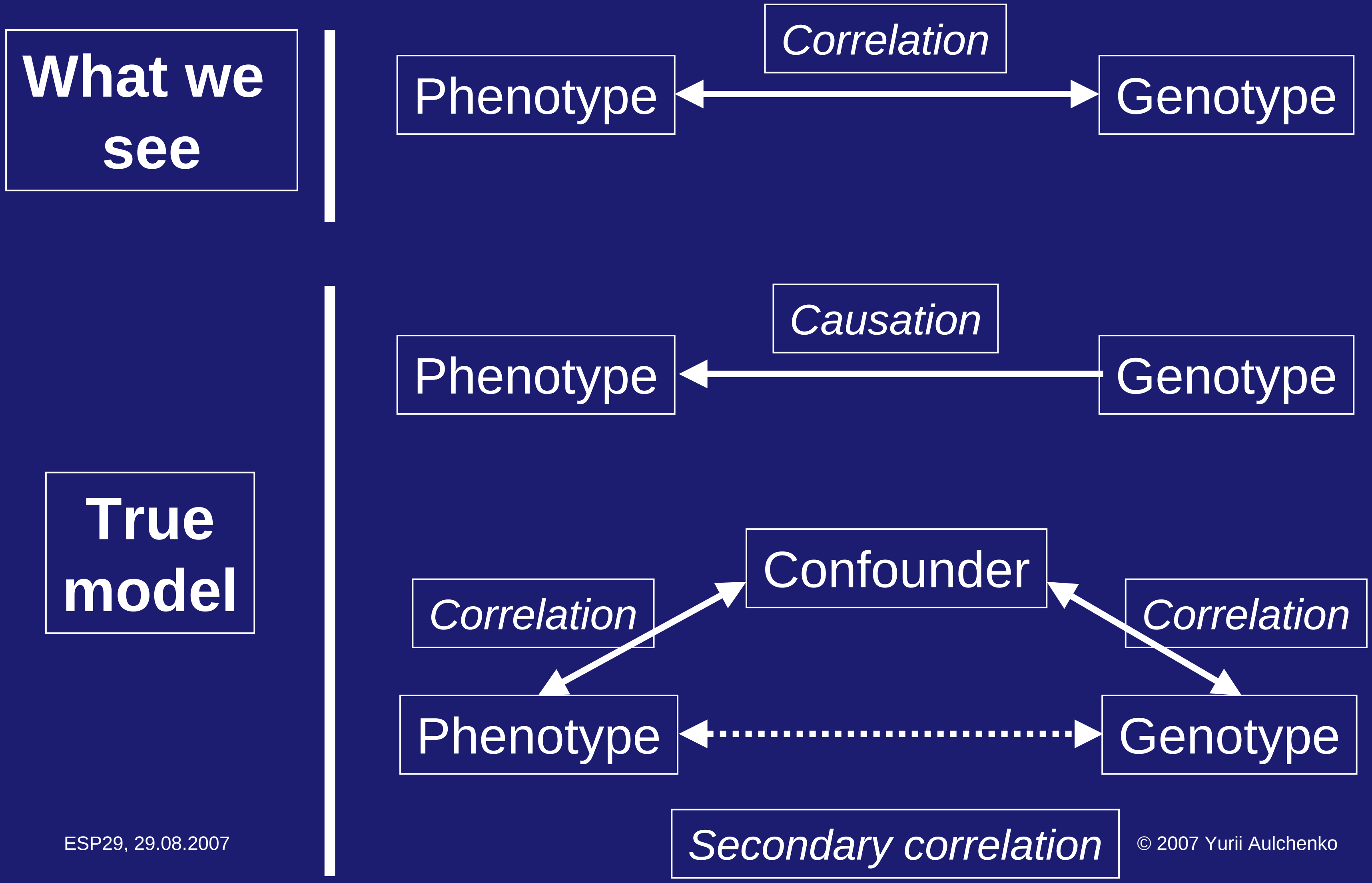
Confounding and stratification in GWA studies

Genomic Control and Structured Association

PCA correction (EIGENSTRAT)

Quality Control (QC) of genetic data

Reasons for genetic association



What we see

Phenotype

Correlation

Genotype

True model

Phenotype

Causation

Genotype

Confounder

Correlation

Correlation

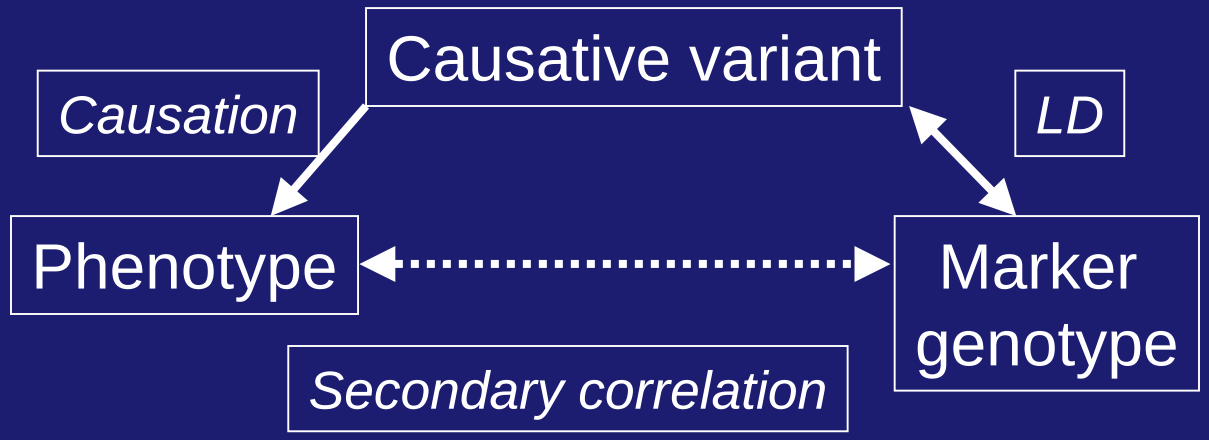
Phenotype

Genotype

Secondary correlation

Confounding in genetic studies

LD mapping



Causative variant

Causation

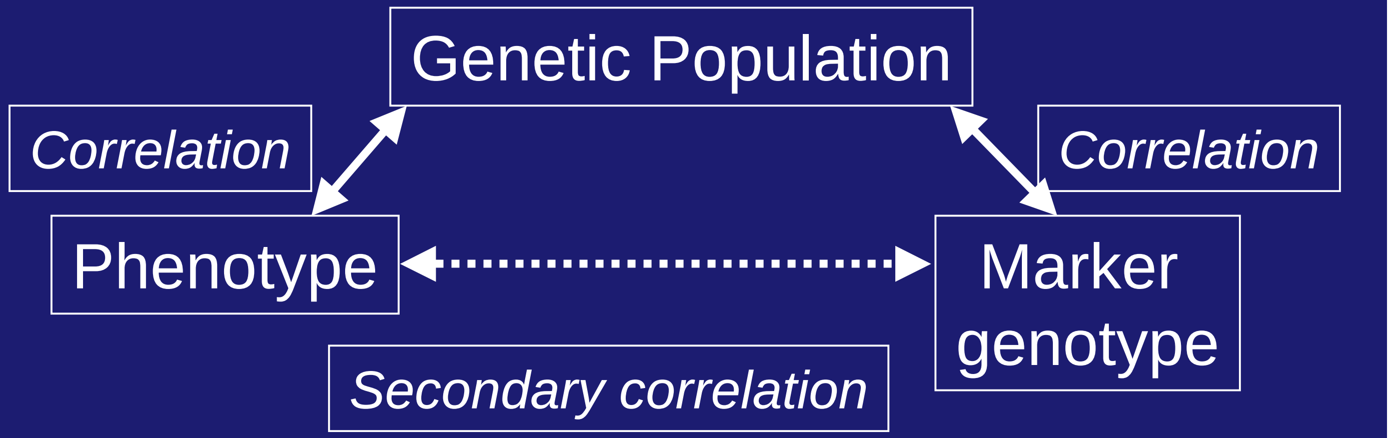
LD

Phenotype

Marker genotype

Secondary correlation

Stratification



Genetic Population

Correlation

Correlation

Phenotype

Marker genotype

Secondary correlation

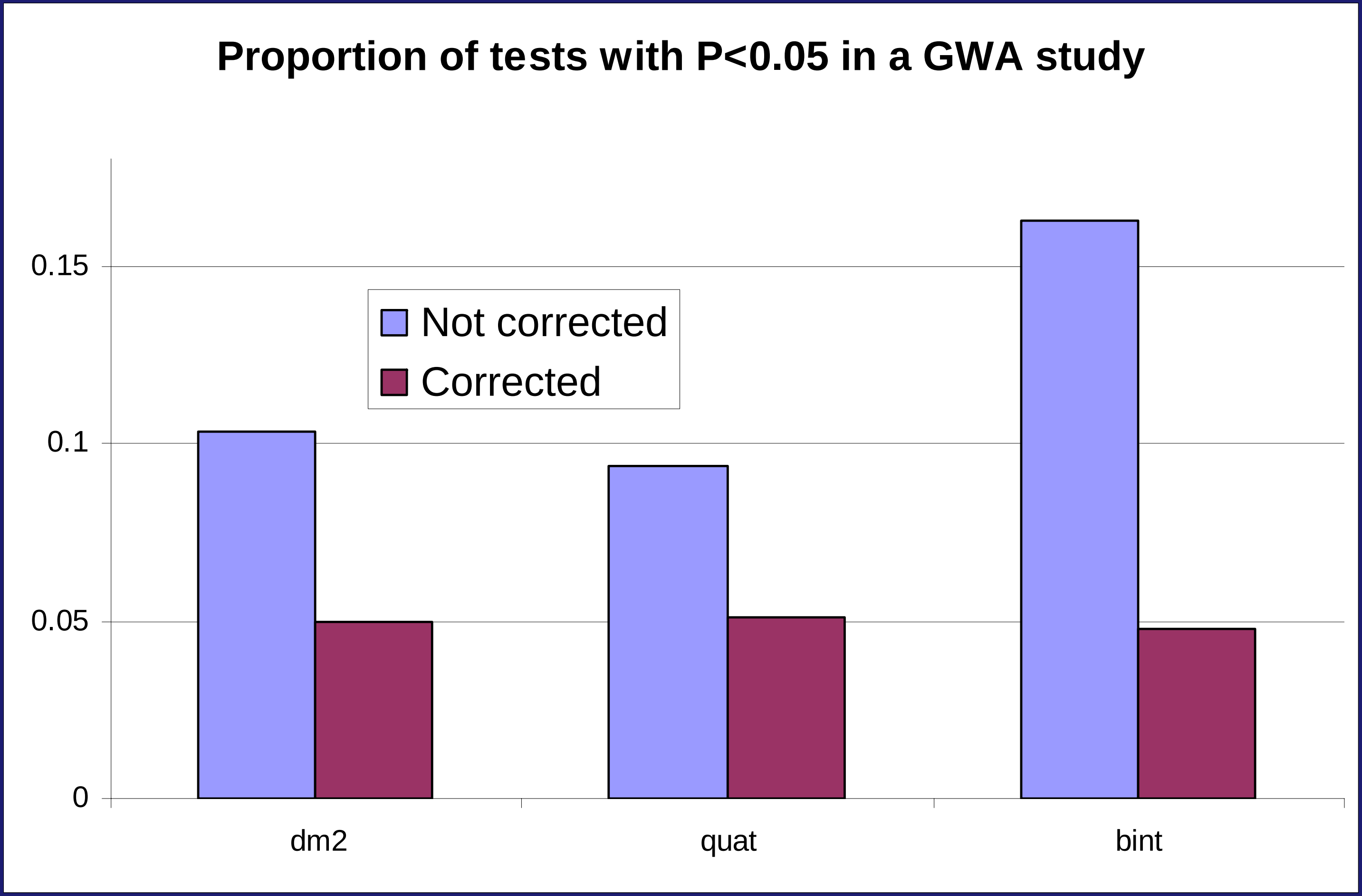
Stratification

Some factor is a confounder for genotypes and disease prevalence

- Chopstick eating behavior is more prevalent in Japanese than in Europeans. The genotypic frequencies are also different between two populations.
- A study of eating habits, which would mix Japanese 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

Consequences of stratification



Outline

Confounding and stratification in GWA studies

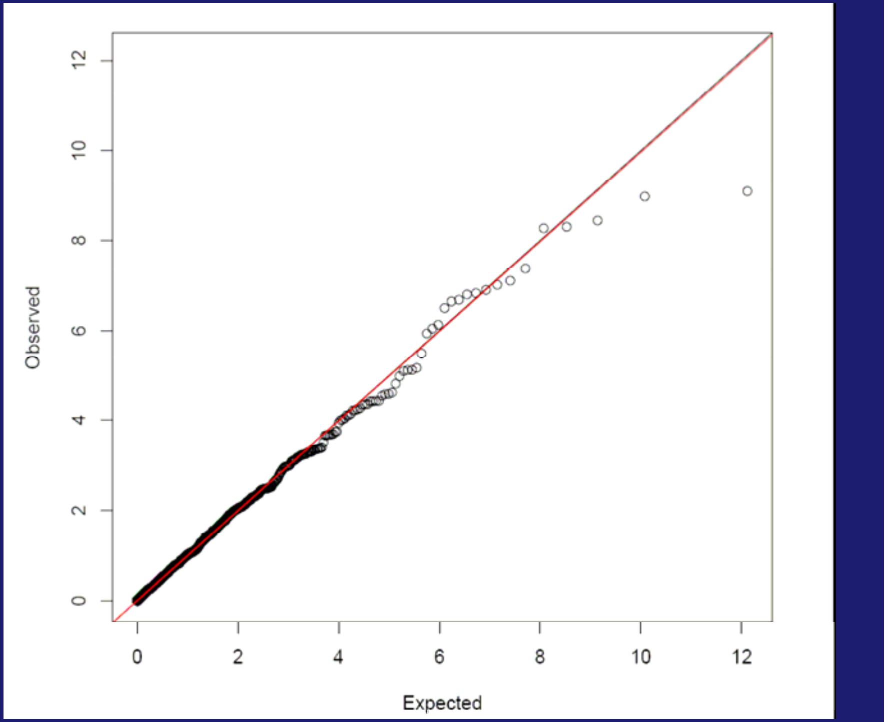
Genomic Control and Structured Association

PCA correction (EIGENSTRAT)

Quality Control (QC) of genetic data

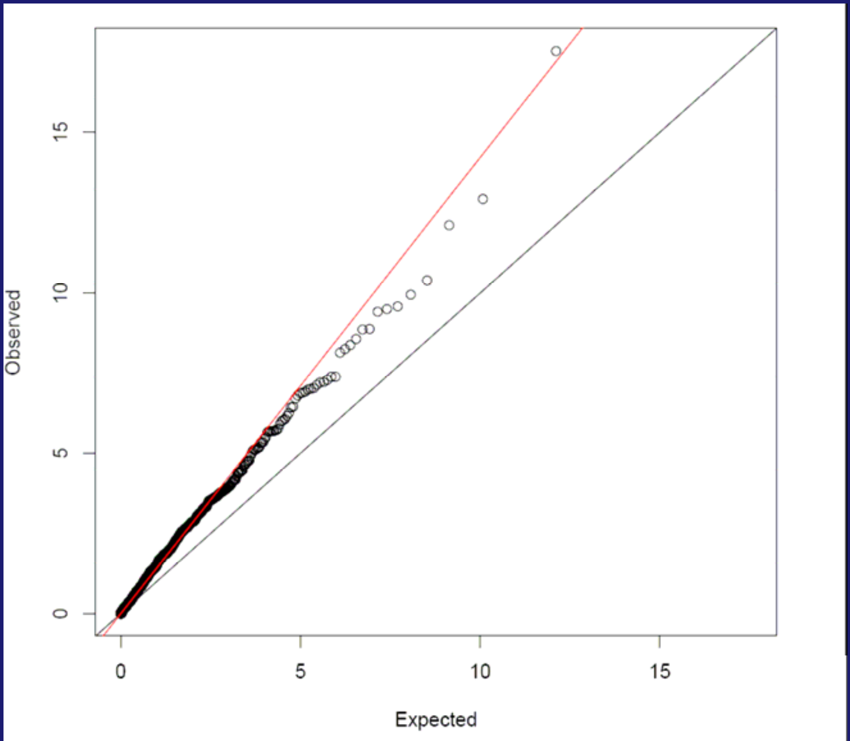
Distribution of the test statistics under the null hypothesis

- 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
  - $\text{Median}\{T^2_1, T^2_2, T^2_3, \ldots , T^2_{N-1}, T^2_N\} /0.456$
  - 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

When GC does not work (well)?

When stratification is large (say, $\lambda > 1.1$) 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

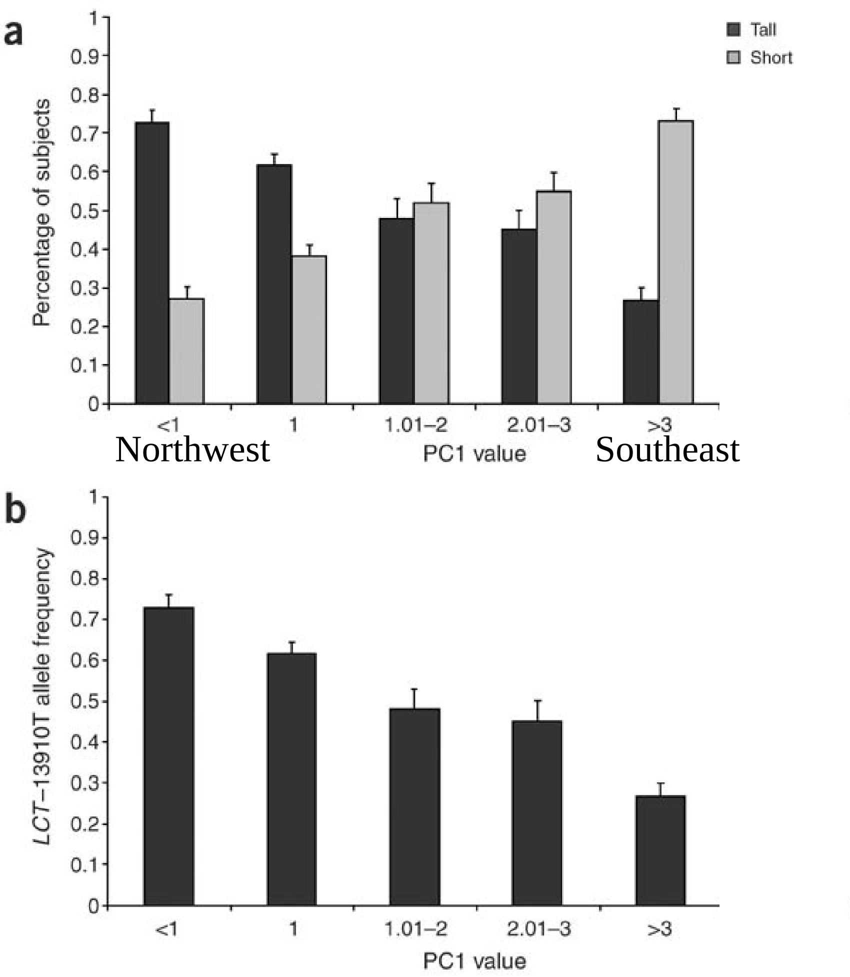
Example: association of stature to LCT

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

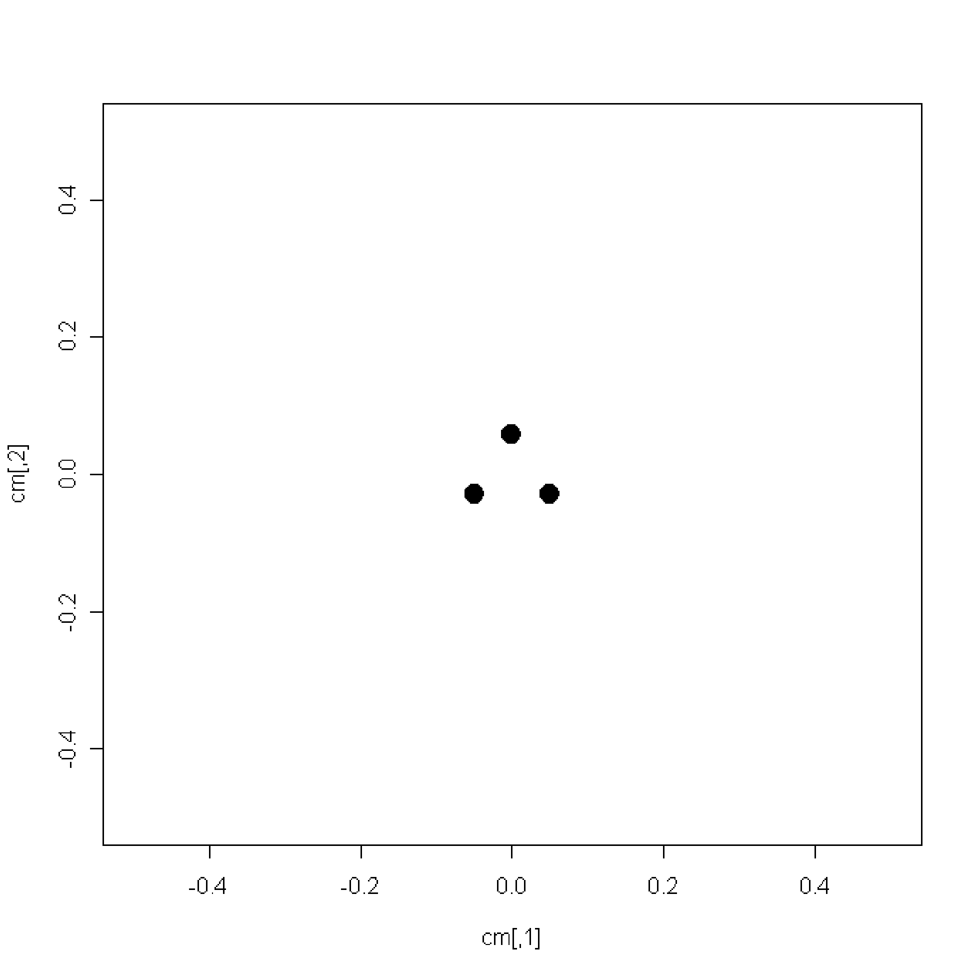
Example CMDS

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



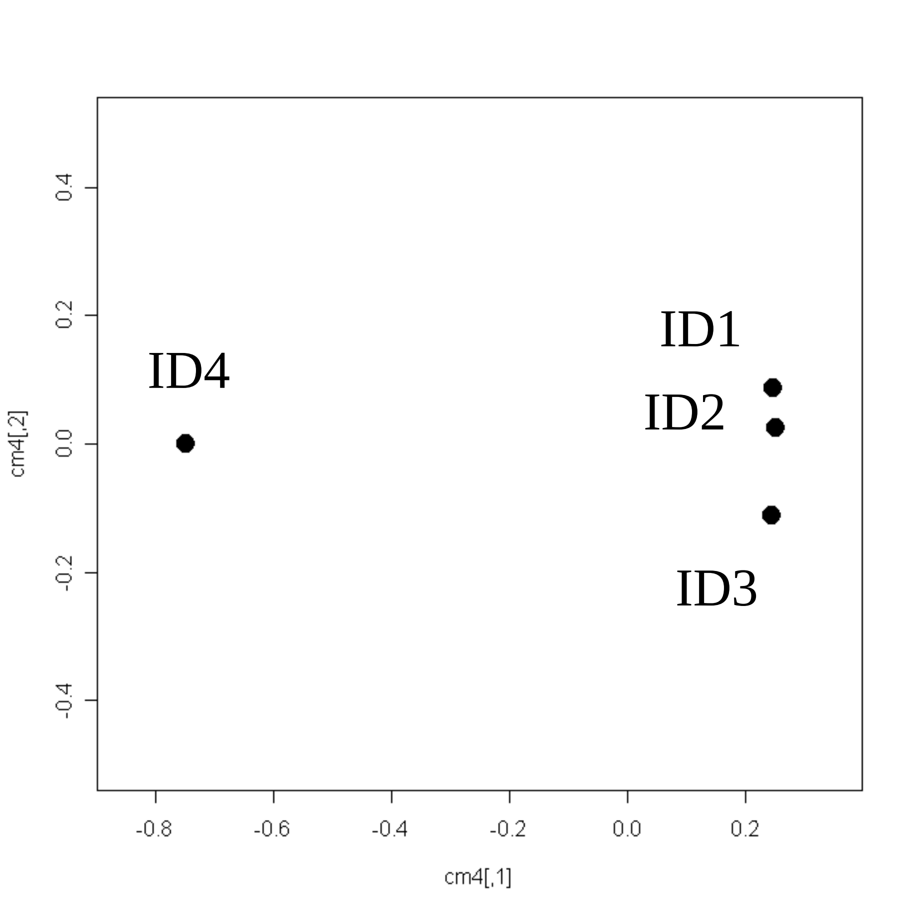
Example CMDS

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

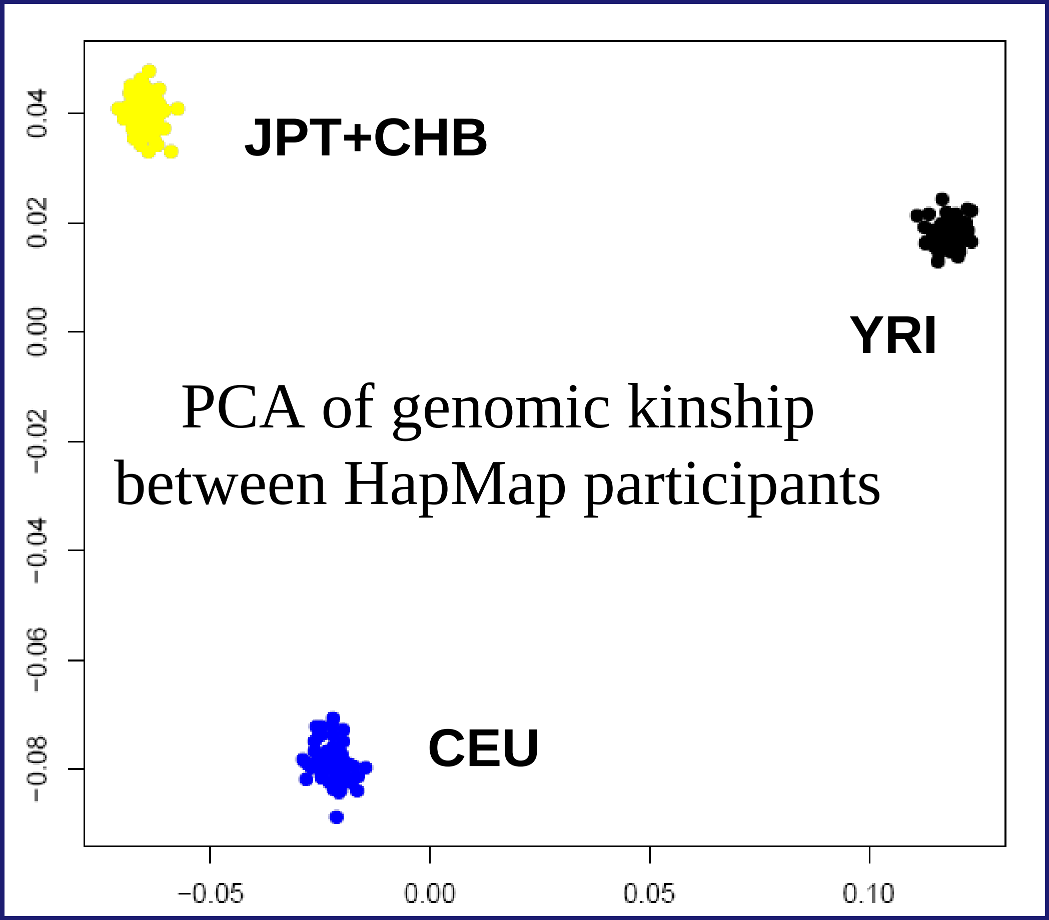
Genomic estimate of kinship between *i* and *j* is computed with

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

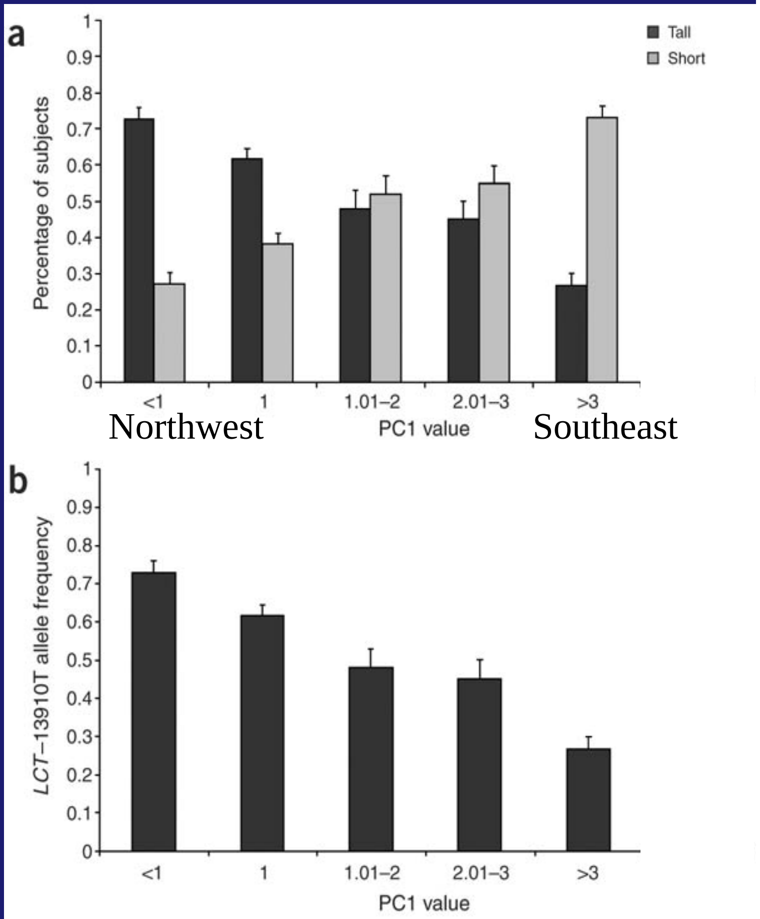
g_{ik} is the genotype (0, 0.5, 1) of the *i*-th person at *k*-th SNP

p_k is the frequency of “1” allele

PCA of genomic kinship



PCA of genomic kinship between HapMap participants



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

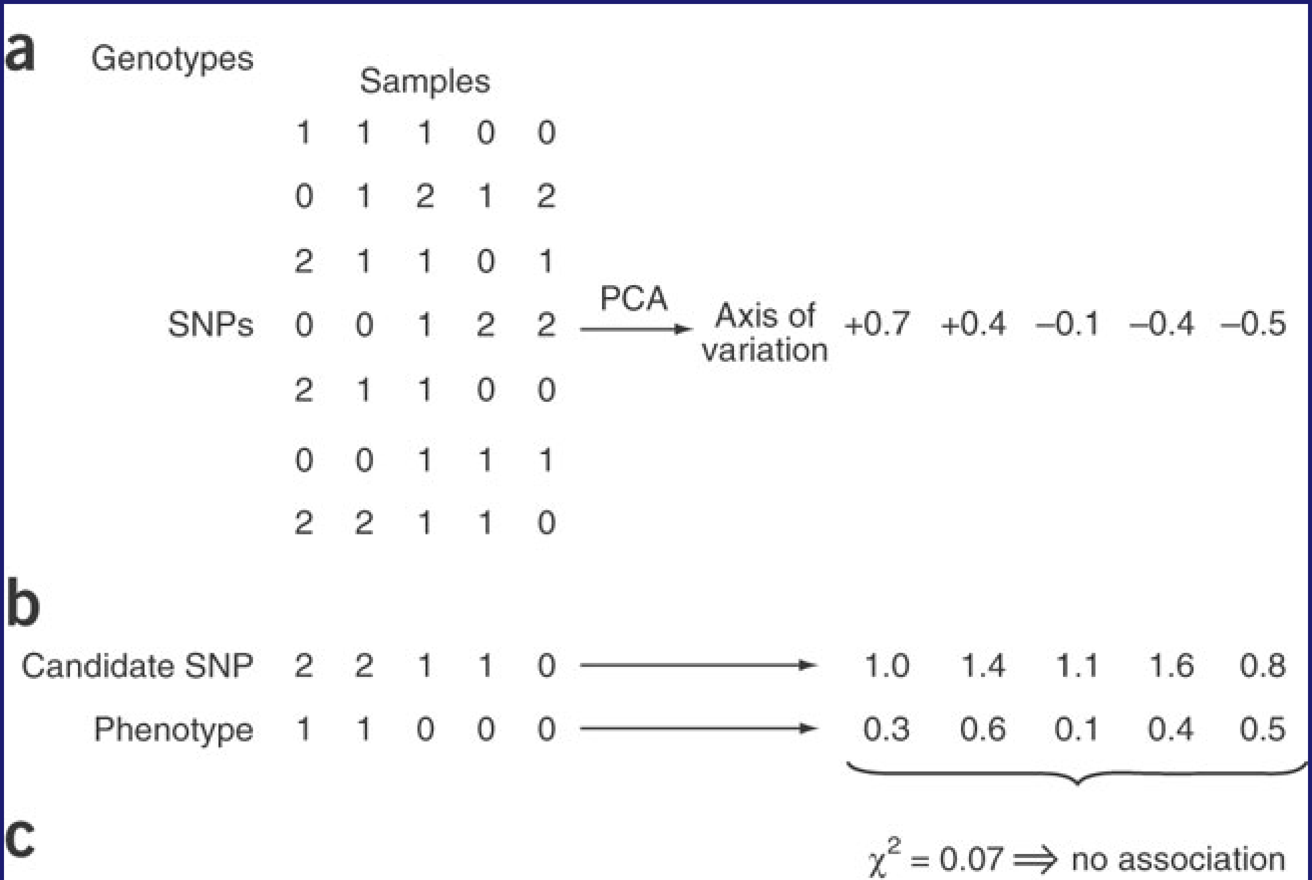
In analysis of association, adjust both phenotypes and genotypes for these principal axes of variation

Apply GC to correct for residual inflation ($1 < \lambda < 1.1$)

Apparently EIGENSTRAT can also pick up and correct for differences between genotyping cohorts

Price et al, Nat Genet 2006

EIGENSTRAT method



a Genotypes

	Samples				
	1	1	1	0	0
	0	1	2	1	2
	2	1	1	0	1
SNPs	0	0	1	2	2
	2	1	1	0	0
	0	0	1	1	1
	2	2	1	1	0

$\xrightarrow{\text{PCA}}$ Axis of variation +0.7 +0.4 −0.1 −0.4 −0.5

b

Candidate SNP	2	2	1	1	0	→	1.0	1.4	1.1	1.6	0.8
Phenotype	1	1	0	0	0	→	0.3	0.6	0.1	0.4	0.5

c

$\chi^2 = 0.07 \Rightarrow$ no association

Price et al, Nat Genet 2006

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

Outline

Confounding and stratification in GWA studies

Genomic Control and Structured Association

PCA correction (EIGENSTRAT)

Quality Control (QC) of genetic data

Sources of genetic data errors

DNA sample swaps

- Same DNA twice
- Plate swap (180°)

Bad quality of material

- Low concentration/amount of DNA
- Contaminated DNA

Imperfect technology

- Calling errors
- “Failed” SNPs
- Sporadic errors

Errors in design

- Unexpected population stratification
- Unexpected presence of related individuals

Consequences

Source	Consequence	Detection	How to deal with
DNA swaps	$(1-\beta)$ ↓	Identical genotypes GW	Remove
Sex errors	$(1-\beta)$ ↓	Male X het, Female X hom	Remove or fix
Low DNA	$(1-\beta)$ ↓	Low personal call rate	Remove
Contam. DNA	$(1-\beta)$ ↓	High heterozygosity	Remove
Calling errors	$(1-\beta)$ ↓	SNP is out of HWE	Remove or fix
Failed SNPs	$(1-\beta)$ ↓	Low SNP call rate	Remove
Sporadic err.	$(1-\beta)$ ↓	Possible only for X	Remove
Genetic strat.	α ↑	Multiple SNPs out of HWE; Special methods	Remove or special

It is assumed that genotyping errors occur at random
α: type 1error
$(1-\beta)$: power

QC procedure

(1) selection of people checks based on

- Selection of SNPs
  - Per-SNP call rate
  - X-markers with multiple heterozygous males
  - *Low Minor allele frequency (???)*
- Selection of people
  - Per-person call rate
  - Males heterozygous for multiple X-markers
  - Females homozygous for multiple X-markers
  - Heterozygosity
  - GW identity of genotypes between people

(2) Detection of possible genetic outliers/strata

(3) Repeat (1) + *HWE checks (???)*, fix sporadic X errors