Confounding in genome-wide studies

Yurii Aulchenko

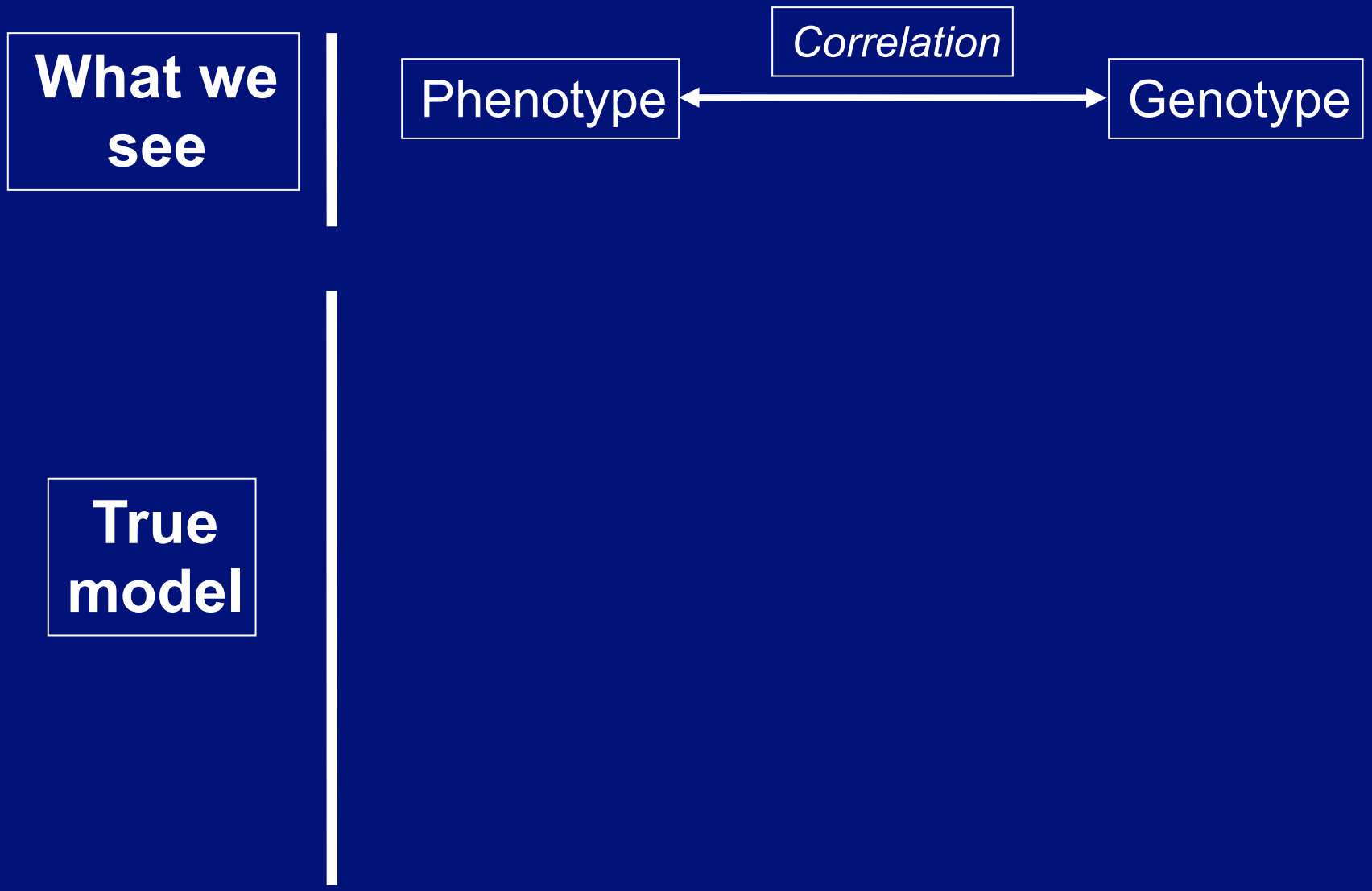
Outline

- Confounding in GWA studies
- Genomic Control
- Structured Association
- EIGENSTRAT
- Mixed Models

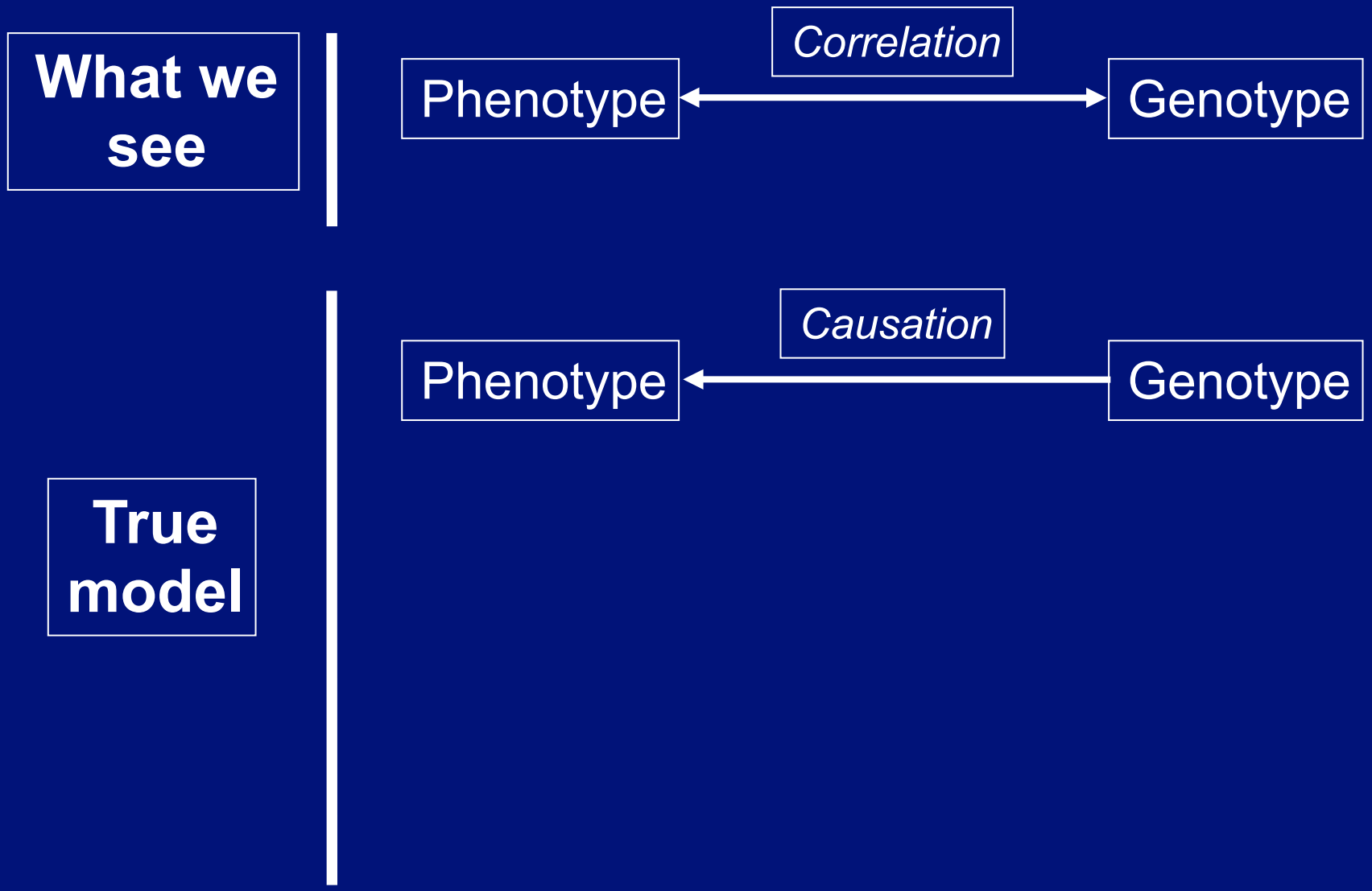
Reasons for genetic association

What we see

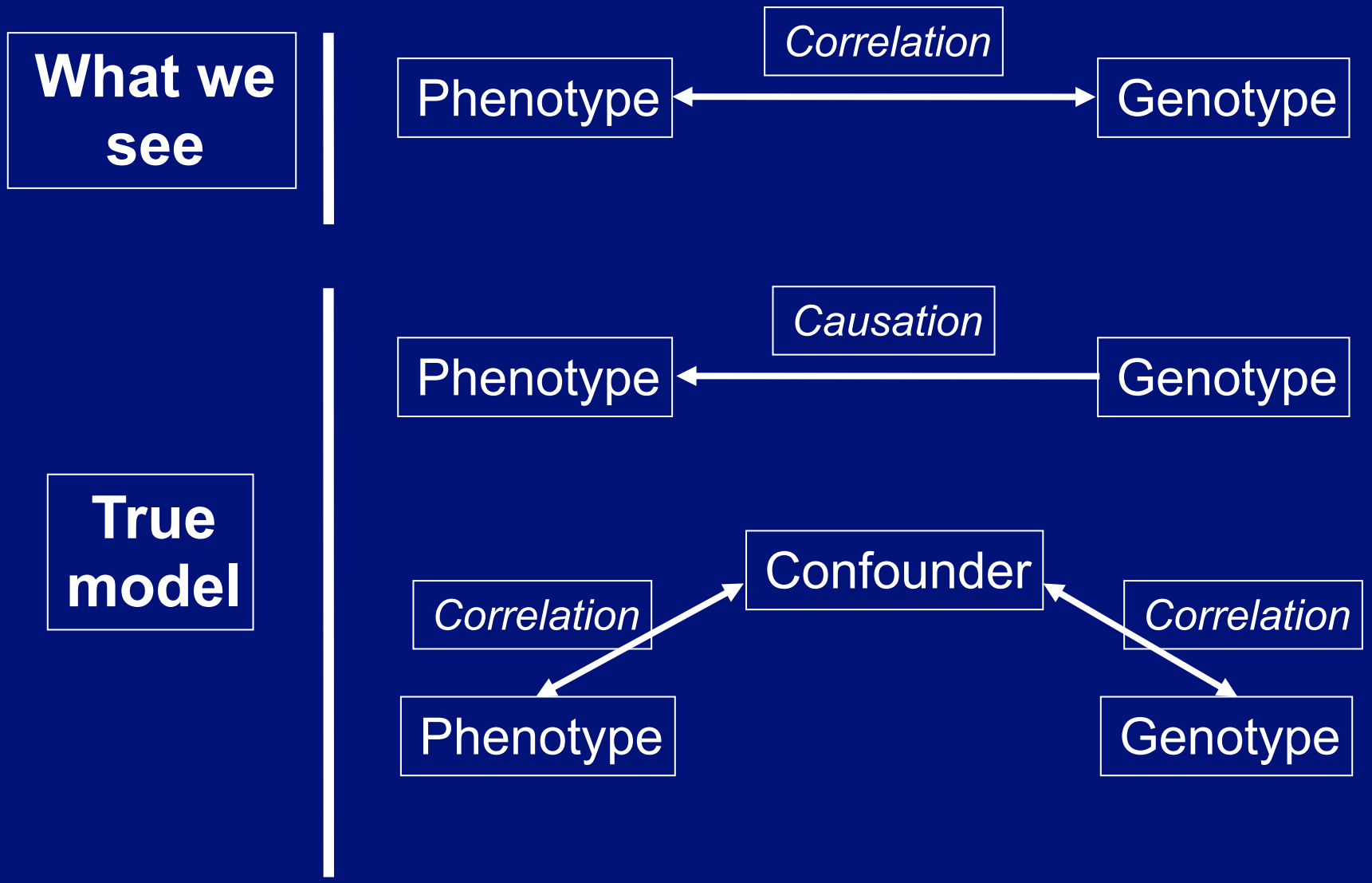
Reasons for genetic association



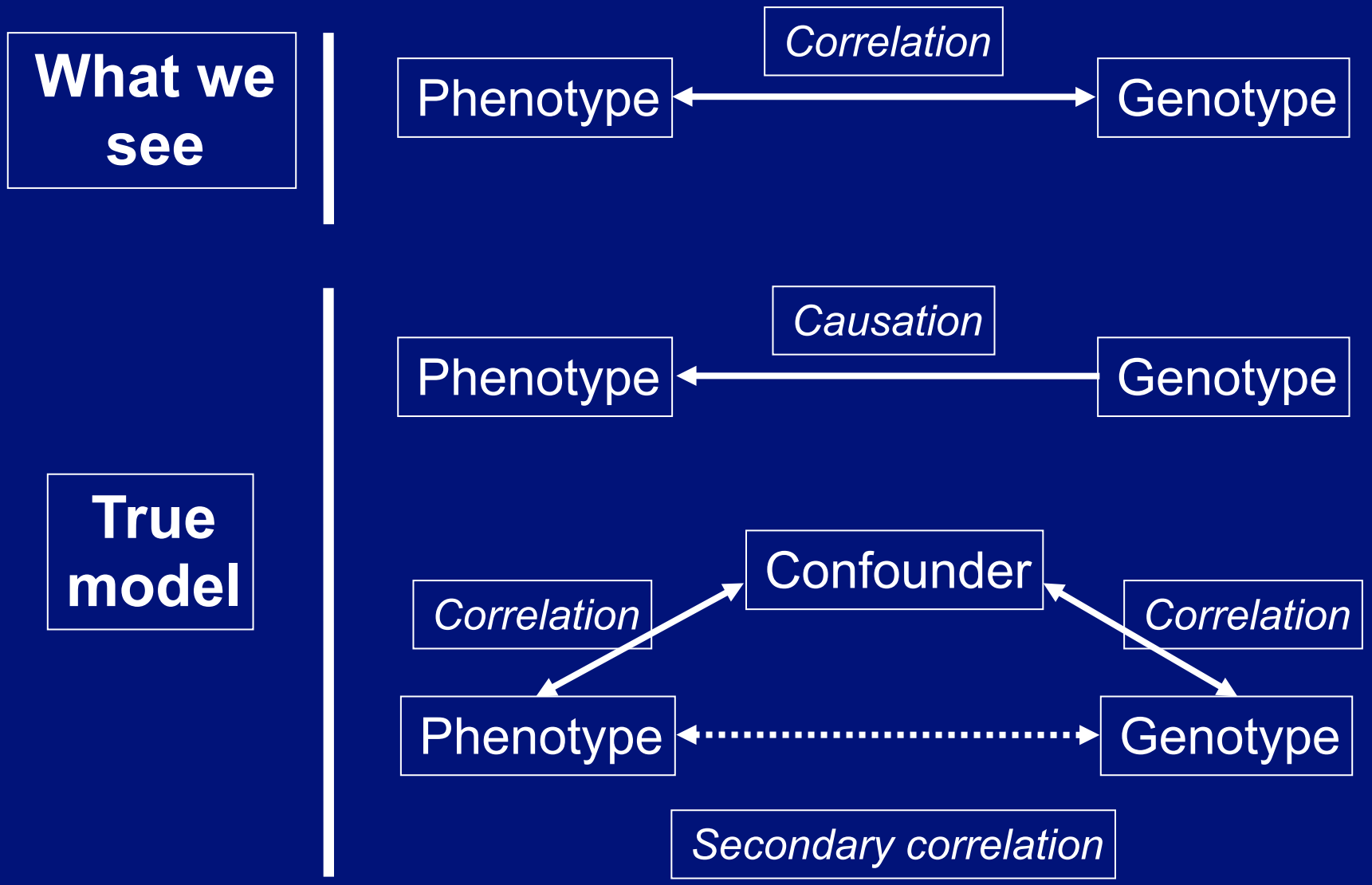
Reasons for genetic association



Reasons for genetic association

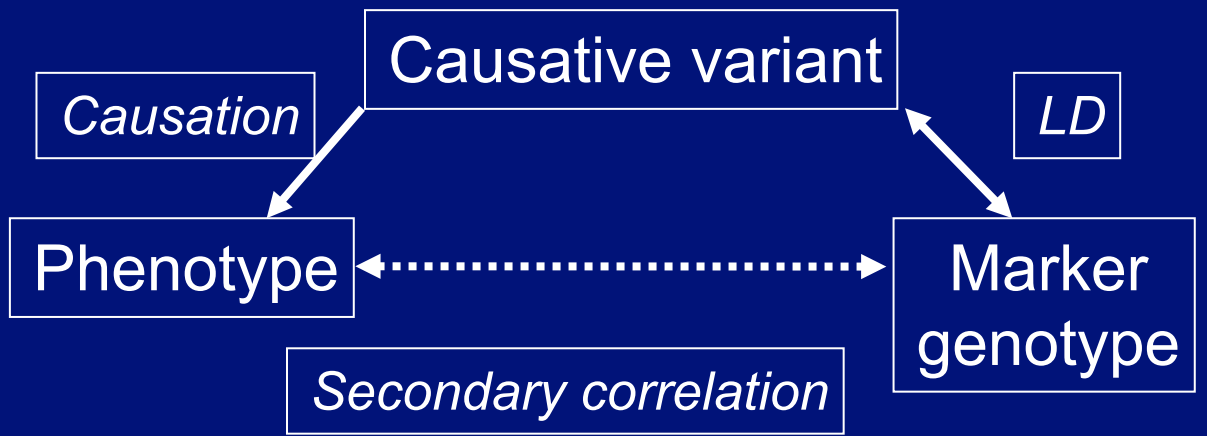


Reasons for genetic association

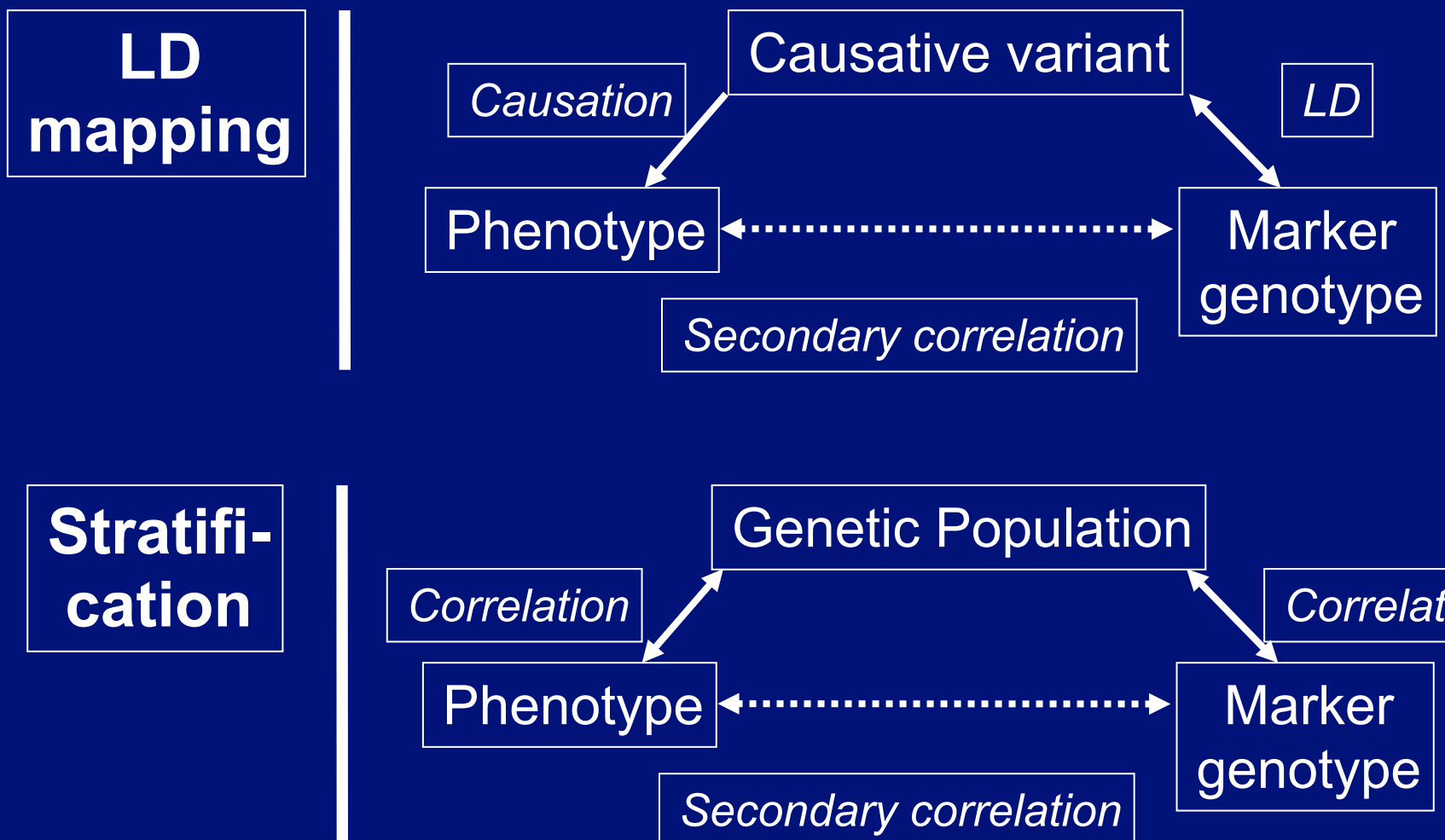


Confounding in genetic studies

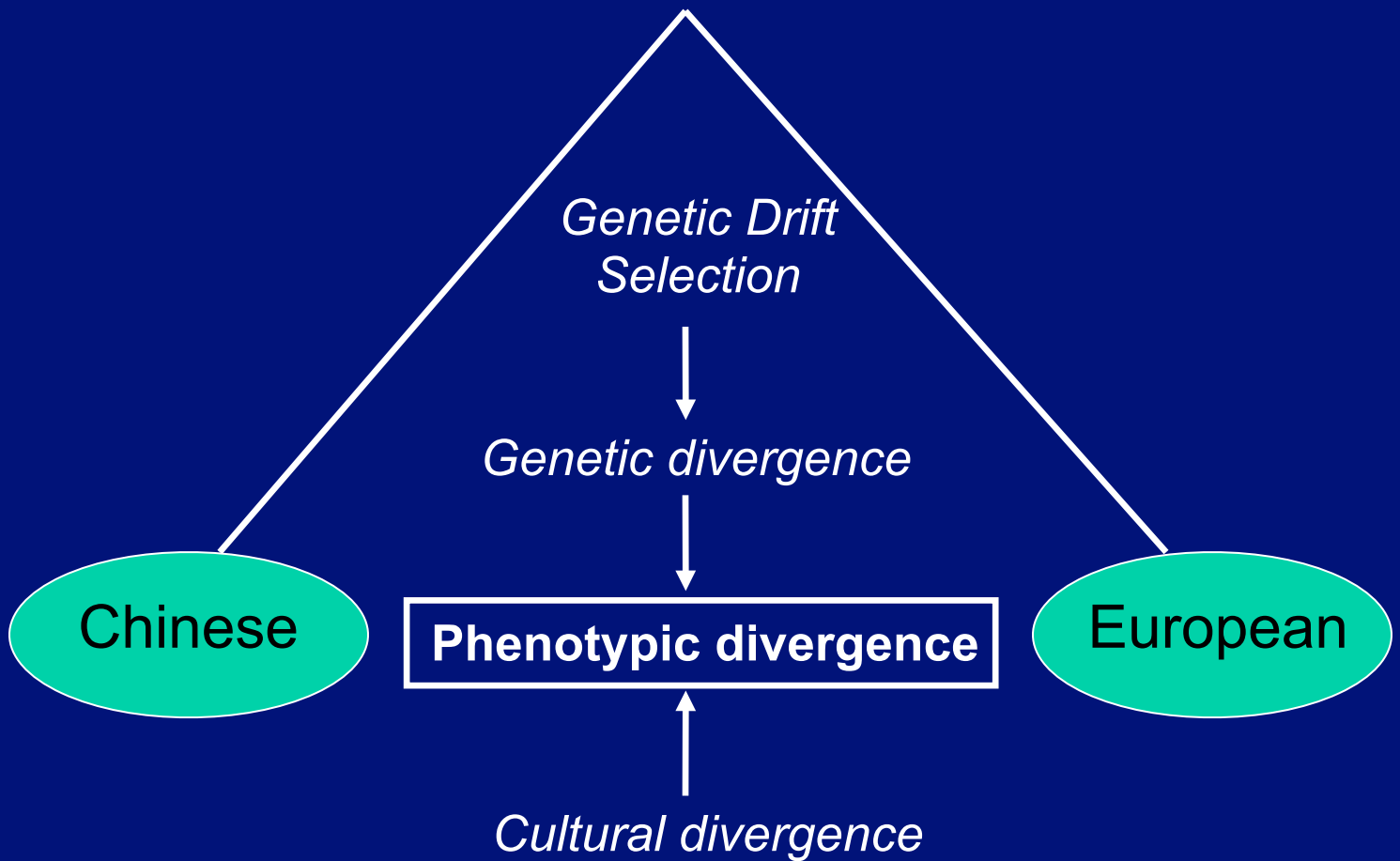
LD mapping



Confounding in genetic studies



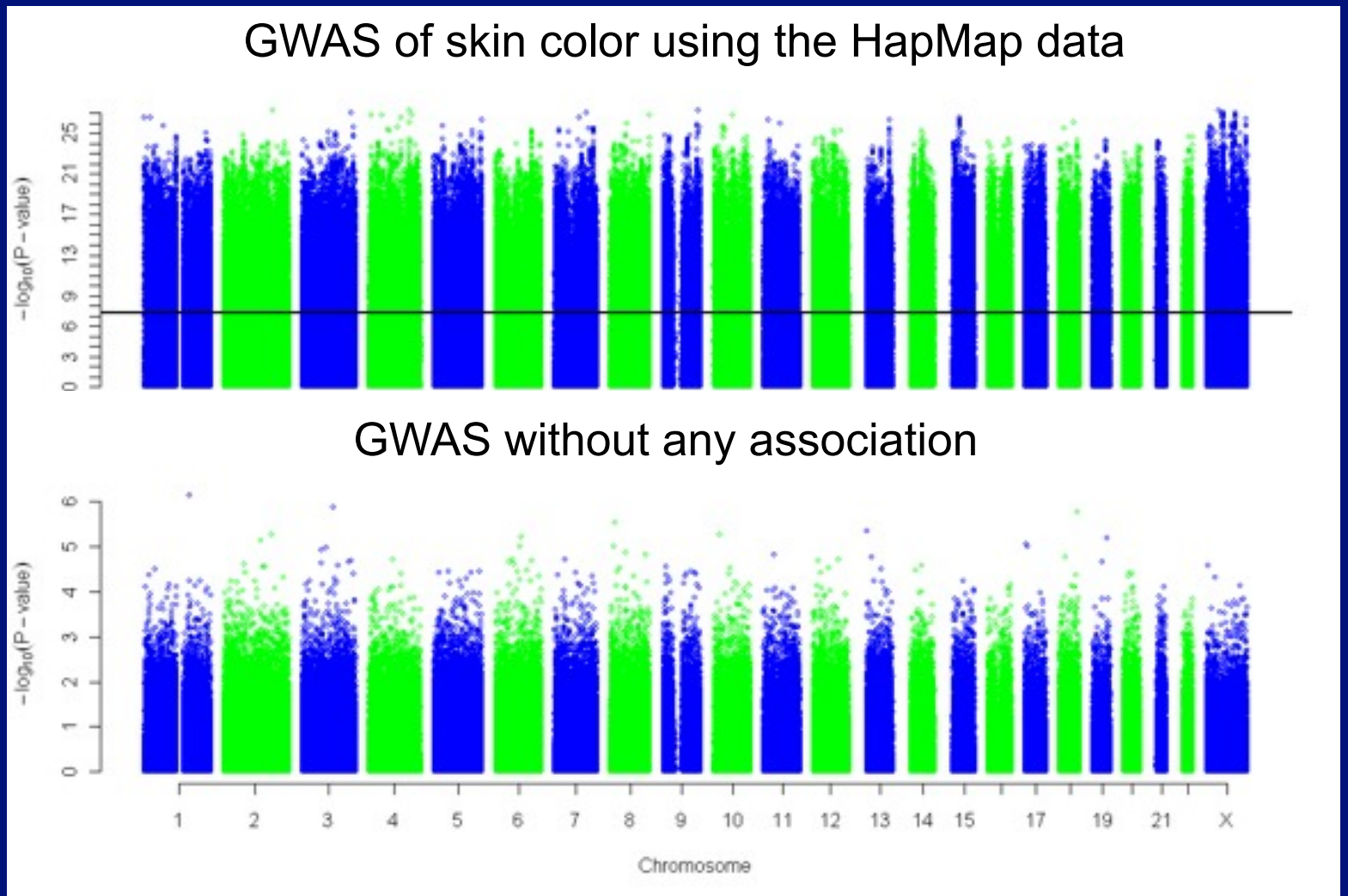
Genetic origin is a major confounder



Confounding in GWAS

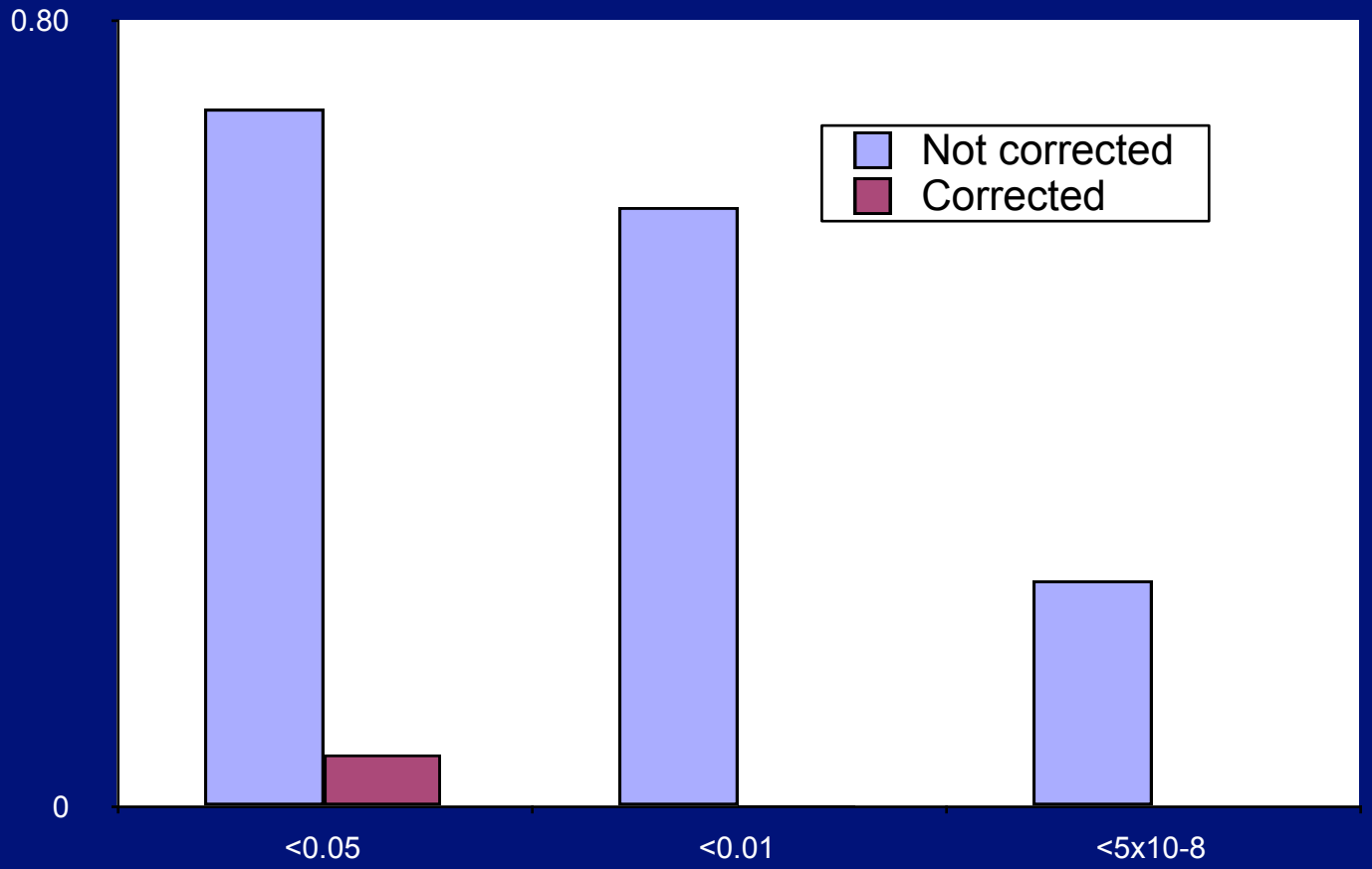
- Some factor is a confounder for genotypes and disease prevalence
  - Dark skin is more prevalent in Africans than in Europeans. The genotypic frequencies are also different between two populations.
  - A study of skin color, which would mix Africans and Europeans is likely to generate multiple false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

Skin color scan

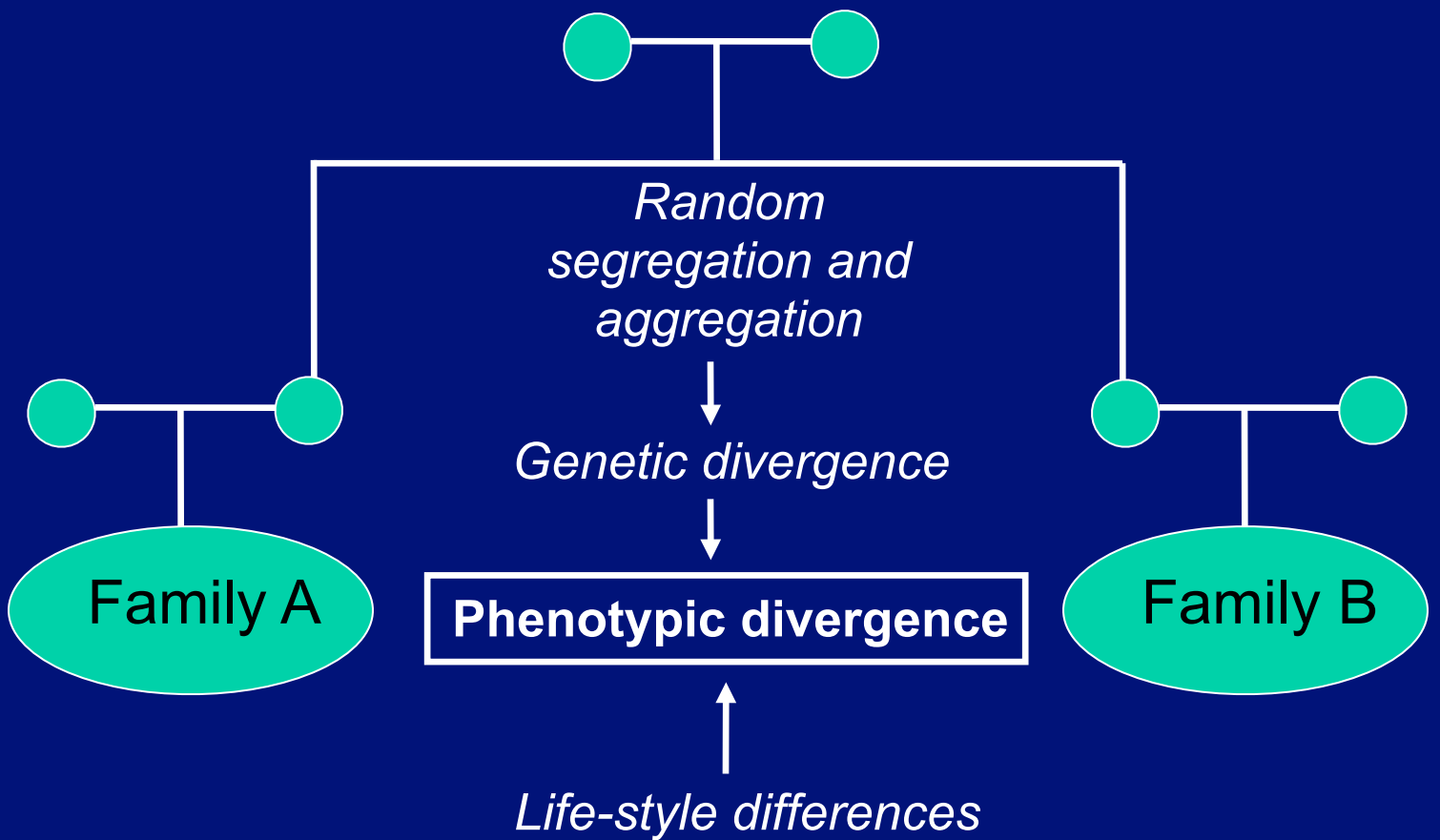


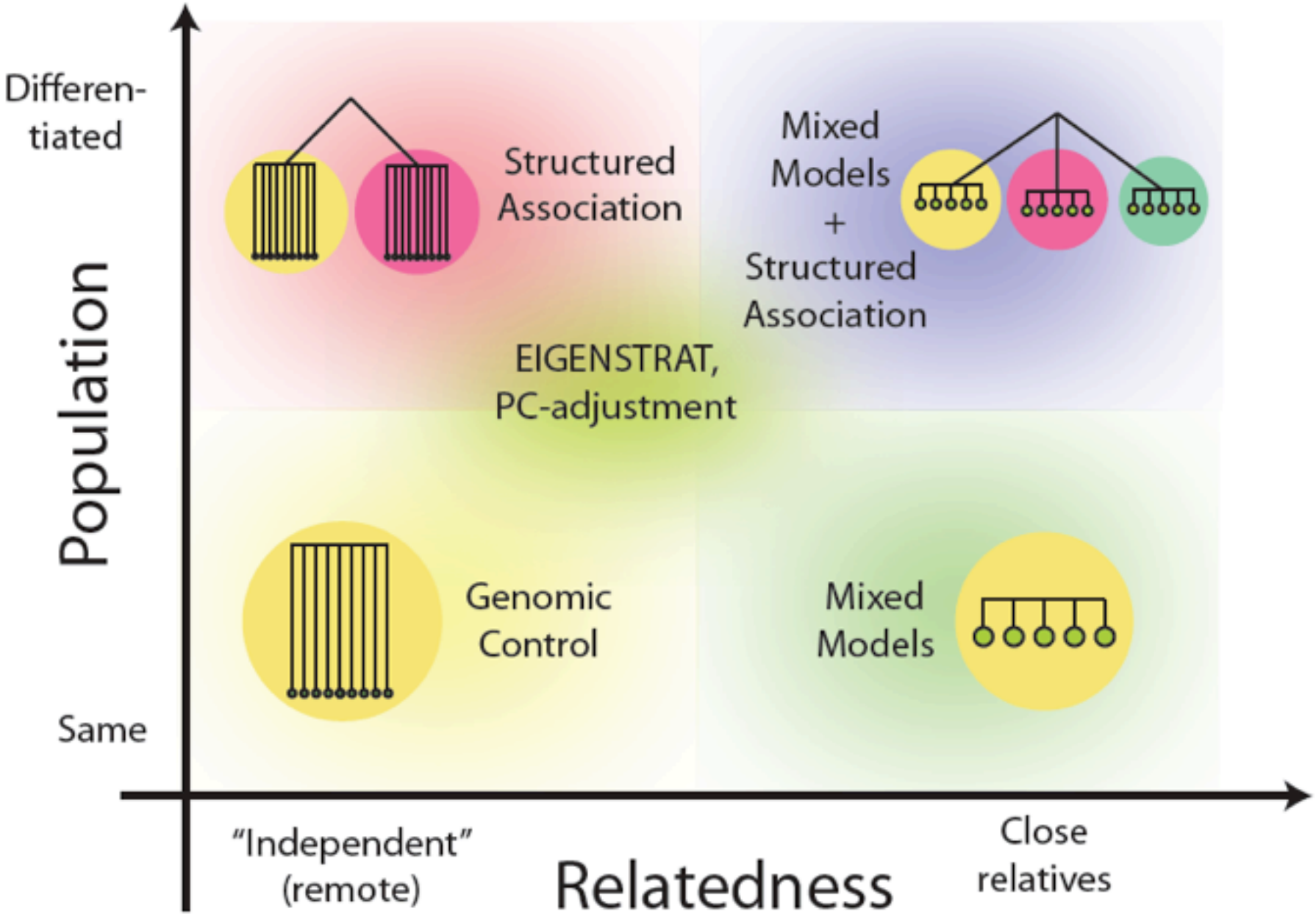
Consequences of stratification

Proportion of P less than some threshold in the skin color GWA



Pedigree is a major confounder




Differen-
tiated
Population
Same
Structured
Association
Mixed
Models
+
Structured
Association
EIGENSTRAT,
PC-adjustment
Genomic
Control
Mixed
Models
"Independent"
(remote)
Relatedness
Close
relatives

Methods to deal with stratification

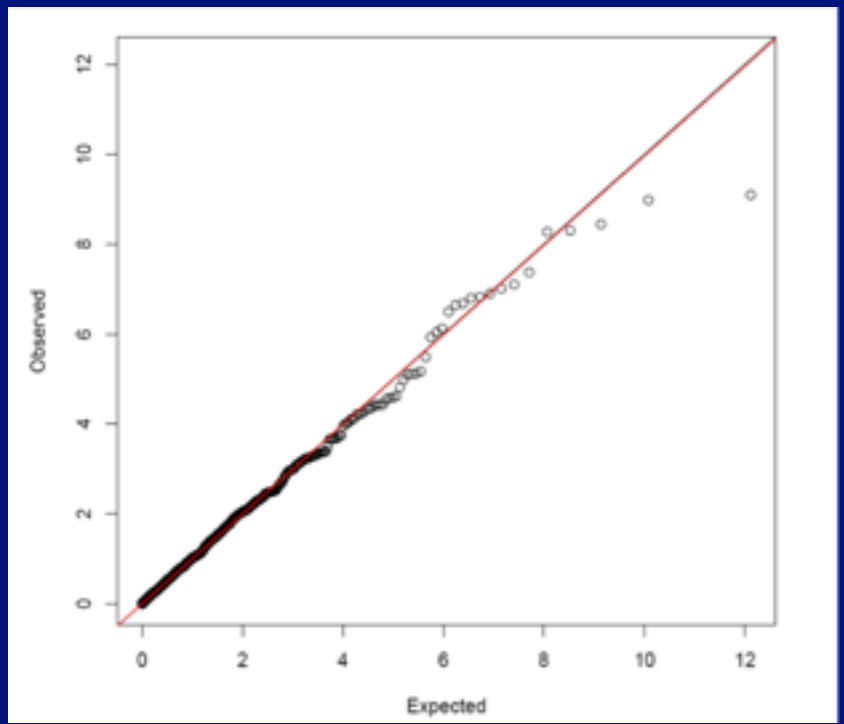
- Confounding: violates the "null" assumption of independence between genotype and phenotype
- Structured association
  - Scope: populations are well-defined, well-separated
- EIGENSTRAT
  - Scope: populations may be less well-defined and separated
- Mixed models
  - Scope: relatives, genetic isolates
- Genomic control
  - Is NOT the method to explicitly correct for dependencies
  - Scope: correcting residual, small degree of stratification

Outline

- Confounding in GWA studies
- **Genomic Control**
- Structured Association
- EIGENSTRAT
- Mixed Models

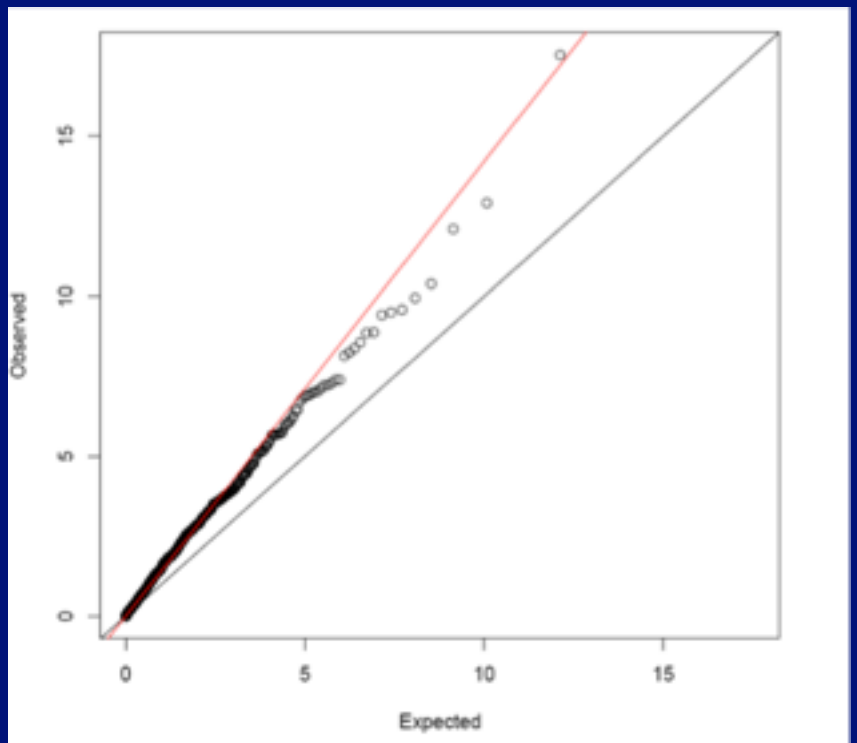
Distribution of the test statistics under the null hypothesis

- 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)
- 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$\{T^2_1, T^2_2, T^2_3, \ldots, T^2_{N-1}, T^2_N\}$ /0.455
  - 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 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

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

Few notes on GC

When inflation is large (say, $\lambda > 1.05$) other, more powerful methods are to be used

GC assumes that stratification acts in the same manner across all loci, which is not always true

In present form, **works only for additive model**

Inflation factor λ depends on samples size. Thus

(1) Report of standardized values (say, per 1,000 cases and 1,000 controls) is recommended

(2) Special methods should be used when number of people typed for different SNPs is different

Outline

- Confounding in GWA studies
- Genomic Control
- **Structured Association**
- EIGENSTRAT
- Mixed Models

Structured association (SA)

- Identify genetic populations (strata)
- Do stratified analysis; e.g. Cochran-Mantel-Haenszel test; or meta-analysis of results obtained in different strata
- Apply GC to correct for residual inflation ($1 < \lambda < 1.05$)
- Potential problems: strata not always known *a priori* or easily identified, they also may be not well-defined

Adjust for strata?

- Inclusion of strata in your linear model
  - Y ~ mu + sex + age+ strata + snp
- accounts for the difference in means

- This is NOT EXACTLY what is meant by stratified analysis, which also allows for different effects of nuisance covariates in different strata. You can do that by model
  - Y ~ mu + strata*(sex + age) + snp
- Still, even this is not exactly the same, as stratified analysis allows for different residual variances across strata
- You can do that with Linear Mixed Models (LMM) or Generalized Estimating Equations (GEE)

Outline

- Confounding in GWA studies
- Genomic Control and Structured Association
- **EIGENSTRAT**
- Mixed Models

Estimation of genetic similarity

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

Basically, this matrix tells how similar are genomes of people involved

Idea of Multidimensional Scaling

- Study of *N* subjects
- *NxN* matrix of pair-wise distances (0 = the same subject, I = 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

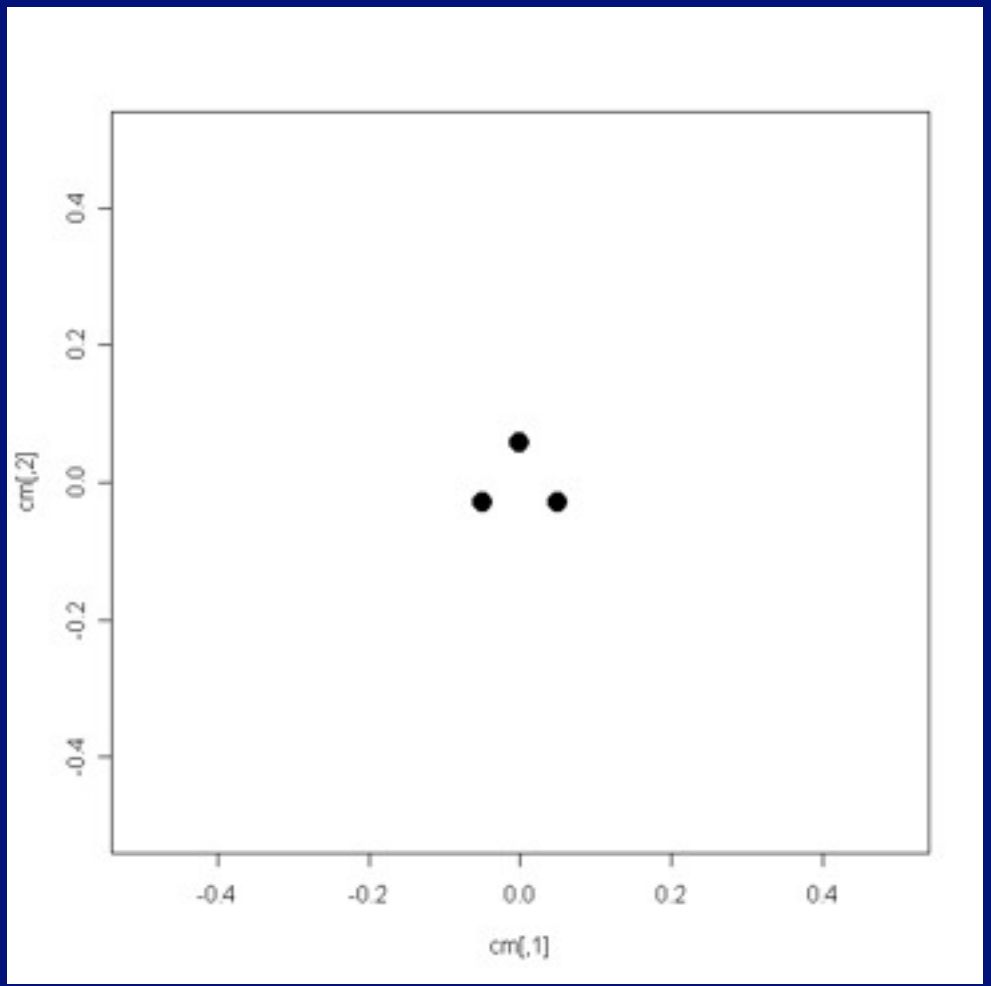
Example CMDS

- 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



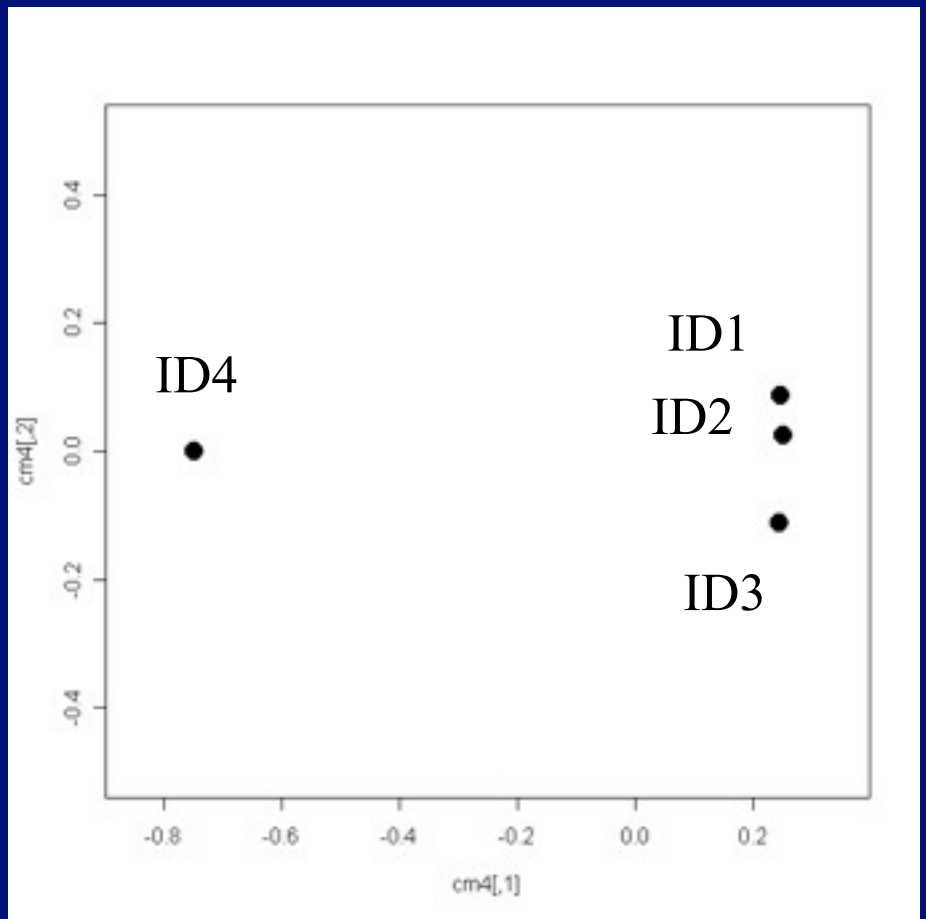
Example CMDS

- Distance matrix

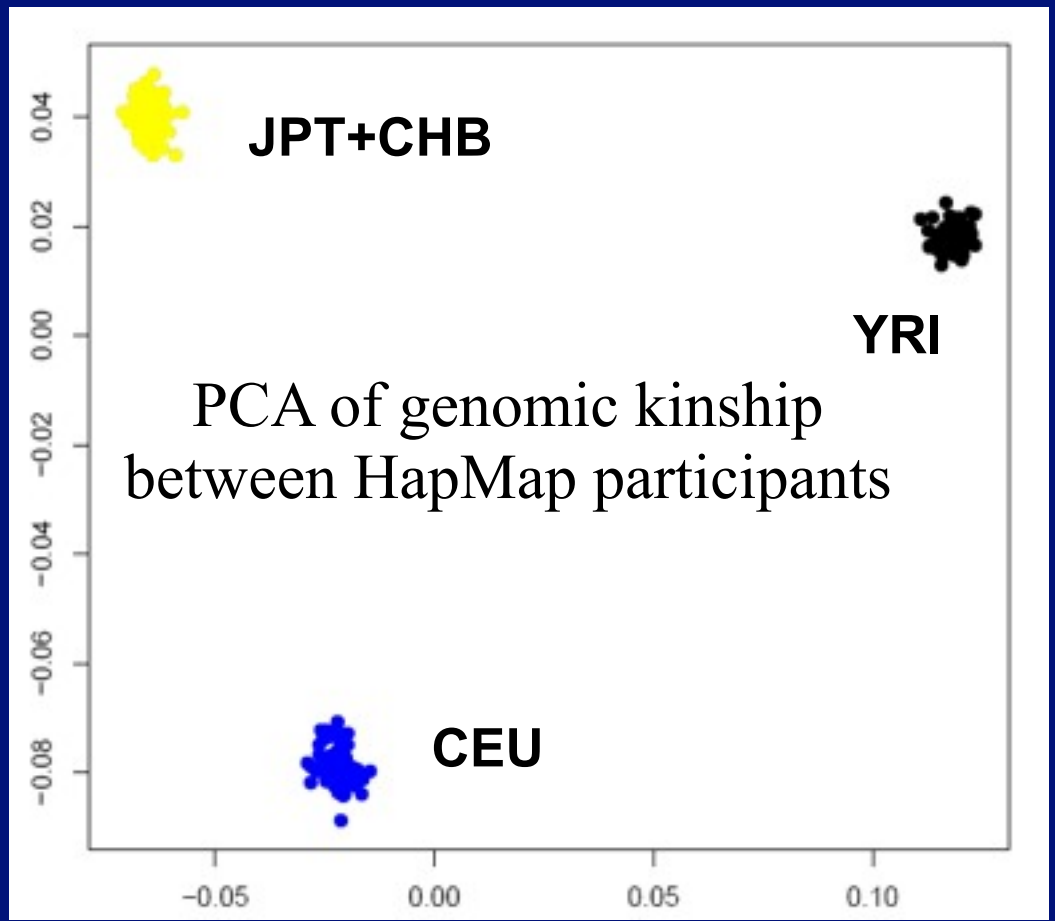
	ID1	ID2	ID3	ID4
ID1	0	0.1	15	1.00
ID2	0.1	0	0.20	1.00
ID3	0.15	0.20	0	1.00
ID4	1.00	1.00	1.00	0

- 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

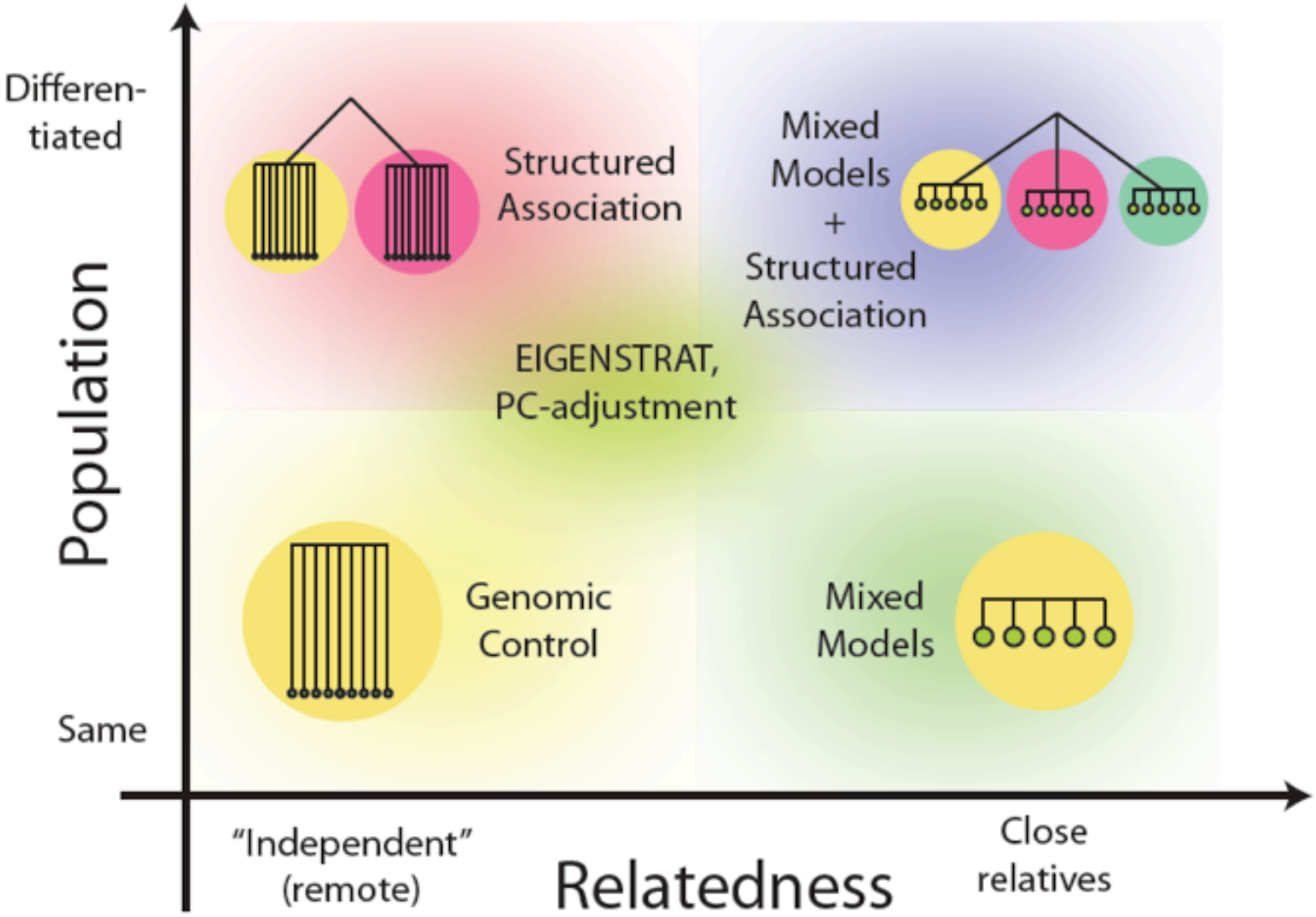


PCA of genomic kinship



Idea of EIGENSTRAT method

- Estimate genetic relations between the study participants using genomic data, compute pair-wise distance matrix
- Extract 3 to 10 principal components (PC) of variation from this matrix
- In analysis of association, adjust both phenotypes and genotypes for these PCs (modification: include principal axes of variation as covariates in regression model)
- Apply GC to correct for residual inflation ($1 < \lambda < 1.05$)
- Problems with ES: accounts for mean, but not variance differences; does not work in case of strong relations (families, isolates)


Differen-
tiated
Population
Same
Structured
Association
Mixed
Models
+
Structured
Association
EIGENSTRAT,
PC-adjustment
Genomic
Control
Mixed
Models
"Independent"
(remote)
Relatedness
Close
relatives

Summary: software & functions

- Genomic control: for additive models, implemented in any GWAS software, or do it yourself. For other models: we work on that … may be released late this year
- Stratified analysis: use any GWA software and then meta-analysis programs (METAL, MANTEL, metaMapper, GWAMA, MetABEL), or write custom scripts
- Genomic kinship matrix (base for EIGENSTRAT, PC-adjustment): PLINK's 'IBD', GenABEL's ibs() function
- EIGENSTRAT analysis: EIGENSTRAT, GenABEL's egscore() function
- Adjustment for PCs: any GWA software supporting covariates
- **Mixed-model based analysis:** GenABEL's mmscore & grammar, Merlin (but with pedigree…); grammar+ and FMM are going to be released later this year (MixABEL)