

# CONFOUNDING IN GWAS

YURII AULCHENKO

YURII [DOT] AULCHENKO [AT] GMAIL [DOT] COM

# OUTLINE

---

Confounding in GWA studies

Genomic Control

Structured Association

Mixed Models

EigenSTRAT

# REASONS FOR GENETIC ASSOCIATION

---

# REASONS FOR GENETIC ASSOCIATION

---

**What we see**



# REASONS FOR GENETIC ASSOCIATION

---

**What we see**



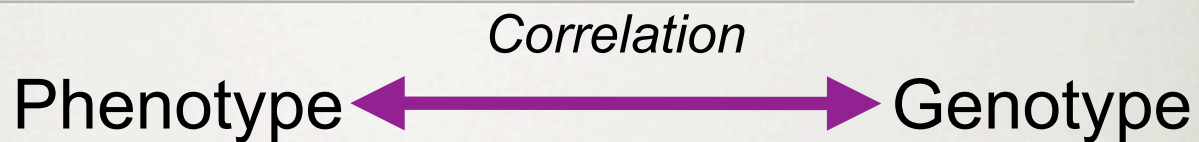
**True model**



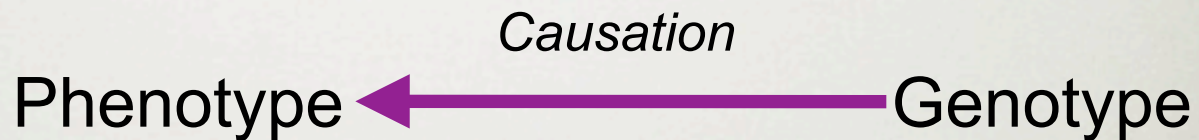
# REASONS FOR GENETIC ASSOCIATION

---

**What we see**

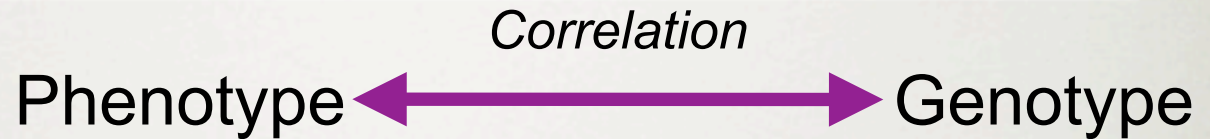


**True model**

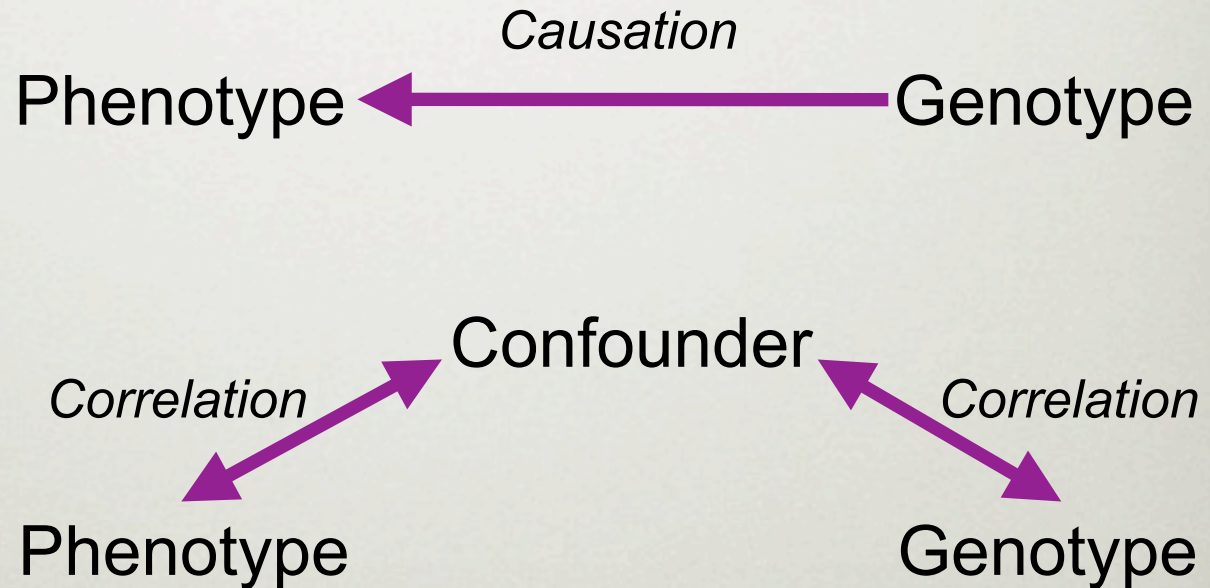


# REASONS FOR GENETIC ASSOCIATION

**What we see**

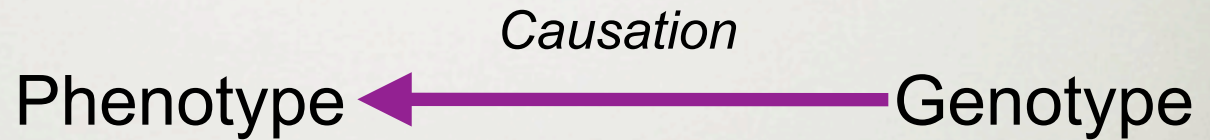
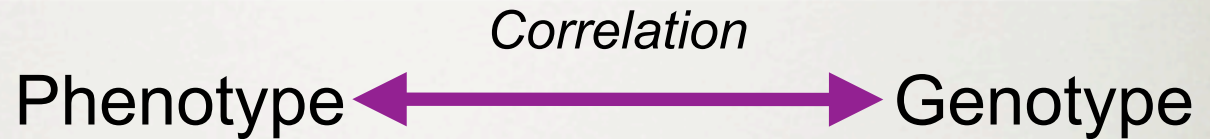


**True model**

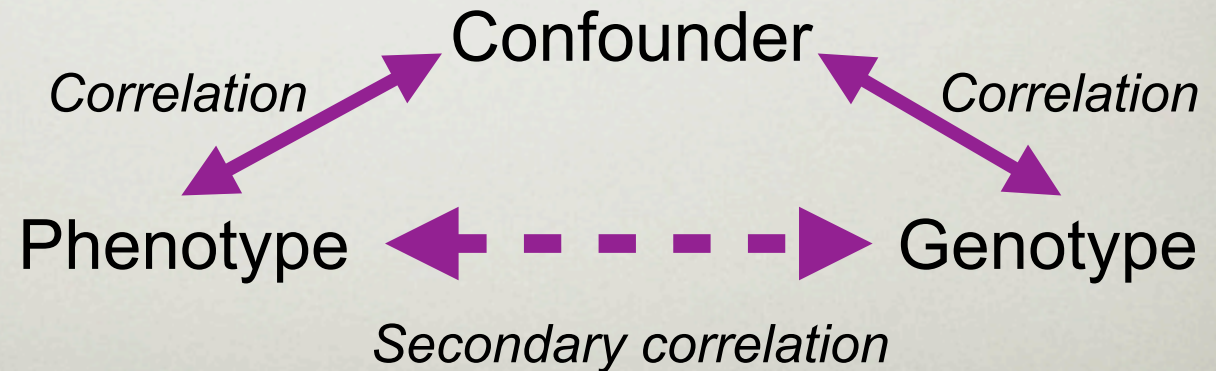


# REASONS FOR GENETIC ASSOCIATION

**What we see**



**True model**






# CONFOUNDING IN GENETIC STUDIES

---

# CONFOUNDING IN GENETIC STUDIES

---

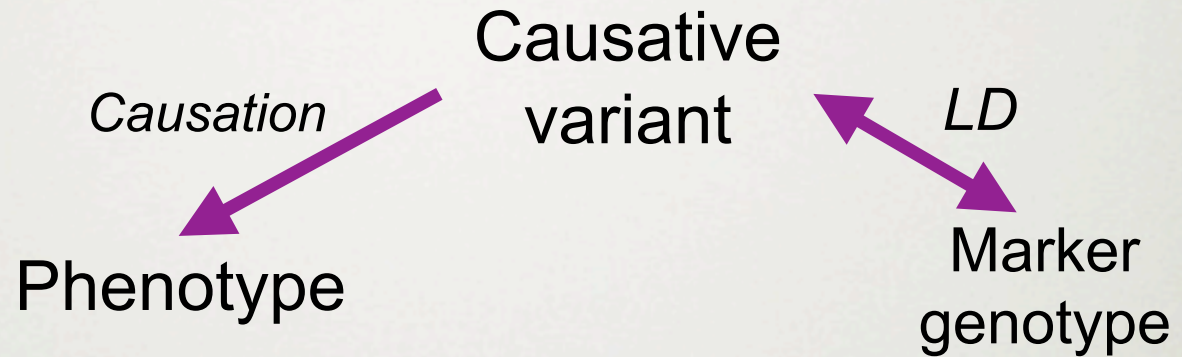
**LD  
mapping**

A vertical pink line is positioned to the right of the text 'LD mapping'.

# CONFOUNDING IN GENETIC STUDIES

---

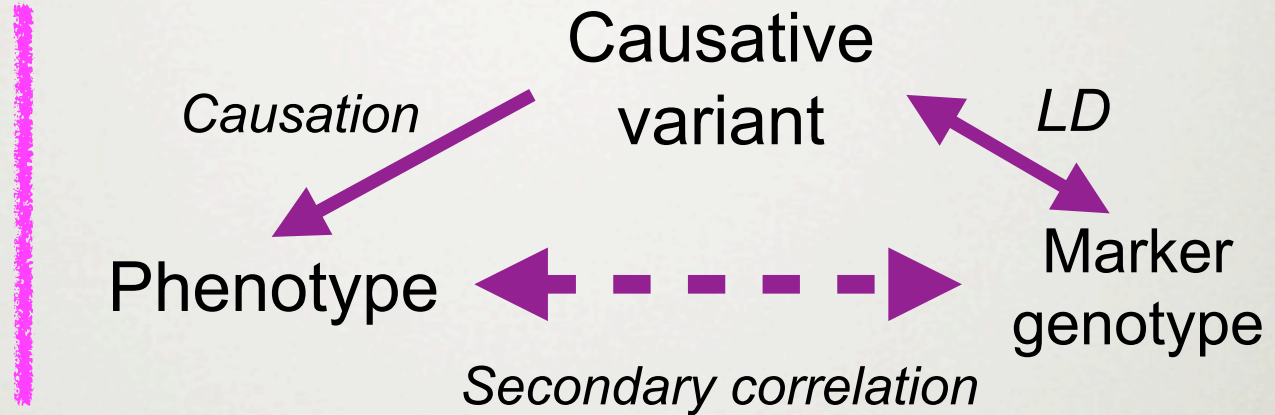
**LD  
mapping**



# CONFOUNDING IN GENETIC STUDIES

---

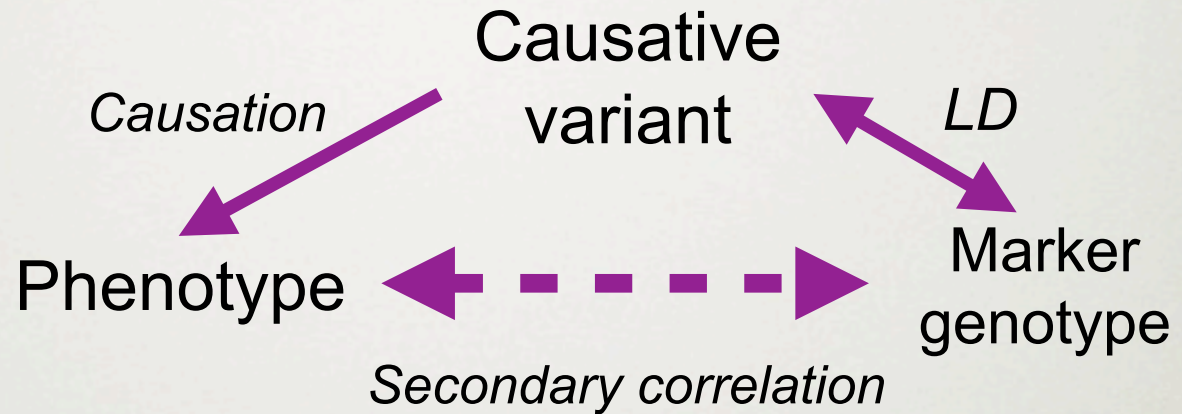
**LD  
mapping**



# CONFOUNDING IN GENETIC STUDIES

---

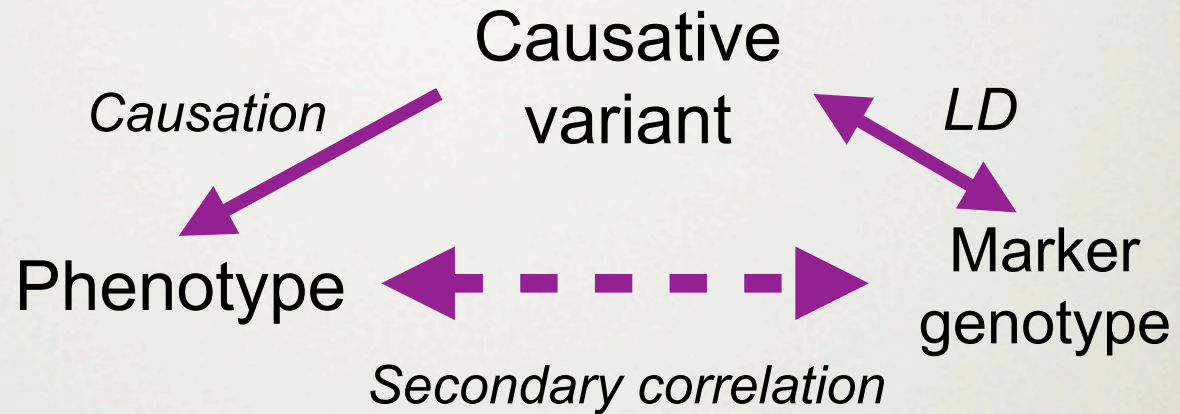
**LD mapping**



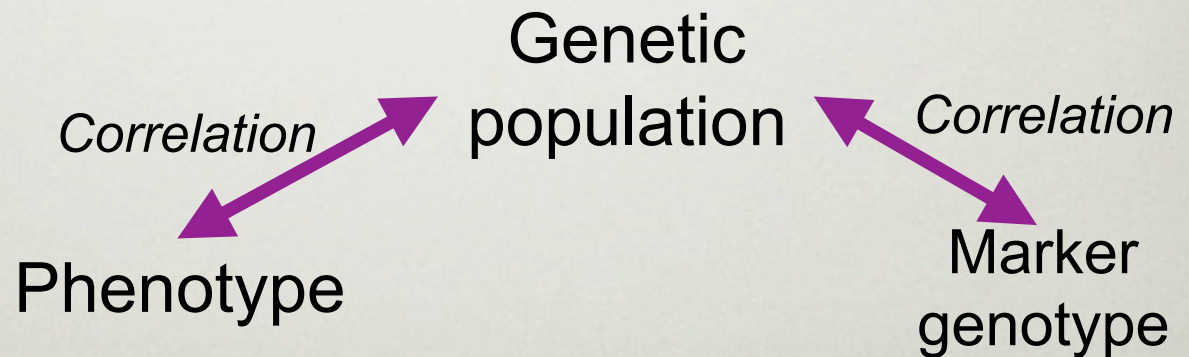
**Stratification**

# CONFOUNDING IN GENETIC STUDIES

**LD mapping**

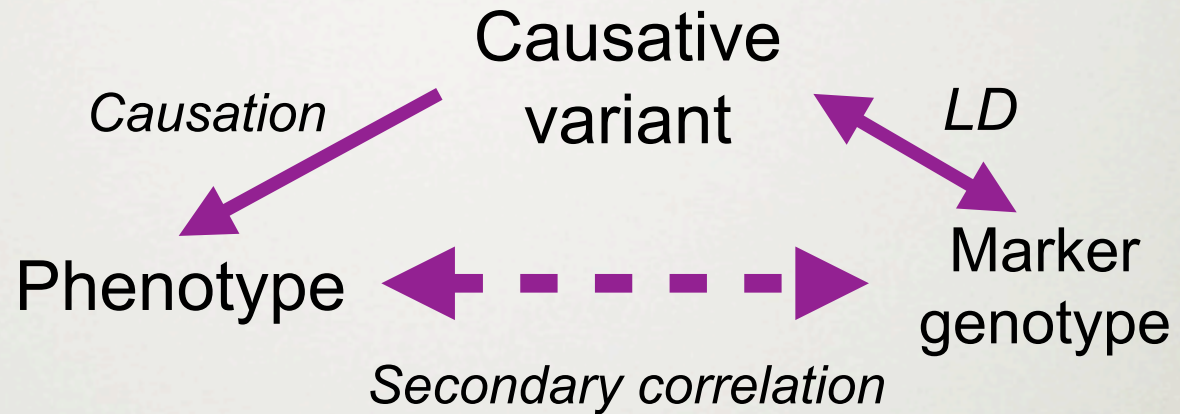


**Stratification**

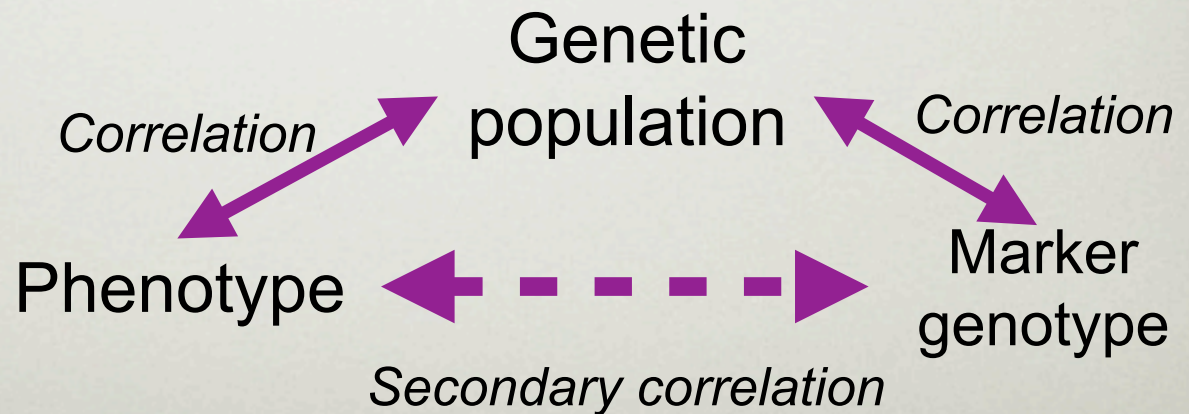


# CONFOUNDING IN GENETIC STUDIES

**LD mapping**

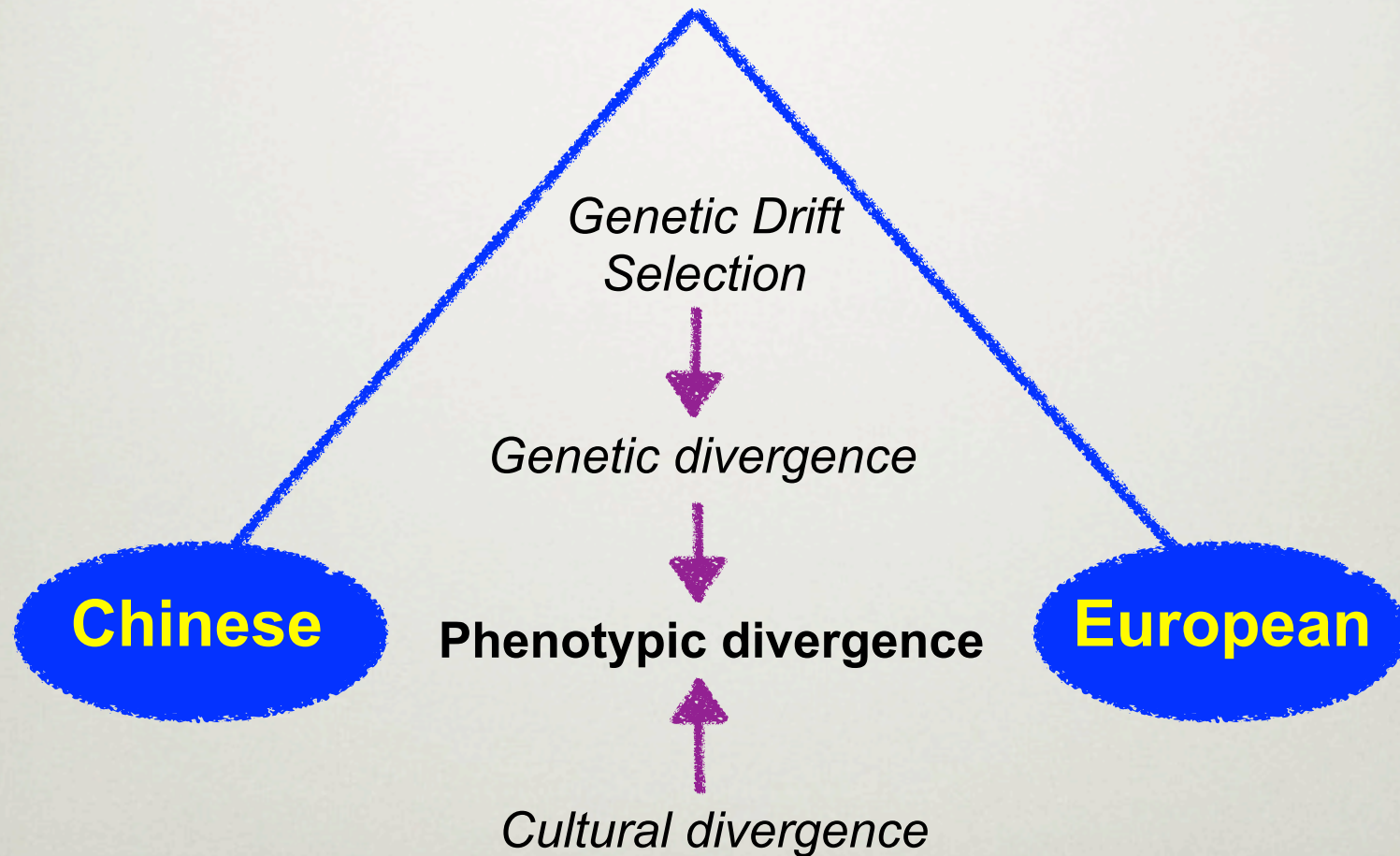


**Stratification**



# GENETIC ORIGIN IS A MAJOR CONFOUNDER

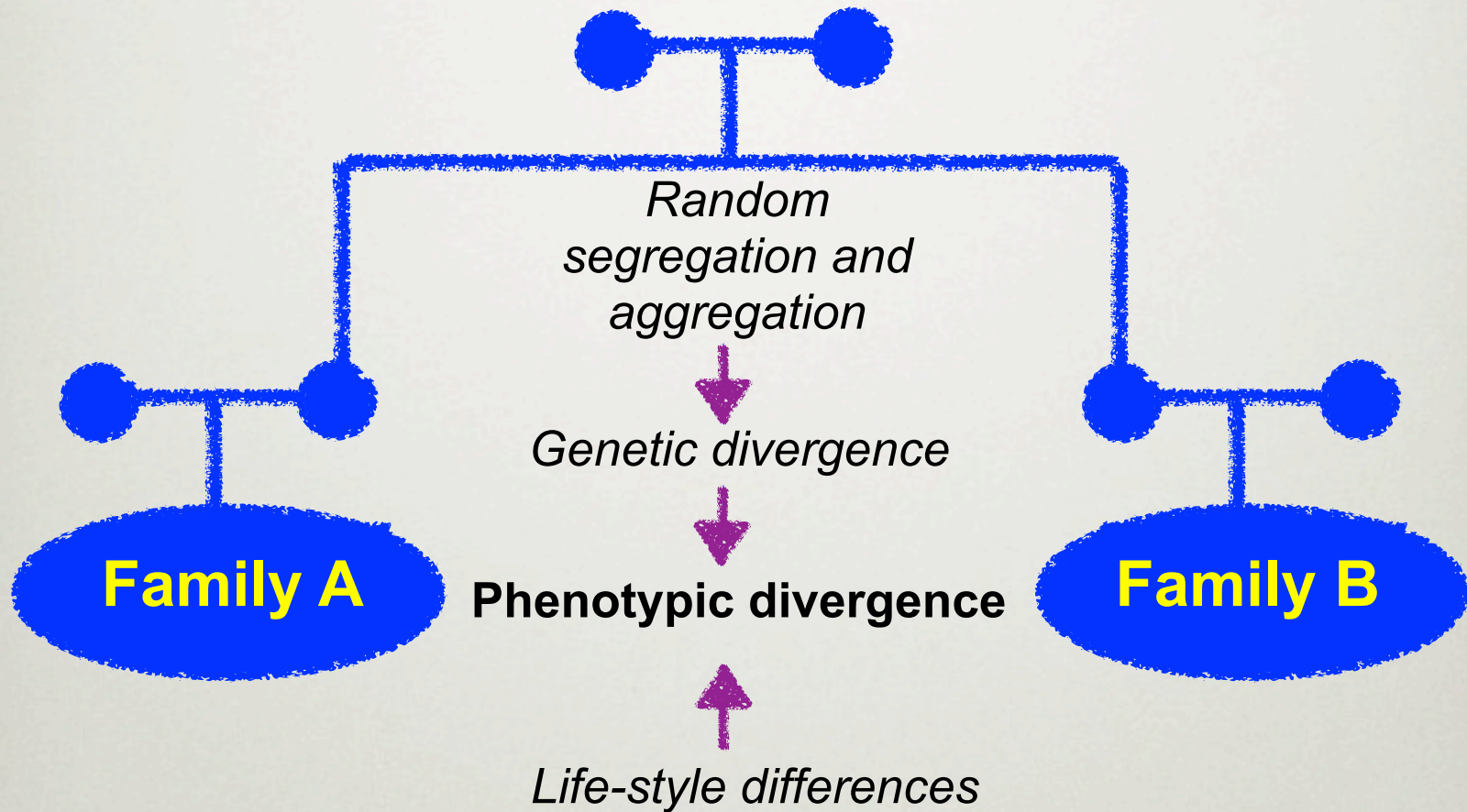
---





# PEDIGREE IS A MAJOR CONFOUNDER

---



# CONFOUNDING IN GWAS

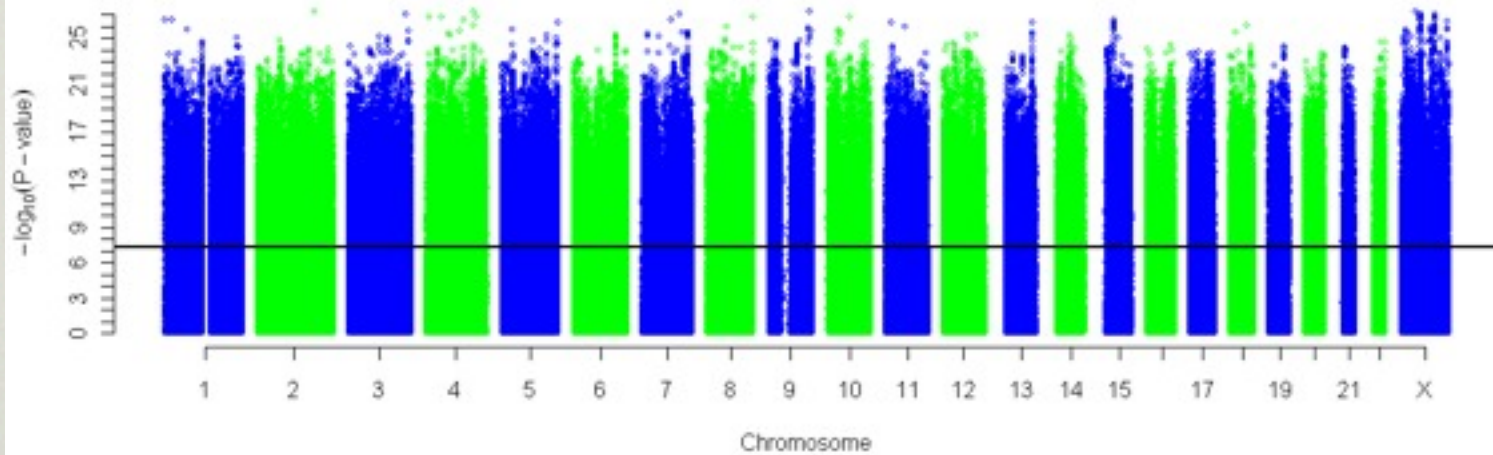
---

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

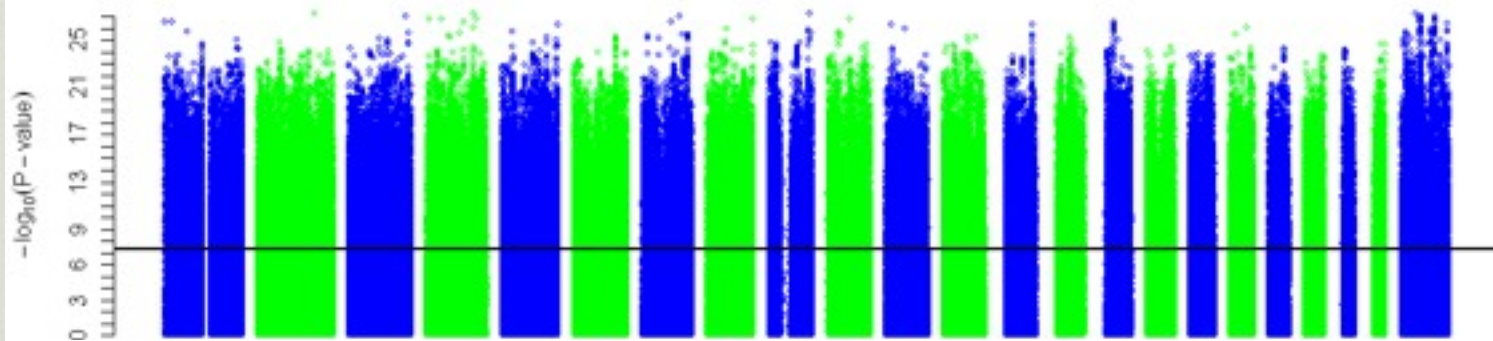
# SKIN COLOR SCAN

GWAS of skin color using the HapMap data

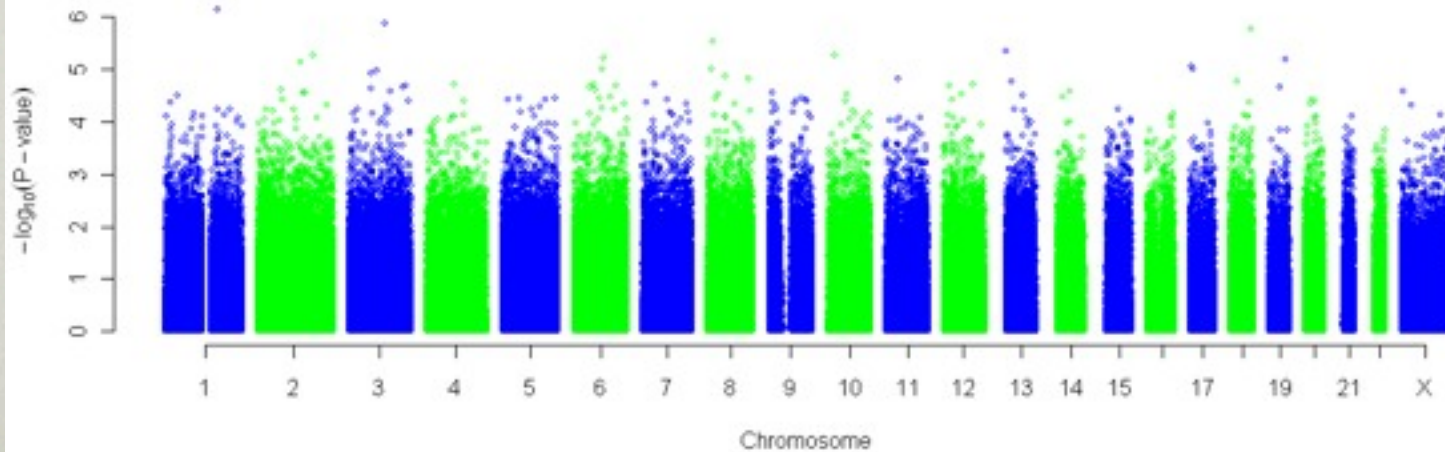


# SKIN COLOR SCAN

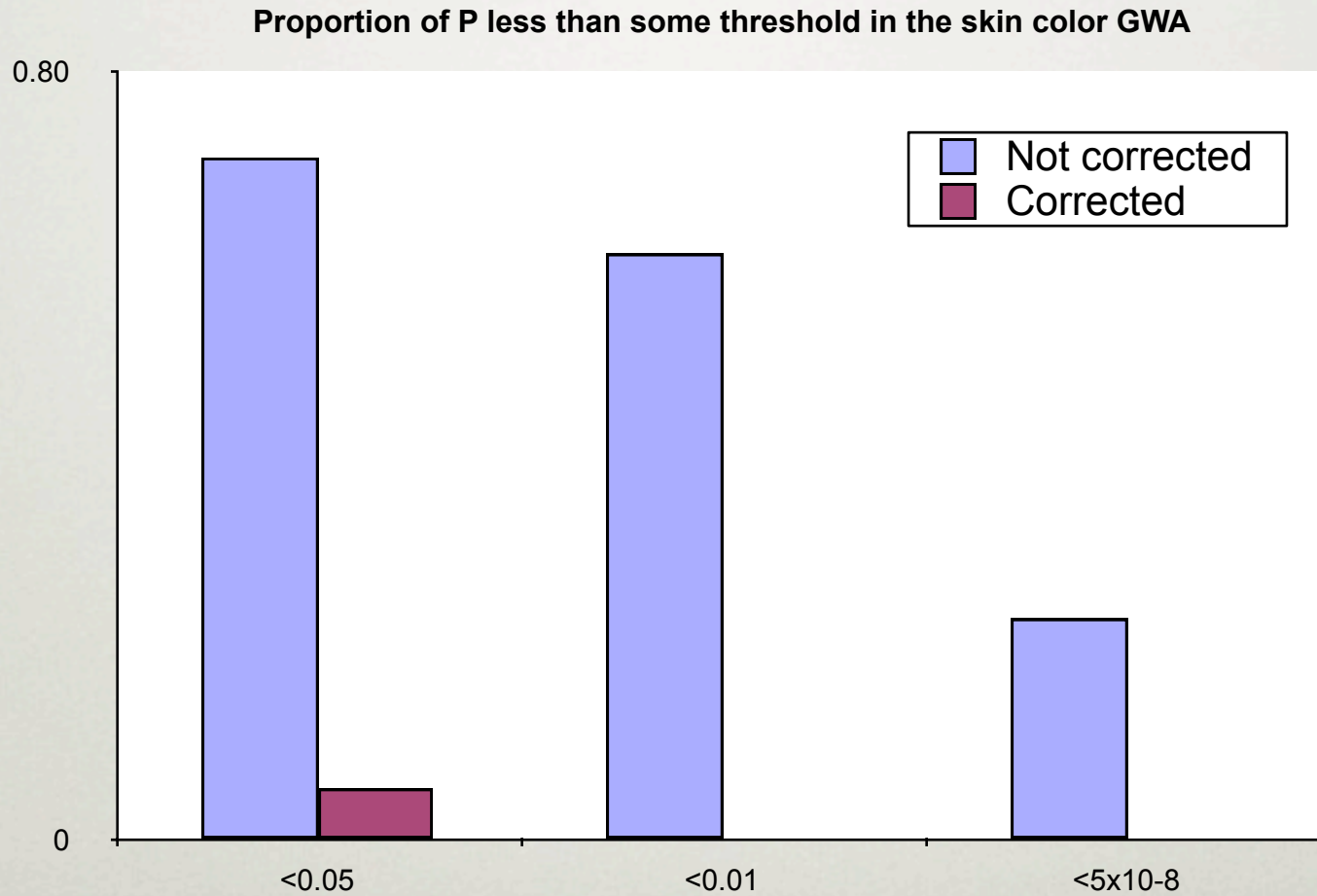
GWAS of skin color using the HapMap data

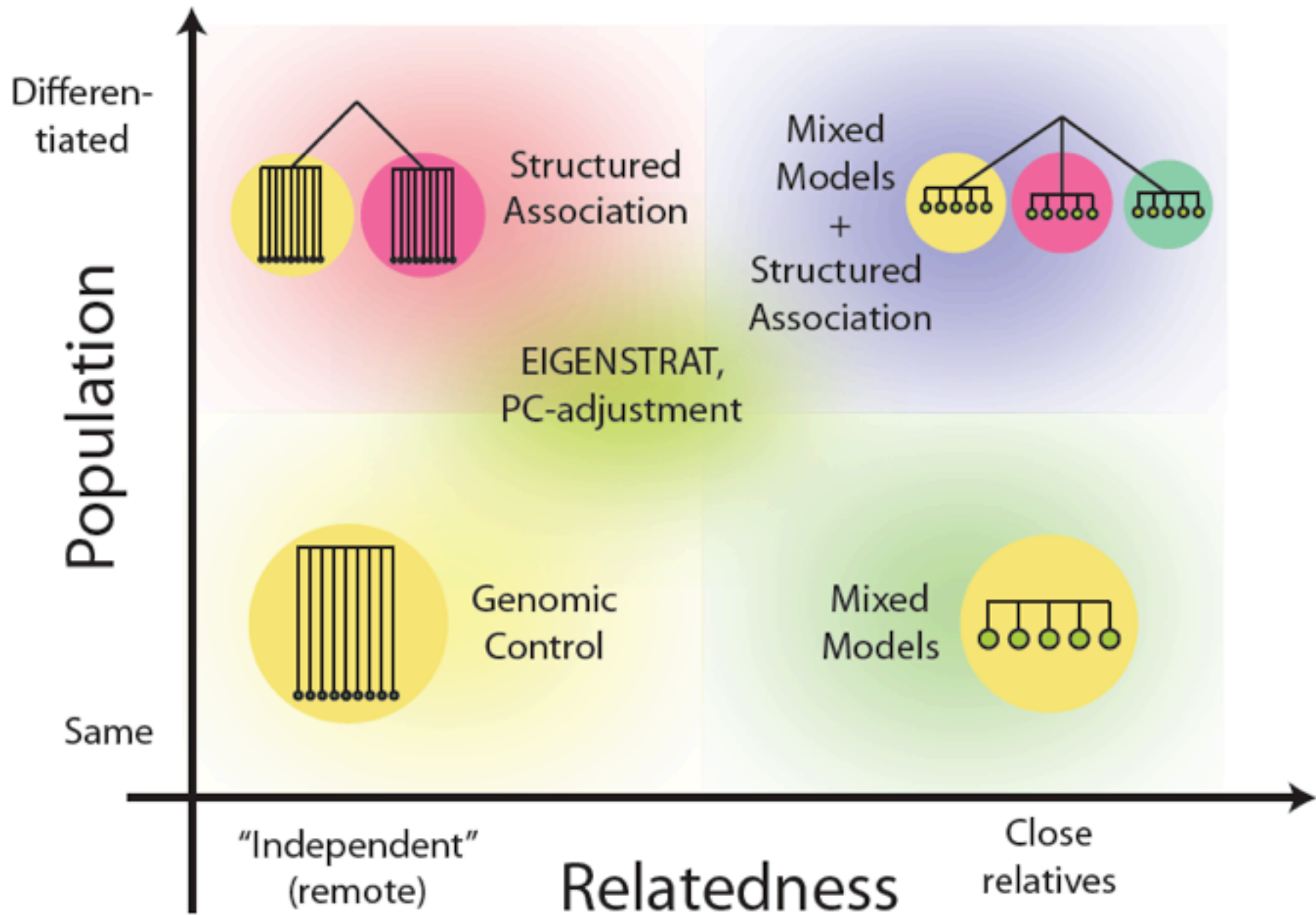


GWAS without any association



# CONSEQUENCES OF STRATIFICATION





# METHODS TO DEAL WITH STRATIFICATION

---

- **Structured association:** populations are well-defined, well-separated
- **EIGENSTRAT:** populations may be less well-defined and separated
- **Mixed models:** very complex structure, relatives, genetic isolates
- **Genomic control** (does not explicitly correct for dependencies): correcting residual, small degree of stratification

# OUTLINE

---

Confounding in GWA studies

**Genomic Control**

Structured Association

Mixed Models

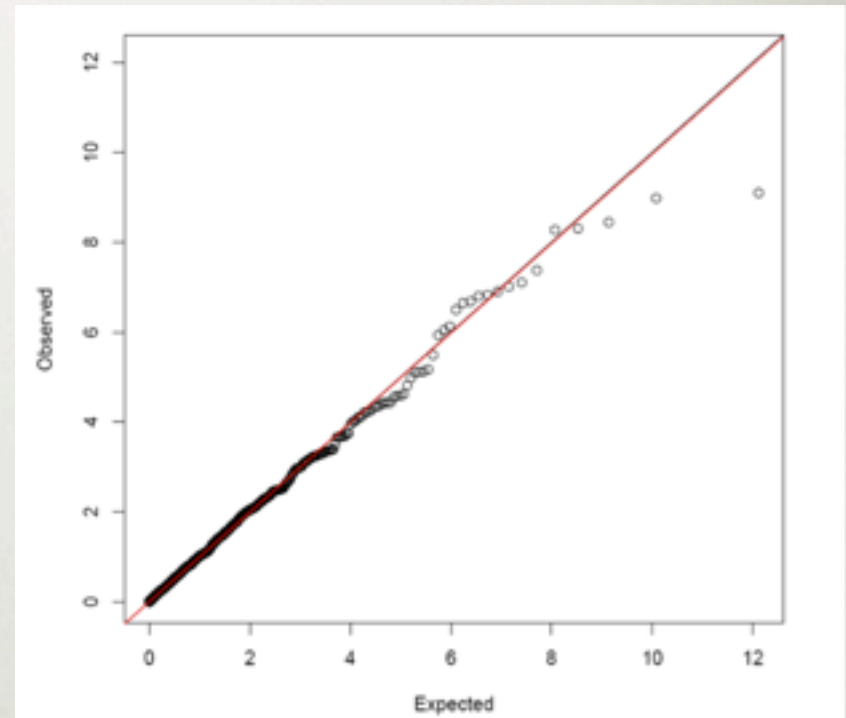
EigenSTRAT



# DISTRIBUTION OF THE TEST

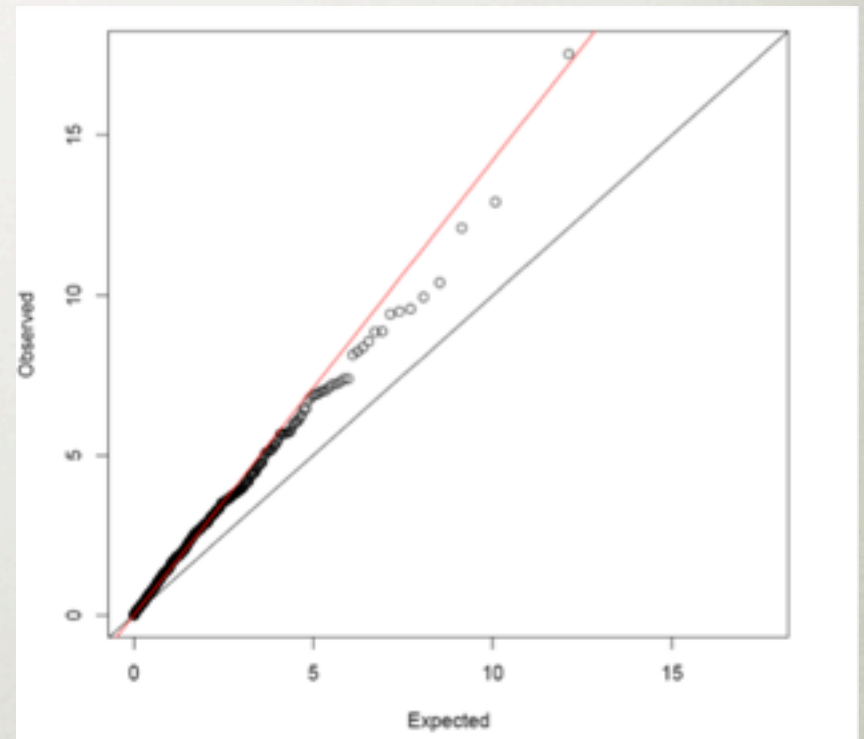
---

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



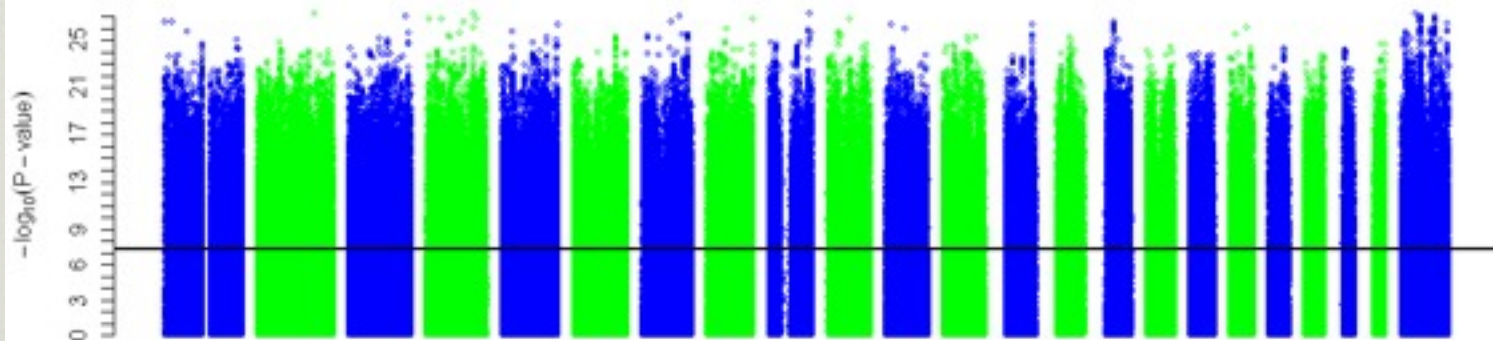
# IDEA OF THE GENOMIC CONTROL

- 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  $\lambda \cdot \chi^2_1$  ( $\lambda \geq 1$ )

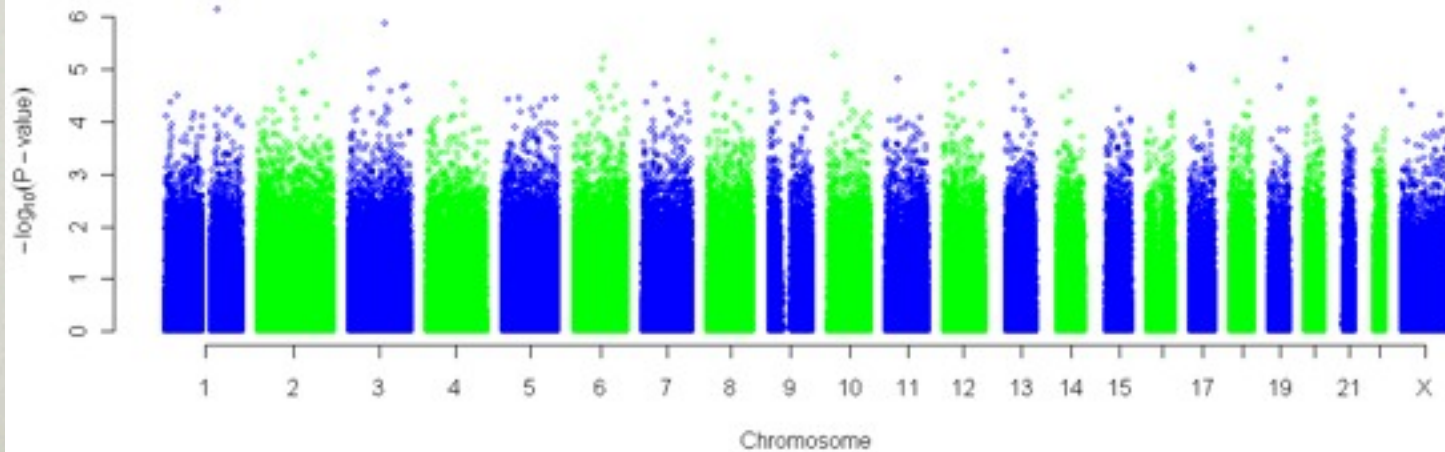


# SKIN COLOR SCAN

GWAS of skin color using the HapMap data



GWAS without any association



# GENOMIC CONTROL

---

- Consider a test distributed as  $\chi^2_1$  under the null (e.g. trend test)
- Compute the vector of test statistics  $\{T^2_1, T^2_2, T^2_3, \dots, T^2_{N-1}, T^2_N\}$
- Estimate  $\lambda$  as
  - ★ Median $\{T^2_1, T^2_2, T^2_3, \dots, T^2_{N-1}, T^2_N\} / 0.455$
  - ★ Slope of regression of observed onto expected
- The GC-corrected test statistic  $T^2 / \lambda \sim \chi^2_1$
- In practice, all (or large proportion of) GW test are used to estimate  $\lambda$

# $\lambda$ IS DEPENDENT ON SAMPLE SIZE

---

$\lambda$  is related to non-centrality parameter, thus it grows with sample size. Therefore  $\lambda$  should be estimated per certain sample size. This is especially important if

- SNP call rate is different between SNPs
- When reporting the results

# STANDARDIZED $\lambda$

---

- For QT analysis,  $\lambda_n = 1 + (\lambda_{n_{ref}} - 1) n / n_{ref}$   
where  $n_{ref}$  is the reference sample size
- For case / control design

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

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

# 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  $\lambda$  depends on samples size. Thus
  - Report of standardized values (say, per 1,000 cases and 1,000 controls) is recommended
  - Special methods should be used when number of people typed for different SNPs is different

# OUTLINE

---

Confounding in GWA studies

Genomic Control

**Structured Association**

Mixed Models

EigenSTRAT



# STRUCTURED ASSOCIATION

---

- 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

$$Y \sim \mu + \text{sex} + \text{age} + \text{strata} + \text{snp}$$

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

$$Y \sim \mu + \text{strata}^*(\text{sex} + \text{age}) + \text{snp}$$

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

Structured Association

**Mixed Models**

EigenSTRAT

# ESTIMATION OF KINSHIP FROM GENOMIC DATA

---

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

# MIXED MODEL

---

Vector of quantitative phenotype  $Y$

$$Y = \mu + \beta_g g + \mathbf{G} + e$$

$g$ : genotype indicator vector  $g_i$  in  $\{0,1,2\}$

$\beta_g$ : additive affect of the allele

$e$ : random residual effect  $\sim \text{MVN}(\mathbf{0}, I\sigma_e^2)$

$\mathbf{G}$ : random polygenic effect  $\sim \text{MVN}(\mathbf{0}, \Phi \sigma_G^2)$

# MIXED MODELS FOR GWAS

---

# MIXED MODELS FOR GWAS

---

- Excellent method to account for complex genetic structure

# MIXED MODELS FOR GWAS

---

- Excellent method to account for complex genetic structure
- May be very computationally extensive



# MIXED MODELS FOR GWAS

---

- Excellent method to account for complex genetic structure
- May be very computationally extensive
- Therefore is normally used only when other methods fail

# OUTLINE

---

Confounding in GWA studies

Genomic Control

Structured Association

Mixed Models

**EigenSTRAT**

# IDEA OF MULTIDIMENSIONAL SCALING

---

- Study of  $N$  subjects
- $N \times N$  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  $N \times N$  matrix
- Classical MDS is also known as Principal Components Analysis

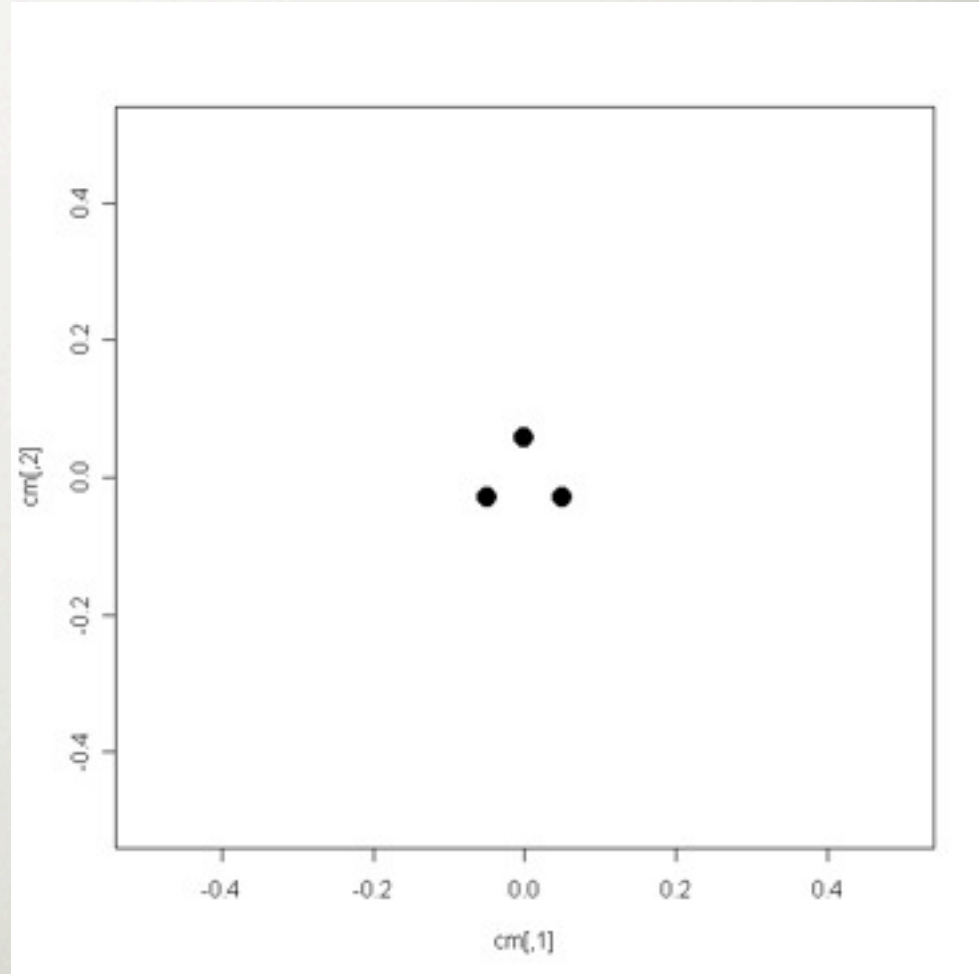
# EXAMPLE CMDS

## 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



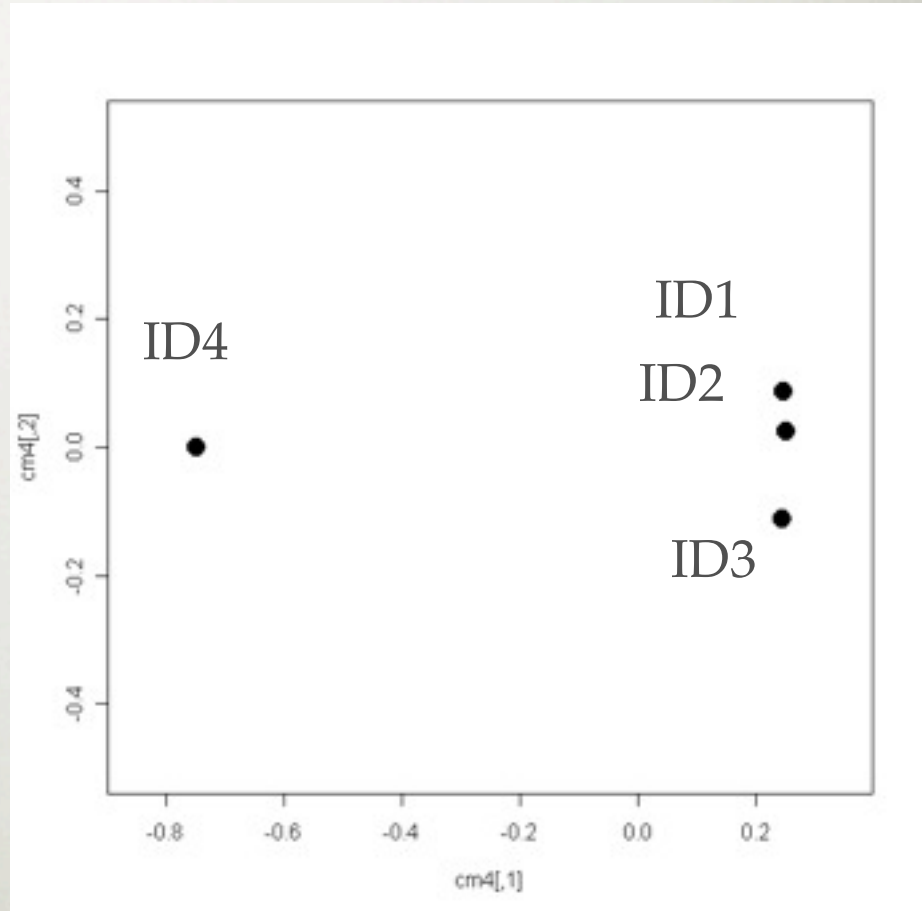
# EXAMPLE CMDS

Distance matrix

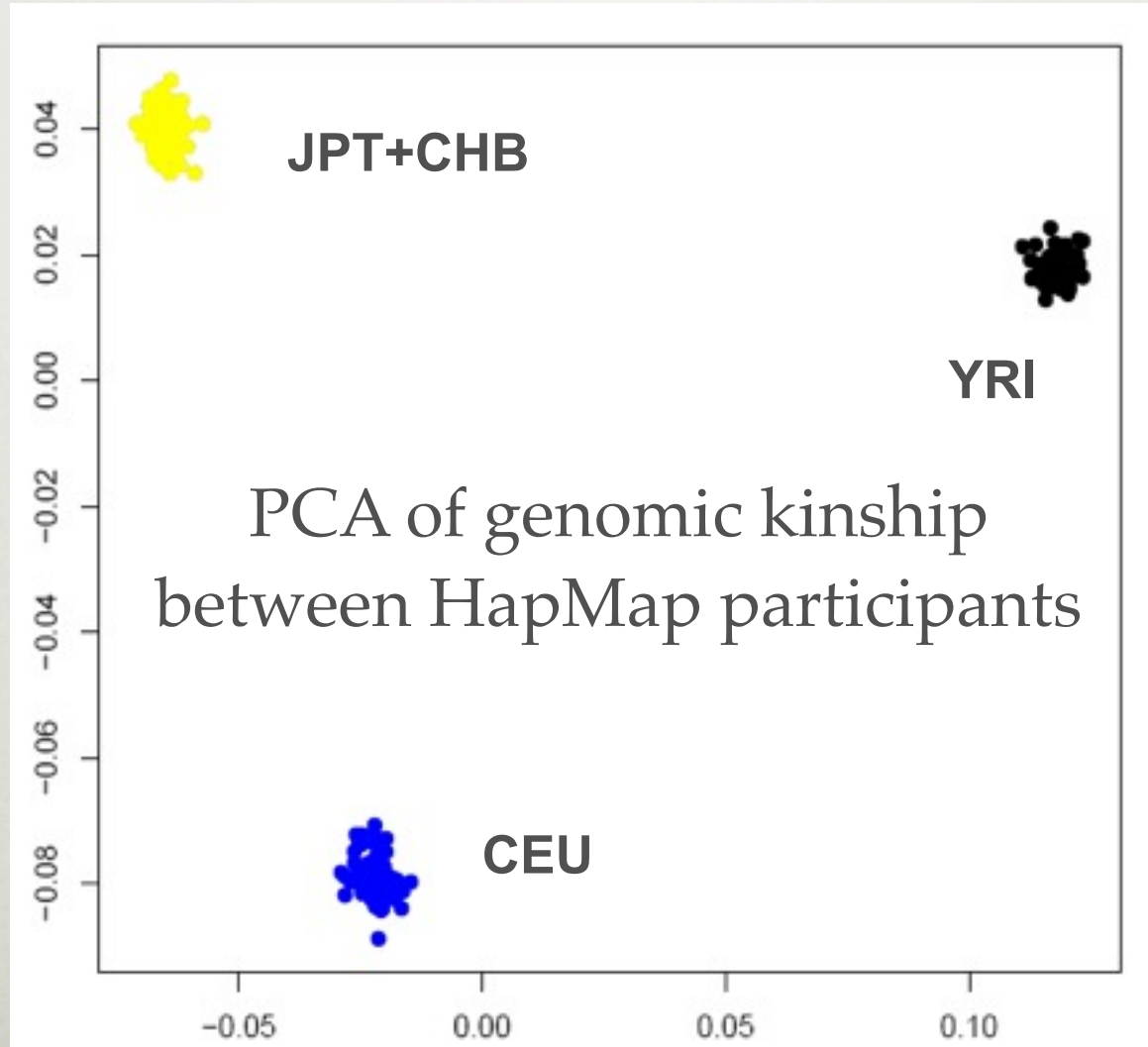
	ID1	ID2	ID3	ID4
ID1	0	0.1	0.15	1.00
ID2	0.1	0	0.20	1.00
ID3	0.15	0.20	0	1.00
ID4	1.00	1.00	1.00	0

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



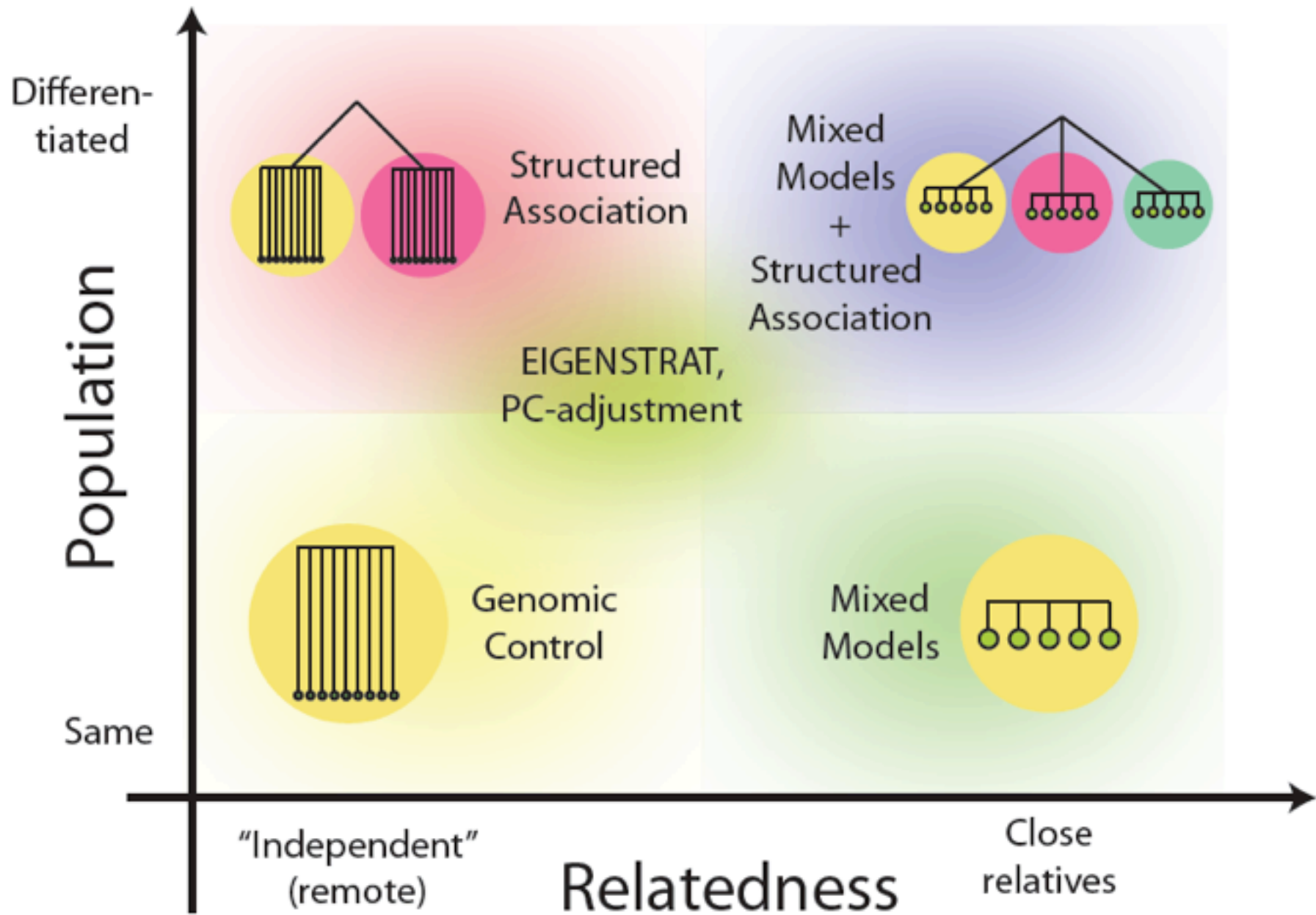
# PCA OF GENOMIC KINSHIP



# IDEA OF EIGENSTRAT

---

- 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 ( $1 < \lambda < 1.05$ )





# 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: qtscore() of GenABEL; also you can do separate analyses and then meta-analyse
- Genomic kinship matrix (base for EIGENSTRAT, PC-adjustment): PLINK's 'IBD', GenABEL's ibs() function
- EIGENSTRAT: EIGENSTRAT, GenABEL's egscore() function
- Adjustment for PCs: any GWA software supporting covariates
- Mixed-models: GenABEL's mmscore & grammar, Merlin (but with pedigree...); MixABEL's GWFGLS

# MIXED MODELS FOR GWAS

---

For convenience, we represent log-likelihood (4) for polygenic model under hypothesis as:

$$\ln L_0 = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2} \sum_{i=1}^n \ln(\eta_i h^2 + 1 - h^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_x)^T \mathbf{T}^T \left\{ \frac{1}{\eta_i h^2 + 1 - h^2} \right\} \mathbf{T} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_x),$$

where  $\left\{ \frac{1}{\eta_i h^2 + 1 - h^2} \right\}$  denote a diagonal matrix with elements  $\frac{1}{\eta_i h^2 + 1 - h^2}$ ,  $i = \overline{1, n}$ .

Then a partial first-order derivative of the log-likelihood (5) on  $\boldsymbol{\beta}_{x_j}$  ( $j = \overline{0, k}$ ) is

$$\frac{\partial(\ln L_0)}{\partial \boldsymbol{\beta}_{x_j}} = -\boldsymbol{\beta}_{x_j} \mathbf{X}_{T_j}^T \left\{ \frac{1}{\eta_i h^2 + 1 - h^2} \right\} \mathbf{X}_{T_j} + \mathbf{X}_{T_j}^T \left\{ \frac{1}{\eta_i h^2 + 1 - h^2} \right\} \left( \mathbf{y}_T - \sum_{m=0, m \neq j}^k \mathbf{X}_{T_m} \boldsymbol{\beta}_{x_m} \right) = 0,$$

where  $\mathbf{X}_{T_j}$  is j-th column-vector of the matrix  $\mathbf{X}_T$ .

# MIXED MODELS FOR GWAS

---

From above it follows that

$$\boldsymbol{\beta}_{\mathbf{X}_j} = \frac{\mathbf{X}_{\mathbf{T}_j}^T \left\{ \frac{1}{\eta_i h^2 + 1 - h^2} \right\} \left( \mathbf{y}_{\mathbf{T}} - \sum_{m=0, m \neq j}^k \mathbf{X}_{\mathbf{T}_m} \boldsymbol{\beta}_{\mathbf{X}_m} \right)}{\mathbf{X}_{\mathbf{T}_j}^T \left\{ \frac{1}{\eta_i h^2 + 1 - h^2} \right\} \mathbf{X}_{\mathbf{T}_j}} \text{ for } j = \overline{0, k}.$$

The system of expressions (6) can be written in matrix form as

$$\boldsymbol{\beta}_{\mathbf{X}} = \left( \mathbf{X}_{\mathbf{T}}^T \left\{ \frac{1}{\eta_i h^2 + 1 - h^2} \right\} \mathbf{X}_{\mathbf{T}} \right)^{-1} \mathbf{X}_{\mathbf{T}}^T \left\{ \frac{1}{\eta_i h^2 + 1 - h^2} \right\} \mathbf{y}_{\mathbf{T}}$$

So the vector  $\boldsymbol{\beta}_{\mathbf{X}}$  is estimated only through  $h^2$ .