

# DEALING WITH GENETIC (SUB)STRUCTURE IN GWAS

YURII AULCHENKO

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

# OUTLINE

---

Confounding in GWA studies

Genomic Control

Structured Association

EigenSTRAT

Mixed Models

# REASONS FOR GENETIC ASSOCIATION

---



# REASONS FOR GENETIC ASSOCIATION

---

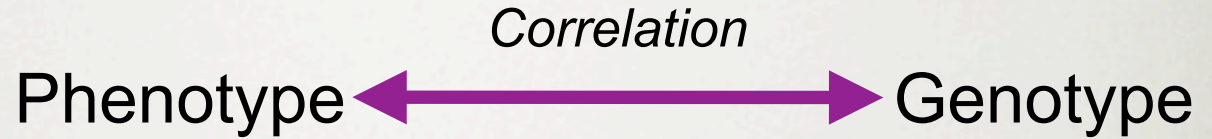
**What we see**



# REASONS FOR GENETIC ASSOCIATION

---

**What we see**



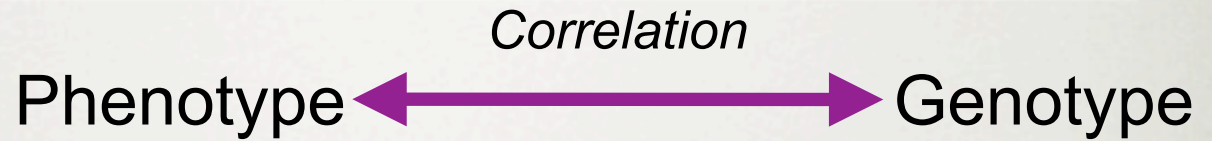
**True model**



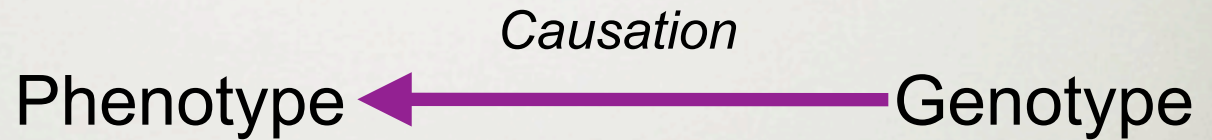
# REASONS FOR GENETIC ASSOCIATION

---

**What we see**

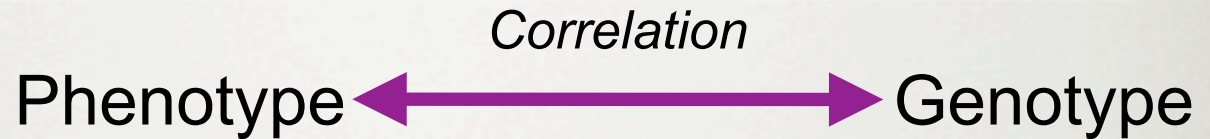


**True model**

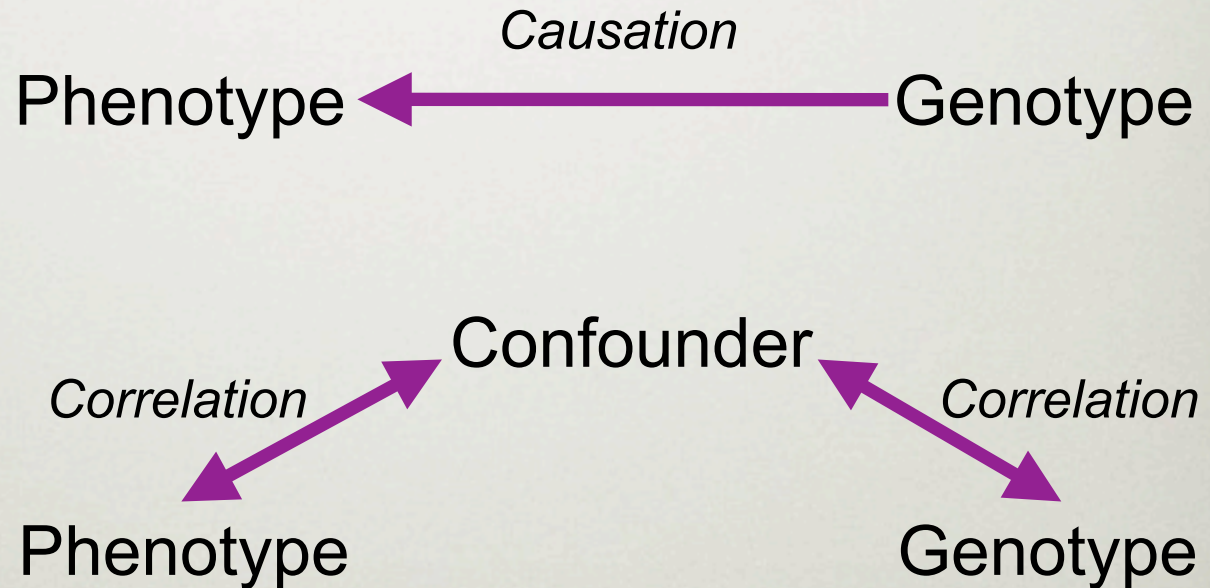


# REASONS FOR GENETIC ASSOCIATION

**What we see**



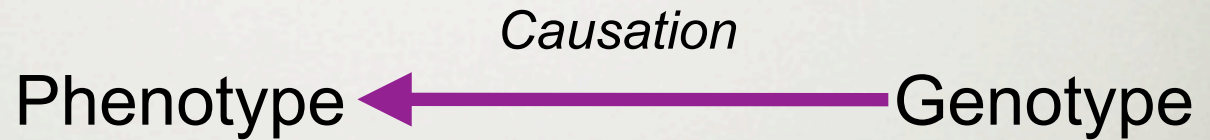
**True model**



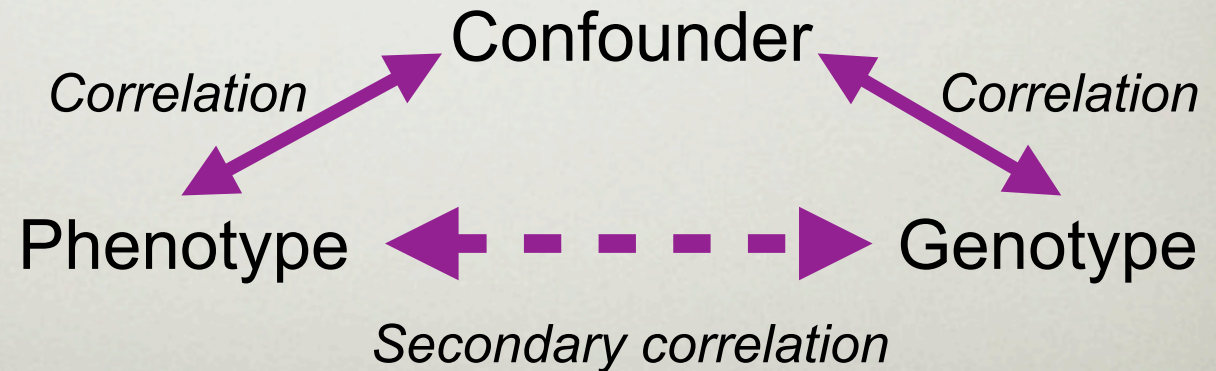


# REASONS FOR GENETIC ASSOCIATION

**What we see**



**True model**






# CONFOUNDING IN GENETIC STUDIES

---

# CONFOUNDING IN GENETIC STUDIES

---

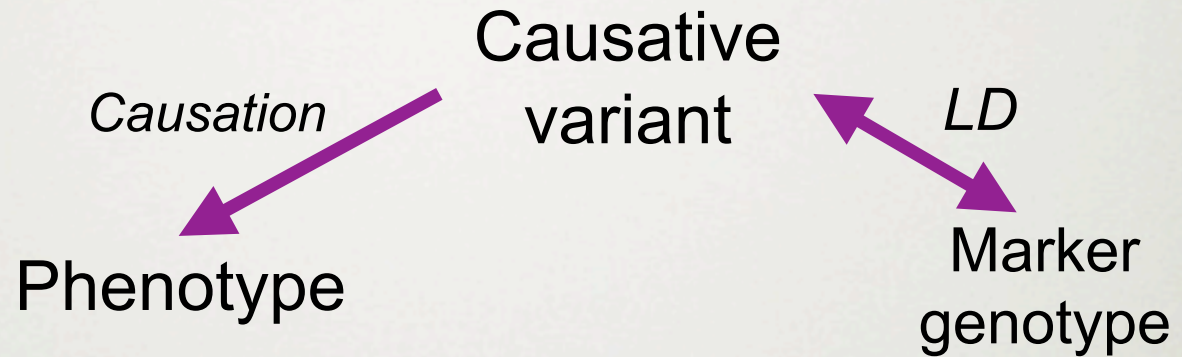
**LD  
mapping**

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

# CONFOUNDING IN GENETIC STUDIES

---

**LD  
mapping**

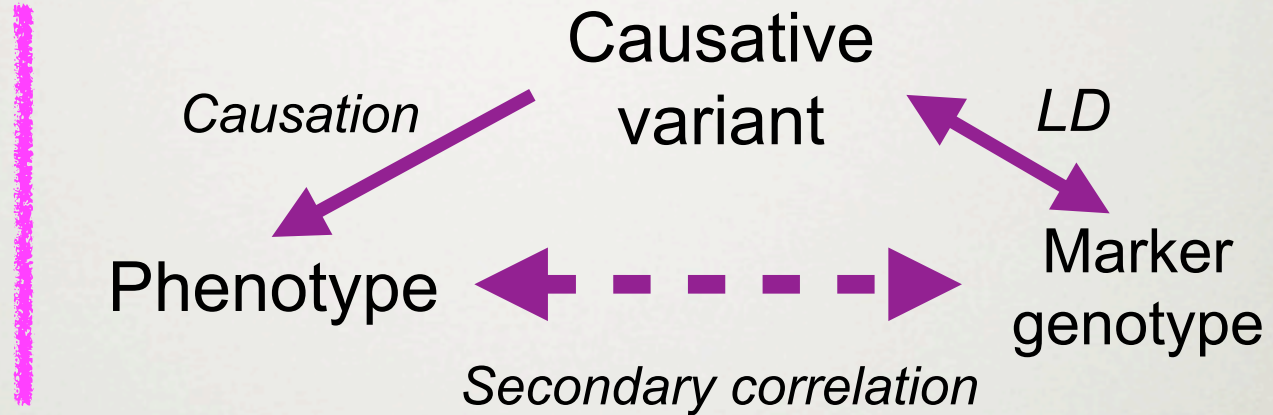




# CONFOUNDING IN GENETIC STUDIES

---

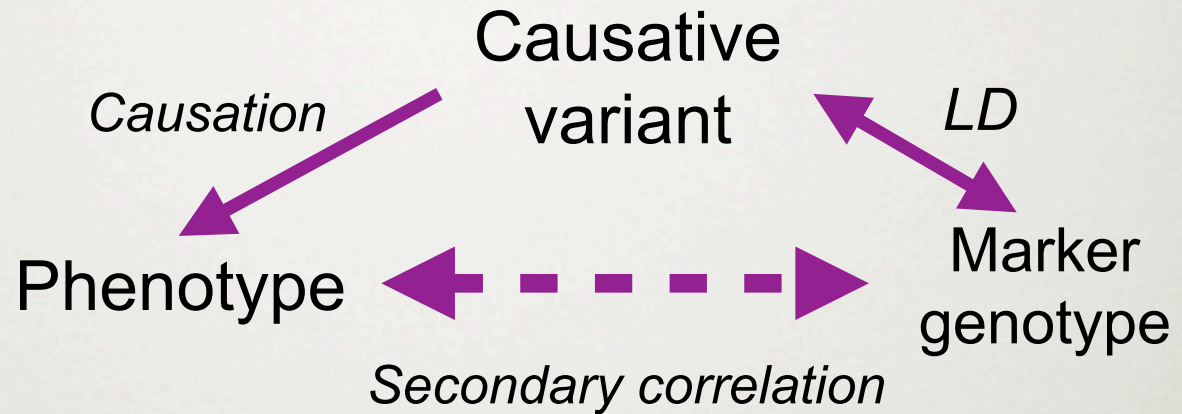
**LD  
mapping**



# CONFOUNDING IN GENETIC STUDIES

---

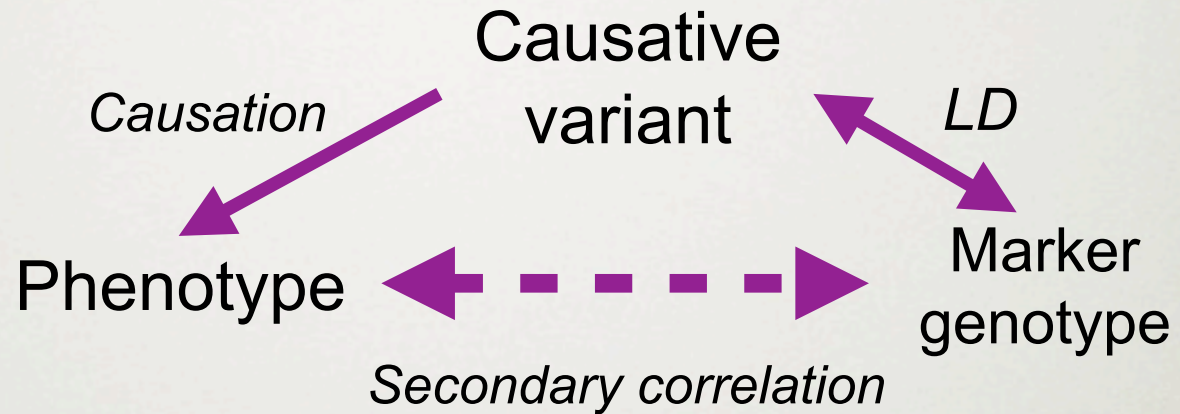
**LD mapping**



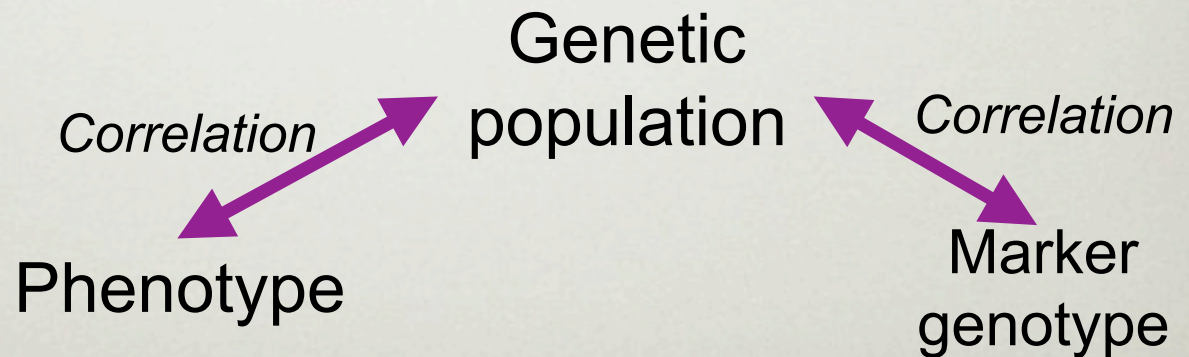
**Stratification**

# CONFOUNDING IN GENETIC STUDIES

**LD mapping**



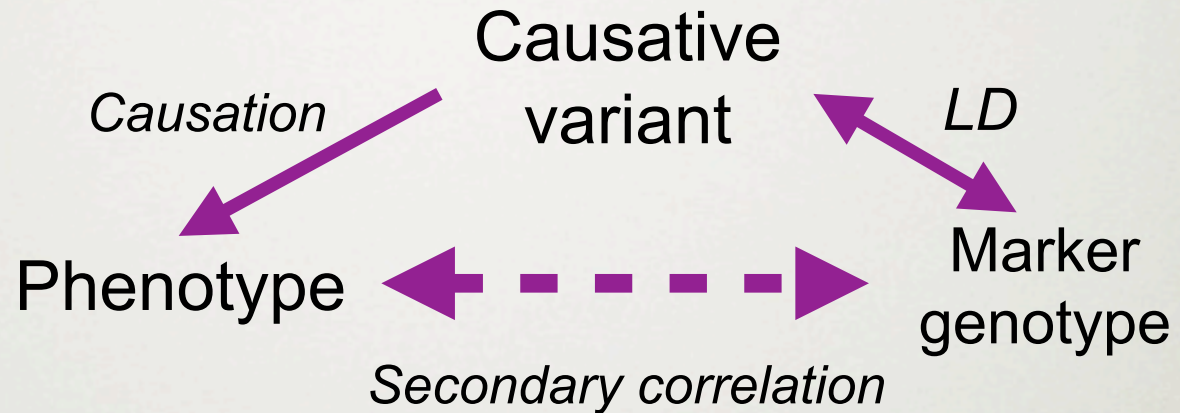
**Stratification**



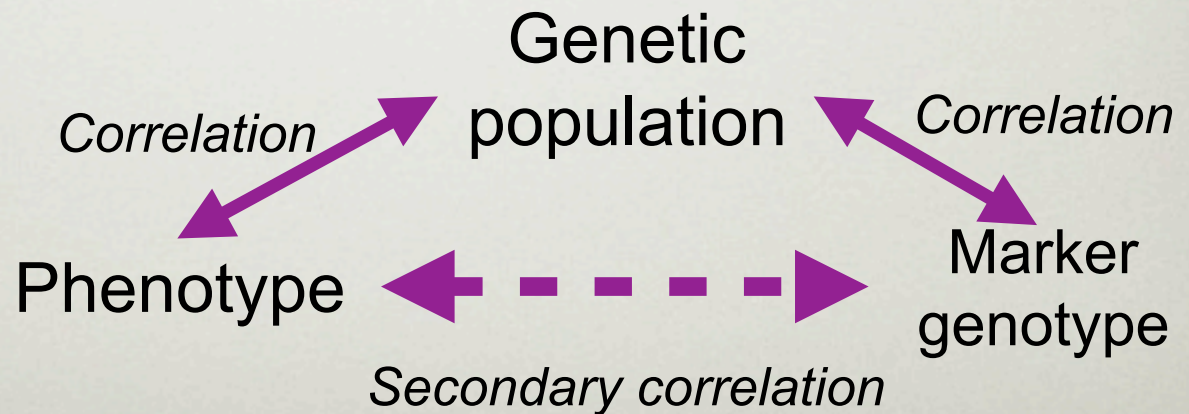


# CONFOUNDING IN GENETIC STUDIES

**LD mapping**

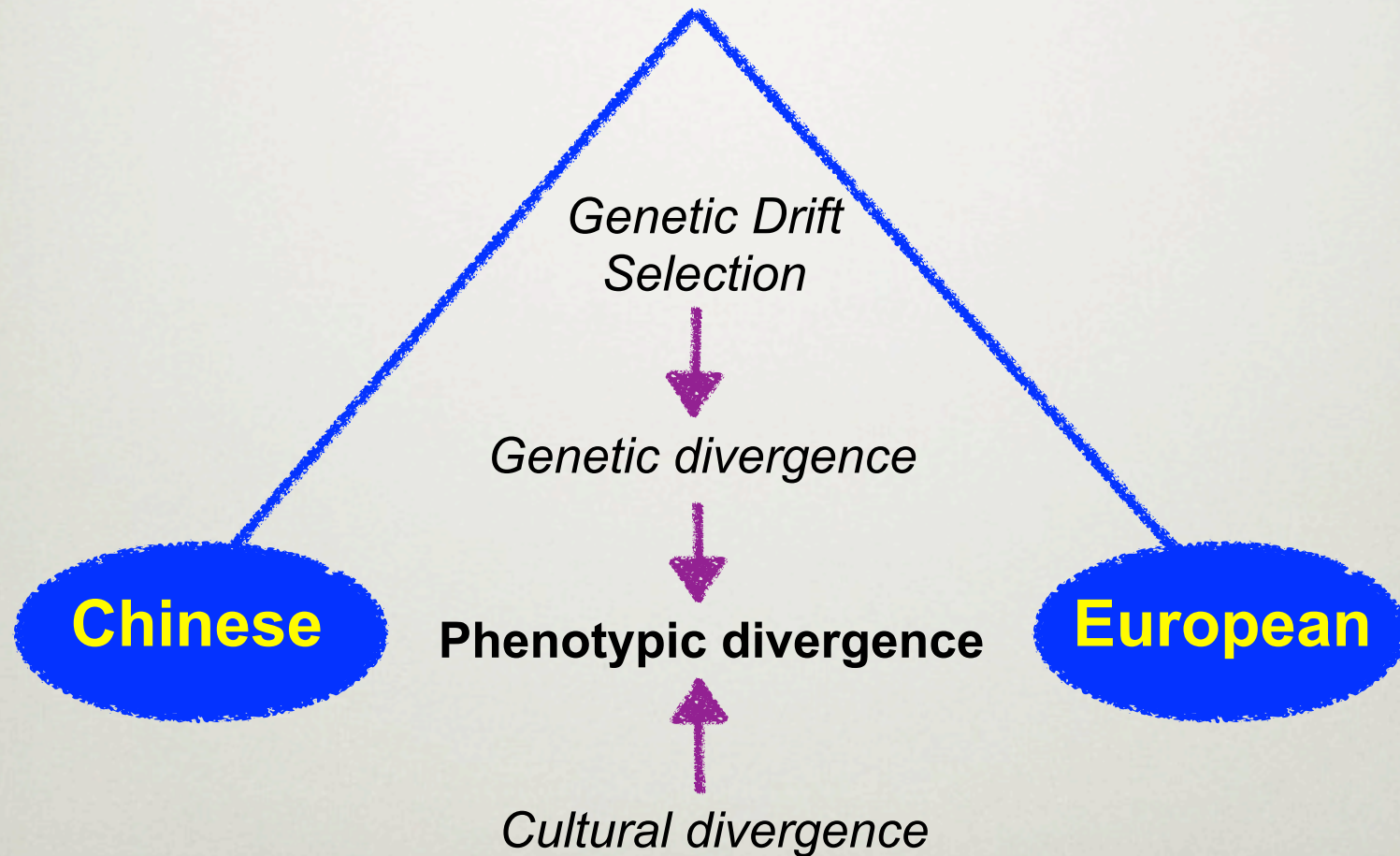


**Stratification**



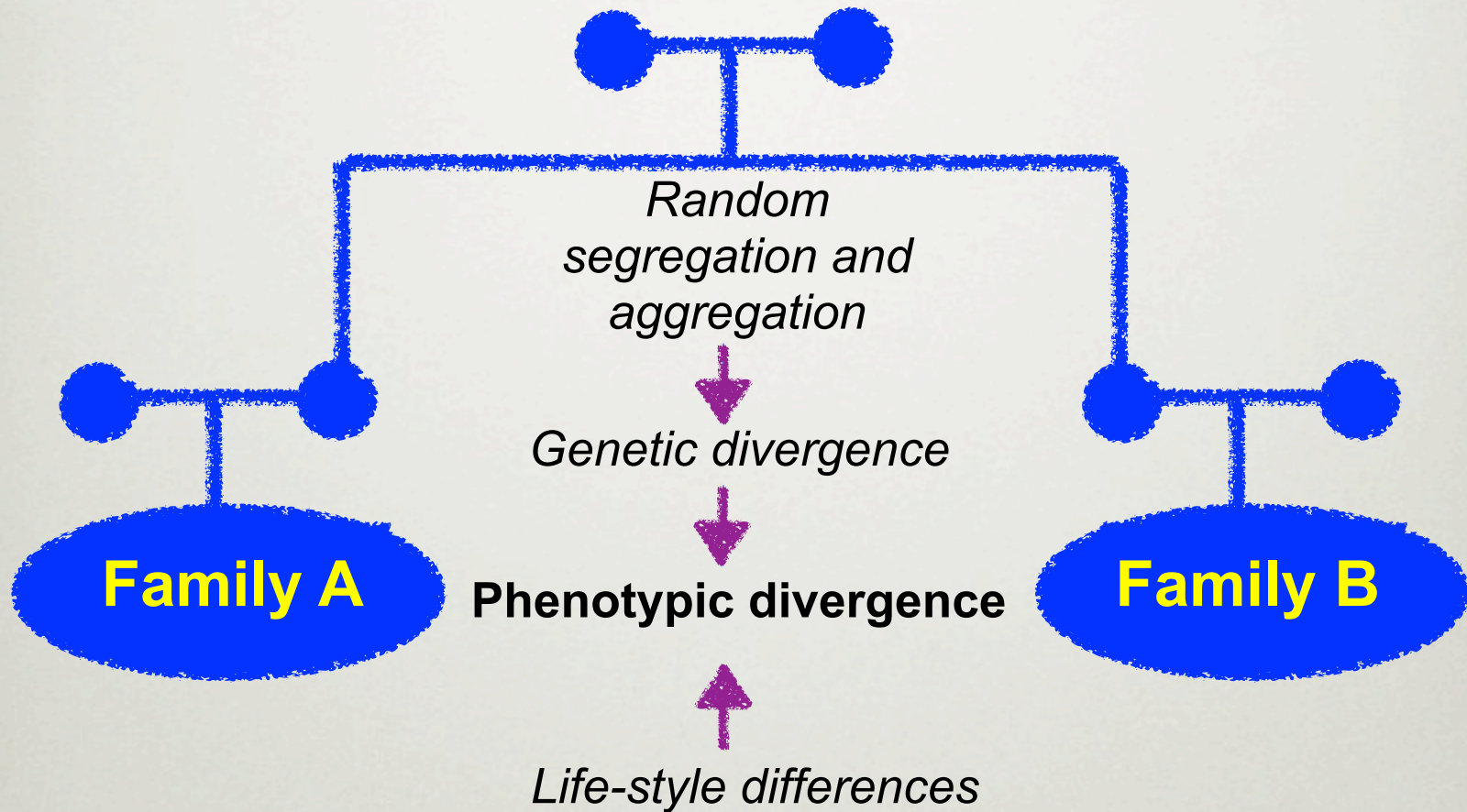
# GENETIC ORIGIN IS A MAJOR CONFOUNDER

---



# 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

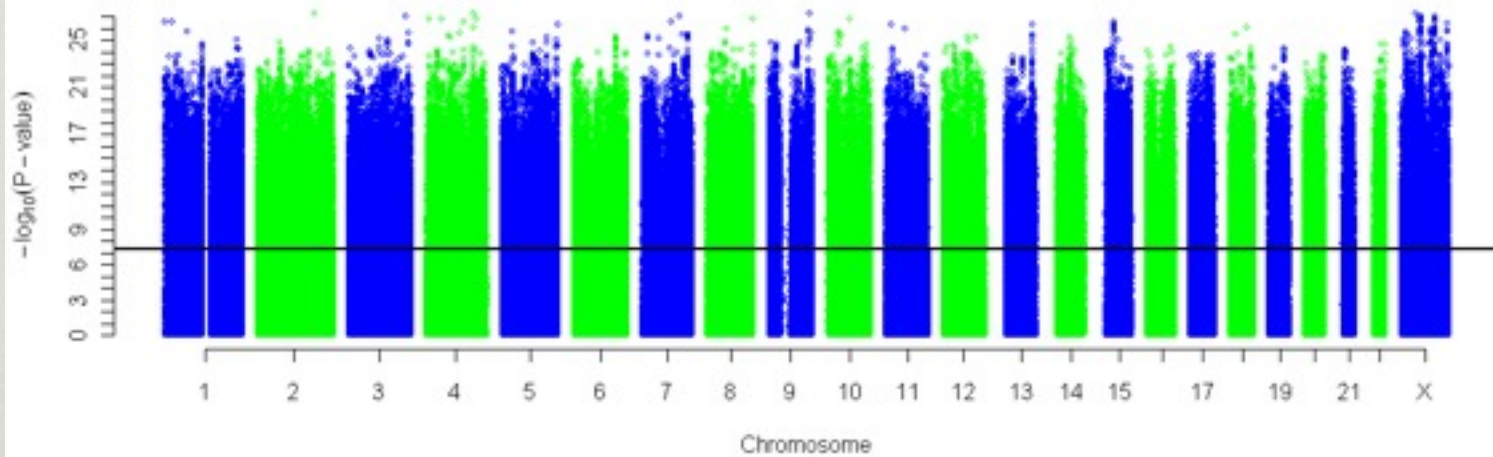
Other causes of genetic stratification are “cryptic” relations or systematic pedigree structure presented in a sample

# SKIN COLOR SCAN

---

# SKIN COLOR SCAN

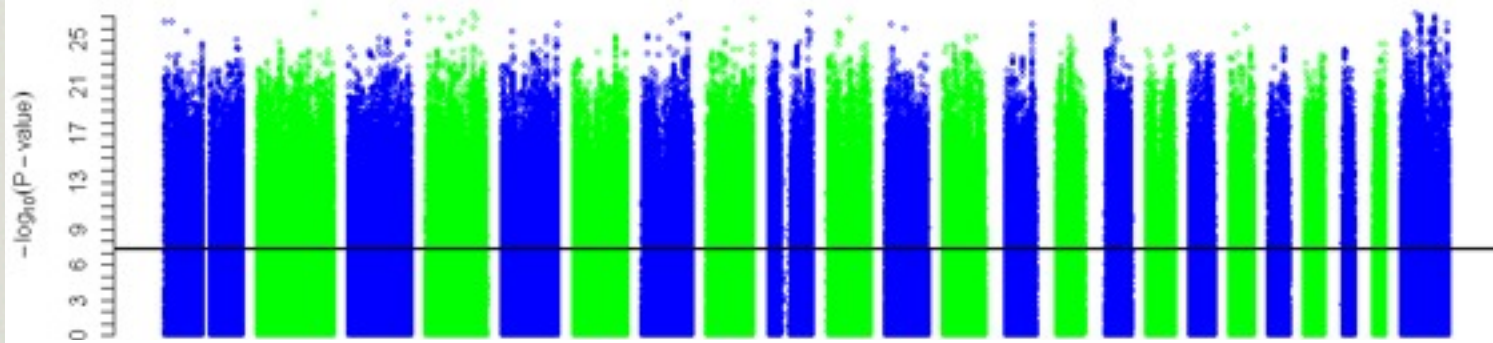
GWAS of skin color using the HapMap data



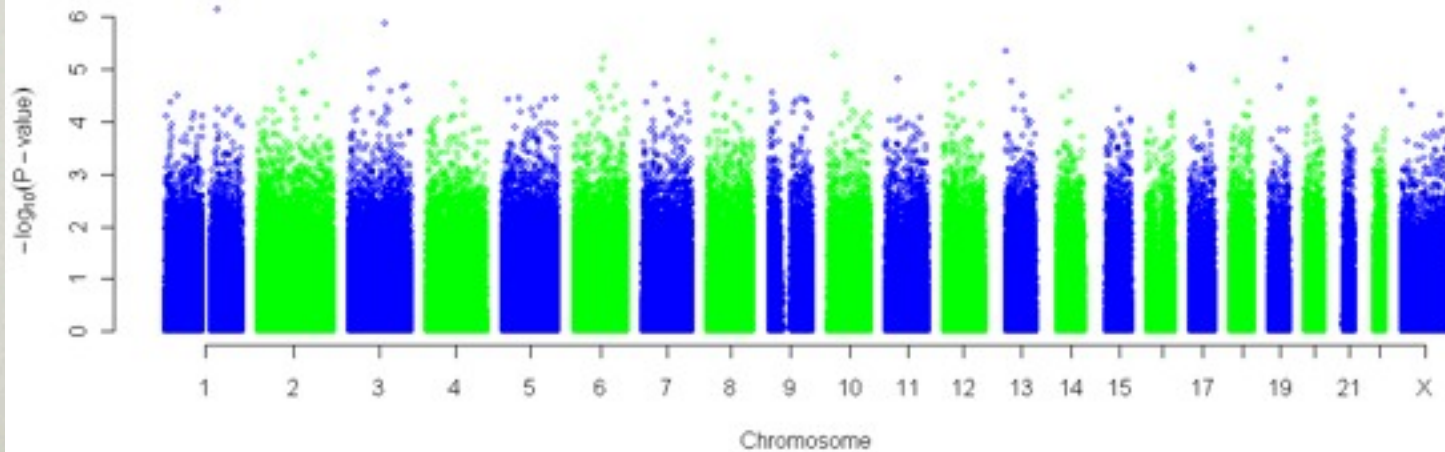


# SKIN COLOR SCAN

GWAS of skin color using the HapMap data



GWAS without any association

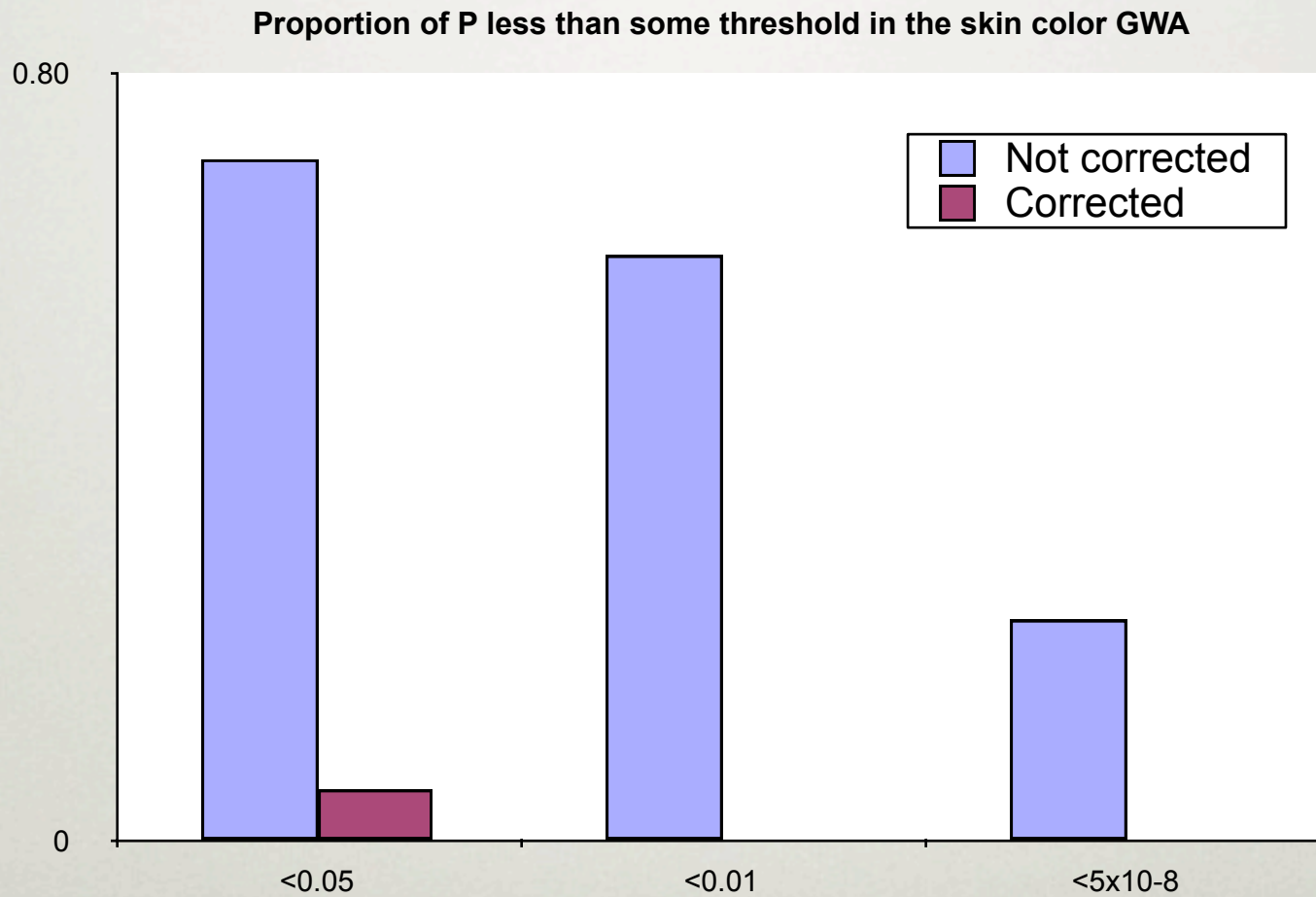


Yurii Aulchenko

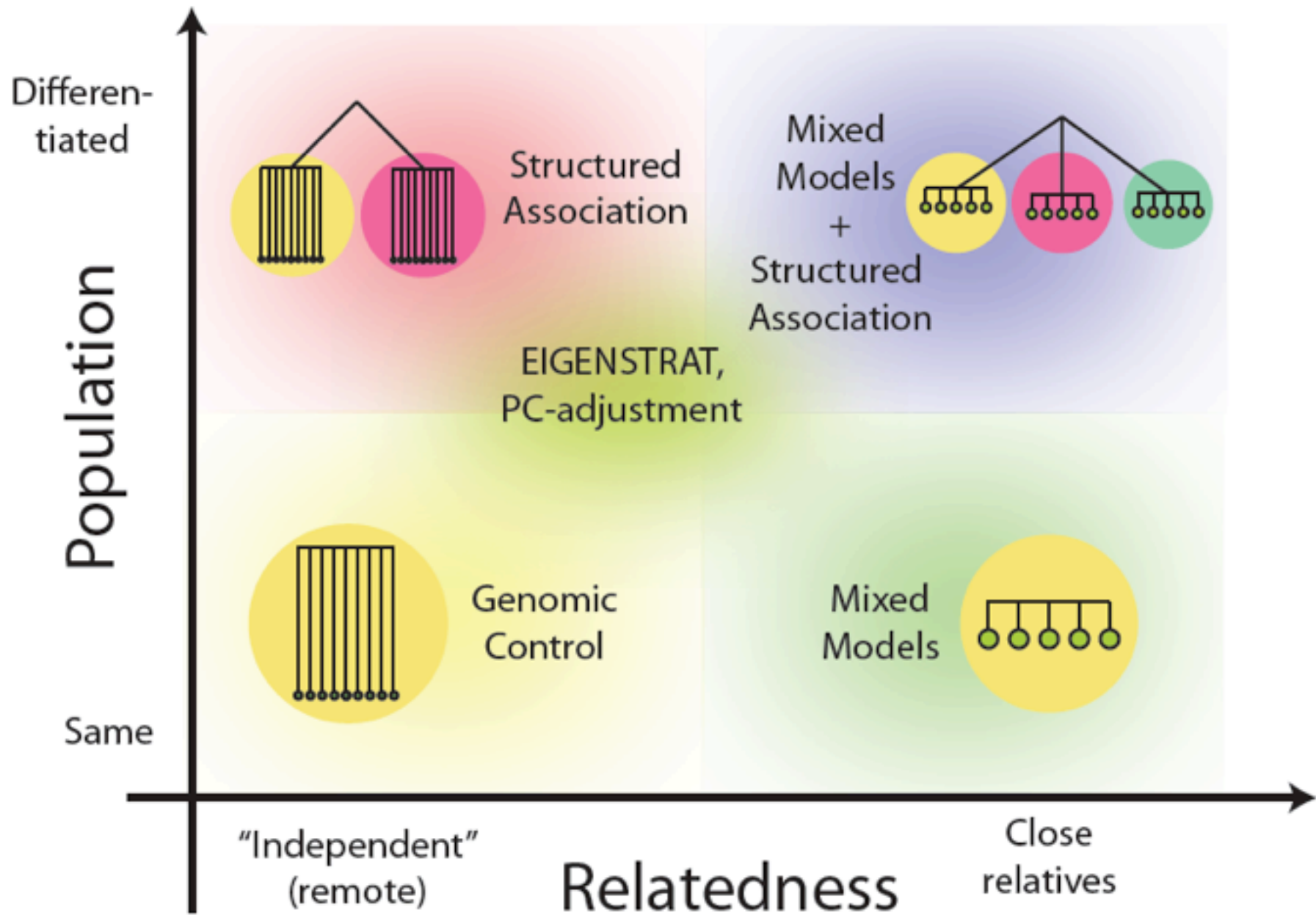
# CONSEQUENCES OF STRATIFICATION

---

# 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

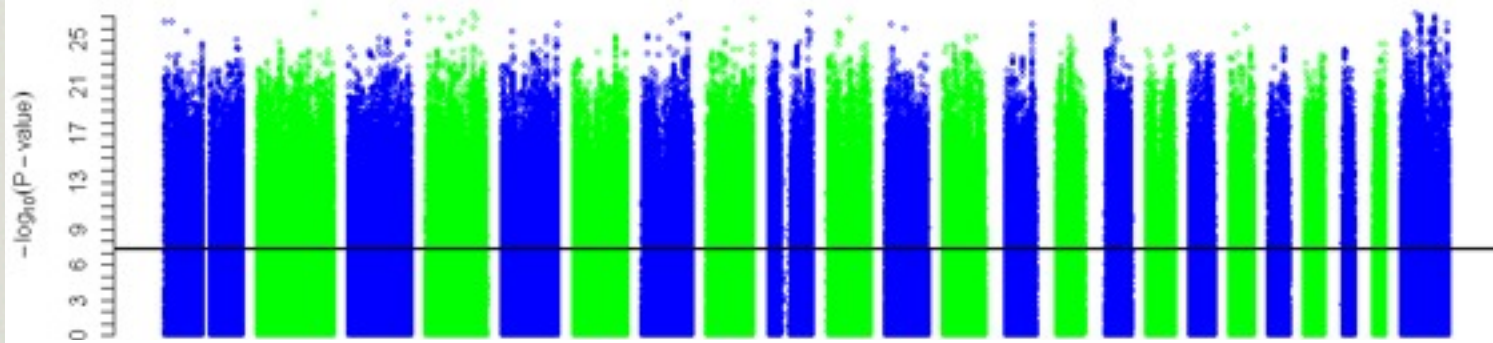
EigenSTRAT

Mixed Models

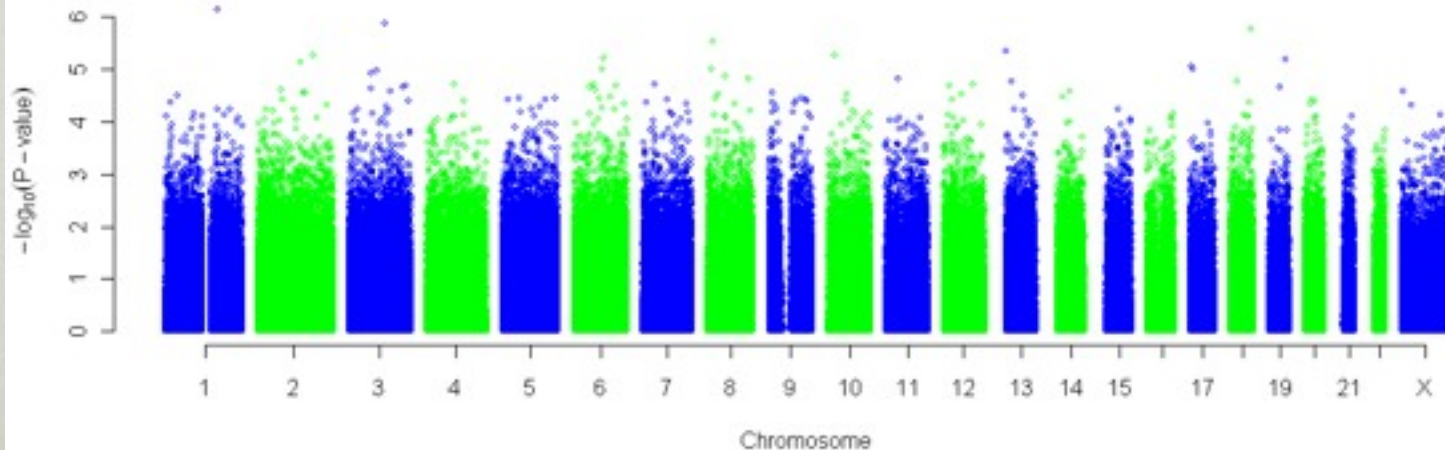


# SKIN COLOR SCAN

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  $\lambda$

# 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. 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

---

- 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

# OUTLINE

---

Confounding in GWA studies

Genomic Control

Structured Association

**EigenSTRAT**

Mixed Models



# HOW SIMILAR ARE GENOMES?

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

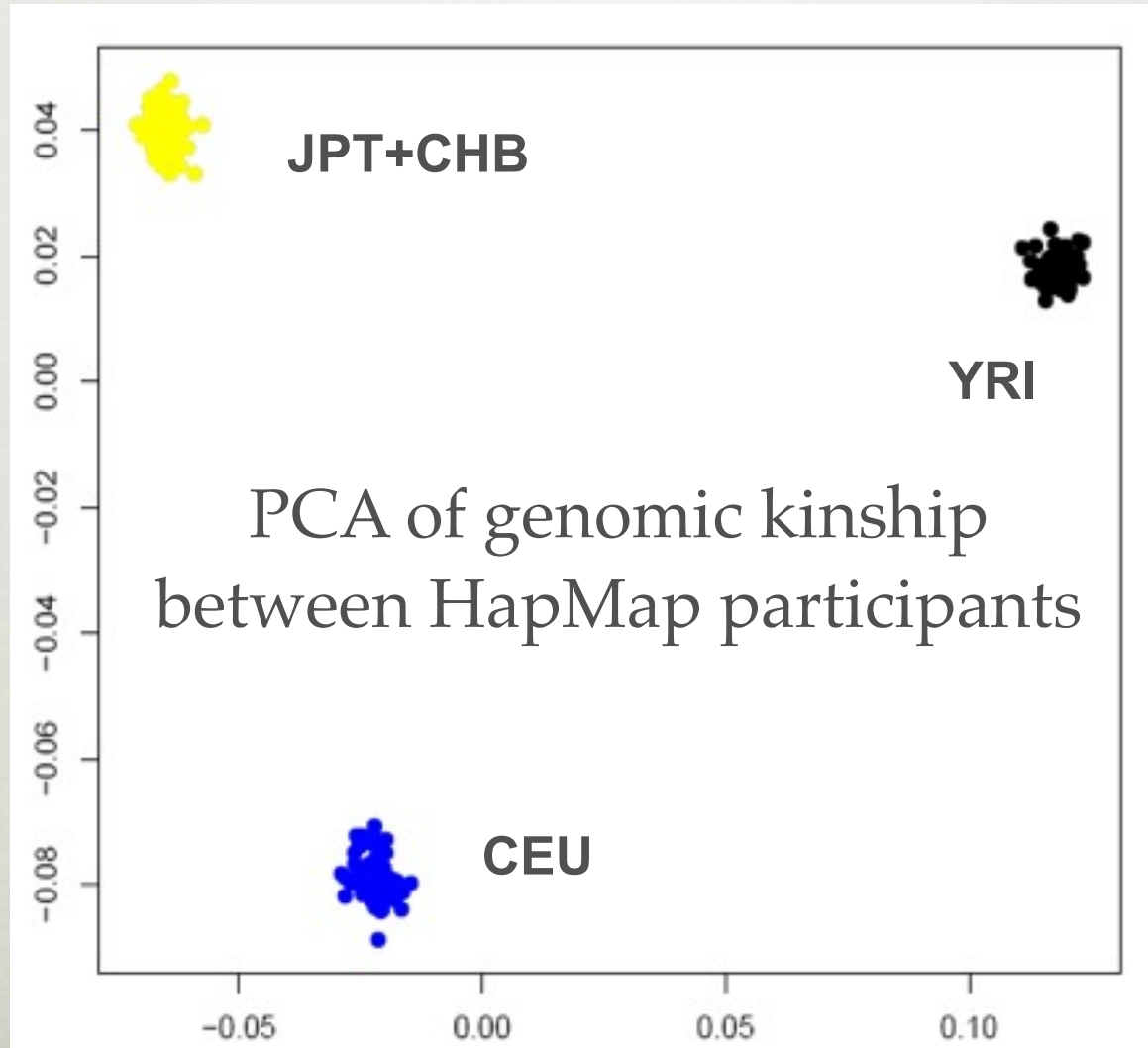
$$f_{ij} = \frac{1}{n} \sum_{k=1}^n \frac{(g_{ik} - p_k)(g_{jk} - p_k)}{p_k(1 - p_k)}$$

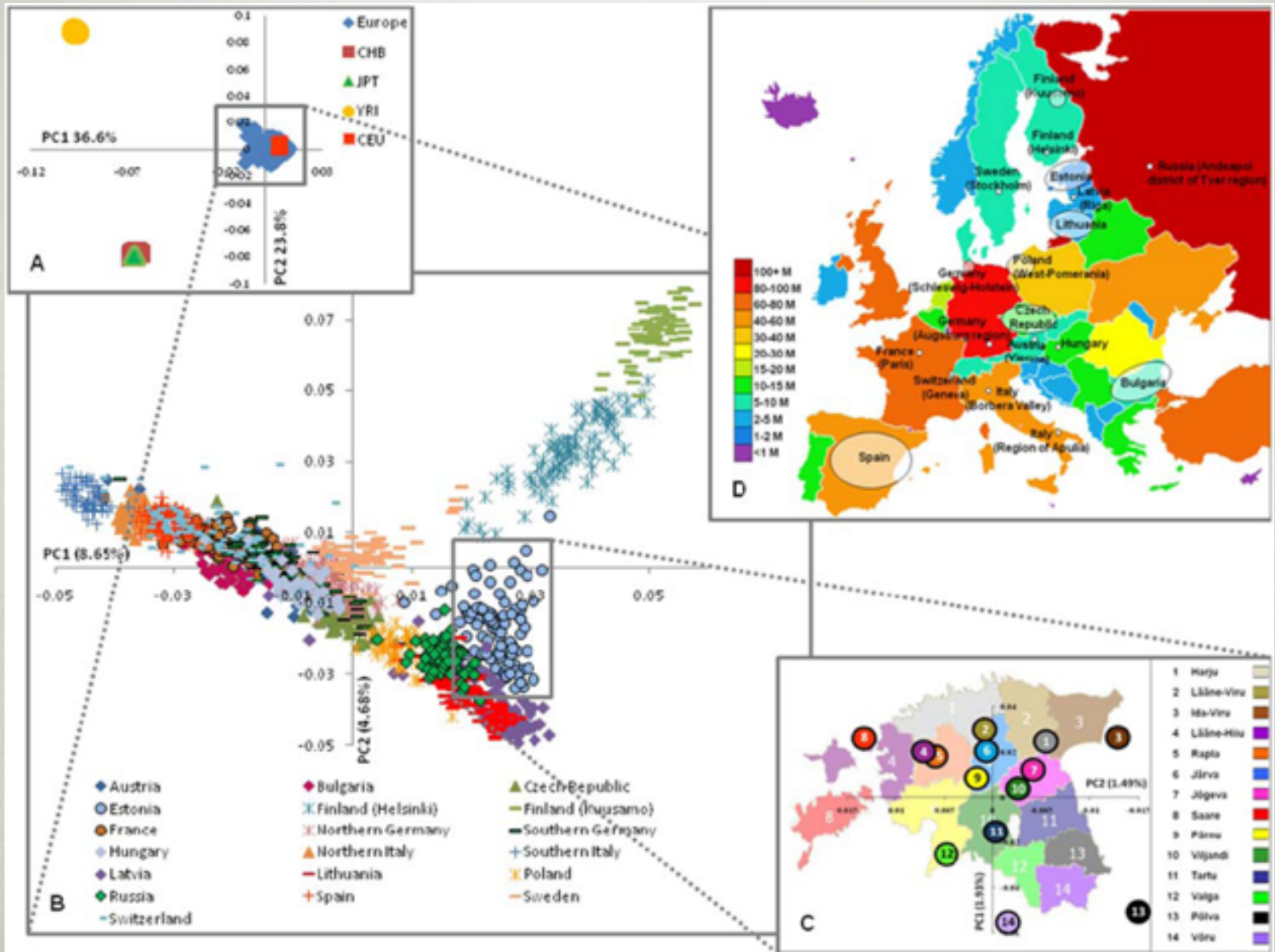
$g_{ik}$  is the genotype (0, 0.5, 1) of the  $i$ -th person at  $k$ -th SNP

$p_k$  is the frequency of the effective allele

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

# PCA OF GENOMIC KINSHIP





Nelis et al., PLoS ONE, 2009



# IDEA OF EIGENSTRAT

---

- Estimate genetic relations between the study participants using genomic data, compute pair-wise distance matrix
- Extract 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)

# OUTLINE

---

Confounding in GWA studies

Genomic Control

Structured Association

EigenSTRAT

**Mixed Models**

# 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  $\sim \text{MVN}(\mathbf{0}, \Phi \sigma_G^2)$



# ESTIMATION OF KINSHIP FROM GENOMIC DATA

---

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

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

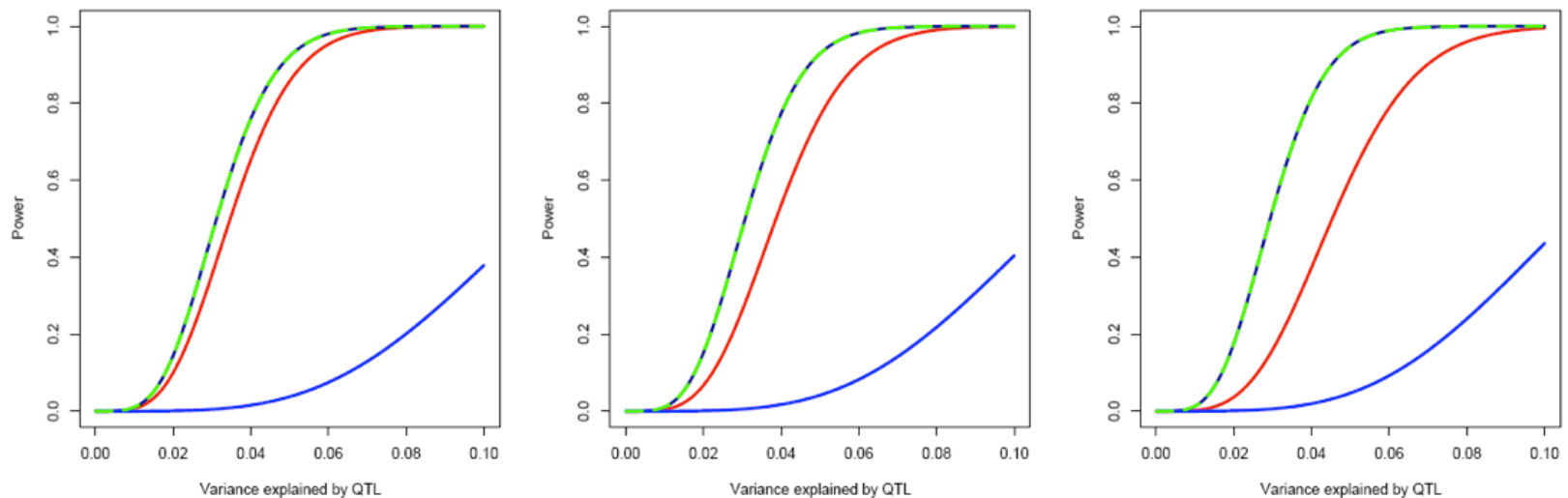
$g_{ik}$  is the genotype (0, 0.5, 1) of the  $i$ -th person at  $k$ -th SNP

$p_k$  is the frequency of "1" allele

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

# COMPARISON FOR AN ISOLATED POPULATION

Comparison of power of FASTA (upper line) and GC-corrected score test (red line). Three panels correspond to different trait heritability (0.3, 0.5, 0.8)



# COMPARISON FOR A “POPULATION-BASED” STUDY

**Table 1 Comparison of genomic control inflation factors obtained with different models**

Phenotype	Genomic control inflation factor			
	Uncorrected	IBD < 0.1	ES100	EMMAX
CRP	1.007	1.007	1.019	0.993
TG	1.023	1.010	1.019	1.002
INS	1.029	1.022	1.013	1.005
DBP	1.031	1.019	1.028	1.007
BMI	1.031	1.024	1.016	0.995
GLU	1.045	1.033	1.030	1.008
HDL	1.052	1.056	1.036	1.004
SBP	1.066	1.056	1.021	1.006
LDL	1.098	1.089	1.040	1.002
Height	1.187	1.151	1.074	1.003

ES100, EIGENSOFT correcting for 100 principal components; IBD < 0.1, uncorrected analysis after excluding 611 individuals whose PLINK's IBD estimates with another individual is greater than 0.1; phenotype abbreviations are CRP, C-reactive protein; TG, triglyceride; INS, insulin plasma levels; DBP, diastolic blood pressure; BMI, body mass index; GLU, glucose; HDL, high-density lipoprotein; SBP, systolic blood pressure; LDL, low density lipoprotein.

*Kang et al., Nat Genet, 2010*



# MIXED MODELS FOR GWAS

---

# MIXED MODELS FOR GWAS

---

- Excellent method to account for complex genetic structure, such as found in special populations or in family-based studies

# MIXED MODELS FOR GWAS

---

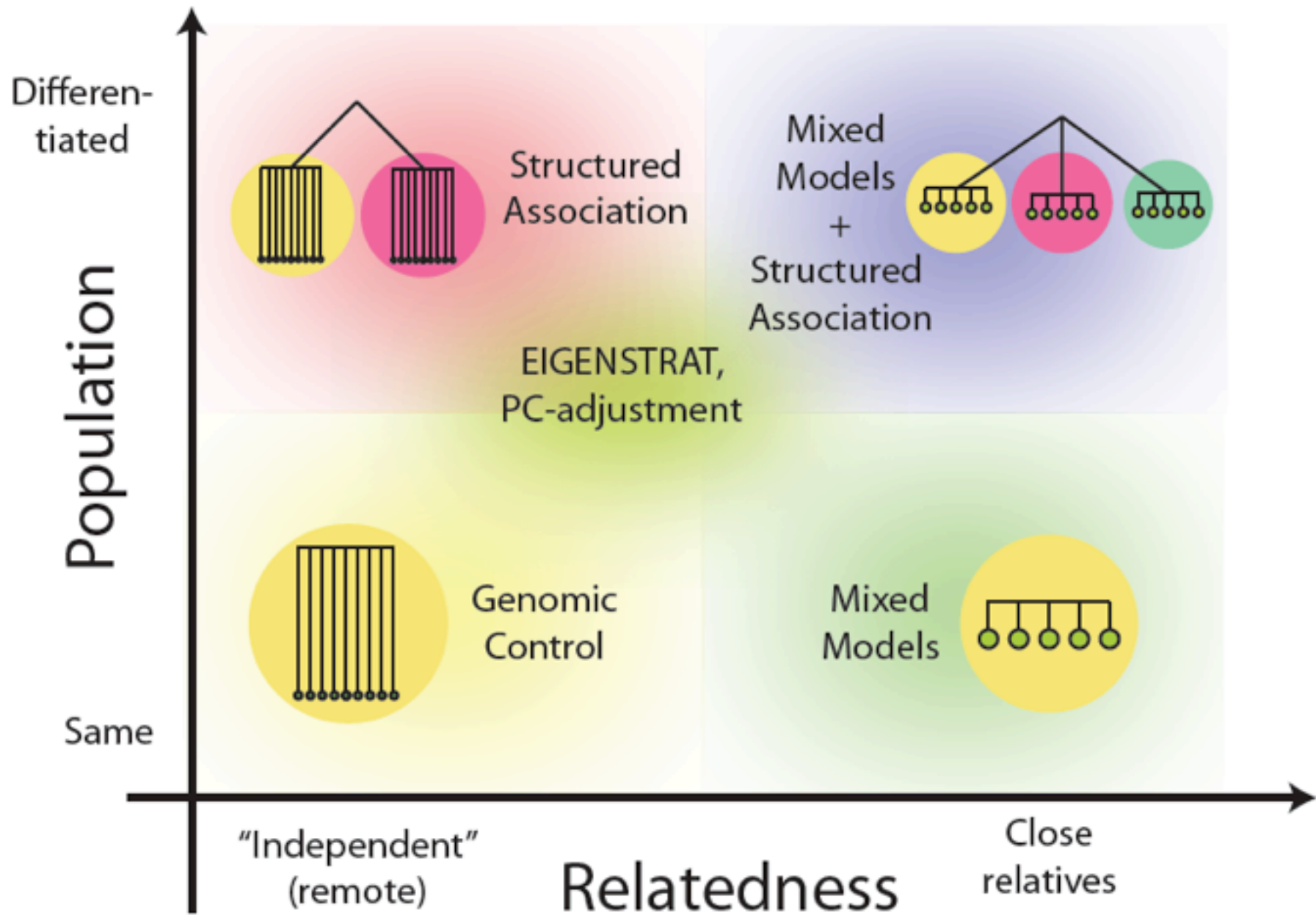
- Excellent method to account for complex genetic structure, such as found in special populations or in family-based studies
- Complex structures found in large “population based” studies



# MIXED MODELS FOR GWAS

---

- Excellent method to account for complex genetic structure, such as found in special populations or in family-based studies
- Complex structures found in large “population based” studies
- May be very computationally extensive



# 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 and FMM; EMMAX; FaST-LMM



# FURTHER TOPICS (GEO3, GEO5)

---

- Advanced Mixed Models analysis:  
power, precision and speed
- Using Mixed Models for analysis of rare variants in the context of whole-genome / exome re-sequencing studies