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

Mixed Models

EigenSTRAT

Reasons for genetic association

Reasons for genetic association

What we see

Reasons for genetic association

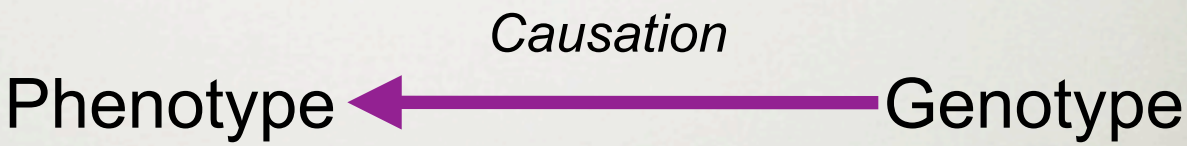
What we see

True model

Reasons for genetic association

What we see

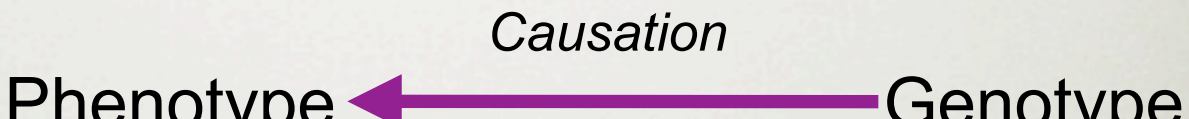
True model

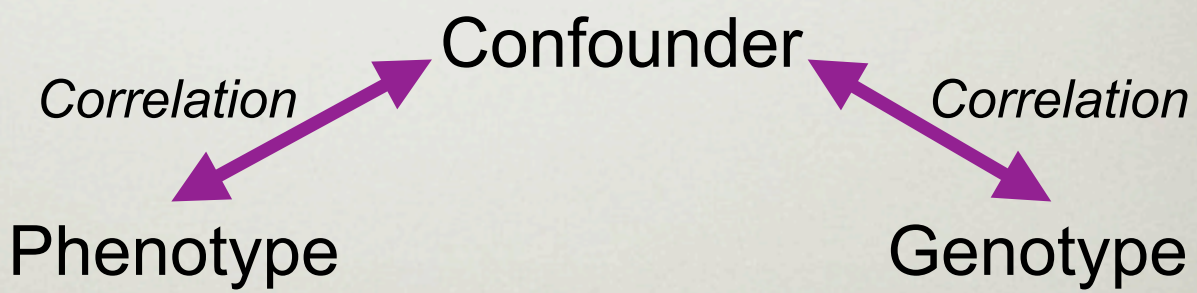


Reasons for genetic association

What we see

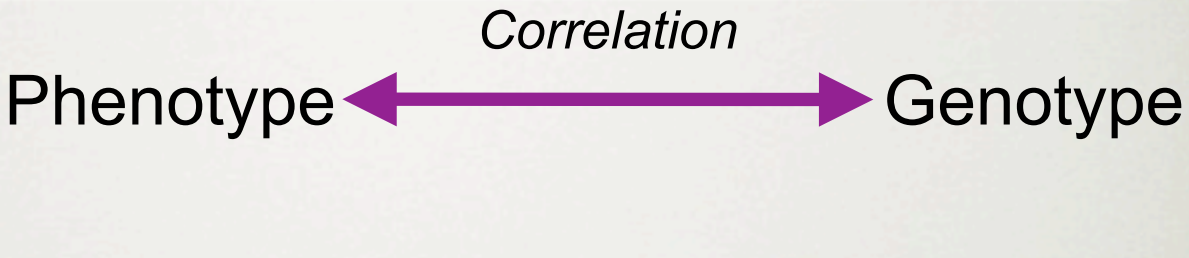
True model



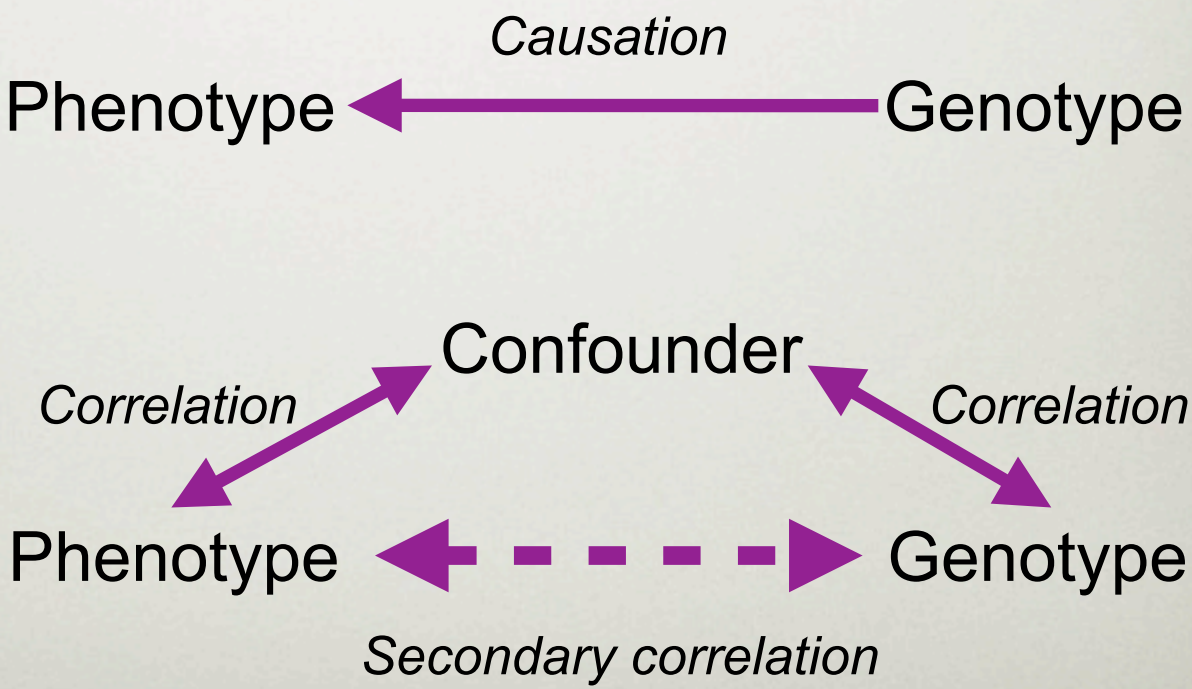


Reasons for genetic association

What we see



True model



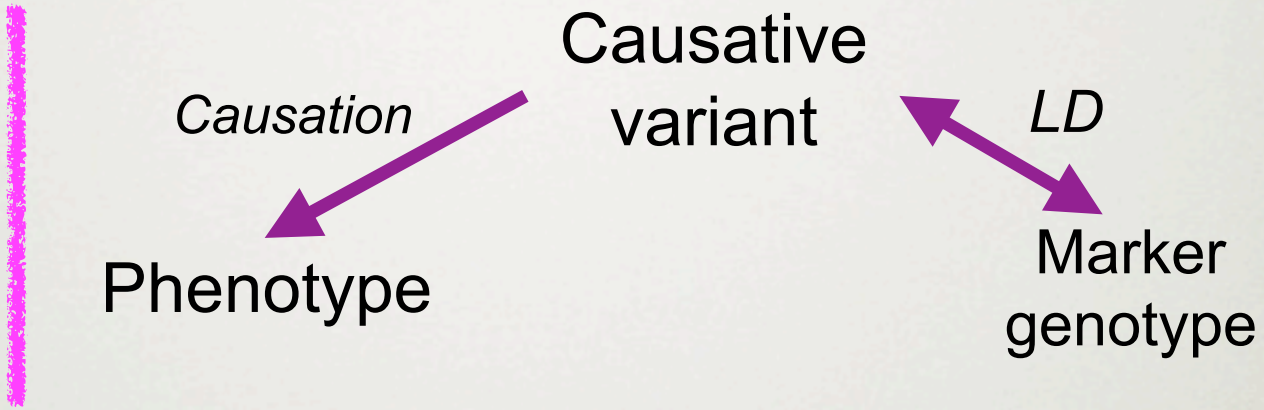
Confounding in genetic studies

Confounding in genetic studies

LD mapping

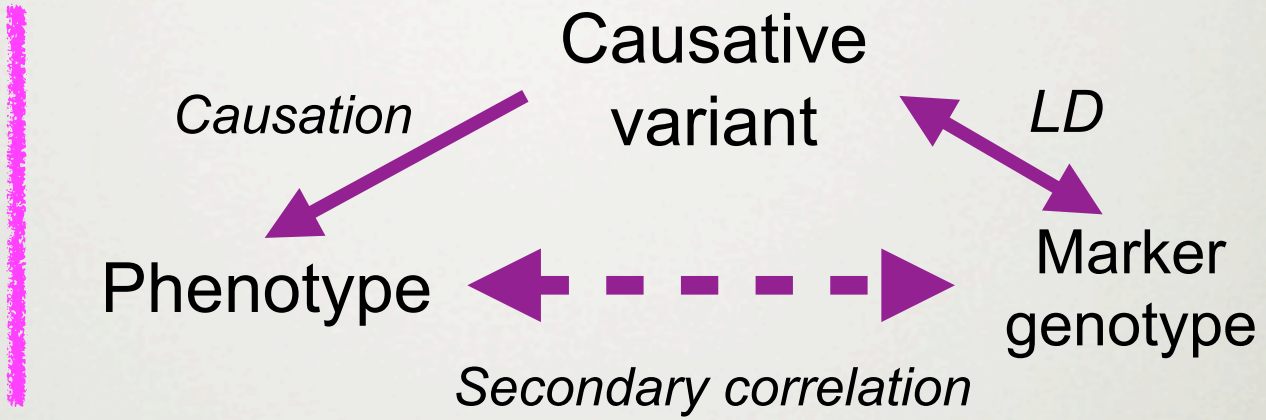
Confounding in genetic studies

LD mapping



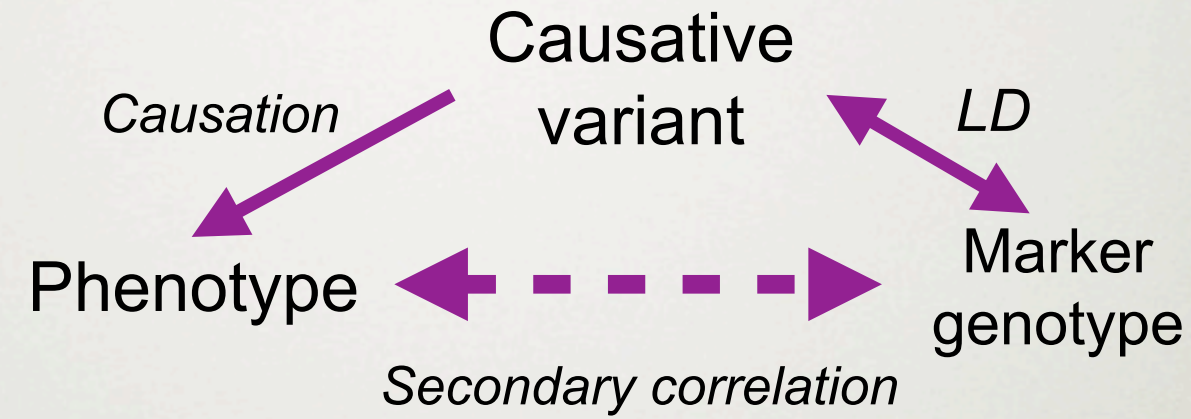
Confounding in genetic studies

LD mapping



Confounding in genetic studies

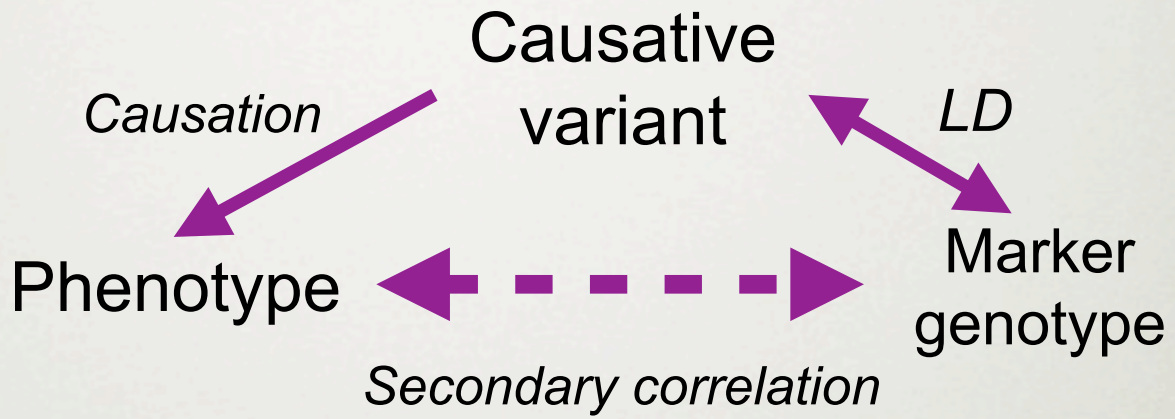
LD mapping



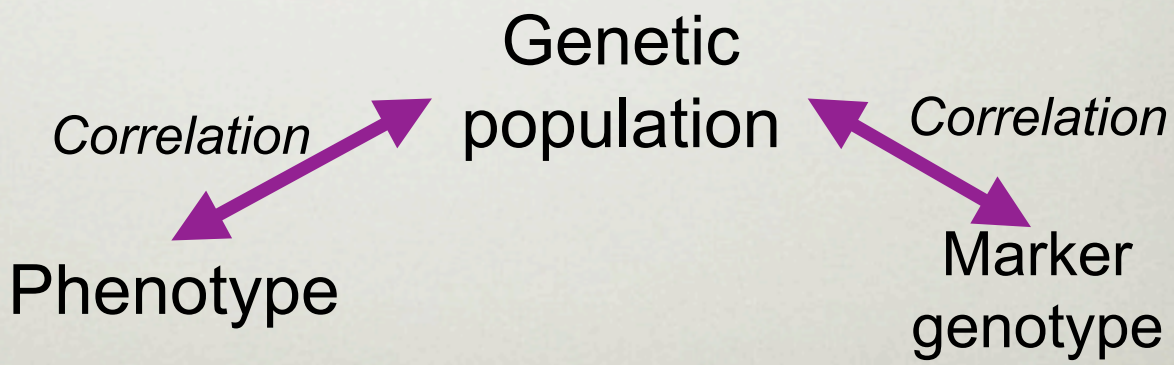
Stratifi-cation

Confounding in genetic studies

LD mapping

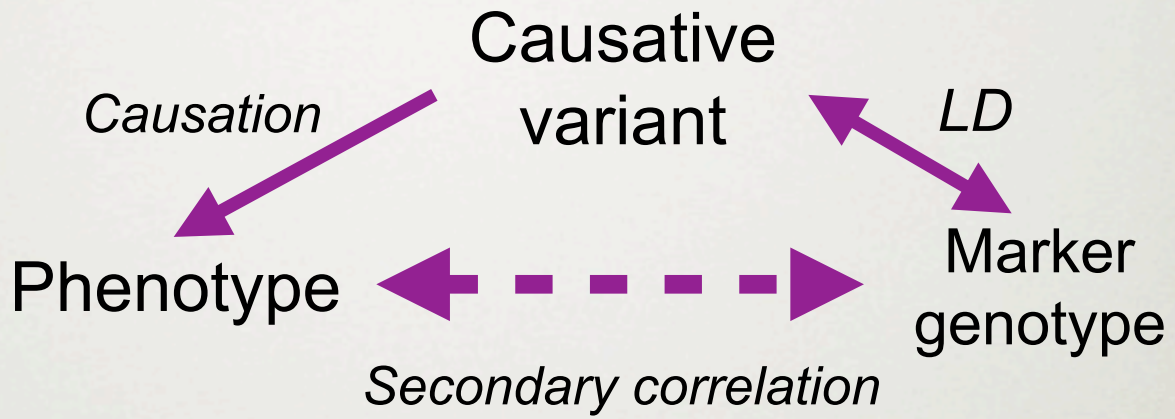


Stratification

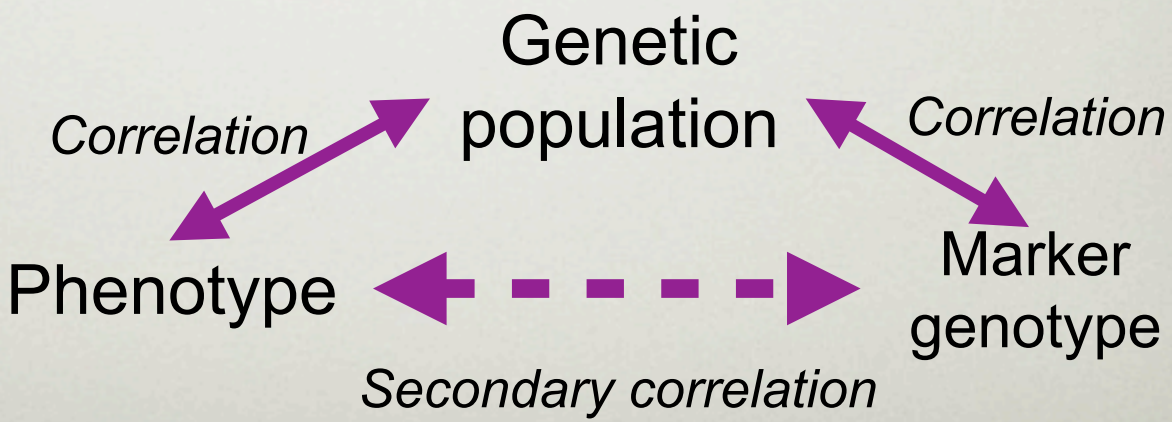


Confounding in genetic studies

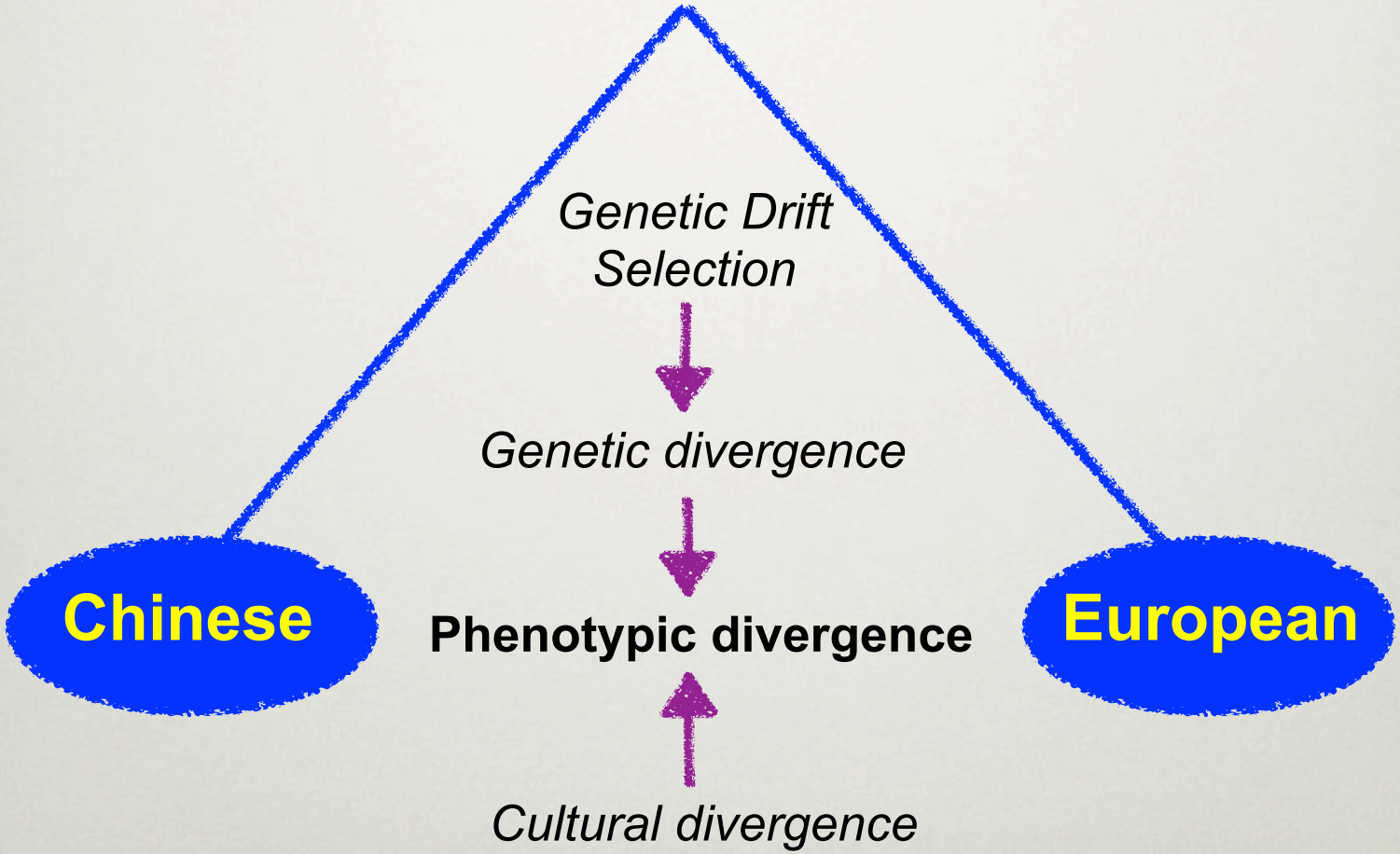
LD mapping



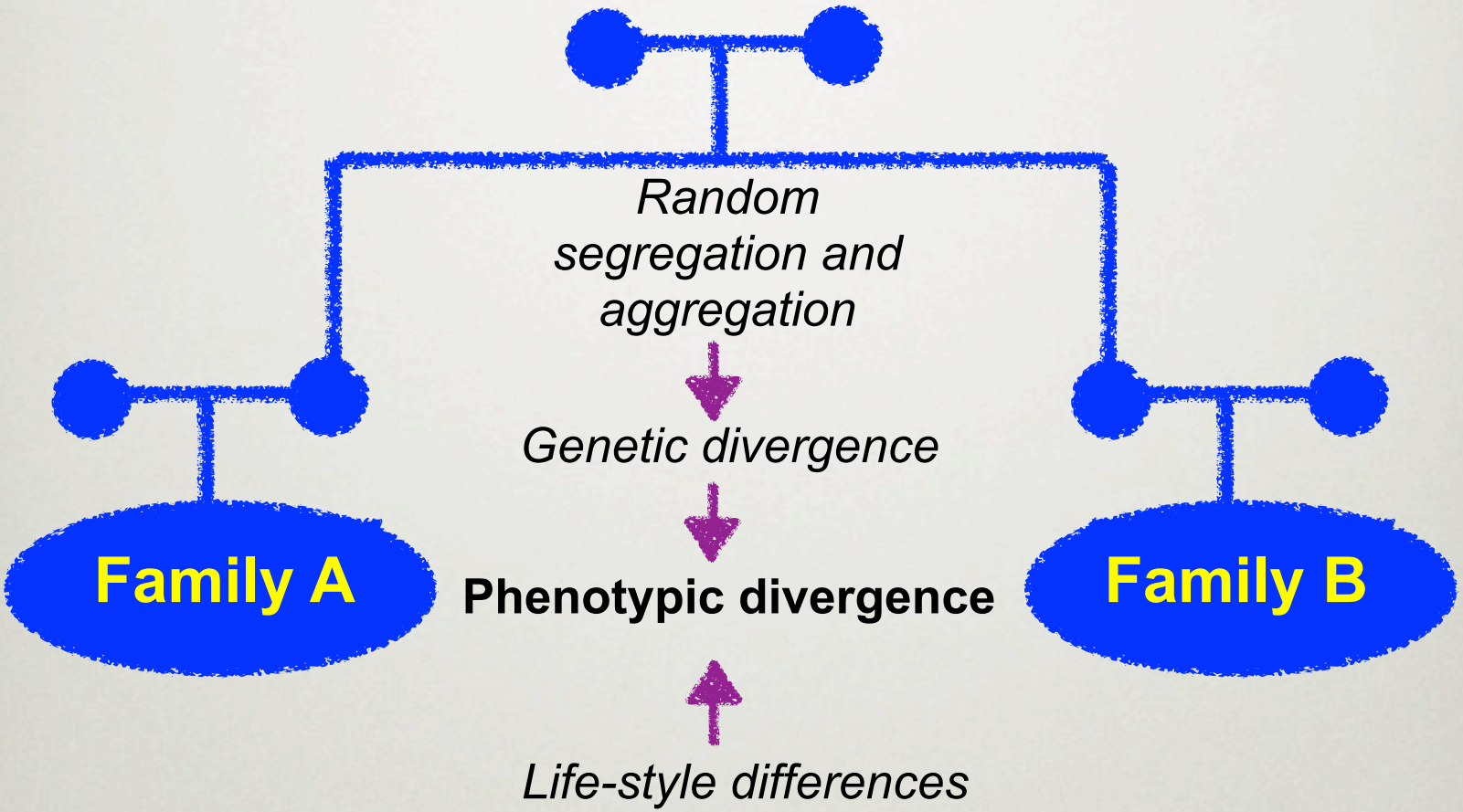
Stratification



Genetic origin is a major confounder



Pedigree is a major confounder



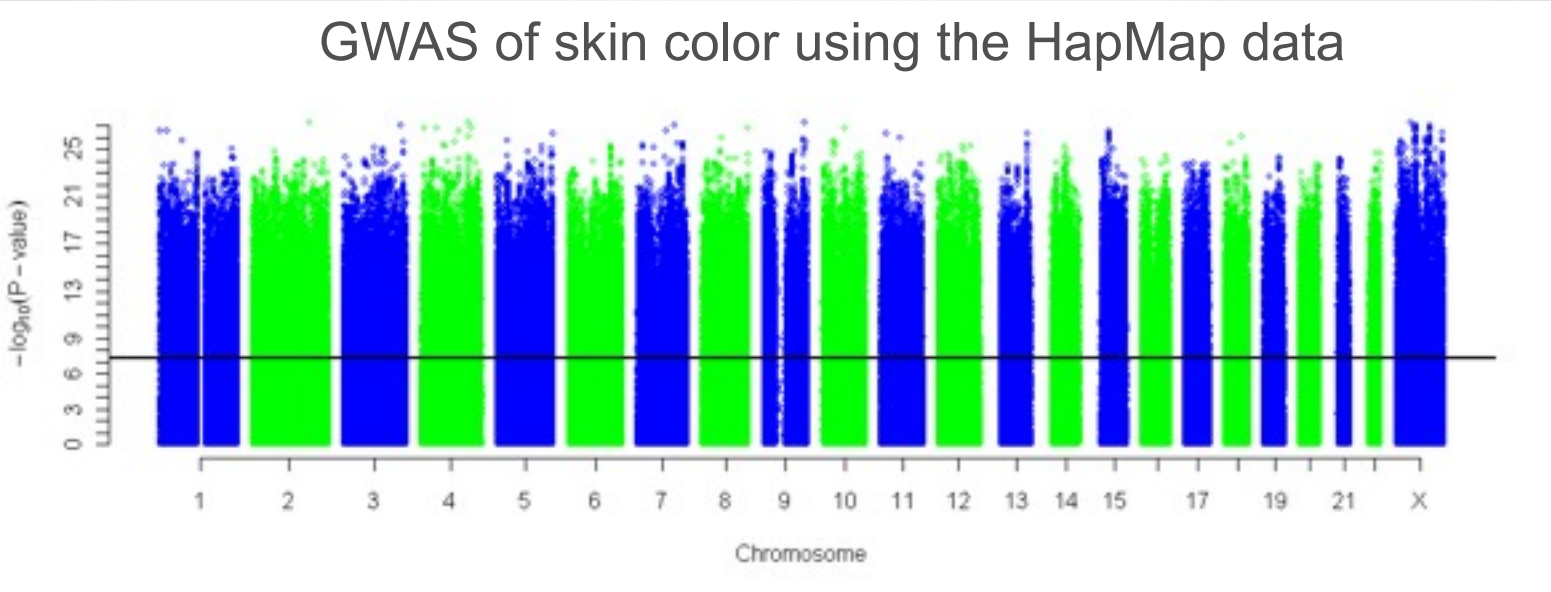
Confounding in GWAS

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

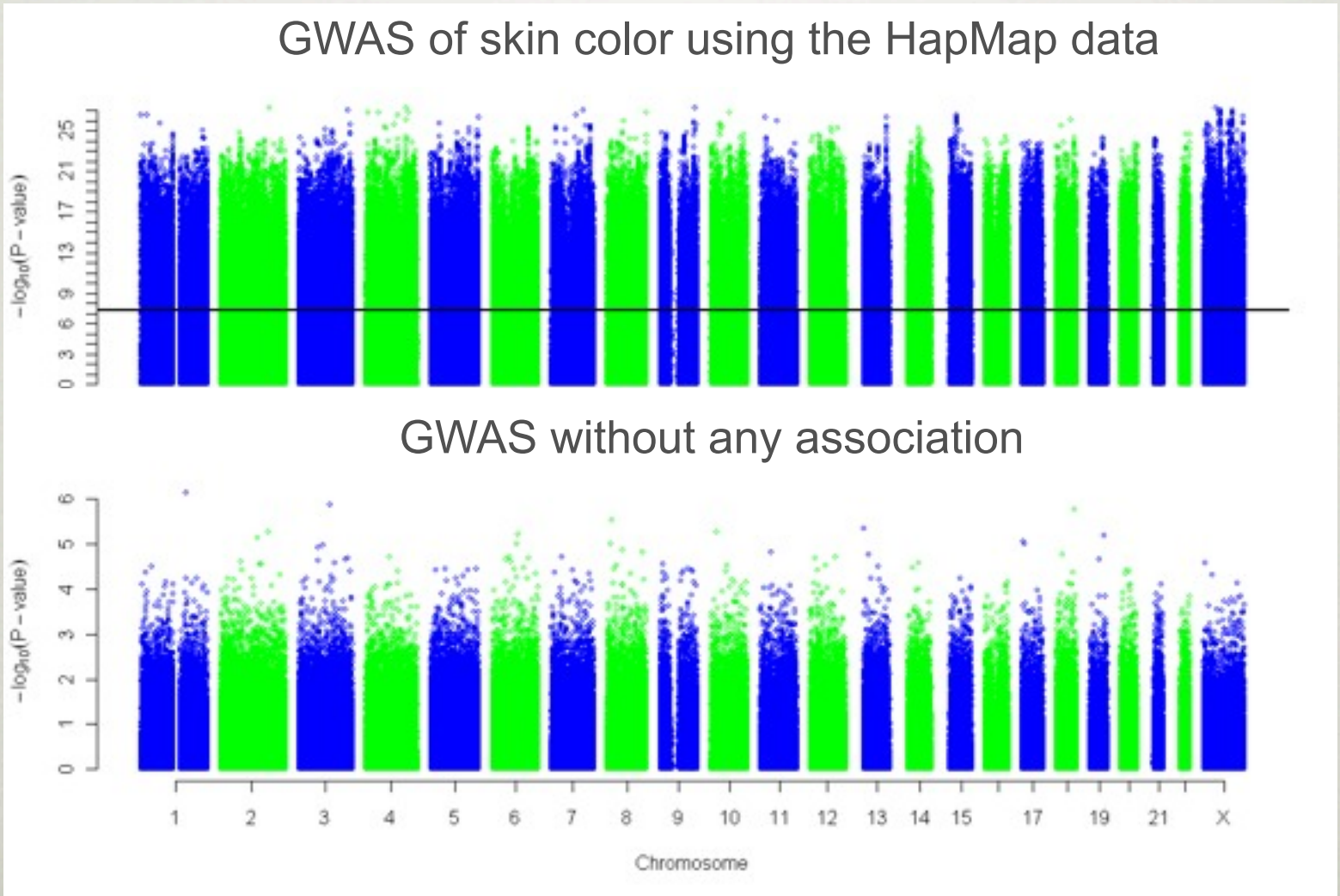
Other causes of genetic stratification are “cryptic” relations or systematic pedigree structure presented in a sample

Skin color scan

Skin color scan

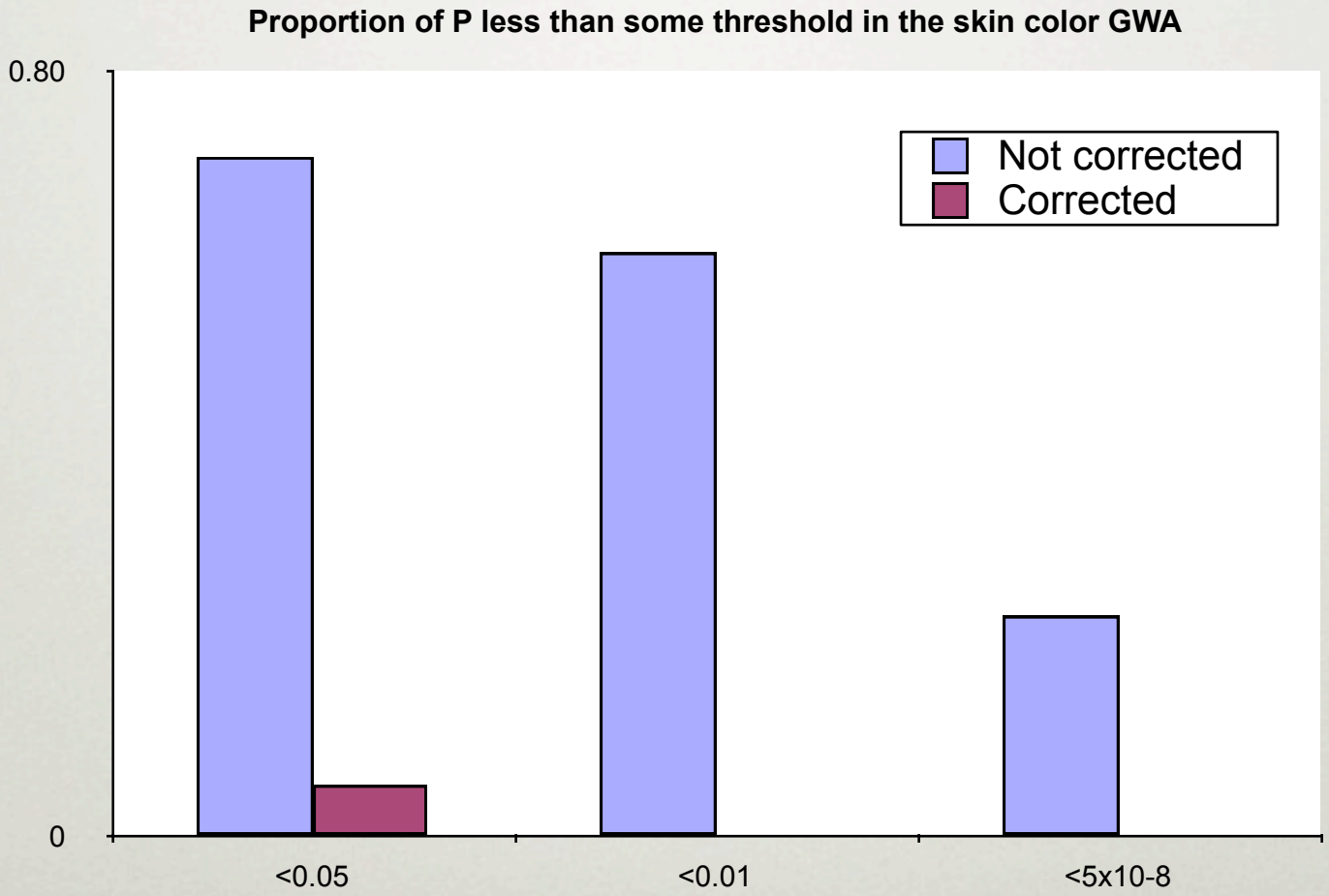


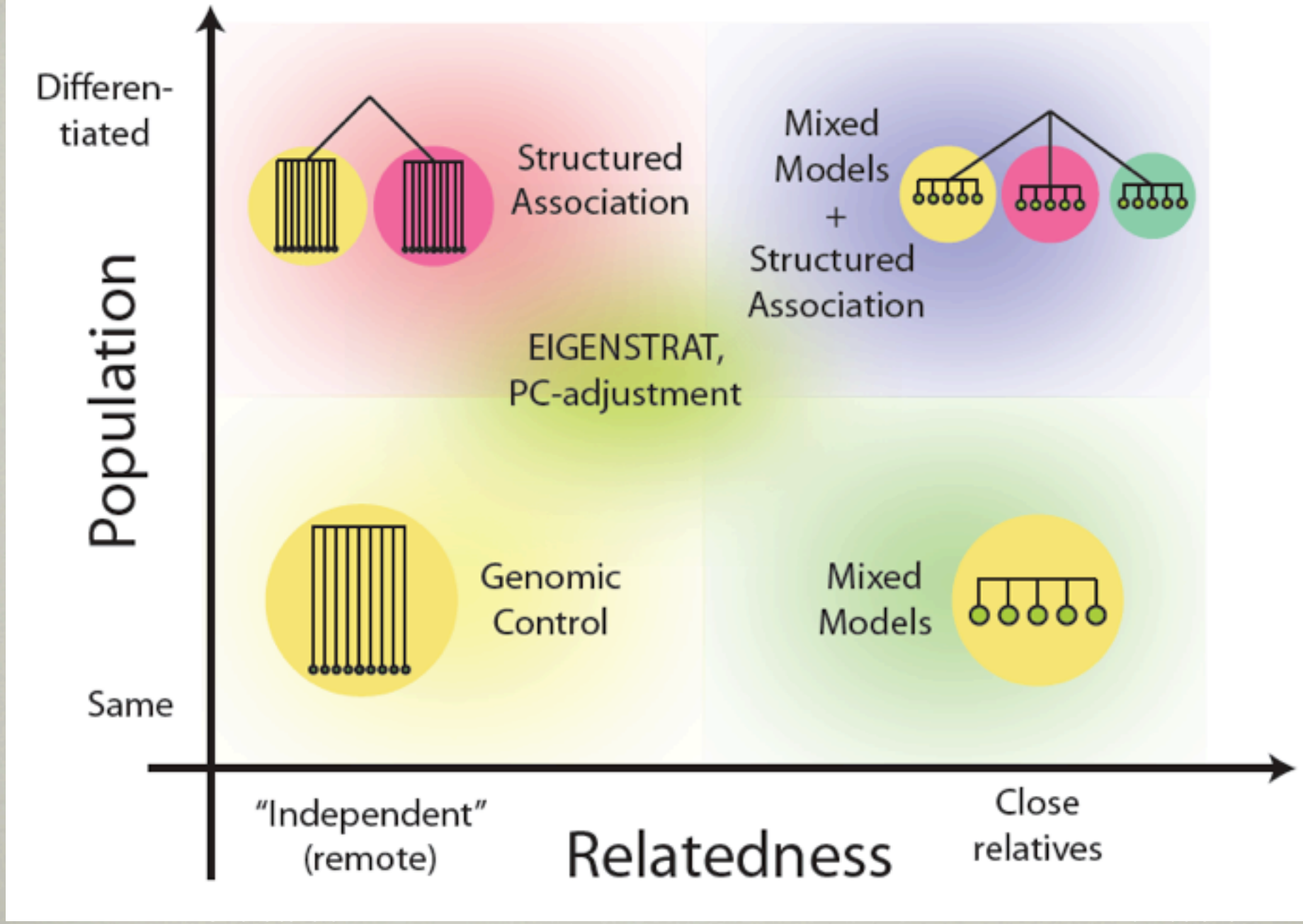
Skin color scan



Consequences of stratification

Consequences of stratification




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

- **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

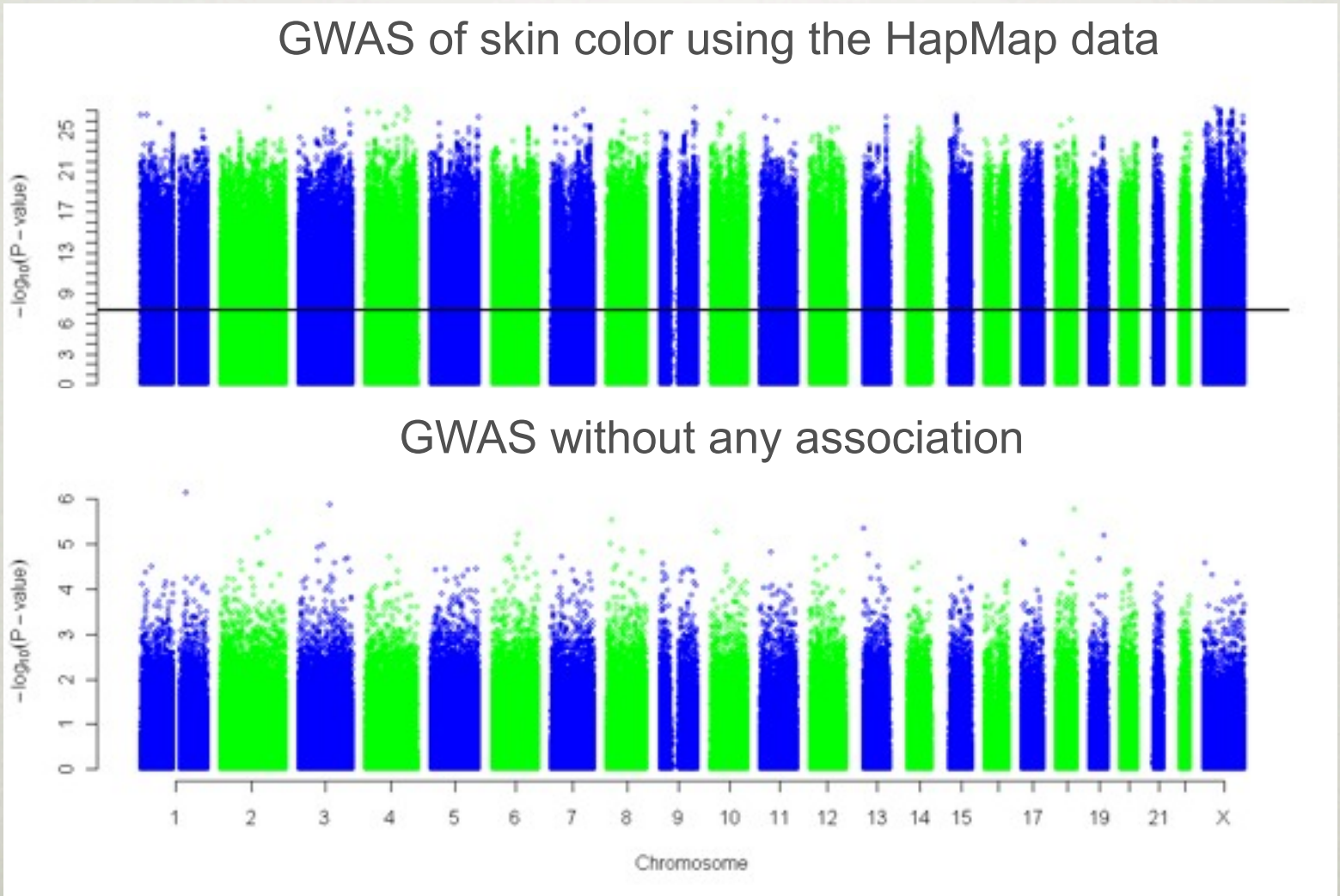
Genomic Control

Structured Association

Mixed Models

EigenSTRAT

Skin color scan



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 statistic $T^2/\lambda \sim \chi^2_1$
- In practice, all (or large proportion of) GW test are used to estimate λ

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. 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

Outline

Confounding in GWA studies

Genomic Control

Structured Association

Mixed Models

EigenSTRAT

How similar are genomes?

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

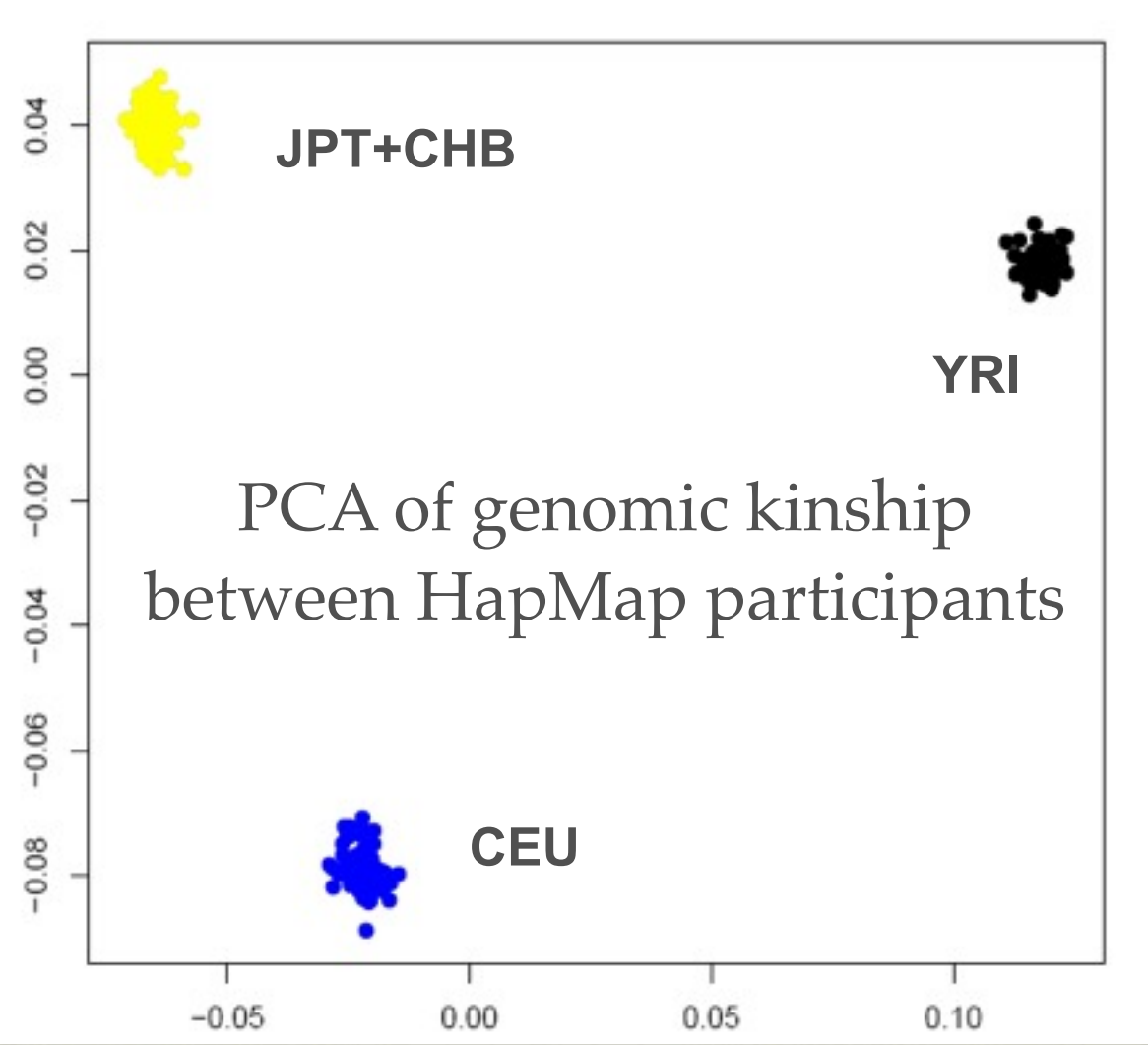
$$f_{ij} = \frac{1}{n}\sum_{k=1}^{n}\frac{(g_{ik} - p_k)(g_{jk} - p_k)}{p_k(1 - p_k)}$$

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



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 ($1 < \lambda < 1.05$)

Outline

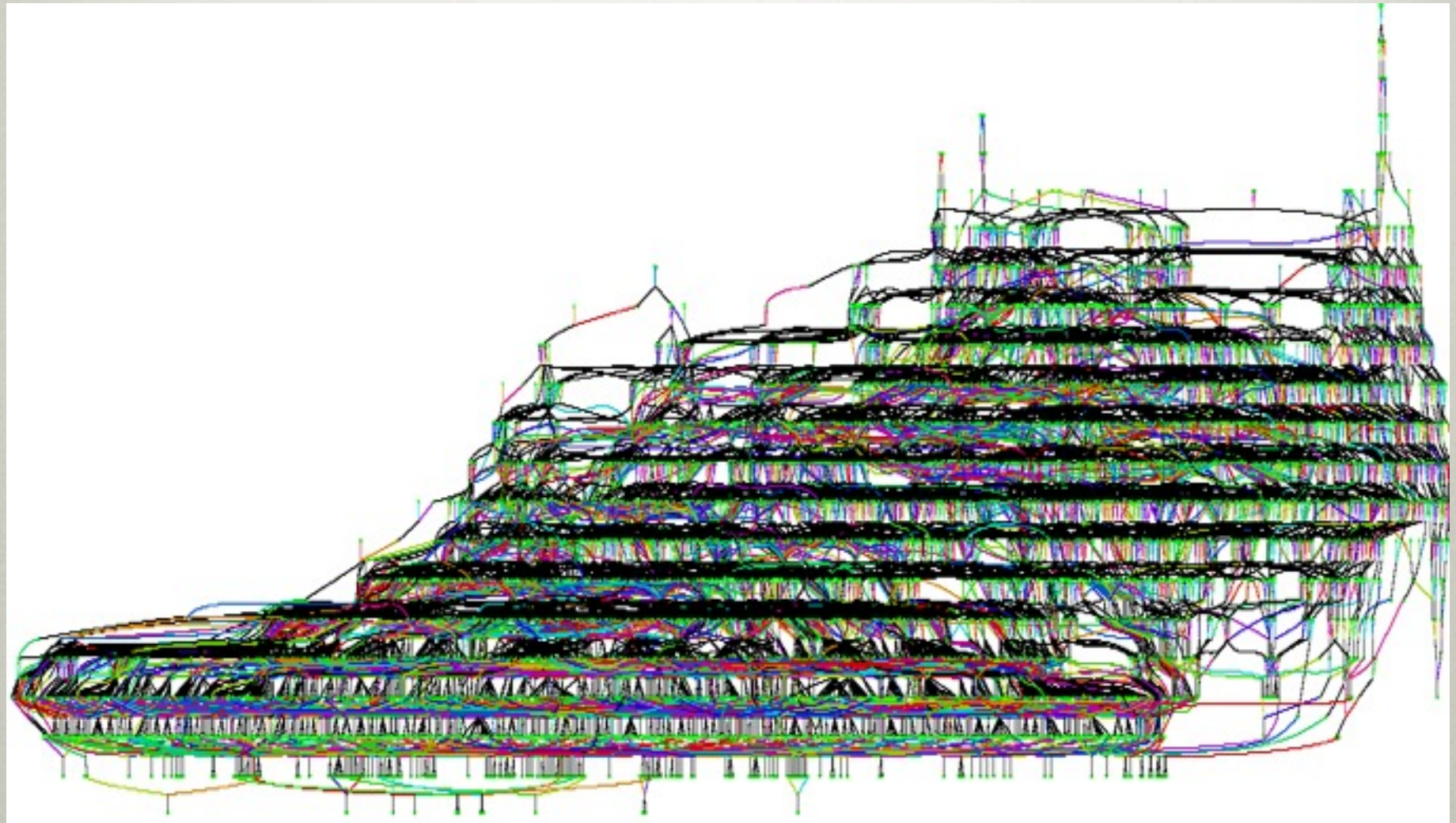
Confounding in GWA studies

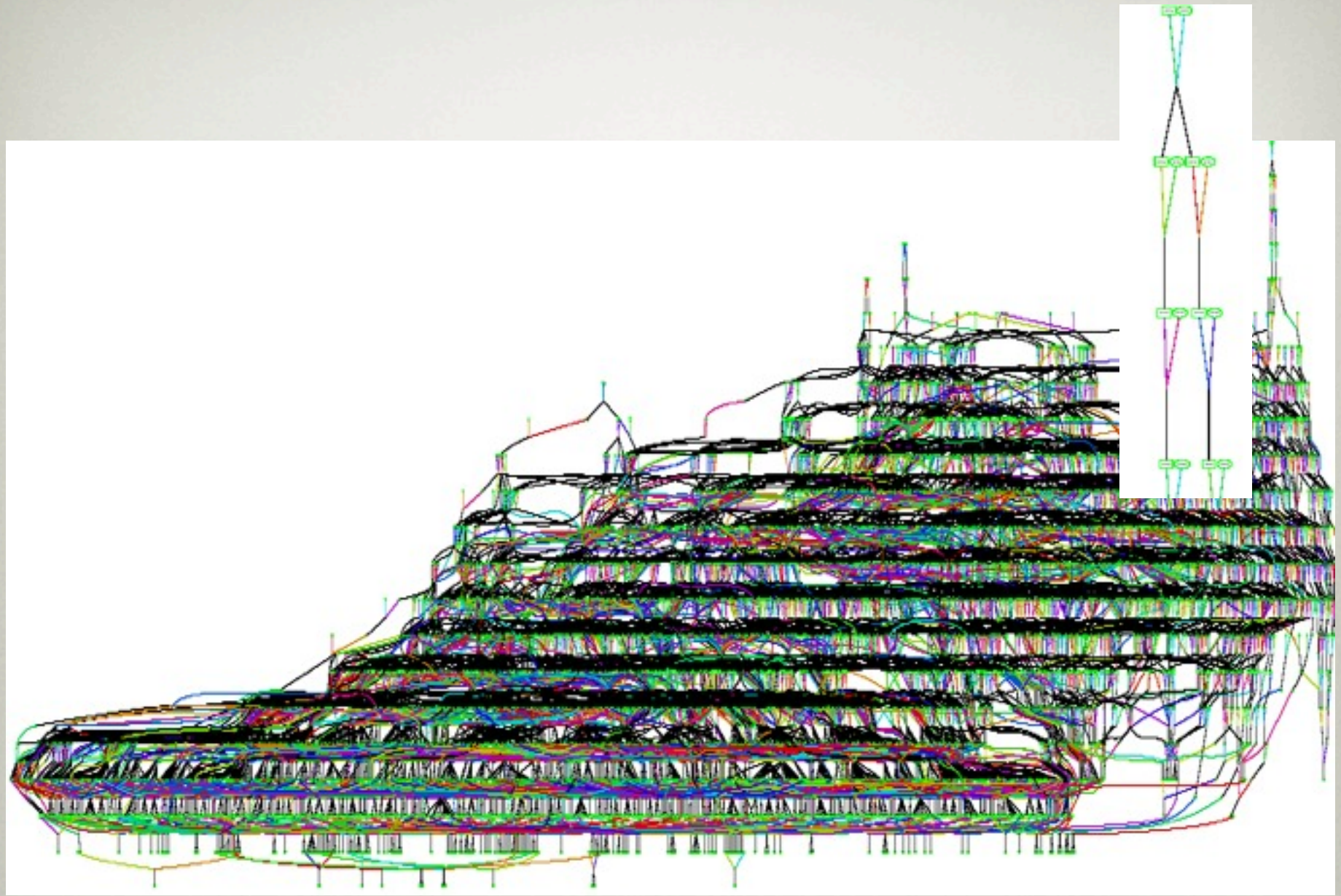
Genomic Control

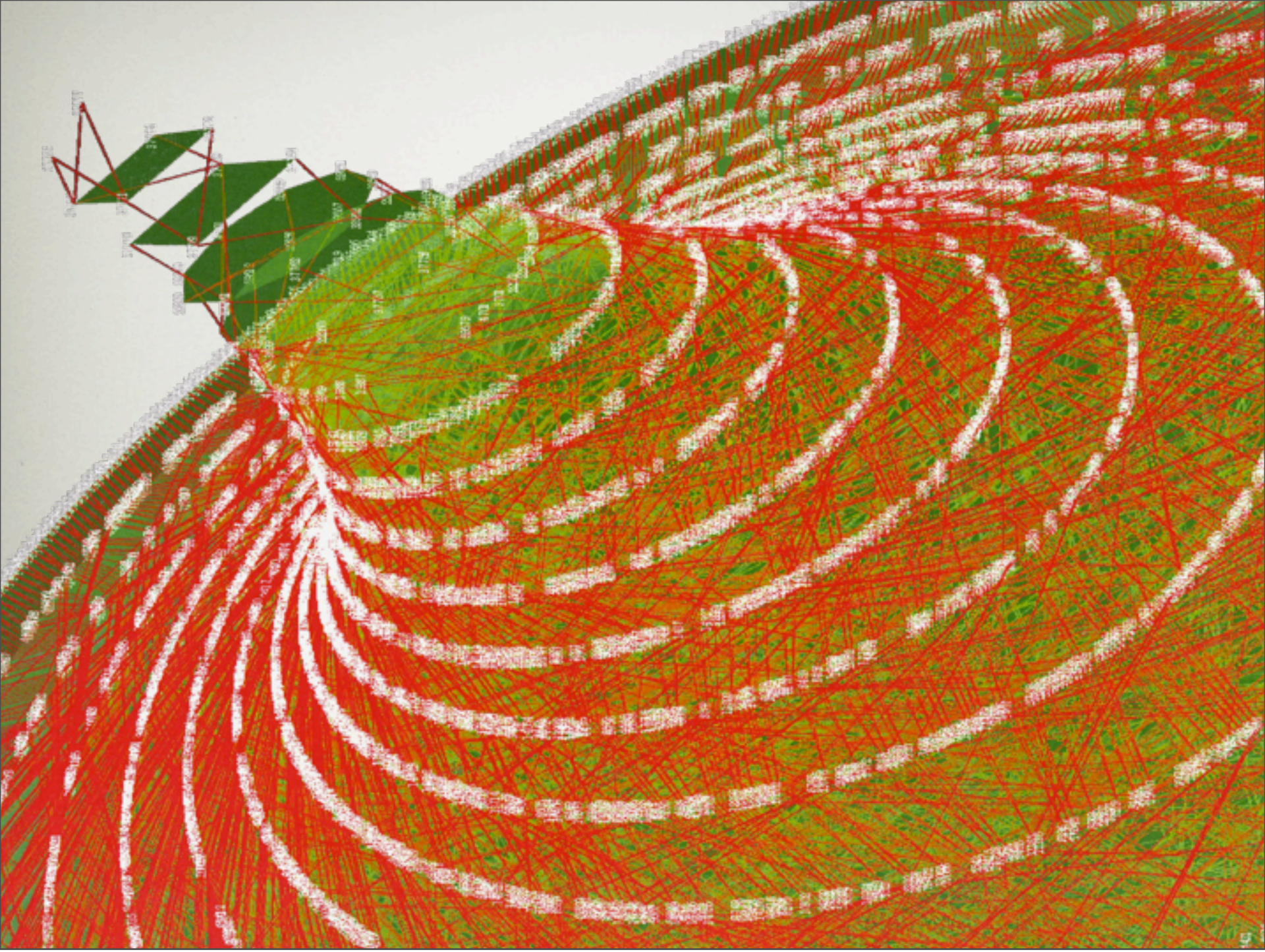
Structured Association

Mixed Models

EigenSTRAT







Mixed model

Vector of quantitative phenotype Y

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

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

β_g: additive affect of the allele

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

G: random polygenic effect ~ MVN(0, Φ σ_G^2)

Estimation of kinship from genomic data

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

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

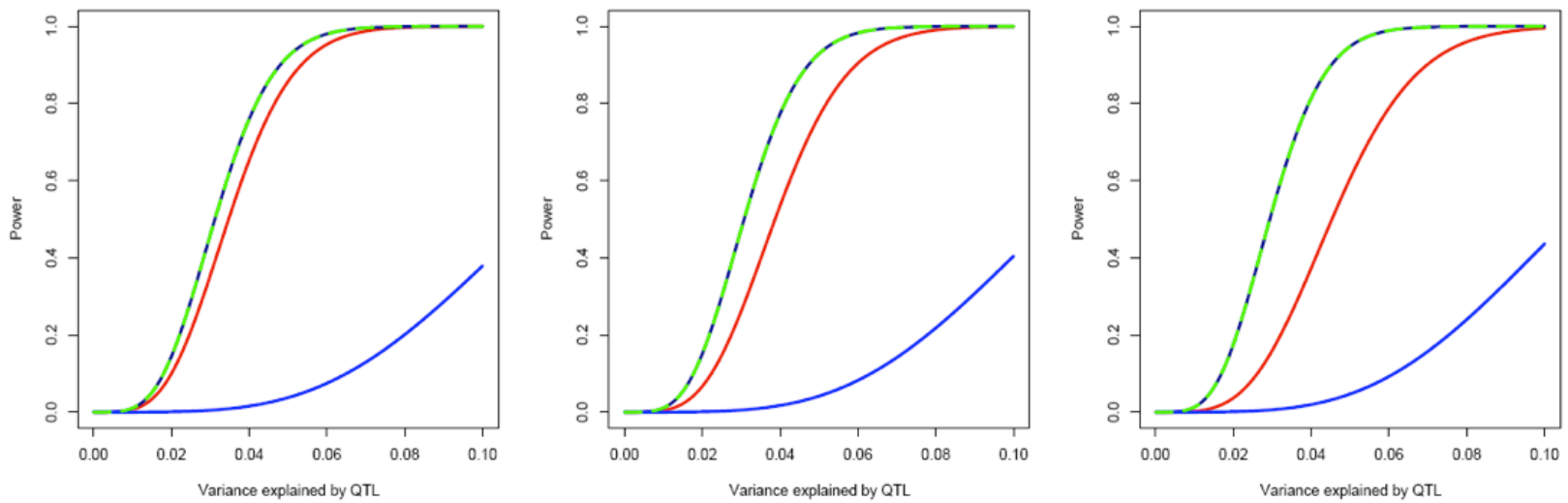
g_{ik} is the genotype (0, 0.5, 1) of the *i*-th person at *k*-th SNP

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

	Genomic control inflation factor			
Phenotype	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

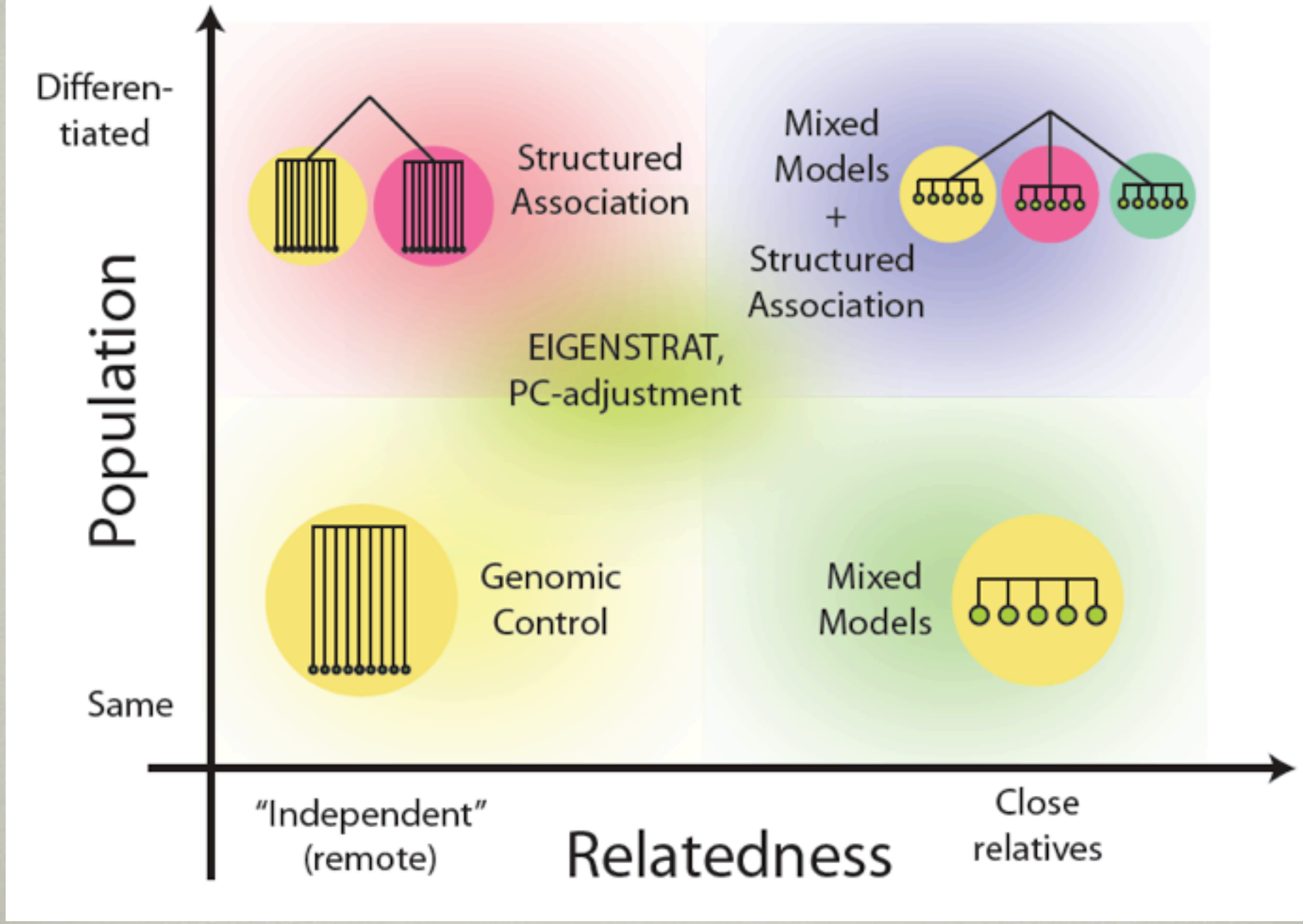
- Excellent method to account for complex genetic structure, such as found in special populations or in family-based studies

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


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: 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