Dealing with genetic (sub)structure in GWAS

Yurii Aulchenko
yurii [dot] aulchenko [at] gmail [dot] com

Genetic structure

- **A population has structure** when there are large-scale systematic differences in ancestry and/or groups of individuals with more, recent shared ancestors than one would expect in a randomly mating population
- **Shared ancestry corresponds to relatedness**, or kinship, so population structure can be described in terms of patterns of kinship among groups of individuals

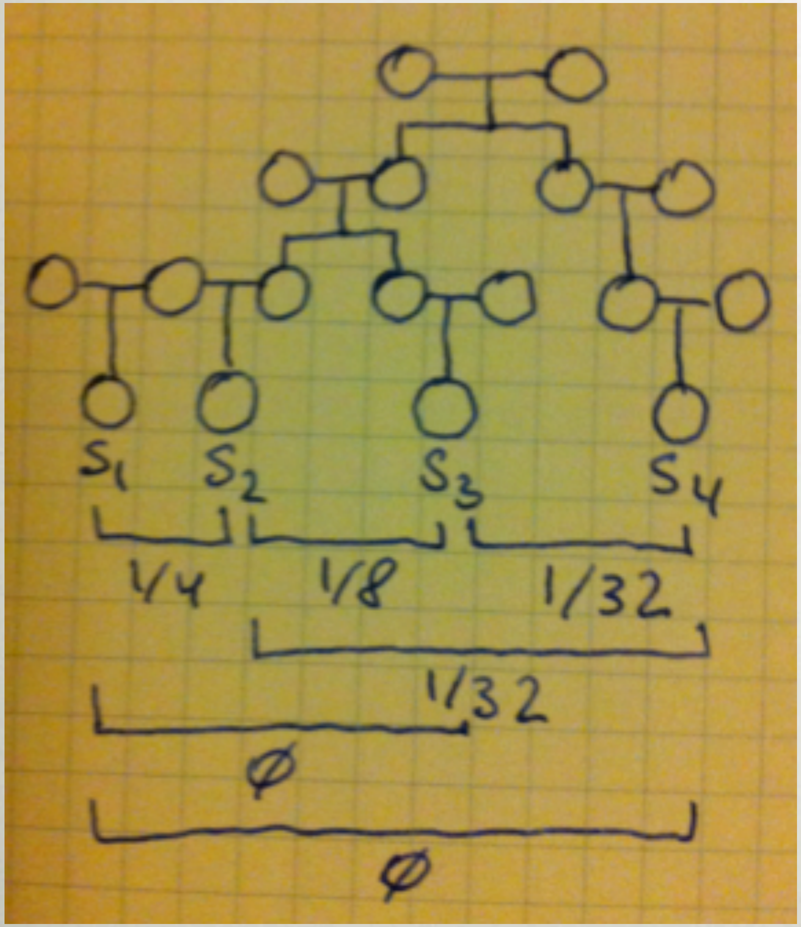
Measuring kinship

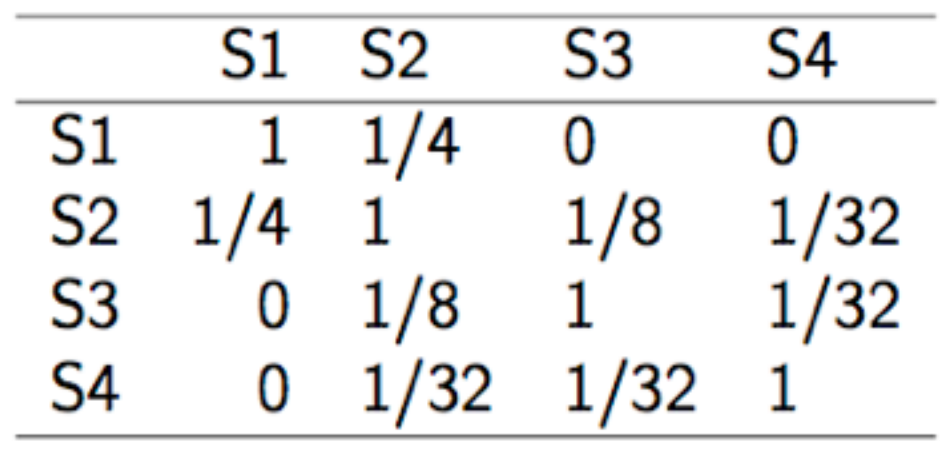
- Alleles that have descended from a single ancestral allele are said to be **identical by descent** (IBD)
- **Coefficient of kinship**, k_{ij}, between two individuals i and j is defined as the probability that two alleles sampled sampled at random from each individual are IBD
- For unrelated individuals, $k = 0$; in inbred lines, $k = 1$

Coefficient of relationship

- In outbred populations (no inbreeding), the **relationship coefficient** defined as $r_{ij}=2 \cdot k_{ij}$, has a simple interpretation as the expected proportion of genome i an j share IBD
- This coefficient is easily computed from pedigree information, e.g. $r = 1/2$ for parent-offspring and sib-pairs; $r = 1/4$ for half-sibs and grandparent-grandchild pairs

Example 1: pedigree





	S1	S2	S3	S4
S1	1	1/4	0	0
S2	1/4	1	1/8	1/32
S3	0	1/8	1	1/32
S4	0	1/32	1/32	1

No pedigree known

- The definition of kinship readily extends to any groups of individuals
- The problem is that we may not know the true underlying "pedigree"
- In case genomic data are available, we can estimate kinship from these

Genotypic correlation estimator of kinship

Kinship between *i* and *j* is computed with

$$\hat{K} = \frac{1}{L}\sum_{l=1}^{L}\frac{(x_l - 2p_l\mathbf{1})(x_l - 2p_l\mathbf{1})^T}{4p_l(1-p_l)}$$

where x_l is the column vector of genotypes (coded as count of "A" alleles) at *l*-th SNP and p_l is the frequency of the "A" allele

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

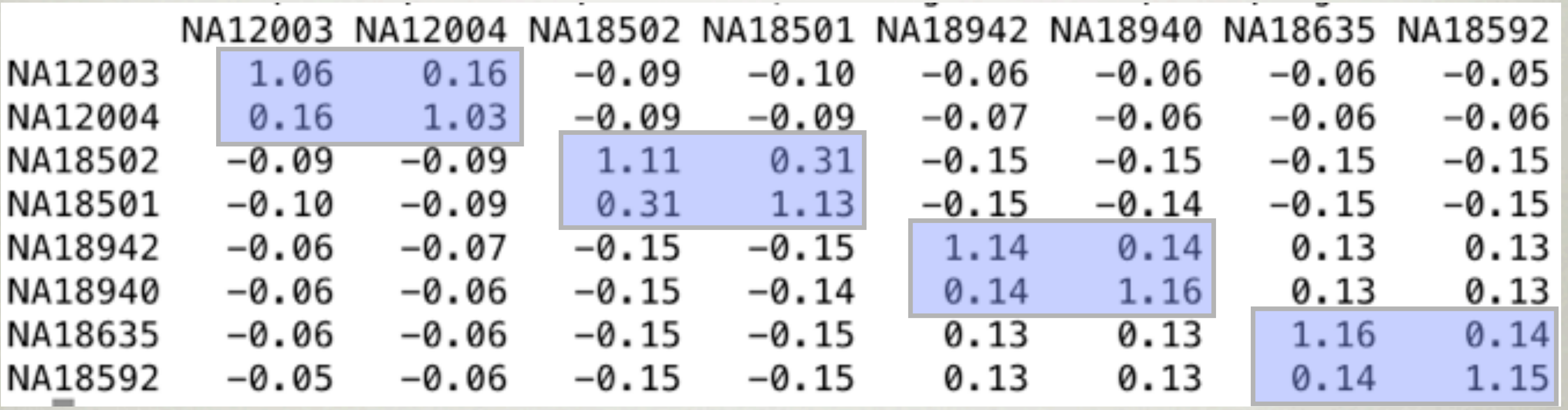
Correlation estimator

- The allele frequencies used are estimated from the sample, but the “true” ancestral allele frequencies *are not known*
- This leads to the fact that the estimates of kinship thus obtained can be negative
- Does not make sense in probability definition of kinship
- Does make sense in interpretation of kinship as an excess allele sharing

Genomic kinship for HapMap individuals

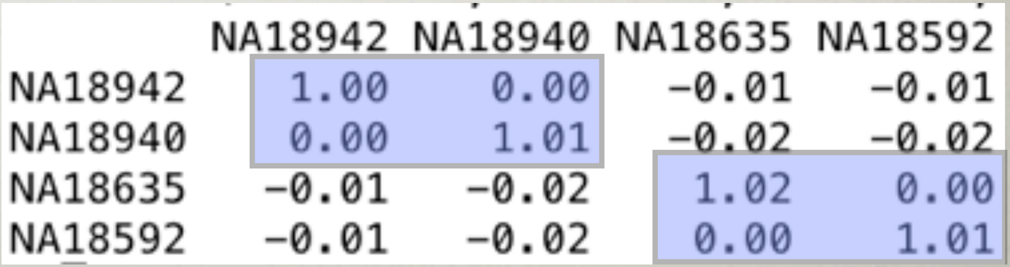
Using all data

CEU YRI JPT CHB



	NA12003	NA12004	NA18502	NA18501	NA18942	NA18940	NA18635	NA18592
NA12003	1.06	0.16	-0.09	-0.10	-0.06	-0.06	-0.06	-0.05
NA12004	0.16	1.03	-0.09	-0.09	-0.07	-0.06	-0.06	-0.06
NA18502	-0.09	-0.09	1.11	0.31	-0.15	-0.15	-0.15	-0.15
NA18501	-0.10	-0.09	0.31	1.13	-0.15	-0.14	-0.15	-0.15
NA18942	-0.06	-0.07	-0.15	-0.15	1.14	0.14	0.13	0.13
NA18940	-0.06	-0.06	-0.15	-0.14	0.14	1.16	0.13	0.13
NA18635	-0.06	-0.06	-0.15	-0.15	0.13	0.13	1.16	0.14
NA18592	-0.05	-0.06	-0.15	-0.15	0.13	0.13	0.14	1.15

Using only JPT+CHB data:



	NA18942	NA18940	NA18635	NA18592
NA18942	1.00	0.00	-0.01	-0.01
NA18940	0.00	1.01	-0.02	-0.02
NA18635	-0.01	-0.02	1.02	0.00
NA18592	-0.01	-0.02	0.00	1.01

IBS estimator of kinship

Kinship between *i* and *j* is computed with

$$\frac{1}{2L}\sum_{l=1}^{L}(x_l - \mathbf{1})(x_l - \mathbf{1})^T + \frac{1}{2}.$$

where x_l is the column vector of genotypes (coded as count of "A" alleles) at *l*-th SNP

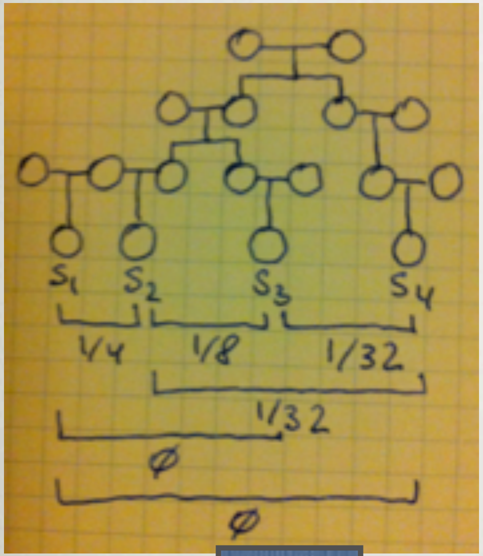
If IBS implies IBD, this is kinship estimator

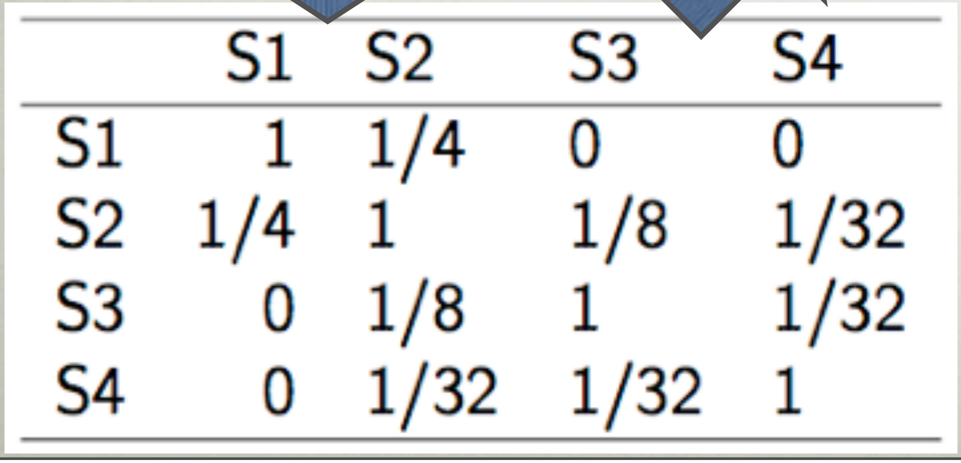
Usually less precise than the correlation estimator

Classical multi-dimensional scaling

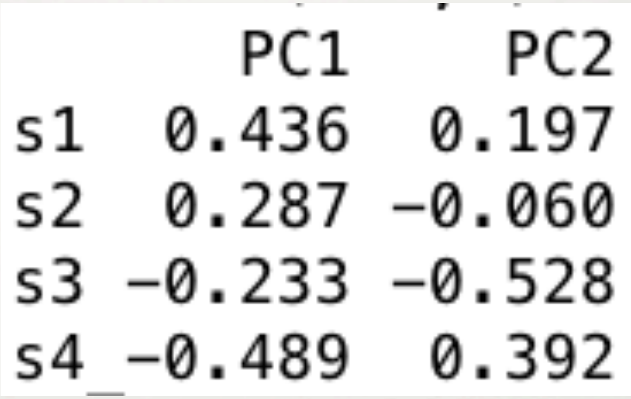
- Given pair-wise distance matrix for a set of entities finds out their coordinates in an *t*-dimensional space so that the distances in this space are as close as possible to the original distances
- Kinship *K* measures "closeness", so CMDS is applied to (0.5-*K*)

CMDS of the pedigree

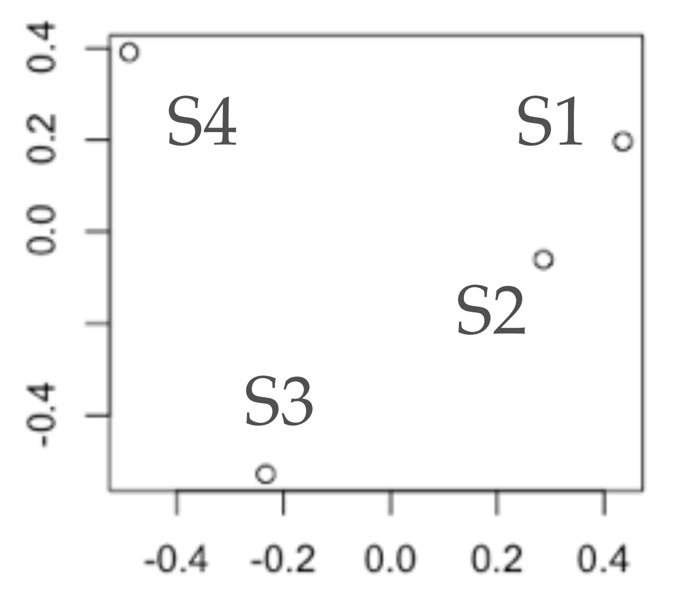




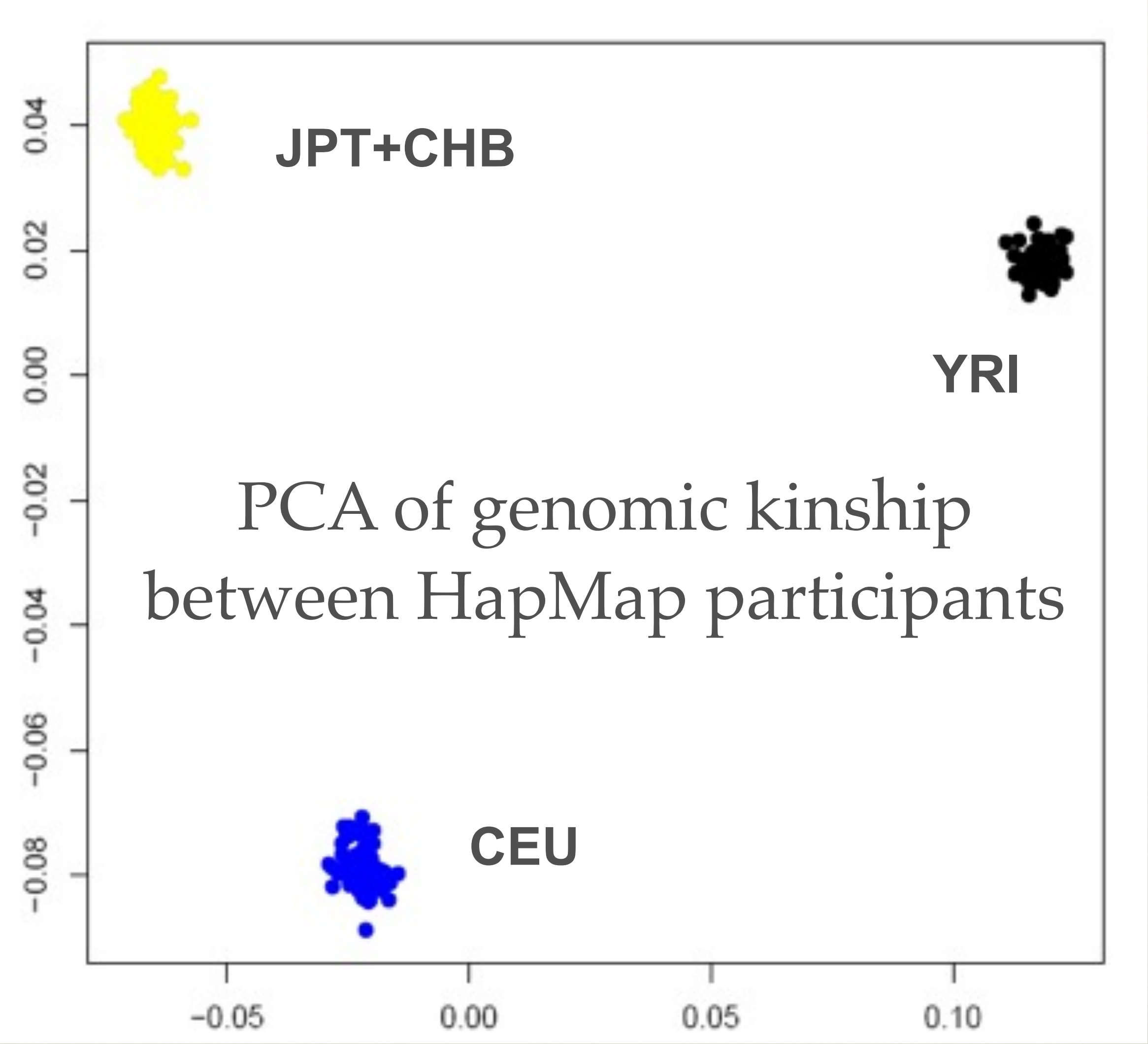
	S1	S2	S3	S4
S1	1	1/4	0	0
S2	1/4	1	1/8	1/32
S3	0	1/8	1	1/32
S4	0	1/32	1/32	1



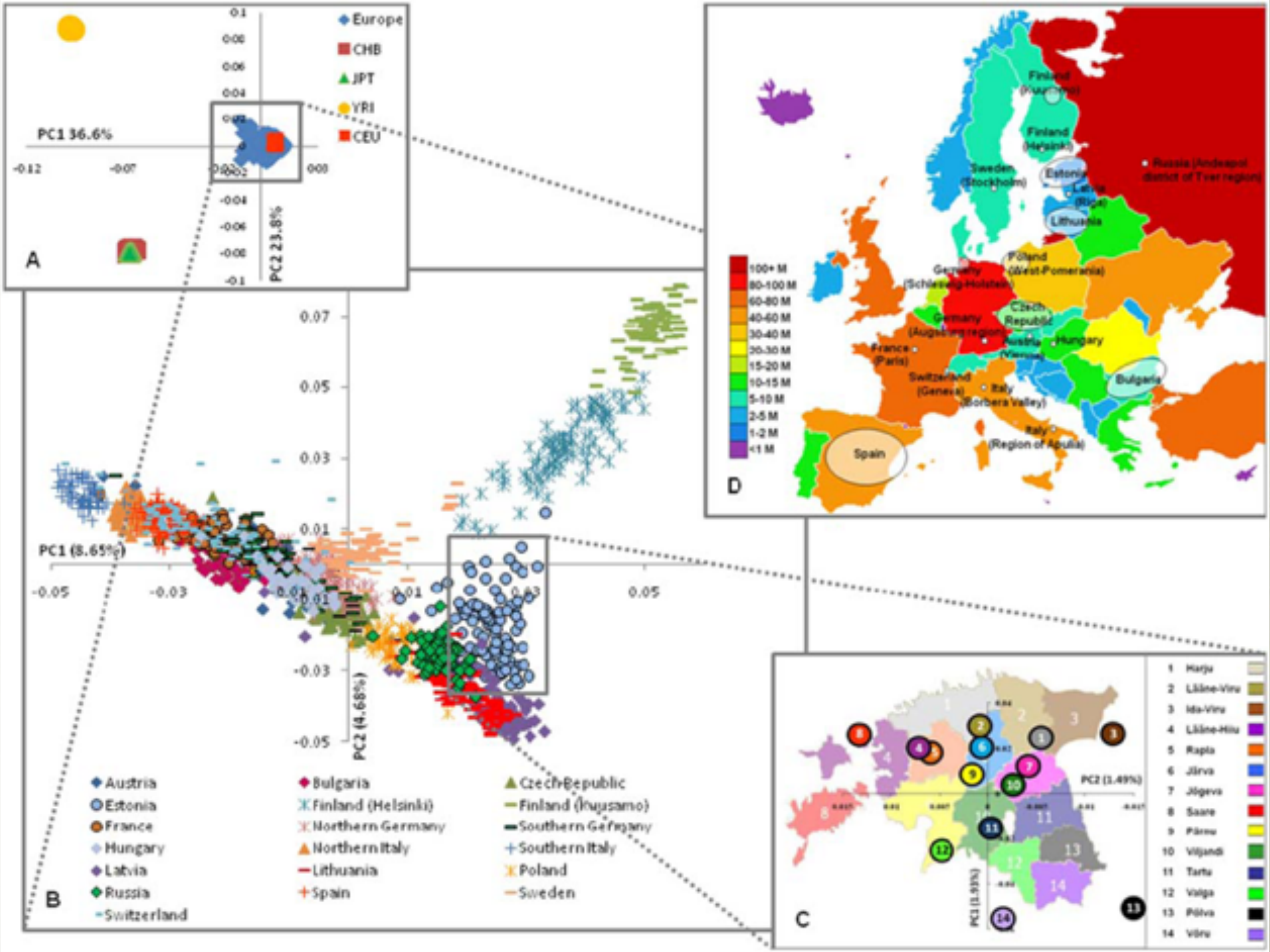
```
       PC1    PC2
s1  0.436  0.197
s2  0.287 -0.060
s3 -0.233 -0.528
s4 -0.489  0.392
```



CMDS of HapMap data



PCA of genomic kinship between HapMap participants



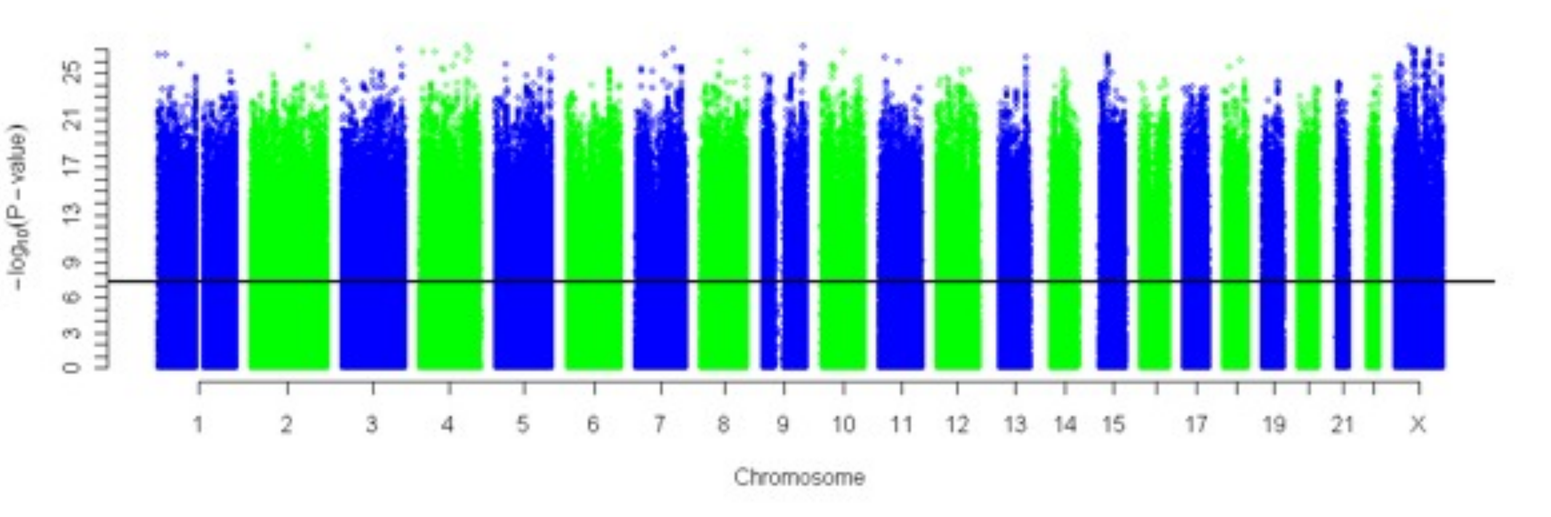
Nelis et al., PLoS ONE, 2009

GWAS: why do we bother about structure?

Yurii Aulchenko

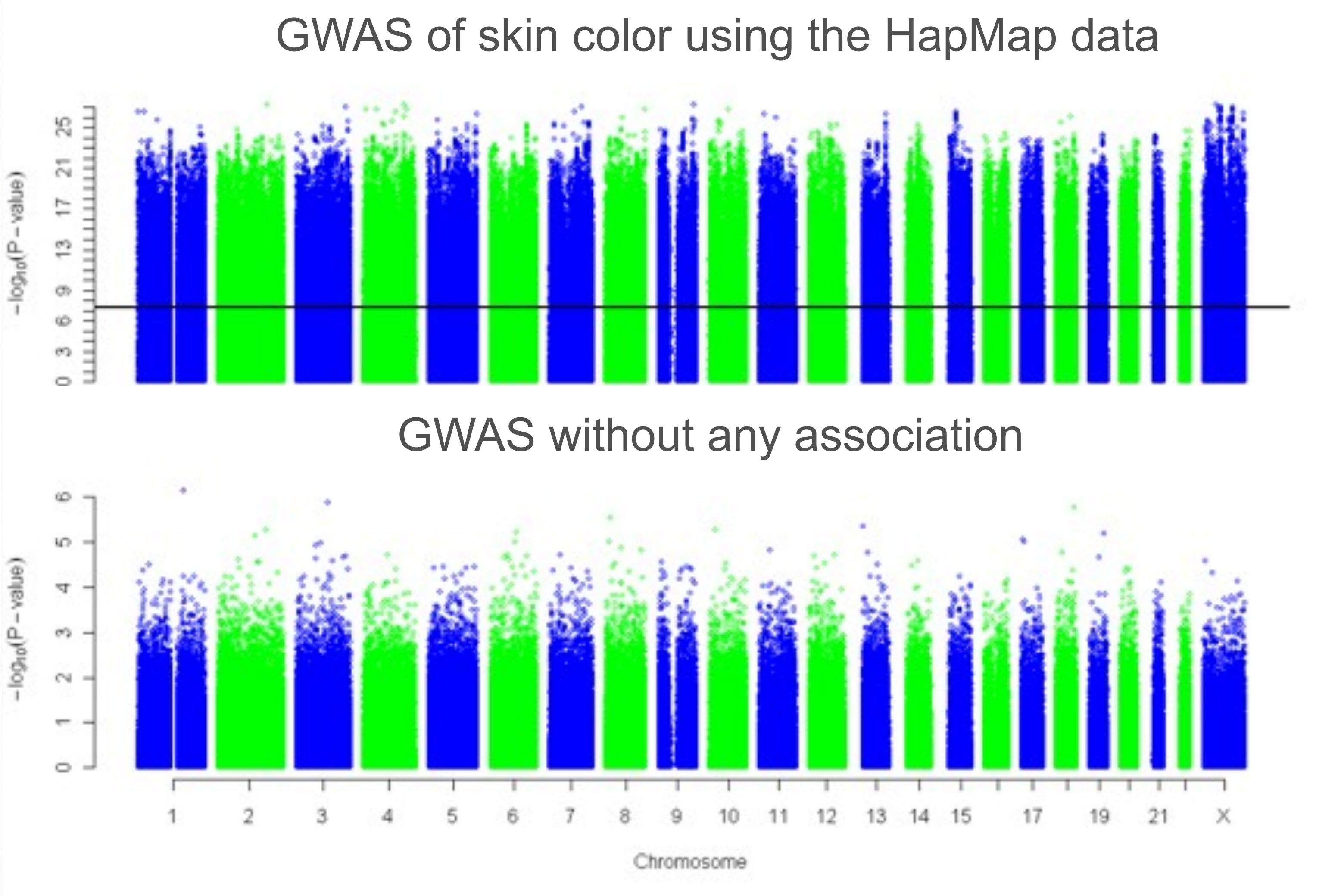
GWAS: why do we bother about structure?

GWAS of skin color using the HapMap data



Yurii Aulchenko

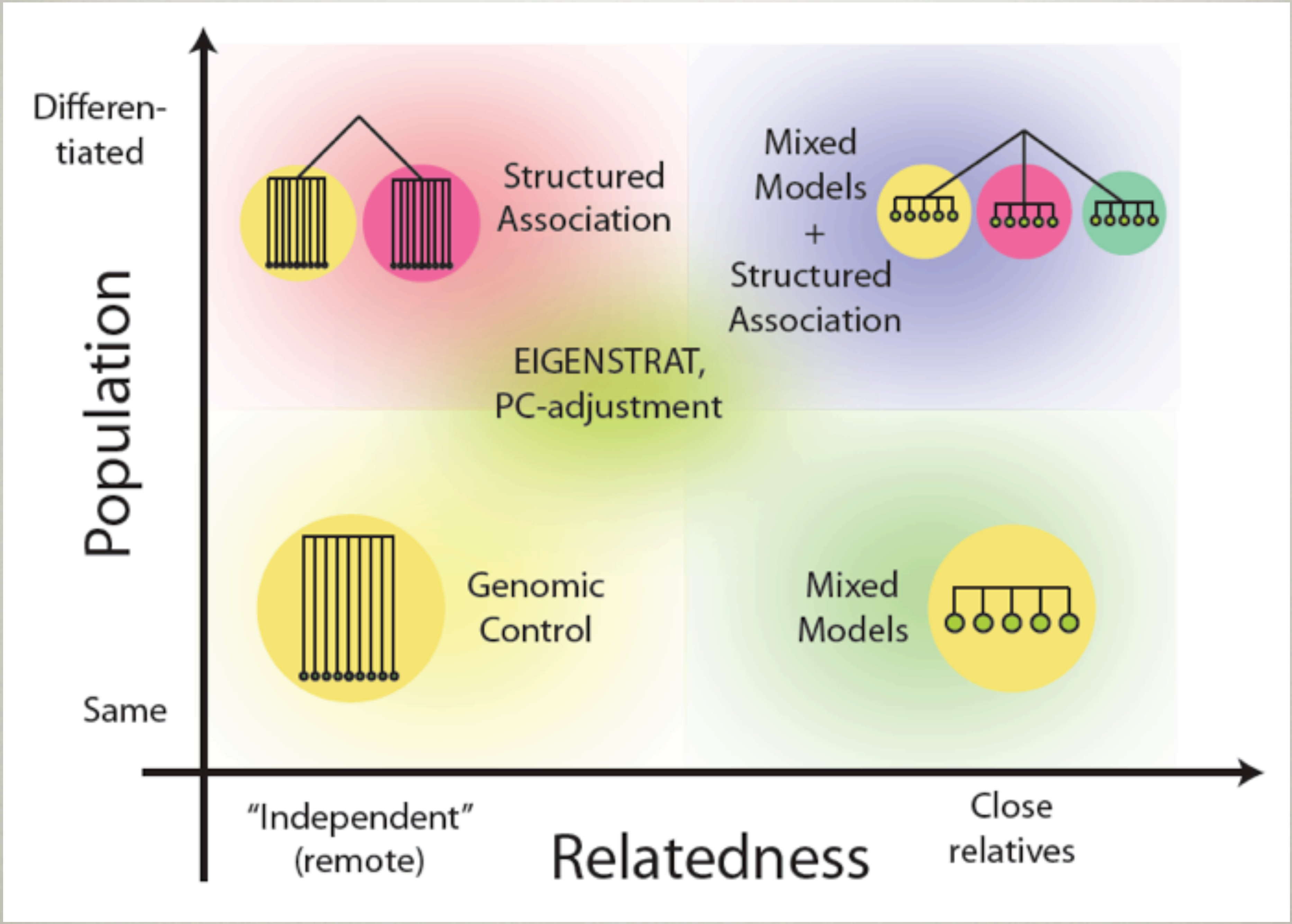
GWAS: why do we bother about structure?



GWAS of skin color using the HapMap data

GWAS without any association

Yurii Aulchenko


Differen-
tiated
Same
Population
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

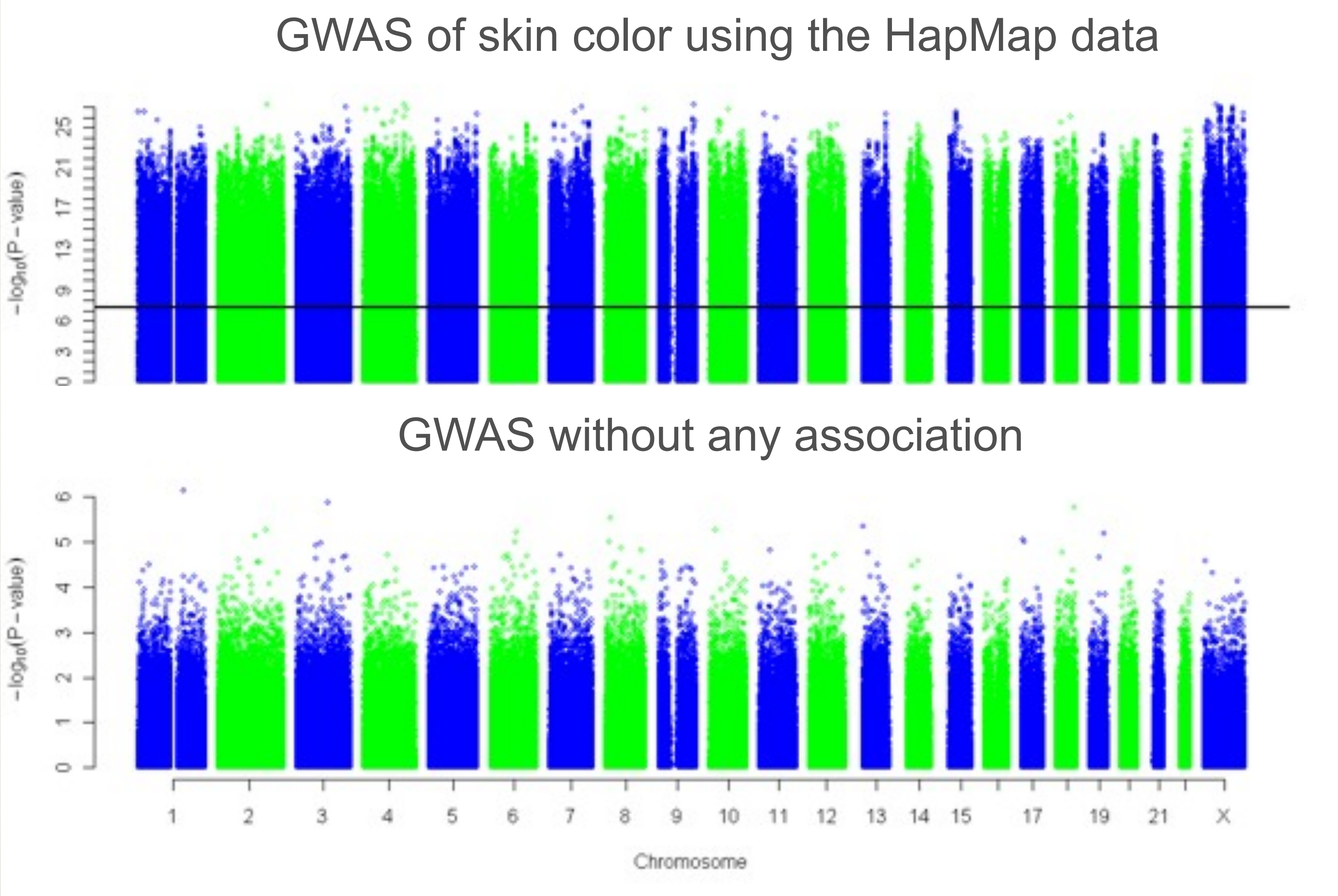
Genomic Control

Structured Association

EigenSTRAT

Mixed Models

Skin color scan



GWAS of skin color using the HapMap data

GWAS without any association

Genomic control

- If a test statistic is distributed as χ^2_1 under the null hypothesis of no association, it has been demonstrated that under stratification, the test statistic is distributed as χ^2_1 up to some scaling constant λ
- Estimate λ from the vector of test statistics $\{T^2_1, T^2_2, T^2_3, \ldots, T^2_{N-1}, T^2_N\}$ obtained from GWAS
- The GC-corrected test statistic T^2/λ is distributed as χ^2_1

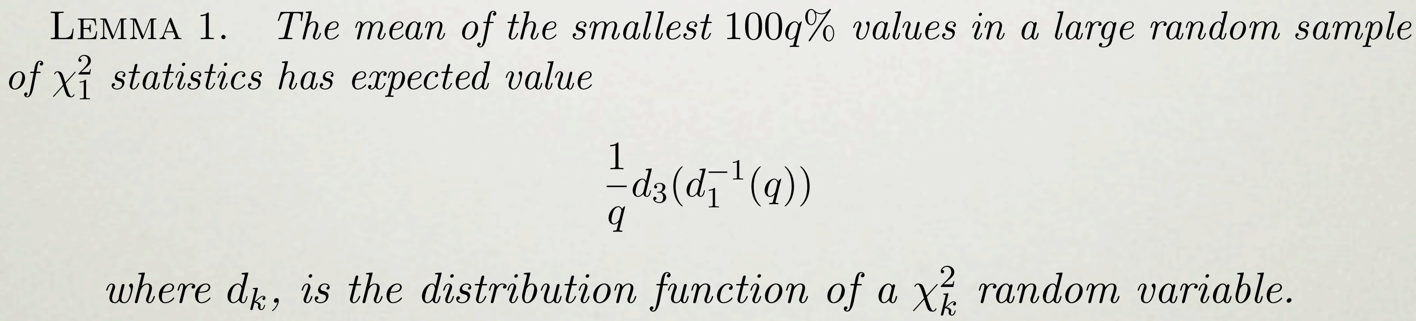
Estimators of λ

- Mean estimator: mean(T^2)
- Median estimator: median(T^2)/0.455
- Regression estimator: slope of regression of observed T^2 on the expected

- Mean is more effective than median *under the null*
- ... but there is a little problem

Trimmed mean estimator

- The idea is to remove the highest test values from consideration, and use the mean estimator then
- Following Astle and Balding (2009)



Lemma 1. *The mean of the smallest* $100q\%$ *values in a large random sample of* χ^2_1 *statistics has expected value*

$$\frac{1}{q} d_3(d_1^{-1}(q))$$

where d_k, *is the distribution function of a* χ^2_k *random variable.*

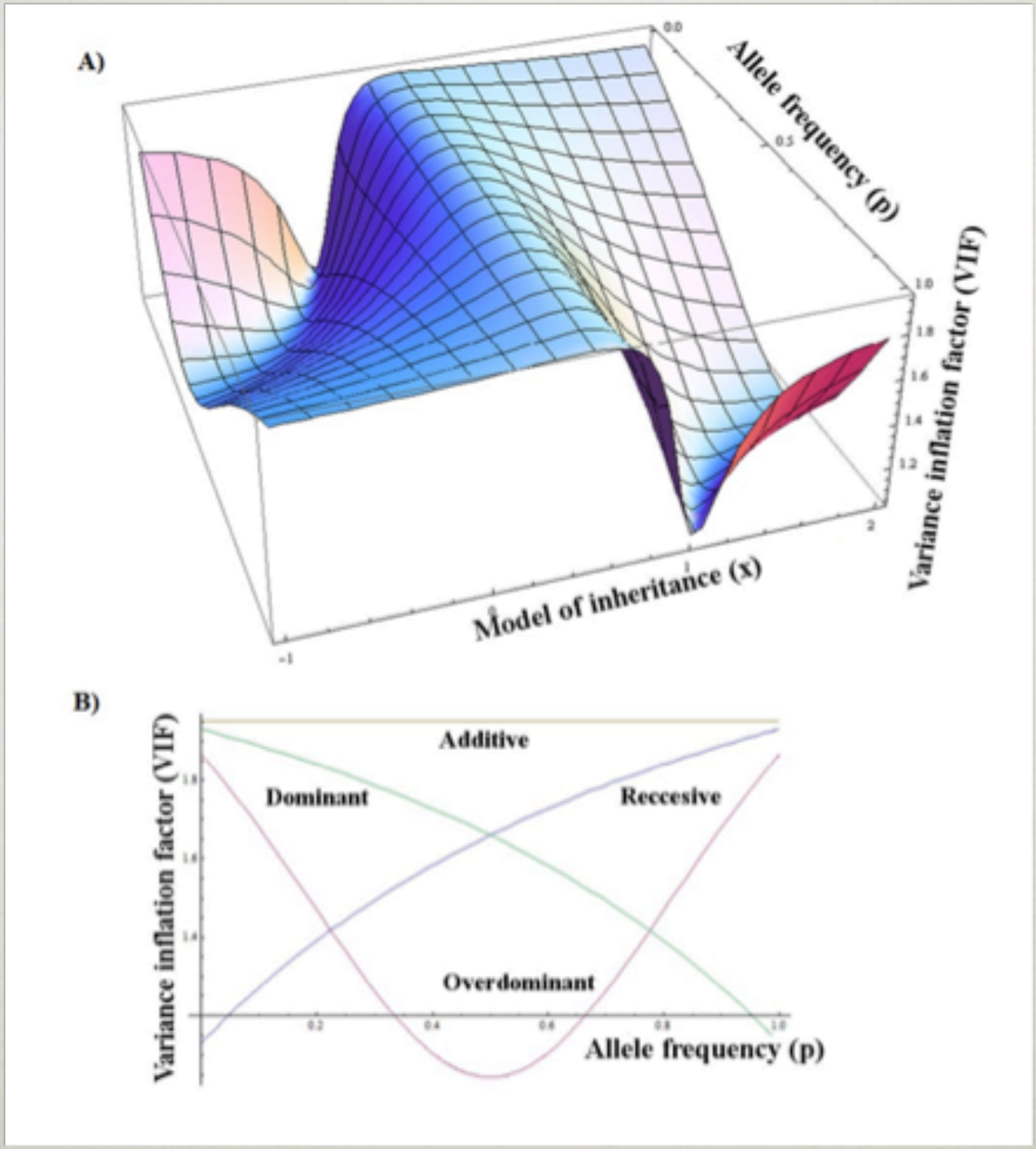
Estimate(λ) = mean(lower 95% of T^2)/0.759

Two uses of the GC

- GC is a method to *correct the test statistic,* and hence have interpretable p-values
- What may be even more important - deviation of λ from 1 tells that something went wrong with the analysis
- For example, high values ($\lambda > 1.05$) *is an indicator* that the analysis model failed to account for the sample structure, and other model should be used

Few notes on GC

- GC assumes that stratification acts in the same manner across all loci, which is not always true
- Inflation factor λ depends on samples size. Special methods should be used when number of people typed for different SNPs is different
- In present form, GC *works only for additive model*


A)
Allele frequency (p)
0.0
0.5
Variance inflation factor (VIF)
1.0
1.8
1.6
1.4
1.2
Model of inheritance (x)
-1
0
1
2
B)
Variance inflation factor (VIF)
1.8
1.6
1.4
Additive
Dominant
Reccesive
Overdominant
0.2
0.4
0.6
0.8
1.0
Allele frequency (p)

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; stratified score test (GenABEL::qtscore with 'strata'); 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

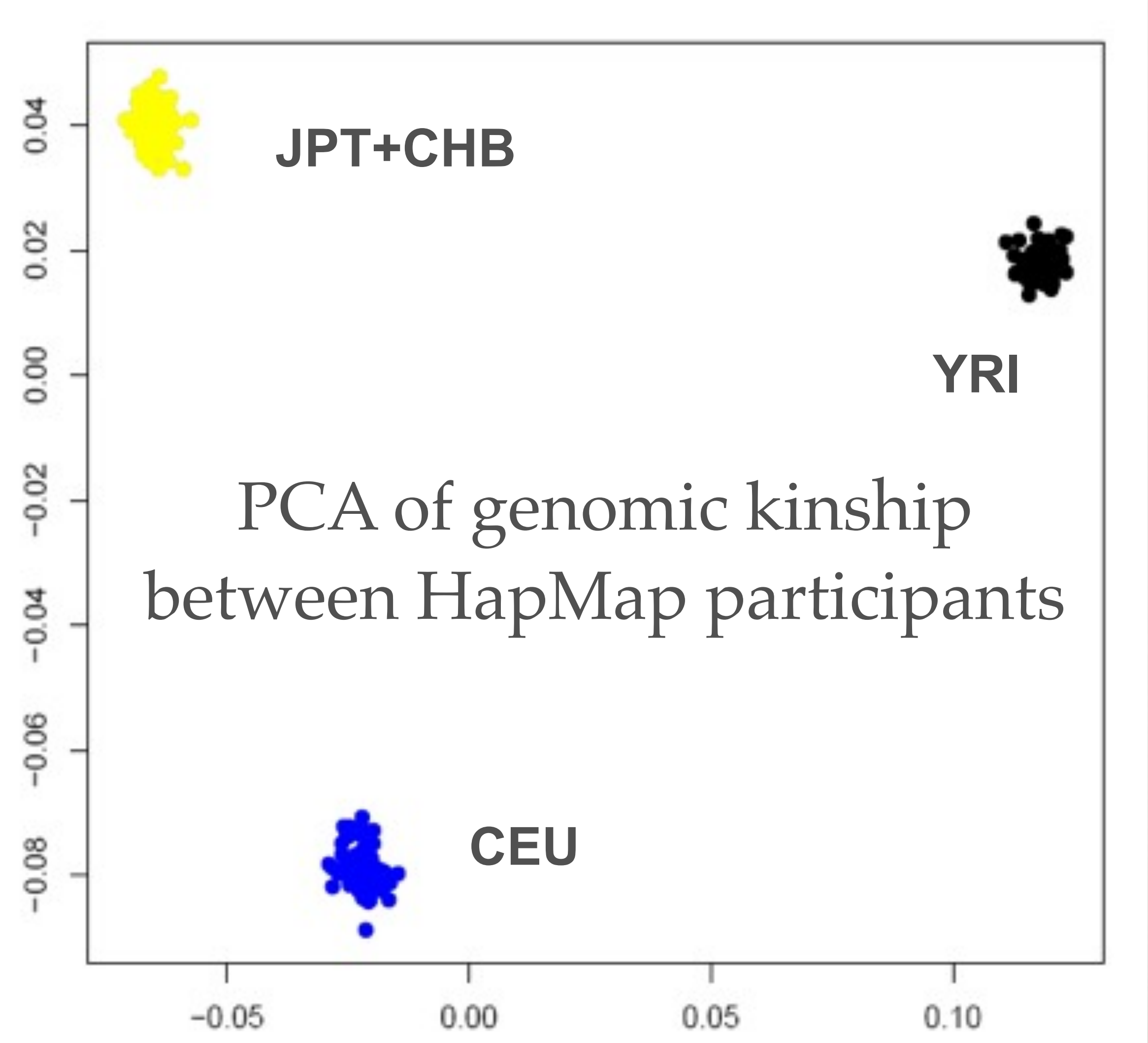
Genomic Control

Structured Association

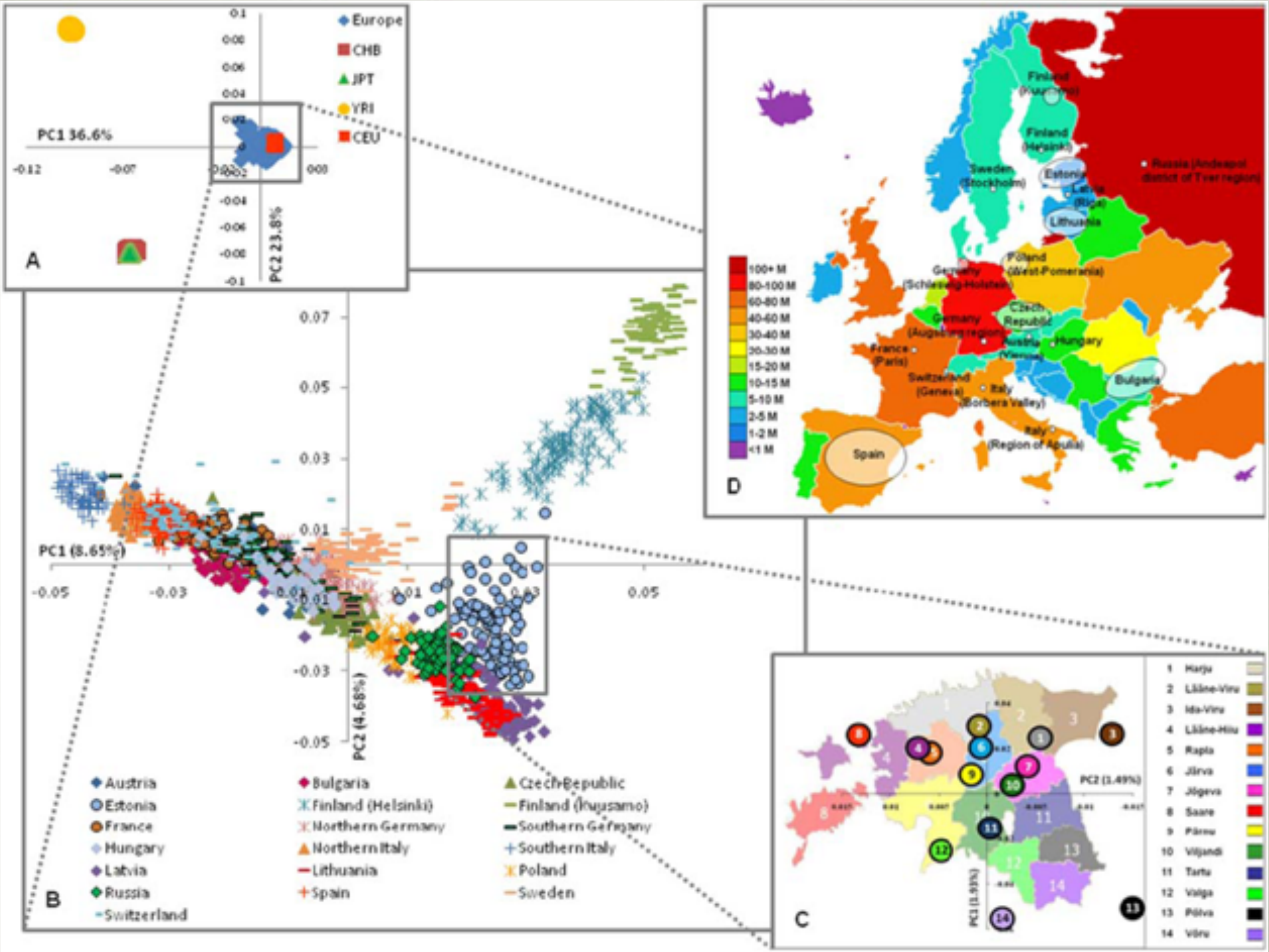
EigenSTRAT

Mixed Models

PCA of genomic kinship



PCA of genomic kinship between HapMap participants



Nelis et al., PLoS ONE, 2009

EigenStrat and PCA-adjustment

- Estimate genetic relations between the study participants using genomic data; compute pair-wise distance matrix; perform CMDS
- Is equivalent to extraction of principal components (PC) of variation from genotypic matrix
- In analysis of association...
  - EIGENSTRAT: adjust both phenotypes and genotypes for these PCs
  - PCA: include principal axes of variation as covariates in regression model
- Apply GC to correct for residual inflation ($1 < \lambda < 1.05$)

How many axes to use?

- Rule of thumb: 10
- Use the ones significantly associated with the trait
- Stop when λ ~ 1
- ...
- If difficult to decide - think of using Mixed Models

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}

β_g: additive affect of the allele

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

***G:* random polygenic effect ~ MVN(0, $\Phi\ \sigma_G^2$)**

Comparison for a population-based study

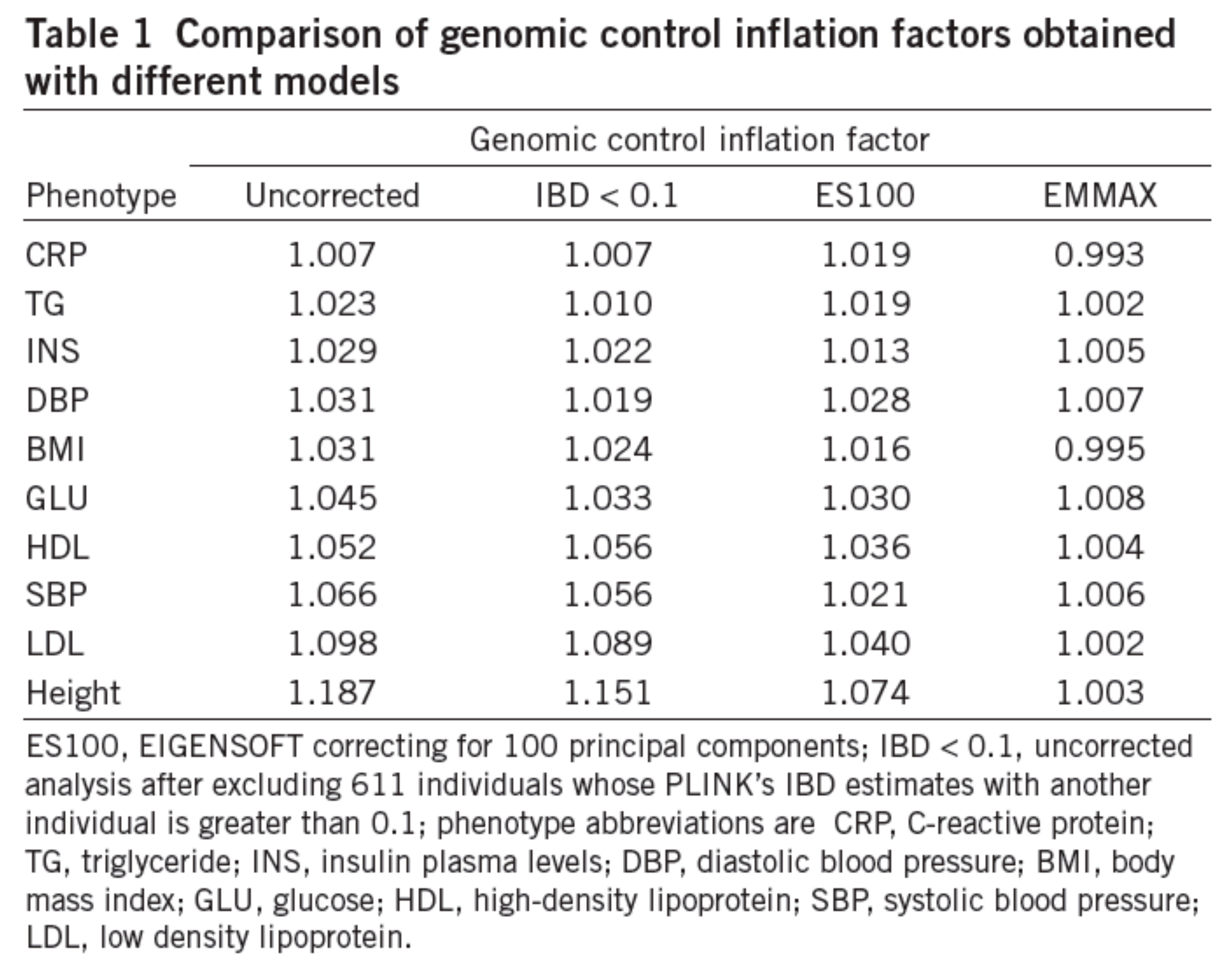


Table 1 Comparison of genomic control inflation factors obtained with different models

	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

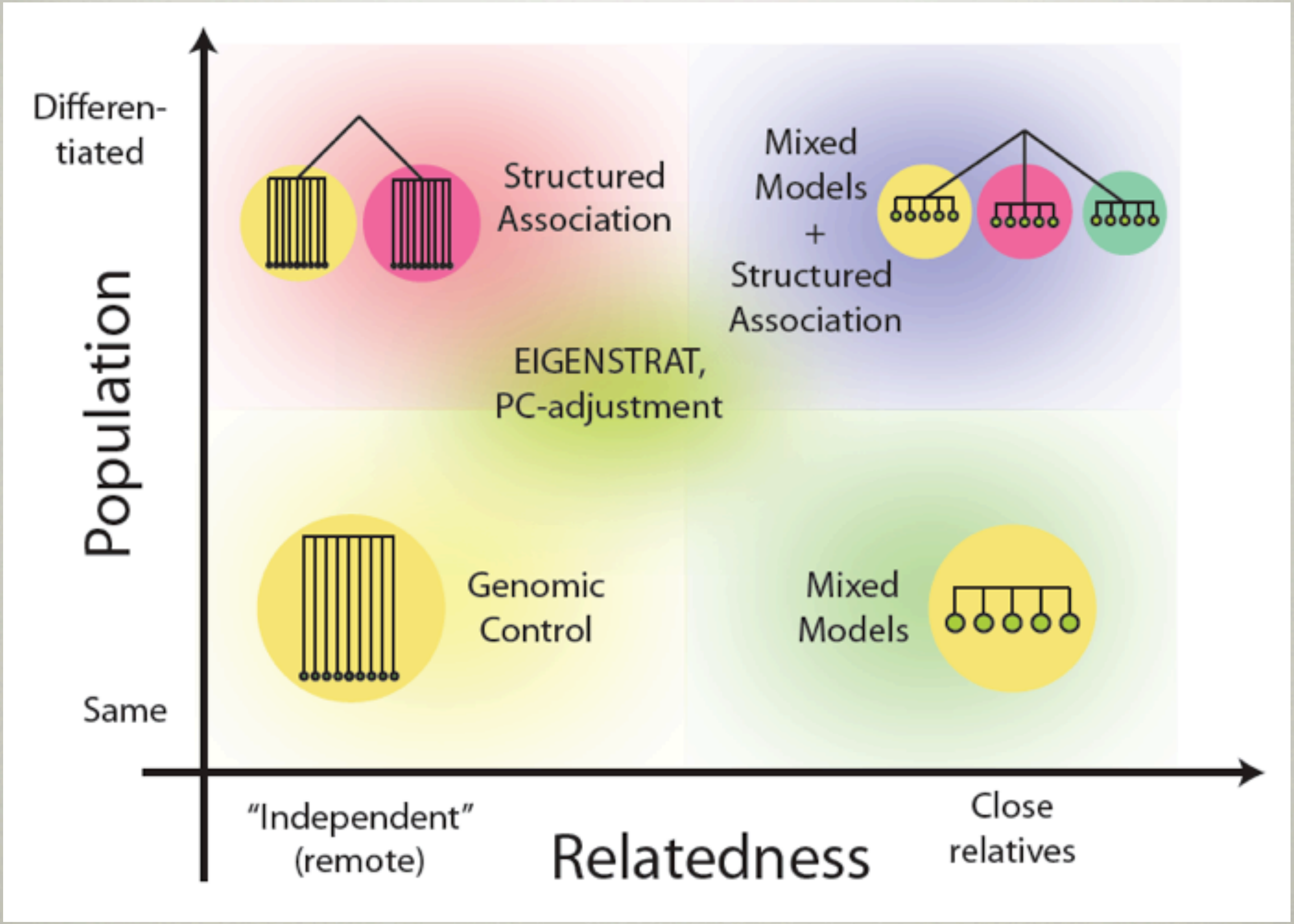
- Excellent method to account for complex genetic structure, such as found in special populations or in family-based studies

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
Same
Population
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