...

Yurii Aulchenko
Erasmus MC Rotterdam

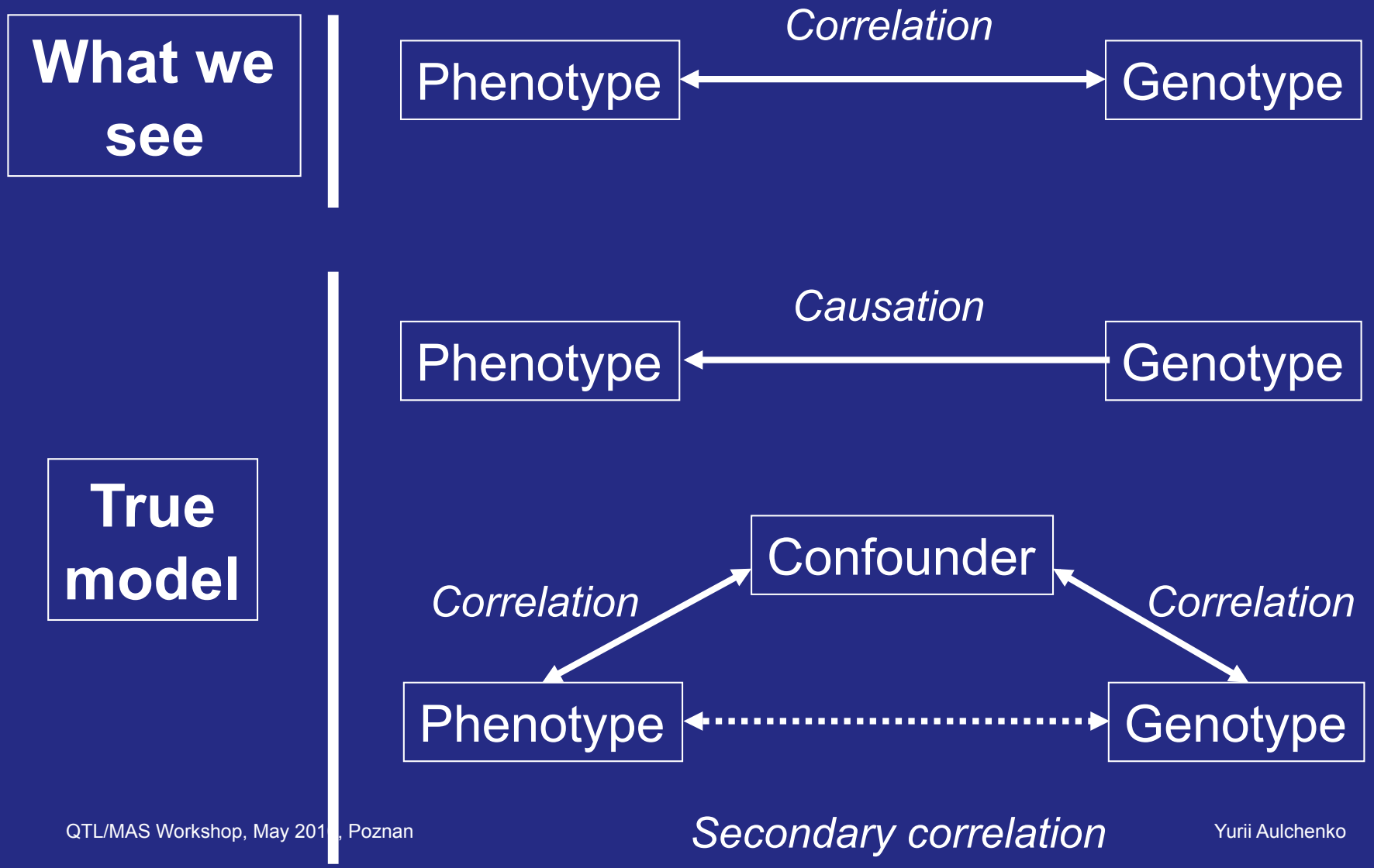
Outline

Reasons for genetic association

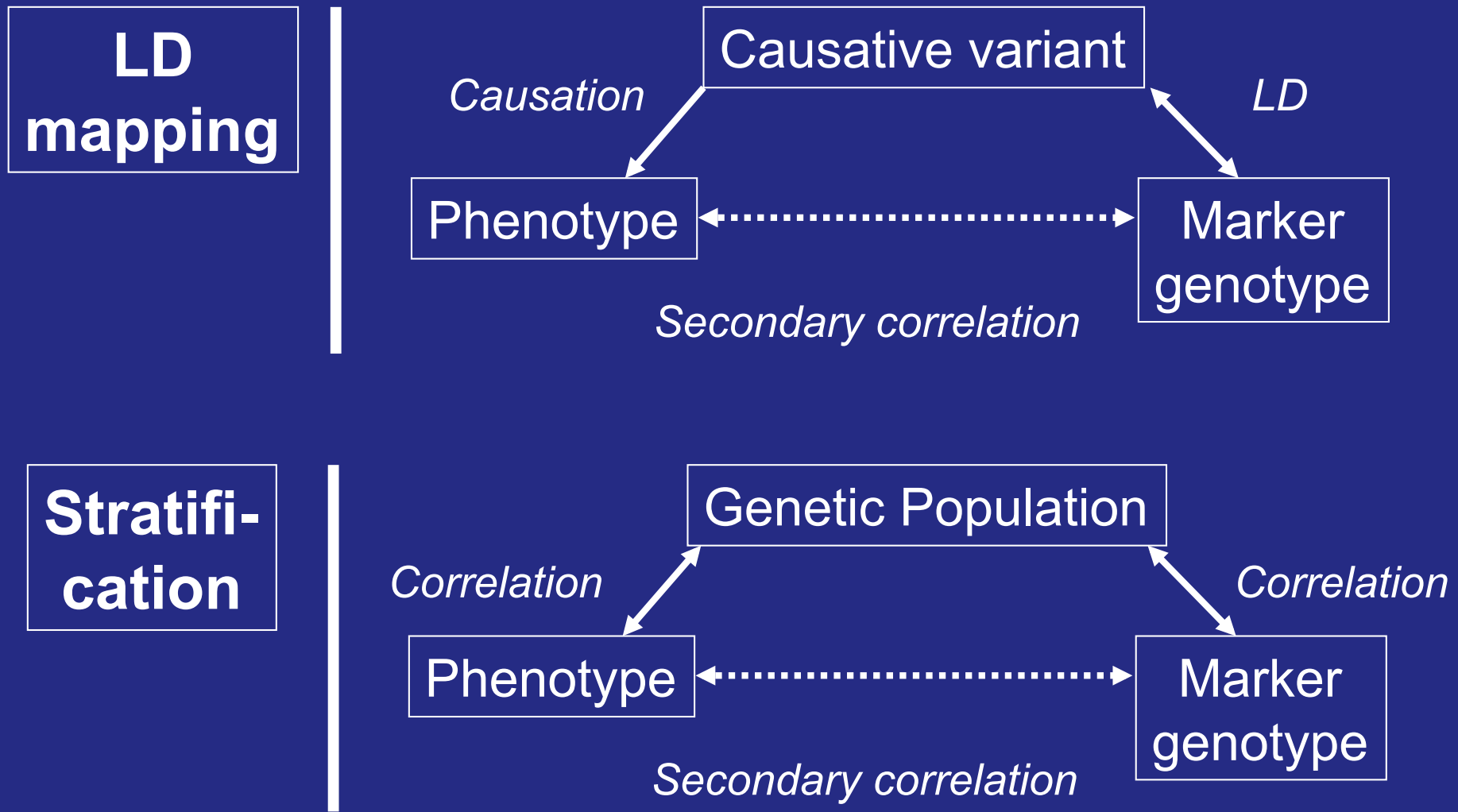
(Genome-Wide) Association analysis in pedigrees

Recent developments

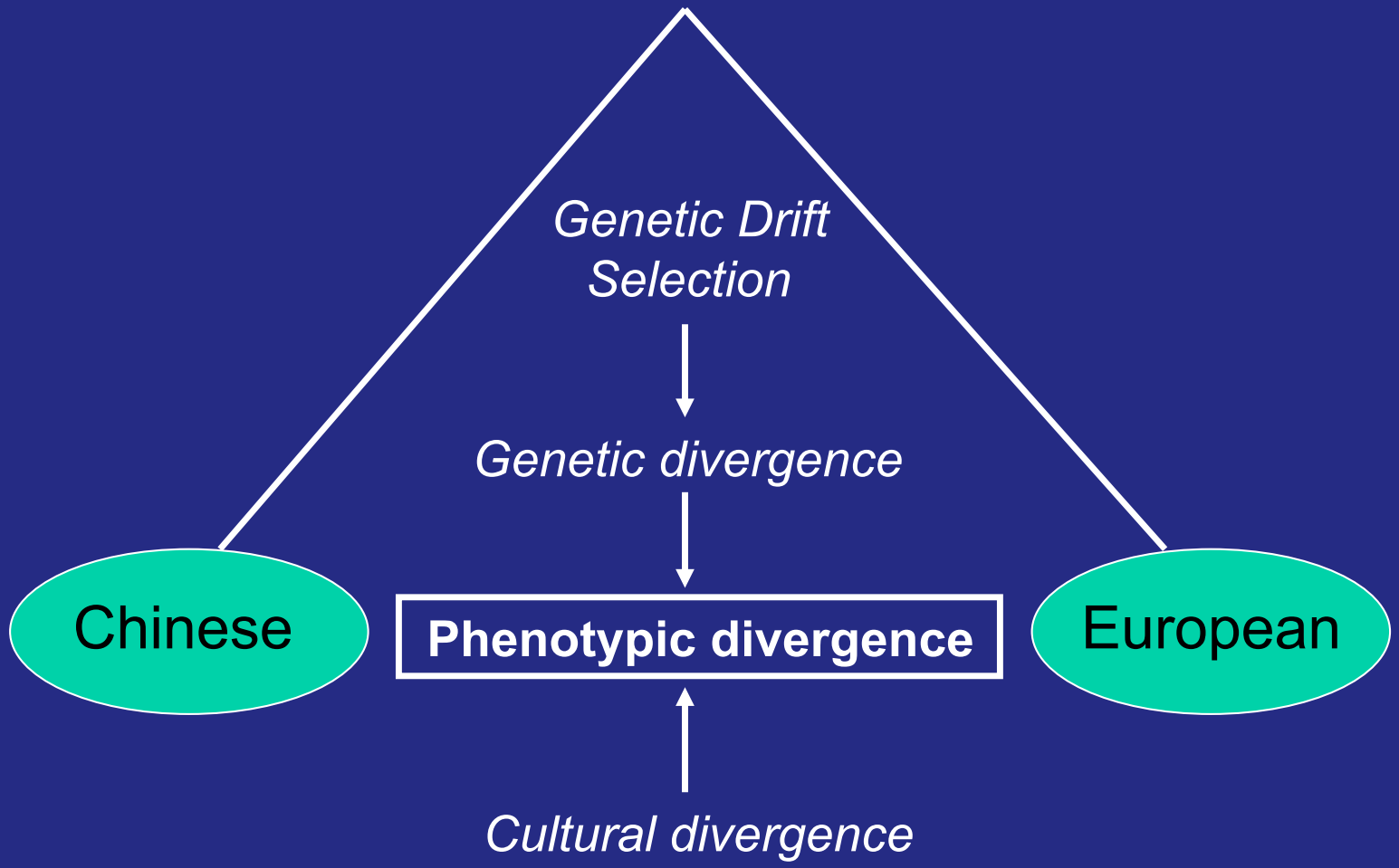
Reasons for genetic association



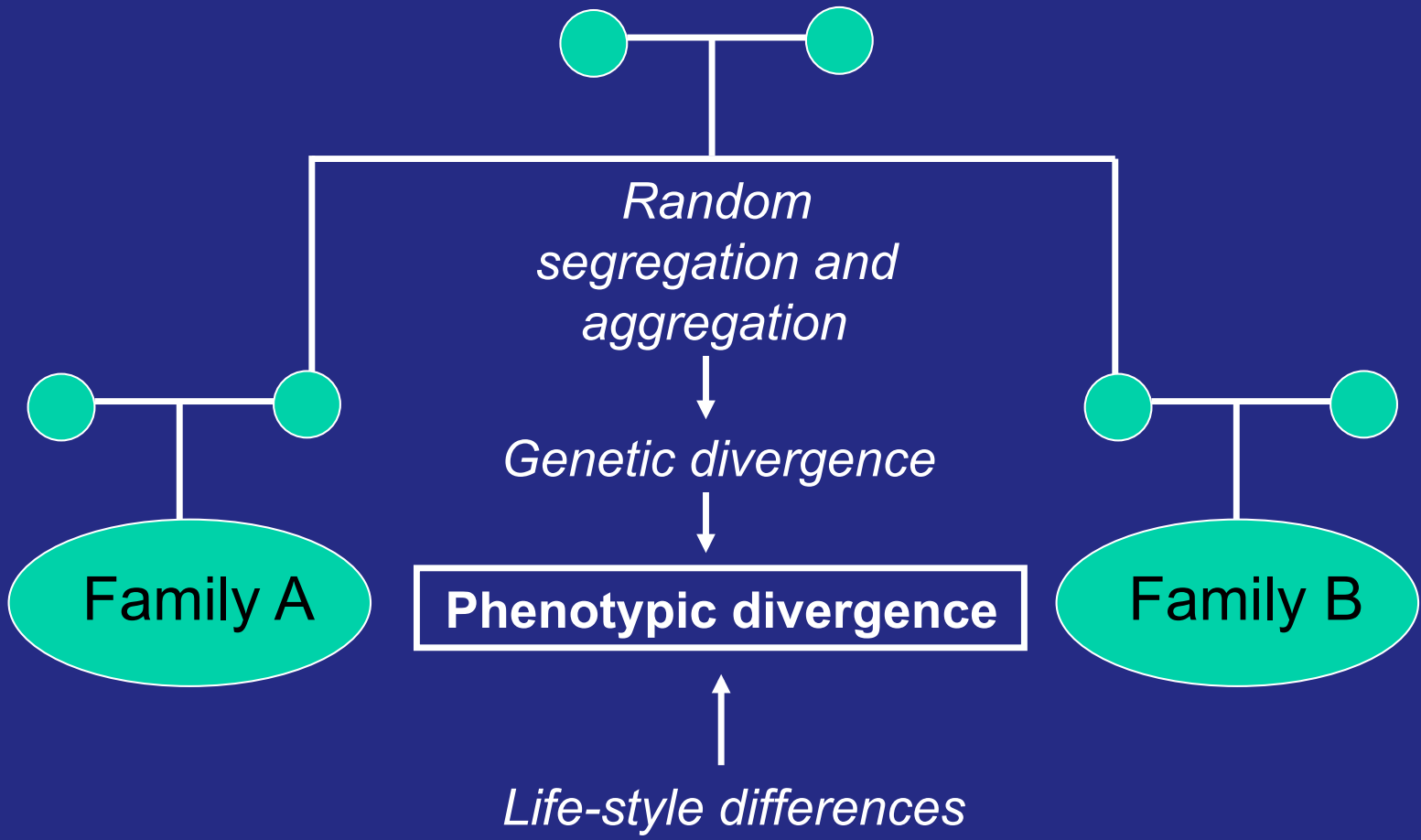
Confounding in genetic studies



Genetic origin is a major confounder

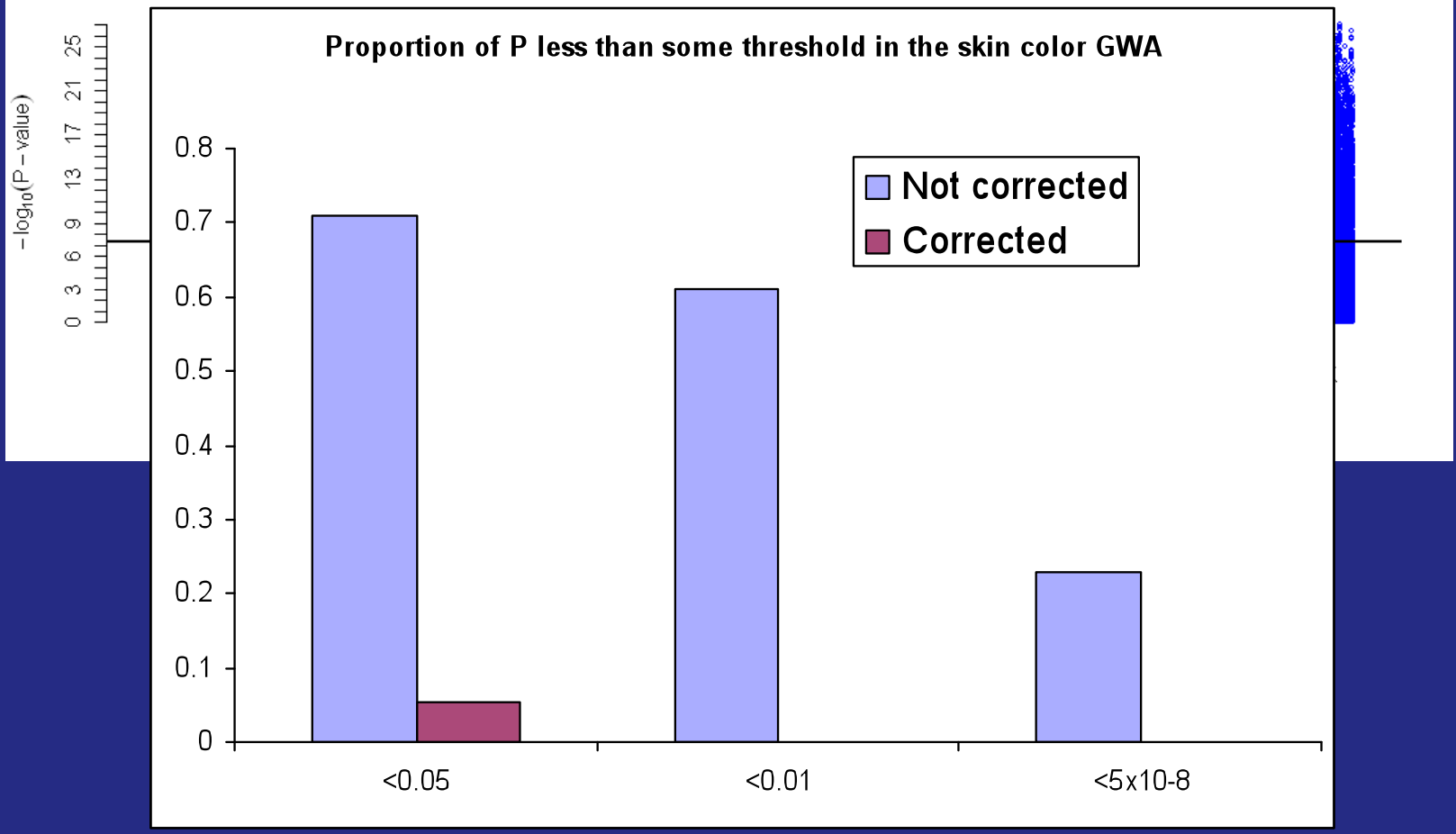


Pedigree is a major confounder



Skin color scan

GWAS of skin color using the HapMap data (European vs African)



Outline

Reasons for genetic association

(Genome-Wide) Association analysis in pedigrees

Recent developments

Genome-Wide Association analysis

100,000s markers distributed over the genome

Each is analysed multiple times:

- Quality control procedures
- Tests for association (including empirical procedures)

Approximately 10^8 – 10^{10} tests to be done

1 test per sec.: 30 years

15,000 tests p.s.: 20 hours

Throughput of population-based tests (e.g. GenABEL): ~40,000 t.p.s.

Mixed (polygenic) model

Vector of quantitative phenotype *Y*

$$Y = \mu + B\,g + G + e$$

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

B: additive affect of the allele

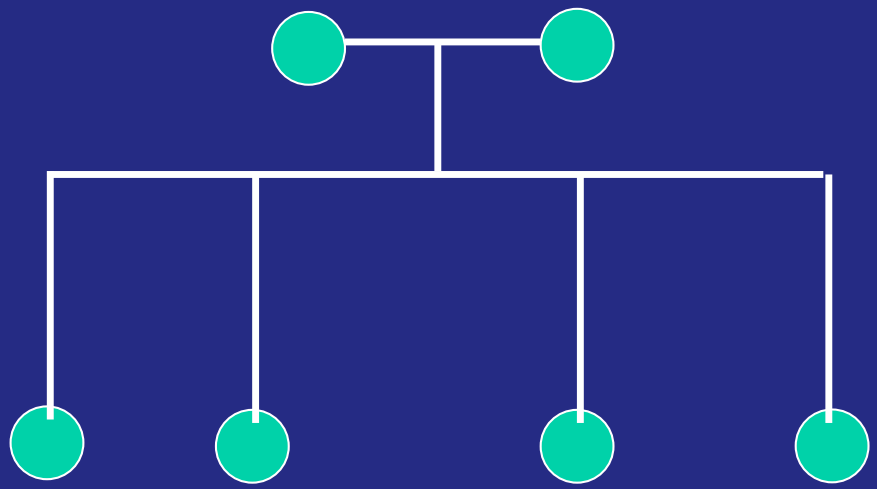
e: is random residual effect ~ MVN(**0**, $\boldsymbol{I}\sigma_e^2$)

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

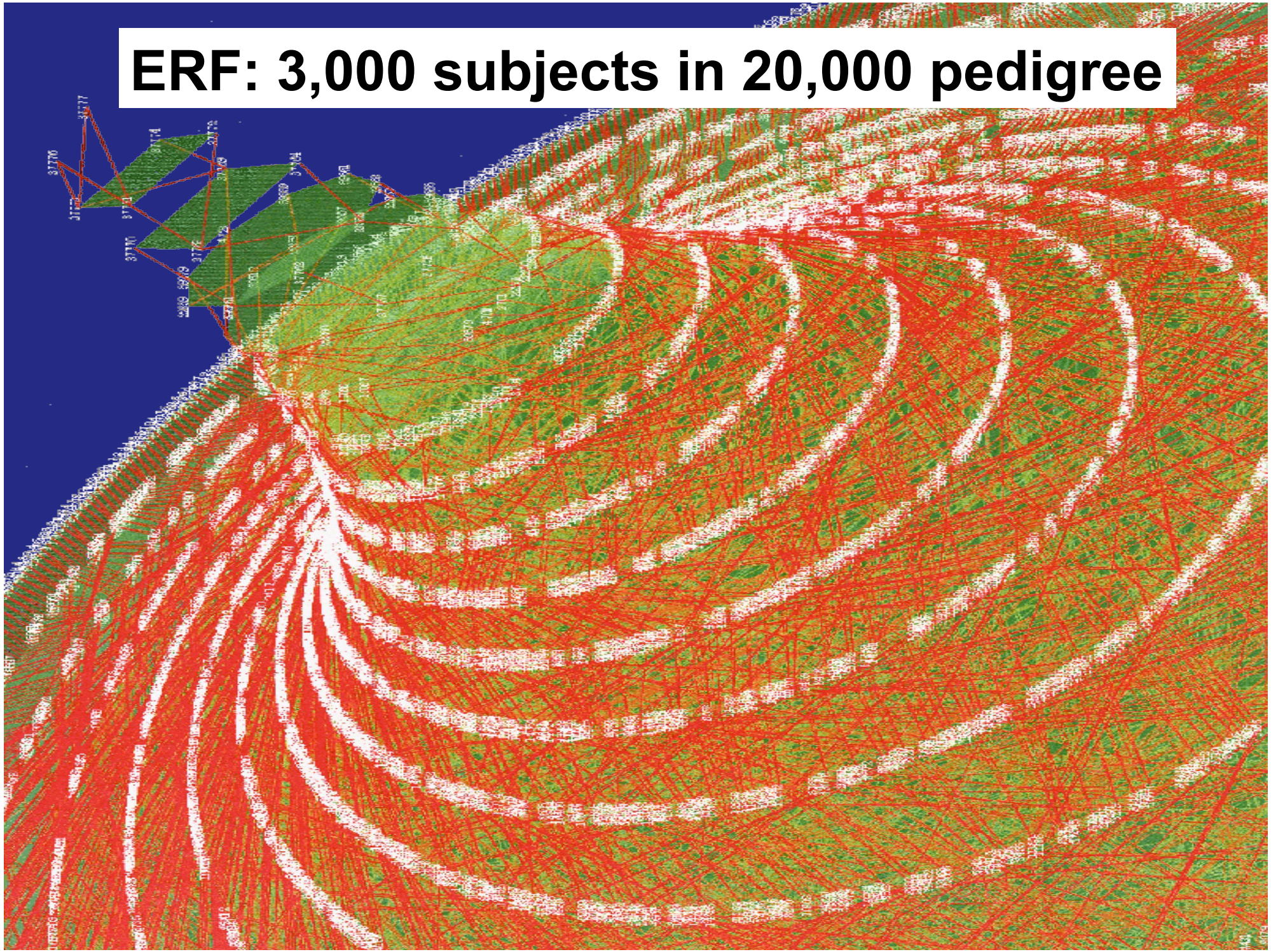
Maximum Likelihood (ML) or Restricted ML (REML)

Software packages available (MERLIN, QTDT, SOLAR, ASReml)

Nuclear pedigree



ERF: 3,000 subjects in 20,000 pedigree



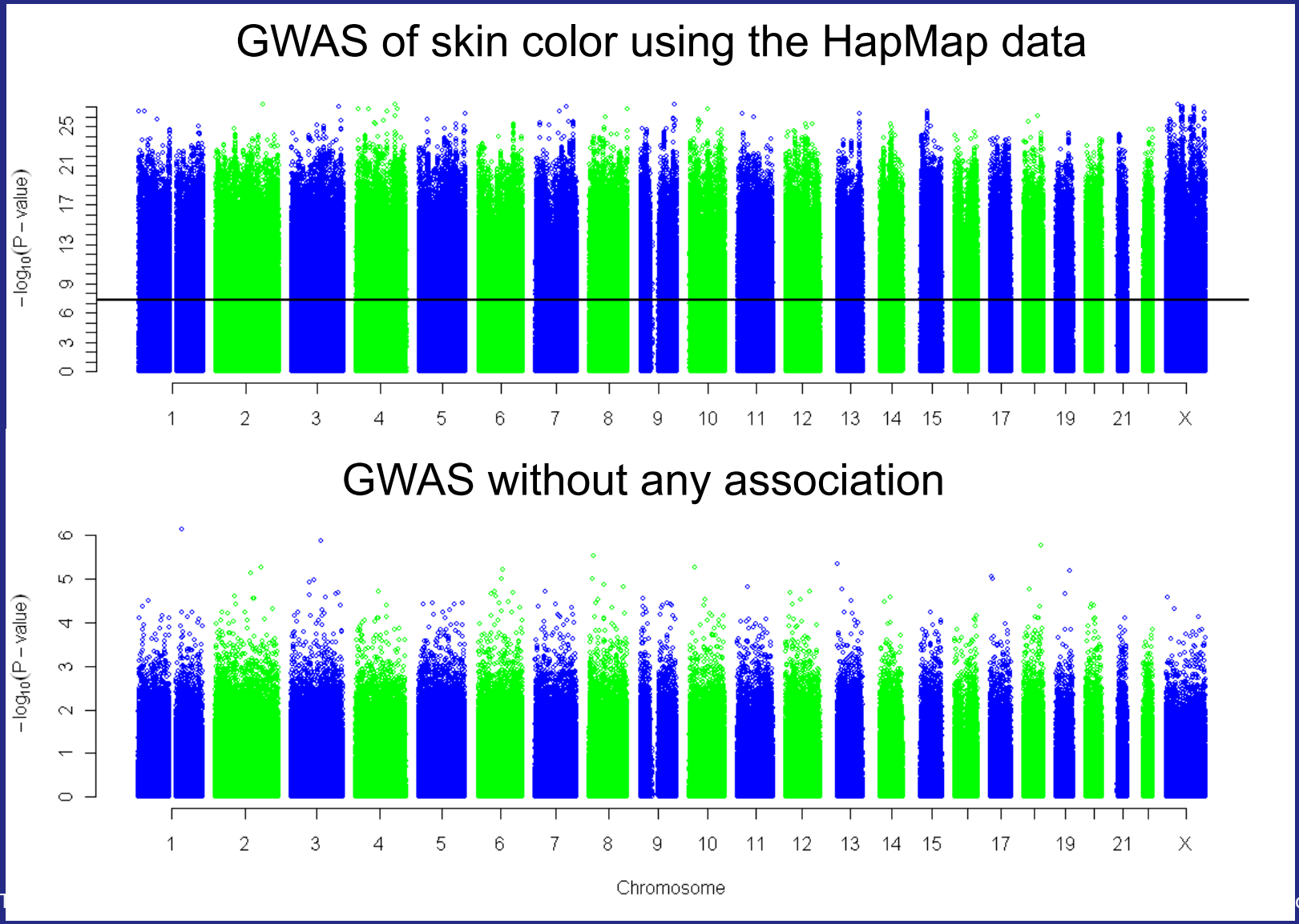
Analysis of large complex pedigrees

ML/REML not feasible: take years for a single GWA scan to be analysed

What are the options?

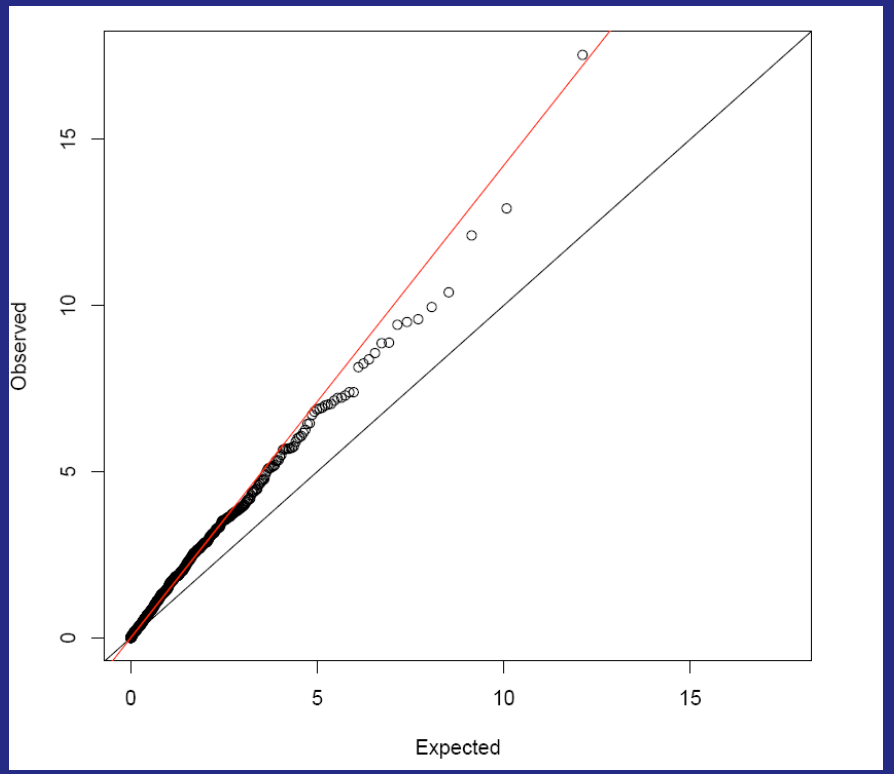
- Genomic control (general technique to account for populational confounding)
- Approximations to ML: FASTA, GRAMMAS, … new staff

Skin color scan



Idea of the genomic control

- Model: $Y = \mu + B\,g + e$
- There is stratification/pedigree
- **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 $\lambda \cdot \chi^2_1$ ($\lambda \geq 1$)



Assumption holds for confounding
If genetic selection: EIGENSTRAT, SA

Solution 1: ignore G, apply GC

Vector of quantitative phenotype *Y*

$$Y = \mu + B\,g + e$$

Score test for association:

$$T^2 = \frac{(g \cdot Y)^2}{g \cdot g} \sim \chi_1^2 \cdot \lambda$$

Lambda is estimated using genomic control (GC):

$$\lambda = \frac{Median\left(T_1^2, T_2^2, T_3^2, \ldots, T_M^2\right)}{0.456}, \quad \lambda \geq 1$$

Computation time ~ O(N)

Few notes on GC

When stratification is large (say, $\lambda > 1.1$) 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 special methods should be used when number of people typed for different SNPs is different

Apparently inflation sets the “detectability limit” on the size of identifiable effect (Holmans’s hypothesis)

FASTA

FAmily Score Test for Association (aka MMSCORE)
Based on the mixed model $Y = \mu + B\,g + G + e$

FASTA test for association:

(a) Estimate polygenic model $Y = \mu + G + e$

(b) Compute FASTA test

$$T^2 = \frac{\left(g \cdot \left(\Phi\hat{\sigma}_G^2 + \mathrm{I}\hat{\sigma}_e^2\right)^{-1} \cdot Y\right)^2}{g \cdot \left(\Phi\hat{\sigma}_G^2 + \mathrm{I}\hat{\sigma}_e^2\right)^{-1} \cdot g} \sim \chi_1^2$$

(c) Apply GC afterwards if λ > 1

[extended in recent ProbABEL paper]

Computation time ~ O(N²)

Chen & Abecasis, 2007; Aulchenko & Struchalin, 2010

GRAMMAS-GC

GW Rapid Association using Mixed Model And Score test

Based on the mixed model $Y = \mu + B\,g + G + e$

GRAMMAS test for association:

(a) Estimate polygenic model $Y = \mu + G + e$

(b) Compute environmental residuals $Y^* = Y - (\hat{\mu} + \hat{G}) = \hat{e}$

(c) Runs score test on residuals

$$T^2 = \frac{\left(g \cdot Y^*\right)^2}{g \cdot g} = \frac{\left(g \cdot \hat{\sigma}_e^2 \cdot \left(\Phi\hat{\sigma}_G^2 + \mathrm{I}\hat{\sigma}_e^2\right)^{-1} \cdot Y\right)^2}{g \cdot g}$$

(d) **Apply GC (λ expected to be < 1) => grammas-gc**

Computation time ~ O(N)

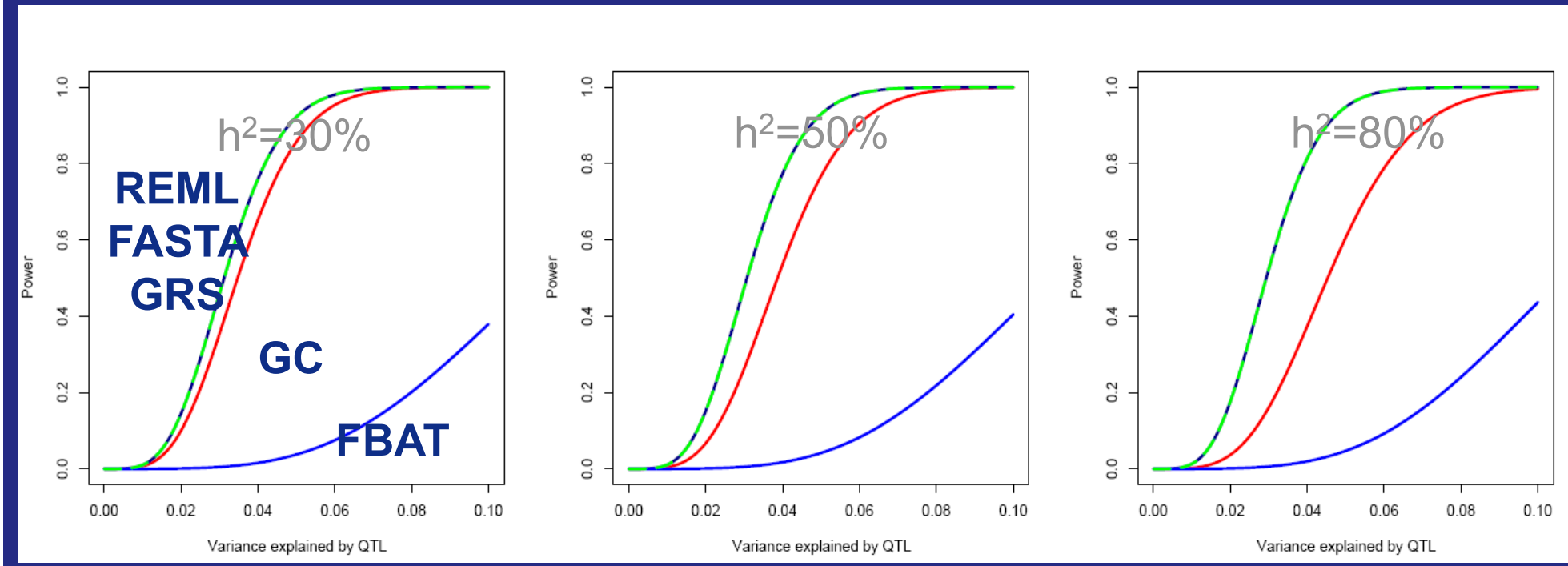
Aulchenko et al, 2007; Amin et al., 2007

Power comparison

Part of ERF pedigree

Associated SNP explained 1, 2 or 3% of variance

Polygenic effect simulated using MVN distribution



Relationship between genomes

The estimate of kinship between *i* and *j* may be obtained from genomic data:

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

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

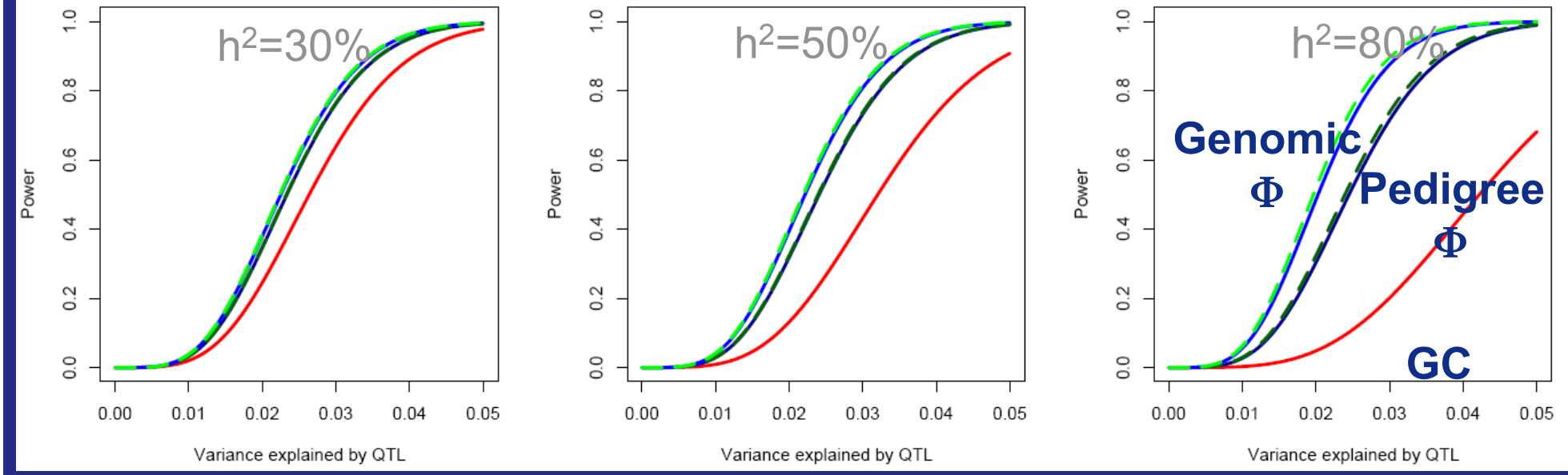
p_k is the frequency of "1" allele

Genomic vs. Pedigree kinship

1,400 ERF people genotyped for 6K Illumina Array

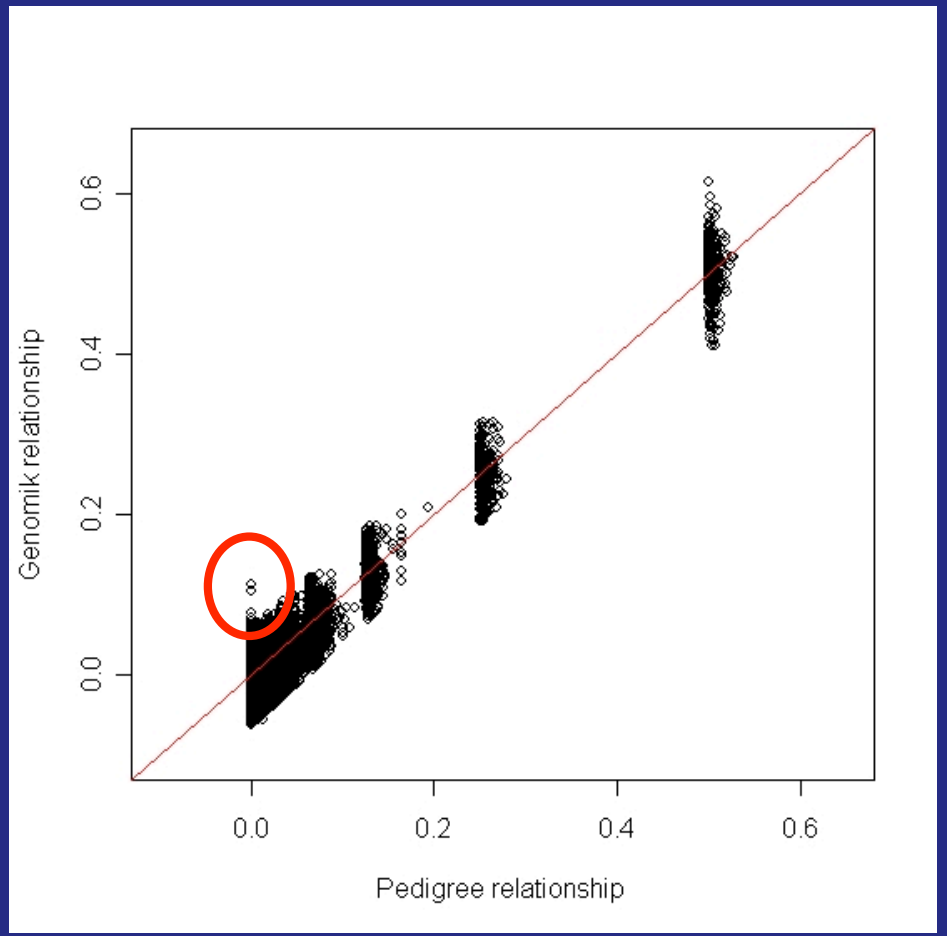
Trait values simulated based on observed genotypes

Associated SNPs explained from 0.3 to 4% of variance



Genomic Φ is better than pedigree Φ

- Pedigree is not guaranteed to be correct
  - Missing links => increased type 1 error
- Pedigree relationship coefficient is the <u>expected</u> proportion of genome shared
  - Genomic relationship may better estimate true sharing



Outline

Reasons for genetic association

(Genome-Wide) Association analysis in pedigrees

Recent developments

Recent developments: even better GRAMMAR

Power of GRAMMAR can be brought to the power of reml/mmscore with “inverse GC”, giving GRAMMAS-GC. The problem remained that the effect estimates were biased downwards.

P. Visscher noticed long time ago that bias seems to be a simple function of heritability. Apparently we now can prove that and get unbiased estimates for GRAMMAR.

Recent developments: FMM

Existing practical methods are approximate (2-step). W. Astle and D. Balding have developed fast mixed models algorithm, which is under implementation in GenABEL

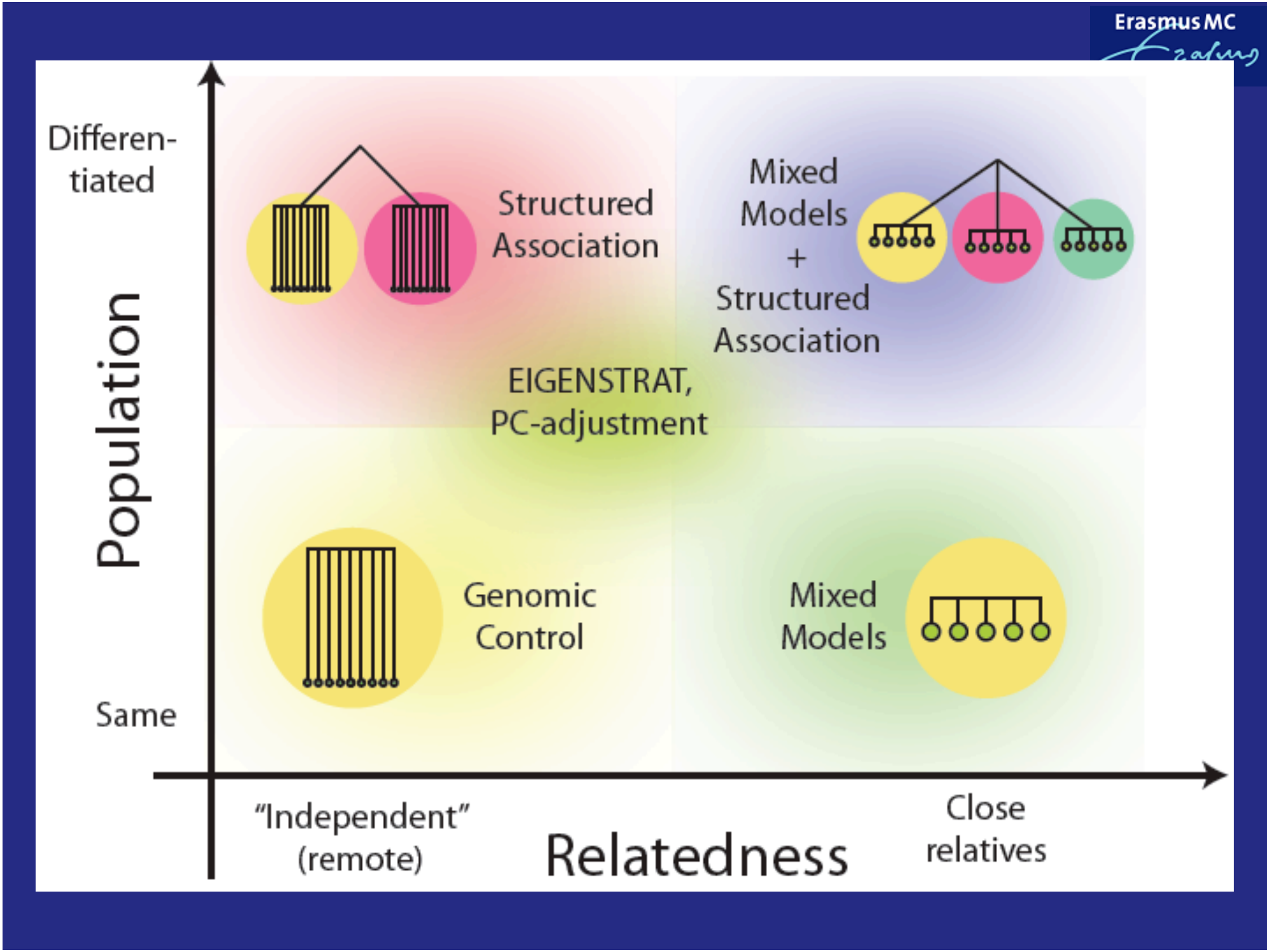
Computation time: for all methods, linear with #markers, with n = #people:

- GRAMMAS-GC ~ O(n)
- MMSCORE ~ $O(n^2)$
- new FMM ~ $O(n^2+kn)$, where k is small (3-10)

Recent developments: binary traits

Up until now the problem of GWA analysis of binary traits in samples of relatives was not (practically) solved. Biggest problem is not a correct p-value, but rather good effect estimate (Odds Ratio)

N. Pirastu is working on simple method for analysis of binary traits; will be implemented in GenABEL in near future. Speed will be the same as the speed for analysis of QTs


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