Erasmus MC
Universitair Medisch Centrum Rotterdam

Genome-wide association analysis in samples of related individuals

Yurii Aulchenko

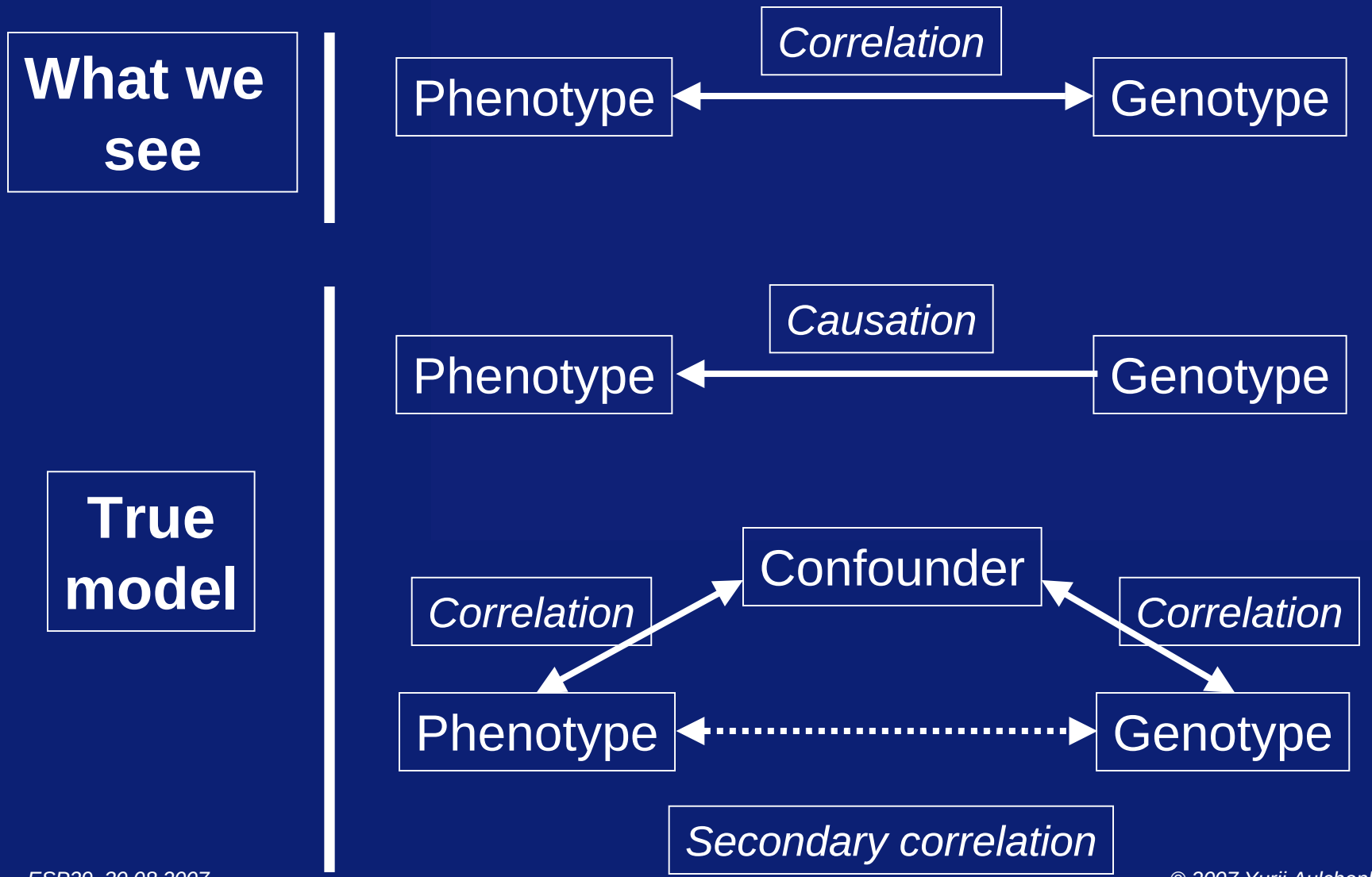
Erasmus MC Rotterdam

The Netherlands

Overview

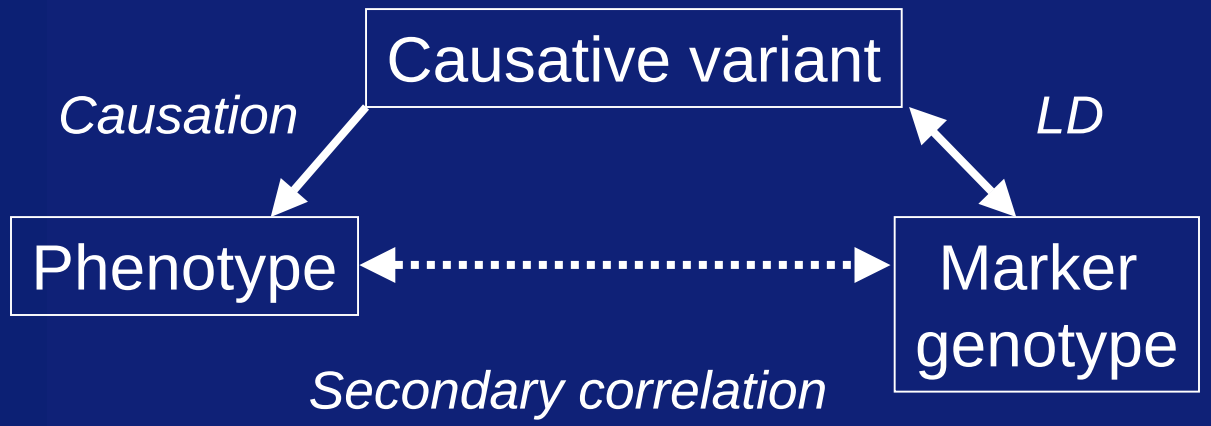
- Confounding in genetic studies
- Analysis of samples of relatives from genetically homogeneous population
- Analysis of samples of relatives from genetically heterogeneous population

Reasons for genetic association

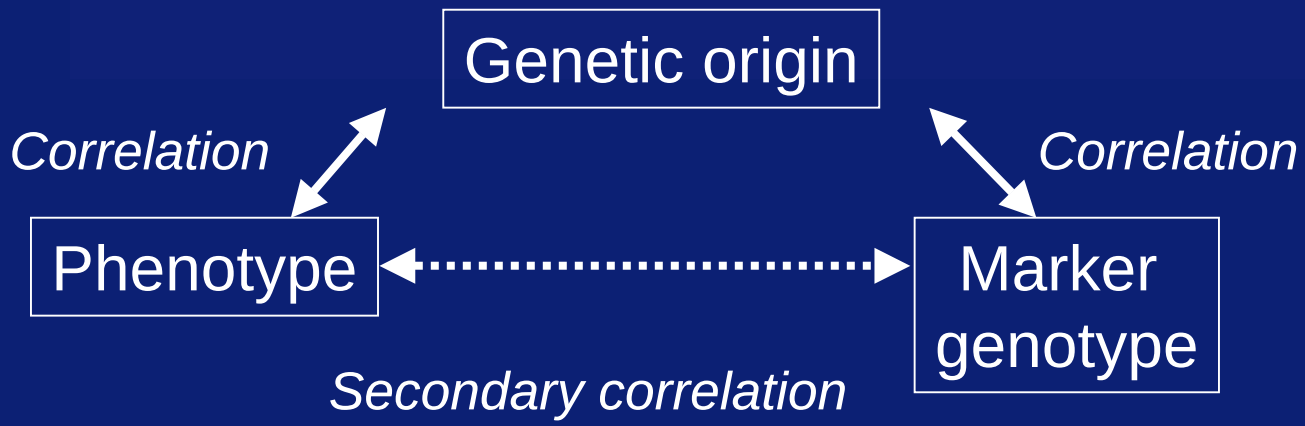


Confounding in genetic studies

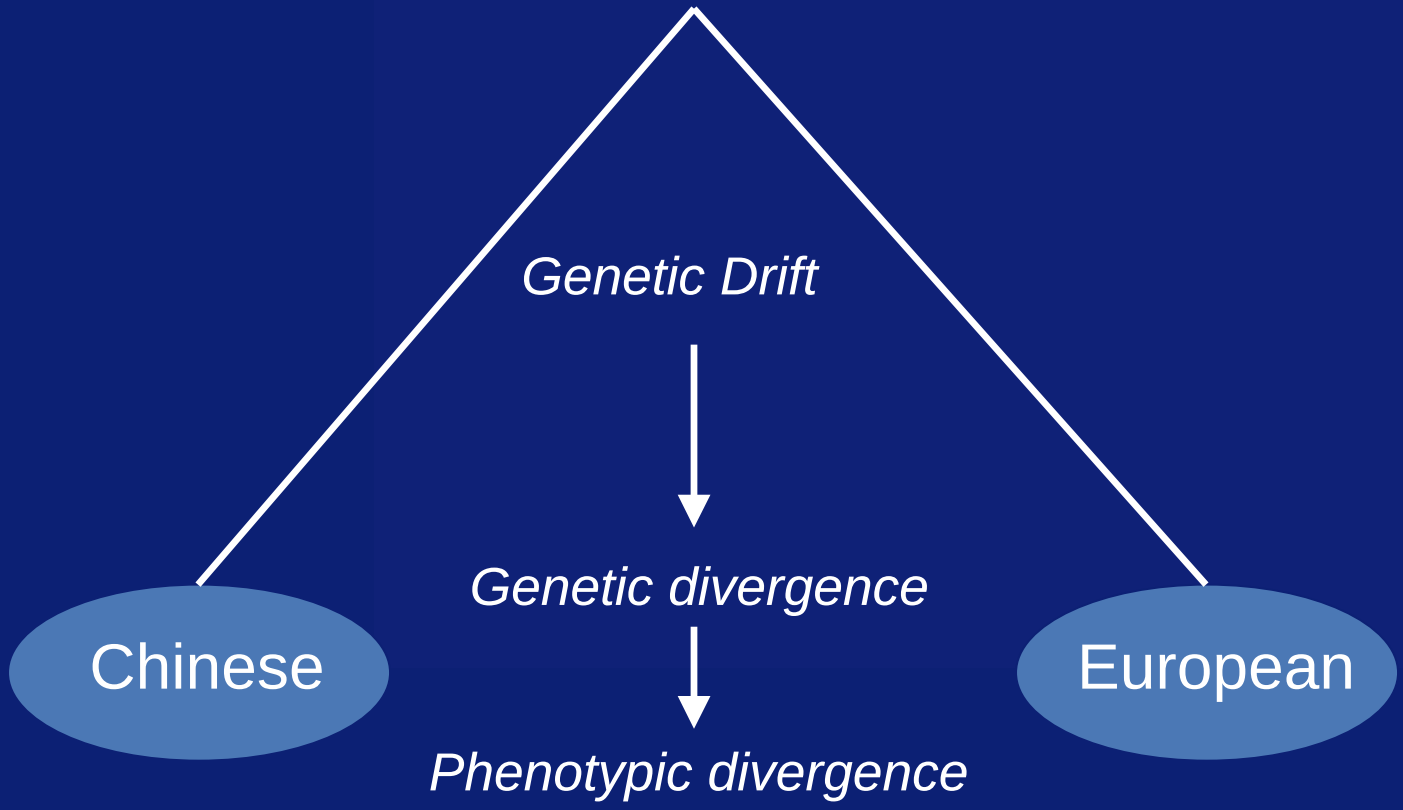
LD mapping



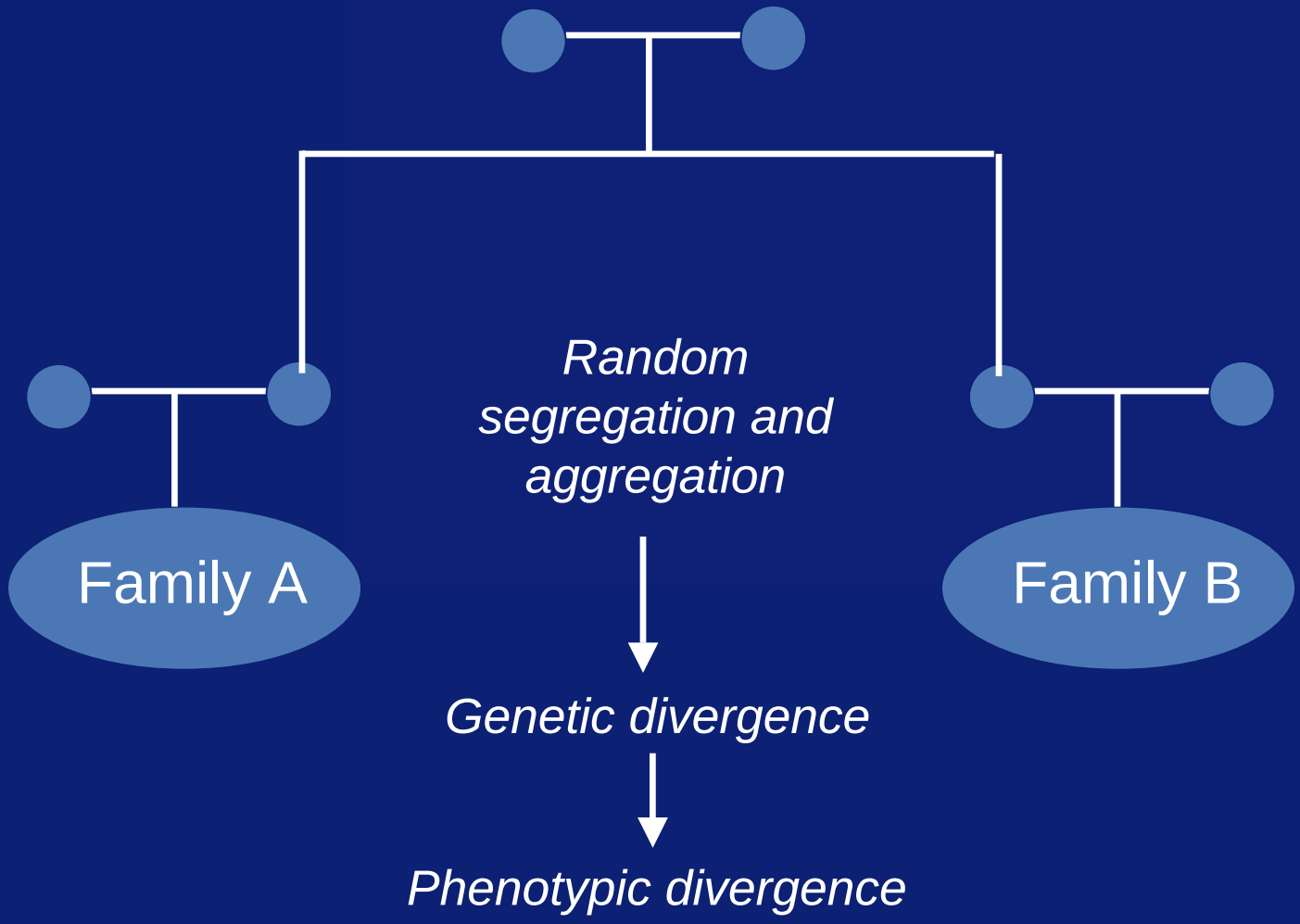
Stratifi-cation



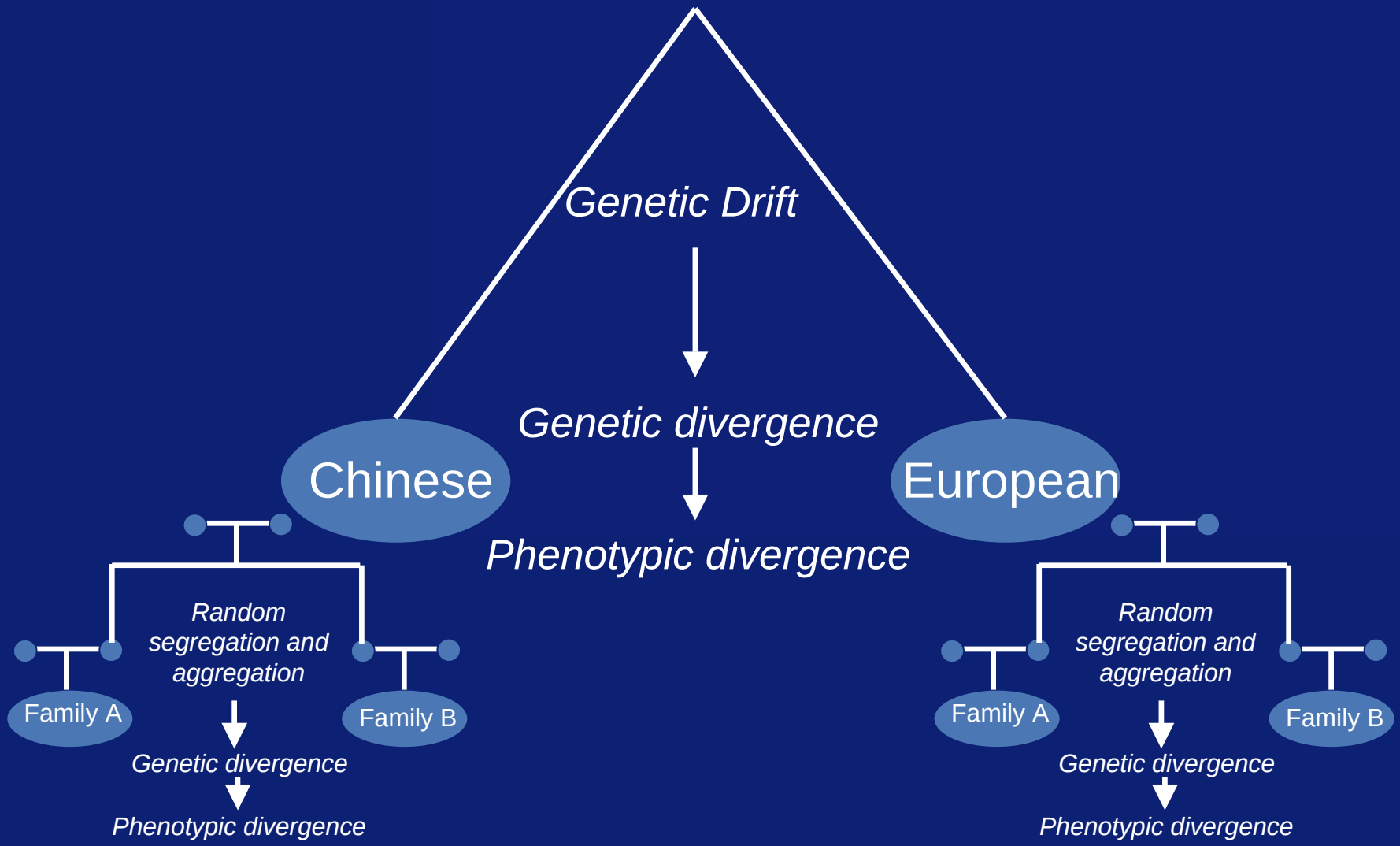
Population is a major confounder



Pedigree is a major confounder



Both population and pedigree are!



Linear model

Vector of quantitative phenotype *Y*

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

g is vector of genotypes (coded 0, 1, 2)

B is additive effect of the genotype

e is the vector of random residuals

Score test for association: $$T^2 = \frac{(g \cdot Y)^2}{g \cdot g} \sim \chi_1^2$$

- Computation time ~ N
- **Generates false positives in presence of pedigree**

Am. J. Hum. Genet. 69:1146–1148, 2001

Erasmus MC

The Importance of Genealogy in Determining Genetic Associations with Complex Traits

DINA L. NEWMAN,[1] MARK ABNEY,[1,2] MARY SARA MCPEEK,[1,2] CAROLE OBER,[1] AND NANCY J. COX[1]

- >750 Hutterites. Association tested between 3 quantitative traits (IgE level, LDL, BMI) and >500 markers with and without modeling the relatedness

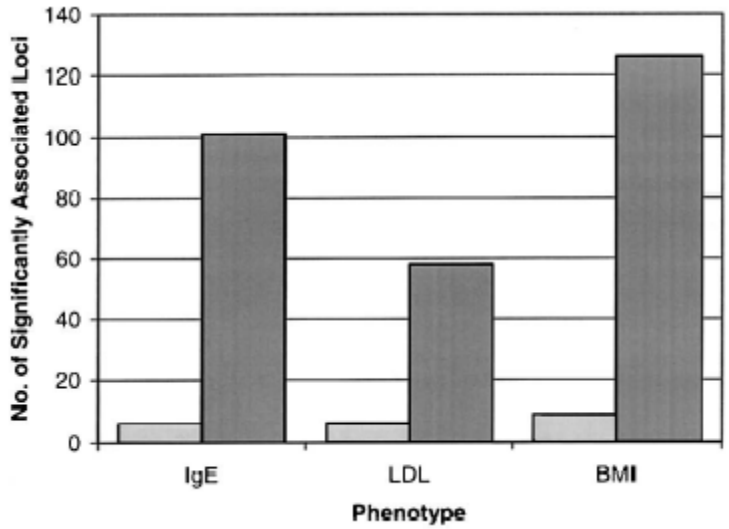


Figure 2 Number of significantly associated ($P < .01$) loci when pedigree structure is included (*lighter bars*) and when pedigree structure is not included (*darker bars*).

➔ High level of false positive signals

Genomic Control (GC)

Compute the vector of test statistics genome-wide

$$\{T^2_1, T^2_2, T^2_3, \ldots, T^2_{N-1}, T^2_N\}$$

Estimate inflation factor λ as

$$\text{Median}\{T^2_1, T^2_2, T^2_3, \ldots, T^2_{N-1}, T^2_N\} / 0.456$$

The GC-corrected test statistics

$$T^2/\lambda \sim \chi^2_1$$

Mixed (animal) model for pedigrees

Vector of quantitative phenotype Y

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

G is random polygenic effect distributed as MVN($\mathbf{0}$, $\Phi\,\sigma_G^2$)

Φ is relationship matrix

σ_G^2 is polygenic variance

Again GWA analysis

- Assessment of 100-1,000K SNPs in thousands of study participants
- Analysis of association between each of these SNPs and traits of interest
- Millions of tests => they should be fast

Potential problems with classical MM

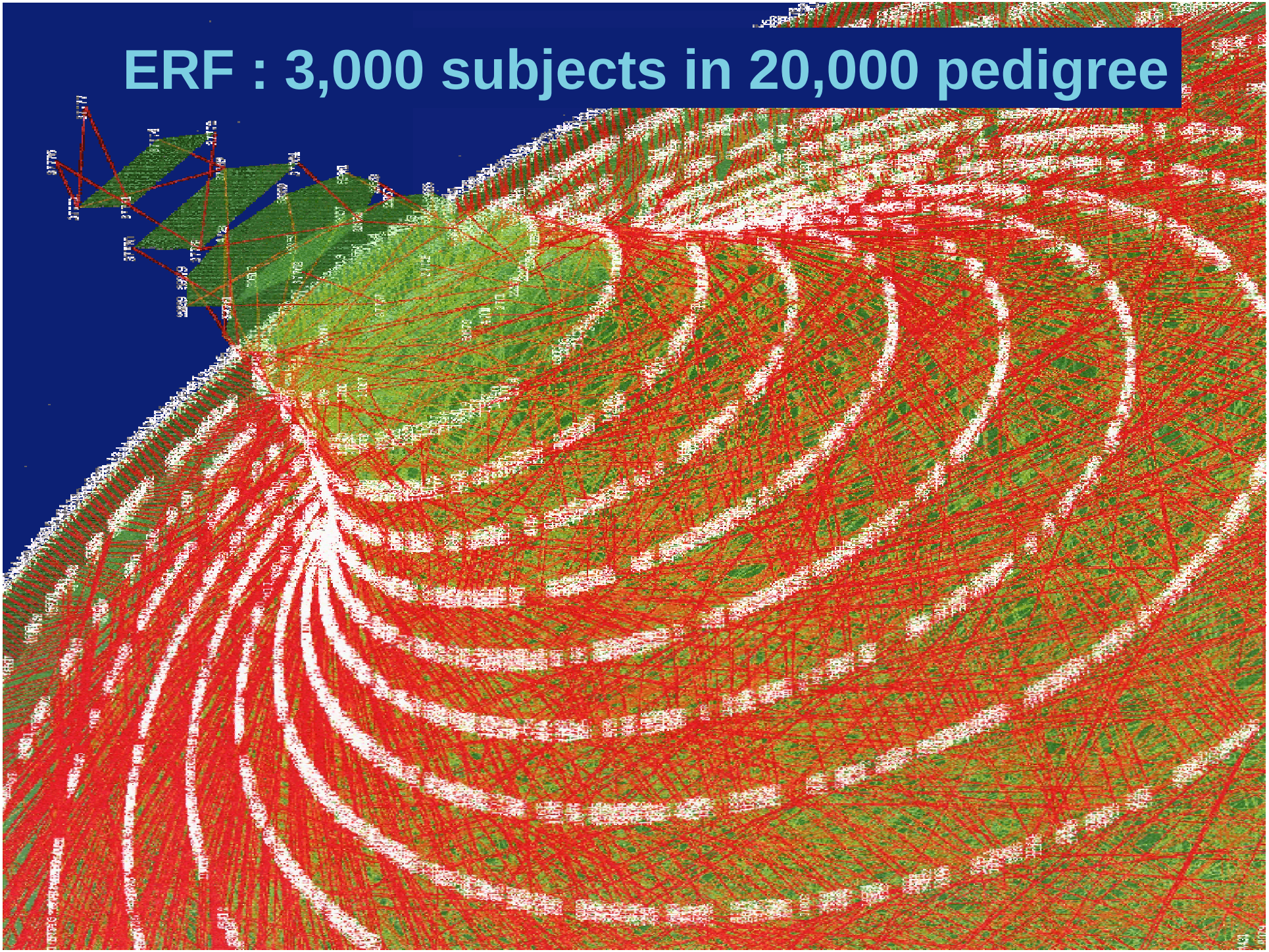
Estimation of relationship matrix Φ

- No problem if pedigree is known
- In most studies, pedigree is only partly known or not known!
- Use of genomic kinship?

Large pedigrees from genetically isolated populations

- Analysis of single SNP may take few minutes
- ERF pedigree: 15 minutes
- GWA with 318K: 9 years

ERF : 3,000 subjects in 20,000 pedigree



Family-based Score Test for Association (FASTA)

Erasmus MC

Estimate polygenic model from the data

FASTA test for association:

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

Apply GC to correct for residual inflation (if any)

Computation time ~ N^2+N (N times slower than GC!)

Chen & Abecasis, Am. J. Hum. Genet., in press

Genome-wide Rapid Association using Mixed Models And Score test (GRAMMAS)

Avoid vector-by matrix multiplication by use of environmental residuals from polygenic analysis

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

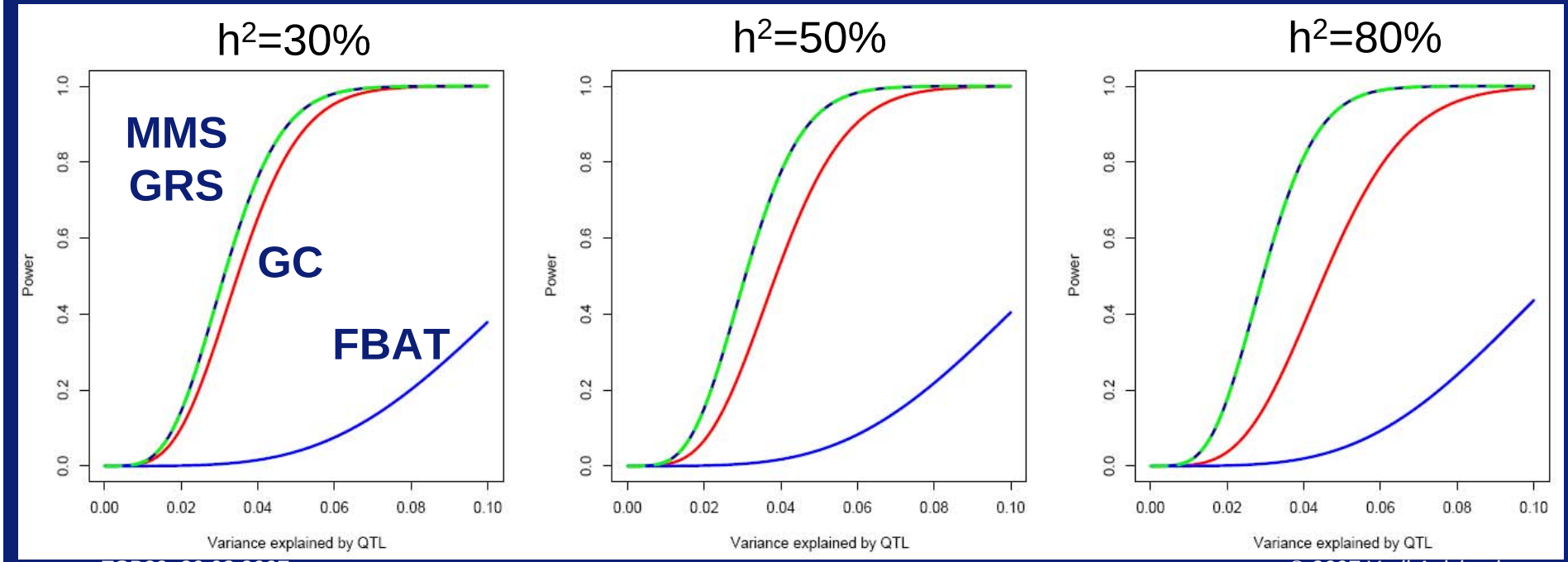
GRAMMAS: Score test + GC

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

Aulchenko et al., Genetics, in press

Comparison of FASTA, GRAMMAS, GC and TDT

- Part of ERF pedigree
- Associated SNP explained 1, 2 or 3% of variance
- Polygenic effect simulated using MVN distribution



Relationship matrix from genomic data

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

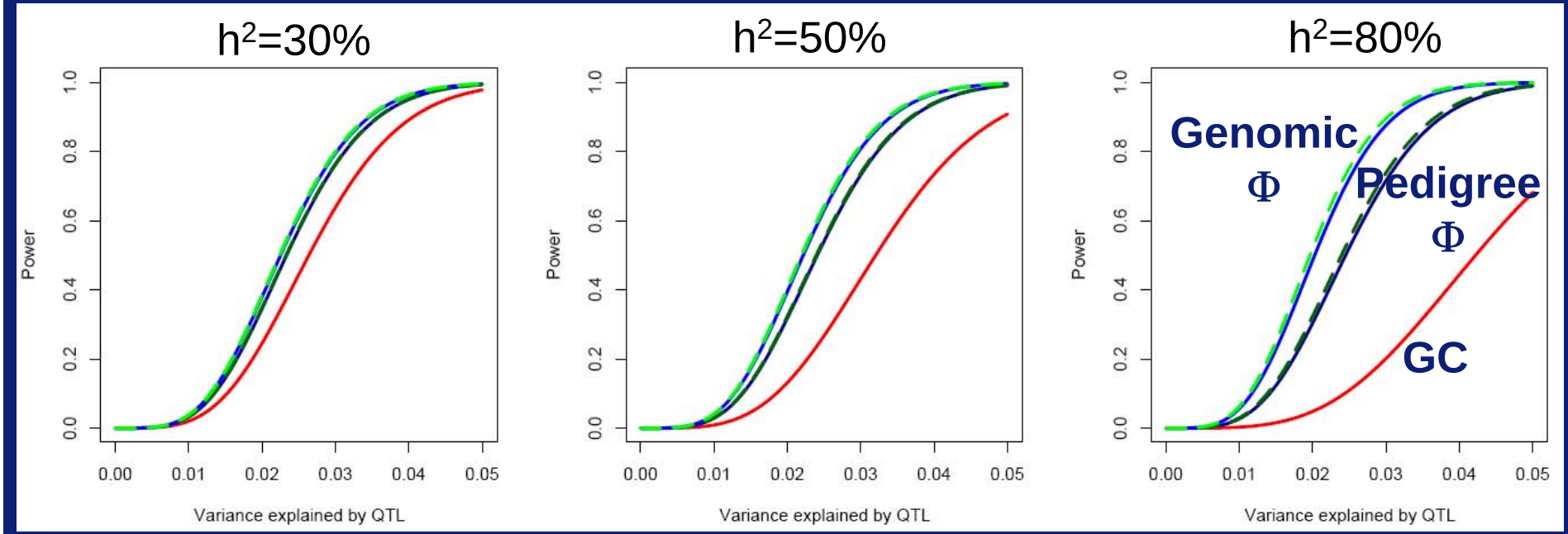
$$f_{ij} = \frac{1}{n}\sum_{k=1}^{n}\frac{(g_{ik} - p_k)(g_{jk} - p_k)}{p_k(1-p_k)}$$

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

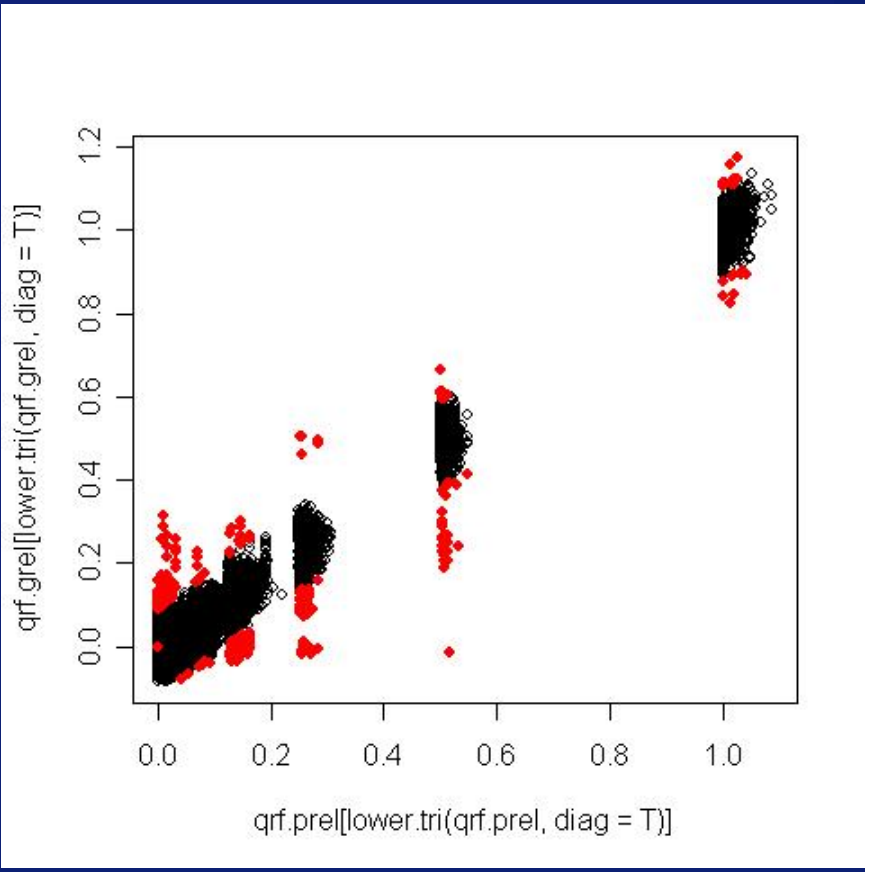
Genomic vs. Pedigree kinship

- 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



Why genomic kinship is better than pedigree kinship?

- Pedigree is not guaranteed to be correct
- Genomic relationship may better estimate true genomic proportion shared
- Genomic kinship:
  - More precise h^2 estimation
  - Better prediction of residuals



Conclusions

- GC and Genomic FASTA/GRAMMAS are the methods for analysis of samples of relatives in absence of pedigree data
- Power Genomic FASTA ~ Power GRAMMAS > Pedigree-based F~G > GC
- **Recommended: genomic FASTA/GRAMMAS**

What if relatives come from different populations?

- Originally considered by Yu et al., Nat Genet 2006

- Combine structured association with previous methods (e.g. FASTA/GRAMMAS)

Transmission-disequilibrium test (TDT, FBAT, QTDT, etc.)

- Analyses effect of SNP on WITHIN-FAMILY variation
- Robust test for association in presence of population stratification
- For sib-pairs:

Partition vector $\vec{g}$ of measured genotypes to within- and between family components. For every person i having sib j define between-family component as

$$(g_b)_i = \frac{g_i + g_j}{2}$$

and within-family component as

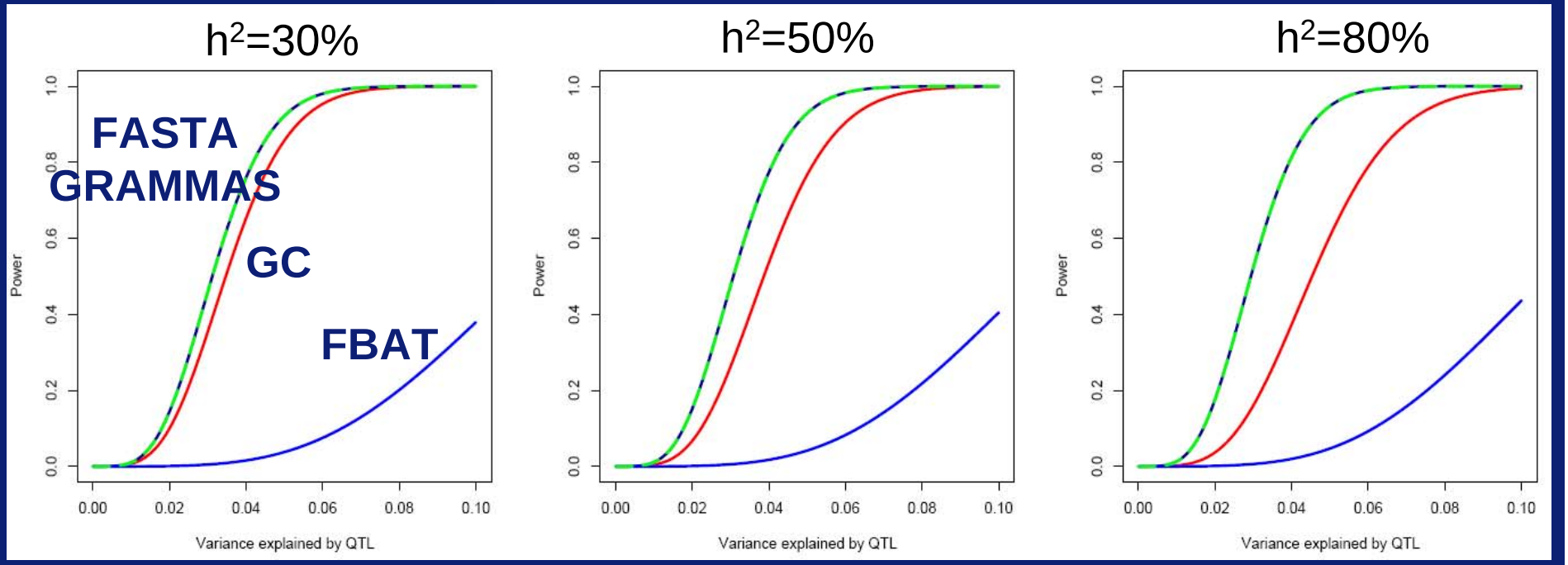
$$(g_w)_i = g_i - (g_b)_i$$

Expected trait value

$$E[x_i] = \mu + a_b \cdot (g_b)_i + a_w \cdot (g_w)_i$$

TDT

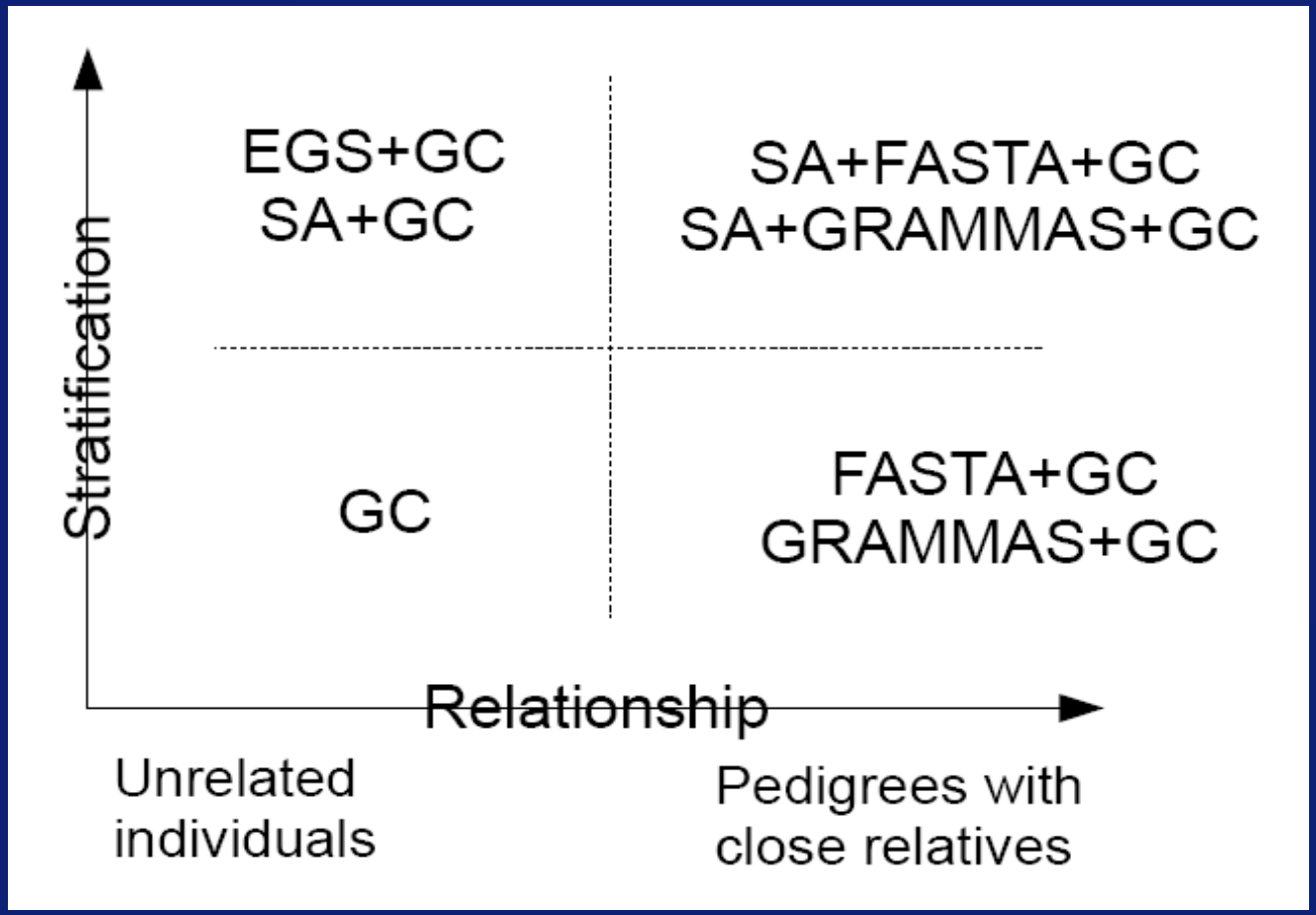
- Never use in homogeneous population
- You will loose 30-75% of NCP (=sample=money)



Relatives from heterogeneous population?

- No systematic analysis TDT vs Yu yet
- All lines point that TDT should be no more powerful than a combination of SA and FASTA/GRAMMS
- Use TDT only if strata can not be identified

Summary of analysis with stratification



Legend

- EGS = EIGENSTRAT (Price et al.)
- FASTA = Family-Based Score Test for Association (Chen & Abecasis)
- GC = Genomic Control (Devlin & Roeder)
- GRAMMAS = Genome-wide Rapid Association using Mixed Models and Score test (Aulchenko et al.)
- SA = Structured Association