GWA analysis of family studies

Yurii Aulchenko
Erasmus MC Rotterdam
27.08.2009

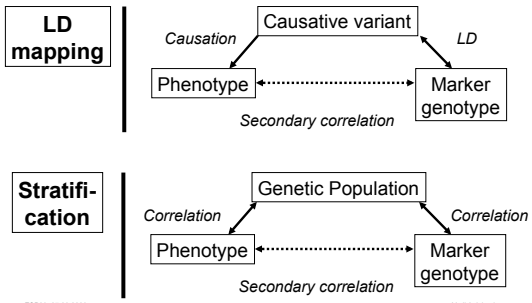
Outline

- **Reasons for genetic association**
- (Genome-Wide) Association analysis in pedigrees
- Conclusions

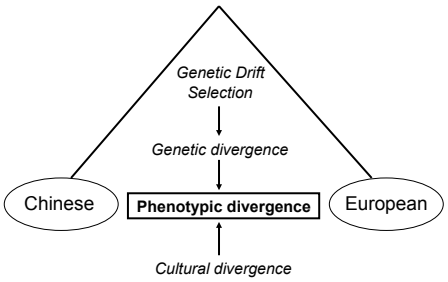
Benefits of studies of genetically isolated populations / family data

- Power to detect genes with rare variation
  - Variants with 0.0001 < MAF < 0.01
    - May become more frequent compared to general population, providing high power
    - Replication issue: consortia of genetically isolated populations
  - "Singleton variants"
    - … in outbred
    - Isolated populations: few copies in close relatives
- Ability to investigate complex genetic models
  - Parent-of-origin effects: imprinting, maternal
  - Parent x Child genotype interactions

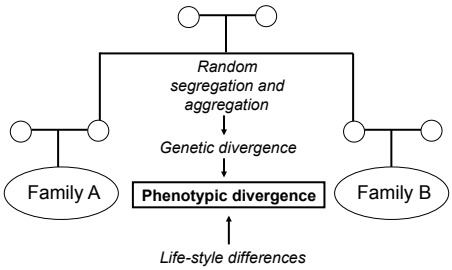
Confounding in genetic studies



Genetic origin is a major confounder



Pedigree is a major confounder



Am. J. Hum. Genet. 69:1146–1148, 2001

Erasmus MC Ezafung

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

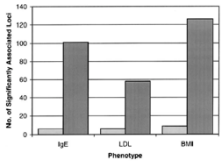


Figure 2 Number of significantly associated ($P < .01$) loci when pedigree structure is included (*lighter bars*) and when pedigree structure is not included (*darker bars*).

➔ High level of false positive signals

Ignore relationship, 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):

$$\lambda = \frac{Median\left(T_1^2, T_2^2, T_3^2, \ldots, T_M^2\right)}{0.455}, \quad \lambda \geq 1$$

Computation time ~ N

When GC does not work (well)?

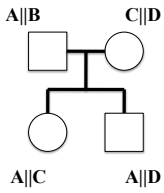
When stratification is large (say, $\lambda_{1000} > 1.1$) other, more powerful methods are to be used

GC assumes that stratification acts in the same manner across all loci

- This is not true for loci differentiated between population e.g. because of selection
- Such loci will still be falsely detected after GC correction

Relationship coefficient

- Two chromosomal regions are called Identical-by-descent (IBD) if they are copies of the same ancestral chromosomal region
- For a pair of people **relationship coefficient r** is the (expected) proportion of genome shared IBD
- Kinship coefficient = ½ relationship coefficient



Correlation between relatives

r	relationship
0.5 (½)	parent-offspring
0.25 (¼)	grandparent-grandchild
0.125 (⅛)	great grandparent-great grandchild
1	identical twins
0.5 (½)	full siblings
0.25 (¼)	half siblings
0.125 (⅛)	first cousins
0.03125 (1/32)	second cousins[2]

- If a trait is controlled by genes only (heritability, h^2=1)
  - Identical twins would have exactly the same trait value (cor = 1)
  - Correlation between the phenotypes of sibs would be 0.5
  - For arbitrary relatives correlation would be **r**

If the proportion of trait's variance explained by genes is h^2

- Correlation between phenotypes of identical twins would be h^2
- Correlation between the phenotypes of sibs would be 0.5 h^2
- For arbitrary relatives correlation would be **r** h^2

Mixed (polygenic) model

Linear Mixed Model (LMM) where the vector of quantitative phenotype *Y* is modeled as

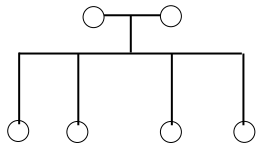
$$Y = \mu + B\,g + G + e$$

g: genotype indicator vector g_i in {0,1,2}
B: additive affect of the allele
e: is random residual effect ~ MVN(**0**, $\boldsymbol{I}\sigma_e^2$)
I: identity matrix
G: is random polygenic effect ~ MVN(**0**, $\boldsymbol{\Phi}\,\sigma_G^2$)
$\boldsymbol{\Phi}$: relationship matrix

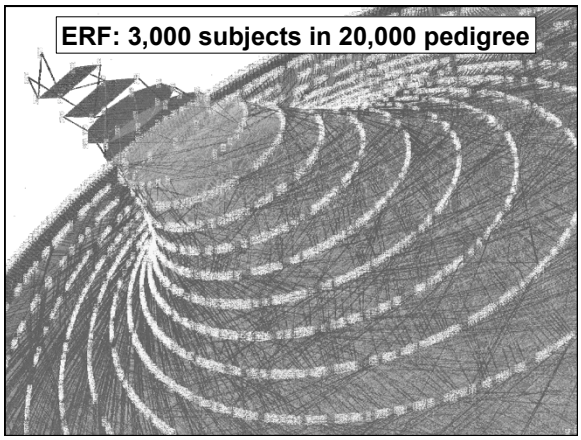
Maximum Likelihood (ML) or Restricted ML (REML)

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

Nuclear pedigree



ERF: 3,000 subjects in 20,000 pedigree



Analysis of large complex pedigrees

- Time required for GWA scan (2.5 million SNPs in 3,000 people)
- ML: 20 minutes per test => 95 years
- REML: 3-5 times faster

FASTA

FAmily Score Test for Association
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

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

(c) Apply GC afterwards if λ > 1

Computation time ~ N^2+N; no permutation testing

Chen & Abecasis, 2007

GRAMMAS

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

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

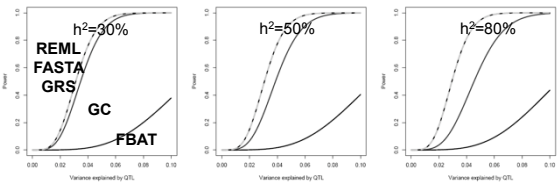
(d) Apply GC (λ expected to be < 1)

Computation time ~ N; permutation testing possible

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:

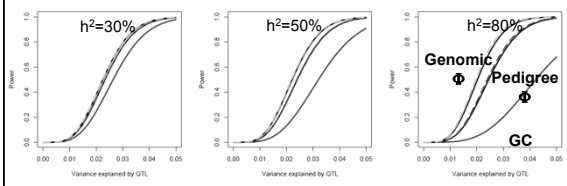
$$f_{ij} = \frac{1}{n}\sum_{k=1}^{n}\frac{(g_{ik} - p_k)(g_{jk} - p_k)}{p_k(1-p_k)}$$

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

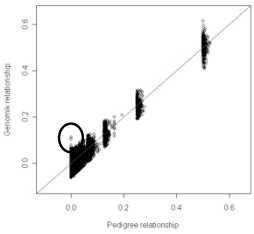
Genomic vs. Pedigree kinship

- 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 expected proportion of genome shared
  - Genomic relationship may better estimate true sharing



Testing association with binary trait in a general pedigree

- Bourgain et al., AJHG, 2003
  - Comparison of allele frequency between cases and controls
  - χ^2_{corr} ~ Genomic Control
  - CC-QLS (case-control quasi-likelihood score test)
    - Frequency is estimated under the null, using BLUP
    - Score test is performed
- William Astle, David Balding
  - Faster, more flexible methods based on Mixed Model