Estimating “Heritability” using Genetic Data

David Evans
University of Queensland

The Majority of Heritability for Most Complex Traits and Diseases is Yet to Be Explained

NEWS FEATURE PERSONAL GENOMES NATURE|Vol 456|6 November 2008

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Maher (2009) Nature

Places the Missing Heritability Could be Hiding

- In the form of common variants of small effect scattered across the genome
- In the form of low frequency variants only partially tagged by common variants
- Estimates of heritability from twin models are inflated (GASP!!!)

http://www.complextraitgenomics.com/software/gcta/

http://www.complextraitgenomics.com/softwar

GCTA Homepage

GCTA

a tool for Genome-wide Complex Trait Analysis

Overview

Download

Tutorial

FAQ

Options

1. Input and output
2. Data management
3. Estimation of the genetic relationships
4. Manipulation of the genetic relationship matrix
5. Principal component analysis
6. Estimation of the variance explained by all the SNPs
7. Estimation of the LD structure
8. GWAS Simulation
9. Raw genotype data

Overview

New version 1.24, more options and much faster!

GCTA Forum http://gcta.freeforums.net

GCTA (Genome-wide Complex Trait Analysis) was originally designed to estimate the proportion of phenotypic variance explained by genome- or chromosome-wide SNPs for complex traits (the GREML method), and has subsequently extended for many other analyses to better understand the genetic architecture of complex traits. GCTA was developed by Jian Yang, Hong Lee, Mike Goddard and Peter Visscher and is maintained in Peter Visscher's lab at the University of Queensland. GCTA currently supports the following functionalities:

- Estimate the genetic relationship from genome-wide SNPs;
- Estimate the inbreeding coefficient from genome-wide SNPs;
- Estimate the variance explained by all the autosomal SNPs;
- Partition the genetic variance onto individual chromosomes;
- Estimate the genetic variance associated with the X-chromosome;
- Test the effect of dosage compensation on genetic variance on the X-chromosome;
- Predict the genome-wide additive genetic effects for individual subjects and for individual SNPs;

http://gcta.freeforums.net/ | GCTA Homepage | Home | GCTA Software For...

GCTA Software Forum

Home Help Search

Welcome Guest. Please Login or Register.

GCTA Software Forum > Home >

General

	Board	Threads	Posts	Last Post
	GCTA Discussion Board Answers & Questions about GCTA analyses. Moderator: Jian Yang	30	100	Interpretation of GCTA results by kc **2 hours ago**

Legend

New Posts No New Posts

Forum Information & Statistics

Threads and Posts
Total Threads: 30 Total Posts: 100
Last Updated: Interpretation of GCTA results by kc (**2 hours ago**)
Recent Threads - Recent Posts - RSS Feed

Members
Total Members: 31
Newest Member: julio
Most Users Online: 21 (*Feb 19, 2014 at 7:23am*)
View today's birthdays

Users Online
0 Staff, 0 Members, 2 Guests.

nature genetics

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

ARTICLE

Estimating Missing Heritability for Disease from Genome-wide Association Studies

Sang Hong Lee,[1] Naomi R. Wray,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1,*]

REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

GCTA- The Mixed Model Framework

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$$

$$\underset{(n \times 1)}{\begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}} = \underset{(n \times m)}{\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}} \underset{(m \times 1)}{\begin{bmatrix} \beta_1 \\ \cdots \\ \beta_m \end{bmatrix}} + \underset{(n \times k)}{\begin{bmatrix} w_{11} & \cdots & w_{1k} \\ \cdots & \cdots & \cdots \\ w_{n1} & \cdots & w_{nk} \end{bmatrix}} \underset{(k \times 1)}{\begin{bmatrix} u_1 \\ \cdots \\ u_k \end{bmatrix}} + \underset{(n \times 1)}{\begin{bmatrix} \varepsilon_1 \\ \cdots \\ \varepsilon_k \end{bmatrix}}$$

where:

y is a vector of phenotypes

X contains covariates

β contains fixed effects regression coefficients

W contains standardized genotype dosages

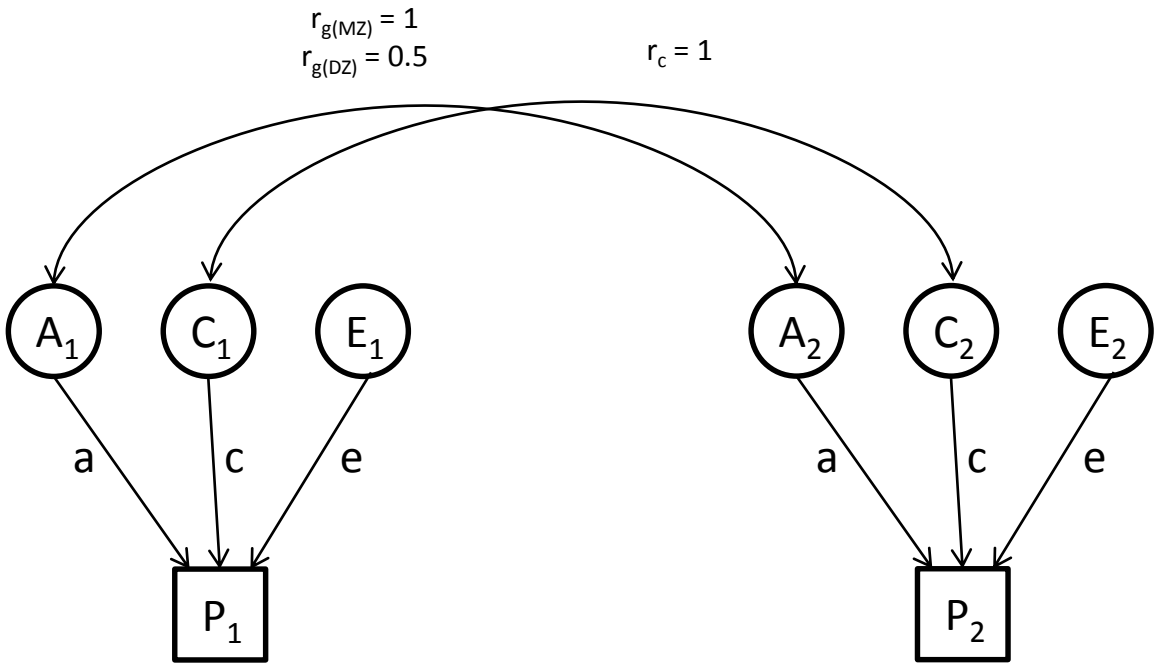
u contains random effects coefficients

k is number of SNPs

m is number of covariates

n is number of individuals

The Classical Twin Design



$$P_1 = aA_1 + cC_1 + eE_1$$

$$P_2 = aA_2 + cC_2 + eE_2$$

$$V_{MZ} = \begin{matrix} a^2 + c^2 + e^2 & a^2 + c^2 \\ a^2 + c^2 & a^2 + c^2 + e^2 \end{matrix}$$

$$V_{DZ} = \begin{matrix} a^2 + c^2 + e^2 & \tfrac{1}{2}a^2 + c^2 \\ \tfrac{1}{2}a^2 + c^2 & a^2 + c^2 + e^2 \end{matrix}$$

Expected Covariance Matrix Twin Pairs (AE Model)

$$\mathbf{V} = \mathbf{R}\sigma^2_A + \mathbf{I}\sigma^2_E$$

$$\underset{(2 \times 2)}{\begin{bmatrix} \sigma^2_1 & \sigma_{12} \\ \sigma_{21} & \sigma^2_2 \end{bmatrix}} = \underset{(2 \times 2)}{\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}} \cdot \sigma^2_A + \underset{(2 \times 2)}{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}} \cdot \sigma^2_E = \underset{(2 \times 2)}{\begin{bmatrix} \sigma^2_A + \sigma^2_E & r\,\sigma^2_A \\ r\sigma^2_A & \sigma^2_A + \sigma^2_E \end{bmatrix}}$$

V is the expected phenotypic covariance matrix

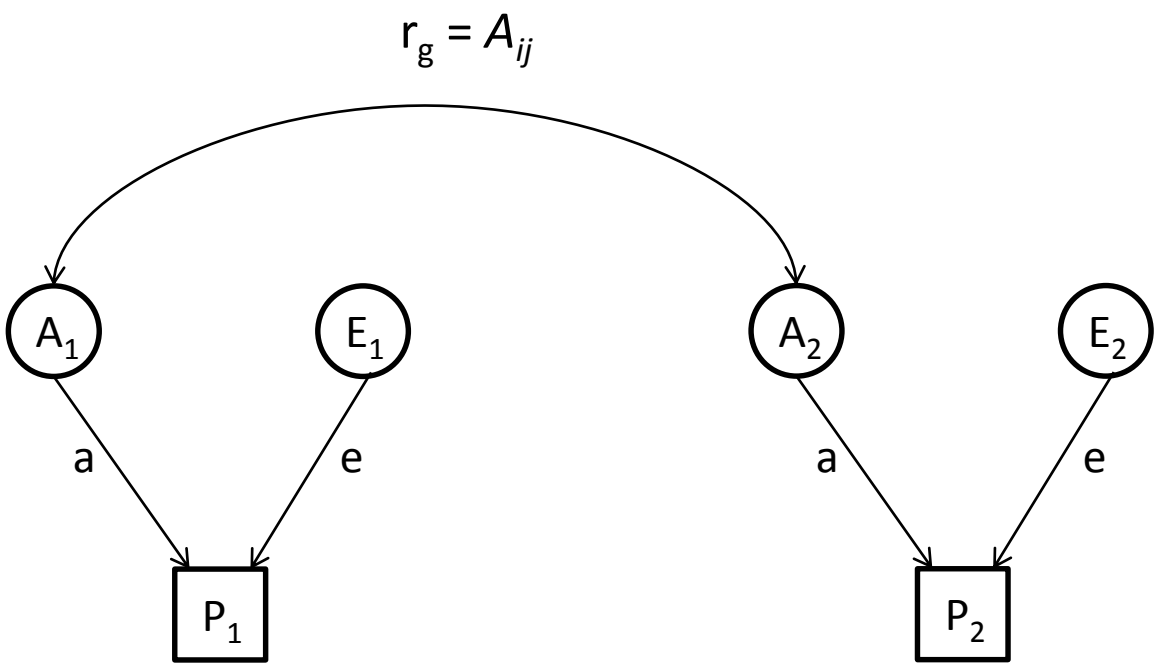
$\boldsymbol{\sigma^2_A}$ is the additive genetic variance

$\boldsymbol{\sigma^2_E}$ is the unique environmental variance

R is a matrix containing twice the kinship coefficient (r = 1 for MZ, r = 0.5 for DZ))

I is an identity matrix

The GCTA Design- Unrelateds



$$P_1 = aA_1 + eE_1$$

$$P_2 = aA_2 + eE_2$$

$$V = \begin{matrix} a^2 + e^2 & A_{ij}a^2 \\ A_{ij}a^2 & a^2 + e^2 \end{matrix}$$

Expected Covariance Matrix - Unrelateds

$$\mathbf{V} = \mathbf{A}\sigma^2_g + \mathbf{I}\sigma^2_e$$

$$\underset{(n \times n)}{\begin{bmatrix} \sigma^2_1 & \cdots & \sigma_{1n} \\ \cdots & \cdots & \cdots \\ \sigma_{n1} & \cdots & \sigma^2_n \end{bmatrix}} = \underset{(n \times n)}{\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a^2_{nn} \end{bmatrix}} \cdot \sigma^2_g + \underset{(n \times n)}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \cdot \sigma^2_e$$

V is the expected phenotypic covariance matrix

$\boldsymbol{\sigma^2_g}$ is the additive genetic variance

$\boldsymbol{\sigma^2_e}$ is the unique environmental variance

A is a **GRM** containing average standardized genome-wide IBS between individual *i* and *j*

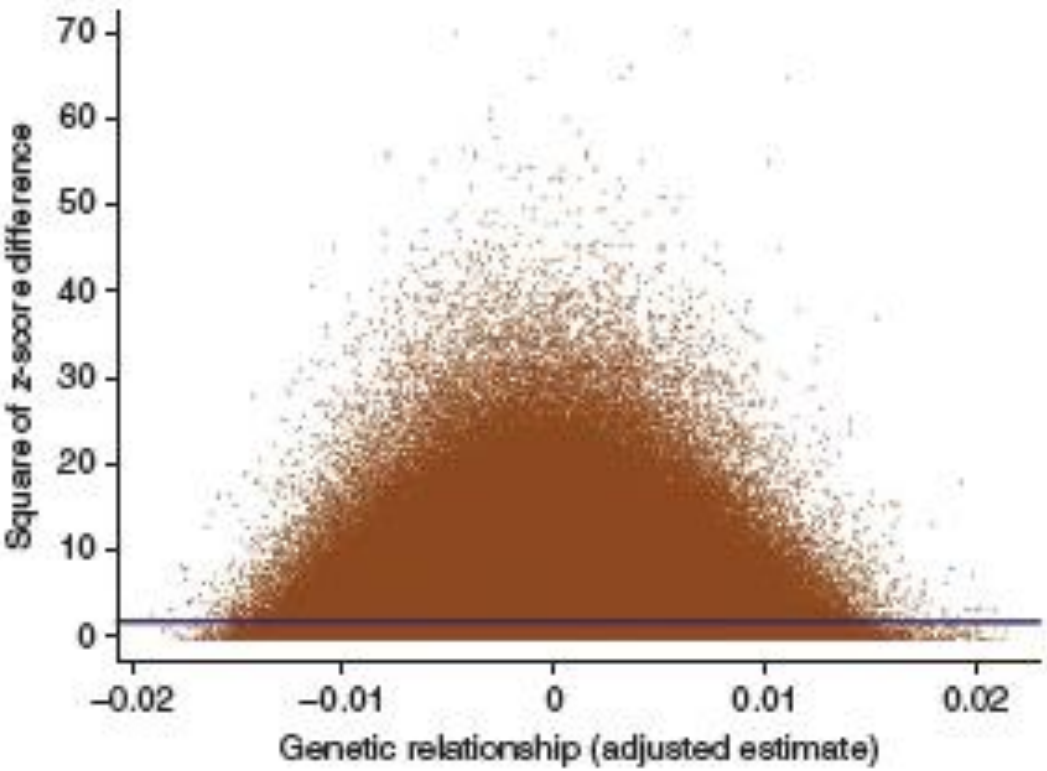
I is an identity matrix

GCTA- Genetic Relationship Matrix

$$A_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}.$$

where x_{ij} is the number of copies of the reference allele for the ith SNP of the jth individual and p_i is the frequency of the reference allele.

Intuitively...



- If a trait is genetically influenced, then individuals who are more genetically similar should be more phenotypically similar
- Can be thought of like a Haseman- Elston regression

GCTA Process

- Two step process
- Estimate GRM
  - Exclude one from each pair of individuals who are >2.5% IBS
- Estimate variance components via “REML”

GCTA- Some Results

Table 1 Estimates of the variance explained by all autosomal SNPs for height, BMI, vWF and QTi

		No PC[a]		10 PCs[b]			
Trait	n	h_G^2 (s.e.)[c]	P	h_G^2 (s.e.)	P	Heritability[d]	GWAS[e]
Height	11,576	0.448 (0.029)	4.5×10^{-69}	0.419 (0.030)	7.9×10^{-48}	80–90%[32]	~10%[23]
* BMI	11,558	0.165 (0.029)	3.0×10^{-10}	0.159 (0.029)	5.3×10^{-9}	42–80%[25,26]	~1.5%[14]
vWF	6,641	0.252 (0.051)	1.6×10^{-7}	0.254 (0.051)	2.0×10^{-7}	66–75%[33,34]	~13%[15]
QTi	6,567	0.209 (0.050)	3.1×10^{-6}	0.168 (0.052)	5.0×10^{-4}	37–60%[35,36]	~7%[16]

Adapted from
Yang et al. (2011) Nat Genet

GCTA Interpretation

- GCTA does not estimate “heritability”
- GCTA does not estimate the proportion of trait variance due to common SNPs
- GCTA tells you nothing definitive about the number of variants influencing a trait, their size or their frequency

GCTA- Some Assumptions

- The GRM accurately reflects the underlying causal variants
- Underlying variants explain the same amount of variance
  - Relationship between MAF and effect size
- Independent effects
  - Contributions to h^2 overestimated by causal variants in regions of high LD and underestimated in regions of low LD

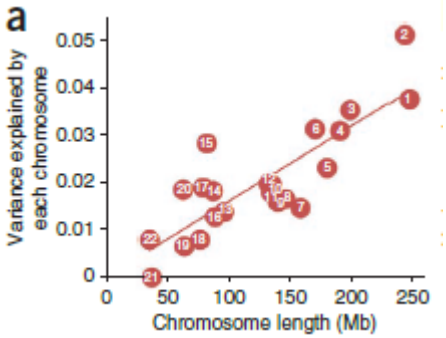
Extending the Model - Genome Partitioning

$$\mathbf{V} = \sum_{c=1}^{22} \mathbf{A}_c \sigma^2_{g,c} + \mathbf{I}\sigma^2_e$$

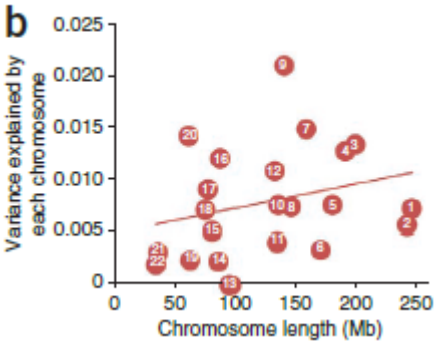
- The genetic component can be partitioned further into e.g. different chromosomes, genic vs non-genic regions
- A different GRM (A_c) needs to be computed for each of these components

Extending the Model - Genome Partitioning

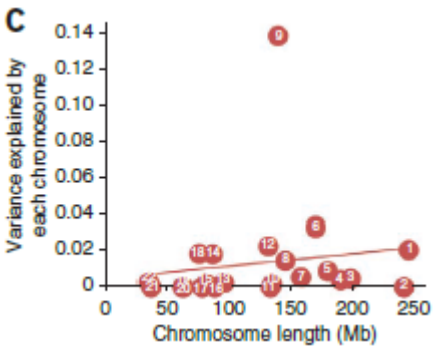
Height



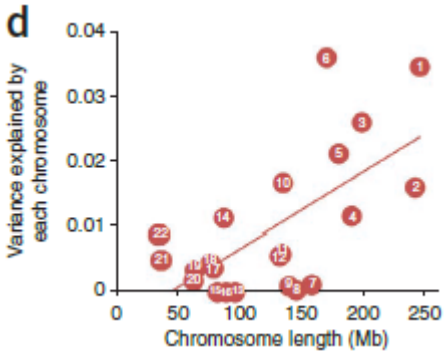
BMI



Von Willebrand Factor



QT Interval



Adapted from
Yang et al. (2011) Nat Genet

Extending the Model: Gene-Environment Interaction

$$\mathbf{V} = \mathbf{A}_g\sigma^2_g + \mathbf{A}_{ge}\sigma^2_{ge} + \mathbf{I}\sigma^2_e$$

- $A_{ge} = A_g$ for pairs of individuals in the same environment and Age = 0 for pairs of individuals in different environments
- "Environmental" factors could be sex or medical treatment for example

Extending the Model - Binary Traits

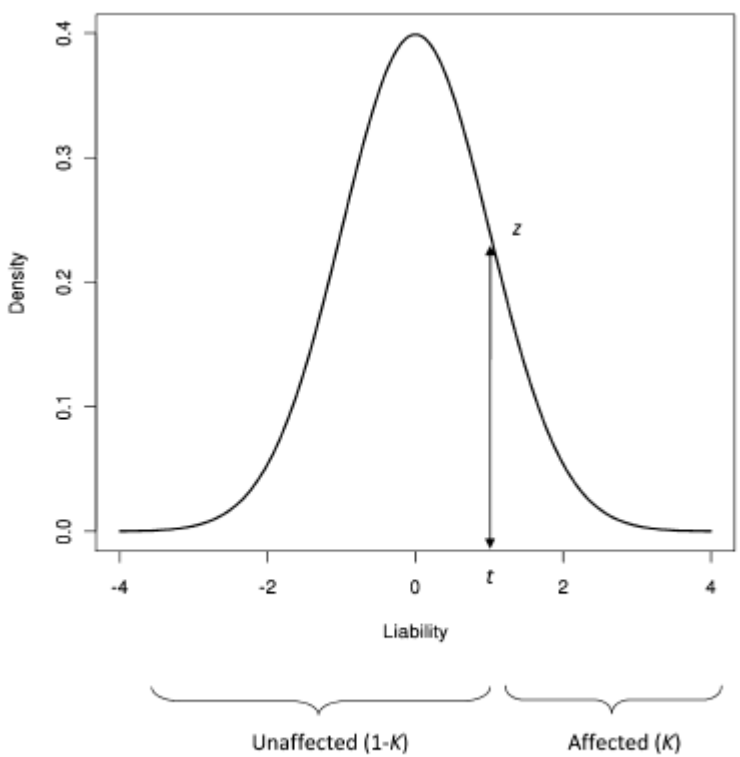


Figure 1. The Liability Threshold Model for a Disease Prevalence of *K*

- Assume an underlying normal distribution of liability
- Transform estimates from the observed scale to the liability scale

$$h_l^2 = h_o^2 K(1-K)/z^2 .$$

Extending the Model – BinaryTraits

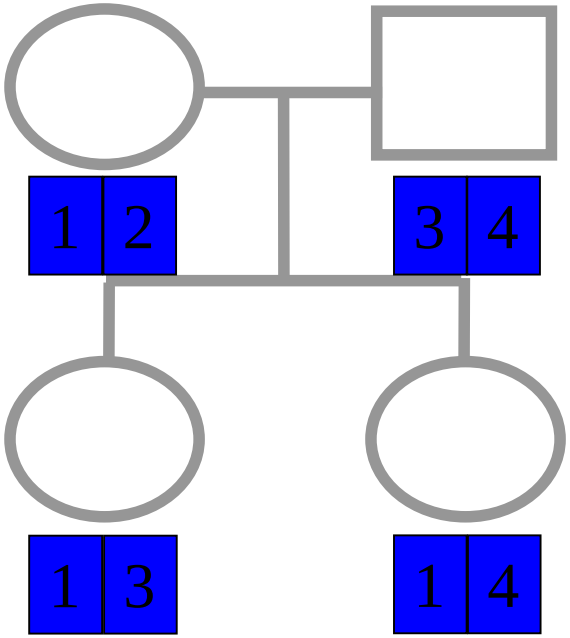
- Estimate GRM
  - Exclude one from each pair of individuals who are >2.5% IBS
- Estimate variance components via “REML”
- Transform from observed scale to liability scale
- Adjust estimates to take account of ascertainment (i.e. the fact that case-control proportions are not the same as in the population)

Extending the Model – Bivariate Association

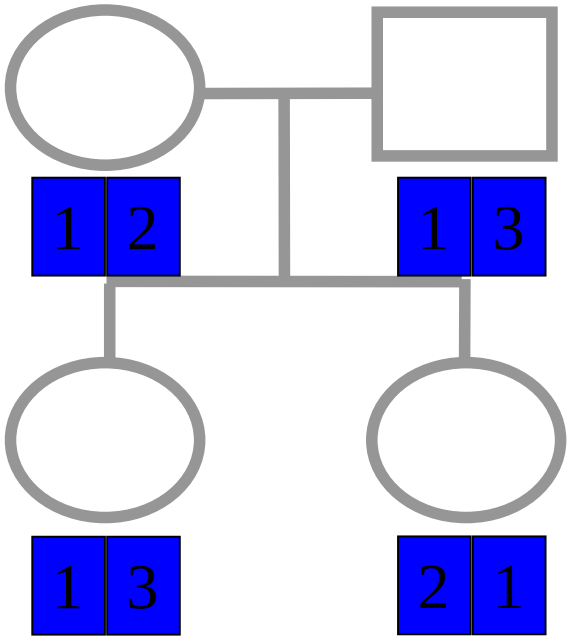
- Estimate the genetic and residual correlation between different traits/diseases

- Individuals need not be measured on both traits

Extending the Model - Identity By Descent (IBD)



Identical by Descent



Identical by state only

Two alleles are IBD if they are descended from the same ancestral allele

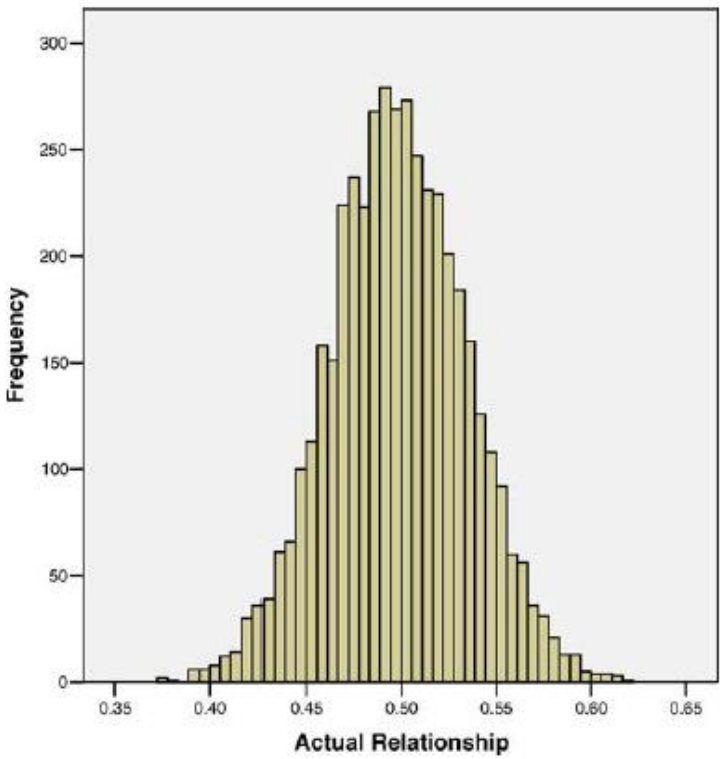
Extending the Model – IBD

$$\mathbf{V} = \boldsymbol{\pi}_{\mathrm{IBD}}\sigma^2_{\mathrm{A}} + \mathbf{C}\sigma^2_{\mathrm{C}} + \mathbf{I}\sigma^2_{\mathrm{e}}$$

$$\underset{(n \times n)}{\begin{bmatrix} \sigma^2_1 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma^2_2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma^2_3 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma^2_4 \end{bmatrix}} = \underset{(n \times n)}{\begin{bmatrix} 1 & \pi_{12} & 0 & 0 \\ \pi_{21} & 1 & 0 & 0 \\ 0 & 0 & 1 & \pi_{34} \\ 0 & 0 & \pi_{43} & 1 \end{bmatrix}} . \sigma^2_{\mathrm{A}} + \underset{(n \times n)}{\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}} . \sigma^2_{\mathrm{C}} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} . \sigma^2_{\mathrm{e}}$$

USE IBD variation within SIBS to estimate heritability

- Use variation in genetic sharing within a relative type rather than different types of relatives
- Gets around problem of the "Equal Environment" assumption in twin studies



Extending the Model – IBD

- Estimate GRM
  - Exclude one from each pair of individuals who are >2.5% IBS
- Estimate variance components via “REML”
- Transform from observed scale to liability scale
- Adjust estimates to take account of ascertainment (i.e. the fact that case-control proportions are not the same as in the population)

Application to Height & BMI

Table 1. Summary of Estimates of BMI and Height Heritabilities from Realized Relationships for All Cohorts

Cohort	QISPs	r	h^2 (SE)	c^2 (SE)	r[a]	h^2 (SE)[b]	c^2 (SE)[c]
QIMR	9,585	0.26	0.76 (0.20)	0.00 (0.10)	0.39	0.80 (0.22)	0.00 (0.10)
Framingham Heart study	4,607	0.30	0.00 (0.34)	0.27 (0.17)	0.47	0.72 (0.28)	0.10 (0.14)
TWINGENE	2,722	0.24	0.00 (0.47)	0.24 (0.24)	0.50	0.75 (0.35)	0.12 (0.18)
Netherlands Twin Registry	1,819	0.37	0.78 (0.47)	0.00 (0.23)	0.48	0.00 (0.42)	0.49 (0.22)
TwinsUK	1,507	0.41	0.31 (0.55)	0.25 (0.28)	0.54	0.56 (0.46)	0.26 (0.23)
Total	20,240	0.29	0.42 (0.17)	0.10 (0.08)	0.44	0.69 (0.14)	0.08 (0.07)

[a]Correlation between phenotypes of QISPs after adjustment for fixed effects.
[b]Heritability estimates.
[c]Proportion of the phenotypic variance attributed to common environmental variance.

OPEN ACCESS Freely available online PLOS | GENETICS

Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits

Noah Zaitlen[1]*, Peter Kraft[2,3,4], Nick Patterson[4], Bogdan Pasaniuc[5], Gaurav Bhatia[2,3,4], Samuela Pollack[2,3,4], Alkes L. Price[2,3,4]*

1 Department of Medicine, Lung Biology Center, University of California San Francisco, San Francisco, California, United States of America, **2** Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America, **3** Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of America, **4** Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **5** Interdepartmental Program in Bioinformatics Pathology and Laboratory Medicine, University of California Los Angeles, Los Angeles, California, United States of America

Table 1. Narrow-sense heritability estimated from IBD (h^2_{IBD}) and thresholding IBS ($h^2_{IBS>0.05}$) for 11 quantitative traits.

Quantitative trait	N[a]	h^2_{IBD}	s.e.	$h^2_{IBS>0.05}$	s.e.	h^2_{Pub}
Body Mass Index (kg/m^2)	20000	0.422	0.018	0.433	0.018	0.4–0.6 [6]
Cholesterol High Density Lipoprotein	19977	0.446	0.017	0.457	0.018	0.5 [6]
Cholesterol Low_Density Lipoprotein	4547	0.196	0.062	0.198	0.063	0.376 [42]
Height (cm)	20000	0.691	0.016	0.704	0.016	0.8 [6]
Menarche Age (years)	15150	0.443	0.022	0.454	0.022	0.4–0.7 [43]
Menopause Age (years)	5540	0.400	0.047	0.409	0.048	0.4–0.6 [44]
Monocyte White Blood Cell	9651	0.343	0.032	0.351	0.032	0.378 [42]
Waist-Hip Ratio	5538	0.181	0.037	0.187	0.038	0.3–0.6 [45]
Sex Ratio of offspring	15000	0.026	0.017	0.021	0.018	-
Total Children	15000	0.103	0.017	0.111	0.018	-
Recombination Rate	10259	0.099	0.023	0.110	0.030	-

[a]N is the number of individuals used in the analysis of each phenotype. h^2_{Pub} are previously published estimates of heritability from different populations.

doi:10.1371/journal.pgen.1003520.t001

Idea…

- It should be obvious now, that pretty much all the models that we have touched on this week can be expressed within this GCTA framework
- Yet only a small proportion of these have been parameterized in GCTA
- Considerable scope exists for parameterization of the GCTA framework in Mx…