Estimation of Heritability in Humans

Jian Yang
Queensland Brain Institute
jian.yang@uq.edu.au

Lecture 11am to 12:50pm, 83-C415

- Genetics
- Statistics: correlation, ANOVA
- Tools: R, Excel

Heritability

- Proportion of phenotypic variation that is due to genetic factors (e.g. genes / genetic variants)

- Specific to a population
  - Allele frequencies
  - Effects of gentic variants
  - Environmental factors
  - …

Heritability

- A trait is heritable if more closely related individuals have more similar phenotypes

- The stronger the relationship between relatedness and phenotypic similarity, the more heritable the trait is.

Estimating heritability

The simplest genetic model:

$Y = G + E$

Y = phenotype

G = genetic value

E = residual

$H^2 = \mathrm{var}(G) / \mathrm{var}(Y)$

Heritability Estimation

- Aim to disentangle genetic and environmental influences on trait variation
- Resemblance between relatives
  - Shared genes
  - Shared environmental factors
- Differences between relatives
  - Non-shared genes
  - Unique environmental factors

Clones

$Y_{j1} = G + E_{j1}$

$Y_{j2} = G + E_{j2}$

Assuming E_{j1} and E_{j2} are independent

$\mathrm{Cov}(Y_{j1}, Y_{j2}) = \mathrm{cov}(G + E_{j1,}\ G + E_{j2}) = \mathrm{var}(G)$

$$\begin{aligned}\mathrm{Cor}(Y_{j1}, Y_{j2}) &= \mathrm{Cov}(Y_{j1}, Y_{j2}) / [\sigma(Y_{j1})\ \sigma(Y_{j2})] \\ &= \mathrm{var}(G) / \mathrm{var}(Y) \\ &= H^2\end{aligned}$$

Twin Design

- A “natural experiment”
  - Gets around inability to use breeding experiments in humans!
- Relatively high frequency
  - ~1 in 80 births in Australia are twins
  - Ratio of MZ/DZ ~1:2 in Caucasians

MZ twins: E_{j1} and E_{j2} are dependent

$Y_{j1} = G + E_{j1}$

$Y_{j2} = G + E_{j2}$

If E_{j1} and E_{j2} are dependent

$$\begin{aligned}
\mathrm{Cov}(Y_{j1}, Y_{j2}) &= \mathrm{cov}(G + E_{j1,}\ G + E_{j2}) \\
&= \mathrm{var}(G) + \mathrm{cov}(E_{j1}, E_{j2}) \\
&> \mathrm{var}(G)
\end{aligned}$$

H^2 overestimated!

A more complicated but realistic model

$$
\begin{aligned}
y &= \mu + G + E \\
&= \mu + (A + D + I) + E_c + E_s
\end{aligned}
$$

$$
\mathrm{var}(y) = V_G + V_E
$$

$$
= \boxed{V_A + V_D + V_I} + \boxed{V_{Ec} + V_{Es}}
$$

Falconer & Mackay, Chapters 7 & 8

MZ covariance

$$\mathrm{Cov}(y_{i1},y_{i2}|MZ) = \mathrm{Cov}(MZ)$$

$$= V_G + V_{Ec(MZ)}$$

$$= V_A + V_D + V_I + V_{Ec(MZ)}$$

DZ covariance

$\mathrm{Cov}(y_{i1}, y_{i2} | DZ) = \mathrm{Cov}(DZ)$

$= \frac{1}{2} V_A + \frac{1}{4} V_D + \frac{1}{4} V_{AA} + \ldots + V_{Ec(DZ)}$

Falconer & Mackay, Chapters 9

Example: Correlations

	Intelligence (IQ)
rMZ	0.81
rDZ	0.51

$\mathrm{Cov}(y_{i1},y_{i2}|MZ) = V_A + V_D + V_I + V_{Ec(MZ)}$

$\mathrm{Cov}(y_{i1},y_{i2}|DZ) = ½\, V_A + ¼\, V_D + ¼V_{AA}+\ldots$
$+ V_{Ec(DZ)}$

Luciano et al (2001) Intelligence 29:443

Analysing Twin Data

- Correlation
- One-way ANOVA
- (Maximum likelihood, structural equation modelling...)

Correlation

$\rho_{MZ} = \mathrm{cov}(MZ) / (\sigma_{y1} \sigma_{y2})$

$= h^2 + c^2 + \ldots$

$\rho_{DZ} = \mathrm{cov}(DZ) / (\sigma_{y1} \sigma_{y2})$

$= \frac{1}{2} h^2 + c^2 + \ldots$

Note: $\sigma^2_y = \sigma^2_{y1} = \sigma^2_{y2}$

ANOVA Overview

- Two separate ANOVAs for MZ and DZ twin pairs
  - Between-pairs and within-pairs components of variance
  - Assumes that trait has same variance in MZ and DZ twins

Linear Model

$$y_{ij} = \mu + b_i + w_{ij}$$

$$\sigma_y^2 = \sigma_b^2 + \sigma_w^2$$

- Balanced: j=1,2 for all groups
- *y*, *b* and *w* are random variables
- $H^2 = \sigma_b^2/\sigma_y^2$
  - Intra-Class Correlation = proportion of total variance attributable to differences between pairs
  - Very similar to direct correlation estimate...

$\sigma_b^2 = \sigma_G^2$

ANOVA table

Source	d.f.	MS	E(MS)
Between pairs	n-1	B	$\sigma^2{}_w + 2\sigma^2{}_b$
Within pairs	n(2-1)	W	$\sigma_w{}^2$

$$V_w = \sigma^2{}_w = E(MS)_W$$

$$V_b = \sigma^2{}_b = [E(MS)_B – E(MS)_W] / 2$$

Why Use ANOVA

- Ordering of pairs does not matter
- Can correct for other variables
  - Age
  - Sex
  - …
- Can test (some) assumptions

Assumption Testing

- Test of equality of variances

  $$F = \frac{MST_{MZ}}{MST_{DZ}}$$

  with $(2n_{MZ}-1,\ 2n_{DZ}-1)$ d.f.

- Test of genetic contribution to trait

  $$F = \frac{MSW_{DZ}}{MSW_{MZ}}$$

  with $(n_{DZ},\ n_{MZ})$ d.f. *[n = # pairs]*

Components of ANOVA

	V_b (Between pairs)	V_w (Within pairs)
MZ	$V_A + V_{Ec(MZ)}$	$V_{Es(MZ)}$
DZ	$\frac{1}{2}V_A + V_{Ec(DZ)}$	$\frac{1}{2}V_A + V_{Es(DZ)}$

Assumption #1:

We ignore the contribution of non-additive genetic variation

BUT!

Still too many unknowns (5) to be estimated from only 4 summary statistics

More Assumptions...

- Assume that environmental variances are equal for MZ and DZ:

$V_{Ec(MZ)} = V_{Ec(DZ)}$ $\quad$ $V_{Es(MZ)} = V_{Es(DZ)}$

	V_b (Between pairs)	V_w (Within pairs)
MZ	$V_A + V_{Ec}$	V_{es}
DZ	$\frac{1}{2}V_A + V_{Ec}$	$\frac{1}{2}V_A + V_{es}$

Variance components estimates

- $V_A = 2\,(V_{b(MZ)} – V_{b(DZ)})$
  $= 2\,[(V_A + V_{Ec}) – (½V_A + V_{Ec})]$
  $= V_A$

- $V_{ec} = 2\,V_{b(DZ)} – V_{b(MZ)}$
  $= [2\,(½V_A + V_{Ec})] – (V_A + V_{ec})$

The equal environments assumption

- We assume that environmental factors causing twin similarity operate at same level in MZ and DZ twins

- If MZ twins experience more similar environment than DZ twins, this will inflate $\hat{h}^2$

Summary of assumptions

- Total variance of the trait same for both types of twins
  - Var(MZ) = Var(DZ)

- Influence of non-additive genetic variation (dominance and epistasis) can be ignored

- Environmental sources of variance are the same in MZs and DZs
  - $V_{Ec(MZ)} = V_{Ec(DZ)}$ & $V_{Es(MZ)} = V_{Es(DZ)}$

Are twins representative?

- Assume twins are representative of the general population but possible that
  - Not genetically representative
    - Risk of congenital malformations
  - Not environmentally representative
    - Parental treatment
    - Sibling co-operation or competition
- Volunteer twin registries generally used so may not be representative of non-volunteers
  - May be especially problematic for some behavioural traits

Different study designs

- Family studies
  - Gene + environment confounded
  - Focus on relative pairs or all individuals
- MZ twins reared apart / Adoptions
  - Could remove environmental confounding
  - Atypical, possible selective placement

Relative Pair Correlations

- Assuming similarity is only due to additive effects....

Pair Type	Correlation
MZ	h^2
DZ	½ h^2
Parent – Offspring	½ h^2
Mid-Parent – Offspring	sqrt(½) h^2
Sib Pair	½ h^2
Half Sibs	¼ h^2
Grandparent - Grandchild	¼ h^2
Avuncular (Uncle - Nephew)	¼ h^2

Examples

- Morphological Measures

Fingerprint Ridges	~90%
Height	~80%
Baldness	~80%
BMI	~65%
Facial Traits	~50%
Birth Weight	~30%

Examples

- Diseases

Schizophrenia	~80%
Type I Diabetes	~80%
Macular Degeneration	~60%
Lupus	~50%
Coronary Heart Disease	~45%
Type II Diabetes	~25%

10 min break

Practical 2pm to 4:50pm, 83-C310

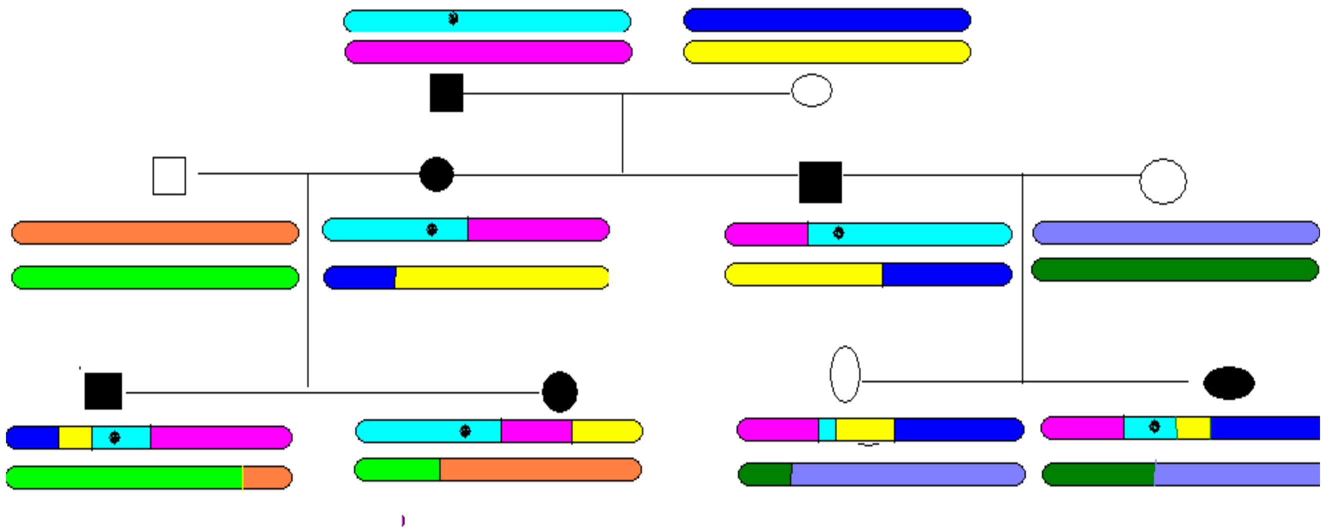
http://ctgg.qbi.uq.edu.au/teaching/UQQG/

Using Variation Within Pairs

- For some relationship pairs, there is variation in the amount of the genome shared

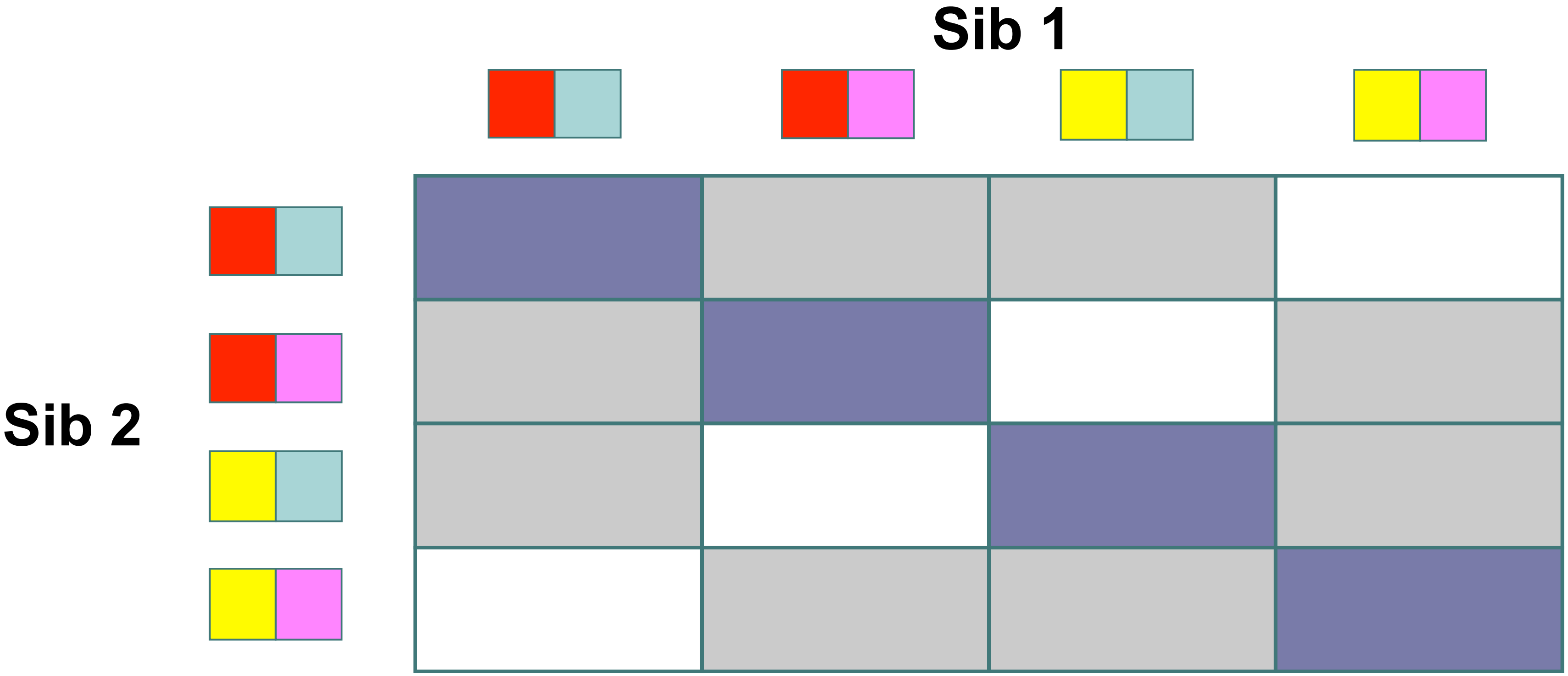
- Parent-offspring – always 50% sharing (ignoring inbreeding...)
- Sib-pairs – average of 50% sharing
  - ¼ IBD 2, ½ IBD 1, ¼ IBD 0

Chromosome Transmission



- Identity By Descent – IBD
- Related individuals share the same allele or haplotype

IBD – Identity By Descent



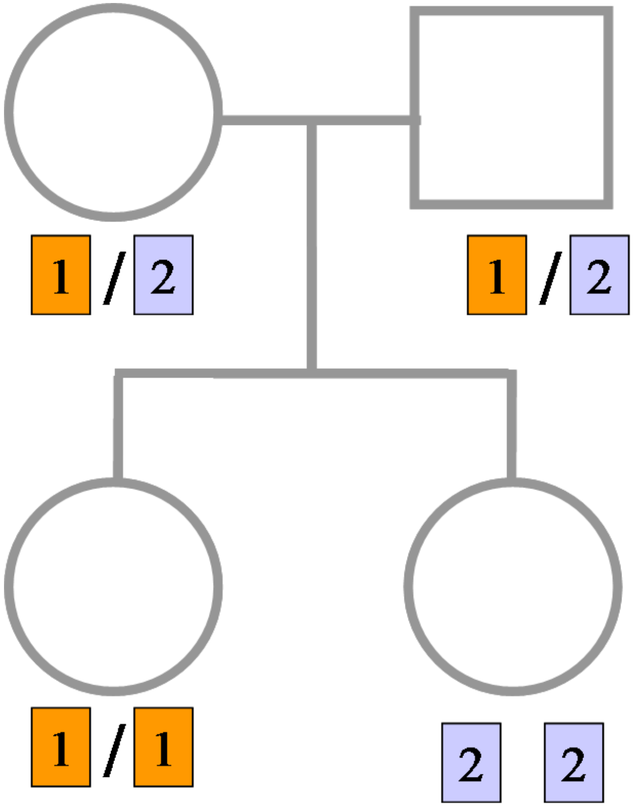
▨ 4/16 = 1/4 sibs share BOTH parental alleles IBD = 2

▨ 8/16 = 1/2 sibs share ONE parental allele IBD = 1

□ 4/16 = 1/4 sibs share NO parental alleles IBD = 0

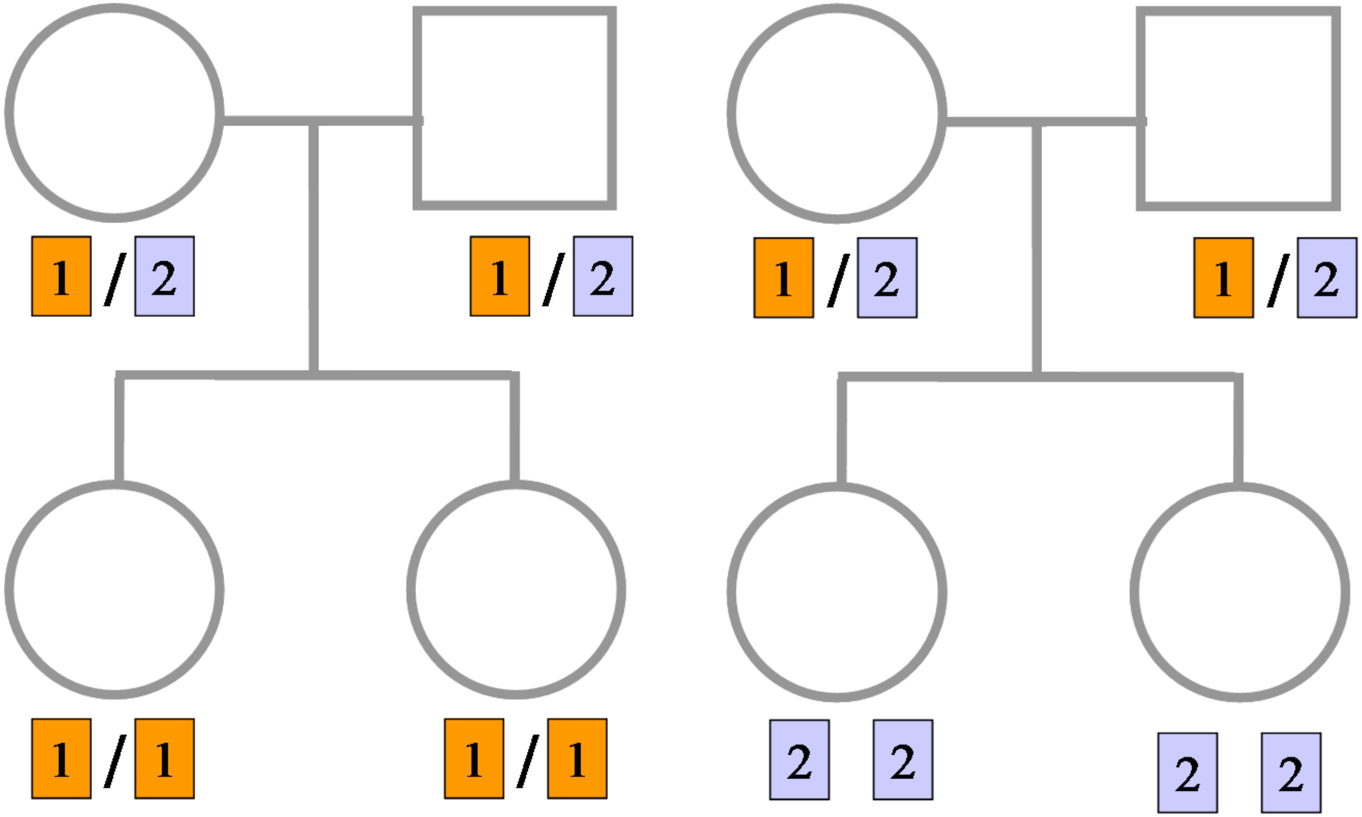
IBD – Identity By Descent

- Simple case: IBD = 0



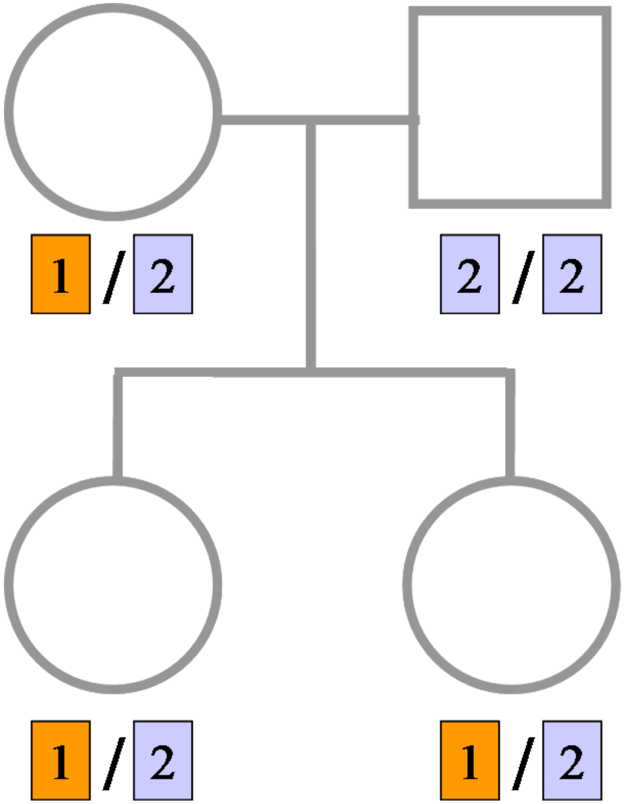
IBD – Identity By Descent

- More simple cases: IBD = 2



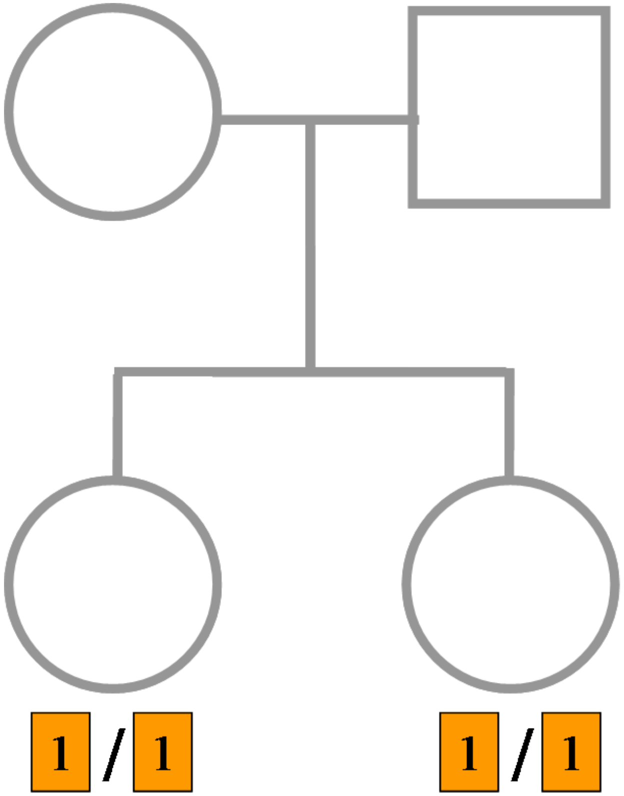
IBD – Identity By Descent

- Not so simple: 50% IBD 1, 50% IBD 2



IBD – Identity By Descent

- Complex case: IBD = ???



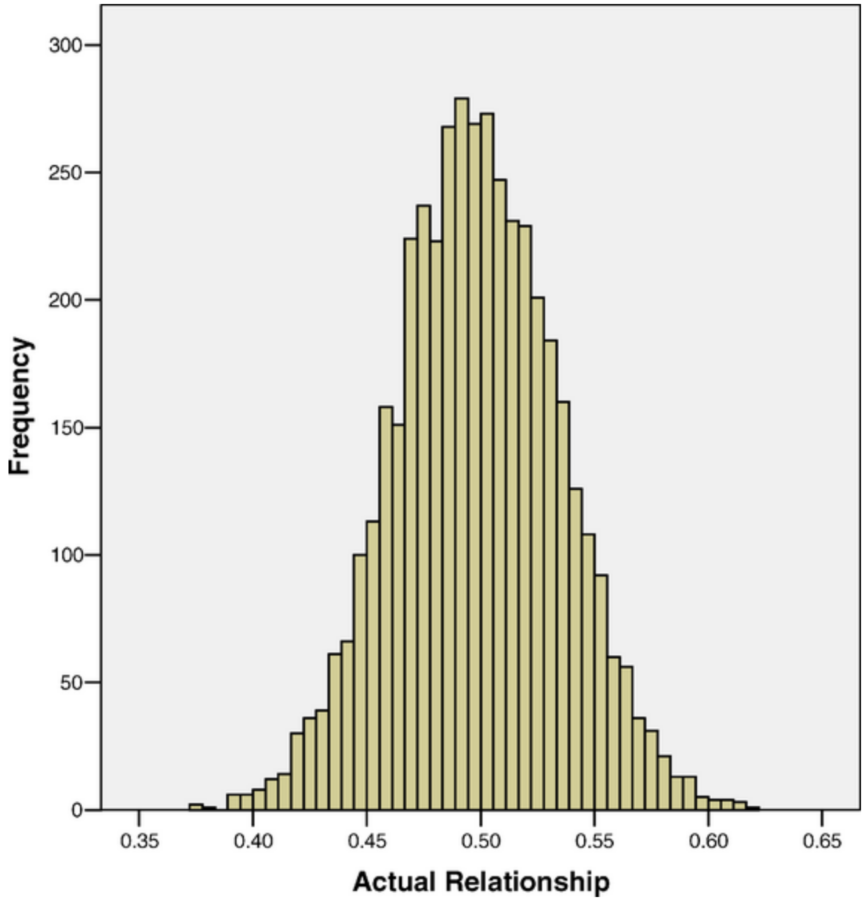
Estimating Relatedness

- Genotype a large number of markers across the genome

- Calculate IBD probabilities across the genome and take the average

- Genetic relatedness = P(IBD=2) + ½ P(IBD = 1)

Relatedness of Sib-Pairs

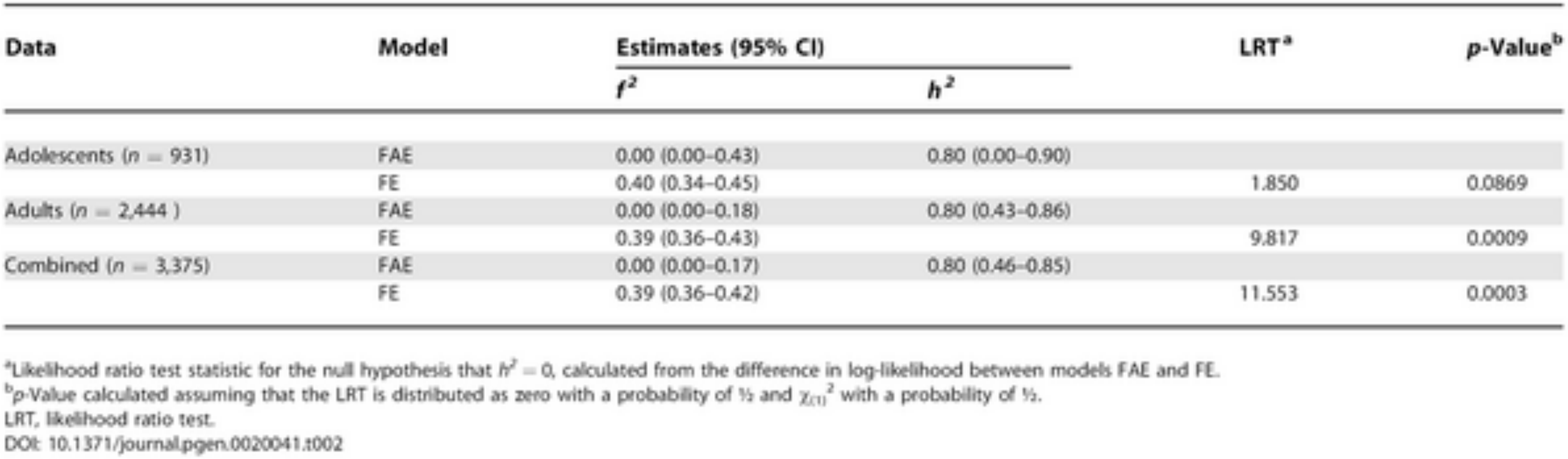


Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings
Visscher et al., PLoS Genet (2006) 2: e41

Heritability Within-Pairs

- Tests if more related people are more phenotypically similar
- Can use variation in relatedness within (e.g.) sib-pairs to estimate heritability

Example - Height



Data	Model	Estimates (95% CI)		LRT[a]	*p*-Value[b]
		f^2	h^2		
Adolescents ($n = 931$)	FAE	0.00 (0.00–0.43)	0.80 (0.00–0.90)		
	FE	0.40 (0.34–0.45)		1.850	0.0869
Adults ($n = 2,444$)	FAE	0.00 (0.00–0.18)	0.80 (0.43–0.86)		
	FE	0.39 (0.36–0.43)		9.817	0.0009
Combined ($n = 3,375$)	FAE	0.00 (0.00–0.17)	0.80 (0.46–0.85)		
	FE	0.39 (0.36–0.42)		11.553	0.0003

[a] Likelihood ratio test statistic for the null hypothesis that $h^2 = 0$, calculated from the difference in log-likelihood between models FAE and FE.

[b] *p*-Value calculated assuming that the LRT is distributed as zero with a probability of ½ and $\chi_{(1)}^2$ with a probability of ½.

LRT, likelihood ratio test.

DOI: 10.1371/journal.pgen.0020041.t002

Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings

Visscher et al., PLoS Genet (2006) 2: e41

Heritability Within Pairs

- Advantage
  - Using differences within a family means no assumptions are made about variation across families
- Disadvantage
  - Estimate has large variance
  - Requires very large numbers of pairs

Population Based Estimation

- “Unrelated” individuals from the population show differing amounts of genetic similarity.
- We can use these differences to estimate a “heritability”.
- Need to measure how related “unrelated” people are.

Genome-wide SNP Chips

Genome-wide SNP Chip

- Measure an individuals genotype at 100s of thousands / millions of SNP

- SNP = Single Nucleotide Polymorphism

- Look at “common” variation
  - Minor allele frequency > 0.05 (0.01)

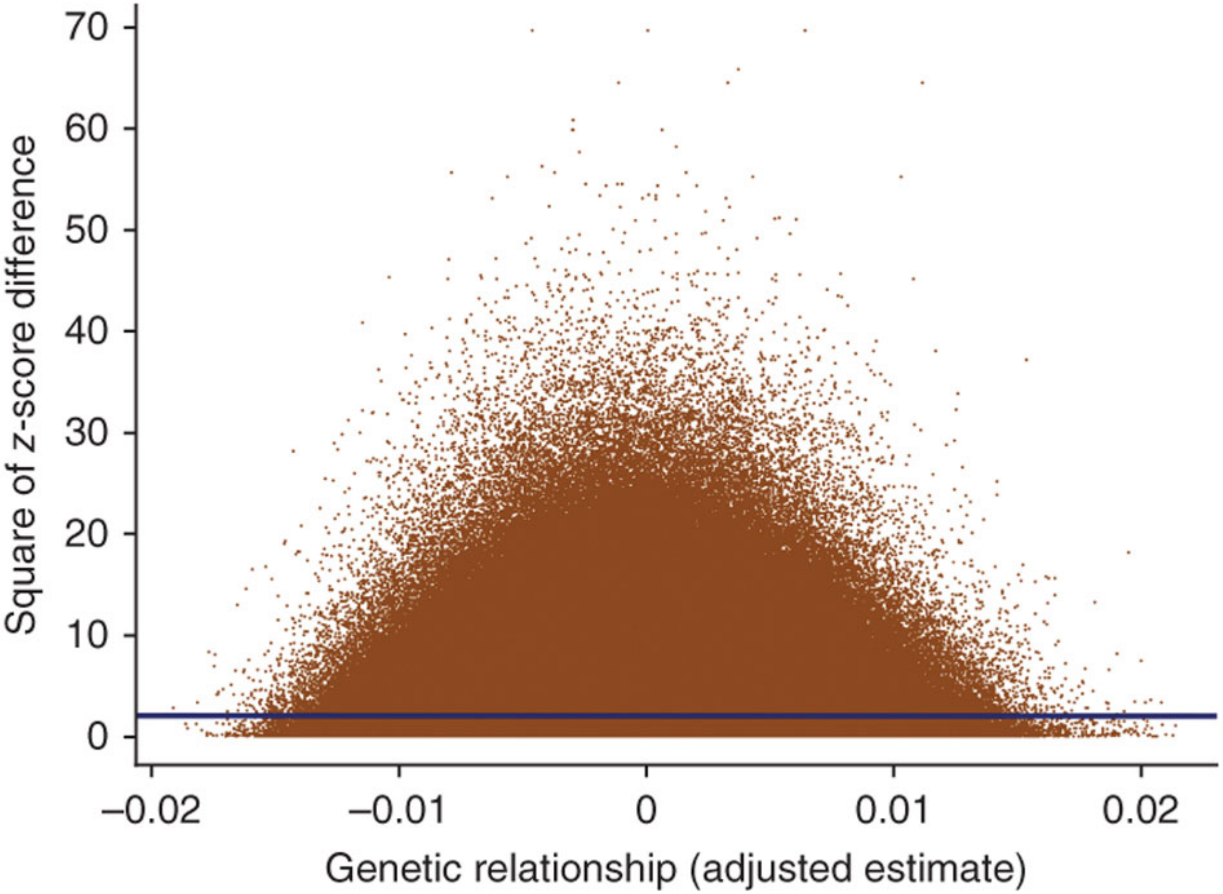
Measuring Relatedness

- Look at similarity of genotypes
- IBS - Identity-by-state
- How similar depends on population allele frequencies

Calculating Relatedness

- Can calculate a measure of relatedness at a SNP using IBS and allele frequency
- Average across all SNPs genotyped

“Unrelated” People



Common SNPs explain a large proportion of the heritability for human height
Yang et al., Nature Genetics (2010) 42, 565–569

Estimating “Heritability”

- Simple regression
  - Squared difference of trait (standardised)
  - Genetic relationship

- Intercept = 2 * V_P
- Slope = -2 * V_A

Example - Height

- From the Yang et al.:
  - Slope = 1.98, Intercept = -1.01
  - → $V_P = 0.990$
  - → $V_A = 0.505$

- $h^2 = V_A / V_P = 0.51$

Not really a heritability...

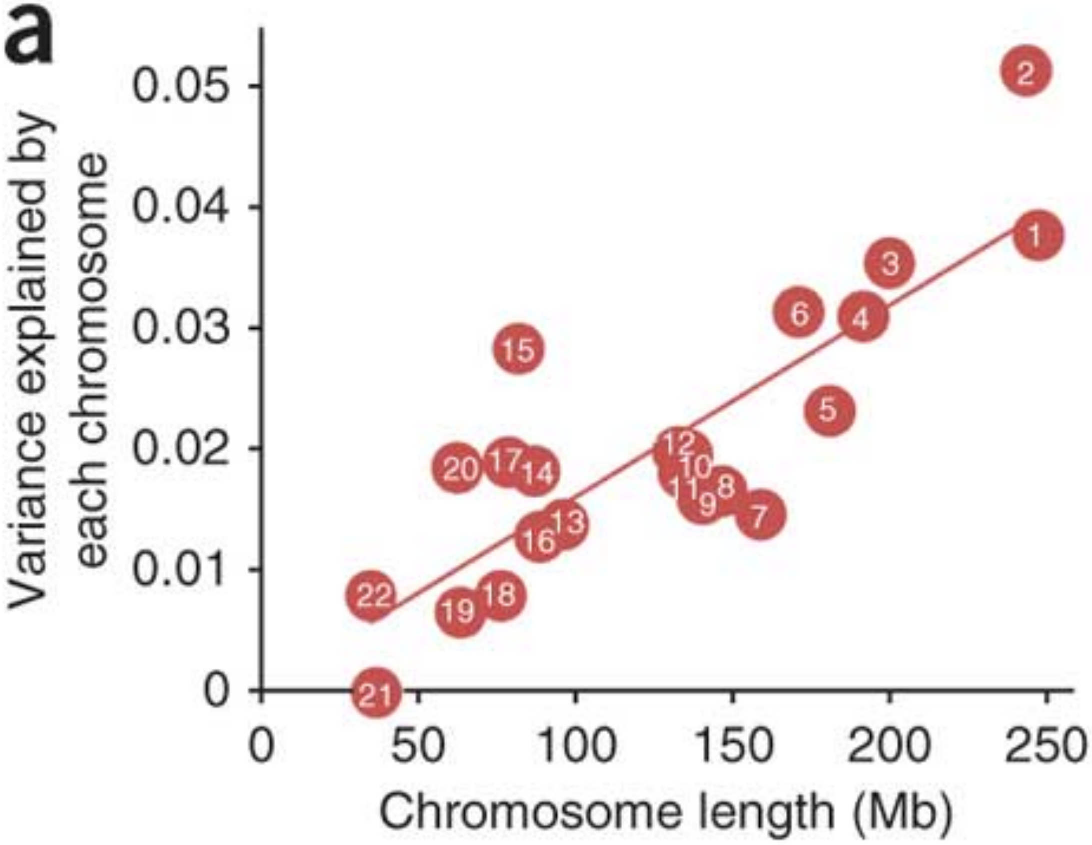
- Variance explained by the SNPs
- ~300,000 SNPs does not capture all variation in the genome
- In particular, rare variation is missed

Further Dissecting

- We can subset the SNPs to ask further questions about the genetic make-up of the trait
- E.g.
  - Do chromosomes contribute equally?
  - Do gene regions contribute more than intergenic regions?

Genome partitioning of genetic variation for complex traits using common SNPs
Yang et al., Nature Genetics (2011) 43, 519–525

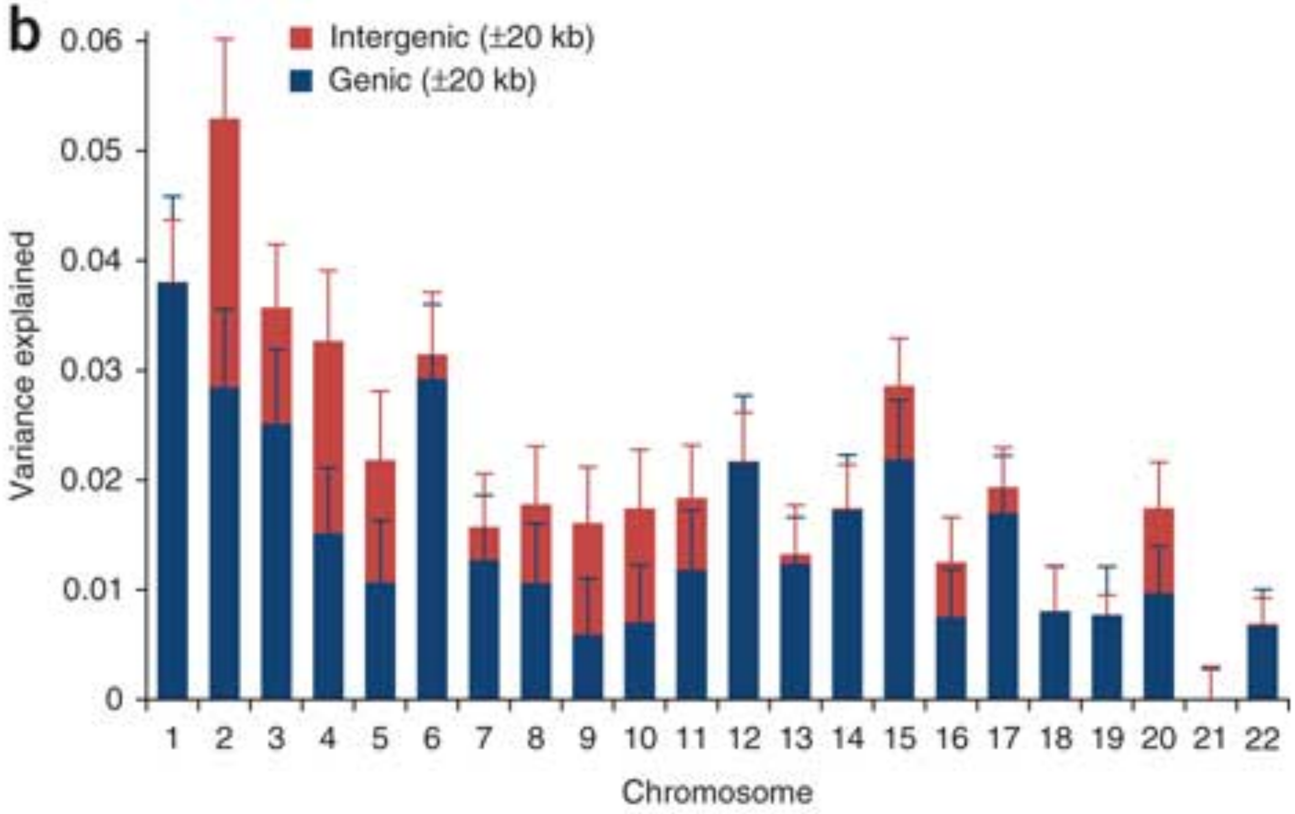
Variance by Chromosome



Genic vs Intergenic Regions

- “Genic” region defined as being from the 5' to the 3' end of a gene +20KB
- Covers 49.4% of the genome
- If random, expect genic region to explain ~50% of variation

Genic vs Intergenic Regions



Genic = 0.328 (72%), Intergenic = 0.126

Heritability

- A trait is heritable if more closely related individuals have more similar phenotypes

- The stronger the relationship between relatedness and phenotypic similarity, the more heritable the trait is.