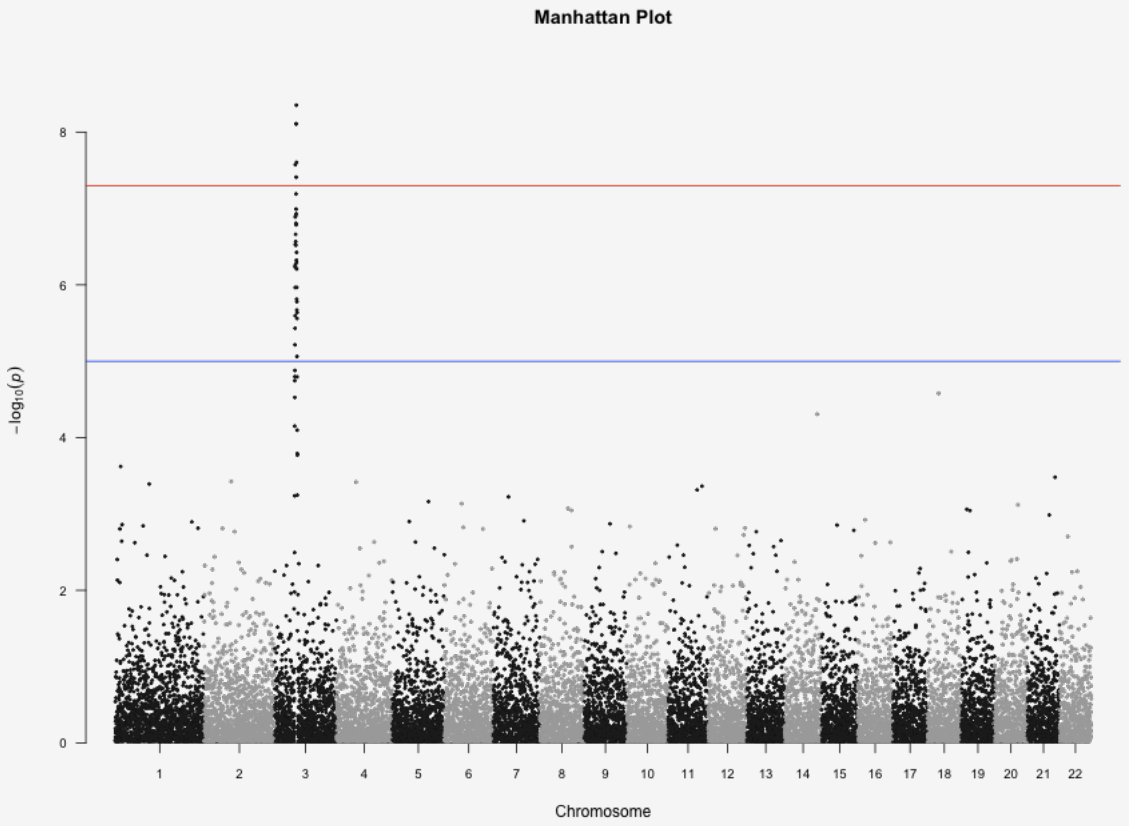
Whole genomes

Gibran Hemani
g.hemani@bristol.ac.uk

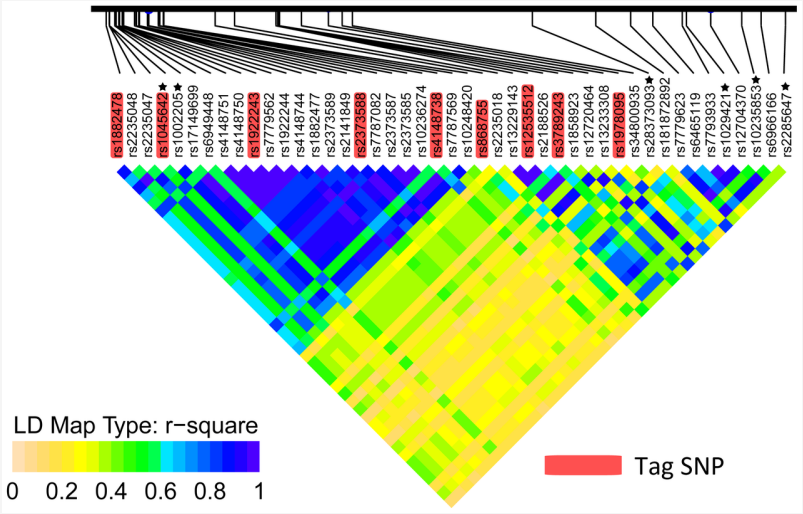
Outline

- Genetic relatedness
- Heritability
- The GCTA approach
- Latest developments

Above and below the threshold

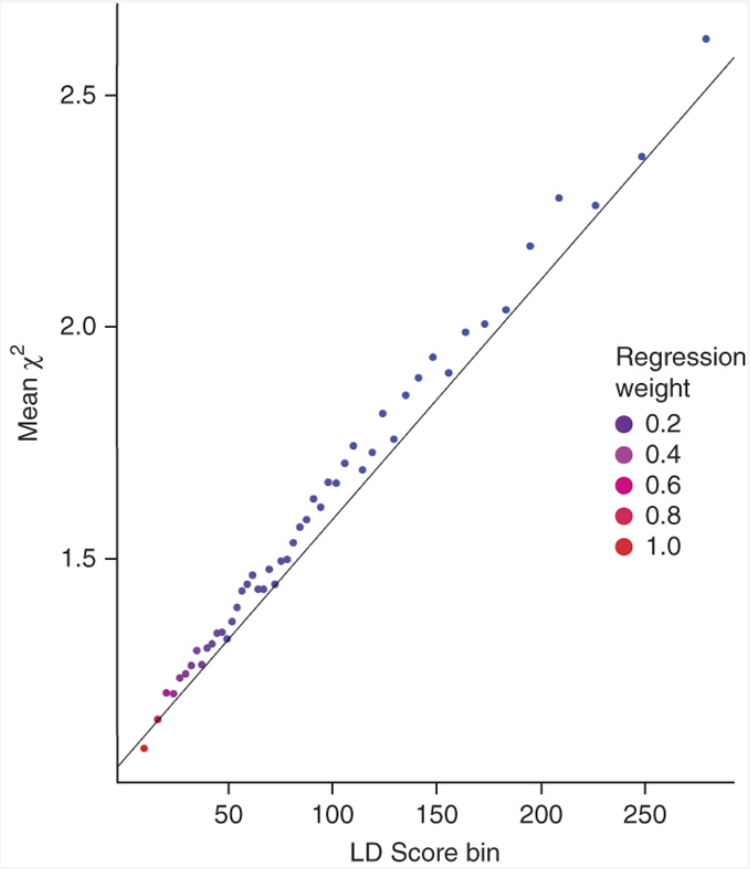


LD score regression



- Imagine every SNP has a random effect
- In GWAS, a SNP with high LD would have larger effects because they would be the combined influence of their own effect and those that are correlated with them

LD score regression



- Plot 'LD score' (average R^2) vs GWAS test statistic
- Slope = SNP heritability
- Intercept = population stratification

Alternatively…

We can also learn about heritability based on how well we can predict a trait based on the results from a GWAS

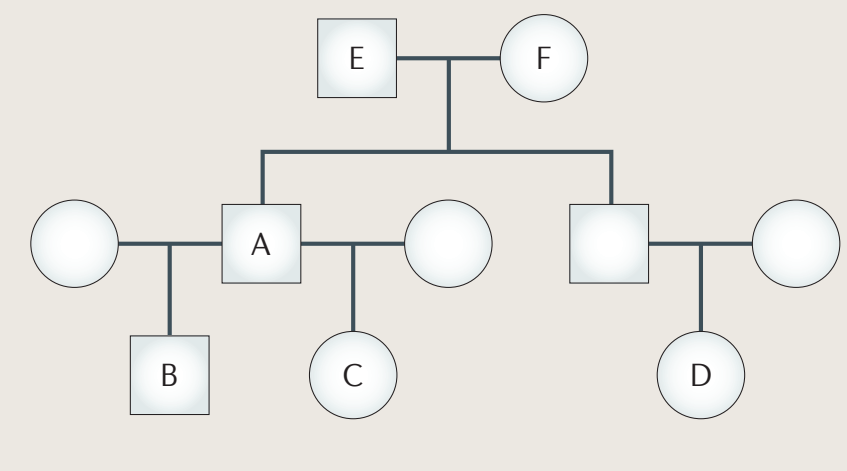
AVENGEME

- Create an allele score based on GWAS results
- Calculate the prediction accuracy of the allele score in an independent sample

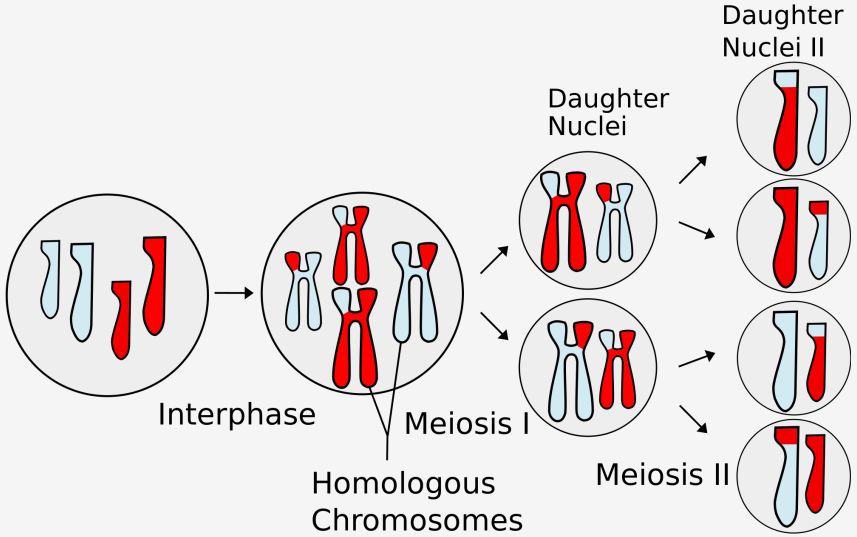
- SNP h2 can be inferred depending on:
  - Training sample size
  - Testing sample size
  - Correlation between PRS and test trait
  - Threshold used to calculate PRS

Family relatedness

Family tree: degrees of relatedness

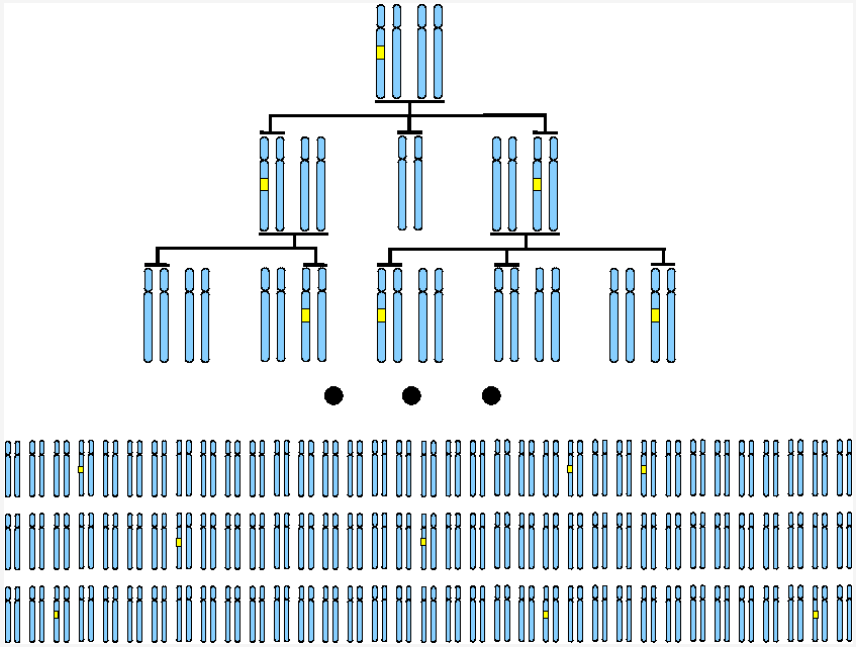


Meiosis: random shuffling



Identity by descent

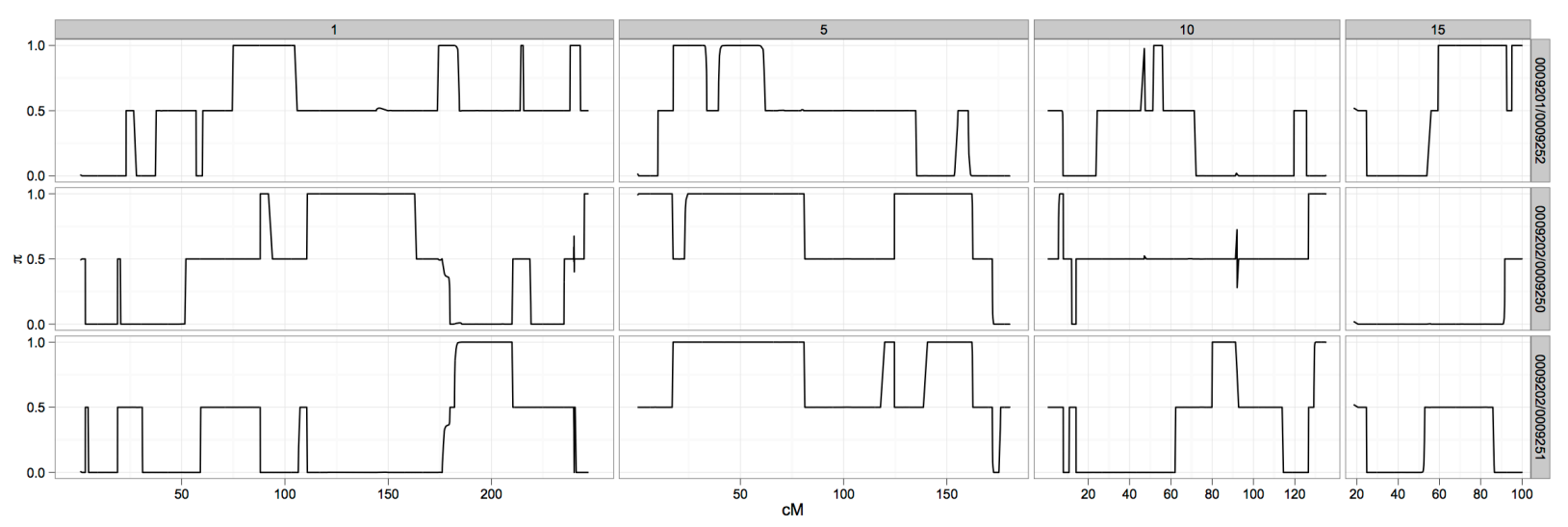
Segment of DNA shared by two or more people that has been inherited from a recent common ancestor without any intervening recombination



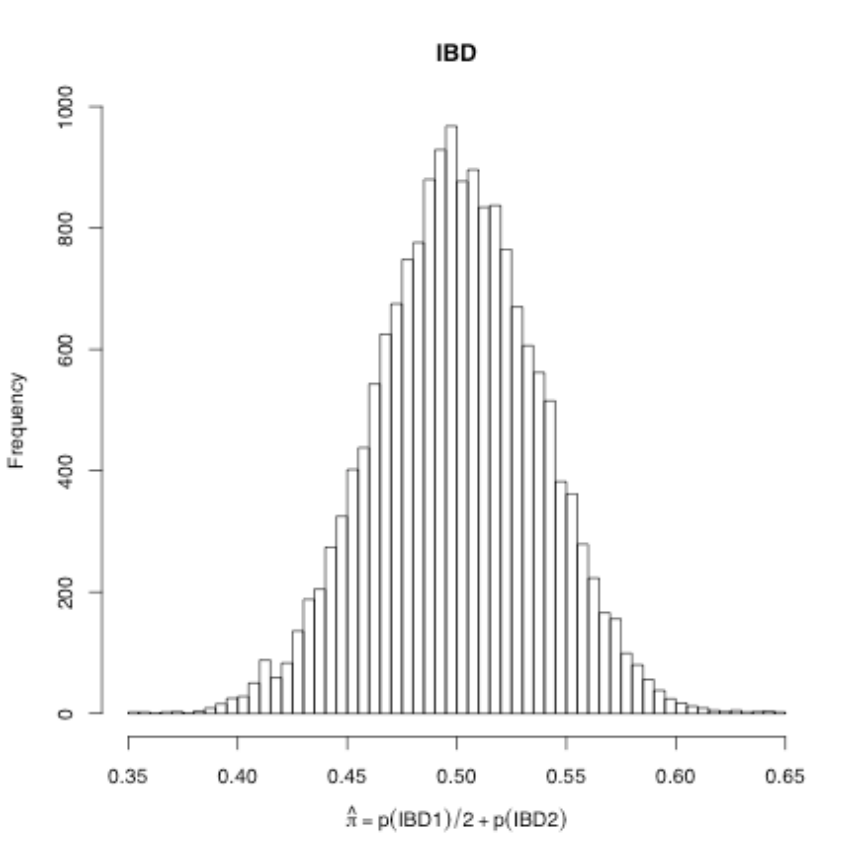
Relatedness (or kinship) coefficients

Relationship to you	relatedness coefficient
yourself	1
identical twin	1
parent, child	1/2
grandparent, grandchild	1/4
great-grandparent, great-grandchild	1/8
n^{th} level ancestor or descendant	$1/2^n$
sibling (sister or brother)	1/2
half-sibling	1/4
aunt, uncle	1/4
niece, nephew	1/4
great-aunt, great-uncle	1/8
great-niece, great-nephew	1/8
first-cousin	1/8
first-cousin-once-removed	1/16
second-cousin	1/32
second-cousin-once-removed	1/64
third-cousin	1/128
n^{th} cousin	$1/2^{2n+1}$
n^{th} cousin, m times removed	$1/2^{2n+m+1}$
stranger	0

Siblings: Realised genetic sharing



Siblings: realised genetic sharing



Variation in complex traits

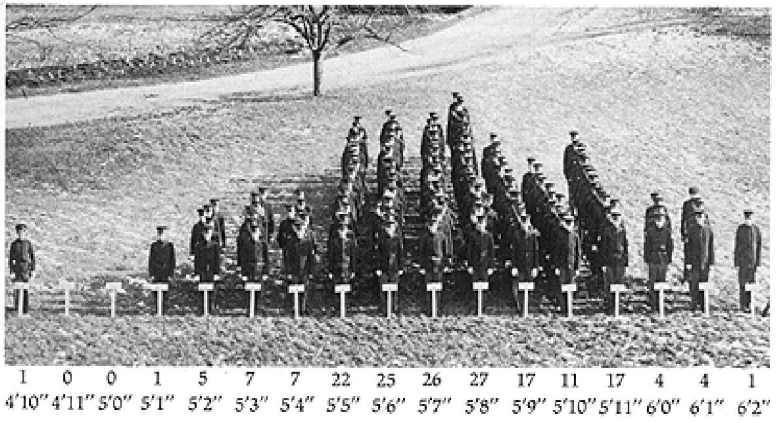


Figure 1.5
Differences in height in the same population: heights of conscripts over 60 years ago. (From A. Blakeslee, *Journal of Heredity*, vol. 5, 1914.)

$$Var(\boldsymbol{Y}) = Var(\boldsymbol{G}) + Var(\boldsymbol{E})$$

$$Var(\boldsymbol{G}) = Var(\boldsymbol{A}) + Var(\boldsymbol{D}) + Var(\boldsymbol{AA}) + Var(\boldsymbol{AD}) + ...$$

$$Var(\boldsymbol{E}) = Var(\boldsymbol{C}) + Var(\boldsymbol{S})$$

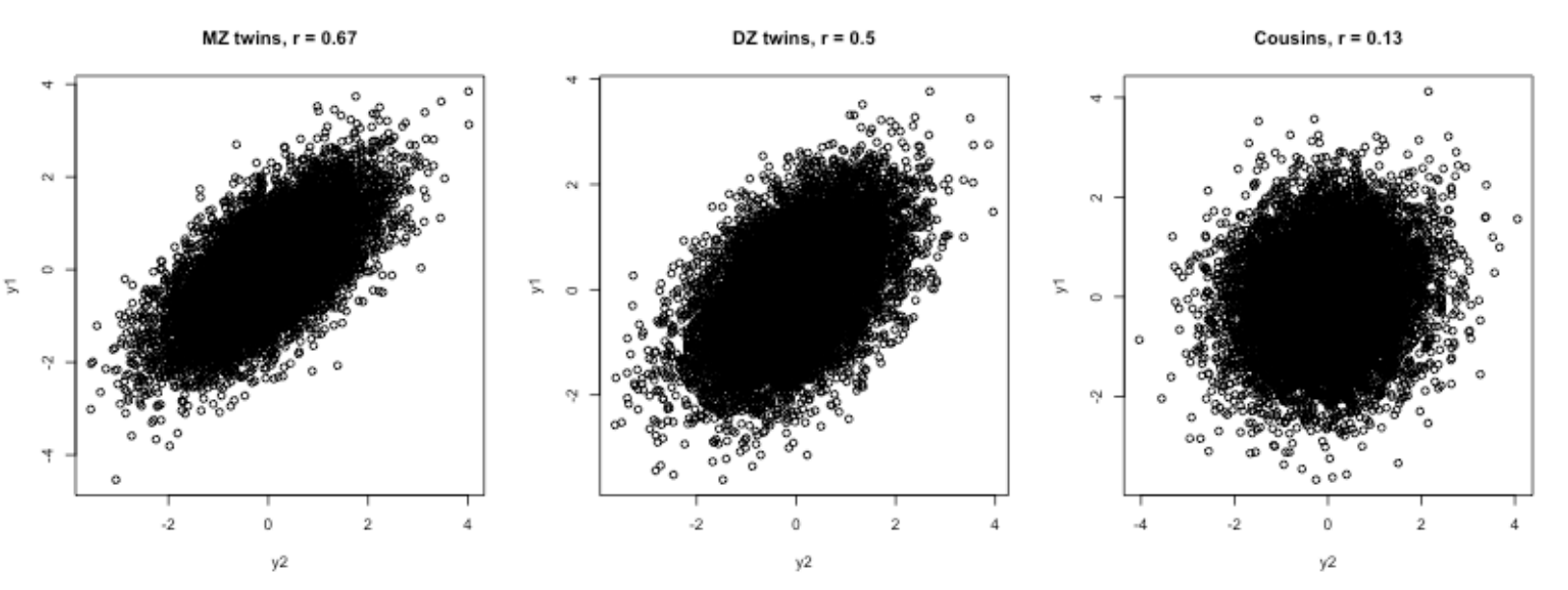
$$H^2 = Var(\boldsymbol{G})/Var(\boldsymbol{Y})$$

$$h^2 = Var(\boldsymbol{A})/Var(\boldsymbol{Y})$$

Heritability tells us:

- The influence of nature vs nurture on the **variation** of a trait in a population
- If there is any *a priori* reason to perform a GWAS
- If heritability estimates of the **same trait** differ in **different populations**
- If heritability estimates of **different traits** differ in the **same population**
- The maximum possible prediction accuracy of a trait using genetic measures alone
- How a trait will respond to natural selection

Phenotypic similarity between relatives



Falconer's method for estimating heritability

What contributes to the phenotypic correlations?

$$r_{mz} = A + C$$

$$r_{dz} = \frac{1}{2}A + C$$

$$r_{cs} = \frac{1}{8}A + xC$$

Let's estimate the additive and common environmental effects

$$0.50 = \frac{1}{2}A + C$$

$$0.13 = \frac{1}{8}A + xC$$

...

$$0.67 - 0.50 = \frac{1}{2}A$$

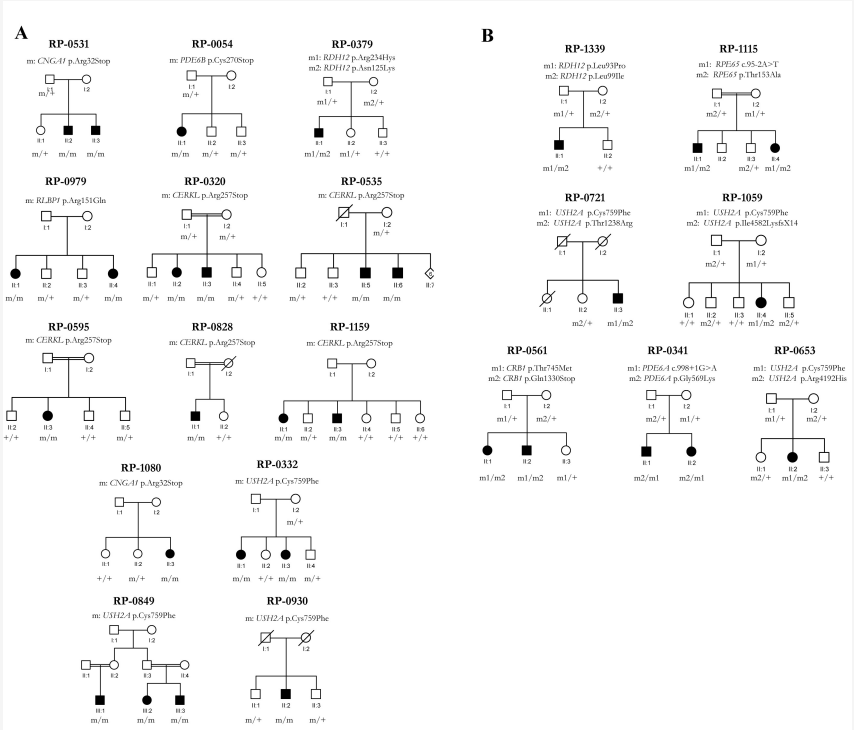
$$A = 0.34$$

Generalising to pedigrees

Lots of different levels of genetic sharing amongst relatives

Relationship to you	relatedness coefficient
yourself	1
identical twin	1
parent, child	1/2
grandparent, grandchild	1/4
great-grandparent, great-grandchild	1/8
n^{th} level ancestor or descendant	$1/2^n$
sibling (sister or brother)	1/2
half-sibling	1/4
aunt, uncle	1/4
niece, nephew	1/4
great-aunt, great-uncle	1/8
great-niece, great-nephew	1/8
first-cousin	1/8
first-cousin-once-removed	1/16
second-cousin	1/32
second-cousin-once-removed	1/64
third-cousin	1/128
n^{th} cousin	$1/2^{2n+1}$
n^{th} cousin, m times removed	$1/2^{2n+m+1}$
stranger	0

Can we use this information to estimate heritability?



Haseman-Elston regression

For each pair of individuals:

To what extent is the phenotypic similarity between pairs of individuals explained by the genetic similarity?

$$(Y_1 - Y_2)^2 = 2(1\text{-}r) - (2A)\theta + e$$

ID1	ID2	y1	y2	$(y1\text{-}y2)^2$	θ
1	2	0.1	-0.3	0.16	0
1	3	0.1	0.2	0.01	0.5
1	4	0.1	-0.1	0.04	0.25
2	3	-0.3	0.2	0.25	0
2	4	-0.3	-0.1	0.04	0

Intuitive, but low statistical power

Mixed model approach

$$\mathbf{y} = \mu + \mathbf{X}\beta + \mathbf{g} + \mathbf{e}$$

Partition variance into covariates, genetic effects and residuals

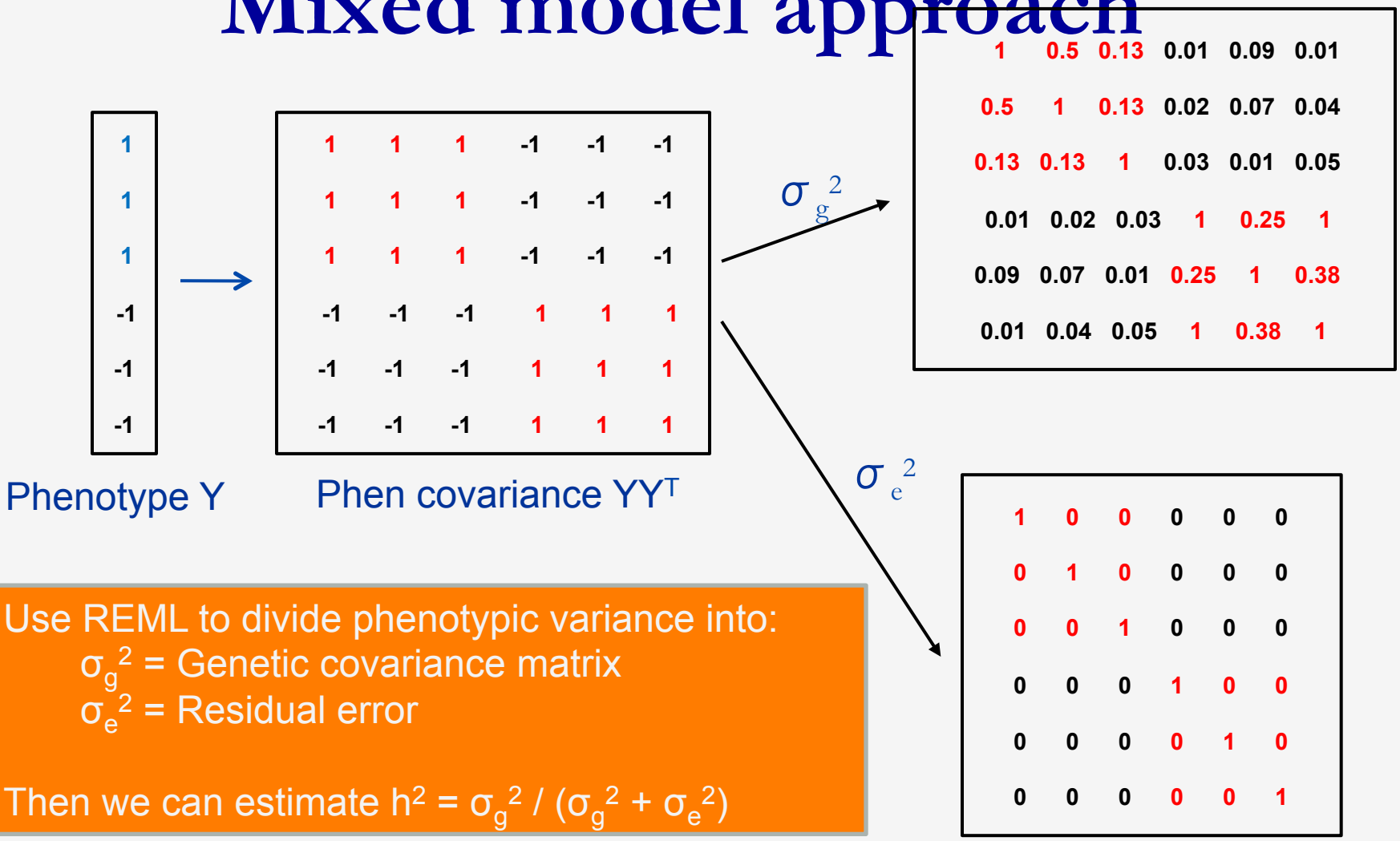
$$\mathbf{e} \sim N(0, \sigma_e^2)$$
$$\mathbf{g} \sim N(0, \sigma_g^2)$$

$$L(\mathbf{y}|\sigma_g^2, \sigma_e^2) = \frac{exp(-0.5(\mathbf{y}-\mu)^T(\mathbf{K}\sigma_g^2 + \mathbf{I}\sigma_e^2)^{-1}(\mathbf{Y}-\mu))}{(2\pi)^{n/2}|\mathbf{K}\sigma_g^2 + \mathbf{I}\sigma_e^2|^{\frac{1}{2}}}$$

Use REML to maximise likelihood for estimation of σ_g^2 and σ_e^2 parameters

Slide courtesy of Doug Speed

Mixed model approach



Use REML to divide phenotypic variance into:

σ_g^2 = Genetic covariance matrix

σ_e^2 = Residual error

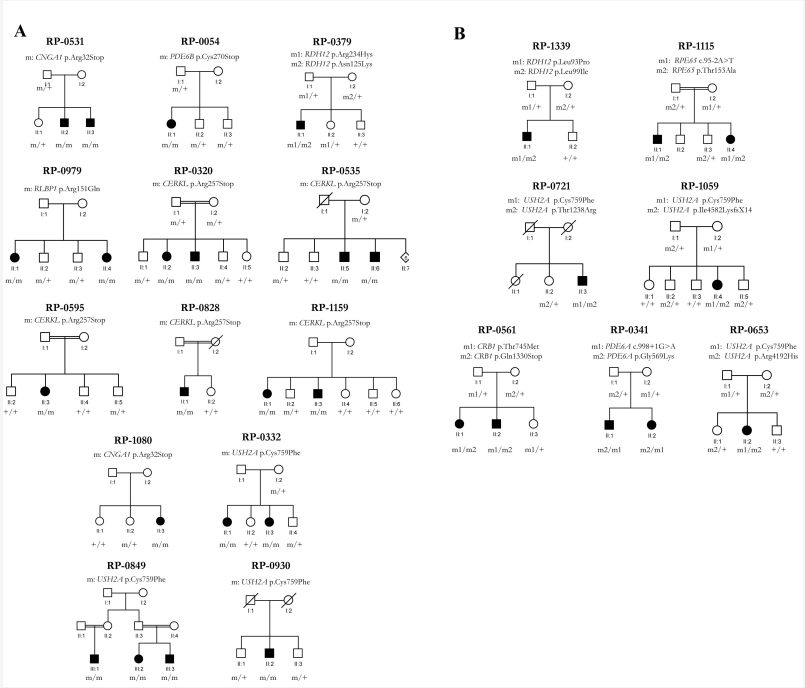
Then we can estimate $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$

Huge number of studies have been performed to estimate heritability

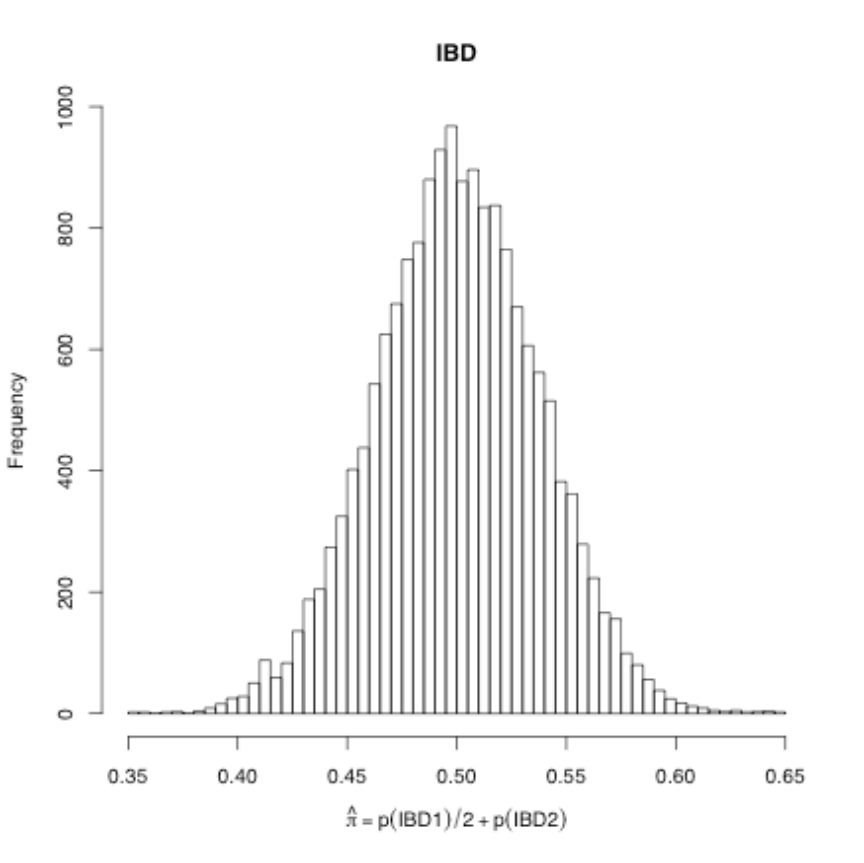
Visscher 2008

Problem: Pedigrees are always incomplete

If pedigrees are incomplete then the relatedness coefficients that we are using are wrong



Estimating the *realised* genomic similarity



Estimating the *realised* genomic similarity using SNPs: Identity by state (IBS)

Objective: For each pair of individuals see how many alleles they have in common as a proportion of all alleles – this is our measure of genetic similarity

Step 1: Take all your SNPs and normalise them so that they have mean 0 and variance 1

Step 2: For each pair of individuals increase their score when they have the same genotype at a SNP, decrease it when they have a different genotype

Estimating the *realised* genomic similarity using SNPs

Raw genetic data, *S* matrix

ID	SNP1	SNP2	SNP3
1	0	2	0
2	1	1	2
3	0	0	0
4	0	2	1
5	1	2	2
6	0	1	1
7	2	0	0

Scale each SNP to have mean 0, variance 1, *Z* matrix

ID	SNP1	SNP2	SNP3
1	-0.7	2.1	-1.0
2	0.5	-0.5	1.3
3	-0.7	-0.4	-1.0
4	-0.7	2.1	0.2
5	0.5	2.1	1.3
6	-0.7	-0.5	0.2
7	1.8	-0.4	-1.0

Estimating the *realised* genomic similarity using SNPs

Z matrix:

ID	SNP1	SNP2	SNP3
1	-0.7	2.1	-1.0
2	0.5	-0.5	1.3
3	-0.7	-0.4	-1.0
4	-0.7	2.1	0.2
5	0.5	2.1	1.3
6	-0.7	-0.5	0.2
7	1.8	-0.4	-1.0

Calculating genetic relationship matrix (or kinship matrix):

$$k_{i,j} = \frac{1}{N}\sum_{k=1}^{N}\frac{(s_{i,k}-2p_k)(s_{j,k}-2p_k)}{2p_k(1-p_k)}$$

$$k_{i,j} = \frac{1}{N}\mathbf{Z}\mathbf{Z}^T$$

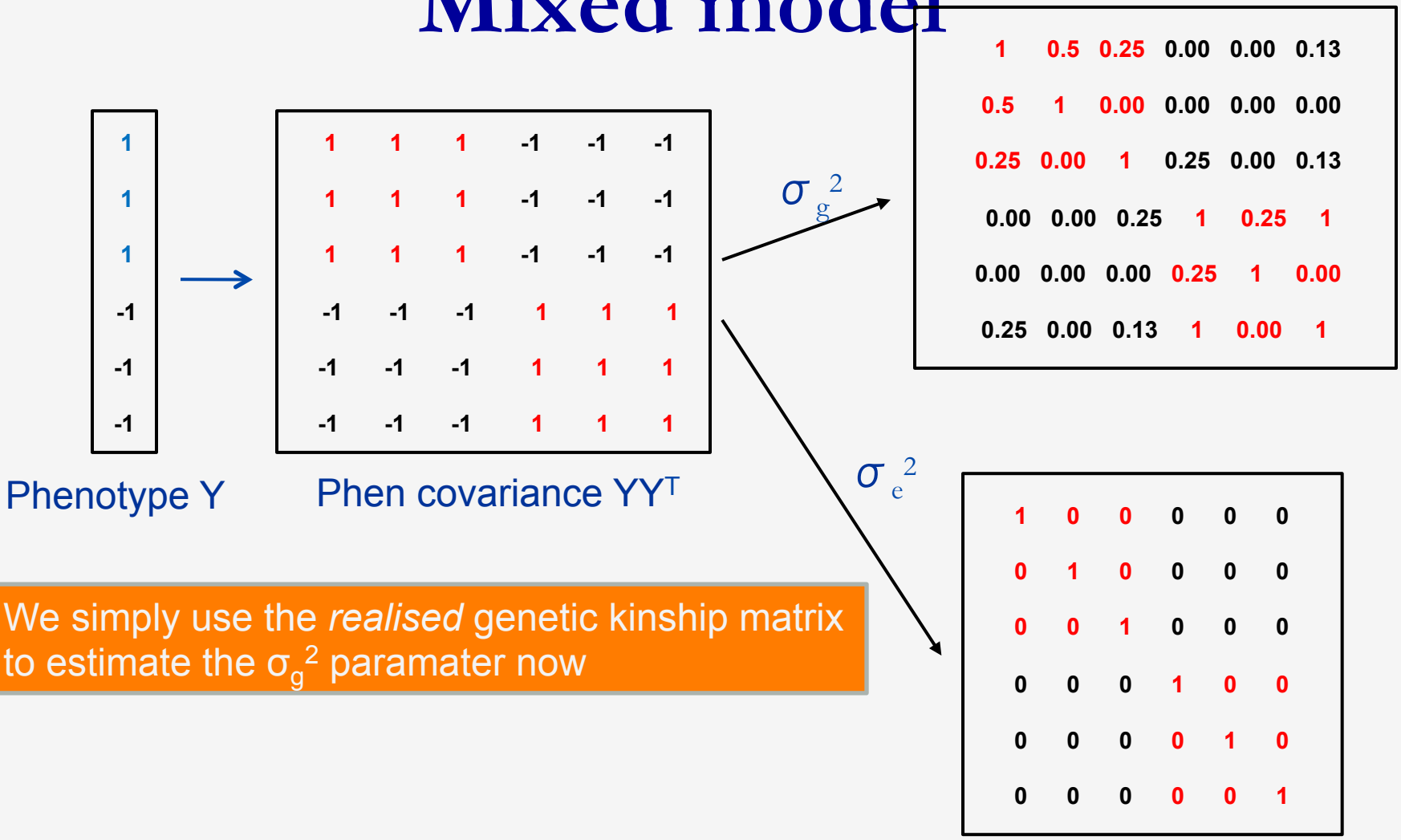
Example:

$$k_{1,2} = \frac{1}{3}(-0.7\times 0.5 + 2.1\times -0.5 + -1.0\times 1.3)$$

$$k_{1,2} = -0.9$$

Slide courtesy of Doug Speed

Mixed model

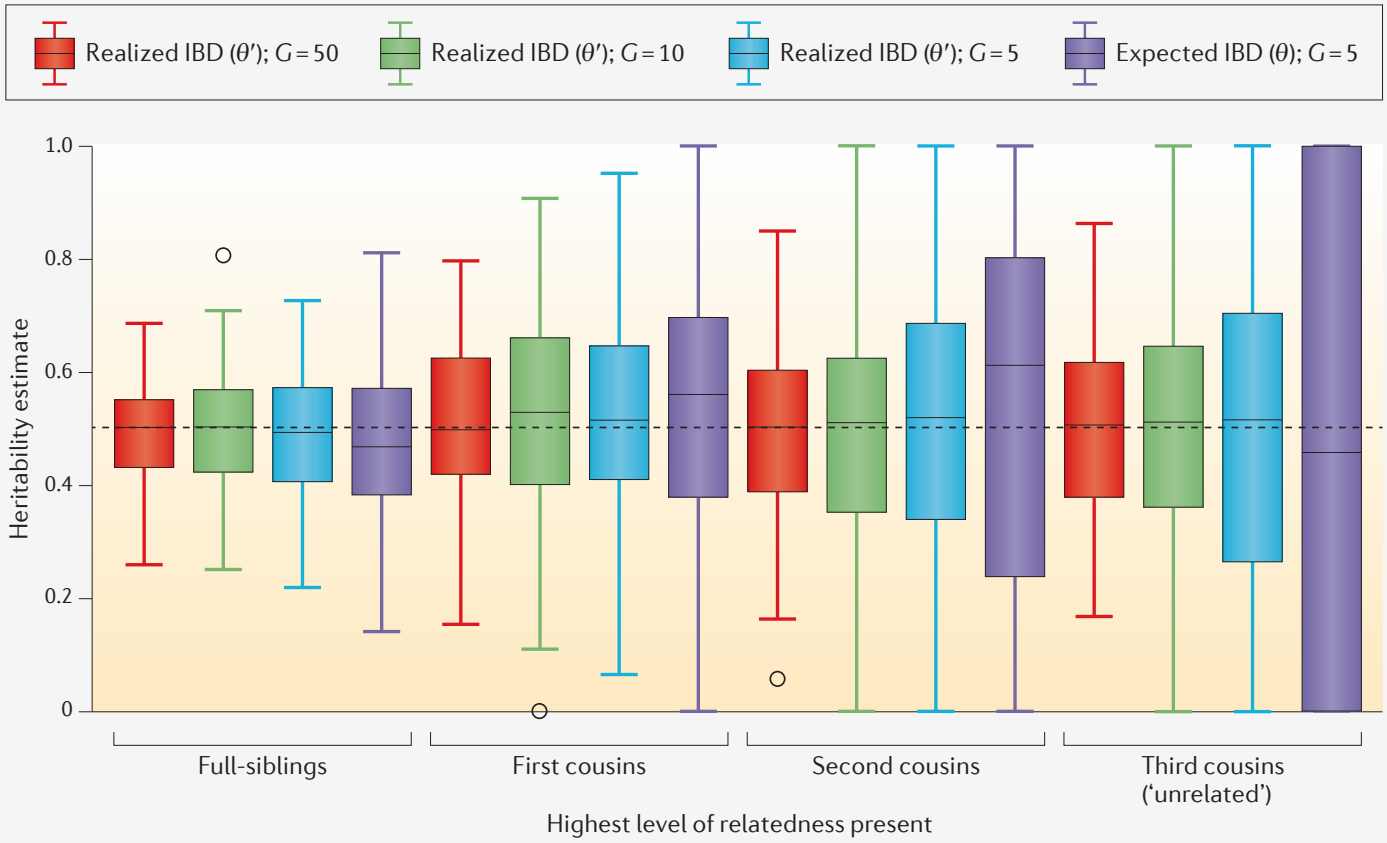


Phenotype Y

Phen covariance YY^T

We simply use the *realised* genetic kinship matrix to estimate the σ_g^2 paramater now

Precision of heritability estimates



Speed & Balding 2015

~~GCTA~~ GREML

Applied SNP-based heritability analysis using only unrelated individuals:

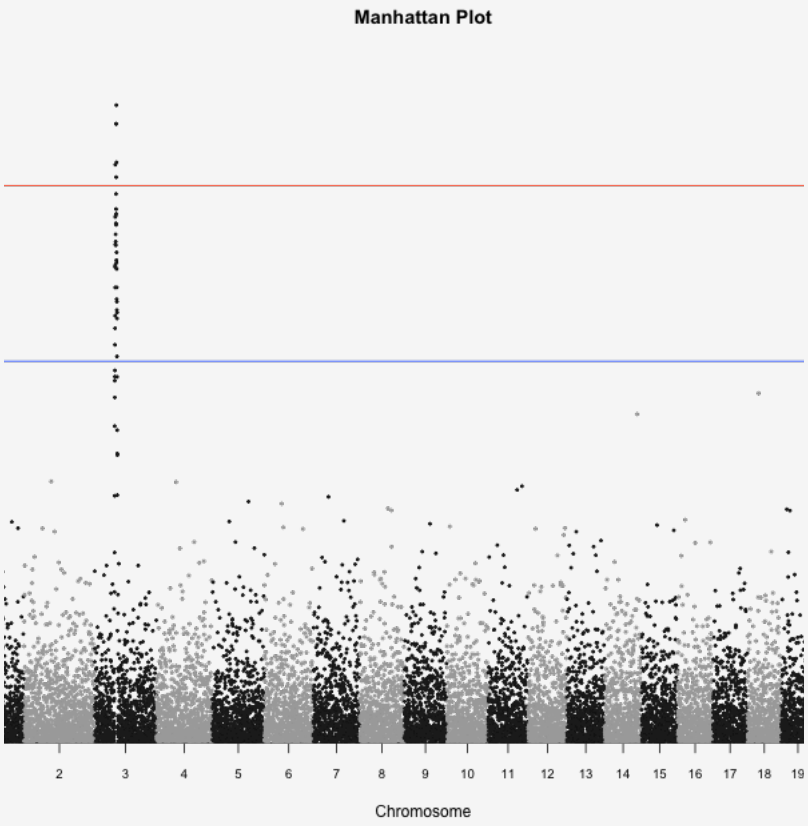
Realised genetic relatedness is required (otherwise all k=0!)

Why go in this direction?

Using unrelateds leads to lower precision!

1. There is a lot of data on unrelated individuals
2. Unrelateds don't have problems of common environment confounding
3. **It's actually trying to answer quite a different question**

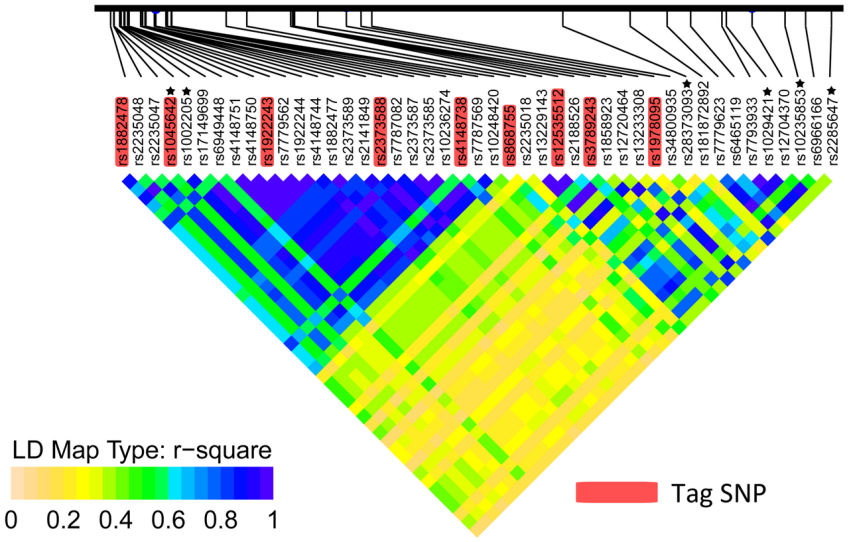
Where is GWAS going?



- How much variance does our significant SNP explain?
  - $V_A/V_P = 2a^2p(1-p)/V_P$
- Usually very small
- Much less than the total heritability predicted to exist from pedigree and twin studies
- AKA the missing heritability problem

Where is GWAS going?

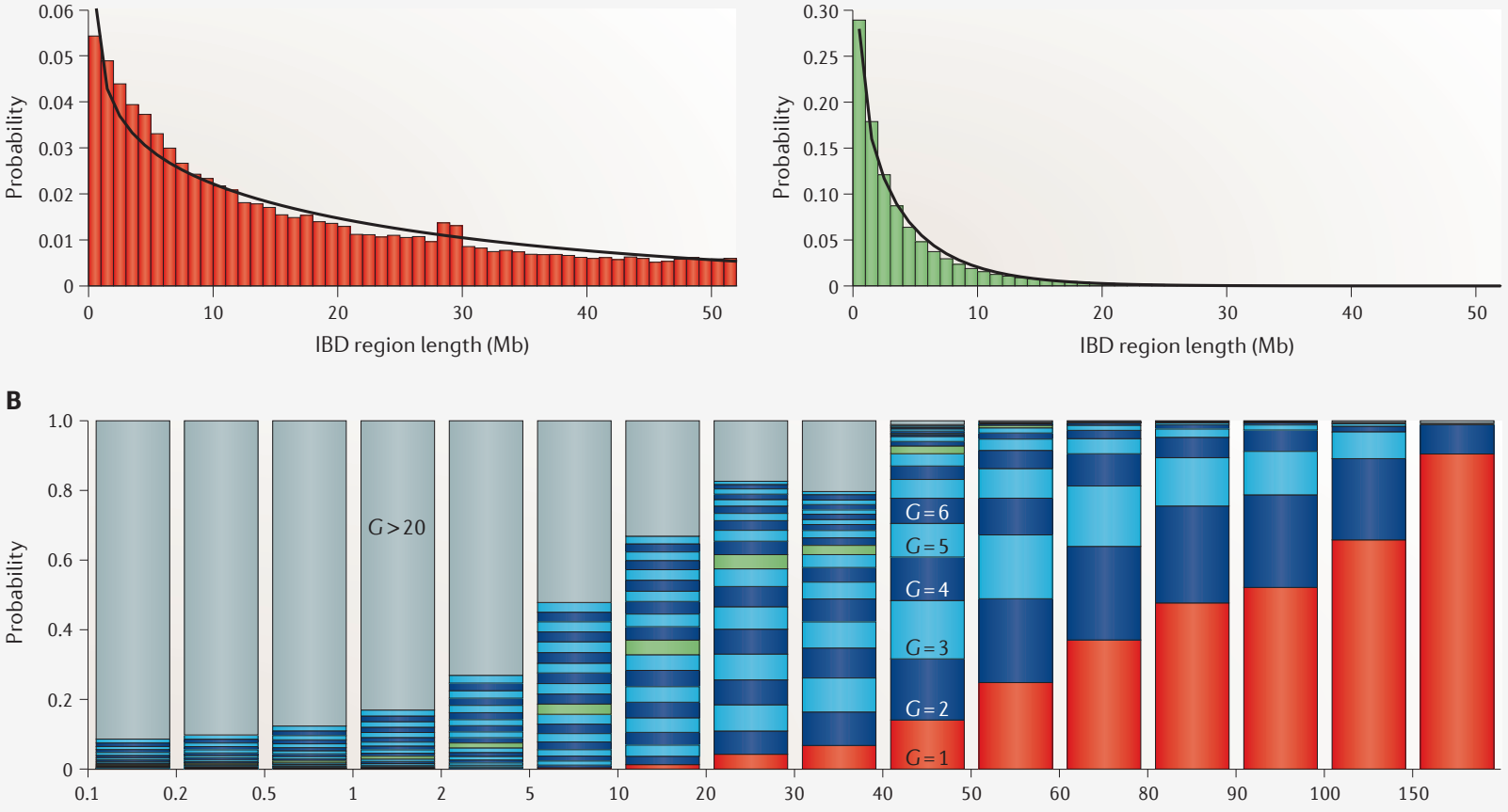
SNP chips use common tagging SNPs



Question: How much of total heritability have we managed to capture using this technology?

- Around 12 million common variants
  - SNP chips usually have fewer than 1 million
  - Estimated to capture at least 80% of genetic variation
- Countless millions of rare variants
  - SNP chips capture very tiny proportion of rare variation in unrelated individuals

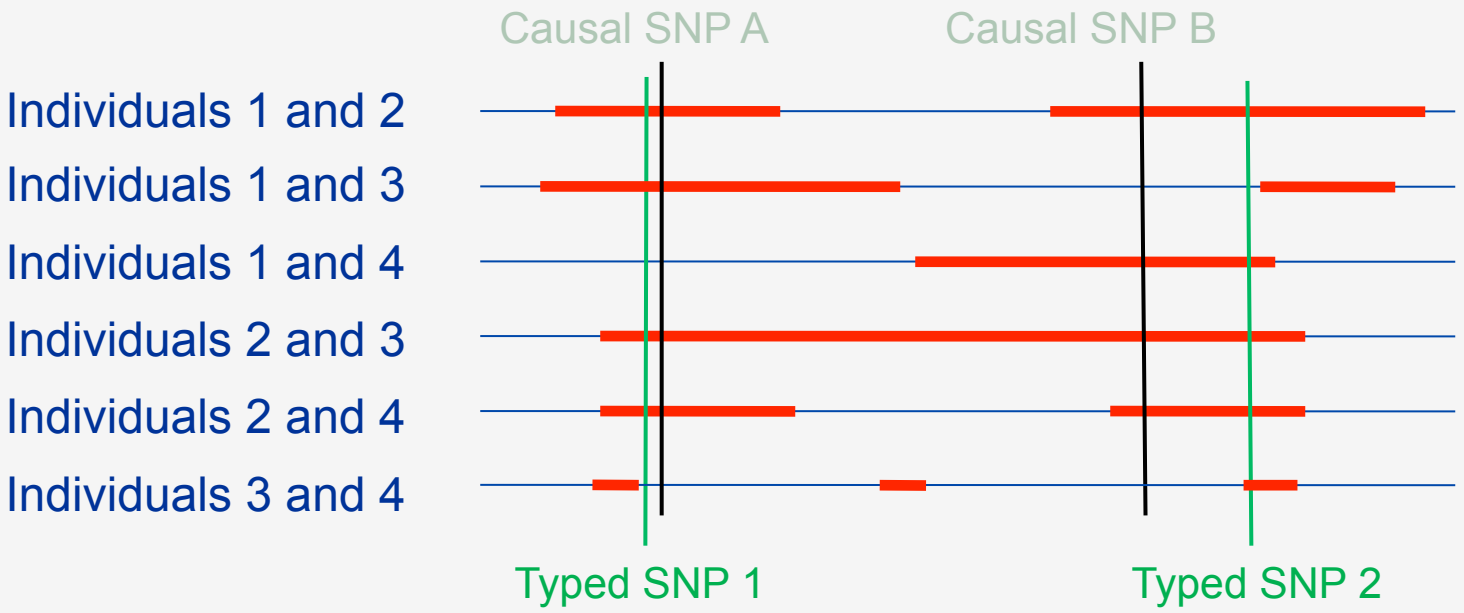
Length of IBD segments reduce when people are less related



Speed & Balding 2015

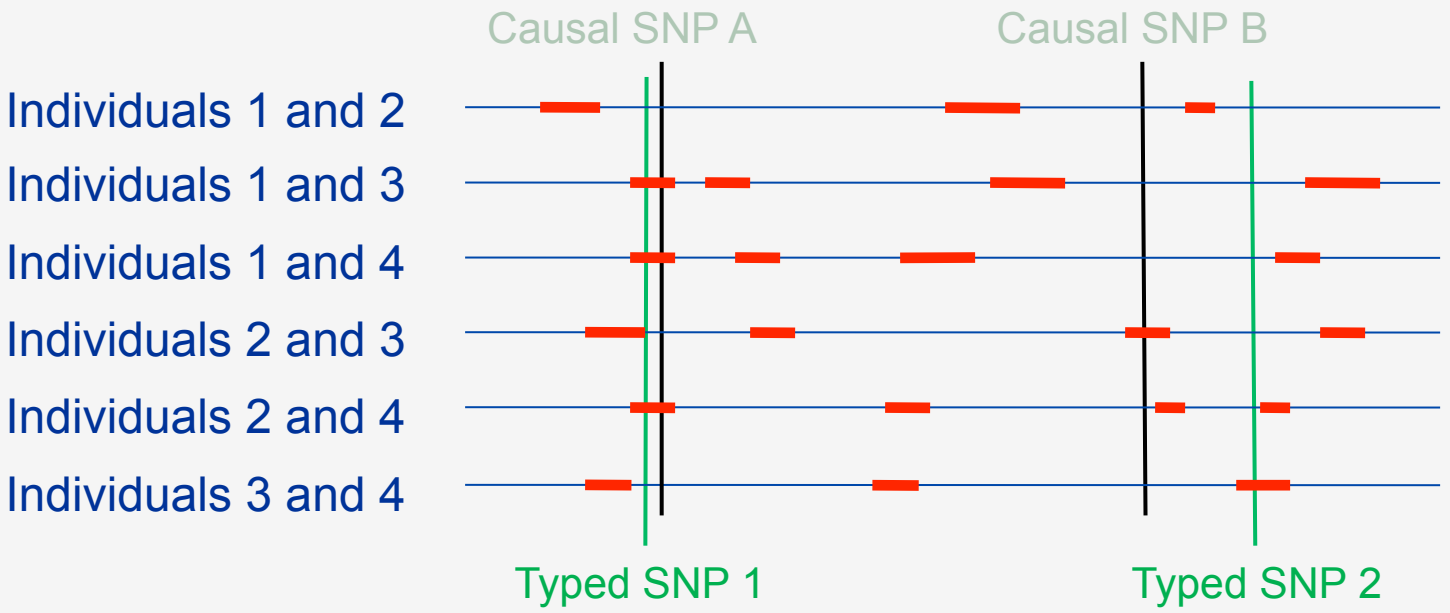
Slide courtesy of Doug Speed

When individuals are related…



When individuals are unrelated…

Slide courtesy of Doug Speed



SNP heritability = proportion of phenotypic variance explained by all measured SNPs

Interpreting SNP heritability

$$
\begin{aligned}
Y = {} & \alpha \\
& + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \\
& + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{15} X_{15} + \beta_{16} X_{16} \\
& + \beta_{17} X_{17} + \beta_{18} X_{18} + \beta_{19} X_{19} + \beta_{20} X_{20} + \beta_{21} X_{21} + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} \\
& + ... + ... + ... + ... + ... + ... + ... + ... \\
& + \beta_{500,000} X_{500,000} \\
& + e
\end{aligned}
$$

Akin to fitting all SNPs in a linear model where every SNP is assumed to have an effect

$$
\begin{aligned}
\mathbf{y} &= \mu + \mathbf{X}\beta + \mathbf{g} + \mathbf{e} \\
\mathbf{g} &= \mathbf{Z}\mathbf{u} \\
\mathbf{e} &\sim N(0, \sigma_e^2) \\
\mathbf{u} &\sim N(0, \mathbf{I}\sigma_u^2) \\
\mathbf{g} &\sim N(0, N\sigma_u^2)
\end{aligned}
$$

Slight difference when fit as a mixed model: assume that the distribution of effects (**u**) are normal

Interpreting SNP heritability

- GCTA has now been used to perform GREML to estimate SNP heritability for thousands of traits

- Typically SNP h^2 is about 50% of 'true' h^2

- This tells us that the capacity of SNP chips used in general populations to map all genetic effects is good, but limited

Interpreting SNP heritability

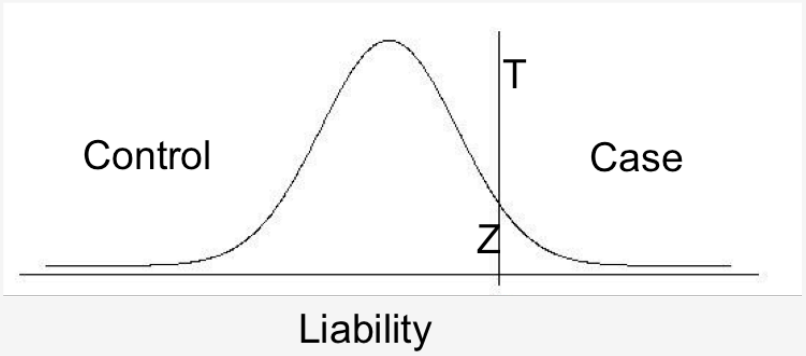
Suppose trait A and trait B both have h^2 = 0.8 estimated from twin studies. We use unrelated individuals to estimate SNP h^2 for each trait in the same population

- Trait A has SNP h^2 = 0.4
- Trait B has SNP h^2 = 0.2

How can we interpret this result?

Case control traits

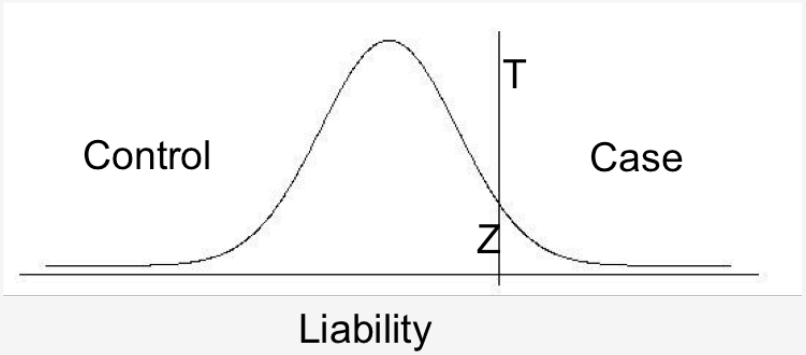
- Normally we ascertain cases to be disproportionately represented in our study
- Can imagine a continuous risk underlying our observed cases or controls
- Once enough risk factors are present above a certain threshold you get the disease



Case control traits

$$h_L^2 = \frac{h_O^2 K^2 (1-K)^2}{P(1-P)Z^2}$$

K = prevalence in population
P = proportion cases in sample
Z = height of curve at threshold T



Genomic partitioning

$$
\begin{aligned}
Y = {} & \alpha \\
& + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \\
& + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{15} X_{15} + \beta_{16} X_{16} \\
& + \beta_{17} X_{17} + \beta_{18} X_{18} + \beta_{19} X_{19} + \beta_{20} X_{20} + \beta_{21} X_{21} + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} \\
& + e
\end{aligned}
$$

Create multiple GRMs, and estimate the variance attributable to each GRM

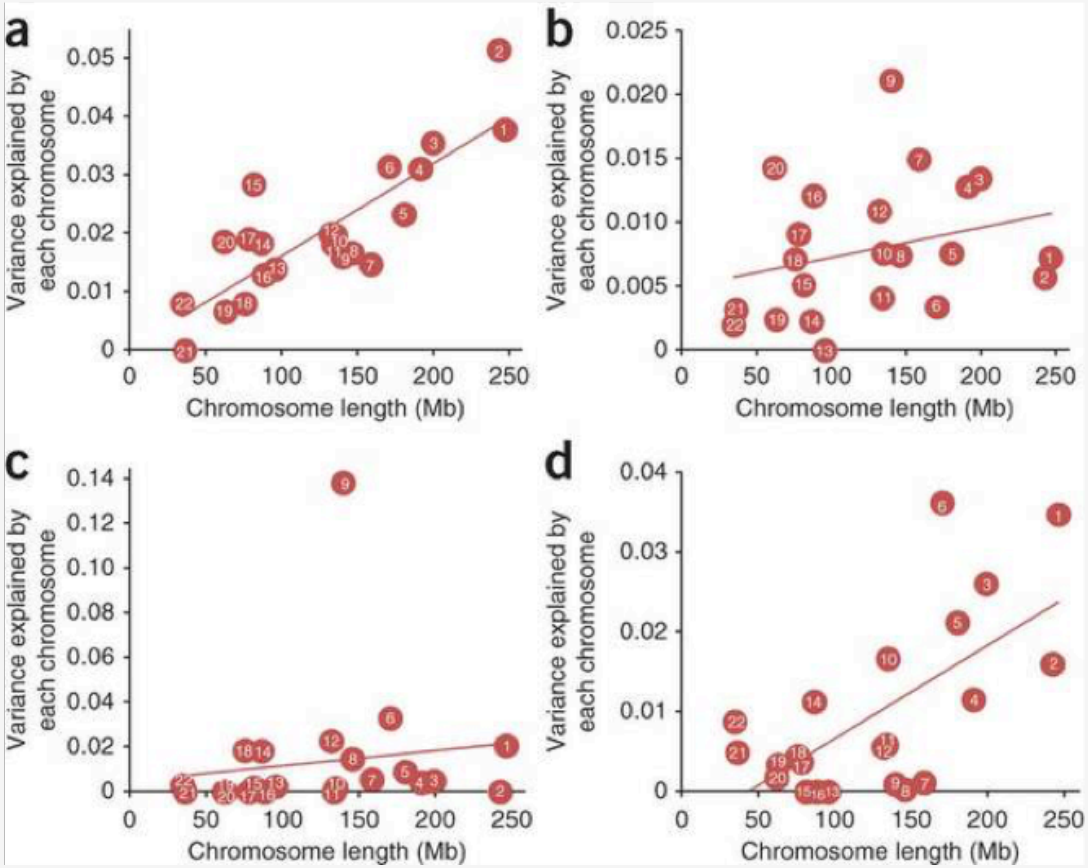
$$\mathbf{y} = \mu + \mathbf{X}\beta + \mathbf{g_1} + \mathbf{g_2} + \mathbf{e}$$

$$\mathbf{g_1} = \mathbf{Z_1 u_1}$$
$$\mathbf{g_2} = \mathbf{Z_2 u_2}$$
$$\mathbf{u_1} \sim N(0, \mathbf{I}\sigma^2_{u_1})$$
$$\mathbf{u_2} \sim N(0, \mathbf{I}\sigma^2_{u_2})$$
$$\sigma^2_g = m\sigma^2_{u_1} + m\sigma^2_{u_2}$$

Examples:

- A partition for each chromosome
- Two partitions to compare genic vs non-genic SNPs
- Divide SNPs into different MAF categories

Genome partitioning: polygenic model



a) Height
b) BMI
c) vWF
d) QTi

Yang et al 2012

Genetic correlations – bivariate analysis

$$Y_1 = \alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \ldots + e_1$$

$$Y_2 = \alpha_2 + \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \lambda_4 X_4 + \lambda_5 X_5 + \lambda_6 X_6 + \ldots + e_2$$

Question: To what extent are the SNPs that influence trait 1 the same that influence trait 2 in terms of concordance of magnitude and direction of effect sizes?

cor(β, λ) = r_g, where values can range from -1 to 1.

Use generalised REML implementation which analyses two traits jointly, estimating σ^2_{g1} and σ^2_{g2} and also r_g.

Interpretation:
Proportion of h^2 that can be explained in trait 1 if perfect knowledge of all SNPs influencing trait 2. And vice versa

a) Is there pleiotropy between two traits?
b) Better understand relationship between sub phenotypes

Interpretation

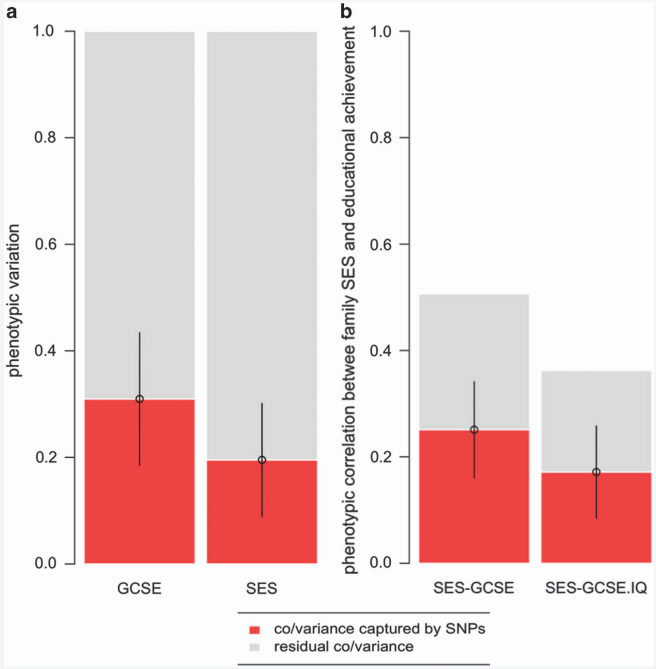


Figure 1. Bivariate genome-wide complex trait analysis (GCTA) of family socioeconomic status (SES) and children's educational achievement (General Certificate of Secondary Education (GCSE)). (**a**) Proportion of phenotypic trait variance tagged by the sampled SNPs in GCSE and family SES, respectively. (**b**) Covariance between family SES and GCSE captured by SNPs, without controlling for intelligence (left bar) and when controlling for intelligence (GCSE.IQ) (right bar). The length of the bar indicates the total phenotypic correlation between SES and GCSE. Solid black lines indicate standard errors.

Linear mixed models for GWAS

- One way to control for population structure is to perform a GWAS where for each SNP you estimate its effect while adjusting for all other SNPs

- There now exist many fast implementations for this
  - FaSTLMM
  - GenABEL
  - GEMMA
  - EMMAX

	GREML	LD Score regression	AVENGEME
Data required	Individual level genetic and phenotypic data	GWAS summary statistics	GWAS summary statistics with replication
Statistical efficiency	High	Low	High
Bivariate analysis	Yes	Yes	Yes
Confounding estimation	No	Yes	No
Estimates of number of causal variants	No	No	Yes
Software available	Yes – GCTA and LDAK	Python scripts and also online platform (LD Hub)	Kind of. R scripts to do calculations but you need to do a few steps first
Reliability of inferences	Biases exist and are somewhat understood	Largely unexplored so far	Largely unexplored so far

References

Useful reviews:

Introduction to Quantitative Genetics. Falconer D.S., Mackay T.F.C. Longman; Harlow, UK: 1996.

Heritability in the genomics era—concepts and misconceptions. Visscher P.M., Hill W.G., Wray N.R. Nat. Rev. Genet. 2008;9:255–266

Relatedness in the post-genomic era: is it still useful? Doug Speed & David J. Balding. Nature Reviews Genetics 16, 33–44 (2015)

The heritability of human disease: estimation, uses and abuses. Tenesa A1, Haley CS. Nat Rev Genet. 2013 Feb;14(2):139-49.

Reconciling the analysis of IBD and IBS in complex trait studies. Powell J.E., Visscher P.M., Goddard M.E. Nat. Rev. Genet. 2010;11:800–805

Finding the missing heritability of complex diseases. Manolio T.A., Collins F.S., Cox N.J., Goldstein D.B., Hindorff L.A., Hunter D.J., McCarthy M.I., Ramos E.M., Cardon L.R., Chakravarti A. Nature. 2009;461:747–753.

Heritability of threshold characters. Dempster E.R., Lerner I.M. Genetics. 1950;35:212–236.

Primary research papers:

Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs. Hemani G et al. Am J Hum Genet. 2013 Nov 7;93(5):865-75.

Common SNPs explain a large proportion of the heritability for human height. Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., Nyholt D.R., Madden P.A., Heath A.C., Martin N.G., Montgomery G.W. Nat. Genet. 2010;42:565–569

Genome-partitioning of genetic variation for complex traits using common SNPs. Yang J, Manolio TA, Pasquale LR, et al. Nature genetics. 2011;43(6):519-525. doi:10.1038/ng.823.

Improved heritability estimation from genome-wide SNPs. Speed D, Hemani G, Johnson MR, Balding DJ. Am J Hum Genet. 2012 Dec 7;91(6):1011-21.

Estimating Missing Heritability for Disease from Genome-wide Association Studies. Lee SH, Wray NR, Goddard ME, Visscher PM. American Journal of Human Genetics. 2011;88(3):294-305. doi: 10.1016/j.ajhg.2011.02.002.

Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Bioinformatics. 2012;28(19):2540-2542. doi: 10.1093/bioinformatics/bts474.

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Bulik-Sullivan et al. Nature Genetics. 2015

A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait. Luigi Palla and Frank Dudbridge. AJHG. 2015