SNP-Based Heritability Analysis

1. SNP-based Heritability Analysis with Unrelated Individuals
2. Equivalence with Random Effects Regression

Recap of Module 7

The focus is on estimating $h^2 = Var(A)/Var(Y)$, the narrow-sense (additive) heritability. Var(Y) is in general easy to estimate

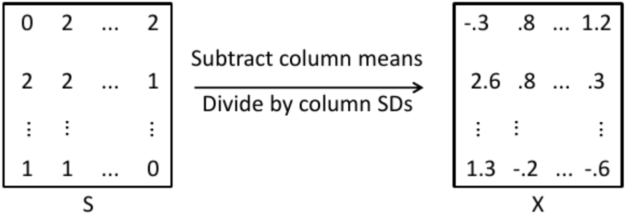
We explained how to estimate h^2 from (a single type of) related pairs: the Covariance Equation explains how phenotypic covariance ($Cov(Y_i, Y_j)$) is related to additive variance (Var(A))

When we have multiple types of related pairs, a generalization of the Covariance Equation is the Mixed Model. Key to this analysis is construction of the kinship matrix **K**, which provides pairwise genetic similarities

Traditionally, K_{ij} represented expected relatedness (twice IBD) between Individuals i and j based on the known pedigree. When SNP data are available, can instead use actual relatedness / genetic similarity

Allelic Correlations

Suppose the matrix S (size n x N) contains for n individuals the genotypes for each of N SNPs. First, for each SNP, we standardise the genotypes so they have mean zero and variance one[†]



$$\underset{S}{\boxed{\begin{matrix}0 & 2 & \ldots & 2\\ 2 & 2 & \ldots & 1\\ \vdots & \vdots & & \vdots\\ 1 & 1 & \ldots & 0\end{matrix}}} \xrightarrow[\text{Divide by column SDs}]{\text{Subtract column means}} \underset{X}{\boxed{\begin{matrix}-.3 & .8 & \ldots & 1.2\\ 2.6 & .8 & \ldots & .3\\ \vdots & \vdots & & \vdots\\ 1.3 & -.2 & \ldots & -.6\end{matrix}}}$$

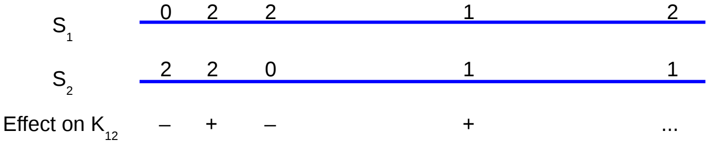
Then we use $\mathbf{K} = \mathbf{X}\mathbf{X}^T/N$ ("allelic correlations") as an estimator of genetic similarities

[†] as discussed in Module 8, other standardizations can be used

Allelic Correlations

$\mathbf{XX}^T/N$ measures pairwise IBS (identity by state) how similar the genotype values are for each pair of individuals

For example, to calculate $K_{1,2}$, correlations between individuals 1 and 2, you can imagine laying their two genomes side by side, then examining for each SNP in turn, how similar their genotypes

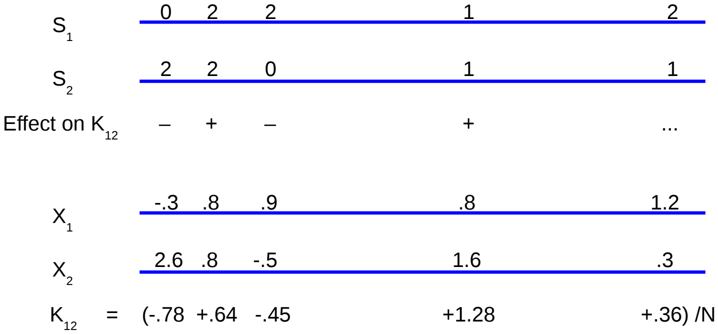


S_1	0	2	2	1	2
S_2	2	2	0	1	1
Effect on K_{12}	–	+	–	+	...

Allelic Correlations

$\mathbf{XX}^T/N$ measures pairwise IBS (identity by state) how similar the genotype values are for each pair of individuals

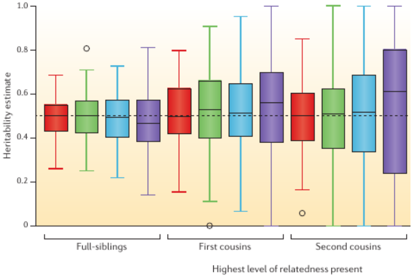
For example, to calculate $K_{1,2}$, correlations between individuals 1 and 2, you can imagine laying their two genomes side by side, then examining for each SNP in turn, how similar their genotypes



S_1		0	2	2	1	2
S_2		2	2	0	1	1
Effect on K_{12}		–	+	–	+	...
X_1		-.3	.8	.9	.8	1.2
X_2		2.6	.8	-.5	1.6	.3
K_{12}	=	(-.78	+.64	-.45	+1.28	+.36) /N

Using actual relatedness improves accuracy

By measuring actual genetic similiarity, rather than relying on expected similarity, we can obtains more precise estimate of h^2



Purple boxes are estimates using expected relatedness; red use actual relatedness (green, blue use less accurate measures of actual relatedness)

Even bigger benefit - we can use “unrelated” individuals

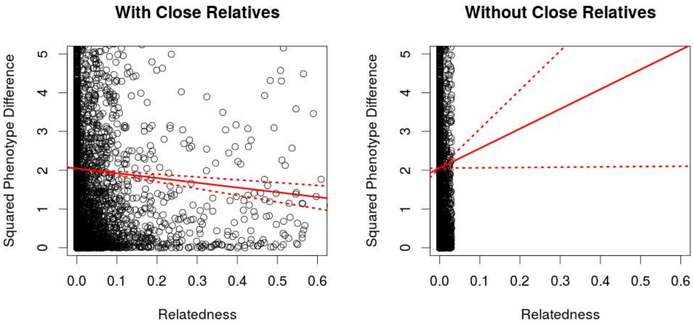
1. SNP-based Heritability Analysis with Unrelated Individuals

2. Equivalence with Random Effects Regression

Using unrelated individuals

In 2010, Jian Yang, Peter Visscher, et al. considered estimating heritability using only "unrelated individuals"

Why? Estimates of h^2 become less precise as number of close relatives in the sample decreases



Using unrelated individuals

However, using unrelated individuals has three key advantages:

Less of a problem that we ignore effects of common environment (and dominance / epistasis)

Can use GWAS data, so sample sizes are much larger than using family data

The resulting estimates, referred to as h^2_{SNP}, are estimates of “SNP heritability”, the total variance explained by all SNPs

This area is referred to as **SNP-based heritability analysis**. The major software is GCTA; our software is LDAK

The missing heritability problem

From about 2006 - 2012 human geneticists were increasingly referring to the missing heritability problem

The case of the missing heritability

[illegible]

Maher, Nature news feature (2008)

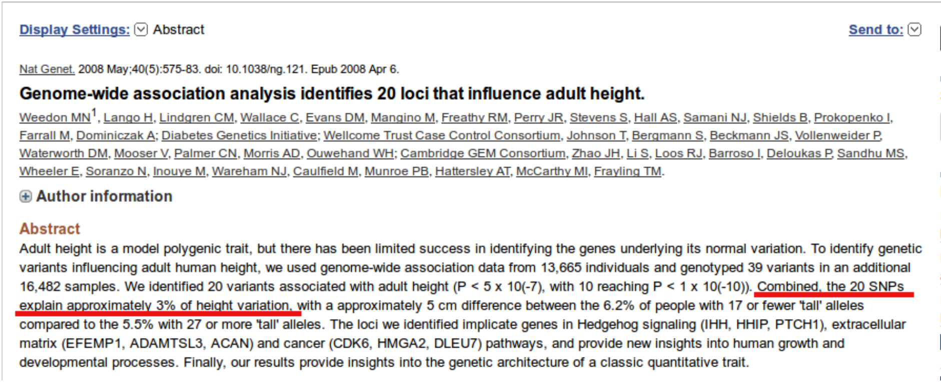
The missing heritability problem

Although GWAS had found a number of associations for a wide range of phenotypes, the proportion of variance explained by the associations for any particular phenotype (h^2_{GWAS}) was typically slight compared to the phenotype's heritability

The classic example was height. The heritability is about 80%
In 2008, 20 associations had been found, but these explained only a few percent of variation (Genome-wide association analysis identifies 20 loci, Nature Genetics - next slide)

A 2014 study by the GIANT consortium increased the number of loci to > 100, but h^2_{GWAS} remains only 10% (Defining the role of common variation ... in human height, Nature Genetics)

The missing heritability problem



Nat Genet. 2008 May;40(5):575-83. doi: 10.1038/ng.121. Epub 2008 Apr 6.

Genome-wide association analysis identifies 20 loci that influence adult height.

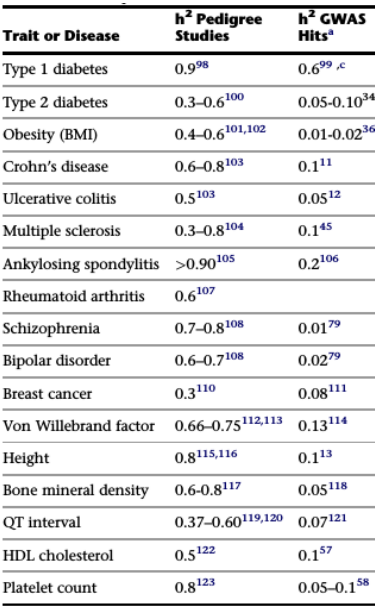
Weedon MN[1], Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JR, Stevens S, Hall AS, Samani NJ, Shields B, Prokopenko I, Farrall M, Dominiczak A; Diabetes Genetics Initiative; Wellcome Trust Case Control Consortium, Johnson T, Bergmann S, Beckmann JS, Vollenweider P, Waterworth DM, Mooser V, Palmer CN, Morris AD, Ouwehand WH; Cambridge GEM Consortium, Zhao JH, Li S, Loos RJ, Barroso I, Deloukas P, Sandhu MS, Wheeler E, Soranzo N, Inouye M, Wareham NJ, Caulfield M, Munroe PB, Hattersley AT, McCarthy MI, Frayling TM.

⊕ **Author information**

Abstract

Adult height is a model polygenic trait, but there has been limited success in identifying the genes underlying its normal variation. To identify genetic variants influencing adult human height, we used genome-wide association data from 13,665 individuals and genotyped 39 variants in an additional 16,482 samples. We identified 20 variants associated with adult height (P < 5 x 10(-7), with 10 reaching P < 1 x 10(-10)). Combined, the 20 SNPs explain approximately 3% of height variation, with a approximately 5 cm difference between the 6.2% of people with 17 or fewer 'tall' alleles compared to the 5.5% with 27 or more 'tall' alleles. The loci we identified implicate genes in Hedgehog signaling (IHH, HHIP, PTCH1), extracellular matrix (EFEMP1, ADAMTSL3, ACAN) and cancer (CDK6, HMGA2, DLEU7) pathways, and provide new insights into human growth and developmental processes. Finally, our results provide insights into the genetic architecture of a classic quantitative trait.

The missing heritability problem



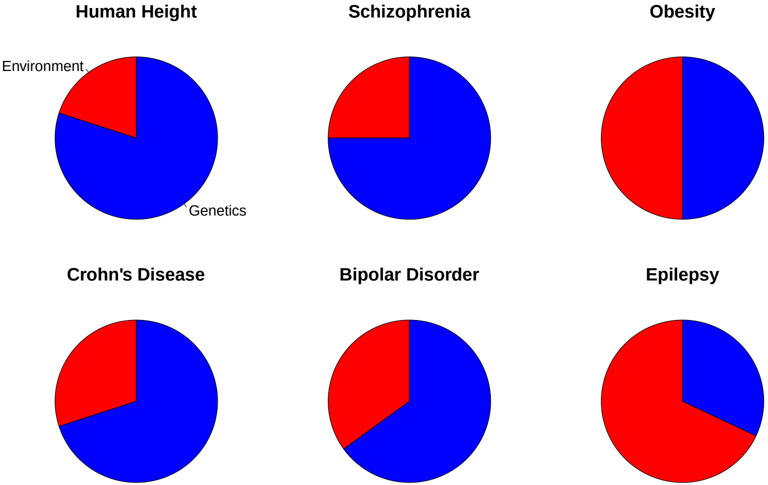
Trait or Disease	h^2 Pedigree Studies	h^2 GWAS Hits[a]
Type 1 diabetes	0.9[98]	0.6[99 ,c]
Type 2 diabetes	0.3–0.6[100]	0.05-0.10[34]
Obesity (BMI)	0.4–0.6[101,102]	0.01-0.02[36]
Crohn's disease	0.6–0.8[103]	0.1[11]
Ulcerative colitis	0.5[103]	0.05[12]
Multiple sclerosis	0.3–0.8[104]	0.1[45]
Ankylosing spondylitis	>0.90[105]	0.2[106]
Rheumatoid arthritis	0.6[107]	
Schizophrenia	0.7–0.8[108]	0.01[79]
Bipolar disorder	0.6–0.7[108]	0.02[79]
Breast cancer	0.3[110]	0.08[111]
Von Willebrand factor	0.66–0.75[112,113]	0.13[114]
Height	0.8[115,116]	0.1[13]
Bone mineral density	0.6-0.8[117]	0.05[118]
QT interval	0.37–0.60[119,120]	0.07[121]
HDL cholesterol	0.5[122]	0.1[57]
Platelet count	0.8[123]	0.05–0.1[58]

h^2 pedigree $=$
narrow sense heritability
(estimated from relateds)

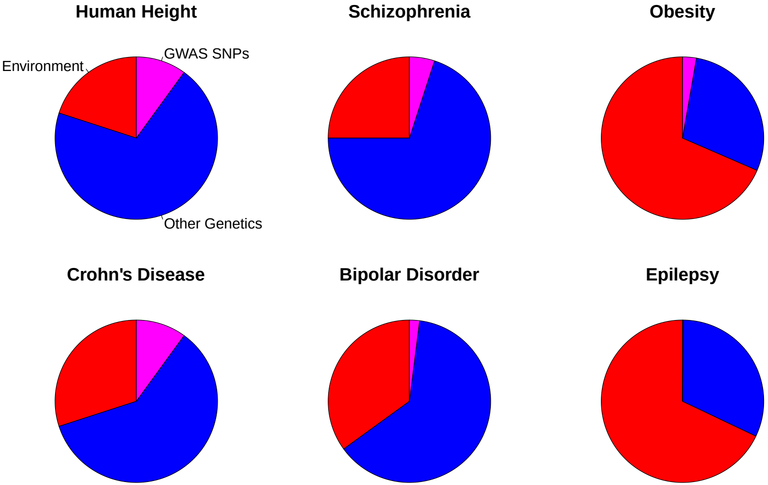
h^2 GWAS hits $= h^2_{GWAS}$

Five years of GWAS Discovery.
AJHG. 2012

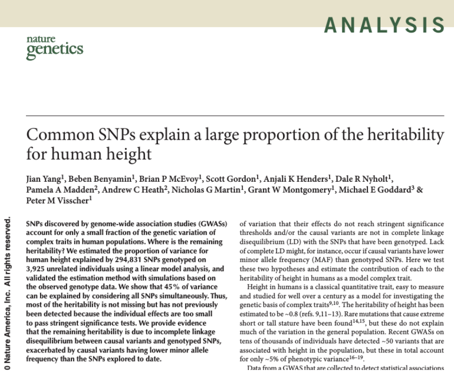
The missing heritability problem



The missing heritability problem



The missing heritability problem SOLVED



ANALYSIS

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]

0 Nature America, Inc. All rights reserved.

SNPs discovered by genome-wide association studies (GWASs) account for only a small fraction of the genetic variation of complex traits in human populations. Where is the remaining heritability? We estimated the proportion of variance for human height explained by 294,831 SNPs genotyped on 3,925 unrelated individuals using a linear model analysis, and validated the estimation method with simulations based on the observed genotype data. We show that 45% of variance can be explained by considering all SNPs simultaneously. Thus, most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests. We provide evidence that the remaining heritability is due to incomplete linkage disequilibrium between causal variants and genotyped SNPs, exacerbated by causal variants having lower minor allele frequency than the SNPs explored to date.

of variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage disequilibrium (LD) with the SNPs that have been genotyped. Lack of complete LD might, for instance, occur if causal variants have lower minor allele frequency (MAF) than genotyped SNPs. Here we test these two hypotheses and estimate the contribution of each to the heritability of height in humans as a model complex trait.

Height in humans is a classical quantitative trait, easy to measure and studied for well over a century as a model for investigating the genetic basis of complex traits[9,10]. The heritability of height has been estimated to be ~0.8 (refs. 9,11–13). Rare mutations that cause extreme short or tall stature have been found[14,15], but these do not explain much of the variation in the general population. Recent GWASs on tens of thousands of individuals have detected ~50 variants that are associated with height in the population, but these in total account for only ~5% of phenotypic variance[16–19].

Data from a GWAS that are collected to detect statistical associations

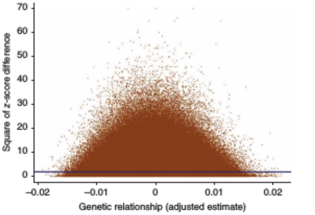
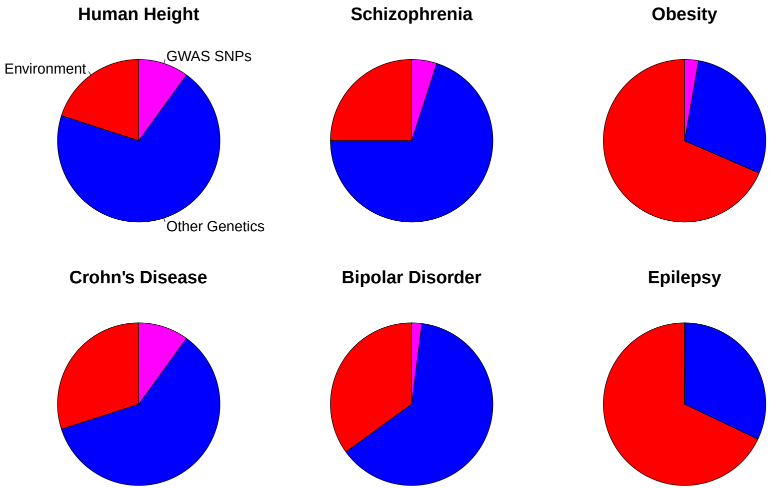


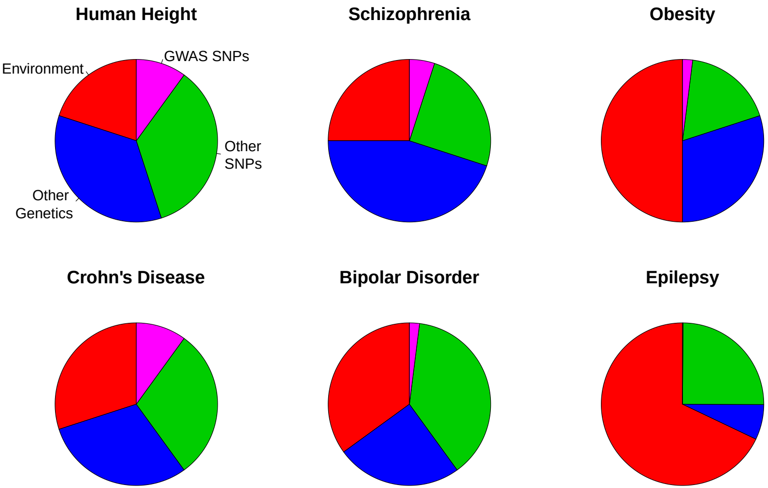
Figure 3 All pairwise comparisons contribute to the estimate of genetic variance. Shown are the squared *z*-score differences between individuals (Δy^2_{jk}) plotted against the adjusted estimates of genetic relationship (A^*_{jk}). The blue line is the linear regression line of Δy^2_{jk} on A^*_{jk}. The intercept and regression coefficient are estimates of twice the phenotypic variance and minus twice the genetic variances[23], respectively. The intercept is 1.98 (s.e. = 0.001), and the regression coefficient is –1.01 (s.e. = 0.27), consistent with estimates of the phenotypic and additive genetic variance of 0.990 and 0.505, respectively, and a proportion of variance explained by all SNPs of 0.51.

Estimating h^2_{SNP} using mixed model analysis with unrelateds found SNPs explain at least 45% of variation in height - over half the heritability

The missing heritability problem SOLVED



The missing heritability problem SOLVED



Why SNP heritability?

When our sample contains related individuals, h^2 was an estimate of narrow-sense heritability, the proportion of variance explained by ANY ADDITIVE genetic variation

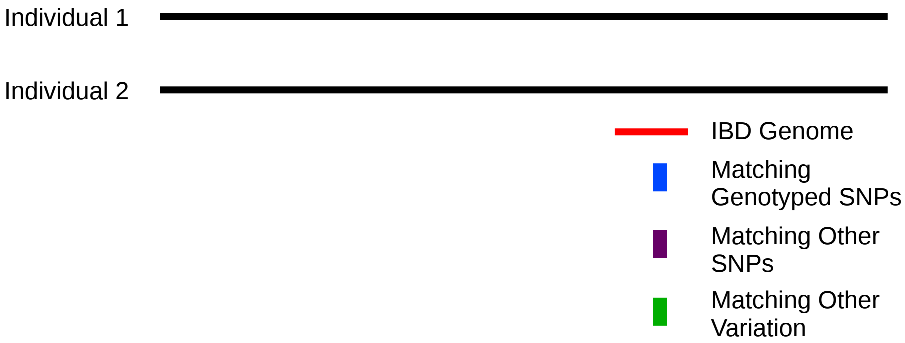
When individuals are unrelated, h^2 becomes an estimate of h^2_{SNP}, the total variance explained by all SNPs

This is because, when two individuals are related, the similarities between their SNP genotypes are (mainly) due to the relatedness, and there will be similar patterns of similarities between other types of genetic variation

When individuals are unrelated, any similarities are due to chance, so similarities observed across SNPs will be independent of (uncorrelated with) similarities across other types

Why SNP heritability? (Explanation 1)

Consider two full-sibs - they share 50% of genetic variation due to IBD



Individual 1

Individual 2

🟥 IBD Genome

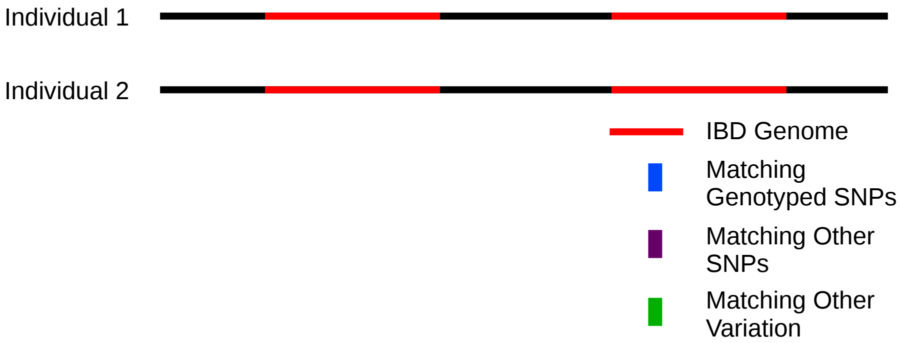
🟦 Matching Genotyped SNPs

🟪 Matching Other SNPs

🟩 Matching Other Variation

Why SNP heritability? (Explanation 1)

Consider two full-sibs - they share 50% of genetic variation due to IBD



Individual 1

Individual 2

🟥 IBD Genome

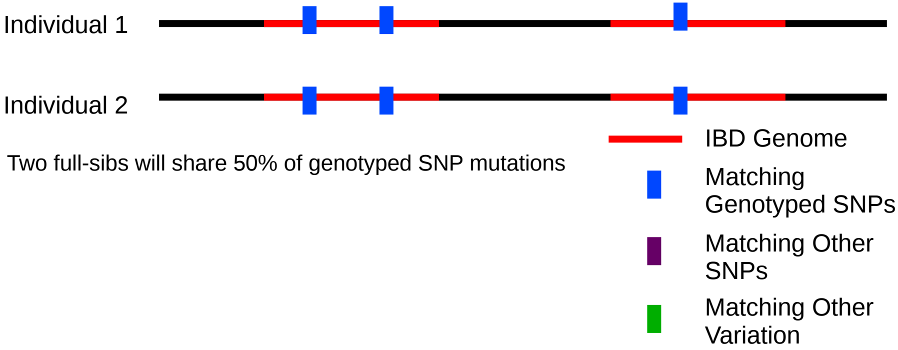
🟦 Matching Genotyped SNPs

🟪 Matching Other SNPs

🟩 Matching Other Variation

Why SNP heritability? (Explanation 1)

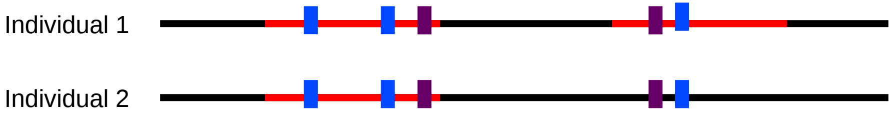
Consider two full-sibs - they share 50% of genetic variation due to IBD



Two full-sibs will share 50% of genotyped SNP mutations

Why SNP heritability? (Explanation 1)

Consider two full-sibs - they share 50% of genetic variation due to IBD



Two full-sibs will share 50% of genotyped SNP mutations

But will also share 50% of unobserved SNP mutations

🟥 IBD Genome

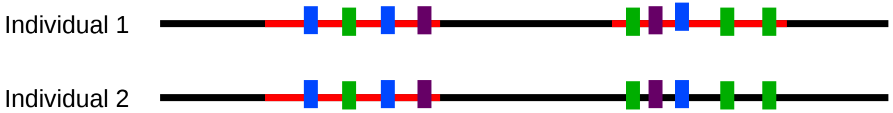
🟦 Matching Genotyped SNPs

🟪 Matching Other SNPs

🟩 Matching Other Variation

Why SNP heritability? (Explanation 1)

Consider two full-sibs - they share 50% of genetic variation due to IBD



- 🟥 IBD Genome
- 🟦 Matching Genotyped SNPs
- 🟪 Matching Other SNPs
- 🟩 Matching Other Variation

Two full-sibs will share 50% of genotyped SNP mutations

But will also share 50% of unobserved SNP mutations

…. and 50% of epigenetic effects

…. and 50% of proteomic effects

So can not tell whether any phenotypic similarity due to SNPs or other genetic variation

Why SNP heritability? (Explanation 1)

When two individuals are unrelated, they will still match for some genetic variation

But these matching variants will occur at random (be independent)

So if they match for one SNP, this does not mean they will match for other SNPs, or for other genetic variation

Therefore, if individuals with a particular SNP mutation tend to have higher phenotype, then the phenotypic similarity must be due to this SNP, rather than being due to a different source of genetic variation correlated with this SNP

Stay tuned for Explanation 2 later :)

GCTA: Recipe for a Nature Paper (2010-2012)

Collect GWAS data for a particular trait (say > 5000 individuals with genome-wide genotyping)

Compute allelic correlations K

Remove individuals so that no pair remains with $K_{i,j} > 0.05$

Perform REML to estimate Var(A) and Var(E)

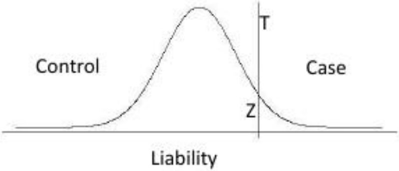
$h^2_{SNP} = Var(A)/Var(Y)$ is an estimate of the total variance explained by all SNPs

Write up paper explaining how much higher h^2_{SNP} than h^2_{GWAS}

The Liability Model

We only observe whether L above or below a threshold T

T determined by disease prevalence K



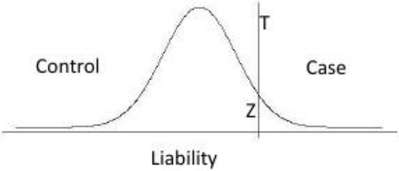
One way to model binary traits is to assume for each individual, there is a normally distributed, underlying liability. We can not observe the liability directly, but only know whether it is above (case) or below (control) a threshold, T

This is the model behind the probit link discussed in Module 3

The Liability Model

We only observe whether L above or below a threshold T

T determined by disease prevalence K



If we knew the liability L, we could fit $L \sim \mathbb{N}(\alpha, K\sigma_L^2 + I\sigma_e^2)$ and estimate $h^2_{Liab} = \sigma^2_L / Var(L)$, the heritability estimate on the liability scale directly

But with L unknown, we instead analyse the phenotype pretending it is continuous, then use the following transformation:

$$h^2_{Liab} = h^2_{SNP} \frac{K^2(1-K)^2}{P(1-P)z^2}$$

where K is the prevalence, P the ascertainment, and z is the "height of the standard normal distribution" at the liability threshold, T

GCTA: Recipe for a Nature Paper (2010-2012):

Since the application to human height in 2010, this approach has been applied to over 40 traits, including:

Crohns Disease
Type I Diabetes
Intelligence
Schizophrenia
Human Personality
Multiple Sclerosis
Cardiovascular Disease
Atherosclerotic Stroke
Executive Functioning
Rheumatoid Arthritis
Eating Disorders
Cannabis Use
Bipolar Disorder
Body Mass Index
Economic & Political Preferences
Parkinsons Disease
Major Depressive Disorder
Cilantro soapy taste detection
Childhood Leukaemia
Adult Antisocial Behaviour
Canine Leishmaniasis
Neuroticism & Extraversion
Life Span
Bird wing span

In two years, there were at least 10 publications in Nature or Nature Genetics, all identical except for the trait considered

h^2_{GWAS} vs h^2_{SNP} vs narrow sense heritability

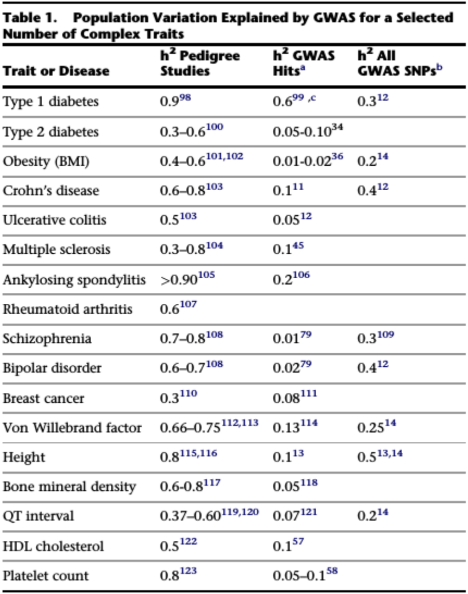


Table 1. Population Variation Explained by GWAS for a Selected Number of Complex Traits

Trait or Disease	**h^2 Pedigree Studies**	**h^2 GWAS Hits[a]**	**h^2 All GWAS SNPs[b]**
Type 1 diabetes	0.9[98]	0.6[99 ,c]	0.3[12]
Type 2 diabetes	0.3–0.6[100]	0.05-0.10[34]	
Obesity (BMI)	0.4–0.6[101,102]	0.01-0.02[36]	0.2[14]
Crohn's disease	0.6–0.8[103]	0.1[11]	0.4[12]
Ulcerative colitis	0.5[103]	0.05[12]	
Multiple sclerosis	0.3–0.8[104]	0.1[45]	
Ankylosing spondylitis	>0.90[105]	0.2[106]	
Rheumatoid arthritis	0.6[107]		
Schizophrenia	0.7–0.8[108]	0.01[79]	0.3[109]
Bipolar disorder	0.6–0.7[108]	0.02[79]	0.4[12]
Breast cancer	0.3[110]	0.08[111]	
Von Willebrand factor	0.66–0.75[112,113]	0.13[114]	0.25[14]
Height	0.8[115,116]	0.1[13]	0.5[13,14]
Bone mineral density	0.6-0.8[117]	0.05[118]	
QT interval	0.37–0.60[119,120]	0.07[121]	0.2[14]
HDL cholesterol	0.5[122]	0.1[57]	
Platelet count	0.8[123]	0.05–0.1[58]	

h^2 pedigree = narrow sense heritability (estimated from relateds)

h^2 GWAS hits = h^2_{GWAS}

h^2 all GWAS SNPs = h^2_{SNP}

Why is $h^2_{SNP} \gg h^2_{GWAS}$

Is it that estimates of h^2_{SNP} are wrong?

Suggested that estimates of h^2_{SNP} are inflated by genotyping errors or population stratification. (Population structure can inflate SNP-based heritability estimates. AJHG. 2011)

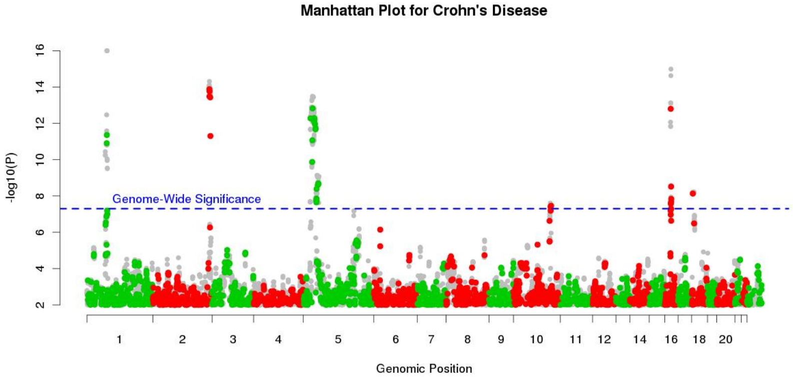
Inflation *is a problem* - you are estimating the total contribution of 500 000+ SNPs, so even if you over-estimate the contribution of each SNP by 0.00001% (one ten-thousandth of a percent), your estimate of h^2_{SNP} will be 50% higher than the truth

But shown that with careful quality control and checks, inflation can be avoided (Improved Heritability Estimation. AJHG. 2012)

Now estimates of h^2_{SNP} are generally accepted ... ish

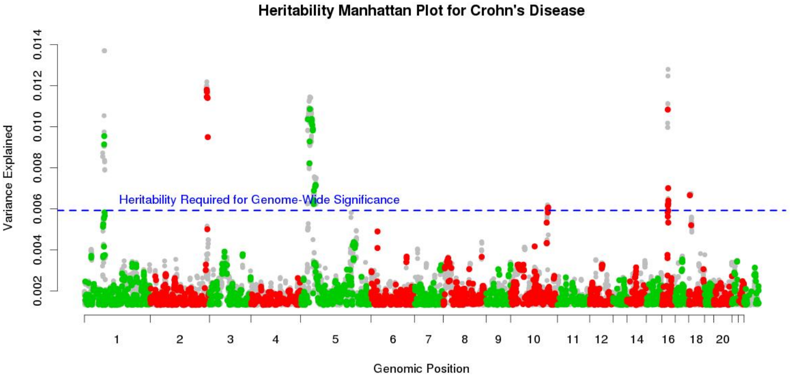
Why is $h^2_{SNP} \gg h^2_{GWAS}$

GWAS are only powered to find strong SNPs with large effect sizes



Why is $h^2_{SNP} \gg h^2_{GWAS}$

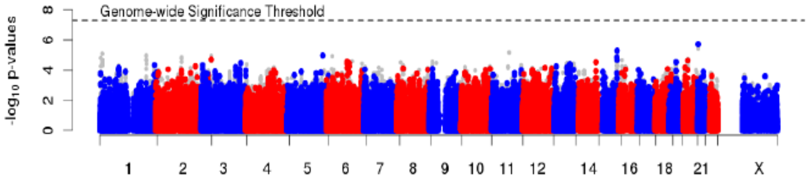
GWAS are only powered to find strong SNPs with large effect sizes



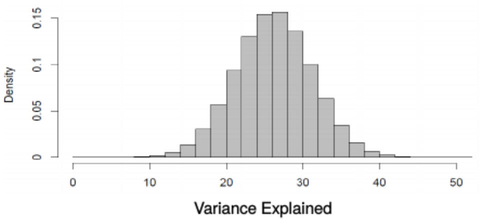
e.g., a GWAS with 5000 individuals can only find SNPs explaining at least 0.6% of variance. So one explanation is that most phenotypes are highly polygenic, with 100s or 1000s of SNPs causal, but most of these contribute only tiny heritability

Epilepsy

For epilepsy, no individual SNPs reach genome-wide significance

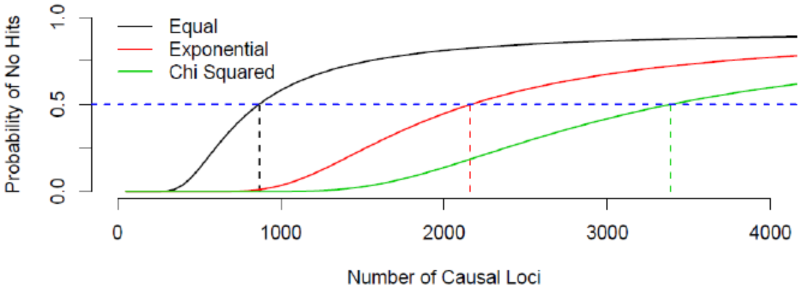


But collectively, all SNPs explain about 25% of (liability) variance



Epilepsy

Consider different ways heritability is distributed across causal loci for each, calculate probability of a GWAS finding no significant SNPs

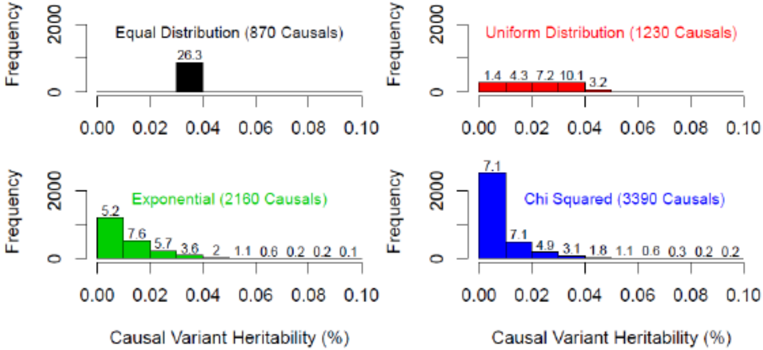


This gives an indication of how many SNPs contribute heritability

Describing the genetic architecture of epilepsy. Brain. 2014

Epilepsy

Which in turn allows us to speculate how heritability is spread across contributing SNPs



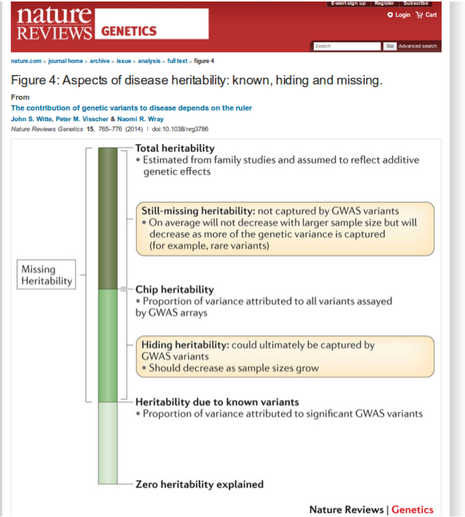
Describing the genetic architecture of epilepsy. Brain. 2014

Is the missing heritability problem solved?

For most phenotypes we have found that h^2_{GWAS} is much less than SNP heritability

...but that SNP heritability remains less than trait h^2

This has led to talk of the "still missing heritability problem"



nature.com » journal home » archive » issue » analysis » full text » figure 4

Figure 4: Aspects of disease heritability: known, hiding and missing.

From

The contribution of genetic variants to disease depends on the ruler

John S. Witte, Peter M. Visscher & Naomi R. Wray

Nature Reviews Genetics **15**, 765–776 (2014) | doi:10.1038/nrg3786

Total heritability
- Estimated from family studies and assumed to reflect additive genetic effects

Still-missing heritability: not captured by GWAS variants
- On average will not decrease with larger sample size but will decrease as more of the genetic variance is captured (for example, rare variants)

Missing Heritability

Chip heritability
- Proportion of variance attributed to all variants assayed by GWAS arrays

Hiding heritability: could ultimately be captured by GWAS variants
- Should decrease as sample sizes grow

Heritability due to known variants
- Proportion of variance attributed to significant GWAS variants

Zero heritability explained

Nature Reviews | Genetics

Reasons for $h^2_{SNP} <$ narrow-sense heritability

Even the latest genotyping arrays do not include all SNPs

SNP genotypes only focus on common variants (present in $> 1\%$ of population), so h^2_{SNP} does not capture the contribution of rare SNPs

SNPs are only one type of genetic variation; there are also structural variants (e.g., CNVs), epigenetic effects, and a whole host of "omics" (e.g., proteomics, lipidomics, transcriptomics, metabolics, etc)

1 SNP-based Heritability Analysis with Unrelated Individuals

2 Equivalence with Random Effects Regression

The Mixed Model

The mixed model of Module 7 states:[†] $Cov(Y) = \mathbf{K}\sigma_g^2 + \mathbf{I}\sigma_e^2$

This corresponds to assuming $Y \sim \mathbb{N}(\mathbf{Z}\boldsymbol{\theta}, \mathbf{K}\sigma_g^2 + \mathbf{I}\sigma_e^2)$

> $\boldsymbol{\theta}$ denotes fixed effects corresponding to $\mathbf{Z}$, a matrix of covariants

Note, that we could extend this to, say,

$$Cov(Y) = \mathbf{K}\sigma_g^2 + \boldsymbol{\gamma} Var(C) + \mathbf{I}\sigma_e^2,$$

but because we are using only unrelated individuals, the shared environment contributions (off-diagonal values of γ) are expected to be negligible

[†] Note, to be consistent with SNP-Based heritability analysis, we have replaced $Var(A)$ by σ_g^2 and $Var(E)$ by σ_e^2

The Mixed Model

With SNP data available, the most common way[†] to compute $\mathbf{K}$ is via allelic correlations, $\mathbf{K} = \mathbf{X}\mathbf{X}^T/N$,

With related individuals, allelic correlations are a good choice, because they provide an (almost) unbiased estimate of the coefficient of relatedness; i.e., the average allelic correlation for full-sibs will be 0.5 (strictly, the average will be 0.5-$1/n$, because SNP MAFs are estimated from the data)

When individuals are unrelated, we can justify the use of allelic correlations as the consequence of assuming a specific random effects regression model

[†] Alternative methods typically attempt to identify (relatively long) shared regions between pairs of individuals (e.g., FASTIBD, Chromopainter)

Motivating allelic correlations, $\mathbf{K} = XX^T/N$

Suppose we assume the following linear model:

$$
\begin{aligned}
\mathbf{Y} = \mathbf{Z}\boldsymbol{\theta} \\
&+ \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\
&+ \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} \\
&+ \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} X_{17} + \beta_{18} X_{18} + \beta_{19} X_{19} + \beta_{20} X_{20} + \beta_{21} X_{21} \\
&+ \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \beta_{25} X_{25} + \beta_{26} X_{26} + \beta_{27} X_{27} + \beta_{28} X_{28} \\
&+ \ldots + \beta_{500\,000} X_{500\,000} \\
&+ e,
\end{aligned}
$$

$$\text{where } \beta_j \sim \mathbb{N}(0, \sigma_g^2/N) \text{ and } e \sim \mathbb{N}(0, \sigma_e^2)$$

Then $g = \sum_{j=1}^{N} \beta_j X_j \sim \mathbb{N}(0, \mathbf{K}\sigma_g^2)$

and therefore $Y \sim \mathbb{N}(\mathbf{Z}\boldsymbol{\theta}, \mathbf{K}\sigma_g^2 + \mathbf{I}\sigma_e^2)$, where $\mathbf{K} = \mathbf{X}\mathbf{X}^T/N$

Motivating allelic correlations, $\mathbf{K} = XX^T/N$

So when we perform mixed model analysis with $Y \sim \mathbb{N}(\mathbf{Z}\boldsymbol{\theta}, \mathbf{K}\sigma_g^2 + I\sigma_e^2)$, where $\mathbf{K} = XX^T/N$, we are asking how much phenotypic variation is explained under a linear model in which every SNP is allowed to contribute towards the phenotype:

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\theta} + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_N X_N$$

and where we assume that each effect size has distribution $\mathbb{N}(0, \sigma_g^2/N)$ and that the noise terms have distribution $\mathbb{N}(0, \sigma_e^2)$

When individuals are "unrelated", each SNP X_j captures only the genetic variation at that basepair and very nearby basepairs (in high LD), and therefore, we end up with an estimate of how much phenotypic variation is explained BY THE SNPs

This is Explaination 2 ... easier, right?

Assumptions in SNP-Based Heritability Analysis

While (implicitly) assuming a specific random effects regression model, provides motivation for using allelic correlations

… it also makes clear that we are making a lot of assumptions

In particular, we assume:

- All SNPs are Causal
- Gaussian Effect Sizes
- Gaussian Noise Terms
- Inverse Relationship between MAF and Effect Size

  i.e., all SNPs contribute equally h^2

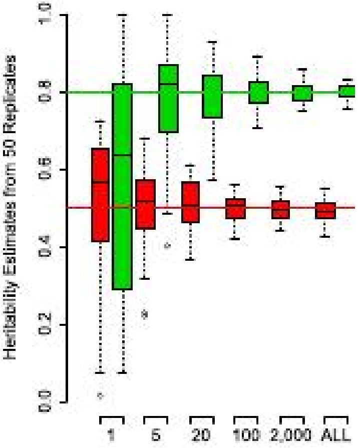
Therefore, we set out to test these assumptions:

Improved Heritability Estimation from Genome-wide SNPs, AJHG (2012)

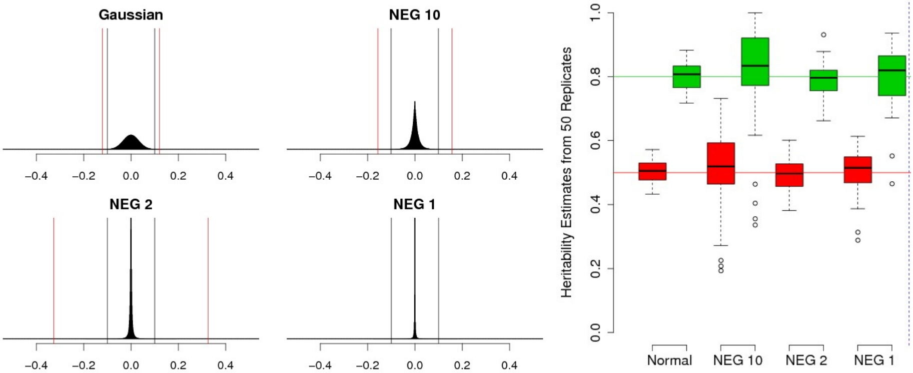
Assuming All SNPs are Causal

We simulated traits with varying numbers of SNPs contributing heritability

Even when relatively few were causal, estimation remained reasonably precise



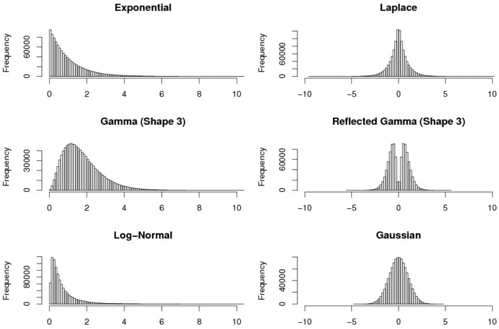
Assuming Gaussian Effect Sizes

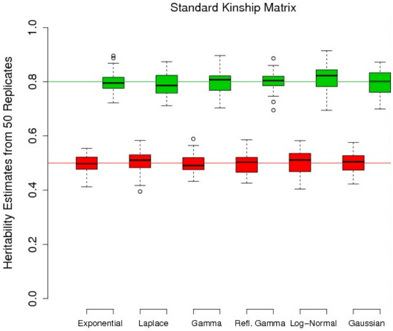


We simulated using alternative distributions for effect size

estimation remained reasonably good

Assuming Gaussian Noise Terms

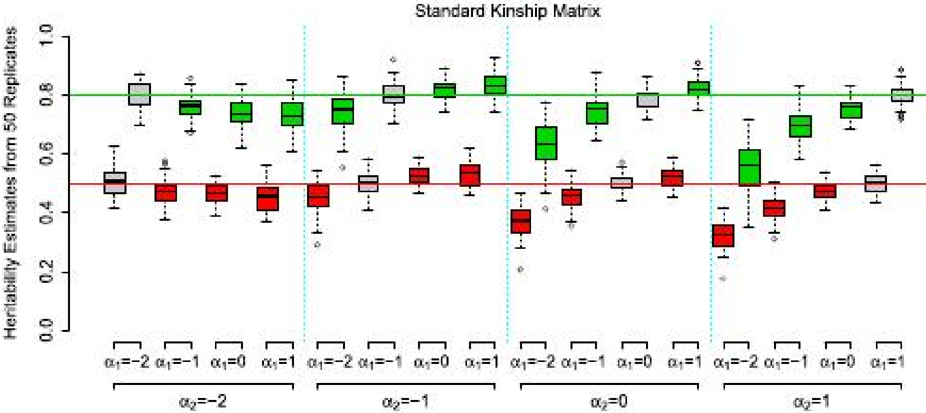




We simulated using alternative distributions for noise terms

estimation remained reasonably good

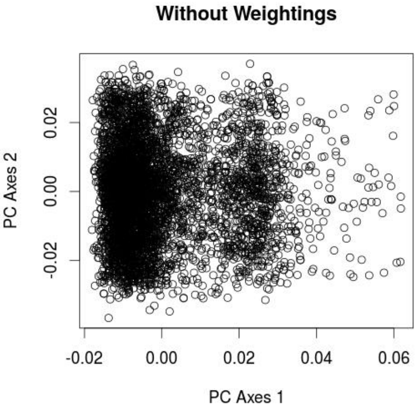
Assuming All SNPs Contribute Equal h^2



We simulated using alternative relationships between MAF and effect size

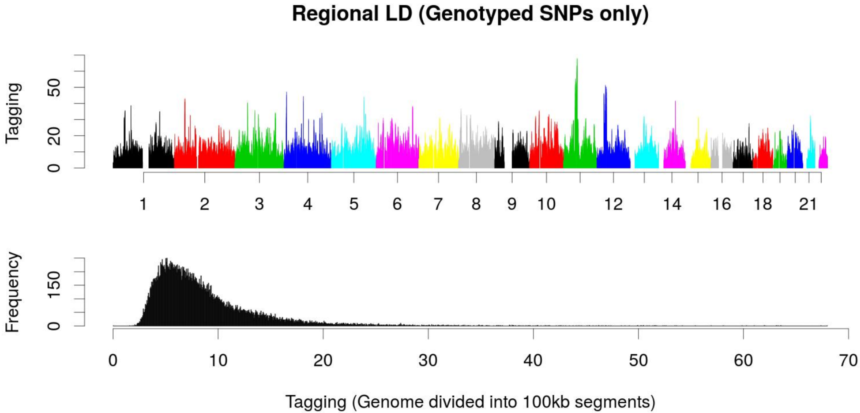
estimation remained reasonably good

Estimates are Sensitive to (Uneven) Linkage Disequilibrium



A common problem when performing principal component analysis

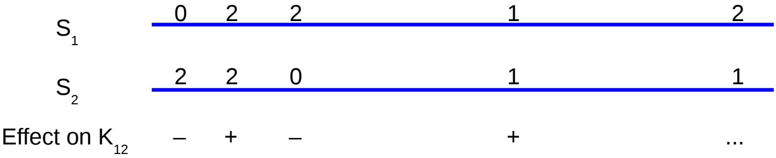
Levels of LD vary greatly across the genome



Values are sums of r^2 between each SNP and neighbours within 100 kb

Estimates are Sensitive to LD

Allelic correlations represent average genome-wide similarity



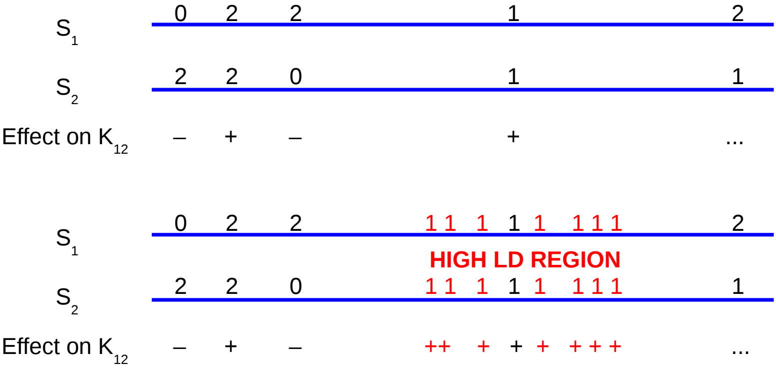
S_1 0 2 2 1 2

S_2 2 2 0 1 1

Effect on K_{12} – + – + ...

Estimates are Sensitive to LD

Allelic correlations represent average genome-wide similarity

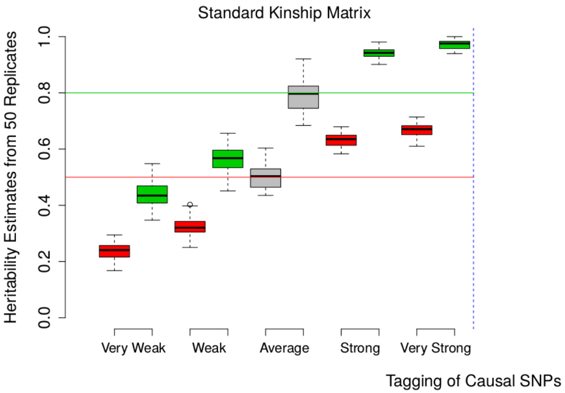


S_1	0	2	2	1	2
S_2	2	2	0	1	1
Effect on K_{12}	–	+	–	+	...

S_1	0	2	2	1 1 1 1 1 1 1 1	2
				HIGH LD REGION	
S_2	2	2	0	1 1 1 1 1 1 1 1	1
Effect on K_{12}	–	+	–	++ + + + + + +	...

More highly tagged genetic variation contributes more to **K**

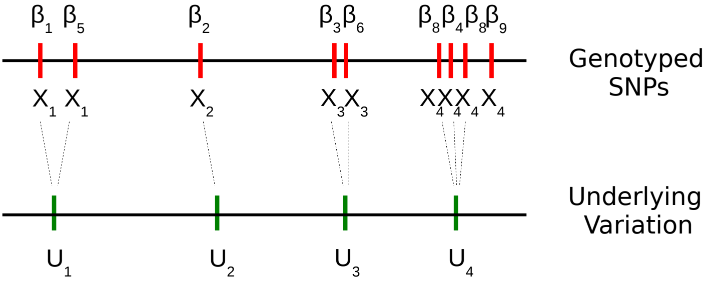
Estimates Can be Sensitive to LD of Causal Variants



Causal variants in high LD areas ⇒ over-estimation of h^2_{SNP}

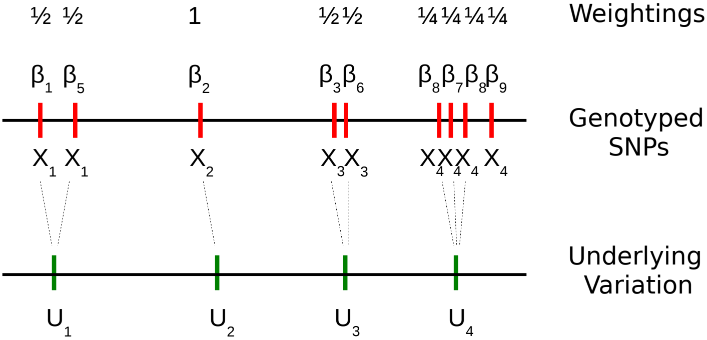
Causal variants in low LD areas ⇒ under-estimation of h^2_{SNP}

Adjusting for Uneven Tagging



LDAK assumes that the observed SNPs are tagging independent underlying signal

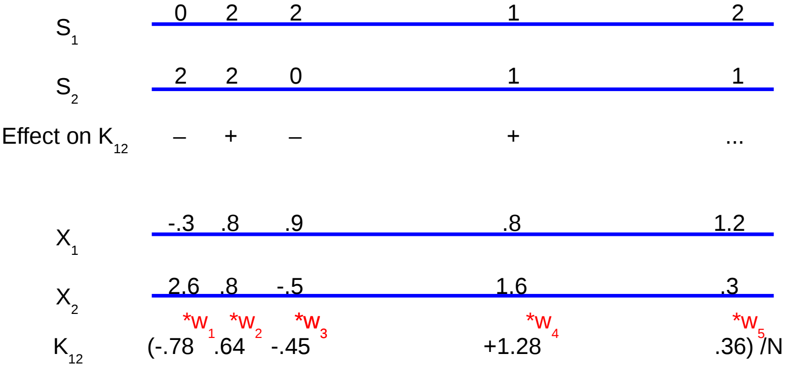
Adjusting for Uneven Tagging



...then calculates SNP weightings so that each underlying signal contributes once

Estimates are Sensitive to LD

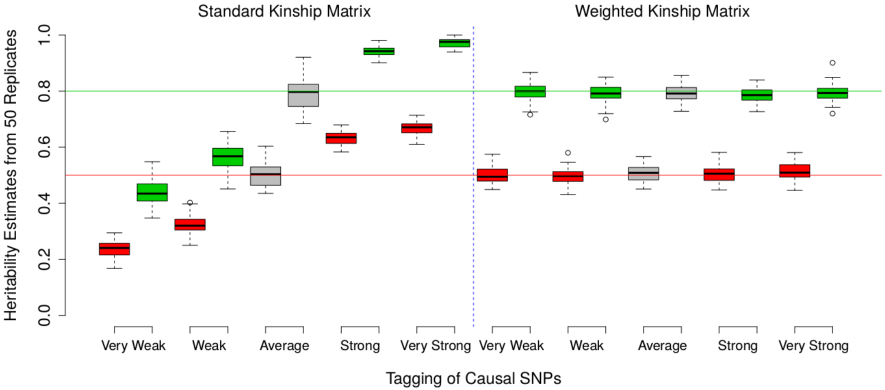
Then instead of computing standard (unweighted) allelic correlations where each SNP contributes evenly to **K**



S_1	0	2	2	1	2
S_2	2	2	0	1	1
Effect on K_{12}	–	+	–	+	...
X_1	-.3	.8	.9	.8	1.2
X_2	2.6	.8	-.5	1.6	.3
	$*w_1$	$*w_2$	$*w_3$	$*w_4$	$*w_5$
K_{12}	(-.78	.64	-.45	+1.28	.36) /N

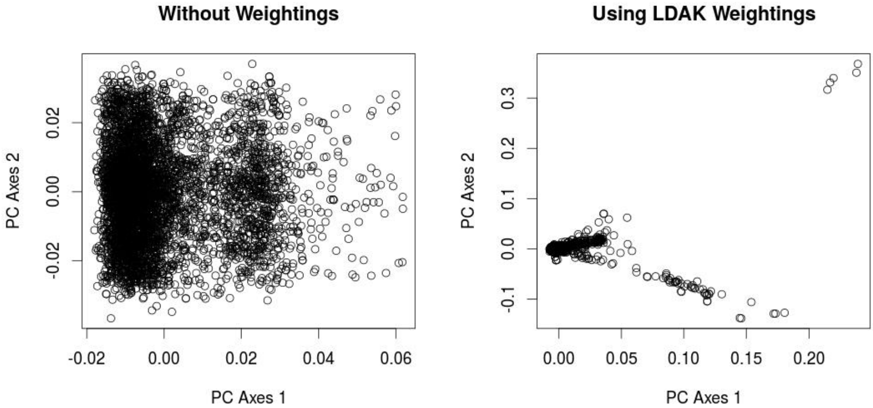
LDAK constructs "LD-Adjusted" Allelic Correlations where the contribution of each SNP is weighted according to local patterns of LD

LDAK: Linkage Disequilibrium Adjusted Kinships



LDAK estimates are unaffected by whether causal variants are in low or high LD regions (e.g., tend to be rare or common)

LDAK: Linkage Disequilibrium Adjusted Kinships

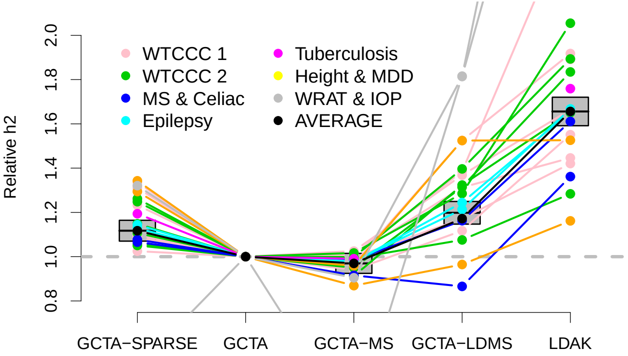


LDAK weights offer an alternative to pruning

e.g., when performing PCA or computing genetic profile risk scores

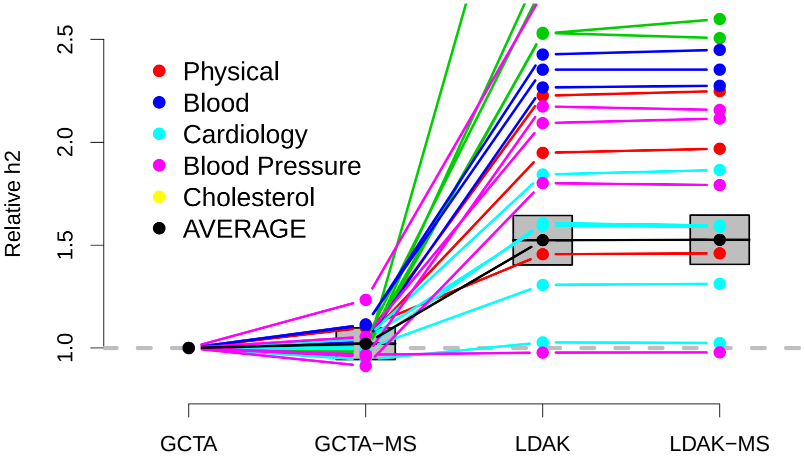
Different Methods Give Different Estimates

For each of 22 (almost) independent GWAS traits (on average 6500 individuals), we estimated h^2_{SNP} from imputed SNP data (N =2-4 M) using LDAK and three versions of GCTA



MS: MAF Stratification; **LDMS**: LD+MAF Stratification.
GCTA-SPARSE is **GCTA** using only genotyped SNPs (N =200-500 k)

Different Methods Give Different Estimates



Also estimated analysed 21 traits where individuals were genotyped using the "Metabochip" (captures approximately 25% of genome-wide variation)

Different Methods Give Different Estimates

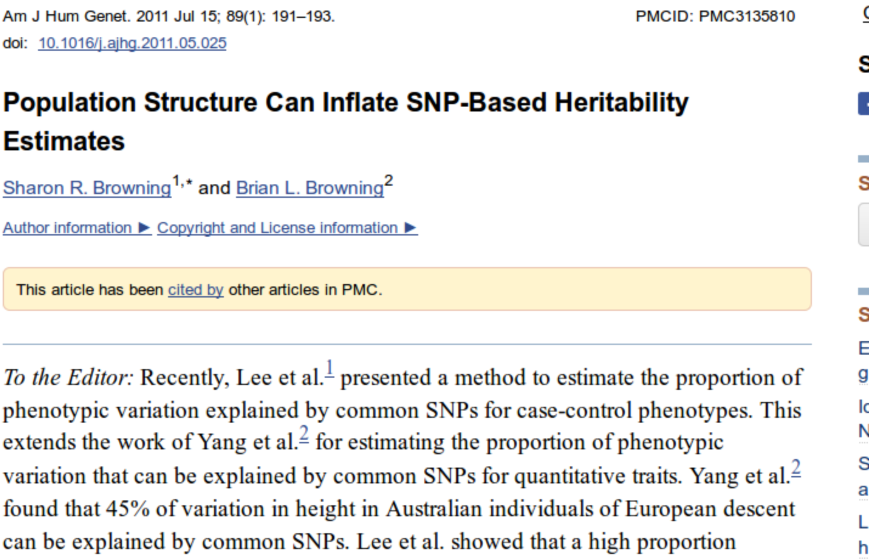
GCTA weights each SNP equally

LDAK (attempts to) weight each source of genetic variation equally

Across 22 (21) traits, estimates of h^2_{SNP} from LDAK are on average 66% (50%) higher than those from GCTA

Neither the GCTA nor LDAK assumption is correct - but if the LDAK model was closer to the truth, then this indicates the "still missing heritability" is even smaller

Inflation due to Population Structure



Am J Hum Genet. 2011 Jul 15; 89(1): 191–193. PMCID: PMC3135810
doi: 10.1016/j.ajhg.2011.05.025

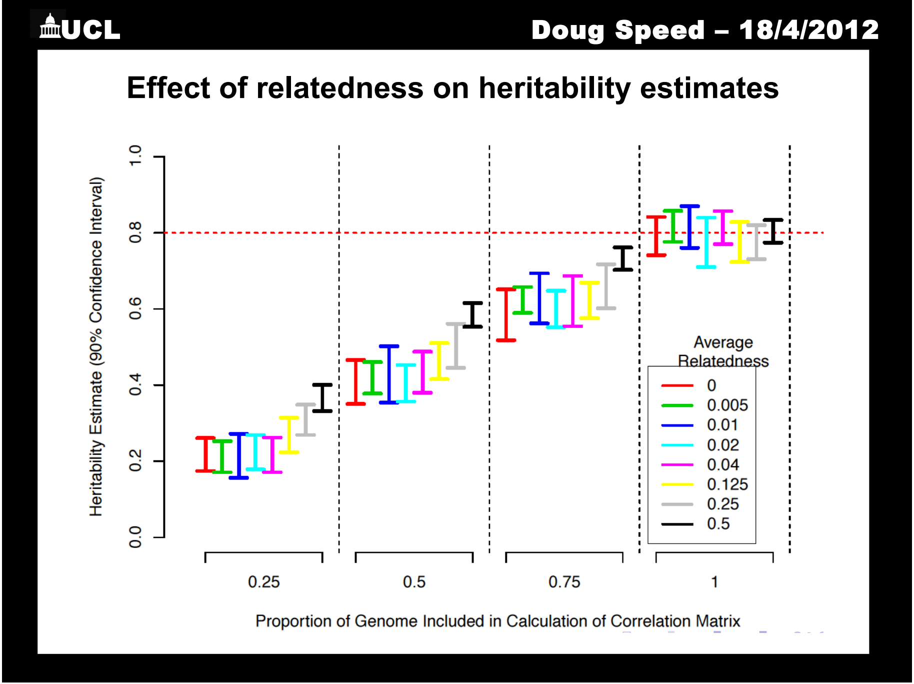
Population Structure Can Inflate SNP-Based Heritability Estimates

Sharon R. Browning[1,*] and Brian L. Browning[2]

Author information ▶ Copyright and License information ▶

This article has been cited by other articles in PMC.

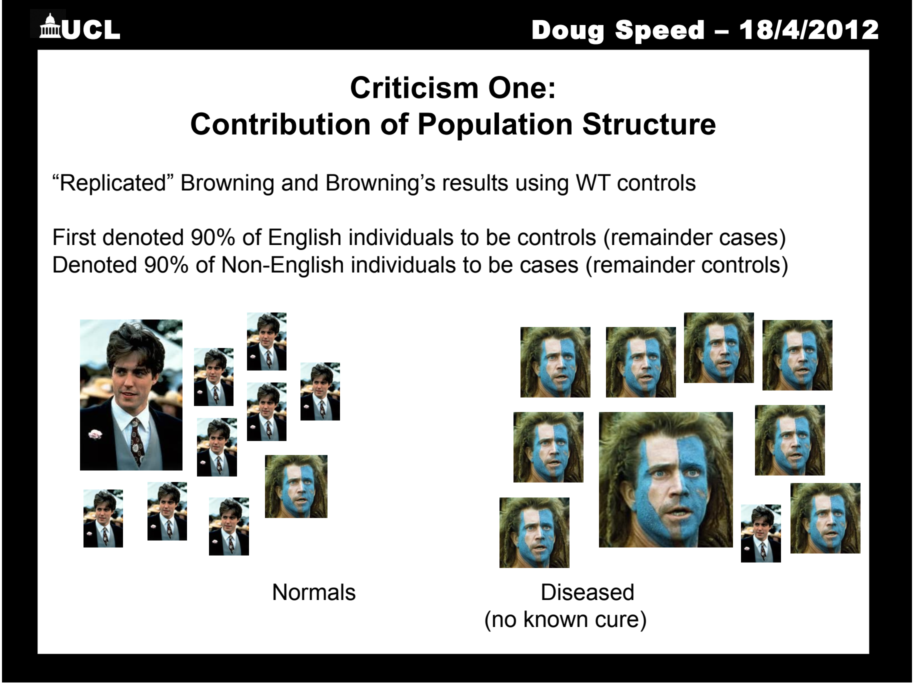
To the Editor: Recently, Lee et al.[1] presented a method to estimate the proportion of phenotypic variation explained by common SNPs for case-control phenotypes. This extends the work of Yang et al.[2] for estimating the proportion of phenotypic variation that can be explained by common SNPs for quantitative traits. Yang et al.[2] found that 45% of variation in height in Australian individuals of European descent can be explained by common SNPs. Lee et al. showed that a high proportion



UCL

Doug Speed – 18/4/2012

Effect of relatedness on heritability estimates



UCL Doug Speed – 18/4/2012

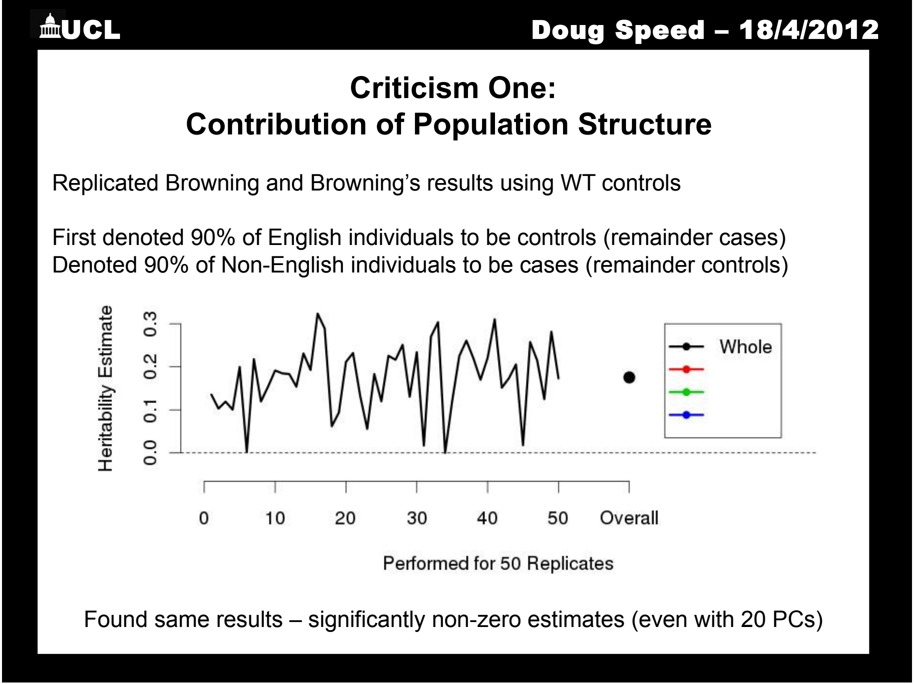
Criticism One: Contribution of Population Structure

“Replicated” Browning and Browning’s results using WT controls

First denoted 90% of English individuals to be controls (remainder cases)
Denoted 90% of Non-English individuals to be cases (remainder controls)

Normals

Diseased
(no known cure)



UCL

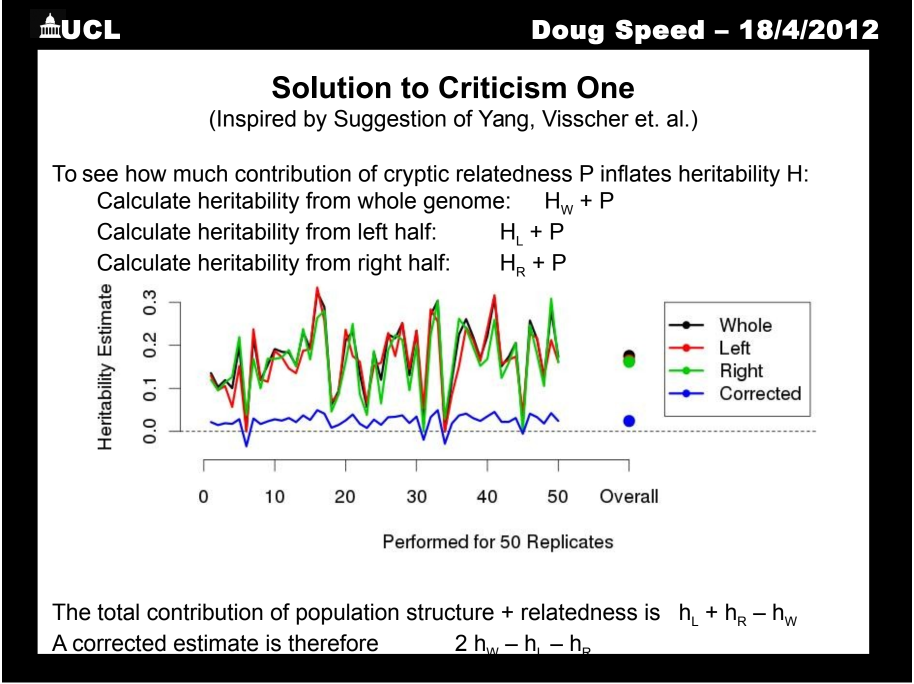
Doug Speed – 18/4/2012

Criticism One:
Contribution of Population Structure

Replicated Browning and Browning's results using WT controls

First denoted 90% of English individuals to be controls (remainder cases)
Denoted 90% of Non-English individuals to be cases (remainder controls)

Found same results – significantly non-zero estimates (even with 20 PCs)



UCL Doug Speed – 18/4/2012

Solution to Criticism One

(Inspired by Suggestion of Yang, Visscher et. al.)

To see how much contribution of cryptic relatedness P inflates heritability H:

Calculate heritability from whole genome: $H_W + P$

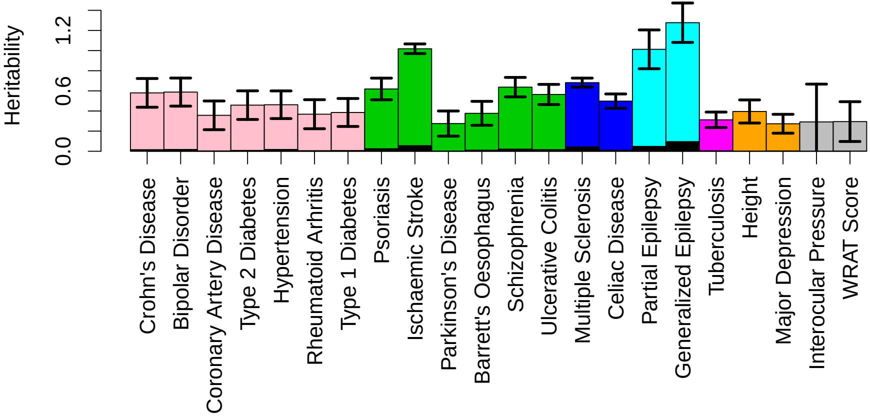
Calculate heritability from left half: $H_L + P$

Calculate heritability from right half: $H_R + P$

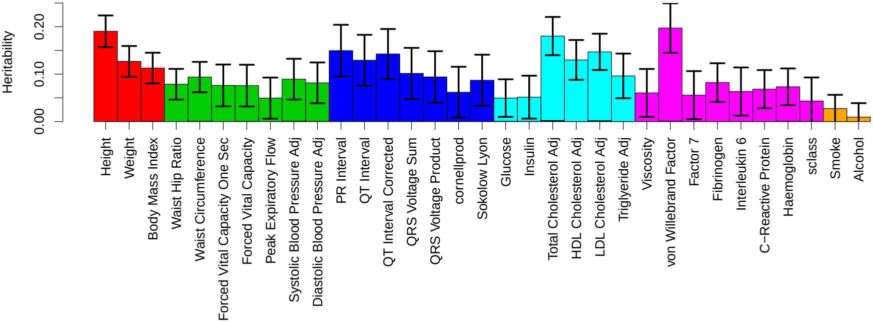
The total contribution of population structure + relatedness is $h_L + h_R – h_W$

A corrected estimate is therefore $2\,h_W – h_L – h_R$

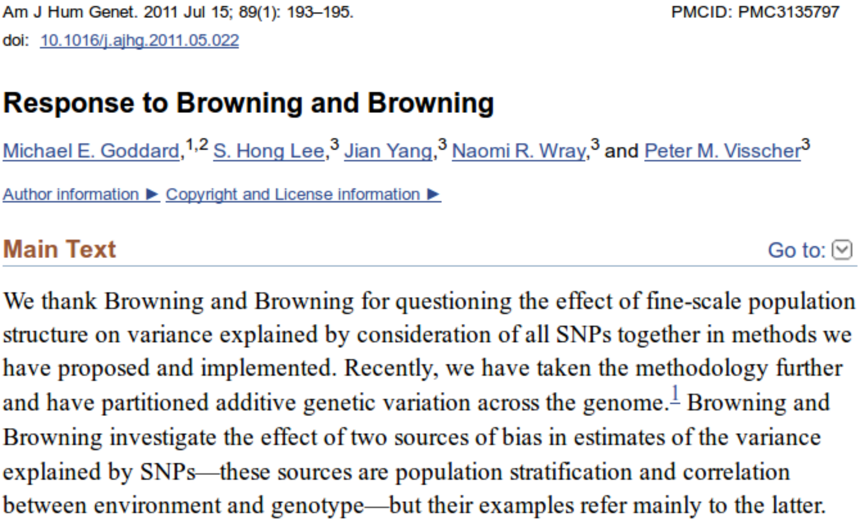
Testing for Inflation due to Population Structure



Testing for Inflation due to Population Structure



Inflation due to Population Structure



Am J Hum Genet. 2011 Jul 15; 89(1): 193–195.
PMCID: PMC3135797
doi: 10.1016/j.ajhg.2011.05.022

Response to Browning and Browning

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

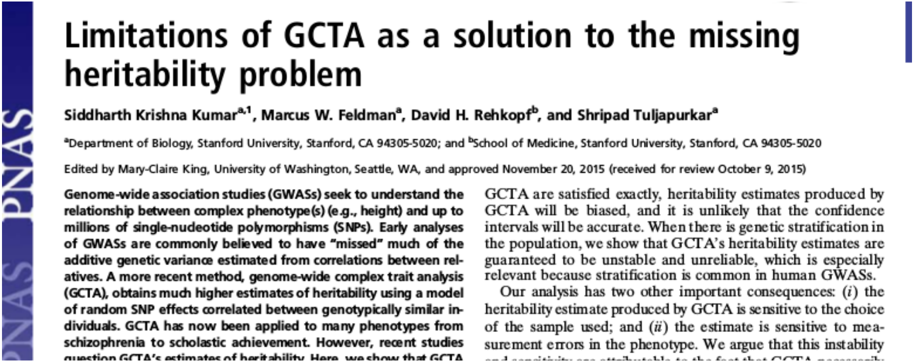
Author information ▶ Copyright and License information ▶

Main Text

Go to: ⌄

We thank Browning and Browning for questioning the effect of fine-scale population structure on variance explained by consideration of all SNPs together in methods we have proposed and implemented. Recently, we have taken the methodology further and have partitioned additive genetic variation across the genome.[1] Browning and Browning investigate the effect of two sources of bias in estimates of the variance explained by SNPs—these sources are population stratification and correlation between environment and genotype—but their examples refer mainly to the latter.

"Limitations of GCTA"



Limitations of GCTA as a solution to the missing heritability problem

Siddharth Krishna Kumar[a,1], Marcus W. Feldman[a], David H. Rehkopf[b], and Shripad Tuljapurkar[a]

[a]Department of Biology, Stanford University, Stanford, CA 94305-5020; and [b]School of Medicine, Stanford University, Stanford, CA 94305-5020

Edited by Mary-Claire King, University of Washington, Seattle, WA, and approved November 20, 2015 (received for review October 9, 2015)

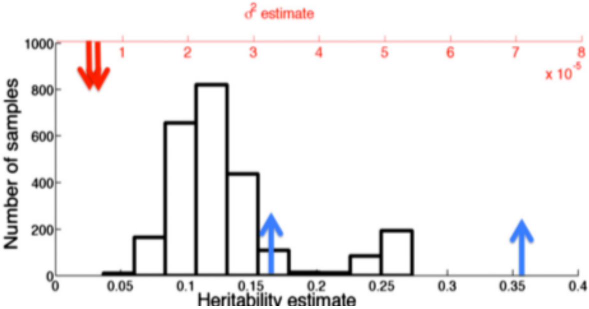
Genome-wide association studies (GWASs) seek to understand the relationship between complex phenotype(s) (e.g., height) and up to millions of single-nucleotide polymorphisms (SNPs). Early analyses of GWASs are commonly believed to have "missed" much of the additive genetic variance estimated from correlations between relatives. A more recent method, genome-wide complex trait analysis (GCTA), obtains much higher estimates of heritability using a model of random SNP effects correlated between genotypically similar individuals. GCTA has now been applied to many phenotypes from schizophrenia to scholastic achievement. However, recent studies question GCTA's estimates of heritability. Here, we show that GCTA

GCTA are satisfied exactly, heritability estimates produced by GCTA will be biased, and it is unlikely that the confidence intervals will be accurate. When there is genetic stratification in the population, we show that GCTA's heritability estimates are guaranteed to be unstable and unreliable, which is especially relevant because stratification is common in human GWASs.

Our analysis has two other important consequences: (*i*) the heritability estimate produced by GCTA is sensitive to the choice of the sample used; and (*ii*) the estimate is sensitive to measurement errors in the phenotype. We argue that this instability and sensitivity are attributable to the fact that GCTA necessarily

Kumar et al, PNAS (December 2015)

"Limitations of GCTA"



Analysed 2 698 individuals from the Framingham study, which (in their words) "is known to be stratified"

They first estimated h^2_{SNP} for Blood Pressure to be 26.3 (SD 5)

Then they repeatly estimated $h^2_{10\%}$, the variance explained by a random 10% of the genome. $h^2_{10\%}$ was typically much higher than 2.6%, which they declared "proof" that the approach is flawed

Summary

For many years, estimating h^2 involved recruiting related individuals

Recently, it was realised that with SNP data, the same methods could be applied to unrelated individuals. The resulting estimates correspond to h^2_{SNP}, the total proportion of phenotypic variance explained by all SNP

For human traits, this approach has largely solved the missing heritability debate, and as we will see in Module 14, this is just the start of things!

SNP-based heritability analysis is largely accepted, although some Americans still don't believe

but then a third of Americans believe in Donald Trump ...