Estimating “SNP Heritability” using GCTA/GREML

David Evans

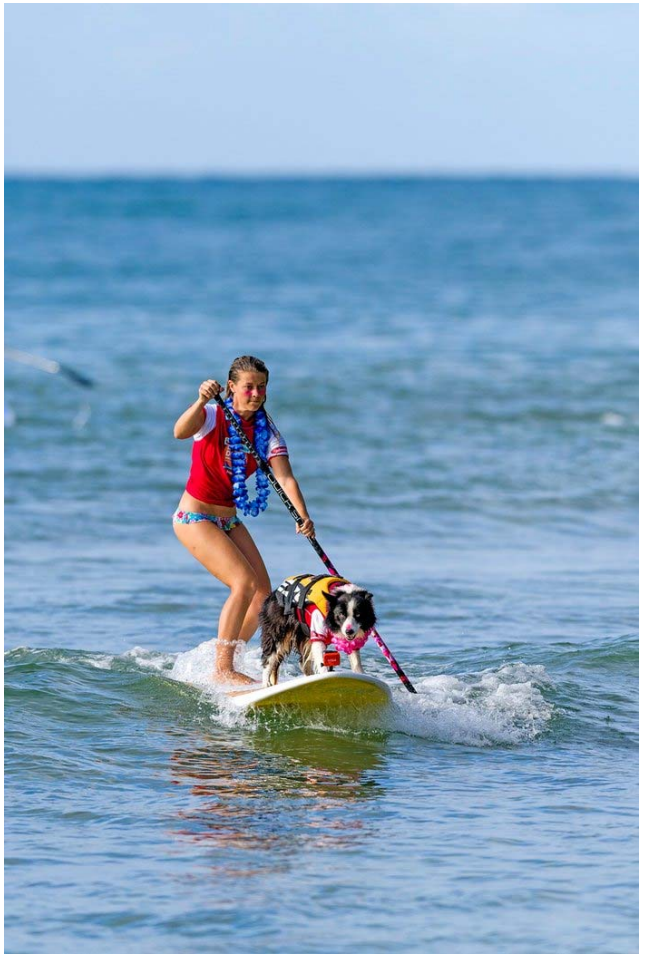
University of Queensland

Behavior Genetics Meeting
Brisbane 2016

View from Evans' Laboratory*

*Presenter makes no guarantees wrt veracity of statements made in the course of this presentation

Behavior Genetics Meeting
Brisbane 2016



Nick Martin

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

The case of the missing heritability

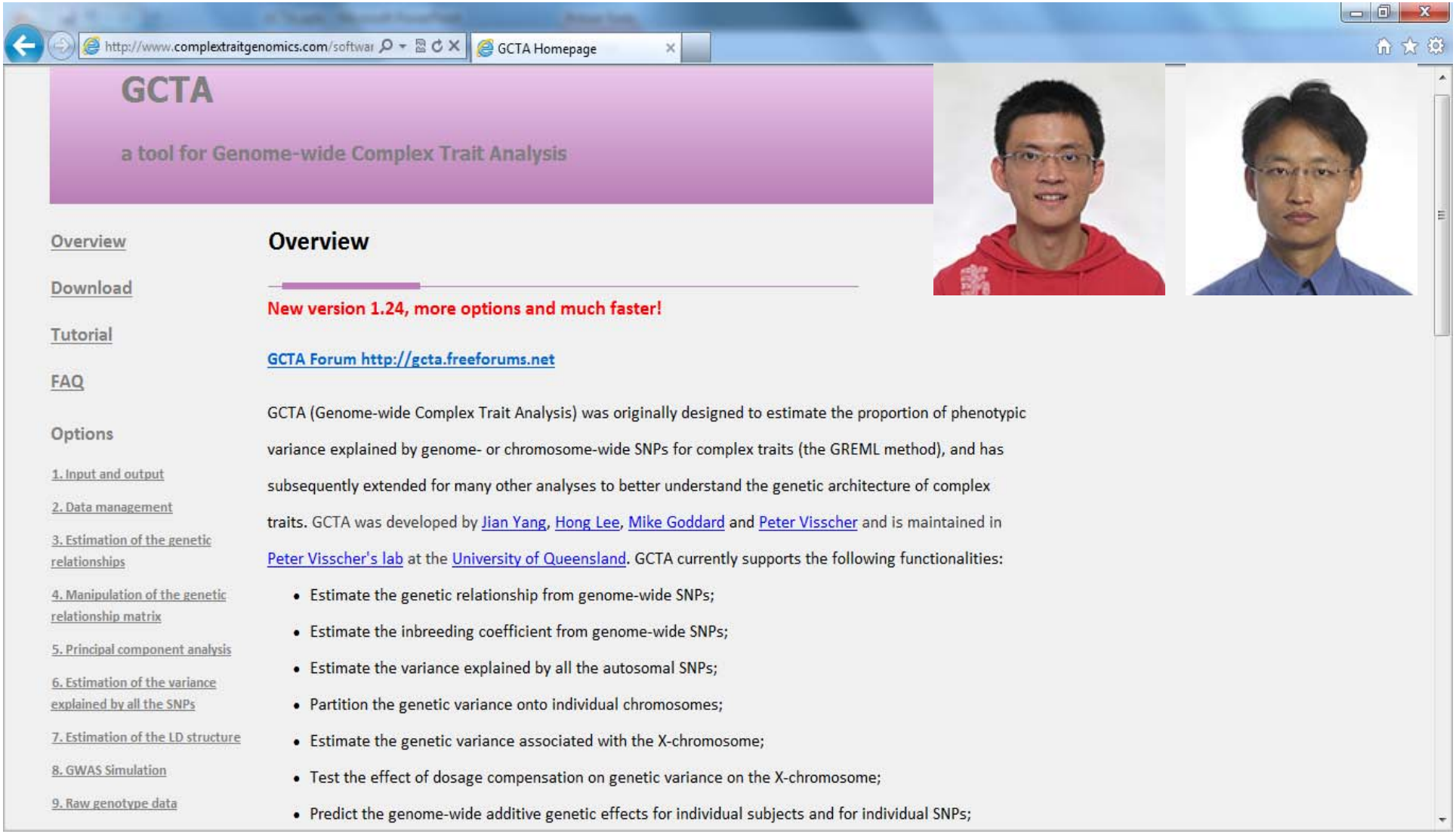
When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

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://cnsgenomics.com/software/gcta/



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

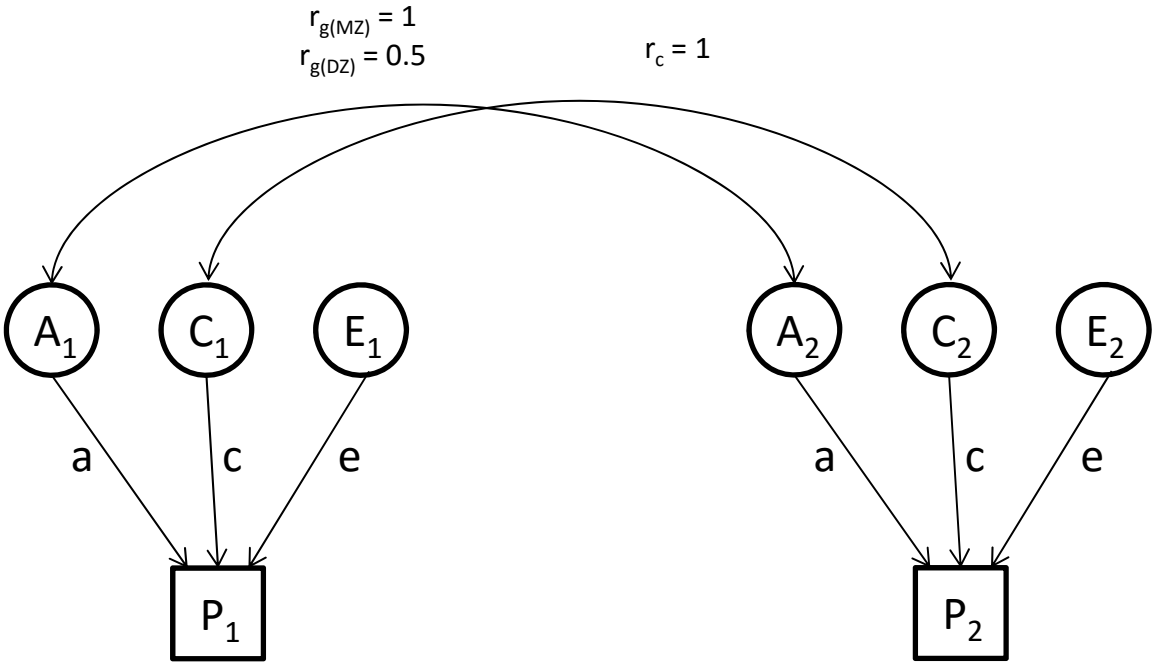
Sang Hong Lee,[1] Naomi R. Wray,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1,*]

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]

The Classical Twin Design



$$P_1 = aA_1 + cC_1 + eE_1$$

$$P_2 = aA_2 + cC_2 + eE_2$$

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

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

Expected Covariance Matrix Twin Pairs (AE Model)

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

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

(2 x 2) (2 x 2) (2 x 2) (2 x 2)

V is the expected phenotypic covariance matrix

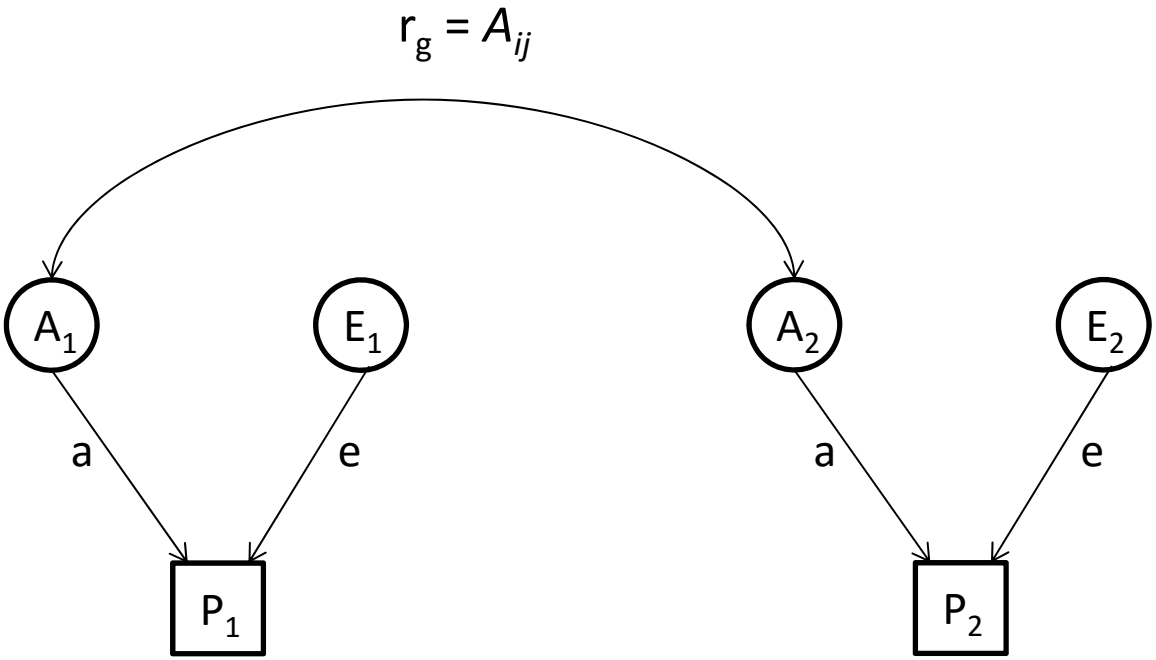
$\boldsymbol{\sigma^2_A}$ is the additive genetic variance

$\boldsymbol{\sigma^2_E}$ 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

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

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

(n x n) (n x n) (n x n)

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*

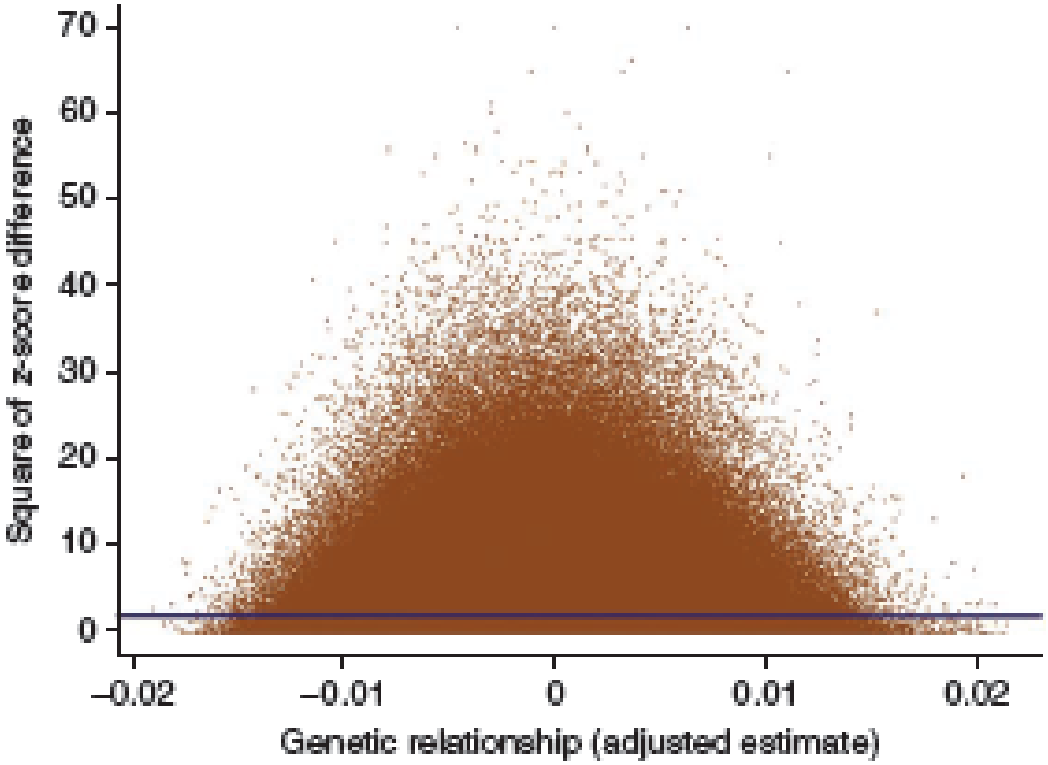
I is an identity matrix

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 i^{th} SNP of the j^{th} 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 (closely related individuals exert undue influence and may share common environments)
- 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^2_G (s.e.)[c]	*P*	h^2_G (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

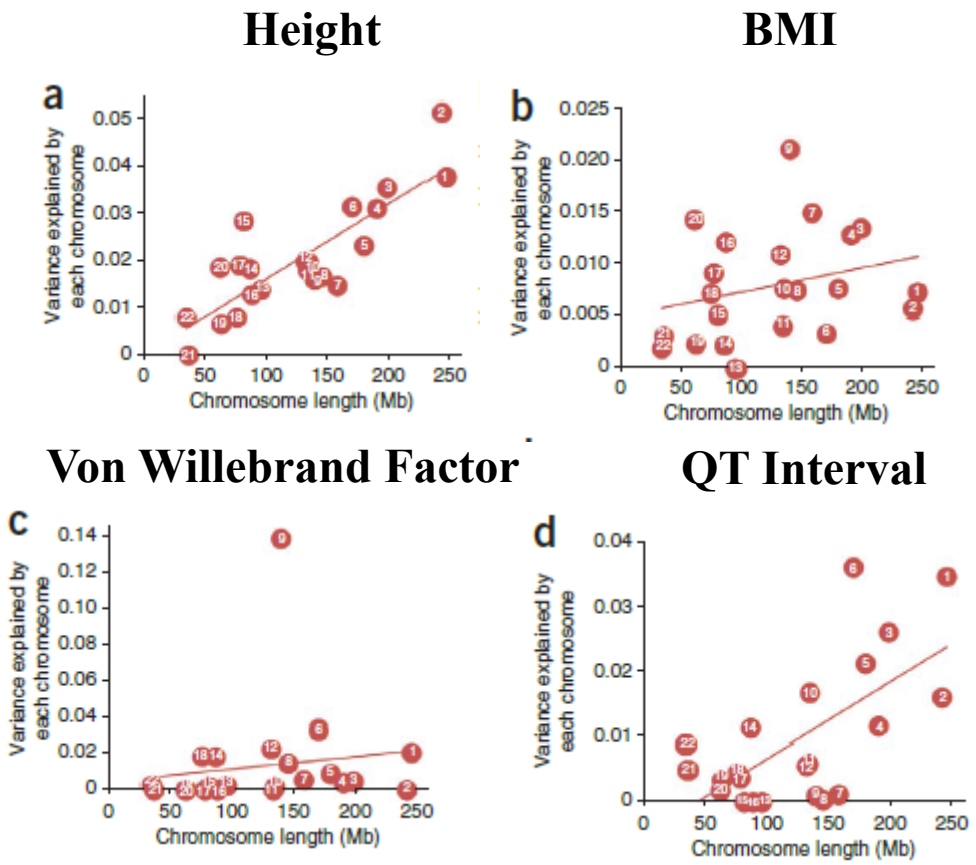
- GCTA explains the proportion of variance explained by SNPs on the microarray ("SNP heritability")
- GCTA does not estimate "heritability"
- GCTA does not estimate the proportion of trait variance due to common SNPs

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



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

Extending the Model - Binary Traits

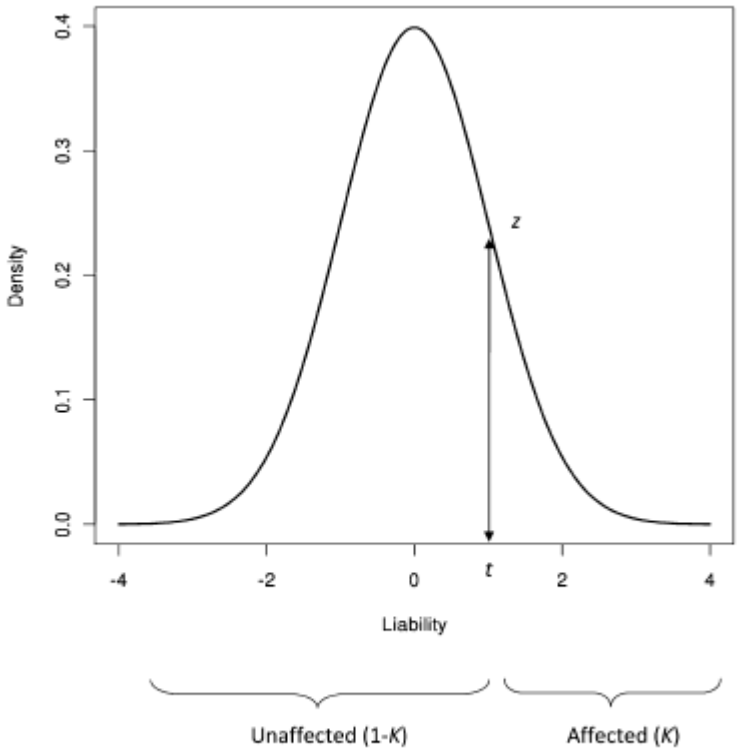


Figure 1. The Liability Threshold Model for a Disease Prevalence of *K*

- 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

GREML-LDMS

nature genetics

Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index

Jian Yang[1,2,24], Andrew Bakshi[1], Zhihong Zhu[1], Gibran Hemani[1,3], Anna A E Vinkhuyzen[1], Sang Hong Lee[1,4], Matthew R Robinson[1], John R B Perry[5], Ilja M Nolte[6], Jana V van Vliet-Ostaptchouk[6,7], Harold Snieder[6], The LifeLines Cohort Study[8], Tonu Esko[9–12], Lili Milani[9], Reedik Mägi[9], Andres Metspalu[9,13], Anders Hamsten[14], Patrik K E Magnusson[15], Nancy L Pedersen[15], Erik Ingelsson[16,17], Nicole Soranzo[18,19], Matthew C Keller[20,21], Naomi R Wray[1], Michael E Goddard[22,23] & Peter M Visscher[1,2,24]

We propose a method (GREML-LDMS) to estimate heritability for human complex traits in unrelated individuals using whole-genome sequencing data. We demonstrate using simulations based on whole-genome sequencing data that ~97% and ~68% of variation at common and rare variants, respectively, can be captured by imputation. Using the GREML-LDMS method, we estimate from 44,126 unrelated individuals that all ~17 million imputed variants explain 56% (standard error (s.e.) = 2.3%) of variance for height and 27% (s.e. = 2.5%) of variance for body mass index (BMI), and we find evidence that height- and BMI-associated variants have been under natural selection. Considering the imperfect tagging of imputation and potential overestimation of heritability from previous family-based studies, heritability is likely to be 60–70% for height and 30–40% for BMI. Therefore, the missing heritability is small for both traits. For further discovery of genes associated with complex traits, a study design with SNP arrays followed by imputation is more cost-effective than whole-genome sequencing at current prices.

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with hundreds of human complex traits and diseases[1]. However, genome-wide significant SNPs often explain only a small proportion of the heritability estimated from family-based studies, in the so-called 'missing heritability' problem[2]. Recent studies have shown that the total variance explained by all common SNPs is a large proportion of the heritability for complex traits and diseases[3,4]. This implies that much of the missing heritability is due to variants whose effects are too small to reach the level of genome-wide significance. This conclusion is supported by recent findings that complex traits and diseases such as height, BMI, age at menarche, inflammatory bowel diseases and schizophrenia are influenced by hundreds or even thousands of genetic variants of small effect[5–9]. Nevertheless, the genetic variance accounted for by all common SNPs is still less than that expected from family-based studies, and there has not been a consensus explanation for the missing heritability problem[2]. There are three major hypotheses. The first hypothesis is that missing heritability is largely due to rare variants of large effect, which are neither on the currently available commercial SNP arrays nor well tagged by the SNPs on the arrays. Here we define rare variants as variants with a minor allele frequency (MAF) of ≤0.01. To genotype rare variants with reasonably high accuracy, whole-genome sequencing with sufficiently high coverage in a large sample is required. The second hypothesis is that the majority of heritability is attributable to common variants (MAF >0.01) of small effect, such that many variants are not detected at the level of genome-wide significance; most of these common variants are either well tagged by genotyped SNPs through linkage disequilibrium (LD) or can be imputed with reasonably high accuracy from whole-genome sequencing reference panels. If the second hypothesis is true, increasing sample size will be more important than extending variant coverage for continued progress in genetic association studies. The third hypothesis is that heritability estimates from family-based studies are biased upward, as a result, for instance, of shared environmental effects. Therefore, quantifying the relative contributions of rare and common variants to trait variation is critical to inform the design of future experiments and to disentangle the genetic architecture of complex traits and diseases. In this study, we seek to quantify the proportion of variation at common and rare sequence variants that can be captured by SNP array genotyping followed by imputation, and we subsequently estimate the proportion of phenotypic variance for the model complex traits height and BMI that can be explained by all imputed variants.

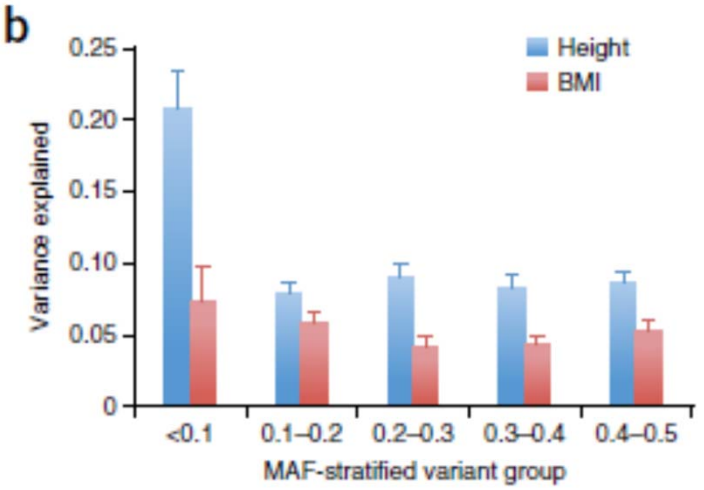
A full list of affiliations appears at the end of the paper.

Received 10 February; accepted 31 July; published online 31 August 2015; doi:10.1038/ng.3390

RESULTS

Unbiased estimate of heritability using whole-genome sequence data

Let h^2_{WGS} denote the narrow-sense heritability (h^2) for a complex trait captured by the sequence variants from whole-genome sequencing and h^2_{1KGP} denote the heritability captured by all variants from



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…