Estimating “Heritability” using Genetic Data

David Evans
University of Queensland
The Majority of Heritability for Most Complex Traits and Diseases is Yet to Be Explained

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.
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!!!)
GCTA

Overview

New version 1.24, more options and much faster!

GCTA Forum http://gcta.freeforums.net

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

- Estimate the genetic relationship from genome-wide SNPs;
- Estimate the inbreeding coefficient from genome-wide SNPs;
- Estimate the variance explained by all the autosomal SNPs;
- Partition the genetic variance onto individual chromosomes;
- Estimate the genetic variance associated with the X-chromosome;
- Test the effect of dosage compensation on genetic variance on the X-chromosome;
- Predict the genome-wide additive genetic effects for individual subjects and for individual SNPs;
<table>
<thead>
<tr>
<th>General</th>
<th>Threads</th>
<th>Posts</th>
<th>Last Post</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Board</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCTA Discussion Board</td>
<td>30</td>
<td>100</td>
<td>Interpretation of GCTA results by kc 2 hours ago</td>
</tr>
</tbody>
</table>

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

Jian Yang¹, Beben Benyamin¹, Brian P McEvoy¹, Scott Gordon¹, Anjali K Henders¹, Dale R Nyholt¹, Pamela A Madden², Andrew C Heath², Nicholas G Martin¹, Grant W Montgomery¹, Michael E Goddard³ & Peter M Visscher¹

ARTICLE

Estimating Missing Heritability for Disease from Genome-wide Association Studies

Sang Hong Lee¹, Naomi R. Wray¹, Michael E. Goddard²,³ and Peter M. Visscher¹,*

REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang¹,* S. Hong Lee¹, Michael E. Goddard²,³ and Peter M. Visscher¹
GCTA - The Mixed Model Framework

\[ y = X\beta + Wu + \varepsilon \]

where:

- \( y \) is a vector of phenotypes
- \( X \) contains covariates
- \( \beta \) contains fixed effects regression coefficients
- \( W \) contains standardized genotype dosages
- \( u \) contains random effects coefficients
- \( k \) is number of SNPs
- \( m \) is number of covariates
- \( n \) is number of individuals
- \( n \) is number of individuals
The Classical Twin Design

\[ r_{g(MZ)} = 1 \]
\[ r_{g(DZ)} = 0.5 \]
\[ r_c = 1 \]

\[
\begin{align*}
P_1 &= aA_1 + cC_1 + eE_1 \\
P_2 &= aA_2 + cC_2 + eE_2
\end{align*}
\]

\[
\begin{align*}
V_{MZ} &= a^2 + c^2 + e^2 & a^2 + c^2 & a^2 + c^2 + e^2 \\
V_{DZ} &= \frac{1}{2}a^2 + c^2 & \frac{1}{2}a^2 + c^2 & a^2 + c^2 + e^2
\end{align*}
\]
Expected Covariance Matrix Twin Pairs
(AE Model)

\[ V = R\sigma^2_A + I\sigma^2_E \]

\[
\begin{pmatrix}
\sigma^2_I & \sigma_{I2} \\
\sigma_{2I} & \sigma^2_2
\end{pmatrix}
= \begin{pmatrix}
l & r \\
r & l
\end{pmatrix} \cdot \begin{pmatrix}
l & 0 \\
0 & l
\end{pmatrix} \cdot \begin{pmatrix}
\sigma^2_A + \sigma^2_E & r \sigma^2_A \\
r\sigma^2_A & \sigma^2_A + \sigma^2_E
\end{pmatrix}
\]

- **V** is the expected phenotypic covariance matrix
- **\sigma^2_A** is the additive genetic variance
- **\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

\[ r_g = A_{ij} \]

\[ P_1 = aA_1 + eE_1 \]
\[ P_2 = aA_2 + eE_2 \]

\[ V = \begin{bmatrix} a^2 + e^2 & A_{ij}a^2 \\ A_{ij}a^2 & a^2 + e^2 \end{bmatrix} \]
Expected Covariance Matrix - Unrelateds

\[ V = A \sigma^2_g + I \sigma^2_e \]

\[
\begin{bmatrix}
\sigma^2_1 & \cdots & \sigma_{1n} \\
\vdots & \ddots & \vdots \\
\sigma_{n1} & \cdots & \sigma^2_n
\end{bmatrix}
= \begin{bmatrix}
a_{11} & \cdots & a_{1n} \\
\vdots & \ddots & \vdots \\
a_{n1} & \cdots & a_{nn}
\end{bmatrix} \cdot \sigma^2_g + \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix} \cdot \sigma^2_e
\]

\( V \) is the expected phenotypic covariance matrix

\( \sigma^2_g \) is the additive genetic variance

\( \sigma^2_e \) 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

• 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

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

Adapted from Yang et al. (2011) Nat Genet

*
GCTA Interpretation

• GCTA does not estimate “heritability”

• GCTA does not estimate the proportion of trait variance due to common SNPs

• GCTA tells you nothing definitive about the number of variants influencing a trait, their size or their frequency
GCTA- Some Assumptions

• The GRM accurately reflects the underlying causal variants

• Underlying variants explain the same amount of variance
  – Relationship between MAF and effect size

• Independent effects
  – Contributions to $h^2$ overestimated by causal variants in regions of high LD and underestimated in regions of low LD
Extending the Model - Genome Partitioning

\[ V = \sum_{c=1}^{22} A_c \sigma^2_{g,c} + I \sigma^2_e \]

- 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

\[ V = A_g \sigma^2_g + A_{ge} \sigma^2_{ge} + I \sigma^2_e \]

- \( A_{ge} = A_g \) for pairs of individuals in the same environment and \( \text{Age} = 0 \) for pairs of individuals in different environments
- “Environmental” factors could be sex or medical treatment for example
Extending the Model - Binary Traits

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

\[ h_1^2 = h_0^2 K (1 - K) / z^2. \]

Figure 1. The Liability Threshold Model for a Disease Prevalence of K
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
Identical by Descent

Identical by state only

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

\[
V = \pi_{IBD}\sigma^2_A + C\sigma^2_C + I\sigma^2_e
\]

\[
\begin{bmatrix}
\sigma^2_1 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\
\sigma_{21} & \sigma^2_2 & \sigma_{23} & \sigma_{24} \\
\sigma_{31} & \sigma_{32} & \sigma^2_3 & \sigma_{34} \\
\sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma^2_4 \\
\end{bmatrix}
= \begin{bmatrix}
1 & \pi_{12} & 0 & 0 \\
\pi_{21} & 1 & 0 & 0 \\
0 & 0 & 1 & \pi_{34} \\
0 & 0 & \pi_{43} & 1 \\
\end{bmatrix} \cdot \begin{bmatrix}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 \\
\end{bmatrix} + \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
\end{bmatrix} \cdot \begin{bmatrix}
\sigma^2_A \\
\sigma^2_C \\
\sigma^2_e \\
\end{bmatrix}
\]
USE IBD variation within SIBS to estimate heritability

- Use variation in genetic sharing within a relative type rather than different types of relatives
- Gets around problem of the “Equal Environment” assumption in twin studies
Extending the Model – IBD

• Estimate GRM
  – Exclude one from each pair of individuals who are >2.5% IBS

• Estimate variance components via “REML”

• Transform from observed scale to liability scale

• Adjust estimates to take account of ascertainment (i.e. the fact that case-control proportions are not the same as in the population)
Application to Height & BMI

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

$^a$Correlation between phenotypes of QISPs after adjustment for fixed effects.
$^b$Heritability estimates.
$^c$Proportion of the phenotypic variance attributed to common environmental variance.
Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits

Noah Zaitlen¹*, Peter Kraft²,³,⁴, Nick Patterson⁴, Bogdan Pasaniuc⁵, Gaurav Bhatia²,³,⁴, Samuela Pollack²,³,⁴, Alkes L. Price²,³,⁴,⁵*

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

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

<table>
<thead>
<tr>
<th>Quantitative trait</th>
<th>N ¹</th>
<th>$h^2_{IBD}$</th>
<th>s.e.</th>
<th>$h^2_{IBS&gt;0.05}$</th>
<th>s.e.</th>
<th>$h^2_{Pub}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body Mass Index (kg/m²)</td>
<td>20000</td>
<td>0.422</td>
<td>0.018</td>
<td>0.433</td>
<td>0.018</td>
<td>0.4-0.6 [6]</td>
</tr>
<tr>
<td>Cholesterol High Density Lipoprotein</td>
<td>19977</td>
<td>0.446</td>
<td>0.017</td>
<td>0.457</td>
<td>0.018</td>
<td>0.5 [6]</td>
</tr>
<tr>
<td>Cholesterol Low Density Lipoprotein</td>
<td>4547</td>
<td>0.196</td>
<td>0.062</td>
<td>0.198</td>
<td>0.063</td>
<td>0.376 [42]</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>20000</td>
<td>0.691</td>
<td>0.016</td>
<td>0.704</td>
<td>0.016</td>
<td>0.8 [6]</td>
</tr>
<tr>
<td>Menarche Age (years)</td>
<td>15150</td>
<td>0.443</td>
<td>0.022</td>
<td>0.454</td>
<td>0.022</td>
<td>0.4-0.7 [43]</td>
</tr>
<tr>
<td>Menopause Age (years)</td>
<td>5540</td>
<td>0.400</td>
<td>0.047</td>
<td>0.409</td>
<td>0.048</td>
<td>0.4-0.6 [44]</td>
</tr>
<tr>
<td>Monocyte White Blood Cell</td>
<td>9651</td>
<td>0.343</td>
<td>0.032</td>
<td>0.351</td>
<td>0.032</td>
<td>0.378 [42]</td>
</tr>
<tr>
<td>Waist-Hip Ratio</td>
<td>5538</td>
<td>0.181</td>
<td>0.037</td>
<td>0.187</td>
<td>0.038</td>
<td>0.3-0.6 [45]</td>
</tr>
<tr>
<td>Sex Ratio of offspring</td>
<td>15000</td>
<td>0.026</td>
<td>0.017</td>
<td>0.021</td>
<td>0.018</td>
<td>-</td>
</tr>
<tr>
<td>Total Children</td>
<td>15000</td>
<td>0.103</td>
<td>0.017</td>
<td>0.111</td>
<td>0.018</td>
<td>-</td>
</tr>
<tr>
<td>Recombination Rate</td>
<td>10259</td>
<td>0.099</td>
<td>0.023</td>
<td>0.110</td>
<td>0.030</td>
<td>-</td>
</tr>
</tbody>
</table>

¹ N is the number of individuals used in the analysis of each phenotype. $h^2_{Pub}$ are previously published estimates of heritability from different populations. doi:10.1371/journal.pgen.1003520.t001
Idea...

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