On “Missing Heritability” in GWAS

Fan Li

BIOSTAT 790

February 18, 2016
Overview

- Heritability analysis
- The problem of “missing heritability” in GWAS
- Two hypotheses
- Evidences to support hypotheses
- Summary
Broad-sense heritability

- Trait of interest is measured on a quantitative scale
  - Height, weight, blood pressure etc.
- Assess the proportion of total phenotypic variation attributable to the genetic component (nature vs nurture)
  - Trait - $Y$; genetic part - $G$ and non-genetic part - $\epsilon$
  - $Y = G + \epsilon$
  - $\text{Var}(Y) = \text{Var}(G) + \text{Var}(\epsilon) + 2 \text{Cov}(G, \epsilon)$
    - assumed to be 0
- Broad-sense heritability $h^2 = \frac{\text{Var}(G)}{\text{Var}(Y)}$
Narrow-sense heritability

- More interested in multiple genes acting jointly
- \( X_m \) - count of disease alleles at the \( m \)th locus \( \in \{0, 1, 2\} \)

\[
Y = \mu + \sum_m \begin{cases} 
  a_m X_m + d_m 1(X_m = 1) & \text{additive} \\
  & \text{codominance}
\end{cases} + \epsilon
\]

- Codominance component - departure from additive mode of inheritance
- \( \text{Var}(G) = V_A + V_D \), partition into additive and dominance genetic variance (Falconer and Mackay, 1996)
- Narrow-sense heritability \( h^2 = V_A / \text{Var}(Y) \)
It can be shown that $V_A$ is a function of the average effect of parents’ genes on the offspring’s phenotype (i.e., breeding values/additive genetic effects).

An advantage of using narrow-sense heritability - can be directly estimated from the phenotypic data on relatives (w/out genotype data).

Given parent-child trio data, $(Y_O, Y_F, Y_M)$
- Linear model $Y_O = \alpha + \beta \left( \frac{Y_F + Y_M}{2} \right) + \epsilon$
- Can show

$$
\beta = \frac{\text{Cov}(Y_O, Y_P)}{\text{Var}(Y_P)} = \frac{V_A}{\text{Var}(Y)}
$$
Narrow-sense heritability - Cont’d

(a) heritability = 0.1

(b) heritability = 0.9

Mid–parental values

Offspring values

Mid–parental values

Offspring values
The (narrow-sense) heritability of height is about 80% (Fisher, 1918; Visscher et al., 2008)

Recent GWASs detected 50 variants that are significantly associated with height in the population, but only account for about 5~10% phenotypic variation (Gudbjartsson et al., 2008; Weeden et al., 2008; Lango Allen et al. 2010)

Where is the “missing heritability” - missing $V_A$?
- Limitation of GWAS?
- Refinement of statistical methodology?
Hypotheses I

- SNPs used in GWAS explain some or all of $V_A$, but most of their individual effects are small (below the stringent genome-wide threshold) and hence not reported
  - Testing association of individual SNP with trait - limit to those with strong associations with trait
  - $V_A$ relies on correctly estimating the individual SNP effects (the classic linear regression framework)

$\Rightarrow$ Refinement of statistical model (pooling effects of all SNPs?)
The actual causal variants are not in complete linkage disequilibrium (LD) with the genotyped SNPs.

- LD: a population concept referring to the associations between alleles at different loci.

- Causal variant $\rightarrow$ Trait $\leftarrow$ marker SNP

$\Rightarrow$ provide evidence for the hypothesis (alternative ways of exploiting GWAS data)
Suppose for now we can genotype $m$ causal variants

- $y_j$ - trait for individual $j$, $g_j$ - total genetic effect, $\epsilon_j$ - normal residual variance

The model

$$y_j = \mu + \sum_{i=1}^{m} z_{ij} u_i + \epsilon_j$$

- $z_{ij}$ - normalized allele count
  $$\in \{(x_{ij} - 2f_i)/\sqrt{f_i(1-f_i)}; x_{ij} = 0, 1, 2\}$$
- $u_i$ - scaled additive genetic effect, assumed random
  $$\sim N(0, \sigma_u^2)$$
- $g_j \sim N(0, \sigma_g^2 = m\sigma_u^2)$ since $\text{Var}(g_j) = \sum_{i=1}^{m} \text{Var}(z_{ij})\text{Var}(u_i)$
In matrix notation, $\mathbf{y} = \mu \mathbf{1} + \mathbf{Zu} + \mathbf{\epsilon}$

$\text{Var}(\mathbf{y}) = (\mathbf{ZZ'}/m) \sigma_g^2 + \mathbf{I} \sigma_\epsilon^2$,

where $\mathbf{G}$ is named the genetic relationship (correlation) matrix.

⇒ the model is parameterized by the narrow-sense heritability

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2}$$

With known $\mathbf{G}$, we can use REML to estimate variance components.
Obtain \( \mathbf{G} \)

- The number and positions of causal variants are unknown
- Obtain the \( \mathbf{G} \) matrix from the genome-wide sample of SNPs \( \mathbf{A} \) instead by \( \mathbf{A} = \mathbf{W} \mathbf{W}' / N \)
  - \( N \) - # of SNPs; \( w_{ij} = (x_{ij} - 2p_i) / \sqrt{2p_i(1 - p_i)} \); \( p_i \) - allele frequency
- Accounting for sampling error associated with each SNP, improve \( \mathbf{A} \) by

\[
A_{jk} = \begin{cases} 
\frac{1}{N} \sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1- p_i)}, & j \neq k, \\
1 + \frac{1}{N} \sum_i \frac{x_{ij}^2 - (1+2p_i)x_{ij} + 2p_i^2}{2p_i(1- p_i)}, & j = k
\end{cases}
\]

- Note: (i) \( p_i \) is estimated from the current population (base/reference population); (ii) \( A_{jj} \) is modified to minimize the sampling variation
Example with human height

Selected 3925 unrelated (3248 adults, 677 16-year-olds) individuals from a number of GWASs

- Leaving out close relatives: avoid the possibility that the resemblance between them could be due to shared environment

They were measured for height and genotyped for 300K to 600K SNPs

\[ \hat{h}^2 = 0.45 \text{ (s.e. } = 0.08) \]

- substantially more than the published \( \approx 10\% \) results
- not focusing on individual SNPs, but on the total variation explained by SNPs (\( \sigma^2_g \))
- conclude that difference is due to many SNPs with small effects (not individually significant in GWAS) – addresses \( H_1 \)
Correcting for incomplete LD

- 45% is still less than the known 80% results, where is the remaining heritability?

- $H_2$ – the ability of SNPs to explain phenotypic variation caused by causal variants depends on the LD between all causal variants and all the SNPs
  - cannot measure LD directly
  - can estimate LD between SNPs
  - lack of LD $\Rightarrow$ larger difference between $G_{jk}$ and $A_{jk}$

- If the causal variants have similar characteristics to the SNPs in terms of allele frequency spectra and LD

- Can mimic LD between causal variants and SNPs using LD between the genotyped SNPs – correcting for incomplete LD and possibly address $H_2$
  - use this as an empirical guide to calibrate $A$ (get closer to $G$)
Calibrating \( A \)

**Steps**

1. Randomly sample \( 2N \) SNPs from all SNPs across the genome and randomly split them into two groups (\( N \) each);
2. Calculate \( A_{jk} \) from group 1;
3. Calculate \( G_{jk} \) using SNPs with MAF \( \leq \theta \) in group 2 \( \Rightarrow \) mimicking the relationship between (proxy) causal variants
4. Regress \( G_{jk} \) on \( A_{jk} \) for \( j \leq k \) (use \( G_{jk} - 1 \) and \( A_{jk} - 1 \) when \( j = k \)), and obtain the slope \( \beta \);
5. Repeat with different values of \( N \)

\( \Rightarrow \) essentially investigating how difference between \( G_{jk} \) and \( A_{jk} \) varies with different choices of \( \theta \) and \( N \)
Randomly sampled five sets of SNPs (from 50K to 250K) in the adult data set and ten sets of SNPs (from 50K to 500K) in the adolescent data set; split each set into two groups

Let $\theta$ range from 0.1 to 0.5

Found the following relationship

$$\beta = 1 - \frac{(c + 1/N)}{\text{Var}(A_{jk})}$$

$c$ depends on $\theta$:

- cause loci have the same spectrum of allele frequency as the genotyped SNPs ($\theta = 0.5$) $\Rightarrow c = 0$;
- cause polymorphisms tend to have lower MAF (e.g., $\theta = 0.1$) $\Rightarrow c = 6.2 \times 10^{-6}$
Prediction error of genetic relationship

On “Missing Heritability” in GWAS
A closer look at $\beta = 1 - (c + 1/N)/\text{Var}(A_{jk})$

- $A_{jk}$ tends to overestimate $G_{jk}$
  - if $\theta = 0.5$, $c = 0$, with $N$ approaching infinity, $\beta = 1$
  - causal variants may have smaller MAF, and $c > 0$

Calibrate $A_{jk}$ with

$$A^*_{jk} = \begin{cases} 
\beta A_{jk}, & j \neq k \\
1 + \beta (A_{jk} - 1), & j = k 
\end{cases}$$

- $\hat{h}^2 = 0.54$ (s.e. = 0.1) if $c = 0$ and $\hat{h}^2 = 0.84$ (s.e. = 0.16) if $c = 6.2 \times 10^{-6}$

- $\theta = 0.1$ is a scenario consistent with the causal variants, on average, being at lower frequency than the SNPs on commercial arrays

- do not prove the causal variants have MAF $\leq 0.1$, but could explain the missing heritability if this is the case
Additional results

- After calibration, the narrow-sense heritability estimates do not depend on the number of SNPs used
  - but standard error will be larger with fewer SNPs

- Used simulation studies to validate the method of estimating heritability using genome-wide SNPs
Simulate a trait based on genotype data of 3925 individuals with around 300K SNPs

(i) randomly sample $m$ causal variants; (ii) randomly sample $m$ causal variants with MAF $\leq 0.1$

Model:

$$y_j = \mu + \sum_{i=1}^{m} z_{ij} u_i + \epsilon_j$$

- causal effect $u_i \sim N(0, 1)$
- generate residual $\epsilon_j \sim N(0, \text{var}(g)(1/h^2 - 1))$
- set $m = 2000, 3000$ and $h^2 = 0.5, 0.8$

run for 30 replicates
## Simulation results

<table>
<thead>
<tr>
<th>MAF ≤ 0.5&lt;sup&gt;e&lt;/sup&gt;</th>
<th>No. causal variants</th>
<th>$h^2$&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Est. $h^2$ (s.e.m.)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Est. $h^2$ (s.e.m.)&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Est. $h^2$ (s.e.m.)&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,000</td>
<td>0.8</td>
<td>0.817 (0.014)</td>
<td>0.678 (0.014)</td>
<td>0.812 (0.014)</td>
<td></td>
</tr>
<tr>
<td>2,000</td>
<td>0.5</td>
<td>0.513 (0.015)</td>
<td>0.428 (0.015)</td>
<td>0.512 (0.015)</td>
<td></td>
</tr>
<tr>
<td>3,000</td>
<td>0.8</td>
<td>0.831 (0.015)</td>
<td>0.693 (0.016)</td>
<td>0.831 (0.016)</td>
<td></td>
</tr>
<tr>
<td>3,000</td>
<td>0.5</td>
<td>0.510 (0.016)</td>
<td>0.424 (0.017)</td>
<td>0.507 (0.017)</td>
<td></td>
</tr>
<tr>
<td>MAF ≤ 0.1</td>
<td>2,000</td>
<td>0.8</td>
<td>0.591 (0.015)</td>
<td>0.433 (0.014)</td>
<td>0.804 (0.026)</td>
</tr>
<tr>
<td>2,000</td>
<td>0.5</td>
<td>0.367 (0.016)</td>
<td>0.271 (0.016)</td>
<td>0.504 (0.030)</td>
<td></td>
</tr>
<tr>
<td>3,000</td>
<td>0.8</td>
<td>0.620 (0.016)</td>
<td>0.462 (0.016)</td>
<td>0.856 (0.029)</td>
<td></td>
</tr>
<tr>
<td>3,000</td>
<td>0.5</td>
<td>0.384 (0.020)</td>
<td>0.287 (0.019)</td>
<td>0.533 (0.036)</td>
<td></td>
</tr>
</tbody>
</table>
Summary

- Statistical approaches to address the two proposed hypotheses to explain “missing heritability” in GWAS
  - Refinement of statistical model
    - a model fitting random effects of individuals and using all SNPs to estimate relationship between individuals
    - contrast to standard GWAS models which select SNPs based upon test statistics for association between height and SNPs
  - correcting for imperfect LD
    - essentially adjusting for $A \Leftrightarrow$ calibrating the prediction for random effects
    - the interplay between prediction of random effects and the estimation of variance components due to these latent random effects