Genome-wide estimates of heritability

Ben Domingue
Institute of Behavioral Science
University of Colorado Boulder
ben.domingue@gmail.com
- Genes → behaviors & outcomes of interest.
- Genome-wide data: FHS, HRS, AddHealth, etc....
- Hard to get a handle on genotype/phenotype connection.
  - GWAS results help, but have limited availability.
  - Even when available, polygenic scores have limited predictive value.

What else can we do?
GCTA

Genome-wide Complex Trait Analysis (GCTA) tells us about heritability.

- GCTA estimates heritability without knowledge of causal variants.
- Instead uses “genetic similarity” (similar to logic of twin studies).
Method

1. Estimate genome-wide similarity:

\[ A_{jk} = \frac{1}{N} \sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)} \]
Method

1. Estimate genome-wide similarity:

\[ A_{jk} = \frac{1}{N} \sum_{i} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)} \]

2. Then estimate mixed model:

\[ y = X\beta + g + \epsilon \]

where \( g \sim \text{MVN}[0, \sigma_g^2 A] \).
Method

1. Estimate genome-wide similarity:

\[ A_{jk} = \frac{1}{N} \sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)} \]

2. Then estimate mixed model:

\[ y = X\beta + g + \epsilon \]

where \( g \sim \text{MVN}[0, \sigma_g^2 A] \).

3. Heritability:\n\[ \frac{\hat{\sigma}_g^2}{\sigma_g^2 + \sigma_\epsilon^2} \].
Method

1. Estimate genome-wide similarity:

\[ A_{jk} = \frac{1}{N} \sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)} \]

2. Then estimate mixed model:

\[ y = X\beta + g + \epsilon \]

where \( g \sim \text{MVN}[0, \sigma^2_g A] \).

3. Heritability:

\[ \hat{\sigma}^2_g = \frac{\hat{\sigma}^2_g}{\hat{\sigma}^2_g + \hat{\sigma}^2_\epsilon} \]

Complicated model & not the DGP.
Sensitivity to genetic architecture?

- Robust to # of causal variants.

Speed et al., 2012, AJHG,
Sensitivity to genetic architecture?

▶ Robust to # of causal variants.
▶ Sensitive to LD.

Speed et al., 2012, AJHG,
Sensitivity to environment?

Could genetic similarity just be a proxy for environmental similarity?

Table 1  GREML heritability estimates for shared childhood urbanicity, height, BMI and education

|                    | $h^2$ No controls | $h^2$ Urban control | $|\Delta| A - B$ | $h^2$ No controls | $h^2$ Urban control | $|\Delta| C - D$ |
|--------------------|-------------------|---------------------|---------------|-------------------|---------------------|---------------|
| Urban childhood $N=6439$ | 0.29155 (0.0574)  | NA                  | NA            | 0.14767 (0.0622)  | NA                  | NA            |
| Height $N=6379$        | 0.32489 (0.0644)  | 0.32510 (0.0644)    | 0.00022 (0.0910) | 0.30397 (0.0659)  | 0.30397 (0.0659)    | 0.02092 (0.0921) |
| BMI $N=6320$           | 0.31300 (0.0674)  | 0.31323 (0.0675)    | 0.00023 (0.0953) | 0.31300 (0.0674)  | 0.3190 (0.0678)     | 0.00596 (0.0956)  |
| Education $N=6414$     | 0.17493 (0.0650)  | 0.15217 (0.0652)    | 0.02276 (0.0921) | 0.1749 (0.0650)   | 0.15939 (0.0656)    | 0.01554 (0.0923)  |

Conley et al., 2014, JHG
My goal: Offer intuition and basic guidance on when GCTA estimates may be reliable.
Data

HRS: 4950 non-Hispanic whites, \(\approx 1.5\text{M}\) autosomal SNPs.

- Height: 0.40
Q1: Gen sim as function of SNPs

Correlation
Q1: Gen sim as function of SNPs

<table>
<thead>
<tr>
<th>Sample</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% Sample</td>
<td>0.98</td>
</tr>
<tr>
<td>30% Sample</td>
<td>0.95</td>
</tr>
<tr>
<td>10% Sample</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Q1: Gen sim as function of SNPs

<table>
<thead>
<tr>
<th>Sample</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% Sample</td>
<td>0.98</td>
</tr>
<tr>
<td>30% Sample</td>
<td>0.95</td>
</tr>
<tr>
<td>10% Sample</td>
<td>0.83</td>
</tr>
</tbody>
</table>

\[ r^2 = 0.01 \] \quad \text{0.57} \\
\[ r^2 = 0.2 \] \quad \text{0.75} \\
\[ r^2 = 0.5 \] \quad \text{0.88}
Q2: GWAS (height) variants
Q2: GWAS (height) variants
Q2: GWAS (height) variants
Q2: GWAS (height) variants
Q3: Heteroskedasticity

Heteroskedasticity is common problem.

- weight on height.
- own education on paternal education.

Of concern here since we’re estimating variance components.
Q3: Heteroskedasticity

Heteroskedasticity is common problem.

- weight on height.
- own education on paternal education.

Of concern here since we’re estimating variance components.

- Simulate outcome based on GCTA model.
- $y = 0.5 \cdot \text{height} + g + \epsilon$.
- $\epsilon_i$ has variance $\exp(\alpha \cdot \text{height} \cdot \sigma^2_\epsilon)$, where $\alpha$ controls level of heteroskedasticity and $\sigma^2_\epsilon$ controls heritability.
Q3: Heteroskedasticity

Heteroskedasticity is a common problem.

- weight on height.
- own education on paternal education.

Of concern here since we’re estimating variance components.

- Simulate outcome based on GCTA model.
- \( y = 0.5 \cdot \text{height} + g + \epsilon. \)
- \( \epsilon_i \) has variance \( \exp(\alpha \cdot \text{height} \cdot \sigma^2_\epsilon) \), where \( \alpha \) controls level of heteroskedasticity and \( \sigma^2_\epsilon \) controls heritability.

Examine recovery of heritability, but def’n no longer simple.
Q3: Heteroskedasticity
Q3: Heteroskedasticity
Heritability not constant: What are implications for GCTA?

- Standard GCTA: $g \sim \text{MVN}[0, \sigma_g^2 A]$. 
Q4: Environmental Moderation

Heritability not constant: What are implications for GCTA?

- Standard GCTA: $g \sim \text{MVN}[0, \sigma_g^2 A]$.
- We simulate data using $g \sim \text{MVN}[0, A']$ where the $(i, j)$-th entry of $A'$ is $h_i h_j A_{ij}$. 
Q4: Environmental Moderation

What if we ignore environment?
Q4: Environmental Moderation

What if we allow for environmental variation?
- LD is important consideration (aside: I’m skeptical about using KING or REAP estimates).
- Heteroskedasticity leads to inflation of $h^2$ estimates.
- Environmental differences are likely to be problematic (and yet may be rampant?).
LD is important consideration (aside: I’m skeptical about using KING or REAP estimates).

Heteroskedasticity leads to inflation of $h^2$ estimates.

Environmental differences are likely to be problematic (and yet may be rampant?).

In closing: GCTA is like a table saw.