Complex traits – heritability and missing heritability

Shamil Sunyaev

Division of Genetics
Department of Medicine
Brigham and Women's Hospital / Harvard Medical School
Broad Institute of M.I.T. and Harvard

Simple and Complex Phenotypes

Simple phenotype

Complex phenotype

Complex traits are heritable but not in Mendelian fashion

Table/Ouetch, Asking About Life, 29, Figure 10.6

Quantitative Trait Loci (QTLs)

Inheritance at each locus is Mendelian. Loci are independent

Phenotype is additive over locus effects -> normal distribution
Dichotomous complex traits such as disease

Population variation is fully described by variance

\[ V = V_G + V_E \]

Genetic contribution

Everything else

Components of genetic variance

\[ V_G = V_A + V_D + V_I + V_M \]

Main (additive) effects

Genetic interactions

Dominant effects

New mutations

Variance decomposition

Additive variance

Additive variance \( V_A \) is variance explained by the model

\[
Y_j = \alpha + \sum_i X_{ij} + \epsilon
\]

\[
V_A = 2 \sum_i^2 x_i (1 - x_i)
\]
Heritability

Broad sense

\[ H^2 = \frac{V_G}{V} \]

Narrow sense

\[ h^2 = \frac{V_A}{V} \]

Estimating heritability

\[ \text{Cov}(MP, O) = \frac{1}{2} V_A + \frac{1}{4} V_I \]

Narrow sense heritability

\[ h^2 = \frac{V_A}{V} \left( \frac{\text{Cov}(MP, O)}{V(MP)} \right) \]

Gene mapping by linkage

How to find genes behind the phenotype?

Linkage studies

Mapping by linkage
Gene mapping by association

Association studies

Array based genotyping can determine case/control allele frequencies

In a typical GWAS, we test ~2 million common SNPs (HapMap)

courtesy: PIW de Bakker

With genotypic information in hand

Regress phenotype on genotype

\[ Y = \sum_i X_i + \epsilon \]

Additive variance

\[ V_A = 2 \sum_i x_i (1 - x_i) \]

Narrow sense heritability

\[ h^2 = \frac{V_A}{V} \]

In the Ideal World

Regress phenotype on genotype

\[ Y = \sum_i X_i + \epsilon \]

Identify significant and reproducible associations. Estimate effect sizes. Estimate additive variance.

\[ V_A = \sum_i x_i (1 - x_i) \]

Reality: missing heritability

\[ h^2 = \frac{V_A}{V} \ll \frac{Cov(MP,O)}{V(MP)} \]
Current GWAS explain a minor fraction of heritability

The case of the missing heritability

Height – 10%, Blood lipids – 12%

Likely reasons for missing heritability

1. Common variants of weak effect
2. Rare variants of larger effect
3. Epistatic interactions

\[ \text{Cov}(MP,O) = \frac{1}{2} V + \frac{1}{4} V_i \]

Questions about allelic architecture

- How many loci are involved?
- Is variation underlying the trait rare or common?
- What is the distribution of effect sizes of variants involved in the trait?
- What is the role of epistasis and dominance?

GxG interactions

Why is epistatic variance commonly disregarded?

- In human genetics, epistatic interactions between common variants have not been observed.
- In a model with two (or several) loci, contribution of epistatic variance is relatively small.
- Long term response to selection in model organisms seems to contradict the importance of epistasis.

Any evidence for or against epistasis?
Why might epistatic variance be of importance?

- A non-linear model involving many loci would generate a large epistatic variance.
- Interactions would be statistically undetectable.
- The model would not generate significant deviations from the observations.
- As an example, we may consider a model with multiple pathways involved.

Many variants of small effect

Evidence in favor of the highly polygenic model

![Graph showing mean height over generations](image)

Evidence in favor of the highly polygenic model

![Graph showing body weight over generations](image)

Evidence in favor of the highly polygenic model

![Graph showing SNP distribution](image)

Evidence in favor of the highly polygenic model

![Graph showing mean height over generations](image)

ANALYSIS

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

Jian Yang1, Babak Barzegar1, Brian P Molloy1, Scott Gordon1, Anjali K Harit1, Dale R Nyholt1, Pamela A Madden2, Andrew C Heath2, Nicholas G Martin3, Grant W Montgomery4, Michael E Goddard5 & Peter M Visscher1
A simple genetic predictor

We can construct a simple additive estimate of the phenotype using a multitude of markers

\[ Y_j = \sum_{i} \hat{\theta_i} X_{ij} \]

A multitude of small effects

Distribution of apparent effect sizes

Model selection

Evidence in favor of the highly polygenic model

Enrichment of GWAS signals in putative regulatory elements

Epigenome Roadmap

- Epigenetic mechanisms
- Health endpoints
- Chromatin
- Methylation
- DNA

- Histone modification
- Chromatin remodeling through transcription factors
- DNA accessibility, gene active
- DNA methylation
- Epigenetic regulation
- Genetic variation
- Environmental exposures
- Genetic risk

- Inflammatory disease
- Autoimmune disease
- Neurodegenerative disease
- Breast cancer
- Intronic (n=10)
- Exonic (n=30)
- Ctrl (n=50)
- Other (n=100)

- Multiple sclerosis
- Genetic risk
- GWAS P-value threshold
- GWAS Q-value threshold
Enrichment of GWAS signals in putative regulatory elements

Mendelian genetics vs. complex trait genetics

Example: HDL-Cholesterol

Stabilizing selection is the most common type of selection on a quantitative trait

Rare variants

Stabilizing selection

Selection may be related or unrelated to the trait
Technically, non-neutral genetic variation should not exist!

Forces to maintain variation:

Selection

Mutation

Common diseases are due to multiple deleterious alleles in mutation-selection balance

- Weak selection
- High mutation rate

CURRENT ESTIMATE:

~100 new mutations per genome
~1-2 new amino acid changes per genome

Combine all non-synonymous variants in a single test

Theory:
1) Most new missense mutations are functional (mutagenesis, population genetics, comparative genomics)
2) Most new missense mutations are only weakly deleterious (population genetics)
3) Most functional missense mutations are likely to influence phenotype in the same direction (mutagenesis, medical genetics)

Data: multiple candidate gene studies
HDL-C, LDL-C, Triglycerides, BMI, Blood pressure, Colorectal adenomas

Study design

This is a direct association!

Disease          Control

This is a direct association!

Disease          Control

Functional variants
Neutral variants

Simulations suggest that

- Sequencing studies would be able to identify many genes involved in biology of the trait under study.
- Studies of large populations (many thousands of individuals) are required to achieve high statistical power.

Mendelian genetics vs. complex trait genetics

Kryukov et al., PNAS 2009, Kiezun et al., Nature Genetics
**Example: HDL-Cholesterol**

Adopted from Brewer et al., 2003

---

**Variance explained by rare variants**

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Trait</th>
<th>Candidate Genes</th>
<th>Prop. functional</th>
<th>Class</th>
<th>Enrichment (fold)</th>
<th>Variance Explained (Per Gene)</th>
</tr>
</thead>
</table>
| Romero 2009   | TG    | [ANGPTL
            |       | 5/4/5]          | 0%    | 3.6%             | 1.96                           | 0.8%                           |
|               |       |                 |                   | 0%    |                   | 0.5%                           |
|               |       |                 |                   | 0%    |                   | 0%                             |
|               |       |                 |                   | 100%  |                   | 0.1%                           |
| Ji 2008       | BP    | [SCD,AIP2]      | 0%                | 5.5%  | 3.19              | 0.7%                           |
|               |       |                 |                   | 0%    |                   | 0.5%                           |
|               |       |                 |                   | 0%    |                   | 0.5%                           |
|               |       |                 |                   | 100%  |                   | 0.5%                           |
| Johansen 2010 | TG    | [APOA5]         | 0%                | 8.1%  | 2.17              | 1.8%                           |
|               |       |                 |                   | 0%    |                   | 1.4%                           |
|               |       |                 |                   | 0%    |                   | 1%                             |
|               |       |                 |                   | 0%    |                   | 0.8%                           |

---

**Exome Sequencing Project**

Program Officer: Deborah Applebaum-Bowden

<table>
<thead>
<tr>
<th>Heart GO</th>
<th>Lung GO</th>
<th>Women’s Health Initiative Sequencing Program</th>
</tr>
</thead>
</table>

Heart GO
- PI: Stephen Rich
- University of Virginia

Lung GO
- PI: Michael Bamshad
- U. of Washington
- Kathleen Barnes, Johns Hopkins University

Women’s Health Initiative Sequencing Program
- PI: Rebecca Jackson
- Ohio State University

Pls: Debbie Nickerson, Mark Rieder, Jay Shendure, & Phil Green

Pls: Stacey Gabriel & David Altshuler

---

**Study design**

1000 Early onset MI cases: Men < 50
Women < 60

Most likely to be genetically influenced

1000 Controls without MI:
Men > 60
Women > 70

---

**Results: Aggregate of very rare variants**

Burden of risk alleles: 1% threshold

---

**Sequencing results for APOA5**

<table>
<thead>
<tr>
<th>Study</th>
<th>N cases</th>
<th>N controls</th>
<th>T1 cases</th>
<th>T1 controls</th>
<th>Freq cases</th>
<th>Freq controls</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>EOMI interim</td>
<td>466</td>
<td>436</td>
<td>13</td>
<td>1</td>
<td>2.8%</td>
<td>0.2%</td>
<td>0.0007</td>
</tr>
</tbody>
</table>
An APOA5 mutation confers a 2.2 fold increased risk for MI.

### Sequencing results for APOA5

<table>
<thead>
<tr>
<th>Study</th>
<th>N</th>
<th>% cases</th>
<th>% controls</th>
<th>OR</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESP EOMI 1</td>
<td>902</td>
<td>2.79</td>
<td>0.23</td>
<td>12.5</td>
<td>0.0007</td>
</tr>
<tr>
<td>VHS</td>
<td>1,327</td>
<td>1.23</td>
<td>0.57</td>
<td>2.2</td>
<td>0.18</td>
</tr>
<tr>
<td>OHS</td>
<td>1,138</td>
<td>0.91</td>
<td>0.17</td>
<td>5.4</td>
<td>0.05</td>
</tr>
<tr>
<td>ATVB</td>
<td>2,122</td>
<td>1.4</td>
<td>0.76</td>
<td>1.9</td>
<td>0.08</td>
</tr>
<tr>
<td>ESP EOMI 2</td>
<td>1,235</td>
<td>2.07</td>
<td>1.39</td>
<td>1.5</td>
<td>0.23</td>
</tr>
<tr>
<td>CCHS</td>
<td>2,830</td>
<td>1.04</td>
<td>0.51</td>
<td>2.1</td>
<td>0.11</td>
</tr>
<tr>
<td>PROCARDIS</td>
<td>2,884</td>
<td>1.16</td>
<td>0.6</td>
<td>1.9</td>
<td>0.06</td>
</tr>
<tr>
<td>Meta-analysis</td>
<td>12,319</td>
<td>1.38</td>
<td>0.62</td>
<td>2.2</td>
<td>1.2e-6</td>
</tr>
</tbody>
</table>

An APOA5 mutation confers a 2.2 fold increased risk for MI.

What defines allelic architecture?

Under the assumption of additivity, the allelic architecture is fully defined by the mutation target $\theta$ and by the joint distribution of effect sizes and selection coefficients $g(\beta, S)$.

**Variance explained**

Effective population size: N=10,000
**Observations:**
- Mean = 0
- Variance = 1
- Skew = 0.24
- Ex. Kurtosis = 0.034
- Intermediate heritability: $h^2 \approx 0.5$

**The simplest model: two effect sizes**

Using Asymptotics for strong selection (large $-S$). Assuming intermediate heritability, $h^2 = 0.5$.

Suppose measured maximum

$$\text{Max } \beta = \frac{\kappa}{h^2}$$

$\theta$ corresponds to 20,000 genes

**Constraining allelic architecture based on results of genomic studies**

**Phenotype of an individual**

$$Y_i = \sum_j \beta_j X_{ij} + \epsilon$$

**Two effect sizes model**

Using Asymptotics for strong selection (large $-S$). Assuming intermediate heritability, $h^2 = 0.5$.

Suppose measured maximum

$$\text{Max } \beta = \frac{\kappa}{h^2}$$

$\theta$ corresponds to 20,000 genes

**Population genetic model**

- Demographic history
- Mutation rate
- Recombination rate
- Distribution of selection coefficients

**Model parameters**

- Mutational target size ($\#$ of base pairs, 7)
- Degree of coupling ($\gamma$) between selection coefficients and phenotypic effects

**Empirically observed values**

- Affected sib-pair linkage LOD scores
- Discovery GWAS p-value distribution
- Number of GWAS loci replicated
- Polygenic score $R^2$

- Disease prevalence
- Disease heritability
- Sibling relative risk

- Exome and genome sequencing and genotyping studies
Extreme models are excluded. A broad range of models is still consistent with the data. Rare variants may explain less than 25% of heritability and more than 80% of heritability.