Missing Heritability

02-223 Personalized Medicine:
Understanding Your Own Genome
Fall 2014
Heritability

• **Nature vs Nurture**
  – Whether observed variation in a particular trait is due to environmental or to biological factors
  – The resemblance between parents and their offspring
Heritability

- Heritability can be estimated from the regression of offspring phenotypic values on the average of parental phenotypic values without using genome information.
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.
Missing Heritability

• Height is 80-90% heritable

• Using univariate regression analysis
  – In GWAS over 30,000 people, more than 40 loci were found to be associated with the height
  – Only 5% of the height’s heritability is explained by the genetic variation at those 40 loci

• Analyzing all SNPs jointly for their influence on height (multivariate regression analysis)
  – 45% of the height’s heritability is explained by SNPs

• How can we explain the missing heritability?
Missing Heritability

• Studies looking at similarities between identical and fraternal twins estimate heritability at more than 90% for autism and more than 80% for schizophrenia

• Often, the genetic loci found by GWAS explain a small fraction of the heritability

<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of loci</th>
<th>Proportion of heritability explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age-related macular degeneration</td>
<td>5</td>
<td>50%</td>
</tr>
<tr>
<td>Crohn’s disease</td>
<td>32</td>
<td>20%</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>6</td>
<td>15%</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>18</td>
<td>6%</td>
</tr>
<tr>
<td>HDL cholesterol</td>
<td>7</td>
<td>5.2%</td>
</tr>
<tr>
<td>Height</td>
<td>40</td>
<td>5%</td>
</tr>
<tr>
<td>Early onset myocardial infarction</td>
<td>9</td>
<td>2.8%</td>
</tr>
<tr>
<td>Fasting glucose</td>
<td>4</td>
<td>1.5%</td>
</tr>
</tbody>
</table>
Explaining Missing Heritability

• Undetected epistasis – Interactions between multiple genetic loci

• Pleiotropy – multiple phenotypes affected by the same genetic loci

• From common variants to rare variants
Epistasis

• Epistasis: The effect of one locus depends on the genotype of another locus
  – Epistatic effects of genetic loci can be detected only if we consider the multiple loci jointly

• In contrast, marginal effects of a locus refers to the genetic effect of the locus that is independent of other loci
  – Most studies assume the phenotype can be predicted as a sum of single-locus effects
Epistasis for Mendelian Traits

Dominant epistasis (Mendelian)

<table>
<thead>
<tr>
<th>Dominant white genotype (K/I)</th>
<th>EE</th>
<th>Ee</th>
<th>ee</th>
</tr>
</thead>
<tbody>
<tr>
<td>ii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extension genotype (MC1R)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Carlberg & Haley, Nature Reviews Genetics 2004
When There is No Epistasis

- Two additive (non-epistatic) loci
- The three lines run in parallel
Epistasis Example

- Dominant epistasis
- One locus in a dominant way suppresses the allelic effects of a second locus
Challenges for Detecting Epistatic Effects

• Many studies ignore epistasis among multiple genetic loci mainly due to the high computational cost for detecting it, but epistasis is believed to be prevalent and thus important.

• A typical study for identifying genetic effects on phenotypes involves about a million SNPs
  – The number of candidate epistatic interactions between two SNPs?
  – Multiple testing problems

• Approximation: Screen for SNPs with marginal effects and consider interactions only among those SNPs
Detecting Epistatic Effects

• Multivariate regression for detecting marginal effects

\[ y = \sum_{j=1}^{J} x^j \beta^j + \beta_0 + \varepsilon \]

• Multivariate regression for detecting epistatic effects

\[ y = \sum_{j=1}^{J} x^j \beta^j + \sum_{(k,m) \in S_m} x^k x^m \beta^{km} + \beta_0 + \varepsilon \]
Explaining Missing Heritability

• Undetected epistasis – Interactions between multiple genetic loci

• **Pleiotropy** – multiple phenotypes affected by the same genetic loci

• From common variants to rare variants
Pleiotropy

• **Pleiotropic effects**
  – Definition: A genetic variation influences multiple phenotypes at the same time
  – Instead of assessing the effect of genotype on a single phenotype, we can consider multiple related phenotypes (e.g., genes in the same pathway) to detect pleiotropic effects

• **Examples of pleiotropy**
  – A mutation that influences your cholesterol level can also influence your body-mass index and blood pressure
  – A mutation that influences your blood IgE level can also influence your allergy
Detecting SNPs with Pleiotropic Effects on Phenotypes

SNPs influencing single phenotype

ACGTTTTACTGTACAATT

causal SNP

a univariate phenotype

SNPs influencing multiple phenotypes

ACGTTTTACTGTACAATT

Multivariate complex syndrome (e.g., asthma)
age at onset, history of eczema
Asthma Study

- 543 severe asthma patients from the Severe Asthma Research Program (SARP)

- Genotypes: 34 SNPs in *IL4R* gene
  - 40kb region of chromosome 16
  - Impute missing genotypes with *PHASE* (Li and Stephens, 2003)

- Traits: 53 asthma-related clinical traits
  - Quality of Life: emotion, environment, activity, symptom
  - Family history: number of siblings with allergy, does the father has asthma?
  - Asthma symptoms: chest tightness, wheeziness
Asthma and *IL4R* Gene

Allergen (ragweed, grass, etc.)

- Interleukin-4 Receptor regulates IgE antibody production

In asthma patients

- T cell
- Th2 cell
- B cell

Immunoglobulin E (IgE) antibody

Inflammation!

- Airway in lung narrows
- Difficult to breathe
- Asthma symptoms
**IL4R Gene**

Chr16: 27,325,251-27,376,098

- SNPs in intron
- SNPs in exon
- SNPs in promoter region
Genetic Association for Asthma Clinical Traits

TCGACGTTTACTGTACAATT

Subnetworks for lung physiology

Subnetwork for quality of life
SNPs with Pleiotropic Effects

- **Lung physiology-related traits I**
  - Baseline FEV1 predicted value: MPVLung
  - Pre FEF 25-75 predicted value
  - Average nitric oxide value: online
  - Body Mass Index
  - Postbronchodilation FEV1, liters: Spirometry
  - Baseline FEV1 % predicted: Spirometry
  - Baseline predrug FEV1, % predicted
  - Baseline predrug FEV1, % predicted

- **Q551R SNP**
  - Codes for amino-acid changes in the intracellular signaling portion of the receptor
  - Exon 11

Are they SNPs that are simply in LD?

- **-log(p-values)** from univariate regression analysis
Linkage Disequilibrium Structure in *IL-4R* gene

- SNP rs3024622
- SNP rs3024660
- SNP Q551R

\[ r^2 = 0.07 \]

\[ r^2 = 0.64 \]
IL4R Gene

Chr16: 27,325,251-27,376,098

SNPs in intron

SNPs in exon

SNPs in promotor region

Q551R
rs3024622
rs3024660
Explaining Missing Heritability

- Undetected epistasis – Interactions between multiple genetic loci

- Pleiotropy – multiple phenotypes affected by the same genetic loci

- From common variants to rare variants
Common Variants vs. Rare Variants

- First-generation genome-wide association study (GWAS): common variant common disease hypothesis
  - Common low-penetrance variants
    (Penetrance: the likelihood that the causal loci will in fact have an effect on the phenotype)

- Common variants with minor allele frequency (MAF)>5%
  - dbGap: ~11 million SNPs
  - HapMap: 3.5 million SNPs
  - A successful GWAS requires a more complete catalogue of genetic variations

- Rare variants (MAF<0.5%), low-frequency variants (MAF:0.5%~5%)
  - Captured by sequencing with next-generation sequencing technology
  - Rare, moderately penetrant alleles
  - Possibly significant contributors to the genetic architecture of disease
    - Causal variants are subject to negative selection
1000 Genome Project
(The 1000 Genome Project Consortium, Nature 2010)

The **goal** is to characterize over **95% of variants** that are in genomic regions accessible to current high-throughput sequencing technologies and that have **allele frequency of 1% or higher** (the classical definition of polymorphism) in each of **five major population groups** (populations in or with ancestry from Europe, East Asia, South Asia, West Africa and the Americas)

- 179 individuals (low coverage)
- 6 individuals in two trios (deep sequencing)
- 697 individuals (exon sequencing of 8,140 exons)
1000 Genome Projects: Known vs. Novel Variants
## Associations to Rare Variants

- Often GWA studies are underpowered for functional rare variants

<table>
<thead>
<tr>
<th>Common Variant Association</th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele a</td>
<td>60</td>
<td>20</td>
</tr>
<tr>
<td>Allele A</td>
<td>40</td>
<td>80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rare Variant Association</th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele a</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Allele A</td>
<td>93</td>
<td>98</td>
</tr>
</tbody>
</table>

- Common variant GWA approaches are appropriate only for common variants
Recent Methods for Detecting Rare Variant Associations

• Test **combined effect** of multiple rare variants
• **Fixed-Threshold Approach** (Li & Leal, AJHG 2008)
  – Include only the SNPs with allele frequency below a fixed threshold
  – For SNP $i=1,...,m$ in a genomic region or a gene,

\[
Score = \sum_{i=1}^{m} \mathcal{E}_i C_i
\]

  » $C_i$: The allele frequency of SNP $i$ in cases

  » $\mathcal{E}_i = \begin{cases} 
1 & \text{If the allele frequency of SNP } i \text{ is below a specified threshold} \\
0 & \text{Otherwise}
\end{cases}$

  – Evaluate the significance of $Score$
Feasibility of Identifying Disease Loci

- Rare alleles causing Mendelian disease
- Few examples of high-effect common variants influencing common disease
- Low-frequency variants with intermediate effect
- Rare variants of small effect very hard to identify by genetic means
- Common variants implicated in common disease by GWA

Effect size:
- High
- Intermediate
- Modest
- Low

Allele frequency:
- Very rare
- Rare
- Low frequency
- Common
Explaining Missing Heritability

• Going beyond SNPs
  – Structural variants: insertions, deletions, duplication, copy number variants

• Heritability estimate could be wrong. Environment (e.g., diet) can be heritable within families

• Epigenetics changes in gene expression that are inherited but not caused by changes in genetic sequence
  – Feeding a mouse a certain diet, for example, can alter the coat color not only in its children, but also in its children’s children. Here, the expression of a coat-color gene is controlled by a type of DNA modification (epigenetic modification) called methylation.
Tag SNPs and GWAS: Using Reference Datasets for Genotype Imputation

- Reference data: dense SNP data from HapMap III, or 1000 genome project
- New data: SNP data for individuals in a given study
- Data after imputation
Genotype Imputation

Reference set of haplotypes, for example, HapMap

Each sample is phased and the haplotypes are modeled as a mosaic of those in the haplotype reference panel

The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

Genotype data with missing data at untyped SNPs (grey question marks)

PHASE can be used for imputation!
Tag SNPs and GWAS
(Servin & Stephens, 2007)

- △ Tag SNP
- ○ Non-tag SNP
Summary

• Missing heritability and how to explain it
  – Epistasis, pleiotropy, rare variant effects, etc.

• 1000 Genome Project for obtaining a more complete catalogue of genetic variations in human genomes