Evolution, maintenance and allelic architecture of complex traits

Shamil Sunyaev

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

Why are we doing genetics?

Simple and Complex Phenotypes

Complex traits are heritable but not in Mendelian fashion

Complex traits are heritable but not in Mendelian fashion

Complex traits are heritable but not in Mendelian fashion
Quantitative Trait Loci (QTLs)

Inheritance at each locus is Mendelian. Loci are independent.

Phenotype is additive over locus effects -> normal distribution

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

Dichotomous complex traits such as disease

Liability distribution

Liability threshold

Variance decomposition

Variance around the mean

Variance of means

Regression

\[ Y \]

AA  Aa  aa
Additive variance

Additive variance $V_A$ is variance explained by the model

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

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

Heritability

Broad sense

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

Narrow sense

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

Estimating heritability

Narrow sense heritability

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

With genotypic information in hand

Regress phenotype on genotype

\[ Y_j = \sum_i \beta_i X_{ij} + \epsilon \quad V_A = 2 \sum_i \beta_i^2 x_i (1 - x_i) \]

Narrow sense heritability

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

In the Ideal World

Regress phenotype on genotype

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

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

\[ \hat{V}_A = 2 \sum_i \hat{\beta}_i^2 x_i (1 - x_i) \]

Reality: missing heritability

\[ \hat{h}^2 = \frac{\hat{V}_A}{V} \ll \frac{\text{Cov}(MP,O)}{V(MP)} \]

Current GWAS explain a minor fraction of 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

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

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 is epistatic variance might 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.
Evidence in favor of the highly polygenic model

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

- We can model effects of individual variants as random effects distributed as $N(0, \sigma^2)$.
- Random effect model is a model with error terms drawn from a multivariate normal distribution.
- In the infinitesimal model, co-variance matrix can be approximated using IBS (not IBD).

\[
y_i = \sum_j \beta_j X_{ij} + \epsilon
\]

$X_{ij}$—Normalized genotype of individual $i$ at SNP $j$

In the matrix form:

\[
\bar{y} = X\hat{\beta} + \epsilon
\]

Two important matrices:

\[
LD = \frac{1}{M} X^T X
\]

\[
GRM = \frac{1}{N} X X^T
\]
Linear Mixed Model (LMM)

For each SNP we can fit the model

\[ Y_i = \beta_j X_{ij} + \varepsilon \]

\[ \varepsilon \sim N(0, \sigma^2) \]

\[ u \sim MVN(0, GRM) \]

Polygenic scores

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

\[ \hat{Y}_i = \sum \hat{\beta}_j X_{ij} \]

A multitude of small effects

Distribution of apparent effect sizes

Estimates of heritability from exome sequencing studies

Back-calculating heritability from prediction in independent samples

\[ r^2(\hat{Y}, Y) \approx \frac{h^2}{h^2 + M/N} \]

Evidence in favor of the highly polygenic model

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

Stabilizing selection is the most common type of selection on a quantitative trait. Selection may be related or unrelated to the trait.

**Technically, non-neutral genetic variation should not exist!**

Forces to maintain variation:

*Selection*

*Mutation*

**Why does a common genetic disease exist?**

*From evolutionary perspective common genetic disease should not exist: natural selection should remove disease-causing alleles from the population*

**Theory 1:** MEDICALLY detrimental polymorphisms are not EVOLUTIONARY deleterious

- Disease late onset (after the reproductive age)
- Changed environment and lifestyle (selection direction reversal)
- Compensatory positive effect

Balancing selection
Frequency dependent selection
Antagonistic pleiotropy (Trade Off)

**Examples:** APOE (Alzheimer’s disease), AGT (Hypertension), CYP3A (Hypertension)

**Mutation/selection balance**

**Theory 2:** 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*

**Examples:** LDL-C, HDL-C, Triglyceride, Blood pressure, Colorectal adenomas

---

![Graph showing cumulative number of samples](image)

**Kiezun, Garimella, Do, Stitziel, et al. Nature Genetics 2012**

---

![Graph showing cumulative number of samples](image)

**Kiezun, Garimella, Do, Stitziel, et al. Nature Genetics 2012**
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**

- Lowest values
- Highest values
- Quantitative trait

---

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.

Kryukov et al., PNAS 2009; Kiezun et al., Nature Genetics 2012

Example: HDL-Cholesterol

Adopted from Brewer et al., 2003

What can we learn about allelic architecture from data available in 2013?

- Simulate various architectures
- Match the results with GWAS studies
- Match the results with polygenic scores
- Match the results with sequencing data

Variance explained by rare variants

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Trait</th>
<th>Candidate Genes</th>
<th>Prop. functional</th>
<th>Class Freq.</th>
<th>Enrichment (fold)</th>
<th>Variance Explained (Per Gene)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Romeo 2009</td>
<td>TG</td>
<td>[ANGPTL, (5/45)]</td>
<td>90%</td>
<td>2.6%</td>
<td>1.56</td>
<td>0.0%</td>
</tr>
<tr>
<td>Ji 2008</td>
<td>BP</td>
<td>[SC11A2, SCL22A1, KCNQ1]</td>
<td>70%</td>
<td>3.12</td>
<td>0.5</td>
<td>0.4%</td>
</tr>
<tr>
<td>Johansen 2010</td>
<td>TG</td>
<td>[APOL1, GCRB, LPL, APOE]</td>
<td>90%</td>
<td>3.12</td>
<td>0.5</td>
<td>0.4%</td>
</tr>
</tbody>
</table>
Regulatory variants

- Regulation: variants in promoters, enhancers, silencers, insulators

What can we learn about allelic architecture from first principles and data available in 1918?

- Phenotypic variation is stable in population
- Many complex and quantitative traits have intermediate heritability
- Quantitative traits are normally distributed

Selection can be directional and stabilizing

Fitness optimum

Mean phenotype

Selection can be direct and apparent (pleiotropic)

Fitness optimum

Mean phenotype

Pleiotropic selection

Beneficial allele

Direct selection

Parameters of the model

- How many sites? Target size $L$
- Strength of phenotypic effects? Effect size $\beta$
- Are these alleles under selection? Rare or common?

Selection $s$

Allele frequency $x$

Heritability by allele frequency

Effective population size: N=10,000
Quantitative trait variation is

Stable over time

Has an intermediate heritability

Normally distributed

Can we constrain allelic architecture based on the trait’s distribution?

Properties of trait distributions

Heritability and Ex. Kurtosis for several traits

Phenotype of an individual

Intermediate heritability

Integral constraints on genotype-phenotype map

Intermediate heritability

\[ h^2 = E[(Y_i - \bar{Y})^2] - \sigma_e^2 = 0.5 \]

\[ h^2 \approx 0.5 = \int ds \int d\beta \beta^2 M(\beta, s) \rho(s) \int dx 2x(1-x) \Phi(x, s, N) \]

Skew is finite and small

\[ \delta_{skew} \geq \left| \int ds \int d\beta \beta^3 M(\beta, s) \rho(s) \int dx 2x(1-x)(1-2x) \Phi(x, s, N) \right| \]

Excess kurtosis is finite and small

\[ \delta_{kurt} \geq \left| \int ds \int d\beta \beta^4 M(\beta, s) \rho(s) \int dx 2x(1-x)(1-6x(1-x)) \Phi(x, s, N) \right| \]
Parameterization

Eyre-Walker (PNAS 2010) parameterization

$\beta(s) \propto s^T$

phenotypic effect

selection strength

power falloff

The simplest model: two effect sizes

Two effect sizes model

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

$\theta$ corresponds to 20,000 genes

$\text{Max } \beta = k/h^2$

Constraining allelic architecture based on results of genomic studies

Parameterization

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Exact Eq.</th>
<th>$-S \rightarrow \infty$</th>
<th>$-S \rightarrow 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heritability</td>
<td>$2\beta^2 \frac{h^2}{1 + 2\beta^2} = h^2$</td>
<td>$\frac{1}{h^2}$</td>
<td>$0$</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>$2\beta^2 \frac{h^2}{1 + 2\beta^2} - \frac{5}{3} = \kappa$</td>
<td>$\frac{2\beta^2}{3}$</td>
<td>$-\frac{5}{3} = \kappa$</td>
</tr>
<tr>
<td>Allele Freq.</td>
<td>Rare</td>
<td>Common</td>
<td></td>
</tr>
</tbody>
</table>

Empirically observed values

Step 1 Forward evolutionary simulation

DNA sequence variation

Population genetic model

Step 2 Disease models

Genetic architecture

Frequency

Effect size

Step 3 Case/control sampling

In silico genetic studies

Linkage

GWAS

Polygenic score

Sequencing studies

Empirical data

Model parameters

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

- # singletons / Mb
- Frequency spectra for synonymous and missense sites
- Pairwise linkage
Complex phenotypes (common diseases)

Exome Sequencing Project

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

PIs: Stacey Gabriel & David Altshuler

Women's Health Initiative Sequencing Program

PI: Rebecca Jackson

Ohio State University

Heart GO

Pl: Stephen Rich

University of Virginia

Lung GO

Pls: Michael Barshad

U. of Washington

Kathleen Barnes,

Johns Hopkins University

Study design

Results: Aggregate of very rare variants

Burden of risk alleles: 1% threshold

1000 Early onset MI

1000 Controls without MI:

Men < 50

Women < 60

Most likely to be genetically influenced

Age

Men > 60

Women > 70

Allelic architecture constraints

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.
Results: Aggregate of very rare variants

Burden of risk alleles: 1% threshold

Due to lack of association?  
Due to lack of power?

84 mutations found in 6,078 cases
39 mutations found in 6,241 controls

Due to lack of association?
Due to lack of power?

Hyperlipidemia in Coronary Heart Disease
HL. GENETIC ANALYSIS OF LIPID LEVELS IN 17 FAMILIES AND DETERMINATION OF A NEW INHERITED DISORDER, COMBINED HYPERLIPIDEMIA

TABLE XII
Frequency of Hyperlipidemia

<table>
<thead>
<tr>
<th>Disorder</th>
<th>c. Age 60</th>
<th>≥ Age 60</th>
<th>Ratio 60/90</th>
<th>General population</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Monogenic hyperlipidemia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Familial hyperchylomicronemia</td>
<td>4.1</td>
<td>0.7</td>
<td>5.9</td>
<td>~0.1-0.2</td>
</tr>
<tr>
<td>Familial hypertriglyceridemia</td>
<td>5.2</td>
<td>2.5</td>
<td>2.8</td>
<td>~0.2-0.3</td>
</tr>
<tr>
<td>Combined hyperlipidemia</td>
<td>11.3</td>
<td>4.1</td>
<td>2.8</td>
<td>~0.4-0.5</td>
</tr>
<tr>
<td>Total</td>
<td>20.6</td>
<td>7.5</td>
<td>~0.6-1.0</td>
<td></td>
</tr>
<tr>
<td>B. Polygenic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypercholesterolemia</td>
<td>5.5</td>
<td>5.5</td>
<td>1.0</td>
<td>—</td>
</tr>
<tr>
<td>C. Sporadic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertriglyceridemia</td>
<td>5.8</td>
<td>6.9</td>
<td>0.8</td>
<td>—</td>
</tr>
</tbody>
</table>

Goldstein et al, JCI, 52:1544, 1973