Lecture 16: Complex Trait Heritability

Module 4: Population and Disease Genetics

• **L13: Population genetics:**
  - Measuring and understanding human variation

• **L14: Disease association mapping:**
  - Molecular basis of human phenotypic variation and disease

• **L15: Quantitative trait mapping:**
  - Intermediate phenotypes bridging the genotype-phenotype gap

• **L16: Heritability:**
  - Whole-genome disease association beyond top hits

**Today: heritability**

1. **Heritability definition and key concepts:** partitioning variance, estimating variances, narrow sense vs. broad sense
2. **Genetic architecture of complex traits:** polygenic risk scores, linear mixed models, heritability partitioning, omnigenic model of complex traits, omnigenic model of complex traits
3. **From genetic architectures to systems biology:** rank-based enrichments, genes, pathways, regulators
4. **Phenotype prediction:** imputing intermediate phenotypes, large-scale models, inference algorithms

**Lessons of GWAS**

1. **We haven't found all causal loci:** known loci explain little phenotypic variance
2. **Most loci affect transcriptional regulation:** they don’t tag coding variation

**Components of phenotypic variance**

- Assume $p$ (phenotype) = $g$ (genetic) + $e$ (environment)

**Example:** one causal variant

- Three possible genetic values in the population
- Intuition: $V[g]$ is the variance of mean phenotype across different genetic values
- $V[e]$ is the variance of phenotype for the same genetic value
Components of genetic variance

- Assume $V[g] = V[a]$ (additive) + $V[d]$ (dominance) + $V[i]$ (interactions)
- The additive component corresponds to a linear model
- As we add more causal variants, phenotypes become closer to Gaussian
- We could further decompose interactions
- We could include variance due to de novo mutations

Heritability is a ratio of variances

- $V[p] = V[g] + V[e]$
- $V[g] = V[a] + V[d] + V[i]$
- **Broad sense heritability**
  $H^2 = \frac{V[g]}{V[p]}$
- Broad sense captures all genetic factors
- **Narrow sense heritability**
  $h^2 = \frac{V[a]}{V[p]}$
- Narrow sense captures only additive effects
- Ongoing debate about the relative importance of additive vs. other effects in disease, selection, etc.

Why study heritability?

- Quantify the importance of genetics vs. environment in traits of interest
- Learn about *genetic architecture*: how many causal variants, effect sizes, allele frequencies
- Narrow sense heritability is the fundamental parameter needed for phenotype prediction (and is the theoretical best possible prediction performance with a linear model)

Estimating heritability in relatives

- Intuition: heritability relates phenotypic correlations to genotypic correlations
- If two individuals have the same allele at each of the causal variants, they will have the same phenotype
- **Haseman-Elston regression**: fit linear regression of phenotypic correlations against genotypic correlations
- Derive genotypic correlation from family relationships: monozygotic twins share 100% of genome, siblings share 50%, etc.
- Example (height): $h^2 = 0.73$

Estimating heritability from GWAS

- Linear model $g = X\beta$
- We can estimate SNP effect sizes $\beta$ from GWAS
- The variance explained by each SNP depends on effect size and MAF
- $V[X_\beta] = 2f(1-f)\beta^2$ if we do this with genome-wide significant SNPs, we usually $h^2_{GWAS} < h^2$
- Example (height): 253,288 samples; 697 genome-wide significant loci; $h^2_{GWAS}=0.16$, $h^2 = 0.73$
- Known as the *missing heritability problem*

Sources of missing heritability

- Ongoing debate about several possible explanations for the missing heritability problem.
  1. Many common variants, small effects
  2. Unobserved rare variants, large effects
  3. Wrong model assumptions

Each has very different implications for the future of human genetics studies.
Today: complex trait heritability

- **Fundamental concepts:** partitioning variance, estimating variances, narrow sense vs. broad sense
- **Genetic architecture of complex traits:** polygenic risk scores, linear mixed models, heritability partitioning, omni-locus model of complex traits
- **From genetic architectures to systems biology:** rank-based enrichments, genes, pathways, regulators
- **Phenotype prediction:** imputing intermediate phenotypes, large-scale models, inference algorithms

Evidence for many common variants

1. Polygenic risk scores
2. Linear mixed models
3. Variable selection in regression
4. Forward simulation of risk loci

Polygenic risk scores

- Aggregate burden of sub-threshold SNPs to improve prediction performance (Stahl 2012)
- As we include more SNPs in the risk score, the association with RA, celiac disease, MI, CAD gets stronger
- In practice, requires tuning of p-value threshold, LD pruning threshold

Linear mixed models

\[ p \sim \mathcal{N}(0, h^2 G - (1 - h^2) I) \]
\[ G = XX' / p \]

- Joint model of all SNPs explains more heritability (Yang 2010)
- Idea: under suitable assumptions, \( V[a] = \Sigma \beta_j^2 \)
- Under the infinitesimal assumption \( \beta_j \sim \mathcal{N}(0, h^2 / p) \), we can estimate \( V[a] \) without estimating individual \( \beta_j \) using residual maximum likelihood (REML)
- REML avoids using ML fit of parameters, instead uses transformed data so that nuisance parameters have no effect.
- In variance components analysis (random effects model), transformation focuses on differences, sum of variances
- **This works despite not knowing the causal variants**
- Example (height): \( h^2_{\text{GWAS}} = 0.16, h^2 = 0.73, h^2_g = 0.5 \)

Partitioning heritability

- Extend the model so chromosomes can explain different proportions of variance
- Intuition: genetic relationship matrix \( G \) captures identity by state in unrelated individuals
- This is again the probability of sharing the same allele at the causal variants
- This is called **PCGC regression** (Golan 2015) (phenotype correlation – genotype correlation regression)
Partitioning heritability

- Fit a model with one component per 1MB window (Loh 2015)
- Bound cumulative heritability explained to estimate number of regions
- Most of the genome explains non-zero heritability

Bayesian variable selection

- Directly fitting the underlying linear model is ill-posed: we have $n < p$ so there are infinitely many solutions
- Idea: use spike and slab prior to force many effects to be exactly 0 and regularize the problem (one solution)
- Inference goal: estimate the effect sizes and the level of sparsity (Carbonetto 2013)

Pathways-informed prior from enrichments

- Extension: some pathways contain more causal variants than the rest of the genome
- Incorporate into the prior
- Identifies relevant immune signaling pathways which are not found using existing methods
- Identifies tens of thousands of SNPs which could be affecting those pathways

Forward simulation of T2D

- Simulate realistic loci using known population/evolutionary parameters (Agarwala 2012)
- Simulate disease phenotypes varying number of causal loci, heritability, prevalence, strength of selection
- Perform twin studies, GWAS and compare predicted results to observed results on real data

Distinguish disease architectures

- Only some architectures consistent with observed data

Evidence for other explanations

- Incorporating Identity by Descent (IBD) in unrelated individuals
- Partitioning SNPs by MAF, LD
- Assumptions do not hold in real data
Estimating heritability: shared haplotypes

- Shared haplotypes explain more heritability than tag SNPs
- There is a still a discrepancy between \( h^2_g \) and \( h^2 \)
- If two individuals share a chromosomal segment, unobserved variants should also be shared (Bhatia 2015)
- Idea: Identify IBD segments by quickly scanning SNPs and finding stretches of identical alleles
- Inferring shared segments captures rarer variants more effectively than LD

Image credit: http://gcbias.org/european-genealogy-faq/

Partitioning SNPs by MAF/LD

- Low frequency/low LD variants are poorly tagged by observed/imputed variants, so estimate variance for them separately (Yang 2015)
- Partitioning appears to explain all of the heritability of height using only common/low frequency variants!

Examining model assumptions

- Phenotypes might not be Gaussian
- GWAS samples are not independent and identically distributed
- SNPs are not independent
- Not all SNPs have an effect
- Not all causal SNPs have equal effects
- There are gene-environment interactions
- There are gene-gene interactions

Limitations of heritability

- Explaining all of the heritability of complex traits is not enough
- As sample size goes to infinity, will the entire genome be associated with all traits? (Goldstein 2009)
- Goal: Find biological pathways recurrently disrupted by non-coding variation

Regulatory enrichments

- Weakly associated variants overlap accessible chromatin more often than expected by chance (Maurano 2012)
- Same trend observed in other predicted regulatory elements: histone peaks, ChromHMM segments, super enhancer clusters

Joint model of SNPs and annotations

- Use penalized stepwise regression to pick relevant annotations (Pickrell 2014)
- Use approximate Bayes factors to compute posterior probability of association
- Forward steps: add annotations to the model until they don’t explain enough variance
- Backward steps: remove annotations from the fitted model until variance explained drops too much
Joint model of SNPs and annotations

- Use approximate Bayes factors to compute posterior probability of association
- Posterior probability of association re-prioritizes new GWAS loci

Partitioning heritability by annotation

- Accessible chromatin explains more heritability
- Combine DHS in >100 cell types: 70% of genome is accessible in some cell type, but only 16% is accessible in multiple cell types
- Implies non-coding SNPs explain more variance per SNP than coding SNPs

LD score regression

$$E[z^2] = N h^2 / M$$

- Intuition: Causal variants drawn uniformly at random from the genome are more likely to come from larger LD blocks (Bulik-Sullivan 2014)
- Linear regression of summary statistics against LD score gives h^2 without access to genotypes

Partitioning heritability by cell type

- LD score regression generalizes easily to multiple components (Finucane 2014)
- Fit different annotations separately (different heritability parameter)

Omnigenic model of heritability

- (A) Genome-wide inflation of small p values from the GWAS for height, with particular enrichment among expression quantitative trait loci and single-nucleotide polymorphisms (SNPs) in active chromatin (H3K27ac).
- (B) Estimated fraction of SNPs associated with non-zero effects on height (Stephens, 2017) as a function of linkage disequilibrium score (i.e., the effective number of SNPs tagged by each SNP; Bulik-Sullivan et al., 2015b). Each dot represents a bin of 1% of all SNPs, sorted by LD score. Overall, we estimate that 62% of all SNPs are associated with a non-zero effect on height. The best-fit line estimates that 3.8% of SNPs have causal effects.
- (C) Estimated mean effect size for SNPs, sorted by GIANT p value with the direction (sign) of effect ascertained by GIANT. Replication effect sizes were estimated using data from the Health and Retirement Study (HRS). The points show averages of 1,000 consecutive SNPs in the p-value-sorted list. The effect size on the median SNP in the genome is about 10% of that for genome-wide significant hits.

More heritability in broad classes

- Contributions to heritability (relative to random SNPs) as a function of chromatin context. There is enrichment for signal among SNPs that are in chromatin active in the relevant tissue, regardless of the overall tissue breadth of activity.
- Genes with brain-specific expression show the strongest enrichment of schizophrenia signal (left), but broadly expressed genes contribute more to total heritability due to their greater number (right).
Most GO categories are enriched

- Gene Ontology Enrichments for Three Diseases, with Categories of Particular Interest Labeled. The x axis indicates the fraction of SNPs in each category; the y axis shows the fraction of heritability assigned to each category as a fraction of the heritability assigned to all SNPs. Note that the diagonal indicates the genome-wide average across all SNPs; most GO categories lie above the line due to the general enrichment of signal in and around genes. Analysis by stratified LD score regression.

Boyle, Li, Pritchard, Cell, 2017

Today: complex trait heritability

- **Fundamental concepts:** partitioning variance, estimating variances, narrow sense vs. broad sense
- **Genetic architecture of complex traits:** polygenic risk scores, linear mixed models, heritability partitioning, omnigenic model of complex traits
- **From genetic architectures to systems biology:** rank-based enrichments, genes, pathways, regulators
- **Phenotype prediction:** imputing intermediate phenotypes, large-scale models, inference algorithms

Boyle, Li, Pritchard, Cell, 2017

How far down the SNP list does enrichment go?

- Use functional enrichment to gain insight into genetic architecture (Sarkar 2016)
- Idea: as we consider more SNPs beyond genome-wide significance, relevant regulatory regions will be disrupted more often than irrelevant regions

Long tails of enrichment for 8 diseases

- Use functional enrichment to gain insight into genetic architecture (Sarkar 2016)
- Idea: as we consider more SNPs beyond genome-wide significance, relevant regulatory regions will be disrupted more often than irrelevant regions

Enhancer modules: constitutive, cell type specific

- Challenge: annotations learned one cell type at a time can’t account for sharing of elements across cell types
- Use k-means clustering to define modules of enhancer activity
- Functional enrichments highlight importance of both constitutive and lineage-specific enhancers
From enhancers to genes to pathways

<table>
<thead>
<tr>
<th>Trait</th>
<th>Known pathways</th>
<th>Total genes</th>
<th>Total pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>Cyclic GMP signaling, immune response</td>
<td>230</td>
<td>218</td>
</tr>
<tr>
<td>BIP</td>
<td>Glucocorticoid signaling</td>
<td>257</td>
<td>230</td>
</tr>
<tr>
<td>CAD</td>
<td>Cholesterol/triglyceride metabolism, IgA</td>
<td>288</td>
<td>215</td>
</tr>
<tr>
<td>CD</td>
<td>CD8 T cell proliferation, IgE, IL4</td>
<td>224</td>
<td>359</td>
</tr>
<tr>
<td>RA</td>
<td>NFKB, actin nucleation</td>
<td>196</td>
<td>146</td>
</tr>
<tr>
<td>SCZ</td>
<td>Dendritic spine development</td>
<td>271</td>
<td>183</td>
</tr>
<tr>
<td>T1D</td>
<td>MHC I/II, JAK-STAT, IFNG</td>
<td>266</td>
<td>245</td>
</tr>
<tr>
<td>T2D</td>
<td>Pancreatic beta cell apoptosis</td>
<td>281</td>
<td>177</td>
</tr>
</tbody>
</table>

• Link enhancers to their downstream target genes
• Target genes enriched in known disease pathways, but through previously unknown mechanisms
• Reveals broad similarities at pathway level between classes of diseases (e.g. signaling in autoimmune traits), but also specific pathways important to each disease
• Potentially implicate novel genes in enriched pathways

From genes/pathways to upstream regulators

• Challenge: heritability-based methods can’t identify specific enhancer regions
• Our method can implicate specific enhancers, so we can dissect their mechanism
• Predict the upstream regulator using sequence-based enrichment (Kheradpour 2013) without considering GWAS
• Find master regulators recurrently disrupted by sub-threshold SNPs
• Many disease-specific regulators, but interesting shared regulators

Regulator → gene networks across diseases

• GWAS associated SNP often does not directly disrupt the predicted master regulator
• Instead, falls in a different motif instance for a putative co-factor
• Explains how master regulators can be shared across very different phenotypes (NFKB in schizophrenia, T1D)

Upstream regulators add cell-type-specificity

• Many predicted master regulators found in predicted constitutive enhancers rather than cell type-specific regulators
• Although enhancers might be constitutively marked, expression of the upstream regulator is cell type-specific
• Additional insight into transcriptional regulation needed to interpret non-coding disease associations

Today: complex trait heritability

• Fundamental concepts: partitioning variance, estimating variances, narrow sense vs. broad sense
• Genetic architecture of complex traits: polygenic risk scores, linear mixed models, heritability partitioning, omnigenic model of complex traits
• From genetic architectures to systems biology: rank-based enrichments, genes, pathways, regulators

• Phenotype prediction: imputing intermediate phenotypes, large-scale models, inference algorithms
**Key Idea:**
- Learn G→M model (ROSMAP n=800) Fewer indiv. Simpler phenotype
- Impute methylation IM for GWAS cohort (n=74k)
- iMWAS between genotype-driven M and AD phenotype (n=47k)

**Advantage:**
- Much larger GWAS cohorts (>>MWAS): increased power
- Genetic component of methyl. variation

**Logistical challenge:**
- Summary stats, not full genotypes  ➔  Linear model, impute stats direct

---

**iMTWAS: Imputation across multiple intermediate variables**

Model multiple mediator variables  
SNP → Methylation → Expression → Disease  
Predict new loci, increased power  
Predict regulatory regions & target genes

---

**Goal 2: Dissect molecular and cellular phenotypes**

<table>
<thead>
<tr>
<th>Whole tissue</th>
<th>1,000 individuals x 8 tissues / cell types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurons</td>
<td></td>
</tr>
<tr>
<td>Astrocytes</td>
<td></td>
</tr>
<tr>
<td>Oligodendrocytes</td>
<td></td>
</tr>
<tr>
<td>Microglia</td>
<td></td>
</tr>
</tbody>
</table>

| GWAS: G → D | N=74k  | Learn G→D directly (complex phenotype) |
| meQTLG → M  | N=800  | Learn G→M (simpler phenotype)          |
| eQTL: G → T  | N=800  | Learn G→M (simpler phenotype)          |
| MWAS: M → D  | N=800  | M→D (no causality)                     |
| TWAS: T → D  | N=800  | M→T (no causality)                     |
| NWAS: G → IM | N=74k  | IM→D (causality)                       |
| iMWAS: G → IM |        |                                          |

---

**Multivariate regression for prediction**

- Recent successes in recent TWAS highlighted importance of sub-threshold GWAS loci and novel genes associated with traits.

- Highly heritable genes are remarkably well-predicted SNPs in cis-regulatory regions.

- If the regression overfit to the reference panel...

- No TWAS

---

**Imputation of inaccessible tissues from surrogates (e.g. blood)**

- Imputed data surpasses observed data quality
- Incorporate cis, trans, multi-tissue effects
- Common meta-sample decomposition matrix
- Combine genetic and environmental effects
Formal definition of a linear model

\[ y = (y_1, y_2, \ldots, y_n), \quad X = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ X_{21} & \cdots & X_{2p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix} \]

In matrix notation, expression \( y \) as a factor of genetic information \( x \)

\[ y = X\theta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I). \]

\( \theta \) = effect size (can be itself sampled from a normal prior)

Maximum likelihood estimation w/o prior distribution

Here, we work on Gaussian for simplicity, but other linear models can be written as log-likelihood as easily as this.

\[ y = X\theta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I). \]

Data log-likelihood (as func. of \( \theta \))

\[ \ln p(y|X, \theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i\theta)^2 + \text{const}. \]

Ignore other constant terms with respect to \( \theta \) for simplicity.

Intractability of MLE for large-scale linear models

- If \( n \geq p \), we could have point estimation.
- If \( n < p \), we have line or surface, as solution.

Variable selection

- Intuitive idea: choose the best combination of variables. \( \rightarrow 2^p \) choices (even harder).
- Alternative idea: make as many \( \theta_j \)'s nearly zero values.
- What prior does: penalize \( |\theta_j| > 0 \) so that only the strong enough variables take non-zero values.
Variable selection by shrinkage of coeff. $\theta$

Consider regression model: $y_i = \theta_1 x_{i1} + \theta_2 x_{i2}$.

L1-regularization and Laplace prior

Maximizing a posteriori 

\[
\ln p(y|x, \theta) + \ln p(\theta | \lambda) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i \theta)^2 + \lambda \| \theta \|_1
\]

Minimize $L_1$-regularized error 

\[
-\sum_{i=1}^{n} (y_i - x_i \theta)^2 + \lambda \| \theta \|_1
\]

Prior distribution 

\[
p(\theta | \lambda) \propto \exp(-\lambda \| \theta \|_1)
\]

No analytical solution to regression with Laplace prior ($L_1$-regularized)

- L1-regularized (or Laplace prior) regression can select variables with statistical guarantee if variables are not colinear to each other.
- However, there is no closed form solution of point estimation (MAP).
- Moreover, there is no closed form solution of posterior probability.
- In order to identify confidence interval of the point estimates, we can utilize bootstrapping (active research area; conformal inference).

L2-regularization and Gaussian prior

Maximizing a posteriori 

\[
\ln p(y|x, \theta) + \ln p(\theta | \lambda) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i \theta)^2 + \frac{\lambda}{2} \| \theta \|^2
\]

Minimize $L_2$-regularized error 

\[
-\sum_{i=1}^{n} (y_i - x_i \theta)^2 + \frac{\lambda}{2} \| \theta \|^2
\]

Prior distribution 

\[
N(\theta | 0, \lambda^{-1}I) \propto \exp\left(-\frac{\lambda}{2} \| \theta \|^2\right)
\]

\[
\hat{\theta} = (X^T X + \lambda \sigma^2 I)^{-1}X^T y
\]

(See next slides for derivation).

Comparison between MLE and Bayesian regression

$\hat{\theta}_{MLE} = (X^T X)^{-1}X^T y$.

$\theta_{Bayes} = (X^T X + \lambda \sigma^2 I)^{-1}X^T y$

(a) This is classically called Ridge regression.
(b) Matrix inverse can exists with a proper choice of tuning parameter $\lambda$, or hyper-parameter in our Bayesian framework.
(c) Prior $\lambda$ and noise StdDev $\sigma$ can co-adapt to each other.
(d) We can first set $\sigma$ using the null model (setting $\theta$ = zero) or with MLE $\theta$, and find $\lambda$ accordingly.
L1 and L2-regularized regression effectively handle issues in practice

Combining both L1 and L2, we can perform variable selection, alleviate collinearity.
(Original Elastic Net paper: Zhou & Hastie, 2005)
We will see further discuss on collinearity later (section).

Spike-slab prior model effectively avoid colinearity

Can L1-regularized one handle this?

If correlation between $X_1 \sim X_2$ is strong, probably not ...
(best solution within the box is still non-zero for both vars).

Today: complex trait heritability

- **Fundamental concepts**: partitioning variance, estimating variances, narrow sense vs. broad sense
- **Genetic architecture of complex traits**: polygenic risk scores, linear mixed models, heritability partitioning, omnigenic model of complex traits
- **From genetic architectures to systems biology**: rank-based enrichments, genes, pathways, regulators
- **Phenotype prediction**: imputing intermediate phenotypes, large-scale models, inference algorithms