Lecture 16: Systems Genetics
Heritability, Polygenicity, Pleiotropy

Slides credit:
Abhishek Sarkar,
Yongjin Park
<table>
<thead>
<tr>
<th>Project</th>
<th>Psets</th>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
<th>Loc</th>
<th>Topic</th>
<th>Read*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>Thu, Sep 7</td>
<td>Introduction</td>
<td>L1</td>
<td>Intro: Biology, Algorithms, Machine Learning, Course Overview</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>Fri, Sep 8</td>
<td></td>
<td>R1</td>
<td>Recitation 1: Biology and Probability Review</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Thu, Sep 12</td>
<td>Module I: Aligning and Modeling Genomes</td>
<td>L2</td>
<td>Alignment I: Dynamic Programming, Global and local alignment</td>
<td>2,3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Thu, Sep 14</td>
<td></td>
<td>R2</td>
<td>Recitation 2: Deriving Parameters of Alignment, Multiple Alignment</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Fri, Sep 15</td>
<td></td>
<td>L3</td>
<td>Alignment II: Database search, Rapid string matching, BLAST, BLOSUM</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Fri, Sep 19</td>
<td></td>
<td>L4</td>
<td>Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Sat, Sep 21</td>
<td></td>
<td>L5</td>
<td>Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Sun, Sep 22</td>
<td></td>
<td>L6</td>
<td>No classes - student holiday</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Mon, Sep 25</td>
<td></td>
<td>L7</td>
<td>Project Intro: about the projects, self introductions, mentor intro, example projects, teamwork 32D-507</td>
<td>15,16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Thu, Sep 28</td>
<td></td>
<td>L8</td>
<td>Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Fri, Sep 29</td>
<td></td>
<td>L9</td>
<td>Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Tue, Oct 3</td>
<td></td>
<td>L10</td>
<td>Recitation 3: Affinity Propagation Clustering and Random Forest Classification</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Tue, Oct 4</td>
<td></td>
<td>L11</td>
<td>Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Tue, Oct 5</td>
<td></td>
<td>L12</td>
<td>Three-dimensional chromatin interactions: 3C, 5C, HiC, HiChIP</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Fri, Oct 6</td>
<td></td>
<td>L13</td>
<td>Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Fri, Oct 6</td>
<td></td>
<td>L14</td>
<td>Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Tue, Oct 10</td>
<td></td>
<td>L15</td>
<td>No Courses - Columbus Day Holiday</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Thu, Oct 12</td>
<td></td>
<td>L16</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Fri, Oct 13</td>
<td></td>
<td>L17</td>
<td>Recitation 5: Gapped Motif Discovery, DNAXShape, PBMs, Selex</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Thu, Oct 17</td>
<td></td>
<td>L18</td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td>20,21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Thu, Oct 19</td>
<td></td>
<td>L19</td>
<td>Deep Learning, Neural Networks, Convolutional NNs, Recurrent NNs, Autoencoder</td>
<td>20,22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Fri, Oct 20</td>
<td></td>
<td>L20</td>
<td>Recitation 6: Networks review, Recommendation systems, EHR, PhewAS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>Fri, Oct 6</td>
<td></td>
<td>L21</td>
<td>No recitation, Veterans Day</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>Fri, Oct 27</td>
<td></td>
<td>L22</td>
<td>Project feedback: Prepare 2-3 slide presentation of your term project for your mentor, 320-507 at 4-5pm</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Tue, Oct 24</td>
<td></td>
<td>L23</td>
<td>Module IV: Population Genetics and Disease Genomics</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Tue, Oct 26</td>
<td></td>
<td>L24</td>
<td>Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Fri, Oct 27</td>
<td></td>
<td>L25</td>
<td>Disease Association Mapping, GWAS, organismal phenotypes</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Fri, Nov 3</td>
<td></td>
<td>L26</td>
<td>Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Thu, Nov 2</td>
<td></td>
<td>L27</td>
<td>Panel Discussion: reconciling critiques, strategies for improvement, feedback to author 32D-507</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Fri, Nov 3</td>
<td></td>
<td>L28</td>
<td>Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, IMWAS</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Fri, Nov 3</td>
<td></td>
<td>L29</td>
<td>Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Fri, Nov 3</td>
<td></td>
<td>L30</td>
<td>No recitation, Veterans Day</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Thu, Nov 7</td>
<td></td>
<td>L31</td>
<td>Recitation 8: Rare Variants, ExAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Fri, Nov 10</td>
<td></td>
<td>L32</td>
<td>Comparative genomics and evolutionary signatures</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Thu, Nov 8</td>
<td></td>
<td>L33</td>
<td>Genome Scale Evolution, Genome Duplication</td>
<td>4,5,7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Thu, Nov 14</td>
<td></td>
<td>L34</td>
<td>No recitation, Veterans Day</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Fri, Nov 17</td>
<td></td>
<td>L35</td>
<td>No recitation, Veterans Day</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Thu, Nov 16</td>
<td></td>
<td>L36</td>
<td>In Class Quiz (the only quiz - the class has no final exam) - covers L1-L20,R1-R9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Fri, Nov 17</td>
<td></td>
<td>L37</td>
<td>Quiz</td>
<td>No lecture, thanksgiving break - Thu Nov 26, 2015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>Thu, Nov 23</td>
<td></td>
<td>L38</td>
<td>No recitation, thanksgiving break</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>Fri, Nov 24</td>
<td></td>
<td></td>
<td>No recitation, thanksgiving break</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>Thu, Nov 28</td>
<td></td>
<td>L39</td>
<td>Module VI: Current Research Directions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>Thu, Nov 30</td>
<td></td>
<td>L40</td>
<td>Single-cell genomics: technology, analysis, microfluidics, applications, insights</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Fri, Dec 1</td>
<td></td>
<td>L41</td>
<td>Mining human phenotypes, PhewAS, UK Biobank, meta-phenotypes+imputation</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Thu, Dec 5</td>
<td></td>
<td>L42</td>
<td>Recitation 10: Project Feedback, results, interpretation, directions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>Thu, Dec 7</td>
<td></td>
<td>L43</td>
<td>Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>Fri, Dec 8</td>
<td></td>
<td>L44</td>
<td>Genome Engineering with CRISPR/Cas9 and related technologies</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>Tue, Dec 12</td>
<td></td>
<td>L45</td>
<td>No recitation, Veterans Day</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>Tue, Dec 12</td>
<td></td>
<td>L46</td>
<td>Final Presentations - Part I (11am), 32-G8 reading room</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>Tue, Dec 12</td>
<td></td>
<td>L47</td>
<td>Final Presentations - Part I (1pm), 32-141</td>
<td></td>
</tr>
</tbody>
</table>

* readings refer to chapters in compiled 2016 scribe notes, available in the materials folder on Stellar
** recitation topics will be adjusted to respond to lecture and student needs
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: Systems Genetics

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
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
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
Today: Systems Genetics

- Heritability
- Polygenic risk scores
- Disease architecture
- Linear Mixed Models (LMMs)
- Partitioning Heritability
- LD score regression
- Polygenicity
Components of phenotypic variance

• Assume \( p \) (phenotype) = \( g \) (genetic) + \( e \) (environment)

• Then, \( V[p] = V[g] + V[e] + 2\text{Cov}(G,E) \)
  (assume no gene-environment interactions)

- 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

\[ p = g + e \]
\[ E[p_i p_j] = h^2 E[g_i g_j] \]

• 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
  \[ \text{Var}(X_j \beta_j) = 2 f_j (1 - f_j) \beta_j^2 \]
- If we do this with genome-wide significant SNPs, we usually \( h^2_{\text{GWAS}} < h^2 \)
- Example (height): 253,288 samples; 697 genome-wide significant loci; \( h^2_{\text{GWAS}} = 0.16 \), \( h^2 = 0.73 \)
- 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: Systems Genetics

- Heritability
- **Polygenic risk scores**
  - Disease architecture
  - Linear Mixed Models (LMMs)
  - Partitioning Heritability
  - LD score regression
  - Polygenicity
Estimate absolute risk combining genetic and environmental risk factors

Possible clinical decisions

- General advice on having a healthy lifestyle
- Mammography screening frequency tailored to risk
- Lifestyle changes
- Frequent mammography screening
- Discuss preventive therapies
- Individual counselling in primary care and referral to secondary or tertiary care
- Enhanced screening and surveillance
- Chemoprevention and/or endocrine therapy
- Risk-reducing surgery (mastectomy, salpingo-oophorectomy)

Absolute risk

Possible risk factor profile

- No family history of breast cancer, low to moderate polygenic risk, and none or few environmental risk factors
- No family history of breast cancer, moderate polygenic risk and several environmental risk factors
- Moderate to high polygenic risk with family history of breast cancer and many environmental risk factors, or known BRCA1 and BRCA2 or TP53 mutation carriers for very high risk

Chatterjee et al. Nature Reviews Genetics (2016)
How do we estimate polygenic risk score?

Univariate GWAS statistics teach us:

\[ \beta_j = \log(\text{odds ratio of SNP } j) \]
\[ g_j = \text{genotype (dosage)} \]

Predict overall risk by combining many, many variants!

\[ \text{PRS} = \sum_{j \in \{\text{SNPs}\}} \beta_j g_j \]

**Can we just combine all the SNPs? Why not?**

- Is correlation between \( g_1 \) and \( g_2 \) zero?
- Can we trust the estimate \( \beta \) of all the SNPs?
- Can we just select GWAS significant SNPs?
A common practice of PRS estimation

Univariate GWAS statistics:

\[ \beta_j = \log(\text{OR of SNP } j) \]
\[ g_j = \text{genotype (dosage)} \]

PRS model:

\[ \text{PRS}[i] = \sum_{j \in \{\text{SNPs}\}} \beta_j g_j[i] \]

Goal: Tuning this parameter

Filter #1: p-value thresholding

Filter #2: LD pruning
A common practice of PRS estimation: Cross-validation with observed phenotype

Univariate GWAS statistics:

\[ \beta_j = \log(\text{OR of SNP } j) \]
\[ g_j = \text{genotype (dosage)} \]

PRS model:

\[ \text{PRS}[i] = \sum_{j \in \{\text{SNPs}\}} \beta_j g_j[i] \]

Goal: Tuning this parameter

How do we know the selected SNPs are good?

How do we know the selected SNPs are good?

Observed risk

Predicted risk

AUROC

- GWAS heritability (AUC=71.9%)
- 500K/500K (AUC=69.7%)
- 200K/200K (AUC=65.9%)
- 59K/59K (AUC=62.3%)
An alternative method for estimating PRS (and a simpler and more powerful way)

Univariate GWAS statistics:

\[ \beta_j = \log(\text{OR of SNP } j) \]
\[ g_j = \text{genotype (dosage)} \]

PRS model:

\[ \text{PRS}[i] = \sum_{j \in \{\text{SNPs}\}} \beta_j g_j[i] \]

What’s wrong with using all the SNPs? LD between them. Adjust spurious weak effects.

Chun .. Sunyeav, BioRxiv (2019)
Baker et al., Genetic Epidemiology (2017)
Idea: Decorrelate LD structure

- Transform SNP space to multi-SNP space (SVD)
- Select independent & orthogonal factors.
- Or regularize eigenvalues to smooth out spurious associations.
- We don’t need much tuning with regularization.

Baker et al., Genetic Epidemiology (2017)
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
Today: Systems Genetics

- Heritability
- Polygenic risk scores
  - Disease architecture
- Linear Mixed Models (LMMs)
- Partitioning Heritability
- LD score regression
- Polygenicity
Partitioning heritability

- Extend the model so chromosomes can explain different proportions of variance
- Intuition: add more variance parameters for each partition of SNPs
- Each partition induces a different genetic relationship matrix
- Longer chromosomes explain more heritability
- Suggests causal variants are spread uniformly through the genome
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

<table>
<thead>
<tr>
<th>enriched pathway</th>
<th>Bayes factor</th>
<th>number of genes</th>
<th>number of SNPs</th>
<th>genome-wide log-odds $\theta_0$</th>
<th>log$_{10}$-fold enrichment $\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1D IL-2 signaling*</td>
<td>1.2 x 10$^{12}$</td>
<td>52</td>
<td>1964</td>
<td>-6</td>
<td>0</td>
</tr>
<tr>
<td>Measles*</td>
<td>2.0 x 10$^{11}$</td>
<td>10</td>
<td>3494</td>
<td>-5</td>
<td>1</td>
</tr>
<tr>
<td>CD Cytokine signaling</td>
<td>9.0 x 10$^5$</td>
<td>66</td>
<td>6711</td>
<td>-4</td>
<td>2</td>
</tr>
<tr>
<td>IL-23 signaling</td>
<td>1.4 x 10$^4$</td>
<td>111</td>
<td>2218</td>
<td>-3</td>
<td>3</td>
</tr>
<tr>
<td>IL-12 signaling</td>
<td>8417</td>
<td></td>
<td>3641</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>Immune system (BS)</td>
<td>1897</td>
<td></td>
<td></td>
<td>-1</td>
<td>5</td>
</tr>
<tr>
<td>Immune system (PC)</td>
<td>1168</td>
<td></td>
<td></td>
<td>-0</td>
<td>0</td>
</tr>
<tr>
<td>RA Measles*</td>
<td>2576</td>
<td>130</td>
<td>15074</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Release of eIF4E*</td>
<td>713</td>
<td>6</td>
<td>3488</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>T2D Incretin regulation</td>
<td>259</td>
<td>18</td>
<td>216</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Regulation of GLP-1</td>
<td>241</td>
<td>52</td>
<td>689</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Id signaling</td>
<td>241</td>
<td></td>
<td>593</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

- 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 still a discrepancy between $h_g^2$ 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
Today: Systems Genetics

- Heritability
- Polygenic risk scores
- Disease architecture
- Linear Mixed Models (LMMs)
- Partitioning Heritability
- LD score regression
- Polygenicity
Formal definition of a linear model

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

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

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

\( \theta = \text{effect size (can be itself sampled from a normal prior)} \)
What are we missing in the previous multivariate model?

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

Assume IID individuals. This may not be true.

\[ y = X\theta + u + \epsilon. \]

Add random effects to account for the unknown

\[ u \sim \mathcal{N}(0, K) \]

We assume this random effect can be captured by Kinship covariance.

In GWAS problems, the most influential/spurious random effect stems from population structure.
Why do we need a random effect?

- Unknown population structure
- Influence to many SNPs
- Phenotypic variation due to both pop. struct. & actual association

\[ u \rightarrow y \]
\[ x_1, \ldots, x_p \rightarrow y \]
A Bayesian approach to account for the random effect $u$

Likelihood model:

$$y = X\theta + u + \epsilon.$$  

(Empirical) prior knowledge: $u \sim \mathcal{N}(0, K)$

A Bayesian method $\approx$ Address/remove uncertainty by averaging out

$$p(y|X\theta) = \int p(y|X\theta, u)p(u)du$$

A Linear mixed effect model:

$$y = X\theta + \tilde{\epsilon}$$ with $\tilde{\epsilon} \sim \mathcal{N}(0, \sigma^2 I + \tau^2 K)$

two components in covariance matrix

two components in covariance matrix

IID error

Kinship components
Linear mixed models

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

- Joint model of all SNPs explains more heritability (Yang 2010)
- Idea: under suitable assumptions, \( V[a] = \Sigma \beta_j^2 \)
- Under the infinitesimal assumption \( \beta_j \sim N(0, h^2/p) \), 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): \( h^2_{GWAS} = 0.16, h^2 = 0.73, h^2_g = 0.5 \)
Linear mixed models

\[ p \sim N(0, h^2 G - (1 - h^2) I) \]
\[ G = XX' / p \]
\[ E[p_i p_j] = h^2 G_{ij} \]

• We can generalize Haseman-Elston regression to estimate heritability for unrelated individuals using LMM
• 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)
Today: Systems Genetics

- Heritability
- Polygenic risk scores
- Disease architecture
- Linear Mixed Models (LMMs)
  - Partitioning Heritability
- LD score regression
- Polygenicity
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
Today: Systems Genetics

- Heritability
- Polygenic risk scores
- Disease architecture
- Linear Mixed Models (LMMs)
- Partitioning Heritability
- LD score regression
- Polygenicity
LD score regression

\[ E[z_j^2] = N l_j h^2 / M \]

- 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 individual-level genotype matrix

Image credit: Simoni 2008
LD score regression estimates heritability from summary data

A multivariate model for phenotype variation

\[ y_i = \sum_{j} X_{ij} \beta_j + \varepsilon_i \]

Assuming \( \mathbb{E}[X_j] = 0 \) and \( \text{V}[X_j] = 1 \), heritability = \( \text{V}[X\beta] \approx \Sigma X^2 \beta^2 \approx \Sigma \beta^2 \)

Heritability by partitioning (restricting on a set \( C \)):

\[ h^2 (C) = \sum_{j \in C} \beta_j^2 \]

Finucane et al. (2015)
LD score regression estimates heritability from summary data

A multivariate model

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

Assuming \( \mathbb{E}[X_j] = 0 \) and \( \mathbb{V}[X_j] = 1 \), heritability = \( \mathbb{V}[X \beta] \approx \Sigma X^2 \beta^2 \approx \Sigma \beta^2 \)

Heritability by partitioning (restricting on a set \( C \)):

\[ h^2(C) = \sum_{j \in C} \beta_j^2 \]

Summary statistics data

(1) \( \chi_j^2 \) - square tests statistic for all SNP \( j \)
(2) \( r_{jk}^2 \) - LD matrix (or correlation between SNP \( j \) and \( k \))

Finucane et al. (2015)
Idea: Reverse-engineer summary data to find multivar. parameters

A univariate effect (GWAS)

\[ \hat{\beta}_j = \frac{1}{N} X_j^T (X \beta + \epsilon) \]

\[ = \sum_k \hat{r}_{jk} \beta_k + \epsilon'_j \]

LD between SNP \(j\) and \(k\)

A univariate chi-square (GWAS)

\[ \chi^2_j = N \hat{\beta}_j^2 \]

\[ \text{E}[\chi^2_j] = NE \left( \sum_k \hat{r}_{jk} \beta_k + \epsilon'_j \right)^2 \]

Finucane et al. (2015)
A univariate effect (GWAS)

\[ \hat{\beta}_j = \frac{1}{N} X_j^T (X \beta + \epsilon) \]

\[ = \sum_k \hat{r}_{jk} \beta_k + \epsilon'_j \]

LD between SNP \( j \) and \( k \)

Per SNP variance (heritability)

\[ \text{Var}(\beta_j) = \sum_{c: j \in C_c} \tau_c \]

\[ = E[\beta_j^2] \text{ (assuming } E[\beta_j] \approx 0) \]

A univariate chi-square (GWAS)

\[ \chi_j^2 = N \hat{\beta}_j^2 \]

\[ E[\chi_j^2] = NE \left( \sum_k \hat{r}_{jk} \beta_k + \epsilon'_j \right)^2 \]

\[ = N \sum_k \hat{r}_{jk}^2 E[\beta_k^2] + NE[\epsilon'_j^2] \]
Idea: Reverse-engineer summary data to find multivar. parameters

**A univariate effect (GWAS)**

\[
\hat{\beta}_j = \frac{1}{N} X_j^T (X \beta + \epsilon) \\
= \sum_k \hat{r}_{jk} \beta_k + \epsilon_j
\]

LD between SNP \(j\) and \(k\)

**A univariate chi-square (GWAS)**

\[
\chi_j^2 = N \hat{\beta}_j^2 \\
\mathbb{E}[\chi_j^2] = N \mathbb{E} \left( \sum_k \hat{r}_{jk} \beta_k + \epsilon_j' \right)^2 \\
= N \sum_k \hat{r}_{jk}^2 \mathbb{E}[\beta_k^2] + N \mathbb{E}[\epsilon_j'^2]
\]

**Per SNP variance (heritability)**

\[
\text{Var}(\beta_j) = \sum_{c:j \in C_c} \tau_c \\
= \mathbb{E}[\beta_j^2] \text{ (assuming } \mathbb{E}[\beta_j] \approx 0) \\
\]

\[
\mathbb{E}[\chi_j^2] = N \sum_c \tau_c \sum_{k \in C_c} \hat{r}_{jk}^2 + \sigma_e^2
\]

Finucane et al. (2015)
Regression of chi-square statistics on LD scores

\[ E[\chi_j^2] = N \sum_c \tau_c \sum_{k \in C_c} \hat{\tau}_{jk}^2 + \sigma_e^2 \]

\[ E[\chi_j^2] = N \sum_c \tau_c \ell(j, c) + 1 \]

\[ \ell(j, c) = \sum_{k \in C_c} r_{jk}^2 \]

Intuition: Remove unwanted “double-counting” of annotation enrichment due to LD

\[ \chi_1^2 \sim \sum_c \tau_c \]

\[ \chi_2^2 \sim \sum_c \tau_c \]

\[ \chi_{p-1}^2 \sim \sum_c \tau_c \]

\[ \chi_p^2 \sim \sum_c \tau_c \]

\[ \sum_c \tau_c \]

\[ p \text{ SNPs} = p \text{ observations} \]

Finucane et al. (2015)
Stratified LDSC partitions heritability of complex trait GWAS summary

Finucane et al. (2015)
Today: Systems Genetics

- Heritability
- Polygenic risk scores
- Disease architecture
- Linear Mixed Models (LMMs)
- Partitioning Heritability
- LD score regression

- Polygenicity
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>220</td>
<td>216</td>
</tr>
<tr>
<td>BIP</td>
<td>Glucocorticoid signaling</td>
<td>217</td>
<td>230</td>
</tr>
<tr>
<td>CAD</td>
<td>Cholesterol/triglyceride metabolism, IgA</td>
<td>248</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
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.

Boyle, Li, Pritchard, Cell, 2017
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).

Boyle, Li, Pritchard, Cell, 2017
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
Core genes vs. periphery

- Omnigenic Model of Complex Traits
  - (A) For any given disease phenotype, a limited number of genes have direct effects on disease risk. However, by the small world property of networks, most expressed genes are only a few steps from the nearest core gene and thus may have non-zero effects on disease. Since core genes only constitute a tiny fraction of all genes, most heritability comes from genes with indirect effects.

- (B) Diseases are generally associated with dysfunction of specific tissues; genetic variants are only relevant if they perturb gene expression (and hence network state) in those tissues. For traits that are mediated through multiple cell types or tissues, the overall effect size of any given SNP would be a weighted average of its effects in each cell type.