Lecture 16: Complex Trait Heritability
<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>PS1 on L1-L5</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></td>
<td>Fri, Sep 8</td>
<td></td>
<td></td>
<td>R1 Recitation: Biology and Probability Review</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Tue, 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></td>
<td>Thu, Sep 14</td>
<td></td>
<td></td>
<td>R2 Recitation 2: Deriving Parameters of Alignment, Multiple Alignment</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Sep 15</td>
<td></td>
<td></td>
<td>R3 Alignment II: Database search, Rapid string matching, BLAST, BLOSUM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Tue, Sep 19</td>
<td>Foundations</td>
<td>L4</td>
<td>Hidden Markov Models Part 1: Evaluation/ Parsing, Viterbi, Forward algorithms</td>
<td>7, 8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Sep 21</td>
<td></td>
<td></td>
<td>L5 Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Sep 22</td>
<td></td>
<td></td>
<td>No classes - student holiday</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mon, Sep 25</td>
<td></td>
<td></td>
<td>Project Intro: about the projects, self introductions, mentor intro, example projects, teamwork 32D-507</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PS2 on L6-R4</td>
<td>4</td>
<td>Thu, Sep 26</td>
<td>Expression Analysis: Clustering/ Classification, K-means, Hierarchical, Bayesian</td>
<td>L6</td>
<td>Expression Analysis: Clustering/ Classification, K-means, Hierarchical, Bayesian</td>
<td>15, 16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Sep 28</td>
<td>Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr</td>
<td>L7</td>
<td>Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr</td>
<td>14, 15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Tue, Oct 3</td>
<td>Foundations</td>
<td>L8</td>
<td>Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Oct 6</td>
<td></td>
<td></td>
<td>L9 Three-dimensional chromatin interactions: 3C, 5C, HIC, CHIA-Pet</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Oct 7</td>
<td></td>
<td></td>
<td>R4 Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Oct 8</td>
<td>Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Tue, Oct 10</td>
<td>Module III: Regulatory Genomics and Networks</td>
<td>L10</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Oct 12</td>
<td></td>
<td></td>
<td>R5 Recitation 5: Gapped Motif Discovery, DNAscope, PBMs, Selex</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Oct 13</td>
<td></td>
<td></td>
<td>R6 Recitation 6: Networks review, Recommendation systems, EHR, PhewAS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>Tue, Oct 17</td>
<td>Frontiers</td>
<td>L11</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></td>
<td>Thu, Oct 19</td>
<td></td>
<td></td>
<td>L12 Deep Learning, Neural Networks, Convolutional NNs, Recurrent NNs, Autoencoder</td>
<td>20, 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Oct 20</td>
<td></td>
<td></td>
<td>R7 Recitation 7: Linkage Disequilibrium, haplotype phasing, Genotype Imputation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Oct 21</td>
<td>Project feedback: Prepare 2-3 slide presentation of your term project for your mentor, 320-507 at 4-5pm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PS4 on L13-R8</td>
<td>8</td>
<td>Tue, Oct 24</td>
<td>Foundations</td>
<td>L13</td>
<td>Population genetics: Linkage disequilibrium, pop struc, 1000genomes, allele freq</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Oct 26</td>
<td></td>
<td></td>
<td>L14 Disease Association Mapping, GWAS, organismal phenotypes</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Thu, Nov 2</td>
<td>Frontiers</td>
<td>L15</td>
<td>Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, IMWAS</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 3</td>
<td></td>
<td></td>
<td>R8 Recitation 8: Rare Variants, eXAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 4</td>
<td>Panel Discussion: reconciling critiques, strategies for improvement, feedback to author 32D-507</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 7</td>
<td>No recitation, Veterans Day</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Thu, Nov 14</td>
<td>Frontiers</td>
<td>L17</td>
<td>Comparative genomics and evolutionary signatures</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 17</td>
<td></td>
<td></td>
<td>L18 Genome Scale Evolution, Genome Duplication</td>
<td>4, 5, 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 18</td>
<td>No recitation, Veterans Day</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No more pssets! work on your final project</td>
<td>12</td>
<td>Thu, Nov 21</td>
<td>Quiz</td>
<td>Quiz</td>
<td>In Class Quiz (the only quiz - the class has no final exam) - covers L1-L20, R1-R9</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Nov 23</td>
<td>No lecture, thanksgiving break - Thu Nov 26, 2015</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Fri, Dec 1</td>
<td>No recitation, thanksgiving break</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Dec 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Dec 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Dec 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>Tue, Dec 12</td>
<td>Frontiers</td>
<td>L21</td>
<td>Single-cell genomics: technology, analysis, microfluidics, applications, insights</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tue, Dec 15</td>
<td></td>
<td></td>
<td>L22 Mining human phenotypes, phewas, UK Biobank, meta-phenotypes+imputation</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Dec 18</td>
<td></td>
<td></td>
<td>R10 Recitation 10: Project Feedback, results, interpretation, directions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Dec 21</td>
<td></td>
<td></td>
<td>R11 Recitation 11: Presentation Tips - Intro, discussion, Slides, Presentation skills</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Dec 22</td>
<td></td>
<td></td>
<td>R12 Recitation 12: Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Dec 23</td>
<td></td>
<td></td>
<td>L24 Genome Engineering with CRISPR/Cas9 and related technologies</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tue, Dec 25</td>
<td>Final Presentations - Part I (11am), 32G-8 reading room</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tue, Dec 25</td>
<td>Final Presentations - Part I (1pm), 32-141</td>
<td></td>
<td></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

| Module IV: Population and Disease Genetics | Foundations | L13 Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq | 30 |
|                                           | L14 Disease Association Mapping, GWAS, organismal phenotypes | 31 |
|                                           | R7 Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation |  |
| Panel Discussion: reconciling critiques, strategies for improvement, feedback to author 32D-507 | Frontiers | L15 Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS | 32 |
|                                           | L16 Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment | 33 |
|                                           | R8 Recitation 8: Rare Variants, ExAC |  |

- **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

3. **From genetic architectures to systems biology:**
   functional enrichments, rank-based enrichment

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)

• Then, \( V[p] = V[g] + V[e] + 2Cov(G,E) \)
  (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 $g = \mathbf{X}\beta$
- We can estimate SNP effect sizes $\beta$ from GWAS
- The variance explained by each SNP depends on effect size and MAF
- $V[\mathbf{X}_j \beta_j] = 2f_j(1-f_j)\beta_j^2$
- 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: 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

- **From genetic architectures to systems biology**: functional enrichments, rank-based enrichment

- **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

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

- 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).
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).
Bayesian variable selection

- 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
Forward simulation of T2D

- 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 $h^2_g$ 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

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

- 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 genotypes
Partitioning heritability by cell type

- LD score regression generalizes easily to multiple components (Finucane 2014)
- Fit different annotations separately (different heritability parameter)
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
- **From genetic architectures to systems biology**: functional enrichments, rank-based enrichment
- **Phenotype prediction**: imputing intermediate phenotypes, large-scale models, inference algorithms
Functional enrichments across 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
Functional enrichments across 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
Functional enrichments across 8 diseases

• 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
Functional enrichments across 8 diseases

<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
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
Functional enrichments across 8 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).
Functional enrichments across 8 diseases

- 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
- **From genetic architectures to systems biology**: functional enrichments, rank-based enrichment
- **Phenotype prediction**: imputing intermediate phenotypes, large-scale models, inference algorithms
**Imputed MWAS: increased power, genetic component**

<table>
<thead>
<tr>
<th>GWAS: G</th>
<th>→</th>
<th>D</th>
<th>N=74k</th>
</tr>
</thead>
<tbody>
<tr>
<td>meQTL: G</td>
<td>→</td>
<td>M</td>
<td>N=800</td>
</tr>
<tr>
<td>MWAS:</td>
<td>M</td>
<td>←</td>
<td>D</td>
</tr>
<tr>
<td>iMWAS: G</td>
<td>→</td>
<td>iM</td>
<td>→</td>
</tr>
</tbody>
</table>

**Learn G→D directly**
- (complex phenotype)
- (simpler phenotype)
- M↔D (no causality)
- Apply G→M to get iM
- iM→D (causality)

**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
iMWAS results: new loci, multiple contributing SNPs

Chromosome 6
iMTWAS: Imputation across multiple intermediate variables

Model multiple mediator variables
SNP → Methylation → Expression → Disease
Predict new loci, increased power
Predict regulatory regions & target genes
Imputation of inaccessible tissues from surrogates (e.g., blood)

- Imputed data surpasses observed data quality
- Incorporate \textit{cis}, \textit{trans}, multi-tissue effects
- Common meta-sample decomposition matrix
- Combine genetic and environmental effects

Ernst and Kellis, Nature Biotechnology, 2015
Goal 2: Dissect molecular and cellular phenotypes

1,000 individuals x 8 tissues / cell types

Whole tissue
Neurons
Astrocytes
Oligodendrocytes
Microglia

200 individuals

GWAS:G → D  N=74k  Learn G→D directly (complex phenotype)
meQTL:G → 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 → iT → D  N=74k  Apply G→M to get iT

Learn G→M (simpler phenotype)
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 genes

Gusev et al. Nature Genetics (2016)

Formal definition of a linear model

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

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

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

\( \mathbf{\theta} \) = effect size (can be itself sampled from a normal prior)
Maximum likelihood estimation w/o prior distribution

\[ f(y|\theta, X) \]

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

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

With isotropic Gaussian noise, we can easily write out log-likelihood of the model as sum of individuals.

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.
MLE $\iff$ convex optimization

Maximum likelihood

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

Convex loss function

$$L(\theta) = (y - X\theta)^T(y - X\theta)$$

Can optimize by taking gradient, set to 0

$$\nabla_{\theta} L = X^T(y - X\theta) = 0$$

Minimum square error

$$\sum_{i=1}^{n} (y_i - x_i \theta)^2.$$
Intrinsic difficulty of MLE

We can gain insights into limitation of ordinary linear regression by looking at its maximum likelihood solution.

1. Each j-th element of the vector \( x_j^T y \)
2. Assuming \( x \) and \( y \) are normalized, this dot product = cosine of the angle between two vectors.
3. SNP by SNP association test relies on this statistic.
4. If \( p \gg n \), prob. of this dot product \( x_j^T y \) significantly deviating from zero increases.
Intractability of MLE for large-scale linear models

High dimensionality (not enough samples / too many parameters)

Colinearity (dependency between variables)

\[ y \sim X \theta \]

\[ n \ll p \quad n \approx 500 \]

\[ p \approx 10,000 \]
Variable selection

Regression analysis = projecting the observed \( \mathbf{y} \) vector on to column space of \( \{ \mathbf{x}_j : j \in [p] \} \),

\[
\begin{pmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
\end{pmatrix}
= \theta_1
\begin{pmatrix}
  X_{11} \\
  X_{21} \\
  \vdots \\
  X_{n1}
\end{pmatrix}
+ \ldots \theta_p
\begin{pmatrix}
  X_{1p} \\
  X_{2p} \\
  \vdots \\
  X_{np}
\end{pmatrix}.
\]

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

- 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

- $|\theta_1| = 0$
- likelihood landscape $P(y|X,\theta)$
- prior $P(||\theta||_1 > c$ or $||\theta||_1 < t$

L2

- $|\theta_1| > 0$
- $P(y|X,\theta)$
- prior $P(||\theta||_2 > c$ or $||\theta||_2 < t$

Other priors

L0.5

L0.1
L1-regularization and Laplace prior

Maximizing \textit{a posteriori}

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

Minimize \textit{L}_1\text{-regularized error}

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

Prior distribution

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

where \( \| \theta \| = \sum p |\theta_j| \).

(a) There is no known closed form solution to this problem.
(b) This is a convex optimization problem.
(c) Numerous algorithms were developed to address this problem over the past 2 decades.
No analytical solution to regression with Laplace prior (L1-regularized)

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

a) L1-regularized (or Laplace prior) regression can select variables with statistical guarantee if variables are not colinear to each other.

\[
\begin{pmatrix}
    y_1 \\
    y_2 \\
    \vdots \\
    y_n
\end{pmatrix} = \theta_1 \begin{pmatrix}
    X_{11} \\
    X_{21} \\
    \vdots \\
    X_{n1}
\end{pmatrix} + \cdots + \theta_p \begin{pmatrix}
    X_{1p} \\
    X_{2p} \\
    \vdots \\
    X_{np}
\end{pmatrix}
\]

b) However, there is no closed form solution of point estimation (MAP).

c) Moreover, there is no closed form solution of posterior probability.

d) In order to identify confidence interval of the point estimates, we can utilize bootstrapping (active research area; conformal inference).

LARS, Efron et al. (2002)
L2-regularization and Gaussian prior

Maximizing a posteriori

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

Minimize \( L_2 \)-regularized error

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

Prior distribution

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

\[
\| \theta \|^2 = \sum_{j \in [p]} \theta_j^2.
\]

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

(See next slides for derivation).
Analytical derivation of Bayesian regression with Gaussian prior

Log-likelihood + log prior:

\[
\ln p(\theta|y, X) = - \frac{1}{2\sigma^2} (y - X\theta)^T(y - X\theta) - \frac{\lambda}{2} \theta^T\theta + \text{const.}
\]

By taking derivative with respect to \(\theta\) and setting it the zero vector:

\[
\nabla_\theta = - \frac{1}{\sigma^2} X^T(y - X\theta) - \lambda \theta = 0
\]

\[
\nabla^2_\theta = X^TX\theta - \lambda I
\]

Rearranging the equation:

\[
X^Ty = (X^TX + \lambda \sigma^2 I)\theta \implies \hat{\theta} = (X^TX + \lambda \sigma^2 I)^{-1}X^Ty
\]

Approximately, \(p(\theta|y, X, \lambda) \approx \mathcal{N}(\hat{\theta}, (X^TX + \lambda \sigma^2 I)^{-1}).\)

(In Gaussian, 2\textsuperscript{nd} order derivative \(\approx\) inverse of variance)
Comparison between MLE and Bayesian regression

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

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

(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).

Extension: spike-slab prior on $\theta$

$$p(\theta|z=1) \sim \mathcal{N}(0, 1/\tau)$$

Fat Gaussian for true effects (slab; magnitude and direction)

$$p(\theta|z=0) = \delta(\theta)$$

Completely set to zero if not selected

$z = 1 \sim \text{Bernoulli}(\pi)$

$\pi$ determines prior prob. of including variables (usually < .1; spike; prescribed or optimized)

$$p(\theta) \sim \exp(-\lambda|\theta|)$$

Figure: Hernandez-Lobato (2014)
Spike-slab prior model effectively avoid colinearity

Simulated model:
\[ y \sim X_1 \theta_1 \]
\[ X_2 \sim X_1 \gamma \]

OLS model:
\[ y \sim X_1 \theta_1 + X_2 \theta_2 \]

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).

MLE is overfitting

True effect locates little deeper in likelihood contour

- Rockova & George, *Metron* (2014)
Spike-slab prior fundamental to success of iMWAS
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
- **From genetic architectures to systems biology:** functional enrichments, rank-based enrichment
- **Phenotype prediction:** imputing intermediate phenotypes, large-scale models, inference algorithms