Lecture 17 - Systems genetics: Deep Learning, eQTLs, Polygenicity, Heritability, LMMs, LDSC, PRS, Networks

Prof. Manolis Kellis

Slides credit: Yongjin Park, Abhishek Sarkar, Mark Daly, David Gifford, et al
1. Review: GWAS, fine-mapping, Bayesian variant prioritization
2. Deep Learning for GWAS: calling SNPs, prioritize function
3. eQTLs/Mediation: intermediate molecular phenotypes
4. Linear Mixed Models (LMMs) for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): summing over many variants
6. Heritability: definition(s), missing heritability, partitioning
7. LD Score regression (LDSC) for fast heritability partitioning
8. Polygenic/Omnigenic disease models: core vs. periphery
9. Disease gene networks from GWAS evidence boosting
1. **Review**: GWAS, fine-mapping, Bayesian methods for variant prioritization
Monogenic vs. oligogenic vs. polygenic disorders

- **Linkage analysis**: Few variants of large effects
- **Combination of large/small effects**: Prevalence of the disease
- **GWAS**: Many variants of small effects

- **Single Gene Disorders**: Mostly coding
- **Oligogenic Disorders**: Number and effects sizes of determining alleles
- **Polygenic Disorders**: Mostly non-coding

Molecular genetic studies of complex phenotypes. Marian - Translational Research, 2012
Common variants (SNPs) live in Haplotypes

- Common SNPs only once every 1000 nucleotides or so
- These are co-inherited, so only need to profile a subset
- Markers selected for haplotype profiling are “tag” SNPs
Dissecting non-coding genetic associations

1. Establish relevant tissue/cell type(s)
2. Establish downstream target gene(s)
3. Establishing causal nucleotide variant
4. Establish upstream regulator causality
5. Establish cellular phenotypic consequences
6. Establish organismal phenotypic consequences

Goal:
Apply these to the FTO locus in obesity
Manipulate circuitry $\rightarrow$ reverse disease phenotypes

Thermogenic stimuli (e.g. cold)

- ARID5B
- IRX3
- IRX5

**Incr. ARID5B $\rightarrow$ Lean**

**Decr ARID5B $\rightarrow$ Obese**

C-to-T $\rightarrow$ Lean

T-to-C $\rightarrow$ Obese

Decrease IRX3, IRX5 $\rightarrow$ Lean

Increase IRX3, IRX5 $\rightarrow$ Obese

CRISPR-edit human fat cells $\rightarrow$ able to burn calories again

IRX3 KD $\rightarrow$ Burn calories in their sleep $\rightarrow$ 54% weight loss. Can’t gain weight
Dissect mechanisms of disease-associated regions

1. Disease genetics reveals common + rare variants/regions

2. Profile RNA + Epigenome in healthy + disease samples

3. Integrate data to predict driver genes, regions, cell types

4. Validate predictions in human cells + mouse models

5. Disseminate results
Disease hits in enhancers of relevant cell types

<table>
<thead>
<tr>
<th>Trait</th>
<th>Abbrev</th>
<th>-logF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>ESC</td>
<td>4.7</td>
</tr>
<tr>
<td>Height</td>
<td>ESC</td>
<td>4.0</td>
</tr>
<tr>
<td>Chronic lymphocytic leukaemia</td>
<td>Tcor</td>
<td>4.9</td>
</tr>
<tr>
<td>Type 1 diabetes autoantibodies</td>
<td>Treg</td>
<td>4.6</td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>Treg</td>
<td>4.1</td>
</tr>
<tr>
<td>Platelet counts</td>
<td>Th.tcm</td>
<td>4.8</td>
</tr>
<tr>
<td>Chronic lymphocytic leukaemia</td>
<td>Th.tcm</td>
<td>4.9</td>
</tr>
<tr>
<td>Self-reported allergy</td>
<td>Th.tcm</td>
<td>4.3</td>
</tr>
<tr>
<td>Chronic disease</td>
<td>Th.17s</td>
<td>6.9</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>Th.17s</td>
<td>4.2</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>Th.mm</td>
<td>1.16</td>
</tr>
<tr>
<td>Celiac disease + rheumat. arthritis</td>
<td>Th.mm</td>
<td>5.5</td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>Th.mm</td>
<td>5.5</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>Th.mm</td>
<td>4.8</td>
</tr>
<tr>
<td>Primary biliary cirrhosis</td>
<td>Beor</td>
<td>3.3</td>
</tr>
<tr>
<td>Red blood cell traits</td>
<td>HScmb</td>
<td>5.9</td>
</tr>
<tr>
<td>Platelet counts</td>
<td>HScmb</td>
<td>8.0</td>
</tr>
<tr>
<td>Mean platelet volume</td>
<td>HScmb</td>
<td>5.0</td>
</tr>
<tr>
<td>Mean platelet volume</td>
<td>HScmb</td>
<td>3.9</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>Bper</td>
<td>8.5</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>Bper</td>
<td>4.7</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>Nkper</td>
<td>6.0</td>
</tr>
<tr>
<td>Mean platelet volume</td>
<td>Fat</td>
<td>4.2</td>
</tr>
<tr>
<td>HDL cholesterol</td>
<td>Fat</td>
<td>4.9</td>
</tr>
<tr>
<td>Height</td>
<td>Plate</td>
<td>4.8</td>
</tr>
<tr>
<td>Multiple myeloma</td>
<td>Thym</td>
<td>4.2</td>
</tr>
<tr>
<td>Adiponectin levels</td>
<td>Brain</td>
<td>4.3</td>
</tr>
<tr>
<td>Attention deficit hyperact. disord.</td>
<td>Brain</td>
<td>4.5</td>
</tr>
<tr>
<td>Ph interval</td>
<td>Heart</td>
<td>4.7</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>Heart</td>
<td>4.5</td>
</tr>
<tr>
<td>Aortic root size</td>
<td>Vasec</td>
<td>4.1</td>
</tr>
<tr>
<td>Pulmonary function</td>
<td>Smmuc</td>
<td>4.2</td>
</tr>
<tr>
<td>Liver enzyme levels (g-glut 1x)</td>
<td>Gl.int</td>
<td>4.9</td>
</tr>
<tr>
<td>Urate levels</td>
<td>Gl.int</td>
<td>4.5</td>
</tr>
<tr>
<td>Adv. resp. to chem.(neut/leuc)</td>
<td>Gl.muc</td>
<td>4.0</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>Stoma</td>
<td>4.3</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>Stoma</td>
<td>4.3</td>
</tr>
<tr>
<td>Insulin-like growth factors</td>
<td>Placnt</td>
<td>4.2</td>
</tr>
<tr>
<td>Renal disease-related traits</td>
<td>Kidne</td>
<td>3.1</td>
</tr>
<tr>
<td>LDL cholesterol</td>
<td>Liver</td>
<td>10.1</td>
</tr>
<tr>
<td>Cholesterol, total</td>
<td>Liver</td>
<td>9.0</td>
</tr>
<tr>
<td>Cholesterol, total</td>
<td>Liver</td>
<td>7.1</td>
</tr>
<tr>
<td>LDL cholesterol</td>
<td>Liver</td>
<td>6.8</td>
</tr>
<tr>
<td>Lipid metabolism phenotypes</td>
<td>Liver</td>
<td>5.8</td>
</tr>
<tr>
<td>HDL cholesterol</td>
<td>Liver</td>
<td>5.7</td>
</tr>
<tr>
<td>Cholesterol, total</td>
<td>Liver</td>
<td>4.8</td>
</tr>
<tr>
<td>HDL cholesterol</td>
<td>Liver</td>
<td>3.0</td>
</tr>
<tr>
<td>Metabolite levels</td>
<td>Liver</td>
<td>3.9</td>
</tr>
<tr>
<td>Platelet counts</td>
<td>Leuk</td>
<td>4.5</td>
</tr>
<tr>
<td>Primary biliary cirrhosis</td>
<td>Leuk</td>
<td>4.7</td>
</tr>
<tr>
<td>Mean corpuscular volume</td>
<td>Leuk</td>
<td>4.7</td>
</tr>
<tr>
<td>Inflammatory bowel disease</td>
<td>Mncyt</td>
<td>4.6</td>
</tr>
<tr>
<td>Pre-eclampsian</td>
<td>Mncyt</td>
<td>4.6</td>
</tr>
<tr>
<td>Alzheimer’s disease (late onset)</td>
<td>Bone</td>
<td>4.5</td>
</tr>
</tbody>
</table>
Scale up: 834 tissue/cell types $\rightarrow$ 30k GWAS SNPs in 534 traits

127 Epigenomes (Roadmap 2015) $\downarrow$ 834 Epigenomes (EpiMap 2019)

54 enriched GWAS traits (2015)

534 enriched traits

30,247 SNPs in enriched enhancers $\rightarrow$ Highly-specific associations Emerge $\rightarrow$ Precise biological hypotheses on mechanistic basis

Tissue enrich/co-enrichments $\rightarrow$ trait clustering, trait-tissue network
Partitioning complex traits across multiple tissues of action
Bayesian fine-mapping: Predict causal variant and cell type

RiVIERA: multi-trait GWAS integration

Capture conserved elements

Predict causal variants and cell types

Capture eQTLs from GTEx
Combine GWAS+Epig to find new target genes/SNPs

Prioritize sub-threshold loci (<10⁻⁴)

Validate new enhancers: allelic activity, enh-prom looping

Machine learning predictive features

Validate new genes in hum/mou/zb
Today: Deep Learning for Human Genetics and Disease

1. Review: GWAS, fine-mapping, Bayesian variant prioritization
2. Deep Learning for GWAS: calling SNPs, prioritize function
3. eQTLs/Mediation: intermediate molecular phenotypes
4. Linear Mixed Models (LMMs) for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): summing over many variants
6. Heritability: definition(s), missing heritability, partitioning
7. LD SCore regression (LDSC) for fast heritability partitioning
8. Polygenic/Omnigenic disease models: core vs. periphery
9. Disease gene networks from GWAS evidence boosting
2. Deep Learning methods for GWAS
Calling variants, prioritizing functional SNPs
CADD: combine evidence to predict variant function

A

Phylogenetic Tree

- conservation
  - phastCons, phyloP, GERP
- epigenetic modification
  - H3K4Me1, H3K9Ac, DNase-Seq
- functional prediction
  - amino acid change, TF motif disruption
- genetic context
  - GC content, CpG content, base transversion, InDel length

Multiple Species Alignment

generate

derive

changes between human and inferred ancestor

use allele composition to

simulate variants

proxy-neutral variants

train logistic regression model

proxy-deleterious variants

B

User provided variants

annotate

score variants

PHRED scaling

use model to

CADD score

AUROC - ClinVar pathogenic vs frequent ExAC (MAF>0.05)
31,815 vs. 69,894 SNV

AUROC - ClinVar pathogenic vs frequent ExAC (MAF>0.05)
1288 missense SNV each matched by gene

CADD: predicting the deleteriousness of variants throughout the human genome

Philipp Rentzsch, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher
# Large number of methods for variant prioritization

<table>
<thead>
<tr>
<th>Method</th>
<th>Data sources</th>
<th>Approach</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigen</td>
<td>- Uses data from the ENCODE and Roadmap Epigenomics projects</td>
<td>- Weighted linear combination of individual annotations</td>
<td>(14)</td>
</tr>
<tr>
<td></td>
<td>- Enhancer–gene linkage</td>
<td>- Unsupervised learning method</td>
<td></td>
</tr>
<tr>
<td>FunSeq2</td>
<td>- Inter- and Intra-species conservation</td>
<td>- Weighted scoring system</td>
<td>(15)</td>
</tr>
<tr>
<td></td>
<td>- Loss- and gain-of-function events for transcription factor binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LINSIGHT</td>
<td>- Conservation scores (phastCons, phylop), predicted binding sites (TFBS, RNA), regional annotations (ChIP-seq, RNA-seq)</td>
<td>- Graphical model</td>
<td>(16)</td>
</tr>
<tr>
<td></td>
<td>- GC content, CpG content, histone methylation</td>
<td>- Selection parameter fitting using generalized linear model based on 48 genomic features</td>
<td></td>
</tr>
<tr>
<td>CADD</td>
<td>- Ensembl variant effect predictor</td>
<td>- Support vector machine</td>
<td>(11)</td>
</tr>
<tr>
<td></td>
<td>- Protein-level scores: Grantham, SIFT, PolyPhen</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- DNase hypersensitivity, TFBS, transcript information</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- GC content, CpG content, histone methylation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FATHMM</td>
<td>- 46-way sequence conservation</td>
<td>- Hidden Markov models</td>
<td>(17)</td>
</tr>
<tr>
<td></td>
<td>- ChIP-seq, TFBS, DNase-seq</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- FAIRE, footprints, GC content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReMM</td>
<td>- Predict potential of non-coding variant to cause a Mendelian disease if mutated</td>
<td>- Random forest classifier</td>
<td>(18)</td>
</tr>
<tr>
<td></td>
<td>- 26 features: PhastCons, PhyloP, CpG, GC, regulation annotations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orion</td>
<td>- Predict potential of non-coding variant to cause a Mendelian disease if mutated</td>
<td>- Expected and observed site-frequency spectrum of a given stretch of sequence</td>
<td>(19)</td>
</tr>
<tr>
<td></td>
<td>- Independent from annotation and features</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Expected and observed site-frequency spectrum of a given heptamer</td>
<td></td>
<td>(8)</td>
</tr>
<tr>
<td>CDTS</td>
<td>- Identify constrained non-coding regions in the human genome and deleteriousness of variants</td>
<td>- Expected and observed site-frequency spectrum of a given heptamer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Independent from annotation and features. Uses k-mers</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Whole genome variant calling: GATK HaplotypeCaller

1. Use heuristic to find mismatches not explained by noise
2. Use assembly graph to identify possible haplotypes
3. For each haplotype, estimate: $P(\text{read} \mid \text{haplotype})$ using probabilistic sequence alignment
   - Hidden Markov Model
   - States: insertion, deletion, substitution
   - Emissions: pairs of aligned nucleotides/gaps
   - Transitions: equivalent to insertion/deletion/gap penalties from Smith-Waterman algorithm (DP alignment)
   - Get $P(\text{read} \mid \text{haplotype})$ using forward-backward algorithm
4. Use Bayes rule to get $P(\text{haplotype} \mid \text{read})$
5. Assign genotypes to each sample based on the max a posteriori haplotypes

Tour de Force, combining many methods:
- Logistic regression to model base errors
- Hidden Markov models to compute read likelihoods
- Naive Bayes classification to identify variants
- Gaussian mixture model with hand-crafted features to filter likely false positive variants, capturing common error modes

http://gatkforums.broadinstitute.org/gatk/discussion/4148/hc-overview-how-the-haplotypecaller-works
Exome variant calling: atlas2

Motivation: the exome has different sequence properties than the rest of the genome (e.g., substitution rates, GC content).

Train logistic regression classifier to predict which mismatches are errors and which are variants:
- Training data: 1KG Exome project sequencing reads where >2 reads align with a mismatch
- True positives: Reads where mismatch is also discovered in 1KG Exon pilot project
- True negatives: Remaining reads
- Features: mismatch quality score, flanking quality score, whether neighboring nucleotides were swapped, normalized distance to 3’ end of the read

Much faster than full Bayesian model (e.g. HaplotypeCaller), lower false positive rate in validation data

Bamshad et al. Nat Rev Genet 2011
DeepVariant: Combine evidence to call variants

A universal SNP and small-indel variant caller using deep neural networks

Ryan Poplin$^{1,2}$, Pi-Chuan Chang$^3$, David Alexander$^2$, Scott Schwartz$^2$, Thomas Colthurst$^2$, Alexander Ku$^2$, Dan Newburger$^1$, Jojo Dijamco$^1$, Nam Nguyen$^1$, Pegah T Afshar$^1$, Sam S Gross$^1$, Lizzie Dorfman$^{1,2}$, Cory Y McLean$^{1,2}$ & Mark A DePristo$^{1,2}$
Predicting disease mutations

1. High-throughput experiments
   - PBM
   - SELEX
   - ChIP/CLIP

2. Massively parallel deep learning
   - Automatic model training
   - New motifs
   - Prediction network
   - DeepBind models
   - Large-scale data sets
   - GPU server

3. Community needs
   - Gene regulation
   - Precision medicine
   - Detect binding sites

[Alipanahi et al., 2015]
DeepBind summary

The key deep learning techniques:
- Convolutional learning
- Representational learning
- Back-propagation and stochastic gradient
- Regularization and dropout
- Parallel GPU computing especially useful for hyperparameter search

Limitations in DeepBind:
- Require defining negative training examples, which is often arbitrary
- Using observed mutation data only as post-hoc evaluation
- Modeling each regulatory dataset separately
DeepSea:

- Similar as DeepBind but trained a separate CNN on each of the ENCODE/Roadmap Epigenomic chromatin profiles 919 chromatin features (125 DNase features, 690 TF features, 104 histone features).

- It uses the $\Delta s$ mutation score as input to train a linear logistic regression to predict GWAS and eQTL SNPs defined from the GRASP database with a P-value cutoff of 1E-10 and GWAS SNPs from the NHGRI GWAS Catalog

[Zhou and Troyanskaya, 2015]
CNNs for DNA-binding prediction from sequence

DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Uses convolution layers to capture regulatory motifs, and a recurrent layer to discover a ‘grammar’ for how these single motifs work together. Based on Keras/Theano.

Basset—learning the regulatory code of the accessible genome with deep convolutional neural networks. CNN to discover regulatory sequence motifs to predict the accessibility of chromatin. Accounts for cell-type specificity using multi-task learning.

DeepBind and DeeperBind—predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Based on ChIP-seq, ChIP-chip, RIP-seq, protein-binding microarrays and others. Deeperbind adds a recurrent sequence learning module (LSTM) after the convolutional layer(s).


Convolutional neural network architectures for predicting DNA–protein binding. Systematic exploration of CNN architectures for predicting DNA sequence binding using a large compendium of transcription factor data sets.

Predicting enhancers, 3d interactions and cis-regulatory regions

PEDLA: predicting enhancers with a deep-learning-based algorithmic framework. Predicting enhancers based on heterogeneous features from (e.g.) the ENCODE project using a deep learning, HMM hybrid model.

DEEP: a general computational framework for predicting enhancers. Predicting enhancers based on data from the ENCODE project.

Genome-wide prediction of cis-regulatory regions using supervised deep-learning methods. toolkit based on the Theano) for applying different deep-learning architectures to cis-regulatory elements.


DNA methylation


Variant callers, pathogenicity scores and identification of genomic elements

DeepVariant—a variant caller in germline genomes. Uses a deep neural network architecture (Inception-v3) to identify SNP and small indel variants from next-generation DNA sequencing data.


evoNet—deep learning for population genetic inference [code][paper]. Jointly inferring natural selection and demographic history

DANN. Uses the same feature set and training data as CADD to train a deep neural network

DeepSEA—predicting effects of non-coding variants with deep-learning-based sequence model. Models chromatin accessibility as well as the binding of transcription factors, and histone marks associated with changes in accessibility.
Today: Deep Learning for Human Genetics and Disease

1. Review: GWAS, fine-mapping, Bayesian variant prioritization
2. Deep Learning for GWAS: calling SNPs, prioritize function
3. eQTLs/Mediation: intermediate molecular phenotypes
4. Linear Mixed Models (LMMs) for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): summing over many variants
6. Heritability: definition(s), missing heritability, partitioning
7. LD Score regression (LDSC) for fast heritability partitioning
8. Polygenic/Omnigenic disease models: core vs. periphery
9. Disease gene networks from GWAS evidence boosting
3. eQTLs and Mediation Analysis
Intermediate molecular phenotypes to disease
Genetic Variant

CATGACTG

CATGCCTG

Tissue/cell type

Heart

Brain

Cortex

Lung

Blood

Skin

Nerve

Molecular Phenotypes

Epigenetic Changes

Methyl.

DNA access.

Enhancer

H3K27ac

Promoter

Insulator

Gene Expression Changes

Gene expr.

Gene expr.

Gene expr.

Gene expr.

Gene expr.

Organismal phenotypes

Endo phenotypes

Lipids

Tension

Amyloidβ

Metabol.

Drug resp

Environment

Feedback from environment / disease state

Disease
Methylation in 750 Alzheimer patients/controls

- **ROS-MAP cohort (RUSH: David Bennett, HMS: Phil De Jager)**
  - Patients followed for 10+ years with cognitive evaluations
  - Brain samples donated post-mortem methylation/genotype

- Seek predictive features: SNPs, QTLs, mQTLs, regulation

Methylation variation in 723 individuals

Relate to genotype and AD variation

**Diagram:**
- Genome → Epigenome → Phenotype
- 1. Classification MWAS
- 2. meQTL
- 3. Evaluate causality
50,000 significant meQTLs after Bonferroni

- Strong effects across entire range of discovery values
**Imputed MWAS: increased power, genetic component**

<table>
<thead>
<tr>
<th>GWAS:</th>
<th>G → D</th>
<th>N=74k</th>
<th>Learn G→D directly (complex phenotype)</th>
</tr>
</thead>
<tbody>
<tr>
<td>meQTL:</td>
<td>G → M</td>
<td>N=800</td>
<td>Learn G→M (simpler phenotype)</td>
</tr>
<tr>
<td>MWAS:</td>
<td>M ↔ D</td>
<td>N=800</td>
<td>M↔D (no causality)</td>
</tr>
<tr>
<td>iMWAS:</td>
<td>G → iM → D</td>
<td>N=74k</td>
<td>Apply G→M to get iM iM→D (causality)</td>
</tr>
</tbody>
</table>

**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
iMTWAS: Imputation across multiple intermediate variables

Model multiple mediator variables
SNP → Methylation → Expression → Disease
Predict new loci, increased power
Predict regulatory regions & target genes
CaMMEL: 206 significant mediating genes in AD

Small expression change (short), large variance explained (big circle)

Large expression change (tall), little variance explained (small circle)

Higher-expression in AD (risk increasing)

Lower-expression in AD (protective)

Genome-wide significant locus (purple)

Sub-threshold locus (grey)
The nuts and bolts of an eQTL study

**Cell isolation**

**RNA isolation**

**Expression measurement**

**Filter transcripts**

**Subjects**

**Genes**

**DNA**

**Millions of SNP Genotyping QC**

**Linear Regression**

**Expression = genotype + covariates**

**Annotation Visualization Interpretation**

**Determine significance threshold**

**Age, gender Pop stratification Technical Covs**
Expanded eQTL models

\[ Y_{ij} = \alpha + \beta_{ijs} \text{genotype} + \varepsilon \]

\[ Y_{ij} = \alpha + \beta_{1ijs} \text{genotype} + \beta_{2i} \text{gender} + \beta_{3i} \text{age} + \]

\[ + \beta_{4i} \text{gPC1} + \beta_{5i} \text{gPC2} + \beta_{6i} \text{gPC3} + \beta_{7i} \text{gPC4} + \]

\[ + \beta_{8i} \text{ePC1} + \beta_{9i} \text{ePC2} + \beta_{10i} \text{ePC3} + \beta_{11i} \text{ePC4} + \]

\[ + \beta_{12i} \text{ePC5} + \beta_{13i} \text{ePC6} + \beta_{14i} \text{ePC7} + \varepsilon \]
Allelic analysis complements eQTLs

Distinguish reads within the same heterozygous individual
**Combined Haplotype Test**

(Bryce van de Geijn, et.al *Nature Method* 2015)

Maximize likelihood of two observed components:

\[
L(\alpha_h, \beta_h, \phi_j \mid D) = \prod_i \left[ \Pr_{\text{BNB}}(X = x_{ij} \mid \lambda_{hi}, \Omega_i, \phi_j) \prod_k \Pr_{\text{BB-mix}}(Y = y_{ik} \mid p_h, n_{ik}, \gamma_i) \right]
\]

- **Total Read-depth**

- **Allelic imbalance**
“Response eQTLs”: Trait-conditional eQTLs
Today: Deep Learning for Human Genetics and Disease

1. Review: GWAS, fine-mapping, Bayesian variant prioritization
2. Deep Learning for GWAS: calling SNPs, prioritize function
3. eQTLs/Mediation: intermediate molecular phenotypes
4. Linear Mixed Models (LMMs) for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): summing over many variants
6. Heritability: definition(s), missing heritability, partitioning
7. LD SCore regression (LDSC) for fast heritability partitioning
8. Polygenic/Omnigenic disease models: core vs. periphery
9. Disease gene networks from GWAS evidence boosting
4. Linear Mixed Models (LMMs) for GWAS and for eQTL calling
Formal definition of a linear model

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

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

\( \theta = \) 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
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 ≈ 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

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): \( h_{GWAS}^2 = 0.16, h^2 = 0.73, h_g^2 = 0.5 \)
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).
Imputation-based association

1 = learn eQTLs in reference panel

2 = impute expression for each person in a genotyped cohort

3 = use summary statistics to get to associations directly

Gusev et al. “Integrative approaches for large-scale transcriptome-wide association studies” 2016 Nature Genetics
Bayesian linear regression for eQTL modeling

\[ u = \text{magnitude \\& direction of effect size} \]

\[ s = \text{SNP dosage} \]

\[ b = \text{binary indicator of including QTL genes} \]

\[ f = \text{known factors} \]
\[ v = \text{factor loading} \]
\[ \alpha = \text{prior variance} \]

\[ x = \text{hidden factors} \]
\[ w = \text{factor loading} \]
\[ \beta = \text{prior variance} \]

\[ \alpha_c, \beta_k \]
Bayesian extension to ordinary regression models

1. Spike-slab prior to select relevant variables
2. Random effect models
3. Bayesian sparse linear mixed effect model
4. Fine mapping causal variants in LD correlation
Extension 1: 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)

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

$\lambda$ determines slab width of magnitude and direction

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

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

Fitted model:
\[ y \sim X_1 \theta_1 + X_2 \theta_2 \]
\[ \theta_j \sim \text{spike-slab} \]

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

MLE is overfitting.

True effect locates little deeper in likelihood contour.

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).

Rockova & George, *Metron* (2014)
Additive effect of random vector $u$ ($n \times 1$):

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

The random effect captures population structure $K$ (kinship matrix):

$$u \sim \mathcal{N}(0, \tau^2 K)$$

Integrate out uncertain random effect $u$:

$$\int p(y|X, \theta, u)p(u|\tau, K)du = \mathcal{N}(y|X\theta, \tau^2 K + \sigma^2 I)$$

Linear Gaussian model with two variance components.

Inflated statistics due to unknown population structure (almost all loci are significant)

Adjusted GWAS qq-plot with correct structure

Linear mixed-effect calibrated the null distrib.

LMM can correctly capture significant ones.

Extension 3: Bayesian sparse linear mixed effect model

Random effect

\[ y = X\theta + u + \epsilon, \]
\[ u \sim \mathcal{N}(0, K), \]

A sort of spike-slab (two mixture model)

\[ \theta_j \sim \pi \mathcal{N}(0, \tau_1^2) + (1 - \pi) \mathcal{N}(0, \tau_2^2) \]

causal effect
infinitesimal background effect

Extension 4: Fine-mapping causal variants

Hormozdiari et al. (2014)
Extension 4: Fine-mapping under the hood

\[ \mathbf{z} \approx \mathbf{X}^\top \mathbf{y} / \sqrt{n\sigma} \]

We assume phenotype vector were generated by

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

Therefore \( p \times 1 \) vector follows

\[ \mathbf{z} \sim \mathcal{N} \left( \frac{\mathbf{X}^\top \mathbf{X} \boldsymbol{\theta}}{\sqrt{n\sigma}}, \frac{\mathbf{X}^\top \mathbf{X}}{n} \right) \approx \mathcal{N}(\lambda \mathbf{R} \boldsymbol{\theta}, \mathbf{R}). \]

where LD matrix \( \mathbf{R} = n^{-1} \mathbf{X}^\top \mathbf{X} \) and \( \lambda = (n\sigma^2)^{-1/2} \) absorbs all scaling factors.

(a) Considering potential colinearity embedded in the \( \mathbf{R} \) matrix, \( \boldsymbol{\theta} \) desperately needs spike-slab prior.

(b) For computational efficiency, previously developed algorithms restrict number of causal variants (e.g., at most 3).

Hormozdiari et al. (2014)
## Bayesian inference algorithms

<table>
<thead>
<tr>
<th></th>
<th>Exact inference</th>
<th>Markov Chain Monte Carlo</th>
<th>Variational Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>correct</td>
<td>approximate, stochastic</td>
<td>approximate, deterministic</td>
</tr>
<tr>
<td><strong>Convergence</strong></td>
<td>sure</td>
<td>Global optima at equilibrium</td>
<td>Local optima in finite time</td>
</tr>
<tr>
<td><strong>Flexibility</strong></td>
<td>very limited</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td><strong>Examples</strong></td>
<td>HMM’s forward-backward, Dynamic programming</td>
<td>Importance sampling, Metropolis-Hastings, Gibbs, Hamiltonian MC, Elliptical slice sampling</td>
<td>Laplace, Mean-field approx., Belief propagation, Expectation propagation</td>
</tr>
</tbody>
</table>
Today: Deep Learning for Human Genetics and Disease

1. Review: GWAS, fine-mapping, Bayesian variant prioritization
2. Deep Learning for GWAS: calling SNPs, prioritize function
3. eQTLs/Mediation: intermediate molecular phenotypes
4. Linear Mixed Models (LMMs) for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): summing over many variants
6. Heritability: definition(s), missing heritability, partitioning
7. LD Score regression (LDSC) for fast heritability partitioning
8. Polygenic/Omnigenic disease models: core vs. periphery
9. Disease gene networks from GWAS evidence boosting
5. Polygenic Risk Scores (PRS):
Summing over all variants (and more)
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)

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

How do we know the selected SNPs are good?

Goal: Tuning this parameter

How do we know the selected SNPs are good?

Observed risk vs. 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.

Chun .. Sunyeav, BioRxiv (2019)
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
Phasing diploid genomes is hard

- Humans are diploid organisms
- Each individual carries two homologous copies of each chromosome
- Therefore, they carry two copies of each variant (called the maternal/paternal allele)
- Variants co-occur in haplotypes which are inherited as a unit
- Experimentally possible, but currently infeasible, to directly measure haplotypes over the whole genome
- Cheaper and more efficient to measure genotypes (counts of minor allele)
- Genotyping loses information, which we need algorithms and statistical models to recover (phasing, imputation)

<table>
<thead>
<tr>
<th>Haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 0 1 1 0 (maternal)</td>
</tr>
<tr>
<td>0 1 1 0 0 1 0 (paternal)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 0 1 2 0</td>
</tr>
</tbody>
</table>
Molecular diagnostics in IBD

‘Molecular’ diagnosis (based on GWAS SNPs & serologic biomarkers) concordant with GI dx: CD & UC patients can be distinguished accurately

>90% of patients correctly classified with >90% reliability

Jonah Essers (MGH/CHB), Dermot McGovern (CSMC)
Molecular diagnostics flag patients with worst outcome

Black dots represent patients diagnosed with UC who later underwent colectomy and then developed full-blown Crohn’s disease.
Today: Deep Learning for Human Genetics and Disease

1. Review: GWAS, fine-mapping, Bayesian variant prioritization
2. Deep Learning for GWAS: calling SNPs, prioritize function
3. eQTLs/Mediation: intermediate molecular phenotypes
4. Linear Mixed Models (LMMs) for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): summing over many variants
6. Heritability: definition(s), missing heritability, partitioning
7. LD Score regression (LDSC) for fast heritability partitioning
8. Polygenic/Omnigenic disease models: core vs. periphery
9. Disease gene networks from GWAS evidence boosting
6. Heritability: Definition, Missing Heritability, Partitioning
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, $V[p] = V[g] + V[e] + 2\text{Cov}(G,E)$
  (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
  
  \[ V[X_j \beta_j] = 2 f_j (1 - f_j) \beta_j^2 \]
• If we do this with genome-wide significant SNPs, we usually $h^2_{GWAS} < h^2$
• 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.
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

- 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
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
Today: Deep Learning for Human Genetics and Disease

1. Review: GWAS, fine-mapping, Bayesian variant prioritization
2. Deep Learning for GWAS: calling SNPs, prioritize function
3. eQTLs/Mediation: intermediate molecular phenotypes
4. Linear Mixed Models (LMMs) for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): summing over many variants
6. Heritability: definition(s), missing heritability, partitioning
7. LD Score regression (LDSC) for fast heritability partitioning
8. Polygenic/Omnigenic disease models: core vs. periphery
9. Disease gene networks from GWAS evidence boosting
7. LD SCore regression (LDSC):
Computing and partitioning* heritability quickly
(* with stratified LD SCore regression)
LD SCore regression (LDSC)

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

- 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
Intuition: LD score $\leftrightarrow$ heritability

Under pure drift, LD is uncorrelated to magnitude of allele frequency differences between populations.

Assuming i.i.d. (standardized) effect sizes, more LD yields higher chi-square (on average) 
More tags $\Rightarrow$ more causal SNPs. 
More shots $\Rightarrow$ more shots on goal

Simulation under stratification
- $\lambda_{GC} = 1.30$; LD Score Regression intercept = 1.32

Simulation under association
- $\lambda_{GC} = 1.30$; LD Score Regression intercept = 1.02
**Linkage disequilibrium: D and D’**

- Genetic variants do not segregate independently
- \( D = \) coeff. of linkage disequilibrium between alleles A and B at loci L1 and L2
  - \( D_{AB} = P_{11}P_{00} - P_{10}P_{01} = 0.07 \)
  - Property of the specific **alleles**. Different alleles at these loci will have diff \( D_{AB} \)
- If independent, then \( D_{AB} = 0 \) \((P_{11}P_{00} = P_{10}P_{01})\)
- Linkage disequilibrium measures the degree of departure from Mendel’s laws of independent assortment

**How to interpret actual values?**

- Relative to \( D_{AB_{max}} \), which depends on frequencies of individual alleles at A, B
  - \( D_{AB_{max}} = P_{0}\ast P_{1}\ast P_{1}\ast P_{0} = 0.138 \)
  - \( D' = D/D_{max} = 0.51 \)
  - \( \Rightarrow 51\% \) of max possible disequilibrium

<table>
<thead>
<tr>
<th>Haplotype AB</th>
<th>Marginal allele frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0*</td>
<td>0.54</td>
</tr>
<tr>
<td>1*</td>
<td>0.46</td>
</tr>
<tr>
<td>*0</td>
<td>0.30</td>
</tr>
<tr>
<td>*1</td>
<td>0.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>0.162</td>
<td>0.24**</td>
</tr>
<tr>
<td>01</td>
<td>0.324</td>
<td>0.31</td>
</tr>
<tr>
<td>10</td>
<td>0.138</td>
<td>0.07**</td>
</tr>
<tr>
<td>11</td>
<td>0.276</td>
<td>0.39**</td>
</tr>
</tbody>
</table>
Linkage disequilibrium: $r^2$

- **Define**
  
  \[ r^2 = \frac{D^2}{P(A=0)P(B=0)P(A=1)P(B=1)} = 0.37 \]

- This really is the squared Pearson correlation of the two SNPs

- In practice, Pearson correlation is efficiently computed for all SNPs in windows as $X'X/n$

- This is a fundamental quantity for modeling GWAS z-scores

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Marginal allele frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0*</td>
<td>0.54</td>
</tr>
<tr>
<td>1*</td>
<td>0.46</td>
</tr>
<tr>
<td>*0</td>
<td>0.30</td>
</tr>
<tr>
<td>*1</td>
<td>0.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>0.162</td>
<td>0.24</td>
</tr>
<tr>
<td>01</td>
<td>0.324</td>
<td>0.31</td>
</tr>
<tr>
<td>10</td>
<td>0.138</td>
<td>0.07</td>
</tr>
<tr>
<td>11</td>
<td>0.276</td>
<td>0.39</td>
</tr>
</tbody>
</table>

**Key property:** $r^2$ correlation for individual SNPs is exactly the $r^2$ of the GWAS association summary statistics of these SNPs
LD score regression estimates heritability from summary data

A multivariate model for phenotype variation

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

non-genetic for indiv. \(i\)

multivar. effect on SNP \(j\)

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

\[ h^2 = \sum_j \beta_j^2 \]

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

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

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

\[ h^2 = \sum_j \beta_j^2 \]

Summary statistics data

(1) \( \chi^2 \) statistic for all SNP \( j \)

(2) LD matrix (or correlation between SNP \( j \) and \( k \))

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

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

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

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 \( k \)

Per SNP variance (heritability)

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

A univariate chi-square (GWAS)

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

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 \( k \)

Per SNP variance (heritability)

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

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

A univariate chi-square (GWAS)

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

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

\[ = N \sum_k \hat{r}_{jk}^2 \text{E}[\beta_k^2] + N \text{E}[\epsilon'_j^2] \]

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

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

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

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

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

<table>
<thead>
<tr>
<th>Trait</th>
<th>-log₁₀(P)</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schizophrenia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crohn's disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasting glucose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ever smoked</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at menarche</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Finucane et al. (2015)
Today: Deep Learning for Human Genetics and Disease

1. Review: GWAS, fine-mapping, Bayesian variant prioritization
2. Deep Learning for GWAS: calling SNPs, prioritize function
3. eQTLs/Mediation: intermediate molecular phenotypes
4. Linear Mixed Models (LMMs) for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): summing over many variants
6. Heritability: definition(s), missing heritability, partitioning
7. LD Score regression (LDSC) for fast heritability partitioning
8. Polygenic/Omnigenic disease models: core vs. periphery
9. Disease gene networks from GWAS evidence boosting
8. Polygenic \(\rightarrow\) Omnigenic models of disease
Recognizing “core” vs. “periphery” pathways
Schizophrenia GWAS: Number of significant loci

3,500 cases ⇔ 0 loci

10,000 cases ⇔ 5 loci

35,000 cases ⇔ 62 loci!

65,000 cases ⇔ 265 loci!
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
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

(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.

Boyle, Li, Pritchard, Cell, 2017
Today: Deep Learning for Human Genetics and Disease

1. Review: GWAS, fine-mapping, Bayesian variant prioritization
2. Deep Learning for GWAS: calling SNPs, prioritize function
3. eQTLs/Mediation: intermediate molecular phenotypes
4. Linear Mixed Models (LMMs) for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): summing over many variants
6. Heritability: definition(s), missing heritability, partitioning
7. LD SCore regression (LDSC) for fast heritability partitioning
8. Polygenic/Omnigenic disease models: core vs. periphery
9. Disease gene networks from GWAS evidence boosting
9. GWAS networks for evidence boosting
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 $\rightarrow$ 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
Hypothesis: Many associated genes implicate limited number of pathways

Proof: Statistically significant excess connectivity of genes in GWAS regions
Computational tools enable the integration of ‘human genetic screens’ with other genome-scale screening data.

Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology

Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits
Evaluating Significance

Repeat full permutation 50,000 times

...keep moving labels until the network has been fully permuted

Empirical Null Distribution

Number of Edges in Network

Frequency

Disease Network
PPI Networks identify specific genes and pathways

Fanconi anemia
9 synthetic loci

Rheumatoid arthritis
27 loci

Crohn’s disease
25 loci

Direct connectivity

Fanconi anemia:
p $<\!\!< 2 \times 10^{-5}$

Rheumatoid arthritis:
p $= 3 \times 10^{-4}$

Crohn’s disease:
p $= 1.11 \times 10^{-3}$
Validation of PPI networks
Further experimental support that the non-random networks are truly implicating the underlying genes

Network genes are co-expressed

Connected proteins are enriched for newly confirmed associated genes ($p=6.5\times10^{-4}$)
Integrating Autoimmune Risk Loci with Gene-Expression Data Identifies Specific Pathogenic Immune Cell Subsets

Xinli Hu,1,2,3,4 Hyun Kim,1,2 Eli Stahl,1,2,3 Robert Plenge,1,2,3 Mark Daly,3,5 and Soumya Raychaudhuri1,2,3,6,*

The American Journal of Human Genetics 89, 481–482, October 7, 2011

ImmGen data set:
223 murine immune cell subsets
Expression measured on 15,149 human homologs

Are human GWAS hits harboring loci significantly co-expressed in specific immune cell subsets?
GWAS hits significantly co-expressed in specific immune cell subsets
Other opportunities:
Cross-disease information

Genes coordinately associated to multiple disease are tightly functionally linked

Cotsapas et al, August 2011 *PLoS Genetics*