Lecture 15 - eQTLs
Molecular variation and mediation analysis

Slides credit:
Benjamin Raby, MD, MPH
Yongjin Park, PhD

Module 4: Population and Disease Genetics

Today: eQTLs, Mediation analysis

- Motivations for mapping regulatory variants
- Ex: Methylation QTLs in Alzheimer’s Disease
- eQTL conceptual frameworks
- eQTL discovery: methodologies
- Insights on eQTL function and biology
- Allele-specific analysis
- Mediation analysis & Bayesian inference

Motivations for eQTL mapping studies:
Phenotypic differences mostly not explained by coding variation

<table>
<thead>
<tr>
<th>Protein</th>
<th># amino acid differences</th>
<th>Total amino acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibrinopeptides</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Cytochrome c</td>
<td>0</td>
<td>104</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>0</td>
<td>130</td>
</tr>
<tr>
<td>Hemoglobin a</td>
<td>0</td>
<td>141</td>
</tr>
<tr>
<td>Hemoglobin b</td>
<td>0</td>
<td>146</td>
</tr>
<tr>
<td>Hemoglobin y</td>
<td>0</td>
<td>146</td>
</tr>
<tr>
<td>Hemoglobin y'</td>
<td>0</td>
<td>146</td>
</tr>
<tr>
<td>Hemoglobin δ</td>
<td>1</td>
<td>146</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>1</td>
<td>153</td>
</tr>
<tr>
<td>Carbonic Anhydrase</td>
<td>~3</td>
<td>264</td>
</tr>
<tr>
<td>Albumin</td>
<td>~6</td>
<td>580</td>
</tr>
<tr>
<td>Transferrin</td>
<td>~8</td>
<td>647</td>
</tr>
<tr>
<td>Total</td>
<td>~19 (0.7%)</td>
<td>2633</td>
</tr>
</tbody>
</table>

1975:
“We suggest that evolutionary changes in anatomy and way of life are more often based on changes in the mechanisms controlling the expression of genes than on sequence changes in the proteins.”

“We therefore proposed that regulatory mutations account for the major biological differences between humans and chimpanzees.”

Motivations for eQTL mapping studies:
Most genetic variation is non-coding

1000 Genomes Data from initial 179 individuals sequenced
Relative contribution of coding and non-coding variation to development of disease

Monogenic / Mendelian Disease

Polygenic / Complex Disease

Coding
Non-coding

11%
88%

Human Genetic Mutation Database
April 2010 release

Catalog of GWAS studies
Hindorff et al. PNAS 2009

eQTL mapping to facilitate disease susceptibility gene mapping

Dissection disease mechanism (e.g. FTO/obesity)

1. Tissue: pre-adipocytes
2. Establish downstream target gene(s): IRX3 and IRX5
3. Establishing causal nucleotide variant: rs1421085
4. Establish upstream regulator causality: ARID5B
5. Establish cellular phenotypic consequences: thermogenesis
6. Establish organismal phenotypic consequences: body weight

Understanding intermediate phenotypes enable perturbations, therapeutics

ARID5B KD (obesity)
IRX3, IRX5 knock-down (anti-obesity phenotypes)
IRX3, IRX5 overexpression (pro-obesity phenotypes)

C-to-T motif rescue (anti-obesity phenotypes)
T-to-C motif disruption (pro-obesity phenotypes)

Topological domains span 2.5Mb Implicate 8 candidate genes

CATGACTG
CATGCCCTG

2. Targets: 3D folding and expr. genetics indicate IRX3+IRX5

Risk allele: increased expression (gain-of-function)

Dixon, Nature 2012

GWAS region

1. Establish relevant tissue/cell type: pre-adipocytes
2. Establish downstream target gene(s): IRX3 and IRX5
3. Establishing causal nucleotide variant: rs1421085
4. Establish upstream regulator causality: ARID5B
5. Establish cellular phenotypic consequences: thermogenesis
6. Establish organismal phenotypic consequences: body weight

Understanding intermediate phenotypes enable perturbations, therapeutics

Thermogenic stimuli (e.g. cold)

Lipid storage → white adipocytes

ARID5B
PRDM16
PGC1α
UCP1

ARID5B OE (anti-obesity)
IRX3, IRX5 overexpression (pro-obesity phenotypes)

C-to-T motif rescue (anti-obesity phenotypes)

 Lean
Obese
Causality: SNP → Gene → Cellular phenotype → Organism

rs1421085 editing alters IRX3+IRX5 expression
(500,000 and 1 million nucleotides away!)

rs1421085 editing restores thermogenesis

Irx3 knock-down leads to weight-loss, tolerance

Today: eQTLs, Mediation analysis

- Motivations for mapping regulatory variants
- Ex: Methylation QTLs in Alzheimer’s Disease
- eQTL conceptual frameworks
- eQTL discovery: methodologies
- Insights on eQTL function and biology
- Allele-specific analysis
- Mediation analysis & Bayesian inference

Dataset overview
- Chromatin state
  - 18 states
  - 6 marks
  - DLPCF
  - Joint w/ 127 tissues
- Methylation level
  - 450k Illumina array
  - Brain Cortex DLPCF
  - 708 individuals
- Genotype
  - 620k SNPs
  - 586 individuals
  - Blood

Methylation in 750 Alzheimer patients/controls

750 individuals (~50% w/AD)

Philip deJager, Epigenomics Roadmap

Memory and Aging Project
Religious Order Study

1. Active promoter
2. Promoter/Ranking
3. Active enhancer
4. Weak enhancer
5. Gene bodies
6. Enhancers
7. Repetitive
8. Heterochromatin

- Patients followed for 10+ years with cognitive evaluations
- Brain samples donated post-mortem methylation/genotype
- Seek predictive features: SNPs, QTLs, mQTLs, regulation

Most methylation probes are high or low, with little variability

- Chromatin state definitions
- Distribution of CpG avg methylation levels (in Illumina 450k array)
  - Average methylation across 708 individuals
- Distribution of CpG methylation variance across individuals
  - Log: Very few probes show high variance
- 2D distribution: average vs. variance
  - Highest variance immediate-methylation

- However: Intermediate methylation is not just an artifact of averaging bimodal levels between individ.
  - Intermediate methylation is truly intermediate
Enhancer regions show intermediate methylation

- Enhancer states: Intermediate (EnhG1/G1/A1/A2/Wk)
- Active states: Promoters: low, Tx: high.
- Repressed states: TssBiv/EnhBiv/ReprPC: low. Quies/ReprPCWk: high

Enhancers are most variable, promoters least

- Chromatin states vary 10-fold in methylation variance, 3-fold in stdev
- Active states: EnhA > EnhWk > EnhG > TxWk > TssFlnk >> TssA
- Repressed states: Quies > ReprPC > EnhBiv >> TssBiv

Discover 50,000 methylation QTLs after Bonferroni

- Overlay meQTL discovery plot

meQTL discovery vs. distance vs. cohort size

- Vary: (1) distance from CpG; (2) effect size; (3) cohort size
- Strongest effects within 20 kb of tested CpGs
- Expectation for 100, 150, 200 individuals (if searching a 1Mb region)

Selection of the number of individuals

- More individuals ➔ linearly more meQTLs, but smaller effect size
- Strongest effects concentrated within 20 kb of tested CpGs ➔ can be used to increase power for smaller sample sizes.

# of individuals ↔ MAF of meQTL SNPs

- Focusing on 100-150 individuals, MAF > 0.1, as expected
- Large number of SNPs never probed even with 600 indiv
meQTL probes are enriched in enhancers + TssFlnk

Enhancer variation correlated with AD diagnosis

- Prioritize EnhA, EnhWk, TssFlnk regions for meQTLs
- Profile variation in H3K27ac directly (ChIP-seq component)

Enrichment for meQTLs

- Enhancer variation is actually biologically meaningful (not just an artifact of meaningless variation)
- Enhancers > all methylation > Promoters > APOE4 >> SNPs

Functional enrichments persist across 1000 probes

- AD-associated probes in enhancers. Age-assoc in Polycomb
- 10,000 phenotype permutations ➔ Statistical significance
- AD top 1k GWAS enrichment persists across 100k+ probes

Imputed MWAS: increased power, genetic component

| MeQTL: G → M | N=800 |
| GWAS: G → D | N=74k |
| MWAS: M ← D | N=800 |
| iMWAS: G → iM | N=74k |

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

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

Today: eQTLs, Mediation analysis

- Motivations for mapping regulatory variants
- Ex: Methylation QTLs in Alzheimer’s Disease
  - eQTL conceptual frameworks
  - eQTL discovery: methodologies
  - Insights on eQTL function and biology
  - Allele-specific analysis
  - Mediation analysis & Bayesian inference
A conceptual framework for mapping regulatory variants

Gene expression as a quantitative trait

Gene expression is a heritable trait

Types of regulatory variants

Cis vs. Trans elements

- **cis-eQTL**: variant resides in close proximity to target gene location
  - Multiple mechanisms implicated
    - Promoter
    - Splicing
    - Methylation
    - Chromatin modification

- **trans-eQTL**: variant resides very distant to the target
  - Alternative chromosome
  - Same chromosome, but far away
  - Mechanisms less clear

Today: eQTLs, Mediation analysis

- Motivations for mapping regulatory variants
- Ex: Methylation QTLs in Alzheimer’s Disease
- eQTL conceptual frameworks
- eQTL discovery: methodologies
- Insights on eQTL function and biology
- Allele-specific analysis
- Mediation analysis & Bayesian inference
**eQTL methodology**

**Regression and beyond**

---

**The nuts and bolts of an eQTL study**

- **Cell isolation**
- **RNA isolation**
- **Expression measurement**
- **DNA Genotyping**

**Expression ~ genotype**

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

**Linear Regression Equation**

---

**Considerations: Expression data**

- Array vs. Sequencing based
- Pre-processing and normalization

**Feature selection:**

- Uniquely mappable elements
- SNP under probe effects

**Technical considerations / bias**

- Minimum expression levels
- Minimum population variance

**Statistical considerations / power**

---

**SNP-under-probe effects**

Several variants identified within 10 kb of transcript with significant associations with transcript abundance (p=3x10^{-7}).

All SNP had allele frequencies of 0.44-0.48

Inspection of distribution of intensity values suggests binary pattern of expression.

Inspection of target sequence for ILMN_15237 reveals a variant (rs7326) with frequency of 0.44-0.48

---

**SNP-under-probe effects: overrated?**

Study of cis-eQTL in CD4+ lymphocytes:

Trend to higher prevalence of SNP-under-probe effects for eQTL transcripts:

7.4% (123 of 1662) in associated transcripts vs. 6.2% (1105/17789) for non-associated transcripts (Fisher’s exact test, p = 0.06).

Removal of probes with known sequence variation did not change our results & Regulatory variation confirmed by other means for many variants with SNP under probe effects.

Similar to observations by others


Murphy et al. Hum Mol Genet 2010
**Variability in gene expression**

Mean Intensity

On average, most cells express fewer than 50% of known genes.

Testing such targets is inefficient and essentially useless.

Median (6.6)

Variance in intensity

---

**Considerations: Genotype data**

- Minor allele frequency
- Defining search space:
  - Cis-only vs. all SNP (truly genome-wide)
  - Search radius

---

**Considerations: Covariate adjustment**

- Demographic covariates
  - Age, gender

- Population stratification
  - Structure analysis
  - Genomic control

- Technical covariates
  - Batch effects
  - Unobserved confounders

---

**Population stratification**

Ethnic-specific differences in gene expression

Ethnic-specific differences in genetic architecture


**Estimating population stratification**

Genotypes summarized as principal components can be used for covariate adjustment in genetic analysis

Expression = genotype + PC1 + PC2 + PC3 ... + other covariates
There are numerous factors that can contribute to technical variance of gene expression, particularly for large population studied conducted over time...

Adjusting for expression PCs results in incremental gains in number of eQTL identified (chromosome 22 here)

Note there is an optimal number of PCs beyond which yield decreases

Expanded eQTL models

$$Y_{ij} = \alpha + \beta_{ij}\text{genotype} + \epsilon$$

$$Y_{ij} = \alpha + \beta_{1ij}\text{genotype} + \beta_{2ij}\text{gender} + \beta_{3ij}\text{age} +$$

$$\beta_{4ij}\text{PC1} + \beta_{5ij}\text{PC2} + \beta_{6ij}\text{PC3} + \beta_{7ij}\text{PC4} +$$

$$\beta_{8ij}\text{PC1} + \beta_{9ij}\text{PC2} + \beta_{10ij}\text{PC3} + \beta_{11ij}\text{PC4} +$$

$$\beta_{12ij}\text{PC5} + \beta_{13ij}\text{PC6} + \beta_{14ij}\text{PC7}$$

$$+ \epsilon$$

Decisions matter: impacts of MAF and distance

Rare variants & statistical variance
Setting parameters is tough!

Optimal strategies are dataset specific

Dataset 1

Dataset 2

Number of Principal Components

Optimal strategies are dataset specific

Gene Yield

O

P

Distance from gene

Search radius impacts FDR

Opportunities for greedy tuning of cis-eQTL search

GGtools: iterative eQTL research

Today: eQTLs, Mediation analysis

- Motivations for mapping regulatory variants
- Ex: Methylation QTLs in Alzheimer’s Disease
- eQTL conceptual frameworks
- eQTL discovery: methodologies
- Insights on eQTL function and biology
- Allele-specific analysis
- Mediation analysis & Bayesian inference
Insights on eQTL function and biology

eQTL prevalence and impact

- eQTL positioning
- Reproducibility
- Tissue-specificity
- Gene annotation
- Co-regulation
- Trans-acting eQTL
- Disease-enrichment
- Directionality
- Disease-relevance

- Identification of regulatory variants for 1,585 genes (~20% of genes tested)
- Variants explain ~10% of transcript abundance (IQR of 7.6% - 16.1%)


<table>
<thead>
<tr>
<th>Study</th>
<th>Sample size</th>
<th>Cell type</th>
<th>Cis-eQTL</th>
<th>Trans-eQTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stranger</td>
<td>270</td>
<td>LCL</td>
<td>1,348</td>
<td>180</td>
</tr>
<tr>
<td>Dixon</td>
<td>378</td>
<td>LCL</td>
<td>536</td>
<td>1,453</td>
</tr>
<tr>
<td>Emilsson</td>
<td>1,002</td>
<td>Blood</td>
<td>1,970</td>
<td>52</td>
</tr>
<tr>
<td>Emilsson</td>
<td>673</td>
<td>Adipocytes</td>
<td>1,215</td>
<td>25</td>
</tr>
<tr>
<td>Murphy</td>
<td>200</td>
<td>CD4+ cells</td>
<td>1,585</td>
<td>-</td>
</tr>
<tr>
<td>Westra</td>
<td>5,311</td>
<td>Blood</td>
<td>6,418</td>
<td>430</td>
</tr>
</tbody>
</table>


eQTL for thousands of genes identified

The overwhelming majority of significant eQTL localize to within a few kb of the target gene. Here we observe a 34-fold enrichment for eSNP within 5kb.

Though significant and validated regulatory variants have been observed at a distance, detecting these is more difficult.


Insights on eQTL function and biology

- eQTL prevalence and impact
- eQTL positioning
- Reproducibility
- Tissue-specificity
- Gene annotation
- Co-regulation
- Trans-acting eQTL
- Disease-enrichment
- Directionality
- Disease-relevance

Reproducibility

In general, there is a 30-40% reproducibility rate across datasets that are well powered (at least 200 subjects, MAF cutoff of 0.10).

Importantly, despite ethnic-specific differences in eQTL, when eQTL detected in two or more cohorts, the direction and magnitude of effect is very consistent across ethnic groups.

Tissue specificity

The question of tissue specificity is both intuitive and controversial. Some papers claim tissue specificity is the rule, while others find a great deal of similarities.

Approximately 30-35% are shared cross-tissues

GTEx eQTLs across 44 tissues

- Sample size matters

Expected diversity of meQTLs based on gene expression diversity of GTEx samples. We have selected tissues (red lines) that are maximally distinct in their expression patterns, in order to maximize the expected diversity of meQTLs discovered, and to maximally capture the diversity of gene expression regulatory programs.
eQTL sharing across tissues

- Bimodal distribution of sharing

Defining gene-variant relationships

Establishing a molecular link

GWAS and eQTL

CD4+ T-cell co-expression network

eQTL-enabled genic annotation of SNP

Insights on eQTL function and biology

GWAS and eQTL

CD4+ T-cell co-expression network

GWAS and eQTL

CD4+ T-cell co-expression network

eQTL prevalence and impact

eQTL positioning

Reproducibility

Tissue-specificity

Gene annotation

Co-regulation

Trans-acting eQTL

Disease-enrichment

Directionality

Disease-relevance

Verlaan, Berlivet, Hunninghake, et al. AJHG 2009


Co-regulation

17q locus regulates 3 genes, not 1

Trans-acting eQTL

- Trans-eQTL are harder to find than cis-eQTL
  - Pre-test probability lower → higher significance threshold
  - Effect sizes are substantially smaller → lower power
  - Impact of epistasis and environmental factors are likely greater
- Fewer trans-eQTL detected per study
- Few replicate reliably

Graphical model approach to map trans eQTLs

Forfax et al. Science. 2014; 343:1246949

Insights on eQTL function and biology

eQTL prevalence and impact
eQTL positioning
Reproducibility
Tissue-specificity
Gene annotation
Co-regulation

Trans-acting eQTL

Disease-enrichment
Directionality
Disease-relevance

Trans-acting eQTL are sensitive to environmental stimuli


Identifying the MiGAQK Protein Carrol-I as a Central Regulator of Neural Immune Reagons and Mice to Gamma-White Mice. MetabMol

Insights on eQTL function and biology

- eQTL prevalence and impact
- eQTL positioning
- Reproducibility
- Tissue-specificity
- Gene annotation
- Co-regulation
- Trans-acting eQTL
- Disease-enrichment
- Directionality
- Disease-relevance

Remember our motivation: Are disease-associated variants more likely to be eQTL?

- Monogenic / Mendelian Disease
  - Coding: 11%
  - Non-coding: 88%

- Polygenic / Complex Disease

Human Genetic Mutation Database
April 2010 release

Catalog of GWAS studies
Hindorff et al. PNAS 2009

Disease-associated variants are more likely to be eQTL

- Crohn’s Disease
- Th17
- Multiple Sclerosis
- CD3+
- QRS Duration
- Heart


Can we leverage eQTL data for disease-variant identification?

- Multivariate regression model building:
  - Outcome: Disease-susceptibility variants identified by GWAS and all variants with $r^2 > 0.8$
  - Variables: Distance from TSS, MAF, Chromatin state, and eQTL FDR
- Training set: random set of 2.7 million variants mapping to within 50 kb of transcript
- Test set: remaining ~2.0 million variants

Utility #3: Identify new disease-associated loci

Regression Model building:

Outcome

Predictors

Model training:

Even-numbered chromosomes (~2.4 million SNPs)

Odd-numbered chromosomes (~2.3 million SNPs)
### Multivariate Predictors of GWAS-tagging variants

<table>
<thead>
<tr>
<th>Distance fromGene (25-50 kb ref)</th>
<th>Within gene</th>
<th>&gt;1kb</th>
<th>&gt;5kb</th>
<th>&gt;10kb</th>
<th>&gt;25kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minor Allele Frequency (0.25–0.5 ref)</td>
<td>0.00 – 0.025</td>
<td>0.025 - 0.05</td>
<td>0.05-0.10</td>
<td>0.10-0.20</td>
<td></td>
</tr>
<tr>
<td>eQTL FDR (&gt;2 ref)</td>
<td>0.001 – 0.01</td>
<td>0.01-0.05</td>
<td>0.05-0.10</td>
<td>0.10-0.20</td>
<td></td>
</tr>
</tbody>
</table>

### eQTLs Improve GWAS variant prediction

<table>
<thead>
<tr>
<th>Odds Ratio</th>
<th>Proportion of disease-linked SNP detected in training set</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.009</td>
<td>3.10%</td>
</tr>
<tr>
<td>0.026</td>
<td>7.80%</td>
</tr>
<tr>
<td>0.036</td>
<td>12.90%</td>
</tr>
</tbody>
</table>

### Insights on eQTL function and biology

- **eQTL prevalence and impact**
- **eQTL positioning**
- Reproducibility
- Tissue-specificity
- Gene annotation
- Co-regulation
- Trans-acting eQTL
- Disease-enrichment
- Directionality
- Disease-relevance

### Today: eQTLs, Mediation analysis

- **Motivations for mapping regulatory variants**
- **Ex: Methylation QTLs in Alzheimer’s Disease**
- **eQTL conceptual frameworks**
- **eQTL discovery: methodologies**
- **Insights on eQTL function and biology**
  - **Allele-specific analysis**
  - **Mediation analysis & Bayesian inference**

### eQTL calling

**rs17366947**
Ref: GAGCCAGAGGCCCCCTGATATTGCAGCACAATGGGTCTTCTCTCTCTCTGAGAACTAAAGCTGG
Alt: GAGCCAGAGGCCCCCTGATATTGCAGCACAATGGGTCTTCTCTCTCTCTGAGAACTAAAGCTGG

**SOX2 Motif**

**Wilcoxon rank sum test:**

- **H0:** the haplotype imbalance ratio distribution are same no matter the eQTL snp is heterozygous or not.
- **H1:** the haplotype imbalance ratio are higher when eQTL snp is heterozygous comparing to homozygous

### aseQTL calling

**Example aseQTL**

(Alonso Battle, et al Genome Research, 2013)
Combined Haplotype Test

Maximize likelihood of two observed components:

$$L(a_k, \beta_k, \psi_j | D) = \prod_i \Pr_{\text{Beta-Binomial}}(X = x_{ij} | \lambda_k, \Omega, \phi_j) \prod_k \Pr_{\text{Beta-Negative-Binomial}}(Y = y_k | \mu_k, n_{jk}, T_k)$$

Allele-specific methylation enriched in GWAS loci

Today: eQTLs, Mediation analysis

• Motivations for mapping regulatory variants
  • Ex: Methylation QTLs in Alzheimer’s Disease
  • eQTL conceptual frameworks
  • eQTL discovery: methodologies
  • Insights on eQTL function and biology
  • Allele-specific analysis
  • Mediation analysis & Bayesian inference

Imputed MWAS: increased power, genetic component

GWAS: G → D (N=74k)
meQTL: G → M (N=800)
MWAS: M → D (N=800)
iMWAS: G → iM → D (N=74k)

Key Idea:
• Learn G→D directly (complex phenotype)
• Learn G→M (simpler phenotype)
• 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

1,000 individuals x 8 tissues / cell types

200 individuals

GWAS: G → D N=74k
meQTL: G → M N=800
eQTL: G → T N=800
MWAS: M → T D N=800
TWAS: T → D N=800
NWAS: G → iM → iT D N=74k

Learn G→D directly (complex phenotype)
Learn G→M (simpler phenotype)
Learn G→H (simpler phenotype)
M→D (no causality)
M→T (no causality)
Apply G→M to get iMiT

Learn G→M directly (complex phenotype)
Learn G→M (simpler phenotype)
Learn G→H (simpler phenotype)
M→D (no causality)
M→T (no causality)
Apply G→M to get iMiT

**Expand to multiple intermediate phenotypes**

GWAS: G → D N=74k
meQTL: G → M N=800
eQTL: G → T N=800
MWAS: M → T D N=800
TWAS: T → D N=800
NWAS: G → iM → iT D N=74k

Learn G→D directly (complex phenotype)
Learn G→M (simpler phenotype)
Learn G→H (simpler phenotype)
M→D (no causality)
M→T (no causality)
Apply G→M to get iMiT

Learn G→M directly (complex phenotype)
Learn G→M (simpler phenotype)
Learn G→H (simpler phenotype)
M→D (no causality)
M→T (no causality)
Apply G→M to get iMiT

**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 (eg: blood)**

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

**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

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

Gusev et al. “Integrative approaches for large-scale transcriptome-wide association studies” 2016 Nature Genetics

**Multivar. regression, a workhorse of imputed TWAS.**

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

Bayesian linear regression

Formal definition of a linear model

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

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

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

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

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

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

Maximum likelihood estimation w/o prior distribution

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

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

\[ \hat{\theta}_{MLE} = (X^TX)^{-1}X^Ty. \]

Intrinsic difficulty of MLE

\[ \ln \mathcal{L}(\theta | X, y). \]

\[ \ln \mathcal{L}(\theta | X, y). \]

\[ \hat{\theta}_{MLE} = (X^TX)^{-1}X^Ty. \]

\[ \hat{\theta}_{MLE} = (X^TX)^{-1}X^Ty. \]

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

(1) \( p \times p \) is huge symmetric matrix for GWAS.

(2) We should also ask whether this inversion even exists, meaning there is a unique solution to this.

(3) Rank of \( XX^T \) matrix is the same as \( \text{rank}(X) \leq \min\{\text{rows, columns} \} = n \) (for \( n \ll p \)).

(4) If \( p \gg n \), prob. of this dot product \( X^Ty \) significantly deviating from zero increases.

Think about each element of \( X^Ty \):

(1) each \( j \)-th element of the vector \( = x_j^Ty \)

(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 \( x^Ty \) 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)

Variable selection

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

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

Variable selection by shrinkage of coeff. \( \theta \)

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

L1-regularization and Laplace prior

Maximizing a posteriori

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

Equivalent \( L_1 \)-regularized error

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

Prior distribution

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

L2-regularization and Gaussian prior

Maximizing a posteriori

\[
\ln p(y|x, \theta) + \ln p(\theta | \lambda) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i \theta)^2 - \frac{\lambda}{2} \| \theta \|_2^2
\]

Equivalent \( L_2 \)-regularized error

\[
\sum_{i=1}^{n}(y_i - x_i \theta)^2 + \frac{\lambda}{2} \| \theta \|_2^2
\]

Prior distribution

\[
N(\theta | 0, \lambda^{-1} I) \propto \exp\left(-\frac{\lambda}{2} \| \theta \|_2^2\right)
\]

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

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

Log-likelihood + log prior:

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

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

\[ \nabla_{\Theta} \ln p(\mathbf{y}|\mathbf{X}, \Theta) = -\frac{1}{\sigma^2} \mathbf{X}^T (\mathbf{y} - \Theta\mathbf{X}) - \lambda \Theta = 0 \]

Rearranging the equation:

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

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

(In Gaussian, 2nd order derivative = inverse of variance)

Comparison between MLE and Bayesian regression

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

\[ p \times p \text{ covariance matrix from MLE} \]

\[ \text{from prior prob of } \theta \sim N(0, \lambda^2) \]

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

(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 \( \Theta = 0 \)) or with MLE \( \Theta \), and find \( \lambda \) accordingly.

Intractability of MLE for large-scale linear models

\( n \times p \)

High dimensionality (not enough sample / too many parameters)

Colinearity drives spurious associations in MLE

Simulated model:

\( y \sim X_1 \theta_1, X_2 \sim X_1 \theta_2 \)

Fitted model:

\( y \sim X_1 \theta_1 + X_2 \theta_2 \)

\( R^2 = 1 \) correlation \( X_1 \sim X_2 \)

\( \theta_1 \)

\( \theta_2 \)

\( \theta_{12} \)

L1 and L2-regularized regression effectively handle issues in practice

1 best SNP

All SNPs

Models by selected set of SNPs

Gray = best possible \( R^2 \) aka heritability

Red = Cross-validated \( R^2 \)

Almost all genes attain maximum heritability

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).

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 $\theta$

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

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)

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

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

Gaussian

$p(\theta | z=0) = (1-\pi) \delta(\theta)$

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

Spike-slab prior model effectively avoid colinearity

Simulated model: $y \sim X_1 \theta_1 + X_2 \theta_2$

OLS model: $y \sim X_1 \theta_1 + X_2 \theta_2$

MLE is overfitting

True effect locates little deeper in likelihood contour

Fitted model: $y \sim X_1 \theta_1 + X_2 \theta_2$

$\theta_2 \sim \text{spike-slab}$

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).

Figure: Hernandez-Lobato (2014)

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

Completely set to zero if not selected

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

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

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

Gaussian

$p(\theta | z=0) = (1-\pi) \delta(\theta)$

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

Spike-slab prior model effectively avoid colinearity

Simulated model: $y \sim X_1 \theta_1 + X_2 \theta_2$

OLS model: $y \sim X_1 \theta_1 + X_2 \theta_2$

MLE is overfitting

True effect locates little deeper in likelihood contour

Fitted model: $y \sim X_1 \theta_1 + X_2 \theta_2$

$\theta_2 \sim \text{spike-slab}$

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).

Figure: Hernandez-Lobato (2014)

Spike-slab prior model effectively avoid colinearity

Simulated model: $y \sim X_1 \theta_1 + X_2 \theta_2$

OLS model: $y \sim X_1 \theta_1 + X_2 \theta_2$

MLE is overfitting

True effect locates little deeper in likelihood contour

Fitted model: $y \sim X_1 \theta_1 + X_2 \theta_2$

$\theta_2 \sim \text{spike-slab}$

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).

Figure: Hernandez-Lobato (2014)

Spike-slab prior model effectively avoid colinearity

Simulated model: $y \sim X_1 \theta_1 + X_2 \theta_2$

OLS model: $y \sim X_1 \theta_1 + X_2 \theta_2$

MLE is overfitting

True effect locates little deeper in likelihood contour

Fitted model: $y \sim X_1 \theta_1 + X_2 \theta_2$

$\theta_2 \sim \text{spike-slab}$

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).

Figure: Hernandez-Lobato (2014)

Spike-slab prior model effectively avoid colinearity

Simulated model: $y \sim X_1 \theta_1 + X_2 \theta_2$

OLS model: $y \sim X_1 \theta_1 + X_2 \theta_2$

MLE is overfitting

True effect locates little deeper in likelihood contour

Fitted model: $y \sim X_1 \theta_1 + X_2 \theta_2$

$\theta_2 \sim \text{spike-slab}$

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).

Figure: Hernandez-Lobato (2014)

Spike-slab prior model effectively avoid colinearity

Simulated model: $y \sim X_1 \theta_1 + X_2 \theta_2$

OLS model: $y \sim X_1 \theta_1 + X_2 \theta_2$

MLE is overfitting

True effect locates little deeper in likelihood contour

Fitted model: $y \sim X_1 \theta_1 + X_2 \theta_2$

$\theta_2 \sim \text{spike-slab}$

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).

Figure: Hernandez-Lobato (2014)

Spike-slab prior model effectively avoid colinearity

Simulated model: $y \sim X_1 \theta_1 + X_2 \theta_2$

OLS model: $y \sim X_1 \theta_1 + X_2 \theta_2$

MLE is overfitting

True effect locates little deeper in likelihood contour

Fitted model: $y \sim X_1 \theta_1 + X_2 \theta_2$

$\theta_2 \sim \text{spike-slab}$

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).

Figure: Hernandez-Lobato (2014)

Spike-slab prior model effectively avoid colinearity

Simulated model: $y \sim X_1 \theta_1 + X_2 \theta_2$

OLS model: $y \sim X_1 \theta_1 + X_2 \theta_2$

MLE is overfitting

True effect locates little deeper in likelihood contour

Fitted model: $y \sim X_1 \theta_1 + X_2 \theta_2$

$\theta_2 \sim \text{spike-slab}$

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).

Figure: Hernandez-Lobato (2014)

Spike-slab prior model effectively avoid colinearity

Simulated model: $y \sim X_1 \theta_1 + X_2 \theta_2$

OLS model: $y \sim X_1 \theta_1 + X_2 \theta_2$

MLE is overfitting

True effect locates little deeper in likelihood contour

Fitted model: $y \sim X_1 \theta_1 + X_2 \theta_2$

$\theta_2 \sim \text{spike-slab}$

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).

Figure: Hernandez-Lobato (2014)

Spike-slab prior model effectively avoid colinearity

Simulated model: $y \sim X_1 \theta_1 + X_2 \theta_2$

OLS model: $y \sim X_1 \theta_1 + X_2 \theta_2$

MLE is overfitting

True effect locates little deeper in likelihood contour

Fitted model: $y \sim X_1 \theta_1 + X_2 \theta_2$

$\theta_2 \sim \text{spike-slab}$

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).

Figure: Hernandez-Lobato (2014)
We assume phenotype vector were generated by
\[ y \sim \mathcal{N}(X\theta, \sigma^2 I). \]
Therefore, \( p(x | \theta, \sigma^2, I) \) vector follows
\[ z \sim \mathcal{N}\left(\frac{X^TX\theta}{\sqrt{\sigma^2}}, \frac{1}{n}\right) \approx \mathcal{N}(R\theta, R), \]
where LD matrix \( R = n^{-1}X^TX \) and \( \lambda = (n\sigma^2)^{-1/2} \) absorbs all scaling factors.

Hormozdiari et al. (2014)
MCMC: Metropolis-Hastings

For iteration $t = 1, \ldots, T_{\text{max}}$

Determining $T_{\text{max}}$ is an art.

(a) Scheduling should not matter much, but sometimes different scheduling can make big difference.
(b) Blocked scheduling usually helps (blocked Gibbs sampling; type by type).
(c) Researchers usually repeat multiple simulations starting from different random restarts (multiple Markov chains) to determine convergence.

---

MCMC: results of example data

MLE solution w/ statistical power

MCMC with high sensitivity

Some false positives (slightly deviate from zero but significant)

Variables

Method

Bayesian (MCMC posterior) ▼ Frequentist (MLE ± 2SE)

---

MCMC: multiple Markov chains

Different random start

Different random start

converge to the same distribution

---

Advanced MCMC: collapsed sampling

When we sample $\theta_\omega$ collapse out uninteresting $\theta$'s and $\sigma$,

Pros:
- Local sampling is less tied with current configuration.
- Much faster convergence.
- Mathematically elegant.

Cons:
- Math is harder, sometimes impossible.
- Averaging over uncertainty is not necessarily the best.

---

Advanced MCMC: elliptical slice sampling

Slice sampling:
- Slice probability (log-probability) horizontally to next sampling points.
- Propose a next point to land on the acceptable region (area under the slice).
- Metropolis-Hastings with much better proposal (but not so scalable for high-dimensional).

Moving along the axis

Slice sampling along the ellipse

1st proposal
2nd proposal
3rd attempt

1st proposal auxiliary (make ellipse)

2nd proposal shrink $\theta_{\text{net}}$

3rd attempt shrink $\theta_{\text{net}}$

Finally got it! (next x)

 Murray, Adams, MacKay, AISTATS (2009)

---

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>Accuracy</td>
<td>correct</td>
<td>approximate, stochastic</td>
<td>approximate, deterministic</td>
</tr>
<tr>
<td>Convergence</td>
<td>sure</td>
<td>Global optima at equilibrium</td>
<td>Local optima in finite time</td>
</tr>
<tr>
<td>Flexibility</td>
<td>very limited</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>Examples</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>
Variational Bayes (mean field theory)

Construct fully factored (independent),
\[ q(\mathbf{\theta}) = \prod N(\mathbf{\theta}_j | \mathbf{\mu}_j, \lambda_j) \]

Estimate \( p(\mathbf{\theta} | \text{Data}) \) by finding \( q(\mathbf{\theta}) \) as close as possible to \( p(\mathbf{\theta} | \text{Data}) \).

Variational param. Estimate \( p(\mathbf{\theta} | \text{Data}) \) by finding \( q(\mathbf{\theta}) \) as close as possible to \( p(\mathbf{\theta} | \text{Data}) \).

Minimization of Kullback-Leibler (with respect to variational \( \lambda, \mu \)).

Minimization of Kullback-Leibler (with respect to variational \( \lambda, \mu \)).

VB: iterative mean field update

Initialize all the parameters, e.g., \( E[\theta_1], ..., E[\theta_p], \) and \( E[\sigma] \)

At iteration \( t \), update \( \theta_1 \):

Replace \( \theta \)'s and \( \sigma \) with their expectation (mean field)

Optimize \( q(\theta_1 | ...) \) as if we are only concerning \( \theta_1 \)

Under the mean field

Find optimal distribution of \( \theta_j \)

(a) Optimal \( q(\theta) \) is relatively straightforward for exponential family distributions (e.g., Gaussian, Poisson, Gamma, etc.)
(b) Use the same tricks, take derivatives w.r.t. \( \mu \), \( \lambda \), and set them to zero.

Using VB inference with spike-slab prior

VB captures true effect

So many false positives by MCMC w/o spike-slab prior

MLE is overfitting

True

VB generally yields more conservative PIPs

Strong agreements between MCMC and VB on known causal variants

Carbonetto & Stephens, Bayesian Analysis (2012)
Make use of statistical toolboxes!

<table>
<thead>
<tr>
<th>Toolbox</th>
<th>Inference</th>
<th>Programming language</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>stan</td>
<td>Mostly MCMC</td>
<td>R, Python, Julia</td>
<td>No spike-slab :-(</td>
</tr>
<tr>
<td>PyMC3</td>
<td>Mostly MCMC, weak VB</td>
<td>Python (Theano)</td>
<td>Similar functionality as Stan</td>
</tr>
<tr>
<td>edward</td>
<td>Strong VB, many MCMC alg.</td>
<td>Python (TensorFlow)</td>
<td>Provides wrappers to previously developed methods</td>
</tr>
</tbody>
</table>

Other tools available (a little lower-level implementation):
1. Generally Theano, TensorFlow, and Torch7 are useful library.
2. For class projects, we recommend R, Matlab, Python + stan or PyMC3.
3. C++ matrix libraries : e.g., Armadillo, Eigen, Rcpp wrappers.

Today: eQTLs, Mediation analysis

- Motivations for mapping regulatory variants
- eQTL conceptual frameworks
- eQTL discovery: methodologies
- Insights on eQTL function and biology
- Allele-specific analysis
- Mediation analysis & Bayesian inference