Advanced statistical genetics methods

Yongjin Park
University of British Columbia

22 March, 2022

Learning objective

- Population structures in genetics data
  - Admixture model
  - Linear mixed effect model
- Linkage disequilibrium
  - Rare variant burden tests
  - Fine-mapping causal variants
- GWAS summary statistics
  - Transcriptome-wide association studies
  - LD-score regression

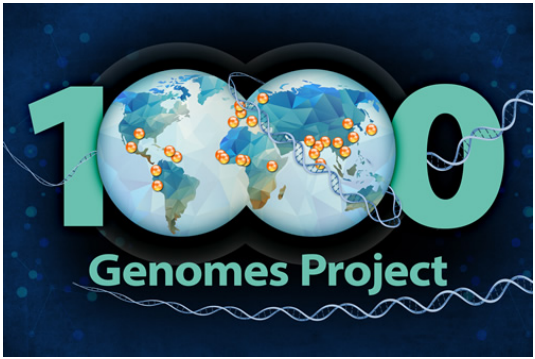
Today's lecture

Population structures in human genetics data

Linkage Disequilibrium: blessing and curse

Systems genetics and summary statistics-based inference

The 1000 Genomes Project to investigate Human Genetic Variation



1KG contains whole genome sequencing data of 2,490 individuals sampled from 26 groups based on the origins and geographical locations (as of 2013 phase3).

Single Nucleotide Polymorphism (SNP) genotype information

0	1	0	1	1	1	1	0	0	0
0	2	0	2	2	2	2	0	0	0
0	1	0	1	1	1	1	0	0	0
0	2	0	2	2	2	2	0	0	1
0	2	0	2	2	2	2	0	1	0
0	2	0	2	2	2	2	0	0	1
0	2	0	2	2	2	2	0	0	1
0	2	0	2	2	2	2	0	0	1
0	2	0	2	2	2	2	0	0	1
0	1	0	1	1	1	1	0	0	1

first 10 individuals and 10 variants

Previously on the lecture 18:

- We will focus on biallelic variant (two allele, two different forms)
- Major and minor characters (depending on the frequency in reference data)
- We keep track of the number of the minor allele (0 to 2, due to diploid genome)

Variant-level variation across individuals

Using the minor allele frequency (MAF), let f_j be a minor allele frequency (MAF) of a variant j. In Binomial distribution,

- What is the mean of this variant?

Remark: Technically, the dosage (0,1,2) does not follow binomial distribution. Why? The underlying data generation process involves haplotypes (separating the maternal and paternal 0/1 counts) and dependency along the genomic axis.

Variant-level variation across individuals

Using the minor allele frequency (MAF), let f_j be a minor allele frequency (MAF) of a variant j. In Binomial distribution,

- What is the mean of this variant?

$$\hat{\mathbb{E}}\left[X_j\right] = 2f_j$$

Remark: Technically, the dosage (0,1,2) does not follow binomial distribution. Why? The underlying data generation process involves haplotypes (separating the maternal and paternal 0/1 counts) and dependency along the genomic axis.

Variant-level variation across individuals

Using the minor allele frequency (MAF), let f_j be a minor allele frequency (MAF) of a variant j. In Binomial distribution,

- What is the mean of this variant?

$$\hat{\mathbb{E}}[X_j] = 2f_j$$

- What is the variance of this variant?

Remark: Technically, the dosage (0,1,2) does not follow binomial distribution. Why? The underlying data generation process involves haplotypes (separating the maternal and paternal 0/1 counts) and dependency along the genomic axis.

Variant-level variation across individuals

Using the minor allele frequency (MAF), let f_j be a minor allele frequency (MAF) of a variant j. In Binomial distribution,

- What is the mean of this variant?

$$\hat{\mathbb{E}}\big[X_j\big] = 2f_j$$

- What is the variance of this variant?

$$\hat{\mathbb{V}}\big[X_j\big] = 2f_j(1 - f_j)$$

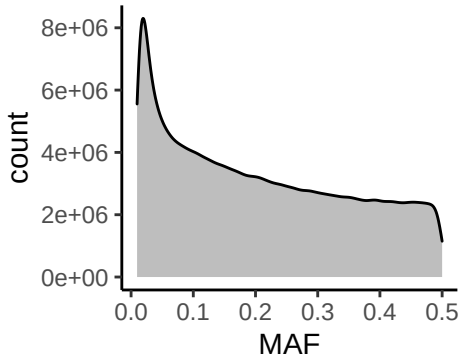
Remark: Technically, the dosage (0,1,2) does not follow binomial distribution. Why? The underlying data generation process involves haplotypes (separating the maternal and paternal 0/1 counts) and dependency along the genomic axis.

Variant-level variation across individuals

We can easily calculate MAF using `bigsnpr`

```
maf <- snp_MAF(X$genotypes,
               ncores=8)
```

What is your interpretation?



Much of human genetics problems centre on two covariance matrices

For a standardized $n \times p$ genotype matrix X,

1. Genetic relatedness matrix (GRM)

a $n \times n$ matrix

$$K \approx XX^\top/n$$

The matrix K captures population structure/correlation across different individuals.

- Kinship matrix; population admixture
- Human migration history

2. Linkage disequilibrium (LD)

a $p \times p$ matrix

$$R \approx X^\top X/n$$

The matrix R captures localized correlation patterns along the genomic axis within a chromosome.

- LD matrix
- The results of many, many recombination events

Recall: SVD captures principal components

$$X = UDV^\top$$

Recall: SVD captures principal components

$$X = UDV^\top$$

What is this?

$$\frac{1}{n}X^\top X = \frac{1}{n}VDU^\top UDV^\top = \frac{1}{n}VD^2V^\top$$

variant × variant

Recall: SVD captures principal components

$$X = UDV^\top$$

What is this?

$$\frac{1}{n}X^\top X = \frac{1}{n}VDU^\top UDV^\top = \frac{1}{n}VD^2V^\top$$

variant × variant

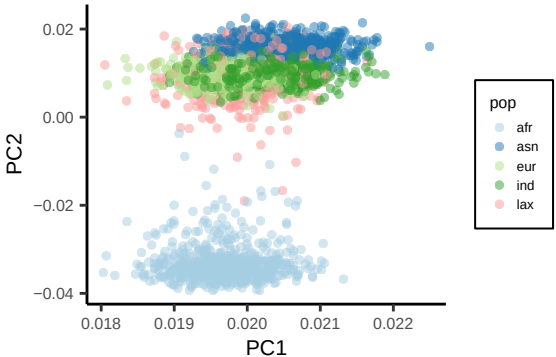
What is this?

$$\frac{1}{n}XX^\top = \frac{1}{n}UDV^\top VDU^\top = \frac{1}{n}UD^2U^\top$$

sample × sample

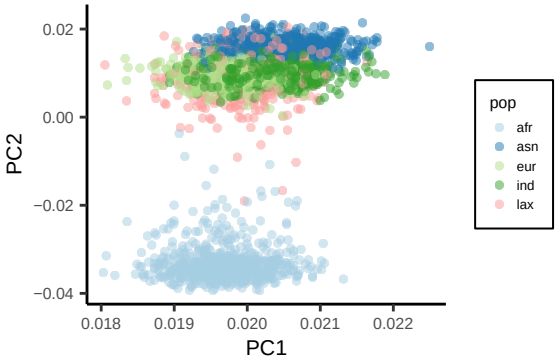
Let's take top 1000 most frequent variants

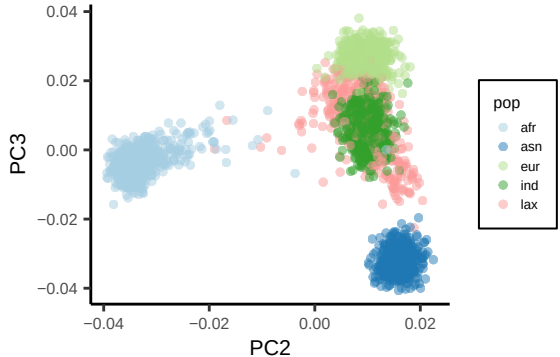
PCA already teaches us something interesting...



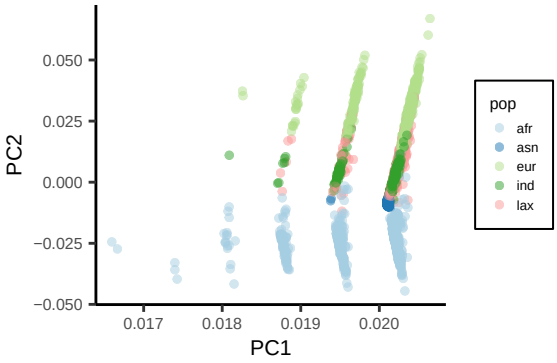
Let's take top 1000 most frequent variants

PCA already teaches us something interesting...

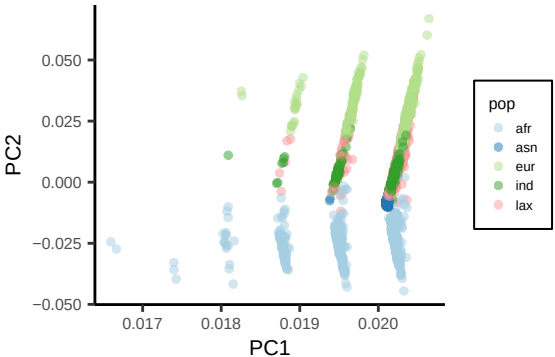


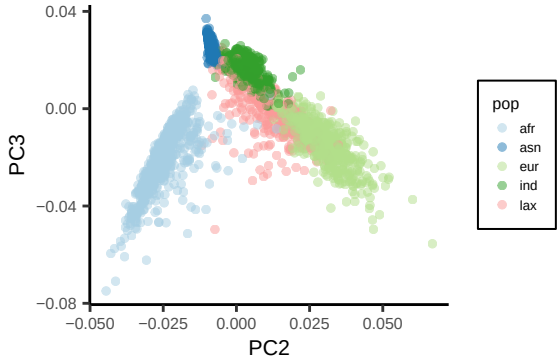


What about the 1000 least frequent variants?



What about the 1000 least frequent variants?





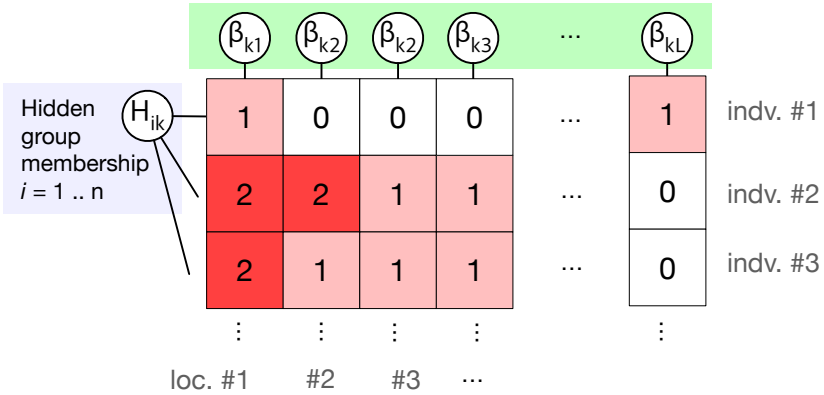
Why do we study population structures in human genetics?

- If there is no mutation/variation, there is no genetic association.
- Without knowing a macro-level dependency structures across cohorts, it is hard to dissect micro-level, perhaps disease-specific patterns.
- *Causal inference*: It also serves as a natural way to stratify/divide cohorts in a population genetics study to edify causal relationships that hold invariantly across multiple strata.
- *Precision health*: Characterization of population-invariant or specific variation is one of the first steps toward precision medicine.

An admixture model to identify hidden groups in a genotype matrix X

$H_{ik} \in (0, 1)$: hidden (probabilistic) membership of an individual i to a group k

$\beta_{kl} \in (0, 1)$: a group k-specific allele frequency in a locus l.

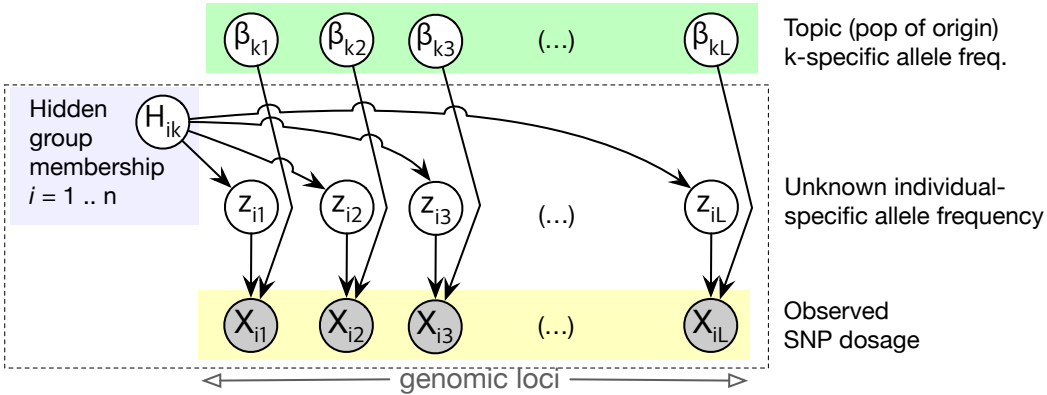


Related work: Pritchard, Stephens, Donnelly, *Genetics* (2000)

An admixture model to identify hidden groups in a genotype matrix X

$H_{ik} \in (0, 1)$: hidden (probabilistic) membership of an individual i to a group k

$\beta_{kl} \in (0, 1)$: a group k-specific allele frequency in a locus l.



Related work: Pritchard, Stephens, Donnelly, *Genetics* (2000)

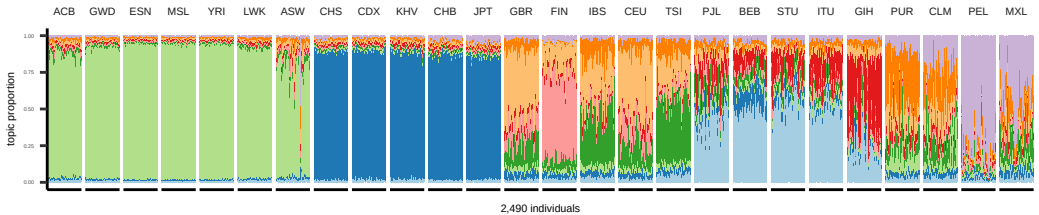
Population admixture learned from top 10k high MAF variants

A generative model:

- Sample each individual's topic proportion H_i
- Sample a topic membership for each variant j, say $Z_{ij} = k$ (could be implicitly handled)
- Genotype $X_{ij}|Z_{ij} = k \sim$ topic-specific β_{kj}

$$\mathcal{L} = \prod_{i=1}^{n} \prod_{j=1}^{10k} \left(\sum_{k=1}^{9} H_{ik}\beta_{kj} \right)^{X_{ij}}$$

topic proportion H



Related work: Pritchard, Stephens, Donnelly, *Genetics* (2000)

What is the benefit of learning an admixture model in GWAS data?

- Probabilistic interpretation of latent states
- Bayesian missing data imputation
- Potentially, a scalable approach for a biobank-scale data

How do we deal with such a population structure in GWAS?

1. Consider these population structures as a "backdoor" variable and adjust or include them in the regression model

How do we deal with such a population structure in GWAS?

1. Consider these population structures as a "backdoor" variable and adjust or include them in the regression model

- We need to first estimate the population structures... Are there any uncertainty? Can we propagate the measurement errors?

How do we deal with such a population structure in GWAS?

1. Consider these population structures as a "backdoor" variable and adjust or include them in the regression model

- We need to first estimate the population structures... Are there any uncertainty? Can we propagate the measurement errors?
- Which variants are okay to include in the latent topic model?

How do we deal with such a population structure in GWAS?

1. Consider these population structures as a "backdoor" variable and adjust or include them in the regression model

2. Treat such population structure-related effects as a "random" effect

- We need to first estimate the population structures... Are there any uncertainty? Can we propagate the measurement errors?
- Which variants are okay to include in the latent topic model?

How do we deal with such a population structure in GWAS?

1. Consider these population structures as a "backdoor" variable and adjust or include them in the regression model

- We need to first estimate the population structures... Are there any uncertainty? Can we propagate the measurement errors?
- Which variants are okay to include in the latent topic model?

2. Treat such population structure-related effects as a "random" effect

- We can include some proxy random variables for population structures in a linear GWAS model

How do we deal with such a population structure in GWAS?

1. Consider these population structures as a "backdoor" variable and adjust or include them in the regression model

- We need to first estimate the population structures... Are there any uncertainty? Can we propagate the measurement errors?
- Which variants are okay to include in the latent topic model?

2. Treat such population structure-related effects as a "random" effect

- We can include some proxy random variables for population structures in a linear GWAS model
- We might still need to consider uncertainty of the random effects

(digression) Useful facts on multivariate Gaussian distribution - 1

If we have $\mathbf{y}$

$$\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$$

then

$$\mathbb{E}[U^\top \mathbf{y}] = U^\top \mu, \quad \mathbb{V}[U^\top \mathbf{y}] = U^\top \Sigma U$$

and (affine transformation)

$$U^\top \mathbf{y} \sim \mathcal{N}(U^\top \mu, U^\top \Sigma U)$$

(digression) Useful facts on multivariate Gaussian distribution - 2

If we have two Gaussian random vectors, $\mathbf{y} \sim \mathcal{N}\left(\mu + \mathbf{u}, \Sigma_y\right)$ and $\mathbf{u} \sim \mathcal{N}(\mathbf{u}|\mathbf{0}, \Sigma_u)$

Bayesian integration:

$$\int \mathcal{N}\left(\mathbf{y}|\mu + \mathbf{u}, \Sigma_y\right) \mathcal{N}(\mathbf{u}|\mathbf{0}, \Sigma_u)\, d\mathbf{u} = \mathcal{N}\left(\mathbf{y}|\mu, \Sigma_y + \Sigma_u\right)$$

A key idea in the proof:

$$\left[\Sigma_y^{-1} - \Sigma_y^{-1}\left(\Sigma_y^{-1} + \Sigma_u^{-1}\right)^{-1}\Sigma_y^{-1}\right]^{-1} = \Sigma_y + \Sigma_u$$

by Woodbury identity.

A linear model with population-driven random effects

A linear regression model:

$$\mathbf{y} = \mathbf{x}_j \qquad \beta_j \qquad + \epsilon$$

a fixed genetic effect

What are we missing? Can we assume homo-scedasticity, i.e.,

$$\epsilon \overset{?}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

A linear model with population-driven random effects

A linear regression model:

$$\mathbf{y} = \mathbf{x}_j \quad \beta_j \quad + \epsilon$$

a fixed genetic effect

What are we missing? Can we assume homo-scedasticity, i.e.,

$$\epsilon \overset{?}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

A linear model with a random effect:

$$\mathbf{y} = \mathbf{x}_j \, \beta_j \; + \; \mathbf{u} \; + \epsilon$$

fixed

random effect

Note: There is no specific parameterization for this $n \times 1$ random vector $\mathbf{u}$. Now, we assume:

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

A linear model with population-driven random effects - 2

We want to capture unwanted population, cohort-specific random effects by $n \times 1$ vector $\mathbf{u}$ and **remove** since our **goal** is to estimate the fixed genetic effect of a particular variant j.

$$\mathbf{y} = \underset{\text{goal}}{\mathbf{x}_j \beta_j} + \underset{\text{remove}}{\mathbf{u}} + \epsilon$$

A linear model with population-driven random effects - 2

We want to capture unwanted population, cohort-specific random effects by $n \times 1$ vector $\mathbf{u}$ and **remove** since our **goal** is to estimate the fixed genetic effect of a particular variant j.

$$\mathbf{y} = \underset{\text{goal}}{\mathbf{x}_j \beta_j} + \underset{\text{remove}}{\mathbf{u}} + \epsilon$$

1. Note that $\mathbf{u}$ shouldn't be tied to a particular variant (by definition)

A linear model with population-driven random effects - 2

We want to capture unwanted population, cohort-specific random effects by $n \times 1$ vector $\mathbf{u}$ and **remove** since our **goal** is to estimate the fixed genetic effect of a particular variant j.

$$\mathbf{y} = \underset{\text{goal}}{\mathbf{x}_j \beta_j} + \underset{\text{remove}}{\mathbf{u}} + \epsilon$$

1. Note that $\mathbf{u}$ shouldn't be tied to a particular variant (by definition)
2. Also, the covariation of $\mathbf{u}$ is primarily driven by relatedness among individuals, not the variants.

$$\mathbf{u} \sim \mathcal{N}\left(\mathbf{0}, \tau^2 K\right), \quad K \approx \frac{1}{n} X X^\top$$

A linear mixed effect model (LMM) to test associations while adjusting population structure

We can define a hierarchical model:

$$\mathbf{y}|X,\beta,\mathbf{u},\sigma \sim \mathcal{N}(X\beta+\mathbf{u},\sigma^2 I) \quad (1)$$
$$\mathbf{u}|\tau,K \sim \mathcal{N}(\mathbf{0},\tau^2 K) \quad (2)$$

If we integrate out $\mathbf{u}$,

$$\mathbf{y}|X,\beta \sim \mathcal{N}\left(\mathbf{y}\middle| X\beta, \underbrace{\tau^2 K}_{\text{genetic-relatedness matrix}} + \underbrace{\sigma^2 I}_{\text{irreducible}}\right)$$

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects
- Cumbersome computation required for matrix factorization or other latent variable modelling on a large genotype matrix

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects
- Cumbersome computation required for matrix factorization or other latent variable modelling on a large genotype matrix
- We many not have a large matrix to learn about non-genetic confounders...

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects
- Cumbersome computation required for matrix factorization or other latent variable modelling on a large genotype matrix
- We many not have a large matrix to learn about non-genetic confounders...
- One LMM estimation can substitute multiple matrix factorization steps

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects
- Cumbersome computation required for matrix factorization or other latent variable modelling on a large genotype matrix
- We many not have a large matrix to learn about non-genetic confounders...
- One LMM estimation can substitute multiple matrix factorization steps
- We may have a good idea about relationships induced by random effects!

FaST Linear Mixed Model (Lippert *et al.* 2011)

We can resolve maximum likelihood estimate of the parameters, β, τ, σ,

$$\max \log \mathcal{N}(\mathbf{y} \,|\, X\beta, \sigma_2 \left(\delta K + I\right))$$

where $\tau^2 = \delta\sigma^2$.

Lippert, Listgarten, .. , Heckerman, *Nature Methods* (2011)

FaST Linear Mixed Model (Lippert *et al.* 2011)

We can resolve maximum likelihood estimate of the parameters, β, τ, σ,

$$\max \log \mathcal{N}(\mathbf{y} \,|\, X\beta, \sigma_2 \left(\delta K + I\right))$$

where $\tau^2 = \delta\sigma^2$.

We need to deal with this unfriendly form of likelihood:

$$-\frac{1}{2}\left(n \log(2\pi\sigma^2) + \log|I + \delta K| + \frac{1}{\sigma^2}\left[\mathbf{y} - X\beta\right]^\top \left(I + \delta K\right)^{-1}\left[\mathbf{y} - X\beta\right]\right)$$

Lippert, Listgarten, .. , Heckerman, *Nature Methods* (2011)

FaST Linear Mixed Model (Lippert *et al.* 2011)

Instead, we can transform the underlying distribution using spectral decomposition of the genetic-relatedness matrix (GRM),

$K = USU^\top$ where $U^\top U = I$, and S is a diagonal matrix.

$$\underbrace{U^\top \mathbf{y}}_{\text{projected output}} \sim \mathcal{N}\Big(\underbrace{U^\top X}_{\text{projected genotype}} \beta, \sigma^2 U^\top (I + \delta K) U\Big)$$

Lippert, Listgarten, .. , Heckerman, *Nature Methods* (2011)

FaST Linear Mixed Model (Lippert *et al.* 2011)

Instead, we can transform the underlying distribution using spectral decomposition of the genetic-relatedness matrix (GRM),

$K = USU^\top$ where $U^\top U = I$, and S is a diagonal matrix.

$$\underset{\text{projected output}}{U^\top \mathbf{y}} \sim \mathcal{N}\left(\underset{\text{projected genotype}}{U^\top X}\ \beta, \sigma^2 U^\top (I + \delta K) U\right)$$

$$\text{(by the affine transformation)} \sim \mathcal{N}\left(\underset{\text{projected genotype}}{U^\top X}\ \beta,\ \sigma^2 \underset{\text{diagonal matrix}}{(I + \delta S)}\right)$$

- We can find β by weighted least square
- We can find σ^2 and δ by fixing β

Lippert, Listgarten, .. , Heckerman, *Nature Methods* (2011)

A key research question in LMM: What covariance matrix?

If there were many types of random effects,

$$\mathbf{y} = \underbrace{X\beta}_{\text{fixed}} + \underbrace{\mathbf{u} + \mathbf{w} + ...}_{\text{random effects}} + \underbrace{\epsilon}_{\text{unknown}}$$

A key research question in LMM: What covariance matrix?

If there were many types of random effects,

$$\mathbf{y} = \underbrace{X\beta}_{\text{fixed}} + \underbrace{\mathbf{u} + \mathbf{w} + ...}_{\text{random effects}} + \underbrace{\epsilon}_{\text{unknown}}$$

We would need to many covariance matrices:

$$\mathbf{y}|\cdot \sim \mathcal{N}\left(X\beta,\, \sigma^2(I + \underbrace{\delta_u K_u + \delta_w K_w + ...}_{\text{random effects}})\right)$$

A key research question in LMM: What covariance matrix?

If there were many types of random effects,

$$\mathbf{y} = \underbrace{X\beta}_{\text{fixed}} + \underbrace{\mathbf{u} + \mathbf{w} + ...}_{\text{random effects}} + \underbrace{\epsilon}_{\text{unknown}}$$

We would need to many covariance matrices:

$$\mathbf{y}|\cdot \sim \mathcal{N}\left(X\beta,\, \sigma^2(I + \underbrace{\delta_u K_u + \delta_w K_w + ...}_{\text{random effects}})\right)$$

If we only care about variance decomposition $\beta_j \sim \mathcal{N}(0, \tau)$:

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{0},\, \sigma^2\left(\underbrace{\frac{\sigma^2_{\text{genetic}}}{n} XX^\top}_{\text{observed genetic}} + I + \underbrace{\delta_u K_u + \delta_w K_w + ...}_{\text{random effects}}\right)\right)$$

Should we worry about "over-fitting" in LMM?

An equivalent question for PCA-based confounder adjustment:

How many PCs to adjust in GWAS?
Can we include a candidate SNP in the GRM K matrix?

Should we worry about "over-fitting" in LMM?

An equivalent question for PCA-based confounder adjustment:

How many PCs to adjust in GWAS?
Can we include a candidate SNP in the GRM K matrix?

For each chromosome $c \in \{1, \ldots, 22, \mathrm{X}, \mathrm{Y}\}$, build a leave-one-chromosome-out (LOCO) kinship matrix, say K_{-c}:

$$\mathcal{N}\big(\mathbf{y}|\mathbf{x}_j\beta_j, \sigma^2(\delta K_{-c} + I)\big)$$

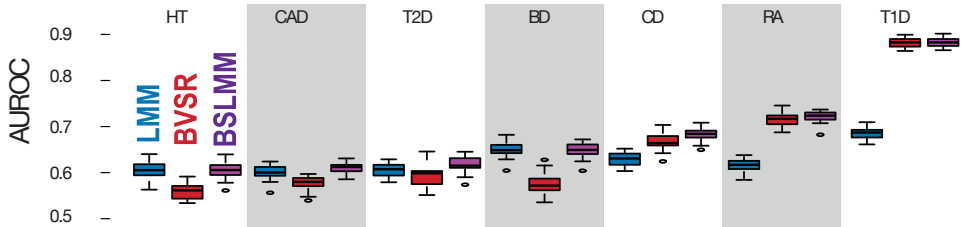
Yang, *et al.*, *Nature Genetics* (2014)

Tucker, Price, Berger, *Genetics* (2014)

BSLMM: What will be a good prior for the effect variable in a LMM?

Bayesian Sparse LMM (BSLMM): Causal variants should have a higher, additional level of effect size (σ_a^2) than the background ones (σ_b^2).

$$\beta_j \sim \pi \mathcal{N}\left(0, \frac{\sigma_a^2 + \sigma_b^2}{p\tau}\right) + (1-\pi)\mathcal{N}\left(0, \frac{\sigma_b^2}{p\tau}\right)$$



BVSR: spike-and-slab; BSLMM: mixture of two Gaussians; LMM: Gaussian

Zhou, Carbonetto, Stephens, *PLoS Genetics* (2014)

Today's lecture

Population structures in human genetics data

Linkage Disequilibrium: blessing and curse

Systems genetics and summary statistics-based inference

Much of human genetics problems centre on two covariance matrices

For a standardized $n \times p$ genotype matrix X,

1. Genetic relatedness matrix (GRM)

a $n \times n$ matrix

$$K \approx XX^\top/n$$

The matrix K captures population structure/correlation across different individuals.

- Kinship matrix; population admixture
- Human migration history

2. Linkage disequilibrium (LD)

a $p \times p$ matrix

$$R \approx X^\top X/n$$

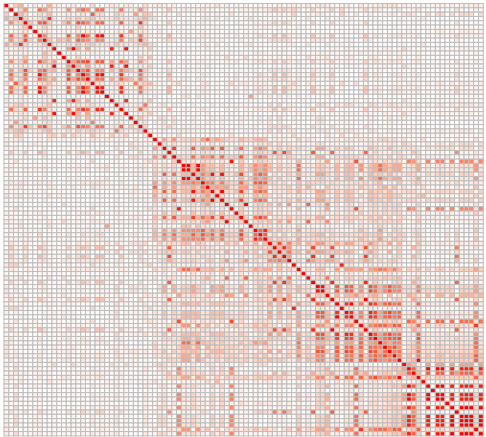
The matrix R captures localized correlation patterns along the genomic axis within a chromosome.

- LD matrix
- The results of many, many recombination events

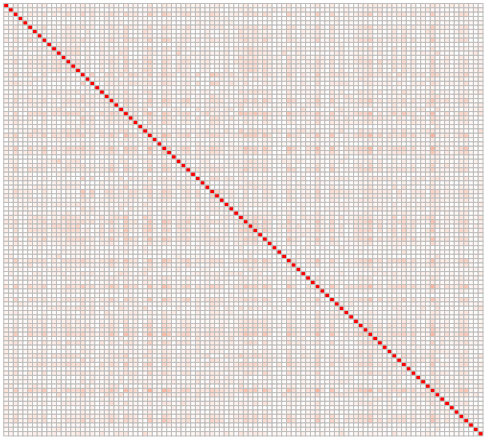
Let's discuss LD structures

Pairwise correlations between SNPs

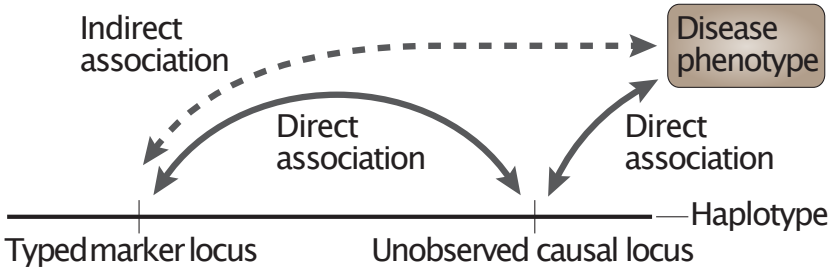
consecutive 100 SNPs



random 100 SNPs



GWAS fail to pinpoint exact locations associated with a disease



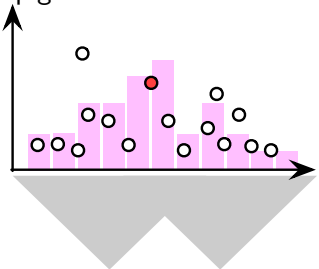
Balding, *Nature Genetics Review* (2006)

Common strategies to deal with LD structures

Strategy 1.

Fine-mapping to find a handful of causal ones

- Bayesian posterior estimation
- Overlap with epigenomics data



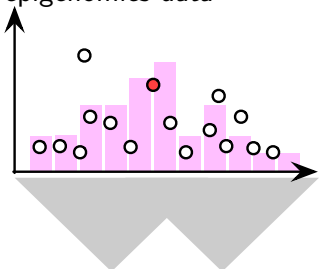
x-axis: genomic location; y-axis: -log10 p-value

Common strategies to deal with LD structures

Strategy 1.

Fine-mapping to find a handful of causal ones

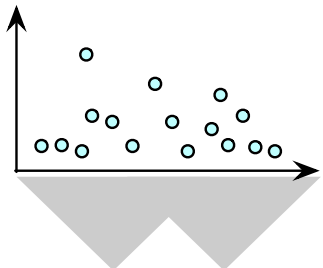
- Bayesian posterior estimation
- Overlap with epigenomics data



Strategy 2.

Aggregation to combine all the information:

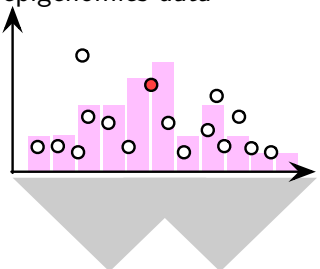
- Rare variant analysis
- Gene-level enrichment/association



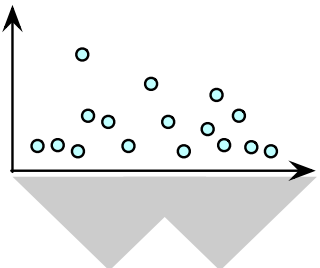
x-axis: genomic location; y-axis: -log10 p-value

Common strategies to deal with LD structures

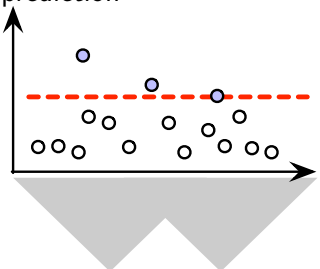
Strategy 1.
Fine-mapping to find a handful of causal ones
- Bayesian posterior estimation
- Overlap with epigenomics data



Strategy 2.
Aggregation to combine all the information:
- Rare variant analysis
- Gene-level enrichment/association



Strategy 3. **Pruning** to remove somewhat redundant information (heuristics)
- p-value thresholding
- Useful in polygenic risk prediction



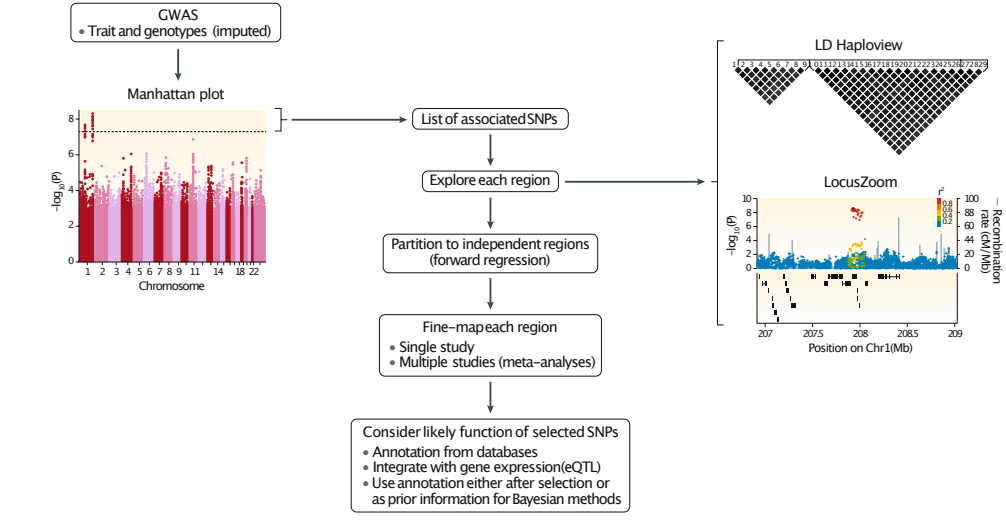
x-axis: genomic location; y-axis: -log10 p-value

Why fine-mapping?

A lead SNP[1] within a locus $\neq$ a causal SNP

[1] lowest p-value

Fine-mapping typically follows GWAS meta-analysis



Schaid *et al.*, *Nature Review Genetics* (2018)

Fine-mapping could be done by a variable selection problem

If we had fully observed X and Y for the $>$ 10k samples,

$$\mathbf{y} \sim \sum_{j=1}^{p} \mathbf{x}_j \beta_j + \epsilon$$

1. **A greedy forward selection method**

- $\mathbf{y} \sim \mathbf{x}_k \beta_k$ (find the best)
- $\mathbf{y} \leftarrow \mathbf{y} - \mathbf{x}_k \beta_k$ (take the residual)

(technically not a fine-mapping method)

Fine-mapping could be done by a variable selection problem

If we had fully observed X and Y for the $>$ 10k samples,

$$\mathbf{y} \sim \sum_{j=1}^{p} \mathbf{x}_j \beta_j + \epsilon$$

1. **A greedy forward selection method**
- $\mathbf{y} \sim \mathbf{x}_k \beta_k$ (find the best)
- $\mathbf{y} \leftarrow \mathbf{y} - \mathbf{x}_k \beta_k$ (take the residual)

(technically not a fine-mapping method)

2. **A brute-force combinatorial search**

For all possible subsets of non-zero's $S \subset [p]$:

$\max_S p(\mathbf{y} | \sum_{j \in S} \mathbf{x}_j \beta_j)$

(usually limit the search space $|S| < k$)

Fine-mapping could be done by a variable selection problem

If we had fully observed X and Y for the $>$ 10k samples,

$$\mathbf{y} \sim \sum_{j=1}^{p} \mathbf{x}_j \beta_j + \epsilon$$

1. **A greedy forward selection method**
- $\mathbf{y} \sim \mathbf{x}_k \beta_k$ (find the best)
- $\mathbf{y} \leftarrow \mathbf{y} - \mathbf{x}_k \beta_k$ (take the residual)

(technically not a fine-mapping method)

2. **A brute-force combinatorial search**
For all possible subsets of non-zero's $S \subset [p]$:
$\max_S p(\mathbf{y} | \sum_{j \in S} \mathbf{x}_j \beta_j)$

(usually limit the search space $|S| < k$)

3. **Bayesian prior**, L1 or spike-slab
- with a sparse prior $p(\beta)$

$$\min \|\mathbf{y} - X\beta\|^2 - \log p(\beta)$$

(normalization is required)

In practice, we don't have a full panel of genotypes!

But we have summary statistics of meta-analysis:

A generative model of SNP-level statistics

For each β_j's, effect size, variance, z-score:

$$\hat{\beta}_j = \frac{\sum_i X_{ij} Y_i}{\sum_i X_{ij}^2}, \quad \hat{\mathbb{V}}[\beta_j] = \frac{\sigma_\epsilon^2}{\sum_{i=1} X_{ij}^2}, \quad Z_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\mathbb{V}}[\beta_j]}}$$

In practice, we don't have a full panel of genotypes!

But we have summary statistics of meta-analysis:

A generative model of SNP-level statistics

For each β_j's, effect size, variance, z-score:

$$\hat{\beta}_j = \frac{\sum_i X_{ij} Y_i}{\sum_i X_{ij}^2}, \quad \hat{\mathbb{V}}[\beta_j] = \frac{\sigma_\epsilon^2}{\sum_{i=1} X_{ij}^2}, \quad Z_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\mathbb{V}}[\beta_j]}}$$

Although **the underlying multivariate regression model**:

$$\mathbf{y} = X\theta + \epsilon$$

where $\theta_j \neq \beta_j$

What is the relationship between the summary (univariate) and multivariate effects?

For simplicity, let's assume standardized genotype matrix X, i.e., $\bar{X}_j = 0$ and $\hat{\sigma}^2_{X_j} = 1$. The we have z-score

$$\hat{Z}_j = \sum_{i=1}^{n} X_{ij} Y_i / \sigma_\epsilon \sqrt{n}$$

for all $j \in [p]$.

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

What is the relationship between the summary (univariate) and multivariate effects?

For simplicity, let's assume standardized genotype matrix X, i.e., $\bar{X}_j = 0$ and $\hat{\sigma}^2_{X_j} = 1$. The we have z-score

$$\underset{p \times 1 \text{ univariate}}{\mathbf{z}} = \frac{1}{\sigma\sqrt{n}} X^\top \mathbf{y}$$

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

What is the relationship between the summary (univariate) and multivariate effects?

For simplicity, let's assume standardized genotype matrix X, i.e., $\bar{X}_j = 0$ and $\hat{\sigma}^2_{X_j} = 1$. The we have z-score

$$\underset{p \times 1 \text{ univariate}}{\mathbf{z}} = \frac{1}{\sigma\sqrt{n}} X^\top \mathbf{y} = \frac{1}{\sigma\sqrt{n}} X^\top \underbrace{(X\theta + \epsilon)}_{\text{a multivariate model}}$$

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

What is the relationship between the summary (univariate) and multivariate effects?

For simplicity, let's assume standardized genotype matrix X, i.e., $\bar{X}_j = 0$ and $\hat{\sigma}^2_{X_j} = 1$. The we have z-score

$$\underset{p \times 1 \text{ univariate}}{\mathbf{z}} = \frac{1}{\sigma\sqrt{n}} X^\top \mathbf{y} = \frac{1}{\sigma\sqrt{n}} X^\top \underbrace{(X\theta + \epsilon)}_{\text{a multivariate model}}$$

$$= \frac{\sqrt{n}}{\sigma} \underbrace{\left(\frac{1}{n} X^\top X\right)}_{\text{LD}} \theta + \frac{1}{\sigma\sqrt{n}} X^\top \epsilon$$

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

What is the relationship between the summary (univariate) and multivariate effects?

For simplicity, let's assume standardized genotype matrix X, i.e., $\bar{X}_j = 0$ and $\hat{\sigma}^2_{X_j} = 1$. The we have z-score

$$\underset{p \times 1 \text{ univariate}}{\mathbf{z}} = \frac{1}{\sigma\sqrt{n}} X^\top \mathbf{y} = \frac{1}{\sigma\sqrt{n}} X^\top \underbrace{(X\theta + \epsilon)}_{\text{a multivariate model}}$$

$$= \mathbf{R}\frac{\sqrt{n}}{\sigma}\theta + \tilde{\epsilon}, \quad \tilde{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$$

where $\mathbf{R} = n^{-1}X^\top X$ is an empirical LD matrix.

Hormozdiari *et al.*, Genetics (2014); Zhu and Stephens, Annals of Applied Statistics (2017)

Fine-mapping is to find a sparse multivariate θ

Input:

- Summary statistics $p \times 1$ z-score vector: $\mathbf{z}$
- Reference panel $p \times p$ LD matrix: R

Goal:

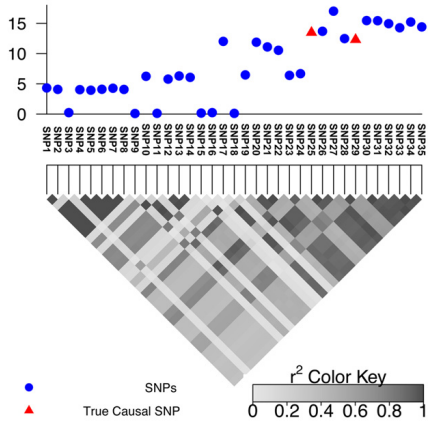
$$\max_{\theta} \mathcal{N}\left(\mathbf{z} \middle| \frac{\sqrt{n}}{\sigma} R\theta, R\right)$$

where

$$\theta_j \sim \pi\delta_0(\theta_j) + (1-\pi)\mathcal{N}(0, \tau^{-1})$$

Zhu and Stephens, Annals of Applied Statistics (2017)

Not all GWA-significant variants are causal



A reasonable fine-mapping approach:

- Identify GWAS loci (with p $< 5 \times 10^{-8}$)
- Find neighbouring SNPs in each GWAS locus
- Convert p-values to z-scores (caution: We should take into account major/minor allele directions)
- Take an appropriate local LD matrix R
- Estimate θ in the following model:

$$\mathbf{z} \sim \mathcal{N}(R\theta, R)$$

- We may construct 95% credible set:

$$\{j : \hat{p}(\theta_j \neq 0 | R, \mathbf{z}) > .95\}$$

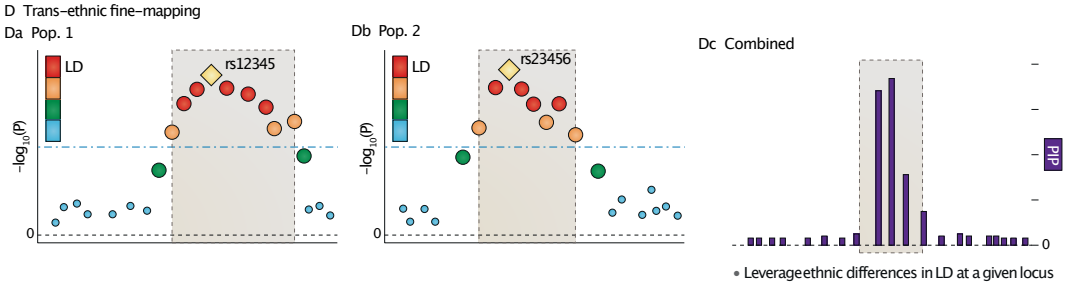
Hormozdiari *et al.*, Genetics (2014)

Fine-mapping approaches - 1: heuristics and Bayesian approach

- In almost all cases, Bayesian method outperforms
- Note: This is a "statistical" fine-mapping

Schaid *et al.*, *Nature Review Genetics* (2018)

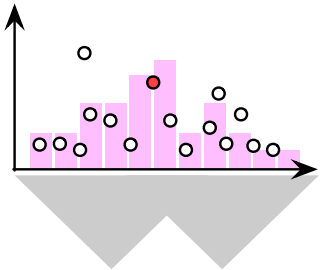
Fine-mapping approaches - 2: *trans*-ethnic analysis



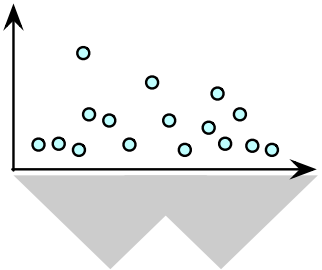
- Additional information across multiple GWAS summary statistics
- "Causal triangulation" to combine multiple lines of orthogonal evidence

Schaid *et al.*, *Nature Review Genetics* (2018)

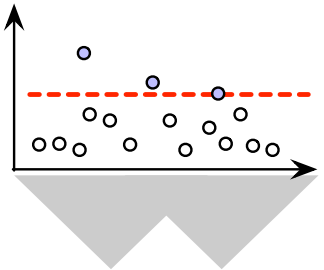
How should we deal with LD structures?



fine-mapping



Aggregate information within an independent LD block!



will not discuss this

Burden test: Can we aggregate all the SNPs to boost power?

Motivation:

- Summary statistics $p \times 1$ z-score vector: $\mathbf{z}$
- Reference panel $p \times p$ LD matrix: R
- Unfortunately, none of the SNPs make GWA significance

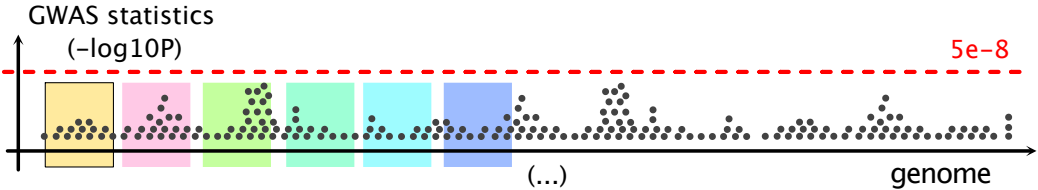
Question:

- Should we give up on this GWAS result (z-scores)?
- Alternatively, is there any way to reduce the number of hypothesis?

Gene-level aggregate test statistics

Assume underlying multivariate effect θ for the observed z-score vector:

$$\mathbf{z} \sim \mathcal{N}(R\theta, R)$$



- Can we aggregate information over many SNPs within each gene (box)?
- Although each genetic variant can occur rarely (hence, very weak association statistics), they may implicate the same target gene.

Sequence Kernel Association Test (SKAT) to aggregate rare variant info

In a model $\mathbf{y} \sim X^{(g)}\theta_g$, where X is constructed within a window around a specific gene g, we want to test

$$H_0 : \theta_g = 0 \quad \text{vs.} \quad H_1 : \theta_g = 0$$

In a sense, it is the same as doing model comparison (after integrate out θ):

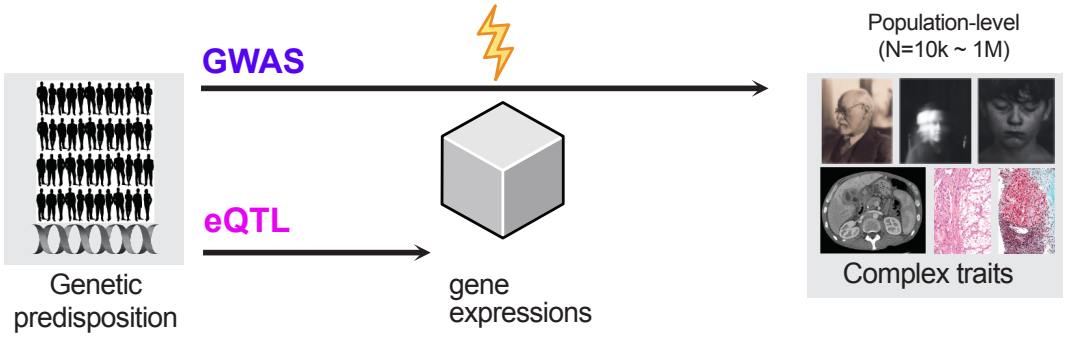
$$H_0 : \mathcal{N}\left(\mathbf{y} \middle| \mathbf{0}, \underbrace{\tau^2 \mathbf{K}}_{\text{e.g., local kinship}} + \sigma^2 I\right) \quad \text{vs.} \quad H_1 : \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2 I)$$

where $K \propto n^{-1} X^{(g)} {X^{(g)}}^\top$ or we can substitute it with a different type of kernel matrix. $\implies$ We want to test $H_0 : \tau = 0$ or not.

Wu, Lee, ..., Lin, *AJHG* (2011)

Can we aggregate genetic association statistics using prior knowledge?

Expression Quantitative Trait Loci (eQTL) results provide necessary context to interpret GWAS signals



$$\mathbf{m}_g \sim X\alpha_g + \epsilon_g \quad \Longrightarrow \quad \mathbf{y}_{\text{GWAS}} = \sum_{g \in \text{causal genes}} \mathbf{m}_g \beta_g + \epsilon_y$$

Transcriptome-wide association study to test gene-level correlations

The same type of a linear model for a phenotype vector $\mathbf{y}$

$$\mathbf{y} \sim X\theta + \epsilon_y,$$

where X is constructed within a cis-window around a specific gene g (say $\pm$ 500kb).

Gusev, ..., Price, *Nature Genetics* (2016)

Transcriptome-wide association study to test gene-level correlations

The same type of a linear model for a phenotype vector $\mathbf{y}$

$$\mathbf{y} \sim X\theta + \epsilon_y,$$

where X is constructed within a cis-window around a specific gene g (say $\pm$ 500kb).

We also have a gene expression vector $\mathbf{m}$:

$$\mathbf{m}_g \sim X\alpha_g + \epsilon_m$$

Gusev, ..., Price, *Nature Genetics* (2016)

Transcriptome-wide association study to test gene-level correlations

The same type of a linear model for a phenotype vector $\mathbf{y}$

$$\mathbf{y} \sim X\theta + \epsilon_y,$$

where X is constructed within a cis-window around a specific gene g (say $\pm$ 500kb).

We also have a gene expression vector $\mathbf{m}$:

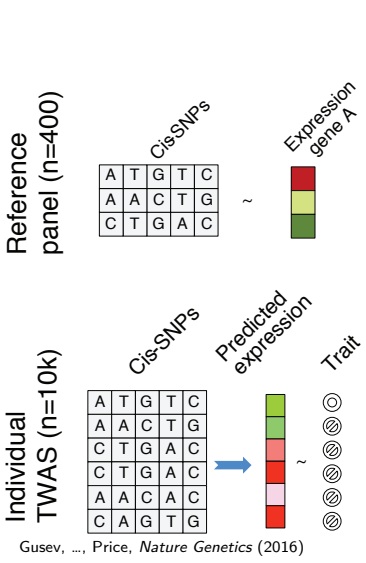
$$\mathbf{m}_g \sim X\alpha_g + \epsilon_m$$

A key question: Are they correlated?

$$H_0 : \mathbb{E}[\mathbf{y}^\top \mathbf{m}] = 0 \quad \text{vs.} \quad H_1 : \mathbb{E}[\mathbf{y}^\top \mathbf{m}] \neq 0$$

Gusev, ..., Price, *Nature Genetics* (2016)

Transcriptome-wide association study to test gene-level correlations

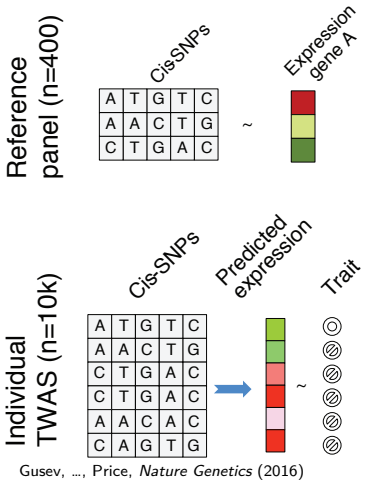


Gusev, ..., Price, *Nature Genetics* (2016)

Goal: hypothesis testing of non-zeroness

$$\frac{1}{n}\mathbf{m}^\top \mathbf{y} \quad = \quad \frac{1}{n}(X\alpha + \epsilon_m)^\top (X\theta + \epsilon_y)$$

Transcriptome-wide association study to test gene-level correlations

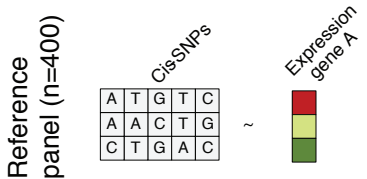


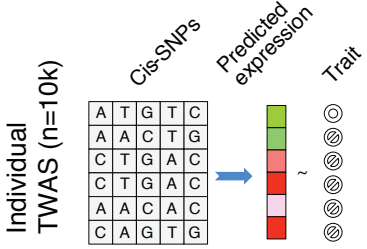
Gusev, ..., Price, *Nature Genetics* (2016)

Goal: hypothesis testing of non-zeroness

$$\begin{aligned}\frac{1}{n}\mathbf{m}^\top\mathbf{y} &= \frac{1}{n}(X\alpha+\epsilon_m)^\top(X\theta+\epsilon_y)\\ &= \alpha^\top\Big(\underbrace{\frac{1}{n}X^\top X}_{LD}\Big)\theta+\ldots\end{aligned}$$

Transcriptome-wide association study to test gene-level correlations





Gusev, ..., Price, *Nature Genetics* (2016)

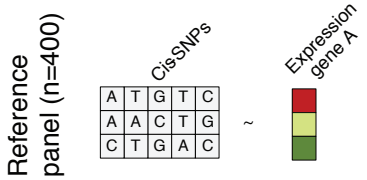
Goal: hypothesis testing of non-zeroness

$$\frac{1}{n}\mathbf{m}^\top \mathbf{y} = \frac{1}{n}(X\alpha + \epsilon_m)^\top (X\theta + \epsilon_y)$$

$$= \alpha^\top \underbrace{\left(\frac{1}{n}X^\top X\right)}_{LD}\theta + \ldots$$

(we saw this) $$= \alpha^\top \underbrace{\left(\frac{1}{n}X^\top X\right)}_{\text{GWAS z-score}}\theta + \ldots$$

Transcriptome-wide association study to test gene-level correlations



Gusev, ..., Price, *Nature Genetics* (2016)

Goal: hypothesis testing of non-zeroness

$$\begin{aligned}\frac{1}{n}\mathbf{m}^\top\mathbf{y} &= \frac{1}{n}(X\alpha+\epsilon_m)^\top(X\theta+\epsilon_y)\\ &= \alpha^\top\underbrace{\left(\frac{1}{n}X^\top X\right)}_{LD}\theta+\ldots\\ \text{(we saw this)} &= \alpha^\top\underbrace{\left(\frac{1}{n}X^\top X\right)}_{\text{GWAS z-score}}\theta+\ldots\end{aligned}$$

TWAS statistic:

$$T_g = \frac{\alpha_g^\top\mathbf{z}}{\sqrt{\alpha_g^\top R\alpha_g}} \sim \mathcal{N}(0,1)$$

where α_g = multivariate eQTL for a gene g.

For a standardized $n \times p$ genotype matrix X,

1. Genetic relatedness matrix (GRM)

a $n \times n$ matrix

$$K \approx XX^\top/n$$

The matrix K captures population structure/correlation across different individuals.

- Kinship matrix; population admixture
- Human migration history

2. Linkage disequilibrium (LD)

a $p \times p$ matrix

$$R \approx X^\top X/n$$

The matrix R captures localized correlation patterns along the genomic axis within a chromosome.

- LD matrix
- The results of many, many recombination events

Today's lecture

Population structures in human genetics data

Linkage Disequilibrium: blessing and curse

Systems genetics and summary statistics-based inference

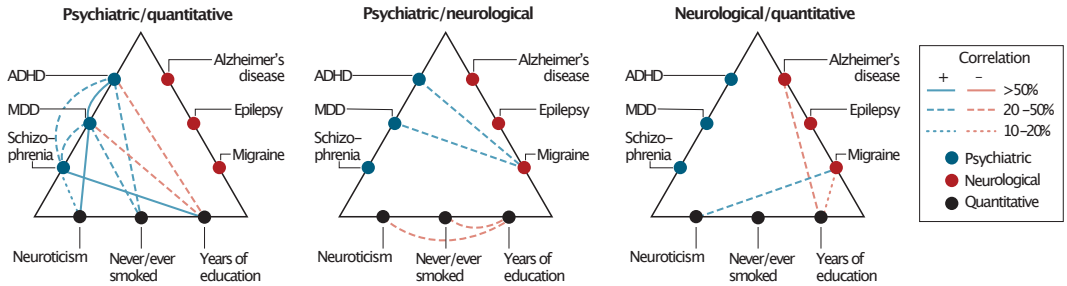
GWAS is only the beginning of a post-GWAS analysis.

Post-GWAS analysis example: genetic correlations across many traits

Psychiatric disorders				Neurological disorders			
Disorder	**Source**	**Cases**	**Controls**	**Disorder**	**Source**	**Cases**	**Controls**
Attention deficit hyperactivity disorder	PGC-ADD2	12,645	84,435	Alzheimer's disease	IGAP	17,008	37,154
Anorexia nervosa	PGC-ED	3495	10,982	Epilepsy	ILAE	7779	20,439
Anxiety disorders	ANGST	5761	11,765	Focal epilepsy	"	4601*	17,985*
Autism spectrum disorder	PGC-AUT	6197	7377	Generalized epilepsy	"	2525*	16,244*
Bipolar disorder	PGC-BIP2	20,352	31,358	Intracerebral hemorrhage	ISGC	1545	1481
Major depressive disorder	PGC-MDD2	66,358	153,234	Ischemic stroke	METASTROKE	10,307	19,326
Obsessive-compulsive disorder	PGC-OCDTS	2936	7279	Cardioembolic stroke	"	1859*	17,708*
Posttraumatic stress disorder	PGC-PTSD	2424	7113	Early onset stroke	"	3274*	11,012*
Schizophrenia	PGC-SCZ2	33,640	43,456	Large-vessel disease	"	1817*	17,708*
Tourette syndrome	PGC-OCDTS	4220	8994	Small-vessel disease	"	1349*	17,708*
				Migraine	IHGC	59,673	316,078
				Migraine with aura	"	6332*	142,817*
				Migraine without aura	"	8348*	136,758*
				Multiple sclerosis	IMSGC	5545	12,153
				Parkinson's disease	IPDGC	5333	12,019
Total psychiatric		158,028	365,993	Total neurologic		107,190	418,650

The Brainstorm Consortium, *Science* (2018)

Post-GWAS analysis example: genetic correlations across many traits



How did they measure correlations?

The Brainstorm Consortium, *Science* (2018)

Several benefits of post-GWAS (or systems genetics) analysis

- Inherited genetic information is usually stable over a lifetime.
- It is hard to test/measure all the disease-related phenotypes for all individuals.
- Full, unlimited access to individual-level information is often unnecessary in a post-GWAS analysis.
- A post-GWAS analysis is often computationally more efficient than an individual-level analysis.

LD-score regression: a model-based summary-statistics analysis

What is a generative model for a $\chi^2_j (\equiv Z^2_j)$ statistics vector?

We have seen that

$$Z_j = \frac{\sqrt{n}}{\sigma} \sum_k R_{jk} \theta_k + \epsilon_j$$

where $\epsilon \sim \mathcal{N}(0, 1)$.

Bulik-Sullivan *et al.*, *Nature Genetics* (2014); Finucane *et al.*, *Nature Genetics* (2015)

LD-score regression: a model-based summary-statistics analysis

What is a generative model for a $\chi^2_j (\equiv Z^2_j)$ statistics vector?

$$\mathbb{E}\left[\chi^2_j\right] = \mathbb{E}\left[Z^2_j\right] = \mathbb{E}\left(\sqrt{n}\sum_k R_{jk}\theta_k + \epsilon_j\right)^2$$

Bulik-Sullivan *et al.*, *Nature Genetics* (2014); Finucane *et al.*, *Nature Genetics* (2015)

LD-score regression: a model-based summary-statistics analysis

What is a generative model for a $\chi^2_j (\equiv Z^2_j)$ statistics vector?

$$\mathbb{E}\left[\chi^2_j\right] = \mathbb{E}\left[Z^2_j\right] = \mathbb{E}\left(\sqrt{n}\sum_k R_{jk}\theta_k + \epsilon_j\right)^2$$

If "true" multivariate effect for each variant is independent of other variants' effects, i.e., $\mathbb{E}\left[\theta_k\theta_j\right] = 0$ for all $k \neq j$,

$$\mathbb{E}\left[\chi^2_j\right] = n\underbrace{\sum_k R^2_{jk}}_{\text{LD-score}}\mathbb{E}\left[\theta^2_k\right] + 1$$

Bulik-Sullivan *et al.*, *Nature Genetics* (2014); Finucane *et al.*, *Nature Genetics* (2015)

Baseline LD-score regression to measure polygenic heritability

If we assume $\theta_k = \tau/p$, assuming all the variants can exert weak effects, and defining an LD score for a variant j, $l_j = \sum_k R_{jk}^2$, we have

$$\mathbb{E}\left[\chi_j^2\right] = n \underbrace{\sum_k R_{jk}^2}_{\text{LD-score}} \mathbb{E}\left[\theta_k^2\right] + 1$$

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Baseline LD-score regression to measure polygenic heritability

If we assume $\theta_k = \tau/p$, assuming all the variants can exert weak effects, and defining an LD score for a variant j, $l_j = \sum_k R_{jk}^2$, we have

$$\mathbb{E}\left[\chi_j^2\right] = n \underbrace{\sum_k R_{jk}^2}_{\text{LD-score}} \mathbb{E}\left[\theta_k^2\right] + 1 = \underset{\text{sample size}}{n} \underset{\text{LD score}}{l_j} \underset{\text{per SNP heritability}}{\frac{\tau}{p}} + 1$$

where p is the total number of SNPs.

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Baseline LD-score regression to measure polygenic heritability

If we assume $\theta_k = \tau/p$, assuming all the variants can exert weak effects, and defining an LD score for a variant j, $l_j = \sum_k R^2_{jk}$, we have

$$\mathbb{E}\left[\chi^2_j\right] = \underbrace{n}_{\text{sample size}} \underbrace{l_j}_{\text{LD score}} \underbrace{\frac{\tau}{p}}_{\text{per SNP heritability}} + 1$$

We estimate the parameters by regressing the observed χ^2 statistics on the reference LD scores l_j:

$$\begin{pmatrix} \chi^2_1 \\ \vdots \\ \chi^2_j \\ \vdots \end{pmatrix} \sim n \begin{pmatrix} l_1 \\ \vdots \\ l_j \\ \vdots \end{pmatrix} \underbrace{\frac{\tau}{p}}_{\text{per SNP heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

Bulik-Sullivan *et al.*, *Nature Genetics* (2014)

Stratified LD-score regression

When genome is partitioned by annotations (e.g., epigenetic tracks)

$$\mathbb{E}\left[\chi_j^2\right] = \frac{n}{p}\sum_t l_{jt}\underbrace{\tau_t}_{\text{stratified heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

where we use partitioned LD-scores for each annotation type t

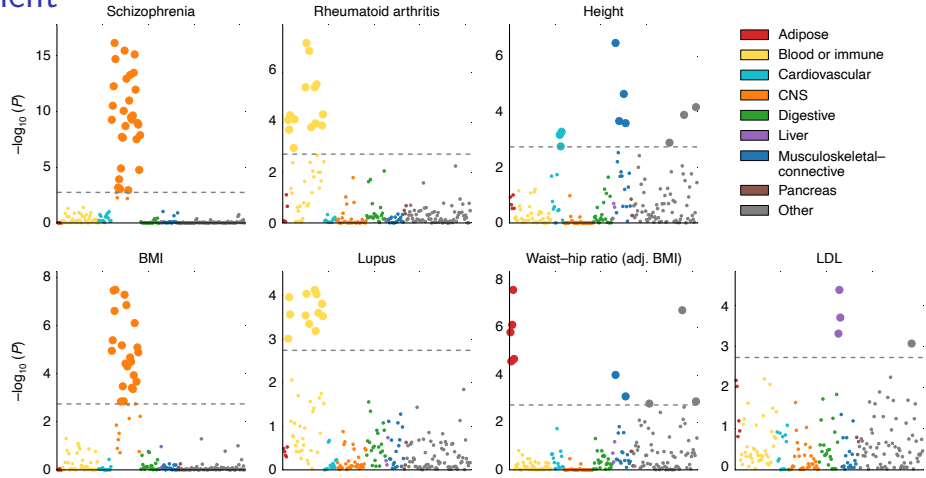
$$l_{jt} = \sum_k R_{jk}^2 I\left\{k \in \mathcal{A}_t\right\}.$$

More explicitly,

$$\begin{pmatrix} \chi_1^2 \\ \vdots \\ \chi_j^2 \\ \vdots \end{pmatrix} \sim \frac{n}{p}\begin{pmatrix} l_{11} & l_{12} & l_{1t} & \cdots \\ & \vdots & & \\ l_{j1} & l_{j2} & l_{jt} & \cdots \\ \vdots & & & \end{pmatrix} \underbrace{\begin{pmatrix} \tau_1 \\ \vdots \\ \tau_t \\ \vdots \end{pmatrix}}_{\text{stratified heritability}} + \underbrace{n\phi}_{\text{genomic inflation}} + \underbrace{1}_{\text{null}}$$

Finucane *et al.*, *Nature Genetics* (2015)

Stratified LD-score regression can identify tissue-specific heritability enrichment



Finucane *et al.*, *Nature Genetics* (2018)

Bivariate LD-score regression

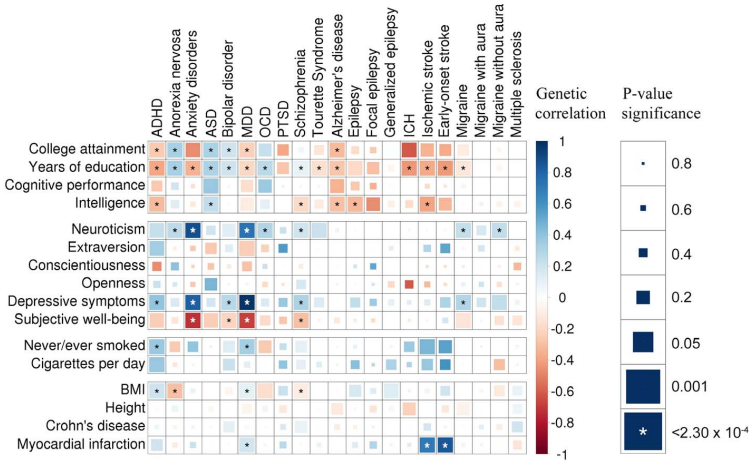
Instead of one χ^2 vector, we need to deal with the element-wise product of two vectors of z-scores (between a trait 1 and 2):

$$\begin{pmatrix} z_1^{(1)} z_1^{(2)} l_1 \\ \vdots \\ z_j^{(1)} z_j^{(2)} l_j \\ \vdots \end{pmatrix} \sim \frac{\sqrt{N_1 N_2}}{p} \begin{pmatrix} l_1 \\ \vdots \\ l_j \\ \vdots \end{pmatrix} \underset{\text{genetic correlation}}{\rho} + \underset{\text{sample sharing}}{\frac{\rho_0 N_s}{\sqrt{N_1 N_2}}}$$

where N_1 and N_2 count sample size of the GWAS 1 and 2; N_s is the number of control individuals shared between the two traits.

Bulik-Sullivan *et al.*, *Nature Genetics* (2015)

Bivariate LD-score regression to test genetic correlations



The Brainstorm Consortium, *Science* (2018)

Learning objective

- Population structures in genetics data
  - Admixture model
  - Linear mixed effect model
- Linkage disequilibrium
  - Rare variant burden tests
  - Fine-mapping causal variants
- GWAS summary statistics
  - Transcriptome-wide association studies
  - LD-score regression