Tissue-specific Polygenic Risk Score

Tuo Wang

Outline:

- Short introduction about polygenic risk score
- How to formulate Tissue-specific polygenic risk score
- Results

Let's review some concepts

SNPs

SNP is the unit of genetic variation, which is a single base-pair changes in the DNA sequence that occur with high frequency in the human genome.

Annotations:

Functional annotations is the process of identifying a gene's biological roles and subcellular location.

GWAS summary statistics

$\hat{\beta}_j$ is least square estimator of Y regressing on X_j, where X_j is the jth column of X.

Linkage disequilibrium:

LD is a property of SNPs on a contiguous stretch of genomic sequence that describes the degree to which an allele of one SNP is inherited or correlated with an allele of another SNP within a population Mathematically, it is the correlation between SNPs.

polygenic risk score

Genome-wide association studies (GWAS) :

$$Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{iM}\beta_M + \epsilon_i, \qquad i = 1, \cdots, N$$

Matrix form: $Y = X\beta + \epsilon$

- ϵ_i is the error term.
- Y_i is the response variable for individual i
- $X_{i1} \cdots X_{iM}$ are single nucleotide polymorphisms (SNPs). SNP is the unit of genetic variation, which is a single base-pair changes in the DNA sequence that occur with high frequency in the human genome.

Polygenic risk score:

$$PRS_i = \sum_{j=1}^{m} X_{ij} w_j$$

How to find w_j's?

polygenic risk score

Model:

$$Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{iM}\beta_M + \epsilon_i\,, \qquad i = 1, \cdots, N$$

Matrix form: $Y = X\beta + \epsilon$

GWAS summary statistics: $\hat{\beta}_j$ is least square estimator of Y regressing on X_j, where X_j is the jth column of X. Assume we M SNPs, then we will end up with $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_M$.

Polygenic risk score:

$$\mathrm{PRS_i} = \sum_{j=1}^{M} X_{i1}\,\hat{\beta}_j$$

Tissue-specific polygenic risk score

Integrating function annotations data could generate biological insights.

Assume we know which categories or tissues the SNPs belong to. For example, $S_1, \cdots, S_K$ are K categories or tissues, $S_1 = \{1,3,5\}$

Tissue-specific polygenic risk score:

$$PRS_{i,k} = \sum_{j=1}^{m} X_{i,j} w_{j,k}$$

Tissue-specific polygenic risk score

Why is tissue-specific polygenic risk score useful:

Because we want to know the pathway of a certain trait.

For example, for obesity, there may be two reasons. 1. addiction to food, which associated with brain 2. digestive problem, which associated with digestive system.

Suppose Tom and Jerry both have obesity, and the tissue specific PRS is:

	PRS for brain	PRS for digestive system
Tom	0.3	0.005
Jerry	0.007	0.25

The ultimate goal for Tissue-specific PRS: precision medicine

formulation

Model:

$$\mathrm{Y_i} = \mathrm{X_{i1}}\beta_1 + \mathrm{X_{i2}}\beta_2 + \cdots + \mathrm{X_{iM}}\beta_\mathrm{M} + \epsilon_\mathrm{i}$$

SNPS: $j = 1, \cdots, M$

Classes: $k = 1, \cdots, K$

Individuals: $i = 1, \cdots, N$

Assumption: $\beta_j = \sum_{k=1}^{K} \beta_{j,k} I(j \in S_k)$

Priori:

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_M \end{pmatrix} \sim MVN\left(0, \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_M^2 \end{pmatrix}\right)$$

$$\beta_{j,k} \sim \begin{cases} N\left(0, \sigma_{j,k}^2\right), & if\ j \in S_k \\ 0, & if\ j \notin S_k \end{cases}$$

Note that: σ_j^2 and $\sigma_{j,k}^2$ are known. They are functions of per-SNP heritability

We use a Bayesian framework to estimate the **posterior expectation** $\boldsymbol{E\left(\beta_{j,k} \middle| \widehat{\beta}_j\right) \overset{\text{def}}{=} \widetilde{\beta}_{j,k}}$

Tissue-specific polygenic risk score: $\boldsymbol{\mathrm{PRS}_{\mathrm{i},k} = \sum_{\mathrm{j}=1}^{m} X_{\mathrm{i},j} \widetilde{\beta}_{j,k}}$

formulation

Priori:

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_M \end{pmatrix} \sim MVN\left(0, \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_M^2 \end{pmatrix}\right)$$

$$\beta_{j,k} \sim \begin{cases} N\left(0, \sigma_{j,k}^2\right), & if\ j \in S_k \\ 0, & if\ j \notin S_k \end{cases}$$

Note that: σ_j^2 and $\sigma_{j,k}^2$ are known. They can functions of per-SNP heritability

Define τ_k is the per-SNP heritability in Category k. $\sigma_{j,k}^2$

$\sigma_j^2 = \sum_{k=1}^{K} \tau_k I(j \in S_k)$; $\sigma_{j,k}^2 = f_k(\tau_1, \cdots, \tau_K)$ has no relationship with SNP j, for example $\sigma_{j,k}^2 = \tau_k$

Thus,

$$\sigma_{1,k}^2 = \sigma_{2,k}^2 = \cdots = \sigma_{M,k}^2 = \sigma_{\cdot,k}^2$$

Formulation, No linkage disequilibrium (LD)

Linkage disequilibrium: correlation between SNPs. When there is no LD, $Cor(X_i, X_j) = 0$.

Assume no LD:

To estimate the posterior expectation $\tilde{\beta}_{j,k} = E(\beta_{j,k} | \hat{\beta}_j)$, we need to know the posterior distribution

$$\beta_{j,k}|\hat{\beta}_j \sim \begin{cases} 0, & \text{if } j \notin S_k \\ f(\beta_{j,k}|\hat{\beta}_j), & \text{if } j \in S_k \end{cases}$$

From Bayes rule, we know:

$$f(\beta_{j,k}|\hat{\beta}_j) \propto f(\hat{\beta}_j|\beta_{j,k}) f(\beta_{j,k})$$

Formulation, No linkage disequilibrium (LD)

For SNP j, assume $j \in S_c$, distribution of $\hat{\beta}_j|\beta_{j,c}$:

$$\hat{\beta}_j = \frac{1}{N} X_j^T Y = \beta_j + \frac{1}{N} X_j^T \epsilon$$

$$E(\hat{\beta}_j|\beta_j) = \beta_j$$

$$Var(\hat{\beta}_j|\beta_j) = \frac{1}{N}\left(1 - \sum_{j=1}^{M} \sigma_j^2\right) = \frac{1}{N}(1 - h^2)$$

Because $\beta_j = \sum_{k=1}^{K} \beta_{j,k} I(j \in S_k)$,

$$E(\hat{\beta}_j|\beta_{j,c}) = \beta_{j,c}$$

$$Var(\hat{\beta}_j|\beta_{j,c}) = \sum_{k=1,k\neq c}^{K} Var(\beta_{j,k}) I(j \in S_k) + \frac{1}{N}(1 - h^2)$$

$$= \sum_{k=1,k\neq c}^{K} \sigma_{j,k}^2 I(j \in S_k) + \frac{1}{N}(1 - h^2) \stackrel{\text{def}}{=} \gamma_{j,c}^2$$

Formulation, No linkage disequilibrium (LD)

Therefore,

$$\hat{\beta}_j|\beta_{j,c} \sim N(\beta_{j,c}, \gamma^2_{j,c})$$

$$\gamma^2_{j,c} \overset{\text{def}}{=} \sum_{k=1,k\neq c}^{K} \sigma^2_{j,k} I(j \in S_k) + \frac{1}{N}(1-h^2)$$

$$h^2 = 1 - \sum_{j=1}^{M} \sigma^2_j$$

Formulation, No linkage disequilibrium (LD)

Also, we have the prior: $\beta_{j,k} \sim N\left(0, \sigma_{j,k}^2\right)$ then the posterior distribution is:

$$
\begin{aligned}
&f\left(\beta_{j,c}|\hat{\beta}_j\right) \propto f\left(\hat{\beta}_j|\beta_{j,c}\right) f\left(\beta_{j,c}\right) \\
&= \frac{1}{\sqrt{2\pi\gamma_{\cdot,c}^2}} \exp\left[-\frac{1}{2\gamma_{\cdot,c}^2}\left(\hat{\beta}_j - \beta_{j,c}\right)^2\right] \frac{1}{\sqrt{2\pi\sigma_{j,c}^2}} \exp\left[-\frac{1}{2\sigma_{j,c}^2}\beta_{j,c}^2\right] \\
&\propto \exp\left[-\frac{\gamma_{\cdot,c}^2 + \sigma_{j,c}^2}{2\gamma_{\cdot,c}^2\sigma_{j,c}^2}\left(\beta_{j,c} - \frac{\sigma_{j,c}^2}{\gamma_{\cdot,c}^2 + \sigma_{j,c}^2}\hat{\beta}_j\right)^2\right]
\end{aligned}
$$

Thus,

$$
\beta_{j,c}|\hat{\beta}_j \sim N\left(\frac{\sigma_{j,c}^2}{\gamma_{j,c}^2 + \sigma_{j,c}^2}\hat{\beta}_j, \frac{\gamma_{j,c}^2\sigma_{j,c}^2}{\gamma_{j,c}^2 + \sigma_{j,c}^2}\right)
$$

$$
E\left[\beta_{j,c}|\hat{\beta}_j\right] = \frac{\sigma_{j,c}^2}{\gamma_{j,c}^2 + \sigma_{j,c}^2}\hat{\beta}_j
$$

Formulation, No linkage disequilibrium (LD)

SNP ID	$\widehat{\beta}$	Brain	Liver
1	0.01	1	0
2	0.02	1	0
3	0.03	1	1
4	0.04	0	1
5	0.05	0	1

Assume, N=100

$\sigma_{1,1}^2 = \sigma_{2,1}^2 = \sigma_{3,1}^2 = \sigma_{4,1}^2 = \sigma_{5,1}^2 = 0.001$

$\sigma_{1,2}^2 = \sigma_{2,2}^2 = \sigma_{3,2}^2 = \sigma_{4,2}^2 = \sigma_{5,2}^2 = 0.002$

Then, $\sigma_1^2 = 0.001$, $\sigma_2^2 = 0.001$, $\sigma_3^2 = 0.003$, $\sigma_4^2 = 0.002$, $\sigma_5^2 = 0.002$

Then, $h^2 = 0.009$

$$\gamma_{1,1}^2 = \sigma_{1,2}^2 I(1 \in S_2) + \frac{1}{N}(1 - h^2) = \frac{1}{N}(1 - h^2) = 0.00991$$

$$\tilde{\beta}_{1,1} = E\left(\beta_{1,1} \middle| \hat{\beta}_1\right) = \frac{\sigma_{1,1}^2}{\gamma_{1,1}^2 + \sigma_{1,1}^2}\hat{\beta}_1 = \frac{0.001}{0.00991 + 0.001} \times 0.01 = 9 \times 10^{-4}$$

$$\gamma_{3,1}^2 = \sigma_{1,2}^2 I(1 \in S_2) + \frac{1}{N}(1 - h^2) = 0.002 + \frac{1}{N}(1 - h^2) = 0.01191$$

$$\tilde{\beta}_{3,1} = E\left(\beta_{3,1} \middle| \hat{\beta}_3\right) = \frac{\sigma_{3,1}^2}{\gamma_{3,1}^2 + \sigma_{3,1}^2}\hat{\beta}_3 = \frac{0.001}{0.01191 + 0.001} \times 0.03 = 2 \times 10^{-2}$$

Formulation, No linkage disequilibrium (LD)

In this circumstatnce, $Cov(X_i, X_j) \neq 0$. But we can used the **reference panel** to get the LD matrix D. Mathematically speaking, $D = \frac{1}{N} X^T X$

Goal: estimate the posterior expectation $\tilde{\beta}_{\cdot,c} = E[\beta_{\cdot,c} \mid \hat{\beta}, D]$, where $\beta_{\cdot,c} = \begin{pmatrix} \beta_{1,c} \\ \vdots \\ \beta_{M,c} \end{pmatrix}$

In the following formula, we consider a region of LD matrix to ensure D is invertible.

$$\hat{\beta} = \frac{1}{N} X^T Y = \frac{1}{N} X^T X \beta + \frac{1}{N} X^T \epsilon = D\beta + \frac{1}{N} X^T \epsilon$$

$$E\left(\hat{\beta} \middle| \beta, D\right) = D\beta$$

$$Var\left(\hat{\beta} \middle| \beta, D\right) = \frac{1 - h^2}{N} D$$

Formulation, No linkage disequilibrium (LD)

In this circumstatnce, $Cov(X_i, X_j) \neq 0$. But we can used the **reference panel** to get the LD matrix D. Mathematically speaking, $D = \frac{1}{N} X^T X$

In this scenario, the posterior distribution is:

$$\beta_{\cdot,c} = \begin{pmatrix} \beta_{1,c} \\ \vdots \\ \beta_{M,c} \end{pmatrix} \mid \hat{\beta}, D \sim MVN\left(\left(D^T \Gamma_c^{-1} D + \frac{1}{\sigma_{\cdot,c}^2} I_M\right)^{-1} D\Gamma_c^{-1}\hat{\beta}, \left(D^T \Gamma_c^{-1} D + \frac{1}{\sigma_{\cdot,c}^2} I_M\right)^{-1}\right)$$

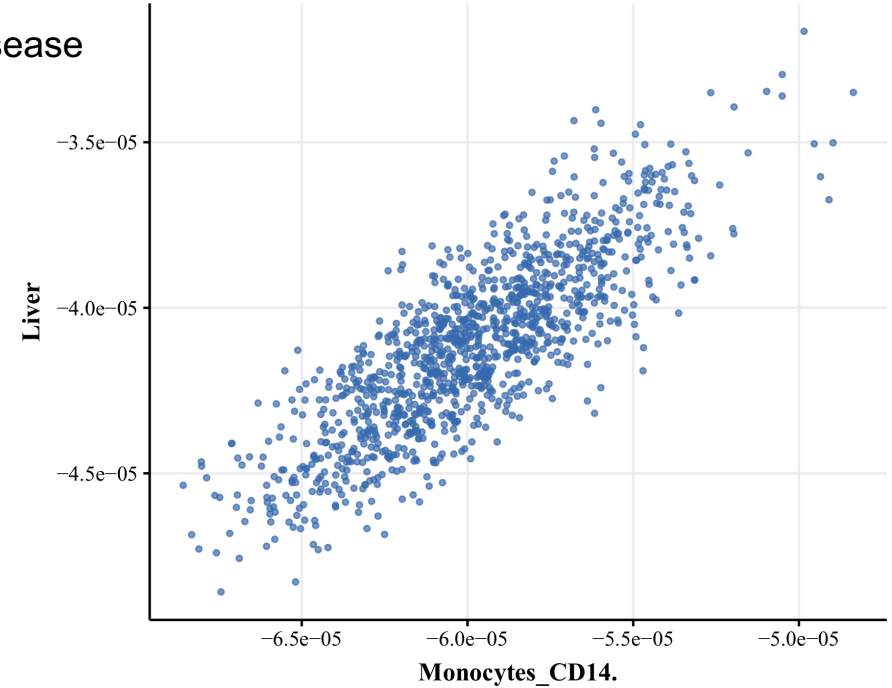
$$\Gamma_c \overset{\text{def}}{=} \sum_{k=1, k\neq c}^{K} \begin{bmatrix} \sigma_{\cdot,k} I(1 \in S_k) & & \\ & \sigma_{\cdot,k} I(2 \in S_k) & \\ & & \ddots \\ & & & \sigma_{\cdot,k} I(m \in S_k) \end{bmatrix} + \frac{1-h^2}{N} D$$

Results

GWAS summary statistics for Alzheimer's disease

Annotations: Monocyte and Liver

Correlation between PRS_Liver and PRS_Monocyte: 0.8



Thanks!

Any questions ?