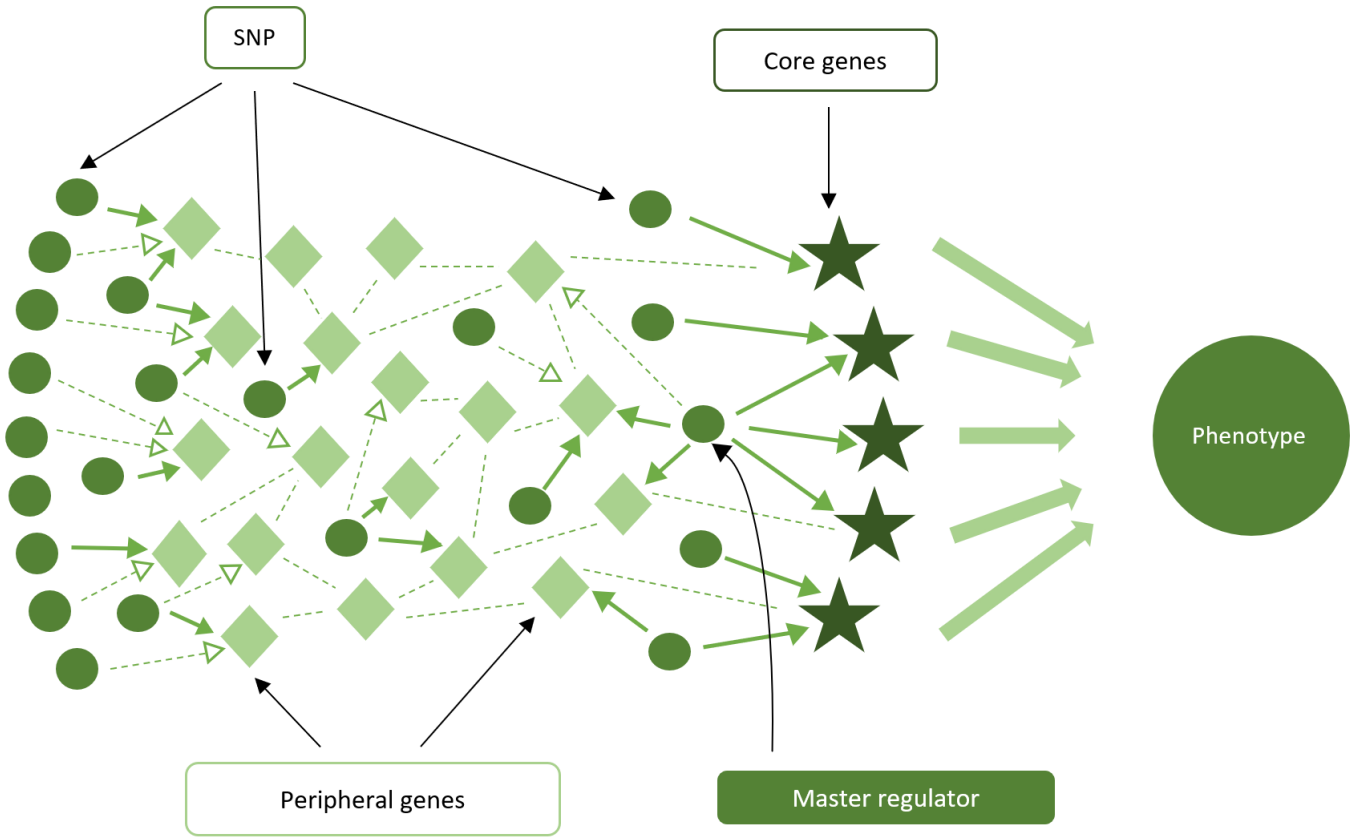
# Decoding the Genomic Network: A Dependence-Adjusted Approach for Functional Analysis of Omnigenic Models

Samhita Pal, Xinge Jessie Jeng

North Carolina State University

Oct 26, 2023

# Motivation : Omnigenic Network



# Motivation : Why consider trans effects?

$$\text{Phenotypic variance} = \underbrace{\sum_{j=1}^{M} \mu_j^2 V_{j;\text{cis}}}_{M \text{ core terms}} + \underbrace{\sum_{j=1}^{M} \mu_j^2 V_{j;\text{trans}}}_{M \text{ trans terms}}$$

$$+ \underbrace{\sum_{j=1}^{M} \sum_{k=1}^{j} 2\mu_j \mu_k C_{jk}}_{M^2 - M \text{ covariance terms that mainly arise due to trans effect}} \quad \text{where } M^2 - M \geq M.$$

# Statistical Modeling of the Biological Problem

- Suppose there are $p$ SNPs in total across the entire genome.
- $Z_j$ is the summary-statistic for the $j^{\text{th}}$ SNP.
- Suppose, WLG, $S_0 = \{1, 2, \ldots, p-2\}$ is the set of influential SNPs for a particular gene.
- $Z_j|\mu_j = \mu_j \mathbb{1}\{j \in S_0\} + \delta_j$, where

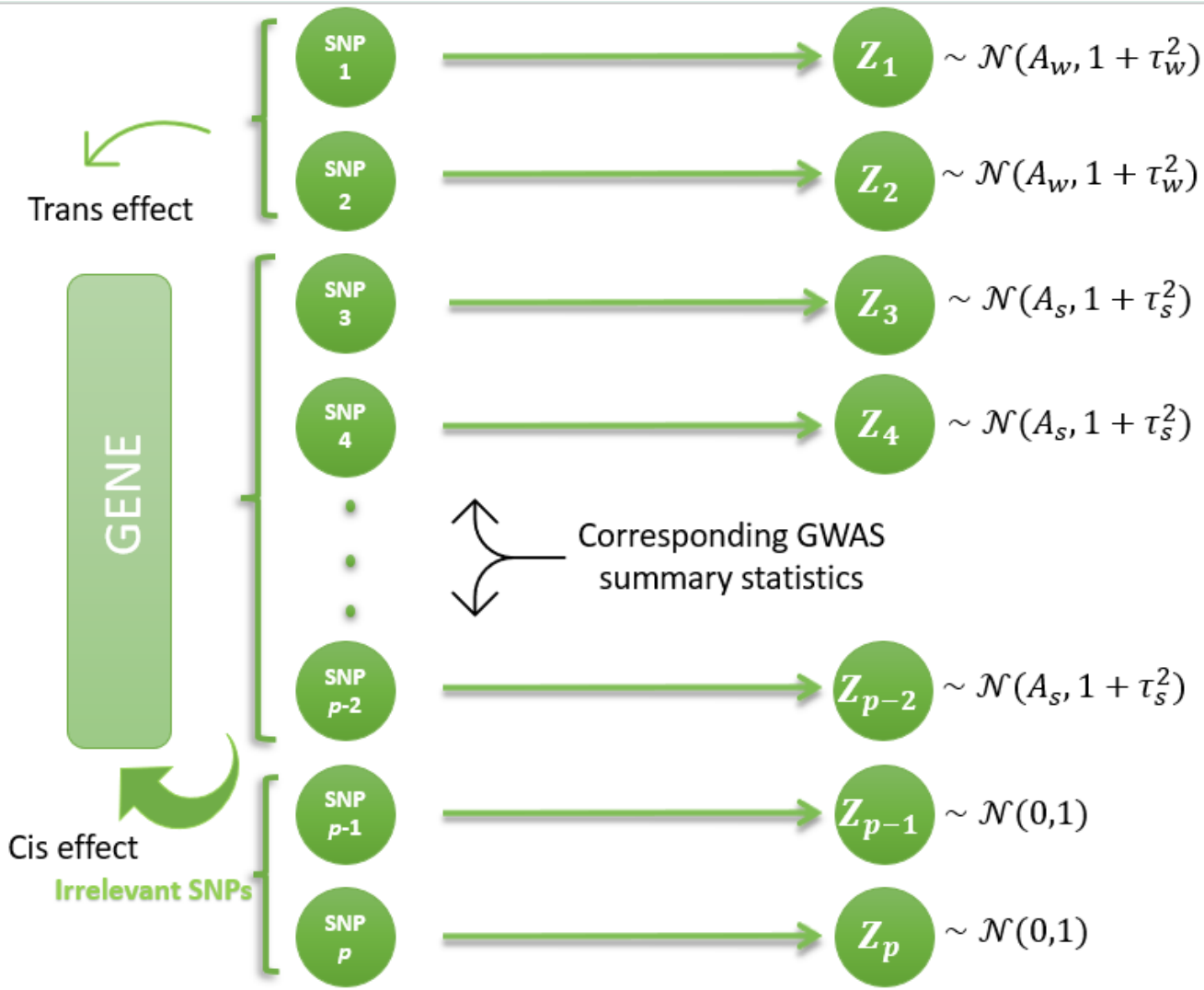
$$\delta_j \sim \mathcal{N}(0,1) \text{ and } \mu_j \sim (1-\epsilon)\delta_0 + \epsilon\mathcal{N}(A, \tau^2).$$

- Marginally,

$$Z_j \sim \begin{cases} (1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(A, 1+\tau^2), & \text{if } j \in S_0, \\ \mathcal{N}(0,1), & \text{if } j \notin S_0. \end{cases}$$

- $A$ and $\tau$ differ based on different kinds influential SNPs.

# One Gene System



# Model

- The formal statistical hypothesis then looks like

$$H_0: \quad Z_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \quad \text{vs.}$$

$$H_1: \quad Z_j \overset{\text{iid}}{\sim} (1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(A, 1+\tau^2), \quad \forall j = 1, \ldots, p.$$

- HC and BJ tests can be used to determine how much each element of the vector of test-statistics digresses from $\mathcal{N}(0,1)$.

**HC**

$$\max_{1 \le j \le p} \frac{\sqrt{p}(j/p - p_{(j)})}{\sqrt{(p_{(j)}(1 - p_{(j)}))}}$$

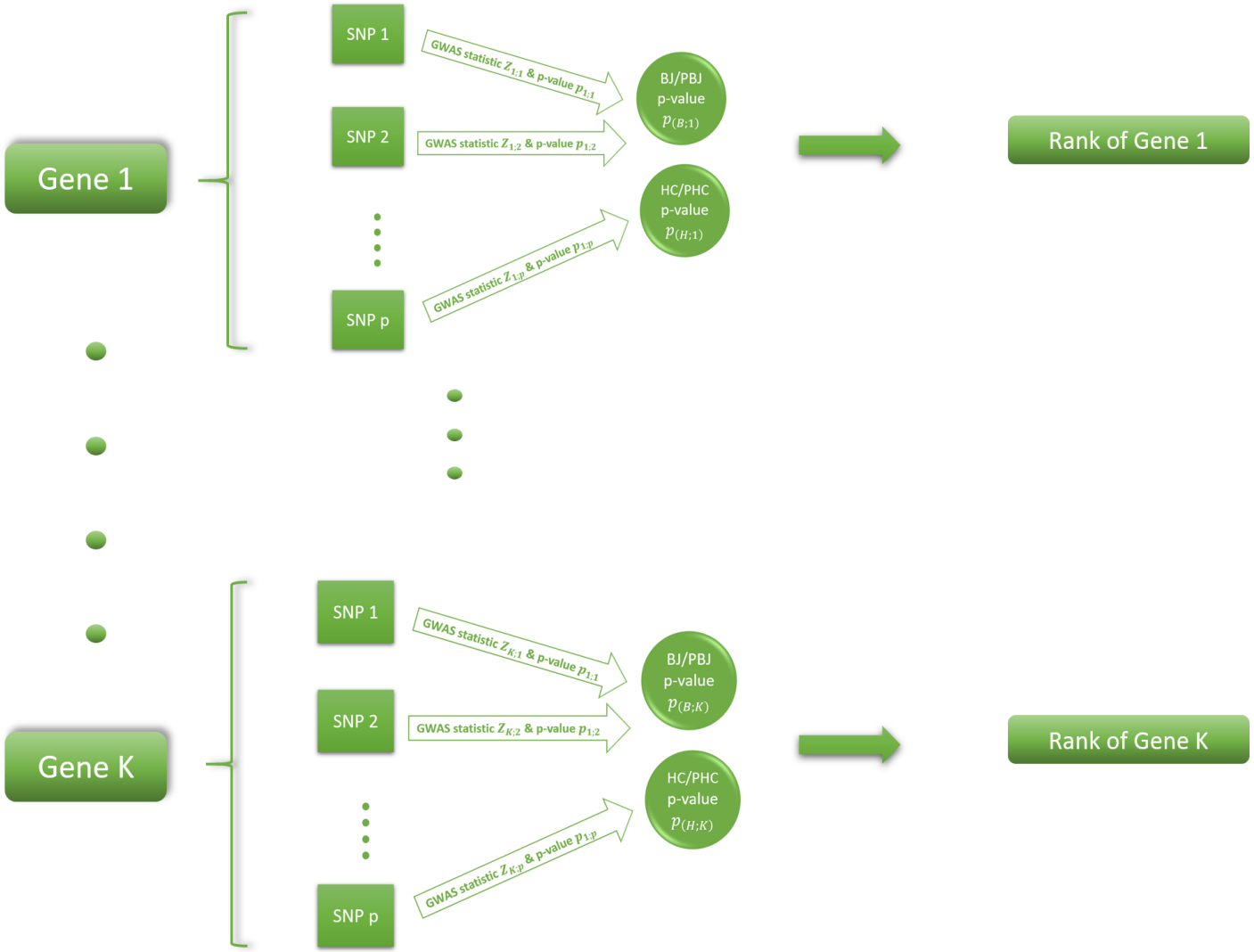
$$\sup_{t \in \mathbb{R}} \frac{\sqrt{p}(\overline{F_p}(t) - \bar{\Phi}(t))}{\sqrt{(\bar{\Phi}(t)(1 - \bar{\Phi}(t)))}}$$

**BJ**

$$\max_{1 \le j \le p} \frac{j}{p} \log\left(\frac{j/p}{p_{(j)}}\right) + (1 - \frac{j}{p}) \log\left(\frac{1 - j/p}{1 - p_{(j)}}\right)$$

$$\sup_{t \in \mathbb{R}} F_p(t) \log\left(\frac{F_p(t)}{\Phi(t)}\right) + \overline{F_p}(t) \log\left(\frac{\overline{F_p}(t)}{\bar{\Phi}(t)}\right)$$

# Ranking Scheme



# Dependence Adjustment

- For a given gene, let $\mathbf{\Sigma}$ be the covariance matrix of the $p$ SNPs within the gene.
- Then, the spectral decomposition of $\mathbf{\Sigma}$ can be split into two parts as follows:

$$\mathbf{\Sigma} = \sum_{m=1}^{k} \lambda_m \boldsymbol{\gamma}_m \boldsymbol{\gamma}_m^\top + \sum_{m=k+1}^{p} \lambda_m \boldsymbol{\gamma}_m \boldsymbol{\gamma}_m^\top = \sum_{m=1}^{k} \mathbf{b}_m \mathbf{b}_m^\top + \mathbf{A}$$

- **PFA:** $Z_j = \mu_j + \mathbf{b}_j^\top \mathbf{W} + K_j; \;\; j = 1, 2, \ldots, p$ where $\mathbf{W} \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k) \perp (K_1, \ldots, K_p)^\top \sim \mathcal{N}_p(\mathbf{0}, \mathbf{A})$.
- The dependence-adjusted p-values are then given by

$$p_i^* = 2\Phi\left(-a_j\left(Z_j - \mathbf{b}_j^\top \hat{\mathbf{W}}\right)\right)$$

where $a_j$ is the suitable normalizing factor and $\hat{\mathbf{W}}$ is a suitable estimator of $\mathbf{W}$.

# Detection Boundary

- **Reparameterization**: $(A, \epsilon) \to (h, \gamma)$, where $h$ is the intensity parameter and $\gamma$ is the sparsity parameter.
- Define $m(\tau) = \min\left(\frac{1+\tau^2}{4}, \frac{1}{1+\tau^2}\right)$. Then, the **detection boundary** is defined for $1/2 \leq \gamma < 1$ as

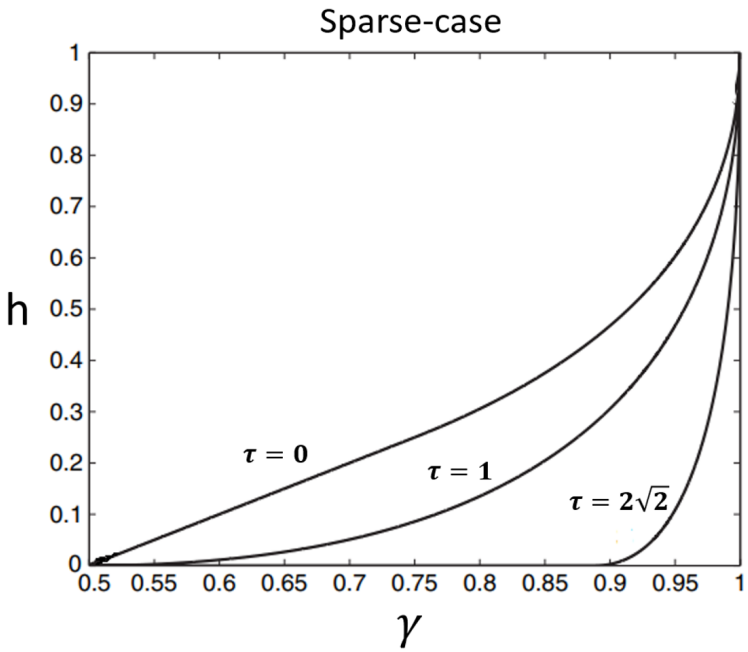
$$\rho(\gamma, \tau) = \begin{cases} (1 - (1+\tau^2)\sqrt{1-\gamma})^2 & , 1 - m(\tau) < \gamma < 1 \\ (1-\tau^2)(\gamma - 1/2) & , 1/2 < \gamma < 1 - m(\tau) \end{cases}$$

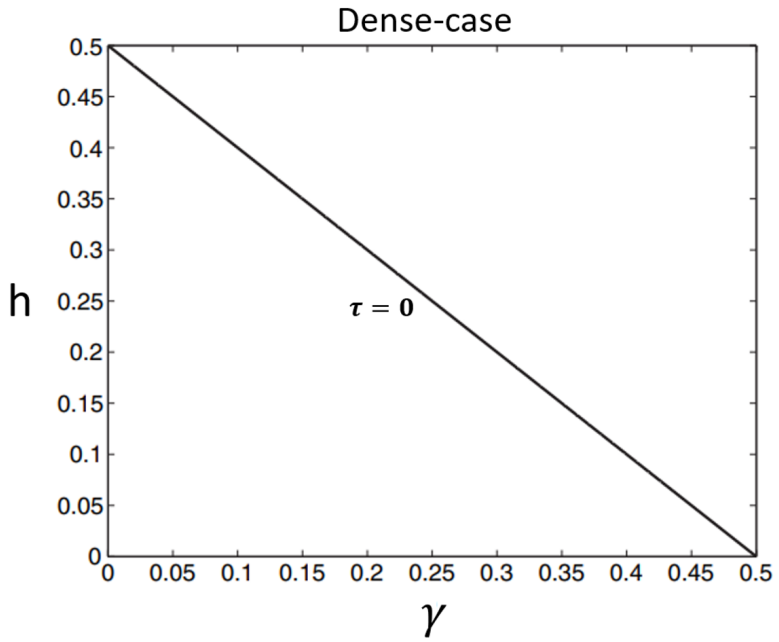
and for $0 \leq \gamma < 1/2$ as

$$\rho(\gamma, \tau) = \begin{cases} 1/2 - \gamma & , \tau = 0 \\ \infty & , \tau > 0 \end{cases}$$

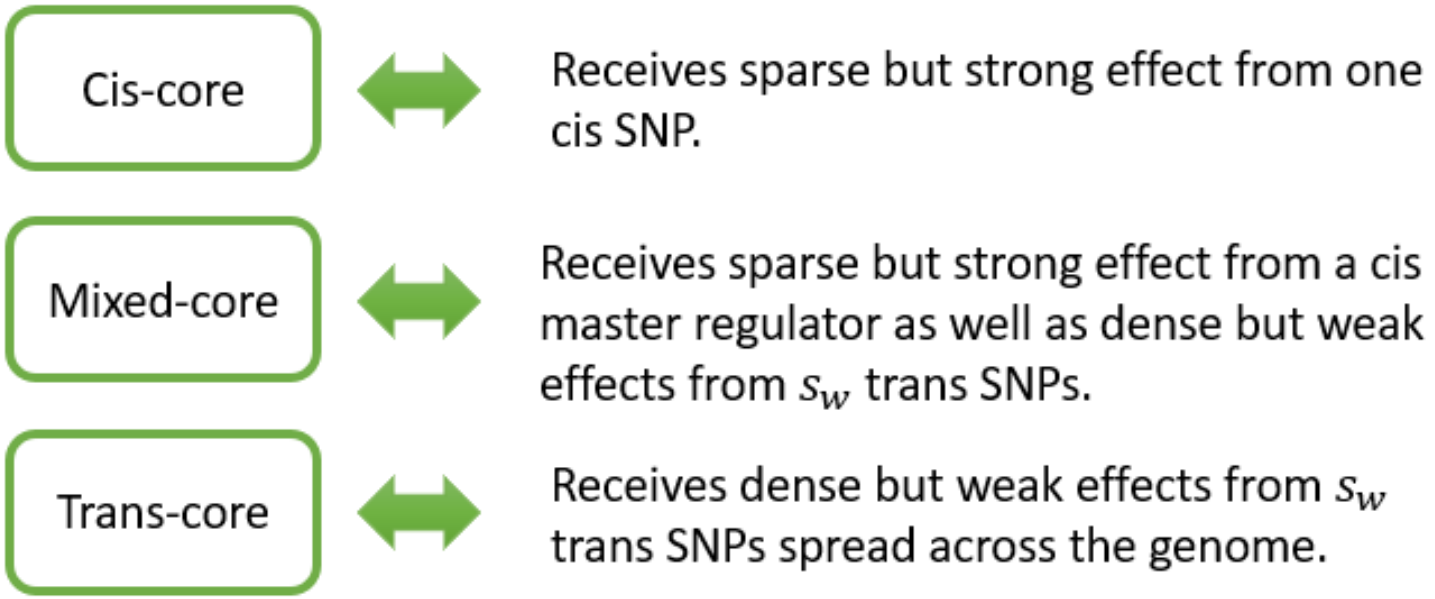
- The for the gene under consideration, HC and BJ tests attain **full asymptotic power** in the detectable region $(h > \rho(\gamma, \tau))$ and the **type-I error vanishes as** $n \to \infty$.

# Detection Region



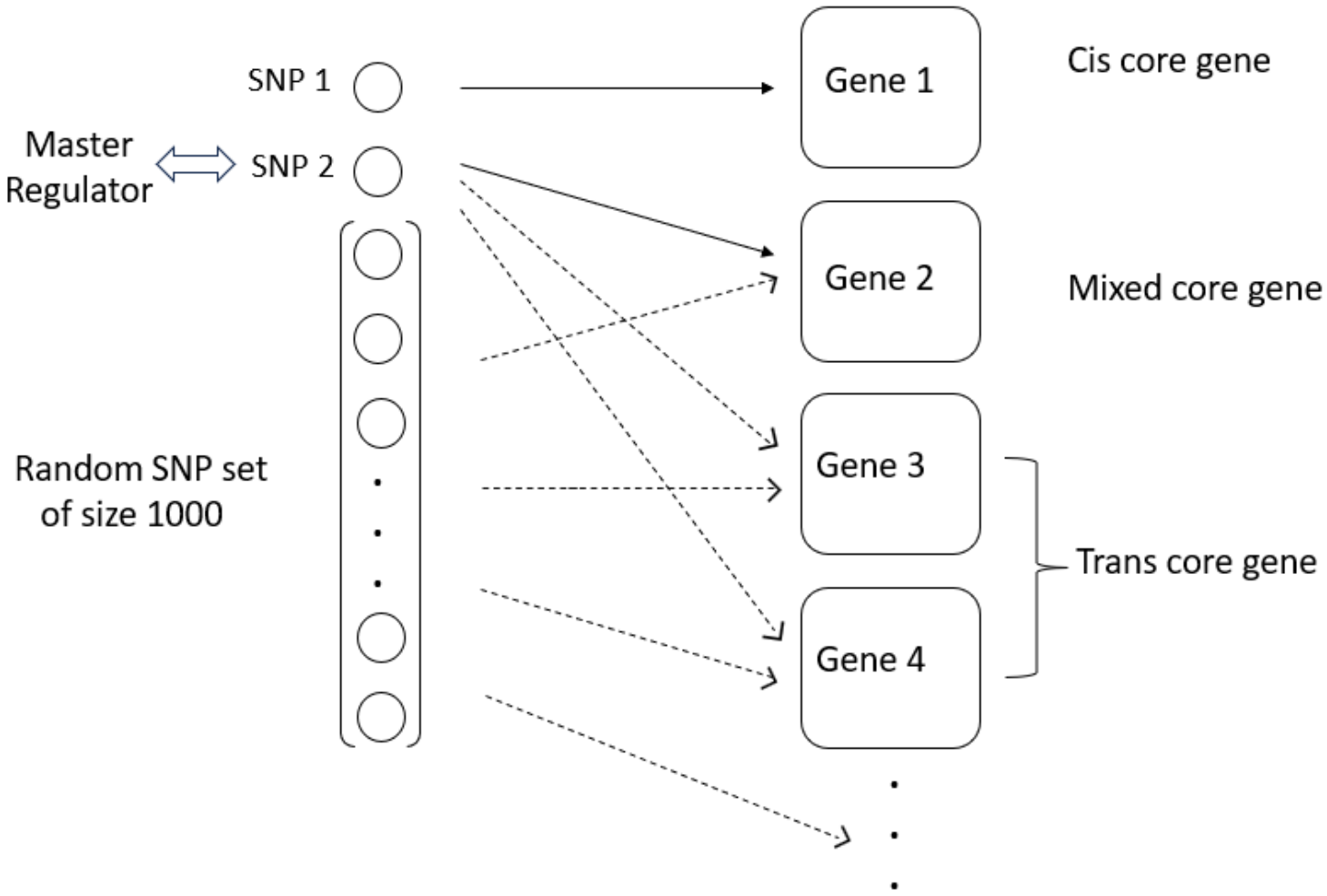



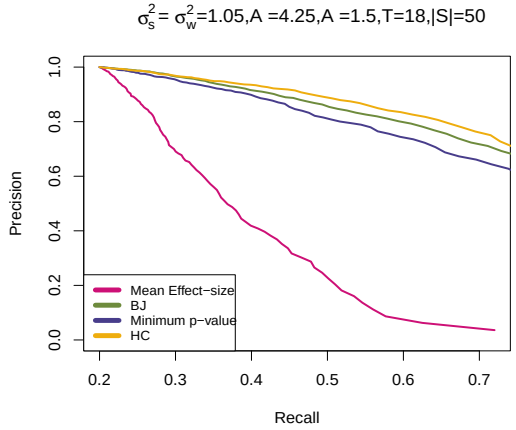
# Simulation Architecture



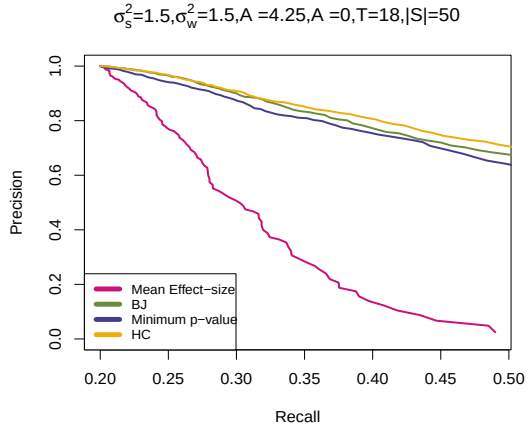

| Genes | Index | Stong cis effect | Weak trans effect |
|---|---|---|---|
| Cis-core | k = 1 | $\mu_{jk} \sim \mathcal{N}(A_s, \tau_s^2), j = 1,$ | - |
| Mixed-core | k = 2 | $\mu_{jk} \sim \mathcal{N}(A_s, \tau_s^2), j = 2$ | $\mu_{jk} \sim \mathcal{N}(A_w, \tau_w^2), j \in \tau_2$ |
| Trans-core | k = 3,4,. . . ,18 | - | $\mu_{jk} \sim \mathcal{N}(A_w, \tau_w^2), j \in \tau_k$ |

# Simulation Setup

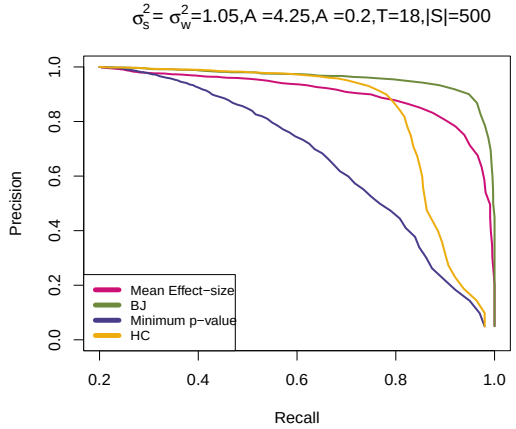




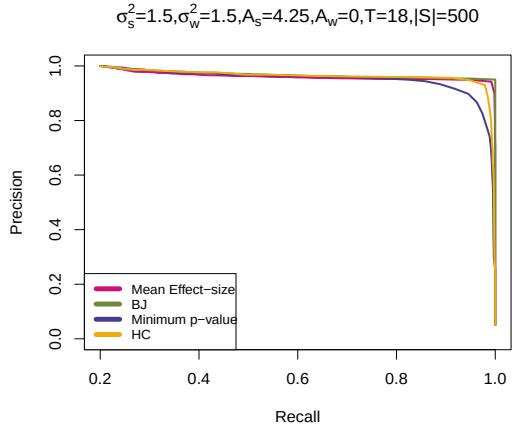

**(a)** Sparse Case 1




**(b)** Sparse Case 2




**(c)** Dense Case 1




**(d)** Dense Case 2

*Thank you!*