Statistical Methods for Analysis of Correlated Data

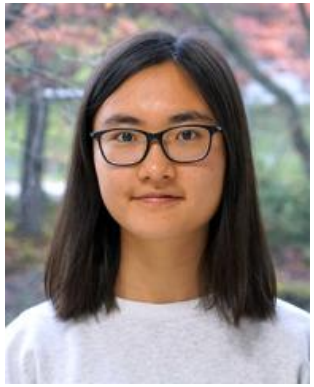
Jing Ma

Division of Public Health Sciences
Fred Hutchinson Cancer Center

Genetic Analysis of Mendelian and Complex Disorders Course
27 July 2022

Collaborators

FRED HUTCH
CURES START HERE™



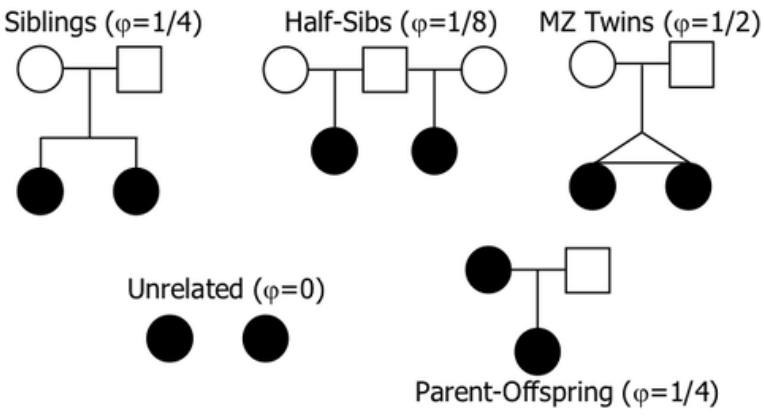
Kun Yue

Mike Hellstern

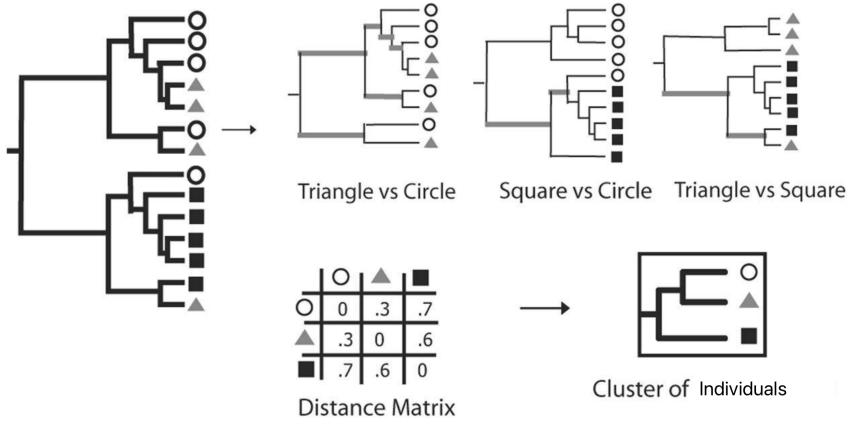
Ali Shojaie

George Michailidis

Correlation among Samples



Genetic relatedness



Phylogenetic relatedness[1]

Human genetics: kinship

Microbiome: phylogenetic distances

[1] Lozupone and Knight. *Appl Environ Microbiol.* '05

Correlation among Variables

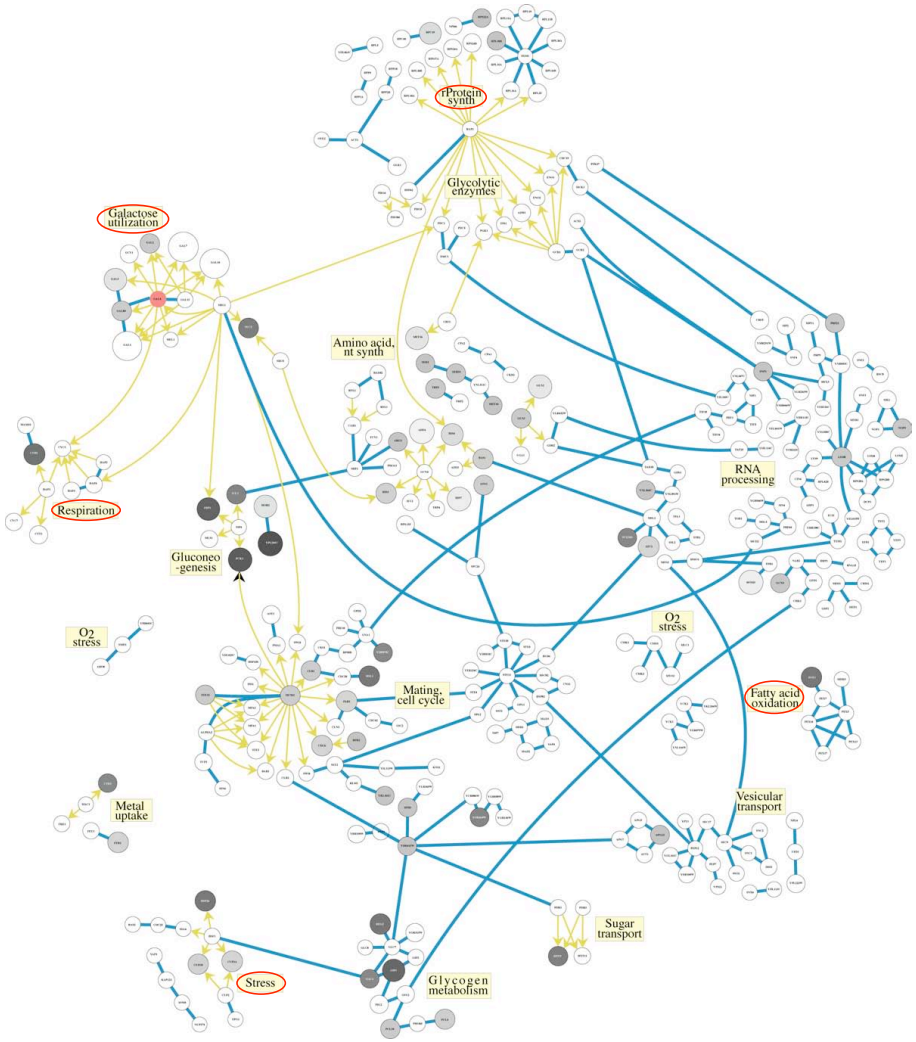


Fig: Integrated physical interaction network in yeast *Saccharomyces cerevisiae*[2].

- Nodes: genes
- Edges: protein $\rightarrow$ DNA and protein – protein
- Genes form functional modules

[2] Ideker et al. *Science.* 01'

Today's Talk

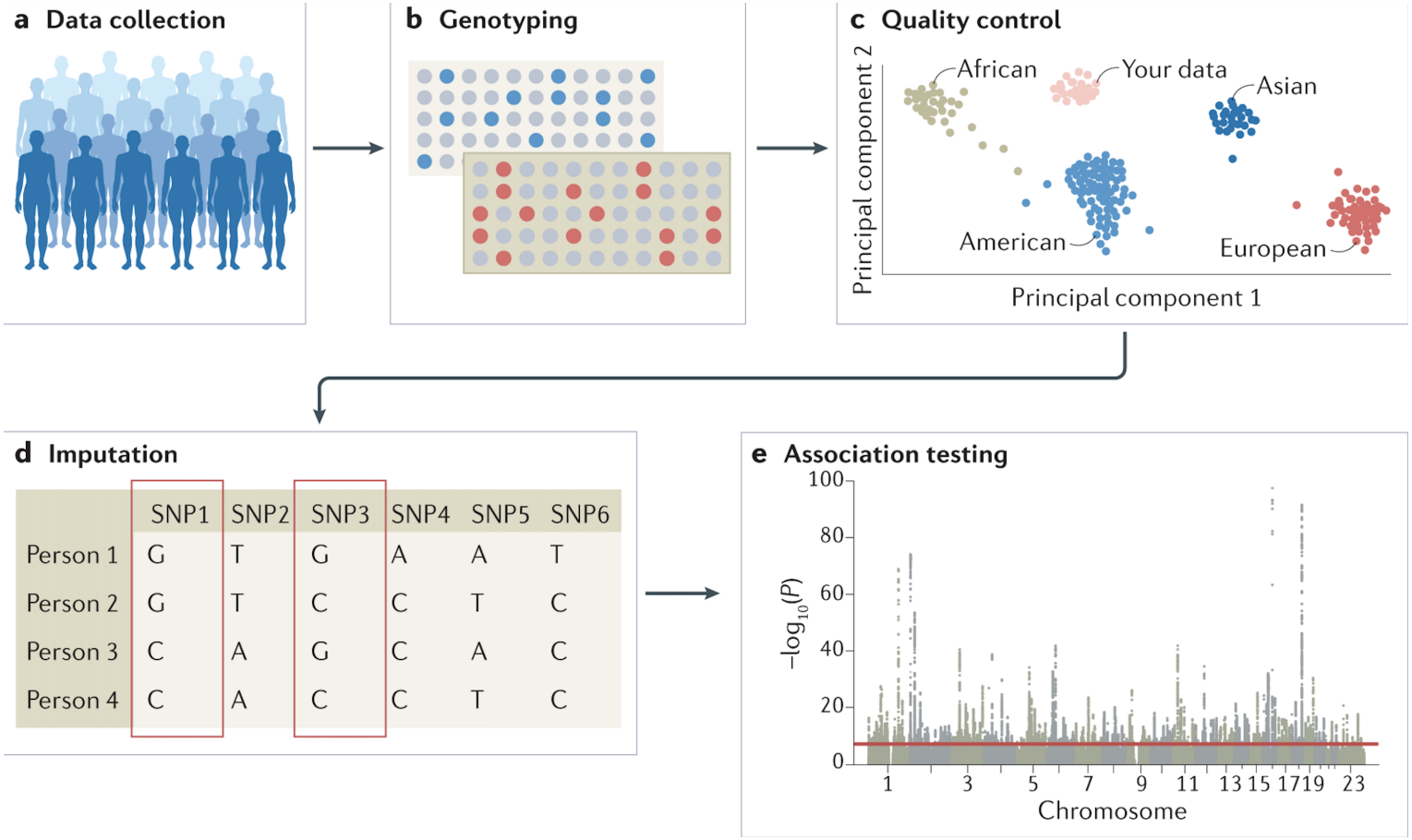
FRED HUTCH
CURES START HERE™

Genome-wide Association Analysis

Gene Set Analysis

Genome-wide Association

Scientific Question: to identify associations of genotypes with phenotypes.



	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Person 1	G	T	G	A	A	T
Person 2	G	T	C	C	T	C
Person 3	C	A	G	C	A	C
Person 4	C	A	C	C	T	C

Fig: Steps of a GWAS experiment[3].

[3] Uffelmann et al. *Nat Rev Methods Primers*. '21

FRED HUTCH
CURES START HERE™

Statistical model

$$y = W\alpha + X_s\beta_s + \gamma + \epsilon$$
$$\gamma \sim N(0, \sigma^2_\gamma K)$$
$$\epsilon \sim N(0, \sigma^2_\epsilon I_n)$$

This is a linear mixed model where

FRED HUTCH
CURES START HERE™

Statistical model

$$y = W\alpha + X_s\beta_s + \gamma + \epsilon$$
$$\gamma \sim N(0, \sigma^2_\gamma K)$$
$$\epsilon \sim N(0, \sigma^2_\epsilon I_n)$$

This is a linear mixed model where

- y: an $n \times 1$ vector of quantitative traits (e.g., red blood cell count)

FRED HUTCH
CURES START HERE™

Statistical model

$$y = W\alpha + X_s\beta_s + \gamma + \epsilon$$
$$\gamma \sim N(0, \sigma^2_\gamma K)$$
$$\epsilon \sim N(0, \sigma^2_\epsilon I_n)$$

This is a linear mixed model where

- y: an $n \times 1$ vector of quantitative traits (e.g., red blood cell count)
- W: a matrix of covariates (e.g. age, sex, ancestry)

FRED HUTCH
CURES START HERE™

Statistical model

$$y = W\alpha + X_s\beta_s + \gamma + \epsilon$$
$$\gamma \sim N(0, \sigma^2_\gamma K)$$
$$\epsilon \sim N(0, \sigma^2_\epsilon I_n)$$

This is a linear mixed model where

- y: an $n \times 1$ vector of quantitative traits (e.g., red blood cell count)
- W: a matrix of covariates (e.g. age, sex, ancestry)
- X_s: an $n \times 1$ vector of genotype values at SNP s

FRED HUTCH
CURES START HERE™

Statistical model

$$y = W\alpha + X_s\beta_s + \gamma + \epsilon$$
$$\gamma \sim N(0, \sigma^2_\gamma K)$$
$$\epsilon \sim N(0, \sigma^2_\epsilon I_n)$$

This is a linear mixed model where

- y: an $n \times 1$ vector of quantitative traits (e.g., red blood cell count)
- W: a matrix of covariates (e.g. age, sex, ancestry)
- X_s: an $n \times 1$ vector of genotype values at SNP s
- β_s: the strength of association between SNP s and y

FRED HUTCH
CURES START HERE™

Statistical model

$$y = W\alpha + X_s\beta_s + \gamma + \epsilon$$
$$\gamma \sim N(0, \sigma_\gamma^2 K)$$
$$\epsilon \sim N(0, \sigma_\epsilon^2 I_n)$$

This is a linear mixed model where

- y: an $n \times 1$ vector of quantitative traits (e.g., red blood cell count)
- W: a matrix of covariates (e.g. age, sex, ancestry)
- X_s: an $n \times 1$ vector of genotype values at SNP s
- β_s: the strength of association between SNP s and y
- γ: a random effect that captures the polygenic effect of other SNPs

FRED HUTCH
CURES START HERE™

Statistical model

$$y = W\alpha + X_s\beta_s + \gamma + \epsilon$$
$$\gamma \sim N(0, \sigma^2_\gamma K)$$
$$\epsilon \sim N(0, \sigma^2_\epsilon I_n)$$

This is a linear mixed model where

- y: an $n \times 1$ vector of quantitative traits (e.g., red blood cell count)
- W: a matrix of covariates (e.g. age, sex, ancestry)
- X_s: an $n \times 1$ vector of genotype values at SNP s
- β_s: the strength of association between SNP s and y
- γ: a random effect that captures the polygenic effect of other SNPs
- K: $n \times n$ kinship matrix

Problems of Interest

Input data: (W, X_s, y, K)

Association testing

$$H_0 : \beta_s = 0$$

Heritability estimation

$$h^2 = \frac{\sigma^2_\gamma}{\sigma^2_\gamma + \sigma^2_\epsilon}$$

Association Testing

$$y = W\alpha + X_s\beta_s + u,$$

where $u \sim \mathcal{N}(0, \sigma_\gamma^2 V)$ and $V = K + \sigma_\epsilon^2/\sigma_\gamma^2 I_n$.

Association Testing

$$y = W\alpha + X_s\beta_s + u,$$

where $u \sim \mathcal{N}(0, \sigma^2_\gamma V)$ and $V = K + \sigma^2_\epsilon / \sigma^2_\gamma I_n$.

Generalized least squares

$$\begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta}_s \end{bmatrix} = (X^\intercal V^{-1} X)^{-1} X^\intercal V^{-1} y$$

Association Testing

$$y = W\alpha + X_s\beta_s + u,$$

where $u \sim \mathcal{N}(0, \sigma^2_\gamma V)$ and $V = K + \sigma^2_\epsilon / \sigma^2_\gamma I_n$.

Generalized least squares

$$\begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta}_s \end{bmatrix} = (X^\intercal V^{-1} X)^{-1} X^\intercal V^{-1} y$$

Both analysis tasks require estimating the variance components!

Kinship Matrix

Let $Z \in \{0, 1, 2\}^{n \times q}$ denote the remaining q SNPs (i.e. excluding X_s).

Kinship Matrix

Let $Z \in \{0, 1, 2\}^{n \times q}$ denote the remaining q SNPs (i.e. excluding X_s).
Consider the model

$$y = W\alpha + X_s\beta_s + Z\tau + \epsilon$$

Kinship Matrix

Let $Z \in \{0, 1, 2\}^{n \times q}$ denote the remaining q SNPs (i.e. excluding X_s). Consider the model

$$y = W\alpha + X_s\beta_s + Z\tau + \epsilon$$

The coefficient β_s is the effect of SNP s on y after adjusting the effects introduced by other SNPs Z.

Kinship Matrix

Let $Z \in \{0, 1, 2\}^{n \times q}$ denote the remaining q SNPs (i.e. excluding X_s). Consider the model

$$y = W\alpha + X_s\beta_s + Z\tau + \epsilon$$

The coefficient β_s is the effect of SNP s on y after adjusting the effects introduced by other SNPs Z.

Fitting the model

$$\tau_j \sim \mathcal{N}(0, \sigma^2_\gamma), \quad j = 1, \ldots, q$$

Kinship Matrix

Let $Z \in \{0, 1, 2\}^{n \times q}$ denote the remaining q SNPs (i.e. excluding X_s). Consider the model

$$y = W\alpha + X_s\beta_s + Z\tau + \epsilon$$

The coefficient β_s is the effect of SNP s on y after adjusting the effects introduced by other SNPs Z.

Fitting the model

$$\tau_j \sim \mathcal{N}(0, \sigma^2_\gamma), \quad j = 1, \ldots, q$$

Averaging over the distribution of τ_j's, we obtain

$$y \sim \mathcal{N}(W\alpha + X_s\beta_s, \sigma^2_\gamma ZZ^\top + \sigma^2_\epsilon I_n)$$

Kinship Matrix

Let $Z \in \{0, 1, 2\}^{n \times q}$ denote the remaining q SNPs (i.e. excluding X_s). Consider the model

$$y = W\alpha + X_s\beta_s + Z\tau + \epsilon$$

The coefficient β_s is the effect of SNP s on y after adjusting the effects introduced by other SNPs Z.

Fitting the model

$$\tau_j \sim \mathcal{N}(0, \sigma^2_\gamma), \quad j = 1, \ldots, q$$

Averaging over the distribution of τ_j's, we obtain

$$y \sim \mathcal{N}(W\alpha + X_s\beta_s, \sigma^2_\gamma ZZ^\top + \sigma^2_\epsilon I_n)$$

The kinship $K = ZZ^\top$ is a natural choice.

VC Estimation

Maximum likelihood (null)

$$\max_{\sigma^2_\gamma, \sigma^2_\epsilon/\sigma^2_\gamma} \{-\frac{1}{2}\log|\sigma^2_\gamma V| - \frac{1}{2}\sigma^{-2}_\gamma (y - W\widehat{\alpha})^\top V^{-1}(y - W\widehat{\alpha})\}$$

VC Estimation

Maximum likelihood (null)

$$\max_{\sigma^2_\gamma, \sigma^2_\epsilon/\sigma^2_\gamma} \{-\frac{1}{2}\log|\sigma^2_\gamma V| - \frac{1}{2}\sigma^{-2}_\gamma (y - W\widehat{\alpha})^\mathsf{T} V^{-1}(y - W\widehat{\alpha})\}$$

Restricted maximum likelihood (REML)

$$\max_{\sigma^2_\gamma, \sigma^2_\epsilon/\sigma^2_\gamma} \{\text{likelihood of } L^\mathsf{T} y\}$$

where $L^\mathsf{T} W = 0$ and L^T has full row rank.

VC Estimation

Maximum likelihood (null)

$$\max_{\sigma^2_\gamma, \sigma^2_\epsilon/\sigma^2_\gamma} \{-\frac{1}{2}\log|\sigma^2_\gamma V| - \frac{1}{2}\sigma^{-2}_\gamma (y - W\widehat{\alpha})^\mathsf{T} V^{-1}(y - W\widehat{\alpha})\}$$

Restricted maximum likelihood (REML)

$$\max_{\sigma^2_\gamma, \sigma^2_\epsilon/\sigma^2_\gamma} \{\text{likelihood of } L^\mathsf{T} y\}$$

where $L^\mathsf{T} W = 0$ and L^T has full row rank.

☺ REML estimator has the smallest variance among all estimators

VC Estimation

Maximum likelihood (null)

$$\max_{\sigma^2_\gamma, \sigma^2_\epsilon/\sigma^2_\gamma} \{-\frac{1}{2}\log|\sigma^2_\gamma V| - \frac{1}{2}\sigma^{-2}_\gamma (y - W\widehat{\alpha})^\top V^{-1}(y - W\widehat{\alpha})\}$$

Restricted maximum likelihood (REML)

$$\max_{\sigma^2_\gamma, \sigma^2_\epsilon/\sigma^2_\gamma} \{\text{likelihood of } L^\top y\}$$

where $L^\top W = 0$ and $L^\top$ has full row rank.

☺ REML estimator has the smallest variance among all estimators

☹ REML is computationally expensive: need to invert $n \times n$ matrices where $n > 100K$ in large studies

VC Estimation

Maximum likelihood (null)

$$\max_{\sigma^2_\gamma, \sigma^2_\epsilon/\sigma^2_\gamma} \{-\frac{1}{2}\log|\sigma^2_\gamma V| - \frac{1}{2}\sigma^{-2}_\gamma (y - W\widehat{\alpha})^\top V^{-1}(y - W\widehat{\alpha})\}$$

Restricted maximum likelihood (REML)

$$\max_{\sigma^2_\gamma, \sigma^2_\epsilon/\sigma^2_\gamma} \{\text{likelihood of } L^\top y\}$$

where $L^\top W = 0$ and $L^\top$ has full row rank.

☺ REML estimator has the smallest variance among all estimators

☹ REML is computationally expensive: need to invert $n \times n$ matrices where $n > 100K$ in large studies

Need alternatives that can balance statistical and computational efficiency.

Method of Moments

Assume no fixed effects for the moment. The model is

$$y = \gamma + \epsilon.$$

Method of Moments

Assume no fixed effects for the moment. The model is

$$y = \gamma + \epsilon.$$

The second moment of y is

$$\mathrm{E}(yy^{\mathsf{T}}) = \sigma_\gamma^2 K + \sigma_\epsilon^2 I_n.$$

Method of Moments

Assume no fixed effects for the moment. The model is

$$y = \gamma + \epsilon.$$

The second moment of y is

$$\mathrm{E}(yy^{\mathsf{T}}) = \sigma_\gamma^2 K + \sigma_\epsilon^2 I_n.$$

yy^{T} is a linear function of K and I_n!

Haseman-Elston Regression

Let vec(K) denote the vectorization of K by stacking its columns. Let $n^* = n^2$ and

$$\widetilde{Y} = \text{vec}(yy^\top) \in \mathbb{R}^{n^*}, \quad \widetilde{X} = [\text{vec}(I_n), \text{vec}(K)] \in \mathbb{R}^{n^* \times 2}.$$

[4] Haseman and Elston. *Behavior Genetics*. '72; Sofer T. *Stat. Appl. Genet. Mol. Biol.* '17

Haseman-Elston Regression

Let vec(K) denote the vectorization of K by stacking its columns. Let $n^* = n^2$ and

$$\widetilde{Y} = \text{vec}(yy^\top) \in \mathbb{R}^{n^*}, \quad \widetilde{X} = [\text{vec}(I_n), \text{vec}(K)] \in \mathbb{R}^{n^* \times 2}.$$

HE regression[4] solves for σ_j^2 by minimizing

$$\frac{1}{n^*}(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2)^\top(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2)$$

[4] Haseman and Elston. *Behavior Genetics*. '72; Sofer T. *Stat. Appl. Genet. Mol. Biol.* '17

Haseman-Elston Regression

Let vec(K) denote the vectorization of K by stacking its columns. Let $n^* = n^2$ and

$$\widetilde{Y} = \text{vec}(yy^\mathsf{T}) \in \mathbb{R}^{n^*}, \quad \widetilde{X} = [\text{vec}(I_n), \text{vec}(K)] \in \mathbb{R}^{n^* \times 2}.$$

HE regression[4] solves for σ_j^2 by minimizing

$$\frac{1}{n^*}(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2)^\mathsf{T}(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2)$$

☺ The HE estimator is unbiased.

☺ HE is computationally efficient: $O(dn^2)$ as opposed to $O(n^3)$ for REML

[4] Haseman and Elston. *Behavior Genetics*. '72; Sofer T. *Stat. Appl. Genet. Mol. Biol.* '17

Haseman-Elston Regression

Let vec(K) denote the vectorization of K by stacking its columns. Let $n^* = n^2$ and

$$\widetilde{Y} = \text{vec}(yy^\mathsf{T}) \in \mathbb{R}^{n^*}, \quad \widetilde{X} = [\text{vec}(I_n), \text{vec}(K)] \in \mathbb{R}^{n^* \times 2}.$$

HE regression[4] solves for σ_j^2 by minimizing

$$\frac{1}{n^*}(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2)^\mathsf{T}(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2)$$

☺ The HE estimator is unbiased.

☺ HE is computationally efficient: $O(dn^2)$ as opposed to $O(n^3)$ for REML

☹ May get negative estimates: truncation to zero?

[4] Haseman and Elston. *Behavior Genetics*. '72; Sofer T. *Stat. Appl. Genet. Mol. Biol.* '17

Restricted HE Regression

☺ Avoid negative estimates by non-negative least squares (NNLS)

Restricted HE Regression

☺ Avoid negative estimates by non-negative least squares (NNLS)

REHE solves for the variance components by minimizing

$$\frac{1}{n^*}(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2)^\intercal(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2) = \frac{1}{n^*}\left\{(\boldsymbol{\sigma}^2)^\intercal\widetilde{X}^\intercal\widetilde{X}\boldsymbol{\sigma}^2 - 2(\boldsymbol{\sigma}^2)^\intercal\widetilde{X}^\intercal\widetilde{Y}\right\},$$

subject to $\boldsymbol{\sigma}^2 \geq 0$.

Restricted HE Regression

☺ Avoid negative estimates by non-negative least squares (NNLS)

REHE solves for the variance components by minimizing

$$\frac{1}{n^*}(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2)^\intercal(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2) = \frac{1}{n^*}\left\{(\boldsymbol{\sigma}^2)^\intercal\widetilde{X}^\intercal\widetilde{X}\boldsymbol{\sigma}^2 - 2(\boldsymbol{\sigma}^2)^\intercal\widetilde{X}^\intercal\widetilde{Y}\right\},$$

subject to $\boldsymbol{\sigma}^2 \geq 0$.

☺ Global minimizer is guaranteed due to convexity.

Restricted HE Regression

☺ Avoid negative estimates by non-negative least squares (NNLS)

REHE solves for the variance components by minimizing

$$\frac{1}{n^*}(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2)^\intercal(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2) = \frac{1}{n^*}\left\{(\boldsymbol{\sigma}^2)^\intercal\widetilde{X}^\intercal\widetilde{X}\boldsymbol{\sigma}^2 - 2(\boldsymbol{\sigma}^2)^\intercal\widetilde{X}^\intercal\widetilde{Y}\right\},$$

subject to $\boldsymbol{\sigma}^2 \geq 0$.

☺ Global minimizer is guaranteed due to convexity.

☺ Computational cost of REHE is comparable to HE, both faster than REML.

Restricted HE Regression

☺ Avoid negative estimates by non-negative least squares (NNLS)

REHE solves for the variance components by minimizing

$$\frac{1}{n^*}(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2)^\intercal(\widetilde{Y} - \widetilde{X}\boldsymbol{\sigma}^2) = \frac{1}{n^*}\left\{(\boldsymbol{\sigma}^2)^\intercal\widetilde{X}^\intercal\widetilde{X}\boldsymbol{\sigma}^2 - 2(\boldsymbol{\sigma}^2)^\intercal\widetilde{X}^\intercal\widetilde{Y}\right\},$$

subject to $\boldsymbol{\sigma}^2 \geq 0$.

☺ Global minimizer is guaranteed due to convexity.

☺ Computational cost of REHE is comparable to HE, both faster than REML.

☹ May get zero estimates

REHE with Resampling

FRED HUTCH
CURES START HERE™

$$\frac{1}{n^*}\widetilde{X}^\intercal\widetilde{X} = \frac{1}{n^*}\sum_{i=1}^{n^*}\widetilde{X}_i^\intercal\widetilde{X}_i, \quad \frac{1}{n^*}\widetilde{X}^\intercal\widetilde{Y} = \frac{1}{n^*}\sum_{i=1}^{n^*}\widetilde{X}_i^\intercal\widetilde{Y}_i.$$

REHE with Resampling

$$\frac{1}{n^*}\widetilde{X}^\mathsf{T}\widetilde{X} = \frac{1}{n^*}\sum_{i=1}^{n^*}\widetilde{X}_i^\mathsf{T}\widetilde{X}_i, \quad \frac{1}{n^*}\widetilde{X}^\mathsf{T}\widetilde{Y} = \frac{1}{n^*}\sum_{i=1}^{n^*}\widetilde{X}_i^\mathsf{T}\widetilde{Y}_i.$$

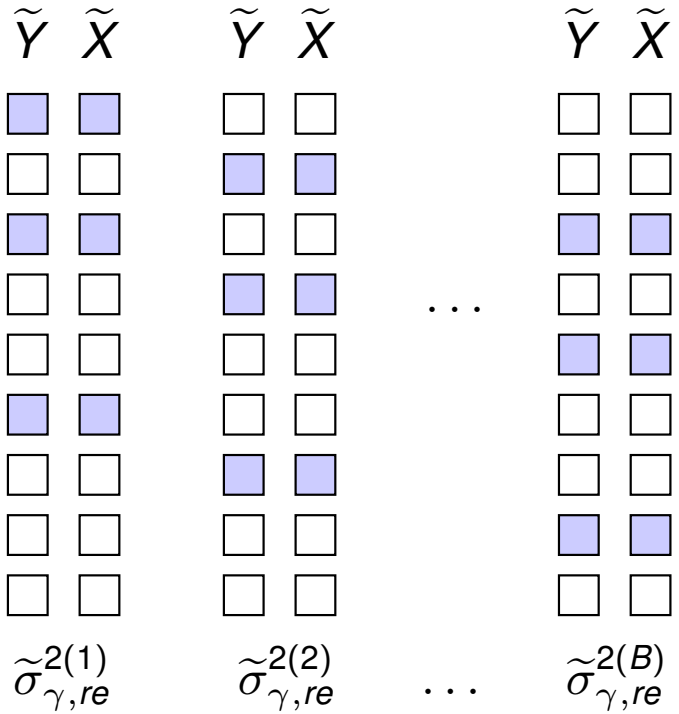
We can approximate these inner products by subsampling rows of $\widetilde{X}$ and $\widetilde{Y}$.

REHE with Resampling

$$\frac{1}{n^*}\widetilde{X}^\intercal\widetilde{X} = \frac{1}{n^*}\sum_{i=1}^{n^*}\widetilde{X}_i^\intercal\widetilde{X}_i, \quad \frac{1}{n^*}\widetilde{X}^\intercal\widetilde{Y} = \frac{1}{n^*}\sum_{i=1}^{n^*}\widetilde{X}_i^\intercal\widetilde{Y}_i.$$

We can approximate these inner products by subsampling rows of $\widetilde{X}$ and $\widetilde{Y}$.

REHE with Resampling (reREHE)



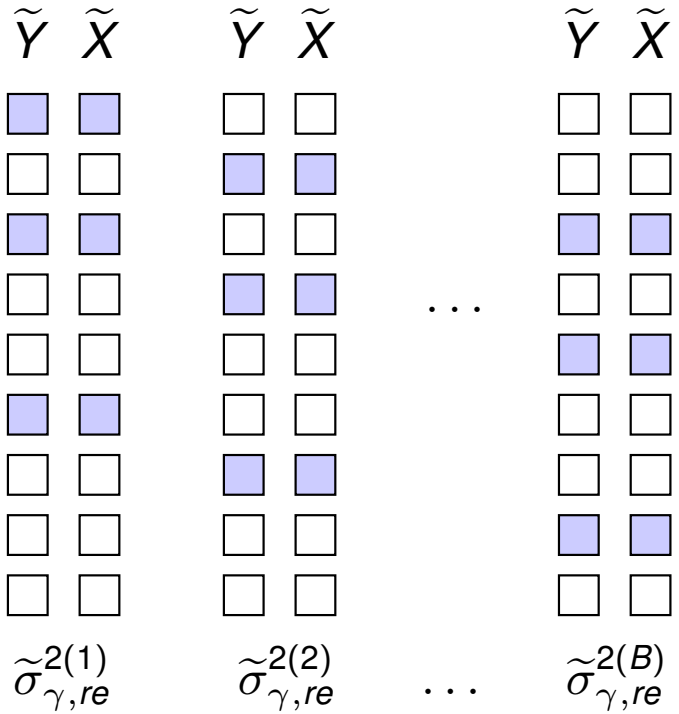
$$\widetilde{\sigma}^2_{\gamma,re} = \frac{1}{B}\sum_{b=1}^{B}\widetilde{\sigma}^{2(b)}_{\gamma,re}$$

REHE with Resampling

$$\frac{1}{n^*}\widetilde{X}^\intercal\widetilde{X} = \frac{1}{n^*}\sum_{i=1}^{n^*}\widetilde{X}_i^\intercal\widetilde{X}_i, \quad \frac{1}{n^*}\widetilde{X}^\intercal\widetilde{Y} = \frac{1}{n^*}\sum_{i=1}^{n^*}\widetilde{X}_i^\intercal\widetilde{Y}_i.$$

We can approximate these inner products by subsampling rows of $\widetilde{X}$ and $\widetilde{Y}$.

REHE with Resampling (reREHE)



$$\widetilde{\sigma}^2_{\gamma,re} = \frac{1}{B}\sum_{b=1}^{B}\widetilde{\sigma}^{2(b)}_{\gamma,re}$$

☺ reREHE estimates are strictly positive and can be faster to compute.

REHE with Fixed Effects

Suppose we have covariates. The null model

$$y = W\alpha + \gamma + \epsilon.$$

[5] $K^\dagger$ can be replaced by K when n is large.

REHE with Fixed Effects

Suppose we have covariates. The null model

$$y = W\alpha + \gamma + \epsilon.$$

Let $P^{\perp} = I_n - W(W^{\mathsf{T}}W)^{-1}W^{\mathsf{T}}$ denote the projection matrix onto the orthogonal complement of the column space of W. Let

$$y^{\dagger} = P^{\perp}y, \quad \gamma^{\dagger} = P^{\perp}\gamma, \quad \epsilon^{\dagger} = P^{\perp}\epsilon$$

[5] $K^{\dagger}$ can be replaced by K when n is large.

REHE with Fixed Effects

Suppose we have covariates. The null model

$$y = W\alpha + \gamma + \epsilon.$$

Let $P^{\perp} = I_n - W(W^{\mathsf{T}}W)^{-1}W^{\mathsf{T}}$ denote the projection matrix onto the orthogonal complement of the column space of W. Let

$$y^{\dagger} = P^{\perp}y, \quad \gamma^{\dagger} = P^{\perp}\gamma, \quad \epsilon^{\dagger} = P^{\perp}\epsilon$$

We obtain a new model with no covariates

$$y^{\dagger} = \gamma^{\dagger} + \epsilon^{\dagger}, \quad \gamma^{\dagger} \sim \mathcal{N}(0, \sigma_{\gamma}^2 K^{\dagger})$$

where $K^{\dagger} = P^{\perp}KP^{\perp}$[5].

[5] $K^{\dagger}$ can be replaced by K when n is large.

Constructing Confidence Intervals

[6] Can also construct quantile confidence interval

Constructing Confidence Intervals

Parametric Bootstrap

- Compute REHE estimates $\widetilde{\sigma}_\gamma^2, \widetilde{\sigma}_\epsilon^2$ based on $\widetilde{Y}, K, I_n$;
- For $b = 1$ to B
  - Generate response vector $\widetilde{Y}^{*(b)}$ from $\mathcal{N}\left(0, \widetilde{\sigma}_\gamma^2 K + \widetilde{\sigma}_\epsilon^2 I_n\right)$;
  - Compute REHE estimates $\widetilde{\sigma}_\gamma^{2(b)}$, $\widetilde{\sigma}_\epsilon^{2(b)}$, based on $\widetilde{Y}^{*(b)}, K, I_n$;

[6] Can also construct quantile confidence interval

Constructing Confidence Intervals

FRED HUTCH
CURES START HERE™

Parametric Bootstrap

- Compute REHE estimates $\widetilde{\sigma}_\gamma^2, \widetilde{\sigma}_\epsilon^2$ based on $\widetilde{Y}, K, I_n$;
- For $b = 1$ to B
  - Generate response vector $\widetilde{Y}^{*(b)}$ from $\mathcal{N}\left(0, \widetilde{\sigma}_\gamma^2 K + \widetilde{\sigma}_\epsilon^2 I_n\right)$;
  - Compute REHE estimates $\widetilde{\sigma}_\gamma^{2(b)}$, $\widetilde{\sigma}_\epsilon^{2(b)}$, based on $\widetilde{Y}^{*(b)}, K, I_n$;

Wald-type confidence interval[6]

$$\left[\widetilde{\sigma}_\gamma^2 - z_{\alpha/2} \times \text{s.e.}\left(\widetilde{\sigma}_\gamma^{2(b)}\right), \widetilde{\sigma}_\gamma^2 + z_{\alpha/2} \times \text{s.e.}\left(\widetilde{\sigma}_\gamma^{2(b)}\right)\right],$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$-th percentile of the standard normal distribution.

[6] Can also construct quantile confidence interval

GWAS for HCHS/SOL

FRED HUTCH
CURES START HERE™

- $n = 12,502$ after removing observations with missing values

GWAS for HCHS/SOL

FRED HUTCH
CURES START HERE™

- $n = 12,502$ after removing observations with missing values
- y: red blood cell count

GWAS for HCHS/SOL

FRED HUTCH
CURES START HERE™

- $n = 12,502$ after removing observations with missing values
- y: red blood cell count
- X: 4,100,028 SNPs

GWAS for HCHS/SOL

FRED HUTCH
CURES START HERE™

- $n = 12,502$ after removing observations with missing values
- y: red blood cell count
- X: 4,100,028 SNPs
- Covariates W: age, sex, cigarette use, field center indicator, genetic subgroup indicator, ancestry, sampling weights

GWAS for HCHS/SOL

FRED HUTCH
CURES START HERE™

- $n = 12,502$ after removing observations with missing values
- y: red blood cell count
- X: 4,100,028 SNPs
- Covariates W: age, sex, cigarette use, field center indicator, genetic subgroup indicator, ancestry, sampling weights
- Variance components: genetic relatedness, membership of household, and membership of community group

GWAS for HCHS/SOL

FRED HUTCH
CURES START HERE™

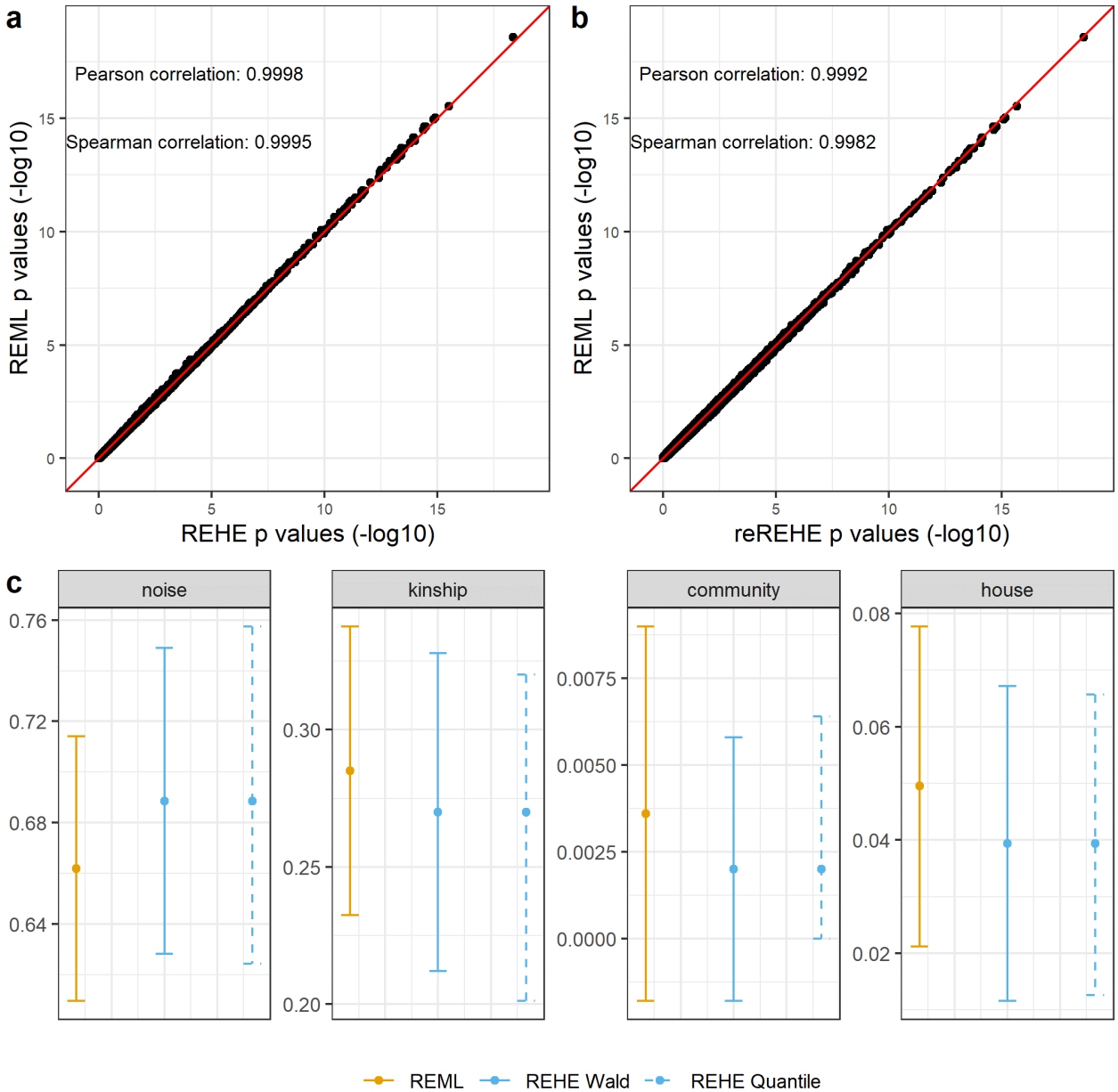
- $n = 12,502$ after removing observations with missing values
- y: red blood cell count
- X: 4,100,028 SNPs
- Covariates W: age, sex, cigarette use, field center indicator, genetic subgroup indicator, ancestry, sampling weights
- Variance components: genetic relatedness, membership of household, and membership of community group
- Bonferroni correction for multiple testing

GWAS for HCHS/SOL

FRED HUTCH
CURES START HERE™

- $n = 12,502$ after removing observations with missing values
- y: red blood cell count
- X: 4,100,028 SNPs
- Covariates W: age, sex, cigarette use, field center indicator, genetic subgroup indicator, ancestry, sampling weights
- Variance components: genetic relatedness, membership of household, and membership of community group
- Bonferroni correction for multiple testing
- REHE took 2.4 min for estimation and 18 min for inference; REML 23.9 min

GWAS for HCHS/SOL Results



Simulation

Synthetic data were generated from

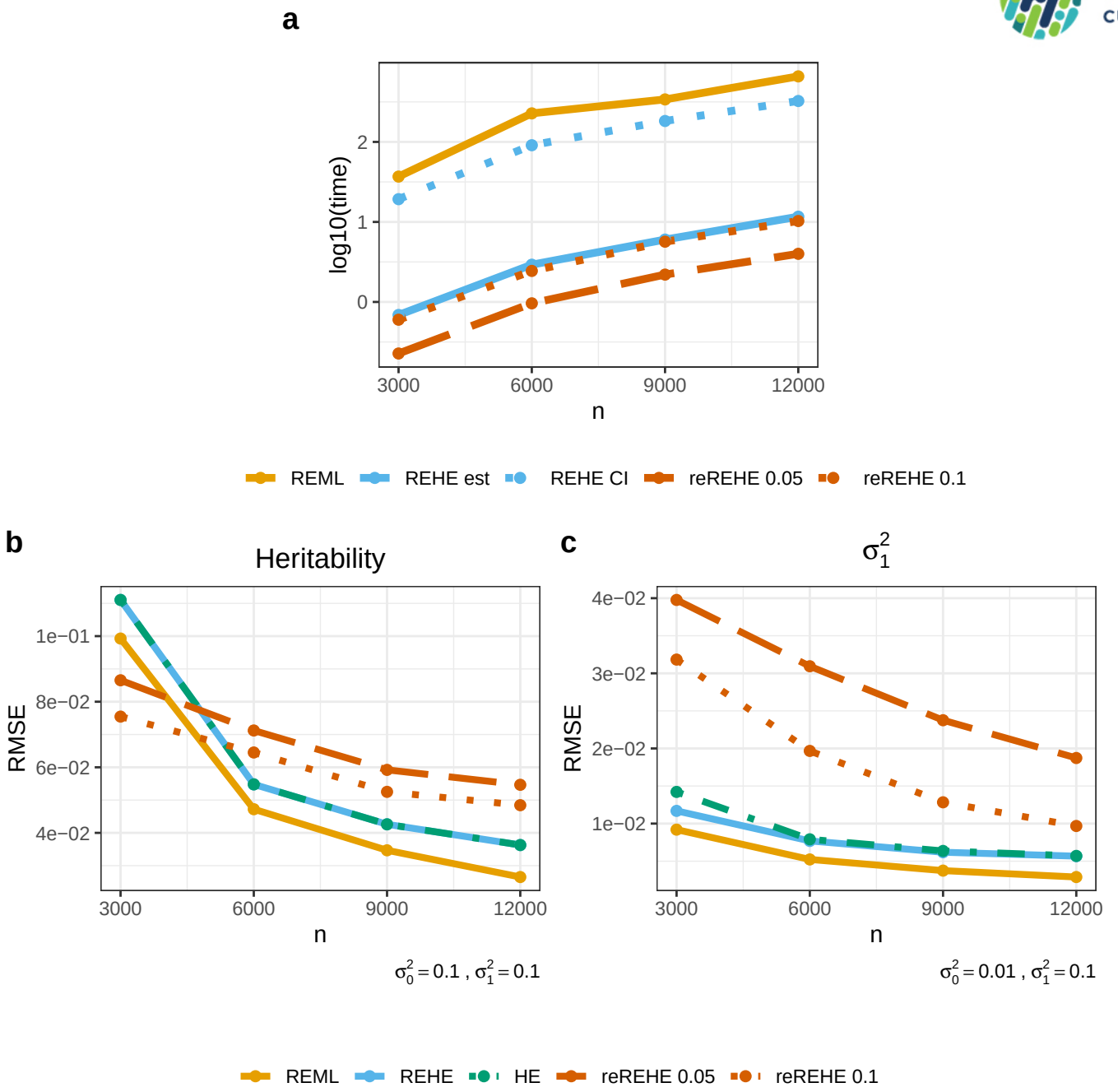
$$y = \sigma_0^2 I_n + \sigma_1^2 K_1,$$

where K_1 is a submatrix of the genetic relatedness matrix from HCHS/SOL.

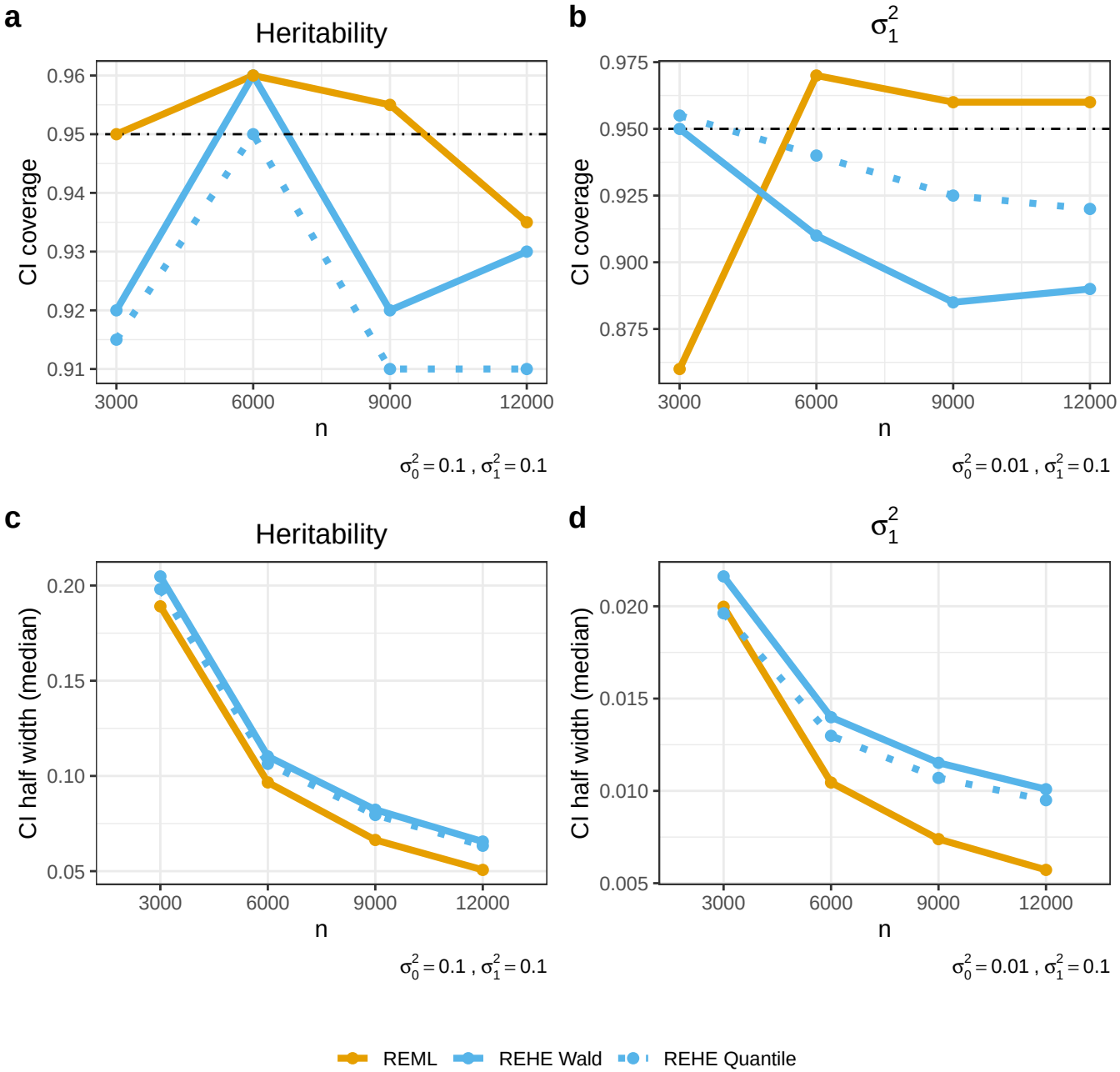
- $n \in \{3,000, 6,000, 9,000, 12,000\}$
- $(\sigma_0^2,\ \sigma_1^2) \in \{(0.1,\ 0.1), (0.01,\ 0.1)\}$

23% HE estimates were negative before truncation at zero ($n = 3000, \sigma_0^2 = 0.01$).

Estimation Results



Confidence Interval Results



Gene Set Analysis

Gene Set Analysis

FRED HUTCH
CURES START HERE™

Gene Set: a set of all SNPs located near a list of related genes.

Gene Set Analysis

FRED HUTCH
CURES START HERE™

Gene Set: a set of all SNPs located near a list of related genes.

Scientific Question: whether a *gene set* is associated with a trait.

Gene Set Analysis

FRED HUTCH
CURES START HERE™

Gene Set: a set of all SNPs located near a list of related genes.

Scientific Question: whether a *gene set* is associated with a trait.

Motivation: many biological processes are driven by mechanisms involving more than one SNP

Gene Set Analysis

FRED HUTCH
CURES START HERE™

Gene Set: a set of all SNPs located near a list of related genes.

Scientific Question: whether a *gene set* is associated with a trait.

Motivation: many biological processes are driven by mechanisms involving more than one SNP

☺ Easy interpretation

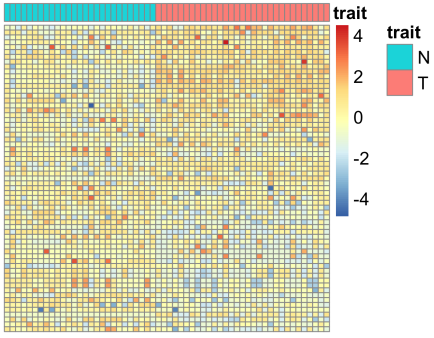
☺ Fewer number of gene sets compared to number of genes/SNPs

☺ More power by pooling many weaker signals

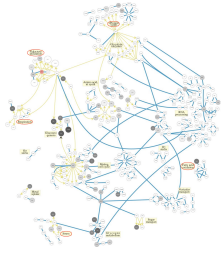
Gene Set Analysis

FRED HUTCH
CURES START HERE™

Input



Pathway Database



$\Rightarrow$

Methods

- GSEA
- SPIA
- DEGraph
- NetGSA
- ...

$\Rightarrow$

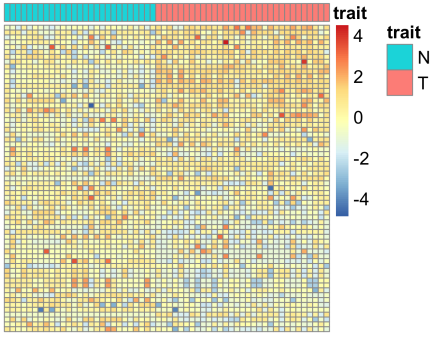
Output

List of significant pathway

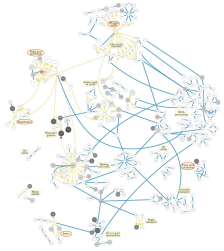
Gene Set Analysis

FRED HUTCH
CURES START HERE™

Input



Pathway Database



Methods

GSEA

SPIA

DEGraph

NetGSA

...

Output

List of significant pathway

Pathway Database

KEGG, MSigDB, BioCarta, Reactome, MetaCyc, etc.

Topology-based Gene Set Analysis

FRED HUTCH
CURES START HERE™

Motivation: genes are not independent

FRED HUTCH
CURES START HERE™

Motivation: genes are not independent

Most existing methods rely on curated interactions from pathway databases.

☹ Curated networks can be incomplete and/or inaccurate

☹ Curated networks lack condition/disease-specific alterations in interactions

Topology-based Gene Set Analysis

FRED HUTCH
CURES START HERE™

Motivation: genes are not independent

Most existing methods rely on curated interactions from pathway databases.

☹ Curated networks can be incomplete and/or inaccurate

☹ Curated networks lack condition/disease-specific alterations in interactions

Which null hypothesis?

FRED HUTCH
CURES START HERE™

Motivation: genes are not independent

Most existing methods rely on curated interactions from pathway databases.

☹ Curated networks can be incomplete and/or inaccurate

☹ Curated networks lack condition/disease-specific alterations in interactions

Which null hypothesis?

- The genes in a given pathway are at most as differentially expressed as those outside the pathway (camera, PathNet).

FRED HUTCH
CURES START HERE™

Motivation: genes are not independent

Most existing methods rely on curated interactions from pathway databases.

☹ Curated networks can be incomplete and/or inaccurate

☹ Curated networks lack condition/disease-specific alterations in interactions

Which null hypothesis?

- The genes in a given pathway are at most as differentially expressed as those outside the pathway (camera, PathNet).
- The observed number of DE genes is just by chance and the DE genes are randomly located in the pathway (SPIA, Pathway-Express)

FRED HUTCH
CURES START HERE™

Motivation: genes are not independent

Most existing methods rely on curated interactions from pathway databases.

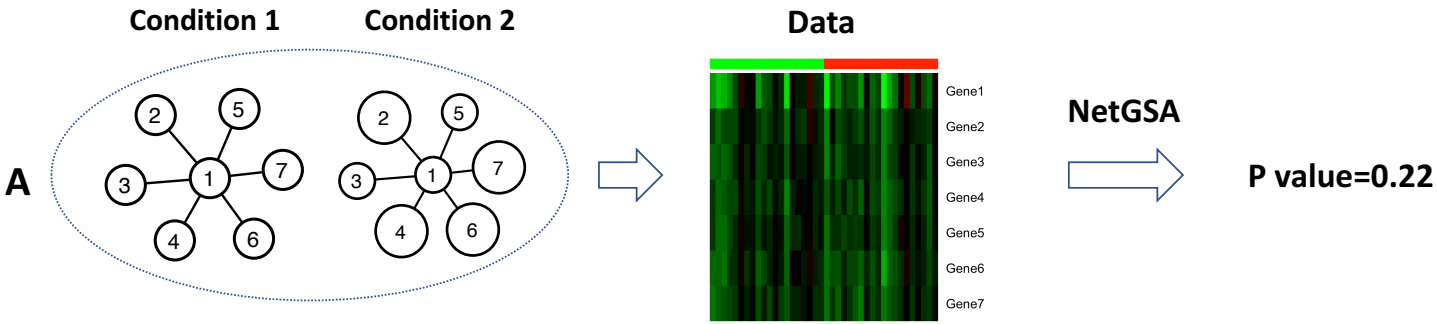
☹ Curated networks can be incomplete and/or inaccurate

☹ Curated networks lack condition/disease-specific alterations in interactions

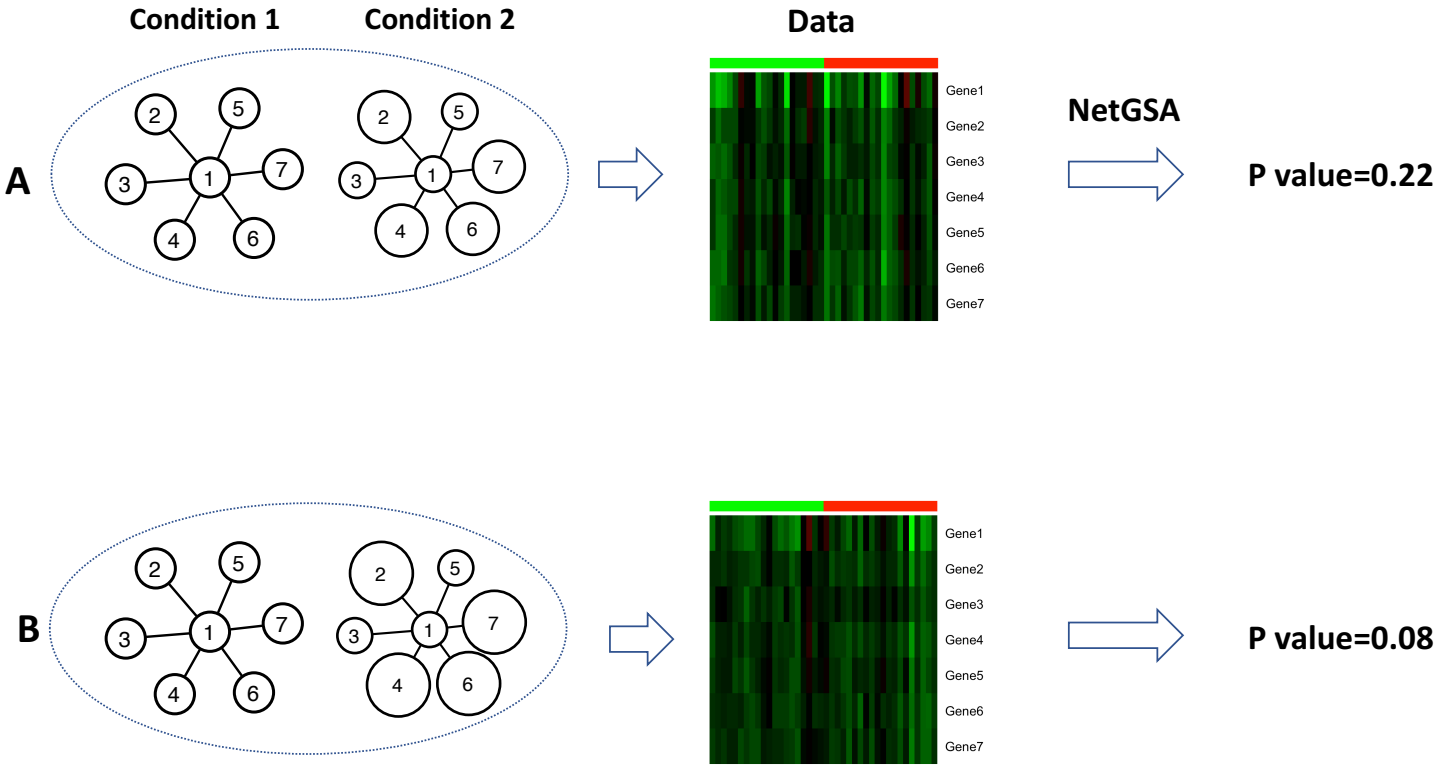
Which null hypothesis?

- The genes in a given pathway are at most as differentially expressed as those outside the pathway (camera, PathNet).
- The observed number of DE genes is just by chance and the DE genes are randomly located in the pathway (SPIA, Pathway-Express)
- Self-contained null (NetGSA, DEGraph and topologyGSA)

NetGSA - Toy Example

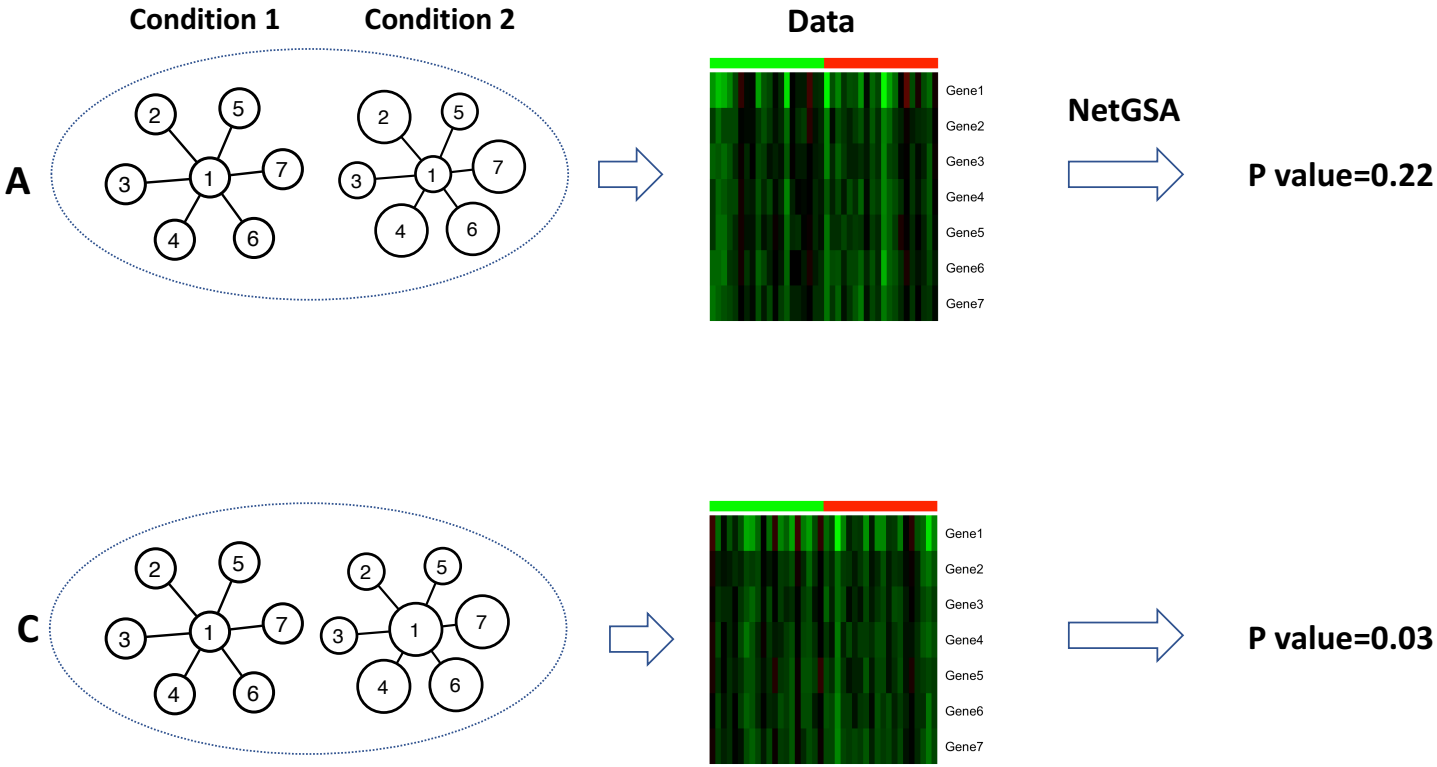


NetGSA - Toy Example



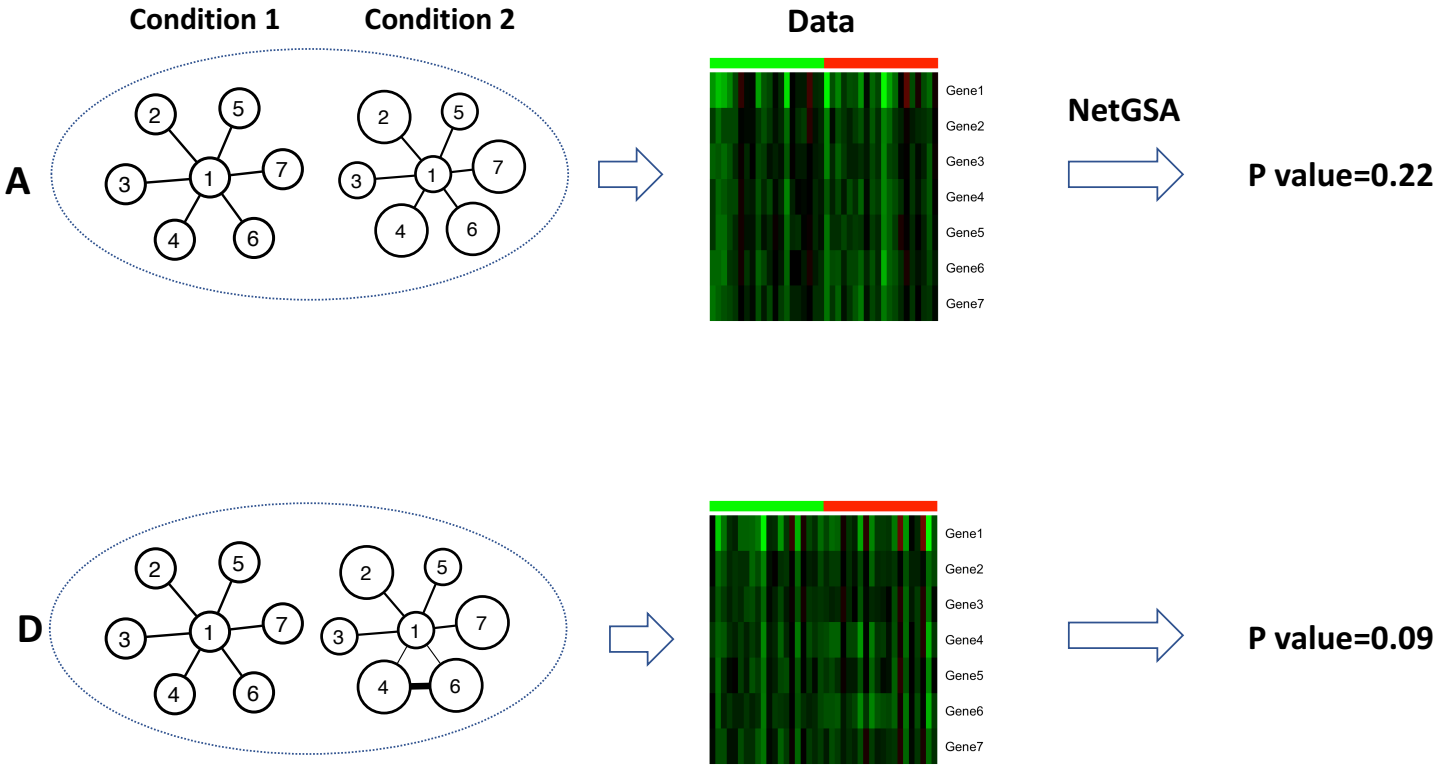
Nodes 2, 4, 6, 7 have larger changes in mean in case B than in case A.

NetGSA - Toy Example



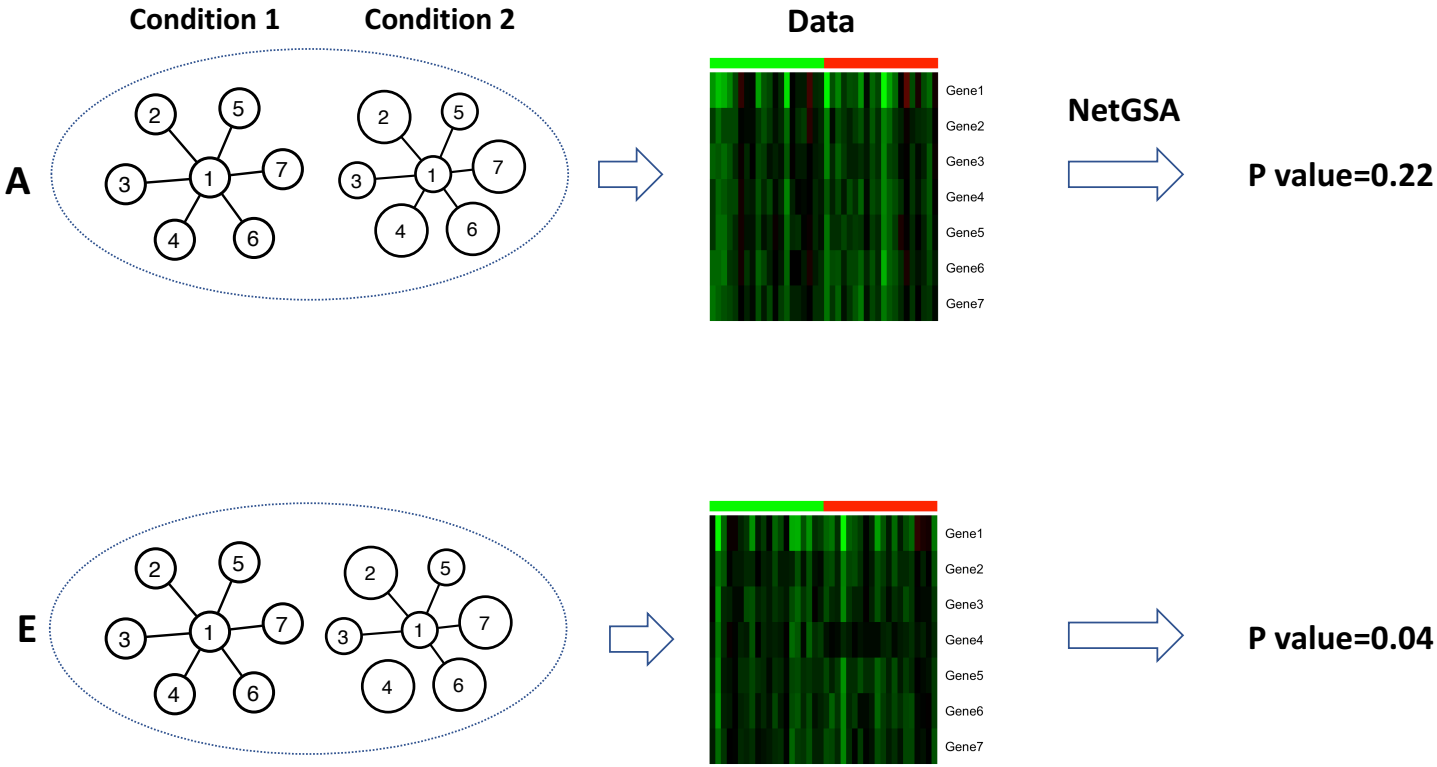
Node 1 as opposed to node 2 has change in mean in case C.

NetGSA - Toy Example



There is an additional change in correlation between nodes 4 and 6 in case D.

NetGSA - Toy Example



There is an additional change in correlation between nodes 1 and 4 in case E.

What Drives Gene Set Significance

FRED HUTCH
CURES START HERE™

- Change in mean values of genes in the set
- Position of genes: hub genes are more important
- Change in gene-gene interaction

What Drives Gene Set Significance

FRED HUTCH
CURES START HERE™

- Change in mean values of genes in the set
- Position of genes: hub genes are more important
- Change in gene-gene interaction

NetGSA captures all three factors!

NetGSA Model

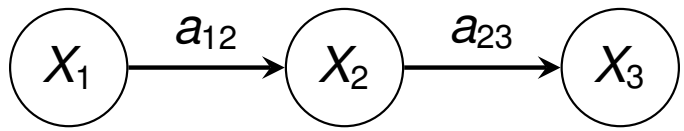
Let $Y \in \mathbb{R}^p$ denote the expression values of p genes from an arbitrary sample. Suppose $Y = X + \epsilon$, where X is signal and ϵ is noise.

[7] Shojaie and Michailidis. *JCB*. '09

NetGSA Model

Let $Y \in \mathbb{R}^p$ denote the expression values of p genes from an arbitrary sample. Suppose $Y = X + \epsilon$, where X is signal and ϵ is noise.

Assume the p genes are related via a network $A = (a_{ij})$ where a_{ij} denotes the strength of association between genes i and j.

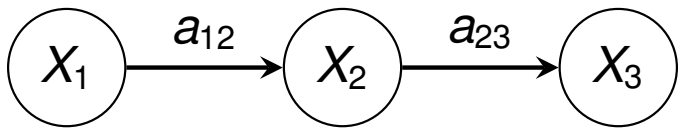


[7] Shojaie and Michailidis. *JCB*. '09

NetGSA Model

Let $Y \in \mathbb{R}^p$ denote the expression values of p genes from an arbitrary sample. Suppose $Y = X + \epsilon$, where X is signal and ϵ is noise.

Assume the p genes are related via a network $A = (a_{ij})$ where a_{ij} denotes the strength of association between genes i and j.



We model X via the latent variable model[7]

$$
\begin{aligned}
X_1 &= \gamma_1 \\
X_2 &= a_{12} X_1 + \gamma_2 \\
X_3 &= a_{23} X_2 + \gamma_3 = a_{12} a_{23} \gamma_1 + a_{23} \gamma_2 + \gamma_3
\end{aligned}
$$

where $\gamma_j \sim \mathcal{N}(\mu_j, \sigma^2_\gamma)$ represents the baseline expression of gene j.

[7] Shojaie and Michailidis. *JCB*. '09

NetGSA Model

$$Y = \Lambda\gamma + \epsilon, \quad \gamma \sim \mathcal{N}(\mu, \sigma_\gamma^2 I_p), \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I_p)$$

where

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ a_{12} & 1 & 0 \\ a_{12}a_{23} & a_{23} & 1 \end{pmatrix}$$

is the influence matrix of the gene network $\Lambda = (I_p - A)^{-1}$.

NetGSA Model

$$Y = \Lambda\gamma + \epsilon, \quad \gamma \sim \mathcal{N}(\mu, \sigma_\gamma^2 I_p), \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I_p)$$

where

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ a_{12} & 1 & 0 \\ a_{12}a_{23} & a_{23} & 1 \end{pmatrix}$$

is the influence matrix of the gene network $\Lambda = (I_p - A)^{-1}$.

Statistical Inference

Given data Y_i $(i = 1, \ldots, n)$ and network A, test for a gene set G

$$H_0 : \mu_G^{(1)} = \mu_G^{(2)}$$

NetGSA Model

$$Y = \Lambda\gamma + \epsilon, \quad \gamma \sim \mathcal{N}(\mu, \sigma_\gamma^2 I_p), \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I_p)$$

where

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ a_{12} & 1 & 0 \\ a_{12}a_{23} & a_{23} & 1 \end{pmatrix}$$

is the influence matrix of the gene network $\Lambda = (I_p - A)^{-1}$.

Statistical Inference

Given data Y_i $(i = 1, \ldots, n)$ and network A, test for a gene set G

$$H_0 : \mu_G^{(1)} = \mu_G^{(2)}$$

or

$$H_0^{net} : (\Lambda^{(1)}\mu^{(1)})_G = (\Lambda^{(2)}\mu^{(2)})_G$$

Network Information

A can be directed acyclic or undirected.

[8] Ma et al. *Bioinformatics*. '16

Network Information

A can be directed acyclic or undirected.

A is weighted.

[8] Ma et al. *Bioinformatics*. '16

Network Information

A can be directed acyclic or undirected.

A is weighted.

NetGSA infers the weights from data (independent from Y) using graphical models.

[8] Ma et al. *Bioinformatics*. '16

Network Information

A can be directed acyclic or undirected.

A is weighted.

NetGSA infers the weights from data (independent from Y) using graphical models.

☺ Many RNA-seq data are available

[8] Ma et al. *Bioinformatics*. '16

Network Information

A can be directed acyclic or undirected.

A is weighted.

NetGSA infers the weights from data (independent from Y) using graphical models.

☺ Many RNA-seq data are available

☺ Can use curated networks as side information to improve data-driven network inference[8]

[8] Ma et al. *Bioinformatics*. '16

Incomplete Network Information

$$A = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{pmatrix} \cdot & ? & 1 & 0 & ? & 0 \\ ? & \cdot & ? & ? & 0 & ? \\ 1 & ? & \cdot & ? & 0 & 0 \\ 0 & ? & ? & \cdot & ? & 1 \\ ? & 0 & 0 & ? & \cdot & ? \\ 0 & ? & 0 & 1 & ? & \cdot \end{pmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}$$

- 0: there is no interaction; 1: there is interaction; ?: unknown

Incomplete Network Information

$$A = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{pmatrix} \cdot & ? & 1 & 0 & ? & 0 \\ ? & \cdot & ? & ? & 0 & ? \\ 1 & ? & \cdot & ? & 0 & 0 \\ 0 & ? & ? & \cdot & ? & 1 \\ ? & 0 & 0 & ? & \cdot & ? \\ 0 & ? & 0 & 1 & ? & \cdot \end{pmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}$$

- 0: there is no interaction; 1: there is interaction; ?: unknown
- Given data, we use graphical models to incorporate existing information using a constrained optimization framework.

Incomplete Network Information

$$A = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{pmatrix} \cdot & ? & 1 & 0 & ? & 0 \\ ? & \cdot & ? & ? & 0 & ? \\ 1 & ? & \cdot & ? & 0 & 0 \\ 0 & ? & ? & \cdot & ? & 1 \\ ? & 0 & 0 & ? & \cdot & ? \\ 0 & ? & 0 & 1 & ? & \cdot \end{pmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}$$

- 0: there is no interaction; 1: there is interaction; ?: unknown
- Given data, we use graphical models to incorporate existing information using a constrained optimization framework.
- Can estimate novel interactions and validate existing information.

Incomplete Network Information

$$A = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{pmatrix} \cdot & ? & 1 & 0 & ? & 0 \\ ? & \cdot & ? & ? & 0 & ? \\ 1 & ? & \cdot & ? & 0 & 0 \\ 0 & ? & ? & \cdot & ? & 1 \\ ? & 0 & 0 & ? & \cdot & ? \\ 0 & ? & 0 & 1 & ? & \cdot \end{pmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}$$

- 0: there is no interaction; 1: there is interaction; ?: unknown
- Given data, we use graphical models to incorporate existing information using a constrained optimization framework.
- Can estimate novel interactions and validate existing information.
- Consistent estimation of network requires fewer observations, depending on the available external information.

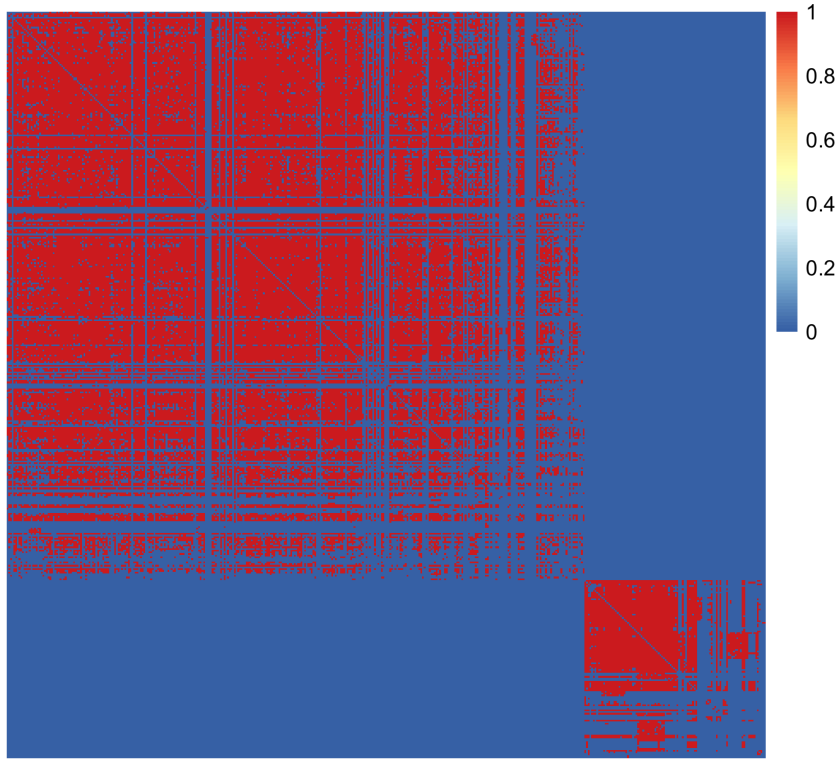
Large Networks

FRED HUTCH
CURES START HERE™

[9] Hellstern et al. *PLoS Comp Bio*. '21

Large Networks

Partition large networks into smaller ones by estimating a block diagonal network.



This strategy improves computational speed with little loss in performance[9].

[9] Hellstern et al. *PLoS Comp Bio*. '21

Incomplete Pathway Information

Pathway memberships may be unknown.

[10] Ma et al. *Bioinformatics*. '19

Incomplete Pathway Information

Pathway memberships may be unknown.

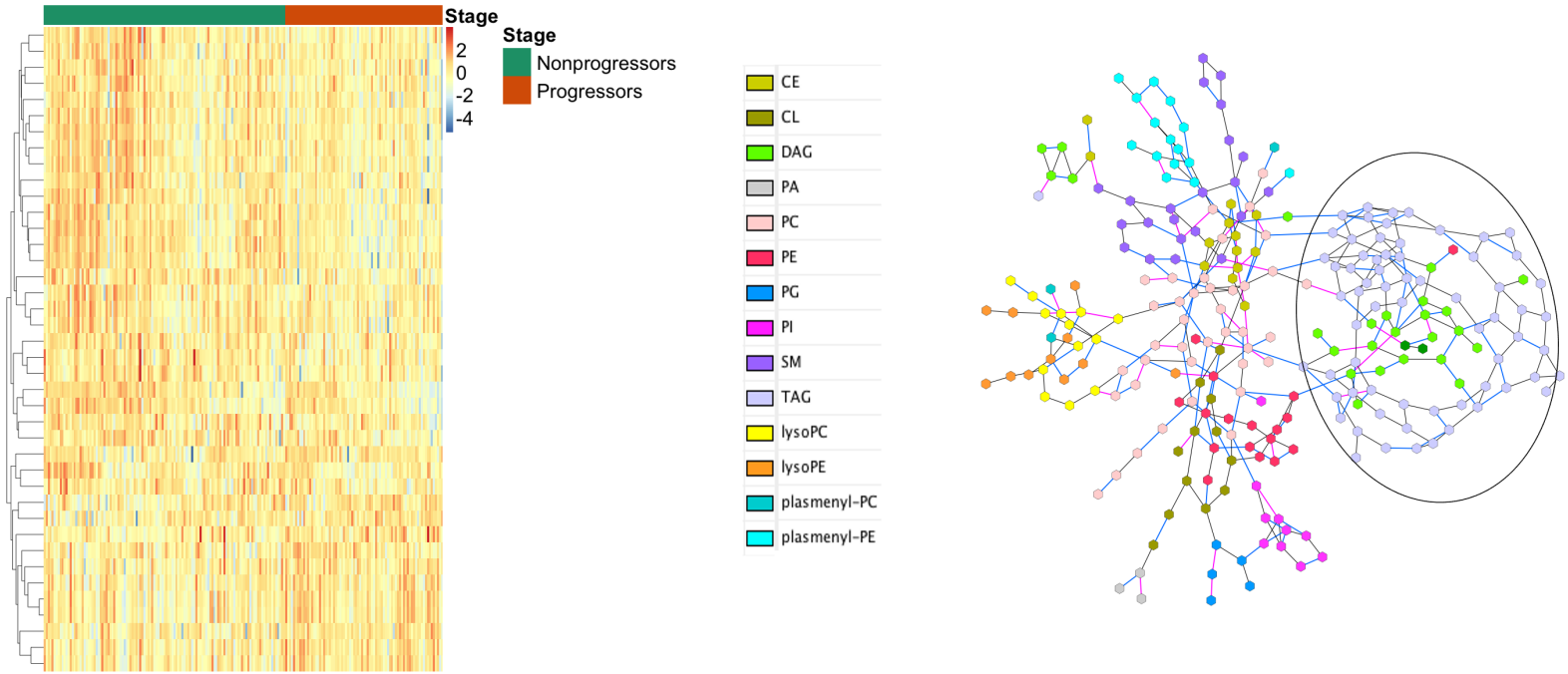


Fig: Inferred lipid interaction network in Chronic Kidney Disease progression

DNEA[10] uses data to estimate the network topology, identify modules by consensus clustering of the network, and perform enrichment analysis.

[10] Ma et al. *Bioinformatics*. '19

Topology-based Methods

FRED HUTCH
CURES START HERE™

Competitive null:

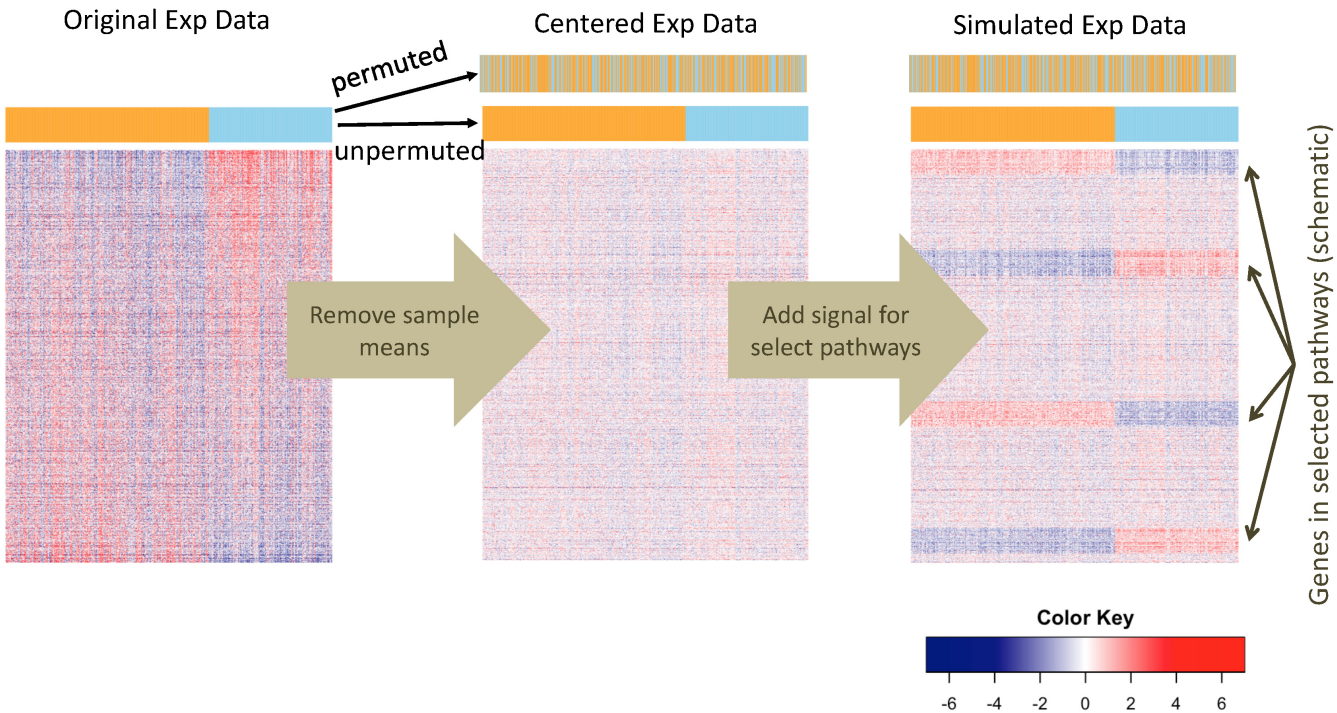
- SPIA (Tarca et al. '09)
- camera (Wu and Smyth, '12)
- PathNet (Dutta, et al. '12)

Self-contained null:

- topologyGSA (Massa et al. '10)
- DEGraph (Jacob et al. '12)
- NetGSA (Ma et al. '16)

Simulation I

Synthetic data were generated from TCGA[11]. $p = 2598$ genes; $n_1 = 403$ ER positive samples; $n_2 = 117$ ER negative samples.



Permuting the sample labels removes any difference in gene-gene correlation.

[11] TCGA. *Nature*. '12

Type I Error

100 KEGG pathways (`graphite` R package).

Table 2 Average type I errors over multiple pathways, grouped by pathway sizes, for the TCGA breast cancer study [26].

	Pathway size	
Method	$<= 75$	> 75
Pathway-Express	0*	0*
NetGSA	0.052	0.103
SPIA	0*	0*
topologyGSA	0.506	0.754
CAMERA	0.002	0.003
DEGraph	0.001	0.001
PathNet	0.048	0.057

* Under the self-contained null, the number of DE genes is zero. SPIA and Pathway-Express can not assess the impact of pathways that do not have any DE genes.

Power of Selected Pathways

Clockwise from top left to bottom left: *Glucagon signaling pathway, AMPK signaling pathway, Insulin signaling pathway*, and *B cell receptor signaling pathway*.

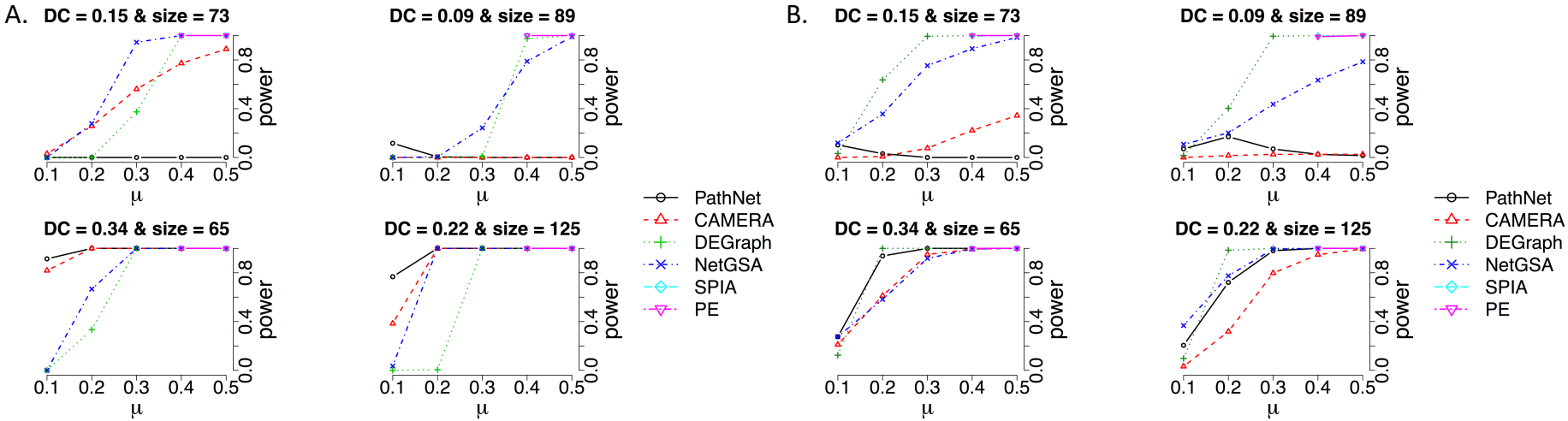


Fig: A: sample labels same as in TCGA; B: sample labels permuted.

Average Power

Powers are averaged over multiple pathways that have similar proportion of affected genes.

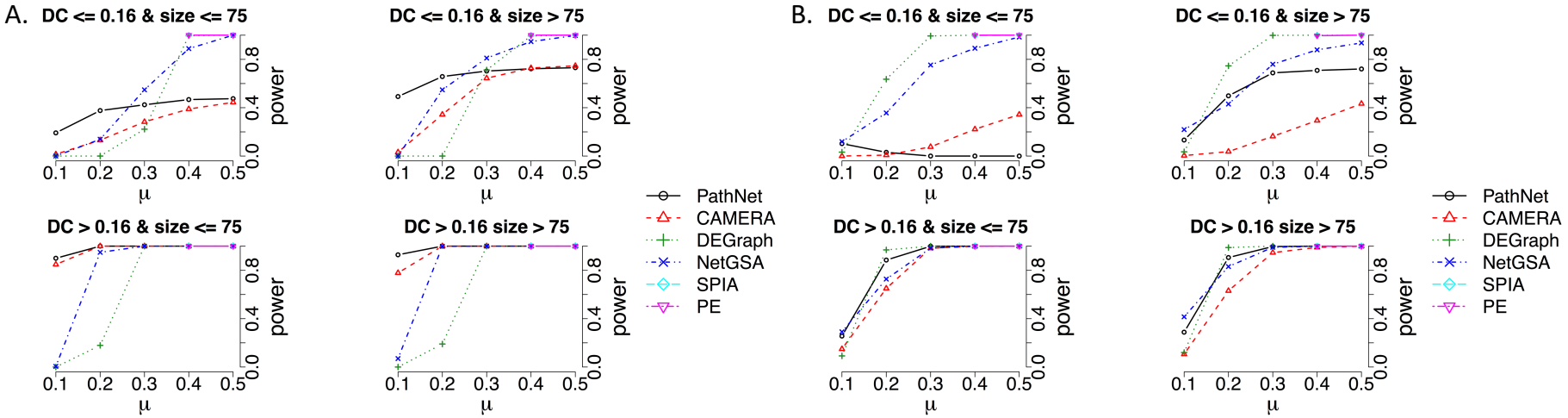
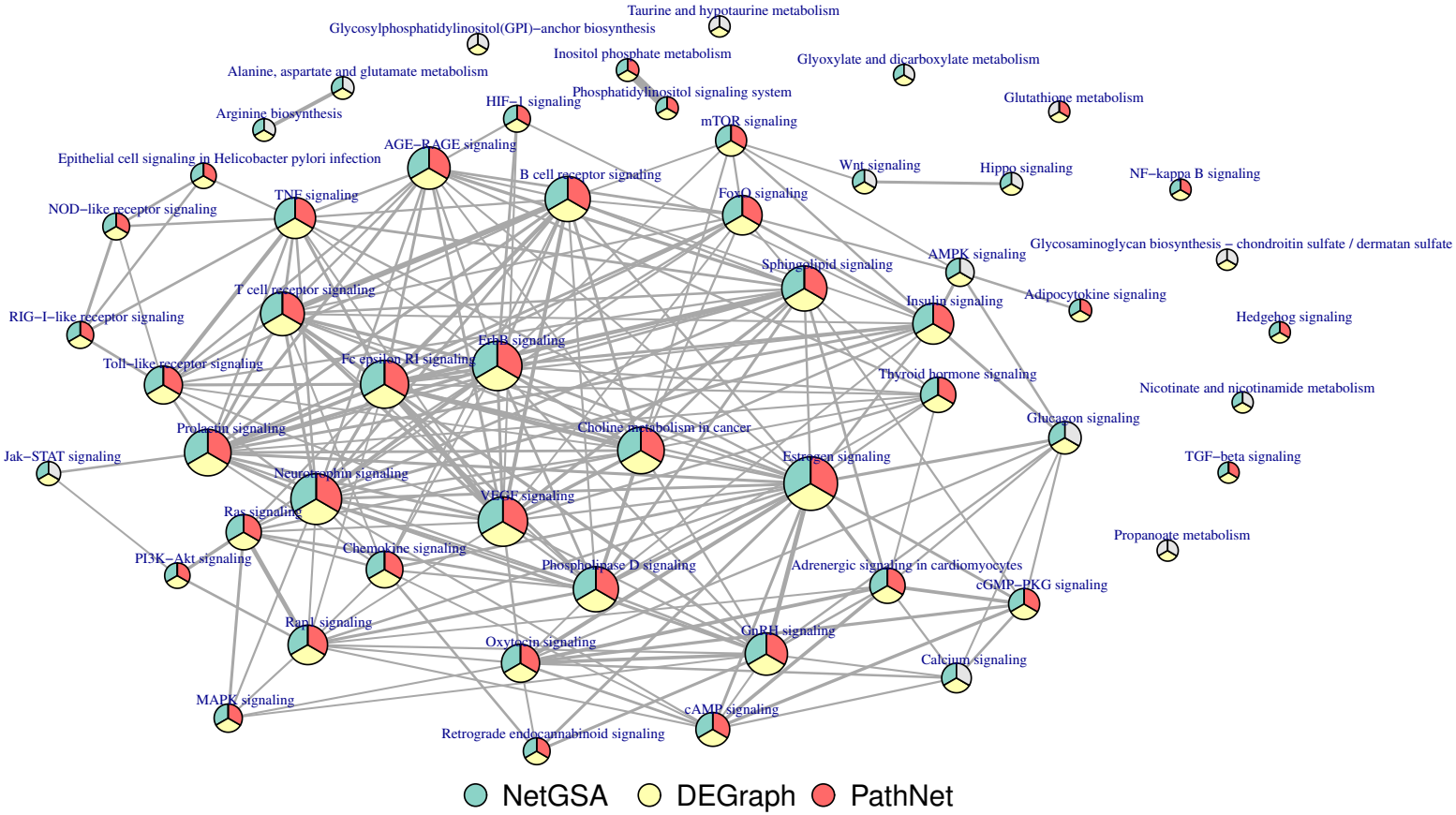


Fig: A: sample labels same as in TCGA; B: sample labels permuted.

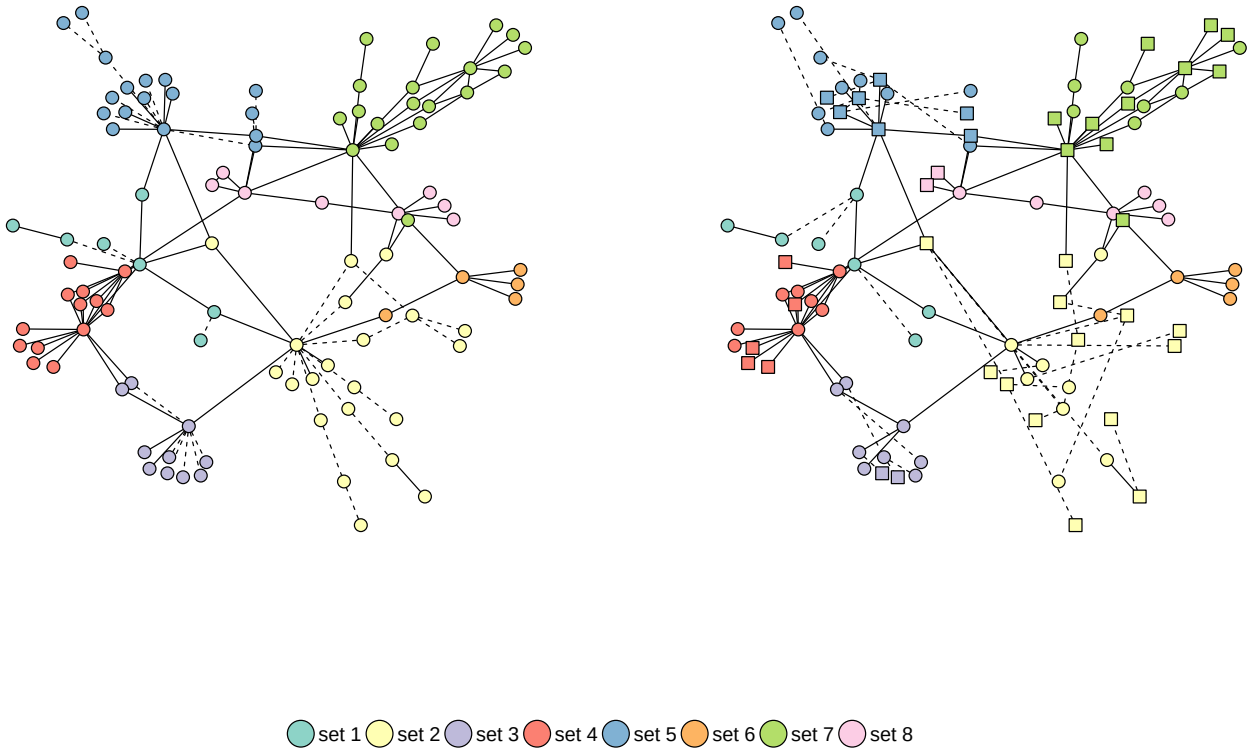
Analysis of TCGA Data



- Nodes: pathways
- Edges: share of genes (top 5%)

Simulation II

Synthetic data were generated from a DREAM network with changes in network topology.



Simulation II Results

sets 1, 6: no change

sets 3, 8: 20% nodes with differential means

sets 4, 5: 40% nodes with differential means

sets 2, 7: 60% nodes with differential means

sets 1, 2, 3, 5: also have changes in topology

Table: Empirical powers averaged in 100 replications.

Method	1	2	3	4	5	6	7	8
NetGSA	0.08	0.89	0.96	0.14	0.99	0.02	0.94	0.03
DEGraph	0.18	1.00	1.00	0.49	1.00	0.06	0.62	0.31
true power	0.12	0.93	0.98	0.11	0.99	0.05	0.95	0.10

Summary

- REHE offers gain in computational efficiency with little loss in accuracy for fitting large-scale linear mixed models.

Summary

FRED HUTCH
CURES START HERE™

- REHE offers gain in computational efficiency with little loss in accuracy for fitting large-scale linear mixed models.
- NetGSA tests for gene set enrichment by incorporating the topology.

Summary

FRED HUTCH
CURES START HERE™

- REHE offers gain in computational efficiency with little loss in accuracy for fitting large-scale linear mixed models.
- NetGSA tests for gene set enrichment by incorporating the topology.
- NetGSA can leverage existing network information and expression data.

Summary

FRED HUTCH
CURES START HERE™

- REHE offers gain in computational efficiency with little loss in accuracy for fitting large-scale linear mixed models.
- NetGSA tests for gene set enrichment by incorporating the topology.
- NetGSA can leverage existing network information and expression data.
- Caveat in gene set analysis: null hypothesis

References

1 **Ma J**, Shojaie, A and Michailidis, G. Network-based pathway enrichment analysis with incomplete network information. *Bioinformatics*. 32(20):3165–3174, 2016.

2 **Ma J**[†], Shojaie A and Michailidis G. A comparative study of topology-based pathway enrichment analysis methods. *BMC Bioinformatics*. 20 (546). 2019

3 **Ma J**, Karnovsky A, Afshinnia F, Wigginton J, Feldman H, Rader D, Shama K, Porter A, Rahman M, He J, Hamm L, Shafi T, Pennathur S, Michailidis G. Differential network-based enrichment analysis of lipid pathways altered in Chronic Kidney Disease progression. *Bioinformatics*. 35(18):3441–3452, 2019.

4 Hellstern M, **Ma J**, Yue K and Shojaie A. netgsa: Fast computation and interactive visualization for topology-based pathway enrichment analysis. *PLoS Computational Biology*. 17(6): e1008979, 2021.

5 Yue K, **Ma J**, Thornton T and Shojaie A. REHE: fast variance components estimation for linear mixed models. *Genetic Epidemiology*. 45(8):891–905, 2021.