Variance Components Estimation for Linear Mixed Models

Jing Ma

Public Health Sciences Division
Fred Hutch Cancer Research Center

Banff Workshop
February 5, 2019

Heritability Estimation

The ACE model

$$y = \mathbf{X}\beta + \gamma + c + e$$

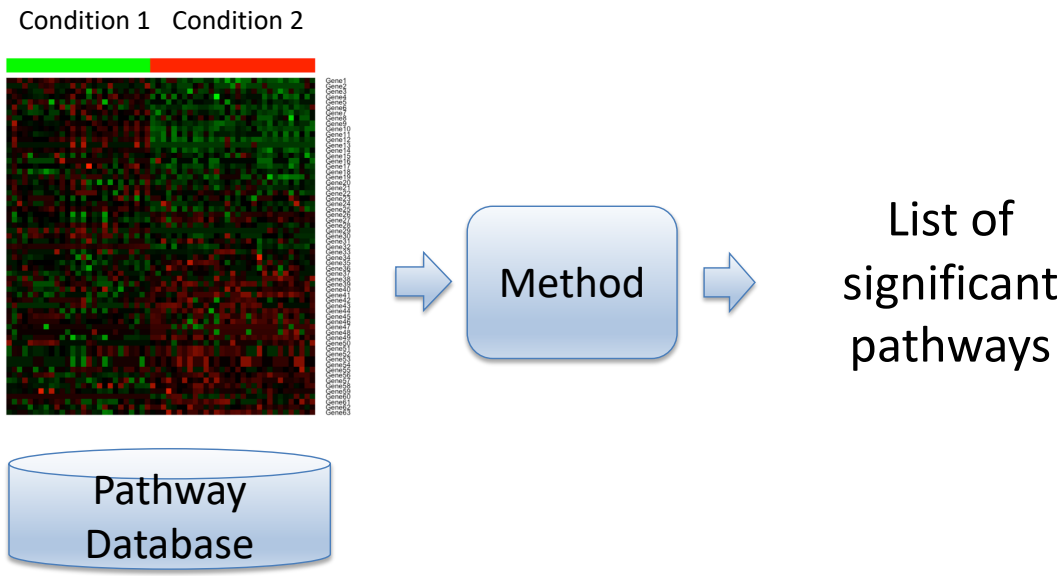
$$\gamma \sim N(0, \sigma_\gamma^2 A), \quad c \sim N(0, \sigma_c^2 C), \quad e \sim N(0, \sigma_e^2 I_n)$$

- y: $n \times 1$ vector of quantitative traits
- A: $n \times n$ genetic related matrix
- C: $n \times n$ matrix for shared environment
- $h^2 = \sigma_\gamma^2/(\sigma_\gamma^2 + \sigma_c^2 + \sigma_e^2)$

Pathway Enrichment Analysis

Scientific Question

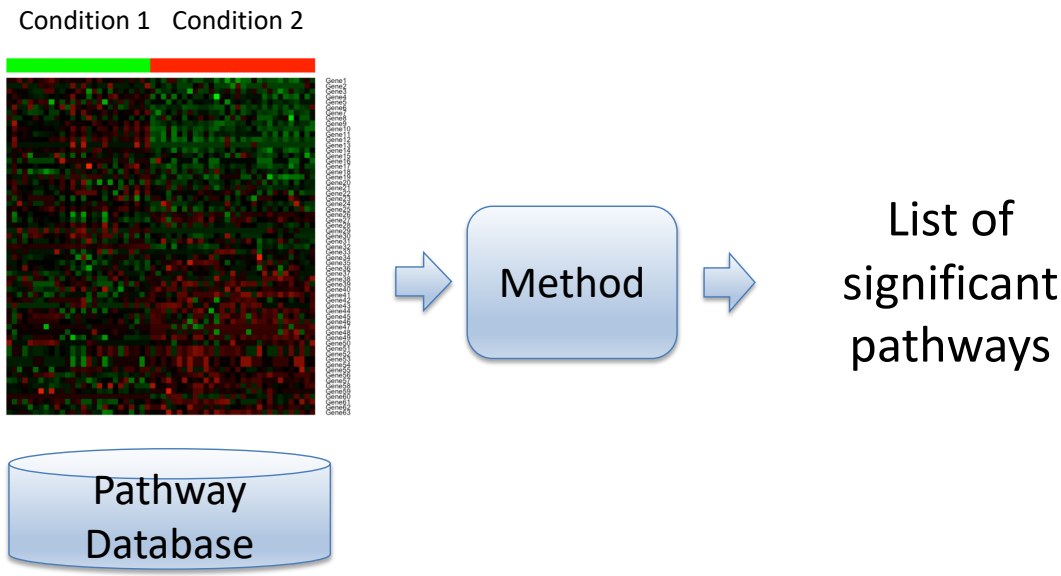
Whether a *genetic/metabolic pathway* is involved in responding to changes in environmental conditions or in specific cell functions.



FRED HUTCH
CURES START HERE™

Scientific Question

Whether a *genetic/metabolic pathway* is involved in responding to changes in environmental conditions or in specific cell functions.



- Reduce the complexity; more explanatory power.

Pathway Enrichment Analysis

The NetGSA model[1] (for $t = 1, 2, i = 1, \ldots, n_t$)

$$y_i^{(t)} = \mu^{(t)} + \gamma_i^{(t)} + e_i^{(t)}$$
$$\gamma^{(t)} \sim N(0, A_t)$$
$$e_i^{(t)} \sim N(0, \sigma_e^2 I_p)$$

- $y_i^{(t)}$: $p \times 1$ vector of observation for individual i in group t
- A_t^{-1}: $p \times p$ network information matrix
- Test statistic for $H_0 : \mu_G^{(1)} = \mu_G^{(2)}$ depends on the variance components.

[1] Ma J, et al. Bioinformatics. 2016

VC Estimation with REML

FRED HUTCH
CURES START HERE™

Pros

- Statistically efficient

Cons

- Need to invert $n \times n$ matrices (or $p \times p$ matrices in NetGSA) $\rightarrow$ computationally expensive, e.g.
  - $n > 100K$ in heritability estimation
  - $p \approx 3K$ in enrichment analysis

Haseman-Elston Regression[2]

The residual after removing fixed effects is

$$\varepsilon = y - \mathbf{X}\beta = \gamma + c + e,$$

whose second moment is

$$\mathbb{E}(\varepsilon\varepsilon') = \sigma_\gamma^2 A + \sigma_c^2 C + \sigma_e^2 I_n.$$

[2] Sofer T. Stat. Appl. Genet. Mol. Biol. 2017

Haseman-Elston Regression[2]

The residual after removing fixed effects is

$$\varepsilon = y - \mathbf{X}\beta = \gamma + c + e,$$

whose second moment is

$$\mathbb{E}(\varepsilon\varepsilon') = \sigma_\gamma^2 A + \sigma_c^2 C + \sigma_e^2 I_n.$$

Let $\mathrm{Vec}(A)$ denote the upper triangular part of a matrix A including the diagonal. The HE method solves for σ_j^2 by regressing

$$\widetilde{Y} = \mathrm{Vec}(\hat{\varepsilon}\hat{\varepsilon}') \in \mathbb{R}^{n^*}, \quad n^* = \frac{n(n+1)}{2}$$

on the design matrix

$$\widetilde{X} = [\mathrm{Vec}(I_n), \mathrm{Vec}(A), \mathrm{Vec}(C)] \in \mathbb{R}^{n^* \times 3}.$$

[2] Sofer T. Stat. Appl. Genet. Mol. Biol. 2017

Haseman-Elston Regression

FRED HUTCH
CURES START HERE™

Pros

- Only need to invert a 3×3 matrix

Cons

- May get negative estimates
- Computational cost is $O(n^2)$ in the ACE model and $O(np^2)$ in NetGSA $\rightarrow$ can be inefficient if n and p are large

Restricted HE Regression

Negative estimates

↓

Use non-negative least squares (NNLS)

NNLS

NNLS solves for the variance components by minimizing

$$\frac{1}{n^*}\left\{\theta'\widetilde{X}'\widetilde{X}\theta - 2\theta'\widetilde{X}'\widetilde{Y}\right\}, \quad s.t.\ \theta \geq 0,$$

where

$$\frac{1}{n^*}\widetilde{X}'\widetilde{X} = \frac{1}{n^*}\sum_{i=1}^{n^*}\widetilde{X}_i'\widetilde{X}_i,$$

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

Subsampling

FRED HUTCH
CURES START HERE™

- NNLS depends only on the average products between rows of $\widetilde{X}$ and between rows of $\widetilde{X}$ and $\widetilde{Y}$.
- We can approximate these values by subsampling rows of $\widetilde{X}$ and $\widetilde{Y}$ multiple times to get robust NNLS estimates.

Subsampling for NetGSA

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{Y} = \begin{pmatrix} y_1^{(1)} \\ \vdots \\ y_{n_1}^{(1)} \\ y_{n_1+1}^{(2)} \\ \vdots \\ y_n^{(2)} \end{pmatrix}, \quad \mathrm{Var}(\boldsymbol{\gamma}) = \begin{pmatrix} A_1 & & & & & \\ & \ddots & & & & \\ & & A_1 & & & \\ & & & A_2 & & \\ & & & & \ddots & \\ & & & & & A_2 \end{pmatrix}$$

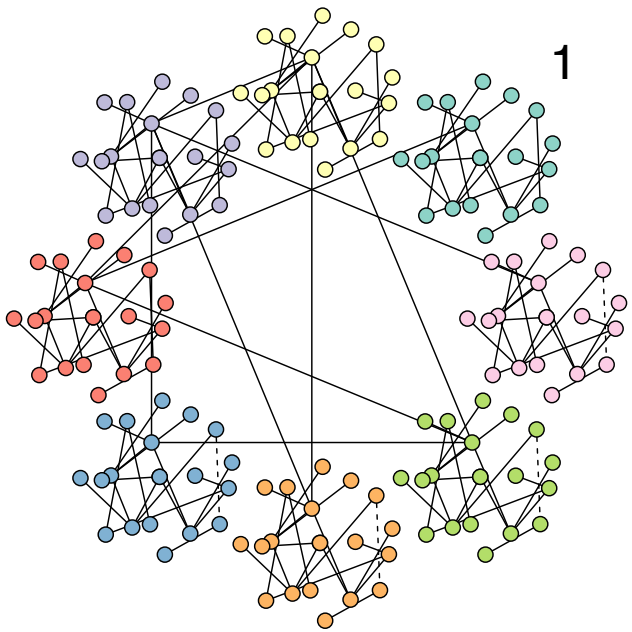
- $\mathrm{Var}(\boldsymbol{\gamma})$ has $n = n_1 + n_2$ diagonal block.
- We can subsample observations and genes simultaneously!

Consistency

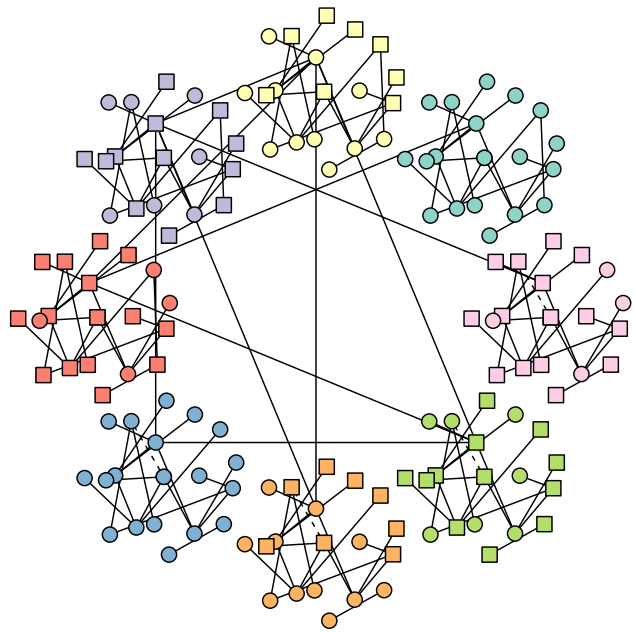
- Classical CLT fails because entries in $\widetilde{Y}$ are dependent
- We evoke the CLT for weakly dependent processes to conclude consistency for the HE/REHE estimator

Simulations: Setup

- 8 subnetworks with varying degrees of enrichment



Null

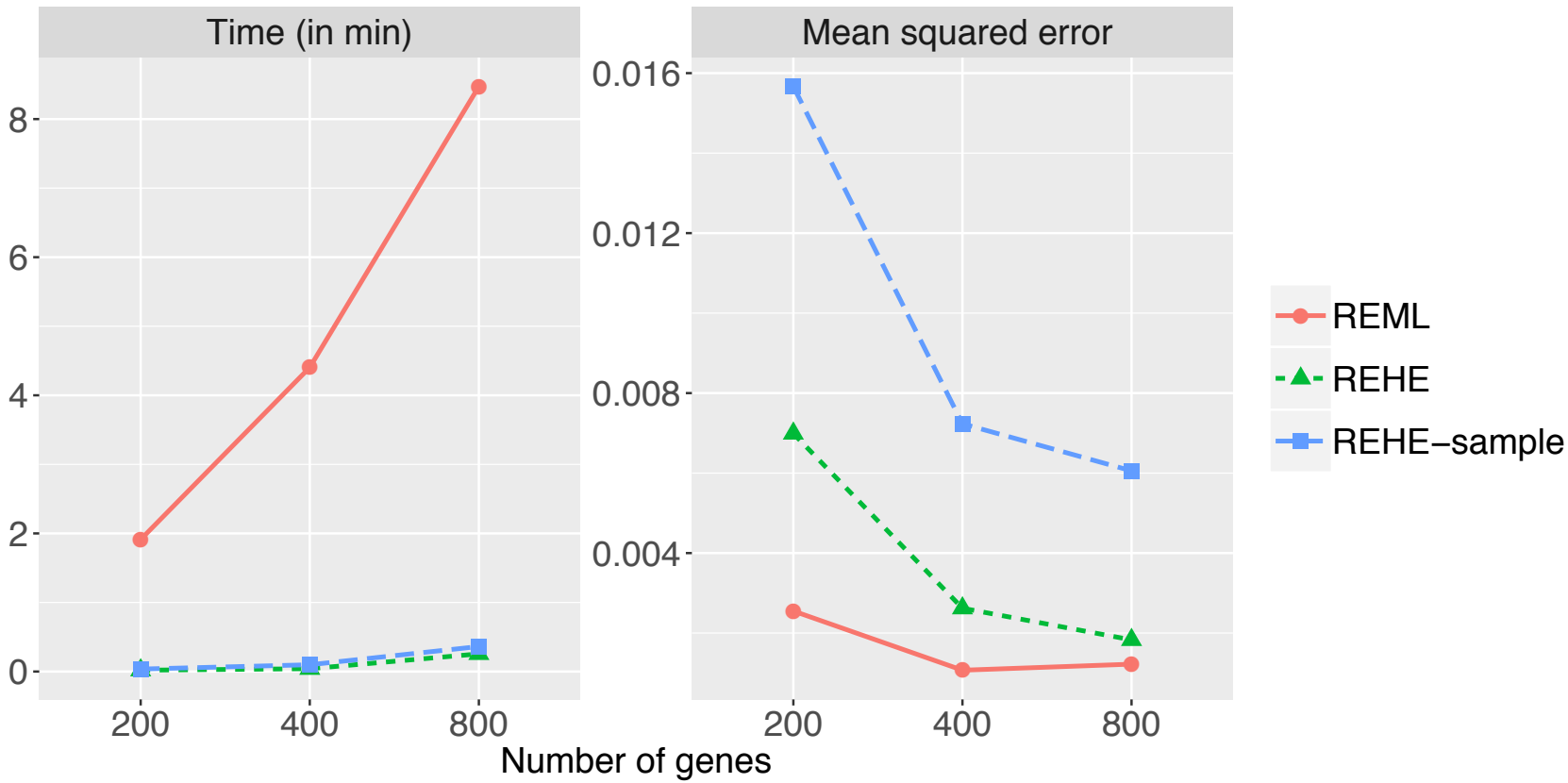


Alt

set 1 set 2 set 3 set 4 set 5 set 6 set 7 set 8

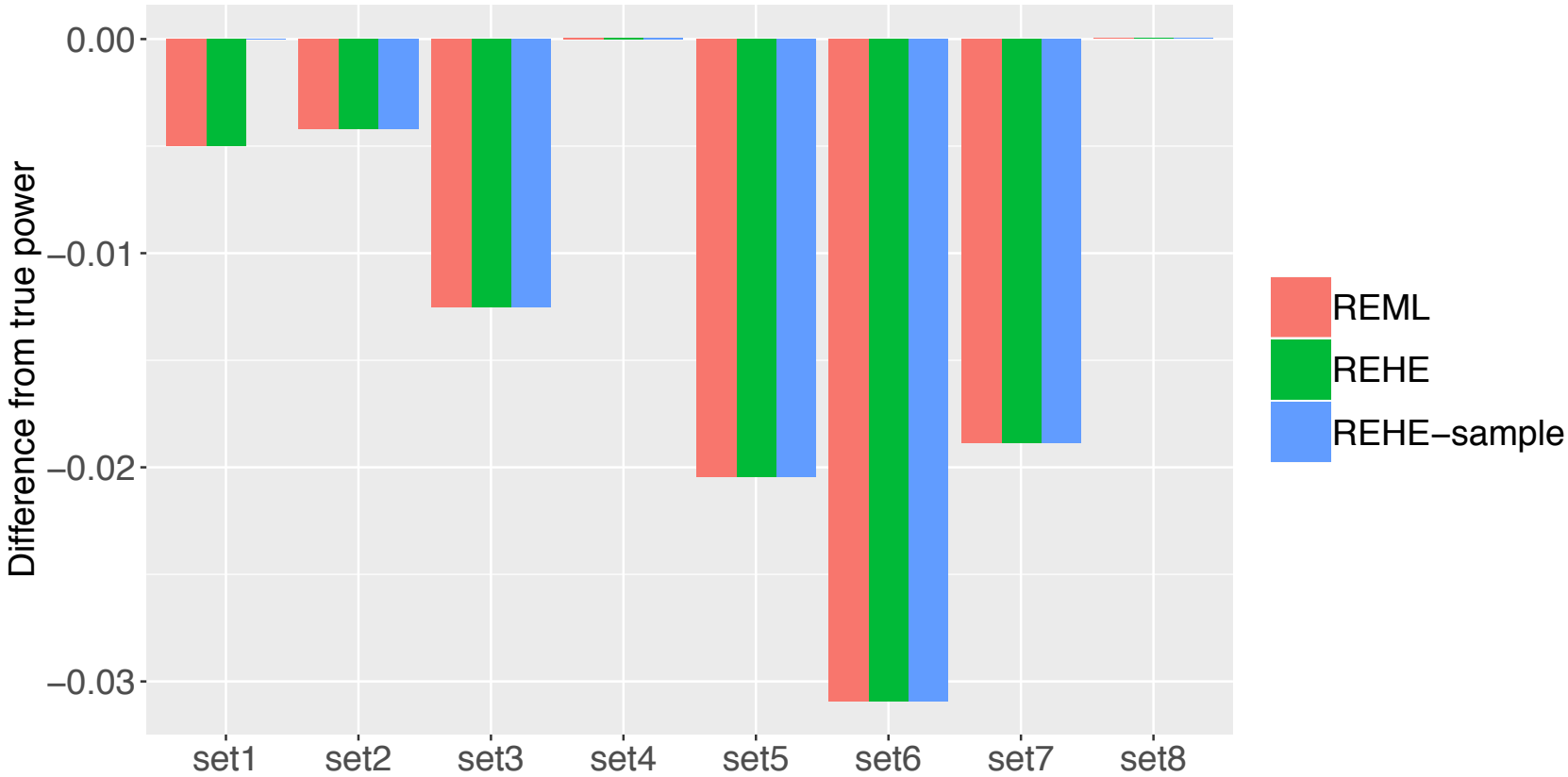
Simulations: VC Estimation

- $n_1 = n_2 = 200, \sigma^2_\gamma = \sigma^2_\epsilon = 1$



Simulations: Power

- $p = 800, n_1 = n_2 = 200, \sigma^2_\gamma = \sigma^2_\epsilon = 1, \mu^{(2)} - \mu^{(1)} = 0.1$



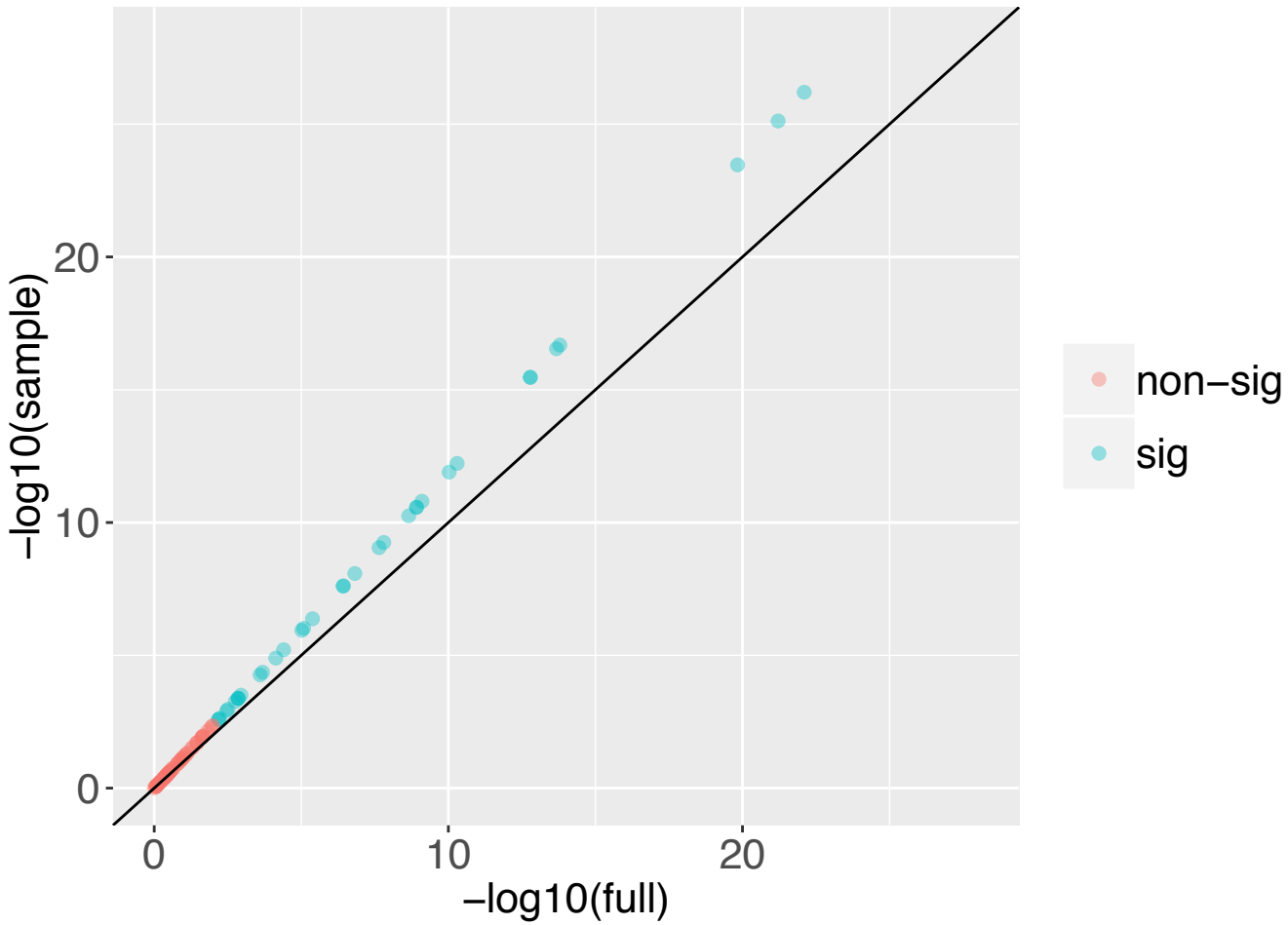
TCGA Prostate Cancer Pathway Analysis

FRED HUTCH
CURES START HERE™

- Gene expression from 160 normal vs 264 tumor samples
- $p = 2800$ genes with Entrez IDs
- Network topology information extracted from BioGrid
- Analysis of 96 KEGG signaling pathways

TCGA Prostate Cancer Pathway Analysis

- Each dot represents one pathway; FDR *p*-value threshold at 0.01



Extensions

The ACE model with $\mathbf{X} \in \mathbb{R}^{n \times m}$

$$y = \mathbf{X}\beta + \gamma + c + e.$$

Let L be the $(n - m) \times n$ matrix with its rows spanning the kernel space of $\mathbf{X}'$. Then

$$\mathbb{E}[L\ yy'\ L'] = L(\sigma_\gamma^2 A + \sigma_c^2 C + \sigma_e^2 I_n)L'.$$

Extensions

The ACE model with $\mathbf{X} \in \mathbb{R}^{n \times m}$

$$y = \mathbf{X}\beta + \gamma + c + e.$$

Let L be the $(n - m) \times n$ matrix with its rows spanning the kernel space of $\mathbf{X}'$. Then

$$\mathbb{E}[L\ yy'\ L'] = L(\sigma_\gamma^2 A + \sigma_c^2 C + \sigma_e^2 I_n)L'.$$

- yy' is a sample outer product derived from the Euclidean distance.

Extensions

The ACE model with $\mathbf{X} \in \mathbb{R}^{n \times m}$

$$y = \mathbf{X}\beta + \gamma + c + e.$$

Let L be the $(n - m) \times n$ matrix with its rows spanning the kernel space of $\mathbf{X}'$. Then

$$\mathbb{E}[L\ yy'\ L'] = L(\sigma^2_\gamma A + \sigma^2_c C + \sigma^2_e I_n)L'.$$

- yy' is a sample outer product derived from the Euclidean distance.
- If we do not observe y but have an outer product matrix M defined from a distance measure suitable for microbiome data, we can detect heritable microbial communities by

$$\mathbb{E}[L\ M\ L'] = L(\sigma^2_\gamma A + \sigma^2_c C + \sigma^2_e I_n)L'.$$

Collaborators

- Kun Yue (UW)
- Ali Shojaie (UW)