Network-based Enrichment Analysis

Jing Ma

Public Health Sciences Division
Fred Hutch Cancer Research Center
jingma@fredhutch.org

January 24, 2018

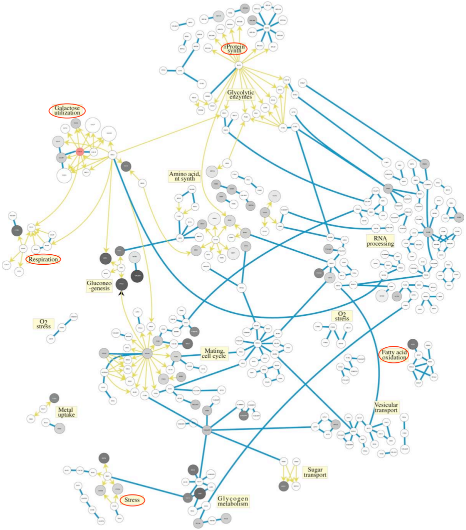
Collaborators

Ali Shojaie

George Michailidis

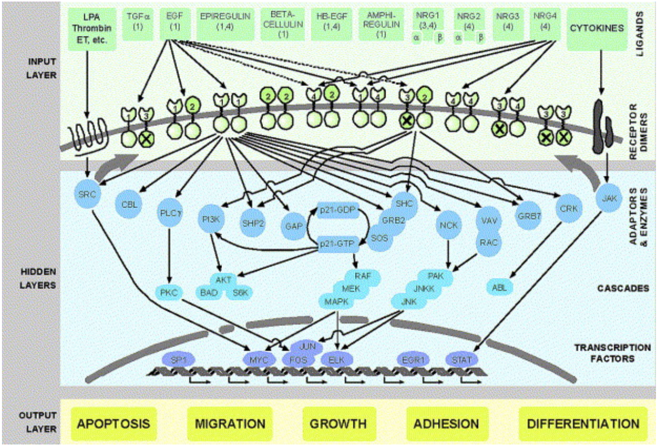
Yeast GAL Pathway[1]

- Physical-interaction network
- Nodes: genes
- Edges: DNA binding, protein-protein interaction
- Highly interconnected groups of genes have common biological function



[1] Ideker et al. Science. 2001

ERBB Signaling Network[2]



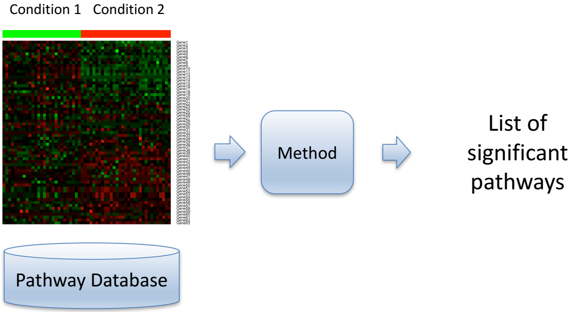
[2] Yarden & Sliwkowski Nat. Rev. Mol. Cell Biol. 2001

Pathway Enrichment Analysis

Scientific Question: whether a genetic/metabolic pathway is involved in responding to changes in environmental conditions or in specific cell functions.

Pathway Enrichment Analysis

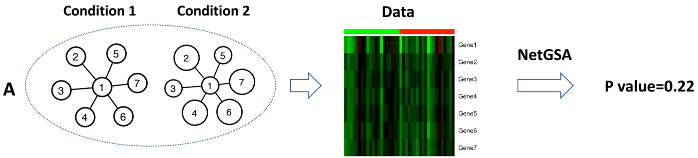
Scientific Question: whether a genetic/metabolic pathway is involved in responding to changes in environmental conditions or in specific cell functions.



Pathway Enrichment Analysis

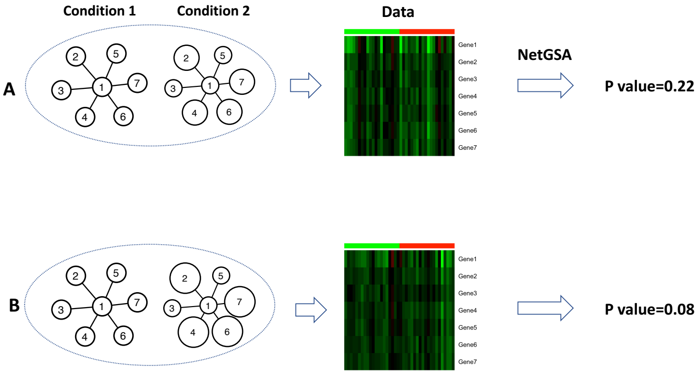
- Reduce the complexity.
- More explanatory power.

Toy Example

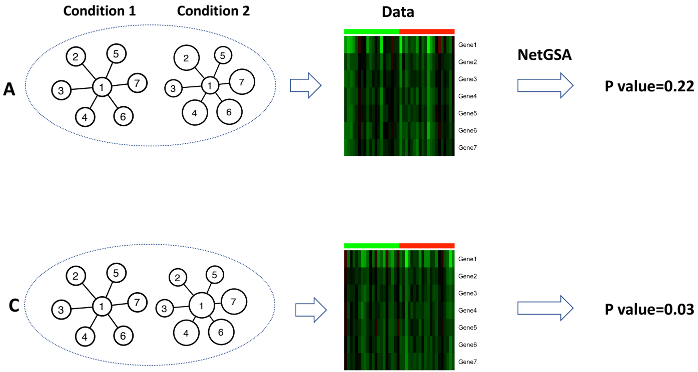


Condition 1
Condition 2
A
Data
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
Gene7
NetGSA
P value=0.22

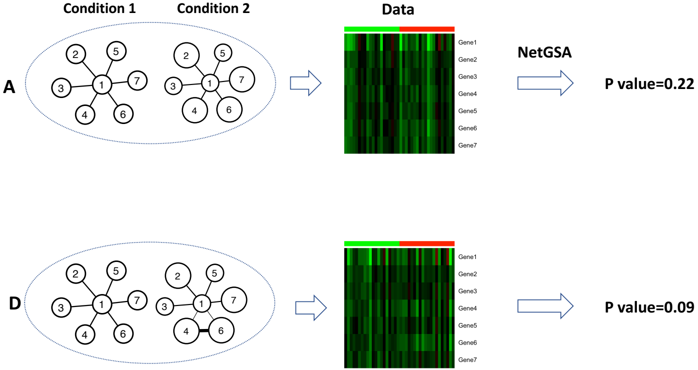
Toy Example


Condition 1
Condition 2
Data
A
B
1
2
3
4
5
6
7
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
Gene7
NetGSA
P value=0.22
P value=0.08

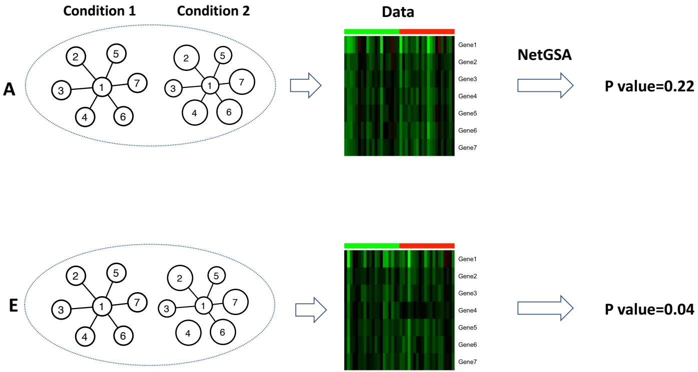
Toy Example


Condition 1
Condition 2
Data
A
NetGSA
P value=0.22
C
P value=0.03
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
Gene7

Toy Example


Condition 1
Condition 2
Data
A
D
NetGSA
P value=0.22
P value=0.09
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
Gene7

Toy Example



Condition 1
Condition 2
Data
A
2
5
3
1
7
4
6
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
Gene7
NetGSA
P value=0.22
E
P value=0.04

What Drives Pathway Significance?

- Mean expression levels of all genes.
- Gene position: hub gene?
- Gene-gene interactions.

What Drives Pathway Significance?

- Mean expression levels of all genes.
- Gene position: hub gene?
- Gene-gene interactions.

NetGSA captures all three factors!

NetGSA: Network-based Gene Set Analysis

- Let Y be an arbitrary sample in the expression data.

NetGSA: Network-based Gene Set Analysis

- Let Y be an arbitrary sample in the expression data.
- $Y = X + \varepsilon$, with X the signal and $\varepsilon \sim N_p(0, \sigma_\varepsilon^2 I_p)$ the noise.

NetGSA: Network-based Gene Set Analysis

- Let Y be an arbitrary sample in the expression data.
- $Y = X + \varepsilon$, with X the signal and $\varepsilon \sim N_p(0, \sigma_\varepsilon^2 I_p)$ the noise.
- $X \sim MVN$ with partial correlation A.

NetGSA: Network-based Gene Set Analysis

- Let Y be an arbitrary sample in the expression data.
- $Y = X + \varepsilon$, with X the signal and $\varepsilon \sim N_p(0, \sigma_\varepsilon^2 I_p)$ the noise.
- $X \sim MVN$ with partial correlation A.
- e.g. $X = (X_1, \ldots, X_p)$ the log concentration of p genes. The network A captures gene-gene interactions.

NetGSA: Network-based Gene Set Analysis

- Let Y be an arbitrary sample in the expression data.
- $Y = X + \varepsilon$, with X the signal and $\varepsilon \sim N_p(0, \sigma_\varepsilon^2 I_p)$ the noise.
- $X \sim MVN$ with partial correlation A.
- e.g. $X = (X_1, \ldots, X_p)$ the log concentration of p genes. The network A captures gene-gene interactions.
- Assume the network A is known.

Linear Recursive Equations

$$\begin{aligned} X_1 + \lambda_{12} X_2 + \lambda_{13} X_3 &= \gamma_1 \\ X_2 + \lambda_{23} X_3 &= \gamma_2 \\ X_3 &= \gamma_3 \end{aligned}$$

Linear Recursive Equations

$$\begin{aligned} X_1 + \lambda_{12} X_2 + \lambda_{13} X_3 &= \gamma_1 \\ X_2 + \lambda_{23} X_3 &= \gamma_2 \\ X_3 &= \gamma_3 \end{aligned}$$

- γ_i: baseline expression levels.

Linear Recursive Equations

$$\begin{aligned} X_1 + \lambda_{12} X_2 + \lambda_{13} X_3 &= \gamma_1 \\ X_2 + \lambda_{23} X_3 &= \gamma_2 \\ X_3 &= \gamma_3 \end{aligned}$$

- γ_i: baseline expression levels.
- $X = \Lambda\gamma$ where

$$\Lambda^{-1} = \begin{pmatrix} 1 & \lambda_{12} & \lambda_{13} \\ 0 & 1 & \lambda_{23} \\ 0 & 0 & 1 \end{pmatrix}, \quad \Lambda\Lambda' = (I_p - A)^{-1}.$$

Linear Recursive Equations

$$\begin{aligned} X_1 + \lambda_{12} X_2 + \lambda_{13} X_3 &= \gamma_1 \\ X_2 + \lambda_{23} X_3 &= \gamma_2 \\ X_3 &= \gamma_3 \end{aligned}$$

- γ_i: baseline expression levels.
- $X = \Lambda\gamma$ where

$$\Lambda^{-1} = \begin{pmatrix} 1 & \lambda_{12} & \lambda_{13} \\ 0 & 1 & \lambda_{23} \\ 0 & 0 & 1 \end{pmatrix}, \quad \Lambda\Lambda' = (I_p - A)^{-1}.$$

- $Y = \Lambda\gamma + \varepsilon$.

Linear Mixed Effects Model

- Matrix representation: letting $\mathbf{Y} \in \mathbb{R}^{np \times 1}$,

$$\mathbf{Y} = (\Psi\beta + \Pi\mathcal{G}) + \mathcal{E}$$

where β and $\mathcal{G}$ are fixed and random effect parameters and

$$\mathcal{G} \sim MVN(0, \sigma_\gamma^2 I_{np}), \quad \mathcal{E} \sim MVN(0, \sigma_\varepsilon^2 I_{np}).$$

Linear Mixed Effects Model

- Matrix representation: letting $\mathbf{Y} \in \mathbb{R}^{np \times 1}$,

$$\mathbf{Y} = (\Psi\beta + \Pi\mathcal{G}) + \mathcal{E}$$

  where β and $\mathcal{G}$ are fixed and random effect parameters and

$$\mathcal{G} \sim MVN(0, \sigma_\gamma^2 I_{np}), \quad \mathcal{E} \sim MVN(0, \sigma_\varepsilon^2 I_{np}).$$

- Design matrices Ψ and Π are defined as functions of Λ.

Estimation

- Estimation of fixed effect parameter β is done using generalized least squares.

Estimation

- Estimation of fixed effect parameter β is done using generalized least squares.
- Estimation of variance components $\sigma_\gamma^2, \sigma_\varepsilon^2$ can be done using restricted maximum likelihood (REML).

Inference

- Let ℓ be an arbitrary linear combination (contrast vector). Consider a test of the form:

$$H_0 : \ell\beta = 0 \quad \textit{vs.} \quad H_1 : \ell\beta \neq 0.$$

Inference

- Let ℓ be an arbitrary linear combination (contrast vector). Consider a test of the form:

$$H_0 : \ell\beta = 0 \quad vs. \quad H_1 : \ell\beta \neq 0.$$

- e.g. $\ell = (\mathbf{1}', -\mathbf{1}')$ and $\ell\beta = \mathbf{1}'\beta^C - \mathbf{1}'\beta^T$.

Inference

- Let ℓ be an arbitrary linear combination (contrast vector). Consider a test of the form:

$$H_0 : \ell\beta = 0 \quad vs. \quad H_1 : \ell\beta \neq 0.$$

- e.g. $\ell = (\mathbf{1}', -\mathbf{1}')$ and $\ell\beta = \mathbf{1}'\beta^C - \mathbf{1}'\beta^T$.
- Use a *t*-test to test the significance of each hypothesis separately.

Choice of the Contrast Vector

- The hypotheses and performance of the test depend on the choice of ℓ.

Choice of the Contrast Vector

- The hypotheses and performance of the test depend on the choice of ℓ.
- Intuitively, one can use the indicator of pathway membership; however, this only reflects changes in the expression levels.

Choice of the Contrast Vector

- The hypotheses and performance of the test depend on the choice of ℓ.
- Intuitively, one can use the indicator of pathway membership; however, this only reflects changes in the expression levels.
- The appropriate test, should account for changes in means as well as the network (differential network biology).

Choice of the Contrast Vector

- The hypotheses and performance of the test depend on the choice of ℓ.
- Intuitively, one can use the indicator of pathway membership; however, this only reflects changes in the expression levels.
- The appropriate test, should account for changes in means as well as the network (differential network biology).
- NetGSA combines pathway membership with the influence matrix, which also allows us to test for changes in the network.

Incomplete Network Information

$$
\begin{array}{ccccccc}
1 & 2 & 3 & 4 & 5 & 6 & \\
\end{array}
$$

$$
A = \left(\begin{array}{cccccc}
\cdot & ? & 1 & 0 & ? & 0 \\
? & \cdot & ? & ? & 0 & ? \\
1 & ? & \cdot & ? & 0 & 0 \\
0 & ? & ? & \cdot & ? & 1 \\
? & 0 & 0 & ? & \cdot & ? \\
0 & ? & 0 & 1 & ? & \cdot
\end{array}\right)
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array}
$$

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

Incomplete Network Information

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

- 0: there is no interaction; 1: there is interaction; ?: unknown
- Available network information can be incomplete and/or inaccurate (especially true for metabolomic studies).

Incomplete Network Information

$$A = \begin{array}{cc} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} & \\ \left(\begin{array}{cccccc} \cdot & ? & 1 & 0 & ? & 0 \\ ? & \cdot & ? & ? & 0 & ? \\ 1 & ? & \cdot & ? & 0 & 0 \\ 0 & ? & ? & \cdot & ? & 1 \\ ? & 0 & 0 & ? & \cdot & ? \\ 0 & ? & 0 & 1 & ? & \cdot \end{array}\right) & \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \end{array}$$

- 0: there is no interaction; 1: there is interaction; ?: unknown
- Available network information can be incomplete and/or inaccurate (especially true for metabolomic studies).
- Lacking condition/disease-specific alterations in interactions.

Incomplete Network Information

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

- Given data, we use graphical models to incorporate existing information using a constrained optimization framework.

Incomplete Network Information

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

- 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

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

- 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.

Efficient Computation for Large Networks

- Estimation of variance components is the computational bottleneck.

Efficient Computation for Large Networks

- Estimation of variance components is the computational bottleneck.
- REML uses a Newton-Raphson method for fast convergence.

Efficient Computation for Large Networks

- Estimation of variance components is the computational bottleneck.
- REML uses a Newton-Raphson method for fast convergence.
- New version implements the (restricted) Haseman-Elston regression method for fast computation (coming soon!).

$$\mathrm{Var}(Y) = \sigma_\gamma^2 \Lambda\Lambda^T + \sigma_\varepsilon^2 I_p.$$

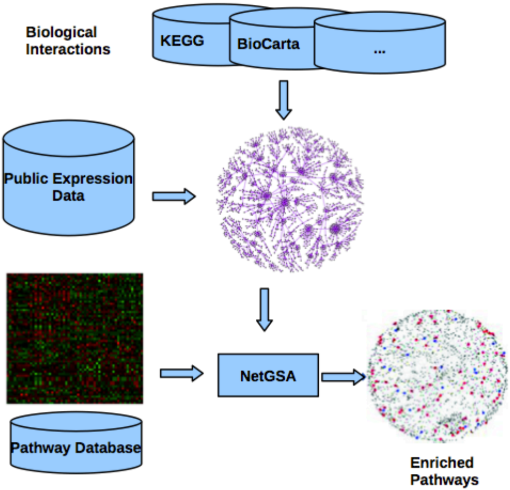
Efficient Computation for Large Networks

- Estimation of variance components is the computational bottleneck.
- REML uses a Newton-Raphson method for fast convergence.
- New version implements the (restricted) Haseman-Elston regression method for fast computation (coming soon!).

$$\mathrm{Var}(Y) = \sigma_\gamma^2 \Lambda\Lambda^T + \sigma_\varepsilon^2 I_p.$$

- Integrative analysis of multiple Omics data can be done using a permutation test.

A Flowchart for NetGSA



Introduction

The NetGSA Model

Extensions of NetGSA

Applications

Pathway Topology-based Methods

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)

Analysis of RNA-seq Data

- RNA-seq data for 2598 genes (TCGA '12).
- 403 ER positive samples; 117 ER negative samples.
- 100 KEGG pathways (`graphite`).

Analysis of RNA-seq Data

- RNA-seq data for 2598 genes (TCGA '12).
- 403 ER positive samples; 117 ER negative samples.
- 100 KEGG pathways (`graphite`).

- Each sample is standardized to have mean 0 and unit variance.

Analysis of RNA-seq Data

- RNA-seq data for 2598 genes (TCGA ’12).
- 403 ER positive samples; 117 ER negative samples.
- 100 KEGG pathways (`graphite`).

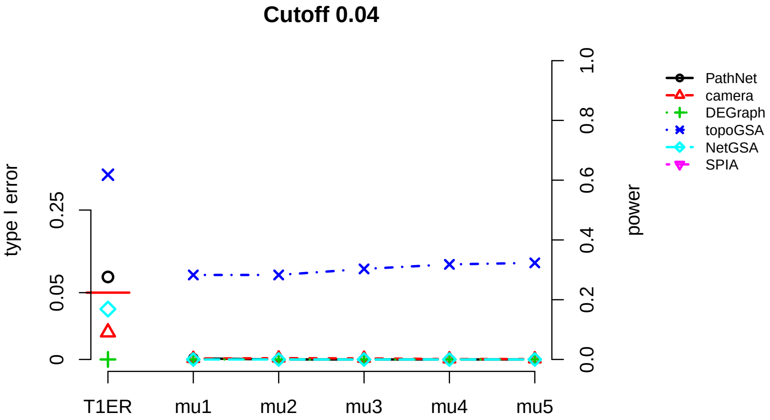
- Each sample is standardized to have mean 0 and unit variance.
- A nonzero mean signal is added to varying proportions of genes in each pathway.

Analysis of RNA-seq Data

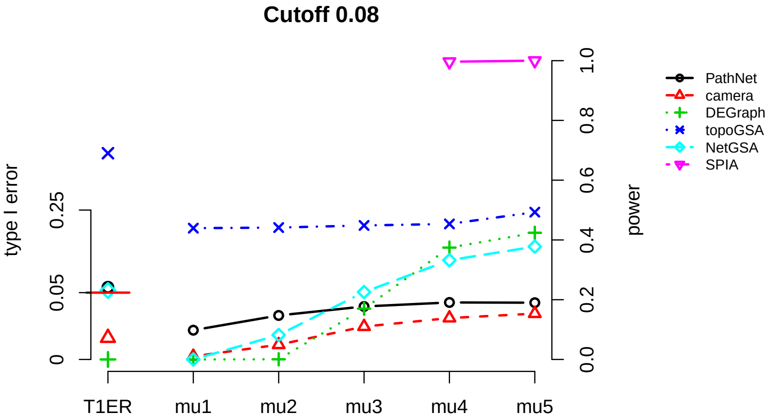
- RNA-seq data for 2598 genes (TCGA '12).
- 403 ER positive samples; 117 ER negative samples.
- 100 KEGG pathways (`graphite`).

- Each sample is standardized to have mean 0 and unit variance.
- A nonzero mean signal is added to varying proportions of genes in each pathway.
- Powers are averaged over multiple pathways that have similar proportion of affected genes.

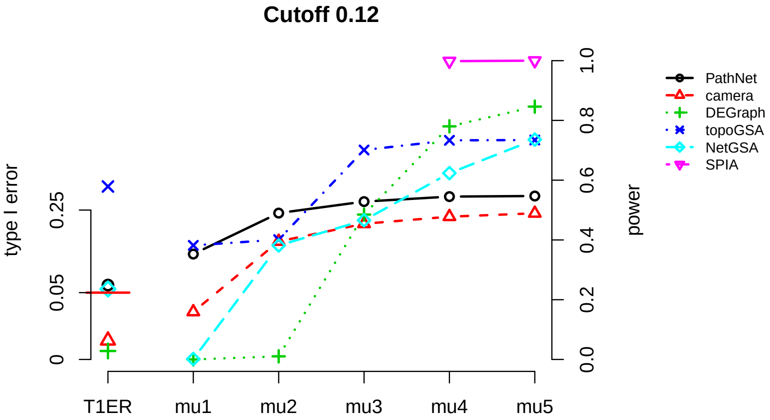
Analysis of RNA-seq Data



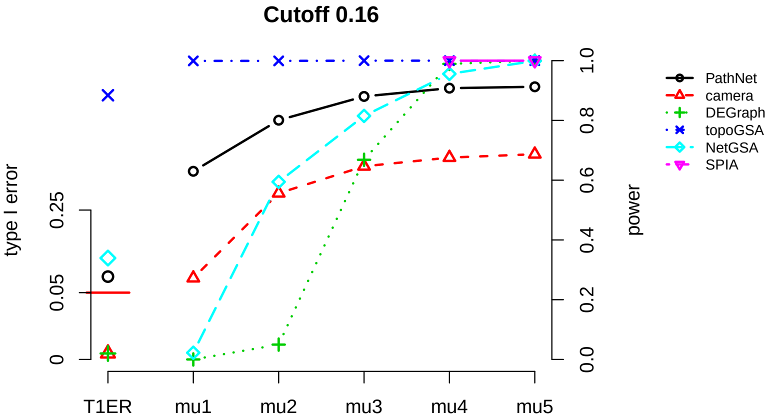
Analysis of RNA-seq Data



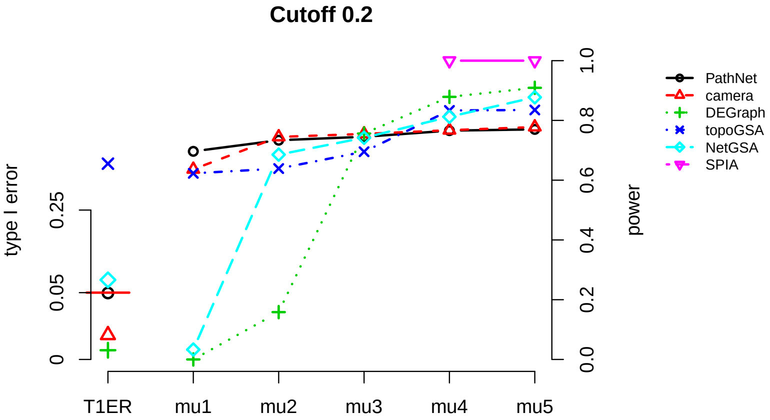
Analysis of RNA-seq Data



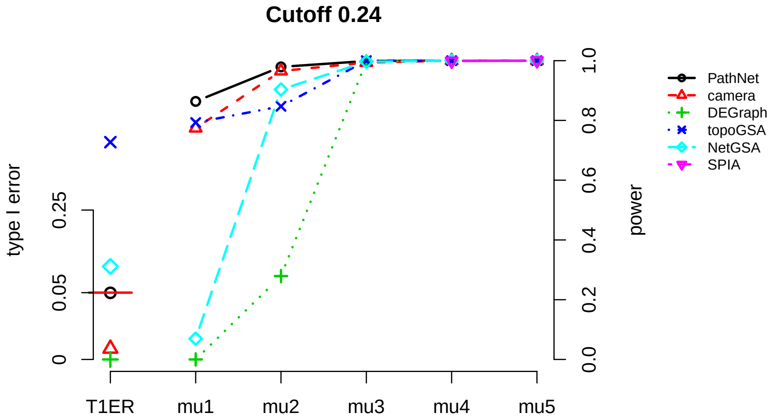
Analysis of RNA-seq Data



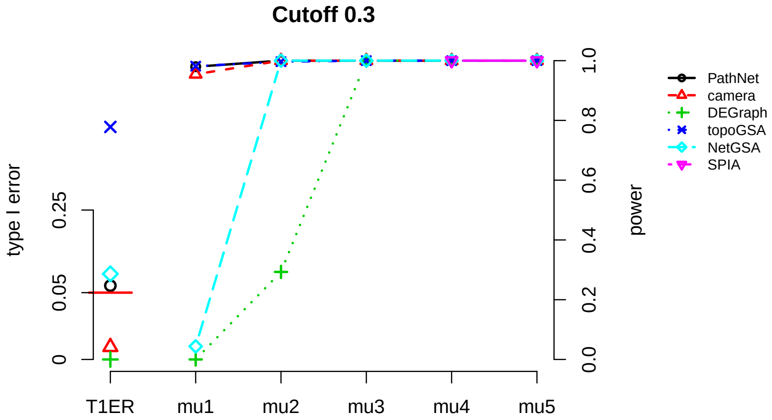
Analysis of RNA-seq Data



Analysis of RNA-seq Data



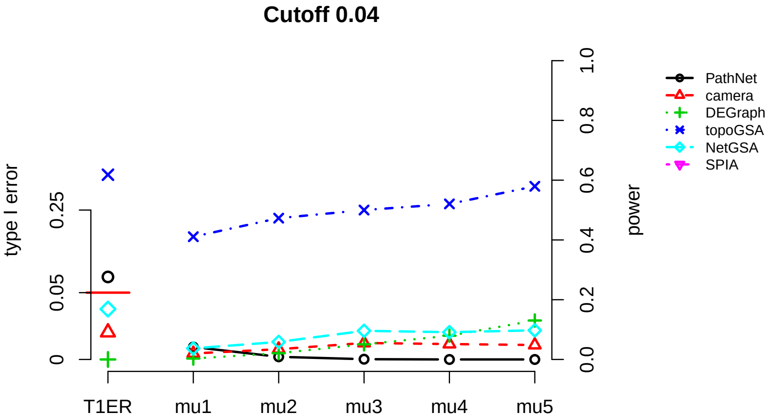
Analysis of RNA-seq Data



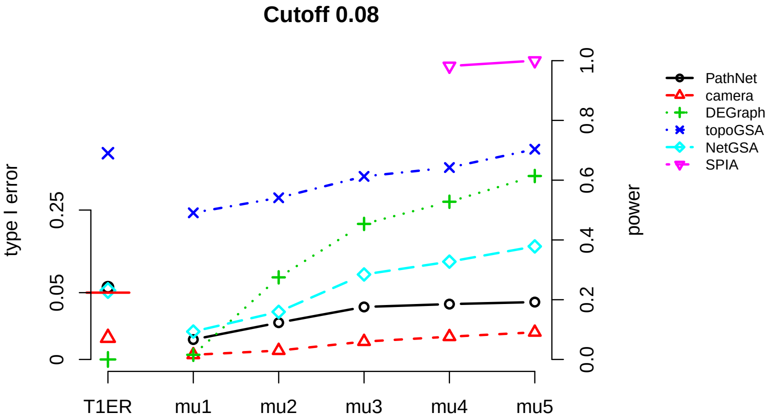
Analysis of RNA-seq Data

What if we permute the samples?

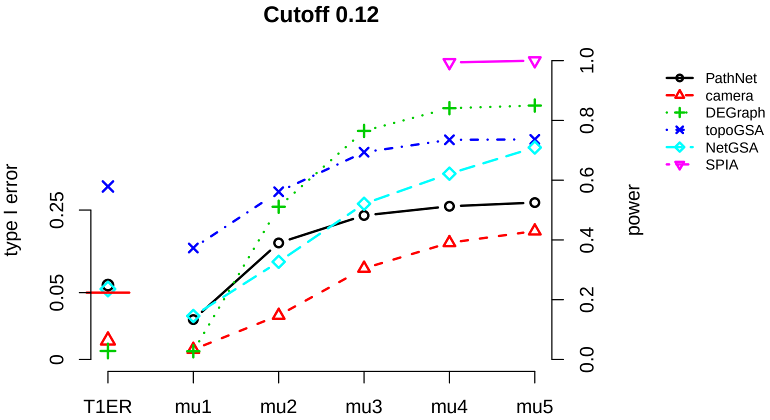
Analysis of RNA-seq Data



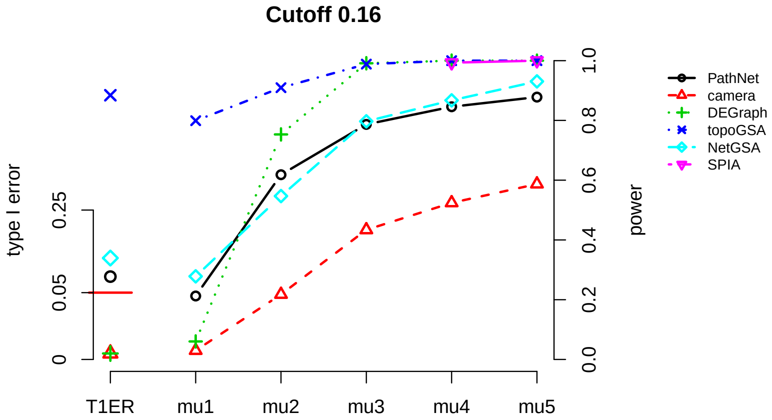
Analysis of RNA-seq Data



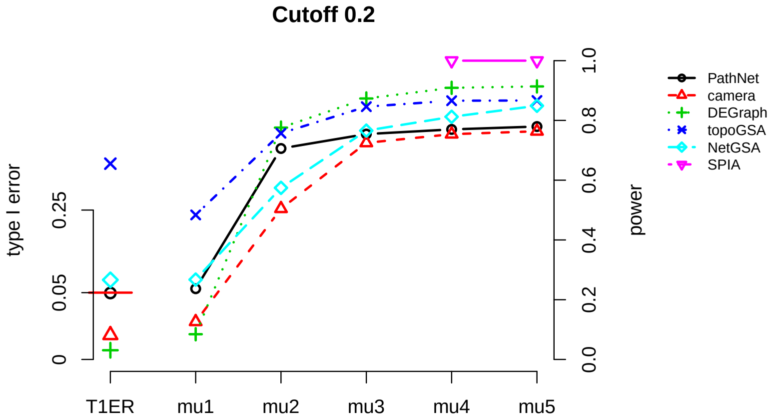
Analysis of RNA-seq Data



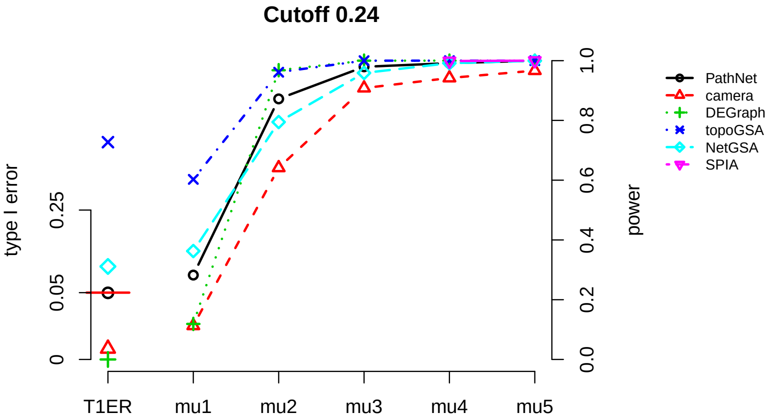
Analysis of RNA-seq Data



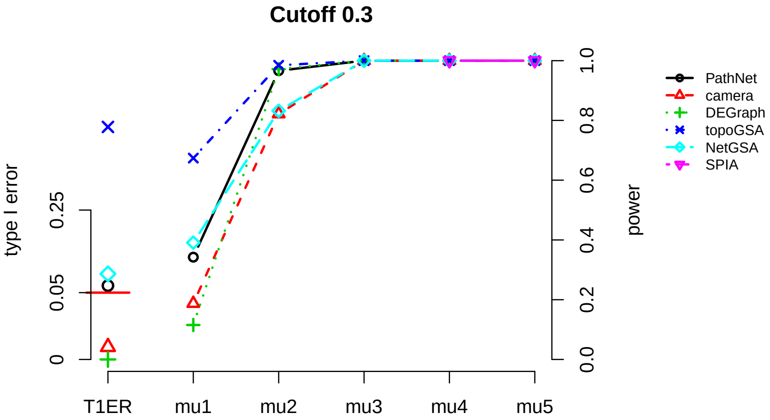
Analysis of RNA-seq Data



Analysis of RNA-seq Data



Analysis of RNA-seq Data



Analysis of Metabolomic Data

- Metabolomic profiles for 100 named metabolites (Fahrmann et al. '15).
- 41 nondiabetic mice; 30 diabetic mice.
- 65 KEGG pathways (`graphite`).
- Limited network information is available.

Analysis of Metabolomic Data

- Metabolomic profiles for 100 named metabolites (Fahrmann et al. '15).
- 41 nondiabetic mice; 30 diabetic mice.
- 65 KEGG pathways (`graphite`).
- Limited network information is available.
- Each sample is standardized to have mean 0 and unit variance.

Analysis of Metabolomic Data

- Metabolomic profiles for 100 named metabolites (Fahrmann et al. '15).
- 41 nondiabetic mice; 30 diabetic mice.
- 65 KEGG pathways (`graphite`).
- Limited network information is available.

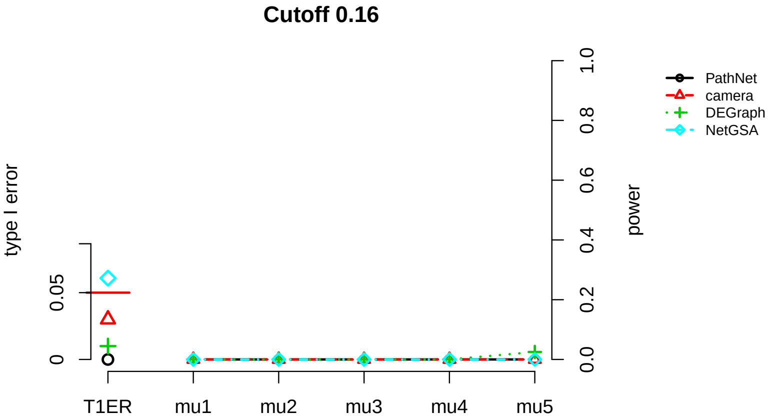
- Each sample is standardized to have mean 0 and unit variance.
- A nonzero mean signal is added to varying proportions of metabolites in each pathway.

Analysis of Metabolomic Data

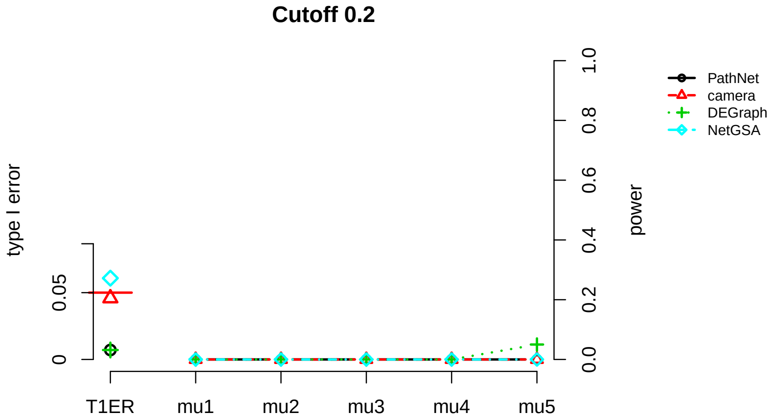
- Metabolomic profiles for 100 named metabolites (Fahrmann et al. '15).
- 41 nondiabetic mice; 30 diabetic mice.
- 65 KEGG pathways (`graphite`).
- Limited network information is available.

- Each sample is standardized to have mean 0 and unit variance.
- A nonzero mean signal is added to varying proportions of metabolites in each pathway.
- Powers are averaged over multiple pathways that have similar proportion of affected metabolites.

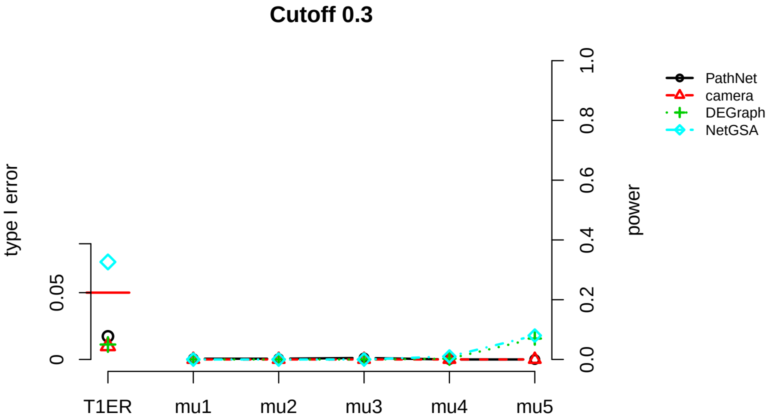
Analysis of Metabolomic Data



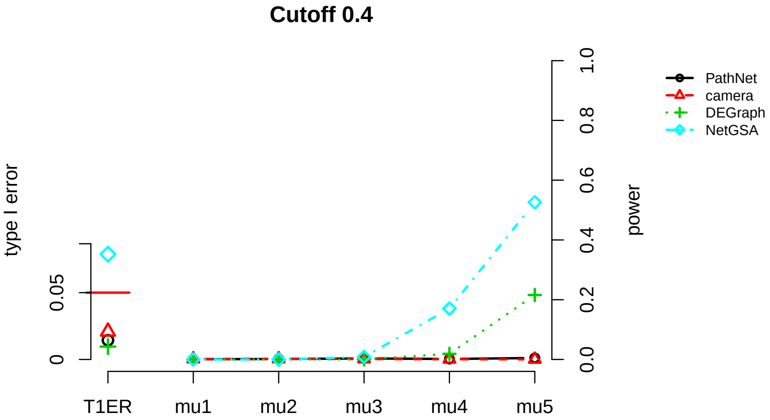
Analysis of Metabolomic Data



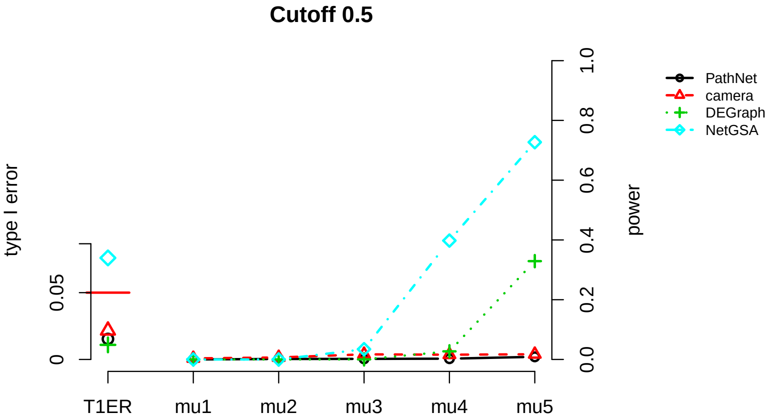
Analysis of Metabolomic Data



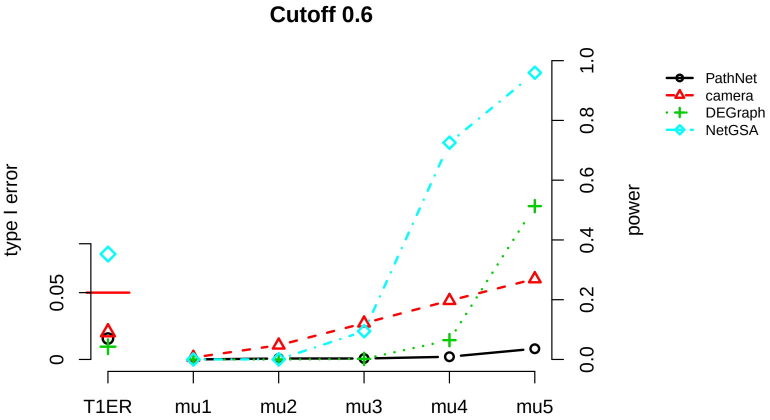
Analysis of Metabolomic Data



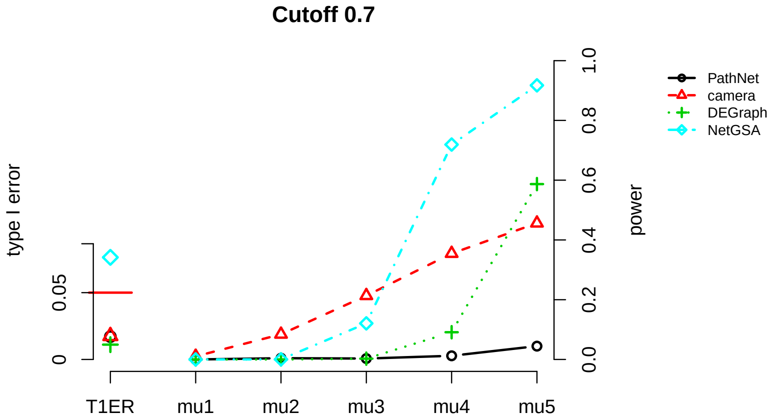
Analysis of Metabolomic Data



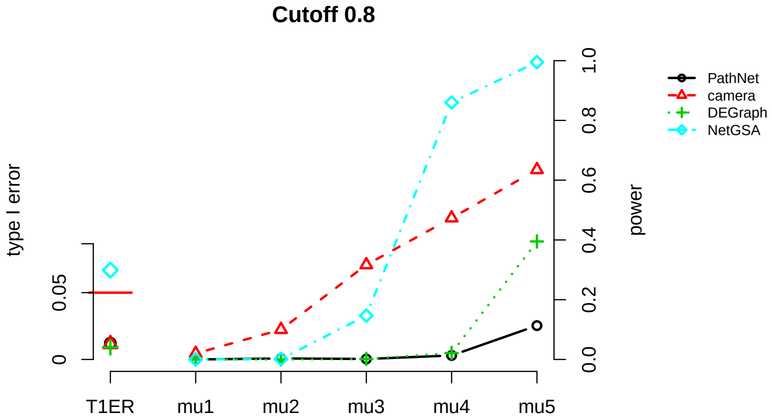
Analysis of Metabolomic Data



Analysis of Metabolomic Data



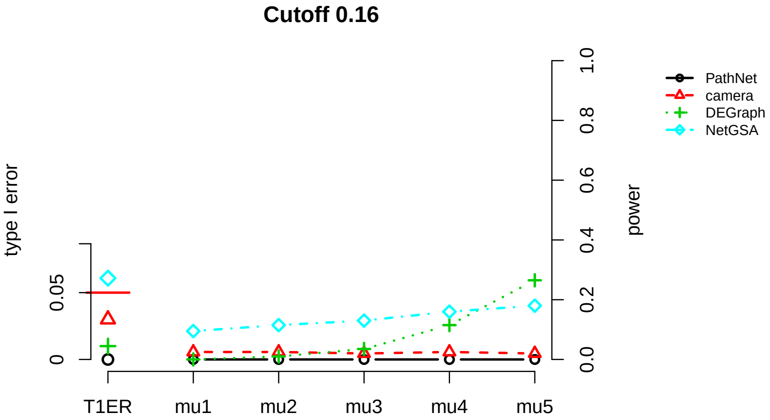
Analysis of Metabolomic Data



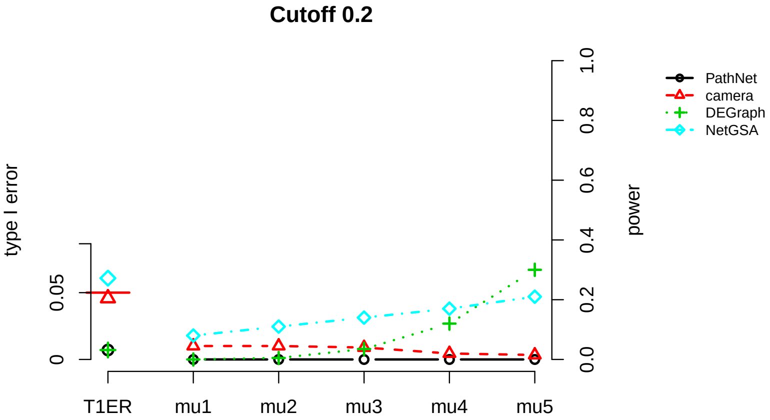
Analysis of Metabolomic Data

What if we permute the samples?

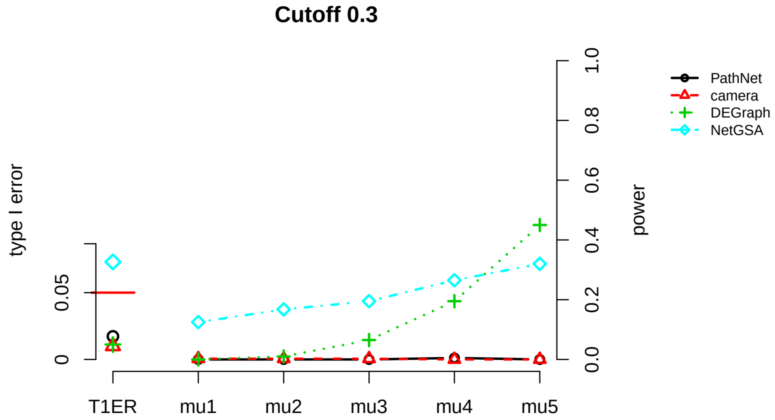
Analysis of Metabolomic Data



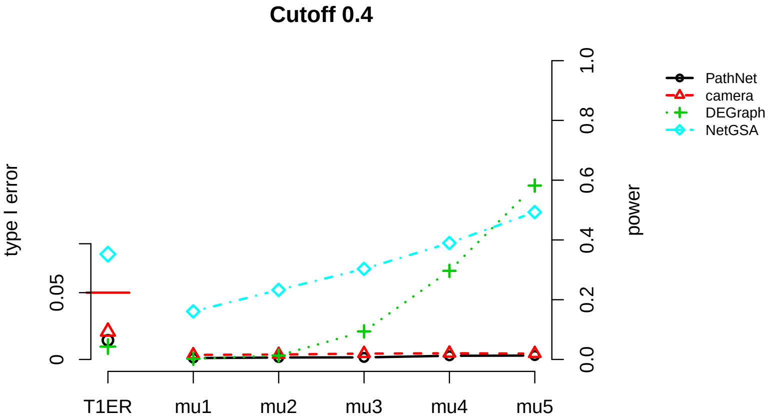
Analysis of Metabolomic Data



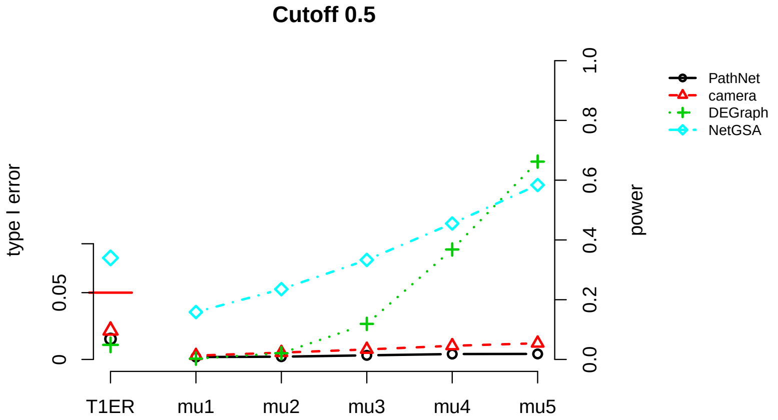
Analysis of Metabolomic Data



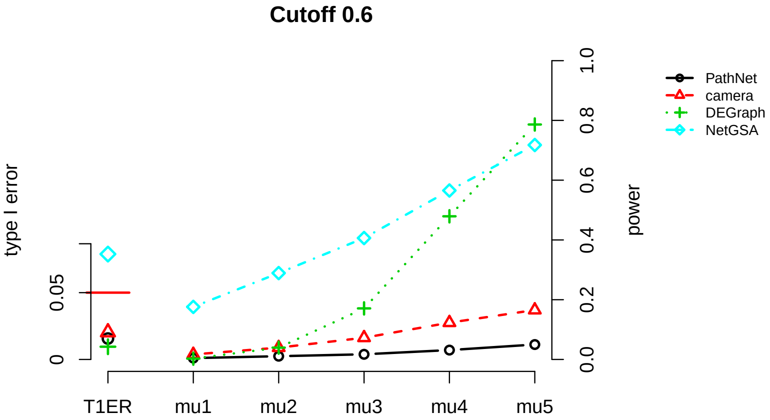
Analysis of Metabolomic Data



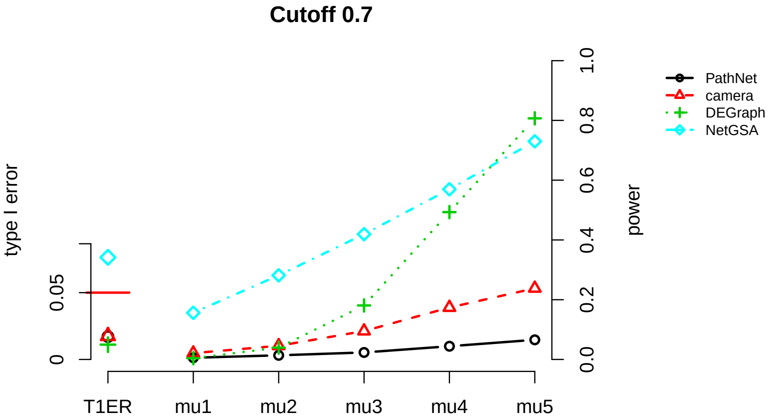
Analysis of Metabolomic Data



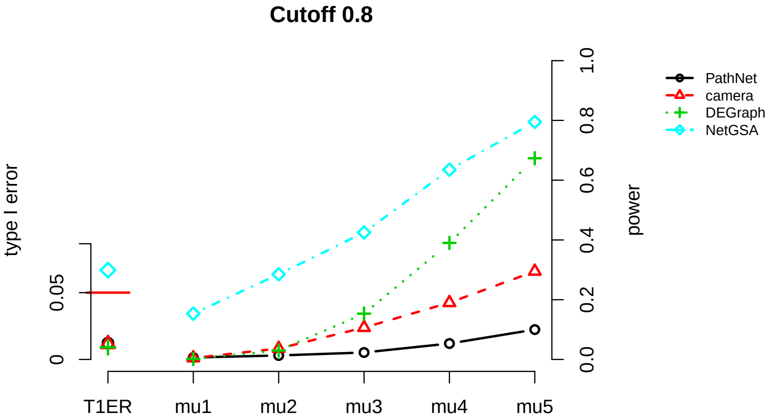
Analysis of Metabolomic Data



Analysis of Metabolomic Data



Analysis of Metabolomic Data



Summary of Results

- Performance of NetGSA is reliable in both studies, especially in the metabolomic data analysis.

Summary of Results

- Performance of NetGSA is reliable in both studies, especially in the metabolomic data analysis.
- DEGraph is also reliable compared to the others.

Summary of Results

- Performance of NetGSA is reliable in both studies, especially in the metabolomic data analysis.
- DEGraph is also reliable compared to the others.
- topologyGSA can have inflated type I errors (and is very slow!).

Summary of Results

- Performance of NetGSA is reliable in both studies, especially in the metabolomic data analysis.
- DEGraph is also reliable compared to the others.
- topologyGSA can have inflated type I errors (and is very slow!).
- SPIA may not work if the mean signal is very small.

Implementation

R-package:

- netgsa
- New version that implements the HE regression method will be released soon.

Source code: `https://github.com/drjingma/netgsa`

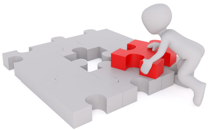
Summary

- NetGSA tests for pathway enrichment by incorporating the pathway topology.

Summary

- NetGSA tests for pathway enrichment by incorporating the pathway topology.
- NetGSA can leverage existing network information and expression data.

Summary



- NetGSA tests for pathway enrichment by incorporating the pathway topology.
- NetGSA can leverage existing network information and expression data.
- Caveat:
  - null hypothesis.
  - sample sizes.