DISNEP simulation

ZHUOHUI Liang

7/15/2021

Simulation Settings

scenario	1.00	2.00	3.00	4.00	5.00
n_signal	50.00	50.00	50.00	50.00	50.00
n_noise	1950.00	1950.00	1950.00	1950.00	1950.00
strong_signal_mu	10.00	10.00	10.00	11.00	12.00
median_signal_mu	9.00	9.00	9.00	10.00	11.00
noise_mu	8.00	8.00	8.00	8.00	8.00
strong_corr	0.10	0.10	0.10	0.10	0.10
median_corr	0.02	0.02	0.02	0.02	0.02
noise_corr	0.01	0.01	0.01	0.01	0.01
sigma	3.00	2.00	1.00	1.00	1.00

For each simulation, there will be two data/matrix simulated, namely gene-gene-interaction network and gene-expression data. Each simulation we will compare t-test, GeneWanderer(random walk with restart on gene-gene-interaction network) and Disnep(random walk with restart on disease-enhanced-gene-gene-interaction network[Ruan,2019])

Gene-gene-interaction network

In each simulation, network is sample from existing database, we have homo sapiens(human) gene-gene(protein-protein) interaction network(PPI) downloaded from V.11 STRING. Since Genewanderer only consider gene-gene-interaction score $>$ 0.4, we use Cytoscape to pull STRING node information from STRING containing only score $>$ 0.4 and map Gene name to PPI.

And we set the gene from DISGENET V7 renal carcinoma disease(C1378703) as disease/signal gene list and the rest of the gene not in DISGENET list as noise gene.

We sample 50 gene from disease gene list and 1950 as noise gene. We keep the edge of the gene from STRING PPI network and set all missing value as 0. And we order the first 50 gene as signal gene just for convenient.

Gene Expression network

In each simulation, we generate 300 case and 50 control sample.

Case

The case set to have 25 strong signal gene, which each gene's mean is generated from N(μ_1,1), and use this mean to generate gene expression level. case has another 25 median signal gene, which each gene's mean is generated from N(μ_2,1). the rest 1950 gene are set as noise gene, which's mean is generate from N(8,1).

The correlation between strong signal is sampled from U(0,0.1) and correlation between median signal is sampled from U(0,0.02) and correlation between noise is sampled from U(0,0.01).

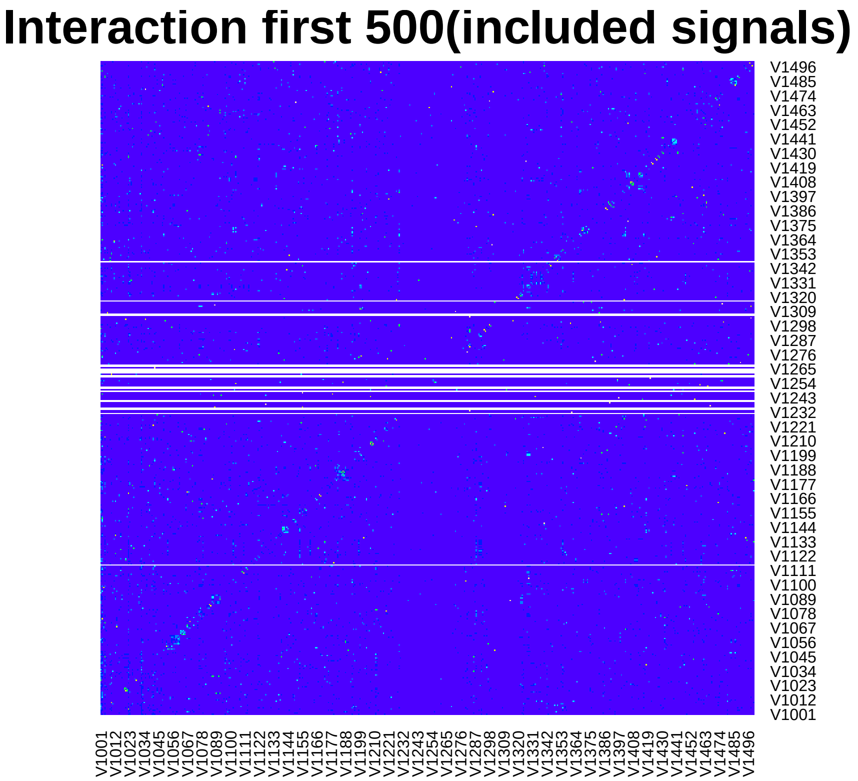
$$Cov = sigma^2 * \begin{pmatrix} 1 & \rho_{12} & 0 & 0 \\ \rho 21 & 1 & \rho 23 & 0 \\ 0 & \rho 32 & 1 & \rho 34 \\ 0 & 0 & \rho 43 & 1 \end{pmatrix} \tag{1}$$

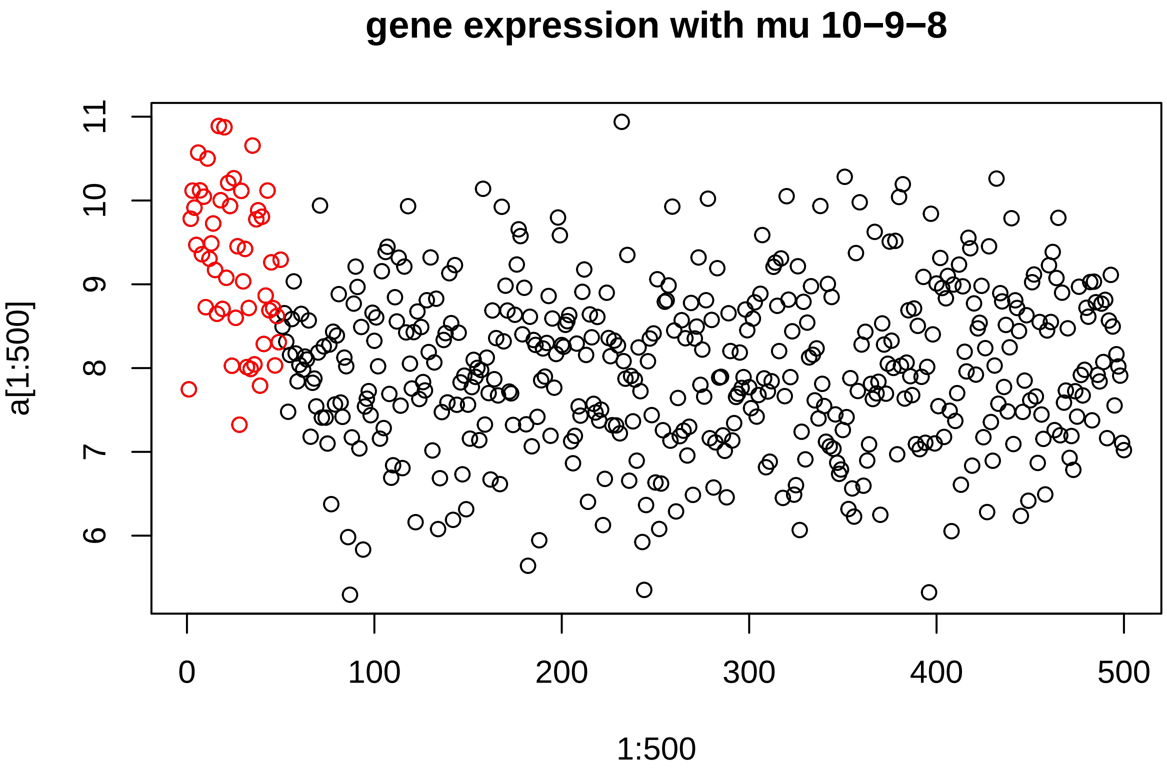
where $\rho_{12} = \rho_{21}^t$ is the triagle block sample from U(0,0.1) and vise versa for $\rho_{23}, \rho 34$

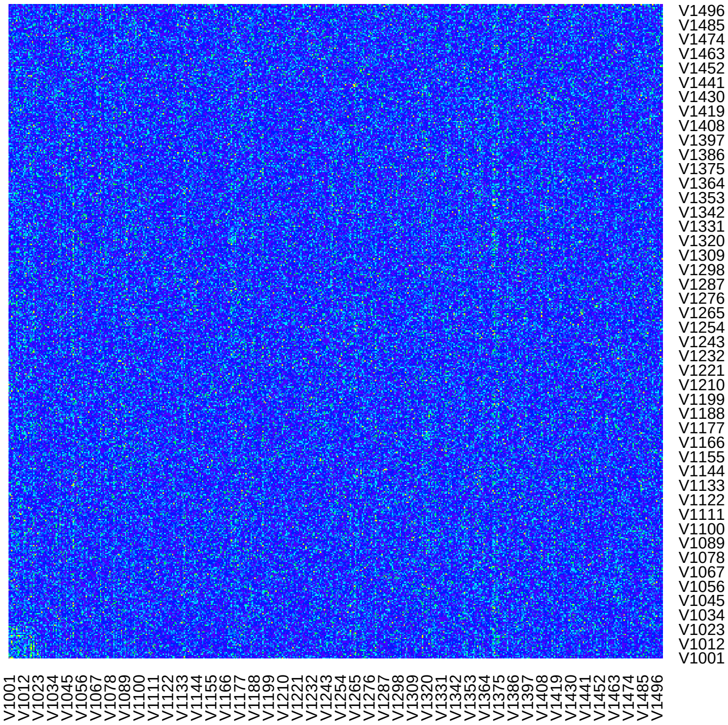
control

All genes in control is set to be noise and its mean generate from N(8,1) and correlation sample from U(0,0.01)

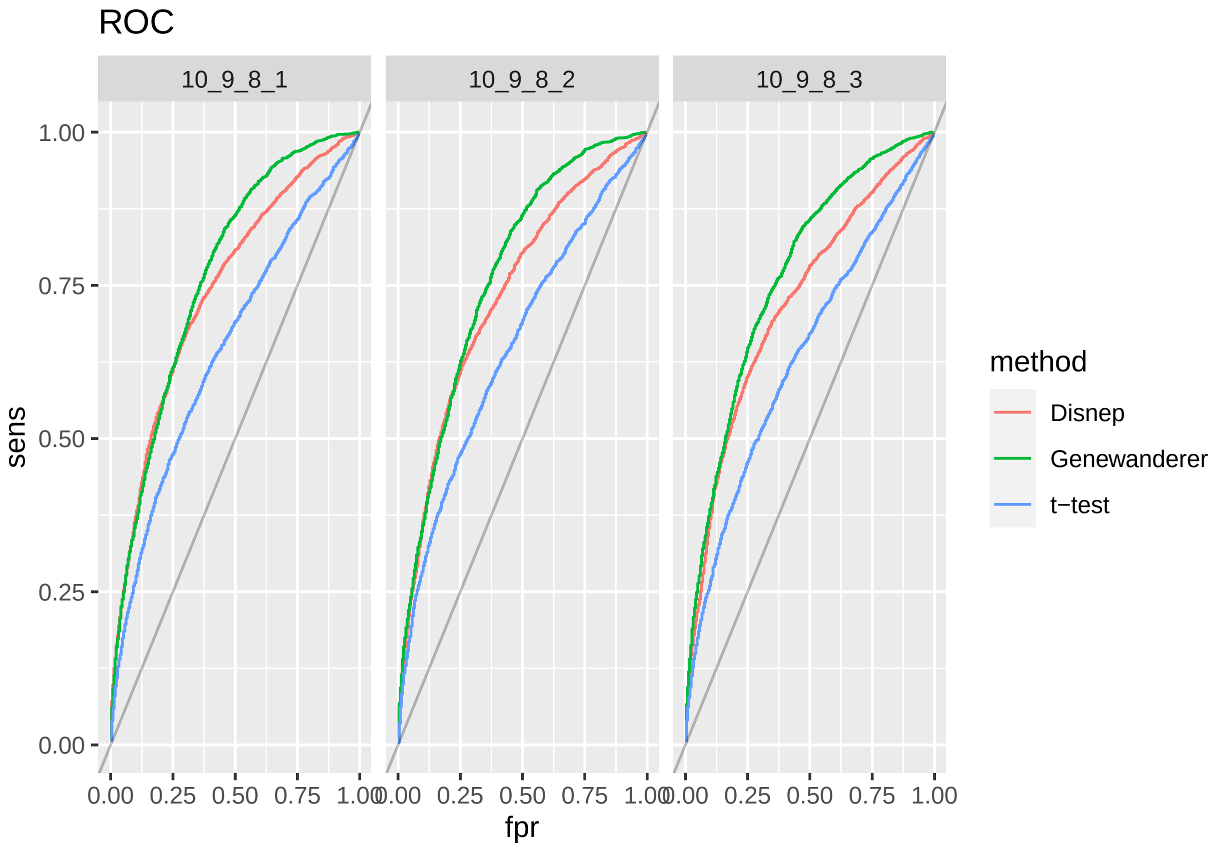
all gene expression is generated from multivariate normal, with generated means and covariate as mention above.

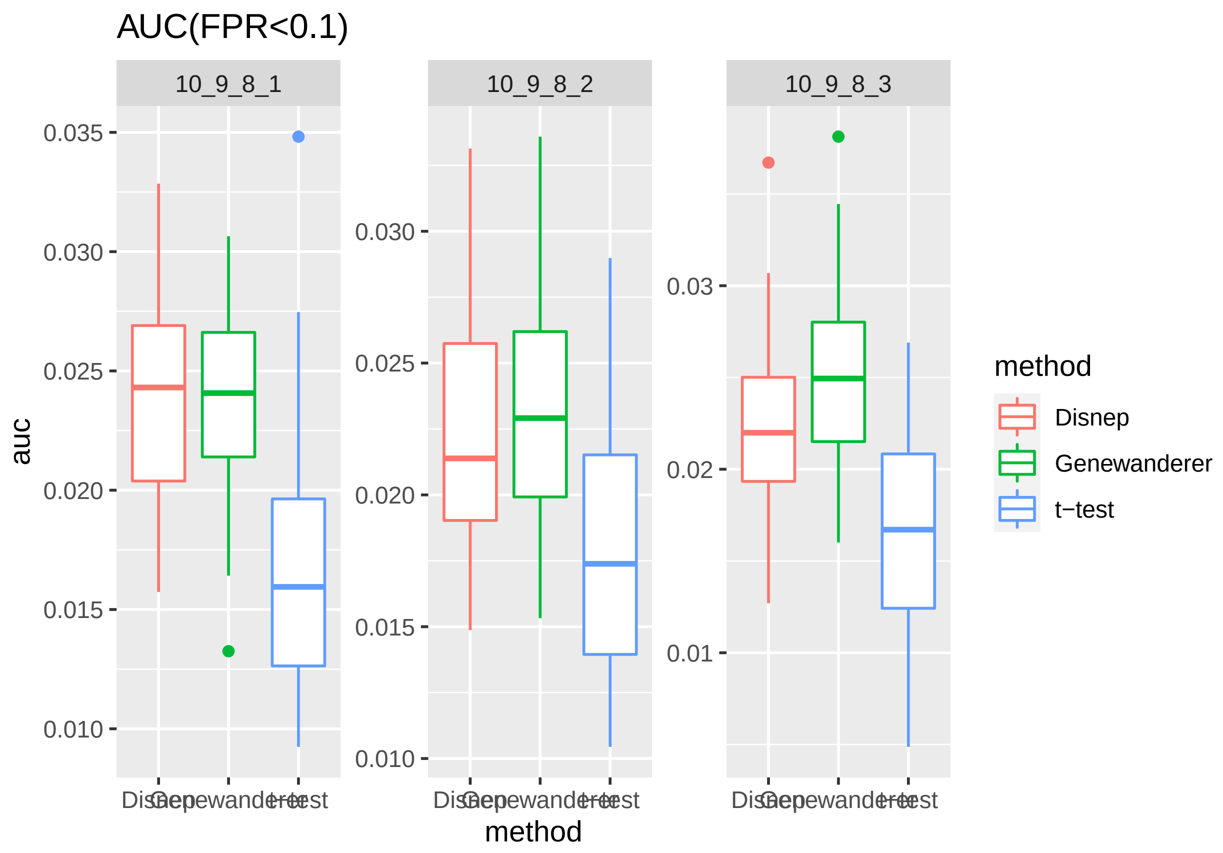



gene expression with mu 10−9−8
a[1:500]
6
7
8
9
10
11
0
100
200
300
400
500
1:500



Correlation first 500(included signals)


ROC
10_9_8_1
10_9_8_2
10_9_8_3
sens
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
fpr
method
Disnep
Genewanderer
t−test


AUC(FPR<0.1)
10_9_8_1
10_9_8_2
10_9_8_3
auc
0.035
0.030
0.025
0.020
0.015
0.010
0.030
0.025
0.020
0.015
0.010
0.03
0.02
0.01
Disnep
Genewanderer
t−test
method
method
Disnep
Genewanderer
t−test