# Effects of Different Normalization Methods on Gene Set Enrichment Analysis (GSEA)

**Ming Zhao**

**Tutor: Dr. Mengjin Zhu**

12th, Nov, 2011

**Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education HZAU**

# Outline

- **Introduction**
  - Normalizations under R/Bioconductor environment
  - Comparison of differentially expressed genes and gene set tests
  - An introduction to Gene Set Enrichment Analysis(GSEA)
  - Normalizations and GSEA
- **Results**
- Detection rate of different normalization methods
- Correlations among different studies base on DEGs
- **Discussion**

# Introduction

## Normalizations under R/Bioconductor environment

- Most Extensive and Flexible Environment
  - Oligonucleotide & cDNA Arrays
  - Multiple methods for
    - Background correction
    - Probe-specific correction
    - **Normalization over multiple chips**
- Open Source
- Best Platform for Development
- **FREE!**

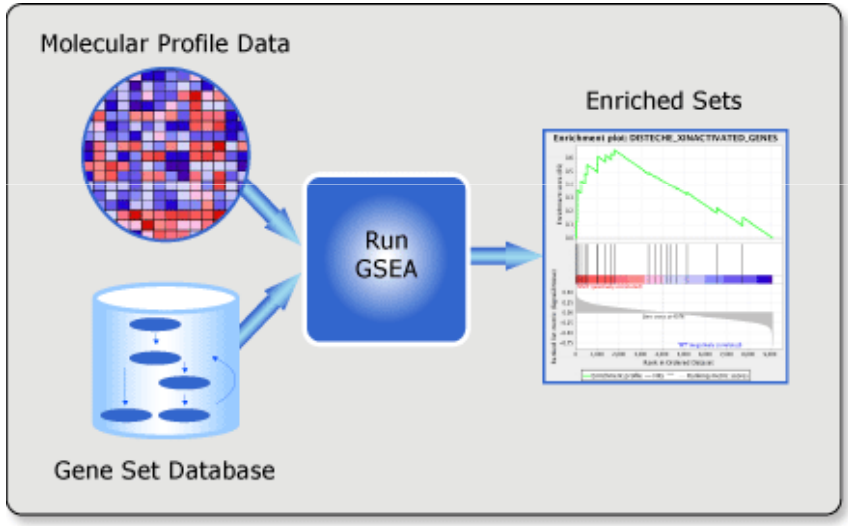
# Normalize Method Options

- Background
  - Basics of Affymetrics Chip Design
    - Need to correct for Background Noise
    - Need for Normalization
- > normalize methods
  - constant
  - contrasts
  - invariantset
  - loess
  - qspline
  - quantiles ?
  - quantiles.robust

# Comparison of Differentially Expressed Genes and Gene-set Tests

- In order to allow direct array-to-array comparisons, normalization is a prerequisite
- Comparison of differently expressed genes(DEGs) by various normalization methods
- On Gene-sets?

# An introduction to Gene Set Enrichment Analysis(GSEA)




http://www.broadinstitute.org/gsea/index.jsp

- Two platforms

  GSEA-P and GSEA-P-R

- Prepare your data files:

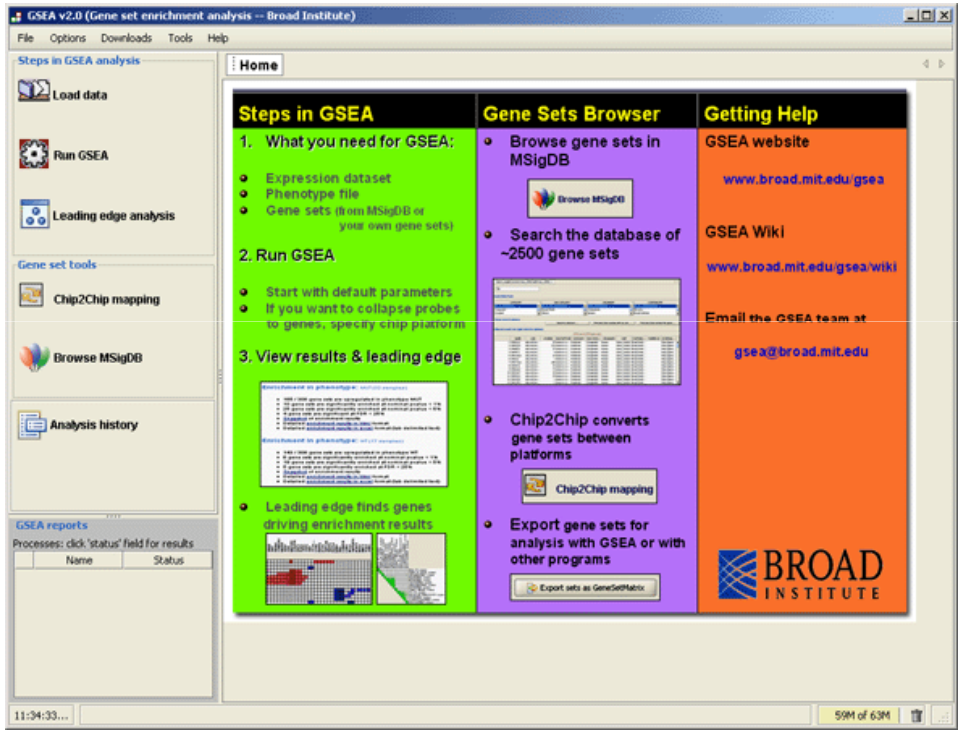
  Expression dataset file (res, gct, pcl, or txt)

  Phenotype labels file (cls)

  Gene sets file (gmx or gmt)

  Chip (array) annotation file (chip)

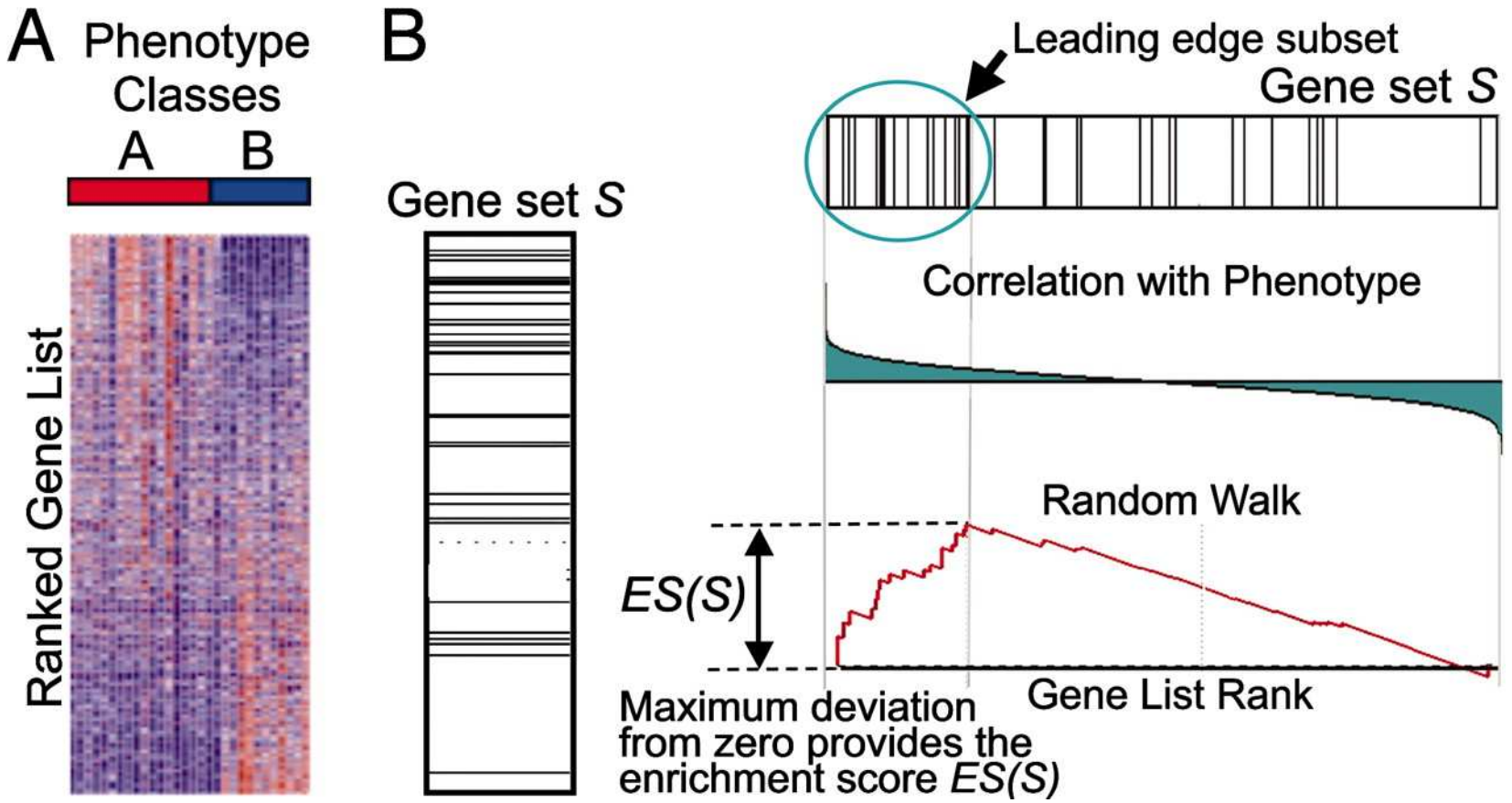
# How to Use GSEA-P




- Load your data files into GSEA
- Set the analysis parameters and run the analysis
- Run a Leading Edge Analysis
- View the analysis results

http://www.broadinstitute.org/gsea/index.jsp

# A GSEA overview illustrating the method




Subramanian A et al. PNAS 2005;102:15545-15550

# Why We Focus On GSEA

- **Single-gene analysis has limitation**
- No individual gene meet the threshold for statistical significance
- One may be left with a long list of statistically significant genes without any unified biological theme
- Single-gene analysis may miss important effects on pathways
- The lists of statistically significant genes from the two studies may show little overlap
- **GSEA can address these analytical obstacles**
- **Chip gradually becomes an common tool**

# Normalization and GSEA

- The GSEA algorithm expects different levels of expression and provides better results when given all of the data
- When we get the expression files, we'd better ignores Present/Marginal/Absent calls and do not filter the data
- Consequently, the method for normalization becomes very important

# Results

- Experiment design
- Six samples, three for control and three for treated
- Expression dataset file (gct)
- RMA background correct, PMONLY pmcorrect
- Then 7 normalize method
- Phenotype labels file (cls)
- Control VS treated
- Gene sets file (gmt)
- 62 immune pathways
- Chip (array) annotation file (chip)

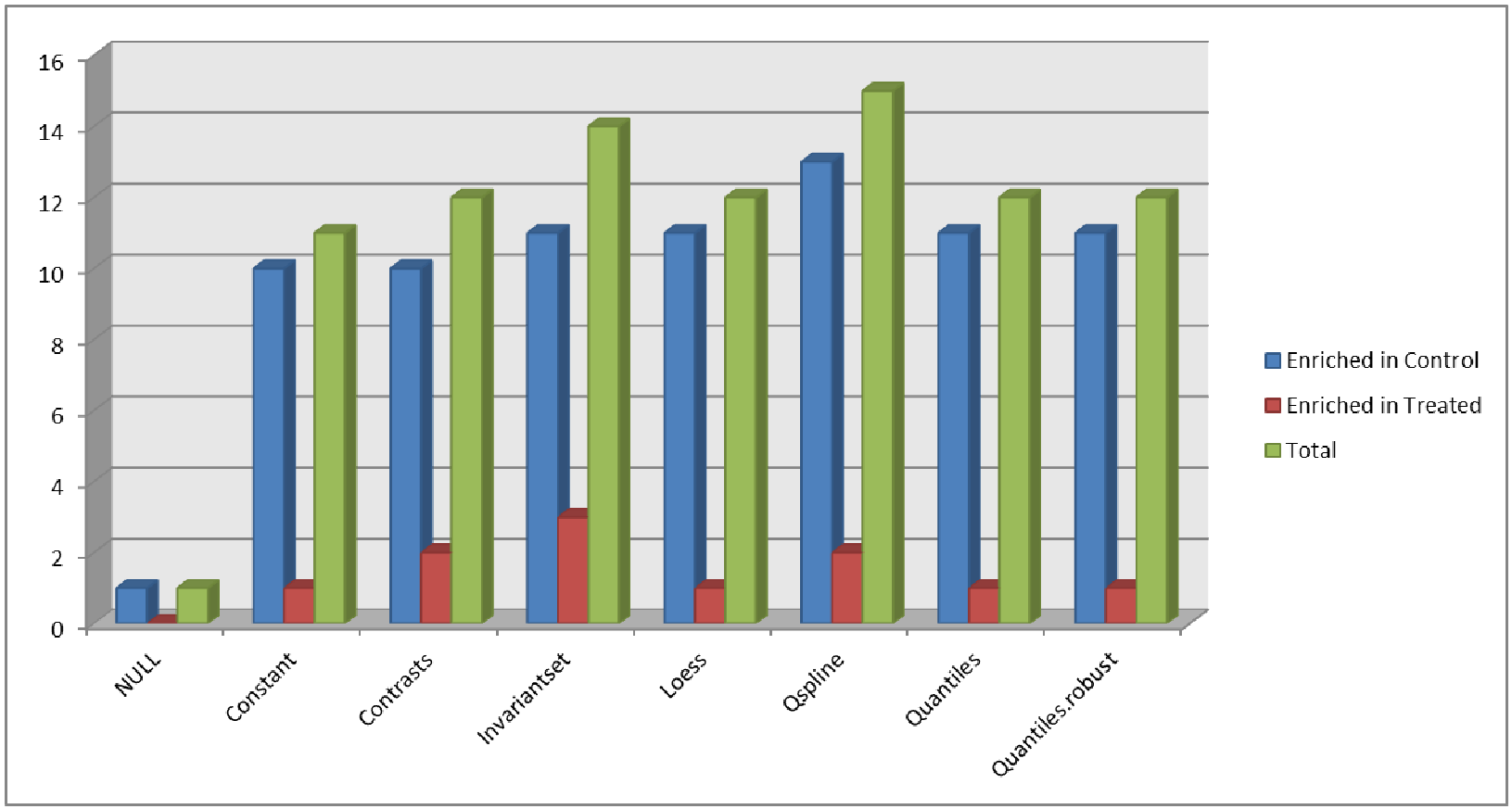
  Affymetrix Porcine Genome Array
- GSEA-P or GSEA-P-R

# Significance of Up-regulated pathways in treated

| Methods | Gene set | Source | Size | ES | NES | NOM p-val | FDR q-val |
|---|---|---|---|---|---|---|---|
| **NULL** | Non | Non | Non | Non | Non | Non | Non |
| **Constant** | TRAF6 mediated NF-KB activation | Reactome | 24 | 0.53 | 1.74 | 0.004 | 0.049 |
| **Contrasts** | Advanced glycosylation endproduct receptor signaling | Reactome | 18 | 0.66 | 1.92 | 0.000 | 0.011 |
| | TRAF6 mediated NF-KB activation | Reactome | 24 | 0.54 | 1.71 | 0.013 | 0.041 |
| **Invariantset** | Advanced glycosylation endproduct receptor signaling | Reactome | 18 | 0.68 | 1.74 | 0.001 | 0.033 |
| | Complement and coagulation cascades | KEGG | 80 | 0.49 | 1.7 | 0.000 | 0.027 |
| | Complement cascade | Reactome | 22 | 0.6 | 1.63 | 0.013 | 0.043 |
| **Loess** | Advanced glycosylation endproduct receptor signaling | Reactome | 18 | 0.64 | 1.89 | 0.000 | 0.019 |
| **Qspline** | Advanced glycosylation endproduct receptor signaling | Reactome | 18 | 0.69 | 1.92 | 0.000 | 0.007 |
| | TRAF6 mediated NF-KB activation | Reactome | 24 | 0.58 | 1.74 | 0.007 | 0.047 |
| **Quantiles** | Advanced glycosylation endproduct receptor signaling | Reactome | 18 | 0.7 | 1.84 | 0.004 | 0.018 |
| **Quantiles.robust** | Advanced glycosylation endproduct receptor signaling | Reactome | 18 | 0.7 | 1.85 | 0.000 | 0.017 |

**Entries of all sets with nominal p value ≤ 0.05, false discovery rate (FDR) ≤ 0.05.**
**ES, enrichment score; NES, normalized enrichment score**

# Detection Rate of Different Normalization Method




The abscissa on behalf of different normalization methods, the ordinate represent the number of  significant pathways

# Connectivity Map Based on DGEs

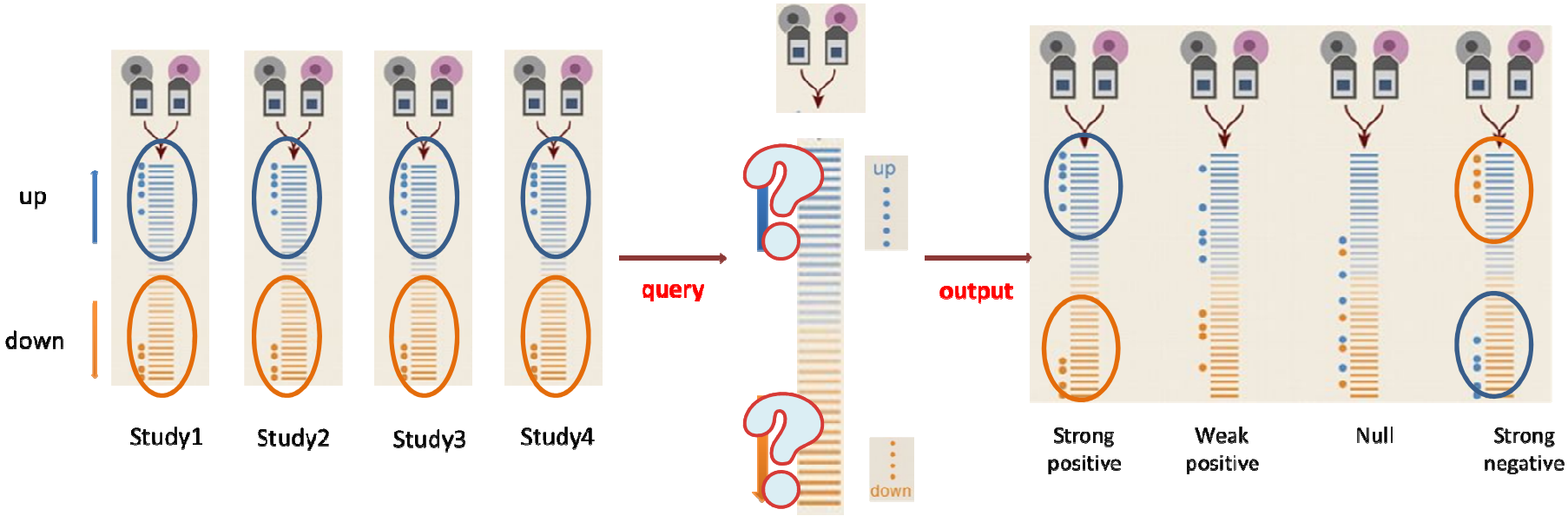
- Comparison of correlations among different studies by GSEA
- Considering up and down regulated DGEs as a gene-set
- Quantify how well the up (and down) regulated genes rank in the ordered list
- How to ?

# Connectivity Map

Query signature : up- and down-regulated genes of studies

Reference signatures : ranked gene lists for a reference study

Output : lists of high and low scoring correlation



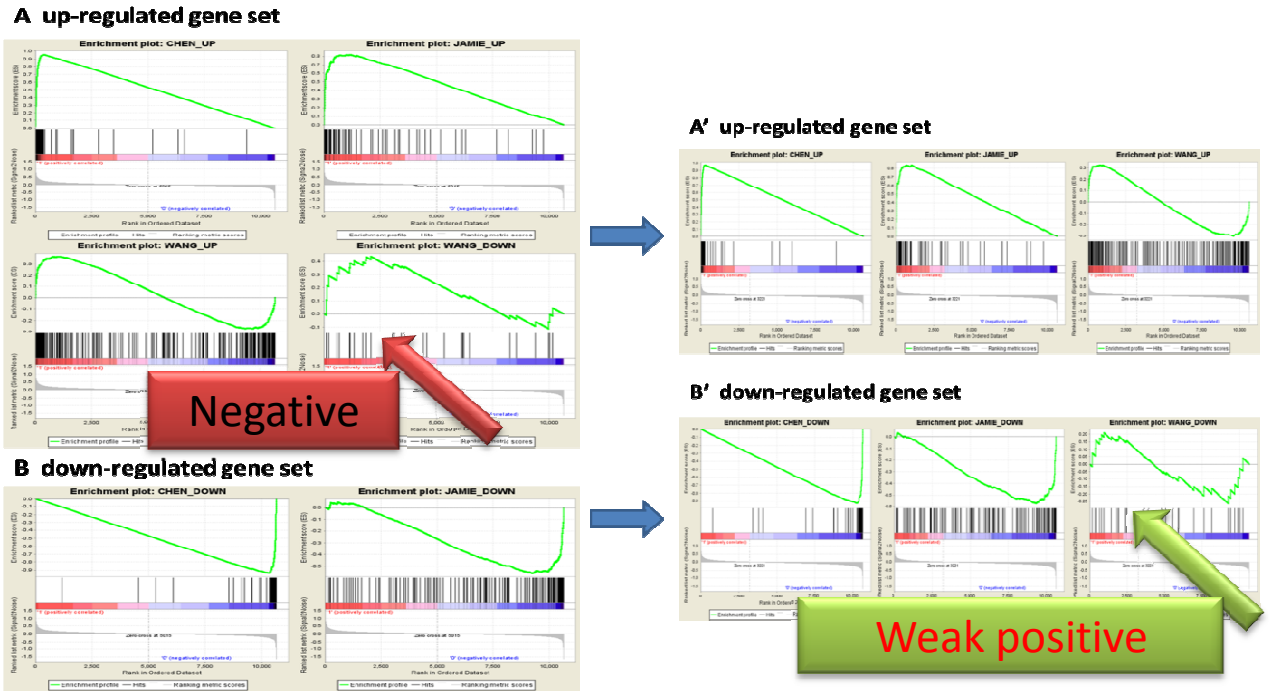
# Connective Score of Published Studies

GESA analysis of published transcriptomes with our data by **quantiles**

| Gene set | SIZE | ES | NES | NOM p-val | FDR q-val | CS |
|---|---|---|---|---|---|---|
| CHEN_UP | 124 | 0.95 | 3.780 | 0.00 | 0.00 | 0.945 |
| CHEN_DOWN | 89 | -0.94 | -3.550 | 0.00 | 0.00 | |
| JAMIE_UP | 83 | 0.82 | 3.000 | 0.00 | 0.00 | 0.69 |
| JAMIE_DOWN | 146 | -0.56 | -2.300 | 0.00 | 0.00 | |
| WANG_UP | 229 | 0.37 | 1.570 | 0.00 | 0.008 | 0.405 |
| WANG_DOWN | 31 | **0.44** | 1.33 | 0.124 | 0.059 | |

ES: Enrichment Score; NES: Normalized Enrichment Score; FDR: False Discovery Rate

# Correlations among Different Studies Base on DEGs




Invariantset，Qspline，Quantiles，Quantiles.robust

Constant，Contrasts，Loess

# Summary

- In our study, we compared the effect of different normalization method from the aspect of detection rate and connection map
- **Detection rate**: invariantset ,qspline
- **GSEA connecting map**: invariantset ,qspline quantiles,quantiles.robust
- Quantiles?

# Discussion

- **A more reliable data set：**
- More duplicates（affymetrix spike-in data ?)
- Reasonable experiment design（typical pathway variation）
- **An appropriate normalization procedure ?**
- On which level should the normalization be performed
- How to actually perform the normalization

# Enlightenments

- How to keep the constant competence against novel sequencing technologies?
- We need more feasible tools to dig out more information in mass chip data
- Biological background challenges the most
- R is the optimal tool

# Thank you!

Questions?

zhaoming159753@gmail.com