Statistical Assessment of Depth Normalization Methods for MicroRNA Sequencing

Jian Zou

Department of Biostatistics

University of Pittsburgh

Advisor: **Li-Xuan Qin**

Department of Epidemiology and Biostatistics

Memorial Sloan Kettering Cancer Center

Background

Data artifacts due to disparate experimental handling is a serious issue for molecular profiling data, which demonstrates the necessity of normalization

NATURE REVIEWS | GENETICS 2010

OPINION

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry

Challenge

- One major and unique aspect of RNA sequencing data normalization is the *depth of coverage*
- MicroRNAs are molecules regulating gene expression and the assumption of depth normalization methods may not hold for microRNA sequencing

self-assessment Trap

Our Study

- We perform a study to assess the performance of existing popular depth normalization methods
- Both a pair of datasets on the same set of tumor samples and data simulated from the paired datasets under various scenarios of differential expression are used.

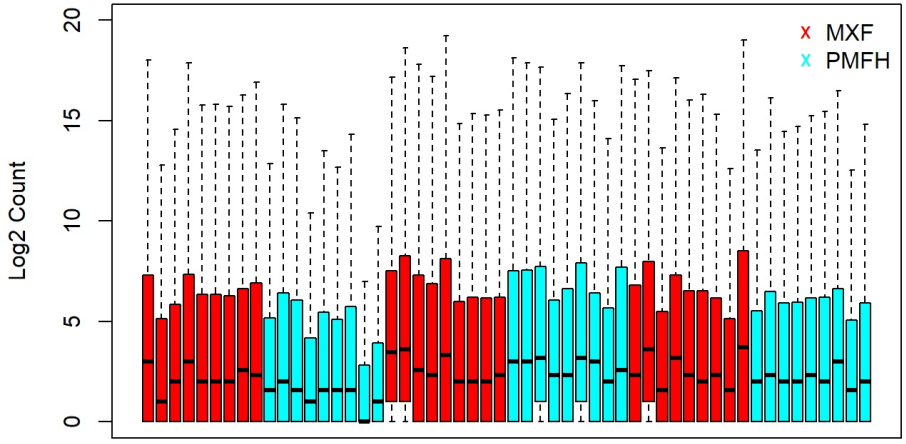
Method	**Reference**
Total-count	Dillies
Upper-quartile	Bullard
Median	Dillies
TMM	Robinson
DESeq	Anders
PoissonSeq	Li
Quantile Normalization	Bolstad
SVAseq	Leek
RUV-seq	Risso

Empirical Data Preparation

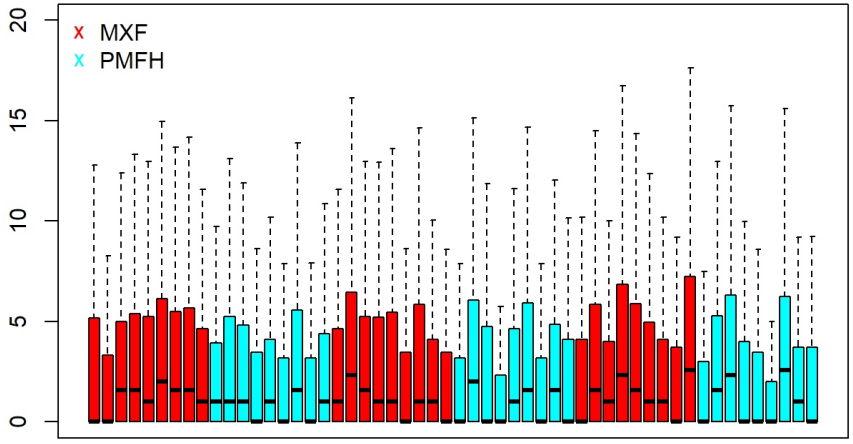
We collect *two datasets* for the same set of 54 samples

- First dataset (test data)
  - First come first serve
  - Collected over several years
- Second dataset (benchmark data)
  - **Balanced library-assignment** for the samples to avoid confounding
  - **Uniform handling**
  - **Three quality control measures**:
    1. Calibrators
    2. Pooled samples
    3. Technical replication

Empirical Data Overview

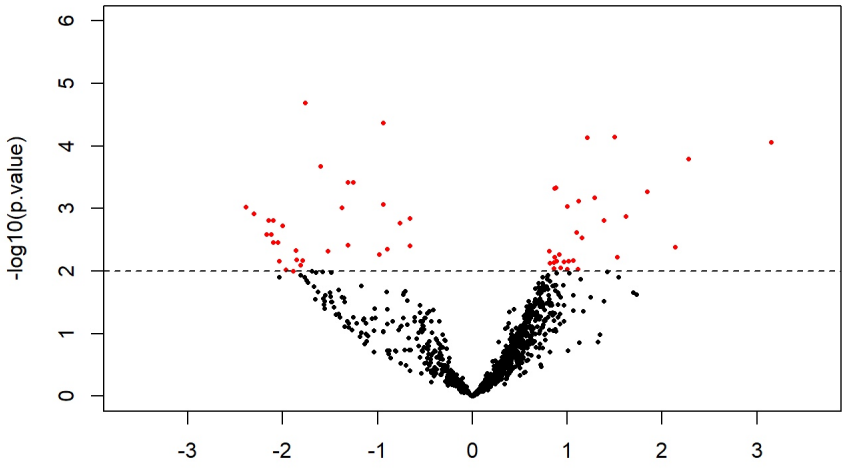


Distribution for the Benchmark Data

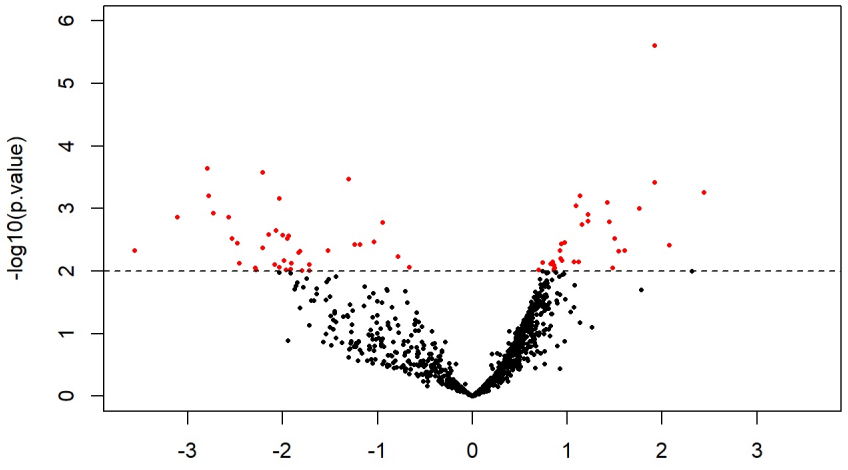


Distribution for the Test Data

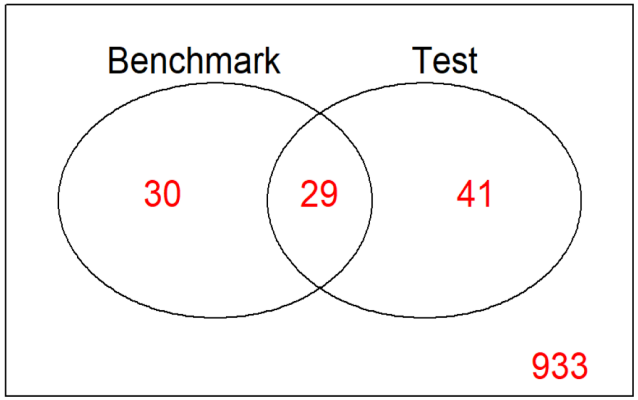
DEA Comparison: Benchmark V.S. Test



Volcano Plot for Benchmark Data

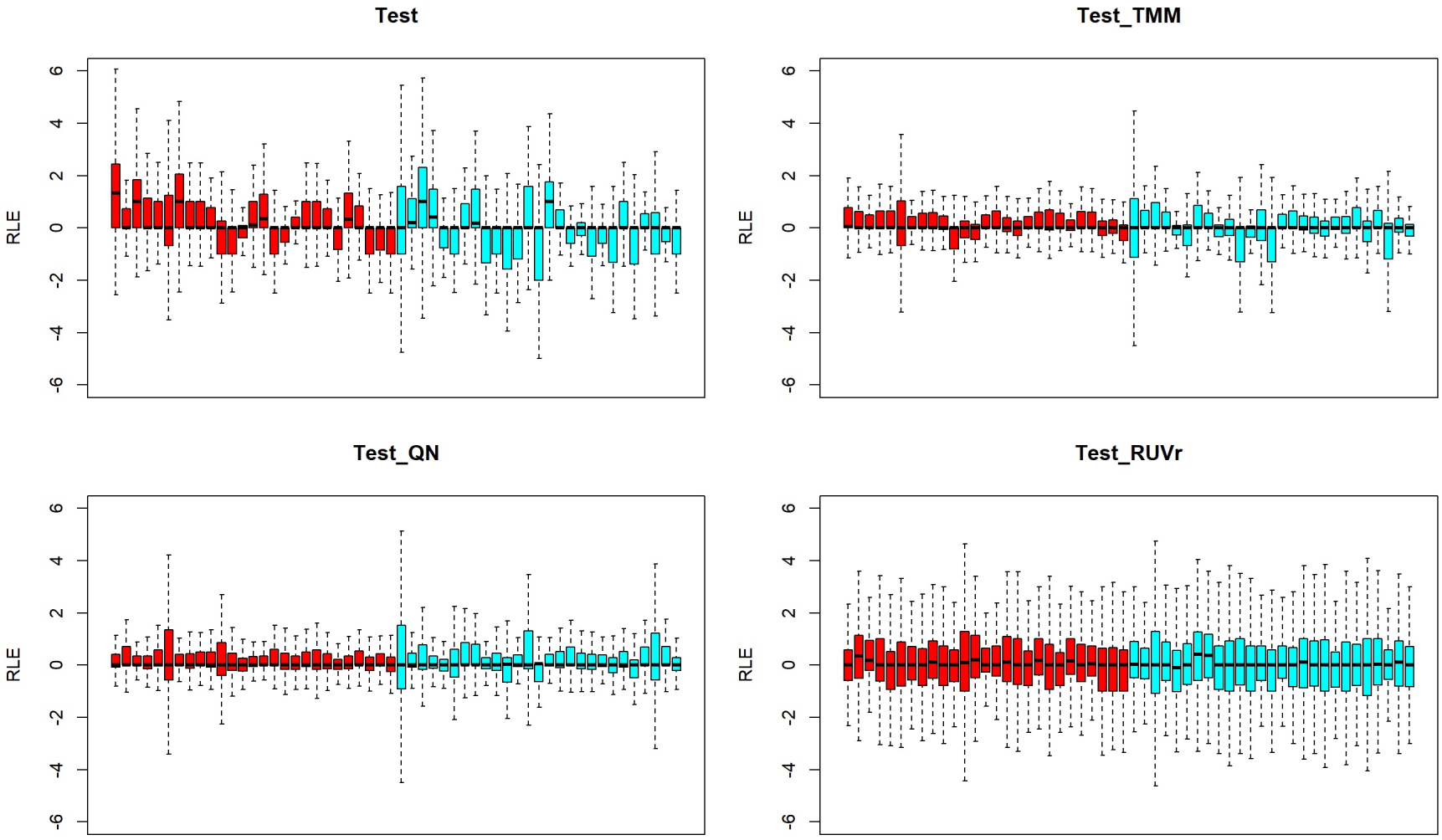


Volcano Plot for Test Data



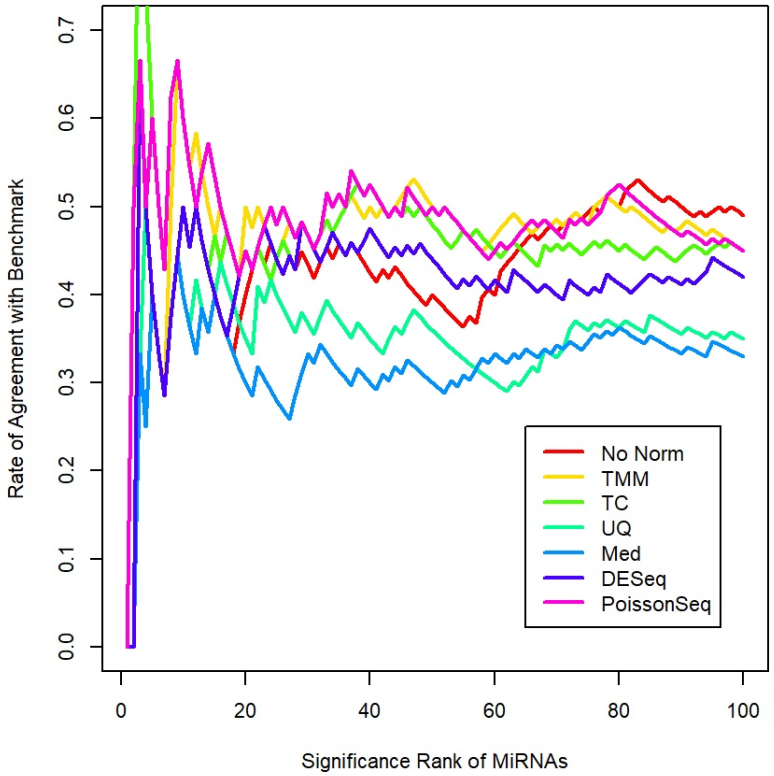
Venn Diagram

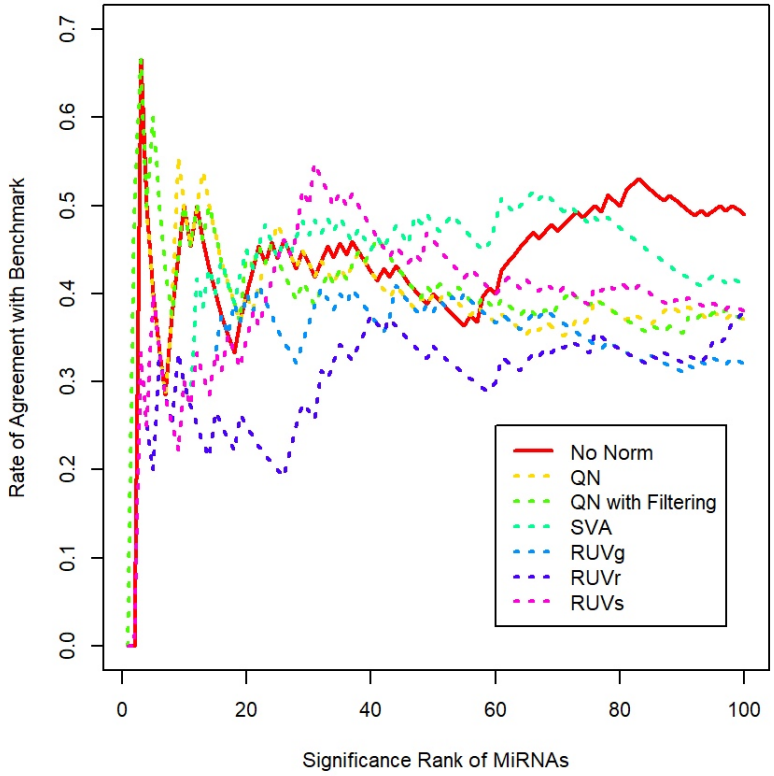
Empirical Data Normalization



Relative Log Expression for Normalized Data

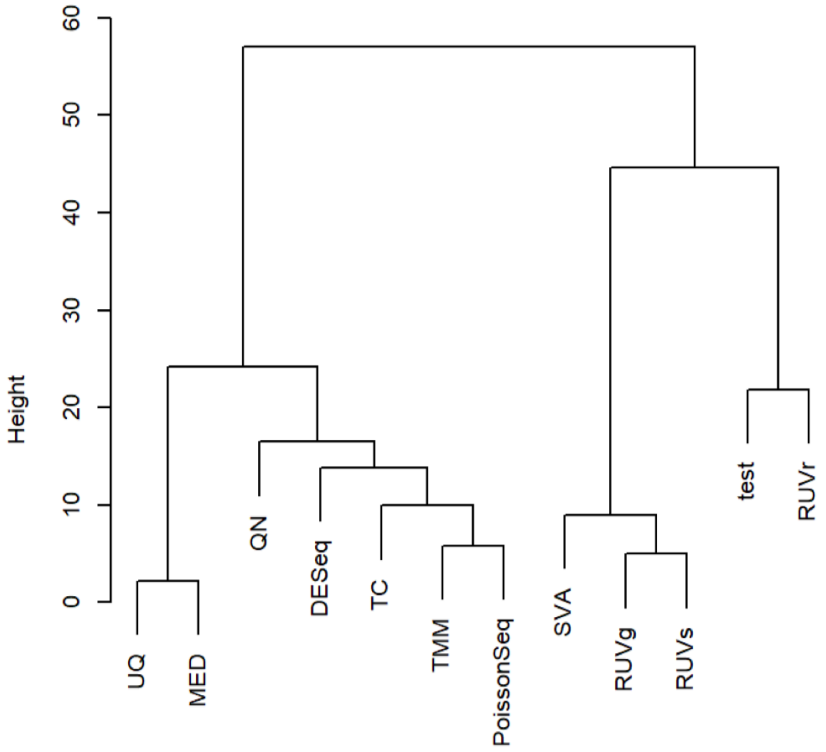
DEA Comparison: CATPlots



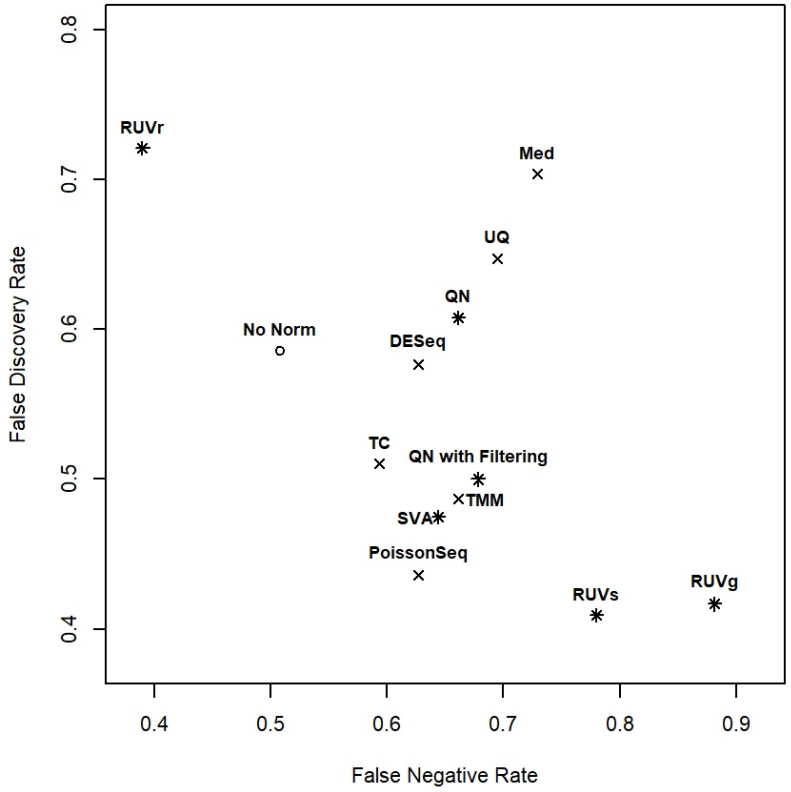


Concordance At The Top Plot for the Significance Levels

DEA Comparison: Dendrogram and Scatterplot



Dendrogram



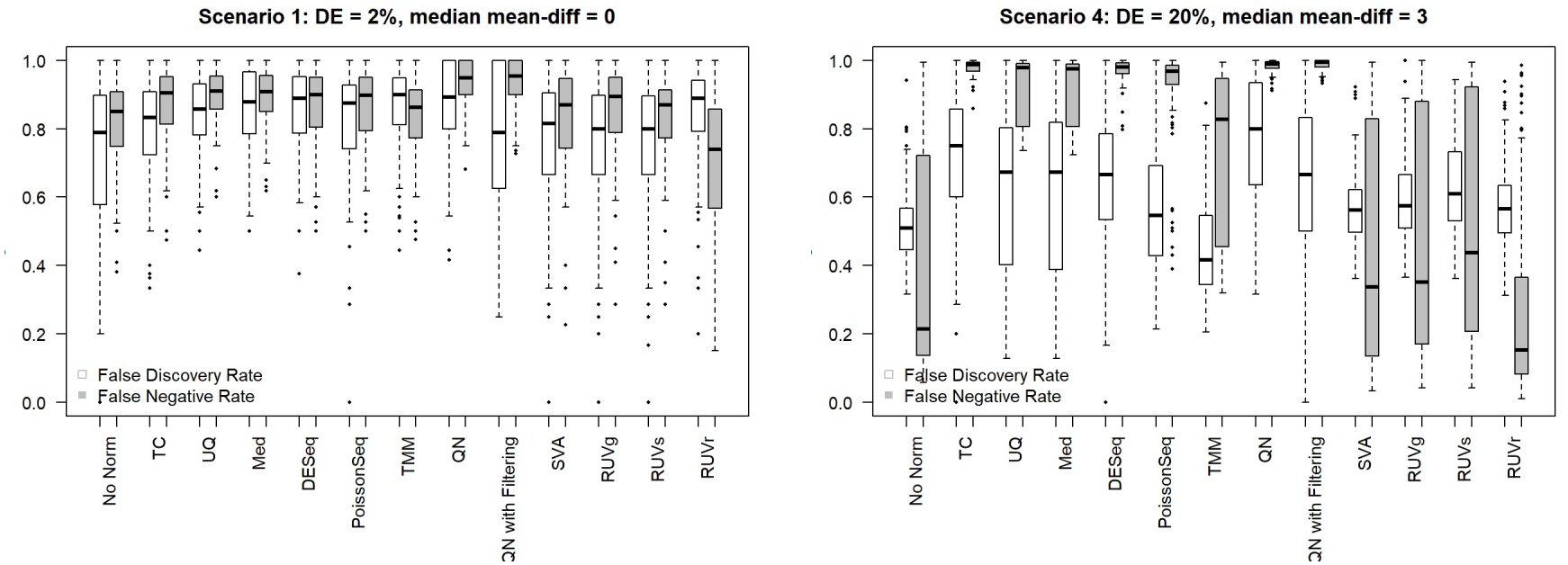
Scatterplot for FNR and FDR

Simulated Data Preparation

We simulate datasets for different scenarios of DE proportion and median of mean differences

- **Clustering** 54 empirical samples of benchmark data into two groups
- Randomly **selecting** 9 samples from each cluster, with each three of them from the same sequencing library
- **Allocating** the remaining 36 samples into two groups randomly, with ensuring same number of samples from same sequencing library
- **Generating** the corresponding simulated test data using the same allocation of the simulated benchmark data

Simulated Data Analysis: Boxplot



Boxplot of FDR and FNR for different methods in different scenarios

Conclusions

- Performance of normalization methods depends on the specific pattern of differential expression and in general only brought limited benefits to the analysis of differential expression
- *TMM* tends to outperform the other scaling-based normalization methods, and *RUVr* tended to outperform the other regression-based normalization methods
- *Median* and *upper-quartile are* consistently the worst performers across all methods examined in our study
- We have developed an R package including paired datasets, empirical analysis and simulations

Acknowledgement

- Li-Xuan Qin, PI

- Graduate Students
  - Jian Zou @University of Pittsburgh
  - Yannick Duren @Ruhr-University Bochum

- Collaborators
  - Samuel Singer, Surgical Oncologist @MSK
  - Thomas Tuschl, Molecular Biologist @Rockefeller University