An evaluation of statistical differential analysis methods in single-cell RNA-Seq data

Dongmei Li, PhD

Clinical and Translational Science Institute
University of Rochester School of Medicine and Dentistry

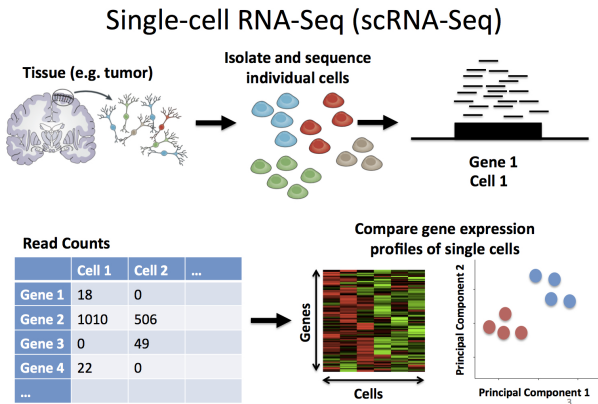*dongmei_li@urmc.rochester.edu*

The 2019 Meeting of The International Society for Data Science and Analytics

July 6, 2019

Overview

1 Introduction

2 Single-cell RNA Sequencing analysis methods

3 Simulation Studies

4 Real data example

5 Discussion

Single-cell RNA Sequencing Overview

Figure: *A general overview of scRNA-seq. Source:*
https : //learn.gencore.bio.nyu.edu/single − cell − rnaseq/
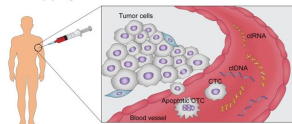
Use of Single-cell RNA sequencing

- Bulk RNA-seq measure the average level of gene expression of multiple cells

- Single-cell RNA-seq allow us to understand gene expression pattern within the cell

- Single-cell RNA-seq can identify cell heterogeneity, cell population and sub-population

- Single-cell RNA-seq can examine the effects of low copy mRNA distribution and transcriptional regulation
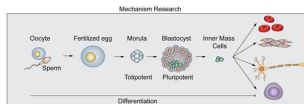
Single-cell RNA Sequencing Applications
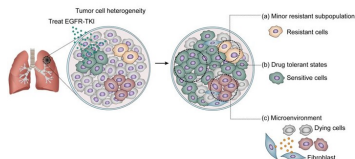
Figure: *Application of single-cell RNA sequencing technology in biological and biomedical research. More could be found from:*
https : //www.nature.com/articles/s12276 − 018 − 0071 − 8

Single-cell RNA sequencing differential analysis methods in Bioconductor/R

- DEsingle (Bioinformatics, 2018)

- Linnorm (Nucleic Acids Research, 2017)

- Monocle2 (Nature Methods, 2017)

- MAST (Genome Biology, 2015)

- DESeq2 (Genome Biology, 2014)

Motivation for evaluating RNA sequencing data analysis methods

- Multiple methods provide inconsistent results for the same dataset

- Exploring false discovery rate control (FDR), sensitivity, specificity, accuracy, and AUC under the ROC curve across different methods for the same dataset

- Proving guidance for investigators to choose appropriate method for their singlle-cell RNA sequencing data analysis

Definition of FDR, Sensitivity, Specificity, Accuracy

	number not rejected	number rejected	
true null hypotheses	U	V	m_0
non-true null hypotheses	T	S	m_1
total	$m - R$	R	m

Table: *Possible outcomes from m hypotheses tests*

$$FDR = E(\frac{V}{R})Pr(R > 0),$$

$$Sensitivity = E(\frac{S}{m_1}), \quad Specificity = E(\frac{U}{m_0})$$

$$Accuracy = E(\frac{U + S}{m})$$

AUC under the ROC curve is calculated using the AUC function in the pROC package in R.

DEsingle

DEsingle use the Zero-Inflated Negative Binomial (ZINB) model to decrible the read counts and excess zeros in single-cell RNA sequencing data. The count data for gth gene in a group of cell are assumed to follow ZINB distribution:

$$Pr(Y_g = y | \theta, r, p) = \theta \times I(y = 0) + (1 - \theta) \times f_{NB}(r, p),$$

where θ is the proportion of constant zeros of gene g in the group of cells, $I(y = 0)$ is an indicator function, $f_{NB}(r, p)$ is the probability mass function of Negative Binomial distribution with parameters r and p.

- DEsingle use likelihood-ratio tests for gene differential analysis.

Linnorm

- Linnorm proposed a novel normalization and transformation method for single-cell RNA-seq analysis.

- The normalization and transformation parameters were calculated based on stably expressed genes across different cells.

- The moderated t test statistics in the limma package for differential analysis through the empirical Bayes approach.

Monocle2

- Monocle2 use census algorithm to convert relative RNA-seq expression levels into relative transcript counts without the need for experimental spike-in controls.

- The census algorithm calculate the total number of single-mRNA genes and divide this number by the fraction of the library contributed by them to estimate the total number of captured mRNAs in the cell and then rescale the transcript per million (TPM) in single cell values into mRNA counts for each gene.

- Monocle2 tests gene differential analysis through a likelihood ratio test for comparing a full generalized linear model with additional effect to a reduced generalized linear model based on negative binomial distributions.

MAST

- MAST propose a hurdle model approach for scRNA-seq data analysis.

- In the scRNA-seq expression data Y_{ig}, the rate of expression and the level of expression for the expressed cells are assumed conditionally independent for each gene g.

- MAST use an indicator variable Z_{ig} to denote whether gene g is expressed in cell i ($z_{ig} = 0$ if $y_{ig} = 0$ and $z_{ig} = 1$ if $y_{ig} > 0$).

- MAST fits a logistic regression for the discrete variable Z and a normal distributed linear model for the continuous variable ($Y|Z = 1$) independently.

$$logit(Pr(Z_{ig} = 1) = X_i\beta_g^D, \quad Pr(Y_{ig} = y|Z_{ig} = 1) = N(X_i\beta_g^C, \sigma_g^2)$$

DESeq2

- DESeq2 uses a generalized linear model approach to accommodate complex study designs.

- DESeq2 uses a logarithm link between relative gene abundance and design matrix.

- DESeq2 integrates the dispersion estimate and fold change estimate using empirical Bayes approach and test the differential expression using a Wald test.

Simulation Set up

- RNA sequencing count data are generated from negative binomial (NB), Poisson, and log normal distributions using RnaXSim function in R

- The means and variance are generated using two sets of real RNA sequencing count data with estimated different proportions of zeros in the data

- 1000 genes and 20 independent simulations

- Fraction of differentially expressed genes (π_1) were set at 5%, 10%, 20%, 30%, 40% and 50%

- Sample sizes are 5, 10, and 15 in each group with two-group comparisons

Case One

- Proportion of all zeros in the simulated data set ranged from 4% to 6% for negative binomial distribution, 2% - 3% for Poisson distribution, and 1% for log normal distribution.

- The proportion of all zeros are pretty constant across different sample sizes.

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group with NB distribution
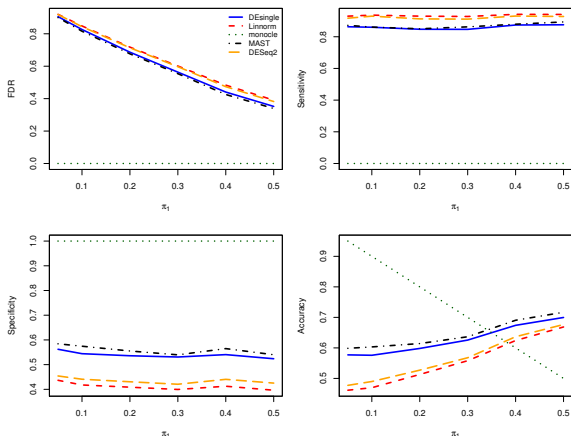
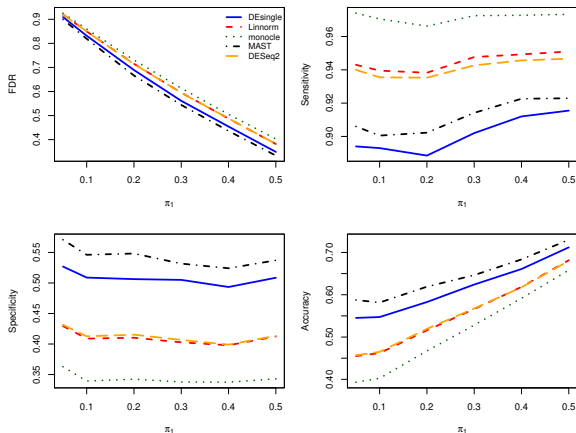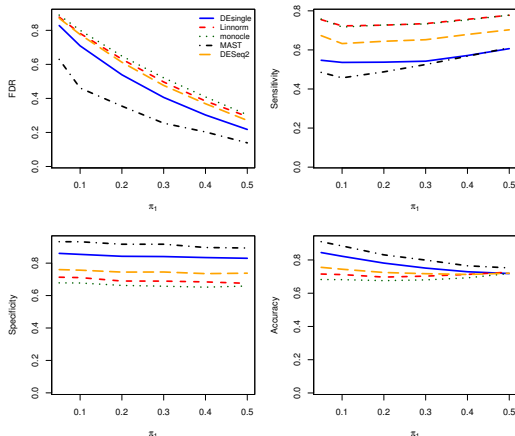**Figure:** *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group*

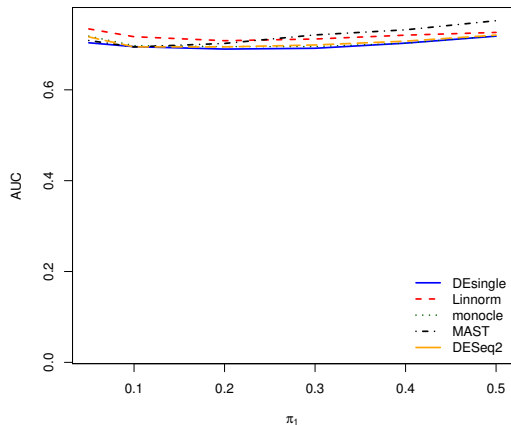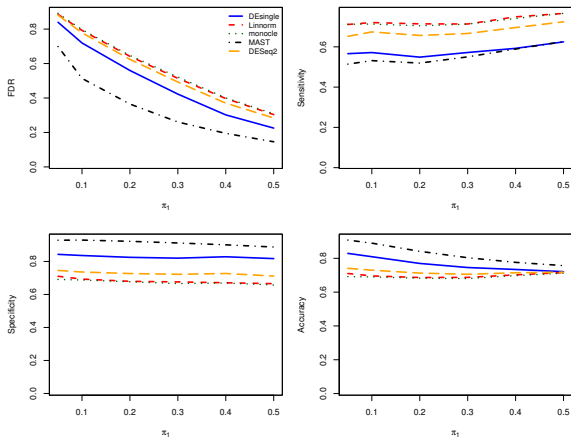# AUC under ROC curves comparisons for sample size 5 in each group with NB distribution

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 10 in each group with NB distribution
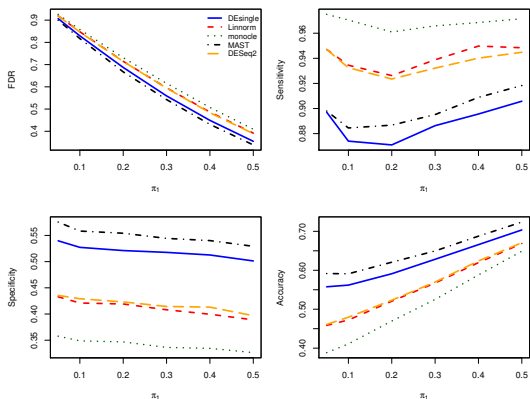
Figure: *Box plots of simulation results for $n = 6$ and equal library size*

AUC under ROC curves comparisons for sample size 10 in each group with NB distribution
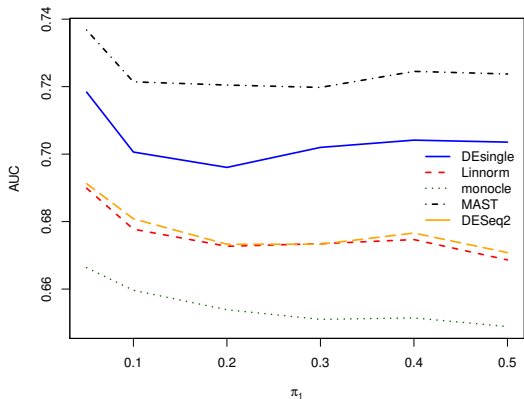
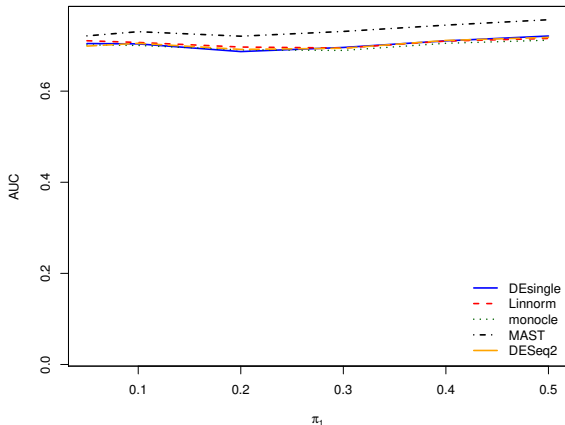**Figure:** *AUC under ROC curves comparisons for sample size 10 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 15 in each group with NB distribution

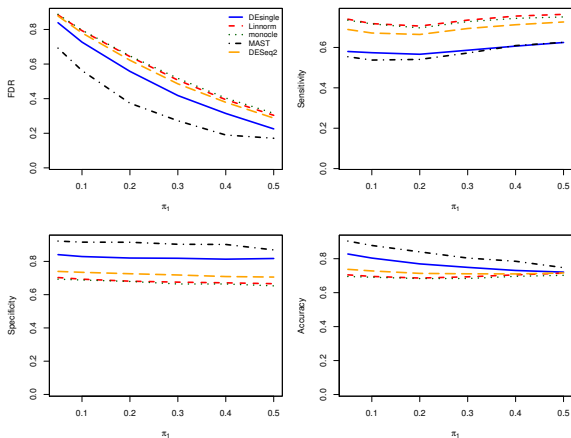**Figure:** *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size* 15 *in each group*

AUC under ROC curves comparisons for sample size 15 in each group with NB distribution

Figure: *AUC under ROC curves comparisons for sample size 15 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group with Poisson distribution

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group*

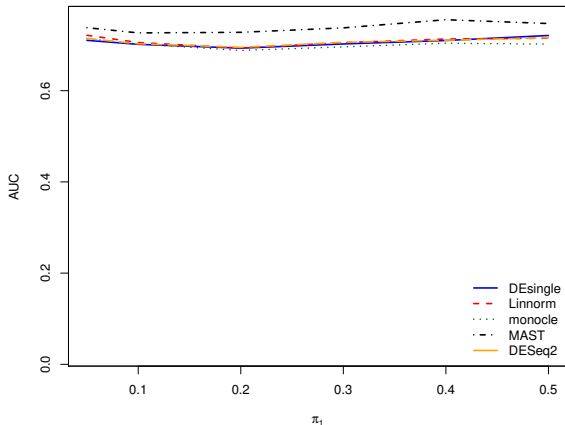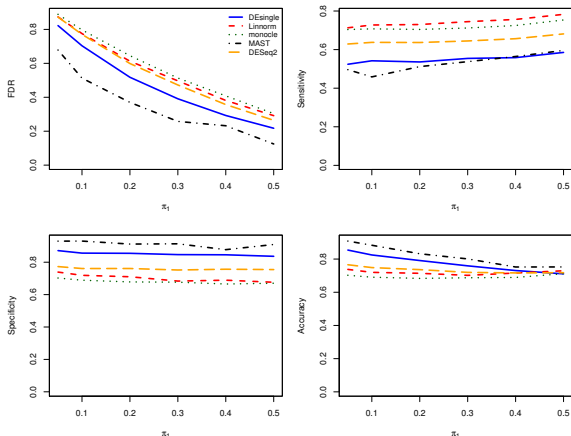# AUC under ROC curves comparisons for sample size 5 in each group with Poisson distribution

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 10 in each group with Poisson distribution

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size* 5 *in each group*

AUC under ROC curves comparisons for sample size 10 in each group with Poisson distribution
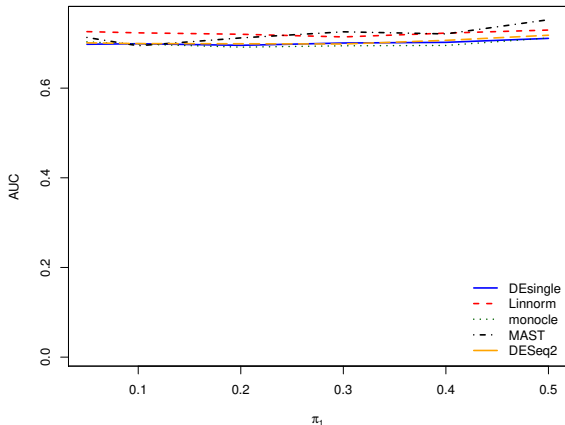

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 15 in each group with Poisson distribution
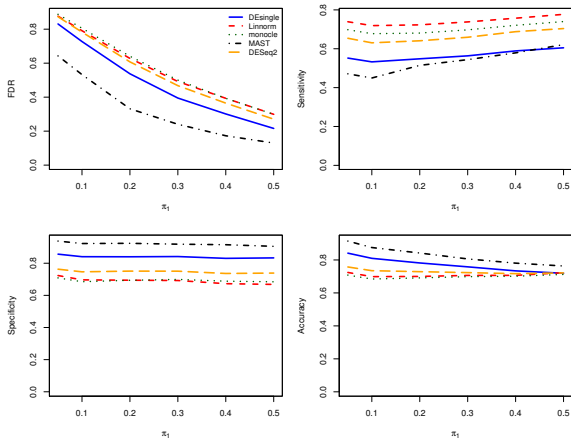

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size* 5 *in each group*

AUC under ROC curves comparisons for sample size 15 in each group with Poisson distribution
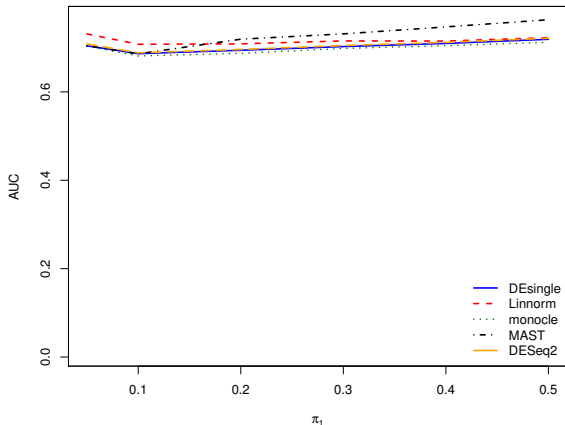

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group with Log Normal distribution
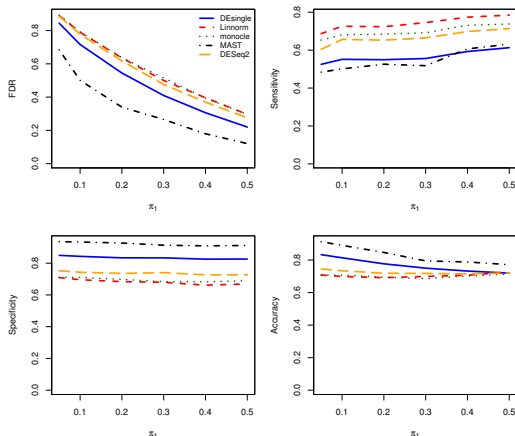

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group*

AUC under ROC curves comparisons for sample size 5 in each group with Log Normal distribution

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*
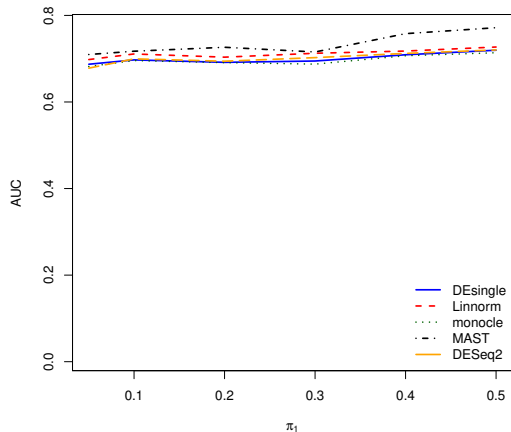
FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 10 in each group with Log Normal distribution

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size* 5 *in each group*

AUC under ROC curves comparisons for sample size 10 in each group with Log Normal distribution
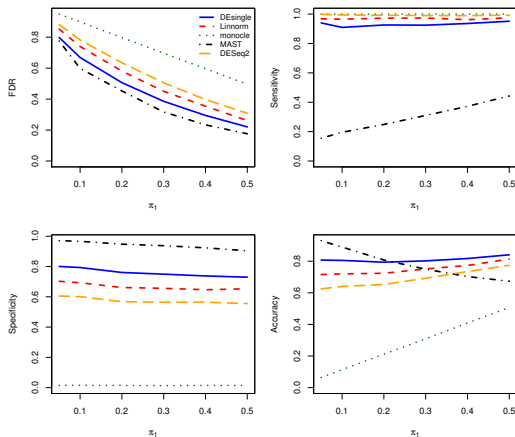

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 15 in each group with Log Normal distribution

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group*

AUC under ROC curves comparisons for sample size 15 in each group with Log Normal distribution

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*
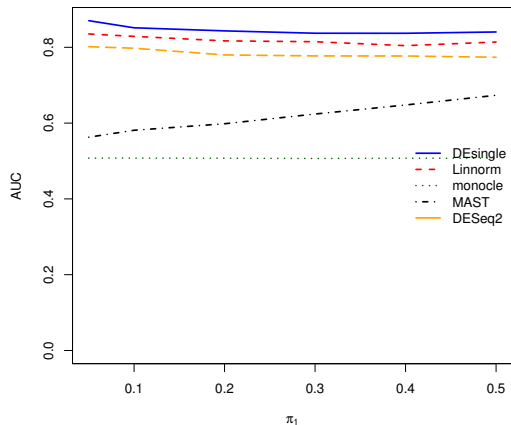
Case Two

- Proportions of all zeros in the simulated data set are less than 0.4% for negative binomial distribution and close to zero for Poisson and log normal distribution.

- The proportion of all zeros are pretty constant across different sample sizes.

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group with NB distribution
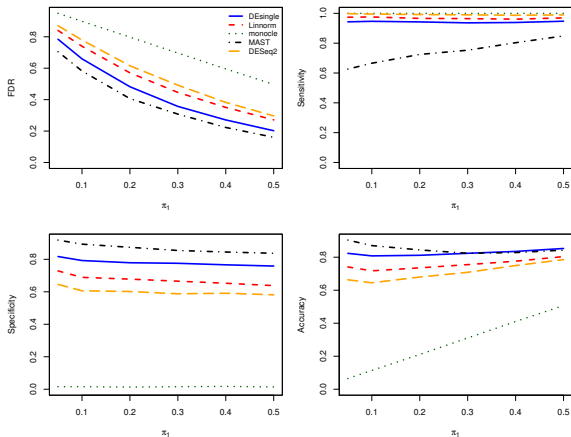

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group*

AUC under ROC curves comparisons for sample size 5 in each group with NB distribution
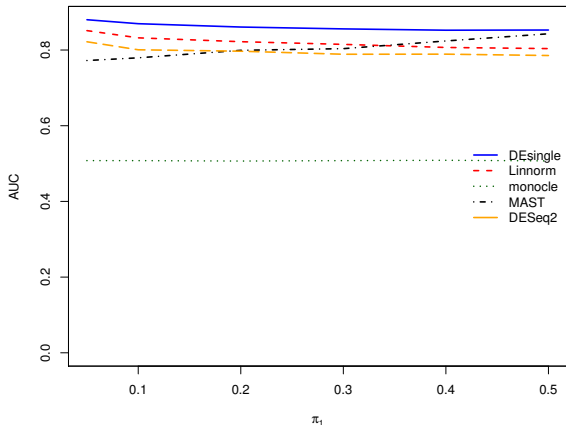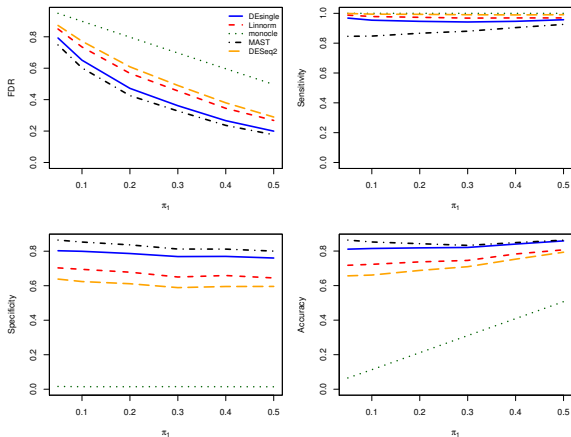

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 10 in each group with NB distribution

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size* 5 *in each group*

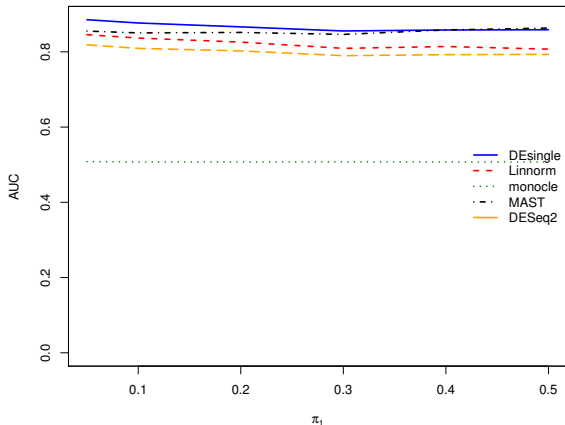# AUC under ROC curves comparisons for sample size 10 in each group with NB distribution

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 15 in each group with NB distribution

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size* 5 *in each group*
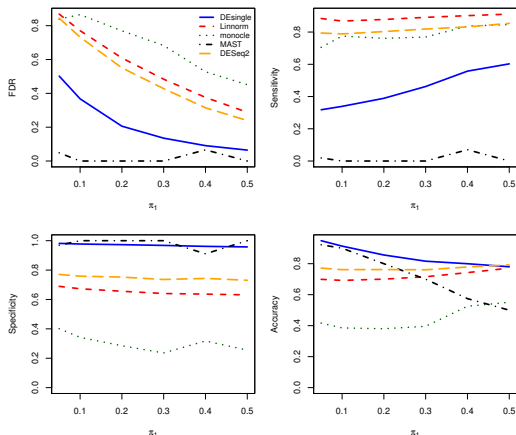
AUC under ROC curves comparisons for sample size 15 in each group with NB distribution

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*
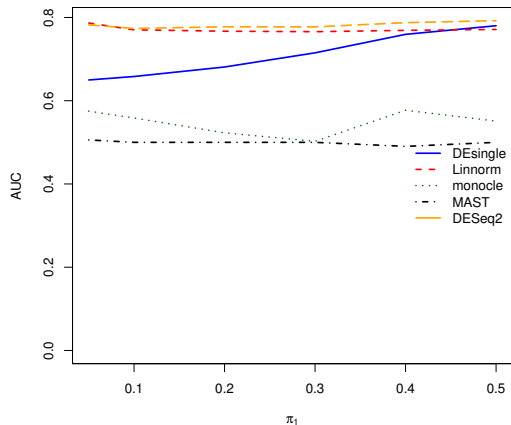
FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group with Poisson distribution
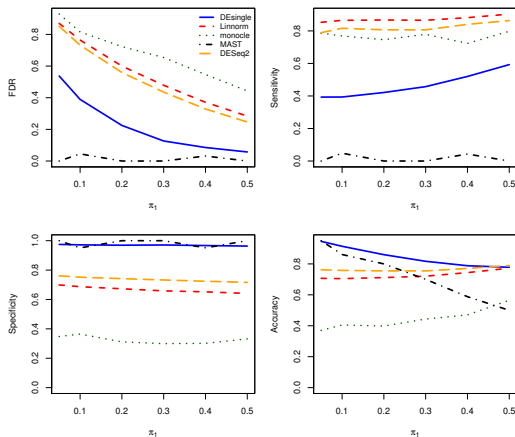

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size* 5 *in each group*

AUC under ROC curves comparisons for sample size 5 in each group with Poisson distribution
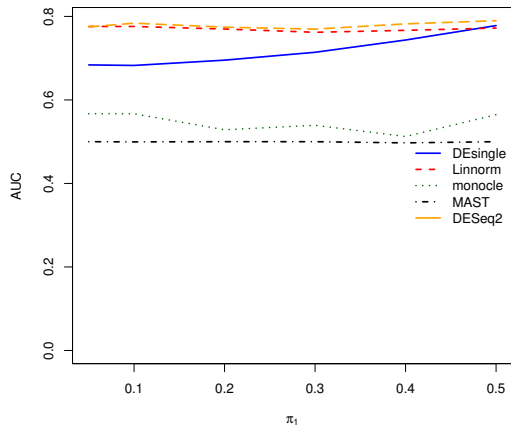

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 10 in each group with Poisson distribution
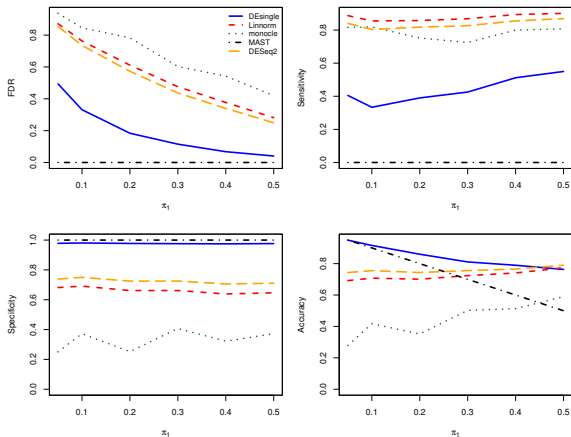

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group*

AUC under ROC curves comparisons for sample size 10 in each group with Poisson distribution
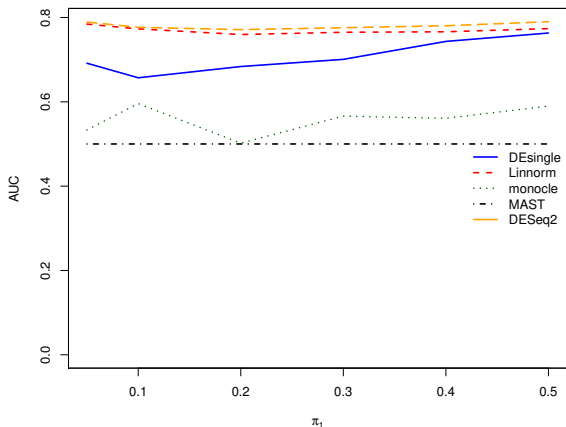

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 15 in each group with Poisson distribution
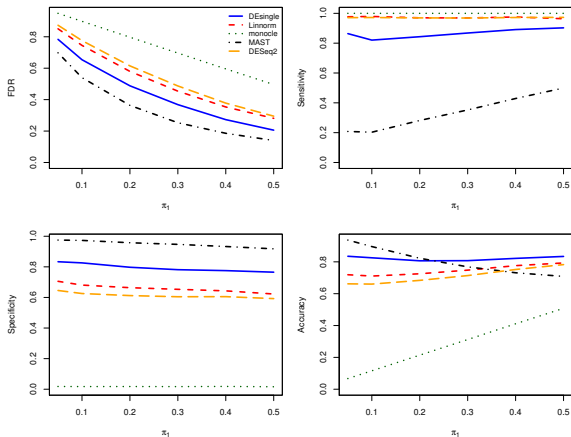

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size* 5 *in each group*

AUC under ROC curves comparisons for sample size 15 in each group with Poisson distribution

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group with Log Normal distribution

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group*

AUC under ROC curves comparisons for sample size 5 in each group with Log Normal distribution
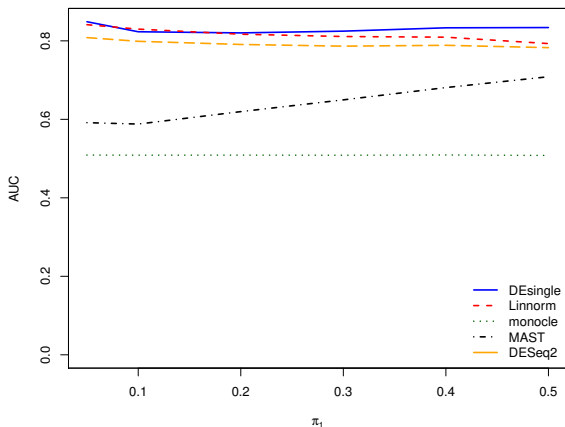

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*

FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 10 in each group with Log Normal distribution

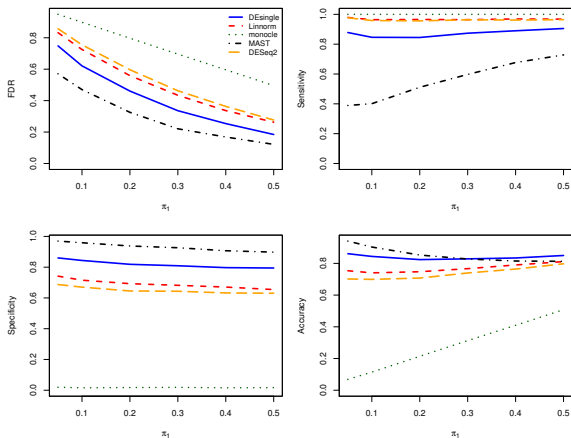**Figure:** *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group*

AUC under ROC curves comparisons for sample size 10 in each group with Log Normal distribution

Figure: *AUC under ROC curves comparisons for sample size 5 in each group*
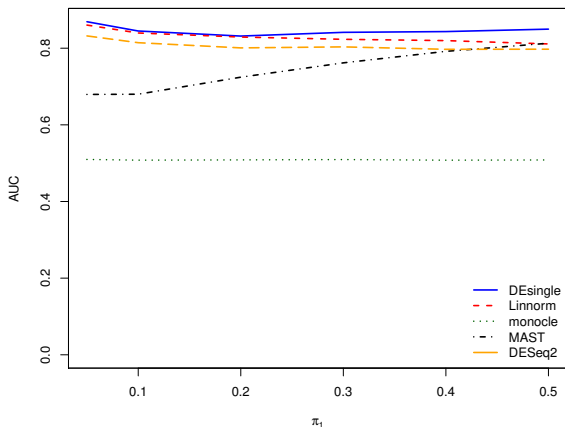
FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 15 in each group with Log Normal distribution
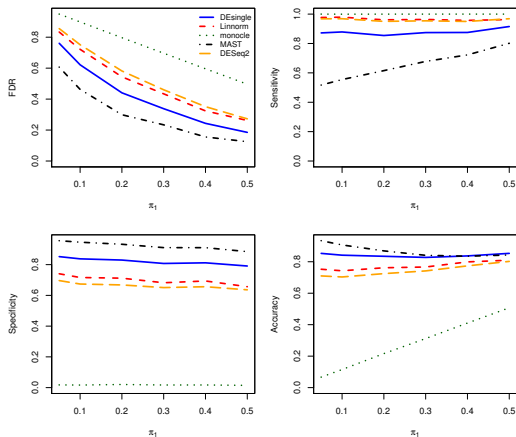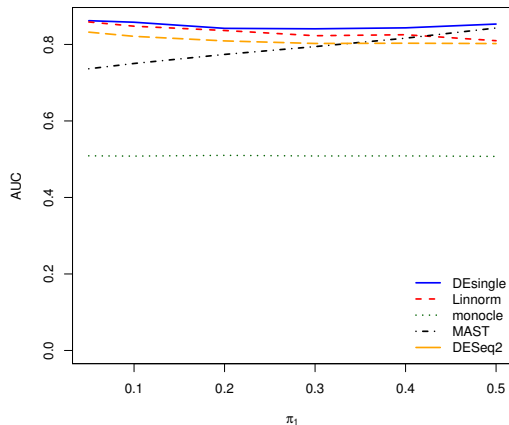

Figure: *FDR, Sensitivity, Specificity, and Accuracy comparisons for sample size 5 in each group*

AUC under ROC curves comparisons for sample size 15 in each group with Log Normal distribution


Figure: *AUC under ROC curves comparisons for sample size* 5 *in each group*

Real data example

- Islam dataset with 92 samples of scRNA-seq raw count data downloaded from GEO website with accession no. GSE29087.

- 48 samples are embryonic stem cells and 44 are embryonic fibroblasts from mouse.

- All methods were used for selecting differentially expressed genes between the two types of cells from 14905 genes.

- The raw p-values from all methods were adjusted using the Benjamini-Hochberg procedure to control FDR at 5%.

Empirical power of different scRNA sequencing differential analysis methods
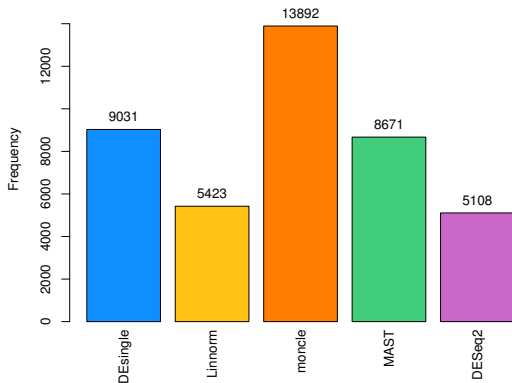
Figure: *DEsingle: dodgerblue; Linnorm: goldenrod1; monocle:darkorange1; MAST:seagreen3; DESeq2:orchid3.*

Venn diagram of selected differentially expressed genes by different scRNA sequencing differential analysis methods
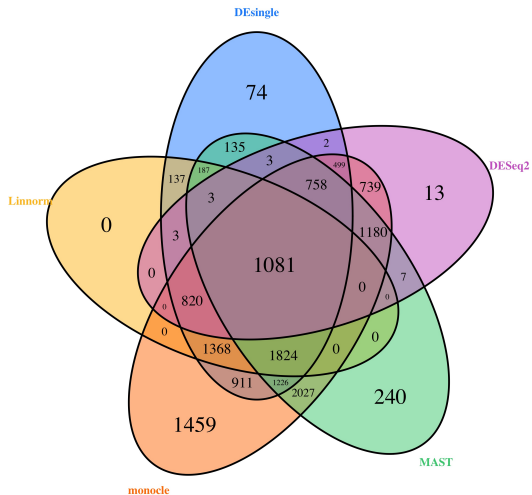
Discussion

- For all methods compared, the FDR seems decrease as the proportions of differential expressed gene increase

- Both sensitivity and specificity seems relatively stable across different proportions of differential expressed genes

- For all five methods compared, AUC under the ROC curve seems relatively stable across different proportions of differential expressed genes

- The performance of different methods seems related to the proportion of total zeros in the data set.

Discussion

- The performance of the monocle method seems depend on data distributions (Poisson and Log Normal) and proportion of total zeros in the data set (1 - 5 %).

- All other methods slightly improved with the increase of sample size

- The MAST method performed well when there are some proportions of zeros in the data set regardless of data distribution.

- The MAST method did not perform well when there is almost no zeros count in the data set for all three distributions.

Discussion

- When there is almost no zeros in the data set, DEsingle, DESeq2, and Linnorm Performs relatively well for all three data distribution.

- The similar performance of the Linnorm and the DESeq2 methods is likely due to the application of empirical Bayes approach in both methods.

- The performance of the Linnorm method seems does not affect much by proportion of zeros and different data distributions.

Conclusion

- The MAST and Linnorm methods have the best AUC under ROC curves than other methods when there are some proportion of zeros in the data set.

- The DEsingle, Linnorm, and DESeq2 methods perform relatively well when the proportion of zero counts close to zero in the data set.

Acknowledge