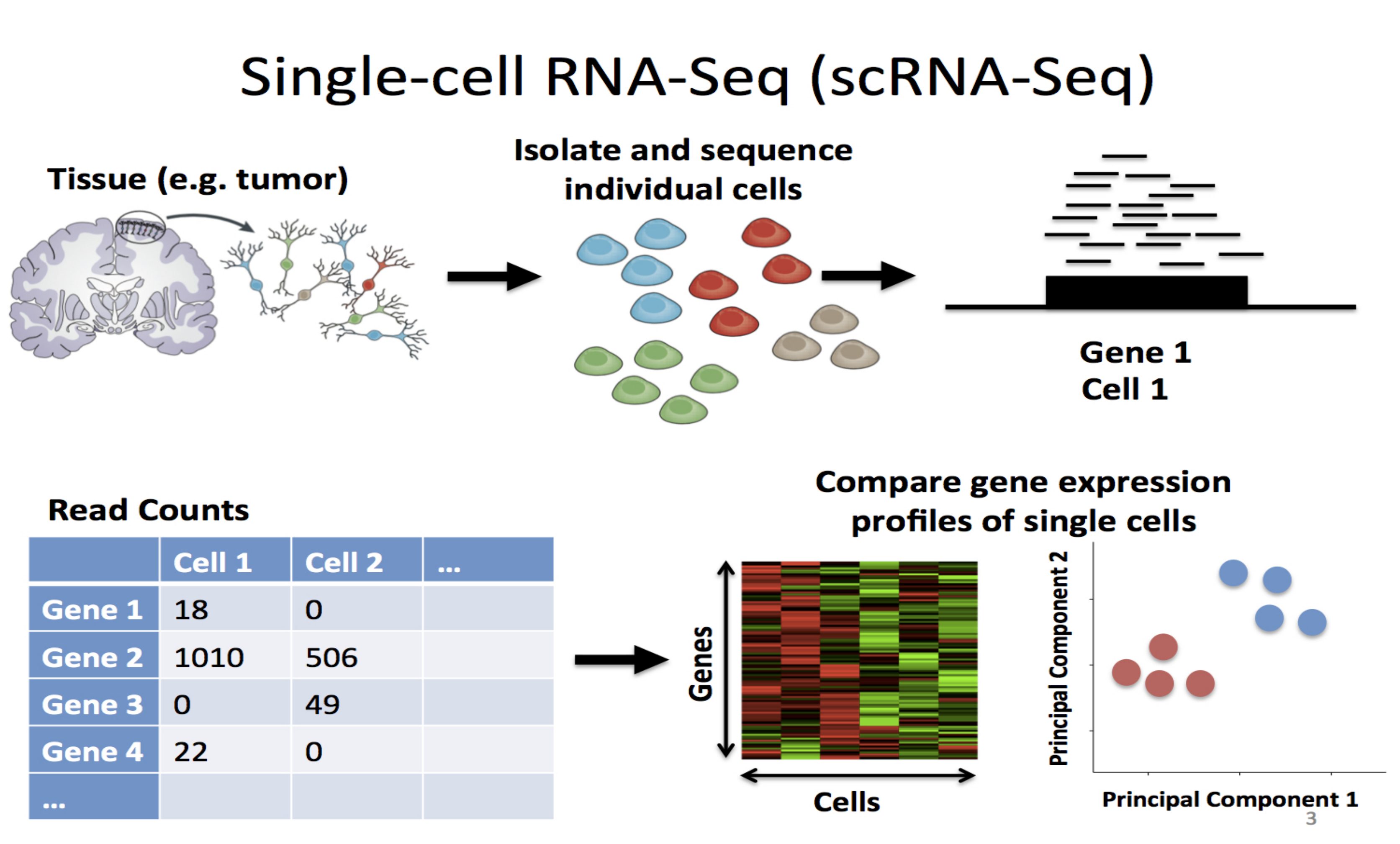
An evaluation of statistical differential analysis methods in single-cell RNA-Seq data

Dongmei Li, PhD (Dongmei_Li@urmc.rochester.edu)
Clinical and Translational Science Institute, University of Rochester School of Medicine and Dentistry, Rochester, NY 14642-0708

Introduction


Single-cell RNA-Seq (scRNA-Seq)
Tissue (e.g. tumor)
Isolate and sequence individual cells
Gene 1
Cell 1
Read Counts
Compare gene expression profiles of single cells
Genes
Cells
Principal Component 2
Principal Component 1

	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			

https://learn:gencore:bio:nyu:edu=single-cell-rnaseq/

Single-cell RNA-Seq is gaining popularity in recent years. Compared to bulk RNA-Seq, single-cell RNA-Seq allows the gene expression being measured within individual cells instead of mean gene expression levels across all cells. Thus, cell-to-cell variation of gene expressions could be examined. Gene differential expression analysis remains the major purpose in most Single-cell RNA-Seq experiments and many tools have been developed in recent years to conduct gene differential expression analysis for Single-cell RNA-Seq data.

Methods

Using simulation studies and real data examples, we evaluate the performance of five open-source popular methods for gene differential expression analysis.

- **DEsingle** (Zero-inflated negative binomial model)
- **Linnorm** (Empirical Bayes method on transformed count data using the limma package)
- **Monocle2** (Approximate Chi-Square likelihood ratio test)
- **MAST** (A generalized linear hurdle model)
- **DESeq2** (A generalized linear model with empirical Bayes approach)

Simulation Results

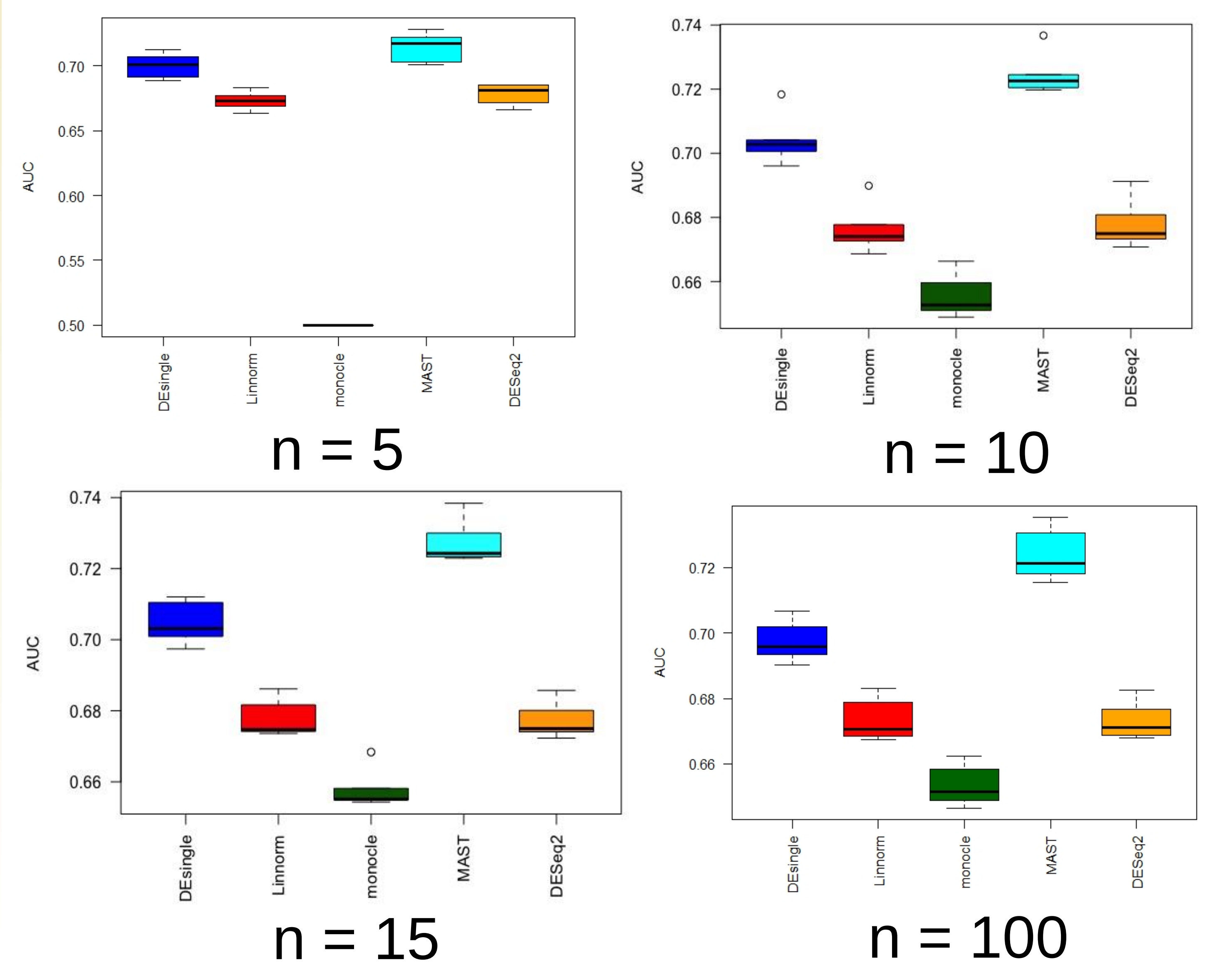


Figure 1: AUC of different RNA-Seq differential analysis methods with various sample sizes in each group from simulated data following Negative Binomial distribution with greater than 0 proportion of zero counts.

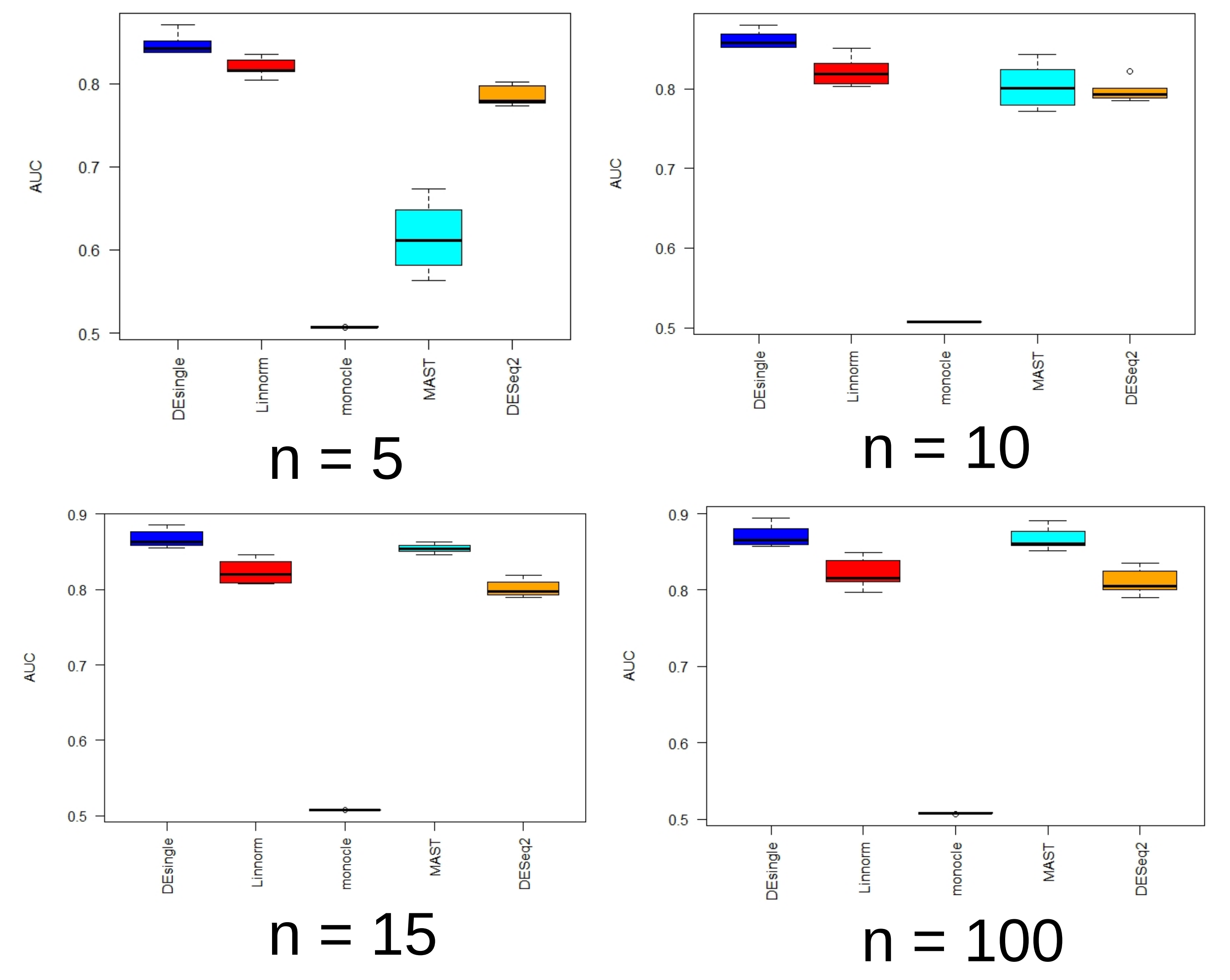


Figure 2: AUC of different RNA-Seq differential analysis methods with various sample sizes in each group from simulated data following Negative Binomial distribution with zero proportion of zero counts.

Real Data Example

- Single-cell RNA-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 five methods used to select significant differentially expressed genes from 14,905 genes.

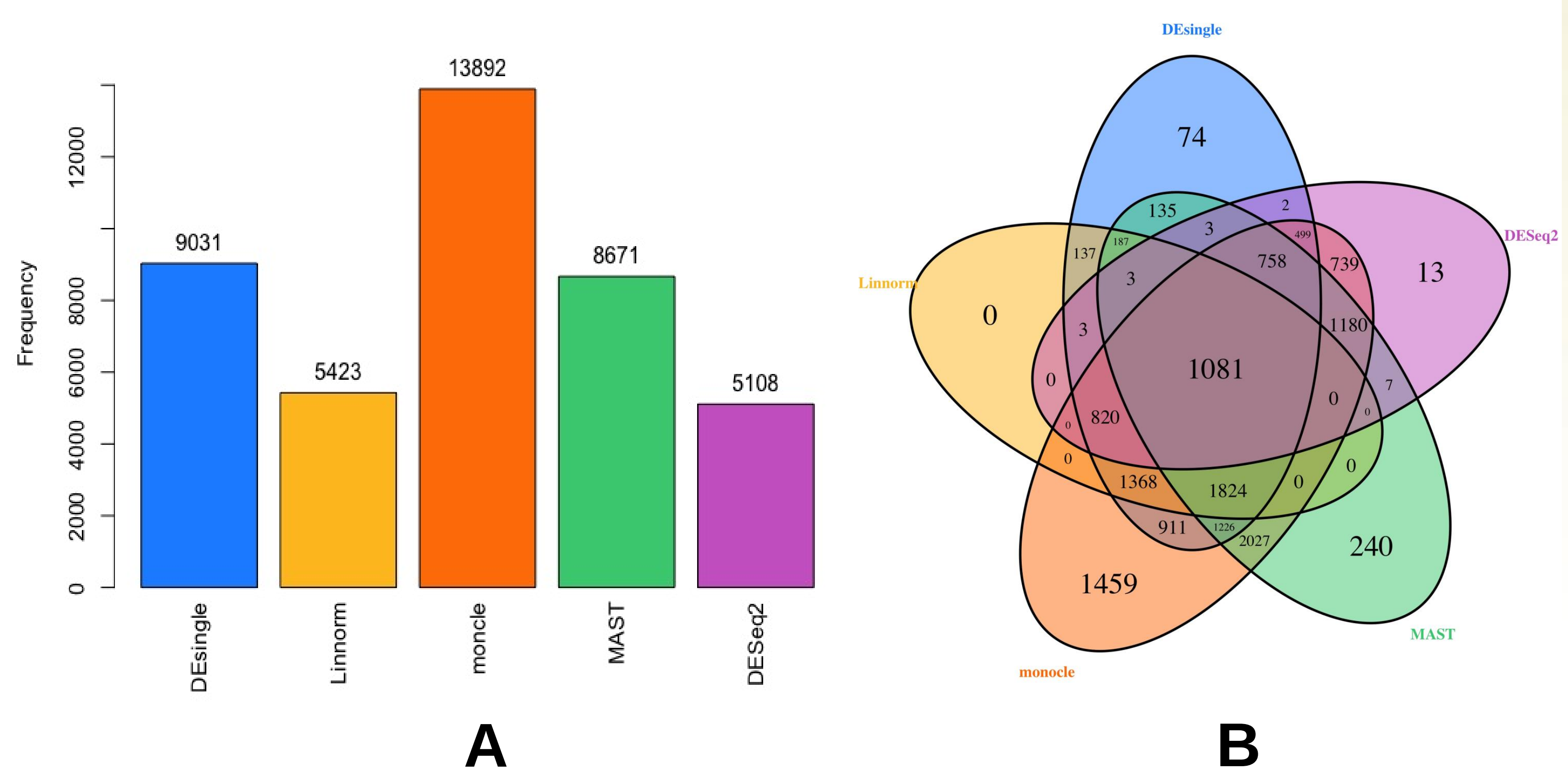


Figure 3: Empirical power (**A**) and venn diagram (**B**) of different single-cell RNA-Seq differential analysis methods using the real data example.

Conclusions

- MAST and Linnorm performs relatively better than other methods with higher AUC, when there are some proportion of zeros in the single-cell RNA-Seq data after filtering.
- DESingle, Linnorm, and DESeq2 performs relatively better than others with higher AUC when the proportions of zeros are close to zero.
- When sample size increases to 100 in each group, MAST shows the best performance with the highest AUC regardless of the proportion of zeros in the data.

Acknowledgements

- U54CA228110: WNY Center for Research on Flavored Tobacco Products (CRoFT) from FDA/NCI
- UL1 TR002001: CTSA award from NCATS/NIH