Detecting Differentially Expressed Genes along Single-cell Pseudotime through QGAM

Tianyang Liu[2], Huy Nguyen[1,2]

1 BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
2 The Junction of Statistics and Biology, UCLA

August 13th, 2021

Background:

- Analyzing a single cell makes it possible to discover molecular mechanisms behind cell state changes that are not seen when studying a bulk population of cells
- A crucial step is identifying differentially expressed (DE) genes along inferred single-cell pseudotime
- QGAM provides more flexibility by modelling the quantiles of conditional response distribution individually, thus avoiding any parametric assumption on the distribution of the response variable
- Single-cell expression data:
  - Dyntoy Simulator
  - Our group’s simulation

Methods from PseudotimeDE

The baseline model that describes the relationship between every gene's expression in a cell and the cell's pseudotime is the negative binomial–generalized additive model. For gene $j(j = 1, \ldots, m)$, its expression Y_{ij} in cell i and the pseudotime T_i of cell $i(i = 1, \ldots, n)$ are assumed to follow:

$$\begin{cases} Y_{ij} \sim NB(\mu_{ij}, \phi_j) \\ \log(\mu_{ij}) = \beta_{j0} + f_j(T_i) \end{cases}$$

where $NB(\mu_{ij}, \phi_j)$ denotes the negative binomial distribution with mean μ_{ij} and dispersion ϕ_j, and $f_j(T_i) = \sum_{k=1}^{K} b_k(T_i)\beta_{jk}$ is a cubic spline function.

To account for excess zeros in scRNA-seq data that may not be explained by the NB-GAM, we introduce a hidden variable Z_{ij} to indicate the "dropout" event of gene j in cell i, and the resulting model is called the zero-inflated negative binomial–generalized additive model (ZINB-GAM):

$$\begin{cases} Z_{ij} \sim Ber(p_{ij}) \\ Y_{ij} \mid Z_{ij} \sim Z_{ij} \cdot NB(\mu_{ij}, \phi_j) + (1 - Z_{ij}) \cdot 0 \\ \log(\mu_{ij}) = \beta_{j0} + f_j(T_i) \\ logit(p_{ij}) = \alpha_{j0} + \alpha_{j1} \log(\mu_{ij}) \end{cases}$$

Statistical test and p-value calculation:

To test if gene j is DE along cell pseudotime, PseudotimeDE defines the null and alternative hypotheses as:

$H_0 : f_j(\cdot) = 0$ vs. $H_1 : f_j(\cdot) \neq 0$

Test statistic is:

$S_j = \hat{f}_j^{\top} \hat{V}_{f_j}^{r-} \hat{f}_j$, where $\hat{f}_j = (f_j(T_1), \ldots, f_j(T_n))^{\top}$ and $\hat{V}_{f_j}^{r-}$ is the estimated covariance matrix.

Methods from QGAM

Additive structure $\mu(x) = \sum_{j=1}^{m} \sum_{k=1}^{r} \beta_{jk} b_{jk}(x_j)$, where β_{jk} are unknown coefficients and $b_{jk}(x_j)$ are known spline basis functions. $\mu(x_i)$ can also be express as $\mu(\mathrm{x}_i) = \mathbf{x}_i^{\top}\beta$. Let V be the covariance matrix for β, then $x^{\top}Vx$ becomes the variance of $\mu(x)$. If we take the analogue to PseudotimeDE for QGAM, let's set $\mu(x)_j$ as the additive structure for gene j. The model can be set as:

$$\begin{cases} Y_{ij} \sim NB(\mu_{ij}, \phi_j) \\ \log(\mu_{ij}) = \beta_{j0} + \mu_j(T_i) \end{cases}$$

Then The null and alternative hypotheses can be :

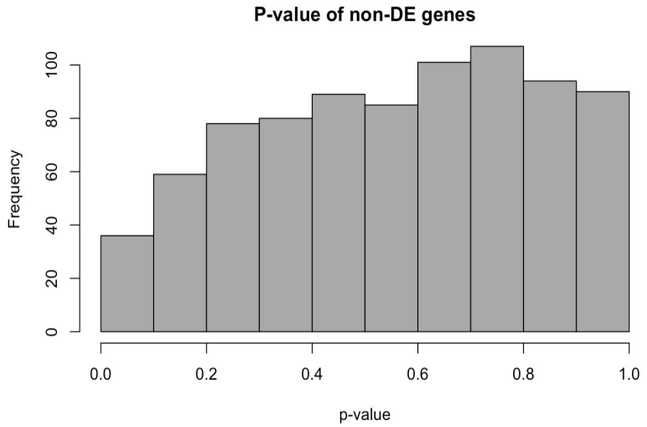
$H_0 : \mu_j(.) = 0$ v.s $H_1 : \mu_j(.) \neq 0$

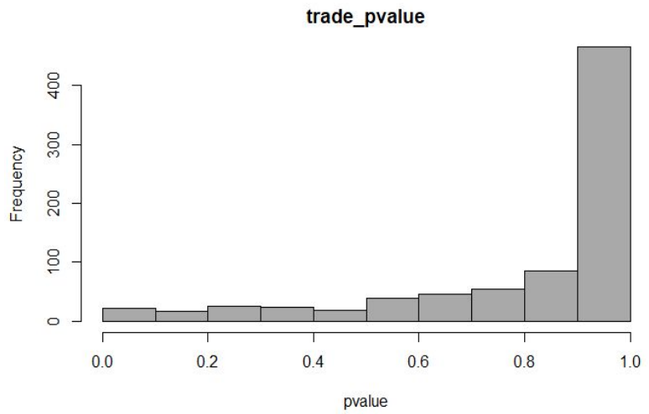
Data Simulation

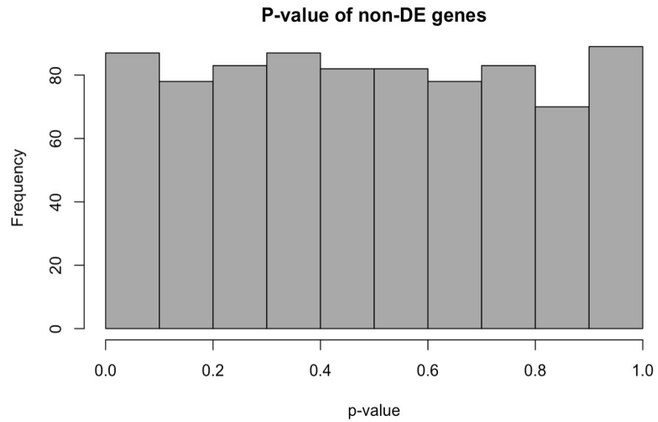
We simulated both single-lineage and bifurcation single-cell data. And we have low dispersion level and medium dispersion level for the single-lineage data. The bifurcation data is in low dispersion level. For one set of dataset, single-lineage low-dispersion data, single-lineage medium-dispersion data, and bifurcation data, we added 10 doublets as outliers to the data. Another set of dataset remains the same as the control group.

For the single-lineage data without outliers, we first generated 510 true time variable from uniform distribution. Then we generated 1000 genes from the negative binomial distribution. For the parameters of negative binomial distribution, μ is generated by $log_\mu = a + b * cos(t) + c * t^2 + d * sin(t^3) + 1$, where t is the previous true time, a is generated by uniform distribution from range of 0 1, b is generated by uniform distribution from range of 2 3, c is generated by uniform distribution from range of -2 2, and d s generated by uniform distribution from range of -4 4. The reason why we chose different range for those coefficients of terms in log_μ is that we want the genes to have different trend along the true time.

Dyntoy’s Medium Dispersion



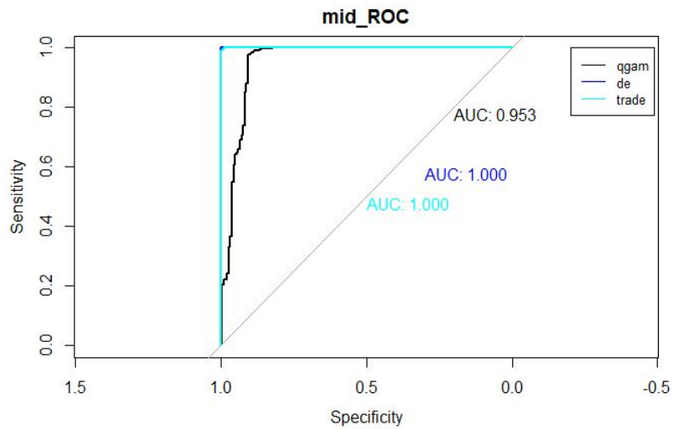


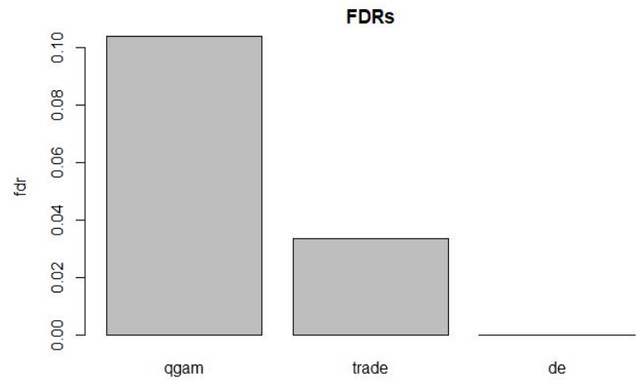


PseudotimeDE (No subsampling)

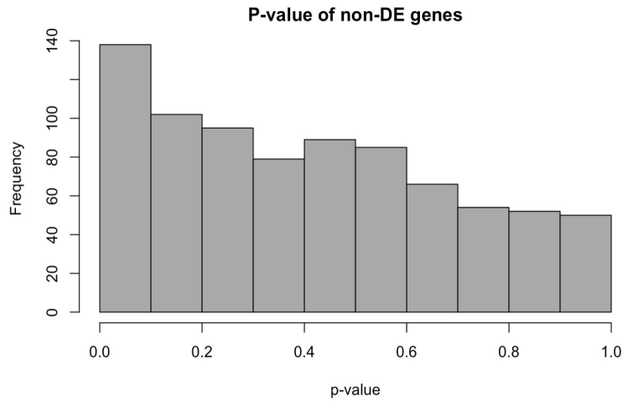
TradeSeq

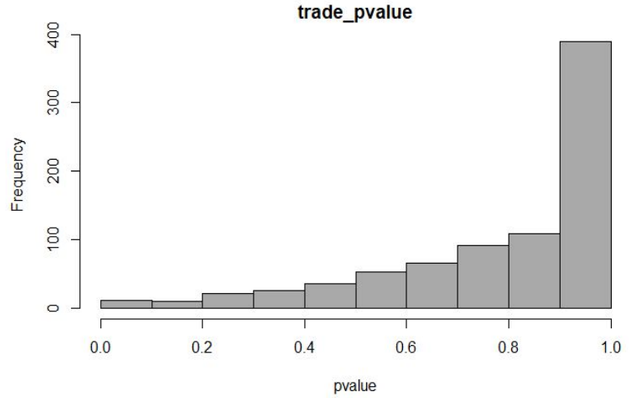
QGAM

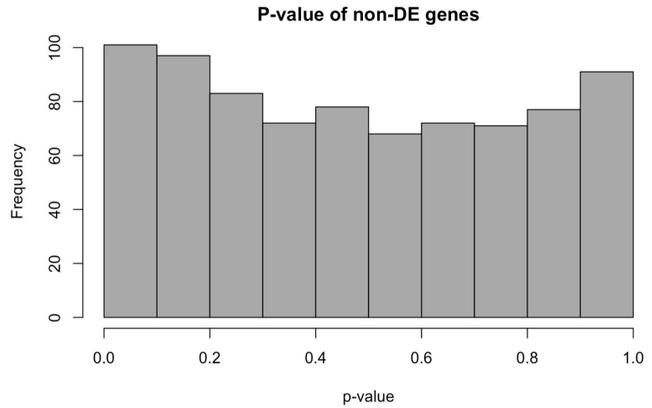




Dyntoy’s High Dispersion



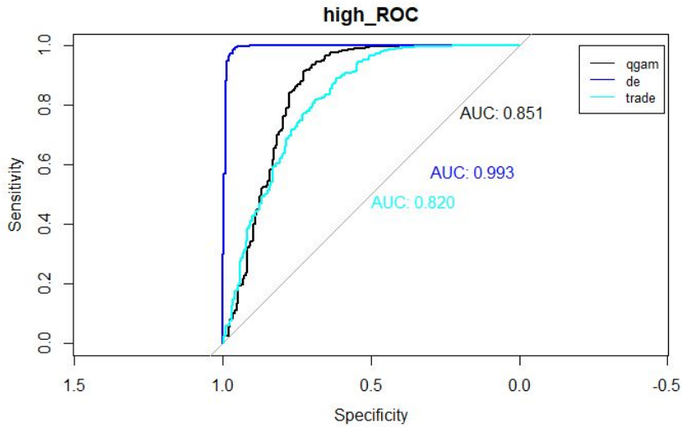


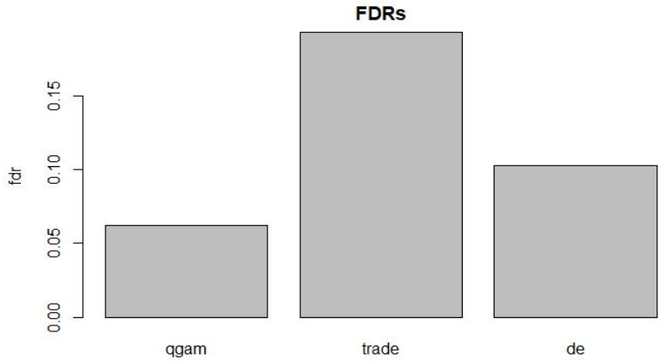


PseudotimeDE (No subsampling)

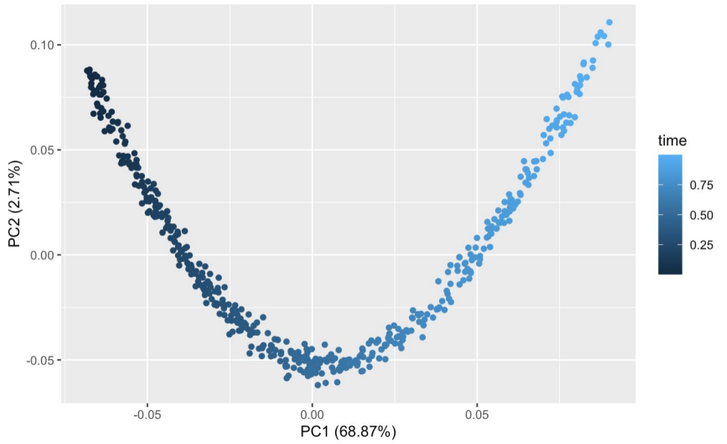
TradeSeq

QGAM

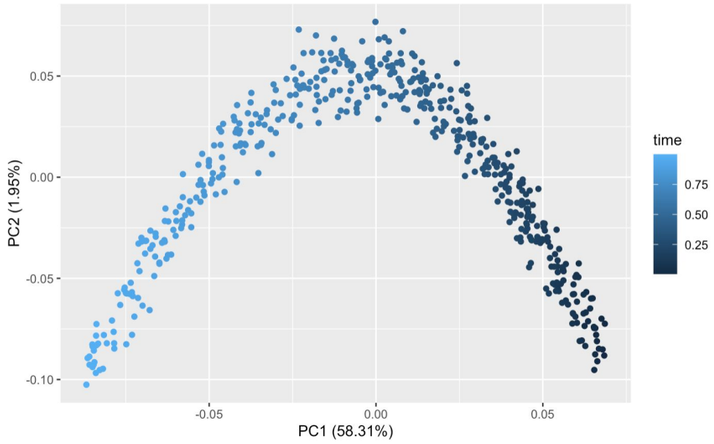




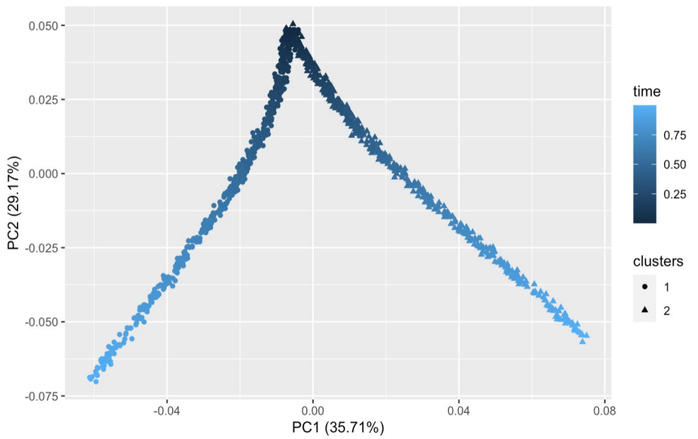
Our Simulations' PCA



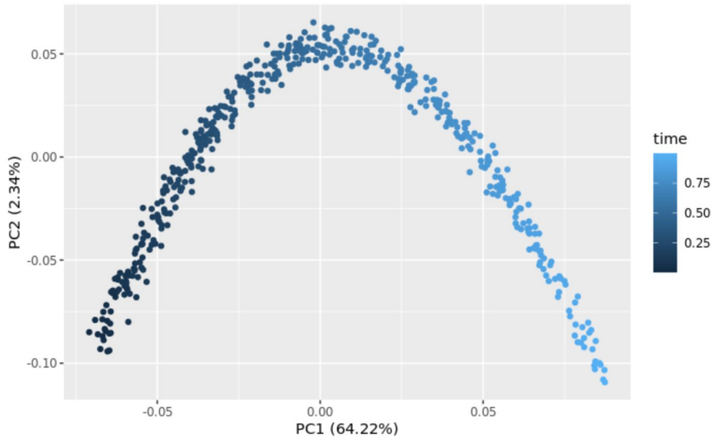
Low (No Outlier)



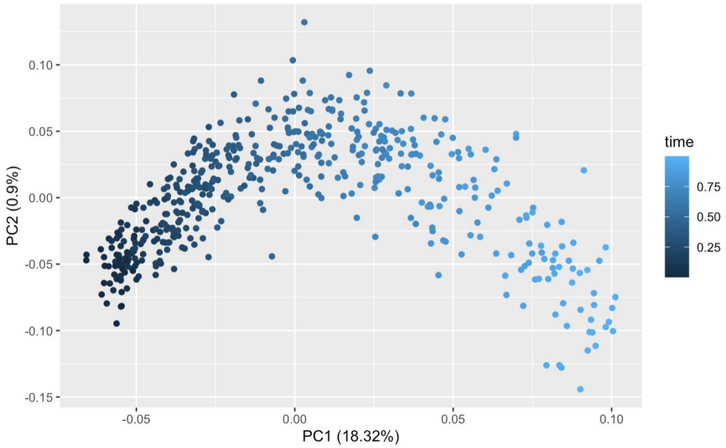
Medium (No Outlier)



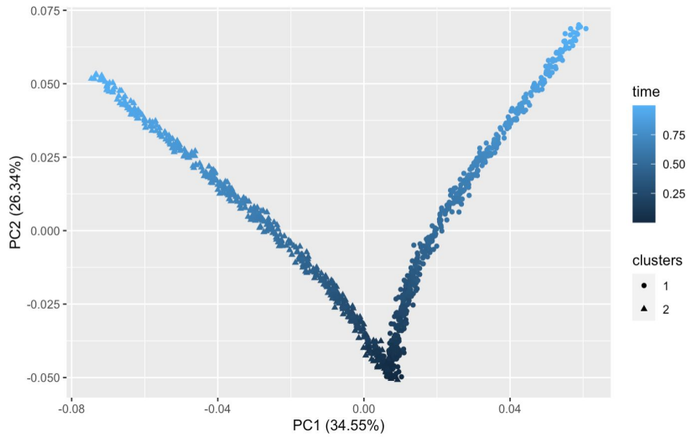
Bifurcation (No Outlier)



Low (With Outliers)

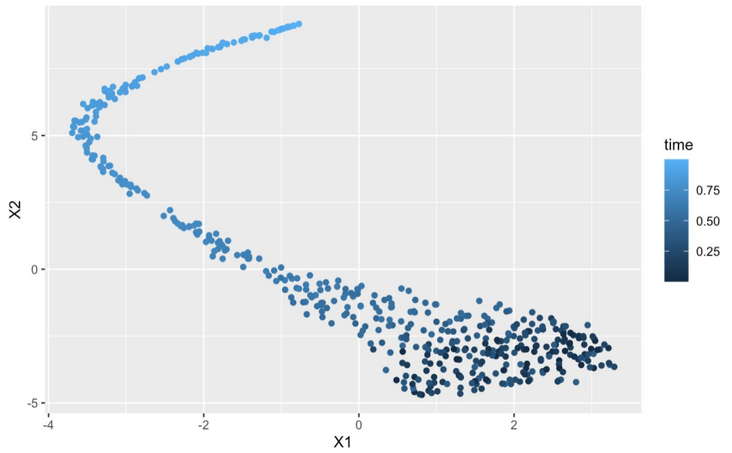


Medium (With Outliers)

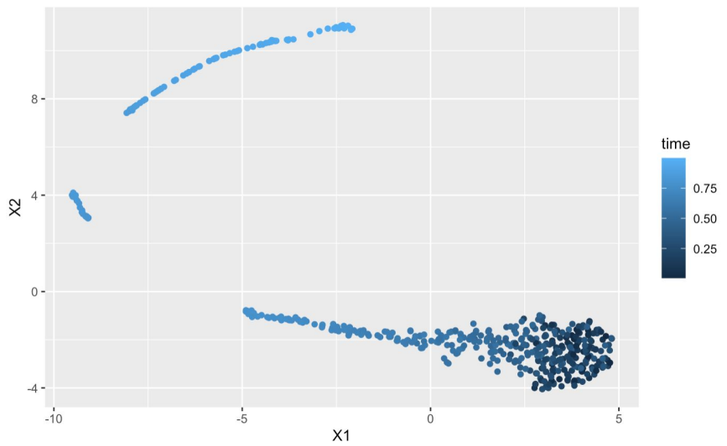


Bifurcation (With Outliers)

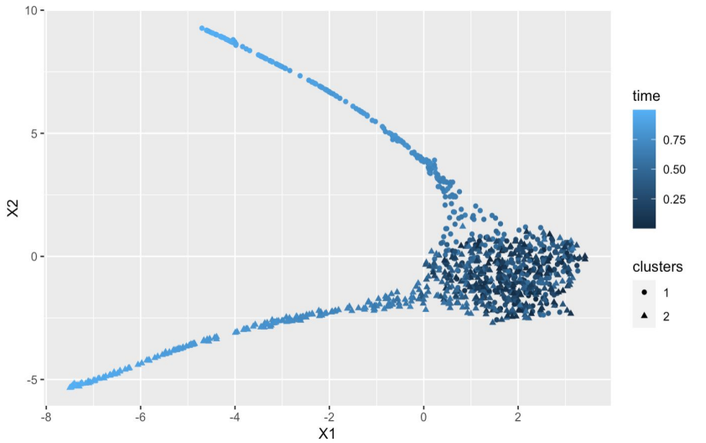
Our Simulations' UMAP



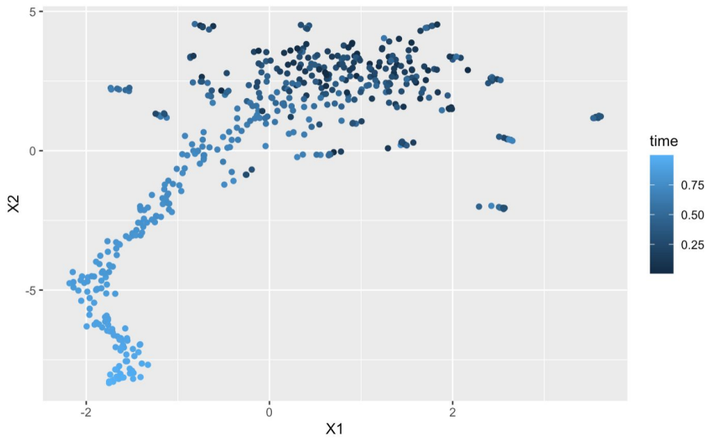
Low (No Outlier)



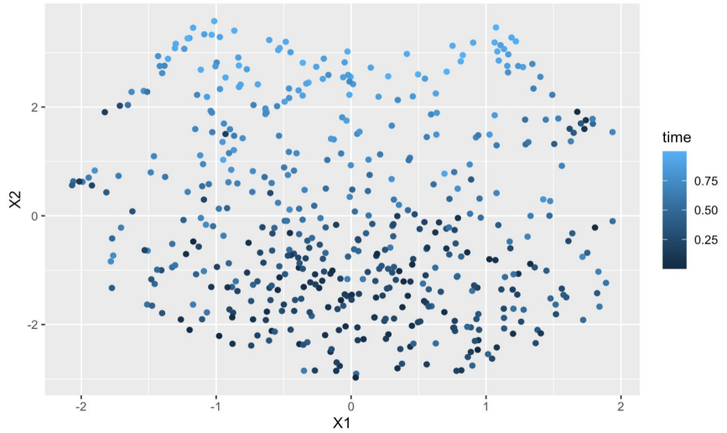
Medium (No Outlier)



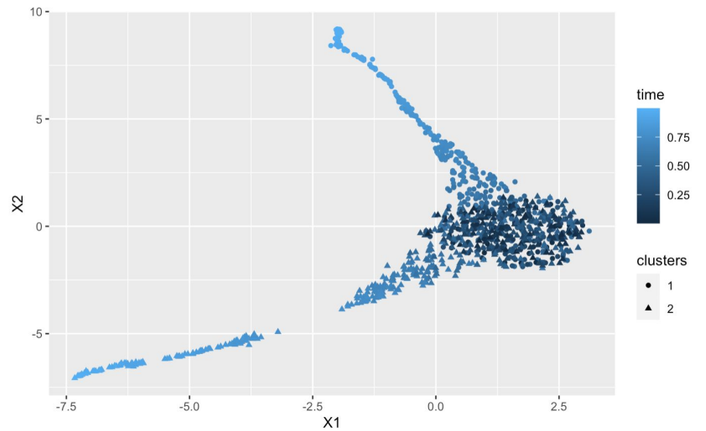
Bifurcation (No Outlier)



Low (With Outliers)



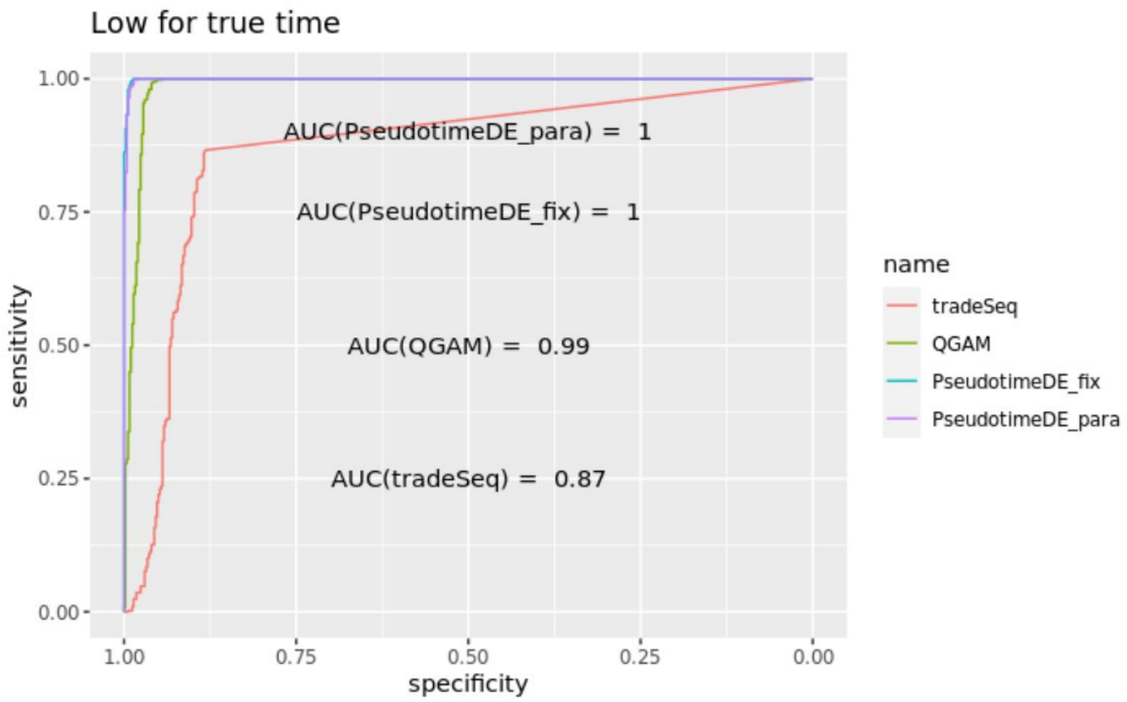
Medium (With Outliers)

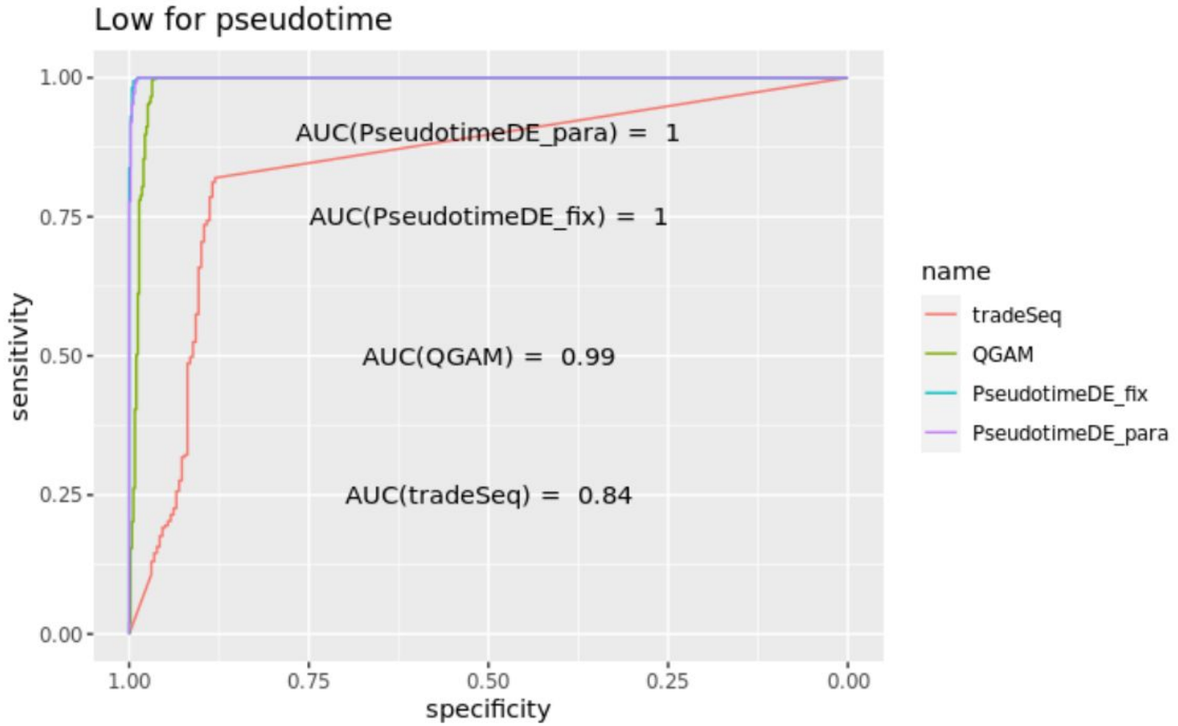


Bifurcation (With Outliers)

Our Simulations' ROC curves

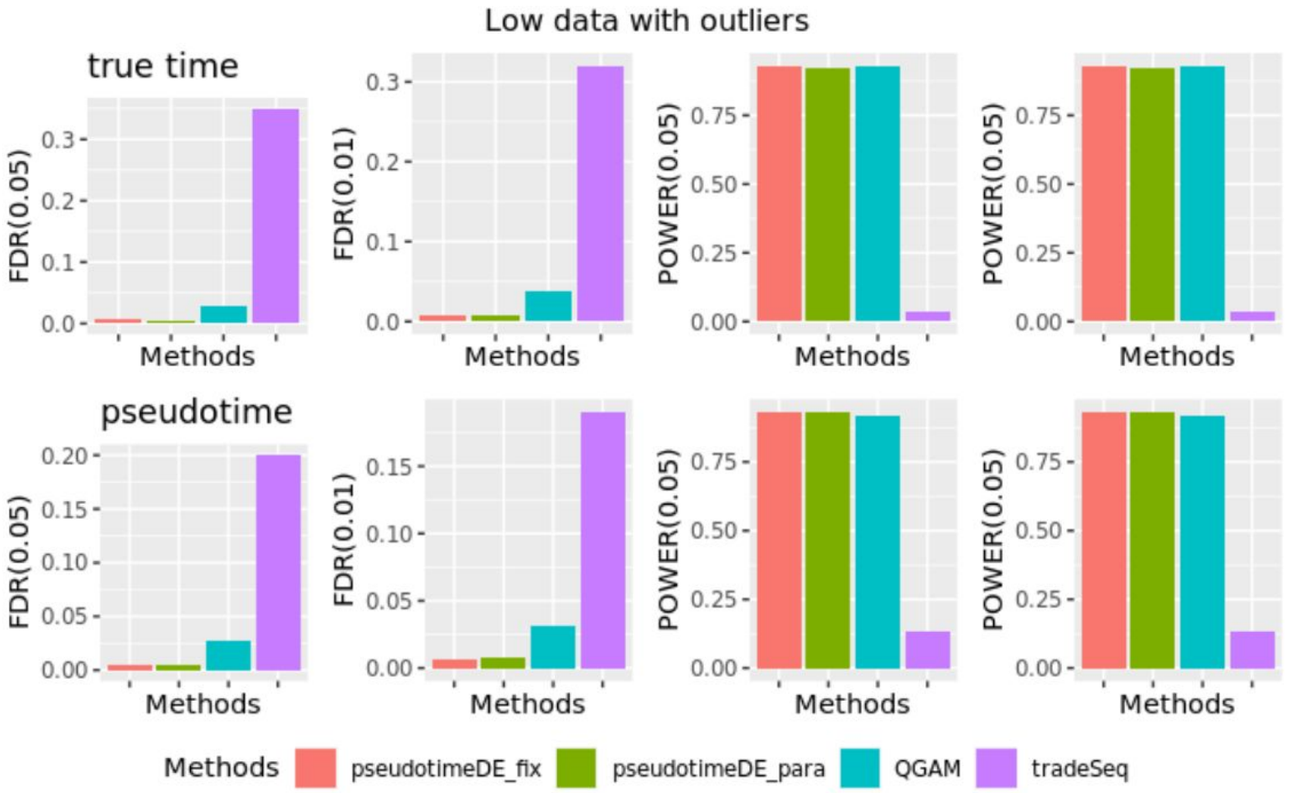
Subtitle



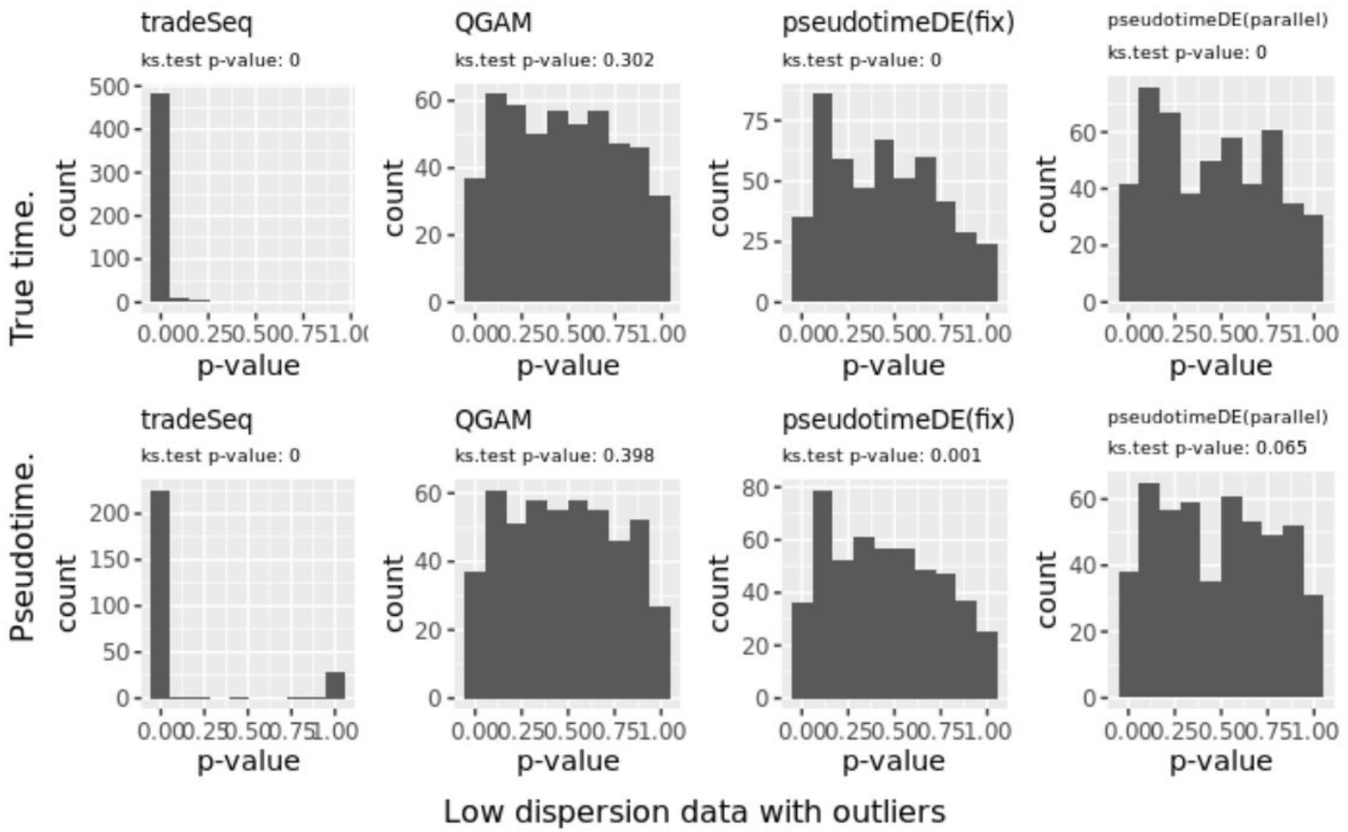


Our Simulations' FDR

Here

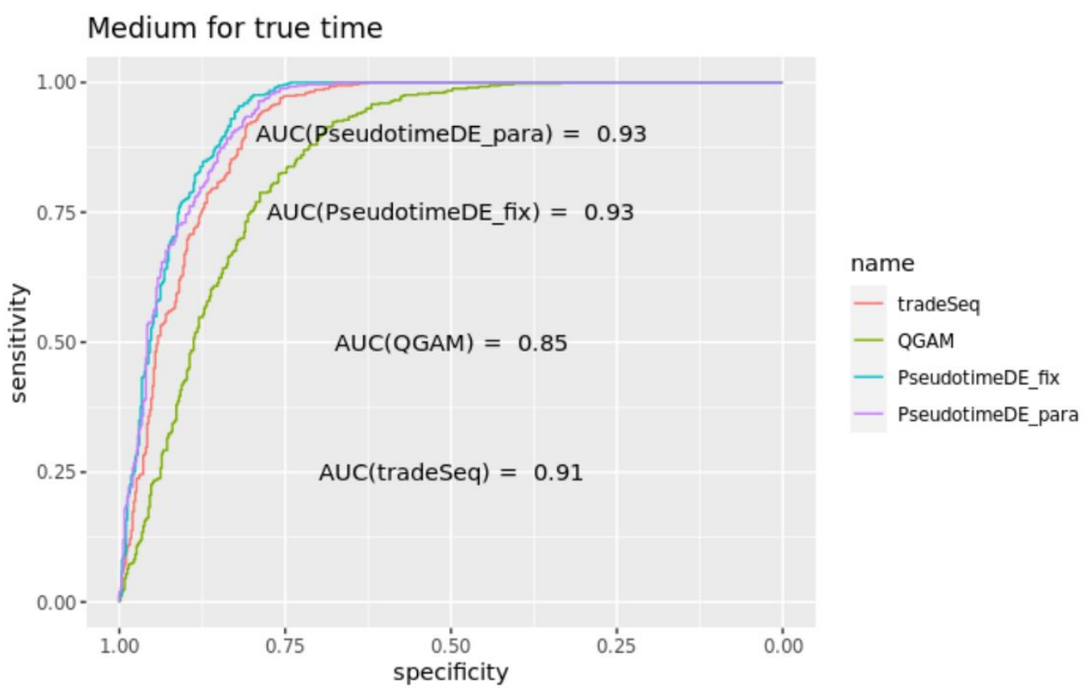


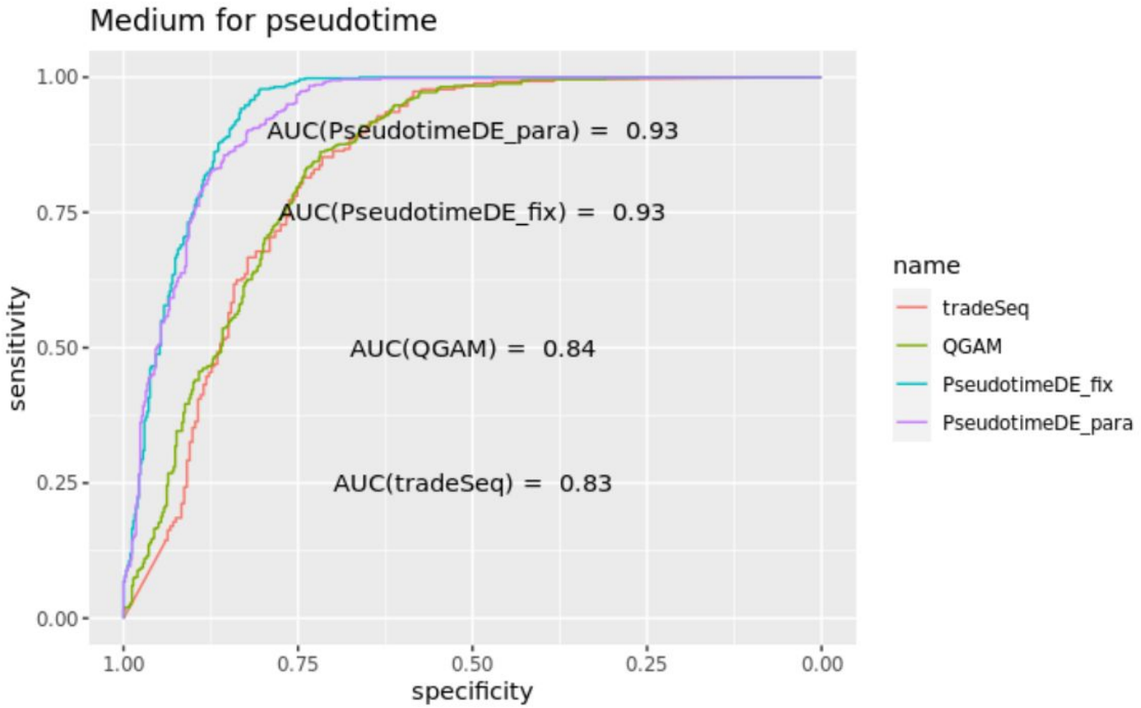
Our Simulations' Distribution of non-DE genes



Our Simulations' ROC curves

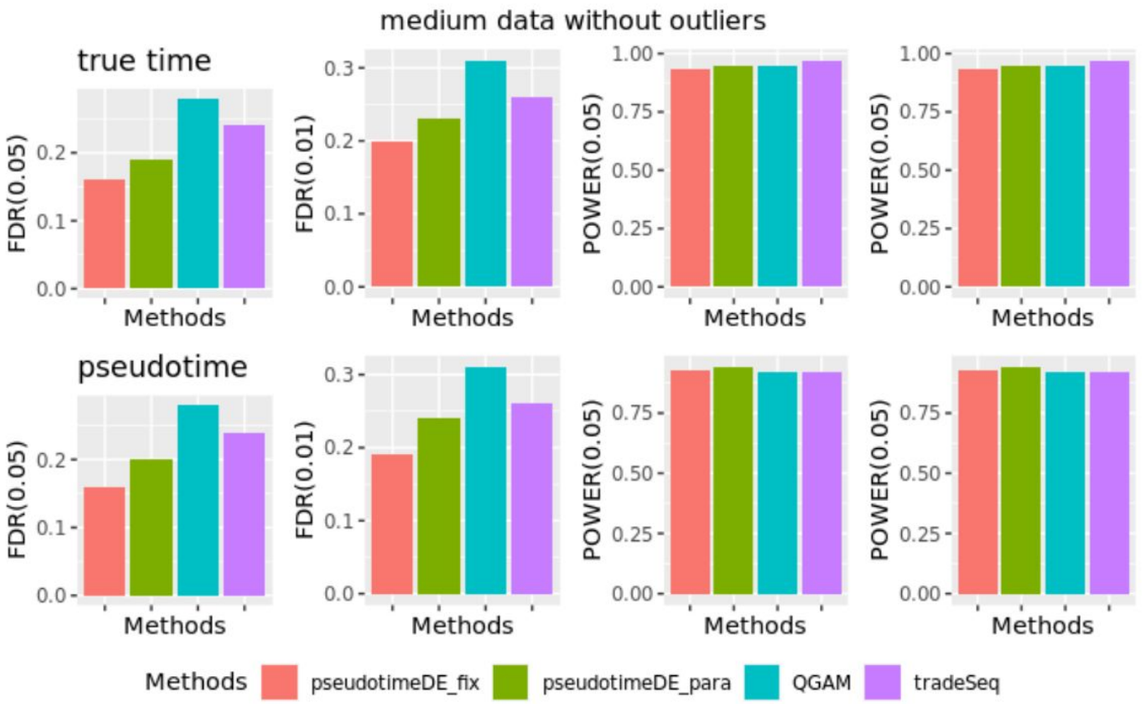
Subtitle

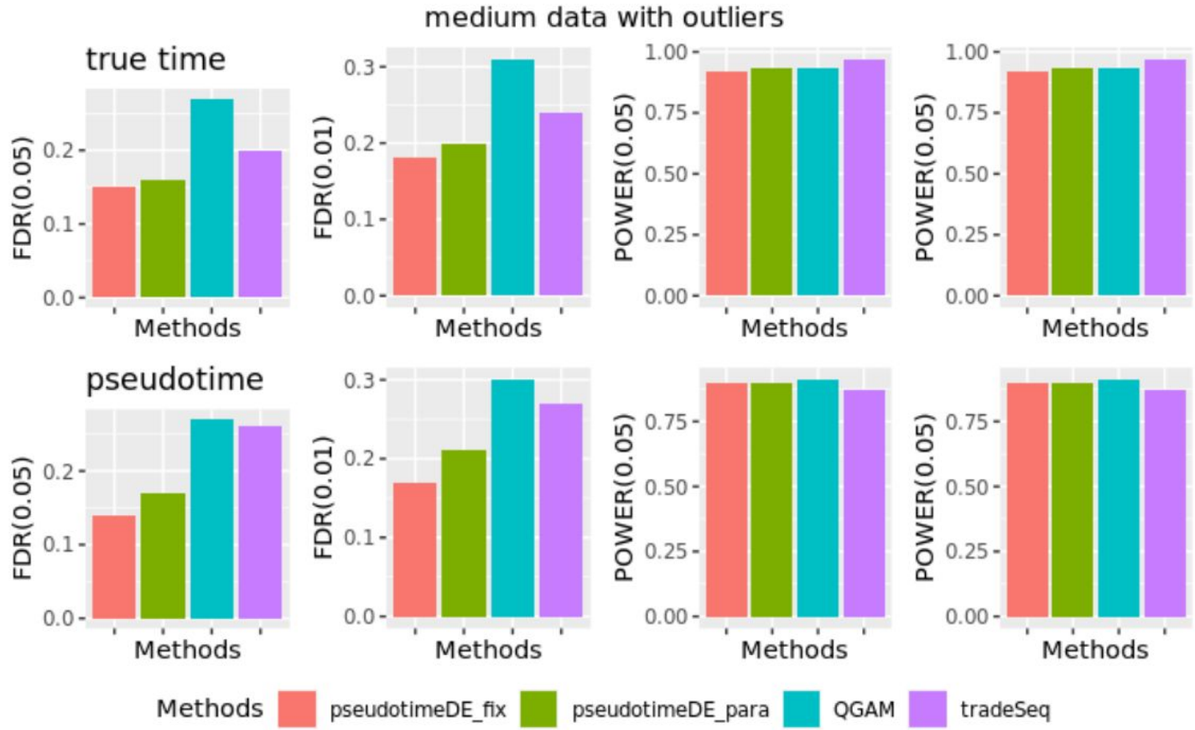




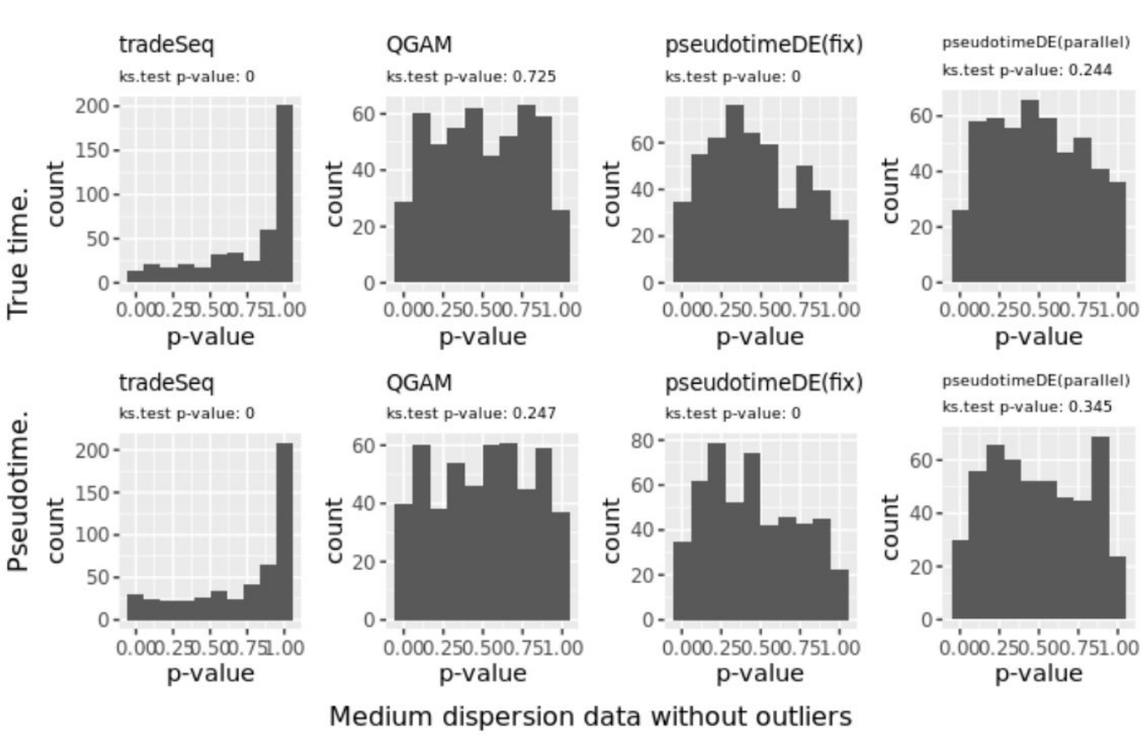
Our Simulations' FDR

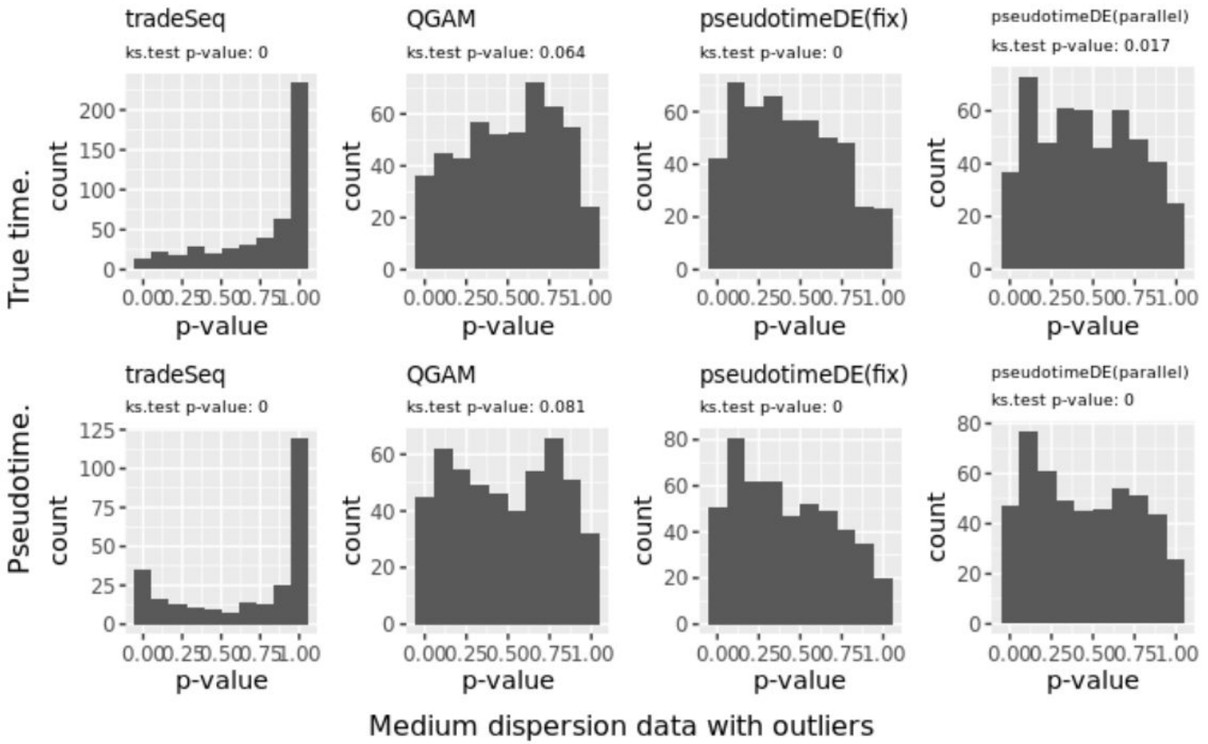
Since this is medium dispersion data, we also need to benchmark when the data has no outliers





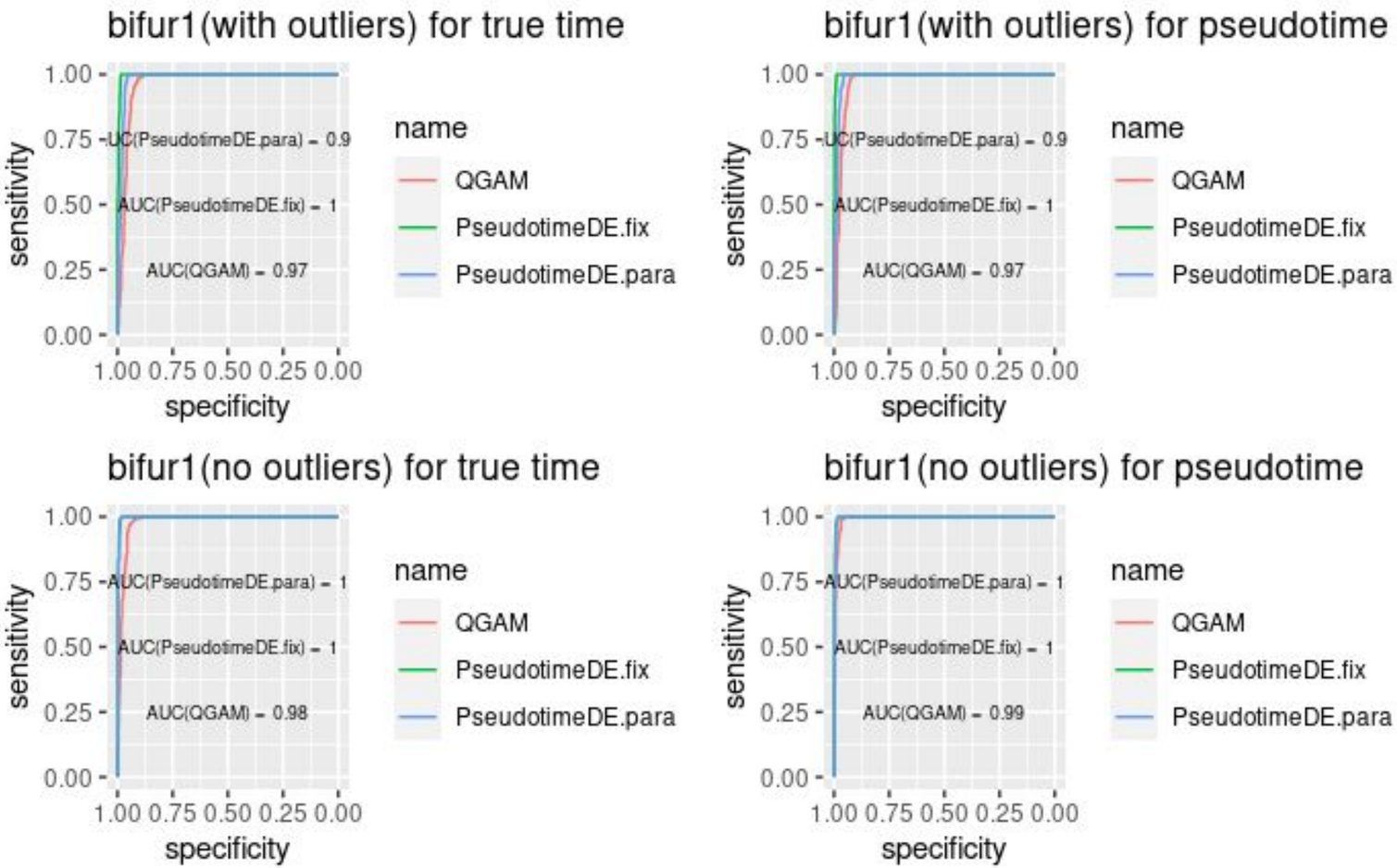
Our Simulations' Distribution of non-DE genes





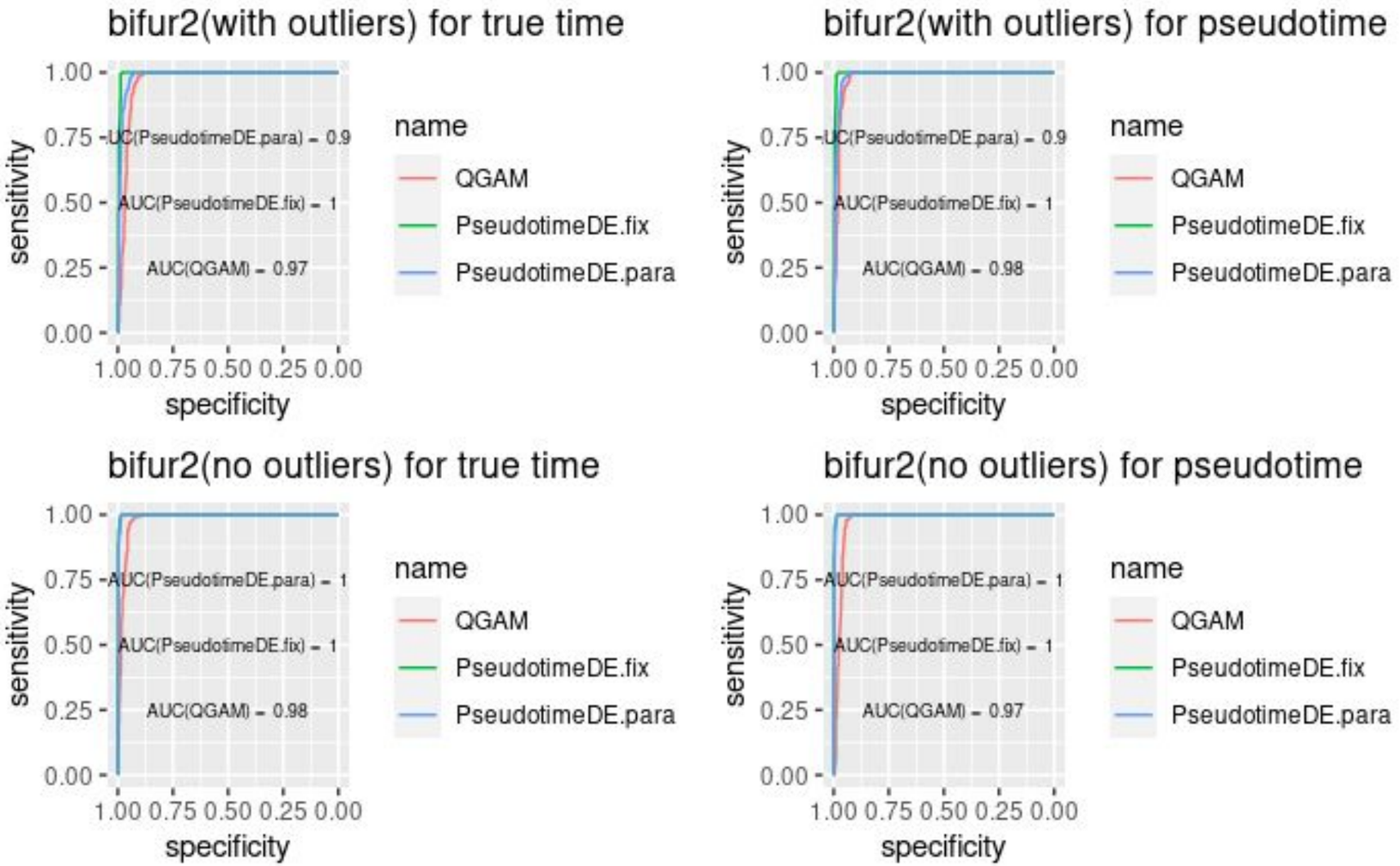
Our Simulations' ROC curves

Subtitle



Our Simulations' ROC curves

Subtitle



Conclusions

- QGAM is not always better than PseudotimeDE and TradeSeq
- QGAM has good distribution of non-DE genes' p-values
- QGAM has good AUC scores both in Dyntoy and our simulated data
- For bifurcation data, QGAM yields the best AUC.
- QGAM is relatively fast to yield results

Future Directions

- Improve our data simulation algorithms, especially to work better with TradeSeq
- Generate high and ultra-high dispersion data
- Improve bifurcation data so that a gene needs not be DE in both lineages
- Provide FDR Barplots, non-DE genes' Histograms for our bifurcation data (benchmarking)
- Begin our manuscript

Citations

Matteo Fasiolo, Simon N. Wood, Margaux Zaffran, Raphaël Nedellec, and Yannig Goude. "Fast Calibrated Additive Quantile Regression", Journal of the American Statistical Association, 7 Mar. 2020.

Dongyuan Song, and Jingyi Jessica Li. "PseudotimeDE: Inference of Differential Gene Expression along Cell Pseudotime with Well-Calibrated p-Values from Single-Cell RNA Sequencing Data." *Genome Biology*, BioMed Central, 29 Apr. 2021.

Acknowledgements

- PI: Dr. Jingyi Jessica Li
- Direct research mentor: Dongyuan Song
- B.I.G Summer Program
- Melody Zhang for her initial data transformations

“Develop a passion for learning. If you do, you will never cease to grow.”
– Anthony J. D’Angelo

Thank you!

Presenter: Huy Nguyen
The Junction of Statistics and Biology
University of California, Los Angeles
huynguyen012016@gmail.com
http://jsb.ucla.edu

UCLA
Statistics