Parametric versus rank-based, non-parametric methods for differential expression analysis in single-cell RNA sequencing data

Olympia Hardy

GUID: 2505621H

Supervisor: Dr. Pawel Herzyk

Course: MSc Bioinformatics

A report submitted in partial fulfilment of the requirements for the MSc Bioinformatics Degree at The University of Glasgow

August 2020

Summary

The progression of transcriptomics calls for the development of robust bioinformatic tools to process and analyse scRNA-seq data to obtain biologically meaningful results. Although there are similarities between the pipeline of bulk RNA-seq and scRNA-seq data, the downstream analyses and normalisation steps of the data differs substantially due to the variability, multimodality and sparsity of scRNA-deq data. Currently, there seems to be little agreement among the scientific community of which bioinformatics tools can accurately characterise scRNA-deq data resulting in the need for comparative studies of existing tools to evaluate their performance. This can be achieved through generating simulated scRNA-seq data where the 'truth' of the data is known and thus we can evaluate the results of existing analysis tools to see how accurately they perform. In this project we have divided into two main components. The first is the comparison and evaluation of four existing scRNA-seq data simulators to identify the optimal simulator to generate our simulated data for analysis. We will be examining SPARSim, scDesign, Splatter and SPsimSeq to assess how well they can mimic the characteristics of real droplet-based scRNA-seq data generated from the 10X protocol. Once we have assessed our optimal data simulator performance, we can move onto analysing the performance of methods designed to identify differentially expressed genes. Here we will focus on evaluating the performance of the non-parametric method RankStat, a derivative of RankProducts developed by Glasgow Polyomics compared to the performance of several other existing differential expression methods both parametric and non-parametric also. Considering we know the 'truth' of our simulated data we can then evaluate how well these methods perform with scRNA-seq data using several performance markers such as the precision, receiver operating characteristic, and false positive rate. Differential expression methods have successfully allowed for identification of differentially expressed genes in bulk RNA-seq data however we will focus on scRNA-seq data which comes with added challenges such as increased noise in the data, poor documentation of existing methods and the absence of reproducible code. We hope to not only evaluate the performance of RankStat against existing

differential expression methods as an exciting alternative to characterise scRNA-seq data, but also to provide reproducible documentation of the analysis pipeline facilitating accessibility of this newly emerged computational challenge that may revolutionise transcriptomic studies in the future.

Acknowledgements

I would like to thank Dr. Pawel Herzyk for his commitment and support shown throughout the completion of this project. Through times of adversity and remote communications his guidance has not faltered and has provided me with great insight into the subject at hand. I would also like to show my appreciation for him allowing me to be involved in this pilot study featuring a methodology he played a role in creating.

Table of Contents

Abbreviations

- Differential Expression/Expressed (DE)
- RNA sequencing (RNA-seq)
- Single-cell RNA sequencing (scRNA-seq)
- Peripheral blood mononuclear cell (PBMC)
- Transforming growth factor beta-1 (TGFB1)
- Receiver Operating Characteristics (ROC)
- Area under the curve (AUC)
- False Discovery Rate (FDR)
- True Positive Rate (TPR)
- Maximum a posteriori (MAP)
- Maximum likelihood estimation (MLE)
- Zero-inflated negative binomial (ZINB)
- Cell detection rate (CDR)
- Earth movers distance (EMD)
- Median absolute deviation (MAD)
- Rank Products (RP)
- Rank Sum (RS)
- Rank distance (RD)
- Reverse-rank distance (RRD)
- Difference in Rank product (DRP)

Introduction

The field of transcriptomics has provided valuable biological insight into gene expression patterns through the application of bulk-RNA sequencing (RNA-seq) techniques. Advancement in methodologies has given rise to the emergence of single-cell RNA sequencing (scRNA-seq) that allows us to explore gene expression at a singular cell level. Unlike bulk-RNA sequencing that generates a gene expression profile for each gene by averaging expression levels across a large population of cells in a sample, scRNA-seq allows us to understand the variability of cell-to-cell gene expression(1). Evident advantages of scRNA-seq allow a more comprehensive expression profile of heterogenous samples with divergent transcriptomes, spatiotemporal characterisation of cells including the identification of novel or rare cell types and better representation of the inherent stochasticity of gene expression(2). However, key characteristics of scRNA-seq data make downstream computational analyses prove difficult. The first issue is the sparsity of the data due to 'drop out' events which can be caused by either technical causes where the transcript has not been captured or biological reasons such as transcriptional bursting where temporal fluctuation of transcript levels leads to a high number of zero values in the data(3). This phenomenon also contributes to the biological variability expected when analysing heterogenous samples of cells leading to variability in counts where different cells are expressing different genes. Finally, the distribution of read counts are challenging when analysing scRNA-seq data. When dealing with a heterogenous sample there may be multimodality in the distribution of transcripts due to different cells expressing different genes(3). As a result, there have been many developments of bioinformatics tools to facilitate scRNA-seq analysis however many share little agreement and lack clarity in the documentation that may be confusing to those that are unfamiliar with complex bioinformatic pipelines therefore limiting the accessibility of scRNA-seq data. For example, there are many parametric tools that attempt to model scRNA-seq data usually adopting a two-part model to reduce sensitivity towards the sparse nature of the data and complex multimodality to calculate gene expression. Alternatively, there are also non-parametric methods that can better capture multimodality in scRNA-seq data as there is no need for parameter estimation of the model resulting in lower true positive rates than parametric methods(4). Few benchmarking studies have been carried out to test the efficacy and validity of these packages in what the claim to achieve. One way this can be done is with the use of data simulators where we can use a real biological dataset as an input to create a synthetic dataset where a known truth is defined to evaluate the performance of these analysis packages. Unfortunately, many data simulator packages share the same issue where they are usually poorly documented and hard to implement. Therefore we propose a two-stage study where four current scRNA-seq data simulators (Splat(5), scDesign(6), SPARSim(7) and SPsimSeq(8)) will be evaluated using scRNA-seq data generated by the 10X Chromium protocol(9). The 10X Chromium approach is a popular droplet-based technique that boasts high-throughput and low costs when compared to previous well-based methods(10) however is substantially more noisy and subsequently harder to handle in downstream analyses. The outperforming package will then generate a synthetic dataset that will test the performance of RankStat(11), a non-parametric set of differential expression analysis methods against popular differential analysis methods that are widely used by bioinformaticians in scRNA-seq data analysis pipelines.

Methods

Experimental Design

The study performed can be broadly categorised into two distinct sections; the evaluation of data simulator packages where the optimal package was selected and used to then investigate the efficacy of differential expression packages. The seven datasets used in this study was scRNA-seq data obtained from the 10X Chromium protocol, four of which are freely available to download from the 10X Genomics website (10X Brain, 10X PBMC, 10X T-cell, and 293T-cell/Jurkat(9)) and the remainder were taken with the permission of the Glasgow Polyomics Institute (Sheep, Drosophila and human TFGB1). For this study the independent datasets were selected exclusively as they were generated by the 10X droplet-based protocol, additional aspects of the data such as cell type and sample size were irrelevant. The various data simulator packages were then compared against the real datasets and tested in seven different metrics. The dataset selected to produce the synthetic count matrix to examine the differential expression packages was the Drosophila dataset where cells were simulated in a two-condition framework to observe which genes are differentially expressed across the two conditions. To investigate the effect of sample size on the performance of the DE packages four separate simulations were carried out with 25, 50, 100, and 200 cells per condition. In each simulation 10% of the total genes were differentially expressed with equal proportions being up- and down-regulated at a p-value threshold of <0.05. As we generated a synthetic dataset there were no necessary read-alignment steps, all pre-processing was the initial filtering of genes that showed zero gene expression across all cells and the mandatory normalisation of input data required by various packages. The performance of the differential expression packages was benchmarked using six statistical metrics explained in detail below. The results were also compared against one another to observe similarities between packages such as the ranking methods that each package implements and the overlap of differentially expressed genes returned.

Data Simulators

scDesign

Generation of a synthetic dataset requires four steps in the scDesign framework. The first is the estimation of parameters from the real data whereby a mixed Gamma-Normal model calculated library size, the mean and standard deviation of gene expression. Next each gene expression value is simulated independently using the previous estimated parameters. Following this dropout events are introduced through sampling without replacement from dropout events that occur in the real dataset. Finally, the synthetic count matrix is produced adjusting for sequencing depth from a multinomial distribution(6).

SPARSim

The SPARSim package uses a Gamma-Multivariate Hypergeometric model to first model the biological variability between the cells in the data using a gamma distribution, then using a multivariate hypergeometric distribution to account for technical variability that may be present from experimental procedures such as the sequencing process. In this study, the parameters required by SPARSim were estimated from the raw count table of the real data which are then used to produce the synthetic dataset(7).

Splatter

Splatter carries out the simulation in two main steps; first the package estimates parameters from the real input dataset, then it uses these estimated parameters to generate the simulated dataset. The Splatter package contains five wrapper functions for various simulation packages however this study implements the core simulation model from the package Splat. The Splat simulation model uses a Gamma-Poisson hierarchical model where using a gamma distribution the mean expression for each gene is simulated, followed by the count for each cell being sampled from a Poisson distribution. To capture features typical of scRNA-seq data Splat estimates high expression gene outliers, varying library sizes between cells and zero-inflation to account for technical dropout. The Splatter package also contains a convenient functionality to compare the results of simulator packages that was implemented in the evaluation of the packages used in this study(5).

SPsimSeq

This package implements a semiparametric model the fast log-linear model-based density estimation method that combines both parametric gamma distribution and non-parametric kernel density estimates to generate gene-wise distribution. This is followed by the addition of zero counts to reflect scRNA-seq data based on the mean expression and library size relationship from the real data(8).

Differential Expression Packages

DESeq2

The parametric method DESeq2, mainly used in bulk RNA-seq analyses, uses a generalised linear model where gene counts are sampled from a negative binomial distribution. It also estimates a dispersion parameter using a gene-specific shrinkage estimation using average expression values. Here we also implemented a zero-centred Normal prior to provide maximum *a posteriori* (MAP) estimates to more accurately capture the true dispersion of the data(12).

DEsingle

DEsingle is a parametric method designed to analyse scRNA-seq data using a zero-inflated negative binomial (ZINB) model to factor in proportions of real zero counts and dropout estimates in the data. It then uses an integrated framework from DESeq to normalise the input count matrix, and then uses maximum likelihood estimations (MLE) to detect differentially expressed genes which are subsequently sorted into three distinct categories(13).

Model-based Analysis of Single-cell Transcriptomics (MAST)

The MAST package analyses scRNA-seq data in a two-part generalised linear model which takes into account an estimated cell detection rate (CDR) as a co-variate derived from the data. The CDR is defined as the fraction of genes expressed in each cell and corrects for technical and biological variability in the data. The first step in the hurdle model is using a logistic regression to model expression rate of a given gene followed by the second where a positive expression mean is modelled based on a Gaussian linear model(14).

SigEMD

SigEMD is the only non-parametric method outside of the RankStat package and uses Earth Mover's Distance (EMD) to measure the distances between distribution of expression for genes in the data. Instead of computing the large amount of zero values common in scRNA-seq data it instead uses a logistic regression to model them in addition to a Wald test to then decide whether the factor should be considered in the EMD calculation. Where this is true the package uses a Lasso regression model to impute zero values based on the expression values of similar genes in the data(4).

RankStat Methods

The RankStat package consists of five non-parametric methodologies that calculates the log-fold change for each gene comparatively which are then sorted and ranked in descending order.

The first method is RankProducts (RP) which is the product of all the ranks that takes into account whether the gene is up- or down-regulated by conducting two one-sided tests. Given a gene g is upregulated the RP statistic is calculated as follows:

$$RP_g^{up} = \prod_{i=1}^{k} r_{g,i}$$

Where k represents the number of all ranks in the data and $r_{g,i}$ is denoted as the rank of gene g in every i-th comparison. Similarly, RP calculates down-regulated genes using the formula below where n marks the number of all the genes in the data:

$$RP_g^{down} = \prod_{i=1}^{k} (n + 1 - r_{g,i})$$

The next method RankSums (RS) is similar to RP however it is the sum of all the ranks with the hypothesis denoted as:

$$RS_g^{up} = \sum_{i=1}^{k} r_{g,i}$$

$$RS_g^{down} = \sum_{i=1}^{k} (n + 1 - r_{g,i})$$

The third method is Difference of Rank Products (DRP) which builds on the assumption that if any given gene g is not differentially expressed then the RP statistic of the gene is approximately equal to that of the RP statistic when ranking was performed in the opposite direction. Therefore, the difference in RP_g^{up} and RP_g^{down} will be larger if gene g is indeed differentially expressed in the data:

$$DRP_g^{up} = \frac{\prod_{i=1}^{k} r_{g,i}}{\prod_{i=1}^{k} (n + 1 - r_{g,i})} = \prod_{i=1}^{k} \frac{r_{g,i}}{n + 1 - r_{g,i}}$$

$$DRP_g^{down} = \frac{\prod_{i=1}^{k} (n + 1 - r_{g,i})}{\prod_{i=1}^{k} r_{g,i}} = \prod_{i=1}^{k} \frac{n + 1 - r_{g,i}}{r_{g,i}}$$

The last two methods in the RankStat package are Rank Distances (RD) and Reverse Rank Distances (RRD) and are based on the calculation of Euclidean distance. The hypothesis is identical to RS where the ranks are summed however here they are raised to the power of two:

$$RD_g^{up} = \sum_{i\,=\,1}^{k} r_{g,i}^2$$

$$RD_g^{down} = \sum_{i\,=\,1}^{k} (n - r_{g,i} + 1)^2$$

$$RRD_g^{up} = \sum_{i\,=\,1}^{k} (n - r_{g,i} + 1)^2$$

$$RRD_g^{down} = \sum_{i\,=\,1}^{k} r_{g,i}^2$$

Here the smaller RD statistic the more likely gene g will be differentially expressed, conversely the RRD statistic must be higher in order to suggest there has been a significant change in gene expression.

Evaluation Metrics

The results of the DE analysis packages will be evaluated looking at the following metrics. The receiver receiving operating characteristic (ROC) curve is a probability curve that shows how well something distinguishes one thing from another. The area under the curve (AUC) is a numerical value indicating the effectiveness of separability with a value of 1 being the optimum(15) i.e. ability of the package to distinguish genes to be DE or non-DE. The next statistic is the false discovery rate (FDR) which is at a threshold of 0.05 so the ability of a package to control falsely identified DE genes at 5%. The true positive rate (TPR/recall) is defined as the ability of a package to correctly identify true DE genes i.e. the fraction of truth. Precision is the indicator of how much of our discovered truth is in fact true and finally the F1 score is a balance of recall and precision where both are equally weighted. How these are calculated can be seen in Figure X in the Results section.

Results

Using the Splatter package seven different parameters of the scRNA-seq datasets were evaluated. The parameters from the real dataset are estimated then compared against the estimated parameters from the four simulated datasets. The first parameter is the distribution of the mean gene expression values across the dataset. The next is the distribution of gene variance that reflects the variability in the dataset that could be down to either biological or technical factors. Library size is also evaluated that refers to the total count across all genes for each cell. Another metric is the mean-variance relationship that refers to the fact that lowly expressed genes are more variable and the converse for highly expressed genes. To evaluate the presence of zeros in the data we look at sparsity by gene characterised by the percentage of genes that have zero counts in the data. Similarly, sparsity by cell is the evaluation of the percentage of zero values for each cell in the data. Lastly, we look at the mean-sparsity relationship to characterise the effect of zero values in the data on the mean gene expression.

Here we present the results from the Drosophila 10X dataset including the estimated parameters of the real data and parameters from each synthetic dataset produced from four separate data simulations; the results for the additional datasets can be found in the Appendix. The comparison plots include the real data and we can compare the similarities of the synthetic datasets from each simulator package. On the other hand, the difference plots allow us to observe differences in how each package could recapitulate the parameters relative to the real data. These are shown below in Figure 1.

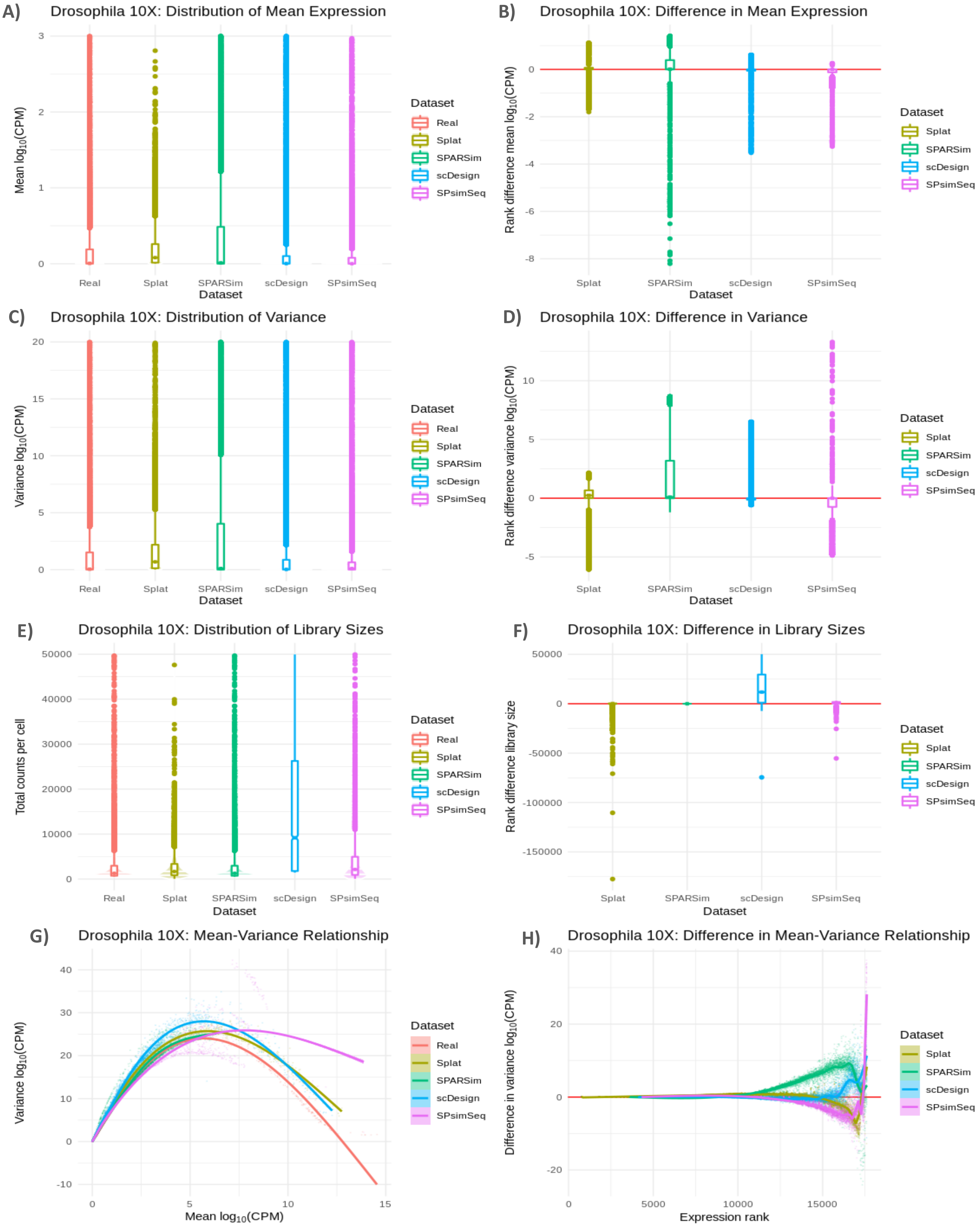


Figure 1: Comparison of simulated datasets using the Drosophila 10X real data. The left-hand plots show the similarities of each simulated dataset alongside the real data in four estimated parameters; boxplot showing distribution of mean gene expression ***A)****, boxplot showing distribution of variance* ***C)****, boxplot showing the distribution of library sizes* ***E)*** *and a scatterplot of the mean-variance relationship* ***G)****. Each dataset is colour-coded indicated by the key on the right of the graph. The right-hand plots show the ranked differences in each simulated dataset relative to the real data;* ***B)*** *boxplot showing the difference in mean gene expression from the real data,* ***D)*** *boxplot showing the difference in variance from the real data,* ***F)*** *boxplot showing the difference in library size from the real data,* ***H)*** *scatterplot showing the difference in mean-variance relationship from the real data. The real data is represented as the red line at 0 and the dataset generated by each simulator is colour coded indicated by the key to the right of the graph. The Y-axis of the boxplots have been adjusted to show the results in more detail.*

When looking at the gene mean distribution, we can see that SPARSim performs poorly and overestimates the mean when compared to the real data. However, scDesign simulates gene means that are the closest to the real data. Similarly, the variance yields the same results where SPARSim overestimates the variance of the data as does the Splat simulation. Library size estimation on the other hand, is where SPARSim clearly outperforms all other packages in particular scDesign. The mean-variance relationship metric is most accurately characterised by scDesign when looking at the difference from the real data with Splat and SPsimSeq performing poorly.

Looking at the zero-characteristics of our data in Figure 2 it seems that sparsity by gene, defined as zeros per gene feature, and sparsity by cell were best defined by the Splat simulation with SPARSim underestimating both metrics. However, none of the packages performed particularly well in accurately representing a vital characteristic of scRNA-seq data. Lastly in the mean sparsity relationship we can see that the Splat package seems to follow the data well in addition to scDesign with SPARSim and SPsimSeq performing poorly.

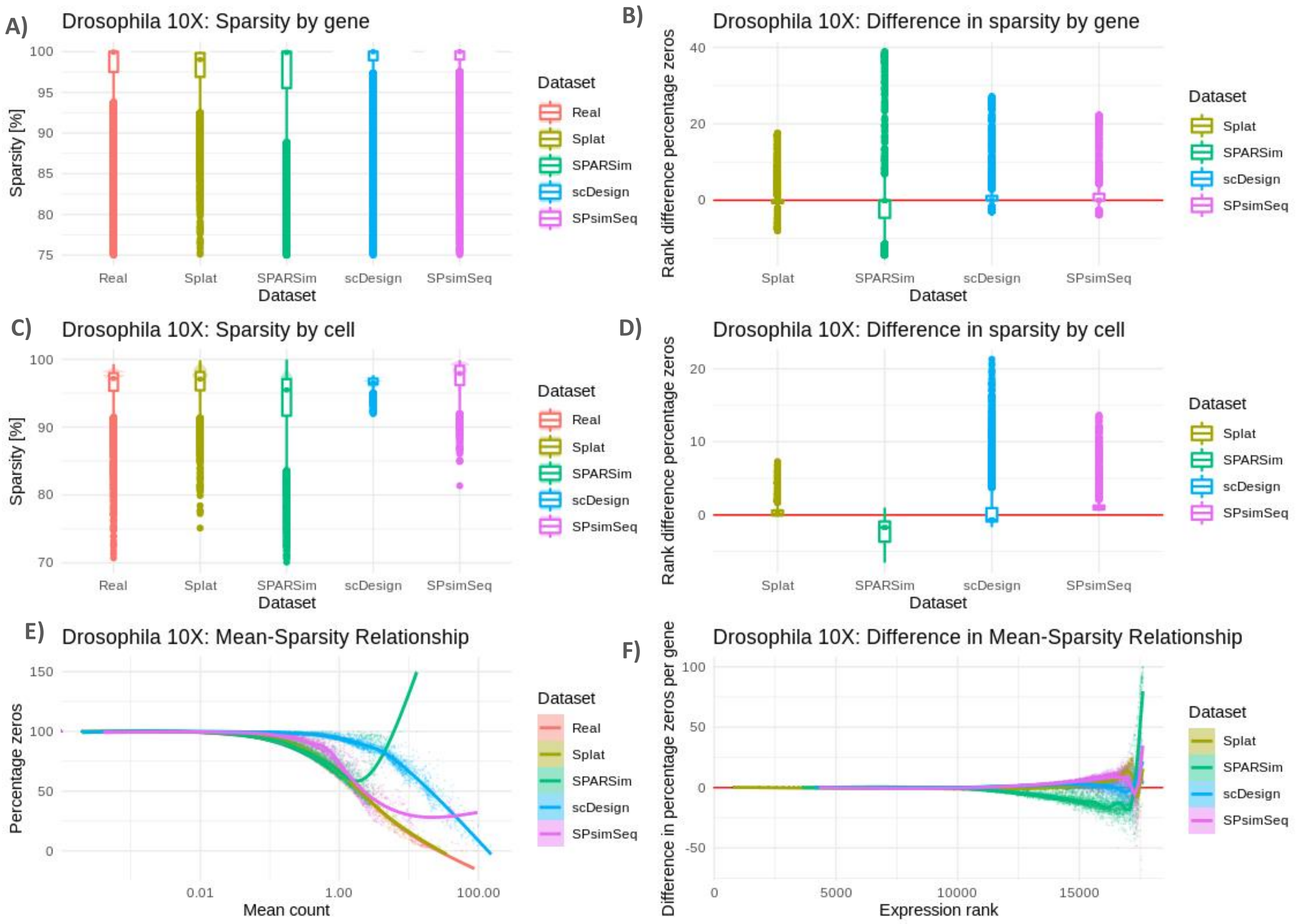


Figure 2: Comparison of zero-characteristics in the simulated datasets using the Drosophila 10X real data. The left-hand plots show the similarities of each simulated dataset alongside the real data in three estimated parameters; ***A)*** *a boxplot showing the distribution of the percentage of zeros per gene (sparsity by gene),* ***C)*** *a boxplot showing the distribution of the percentage of zeros per cell (sparsity by cell) and* ***E)*** *a scatterplot showing the mean-sparsity relationship. Each dataset is colour-coded indicated by the key on the right of the graph. The right-hand plots show the ranked differences in each simulated dataset relative to the real data;* ***B)*** *a boxplot showing the difference in the percentage of zeros per gene (sparsity by gene) from the real data,* ***D)*** *a boxplot showing the difference in the percentage of zeros per cell from the real data (sparsity by cell),* ***F)*** *a scatterplot showing the difference in the mean-sparsity relationship from the real data. The real data is represented as the red line at 0 and the dataset generated by each simulator is colour coded indicated by the key to the right of the graph. The Y-axis of the boxplots have been adjusted to show the results in more detail.*

To more accurately determine the performance of each simulator across the seven real datasets that were used in the study the median absolute deviation (MAD) statistic for each metric was calculated. The MAD statistic for distribution of mean expression for example is obtained by calculating the mean expression values from the real data and each of the simulated datasets. These values are sorted, and the real expression value is subtracted from the simulated expression value. The median of the difference between the two values is then used as the MAD statistic(5). The MAD values were then assigned a ranking (MADRank) between 1-4 with 1 being the most representative of the real data and 4 being the least that are summarised in Figure 3.

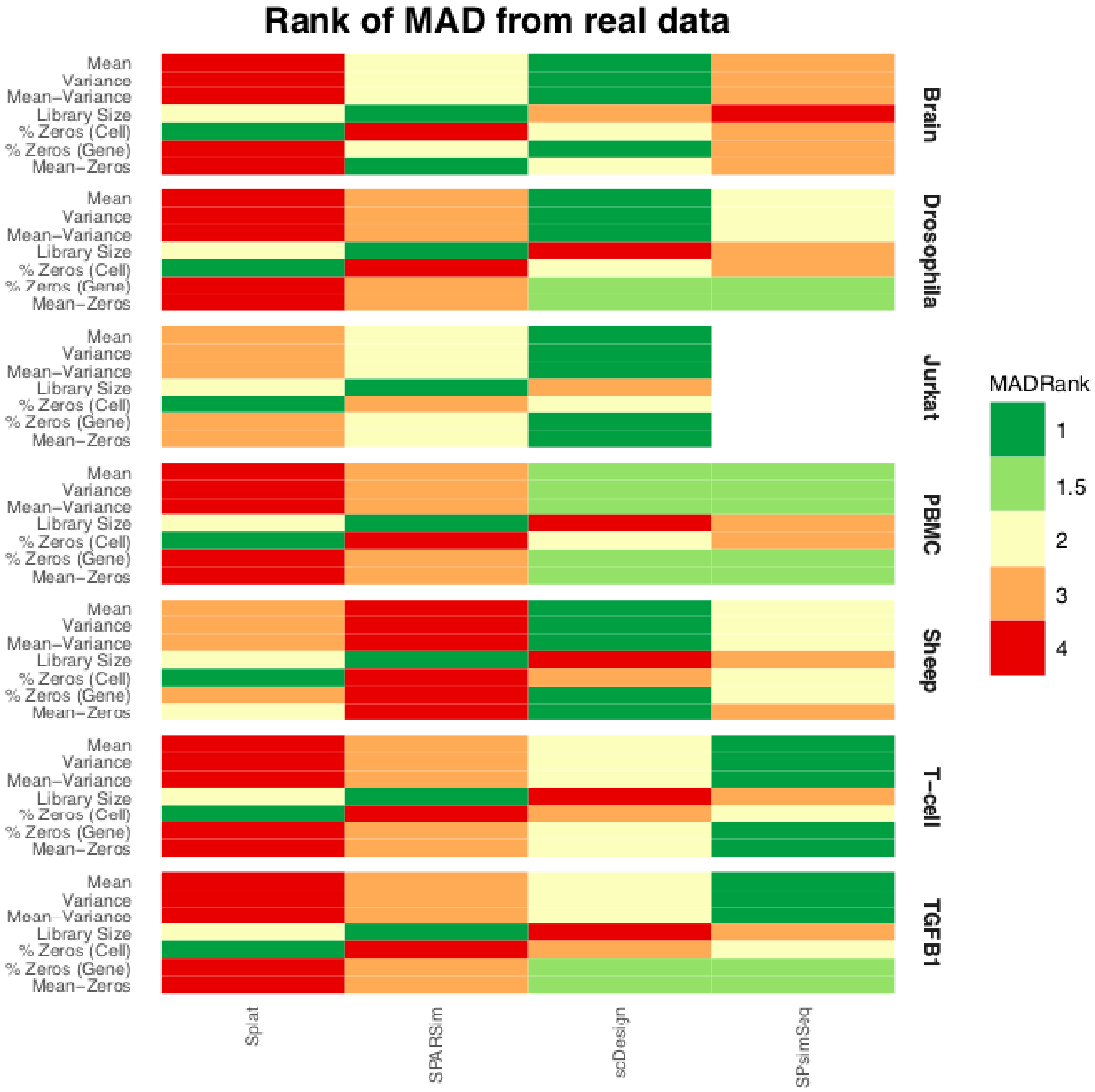


Figure 3: Overall comparison of the performance of each of the four data simulator packages in recapitulating the real 10X datasets. For each dataset the metrics that were estimated are listed down the left-hand side. For each metric result for all the datasets the median absolute deviation (MAD) statistic was calculated relative to the real data. These values were then assigned a rank (MADRank) from 1-4 and are colour coded indicated by the key on the right-hand side of the plot. A MADRank of 1 indicates accurate estimation of the metric when compared to the real data. SPsimSeq was unable to simulate a synthetic dataset for the 293T/Jurkat dataset therefore these results are missing from the heatmap.

From the heatmap we can observe that overall the simulations from scDesign are ranked highly across most datasets with the exception of the 10X T-cell and human TGFB1 datasets where the package did not perform as well. scDesign also simulated the 293T/Jurkat dataset well which is the most complex of the datasets in this study due to the different groups of cells included in the data. A consistent metric however that scDesign consistently failed to accurately replicate is the library size of each real dataset tested. In this metric, the outstanding performer is SPARSim which seems to be able to successfully reflect library size in the simulated datasets however aside from this the package performed poorly in all other areas. The lowest ranked simulator package is clearly Splat which did not handle any of the droplet-based datasets well with the exception of sparsity by cell which was accurately simulated in each synthetic dataset. This may be down to the increased dropout rate in droplet-based scRNA-seq protocols that may have affected parameter estimation in the simulation despite Splat accounting for this. The final simulator package SPsimSeq performed quite well with the obvious exception of the 10X Brain dataset however it was unable to successfully simulate the Jurkat dataset. This was down to small sample sizes of the different cell types where the package was unable to estimate the necessary parameters to perform the simulation. As a result of this comparative study of simulators it was decided that scDesign would be the optimal package to use for the generation of a synthetic dataset. For the differential expression analysis, the Drosophila dataset was chosen for the basis of our simulated datasets.

Evaluation of Differential Expression Analysis Tools

The synthetic datasets from the real Drosophila data consist of four different sample sizes defined as cells per condition across a two-condition framework and are as follows; 25, 50, 100, and 200 cells per condition. In each dataset there are a total of 17,608 genes with 1,760 being differentially expressed. These are divided equally into 880 up- and 880 down-regulated genes.

ROC/AUC

When examining the ROC curve the ideal performance of a given package should closely follow the y-axis and not drop towards the x-axis, this gives a high AUC close to 1 indicating good performance of the package in identifying differentially expressed genes.

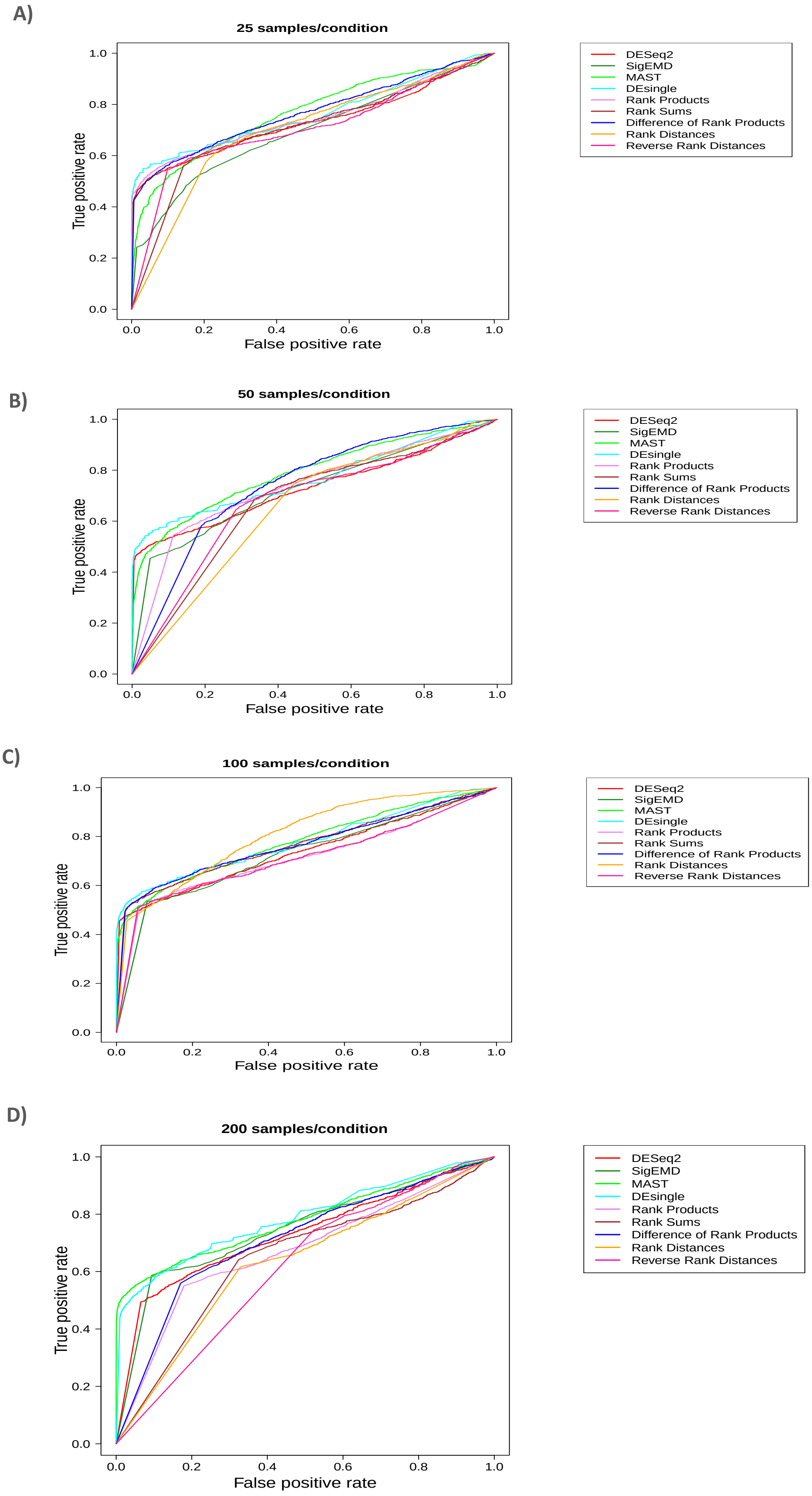


Figure 4: Receiver operating characteristic (ROC) curve showing the ability of each package to be able to distinguish between a true DE and non-DE gene. Differential expression packages are colour-coded as per the key to the right of the graphs. ***A)*** *ROC curve for 25 cells per condition,* ***B)*** *ROC curve for 50 cells per condition,* ***C)*** *ROC curve for 100 cells per condition and* ***D)*** *ROC curve for 200 cells per condition.*

Figure 4 shows the ROC for each of the different sample sizes tested where we can clearly see that sample size does indeed have an effect on performance. As the ROC curves are hard to distinguish the AUC value is more informative shown in Figure 5.

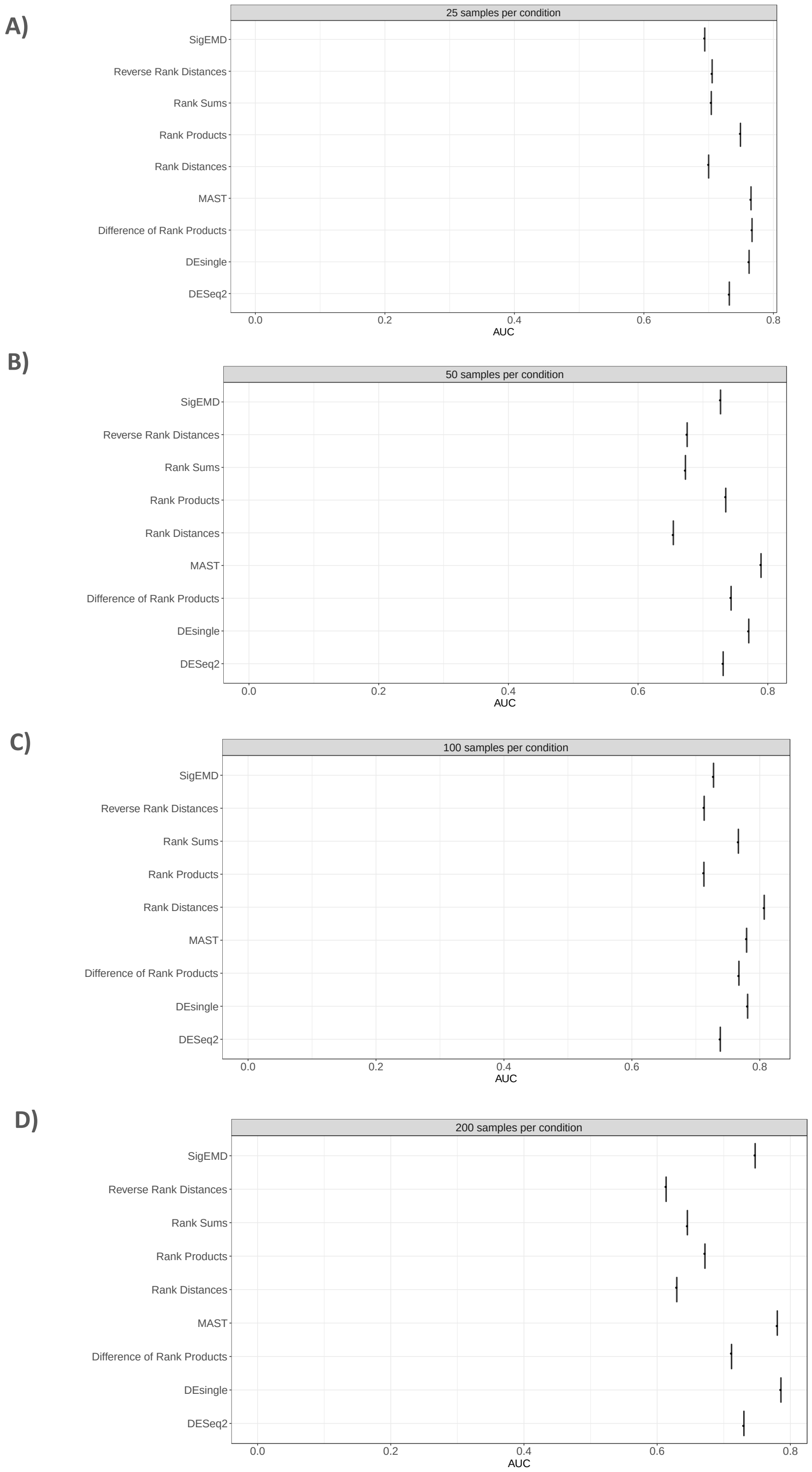


Figure 5: Area under the curve (AUC) plots for each of the differential expression analysis packages; ***A)*** *AUC plot for 25 cells per condition,* ***B)*** *AUC plot for 50 cells per condition,* ***C)*** *AUC plot for 100 cells per condition,* ***D)*** *AUC plot for 200 cells per condition*

At a small sample size of 25 cells we can observe that Difference of Rank Products (DRP) and MAST performed the best with the highest AUC value out of all the packages. From the non-parametric methods, we can see that all RankStat methods outperformed SigEMD. As sample size increases the overall performance of the RankStat methodologies decreases with the exception at 100 samples per condition where Rank Distances has the highest AUC value. Furthermore, despite outperforming SigEMD at the smallest sample size, we can see that the alternative non-parametric method outperforms RankStat at the largest sample size 200 cells. This suggests RankStat, as a non-parametric method for analysing scRNA-seq data, could be more efficient when sample size per condition is very small and SigEMD may be more useful when sample size is large. When we compare the AUC of the RankStat methods with DESeq2 the bulk-RNA package DRP yields a higher AUC value with the exception of 200 cells/condition. At each sample size however MAST and DEsingle consistently have the highest AUC values of all the packages outperforming the rest.

FDR

The next metric investigated was the FDR which is representative of the fraction of true non-differentially expressed genes that we expect to find in our differentially expressed genes results. Here the FDR threshold was set at an adjusted p-value threshold of 0.05 thus any packages with an FDR strongly exceeding the 0.05 threshold have lower control in correctly identifying true differentially expressed genes. The results can be seen below in Figure 6.

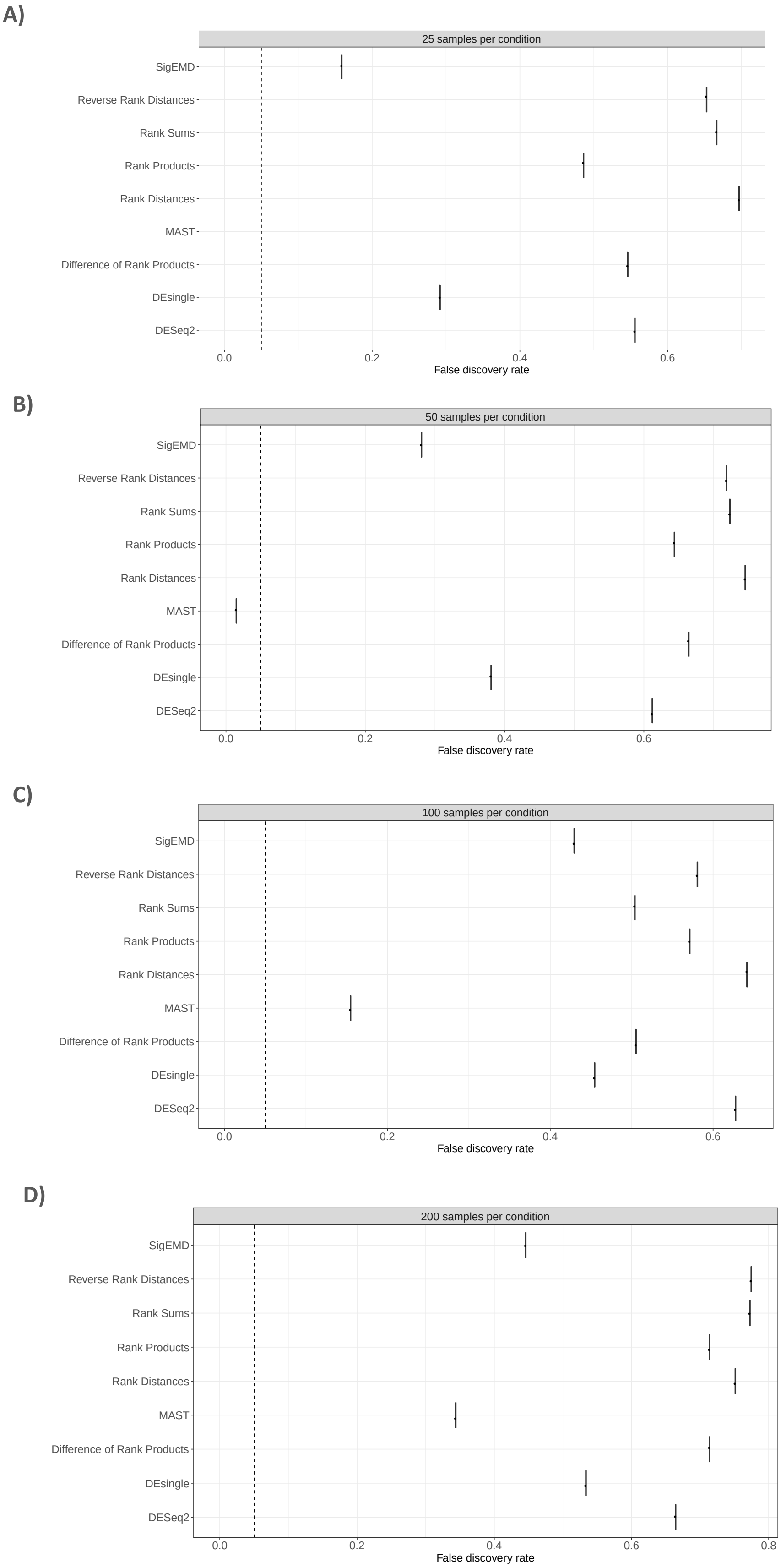


Figure 6: Plots showing the false discovery rate (FDR) control for each differential analysis package. The dotted line indicates the desired FDR threshold set at 0.05. ***A)*** *FDR plot for 25 cells per condition,* ***B)*** *FDR plot for 50 cells per condition,* ***C)*** *FDR plot for 100 cells per condition and* ***D)*** *FDR plot for 200 cells per condition.*

At the lowest sample size, we can see that none of the packages performed particularly well in their ability to control FDR with SigEMD obtaining the lowest FDR at 0.376. Rank Distance and Rank Sum obtained the highest FDR at 0.697 and 0.666 respectively. At 50 samples per condition MAST had the smallest FDR falling below the 0.05 threshold with RRD, RS and RD being the top 3 worst performing in FDR control. The FDR is also very high at 100 cells per condition with SigEMD showing the best control out of the non-parametric methods and DESeq2 having the highest FDR out of the parametric methods. Similarly, at the largest sample size, the control of FDR is depreciated across all packages with none achieving on or below the 0.05 threshold. Across all sample sizes the worst package for controlling FDR is RD that consistently has a large FDR value.

TPR/Recall

Similarly, the TPR is defined as the fraction of true differentially expressed genes that are counted as significant at an adjusted p-value threshold of 0.05. The higher the TPR value, coupled with the consideration of the FDR, the better the package is at identifying true differentially expressed genes. The results of TPR for the packages are summarised in Figure 7 below.

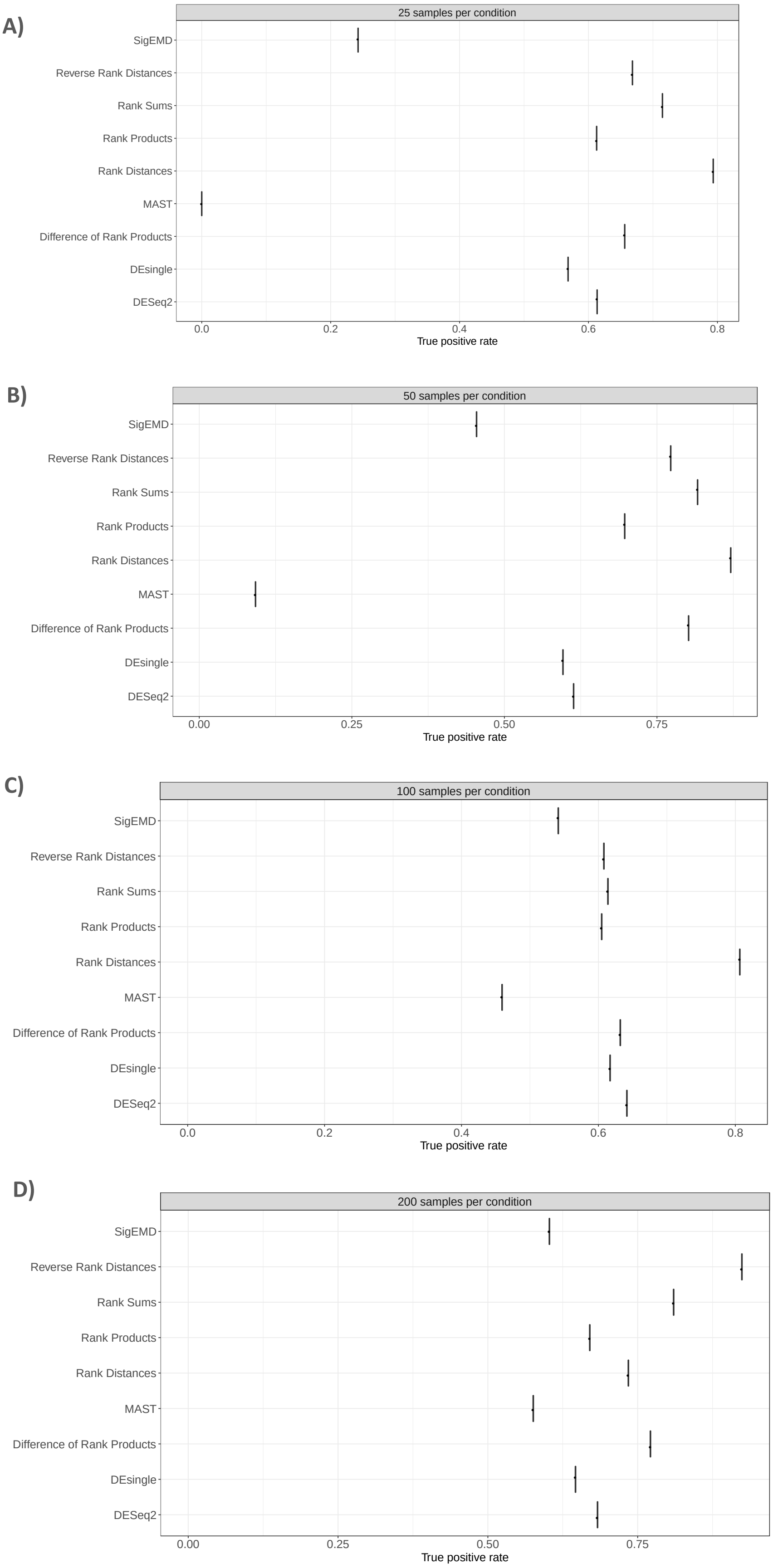
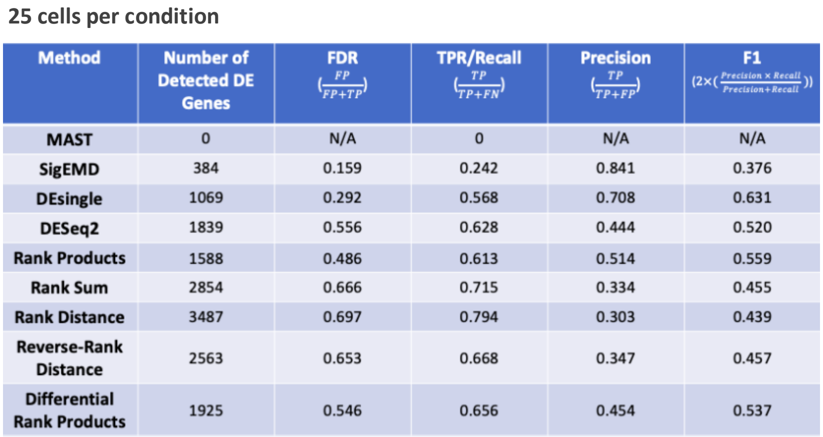


*Figure 7: Plots summarising the true positive rate (TPR) for each differential expression analysis package. **A)** TPR plot for 25 cells per condition, **B)** TPR plot for 50 cells per condition, **C)** TPR plot for 100 cells per condition and **D)** TPR plot for 200 cells per condition.*

In the smallest sample size, the TPR for the RankStat methodologies were among the highest in the comparison with MAST failing to identify any differentially expressed genes. Similarly, in 50 cells per condition the TPR stays consistent for the RankStat methods however it is worth noting the FDR in both cases was far above the 0.05 threshold. MAST shows some improvement in TPR at 100 cells/condition and had the lowest FDR control at this sample size. The results for the largest sample size are quite varied across the packages indicating that larger sample sizes influence the efficacy of differential gene detection.

Precision/F1 score

Summarised in Tables 1-4 below are the mathematical calculations of precision and F1 score for each sample size, the full summary statistic table can be found in the Appendix.

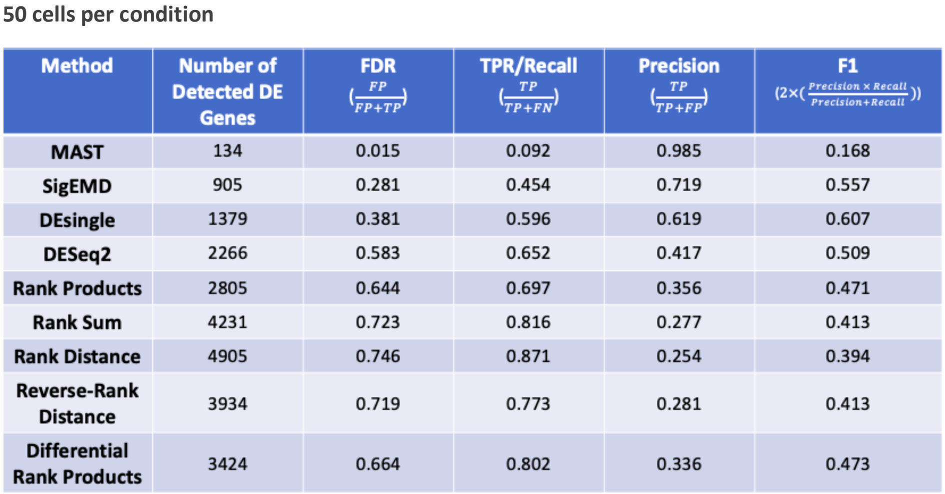


25 cells per condition

Method	Number of Detected DE Genes	FDR $(\frac{FP}{FP+TP})$	TPR/Recall $(\frac{TP}{TP+FN})$	Precision $(\frac{TP}{TP+FP})$	F1 $(2\times(\frac{Precision \times Recall}{Precision+Recall}))$
MAST	0	N/A	0	N/A	N/A
SigEMD	384	0.159	0.242	0.841	0.376
DEsingle	1069	0.292	0.568	0.708	0.631
DESeq2	1839	0.556	0.628	0.444	0.520
Rank Products	1588	0.486	0.613	0.514	0.559
Rank Sum	2854	0.666	0.715	0.334	0.455
Rank Distance	3487	0.697	0.794	0.303	0.439
Reverse-Rank Distance	2563	0.653	0.668	0.347	0.457
Differential Rank Products	1925	0.546	0.656	0.454	0.537

Table 1: A summary statistics table showing the total number of detected differentially expressed genes by each package for 25 cells per condition. The calculated false discovery rate (FDR), true positive rate (TPR/Recall), precision and F1 score. Calculation formulas are included in the header where appropriate with number of false positives denoted as FP, the number of true positives as TP and the number of false negatives as FN.

MAST failed to detect any DE genes at this sample size so the precision and F1 was unable to be calculated. SigEMD showed the highest precision however had the lowest F1 score with DEsingle having the highest F1 score followed by RP.

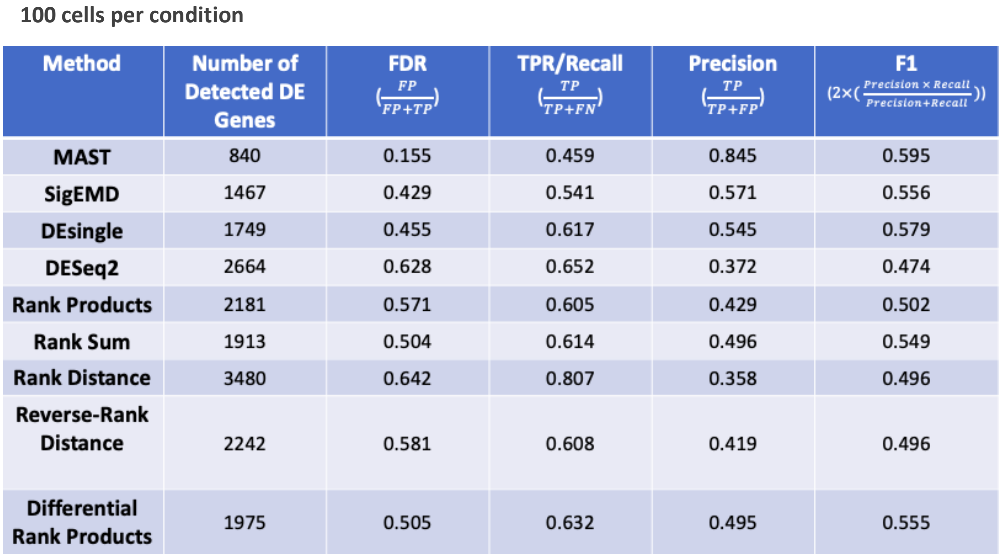


50 cells per condition

Method	Number of Detected DE Genes	FDR $(\frac{FP}{FP+TP})$	TPR/Recall $(\frac{TP}{TP+FN})$	Precision $(\frac{TP}{TP+FP})$	F1 $(2\times(\frac{Precision \times Recall}{Precision+Recall}))$
MAST	134	0.015	0.092	0.985	0.168
SigEMD	905	0.281	0.454	0.719	0.557
DEsingle	1379	0.381	0.596	0.619	0.607
DESeq2	2266	0.583	0.652	0.417	0.509
Rank Products	2805	0.644	0.697	0.356	0.471
Rank Sum	4231	0.723	0.816	0.277	0.413
Rank Distance	4905	0.746	0.871	0.254	0.394
Reverse-Rank Distance	3934	0.719	0.773	0.281	0.413
Differential Rank Products	3424	0.664	0.802	0.336	0.473

Table 2: A summary statistics table showing the total number of detected differentially expressed genes by each package for 50 cells per condition. The calculated false discovery rate (FDR), true positive rate (TPR/Recall), precision and F1 score. Calculation formulas are included in the header where appropriate with number of false positives denoted as FP, the number of true positives as TP and the number of false negatives as FN.

Here the RankStat methodologies had the lowest precision values meaning they returned many false positives. MAST shows the lowest F1 score but high precision as although the FDR was low, few DE genes were identified in total.

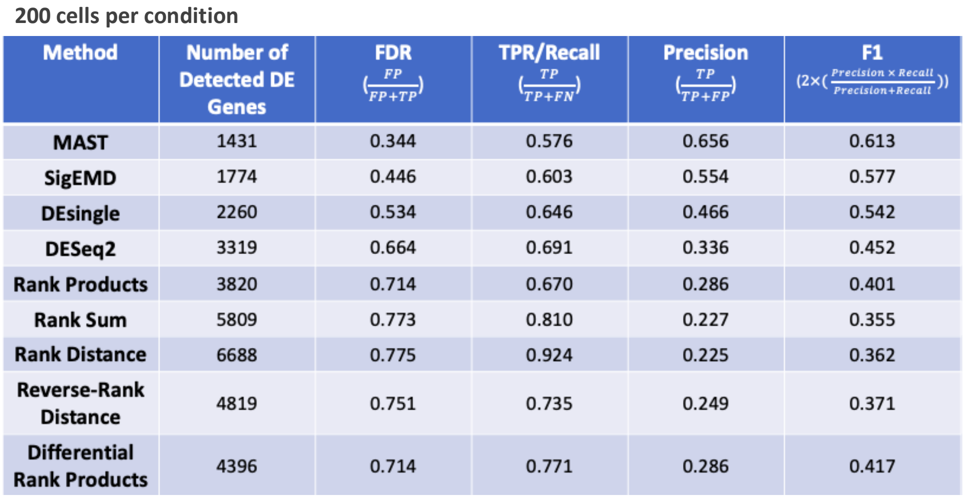


100 cells per condition

Method	Number of Detected DE Genes	FDR $(\frac{FP}{FP+TP})$	TPR/Recall $(\frac{TP}{TP+FN})$	Precision $(\frac{TP}{TP+FP})$	F1 $(2\times(\frac{Precision \times Recall}{Precision+Recall}))$
MAST	840	0.155	0.459	0.845	0.595
SigEMD	1467	0.429	0.541	0.571	0.556
DEsingle	1749	0.455	0.617	0.545	0.579
DESeq2	2664	0.628	0.652	0.372	0.474
Rank Products	2181	0.571	0.605	0.429	0.502
Rank Sum	1913	0.504	0.614	0.496	0.549
Rank Distance	3480	0.642	0.807	0.358	0.496
Reverse-Rank Distance	2242	0.581	0.608	0.419	0.496
Differential Rank Products	1975	0.505	0.632	0.495	0.555

Table 3: A summary statistics table showing the total number of detected differentially expressed genes by each package for 100 cells per condition. The calculated false discovery rate (FDR), true positive rate (TPR/Recall), precision and F1 score. Calculation formulas are included in the header where appropriate with number of false positives denoted as FP, the number of true positives as TP and the number of false negatives as FN.

For 100 cells per condition MAST obtained the highest precision value and F1 score at 0.845 and 0.595 respectively. SigEMD had the highest precision values and F1 score out of the non-parametric methods. DESeq2 and RD had the lowest precision indicating both methods returned many false positives.



200 cells per condition

Method	Number of Detected DE Genes	FDR ($\frac{FP}{FP+TP}$)	TPR/Recall ($\frac{TP}{TP+FN}$)	Precision ($\frac{TP}{TP+FP}$)	F1 (2×($\frac{Precision \times Recall}{Precision+Recall}$))
MAST	1431	0.344	0.576	0.656	0.613
SigEMD	1774	0.446	0.603	0.554	0.577
DEsingle	2260	0.534	0.646	0.466	0.542
DESeq2	3319	0.664	0.691	0.336	0.452
Rank Products	3820	0.714	0.670	0.286	0.401
Rank Sum	5809	0.773	0.810	0.227	0.355
Rank Distance	6688	0.775	0.924	0.225	0.362
Reverse-Rank Distance	4819	0.751	0.735	0.249	0.371
Differential Rank Products	4396	0.714	0.771	0.286	0.417

Table 4: A summary statistics table showing the total number of detected differentially expressed genes by each package for 200 cells per condition. The calculated false discovery rate (FDR), true positive rate (TPR/Recall), precision and F1 score. Calculation formulas are included in the header where appropriate with number of false positives denoted as FP, the number of true positives as TP and the number of false negatives as FN.

At the largest sample size, we see MAST outperform all differential analysis packages in terms of precision and F1 score being 0.656 and 0.613 respectively. The RankStat methods performed the worst with low precision and low F1 scores, detecting a large number of DE genes with poor FDR control.

Agreement Among Methods

Lastly to investigate the agreement amongst the differential expression methods tested in this study we looked at how similar each method is in the ranking scores they assign genes in the data. This is done by doing a pair-wise comparison for each gene and calculating the Spearman correlation statistic. The methods are then subsequently clustered in a hierarchical dendrogram shown in Figure 8 to show methods that rank genes similarly.

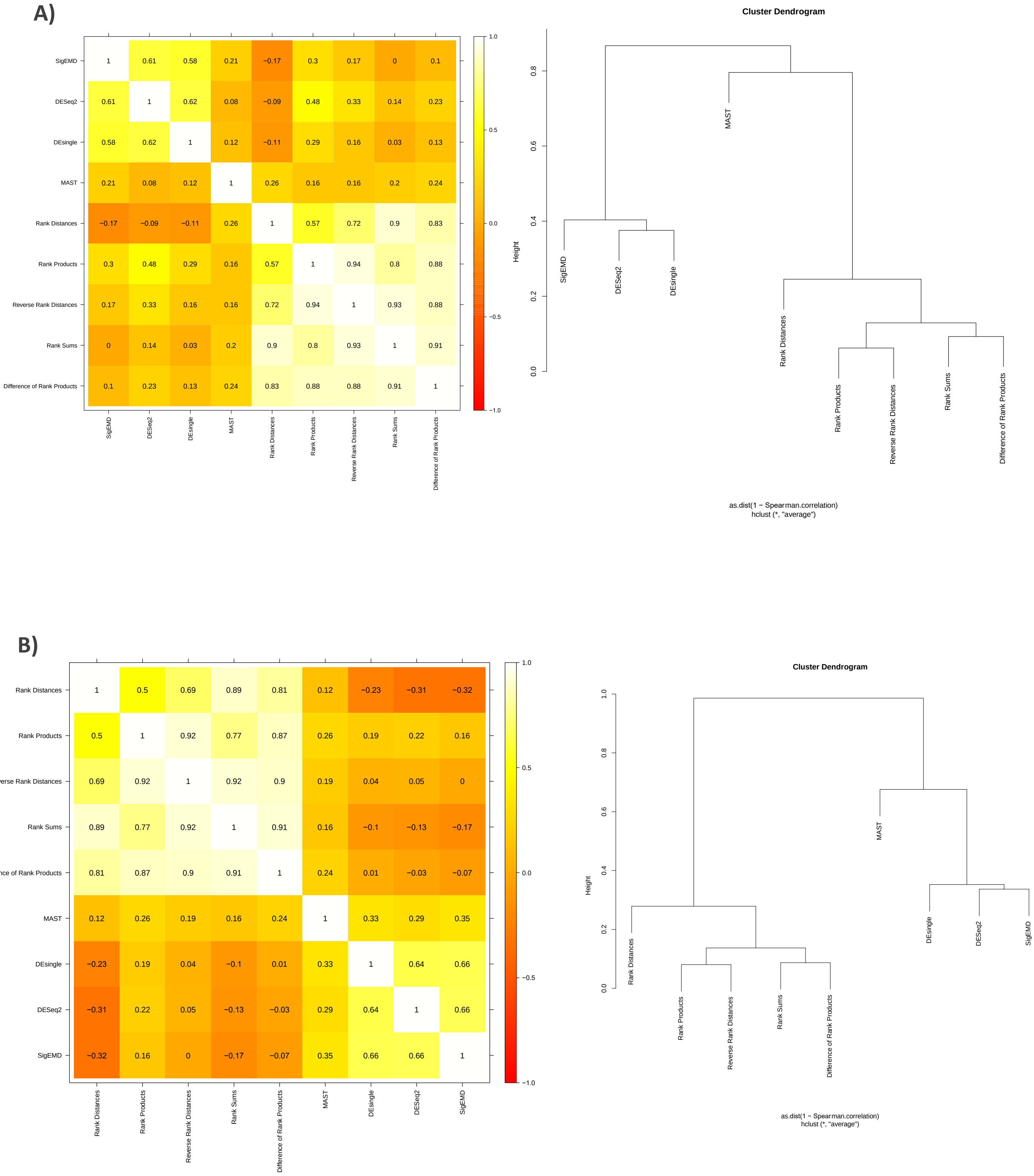


Figure 8a: Spearman correlation heatmaps showing the similarity of ranking of genes between differential analysis packages and the corresponding cluster dendrogram. A high positive value close to 1 indicates close similarity and a negative value close to -1 indicates dissimilarity between packages. Packages closer to each other on the dendrogram indicate they rank genes similarly and the more distant a package is from another indicates difference in gene ranking. ***A)*** *Spearman correlation heatmap and clustered dendrogram for 25 cells per condition and* ***B)*** *Spearman correlation heatmap and clustered dendrogram for 50 cells per condition.*

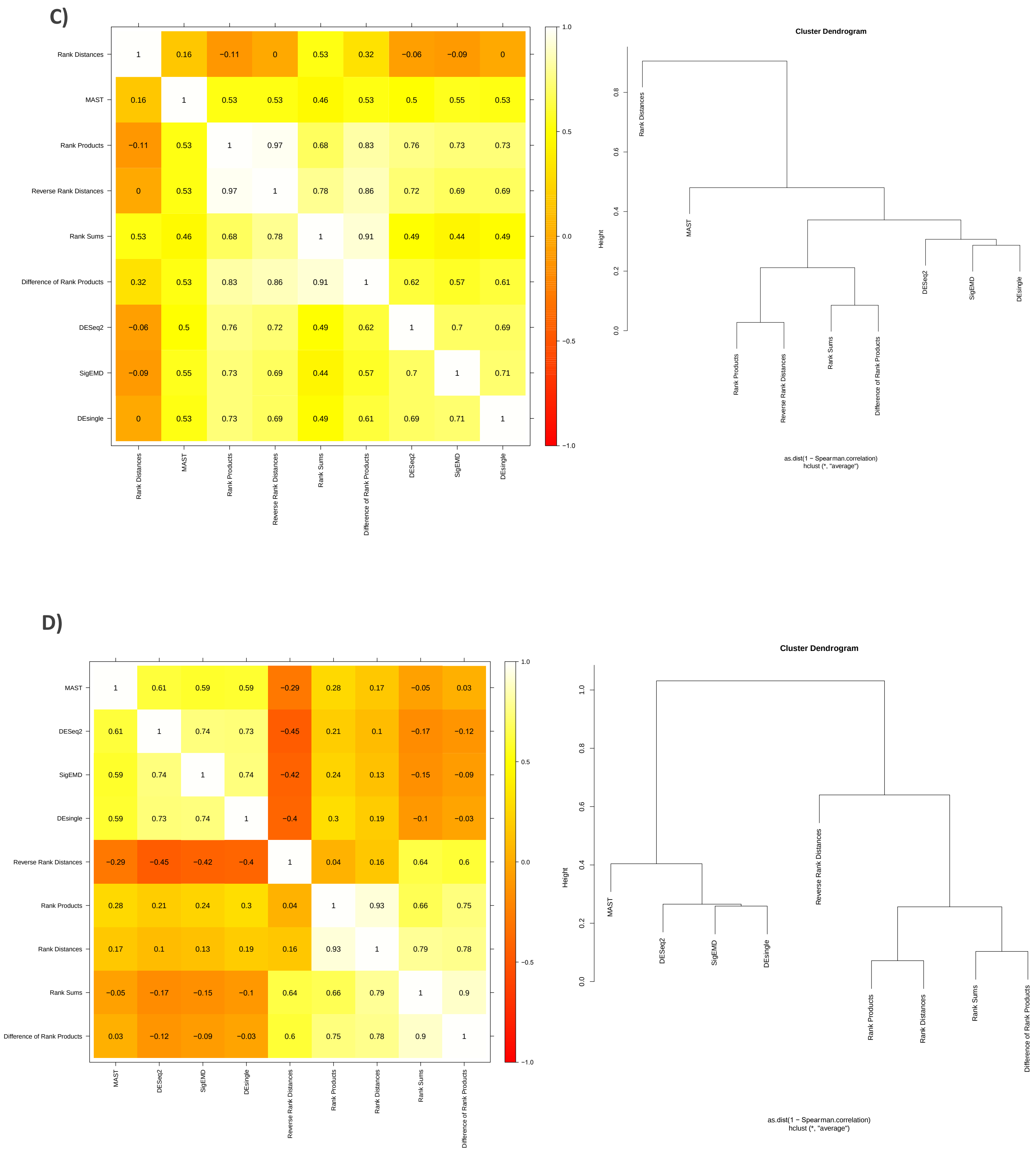


Figure 8b: Spearman correlation heatmaps showing the similarity of ranking of genes between differential analysis packages and the corresponding cluster dendrogram. A high positive value close to 1 indicates close similarity and a negative value close to -1 indicates dissimilarity between packages. Packages closer to each other on the dendrogram indicate they rank genes similarly and the more distant a package is from another indicates difference in gene ranking. ***C)*** *Spearman correlation heatmap and clustered dendrogram for 100 cells per condition and* ***D)*** *Spearman correlation heatmap and clustered dendrogram for 200 cells per condition.*

We can see that RankStat packages rank genes similarly while SigEMD ranks genes more similar to DESeq2 and DEsingle. In addition to this we also looked at the similarity in returned genes by calculating the Sorenson index which shows the overlap between significant genes returned at an adjusted p-value of 0.05.

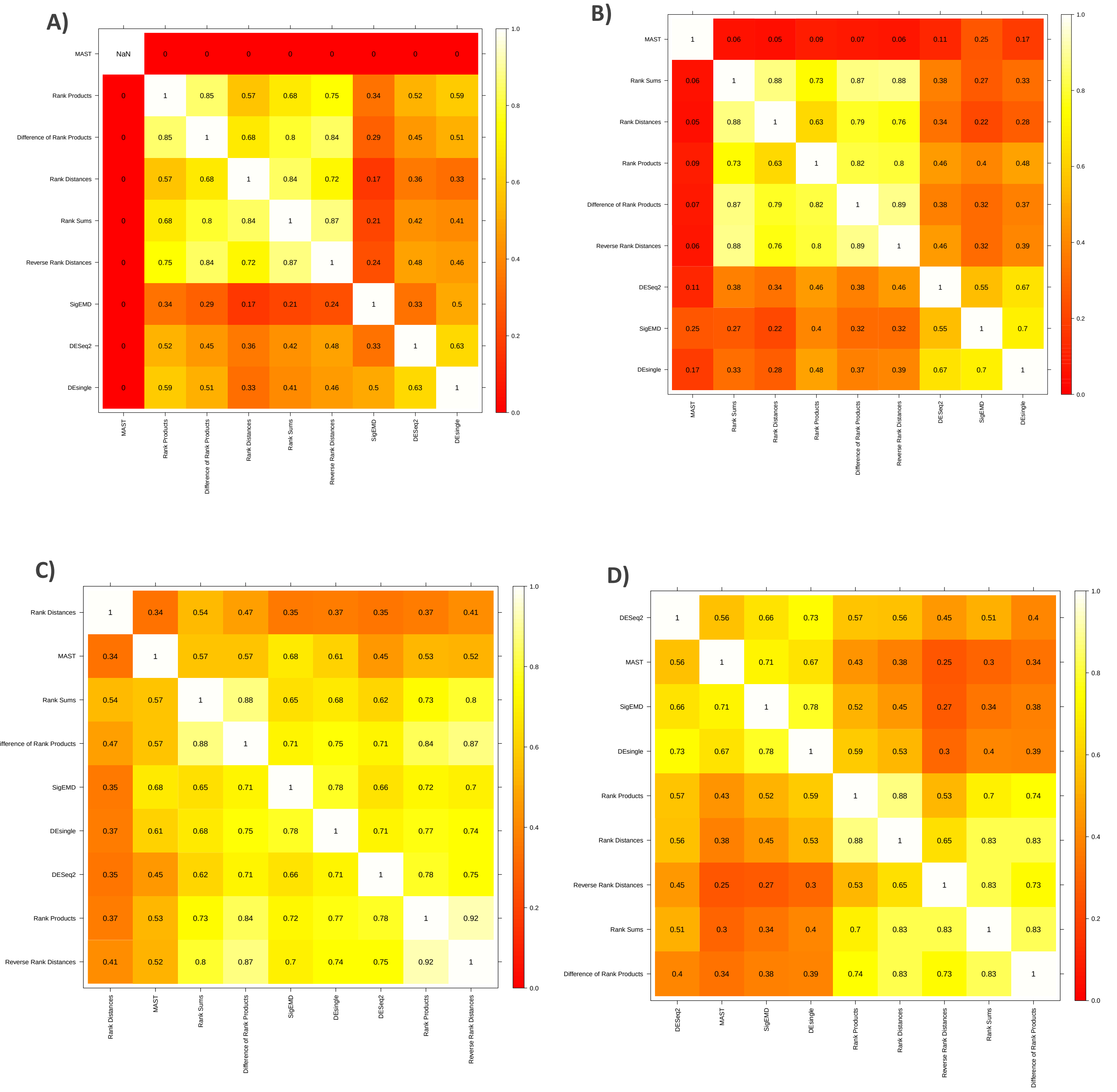


Figure 9: Sorensen index heatmap that shows the overlap of returned DE genes between each differential analysis package. The Sorensen index is the ratio between the number of genes that are shared between two packages and the average number of genes returned by the packages. An index closer to 1 indicates strong overlap. ***A)*** *Sorensen index heatmap for 25 cells per condition,* ***B)*** *Sorensen index heatmap for 50 cells per condition,* ***C)*** *Sorensen index heatmap for 100 cells per condition and* ***D)*** *Sorensen index heatmap for 200 cells per condition.*

Out of the RankStat methods RRD and RS had the most overlap in returned DE genes across all sample sizes with the exception of 100 cells per condition where the highest overlap was seen between RRD and RP.

Discussion

Data Simulator Packages

The necessity for competent and accurate scRNA-seq analysis methods is of great importance especially due to the complex and difficult nature of the data. The large presence of zero values due to the noisiness of scRNA-seq datasets seemed to be the biggest obstacle in the analysis pipeline for both the data simulators and the differential expression analysis packages. With many advancing developments in novel bioinformatic tools to analyse scRNA-seq data the need for a reliable, robust simulator package must first be selected to evaluate the viability of their performances. Not exempt from this is the simulator packages themselves which must also be thoroughly investigated to their claim to simulate real datasets in particular, in this study, exclusively from the 10X platform, a technology implemented at Glasgow Polyomics. Some of the data we used were also used in the recently published SPARSim simulator paper(7) however this study did not report the same success and the additional 10X datasets used further confirmed a poor performance overall of the package. This highlights another issue in the analysis of scRNA-seq data where different datasets yield different results. A possible explanation of this could be down to biological variability of datasets where more complex scRNA-seq data containing many different cell types displaying expression of different genes may make it harder to estimate parameters in the synthetic dataset that accurately reflect the real data. Also, it is worth mentioning that different experimental techniques may introduce technical variability for example the increased noise in data that comes with droplet-based protocols when compared to well-based techniques such as SMART-Seq2(16). The intrinsic variability of scRNA-seq data creates the need for preliminary comparative analysis of data simulator packages a highly valuable functionality which Splatter provides. Splatter currently only provides six simulation models within the package however over 14 other data simulators have been proposed to be added including scDesign pending demand from the scientific community. This provides an opportunity for multiple scRNA-seq data simulators to be available for use on a singular platform increasing accessibility of these types of analyses and ease of comparison as demonstrated in this study. Another advantage to Splatter although not the best performing package on the datasets used in this study, it was the most well documented and user-friendly package to use from the four packages evaluated. The package manual provides well commented reducible code for all functionalities which could potentially open up scRNA-seq analysis to those unfamiliar with hard bioinformatics pipelines. The top performing package however in this case was scDesign which has other functionalities within the package not described in this study. As well as simulating data to benchmark differential expression analysis tools scDesign also provides the framework to perform power analysis to aid rational experimental design. For example, one of the simulation models included in the package allow the estimation of the optimal cell number necessary at a given fixed sequencing depth to best perform differential expression analysis. An experimental trade-off in scRNA-seq analysis is the more cells that are sequenced the shallower the sequencing depth must be, therefore without unnecessary costs risking the quality of the real data one may use synthetic data instead to determine suitable experimental design(6).

Differential Analysis Methods

When subsequently examining the performance of the differential expression analysis methods, out of a total 1,760 true differentially expressed genes in the simulated dataset the number of detected genes provided by each package varied. The most conservative method across all sample sizes with the exception of 25 cells/condition was MAST. In the identification of marker genes across two populations of cells in scRNA-seq, precision is more important than recall as we want to ensure the identified DE genes are indeed true. The RankStat packages demonstrated the worst FDR control and precision across most sample sizes meaning that there were a large number of false positives returned by the methods. It is worth pointing out that RankStat is designed for microarray data and not scRNA-seq. The implementation of using fold-changes to quantify gene expression can lead to a large number of false positives especially in data where there are low abundance transcripts that can have a large effect on gene expression(17). When we look at the development of DEsingle the underlying statistical framework of DESeq has been integrated alongside the implementation of a ZINB model to account for drop-out effects(13)(18). Perhaps RankStat may be further developed to provide an additional model to account for the zero-inflation of scRNA-seq data that leads to increased variability that arbitrary fold-change cut-off methods have insufficient power to compute(17).

Effect of Sample Size

When considering the effect of sample size on the performance of the differential expression analysis packages there is clearly some bias introduced. At the largest sample size none of the methods tested exhibited good control of FDR at the specified threshold resulting in many false positives being detected. Similarly, the same can be observed in the smaller sample sizes of 25 and 50 cells per condition where there were poor control of FDR and therefore lower precision in detecting true differentially expressed genes. When deciding which differential expression analysis tool to use when analysing scRNA-seq data it is apparent that considering the effect sample size will have on the performance of DE gene detection. Studies have concluded that non-parametric methods may be advantageous to analysing scRNA-seq datasets in particular of larger sample sizes due to their simplistic nature and lack of constraint from shape distribution parameters(19). Due to the multimodality of scRNA-seq data, underlying assumptions imposed by the statistical framework of parametric methods may not fit the distribution of the data between cell groups for example making non-parametric methods potentially a more robust alternative(20). Despite the SigEMD outperforming RankStat methods, it is apparent that for the dataset simulated in this study the parametric methods MAST and DEsingle tailored to scRNA-seq analysis outperformed their non-parametric counterparts. However, given the dataset was a simple two-condition design perhaps when analysing a more complex framework including multiple cell sub-populations and evident multimodality of distribution between the different groups, simpler non-parametric methods may be more attractive.

Agreement of Methods

Overall there were very little agreement between the methods in the differentially expressed genes that they returned with the exception of all the RankStat methods. This is to be expected as the rank-based methodologies are all derivatives of the first proposed method Rank Products(11). Therefore, due to the range of results between the methods it is hard to determine if there is a clear winner in the context of the identification of differentially expressed genes. This may be down to differences between the various packages in the approach to handling the large number of zeros in the data, the multimodality of gene expression and drop out events that may occur in real data or are simulated within synthetic datasets.

Limitations

Since the study conducted here is simply a pilot study, a more thorough analysis must be conducted to obtain a definitive conclusion exploring the topics described in more depth. For example, when evaluating the data simulators, despite this study using the highest number of 10X droplet-based datasets compared to current papers there it would be advantageous to use more datasets to further validate the results of this study. Here we conclude scDesign was the best performing simulator however in addition to extending the variety of data we could also test more data simulator packages. Although recently published packages were selected there are a couple of packages that were left out of the comparison for example powsimR(21) and SymSim(22) due to time constraints and installation issues the powsimR package. However results from the scDesign paper(6) show that scDesign outperformed powsimR in the 10X datasets they used although this would be interesting to investigate in relation to 10X datasets. When examining the DE analysis component of this study the biggest limitation in our results is that only one synthetic dataset was used where if more datasets were introduced this will give a more accurate representation of DE package performance. Following this the dataset we used was simple in that it represented a two-condition framework; however, we could explore the effect of multi-group scenarios on performance. Furthermore, we only used four sample sizes therefore a potential development could be to run additional data simulations for more sample sizes. This paper mainly focuses on the performance of the RankStat methods and how they compare to existing DE methods for scRNA-seq data, a further study could use more DE analysis methods such as SCDE(23) and scDD(24).

Conclusion

As the potential benefits of scRNA-seq are becoming more apparent the requirement for robust, easily reproducible and flexible analysis tools will necessitate. Data simulator packages are a valuable tool and the Splatter documentation and interoperability eases the determination of many aspects of the scRNA-seq analysis pipeline. Differential expression packages still struggle to find common ground in the identification of DE genes within data of such complex nature. Non-parametric methods provide some potential for larger more complex datasets however, this small-scale analysis study cannot fully determine the efficacy of the RankStat methodologies with respect to scRNA-seq data. Perhaps with improved documentation and statistical framework development tailored to handling the difficult nature of scRNA-seq datasets, future progress in this field will yield more accurate and robust results.

References

1. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front Genet [Internet]. 2019 [cited 2020 Aug 11];10. Available from:

2. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015 Mar;16(3):133–45.

3. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017 Dec;9(1):75.

4. Wang T, Nabavi S. SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data. Methods. 2018 Aug 1;145:25–32.

5. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 2017 Sep 12;18(1):174.

6. Li WV, Li JJ. A statistical simulator scDesign for rational scRNA-seq experimental design. Bioinformatics. 2019 Jul 15;35(14):i41–50.

7. Baruzzo G, Patuzzi I, Di Camillo B. SPARSim single cell: a count data simulator for scRNA-seq data. Bioinformatics. 2020 Mar 1;36(5):1468–75.

8. Assefa AT, Vandesompele J, Thas O. SPsimSeq: semi-parametric simulation of bulk and single-cell RNA-sequencing data. Bioinformatics. 2020 May 1;36(10):3276–8.

9. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017 16;8:14049.

10. Zhang X, Li T, Liu F, Chen Y, Yao J, Li Z, et al. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. Mol Cell. 2019 03;73(1):130-142.e5.

11. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett. 2004 Aug 27;573(1–3):83–92.

12. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12):550.

13. Miao Z, Deng K, Wang X, Zhang X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Bioinformatics. 2018 Sep 15;34(18):3223–4.

14. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015 Dec 10;16(1):278.

15. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Casp J Intern Med. 2013;4(2):627–35.

16. Wang X, He Y, Zhang Q, Ren X, Zhang Z. Direct Comparative Analysis of 10X Genomics Chromium and Smart-seq2. bioRxiv. 2019 Apr 22;615013.

17. Subramaniam S, Hsiao G. Gene-expression measurement: variance-modeling considerations for robust data analysis. Nat Immunol. 2012 Feb 16;13(3):199–203.

18. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010 Oct 27;11(10):R106.

19. Sengupta D, Rayan NA, Lim M, Lim B, Prabhakar S. Fast, scalable and accurate differential expression analysis for single cells [Internet]. Genomics; 2016 Apr [cited 2020 Aug 14]

20. Zhu A, Srivastava A, Ibrahim JG, Patro R, Love MI. Nonparametric expression analysis using inferential replicate counts. Nucleic Acids Res. 2019 Oct 10;47(18):e105–e105.

21. Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimR: power analysis for bulk and single cell RNA-seq experiments. Bioinforma Oxf Engl. 2017 Nov 1;33(21):3486–8.

22. Zhang X, Xu C, Yosef N. Simulating multiple faceted variability in single cell RNA sequencing. Nat Commun. 2019 Jun 13;10(1):2611.

23. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014 Jul;11(7):740–2.

24. Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, Stewart R, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016 Oct 25;17(1):222.