Quantitative Biology 2016, 4(4): 243–260
DOI 10.1007/s40484-016-0089-7

RESEARCH ARTICLE

Differential expression analyses for single-cell RNA-Seq: old questions on new data

Zhun Miao[1] and Xuegong Zhang[1,2,*]

[1] MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China
[2] School of Life Sciences, Tsinghua University, Beijing 100084, China
* Correspondence: zhangxg@tsinghua.edu.cn

Received July 13, 2016; Revised August 20, 2016; Accepted August 26, 2016

***Background*: Single-cell RNA sequencing (scRNA-seq) is an emerging technology that enables high resolution detection of heterogeneities between cells. One important application of scRNA-seq data is to detect differential expression (DE) of genes. Currently, some researchers still use DE analysis methods developed for bulk RNA-Seq data on single-cell data, and some new methods for scRNA-seq data have also been developed. Bulk and single-cell RNA-seq data have different characteristics. A systematic evaluation of the two types of methods on scRNA-seq data is needed.**

***Results*: In this study, we conducted a series of experiments on scRNA-seq data to quantitatively evaluate 14 popular DE analysis methods, including both of traditional methods developed for bulk RNA-seq data and new methods specifically designed for scRNA-seq data. We obtained observations and recommendations for the methods under different situations.**

***Conclusions*: DE analysis methods should be chosen for scRNA-seq data with great caution with regard to different situations of data. Different strategies should be taken for data with different sample sizes and/or different strengths of the expected signals. Several methods for scRNA-seq data show advantages in some aspects, and DEGSeq tends to outperform other methods with respect to consistency, reproducibility and accuracy of predictions on scRNA-seq data.**

Keywords: single-cell; RNA-Seq; differential expression

INTRODUCTION

In recent years, RNA sequencing (RNA-Seq) technology has been widely used for studying transcriptomes [1]. Standard RNA-Seq experiments need millions of cells for sequencing [2,3], and therefore can only get averaged measurements of gene expressions of the cells sequenced. Many recent studies have shown that even phenotypically identical cells can have very different transcriptomic profiles [4,5]. Such heterogeneities between cells cannot be studied with standard RNA-Seq experiments [6].

The rapid development of technologies such as cell separation, selection and amplification of minimal amounts of mRNA has enabled the sequencing of RNAs from an individual cell [7]. This is called single-cell RNA-Seq or scRNA-seq. In contrast, the standard RNA-Seq technology that needs many cells is called bulk RNA-Seq. A typical workflow for scRNA-seq experiment is cell capture, cell lysis, reverse transcription, pre-amplification, library preparation and sequencing [7]. With the development of new technologies, scRNA-seq has become a more and more popular technology to study many questions that cannot be addressed by bulk RNA-Seq, such as investigating transcriptome heterogeneities between individual cells, identifying novel cell types or cellular states and studying the transcriptomes of rare cell types [2,6–9]. An important task in the analyses of scRNA-seq data is to detect genes which are differentially expressed (DE) between individual cells or clusters of cells, and defining marker genes from the most differentially expressed genes [2].

The study of differentially expressed genes (DEG) has

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

been a major theme in transcriptome studies. Many methods have been developed for the detection of DEG based on bulk RNA-Seq data, such as DESeq [10], edgeR [11], DEGseq [12]. Recently there have been some DE analysis methods specifically designed for scRNA-seq data [13–17], but there are still many single-cell studies using existing DE analysis methods developed for bulk RNA-Seq on scRNA-seq data [18–23]. However, scRNA-seq has many characteristics that are different with bulk RNA-Seq data. For example, scRNA-seq data tend to be more noisy than bulk RNA-Seq data, which is caused by the tiny amount and low capture efficiency of mRNA molecules in single cells [4]. In typical cases, the tiny amount of starting mRNA in single cells will make it more likely to randomly miss some transcripts during reverse transcription. These missed transcripts will not be detected in the following sequencing step. This phenomenon exists widely in scRNA-seq and causes the so-called 'dropout' phenomena: a slight difference in gene expression or in some random factor may cause a gene to be undetected in one cell but have a moderate or high expression in another cell [13]. ScRNA-seq data also suffer more from problems in severe 3' bias, partial coverage and uneven depth than bulk data. It deserves systematic investigations that how well existing methods can be applied on scRNA-seq data.

Until now, there has been no published evaluation on the applicability and performance of existing DE analysis methods on scRNA-seq data. There are many existing methods one may use, and the choices under different scenario can be hard without a systematic comparison. In this paper, we performed a computational analysis systematically of 14 representative methods for detecting DEG on several large scale scRNA-seq datasets to quantitatively evaluate the characteristics and performance of the methods under different scenario. We found that results from different methods can be very different in general. Some methods are more suitable than others for certain situations with regard to sample sizes and degree of difference between the compared groups, and the reproducibility of each method on different subsamples of the data also varies. These comparisons can be a basic reference for choosing existing methods in a particular study, and also suggest directions for the future development of DE analysis methods that are more suitable for scRNA-seq data.

RESULTS

The methods we evaluated include SCDE [13], monocle [14], D3E [15], BPSC [16], DESeq [10], edgeR [11], baySeq [24], NBPSeq [25], Cuffdiff [26], DEGseq [12], TSPM [27], limma [28], ballgown [29] and SMAseq [30] as shown in Table 1. The first 4 methods were designed for scRNA-seq data specifically while the rest methods were developed based on bulk RNA-seq data. DE analysis methods can be divided into parametric methods and non-parametric methods according to whether assuming the data come from a parameterized probability distribution. Most of the methods we chose are parametric methods, while D3E and SAMseq are the two representative non-parametric methods we chose. The D3E method is based on test of distribution and the SAMseq method is based on resampling. According to the type of models used, parametric methods can be mainly divided into negative binomial model (SCDE, DESeq, edgeR,

Table 1. Information of gene differential expression analysis methods used.

Method	Model	Input	Platform	Threshold	Run time	Ref.
SCDE	Poisson and negative binomial model	Read counts matrix	R(package)	*p*-value	Minutes	[13]
monocle	Generalized additive models	Read counts matrix	R(package)	*p*-value	Minutes	[14]
D3E	Non-parametric (test of distribution)	Read counts matrix	Python(package)	*p*-value	1 hour	[15]
BPSC	Beta-Poisson model	Read counts matrix	R(package)	*p*-value	1 hour	[16]
DESeq	Negative binomial model	Read counts matrix	R(package)	*p*-value	Minutes	[10]
edgeR	Negative binomial model	Read counts matrix	R(package)	*p*-value	Minutes	[11]
baySeq	Negative binomial model	Read counts matrix	R(package)	Likelihood	12 hours	[24]
NBPSeq	Negative binomial model	Read counts matrix	R(package)	*p*-value	Minutes	[25]
Cuffdiff	Beta negative binomial model	Sam file	Linux	*p*-value	13 hours	[26]
DEGseq	Poisson model	Read counts matrix	R(package)	*p*-value	Minutes	[12]
TSPM	Poisson model	Read counts matrix	R(script)	*p*-value	1 hour	[27]
limma	Linear models	Read counts matrix	R(package)	*p*-value	Seconds	[28]
ballgown	Nested linear models	Read counts matrix /ctab file	R(package)	*p*-value	Seconds	[29]
SAMseq	Non-parametric (resampling)	Read count matrix	R(package)	*p*-value	Minutes	[30]

Run time is measured by one experiment of 40 samples vs 40 samples, and the used parameters and settings are shown in the materials and methods part.

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

baySeq, NBPSeq, Cuffdiff), Poisson model (BPSC, DEGseq, TSPM) and linear model (monocle, limma, ballgown). Almost all the methods' input is read counts matrix stored in a txt file, except Cuffdiff's input is sam format file [31] and ballgown's input could be either read counts matrix or a ctab format file generated by Tablemaker [29], which is more flexible. Most of the methods' thresholds for calling DEG are based on *p*-values, except that baySeq uses threshold on the likelihood but it also provides adjusted *p*-values for DE in result table.

We first used dataset GSE48968 from Gene Expression Omnibus (GEO) in our study. It contains scRNA-seq data of more than 1,700 primary mouse bone-marrow-derived dendritic cells, and has an average depth of 4.5±3.0 million read pairs per sample [32]. We used data of four groups of cells from the dataset. Three of them were stimulated with lipopolysaccharide (LPS, a component of Gram-negative bacteria) for 4 hours. We call them the group of stimulated cells with biological replicates (SBR), stimulated cells with technical replicates group 1 (STR1) and stimulated cells with technical replicates group 2 (STR2), respectively. And the fourth group was unstimulated cells with biological replicates (UBR). We then used another two scRNA-seq datasets for confirmation and verification of our observations. They were measured with different protocols and have different sequencing depths [33,34]. The information of all data we used are summarized in Table 2 and more details can be found in Materials and Methods. In all the scRNA-seq data we used, we observed that detected expression levels of 95%–99% genes are 0 in the data of one cell, and those of 75%–90% genes are 0 in all cells of the same group (groups in Table 2). This shows that dropout events are widespread in all scRNA-seq data. And we also observed strong 3' bias, partial coverage and uneven depth in all three datasets.

The main idea of our experiments is to use each of the 14 methods to detect DEG under different experimental settings with regard to the type of samples to be compared and the number of samples to be compared. We applied the 14 methods to detect DEG between groups and between subgroups within each group to study their performances when comparing samples between different treatments, between biological replicates of the same treatment, and between technical replicates. All the comparisons were conducted using different sample sizes of 1 sample vs 1 sample, 2 vs 2, 5 vs 5, 10 vs 10, 20 vs 20 and 40 vs 40, to see the impact of sample size on the results. For experiments of sample sizes from 1 vs 1 to 20 vs 20, we repeated each experiment 20 times by random sampling from the whole sample set to study the influences of randomness in the samples. For the experiment of 40 vs 40 samples, we only conducted the experiment once for between group comparison with each method due to the limitation of the total number of samples in the dataset. In total, we designed 7 sets of main experiments on dataset GSE48968 to analyze the performance of the 14 methods on scRNA-seq data, and designed another 9 sets of experiments on the other two datasets for confirmation and verification. The 7 sets of main experiments are summarized in Table 3 and the information of the other 9 sets of experiments for verification are summarized in Supplementary Table S1. The first 6 sets of main experiments on dataset GSE48968 were direct comparison between or within groups. The 7th set of experiment was carried out using two fake subgroups of samples with reads randomly extracted from a same sample set, which was designed for assessing the false discoveries of each method. To eliminate the possible influence of overlapping samples between replicate experiments in the evaluation of the reproducibility of each method, we also designed extra experiments with mutual exclusive samples. The detailed design of the experiments are in Materials and Methods.

Numbers of differential expression genes detected

We first studied the number of DEG detected by different methods in all the experiments. Table 4 summarizes the average number of DEG of each experiment by each method. The threshold used were adjusted *p*-value of false

Table 2. Information of samples for experiments.

Dataset	Cell type	Group	Group description	Sample size	Ref.
GSE48968	Mouse bone-marrow-derived dendritic cells	SBR	LPS stimulation for 4 h, biological replicate	96	[32]
		STR1	LPS stimulation for 4 h, technical replicate 1	81	
		STR2	LPS stimulation for 4 h, technical replicate 2	56	
		UBR	Unstimulated, biological replicate	96	
GSE59127	Mouse kidney cells	E11.5	Embryonic day 11.5 total kidney	49	[33]
GSE59129	Mouse kidney cells	E12.5	Embryonic day 12.5 total kidney	86	
GSE59130	Mouse kidney cells	P4	Renal vesicle cells from post-natal day (P4) kidneys	57	
GSE74923	Mouse CD8 + T-cells	CD8	Activated murine CD8 + T-cells	106	[34]
	Mouse lymphocytic leukemia cells	L1210	Lymphocytic leukemia cell line lineages	88	

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

Table 3. Information of experiments on dataset GSE48968.

Experiments	Experiments type	Abbreviation	Group 1	Group 2
1		SBR_v_UBR	SBR	UBR
2	Between group comparison	STR1_v_UBR	STR1	UBR
3		STR1_v_STR2	STR1	STR2
4		SBR_v_SBR	SBR	SBR
5	Within group comparison	UBR_v_UBR	UBR	UBR
6		STR1_v_STR1	STR1	STR1
7	Identical comparison	SBRsplit	SBR(split)	SBR(split)

discovery rate (FDR) < 0.05. Because SCDE, D3E, BPSC, edgeR, NBPSeq, limma, ballgown and SAMseq could not do the analysis of one sample vs one sample [11,13,15,16,25,28–30], we use NA in the table to indicated that the method cannot work in that particular experiment.

As we can see in Table 4, the average numbers of DEG found by the 14 methods differ very much. Although the FDR threshold was set at the same level, the numbers of reported DEG can be at different magnitudes. Most methods reported nothing to several hundred DEG in most experiments, like SCDE, monocle, D3E, BPSC, DESeq, edgeR, Cuffdiff, ballgown and SAMseq while other methods such as baySeq, NBPSeq, DEGseq, TSPM and limma could call up to thousands of DEG. According to the numbers of DEG they found, we can divide the methods into three categories. The first category includes SCDE, monocle, D3E, BPSC, DESeq, edgeR, Cuffdiff, ballgown and SAMseq which report relatively small numbers of DEG (usually less than a few hundreds). The second category includes NBPSeq, DEGseq and limma which tend to report many DEG (from hundreds to thousands). BaySeq and TSPM are of the third category, which can report numbers of DEG from nothing to several thousands.

The manners of the changes of each method's DEG when sample size increases also differ very much. SCDE, monocle, BPSC, DESeq, edgeR and SAMseq report zero or few DEG when sample size is very small and the numbers of reported genes increase gradually when sample size increases. The numbers of DEG found by baySeq decrease first and then increase sharply. Similarly, the numbers of DEG found by TSPM decrease rapidly and sometimes increase again when sample size increases. Because of the high variation of results when sample size increases, baySeq and TSPM tend to have low consistency with scRNA-seq data. The numbers of DEG found by Cuffdiff decrease first and then increase; the DEG NBPSeq and DEGseq detected increase gradually when sample size increases, however the numbers of DEG found by D3E, limma and ballgown seems to not have a coherent pattern under such condition. According to the manner that the numbers of DEG change when sample size increases, we could classify these methods into three categories: the increasing category, decreasing-then-increasing category and inconsistent category. The increasing category includes SCDE, monocle, BPSC, DESeq, edgeR, NBPSeq, DEGseq and SAMseq for the DEG they find will gradually increase when sample size increases. The decreasing-then-increasing category contains baySeq and Cuffdiff. The inconsistent category includes D3E, TSPM, limma and ballgown, for the DEG they find do not have a coherent pattern across the experiments.

In the SBRsplit experiment which was designed for studying false positive detections by every method, we found that most of methods report zero DEG. It is as expected because the compared samples in this experiment were composed of reads randomly extracted from the same sample and there should not be any DEG. All DEG reported in this experiment are false positives. In our results, most methods give almost no false discoveries in this experiment when sample size is not too small, except the NBPSeq method that give many false positive. When sample size is small, monocle, ballgown and Cuffdiff also report some false positives.

The DEG found by ballgown show an inconsistent pattern between different experiments (Table 4), which seems to be caused by the special pattern of ballgown's *p*-values. The histogram of *p*-values of all genes analyzed by ballgown contains randomly distributed pulses and the number of pulses is approximately less than or equal to the number of samples compared (Supplementary Figure S1). This phenomenon leads to the number of DEG found by ballgown shows an unstable pattern. We speculate that the special pattern of ballgown's *p*-value is caused by its nested linear models [29].

The number of DEG Cuffdiff detected in almost every 2 vs 2 experiments equals to zero (Table 4), which is very strange and seems to be largely determined by the distribution of its *p*-values near zero. The distribution of Cuffdiff's *p*-values of all genes near zero differs a lot between different sample sizes. The *p*-values near zero of 2 vs 2 experiments are much lower than those of

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

Table 4. Average numbers of differential expression genes detected — adjusted *p*-value (FDR) < 0.05.

DE genes mean	SCDE	monocle	D3E	BPSC	DESeq	edgeR	baySeq	NBPSeq	Cuffdiff	DEGseq	TSPM	limma	ballgown	SAMseq
SBR_v_UBR-1v1	NA	0	NA	NA	0	NA	128.6	NA	62.1	2663.4	2755.4	NA	NA	NA
SBR_v_UBR-2v2	0.4	0.2	127.2	0	2	0	1.1	899.6	7	3815.8	137.1	268.6	27.9	0
SBR_v_UBR-5v5	31.8	10.9	0	2.1	10.6	9.6	38.2	1549.9	252.1	5711.8	5.5	1592.8	17.4	4.7
SBR_v_UBR-10v10	120.4	35.3	172.7	172.2	169.2	35.8	1554	1991.4	439	7612.2	11.8	3235.4	109.7	152.8
SBR_v_UBR-20v20	253.1	108.2	688.5	692.2	277.4	161.8	5972.2	2572.9	627.5	9972.3	133.5	8250.6	250.5	490.4
SBR_v_UBR-40v40	487	276	1267	1380	458	2209	8537	3430	956	12862	496	13834	511	963
STR1_v_UBR-1v1	NA	0.3	NA	NA	0	NA	32.3	NA	47.5	2172.8	2306.7	NA	NA	NA
STR1_v_UBR-2v2	1.4	17.8	768.1	0	6.2	0	1	862.3	15.7	3100.1	109.8	243.1	125.3	0
STR1_v_UBR-5v5	38.6	8.3	0	14.1	20.1	13.7	35.8	1563.8	170.2	4664.4	10.7	161.2	15.5	53.2
STR1_v_UBR-10v10	129.9	59.6	86.5	362.2	276.9	39	375.4	1988	266.2	6191	40.5	803.9	75.5	204.4
STR1_v_UBR-20v20	287.8	188.8	386.9	756.2	462.6	143.4	4619.6	2534.3	393.4	7979.8	183.8	2896.1	218.5	543.1
STR1_v_UBR-40v40	630	412	787	1373	804	539	6660	3205	600	10285	503	6877	516	1077
STR1_v_STR2-1v1	NA	0.1	NA	NA	0.2	NA	3.1	NA	100.2	1762.7	1850.4	NA	NA	NA
STR1_v_STR2-2v2	0	10.9	84	0	0.6	0	0	695.5	0	2523.4	70.1	113.8	54.5	0
STR1_v_STR2-5v5	0	0.9	0	0.3	0.2	0	0.4	1135.9	113.3	3710.3	0.6	20.6	0.1	0
STR1_v_STR2-10v10	0.9	1	0	75.5	10.8	0	13.2	1405	129.2	4847.1	0	256.1	0.2	3.4
STR1_v_STR2-20v20	11.3	1.3	0.4	172.1	4.5	0.2	2821.3	1674.2	170.6	6124.2	0	2029.3	0.6	2.5
STR1_v_STR2-40v40	21	10	23	281	14	0	3952	1930	180	7590	1	5318	11	6
SBR_v_SBR-1v1	NA	0.3	NA	NA	0.2	NA	171.6	NA	126.5	3004.9	3119.8	NA	NA	NA
SBR_v_SBR-2v2	0.2	0.7	344.8	0	0.7	0	0.2	837.4	0	4128.1	116.2	168.5	66	0
SBR_v_SBR-5v5	0.2	1.3	0	0.1	0.2	0	1.4	1037	230.2	6244.1	0.7	329.6	25.5	0
SBR_v_SBR-10v10	0.2	1	0	24.6	15.2	0	137.9	1273.9	342.5	8312	0	1223	0	1.1
SBR_v_SBR-20v20	6.8	0.8	0	281.6	5.5	0.1	5772.6	1261.8	442.9	10945.2	0	1502	0	1.6
UBR_v_UBR-1v1	NA	0.2	NA	NA	0	NA	22.4	NA	110.7	2301.6	2444	NA	NA	NA
UBR_v_UBR-2v2	0	0.6	235.9	0	0.2	0	0	875.1	0	3314.8	78.2	131.7	46	0
UBR_v_UBR-5v5	0	0.7	0	0.1	0.2	0	0.8	1469.6	149.6	5024.9	0.8	12.6	0.4	0
UBR_v_UBR-10v10	0.4	0.4	0	49.2	7	0.1	58.3	1847	201.2	6646.9	0	17.1	0	12.6
UBR_v_UBR-20v20	11.1	0.6	0	153.1	1.8	0	4414.4	2202.9	245.8	8533.8	0	139.9	0	4.8
STR1_v_STR1-1v1	NA	0.2	NA	NA	0.2	NA	2.2	NA	101.7	1801.2	1899	NA	NA	NA
STR1_v_STR1-2v2	0	4	134.6	0	0.1	0	0	619.2	0	2508.4	65	103.8	14.2	0
STR1_v_STR1-5v5	0	0.6	0	0.2	0.2	0.2	0.4	1037.9	95.3	3708.8	0.4	6.1	0.3	0
STR1_v_STR1-10v10	0.2	1.1	0	67.8	6.6	0	7.5	1300.4	106	4766.9	0	24.1	0	2
STR1_v_STR1-20v20	6.1	0.4	0	153.9	1.8	0	2719.1	1538.5	137.1	6035.4	0	144.9	0	1.6
SBRsplit_v_SBRsplit-1v1	NA	53.5	NA	NA	0	NA	0	NA	12.7	0	0	NA	NA	NA
SBRsplit_v_SBRsplit-2v2	0	43.9	0	0	0	0	0	372	0	0	0	0.1	30.8	0
SBRsplit_v_SBRsplit-5v5	0	0.8	0	0	0	0	0	1357.4	0	0.7	0	0	0	0
SBRsplit_v_SBRsplit-10v10	0	1	0	0	0	0	0	2498.7	0	0.6	0	0	0	0
SBRsplit_v_SBRsplit-20v20	0	1	0	0	0	0	0	3929.8	0	1	0	0	0	0

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

experiments with other sample sizes (Supplementary Figure S2), which determines that Cuffdiff finds the least number of DEG when the comparison sample size is 2. Cuffdiff uses the beta distribution to model the uncertainty in the assignment of fragments to isoforms and uses negative binomial distribution to model over-dispersion of fragment counts, and then it combines them to a single model of fragment count variability [26]. It is likely that this special characteristic of Cuffdiff's *p*-values' distribution is caused by the mixture of beta distribution and negative binomial distribution. Anyway, the strange pattern of Cuffdiff results implies that it may not be very reliable for comparison of two samples with two samples on scRNA-seq data. The histograms of other methods' *p*-values are also shown in Supplementary Figures S3–S12 , except for baySeq and SAMseq that do not output original *p*-values.

In summary, the number of DEG detected by the 14 methods differs very much in their scales and the manners they change when sample size varies. According to experiments on the SBRsplit data, all methods except NBPSeq can avoid false positive detections well on completely non-separable data. The number of DEG found by baySeq and TSPM varies at a very wide range when sample size changes, and ballgown and Cuffdiff show some unreasonable patterns in the change of DEG numbers with regard to sample size. The other methods, albeit the big difference between results of different methods, perform in a predictable manner when sample size changes. When we change the FDR cutoff from 0.05 to 0.04, 0.03, 0.02 and 0.01, we found that the numbers of DEG detected decrease with the cutoff as expected, but the scale and the changing manners of DEG with sample sizes remain unchanged (Supplementary Tables S2–S5), suggesting that our observations are not specific to a particular cutoff. We got similar results on the other two datasets (SupplementaryTables S6–S10).

Similarities between results of different methods

The *p*-values of different methods are not directly comparable, so we used the number of genes in the intersection of top 1,000 DEG (ranked by ascending *p*-values) of two results as a measure of their similarity. Figure 1 shows the similarity heatmaps illustrating the intersection numbers between the compared methods in the experiments with different sample sizes. The numbers were averaged among the 20 replicated experiments with the same settings except sample size 40.

When analyzing the similarity among results from different methods, we should keep in mind that people had observed even for data with a bulk of cells obtained with RNA-Seq or microarrays, it is not unexpected that two results on the same data only have about half overlap in some scenarios [35–37]. Given this context, we can understand that an intersection number of ~500 genes among the top 1,000 genes can be seen as an indication of reasonable similarity. We can see from Figure 1 that, when sample size is small (1 vs 1), the methods that can work for this scenario (monocle, DESeq, baySeq, Cuffdiff, DEGseq and TSPM) give pretty similar results except for Cuffdiff. When sample size increases to 2 vs 2 and 5 vs 5, the similarity among results of DESeq, edgeR, NBPSeq, DEGseq, limma and SCDE are reasonably high (around or above 50%). Considering the observation on the dramatic difference in the number of reported DEG between methods, we can see that a large part of the difference can be caused by the different estimate of *p*-values by different models and algorithms. When the sample size further increases to 10 vs 10, 20 vs 20 and 40 vs 40, the results of DEGseq is becoming less similar with those of edgeR and DESeq, which implies that the difference between the simple Poisson model (used by DEGseq) and the negative binomial model (used by edgeR and DESeq) becomes more significant when sample size is getting large. And when sample size get larger, the results of methods specifically designed for scRNA-seq (SCDE, monocle, D3E, BPSC), DESeq and SAMseq become more similar to each other. DEGseq and BPSC have quite high similarity all the way because both of them use the Poisson model. The similarities between limma and other methods are quite high when sample size is 2 vs 2 but drop rapidly when sample size increases. This implies what limma uses a very different model with other methods and the difference between models and implementations becomes more obvious when sample sizes increase. Similar results of similarities between different methods are also seen in other experiments and the other two datasets (Supplementary Figures S13–S26).

We also studied the situations that using the top 500 DEG and top 100 DEG separately. We found that the similarity and the relative order of similarity of different methods almost remain the same for top 500 genes as with top 1,000 genes (Supplementary Figures S27–S56). For the top 100 genes, the similarity between methods decreased some in general, and some of the relative orders of similarities also changed. This inconsistent result is mainly because the used gene number is too small and the influences of stochastic factor become larger.

Consistency of the same methods with different data sizes

We define the consistency of a method as the overlap of top 1,000 DEG (ranked by ascending *p*-value) it found between results on experiments with different sample sizes. This can be illustrated as consistency heatmaps in a similar way as in Figure 1 for each method and for each

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

set of experiments. Figure 2A shows only the example of consistency heatmaps of edgeR, DESeq and DEGseq on the experiment SBR_v_UBR. All the consistency heatmaps of all experiments and methods are given in the Supplementary Figure S57.

By eyeballing through all these consistency heatmaps in Supplementary Figure S57, we can get the general feeling that some methods have high consistency in most experiments, and some only have high consistency in certain datatypes. But it is not straightforward to make any direct comparison between methods. To make the observations more comparable, we define a "mean consistency measure" for each method in the following way: firstly, we performed a union of the top 1,000 DEG found with sample size 2, 5 and 10. Then, we found the intersection on this union list with the top 1,000 DEG found with sample size 20, and counted the number of genes in the intersection and divide it by 1,000 to get the percentage of intersection. We did these steps on all the 20 replicated experiments in each setting and define the average of the percentages of intersection as the mean consistency measure of the method on each type of data. The mean consistency measures of all 14 methods on the 6 informative experiments are shown in Figure 2B. Figure 2C shows the mean consistency measures of all methods that are further averaged over experiments on different data types.

From heatmap in Figure 2B, we can have a few interesting observations. We can see that the consistency of most methods are good in general considering that a ~50% consistency between results two parallel datasets of bulk expression data is already reasonable. Some methods like DEGseq and D3E have higher consistency than others. Because the comparison is based on the top 1,000 DEG regardless of whether they are called as significant according to some cutoff, this consistency measure is not affected by whether a method tends to call more or less positives. It is interesting to note that the experiments on the comparison of SBR vs UBR samples and STR1 vs UBR samples show relatively higher consistency for all methods, and the comparison of subgroups of SBR samples (SBR vs SBR) has the lowest consistency for all methods. For SBR_v_UBR and STR1_v_UBR experiments, the dominating difference between the two groups are whether the cells were stimulated or untreated. The observation tells that this signal is strong so that the consistency of discoveries from samples of difference size is high. On the other hand, the samples in the two groups of the SBR_v_SBR experiment are actually cells from the same treatment. The reported DEG between the two groups are mostly due to expression heterogeneity between the cells. This explains why the consistency between results on different sample sizes is low. Similar observations on the other two datasets can be seen in Supplementary Figure S58. In Supplementary Figure S58B, the mean consistency measure of experiment CD8_v_L1210 is significantly higher than other experiments because the cell types for comparison are totally different. This verifies that the consistency of these methods will be higher when the difference between the compared samples is stronger.

We also use the top 500 DEG and the top 100 DEG to calculate the consistency measure. The consistency becomes slightly lower than that with the top 1,000 DEG, but the observed patterns and relatively orders of methods almost remain unchanged (see Supplementary Figures S59–S62).

Reproducibility of the methods

We define the reproducibility of a method as the average of the overlap of top 1,000 DEG (ranked ascending with *p*-value) it found between all pairs of the 20 replicated experiments with different random seeds for sampling divided by 1,000. So the reproducibility of a method in a specific experiment and specific sample size will be in the range of [0, 1]. This can be illustrated as reproducibility heatmaps in a similar way as in Figure 2B for each method, each sample size and each set of experiments. Figure 3A shows the reproducibility heatmaps of each method in each set of experiments of different sample sizes, and Figure 3B is an average over experiments of the six heatmaps in Figure 3A. Figure 3C is an average over sample sizes of the heatmap in Figure 3B.

From these heatmaps, we can have some interesting findings. We can see that in most situation, the reproducibility is in the range of [0.2, 0.4], which means the overall reproducibility of each method on scRNA-seq data tend to be in a low level. In general, the reproducibility will increase as sample size increases. Some methods like DEGseq and SCDE show higher reproducibility than others. Similar to consistency in Figure 2, because the comparison is based on the top 1,000 DEG regardless of whether they are called as significant according to some cutoff, this reproducibility measure is also not affected by whether a method tends to call more or less positives. It is also worth noting that the set of experiments SBR_v_UBR and the set of experiments STR1_v_UBR show higher reproducibility than other sets of experiments for all methods, similar with the observation of consistency in Figure 2. This reflects that the signal of differential expression of genes because of treatment is stronger and more stable than that caused by the heterogeneity between cells. We got consistent observations of reproducibility on the other two datasets (Supplementary Figure S63).

We also used the top 500 DEG and the top 100 DEG to calculate the reproducibility of each method. The reproducibility became smaller than that calculated with

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

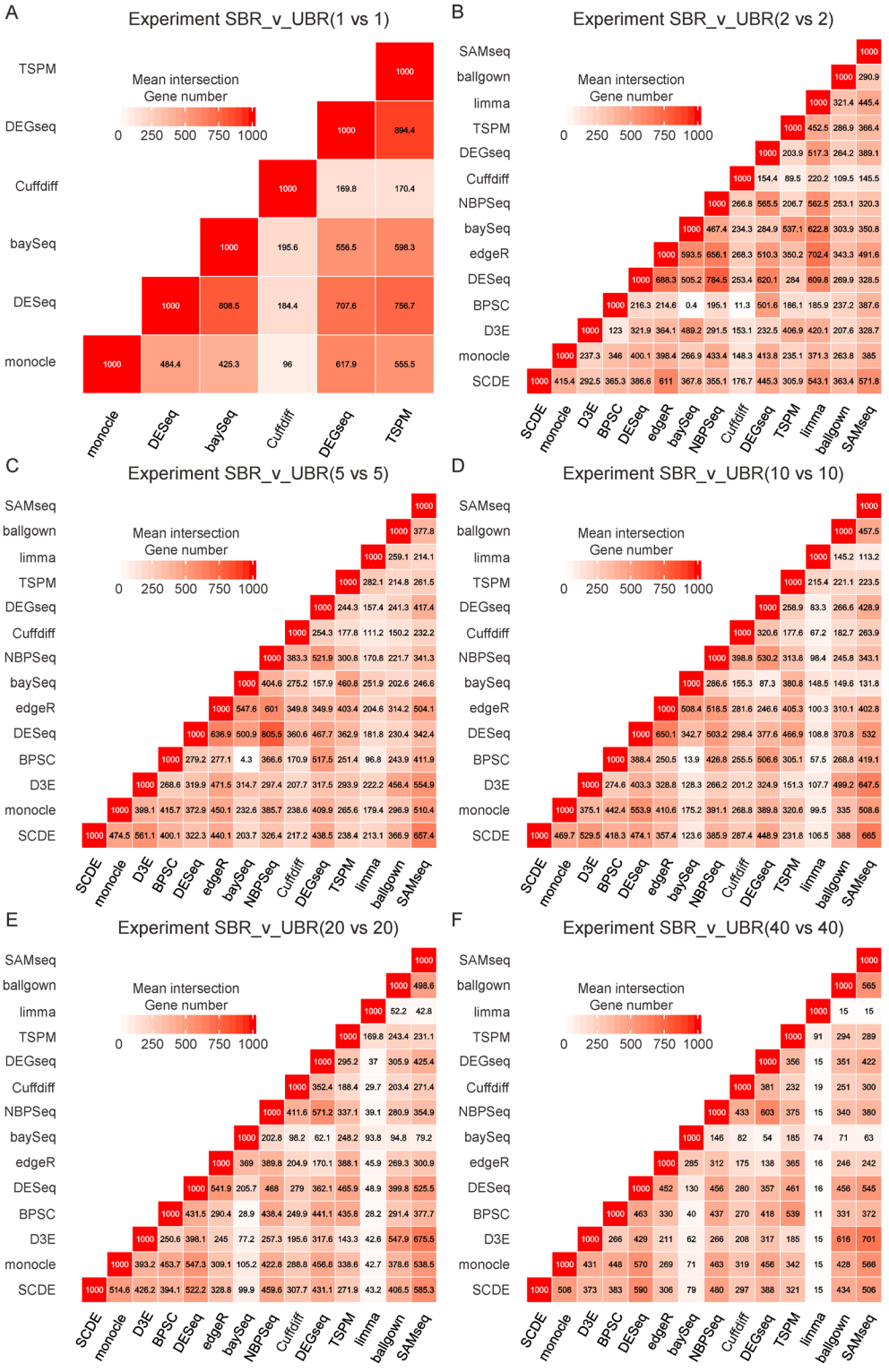


Figure 1. Mean intersection numbers of top 1,000 DEG between different methods of experiment SBR_v_UBR. Rows and columns stand for methods, and each cell of each table is a mean intersection number of top 1,000 DEG (ranked ascending with *p*-value) detected by the methods of row and column corresponding to respectively in 20 trials with random sampling. Sample numbers are 1 vs 1, 2 vs 2, 5 vs 5, 10 vs 10, 20 vs 20 and 40 vs 40 respectively from A to F. When sample number equals to 1, the table size is 6×6 because only monocle, DESeq, baySeq, Cuffdiff, DEGseq and TSPM could do the DE analysis of one sample versus one sample. And when sample number equals to 2, 5, 10, 20 or 40, the table size is 14×14 of all the 14 methods.

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

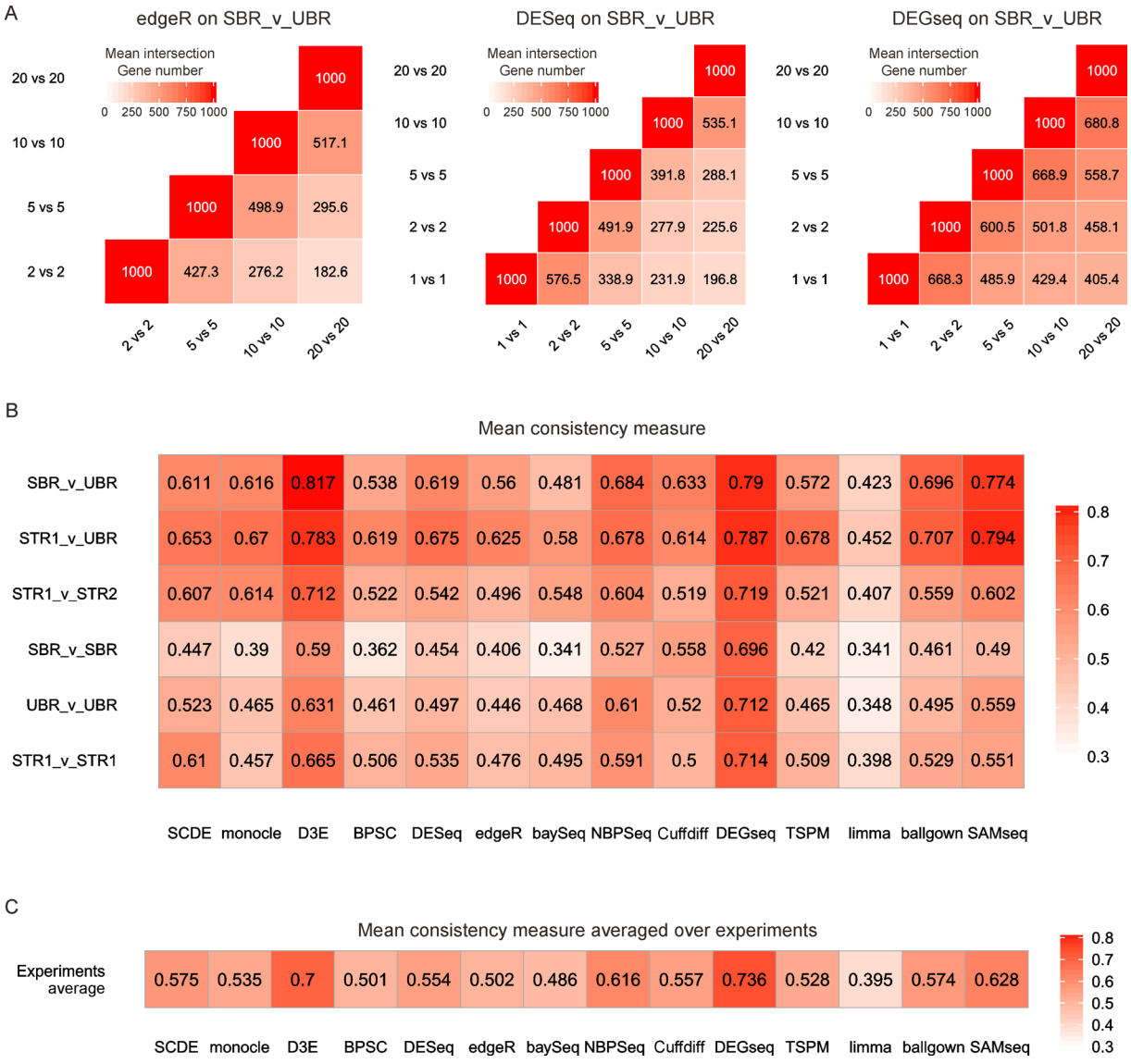


Figure 2. Consistency heatmaps. (A) Mean intersection numbers of top 1,000 DEG between different sample numbers. Rows and columns stand for sample sizes, and each cell of every table is a mean intersection number of the top 1,000 DEG (ranked ascending with *p*-value) detected by the specific method of the sample sizes row and column corresponding to respectively in 20 trials with random sampling. The method is edgeR, DESeq and DEGseq from left to right. For each method of each experiment, the table size is 4×4 or 5×5, which depends on the method whether could do the DE analysis of one sample versus one sample. (B) Mean consistency measure. The mean intersection percentage between the top 1,000 DEG of sample number 20 and the union of the top 1,000 DEG of sample numbers 2, 5 and 10. Rows stand for different experiments and columns stand for different methods. The number in each cell of the table is calculated by three steps: get the union of the top 1,000 DEG of sample number 2, 5 and 10 of a specific method and specific experiment; get the intersection of the union last step get with the top 1,000 DEG of sample number 20, and counted the number of genes in the intersection and divide it by 1,000 to get the percentage of intersection; repeat these steps on all the 20 replicated experiments in each setting and get the average. (C) Mean consistency measure averaged over experiments. This table is a column average of the table in Figure 2B.

the top 1,000 DEG, but the observed patterns and relatively orders between methods remained almost unchanged (Supplementary Figures S64–S67).

To eliminate the possible influences of overlapping samples between replicate experiments, we also designed parallel exclusive experiments whose samples are mutual exclusive to calculate the reproducibility (Supplementary Figure S68). We found that the reproducibility of each

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

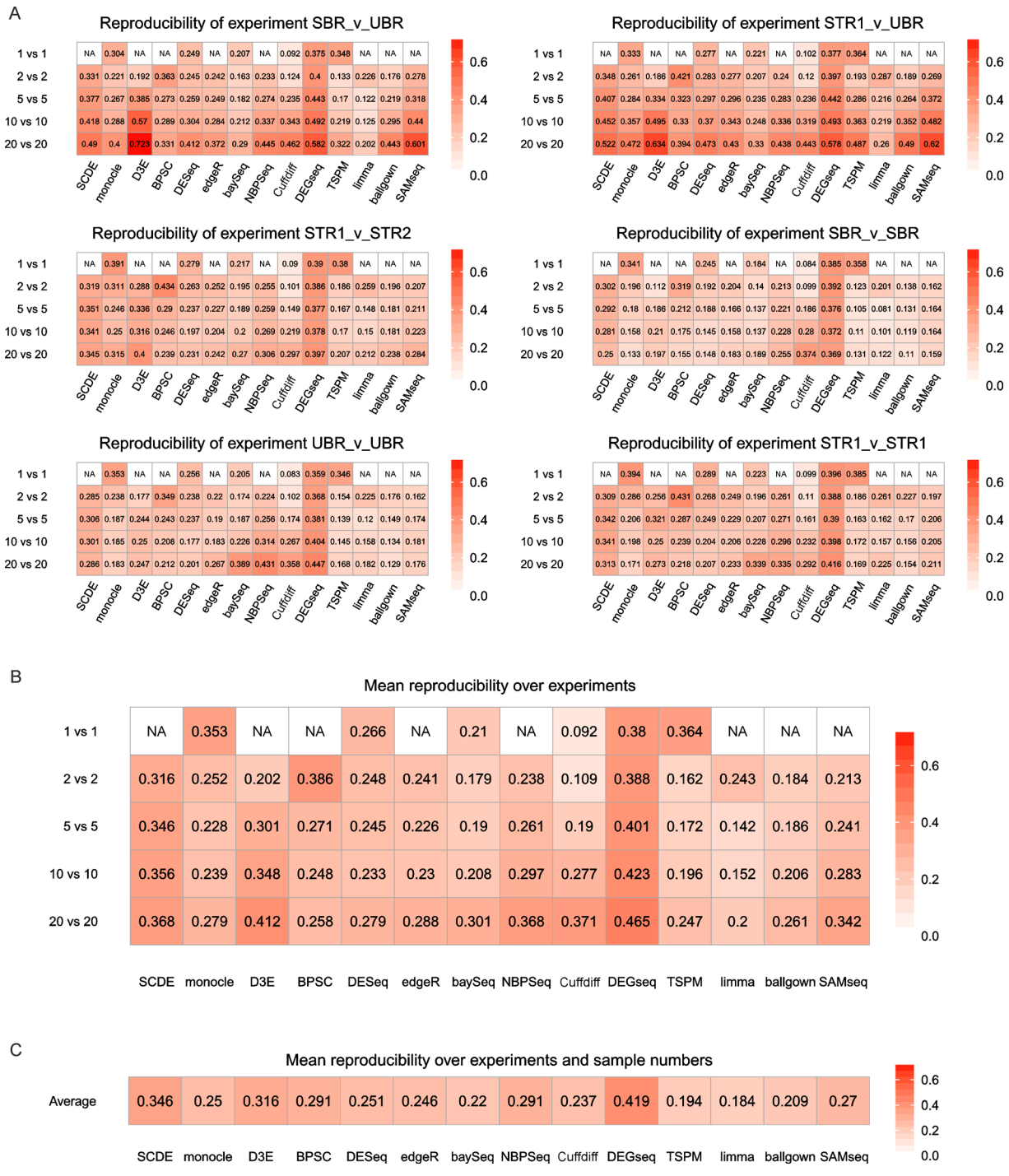


Figure 3. Reproducibility heatmaps. (A) Reproducibility of each method in each set of experiments of different sample sizes. Every table stands for a different set of experiments. In each table, rows stand for different sample sizes and columns stand for different methods. NA means that the method could not conduct the analysis of one sample versus one sample. (B) Mean reproducibility over experiments of each method. This table is an average of the six tables in Figure 3A. Rows stand for different sample sizes and columns stand for different methods. NA means that the method could not conduct the analysis of one sample versus one sample. (C) Mean reproducibility over experiments and sample numbers of each method. This table is a column average of the last four rows of the table in Figure 3B.

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

method still increases with sample size, which verifies our conclusions about reproducibility.

Accuracy of predictions indicated by ROC-like curves

A problem for studying DEG from single cell data is that we do not have access to the ground truth. Considering the heterogeneity in gene expression among single cells, which is the key reason why people want to use single-cell technology, it may even be difficult to validate results obtained on one group of samples by doing biological experiments on another group of samples before we have a good understanding on the nature of heterogeneity among the cells. Even the heterogeneity in technical replicated samples cannot be guaranteed due to imperfectness in the multiple steps of single cell experiments. Reliable simulation models can also not be built. In this study, we introduced the SBRsplit experiments on man-made samples that were generated by randomly extracting 50% reads of one scRNA-seq sample two times to two samples. We can be sure that any difference detected between two groups of artificial samples generated in this way must be false positive detections.

Using detections in the SBRsplit experiment as false positive discoveries drawing as the x-axis, and using the number of detected DEG in the SBR_v_UBR experiment to draw the y-axis, we generated ROC-curve-like plots to study the relative relations of false discoveries with probable true discoveries when the *p*-value cutoffs change from 0 to 1. The baySeq does not provide *p*-values and SAMseq reported no DEG in SBRsplit experiment even when threshold of FDR is set to 1, this curve cannot be drawn for them. The ROC-like curves and their AUC of the other 12 methods are given in Figure 4.

We can see from Figure 4 that most methods' ROC-like curves will become better when sample sizes increase and the curves of BPSC, DESeq, edgeR, Cuffdiff, DEGseq and limma are quite similar when sample size is 5 or larger. Among these 6 methods, BPSC, edgeR and limma cannot work on the 1 vs. 1 comparison; the area under the curve is small for DESeq and Cuffdiff when sample size is 1 or 2, which reflects that the estimation of the over-dispersion parameter is noisy when sample size is too small. The areas under the curves of the other methods are smaller than the above 6 methods, and there are some abrupt changes at some points of those curves. The reasons that cause those special curve shapes must be rooted in the models and algorithms used in those methods and deserve further investigations.

DEG found by DEGseq

As an example, we checked the DEG found by DEGseq using MA-plot visualization to have a direct feeling on the signals in the detected genes. Supplementary Figure S69 shows the example of 1 versus 1 results in the SBR_v_UBR experiment and in the SBRsplit experiment. The boundary for calling significant differential expression is defined by the *p*-value cutoff of FDR < 0.05. From these plot, we can perceive that the number of DEG between two samples is large.

We used heatmaps to visualize the expression of the detected DEG in the experiments with multiple samples. As there are too many detected DEG, we choose the top 50, middle 50 and bottom 50 among the detected DEG, according to the rank of the *p*-values. Supplementary Figure S70 shows the heatmaps in the SBR_v_UBR experiments with sample size 2, and those with sample sizes 5, 10 and 20 are given in Supplementary Figures S71–S73. We can observe that the heterogeneity of gene expression within each group is large, but a strong difference between the two groups can be perceived from the top 50 genes. For the middle 50 genes, the difference between the two groups is still obvious, but it tend to be due to one or two particular samples. And for the bottom 50 genes, the signals are weaker and often exist in only few or even single sample, which exhibits severe "dropout" phenomenon. Due to the large heterogeneity of gene expression within each group, it is questionable whether the bottom or even the middle genes in the list should be really regarded as having systematic differential expression although the *p*-values under the current model is small. This suggests that, when using methods like DEGseq that reports too many DEG based on signals in very few samples, we should set a more stringent threshold to only report the top genes for a more conservative result. And to solve this problem fundamentally, we should develop new methods based on novel models to deal with the severe 'dropout' events and the other particular properties for scRNA-seq data.

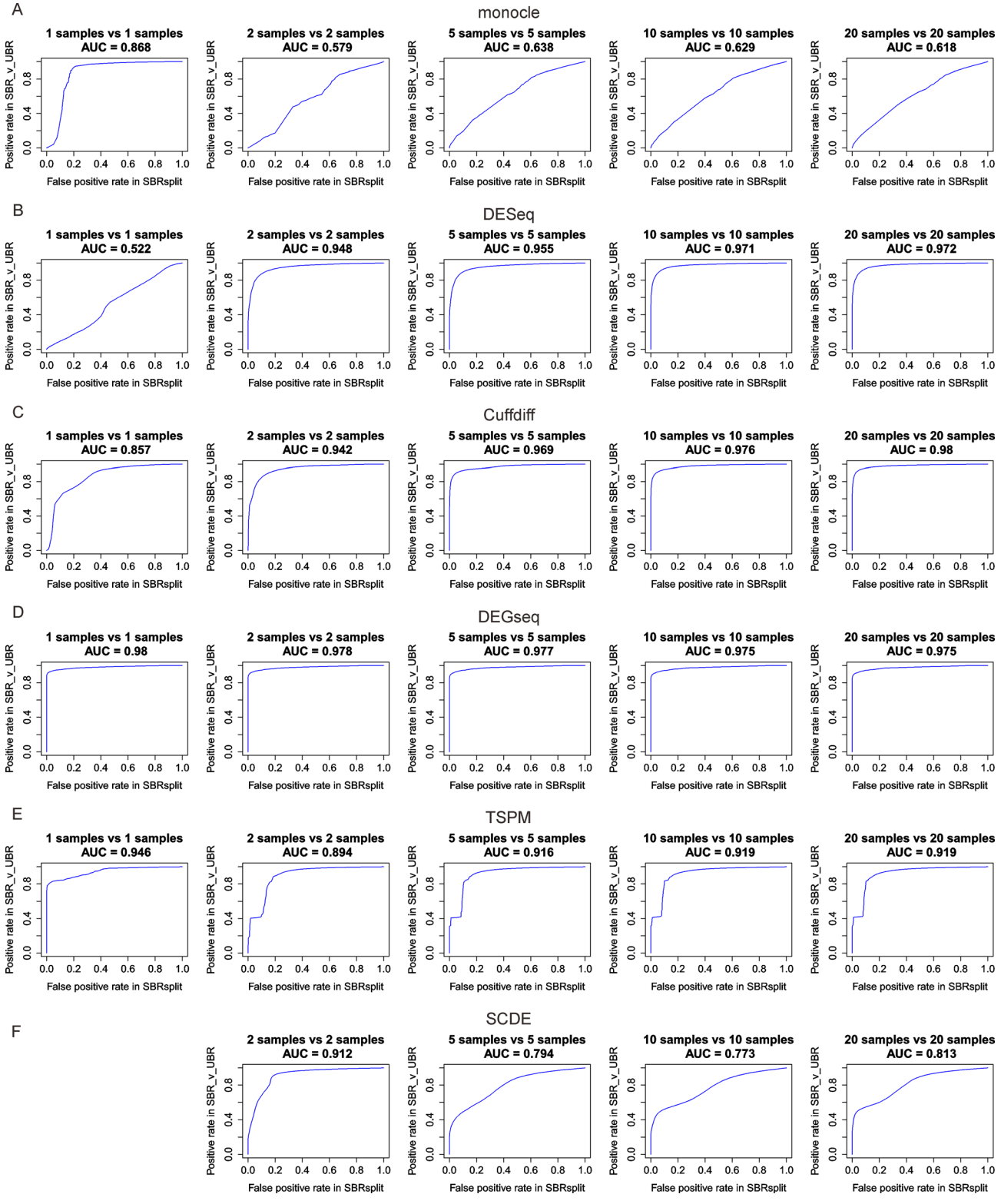
DISCUSSION AND CONCLUSIONS

In this study, after 16 sets of experiments on 3 scRNA-seq datasets with the 14 methods designed based on bulk RNA-seq data or scRNA-seq data, we observed that the methods can behave differently in the number of DEG each method tend to report, and also in the variation in this number when the sample size of the compared groups changes. Some methods tend to give very different reports at different experiment settings and may thus be less reliable. When comparing the similarities between results of different methods, we found that some methods give very similar results when sample size is small. The overall similarity between results of different methods drops when sample size increases, which implies that the difference between models and implementations becomes

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

more obvious. More work is needed to investigate what model can fit the scRNA-seq data better.

We assessed the consistency of each method on data of different sample sizes, and found that the consistency of most methods are in general good, especially when the difference between the compared samples is strong. DEGseq and D3E showed higher consistency than others in most experiments. We also checked the reproducibility



© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

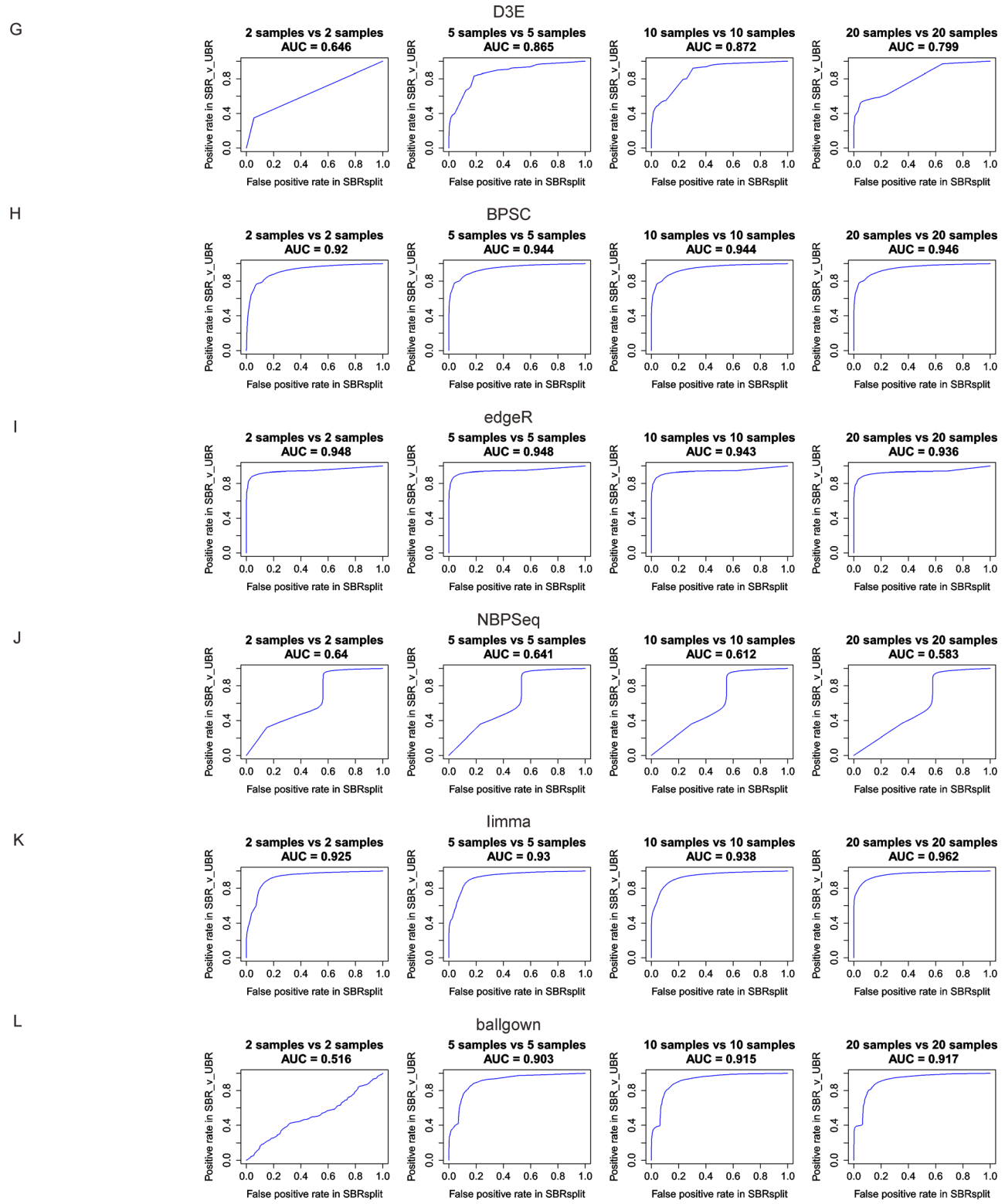


Figure 4. ROC-like curves of every method. The methods are monocle, DESeq, Cuffdiff, DEGseq, TSPM, SCDE, D3E, BPSC, edgeR, NBPSeq, limma and ballgown from A to H respectively. For every method, sample number is 1, 2, 5, 10, 20 from left to right. Some methods don't have the figure of one sample versus one sample because of nonsupport of 1 vs 1 comparison of the methods.

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

of the same method on two random subsets of the data with equal size. We found that random sampling of the samples can affect the detection of DEG, which highlights the high heterogeneity in gene expression among single cells. DEGseq has the highest reproducibility among the compared methods, and the reproducibility is higher for most methods when the difference signal between the two groups is stronger and when the sample size is larger. D3E and SAMseq's consistency and reproducibility are especially good on data with strong between-group differences and large sample sizes, but they perform just at average level on data with other situations. Meanwhile, we found that the methods developed for scRNA-seq data showed above-average or close-to-the-best consistency and reproducibility among all methods. This exhibits the advantages of the specifically designed methods over most of the traditional DE analysis methods.

We introduced an ROC-like curve and its AUC to quantitatively study the potential accuracy of each method at each sample size. We found that the performances of most methods are severely affected by the compared sample sizes and perform less satisfying overall when sample sizes are small. We saw that BPSC, DESeq, edgeR, Cuffdiff, DEGseq and limma perform well on their curves when sample size is not very small, and BPSC is the best among the 4 scRNA-seq specific methods. When there are only two samples in each group, edgeR, DEGseq and DESeq still perform well, and DEGseq outperforms other methods when sample size is one in each group. More detailed study on the DEG detected by DEGseq showed that it tends to be over-sensitive to genes that have expression in only few samples. In practice, more stringent threshold should be used to only report the top DEG when using DEGseq to avoid false positives caused by the abnormal expression in one or few samples, especially when the compared sample size is not very small.

From these comparative experiments in different experimental settings, we conclude that great caution should be taken with regard to the situation of data when choosing DE analysis methods for scRNA-seq data. Different strategies should be used for different sample sizes and different strengths of the expected signals. When the compared sample size is small, say less than or equal to 5, especially for the comparison of 1 vs 1, or when the compared DE signal is very weak, we recommend to use DEGseq, whose consistency and reproducibility both are high on scRNA-seq data. Since DEGseq tends to be a little oversensitive for some genes, a more stringent threshold is also recommended. For data with strong between-group differences and large sample sizes($\geqslant 20$), D3E and SAMseq will be good choices since their consistency and reproducibility are especially good for such scenario. For other situations, BPSC, DESeq and edgeR will be excellent choices for DE analysis of scRNA-seq data. But we also should keep in mind that, none of these methods perform consistently good for scRNA-seq data, these recommendations are only expedient in the existing methods. The low overall agreement between results with different methods and under different situations highlights the urgent need for models that can better capture the nature of single cell gene expression data.

Bulk RNA-Seq data have in general much smaller within-group variations than single-cell data due to the average effects of millions of cells. Therefore, methods for DEG detection on bulk data are all based on test of means. Due to the high heterogeneity of single cells, gene expression has higher variation among cells in the same group for single-cell data. In many cases, what we need to do for detecting differential gene expression between two group of single cells is to compare two distributions rather than only two means, like D3E does. Besides, the widespread 'dropout' events and other characteristics shown in scRNA-seq data should also be considered cautiously when conduct DEG analysis. The latest methods specifically designed for single-cell data showed advantages over most existing methods in some aspects, but there is no obvious winner that can perform the best in all major aspects. New models and new methods that can combine advantages of multiple methods need to be developed.

MATERIALS AND METHODS

Dataset

The dataset we mainly used in this study is from GEO of accession number GSE48968, which is from the paper published on *Nature* 2014 [32]. The dataset contains more than 1,700 primary mouse bone-marrow-derived dendritic cells' scRNA-seq data, and has an average depth of 4.5 ± 3.0 million read pairs. We used four groups of data from the dataset, they are LPS 4 h biological replicate, LPS 4 h technical replicate 1, LPS 4 h technical replicate 2 and unstimulated biological replicate respectively. To keep it simple, we rename them as SBR, STR1, STR2 and UBR respectively in our study, as shown in Table 2. The samples in group SBR, STR1 and STR2 are all stimulated with lipopolysaccharide (LPS, a component of Gram-negative bacteria) for 4 hours and the samples in group UBR are unstimulated. The sample sizes of the four group are 96, 81, 56, 96 respectively. And in order to control the quality of the sequencing data, we only use the samples whose coverages >1 million read pairs for experiment 1 to experiment 6, and the used sample sizes of SBR, STR1, STR2 and UBR become 80, 81, 55 and 57 respectively after quality control. And in experiment 7 (Experiment

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

SBRsplit), we only use the samples from group SBR whose coverages >2 million read pairs (77 samples) because in this experiment, samples will be extracted 50% reads two times for comparison.

The other two dataset we used in our study for confirmation and verification are also from GEO, and their accession numbers are GSE59127, GSE59129, GSE59130 and GSE74923 [33,34]. The average sequencing depth of GSE59127, GSE59129 and GSE59130 is 2.6 million read per cell, while the average depth of GSE74923 is 1.2 ± 0.06 million read pairs per cell. So we used the samples whose coverages >1 million read in GSE59127, GSE59129 and GSE59130, and the samples whose coverages >0.1 million read pairs in GSE74923. The other information of the two datasets is shown in Table 2.

Design of experiments

The main idea of experimental design is to use the 14 methods separately to detect the DEG between two groups of randomly selected samples with several sets of sample numbers. In experiment 1 (experiment SBR_v_UBR), we conduct the DE analysis between group SBR and group UBR, with sample number of 1 vs 1, 2 vs 2, 5 vs 5, 10 vs 10, 20 vs 20 and 40 vs 40 respectively. And the samples selected will have a nested relation when sample number increases, that is, when sample number is 1 vs 1, we just randomly draw out one sample from each group for comparison; when sample number comes to 2 vs 2, we keep the samples 1 vs 1 used and randomly draw one new sample from each group then add them to the comparison; when sample number is 5 vs 5, we keep the samples 2 vs 2 used and randomly draw three new samples from each group then add them to the comparison; and so on. And every experiment is repeated for 20 times except for 40 vs 40 because of the limitation of sample number of each group in the datasets, using different random number seeds for random sampling in R. The design of nested relation between different sample numbers is for studying the consistency of every experiment when sample number increases. Similarly, experiment 2 (experiment STR1_v_UBR) is group STR1 vs UBR. Experiment 3 (experiment STR1_v_STR2) is group STR1 vs STR2. Experiment 4 (experiment SBR_v_SBR) is group SBR vs SBR. Experiment 5 (experiment UBR_v_UBR) is group UBR vs UBR. Experiment 6 (experiment STR1_v_STR1) is group STR1 vs STR1. Experiment 7 (experiment SBRsplit) is a little different from the others. The design of experiment 7 is as following. Firstly, we randomly draw specific number of samples for comparison from group SBR, which also have the nested relation between different sample numbers. Secondly, we randomly extract 50% reads two times from the samples chosen, so that we could get two set of small samples, between which there would not exist any DEG. Thirdly, we carry out the detection of DEG between the two set of small samples and the result of significant DEG could be used as false positives of the method. The information of every experiment is listed in Table 3.

The other 9 sets of experiments on the other two datasets for confirmation and verification are summarized in Table S1. Their experimental settings for sample size and replicated times are same with the main experiments.

To eliminate the influences of overlap of samples between replicate experiments when we evaluate the reproducibility of each method, we also designed parallel exclusive experiments whose samples are mutual exclusive. And considering the impact of the needed sample number, we designed the parallel exclusive experiments only for the experiments of between group comparison, i.e., experiment SBR_v_UBR, experiment STR1_v_UBR, experiment STR1_v_STR2, experiment E11.5_v_E12.5, experiment E11.5_v_P4, experiment E12.5_v_P4 and experiment CD8_v_L1210. Concretely speaking, for these experiments, we conduct each trial twice and guarantee the samples used in them are mutual exclusive, and we call them a pair of exclusive trials. And we get the intersection number between the top 1000 DEG of each trial in the pair of exclusive trials as its reproducibility. Then we repeat each pair of exclusive trial 10 times, using different random seeds for randomly sampling, to calculate their average reproducibility. Then similarly we calculate the average reproducibility for every sample size of every experiment.

Sequence alignment and gene read counts

All sequencing data is mapped to the mouse genome (mm9, NCBI Build 37) using Tophat (v2.0.12) [38] with default parameters except the parameter for number of threads (‘-p 10’). Every alignment file of group SBR is randomly extracted 50% aligned reads for two times to two new alignment files using samtools [31] view with parameters of ‘-h-s 1.50’ and ‘-h-s 2.50’. All alignment files are sorted by read names using samtools sort with parameter ‘-n’. Then the alignment files and a GTF file of mm9 Ensembl Genes from UCSC are used to generate read count matrix for every sample using HTSeq (version 0.6.0)[39] with the following parameters: ‘python-m HTSeq.scripts.count-s no-f bam-r name’. All the alignment files in bam format are converted to sam format using samtools view for the following use of Cuffdiff.

Analysis of differential expression

The read count matrixes got from HTSeq last step are

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

used for analysis of DE except for Cuffdiff which use the sam format files converted from the bam format files. All the analyses are carried out with the standard procedures and default parameters as the documentation of every method suggest unless otherwise stated. Adjusted *p*-value of FDR of differential expression for each gene are calculated by each method. The information and details of every method used in our study are listed as following.

SCDE (v.1.99.1): The parameters of function clean. counts are 'min.reads = 0' and 'min.detected = 0'. We set paramrter 'n.cores = 8' for function scde.error.models and function scde.expression.difference. Other parameters are as suggested in SCDE's tutorials.

monocle (v.1.99.0): The parameter expressionFamily for function newCellDataSet is 'expressionFamily = negbinomial()'. The parameter cores for function differentialGeneTest is 'cores = 8'. Other parameters are as vignette of monocle suggested.

D3E (Latest commit 6727adf on 21 Oct 2015): All the parameters are as suggested in the example of D3E's wiki pages, except '-m 0', because mode 1 runs too slow for the large scale experiments.

BPSC(v.0.99.0): The analysis is conducted as the examples shown in BPSC package introduction.

DESeq (v.1.18.0): The dispersion estimation procedure call function of estimateDispersions with different parameters for different sample numbers. When there is no replicate, the parameters are 'method = "blind", sharingMode ="fit-only", fitType ="local"'. when sample number of replicates is larger than 5, the parameters are 'sharingMode ="gene-est-only", fitType ="local"' as recommended by the documentation for large replicates. For other conditions, the set parameter is 'fitType = "local"'.

edgeR (v.3.8.6): When there is no replicates, as suggested by the documentation of edgeR, we did not carry the significance analysis of DE.

baySeq (v.2.0.50): Prior parameters of negative binomial are estimated using getPriors.NB with parameter of 'estimation = "QL"' which stands for quasi-likelihood estimation. Note that baySeq will use posterior probabilities as lower threshold instead of using adjusted *p*-values as upper threshold so we use parameter 'likelihood = 0' in topCounts to get a result table of all the genes. And in the table there is also a column of 'FDR.DE' which contains the FDR of differential expression so we could use FDR as upper threshold again.

NBPSeq (v.0.3.0): The analysis is conducted as the examples of exact.nb.test shown in documentation of NBPSeq.

Cuffdiff (v2.2.1 (4237)) with the default parameters and GTF file used in HTSeq before except the parameter for number of threads ('-p 8').

DEGseq (v.1.20.0): The parameter used in DEGexp is 'method ="MARS"', and other parameters are as the examples of DEGexp shown in documentation of DEGseq.

TSPM (February 2011): The analysis is conducted as the examples shown in the script of TSPM (http://www.stat.purdue.edu/~doerge/software/TSPM.R).

Limma (v.3.22.7): The analysis is conducted as the examples shown in 'limma: Linear Models for Microarray and RNA-Seq Data, User's Guide'(Last revised 8 September 2015).

Ballgown (v.1.0.4): The command we used for DE analysis is 'stattest (gowntable = counts, feature ='gene', pData = pData, covariate ='group', getFC = TRUE)'.

SMAseq (v.2.0): We exclude the genes whose read count is 0 in all samples first. We set 'resp.type = "Two class unpaired"' and specific random seed for every round of analysis.

For each method and each experiment shown in Table 3 and Table S1, comparisons were conducted between Group 1 and Group 2 with the sample sizes of 1, 2, 5, 10 and 20 respectively, and each sample size's experiment was repeated for 20 times with different random seeds for random sampling.

ABBREVIATIONS

RNA-Seq,	RNA sequencing
scRNA-seq,	single-cell RNA sequencing
DE,	differential expression
DEG,	differentially expressed genes
LPS,	lipopolysaccharide
FDR,	false discovery rate
ROC,	receiver operating characteristic
NDE,	non-differentially expressed
GEO,	Gene Expression Omnibus

AVAILABILITY OF SUPPORTING DATA

The data sets we used in our study are all come from Gene Expression Omnibus (GEO) and their accession numbers are GSE48968, GSE59127, GSE59129, GSE59130 and GSE74923 respectively.

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at DOI 10.1007/s40484-016-0089-7.

AUTHORS' CONTRIBUTIONS

XZ conceived the study. ZM and XZ designed the experiments and analyzed the data. ZM implemented the experiments. ZM and XZ wrote the manuscript.

ACKNOWLEDGEMENTS

The authors greatly acknowledge the contributions and suggestions from

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

Drs. Ke Deng, Xiaowo Wang, Jun Li, Xi Wang and Zhixing Feng. This work is partially supported by the National Basic Research Program of China (2012CB316504).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Zhun Miao and Xuegong Zhang declare that they have no conflict of interests. All the data sets the authors used are from public repositories.

REFERENCES

1. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods, 5, 621–628
2. Stegle, O., Teichmann, S. A. and Marioni, J. C. (2015) Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet., 16, 133–145
3. Shapiro, E., Biezuner, T. and Linnarsson, S. (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat. Rev. Genet., 14, 618–630
4. Macaulay, I. C. and Voet, T. (2014) Single cell genomics: advances and future perspectives. PLoS Genet., 10, e1004126
5. Tang, F., Lao, K. and Surani, M. A. (2011) Development and applications of single-cell transcriptome analysis. Nat. Methods, 8, S6–S11
6. Kanter, I. and Kalisky, T. (2015) Single cell transcriptomics: methods and applications. Front. Oncol., 5, 53
7. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. and Teichmann, S. A. (2015) The technology and biology of single-cell RNA sequencing. Mol. Cell, 58, 610–620
8. Sandberg, R. (2014) Entering the era of single-cell transcriptomics in biology and medicine. Nat. Methods, 11, 22–24
9. Saliba, A. E., Westermann, A. J., Gorski, S. A. and Vogel, J. (2014) Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res., 42, 8845–8860
10. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. Genome Biol., 11, R106
11. Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 139–140
12. Wang, L., Feng, Z., Wang, X., Wang, X. and Zhang, X. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics, 26, 136–138
13. Kharchenko, P. V., Silberstein, L. and Scadden, D. T. (2014) Bayesian approach to single-cell differential expression analysis. Nat. Methods, 11, 740–742
14. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. and Rinn, J. L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol., 32, 381–386
15. Delmans, M. and Hemberg, M. (2016) Discrete distributional differential expression (D3E)—a tool for gene expression analysis of single-cell RNA-seq data. BMC Bioinformatics, 17, 110
16. Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen, M. and Pawitan, Y. (2016) Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics, 32, 2128–2135
17. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol., 16, 278
18. Wu, L., Zhang, X., Zhao, Z., Wang, L., Li, B., Li, G., Dean, M., Yu, Q., Wang, Y., Lin, X., *et al.* (2015) Full-length single-cell RNA-seq applied to a viral human cancer: applications to HPV expression and splicing analysis in HeLa S3 cells. Gigascience, 4, 51
19. Freeman, B. T., Jung, J. P. and Ogle, B. M. (2015) Single-cell RNA-seq of bone marrow-derived mesenchymal stem cells reveals unique profiles of lineage priming. PLoS One, 10, e0136199
20. Avraham, R., Haseley, N., Brown, D., Penaranda, C., Jijon, H. B., Trombetta, J. J., Satija, R., Shalek, A. K., Xavier, R. J., Regev, A., *et al.* (2015) Pathogen cell-to-cell variability drives heterogeneity in host immune responses. Cell, 162, 1309–1321
21. Blakeley, P., Fogarty, N. M. E., Valle, I. D., Wamaitha, S. E., Hu, T. X., Elder, K., Snell, P., Christie, L., Robson, P. and Niakan, K. K. (2015) Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. Development, 142, 3613
22. Fan, X., Zhang, X., Wu, X., Guo, H., Hu, Y., Tang, F. and Huang, Y. (2015) Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol., 16, 148
23. Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., Levi, B., Gray, L. T., Sorensen, S. A., Dolbeare, T., *et al.* (2016) Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci., 19, 335–346
24. Hardcastle, T. J. and Kelly, K. A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics, 11, 422
25. Di, Y., Schafer, D. W., Cumbie, J. S. and Chang, J. H. (2011) The NBP negative binomial model for assessing differential gene expression from RNA-Seq. Stat. Appl. Genet. Mol. Biol., 10, 1–28
26. Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. and Pachter, L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol., 31, 46–53
27. Auer, P. L. and Doerge, R. W. (2011) A two-stage Poisson model for testing RNA-Seq data. Stat. Appl. Genet. Mol. Biol., 10, doi: 10.2202/1544-6115.1627
28. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res., 43, e47
29. Frazee, A. C., Pertea, G., Jaffe, A. E., Langmead, B., Salzberg, S. L. and Leek, J. T. (2014) Flexible analysis of transcriptome assemblies with Ballgown. Biorxiv: http://dx.doi.org/10.1101/003665
30. Li, J. and Tibshirani, R. (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat. Methods Med. Res., 22, 519–536
31. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.Durbin, R., and the 1000 Genome Project Data Processing Subgroup. (2009) The sequence alignment/map format and SAMtools. Bioinformatics, 25, 2078–2079
32. Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., *et al.* (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature, 510, 363–369
33. Brunskill, E. W., Park, J. S., Chung, E., Chen, F., Magella, B. and Potter, S. S. (2014) Single cell dissection of early kidney development:

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

multilineage priming. Development, 141, 3093–3101

34. Kimmerling, R. J., Lee Szeto, G., Li, J. W., Genshaft, A. S., Kazer, S. W., Payer, K. R., de Riba Borrajo, J., Blainey, P. C., Irvine, D. J., Shalek, A. K., *et al.* (2016) A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. Nat. Commun., 7, 10220
35. Su, Z., Łabaj, P. P., Li, S., Thierry-Mieg, J., Thierry-Mieg, D., Shi, W., Wang, C., Schroth, G. P., Setterquist, R. A., and Thompson, J. F. (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat. Biotechnol., 32, 903–914
36. Tan, P. K., Downey, T. J., Spitznagel, E. L. Jr, Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M. and Cam, M. C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res., 31, 5676–5684
37. Shi, L., Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat. Biotechnol., 24, 1151–1161
38. Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25, 1105–1111
39. Anders, S., Pyl, P. T., Huber, W (2015) HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics 2015, 31, 166–169

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2016