BMC Bioinformatics

SOFTWARE

Open Access

RankCompV3: a differential expression analysis algorithm based on relative expression orderings and applications in single-cell RNA transcriptomics

Jing Yan[1], Qiuhong Zeng[1] and Xianlong Wang[1,2*]

*Correspondence:
wang.xianlong@139.com

[1] Department of Bioinformatics, Fujian Key Laboratory of Medical Bioinformatics, School of Medical Technology and Engineering, Fujian Medical University, Fuzhou 350122, China
[2] The Second Affiliated Hospital, Fujian Medical University, Quanzhou 362000, China

Abstract

Background: Effective identification of differentially expressed genes (DEGs) has been challenging for single-cell RNA sequencing (scRNA-seq) profiles. Many existing algorithms have high false positive rates (FPRs) and often fail to identify weak biological signals.

Results: We present a novel method for identifying DEGs in scRNA-seq data called RankCompV3. It is based on the comparison of relative expression orderings (REOs) of gene pairs which are determined by comparing the expression levels of a pair of genes in a set of single-cell profiles. The numbers of genes with consistently higher or lower expression levels than the gene of interest are counted in two groups in comparison, respectively, and the result is tabulated in a 3×3 contingency table which is tested by McCullagh's method to determine if the gene is dysregulated. In both simulated and real scRNA-seq data, RankCompV3 tightly controlled the FPR and demonstrated high accuracy, outperforming 11 other common single-cell DEG detection algorithms. Analysis with either regular single-cell or synthetic pseudo-bulk profiles produced highly concordant DEGs with the ground-truth. In addition, RankCompV3 demonstrates higher sensitivity to weak biological signals than other methods. The algorithm was implemented using Julia and can be called in R. The source code is available at https://github.com/pathint/RankCompV3.jl.

Conclusions: The REOs-based algorithm is a valuable tool for analyzing single-cell RNA profiles and identifying DEGs with high accuracy and sensitivity.

Key points

- RankCompV3 is a method for identifying differentially expressed genes (DEGs) in either bulk or single-cell RNA transcriptomics. It is based on the counts of relative expression orderings (REOs) of gene pairs in the two groups. The contingency tables are tested using McCullagh's method.

- RankCompV3 has comparable or better performance than that of other conventional methods. It has been shown to be effective in identifying DEGs in both single-cell and pseudo-bulk profiles.
- Pseudo-bulk method is implemented in RankCompV3, which allows the method to achieve higher computational efficiency and improves the concordance with the bulk ground-truth.
- RankCompV3 is effective in identifying functionally relevant DEGs in weak-signal datasets. The method is not biased towards highly expressed genes.

Keywords: Single-cell RNA sequencing, Differential expression analysis, Differentially expressed genes, Relative expression orderings

Background

High-throughput transcriptomic sequencing (RNA-seq) is a powerful tool for comprehensive expression profiling, which is essential for understanding biological and medical problems. A key step in RNA-seq analysis is the detection of differentially expressed genes (DEGs) [1, 2]. However, the use of RNA-seq data often involves joint analysis of data across multiple platforms, which can introduce batch effects [3], systematic differences between samples that are not due to biological variation. Many batch effect adjustment methods have been proposed, such as SVA [4] for microarray data and svaseq [5] for RNA-seq data. However, normalization of different batch datasets can also distort true biological signals [6], especially when the samples are not evenly distributed between different batches. It can distort true biological signals and lead to high false positive rates (FPRs) for DEGs. Therefore, it is important to carefully consider the normalization step when analyzing RNA-seq data.

The relative expression orderings (REOs) of gene pairs within a profile are often stable in samples of the same phenotype, but differ significantly in different phenotypes. This can be used to construct biomarkers and identify DEGs. Our lab previously developed two versions of an algorithm based on REOs, RankComp [7] and RankCompV2 [8]. We successfully applied these algorithms to microarray, RNA-seq, methylation, and proteomic data [7, 9–14]. RankComp can be used to identify DEGs at the population and individual levels, and it is insensitive to batch effects and normalization.

Previous versions of the algorithm used Fisher's exact test to calculate the significance level of a 2×2 contingency table. The two rows of the table represent two groups in comparison, such as control and treatment groups, and the two columns represent the two REO outcomes: whether the expression level of the paired gene is higher or lower than that of the current gene. The test evaluates whether there is a significant correlation between the treatment conditions or phenotypes and the distribution of REO outcomes in the two groups.

RankCompV2 evaluates the stability of DEGs by including background genes updating cycles to filter out unstable DEGs. This improvement addresses the problem that upregulation or downregulation of a gene may be incorrectly indicated by its paired gene if the paired gene is dysregulated. However, both RankComp and RankCompV2 neglect the matched pair relationship of REOs in the two compared groups. As a result, REOs that are consistent in both groups are also included in the construction

of contingency tables. This negligence may lead to non-differential genes being identified as DEGs, resulting in high FPRs.

McNemar's test is a statistical test that takes into account the matched experiment design in the test of 2×2 contingency tables. McNemar–Bowker's test extends this to the general situation with more than 2 categories. However, both of these tests do not take into account the ordered relationship of the three possible outcomes of REOs.

Another limitation of the previous two versions of RankComp is that the contingency tables do not consider the contribution of gene pairs with approximately equal expression levels (non-stable REOs based on the binomial distribution) in one group but stable REOs in the other group.

In RankCompV3, we count the frequency of all possible 9 REO outcomes in a matched pairs design. McCullagh's test [15], which is designed for the matched designs with ordered categories, is applied to test the 3×3 contingency tables.

We implemented RankCompV3 and tested it on single-cell RNA-seq (scRNA-seq) data. Many methods have been developed for differential expression analysis of scRNA-seq data. Some were specifically developed for scRNA-seq, such as MAST [16], DEsingle [17], Wilcoxon signed-rank test [18], Monocle2 [19], SigEMD [20], and scDD [21]. Others were originally developed for microarray and bulk RNA-seq data, such as limma [22, 23], edgeR [24], and DESeq2 [25], but have been found to perform well in scRNA-seq data using the pseudo-bulk analysis scheme [26]. However, the consistency of these algorithms is not high, and many algorithms have problems with insufficient sensitivity and high FPRs. Additionally, it has been found that algorithms specifically developed for scRNA-seq data do not necessarily perform better than those designed for bulk profiles [27–29].

We evaluated the performance of RankCompV3 on scRNA-seq datasets by comparing it with several common DEG identification algorithms using both simulated multimodal single-cell datasets and real datasets. Our analysis results showed that RankCompV3 performed well on these datasets and was sensitive to weak biological signals. Additionally, we obtained concordant DEGs between the analysis comparing the single-cell profiles directly and the pseudo-bulk analysis, in which single-cell profiles were aggregated randomly into pseudo-bulk profiles [26].

Methods and materials

The RankCompV3 algorithm

The RankCompV3 method for identifying DEGs in either bulk or scRNA-seq data consists of the following steps:

1. *Identification of significantly stable REOs* For each gene pair (a, b), the observed REO outcome is either $a > b$ or $a < b$ in a single-cell profile or a bulk profile, depending on which gene has the higher expression level. If the two genes have the equal expression levels within the measurement uncertainty, the REO is randomly assigned to either $a > b$ or $a < b$ with equal probability.

The binomial distribution model is then used to test if the observed REOs are stable across all the profiles in each group, respectively. The null hypothesis is that the two genes have the same expression levels. This means that the probability of observing each REO outcome ($a > b$ or $a < b$) is equal, $p_0 = 0.5$. The P value is the probability of observing the major REO outcome in m or more profiles out of a total of n profiles by chance (where 'major' means $m > n/2$) in one group, $P = 1 - \sum_{i=0}^{m} \binom{n}{i} p_0^i (1 - p_0)^{n-i}$. If the P value is less than the preset threshold, e.g., $\alpha = 0.01$, the major REO outcome is considered to be significantly stable, $a > b$ or $a < b$. Otherwise, it is considered that the REO is not stable, denoted as $a \sim b$.

2. *Construction of contingency tables of stable REOs* For each gene, the contingency tables of REO counts are constructed. The contingency tables summarize the comparison results of the gene with reference genes in two groups, e.g., control and treatment groups (Table 1). The diagonal elements of the contingency tables are the numbers of REOs which are consistent in the two groups, while the off-diagonal elements are the numbers of inconsistent REOs in the two groups. Lower triangular elements support that a is down-regulated in the treatment group compared with the control group, while the upper triangular elements support that a is up-regulated.
3. *Significance test of contingency tables* McCullagh's method is applied to test the significance of the contingency tables. The method applies a logistic model for matched comparisons with ordered categorical data. The null hypothesis H_0 is that the distribution of the REO outcomes has no association with the grouping and the contingency table should be symmetric. The Benjamini–Hochberg (BH) method is used to adjust P values for multiple comparisons. If the adjusted P value < 0.05, the null hypothesis H_0 is rejected, and the gene is considered as a candidate DEG.
4. *Iteration* Initially, each gene is compared with a list of housekeeping genes to construct the contingency tables. If such gene list is not provided, all the genes are used as the reference genes. After the first cycle, all non-differentially expressed genes from the previous cycle are used as the reference list to construct new contingency tables. The cycle ends when the list of candidate DEGs does not change any more.

The overall workflow of the RankCompV3 method is shown in Fig. 1.

Table 1 3 × 3 contingency table of REO counts for one gene (a)

Control	Treat		
	$a < b$	$a \sim b$	$a > b$
$a < b$	n_{11}	n_{12}	n_{13}
$a \sim b$	n_{21}	n_{22}	n_{23}
$a > b$	n_{31}	n_{32}	n_{33}

The table shows the number of REOs for gene a that form either a significantly stable REO or a nonstable REO with a reference gene b in the two groups, e.g., control and treatment. The row and column headers indicate the different types of REOs. Each cell sums the number of REOs forming 9 possible paired outcomes in the two groups

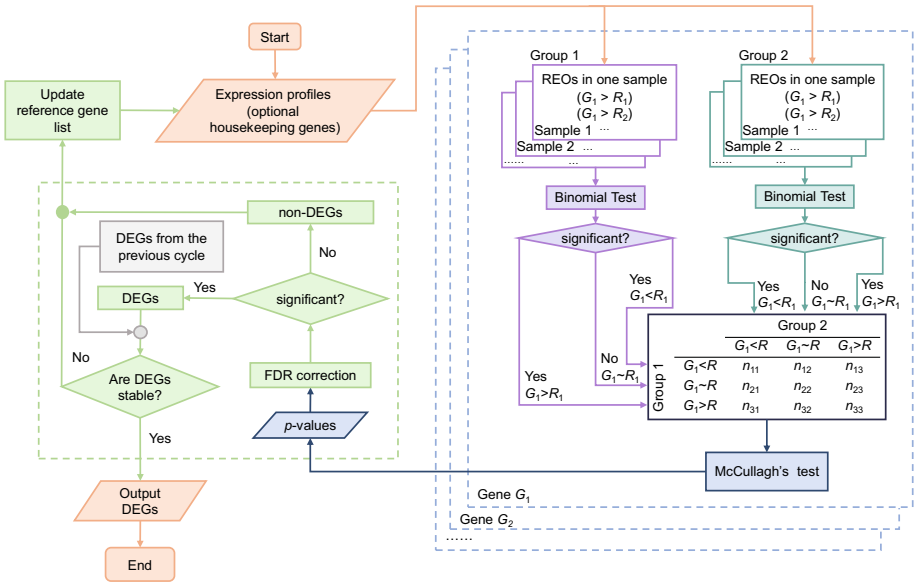

Fig. 1 Flowchart of the RankCompV3 method. The RankCompV3 method is a three-step process for identifying differentially expressed genes (DEGs) in RNA transcriptomics. The first step is to calculate the relative expression orderings (REOs) of gene pairs in the two groups. The second step is to test the contingency tables of REOs using McCullagh's method. The third step is to filter the DEGs using a significance threshold

Other DEG identification methods

We compared the performance of RankCompV3 with the performance of 11 other methods, including 7 independent methods, limma, DESeq2, edgeR, DEsingle, SigEMD, Monocle2 and scDD, and 4 methods implemented in the Seurat package, Wilcoxon rank sum test ("wilcox"), likelihood-ratio test ("bimod"), logistic regression ("LR") and MAST. The methods were compared using a simulated scRNA-seq dataset and real scRNA-seq datasets.

Performance evaluation

The performance of RankCompV3 was evaluated using the following metrics: true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), false negative rate (FNR), precision, accuracy, AUC (area under the receiver operating characteristic curve) [30] and AUCC (area under the concordance curve) [26]. ROC curve and AUC calculation were performed with the R package of pROC. AUCC is a more stringent metric than AUC that is used to evaluate the concordance between single-cell or pseudo-bulk DEGs and bulk ground truth.

Benchmark dataset

We used the 'gold standard' scRNA-seq dataset collected in Squair et al.'s study [26] to compare the performance of our method and to the benchmark results in that study. The dataset was compiled from four studies [26, 31–34] in which matched bulk and scRNA-seq were carried out on the same population of purified cells under the same

conditions, where the bulk results were considered as the 'ground-truth', as detailed in Supplementary file 2: Table S1.

Simulation dataset

The simulation dataset was generated using the scDD package ("simulateSet" function). The synthetic dataset scDatEx provided by the author was taken as the starting point. The package models single-cell expression profiles of cells under two conditions using Bayesian mixture models to accommodate the main characteristics of scRNA-seq data, including heterogeneity, multimodality, and sparsity (a large number of 0 counts).

The profiles were generated for 75 cells and 20,000 genes under each condition, including 2000 DEGs and 18,000 non-DEGs. The 2000 DEGs were equally distributed among four modes: DE (differential expression of unimodal genes): The expression levels of these genes are different between the two conditions, and the genes are unimodal (i.e., they belong to a single cluster); DP (differential proportion for multimodal genes): The expression levels of these genes are different between the two conditions, but the genes are multimodal (i.e., they belong to multiple clusters); DM (differential modality genes): The expression levels of these genes are different between the two conditions, and the genes change modality (i.e., they move from one cluster to another); DB (both differential modality and different component genes): The expression levels of these genes are different between the two conditions, and the genes change modality and also change the cluster they belong to. Among the 18,000 non-DEGs, one half are EE (equivalent expression for unimodal) genes and the other half are EP (equivalent proportion for multimodal) genes.

All simulation parameters were set to default values, and the resulting data were rounded to the nearest integers. The specific types and distributions of the simulated data are shown in Supplementary file 1: Fig. S1.

Real datasets

We used public datasets downloaded from the Gene Expression Omnibus Database (GEO, http://www.ncbi.nlm.nih.gov/geo/). The datasets included bulk RNA sequencing (RNA-seq) dataset (GSE82158) and single-cell RNA sequencing (scRNA-seq) datasets (GSE54695, GSE29087 and GSE59114) measured on different platforms.

Negative test dataset

We used 80-cell profiles from the GSE54695 dataset provided by Grün et al. [35] as a negative test dataset to evaluate FPR. The profiles were obtained by lysing frozen mouse embryonic stem cells (mESCs) under the same culture conditions and measured with the CEL-seq technique. According to Wang et al.'s study [36], the top 7277 genes were retained for analysis with the highest number of nonzero expression in all cells. We randomly divided 80 profiles into two subsets with 40 in each. Since all the profiles were generated under the same condition, there should be no DEGs between the two subsets. The randomized experiments were repeated 10 times and the average FPR was calculated [37].

Positive test dataset

The positive test dataset (GSE29087) of scRNA-seq were provided by Islam et al. [38], consisting of 22,936 genes from 48 mouse ES cells and 44 mouse embryonic fibroblasts. We used the top 1000 DEGs as the gold standard DEGs between the two cell types, which were verified by the real-time quantitative reverse transcription PCR (RT-qPCR) experiments [27, 39].

Influence of sample size

The scRNA-seq dataset (GSE59114) provided by Kowalczyk et al. [40] were used to evaluate the influence of sample size (number of profiles in a group) to the performance. The dataset consisted of the profiles of 89 long-term hematopoietic stem cells (LTHSCs) from young mice (2–3 months) and 135 LTHSCs from old mice (20 months). We randomly sampled the profiles of two conditions into subsets with sizes of 10, 30, 50, and 70, respectively. The DEGs were identified between the subsets with the same sample size. DEGs identified by the same algorithm in the entire dataset were used as the gold standard. The random sampling experiments were repeated 10 times for each sample size and the average performance indices were calculated as the final result.

Weak-signal test dataset

To test the performance of our algorithm on weak-signal detection, we used the GSE82158 dataset provided by Misharin et al. [41]. We compared the profiles of 4 monocyte-derived alveolar macrophages (Mo-AMs) and 4 tissue-resident macrophages (TR-AMs) from mice treated with bleomycin for 10 months. Pathway enrichment analysis was performed on the DEGs to detect biological signals.

Pseudo-bulk method

In addition to comparing the input single-cell profiles directly, the pseudo-bulk method is also available in the implementation of RankCompV3. The single-cell profiles are randomly partitioned into a number of subsets of equal sample size, respectively, in each group. The profiles in each subset are aggregated into a single pseudo-bulk profile. Differential expression analysis is then conducted with these synthetic profiles from two groups.

For the benchmark datasets from Squair et al.'s study [26], the pseudo-bulk profiles were generated in the same manner with the original study, where the cells were aggregated in each replicate.

Data preprocessing

The contingency tables were constructed from the raw counts except for GSE59114 and GSE82158 where the counts were not available. In GSE59114, the profiles were provided as $\log_2(\mathrm{TPM}+1)$ and in GSE82158 the profiles were normalized with

edgeR. We removed the genes that were not expressed in most of the cells and the cells expressing very few genes. No other preprocessing steps were applied for RankCompV3.

Pathway enrichment analysis

Pathway enrichment analysis was performed against the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [42] with the hypergeometric distribution model [43] through the clusterProfiler package ("enrichKEGG" function). The *P* values were corrected by the BH method for multiple tests.

Results

Performance on the null dataset

We tested the FPR using the null dataset of profiles of 80 cells lysed from frozen mouse embryonic stem cells under the same culture conditions, provided by Grün et al. [35]. The dataset was randomly divided into two subsets, each with 40 profiles. In 10 randomized trials with a false discovery rate (FDR) < 0.05 as the threshold, no or very few DEGs were detected by RankCompV3, Bimod, LR, MAST, DEsingle, Wilcoxon, edgeR, limma, and scDD algorithms. On average, 43.9, 7.5, 0, 0, 4.7, 0, 0, 0, and 77.5 DEGs were detected, respectively. The Monocle2 and SigEMD algorithms detected 132.6 and 1291 DEGs, which showed relatively high FPRs. These results show that the RankCompV3 method performs well in controlling the FPR in the negative dataset.

Performance comparison with Squair et al.'s benchmark test

Squair et al. performed a benchmark test on various DEG algorithms. It included the algorithms specially designed for single-cell mode and the algorithms designed for bulk profiles. The results from bulk RNA-seq profiles of the same samples were used as the ground truth to test the performance of the algorithms in single-cell and pseudo-bulk

Table 2 Area under the concordance curve (AUCC) for RankCompV3 for pseudo-bulk analysis scheme and single-cell analysis scheme in the benchmark datasets from Squair et al. [26]

Dataset	Ground-truth[a]		edgeR-LRT[b]	
	Pseudo-bulk	Single-cell	Pseudo-bulk	Single-cell
Hagai2018_mouse-lps	0.742	0.502	0.457	0.504
Hagai2018_mouse-pic	0.651	0.590	0.560	0.533
Hagai2018_pig-lps	0.543	0.483	0.489	0.455
Hagai2018_rabbit-lps	0.453	0.385	0.447	0.335
Hagai2018_rat-lps	0.666	0.521	0.453	0.453
Hagai2018_rat-pic	0.536	0.420	0.364	0.289
Angelidis2019_alvmac	0.039	0.028	0.053	0.049
Angelidis2019_pneumo	0.216	0.195	0.226	0.176
CanoGamez2020_memory-iTreg	0.382	0.306	0.186	0.177
Reyfman2020_alvmac[c]	0.170	0.094	0.004	0.004
Reyfman2020_pneumo[c]	0.269	0.099	0.004	0.004

[a] Concordance with the 'ground-truth' DEGs obtained with the bulk datasets using RankCompV3;

[b] Concordance with the 'ground-truth' DEGs obtained with the bulk datasets using edgeR-LRT, the best method in Ref. [26];

[c] Concordance with the given bulk DEGs and no bulk profiles were provided

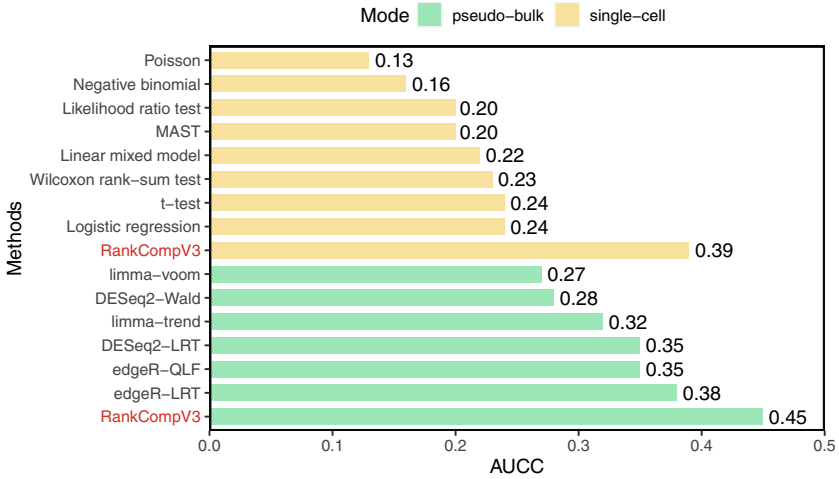

Fig. 2 Median AUCC for RankCompV3 and 14 other DEGs recognition algorithms for pseudo-bulk analysis scheme in the benchmark datasets from Squair et al.

modes. Eleven gold standard datasets were used to evaluate the performance of RankCompV3 using the same benchmark protocol as in Ref. [26]. The results are listed in Table 2. In most datasets, the concordance between the pseudo-bulk and the 'ground-truth' is very high. The median AUCC of RankCompV3 in the pseudo-bulk mode is 0.45, which is higher than that of the best method in Ref. [26] (edgeR-LRT, median AUCC = 0.38) (Fig. 2). More importantly, the concordance is also high for RankCompV3 applied to single-cell profiles directly (median AUCC = 0.39), which is better than the best method for single-cell in the Ref. [26] (*t*-test and Wilcoxon rank-sum test, the median AUCC is 0.24) (Fig. 2).

Only in two datasets, Angelidis2019_alvmac and Reyfman2022_pneumo, did RankCompV3 perform poorly. The data quality of Angelidis2019_alvmac dataset was very poor and many genes were not detected in either the bulk or single-cell profiles. For the Reyfman2022 datasets, no bulk profiles were provided, and the given bulk DEGs were used as the 'ground-truth'.

Furthermore, in comparison with the 'ground-truth' obtained with edgeR-LRT, our results also show strong concordance. The median AUCC metric is 0.36 for the pseudo-bulk method and 0.29 for the single-cell method.

These results suggest that RankCompV3 can be effectively used for pseudo-bulk analysis, which can improve the performance of differential expression analysis in scRNA-seq data.

Performance evaluation with the simulated single-cell dataset

Since we could not obtain all the truly differentially expressed genes in a real dataset, we simulated multimodal scRNA-seq profiles with scDD for cells under two conditions with 75 cells in each to test the performance. The dataset contains 2000 DEGs equally distributed in four different modes and 18,000 non-DEGs equally distributed in two modes. The ROC curves are shown in Fig. 3A. RankCompV3 has an AUC of 0.865, which is in the middle-tier of all 12 algorithms. The AUCs of Monocle2,

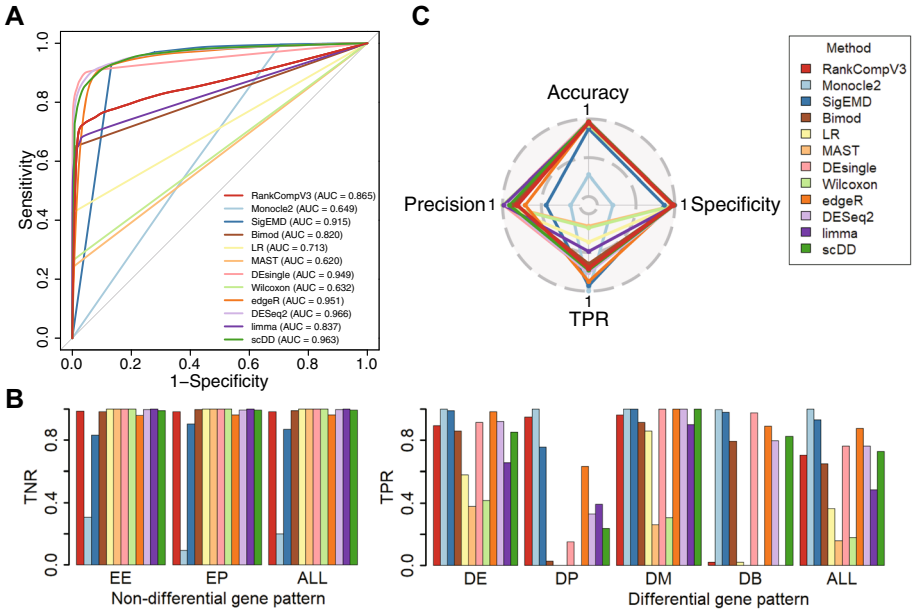

Fig. 3 Simulated single-cell dataset. **A** ROC curves of RankCompV3 and other algorithms in the simulated scRNA-seq dataset. The area under each ROC curve (AUC) is shown in the legend. **B** TNRs and TPRs for the non-differential and differential genes with different expression modes in the simulated scRNA-seq dataset. **C** The radar map shows the Accuracy, Specificity, True positive rate and Precision of each algorithm

Table 3 Detection performance of RankCompV3 and 11 other algorithms in the simulated scRNA-seq dataset (FDR < 0.05)

Method	# DEGs	TPR	FPR	Precision	Accuracy
RankCompV3	1731	0.706	0.018	0.815	0.955
Monocle2	16,383	0.998	0.799	0.122	0.280
SigEMD	4252	0.930	0.133	0.437	0.873
Bimod	1497	0.648	0.011	0.866	0.955
LR	732	0.366	0.000	1.000	0.937
MAST	320	0.160	0.000	1.000	0.916
DEsingle	1546	0.761	0.001	0.984	0.975
Wilcoxon	363	0.182	0.000	1.000	0.918
edgeR	2474	0.876	0.040	0.708	0.951
DESeq2	1634	0.762	0.006	0.933	0.971
limma	988	0.487	0.001	0.985	0.948
scDD	1599	0.728	0.008	0.910	0.966

SigEMD, Bimod, LR, MAST, DEsingle, Wilcoxon, edgeR, DESeq2, limma, and scDD are 0.649, 0.915, 0.820, 0.713, 0.620, 0.949, 0.632, 0.951, 0.966, 0.837, and 0.963, respectively.

The performances of the different algorithms are compared in Table 3 at a FDR < 0.05 threshold. RankCompV3 had a FPR of 0.018, a precision of 0.815, and an accuracy of 0.955, which were all better than or similar to the other algorithms. Compared to LR, MAST, and Wilcoxon, RankCompV3 showed a higher TPR while maintaining an extremely low FPR. Monocle2, SigEMD, and edgeR obtained extremely high TPRs, but they also contained a large number of false positives (FPs). Monocle2 had a TPR of

0.998, but its accuracy was only 0.122. This is because Monocle2 identified 16,383 DEGs among 20,000 genes, which introduced a larger FPR of 0.799 (Fig. 3C).

We evaluated the genes of different modes separately and compared the TNRs of non-DEGs of the two different modes and the TPRs of DEGs of the four different modes, as shown in Fig. 3B. The results showed that the average TNR of the two modes of non-DEGs, EE and EP, was 0.982 using RankCompV3. The highest TPR of the four modes of DEGs, DD, DP, DB, and DM, was 0.962. For the DEG modes with no or low multimodality, DE and DM, the average TPR was 0.927. For the highly pleiotropic DEGs, DP and DB, the identification ability for DP was also very high (0.946), but it failed to detect DB genes. These results demonstrate that RankCompV3 strictly controls the FPR in the single-cell simulation dataset and has a good ability to detect DEGs of low multimodality and some DEGs with high multimodality.

In the simulated dataset, we randomly generated 10 pseudo-bulk profiles for each group to identify DEGs. The mean number of detected DEGs from the 10 random experiments was 1737.5 (standard deviation, SD = 13.3), and the mean number of true DEGs was 1413.3 (SD = 3.1). These numbers show a slight improvement compared to those obtained with the single-cell profiles directly (1731 and 1411, respectively). The mean AUCC metric was 0.884 (SD = 0.01), showing a strong concordance between the pseudo-bulk and the single-cell methods.

Performance evaluation in real scRNA-seq dataset

Although the simulated dataset mimics numerous aspects of single-cell expression profiles, it cannot fully capture the complex characteristics of real data. To evaluate the performance of RankCompV3 on real data, we used the scRNA-seq dataset provided by Islam et al. [38], which consisted of the profiles of 48 mouse ES cells and 44 mouse

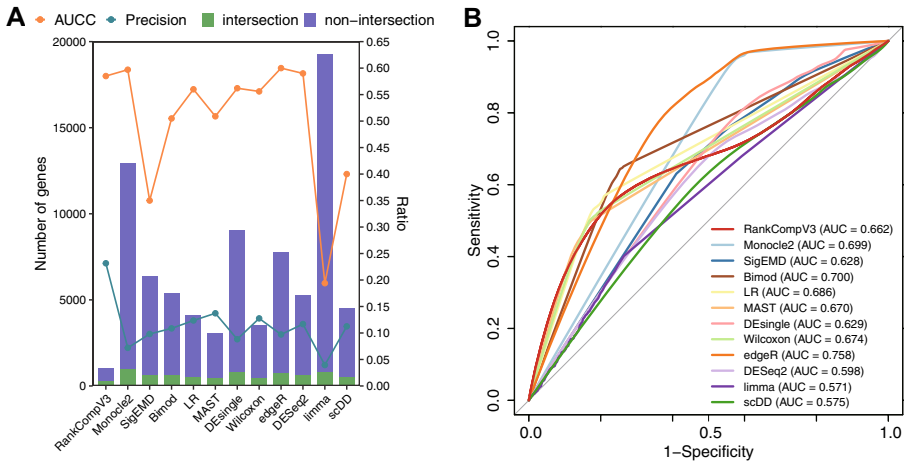

Fig. 4 Real scRNA-seq dataset. **A** Performance in identifying DEGs in the scRNA-seq test dataset of GSE29087 where the top1000 genes confirmed by RT-qPCR were used as the gold standard. The purple column represents the number of DEGs identified by each algorithm. The green area represents the number of intersection genes between the DEGs identified by the algorithm and the top1000 genes. The connected blue dots indicate TPRs and the connected orange lines indicate the area under the concordance curve (AUCC) metric between the top 1000 DEGs identified by the algorithm and the gold standard. **B** ROC curves in the scRNA-seq positive test dataset of GSE29087. The top1000 genes were used as the gold standard and the AUC values are shown in the legend

embryonic fibroblasts. The top 1000 differential genes verified by RT-qPCR experiments were used as the gold standard DEGs (top1000).

Figure 4A shows the number of DEGs identified by each algorithm, the true number of DEGs (the number of genes that intersect with the top1000 genes), AUCC, and precision. RankCompV3 identified 1045 DEGs, of which 242 were true DEGs (TPR = 0.242). Although the number of true DEGs is the smallest, RankCompV3 has the highest precision and accuracy. RankCompV3 has a strictly conservative FPR (0.037), while maintaining a higher accuracy (0.932) than the other 11 methods which show high FPRs, ranging from 0.122 to 0.873. Among the other 11 algorithms, limma, Monocle2, DEsingle, and edgeR have the highest TPRs, which are 0.761, 0.931, 0.797, and 0.753, respectively. However, these algorithms also have high FPRs, ranging from 0.330 to 0.873. As a result, their accuracies are relatively low, ranging from 0.150 to 0.651.

Since the top1000 is a partial list of true DEGs which might cause higher FPRs in some algorithms, we further evaluated the performance with AUC and AUCC. The ROC curves are shown in Fig. 4B. The AUC of RankCompV3 is 0.662, which is comparable to the other 11 algorithms. The AUCs of Monocle2, SigEMD, Bimod, LR, MAST, DEsingle, Wilcoxon, edgeR, DESeq2, limma, and scDD are 0.699, 0.628, 0.700, 0.686, 0.670, 0.629, 0.674, 0.758, 0.598, 0.571, and 0.575, respectively. The AUCC in Fig. 4A measures the concordance between the top 1000 DEGs detected by each algorithm and the gold standard. EdgeR and Monocle2 obtained the best concordance scores (0.60), followed by RankCompV3 and DESeq2 (0.59).

In conclusion, RankCompV3 can identify DEGs in scRNA-seq positive datasets. Compared to the high FPRs of many other methods, RankCompV3 may be more suitable for studies that require strict control of FPR.

Effect of sample size to performance

To investigate the dependence of performance on sample size, we used the scRNA-seq dataset of LTHSC of two age-group mice provided by Kowalczyk et al. [40]. Subsets with sizes of 10, 30, 50, and 70 were randomly sampled for each age-group, and the sampling experiments were repeated 10 times. DEGs identified by the same algorithm in the entire dataset were taken as the gold standard. We assessed the concordance of the gene list sorted according to the significance level between the entire dataset and the subsets. Figure 5 shows the trend of the AUCC for the complete gene list and the top 1000 genes. The AUCC for the top 1000 genes is more meaningful since they are concentrated with

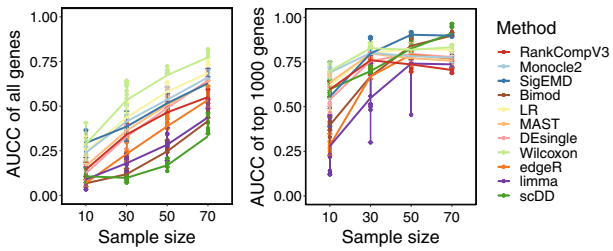

Fig. 5 Performance dependence on sample size. The AUCC was calculated for all the genes in the left panel and for the top 1000 genes in the right panel. The genes were sorted according to their significance levels given by each method in each dataset

more DEGs. As the sample size increased, the TPR and AUCC for the complete gene list increased whereas the FNR decreased (Fig. 5, see also Supplementary file 1: Fig. S2). At a sample size of 30, RankCompV3 identified approximately half of the gold standard DEGs while the AUCC for the top 1000 genes reached a plateau. This indicates that most of the genes of interest have been correctly placed at the top of the gene list, despite some not being identified as DEGs.

Application in weak-signal datasets

In Misharin et al.'s study [44], they found that the differential expression of Siglecf can reliably distinguish Mo-AMs from TR-AMs in the bleomycin-induced early fibrosis. However, after 10 months of treatment with bleomycin, TR-AMs and Mo-AMs expressed similar levels of Siglecf and could not be distinguished by flow cytometry [41]. Only 330 DEGs were identified by two-way ANOVA (FDR < 0.05). With FDR < 0.05, RankCompV3 identified 5023 DEGs. SigEMD, Monocle2, edgeR, scDD, Bimod, DESeq2 and limma detected 2688, 1456, 1224, 720, 182, 111 and 79 DEGs, respectively. Other algorithms failed to detect or detected very few DEGs.

Through functional analysis, we confirmed that many DEGs identified by RankCompV3 have been shown to be associated with the differentiation of Mo-AMs and TR-AMs and the development of pulmonary fibrosis. For example, *Siglecf* is a reliable marker to distinguish Mo-AMs from TR-AMs [44]. Vcam-1 is a TGF-β1 responsive mediator that is upregulated in idiopathic pulmonary fibrosis [45]. SPARC drives pathological responses in non-small cell lung cancer and idiopathic pulmonary fibrosis by promoting microvascular remodeling and the excessive deposition of ECM proteins [46]. FGF2 inhibits pulmonary fibrosis through FGFR1 receptor action [47]. Adam8 deficiency increases CS-induced pulmonary fibrosis [48]. Macrophages expressing SPP1 proliferate during pulmonary fibrosis [49]. *Sparc*, *Fgfr1* and *Adam8* were identified by RankCompV3 only. Many other algorithms failed to detect any of these DEGs, and only Monocle2 (*Siglecf*, *Spp1*, and *Vcam1*), SigEMD (*Spp1*), edgeR (*Siglecf*, *Spp1*, and *Vcam1*), and DESeq2 (*Spp1*) identified 1 to 3 of the above functional-meaningful genes.

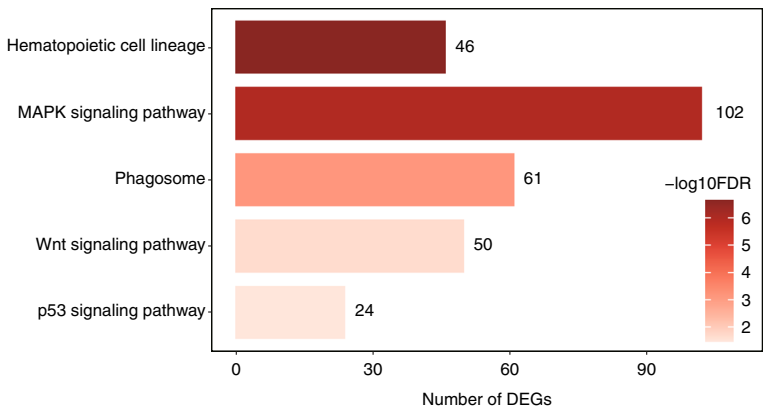

Fig. 6 Pulmonary fibrosis-related pathways enriched with DEGs identified by RankCompV3. The x-axis value (also shown next to the bars) indicates the number of DEGs and color indicates the FDR value of the enriched pathway

We performed pathway enrichment analysis (FDR < 0.05) for the DEGs identified by RankCompV3. The DEGs were enriched in 82 pathways, of which five were associated with pulmonary fibrosis (Fig. 6). Similarly, we performed pathway enrichment analysis on the DEGs identified by the other algorithms. Only five algorithms, Monocle2, SigEMD, DEsingle, edgeR, and DESeq2, recognized 1 to 3 of these functional pathways. Even when the significance threshold was relaxed to FDR < 0.2, the DEGs identified by Bimod and MAST were still not enriched in any of these pathways.

The Wnt signaling pathway was specifically identified by RankCompV3. The Wnt signaling pathway is an important pathway promoting pulmonary fibrosis [50], and studies have shown that targeting the Wnt pathway is a potential new treatment option for fibrosis [51].

In conclusion, RankCompV3 can detect weak biological signals that are functionally meaningful.

Discussion and conclusion

REOs are stable in normal samples, but they are often disrupted in diseased samples [52]. This led us to develop two versions of RankComp [7], which can be used to identify DEGs at both the population and individual levels. These algorithms are insensitive to batch effects and have the advantage of being able to integrate datasets from different sources.

Fisher's exact test was used in the original version of RankComp to calculate the significance level of the 2×2 contingency tables. RankCompV2 [8] added a filtering step to obtain a stable list of DEGs using non-DEGs as the background. This improvement circumvents the correlation effect between the true DEGs and their paired genes. But we failed to recognize that the REO analysis is a matched pairs design. The stability and direction of the REO of a gene-pair are interrelated in the two groups; they should not be counted independently.

In this study, we modified the contingency tables to tabulate the counts of 9 possible REO combinations for matched pairs design. McNemar's test is designed for the 2×2 contingency table of a matched pairs experiment. The McNemar–Bowker test extends it to the variables of general k categories other than dichotomous variables and it is also called symmetry test of contingency tables. However, in this study, the three REO outcomes are ranked categories, and McCullagh's method is a more appropriate choice. This method uses a logistic model to compare ordered categorical data in matched pairs experiments.

Previously, we showed that our RankComp algorithms based on REOs can be applied to microarray and bulk RNA-seq profiles and proteome profiles [11, 12, 53]. However, their applicability to scRNA-seq data has not been explored.

The scRNA-seq profiles tend to exhibit multi-mode expression patterns, heterogeneity, and sparsity compared to the bulk RNA-seq profiles. This makes it challenging to identify DEGs in scRNA-seq data. Many algorithms [54, 55] have been developed for scRNA-seq data to deal with dropouts [56, 57] or multi-mode patterns. However, these algorithms often cannot deal with both simultaneously. Additionally, many algorithms developed specifically for scRNA-seq have high FPRs and are affected by the number of cells and signal strength of the datasets. Algorithms developed for bulk RNA-seq data

have also been shown to perform well in scRNA-seq data [27, 28]. For example, limma, edgeR, and DESeq2 are all effective at identifying DEGs in scRNA-seq data.

In our study, we found that RankCompV3 exhibits an extremely low FPR in both simulated scRNA-seq data and a real negative dataset. Although DEseq2 and DEsingle perform well on the simulation dataset, but they have moderate or low performance on other datasets. This may indicate that they have a performance advantage in those datasets that satisfy an ideal distribution, but in real scenarios where the distribution is more complex, these methods often result in poorer performance. Additionally, RankCompV3 showed a lower FPR and higher accuracy than limma, edgeR, and DESeq2 in a positive test dataset of scRNA-seq.

Another advantage of the RankComp algorithms is that either the counts or normalized data (such as RPKM, FPKM, or TPM) can be used. As a heuristic method, RankCompV3 does not rely on a particular distribution of the gene profiles. The implication of normalization to DEG identification was discussed in a previous work [7].

Pseudo-bulking is a method that has been found to be effective in improving differential expression analysis in scRNA-seq data [26]. In pseudo-bulking, cells of the same type within a biological replicate are aggregated to a pseudo-bulk profile. This helps to lower the impact of dropouts in scRNA-seq, which are common due to the low sequencing depth of single cells.

Using the bulk profiles as the ground truth, pseudo-bulk analysis schemes can capture more true lowly expressed DEGs while lowering false positive with high expression. We have also implemented this method in RankCompV3. Tests with simulated and benchmark datasets show that our method produces concordant results using either single-cell or pseudo-bulk methods. The pseudo-bulk method slightly improves the performance compared to the single-cell method. The results in Squair et al.'s show that the performance of edgeR-LRT is better than 13 algorithms, especially the results of pseudo-bulk analysis are better than those of single-cell analysis. Through our previous analysis, edgeR also maintains better performance advantages than other 10 algorithms in multiple datasets. Therefore, we conducted comparative tests against edgeR-LRT using single-cell and pseudo-bulk data, and the results showed that our algorithm RankCompV3 performed better than edgeR.

In RankCompV3, the quantitative expression level is not used. This means that the method is not biased towards highly expressed genes. The advantage of denser read counts, which leads to a more accurate ordering of genes in the pseudo-bulk profiles, is largely compensated by the larger number of single-cell profiles through the binomial test of the stability of a REO in a group.

The influence of sample size on the performance of the evaluated methods was investigated. Most methods showed little improvement with increasing sample size. However, RankCompV3 showed a gradual increase in TPR with increasing sample size. In terms of TPR and FNR, RankCompV3 was second only to Monocle2. However, Monocle2 achieved its high TPR and FNR by including a large proportion of genes as DEGs in the full dataset. This resulted in extremely high FPR and low accuracy.

In contrast, RankCompV3 achieved strict FPR control while maintaining high precision and accuracy. This was possible because RankCompV3 does not rely on the quantitative expression level of genes. Instead, it uses a binomial test to assess the stability

of REOs in single-cell profiles. This makes RankCompV3 more robust to dropouts and other technical artifacts, and allows it to identify DEGs with high accuracy even in small datasets.

For weak-signal datasets, some common algorithms cannot capture differential expression signals. For example, in the pulmonary fibrosis dataset, LR, MAST, DEsingle, and Wilcoxon algorithms failed to identify or identify very few DEGs. However, our algorithm identified more DEGs than the other 11 algorithms, and the DEGs are significantly enriched with pulmonary fibrosis-related pathways.

One reason for detecting a relatively high number of DEGs is due to the less stringent control of the REO stability when the sample size is very small. In the case of a sample size of 4, the probability of observing all identical REO outcomes, e.g., $a < b$ in 4 profiles, is 6.25% using the binomial model, even if the two genes have the same expression levels. This implies that the off-diagonal elements of the contingency tables may be higher than expected with the preset significance threshold. This is a limitation of the method that might lead to high FPRs when the sample sizes are too small.

However, the functional study of the DEGs identified by RankCompV3 indicates that they are reliable in this weak signal dataset. This suggests that the algorithm is able to identify true DEGs even in the presence of noise and other technical artifacts.

In summary, the results of this study demonstrate that RankCompV3 is a promising algorithm for identifying DEGs in scRNA-seq data, even in small datasets with weak biological signals. It is able to achieve strict FPR control while maintaining high precision and accuracy, which makes it a valuable tool for identifying biologically relevant DEGs.

Implementation and runtime analysis

For computational efficiency reason, we implemented RankCompV3 in Julia, a modern scientific computing language that is both easy to use and has performance on par with C [58]. The RankCompV3 package can be directly added using the Pkg.add("RankCompV3") function in Julia and it can also be called in R via julia_installed_package("RankCompV3"). The source code is available at https://github.com/pathint/RankCompV3.jl.

A typical analysis takes a few minutes, depending on the sample sizes (number of profiles in each group), number of genes, and number of execution threads. The most time-consuming step in RankCompV3 is the comparison of a large number of gene pairs. The time complexity of a naïve implementation is $O(nN^2)$, where N is the number of genes passing the filtering step and n is the total sample size.

In Supplementary file 2: Table S2, we show the average runtimes of the 12 tools for the test in Fig. 3. Even using a single thread, RankCompV3 is able to achieve faster or similar speed compared with other algorithms. Furthermore, the time cost can be significantly reduced by increasing the number of execution threads. In Supplementary file 1: Fig. S3, we show the parallel runtimes for 1 to 8 threads on a single node.

Abbreviations

DEGs	Differentially expressed genes
scRNA-seq	Single-cell RNA sequencing
FPRs	False positive rates

REOs	Relative expression orderings
RNA-seq	High-throughput transcriptomic sequencing
BH	Benjamini–Hochberg
TPR	True positive rate
TNR	True negative rate
FPR	False positive rate
FNR	False negative rate
AUC	Area under the receiver operating characteristic curve
AUCC	Area under the concordance curve
DE	Differential expression of unimodal genes
DP	Differential proportion for multimodal genes
DM	Differential modality genes
DB	Both differential modality and different component genes
EE	Equivalent expression for unimodal
EP	Equivalent proportion for multimodal
GEO	Gene Expression Omnibus Database
mESCs	Mouse embryonic stem cells
RT-qPCR	Real-time quantitative reverse transcription PCR
LTHSCs	Long-term hematopoietic stem cells
Mo-AMs	Monocyte-derived alveolar macrophages
TR-AMs	Tissue-resident macrophages
KEGG	Kyoto Encyclopedia of Genes and Genomes
FDR	False discovery rate
FPs	False positives

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05889-1.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Availability and Requirements

Project name: RankCompV3. Project home page: https://github.com/pathint/RankCompV3.jl. Operating systems: Platform independent. Programming language: Julia. Other requirements: Julia 1.7.1 or higher. License: MIT License. Restrictions for use by non-academics: None.

Author contributions

JY: Methodology, Coding, Formal analysis, Data Curation, Investigation and Writing. QZ: Data Curation and Investigation. XW: Conceptualization, Methodology, Coding, Writing—Review & Editing, Supervision and Funding acquisition.

Funding

XW was supported by Fujian Medical University (Grant No. XRCZX2017001) and the Natural Science Foundation of Fujian Province (Grant No. 2019J01294).

Availability of data and materials

The scRNA-seq simulation dataset was generated using the scDD package. Other datasets were downloaded from the Gene Expression Omnibus Database (GEO, http://www.ncbi.nlm.nih.gov/geo/) and the accession codes are GSE82158, GSE54695, GSE29087 and GSE59114. The source code is publicly available at https://github.com/pathint/RankCompV3.jl.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 5 December 2023 Accepted: 30 July 2024

Published online: 07 August 2024

References

1. Siavoshi A, Taghizadeh M, Dookhe E, et al. Gene expression profiles and pathway enrichment analysis to identification of differentially expressed gene and signaling pathways in epithelial ovarian cancer based on high-throughput RNA-seq data. Genomics. 2022;114(1):161–70.
2. Wang H, Nie X, Li X, et al. Bioinformatics analysis and high-throughput sequencing to identify differentially expressed genes in nebulin gene (NEB) mutations mice. Med Sci Monit. 2020;26: e922953.
3. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11(10):733–9.
4. Leek JT, Johnson WE, Parker HS, et al. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882–3.
5. Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res. 2014;42(21): e161.
6. Wang D, Cheng L, Zhang Y, et al. Extensive up-regulation of gene expression in cancer: the normalised use of microarray data. Mol Biosyst. 2012;8(3):818–27.
7. Cai H, Li X, Li J, et al. Identifying differentially expressed genes from cross-site integrated data based on relative expression orderings. Int J Biol Sci. 2018;14(8):892–900.
8. Li X, Cai H, Wang X, et al. A rank-based algorithm of differential expression analysis for small cell line data with statistical control. Brief Bioinform. 2019;20(2):482–91.
9. Xie J, Xu Y, Chen H, et al. Identification of population-level differentially expressed genes in one-phenotype data. Bioinformatics. 2020;36(15):4283–90.
10. Yan H, Guan Q, He J, et al. Individualized analysis reveals CpG sites with methylation aberrations in almost all lung adenocarcinoma tissues. J Transl Med. 2017;15(1):26.
11. Song K, Su W, Liu Y, et al. Identification of genes with universally upregulated or downregulated expressions in colorectal cancer. J Gastroenterol Hepatol. 2019;34(5):880–9.
12. Hu G, Cheng Z, Wu Z, et al. Identification of potential key genes associated with osteosarcoma based on integrated bioinformatics analyses. J Cell Biochem. 2019;120(8):13554–61.
13. Wang R, Zheng X, Wang J, et al. Improving bulk RNA-seq classification by transferring gene signature from single cells in acute myeloid leukemia. Brief Bioinform. 2022;23(2): bbac002.
14. Wu Q, Zheng X, Leung KS, et al. meGPS: a multi-omics signature for hepatocellular carcinoma detection integrating methylome and transcriptome data. Bioinformatics. 2022;38(14):3513–22.
15. McCullagh P. A logistic model for paired comparisons with ordered categorical data. Biometrika. 1977;64(3):449–53.
16. Finak G, McDavid A, Yajima M, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16:278.
17. Miao Z, Deng K, Wang X, et al. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Bioinformatics. 2018;34(18):3223–4.
18. Wilcoxon F. Individual comparisons by ranking methods. Biom Bull. 1945;1(6):80–3.
19. Qiu X, Hill A, Packer J, et al. Single-cell mRNA quantification and differential analysis with Census. Nat Methods. 2017;14(3):309–15.
20. Wang T, Nabavi S. SigEMD: a powerful method for differential gene expression analysis in single-cell RNA sequencing data. Methods. 2018;145:25–32.
21. Korthauer KD, Chu LF, Newton MA, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016;17(1):222.
22. Ritchie ME, Phipson B, Wu D, et al. LIMMA powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7): e47.
23. Law CW, Chen Y, Shi W, et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29.
24. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
25. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
26. Squair JW, Gautier M, Kathe C, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021;12(1):5692.
27. Jaakkola MK, Seyednasrollah F, Mehmood A, et al. Comparison of methods to detect differentially expressed genes between single-cell populations. Brief Bioinform. 2017;18(5):735–43.
28. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018;15(4):255–61.
29. Van den Berge K, Perraudeau F, Soneson C, et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. Genome Biol. 2018;19(1):24.
30. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
31. Angelidis I, Simon LM, Fernandez IE, et al. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. Nat Commun. 2019;10(1):963.
32. Cano-Gamez E, Soskic B, Roumeliotis TI, et al. Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4(+) T cells to cytokines. Nat Commun. 2020;11(1):1801.
33. Hagai T, Chen X, Miragaia RJ, et al. Gene expression variability across cells and species shapes innate immunity. Nature. 2018;563(7730):197–202.
34. Reyfman PA, Walter JM, Joshi N, et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. Am J Respir Crit Care Med. 2019;199(12):1517–36.
35. Grun D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014;11(6):637–40.
36. Wang T, Li B, Nelson CE, et al. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinform. 2019;20(1):40.

37. Wang X, Chen H. Prognosis prediction through an integrated analysis of single-cell and bulk RNA-sequencing data in triple-negative breast cancer. Front Genet. 2022;13: 928175.
38. Islam S, Kjallquist U, Moliner A, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011;21(7):1160–7.
39. Moliner A, Enfors P, Ibanez CF, et al. Mouse embryonic stem cell-derived spheres with distinct neurogenic potentials. Stem Cells Dev. 2008;17(2):233–43.
40. Kowalczyk MS, Tirosh I, Heckl D, et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res. 2015;25(12):1860–72.
41. Misharin AV, Morales-Nebreda L, Reyfman PA, et al. Monocyte-derived alveolar macrophages drive lung fibrosis and persist in the lung over the life span. J Exp Med. 2017;214(8):2387–404.
42. Kanehisa M, Goto S, Sato Y, et al. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40(Database issue):D109-114.
43. Fury W, Batliwalla F, Gregersen PK, et al. Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. Conf Proc IEEE Eng Med Biol Soc. 2006;2006:5531–4.
44. Misharin AV, Morales-Nebreda L, Mutlu GM, et al. Flow cytometric analysis of macrophages and dendritic cell subsets in the mouse lung. Am J Respir Cell Mol Biol. 2013;49(4):503–10.
45. Agassandian M, Tedrow JR, Sembrat J, et al. VCAM-1 is a TGF-beta1 inducible gene upregulated in idiopathic pulmonary fibrosis. Cell Signal. 2015;27(12):2467–73.
46. Wong SL, Sukkar MB. The SPARC protein: an overview of its role in lung cancer and pulmonary fibrosis and its potential role in chronic airways disease. Br J Pharmacol. 2017;174(1):3–14.
47. Koo HY, El-Baz LM, House S, et al. Fibroblast growth factor 2 decreases bleomycin-induced pulmonary fibrosis and inhibits fibroblast collagen production and myofibroblast differentiation. J Pathol. 2018;246(1):54–66.
48. Polverino F, Rojas-Quintero J, Wang X, et al. A Disintegrin and metalloproteinase domain-8: a novel protective proteinase in chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2018;198(10):1254–67.
49. Morse C, Tabib T, Sembrat J, et al. Proliferating SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis. Eur Respir J. 2019;54(2): 1802441.
50. Kulkarni YM, Dutta S, Iyer AK, et al. A proteomics approach to identifying key protein targets involved in VEGF inhibitor mediated attenuation of bleomycin-induced pulmonary fibrosis. Proteomics. 2016;16(1):33–46.
51. Burgy O, Konigshoff M. The WNT signaling pathways in wound healing and fibrosis. Matrix Biol. 2018;68–69:67–80.
52. Guan Q, Chen R, Yan H, et al. Differential expression analysis for individual cancer samples based on robust within-sample relative gene expression orderings across multiple profiling platforms. Oncotarget. 2016;7(42):68909–20.
53. Yan H, He J, Guan Q, et al. Identifying CpG sites with different differential methylation frequencies in colorectal cancer tissues based on individualized differential methylation analysis. Oncotarget. 2017;8(29):47356–64.
54. Sekula M, Gaskins J, Datta S. Single-cell differential network analysis with sparse Bayesian factor models. Front Genet. 2021;12: 810816.
55. Mallick K, Chakraborty S, Mallik S, et al. A scalable unsupervised learning of scRNAseq data detects rare cells through integration of structure-preserving embedding, clustering and outlier detection. Brief Bioinform. 2023;24(3): bbad125.
56. Seth S, Mallik S, Islam A, et al. Identifying genetic signatures from single-cell RNA sequencing data by matrix imputation and reduced set gene clustering. Mathematics. 2023;11(20):4315.
57. Seth S, Mallik S, Bhadra T, et al. Dimensionality reduction and Louvain agglomerative hierarchical clustering for cluster-specified frequent biomarker discovery in single-cell sequencing data. Front Genet. 2022;13: 828479.
58. Roesch E, Greener JG, MacLean AL, et al. Julia for biologists. Nat Methods. 2023;20(5):655.

Publisher's Note