Title

Differential variability analysis of single-cell gene expression data.

Permalink
https://escholarship.org/uc/item/9f33c997

Journal
Briefings in Bioinformatics, 24(5)

Authors
Liu, Jiayi
Kreimer, Anat
Li, Wei Vivian

Publication Date
2023-09-20

DOI
10.1093/bib/bbad294

Peer reviewed

Briefings in Bioinformatics, 2023, **24(5)**, 1–13
https://doi.org/10.1093/bib/bbad294
Advance access publication date 20 August 2023
Problem Solving Protocol

Differential variability analysis of single-cell gene expression data

Jiayi Liu, Anat Kreimer and Wei Vivian Li

Corresponding authors. Anat Kreimer, Center for Advanced Biotechnology and Medicine, Rutgers, The State University of New Jersey, 679 Hoes Lane West, Piscataway, 08854, NJ, USA. Email: kreimer@cabm.rutgers.edu; Wei Vivian Li, Department of Statistics, University of California, Riverside, 900 University Ave, Riverside, CA 92521, USA. Email: weil@ucr.edu

Abstract

The advent of single-cell RNA sequencing (scRNA-seq) technologies has enabled gene expression profiling at the single-cell resolution, thereby enabling the quantification and comparison of transcriptional variability among individual cells. Although alterations in transcriptional variability have been observed in various biological states, statistical methods for quantifying and testing differential variability between groups of cells are still lacking. To identify the best practices in differential variability analysis of single-cell gene expression data, we propose and compare 12 statistical pipelines using different combinations of methods for normalization, feature selection, dimensionality reduction and variability calculation. Using high-quality synthetic scRNA-seq datasets, we benchmarked the proposed pipelines and found that the most powerful and accurate pipeline performs simple library size normalization, retains all genes in analysis and uses denSNE-based distances to cluster medoids as the variability measure. By applying this pipeline to scRNA-seq datasets of COVID-19 and autism patients, we have identified cellular variability changes between patients with different severity status or between patients and healthy controls.

Keywords: single-cell genomics, differential variability analysis, hypothesis testing, data normalization

INTRODUCTION

In multi-cellular organisms, differences among single cells manifest in the forms of phenotypic, genetic, transcriptomic and epigenetic heterogeneity [1, 2]. It is now well acknowledged that bulk sequencing technologies prevent the precise quantification of cellular heterogeneity when a population of cells is sequenced together. In contrast, single-cell omics profiling techniques provide a much higher resolution for revealing cell-to-cell variability. Among these single-cell techniques, the single-cell RNA sequencing (scRNA-seq) technologies have been extensively utilized to investigate transcriptional variation and regulation among individual cells [3–5].

Data generated by scRNA-seq experiments have shown that individual cells of the same cell type still present cell-to-cell variability in gene expression, and such variability is required for cell-type-specific, higher level system functions [6–8]. For example, differentially regulated genes and distinct nutrient absorption preferences have been found among Paneth cells from different regions of the intestine [9, 10]. Furthermore, transcriptional variability of the same cell type is observed to change in response to biological conditions. Several studies have reported increased cell-to-cell variability preceding irreversible commitment in differentiation processes [11, 12]. In addition, T cells and cardiomyocytes were found to exhibit high transcriptional variability in aged mice [13, 14].

Although cell-to-cell transcriptional variability has been frequently observed in single-cell studies, there have been limited attempts to quantify to what extent transcriptional variability changes between biological conditions. A handful of computational methods have been developed to detect differential variability of individual genes from bulk-tissue or single-cell gene expression data [15, 16], but their reliance on specific parametric models could lead to reduced power when model assumptions are violated. New visualization tools, denSNE and densMAP [17], enable visual comparison of relative transcriptional variability of cell groups in a lower dimensional space. However, these tools do not offer a rigorous computational comparison.

Motivated by the scarcity of computational tools to study differential variability between two groups of single cells, we proposed and benchmarked 12 pipelines for quantifying and comparing the transcriptional variability of two cell groups (e.g. one group of T cells from healthy donors and one group from cancer patients). Rather than focusing on individual genes, our study emphasizes the overall cellular variability of a cell group. Hence, we consider a distance-based variability measure as the proxy of transcriptomic dissimilarities of the cells within a cell group. Then, we further define differential variability as statistically significant

Jiayi Liu is a second-year Ph.D. student at the Rutgers University. She is interested in computational biology and translational research.
Anat Kreimer is an Assistant Professor at the Rutgers University. She completed her Ph.D. in Biomedical Informatics at Columbia University. Her main research interests are focused on developing predictive models of transcriptional regulation by integrating large-scale datasets to shed light on regulatory processes that are condition-specific.
Wei Vivian Li is an Assistant Professor in the Department of Statistics at the University of California, Riverside. She has a Ph.D. in Statistics and her research focuses on statistical and machine learning methods for biological and biomedical data. Dr. Li was at Rutgers School of Public Health, where part of this work was done.

Received: April 10, 2023. **Revised:** July 18, 2023. **Accepted:** July 29, 2023

differences in this variability measure between two cell groups. All 12 pipelines for detecting differential variability take read counts and cell group labels as input and output *P*-values indicating the statistical significance of observed variability changes. In these pipelines, we consider different combinations of methods for normalization, dimensionality reduction and variability calculation.

We conducted a comprehensive evaluation of the 12 pipelines using high-quality synthetic datasets, which were generated with ground truth information of differential variability. Based on our analyses, we provide guidance on the selection of normalization and variability calculation methods in differential variability analysis. In particular, our analyses have identified the most powerful pipeline, which performs simple library size normalization, retains all genes in analysis and calculates the distances to medoids using the density-preserving t-SNE embeddings as the variability measure. Furthermore, we demonstrated the applicability of this pipeline in a real-data analysis of immune cells from COVID-19 patients. Our findings showed that 14 out of 17 immune cell types differed significantly in cellular variability between mild and severe disease stages. A second application on autism spectrum disorder (ASD) revealed 14 out of 17 cortical cell types that differed significantly in cellular variability between ASD patients and healthy controls. To the best of our knowledge, there is currently no established pipeline that can identify and quantify changes in overall cellular variability using single-cell gene expression data. Our study bridges this gap by constructing and evaluating computational pipelines, as well as investigating the effectiveness of individual steps in these pipelines. Our findings provide a new perspective for comparing single-cell populations between diverse biological conditions.

METHODS

Pipelines for differential variability analysis of scRNA-seq data

Since there are no existing methods for comparing cellular variability based on single-cell gene expression data of two conditions, we propose multiple differential variability analysis pipelines, all of which consist of the following five key steps: (1) normalization, (2) feature selection, (3) dimensionality reduction, (4) variability calculation and (5) statistical testing. In particular, we consider three methods of normalization, two methods of feature selection and two methods of variability calculation, resulting in 12 distinct pipelines for evaluation and comparison.

Normalization

Normalization is a critical step in scRNA-seq analysis to address various technical effects in the sequencing process [18]. We consider three normalization methods for the count data, TP10K, LogTP10K and SCT, and we refer to the normalized counts as gene expression levels (Figure 1).

The TP10K method calculates the number of transcripts per 10,000 transcripts in a cell. For each gene in each cell, its UMI count is first divided by the total UMI count in that cell, and then scaled by multiplying 10,000. The LogTP10K method represents the TP10K calculation followed by the logarithm transformation. For each gene in each cell, its TP10K value is first added by a pseudo-count of 1, and then transformed using the natural logarithm. The SCT method stands for sctransform, which is a normalization method proposed by Hafemeister and Satija [19]. For each gene, sctransform fits a Negative Binomial regression model on its UMI counts, using the total number of UMI counts in individual cells as an explanatory variable. Based on the fitted regression models, the Pearson residuals are calculated for each gene in each cell and then treated as the normalized gene expression levels.

Feature selection

Selection of gene features is a common step in scRNA-seq analysis before model-based dimensionality reduction [20, 21]. Even though feature selection has been shown to be effective in clustering and classification analyses of single-cell gene expression data, it is not clear if it also contributes to differential variability analysis. In our study, we consider two methods for comparison. The first method is to use all detected genes without feature selection. The second method is to select the top 2000 variable genes using the 'vst' method implemented in the Seurat package [22]. Then, only the normalized expression levels of these genes are used in subsequent steps.

Dimensionality reduction

We consider two methods for dimensionality reduction: principal component analysis (PCA) and density-preserving stochastic neighbor embedding (denSNE) [17]. First, for each gene, the normalized expression levels across individual cells are centered and scaled, so the mean expression becomes 0, and the variance becomes 1. When PCA is used, it is performed on the scaled data to calculate the principal components (PCs) and corresponding scores. The top 200 PCs are then used as the input for downstream variability calculation.

The denSNE method modifies the objective function used by tSNE [17] to better preserve the local density in data. Following the original work of denSNE to first produce lower dimensional representations of individual cells using PCA, denSNE is applied on the top 200 PCs (calculated as described above) to obtain two-dimensional denSNE embeddings.

Variability calculation

We use a matrix X to denote the cells' coordinates obtained from the dimensionality reduction step. X is either a $200 \times J$ or a $2 \times J$ matrix with columns representing cells and rows representing 200 PCs or two embedded denSNE features. For cell j, its coordinate vector in the reduced-dimensional space is denoted as X_j. These vectors are used to calculate the variability measures.

We use the distance to the medoid (DM) as the variability measure. As a distance-based measure, the DM represents dissimilarity between the transcriptomic profile of each individual cell and the 'average' transcriptomic profile of the same cell type. If a cell type has increased variability in one condition than another, the distribution of DM is expected to have a larger mean in the former condition. The calculation of the DM measure requires cell type labels so that the cell type medoids can be obtained. In our simulation study, we directly used the K cell-type labels obtained from the simulation process. In real practice, the labels could be provided by the users as the input information of the pipelines.

For each cell type, we first identify its medoid cell, assuming that there are C cells in the cell type. The index of the medoid cell can be determined as follows:

$$j^* = \underset{j_0 \in \{1,\ldots,C\}}{\operatorname{argmin}} \sum_{j \in \{1,\ldots,C\}} d(X_{j_0}, X_j), \tag{1}$$

where $d(\cdot)$ represents the Euclidean distance function. Then, cell j's distance to medoid is defined as its euclidean distance to the

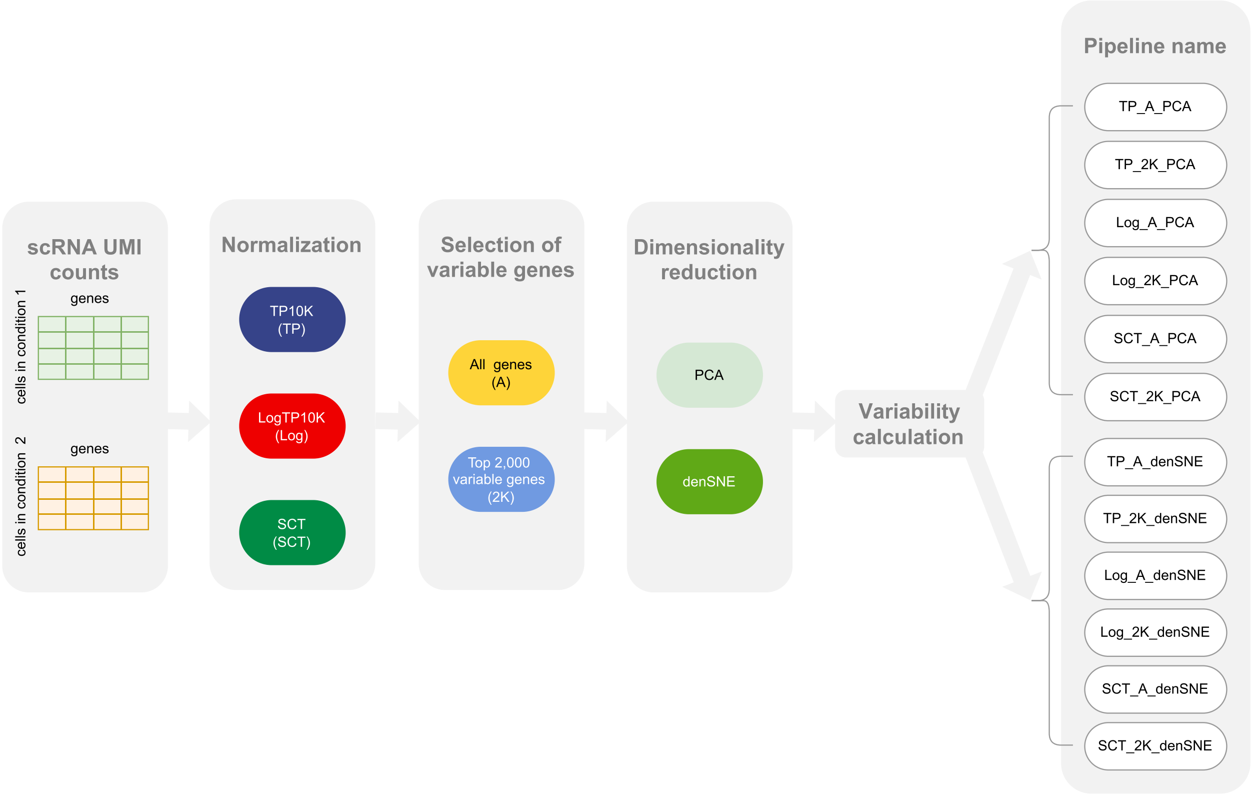

Figure 1. An outline of the 12 pipelines for differential variability analysis. For simplicity, we name the pipelines as X_Y_Z, where X stands for the normalization method ('TP' for TP10K, 'Log' for LogTP10K and 'SCT' for sctransform), Y stands for the feature selection method ('A' for all genes and '2K' for top 2000 variable genes) and Z stands for the dimensionality reduction method ('PCA' for principal component analysis and 'denSNE' for density-preserving t-SNE). The variability measure is defined as the distance to the medoid (DM), which is the distance between a single cell and the medoid of the cell type it belongs to (see Methods for details). All the pipelines take count matrices as input and output *P* values for the differential variability analysis.

corresponding medoid cell:

$$\mathrm{DM}_j = d(X_j, X_{j^*});\ j \in \{1, \ldots, C\}\,. \qquad (2)$$

The above calculation process is repeated for every cell type.

Testing of differential variability

We consider the problem of comparing transcriptional variability of the same cell type between two conditions. First, the count data of cells from different cell types and conditions are analyzed jointly to perform normalization, feature selection and dimensionality reduction. Second, the cells are separated by cell types and conditions before the calculation of distances to medoids. After we obtain the DMs of individual cells in the two conditions, a Wilcoxon rank sum test is performed for each cell type to decide if there is a significant difference in cell type variability between the two conditions.

Simulation of scRNA-seq data

We utilized the scDesign2 simulator (v0.1.0) to generate single-cell gene expression data, so that we can evaluate the pipelines on data with ground truth information about the change of transcriptional variability [23, 24]. Briefly, we simulated single-cell gene expression data of six cell types for two biological conditions, C1 and C2. In detail, our simulation assumed that two cell types (I1 and I2) had increased variability in C2 compared with C1, two cell types had decreased variability (D1 and D2) and two cell types (R1 and R2) had the same variability between the two conditions. We introduce our simulation procedure in detail below.

First, to mimic real data characteristics, we used scDesign2 to learn gene expression parameters from a single-cell gene expression dataset of six immune cell types (B cells, monocytes, neutrophils, platelets, red blood cells and T cells). The dataset was obtained from the immune cells of primary non-small-cell lung cancer patients [25]. The expression parameters learned from these real data were used to set up the expression parameters in conditions C1 and C2 in the next step.

Second, we generated two sets of gene expression parameters for the two conditions, C1 and C2. For cell types I1 and I2, we directly used the estimated gene expression parameters from real data as the parameters in condition C1. Since their variability increased in condition C2, we randomly selected half of the genes and increased their dispersion parameters by factors randomly drawn from Unif(1.5, 3.0). For cell types D1 and D2, we directly used the estimated gene expression parameters from real data as the parameters in condition C2, and modified the dispersion parameters in condition C1 as described above. For cell types R1 and R2, their parameters were the same in the two conditions, and directly set to the estimated ones from real data.

Finally, we used scDesign2 to simulate gene expression count matrices of the six cell types in the two conditions based on the above gene expression parameters. As an example, we randomly selected one simulated dataset with 200 cells per cell type and visualized the cells in Figure 2 and Supplementary Figure 1.

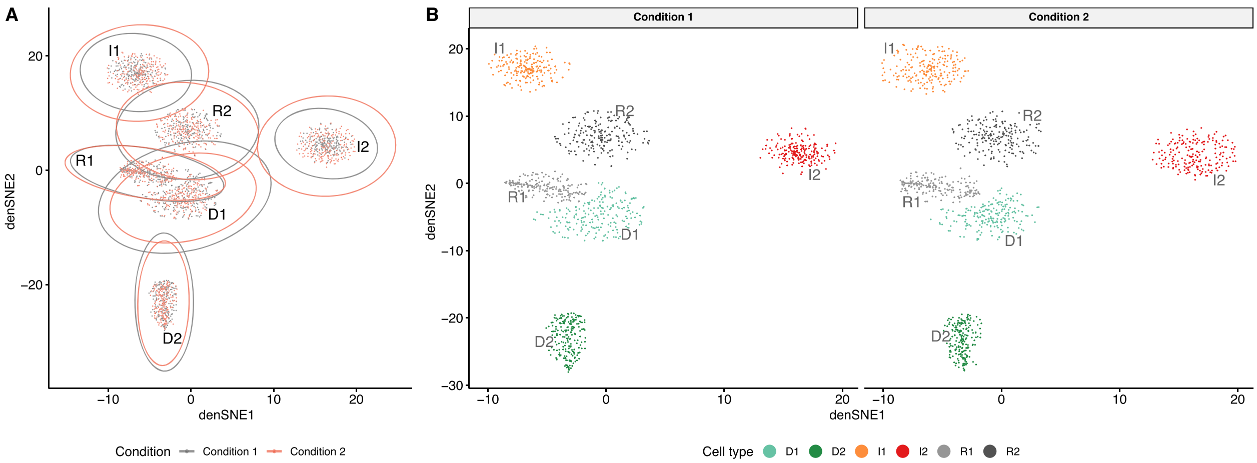

Figure 2. The denSNE embeddings of an example simulated dataset. The TP10K normalization method and all genes were used for the dimensionality reduction step. (**A**): Cells are colored by the two conditions. For each cell type in each condition, the ellipse outlines the region that covers around 95% of cells from that cell type (estimated using a multivariate normal distribution). (**B**): Cells are colored by the six cell types, and cells are separately shown for the two conditions.

As expected, cell types D1 and D2 had smaller variability in condition C2, so the cells had increased densities in the denSNE and PCA plots, compared with cells from the same type in condition C1; cell types I1 and I2 had larger variability in condition C2, so the cells had decreased densities compared with cells from the same type in condition 1.

To evaluate the effect of available cell numbers on the performance of the pipelines, we simulated data in four distinct scenarios, each comprising 100, 200, 500 and 1000 cells per cell type and condition. For each scenario, the simulation was independently repeated 1000 times to allow for the calculation of type I error and statistical power in differential variability analysis.

In our simulation study, we applied two-sided Wilcoxon rank sum tests to data of cell types R1 and R2, so as to evaluate the pipelines' type I error given no change in cellular variability between conditions. In addition, we applied one-sided Wilcoxon tests to data of the remaining four cell types (D1, D2, I1 and I2) in order to evaluate the ability of different pipelines to correctly detect significant variability changes and their respective directions. Specifically, for cell types D1 and D2 (I1 and I2), the alternative hypothesis posits that the cellular variability is lower (greater) in C2 than in C1.

Real data analysis

COVID-19 data analysis

In this real data application, we used the scRNA-seq count data of 36 COVID-19 patients [26] who exhibited mild or severe symptoms at the time of blood sample collection. To perform the differential variability analysis, we leveraged the cell type annotations from the original publication and considered 17 cell types, including B cells, CD4$^+$ T cells (CD4), CD8$^+$ T cells (CD8), dendritic cells (DC), $\gamma\delta$ T cells (gdT), hematopoietic stem cells (HSC), proliferating lymphocytes (Lymph_prolif), mucosal-associated invariant T cells (MAIT), CD14$^+$ monocytes (Mono1), CD16$^+$ monocytes (Mono2), CD16$^+$ natural killer cells (NK_16hi), CD56$^+$ natural killer cells (NK_56hi), plasmacytoid dendritic cells (pDC), plasmablasts, platelets, red blood cells (RBC) and regulatory T cells (Treg). We utilized the TP_A_denSNE pipeline to compare the variability of each cell type in patients between two clinical conditions of COVID-19 patients: mild and severe disease phase. The two-sided Wilcoxon rank sum tests were used in this analysis.

ASD data analysis

In this application, we used the scRNA-seq count data of the cortical tissue from 14 ASD patients diagnosed with ASD and 16 healthy controls [27]. To perform the differential variability analysis, we used the 17 cell types annotated by the original publication, including fibrous astrocytes (AST-FB), protoplasmic astrocytes (AST-PP), endothelial cells (Endothelial), parvalbumin interneurons (IN-PV), somatostatin interneurons (IN-SST), synaptic vesicle glycoprotein 2C-expressing interneurons (IN-SV2C), vasoactive intestinal polypeptide–expressing interneurons (IN-VIP), layer 2/3 excitatory neurons (L2/3), layer 4 excitatory neurons (L4), layer 5/6 corticofugal projection neurons (L5/6), layer 5/6 corticocortical projection neurons (L5/6-CC), microglia, maturing neurons (Neu-mat), neurogranin-expressing neurons subpopulation I (Neu-NRGN-I), neurogranin-expressing neurons subpopulation II (Neu-NRGN-II), oligodendrocytes and oligodendrocyte precursor cells. Then, we utilized the TP_A_denSNE pipeline to compare the variability of each cell type between autism patients and healthy donors. The two-sided Wilcoxon rank sum tests were used in this analysis.

Gene set enrichment analysis

To investigate the individual genes that contributed to the cell type variability change and their biological functions, we performed the gene set enrichment analysis on the COVID-19 and ASD datasets, respectively. First, we calculated the expression variance of each gene in each cell type based on the TP10K-normalized expression levels, and then selected the top 25% genes that exhibited higher (or lower) variances in cell types with increased (or decreased) variability. For the COVID-19 dataset, we performed the enrichment analysis using pathways in the Pathway Interaction Database (MSigDB v2023.1.Hs C2:PID) [28, 29]. For the ASD dataset, we performed the enrichment analysis using pathways in the Pathway Interaction Database (MSigDB v2023.1.Hs C2:PID) and the Human Phenotype Ontology (MSigDB v2023.1.Hs C5:HPO) [28, 30]. The analyses were implemented using the R package ClusterProfiler [31]. The enriched pathways were defined as those with an FDR-adjusted *P*-value < 0.05.

RESULTS

Effect of normalization methods in differential variability analysis

To investigate the effect of various normalization methods in differential variability analysis of single cells, we first compared the pipelines' performance with a focus on comparing the three normalization methods, TP10K, LogTP10K and SCT, as described in Methods. For this comparison, all gene features were used when performing PCA.

When the denSNE embeddings were used to calculate the variability measure, the statistical power of the TP10K-based pipeline (blue bars) was consistently better than the pipelines based on LogTP10K (red bars) or SCT (green bars) (Figure 3A). It is also worth noting that, when the cell number was relatively low (100 cells per cell type), the TP10K-based pipeline had difficulty in detecting decreasing variability from condition C1 to C2, but its statistical power was between 0.642 and 1 when the cell number was at least 200 cells per cell type. In contrast, the performance of the SCT-based pipeline was not consistent across cell types. It successfully detected increasing variability (cell types I1 and I2) when the cell number was at least 200 per cell type, but it failed to detect decreasing variability (cell types D1 and D2) regardless of cell numbers. As for the LogTP10K-based pipeline, it performed poorly in all combinations of cell types and cell numbers. In terms of the type I error, all three pipelines were able to control the type I error below the nominal level of 0.05 (Figure 3B and Supplementary Table 1).

When the PC scores were used to calculate the variability measure, the TP10K normalization still achieved the best power in most cases, followed by SCT and LogTP10K (Figure 3C). However, the PCA-based pipelines often led to high type I errors regardless of the normalization method used (Figure 3D and Supplementary Table 1). In addition, we observed that the LogTP10K normalization yielded better power when used with the PCA-based variability measure, while the TP10K normalization tended to have a better power when used with the denSNE-based variability measure.

To explore the reasons why the LogTP10K-based pipelines performed the worst, and why the SCT-based pipelines had varying power across the four cell types, we compared the calculated gene variances between conditions C2 and C1 based on the normalized expression levels (Figure 4). According to the ground truth from the simulation process, the log-ratios of median gene variances were expected to be negative in cell types D1 and D2 and positive in cell types I1 and I2. In our analyses, the signs of log-ratios after TP10K normalization were consistent with the expected signs in all the scenarios. However, the log-ratios after LogTP10K normalization always had opposite signs. This phenomenon explains the poor performance of LogTP10K-based pipelines in detecting differential variability of cells. Besides, the SCT-based log-ratios were always positive except when the cell number was 500 or 1000 in cell type D1, which explains why the SCT-based pipelines only had good power on cell types I1 and I2.

Next, we further investigated if the performance of the normalization methods depends on the mean expression levels of genes, since the dependence between expression mean and variance is widely observed in scRNA-seq data [32]. We used one simulation dataset with 100 cells per cell type as an example, and compared the calculated gene variances based on the normalized data between two gene groups: the low expression group and the high expression group (Figure 5). We found that, for the high expression group, the normalized data always presented variance changes consistent with the ground truth directions, regardless of normalization methods used. Yet, with the LogTP10K normalization, the gene variances calculated for the low expression group could not reflect the actual variance changes between the two conditions for any of the four cell types. With the SCT normalization, the calculated variances of the low expression group always increased in condition C2 regardless of the ground truth changes. These comparisons suggest that the LogTP10K and SCT normalization methods cannot guarantee to retain the relative relationships of gene variances in the normalization process, especially for genes with relatively low expression levels.

Effect of feature selection in differential variability analysis

Since TP10K and SCT have demonstrated obviously better performance in the aforementioned results, we next evaluated how the selection of gene features affected the differential variability analysis based on these two normalization methods. For simplicity, we name the pipelines as X_Y_Z, where X refers to the normalization method ('TP' for TP10K and 'SCT' for sctransform), Y stands for the feature selection method ('A' for all genes and '2K' for top 2000 variable genes) and Z stands for the dimensionality reduction method.

First, we evaluated the impact of feature selection using the pipelines based on TP10K normalization, by comparing TP_2K_denSNE versus TP_A_denSNE and TP_2K_PCA versus TP_A_PCA (Figure 6). Using the denSNE-based variability, the pipelines based on both feature sets (TP_2K_denSNE and TP_A_denSNE) performed comparably in detecting increasing variability (cell types I1 and I2), whereas TP_A_denSNE performed better in detecting decreasing variability (cell types D1 and D2) when the cell number was relatively low (Figure 6A). Moreover, both pipelines showed well-controlled type I errors (Figure 6B and Supplementary Table 2). However, when using the PCA-based variability measure, TP_2K_PCA consistently had larger power and smaller type I errors than TP_A_PCA. Yet, the type I errors could not be controlled under the target level, regardless of the gene feature set being used (Figure 6C-D and Supplementary Table 2).

Next, we evaluated SCT normalization-based pipelines by comparing SCT_2K_denSNE versus SCT_A_denSNE and SCT_2K_PCA versus SCT_A_PCA (Supplementary Figure 2). The two pipelines using denSNE-based measures (SCT_2K_denSNE and SCT_A_denSNE) were able to control type I errors (Supplementary Figure 2A-B and Supplementary Table 3), but both had an unstable power in detecting variability change (Supplementary Figure 2A). The SCT_A_denSNE pipeline performed slightly better, but exhibited a relatively high power only in cell types I1 and I2 when the cell number was at least 200 per cell type. This is consistent with our prior finding that the SCT normalization was not able to preserve relative gene variance in all combinations of cell types and cell numbers. Besides, between the two PCA-based pipelines, SCT_A_PCA and SCT_2K_PCA, SCT_A_PCA had slightly higher power in most cases, but the type I errors were again not well controlled (Supplementary Figure 2C-D and Supplementary Table 3).

Comparison of the 12 pipelines for differential variability analysis

After evaluating the individual effects of normalization methods and feature selection in differential variability analysis, we sought to determine the pipelines with the best accuracy among all possible combinations of normalization (TP10K or SCT), feature selection and dimensionality reduction methods. We compared eight

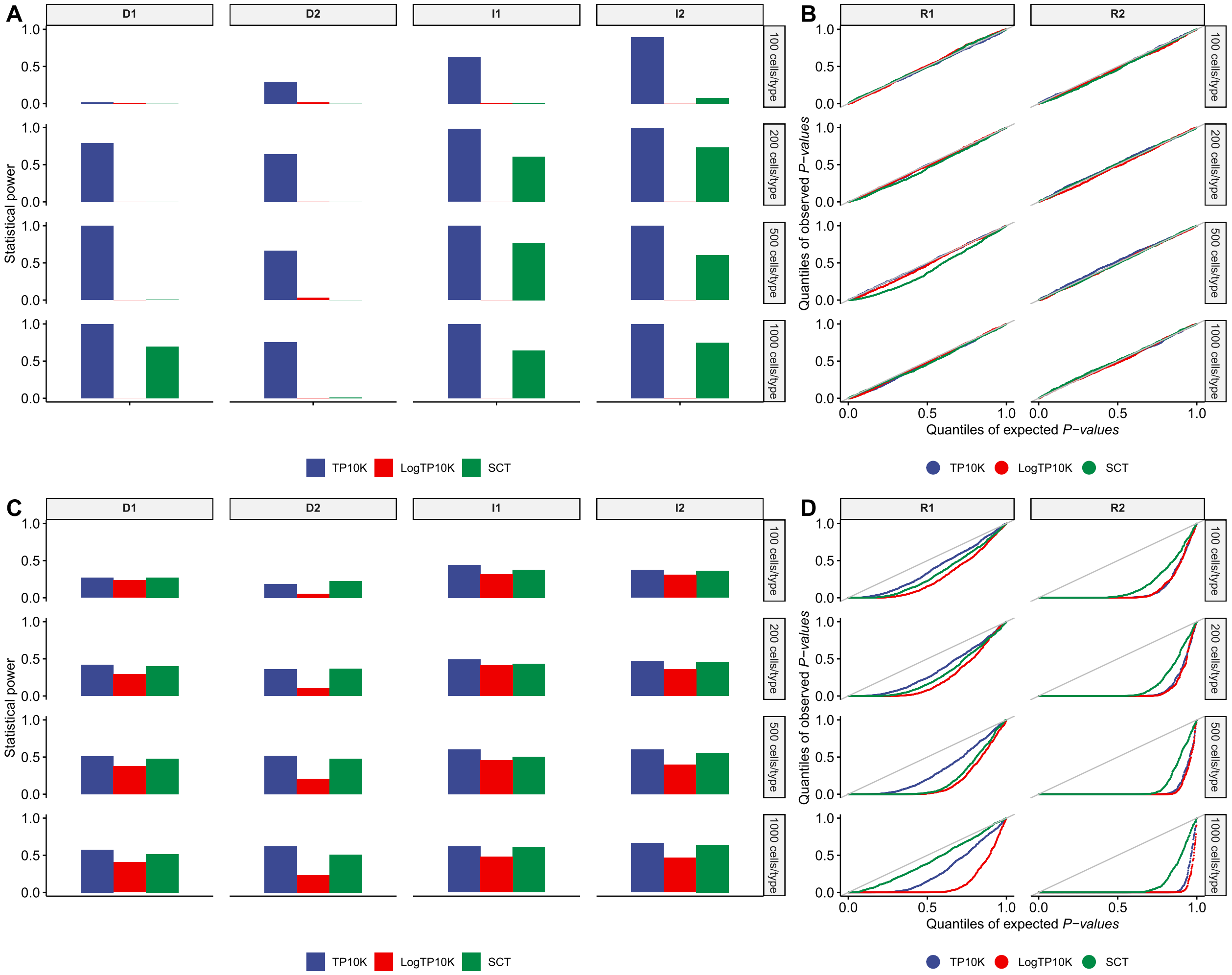

Figure 3. A comparison of the three normalization methods in differential variability analysis. (**A**): Statistical power of denSNE-based pipelines for cell types D1, D2, I1 and I2. (**B**): Quantile–quantile plots comparing distributions of expected *P*-values and *P*-values from denSNE-based pipelines under the null hypothesis. (**C**): Statistical power of PCA-based pipelines for different cell types. (**D**): Quantile–quantile plots comparing distributions of expected *P*-values and *P*-values from PCA-based pipelines under the null hypothesis.

pipelines (TP_2K_denSNE, TP_A_denSNE, TP_2K_PCA, TP_A _PCA, SCT_2K_denSNE, SCT_A_denSNE, SCT_2K_PCA and SCT_A_PCA), and excluded the LogTP10K-based pipelines as previous results have shown that LogTP10K normalization cannot correctly retain the relative relationship among gene variances.

The comparison of statistical power shows that the TP10K-based pipelines overall had a better power in detecting differential variability than the SCT-based pipelines accounting for different cell types and cell numbers (Figure 7A). The TP_A_denSNE and TP_2K_denSNE pipelines had the largest average power across different scenarios, followed by the TP_2K_PCA and TP_A_PCA pipelines. In contrast, the power of SCT-based pipelines was not stable across different cell types. In terms of controlling type I errors, all of the denSNE-based pipelines were able to consistently control type I errors across scenarios, while the PCA-based pipelines frequently had large type I errors (Figure 7). In summary, the TP_A_denSNE pipeline had the best performance in differential variability analysis because of its high statistical power and low type I error rate.

After identifying TP_A_denSNE as the best pipeline for differential variability analysis, we further evaluated an extended version of this pipeline on single-cell gene expression data with batch effects. For this evaluation, we designed another simulation study, in which we generated data from two biological conditions, and cells in both conditions came from two batches (Supplementary Methods). Subsequently, we used the extended TP_A_denSNE pipeline to perform differential variability analysis. First, the variability measure was calculated as described in Methods for each cell type, batch and condition. Second, to consider batch effects in the differential analysis, we built a linear regression model for each cell type with the variability measure as the response. The condition and batch labels were treated as the independent variables. Finally, differential variability was determined based on the statistical significance of the condition's regression coefficients (Supplementary Methods). Our results based on 500 independently generated datasets show that the extended TP_A_denSNE pipeline was able to maintain a high statistical power and achieve control of the type I error in the presence of batch effects (Supplementary Figure 3).

The TP_A_denSNE pipeline uses the denSNE method for dimensionality reduction and cells' distances to medoids as the variability measure. In the denSNE publication (Narayan et al.), a measure called the local radius (LR) was proposed, which intuitively represents the average distance of a cell to its nearest neighbors [17]. Therefore, we also evaluated the performance of this variability measure. We refer to this new pipeline as TP_A_LR,

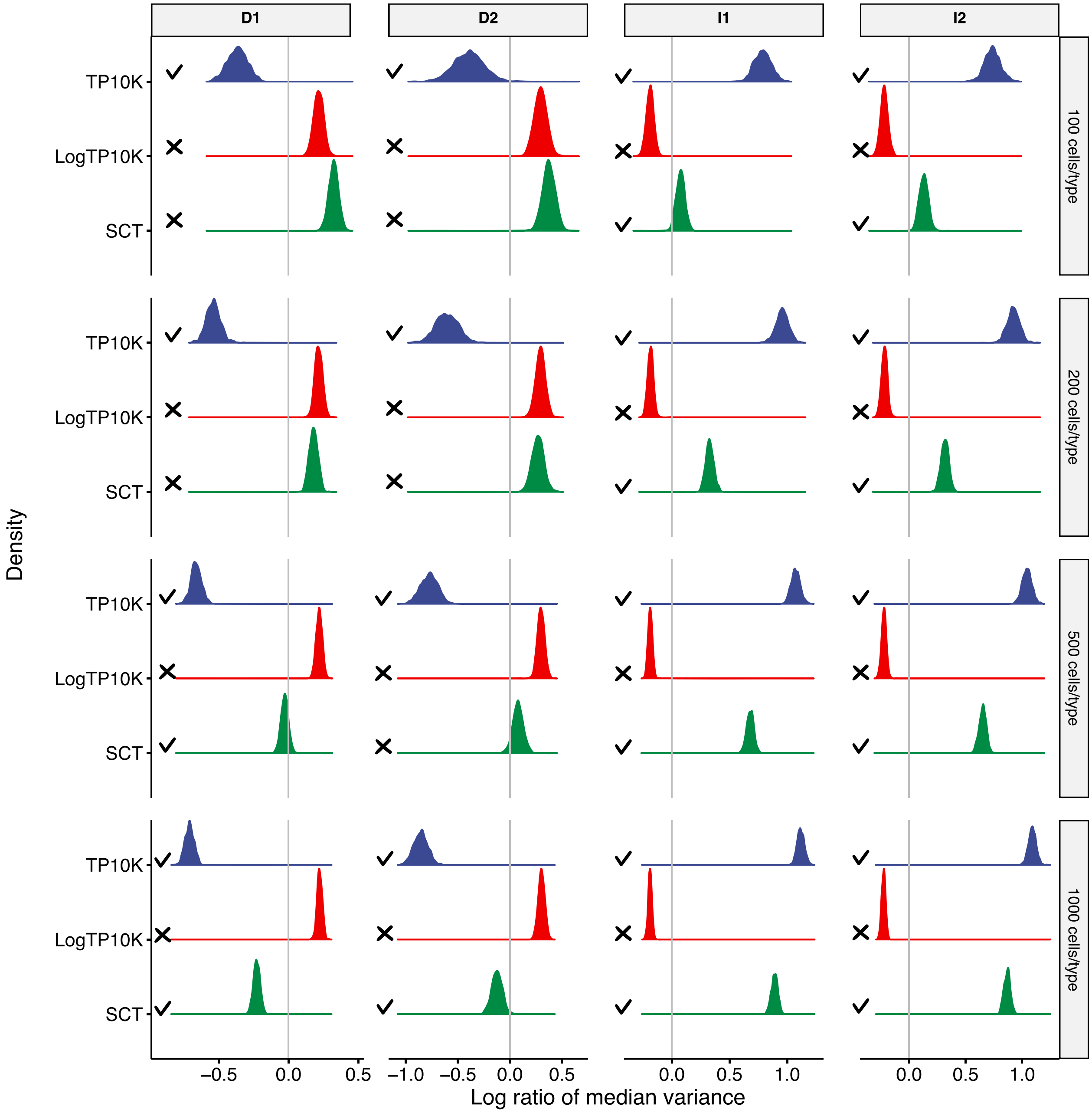

Figure 4. A comparison of gene variances calculated using different normalization methods. As 1000 independent datasets were generated for each combination of cell type and cell number, for each dataset, we calculated the log-ratio of median gene variances (between condition 2 and condition 1) based on the normalized expression values. The density plots illustrate the distribution of the log-ratios across the 1000 datasets. The expected sign of log-ratios is negative for cell types D1 and D2 and positive for I1 and I2. Normalization methods that led to the correct signs are marked with ✓, and those leading to incorrect signs are marked with ✗.

which followed Narayan et al. to calculate the cells' LR and then used the Wilcoxon test to identify differential variability. Our results show that TP_A_denSNE outperformed TP_A_LR in terms of statistical power when cell number within each cell type was at least 200 (Supplementary Figure 4A). Moreover, TP_A_LR failed to control the type I error regardless of cell numbers (Supplementary Figure 4B).

In addition to the differential variability analysis considered by this article, the BASiCS [16] method is able to perform differential variability testing on individual genes. Even though developed for a different purpose, we applied the BASiCS method to our simulated data (with 200 cells per cell type) to evaluate its applicability. For each gene, BASiCS will determine whether its variability is significantly different between the biological conditions. Based on our simulation process, cell types I1 and I2 only had genes with higher variability in C2; cell types D1 and D2 only had genes with higher variability in C1; cell types R1 and R2 did not have genes with differential variability. However, the testing results of BASiCS did not reflect this setting (Supplementary Table 4).

The TP_A_denSNE pipeline identifies changes in immune cell variability in COVID-19 patients

Since the TP_A_denSNE pipeline has demonstrated the best accuracy in the simulation study, we applied it to study the peripheral blood cell populations of COVID-19 patients [26]. In detail, we used the scRNA-seq data of 36 patients, and tested the difference

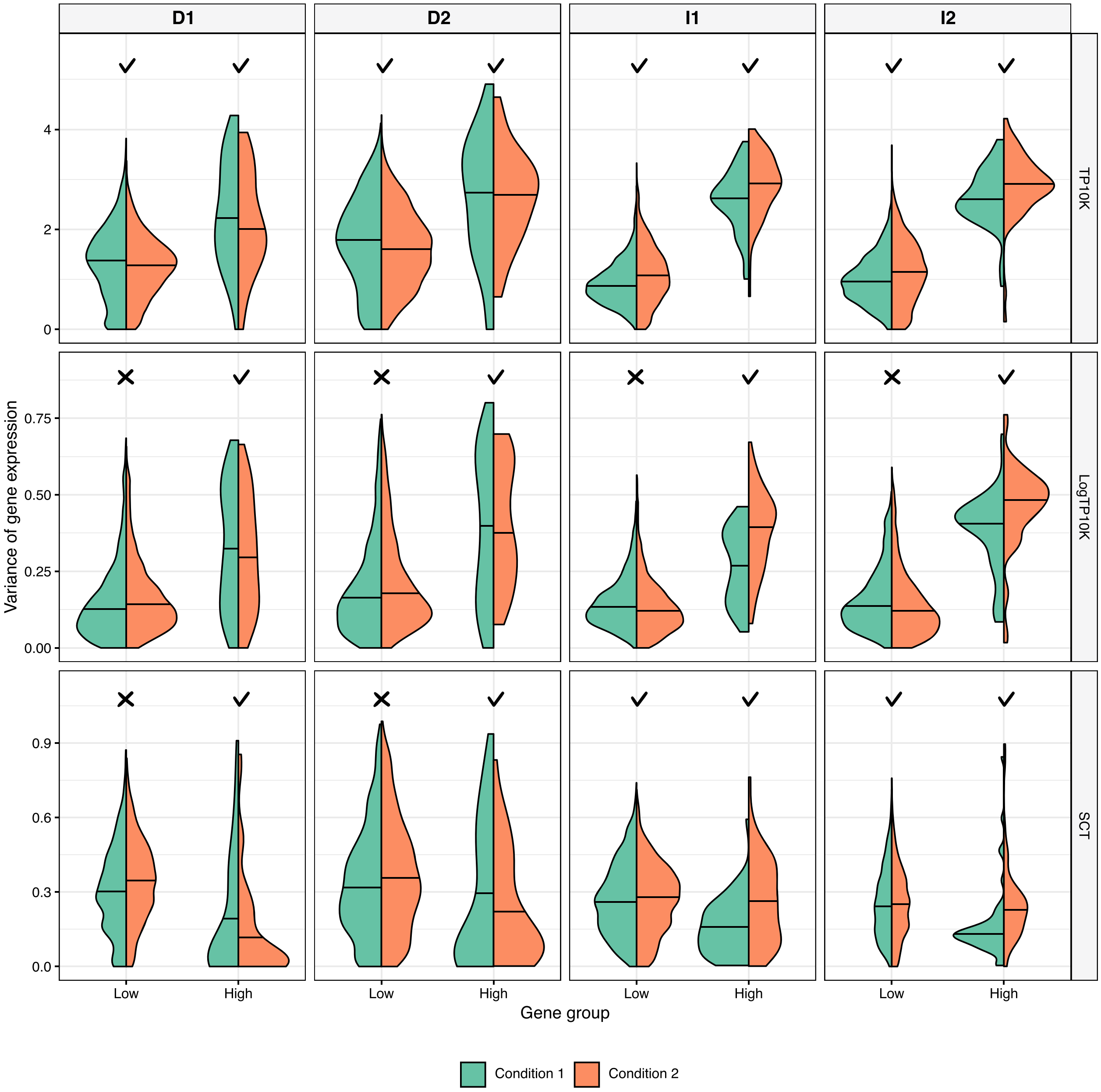

Figure 5. A comparison of calculated gene variances between genes with low and high mean expression levels (based on one simulation dataset). Lowly and highly expressed genes were classified based on their $\log_{10}$-transformed mean UMI counts, using 0.5 as the cutoff. Each column represents one cell type, and each row represents one normalization method. Only genes with different variances between the two conditions are included. The ground truth gene variances in cell types I1 and I2 are higher under condition 2, and the ground truth gene variances in cell types D1 and D2 are higher under condition 1. Normalization methods that lead to the correct orders are marked with ✓, and those leading to incorrect orders are marked with ✗.

in variability of their blood immune cells between two clinical conditions: the mild and severe disease status (see Methods). Our analysis revealed that 14 out of 17 cell types presented significant differential variability in COVID-19 patients with different disease severity, using a threshold of 0.05 and Benjamini–Hochberg correction (Figure 8). Among these, nine cell types have increased variability and five cell types have decreased variability in the severe stage.

Further, we identified top genes that had gene expression variance change in the same direction as the cell type variability change, and performed the pathway enrichment analysis (Methods). The enriched pathways helped explain the biological variation of immune cells between mild and severe disease stages. For example, the B cell population had increased variability in the severe stage and its enriched pathways included the BCR pathway, PDGFRB pathway, CDC42 pathway and MYC pathway, which have important roles in the activation and proliferation of B cells [33–36] (Supplementary Figure 5A). These enriched pathways likely explain the clonal expansion of B cells in severe COVID-19 patients [26]. Besides, CD4$^+$ T cells also had increased variability in the severe stage, and we observed multiple enriched pathways related to T cell activation and development, such as the TCR pathway, HDAC Class I pathway and MYC pathway [37, 38]). Further, we found that the genes that contributed to the decreased variability of CD8$^+$ T cells had enriched pathways responsible for the effector function and exhaustion of CD8 T cells, such

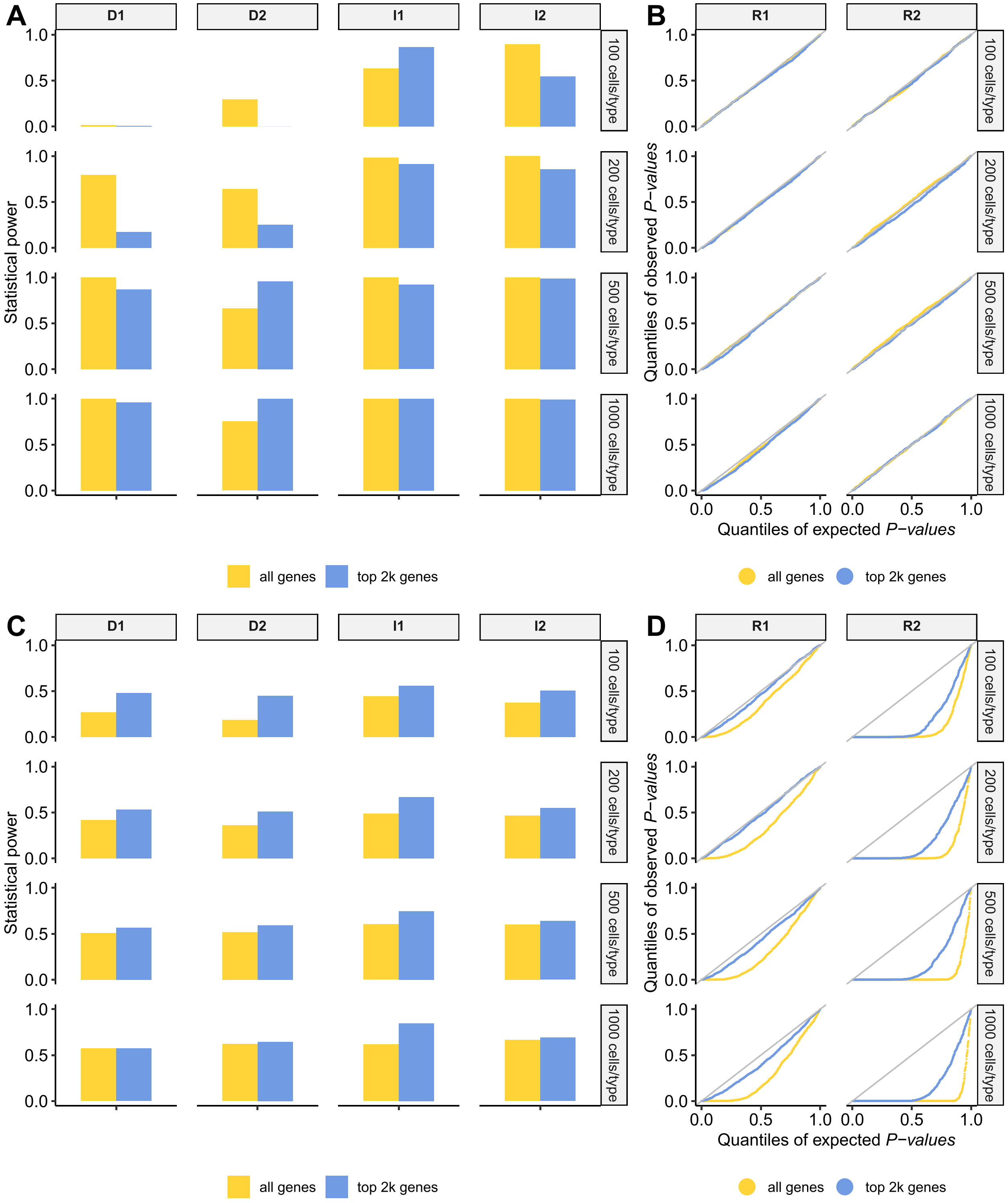

Figure 6. Effect of feature selection in differential variability analysis. (**A**): Statistical power of the TP_A_denSNE and TP_2K_denSNE pipelines. (**B**): Quantile–quantile plots comparing distributions of expected *P*-values and *P*-values from TP_A_denSNE and TP_2K_denSNE pipelines under the null hypothesis. (**C**): Statistical power of the TP_A_PCA and TP_2K_PCA pipelines. (**D**): Quantile–quantile plots comparing distributions of expected *P*-values and *P*-values from TP_A_PCA and TP_2K_PCA pipelines under the null hypothesis.

as the HIF-1 pathway, lysophospholipid-related pathways, NFAT pathway, and AP-1 pathway [39–41] (Supplementary Figure 5B).

Based on the top genes used in the pathway enrichment analysis, we also investigated to what extent these genes demonstrated consistent variance change across different patients. For each cell type that had significantly different variability between mild and severe stages and each top gene, we calculated the genes's variance in each patient and found the median variances in the two stages. If the cell type had increased variability in a condition and a gene's median variance was also higher in the same condition,

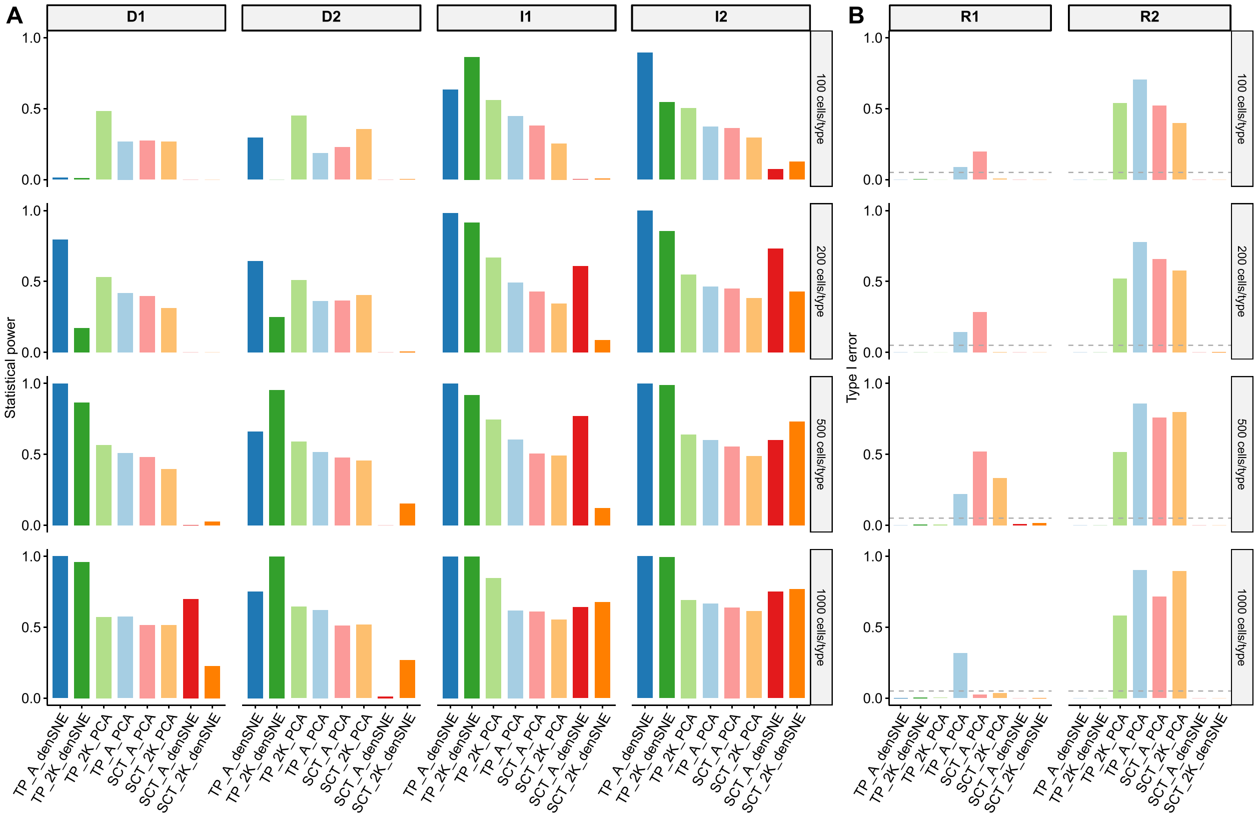

Figure 7. A comparison of the eight TP10K- or SCT- based pipelines in differential variability analysis. (**A**): Statistical power of the eight pipelines. (**B**): Type I errors of the eight pipelines under the null hypothesis. The pipelines are ordered by their average statistical power across all the combinations of cell types and cell numbers.

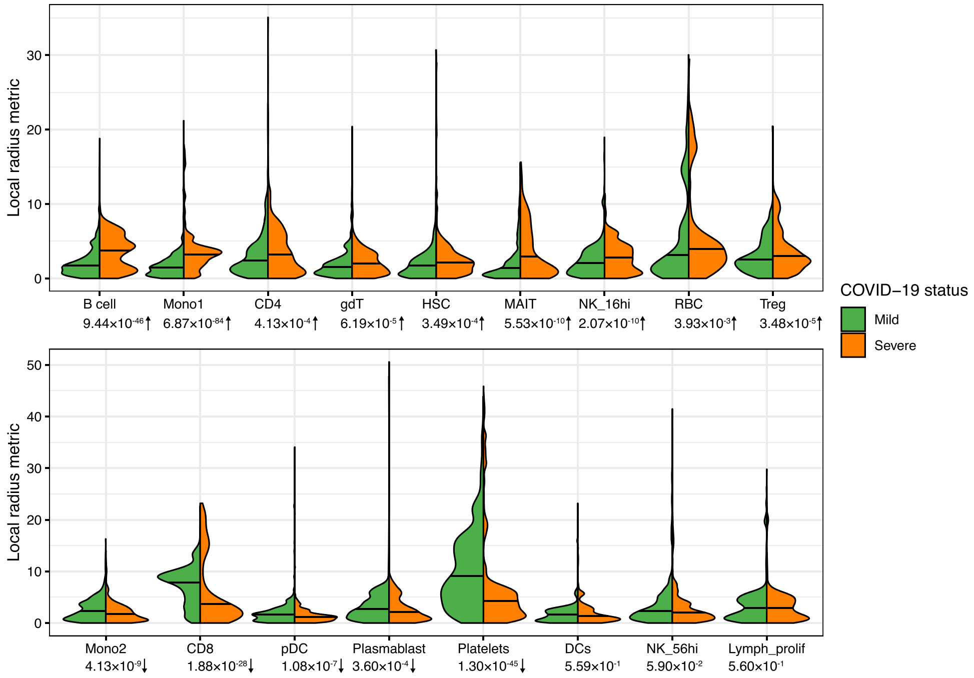

Figure 8. A comparison of immune cell variability in patients at different COVID-19 severity status. The vertical axis represents the variability measure of the immune cell variability calculated by the TP_A_denSNE pipeline. The plots are colored by the severity of COVID-19 symptoms. The adjusted *P*-values from the TP_A_denSNE pipeline are shown below the cell type names. For significant changes (adjusted *P*-value < 0.05), the directions of variability change from the mild to the severe stage are marked as ↑ (increase) or ↓ (decrease). CD4: $CD4^+$ T cells; CD8: $CD8^+$ T cells; DCs: dendritic cells; gdT: $\gamma\delta$ T cells; HSC: hematopoietic stem cells; Lymph_prolif: proliferating lymphocytes; MAIT: mucosal-associated invariant T cells; Mono1: $CD14^+$ monocytes; Mono2: $CD16^+$ monocytes; NK_16hi: $CD16^+$ natural killer cells; NK_56hi: $CD56^+$ natural killer cells; pDC: plasmacytoid dendritic cells; RBC: red blood cells; Treg: regulatory T cells.

this indicated that the majority of subjects had consistent change in this gene. We found that most cell types had over 70% top genes that demonstrated a consistent change (Supplementary Table 5).

This real data application highlights the potential of differential variability analysis to analyze and compare single-cell gene expression data from various disease phases, revealing cellular variability changes in blood immune cells associated with COVID-19 severity.

The TP_A_denSNE pipeline identifies changes in cortical cell type variability in autism patients

We further applied the TP_A_denSNE pipeline to compare the cortical cell types of ASD patients and healthy controls [27]. In detail, we used the scRNA-seq data of 15 ASD patients and 16 healthy controls, and tested the difference in variability of 17 cell types from their cortical tissues (see Methods). Our results revealed that 14 out of 17 cell types presented significant differential variability in ASD patients compared with healthy controls, using a threshold of 0.05 and Benjamini–Hochberg correction (Supplementary Figure 6). Among these, eight cell types have increased variability and six cell types have decreased variability in the ASD patients, compared with the healthy donors.

Next, we also extracted cell-type-specific genes that contributed to the cell type variability change and performed pathway enrichment analysis using these genes (see Methods). In this analysis, we considered pathways in the Pathway Interaction Database and the Human Phenotype Ontology [28, 30]. Among cell types with significantly increased variability in ASD patients, the ERBB1 and PDGFRB pathways were enriched in six and four cell types, respectively. They have important functions in nerve cell development and repair, and were reported to be associated with ASD symptom severity [42, 43] (Supplementary Figure 7A). In addition, we found that several phenotype ontology terms, such as ‘Hyperactivity’, ‘Intellectual disability, severe’ and ‘Abnormal emotion/affect behavior’, were enriched in multiple cell types with increased variability (Supplementary Figure 8A). As for cell types with significantly decreased variability in ASD patients, the ERBB1 and PDGFRB pathways were most enriched in the endothelial cells. Besides, in parvalbumin interneurons (IN-PV), the CDC42 and mTOR signaling pathways were enriched, both of which are indicative of neurodevelopmental abnormalities in ASD patients [44, 45] (Supplementary Figure 7B). The phenotype ontology ‘Abnormality of the cerebral cortex’ was also enriched in genes with decreased variance in IN-PV of ASD patients (Supplementary Figure 8B).

On this dataset, we also investigated to what extent the top genes demonstrated consistent variance change across different patients, using the same approach as on the COVID-19 data. Our results show that all cell types had over 80% top genes that demonstrated a consistent change (Supplementary Table 6).

CONCLUSIONS

Quantifying and comparing single-cell heterogeneity are pivotal to understanding cell fate decisions and patterns of distinct cellular behaviors. Although there have been some computational efforts to detect and quantify gene-level variability from scRNA-seq data, methods for differential variability analysis between cell populations are not yet available.

Instead of studying the variability of individual genes, our work focuses on quantifying and comparing the cellular variability of cell populations between different biological conditions. To this end, we formalized and evaluated 12 pipelines for differential variability analysis of scRNA-seq data, which accounted for different combinations of methods for normalization, feature selection and dimensionality reduction.

Evaluation based on high-fidelity synthetic data with ground truth suggests that the best pipeline performs library size normalization without logarithm transformation, retains all genes in analysis and uses the denSNE-based distances to population medoids as the variability measure.

In summary, our work complements existing research focused on gene-level variability changes or visualization methods. We anticipate that our findings, together with the proposed pipelines, will offer a new perspective to compare single-cell populations from different cell types or biological conditions based on the overall transcriptional variability. This approach will contribute to a deeper understanding of cellular heterogeneity in various biological and biomedical contexts.

Key Points

- We evaluated 12 pipelines for the differential variability analysis of single-cell RNA-seq data. The pipelines are composed of various combinations of methods for normalization, feature selection, dimensionality reduction and variability calculation.
- To benchmark the 12 pipelines, we generated high-fidelity synthetic data with ground truth variability alterations, and compared the statistical power and false discovery control of the pipelines in diverse scenarios.
- By evaluating the effectiveness of individual steps within the pipelines, we found that the TP10K-based normalization method is more accurate than LogTP10K and SCT in preserving gene variability.
- Utilizing density-based tSNE embeddings to define cellular distances improves differential variability analysis compared with scores of PCs.
- Among the 12 pipelines, the TP_A_denSNE pipeline is the most powerful in identifying changes in transcriptional variability in scRNA-seq data. Applying this pipeline to real data uncovered biological variability distinctions between COVID-19 patients with varying degrees of disease severity, and between autism patients and healthy controls.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

ACKNOWLEDGMENTS

We would like to thank Dr. Ching-On Wong at Rutgers, The State University of New Jersey and members of the Vivian Li Lab for their helpful comments and suggestions. We would also like to acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here.

FUNDING

This work was supported by the National Institute of General Medical Sciences [R35GM142702] (to W.V.L.) and the National Institute of Mental Health [R00MH117393] (to A.K.)

AUTHORS' CONTRIBUTIONS

W.V.L. conceived the study. J.L. and W.V.L. designed and performed the data analysis. J.L., W.V.L. and A.K. wrote and reviewed the manuscript.

AVAILABILITY OF DATA AND MATERIALS

The source code for data simulation and differential variability analysis is available at our GitHub repository (https://github.com/jiayiliujiayi/scRNA_Seq-Differential_Variability_Analysis). The data of immune cells from lung cancer are available at GEO with accession number GSE127465 [25]. The COVID-19 data are available from https://covid19cellatlas.org/ [26]. The data of ASD patients and healthy controls are downloaded from https://autism.cells.ucsc.edu/ [27].

REFERENCES

1. Altschuler SJ, Lani FW. Cellular heterogeneity: when do differences make a difference? *Cell* 2010;**141**(4):559–63.
2. Lelièvre SA, Hodges KB, Vidi P-A. Chapter 26 - Application of Theranostics to Measure and Treat Cell Heterogeneity in Cancer. In: Chen X, Wong S (eds). *Cancer Theranostics*. Oxford: Academic Press, 2014, 493–516.
3. Regev A, Teichmann SA, Lander ES, *et al.* The human cell atlas. *Elife* 2017;**6**:e27041.
4. Rozenblatt-Rosen O, Regev A, Oberdoerffer P, *et al.* The human tumor atlas network: charting tumor transitions across space and time at single-cell resolution. *Cell* 2020;**181**(2):236–49.
5. Qian K, Shiwei F, Li H, Li WV. scINSIGHT for interpreting single-cell gene expression from biologically heterogeneous data. *Genome Biol* 2022;**23**(1):82.
6. Osorio D, Xue Y, Zhong Y, *et al.* Single-cell expression variability implies cell function. *Cell* 2019;**9**(1):14.
7. Li WV. Phitest for analyzing the homogeneity of single-cell populations. *Bioinformatics* 2022;**38**(9):2639–41.
8. Zhang X, Chenling X, Yosef N. Simulating multiple faceted variability in single cell rna sequencing. *Nat Commun* 2019;**10**(1): 1–16.
9. Wang Y, Song W, Wang J, *et al.* Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. *J Exp Med* 2019;**217**(2):e20191130.
10. Joseph Burclaff R, Bliton J, Breau KA, *et al.* A proximal-to-distal survey of healthy adult human small intestine and colon epithelium by single-cell Transcriptomics. *Cellular and molecular*. *Gastroenterol Hepatol* 2022;**13**(5):1554–89.
11. Richard A, Boullu L, Herbach U, *et al.* Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. *PLoS Biol* 2016;**14**(12):e1002585.
12. Mojtahedi M, Skupin A, Zhou J, *et al.* Cell fate decision as high-dimensional critical state transition. *PLoS Biol* 2016;**14**(12):e2000640.
13. Martinez-Jimenez CP, Eling N, Chen H-C, *et al.* Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* 2017;**355**(6332):1433–6.
14. Bahar R, Hartmann CH, Rodriguez KA, *et al.* Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature* 2006;**441**(7096):1011–4.
15. Ho JWK, Stefani M, dos Remedios CG, Charleston MA. Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics* 2008;**24**(13):i390–8.
16. Eling N, Richard AC, Richardson S, *et al.* Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. *Cell Syst* 2018;**7**(3):284–294.e12.
17. Narayan A, Berger B, Cho H. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nat Biotechnol* 2021;**39**(6):765–74.
18. Lytal N, Ran D, An L. Normalization methods on single-cell RNA-seq data: An empirical survey. *Front Genet* 2020;**11**:41.
19. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* 2019;**20**(1):296.
20. Brennecke P, Anders S, Kim JK, *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 2013;**10**(11):1093–5.
21. Sheng J, Li WV. Selecting gene features for unsupervised analysis of single-cell gene expression data. *Brief Bioinform* 2021;**22**(6):bbab295.
22. Stuart T, Butler A, Hoffman P, *et al.* Comprehensive integration of single-cell data. *Cell* 2019;**177**(7):1888–1902.e21.
23. Sun T, Song D, Li WV, Li JJ. scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biol* 2021;**22**(1):163.
24. Li WV, Li JJ. A statistical simulator scDesign for rational scRNA-seq experimental design. *Bioinformatics* 2019;**35**(14):i41–50.
25. Zilionis R, Engblom C, Pfirschke C, *et al.* Single-cell Transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species. *Immunity* 2019;**50**(5):1317–1334.e10.
26. Stephenson E, Reynolds G, Botting RA, *et al.* Single-cell multi-omics analysis of the immune response in covid-19. *Nat Med* 2021;**27**(5):904–16.
27. Velmeshev D, Schirmer L, Jung D, *et al.* Single-cell genomics identifies cell type–specific molecular changes in autism. *Science* 2019;**364**(6441):685–9.
28. Schaefer CF, Anthony K, Krupa S, *et al.* Pid: the pathway interaction database. *Nucleic Acids Res* 2009;**37**(suppl_1):D674–9.
29. Liberzon A, Birger C, Thorvaldsdóttir H, *et al.* The molecular signatures database hallmark gene set collection. *Cell Syst* 2015;**1**(6):417–25.
30. Köhler S, Gargano M, Matentzoglu N, *et al.* The human phenotype ontology in 2021. *Nucleic Acids Res* 2021;**49**(D1):D1207–17.
31. Tianzhi W, Erqiang H, Shuangbin X, *et al.* Clusterprofiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation* 2021;**2**(3):100141.
32. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**(10):R106.
33. Liu W, Tolar P, Song W, Kim TJ. Bcr signaling and b cell activation. *Frontiers in immunology* 2020;**11**:45.
34. Trink B, Wang G, Shahar M, *et al.* Functional platelet-derived growth factor-beta (pdgf-β) receptor expressed on early b-lineage precursor cells. *Clin Exp Immunol* 1995;**102**(2):417–24.
35. Burbage M, Keppler SJ, Gasparrini F, *et al.* Cdc42 is a key regulator of b cell differentiation and is required for antiviral humoral immunity. *J Exp Med* 2015;**212**(1):53–72.
36. Habib T, Park H, Tsang M, *et al.* Myc stimulates b lymphocyte differentiation and amplifies calcium signaling. *J Cell Biol* 2007;**179**(4):717–31.

37. Akimova T, Beier UH, Liu Y, *et al.* Histone/protein deacetylases and t-cell immune responses. *Blood* 2012;**119**(11):2443–51.
38. Marchingo JM, Sinclair LV, Howden AJM, Cantrell DA. Quantitative analysis of how myc controls t cell proteomes and metabolic pathways during t cell activation. *Elife* 2020;**9**:e53725.
39. Doedens AL, Phan AT, Stradner MH, *et al.* Hypoxia-inducible factors enhance the effector responses of cd8+ t cells to persistent antigen. *Nat Immunol* 2013;**14**(11):1173–82.
40. Oda SK, Strauch P, Fujiwara Y, *et al.* Lysophosphatidic acid inhibits cd8 t-cell activation and control of tumor progressionlpa inhibits cd8 t-cell activation and tumor control. *Cancer Immunol Res* 2013;**1**(4):245–55.
41. Seo W, Jerin C, Nishikawa H. Transcriptional regulatory network for the establishment of cd8+ t cell exhaustion. *Exp Mol Med* 2021;**53**(2):202–9.
42. Russo AJ. Increased epidermal growth factor receptor (EGFR) associated with hepatocyte growth factor (HGF) and symptom severity in children with autism spectrum disorders (ASDs). *J Cent Nerv Syst Dis* 2014;**6**:JCNSD–S13767.
43. Zakareia FA, Al-Ayadhi LY, Al-Drees AMA. Study of dual angiogenic/neurogenic growth factors among saudi autistic children and their correlation with the severity of this disorder. *Neurosci J* 2012;**17**(3):213–8.
44. Martinelli S, Krumbach OHF, Pantaleoni F, *et al.* Functional dysregulation of cdc42 causes diverse developmental phenotypes. *Am J Hum Genet* 2018;**102**(2):309–20.
45. Wang B, Qin Y, Qunyan W, *et al.* Mtor signaling pathway regulates the release of proinflammatory molecule ccl5 implicated in the pathogenesis of autism spectrum disorder. *Front Immunol* 2022;**13**:818518.