SIEVE: identifying robust single cell variable genes for single-cell RNA sequencing data

Yinan Zhang[a,b], Xiaowei Xie[a,b,c], Peng Wu[a,b,c,*], Ping Zhu[a,b,c,*]

[a]State Key Laboratory of Experimental Hematology, Institute of Hematology and Blood Diseases Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Tianjin 300020, China; [b]Department of Stem Cell & Regenerative Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Tianjin 300020, China; [c]National Clinical Research Center for Blood Diseases, Center for Stem Cell Medicine, Chinese Academy of Medical Sciences, Tianjin 300020, China

Abstract

Single-cell RNA-seq data analysis generally requires quality control, normalization, highly variable genes screening, dimensionality reduction and clustering. Among these processes, downstream analysis including dimensionality reduction and clustering are sensitive to the selection of highly variable genes. Though increasing number of tools for selecting the highly variable genes have been developed, an evaluation of their performances and a general strategy are lack. Here, we compare the performance of nine commonly used methods for screening variable genes by using single-cell RNA-seq data from hematopoietic stem/progenitor cells and mature blood cells, and find that SCHS outperforms other methods regarding to reproducibility and accuracy. However, this method prefers the selection of highly expressed genes. We further propose a new strategy SIEVE (SIngle-cEll Variable gEnes) by multiple rounds of random sampling, therefore minimizing the stochastic noise and identifying a robust set of variable genes. Moreover, SIEVE recovers lowly expressed genes as variable genes and substantially improves the accuracy of single cell classification, especially for the methods with lower reproducibility. The SIEVE software is freely available at https://github.com/YinanZhang522/SIEVE.

Keywords: Hematopoietic stem/progenitor cells, Highly variable genes, Single-cell RNA-seq

1. INTRODUCTION

In the past decades, next-generation sequencing technologies have provided unprecedented opportunities to explore single cell heterogeneity at genomics, transcriptomics and epigenomics levels.[1–4] Single-cell RNA sequencing (scRNA-seq) enables us to investigate the transcriptomes of thousands of single cells in complex multicellular organisms with lower costs.[5,6] How to process the scRNA-seq data to reflect the biological significance has become a big challenge. In general, scRNA-seq data needs to be filtered out of low-quality data and normalized to eliminate background noise. Then, highly variable genes (HVGs) are selected and fed to dimensionality reduction and followed by clustering analysis.[7,8] Notably, advanced downstream analyses depend on the accurate determination of HVGs.

Currently, multiple tools have been developed to identify HVGs. Representative methods include the M3Drop,[9,10] Scmap,[11] Scran,[12] Vst, SCTransform (SCT) and Disp in Seurat,[13,14] ROGUE, ROGUE using normalized value (ROGUE_n)[15] and singleCellHaystack (SCHS).[16] The M3Drop method identifies genes with higher variability than expected by analyzing the relationship between the square of the variation coefficient and the average expression. Scmap first checks the dropout rate of the single-cell matrix and then selects genes by the relationship between the average expression and dropout rate. Scran obtains HVGs through the variance of the logarithmic expression value and the mean expression value. Vst, SCT and Disp are all sourced from Seurat package. Vst uses local polynomial regression to fit the logarithmic variance to logarithmic mean value and calculates the variance based on the normalized values; SCT applies the regularized negative binomial regression to calculate the technical noise model and selects genes by Pearson residuals; Disp directly selects genes with the highest discrete values. A new method ROGUE recently published selects meaningful genes by fitting the relationship between the expression entropy values and the average expression levels. It can handle both the original counts and the normalized data, and we use ROGUE and ROGUE_n to represent them respectively. SCHS is initially designed to detect differentially expressed genes and identifies genes by the spatial distribution of cells.

Here, we compare the performance of nine methods for identifying HVGs on scRNA-seq data of hematopoietic stem/progenitor cells (HSPCs) and mature blood cells separately, in terms of purity, reproducibility and accuracy. We further propose a new strategy SIEVE (SIngle-cEll Variable gEnes) by multiple rounds of random sampling to identify a robust set of variable

* *Address correspondence: Peng Wu, PhD, and Ping Zhu, PhD, State Key Laboratory of Experimental Hematology, Institute of Hematology and Blood Diseases Hospital, No. 288 Nanjing Road, Heping District, Tianjin 300020, China. E-mail address: wupeng1@ihcams.ac.cn (P. Wu); zhuping@ihcams.ac.cn (P. Zhu)*

This work has been supported by the National Natural Science Foundation of China (82022002, 81900117, 81890993, 81890990, 32000803), National Key Research and Development Program of China (2018YFA0107804), CAMS Initiative for Innovative Medicine (2017-I2M-1-015, 2017-I2M-3-009, 2019-I2M-2-001) and Fundamental Research Funds for the Central Research Institutes (2020-RC310-005).

The authors declare no conflicts of interest.

Blood Science, (2021) 3, 35-39

Received February 5, 2021; Accepted March 30, 2021.

http://dx.doi.org/10.1097/BS9.0000000000000072

genes. We show that SIEVE substantially improves the accuracy of single cell classification.

2. RESULTS

2.1. Overall experimental design

In order to evaluate the performance of different methods for identifying HVGs, a general pipeline of scRNA-seq data analysis, including data preprocessing, HVGs selection, dimensionality reduction and clustering, was carried out on single cell transcriptomes of hematopoietic cells we recently published[17] (Fig. 1A). Only in the HVGs selection step, nine different methods were applied, while the other steps followed the default settings in Seurat. The dataset covered a total of 7,551 single cells with 32 immunophenotypic cell types from fluorescent-activated cell sorting. The evaluation indicators used for the comparison of different methods were the purity of cell clustering, reproducibility of HVGs, and accuracy of cell classification.

2.2. Characterization of HVGs by different methods

Most HVGs selection methods depend on the calculation of the relationship between the mean and variance of gene expression. We analyzed the distribution of 2,000 HVGs obtained by each method in hematopoietic cells and found that the average expression levels of these gene sets varied among these methods (Fig. 1B and Supplementary Fig. S1A, http://links.lww.com/BS/A27). About a quarter of genes reported by four methods, including M3Drop, Seurat_vst, Seurat_sct and Seurat_disp, expressed at a low level. In contrast, ROGUE, ROGUE_n and Scran methods chose almost no lowly expressed genes. However, over 85% of Scmap- and SCHS-derived genes were highly expressed genes (Supplementary Fig. S1B, http://links.lww.com/BS/A27). In addition, the distribution of HVGs obtained by different methods in mature blood cells was consistent with that of stem/progenitor cells (Supplementary Fig. S1C, http://links.lww.com/BS/A27).

These distinct gene sets would inevitably lead to different results for downstream advanced analyses. When using HVGs selected by M3Drop to cluster HSPCs, clusters were separated from each other and there was more than one immunophenotypic cell type in each cluster. This might be relevant to the similarity of the transcriptome among HSPCs, and we cannot see the transcriptional continuum during hematopoietic stem cell (HSC) differentiation from the clusters resulted from M3Drop. While other methods had a relatively good distinguishing ability, even for HSPCs with very similar expression profiles. HSC was located at the apex of hematopoietic differentiation and produced complete progenitor cells of erythroid, myeloid and lymphoid lines, reflecting that the hematopoietic process was a continuous progress (Supplementary Fig. 1D, http://links.lww.com/BS/A27). While each type of mature blood cell had its own relatively specific expression profile, a complete separation of erythrocyte cells, T cells, B cells and granulocytes was revealed by various methods (Supplementary Fig. 1E, http://links.lww.com/BS/A27). We used the Calinski-Harabasz index and Davies-Bouldin index[18] to evaluate the different methods, where a higher Calinski-Harabasz index and a lower Davies-Bouldin index relate to a model with better separation between the clusters. M3Drop have the lowest Calinski-Harabasz index and the highest Davies-Bouldin index, indicating a relatively lower distinguishing capability. (Supplementary Fig. 1F, G, http://links.lww.com/BS/A27). Conclusively, single cells containing very different transcriptome spectrum could be clustered better no matter which HVG selection method was applied.

2.3. Robustness of HVGs and downstream effects

To compare the effects of HVGs on downstream analyses, we first calculated the purity of each cluster. The pure cluster here

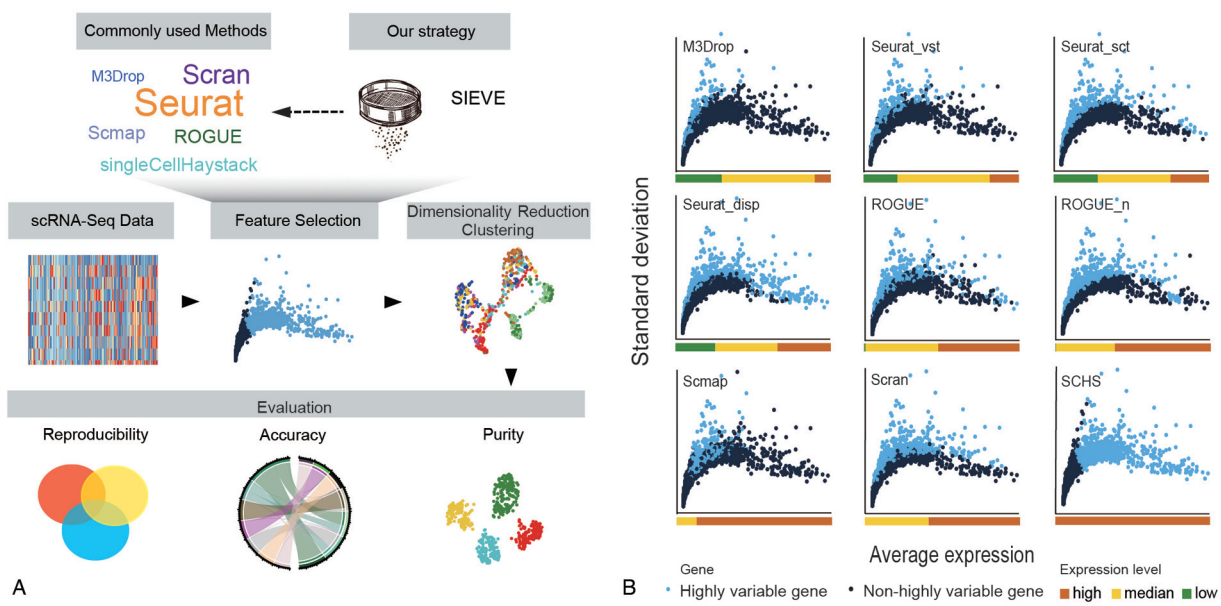

Figure 1. Characterization of highly variable genes by different methods in the HSPC dataset. (A) Workflow for analysis of single-cell RNA-seq data and evaluation of HVG selection methods. A new strategy SIEVE (SIngle-cEll Variable gEnes) is proposed to improve the HVG selection. (B) HVGs by different methods are viewed in the relationship between standard deviation and average expression of genes. Each point represents one gene, light blue points mean HVGs and dark blue points mean other genes. The bar plot below each scatter plot presents the gene expression levels for HVGs. High, higher than the 3rd quartile (orange); median, between the 1st quartile and the 3rd quartile (yellow); low, lower than the 1st quartile (green).

was defined as a cluster that had the identical function and state without variable genes. We observed that the purity of cell clusters generated by all methods was relatively high, maintaining above 90% in both HSPCs and mature blood cells (Fig. 2**A**). ROGUE, Scmap and Scran were slightly inferior to the other methods.

Next, we randomly sampled 70% of the total number of cells 50 times. Cells extracted each time were used as the reference set, and the remaining 30% cells were used as the query set. Applying each method on the reference set resulted in 2,000 HVGs each test. We first examined the reproducibility of different methods in identifying HVGs, the proportion of overlapped genes that reported by every two tests was used to indicate the reproducibility. We found that SCHS outperformed other methods in both HSPCs and mature blood cells (Fig. 2**B**). Scran, Scmap and ROGUE had a relatively lower performance, with a stable reproducibility of 80% to 90%, while others showed the lowest reproducibility ranging from 50% to 70%. We also tested the effect of different sampling sizes on the reproducibility and found that the reproducibility was increased with the sampling size of reference set. However, when the sample size is the same, the differential performances of distinct methods were still consistent. (Supplementary Fig. 2A, B, http://links.lww.com/BS/A28).

Cell classification is a routine procedure to annotate cell types in newly generated datasets for integrative analyses. Therefore, we assessed the accuracy of cell classification that confounded by HVGs. During each round of random sampling, the HVGs were selected for the reference set by different methods. The random forest classifier was trained using the reference set and the corresponding selected genes, and then used to predict cell types in the remaining query set. Compared to the cell labels predefined using the total cell set, the classification accuracy could be calculated (Supplementary Fig. S2C, D, http://links.lww.com/BS/A28). In HSPCs, the classification accuracy of majority methods ranged from 85 to 90% (Fig. 2**C** left). Due to the high gene expression specificity of mature blood cells, the classification accuracy reached over 90% except for the M3Drop method (Fig. 2**C** right). We further showed that the classification accuracy of these methods except M3Drop was not greatly affected by the number of HVGs (Fig. 2**D**). Collectively, the selection of HVGs affected single cell classification to various degrees in these methods.

2.4. Optimized strategy for HVGs screening

The low reproducibility of HVGs means that the corresponding gene selection method is unstable, affecting the purity of cell clustering and accuracy of cell classification. To further improve the performance of existing HVGs selection methods, we developed a strategy named SIEVE (SIngle-cEll Variable gEnes) (Fig. 3**A**). Based on random sampling for all single cells in a scRNA-seq dataset, SIEVE provided the reproducibility estimation for HVGs selection method and screened robust HVGs for the following analysis. The average expression levels of HVGs after using SIEVE increased as compared to that before using SIEVE (Fig. 3**B**, Supplementary Fig. S3C, http://links.lww.com/BS/A29). The methods, such as M3Drop, Seurat, ROGUE and Scran, chose more median expression level genes than Scmap and SCHS, illustrating that HVGs selected by them had a wider dynamic range of expression. In addition, we clustered cells using different HVGs and calculated the cluster markers. For majority methods, the HVGs obtained by using SIEVE were more enriched in cell cluster markers than those without SIEVE (Fig. 3**C**), demonstrating that SIEVE selected more biologically relevant genes.

Furthermore, SIEVE significantly enhanced the classification accuracy, in spite of the number of random sampling (Fig. 3**D**). SIEVE geared the commonly used methods to select the most

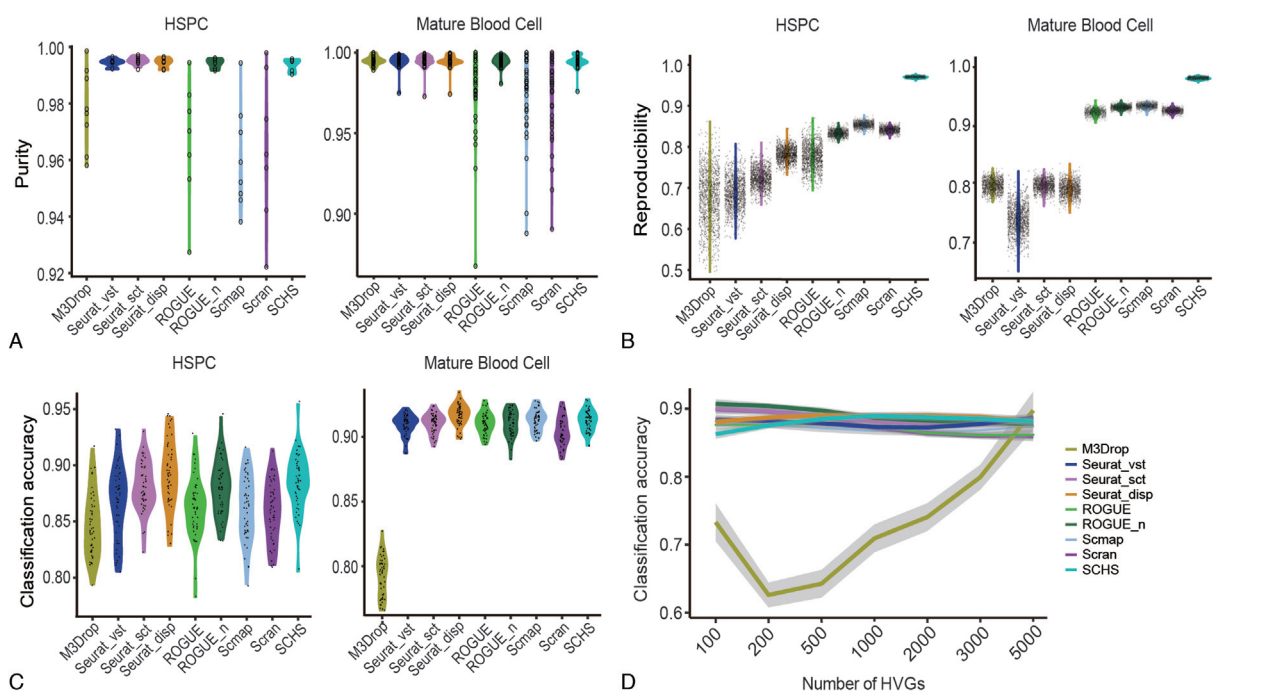

Figure 2. Evaluation of different HVG selecting methods. (A) The purity of clusters defined by genes selected by different methods. Each point represents a value. (B) Reproducibility of HVGs selected by different methods. (C) Cell classification accuracy of different methods. (D) The relationship between the number of HVGs and classification accuracy. A smooth line was fitted for each method. The middle line shows the average value and the grey region indicates interquartile range.

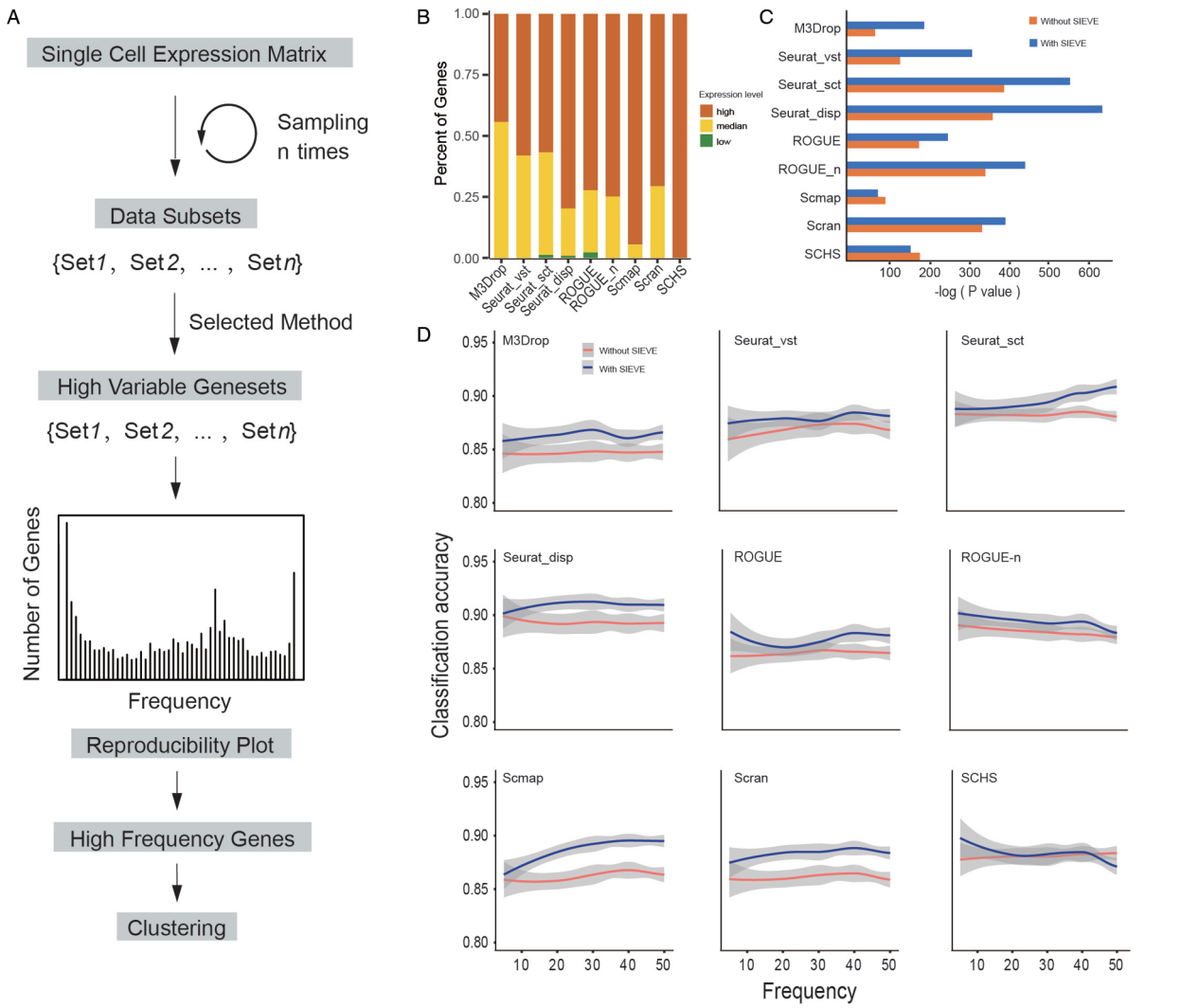

Figure 3. SIEVE improves the single-cell clustering. (A) Workflow for SIEVE. Based on random sampling for a scRNA-seq dataset, SIEVE provides the reproducibility estimation for HVG selection method and screens high frequency variable genes as the last HVGs used to the following dimensionality reduction. (B) The gene expression levels for HVGs after using SIEVE. High, higher than the 3rd quartile (orange); median, between the 1st quartile and the 3rd quartile (yellow); low, lower than the 1st quartile (green). (C) Comparation of the enrichment significances of HVGs on cluster markers between using SIEVE and without SIEVE. (D) The relationship between the number of repetitions and classification accuracy. The middle line shows mean and the grey region indicates interquartile range.

representative HVGs and effectively improve the cell clustering visualization, especially for those with lower reproducibility (Supplementary Fig. S3A, B, http://links.lww.com/BS/A29). After applying SIEVE, the cell clustering of M3Drop and Seurat_disp were substantially advanced, the hematopoietic differentiation of erythroid, myeloid lines became clear.

3. DISCUSSION

The performance of HVGs selection methods varies for different cell type data. For mature blood cells, these methods show highly consistent performances on purity of cell clustering and accuracy of cell type classification. By contrast, they perform differently in stem/progenitor cells. Probably, these differences are relevant to the transcriptional continuum during HSC differentiation. Hence, we adopt the strategy through multiple rounds of random sampling and sieve genes that are highly variable in single cells. These HVGs are more likely to represent cell heterogeneity and relevant to biological processes.

As single-cell data sets may be different in cell types or data volumes, our proposed SIEVE is more flexible and generalized to commonly used methods during HVGs screening. For instance, when dealing with a data set from a large number of single cells, we can extract a smaller proportion of the data subset to shorten the calculation time and ensure the selection of robust HVGs. Our strategy is expected to be widely applied in a large amount of data and more gene selection methods.

4. MATERIALS AND METHODS

4.1. HVG selection methods

The M3Drop method used the BrenneckeGetVariableGenes function. The Seurat_sct method was implemented with the SCTransform function in Seurat package. The Seurat_vst and Seurat_disp method were implemented with the FindVariable-Features function in Seurat package. The ROGUE and ROGUE_n method were implemented with the SE_fun function in ROGUE package. The feature selection step in Scmap was implemented

according to the selectFeatures function in scmap package. The step of the Scran method was copied from the source code of the pipeline of Scran. In addition, we implemented the SCHS method with haystack function in singleCellHaystack package.

4.2. Single-cell dimensionality reduction and cell clustering

The read count of a given gene was quantified by the total number of distinct UMIs, and the raw UMIs of protein-coding genes were normalized by $\log_2$ (TPM/10 + 1) (TPM: transcripts per million) for downstream analysis. The mean of the normalized data of one gene across all single cells was calculated as the average expression level, which was further divided into three levels: high (higher than the 3rd quartile), median (between the 1st quartile and the 3rd quartile) and low (lower than the 1st quartile). Single-cell dimensionality reduction and cell clustering were analyzed using different HVG sets by Seurat. We performed RunUMAP (dims = 1:8) and FindClusters (resolution = 0.6) to cluster single cells.

4.3. Comparison of purity, reproducibility and classification accuracy

The purity index of cell cluster is calculated with the rogue function in ROGUE package. When the purity index is 1, it means there's no differentially expressed genes in the cluster. Gene expression matrix and cell labels were input with the parameter of that "sample" of all cells was set to "1".[15]

To illustrate the performance of different methods, we adapt the random sampling approach. Firstly, we randomly selected 70% of the total cells as the reference set and identified HVGs based on the reference set with different HVG selection methods respectively. Then, we further trained the random forest classifier using the reference set with only informative genes selected by different methods and predicted the remaining 30% cells as query set with the trained classifier. Finally, we repeated this entire procedure for n = 50 times for each method. The reproducibility score is the overlapped ratio between the different genes selected during every two repetitions. The classification accuracy was quantified with the classification accuracy score, which is the similarity between the predicted cell types and the original cell types of the query set.

We calculated the reproducibility by intersecting the corresponding sets of variable genes as

$$\text{Reproducibility} = \frac{Geneset_1 \cap Geneset_2}{n}$$

where n is the number of HVGs.

We calculated the reproducibility by the similarity between the predicted cell types and the original cell types of the query set.

$$\text{Classification accuracy} = \frac{n_r}{N}$$

where n_r is the number of rightly-predicted cells, N is the number of cells of the query set.

4.4. Evaluation of SIEVE

Single-cell dimensionality reduction and cell clustering using HVGs obtained by SIEVE were performed by Seurat with the same parameters as described above. Cluster markers were calculated using the FindAllMarkers function of Seurat and defined with the FDR < 0.05 and |avg_logFC| > 1. The enrichment analysis of HVGs on cell cluster markers was performed by Fisher's exact test. To confirm the effects of gene sets before and after using SIEVE on the dimensionality reduction and cell clustering, we calculated classification accuracy by choosing genes with different times of random sampling.

4.5. Sequencing data acquisition

The single-cell RNA sequencing data used in this study were downloaded from the Gene Expression Omnibus database with the accession numbers GSE137864 and GSE149938.[17]

ACKNOWLEDGMENTS

This work has been supported by the National Natural Science Foundation of China (82022002, 81900117, 81890993, 81890990, 32000803), National Key Research and Development Program of China (2018YFA0107804), CAMS Initiative for Innovative Medicine (2017-I2M-1-015, 2017-I2M-3-009, 2019-I2M-2-001) and Fundamental Research Funds for the Central Research Institutes (2020-RC310-005).

REFERENCES

[1] Chappell L, Russell AJC, Voet T. Single-cell (multi)omics technologies. *Annu Rev Genomics Hum Genet* 2018;19:15–41. doi: 10.1146/annurev-genom-091416-035324.
[2] Song Y, Xu X, Wang W, Tian T, Zhu Z, Yang C. Single cell transcriptomics: moving towards multi-omics. *Analyst* 2019;144 (10):3172–3189. doi: 10.1039/c8an01852a.
[3] Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;6(5):377–382. doi: 10.1038/nmeth.1315.
[4] Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* 2013;23 (12):2126–2135. doi: 10.1101/gr.161679.113.
[5] Picelli S. Single-cell RNA-sequencing: the future of genome biology is now. *RNA Biol* 2017;14(5):637–650. doi: 10.1080/15476286.2016.1201618.
[6] Wu AR, Wang J, Streets AM, Huang Y. Single-cell transcriptional analysis. *Ann Rev Anal Chem* 2017;10(1):439–462. doi: 10.1146/annurev-anchem-061516-045228.
[7] Choi YH, Kim JK. Dissecting cellular heterogeneity using single-cell RNA sequencing. *Mol Cells* 2019;42:189–199. doi: 10.14348/molcells.2019.2446.
[8] Haque A, Engel J, Teichmann SA, Lonnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* 2017;9:75. doi: 10.1186/s13073-017-0467-4.
[9] Brennecke P, Anders S, Kim JK, et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 2013;10(11):1093–1095. doi: 10.1038/nmeth.2645.
[10] Andrews TS, Hemberg M. M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics* 2019;35(16):2865–2867. doi: 10.1093/bioinformatics/bty1044.
[11] Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods* 2018;15(5):359–362. doi: 10.1038/nmeth.4644.
[12] Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. *F1000Research* 2016;5:2122. doi: 10.12688/f1000research.9501.2.
[13] 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. doi: 10.1186/s13059-019-1874-1.
[14] Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. *Cell* 2019;177(7):1888–1902.e21. doi: 10.1016/j.cell.2019.05.031.
[15] Liu B, Li C, Li Z, Wang D, Ren X, Zhang Z. An entropy-based metric for assessing the purity of single cell populations. *Na Commun* 2020;11 (1):3155. doi: 10.1038/s41467-020-16904-3.
[16] Vandenbon A, Diez D. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nat Commun* 2020;11(1):4318. doi: 10.1038/s41467-020-17900-3.
[17] Xie X, Liu M, Zhang Y, et al. Single-cell transcriptomic landscape of human blood cells. *Natl Sci Rev* 2021;8(3):nwaa180. https://doi.org/10.1093/nsr/nwaa180.
[18] Pedregosa , et al. Scikit-learn: machine learning in python. *JMLR* 2011;12:2825–2830. doi: 10.5555/1953048.2078195.