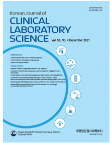
Korean Society for Clinical Laboratory Science

https://doi.org/10.15324/kjcls.2024.56.1.10

Korean Journal of
CLINICAL LABORATORY SCIENCE



REVIEW ARTICLE

The Workflow for Computational Analysis of Single-cell RNA-sequencing Data

Sung-Hun WOO[1], Byung Chul JUNG[2]

[1]Department of Biomedical Laboratory Science, College of Software and Digital Healthcare Convergence, Yonsei University, Wonju, Korea
[2]Department of Nutritional Sciences and Toxicology, University of California, Berkeley, California, USA

단일 세포 RNA 시퀀싱 데이터에 대한 컴퓨터 분석의 작업과정

우성훈[1], 정병출[2]

[1]연세대학교 소프트웨어디지털헬스케어융합대학 임상병리학과, [2]캘리포니아대학교 버클리캠퍼스 영양과학 및 독성학과

ARTICLE INFO

Received January 1, 2024
Revised 1st January 29, 2024
Revised 2nd February 1, 2024
Accepted February 2, 2024

Key words
Big data
Computational biology
Single-cell gene expression analysis

ABSTRACT

RNA-sequencing (RNA-seq) is a technique used for providing global patterns of transcriptomes in samples. However, it can only provide the average gene expression across cells and does not address the heterogeneity within the samples. The advances in single-cell RNA sequencing (scRNA-seq) technology have revolutionized our understanding of heterogeneity and the dynamics of gene expression at the single-cell level. For example, scRNA-seq allows us to identify the cell types in complex tissues, which can provide information regarding the alteration of the cell population by perturbations, such as genetic modification. Since its initial introduction, scRNA-seq has rapidly become popular, leading to the development of a huge number of bioinformatic tools. However, the analysis of the big dataset generated from scRNA-seq requires a general understanding of the preprocessing of the dataset and a variety of analytical techniques. Here, we present an overview of the workflow involved in analyzing the scRNA-seq dataset. First, we describe the preprocessing of the dataset, including quality control, normalization, and dimensionality reduction. Then, we introduce the downstream analysis provided with the most commonly used computational packages. This review aims to provide a workflow guideline for new researchers interested in this field.

INTRODUCTION

Genomic transcription is tightly regulated, and assessing its expression can be a key to understanding pathological conditions. Currently, bulk RNA-sequencing (RNA-seq) is the most commonly used technique for analyzing transcriptomic profiling in a variety of fields, including neuroscience, diabetes, and oncology [1-4]. Although bulk RNA-seq is a powerful tool for estimating global transcriptomic profiling in the target samples, it can only provide an average of gene expression across entire populations within the samples. This becomes particularly problematic when dealing with heterogeneous samples such as blood and biopsies, as the uniqueness of each cell within the whole population is masked [5]. To overcome this issue, a technique for RNA-seq at the single-cell level, known as single-cell RNA-Seq (scRNA-seq), has been invented [6]. Since its first introduction, scRNA-seq has rapidly gained attention due to its ability to facilitate the deconvolu-

Corresponding author: Byung Chul JUNG
Department of Nutritional Sciences and Toxicology, University of California, Berkeley, California 94720, USA
E-mail: sandbag9@berkeley.edu
ORCID: https://orcid.org/0000-0003-0732-0122

pISSN 1738-3544 eISSN 2288-1662

tion of cell types in heterogeneous samples. In particular, scRNA-seq allows us to identify rare populations that were unable to be addressed using traditional bulk RNA-seq [7]. Moreover, scRNA-seq can be used to trace the trajectories of distinct cell lineage in tissue development, cancer, and immune cells [8]. Currently, the most widely used and most common commercially available platform is a droplet-based microfluidics system by 10x Genomics [9]. This platform captures single cells into each droplet containing beads conjugated to the primers with both common and unique barcodes and enzymes for library preparation. Each droplet acts as an individual reaction chamber where cell lysis and library generation take place. Since each generated library has unique barcodes, each cell can be distinguished and analyzed at a single-cell level. However, the data generated from scRNA-seq is much more complex than bulk RNA-seq and there is not a universal standardization to analyze the dataset. For analyzing scRNA-seq, we need to use a bunch of bioinformatic tools (Table 1). In this article, we aim to provide general considerations when dealing with analyzing the scRNA-seq datasets. In addition, we focused on introducing relatively new and most widely used tools for scRNA-seq rather than the explaining complex algorithm and mathematics underlying each tool. For new researchers interested in this field, this paper can serve as a reference to build a workflow for scRNA-seq analysis that can be modified to their own research context.

Tabel 1. The descriptions of the popular methods for scRNA-seq pipelines

Task	Tool (language)	Year	Description
General scRNA-seq	Seurat (R)	2015~2023	A popular R package for the preprocessing and explorative downstream analysis of single-cell RNA sequencing data. Commonly, it is used with a variety of R-based package
General scRNA-seq	Scanpy (Python)	2018	A popular Python package for the preprocessing and explorative downstream analysis of single-cell RNA sequencing data. Commonly, it is used with a variety of Python-based package
Empty-drop identification	EmptyDrops (R)	2019	It estimates the background levels of RNA present in empty droplets, then identifies droplets containing cell that significantly deviate from the background
Ambient RNA identification	DecontX (Python)	2020	It utilizes Bayesian method to estimate the percentage of contaminating transcripts from ambient RNA, then removing contaminated transcripts in each cell data
Doublet identification	DoubletFinder (R) Scrublet (Python)	2019 2019	Generate artificial doublets using a nearest-neighbor algorithm, then identifies the doublets that are similar to artificial doublets
Normalization	SCtransform (R)	2019	It utilizes regularized negative binomial regression, which represent normalized data value without affected by technical issues
Visualization	t-SNE (R, Python) UMAP (R, Python)	2008 2018	Both are unsupervised non-linear dimensionality reduction method for visualization. UMAP has been rapidly overtaking t-SNE due to its superior ability to preserve large-scale structures
Differential expression testing	ROTSvoom (R) D3E (Python) Limma-trend (R) Wilcoxon rank-sum (R, Python)	2020	They are famous packages for differential expression testing, which show good performance after prefiltering lowly expressed genes. Wilcoxon rank-sum test is most widely used option
Pseudotime	Monocle3 (R) scTEP (R)	2019 2023	Monocle3 is the most popular package for pseudotime while scTEP is the most recently developed package, which may show better accuracy

Abbreviations: scRNA-seq, single-cell RNA-sequencing; t-SNE, t-distributed stochastic neighbor embedding; UMAP, uniform manifold approximation and projection; scTEP, single-cell data trajectory inference method using ensemble pseudotime.

MAIN ISSUE

1. Preprocessing of Data from Single-cell RNA-sequencing

Unlike traditional bulk RNA-seq, droplet-based scRNA-seq can generate a variety of artifacts such as empty droplets and ambient RNA. In addition, we need to remove the data generated from damaged cells and non-single cells (i.e. duplet or multiplets) by computational tools. In the first section, we will discuss how to remove the data from low-quality cells. Then, we will discuss data normalization and visualization in low-dimensional space.

1) Quality Control of the Data

The first step of scRNA-seq data analysis is to ensure that each transcriptomic data corresponds to intact live cells. With a droplet-based microfluidics system, it is unavoidable to have empty drops as cells in the target sample are highly diluted to achieve a single cell in each droplet. In addition, ambient RNA released from damaged and dead cells in the microfluidics system can be encapsulated in empty drops. As a result, ambient RNA also can be amplified and have its own barcode, which can be incorrectly considered as data from real cells. Therefore, it is necessary to filter out the data obtained from ambient RNA and should not be included for further analysis. Initially, this was performed by removing all barcodes with lower transcripts [10]. Although this method is simple and straightforward, it can wrongly filter out the droplets containing small cells with low RNA amounts. Recently, several computational tools were developed including EmptyDrops to address this issue [11]. This method first estimates expression profiles of ambient RNA in empty droplets and then identifies cell-containing droplets that significantly differ from ambient RNA. Moreover, Yang et al [12] developed computational tools called DecontX, which are designed to remove transcripts from ambient RNA. DecontX used a Bayesian method to estimate the percentage of contaminating transcripts from ambient RNA. After estimation, DecontX can remove contaminated transcripts in each cell. With these approaches, researchers will have better cell type recovery as it does not filter out the data obtained from cells with low RNA content. Next, the data generated by the damaged cell should be filtered out. Since mitochondrial RNA (mtRNA) is more likely to be retained in damaged cells due to the mitochondrial membranes, a higher proportion of mtRNA is a widely used criterion to determine damaged cells [13]. The cutoff value for the acceptable proportion of mtRNA is varied depending on the tissue types, technical factors, etc [14]. In general, 5%~10% of the threshold can be a starting point to adjust the cut-off value to distinguish normal and damaged cells [14, 15]. Another important aspect to take into account is the removal of multiplet artifacts, which represent two or more cells in a single droplet, from the dataset for further analysis. Typically, the frequency of multiplets is positively correlated with the concentration of the input cell. Although diluted cell suspension can reduce the frequency of multiplets, it is not feasible to dilute too much as it reduces cell recovery. In this regard, several computational tools such as DoubletFinder and Scrublet were proposed [16, 17]. The algorithms of the two tools are similar. The tools generate artificial doublets by combining two randomly selected droplets of gene expression data. Subsequently, doublet scores in real scRNA-seq data are calculated based on the similarity of artificial droplets using k-nearest neighbors algorithm [18]. One of the main differences between these tools is the hyperparameter setting such as the number of artificial droplets and the number of principal components to determine nearest neighbors [18]. Although these two tools have their own strength, the recent benchmarking result indicates DoubletFinder method shows the best doublet detection accuracy among the 9 cutting-edge computational tools [18]. An overview of the computational methods to isolate intact live cells is illustrated in Figure 1.

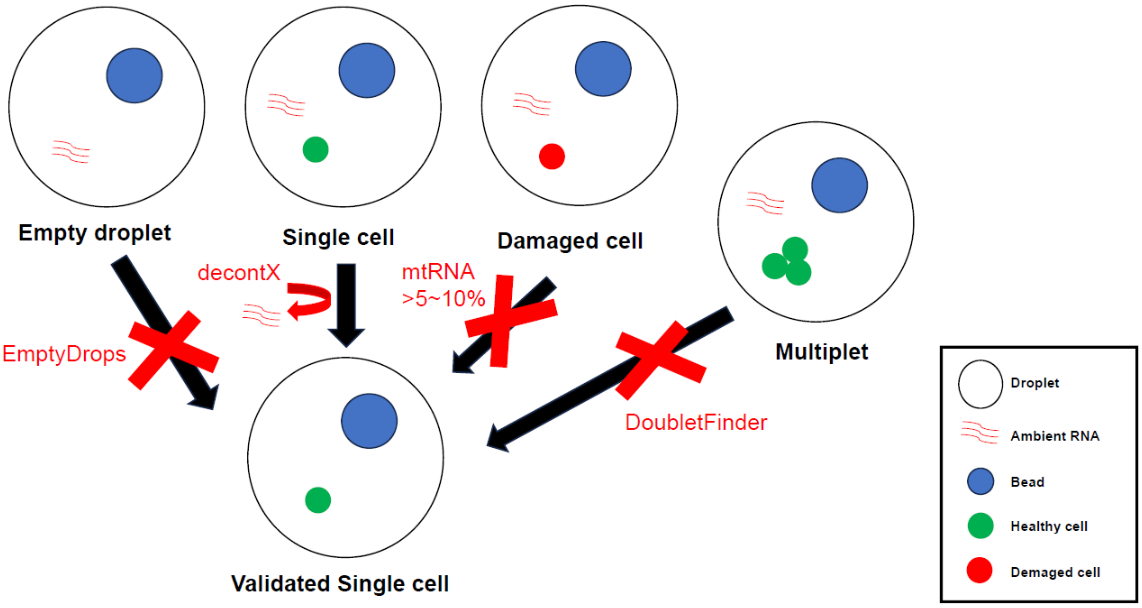


Figure 1. Isolation of dataset derived from single cell by computational methods.
Abbreviation: mtRNA, mitochondrial RNA.

2) Normalization of the Data

After cell-level quality control, data normalization is a prerequisite to correct for cell-to-cell differences due to technical variability such as capture efficiency, amplification biases, and sequencing depth (number of transcripts detected per cell) [19]. If data is not normalized properly, downstream analysis such as comparison of gene expression and clustering of subpopulations would be biased. One of the common normalized methods is library size (total counts) normalization [20]. This process involves dividing the read count of each cell by a size factor to normalize the library size across all cells. For example, let us assume that we have scRNA-seq data derived from mouse tissue. Then, we can suppose that cell A has 10 reads from the *Pparg* gene with a total of 1×10^6 reads and cell B has 20 reads from the *Pparg* gene with a total of 2×10^6 reads. Since cell B has a 2-fold library size compared to cell A, the actual expression level of *Pparg* is the same after normalization. While this approach is simple, it is based on the assumption that variations in library size across cells are due to technical issues, not genuine biological heterogeneity. However, it is possible that certain cell types have more transcripts than others [21]. Moreover, library size normalization can be skewed by highly expressed genes, which have more read counts and therefore contribute more to the normalization process [22]. To overcome this issue, several new normalization methods designed specifically for single-cell studies have been published. SCtransform is one of the most widely used normalization tools for scRNA-seq [23, 24]. It calculates Pearson residuals from a regularized negative binomial regression, which can eliminate variations derived from technical sources such as sequencing depth, but conserve biological variations [23]. The data normalized by SCtransform showed more accurate downstream analysis, compared to the data normalized by other size-factor-based methods such as Scran and standard log-normalization [23]. SCtransform is a freely available R package and has been integrated into the single-cell toolkit, Seurat.

3) Feature Selection and Dimensionality Reduction

In the data generated from scRNA-seq, the expression value of each gene represents a dimension. While high-dimensional data is more informative compared to traditional approaches that rely on low-throughput techniques, interpreting high-dimensional data can be much more difficult. Feature selection is the computational technique, which selects genes that efficiently describe the original dataset. In other words, it excludes genes that act as noise and redundancy. For example, if two genes (features) are perfectly correlated, selecting one feature can be enough to describe the original dataset and the other feature is no longer informative but can serve as a noise. Therefore, feature selection (excluding the irrelevant features) can help in not only reducing computational burden but also providing a better understanding of the data to improve downstream analysis [25]. To exclude irrelevant features, it is necessary to estimate the relevance of each feature to the dataset. A commonly used approach is selecting highly variable genes (HVGs) because it does not require any prior knowledge such as predefined cell marker genes [26]. Several tools, like Seurat, can calculate the mean expression and dispersion of genes to identify HVGs [27]. Typically, genes with the highest variance-to-mean ratio are selected first as HVGs, with the number ranging from 1,000 to 5,000 depending on the complexity of the dataset [28].

After selecting HVGs, the dimensions of the dataset can be further reduced by several dimensionality reduction methods for visualization in two- or three-dimensional space. For the data generated from scRNA-seq, t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) are the most widely used dimensionality reduction techniques for two dimensions visualization (Figure 2) [29]. t-SNE is an unsupervised non-linear dimensionality reduction method that maps high-dimensional data points to two dimensions while preserving the local structures [30]. The algorithm of t-SNE is creating a probability distribution for the high-dimensional data that represents data (each cell) similarities a high-dimensional space. For example, similar cells are assigned a higher probability that the cell would choose another similar cell as its neighbor in high dimensions. Then, t-SNE defines a probability distribution that represents data similarities in the lower dimensions (usually two dimension) and try to minimize the Kullback–Leibler divergence (KL divergence) between the two distributions [31]. By minimizing KL divergence, the data in low dimensional space becomes similar to the original structure of the data (the data in high dimensional space) [32]. However, the disadvantages of t-SNE algorithms are (1) inaccuracies of the global structure, (2) slow computation time due to the complex calculations, and (3) computationally unfeasibility with large datasets, such as those with over 10,000 cells [33]. In this regard, UMAP has been introduced and has become preferable over t-SNE [34]. The underlying algorithm of UMAP is constructing a fuzzy topological structure that represents the likelihood of a connection between the cells in high dimensional space. Then it optimizes the low dimensional representation to make its fuzzy topological structure as similar as possible to the original dataset [35]. Recently, UMAP has been rapidly overtaking t-SNE due to its superior ability to preserve large-scale structures and its better scaling performance, which results in faster processing time [36]. However, a recent paper argues that an advanced t-SNE algorithm can perform at a similar speed and can preserve the global structure in a manner similar to UMAP [37]. Therefore, it is still controversial which algorithm is more suitable for visualizing the data from scRNA-seq.

2. Downstream Analysis of Single-cell RNA-sequencing

After applying appropriate preprocessing of the scRNA-seq dataset, the processed dataset can provide a vast amount of information including inference of the cell types and the ordering of cells along a lineage based

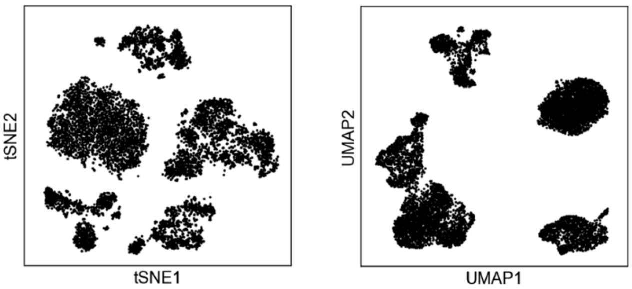


Figure 2. t-SNE and UMAP plots, generated from the artificial dataset. Abbreviations: t-SNE, t-distributed stochastic neighbor embedding; UMAP, uniform manifold approximation and projection.

on the gene expression at the single cell level. In the following section, we provide information regarding the most popular and widely used downstream analyses.

1) Unsupervised Clustering and Cell Type Annotation

A common analysis step after dimensionality reduction is unsupervised clustering to identify the groups of cells with similar characteristics based on the expression profiles. Although there are several tools were developed, Leiden clustering is recently developed and shows better performance, compared to classical methods such as Louvain clustering [38]. The Leiden algorithm is based on modularity and hierarchical clustering. Modularity is a measure of the strength of communities and shows dense connections between the nodes when communities have high modularity [39]. Hierarchical clustering is an unsupervised machine learning algorithm that pairs objects based on the similarity between the dataset [40]. After running the Leiden algorithm, the clustering can be visualized with t-SNE or UMAP. When running Leiden clustering for your dataset, it is important to note that resolution parameters used for computing the modularity can specify the number of communities (clusters). Figure 3, which was implemented by Scanpy (python based), shows the Leiden clustering visualized in t-SNE and UMAP representation at two different resolution settings [41]. The resolution parameters are generally optimized with cell type annotations by checking gene expression profiles in each cluster and the researcher needs to decide the degree of annotation detail. For example, a satisfactory cell label could be “natural killer (NK) cells,” but it can be further divided into “NK cells” and “activated NK cells” [42]. To identify cell types, researchers need to explore the scRNA dataset, which shows transcriptomic profiles of clusters, and manually inspect whether established marker genes that have been reported for a given cell type are expressed in certain clusters. For example, CD31$^-$, CD45$^-$, Ter119$^-$ and PDGFRα$^+$ can be the cell maker for adipose precursor cells in stromal vascular fraction [2].

2) Differential Expression Testing

Differential expression testing can provide biological insights into differences between two experimental conditions, and it is very well-documented in bulk RNA-seq [1]. Briefly, differential expression testing can be conducted using three generational statistical tests: (1) over-representation analysis, (2) functional class scoring, and (3) topology-based pathway analysis. Through differential expression testing, researchers can identify important biological pathways from the

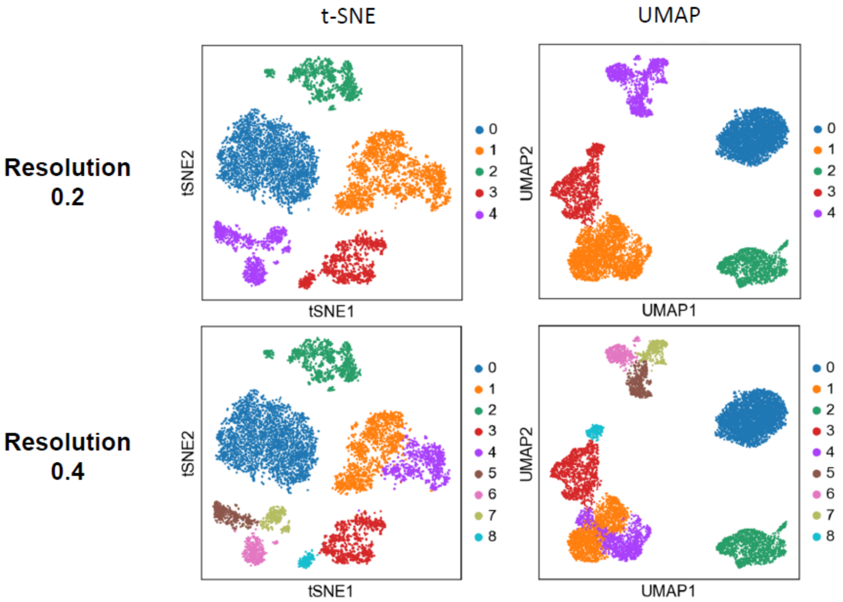


Figure 3. Leiden clustering visualized in t-SNE and UMAP plots, generated from the artificial dataset.
Abbreviations: t-SNE, t-distributed stochastic neighbor embedding; UMAP, uniform manifold approximation and projection.

gene list of differentially expressed genes or the total gene list assigned with gene scores (ranking). With the dataset from scRNA-seq, it is particularly useful to dissect cell-type-specific responses to perturbations such as chemical treatment, disease, or genetical modification. For example, imagine a scenario where we have identified immune cells in the whole blood from wild-type and genetically altered mice through scRNA-seq analysis. Then, we can select certain cell types, such as monocyte only, from the dataset and identify relevant biological pathways from the two genotypes of mice. This allows us to determine which cell types are most perturbed or mostly involved in phenotypic change by genetic alteration. However, differential expression testing for scRNA-seq is much more challenging, compared to bulk RNA-seq because scRNA-seq data typically has lower library sizes, high noise levels, and "dropout" events, where certain genes are detected in small populations but not detected in other populations (i.e. lots of the zero value of certain genes in certain populations) [43]. Moreover, given the clustering is already defined based on the gene profiling before differential expression testing in scRNA-seq analysis, clustering and differential expression testing are not independent which results in artificial false discoveries [44]. In this regard, Soneson and Robinson [45] compared several methods to find differential expression testing in the scRNA-seq dataset. They suggest that prefiltering of lowly expressed genes can be beneficial to type I error control and ROTSvoom, D3E, limma-trend, the t-test, and the Wilcoxon test performed well in terms of lower false discovery proportions. Indeed, the Wilcoxon rank-sum test is the most widely used statistical method for differential expression testing employed in recent scRNA-seq analysis [44]. Wilcoxon rank-sum test is the default option for differential expression testing in the single-cell toolkit, Seurat.

3) Pseudotime

One of the most revolutionary uses of scRNA-seq is to assess cellular state transitions, which are characterized by cascades of gene expression changes [7]. For example, cascades of gene expression changes are well documented during the cell cycle and cell differentiation [46, 47]. Pseudotime analysis can assign each cell to a specific position based on its gene expression patterns, which provides an ordering of cells along a trajectory or lineage [48]. More than 70 pseudotime analysis tools have already been developed, but the most famous and widely used method is an R package, Monocle [49, 50]. Monocle3, the most recent version of Monocle, uses a principal graph algorithm and calculates the geodesic distance of cells from the user-selected root node in the trajectory as a hypothesized time course, pseudotime [51]. Recently, the single-cell data trajectory inference method using ensemble pseudotime inference (scTEP) is proposed by Zhang et al [52]. It takes advantage of the multiple clustering results and fine-tuning algorithm for improving pseudotime accuracy. They compared the scTEP with other pseudotime analysis methods using 41 real scRNA-seq data sets and showed better robustness and accuracy.

3. Clinical Applications of Single-cell RNA-sequencing

Since the introduction of scRNA-seq, numerous studies have been conducted mostly by specialized research groups with expertise in single-cell isolation and computational analysis. Recently, scRNA-seq has become more accessible to the broader research community including biomedical researchers and clinical researchers [8]. Especially, the field of oncology has been actively studied using scRNA-seq [53]. Since one of the major causative factors for cancer is genomic disruption, cancer cells likely have distinct transcriptome profiling from their normal counterparts [54]. Therefore, cancer cells would be shown in a distinct cluster whereas normal cells would be located in broader clusters in the UMAP of scRNA-seq [55]. Then, we could further analyze the cancer-specific cluster if there are certain mutations converting normal cells to cancer cells. In Figure 4, we can consider that cluster 8 is a cancer-specific cluster. If we use bulk RNA-seq, we may not

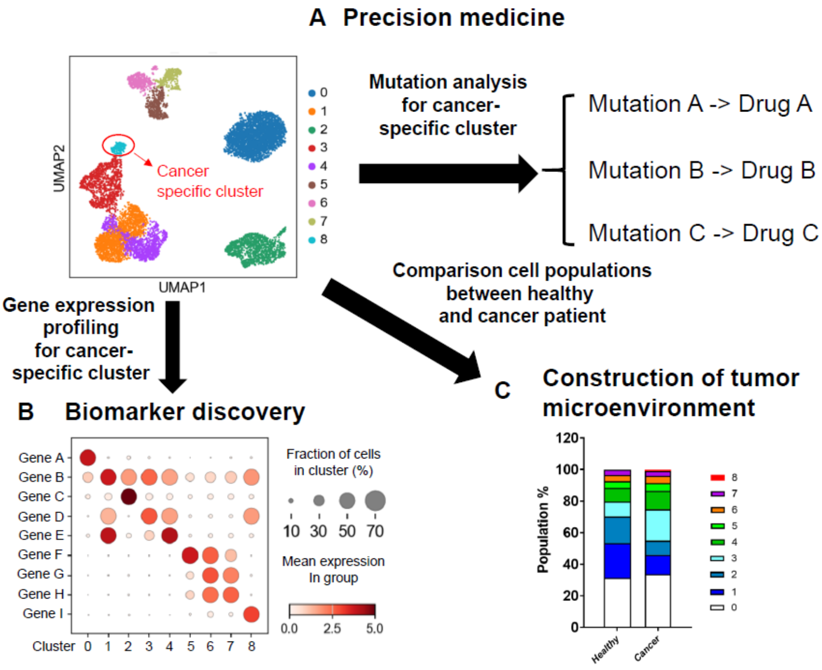


Figure 4. Clinical applications of scRNA-seq.
Abbreviations: scRNA-seq, single-cell RNA-sequencing; UMAP, uniform manifold approximation and projection.

find the mutation because cancer-specific cluster 8 is a relatively small population, which could be masked from other major populations. With scRNA-seq, we can specifically characterize the cancer cell and apply appropriate therapeutic options (Figure 4A). For example, mutations in epidermal growth factor receptor (EGFR) are almost exclusively found in lung adenocarcinomas and the drug for each mutation varies (e.g. EGFR exon 19 deletions and exon 21 L858R alterations-erlotinib, gefitinib, and afatinib; EGFR exon 20 insertions-osimertinib and poziotinib) [56, 57]. In addition, we may find specific biomarkers after differential expression testing. In dot plots, a gene I exclusively expressed in the cancer-specific cluster, which could be considered as a potential biomarker gene (Figure 4B). Furthermore, the construction of tumor microenvironment is available through scRNA-seq. Since tumor tissue has different microenvironments consisting of heterogeneous cell types such as tissue-resident cells, endothelial cells, and tumor-infiltrating immune cells, the population ratio of each cell type can be valuable information to understand characterization of the cancers [58]. In Figure 4C, we can compare cell populations between healthy and cancer patients to construct the tumor microenvironment. Among the 9 clusters, clusters 3 and 4 are increased whereas cluster 1 is decreased in cancer patients (Figure 4C). Therefore, comparison between these clusters can be interesting in understanding the mechanisms of disease and finding new therapeutic targets.

CONCLUSION

scRNA-seq technology has opened the avenue to investigate cellular heterogeneity of RNA transcripts at the single-cell level. The field of scRNA-seq analysis is rapidly expanding, with advanced platforms and computational tools emerging regularly. In this review, we provide an overview of analyzing the scRNA-seq datasets with the introduction of the most commonly used R packages and recently developed tools for preprocessing of the data and downstream analysis. However, researchers need to be aware that none of the analytical tools is flawless and the tools of scRNA-seq analysis are rapidly emerging to overcome current limitations. Therefore, researchers should expect there will be other improvements to the packages shortly although the computational tools described here remain valid over time.

요 약

RNA-시퀀싱은 표본에 대한 전사체 전체의 패턴을 제공하는 기법이다. 그러나 RNA-시퀀싱은 표본 내 전체 세포에 대한 평균 유전자 발현만 제공할 수 있으며, 표본 내의 이질성(heterogeneity)에 대한 정보는 제공하지 못한다. 단일 세포 RNA-시퀀싱 기술의 발전을 통해 우리는 표본의 단일 세포 수준에서 이질성과 유전자 발현의 동역학(dynamics)에 대한 이해를 할 수 있게 되었다. 예를 들어, 우리는 단일 세포 RNA-시퀀싱을 통해 복잡한 조직을 구성하는 다양한 세포 유형을 식별할 수 있으며, 특정 세포 유형의 유전자 발현 변화와 같은 정보를 알 수 있다. 단일 세포 RNA-시퀀싱은 처음 도입된 이후 많은 이들의 관심을 끌게 되었으며, 이를 활용하기 위한 대규모 생물정보학(bioinformatics) 도구가 개발되었다. 그러나 단일 세포 RNA-시퀀싱에서 생성된 빅데이터 분석에는 데이터 전처리에 대한 이해와 전처리 이후 다양한 분석 기술에 대한 이해가 필요하다. 본 종설에서는 단일 세포 RNA-시퀀싱 데이터분석과 관련된 작업 과정의 개요를 제시한다. 먼저 데이터의 품질 관리, 정규화 및 차원 감소와 같은 데이터의 전 처리 과정에 대해 설명한다. 그 이후, 가장 일반적으로 사용되는 생물정보학 도구를 활용한 데이터의 후속 분석에 대해 설명한다. 본 종설은 이 분야에 관심이 있는 새로운 연구자를 위한 가이드라인을 제공하는 것을 목표로 한다.

Funding: None

Acknowledgements: None

Conflict of interest: None

Author's information (Position): Woo SH[1], Graduate student; Jung BC[2], Researcher.

Author Contributions

- Writing - original draft: Woo SH, Jung BC.
- Writing - review & editing: Woo SH, Jung BC.

Ethics approval

This article does not require IRB/IACUC approval because there are no human and animal participants.

ORCID

Sung-Hun WOO	https://orcid.org/0000-0002-1642-9341
Byung Chul JUNG	https://orcid.org/0000-0003-0732-0122

REFERENCES

1. Woo SH, Jung BC. Big data analytics in RNA-sequencing. Korean J Clin Lab Sci. 2023;55:235-243. https://doi.org/10.15324/kjcls.2023.55.4.235
2. Jung BC, You D, Lee I, Li D, Schill RL, Ma K, et al. TET3 plays a critical role in white adipose development and diet-induced remodeling. Cell Rep. 2023;42:113196. https://doi.org/10.1016/j.celrep.2023.113196
3. Kim TK, Bae EJ, Jung BC, Choi M, Shin SJ, Park SJ, et al. Inflammation promotes synucleinopathy propagation. Exp Mol Med. 2022;54:2148-2161. https://doi.org/10.1038/s12276-022-00895-w
4. Park S, Lee C, Ku BM, Kim M, Park WY, Kim NKD, et al. Paired analysis of tumor mutation burden calculated by targeted deep sequencing panel and whole exome sequencing in non-small cell lung cancer. BMB Rep. 2021;54:386-391. https://doi.org/10.5483/bmbrep.2021.54.7.045
5. Li X, Wang CY. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci. 2021;13:36. https://doi.org/10.1038/s41368-021-00146-0
6. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6:377-382. https://doi.org/10.1038/nmeth.1315
7. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50:1-14. https://doi.org/10.1038/s12276-018-0071-8
8. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9:75. https://doi.org/10.1186/s13073-017-0467-4
9. Mair F, Erickson JR, Voillet V, Simoni Y, Bi T, Tyznik AJ, et al. A targeted multi-omic analysis approach measures protein expression and low-abundance transcripts on the single-cell level. Cell Rep. 2020;31:107499. https://doi.org/10.1016/j.celrep.2020.03.063
10. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049. https://doi.org/10.1038/ncomms14049
11. Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, Marioni JC. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. 2019;20:63. https://doi.org/10.1186/s13059-019-1662-y
12. Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. 2020;21:57. https://doi.org/10.1186/s13059-020-1950-6
13. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016;17:29. https://doi.org/10.1186/s13059-016-0888-1
14. Osorio D, Cai JJ. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. Bioinformatics. 2021;37:963-967. https://doi.org/10.1093/bioinformatics/btaa751

15. Emont MP, Jacobs C, Essene AL, Pant D, Tenen D, Colleluori G, et al. A single-cell atlas of human and mouse white adipose tissue. Nature. 2022;603:926-933. https://doi.org/10.1038/s41586-022-04518-2
16. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Syst. 2019;8:281-291.e9. https://doi.org/10.1016/j.cels.2018.11.005
17. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst. 2019;8:329-337.e4. https://doi.org/10.1016/j.cels.2019.03.003
18. Xi NM, Li JJ. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. Cell Syst. 2021;12:176-194.e6. https://doi.org/10.1016/j.cels.2020.11.008
19. Lu J, Sheng Y, Qian W, Pan M, Zhao X, Ge Q. scRNA-seq data analysis method to improve analysis performance. IET Nanobiotechnol. 2023;17:246-256. https://doi.org/10.1049/nbt2.12115
20. Wu Y, Zhang K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. Nat Rev Nephrol. 2020;16:408-421. https://doi.org/10.1038/s41581-020-0262-0
21. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17:75. https://doi.org/10.1186/s13059-016-0947-7
22. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods. 2017;14:565-571. https://doi.org/10.1038/nmeth.4292
23. Choudhary S, Satija R. Comparison and evaluation of statistical error models for scRNA-seq. Genome Biol. 2022;23:27. https://doi.org/10.1186/s13059-021-02584-9
24. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20:296. https://doi.org/10.1186/s13059-019-1874-1
25. Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng. 2014;40:16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024
26. Yang P, Huang H, Liu C. Feature selection revisited in the single-cell era. Genome Biol. 2021;22:321. https://doi.org/10.1186/s13059-021-02544-3
27. Sheng J, Li WV. Selecting gene features for unsupervised analysis of single-cell gene expression data. Brief Bioinform. 2021;22:bbab295. https://doi.org/10.1093/bib/bbab295
28. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15:e8746. https://doi.org/10.15252/msb.20188746
29. Do VH, Canzar S. A generalization of t-SNE and UMAP to single-cell multimodal omics. Genome Biol. 2021;22:130. https://doi.org/10.1186/s13059-021-02356-5
30. Meyer BH, Pozo ATR, Nunan Zola WM. Global and local structure preserving GPU t-SNE methods for large-scale applications. Expert Syst Appl. 2022;201:116918. https://doi.org/10.1016/j.eswa.2022.116918
31. Lee JA, Renard E, Bernard G, Dupont P, Verleysen M. Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. Neurocomputing. 2013;112:92-108. https://doi.org/10.1016/j.neucom.2012.12.036
32. Ge H, Zhu Z, Lou K, Wei W, Liu R, Damaševičius R, et al. Classification of infrared objects in manifold space using Kullback-Leibler divergence of gaussian distributions of image points. Symmetry. 2020;12:434. https://doi.org/10.3390/sym12030434
33. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019;10:5416. https://doi.org/10.1038/s41467-019-13056-x
34. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018. [Epub ahead of print]. https://doi.org/10.1038/nbt.4314
35. Sainburg T, McInnes L, Gentner TQ. Parametric UMAP embeddings for representation and semisupervised learning. Neural Comput. 2021;33:2881-2907. https://doi.org/10.1162/neco_a_01434
36. Andrews TS, Kiselev VY, McCarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat Protoc. 2021;16:1-9. https://doi.org/10.1038/s41596-020-00409-w
37. Kobak D, Linderman GC. Initialization is critical for preserving global data structure in both t-SNE and UMAP. Nat Biotechnol. 2021;39:156-157. https://doi.org/10.1038/s41587-020-00809-z
38. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9:5233. https://doi.org/10.1038/s41598-019-41695-z
39. Hairol Anuar SH, Abas ZA, Yunos NM, Mohd Zaki NH, Hashim NA, Mokhtar MF, et al. Comparison between Louvain and Leiden algorithm for network structure: a review. J Phys Conf Ser. 2021;2129:012028. https://doi.org/10.1088/1742-6596/2129/1/012028
40. El Bouchefry K, de Souza RS. Learning in big data: introduction to machine learning. In: Škoda P, Adam F, editors. Knowledge discovery in big data from astronomy and earth observation: AstroGeoInformatics. Elsevier: 2020. p. 225-249.
41. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:15. https://doi.org/10.1186/s13059-017-1382-0
42. Martini E, Kunderfranco P, Peano C, Carullo P, Cremonesi M, Schorn T, et al. Single-cell sequencing of mouse heart immune infiltrate in pressure overload-driven heart failure reveals extent of immune activation. Circulation. 2019;140:2089-2107. https://doi.org/10.1161/circulationaha.119.041694
43. Qiu P. Embracing the dropouts in single-cell RNA-seq analysis. Nat Commun. 2020;11:1169. https://doi.org/10.1038/s41467-020-14976-9
44. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021;12:5692. https://doi.org/10.1038/s41467-021-25960-2
45. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018:15:255-261. https://doi.org/10.1038/nmeth.4612
46. Rauch A, Mandrup S. Transcriptional networks controlling stromal cell differentiation. Nat Rev Mol Cell Biol. 2021;22:465-482. https://doi.org/10.1038/s41580-021-00357-7
47. Bertoli C, Skotheim JM, de Bruin RA. Control of cell cycle transcription during G1 and S phases. Nat Rev Mol Cell Biol. 2013;14:518-528. https://doi.org/10.1038/nrm3629

48. Song D, Li JJ. PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data. Genome Biol. 2021;22:124. https://doi.org/10.1186/s13059-021-02341-y
49. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37:547-554. https://doi.org/10.1038/s41587-019-0071-9
50. Greenblatt MB, Ono N, Ayturk UM, Debnath S, Lalani S. The unmixing problem: a guide to applying single-cell RNA sequencing to bone. J Bone Miner Res. 2019;34:1207-1219. https://doi.org/10.1002/jbmr.3802
51. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32:381-386. https://doi.org/10.1038/nbt.2859
52. Zhang Y, Tran D, Nguyen T, Dascalu SM, Harris FC Jr. A robust and accurate single-cell data trajectory inference method using ensemble pseudotime. BMC Bioinformatics. 2023;24:55. https://doi.org/10.1186/s12859-023-05179-2
53. Van de Sande B, Lee JS, Mutasa-Gottgens E, Naughton B, Bacon W, Manning J, et al. Applications of single-cell RNA sequencing in drug discovery and development. Nat Rev Drug Discov. 2023;22:496-520. https://doi.org/10.1038/s41573-023-00688-4
54. Duncavage EJ, Bagg A, Hasserjian RP, DiNardo CD, Godley LA, Iacobucci I, et al. Genomic profiling for clinical decision making in myeloid neoplasms and acute leukemia. Blood. 2022;140:2228-2247. https://doi.org/10.1182/blood.2022015853
55. Kim N, Eum HH, Lee HO. Clinical perspectives of single-cell RNA sequencing. Biomolecules. 2021;11:1161. https://doi.org/10.3390/biom11081161
56. Riess JW, Gandara DR, Frampton GM, Madison R, Peled N, Bufill JA, et al. Diverse EGFR exon 20 insertions and co-occurring molecular alterations identified by comprehensive genomic profiling of NSCLC. J Thorac Oncol. 2018;13:1560-1568. https://doi.org/10.1016/j.jtho.2018.06.019
57. Vincent MD, Kuruvilla MS, Leighl NB, Kamel-Reid S. Biomarkers that currently affect clinical practice: EGFR, ALK, MET, KRAS. Curr Oncol. 2012;19(Suppl 1):S33-S44. https://doi.org/10.3747/co.19.1149
58. Qian J, Olbrecht S, Boeckx B, Vos H, Laoui D, Etlioglu E, et al. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. Cell Res. 2020;30:745-762. https://doi.org/10.1038/s41422-020-0355-0