SC-JNMF: single-cell clustering integrating multiple quantification methods based on joint non-negative matrix factorization

Mikio Shiga, Shigeto Seno, Makoto Onizuka and Hideo Matsuda

Graduate School of Information Science and Technology, Osaka University, Osaka, Japan

ABSTRACT

Single-cell RNA-sequencing is a rapidly evolving technology that enables us to understand biological processes at unprecedented resolution. Single-cell expression analysis requires a complex data processing pipeline, and the pipeline is divided into two main parts: The quantification part, which converts the sequence information into gene-cell matrix data; the analysis part, which analyzes the matrix data using statistics and/or machine learning techniques. In the analysis part, unsupervised cell clustering plays an important role in identifying cell types and discovering cell diversity and subpopulations. Identified cell clusters are also used for subsequent analysis, such as finding differentially expressed genes and inferring cell trajectories. However, single-cell clustering using gene expression profiles shows different results depending on the quantification methods. Clustering results are greatly affected by the quantification method used in the upstream process. In other words, even if the original RNA-sequence data is the same, gene expression profiles processed by different quantification methods will produce different clusters. In this article, we propose a robust and highly accurate clustering method based on joint non-negative matrix factorization (joint-NMF) by utilizing the information from multiple gene expression profiles quantified using different methods from the same RNA-sequence data. Our joint-NMF can extract common factors among multiple gene expression profiles by applying each NMF under the constraint that one of the factorized matrices is shared among multiple NMFs. The joint-NMF determines more robust and accurate cell clustering results by leveraging multiple quantification methods compared to conventional clustering methods, which use only a single gene expression profile. Additionally, we showed the usefulness of discovering marker genes with the extracted features using our method.

Subjects Bioinformatics, Computational Biology, Genomics, Computational Science

Keywords Single-cell, RNA-seq, Non-negative matrix factorization, Clustering

Submitted 10 March 2021
Accepted 7 August 2021
Published 27 August 2021

Corresponding author
Shigeto Seno, senoo@ist.osaka-u.ac.jp

Academic editor
Kenta Nakai

Additional Information and Declarations can be found on page 14

DOI **10.7717/peerj.12087**

Copyright
2021 Shiga et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

Advances in technology have made it possible to isolate individual cells from a population of cells and to sequence their transcriptomes at the single-cell level, known as single-cell RNA-sequencing (scRNA-seq). This technology has reached a surprising level of resolution that reveals the regulation of gene expression within cells. scRNA-seq measures gene expression on a cell-by-cell basis and allows the analysis of the functions and properties of cells using this information. Many experimental protocols and computational analyses exist for scRNA-seq, and they have different goals such as differential expression analysis, cell

clustering, cell classification, and trajectory reconstruction. Therefore, single-cell analysis forms a pipeline (a series of procedures) that mainly consists of two parts, quantification of gene expression and downstream analysis depends on the goals. In the quantification part of a pipeline, gene expression levels measured for each cell (raw sequence data) are converted to a matrix called the ''gene expression profile''. The quantification part also forms a pipeline including read quality control, adaptor trimming, demultiplexing, deduplicating using barcodes, mapping to genome, counting transcripts.

One of the objectives of single-cell analysis is to identify cell types by applying unsupervised clustering and extract a group of characteristic genes for a specific cell type as marker genes. Clustering is a useful method for the classification and identification of unknown cell groups and discovering the diversity and subpopulation of known cell types, and it is a fundamental step in scRNA-seq data analysis. It is the key to understanding cell function and constitutes the basis of other advanced analyses. Quantification is a critical factor for the subsequent clustering of analysis results. Different quantification methods result in different gene expression values even when the same RNA-seq library is processed. The greatest difference in the quantification methods is using an alignment-based method or an alignment-free method. As examples of alignment-based methods, short reads of the RNA-seq library are aligned to a reference genome using alignment tools, such as Bowtie2 (*Langmead & Salzberg, 2012*) and STAR (*Dobin et al., 2013*). Then, the gene expression values are obtained from the results using gene expression estimation tools such as RSEM (*Li et al., 2010*) and Cufflinks (*Trapnell et al., 2010*). In contrast, alignment-free tools such as Salmon(*Patro et al., 2017*) and kallisto(*Bray et al., 2016*) estimates mRNA abundances with k-mer counting approach (pseudo-alignment for transcriptome indices). Depends on the methods and the property of the RNA-seq library, several units of gene expression values are used such as RPKM (Reads per Kilobase Million), FPKM (Fragments per Kilobase Million), TPM (Transcripts per Million), UMI (Unique Molecular Identifiers) counts, and raw read counts. In addition, there are specialized quantification pipelines for the platform, such as Cell Ranger (*Zheng et al., 2017*) for 10X Genomics, and mappa for the ICELL8 system. To date, many methods for quantifying gene expression from RNA-seq library have been proposed; however, a consensus has not yet been reached on the best quantification method for all data (*Costa-Silva, Domingues & Lopes, 2017*; *Vieth et al., 2019*; *Wu et al., 2018*). Moreover, since each quantification method has different measurable genes, conventional analysis methods for cell clustering using gene expression quantified *via* only one method are strongly biased.

After quantification, feature selection or dimension reduction is required to analyze the gene expression profile because it has a large number of rows and columns (cells and genes). Additionally, single-cell analysis involves some differences among cells (*e.g.*, different gene expression values derived from different phases in the cell cycle and errors in measurement, such as missing values), even if these cells belong to the same cell type. For these reasons, we need a robust and data-driven clustering method for features that reflect variation in gene expression within individual cells. To date, many unsupervised feature selection, dimensionality reduction, and clustering methods have been developed (*Sun et al., 2019*; *Kiselev, Andrews & Hemberg, 2019*; *Freytag et al., 2018*). In particular, SC3 (*Kiselev et al.,*

2017) and Seurat (Satija et al., 2015) are useful data-driven analysis tools for single cells. SC3 is a clustering method that uses an ensemble of multiple analysis results using the algorithm based on k-means. Seurat is a method to analyze single cells, and Louvain clustering is used in cell clustering. Matrix factorization, as a method of unsupervised learning, is another efficient method for cell clustering and is excellent in data dimension reduction or the extraction of latent factors. In particular, non-negative matrix factorization(NMF) (Lee & Seung, 1999) is a suitable method for dimension reduction to extract the features of gene expression profiles because NMF interprets the data as a superposition of the gene functions and cell characteristics. NMF factorizes a matrix into multiple matrices (basis and coefficients) under the constraint that all elements are non-negative. The product of these matrices includes an approximate matrix of the input matrix. NMF is applied to various real data because of its non-negative feature and has been adapted to microarray data for clustering or feature extraction (Brunet et al., 2004; Zheng et al., 2011; Nik-Zainal et al., 2012; Zhang et al., 2012) and scRNA-seq data (Zhu et al., 2017; Wu et al., 2020). Especially, Shao & Höfer (2017) used LSNMF, using the projected gradient method to optimize the objective function (Lin, 2007), for single-cell clustering. By performing NMF, we can analyze more details of genes and cells (*e.g.*, finding marker genes and performing unsupervised cell clustering) with the feature matrices extracted.

However, these clustering methods do not consider the differences in quantification methods that are upstream of this series of procedures. We usually observe that different quantification pipelines produce similar but also different gene expression profiles even when we apply them to the same RNA-seq library. Consequently, different clusters are obtained and misunderstanding might be produced. Our objective is to develop a method that integrates the gene expression profiles from different quantification methods to extract reliable feature matrices for clustering. Joint-NMF, which performs multiple NMF with a shared matrix, is one of the most suitable methods for integrating such different but potentially similar data. Joint-NMF has also been studied for genomic applications, Wang, Zheng & Zhao (2015) simultaneously decompose multiple transcriptomics data matrices. Zhang et al. (2012), Yang & Michailidis (2016), Duren et al. (2018), Jin, Zhang & Nie (2020) integrated heterogeneous omics multi-modal data (*e.g.*, DNA methylation, miRNA expression, and gene expression) to detect modules, and Fujita et al. (2018) also integrated multi-omics data to discover biomarkers with a combination of Joint-NMF and pathway analysis. These methods decompose the data into a shared matrix of "genes × modules" and some dedicated matrices of "modules × samples" for each omics data. Genes are regarded as common variables across data matrices and samples are different. In contrast, we decompose the data into some matrices of "genes × modules" and a shared matrix of "modules × cells" because the cells included in the gene expression profiles are perfectly the same in our problem setting.

In this article, we propose SC-JNMF, a novel unsupervised clustering method using NMF to eliminate the differences in quantification methods and extract the common factors over multiple gene expression profiles. The features of the data can be decomposed into gene-derived factors that contain bias dependent on each quantification method and common cell-derived factors, and cell clustering can be performed based on these common

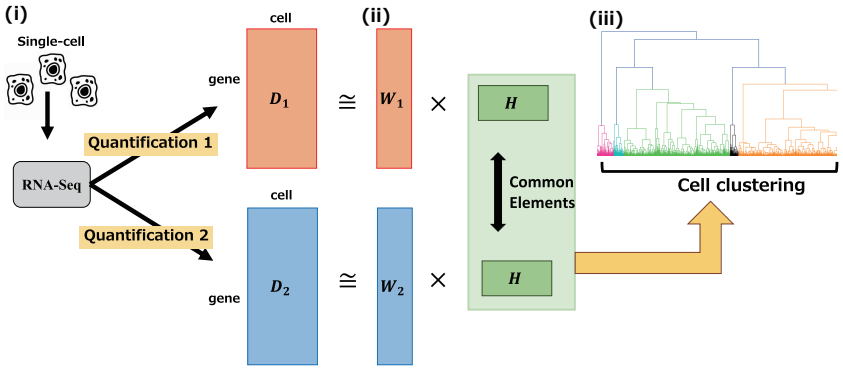

Figure 1 **A workflow of SC-JNMF.** (i) Creating multiple gene expression profiles using different quantification methods. (ii) Extracting the common factor among these gene expression profiles from the same RNA-seq library. (iii) Cell clustering using hierarchical clustering with the extracted common factor.
Full-size DOI: 10.7717/peerj.12087/fig-1

factors to obtain more essential biological information. To our knowledge, this study is the first to incorporate multiple quantification methods into the clustering analysis of scRNA-seq data.

MATERIALS & METHODS

The outline of SC-JNMF

We proposed SC-JNMF, a method that extracts latent factors from different gene expression profiles at the same time using joint-NMF, which can express matrix data as the product of lower-dimensional matrices, one of which is shared. SC-JNMF extracted the latent factors in different gene expression profiles using a similar approach to NMF and used them for cell clustering and gene analysis.

The outline of our method is showed in Fig. 1. We created two different gene expression profiles using different quantification methods from the same RNA-seq library. Then, we extracted the common factor using joint-NMF and extended it to perform multiple NMFs in parallel with two different basis factors $\boldsymbol{W}_1, \boldsymbol{W}_2$ (derived from the different methods) and shared factors $\boldsymbol{H}$ (derived from the original RNA-seq library). Finally, we performed cell clustering using these extracted factors and any appropriate clustering methods, such as hierarchical clustering. Although the conventional NMF clustering methods that perform clustering directly with the factorized matrix strongly depend on the rank, our method performs robust clustering by using hierarchical clustering.

Quantification and normalization

As the first step of SC-JNMF, we quantified gene expressions with different quantification pipelines, and obtained a set of gene expression profiles. Parameters and references could be set as recommended by each quantification pipeline.

For preprocessing before performing joint matrix factorization, we applied a gene filter to the input data, similar to the SC3 method (*Kiselev et al., 2017*), and then $\log_2(x+1)$ transformed the data. Finally, we normalized the data so that the sum of each gene L^1 norm

was 1 because it should be compared the relative expression values between cells for each gene to perform the cell clustering. This L_1 normalization is possible to consider smaller expression values than using the L_2 norm and to keep the values non-negative.

Joint-NMF

We considered the given data as a non-negative matrix $\boldsymbol{D} \in \mathbb{R}_+^{N \times M}$. Non-negative matrix factorization found the basis matrix $\boldsymbol{W} \in \mathbb{R}_+^{N \times k}$ and coefficient matrix $\boldsymbol{H} \in \mathbb{R}_+^{k \times M}$, where all the elements were non-negative, such that these matrix products approximated the input data matrix.

$$\boldsymbol{D} \approx \boldsymbol{WH} \tag{1}$$

NMF minimized the distance between matrix $\boldsymbol{D}$ and matrix $\boldsymbol{WH}$. Here, we considered a Euclidean norm distance. Thus, the objective function that NMF minimizes was as follows:

$$\min_{\boldsymbol{W},\boldsymbol{H} \geq 0} L := \|\boldsymbol{D} - \boldsymbol{WH}\|_{\mathcal{F}}^2 \tag{2}$$

Here, we considered a simultaneous matrix factorization of two matrices using NMF. Given the two input data as non-negative matrices $\boldsymbol{D}_1 \in \mathbb{R}_+^{N_1 \times M}$ and $\boldsymbol{D}_2 \in \mathbb{R}_+^{N_2 \times M}$, joint-NMF found the basis matrices $\boldsymbol{W}_1 \in \mathbb{R}_+^{N_1 \times k}$, $\boldsymbol{W}_2 \in \mathbb{R}_+^{N_2 \times k}$ corresponding to each approximated input matrix and the common coefficient matrix $\boldsymbol{H} \in \mathbb{R}_+^{k \times M}$, minimizing the distances between the given matrices and the approximated matrices. In our proposed method, we added the L1 norm constraint to this objective function for $\boldsymbol{H}$ so that the factorized shared matrix was sparse. Thus, SC-JNMF found the matrix $\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{H}$ that minimized the following objective function:

$$\min_{\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{H},\lambda_n \geq 0} L := \|\boldsymbol{D}_1 - \boldsymbol{W}_1\boldsymbol{H}\|_{\mathcal{F}}^2 + \lambda_1 \|\boldsymbol{D}_2 - \boldsymbol{W}_2\boldsymbol{H}\|_{\mathcal{F}}^2 + \lambda_2 \sum_k \|\boldsymbol{H}^{k,*}\|_1 \tag{3}$$

where, λ_1 is a balance parameter for the losses of reconstruction matrices and λ_2 is a parameter for row vector sparsity regularization of shared factorized matrix.

We applied a multiplicative update algorithm to optimize the objective function same as conventional NMF. By applying Jensen's inequalities to the first and second terms of the objective function, the function to be minimized could be rewritten as follows:

$$\begin{aligned}\min_{\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{H},\lambda_n \geq 0} L := & \sum_{i,j} \left(|\boldsymbol{D}_1^{i,j}|^2 - 2\boldsymbol{D}_1^{i,j} \sum_k \boldsymbol{W}_1^{i,k}\boldsymbol{H}^{k,j} + \sum_k \frac{(\boldsymbol{W}_1^{i,k})^2 (\boldsymbol{H}^{k,j})^2}{c_1^{i,j,k}} \right) \\ & + \lambda_1 \sum_{i,j} \left(|\boldsymbol{D}_2^{i,j}|^2 - 2\boldsymbol{D}_2^{i,j} \sum_k \boldsymbol{W}_2^{i,k}\boldsymbol{H}^{k,j} + \sum_k \frac{(\boldsymbol{W}_2^{i,k})^2 (\boldsymbol{H}^{k,j})^2}{c_2^{i,j,k}} \right) \\ & + \lambda_2 \|\boldsymbol{H}\|_1 \end{aligned} \tag{4}$$

where,

$$c_1^{i,j,k} = \frac{\boldsymbol{W}_1^{i,k}\boldsymbol{H}^{k,j}}{\sum_{k'} \boldsymbol{W}_1^{i,k'}\boldsymbol{H}^{k',j}} \tag{5}$$

$$c_2^{i,j,k} = \frac{\boldsymbol{W}_2^{i,k}\boldsymbol{H}^{k,j}}{\sum_{k'}\boldsymbol{W}_2^{i,k'}\boldsymbol{H}^{k',j}} \tag{6}$$

We found each element of $\boldsymbol{W}_1, \boldsymbol{W}_2$, and $\boldsymbol{H}$ that minimized the objective function by performing partial differentiation.

$$\frac{\partial L}{\partial W_1^{i,k}} = \sum_j (-2D_1^{i,j}H^{k,j} + \frac{2W_1^{i,k}(H^{k,j})^2}{c_1^{i,j,k}}) \tag{7}$$

$$\frac{\partial L}{\partial W_2^{i,k}} = \lambda_1 \sum_j (-2D_2^{i,j}H^{k,j} + \frac{2W_2^{i,k}(H^{k,j})^2}{c_2^{i,j,k}}) \tag{8}$$

$$\begin{aligned}\frac{\partial L}{\partial H^{k,j}} &= \sum_i (-2D_1^{i,j}W_1^{i,k} + \frac{2(W_1^{i,k})^2 H^{k,j}}{c_1^{i,j,k}}) \\ &\quad + \lambda_1 \sum_{i'} (-2D_2^{i',j}W_2^{i',k} + \frac{2(W_2^{i',k})^2 H^{k,j}}{c_2^{i',j,k}}) + \lambda_2\end{aligned} \tag{9}$$

The objective function are minimized when these are 0. Thus, the variable updates became:

$$\boldsymbol{W}_1 = \boldsymbol{W}_1 \frac{\boldsymbol{D}_1\boldsymbol{H}^\top}{[\boldsymbol{H}[\boldsymbol{W}_1\boldsymbol{H}]^\top]^\top} \tag{10}$$

$$\boldsymbol{W}_2 = \boldsymbol{W}_2 \frac{\lambda_1\boldsymbol{D}_2\boldsymbol{H}^\top}{\lambda_1[\boldsymbol{H}[\boldsymbol{W}_2\boldsymbol{H}]^\top]^\top} \tag{11}$$

$$\boldsymbol{H} = \boldsymbol{H} \frac{[\boldsymbol{D}_1^\top\boldsymbol{W}_1]^\top + \lambda_1[\boldsymbol{D}_2^\top\boldsymbol{W}_2]^\top - \lambda_2/2}{\boldsymbol{W}_1^\top\boldsymbol{W}_1\boldsymbol{H} + \lambda_1\boldsymbol{W}_2^\top\boldsymbol{W}_2\boldsymbol{H}} \tag{12}$$

Applications

Cell clustering

The factorized matrices were used to perform highly accurate clustering. In our proposed method, we used hierarchical clustering (Ward's method (*Ward, 1963*), implemented in SciPy (*Virtanen et al., 2020*)). In this study, the adjusted Rand index (ARI, implemented in scikit-learn (*Pedregosa et al., 2011*)) was used to evaluate clustering performance.

Given a set of n elements (*i.e.,* cells) $S = \{o_1, \ldots, o_n\}$, and supposing that $U = \{u_1, \ldots, u_R\}$ and $V = \{v_1, \ldots, v_C\}$ represent two different partitions of S, define the following:

1. a, the number of cell pairs which are assigned to the same class in both U and V
2. b, the number of cell pairs which are assigned to different classes by both U and V

The Rand index was calculated by

$$RI = \frac{a+b}{\binom{n}{2}} \tag{13}$$

In addition, the ARI considered adjusting measures of clustering accuracy for chance. The ARI was defined using the following formula:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}/\binom{n}{2}}{\binom{n}{2} - \sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}/\binom{n}{2}} \tag{14}$$

where, $n_{ij} = |U_i \cap V_j|$, $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot j} = \sum_i n_{ij}$.

Table 1 Dataset and quantification methods in previous studies.

Dataset (citaion)	Tissue/process	Quantification	Reference	The number of cells	The number of classes
Treutlein (*Treutlein et al., 2014*)	Lung epithelium	TopHat v2.0.8, Cufflinks v2.0.2	mm10	80	5
Pollen (*Pollen et al., 2014*)	Brain	TopHat v2.0.4, RSEM v1.2.4	hg19	259	10
Segerstolpe (*Segerstolpe et al., 2016*)	Pancrease	STAR v2.3.0e, rpkmforgenes	hg19	2166	12
Xin (*Xin et al., 2016*)	Pancrease	CLC Bio Genomics Workbench v7.0	GRCh37	1492	4
Monaco (*Monaco et al., 2019*)	Immune cells	Kallisto v0.43.1, Tximport v1.6.0	GENCODE Human 26	114	10
CellBench (sc_10x_5cl) (*Tian et al., 2019*)	Lung adenocarcinoma cell lines	scPipe v.1.3.0	GRCh38	3918	5

Gene analysis

Our proposed method was used not only for cell clustering but also for gene analysis using the extracted factors. The factors in the coefficient matrix that showed higher values in specific clusters than those in other clusters indicated latent factors of the cluster. Therefore, the same factors in the basis matrices also showed latent factors of the cluster, and the genes that showed higher values in the basis matrices reflected characteristic features of the cluster, in other words, marker genes.

RESULTS

To assess the accuracy of our method, we performed cell clustering using five scRNA-seq datasets and one bulk RNA-seq dataset. In this clustering, we estimated the optimal parameters (the ranks in matrix factorization) using the trade-off relationship between sparseness and loss. We also evaluated the effect of the combinations of quantification methods and the sample size for the clustering accuracy. Additionally, we analyzed more details of the factorized matrices by showing their relevance to marker genes.

Dataset

We used six different datasets including RNA-seq library and quantified gene expression levels measured using each method in previous studies (Table 1). Only Monaco dataset is a bulk RNA-seq, the others are scRNA-seq libraries. Because Monaco dataset is a set of sorted cell types, it is suitable for the evaluation of clustering accuracy. CellBench is a set of datasets designed to benchmark various single-cell data analysis. In this study, we used the dataset named "sc_10x_5cl", single cells from the mixture of five cell lines. In advance, we removed any cell types that were unclear in previous studies.

To perform our method SC-JNMF and compare the accuracy of the clustering methods, We alternatively quantified gene expression values in each cell from the RNA-seq library using Salmon(v1.0.0) with GENCODE references and annotations (Mouse Release 21/ Human Release 32). Only CellBench dataset was quantified by Using Salmon with Alevin

(*Srivastava et al., 2019*), kallisto with BUStools and STAR with Cell Ranger to process barcodes in short reads.

In addition, for the first 4 datasets, we also prepared other gene expression profiles using kallisto(v0.46.2), STAR(v2.7.3), to evaluate our method for the combinations of quantification methods and the sample size.

Settings of clustering methods

We compared the accuracy of cell clustering using our proposed method to that obtained using other major unsupervised clustering methods, including LSNMF (*Lin, 2007*; *Zitnik & Zupan, 2012*), SC3 (*Kiselev et al., 2017*), and Seurat (*Satija et al., 2015*).

For our proposed method, each run consisted of ten runs and extracted an ARI score of the top combined rank in terms of sparseness and loss (the top ranking of sparseness is maximum and the top ranking of loss is minimum). The rank was determined according to the results of our experiment described in the Supplemental Information and Fig. S1 as follows: Treutlein, $k = 5$; Pollen, $k = 8$; Segerstolpe, $k = 25$; Xin, $k = 16$; Monaco, $k = 8$; CellBench, $k = 5$. In addition, we determined the regularization parameter $\lambda_1 = |geneset\,1|/|geneset\,2|$. Hierarchical clustering could not determine the number of clusters; therefore, we set the same number reported in previous studies in advance.

In the LSNMF method, similar to our joint-NMF, classification by hierarchical clustering was performed using a matrix of factors of the cells, and the rank and the number of clusters were set to the same values as ours. SC3 method was performed with the default parameters. For LSNMF and SC3, each run was performed ten times, considering random initial values. For Seurat, we plotted only the highest ARI score in the runs as the resolution parameter in the FindClusters function was increased from 0 to 1 in 0.1 steps. The parameters of the FindNeighbors function were set to the default values (dims = 1:10, k.param = 20).

Accuracy of cell clustering

Figure 2 shows the ARI score of the original (quantified previously) and alternative (quantified using Salmon) data for each clustering method. We ran a two-way experiment using $\lambda_2 = 1$ and $\lambda_2 = 10$.

In the Pollen, Xin, and CellBench datasets, our proposed method performed accurate cell clustering. In contrast, in the Treutlein, Segerstolpe, and Monaco datasets, the ARI score of the proposed method was higher than that of LSNMF, but it was not the highest score. Additionally, the resulting clusters of our method were stable compared to the conventional NMF method (except for the CellBench dataset).

Comparison of quantification methods in SC-JNMF

SC-JNMF performed cell clustering using one or two gene expression profiles. In this study, we compared the combinations of quantification methods (kallisto, STAR, Salmon) by performing cell clustering and calculating ARI using SC-JNMF. To evaluate the impact of "joint" NMF, we also performed factorization and clustering using a single gene expression profile with setting to the parameter $\lambda_1 = 0$. We ran SC-JNMF repeatedly for ten trials with random initial values and the parameter $\lambda_2 = 10$.

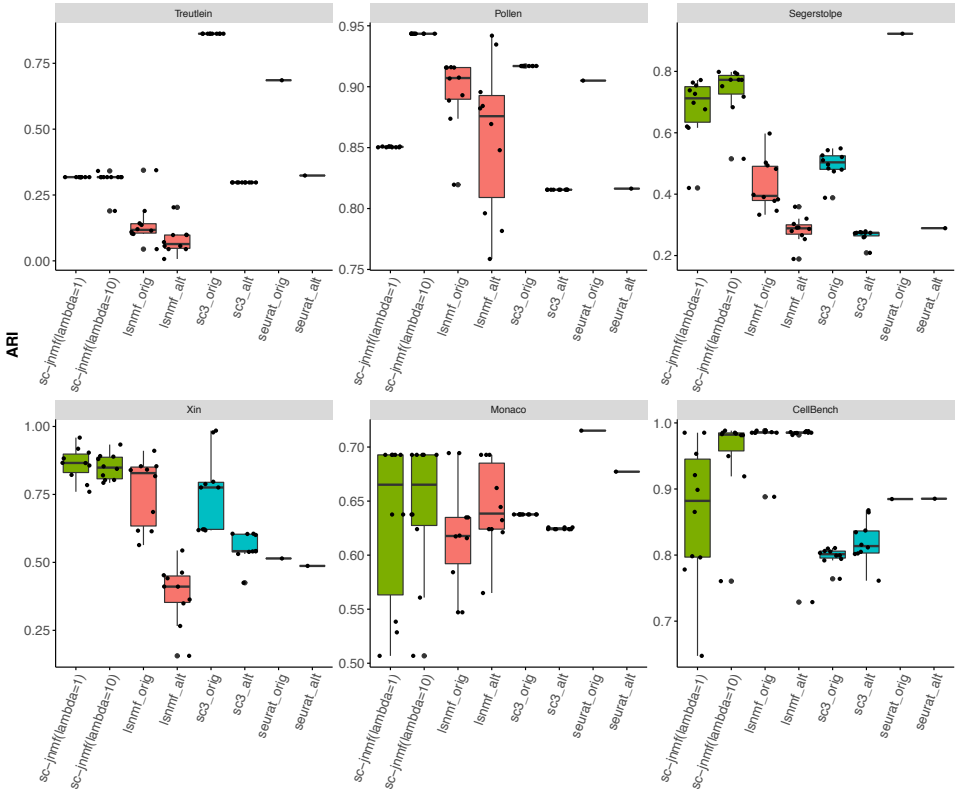

Figure 2 The accuracy (ARI) of cell clustering in each method. For comparison methods, we performed experiments using the original gene expression profile (quantified previously) and alternative gene expression profile (quantified using Salmon). For our proposed method, we performed the experiment in two ways with different regularization parameters ($\lambda_2 = 1$ and $\lambda_2 = 10$).

Full-size DOI: 10.7717/peerj.12087/fig-2

Figure 3 shows the ARI of the cell clustering using SC-JNMF in each quantification method and their combination. As a result, for each dataset, the accuracies of kallisto & STAR and Salmon & STAR combination outperformed using the single quantification method. The accuracy of Salmon & kallisto combination was lower than that of other combinations.

Additionally, we counted the frequency of correlation coefficients for each cell and each gene in each dataset and combination (Fig. 4). In Salmon & kallisto combination, the correlation coefficients were inclined toward a higher value than the other combinations. These results indicated that SC-JNMF had a lower accuracy when using similar gene expression profiles (*e.g.*, Salmon & kallisto combination). Therefore, two gene expression profiles with different properties, having a lower correlation of cells and genes, were suitable for SC-JNMF.

Size effect for SC-JNMF and NMF

To further understand the benefit of NMF-based methods, the effect of sample size (the number of cells in the RNA-seq library) against clustering accuracy was evaluated. We

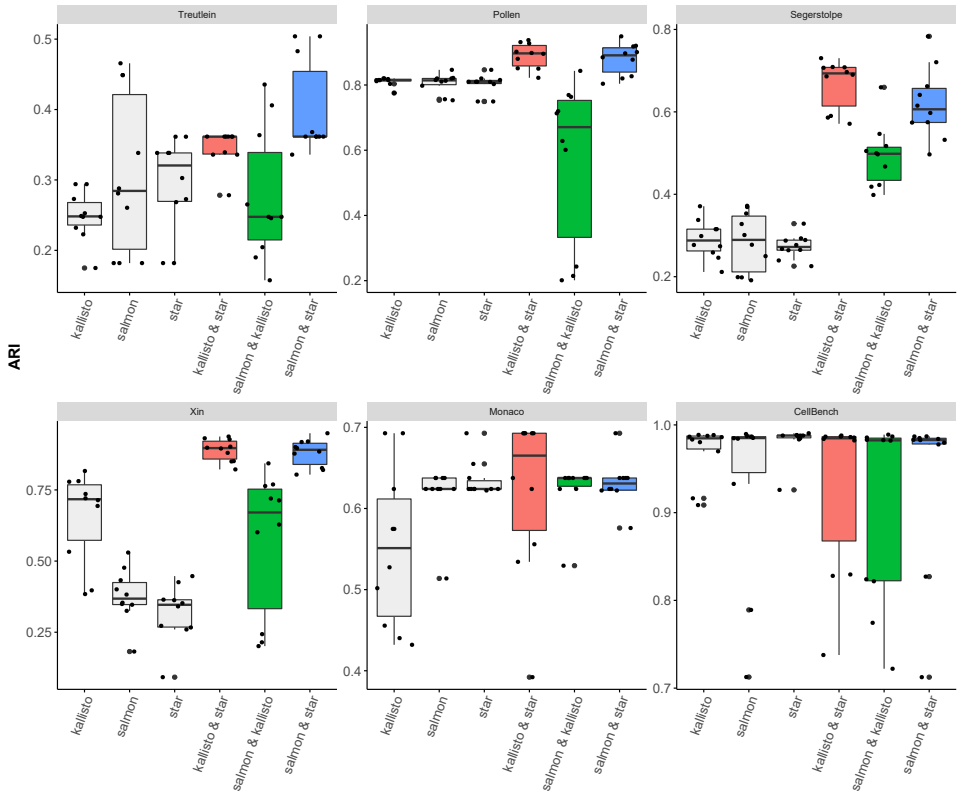

Figure 3 The accuracy (ARI) of cell clustering in each quantification method and their combination.
Full-size DOI: 10.7717/peerj.12087/fig-3

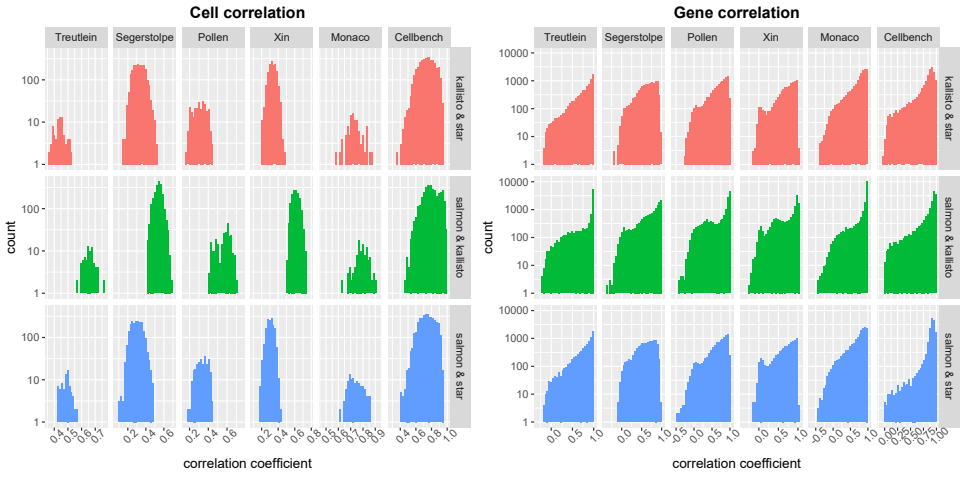

Figure 4 Histograms ($\log_{10}$ scale) of correlation coefficients in each gene and cell.
Full-size DOI: 10.7717/peerj.12087/fig-4

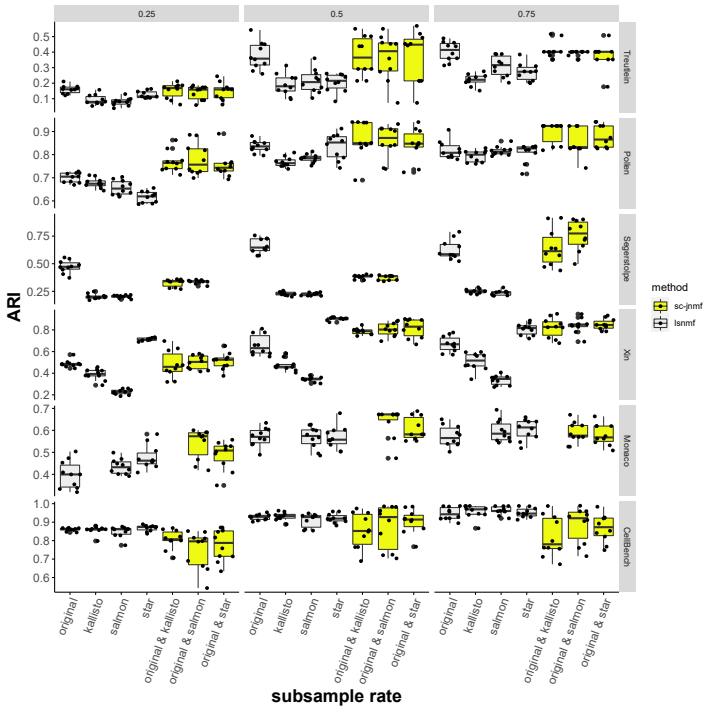

Figure 5 The accuracy (ARI) of cell clustering against subsample rate. Because the original quantification method of Segerstolpe dataset is STAR, the plots for the "star" and the combination of "original & star" are blank. For the same reason, "kallisto" and "original & kallisto" are blank for the Monaco dataset.
Full-size DOI: 10.7717/peerj.12087/fig-5

randomly subsampled the 25%, 50%, and 75% cells from gene expression profiles, and ran SC-JNMF ($\lambda_2 = 10$) and LSNMF. Each trial was performed ten times.

Figure 5 shows the ARI of the cell clustering using LSNMF and SC-JNMF for each dataset and subsample rate. As a result, for each dataset, the accuracies of both LSNMF and SC-JNMF were increased depends on the subsample rate. This result indicated that the methods based on NMF benefit more from a dataset including a larger number of cells.

Gene analysis using factorized matrix

We found marker genes of the Xin dataset using factorized matrices. The Xin dataset contained data on 1492 single cells and four classes (alpha, beta, delta, and PP) in the pancreas. We showed the results in Fig. 6. The factor of the coefficient matrix showed some characteristic patterns in each cell cluster (*e.g.*, Factor 2 and Factor 5 showed higher values in delta and PP cells) (Fig. 6A). Next, we calculated the correlation of factors between the basis matrices in the common genes (Fig. 6B). The factors in the basis matrices showed a similar tendency. We also showed the loadings of marked factors for delta cells in the coefficient matrix in Fig. 6C. Almost all genes showed similar factor loadings between base1 and base2; however, we confirmed that some of the genes only observed in either gene expression profile also had high values. We also showed the marker genes of delta cells detected using Scanpy *Wolf, Angerer & Theis (2018)* in the scatter plot. The factor loadings of these genes tended to be higher than those of others, regardless of whether the gene was

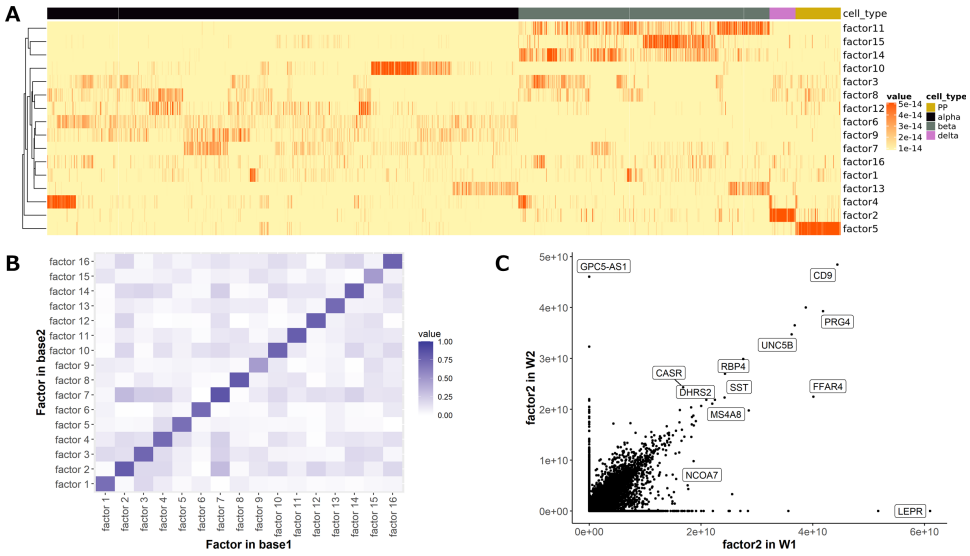

Figure 6 **Gene analysis of the Xin dataset using SC-JNMF.** (A) A heatmap of the common coefficient matrix generated using SC-JNMF. (B) The correlation of the factor loadings in genes common to the basis matrix1 and basis matrix2. (C) "factor 2" loadings of each gene in base1 and base2 and marker genes of delta cells in original and alternative gene expression profiles detected using Scanpy (*Wolf, Angerer & Theis, 2018*).

Full-size DOI: 10.7717/peerj.12087/fig-6

observed in both gene expression profiles or not. All factors of Xin dataset were shown in Fig. S2.

DISCUSSION

Thus far, we have shown the possibility that our unsupervised clustering method had high accuracy when using the differences in gene expression quantification methods compared with previous studies. However, the proposed method showed worse accuracy than SC3 in the original Treutlein dataset, as well as Seurat in the original Segerstolpe dataset, and showed differences in the accuracy due to the differences in regularization parameters in the Pollen dataset. Treutlein and Pollen datasets have fewer cells than the others, which was one of the most important features of our experiments. Dataset size is an important factor (common to machine learning approaches) for high accuracy, and this characteristic is also applied to our method.

We showed the effective combinations of quantification methods by comparing the accuracy of cell clustering for each combination. In particular, we suggested that the combination of gene expression profiles that have similar properties (*e.g.*, Salmon & kallisto) had lower accuracy in SC-JNMF. Compared to the combination, including STAR that maps RNA-seq reads to a reference genome, Salmon and kallisto are similar methods in the quantification algorithm, as these are alignment-free quantification methods. Therefore, these gene expression profiles have similar properties. In SC-JNMF using similar gene expression profiles, it is difficult to separate common factors derived from cells ($\boldsymbol{H}$) and factors derived from genes ($\boldsymbol{W}_1$, $\boldsymbol{W}_2$). Meanwhile, although it has been

reported that the pseudo-alignment method loses many reads, leading to a lower mean expression (*Vieth et al., 2019*), it is possible to improve the clustering performance by incorporating them into our method with the other quantification method.

We presented more details about the characteristics of the factorized matrix and the relationships of marker genes. The coefficient matrix showed characteristic factors in each cell cluster, and both basis matrices had similar factors. The marker genes of a cell cluster showed high factor loadings in the basis matrices that characterized a specific cell cluster in the coefficient matrix, regardless of including both gene expression profiles or only one. In other words, factor loadings in basis matrices that characterize a specific cell cluster in the coefficient matrix are related to the marker genes for that cluster. This result suggested that genes showing high factor loadings in the basis matrices probably had some important features in the cluster with high factor loadings in the coefficient matrix. In particular, we should pay particular attention to those genes observed only in either gene expression profile because they are not considered in conventional methods.

In summary, our method is effective in the following cases.

- The number of cells is sufficient large.
- Different quantification methods yield gene expression profiles with different characteristics.

Meanwhile, there are some limitations. As shown in the result of CellBench dataset, the number of cells is quite large and the nature of the cell type is well defined, jointless NMF gives sufficiently good results. In such a case, our joint-NMF may deteriorate the stability of the solution. Moreover, we also observed cases that the joint-NMF worsened the accuracy (‘‘salmon & kallisto’’ combination in Pollen dataset, Fig. 3), although the effect of the ‘‘joint’’ was either better or unchanged in many cases. It should be avoided to input expression profiles quantified by similar approaches.

CONCLUSION

We proposed SC-JNMF, which performs cell clustering using common factors extracted from multiple gene expression profiles quantified using different methods. As a result, it is possible to perform robust analysis compared with the case in which only a single quantification method is used because it is unnecessary to consider the differences in gene expression profiles. The accuracy (ARI) of cell clustering obtained using our method was higher than that of other major clustering methods. Additionally, we showed that the combination of different quantification methods increases the accuracy of cell clustering compared to that of a similar quantification. Moreover, we showed the details of the extracted factors. The genes characteristic to specific cell groups (marker genes) showed remarkable factor loadings in terms of the factorized matrices; in other words, these results suggest a potential for identifying important genes in the dataset. Some genes may not be counted depending on the quantification methods used; they can be detected using multiple gene expression profiles generated using different quantification methods and SC-JNMF if they are potential markers.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by JSPS KAKENHI Grant Number 19H04207, Japan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
JSPS KAKENHI, Japan: 19H04207.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Mikio Shiga and Shigeto Seno conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Makoto Onizuka and Hideo Matsuda analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

SC-JNMF is implemented in Python. The source code and documentation are available at GitHub: https://github.com/agis09/sc-jnmf.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.12087#supplemental-information.

REFERENCES

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34(5)**:525–527 DOI 10.1038/nbt.3519.

Brunet J-P, Tamayo P, Golub TR, Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America* **101(12)**:4164–4169 DOI 10.1073/pnas.0308531101.

Costa-Silva J, Domingues D, Lopes FM. 2017. RNA-Seq differential expression analysis: an extended review and a software tool. *PLOS ONE* **12(12)**:e0190152 DOI 10.1371/journal.pone.0190152.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29(1)**:15–21 DOI 10.1093/bioinformatics/bts635.

Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, Wang Y, Wong WH. 2018. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proceedings of the National Academy of Sciences of the United States of America* **115(30)**:7723–7728 DOI 10.1073/pnas.1805681115.

Freytag S, Tian L, Lönnstedt I, Ng M, Bahlo M. 2018. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research* **7** DOI 10.12688/f1000research.15809.2.

Fujita N, Mizuarai S, Murakami K, Nakai K. 2018. Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. *Scientific Reports* **8(1)**:9743 DOI 10.1038/s41598-018-28066-w.

Jin S, Zhang L, Nie Q. 2020. scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biology* **21(1)**:1–19 DOI 10.1186/s13059-019-1906-x.

Kiselev VY, Andrews TS, Hemberg M. 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* **20(5)**:273–282 DOI 10.1038/s41576-018-0088-9.

Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, Hemberg M. 2017. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods* **14(5)**:483–486 DOI 10.1038/nmeth.4236.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9(4)**:357 DOI 10.1038/nmeth.1923.

Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* **401(October 1999)**:788–791 DOI 10.1038/44565.

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26(4)**:493–500 DOI 10.1093/bioinformatics/btp692.

Lin CJ. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural Computation* **19(10)**:2756–2779 DOI 10.1162/neco.2007.19.10.2756.

Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carre C, Burdin N, Visan L, Ceccarelli M, Poidinger M, Zippelius A, Pedro de Magalhães J, Larbi A. 2019. RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Reports* **26(6)**:1627–1640 DOI 10.1016/j.celrep.2019.01.041.

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jönsson G, Garber JE, Silver D, Miron P, Fatima A, Boyault S, Langerod A, Tutt A, Martens JW, Aparicio SA, Borg A, Salomon AV, Thomas G, Borresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR. 2012. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149(5)**:979–993 DOI 10.1016/j.cell.2012.04.024.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14(4)**:417–419 DOI 10.1038/nmeth.4197.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay é. 2011. Scikit-learn: machine learning in Python. *The Journal of Machine Learning Research* **12**:2825–2830.

Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, Li N, Szpankowski L, Fowler B, Chen P, Ramalingam N, Sun G, Thu M, Norris M, Lebofsky R, Toppani D, Kemp DW, Wong M, Clerkson B, Jones BN, Wu S, Knutsson L, Alvarado B, Wang J, Weaver LS, May AP, Jones RC, Unger MA, Kriegstein AR, West JA. 2014. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology* **32(10)**:1053–1058 DOI 10.1038/nbt.2967.

Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33(5)**:495–502 DOI 10.1038/nbt.3192.

Segerstolpe Å, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK, Smith DM, Kasper M, Ämmälä C, Sandberg R. 2016. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metabolism* **24(4)**:593–607 DOI 10.1016/j.cmet.2016.08.020.

Shao C, Höfer T. 2017. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics* **33(2)**:235–242 DOI 10.1093/bioinformatics/btw607.

Srivastava A, Malik L, Smith T, Sudbery I, Patro R. 2019. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biology* **20(1)**:1–16 DOI 10.1186/s13059-018-1612-0.

Sun S, Zhu J, Ma Y, Zhou X. 2019. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology* **20(1)**:1–21 DOI 10.1186/s13059-018-1612-0.

Tian L, Dong X, Freytag S, Lê Cao K-A, Su S, JalalAbadi A, Amann-Zalcenstein D, Weber TS, Seidi A, Jabbari JS, Naik SH, Ritchie ME. 2019. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature Methods* **16(6)**:479–487 DOI 10.1038/s41592-019-0425-8.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28(5)**:511–515 DOI 10.1038/nbt.1621.

Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR. 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509(7500)**:371–375 DOI 10.1038/nature13173.

Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. 2019. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nature Communications* **10(1)**:1–11 DOI 10.1038/s41467-018-07882-8.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, Van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson AR, Jones E, Kern R, Larson E, Carey CJ, Polat I, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, Van Mulbregt P, Vijaykumar A., Bardelli AP, Rothberg A, Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN, Fulton C, Masson C, Häggström C, Fitzgerald C, Nicholson DA, Hagen DR, Pasechnik DV, Olivetti E, Martin E, Wieser E, Silva F, Lenders F, Wilhelm F, Young G, Price GA, Ingold GL, Allen GE, Lee GR, Audren H, Probst I, Dietrich JP, Silterra J, Webber JT, Slavič J, Nothman J, Buchner J, Kulick J, Schönberger JL, De Miranda Cardoso JV, Reimer J, Harrington J, Rodríguez JLC, Nunez-Iglesias J, Kuczynski J, Tritz K, Thoma M, Newville M, Kümmerer M, Bolingbroke M, Tartre M, Pak M, Smith NJ, Nowaczyk N, Shebanov N, Pavlyk O, Brodtkorb PA, Lee P, McGibbon RT, Feldbauer R, Lewis S, Tygier S, Sievert S, Vigna S, Peterson S, More S, Pudlik T, Oshima T, Pingel TJ, Robitaille TP, Spura T, Jones TR, Cera T, Leslie T, Zito T, Krauss T, Upadhyay U, Halchenko YO, Vázquez-Baeza Y. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17(3)**:261–272 DOI 10.1038/s41592-019-0686-2.

Wang HQ, Zheng CH, Zhao XM. 2015. JNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data. *Bioinformatics* **31(4)**:572–580 DOI 10.1093/bioinformatics/btu679.

Ward JH. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58(301)**:236–244 DOI 10.1080/01621459.1963.10500845.

Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19(1)**:1–5 DOI 10.1186/s13059-017-1381-1.

Wu P, An M, Zou HR, Zhong CY, Wang W, Wu CP. 2020. A robust semi-supervised NMF model for single cell. *PeerJ* **8**:e10091 DOI 10.7717/peerj.10091.

Wu DC, Yao J, Ho KS, Lambowitz AM, Wilke CO. 2018. Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics* **19(1)**:1–14 DOI 10.1186/s12864-017-4368-0.

Xin Y, Kim J, Okamoto H, Ni M, Wei Y, Adler C, Murphy AJ, Yancopoulos GD, Lin C, Gromada J. 2016. RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metabolism* **24(4)**:608–615.

Yang Z, Michailidis G. 2016. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **32(1)**:1–8 DOI 10.1093/bioinformatics/btv544.

Zhang S, Liu CC, Li W, Shen H, Laird PW, Zhou XJ. 2012. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research* **40(19)**:9379–9391 DOI 10.1093/nar/gks725.

Zheng CH, Ng TY, Zhang L, Shiu CK, Wang HQ. 2011. Tumor classification based on non-negative matrix factorization using gene expression data. *IEEE Transactions on Nanobioscience* **10(2)**:86–93 DOI 10.1109/TNB.2011.2144998.

Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH. 2017. Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8(1)**:1–12 DOI 10.1038/s41467-016-0009-6.

Zhu X, Ching T, Pan X, Weissman SM, Garmire L. 2017. Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. *PeerJ* **5**:e2888 DOI 10.7717/peerj.2888.

Zitnik M, Zupan B. 2012. Nimfa: a python library for nonnegative matrix factorization. *Journal of Machine Learning Research* **13**:849–853.