A robust semi-supervised NMF model for single cell RNA-seq data

Peng Wu, Mo An, Hai-Ren Zou, Cai-Ying Zhong, Wei Wang and Chang-Peng Wu

Department of Neurosurgery, The People's Hospital of Longhua District, Shenzhen, Guangdong Province, China

ABSTRACT

Background. Single-cell RNA-sequencing (scRNA-seq) technology is a powerful tool to study organism from a single cell perspective and explore the heterogeneity between cells. Clustering is a fundamental step in scRNA-seq data analysis and it is the key to understand cell function and constitutes the basis of other advanced analysis. Nonnegative Matrix Factorization (NMF) has been widely used in clustering analysis of transcriptome data and achieved good performance. However, the existing NMF model is unsupervised and ignores known gene functions in the process of clustering. Knowledges of cell markers genes (genes that only express in specific cells) in human and model organisms have been accumulated a lot, such as the Molecular Signatures Database (MSigDB), which can be used as prior information in the clustering analysis of scRNA-seq data. Because the same kind of cells is likely to have similar biological functions and specific gene expression patterns, the marker genes of cells can be utilized as prior knowledge in the clustering analysis.

Methods. We propose a robust and semi-supervised NMF (rssNMF) model, which introduces a new variable to absorb noises of data and incorporates marker genes as prior information into a graph regularization term. We use rssNMF to solve the clustering problem of scRNA-seq data.

Results. Twelve scRNA-seq datasets with true labels are used to test the model performance and the results illustrate that our model outperforms original NMF and other common methods such as KMeans and Hierarchical Clustering. Biological significance analysis shows that rssNMF can identify key subclasses and latent biological processes. To our knowledge, this study is the first method that incorporates prior knowledge into the clustering analysis of scRNA-seq data.

Subjects Bioinformatics, Molecular Biology, Computational Science
Keywords Semi-supervised, NMF model , Single cell RNA-seq

Submitted 27 April 2020
Accepted 13 September 2020
Published 16 October 2020

Corresponding author
Chang-Peng Wu,
changpeng_wu@163.com

Academic editor
Gökhan Karakülah

Additional Information and Declarations can be found on page 14

DOI **10.7717/peerj.10091**

INTRODUCTION

Single cell RNA-seq (scRNA-seq) is a powerful tool enabling the transcriptional profiles at cellular resolution, comparing with "bulk" RNA-seq which can only measure the average gene expression among a group of cells. Compared with traditional high density microarray in single-cell transcriptomes, scRNA-seq has an obvious advantage that it can profile all transcripts in one cell rather than only detecting known genes by complement sequence probes. Besides, by scRNA-seq we can assess the heterogeneity between cells and identify

the hidden biological process, such as embryonic development and the origin of cancer cells (*Tirosh et al., 2016*; *Zeisel et al., 2015*).

Downstream analysis is the most important step in the workflow of scRNA-seq analysis and necessary for solving specific biological question. Clustering plays a fundamental and important role in many downstream analysis methods since it has a substantial impact on the outcome. There are considerable clustering algorithms for scRNA-seq and most of them can be applied to any type of data, such as KMeans and Hierarchical Clustering. However, most of these methods are unsupervised and does not consider known biological knowledge such as cell marker genes. In this paper, we proposed a semi-supervised NMF model for the clustering analysis of scRNA-seq data, which incorporates cell marker information and significantly improves the accuracy of clustering analysis.

Related work

Nonnegative matrix factorization (NMF) is an effective method for unsupervised clustering analysis of gene expression data. Given a nonnegative matrix X of size $m \times n$, NMF aims to find two non-negative matrices W and H such that:

$$X \approx WH$$

Where $W \in R^{m \times k}$ is a basis matrix, $H \in R^{k \times n}$ is a coefficient matrix. The solution to this problem can be obtained by solving the following optimization problem:

$$\min_{W,H} O(W,H) = min_{W,H} \frac{1}{2} ||X - WH||_F^2$$

$||.||_F$ means Frobenius norm and W and H satisfy nonnegative constrains. Since many practical problems in data mining, pattern recognition and machine learning require non-negativity constraints and appropriate low dimensional representation of original data, NMF has been successfully applied to these fields, which obtains the parts-based representation as well as enhancing the interpretability of the data (*Berry et al., 2007*; *Chagoyen et al., 2006*; *Lee & Seung, 1999*). In bioinformatics, it has been used to extract meaningful features that belongs to different cell types from microarray and scRNA-seq data and identify mRNA isoforms (*Brunet et al., 2004*; *Shao & Hofer, 2017*; *Ye & Li, 2016*).

Different variants of NMF has been put forward such as sparse NMF (SNMF) and discriminant NMF (DNMF) for microarray and RNA-seq data (*Jia et al., 2015*; *Kim & Park, 2007*). SNMF introduces a regularization term on W or H to control the degree of sparsity and generate sparser representation. DNMF incorporates Fisher's discriminant criterion in the coefficient matrix by maximizing the distance among any samples from different classes meanwhile minimizing the dispersion between any pair of samples in the same class. The DNMF requires discriminant information to construct the objective function and has been applied in various scenes such as face recognition and facial expression recognition. Yet in most cases of scRNA-seq, we cannot know the exact class information. Moreover, none of these methods consider the technical factors including amplification, library size differences and dropouts (*Buettner et al., 2015*; *Kharchenko, Silberstein & Scadden, 2014*). In addition to the technical factors, scRNA-seq data exhibit high cell-to-cell variability in gene expression, which impeding the analysis. Overall, the

technical effects and inherent variability in scRNA-seq introduce substantial noise and may corrupt the analysis of underlying biological process.

In this paper, we propose a robust semi-supervised NMF model that is robust to noises and uses marker information of cells as prior knowledge. We already have prior knowledge about the marker genes of different cells (such as MSigDB database; *Liberzon et al., 2011*), for instance, metastatic melanoma is a mixture of tumor cells and a variety of normal cell including T-cells, B-cells, macrophages and NK cells; thus, we can know the prior marker genes of different groups. We will show that incorporating such information into the NMF model, the clustering accuracy can be increased significantly.

MATERIALS & METHODS

Original NMF algorithm

The original NMF algorithm was introduced by Lee and Seung and the objective function is non-increasing if follows the multiplicative update rules (*Lee & Seung, 2001*):

$$W_{ij} \longleftarrow W_{ij}\frac{(XH^T)_{ij}}{(WHH^T)_{ij}} H_{ij} \longleftarrow H_{ij}\frac{(W^TX)_{ij}}{(W^TWH)_{ij}}$$

The author proved that by repeating iteration of the update rules is guaranteed to converge to a locally optimal matrix factorization. By using the non-negative constrains, NMF can learn a low dimensional representation of the original data.

Roubst NMF algorithm

Robust NMF (rNMF) first proposed by Kong and Ding and was designed to handle outliers and noises that original NMF fail to cope with (*Kong, Ding & Huang, 2011*). The main difference of rNMF is using $L_{2,1}$ norm loss function instead of Frobenius norm which is defined as follow:

$$||X - WH||_{2,1} = \sum_{j=1}^{n}\sqrt{\sum_{h=1}^{m}(X - WH)_{hj}^2} = \sum_{j=1}^{n}||x_j - Wh_j||$$

In this formulation, the error of each data point is $||x_j - Wh_j||$ rather than squared, thus the errors caused by outliers and noises do not dominate the objective function compared with the squared one. Different from the revision on the objective function, Zhang proposed a method which introduce an error matrix $S \in \mathbb{R}^{n\times m}$ to capture the errors and handle the extreme data points (*Zhang et al., 2011*):

$$\min_{W,H,S}||X - WH - S||_F^2$$
$$s.t. W \geq 0, H \geq 0, ||S||_0 \leq \upsilon$$

Where υ is the parameter that specifies the maximum number of nonzero elements in S.

Robust semi-supervised NMF algorithm

Motivated by robust NMF formulation, we adopt similar method to deal with the noises in scRNA-seq data under the assumption that the data matrix may be corrupted by noises and the noises are sparse (we use rNMF to represent this model below). In addition, we

incorporate the prior information into a graph regularization term to maintain the intrinsic geometrical and discriminating structure of the data in the parts-based representation:

$$\min_{W,H,S} ||X - WH - S||_F^2 + \alpha||S||_1 + \beta Tr\left(HLH^T\right)$$
$$s.t. W \geq 0, H \geq 0$$

Where Tr(.) denotes the trace of a matrix and α, β is the regularization parameters. Since the $\mathfrak{l}_0 - norm$ is difficult to solve, we replace the $\mathfrak{l}_0 - norm$ constraint with a $\mathfrak{l}_1 - norm$ regularizer, which is a common technique for sparse solution. The second graph regularization term maintains the local geometrical structure of the original data matrix X in the low dimensional representation, i.e., the matrix H. Briefly, if two data points x_i and x_j are close to each other in matrix X, then h_i and h_j, points in the low dimensional representation matrix H, are also close to each other. This property is usually referred to as *local invariance assumption* (*Belkin & Niyogi, 2002*; *Cai, Wang & He, 2009*) and has been applied in the development of dimensionality reduction algorithms and semi-supervised learning algorithms (*Belkin & Niyogi, 2002*; *Zhu & Lafferty, 2005*).

Chung et al. demonstrate that the local geometrical structure can be modeled by a nearest neighbor graph on a scatter of data points (*Chung & Graham, 1997*). For each data point x_i in X, we need to find its neighbors and put edges between x_i and its neighbors. In our scRNA-seq problem, we suppose that all the cells are the nodes in a graph, but different group of cells have different weighting. We use marker genes of cell group to construct the weight matrix Q, which can be defined by various ways. In this paper, we chose the heat kernel weighting and the weighting between x_i and x_j are:

$$Q_{ij} = e^{-\frac{||x_i - x_j||^2}{\sigma}}$$

where σ is parameter that controls the weighting and we set $\sigma = 1$ in the experiment. Note that only the rows (genes in rows and samples in columns for matrix X) that we select as cell markers of X will be used to compute the weighting. As shown in Fig. 1, the expression matrix X includes five groups and each group has its own highly-expressed genes (yellow rectangular block). The rows which are the markers of different groups are selected to construct the weight matrix.

We use Euclidean distance to measure the dissimilarity in the low dimensional representation of X:

$$D(h_i, h_j) = ||h_i - h_j||^2$$

then the local geometrical structure preserving criterion can be written as:

$$\begin{aligned}\Re &= \sum_{i,j=1}^{n} ||h_i - h_j||^2 Q_{ij} \\ &= \sum_{i=1}^{n} h_i^T h_i D_{ii} - \sum_{i,j=1}^{n} h_i^T h_j Q_{ij} \\ &= Tr(HDH^T) - Tr(HQH^T) = Tr(HLH^T)\end{aligned}$$

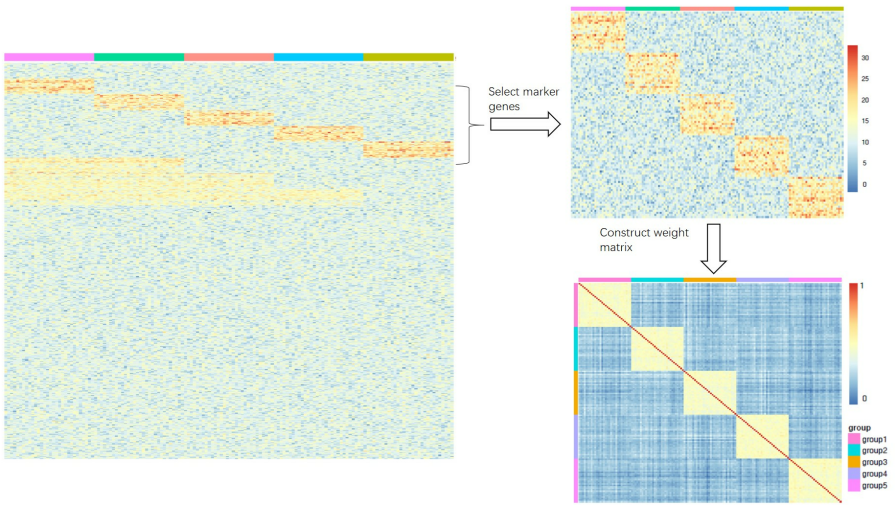

Figure 1 **(A-B) An illustration of how to construct the weight matrix Q (C).** (A) The heatmap is an ideal simulated gene expression matrix X; (B) the heatmap is part of the matrix X that only selects rows of marker genes.

Full-size DOI: 10.7717/peerj.10091/fig-1

Where D is a diagonal matrix and $D_{ii} = \sum_j Q_{ij}$. L is called graph Laplacian and $L = D - Q$. We can see that by minimizing $\Re$, if data point x_i and x_j are close (Q_{ij} is big), h_i and h_j will also be close to each other, so the distance relation between points in original data matrix X can be preserved in low dimensional matrix H. This technique was first designed by Cai et al. and introduced to solve the embedded structure problem, which assume that the data is usually sampled from a low dimensional manifold embedded in a high dimensional ambient space (*Cai et al., 2011*). We show that by encoding the geometrical information in the NMF model, the clustering accuracy can be greatly improved.

Optimization algorithm for robust semi-supervised NMF

In this section, we derive the iterative multiplicative update rules for the robust semi-supervised NMF based on the coordinate descent method (*Sha et al., 2007*). The objective function is not convex for W, H and S together, so it is unrealistic to expect an algorithm to find the global minima. It is natural to optimize W, H and S separately since the objective function is convex while holding the other two variables as constant.

The objective function of rssNMF can be rewritten as:

$$\begin{aligned}
\mathfrak{O} &= ||X - WH - S||_F^2 + \alpha||S||_1 + \beta Tr\left(HLH^T\right) \\
&= Tr((X - WH - S)(X - WH - S)^T) + \beta Tr\left(HLH^T\right) + \alpha||S||_1 \\
&= Tr(XX^T) - 2Tr(XH^TW^T) - 2Tr(XS^T) + 2Tr(WHS^T) + Tr(WHH^TW^T) \\
&\qquad + Tr(SS^T) + \beta Tr\left(HLH^T\right) + \alpha||S||_1
\end{aligned}$$

After introducing the Lagrange multiplier Ψ and Φ for the constrains, the Lagrange function $\mathfrak{L}$ is stated as

$$\mathfrak{L} = Tr(XX^T) - 2Tr(XH^TW^T) - 2Tr(XS^T) + 2Tr(WHS^T) + Tr(WHH^TW^T)$$

$$+Tr(SS^T)+\beta Tr\left(HLH^T\right)+\alpha||S||_1+Tr(\Psi W)+Tr(\Phi H)$$

and the partial derivatives of $\mathfrak{L}$ with respect to W and H are:

$$\frac{\partial \mathfrak{L}}{\partial \mathrm{W}}=-2XH^T+2SH^T+2WHH^T+\Psi$$
$$\frac{\partial \mathfrak{L}}{\partial \mathrm{H}}=-2W^TX+2W^TS+2W^TWH+2\beta HD-2\beta HQ+\Phi$$

By KKT conditions $\psi_{hi}w_{hi}=0 \text{and} \phi_{ij}h_{ij}=0$, we can obtain:

$$-(XH^T)_{ij}w_{ij}+(SH^T)_{ij}w_{ij}+(WHH^T)_{ij}w_{ij}=0$$
$$-(W^TX)_{ij}h_{ij}+(W^TS)_{ij}h_{ij}+(W^TWH)_{ij}h_{ij}+(\beta HD)_{ij}h_{ij}-(\beta HQ)_{ij}h_{ij}=0$$

Then the updating rules for W and H are:

$$W_{ij}\longleftarrow W_{ij}\frac{(XH^T)_{ij}}{(WHH^T+SH^T)_{ij}} \tag{1}$$

$$H_{ij}\longleftarrow H_{ij}\frac{(W^TX+\beta HQ)_{ij}}{(W^TWH+W^TS+\beta HD)_{ij}} \tag{2}$$

For fix W and H, we update S via the *soft-thresholding operator* (*Hale, Yin & Zhang, 2008*) and the optimization problem for S is:

$$\min_S||X-WH-S||_F^2+\alpha||S||_1$$

Theorem 1: Define the *soft-thresholding operator* as below:

$$T_\upsilon(z)=\begin{cases}z-\upsilon, if\ z>\upsilon\\ z+\upsilon, if\ z<-\upsilon\\ 0, otherwise\end{cases}$$

Where $z\in\mathbb{R}$ and $\upsilon>0$ and this operator can also be applied to vectors or matrices by element-wise operation. For the following $\mathfrak{l}_1-norm$ problem:

$$\min_v\frac{1}{2}||x-v||_F^2+\alpha||v||_1$$

the unique solution of v is given by $T_\alpha(\mathrm{x})$ (*Hale, Yin & Zhang, 2008*). Similarly, we can get the update rule for S:

$$\mathrm{S}\longleftarrow T_{\frac{\alpha}{2}}(X-WH) \tag{3}$$

Based on above analysis, our algorithm for solving rssNMF is presented in Algorithm 1.

Algorithm 1 Multiplicative Updating Rules for Semi-Supervised NMF

Input: Gene expression matrix $X \in R^{m \times n}$, graph Laplacian L, clustering number k, parameter α and β

1: Initial W, H and S

2: **Repeat**

3: Update W using Eq. (1)

4: Update H using Eq. (2)

5: Update S using Eq. (3)

6: **until** a predefined stopping criterion is satisfied

Output: $w \in R^{m \times k}$, $H \in R^{k \times n}$

The proof of convergence analysis of our algorithm essentially follows the idea of graph NMF and ensures that the objective function of rssNMF is nonincreasing under the updating rules in Eqs. (1), (2) and (3) and could converge to a stationary point (*Cai et al., 2010*). Since the idea is similar to the method in the paper (*Cai et al., 2010*) so we did not prove it here.

RESULTS

Dataset

Cell markers, usually referring to surface molecules on cell membrane, are different in different kinds of cells. Surface molecules that appear only on a particular type of cell are called cell markers and often used for cell type identification. For example, cancer cell specific markers often appear on the surface of cancer cells and are utilized as targets for anticancer drugs. Cell markers have been a hotspot in molecular biology study and researchers have accumulated lots of marker data for different kinds of cells, such as MSigDB (*Liberzon et al., 2011*). Here, we extend the concept of cell markers. We use ''cell marker genes'' to represent those genes that can distinguish different types of cell, which are usually only expressed in specific type of cells or relatively highly expressed, such as differentially expressed genes. In single-cell RNA-sequencing, we already know where the sequencing samples come from and then we can have a modest degree of prior knowledge. That is, we have known what kinds of cells are in the experimental samples although there are still many kinds of cells remaining unknown. For instance, when the cancer tissue was sequenced, the samples contained a large number of normal cells in addition to cancer cells, such as lymphocytes, myeloid populations and cancer-associated fibroblasts. Taking melanoma as an example, melanoma is often known as malignant melanoma, which is a kind of cancer that stems from the pigment-containing cells known as melanocytes. Melanoma is a mixture of tumor cells and a variety of normal cells including T-cells, B-cells, macrophages and NK cells. This feature has been found in many studies, thus we can know the prior marker genes of different groups of cells in advance. We will show that incorporating such information into the NMF model, the clustering accuracy can be increased significantly.

Table 1 Published ten scRNA-seq datasets used to test rssNMF model. All the datasets are scRNA-seq data of human or mouse embryos.

Dataset	Units	GSE/ArrayExpress Number	Number of cells	Species	Number of Clusters
Biase	FPKM	GSE57249	56	Mouse	5
Goolam	CPM	E-MTAB-3321	124	Mouse	5
Yan	RPKM	GSE36552	124	Human	9
Shin	RPKM	GSE71485	256	Mouse	10
Deng	RPKM	GSE45719	259	Mouse	10
Leng	Normalized counts	GSE64016	460	Human	4
Kowalczyk	TPM	GSE59114	564	Mouse	8
Camp	FPKM	GSE75140	734	Human	9
Chu_1	TPM	GSE75748	758	Human	6
Chu_2	TPM	GSE75748	1,018	Human	7
Tasic	RPKM	GSE71585	71,585	Mouse	7
Zeisel	Counts	GSE60361	60,361	Mouse	8

Notes.
FPKM, fragments per kilobase of transcript per million mapped reads; RPKM, reads per kilobase of transcript per million mapped reads; CPM, counts per million mapped reads.

Twelve scRNA-seq datasets are used in the experiment. All the datasets are normalized expression level (FPKM, RPKM or CPM) or counts and number of cells range from 56 to 3005 (Table 1). Almost allthe datasets except dataset "Zeisel" have true labels since they are collected from different time points during embryonic development and we provide the accession number in Table 1. The labels of dataset "Zeisel" comes from computation and can be seen as a silver dataset.

Compared algorithms

To benchmark rssNMF, six clustering algorithms are used: NMF, rNMF, ssNMF, KMeans, Hierarchical Clustering and SC3.

NMF: F-norm formulation is adopted to cluster scRNA-seq data.

rNMF: similar to the proposed rssNMF but without the graph regularization term:

$$\min_{W,H,S} ||X - WH - S||_F^2 + \alpha ||S||_1$$
$$s.t. W \geq 0, H \geq 0$$

ssNMF: ssNMF removes the matrix S which copes with noises and outliers but keep the graph regularization term:

$$\min_{W,H} ||X - WH||_F^2 + \beta Tr\left(HLH^T\right)$$
$$s.t. W \geq 0, H \geq 0$$

KMeans: a canonical distance-based iterative algorithm. Euclidean distance is used in the experiment.

Hierarchical clustering: we use a division-based algorithm which initially starts with all observations in a single cluster and divide samples until each cluster only contain one observation. Euclidean distance is used in the experiment and "ward.D2" method was used in R hclust() function.

SC3: consensus clustering algorithm for single-cell RNA-seq data, which is a benchmark in comparison to various clustering methods in scRNA-seq clustering analysis (*Duò, Robinson & Soneson, 2018*; *Kiselev et al., 2017*).

After factorization, we use KMeans to cluster matrix H and obtain the clustering results. The other common method is categorizing sample j directly by the largest coefficient in column h_j.

To evaluate the clustering performance of robust semi-supervised NMF, we use adjusted Rand index (ARI, Supplementary File) which ranges from -1 to $+1$. Since cell labels are available so ARI can be calculated to measure the similarity between two data clustering results.

Prior genes selection and parameter setting

For all the datasets, we use differentially expressed genes as marker genes. For all the clusters in each dataset, we use R package "Deseq2" to identify the differentially expressed genes and select 20 differentially expressed genes with largest variance as cell marker genes for each group to construct weighting matrix. We set the same parameter value for rNMF, ssNMF and rssNMF and results are obtained by running the six methods 30 times on each dataset. For the factoring matrices W and H, we use standard normal distribution for random initialization.

Experimental results

Table 2 shows the clustering results on twelve datasets and we can see that rssNMF performs better than the most methods on almost all datasets. Here, we consider a basic circumstance that we only know one group of cells in the experiment and shows the results that one group of genes is added. All these datasets can be downloaded on NCBI Gene Expression Omnibus (GEO). The final results are evaluated by taking the average value of ARI over 30 runs. Compared with SC3, rssNMF gets better performance over seven datasets. We notice that the overall ARI on these seven datasets is relatively low (smaller than 0.4) which indicates that the cells hard to discriminate. Under such circumstance, the prior information can help improve the clustering accuracy than purely unsupervised clustering algorithm.

Our rssNMF model has two critical parameters: regularization parameter α and β. Figure 2 and Fig. S1 shows the how the average performance of rssNMF varies with the parameters α and β. As we can see in the Fig. 2, the performance of rssNMF is quite stable with respect to the parameter α on all four datasets. If $\alpha/2 > max_{ij}(X - WH)_{ij}$, all the elements in S will be zero, thus when the parameter α is large enough, rNMF will be equivalent to the original NMF.

The parameter β controls the weights of the graph regularization term and we test different weights range from 2 to 20,000 (Fig. S1). The performances are stable when β is smaller than 200 in the four datasets. When β is too large, clustering accuracy decreases a lot for the graph regularization term dominates the whole formulation and covers the main information.

To investigate how the prior information influences the clustering accuracy, we test the performance of rssNMF with respect to different number of group information. For each

Table 2 Benchmarking of rssNMF against other clustering method. All the algorithms were applied 50 times to each dataset. Parameter α for rNMF and rssNMF: 2. Parameter β for rssNMF: 2. Prior information: for each dataset, we randomly select one cluster and use 20 marker genes of the selected cluster to construct the weight matrix.

Dataset	KMeans	HC	NMF	SC3	rNMF	ssNMF	rssNMF
Biase	0.712	0.761	0.774	0.844	0.806	0.796	0.862
Goolam	0.304	0.310	0.387	0.731	0.43	0.642	0.657
Yan	0.375	0.570	0.533	0.805	0.572	0.675	0.710
Shin	0.167	0.217	0.282	0.366	0.282	0.327	0.370
Deng	0.42	0.399	0.466	0.775	0.52	0.547	0.682
Leng	0.057	0.009	0.112	0.179	0.14	0.165	0.213
Kowalczyk	0.182	0.176	0.269	0.307	0.304	0.293	0.365
Camp	0.232	0.225	0.274	0.327	0.3	0.297	0.305
Chu_1	0.177	0.199	0.22	0.205	0.241	0.326	0.369
Chu_2	0.204	0.242	0.314	0.312	0.314	0.322	0.357
Tasic	0.51	0.284	0.705	0.822	0.711	0.791	0.790
Zeisel	−4.97E−05	−9.36E−04	2.43E−03	−5.60E−04	2.65E−03	0.007	0.014

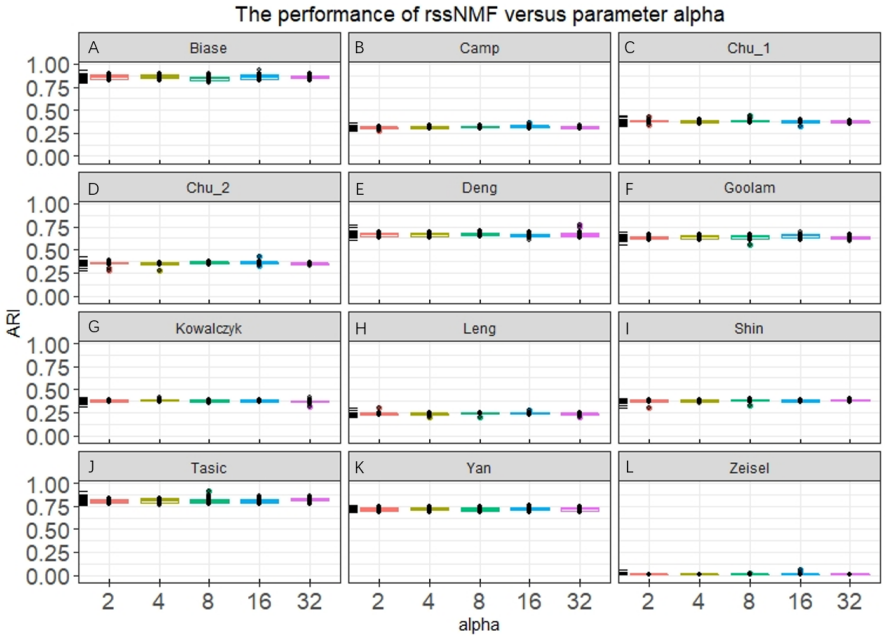

Figure 2 Performance of rssNMF versus parameter. The rssNMF is stable with respect to the parameter—and achieve good performance varies from 2 to 32.

Full-size DOI: 10.7717/peerj.10091/fig-2

group in one dataset, 20 marker genes are selected as prior information to compute the weighting matrix. Figure S2 shows how the ARI changes as the group marker information increases (20 markers genes of one group, 40 markers gene for two groups and so on). We can see that as the number of group information increases, the performance fluctuates but clustering accuracy is higher than only adding one group information.

To evaluate the clustering stability, we calculate the consensus matrix for NMF, rNMF and rssNMF over 30 runs. A consensus matrix M is a $n \times n$ matrix that stores, for each pair of samples, the proportion of clustering runs in which two samples are clustered together. A consensus matrix can be obtained by taking the average over the connectivity matrices for all runs. The connectivity matrix C is also an $n \times n$ matrix and defined based on a single run:

$$C_{ij} = \begin{cases} 1, & i \text{ and } j \text{ are in the same clulster} \\ 0, & i \text{ and } j \text{ are not in the same cluster} \end{cases}$$

So, if sample i and j are in the same cluster for 30 runs, then M_{ij} will equal to 1, corresponding to red (0 corresponds to blue). Figure 3 present the factorizing matrices W, H and consensus matrix taking the average over 30 runs and the consensus matrix from NMF, rNMF and rssNMF on the dataset Yan. The color bar represents different group of cells or genes. More specifically, we assign genes to different cell groups according to the gene score, which is described in section Biological Significance Analysis section. The values in each row of W an column of H have been normalized between 0 and 1. We can see that all of the three methods have a stable clustering result on most samples and only small part of samples are clustered into multiple clusters in dataset Yan. Generally, rssNMF has the highest clustering accuracy and better stability than the other two methods.

Biological significance analysis

In this section, we choose dataset Yan for biological significance analysis and dataset Yan are human embryonic stem cells including zygote, 2-cell, 4-cell, 8-cell, morula and blastocyst. In Fig. 3, each column of matrix W defines a metagene (biological procesess or pathways) or a metasample (cell cluster) then entry w_{ij} can be regarded as the coefficient of gene I in metagene or metasample j (*Brunet et al., 2004*). Correspondingly, a column vector h_j represents the expression contribution of sample j to k biological processes, and H_{ij} can be seen as the contribution of metagene or metasample i in sample j. The class of samples is denoted by column annotation bar. We can see that for some samples (columns of H), it is hard to visually determine which cluster it belongs to since the depth of color is close in different clusters. Another obvious feature is that compared with NMF and rNMF, the pattern of H in rssNMF is clearer and more concentrated, which makes the classification characteristics of samples more explicit. We suppose that the addition of prior information makes the differences between different clusters more obvious, thus achieving better classification results.

For the basis matrix W, the ith row of W represents the contribution of gene i to all metagenes so the expression level of a specific gene in one cell is determined by the linear combination of its contribution to all biological processes. From Fig. S3 we can see that a gene can participate in multiple biological processes, but it is more reasonable to focus on the process with the largest value. We define a gene score to determine cluster-specific genes and the assignment of cluster-specific genes is shown in row annotation bar of matrix

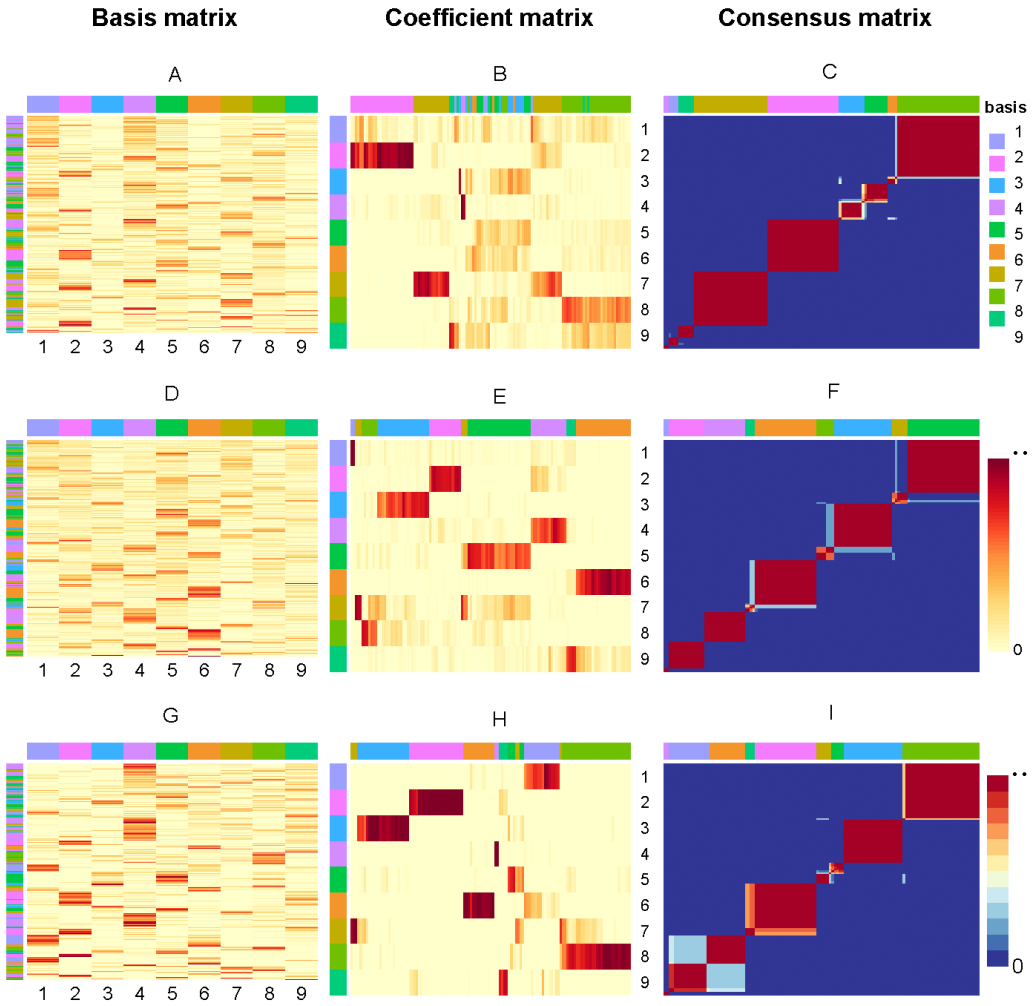

Figure 3 Factorizing matrices W (basis matrix: A, D, G), H (coefficient matrix: B, E, H) and consensus matrix (C, F, I) respectively obtained from NMF (A, B, C), rNMF (D, E, F) and rssNMF (G, H, I) for dataset Yan with 124 cells and nine clusters. The annotation color bar denotes nine clusters. The rows annotation of W and columns of H indicate the assignment of genes and samples for clusters. The paramete $\alpha = 2$ for rNMF and rssNMF and $\beta = 2$ for rssNMF.

Full-size DOI: 10.7717/peerj.10091/fig-3

W. The gene score for the ith gene is:

$$Gene\_score(i) = 1 + \frac{1}{\log 2(k)} \sum_{j=1}^{k} p(i,j) \log 2\left(p(i,j)\right),$$

where $p(i,\Omega)$ is the probability that the ith gene contributes to cluster Ω, i.e., $p(i,\Omega) = W_{i\Omega} / \sum_{j=1}^{k} W_{ij}$. The gene score is a real value and ranges from $[0,1]$. The higher gene score is, the more cluster-specific the corresponding gene. Simply we can think that a gene is highly-expressed if the gene has high Gene_score in a specific cluster over other clusters.

As shown in the A, D and G which are the basis matrix W from NMF, rNMF and rssNMF in the Fig. 3, we assign all the genes to a specific cluster using gene score by

different colors as shown in the left color bar. For the basis matrix W, the rows indicate genes and the columns indicate clusters andthe depth of the color in the matrix indicates the size of the value in W. Deep color represents a larger value and we can assign the cluster genes visually. After classifying each gene, we need to validate the biological significance of feature genes detected in different cluster. Table S1 present the results of enrichment analysis for cluster-specific genes.

We use the KEGG database to investigate the functional genes selected for all clusters. KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. Through functional enrichment analysis of genes in each cell cluster, we can understand what biological processes are involved in the cell cluster and further infer the role of the cluster in the biological microenvironment.

Embryonic stem cells are highly undifferentiated cells. It has developmental omnipotence and can differentiate all tissues and organs of adult animals, including germ cells. The research and utilization of ES cells is one of the core topics in the field of bioengineering. The enrichment results (Table S1) show that embryonic stem cells are in an active metabolic and proliferative state, such as Ribosome pathway in cluster 3, 5, 6, 7 and 8. Ribosomes are organelles that are responsible for making proteins and are widespread. The manufacture of ribosomes requires hundreds of cytokines that are not found in mature particles. In the absence of these factors, ribosome production will stagnate. Once ribosome production is stagnant, cell growth is terminated even under optimal growth conditions. Besides, various metabolic pathways such as TCA cycle, oxidative Phosphorylation and cholesterol metabolism also occur in multiple clusters, indicating that these clusters are in an active proliferation state. In addition, we can infer the differentiation path of stem cells from the results of enrichment. Cluster 6 contains two distinctive pathways, Cardiac muscle contraction and Retrograde endocannabinoid signaling. Both Cardiac muscle contraction and Retrograde endocannabinoid signaling are involved in nerve signal transduction and are usually active in the nervous system and cardiac conduction system. Endogenous cannabinoids (endocannabinoids) serve as retrograde messengers at synapses in various regions of the brain. Therefore, it can be inferred that cells in cluster 6 are differentiating toward the ectoderm and will form epidermis and nervous system in the future. Thermogenesis pathway appears in cluster 5, 6 and 7 and this pathway usually activates in brown adipose tissues. Brown adipose tissues originate from mesenchymal stem cells which differentiate from ectoderm. This is consistent with our view that cluster 7 belongs to ectoderm cell groups.

DISCUSSION

Considering data noises and absorbing prior knowledge at the same time in single-cell RNA-sequencing data, the clustering accuracy is significantly improved compared with NMF, which is an unsupervised clustering method. Compared to NMF, our model has higher clustering accuracy and can discover hidden structures in the data after adding the prior

information. Our method also achieves better performance in part of the datasets compared to other unsupervised clustering method such as SC3, which is seen as a benchmark in scRNA-seq clustering analysis. We notice that rssNMF gets higher accuracy upon datasets that most unsupervised methods perform bad such as datasets Leng, Kowalczyk and Zeisel. We supposed that in the situation that cells are much alike each other, semi-supervised approach can get better performance over unsupervised clustering method. In addition, our approach can be used in a wider range of areas, such as medical text information mining and biological network motif identification, provided that there are noises in the data and that the characteristics of some of the samples are known. However, our method still has some limitations. Firstly, the objective function of rssNMF is based on Frobenius norm, without considering other objective functions, such as KL divergence and $L_{2,1}$ norm. Some studies have found that KL divergence can achieve higher accuracy in clustering analysis of gene expression data than Frobenius norm. Therefore, a semi-supervised NMF model based on different objective functions deserves further exploration. Second, the choice of weighting matrix in rssNMF is arbitrary and we use heat kernel weighting in our model. Other weight matrix such as 0-1 weighting, and Dot-Product weighting have not been tried. In general, the choice of weighting matrix is empirical and depends on the dataset. Third, rssNMF does not take into account the statistical dependency between latent variables. In rssNMF, the gene expression level of single cell is determined by the linear combination of potential biological processes (metagenes), and the dependence relationship between these biological processes is neglected. In fact, different biological processes are interrelated and regulated with each other. In Fig. S3, we can see that the CELL CYCLE intersects the MAPK signaling pathway, Apoptosis and Ubiquitous mediated proteolysis, and NMF cannot identify the relationship between them.

CONCLUSIONS

We present a novel robust semi-supervised NMF model called rssNMF for scRNA-seq data and the model remarkably improve clustering accuracy when the cell marker information for one group or more groups are available. Compared to NMF and other common clustering algorithms, our model has higher clustering accuracy and can discover hidden structures in the data after adding the prior information. Our model also outperforms SC3 in part of the datasets while the prior information is available. What's more, the model is robust to noises and outliers compared with standard NMF model. Our model can also be applied into other clustering tasks as long as feature information of some samples are known, such as electronic health records, single cell methylation sequencing and proteomics.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Peng Wu conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Mo An analyzed the data, prepared figures and/or tables, and approved the final draft.
- Hai-ren Zou performed the experiments, prepared figures and/or tables, and approved the final draft.
- Cai-Ying Zhong and Wei Wang analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Chang-Peng Wu conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

R code of NMF and NMF related variants are available as a Supplemental File. The experimental RNA-seq data are available at GEO and all the accession numbers are available in Table 1.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10091#supplemental-information.

REFERENCES

Belkin M, Niyogi P. 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. Cambridge: MIT Press, 585–591.

Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ. 2007. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis* **52**:155–173 DOI 10.1016/j.csda.2006.11.006.

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**:4164–4169 DOI 10.1073/pnas.0308531101.

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O. 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33**:155–160 DOI 10.1038/nbt.3102.

Cai D, He X, Han J, Huang TS. 2010. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**:1548–1560.

Cai D, He X, Han J, Huang TS. 2011. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**:1548–1560 DOI 10.1109/TPAMI.2010.231.

Cai D, Wang X, He X. 2009. Probabilistic dyadic data analysis with local and global consistency. In: *Proceedings of the 26th annual international conference on machine learning*. ACM, 105–112.

Chagoyen M, Carmona-Saez P, Shatkay H, Carazo JM, Pascual-Montano A. 2006. Discovering semantic features in the literature: a foundation for building functional associations. *BMC Bioinformatics* **7**:41 DOI 10.1186/1471-2105-7-41.

Chung FR, Graham FC. 1997. *Spectral graph theory*. Providence, Rhode Island: American Mathematical Soc.

Duò A, Robinson MD, Soneson C. 2018. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* **7**.

Hale ET, Yin W, Zhang Y. 2008. Fixed-point continuation for \ell_1-minimization: methodology and convergence. *SIAM Journal on Optimization* **19**:1107–1130 DOI 10.1137/070698920.

Jia Z, Zhang X, Guan N, Bo X, Barnes MR, Luo Z. 2015. Gene ranking of RNA-seq data via discriminant non-negative matrix factorization. *PLOS ONE* **10**:e013778 DOI 10.1371/journal.pone.0137782.

Kharchenko PV, Silberstein L, Scadden DT. 2014. Bayesian approach to single-cell differential expression analysis. *Nature Methods* **11(7)**:740–742 DOI 10.1038/nmeth.2967.

Kim H, Park H. 2007. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* **23**:1495–1502 DOI 10.1093/bioinformatics/btm134.

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

Kong D, Ding C, Huang H. 2011. Robust nonnegative matrix factorization using l21-norm. In: *Proceedings of the 20th ACM international conference on information and knowledge management*. ACM, 673–682.

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

Lee DD, Seung HS. 2001. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems* 556–562.

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**:1739–1740 DOI 10.1093/bioinformatics/btr260.

Sha F, Lin Y, Saul LK, Lee DD. 2007. Multiplicative updates for nonnegative quadratic programming. *Neural Computation* **19**:2004–2031 DOI 10.1162/neco.2007.19.8.2004.

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

Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, Fisher JM, Rodman C, Mount C, Filbin MG, Neftel C, Desai N, Nyman J, Izar B, Luo CC, Francis JM, Patel AA, Onozato ML, Riggi N, Livak KJ, Gennert D, Satija R, Nahed BV, Curry WT, Martuza RL, Mylvaganam R, Iafrate AJ, Frosch MP, Golub TR, Rivera MN, Getz G, Rozenblatt-Rosen O, Cahill DP, Monje M, Bernstein BE, Louis DN, Regev A, Suva ML. 2016. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**:309–313 DOI 10.1038/nature20123.

Ye Y, Li JJ. 2016. NMFP: a non-negative matrix factorization based preselection method to increase accuracy of identifying mRNA isoforms from RNA-seq data. *BMC Genomics: BioMed Central* **17(1)**:127–140.

Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C, Rolny C, Castelo-Branco G, Hjerling-Leffler J, Linnarsson S. 2015. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**:1138–1142 DOI 10.1126/science.aaa1934.

Zhang L, Chen Z, Zheng M, He X. 2011. Robust non-negative matrix factorization. *Frontiers of Electrical and Electronic Engineering in China* **6**:192–200 DOI 10.1007/s11460-011-0128-0.

Zhu X, Lafferty J. 2005. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In: *Proceedings of the 22nd international conference on Machine learning*. ACM, 1052–1059.