Computer Modeling in Engineering & Sciences CMES, vol.122 no.1, pp.377-397, 2020

FSPAM: A Feature Construction Method to Identifying Cell Populations in ScRNA-seq Data

Amin Einipour[1], Mohammad Mosleh[1, *] and Karim Ansari-Asl[1, 2]

Abstract: The emergence of single-cell RNA-sequencing (scRNA-seq) technology has introduced new information about the structure of cells, diseases, and their associated biological factors. One of the main uses of scRNA-seq is identifying cell populations, which sometimes leads to the detection of rare cell populations. However, the new method is still in its infancy and with its advantages comes computational challenges that are just beginning to address. An important tool in the analysis is dimensionality reduction, which transforms high dimensional data into a meaningful reduced subspace. The technique allows noise removal, visualization and compression of high-dimensional data. This paper presents a new dimensionality reduction approach where, during an unsupervised multistage process, a feature set including high valuable markers is created which can facilitate the isolation of cell populations. Our proposed method, called fusion of the Spearman and Pearson affinity matrices (FSPAM), is based on a graph-based Gaussian kernel. Use of the graph theory can be effective to overcome the challenge of the nonlinear relations between cellular markers in scRNA-seq data. Furthermore, with a proper fusion of the Pearson and Spearman correlation coefficient criteria, it extracts a set of the most important features in a new space. In fact, the FSPAM aggregates the various aspects of cell-to-cell similarity derived from the Pearson and Spearman metrics, and reveals new aspects of cell-to-cell similarity, which can be used to extract new features. The results of the identification of cell populations via k-means++ clustering method based on the features extracted from the FSPAM and different datasets of scRNA-seq suggested that the proposed method, regardless of the characteristics that govern each dataset, enjoys greater accuracy and better quality compared to previous methods.

Keywords: Single cell RNA sequencing, cell population, feature extraction, fusion.

1 Introduction

With the advent of the new generation of DNA sequencing method, known as next generation sequencing (NGS), the quantitative and qualitative knowledge of transcriptomes progressed remarkably, through which researchers were able to extract the gene expression of cells completely. The parallel and automatic nature of these new

[1] Department of Computer Engineering, Dezful Branch, Islamic Azad University, Dezful, Iran.

[2] Department of Electrical Engineering, Faculty of Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran.

[*] Corresponding Author: Mohammad Mosleh. Email: mosleh@iaud.ac.ir.

Received: 30 August 2019; Accepted: 12 November 2019.

processes such as RNA sequencing (RNA-seq) causes the production of millions of sequences concurrently, resulting in significant amplification of the operational power. Also, sequencing technologies with a high power have considerably reduced the costs of sequencing [Wang, Gerstein and Snyder (2010); Nagalakshmi, Waern and Snyder (2010)].

Single-cell RNA-sequencing (scRNA-seq) is known as a novel technology first presented in 2009 [Tang, Barbacioru, Wang et al. (2009)]. This method did not gain popularity until 2014, i.e., when new protocols were developed, and its sequencing costs diminished. ScRNA-seq technology measures the distribution of expression levels for every gene throughout the entire population of cells and allows new biological questions to be studied, where specific cellular changes in transcriptomes are important. For example, one can mention the identification of cell types, heterogeneity of cell responses, confirmation of gene expression, and deduction of gene regulatory networks across the cells. There are several protocols for using scRNA-seq such as SMART-seq2 [Picelli, Björklund, Faridani et al. (2013)], CELL-seq [Hashimshony, Wagner, Sher et al. (2012)], and Drop-seq [Macosko, Basu, Satija et al. (2015)].

In recent years, the use of scRNA-seq has allowed researchers to describe phenotypic heterogeneities observed in certain groups of cells and tissues through cell-by-cell indexing from transcriptome heterogeneity [Pouyan and Nourani (2017)]. One of the important uses of the results of scRNA-seq is identifying cell populations. In some cases, it results in the detection of rare and new cell subsets [Shalek, Satija, Adiconis et al. (2013); Buettner, Natarajan, Casale et al. (2015); Grün, Lyubimova, Kester et al. (2015); Nelson, Mould, Bikoff et al. (2016); Pellegrino, Sciambi, Yates et al. (2016)], which cannot be identified via previously known factors. It can further be employed in various areas such as cancer. For example, single-cell RNA-sequencing has been used for the identification of new cell subsets in the colon [Grün, Lyubimova, Kester et al. (2015)], fetus [Nelson, Mould, Bikoff et al. (2016)], cancer [Patel, Tirosh, Trombetta et al. (2014)], brain [Liu, Nowakowski, Pollen et al. (2016); Tasic, Menon, Nguyen et al. (2016)], pancreas [Segerstolpe, Palasantza, Eliasson et al. (2016); Wang, Schug, Won et al. (2016)], and immune cells [Villani, Satija, Reynolds et al. (2017)].

In spite of the hopes developed in this regard, there are challenges that complicate the analysis of scRNA-seq data. Some of these obstacles include the stochastic nature of the expression of genes, the existence of noise, dropout events, and high dimensions of these data. Nevertheless, in recent years, many attempts have been made to overcome these computational challenges.

The proposed method in this paper, which is based on a graph-based Gaussian kernel, extracts a set of high-quality features before clustering through a proper fusion of the Pearson and Spearman criteria in a new nonlinear space. In fact, the FSPAM aggregates the various aspects of cell-to-cell similarity derived from the Pearson and Spearman metrics, and reveals new aspects of cell-to-cell similarity, which can be used to extract valuable features. In summary, it can be said that, the proposed method can be used both for extracting a high-quality feature and for identifying accurate cell populations, can be used as a useful tool for analyzing and visualizing scRNA-seq data for bioinformatics researchers.

This paper is organized as follows: Section 2 provides a review of related work. Section 3 explains the details of the proposed method. In Section 4, experimental results are reported

clearly and, by comparing the results with other state-of-the-art methods, we evaluate the proposed method. Finally, the conclusion and discussion are presented in Section 5.

2 Related works

In most works related to the identification of cell populations in scRNA-seq data, attempts have been made to perform cell clustering by developing machine learning techniques. Partition clustering methods such as k-means and other distance-based clustering algorithms such as hierarchical clustering have been widely used for identifying cell populations in scRNA-seq datasets. For example, Jaitin et al. combined a hierarchical clustering method and probabilistic hybrid models to classify single-cells of different tissues [Jaitin, Kenigsberg, Keren-Shaul et al. (2014)].

Kiselev et al. [Kiselev, Kirschner, Schaub et al. (2017)] proposed a clustering method called single-cell consensus clustering (SC3), which integrates multiple cluster labels by a consensus approach and can improve cell type identification. SC3 combines all the different clustering outcomes into a consensus matrix that summarizes how often each pair of cells is located in the same cluster. The final result is determined by complete-linkage hierarchical clustering of the consensus matrix into k groups.

Žurauskienė et al. [Žurauskiene and Yau (2015)] presented a modified clustering method called pcaReduce for the scRNA-seq data which repeatedly combined the PCA with k-means to generate a hierarchical tree of cells. This method seeks to establish a connection between the reduced representations given by principal components analysis (PCA) and the number of resolvable cell types (clusters).

SINCERA package is another example employing hierarchical clustering, in which the Pearson correlation is used for the similarity criterion, while linkage mean is employed for the linkage method in default settings. This package presents a generally applicable analytic pipeline for processing scRNA-seq data from a whole organ or sorted cells [Guo, Wang, Potter et al. (2015)].

SNN-cliq method presented by Xu et al. [Xu and Su (2015)] uses the shared nearest neighbor (SNN) for defining similarity between data points (cell), which performs clustering according to an algorithm based on the graph theory. This method models data as an SNN graph, with nodes corresponding to data points and weighted edges reflecting the similarities between data points. It then finds the ultimate clustering solution by using graph-theoretic techniques to cluster the sparse SNN graph.

Pouyan et al. [Pouyan and Nourani (2016); Pouyan and Kostka (2018)] introduced methods called RAFSIN and RAFSIL, employing a random forest algorithm for clustering cell populations. RAFSIN uses random forests for identifying the dependence of cell markers and modeling cell populations based on the cell network concept. This cellular network helps to discover what types of cells exist in the tissue. RAFSIL method is also an approach based on the random forest for learning cell-to-cell similarities from scRNA-seq data. RAFSIL runs a two-stage method in which the features related to the scRNA-seq data are created after learning the similarities. This method is designed such that it can be adapted and developed, whereby the similarities obtained from RAFSIL can be used in projects associated with data analysis such as dimension reduction, visualization, and clustering.

Li et al. [Li, Zhang, Wong (2019)] introduced clustering methods based on evolutionary multiobjective. They proposed an evolutionary multiobjective ensemble pruning algorithm (EMEP) that addresses those realistic restrictions. The EMEP algorithm first applies the unsupervised dimensionality reduction to project data from the original high dimensions to low-dimensional subspaces; basic clustering algorithms are applied in those new subspaces to generate different clustering results to form cluster ensembles. Also, Li et al. [Li and Wong (2019)] provided a multiobjective evolutionary clustering based on adaptive non-negative matrix factorization (MCANMF) for multiobjective single-cell RNA-seq data clustering. Firstly, adaptive non-negative matrix factorization is proposed to decompose data for feature extraction. After that, a multiobjective clustering algorithm based on learning vector quantization is proposed to analyze single-cell RNA-seq data.

The mentioned methods, for precise the identification of cell populations, have used solutions for overcoming some computational challenges associated with scRNA-seq data. One of these challenges is the high dimension of this type of data, i.e., the sheer number of features (genes). The process of reducing the number of features and removing noise from data, which is the outcome of the data dimension reduction process, can significantly improve the ability to separate cell populations [Van Der Maaten, Postma and Van Den Herik (2009)]. To achieve this aim, various methods have been developed trying to visualize scRNA-seq data to identify cell populations through dimension reduction. Most of these methods use well-known tools such as PCA and t-SNE [Van Der Maaten and Hinton (2008)] for this aim. The PCA is considered a linear conversion method of data. Therefore, it may not be practical in many scRNA-seq datasets with a nonlinear nature where one cannot present the gene expression data as a linear combination of interrelationships between two cells.

One of the nonlinear techniques that is currently used is t-SNE coupled with Euclidean distance, which can unveil the global structure and obtain many local structures of data with large dimensions. For example, a method called viSNE has been presented for reducing the dimensions of scRNA-seq data, which operates based on t-SNE, and maps high dimension cytometry data to two dimensions while preserving the structure [Amir, Davis, Tadmor et al. (2013)]. Nevertheless, unlike PCA, t-SNE does not learn an explicit map between high and low dimension spaces. This suggests that the points that are close to each other in high dimension spaces will also be close to each other in the low dimensions obtained, while most global relations cannot be interpreted directly [Wagner, Regev and Yosef (2016)]. Note that, according to the developers, t-SNE is a global visualization tool, and not a dimension reduction method, which has not been designed for dimension reduction in the scRNA-seq data. Furthermore, the Euclidean distance criterion which is used as the default distance criterion in most methods functions poorly on the data with high dimensions [Aggarwal, Hinneburg and Keim (2001); Beyer, Goldstein, Ramakrishnan et al. (1999)], and may not be useful for scRNA-seq data [Xu and Su (2015)]. Therefore, new and sometimes hybrid machine learning methods and the combination of distance criteria should be used in this type of data so that one can reduce the data dimensions more appropriately to perform a suitable and precise clustering.

3 Proposed method

One of the ultimate goals of analyzing scRNA-seq data is identifying cell populations. Analyzing scRNA-seq data is a sophisticated procedure faced with issues such as the intrinsic probability of gene expression, existence of noise data, dropout events, and high dimensions. Each of these issues can cause diminished efficiency and accuracy in identifying cell populations. Therefore, they should be prepared for final clustering using preprocessing techniques such as filtering, dimension reduction, and data reconstruction.

The input of the process of scRNA-seq data analysis is typically a matrix called the normalized gene expression matrix as $X_{g\times n}$, which has g rows and n columns. In the mentioned matrix, g and n represent the number of genes and number of cells, respectively, where the number of genes amounts to tens of thousands of genes, while the number of cells varies between hundreds to millions of cells in some datasets. This sheer number of genes, which is considered as features of the problem, results in excessive dimensions for this type of data. Therefore, the process of their analysis for identifying the cell populations becomes complicated. One of the most important measures taken in identifying cell populations is dimension reduction, where using machine learning techniques, a set of features are extracted which can support separation of cell populations. Considering the intrinsic complexity of scRNA-seq datasets, use of classic machine learning methods may not prove very effective. Therefore, by developing these methods and presenting novel approaches, this complexity should be overcome.

The proposed method, called FSPAM here, presents a complete preprocessing step for clustering and identifying cell populations from scRNA-seq data, whose focus and contribution are related to extracting proper features for reducing the dimensions of this type of data. Since typically in scRNA-seq data, one cannot define a linear relationship between the important cellular markers, in FSPAM attempts have been made to overcome this problem through the graph theory. The proposed method, based on a graph-based Gaussian kernel and PCA, extracts a final set of features in three stages with a proper fusion of different correlation criteria. It is indeed a reduced set of markers or highly important genes which can help us in identifying cell populations in the clustering stage. The general procedure employed in the proposed method has been demonstrated in Fig. 1.

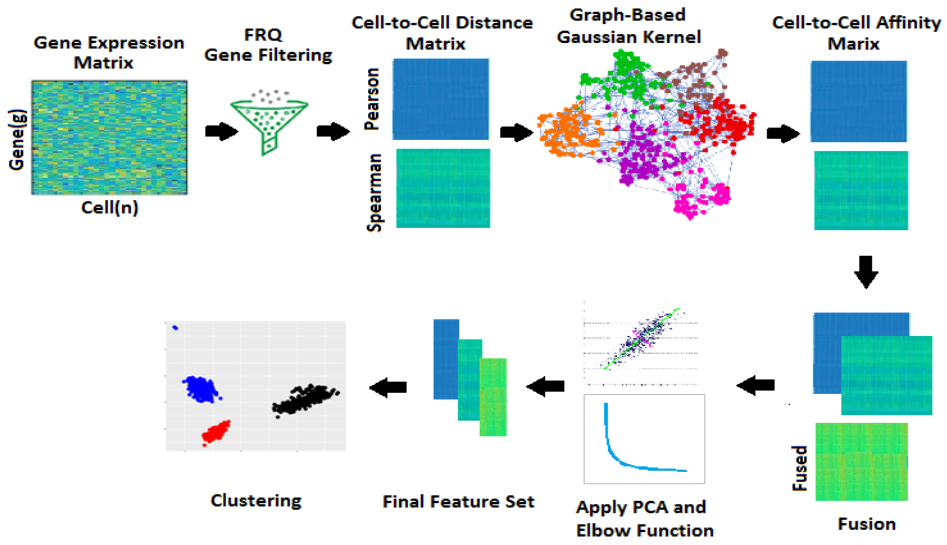

Figure 1: The overall schema of the proposed approach for identifying cell populations

In the rest of this section, each of the stages of the proposed method is explained thoroughly.

3.1 Gene filtering

One of the challenges in scRNA-seq data analysis is the noise propagation which is due to amplification error during the inverse transcription stage in RNA-seq experiments. Noise propagation emerges as excessive growth of zero and close to zero values in the dataset, creating problems in scRNA-seq data analysis. Therefore, typically in the first stage of the analysis of this type of data, a filter is applied to these data so that the features and genes that are most probably noise would be removed from the dataset of interest. The rest of the operations are then performed on the genes with a high degree of importance.

Here, we have used frequency filtering (FRQ) [Pouyan and Kostka (2018)] to select the genes, in which we consider only the genes that are expressed in a specific fraction of cells. Specifically, here we experimentally identify and eliminate the genes regarded as noise that have been expressed in less than 5% of all cell samples, and keep the remaining cells as significant features to be used in the subsequent stages.

3.2 Computing cell-to-cell distance matrix

In this stage, using the Pearson and Spearman correlation coefficient criteria, the cell-to-cell distance matrix is calculated. Each of these matrices shows an aspect of the correlation and cell-to-cell relationship in each dataset.

The Pearson correlation coefficient, known as the moment correlation coefficient or zero order correlation coefficient, is used to determine the magnitude, type, and direction of the relationship between two distance or relative variables, or a distance variable and relative variable. It is calculated by the Eq. (1).

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} \quad (1)$$

where, x and y are the variables of interest and $\overline{x}$ and $\overline{y}$ are their mean. The closer the absolute value of the correlation coefficient to 1, the stronger the relationship between the two variables is. In contrast, the correlation coefficient close to zero indicates that there is a very weak relationship between x and y variables. The Pearson correlation coefficient is a parametric method which is typically used for data with a normal distribution or a large amount of data. If we encounter ranked data or abnormally distributed data, the Spearman correlation coefficient is usually used, in which the rank of variables is used to calculate the magnitude of the relationship between two variables. In some way, it can be considered equivalent to the Pearson coefficient nonparametric method. Accordingly, the related equation can be considered as Eq. (1), in which the rank is used instead of the value of a variable [Hauke and Kossowski (2011); Kowalski (1972)].

Since we intend to present a data-driven approach in this paper, which deals with identifying cell populations regardless of any initial assumption, therefore it is assumed that in the proposed method, no previous information is available on the distribution governing the data as well as the number of data, through which one can select the proper correlation coefficient criterion. Thus, in the following, using an efficient method, the affinity matrices resulting from the Pearson and Spearman coefficient are fused and further used for extracting suitable features.

3.3 Computing cell-to-cell affinity matrix

PCA is one of the well-known and practical methods for dimension reduction in a linear fashion which tries to represent the covariance structure of a group of variables by a small set of variables. Note that this new set is a linear combination of the initial set. PCA is a method based on analyzing eigenvector decomposition (EVD), which divides the problem into principal components.

The main disadvantage of the linear conversion methods such as the PCA is that if data have a nonlinear and more complex structure, this type of methods cannot be useful. One of the solutions for overcoming this problem is applying the kernel function trick. By utilizing kernel functions, one can well calculate the principal components in spaces with high dimensions, where these feature spaces are associated with the input space through a nonlinear mapping.

In kernel-based cases, indeed a linear transformation is learned in Reproducing Kernel Hilbert Space (RKHS) [Mingtao, Zheng and Haixia (2010)]. However, since the kernel reproducing space, such as the Gaussian kernel, has nonlinear statistics from the normal data space, it yields a nonlinear conversion on the initial feature space. Note that in these methods, we do not directly move to kernel reproducing space; rather, we learn transformation as follows through a kernel function which can be applied to any data pair and implicitly in the kernel space (Eq. (2)).

$$\underset{\substack{Input\ Data\ Set \\ [\ ]_{n\times p}}}{\underset{\sim}{X}} \xrightarrow{\varphi} \underset{\substack{Affinity\ Matrix \\ [\ ]_{n\times n}}}{\underset{\sim}{Z}} \xrightarrow{Q} \underset{\substack{Features(PCs) \\ (k)}}{S} \qquad k \le p \tag{2}$$

where, p represents the number of initial features, n shows the number of samples present in the dataset, k denotes the number of final features obtained, and φ is the kernel function which is defined as Eq. (3).

$$Z_i = \varphi(x_i) = K = [K_{ij}]_{n\times n} \quad : \quad K_{ij} = K(x_i, x_j) \tag{3}$$

where, x_i and x_j show a pair of data samples present in the dataset, and K is the kernel function of interest.

The graph-based Gaussian kernel function used here receives the cell-to-cell distance matrix and calculates the affinity matrix using the following relation for each data sample based on k-nearest neighbors (Eq. (4)).

$$K(x_i, x_j) = \exp(-\frac{d^2(x_i, x_j)}{\mu\varepsilon_{i,j}}) \tag{4}$$

where, $d(x_i, x_j)$ is the distance between two samples calculated based on one of the criteria such as the Euclidean, Pearson, Spearman, etc. Further, μ is a parameter which is typically adjusted experimentally. Eventually, $\varepsilon_{i,j}$ refers to a term obtained by the Eq. (5) based on the locality of the k-neighbor of each data sample.

$$\varepsilon_{i,j} = \frac{mean(d(x_i, k_i)) + mean(d(x_j, k_j)) + d(x_i, x_j)}{3}) \tag{5}$$

where, $mean(d(x_i, k_i))$ is the mean distance between the sample x_i and its k-neighbor, $mean(d(x_j, k_j))$ represents the mean distance between the sample x_j and its k-neighbor, and $d(x_i, x_j)$ shows the distance between x_i and x_j samples.

Briefly, the introduced graph-based Gaussian kernel, based on the locality of k-neighbor of each cell, calculates the affinity matrix from the distance matrix. Through this, the input features' space is transferred to a new space via the nonlinear mapping. Then, using PCA in this new space, the eigenvectors and eigenvalues, which are the principal components, are extracted. Via this technique, one can overcome the linearity of PCA.

3.4 Fusion and feature construction

In the previous stages, two aspects of cell-to-cell similarity were obtained using the Pearson and Spearman metrics. In this step, by fusing these criteria, new aspects of this similarity will be discovered, which can lead to the extraction of new features, and help us to identifying cell populations.

In this step, the affinity matrices resulting from the Pearson and Spearman coefficients is fused by the similarity network fusion (SNF) as presented by Wang et al. [Wang, Mezlini, Demir et al. (2014)]. Concerning the SNF presented for integrating different types of data on the genome scale, it fuses the affinity networks equivalent to affinity

matrices more effectively, such that the resulting network offers a complete view about the essential relationships between the samples (cells). The resulting fused affinity matrix, which is very significant, is used for different purposes. Here, we used it for extracting high-importance features.

Generally, the SNF involves two main stages: 1) creating an affinity-sample network for each type of data, and 2) integrating these networks in the form of a single affinity network through a nonlinear and graph-based fusion method. The procedure of SNF involves first calculating the sample-to-sample affinity matrix for each dataset based on an affinity criterion. This matrix is equivalent to an affinity network whose nodes are samples, while the weighted edges represent the extent of similarity of each pair of samples. For the stage of combining networks, the SNF uses a nonlinear approach based on the message sending theory, which frequently updates every network by receiving information from other networks, where each repetition increases the extent of affinity. After several repetitions, the SNF converges to an integrated network. The main advantage of this type of integration method is that the weak affinities, which are the low-weight edges, disappear, thereby supporting noise reduction. On the other hand, the strong affinities or the high-weight edges observed in one or several networks are summed up together, strengthening strong similarities. Also, the low-weight edges supported by all networks are preserved, given the extent of their strong connection to neighbors. Such a nonlinear fusion allows the SNF to use it more completely by integrating the shared and complementary information of a local network structure.

In the following, PCA is applied to the affinity matrices, and after achieving the principal components (PC), the best components are extracted as a set of features with high importance (PC_i) using the Elbow method [Thorndike (1953)].This method plots the PCs obtained based on the value and in a descending order on the coordinate axes. The point where the break occurs in the diagram is called the elbow point. The PCs located before the elbow point are kept as the PCs or the best set of features obtained from each stage (Fig. 2).

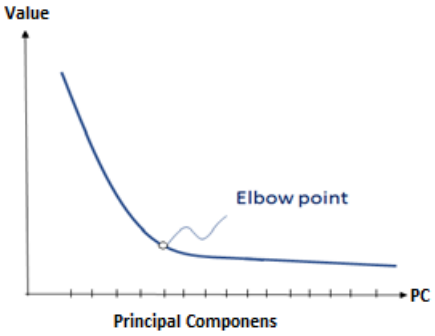

Figure 2: Use of the Elbow method for selecting more important principal components

If the Elbow method is not used for choosing more important principal components, we should consider the number of components as a predetermined constant number, which questions the flexibility of the proposed approach. Nevertheless, usage of the Elbow method helps the system to obtain highly important components after the application of PCA given the dataset of interest. This makes our proposed approach flexible and data-driven.

After extracting the PC_1, PC_2, and PC_3 feature sets, the final feature set is obtained by juxtaposing these features and creating the ultimate eigenvector (Eq. (6)).

$$PC_F = PC_1 \cup PC_2 \cup PC_3 \quad (6)$$

PC_F contains a set of valuable features which can help us in identifying cell populations. Note that this feature set is obtained without any previous knowledge about the distribution governing the data and label of samples. Indeed, it is considered a kind of unsupervised feature extraction approach, which is far more valuable than supervised algorithms. Notably, the set and number of the final features constructed by the FSPAM are different from one dataset to another, where it extracts the best collection of markers in line with the dataset of interest, making the proposed approach flexible and data-oriented.

3.5 Clustering

After the feature extraction, in the next step, the clustering process is performed to identify the cell populations in the new space. One of the most popular clustering methods used in the clustering of scRNA-seq data is the k-means method. In this paper, due to the simplicity and high speed, the k-means clustering method has been used. One of the problems of k-means, however, is its instability, which happens to the random selection of centroids. To overcome this challenge, a method called k-means++ [Arthur and Vassilvitskii (2007)] is presented which partly addresses this problem and provides a more stable clustering algorithm. In this method, first, during an iterative process, by selecting and testing different centroids, the best centroids are identified, after which the standard clustering of k-means is performed using these points. Although it is time-consuming to find these centroids, as it reduces the convergence time of the standard k-means, it will also compensate for that extra time.

4 Results and discussion

The proposed method, the FSPAM, has been developed by the R language and the experiments were run on an intel Core i7 CPU 2.67 GHz computer with 6 GB RAM.

In this section, we first discuss the results of the proposed method from various aspects such as flexibility, accuracy, quality, and stability, and then compare the final results with different and well-known methods in this regard. Before that we briefly review the datasets and evaluation parameters used in this experiment.

4.1 Datasets

The FSPAM has been implemented on Buettner, Kolod, and Usoskin scRNA-seq datasets, and the results were obtained in approximately 10, 90, and 60 seconds, respectively [Buettner, Natarajan, Casale et al. (2015); Kolodziejczyk, Kim, Tsang et al. (2015); Usoskin, Furlan, Islam et al. (2015)]. The properties of these datasets are provided in Tab. 1. These datasets were downloaded from https://github.com/BatzoglouLabSU/SIMLR. For our analysis, the Usoskin and Kolod datasets were re-downloaded to obtain the normalized expression values without batch corrections. For Usoskin, the data were downloaded from the ‘External resource’, available at http://linnarssonlab.org/drg/; for Kolod, the data were downloaded from https://www.ebi.ac.uk/teichmann-srv/espresso/.

Table 1: The properties of scRNA-seq datasets used for evaluating the FSPAM proposed method

Dataset	# Cell	# Gene	# Cluster	Ref.
Buettner	182	9573	3	[Buettner, Natarajan, Casale et al. (2015)]
Kolod	704	13473	3	[Kolodziejczyk, Kim, Tsang et al. (2015)]
Usoskin	622	17772	4	[Usoskin, Furlan, Islam et al. (2015)]

4.2 Evaluation metrics

To evaluate the quality of the clustering, we used three well-known clustering criteria, i.e., ARI, NMI, and Purity. Each of them is explained briefly further.

Adjusted Rand Index (ARI): assume that we divide n cells by k clusters, where $\{u_i\}_{i=1}^{n}$ represents the final labels produced by the clustering method. Also, assume that $\{v_i\}_{i=1}^{n}$ reflects the real labels of each cell (correct cell type). Based on the two mentioned definitions, ARI is calculated according to Eq. (7):

$$ARI = \frac{\sum_{ls}\binom{n_{ls}}{2} - (\sum_{l}\binom{n_l}{2}\sum_{s}\binom{n_s}{2})/\binom{n}{2}}{(\sum_{l}\binom{n_l}{2} + \sum_{s}\binom{n_s}{2})/2 - (\sum_{l}\binom{n_l}{2}\sum_{s}\binom{n_s}{2})/\binom{n}{2}} \tag{7}$$

In this relation, l and s are the indices referring to k clusters. $n_l = \sum_i^n I(u_i = l)$, $n_s = \sum_i^n I(v_i = s)$, and $n_{ls} = \sum_{i,j} I(u_i = l)\, I(v_i = s)$. In these relations, $I(x = y)$ is the indicator function, whose value is 1 when $x = y$,; otherwise, it is zero. Briefly, if the label of clusters produced by a clustering algorithm fully corresponds to original labels, then the ARI value is 1; otherwise, the ARI value declines in proportion with the inconsistencies that exist.

Normalized mutual index (NMI): assume that $p_l = \frac{n_l}{n}$, $q_s = \frac{n_s}{n}$, and $z_{ls} = \frac{n_{ls}}{n}$ .Now, the entropy related to each clustering solution (related to v and u) can be defined as follows: $h(u) = -\sum_l p_l * log p_l$ and $h(v) = -\sum_s q_s * log q_s$. Furthermore, the extent of mutual information between these two clustering solutions is defined as $i(u, v) = \sum_{l,s} z_{ls} \log(z_{ls}/p_l/q_s)$. Now, based on these relations, the NMI criterion is defined as Eq. (8).

$$NMI = i(u, v)/\sqrt{h(u)h(v)} \tag{8}$$

As with ARI, if there is 100% correspondence between u and v clustering solutions, the NMI value becomes 1. Briefly, the closer the ARI and NMI values related to a clustering method applied to a dataset to 1, the higher the quality of the clustering will be.

Purity: the criterion of purity is measured for clusters with a unit class. For its calculation, for every cluster, the number of data points from the typical class is counted in the cluster of interest. Then, all clusters are summed up together and divided by the number of data points (Eq. (9)).

$$Purity(\Omega, C) = \frac{1}{N}\sum_{k,j} \max\left|w_k \cap c_j\right| \tag{9}$$

where, $\Omega = \{w_1, w_2, \ldots, w_k\}$ and $C = \{c_1, c_2, \ldots, c_j\}$ are the sum of groups (correct

clustering solutions).

4.3 Parameters adjustment

In the feature extraction step, when we convert the distance matrix to a similarity matrix, we use a graph-based Gaussian kernel, which operates according to Eqs. (4) and (5). These relations convert all distance values to similarity values based on a k-nearest neighbor and a parameter μ. Here, we used an empirical method to obtain the best results for which the values were as follows $\mu = 0.5$ and $k = 40$. The results related to the adjustment of these parameters are presented in Fig. 3.

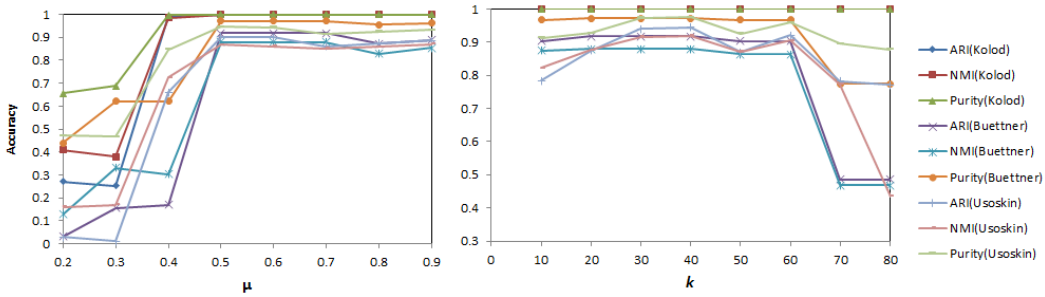

Figure 3: Selecting the value of μ and $\boldsymbol{k}$ parameters experimentally ($\boldsymbol{\mu = 0.5}$, $\boldsymbol{k = 40}$)

4.4 Results

A notable point about the FSPAM is that the proposed method produces a variable and high-quality set of features and principal components in line with each dataset, which helps in more accurate identification of cell populations. This characteristic has been shown in Fig. 4, in which PC1, PC2, and PC3 represent the number of features extracted according to the Pearson, Spearman, and their fusion affinity matrices respectively. Also, PC_F represents the number of final constructed features, which has been extracted as 35, 25, and 67 for Kolod, Buettner, and Usoskin datasets respectively. Also, in order to find the success of the proposed method for the proper fusion of the Spearman and Pearson correlation coefficients, we examine the results of the FSPAM with the results of the Spearman and Pearson correlation coefficients separately. The results show that the use of these coefficients depends on the datasets and their governing distribution, while the FSPAM method obtains the best results, independent of the dominant characteristics of the data (Fig. 5).

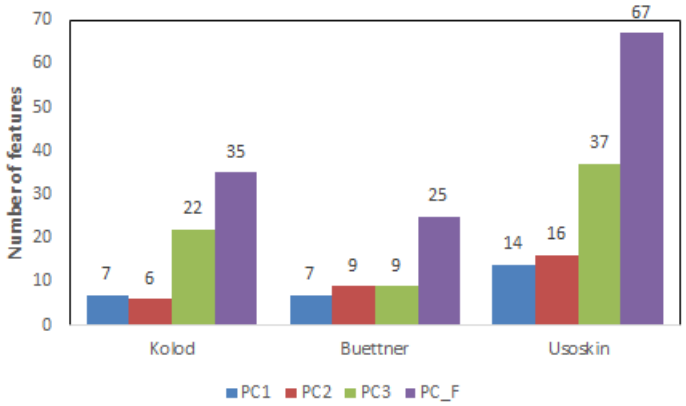

Figure 4: The number of features extracted by the FSPAM

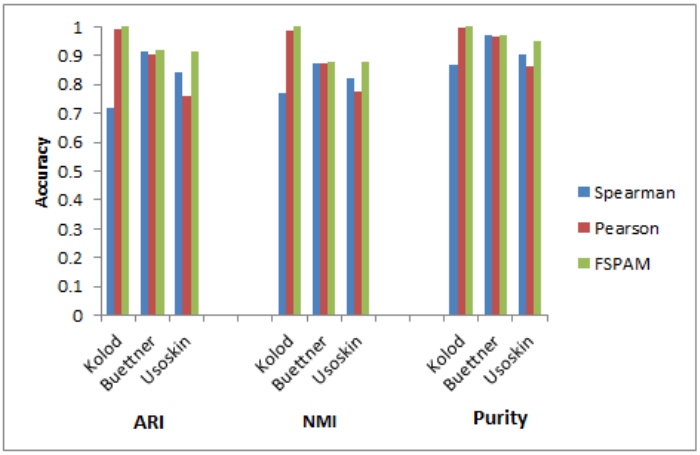

Figure 5: The FSPAM obtains the best results, independent of the dominant characteristics of the data: (1) The Pearson correlation coefficient in the Kolod dataset is better than the Spearman, (2) the Spearman correlation coefficient in the Buettner and Usoskin datasets yields better results relative to the Pearson, (3) the FSPAM method is consistent with the proper fusion of these coefficients gets the best results

In the following, we first obtain the clustering results and identify the cell populations using different clustering methods based on the extracted features. Then, we examine the clustering quality of the proposed method by visualizing the data, and finally discuss the stability and robustness of the proposed method.

In the performed implementation, after extracting the set of final features, to identify the cell populations, we tested different methods of clustering including hierarchical (HR), GMM, DBSCAN, and k-means. The obtained results indicated that the k-means clustering offers the best results (Tab. 2). Therefore, for the rest of the work and in the other stages of testing and evaluating the FSPAM, in the clustering stage, a method developed based on k-means called k-means++ has been used.

Table 2: The results of implementing the FSPAM using different clustering methods (ARI%)

Data/Method	k-means	GMM	HR	DBSCAN
Kolod	**100**	100	100	81.4
Buettner	**91.82**	66.3	88.2	75.4
Usoskin	**91.39**	53.8	66.5	73.2

In the next step, we show that the set of features obtained from FSPAM enhances the quality of clustering cells and can visualize cells with a far better quality by the PCA dimension reduction method. For this purpose, we map every dataset twice by the PCA on a two-dimensional space to visualize the cellular space. In the first state, the PCA is directly applied to the original dataset, and all cells are visualized on the resulting two-dimensional space. In this second state, first the FSPAM is applied to the dataset, and then the features of interest are extracted. Next, the PCA is applied to visualize data on these extracted features, with the results summarized in Fig. 6. In every panel of this figure, every point represents a cell, and each color refers to a type of cell. As can be observed, the features extracted by the FSPAM differentiate different types of cells with a far higher quality.

Furthermore, to investigate the stability, we replicated the FSPAM 50 times for each of the three mentioned datasets, such that in every replication, 90% of data were chosen randomly. The results related to the ARI criterion obtained in these replications have been summarized as a box diagram in Fig. 7, in which blue, red, and green boxes represent variations of the results in the three datasets of Kolod, Buettner, and Usoskin, respectively. As can be observed in this diagram, for the dataset Kolod, FSPAM yields the minimum extent of change, where the accuracy of the results fluctuates within the quartile range of zero. In this dataset, only two different values of 0.98 and 0.993 were obtained, while the other results in the other 48 replications were equal to 1. Furthermore, in the other datasets, again favorable results were obtained, such that in Buettner and Usoskin data, the results fluctuated within the quartile range of 0.0522 and 0.0097, respectively. These results suggest that generally the FSPAM enjoys high robustness and stability, and one can rely on the obtained results to a large extent.

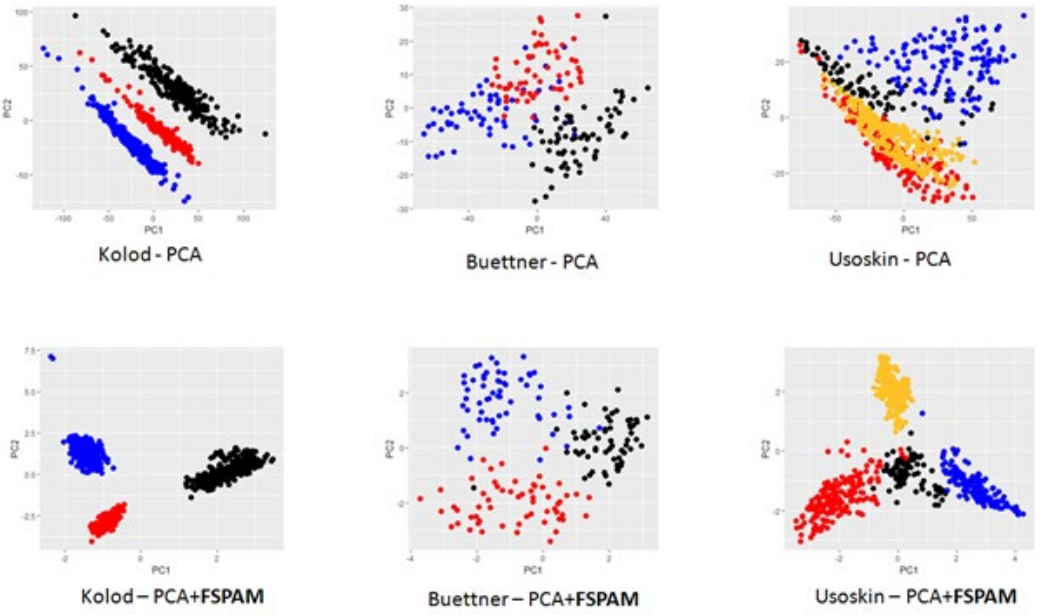

Figure 6: Visualization of data with and without the FSPAM

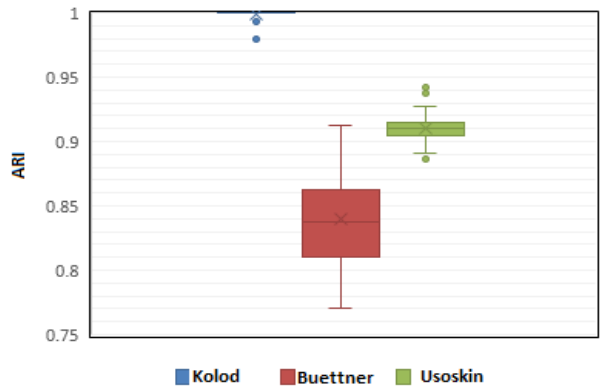

Figure 7: Robustness and stability of the FSPAM based on 50 replications on the tested datasets

Finally, for the quantitative assessment related to the clustering quality, we used the kNN classification method in the resulting two-dimensional space, whereby the resulting classification error (the rate of wrongly classified cells or NNE) has been calculated as the overall error of mapping. Since the kNN algorithm is dependent on k parameter, the number of nearest neighbors, we computed the resulting two-dimensional mapping error per different values of k (3,5,7,9), with the results being summarized in Tabs. 3 and 4. As can be observed, the NNE error when using the features extracted by the FSPAM has been far lower than the case when the PCA has been directly applied to the scRNA-seq gene expression matrix. This suggests the success of the proposed method in constructing the

valuable features and reducing the dimensions of the problem.

Table 3: PCA-(%) NNE error on the original dataset

Data/Method	K=1	K=3	K=5	K=7	K=9	Avg
Kolod	1.1	1.0	0	0	0	0.4
Buettner	21.4	13.1	11.5	10.9	10.3	13.44
Usoskin	30.1	31.2	32.0	28.3	29.7	30.3

Table 4: PCA-(%) NNE error on the set of features extracted by the FSPAM method

Data/Method	K=1	K=3	K=5	K=7	K=9	Avg
Kolod	0	0	0	0	0	0
Buettner	7.1	5.4	5.4	5.4	5.1	5.7
Usoskin	1.7	1.4	1.6	1.7	1.7	1.6

4.5 Evaluation of the proposed FSPAM method with other methods

In this section, we compared the FSPAM with three traditional clustering methods including k-means, GMM, and hierarchical clustering method (HCLUST) as well as new clustering datasets including the SINCERA [Guo, Wang, Potter et al. (2015)], SNN-Cliq [Xu and Su (2015)], and pcaReduce [Žurauskiene and Yau (2015)], which have been designed for scRNA-seq data.

The FSPAM and other six mentioned methods were applied to three datasets of scRNA-seq in Tab. 1, where for each method, the ARI, NMI, and Purity were calculated separately. Fig. 8 summarizes the comparison of the obtained results. Also, its details are provided in Tabs. 5, 6, and 7, representing the high accuracy of the proposed FSPAM method.

As shown in Tab. 5, the ARI parameter value obtained from the proposed FSPAM method, as one of the most important clustering parameters, was obtained for the Kolod, Buettner, and Usoskin datasets. The values were 100, 91.82, and 91.38, respectively which are considerably higher than the values of other methods. The SINCERA is the only method to achieve something equal to the FSPAM method for the Kolod dataset, but in other datasets, it lags behind the FSPAM method.

The values obtained from the implementation of the FSPAM method for the NMI parameter on the Kolod, Buettner, and Usoskin datasets were 100, 87.93, 87.73, respectively. This suggests that the proposed method's accuracy is equal to that of Kolod dataset with the SINCERA method, while in other datasets, it has been more accurate than in other methods (Tab. 6).

The results for the purity parameter also showed that the proposed method presented values of 100, 97.25 and 95.18. As with all other parameters, they were equal to the Kolod dataset values with the SINCERA method, while in other datasets, higher values were obtained (Tab. 7).

With regard to the results obtained in this section and previous sections, it can be seen how the FSPAM method can extract valuable and, of course, proportional features to each scRNA-seq dataset, with a high accuracy and quality to identify cell populations.

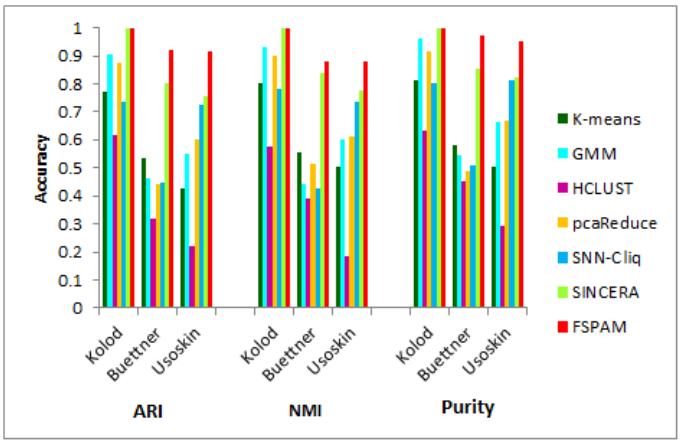

Figure 8: Summary of the results obtained from different methods and the FSPAM

Table 5: The results of implementing the FSPAM and comparing it with well-known methods (ARI)

Data/Method	k-means	GMM	HCLUST	pcaReduce	SNN-Cliq	SINCERA	**FSPAM**
Kolod	77.3	90.6	61.8	87.6	73.5	**100**	**100**
Buettner	53.3	46.3	31.8	44.1	44.8	80.3	**91.82**
Usoskin	42.6	55.2	22.3	60.4	72.6	75.6	**91.39**

Table 6: The results of implementing the FSPAM and comparing it with well-known methods (NMI)

Data/Method	k-means	GMM	HCLUST	pcaReduce	SNN-Cliq	SINCERA	**FSPAM**
Kolod	80.11	93.33	57.78	90.23	78.36	**100**	**100**
Buettner	55.45	44.11	38.93	51.63	42.55	83.71	**87.93**
Usoskin	50.49	60.23	18.26	61.37	73.55	77.53	**87.73**

Table 7: The results of implementing the FSPAM and comparing it with well-known methods (purity)

Data/Method	k-means	GMM	HCLUST	pcaReduce	SNN-Cliq	SINCERA	**FSPAM**
Kolod	81.23	96.18	63.42	91.51	80.32	**100**	**100**
Buettner	58.14	54.73	45.25	48.93	50.79	85.22	**97.25**
Usoskin	50.39	66.08	29.32	66.72	81.38	82.17	**95.18**

5 Conclusion

In this paper, we dealt with identifying cell populations from scRNA-seq data. The analysis of this type of data has some challenges including the existence of noise, dropout events, and their high dimensions. Our proposed method, which we called, the fusion of the Spearman and Pearson affinity matrices, FSPAM, was an unsupervised method according to the graph-based Gaussian kernel. It extracted a suitable feature set for every scRNA-seq dataset without any previous knowledge about the type of cells and through proper fusion of the affinity matrices resulting from the Spearman and Pearson correlation criteria. They were used as valuable markers in the clustering process to identify cell populations. The results on three different datasets indicated that the set of features obtained from the FSPAM enhanced the quality of clustering cells and could visualize cells with a far better quality by the PCA dimension reduction method. The notable point about the FSPAM was the variable set of features extracted from the feature construction step in line with each dataset, making the FSPAM a data-driven approach. Also, to investigate the stability, we replicated the FSPAM 50 times for each of the three mentioned datasets. The results indicated that generally the FSPAM enjoys great robustness and stability, and one can rely on the obtained results to a large extent. Finally, for quantitative assessment related to the clustering quality, we used the kNN classification method in the resulting two-dimensional space. When using the features extracted by the FSPAM, NNE error was far lower than in the case when the PCA was directly applied to the scRNA-seq gene expression matrix.

Also, to evaluate the proposed FSPAM method, we examined it against other well-known methods and from different aspects. To do this, we used three methods of classical clustering (k-means, HCLUST, and GMM) and three state-of-the-art methods presented to identify cell populations in the scRNA-seq data (pcaReduce, SINCERA, and SNN-Cliq). The results revealed that through a valuable feature set tailored to any data, the FSPAM method can be very accurate in clustering and identifying cell populations.

In summary, we can say that in this paper a method called FSPAM was presented which can be used both for extracting a variable set of valuable features and for identifying accurate cell populations. Indeed, the proposed method, which is an accurate, quality, and stable approach, can be used as a useful tool for analyzing scRNA-seq data for bioinformatics researchers in this area.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Aggarwal, C. C.; Hinneburg, A.; Keim, D. A. (2001): On the surprising behavior of distance metrics in high dimensional space. *International Conference on Database Theory, LNCS 1973*, pp. 420-434. Springer, Berlin, Heidelberg.

Amir, el A. D.; Davis, K. L.; Tadmor, M. D.; Simonds, E. F.; Levine, J. H. et al. (2013): viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, vol. 31, no. 6, pp. 545-552.

Arthur, D.; Vassilvitskii, S. (2007): k-means++: the advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027-1035.

Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. (1999): When is 'Nearest Neighbor' meaningful? *International Conference on Database Theory*, pp. 217-235.

Buettner, F.; Natarajan, K.; Casale, F.; Proserpio, V.; Scialdone, A. et al. (2015): Computational analysis of cell-to-cell heterogeneity in single-cell RNA-Sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, vol. 33, no. 2, pp. 155-160.

Grün, D.; Lyubimova, A.; Kester, L.; Wiebrands, K.; Basak, O. et al. (2015): Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, vol. 525, no. 7568, pp. 251-255.

Guo, M.; Wang, H.; Potter, S. S.; Whitsett, J. A.; Xu, Y. (2015): SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Computational Biology*, vol. 11, no. 11, e1004575.

Hashimshony, T.; Wagner, F.; Sher, N.; Yanai, I. (2012): CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports*, vol. 2, no. 3, pp. 666-673.

Hauke, J.; Kossowski, T. (2011): Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, vol. 30, no. 2, pp. 87-93.

Jaitin, D. A.; Kenigsberg, E.; Keren-Shaul, H.; Elefant, N.; Paul, F. et al. (2014): Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, vol. 343, no. 6172, pp. 776-779.

Kiselev, V. Y.; Kirschner, K.; Schaub, M. T.; Andrews, Y.; Yiu, A. et al. (2017): SC3-consensus clustering of single-cell RNA-Seq data. *Nature Methods*, vol. 14, no. 5, pp. 483-486.

Kolodziejczyk, A. A.; Kim, J. K.; Tsang, J. C.; Ilicic, T.; Henriksson, J. et al. (2015): Single-cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, vol. 17, no. 4, pp. 471-485.

Kowalski, C. (1972): On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, vol. 21, no. 1, pp. 1-12.

Li, X.; Wong, K. C. (2019): Single-cell RNA-seq interpretations by evolutionary multiobjective clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.*

Li, X.; Zhang, S.; Wong, K. C. (2019): Single-cell RNA-seq interpretations using evolutionary multiobjective ensemble pruning. *Bioinformatics*, vol. 35, no. 16, pp. 2809-2817.

Liu, S. J.; Nowakowski, T. J.; Pollen, A. A.; Lui, J. H.; Horlbeck, M. A. et al. (2016): Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biology*, vol. 17, pp. 67.

Macosko, E. Z.; Basu, A.; Satija, R.; Nemesh, J.; Shekhar, K. et al. (2015): Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, vol. 161, no. 5, pp. 1202-1214.

Mingtao, D.; Zheng, T.; Haixia, X. (2010): Adaptive kernel principal component analysis. *Signal Processing*, vol. 90, no. 5, pp. 1542-1553.

Nagalakshmi, U.; Waern, K.; Snyder, M. (2010): RNA-Seq: a method for comprehensive transcriptome analysis. *Current Protocols in Molecular Biology*, vol. 89, no. 1, pp. 4-11.

Nelson, A. C.; Mould, A. W.; Bikoff, E. K.; Robertson, E. J. (2016): Single-cell RNA-seq reveals cell type-specific transcriptional signatures at the maternal-foetal interface during pregnancy. *Nature Communication*, vol. 7, pp. 11414.

Patel, A. P.; Tirosh, I.; Trombetta, J. J.; Shalek, A. K.; Gillespie, S. M. et al. (2014): Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, vol. 344, no. 6190, pp. 1396-1401.

Pellegrino, M.; Sciambi, A.; Yates, J. L.; Mast, J. D.; Silver, C. et al. (2016): RNA-Seq following PCR-based sorting reveals rare cell transcriptional signatures. *BMC Genomics*, vol. 17, pp. 361.

Picelli, S.; Björklund, A. K.; Faridani, O. R.; Sagasser, S.; Winberg, G. et al. (2013): Smart-Seq2 for sensitive full-length transcriptome profiling in single-cells. *Nature Methods*, vol. 10, no. 11, pp. 1096-1098.

Pouyan, M. B.; Nourani, M. (2016): Clustering single-cell expression data using random forest graphs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 21, no. 4, pp. 1172-1181.

Pouyan, M. B.; Nourani, M. (2017): Identifying cell populations in flow cytometry data using phenotypic signatures. *IEEE/ACM Trans Comput Biol Bioinform*, vol. 14, no. 4, pp. 880-891.

Pouyan, M. B.; Kostka, D. (2018): Random forest based similarity learning for single-cell RNA sequencing data. *Bioinformatics*, vol. 34, no. 13, pp. 79-88.

Segerstolpe, Å.; Palasantza, A.; Eliasson, P.; Andersson, E. M.; Andreasson, A. C. et al. (2016): Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metabolism*, vol. 24, no. 4, pp. 593-607.

Shalek, A. K.; Satija, R.; Adiconis, X.; Gertner, R. S.; Gaublomme, J. T. et al. (2013): Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, vol. 498, pp. 236-240.

Tang, F.; Barbacioru, C.; Wang, Y.; Nordman, E.; Lee, C. et al. (2009): mRNA-seq whole-transcriptome analysis of a single-cell. *Nature Methods*, vol. 6, no. 5, pp. 377-382.

Tasic, B.; Menon, V.; Nguyen, T. N.; Kim, T. K.; Jarsky, T. et al. (2016): Adult mouse cortical cell taxonomy revealed by single-cell transcriptomics. *Nature Neuroscience*, vol. 19, no. 2, pp. 335-346.

Thorndike, R. L. (1953): Who belongs in the family? *Psychometrika*, vol. 18, no. 4, pp. 267-276.

Usoskin, D.; Furlan, A.; Islam, S.; Abdo, H.; Lönnerberg, P. et al. (2015): Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature Neuroscience*, vol. 18, no. 1, pp. 145-153.

Van Der Maaten, L.; Hinton, G. (2008): Visualizing data using t-SNE. *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605.

Van Der Maaten, L.; Postma, E.; Van Den Herik, J. (2009): Dimensionality reduction: a comparative review. *Tilburg University Technical Report, TiCC-TR*.

Villani, A. C.; Satija, R.; Reynolds, G.; Sarkizova, S.; Shekhar, K. et al. (2017): Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, vol. 356, no. 6335, eaah4573.

Wagner, A.; Regev, A.; Yosef, N. (2016): Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, vol. 34, pp. 1145-1160.

Wang, Z.; Gerstein, M.; Snyder, M. (2010): RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, vol. 10, pp. 57-63.

Wang, B.; Mezlini, A. M.; Demir, F.; Fiume, M.; Tu, Z. et al. (2014): Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, vol. 11, no. 3, pp. 333-337.

Wang, Y. J.; Schug, J.; Won, K. J.; Liu, C.; Naji, A. et al. (2016): Single-cell transcriptomics of the human endocrine pancreas. *Diabetes*, vol. 65. no. 10, pp. 3028-3038.

Xu, C.; Su, Z. (2015): Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, vol. 31, no. 12, pp. 1974-1980.

Žurauskiene, J.; Yau, C. (2015): pcaReduce: hierarchical clustering of single-cell transcriptional profiles. *BMC Bioinformatics*, vol. 17, pp. 140.