Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Single Cell Clustering Based on Cell-Pair Differentiability Correlation and Variance Analysis

Hao Jiang [1], Lydia Sohn[2], Haiyan Huang[2,*] and Luonan Chen[3,4,*]

[1]Department of Mathematics, School of Information, Renmin University of China, Beijing 100872, China
[2]Department of Statistics, University of California, Berkeley, USA
[3]Key Laboratory of Systems Biology, CAS Center for Excellence in Molecular Cell Science, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[4]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

*To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The rapid advancement of single cell technologies has shed new light on the complex mechanisms of cellular heterogeneity. Identification of intercellular transcriptomic heterogeneity is one of the most critical tasks in single-cell RNA-sequencing studies.

Results: We propose a new cell similarity measure based on cell-pair differentiability correlation, which is derived from gene differential pattern among all cell pairs. Through plugging into the framework of hierarchical clustering with this new measure, we further develop a variance analysis based clustering algorithm '**Corr**' that can determine cluster number automatically and identify cell types accurately. The robustness and superiority of the proposed algorithm are compared with representative algorithms: SNN-Cliq and several other state-of-the-art clustering methods, on many benchmark or real scRNA-Seq datasets in terms of both internal criteria (clustering number and accuracy) and external criteria (purity, adjusted rand index, F1-measure). Moreover, differentiability vector with our new measure provides a new means in identifying potential biomarkers from cancer related single cell data sets even with strong noise. Prognosis analyses from independent datasets of cancers confirmed the effectiveness of our '**Corr**' method.

Implementation and Availability: The source code (Matlab) is available at http://sysbio.sibcb.ac.cn/cb/chenlab/soft/Corr--SourceCodes.zip.

Contact: lnchen@sibs.ac.cn or hyh0110@berkeley.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A transcriptome analysis utilizing single cell RNA sequencing (scRNA-Seq) has been one of the most attractive topics among recent research activities. The core technique of scRNA-Seq is to use the next-generation sequencing technologies to sequence cDNAs prepared from a single cell to get information about the cell's RNA content. scRNA-Seq thus offers gene expression measurements at single cell resolution. This unprecedented resolution into cell states allows the investigation of population heterogeneity as well as genetic and epigenetic variabilities in a cellular system (Eberwine *et al*., 2014; Stegle *et al*., 2015). For example, a popular study has conducted clustering analysis to identify cell subpopulations from a set of heterogeneous cells, hoping for a better understanding of cell function and dysfunction (Eisenberg and Levanon, 2013).

Despite the rapid advancement in scRNA-Seq technologies, multiple types of noise prevail in the single-cell experiments and cannot be overlooked. One source of noise is the biological fluctuation in both global and local perspectives (Shapiro *et al*., 2013), e.g., unwanted cell-to-cell variations. There are also noises from the scRNA-Seq protocols, e.g., the technical biases caused by pipetting errors, temperature difference, PCR amplifications, etc. These noises have led to many challenging issues in scRNA-Seq data analysis, including but not limited to high rates of dropout events, batch effects and unwanted cell-to-cell variations (Cheng *et al*., 2017). The prevalence of dropout events would lead to zero-inflated data. Batch effects usually occur due to inconsistencies in library preparation of RNA samples (for sequencing) across different biological labs and thus confound true gene expression differences. Unwanted cell-to-cell variations are attributable to cell size, cell cycle state, and other factors that are irrelevant for cell type identification. If these issues were not managed properly, the results can be significantly affected. We refer to (Yuan *et al*., 2017) for a comprehensive review and discussion on issues in single-cell analysis.

scRNA-Seq data is noisy and of high dimensionality. Many clustering methods have been proposed to deal with data structure in high dimensionality and noise distributions. Among those, some efforts designed new (dis)similarity measures which were implemented in traditional clustering algorithms such as hierarchical or k-means clustering. Shared nearest neighbor (SNN), is a commonly used secondary similarity measure (Guha *et al*., 1999; Ertoz *et al*., 2003), expressed as a function of shared fixed-sized neighborhoods determined by the primary measure, e.g. Euclidean norm. It has been proven to be robust and produce relatively stable clustering results compared to those based on primary measures (Houle *et al*., 2010). However, Guha *et al.* proposed a robust hierarchical clustering algorithm for dealing with categorical attributes such as attributes described by different colors. The algorithm used the number of neighborhood information to measure the similarity of samples instead of merely using the attribute values of sample pairs. However, the algorithm is not quite fit for single cell data sets where scRNA-Seq values are not categorical. On the other hand, in Ertoz *et al.*, SNN clustering is proposed for handling high dimensional data which is constructed based on the notion of DBSCAN by a new definition of density and the core points, and hence the number of optimal clusters is determined automatically. SNN-Cliq (Xu *et al*., 2015) as an extension of SNN clustering algorithm, was further developed based on a quasi-cliq clustering method

to identify tight groups, where similarity measure is constructed based on neighborhood ranking. However, the drawback of the algorithm lies in large scale single cell clustering where many noisy clusters are detected.

Some used well known Pearson correlations or Euclidian distances as measures but implemented them in a newly designed clustering algorithm. The Rare Cell Type Identification algorithm identifies disease-specific cells through k-means clustering (Grun *et al*., 2015), where the first step involves cell similarity construction under Pearson correlation coefficient. One of the drawbacks lies in convergence to a local minimum which may produce counterintuitive results. The phenongraph algorithm (Levine *et al*., 2015) constructs k-nearest neighbor graph to study cell phenotypes, where the distance is measured under Euclidean distance. This algorithm is a supervised one based on the construction of Jaccard graph and the dissimilarity between cells is evaluated from a local perspective. The application to unsupervised cases is not clear. Bo *et al*. (2017) proposed a kernel based similarity learning algorithm for analysis of scRNA-Seq data, where RBF kernel is utilized with Euclidean norm as distance measurement. Teschendorff and Enver (2017) introduced partitioning-around-medoids (PAM) algorithm with Pearson distance correlation metric to infer cell clusters. It works with a generalization of the Manhattan Norm to define distance between data-points instead of L_2 norm. However, the algorithm needs to firstly know the number of cell populations, which restricts the generalization ability to unknown cell type cases. Vladimir *et al* (2017) recently developed a consensus clustering framework for single cell clustering, where a number of distance measures are considered in cell dis(similarity) construction, and a final consensus similarity matrix is obtained for hierarchical clustering. However, all these methods evaluate the dissimilarity of cells locally without taking the micro-environment into consideration.

As discussed above, a central problem in clustering analysis of single cells is quantifying the relationships between cells. However, in general, scRNA-seq data is of high dimensionality, noisy, sparse and heterogeneous (Xu *et al*., 2015). These properties make conventional (dis)similarity measures less effective and reliable (Beyer *et al*., 1999). In this paper, inspired by the previous methods of measuring cell-to-cell similarity using neighborhood information, we propose a new measure for any pair of cells called '**Corr**', which defines a cell-to-cell "differentiability correlation" for scRNA-Seq data taking into consideration on the expression patterns of surrounding cells from a global perspective. In particular, this "differentiability correlation" assesses cell-to-cell relationships based on gene differential patterns. It has a form similar to Gamma correlation, which evaluates all pair-wise similarities between cells from a global viewpoint to ensure a robust association assessment on any two cells. To perform clustering analysis of cells, we borrowed the framework of hierarchical clustering, but using our '**Corr**' as the cell-cell similarity measure and a new rule based on variance analysis to decide where to cut the hierarchical dendrogram. The newly developed dendrogram-cutting rule can be used to determine the number of clusters (or cell subpopulations) automatically. This proposed algorithm is compared with SNN-Cliq and several other state-of-the-art clustering methods for their performance in recovering biologically meaningful cell types when applied to a number of real scRNA-Seq datasets. The effectiveness of the algorithms was evaluated by internal criteria (clustering number and accuracy) and external criteria (purity, adjusted rand index (ARI), F1-measure). Moreover, when computing'**Corr**' , an intermediate step generates the "differentiability vectors" that reveal global expression pattern in the whole environment, thus potentially providing a new means to identify biomarkers of diseases from single cell data. Prognosis analyses on independent datasets of cancers validated the effectiveness of our method in identification of robust biomarkers for Kaplan Meier test.

2 Method

It is known that cell microenvironment or cell-cell interaction plays a critical role in many biological processes (Mario and Antonio., 2010), and can be engineered to improve cell based drug testing (Bhadriraju and Chen., 2002) or targeted therapies for diseases (Gong *et al*., 2005). It is hence desirable, at least to some extent, to account for the effects of microenvironment or cell-cell interaction when assessing the relationships between cells. In this paper, under the belief that cells with similar expression patterns would likely have similar functions, we define a new pairwise similarity measure based on scRNA-Seq data to assess potential functional relationships between cells. This measure evaluates "differentiability correlation", i.e., the correlation between differential expression statuses of the genes in two cells (when each cell is compared against most other cells). It incorporates the microenvironment that evaluates the relationship of two considered cells by taking into consideration on all the other surrounding cells from a global perspective.

With the new "differentiability correlation" employed, we will identify clusters of cells (i.e., cell type identification) using a hierarchical clustering analysis (HCA) procedure. As a popular clustering method, HCA not only groups together cells but also provides a natural way to graphically represent all the cells in a hierarchical structure, allowing a thorough inspection on the relationships between cells and clusters of cells. We also develop a new criterion for determining the number of clusters under HCA. The method is described in detail in the following section.

2.1 Pairwise cell similarity measure – differentiability correlation

We assume there are n cells, and denote the gene expression profile for cell i by

$$\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}),$$

where p is the number of genes and $i = 1, 2, \ldots, n$.

To define the "differentiability correlation" between cells i and j, we first identify two "feature" gene sets for each cell. In more details, for cell i, we will identify the gene sets $\mathbf{V}_{i_j}^+$ and $\mathbf{V}_{i_j}^-$, where $\mathbf{V}_{i_j}^+$ consists of genes that show a relatively higher expression level in cell i than the gene's average expression level across all other cells excluding cell j, and similarly $\mathbf{V}_{i_j}^-$ consists of genes with relatively lower expressions in cell i:

$$\mathbf{V}_{i_j}^+ = \{k | x_{ik} > \frac{(\sum_{l=1}^n x_{lk}) - x_{ik} - x_{jk}}{n-2}\}, \tag{1}$$

$$\mathbf{V}_{i_j}^- = \{k | x_{ik} < \frac{(\sum_{l=1}^n x_{lk}) - x_{ik} - x_{jk}}{n-2}\}. \tag{2}$$

Using $\mathbf{V}_{i_j}^+$ and $\mathbf{V}_{i_j}^-$, we further denote the differential status for gene k in cell i as:

$$U_{i_j k} = \begin{cases} 1, & k \in V_{i_j}^+ \\ -1, & k \in V_{i_j}^- \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Similarly, we can define $U_{j_i k}, k = 1, ..., p$ for cell j. Then we define the dissimilarity (or semi-distance) between cell i and cell j by 1 minus the "differentiability correlation" between U_{i_j} and U_{j_i}, expressed as

$$S_{ij} = 1 - \frac{\sum_{k=1}^p (U_{i_j k} - \overline{U_{i_j}})(U_{j_i k} - \overline{U_{j_i}})}{\sqrt{\sum_{k=1}^p (U_{i_j k} - \overline{U_{i_j}})^2} \sqrt{\sum_{k=1}^p (U_{j_i k} - \overline{U_{j_i}})^2}} \tag{4}$$

where $\overline{U_{i_j}} = \sum_{k=1}^p U_{i_j k}/p$ and $\overline{U_{j_i}} = \sum_{k=1}^p U_{j_i k}/p$. Note that the differentiability correlation $(1 - S_{ij})$ relies on U_{i_j} and U_{j_i}, which were defined through comparing cells i and j against all other cells in the population, leading to an evaluation of the expression relationship between cells i and j through a global perspective. This measure shares many nice properties of the SIDEseq measure in (Huang *et al*., 2017), as both methods are able to incorporate information from all cells when evaluating the similarity between any two cells. But compared to SIDEseq, this measure is expected to be even more robust against cell heterogeneity and data noise since it considered the relationship of cell-specific gene expression patterns over the cell populations. This measure is also related to Gamma coefficient (Goodman., 1954), where expression patterns resemble the order vector. However, our measure calculates the correlation between two pattern vectors while the gamma coefficient considers the ratio of different order number.

2.2 Determining the number of clusters based on variance analysis

If we determine the number of clusters by simply using a divisive (or top-down) hierarchical clustering algorithm (HCA), all cells start in one cluster, and splits are performed recursively as one moves down the hierarchy until every cell is separated. However, to effectively determine the number of clusters in HCA, we here propose an optimal cutting of the HCA dendrogram based on variance analysis.

2.2.1 Review of Variance Analysis

Let $\mathbf{Y}_{ij}$ denote the random response for the ith ($i = 1, 2, \ldots, n_i$) observation in the jth ($j = 1, 2, \ldots, s$) treatment group. Assume $\mathbf{Y}_{ij}$ follows the following linear model:

$$\mathbf{Y}_{ij} = \mu + \delta_j + \epsilon_{ij},$$

where μ represents the average effect which is common to all treatments, δ_j denotes the jth treatment effect (with constraint $\sum_{j=1}^s n_j \delta_j = 0$), and ϵ_{ij}'s are the i.i.d. random error terms that are distributed as $N(0, \sigma^2)$.

Based on the above model, we can determine the differences among various treatment groups by the analysis of variance (ANOVA) followed by an F-test with $H_0 : \delta_1 = \delta_2 = \ldots = \delta_s = 0$. The ANOVA is based on the decomposition of total variance in data into within-group variance (considered error) and between-group variance (considered meaningful difference among group means). : $SST = SSW + SSB$ where

$$\begin{cases} SST = & \sum_{j=1}^{s}\sum_{i=1}^{n_j}(\mathbf{Y}_{ij} - \overline{\mathbf{Y}})^2 \\ SSW = & \sum_{j=1}^{s}\sum_{i=1}^{n_j}(\mathbf{Y}_{ij} - \overline{\mathbf{Y}}_{.j})^2 \\ SSB = & \sum_{j=1}^{s} n_j(\overline{\mathbf{Y}}_{.j} - \overline{\mathbf{Y}})^2 \end{cases} \quad (5)$$

The F-test statistic is defined as $\frac{SSB/(s-1)}{SSW/(n-s)}$, the ratio of two measures of variability. When H_0 is true, or when there are no significant differences among the means of various treatment groups, the *average variability between groups* ($\frac{SSB}{s-1}$) will be expected to be comparable to the *average variability within groups* ($\frac{SSW}{(n-s}$) and so the test statistic will likely be close to 1. Under the null model described above, the test statistic follows an F-distribution with degrees of freedom $(s-1, n-s)$. When H_0 is not true, the variability between groups will likely be large and so the test statistic will tend to be much larger than 1.

2.2.2 Optimal cluster number

Motivated by variance analysis, we here introduce a new measure for deciding an optimal number of clusters in HCA.

Following the notations in Section 2.1, we denote the expression profile for gene j across n cells by $(x_{1j}, ..., x_{nj})$. Now assume there are s cell subpopulations and each cell belongs to one and only one subpopulation. We model the expected expression value of gene j in cell i by $E(X_{ij}) = \alpha_j + \sum_{k=1}^{s} z_{ik}\beta_k$, where z_{ik} is a binary membership indicator (i.e., $z_{ik} = 1$if cell i belongs to subpopulation k, and $z_{j,k} = 0$ otherwise). For each gene j, treating each cell subpopulation as a treatment group, we can define SST_j, SSB_j and SSW_j analogously as (5). If the s cell subpopulations are well separated, $r_j = \frac{SSB_j}{SST_j}$ is likely to be large.

Based on m pre-selected genes, we determine a potentially optimal number of clusters (denoted by C_m) under HCA by checking the changes of $R = \sum_{j=1}^{m} r_j$ as s increases (or equivalently, as the cutting threshold of the dendrogram decreases): C_m is the one when the first local maximum achieves. We compute R based on m top ranked differentially expressed genes across all the cells. To achieve a stable result, we repeatedly compute R and determine C_m across many different choices of m and the set of pre-selected genes. We finally determine a stable optimal number of clusters C_{opt} as the most frequent among many C_m's we obtained.

2.2.3 Computational complexity of the algorithm

In this algorithm, one major computation lies in the differential gene set computation, which is of computational complexity $O(p)$. On the other hand, permuting the algorithm for all possible pairs of cells, we can see that computational complexity of the algorithm is of $O(n^2p)$ which depends on the number of attributes (e.g. genes) as well as the number of cells involved and thus is relatively time-consuming.

Note that differential sets $\mathbf{V}_{ij}^{+}$ and $\mathbf{V}_{ij}^{-}$ between cell i and cell j depend on both i and j, which take major computational cost, because we intend to isolate cell i and cell j and compare them with the remaining cells. However, if we relax such a condition by making them only depend on i, we can simply replace or approximate Eqns.(1)-(2) with $\mathbf{V}_i^{+} = \{k|x_{ik} > \frac{(\sum_{l=1}^{n} x_{lk}) - x_{ik}}{n-1}\}$, $\mathbf{V}_i^{-} = \{k|x_{ik} < \frac{(\sum_{l=1}^{n} x_{lk}) - x_{ik}}{n-1}\}$ when n is relatively large. Clearly when n is large, $\frac{(\sum_{l=1}^{n} x_{lk}) - x_{ik} - x_{jk}}{n-2}$ is very similar to $\frac{(\sum_{l=1}^{n} x_{lk}) - x_{ik}}{n-1}$. This kind of amendment in the algorithm would largely shorten the time complexity from $O(n^2p)$ to $O(np)$ which makes large scale clustering feasible. For such a case, each set is not symmetric to i and j, i.e., generally $\mathbf{V}_i^{+} \neq \mathbf{V}_j^{+}$ and $\mathbf{V}_i^{-} \neq \mathbf{V}_j^{-}$ with respect to cell-pair (i, j).

The algorithm of the proposed method '**Corr**' is illustrated in the following with flowchart attached in Fig.S1 in supplementary file1.

3 Results

To evaluate our algorithm '**Corr**' with differentiability correlation and variance based on hierarchical clustering for single cell RNA-Seq data, we introduce the following state-of-the-art algorithms for comparison:

Algorithm 1 Framework of our algorithm '**Corr**' based on differentiability correlation and variance.

Input: The set of single cells, $P_n \in R^{n\times p}$;

Output: Cell types for the set of single cells, $L_n \in R^{n\times 1}$;

Evaluate cell-pair (i, j) differentiability correlation for $i, j \in \{1, 2, \ldots, n\}$;

Construct dissimilarity measure based on cell-pair differentiability correlations by eqn.(4);

Determine Cluster Number C_{opt} based on variance analysis;

Hierarchical Clustering with the above dissimilarity measure and cluster number.

Return: L_n;

- **SNN-Cliq** (Xu *et al*., 2015) is a recently proposed graph theory based clustering method that utilizes the concept of shared nearest neighbor (SNN) to define cell similarity. The clustering output on single cell RNA-seq data is highly in accordance with the cell type origins.
- **Partitioning Around Medoids (PAM)** (Teschendorff and Enver., 2017) is the most common realization in k-medoids clustering. In the algorithm, centers are chosen from the given data points and a generalization of the Manhattan Norm is used to define distance between data points. The average silhouette width is introduced to determine the optimal number of clusters after 200 rounds of k-medoids clustering are done. The final best clustering result is then generated with the selected optimal number.
- **K-means Clustering** is the most commonly used algorithm in clustering. Supposing that there are n observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, K-means clustering aims to cluster the data set into K clusters $S = \{S_1, S_2, \ldots, S_K\}$ to minimize within cluster variance which can be described as the following optimization problem:

$$\arg\min S \sum_{i=1}^{K} \sum_{\mathbf{x}\in S_i} ||\mathbf{x} - \mu_{\mathbf{i}}||^2 = \sum_{i=1}^{K} |S_i| \text{var} S_i$$

  where $\mu_{\mathbf{i}}$ is the mean of points in S_i. In this algorithm, the optimal number of clusters is determined using the same rule as PAM.
- **Hierarchical Clustering with Euclidean Distance** Euclidean distance is the most frequently used measure in data (dis)similarity evaluation. We incorporate it as a comparison partner and the optimal cluster number is determined using the same method as proposed in '**Corr**'.
- **Hierarchical Clustering with Spearman Distance** Spearman distance evaluates data (dis)similarity from a correlation perspective. We also incorporate it as a comparison partner and the optimal cluster number is determined using the same method as proposed in '**Corr**'.

We report the results of various methods for the considered data sets. The data sets used can be downloaded from NCBI (National Center for Biotechnology Information) GEO, and we will have the detailed descriptions on the data sets. The clustering results are summarized in Fig.1.

- **Islet single cells**
  Six known human islet cell types (alpha cells, beta cells, delta cells, pp cells, acinar cells and duct cells) were identified from this single-cell RNA-Seq data set based on the expression of known marker genes (Li *et al*., 2016). In total, there are 72 single cells involved, 12 of which are of unknown types, including 2 delta cells. Hence we exclude 12 undefined single cells, leaving 60 single cells for verification of the methods. In the data set, there are 18 alpha cells, 12 beta cells, 11 acinar cells, 8 duct cells , 2 delta cells and 9 pp cells. Following the filtering method of (Xu *et al*., 2015 and Ramskold *et al*., 2012), we discarded genes with average RPKM less than 20 across all 60 cells, leaving over 4000 genes.
- **Human Cancer Cells**
  The data set from (Ramskold *et al*., 2012) used a single-cell RNA-Seq platform named Smart-Seq for data extraction. RPKM gene expressions are used to quantify the single cells. There are 8 human embryonic stem cells hESC (8 cells), 4 cells from prostate cancer cell lines LNCap (4 cells), 4 PC3 (4 cells), 6 putative melanoma CTCs (6 cells) from peripheral blood, 4 cells from melanoma cell lines SKMEL5 (4 cells), 3 UACC257 cells (3 cells), and 4 cells from bladder cancer cell line T24 (4 cells). Following the filtering method of (Xu *et al*., 2015) and (Ramskold *et al*., 2012), we discarded genes with average RPKM less than 20 across all 33 cells, leaving over 3000 genes.
- **Human Embryo Stem Cells**
  This data set is obtained from (Yan *et al*., 2013) and uses 124 individual cells in various developmental stages from human pre-implantation

embryos. The human embryonic stem cells are extracted using a highly sensitive sequencing technique. The data set covers 7 early developmental stages: metaphase II ooocyte (3 cells), zygote (3 cells), 2-cell-stage (6 cells), 4-cell stage (12 cells), 8-cell-stage (20 cells), morula (16 cells) and late blastocyst at hatching stage (30 cells). The data set also includes an eighth stage of development of primary outgrowth during human embryonic stem cell (hESC) derivation (34 cells). Different from (Xu *et al*., 2015) who only used cells from the first seven early developmental stages, we used all 124 cells for the clustering analysis.

- **Allodiploid embryonic stem cells**
  This data set involves mRNA profiles of allodiploid embryonic stem cell lines from mice and rats, analyzed using the Illumina HiSeq 2000 platform (Xi *et al*., 2016). Single-cell RNA-Seq was conducted to check the transcriptomes of single allodiploid embryonic and differentiated cells. The quality of sequencing reads was assessed using FastQC. Low quality bases were trimmed with a cutoff of Phred quality score 20 using Fastq Quality Trimmer program in FASTX-Toolkit. We focus on allodiploid embryonic stem cell line MR1-1 in this paper where cell types contain G0/G1 stage embryonic stem cell and fibroblast single cells derived from MR1-1 mouse chimaera, labelled by ESC and Mfibroblast.
- **Mouse Embryo Stem Cells**
  This data set refers to the existence of inter-blastomere differences in 2 and 4-cell mouse embryos (Biase *et al*., 2014). The hypothesis has been certified by the related experiments. Within the dataset, 1857 million SMART-Seq reads were generated from 49 single cells composed of nine 1-cell (zygote), 10 midstage 2-cell, and five 4-cell embryos.

3.1 Results for Islet Data

In Islet data set, there are 6 types of cells considered: alpha, beta, acinar, duct, delta and pp. From the results as shown in Fig.1 (A), we found that our proposed algorithm '**Corr**' yields the best result. In the K-medoids algorithm where only 2 clusters are generated, pp cells are regarded as significantly different from other types of cells, and the difference among other types of cells cannot be detected. In the K-means algorithm, results show only 2 clusters in the considered data set. Slightly different from the K-medoids algorithm, K-means algorithm cannot clearly distinguish pp cells from other type of cells, where 2 pp cells are clustered into the remaining cluster. SNN-Cliq finds 11 clusters where delta cells are clustered together with duct cells. Moreover, the algorithm grouped alpha cells into 3 clusters, beta cells into 2 clusters, and pp cells into 3 clusters. On the other hand, our method also has better performance than those hierarchical clustering based algorithms. In the 'Spearman' algorithm, 3 clusters are found, whereas alpha, beta, delta, and most of the pp cells are merged into one single cluster. Duct cells and acinar cells are regarded to be explicitly different from other cells. There are only 2 clusters found in 'Euclidean' algorithm. Similar to the K-means algorithm, the 'Euclidean' algorithm regards almost all types of cells as one single type, with pp cells being divided into two clusters. Our '**Corr**' algorithm has been shown to accurately cluster alpha, acinar, beta, duct and pp cells into different clusters, with only one acinar cell having been wrongly clustered. 3 alpha cells with 1 pp cell are clustered into a single scattered cluster. delta cells are regarded to cluster near pp cells. Looking further at the results of the '**Corr**' algorithm, we found that the wrongly clustered 'acinar' cell was also considered to be isolated from other 'acinar' cells in the SNN-Cliq algorithm, and it was clustered in the 'duct' cell cluster in both algorithms. Comparing all the considered algorithms, we can conclude that the '**Corr**' algorithm can correctly find the right cluster number and further cluster the single cells into the desired clusters. Other algorithms like 'Euclidean' and 'Kmeans' algorithms underestimate the cluster number, while the SNN-Cliq algorithm over-estimates the cluster number within the data set.

Regarding the optimal cluster number determination, we choose the optimal cluster number based on factorial analysis of variance for the three hierarchical clustering based algorithms : '**Corr**', 'Euclidean' and 'Spearman'. The results are shown in Figure S2(a), where the subfigures on the left hand side show the optimal cluster number distribution in '**Corr**', 'Euclidean' and 'Spearman'. The final optimal cluster number is determined as the most frequently appeared one. The subfigures on the right-hand side report the changes in the suggested optimal cluster number when the number of differentially expressed attributes increases. Figure S2(a) shows the distribution of the suggested optimal cluster number when the number of involved differential attributes permutes from 1 to 100 (m=100). In the '**Corr**' algorithm, the most frequently selected cluster number is 6, achieving almost 80% in all the suggested optimal cluster numbers. In the 'Euclidean' algorithm, the most frequently selected cluster number is

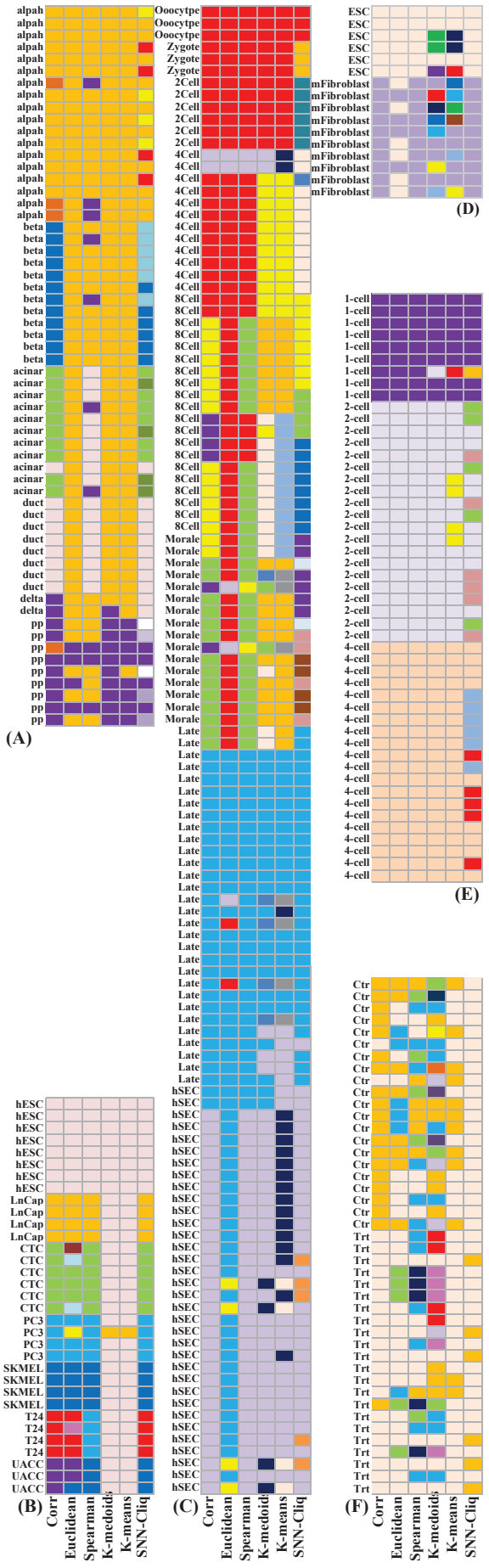


Fig. 1. Clustering Results from different algorithms for considered datasets. The compared algorithms are '**Corr**', 'Euclidean', 'Spearman', 'Kmedoids', 'Kmeans' and 'SNN-Cliq'. The first column under name '**Corr**' stands for our '**Corr**' algorithm, the second column named Euclidean refers to hierarchical clustering with 'Euclidean' distance, the third column represents hierarchical clustering with 'Spearman' distance, the last column stands for SNN-Cliq, and similarly the fourth and fifth columns represent K-medoids clustering and K-means clustering. The column on the leftmost lists the cell type, and the columns indicated with different colors refer to the clustering results of cell type. And the involved datasets are **(A)**: Islet Data; **(B)**: Human Cancer Data; **(C)**: Human Embryo Data; **(D)**: Allodiploid Data; **(E)**:Mouse Embryo Data; **(F)**: Ovarian Cancer Data. In the heatmap, each row stands for an individual cell; each column corresponds to the clustering result produced by one of the six methods. Cells that are grouped in the same cluster by a method are displayed in the same color in the column.

2, achieving almost 80% in all the suggested optimal cluster numbers. In the 'Spearman' algorithm, the most frequently selected cluster number is 3, achieving around 70% in all the suggested optimal cluster numbers. Figure S2(b) shows the optimal number determination for K-medoids and K-means algorithms. For K-medoids algorithm, 10 runs of k-medoids clustering were conducted on islet data to generate a relatively robust and stable result. The top left subfigure corresponds sihullate value distribution and the top right subfigure corresponds the corresponding average sihullate value distribution. The figure shows that the optimal cluster number is 2 for 'K-medoids' algorithm. Similarly we found that the optimal cluster number for 'K-means' algorithm is 2 as well.

Further analysis at the hierarchical clustering results for islet data are shown in Fig.S3. '**Corr**' and 'Spearman' algorithms obviously outperform 'Euclidean' algorithm as we could not find tightly clustered cells using 'Euclidean' algorithm. In the '**Corr**' and 'Spearman' algorithms, several tight clusters can be found. Employing variance analysis, we determined the optimal cluster number for each algorithm. We found that '**Corr**' algorithm is the most accurate with only 1 pp cell and 3 alpha cells clustered in the right hand side as noise clusters. When we further assessed the hierarchical relationships among the clusters, we found that acinar and duct cells are more closely clustered. Studies in mice have demonstrated that acinar cells can transdifferentiate to ductal cells (Mukhi *et al.*, 2011) because of inherited relationship between those types of cells. Alpha , beta , delta and pp cells were more closely clustered in '**Corr**' algorithm. We further found that alpha, beta and delta cells are actually T cells. This further illustrates that '**Corr**' algorithm can help to determine the cell types and hierarchical relationship among the clusters.

3.2 Results for Human Cancer Data

There are 7 types of cells in Human Cancer Data set. Following the data preprocessing (Xu *et al.*, 2015), we selected attributes with RPKM $\geq$ 20 in at least one cell for data analysis. Considering SNN-Cliq algorithm, log transformation ($\log_2(x+1)$) was introduced to reduce the effect of highly expressed genes. However, we did not do any transformation on the data in our '**Corr**' algorithm.

Looking at the clustering results for different algorithms (Fig.1(B)), we made some conclusions. Using '**Corr**' algorithm, the resulting cluster number is 7 with 7 clusters each corresponding to a unique cell type. Each cell type was clearly identified. For 'Euclidean' algorithm, the number of clusters is defined to be 10 after calculation. There are a number of single data clusters, and for 'CTC' cell line alone, 3 clusters are detected. For 'PC3', 'UACC' and 'T24' cell lines respectively, 2 clusters are identified. For 'Spearman' algorithm, the clustering results were better than 'Euclidean' algorithm, with 5 clusters identified. Similar to 'Corr' algorithm, 'hESC', 'LnCap' and 'CTC' types were successfully clustered. But the algorithm groups 'PC3' and 'T24' in the same cluster, 'SKMEL' and 'UACC' in the same cluster. Using SNN-Cliq algorithm, 6 clusters were yielded, with SKMEL5 and UACC257 cells grouped into one single cluster. Note that SKMEL5 and UACC257 are melanoma cell lines and the difference between them should be relatively small. Hence, SNN-Cliq could not detect the slight differences between the two related cell lines. Notably, '**Corr**' algorithm could capture the slight difference between these two types, successfully splitting them into two different types. K-medoids algorithm and K-means algorithm showed only 2 clusters, with all but one cells grouped into a cluster, leaving one 'PC3' cell as a singleton.

Figure S4 illustrates the determination of optimal cluster numbers for all algorithms excluding 'SNN-Cliq'. In Figure S4(a), the subfigures on the left hand side show the optimal cluster number distribution using '**Corr**', 'Euclidean' and 'Spearman' algorithms, and the final optimal cluster number is determined as the most frequently appeared one. Subfigures on the right hand side report the changes in the suggested optimal cluster number when the number of the involved differentially expressed attributes increases where the number of involved differential attributes permutes from 1 to 100 (m=100). Similarly, on Figure S4, it can be seen that the optimal cluster numbers in '**Corr**', 'Euclidean', 'Spearman', 'K-medoids' and 'Kmeans' algorithms are 7, 10, 5, 2, and 2, respectively.

We further analyzed the hierarchical clustering results for human cancer data as shown in Figure S5. We found that '**Corr**' algorithm can clearly distinguish various cell status, besides 'hESC' cells cluster separately from other types of cells. On the contrary, 'Euclidean' algorithm did not allow to capture cell variability. 'Spearman' algorithm has an advantage over 'Euclidean' algorithm, but we found that 'hESC' cells were not isolated from other cells, rather clustered near T24 and PC3 cells. Further analysis of the hierarchical relationship by '**Corr**' algorithm indicates that hESC cells can be isolated from other types of cells, which is consistent with literature and our usual understanding. SKMEL and UACC are closely clustered as they are from melanoma cell line. PC3 and LNCaP are closely clustered because they are both prostate cancer cells. Notably, the PC3 prostate cancer cells are clustered nearer to the T24 bladder cancer cells than to the other LNCap prostate cancer cells. This might contain significant biological meaning and should be further investigated using gene ontology or functional experimental analysis.

3.3 Results for Human Embryo data

In single cell RNA-Seq clustering analysis for Human Embryo data, attributes with RPKM$>$0.1 in at least one cell are selected. We can see the comparison results for different methods as shown in Fig.1(C). Using '**Corr**' algorithm, the 124 single cells were clustered into 6 major clusters. OOOcytpe, Zygote, 2-cell and 4-cell at very early developmental stages were clustered into a single cluster. 8-cell and Morule were clustered into another single cluster. hESC cells were well separated from late blastocyst cells. However, 4 8-cell cells and 2 Morule cells were wrongly clustered. Results received using 'Spearman' algorithm and '**Corr**' were similar where OOOcytpe, Zygote, 2 cell and 4-cell in very early developmental stages were clustered into a single cluster. The number of clusters is 5, 'Late' and 'hESC' cell types were well distinguished and most of the '8-cell', 'Morule' cells are merged into one single cluster, with 2 'hESC' cells and two 'Morule' cells separately designated as one cluster. Using 'Euclidean' algorithm, we detected only 4 clusters, while the rest of the cells were pooled in by 2 clusters. Morule cells were clustered into a single cluster. Moreover, 4 'hESC' cells and 2 'Late' cells separately were designated as one cluster as well. For SNN-Cliq algorithm, the default parameters were introduced $r = 0.7, m = 0.5$; that the final clustering yielded 13 clusters. The method could distinguish cells at the very early stage of development: OOOcytpe, Zygote and 2-cell. However, a number of clusters were generated within 8-cell, Morule and hESC stage cells separately. Late cells were separately clustered. The algorithm could not clearly differentiate stem cells from other type of cells.

In K-medoids algorithm, 9 clusters were detected and 4-cell stage cells were well separated from other cells. The 'hESC' cells were partitioned into two clusters, very similar to SNN-Cliq algorithm. Similar to SNN-Cliq, K-medoids algorithm could not clearly identify Morule cells, generating a number of noise clusters. Using K-means algorithm, the cluster number was 9, OOOcytpe, Zygote and 2-cell cells were grouped into one cluster, and 4-cell stage cells were well separated. For cells in other stages, more than two clusters were detected even in a single stage, e.g., for hESC cells, more than 3 clusters were observed.

The cluster number determination and hierarchical clustering results are shown in Figures S6 and S7, respectively.

3.4 Results for allodiploid data

The results of Allodiploid data analysis are summarized in Fig.1(D). 16 cells were assessed, of which 6 are embryonic stem cells and 10 are mouse fibroblast cells. Using '**Corr**' algorithm, we could perfectly partition the cells into two separated groups. SNN-Cliq and 'Spearman' algorithm generated the same results. Other methods brought unsatisfactory results. The cluster number was 2 for Euclidean Hierarchical Clustering, but the majority of mouse fibroblast cells were included into the 'ESC' cells cluster. The cluster number for 'K-medoids' algorithm and 'Kmeans' algorithm were both 10, generating a number of noise clusters. For mouse fibroblast cells cluster that includes only 10 cells, both algorithms detected 7 clusters.

The determination of optimal cluster number was shown in Fig.S8, where the number of involved differential attributes permutes from 1 to 100(m=100). Fig. S8(a) shows that the optimal number is 2 for '**Corr**', 'Euclidean', and 'Spearman' algorithms . Figure S8(b) records the average silhouette width of K-medoids clustering and K-means clustering algorithm for MR11 cell line. The silhouette value is a measure of similarity for cell in its own cluster compared to cells in other clusters, ranging from -1 to 1. Hence, a large average silhouette value would indicate a better clustering result. For K-medoids clustering in MR11, optimal cluster number is 10:(9+1) and for K-means clustering is 10:(9+1).

Figure S9 reports the results on the hierarchical clustering algorithms including '**Corr**', 'Euclidean', and 'Spearman'. Looking at the hierarchical clustering result for each method, we found that for '**Corr**' algorithm helps to separate cell types clearly. Alternatively, clusters are not so tightly grouped using 'Euclidean' algorithm. Variance analysis finally indicates the number to be 2 and most of the fibroblast cells were clustered near hESC cells under

'Euclidean' distance measure. Using 'Spearman' algorithm, Embryo stem cells were clustered separately as shown in Fig.S9. Therefore, '**Corr**' and 'Spearman' algorithms give a correct description of the relationship between fibroblast cells and ESC cells, while 'Euclidean' algorithm failed to do so. Looking at the hierarchical structure provided by '**Corr**' and 'Spearman' algorithms, we observed that the structure of fibroblast cell cluster differs from each other, which may require further investigation.

3.5 Results for mouse embryo stem cells

For mouse embryo stem cells, hierarchical clustering based methods performed similarly and the optimal cluster number determined was uniformly 3, as shown in Fig.1(E). The number of involved differential attributes equals 100 (m=100). In addition, '**Corr**', 'Euclidean' and 'Spearman' correctly distinguished the 3 types of cells. In contrast, for 'Kmeans' and 'Kmedoids' algorithms, the optimal cluster number were 4 and 3, respectively. 'Kmedoids' algorithm misclassified one 1-cell to 2-cell cluster. For 'Kmeans' algorithm, 2-cell cluster was divided into two different groups. For 'SNN-Cliq' algorithm, the clustering result was not satisfactory. The total number of clusters was 7, although there were only 3 truly different clusters. The corresponding graphical results are presented in Fig.1(E). And the number determination process is indicated on Figure S10 in supplementary file1. Regarding the hierarchical clustering results, we could find differences from Fig.S11. '**Corr**' and 'Euclidean' algorithms grouped 1-cell and 2-cell more closely located, and 'Spearman' algorithm considered 2-cell and 4-cell types more similar than others. It was shown previously in (Yan *et al*., 2013) that inter-blastomere differences between 2- and 4-cell mouse embryos exist. Therefore, the hierarchical structures presented using '**Corr**' and 'Euclidean' algorithms are more reasonable.

4 Discussions

4.1 Determination Methods of Optimal Cluster Number

We proposed variance analysis based algorithm for optimal cluster number determination in our 'Corr' method. State-of-the-art methods in optimal cluster number determination include Calinski-Harabasz index (Calinski and Harabasz, 1974), where a method for identifying clusters of points in a multidimensional Euclidean space is described and a dendrite method for cluster analysis is proposed. Davies-Bouldin index (Davis and Bouldin, 1979) proposed a measure indicating the similarity of clusters. The clusters were assumed to have a data density defined as a decreasing function of distance from a vector characteristic of the cluster. This measure can be used to infer the appropriateness of data partitions and can therefore be used to compare relative appropriateness of various divisions of the data. One of the frequently used measure, the average silhouette width (Rouseeuw, 1987), provides an evaluation of clustering validity, and can be used to select an appropriate number of clusters. We compared our variance-analysis-based measure with the above 3 measures for determining an optimal number of clusters. The comparison results are included in Table 1 with detailed illustrations in supplementary files. It can be seen that in the presented examples, our new measure always works better than or at least equivalent to the other three measures. In particular, Silhouette and Davies-Bouldin tend to get more clusters. Calinski-Harabasz (CH index) is more comparable to ours; this is expected since both measures are based on variance decomposition (or ANOVA). However, for its best performance, CH index requires a constant variance across the subpopulations (or individual clusters), for which scRNA-seq data usually do not have such a property. We consider this the main reason for the outperformance of our measure over CH index.

4.2 External clustering evaluation measures

In addition to the clustering results with internal criteria such as clustering number and accuracy, we introduced external evaluation measures (i.e. purity, adjusted rand index (ARI), and F1-measure) on various scRNA-Seq methods as suggested in (Xu *et al*., 2015) for comparing the clustering ability.

Table 2 lists the evaluation measures on representative scRNA-Seq methods in islet data, human cancer and human embryo data, Allodiploid data with cell line MR11 and mouse embryo data. The graphical results are shown in Figure S12, supplementary file1. The results for our '**Corr**' method were marked in bold font. Clearly, our '**Corr**' algorithm demonstrated the best overall performances when compared to other algorithms applied to all scRNA-seq datasets. For example, in Figure S12 (a) and Table 2 for Islet Data, we can see that '**Corr**' algorithm outperforms all the other algorithms including SNN-Cliq.

Row indicated with Islet(Table 2) demonstrates that ARI and F1-measure for '**Corr**' algorithm are clearly superior to other methods except for purity (Purity is slightly less than that of SNN-Cliq). Although SNN-Cliq is slightly better than '**Corr**' algorithm in purity, but from the other two robust measures 'ARI' and F1-measure, we can see that '**Corr**' algorithm is better. The reason why 'SNN-Cliq' algorithm has a larger purity value is probably that 'SNN-Cliq' algorithm generates more clusters (11 clusters) than the real number of analyzed cell types(6 types of cells).

In human cancer cells data as shown in Table 2 and Figure S12(b), we can see that '**Corr**' algorithm clearly identifies all the cell identities and hence the three considered measures in '**Corr**' algorithm are uniformly 1. The second best algorithm for this data set is 'SNN-Cliq' algorithm, while 'K-medoids' and 'Kmeans' algorithms seem not suitable for handling this data set.

For human embryo cells, however, we that 'Euclidean' and 'Spearman' algorithms could not capture the cell variability. 'K-medoids' algorithm is better than 'K-means' algorithm. Regarding the other four algorithms, we found that 'SNN-Cliq' algorithm delivers the largest purity value , but '**Corr**' algorithm has the largest 'ARI' value and F1-measure. Despite SNN-Cliq produces the highest purity value, we found that it has too many clusters, making each cluster having few samples, increasing the purity of the algorithm. Hence, large value of purity does not correspond to good performance of the method. If we combine the three measures together (either by arithmetic mean or by geometric mean), '**Corr**' outperforms other algorithms including SNN-Cliq. There is no dominant algorithm in discriminating the cell types within the data set. Conclusively, '**Corr**' and 'SNN-Cliq' algorithms perform relatively better.

Table 2 shows the external measure for the 6 algorithms in Allodiploid data with cell line MR11. It was found that '**Corr**', 'Spearman' and 'SNN-Cliq' algorithms can correctly determine the cell types within the data set, yielding 100% accuracy for all considered measures. But for 'Euclidean', 'K-medoids' and 'K-means' algorithms, the results are not satisfactory.

Figure S12(e) shows the external measure for the 6 algorithms in mouse embryo data. It can be found that '**Corr**', 'Spearman' and 'Euclidean' algorithms can correctly determine the cell types within the data set, yielding 100% accuracy in all considered measures, which implies that variance analysis based hierarchical clustering is a reliable option in cell type determination in this case. But for 'K-medoids' and 'K-means' algorithms, the results are relatively inferior although 'K-medoids' is better than 'K-means'. 'SNN-Cliq' algorithm is not satisfactory, in particular for ARI and F1-Measure.

Table 1. Optimal Cluster Number Determination Method

Datasets \ Methods	Measures	'**Variance Analysis**'	'Calinski-Harabasz'	'Silhouette'	'Davies-Bouldin'
Islet	Number	**6**	6	7	7
	Purity	**0.9333**	0.9333	0.9333	0.9333
	ARI	**0.8289**	0.8289	0.7703	0.7703
	F1-Measure	**0.9102**	0.9102	0.8817	0.8817
Human Cancer	Number	**7**	10	10	10
	Purity	**1**	1	1	1
	ARI	**1.0000**	0.8337	0.8337	0.8337
	F1-Measure	**1.0000**	0.9307	0.9307	0.9307
Human Embryo	Number	**6**	2	10	2
	Purity	**0.8871**	0.4355	0.8871	0.4355
	ARI	**0.7129**	0.3295	0.4865	0.3295
	F1-Measure	**0.7964**	0.5097	0.6879	0.5097
Allipoloid	Number	**2**	2	10	10
	Purity	**1.0000**	1.0000	1.0000	1.0000
	ARI	**1.0000**	1.0000	0.2667	0.2667
	F1-Measure	**1.0000**	1.0000	0.5833	0.5833
Mouse Embryo	Number	**3**	2	10	3
	Purity	**1.0000**	0.81630	1.0000	1.0000
	ARI	**1.0000**	0.6951	0.6718	1.0000
	F1-Measure	**1.0000**	0.8284	0.7927	1.0000
Ovarian TGFB	Number	**2**	2	10	2
	Purity	**0.9302**	0.9302	0.9302	0.9302
	ARI	**0.7341**	0.7341	0.1603	0.7341
	F1-Measure	**0.9302**	0.9302	0.4887	0.9302

Table 2. External Measures for the considered methods

Datasets \ Methods	Measures	'**Corr**'	'Euclidean'	'Spearman'	'K-medoids'	'K-means'	'SNN-Cliq'
Islet	Purity	**0.9333**	0.3667	0.4500	0.4500	0.3667	0.9500
	ARI	**0.8289**	0.0354	0.2633	0.1532	0.0354	0.5109
	F1-Measure	**0.9102**	0.4047	0.5197	0.4864	0.4047	0.6919
Human Cancer	Purity	**1.0000**	0.8788	0.7879	0.2727	0.2727	0.9091
	ARI	**1.0000**	0.7065	0.8028	0.0057	0.0057	0.9079
	F1-Measure	**1.0000**	0.8023	0.8498	0.2993	0.2993	0.9306
Human Embryo	Purity	**0.8871**	0.4516	0.75	0.8225	0.7983	0.9677
	ARI	**0.7128**	0.2662	0.6149	0.5325	0.4193	0.6405
	F1-Measure	**0.7964**	0.4863	0.7194	0.7112	0.6494	0.7841
Allipoloid	Purity	**1.0000**	0.6250	1.0000	1.0000	1.0000	1.0000
	ARI	**1.0000**	0	1.0000	0.1667	0.1667	1.0000
	F1-Measure	**1.0000**	0.6071	1.0000	0.6071	0.6071	1.0000
Mouse Embryo	Purity	**1.0000**	1.0000	1.0000	0.9796	0.9796	0.9796
	ARI	**1.0000**	1.0000	1.0000	0.9483	0.8532	0.4054
	F1-Measure	**1.0000**	1.0000	1.0000	0.9792	0.9438	0.6711
Ovarian TGFB	Purity	**0.9302**	0.8140	0.6977	0.8140	0.6977	0.6279
	ARI	**0.7341**	0.1857	0.0462	0.0640	0.1397	0.0544
	F1-Measure	**0.9302**	0.6333	0.4651	0.3848	0.6821	0.6311

4.3 Applications

- **Differentiability vector for biomarker detection and prognosis analysis**

As discussed above, differentiability correlation based on gene differential patterns provides a new way for evaluating the relationship between different cells. The method proves to be useful and robust for cell type identification. In this subsection, we will further dig deeper the value of differentiability.

As shown in Section 2, we can reformulate transformed vector U in Eqn.(3) for a given gene expression based single cell data as indicated in the construction of differentiability correlation. The transformed vector can be expressed by differentiability vector or differentiability transformation (DT) as follows:

$$\vec{c}(\mathbf{X_i}) = \frac{1}{\sqrt{\sum_{k=1}^{p}(U_{ik} - \overline{U_i})^2}}(U_{i1} - \overline{U_i}, U_{i2} - \overline{U_i}, \ldots, U_{ip} - \overline{U_i})$$

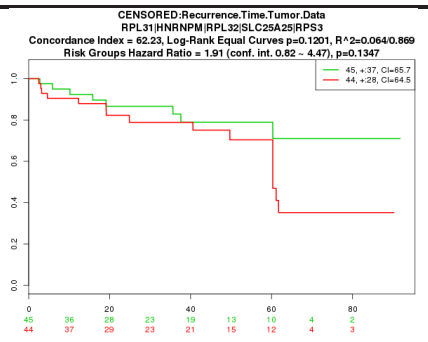
The transformation may provide a new way for data normalization and can help to find some significant gene markers that traditional differential analysis would not be able to identify.

The test data set was derived from (Miyamoto *et al.*, 2015) and it is related to castration resistant prostate cancer. We analyzed a subset of 73 single cells, of which 37 cells are derived from patients receiving no treatment, and the remaining 36 cells were derived from patients who have resistance to enzulatamide treatment. We extracted the differentially expressed genes with DT and without, and then compared them using KEGG pathway analysis. The detailed results are attached in Supplementary file2. The top 100 extracted genes with and without are quite different, where there are 36 overlapped genes. Note that the differentially expressed genes without DT is the traditional differentially expressed genes, but the differentially expressed genes with DT belong to the differentiability vector of genes.

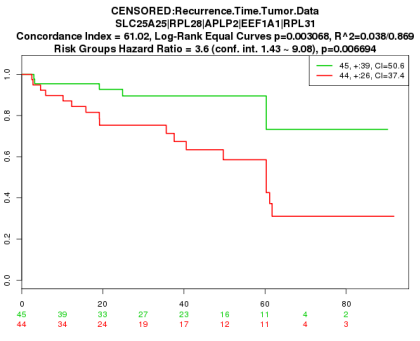
Looking at the pathway analysis results, we can draw the following conclusions. Traditional differential gene analysis can identify some cancer related genes. For instance, DNMT1, DNA (cytosine-5-)-methyltransferase 1, is involved in pathway: MicroRNAs in cancer. GNAI2, which is G protein subunit alpha i2, is involved in pathways in cancer. Another G protein, GNB1(G protein subunit beta 1) is also involved in pathways in cancer. RAC1, RAS-related C3 botulinum substrate 1 is a typical cancer related gene. SLC2A1, which is solute carrier family 2 (facilitated glucose transporter), member 1, is involved in pathways in cancer.

On the other hand, for genes extracted using DT or differentiability vector, we found that GNB1,RAC1 and SLC2A1 are also listed. Besides, we identified some other genes. ARAF, A-Raf proto-oncogene, serine/threonine kinase, is a significant in prostate cancer and involved in pathways in cancer. MCL1, BCL2 family apoptosis regulator, is involved in pathways: MicroRNAs in cancer. GNG10, G protein subunit gamma 10, is involved in pathways in cancer. In addition, two drug metabolism related genes are identified: GSTK1 and GST02, two different versions of glutathione S-transferase. GOLPH3 is identified, and it is involved in transcriptional mis-regulation in cancer. ITPR2, inositol 1,4,5-trisphosphate receptor type 2, is involved in proteoglycans in cancer. All these important genes are identified with DT based analysis while traditional differential gene analysis cannot identify.

We conducted an independent prognostic analysis on the selected markers suggested by both methods where the top 5 ranked genes are selected in both methods. Traditional method selected the top 5 genes: RPL31, HNRNPM, RPL32, SLC25A25, RPS3. And in our differentiability vector method, we selected the top 5 genes as follows: SLC25A25, RPL28, APLP2, EEF1A1, RPL31. The analysis is conducted using SurvExpress (Raul *et al.*, 2013). Fig.2 demonstrates that two methods performed differently. Fig.2 (a) reports the Kaplan Meier curve for traditional method, and Fig.2 (b) reports the Kaplan Meier curve for differentiability vector method. Results show that our selected biomarkers are able to separate risk groups characterized by differences in their gene expression while traditional method failed to do so. For our method, the concordance index was 61.02, the log-rank equal curve p value was 0.003068 and the risk group hazard ratio 3.6 with p value 0.006694, which is significant for the prognosis. In comparison, the concordance index in traditional method is 62.236, the log-rank equal curve p value was 0.1201 and the risk group hazard ratio 1.91 with p value 0.1347, which is not significant. Thus differentiability vector method shows superior performance. Heatmaps and boxplots for both methods are shown in Fig.S13.(a)-(d). All the other supporting files are attached in supplementary file2. In summary, these results indicate that differentiability transformation can provide a new way for biomarker detection and prognosis.



(a) Traditional Method



(b) Differentiability Method

Fig. 2. Prognosis analyses of Kaplan Meier Curve. From the Kaplan Meier Curve in both methods shown in Subfigure (a),(b), we can see that our differentiability method can separate risk groups but traditional method fails to do so.

- **Ovarian Cancer Single Cell Clustering**

We further utilize the new scRNA-Seq dataset consisting of 96 cells derived from the human epithelial ovarian cancer cells line, CAOV-3 (ATCC, Manassas, VA, USA(Huang *et al.*, 2017). The ovarian cancer cells were sequenced in two batches of 48 cells each. One half of the cells in one batch were treated with TGFβ-1, and similarly half of the cells in the second batch were treated with thrombin. The remaining 48 cells in both batches were untreated, control cells. Since TGFβ-1 treatment is a well-studied inducer of EMT, the heterogeneity of the cellular phenotypes resulting from EMT in ovarian cancer cells was thought to likewise lead to an increased ability to escape early detection. We focused on TGFβ-1 batch trying to differentiate treatment cell from control cells. However, it is not an easy task for clearly distinguishing cells at two different conditions. The untreated control cells in both batches should not have significantly different expression profiles, but apparent differences were observed for control cells in the two batches. This is probably associated with technical noise or unwanted biological variability. Therefore, we tested the ability of our method to deal with such noise or biological variability.

Using the differentiability vector (the same as previous subsection) to represent the noisy data, we measured the cell dis-similarity with differentiability correlation and determined cluster number based on variance analysis. The clustering result is reported in Fig.1(F). The clustering numbers determined by the compared 6 methods were 2,4,5,10,2,2 for '**Corr**','Euclidean','Spearman','K-medoids', 'K-means' and 'SNN-Cliq' respectively. We found that our '**Corr**', 'Kmeans' and 'SNN-Cliq' determined the correct cluster numbers. By further analysis on the clustering result, we found that '**Corr**' method yields the best result. It can cluster the cells into two separate types: control and treatment, and only 3 out of the 43 cells were wrongly clustered. For 'Kmeans' and 'SNN-Cliq' algorithms, the clustering results are different, but they share some similarity in that the majority of cells was grouped as one cluster. They can not clearly detect the differences between the two types of cells.

As the number determined by variance analysis was 2, we found that the wrongly clustered 3 cells are: 2 control cells and one treated cell. The purity, adjusted rand index and F1-measure are also reported in Table 2. We found that '**Corr**' algorithm shows the best performance, achieving purity and F1-measure 0.9304. The adjusted rand index was 0.7341. For

other algorithms, the best purity value was 0.8140, in 'Euclidean' and 'K-medoids' algorithms. The best F1-measure value was achieved in 'K-means' clustering algorithm, achieving 0.6821. The best adjusted rand index for other algorithms was achieved in 'Euclidean' algorithm, and was only 0.1857. Figure S14 in supplementary file 1 illustrated cluster number determination in the data set. The hierarchical representations for the '**Corr**','Euclidean' and 'Spearman' algorithms are attached in Supplementary file 1(Figure S15).

From the reported results, we found that '**Corr**' algorithm is the best among the considered methods. In some of the cases, 'SNN-Cliq' shows good performance similarly to '**Corr**' algorithm. However, there are occasions in which 'SNN-Cliq' cannot compete with hierarchical based clustering methods including 'Euclidean' and 'Spearman'. Thus '**Corr**' algorithm can provide an appropriate cell-similarity measure and also identify suitable cluster number.

When constructing our new cell-cell (dis)similarity measure employed in 'Corr', we used the sample mean expression (excluding cell i and cell j) to estimate/approximate the "expected" expression for the cell population. This approximation should be reasonable in general but likely works best for normally distributed data. The '**Corr**' algorithm provides an alternative cell-to-cell measure in DE-vector based correlation construction. More importantly, we also introduced a variance analysis that can systematically determine the number of clusters in the hierarchical clustering analysis.

5 Conclusion

scRNA-Seq analysis provides a new way of cell composition analysis at various developmental stages. As a fundamental problem in grouping individual cells based on their noisy gene expression values, scRNA-seq clustering is developed to understand pathology of developmental processes. We proposed a novel measure of differentiability correlation for evaluating cell dissimilarity that fits the framework of hierarchical clustering. Using variance analysis, we determined optimal cluster number. We therefore propose a new method '**Corr**' for understanding cell functions based on scRNA-seq data even with strong noise or fluctuations.

Our '**Corr**' algorithm is characterized by a list of notable features. Firstly, the algorithm takes into consideration the micro-environment of cell populations in dissimilarity construction, providing a more robust relationship assessment, i.e., cell-pair differentiability correlation. Furthermore, '**Corr**' algorithm incorporates factorial analysis of variance in optimal cluster number determination which is new way in this field. The algorithm can automatically determine the optimal cluster number without a need for number specifications, and it can always suggest a correct cluster number for the respective data set. Notably, there is no requirement for setting extra parameters. '**Corr**' algorithm shows outstanding performance in benchmark or real experimental data sets, handling data with various cluster structures, against various exiting approaches in terms of internal criteria (clustering number and accuracy) and external criteria (purity, ARI, and F1-measure). '**Corr**' algorithm can clearly recognize embryo stem cells and other type of cells in various stages, capturing the cell-to-cell variability. The data normalization based on differentiability transformation (or differentiability vector) proves to be a realiable way to identify significant genes related to the biological processes or phenotypes while traditional differential gene expression fails to do so. And the application in ovarian cancer single cell clustering shows the ability of '**Corr**' algorithm in dealing with noise or unwanted biological variability, was also validated by the prognosis analysis of independent dataset.

Acknowledgements

The authors would like to thank anonymous reviewers for providing valuable comments for our manuscript.

Funding

This research is supported by National key R&D program of China (No. 2017YFA0505500), Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB13040700), and National Natural Science Foundation of China Grant (Nos. 11626229, 91730301, 91529303, 81471047, 31771476).

References

Aguirre-Gamboa, R., Gomez-Rueda, H., Martinez-Ledesma, E., et al. (2013) SurvExpress: An Online Biomarker Validation Tool and Database for Cancer Gene Expression Data Using Survival Analysis, *Plos One*, **8**, e74250.

Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U. (1999) When is "Nearest Neighbor" meaningful? In: Beeri, C. and Buneman, P. (eds) ICDT' 99 Proceedings of the 7th International Conference on Database Theory., Springer-Verlag London, UK., 217-235.

Bhadriraju, K., Chen, C.S. (2002) Engineering cellular microenvironments to improve cell-based drug testing, *Drug Discovery Today*, **7**, 612-20.

Biase, F.H., Cao X., Zhong S. (2014) Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell rna sequencing, *Genome Research*, **24**, 1787.

Bo, W., Zhu, J., Pierson, E., *et al.*, (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature Methods*, **14**, 414.

Caliński, T., Harabasz, J. (1974) A dendrite method for cluster analysis. *Communications in Statistics*, **3**, 1-27.

Cheng, J., Derek, K., Kim, J.Y., Li, M.Y., Zhang, N.R. (2017) Accounting for technical noise in single-cell RNA sequencing analysis, bioRxiv.

Davies D. L., Bouldin D. W. (1979) A Cluster Separation Measure. IEEE Computer Society.

Eberwine, J. et al. (2014) The promise of single-cell sequencing, *Nature Methods*, **11**, 25-27.

Eisenberg E, Levanon EY (2013) Human housekeeping genes, revisited. *Trends in Genetics*, **29**, 569-574.

Ertoz, L., Steinbach, M., Kumar, V. (2003) Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data, In Proceedings of 2nd SIAM Interational Conference on Data Mining.

Federico, M., Giordano, A. (2010) Cancer Stem Cells and Microenvironment. In Cancer Stem Cells and Microenvironment, Springer New York, 169-185.

Gong, Q., Ou, Q., Ye, S., et al. (2005) Importance of cellular microenvironment and circulatory dynamics in B cell immunotherapy, *Journal of Immunology*, **174,** 817-826.

Goodman, L.A., Kruskal, W.H. (1954) Measures of Association for Cross Classifications. *Journal of the American Statistical Association*, **49**, 732-764.

Grun, D., Lyubimova, A., Kester, L., et al. (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types, *Nature*, **525**, 251-255.

Guha, S., Rastogi, R., Shim, K. (1999) ROCK: a robust clustering algorithm for categorical attributes, International Conference on Data Engineering. IEEE Computer Society, 512.

Houle, M.E. et al . (2010) Can shared-neighbor distances defeat the curse of dimensionality? Scientific and Statistical Database Management: 22nd International Conference, SSDBM 2010, Heidelberg, Germany, June 30-July 2, 2010. Springer Berlin Heidelberg, **6187**, 482-500.

Rouseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65.

Kiselev, V.Y., Kirschner, K., Schaub, M. T., et al. (2017) SC3: consensus clustering of single-cell RNA-seq data, *Nature Methods*, **14**,483.

Levine, J.H., Simonds, E.F., Bendal, S.C. et al., (2015) Data-driven phenotypic disection of AML reveals progenitor-like cells that correlate with prognosis, *Cell*, **162**, 184-197.

Li, J., Klughammer, J., Farlik, M., Penz, T. et al. (2016) Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Reports*, **17**,178-87.

Li, X., Cui, X.L., Wang, J.Q., et al. (2016) Generation and Application of Mouse-Rat Allodiploid Embryonic Stem Cells, *Cell*, **164**, 279-92.

Miyamoto, D.T.,Zheng, Y., Wittner,B.S. et al. (2015) RNA-seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science*, **349**, 1351-1356.

Mukhi, S., Brown, D. D. (2011) Transdifferentiation of tadpole pancreatic acinar cells to duct cells mediated by Notch and stromelysin-3. *Developmental Biology*, **351**, 311-317.

Ramskold, D.,, Luo S., Wang, Y.C., et al. (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, **30**, 777-82.

Schiffman, C., Lin, C., Shi, F., et al.(2017) SIDEseq: A Cell Similarity Measure Defined by Shared Identified Differentially Expressed Genes for Single-Cell RNA sequencing Data, *Statistics in Biosciences*, **9**, 200-216.

Shapiro, E., Biezuner, T., Linnarsson,S. (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, **14**, 618-630.

Stegle, O., Teichmann, S. A., Marioni, J. C. (2015) Computational and analytical challenges in single-cell transcriptomics, *Nature Reviews Genetics*, **16**, 133-145.

Teschendorff, A.E., Enver, T. (2017) Single-cell entropy for quantification of differentiation potency from a cell's transcriptome. *Nature Communications*, **8**, 15599.

Xu, .C, Su, Z. (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method, *Bioinformatics*, **31**, 1974-80.

Yan, L., Yang, M., Guo, H. et al. (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells, *Nature Structural and Molecular Biology*, **20**, 1131-1139.

Yuan, G.C., Cai, L., Elowitz, M., et al. (2017) Challenges and emerging directions in single-cell analysis, *Genome Biology*, **18**, 84.