Briefings in Bioinformatics, 00(00), 2018, 1–11

doi: 10.1093/bib/bby076
Advance Access Publication Date: 22 August 2018
Review article

Impact of similarity metrics on single-cell RNA-seq data clustering

Taiyun Kim, Irene Rui Chen, Yingxin Lin, Andy Yi-Yang Wang, Jean Yee Hwa Yang and Pengyi Yang

Corresponding author: Pengyi Yang, Charles Perkins Centre, The University of Sydney, Sydney, NSW 2006, Australia. Tel.: +61 2-9351-3039; Fax: +61 2-9351-4534; E-mail: pengyi.yang@sydney.edu.au

Abstract

Advances in high-throughput sequencing on single-cell gene expressions [single-cell RNA sequencing (scRNA-seq)] have enabled transcriptome profiling on individual cells from complex samples. A common goal in scRNA-seq data analysis is to discover and characterise cell types, typically through clustering methods. The quality of the clustering therefore plays a critical role in biological discovery. While numerous clustering algorithms have been proposed for scRNA-seq data, fundamentally they all rely on a similarity metric for categorising individual cells. Although several studies have compared the performance of various clustering algorithms for scRNA-seq data, currently there is no benchmark of different similarity metrics and their influence on scRNA-seq data clustering. Here, we compared a panel of similarity metrics on clustering a collection of annotated scRNA-seq datasets. Within each dataset, a stratified subsampling procedure was applied and an array of evaluation measures was employed to assess the similarity metrics. This produced a highly reliable and reproducible consensus on their performance assessment. Overall, we found that correlation-based metrics (e.g. Pearson's correlation) outperformed distance-based metrics (e.g. Euclidean distance). To test if the use of correlation-based metrics can benefit the recently published clustering techniques for scRNA-seq data, we modified a state-of-the-art kernel-based clustering algorithm (SIMLR) using Pearson's correlation as a similarity measure and found significant performance improvement over Euclidean distance on scRNA-seq data clustering. These findings demonstrate the importance of similarity metrics in clustering scRNA-seq data and highlight Pearson's correlation as a favourable choice. Further comparison on different scRNA-seq library preparation protocols suggests that they may also affect clustering performance. Finally, the benchmarking framework is available at http://www.maths.usyd.edu.au/u/SMS/bioinformatics/software.html.

Key words: single-cell RNA-seq; scRNA-seq; clustering; similarity metric; distance; correlation

Taiyun Kim is a PhD candidate at the University of Sydney. His research interest is in the area of integrative single-cell omics analysis and its applications in precision medicine.
Irene Rui Chen is a research assistant at Judith and David Coffey Life Lab, Charles Perkins Centre, the University of Sydney. Her research interest is in the broad area of bioinformatics.
Yingxin Lin is a PhD candidate at the University of Sydney. Her research interest is in normalisation and statistical modelling of single-cell RNA-seq data.
Andy Yi-Yang Wang is an MD and clinical senior lecturer at the Sydney Medical School, and a PhD candidate at the School of Mathematics and Statistics, University of Sydney. His research interest is in precision medicine.
Jean Yee Hwa Yang is a professor at the School of Mathematics and Statistics, University of Sydney, and an NHMRC CDF Fellow. She is an applied statistician with expertise in statistical bioinformatics. She was awarded the 2015 Moran Medal in statistics from the Australian Academy of Science in recognition of her work on developing methods for molecular data arising in cutting-edge biomedical research. As a statistician who works in the bioinformatics area, she enjoys research in a collaborative environment, working closely with scientific investigators from diverse backgrounds.
Pengyi Yang is a senior lecturer and an Australian Research Council Discovery Early Career Researcher Award Fellow at the School of Mathematics and Statistics, University of Sydney. He received his PhD from the School of Information Technologies, University of Sydney, in 2012. His was a Research Fellow at National Institutes of Health, USA, from 2012 to 2015 before joining the School of Mathematics and Statistics at the University of Sydney. His research is at the interface of machine learning and data mining, and their applications to bioinformatics and computational biology. For his scientific contribution to the interdisciplinary research in bioinformatics, he was awarded the JG Russell Award from the Australian Academy of Science.
Submitted: 12 June 2018; **Received (in revised form):** 1 August 2018

Introduction

Single-cell RNA sequencing (scRNA-seq) has become the method of choice for profiling transcriptomes of individual cells from complex samples [1, 2]. With the advances of scRNA-seq techniques and their widespread applications, much effort has been directed to methodological development to address computational challenges brought about by the scRNA-seq data [3, 4]. Among these, the characterisation of cell types and/or subpopulation based on their transcriptome signatures is one of the key computational challenges that received much attention [5]. Toward this goal, unsupervised clustering algorithms are frequently applied to partition cells into distinctive clusters to discern cell identity and characterise their function [6, 7]. Subsequently, biological knowledge such as marker genes defined in the literature from published studies is often utilised to precisely define the cell type(s) in each cluster [8] and/or refine clustering results [9–11].

Due to the important role played by the clustering procedure for downstream analysis such as cell-type identification, numerous clustering algorithms have been proposed for scRNA-seq data [12]. These range from simple *k*-means clustering [13], hierarchical clustering [10] and their variants such as RaceID/RaceID2 [14], SC3 [15] and CIDR [16], to more sophisticated methods that utilise likelihood-based mixture modelling (countClust) [17], density-based spatial clustering [18], projection to reference cell types Reference component analysis (RCA) [11] and kernel-based approach such as single-cell interpretation via multi-kernel learning (SIMLR) [19].

Fundamentally, all clustering algorithms mentioned above rely on a similarity metric that compares transcriptome profiles to either assign or partition each cell into different clusters. Similarity metrics [20] can largely be categorised into distance-based metrics (e.g. Euclidean distance) or correlation-based metrics (e.g. Pearson's correlation). Although, previous studies have reviewed and compared different similarity metrics in various machine learning and data mining applications [20–24], very few of them were dedicated to gene expression data analysis. When applied to gene expressions in a scRNA-seq dataset, distance-based metrics capture the level of expression in transcriptome profiles, whereas correlation-based metrics are invariant to scaling and focus primarily on the relativity of gene expressions. While numerous studies have compared the performances of various clustering algorithms proposed for scRNA-seq data analysis [9, 25], studies focusing on assessing the impact of similarity metrics on clustering scRNA-seq data, which is at the core of each clustering algorithm, are lacking.

Here, we present a systematic assessment on the impact of similarity metrics on clustering analyses of scRNA-seq data. Using a collection of well-annotated scRNA-seq datasets, we first benchmarked a panel of widely used similarity metrics that comprised both correlation- and distance-based measures using a standard *k*-means clustering algorithm. A repeated stratified subsampling procedure was introduced to account for the variability in each clustering run. By benchmarking to an array of evaluation measures, we showed that, overall, correlation-based metrics performed better than distance-based metrics. We next sought to test if, by changing the similarity measure from Euclidean distance to Pearson's correlation, an improvement in the performance of a state-of-the-art kernel-based clustering algorithm (SIMLR) can be achieved. Indeed, our result showed that such a modification led to a statistically significant improvement on the performance of its original algorithm, which relies on Euclidean distance in measuring similarities between transcriptome profiles. These findings demonstrate the utility of correlation-based metrics and may provide a guide to future developments of clustering algorithms for scRNA-seq data analysis.

Methods

scRNA-seq data collection

To benchmark the impact of different similarity metrics on clustering of individual cells to their corresponding cell types, we selected scRNA-seq datasets in which each cell was annotated by cell and lineage markers and in some cases also with additional biological information such as morphological, physiological and functional properties in their respective studies. This allowed us to treat the annotation of individual cells from their original studies as gold standards. Datasets that relied purely on clustering for cell-type annotation in their original studies were excluded because they did not have independent cell-type characterisation and therefore could not be used objectively for comparing clustering results.

With the above criteria, we collected 22 scRNA-seq data-sets with cell-type annotations, sourced from three major repositories including Gene Expression Omnibus from the National Center for Biotechnology Information (NCBI-GEO) [26], The EMBL-European Bioinformatics Institute (EMBL-EBI) [27] and Broad Institute single cell databases (https://portals.broadinstitute.org/single_cell) (Table 1). When available, the log2 of transcripts per million or counts per million (CPM) from their original publications was used to quantify full-length gene expression for each dataset. Only raw counts were provided from the Broad Institute for the human and mouse datasets generated by Habib *et al*. [45]. For these, we computed log2 CPM using the edgeR package [47]. Each scRNA-seq dataset was preprocessed by filtering (i) genes that were detected in less than 20% of cells, and then (ii) cells that have less than 1% of genes detected to remove lowly expressed genes and poorly quantified cells. To account for data scaling and zero counts, each dataset was further processed by using Linnorm [48] and SAVER [49]. Additional experiments were also performed on these further processed datasets.

Similarity metrics

Five similarity measures that are widely used by various clustering algorithms and that represent two broad classes of correlation-based metrics and distance-based metrics were chosen in this comparison study. These included Euclidean, Manhattan and maximum distances (distance-based metrics), and Pearson and Spearman's correlation coefficients (correlation-based metrics). Let x_{ig} and x_{jg} denote the expression of a gene $g = 1,...,G$ in cell $i = 1,...,N$ and cell $j = 1,...,N$, where G and N are the total number of genes and cells, respectively. For a distance matrix $D = (d_{ij})$, the element d_{ij} represents the distance between cell i and cell j, which can be calculated as follows:

Euclidean distance,

$$d_{ij} = \sqrt{\sum_{g=1}^{G} \left(x_{ig} - x_{jg}\right)^2};$$

Manhattan distance,

$$d_{ij} = \sum_{g=1}^{G} \mid x_{ig} - x_{jg} \mid;$$

Table 1. scRNA-seq datasets used for clustering comparison. datasets are sorted by the number of profiled cells. TC, time-course; CT, cell types

Source	Publication	Organism	No. of cells	No. of classes
GSE45719	Deng *et al.* [28]	Mouse	300	8
GSE63818	Guo *et al.* [29]	Human	328	37
GSE67835	Darmanis *et al.* [30]	Human	420	8
GSE82187	Gokce *et al.* [31]	Mouse	705	10
GSE75140	Camp *et al.* [32]	Human	734	13
GSE75748 (TC)	Chu *et al.* [33]	Human	758	6
GSE84133	Baron *et al.* [34]	Mouse	822	13
GSE89232	Breton *et al.* [35]	Human	957	4
GSE75748 (CT)	Chu *et al.* [33]	Human	1018	7
GSE94820	Villani *et al.* [36]	Human	1140	5
E-MTAB-4079	Scialdone *et al.* [37]	Mouse	1205	4
GSE84371	Habib *et al.* [38]	Mouse	1402	8
GSE59114	Kowalczyk *et al.* [39]	Mouse	1428	6
E-MTAB-3929	Petropoulos *et al.* [40]	Human	1529	5
GSE93593	Close *et al.* [41]	Human	1733	4
GSE86146	Li *et al.* [42]	Human	2621	45
GSE60361	Zeisel *et al.* [7]	Mouse	3005	7
GSE70630	Tirosh *et al.* [43]	Human	4347	8
GSE72056	Tirosh *et al.* [44]	Human	4645	7
Broad portal	Habib *et al.* [45]	Mouse	13313	26
Broad portal	Habib *et al.* [45]	Human	14963	19
GSE81905	Shekhar *et al.* [46]	Mouse	27499	19

Maximum distance,

$$d_{ij} = \max_g \mid x_{ig} - x_{jg} \mid .$$

Correlation-based metrics are calculated as follows:

1-Pearson's correlation coefficient,

$$d_{ij} = 1 - \frac{\sum_{g=1}^{G} \left(x_{ig} - \bar{x}_i\right)\left(x_{jg} - \bar{x}_j\right)}{\sqrt{\sum_{g=1}^{G} \left(x_{ig} - \bar{x}_i\right)^2}\sqrt{\sum_{g=1}^{G} \left(x_{jg} - \bar{x}_j\right)^2}};$$

1-Spearman's correlation coefficient,

$$d_{ij} = 1 - \frac{\sum_{g=1}^{G} \left(r_{ig} - \bar{r}_i\right)\left(r_{jg} - \bar{r}_j\right)}{\sqrt{\sum_{g=1}^{G} \left(r_{ig} - \bar{r}_i\right)^2}\sqrt{\sum_{g=1}^{G} \left(r_{jg} - \bar{r}_j\right)^2}},$$

where r_{ig} denotes the rank of the expression value of gene g in cell i, and $\bar{x}_i$ and $\bar{r}_i$ define the mean expression and mean rank of the expression of cell i.

k-means and SIMLR clustering algorithms

To assess the impact of similarity metrics performance on clustering algorithms, we first conducted experiments on a standard k-means clustering algorithm with Lloyd's implementation [50]. Given an initial set of random centres $m_1,...,m_K$ and a distance matrix D computed using the above five similarity metrics, the algorithm first finds the closest cluster centres for each of all cells based on their expression profiles $X = x_1,...,x_N$:

$$\text{for } x \in X : c(x) = \text{argmin}_{k=1,...,K}\left\{D\left(x, m_j\right)\right\}$$

Table 2. Confusion matrix for measuring clustering concordance with predefined cell class

		Clustering partition	
		No. of pairs in the same class	No. of pairs in different classes
Cell type/class	No. of pairs in the same class	*a*	*b*
(gold standard)	No. of pairs in different classes	*c*	*d*

and then updates the cluster centres:

$$\text{for } k \in 1,...,K : m_k = \text{mean}\left(\{x \in X | c(x) = k\}\right).$$

The output are the assignments of each cell based on its expression profile x to a cluster $k \in 1,...,K$. We also performed the clustering assessment using a kernel-based approach (SIMLR) proposed by Wang *et al.* [19]. The original implementation of SIMLR uses Euclidean distance as the metric for constructing a Gaussian kernel function as shown below:

$$\mathcal{K}\left(x_i, x_j\right) = \frac{1}{\epsilon_{ij}\sqrt{2\pi}} \exp\left(-\frac{\sqrt{\sum_{g=1}^{G}\left(x_{ig} - x_{jg}\right)^2}}{2\epsilon_{ij}^2}\right),$$

where ϵ_{ij} is the variance and $\sqrt{\sum_{g=1}^{G}\left(x_{ig} - x_{jg}\right)^2}$ is the Euclidean distance between cells i and j, calculated from their expression profiles x_i and x_j across g. To investigate the impact of similarity metrics on SIMLR clustering, we obtained the R source code of SIMLR (Version 1.4.0) from Bioconductor (Release 3.6)

and compared the original implementation that uses Euclidean distance-based Gaussian kernel and a modified version where we replaced the Euclidean distance metric with Pearson's correlation for Gaussian kernel construction.

The parameter k that determines the number of clusters to be created was set according to the number of predefined cell types/classes in each scRNA-seq data.

Cluster evaluation measures

To benchmark the performances of clustering from using the five different similarity metrics, four cluster evaluation measures were employed to quantify concordance of clustering results on each scRNA-seq dataset with respect to their predefined cell-type annotations. These included normalised mutual

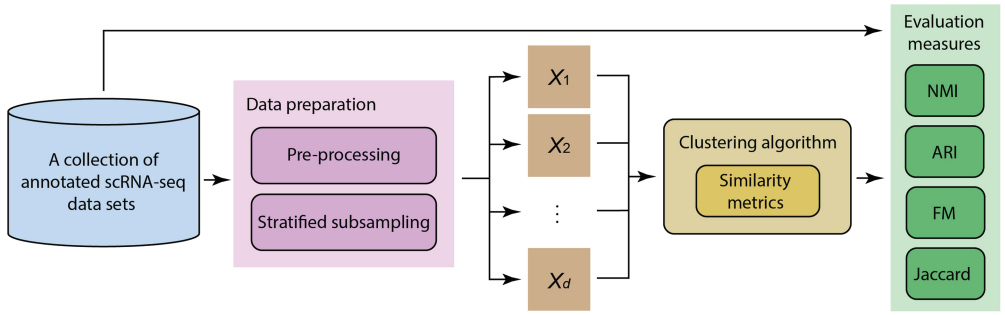

Figure 1. A schematic workflow of the clustering evaluation framework for scRNA-seq dataset.

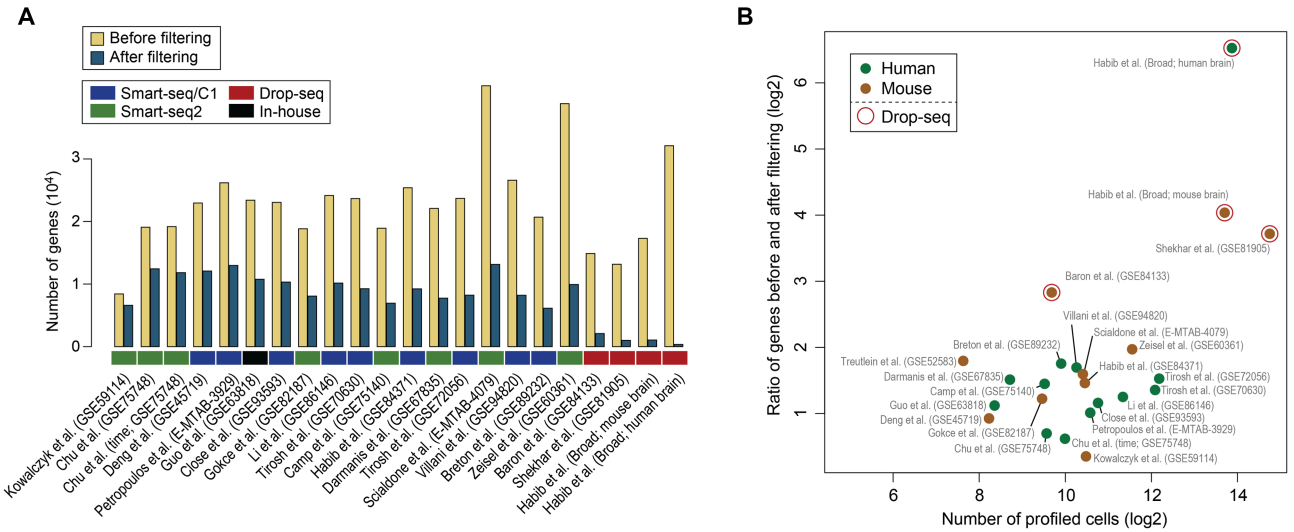

Figure 2. Preprocessing of a collection of annotated scRNA-seq datasets. (A) Datasets ordered decreasingly by the ratio of genes before and after gene filtering. Colour boxes under each bar denote the scRNA-seq library preparation protocol for each dataset. (B) The number of profiled cells in each dataset versus the ratio of genes before and after filtering. Open red circles denote datasets generated by using the Drop-seq-based protocol.

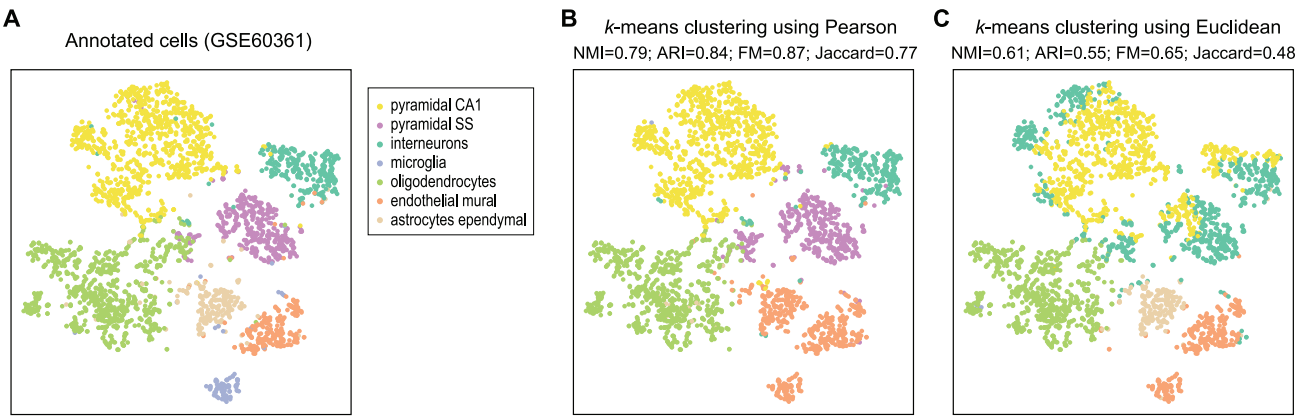

Figure 3. tSNE visualisation of cells from a sample scRNA-seq dataset (GSE60361). (A) Cells coloured by cell-type annotation from the original study [7]. (B) Cells coloured by k-means clustering using Pearson's correlation and concordance quantified by NMI, ARI, FM and Jaccard index with respect to the original cell-type annotation. Panel (C) is the same as panel (B) except for the use of Euclidean distance for k-means clustering.

information (NMI), adjusted Rand index (ARI), Fowlkes–Mallows index (FM) and Jaccard index [51]. Let $U = \{u_1, u_2, \ldots, u_k\}$ and $V = \{v_1, v_2, \ldots, v_k\}$ denote the gold standard and the clustering partition across K classes, respectively.

NMI is defined as follows:

$$\mathrm{NMI} = \frac{2I(U,V)}{H(U)+H(V)},$$

where $I(U,V)$ is the mutual information of U and V, defined as

$$I(U,V) = \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{|u_i \cap v_j|}{N} \log \frac{N|u_i \cap v_j|}{|u_i| \times |v_j|},$$

and $H(U)$ and $H(V)$ are the entropy of partitions U and V calculated as

$$H(U) = -\sum_{i=1}^{K} \frac{u_i}{N} \log \frac{u_i}{N}, \quad H(V) = -\sum_{j=1}^{K} \frac{v_j}{N} \log \frac{v_j}{N},$$

where N is the total number of cells.

Let us define a confusion matrix where, given a gold standard, a is the number of pairs of cells correctly partitioned into the same class by a clustering method, b is the number of pairs of cells partitioned into the same cluster but in fact belongs to different classes, c is the number of pairs of cells partitioned into different clusters but belongs to the same class and d is the number of pairs of cells correctly partitioned into different clusters [52] (Table 2).

ARI, FM and Jaccard index can be calculated as follows:

$$\mathrm{ARI} = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)};$$

$$\mathrm{FM} = \sqrt{\left(\frac{a}{a+b}\right)\left(\frac{a}{a+c}\right)};$$

$$\mathrm{Jaccard} = \frac{a}{a+b+c}.$$

Results

scRNA-seq clustering evaluation framework

An illustration of the clustering evaluation framework used in this study is shown in Figure 1. The input to this framework is the compendium of scRNA-seq datasets we collected for methods

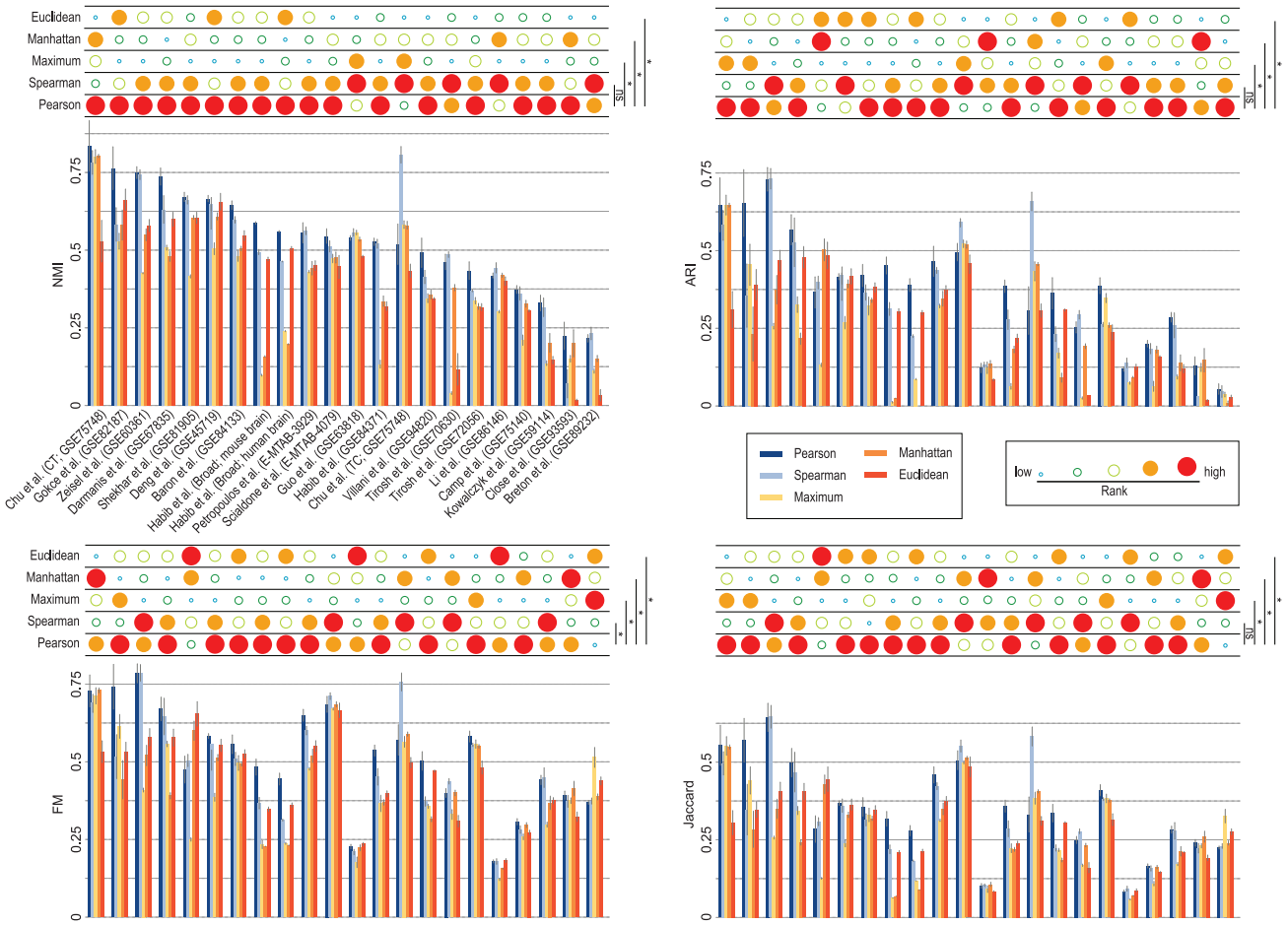

Figure 4. Benchmarking of similarity metrics on k-means clustering of scRNA-seq datasets. Results were ranked in descending order by NMI of Pearson's correlation. Performance of each similarity metric was further ranked for each dataset using coloured solid and open circles. A one-sided binomial test was performed to compare results from using Pearson's correlation metric with other alternative metrics across scRNA-seq datasets. Error bars, S.E.M.; statistical significance of a one-sided binomial test marked by * or NS, not significant.

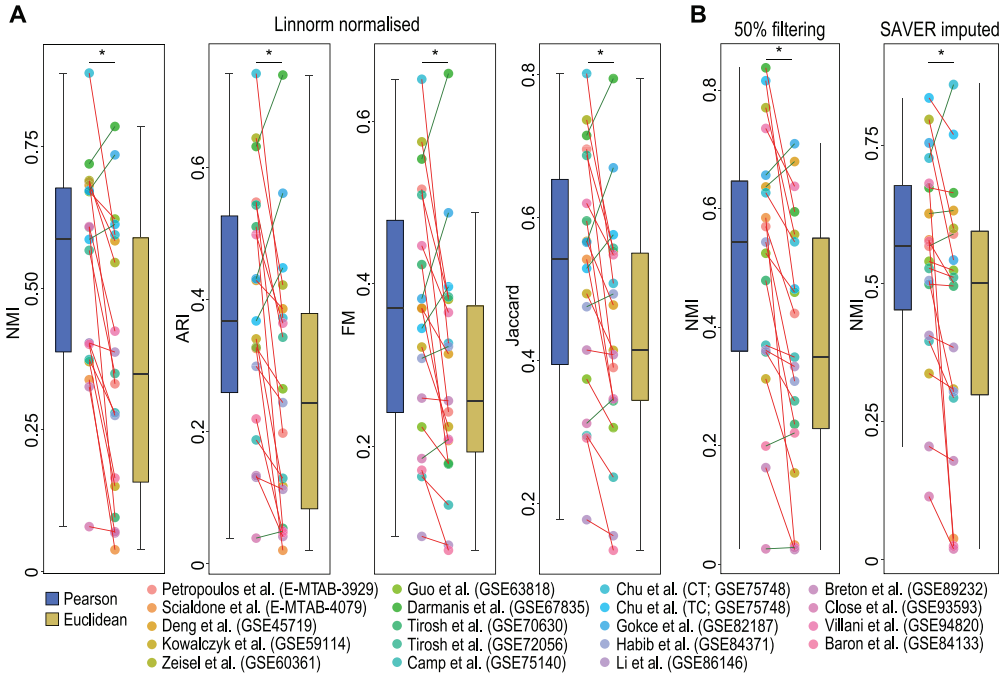

Figure 5. Comparison of *k*-means clustering on scRNA-seq datasets with (A) further Linnorm normalisation, (B) filtering of genes expressed in less than 50% of cells or further SAVER imputation. Clustering results from each dataset using Pearson's correlation or Euclidean distance as the similarity metric are matched by coloured points for each evaluation measure. A red line is drawn between a pair of points if clustering using Pearson's correlation performed better than Euclidean distance, and a green line is drawn vice versa. Statistical significance of a one-sided binomial test computed from the pairs is marked by *.

evaluation (Table 1). Each scRNA-seq dataset is firstly preprocessed (see Methods), and subsequently a repeated stratified subsampling approach was applied to account for the variability in the clustering result. In each stratified subsampling, we randomly selected 80% of the cells from each annotated class and performed clustering on this subsample. This was repeated five times for each similarity metric, for each clustering algorithm and on each scRNA-seq dataset. Finally, the clustering output from a given trial of a clustering algorithm was compared to the predefined annotation of cell types/classes and quantified using NMI, ARI, FM and Jaccard index.

Figure 2A summarises the number of genes in each original dataset we obtained from the three repositories (i.e. NCBI-GEO, EMBL-EBI and Broad Institute) and those after applying data preprocessing steps. Interestingly, we observed an association between the number of genes filtered from a dataset and the type of scRNA-seq protocol utilised (Figure 2A). On average, about half of genes passed filtering for datasets generated by Smart-seq/C1 or Smart-seq2 [53, 54]. In contrast, only 5% of genes on average passed the filtering steps for datasets generated from Drop-seq-based protocols [18, 55]. Consistent with this, it appears that, in general, the larger the number of profiled cells in scRNA-seq data such as with Drop-seq [46] and inDrop [34], the higher the ratio of genes removed by gene filtering (Figure 2B, open red circle). This is in agreement with the current understanding that Drop-seq-based approaches typically generate shallower transcriptome profiles compared to Smart-seq/C1- and Smart-seq2-based approaches [56, 57].

Benchmarking results on *k*-means clustering

We performed *k*-means clustering on a scRNA-seq dataset generated by Zeisel *et al.* [7] (GSE60361) using either Pearson's correlation or Euclidean distance as the similarity metric.

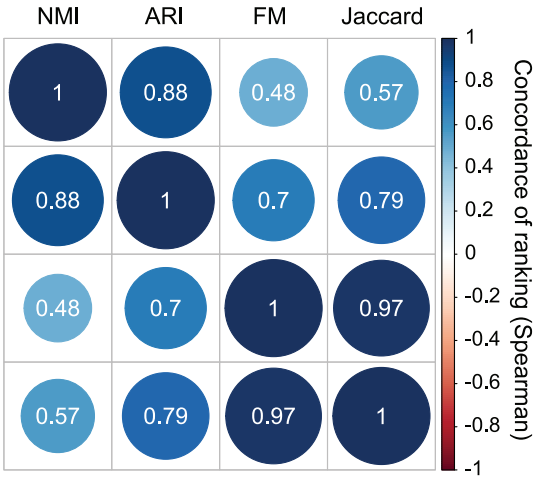

Figure 6. Concordance among the four evaluation measures across all tested scRNA-seq datasets and similarity metrics for *k*-means clustering.

Figure 3 shows the *t*-distributed stochastic neighbour embedding (tSNE) visualisation of the dataset, and cells were colour-coded according to their cell-type annotation (Figure 3A). We found that the classification output from using Pearson's correlation (Figure 3B) shows higher concordance to the predefined cell-type annotation (NMI = 0.79, ARI = 0.84, FM = 0.87 and Jaccard index = 0.77) compared to that from Euclidean distance (NMI = 0.61, ARI = 0.55, FM = 0.65 and Jaccard index = 0.48; Figure 3C).

To systematically assess the impact of each similarity metric on the performance of the *k*-means clustering algorithm, we employed the proposed evaluation framework and summarised the evaluation results in Figure 4. In most cases, Pearson and Spearman's correlation metrics performed favourably in com-

parison to distance-based metrics across all four evaluation measures. On average, our result shows an improvement of correlation-based metrics over distance-based metrics by 31.5% (0.12) for NMI, 39.6% (0.1) for ARI, 16% (0.07) for FM and 23% (0.06) for Jaccard index. To test if such observations are statistically significant, we performed a nonparametric binomial test in which we counted the number of times that *k*-means clustering results from using the Pearson's correlation metric outperformed other alternative metrics across the 22 tested scRNA-seq datasets. We found that regardless of the evaluation measures, the performance of the *k*-means clustering coupled with the Pearson's correlation metric was statistically better compared to those from maximum, Manhattan and Euclidean metrics. Except in the case of FM where *k*-means clustering using Pearson's correlation performed marginally better than that using Spearman's correlation, the two correlation-based metrics otherwise performed similarly according to NMI, ARI and Jaccard index.

To further assess if the improvement from using Pearson's correlation as a similarity metric is caused by data scaling and/or zero counts in the scRNA-seq datasets, we applied Linnorm, a zero-count-aware scRNA-seq normalisation method [48], to perform additional normalisation on all preprocessed datasets prior to clustering. Comparison of *k*-means clustering on Linnorm normalised scRNA-seq datasets showed again that Pearson's correlation performed significantly better than Euclidean distance irrespective of evaluation measures (Figure 5A). We also compared *k*-means clustering performance on (i) datasets after filtering genes that are expressed in less than 50% of the cells and (ii) those with SAVER [49] imputation after preprocessing. We found in both cases that clustering using Pearson's correlation remains better than Euclidean distance (Figure 5B). These results suggest that the performance improvement from using Pearson's correlation as a similarity metric is unlikely simply due to data scaling and/or zero counts in the scRNA-seq datasets.

While all evaluation measures suggested that *k*-means clustering using correlation-based metrics showed better concordance with predefined cell types (gold standard), they did give different rankings in many cases for individual datasets (Figure 4). We therefore tested how consistent the clustering evaluation results are across each of the four evaluation measures. Figure 6 shows the pairwise comparison on the consistency of each evaluation measure in ranking the clustering results from different similarity metrics across all tested scRNA-seq datasets. It appears that the rankings from NMI compared to ARI and FM compared to Jaccard index show high consistency. Overall, the four evaluation metrics largely agreed with each other (Spearman *r* of 0.48 to 0.97).

Clustering of scRNA-seq data using modified SIMLR

To test if the use of a correlation-based metric can benefit the latest clustering algorithm designed for scRNA-seq data, we modified SIMLR (see Methods), a kernel-based clustering algorithm that relies on Euclidean distance for Gaussian kernel construction, to use the Pearson's correlation that appeared to have the most favourable results in *k*-means clustering (Figure 7). We subsequently applied these two versions of implementation to our collection of scRNA-seq datasets and compared their performance using the panel of four evaluation measures. Figure 6 summarises the evaluation results. Similar to our findings from

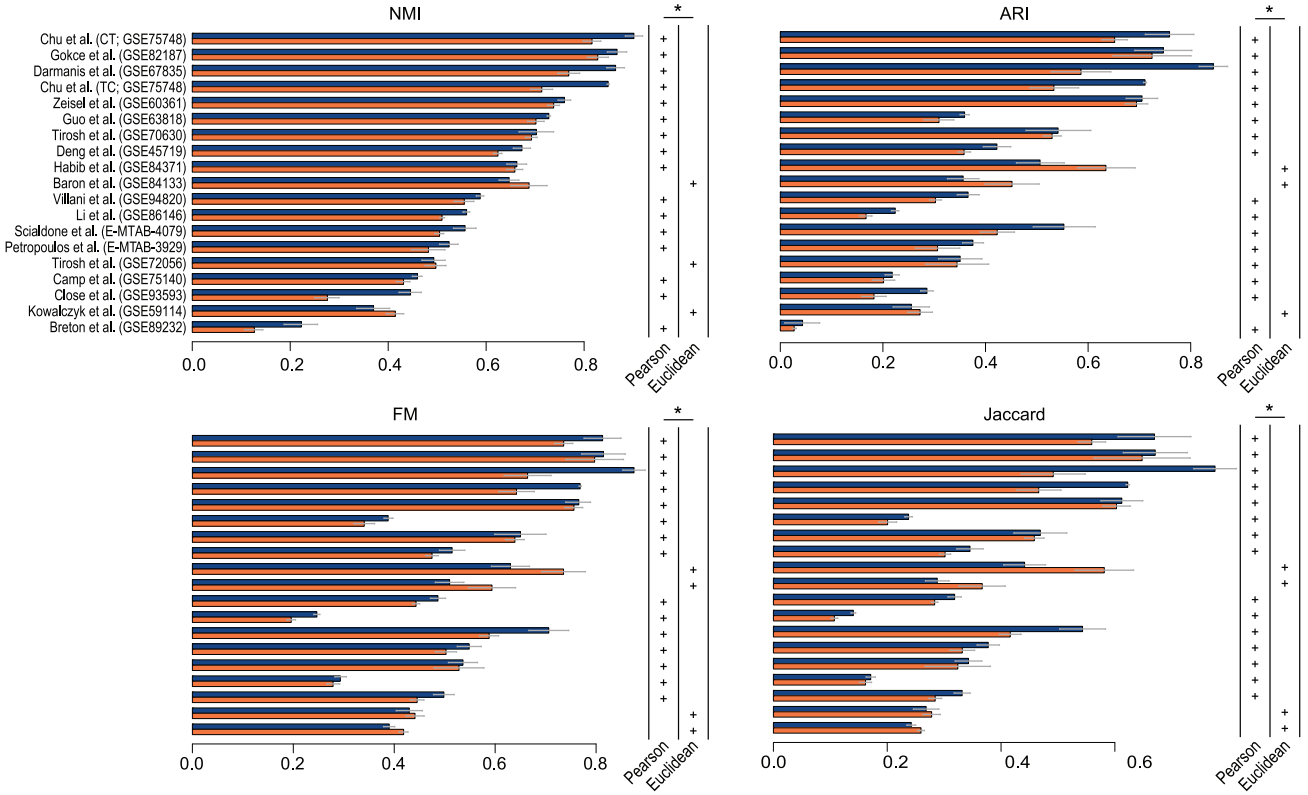

Figure 7. Benchmarking of similarity metrics on kernel-based clustering of scRNA-seq datasets using the SIMLR algorithm. Results were ranked in descending order by NMI of Pearson's correlation. The + sign denotes the outperforming similarity metric in each dataset. Error bars, S.E.M.; statistical significance of a one-sided binomial test is marked by *.

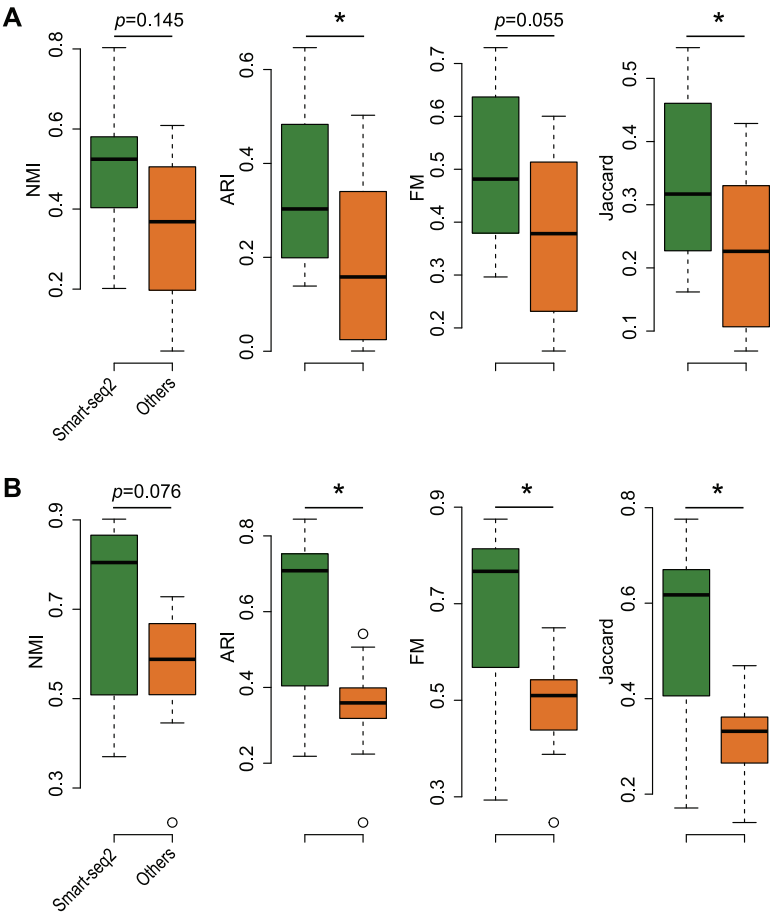

Figure 8. Clustering performance comparison on scRNA-seq library preparation protocols using (A) *k*-means clustering and (B) modified SIMLR both with Pearson's correlation. A one-sided Wilcoxon rank sum test was performed for each comparison with respect to NMI, ARI, FM or Jaccard index evaluation metrics. Statistical significance ($P < 0.05$) is marked by *.

using *k*-means clustering, SIMLR coupled with Pearson's correlation performed significantly better than when it coupled with Euclidean distance according to all four evaluation metrics ($P \leq 0.0096$).

Specifically, the average improvements of the modified SIMLR with a Pearson's correlation metric compared to the original version with a Euclidean distance metric are 7.7% (0.04) for NMI, 12.7% (0.05) for ARI, 6.3% (0.03) for FM and 10.4% (0.04) for Jaccard index. Together, these results demonstrate that the choice of a correlation-based metric such as Pearson's correlation improves not only the standard *k*-means clustering but also more the advanced approach such as SIMLR, suggesting that it may be more suitable for clustering cells by their expression profiles captured in scRNA-seq experiments independent of a clustering algorithm. Note that SIMLR clustering on GSE81905 [46] and two Broad Institute datasets [45] failed to converge within a week presumably due to the large size of cells profiled.

Effect of scRNA-seq library preparation protocols on clustering

Multiple scRNA-seq library preparation protocols are currently in use [57]. To test if the library preparation protocols have any effect on scRNA-seq data clustering, we compared the clustering performance with respect to NMI, ARI, FM or Jaccard index evaluation metrics for scRNA-seq datasets generated from the Smart-seq2 protocol against other alternative protocols (i.e. Smart-seq/C1, Drop-seq, and in-house protocols) (Figure 8). We found that in most cases scRNA-seq data generated from using Smart-seq2 library preparation protocols show better clustering results when clustered using both *k*-means clustering algorithm with Pearson's correlation (Figure 8A) and modified SIMLR with Pearson's correlation (Figure 8B). These data suggest that scRNA-seq datasets generated from using full-length Smart-seq2 protocols may deliver higher sensitivity, which leads to better clustering results.

Discussion

The ultra-high throughput of scRNA-seq techniques has brought about various new computational challenges, including normalisation [58], differential expression analysis [59], cell cycle identification [60] and clustering, a crucial step for cell-type identification. Given the large impact of clustering on downstream analyses, it is important to apply a clustering procedure that is robust to the influence from sources such as data noise and normalisation procedures. In this work, we presented a framework for benchmarking clustering approaches in scRNA-seq data analysis. We demonstrated that the choice of similarity metric could affect the performance of the clustering of scRNA-seq data.

While in some cases differentially expressed genes were selected and/or dimensionality reduction steps (e.g. tSNE or PCA) were applied prior to clustering, the application of clustering to the expression profiles of the transcriptome of cells is less prone to bias (e.g. the selection of gene subsets) and widely performed. Clustering on the transcriptome profile, however, do bear important difference when different similarity metrics are used. For example, distance-based metrics such as Euclidean distance take the gene expression levels into account and therefore may be susceptible to data scaling and normalisation procedures. In contrast, correlation-based metrics such as Pearson's correlation are invariant to scaling and hence are inherently robust.

Taken together, our systematic evaluation on the impact of different similarity metrics suggests that improved cell-type clustering results (according to the predefined gold standard) can be achieved by using correlation-based metrics. Our modified implementation of a state-of-the-art clustering algorithm (SIMLR) using Pearson's correlation highlights its usefulness as a similarity metric for scRNA-seq data clustering.

Key Points

- Similarity metrics are key to clustering algorithms. The comparison of similarity metrics is crucial and will guide the development of scRNA-seq clustering algorithms.
- Using a comprehensive benchmark framework, we demonstrated that the choice of a similarity metric has a significant impact on scRNA-seq data clustering results. In general, correlation-based metrics performed favourably in comparison to distance-based metrics in scRNA-seq data clustering.
- Distance-based metrics such as Euclidean distance are sensitive to data scaling, whereas correlation-based metrics such as Pearson's correlation are invariant to scaling. Such property makes correlation-based metrics robust to data noise and normalisation procedure.
- The modification of a kernel-based clustering algorithm, SIMLR, from using Euclidean distance to Pearson's correlation yields a significant improvement on its performance on scRNA-seq data clustering.
- The scRNA-seq library preparation protocols can also affect the clustering results. Further comparison suggests that, in general, data generated from full-length Smart-seq2 protocols give favourable clustering results irrespective of the clustering algorithms or evaluation metrics.

Acknowledgements

We thank members from the School of Mathematics and Statistics, and School of Life and Environmental Science for their feedback to this work.

Funding

This work has been supported by the Australian Research Council Discovery Early Career Researcher Award (DE170100759 to P.Y.), Australian Research Council Discovery Projects (DP170100654 to J.Y.H.Y. and P.Y.) and National ealth and Medical Research Council Career Development Fellowships (1111338 to J.Y.H.Y). Financial support from Judith and David Coffey Life Lab was given to T.K. and R.C.

References

1. Jaitin DA, Kenigsberg E, Keren-Shaul H, *et al*. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 2014;**343**:776–9.
2. Kolodziejczyk AA, Kim JK, Svensson V, *et al*. The technology and biology of single-cell RNA sequencing. *Mol Cell* 2015; **58**:610–20.
3. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 2015;**16**:133–45.
4. McCarthy DJ, Campbell KR, Lun ATL, *et al*. Scater: preprocessing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 2017;**33**: 1179–86.
5. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol* 2016;**17**:63.
6. Grün D, Lyubimova A, Kester L, *et al*. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 2015;**525**:251–5.
7. Zeisel A, Munoz-Manchado AB, Codeluppi S, *et al*. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 2015;**347**:1138–42.
8. Poulin JF, Tasic B, Hjerling-Leffler J, *et al*. Disentangling neural cell diversity using single-cell transcriptomics. *Nat Neurosci* 2016;**19**:1131–41.
9. Samusik N, Good Z, Spitzer MH, *et al*. Automated mapping of phenotype space with single-cell data. *Nat Methods* 2016;**13**:493–6.
10. Tasic B, Menon V, Nguyen TN, *et al*. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci* 2016;**19**:335–46.
11. Li H, Courtois ET, Sengupta D, *et al*. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* 2017;**49**:708–18.
12. Zappia L, Phipson B, Oshlack A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput Biol* 2018;**14**:e1006245.
13. Buettner F, Natarajan KN, Casale FP, *et al*. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* 2015;**33**:155–60.
14. Grün D, Muraro MJ, Boisset JC, *et al*. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* 2016;**19**:266–77.
15. Kiselev VY, Kirschner K, Schaub MT, *et al*. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 2017;**14**:483–6.
16. Lin P, Troup M, Ho JWK. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol* 2017;**18**:59.
17. Dey KK, Hsiao CJ, Stephens M. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet* 2017;**13**(5):e1006759.
18. Macosko EZ, Basu A, Satija R, *et al*. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;**161**:1202–14.

19. Wang B, Zhu J, Pierson E, *et al*. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* 2017;**14**:414–6
20. Shirkhorshidi AS, Aghabozorgi S, Ying WT. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLoS One* 2015;**10**(12): e0144059.
21. Boriah S, Chandola V, Kumar V. Similarity measures for categorical data: a comparative evaluation. In: *Proceedings of the 2008 SIAM International Conference on Data Mining. SIAM proceedings, Atlanta, Georgia*, 2008.
22. Zhang Z, Huang K, Tan T. Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. In: *Proceedings—International Conference on Pattern Recognition*. IEEE, Hong Kong, China, 2006.
23. Weller-Fahy DJ, Borghetti BJ, Sodemann AA. A survey of distance and similarity measures used within network intrusion anomaly detection. *IEEE Commun Surv Tutor* 2015;**17**:70–91.
24. Irani J, Pise N, Phatak M. Clustering techniques and the similarity measures used in clustering: a survey. *Int J Comput Appl* 2016;**134**(7):9–14.
25. Menon V. Clustering single cells: a review of approaches on high- and low-depth single-cell RNA-seq data. *Brief Funct Genomics* 2018;**17**(4):240–245.
26. Barrett T, Wilhite SE, Ledoux P, *et al*. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013;**41**(Database issue):D991–5.
27. Petryszak R, Burdett T, Fiorelli B, *et al*. Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res* 2014;**42**(Database issue): D926–32.
28. Deng Q, Ramskold D, Reinius B, *et al*. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 2014;**343**:193–6.
29. Guo F, Yan L, Guo H, *et al*. The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell* 2015;**161**:1437–52.
30. Darmanis S, Sloan SA, Zhang Y, *et al*. A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A* 2015;**112**(23):7285-7290.
31. Gokce O, Stanley GM, Treutlein B, *et al*. Cellular taxonomy of the mouse striatum as revealed by single-cell RNA-seq. *Cell Rep* 2016;**16**:1126–37.
32. Camp JG, Badsha F, Florio M, *et al*. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc Natl Acad Sci U S A* 2015;**112**(51):15672–15677.
33. Chu L-F, Leng N, Zhang J, *et al*. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol* 2016;**17**:173.
34. Baron M, Veres A, Wolock SL, *et al*. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst* 2016;**3**:346–60.
35. Breton G, Zheng S, Valieris R, *et al*. Human dendritic cells (DCs) are derived from distinct circulating precursors that are precommitted to become $CD1c^+$ or $CD141^+$ DCs. *J Exp Med* 2016;**213**:2861–70.
36. Villani A-C, Satija R, Reynolds G, *et al*. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*80 2017;**356**:eaah4573.
37. Scialdone A, Tanaka Y, Jawaid W, *et al*. Resolving early mesoderm diversification through single-cell expression profiling. *Nature* 2016;**535**:4–6.
38. Habib N, Li Y, Heidenreich M, *et al*. Div-seq: single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. *Science* 2016;**353**:925–8.
39. Kowalczyk MS, Tirosh I, Heckl D, *et al*. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res* 2015;**25**:1860–72.
40. Petropoulos S, Edsgärd D, Reinius B, *et al*. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* 2016;**165**:1012–26.
41. Close JL, Yao Z, Levi BP, *et al*. Single-cell profiling of an in vitro model of human interneuron development reveals temporal dynamics of cell type production and maturation. *Neuron* 2017;**93**:1035–48 e5.
42. Li L, Dong J, Yan L, *et al*. Single-cell RNA-seq analysis maps development of human germline cells and gonadal niche interactions. *Cell Stem Cell* 2017;**20**:858–73 e4.
43. Tirosh I, Venteicher AS, Hebert C, *et al*. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* 2016;**539**:309–13.
44. Tirosh I, Izar B, Prakadan SM, *et al*. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 2016;**352**:189–96.
45. Habib N, Avraham-Davidi I, Basu A, *et al*. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* 2017;**14**:955–8.
46. Shekhar K, Lapan SW, Whitney IE, *et al*. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* 2016;**166**:1308–23 e30.
47. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40.
48. Yip SH, Wang P, Kocher J-PA, *et al*. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res* 2017;**45**(22):e179.
49. Huang M, Wang J, Torre E, *et al*. SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* 2018;**15**:539–42.
50. Lloyd SP. Least squares quantization in PCM. *IEEE Trans Inf Theory* 1982;**28**:129–37.
51. Wagner S, Wagner D. Comparing clusterings—an overview. *Analysis* 2007;**4769**:1–19.
52. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun* 2018;**9**(1):997.
53. Picelli S, Björklund ÅK, Faridani OR, *et al*. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* 2013;**10**:1096–100.
54. Wu AR, Neff NF, Kalisky T, *et al*. Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* 2014;**11**:41–6.
55. Klein AM, Mazutis L, Akartuna I, *et al*. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;**161**:1187–201.
56. Svensson V, Natarajan KN, Ly LH, *et al*. Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* 2017;**14**(4):381–387.
57. Ziegenhain C, Vieth B, Parekh S, *et al*. Comparative analysis of single-cell RNA sequencing methods. *Mol Cell* 2017;**65**(4):631–643.

58. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 2016;**17**:75.
59. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. *Nat Methods* 2014;**11**:740–2.
60. Scialdone A, Natarajan KN, Saraiva LR, *et al*. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* 2015;**85**:54–61.