Stat Biosci
DOI 10.1007/s12561-017-9194-z

SIDEseq: A Cell Similarity Measure Defined by Shared Identified Differentially Expressed Genes for Single-Cell RNA sequencing Data

Courtney Schiffman[1] · Christina Lin[2] · Funan Shi[3] · Luonan Chen[4] · Lydia Sohn[5] · Haiyan Huang[3]

Received: 10 November 2016 / Accepted: 27 April 2017

Abstract One goal of single-cell RNA sequencing (scRNA seq) is to expose possible heterogeneity within cell populations due to meaningful, biological variation. Examining cell-to-cell heterogeneity, and further, identifying subpopulations of cells based on scRNA seq data has been of common interest in life science research. A key component to successfully identifying cell subpopulations (or clustering cells) is the (dis)similarity measure used to group the cells. In this paper, we introduce a novel measure, named SIDEseq, to assess cell-to-cell similarity using scRNA seq data. SIDEseq first identifies a list of putative differentially expressed (DE) genes for each pair of cells. SIDEseq then integrates the information from all the DE gene lists (corresponding to all pairs of cells) to build a similarity measure between two cells. SIDEseq can be implemented in any clustering algorithm that requires a (dis)similarity matrix. This new measure incorporates information from all cells when evaluating the similarity between any two cells, a characteristic not commonly found in existing (dis)similarity measures. This property is advantageous for two reasons: (a) borrow-

Electronic supplementary material The online version of this article (doi:10.1007/s12561-017-9194-z) contains supplementary material, which is available to authorized users.

✉ Haiyan Huang
hhuang@stat.berkeley.edu

1 Department of Biostatistics, UC Berkeley, Berkeley, USA

2 Department of Chemical Biology and Department of Molecular and Cellular Biology, UC Berkeley, Berkeley, USA

3 Department of Statistics, UC Berkeley, Berkeley, USA

4 Key Laboratory of Systems Biology, Innovation Center for Cell Signaling Network, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

5 Department of Mechanical Engineering, UC Berkeley, Berkeley, USA

Published online: 17 May 2017

ing information from cells of different subpopulations allows for the investigation of pairwise cell relationships from a global perspective and (b) information from other cells of the same subpopulation could help to ensure a robust relationship assessment. We applied SIDEseq to a newly generated human ovarian cancer scRNA seq dataset, a public human embryo scRNA seq dataset, and several simulated datasets. The clustering results suggest that the SIDEseq measure is capable of uncovering important relationships between cells, and outperforms or at least does as well as several popular (dis)similarity measures when used on these datasets.

Keywords single-cell RNA sequencing (scRNA seq) · subpopulation identification · single-cell clustering · similarity measure · ovarian cancer · EMT inducers (Thrombin, TGFB-1)

1 Introduction

RNA sequencing technologies have allowed researchers to explore the genetic processes underlying many biological phenomena. Typical bulk RNA seq data, by pooling cells together, measure average gene expression. Since many studies require a deeper examination of genetic activities due to the complex and mysterious nature of certain diseases, single-cell experiments quickly became a technological standard in life science research [5,25]. Focusing in on the single-cell level allows researchers to investigate the meaningful and illuminating heterogeneity among cells of interest and to discover cell-based biologics, e.g., to identify cellular subpopulations and rare cell types, to identify genes differentially expressed within subpopulations, and to examine genetic regulation networks (e.g., [30]).

Numerous clustering methods with varying degrees of complexity have been introduced to study scRNA seq data. For example, a new algorithm, named GiniClust, was developed to cluster cells using genes with the top normalized Gini indices [14]. SNN-Cliq is a method to identify cell subpopulations by first building a list of the k-nearest-neighbors (KNN) for each cell using Euclidean distance, and then assessing the similarity between any two cells by examining their KNN lists [27]. The RaceID (Rare Cell Type Identification) algorithm was designed to differentiate rare, tissue or disease-specific cells among complex populations of cells through two iterations of k-means clustering [10]. The first k-means clustering is performed with various specified similarity measures (the default measure is Pearson correlation) in order to identify the measure which results in the most robust clusters. Outlier cells are then identified and clustered separately. In the second k-means clustering, the centers of the original clusters are redefined, and cells are reassigned to the nearest cluster center. The PhenoGraph algorithm focuses on identifying the phenotypes of cells based on signaling proteins, whose expressions are used to construct a k nearest-neighbor graph (as defined by Euclidean distance) [17]. The Louvain community detection method is then used to partition the graph in order to find communities of phenotypically similar cells. BackSPIN, a biclustering method which seeks to identify subpopulations of cells while simultaneously finding genetic markers of the clusters, has a correlation matrix at the foundation of its complex sorting and splitting algorithm [30]. There are

many clustering methods to add to this list, and there are surely more to come. We see that most clustering algorithms rely on some (dis)similarity measure as a basis for clustering regardless of subsequent computational or mathematical complexity. For instance, a key component in the SNN-Cliq or PhenoGraph algorithms is the use of KNN, derived from Euclidian distances between cells. However, if Euclidian distance was not an appropriate measure to use due to the nature of the data or the study goal, then the KNN lists as well as the final clustering results would be misleading. Similarly, in other methods, if the employed (dis)similarity measures are not appropriate measures of cell similarity, clustering results from the algorithms may be unreliable. Therefore, the performance and accuracy of many clustering algorithms in the scRNA seq setting depend on the ability of the used (dis)similarity measures to summarize true, subtle relationships between cells.

In this paper, we focus on introducing a novel measure, named SIDEseq (defined by shared identified differentially expressed genes), to evaluate pairwise similarities between cells using scRNA seq data. There are several intriguing and unique ideas behind SIDEseq. Most importantly, the SIDEseq measure incorporates information from all cells in the dataset when defining the similarity between just two cells. What kind of information is important to incorporate from all cells when defining cellular relationships? In scRNA seq datasets, differentially expressed (DE) genes between cells/subpopulations often represent the kinds of relationships and information researchers care about. The SIDEseq measure first identifies the lists of putative DE genes for all pairs of cells and then quantifies the similarity between two cells by examining how much the two cells share in common among their resulting lists of DE genes when they are compared against every other individual cell in the dataset. Note that we attempt to evaluate differential expression for a gene based on only two expression values (or between just two cells). This may seem unreasonable at first glance. However, we consider that the DE genes would likely have vague subpopulation-specific information if they were identified across all cells from multiple subpopulations. It is likely that these DE genes would not be as effective at distinguishing between subpopulations as the genes that carry more explicit subpopulation information. SIDEseq attempts to extract and integrate subpopulation-specific information from all cells. Furthermore, since it considers all possible pairwise comparisons of cells, SIDEseq is expected to be robust against noise in any individual list of identified DE genes. The calculation of the SIDEseq measure involves two key quantifications: how to quantify differential expression for a gene between just two cells and how to evaluate consistency among multiple lists of DE genes. To make SIDEseq computationally feasible, we have introduced two simple yet effective statistics to achieve these quantifications (see “Methods and Materials” for more details).

The development of SIDEseq was motivated in part by our investigation of a scRNA seq dataset consisting of 96 cells from the human epithelial ovarian cancer cell line, CAOV-3. Half of the cells were treated with two factors that are hypothesized to be epithelial-to-mesenchymal (EMT) inducers. There were several motivations behind studying the subpopulations of these cells using their expression profiles. First, such a study could reveal the genetic markers of any subpopulations within the untreated (also referred to as control) or treated cells which could then offer an improved understanding as to why and how they transition to a mesenchymal phenotype. Second, by

attempting to cluster cells by treatment status, we could verify whether such treatments could actually induce the cells to transition from epithelial to mesenchymal [9,31]. Furthermore, the heterogeneous nature of human ovarian cancer cells presented a challenging clustering task which is not only statistically interesting but also biologically interesting in its own right. The cells do not differ by tissue type, cancer type, or other forms of strong biological variation, but they are, by nature, quite heterogeneous. The source of biological variation, which may be the most prominent and noticeable among the cells overall, is their treatment with the two factors. However, when the differences due to treatment are subtle (e.g., when a treatment has only a marginal effect on cells), they could be easily overwhelmed by the cell heterogeneity. This would bring challenges to clustering treated (by different factors) and untreated cells. We explored the human ovarian cancer cell dataset using hierarchical clustering paired with Euclidean distance, Pearson and Spearman correlation, and the SIDEseq similarity measure, for comparison. The traditional similarity measures were unable to clearly cluster the treated and untreated cells. Hierarchical clustering with the SIDEseq measure was able to cluster the cells by treatment status to a greater extent. Clustering of cells by treatment status was especially challenging for one of the two batches/treatment factors. Therefore, our clustering analysis of the human ovarian cancer cells not only allowed for a useful comparison of measures within a challenging clustering context, but also helped to shed light on the effectiveness (or ineffectiveness) of the two treatment factors in inducing EMT.

For further evaluation of the SIDEseq measure, we studied a public scRNA seq dataset involving human embryo cells from Xu et al. We focused on this public dataset because, unlike our human ovarian cancer cell dataset, this dataset consists of cells from different developmental stages. Therefore, we believed that the cells from this dataset could be clustered more successfully and would provide a good comparison of the performance of our proposed measure with current, popular measures. We use both hierarchical clustering and spectral clustering of the dataset to compare the similarity measures. The public dataset also allows for a comparison of clustering with SIDEseq to a more recent clustering algorithm, called SNN-Cliq, which was originally used to cluster the embryo cells [27].

To further explore the benefits of the SIDEseq similarity measure and its ability to exploit the information found in DE genes to define cell similarity, we also simulated several scRNA seq datasets. The subpopulations of cells in each dataset varied in size, proportion of DE genes, mean expression of differentially and non-DE genes, etc. For each simulation, we used the SIDEseq similarity measure with hierarchical clustering to study the measure's ability to react to the various simulation parameters which make clustering of cells into true subpopulations more challenging. We also used the simulation studies to compare the SIDEseq similarity measure with the methods used by Jiang *et al.* in their GiniClust algorithm. In the GiniClust algorithm, genes with the top normalized Gini indices are used for clustering [15]. This is similar to how the SIDEseq measure uses largely the top DE genes to define the similarity between two cells, but different in a significant way in that the GiniClust method does not do pairwise comparisons of all cells in the data when defining the similarity between just two cells. To test the importance of this difference, we used the top Gini index genes as identified by the GiniClust algorithm to perform hierarchical clustering with

Pearson and Spearman correlation and Euclidean distance and compared the clustering results with those resulting from SIDEseq. In all simulations, SIDEseq outperformed the other measures. For a final comparison, we used the full GiniClust algorithm on the simulated datasets, but found that this algorithm was significantly outperformed by the hierarchical clustering methods described above.

This paper is organized as follows: First, we give a more detailed description of our human ovarian cancer cell dataset, and the pre-processing steps we took prior to analysis. We then define our proposed similarity measure, SIDEseq. We use various simulations to compare the methods found in the GiniClust algorithm with the SIDEseq similarity measure and their ability to accurately identify subpopulations. Next, we compare the performance of the SIDEseq measure with common (dis)similarities when used for hierarchical clustering of the human ovarian cancer cell dataset. Finally, we compare the performance of the SIDEseq measure with other common measures when used for the clustering of two public, scRNA seq datasets.

2 Methods and Materials

2.1 The Single-Cell RNA seq Data

The novel dataset of interest in this study consists of 96 cells from the human epithelial ovarian cancer cells line, CAOV-3 (ATCC, Manassas, VA, USA). CAOV-3 cells were plated on 100-mm tissue culture dishes at a subcultivation ratio of 1:5, incubated overnight in supplemented DMEM medium, and then incubated with either thrombin (2.0 U/mL) or TGFβ-1 (5ng/mL) (both from R&D Systems, Minneapolis, MN, USA) for 48 hours. The samples were then prepared per the established protocol for the C1 Single-Cell Auto Prep System (Fluidigm, San Francisco, CA, USA). See Supplementary Materials for more details on cell culture and preparation.

The ovarian cancer cells were sequenced in two batches of 48 cells each. Twenty-four of the cells in one batch were treated with TGFβ-1, and 24 of the cells in the second batch were treated with thrombin. The remaining cells in both batches were untreated, control cells. Throughout the paper, the batch containing the 24 cells treated with thrombin and their corresponding control cells will be referred to as the TGFβ-1 group, and the batch containing the 24 cells treated with thrombin along with their control cells will be referred to as the thrombin group. While TGFβ-1 is a well-established inducer of EMT, there is less evidence to support thrombin's role in EMT [9,31]. Within the context of cancer, EMT is a process in which cell–cell adhesion and basoapical polarity are lost, EpCAM is down-regulated, and the expression of mesenchymal-associated genes is induced [7,8]. There is growing evidence that EMT is activated during, and plays a critical role in, cancer invasion and metastasis formation [2,7,8,16,19,21]. The heterogeneity of the cellular phenotypes resulting from EMT in ovarian cancer cells is thought to likewise lead to an increased ability to evade early detection [31]. There are several motivations behind studying this dataset. Examining the treated ovarian cancer cells and studying whether cells cluster by treatment status can shed light on the effectiveness of the two treatments as EMT inducers, which in turn would lead to a better understanding of the EMT process in ovarian cancer.

As previously stated, the TGFβ-1 treatment is a more well-studied inducer of EMT than the thrombin treatment, which requires additional experimental validation [9,31]. Furthermore, the possibility of a variety of subtle subpopulations within the ovarian cancer cells as a result of EMT brings a statistical challenge of developing sensitive measures for assessing (dis)similarities between cells. If such subpopulations and their associated DE genes could be identified, this would aid in the research of this dangerous, often undetected, gynecologic cancer.

An important source of unwanted biological noise in scRNA seq experiments, especially pertinent to our human ovarian cancer cell dataset, is the variability introduced when cells to be sequenced have different passage numbers [3]. Passage number is defined as the number of times a cell culture was subcultured to maintain continued growth [3]. Passage number has a non-negligible effect on gene expression and regulatory pathways within cell lines [18,20]. Thus, when cells are sequenced in different batches and the passage numbers are different, cells that were supposed to be biological replicates may become biologically different. This was the case for the human ovarian cancer cell dataset, where a difference in passage number (i.e., differed by three) resulted in biologically different cells in the two batches. With a lack of "true" biological replicates between the two batches, this source of unwanted variation makes normalization across batches very challenging and suggests that within-batch normalization is more appropriate.

Since normalizing across batches would be always confounded with the passage number effects (see Supplementary Materials for details), we have focused our analysis within batches. It is interesting that the thrombin group (batch 2) shows much noisier results than what we obtained from the TGFB-1 group (batch 1), suggesting that the thrombin treatment cells did not differentiate significantly from the untreated cells (see Supplementary Materials, Normalization). Furthermore, this observation is supported by immunostaining images, which show that the thrombin-treated cells have a smaller proportion of cells that have transitioned (see Supplementary Materials, Cell Culture). This discovery about the thrombin treatment in the dataset is an important biological observation and merits further investigation. Due to the apparent ineffectiveness of thrombin as an EMT inducer, we focused on the TGFB-1-treated cells in batch 1 in the results section. We used the 'RUVs' normalization method from the 'RUVSeq' package in R to normalize the expressions of cells in batch 1 [23]. The 'RUVSeq' package provides several functions to remove unwanted factors of variation from RNA seq data by using control genes or replicate samples, which are independent of the biological variability of interest, to estimate the factors of unwanted variation using factor analysis. We used the 'RUVs' normalized expressions of the cells in batch 1 for all further analysis.

2.2 A New Similarity Measure, SIDEseq, and Clustering Analysis of Cells

We propose a novel measure, SIDEseq, which is defined by **s**hared **i**dentified **d**ifferentially**e**xpressed genes for single-cell RNA sequencing data.

2.2.1 Method Overview

SIDEseq first chooses DE genes for every cell pair by only comparing the expression levels of the two cells to produce $N(N\text{-}1)/2$ lists of DE genes (one list for one pair of cells; assuming there are N cells in total). Next, SIDEseq assesses the similarity between any two cells by comparing the level of consistency among the relevant lists of DE genes (i.e., to compare cells i and j, SIDEseq evaluates to what level the list of DE genes between cell i and cell t overlaps with the list of DE genes between cell j and cell t, and then integrates such overlapping information across all cells $t \neq i,j$ to define the similarity between cells i and j). The involved integration of multiple DE gene lists in SIDEseq makes it a quite robust measure against noise from any single list of DE genes.

The ideas behind SIDEseq stem from the belief that various subpopulations likely exist within the data and each has a unique gene activity profile. If two cells come from the same subpopulation, it may be easier to cluster them together by comparing their relationships with other cells in different subpopulations than by comparing their expression profiles in isolation. This might be the case, for example, if the noise in some expression profiles strongly affects the similarity assessment by their expression profiles alone. Another advantage of the SIDEseq measure is that, instead of using all genes, it uses mainly genes evaluated as differentially expressed to build the dissimilarities between cells. This should improve the efficiency of the measure by eliminating noise from uninformative genes.

2.2.2 Method Details

The SIDEseq measure involves two main calculations: the quantification of differential expression for a gene between two cells, and the evaluation of the consistency between multiple lists of DE genes. A flow chart demonstrating the calculation steps of the proposed measure is presented in the Supplementary Materials.

The building block for the first calculation in the SIDEseq measure is a simple statistic which is used for a rough evaluation of differential expression between two cells. Suppose one has a matrix of gene expressions of J genes by N cells. We define

$$T_{i,j}^{k} = \frac{\left| x_i^k - x_j^k \right|}{\sqrt{x_i^k + x_j^k}}, \tag{1}$$

where x_i^k is the expression of gene k in cell i and x_j^k is the expression of gene k in cell j. The result of calculating this statistic over all J genes between cell i and cell j is a vector, $V_{i,j}$, of size J. This vector of statistics is computed for all distinct pairs of the N cells in the data, and roughly indicates the difference in gene expressions between each pair. Each vector is then sorted in the decreasing order and truncated to the same length $n \leq J$, so that only the top n genes identified as DE are kept in each vector. For a discussion on the choice of n, refer to the end of this section. As a result, each cell is associated with an $n \times (N - 1)$ differential expression (DE) matrix, where each

column in the matrix is a truncated and sorted vector of statistics comparing the cell's expressions with one of the other $N - 1$ cells.

The above procedure to derive lists of DE genes for all pairs of two cells is a key component of SIDEseq, allowing SIDEseq to evaluate the similarity between two cells through examining their relationships with other cells. This is the novel and promising part of the SIDEseq technique which distinguishes it from other methods. We do not consider our statistic T in (1) to be the best or the only choice to define differential expression using only two expression values. Rather, we consider it a simple yet practical statistic that helps achieve our analysis goal. Furthermore, it generates satisfactory results. If a better statistic was found, we could replace T by it to further improve the performance of SIDEseq.

The second key calculation in SIDEseq, which produces the final similarity measure, is the evaluation of the consistency among the derived vectors or lists of DE genes that are relevant to every pair of cells. In more detail, for each $t = 1, \ldots, N, t \neq i, j$, the number of genes in the intersection of cell i and cell t's DE gene list and cell j and cell t's DE gene list is found. These numbers are summed across all $t \neq i, j$ to get the final SIDEseq measure of similarity between the two cells, which is expected to quantify the level of consistency between the cells' associated differential expression matrices. This measure is the element $S_{i,j}$ (and $S_{j,i}$) of the SIDEseq similarity matrix S. (Note that an alternative statistic is to normalize the number of genes in the intersection of the two lists by the number of genes in the union. This statistic has a monotonic relationship to the one we first introduced and they have generated almost equivalent results based on what we have observed. See Supplementary Materials for more details). To convert the similarity matrix into a dissimilarity matrix, we take the maximum value in the similarity matrix and subtract every value in the similarity matrix from the maximum value. The diagonal elements of the dissimilarity matrix are set to zero. For a further explanation of the SIDEseq measure, see the flowchart and toy example provided in the Supplementary Materials.

2.2.3 Selecting the number of DE genes, n

To determine n for a given dataset, one has to determine a number of genes which is large enough to capture the important biological relationships in the data, but small enough so that uninformative, noisy genes are not included. Plotting the values for several vectors of differential statistics is recommended to get an idea of an appropriate range for n in the dataset of interest. We found that in all three of the scRNA seq datasets focused on in this study, there was a range of genes which worked to give optimal clustering results. For the public human embryo dataset with relatively strong biological signals, anywhere from 150 to 500 genes could be used to get optimal clustering results, corresponding roughly to genes with differential statistics greater than two. More genes may be necessary for datasets with weaker biological variation of interest, such as with the human ovarian cancer cell dataset, where n from 600 to 3000 genes were appropriate. It should be noted that clustering results were stable within a range of choices of n, providing some flexibility when it comes to selecting this parameter.

2.2.4 Some Alternative Statistics

One might propose instead to identify a set of DE genes based on their expression profiles across *all* cells, as in the GiniClust algorithm, and then assess cell similarity based on these genes. However, we believe that genes identified in this way may likely have weaker subpopulation-specific information, since they were identified across all cells from multiple subpopulations. It is likely that these identified DE genes would not be as effective at distinguishing between subpopulations as the genes identified by SIDEseq, which may carry more explicit subpopulation information. One might also propose that separating DE genes by sign (positive versus negative differential expression) may improve the accuracy of the similarity measure, or that comparing two cells by examining their relationships with the *mean* expressions of the pooled, remaining cells to capture robust relationships may improve SIDEseq. However, we present a simulated dataset in the Supplementary Materials for which SIDEseq outperforms a measure such as the one just described above. See Supplementary Materials for a further explanation of the advantages of the SIDEseq measure and simulations which demonstrate some of SIDEseq's benefits over Euclidean distance, Pearson and Spearman correlation, etc.

It might be tempting to consider using varied n's across the lists of DE genes since the true number of DE genes could vary a lot between subpopulations. If the number of true DE genes in a list is much less than n, the list would contain many random non-DE genes and become noisy. However, we want to point out that no matter if the lists are noisy or not, as long as there is a reasonable mix of noisy and informative lists of DE genes, all the lists together will provide useful information to help distinguish cells from the different subpopulations. More discussion on this issue has been included in the Supplementary Materials. Although we have been largely happy with using the DE gene lists of the same length n, we also see that SIDEseq could be further improved if we do not have to pre-determine n and can effectively take into account the varied lengths of different DE gene lists. We consider achieving this by adapting the method in Li et al (2011)[1], where a score called the irreproducible discovery rate (IDR), analogous to a false discovery rate, was derived to measure reproducibility between findings from replicate experiments. We could use this IDR approach with slight modifications to assess how "reproducible" two lists of DE genes are and further to define "similarity" between two DE gene lists. Please see the Supplementary Materials for more details.

3 Results

3.1 Simulated Datasets

We used simulation studies to compare the performance of the SIDEseq measure with methods presented in the GiniClust algorithm of *Jiang et al.* which was designed to detect rare cell types using scRNA seq data. This is an important comparison because both methods rely on sets of identified DE genes to detect subpopulations or rare cell types, but the ways in which the two methods identify these genes are quite different. The GiniClust algorithm calculates a normalized Gini index for each gene by looking

at the gene's expression across all cells, and then selects the top Gini index genes for clustering and rare cell type identification. The SIDEseq measure, however, identifies DE genes between *every* cell pair, and then uses the lists of DE genes from all pairwise comparisons to quantify cell similarity. The similarity between two cells is calculated by looking at how consistent the lists of DE genes are that result from their pairwise comparisons with all other cells in the dataset. This integration of multiple lists of DE genes into the SIDEseq measure makes this novel similarity measure quite robust to the noise present in any single list of DE genes.

To simulate various single-cell datasets, we used the R package 'splatter' [29], with a variety of parameters designed to make the identification of subpopulations more challenging. Each simulated dataset consisted of several subpopulations, with different numbers of cells, probabilities of containing DE genes, mean expression of DE genes, etc. We then ran the GiniClust algorithm on each simulated dataset, with several variations on the parameters specifying minimum cell number, minimum point number, and epsilon. However, regardless of the specified parameters, the GiniClust algorithm only detected "rare cell types," and failed to identify the correct subpopulations of cells, each time clustering all cells not deemed as rare cell types into one large cluster. We believe this is because a relatively small number of genes passed the Gini index cutoff specified in the algorithm, and so there were not enough genes to accurately cluster the cells.

To further compare the GiniClust algorithm with the SIDEseq measure, specifically the way in which they identify and use DE genes, we selected the top Gini index genes (around 150) for each simulation and performed hierarchical clustering of the simulated data with Euclidean distance and Pearson and Spearman correlation. We then used the same number of genes to perform hierarchical clustering with the SIDEseq similarity measure, and compared the clustering results using the Adjusted Rand Index (ARI). Each dendrogram was cut according to the correct number of clusters, and the Adjusted Rand Index was used to compare the resulting clusters with the true subpopulations, with an ARI of one being perfect agreement with the truth and an ARI of zero corresponding to random assignment of cells to clusters. Results are shown in Table 1. In simulations 1 through 3, which all contain the same number of cells, genes, and subpopulations but vary in the degree and probability of differential expression, the SIDEseq measure outperforms all three common similarity measures. Simulations 4 and 5 increase the number of subpopulations, yet the SIDEseq measure still outperforms all others. In simulation 6, where each subpopulation has a different probability for DE genes and is arguably the most realistic model for a scRNA seq dataset, SIDEseq again outperforms all three common similarity measures. The results of the simulation studies suggest several points about the SIDEseq measure: (1) The method used by the SIDEseq measure of identifying and exploiting DE genes often outperforms methods like those found in GiniClust, where genes are identified as DE based on their expressions over all cells and (2) The SIDEseq similarity measure is able to uncover true subpopulations of cells in a variety of scRNA seq datasets, including those in which subpopulations have different probabilities of their genes being DE or have varying degrees of differential expression.

Table 1 Adjusted rand indices for simulated datasets

Simulated dataset index	# of subpopulations	#Cells, #Genes	Differential expression factor	Probability of differential expression	Pearson with Gini genes	Spearman with Gini genes	Euclidean with Gini genes	SIDEseq
1	5	240, 10^4	4	0.1	0.806	0.502	0.148	1
2	5	240, 10^4	4	0.05	0.471	0.653	0.404	1
3	5	240, 10^4	3	0.1	0.36	0.741	0.198	1
4	7	140, 10^4	4	0.05	0.462	0.367	0.222	1
5	7	140, 10^4	3	0.05	0.616	0.405	0.064	0.729
6	5	240, 10^4	4	(0.15, 0.1, 0.12, 0.05, 0.8)	0.446	0.5	0.007	0.552

4 Human Ovarian Cancer Cells

The human ovarian cancer cell dataset presents more of a challenging clustering task than the simulated dataset, due to the uncertain nature of the treatment factors, the passage number effects, the heterogeneous nature of ovarian cancer cells, etc. However, this challenging clustering task is useful for assessing the performance of the SIDEseq similarity measure. We clustered the 'RUVs' normalized counts in the TGFβ-1 group using hierarchical clustering with Euclidean distance, Pearson and Spearman correlation, and the SIDEseq similarity measure (Fig. S11). Since the cells did not cluster well according to treatment status, using the Adjusted Rand Index to compare clustering results from the various similarity measures is not meaningful. Instead, for this dataset we rely on visual inspection of the resulting dendrograms. When Spearman correlation is used for hierarchical clustering of the human ovarian cancer cell dataset, there are one or two resulting clusters of treatment cells, but cells largely fail to cluster by treatment status (Fig. 1). When the SIDEseq measure is used, three loose clusters of interest can be recognized (Fig. 1). One is a large cluster consisting of only untreated cells. Actually most untreated cells are in this cluster. Another cluster consists of a mix of treated and untreated cells. The third is a large cluster of mostly all treated cells. This cluster is on the outside of the sub-dendrogram formed by the other two clusters. In addition to clearer clusters of cells, the organization of the clusters within the dendrogram is also biologically interesting. The cluster that contains a mix of treated and untreated cells may correspond to a group of cells in the beginning stages of EMT or that have not entirely transitioned to the mesenchymal phenotype.

5 Human Embryo Cells

In order to further compare the SIDEseq measure with the common similarity measures, we did hierarchical clustering with different measures on an additional scRNA seq dataset from Yan (2012). They used a highly sensitive sequencing technique to obtain gene expressions from 124 human embryo cells in various stages of development. The dataset covers seven early developmental stages: metaphase II oocyte (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 dataset also includes an eighth stage of development of primary outgrowth during human embryonic stem cell (hESC) derivation (34 cells). Following the filtering method of [27] for this dataset, we used only RefSeq genes with at least one cell with RPKM expression greater than 0.1, resulting in roughly 21 thousand genes. However, while [27] only used cells from the first seven early developmental stages, we used all 124 cells for the clustering analysis.

Hierarchical clustering using Euclidean distance and Pearson and Spearman correlation grouped most cells by developmental stage, with the clusters of cells in the dendrogram following the natural progression of embryonic development (Fig. S14). Cells in later developmental stages (late blastocyst and hESC) clustered together on one side of the dendrogram, while cells in the earlier developmental stages (oocyte to morula) clustered on the other. Euclidean distance and Pearson and Spearman

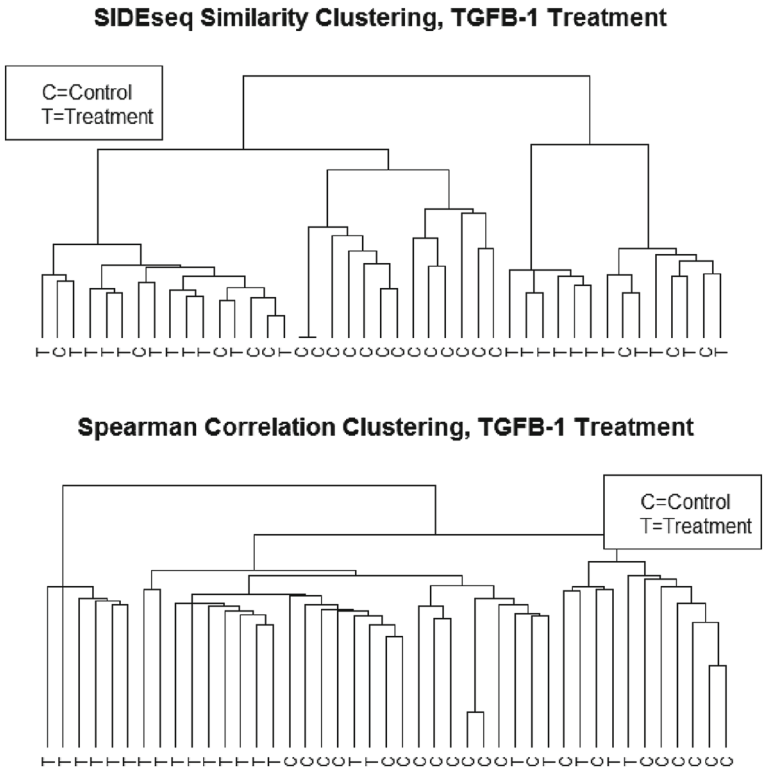

Fig. 1 Hierarchical clustering of the human ovarian cancer cells in the TGFβ-1 group using the SIDEseq measure and Spearman correlation. Cells were normalized with 'RUVs' normalization prior to clustering. Cells are colored by treatment (TGFβ-1) and control status. The SIDEseq measure results in three loose clusters of cells, with one large group of untreated (control) cells and one large group of mostly treated cells. Spearman correlation results in one or two clusters of treated cells, but cells do not appear to cluster as well by treatment and control status as they do with the SIDEseq measure

correlation, however, incorrectly clustered some of the cells in the earlier developmental stages. For example, Spearman correlation grouped four 8-cell stage cells with the earlier stages. Furthermore, Spearman correlation incorrectly clustered the 2-cell stage cells, separating them into two groups and clustering some of the 2-cell stage cells with the zygote cells. Two morula cells were clustered outside of the 8-cell and morula stage cells. It is interesting to note that simple hierarchical clustering using Spearman and Pearson correlation outperformed or matched the performance of more complex clustering methods for this dataset explored by [27]. Xu et al. used their proposed clustering algorithm, SNN-Cliq, to cluster 90 cells from this dataset (all cells except the hESC cells). They also used the k-means and DBSCAN algorithms with Euclidean distance. All methods were either matched in performance or outperformed by hierarchical clustering using Pearson or Spearman correlation.

When we used the SIDEseq measure for hierarchical clustering, it showed a slight improvement over the common similarity measures when clustering the cells in early

Table 2 Adjusted rand indices for hierarchical clustering of the embryo cell dataset

Public dataset	# Clusters	Adjusted Rand Index			
		Pearson correlation	Spearman correlation	Euclidean distance	SIDEseq similarity
Embryo cells, Xu et al.	7	0.740	0.880	0.812	0.880
Embryo cells, Xu et al.	8	0.659	0.770	0.823	0.770
Embryo cells, Xu et al.	9	0.740	0.744	0.740	0.828

developmental stages (Fig. S15). There was again a split between the early and later developmental stages, but with the SIDEseq measure, all of the 8-cell stage cells were clustered together. Two morula cells broke off from the morula cluster to cluster closer to the 8-cell stage cells, indicating that these may be cells in transition. These are the same two morula cells that were separated by Euclidean distance and Pearson and Spearman correlation, but with a different position in the dendrogram. Unlike the traditional similarity measures, the SIDEseq measure successfully clustered all cells in the 8-cell stage together. Furthermore, the SIDEseq measure perfectly clustered the very early stages of oocyte, zygote, and 2-cell stage cells. To provide a more quantitative comparison of the similarity measures, we cut each dendrogram into seven, eight (corresponding to the number of cell types), and nine clusters and calculated the ARI for each clustering method. See Table 2 for a full comparison of all four similarity measures. SIDEseq outperforms most similarity measures for most cluster number values, except in one case where it is outperformed by Euclidean distance. Here, we note that while the ARI values are informative, they should also be analyzed in the context of the original dendrograms. For example, while Euclidean distance has a higher ARI value than SIDEseq when eight clusters are used, SIDEseq outperforms Euclidean distance in terms of correctly classifying the early developmental stages (see Supplementary Figs. S13, S15). These subtle, yet important, clustering details are not taken into account by the ARI when the dendrograms are cut at seven, eight, nine, etc., clusters.

To further explore the subtleties in the clustering of this embryo dataset and compare the performance of the different similarity measures, we also used spectral clustering. For each similarity measure, we specified eight, nine, and ten clusters, performed spectral clustering 100 times (each time obtaining a local, optimal result), and recorded the average ARI values and their standard deviations (Table 3). We chose the number of clusters based on the distributions of the eigenvalues and corresponding eigengaps when different values of epsilon where used to build the epsilon graph. The SIDEseq measure outperformed all three traditional similarity measures for all three cluster values, with Spearman correlation being the second best measure. These results suggest that the SIDEseq similarity measure continues to outperform the common similarity measures when used with other clustering algorithms besides hierarchical clustering.

Table 3 Adjusted rand indices for spectral clustering of the embryo cell dataset

Public dataset	# Clusters	Average Adjusted Rand Index (sd)			
		Pearson correlation	Spearman correlation	Euclidean distance	SIDEseq similarity
Embryo, Xu et al.	8	0.472 (0.006)	0.718 (0.065)	0.681 (0.069)	0.757 (0.1)
Embryo, Xu et al.	9	0.635 (0.056)	0.745 (0.079)	0.670 (0.073)	0.804 (0.12)
Embryo, Xu et al.	10	0.694 (0.052)	0.747 (0.038)	0.670 (0.035)	0.785 (0.082)

Furthermore, when a more principled method is used to choose the number of clusters by using spectral clustering, SIDEseq's performance remains strong, if not improves.

6 Discussion

Exploratory data analysis is crucial in scRNA seq experiments before any analysis such as clustering is performed, and may reveal the need for normalization to remove unwanted sources of variation. This was undoubtedly the case for the human ovarian cancer cell dataset in this study. There was a clear difference between the cells in the two batches, likely due to both technical variation and induced biological variation as a result of a difference in passage number between batches. We see that in studies where passage number effects are present, the normalization task becomes very challenging. In fact, across-batch normalization may become impossible since this variation can be completely confounded with batch effects. The choice of normalization technique *within* batch proved to have an effect on the ability of cells to cluster by treatment status. Technical effects such as those observed in this study need to be kept in mind when performing scRNA seq analysis.

In our study, deriving and integrating lists of DE genes for all pairs of two cells stand as the key component, allowing SIDEseq to evaluate the similarity between two cells through examining their relationships with other cells. This is the novel and promising part of the SIDEseq technique which distinguishes it from other methods. Through studying simulated and real datasets with varying degrees of complexity, we observed the benefits of using the SIDEseq measure. In datasets where there are subtle but important differences between small subpopulations of cells, such as the cells in the early developmental stages of the embryo dataset, SIDEseq is able to very accurately identify subpopulations. Furthermore, in datasets where each subpopulation of cells has a different differential expression probability for its genes, SIDEseq seems to outperform traditional similarity measures. Even with datasets where the biological factor of interest is relatively weak, and may be masked by other sources of variability, the SIDEseq measure performs well compared to the commonly used similarity measures. Furthermore, SIDEseq can be utilized in many different clustering methods, such as hierarchical clustering and spectral clustering, to accurately identify subpop-

ulations. The success of SIDEseq is due to the novel way in which it identifies DE genes between each pair of cells, and uses the consistency among two cells' lists of DE genes (with all other cells) to define their similarity. In this way, SIDEseq is robust to noise in any single list of DE genes, and can investigate the dataset at a deeper level than other common similarity measures or clustering algorithms. These novel features of SIDEseq allow it to perform as well as or to outperform more complex clustering methods such as the GiniClust and SNN-Cliq algorithms, even when the measure is paired with a simple method such as hierarchical clustering.

As another interesting observation resulting from the study of the human ovarian cancer cell data, it seems clear that the thrombin-treated cells did not differentiate from the untreated cells in their batch as well as the TGFβ-1-treated cells diverged from their respective control cells. The findings from this study support the numerous experimental findings that TGFβ-1 is an inducer of EMT, but they do not provide evidence that thrombin is an EMT inducer. The ability of thrombin to induce EMT merits further investigation.

Acknowledgements Thanks to Sandrine Dudoit and Davide Risso for their help with the remove unwanted variation normalization methods. This work is partially supported by NIH U01 HG007031, NSF DMS-11-60319, NIH 5R21CA182375-01A1, Bakar's Fellows Program, Strategic Priority Research Program of the Chinese Academy of Sciences [XDB13040700], and the National Natural Science Foundation of China [91529303, 61134013, 91439103].

References

1. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106
2. Asiedu MK, Beauchamp-Perez FD, Ingle JN, Behrens MD, Radisky DC, Knutson KL (2014) AXL induces epithelial-to-mesenchymal transition and regulates the function of breast cancer stem cells. Oncogene 33(10):1316–1324
3. ATCC (2010) Passage number effects in cell lines. Tech Bull. https://www.atcc.org/sim/media/PDFs/Technical%20Bulletins/tb07.ashx
4. Bullard J et al (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinform 11:94
5. Eberwine J et al (2014) The promise of single-cell sequencing. Nat Methods 11:25–27
6. Eisenberg E et al (2013) Human housekeeping genes revisited. Trends Genet 29(10):569–574
7. Gorges TM, Pantel K (2013) Circulating tumor cells as therapy-related biomarkers in cancer patients. Cancer Immunol Immunother 62:931–939
8. Gorges TM, Tinhofer I, Drosch M, Rose L, Zollner TM, Krahn T, von Ahsen O (2012) Circulating tumour cells escape from EpCAM-based detection due to epithelial-to-mesenchymal transition. BMC Cancer 12:178
9. Gou WF et al (2014) The role of RhoC in epithelial-to-mesenchymal transition of ovarian carcinoma cells. BMC Cancer 14:477
10. Grun D et al (2015) Digital synthesis of plucked-string and drum timbres. Nature 525:251–255
11. Grun D, Kester L, Van Oudenaarden A (2014) Validation of noise models for single-cell transcriptomics. Nat Am 11(6):637–643
12. Hansen KD (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 13(2):204–216
13. Jang H et al (2012) Transformation of epithelial ovarian cancer stemlike cells into mesenchymal lineage via EMT results in cellular heterogeneity and supports tumor engraftment. Mol Med 18:1197–1208
14. Jiang L et al (2016) GiniClust: detecting rare cell types from single-cell gene expression data with Gini Index. Genome Biol 17:144
15. Jiang P et al (2016) Quality control of single-cell RNA-seq by SinQC. Bioinformatics 32(11):1–3

16. Kasimir-Bauer S, Hoffmann O, Wallwiener D, Kimmig R, Fehm T (2012) Expression of stem cell and epithelial-mesenchymal transition markers in primary breast cancer patients with circulating tumor cells. Breast Cancer Res 14(1):R15
17. Levine J et al (2015) Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell 162:184–197
18. Lin H-K et al (2003) Suppression versus induction of androgen receptor functions by the phosphatidylinositol 3-kinase/Akt pathway in prostate cancer LNCaP cells with different passage numbers. J Biol Chem 51:50902–50907
19. Mani SA, Guo W, Liao MJ, Eaton EN, Ayyanan A, Zhou AY, Brooks M, Reinhard F, Zhang CC, Shipitsin M, Campbell LL, Polyak K, Brisken C, Yang J, Weinberg RA (2008) The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell 133(4):704–715
20. O'Driscoll L et al (2006) Phenotypic and global gene expression profile changes between low passage and high passage MIN-6 cells. J Endocrinol 191:665–676
21. Ozkumur E, Shah AM, Ciciliano JC, Emmink BL, Miyamoto DT, Brachtel E, Yu M, Chen PI, Morgan B, Trautwein J, Kimura A, Sengupta S, Stott SL, Karabacak NM, Barber TA, Walsh JR, Smith K, Spuhler PS, Sullivan JP, Lee RJ, Ting DT, Luo X, Shaw AT, Bardia A, Sequist LV, Louis DN, Maheswaran S, Kapur R, Haber DA, Toner M (2013) Inertial focusing for tumor antigen-dependent and -independent sorting of rare circulating tumor cells. Sci Transl Med 5(179):179ra47
22. Ramsköld D et al (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol 30:777–782
23. Risso D et al (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol 32(9):896–902
24. Sandberg R (2014) Entering the era of single-cell transcriptomics in biology and medicine. Nat Methods 11:22–24
25. Stegle O, Teichmann SA, Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 16:133–145
26. Stegle O et al (2012) Using Probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analysis. Nat Protoc 7(3):500–507
27. Xu C, Sui Z (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method., Bioinformatics Advance. Access 31(12):1974–1980
28. Yan L et al (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol 20:1131–1139
29. Zappia L, Phipson B, Oshlack A (2017) splatter: Simple Simulation of Single-cell RNA Sequencing Data. R package version 0.99.10. https://github.com/Oshlack/splatter
30. Zeisel A (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347(6226,):1138–1142
31. Zhong Y-C (2013) Thrombin promotes epithelial ovarian cancer cell invasion by inducing epithelial-mesenchymal transition. J Gynecol Oncol 24(3):265–272