Bioinformatics, 35(6), 2019, 962–971
doi: 10.1093/bioinformatics/bty708
Advance Access Publication Date: 28 August 2018
Original Paper

Gene expression

A cluster robustness score for identifying cell subpopulations in single cell gene expression datasets from heterogeneous tissues and tumors

Itamar Kanter[1], Piero Dalerba[2] and Tomer Kalisky [1,*]

[1]Department of Bioengineering and Bar-Ilan Institute of Nanotechnology and Advanced Materials (BINA), Bar-Ilan University, Ramat Gan 52900, Israel and [2]Department of Pathology and Cell Biology, Department of Medicine (Division of Digestive and Liver Diseases), Herbert Irving Comprehensive Cancer Center (HICCC), and the Columbia Stem Cell Initiative (CSCI), Columbia University, New York, NY 10032, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on September 4, 2017; revised on April 18, 2018; editorial decision on August 14, 2018; accepted on August 23, 2018

Abstract

Motivation: A major aim of single cell biology is to identify important cell types such as stem cells in heterogeneous tissues and tumors. This is typically done by isolating hundreds of individual cells and measuring expression levels of multiple genes simultaneously from each cell. Then, clustering algorithms are used to group together similar single-cell expression profiles into clusters, each representing a distinct cell type. However, many of these clusters result from overfitting, meaning that rather than representing biologically meaningful cell types, they describe the intrinsic 'noise' in gene expression levels due to limitations in experimental precision or the intrinsic randomness of biochemical cellular processes. Consequentially, these non-meaningful clusters are most sensitive to noise: a slight shift in gene expression levels due to a repeated measurement will rearrange the grouping of data points such that these clusters break up.

Results: To identify the biologically meaningful clusters we propose a 'cluster robustness score': We add increasing amounts of noise (zero mean and increasing variance) and check which clusters are most robust in the sense that they do not mix with their neighbors up to high levels of noise. We show that biologically meaningful cell clusters that were manually identified in previously published single cell expression datasets have high robustness scores. These scores are higher than what would be expected in corresponding randomized homogeneous datasets having the same expression level statistics. We believe that this scoring system provides a more automated way to identify cell types in heterogeneous tissues and tumors.

Contact: tomer.kalisky@gmail.com

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

1 Introduction

Single cell technologies can measure gene expression from multiple genes in hundreds or thousands of individual cells (Dalerba *et al.*, 2011; Hashimshony *et al.*, 2016; Jaitin *et al.*, 2014; Klein *et al.*, 2015; Macosko *et al.*, 2015; Treutlein *et al.*, 2014; Villani *et al.*, 2017; Zeisel *et al.*, 2015). A major aim of single cell biology is to identify and characterize important cell subpopulations in heterogeneous biological systems such as a developing embryo, a regenerating tissue, or a tumor. For example, an embryo contains multiple types of stem and progenitor cell populations that create, through carefully regulated interactions, the various lineages that are required for forming properly functioning organs (Chen *et al.*, 2015; Guo *et al.*, 2010; La Manno *et al.*, 2016; Swiers *et al.*, 2013; Treutlein *et al.*, 2014). Similar processes continue throughout the lifetime of the fully developed organism, where many tissues contain small populations of 'tissue-specific stem cells' that are responsible for normal tissue turnover and regeneration (Barker *et al.*, 2007; Lyubimova *et al.*, 2013; Montgomery *et al.*, 2011; Sangiorgi and Capecchi, 2008; Shackleton *et al.*, 2006; Spangrude *et al.*, 1988; Stingl *et al.*, 2006). Similarly, it has been shown that many types of tumors are driven by small populations of 'cancer stem cells' that are thought to sustain the tumor through their ability of self-renewal and pluripotency (Al-Hajj *et al.*, 2003; Bonnet and Dick, 1997; Bussolati *et al.*, 2008; Dalerba *et al.*, 2007; Li *et al.*, 2007; Pode-Shakked *et al.*, 2009, 2013; Prince *et al.*, 2007; Singh *et al.*, 2004). Although these rare cell populations have a large effect on the overall behavior of the system, they are difficult to identify and characterize since they often consist of a small fraction (often less than 1%) of the tissue/tumor, and are therefore likely to be 'averaged out' in bulk gene expression measurements. Therefore, single cell measurements are required.

In a typical single-cell workflow, the tissue or tumor is dissociated into individual cells, and the expression levels of multiple genes are measured in each cell (Fan *et al.*, 2015; Picelli *et al.*, 2014; Sanchez-Freire *et al.*, 2012; Sheng *et al.*, 2017). The resulting single cell expression dataset typically consists of a matrix in which each row (or column) represents an individual cell and each column (or row) represents a specific gene. Then, clustering algorithms are used to identify groups (=clusters) of cells with similar gene expression profiles, where each such group represents a distinct cell type such as a stem cell, a progenitor cell, or a more differentiated cell type with a specialized function. Likewise, it is possible to use clustering algorithms to identify groups of genes with similar expression patterns, which might point to a functional relationship between them (e.g. a common transcriptional activator).

Single cell analysis results in hundreds or thousands of data points, each being a vector of expression levels measured from multiple genes in an individual cell. The objective of clustering analysis is to form groups of data points such that the distances (or more generally, 'dissimilarities') between data points within each group are smaller than the distances between data points from different groups (Kaufman and Rousseeuw, 1990). The dissimilarity $d(i,j)$ between two data points i and j is usually calculated from the Euclidean distance or from the correlation between them.

One approach often used for single cell datasets is agglomerative hierarchical clustering (Dalerba *et al.*, 2011; Guo *et al.*, 2010, 2013; Rothenberg *et al.*, 2012; Treutlein *et al.*, 2014). The agglomerative hierarchical clustering algorithm constructs a hierarchy of clusters 'bottom-up' (Fig. 1A): Initially, each data point is considered a separate cluster. Then, in each step of the algorithm, the two most similar clusters are merged (such that the number of clusters is decreased by one) and the distance between the new set of clusters is recalculated. The distance between two clusters can be calculated, for example, by taking the average of all pair-wise dissimilarities $d(i,j)$, where i is a data point in one cluster and j is a data point in the other (average linkage). The algorithm proceeds until all clusters are merged into a single cluster encompassing all data points. The hierarchical relations between the clusters can be described by a dendrogram (Fig. 1A).

Given N data points, (e.g. cells), the above algorithm creates $N-1$ hierarchically arranged clusters. However, many of these clusters do not represent biologically meaningful groups of cells. For example, the grouping of data points at the bottom of the dendrogram (the leaves) is in many cases a result of 'overfitting': These points are often so close to each other that the relative distances between them are heavily affected by measurement noise (the uncertainty in experimental precision) or by biological noise (the inherent randomness in biochemical cellular processes). This means that the differences in distances between these points—that determines which points are grouped with each other by the clustering algorithm—are irrelevant to the classification of different cell types. Consequentially, these non-meaningful clusters are sensitive to noise: A slight change in gene expression levels from a repeated measurement or a replicate sample will change the relative distances between data points and rearrange the clusters such that points that belonged to the same cluster will now belong to two separate clusters (Fig. 1B). The same may also happen at higher hierarchies. For example (Fig. 1C), a cluster that is almost equally distanced from two other clusters is not robustly grouped to any one of them. A slight change in gene expression of a single data point may regroup this cluster with a different neighboring cluster.

In order to identify biologically meaningful cell populations we propose to use robustness analysis to calculate a 'robustness score' for each cluster: We add increasing amounts of noise (zero mean and progressively increasing variance) and check which clusters are most robust, in the sense that they do not get 'mixed-up' with their neighbors. Our hypothesis is that biologically meaningful clusters will keep their composition under progressively increasing levels of noise and will, therefore, receive high robustness scores. We demonstrate this method by identifying robust cell subpopulations from previously published single cell gene expression datasets for which the cell subpopulation repertoire is relatively well understood (Björklund *et al.*, 2016; Dalerba *et al.*, 2011; Patel *et al.*, 2014; Rothenberg *et al.*, 2012).

2 Materials and methods

2.1 Calculation of cluster robustness scores

Going from the root of the dendrogram to the leaves, each branching point divides a single cluster into two sub-clusters (Fig. 2A). Let us focus on one arbitrarily chosen branching point (Fig. 2A, arrow) and label its sub-clusters (as well as the data points within them) as 'A' and 'B'. The general aim of the clustering algorithm is to assign the labels 'A' and 'B' in an optimal way, such that the distance between data points having the same label (e.g. the distance between 'A' and 'A') is smaller than the distance between data points having different labels (the distance between 'A' and 'B'). We propose to test how robust to noise is the optimality of this partitioning to sub-clusters 'A' and 'B' at each branching point.

One way to evaluate the optimality of a cluster is the 'Silhouette' score (Kaufman and Rousseeuw, 1990; Rousseeuw, 1987), which compares the average 'inner' distance within the cluster to the average 'outer' distance between the cluster and its closest neighbors (Fig. 2B). The Silhouette score can be calculated as follows: For each data point i belonging to a specific cluster 'A', we can calculate $a(i)$, which is the average dissimilarity between data point i to all other data points belonging to the same cluster (in our case, all data points labeled 'A'). We next calculate the average dissimilarity between data point i to all other data points belonging to any other cluster. We recalculate this for all other clusters and take the minimum, which we define as $b(i)$. Thus, $b(i)$ is the minimal averaged distance from data point i to the closest neighboring cluster (which in our case is cluster 'B').

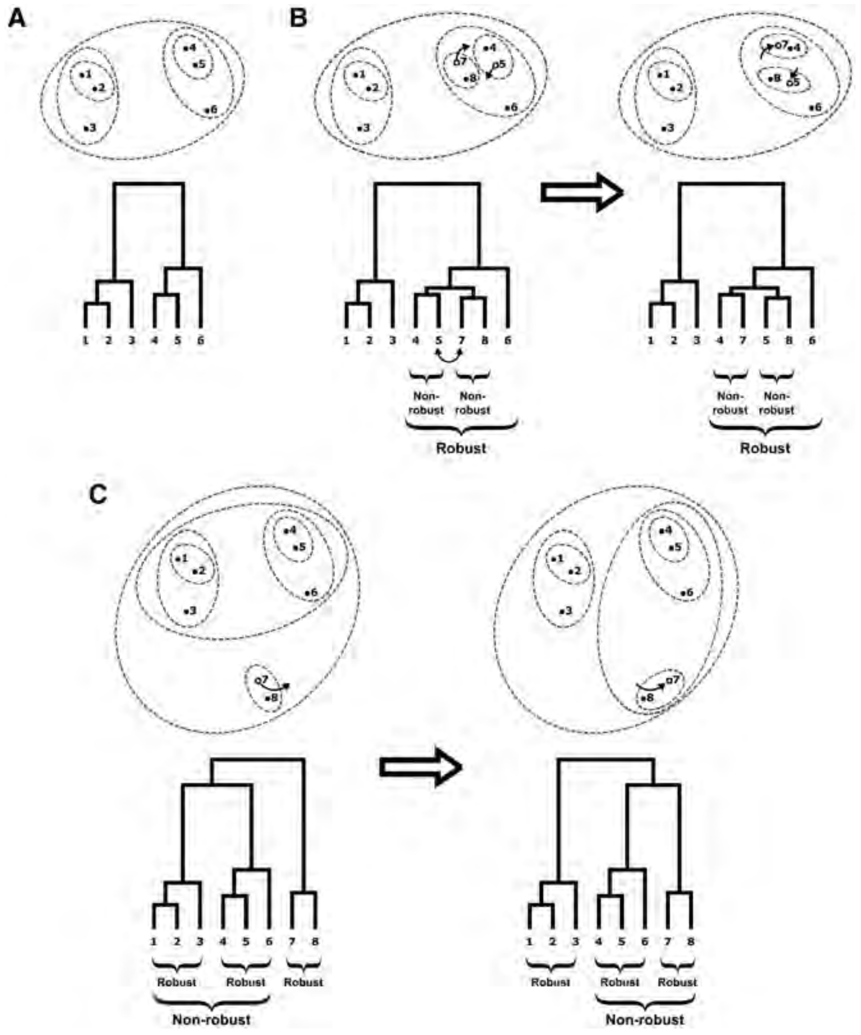

Fig. 1. Hierarchical clustering of single cell gene expression datasets identifies biologically meaningful clusters representing distinct cell types (or gene families), as well as numerous non-meaningful clusters that result from overfitting. The later are non-robust to noise. (**A**) A sketch of the agglomerative hierarchical clustering procedure: At first, each data point (representing a gene expression profile from an individual cell) is considered a separate cluster ('1', '2', '3', …). Then, in each step of the algorithm, the two closest clusters are merged and the distance between all clusters is recalculated. The process continues until all clusters are merged into a single cluster containing all data points. In parallel, a dendrogram is constructed to represent the relationships between the different clusters. In the illustrated example, clusters '1' and '2' are merged first to create the cluster '1-2', then '4' and '5' are merged into cluster '4-5', then point '3' is merged with cluster '1-2' to create cluster '1-2-3' etc. (**B**) Clusters at the base of the dendrogram such as '4-5' and '7-8' are in many cases a result of overfitting and do not represent distinct cell phenotypes. In this example, data points 4, 5, 7, 8 are so close to each other such that the differences in distances between them are below the precision of the measurement apparatus or the inherent biochemical variation of the gene expression mechanism and are not due to an actual difference in cell type. A repeated measurement might result in points '5' and '7' being slightly displaced such that the clustering algorithm will rearrange clusters '4-5' and '7-8' into '4-7' and '5-8'. Therefore, clusters '4-5' and '7-8' are non-robust to noise. However clusters '1-2-3' and '4-5-7-8-6' are much more robust to noise since the distances between them are large and a much larger noise displacement is required to move any data point from one cluster to the other. (**C**) Overfitting may also affect higher hierarchies. In this example clusters '1-2-3', '4-5-6' and '7-8' are almost equidistant from each other such that the differences between their distances are below the precision of the measurement apparatus or the intrinsic biochemical noise. A slight displacement of data point '7' results in separating cluster '4-5-6' from cluster '1-2-3-4-5-6' and joining it with cluster '7-8' to create cluster '4-5-6-7-8'. A similar displacement may reverse this process. Thus, cluster '1-2-3-4-5-6' is non-robust since a small random displacement of one data point may cause it to disappear. However, clusters '1-2-3', '4-5-6' and '7-8' will not be affected and are therefore more robust

The Silhouette score for data point i is: $S(i) = (b(i) - a(i))/max\{a(i), b(i)\}$ and is bound between -1 and 1. $S(i) = 1$ means that point i has been appropriately assigned into cluster 'A', since its average distance to points in any other cluster (in particular, those belonging to the neighboring cluster 'B') is much larger than its average distance to other points within 'A'. $S(i) \approx 0$ implies that it is not clear whether data point i has been appropriately assigned to cluster 'A', since the average distance to cluster 'B' is not much larger than the average distance to cluster 'A'. $S(i) = -1$ indicates that data point i has been 'miss-classified' as belonging to cluster 'A', and that it would have been much more appropriate to assign it to cluster 'B'. To calculate the Silhouette score of a whole cluster, we average the silhouette score over all its members, for example, the Silhouette score of cluster 'A' is the average of $S(i)$ over all points i that belong to cluster 'A' ($S = \langle S(i) \rangle_{i \in A}$). A tightly grouped cluster in which all data points are close to each other and are far from points of neighboring clusters will have a high Silhouette score (Fig. 2B).

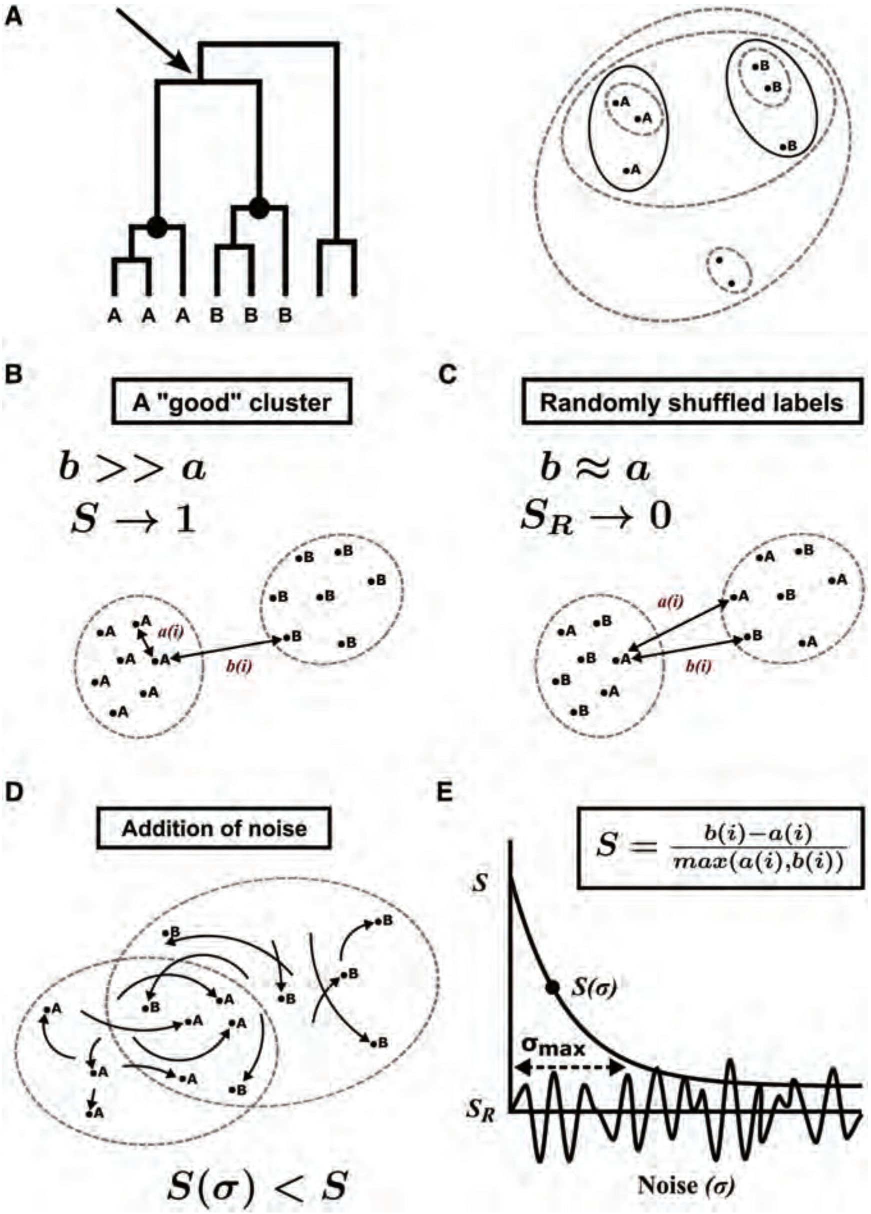

Fig. 2. Calculation of cluster robustness scores. (**A**) Each branching point of the dendrogram, such as the one pointed to by an arrow in the illustrated example, represents a cluster being optimally partitioned into two sub-clusters, here labeled as 'A' and 'B'. The aim of the clustering algorithm is to assign the labels 'A' and 'B' such that the distance between all data points having the same label (e.g. 'A' and 'A') is smaller than the distance between data points having different labels ('A' and 'B'). (**B**) The Silhouette measure for cluster optimality compares the average 'inner' distance within the cluster (*a*) to the average 'outer' distance between the cluster and its closest neighbors (*b*). A tightly grouped cluster in which all data points are close to each other and are far from data points in neighboring clusters ($b \gg a$) will have a high Silhouette score *S*. (**C**) For an optimal cluster, any mixing of the labels 'A' and 'B' (by randomly interchanging labels 'A' and 'B' between pairs of data points from the two clusters) will result in smaller Silhouette values, i.e. $S_R < S$ for any label-mixing realization *R*. (**D**) Addition of noise with zero mean and variance σ^2 will generally 'spread out' the coordinates of points 'A' and 'B', resulting in larger inner distances within a cluster, smaller outer distances between different clusters, and reduced Silhouette scores, i.e. $S(\sigma) < S$. (**E**) Although increased noise will reduce the Silhouette score $S(\sigma)$, the most robust clusters will retain their optimality ($S(\sigma) > S_R(\sigma)$) up to high levels of noise (σ). We define the cluster robustness score as the maximal noise standard deviation σ_{max} such that $S(\sigma) > S_R(\sigma)$ (for a pre-determined fraction of label shuffling realizations *R*)

At any given branching point of the dendrogram, the partitioning of data points into sub-clusters 'A' and 'B' should be optimal such that the Silhouette measure *S* of sub-cluster 'A' (and also sub-cluster 'B') will be maximized. Thus, for an optimal partitioning into subclusters, the labels 'A' and 'B' cannot be randomly shuffled: Any mixing of the labels 'A' and 'B' (by randomly interchanging labels 'A' and 'B' between pairs of data points from the two clusters) will result in smaller Silhouette values, i.e. $S_R < S$ for any label-mixing realization *R* (Fig. 2C). We hypothesize that in clusters that are most biologically meaningful this optimality will be most robust to noise, i.e. *S* will remain higher than S_R even in the presence of large noise.

An addition of noise will randomly 'spread out' the coordinates of data points 'A' and 'B', resulting in larger inner distances and smaller outer distances (Fig. 2D). This will result in lower Silhouette measures. The Silhouette measure *S* of cluster 'A' (or cluster 'B') with noise will be lower than without it, i.e. $S(\sigma) < S$ for any additive (Gaussian) noise with zero mean and variance σ^2. Therefore, in order to evaluate the cluster robustness to noise, we will check how much noise can be added such that its Silhouette score S remains higher than S_R (Fig. 2E). For robust clusters, S will remain larger than S_R even for large levels (=variance) of noise. More formally, for each branching point we are actually testing for the null hypothesis that the labels 'A' and 'B' can be randomly shuffled, and checking if this null hypothesis can be rejected under increasing levels of additive noise.

We propose the following procedure (Fig. 2E): At each branching point of the dendrogram, we first calculate S for both cluster 'A' and cluster 'B' (with respect to each other). Then, for each cluster ('A' and 'B', independently) we compare S to the values S_R resulting from randomly shuffling the labels 'A' and 'B' multiple times. Then, we add Gaussian noise with zero mean and progressively increasing levels of variance (σ^2), recalculate $S(\sigma)$ and $S_R(\sigma)$ for multiple realizations of label-shuffling R, and check the fraction of instances (i.e. label shuffling realizations R) for which the specific labeling of A's and B's remains optimal, i.e. $S(\sigma) > S_R(\sigma)$. The most robust clusters will remain optimal for higher levels of noise σ. Therefore, for each cluster, we define the **cluster robustness score** as the maximal noise standard deviation (σ_{max}) such that $S(\sigma) > S_R(\sigma)$ in more than some predetermined fraction of label shuffling realizations R. This fraction is taken to be 1-p_threshold, where the more label-shuffling realizations R taken, the lower p_threshold can be set. We typically used p_threshold = 0.05 and 100 label-shuffling realizations R, but we found similar results for other values as well (p_threshold = 0.005, 0.01, 0.1; 1000 label-shuffling realizations R).

For single-cell expression datasets we used the Pearson correlation distance or Euclidean distance as the dissimilarity measure between data points. The Pearson correlation distance is calculated as follows: if i and j are vectors of expression levels of multiple genes from two individual cells, then the dissimilarity between them is: $d(i,j) = 1 - corr(i,j)$. For calculating the distance between two clusters we took the average of all pair-wise dissimilarities between them (average linkage), for example, the distance between clusters 'A' and 'B' is: $d(A,B) = 1/(|A||B|)\sum_{i\in A,\ j\in B} d(i,j)$, or Ward's method (Kaufman and Rousseeuw, 1990). We provide a program written in Matlab (2016b) for performing our scoring procedure (Supplementary Material), as well as results from simulated data (Supplementary Material).

We tested the cluster robustness scores of biologically meaningful cell sub-populations that were previously identified by manual inspection of single cell gene expression datasets.

3 Results

We first tested our cluster scoring scheme on three simulated datasets (see Supplementary Material). Then, we tested our method on two published single cell qPCR datasets (Dalerba *et al.*, 2011; Rothenberg *et al.*, 2012) and two published single-cell RNA-seq datasets (Björklund *et al.*, 2016; Patel *et al.*, 2014).

3.1 Example no. 1: A single cell qPCR dataset from a mouse colon epithelium

In a previous study, a combination of multicolor flow cytometry and microfluidic single cell qPCR was used to identify different cell types within the mouse colon crypt (Rothenberg *et al.*, 2012). Briefly, colon epithelial cells from mice were dissociated and isolated by FACS, and then a few hundreds of single cells from the base of the crypt were profiled by microfluidic multiplex qPCR according to a panel of pre-selected genes from the literature. Single cell gene expression values were standardized and hierarchical clustering was performed (Fig. 3A). By manual inspection of the clustering heatmap and dendrogram, five major cell populations were observed, each representing a different cell type or transcriptional state (Rothenberg *et al.*, 2012). The cell populations were labeled as follows:

- Bmi1-high/Lefty1-high cells (Sub-population A): Cells mostly expressing high levels of the genes BmI1, Lefty1 and Gapdh, and mostly lacking immature cell markers (Lgr5, Axin2) and Goblet cell markers (Muc2, Spdef).
- Mature enterocytes (Sub-population B): Expressing high levels of Slc26a3 and Krt20.
- Goblet cells (Sub-population C): Expressing high levels of Spdef, Muc2, Agr2 and Tff3.
- Immature cells (Sub-population D): Containing cells expressing high levels of genes known to be over-expressed in the crypt-base, such as Cftr, Notch1, Axin2 and Ascl2. This cell population contains an additional sub-population of Lgr5-high cells (Sub-population E) that contains nearly all cells over-expressing the gene Lgr5 which is the putative stem cell marker for the colon epithelium.

We first calculated the Silhouette scores for all clusters (Fig. 3B). For each cluster, the Silhouette score was calculated relative to its nearest neighboring cluster originating from the same branching point. We found that there is a large bias with respect to cluster size such that even small meaningless clusters have large Silhouette scores. We next calculated the cluster robustness scores (σ_{max}) for each cluster (Figs 3C and D) and found that the biologically meaningful cell subpopulations—those identified by manual inspection—typically have high robustness scores. Note that although small clusters tend to have higher Silhouette values, they also have a larger dispersion of S_R values for different label-mixing realizations R (Fig. 3D).

We next checked if clusters of comparable robustness also appear in corresponding 'homogeneous' tissue with similar gene expression statistics. We therefore 'homogenized' the dataset as follows: We randomly permuted the gene expression values of each gene separately among the different cells (i.e. randomly permute the values in each column in Fig. 3A). In this way we removed the cross-correlations between different genes while conserving the overall distribution of each gene. As a result, cell subpopulations that are characterized by families of coordinately expressed genes in the heterogeneous tissue are lost. We found that the cluster robustness scores in the 'homogenized' dataset are considerably lower (Fig. 3C). In particular, most biologically meaningful clusters in the original dataset have higher robustness values than clusters of comparable size in the homogenized dataset.

We obtained similar results for single cell qPCR data from the human colon (see Supplementary Fig. S4 and Supplementary Material).

3.2 Example no. 2: A single cell RNA sequencing dataset from innate lymphoid cells (ILCs)

To demonstrate our cluster robustness measure on single cell RNA sequencing data, we downloaded single cell RPKM values from several hundreds of individual tonsil CD127+ Innate Lymphoid Cells (ILC's) and natural killer (NK) collected by Björklund *et al.* (2016). After preprocessing (Supplementary Fig. S5), we obtained a gene expression matrix of 311 highly variable genes from 300 cells (Macosko *et al.*, 2015) (Fig. 4A). In this dataset, four known cell populations (ILC1, ILC2, ILC3 and NK cells) were isolated by flow cytometry using surface markers. mRNA from individual cells was sequenced and all single-cell expression profiles were analyzed together.

We find that the clusters that correspond to the four well-known cell populations (ILC1, ILC2, ILC3 and NK cells) have high cluster

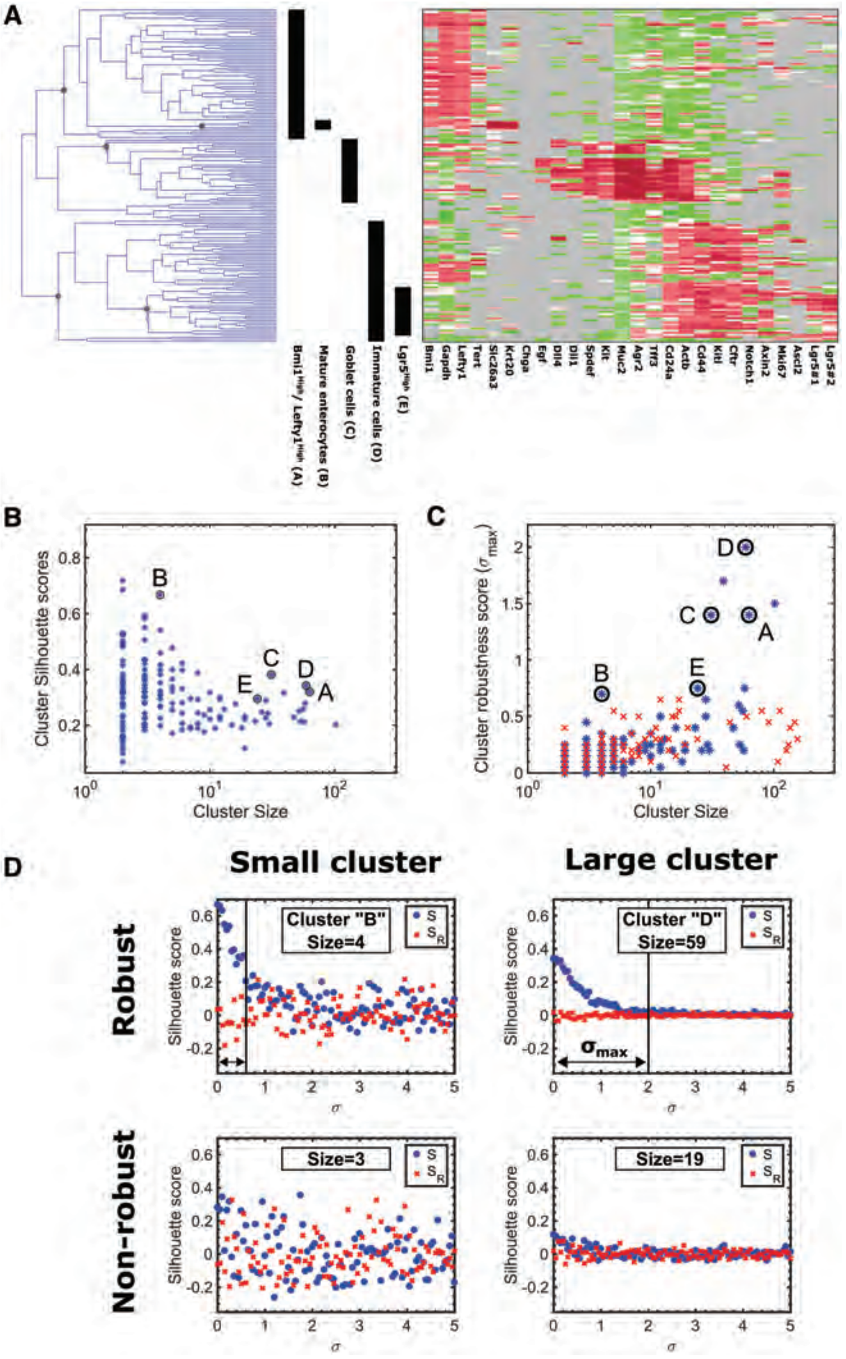

Fig. 3. Biologically meaningful cell sub-populations from a previously published mouse colon epithelium single cell qPCR dataset (Rothenberg *et al.*, 2012) have high cluster robustness scores. (**A**) Hierarchical clustering of single cell gene expression measurements from 161 individual cells isolated from the bases of mouse colon crypts and profiled by microfluidic single cell qPCR as previously described (Rothenberg *et al.*, 2012). Each column represents a gene and each row represents a cell. Each column (=gene) was standardized by subtracting the mean, dividing by 3 times the standard deviation, and truncating to the range [-1, 1]. Manual inspection reveals 5 biologically meaningful cell sub-populations (Red—over-expression, green—under-expression, gray—no expression). (**B**) Cluster Silhouette scores are highly biased with respect to cluster size, such that even meaningless clusters of small size may have high Silhouette values. (**C**) Biologically meaningful cell sub-populations typically have high cluster robustness scores that are higher than expected for clusters of similar size in a corresponding 'homogenized' dataset. Shown is the cluster size versus cluster robustness score (σ_{max}) for all clusters in the single cell gene expression dataset (*, blue) and its 'homogenized' counterpart (x, red). The 'homogenized' dataset was obtained by randomly permuting the values of each column (=gene), such that all cell subpopulations characterized by families of coordinately expressed genes are lost but the overall expression level distribution of each gene is conserved. (**D**) Calculation of robustness scores for selected clusters. Shown is the Silhouette score S (o, blue) and the label-mixed Silhouette score S_R (x, red) as a function of noise with increasing variance σ^2. Robust clusters retain their optimality ($S > S_R$) for higher values of noise (σ). Note that small clusters tend to have higher Silhouette values and a larger dispersion of S_R values (for different label-mixing realizations R). Clustering was done with Pearson correlation distance and average linkage. Cluster Silhouette scores were calculated using Pearson correlation distance. We used p_threshold = 0.05 and 100 label-shuffling realizations R (though the results did not change significantly for 1000 label-shuffling realizations)

robustness scores (Fig. 4C). We find additional putative sub-clusters with relatively high robustness scores (ILC3-A and B, ILC1 CCL5+ and CCL5-), which might represent a finer partition of the four major well-known populations. Note that a subpopulation of CCL5-producing ILC1 cells was previously identified in the small intestine by Gury-BenAri *et al*. (2016).

3.3 Example no. 3: A single cell RNA sequencing dataset from glioblastoma

We tested our method on a second single cell RNA sequencing dataset from 430 cells from five primary glioblastomas collected by Patel *et al*. (2014), using TPM values that were calculated by Townes *et al*. (2017). After preprocessing (Supplementary Fig. S6), we

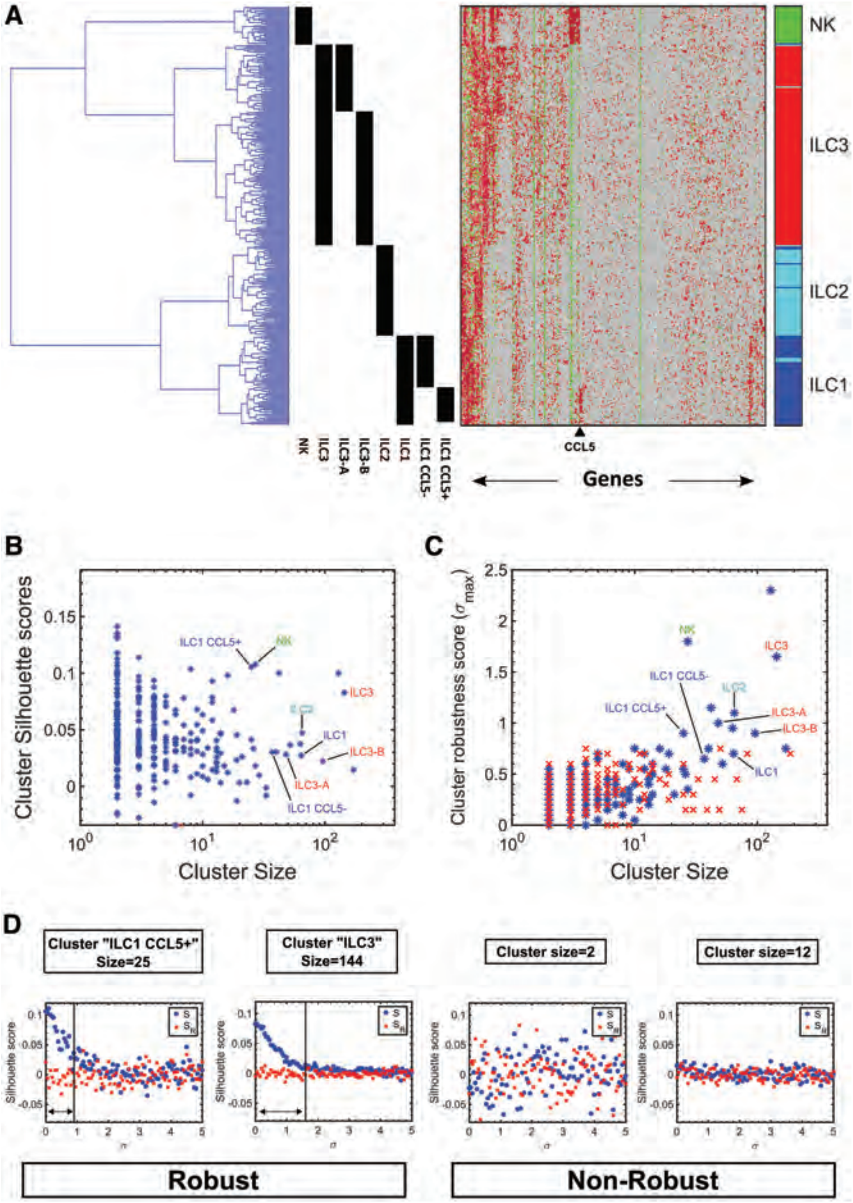

Fig. 4. Biologically meaningful cell sub-populations from a previously published single-cell RNA sequencing dataset of Innate Lymphoid Cells (ILCs) have high cluster robustness scores. (**A**) Hierarchical clustering of single cell gene expression measurements from 300 individual cells isolated from human tonsil tissue and profiled by the Smart-seq2 single cell RNA sequencing protocol (Björklund *et al.*, 2016). After choosing 311 highly variable genes (see Supplementary Fig. S5 and Supplementary Material) and performing a log-plus-one transformation, each gene (=column) was standardized by subtracting the mean, dividing by the standard deviation and truncating to the range [-1, 1] (Red—over-expression, green—under-expression, gray—no expression). The clustering distinguishes between the four known cell populations (ILC1, ILC2, ILC3 and NK cells) that were defined by surface markers and isolated using flow cytometry. Manual inspection reveals additional putative sub-populations (ILC3-A and B, ILC1 CCL5+ and CCL5-). (**B**) Cluster Silhouette scores are highly biased with respect to cluster size, such that even meaningless clusters of small size may have high Silhouette values. (**C**) The four known cell populations (ILC1, ILC2, ILC3 and NK cells), as well as the putative cell subpopulations (e.g. ILC1 CCL5+) have high cluster robustness scores (*, blue) that are higher than expected for clusters of similar size in a corresponding 'homogenized' dataset (x, red). (**D**) Calculation of robustness scores for selected clusters. Shown is the Silhouette score S (o, blue) and the label-mixed Silhouette score S_R (x, red) as a function of noise with increasing variance σ^2. Robust clusters retain their optimality ($S > S_R$) for higher values of noise (σ). Small clusters tend to have higher Silhouette values and a larger dispersion of S_R values (for different label-mixing realizations R). Clustering was done with Euclidean distance and Ward's linkage. Cluster Silhouette scores were calculated using Euclidean distance. We used p_threshold = 0.05 and 100 label-shuffling realizations R

obtained a gene expression matrix of 708 highly variable genes from 430 cells (Macosko *et al.*, 2015) (Fig. 5A). In the original publication for this dataset the authors found that the five tumors (labeled MGH26, MGH28, MGH29, MGH30 and MGH31) were transcriptionally distinct from each other. Moreover, they found two subsets of cells that differed from the rest: A large subpopulation of cycling cells expressing genes typical to the G2/M phase of the cell cycle, and a subset of Oligodendrocytes, mostly from the MGH31 tumor.

We find that the clusters that correspond to the five tumors have high cluster robustness scores, as well as the cycling (G2/M) cells and Oligodendrocytes (Fig. 5C). Notice that the combined cluster (MGH31+Oligodendrocytes) has a low robustness score (Fig. 5D).

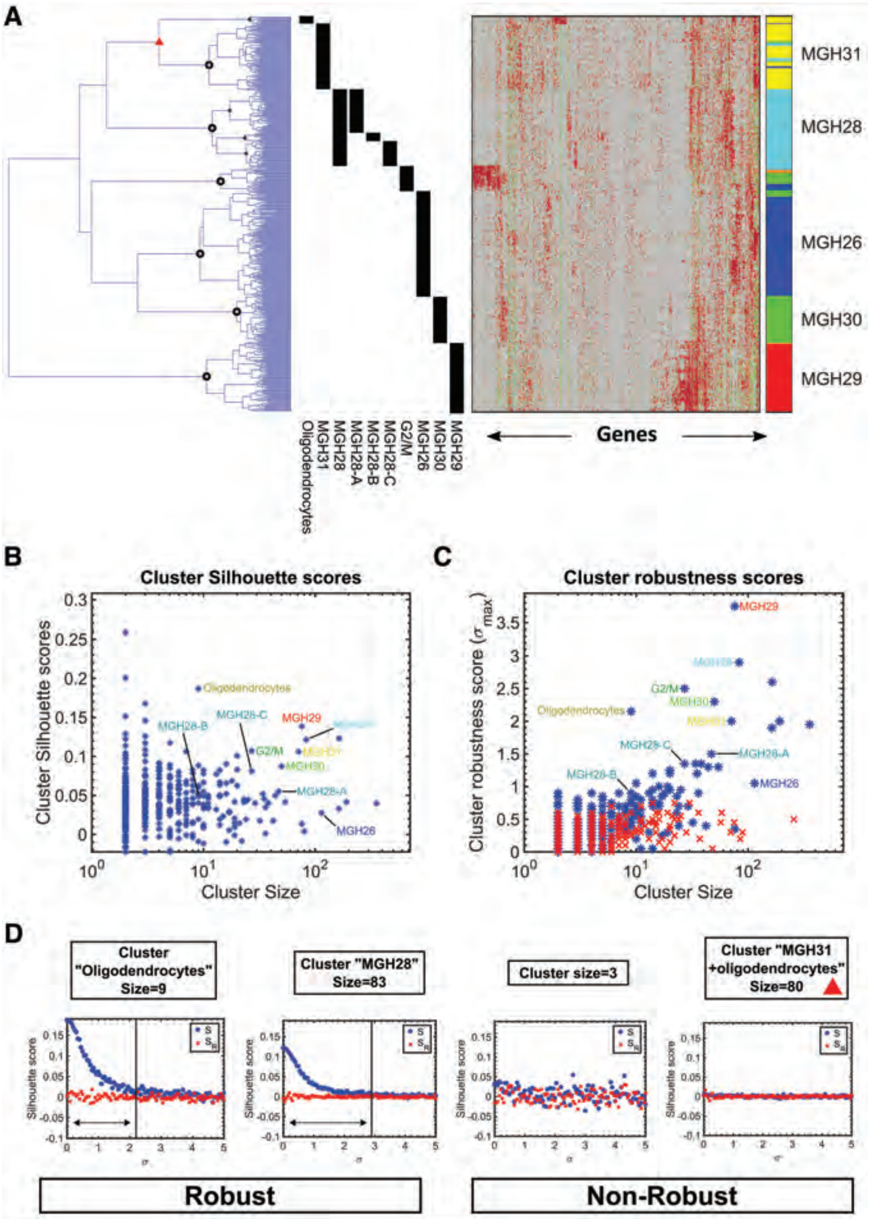

Fig. 5. Biologically meaningful cell sub-populations from a previously published single-cell RNA sequencing dataset of primary glioblastomas have high cluster robustness scores. (**A**) Hierarchical clustering of single cell gene expression measurements from 430 individual cells isolated from five human glioblastoma patients that were profiled by the Smart-seq single cell RNA sequencing protocol (Patel *et al.*, 2014). After choosing 708 highly variable genes (see Supplementary Fig. S6 and Supplementary Material) and performing a log-plus-one transformation, each gene (=column) was standardized by subtracting the mean, dividing by the standard deviation and truncating to the range [-1, 1] (Red—over-expression, green—under-expression, gray—no expression). The clustering distinguishes between the five tumors (MGH26, MGH28, MGH29, MGH30 and MGH31), and identifies two additional cell types: cycling (G2/M) cells and Oligodendrocytes. Manual inspection reveals additional putative sub-populations within tumor MGH28 (MGH28-A, B and C). (**B**) Cluster Silhouette scores are highly biased with respect to cluster size, such that small meaningless clusters have high Silhouette values. (**C**) The clusters representing the five tumors, cycling (G2/M) cells, Oligodendrocytes and the additional putative sub-populations within tumor MGH28, have high cluster robustness scores (*, blue) that are higher than expected for clusters of similar size in a corresponding 'homogenized' dataset (x, red). (**D**) Calculation of robustness scores for selected clusters. Shown is the Silhouette score S (o, blue) and the label-mixed Silhouette score S_R (x, red) as a function of noise with increasing variance σ^2. Robust clusters retain their optimality ($S > S_R$) for higher values of noise (σ). Small clusters tend to have higher Silhouette values and a larger dispersion of S_R values (for different label-mixing realizations R). Notice that the combined cluster (MGH31+Oligodendrocytes, red triangle) has a low robustness score. Clustering was done with Euclidean distance and Ward's linkage. Cluster Silhouette scores were calculated using Euclidean distance. We used p_threshold = 0.05 and 100 label-shuffling realizations R

Likewise, we identify three additional putative sub-populations within tumor MGH28 (MGH28-A, B and C).

4 Discussion

Single cell RNA sequencing datasets have unique properties that require specialized preprocessing steps. Low transcript capture efficiencies that are typical in mRNA sequencing technologies result in large fractions of 'dropout' events, i.e. false zero measurements. Furthermore, the large dimensionality of these datasets requires dimension reduction steps prior to clustering and visualization. Methods for estimating and recovering dropout events include MAGIC (van Dijk *et al.*, 2018) and DrImpute (Gong *et al.*, 2018). Another approach is used in ZIFA (Pierson and Yau, 2015), which performs dimension reduction while assuming a global dropout rate parameter, and VAMF (Townes *et al.*, 2017), which assumes cell-specific dropout rates. Linnorm (Yip *et al.*, 2017) performs normalization and transformation of scRNA-seq data, based on parameters that are calculated on a subset of genes that are homogeneously expressed across different cells. SIMLR (Wang *et al.*, 2017) is an algorithm that learns the similarity measure that best splits the data into a predefined number of clusters. This similarity measure can then be used as an input to dimension reduction, visualization and clustering algorithms. We believe that our cluster robustness scoring method can complement and dovetail nicely with the above algorithms.

Generally, there are two main strategies for clustering (Kaufman and Rousseeuw, 1990): partitioning algorithms and hierarchical algorithms. **Partitioning algorithms** try to partition N data points into k distinct groups in the best possible way. Usually, these algorithms require the user to provide an additional parameter such as the number of clusters k. Examples include k-means and k-medoids (Kaufman and Rousseeuw, 1990) [Note that other partitioning algorithms such as DBSCAN (Ester *et al.*, 1996), Mean-shift (Cheng, 1995), or Affinity propagation (Frey and Dueck, 2007) require the user to select other parameters instead]. **Hierarchical algorithms** construct clusters in the form of a hierarchical tree that splits the N data points into smaller and smaller clusters until a single point is reached. Since hierarchical clustering imposes a tree structure on the data, it is sometimes more appropriate for applications in biology, for example to describe the evolutionary relations between organisms originating from a common ancestor or the hierarchy of differentiating cell types and sub-types originating from a common stem cell. Furthermore, hierarchical clustering algorithms require less external parameters to be optimized by the user, for example, they go over all possibilities for the number of clusters ($k = 1,\ldots, N$) in a single run. However, it remains for the user to correctly interpret the results by identifying the clusters that represent biologically meaningful cell types from all the clusters produced by the algorithm (this is sometimes referred to as the problem of finding the 'termination condition').

In this study we addressed this problem by adding Gaussian noise and using the Silhouette measure to calculate a 'robustness score' for each cluster. The Silhouette measure (Rousseeuw, 1987) was originally developed as a scoring system to evaluate the quality of clusters resulting from partitioning algorithms and to assist the user in selecting the optimal number of clusters k. However, the Silhouette measure S is biased in the sense that small clusters can have large Silhouette measures even when they are non-meaningful (See Figs 3–5B, and simulations in Supplementary Material), which makes it somewhat difficult to identify the most biologically meaningful clusters. Therefore, we took a different approach: We added noise with increasing variance σ^2 and measured the level of noise that can be added such that clusters remain well separated, in the sense that their Silhouette score S remains higher than the label-shuffled Silhouette score S_R. The maximal variance is a measure of the cluster robustness to noise. We found that this measure is considerably less biased with respect to cluster size (Figs 3–5C and simulations in Supplementary Material).

Our underlying assumption is that biologically meaningful cell subpopulations are represented by clusters that have a high robustness to additive noise. On the other hand, clusters that have no real meaning, i.e. those that were formed due to over-fitting, have low robustness. We also assume that in 'homogenized' single cell expression data—when the expression values of each gene are randomly permuted such the mutual relations between the different genes vanish—there are no biologically meaningful cell types. Thus, the cluster robustness score provides us with a semi-automated way to discern between meaningful and non-meaningful clusters that can complement manual inspection of the data and assist in identification and characterization of cell types in tissues and tumors.

One limitation of our method is that it will work only when the Silhouette measure is appropriate for measuring the quality of a cluster, as in the case of clusters that are roughly ball-shaped. Another limitation is the run time. The algorithm can run for ~30 min on a standard PC for ~100 cells and 100 label-randomizing iterations, which can make scaling up to many thousands of cells infeasible. This can be mitigated by performing less label-randomizing iterations or by optimizing the range and number of noise intensities (σ) to be tested. Another possibility is to perform robustness analysis on selected clusters of interest rather than on all the clusters as we did here.

Acknowledgements

We wish to thank Uri Alon, Boaz Nadler, Gur Yaari and all members of our lab for useful comments and suggestions.

Funding

T.K. and I.K. are supported by the Israel Science Foundation ([ICORE number 1902/12] and [Grants numbers 1634/13 and 2017/13]), the Israel Cancer Association [Grant no. 20150911], the Israel Ministry of Health [Grant number 3-10146] and the EU-FP7 [Marie Curie International Reintegration Grant no. 618592]. P.D. is supported by a Runyon-Rachleff Innovator Award [Grant number DRR-44-16; Island Outreach Foundation] from the Damon Runyon Cancer Research Foundation, by the Schaefer Research Scholars Program (2017) of Columbia University's College of Physicians and Surgeons (Dr. Ludwig Schaefer Fund), and by a research grant (CU16-1797) from the Adenoid Cystic Carcinoma Research Foundation (ACCRF). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: P.D. and T.K. are listed as a co-inventors on patents owned by Stanford University and licensed to Quanticel Pharmaceuticals Inc. (now a fully owned subsidiary of Celgene), describing methods for the use of single-cell genomics technologies (e.g. single-cell qPCR, single-cell RNA-seq) in the analysis of human tissues (US-9329170, US-9850483).

References

Al-Hajj,M. *et al.* (2003) Prospective identification of tumorigenic breast cancer cells. *Proc. Natl. Acad. Sci. USA*, **100**, 3983–3988.

Barker,N. *et al.* (2007) Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature*, **449**, 1003–1007.

Björklund,A.K. *et al.* (2016) The heterogeneity of human CD127+ innate lymphoid cells revealed by single-cell RNA sequencing. *Nat. Immunol.*, **17**, 451–460.

Bonnet,D. and Dick,J.E. (1997) Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.*, **3**, 730–737.

Bussolati,B. *et al.* (2008) Identification of a tumor-initiating stem cell population in human renal carcinomas. *Faseb J.*, **22**, 3696–3705.

Chen,S. *et al.* (2015) Intrinsic age-dependent changes and cell-cell contacts regulate nephron progenitor lifespan. *Dev. Cell*, **35**, 49–62.

Cheng,Y. (1995) Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, **17**, 790–799.

Dalerba,P. *et al.* (2007) Phenotypic characterization of human colorectal cancer stem cells. *Proc. Natl. Acad. Sci. USA*, **104**, 10158–10163.

Dalerba,P. *et al.* (2011) Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.*, **29**, 1120–1127.

van Dijk,D. *et al.* (2018) Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, **174**, 716–729.e27.

Ester,M. *et al.* (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. 2nd Int. Conf. Knowl. Discov. Data Min., pp. 226–231.

Fan,H.C. *et al.* (2015) Combinatorial labeling of single cells for gene expression cytometry. *Science*, **347**, 1258367.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Gong,W. *et al.* (2018) DrImpute: Imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*, doi:10.1186/s12859-018-2226-y.

Guo,G. *et al.* (2013) Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell Stem Cell*, **13**, 492–505.

Guo,G. *et al.* (2010) Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell*, **18**, 675–685.

Gury-BenAri,M. *et al.* (2016) The spectrum and regulatory landscape of intestinal innate lymphoid cells are shaped by the microbiome. *Cell*, **166**, 1231–1246.e13.

Hashimshony,T. *et al.* (2016) CEL-Seq2: sensitive highly-multiplexed single--cell RNA-Seq. *Genome Biol.*, **17**, 77.

Jaitin,D.A. *et al.* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, **343**, 776–779.

Kaufman,L. and Rousseeuw,P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics)*, 2005th edn. A John Wiley & Sons, New Jersey.

Klein,A.M. *et al.* (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.

La Manno,G. *et al.* (2016) Molecular diversity of midbrain development in mouse, human and stem cells. *Cell*, **167**, 566–580.

Li,C. *et al.* (2007) Identification of pancreatic cancer stem cells. *Cancer Res.*, **67**, 1030–1037.

Lyubimova,A. *et al.* (2013) Single-molecule mRNA detection and counting in mammalian tissue. *Nat. Protoc.*, **8**, 1743–1758.

Macosko,E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

Montgomery,R.K. *et al.* (2011) Mouse telomerase reverse transcriptase (mTert) expression marks slowly cycling intestinal stem cells. *Proc. Natl. Acad. Sci. USA*, **108**, 179–184.

Patel,A.P. *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.

Picelli,S. *et al.* (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*, **9**, 171–181.

Pierson,E. and Yau,C. (2015) ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, **16**, 1–10.

Pode-Shakked,N. *et al.* (2009) Developmental tumourigenesis: nCAM as a putative marker for the malignant renal stem/progenitor cell population. *J. Cell. Mol. Med.*, **13**, 1792–1808.

Pode-Shakked,N. *et al.* (2013) The isolation and characterization of renal cancer initiating cells from human Wilms' tumour xenografts unveils new therapeutic targets. *EMBO Mol. Med.*, **5**, 18–37.

Prince,M.E. *et al.* (2007) Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma. *Proc. Natl. Acad. Sci. USA*, **104**, 973–978.

Rothenberg,M.E. *et al.* (2012) Identification of a cKit(+) colonic crypt base secretory cell that supports Lgr5(+) stem cells in mice. *Gastroenterology*, **142**, 1195–1205.e6.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Sanchez-Freire,V. *et al.* (2012) Microfluidic single-cell real-time PCR for comparative analysis of gene expression patterns. *Nat. Protoc.*, **7**, 829–838.

Sangiorgi,E. and Capecchi,M.R. (2008) Bmi1 is expressed in vivo in intestinal stem cells. *Nat. Genet.*, **40**, 915–920.

Shackleton,M. *et al.* (2006) Generation of a functional mammary gland from a single stem cell. *Nature*, **439**, 84–88.

Sheng,K. *et al.* (2017) Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods*, **14**, 267–270.

Singh,S.K. *et al.* (2004) Identification of human brain tumour initiating cells. *Nature*, **432**, 396–401.

Spangrude,G.J. *et al.* (1988) Purification and characterization of mouse hematopoietic stem cells. *Science*, **241**, 58–62.

Stingl,J. *et al.* (2006) Purification and unique properties of mammary epithelial stem cells. *Nature*, **439**, 993–997.

Swiers,G. *et al.* (2013) Early dynamic fate changes in haemogenic endothelium characterized at the single-cell level. *Nat. Commun.*, **4**, 2924.

Townes,F.W. *et al.* (2017) Varying-censoring aware matrix factorization for single cell RNA-sequencing. *bioRxiv*, **166736**.

Treutlein,B. *et al.* (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, **509**, 371–375.

Villani,A.-C. *et al.* (2017) Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, **356**, eaah4573.

Wang,B. *et al.* (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods*, **14**, 414–416.

Yip,S.H. *et al.* (2017) Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.*, **45**, e179.

Zeisel,A. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.