Published online 18 January 2018

Nucleic Acids Research, 2018, Vol. 46, No. 6 ***e36***
doi: 10.1093/nar/gky007

dropClust: efficient clustering of ultra-large scRNA-seq data

Debajyoti Sinha[1,2], Akhilesh Kumar[3], Himanshu Kumar[3], Sanghamitra Bandyopadhyay[1,*] and Debarka Sengupta[4,*]

[1]Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, West Bengal, India, [2]Department of Computer Science and Engineering, University of Calcutta, Kolkata 700098, West Bengal, India, [3]Laboratory of Immunology and Infectious Disease Biology, Department of Biological Sciences, Indian Institute of Science Education and Research, Bhopal 462066, Madhya Pradesh, India and [4]Center for Computational Biology and Department of Computer Science and Engineering, Indraprastha Institute of Information Technology, Delhi 110020, India

Received October 21, 2017; Revised December 22, 2017; Editorial Decision January 03, 2018; Accepted January 07, 2018

ABSTRACT

Droplet based single cell transcriptomics has recently enabled parallel screening of tens of thousands of single cells. Clustering methods that scale for such high dimensional data without compromising accuracy are scarce. We exploit Locality Sensitive Hashing, an approximate nearest neighbour search technique to develop a *de novo* clustering algorithm for large-scale single cell data. On a number of real datasets, dropClust outperformed the existing best practice methods in terms of execution time, clustering accuracy and detectability of minor cell sub-types.

INTRODUCTION

Biological systems harbor substantial heterogeneity, which is hard to decode by profiling population of cells. Over the past few years, technological advances enabled genome wide profiling of RNA, DNA, protein and epigenetic modifications in individual cells (1). Amongst the most recent developments, in-drop (within a droplet) barcoding has gained a lot of attention as it enables 3′ mRNA counting of thousands of individual cells in a matter of several minutes to few hours. With the growing popularity of the assay and availability of affordable commercial platforms, a sharp increase is expected in average sample size of future investigations. A recent work produced an unprecedented ∼250k single cell expression profiles as part of a single study (2). This, gives us an idea about the scale of the future single cell experiments. Since the introduction of single cell RNA sequencing (scRNA-seq) technologies, a number of clustering techniques have been devised while accounting for the unique characteristics of the new data type (3–6). However, a majority of these techniques struggle to scale when studies feature several tens of thousands of transcriptomes. In fact, methods developed solely for such ultra large datasets (henceforth referred to as droplet-seq' data) are either computationally expensive (7) or over-simplistic (2).

Network based clustering techniques have been used effectively for clustering sc-RNA-seq data (8,9). An exhaustive nearest neighbour search requires quadratic-time tabulation of pair-wise distances. For large sample sizes, this approach turns out to be significantly slow. Seurat, one of the early-proposed methods for droplet-seq data analysis, performs sub-sampling of transcriptomes prior to nearest-neighbour based network construction. Random sampling can be irreversibly lossy when one of the objectives is to identify rare cell populations. In a recent work, Zheng and colleagues (2) used *k*-means as the method for clustering droplet-seq data. While *k*-means is reasonably fast, it suffers from two major drawbacks: (i) User needs to specify the number of clusters. (ii) The method struggles to identify clusters of non-spherical shapes.

To address the above shortcomings, we developed dropClust, a scalable yet accurate clustering algorithm for droplet-seq data. dropClust uses Locality Sensitive Hashing (LSH) to find nearest neighbours of individual transcriptomes. This neighbourhood information is used to perform *Structure Preserving Sampling* (SPS) of the expression profiles, which retains relatively higher number of representative transcriptomes from smaller sub-populations. The sampling technique used in dropClust helps in accelerating unsupervised cell grouping without compromising accuracy.

We evaluated the efficacy of dropClust first on a large cohort of peripheral blood mononuclear cells (PBMCs), annotated based on similarity with purified, major immune cell sub-types (2). Besides the common cell types, a number of minor immune cell sub-populations were identified by dropClust. In fact, clusters yielded by dropClust were found to be maximally concordant (14% improvement in Adjusted

*To whom correspondence should be addressed. Tel: +91 11 26907446; Email: debarka@iiitd.ac.in
Correspondence may also be addressed to Sanghamitra Bandyopadhyay. Tel: +91 33 2575 3104; Email: sanghami@isical.ac.in

Rand Index or ARI with respect to existing best practice methods) with the available cell type annotations. Its performance was consistent on two more droplet-seq datasets curated from independent studies. We also performed a simulation study leveraging a published droplet-seq data containing expression profiles of Jurkat and 293T cells mixed *in vitro* at equal proportions. Amongst all tested clustering methods, dropClust was found most tolerant to bioinformatic dilution of any of the two cell types, thus providing evidence for its sensitivity to minor cell sub-populations.

MATERIALS AND METHODS

Description of the datasets

We used two datasets from a recent work by Zheng *et al.* (2). The first single-cell-RNA-seq (scRNA-seq) data consists of ∼68 000 PBMCs, collected from a healthy donor. Single cell expression profiles of 11 purified subpopulations of PBMCs are used as reference for cell type annotation. This dataset served as a gold standard for performance assessment of the clustering techniques. The second dataset from the same study contains expression profiles of Jurkat and 293T cells, mixed *in vitro* at equal proportions (50:50). All ∼3200 cells of this data are assigned their respective lineages through SNV analysis (2). Expression matrices for both these datasets were downloaded from www.10xgenomics.com.

Two additional datasets were used to benchmark the performance of the clustering algorithms. The datasets contain expression profiles of ∼49k mouse retina cells (7) and ∼2700 mouse embryonic stem (ES) cells respectively (10).

To evaluate the congruence between dropClust and Seurat, we used a doplet-seq data containing ∼20K transcriptomes sampled from the arcuate-median eminence complex (Arc-ME) region of mouse brain (11).

Data preprocessing, normalization and gene selection

Expression matrices for all the datasets were downloaded from publicly available repositories. For each dataset, the genes whose UMI counts were >3 in at least three cells were retained. For PBMC data, only ∼7000 genes qualified this criterion. The filtered data matrix was then subjected to UMI normalization that involves dividing UMI counts by the total UMI counts in each cell and multiplying the scaled counts by the median of the total UMI counts across cells (2). One thousand most variable genes were selected based on their relative dispersion (variance/mean) with respect to the expected dispersion across genes with similar average expression (2,7). Normalized expression matrix with the selected genes thus obtained was $\log_2$ transformed after addition of 1 as a pseudo count.

dropClust overview

dropClust employs Locality Sensitive Hashing (LSH), a logarithmic-time algorithm to determine approximate neighbourhood for individual transcriptomes. An approximate k nearest neighbour network of individual transcriptomes thus obtained, is subjected to Louvian (12), a widely used network partitioning algorithm. While Louvian based topological clustering delineates majority of the prevalent cell types, finer subpopulations of seemingly similar cells within large clusters are often not separated at a satisfactory precision (data not shown). Clusters found using Louvian are therefore used as points of reference for further down-sampling of the transcriptomes. dropClust uses an exponential decay function to select higher number of expression profiles from clusters of relatively smaller sizes. Simulated annealing is used to perform hyperparameter search with the aim of restricting the sample size close to a number, manageable by hierarchical clustering. The proposed sampling strategy preserves the rare cell clusters even when the sample sizes are fairly small compared to the population size. The complete dropClust work-flow is illustrated in Supplementary Figure S1.

Structure preserving sampling of transcriptomes

It is hard to avoid subsampling while managing high dimensional genomic data. However, random sub-sampling might result in loss of rare sub-populations. The proposed dropClust pipeline introduces a novel data sampling approach that preserves distinct structural properties of the data. This is achieved in two steps: (a) a fairly large (usually minimum of 20 000 and a third of the whole population) number of scRNA-seq profiles are randomly selected from the complete set of transcriptomes and then subjected to a fast, approximate graph based clustering algorithm; (b) the topological clusters thus obtained are used to guide further subsampling of the transcriptomes in a way that retains relatively higher number of cells from smaller clusters, which were otherwise ignored in case of random sub-sampling (Figure 1).

To construct the network, top-k approximate nearest neighbours ($k = 10$ by default) are identified rapidly by employing Locality Sensitive Hashing (LSH) (13). A faster and more accurate implementation of the original LSH, called LSHForest is used for this purpose (14). The nearest neighbour network (NNN) of transcriptomes thus created is subjected to Louvain (12), a widely used method for detecting community structures in networks. Notably, Seurat (7) uses Shared Nearest Neighbours (SNN) for network construction at one specific stage. While construction of NNN using LSH takes $O(n\log n)$ time (14), building SNN requires $O(n^2)$ time (8), where n denotes the number of single cell expression profiles. The choice of LSH to search nearest neighbours leads to a dramatic reduction in computation time. Since LSH is an approximate method for nearest neighbour search, the clusters obtained need further refinement. Moreover, Louvain offers limited control for determining cluster resolution.

Sampling from primary clusters. To ensure selection of sufficient representative transcriptomes from small clusters, an exponential decay function (15) is used to determine the proportion of transciptomes to be sampled from each cluster. For ith cluster, the proportion of expression profiles p_i was obtained as follows.

$$p_i = p_l - e^{-\frac{S_i}{K}}(p_l - p_u) \quad (1)$$

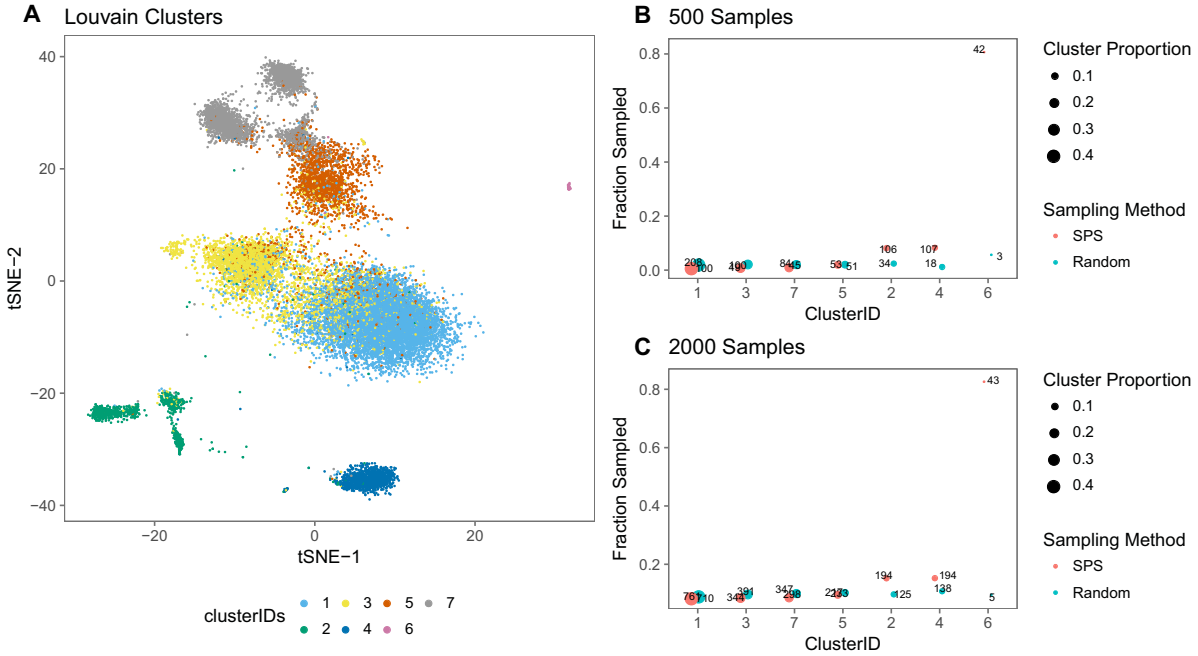

Figure 1. (**A**) 2D embedding of 20K PBMC transcriptomes, chosen randomly from the complete dataset. Separate colours are used for the Louvain-predicted clusters. (**B**) For each cluster, the number of sampled cells is shown using both SPS and random sampling. Size of the Louvain clusters are indicated by the size of the bubbles. X-axis shows cluster ID, whereas Y-axis shows the sampling fraction. True number of cells are also indicated on each of the bubbles. In this case 500 transcriptomes are sampled through SPS. (**C**) Similar figure with 2000 as sample size.

where S_i is the size of cluster i, K is a scaling factor, p_i is the proportion of cells to be sampled from the ith Louvain cluster. p_l and p_u are lower and upper bounds of the proportion value respectively. Based on the above equation we may show the following:

$$\lim_{S_i \to \infty} p_i = p_l \tag{2}$$

$$\lim_{S_i \to 0} p_i = p_u \tag{3}$$

Since Equation (1) does not explicitly impose any upper bound on the final sample size, one may be left with an arbitrarily high or low number of single cell transcriptomes for final clustering. To address this, dropClust allows user specify his preferred sample size and employs simulated annealing (SA) (16) to come up with the right values for p_l, p_u and K. This operation may formally be described as follows:

$$\langle \overset{*}{p_l}, \overset{*}{p_u}, \overset{*}{K} \rangle = \underset{p_l, p_u, K}{\arg\min} \left| \tau - \sum_{\forall i} p_i S_i \right| \tag{4}$$

where τ denotes the user specified sample size. We used simulated annealing implementation from the `GenSA` R package (16).

Choosing the best principal components

For the cells obtained through structure preserving sampling (SPS), gene selection is performed based on Principal Component Analysis (PCA). It is well known that clustering outcome is often improved by careful selection of genes. PCA has widely been used for this purpose (7,17). Traditionally, genes with high loadings on the top few principal components (PCs) are considered to be most informative. This method, in some sense, guarantees selection of the most variable genes. However, expression variability may not necessarily explain cell type heterogeneity. For gene selection based on high PC loadings, dropClust uses PCs that not only explain a sizeable proportion of the observed expression variance but also manifest a large proportion of phenotypic diversity.

To this end dropClust uses mixtures of Gaussians to detect PCs with multi-modal distribution of the projected transcriptomes. For each of the top 50 PCs, we estimate the explained heterogeneity by inspecting the multi-modal nature of its marginal distribution. Gaussian Mixture Model (GMM) (18), supplemented with Bayesian Information Criterion (BIC) (19) is used to determine the number of modes corresponding to each PC. Each of these modes is expected to represent a cell type. R package `mclust` is used for this purpose. PCs, modelled by three or more Gaussians are used for PC-loading based gene selection (17). When applied on the real datasets we commonly encountered cases where a top PC featured a small number of modes whereas a trailing PC featured higher levels of modality (Figure 2, Supplementary Figures S2 and S3). Top 200 high-loading genes are retained for the subsequent clustering step.

Clustering of sampled cells

Average-linkage hierarchical clustering is performed to group the sampled cells based on expression of the 200 selected genes. Euclidean distance is used as the measure of dissimilarity. To cut the dendogram `cutreeDynamic()` is used from the `dynamicTreeCut` R package (20,21).

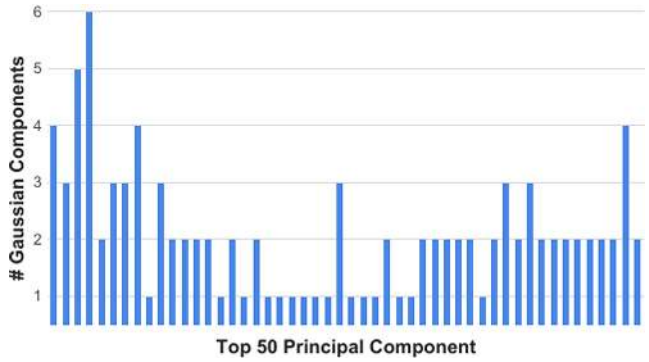

Figure 2. Barplot depicting the number of estimated Gaussian components for each of the top 50 principal components derived from the PBMC data.

Post-hoc cluster assignment for left out transcriptomes. Cells that are not subjected to hierarchical clustering are assigned their respective clusters of origin using a simple post-hoc cluster assignment strategy. To achieve this, locality preserving hash codes are generated for the clustered transcriptomes, using `LSH-Forest`. For each of the left out transcriptomes k ($k = 5$, by default) approximate nearest neighbours are then found through LSH queries. Each unallocated transcriptome is assigned the cluster of origin for which the most number of representatives are found in its corresponding set of k nearest neighbours. Ties for cluster assignment are broken at random.

The rationale behind parameter selection for various methods can be found in the Supplementary Data (Section 1). We probed into the possiblity of producing spurious clusters due to technical bias such as library-depth. However, we didnt find any clear evidence of such events (Supplementary Figure S8).

2D embedding of transcriptomes for visualization

The 2D embedding of samples is carried out in two steps. In the first step t-SNE is applied to transcriptomes obtained through SPS. Top 200 PCA-selected genes are used for this purpose. In the next step, remaining transcriptomes are allocated positions in the pre-existing 2D map of the sampled cells. To perform this, we borrow the sets of k nearest neighbours, found at the time of post-hoc cluster assignment. Coordinates for each newly added point are derived by averaging t-SNE coordinate values of neighbours that belonged to its cluster of origin.

Differential expression of genes

To speed up the differential expression (DE) analyses, we consider 100 randomly chosen transcriptomes from each cluster. Only genes with count >3 in at least 0.5% of these cells are retained for the analysis. Fast nonparametric, DE analysis tool NODES is used to to make DE gene calls with 0.05 as the cut off value for false discovery rate (FDR) and a fold change of 1.2 (BioRxiv: https://doi.org/10.1101/049734). Among the DE genes, ones that are significantly upregulated in a specific cluster with respect to each of remaining clusters are named cell type specific genes.

Simulation of rare cell population

The dataset containing Jurkat and 293T cells at equal ratio was used for performing simulations to assess detectability of minor cell populations. Cell type identity of each transcriptome of this dataset was determined by SNV analysis (2). To introduce rareness, we forcibly reduced the frequency of one of the cell types. To prevent bias, we performed these experiments by treating both cell types as rare in separate simulations. In simulated datasets, the proportion of rare cell transcriptomes was varied among 1%, 2.5%, 5% and 10%. For each of these specified concentrations, 10 datasets were created by independent sub-sampling of transcriptomes of a specific type. The transcriptomes of the major cell type were not subjected to any kind of sampling. Since this procedure was repeated for both the cell types, for each concentration a total of 20 datasets were produced.

We used F_1-score as a measure for detectability of rare cell clusters. The score is defined as follows.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (5)$$

To compute the above score we first associated the predicted cluster that contained the majority of the rare cells to the rare cell group. Following this, *recall* was defined as the ratio between the number of true rare cells within the predicted rare cell cluster and the total number of known rare cells. On the other hand, *precision* was defined as the ratio between the number of known rare cells within the predicted rare cell group and the total number of cells in the predicted rare cell group.

RESULTS AND DISCUSSION

Analysis of ∼68K human PBMC data

We applied dropClust first on a collection of ∼68K human peripheral blood mononuclear cells (PBMC), annotated based on similarity with matched single cell transcriptomes of 11 purified immune cell subpopulations, purified using fluorescence-activated cell sorting (FACS) (2). We used this information to benchmark the performance of the cell clustering methods under investigation. For each method, concordance between cluster assignment and cell type annotation was measured by Adjusted Rand Index (ARI). Among all methods, dropClust maximized the ARI (Figures 3 and 4).

Identification of cell-type specific Marker. Differential expression (DE) analysis was carried out between each pair of clusters to identify the cell type specific genes for each sub-population. (Table 1; Supplementary Figure S9). Details about the mapping of the dropClust predicted clusters to their respective potential cell types can be found in the Supplementary Data.

We could associate the predicted groups of transcriptomes with known cell-types based on marker gene expression. Such associations were not always unambiguous. There are two principal reasons for such ambiguities: (i) Surface protein concentration is not always linearly related to the expression of the corresponding gene. Well known

Table 1. Markers used for cell type inference from the PBMC data

Cluster ID	Potential cell type	Markers
1	Naive T cells	CD27 (22), CCR7 (23), CD8A, CD8B[a]
2	CD4+ memory cells	IL7R (24), CD27 (22), CCR7 (23)
3	NKT cells	ZNF683 (25,26) (UniProtKB - Q8IZ20), CD8A, CD8B[a]
4	B cells	CD79A (27), CD37 (28)
5 & 7	CD8+ T cells	GZMK (29), CD8A, CD8B[a]
6	NK cells	CD160 (30,31), NKG7 (32), GNLY (33), CD247 (34), CCL3 (35), GZMB (36)
8 & 9	CD16+ and CD14+ monocytes	CD68 (37), CD16 (FCGR3A) (38), CD14 (38), S100A12 (39,40)
10	Regulatory T cells	CCR10, CD25(IL2RA)[a], CD52 (41), CMTM7, FOXP3 (42)[a]
11	Monocyte derived dendritic cells	CST3 (43), CD1C (44,45), FCER1A (43)
12	Megakaryocyte progenitors	PF4 (46),PPBP (47), PLA2G12A (48)
13	Progenitor-NK cells	ID2 (49,50)
14	Plasmacytoid dendritic cells	GZMB (51), CD123 (IL3RA) (52)

[a]Markers that are well-known but failed to qualify the gene selection criteria.

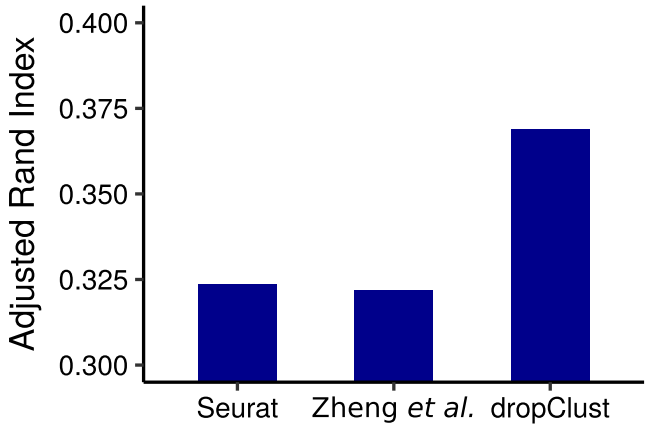

Figure 3. Bars show the ARI indexes obtained by comparing clustering outcomes with cell-type annotations.

surface markers are commonly found having low expression. (ii) High drop-out rates and lack of sequencing depth cause prevalence of zeros as expression estimate. As a result, cell type specific yet low expressed genes are often not detected in single cell assays. Under these constraints, we tried to gather as much evidence as possible to assign a putative cell type to each of the detected PBMC clusters.

We identified all major lymphoid and myeloid subpopulations including a number of minor subtypes. Among the populous cell-subtypes we detected naive, memory and cytotoxic T cells, B cells, natural killers (NK), natural kill T cells (NKT cells), CD14+ and CD16+ blood monocytes and monocyte derived dendritic cells. Besides these we also found a number of minor cell types including plasmocytoid dendritic cells, regulatory T cells (Tregs), progenitor NK cells and circulating megakaryocyte progenitors (Figure 5). Supplementary Figure S11 shows the heatmap of the cell type specific differentially up-regulated genes.Table 1 lists the cell-type markers with respective predicted clusters.

Most notable among the above findings are two crisp subpopulations of Natural Killer progenitors (0.1% of the population) and Regulatory T cells (0.5%) (Figure 5, Table 1 in the main text; Section 6 in Supplementary Data) that other methods failed to resolve (dropClust projection pictures Supplementary Figure S14). To the best of our knowledge, none of the previously published studies reported transcriptomic characterization of these cell types at single cell resolution.

Computation time

Besides improved clustering accuracy, dropClust is designed to provide significant speed up. On ∼68K PBMC data it took ∼8 min to perform preprocessing, clustering and visualization. The *k*-means based pipeline proposed by Zheng *et al.* took around 22 min whereas it took ∼100 min for Seurat for the same. Figure 6 shows the execution time taken by different methods while increasing the number of transcriptomes to analyze.

Of note, execution time of clustering algorithms vary proportionally with algorithm complexity. Seurat requires $\mathcal{O}(N^2)$ time for computing similarity matrix for N input transcriptomes. dropClust, on the other hand, uses $\mathcal{O}(N\log N)$ algorithm for Structure Preserving Sampling (SPS), followed by hierarchical clustering (usually $\mathcal{O}(M^2\log M)$) of M transcriptomes, selected through SPS. In practice, M is much smaller compared to N. Zheng and colleagues (2) applied *k*-means clustering on the 68K PBMC data using *a priori* knowledge about the number of possible clusters ($k = 10$). As discussed before *k*-means has significant drawbacks. In fact, finding the optimal k requires $\mathcal{O}(N^2)$ time (53). Moreover, *k*-means operates on an Expectation Maximization (EM) like algorithm that often requires many iterations to converge. Default setting of the algorithm employed by Zheng and colleagues forces the algorithm to terminate after the 150th iteration. These clearly highlight the theoretical basis for the speed achieved by dropClust. For component-wise time comparison (Section 4, Supplementary Data).

We were compelled to stick to three methods only as some of the other methods including GiniClust (6), BackSpin (5) and RaceID (54) failed to execute on the complete PBMC data (Section 2 in Supplementary Data). Mem-

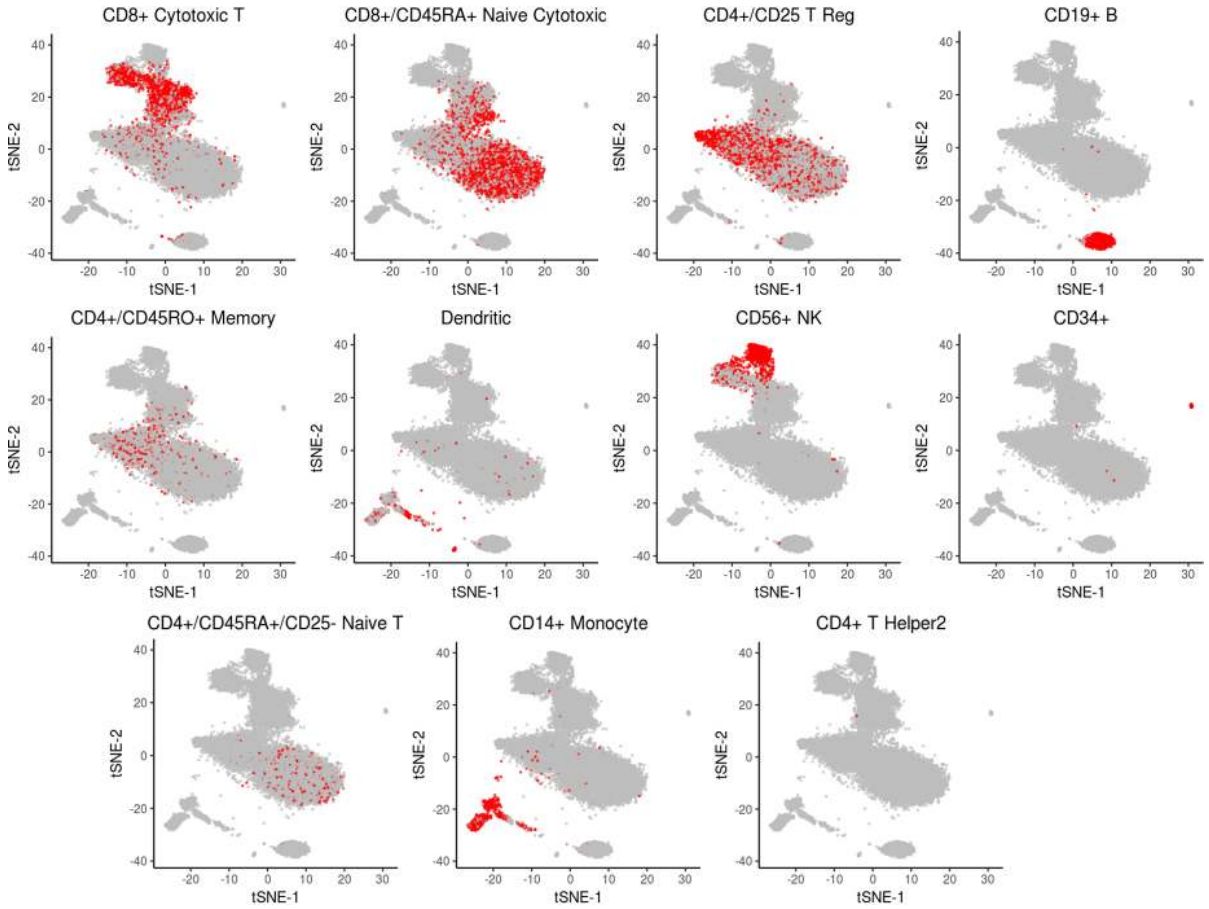

Figure 4. Localization of PBMC transcriptomes of same type (based on annotation) on the 2D embedding produced by dropClust. Each sub-figure corresponds to one of the well known immune cell types considered for benchmarking clustering accuracy by Zheng *et al.* (2).

ory consumption of dropClust for increasing sample size is shown in Supplementary Figure S4.

All time measurements were taken on a system with the following configuration: Intel(R) Xeon(R) CPU E5-2670 v2 @ 2.50GHz with 20 cores and 100GB of RAM.

Detectability of minor cell sub-populations: a simulation study

A major promise of single cell expression profiling at a large scale lies in the possibility of identifying rare cell subpopulations. A cell type may be considered as rare when its abundance in the respective population is $\leq$5% (6,54). The ability of the clustering methods to detect rare cell-types was assessed through a simulation study. For this, we used a collection of ∼3200 scRNA-seq profiles containing Jurkat and 293T cells, mixed *in vitro* at equal proportion (2). The authors tracked the profile of Single Nucleotide Variants (SNVs) to determine the lineage of the individual cells. The ratio of the two cell types was altered *in silico* by downsampling one of the populations. Abundance of the minor cell type was varied between 1% and 10%. A variant of the popular F_1 score was used to measure the algorithm efficacies. dropClust turned out to be the only algorithm that detected the minor clusters nearly accurately at all tested concentrations (Figure 7). The existing methods clearly struggled with the smaller concentrations of the rare cell lineage.

We also assessed rare cell detectability of dropClust by randomly downsampling the Megakaryocyte Progenitor cluster (∼160 cells) in the ∼68K PBMC data. We could identify the cluster correctly even when the cell type constituted 0.04% of the entire population (Supplementary Table S6).

Results on additional datasets

To rule out the possibility of assay sensitivity, we benchmarked the performance of the clustering methods on two additional droplet-seq datasets from independent studies. The first dataset consists of transcriptomes of 49 300 mouse retina cells (`GSE63473`) (7) and the second dataset contains expression profiles of ∼2700 mouse embryonic stem cells (ESC) (`GSE65525`) (10). Both the studies are exploratory in nature and therefore lack any secondary source of information for lineage determination. For these datasets, we, therefore, computed the *Silhouette* scores (a popular unsupervised metric of cluster quality) corresponding to the cell groupings obtained using different clustering methods. *Silhouette* is a non parametric measure of the trade off between cluster tightness and inter-cluster separation (55). For large sample sizes, it takes a long time to compute *Silhouette* score. To this end, we created 100 independent sets of 500

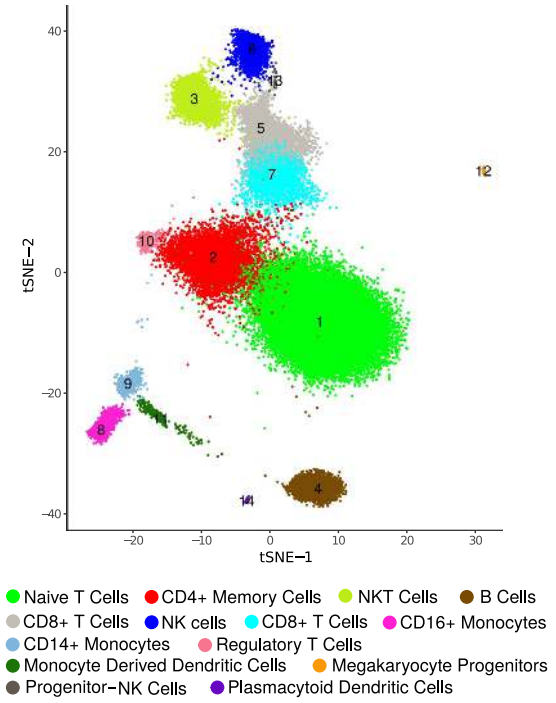

Figure 5. Clustering of ∼68K PBMC data. dropClust based visualization (a modified version of tSNE) of the transcriptomes. Fourteen clusters, retrieved by the algorithm are marked with their respective cluster IDs. Legends show the names of the inferred cell types.

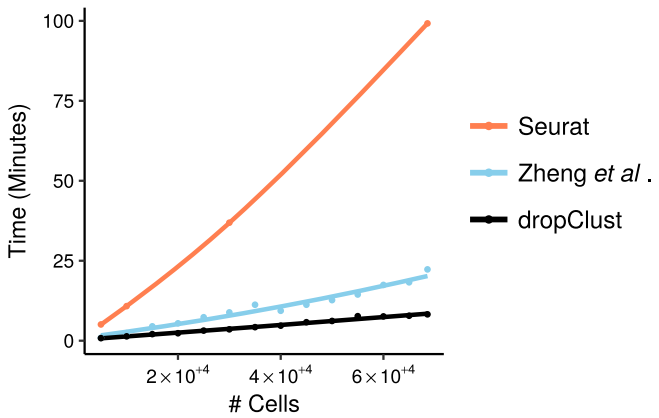

Figure 6. Trend of increase in analysis (preprocessing, clustering and vizualization)) time for different pipelines with growing number of transcriptomes under analysis.

transcriptomes through bootstrapping. Average *Silhouette* scores thus obtained are depicted through the boxplots in Figure 8.

We used an additional droplet-seq dataset featuring ∼20*K* transcriptomes derived from mouse hypothalamus (11), for a qualitative assessment of dropClust's performance with respect to the widely used method Seurat. Despite substantial concordance across the clusters, dropClust was found to be more sensible compared to Seurat in placing the like-cell types geographically closer in their respective 2D maps (Section 8, Supplementary Data for details.)

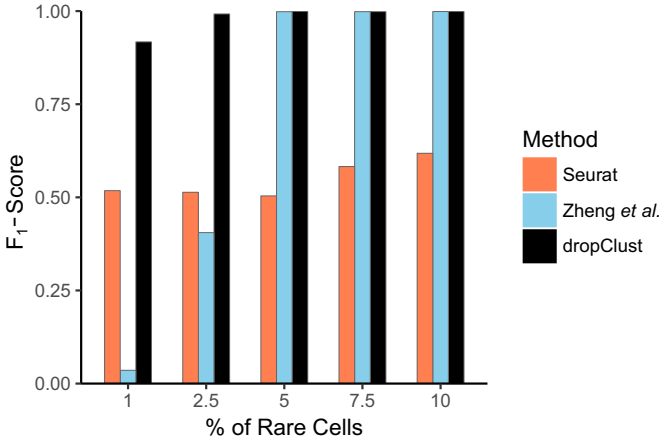

Figure 7. Detectability of minor cell types. Bars showing average of F_1-scores, obtained on 10 simulated datasets at each concentration of the minor population. A dataset containing mixture of Jurkat and 293T cells was used for this study.

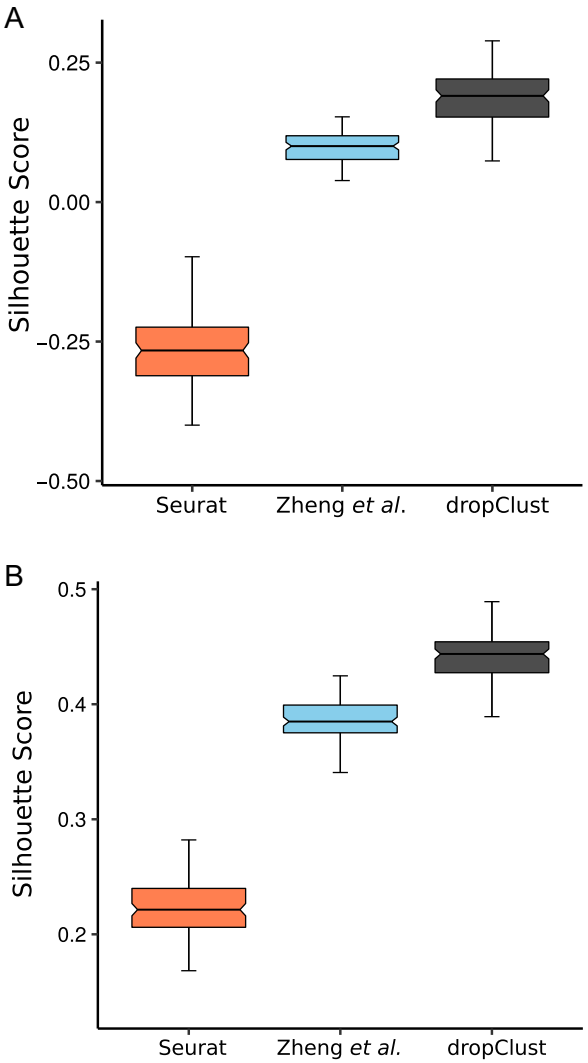

Figure 8. (**A**) Boxplots depicting average *Silhouette* scores computed on 100 bootstrap samples from the mouse retina cell data (7). A separate boxplot is used for each concerned clustering method. (**B**) Similar plots for the mouse ESC dataset (10).

Visualization

Visualizing large-scale scRNA-seq data is challenging. Both Principal Component Analysis (PCA) and t-distributed Stochastic Neighbourhood Embedding (tSNE) are widely used for visualization of scRNA-seq datasets (56). dropClust uses tSNE to obtain the 2D coordinates of a small sub-sample of the data, followed by inferring coordinate pairs of each remaining transcriptome by averaging the coordinates of its nearest neighbours among the sub-sample (Figures 4, 5 and Supplementary Figure S6). This strategy offered significant speedup and improved the correspondence between clustering and low dimensional visualization of the data. As a sharp contrast to dropClust, 2D maps obtained from Seurat and Zheng *et al.* (Supplementary Figures S12 and S13), clearly struggled to mitigate overlap between clusters.

CONCLUSION

In this article we report dropClust, a novel algorithm for clustering and visualization of ultra-large single cell RNA-seq (scRNA-seq) data. Its speed and ability of delineating both major and minor cell types make it uniquely suitable for analysis of large scale scRNA-seq, data produced by droplet based transcriptomics platforms. dropClust is advantageous over the existing methods in detecting minor cell sub-populations. This much sought after feature is attained as a result of a number of careful and innovative design considerations including structure preserving data sampling and subjective selection of the most informative principal components. It is an end-to-end informatics pipeline for downstream analysis of droplet-seq data. The dropClust pipeline enables user to perform speedy analysis of ultra-large scRNA-seq data. The major functionalities of the end-to-end pipeline include data normalization, gene selection, unsupervised clustering of transcriptomes, low dimensional visualization of the complete data and differential expression analysis. With the increase in single cell transcriptomic throughput capabilities and technology availability (Chromium™ by 10× Genomics, ICELL8 by WaferGen Biosystems, similar platform by Illumina and Bio-Rad etc.) we predict unique relevance of the proposed dropClust pipeline.

AVAILABILITY

The dropClust R package is available at: https://github.com/debsin/dropClust.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

FUNDING

INSPIRE Faculty Fellowship [DST/INSPIRE/04/2015/003068 to D.S.] by the Department of Science and Technology (DST), Govt. of India; J.C. Bose Fellowship [SB/S1/JCB-033/2016 to S.B.] by the DST, Govt. of India; SyMeC Project grant [BT/Med-II/NIBMG/SyMeC/2014/Vol. II] given to the Indian Statistical Institute by the Department of Biotechnology (DBT), Govt. of India. Funding for open access charge: Institutional Funding.
Conflict of interest statement. None declared.

REFERENCES

1. Tanay,A. and Regev,A. (2017) Scaling single-cell genomics from phenomenology to mechanism. *Nature*, **541**, 331–338.
2. Zheng,G.X.Y., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.
3. Li,H., Courtois,E.T., Sengupta,D., Tan,Y., Chen,K.H., Goh,J.J.L., Kong,S.L., Chua,C., Hon,L.K., Tan,W.S. *et al.* (2017) Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.*, **49**, 708–718.
4. Kiselev,V.Y., Kirschner,K., Schaub,M.T., Andrews,T., Yiu,A., Chandra,T., Natarajan,K.N., Reik,W., Barahona,M., Green,A.R. *et al.* (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, **14**, 483–486.
5. Zeisel,A., Munoz-Manchado,A.B., Codeluppi,S., Lonnerberg,P., La Manno,G., Jureus,A., Marques,S., Munguba,H., He,L., Betsholtz,C. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.
6. Jiang,L., Chen,H., Pinello,L. and Yuan,G.-C. (2016) GiniClust: Detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol.*, **17**, 144.
7. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.
8. Xu,C. and Su,Z. (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, **31**, 1974–1980.
9. Levine,J.H., Simonds,E.F., Bendall,S.C., Davis,K.L., Amir,E.-a.D., Tadmor,M.D., Litvin,O., Fienberg,H.G., Jager,A., Zunder,E.R. *et al.* (2015) Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, **162**, 184–197.
10. Klein,A.M., Mazutis,L., Akartuna,I., Tallapragada,N., Veres,A., Li,V., Peshkin,L., Weitz,D.A. and Kirschner,M.W. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.
11. Campbell,J.N., Macosko,E.Z., Fenselau,H., Pers,T.H., Lyubetskaya,A., Tenen,D., Goldman,M., Verstegen,A.M., Resch,J.M., McCarroll,S.A. *et al.* (2017) A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.*, **20**, 484–496.
12. Blondel,V.D., Guillaume,J.-L., Lambiotte,R. and Lefebvre,E. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech.*, **2008**, P10008.
13. Gionis,A., Indyk,P. and Motwani,R. (1999) Similarity search in high dimensions via hashing. *VLDB*, **99**, 518–529.
14. Bawa,M., Condie,T. and Ganesan,P. (2005) LSH forest. In: *Proceedings of the 14th international conference on World Wide Web - WWW '05*, ACM Press.
15. Sengupta,D., Pyne,A., Maulik,U. and Bandyopadhyay,S. (2013) Reformulated kemeny optimal aggregation with application in consensus ranking of microRNA targets. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **10**, 742–751.
16. Xiang,Y., Gubian,S., Suomela,B. and Hoeng,J. (2013) Generalized simulated annealing for efficient global optimization: the GenSA Package for R. *R Journal*, **5**, 13–28.
17. Treutlein,B., Brownfield,D.G., Wu,A.R., Neff,N.F., Mantalas,G.L., Espinoza,F.H., Desai,T.J., Krasnow,M.A. and Quake,S.R. (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, **509**, 371–375.
18. Stauffer,C. and Grimson,W. (1999) Adaptive background mixture models for real-time tracking. In: *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. IEEE Comput. Soc.
19. Schwarz,G. (1978) Estimating the Dimension of a Model. *Ann. Statist.*, **6**, 461–464.

20. Langfelder,P., Zhang,B. and Horvath,S. (2007) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, **24**, 719–720.
21. Langfelder,P., Zhang,B. and with contributions from Steve Horvath. (2016) dynamicTreeCut: methods for detection of clusters in hierarchical clustering dendrograms. R package version 1.63-1.
22. Okada,R., Kondo,T., Matsuki,F., Takata,H. and Takiguchi,M. (2008) Phenotypic classification of human CD4+ T cell subsets and their differentiation. *Int. Immunol.*, **20**, 1189–1199.
23. Schiott,A., Lindstedt,M., Johansson-Lindbom,B., Roggen,E. and Borrebaeck,C.A. (2004) CD27(-) CD4(+) memory T cells define a differentiated memory population at both the functional and transcriptional levels. *Immunology*, **113**, 363–370.
24. Colpitts,S.L., Dalton,N.M. and Scott,P. (2009) IL-7 receptor expression provides the potential for long-term survival of both CD62Lhigh central memory T cells and Th1 effector cells during leishmania major infection. *J. Immunol.*, **182**, 5702–5711.
25. Lee,W.-Y., Sanz,M.-J., Wong,C.H.Y., Hardy,P.-O., Salman-Dilgimen,A., Moriarty,T.J., Chaconas,G., Marques,A., Krawetz,R., Mody,C.H. and Kubes,P. (2014) Invariant natural killer T cells act as an extravascular cytotoxic barrier for joint-invading Lyme Borrelia. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 13936–13941.
26. van Gisbergen,K.P., Kragten,N.A., Hertoghs,K.M., Wensveen,F.M., Jonjic,S., Hamann,J., Nolte,M.A. and van Lier,R.A. (2012) Mouse Hobit is a homolog of the transcriptional repressor Blimp-1 that regulates NKT cell effector differentiation. *Nat. Immunol.*, **13**, 864–871.
27. Chu,P.G. and Arber,D.A. (2001) CD79: a review. *Appl. Immunohistochem. Mol. Morphol.*, **9**, 97–106.
28. Oksvold,M.P., Kullmann,A., Forfang,L., Kierulf,B., Li,M., Brech,A., Vlassov,A.V., Smeland,E.B., Neurauter,A. and Pedersen,K.W. (2014) Expression of B-cell surface antigens in subpopulations of exosomes released from B-cell lymphoma cells. *Clin. Therap.*, **36**, 847–862.
29. Bade,B. (2005) Differential expression of the granzymes A, K and M and perforin in human peripheral blood lymphocytes. *Int. Immunol.*, **17**, 1419–1428.
30. Tu,T.C., Brown,N.K., Kim,T.-J., Wroblewska,J., Yang,X., Guo,X., Lee,S.H., Kumar,V., Lee,K.-M. and Fu,Y.-X. (2015) CD160 is essential for NK-mediated IFN- production. *J. Exp. Med.*, **212**, 415–429.
31. Le Bouteiller,P., Tabiasco,J., Polgar,B., Kozma,N., Giustiniani,J., Siewiera,J., Berrebi,A., Aguerre-Girr,M., Bensussan,A. and Jabrane-Ferrat,N. (2011) CD160: a unique activating NK cell receptor. *Immunol. Lett.*, **138**, 93–96.
32. Turman,M.A., Yabe,T., McSherry,C., Bach,F.H. and Houchins,J.P. (1993) Characterization of a novel gene (NKG7) on human chromosome 19 that is expressed in natural killer cells and T cells. *Hum. Immunol.*, **36**, 34–40.
33. Ogawa,K., Takamori,Y., Suzuki,K., Nagasawa,M., Takano,S., Kasahara,Y., Nakamura,Y., Kondo,S., Sugamura,K., Nakamura,M. and Nagata,K. (2003) Granulysin in human serum as a marker of cell-mediated immunity. *Eur. J. Immunol.*, **33**, 1925–1933.
34. Valés-Gmez,M., Esteso,G., Aydogmus,C., Blázquez-Moreno,A., Marn,A.V., Briones,A.C., Garcillán,B., Garca-Cuesta,E.-M., Lpez Cobo,S., Haskologlu,S. *et al.* (2016) Natural killer cell hyporesponsiveness and impaired development in a CD247-deficient patient. *J. Allergy Clin. Immunol.*, **137**, 942–945.
35. Lorenzo,J. (2010) *The Effects of Immune Cell Products (Cytokines and Hematopoietic Cell Growth Factors) on Bone Cells*, Academic Press.
36. Ida,H., Utz,P.J., Anderson,P. and Eguchi,K. (2005) Granzyme B and natural killer (NK) cell death. *Modern Rheumatology*, **15**, 315–322.
37. Barros,M. H.M., Hauck,F., Dreyer,J.H., Kempkes,B. and Niedobitek,G. (2013) Macrophage polarisation: an immunohistochemical approach for identifying M1 and M2 macrophages. *PLoS ONE*, **8**, e80908.
38. Ziegler-Heitbrock,H.L., Passlick,B. and Flieger,D. (1988) The monoclonal antimonocyte antibody My4 stains B lymphocytes and two distinct monocyte subsets in human peripheral blood. *Hybridoma*, **7**, 521–527.
39. Yan,W.X., Armishaw,C., Goyette,J., Yang,Z., Cai,H., Alewood,P. and Geczy,C.L. (2008) Mast cell and monocyte recruitment by S100A12 and its hinge domain. *J. Biol. Chem.*, **283**, 13035–13043.
40. Goyette,J.D. (2008) *The Extracellular Functions of S100A12 PhD thesis Medical Sciences, Faculty of Medicine*, UNSW.
41. Bandala-Sanchez,E., Zhang,Y., Reinwald,S., Dromey,J.A., Lee,B.-H., Qian,J., Bhmer,R.M. and Harrison,L.C. (2013) T cell regulation mediated by interaction of soluble CD52 with the inhibitory receptor Siglec-10. *Nat. Immunol.*, **14**, 741–748.
42. Sugimoto,N., Oida,T., Hirota,K., Nakamura,K., Nomura,T., Uchiyama,T. and Sakaguchi,S. (2006) Foxp3-dependent and-independent molecules specific for CD25+ CD4+ natural regulatory T cells revealed by DNA microarray analysis. *Int. Immunol.*, **18**, 1197–1209.
43. Hruz,T., Laule,O., Szabo,G., Wessendorp,F., Bleuler,S., Oertle,L., Widmayer,P., Gruissem,W. and Zimmermann,P. (2008) Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. *Adv. Bioinformatics*, **2008**, 1–5.
44. Collin,M., McGovern,N. and Haniffa,M. (2013) Human dendritic cell subsets. *Immunology*, **140**, 22–30.
45. Merad,M., Sathe,P., Helft,J., Miller,J. and Mortha,A. (2013) The dendritic cell lineage: ontogeny and function of dendritic cells and their subsets in the steady state and the inflamed setting. *Annu. Rev. Immunol.*, **31**, 563–604.
46. Lambert,M.P., Meng,R., Harper,D., Xiao,L., Marks,M.S. and Poncz,M. (2014) Megakaryocytes exchange significant levels of their alpha-granular PF4 with their environment. *Blood*, **124**, 1432–1432.
47. Sakurai,K., Fujiwara,T., Hasegawa,S., Okitsu,Y., Fukuhara,N., Onishi,Y., Yamada-Fujiwara,M., Ichinohasama,R. and Harigae,H. (2016) Inhibition of human primary megakaryocyte differentiation by anagrelide: A gene expression profiling analysis. *Int. J. Hematol.*, **104**, 190–199.
48. Stafforini,D.M., McIntyre,T.M., Zimmerman,G.A. and Prescott,S.M. (1997) Platelet-activating factor acetylhydrolases. *J. Biol. Chem.*, **272**, 17895–17898.
49. Ikawa,T., Fujimoto,S., Kawamoto,H., Katsura,Y. and Yokota,Y. (2001) Commitment to natural killer cells requires the helix-loop-helix inhibitor Id2. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 5164–5169.
50. Ramirez,K. and Kee,B.L. (2010) Transcriptional regulation of natural killer cell development. *Curr. Opin. Immunol.*, **22**, 193–198.
51. Jahrsdorfer,B., Vollmer,A., Blackwell,S.E., Maier,J., Sontheimer,K., Beyer,T., Mandel,B., Lunov,O., Tron,K., Nienhaus,G.U., Simmet,T., Debatin,K.-M., Weiner,G.J. and Fabricius,D. (2009) Granzyme B produced by human plasmacytoid dendritic cells suppresses T-cell expansion. *Blood*, **115**, 1156–1165.
52. Masten,B.J., Olson,G.K., Tarleton,C.A., Rund,C., Schuyler,M., Mehran,R., Archibeque,T. and Lipscomb,M.F. (2006) Characterization of myeloid and plasmacytoid dendritic cells in human lung. *J. Immunol.*, **177**, 7784–7793.
53. Petrovic,S. (2006) A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In: *Proceedings of the 11th Nordic Workshop of Secure IT Systems*. pp. 53–64.
54. Grn,D., Lyubimova,A., Kester,L., Wiebrands,K., Basak,O., Sasaki,N., Clevers,H. and van Oudenaarden,A. (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, **525**, 251–255.
55. Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.
56. Wagner,A., Regev,A. and Yosef,N. (2016) Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, **34**, 1145–1160.