Domanskyi *et al. BMC Bioinformatics* (2019) 20:369
https://doi.org/10.1186/s12859-019-2951-x

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Polled Digital Cell Sorter (p-DCS): Automatic identification of hematological cell types from single cell RNA-sequencing clusters

Sergii Domanskyi[1*†], Anthony Szedlak[1†], Nathaniel T Hawkins[1], Jiayin Wang[2], Giovanni Paternostro[3] and Carlo Piermarocchi[1]

Abstract

Background: Single cell RNA sequencing (scRNA-seq) brings unprecedented opportunities for mapping the heterogeneity of complex cellular environments such as bone marrow, and provides insight into many cellular processes. Single cell RNA-seq has a far larger fraction of missing data reported as zeros (dropouts) than traditional bulk RNA-seq, and unsupervised clustering combined with Principal Component Analysis (PCA) can be used to overcome this limitation. After clustering, however, one has to interpret the average expression of markers on each cluster to identify the corresponding cell types, and this is normally done by hand by an expert curator.

Results: We present a computational tool for processing single cell RNA-seq data that uses a voting algorithm to automatically identify cells based on approval votes received by known molecular markers. Using a stochastic procedure that accounts for imbalances in the number of known molecular signatures for different cell types, the method computes the statistical significance of the final approval score and automatically assigns a cell type to clusters without an expert curator. We demonstrate the utility of the tool in the analysis of eight samples of bone marrow from the Human Cell Atlas. The tool provides a systematic identification of cell types in bone marrow based on a list of markers of immune cell types, and incorporates a suite of visualization tools that can be overlaid on a t-SNE representation. The software is freely available as a Python package at https://github.com/sdomanskyi/DigitalCellSorter.

Conclusions: This methodology assures that extensive marker to cell type matching information is taken into account in a systematic way when assigning cell clusters to cell types. Moreover, the method allows for a high throughput processing of multiple scRNA-seq datasets, since it does not involve an expert curator, and it can be applied recursively to obtain cell sub-types. The software is designed to allow the user to substitute the marker to cell type matching information and apply the methodology to different cellular environments.

Keywords: Single cell RNA sequencing, Cell type identification, Biomarkers, Bone marrow

*Correspondence: domansk6@msu.edu
†Sergii Domanskyi and Anthony Szedlak contributed equally to this work.
[1]Department of Physics and Astronomy, Michigan State University, 48824 East Lansing, MI, USA
Full list of author information is available at the end of the article

Background

Bulk RNA-sequencing has provided the bioinformatics community with a large volume of high quality data over the past decade. However, bulk measurements make studying the transcriptomics of heterogeneous cell populations difficult and provides limited insight on complex systems composed of interacting cell types. Single cell RNA-seq (scRNA-seq) techniques promise to provide the field of bioinformatics with samples sufficiently large to resolve the subtleties of heterogeneous cell populations [1, 2].

The identification of cell types based on specific molecular signatures is challenging. This is particularly true in samples obtained from ex vivo bone marrow or periferal blood samples, where different types of hematological cells coexist and interact. scRNA-seq of periferal blood mono-nuclear cells (PBMC) and bone marrow mononuclear cells (BMMC) is nowadays possible with high level of sensitivity (see e.g. [3]). Monitoring different cell types and their heterogeneity in these hematological tissues has important applications in precision immunology, and it could help in determining the optimal therapeutic solutions in different hematological cancers.

The classification of the hematopoietic and immune system is predominantly based on a group of cell surface molecular markers named *Clusters of Differentiation* (CD), which are widely used in clinical research for diagnosis and for monitoring disease [4]. These CD markers can play a central role in the mediation of signals between the cells and their environment. The presence of different CD markers may therefore be associated with different biological functions and with different cell types. More recently, these CD markers have been integrated in comprehensive databases that also include intra-cellular markers. An example is provided by CellMarker [5]. This comprehensive database was created by a curated search through PubMed and numerous companies' marker handbooks including R&D Systems, BioLegend (Cell Markers), BD Biosciences (CD Marker Handbook), Abcam (Guide to Human CD antigens), Invitrogen ThermoFisher Scientific (Immune Cell Guide), and eBioscience ThermoFisher Scientific (Cytokine Atlas). Here we use a list of markers of immune cell types taken directly from a published work by Newman et al. [6] where CIBERSORT, a computational tool for deconvolution of cell types from bulk RNA-seq data, was introduced.

Using cell markers on each single cell RNA-seq data for a one-by-one identification would not work for most of the cells. This is fundamentally due to two reasons: (1) The presence of a marker on the cell surface is only loosely associated to the mRNA expression of the associated gene, and (2) single cell RNA-sequencing is particularly prone to dropout errors (i.e. genes are not detected even if they are actually expressed). The first step to address these limitations is unsupervised clustering. After clustering, one can look at the average expression of markers to identify the clusters. Several clustering methods have been recently used for clustering single cell data (for recent reviews see [7, 8]). Some new methods are able to distinguish between dropout zeros from true zeros (due to the fact that a marker or its mRNA is not present) [9], which has been shown to improve the biological significance of the clustering. However, once the clusters are obtained, the cell type identification is typically assigned manually by an expert using a few known markers [3, 10]. While in some cases a single marker is sufficient to identify a cell type, in most cases human experts have to consider the expression of multiple markers and the final call is based on their personal empirical judgment.

An example where a correct cell type assignment requires the analysis of multiple markers is shown in Fig. 1, where we analyzed single cell data from the bone marrow of the first donor from the HCA (Human Cell Atlas) preview dataset. HCA Data Portal [11] After clustering (Fig. 1a), the pattern of CD4 expression (Fig. 1b) suggests that cluster #1 (red) and cluster #2 (light green) are both highly enriched for CD4+, potentially indicating T helper cells. However, a more careful analysis of cluster #2 shows a significant expression of CD68 and CD33 (Fig. 1c, d) that indicates that this cluster consists more likely of Macrophages/Monocyte cells. Figure 1d shows an example of another important marker, CD38, expressed in many immune cells including T cells, B cells and Monocyte cells.

We would like to emphasize our method differences with respect to cell type identification in bulk data, where the main issue is deconvolution, i.e. extracting the relative fraction of cell types in data from a mixture. There are no clusters that have to be labeled in the bulk case and the nature of the problem a little different than in the single cell case. Several deconvolution algorithms have been developed in the past for estimating the relative composition of complex tissues from bulk transcriptomics data. [6, 12–18] These methodologies are based on predefined signature matrices that contain the relative expression of markers, not just the presence/absence of a marker, for different cell types. Regression methods are then used to infer the relative proportions in a mixture. These approaches, however, use lists of markers obtained from the literature as a starting point, and these lists can be integrated in our p-DCS to identify single cells, as we have done here.

In this paper we present a methodology that, after unsupervised clustering, automatically assigns clusters to cell type based on a systematic, unbiased, voting algorithm. Our method does not rely on a human expert empirically selecting a set of markers to interpret the results, but uses all the information available in a large markers database to

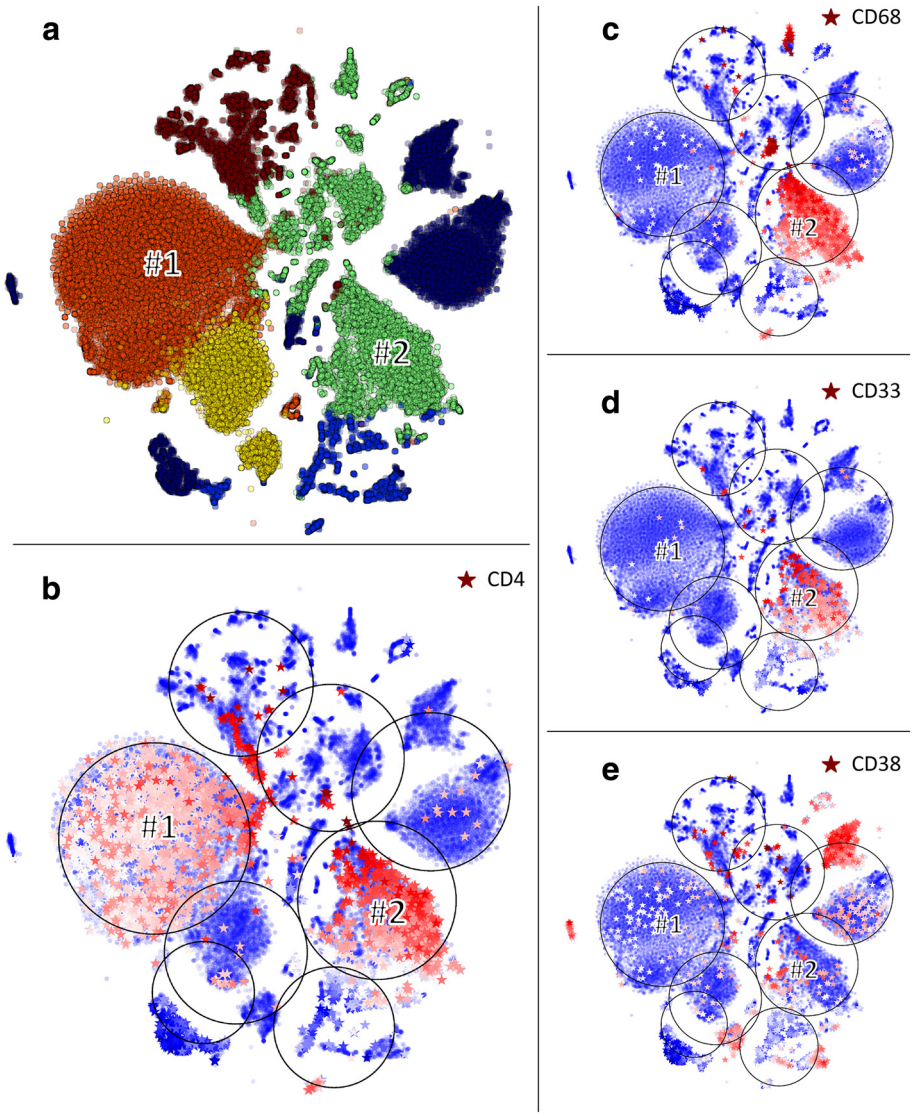

Fig. 1 Markers analysis. **a** t-SNE layout of clusters obtained from the first donor of the HCA preview dataset [11]. **b** CD4 marker expression displayed on a t-SNE layout: cells where CD4 is expressed are shown as stars colored according to the expression level from blue (lowest expression) to red (highest expression), large black circles infer the cluster sizes. Cells in which the marker is not expressed are shown as circles. **c**-**e** Expression of CD68, CD33 and CD38 shown as in (**b**)

predict cell types. While cell type identification by manual interpretation can provide good results, the proposed methodology assures that all the available information is taken into account in an unbiased way, and it allows for the identification of many datasets in parallel. From an algorithmic point of view, voting algorithms are among the simplest and most successful approaches to implement fault tolerance and obtain reliable data from multiple unreliable channels [19]. The idea can be traced back to von Neumann [20], and since then it has been practically used in many error correction computational architectures. The voting algorithm employed here belongs to the class of approval voting algorithms. For a given cluster, each participant (a cell marker) votes for a subset of candidates (cell types) that meet the participant criteria (significant RNA expression) for the position rather than picking just one candidate. The approval vote tally determines the score that we use to assign the cluster to a cell type.

Methods

Overview

Our p-DCS consists of two main modules: (a) clustering and (b) cell type assignment, which are both

based on an unsupervised approach. We demonstrate our methodology using public bone marrow scRNA-seq data from eight donors [11], that will be referred to as BM1--BM8. The data was produced by 10x Genomics with raw counts matrix generated by Cell Ranger with GRCh38, standard 10X reference. The 8 donors average median of

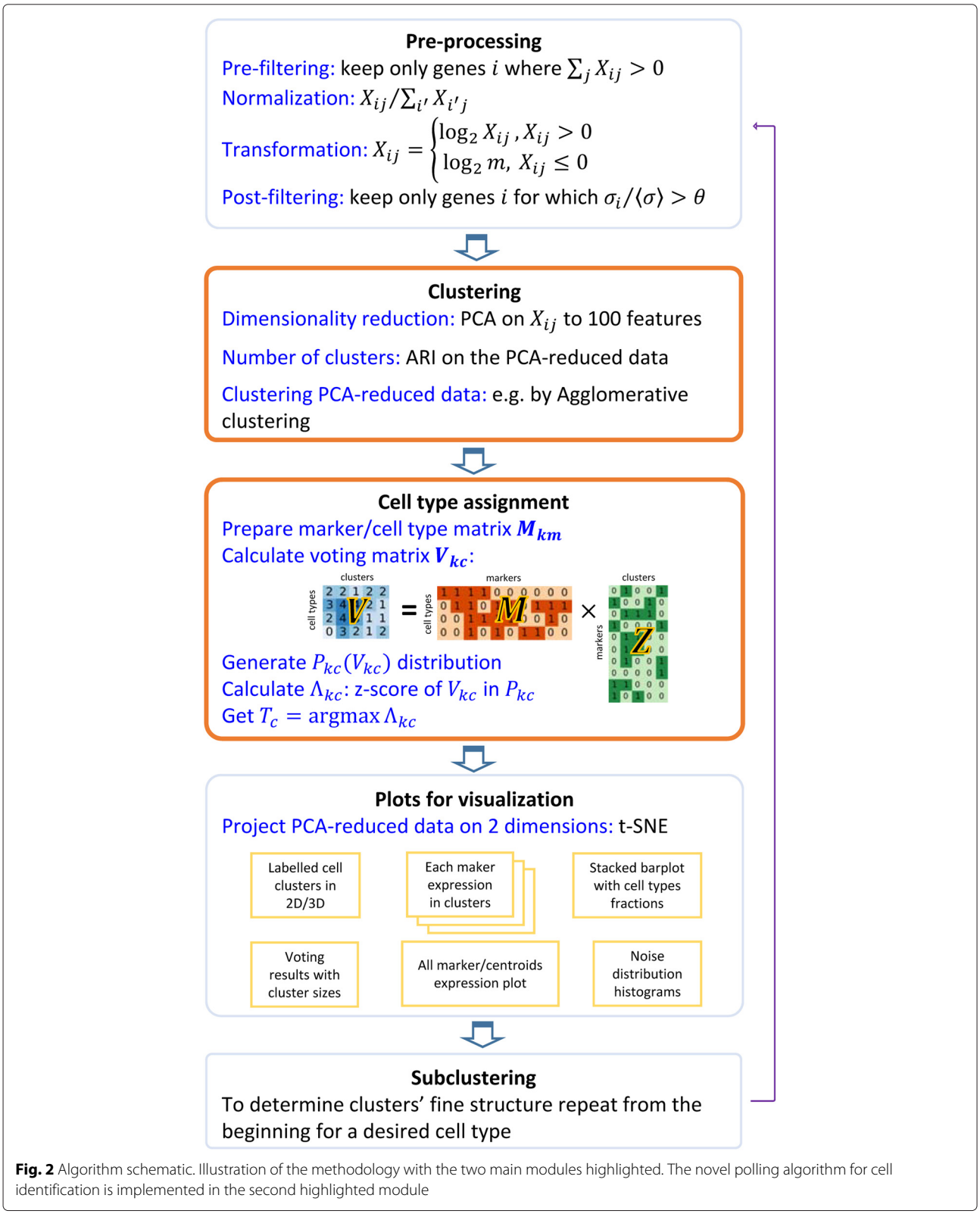

Fig. 2 Algorithm schematic. Illustration of the methodology with the two main modules highlighted. The novel polling algorithm for cell identification is implemented in the second highlighted module

genes per cell is 688, and we did not impute dropout reads. To visualize data the fast interpolation-based t-distributed Stochastic Neighbor Embedding (FIt-SNE) layout recently developed by Linderman et al. [21] can be used. In the software we provide a switch allowing to use either the regular t-SNE (default option) or the FIt-SNE. In this section, we will illustrate the methodology using the first dataset BM1. The remaining bone marrow data along with a large scRNA-seq PBMC dataset, obtained from a different study [3], are analyzed in "Results and discussion" section. In "Results and discussion" section we also show how the proposed methodology can be used recursively, so that for each main cell type one can find the corresponding sub-types. Figure 2 shows the workflow of the methodology. The two main modules are identified by the "Clustering" and "Cell type assignment" labels. The clustering module is preceded by data pre-processing, and a set of visualization tools is included in the software.

Initial gene/cell filtering and normalization

The expression matrix, X_{ij}, the expression of gene i in cell j where $i = 1, \ldots, N$ and $j = 1, \ldots, p$ is normalized following the steps outlined in [3]. The gene expression matrix is first filtered to keep only genes i that are expressed in at least one cell ($\sum_j X_{ij} > 0$). The expression in all cells must then be mapped to the same range of total expression to account for differing yields from PCR amplification. Each cell's expression vector is thus divided by the sum of all its expression values so that

$$X_{ij} \leftarrow X_{ij} \Big/ \sum_{i'} X_{i'j} \ , \tag{1}$$

where the left arrow indicates reassignment of the matrix values. Because gene expression values in RNA-seq measurements tend to span many orders of magnitude, it is helpful to apply a standard $\log_2$ transformation, which is done either to get "fold changes" when comparing groups in differential expression analysis, or to get a "normal" looking statistical distribution. However, the many zeros inherent in single cell RNA-seq data requires the zeros to be replaced with positive values. We choose to replace all zeros with m, the smallest nonzero value in X_{ij}, so that

$$X_{ij} \leftarrow \begin{cases} \log_2 X_{ij} & \text{if } X_{ij} > 0 \\ \log_2 m & \text{otherwise} \end{cases} . \tag{2}$$

Finally, we keep only those genes exhibiting sufficiently high variation as parameterized by a threshold θ,

$$\frac{\sigma_i}{\langle \sigma \rangle} \geq \theta \tag{3}$$

where σ_i is the standard deviation of gene i's expression across all cells and $\langle \sigma \rangle = N^{-1} \sum_i \sigma_i$. For this analysis, we chose $\theta = 0.3$.

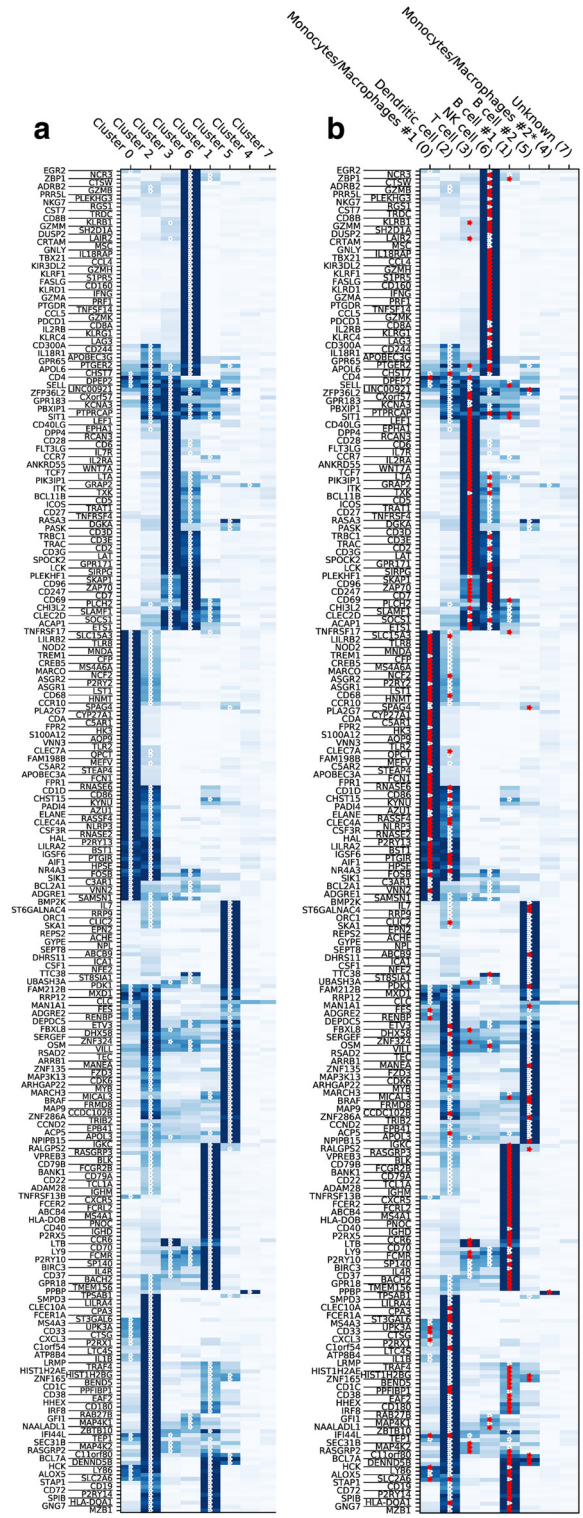

Fig. 3 Marker expression for scRNA-seq HCA BM dataset, subset BM1. **a** Mean expression of marker genes in clusters of yet unidentified cell types. Stars denote genes expressed above a certain z-score threshold. **b** Mean expression of marker genes in clusters with inferred cell type with cluster index in parentheses. Red stars highlight the supporting markers in assigning the cluster cell type

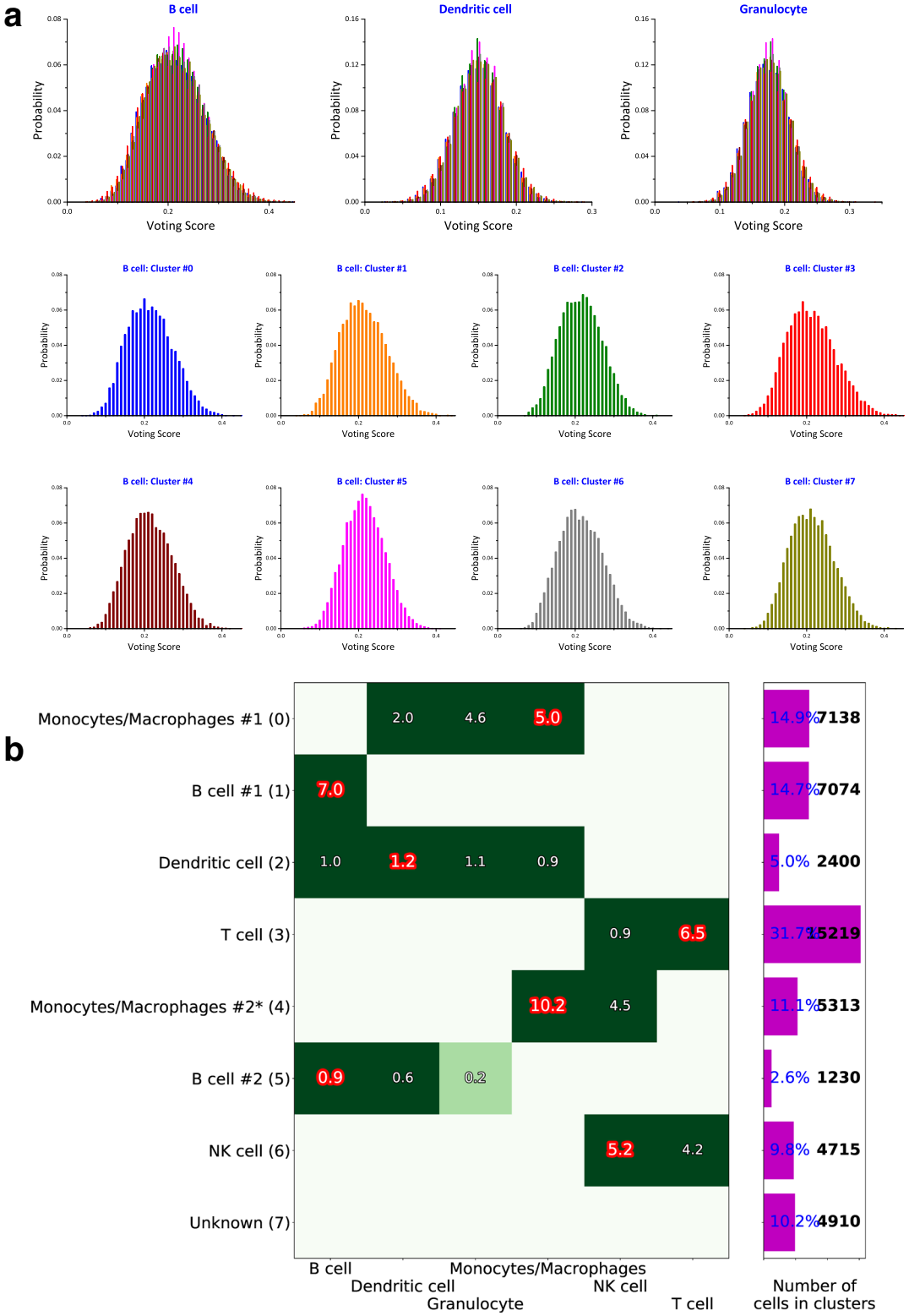

Fig. 4 Voting results visualization. Exemplified on HCA BM1 dataset. **a** $\mathcal{P}_{kc}(V_{kc})$ distributions shown in separate plots for the first three cell types k, different cluster c are shown in different color detailed for cell type "B cell" in the separate 8 histograms, one for each cluster. **b** Visualization of the matrix Λ_{kc}, where columns are the possible cell types and rows are the assigned cell types T_c, with cluster indices 0,1,...,7 in parentheses. The negative z-scores are not shown. The barplot on the right shows relative (%) and absolute (cell count) cluster sizes. Cell clusters that have 3 or less supporting markers are marked with "*", see Fig. 3 for supporting markers

Clustering

The clustering algorithms used in p-DCS require to specify the number of clusters n. The first step is therefore to find a good value for the parameter n. We used the Adjusted Rand Index (ARI) [22] between pairs of clusterings obtained from the same set using a stochastic algorithm (Mini-batch K-Means) and averaging the results to obtain the ARI curve as a function of n. An ARI of one signifies that two clusters are identical. The optimal n corresponds then to the first peak coming from the $n = \infty$ side of the ARI curve (see Fig. 5 below for an example). To remove noisy components and accelerate the procedure, clustering is conducted on a smaller array $\tilde{X}_{ij}$ defined by projecting X_{ij} onto its first 100 principal components (i.e. $\tilde{X}_{ij}$ has $i = 1 \dots 100$). We clustered the cells in $\tilde{X}_{ij}$ using the agglomerative clustering method available in `scikit-learn` [23]. Clustering diagrams such as Fig. 1a are generated by running `scikit-learn`'s t-SNE routine on $\tilde{X}_{ij}$, projecting from 100 to two dimensions (simply for the sake of generating a figure). Cells are colored according to their cluster index. 100 principal components (PCs) were used because the total amount of explained variance increases first rapidly until about 20-25 PCs. Including the top-100 PCs assures that we go beyond this first rapid increase in all samples and capture on average about 25% of the total variance. Note that the two t-SNE dimensions are not equivalent to the first two PCA components. PCA is a linear method, while t-SNE is a nonlinear dimensionality reduction. The layout of the cells in the t-SNE plot is therefore using information from all the 100 PCs.

Cell type assignment

The cell type assignment is based on our voting algorithm idea that uses a database of marker genes. Since this application focuses on bone marrow data, we used a list of markers of immune cells from Newman et al. [6] as our marker/cell type database, D. The latter is used to create a marker/cell type table, specific to a gene expression dataset of interest, e.g. the matrix X of BM1. The table for a given dataset is created after the initial gene filtering and normalization discussed above. For each cell type in D we keep all genes that are expressed. In this way we build a marker/cell type matrix M_{km} where k is the cell type (e.g. T cell), m is the marker gene (e.g. CD4). The element $M_{km} = 1$ if m is an expressed marker of cell type k and 0 otherwise.

Building the matrix M_{km} represents the first step of the voting algorithm. This is equivalent to defining "ballots" in which each qualified voter, i.e. the markers chosen, has a list of candidate cell types they can approve. We normalize $\tilde{M}_{km} = M_{km} / \sum_{m'} M_{km'}$ by the number of markers expressed in each cell type so that the absolute number of known markers in a given cell type is irrelevant. Then we normalize $\tilde{M}_{km}$ by the number of cell types expressing that marker. This second normalization is important because a marker that is unique to a particular cell type will be automatically assigned a large weight. For

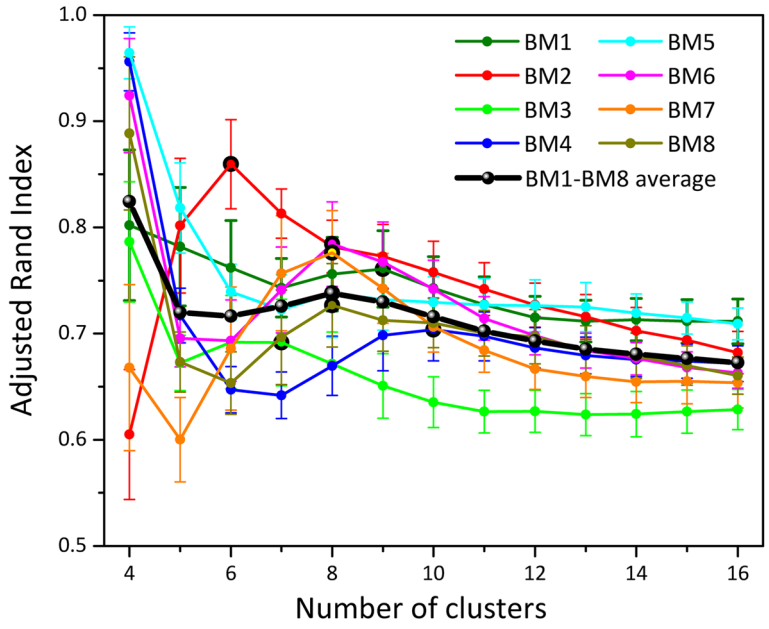

Fig. 5 HCA BM dataset analysis. Adjusted Rand Index (ARI) curves for each dataset BM1-BM8. Clustering was done using Mini-Batch K-Means from scikit-learn. The black line represents the average of the 8 datasets, and the peak at $n = 8$ was used to select the optimal number of clusters

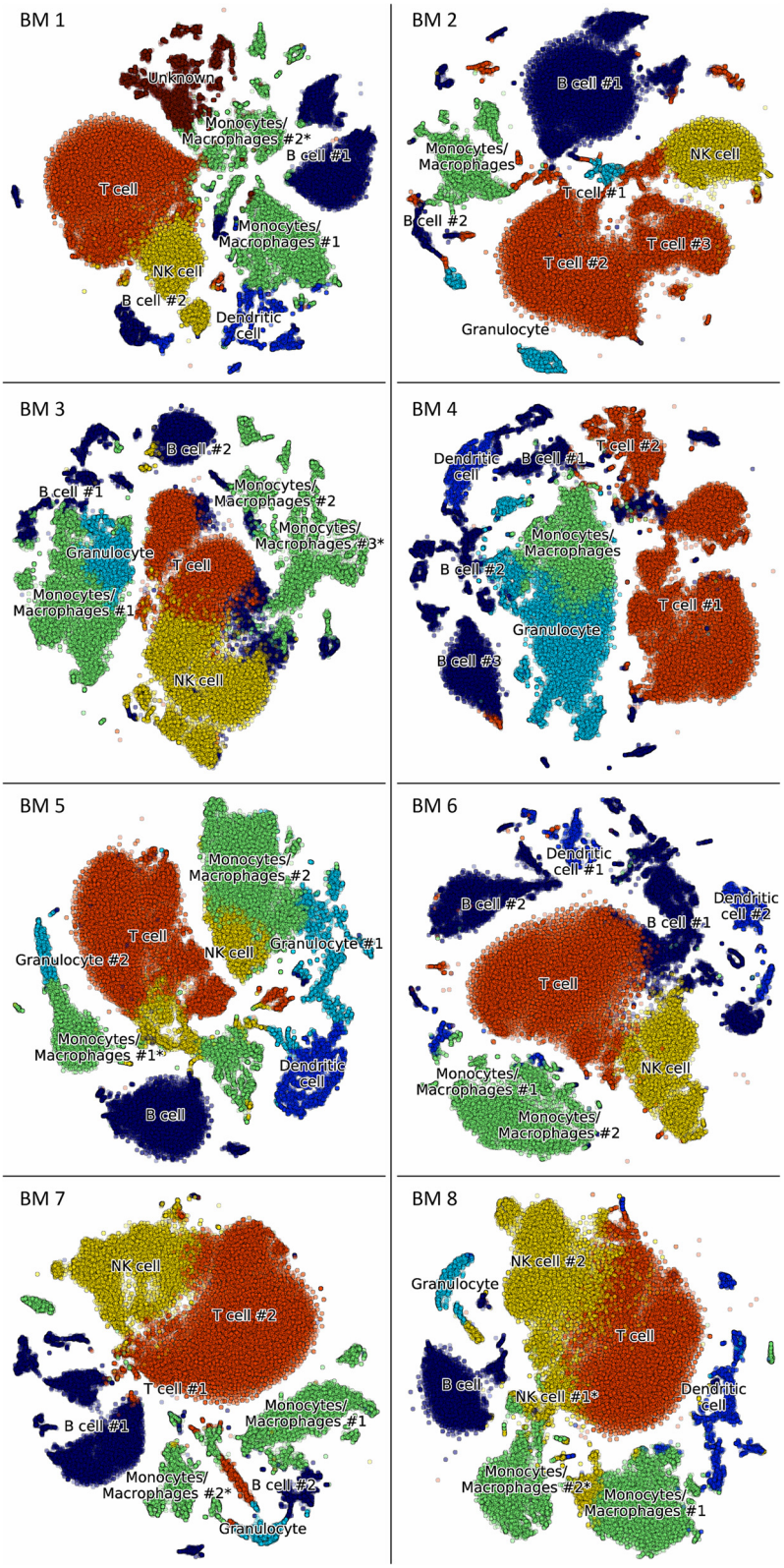

Fig. 6 HCA BM preview dataset analysis. Clustering illustrated with t-SNE plots for each patient in the dataset. The cell type identification is assigned based on the voting algorithm discussed in "Methods" section

each cluster c, the voting algorithm is then implemented as follows:

(i) We build the marker/centroid matrix Y_{mc}, where Y_{mc} is the mean expression of marker m across all cells in cluster c. For each marker m, we use Y_{mc} to compute all cluster centroids' z-scores Z_{mc}. Then we build the matrix $\tilde{Z}_{mc} = 1$ if $Z_{mc} \geq \zeta$ and $\tilde{Z}_{mc} = 0$ otherwise for a given threshold ζ. With increasing values of ζ the number of possible supporting markers decreases. We have varied the parameter ζ in the range 0.1-1.5, and for this application, we chose $\zeta = 0.3$, which provides a reasonable number of markers for all cell types. This procedure is needed to identify markers that are significantly expressed in one cluster compared to the other clusters. Figure 3a shows Y_{mc}, calculated for HCA BM1 dataset: darker blue color corresponds to higher expression of markers, and the stars correspond $\tilde{Z}_{mc} = 1$, i.e. statistically significant markers with z-score larger than ζ among all markers as tested across clusters. The general approach used for selecting ζ has been be to start with $\zeta = 0$ (which does not filter for noise) and increasing its value until the number of matching markers is almost constant.

(ii) We compute the vote matrix according to $V_{kc} = \sum_m \tilde{M}_{km}\tilde{Z}_{mc} / \sum_{mk'} \tilde{M}_{k'm}\tilde{Z}_{mc}$. This is when each voter (the markers) matches a given cluster to a single or more possible cell types. This matrix contains an approval score for each type-cluster pair (k, c).

(iii) To quantify the statistical significance of the approval scores and make the final assignment, we use a stochastic method to quantify the statistical uncertainty associated to each type-cluster pair (k, c). We randomize the clusters by preserving their size and assigning to them cells randomly chosen from the whole dataset, and repeat steps (i) and (ii) to compute the approval scores. This randomization is performed $n = 10^4$ times, recording the voting matrix V_{kc} for each configuration of random clusters. This method accounts for cluster sizes, the overall gene expression distribution of the markers, and imbalances in the number of markers per cell type in estimating the uncertainty. The procedure provides distributions of voting results $\mathcal{P}_{kc}(V_{kc})$ for a null model of random clusters. Figure 4a shows histograms of the distributions $\mathcal{P}_{kc}(V_{kc})$ calculated for the same dataset of Fig. 3. The figure shows three different cell types in separate plots, and each plot

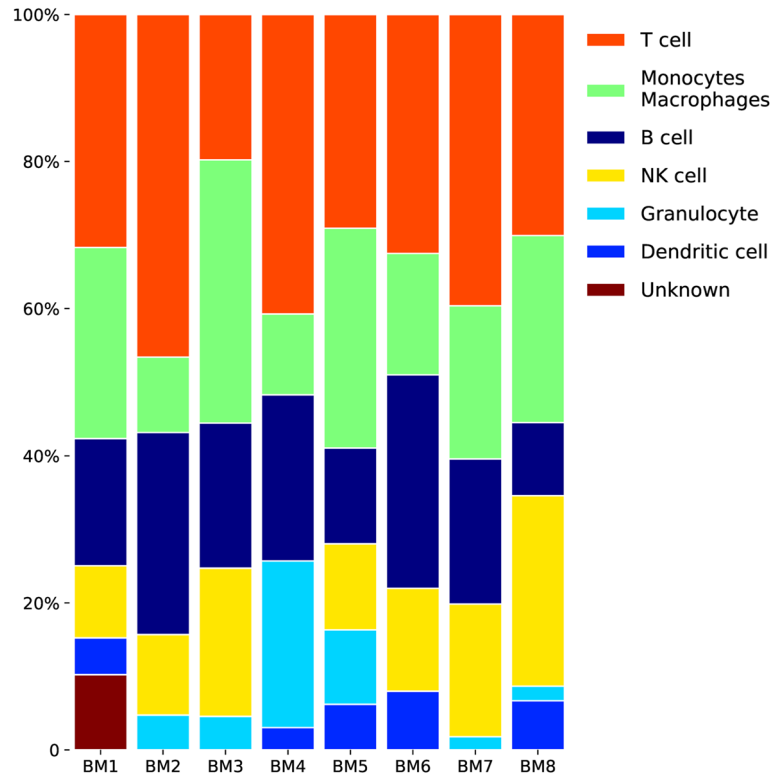

Fig. 7 HCA BM dataset summary. Cell type relative fractions for each BM sample. The cell types are sorted by average (across samples) fraction size, with the exception of the "Unknown" which is moved to the bottom. Color coding for cell types is identical to Fig. 6

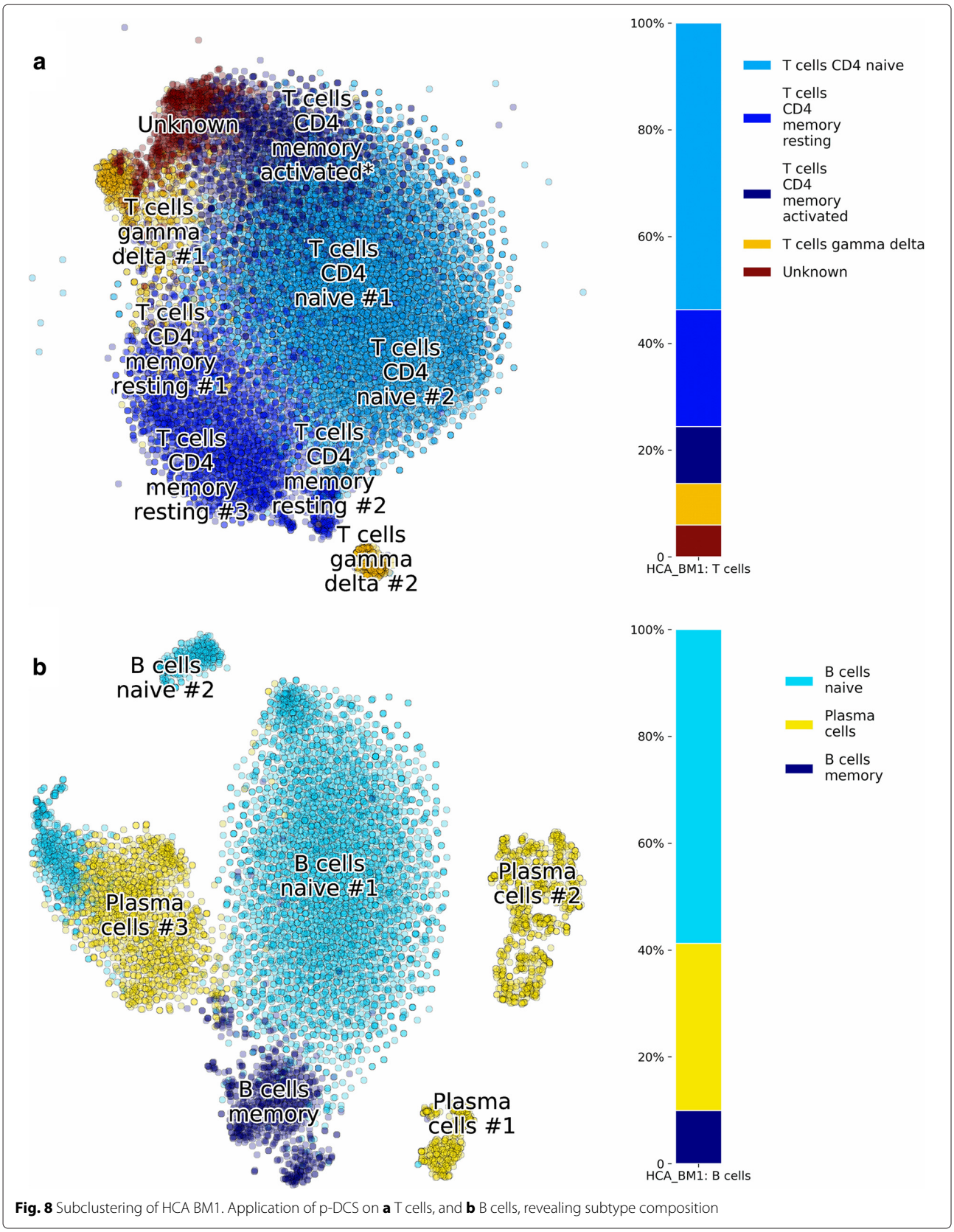

Fig. 8 Subclustering of HCA BM1. Application of p-DCS on **a** T cells, and **b** B cells, revealing subtype composition

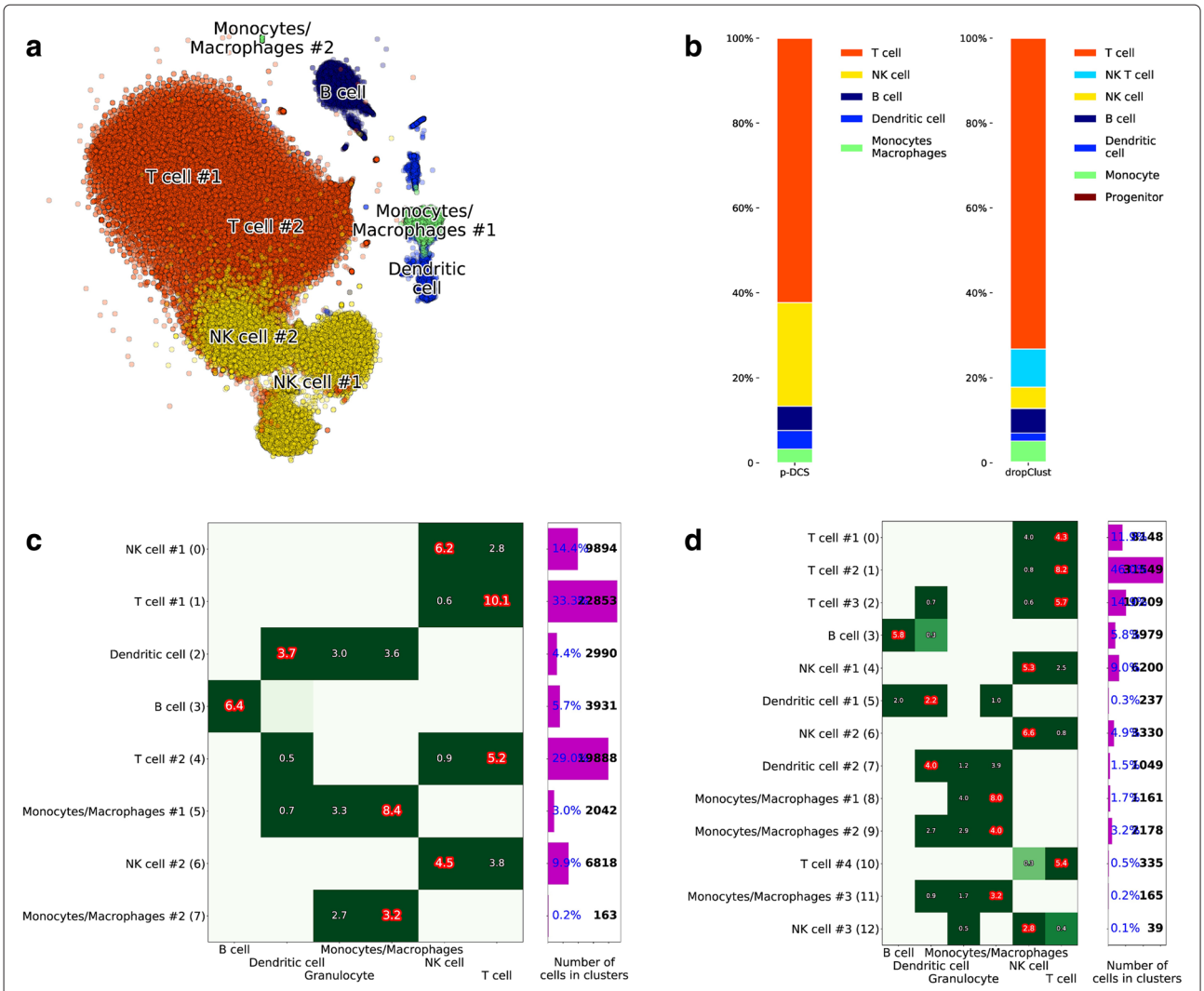

Fig. 9 PBMC dataset processing. **a** Clustering with inferred cell types. **b** Fractions of various cell types obtained by p-DCS in comparison with DropClust manual clusters' annotations [10]. **c** Visualization of the voting results of all possible cell types (columns) and identified clusters (rows), generated from the input marker cell/type table. **d** Same as (**c**) for clustering method by Sinha et al. [10] and 13 clusters

contains the distributions of each cluster in a different color. Note that the distributions do not show a strong dependence on the cluster index c, but they can be very different for different cell types k.

(iv) Finally, we determine the z-scores, Λ_{kc}, of the voting results V_{kc} in (ii), given the null distribution $\mathcal{P}_{kc}(V_{kc})$ calculated in (iii) and assign the cell type according to $T_c = \text{argmax}_k \Lambda_{kc}$. All cells belonging to cluster c are thus identified as cell type T_c. Fig. 4b is a visual representation of Λ_{kc}, shown only for positive values, where the indices k, c are along the x- and y-axis, respectively. After the cell types are determined, the panel (b) of Fig. 3 is produced, with all the markers supporting the assigned identification marked as red stars.

Note that this marker/cell type table is only one of many possible reasonable choices. The software was designed to allow the user to easily substitute this table with a custom table relevant to the particular cell population under investigation. Likewise, the voting scheme outlined above can be replaced with any custom function with the same inputs and outputs. See the documentation for details and examples. [24]

Results and discussion

In this section, we first present the results obtained with our methodology using recently-published data from normal bone marrow samples (the data identified above as BM1-BM8, containing a total of 378k cells). Additionally, we compare our cell type assignment to an existing

Table 1 Comparison of p-DCS and DropClust on PBMC scRNA-seq ~68.6k cells dataset

Cell type	*p-DCS*	*DropClust*
T cell	Cluster #1, 4: 62.3%	Cluster #0, 1, 2, 10: 73.2%
NK T cell		Cluster #4: 9.0%
NK cell	Cluster #0, 6: 24.3%	Cluster #6, 12: 5.0%
B cell	Cluster #3: 5.7%	Cluster #3: 5.8%
Dendritic cell	Cluster #2: 4.4%	Cluster #5, 7: 1.8%
Monocytes/Macrophages	Cluster #5, 7: 3.2%	Cluster #8, 9: 4.9%
Progenitor		Cluster #11: 0.2%

identification of cell types from a large scRNA-seq ~68.6k cells PBMC dataset.

Results on the HCA BM data

Number of clusters

We first calculated the Adjusted Rand Index (ARI) [22] curves for BM1-BM8. For each n between 4 and 16, Mini-batch K-Means clustering was performed 12 times leading to 12 different partitions of the data. The ARI between all the possible 66 pairs of partitions was then calculated and averaged. The procedure was repeated in $N = 200$ independent runs to obtain error bars. The ARI curves are shown in Fig. 5. Note that the ARI curves often have a maximum at or near $n = 1$. This maximum does not provide useful information, and the optimal n is therefore associated to the first peak observed coming from the right side of the plot. In addition to the ARI for each of the BM1-BM8 sets, Fig. 5 displays their average in black. The latter has a peak at $n = 8$, and we therefore select that value for clustering all the datasets.

Clustering and identification in BM1-BM8 datasets

The BM samples were analyzed individually and their cluster plots were combined to demonstrate the similarity between the 8 datasets of bone marrow, see Fig. 6. The color coding is uniform for the cell types across the 8 datasets, i.e. all T cells are colored orange, B cells – dark blue, etc. As some of the clusters overlap on the t-SNE plot [25, 26], it is useful to calculate the relative fractions of cells of various cell types. The latter provide a snapshot of the cellular composition of the 8 bone marrow samples, see Fig. 7.

Clustering of T and B cells sub-types

We applied the methodology illustrated above to identify sub-types of major hematological B and T cells. Additional marker/cell subtype tables M_{km} were prepared for this analysis. Columns of these new matrices indicates sub-types only and rows are the markers/genes that are known to be expressed the these sub-types. We used the same list of immune cell types used above from Newman et al. [6] to build the M_{km} matrices for B and T cells. As above, these matrices M_{km} are created ensuring that only expressed makers are included for each sub-type. Cell sub-types with no expressed makers after pre-processing are discarded.

Clustering with $n = 9$ for T cell subtypes from BM1 is shown in Fig. 8a, revealing Naive T cell and Memory T cell subtypes. In the same way, B cells of BM1 were processed into 6 clusters in Fig. 8b, showing populations of Naive B cells, Memory B cells and a group of Plasma cells.

We have tried different strategies for identifying cell sub-types. The best approach consists in first identifying major cell types and then separately analyzing each of them as shown in this section. We have tried to include major cell types and their sub-types in the matrices M_{km} and have attempted their identification with a larger number of clusters. Such an approach leads often to incorrect results with relative cell frequencies that are incompatible with normal physiological ranges. Also, using only sub-types and removing the major types is not a robust strategy. Depending on the database of markers used, the results are sometimes completely inconsistent with some published expert annotation on the same data (see congruence analysis below).The reason for such incorrect results is that when an arbitrary marker is expressed in several clusters, e.g. sub-populations of T cell, only the cluster(s) with z-score above cutoff ζ will have this marker contribute in voting. Thus, adding many sub-types into the marker-cell type list increases the chance of incorrectly annotating the cluster(s).

Table 2 Cell counts from cell-by-cell validation of p-DCS and dropClust on PBMC scRNA-seq ~68.6k cells dataset

p-DCS cell type (count)	*dropClust cell type*						
	T cell	*NK T cell*	*NK cell*	*B cell*	*Monocyte*	*Dendritic cell*	*Progenitor*
T cell (42741)	42608	73	88	42		1	2
NK cell (16712)	7257	6137	3317	1			
B cell (3931)	84	5	1	3841			
Dendritic cell (2990)	57	7		76	1802	1048	
Monocyte (2205)	235	13	1	19	1537	237	163

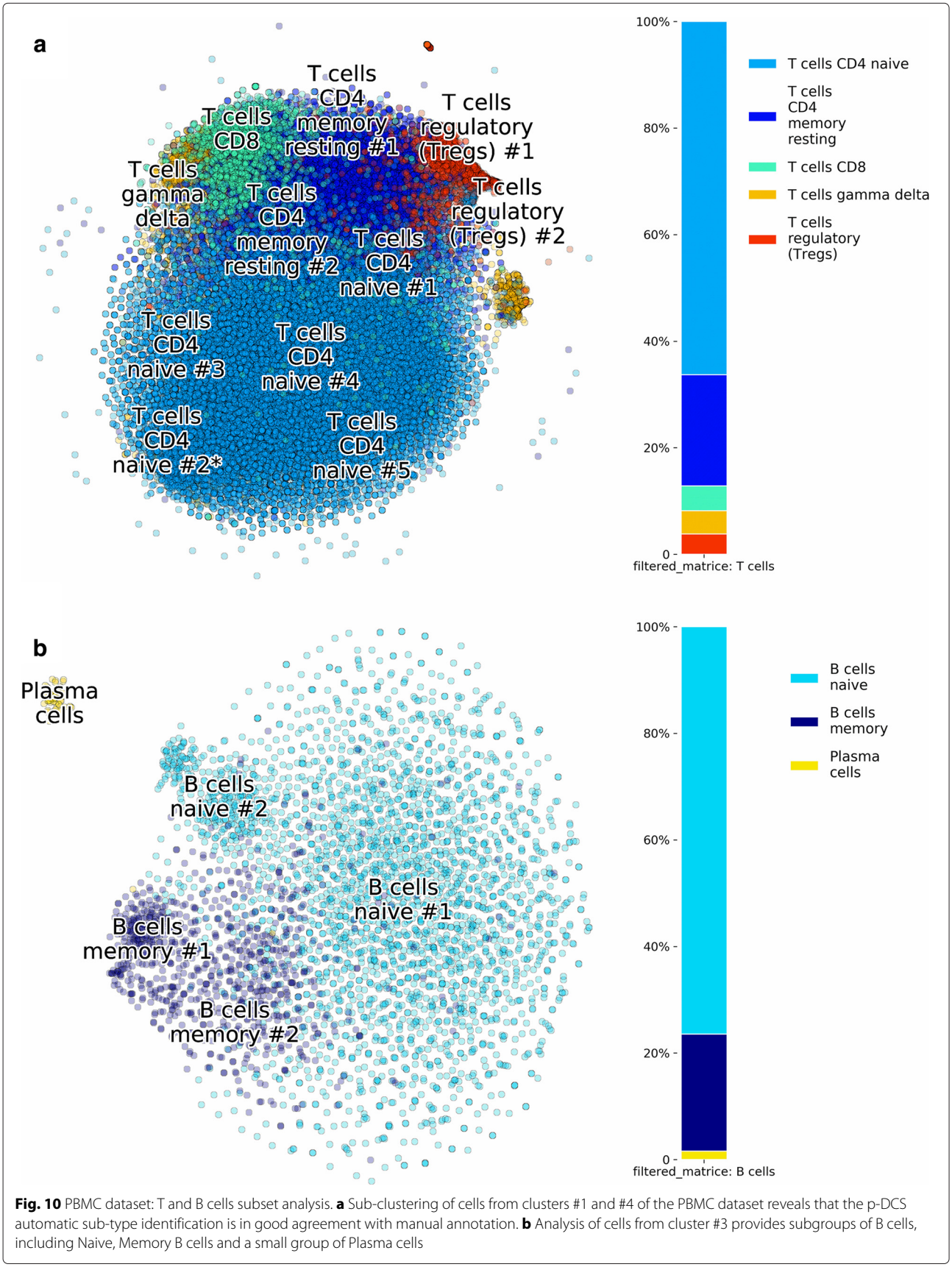

Fig. 10 PBMC dataset: T and B cells subset analysis. **a** Sub-clustering of cells from clusters #1 and #4 of the PBMC dataset reveals that the p-DCS automatic sub-type identification is in good agreement with manual annotation. **b** Analysis of cells from cluster #3 provides subgroups of B cells, including Naive, Memory B cells and a small group of Plasma cells

Congruence with expert annotation on PBMC dataset

In a recent work, Sinha et al. [10] presented their dropClust algorithm to cluster ultra-large scRNA-seq datasets. To illustrate their algorithm, they used data from 68k PBMC from Zheng et al. [3]. The 68k PBMC data was collected with GemCode single-cell technology (GEM - Gel bead in EMulsion) by 10x Genomics using Illumina NextSeq 500 High Output. The median number of genes with nonzero expression per cell for this dataset is 525. Their cluster annotation, obtained from a manual assessment using a few selected markers, is of interest here and can be used to compare the annotation obtained by our automated methodology with one obtained manually by an expert. By pre-processing the whole 68k PBMC dataset, we determined that the optimal number of clusters was 8. The result of the analysis is shown in Fig. 9. The clustering and cell type inference from the automated p-DCS procedure are shown in Fig. 9(a), indicating that T cells constitute the major cell type in this sample. Figure 9b shows a graphical comparison of cell types fractions obtained by p-DCS and by Sinha et al. [10]. The frequencies of various cell types are expected to vary from individual to individual, and the fractions that we determined are within the normal ranges [27]. The main difference in cell type frequencies, Fig. 9b, using the two approaches is in the p-DCS NK cell cluster (yellow), which in Sinha et al. is split into NK (yellow) and NK T (light blue) cells. The NK T cells expresses a combination of T cell and NK cell markers, and therefore distinguishing NK form NK T cells is challenging. Figure 9c displays the candidate cell types used in the voting and the z-scores of the voting scores. A cluster receiving a high z-scores in more that one cell type indicates that it is composed by multiple cell types, e.g. the NK cluster #2 in Fig. 9 with a z-score of 4.5 likely has a significant amount of T cells (z-score 3.8) in addition to NK cells. A full quantitative comparison is also available in Table 1. In addition to comparing the size of cluster between p-DCS and dropClust, we individually analyzed all cells, i.e. their barcodes in the scRNA-seq data, to check if they were assigned to matching cell types. For each cell type annotated by p-DCS we counted how many cells were annotated by Sinha et al. [10] into each of their categories (Table 2). Overall the agreement is strong, with the exception of NK cells and Dendritic cells for which we observed a significant mismatch. Figure 9d shows the annotation of the PBMC dataset for the clustering method dropClust [10] with 13 clusters. Interestingly, all clusters but #4, 11 and 12 are annotated identically to the reference annotations. Cluster #4 is NK T cells in the reference, whereas we did not have this cell type in the list and we labeleld the cluster as NK cells. Similarly we do not have NK progenitors in our matrix of markers, therefore the algorithm assigned cluster #12 to NK cells. Cluster #11 is labelled differently for the same reason of cluster #12.

Table 3 Subclustering of ∼42.7k T cells from a ∼68.6k cell dataset

T cell subtype	*p-DCS*	*DropClust*
Naive T cell	Cluster #1, 5, 6, 8, 9: 41.4%	Cluster #1: 46.0%
$\gamma\delta$ T cell	Cluster #2: 2.7%	
CD8 T cell	Cluster #10: 2.9%	Cluster #0: 11.8%
Memory T cell	Cluster #3, 4: 13.0%	Cluster #2: 14.9%
Regulatory T (Treg) cell	Cluster #0, 7: 2.3%	Cluster #10: 0.5%

Comparison of p-DCS and DropClust subtypes assignment

Sub-clustering of T cells was also done to compare the two approaches. T-cells from clusters #1, 4 (see Fig. 9) were processed with a new list of markers/cell sub-types. The results of cell sub-types annotation are presented in Fig. 10, and the detailed comparison to the results by Sinha et al. [10] are in Tables 3 and 4.

Alternative cell marker input lists

We have used the list of markers from Newman et al. [6]. Alternative lists of markers can be obtained by the Cellmarker database[5], or by the database of the Human Cell Differentiation Molecules (HCDM) organization [28], which is sponsored by a number of large companies. The latter contains detailed information about each CD molecule, including structure, function, and cellular expression. The HCDM and Cellmarker databases provide alternatives to the list of markers used here. We have observed that the marker overlap between these databases is very strong.

Conclusions

We have presented a methodology that, after unsupervised clustering of scRNA-seq data, automatically assigns clusters to cell types based on a voting algorithm without

Table 4 Cell counts from cell-by-cell validation of p-DCS and dropClust on subset (T cells) of PBMC scRNA-seq ∼68.6k cells dataset

p-DCS cell type (count)	*dropClust cell type*				
	CD4 Naive T	*CD4 memory T*	*CD8 T*	*Treg*	*other*
CD4 Naive T (28355)	26345	1779	222	6	3
CD4 memory resting T (8905)	2888	5808	186	19	4
CD8 T (1978)	688	545	727		18
Treg (1622)	301	948	38	300	35
$\gamma\delta$ T (1881)	745	477	582	4	73

manual interpretation by an expert curator. The method provides the classification of individual cells into predefined classes based on a database of known molecular signatures, i.e. cell surface (extracellular) and intracellular markers. The proposed methodology assures that extensive marker/cell type information is taken into account in a systematic way when assigning clusters to cell types. Moreover, the method allows for a high throughput processing of multiple scRNA-seq datasets since it does not involve an expert curator.

In addition to determining major cell types, we have shown how this methodology can be applied recursively to obtain cell sub-types. We have performed a congruence analysis of cluster identification obtained by our method with those obtained by expert curators on the same dataset, showing that the automatic assignment is consistent with expert assignment both of major cell types and cell sub-types. While we have focused on the identification of hematological cell types, the software is designed to allow the user to substitute the marker table to apply the methodology to different tissues.

Abbreviations

ARI: Adjusted rand index; BMMC: Bone marrow mono-nuclear cells; CD: Clusters of differentiation; HCA: Human cell atlas; HCDM: Human cell differentiation molecules; PBMC: Peripheral Blood mono-nuclear cells; PCA: Principal component analysis; p-DCS: Polled digital cell sorter; tSNE: t-distributed Stochastic Neighbor Embedding

Acknowledgements

We thank Prof. George I. Mias and Prof. Michael Bachmann for helpful suggestions.

Authors' contributions

SD, AS, and CP designed the algorithms. SD, AS, JW, and NH wrote the software. GP provided bio-medical analysis. SD, AS, and CP wrote the manuscript. All the authors have read and approved the manuscript.

Funding

This work was supported by National Institutes of Health, Grant No. R01GM122085. The funding body had no role in the design of the study and collection, analysis, interpretation of data and in writing the manuscript.

Availability of data and materials

Analyzed here HCA BM data, available to the research community, was obtained from HCA Data Portal https://preview.data.humancellatlas.org/. The 68k PBMC data, by Zheng et al., used for the p-DCS methodology validation is available at https://support.10xgenomics.com/single-cell-gene-expression/datasets/. The software is available as a Python package at https://github.com/sdomanskyi/DigitalCellSorter.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

JW is an employee of Salgomed Inc., and CP and GP own equity in Salgomed Inc.

Author details

[1]Department of Physics and Astronomy, Michigan State University, 48824 East Lansing, MI, USA. [2]Salgomed, Inc., 92014 Del Mar, CA, USA. [3]Sanford Burnham Prebys Medical Discovery Institute, 92037 La Jolla, CA, USA.

Received: 14 January 2019 Accepted: 13 June 2019
Published online: 01 July 2019

References

1. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol. 2016;34(11):1145.
2. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, et al. Science forum: the human cell atlas. Elife. 2017;6:27041.
3. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
4. Zola H, Swart B, Nicholson I, Voss E. Leukocyte and Stromal Cell Molecules: the CD Markers. Haboken: Wiley; 2007.
5. Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, Luo T, Xu L, Liao G, Yan M, Ping Y, Li F, Shi A, Bai J, Zhao T, Li X, Xiao Y. CellMarker: a manually curated resource of cell markers in human and mouse. https://doi.org/10.1093/nar/gky900. https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky900/5115823. Accessed 17 Oct 2018.
6. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. 12(5):453–7. https://doi.org/10.1038/nmeth.3337. Accessed 14 Nov 2018.
7. Andrews TS, Hemberg M. Identifying cell populations with scRNASeq. 59:114–122. https://doi.org/10.1016/j.mam.2017.07.002. Accessed 31 Aug 2018.
8. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell rna-seq data. Nat Rev Genet. 2019. https://doi.org/10.1038/s41576-018-0088-9.
9. Gong W, Kwak I-Y, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: imputing dropout events in single cell RNA sequencing data. 19(1):. https://doi.org/10.1186/s12859-018-2226-y. Accessed 31 Aug 2018.
10. Sinha D, Kumar A, Kumar H, Bandyopadhyay S, Sengupta D. dropClust: efficient clustering of ultra-large scRNA-seq data. 46(6):36. https://doi.org/10.1093/nar/gky007. Accessed 9 Apr 2018.
11. HCA Data Portal. https://preview.data.humancellatlas.org/. Accessed 9 May 2018.
12. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol. 2013;25(5):571–8.
13. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PloS ONE. 2009;4(7):6098.
14. Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, Bongiovanni S, Szustakowski JD. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PloS one. 2011;6(11):27156.
15. Qiao W, Quon G, Csaszar E, Yu M, Morris Q, Zandstra PW. Pert: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. PLoS Comput Biol. 2012;8(12):1002838.
16. Liebner DA, Huang K, Parvin JD. Mmad: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. Bioinformatics. 2013;30(5):682–9.
17. Zhong Y, Wan Y-W, Pang K, Chow LM, Liu Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics. 2013;14(1):89.
18. Zuckerman NS, Noam Y, Goldsmith AJ, Lee PP. A self-directed method for cell-type identification and separation of gene expression microarrays. PLoS Comput Biol. 2013;9(8):1003189.
19. Parhami B. Voting algorithms. IEEE Trans Reliab. 1994;43(4):617–29.
20. von Neumann J. Probabilistic logics and the synthesis of reliable organisms from unreliable components. Automata Studies. 1956;34:43–99.
21. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. https://doi.org/10.1038/s41592-018-0308-4. Accessed 12 Feb 2019.
22. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66(336):846–50.

23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in python. J Mach Learn Res. 2011;12(Oct):2825–30.
24. Sdomanskyi/DigitalCellSorter: DigitalCellSorter. https://zenodo.org/record/2603265. Accessed 22 Mar 2019.
25. Maaten Lvd, Hinton G. Visualizing data using t-sne. J Mach Learn Res. 2008;9(Nov):2579–605.
26. Amir E-aD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, Pe'er D. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. 31(6):545–52. https://doi.org/10.1038/nbt.2594. Accessed 1 July 2019.
27. Kleiveland CR. In: Verhoeckx K, Cotter P, López-Expósito I, Kleiveland C, Lea T, Mackie A, Requena T, Swiatecka D, Wichers H, editors. Peripheral Blood Mononuclear Cells. Cham: Springer; 2015, pp. 161–7.
28. About HCDM. http://www.hcdm.org/index.php/about-hcdm. Accessed 9 May 2018.

Publisher's Note