Linear-time cluster ensembles of large-scale single-cell RNA-seq and multimodal data

Van Hoan Do[1], Francisca Rojas Ringeling[1], and Stefan Canzar*[1]

[1]Gene Center, Ludwig-Maximilians-Universität München, 81377 Munich, Germany

Abstract

A fundamental task in single-cell RNA-seq (scRNA-seq) analysis is the identification of transcriptionally distinct groups of cells. Numerous methods have been proposed for this problem, with a recent focus on methods for the cluster analysis of ultra-large scRNA-seq datasets produced by droplet-based sequencing technologies. Most existing methods rely on a sampling step to bridge the gap between algorithm scalability and volume of the data. Ignoring large parts of the data, however, often yields inaccurate groupings of cells and risks overlooking rare cell types. We propose method *Specter* that adopts and extends recent algorithmic advances in (fast) spectral clustering. In contrast to methods that cluster a (random) subsample of the data, we adopt the idea of landmarks that are used to create a sparse representation of the *full* data from which a spectral embedding can then be computed in linear time. We exploit Specter's speed in a cluster ensemble scheme that achieves a substantial improvement in accuracy across 16 scRNA-seq datasets and that is sensitive to rare cell types. Its linear time complexity allows Specter to cluster a dataset comprising 2 million cells in just 26 minutes. Furthermore, on CITE-seq data that simultaneously measures gene and protein marker expression we demonstrate that Specter is able to utilize multimodal omics measurements to resolve subtle transcriptomic differences between subpopulations of cells. Specter is open source and available at `https://github.com/canzarlab/Specter`.

1 Background

Single-cell RNA sequencing (scRNA-seq) has dramatically increased the resolution at which important questions in cell biology can be addressed. It has helped to identify novel cell types based on commonalities and differences in genome-wide expression patterns, reconstruct the heterogeneous composition of cell populations in tumors and their microenvironment, and unveil regulatory programs that govern the dynamic changes in gene expression along developmental trajectories.

One of the most fundamental computational tasks in the context of scRNA-seq analysis is the identification of groups of cells that are similar in their expression patterns, i.e. their transcriptomes, and which are at the same time distinct from other cells. Conceptually similar problems have been studied in anthropology [15] and psychology [57] almost a century ago and since then this so-called cluster analysis has become one of the most well-studied problems in unsupervised machine learning. Numerous methods have been proposed for clustering scRNA-seq data sets [16, 48], with Seurat [41] and its underlying Louvain clustering algorithm [6] being arguably the most widely used one. More recently, attempts have been made to design algorithms for the analysis of ultra-large scRNA-seq

*Correspondence: canzar@genzentrum.lmu.de

data sets, owing to the ever-increasing throughput of droplet-based sequencing technologies that allow to profile genome-wide expression for hundreds of thousands of cells at once. At the heart of such methods often lies a sampling technique that reduces the size of the data analyzed by a clustering algorithm. Cluster labels of cells in this so-called sketch are subsequently transferred to the remaining cells using, e.g., a nearest neighbor algorithm. dropClust [45], for example, includes a structure preserving sampling step, but initially picks a small set of cells simply at random. Similarly, Seurat applies random subsampling prior to its nearest neighbor search.

The quality of the final clustering, however, strongly depends on how well the data sketch represents the overall cluster structure and how accurate the cluster labels of cells in the sketch can be inferred from incomplete data. Inaccurate labels of subsampled cells will most likely lead to an inaccurate labeling of the full data, as illustrated in Figure 1a-b. In addition, sampling cells proportional to their abundance might render rare cell types invisible to the algorithm. Geometric sketching was therefore recently proposed as an alternative sampling method that selects cells according to the transcriptomic space they occupy rather than their abundance. Nevertheless, labels need to be inferred from partial data.

Spectral methods for clustering have been applied with great success in many areas such as computer vision, robotics, and bioinformatics. They make few assumptions on cluster shapes and are able to detect clusters that form non-convex regions. On a variety of data types, this flexibility has allowed spectral clustering methods to produce more accurate clusterings than competing methods [43]. The high computational complexity, however, renders its application to large-scale problems infeasible. For n data points, spectral clustering computes eigenvectors of a $n \times n$ affinity matrix, which incurs a computational cost of $\mathcal{O}(n^3)$. For scRNA-seq data sets with n in the order of ten thousands up to the millions this presents a prohibitive cost which has thus prevented the application of spectral clustering to large-scale single-cell data sets.

Furthermore, spectral clustering methods are sensitive to the right choice of parameters used to model the similarity between data points [50], i.e. RNA expression measurements of single cells. Data sets derived from different biological samples exhibiting different cell population structures obtained using different sequencing technologies typically require a different set of parameter values to achieve accurate clustering results.

Here, we address the challenges of computational complexity and parameter sensitivity to allow a tailored version of spectral clustering to be utilized in the analysis of large scRNA-seq data sets. In contrast to methods that cluster only a (random) subsample of the data, we take a fundamentally different approach that avoids learning from unlabeled partial data. We adopt the idea of landmarks [8], a random sample of cells that are used to create a sparse representation of the *full* data from which a spectral embedding can then be computed in linear time. We utilize the speed of this approach to systematically explore the parameter space by a co-association based consensus clustering scheme, also known in literature as cluster ensembles [47]: Instead of picking one set of parameters, we explore different choices of parameters and reconcile the resulting clustering information into a single (consensus) clustering. We combine consensus clustering with a novel *selective sampling* strategy that utilizes clustering information obtained from the *full* data set to achieve overall linear time complexity. Finally, we transfer cluster labels to the remaining cells using supervised k-nearest neighbors classification. Our selective sampling-based consensus clustering approach is illustrated in Figure 1c-d and detailed further in the "Methods" section.

We implemented our landmark-based cluster ensemble approach in the Specter software tool [1]. In a comprehensive benchmark analysis that includes 16 public scRNA-seq data sets we demonstrate that Specter produces substantially more accurate cell clusters than competing methods and is sensitive to rare cell types. Yet its linear time complexity translates to a practical efficiency that allows Specter to cluster a data set comprising more than 2 million cells in just 26 minutes.

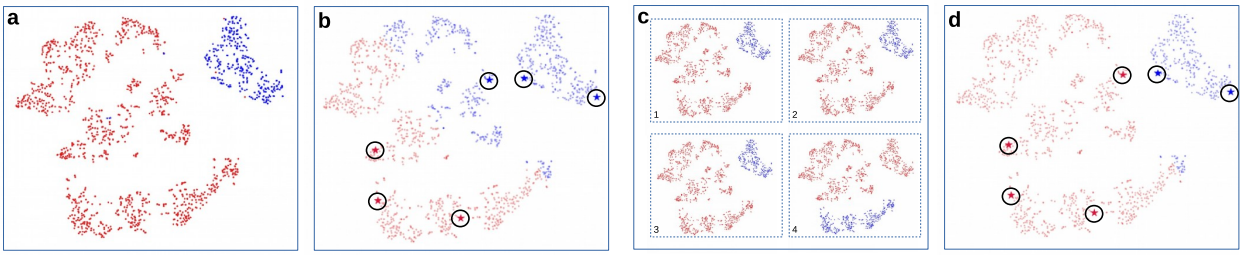

Figure 1: Selective sampling of cells labeled based on full data as compared to learning from unlabeled partial data. (**a**) t-SNE visualization of scRNA-seq data [21] colored to distinguish sorted hematopoetic stem cells (blue) from other mouse bone marrow cells (red). (**b**) 6 cells denoted by circled stars are sampled uniformly at random from the population and clustered by spectral clustering (similar results based on k-means clustering, not shown) into 2 groups denoted by red and blue stars, respectively. Cluster labels are then transferred to the remaining cells by 1-nearest neighbor classification (membership in red and blue). Compared to the true labeling in (a), many cells are incorrectly labeled. (**c**) Four clusterings are produced by our tailored landmark-based spectral clustering algorithm based on different similarity measures between cells. Three clusterings closely resemble the true labeling shown in (a), while one differs substantially. (**d**) For illustrative purposes we cluster the same subsample of 6 cells as in (b) but use labels computed based on the complete data in (c) to derive a consensus clustering. This time, the 4 red stars and the 2 blue stars correctly form 2 groups of cells, based on their co-association in the clusterings shown in (c). Again, cluster labels are propagated to the remaining cells using 1-nearest neighbor classification. In contrast to (b), the final clustering shown in (d) closely resembles the true clustering in (a).

In addition, we illustrate that the consensus clustering scheme applied by Specter is able to utilize multimodal omics measurements. Recent technological advances have allowed to simultaneously measure multiple modalities of single cells [56]. On two CITE-seq data sets that measure both gene expression and surface protein levels of cord blood [46] and peripheral blood mononuclear cells [33], Specter resolves subpopulations of cells that cannot accurately be distinguished based on transcriptomic differences alone.

2 Results

2.1 Specter is more accurate than competing methods

We compare the performance of Specter to state-of-the-art scRNA-seq clustering methods SC3 (v1.10.1) [27], Seurat (v2.3.4) [41], dropClust (v2.1.0) [45] as well as to a geometric sketching based clustering approach [23]. SC3 and Seurat consistently demonstrated superior performance over competing methods in several clustering benchmarks [16, 48] and are routinely used in scRNA-seq based cell type analyses. The graph-based Louvain clustering approach used by Seurat has an additional speed advantage over SC3, which applies a consensus clustering scheme to obtain particularly accurate clusterings. dropClust was recently proposed for the analysis of ultra-large scRNA-seq data sets and follows a strategy outlined above (Figure 1b). It first reduces the size of the data to a maximum of 20,000 cells using random sampling. After a second sampling step based on Louvain clusters, it applies average-linkage hierarchical clustering on the sampled cells. Cluster labels are then transferred to the remaining cells using a Locality Sensitive Hashing forest [4] for approximate nearest neighbor searches. In contrast, the geometric sketching algorithm proposed in

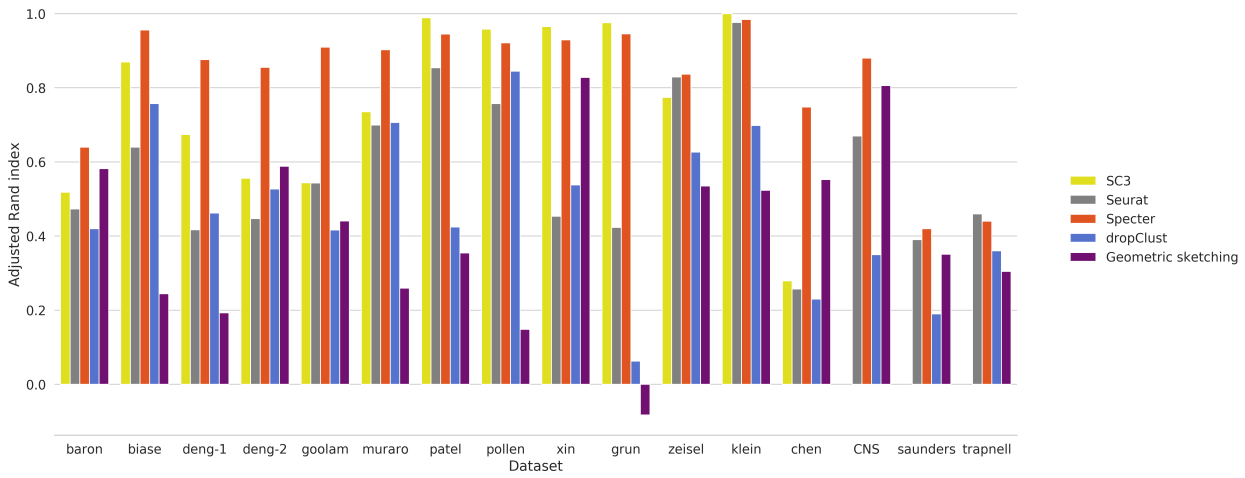

Figure 2: Clustering performance measured in ARI of Specter and competing methods on 15 scRNA-seq data sets. For computational reasons, SC3 was excluded from the comparison on the three largest data sets *CNS*, *saunders*, and *trapnell*. On these three data sets the more efficient SCANPY implementation was used in the Louvain clustering step of the Seurat pipeline. Geometric sketching refers to the Louvain clustering of 10% of the cells sampled using geometric sketching. Results for different sketch sizes are shown in Figures S1.

[23] samples cells evenly across the transcriptional space rather than proportional to the abundance of cell types as uniform sampling schemes do. Experiments in [23] demonstrated that clustering a geometric sketch using the graph-based Louvain algorithm followed by propagating labels to the remaining cells via k-nearest-neighbor classification accelerates clustering analysis and yields more accurate results than uniform sampling strategies. We include the same geometric sketching based clustering method in our benchmark and refer to it simply as geometric sketching throughout the text.

We evaluate Specter and competing methods on 15 public scRNA-seq data sets from mouse and human that vary in size and number of cell populations. Table S1 lists all data sets along with their characteristics. We apply standard and uniform preprocessing [16] on all data sets, including natural log-transformation of gene counts after adding a pseudo-count of 1, followed by dimensionality reduction to 100 principal components. Due to our clustering ensemble scheme, no additional tuning of parameters is required to apply Specter to the 15 data sets. SC3, Seurat, and dropClust are run with their built-in data preprocessing, while the geometric sketching based Louvain clustering is provided with the same preprocessed data as Specter. Consistent with the original publication [23], geometric sketches ranging from 2% to 10% of the original number of cells were computed and clustered as described above. All methods were provided the correct number of clusters or corresponding parameters were tuned accordingly. Due to the cubic dependence of its running time on the number of cells, we had to exclude SC3 from benchmarks on the three largest data sets that included more than $450,000$ cells (Table S1). On data set *chen*, for example, it takes SC3 five hours to cluster 14,000 cells. Similarly, on the three largest data sets we replaced the R implementation of the Louvain clustering algorithm called in the Seurat clustering pipeline by a more efficient python implementation of the same algorithm in the SCANPY package (v1.4.6) [52]. SCANPY was specifically designed for the analysis of large-scale gene expression data sets and was used originally [9] to identify cell types in data set *trapnell* comprising more than 2 million cells.

Consistent with other benchmarks (see e.g. [16]), we use the Adjusted Rand index (ARI) [25] to

measure the similarity between the inferred clusterings and the ground truth clustering that is based on the biological cell types taken from the original study. On all but one data set, Specter obtains more accurate clusterings than Seurat, and without exception achieves higher ARI scores than sampling based methods dropClust and geometric sketching, even when sampling as many as 10% of cells in the latter approach (Figure 2). On many instances, the improvement in ARI is substantial. Results for smaller sketch sizes are shown in Figures S1. Note that the ground truth labeling of cell types in data sets *trapnell*, *CNS*, and *saunders* was obtained in the original publication using Seurat or its underlying Louvain clustering algorithm. Despite the additional manual refinement applied in [9, 42, 54], this might positively impact the evaluation results of Seurat and the geometric sketching based Louvain clustering. On several instances, Specter achieves considerably higher ARI scores than SC3, while on others their performance is comparable. Note, however, that SC3 is not designed to cluster large data sets and had to be excluded from the comparison on the three largest data sets for computational reasons.

In addition, we compared Specter to the original implementation of the landmark-based spectral clustering (LSC) algorithm. Figure S2 demonstrates the effectiveness of our adoptions and extensions of the original algorithm to the analysis of scRNA-seq data (see the "Methods" section). Overall, we observe a substantial improvement of Specter over the classical LSC algorithm in terms of clustering accuracy. On data set *grun* that includes only two distinct populations of cells, for example, the increased number of eigenvectors in the spectral embedding (see the "Methods" section) potentially plays an important role in the boost in ARI score. Again, the only minor improvement on data set *trapnell* might be due to an extensive manual correction applied to a Louvain clustering in the original publication.

Finally, we illustrate in Figure 3 how a higher ARI score translates into a more meaningful representation of cell types.

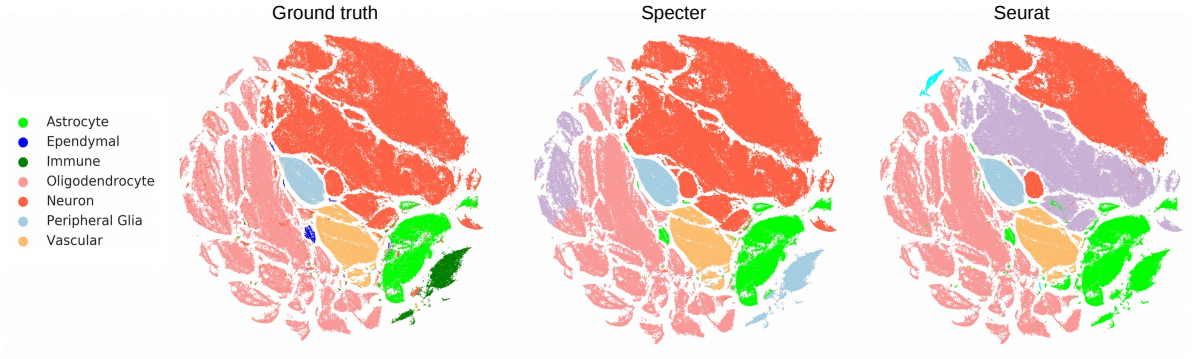

Figure 3: t-SNE visualization of single cells of the mouse nervous system (data set *CNS*). Cells in the ground truth representation (left) are colored by cell type specified by the legend. The visualization of Specter (middle) and Seurat (right) clusterings use the same 2D embedding as the ground truth, but cells are colored according to clusters inferred by the two methods; colors do not directly reflect cell types specified by the legend. As expected by the higher ARI score, Specter makes fewer mistakes. In contrast to Specter, Seurat wrongly splits neurons into 2 populations, is not able to distinguish astrocytes from immune cells, and is similarly not able to distinguish a subpopulation of vascular cells from astrocytes.

2.2 Specter detects rare population of inflammatory macrophages

In [23] the authors report and experimentally validated the discovery of a rare population of inflammatory macrophages by applying Louvain clustering to a geometric sketch of 20,000 cells sampled from a data set of 254,941 umbilical cord blood cells. In their experiments the authors observed that this rare subtype is invisible to Louvain clustering, the algorithm used by Seurat, unless cells are initially sampled evenly across transcriptional space to better balance the abundance of common and rare cell types. In contrast, Specter reveals a similar population of inflammatory macrophages characterized by the same set of marker genes *CD74*, *HLA-DRA*, *B2M* and *JUNB* (AUROC > 0.9) without any prior preprocessing (Figure 4).

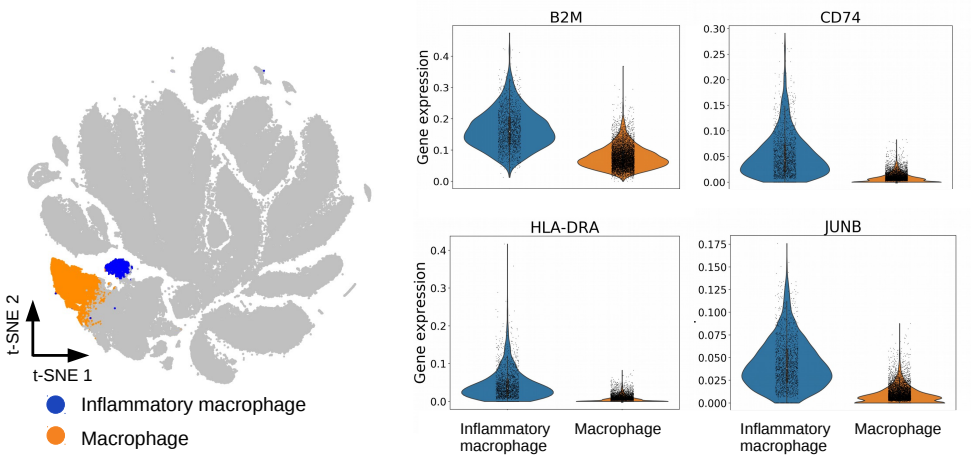

Figure 4: Clustering of 254,941 umbilical cord blood cell by Specter. Among macrophages defined by CD14 and CD68 marker gene expression, Specter detects a rare subpopluation of inflammatory macrophages that was recently discovered [23] *(left)*. This rare subtype can be distinguished in Specter's clustering by the expression of the same set of inflammatory marker gene expression (*CD74*, *HLA-DRA*, *B2M*, and *JUNB*) used for its identification in [23].

2.3 Specter utilizes multi-modal data to resolve subtle transcriptomic differences

In this Section, we demonstrate the ability of Specter to utilize complementary information provided by multi-modal data to refine the clustering of single cells. More specifically, we re-analyzed two public data sets of 4,292 healthy human peripheral blood mononuclear cells (PBMC) [33] and 8,617 cord blood mononuclear cells (CBMC) [46], for which both mRNA and protein marker expressions (ADT, antibody-derived tags) were measured simultaneously using CITE-seq [46]. In these experiments, the authors used 49 and 13 antibodies, respectively, that recognize cell-surface proteins used to classify different types of immune cells.

Consistent with previous analyses of CITE-seq data [40, 26], we used the Seurat R package [7] to preprocess RNA and ADT counts. We normalized ADT expression using centered log-ratio (CLR) transformation and log-transformed RNA counts after adding a pseudocount of 1. After selecting the top 2,000 most variable genes, the expression of each gene was scaled to have mean expression 0 and variance 1, followed by dimensionality reduction to 20 principal components.

Doublets in the PBMC data set were removed using the same cell hashing-based approach with identical parameters as in [26]. Similar to the analysis in [46], a putative cluster of doublets

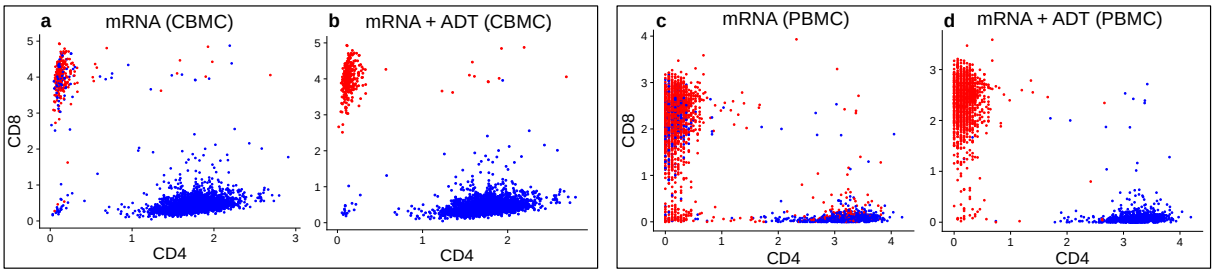

Figure 5: Comparison of unimodal and joint clustering by Specter. CBMCs (left box) and PBMCs (right box) with coordinates of protein expression (ADT) along CD4 and CD8 axis. Cells are clustered by Specter into CD4 T cells (blue) and CD8 T cells (red) either based on mRNA expression alone (**a, c**) or jointly from mRNA and surface protein expression (**b, d**). The mixing of CD4 T cells and CD8 T cells in the mRNA based clustering is corrected through the co-association of both modalities by Specter.

coexpressing different RNA and protein lineage marker were removed from further analysis. On the CBMC data we relied on the doublet removal of Seurat performed in a prior analysis [40] of this data set.

We annotate clusters based on differential expression of marker genes (Wilcoxon rank-sum test) for immune cell types listed in Table 1. The analysis of both data sets is documented at `https://github.com/canzarlab/Specter`.

Table 1: Markers used in the annotation of clusters in the CBMC and PBMC data sets. P-values indicate significance of differential expression according to a Wilcoxon rank-sum test between clusters inferred by Specter from the joint analysis of mRNA and surface protein expression.

Cell-type	Data set	Markers
CD8+CD27-	PBMC	*CD8A* (p = 3.1e-15), *CD8B* (p = 4.3e-6), low *CD27* ADT
CD8+CD27+	PBMC	*CD8B* (p = 3.2e-4), high *CD27* ADT
Naive CD4+ T	PBMC	*SELL* [22] (p = 2.6e-9)
CD4+CD27+	PBMC	*IL7R* [13] (p = 9.4e-11), high *CD27* ADT
CD4+CD27-DR+	PBMC	*IL7R* [13] (p = 4.4e-7), *NKG7* [17] (p = 1.2e-3), *GZMA* [17] (p = 2.0e-4)
CD4+CD27-DR-	PBMC	*IL7R* [13] (p = 1.4e-6), low expression of *NKG7* and *GZMA*; low *CD27* ADT.
CD14+ Mono	PBMC	*LYZ* (p = 7.5e-34), *CST3* (p = 1.4e-32)
FCGR3A+ Mono	PBMC	*FCGR3A* (p = 1.0e-9)
Megakaryocytes	PBMC	*PF4* [30] (p = 1.2e-3)
NK	PBMC	*GNLY* [35] (p = 7.7e-21), *NKG7* [49] (p = 1.1e-15)
Dendritic cells	CBMC	*CST3* [24] (p = 4.7e-29), *CD1C* [12, 32] (p = 1.1e-27), and *FCER1A* [24] (p = 1.3e-27)
Megakaryocytes	CBMC	*PF4* [29] (p = 1.6e-25), *PPBP* [39] (p = 5.8e-24)

On both data sets, both Seurat and Specter fail to accurately distinguish naive CD4 T cells and CD8 T cells based on transcriptomic data alone (Figures 5; Figure S3). Many CD4-/CD8+ T cells identified by protein measurements (ADT) in the CBMC data set are wrongly grouped together with CD4 T cells by Seurat and Specter. Similarly, CD4 and CD8 T cells are mixed in the PBMC data set by both methods.

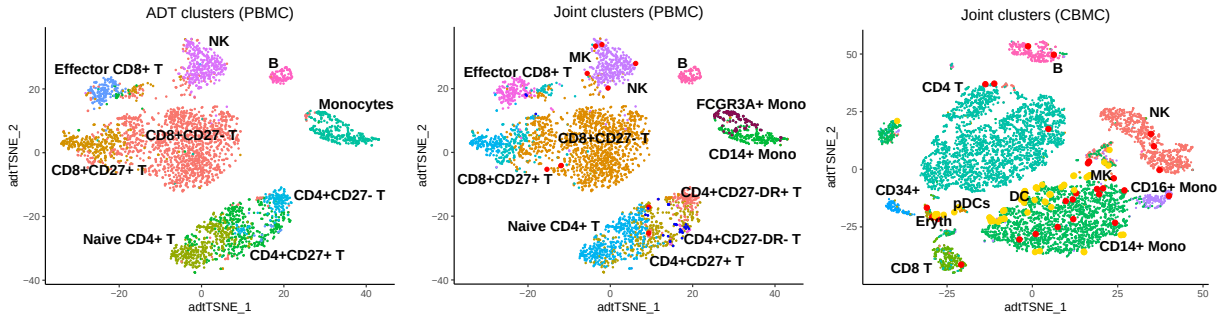

Figure 6: t-SNE visualization of clusters identified by Specter. Clusters of PBM cells were inferred from protein expression (ADT) alone (left) or from combined mRNA and protein expression (middle). In contrast to the joint clustering of both modalities, ADT-based clustering cannot discriminate CD14+ and FCGR3A+ Monocytes, does not detect megakaryocytes (red) and does not allow to discriminate between CD27-DR+ and CD27-DR- subpopulations of CD4+ T cells. The simultaneous clustering of RNA and protein expresssion in CBM cells (right) additionally reveals a rare population of megakaryocytes (red).

On the other hand, dendritic cells and megakaryocytes cannot be identified in the CBMC data set based on protein marker expression, see analysis using Seurat [40]. Similarly, Figure 6 shows that ADT-based clustering by Specter is not able to separate CD14+ from FCGR3A+ Monocytes nor megakaryocytes from other cell types in the PBMC data set. This can be analogously observed in the clustering by Seurat (Figure S3).

We therefore aimed to correct and improve the individual clusterings of RNA and surface marker protein measurements by combining the two distinct species through our clustering ensemble approach. In particular, Specter first produces an identical number of clusterings (here 200) for each modality. It then combines the transcriptome-based clusterings and the protein-based clusterings through a co-association approach (see the "Methods" section).

The joint clustering of RNA and protein expression by Specter profits from both modalities, yet differs from both unimodal analyses: On the PBMC data set, an ARI score of 0.78 comparing multimodal and RNA-based clustering and a score of 0.72 between multimodal and ADT-based clustering indicate complementary aspects of cellular identity utilized in their joint clustering. On the CBMC data set, higher ARI scores of 0.87 and 0.91 between the multimodal clustering and RNA and ADT-based clusterings, respectively, reflect a higher agreement between the two modalities.

More specifically, the joint clustering of RNA and protein expression of CBM and PBM cells allows Specter to more accurately separate CD4 T cells and CD8 T cells compared to a simple transcriptome-based clustering (Figure 5). In contrast to ADT expression based clustering of PBM cells, the joint clustering of RNA and surface protein expression by Specter correctly identifies megakaryocytes, CD14+, and FCGR3A+ Monocytes (Table 1 and Figure 6). In addition, only the combined clustering of ADT and RNA allows Specter to discriminate between CD27- DR+ and CD27- DR- subpopulations of CD4+ memory T cells. In contrast to the clustering of protein data of CBM cells, Specter also correctly detects dendritic cells and megakaryocytes based on the markers listed in Table 1 (see Figure 6).

We compare the joint clustering by Specter to the results of CiteFuse (v0.99.10) [26], a method that was recently proposed specifically for the computational analysis of single cell multimodal profiling data. As proposed initially for the combination of (bulk) genome-wide measurements across, e.g., patients [51], CiteFuse applies the similarity network fusion algorithm to combine

RNA and ADT expression of single cells and then clusters the fused similarity matrix using spectral clustering. We ran CiteFuse as originally described in [26] including the removal of doublets and the (internal) selection of highly variable genes.

Overall, the clusters of CBM and PBM cells as computed by Specter and CiteFuse are highly similar, as indicated by a high ARI score of 0.94 and 0.86 for the two data sets (Figures S4 and S5). In both data sets, however, only Specter is able to identify a rare population of megakaryocytes (Table 1). Furthermore, in contrast to the analysis performed in [26], CiteFuse was not able to discriminate beween CD27- DR+ and CD27- DR- subpopulations of CD4+ memory T cells in the PBMC data set, neither when using identical parameters as in [26] nor when applying more conservative parameters in the doublet removal (parameters taken from CiteFuse tutorial [31]) (Figure S6). The authors of [26] attribute this discrepancy to a different selection of highly variable genes applied in an earlier version of the software used to produce the results in [26].

The major advantage of Specter over CiteFuse, however, is its speed and scalability. CiteFuse requires 15 minutes and nearly 2 hours to jointly cluster the 3,880 PBM cells and 7,895 CBM cells (after doublet removal), respectively, and is thus not expected to scale well on much larger data sets due to the computational expensive fusion of networks. In contrast, Specter returns a high resolution clustering of the two data sets in just 20 and 50 seconds, respectively.

2.4 Scalability

Here, we demonstrate scalability of Specter to large single-cell data sets. Table 2 gives the CPU times in minutes on the four largest data sets used in this study. Seurat was run with a call to the more efficient SCANPY implementation of the Louvain clustering algorithm. Despite solving an optimization problem and employing an highly accurate cluster ensemble scheme, Specter is 3-5 times faster than Seurat whose underlying Louvain algorithm is known for its efficiency. Thanks to the hybrid landmark selection, Specter is even faster than the originally developed landmark-based spectral clustering algorithm. Specter's running time is mostly identical with the running time of dropClust, which was designed for ultra-large scRNA-seq data set. Only on the largest data set comprising around 2 million cells dropClust was with just 12 minutes of computation time the fastest method, compared to 26 minutes it took Specter to cluster the 2 million cells. In contrast to Specter, however, dropClust considers only around 1% of the data (20,000 cells) and its simplified model comes at the cost of a substantial loss in accuracy (see Figure 2).

Table 2: Comparison of running times. Running times of Specter, Seurat, the original Landmark-based Spectral Clustering (LSC) algorithm, dropClust, and the geometric sketching based Louvain clustering are reported in minutes (rounded) on the 4 largest data sets used in this study.

Data set	# Cells	Specter	LSC	Seurat	dropClust	Geometric sketching
cord blood	254,941	1	8	4	1	3
CNS	464,713	2	10	11	2	7
saunders	665,385	3	12	18	3	19
trapnell	2,026,641	26	56	79	12	400

3 Conclusion

We have introduced Specter, a novel method that identifies transcriptionally distinct sets of cells with substantially higher accuracy than existing methods. We adopt and extend algorithmic inno-

vations from spectral clustering, to make this powerful methodology accessible to the analysis of modern single-cell RNA-seq data sets. We have demonstrated the superior performance of Specter across a comprehensive set of public scRNA-seq data sets and illustrate that an overall higher accuracy may also implicate an increased sensitivity towards rare cell types. At the same time, its linear time complexity and practical efficiency makes Specter particularly well-suited for the analysis of large scRNA-seq data sets.

Furthermore, we have illustrated how the flexibility of its underlying optimization model allows Specter to harness multimodal omics measurements of single cells to resolve subtle transcriptomic differences between subpopulations of cells. The application of our cluster ensemble scheme to the joint analysis of multimodal CITE-seq data sets yielded a slightly more fine-grained distinction of cell (sub-)populations compared to the recently proposed multimodal clustering method CiteFuse. More importantly, in contrast to CiteFuse whose running time increased ≈ 8 fold after doubling the number of cells, Specter will scale well to much larger data sets produced by droplet-based approaches that can measure multiple modalities of up to millions of cells together. Taken together, we believe that Specter will be useful in transforming massive amounts of (multiple) measurements of molecular information in individual cells to a better understanding of cellular identity and function in health and disease.

4 Methods

Spectral clustering uses eigenvectors of a matrix derived from the distance between points (here cells) as a low-dimensional representation of the original data, which it then partitions using a method such as k-means. More precisely, given n data points $X = \{x_1, x_2, \ldots, x_n\}$ and a similarity matrix (affinity matrix) $W = (w_{ij})_{n\times n}$, where w_{ij} measures the similarity between points x_i and x_j, the graph Laplacian is defined as $L = D - W$ or $L = I - D^{-1/2}WD^{-1/2}$ in case of a (symmetric) normalized Laplacian. Here, D is a diagonal matrix whose entries are column sums (equivalently row sums) of W. Spectral clustering then uses the top k eigenvectors of L to partition the data into k clusters using the k-means algorithm.

In the following description of our algorithm we assume a given number of clusters k. In Specter we determine the number of clusters based on the Silhouette index [38], which performed particularly well in recent benchmark studies [2, 11].

4.1 Landmark-based spectral clustering of single cells

Several methods have been proposed to accelerate the spectral clustering algorithm [18, 44, 8]. In particular, Landmark-based Spectral Clustering (LSC) has been shown to perform well in terms of efficiency and effectiveness compared to state-of-the-art methods across a large number of data sets [8]. In short, LSC picks a small set of p representative data points, i.e. the landmarks, which it then uses to create a sparse affinity matrix Z between all data points and landmark points via the Gaussian kernel $K(x, y) = \exp(-\|x - y\|^2/2\sigma^2)$. From this landmark-based representation of the *complete* data it computes the Laplacian matrix $L = \hat{Z}\hat{Z}^T$, where $\hat{Z} = D^{-1/2}Z$ and D is the diagonal matrix whose (i, i)-entry equals the sum of the ith row of Z. Then, this graph Laplacian L admits a fast eigendecomposion in time $O(n)$ as oppose to $O(n^3)$ in the general case.

Here, we tailor the idea of landmark-based spectral clustering to the characteristics and scale of modern scRNAseq data sets. In particular, the choice of bandwidth σ used in the (Gaussian) kernel to smooth the measure of similarity between pairs of data points heavily depends on the type of data and can have a strong impact on the final clustering. In the original approach, parameter σ is set to the average Euclidean distance between data points and their K-nearest landmarks, i.e. to the

average value of all elements in matrix Z. We empirically find that setting $\sigma = \gamma \times \text{mean}(\max(Z))$, where $\max(Z)$ denotes a vector of maximum values for each row in Z and γ a carefully chosen parameter, is able to better capture the transcriptional similarity between single cells and yields more accurate clusterings of cells.

Furthermore, we pair the theoretical reduction in time complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(n)$ with a practical speed-up of the LSC algorithm by applying a hybrid strategy when selecting the landmarks. The choice of representative data points, here single cells, plays a crucial role in the quality of the final clustering. Random selection or k-means clustering were originally proposed as procedures for picking landmarks [8]. Random selection of representative cells is very efficient but often yields random sets of cells that do not represent the full data well and thus lead to poor clustering results. k-means, on the other hand, better takes into account the structure of the data when selecting landmark cells but its higher computational cost makes it impractical for large scRNA-seq data sets where it accounts for around 90% of the overall running time in our experiments. Our hybrid strategy seeks to balance the efficiency of random sampling and the accuracy of k-means based landmark selection. It first picks a set of p' candidate landmarks uniformly at random with $p' \ll n$, from which it subsequently selects $p < p'$ final landmark cells using the k-means algorithm. Note that despite the initial random sampling, the *full* data are represented by the final set of landmarks.

Finally, for data sets that contain a small number of clusters, we adjust the spectral embedding based on which the original data is clustered using k-means in the last step of spectral clustering. For a small number of clusters (e.g. $k \leq 4$), the top k eigenvectors used in the original approach typically do not contain enough information to represent the full data well. In this case, we therefore use the top $k + 2$ eigenvectors to compute the spectral embedding.

4.2 Clustering ensembles across parameters and modalities

Different data types require a different choice of parameter values and there is no general rule how to select the best one. To address this issue, we employ consensus clustering, also known in literature as cluster ensembles [47], in the same way as ensemble learning is used in supervised learning. In particular, we generate a series of component clusterings by varying the number of selected landmarks p and the kernel bandwidth. We randomly select parameter γ which controls the bandwidth of the Gaussian kernel from interval [0.1, 0.2] and choose p from interval

$$[\min(8k\log(k), \lceil n/3 \rceil), \min(10k\log(k), \lceil n/2 \rceil)]\,.$$

All clusterings produced by the different runs of our tailored LSC algorithm are then summarized in a co-association matrix H [19] in which entry (i, j) counts the number of runs that placed cells i and j in the same cluster. We compute the final clustering through a hierarchical clustering of matrix H. Our LSC-based consensus clustering approach is summarized in Algorithm 1.

Different parameter choices (e.g. kernel bandwidths) provide different interpretations of the same data. In the same way as clustering ensembles can help unifying these different views on a single modality, they can help reconcile the measurements of multiple modalities, such as transcriptome and proteome, of the same cell. More specifically, Specter produces an identical number of clusterings for each modality in step 2 of Algorithm 1 which it then combines through the same co-association approach (steps 3 and 4).

Time complexity The time complexity of the tailored LSC algorithm is $O(n)$, and single linkage hierarchical clustering requires $O(n^2)$ time, yielding an overall complexity of $O(n^2)$ for Algorithm 1, assuming k is small enough to be considered a constant.

Algorithm 1: LSC ensemble

1 **Input:** Cells $x_1, \ldots, x_n$; number of clusters k
2 Run the tailored LSC algorithm for different kernel bandwidths and varying numbers of landmarks.
3 Summarize all clusterings in a co-association matrix H.
4 Apply the single linkage hierarchical clustering algorithm to H to obtain the final k clusters.

4.3 Selective sampling-based clustering ensemble

With a running time that scales quadratically with the number of cells, the application of Algorithm 1 to large-scale scRNA-seq data sets becomes infeasible. We therefore apply step 3 of our clustering ensemble approach (Algorithm 1) to a carefully selected sketch of the data. Note, however, that the co-association matrix H built in step 3 of the algorithm is based on cluster labels that were learned from the *full* data in step 2 using our tailored LSC algorithm. In addition, we propose a simple sampling technique that uses all clusterings computed in step 2 to guide the selection of cells.

Selective Sampling Sampling cells uniformly at random is naturally fast, since the decision to include a given cell into a sketch does not depend on any other cell. At the same time, these independent decisions ignore the global structure of the data such as the abundance of different cell types and may thus lead to a loss of rare cell types [23]. We therefore propose a sampling approach that utilizes the clusterings of the data computed in step 2 of Algorithm 1 to inform the (fast) selection of cells. More specifically, Let $\Pi = \{\pi_1, \pi_2, \ldots, \pi_m\}$, where $\pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{ik})$ is the ith clustering returned in step 2 of Algorithm 1, $i = 1, 2, \ldots, m$. We select a sketch S of size $[k\sqrt{n}]$ that contains roughly the same number of cells in each cluster π_{ij}, for all i and j. This selective sampling procedure iterates through all clusters contained in all clusterings from which it randomly picks a cell not already contained in the sketch, until the size of the sketch reaches $[k\sqrt{n}]$ (see Algorithm 2).

Algorithm 2: Selective sampling

Input: Component clusterings $\Pi = \{\pi_1, \pi_2, \ldots, \pi_m\}$, number of clusters k.
Initialization: $S = \emptyset$.
while $|S| < k\sqrt{n}$ **do**
  for $i = 1$ *to* m **do**
    for $j = 1$ *to* k **do**
      Randomly select a cell s from $\pi_{ij} \backslash S$
      $S = S \cup \{s\}$
    end
  end
end

Inference Given a selectively sampled sketch S, we apply steps 3 and 4 in Algorithm 1 to cells in S, using labels obtained from the full data in step 2. That is, we construct a co-association matrix whose entries count the number of times the two corresponding cells in S were placed in the same cluster by a run of the LSC algorithm in step 2. From this matrix, we compute

a consensus clustering of S using hierarchical clustering and finally transfer cluster labels to the remaining cells using supervised k-nearest neighbors classification. That is, we assign each cell not in S to the cluster that the majority of its k nearest neighbors were placed in by the preceding consensus clustering of S. Our selective sampling-based cluster ensemble approach is summarized in Algorithm 3.

Algorithm 3: Selective sampling-based clustering ensemble

1 **Input:** Cells $x_1, \ldots, x_n$; number of clusters k
2 Run the tailored LSC algorithm for a varying number of landmarks and different kernel bandwidths. Let $\Pi = \{\pi_1, \pi_2, \ldots, \pi_m\}$ be the set of m clusterings.
3 Run selective sampling (Algorithm 2) on Π to obtain a sketch S of size $|S| = [k\sqrt{n}]$.
4 Summarize all clusterings of cells in S computed in step 2 in a co-association matrix H^S.
5 Apply single linkage hierarchical clustering to H^S to obtain k clusters for S.
6 Transfer labels to full data using k-nearest neighbors classification

Time complexity Landmark-based spectral clustering performed in step 2 of Algorithm 3 takes $O(n)$, see above. Since we selectively sample a sketch of size $|S| = O(\sqrt{n})$ in step 3, the complexity of steps 4 and 5 now reduces to $O(n)$. Together with the k-NN classification that runs in $O(n)$ in step 6, our selective sampling based cluster ensemble scheme scales linearly with the number of cells n.

5 Availability of data and materials

The Specter software is available at `https://github.com/canzarlab/Specter` under the open source MIT license. The Specter repository also includes a step-by-step documentation of the analysis of the PBMC and CBMC CITE-seq data sets [46, 33] as described in this study.

The original publication of data sets used in the "Specter is more accurate than competing methods" subsection of the "Results" section are listed in Table 1. Data sets smaller than 15,000 cells were downloaded at `https://hemberg-lab.github.io/scRNA.seq.datasets`, the remaining 3 largest data sets at `http://mousebrain.org` (*CNS*), `http://dropviz.org` (*saunders*), and `https://oncoscape.v3.sttrcancer.org/atlas.gs.washington.edu.mouse.rna/downloads` (*trapnell*). The umbilical cord blood cell data [23] were downloaded at `http://cb.csail.mit.edu/cb/geosketch`.

6 Author's contributions

V.H.D. and S.C. conceived the algorithm. V.H.D. implemented the software and performed the computational experiments. All authors designed the experiments and wrote the paper.

7 Acknowledgments

We thank Khaled Elbassioni for helpful discussion on the algorithm and Hani Jieun Kim for comments on CiteFuse results.

8 Funding

V.H.D. was supported by a Deutsche Forschungsgemeinschaft fellowship through the Graduate School of Quantitative Biosciences Munich.

9 Competing interests

The authors declare that they have no competing interests.

References

[1] Specter software. `https://github.com/canzarlab/Specter`. Accessed 10 May 2020.

[2] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jess M. Prez, and Iigo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243 – 256, 2013.

[3] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, Douglas A Melton, and Itai Yanai. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Systems*, 3(4):346–360.e4, Oct 2016.

[4] Mayank Bawa, Tyson Condie, and Prasanna Ganesan. Lsh forest: Self-tuning indexes for similarity search. In *Proceedings of the 14th International Conference on World Wide Web*, WWW 05, page 651660, New York, NY, USA, 2005. Association for Computing Machinery.

[5] Fernando H. Biase, Xiaoyi Cao, and Sheng Zhong. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell rna sequencing. *Genome Research*, 24(11):1787–1796, 2014.

[6] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[7] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnoly*, 36(5):411–420, 2018.

[8] Deng Cai and Xinlei Chen. Large scale spectral clustering with landmark-based representation. *AAAI*, pages 313 – 318, 2011.

[9] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, Cole Trapnell, and Jay Shendure. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 34:65–31, February 2019.

[10] Renchao Chen, Xiaoji Wu, Lan Jiang, and Yi Zhang. Single-cell rna-seq reveals hypothalamic cell diversity. *Cell Reports*, 18(13):3227–3241, Mar 2017.

[11] H. Chouikhi, M. Charrad, and N. Ghazzali. A comparison study of clustering validity indices. In *2015 Global Summit on Computer Information Technology (GSCIT)*, pages 1–4, 2015.

[12] Matthew Collin, Naomi McGovern, and Muzlifah Haniffa. Human dendritic cell subsets. *Immunology*, 140(1):22–30, 2013.

[13] Sara L. Colpitts, Nicole M. Dalton, and Phillip Scott. Il-7 receptor expression provides the potential for long-term survival of both cd62l high central memory t cells and th1 effector cells during leishmania major infection. *Journal of Immunology*, 182(9):5702–5711, 2009.

[14] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167), 2014.

[15] H. E. Driver and A. L. Kroeber. Quantitative expression of cultural relationships. 1932.

[16] Angelo Duò, Mark D. Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7:1141, 2018.

[17] Chamith Y. Fonseka, Deepak A. Rao, Nikola C. Teslovich, Ilya Korsunsky, and et al. Mixed-effects association of single cells identifies an expanded effector cd4+ t cell subset in rheumatoid arthritis. *Science Translational Medicine*, 10(463), 2018.

[18] Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik. Spectral grouping using the nystrom method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214 – 225, 2004.

[19] Ana L.N. Fred and Anil K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850, 2005.

[20] Mubeen Goolam, Antonio Scialdone, Sarah J.L. Graham, Iain C. Macaulay, Agnieszka Jedrusik, Anna Hupalowska, Thierry Voet, and John C. Marioni. Heterogeneity in oct4 and sox2 targets biases cell fate in 4-cell mouse embryos. *Cell*, 165(1):61–74, 2016.

[21] Dominic Grün, Mauro J. Muraro, Jean-Charles Boisset, Kay Wiebrands, Anna Lyubimova, Gitanjali Dharmadhikari, Maaike van den Born, Johan van Es, Erik Jansen, Hans Clevers, Eelco J.P. de Koning, and Alexander van Oudenaarden. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*, 19(2):266–277, 2016.

[22] W Nicholas Haining, Jill Angelosanto, Kathleen Brosnahan, Kenneth Ross, Cynthia Hahn, Kate Russell, Linda Drury, Stephanie Norton, Lee Nadler, and Kimberly Stegmaier. High-throughput gene expression profiling of memory differentiation in primary human t cells. *BMC Immunology*, 9(44), 2008.

[23] Brian Hie, Hyunghoon Cho, Benjamin DeMeo, Bryan Bryson, and Bonnie Berger. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell Systems*, 8(6):483 – 493.e7, 2019.

[24] Tomas Hruz, Oliver Laule, Gabor Szabo, Frans Wessendorp, Stefan Bleuler, Lukas Oertle, Peter Widmayer, Wilhelm Gruissem, and Philip Zimmermann. Genevestigator v3: A reference expression database for the meta-analysis of transcriptomes. *Advances in Bioinformatics*, 2008(420747):1–5, 2008.

[25] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[26] Hani Jieun Kim, Yingxin Lin, Thomas A Geddes, Jean Yee Hwa Yang, and Pengyi Yang. CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics*, 04 2020. btaa282.

[27] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, and Martin Hemberg. Sc3: consensus clustering of single-cell rna-seq data. *Nature Methods*, 14(5):483–486, 2017.

[28] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015.

[29] Michele P Lambert, Ronghua Meng, Dawn Harper, Liqing Xiao, Michael S. Marks, and Mortimer Poncz. Megakaryocytes exchange significant levels of their alpha-granular pf4 with their environment. *Blood*, 124(21):1432–1432, 2014.

[30] Michele P. Lambert, Ronghua Meng, Liqing Xiao, Dawn C. Harper, Michael S. Marks, Anna M. Kowalska, and Mortimer Poncz1. Intramedullary megakaryocytes internalize released platelet factor 4 (pf4) and store it in alpha granules. *Journal of Thrombosis and Haemostasis*, 13(10):1888–1899, 2016.

[31] Yingxin Lin and Hani Jieun Kim. Citefuse: getting started, 2020. `https://sydneybiox.github.io/CiteFuse/articles/CiteFuse.html`. Accessed 15 March 2020.

[32] Miriam Merad, Priyanka Sathe, Julie Helft, Jennifer Miller, and Arthur Mortha. The dendritic cell lineage: ontogeny and function of dendritic cells and their subsets in the steady state and the inflamed setting. *Annual Review of Immunology*, 31:563–604, 2013.

[33] Eleni P. Mimitou, Anthony Cheng, Antonino Montalbano, Stephanie Hao, Marlon Stoeckius, Mateusz Legut, Timothy Roush, Alberto Herrera, Efthymia Papalexi, Zhengqing Ouyang, Rahul Satija, Neville E. Sanjana, Sergei B. Koralov, and Peter Smibert. Multiplexed detection of proteins, transcriptomes, clonotypes and crispr perturbations in single cells. *Nature Methods*, 16(5):409–412, 2019.

[34] Mauro J Muraro, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon van Gurp, Marten A Engelse, Francoise Carlotti, Eelco J P. de Koning, and Alexander van Oudenaarden. A single-cell transcriptome atlas of the human pancreas. *Cell Systems*, 3(4):385–394.e3, Oct 2016.

[35] Kazuyuki Ogawa, Yasushi Takamori, Kunou Suzuki, Masayuki Nagasawa, Shoichi Takano, Yoshihito Kasahara, Yoshiko Nakamura, Shigemi Kondo, Kazuo Sugamura, and Kinya Nagata. Granulysin in human serum as a marker of cellmediated immunity. *European Journal of Immunology*, 33(7):1925–1933, 2003.

[36] Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, Brian V. Nahed, William T. Curry, Robert L. Martuza, David N. Louis, Orit Rozenblatt-Rosen, Mario L. Suvá, Aviv Regev, and Bradley E. Bernstein. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[37] Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, Naveen Ramalingam, Gang

Sun, Myo Thu, Michael Norris, Ronald Lebofsky, Dominique Toppani, Darnell W Kemp II, Michael Wong, Barry Clerkson, Brittnee N Jones, Shiquan Wu, Lawrence Knutsson, Beatriz Alvarado, Jing Wang, Lesley S Weaver, Andrew P May, Robert C Jones, Marc A Unger, Arnold R Kriegstein, and Jay A A West. Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology*, 32(10):1053–1058, 2014.

[38] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[39] Kazuki Sakurai, Tohru Fujiwara, Shin Hasegawa, Yoko Okitsu, Noriko Fukuhara, Yasushi Onishi, Minami Yamada-Fujiwara, Ryo Ichinohasama, and Hideo Harigae. Inhibition of human primary megakaryocyte differentiation by anagrelide: a gene expression profiling analysis. *International Journal of Hematology*, 104(2):190–199, 2016.

[40] Rahul Satija. Using seurat with multi-modal data, 2019. `https://satijalab.org/seurat/v3.1/multimodal_vignette.html`. Accessed 15 December 2019.

[41] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015.

[42] Arpiar Saunders, Evan Z. Macosko, Alec Wysoker, Melissa Goldman, Fenna M. Krienen, Heather de Rivera, Elizabeth Bien, Matthew Baum, Laura Bortolin, Shuyu Wang, Aleksandrina Goeva, James Nemesh, Nolan Kamitaki, Sara Brumbaugh, David Kulp, and Steven A. McCarroll. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell*, 174(4):1015–1030.e16, Aug 2018.

[43] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[44] Hiroyuki Shinnou and Minoru Sasaki. Spectral clustering for a large data set by reducing the similarity matrix size. *Proceedings of the Sixth International Language Resources and Evaluation*, 2008.

[45] Debajyoti Sinha, Akhilesh Kumar, Himanshu Kumar, Sanghamitra Bandyopadhyay, and Debarka Sengupta. dropClust: efficient clustering of ultra-large scRNA-seq data. *Nucleic Acids Research*, 46(6):e36–e36, 2018.

[46] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017.

[47] Alexander Strehl and Joydeep Ghosh. Cluster ensembles a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583 – 617, 2003.

[48] Luyi Tian, Xueyi Dong, Saskia Freytag, Kim-Anh L Cao, Shian Su, Abolfazl JalalAbadi, Daniela Amann-Zalcenstein, Tom S. Weber, Azadeh Seidi, Jafar S. Jabbari, Shalin H. Naik, and Matthew E. Ritchie. Benchmarking single cell rna-sequencing analysis pipelines using mixture control experiments. *Nature Methods*, 16(6):479 – 487, 2019.

[49] Martin A. Turman, Toshio Yabe, Cynthia McSherry, Fritz H. Bach, and Jeffrey P. Houchins. Characterization of a novel gene (nkg7) on human chromosome 19 that is expressed in natural killer cells and t cells. *Human Immunology*, 36(1):34–40, 1993.

[50] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[51] Bo Wang, Aziz M. Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, Mar 2014.

[52] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, Feb 2018.

[53] Yurong Xin, Jinrang Kim, Haruka Okamoto, Min Ni, Yi Wei, Christina Adler, Andrew J Murphy, George D Yancopoulos, Calvin Lin, and Jesper Gromada. Rna sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metabolism*, 24(4):608–615, Oct 2016.

[54] Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job van der Zwan, Martin Häring, Emelie Braun, Lars E. Borm, Gioele La Manno, Simone Codeluppi, Alessandro Furlan, Kawai Lee, Nathan Skene, Kenneth D. Harris, Jens Hjerling-Leffler, Ernest Arenas, Patrik Ernfors, Ulrika Marklund, and Sten Linnarsson. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014.e22, Aug 2018.

[55] Amit Zeisel, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

[56] Chenxu Zhu, Sebastian Preissl, and Bing Ren. Single-cell multimodal omics: the power of many. *Nature Publishing Group*, 17(1):11–14, 2020.

[57] J. Zubin. A technique for measuring like-mindedness. *The Journal of Abnormal and Social Psychology*, 33(4):508–516, 1938.

10 Supplementary Tables and Figures

Table S1: Overview of the data sets used in this study. Names listed in the left-most column are used throughout the text. A line separates the three largest data sets.

Data set	# Cells	# Populations	Description	Reference
grun	1502	2	mouse stem cells	[21]
xin	1600	8	human islet cells	[53]
baron	1886	13	human and mouse pancreas	[3]
biase	56	4	mouse embryo devel	[5]
deng	268	6	mouse embryo devel	[14]
goolam	114	5	mouse embryo	[20]
muraro	2126	10	human pancreas	[34]
patel	430	5	human glioblastoma	[36]
pollen	301	11	human developing cortex	[37]
klein	2717	4	mouse embryo stem cells	[28]
zeisel	3005	9	mouse cortex and hippocampus	[55]
chen	14,437	45	mouse brain	[10]
CNS	465,281	7	mouse central nervous system	[54]
saunders	665,858	11	adult mouse brain	[42]
trapnell	2,058,652	38	mouse organogenesis cell atlas	[9]

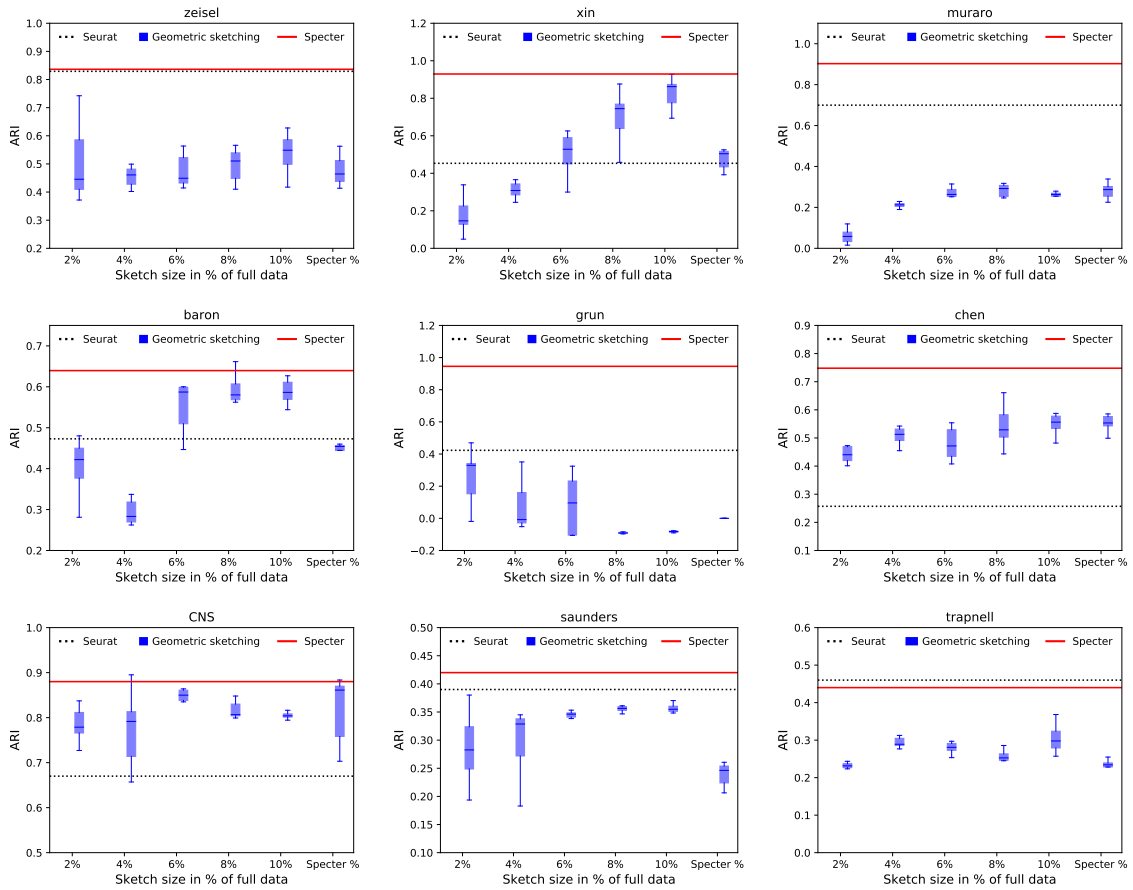

Figure S1: Accuracy (in ARI) of geometric sketching based Louvain clustering for varying sketch sizes. For each sketch size, the results of 10 random trials are shown. "Specter %" uses the same number of cells in the geometric sketch as Specter uses landmarks or cells in the selective sampling step (see the "Methods" section), whichever one is larger. Across all sketch sizes, the landmark-based clustering by Specter more accurately detects cell types than clustering a random geometric sketch of the data. Note that only datasets of more than 1,500 cells are shown.

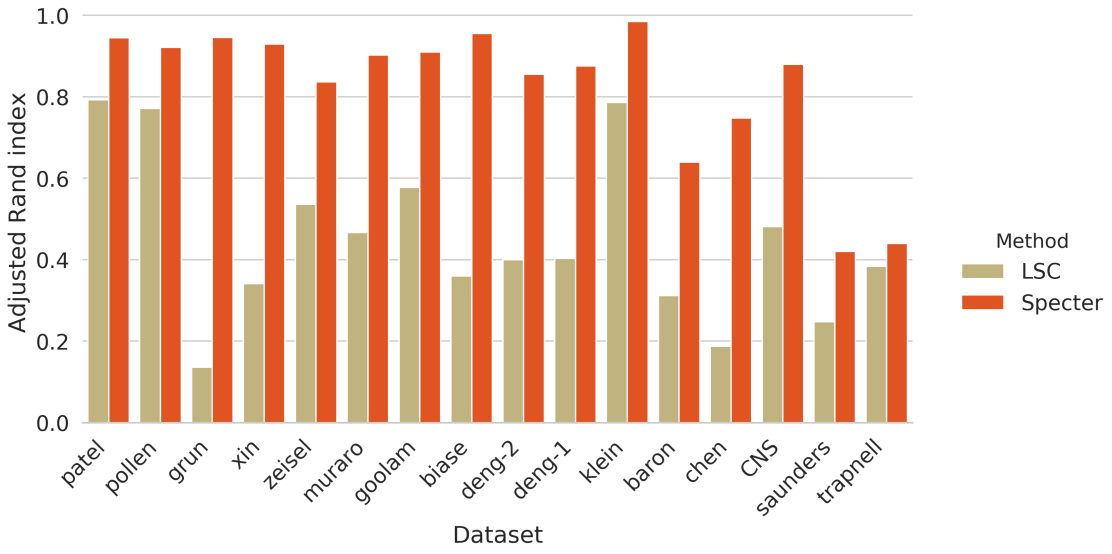

Figure S2: Clustering performance of Specter and the original landmark-based spectral clustering (LSC) algorithm. The Adjusted Rand index is reported for the two methods on the same 15 scRNA-seq data sets as in Figure 2.

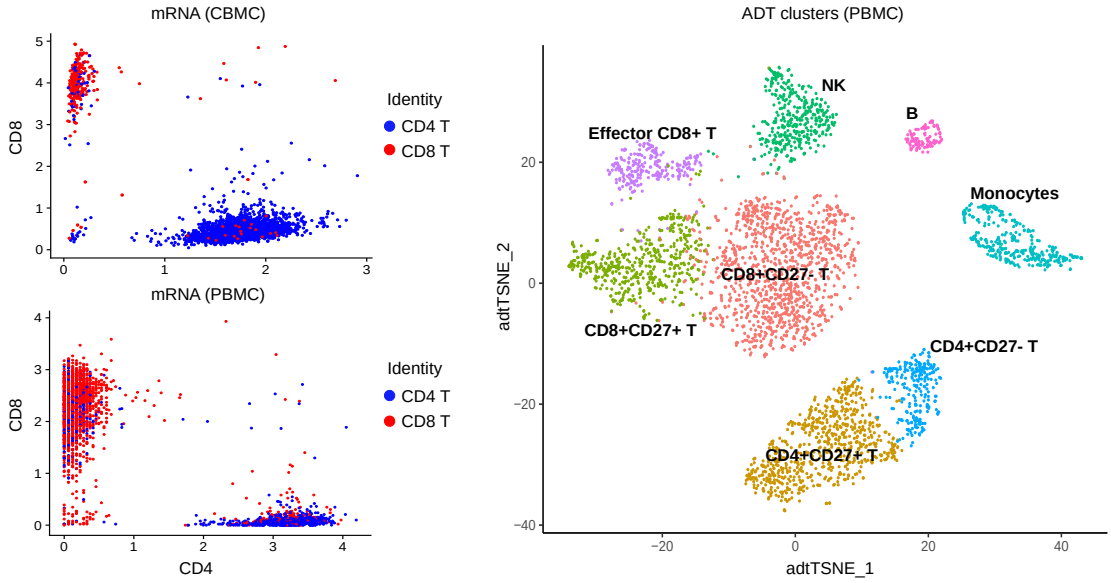

Figure S3: Seurat unimodal clustering. *left:* CBMCs (top) and PBMCs (bottom) with coordinates of protein expression (ADT) along CD4 and CD8 axis. Colors denote clusters computed by Seurat based on mRNA expression which contain a mix of CD4 T cells and CD8 T cells. *right:* t-SNE visualization of clusters identified by Seurat from protein expression (ADT) of PBM cells. CD14+ and FCGR3A+ monocytes cannot be discriminated (compare Figure 6), megakaryocytes are not detected.

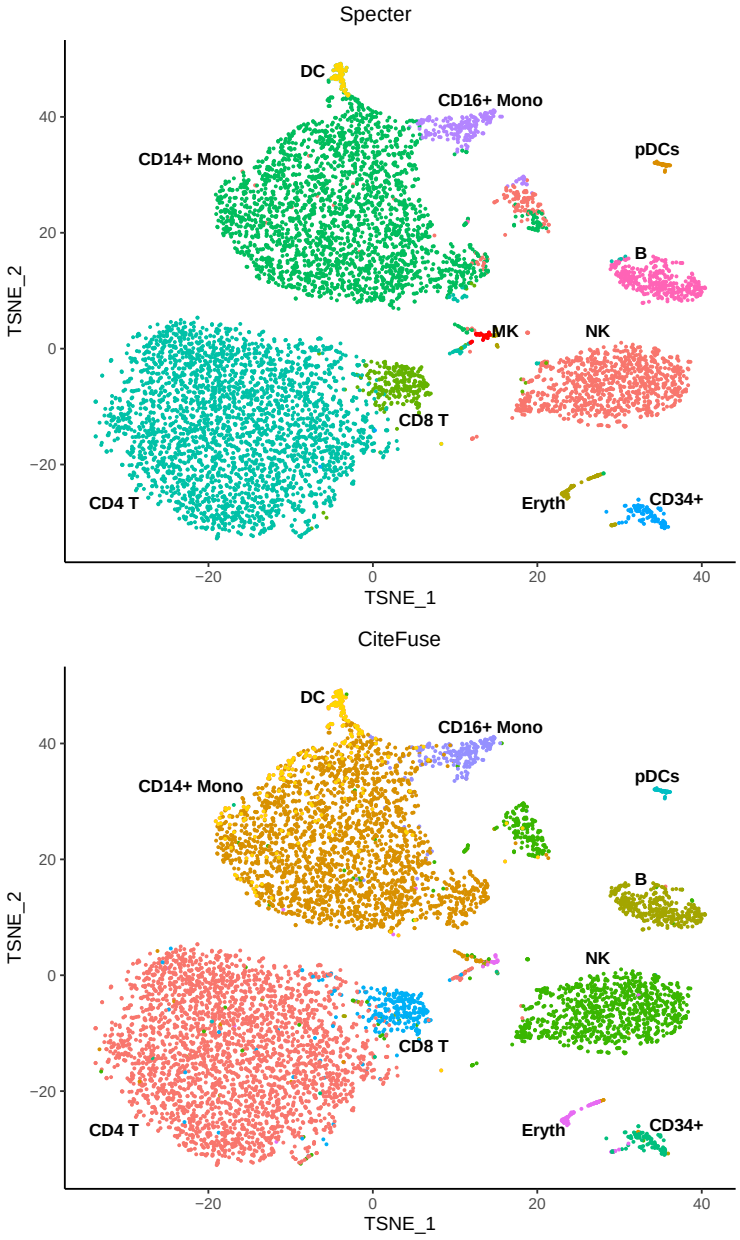

Figure S4: Comparison of multimodal clusterings of CBM cells as computed by Specter (top) and CiteFuse (bottom). Despite an overall high agreement between the two clusterings (ARI 0.94), only Specter detects a rare population of megakaryocytes (red).

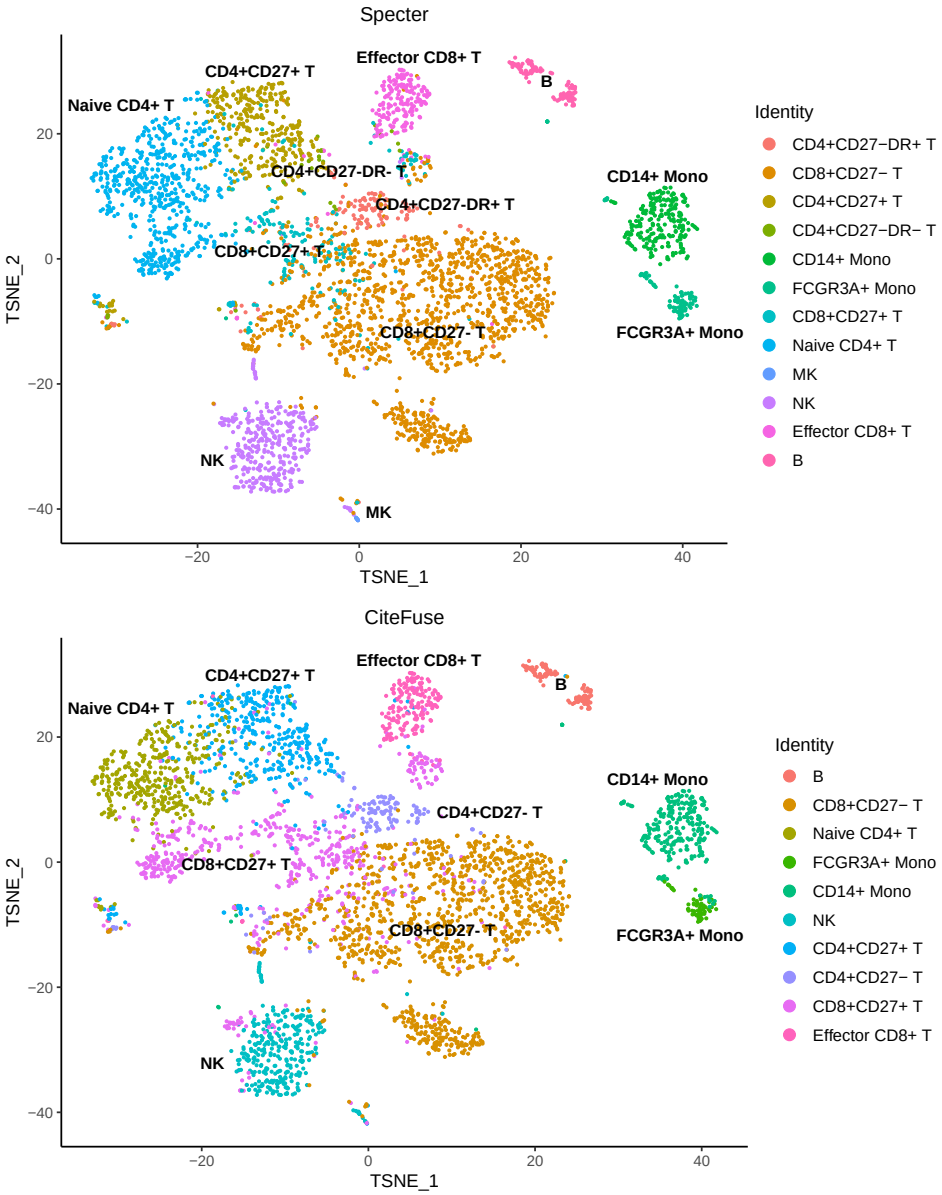

Figure S5: Comparison of multimodal clusterings of PBM cells as computed by Specter (top) and CiteFuse (bottom). Despite an overall high agreement between the two clusterings (ARI 0.86), only Specter detects a rare population of megakaryocytes and can discriminate between CD27-DR+ and CD27-DR- subpopulations of CD4+ memory T cells.

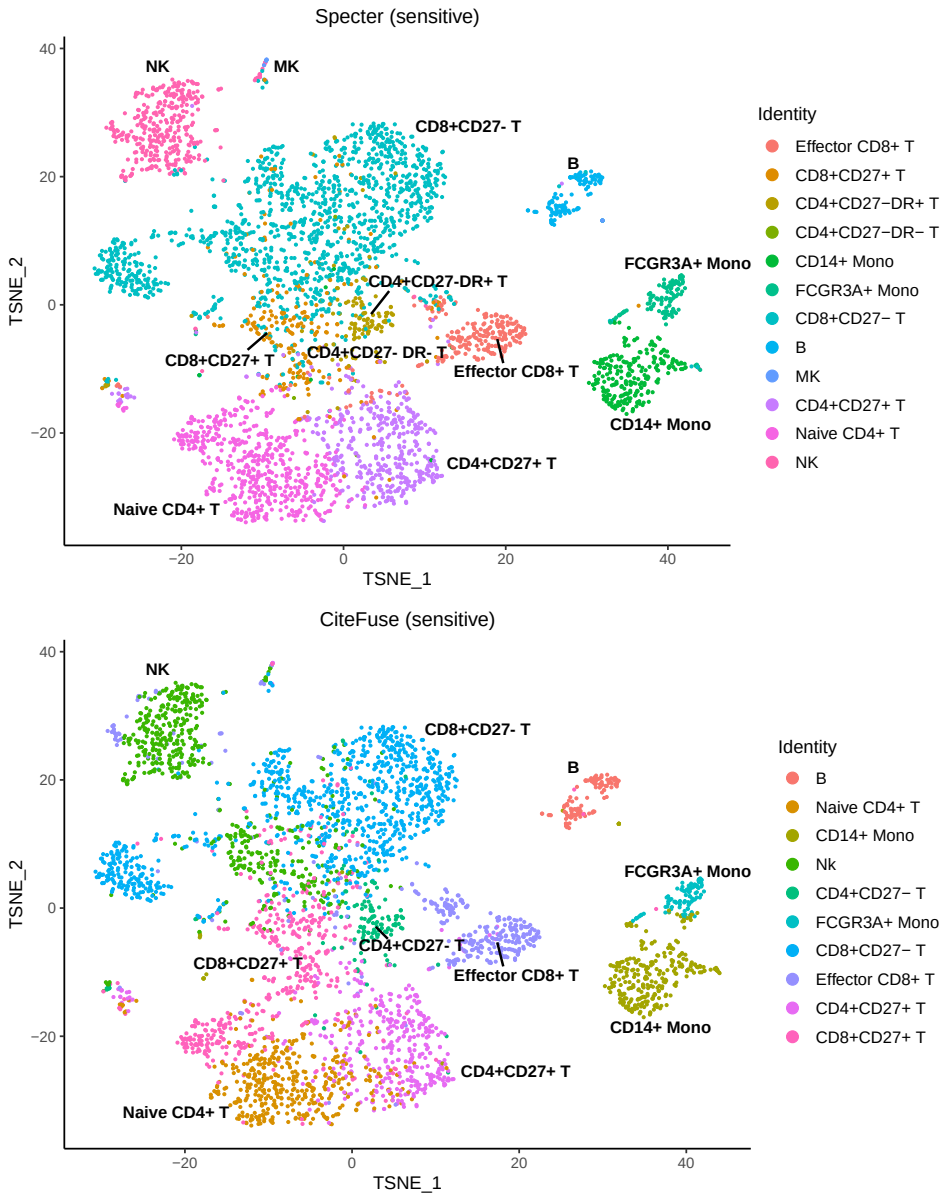

Figure S6: Comparison of multimodal clusterings of PBM cells. Here, Specter (top) and CiteFuse (bottom) use slightly more conservative parameters in the doublet removal (eps = 190, minPts = 10). Again, only Specter is able to discriminate between CD27-DR+ and CD27-DR- subpopulations of CD4+ memory T cells and detects a rare population of megakaryocytes.