Slov Vet Res
DOI 10.26873/SVR-1920-2024
UDC 601.4:577.21:577.27
Pages: 245–61

Original Research Article

Comparative Analysis of Reference-Based Cell Type Mapping and Manual Annotation in Single Cell RNA Sequencing Analysis

Key words

single-cell transcriptomics;
peripheral blood mononuclear cells;
reference mapping;
cell-type annotation;
immune system

Larisa Goričan[1,†], Boris Gole[1,†], Gregor Jezernik[1], Gloria Krajnc[1,3], Uroš Potočnik[1,2,3], Mario Gorenjak[1]*

[1]Centre for Human Genetics and Pharmacogenomics, Faculty of Medicine, University of Maribor, Taborska ulica 8, SI-2000 Maribor, [2]Laboratory for Biochemistry, Molecular Biology and Genomics, Faculty of Chemistry and Chemical Engineering, University of Maribor, Smetanova ulica 17, SI-2000 Maribor, [3]Department for Science and Research, University Medical Centre Maribor, Ljubljanska ulica 5, SI-2000 Maribor, Slovenia, [†]Equal contribution

***Corresponding author:** mario.gorenjak@um.si

Abstract: Single-cell RNA sequencing (scRNA-seq) offers unprecedented insight into cellular diversity in complex tissues like peripheral blood mononuclear cells (PBMC). Furthermore, differential gene expression at a single-cell level can provide a basis for understanding the specialized roles of individual cells and cell types in biological processes and disease mechanisms. Accurate annotation of cell types in scRNA-seq datasets is, however, challenging due to the high complexity of the data. Here, we compare two cell-type annotation strategies applied to PBMCs in scRNA-seq datasets: automated reference-based tool Azimuth and unsupervised Shared Nearest Neighbor (SNN) clustering, followed by manual annotation. Our results highlight the strengths and limitations of the two approaches. Azimuth easily processed large-scale scRNA-seq datasets and reliably identified even relatively rare cell populations. It, however, struggled with cell types outside its reference range. In contrast, unsupervised SNN clustering clearly delineated all the different cell populations in a sample. This makes it well suited for identifying rare or novel cell types, but the method requires time-consuming and bias-prone manual annotation. To minimize the bias, we used rigorous criteria and the collaborative expertise of multiple independent evaluators, which resulted in the manual annotation that was closely related to the automated one. Finally, pseudo-temporal analysis of the major cell types further confirmed the validity of the Azimuth and manual annotations. In conclusion, each annotation method has its merits and downsides. Our research thus highlights the need to combine different clustering and annotation approaches to manage the complexity of scRNA-seq and to improve the reliability and depth of scRNA-seq analyses.

Received: 29 December 2023
Accepted: 23 April 2024

Introduction

Over the past decade, RNA sequencing (RNA-seq) has become an indispensable tool in molecular biology, providing unprecedented insights into the transcriptomic landscape of cells. (1) By deciphering the complexity of human, animal, and plant transcriptomes, this technique has greatly enhanced our understanding of biological processes, disease mechanisms, and therapeutic interventions. (2) However, conventional RNA-seq, which analyses bulk tissue samples, inherently averages the gene expression across many cells and cell types present in the sample, resulting in a loss of resolution at the level of individual cells/cell types. (3) This obscures the understanding of cellular heterogeneity and the roles of rare cell populations in tissue function and disease. (4)

The development of single-cell RNA sequencing (scRNA-seq) has revolutionized the field by providing a lens for exploring the transcriptome at single-cell resolution. (5) The scRNA-seq provides a high-resolution view of tissue cellular diversity. It enables a more detailed understanding of complex biological processes and disease pathogenesis by revealing cell heterogeneity in a given population. Furthermore, scRNA-seq allows for the study of differential gene expression at a single-cell level, which can provide insights into the unique functional roles of individual cells and contribute to a more nuanced understanding of biological processes and disease mechanisms. (6)

Despite its transformative potential, scRNA-seq also introduces unique analytical challenges. Among these, annotation of distinct cell populations in scRNA-seq datasets is a significant hurdle due to the high dimensionality and complexity of single-cell data. (7) To address this, various computational strategies have been developed. Azimuth, a publicly available automated cell-type annotation software (8), employs machine learning algorithms to predict human and murine cell identities based on scRNA-seq data. (9) In parallel, Seurat, a popular R package for scRNA-seq data analysis, offers clustering algorithms that partition single cells into distinct groups based on their transcriptomic profiles, providing an unbiased approach to cell population identification. (10) Manual annotation methods, on the other hand, employ in-depth biological knowledge to assign cell identities based on known marker genes and expression patterns. Such methods can leverage publicly available datasets, such as those available at the Human Protein Atlas (HPA) (11) or the multi-species Single Cell Expression Atlas (12), providing a robust, albeit time-consuming, strategy.

In this study, our primary goal was to perform a comprehensive comparative analysis of different strategies for annotating peripheral blood mononuclear cell (PBMC) populations in single-cell RNA sequencing datasets: Azimuth, an automated reference-based cell type annotation approach; Shared nearest neighbor (SNN) reference annotation naive approach, recommended by the authors of the Seurat single-cell analysis package for R as best practice (10); and manual annotation using two datasets publicly available at the HPA. We evaluated the performance of these methods in terms of accuracy, efficiency, and ability to handle the high dimensionality and complexity of scRNA-seq data. By exploring the strengths and limitations of each method, we aimed to provide critical insights that will help researchers choose the most effective strategy for annotating scRNA-seq datasets.

Material and Methods

A schematic representation of the steps involved in data acquisition and analysis is shown in Figure 1.

Datasets

Datasets- raw sequencing reads were obtained from the publicly available 10X Genomics database portal. (13) To validate PBMC populations, we used single-cell datasets obtained from healthy human donors, containing 10.000 (pbmc10k) and 5.000 (pbmc5k) cells. The datasets used were 5k Peripheral Blood Mononuclear Cells (PBMCs) from a Healthy Donor (v3 chemistry) (https://www.10xgenomics.com/datasets/5-k-peripheral-blood-mononuclear-cells-pbm-cs-from-a-healthy-donor-v-3-chemistry-3-1-standard-3-0-2) and 10k PBMCs from a Healthy Donor - Gene Expression with a Panel of TotalSeq™-B Antibodies (https://www.10xgenomics.com/datasets/10-k-pbm-cs-from-a-healthy-donor-gene-expression-and-cell-surface-protein-3-standard-3-0-0). Both datasets were downloaded on 10.05.2023.

scRNA-seq data analysis

Raw fastq files were first aligned to reference genome GRCh38 using CellRanger 7.1.0 software (10x Genomics). Generated matrices were further analyzed using Seurat package v5 (8) in R environment (14). Matrices were imported using Seurat and converted to Seurat objects containing at least 200 features in 3 cells.

A comprehensive quality control was performed to remove objects indicating multiplets. For the pbmc10k sample, the multiplets rate was estimated at 7.8%, and for pbmc5k, at 3.9%. These rates were also confirmed with DoubletFinder. (15) Thus, for pbmc10k and pbmc5k, all objects with features above 4000 and below 500 (empty droplets) or objects flagged as high-confidence doublets were discarded. Additionally, all objects expressing more than 10% of mitochondrial genes, which is a commonly chosen threshold. (16) Additionally, this threshold was selected based on numbers presented in 10x technical note CG000130. Objects with less than 5% of ribosomal genes were also filtered out to ensure healthy cells are retained as immune cells should have a high fraction of ribosomal proteins (Figure 2a). (16) Subsequently, X- and Y- chromosome genes were removed from the datasets to avoid sex-specific statistical bias due to the unknown genders of the samples. The genes with the highest expression were examined. *MALAT1* (metastasis-associated lung adenocarcinoma transcript 1) was identified as an extensive outlier, most probably representing a common technical issue, and was therefore also removed.

Next, both sample datasets were pooled, and cell cycle genes were flagged to calculate cell cycle scoring. The RNA assay data was first normalized using SCTransform. (17) Then cell cycle scores were calculated on the new SCT assay and used to calculate the S cycle score minus G2M cycle score difference. SCTransform normalization was again performed using the RNA assay and regressed on the difference in cell-cycle scores and the percentage of mitochondrial genes. The new SCT assay was used for

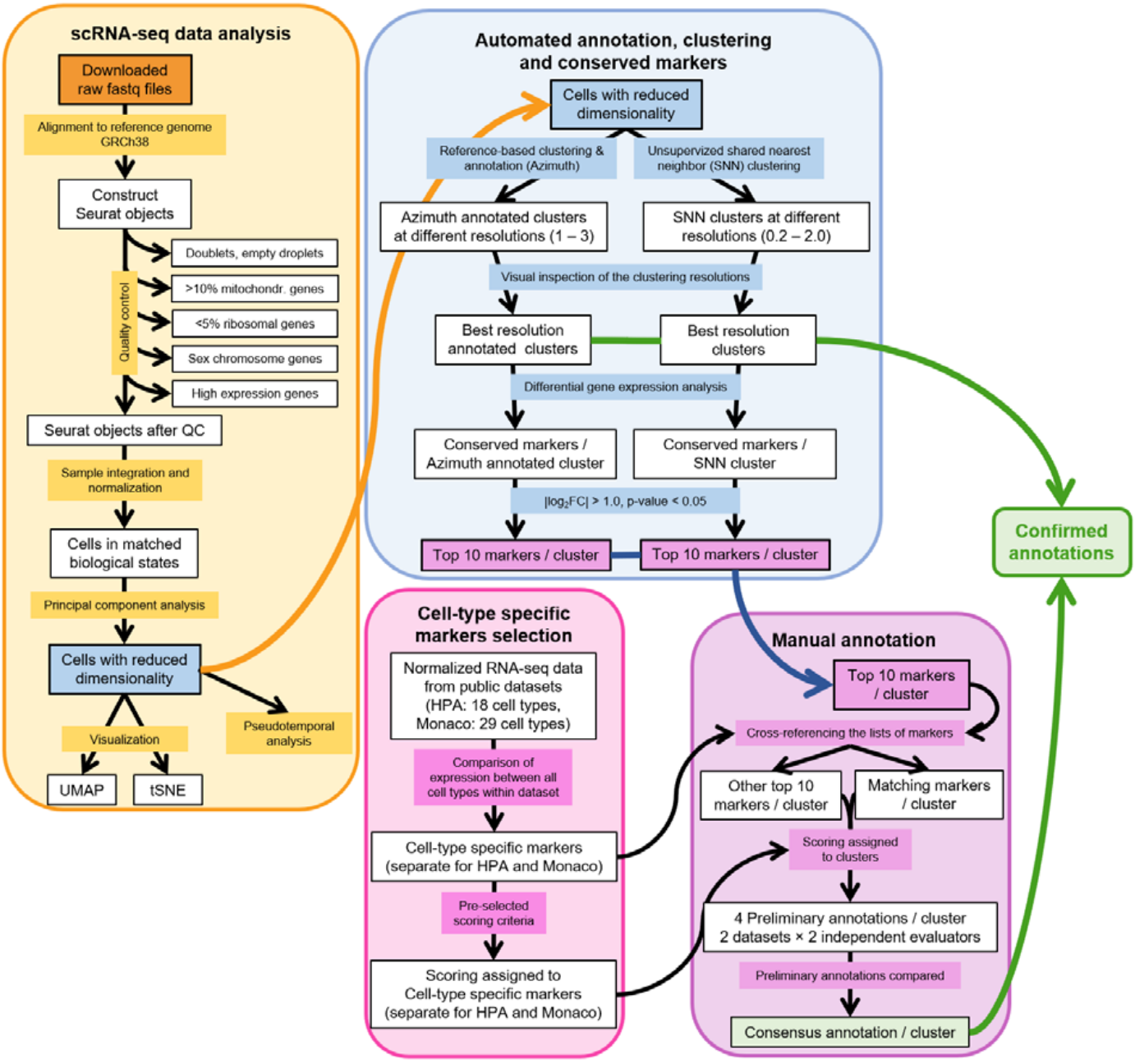

Figure 1: A schematic representation of the steps involved in data acquisition and analysis

downstream analysis and integration. We used at least 5000 features for the final anchor selection out of merged 18913 features across 10608 cells. Using integration, we identified the so-called anchors in the cross-dataset cell pairs that are in a matched biological state. These were used to correct for technical differences between datasets and align the cells between samples for comparative analyses. After integration, we performed principal component analysis for dimensionality reduction with 50 principal components (Figure 2b). Additionally, we performed uniform manifold approximation and projection (UMAP, Figure 2c) and t-distributed stochastic neighbor embedding (tSNE) analyses to visualize the high dimensional data obtained.

Automatic annotation, clustering, and conserved markers

First, automatic annotation was performed using Azimuth reference-based annotation of cells on three levels. (8) The human PBMC reference dataset was generated with 10x Genomics v3 as previously described. (8) Subsequently, the best cluster resolution was determined using the R package clustree. (18) Additionally, shared nearest-neighbor (SNN) modularity optimization clustering was deployed to cluster the cells (19).

The following resolutions were used for cluster granulation: 0.2, 0.4, 0.6, 0.8, 1.0, 1.4, 1.6, 1.8 and 2.0. After identifying

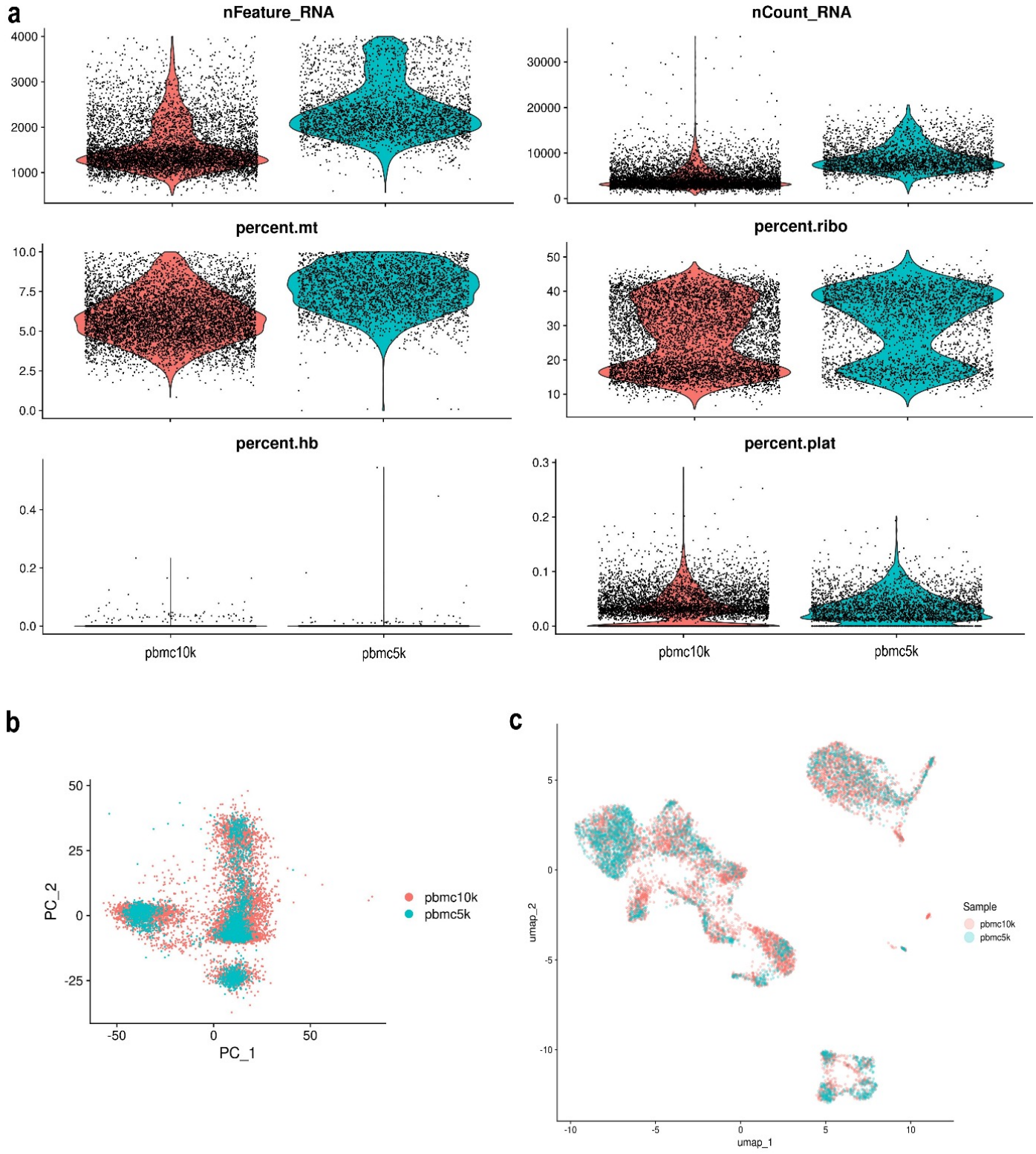

Figure 2: Quality control and dataset integration. (a) Dispersion of cells in datasets after quality control- Nfeature_RNA: number of genes per cell; nCount_RNA: number of transcripts per cell; percent.mt: percent of mitochondrial genes per cell; percent.ribo: percent of ribosomal genes per cell; percent.hb: percent of hemoglobin genes per cell; percent.plat: percent of platelet genes per cell. (b) PCA graph of the two datasets. (c) UMAP plot of aligned and integrated dataset cell landscape- pbmc10k dataset in background and pbmc5k dataset in foreground

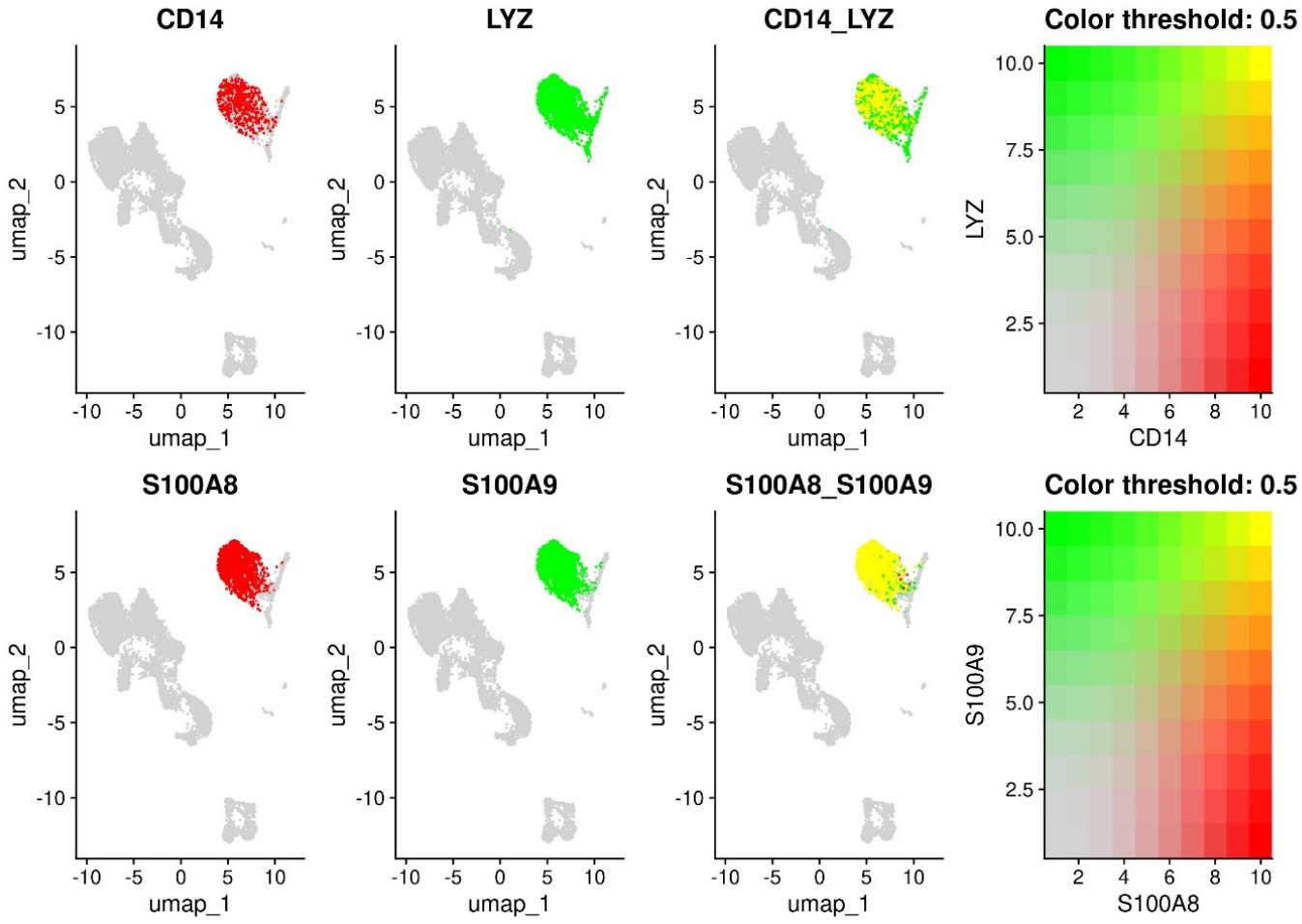

Figure 3: Visualization of selected conserved markers for classical (CD14+) monocytes. UMAP plots of *CD14* (CD14 molecule), *LYZ* (lysozyme), *S100A8* (S100 calcium-binding protein A8) and *S100A9* (S100 calcium-binding protein A9) expression, and of *CD14/LYZ* and *S100A8/S100A9* co-expression across the PBMC populations

the best annotation and cluster resolution, conserved cell-type markers with the same perturbation direction in both datasets were identified by differential gene expression testing. The MetaDE R package embedded in Seurat's FindConservedMarkers function was used for this purpose. (20) Conserved markers (see Figure 3 for an example) were only identified in cell populations where at least three cells were present in an independent sample.

Cell type-specific marker selection and manual cluster annotation

For manual annotation, we used two publicly available human datasets- the RNA HPA immune cell gene data (the HPA dataset) and the RNA Monaco immune cell gene data (the Monaco dataset), which we downloaded from the HPA website (https://www.proteinatlas.org/about/download) on 23.6.2023. The HPA dataset contains transcription data on 18 immune cell types from blood generated within the HPA project (21), while the Monaco dataset is based on the RNA-seq data generated on 29 FACS-sorted immune cell types from the PBMC of healthy donors. (22) The pipeline used to generate both datasets from the raw RNA-seq data, including quality control and normalization, is described on the HPA website. The downloaded datasets are based on The HPA version 23.0 and Ensembl version 109.

For both annotation datasets, we separately determined cell type-specific markers based on the normalized gene expression values, with a cutoff value of 4 as described on the HPA website (https://www.proteinatlas.org/about/assays+annotation#hpa_rna). Genes whose normalized expression levels in a specific immune cell type were at least 4× higher than in any other immune cell type were considered cell type-specific markers for that specific immune cell type. Similarly, genes whose normalized expression in a group of two or three immune cell types was at least 4× higher than in any other immune cell type were considered twin or group markers, respectively. In addition to the markers for the immune cell types defined in the two datasets, we also determined specific markers for several broader groups of immune cell types, for example, total CD8+ T-cells (comprised of Naïve CD8+ T-cell, Central memory CD8+ T-cells, Effector memory CD8+ T-cells and Terminal effector memory CD8+ T-cells in the Monaco dataset). For these marker genes, it was defined that the lowest

Table 1: Immune cell type-specific markers

Marker		Scoring			Nr. of markers	
Type	**Definition**	**Cell type 1**	**Cell type 2**	**Cell type 3**	**HPA dataset**	**Monaco dataset**
Single	nEL in CT1 > 4× nEL in any other CT	8	/	/	1821	1581
Twin 2	nEL in CT1, CT2 > 4× nEL in any other CT nEL in CT1 ≈ nEL in CT2	4	4	/	594	458
Twin 1+1	nEL in CT1, CT2 > 4× nEL in any other CT nEL in CT1 > 4× nEL in CT2	8	4	/	224	149
Group 3	nEL in CT1, CT2, CT3 > 4× nEL in any other CT nEL in CT1 ≈ nEL in CT2 ≈ nEL in CT3	2	2	2	436	273
Group 2+1	nEL in CT1, CT2, CT3 > 4× nEL in any other CT nEL in CT1 ≈ nEL in CT2 nEL in CT1, CT2 >4× nEL in CT3	4	4	2	36	25
Group 1+2	nEL in CT1, CT2, CT3 > 4× nEL in any other CT nEL in CT1 > 4× nEL in CT2, CT3 nEL in CT2 ≈ nEL in CT3	8	2	2	125	71
Group 1+1+1	nEL in CT1, CT2, CT3 > 4× nEL in any other CT nEL in CT1 > 4× nEL in CT2, CT3 nEL in CT2 > 4× nEL in CT3	8	4	2	15	11

(nEL- normalized expression level, CT – cell type)

normalized expression level within the broader group of immune cell types had to be at least 4× higher than in any other immune cell type not included in the specific group. Finally, scores were assigned to all the markers based on marker type. For the twin and group markers, relative differences in normalized gene expression within the pair/group were also considered (Table 1).

Next, the top 10 best-conserved markers for each cluster were determined. To this end, we first selected markers (genes) with a log2FC (fold change) >1.0 and an adjusted p-value <0.05 to ensure that only genes with both high and significant differences in expression levels between clusters were considered. The markers meeting the criteria were then ranked based on the highest $\log_2$FC values.

Then, each cluster's top 10 conserved markers were cross-referenced with the cell type-specific markers from each annotation dataset separately. In this way, clusters were assigned to possible cell types, and each possible cell type was assigned a score based on the scores of the markers identified by the cross-reference. For additional clarification, those of the top 10 conserved markers not identified as cell type specific markers were also considered. If their expression in a particular immune cell type was at least 4× or 2× higher than the average expression of all immune cell types in the annotation dataset, they were assigned a score of 2 or 1, respectively. These scores were added to the above scores of the possible cell types. The final scores obtained for each cluster from the two annotation datasets were then used by two independent evaluators to determine preliminary annotations for each cluster. Finally, the preliminary annotations were compared by the two evaluators and an additional referee to reach a consensus annotation. In ambiguous cases, a broader annotation took precedence over a narrower one (*I.e.*, B-cells vs naïve B-cells) unless multiple clusters shared the same annotation: in such situations, we aimed for consensus with the narrower annotations.

Trajectory and pseudo-time analysis

The trajectory of the cell transitions and the pseudo-temporal arrangement of cells during differentiation was analyzed using the R package monocle3 and the Python implementation. (23–25) The previously constructed Seurat object was pre-processed and partitioned into the main cell types (monocytes, B-cells, T-cells). An explicit principal graph was learned using advanced machine learning called Reverse Graph Embedding to accurately resolve biological processes in individual cells' Pseudo-time. This abstract measure of an individual cell's progress in cell differentiation was calculated as the distance between a cell and the beginning of the trajectory measured along its shortest path. The total length of a trajectory was defined as the total amount of transcriptional changes a cell undergoes on its way from start to end state. The cells with the highest expression of the *CD14* gene for monocytes and the calculated start nodes for T-cells and B-cells were chosen as the roots or so-called beginnings of a biological process. To calculate the start node, the resident cells (double negative T-cells and intermediate B-cells) were first grouped

according to the nearest node of the trajectory graph, and then the proportion of the cells at each node originating from the earliest time point was calculated. The node most heavily occupied by early cells was then selected as the root. Finally, the UMAP visualization was used to identify the pseudo-temporal cell state transition compared to the Azimuth annotation.

Results

Automatic annotation with Azimuth

After quality control and data integration, we used Azimuth's reference-based annotation of cells to automatically determine clusters and immune cell types. First, we evaluated three levels of cluster granulation to determine the best resolution of clusters using a clustering tree diagram. The first and second levels of annotation provide a clear separation between all annotated clusters, while the third level exhibits some over-clustering (Figure 4a). Similarly, UMAP plots of the first two Azimuth annotated clustering levels show clear separations between clusters. At the same time, some over-clustering is evident in level three, for example, populations NK_2, NK_3, and NK_4 (Figure 4b). Overall, the resolution at level one provides information on eight, level two on 28 and level three on 51 distinct PBMC subpopulations. Based on the cluster-tree analysis (*I.e.*, presence of over-clustering and number of distinct subpopulations), level 2 was chosen as the best solution, providing sufficient resolution and the most information.

Unsupervised clustering according to SNN

Additionally, we performed SNN modularity optimization clustering. Again, the best resolution was chosen based on the clustering-tree diagram. Here, the best resolution of granulation was achieved at a resolution of 0.8, with higher and lower resolutions showing at least some over-clustering (Figure 5a). UMAP plots of the smallest (0.2), largest (2.0), and best (0.8) resolution of clustering were also inspected. Clustering at a resolution of 0.2 provided information on 12 unannotated PBMC subpopulations, although more distinct clusters can be observed (for examples, see clusters 2, 3, and 4, Figure 5b). On the other hand, a resolution of 2.0 resulted in 28 distinctive unannotated PBMC subpopulations, with clear signs of poor cluster separation in several instances (for examples, see clusters 4, 5, and 6 or 2, 3, 7, and 19, Figure 5b). Only the best resolution (0.8) shows 18 well-separated PBMC subpopulations (Figure 5b) and was thus chosen as the best resolution for further inspection.

Manual annotation of the Azimuth and SNN clusters

As described above, manual annotation was based on two publicly available datasets and two independent evaluators. Both evaluators cross-referenced the cell-type specific markers defined from the datasets with the top 10 conserved markers from each cluster, thus creating four independent preliminary annotations for all the clusters. The four preliminary annotations were then used to define each cluster's final, consensus annotation. Of note, we could not annotate all the clusters in this way- for 10 Azimuth annotated clusters (for example, classical dendritic cells type 1, plasmablasts or hematopoietic stem/progenitor cells) and 2 of the SNN clusters (clusters 16 and 17) no conserved markers could be defined (see Tables 2 and 3, respectively).

The consensus manual annotation was identical to the Azimuth annotation for 11/19 clusters for which conserved markers could be defined (Table 2). In 5 cases (myeloid dendritic cells instead of type 2 classical dendritic cells; Memory CD4+ T-cells vs Central memory CD4+ T-cells; T cells vs Effector memory and Cytotoxic CD4+ T-cells; natural killer cells vs CD56 bright natural killer cells), the manual annotation identified a super-set and in one case (Exhausted memory B-cells instead of Memory B-cells) a sub-set of the immune cell subtype identified by the Azimuth annotation. In the last cluster, manual annotation identified a different sub-set (Non-switched memory B-cells) of the same super-set (B-cells) than the Azimuth annotation (Intermediate B-cells). Manual annotation of the 16 SNN clusters, for which conserved markers could be defined, identified 15 relatively specific immune cell subtypes, while in one cluster, only a very broad annotation (T-cells) could be determined (Table 3).

Comparison of the Azimuth and SNN clusters with manual annotation

Comparison of the 28 Azimuth annotated clusters, the 18 unsupervised SNN clusters, and the manual annotation of the latter showed good matching for all the Monocytes, Dendritic cells, and B-cells populations/clusters as well as 2/3 natural killer cells populations (Figure 6a-b, Table 4). Also matching are the Naïve CD4+ and CD8+ T-cells, Memory CD8+ T-cells, Mucosal-associated invariant T-cells, and γδ T-cells clusters. The rest of the T-cell populations do not match directly; however, in general, it is evident whether the clusters fall within CD4+ or CD8+ T-cell populations. The hematopoietic stem/progenitor cells, Innate lymphoid cells and platelet populations were not manually annotated due to the lack of appropriate conserved markers (Table 2). In the unsupervised SNN clustering, these populations do not represent separate clusters but are instead distributed among (CD4+) T-cells associated clusters (Figure 6a-b, Table 4).

Pseudo-temporal trajectory analysis

Pseudo-temporal trajectory analysis was used as a final validation method. With this analysis we followed the cell state progress through the differentiation of three distinct Azimuth superclusters. In the partition of the Monocytes supercluster, it's visible that cells start to differentiate in

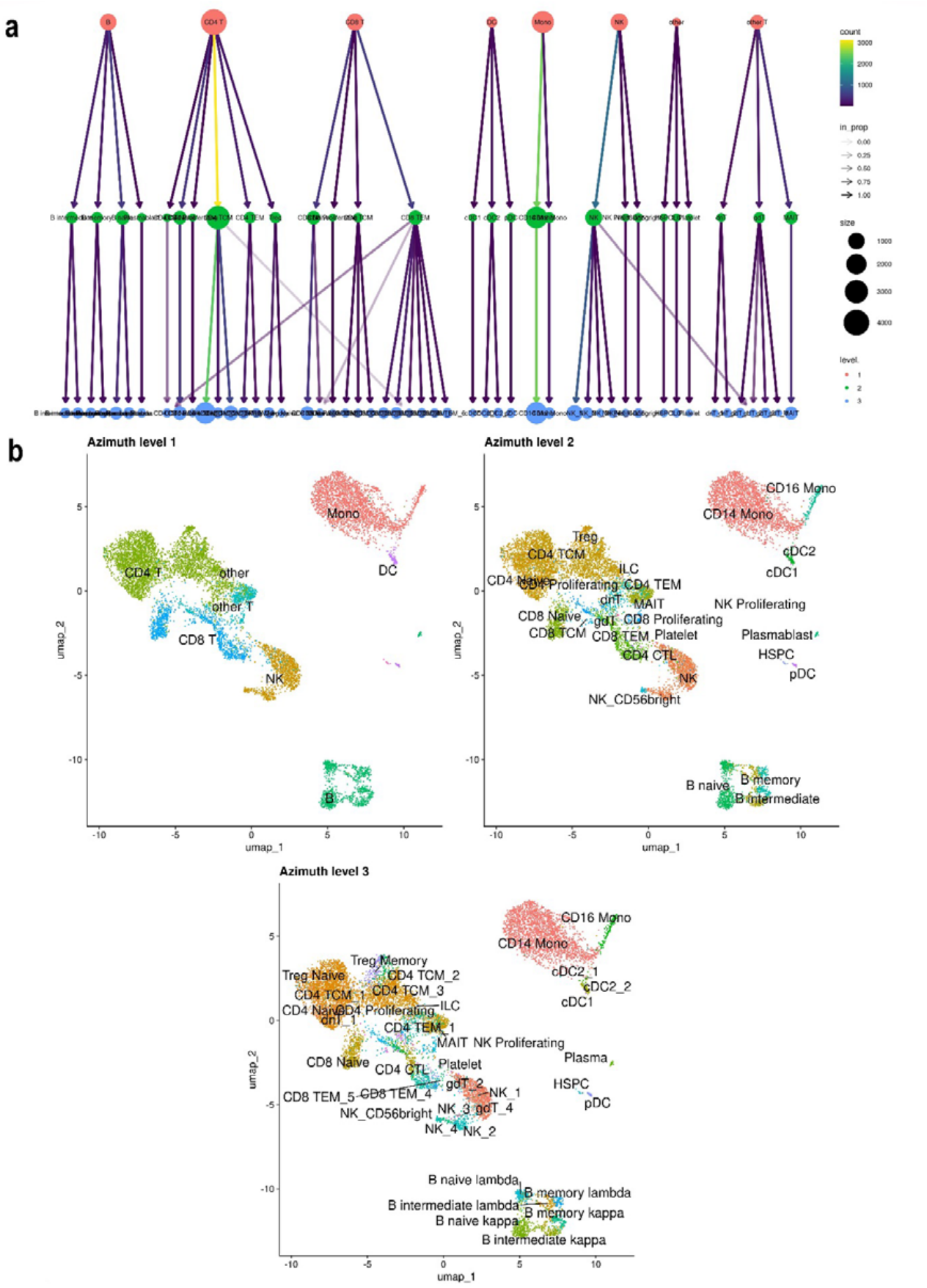

Figure 4: Automated annotation with Azimuth. (a) Evaluation of Azimuth annotation levels using clustering tree. The best resolution is encircled in red. (b) UMAP plots of annotation using all three levels from the Azimuth database. Upper left: level one annotation clusters; Upper right: level two annotation clusters; Lower: level three annotation clusters

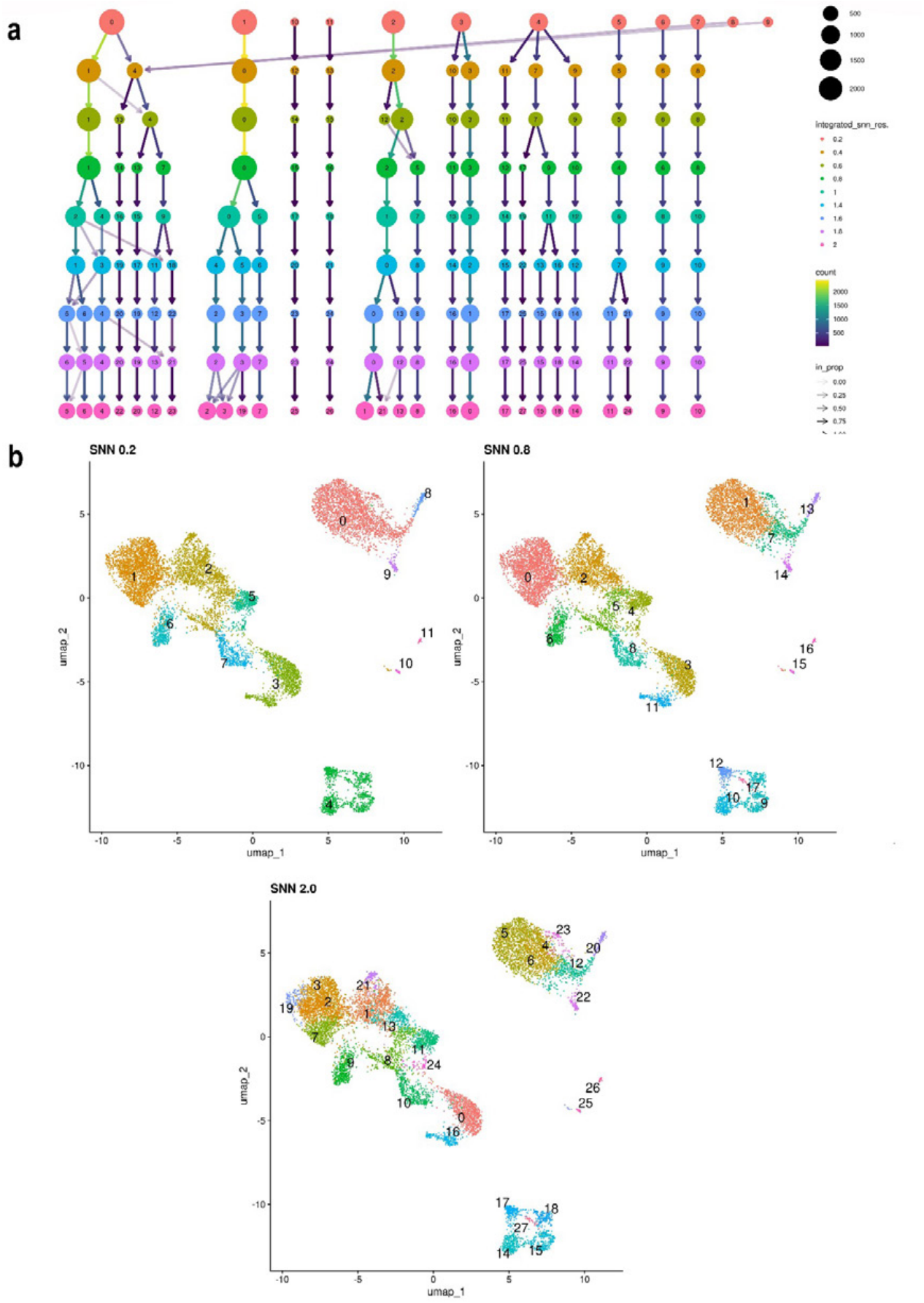

Figure 5: Shared nearest neighbor clustering optimization. (a) Evaluation of clustering levels using clustering tree. The best resolution is encircled in red. (b) UMAP plots of the smallest (0.2; upper left), the largest (2.0; upper right), and the best (0.8; lower) resolution of clustering

Table 2: Comparison between the Azimuth and manual annotation of the best resolution clusters

Azimuth annotation	1st Evaluator's provisional annotation		2nd Evaluator's provisional annotation		Consensus annotation
	HPA dataset	Monaco dataset	HPA dataset	Monaco dataset	
CD14+ Monocytes	Classical Monocytes	Classical Monocytes	Classical Monocytes	Classical Monocytes	Classical Monocytes
CD16+ Monocytes	Non-classical Monocytes	Non-classical / Intermediate Monocytes	Non-classical Monocytes	Non-classical / Intermediate Monocytes	Non-classical Monocytes
cDC, type 1	/	/	/	/	/
cDC, type 2	mDC	mDC	mDC	mDC	mDC
pDC	pDC	pDC	pDC	pDC	pDC
Naïve B-cells	Naïve / Memory B-cells	Naïve B-cells	Naïve B-cells	Naïve B-cells	Naïve B-cells
Intermediate B-cells	Memory / Naïve B-cells	Non-switched memory B-cells	Memory B-cells	Non-switched memory B-cells	Non-switched memory B-cells
Memory B-cells	Memory / Naïve B-cells	Exhausted / Switched memory B-cells	Memory B-cells	Exhausted memory B-cells	Exhausted memory B-cells
Plasmablasts	/	/	/	/	/
Double-negative T-cells	/	/	/	/	/
Naïve CD4+ T-cells	Naïve CD4+ T-cells	Naïve CD4+ T-cells	Naïve CD4+ T-cells	Naïve CD4+ T-cells	Naïve CD4+ T-cells
Proliferating CD4+ T-cells	/	/	/	/	/
TCM CD4+	Naïve / Memory CD4+ T-cells	Naïve / TFH Memory CD4+ T-cells	Naïve / Memory CD4+ T-cells	Naïve CD4+ T-cells	Memory CD+ T-cells
TEM CD4+	MAIT / γδ T-cells	MAIT / Vδ2+ γδ T-cells	MAIT	MAIT	T-cells
CTL CD4+	/	/	/	/	/
Treg	Treg	Treg	Treg	Treg	Treg
Naïve CD8+ T-cells	Naïve CD8+ T-cells	Naïve CD8+ T-cells	Naïve CD8+ T-cells	Naïve CD8+ T-cells	Naïve CD8+ T-cells
Proliferating CD8+ T-cells	/	/	/	/	/
TCM CD8+	Memory CD8+ T-cells	TCM / TEM CD8+	Memory CD8+ T-cells	TCM CD8+	TCM CD8+
TEM CD8+	Memory CD8+ T-cells	TEM CD8+	Memory CD8+ T-cells	TEM CD8+	TEM CD8+
MAIT	MAIT	MAIT	MAIT	MAIT	MAIT
γδ T-cells	γδ T-cells	Vδ2+ γδ T-cells	γδ T-cells	Vδ2+ γδ T-cells	γδ T-cells
NK	NK / γδ T-cells	NK	NK / γδ T-cells	NK	NK
CD56 bright NK	NK	NK / Vδ2+ γδ T-cells	NK	NK	NK
Proliferating NK	/	/	/	/	/
HSPC	/	/	/	/	/
ILC	/	/	/	/	/
Platelets	/	/	/	/	/

cDC (classical Dendritic Cells); CTL (Cytotoxic T-cells); HSPC (Hematopoietic stem/progenitor cells); ILC (Innate lymphoid cells); MAIT (Mucosal-associated invariant T-cells); mDC (myeloid Dendritic Cells); NK (Natural Killer Cells); pDC (plasmacytoid Dendritic Cells); TCM (Central Memory T-cells); TEM (Effector Memory T-cells); Treg (Regulatory T-cells)

Table 3: Manual annotation of the best clusters according to SNN

SNN clusters	1st Evaluator's provisional annotation		2nd Evaluator's provisional annotation		Consensus annotation
	HPA dataset	Monaco dataset	HPA dataset	Monaco dataset	
0	Naïve CD4+T-cells	Naïve CD4+T-cells	Naïve CD4+T-cells	Naïve CD4+T-cells	Naïve CD4+T-cells
1	Classical Monocytes / Neutrophils	Classical Monocytes / Neutrophils	Classical Monocytes	Classical Monocytes	Classical Monocytes
2	Memory CD4+ T-cells / Treg	Th17 Memory CD4+ T-cells	Memory CD4+ T-cells / Treg	Th17 Memory CD4+ T-cells	Memory CD4+ T-cells
3	NK / γδ T-cells	Non-Vδ2+ γδ T-cells / TEM CD8+	NK / γδ T-cells	Non-Vδ2+ γδ T-cells	T-cells
4	γδ T-cells / MAIT	MAIT	γδ T-cells	MAIT	MAIT
5	γδ T-cells / Treg	Vδ2+ γδ T-cells / Non-Vδ2+ γδ T-cells / MAIT	γδ T-cells / Treg	Vδ2+ γδ T-cells	γδ T-cells
6	Naïve CD8+ T-cells	Naïve CD8+ T-cells	Naïve CD8+ T-cells	Naïve CD8+ T-cells	Naïve CD8+ T-cells
7	Intermediate Monocytes	Intermediate Monocytes	Intermediate Monocytes	Intermediate Monocytes	Intermediate Monocytes
8	Memory CD8+ T-cells	TEM CD8+	Memory CD8+ T-cells	TEM CD8+	Memory CD8+ T-cells
9	Memory B-cells	Exhausted memory B-cells	Memory B-cells	Exhausted memory B-cells	Exhausted memory B-cells
10	Naïve B-cells	Naïve B-cells	Naïve B-cells	Naïve B-cells	Naïve B-cells
11	NK	NK	NK	NK	NK
12	Naïve / Memory B-cells	Naïve / Non-switched memory B-cells	Naïve B-cells	Naïve B-cells	Non-switched memory B-cells
13	Non-classical Monocytes	Non-classical / Intermediate Monocytes	Non-classical Monocytes	Non-classical / Intermediate Monocytes	Non-classical Monocytes
14	mDC	mDC	mDC	mDC	mDC
15	pDC	pDC	pDC	pDC	pDC
16	/	/	/	/	/
17	/	/	/	/	/

MAIT (Mucosal-associated invariant T-cells); mDC (myeloid Dendritic Cells); NK (Natural Killer Cells); pDC (plasmacytoid Dendritic Cells); TEM (Effector Memory T-cells); Treg (Regulatory T-cells)

the middle of the CD14+ Monocytes cluster, progressing outwards (Figure 7a). The trajectory distinctively shows progression into Non-classical CD16+ monocytes, whereas type 2 cDC cells are not connected with any trajectory. Within the B-cells supercluster, the starting node resides in the Intermediate B-cells with trajectory soon forking into two arms, both pointing towards Memory and Naïve B-cells (Figure 7b). The starting node within the T-cells supercluster resides in the middle of the cluster (Figure 7c), a position corresponding to the double negative T cells according to the Azimuth annotation (Figure 4b). One trajectory clearly shows differentiation into CD4+ sub-populations and other-T cells, while the other branches early into CD8+ sub-populations and the natural killer cells.

Discussion

ScRNA-seq enables the simultaneous analysis of expression profiles and their interdependencies in multiple cell types present in a tissue of interest. This represents a qualitative leap forward in studying complex biological processes and the role of individual cell subtypes in these processes. Previously, several separate studies were required to achieve the same result. However, reliable and accurate identification of the cell subtypes present in a selected biological sample cannot be taken for granted. (26) In the work presented here, we compared the cell-type annotation techniques/tools used in scRNA-seq to highlight their strengths and potential pitfalls. Specifically, we used an automated reference-based tool, Azimuth, and an SNN clustered reference-naïve approach followed by manual annotation to

Table 4: Comparison of the best Azimuth annotation, the best SNN clustering resolution, and manual annotation.

Azimuth annotation	SNN cluster with cells present in the Azimuth annotation	Manual annotation consensus	
		The best-resolution Azimuth clusters	The best clusters, according to SNN
CD14+ Monocytes	0, **1*** , 2, 3, **7***, 9, 14, 16	Classical Monocytes	Classical Monocytes; Intermediate Monocytes
CD16+ Monocytes	1, 2, 7, **13***	Non-classical Monocytes	Non-classical Monocytes
cDC, type 1	**14**	/	mDC
cDC, type 2	7, 9, **14***	mDC	mDC
pDC	**15**	pDC	pDC
Naïve B-cells	1, **10***, **12***, 17	Naïve B-cells	Naïve B-cells; Non-switched memory B-cells
Intermediate B-cells	**9***, 10, 12, 17	Non-switched memory B-cells	Exhausted memory B-cells
Memory B-cells	**9***, 17	Exhausted memory B-cells	Exhausted memory B-cells
Plasmablasts	**16**	/	/
Double-negative T-cells	**0**, **5**	/	Naïve CD4+T-cells; γδ T-cells
Naïve CD4+ T-cells	**0***, 1, 6	Naïve CD4+ T-cells	Naïve CD4+T-cells
Proliferating CD4+ T-cells	**1**, **2**, **4**	/	Classical Monocytes; Memory CD4+ T-cells; MAIT
TCM CD4+	**0***, 1, **2***, 3, 4, 5, 6, 8, 10	Memory CD+ T-cells	Naïve CD4+T-cells; Memory CD4+ T-cells
TEM CD4+	0, 2, 4, **5***, 8	T-cells	γδ T-cells
CTL CD4+	**3**, **8**	/	T-cells; Memory CD8+ T-cells
Treg	0, **2***, 6	Treg	Memory CD4+ T-cells
Naïve CD8+ T-cells	0, 2, 4, **6***, 8	Naïve CD8+ T-cells	Naïve CD8+ T-cells
Proliferating CD8+ T-cells	**1**, **4**, **8**	/	Classical Monocytes; MAIT; Memory CD8+ T-cells
TCM CD8+	0, 2, **5***, **6***, 8	TCM CD8+	γδ T-cells; Naïve CD8+ T-cells
TEM CD8+	0, 1, 2, 3, 4, 5, 6, **8***	TEM CD8+	Memory CD8+ T-cells
MAIT	2, 3, **4***, 5	MAIT	MAIT
γδ T-cells	0, 2, 3, **4***, **5***, 6, 8, 11	γδ T-cells	MAIT; γδ T-cells
NK	**3***, 4, 8, **11***, 17	NK	T-cells; NK
NK CD56 bright	**11**	NK	NK
NK proliferating	**1**, **3**	/	Classical Monocytes; T-cells
HSPC	**0**, **1**, **2**	/	Naïve CD4+T-cells; Classical Monocytes; Memory CD4+ T-cells
ILC	**2***, 5	/	Memory CD4+ T-cells
Platelets	**3**	/	T-cells

* The most abundant SNN cluster; cDC (classical Dendritic Cells); CTL (Cytotoxic T-cells); HSPC (Hematopoietic stem/progenitor cells); ILC (Innate lymphoid cells); MAIT (Mucosal-associated invariant T-cells); mDC (myeloid Dendritic Cells); NK (Natural Killer Cells); pDC (plasmacytoid Dendritic Cells); TCM (Central Memory T-cells); TEM (Effector Memory T-cells); Treg (Regulatory T-cells)

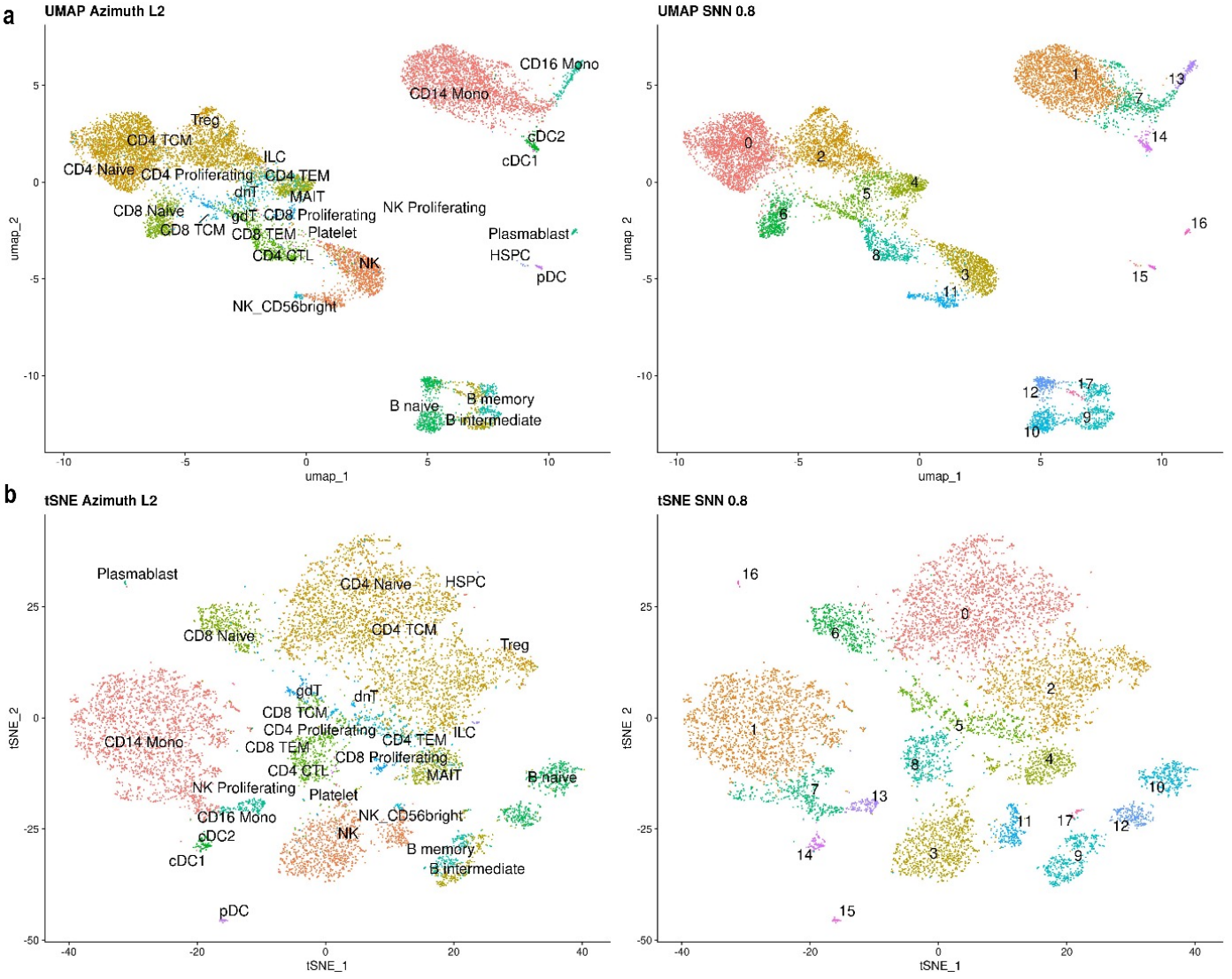

Figure 6: Comparison of Azimuth and SNN clustered cell landscapes (a) UMAP plots of the best resolution Azimuth clusters (left) and the best resolution SNN clusters (right). (b) tSNE plots of the best-resolution Azimuth clusters (left) and the best-resolution SNN clusters (right)

gain insights into their utility and effectiveness in deciphering complex cellular compositions in scRNA-seq datasets.

As a starting point for the analyses, we chose PBMC, a relatively complex but easily accessible biological sample commonly used in medical and veterinary research. We first used the automatic annotation tool Azimuth. Its annotations are based on a reference PBMC dataset generated from 24 samples processed with a CITE-seq (Cellular Indexing of Transcriptomes and Epitopes by Sequencing) panel, which performs RNA sequencing along with obtaining quantitative and qualitative information about proteins (i.e., cell type-specific antigens) on the cell surface. (8) Azimuth automatic annotation has demonstrated the ability to process large scRNA-seq datasets quickly and accurately. The performance of this machine learning-based tool reflects ongoing advances in computational biology, particularly in the automated processing of biological data. (27–29) Performance of the automated annotation tools may decline when confronted with rare cell types, as the classifier may be unable to learn their information during the training phase (30).

In that regard, Azimuth also proved relatively well, as it defined several PBMC populations with low abundance (i.e., classical dendritic cells, plasmacytoid dendritic cells, hematopoietic stem/progenitor cells, Innate lymphoid cells) (31–33) of which the first two we could independently confirm with the manual annotation. Like any other reference-based tool, however, it cannot recognize / annotate populations that lie outside its frame of reference. (34) For example, $CD14^{+}$ and $CD16^{+}$ monocyte populations were annotated that roughly correspond to the classical ($CD14^{+}CD16^{neg}$) and non-classical ($CD14^{dim}CD16^{+}$) monocytes in the HPA and Monaco datasets, respectively, but the intermediate ($CD14^{+}CD16^{+}$) monocytes could not be distinguished.

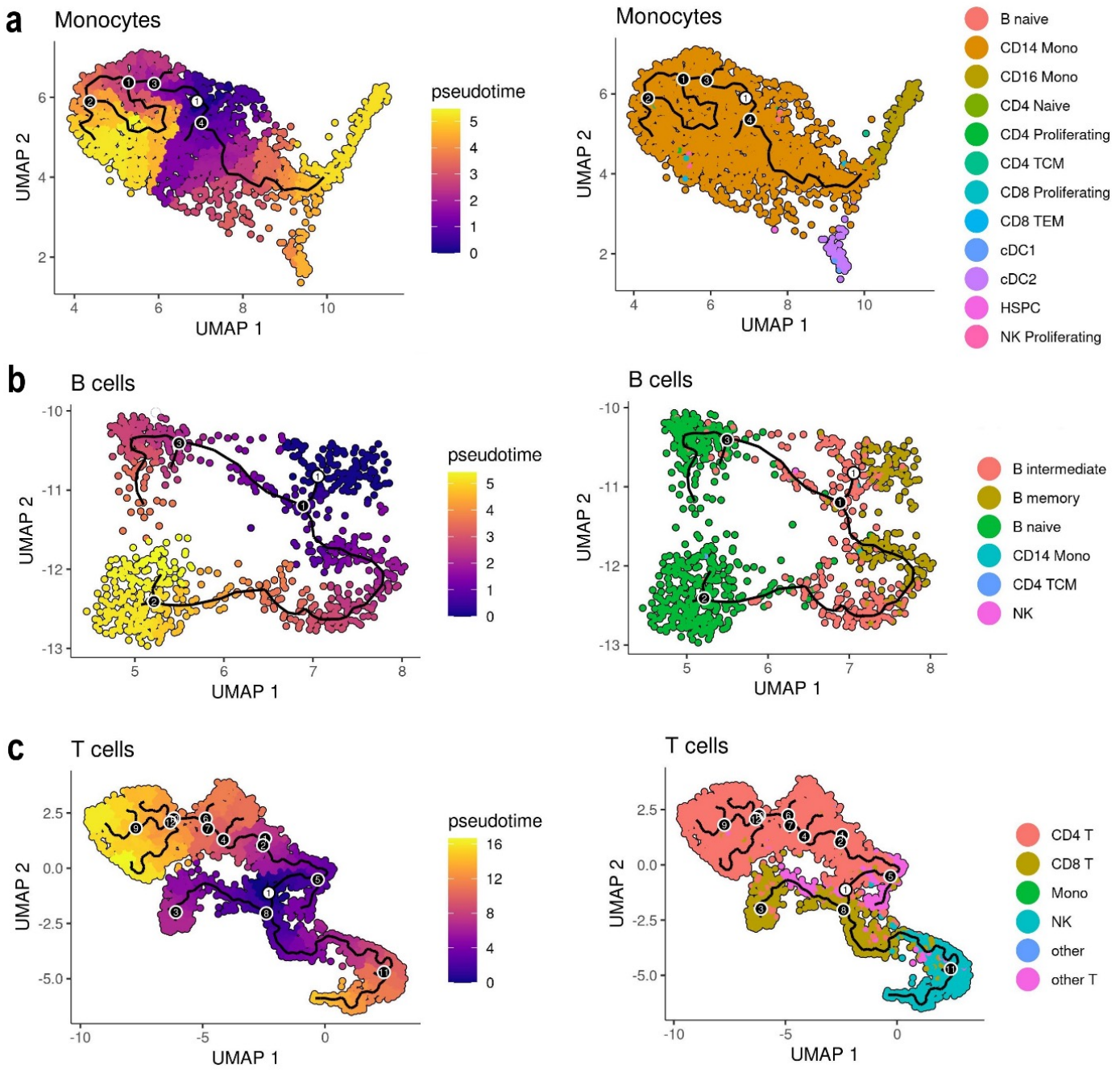

Figure 7: Pseudo-temporal analysis of selected immune cell types. (a) UMAP visualization of pseudo-temporal trajectories (left) and Azimuth annotations (right) of the monocytes partition on levels one and two. (b) UMAP visualization of pseudo-temporal trajectories (left) and Azimuth annotations (right) of the B-cells partition on levels one and two. (c) UMAP visualization of pseudo-temporal trajectories (left) and Azimuth annotations (right) of the T-cells partition on levels one and two

Unsupervised SNN clustering, on the other hand, easily defined three distinct populations at the position corresponding to the monocytes in the Azimuth analysis. Subsequent manual annotation identified them as the three monocyte types mentioned above. The ability of this method to effectively delineate cell populations is well documented (27), and our results confirm its robustness in unsupervised clustering. The main advantage of unsupervised clustering over the reference-based one is that it does not attempt to fit cells into the pre-existing reference frame. Instead, the cells are clustered merely according to their similarity. Unsupervised clustering thus provides more opportunities to recognize rare or even new populations. (34) Conversely, the same fact is also a major disadvantage of the unsupervised method, as one cannot avoid the time-consuming manual annotation of the individual clusters. (34)

In our manual annotation we prioritized biological relevance and statistical rigor. We followed strict criteria, similar to the HPA protocols, for cell type-specific markers selection.

Similarly, we used rigorous criteria to define the top 10 best-conserved markers per cluster, which were then used for comparison with the cell type-specific markers and, thus, cell type annotation of the clusters. We also tried annotation with all the conserved markers (15 – 855 conserved markers per cluster, not shown). A similar approach was used for the scSorter tool, where they combined the expression of marker and non-marker genes for clustering. (35) We found no significant differences in annotation with all versus only the top 10 markers, so we chose the latter, a somewhat less time-consuming approach, for further analysis. Of note is that the stringency of the above criteria for conserved markers resulted in no conserved markers being defined for some clusters. This further meant that these clusters could not be manually annotated; we however decided against loosening the criteria. Besides objective data, manual interpretation also benefits from the evaluator's understanding of the biological processes, but at the same time, it inherently creates bias. (34) To minimize bias, we used two independent annotation datasets, both based on the FACS sorted cell populations (21, 22), and employed two independent evaluators, plus an additional arbiter, to reach consensus annotation for each cluster. This careful approach ensured high accuracy in identifying different cell types, as asserted by the high similarity of the manual and the automated (Azimuth) annotations.

Many discrepancies between the two annotations can be explained by the differences in how specific cell subtypes are defined in the respective reference datasets. Particularly challenging are the phenotypically and functionally highly heterogeneous subsets of the T-cells (36), where the HPA dataset recognizes 7 and the Monaco dataset 15 separate entities (21, 22), which were in turn used to validate the 13 T-cell clusters identified by the Azimuth manually. At the exact coordinates, the unsupervised SNN clustering identified only 6 distinct populations, all manually annotated as various T-cell subsets. Directly comparing the SNN clusters with the Azimuth annotations further emphasizes the invaluability of using multiple approaches when tackling complex populations/clusters. Namely, it clearly shows that a population, coherent at a given level, may consist of several distinct subpopulations. These are not necessarily closely related, and vice-versa, the well-defined cell types may be dispersed over several distinct clusters. In clinical samples related to a specific pathology, such instances can provide opportunities for the identification of important rare and potentially even novel subpopulations. They should thus be more thoroughly investigated at a higher resolution.

The selected resolution of the clustering directly influences the granularity of the identified cell types and, thus, the depth of the biological insights that can be gained. (18) A high resolution can reveal subtle differences between cell populations and possibly visualize rare or transitional states of cells, but at a risk of decreased reliability of clustering- for example, see CD4+ T-cells sub-clusters at resolution level 3 (Figure 4a). Conversely, lower resolution may be highly reliable, but risks conflating cell types with different functions (for example looking at combined CD4+ T-cells instead of the subsets with very distinct roles- regulatory, helper, effector, etc.) and can thus miss important biological differences. (36) Hence, the optimal resolutions we chose for both reference-based and unsupervised clustering are compromises, balancing between distinguishing meaningful cellular subtypes and avoiding fragmentation of homogeneous populations into overly granular clusters. Alas, as with any compromise, the optimal resolution does not satisfy completely, which is most evident when clustering the B-cells. Here, at optimal resolutions, Azimuth and SNN clustering identify 3 and 4 distinct subpopulations, respectively. Visually, though, one can easily distinguish 5-6 entities, suggesting that a higher resolution would be needed here.

The meaningfulness of a granularity higher than the one defined by the optimal resolution for the B-cell subsets was also confirmed by the pseudo-temporal analysis. The pseudo-temporal dimension introduces a framework for mapping progression states and inferring transitional states and lineage relationships. It highlights not only the end states cells reach but also the paths they take to get there. (24) Using this method, we further validated the T-cell and monocyte subsets annotations. The pseudo-temporal trajectories also clearly show that the automatic Azimuth annotations cohere with known cell differentiation stages. The method has previously been instrumental in charting developmental trajectories, and our application further underscores its value in modeling cellular dynamics, as has been explored in other studies focusing on differentiation and immune cells. (37, 38)

Regardless of the sample type, its origin, underlying pathology, and the scientific question, single-cell RNA sequencing has little value if one cannot properly identify the single cells. Novel and ever more powerful tools for accurate and reliable annotation of the cells/clusters are therefore being developed. (39–41) However, as demonstrated here, each method has its merits and downsides. The methods and results of our study have significant implications for the further development of scRNA-seq applications, not only in the field of human medicine but even more in the field of veterinary medicine. Unlike in human medicine, namely, in veterinary medicine, there is often a lack of comprehensive databases for reference-based annotation. (12, 42) This makes using automated annotation tools such as Azimuth difficult and emphasizes the importance of integrating different annotation approaches. In the future, integrated tools may be developed that will combine the efficiency of the automated annotation and expert insight of the manual one, the accuracy of the reference-based annotation with the flexibility of the unsupervised clustering. Until then, a skillful combination of automated and manual annotation techniques is needed to manage the complexity of scRNA-seq data when reference databases are limited or non-existent. This approach is particularly crucial in veterinary science,

where the study of different species requires a customized approach to cell type annotation, given the variability in genetic and cellular profiles of different species. With it, scRNA-seq research can open new avenues for discovery in cell biology and its applications in health and disease.

Acknowledgments

This work was not supported by any specific funding.

References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009; 10(1): 57–63. doi: 10.1038/nrg2484
2. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet 2019; 20(11): 631–56. doi: 10.1038/s41576-019-0150-2
3. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med 2017; 9(1): 75. doi: 10.1186/s13073-017-0467-4
4. Wagner A, Regev A, Yosef NC. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol 2016; 34(11): 1145–60. doi: 10.1038/nbt.3711
5. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 2009; 6(5):3 77–82. doi: 10.1038/nmeth.1315
6. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics 2019; 20(1): 40. doi: 10.1186/s12859-019-2599-6
7. Poirion OB, Zhu X, Ching T, Garmire L. Single-cell transcriptomics bioinformatics and computational challenges. Front Genet 2016; 7: 163. doi: 10.3389/fgene.2016.00163
8. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell 2021; 184(13): 3573–7, e29. doi: 10.1016/j.cell.2021.04.048
9. Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell 2019; 177(7): 1888–902, e21. doi: 10.1016/j.cell.2019.05.031
10. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018; 36(5): 411–20. doi: 10.1038/nbt.4096
11. The Human Protein Atlas. Stockholm: Affinity proteomics, 2023. https://www.proteinatlas.org/ (18. 11. 2023)
12. EMBL-EBI. Single cell expression atlas. Hinxton: European Molecular Biology Laboratory, 2023. https://www.ebi.ac.uk/gxa/sc/home (5. 12. 2023)
13. 10x Genomics Datasets. Pleasanton: 10x Genomics, 2023. https://www.10xgenomics.com/datasets?query=&page=1&configure%5BhitsPerPage%5D=50&configure%5BmaxValuesPerFacet%5D=1000 (3. 7. 2023)
14. R Foundation. The R project for statistical computing. Wien: The R Foundation, 2023. https://www.r-project.org/ (3. 7. 2023)
15. McGinnis CS, Murrow LM, Gartner ZJ. Doubletfinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst 2019; 8(4): 329–37, e4. doi: 10.1016/j.cels.2019.03.003
16. Subramanian A, Alperovich M, Yang Y, Li B. Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. Genome Biol 2022; 23(1): 267. doi: 10.1186/s13059-022-02820-w
17. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol 2019; 20(1): 296. doi: 10.1186/s13059-019-1874-1
18. Zappia L, Oshlack A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. Gigascience 2018; 7(7): giy083. doi: 10.1093/gigascience/giy083
19. Waltman L, van Eck NJ. A smart local moving algorithm for large-scale modularity-based community detection. Eur Phys J B. 2013; 86(11): 471. doi: 10.1140/epjb/e2013-40829-0
20. Lu S, Li J, Song C, Shen K, Tseng GC. Biomarker detection in the integration of multiple multi-class genomic studies. Bioinformatics 2010; 26(3): 333–40. doi: 10.1093/bioinformatics/btp669
21. Uhlen M, Karlsson MJ, Zhong W, et al. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. Science 2019 ; 20; 366(6472): eaax9198. doi: 10.1126/science.aax9198
22. Monaco G, Lee B, Xu W, et al. RNA-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types. Cell Rep 2019; 26(6): 1627–40, e7. doi: 10.1016/j.celrep.2019.01.041
23. Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 2017; 14(10): 979–82. doi: 10.1038/nmeth.4402
24. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014; 32(4): 381–6. doi: 10.1038/nbt.2859
25. Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature 2019; 566(7745): 496–502. doi: 10.1038/s41586-019-0969-x
26. Li X, Wang CY. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci 2021; 13(1): 36. doi: 10.1038/s41368-021-00146-0
27. Lähnemann D, Köster J, Szczurek E, et al. Eleven grand challenges in single-cell data science. Genome Biol 2020; 21(1): 31. doi: 10.1186/s13059-020-1926-6
28. Pasquini G, Rojo Arias JE, Schäfer P, Busskamp V. Automated methods for cell type annotation on scRNA-seq data. Comput Struct Biotechnol J 2021; 19: 961–9. doi: 10.1016/j.csbj.2021.01.015
29. Abdelaal T, Michielsen L, Cats D, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol 2019; 20: 194. doi: 10.1186/s13059-019-1795-z
30. Cheng Y, Fan X, Zhang J, Li Y. A scalable sparse neural network framework for rare cell type annotation of single-cell transcriptome data. Commun Biol 2023; 6: 545. doi: 10.1038/s42003-023-04928-6
31. Flórez-Grau G, Escalona JC, Lacasta-Mambo H, et al. Human dendritic cell subset isolation by magnetic bead sorting: a protocol to efficiently obtain pure populations. Bio Protoc 2023; 13(20): e4851. doi: 10.21769/BioProtoc.4851
32. Nishide M, Nishimura K, Matsushita H, et al. Single-cell multi-omics analysis identifies two distinct phenotypes of newly-onset microscopic polyangiitis. Nat Commun 2023; 14(1): 5789. doi: 10.1038/s41467-023-41328-0
33. Bonne-Année S, Bush MC, Nutman TB. Differential Modulation of Human Innate Lymphoid Cell (ILC) Subsets by IL-10 and TGF-β. Sci Rep. 20191004th ed. 2019 Oct;9(1):14305.

34. Bej S, Galow AM, David R, Wolfien M, Wolkenhauer O. Automated annotation of rare-cell types from single-cell RNA-sequencing data through synthetic oversampling. BMC Bioinformatics 2021; 22(1): 557. doi: 10.1186/s12859-021-04469-x

35. Guo H, Li J. scSorter: assigning cells to known cell types according to marker genes. Genome Biol 2021; 22(1): 69. doi: 10.1186/s13059-021-02281-7

36. Andreatta M, Corria-Osorio J, Müller S, Cubas R, Coukos G, Carmona SJ. Interpretation of T cell states from single-cell transcriptomics data using reference atlases. Nat Commun 2021; 12(1): 2965. doi: 10.1038/s41467-021-23324-4

37. Bendall SC, Davis KL, Amir el-AD, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell. 2014; 157(3): 714–25. doi: 10.1016/j.cell.2014.04.005

38. Yao C, Sun HW, Lacey NE, et al. Single-cell RNA-seq reveals TOX as a key regulator of CD8+ T cell persistence in chronic infection. Nat Immunol 2019; 20(7): 890–901. doi: 10.1038/s41590-019-0403-4

39. Wan H, Chen L, Deng M. scEMAIL: universal and source-free annotation method for scRNA-seq data with novel cell-type perception. Genomics Proteomics Bioinformatics 2022; 20(5): 939–58. doi: 10.1016/j.gpb.2022.12.008

40. Ji X, Tsao D, Bai K, Tsao M, Xing L, Zhang X. scAnnotate: an automated cell-type annotation tool for single-cell RNA-sequencing data. Bioinform Adv 2023; 3(1): vbad030. doi: 10.1093/bioadv/vbad030

41. Nguyen V, Griss J. scAnnotatR: framework to accurately classify cell types in single-cell RNA-sequencing data. BMC Bioinformatics 2022; 23(1): 44. doi: 10.1186/s12859-022-04574-5

42. Yao Z, Liu H, Xie F, et al. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. Nature 2021; 598(7879): 103–10. doi: 10.1038/s41586-021-03500-8

Primerjalna analiza referenčno osnovanega mapiranja celičnih tipov in ročne anotacije pri analizi sekvenciranja RNA posamezne celice

L. Goričan, B. Gole, G. Jezernik, G. Krajnc, U. Potočnik, M. Gorenjak

Izvleček: Sekvenciranje RNA v posamezni celici (scRNA-seq) omogoča edinstven vpogled v celično raznolikost kompleksnih tkiv, kot so mononuklearne celice periferne krvi (PBMC). Dodatno je diferencialno izražanje genov na ravni posameznih celic lahko osnova za razumevanje specializiranih vlog posameznih celic in celičnih tipov v bioloških procesih in bolezenskih mehanizmih. Zaradi velike kompleksnosti pa je točna določitev celičnih tipov v zbirkah podatkov scRNA-seq zahtevna. V članku primerjamo dve strategiji določanja celičnih tipov, ki se uporabljata za PBMC v zbirkah podatkov scRNA-seq: avtomatizirano, na referenčnih bazah podatkov temelječe orodje »Azimuth« in nenadzorovano razvrščanje v grozde »Shared Nearest Neighbour« (SNN), ki mu sledi ročno določanje celičnih tipov. Naši rezultati poudarijo prednosti in omejitve obeh pristopov. »Azimuth« je zlahka obdelal obsežne podatkovne nize scRNAseq in zanesljivo prepoznal tudi razmeroma redke populacije celic. Imel pa je težave s celičnimi tipi izven svojega referenčnega območja. Nasprotno je nenadzorovano razvrščanje SNN jasno razmejilo vse različne celične populacije v vzorcu. Metoda SNN je zato zelo primerna za prepoznavanje redkih ali novih tipov celic, vendar zahteva dolgotrajno ročno določanje celičnih tipov, ki je nagnjeno k pristranskosti. S strogimi merili in skupnim strokovnim znanjem več neodvisnih ocenjevalcev smo to pristranskost minimalizirali. Naše ročno določanje celičnih tipov je tako le malo odstopalo od avtomatiziranega. Nazadnje je veljavnost določitve celičnih tipov z orodjem »Azimuth« in ročno metodo potrdila še psevdočasovna analiza glavnih celičnih tipov. Naša raziskava tako poudarja nujo po kombiniranju različnih pristopov razvrščanja in določanja celičnih populacij za izboljšanje zanesljivosti in globine analiz scRNA-seq.

Ključne besede: transkriptomika posamezne celice; mononuklearne celice periferne krvi; referenčno mapiranje; anotacija celičnih tipov; imunski sistem