SifiNet: A robust and accurate method to identify feature gene sets and annotate cells

Qi Gao*[1], Zhicheng Ji†[1], Liuyang Wang‡[2], Kouros Owzar§[1], Qi-Jing Li¶[3,4], Cliburn Chan‖[1], and Jichun Xie**[1,5]

[1]Department of Biostatistics and Bioinformatics, Duke University, USA
[2]Department of Molecular Genetics and Microbiology, Duke University, USA
[3]Institute of Molecular and Cell Biology, Agency for Science, Technology and Research, Singapore
[4]Singapore Immunology Network, Agency for Science, Technology and Research, Singapore
[5]Department of Mathematics, Duke University, USA

Abstract

SifiNet is a robust and accurate computational pipeline for identifying distinct gene sets, extracting and annotating cellular subpopulations, and elucidating intrinsic relationships among these subpopulations. Uniquely, SifiNet bypasses the cell clustering stage, commonly integrated into other cellular annotation pipelines, thereby circumventing potential inaccuracies in clustering that may compromise subsequent analyses. Consequently, SifiNet has demonstrated superior performance in multiple experimental datasets compared with other state-of-the-art methods. SifiNet can analyze both single-cell RNA and ATAC sequencing data, thereby rendering comprehensive multi-omic cellular profiles. It is conveniently available as an open-source R package.

*qi.gao@duke.edu
†zhicheng.ji@duke.edu
‡liuyang.wang@duke.edu
§kouros.owzar@duke.edu
¶li_qijing@imcb.a-star.edu.sg
‖cliburn.chan@duke.edu
**jichun.xie@duke.edu

Introduction

Single-cell sequencing techniques, including single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq), allow researchers to quantify cellular omic phenotypes. An ideal single-cell data analysis is expected to help researchers understand cellular omic heterogeneity, extract the cell subpopulations of interest, identify feature gene sets corresponding to cell subpopulations, as well as reveal the relationship between cell subpopulations.

Among these analysis tasks, identifying feature gene sets is a crucial step. Feature gene sets are defined as gene sets that are differentially expressed between cell subpopulations. They are often used to annotate cell subpopulations and to perform gene set enrichment analysis. Existing feature gene identification methods often employ a two-step approach (called two-step methods thereafter): cells are first clustered (such as Seurat[1–4], simple Louvain[5], Clustering through Imputation and Dimensionality Reduction (CIDR)[6], and SCANPY[7]) and differentially expressed gene (DEG) analysis (such as DESeq2[8], edgeR[9–13], limma[14,15], limma-voom[16], and MAST[17]) is then performed across cell clusters to identify cluster-specific feature genes. However, this approach has questionable accuracy on data with complex or subtle heterogeneity, since an inaccurate initial clustering step could lead to subsequent erroneous feature gene identification[18]. Alternatively, some methods identify feature genes by detecting highly variable genes (HVG) based on their deviations from the population with respect to deviance of model fitting[19], dropout rates[20], and UMI count distributions[21] (called HVG methods thereafter). However, these methods do not separate feature genes into subpopulation-specific gene sets, limiting their utility for annotating cells.

To overcome the limitations of existing methods, we present SifiNet, a unique approach for direct identification of feature gene sets that eliminates the need for prior cell clustering. Stemming from the crucial observation that genes co-differentially expressed within a cell subpopulation also exhibit co-expression patterns (Suppl. Note 1), SifiNet builds a gene co-expression network and inspects its topology to discern feature gene sets. These gene sets serve to compute cellular gene set enrichment scores and subsequently annotate cells. Additionally, networks among these gene sets imply relationships between cell subpopulations (Fig. 1). Furthermore, SifiNet can optionally integrate scATAC-seq data, as it forms a gene co-open-chromatin network and explores its topology to determine epigenomic feature gene sets. The ability of SifiNet to analyze both scRNA-seq and scATAC-seq data endows researchers with insight into cellular multiomic heterogeneity.

We demonstrate that SifiNet outperforms existing two-step methods and HVG methods in identifying feature gene sets and enhancing cell annotation accuracy. Moreover, we illustrate that SifiNet can identify complex heterogeneity among cells and reveal potential developmental lineages among cell subpopulations. SifiNet can also be scaled to analyze datasets with millions of cells. We applied SifiNet to five published experimental datasets and uncovered several potentially novel discoveries, such as potential new cell cycle markers and senescence markers, the senescent-cell enriched subpopulation, the developmental lineage of myeloid progenitor cells, and the CD8 cell subpopulations along with their possible transition paths.

Materials and Methods

Computation Pipeline of SifiNet

[Figure 1 here]

Figure 1: **SifiNet pipeline.** SifiNet takes the preprocessed feature count matrix as input, and uses gene co-expression network topology to identify feature gene sets (function 1). The identified feature gene sets are used to annotate cells (function 2), and the gene co-expression network is also used to reveal cell subpopulations' transition or developmental relationships (function 3). The bottom row illustrates the main steps for identifying feature gene sets. After multiple testing on quantile associations and removing problematic low-read-count genes, SifiNet obtains a large gene co-expression network. Then, based on the 1st, 2nd, and 3rd order connectivities, SifiNet identifies feature genes marked as red nodes. Then, SifiNet focuses on the positive co-expression network among feature genes and their node topologies to identify core feature genes, marked as blue nodes; the non-core features are marked in yellow. Next, SifiNet clustered the core feature genes into different clusters, and then assigned non-core feature genes and multi-role feature genes to the corresponding clusters. Finally, SifiNet obtains multiple feature gene sets.

SifiNet is a versatile pipeline with multiple functions, including the identification of feature gene sets, recognition of cell subpopulations, and exploration of potential developmental lineage among cell subpopulations (Fig. 1). The main innovative step is to use gene co-expression network topology to identify feature gene sets. Feature genes are genes over-expressed within a cell subpopulation; those under-expressed genes are treated as the feature genes of the remaining cells due to identifiability issues. A set of these genes, exhibiting co-over-expression within the same cell subpopulation, constitutes a feature gene set.

To identify feature gene sets, SifiNet first constructs a gene co-expression network based on quantile association measures. Then, based on topology measures, we choose the genes that form close communities. These genes are likely to be feature genes. Next, we focus on the positive co-expression network of those genes and their community patterns to cluster genes into multiple core feature gene sets. These core feature gene sets will be further enriched by adding additional genes, such as shared feature genes and genes performing multi-role functions. Based on the feature gene sets, we calculate the cellular gene set enrichment scores and annotate cells. Finally, we construct a cell subpopulation network to reveal the potential developmental or transitional relationship among cell subpopulations.

Let Y_{ic} be the feature values of gene i in cell c. For scRNA-seq data, the gene feature value could be the gene UMI count; for scATAC-seq data, it could be the aggregated gene-level cut number. Based on the gene feature values, we construct a gene co-expression network among all the genes. Each node represents a gene, and each edge represents the co-expression between two genes. Let s_c be cell c's library size (such as total read counts). We use quantile associations to define gene co-expressions.

$$D_{ij} = \frac{1}{N} \sum_c P(Y_{ic} \leq q_{ic}, Y_{jc} \leq q_{jc} \mid s_c) - \tau_i \tau_j,$$

where q_{ic} and q_{jc} are the τ_i-th and τ_j-th conditional quantiles of gene i and gene j's expressions. s_c can be further extended to include other cell-specific covariates. Our previous study

shows that quantile associations can effectively measure general dependence, especially for count data such as RNA-seq data[22]. In the gene-co-expression network, if $D_{ij} \neq 0$, there is an edge between nodes i and j.

Here, we establish a standardized estimator S_{ij} for the co-expression D_{ij}:

$$S_{ij} = \frac{\sum_{c=1}^{N}[I(Y_{ic} \leq \hat{q}_{ic}, Y_{jc} \leq \hat{q}_{jc}) - \hat{\tau}_i\hat{\tau}_j]}{\sqrt{N\hat{\tau}_i(1-\hat{\tau}_i)\hat{\tau}_j(1-\hat{\tau}_j)}},$$

where $\hat{q}_{ic}$ is the estimated cell specific τ_i-th conditional quantile of gene i, and $\hat{\tau}_i = \sum_{c=1}^{N} I(Y_{ic} \leq \hat{q}_{ic})/N$. By default, we use the quantile regression model $q_{ic} = \alpha_{i0} + \alpha_{i1}s_c$ to estimate $\hat{q}_{ic}$, where $\tau_i = (1+\hat{p}_{i0})/2$ and $\hat{p}_{i0} = \sum_{c=1}^{N} I(Y_{ic} = 0)/N$ is the proportion of zero observations of gene i. In this default setting, the estimated $\hat{q}_{ic}$ corresponds to the 50%-quantile among the non-zero observations of gene i given cell library size. However, when gene i has very low expression (when the median of Y_{ic} is equal to 0 or Y_{ic} are either 0 or 1), we set $\hat{q}_{ic} = 0$ for stability and robustness consideration.

Assume Y_{ic} and Y_{jc} are independent given a specific cell c. Suppl. Note 1 shows that if gene i or gene j is not a feature gene, then $D_{ij} = 0$, *i.e.*, these two genes are independent. Moreover, Suppl. Note 2 shows that under this case, the standardized estimator S_{ij} asymptotically follows a standard normal distribution.

In reality, even given a cell, some genes may be dependent due to their biological interactions. For example, one gene may regulate another, causing two genes to co-express. Besides, hub genes usually co-express with many other genes. However, we found that feature genes, pathway genes, and hub genes exhibit different local topologies in the co-expression network (Suppl. Fig. 1). Specifically, feature genes tend to form close communities, while pathway genes and hub genes do not.

To capture the community patterns of feature genes, we introduce first-order, second-order, and third-order connectivities – three node topology measures. Denote the gene co-expression network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with node set $\mathcal{V}$ and edge set $\mathcal{E}$. Here, $\mathcal{E} = \cup\mathcal{E}_i$, where $\mathcal{E}_i$ is gene i's edge set. For any set $\mathcal{S}$, $|\mathcal{S}|$ is its cardinality.

- First-order connectivity measures gene i's edge proportions, that is, the number of the observed edges over all possible edges: $C_{1i} = |\mathcal{E}_i|/(p-1)$.
- Second-order connectivity measures the edge proportions among gene i's neighbors: $C_{2i} = |\mathcal{E} \cap \{(j,k) : j,k \in \mathcal{N}_i\}|/|\mathcal{N}_i|(|\mathcal{N}_i|-1)$, where $\mathcal{N}_i = \{j : (i,j) \in \mathcal{E}_i\}$ is gene i's neighbor set.
- Third-order connectivity measures the edge proportions among gene i's 2nd-order neighbors (*i.e.*, neighbors' neighbors), denoted by C_{3i}. Because some genes have many 2nd-order neighbors, leading to a long computation time for C_{3i}, we employ a sampling strategy to improve the computation efficiency (Suppl. Note 3).

As we anticipate that feature genes will form communities with other feature genes, we define a gene as a feature gene if it has high C_{2i} and C_{3i} and moderate or high C_{1i}. Genes with low C_{1i}, C_{2i} and C_{3i} could be genes on the pathways or hub genes. Some feature genes may serve multi-roles as pathway genes and/or hub genes. These genes will have different topology structures from other feature genes. We will identify these multi-role feature genes in the end, after we identify other feature genes.

SifiNet function 1: Identifying feature gene sets

SifiNet rests upon a key observation: co-differentially-expressed genes show co-expression patterns, even if the expressions of those genes are conditionally independent given the cell types (Suppl. Note 1). Here we list the main steps to identify feature gene sets.

Step 1: Initiating a gene co-expression graph. Suppose after preprocessing, the single-cell dataset contains p genes and N cells. Initiate gene co-expression graph by $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with $\mathcal{V} = [p]$. To obtain $\mathcal{E}$, we set up $p(p-1)/2$ hypotheses:

$$\mathrm{H}_{\mathrm{null},ij}: \; Y_{ic} \perp\!\!\!\perp Y_{jc} \mid s_c, \quad \text{for all } i, j \in \mathcal{V} \text{ and } i < j.$$

Then we perform multiple testing on $\mathrm{H}_{\mathrm{null},ij}$ to determine the edges with FDR controlled at 5% (Suppl. Note 3).

Step 2: Eliminating problematic low-count genes and corresponding edges. Some low-count genes exhibit spurious co-expressions with many other low-count genes, and these low-count genes form a spurious community. These genes are removed from the gene co-expression network (Suppl. Note 3).

Step 3. Identifying feature genes The feature gene set $\mathcal{V}_+$ contains the genes with $C_{1i} \geq 5/(p-1)$, $C_{2i} \geq 0.4$ and $C_{3i} \geq 0.3$ (Suppl. Note 3).

Step 4. Identifying core feature genes Core feature genes are those unique feature genes of a cell subpopulation, often showing strong positive co-expressions (Suppl. Note 1). To find core feature genes, we first define a positive co-expression network $\mathcal{G}_+ = \{\mathcal{V}_+, \mathcal{E}_+\}$, where $\mathcal{E}_+$ contains the positive co-expression edges $\mathcal{E}_+ = \{(i,j) : (i,j) \in \mathcal{E}, i, j \in \mathcal{V}_+, S_{ij} > 0\}$. The 1st-, 2nd-, and 3rd-order connectivities for $\mathcal{G}_+$, denoted by $C_{+,1i}$, $C_{+,2i}$, and $C_{+,3i}$, are calculated correspondingly. By default, the core feature genes are those with $C_{+,1i} \geq 5/(|\mathcal{V}_+| - 1)$, $C_{+,2i} \geq 0.7$, and $C_{+,3i} \geq 0.5$. See Suppl. Fig. 2 for an illustration.

Step 5. Clustering core feature genes into gene sets. Taking the positive co-expression network restricted to the core feature genes $\mathcal{G}_{+,\mathrm{core}}$, we apply the Louvain algorithm to cluster these genes into K gene sets, denoted by $\mathcal{V}_k$, where $k \in [K]$.

Step 6. Assigning non-core feature genes. A non-core feature gene may be shared by multiple feature gene sets. It is added to $\mathcal{V}_k$ if it connects to more than 40% of the genes in $\mathcal{V}_k$.

Step 7. Augmenting feature gene sets with additional genes. Multi-role genes will have different connectivities from the sole feature genes. Thus, they might not be identified as feature genes in Step 3. For any gene not included in the gene sets, it will be added to $\mathcal{V}_k$ if it connects more than 90% of the feature genes in $\mathcal{V}_k$.

Some steps involve setting tuning parameters, such as the 1st-, 2nd-, and 3rd-order connectivity thresholds, and the assignment criteria for non-core feature genes, etc. The default values of these parameters are set based on our experiences in analyzing multiple experimental datasets. In Step 5, the Louvain algorithm is applied with the default resolution parameter. Because the core feature genes usually exhibit a clear community structure, different resolution parameters usually lead to the same cluster results.

SifiNet function 2: Annotating cells with feature gene sets

Based on $\mathcal{V}_k$, $k \in [K]$, we calculate the cellular gene set enrichment scores $\alpha_{c,1}, \ldots, \alpha_{c,K}$ for any cell c. For all the data analyses in this paper, we used the gene set variation analysis (GSVA)[23] scores. We define the cell subpopulation corresponding to $\mathcal{V}_k$ by

$$\mathcal{C}_k = \{c : \alpha_{c,k} > 0, \text{ and } \alpha_{c,k} \text{ is the largest among all } \alpha_{c,\ell},\ \ell \in [K]\}.$$

Those cells with all negative $\alpha_{c,k}$ will be assigned to $\mathcal{C}_{K+1}$, called the unassigned group. These cells could be in the transitional stage or simply outliers. To annotate these unassigned cells, we can further apply DEG analysis to find their corresponding feature gene set.

If the feature gene set contains many matching gene markers from known marker gene databases, we can annotate the cell subpopulation with the corresponding cell type. See the Monoclonal dataset analysis for an example. Otherwise, we refer to this cell subpopulation as new, and annotate them with the identified feature genes. See Discussions for more details.

To visualize the cell subpopulations, we derive the GSVA score matrix for all the cells and all the gene sets. Then we calculate the first two principal components (PCs) of the GSVA score matrix. The first two PCs are used to plot the cell subpopulations in a 2D space. See Figs. 3d, 3e, 4e, 4f for examples.

For data with complex heterogeneity structure, after we cluster cells based on the gene set enrichment scores, SifiNet can be applied again to a subset of cells to explore their local heterogeneity structure. By sequentially applying SifiNet, we can identify the hierarchical structure of cell subpopulations. See the BoneMarrow dataset analysis for an example.

SifiNet function 3: Constructing cell subpopulation network

By using the K feature gene sets as nodes, we can construct a gene-set-level graph representing the cell subpopulation lineage. There are a total of $K(K-1)/2$ possible edges. For edge (k, ℓ), its width is defined as edge proportions among all possible edges.

$$W_{k,\ell} = |\mathcal{E} \cap (i,j) : i \in \mathcal{V}_k, j \in \mathcal{V}_\ell| / (|\mathcal{V}_k| \cdot |\mathcal{V}_\ell|).$$

The width indicates the relationship between the two corresponding gene sets. Finally, the gene set network can be transferred to cell subpopulation network after identifying the cell subpopulation corresponding to each feature gene set. Then, a wide edge between two cell subpopulations suggests possible transitional or developmental relationships between them. By default, edges with width less than 0.3 will be removed from the visualization figure.

Competing methods

We compared the performance of SifiNet against 5 two-step methods, 3 HVG identification methods, and 1 trajectory analysis method. See Suppl. Table 1 for details of the competing methods.

The two-step methods first use clustering methods to identify cell subpopulations, and then apply DEG analysis to identify cluster-specific DEGs. They are named after the DEG methods (including DESeq2[8], edgeR[9–13], limma[14,15], limma-voom[16], and MAST[17,24]) and the clustering methods (including Seurat[1–4], simple Louvain[5], and CIDR[6]), followed

by TS. For example, DESeq2-Seurat-TS means that we first apply Seurat to cluster cells, and then use DESeq2 to identify cluster-specific DEGs. We tried all cross-over combinations of DEG and the clustering methods on SD1. CIDR utilizes both gene expression levels and dropout rates to cluster cells. It achieves the higher clustering performance in terms of adjusted rand index (ARI) for SD1 (Suppl. Fig. 3). Also, combined with DEG methods, CIDR achieves higher F1 score in feature gene identification compared with other clustering methods (Suppl. Fig. 4). Thus, we use CIDR as the default clustering method for all other analyses.

The clustering methods are applied at default resolution. The DEG methods are applied in similar ways as described in[25]. False discovery rate (FDR) is controlled at 0.05 with Benjamini-Hochberg (BH) procedure if there are p-values available in the method.

The oracle-cluster methods were also introduced to explain why SifiNet work better than the two-step methods. The oracle-cluster methods assume the ground truth of cell types, and then apply DEG analysis to identify cell-type-specific feature genes. The oracle-cluster methods include DESeq2-OC, edgeR-OC, limma-OC, limma-voom-OC, and MAST-OC.

The HVG identification methods identify feature genes through their deviation from the average gene properties. The HVG identification methods do not rely on knowing cell types or clustering. The methods we included are scry[26], M3Drop[27], and singleCellHaystack[21]. Because scry does not provide gene p-values, when we apply it to the simulated datasets, we selected the top p_1 genes where p_1 is the number of the actual feature genes.

Moreover, we also compared SifiNet with Monocle3[28–33] in their performance to construct cell subpopulation networks. Monocle3 is a comprehensive pipeline with a function to construct graph-based single-cell trajectories.

Data Sets and Analysis

To benchmark SifiNet's performance, we applied SifiNet to simulated and experimental datasets, including two simulated datasets (SD1, SD2) and five experimental datasets (Monoclonal[34], IPF1[35], BoneMarrow[36], CD8[37], MassiveRNA[38]). An additional experimental dataset IPF2[35] was used for validation. Suppl. Table 2 lists the details of the datasets.

Simulated Data

We used the R package `GeneScape`[39] to generate two numerical dataset, SD1 and SD2. `GeneScape` can simulate scRNA-seq data with customized differential expression structures, such as core and shared feature genes, sub-cell types, pathways and hub genes. The parameters used for simulation are shown in `https://github.com/jichunxie/SifiNet_manu_support`.

SD1 simulated 100 repetitions of scRNA-seq datasets with three cell types; each dataset mimics the case where a rare cell subpopulation recently developed from another non-rare cell subpopulation. SD2 simulated five equal-sized cell subpopulations. Among them, two are transitional, and the other three are ending. Transitional cell types share feature genes with their neighboring ending cell subpopulations. The cell developmental lineage has the structure: ending-transition-ending-transition-ending. See Suppl. Note 4 for detailed simulation settings. For preprocessing, for all simulated datasets, we removed the cells with less

than 10% genes expressed and genes expressed in less than 10% cells.

Experimental Data

Monoclonal. The dataset was collected from a monoclonal murine colorectal carcinoma cell line CT26WT. All cells were cloned from a single ancestral cell. We preprocessed the data in the following way. We first filter the genes similar as in the original study[34]. We then removed the cells whose mitochondrial gene expression takes more than 5% of the total gene expression, the cells whose total number of read count is greater than 1.8 times of the mean expression levels, the cells with less than 10% of the genes expressed, and the genes expressed in less than 10% of the cells. No additional normalization was performed before we applied SifiNet. This is because SifiNet constructed a gene co-expression graph based on quantile associations. The quantile regressions used to estimate quantile associations serve a similar role as quantile normalization in this step.

To curate a benchmark gene set to verify the identified genes, we did the following steps. First, we collected human cell cycle related genes using 3 libraries, Reactome[40,41], Biocarta[42], KEGG[43–45], and 32 datasets from single studies or pathway analyses (Suppl. Table 3) with "CELLCYCLE" or "CELL_CYCLE" in their name from GSEA MSigDB C2 curated human gene sets v2023.1[46–48]. Human genes that were included in the related pathway in at least 1 library or at least 3 studies were selected and converted to their mouse equivalents using `biomaRt` R package[49,50]. Second, for the Uniprot database, we downloaded the Uniprot gene annotations on May 1, 2023, and selected those genes whose "Biological process" annotations[51,52] include at least 1 of the following keywords: mitotic, centriole, centrosome, centromere, spindle, chromosome separation, cohesin, DNA replication, DNA repair, microtubule, nuclear chromosome segregation, cell cycle, cell division, cell growth. Finally, the intersection of selected genes in these two databases was chosen as the benchmark cell cycle set (475 genes). Among them, 437 genes remained in the Monoclonal data after preprocessing.

IPF1 and IPF2. These two datasets were collected from the existing study on idiopathic pulmonary fibrosis (IPF)[35]. IPF1 contains the pre-annotated "Ciliate" cells in IPF patients. IPF2 contains the pre-annotated "KRT5-/KRT17+" epithelial cells and the "Proliferating Epithelial Cells". For preprocessing, we removed the cells with less than 5% genes expressed, and the genes expressed in less than 5% cells. No additional normalization was performed before we applied SifiNet.

Next, only for DEG analysis, cell counts were normalized by `sctransform`[53] in Seurat and the log fold changes and p-values are calculated using Seurat `FindMarkers`[8,24,29,54].

BoneMarrow. BoneMarrow was collected from an existing study on myeloid differentiation[36]. We focus on the pre-annotated "Unsorted myeloid" cells. Because BoneMarrow contains multiple batches of single-cell data, we used `sctransform` to regress out the batch effect by including a batch variable. Then, we removed the cells with fewer than 500 total UMI counts, and the genes expressed in fewer than 2% of the cells.

CD8. CD8 is a subset of a 10x scRNA-seq and scATAC-seq multiome datasets[37]. An annotation of cell types is provided by Seurat. CD8 dataset contains cells with "CD8 Naïve", "CD8 TEM_1" and "CD8 TEM_2". In scATAC-seq data, peaks were aggregated to represent the epigenomic features of the corresponding gene if they are located within the 5000-base-pair upstream of the transcription starting site (TSS) of gene. Only the peaks annotated by one gene are considered. Only genes available in both scRNA-seq data and scATAC-seq data were kept. In scRNA-seq data, cells with total read count greater than 1.8 times of the mean total read count were removed. In scATAC-seq data, cells with total epigenomic feature count greater than 1.8 times of the mean total epigenomic feature count were removed. We further removed cells with fewer than 2% of the genes detected in either scRNA-seq or scATAC-seq, and then the genes expressed in fewer than 2% of the cells in either scRNA-seq or scATAC-seq. No additional normalization was performed before we applied SifiNet.

MassiveRNA. MassiveRNA is a large dataset with more than 1 million cells[38]. We removed those cells with fewer than 2% of the genes expressed, and then those genes expressed in fewer than 2% of the cells. After that, 1287071 cells and 10448 genes remain in the dataset. SifiNet was applied to the whole dataset and downsampled datasets with number of cells 10000, 20000, 50000, 100000, 200000, 400000, 600000, 800000, and 1000000. No additional normalization was performed before we applied SifiNet. The time consumption was evaluated using the total elapsed time when running SifiNet on the dataset, and the peak memory usage was evaluated using the `gc` function in R.

Results

Because SifiNet uses intrinsic gene co-expression network topology to identify feature gene sets, it is not affected by the clustering accuracy. Thus, SifiNet can effectively identify feature gene sets even when the cell subpopulation has complex heterogeneity structures and cannot be clustered well. For instance, some cell subpopulations may be transitioning to another state, some may share feature genes, and some may exhibit hierarchical structures. SifiNet can effectively handle these complex cases. By exploring multiple simulated and experimental datasets, we show that SifiNet has significantly improved accuracy in identifying feature gene sets and annotating cells compared to the competing methods, especially when cells display complex heterogeneity structures.

SifiNet identifies feature gene sets with high accuracy. We applied SifiNet to SD1, Monoclonal, and IPF1, to identify the feature gene sets. SifiNet achieved a high accuracy in identifying known feature gene sets, and also identified potential new cell cycle markers and new ciliated cell senescence markers.

First, for the 100 replicated datasets in SD1, SifiNet achieves the highest median F1 score among all methods, almost doubling the second best (Fig. 2a). Among the competing methods, singleCellHaystack, limma-CIDR-TS, limma-voom-CIDR-TS, and MAST-CIDR-TS exhibit higher median sensitivity than SifiNet, but significantly lower median specificity. High specificity is crucial in feature gene identification studies due to the large number of candidate genes. For instance, each replication of SD1 dataset contains, on average, 83

feature genes and 3318 non-feature genes. A specificity as high as 0.9 will falsely label, on average, 332 non-feature genes as feature genes, substantially obscuring the signal of the true feature genes. SifiNet's main advantage is its high specificity (close to 1), indicating that most identified genes are genuine feature genes.

[Figure 2 here]

Figure 2: **(a)-(b) Performance comparison between SifiNet and competing methods on SD1 and Monoclonal.** Sensitivity, specificity, and F1 accuracy scores of 3 types of methods: SifiNet (red), HVG methods (green), and two-step methods (blue). (a) Performance on SD1, a simulated dataset with 100 replications. (b) Performance on Monoclonal to identify cell cycle marker genes. The benchmark cell cycle marker genes were curated from existing literature (Methods). **(c) Verification of SifiNet-identified feature genes with an independent dataset.** SifiNet identified the feature genes in IPF1. Then, DEG analysis was performed on IPF2 to show their log2 fold changes between KRT5-/KRT17+ (senescence-enriched) epithelial cells and proliferating (non-senescent) epithelial cells in IPF2. Stars represent the significance level of the log2 fold change, *: adjusted p-value $<= 10^{-2}$, **: adjusted p-value $<= 10^{-4}$, ***: adjusted p-value $<= 10^{-6}$.

We further investigated why the commonly used two-step methods underperform on SD1. We compared them with the oracle methods, which assume the true cell type labels are known. All oracle-cluster methods perform well on SD1, exhibiting high sensitivity and specificity (Suppl. Fig. 4). This indicates that the key factor is the clustering accuracy. If the resulting clustering is perfect, the DEG step can effectively identify feature genes. However, in practice, exact cell types are unknown, and clustering methods can only provide a reasonable estimate. When these estimates are inaccurate, two-step methods perform poorly (Suppl. Fig. 5).

Second, we applied SifiNet to the Monoclonal dataset to identify the feature genes related to cell cycles. Because all cells were cloned from a single ancestral cell, we reasoned that the major cellular transcriptomic difference stems from their cell cycle difference. Based on the gene co-expression network topology (Suppl. Fig. 6), SifiNet identified two feature gene sets. Set 1 contains many late G1 and S (G1/S) phase genes, such as GINS, MCM, DNA polymerase genes, and RPA complex genes. Interestingly, it also includes HAUS genes (*Haus3*, *Haus5*, *Haus6*, *Haus8*). Previously, HAUS genes were thought to be related to the centrosome cycle by maintaining the integrity of the centrosome. Although many other centrosome-associated genes are considered to be markers for the G2/M phase[55], studies[56,57] indicate that centrosome integrity might also be a checkpoint for the G1 to S phase transition. Without HAUS proteins maintaining the centrosome integrity, the cells may be unable to progress to the late G1/S phase. SifiNet analysis pinpointed that the differential expression of HAUS genes is strongly correlated with other late G1/S phase feature genes. Set 2 features are composed of many well-known G2 and M (G2/M) phase genes, including Tubulin and SKA. The full list of identified genes is in Suppl. Table 4.

We compared the performance of SifiNet and the two-step methods on Monoclonal. Using a benchmark cell cycle marker gene set collected from the MSigDB[46–48] and Uniprot[58] databases, we calculated the sensitivity, specificity, and F1 score of each method. SifiNet has the highest F1 score among all methods (Fig. 2b). The second best, CIDR-DESeq2-TS, has an F1 score lower than half of SifiNet's. This is because CIDR-DESeq2-TS has identified too many irrelevant genes, leading to low specificity. In contrast, a higher proportion of SifiNet-

identified feature genes overlap with the benchmark gene set, leading to high specificity (Suppl. Fig. 7).

Third, we applied SifiNet to IPF1 to identify senescence-related feature gene sets. IPF1 contains 10,375 ciliated cells from Idiopathic Pulmonary Fibrosis (IPF) human lung tissues. Prior research has demonstrated a close relationship between senescence and IPF [59,60]. Therefore, by exploring cell-type-specific feature genes, we expect to identify feature genes associated with senescence. SifiNet identified three feature gene sets (Suppl. Table 4). Set 1 includes numerous ribosomal protein genes, known to accumulate in senescent cells [61]. Set 2 contains *Jun*, *Junb*, *Fos*, and *Fosb* from the Activator Protein-1 (AP-1) transcription factor (TF) family. Other genes in Set 2, such as *Gdf15*, *Egr1*, and *Zfp36*, have been previously found to interact with the AP-1 TF family [62–64]. This family is known to be involved in proliferation, differentiation, and tumorigenesis [65], highly related to the senescence process [66]. Set 3 genes are all mitochondrial genes. Previous studies have shown that mitochondrial genes are related to the senescence-associated secretory phenotype (SASP) [67], and mitochondrial dysfunction is associated with senescence [68]. These three gene sets may relate to different aspects of senescence.

To verify the identified senescence genes, we checked their log fold change and significance levels using an independent dataset, IPF2 (Fig. 2c). IPF2 contains two cell subpopulations: KRT5-/KRT17+ epithelial cells and proliferating epithelial cells; the former are treated as senescent cells, and the latter non-senescent. Among the 84 SifiNet identified genes, 45 are significantly differentially expressed (adjusted p-values < 0.0001, Fig. 2c) between two cell subpopulations. The analysis of independent dataset IPF2 verified the quality of the identified feature genes.

SifiNet identifies cell subpopulations with high accuracy. Here, we aim to identify cell subpopulations corresponding to different cell states, such as different cell cycle phases and senescence levels. Cells with the same functional cell type but different states only have subtle differences; thus, existing clustering methods often fail to separate them well. However, SifiNet achieves high accuracy in calling these cell subpopulations with subtle differences.

[Figure 3 here]

Figure 3: **(a)-(c) UMAPs of Monoclonal cells.** The cells were annotated by (a) SifiNet or (b) Seurat cell cycle scores, or clustered by (c) CIDR. For CIDR, the cell clusters are annotated to best map to the Seurat cell cycle score annotation. Red circles mark the major difference between the CIDR annotation and the benchmark annotation. **(d)-(f) SifiNet-annotated BoneMarrow cell subpopulations.** The x- and y-axis of (d) and (e) are the first two principal components (PCs) of the GSVA scores based on the SifiNet-identified gene sets. (d) When applied to all cells, SifiNet annotated three cell subpopulations: early progenitors, megakaryocyte/erythrocyte progenitor cells, and granulocyte/macrophage progenitor cells. (e) Applied to the granulocyte/macrophage progenitor cells, SifiNet identified four cell subpopulations: Granulocyte progenitor cells, Macrophage progenitor cells, Monocyte and dendritic progenitor cells, and T and B progenitor cells. T and B cells are not myeloid cells, but all these cells share common progenitor cells, hematopoietic stem cells. These cells may get mixed up with the myeloid progenitor cells. (f) Heatmap of some marker gene expressions in the annotated cell subpopulations.

First, we applied SifiNet to Monoclonal to call cells in different cell cycles. Cells with high enrichment scores of feature gene sets 1 and 2 were annotated as late G1/S phase and G2/M phase, respectively, by mapping the feature genes in the gene set to known cell cycle-related markers (Suppl. Table 4); the unassigned cells are annotated as early G1 phase. Fig. 3a displays the Uniform Manifold Approximation and Projection (UMAP)[28] plot of the cells annotated by SifiNet. SifiNet identified gene sets show strong within-set connectivity and relatively weaker between-set connectivity (Suppl. Fig. 6d). To set up a benchmark, we also calculated the Seurat cell cycle scores using known cell cycle markers[4,69] (Fig. 3b). Similar to SifiNet-identified feature genes, these two known cell cycle gene sets exhibit high within-set connectivity and relatively weaker between-set connectivity (Suppl. Fig. 6e). Although SifiNet does not utilize any known marker information, its annotation closely resembles the benchmark annotation (Suppl. Fig. 8). As a comparison, we applied CIDR to cluster cells into three cell clusters. The clustering results are less satisfactory: notably, CIDR clustered many early G1 and G2/M phase cells to the late G1/S phase (Fig. 3c). Overall, CIDR has a lower ARI score than SifiNet when compared with the benchmark (Fig. 3c, Suppl. Fig. 8). Furthermore, CIDR, with the default resolution parameter, yielded 8 clusters for cell clustering; the results were hard to interpret. (Suppl. Fig. 9).

Second, we applied the SifiNet to label the senescent cells in IPF1. To call the SifiNet-identified senescent cells in IPF1, we calculated the GSVA scores of the three gene sets. The 594 cells with top/bottom 40% GSVA scores of all three gene sets are labeled as SifiNet-senescent cells: top scores were used for up-regulated gene set (Gene Set 2), and bottom scores were used for down-regulated gene sets (Gene sets 1 and 3). Then, we check whether the widely known senescence marker, *Cdkn1a*[70], is differentially expressed in the SifiNet-senescent cells versus the others. It turns out that *Cdkn1a* has log2 fold change 0.636, and adjusted p-value= 1.54×10^{-34}. As a comparison, we used SenMayo, a recently published senescence marker set[71], to label the senescence cells. We chose 594 cells (same number of cells as SifiNet-senescent cells) with the top SenMayo-GSVA scores. We found that *Cdkn1a* exhibited lower differential expression in the SenMayo-senescent cells (log2 fold change 0.437, adjusted p-value= 6.66×10^{-9}). This suggests that SifiNet-identified gene sets annotated senescent ciliated cells better than the SenMayo senescence marker set.

SifiNet revealed cellular hierarchical heterogeneity. Cellular heterogeneity is often hierarchical, especially for stem-like cells. We applied SifiNet to BoneMarrow, single-cell transcriptome dataset with heterogenous cells at various development stages isolated from mouse bone marrow. And SifiNet successfully predicted the developmental lineage of the myeloid progenitor cells before they grow into the downstream cells.

Taking the gene expression data from the whole cell set, SifiNet identified two feature gene sets in BoneMarrow, S1 and S2 (Suppl. Table 4) and separated the cells into three major populations with distinct transcriptomic features: early myeloid progenitor cells, megakaryocyte/erythrocyte progenitor cells, and granulocyte/macrophage progenitor cells (Fig. 3d). S1 contains markers for lineages in the granulocyte/macrophage progenitors[72,73], such as *Cebpa*. S2 contains markers for megakaryocyte/erythrocyte progenitors[74–76], such as *Gata1*, *Gfi1b*, and *Klf1* (Fig. 3f). The cells with positive S1 (S2) GSVA scores and negative S2 (S1) GSVA scores were clustered into granulocyte/macrophage progenitor cells (megakaryocyte/erythrocyte progenitor cells); those cells with both negative S1 and S2 GSVA scores

were clustered into early progenitor cells. Early progenitor cell markers such as *Apoe* and *Gata2*[36] are significantly highly expressed in the early progenitor cells (*Apoe*: log2 fold change 1.75, adjusted p-value= 1.19×10^{-169}; *Gata2*: log2 fold change 0.887, adjusted p-value= 2.77×10^{-59}).

When applied to the annotated granulocyte/macrophage progenitor cells, SifiNet identified 3 feature gene sets S1a-S1c (Suppl. Table 4), dividing the cells into 4 subpopulations: granulocyte progenitor cells (corresponding to S1a), macrophage progenitor cells (corresponding to S1b), monocyte and dendritic progenitor cells (corresponding to S1c), and unassigned (Fig. 3e). DEG analysis shows that a marker gene for hematopoietic stem cells with the capacity to differentiate into T cells, *Cd34*[77], and memory B cell and T cell marker gene, *Cd27*[78,79], are significantly over-expressed in the unassigned subpopulation. Thus, we label this subpopulation as lymphoid progenitor cells. These cells may be mixed up with the myeloid progenitor cells when the original study sorted the cells[36].

[Figure 4 here]

Figure 4: **(a)-(c) UMAPs of SD2**, with (a) the SifiNet-built lineage, (b) the true cell subpopulation lineage, and (c) the Monocle3-built lineage. **(d)-(f) Cellular GSVA score PC plots for CD8.** The x- and y-axis are PC directions of the GSVA scores based on the SifiNet-derived gene sets using scRNA-seq and scATAC-seq data. The cells were annotated by (d) Seurat on the multiome data (e) SifiNet on the scRNA-seq data (f) SifiNet on the scATAC-seq data. Cell subpopulation networks are shown in (e) and (f).

SifiNet revealed accurate cell subpopulation network in SD2. We applied SifiNet to SD2, a simulated dataset with five cell subpopulations representing a network with five lineage stages (Fig. 4b). While all the cells are distinctly separated into different clusters, the neighboring cell subpopulations share certain feature genes, indicating a developmental transition at the feature gene level (Suppl. Note 4). SifiNet identified five gene sets, separated the cells into five corresponding subpopulations, and constructed the network. Clearly, the cell subpopulation network built by SifiNet (Fig. 4a) is identical to the actual network (Fig. 4b). As a comparison, we applied Monocle3[28–33], a popular single-cell pipeline for cell lineage analysis. The resulting cell subpopulation network (Fig. 4c) is different from the actual lineage. Thus, SifiNet outperforms Monocle3 in SD2 to reveal cell developmental lineage.

SifiNet accurately annotated CD8 T cells and identified long-lasting T cells. SifiNet is able to differentiate cells into subpopulations based on either single-cell transcriptomic or epigenomic data. Moreover, it can provide useful insights into the relationships between cell subpopulations.

In the CD8 scRNA-seq data, SifiNet identified four gene sets and annotated the cells with five cell subpopulations. Set 1 and Set 3 share *Gzma*, *Klrd1*, and *Klrg1*, markers for effector T cells[80,81]. Set 2 contains many central memory T cell (TCM) markers, including *Foxo1*, *Ccr7*, *Foxp1*, and *Tcf7*[82,83]. Set 4 contains 19 RPL and RPS genes, which are highly expressed in Naïve T cells[84]. Then, SifiNet clustered the cells into five subpopulations, TEM_1, TCM_1, TEM_2, TNaïve, and unassigned (the remaining cells), which are between

the TCM_1 and TEM_2 cells (Fig. 4e). DEG analysis shows that the expression levels of TEM and TCM markers, *Gzma*, *Klrd1*, *Klrg1*, *Foxo1*, *Ccr7*, *Foxp1*, and *Tcf7*, in the unassigned subpopulation exhibit greater similarity to the TCM cells (Suppl. Fig. 10). Thus, we annotate the unassigned subpopulation as "TCM_2". Central memory T cells (TCM_1, TCM_2) and Naïve T cells (TNaive) were previously annotated by Seurat as Naïve cells, and the TCM cells are missing based on Seurat annotations (Fig. 4d). This suggests that SifiNet has a stronger ability in differentiating cells. Further, SifiNet identified two effector T cell (TEM_1, TEM_2) subpopulations, both exhibiting bridging and transitional characteristics between TCM and TNaive cells (Fig. 4e). These results suggest CD8 subpopulation lineage relationships.

Previous studies used gene chromatin accessibility profiles to supplement genes' expression profiles for defining T cell subpopulations.[85–87]. We adopted the same strategy. In the CD8 scATAC-seq data, SifiNet identified two feature gene sets. Set 1 includes TCM markers, such as *Klf7* and *Foxp1* [83]. Set 2 includes TEM markers, such as *Klrg1* [80]. These gene sets separated the cells into TCM and TEM cells, which shows excessive chromatin accessibility on TCM and TEM marker genes. Only one cell is in the unassigned subpopulation, which is not shown in the UMAP. In particular, we identified some CD8 cells with a TCM cell transcriptome profile but also excessive chromatin accessibility on TEM markers (Fig. 4e,f). They were considered as the most long-lasting T cells after yellow fever vaccination in a recent study[88]. Thus, further studying this CD8 cell subpopulation may lead to new insights into long-term immune protection.

SifiNet scales to million-cell datasets. The rapid development of single-cell technology is leading to the generation of large-scale single-cell data collection, making more efficient computational methods necessary. Thus, while SifiNet primarily focuses on exploring subtle differences between similar cell subpopulations, we also applied SifiNet to a super-large-scale scRNA-seq dataset, MassiveRNA, to test its computational performance. MassiveRNA contains 10,448 genes and 1,287,071 cells after preprocessing. SifiNet took 12.2 hours and 101 GB peak memory to complete the analysis. To learn the impact of cell numbers on time cost and peak memory usage, we down-sampled MassiveRNA dataset into smaller datasets with cell numbers ranging between 10,000 to 1,000,000. SifiNet analysis on these datasets showed that the log10 running time is roughly linear with the log10 cell numbers (Suppl. Fig. 11). When the cell number is less than 50,000, the peak memory usage was approximately 4.1 GB, dominated by the multiple testing procedure. As the cell number increases above 50,000, the log10 peak memory usage was also roughly linear with the log10 cell number; the peak memory usage was reached in the co-expression calculation step (Suppl. Fig. 11). Overall, SifiNet is efficient with respect to computation time and memory usage. In general, the computation complexity of SifiNet is $O(Np^2)$, where N and p are the numbers of cells and genes after preprocessing.

Discussion. Cell subpopulation annotation and understanding their relationships pose challenges, particularly when cells exhibit complex and subtle heterogeneity in states such as developmental stages, cell cycles, and senescence levels. While traditional clustering methods perform well in annotating subpopulations when there are clear functional distinctions reflected in cellular transcriptome profiles, they are less reliable in the presence of nuanced

heterogeneity. To address these limitations, traditional methods often rely on domain-knowledge-assisted approaches. For instance, Seurat utilizes known cell cycle markers to derive cell cycle scores. However, for less understood cell statuses like senescence, the availability of marker genes may be limited. SifiNet demonstrates significant potential in these areas to identify new cell status markers and their associated cells.

Although SifiNet is computationally feasible for million-cell datasets, we do not recommend applying SifiNet directly on large-scale cell atlases, where the interested heterogeneity source is likely to be masked by other confounder heterogeneity sources. SifiNet's strength lies in identifying the leading heterogeneity source in cell subpopulations with complex heterogeneity structures. Thus, for large-scale cell atlases, we suggest using UMAP, t-distributed Stochastic Neighbor Embedding (t-SNE)[89], or clustering methods to roughly cluster cells into visually obvious cell subpopulations where the interested heterogeneity is the leading. Then, SifiNet can be applied to the cell subpopulations to extract the feature gene sets of interest.

After SifiNet identifies the feature gene sets and the corresponding enriched cell subpopulation, we can immediately use the feature gene sets to annotate the cells. We can examine the identified feature genes and map them to the known marker genes in the public knowledge database. If we find a consensus among most feature genes, we will annotate the cell subpopulation with the corresponding known cell subpopulation. Some tools can help automate this process. For example, a recent study shows that GPT-4, a multi-modality large language model, can assist human researchers in accurately annotating known cell types based on a given set of feature genes[90]. If we cannot find a consensus in mapping the feature genes, the corresponding cell subpopulation is likely new. We can then use the feature genes and their co-expression network to annotate the cell type. We can also study the cell subpopulation network to understand the relationships between the cell subpopulations and other known cell subpopulations.

Moving forward, there are several avenues for further methodological improvement. Firstly, SifiNet's performance is influenced by the quality of the gene co-expression network. If the interested cell population is very small, say, taking less than 5% of the total population, the current procedure to estimate the co-expression network might not be accurate enough for downstream analysis. Thus, we need new methods to enhance co-expression matrix estimation. Secondly, SifiNet utilizes the Louvain method with default resolution parameters to cluster feature genes, which generally works well when gene sets are well-separated. However, for gene sets with less distinct boundaries, optimizing gene set clustering becomes crucial to further enhance SifiNet's performance on more challenging datasets. Thirdly, the incorporation of domain-knowledge-assisted approaches into SifiNet is being considered. This would allow SifiNet to leverage partially known marker genes, accurately capturing more feature genes and identifying corresponding cell subpopulations.

Data availability

Monoclonal dataset is available at the Genome Sequencing Archive (GSA) with accession CRA008966. IPF1 and IPF2 datasets are collected from study GSE135893. Bone-Marrow dataset is collected from study GSE72857. CD8 dataset is available at 10x geneomics website `https://support.10xgenomics.com/single-cell-multiome-atac-gex/`

datasets/1.0.0/pbmc_granulocyte_sorted_10k, and is also included in SeuratData R package as pbmc.rna and pbmc.atac. Seurat annotations for CD8 are available in SeuratData R package v0.2.2. MassiveRNA is downloaded from 10x genomics website https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons.

Code availability

SifiNet v1.9 is used for the analysis in this study. SifiNet package is available at https://github.com/jichunxie/SifiNet. The codes used for generating the results shown in this paper are available at https://github.com/jichunxie/SifiNet_manu_support. GeneScape v1.0 used for generating numerical datasets is available on R CRAN.

Acknowledgement

The authors thank the Duke CHSI for providing the computation resources.

Funding

QG, ZJ, CC, and JX's research was partially supported by NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director under awards, 1U54AG075936-01 from the National Institute of Aging. LW's research was supported by Duke University. QJL's research was supported by NIH Award R33 CA225328 (early stage of this project) and core funds from IMCB and SIgN, A*STAR. In addition, QG, KO and JX's research was partially supported by the NIH Award 1R01HG012555-01.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, 3rd, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M Fleming, Bertrand Yeung, Angela J Rogers, Juliana M McElrath, Catherine A Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, Jun 2021. doi: 10.1016/j.cell.2021.04.048. PMCID: PMC8238499.

[2] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, Jun 2019. doi: 10.1016/j.cell.2019.05.031. PMCID: PMC6687398.

[3] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and

species. *Nat Biotechnol*, 36(5):411–420, Jun 2018. doi: 10.1038/nbt.4096. PMCID: PMC6700744.

[4] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*, 33(5):495–502, May 2015. doi: 10.1038/nbt.3192. PMCID: PMC4430369.

[5] 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):10008, October 2008. doi: 10.1088/1742-5468/2008/10/P10008.

[6] Peijie Lin, Michael Troup, and Joshua W K Ho. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome Biol*, 18(1):59, Mar 2017. doi: 10.1186/s13059-017-1188-0. PMCID: PMC5371246.

[7] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol*, 19(1):15, Feb 2018. doi: 10.1186/s13059-017-1382-0. PMCID: PMC5802054.

[8] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):550, 2014. doi: 10.1186/s13059-014-0550-8. PMCID: PMC4302049.

[9] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, Jan 2010. doi: 10.1093/bioinformatics/btp616. PMCID: PMC2796818.

[10] Aaron T L Lun, Yunshun Chen, and Gordon K Smyth. It's de-licious: A recipe for differential expression analyses of rna-seq experiments using quasi-likelihood methods in edger. *Methods Mol Biol*, 1418:391–416, 2016. doi: 10.1007/978-1-4939-3578-9_19. PMID: 27008025.

[11] Yunshun Chen, Aaron T L Lun, and Gordon K Smyth. Differential expression analysis of complex rna-seq experiments using edger, 1 2014.

[12] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Res*, 40(10):4288–97, May 2012. doi: 10.1093/nar/gks042. PMCID: PMC3378882.

[13] Yunshun Chen, Aaron T L Lun, and Gordon K Smyth. From reads to genes to pathways: differential expression analysis of rna-seq experiments using rsubread and the edger quasi-likelihood pipeline. *F1000Res*, 5:1438, June 2016. doi: 10.12688/f1000research.8987.2. PMCID: PMC4934518.

[14] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res*, 43(7):e47, Apr 2015. doi: 10.1093/nar/gkv007. PMCID: PMC4402510.

[15] Belinda Phipson, Stanley Lee, Ian J Majewski, Warren S Alexander, and Gordon K Smyth. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann Appl Stat*, 10(2):946–963, Jun 2016. doi: 10.1214/16-AOAS920. PMCID: PMC5373812.

[16] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol*, 15(2):R29, Feb 2014. doi: 10.1186/gb-2014-15-2-r29. PMCID: PMC4053721.

[17] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, Peter S Linsley, and Raphael Gottardo. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biol*, 16:278, Dec 2015. doi: 10.1186/s13059-015-0844-5. PMCID: PMC4676162.

[18] Lucy L. Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, pages 1–11, Aug 2022. doi: 10.1080/01621459.2022.2116331.

[19] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome Biol*, 20(1):295, Dec 2019. doi: 10.1186/s13059-019-1861-6. PMCID: PMC6927135.

[20] Tallulah S Andrews and Martin Hemberg. M3drop: dropout-based feature selection for scrnaseq. *Bioinformatics*, 35(16):2865–2867, Aug 2019. doi: 10.1093/bioinformatics/bty1044. PMCID: PMC6691329.

[21] Alexis Vandenbon and Diego Diez. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nat Commun*, 11(1):4318, Aug 2020. doi: 10.1038/s41467-020-17900-3. PMCID: PMC7455704.

[22] Jichun Xie and Ruosha Li. False discovery rate control for high dimensional networks of quantile associations conditioning on covariates. *J R Stat Soc Series B Stat Methodol*, 80(5):1015–1034, Nov 2018. doi: 10.1111/rssb.12288. PMCID: PMC6497089.

[23] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC Bioinformatics*, 14:7, Jan 2013. doi: 10.1186/1471-2105-14-7. PMCID: PMC3618321.

[24] Andrew McDavid, Greg Finak, and Masanao Yajima. *MAST: Model-based Analysis of Single Cell Transcriptomics*, 2021. R package version 1.20.0.

[25] Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods*, 15(4):255–261, Apr 2018. doi: 10.1038/nmeth.4612. PMID: 29481549.

[26] Kelly Street, F. William Townes, Davide Risso, and Stephanie Hicks. *scry: Small-Count Analysis Methods for High-Dimensional Data*, 2021. R package version 1.6.0.

[27] Tallulah Andrews. *M3Drop: Michaelis-Menten Modelling of Dropouts in single-cell RNASeq*, 2021. R package version 1.20.0.

[28] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, September 2020.

[29] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*, 32(4):381–386, Apr 2014. doi: 10.1038/nbt.2859. PMCID: PMC4122333.

[30] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods*, 14(10):979–982, Oct 2017. doi: 10.1038/nmeth.4402. PMCID: PMC5764547.

[31] 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*, 566(7745):496–502, Feb 2019. doi: 10.1038/s41586-019-0969-x. PMCID: PMC6434952.

[32] V A Traag, L Waltman, and N J van Eck. From louvain to leiden: guaranteeing well-connected communities. *Sci Rep*, 9(1):5233, Mar 2019. doi: 10.1038/s41598-019-41695-z. PMCID: PMC6435756.

[33] Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, El-ad D Amir, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, Rachel Finck, Amanda L Gedman, Ina Radtke, James R Downing, Dana Pe'er, and Garry P Nolan. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–97, Jul 2015. doi: 10.1016/j.cell.2015.05.047. PMCID: PMC4508757.

[34] Jiyuan Fang, Cliburn Chan, Kouros Owzar, Liuyang Wang, Diyuan Qin, Qi-Jing Li, and Jichun Xie. Clustering deviation index (cdi): a robust and accurate internal measure for evaluating scrna-seq data clustering. *Genome Biol*, 23(1):269, Dec 2022. doi: 10.1186/s13059-022-02825-5. PMCID: PMC9793368.

[35] Arun C Habermann, Austin J Gutierrez, Linh T Bui, Stephanie L Yahn, Nichelle I Winters, Carla L Calvi, Lance Peter, Mei-I Chung, Chase J Taylor, Christopher Jetter, Latha Raju, Jamie Roberson, Guixiao Ding, Lori Wood, Jennifer M S Sucre, Bradley W Richmond, Ana P Serezani, Wyatt J McDonnell, Simon B Mallal, Matthew J Bacchetta, James E Loyd, Ciara M Shaver, Lorraine B Ware, Ross Bremner, Rajat Walia, Timothy S Blackwell, Nicholas E Banovich, and Jonathan A Kropski. Single-cell rna sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in

pulmonary fibrosis. *Sci Adv*, 6(28):eaba1972, Jul 2020. doi: 10.1126/sciadv.aba1972. PMCID: PMC7439444.

[36] Franziska Paul, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, Eyal David, Nadav Cohen, Felicia Kathrine Bratt Lauridsen, Simon Haas, Andreas Schlitzer, Alexander Mildner, Florent Ginhoux, Steffen Jung, Andreas Trumpp, Bo Torben Porse, Amos Tanay, and Ido Amit. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7):1663–77, Dec 2015. doi: 10.1016/j.cell.2015.11.013. PMID: 26627738.

[37] 10x Genomics. Pbmc from a healthy donor - granulocytes removed through cell sorting (10k), 2020. Single Cell Multiome ATAC + Gene Exp. Dataset by Cell Ranger ARC 1.0.0.

[38] 10x Genomics. 1.3 million brain cells from e18 mice, 2017. Single Cell Gene Expression Dataset by Cell Ranger 1.3.0.

[39] Qi Gao. Genescape: Simulation of single cell rna-seq data with complex structure, 2023. R package version 1.0.

[40] Bijay Jassal, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, Justin Cook, Marc Gillespie, Robin Haw, Fred Loney, Bruce May, Marija Milacic, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Res*, 48(D1):D498–D503, Jan 2020. doi: 10.1093/nar/gkz1031. PMCID: PMC7145712.

[41] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman, Xiang Shao, Antonio Fabregat, Konstantinos Sidiropoulos, Julieth Murillo, Guilherme Viteri, Justin Cook, Solomon Shorser, Gary Bader, Emek Demir, Chris Sander, Robin Haw, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase 2022. *Nucleic Acids Res*, 50(D1):D687–D692, Jan 2022. doi: 10.1093/nar/gkab1028. PMCID: PMC8689983.

[42] Darryl Nishimura. Biocarta. *Biotech Software & Internet Report*, 2(3):117–120, 2023/05/17 2001. doi: 10.1089/152791601750294344.

[43] M Kanehisa and S Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000. doi: 10.1093/nar/28.1.27. PMCID: PMC102409.

[44] Minoru Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Sci*, 28(11):1947–1951, Nov 2019. doi: 10.1002/pro.3715. PMCID: PMC6798127.

[45] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*, 51(D1):D587–D592, Jan 2023. doi: 10.1093/nar/gkac963. PMCID: PMC9825424.

[46] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43): 15545–50, Oct 2005. doi: 10.1073/pnas.0506580102. PMCID: PMC1239896.

[47] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–40, Jun 2011. doi: 10.1093/bioinformatics/btr260. PMCID: PMC3106198.

[48] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database (msigdb) hallmark gene set collection. *Cell Syst*, 1(6):417–425, Dec 2015. doi: 10.1016/j.cels.2015.12.004. PMCID: PMC4707969.

[49] Steffen Durinck, Paul T Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nat Protoc*, 4(8):1184–91, 2009. doi: 10.1038/nprot.2009.97. PMCID: PMC3159387.

[50] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–40, Aug 2005. doi: 10.1093/bioinformatics/bti525. PMID: 16082012.

[51] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, May 2000. doi: 10.1038/75556. PMCID: PMC3037419.

[52] Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil Dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith

Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The gene ontology knowledgebase in 2023. *Genetics*, 224(1), May 2023. doi: 10.1093/genetics/iyad031. PMCID: PMC10158837.

[53] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, 2019. doi: 10.1186/s13059-019-1874-1.

[54] Andrew McDavid, Greg Finak, Pratip K Chattopadyay, Maria Dominguez, Laurie Lamoreaux, Steven S Ma, Mario Roederer, and Raphael Gottardo. Data exploration, quality control and testing in single-cell qpcr-based gene expression experiments. *Bioinformatics*, 29(4):461–7, Feb 2013. doi: 10.1093/bioinformatics/bts714. PMCID: PMC3570210.

[55] Yingmei Wang, Ping Ji, Jinsong Liu, Russell R Broaddus, Fengxia Xue, and Wei Zhang. Centrosome-associated regulators of the g(2)/m checkpoint as targets for cancer therapy. *Mol Cancer*, 8:8, Feb 2009. doi: 10.1186/1476-4598-8-8. PMCID: PMC2657106.

[56] Arianne Heinrichs. A centrosome-integrity checkpoint. *Nature Reviews Molecular Cell Biology*, 8(2):98–98, February 2007. doi: 10.1038/nrm2107.

[57] Dorothy A Lerit and John S Poulton. Centrosomes are multifunctional regulators of genome stability. *Chromosome Res*, 24(1):5–17, Jan 2016. doi: 10.1007/s10577-015-9506-4. PMCID: PMC4726469.

[58] UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Res*, 51(D1):D523–D531, Jan 2023. doi: 10.1093/nar/gkac1052. PMCID: PMC9825514.

[59] D L Kellogg, D L Kellogg, Jr, N Musi, and A M Nambiar. Cellular senescence in idiopathic pulmonary fibrosis. *Curr Mol Biol Rep*, 7(3):31–40, August 2021. doi: 10.1007/s40610-021-00145-4. PMCID: PMC8358258.

[60] Kaj E C Blokland, David W Waters, Michael Schuliga, Jane Read, Simon D Pouwels, Christopher L Grainge, Jade Jaffar, Glen Westall, Steven E Mutsaers, Cecilia M Prêle, Janette K Burgess, and Darryl A Knight. Senescence of ipf lung fibroblasts disrupt alveolar epithelial cell proliferation and promote migration in wound healing. *Pharmaceutics*, 12(4), Apr 2020. doi: 10.3390/pharmaceutics12040389. PMCID: PMC7238173.

[61] Frédéric Lessard, Sebastian Igelmann, Christian Trahan, Geneviève Huot, Emmanuelle Saint-Germain, Lian Mignacca, Neylen Del Toro, Stéphane Lopes-Paciencia, Benjamin Le Calvé, Marinieve Montero, Xavier Deschênes-Simard, Marina Bury, Olga Moiseeva, Marie-Camille Rowell, Cornelia E Zorca, Daniel Zenklusen, Léa Brakier-Gingras, Véronique Bourdeau, Marlene Oeffinger, and Gerardo Ferbeyre. Senescence-associated ribosome biogenesis defects contributes to cell cycle arrest through the rb pathway. *Nat Cell Biol*, 20(7):789–799, Jul 2018. doi: 10.1038/s41556-018-0127-y. PMID: 29941930.

[62] Luc Rochette, Marianne Zeller, Yves Cottin, and Catherine Vergely. Insights into mechanisms of gdf15 and receptor gfral: Therapeutic targets. *Trends Endocrinol Metab*, 31 (12):939–951, Dec 2020. doi: 10.1016/j.tem.2020.10.004. PMID: 33172749.

[63] Akira Nakashima, Akira Ota, and Esther L Sabban. Interactions between egr1 and ap1 factors in regulation of tyrosine hydroxylase transcription. *Brain Res Mol Brain Res*, 112(1-2):61–9, Apr 2003. doi: 10.1016/s0169-328x(03)00047-0. PMID: 12670703.

[64] R Canzoneri, J Naipauer, M Stedile, A Rodriguez Peña, E Lacunza, N A Gandini, A C Curino, M M Facchinetti, O A Coso, E Kordon, and M C Abba. Identification of an ap1-zfp36 regulatory network associated with breast cancer prognosis. *J Mammary Gland Biol Neoplasia*, 25(2):163–172, Jun 2020. doi: 10.1007/s10911-020-09448-1. PMID: 32248342.

[65] Ines Garces de Los Fayos Alonso, Huan-Chang Liang, Suzanne D Turner, Sabine Lagger, Olaf Merkel, and Lukas Kenner. The role of activator protein-1 (ap-1) family members in cd30-positive lymphomas. *Cancers (Basel)*, 10(4), Mar 2018. doi: 10.3390/cancers10040093. PMCID: PMC5923348.

[66] Yuting Wang, Liping Liu, Yifan Song, Xiaojie Yu, and Hongkui Deng. Unveiling e2f4, tead1 and ap-1 as regulatory transcription factors of the replicative senescence program by multi-omics analysis. *Protein Cell*, 13(10):742–759, Oct 2022. doi: 10.1007/s13238-021-00894-z. PMCID: PMC9233726.

[67] Hélène Martini and João F Passos. Cellular senescence: all roads lead to mitochondria. *FEBS J*, 290(5):1186–1202, Mar 2023. doi: 10.1111/febs.16361. PMCID: PMC9296701.

[68] Ruchi Kumari and Parmjit Jat. Mechanisms of cellular senescence: Cell cycle arrest and senescence associated secretory phenotype. *Front Cell Dev Biol*, 9:645593, March 2021. doi: 10.3389/fcell.2021.645593. PMCID: PMC8039141.

[69] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, 2nd, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia-Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S Genshaft, Travis K Hughes, Carly G K Ziegler, Samuel W Kazer, Aleth Gaillard, Kellie E Kolb, Alexandra-Chloé Villani, Cory M Johannessen, Aleksandr Y Andreev, Eliezer M Van Allen, Monica Bertagnolli, Peter K Sorger, Ryan J Sullivan, Keith T Flaherty, Dennie T Frederick, Judit Jané-Valbuena, Charles H Yoon, Orit Rozenblatt-Rosen, Alex K Shalek, Aviv Regev, and Levi A Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352 (6282):189–96, Apr 2016. doi: 10.1126/science.aad0501. PMCID: PMC4944528.

[70] José Alberto López-Domínguez, Sandra Rodríguez-López, Ulises Ahumada-Castro, Pierre-Yves Desprez, Maria Konovalenko, Remi-Martin Laberge, César Cárdenas, José Manuel Villalba, and Judith Campisi. Cdkn1a transcript variant 2 is a marker of aging and cellular senescence. *Aging (Albany NY)*, 13(10):13380–13392, May 2021. doi: 10.18632/aging.203110. PMCID: PMC8202863.

[71] Dominik Saul, Robyn Laura Kosinsky, Elizabeth J Atkinson, Madison L Doolittle, Xu Zhang, Nathan K LeBrasseur, Robert J Pignolo, Paul D Robbins, Laura J Niedernhofer, Yuji Ikeno, Diana Jurk, João F Passos, LaTonya J Hickson, Ailing Xue, David G Monroe, Tamara Tchkonia, James L Kirkland, Joshua N Farr, and Sundeep Khosla. A new gene set identifies senescent cells and predicts senescence-associated pathways across tissues. *Nat Commun*, 13(1):4827, Aug 2022. doi: 10.1038/s41467-022-32552-1. PMCID: PMC9381717.

[72] Richard Dahl, Jonathan C Walsh, David Lancki, Peter Laslo, Sangeeta R Iyer, Harinder Singh, and M Celeste Simon. Regulation of macrophage and neutrophil cell fates by the pu.1:c/ebpalpha ratio and granulocyte colony-stimulating factor. *Nat Immunol*, 4(10): 1029–36, Oct 2003. doi: 10.1038/ni973. PMID: 12958595.

[73] Catherine V Laiosa, Matthias Stadtfeld, Huafeng Xie, Luisa de Andres-Aguayo, and Thomas Graf. Reprogramming of committed t cell progenitors to macrophages and dendritic cells by c/ebp alpha and pu.1 transcription factors. *Immunity*, 25(5):731–44, Nov 2006. doi: 10.1016/j.immuni.2006.09.011. PMID: 17088084.

[74] Hiromi Iwasaki, Shin-ichi Mizuno, Richard A Wells, Alan B Cantor, Sumiko Watanabe, and Koichi Akashi. Gata-1 converts lymphoid and myelomonocytic progenitors into the megakaryocyte/erythrocyte lineages. *Immunity*, 19(3):451–62, Sep 2003. doi: 10.1016/s1074-7613(03)00242-5. PMID: 14499119.

[75] Lothar Vassen, Hugues Beauchemin, Wafaa Lemsaddek, Joseph Krongold, Marie Trudel, and Tarik Möröy. Growth factor independence 1b (gfi1b) is important for the maturation of erythroid cells and the regulation of embryonic globin expression. *PLoS One*, 9(5): e96636, May 2014. doi: 10.1371/journal.pone.0096636. PMCID: PMC4011847.

[76] Cheng-Tao Yang, Rui Ma, Richard A Axton, Melany Jackson, A Helen Taylor, Antonella Fidanza, Lamin Marenah, Jan Frayne, Joanne C Mountford, and Lesley M Forrester.

Activation of klf1 enhances the differentiation and maturation of red blood cells from human pluripotent stem cells. *Stem Cells*, 35(4):886–897, Apr 2017. doi: 10.1002/stem.2562. PMCID: PMC5396323.

[77] C Schmitt, S Ktorza, S Sarun, M P Verpilleux, C Blanc, M A Deugnier, A Dalloul, and P Debré. Cd34-positive early stages of human t-cell differentiation. *Leuk Lymphoma*, 17(1-2):43–50, Mar 1995. doi: 10.3109/10428199509051702. PMID: 7539659.

[78] J Hendriks, L A Gravestein, K Tesselaar, R A van Lier, T N Schumacher, and J Borst. Cd27 is required for generation and long-term maintenance of t cell immunity. *Nat Immunol*, 1(5):433–40, Nov 2000. doi: 10.1038/80877. PMID: 11062504.

[79] K Agematsu, S Hokibara, H Nagumo, and A Komiyama. Cd27: a memory b-cell marker. *Immunol Today*, 21(5):204–6, May 2000. doi: 10.1016/s0167-5699(00)01605-4. PMID: 10782048.

[80] Nikhil S Joshi, Weiguo Cui, Anmol Chandele, Heung Kyu Lee, David R Urso, James Hagman, Laurent Gapin, and Susan M Kaech. Inflammation directs memory precursor and short-lived effector cd8(+) t cell fates via the graded expression of t-bet transcription factor. *Immunity*, 27(2):281–95, Aug 2007. doi: 10.1016/j.immuni.2007.07.010. PMCID: PMC2034442.

[81] Trine Sundebo Meldgaard, Fabiola Blengio, Denise Maffione, Chiara Sammicheli, Simona Tavarini, Sandra Nuti, Roland Kratzer, Duccio Medini, Emilio Siena, and Sylvie Bertholet. Single-cell analysis of antigen-specific cd8+ t-cell transcripts reveals profiles specific to mrna or adjuvanted protein vaccines. *Front Immunol*, 12:757151, 2021. doi: 10.3389/fimmu.2021.757151. PMCID: PMC8586650.

[82] Myoungjoo V Kim, Weiming Ouyang, Will Liao, Michael Q Zhang, and Ming O Li. The transcription factor foxo1 controls central-memory cd8+ t cell responses to infection. *Immunity*, 39(2):286–97, Aug 2013. doi: 10.1016/j.immuni.2013.07.013. PMCID: PMC3809840.

[83] Hua Lu, Hui Wang, Lijun Yan, Hongwei Shao, Wenfeng Zhang, Han Shen, Huaben Bo, Changli Tao, Shengfang Xia, and Fenglin Wu. Overexpression of early t cell differentiation-specific transcription factors transforms the terminally differentiated effector t cells into less differentiated state. *Cell Immunol*, 353:104118, Jul 2020. doi: 10.1016/j.cellimm.2020.104118. PMID: 32413598.

[84] Laura S Christian, Liuyang Wang, Bryan Lim, Dachuan Deng, Haiyang Wu, Xiao-Fan Wang, and Qi-Jing Li. Resident memory t cells in tumor-distant tissues fortify against metastasis formation. *Cell Rep*, 35(6):109118, May 2021. doi: 10.1016/j.celrep.2021.109118. PMCID: PMC8204287.

[85] Christopher D Scharer, Alexander P R Bally, Bhanu Gandham, and Jeremy M Boss. Cutting edge: Chromatin accessibility programs cd8 t cell memory. *J Immunol*, 198(6):2238–2243, Mar 2017. doi: 10.4049/jimmunol.1602086. PMCID: PMC5341694.

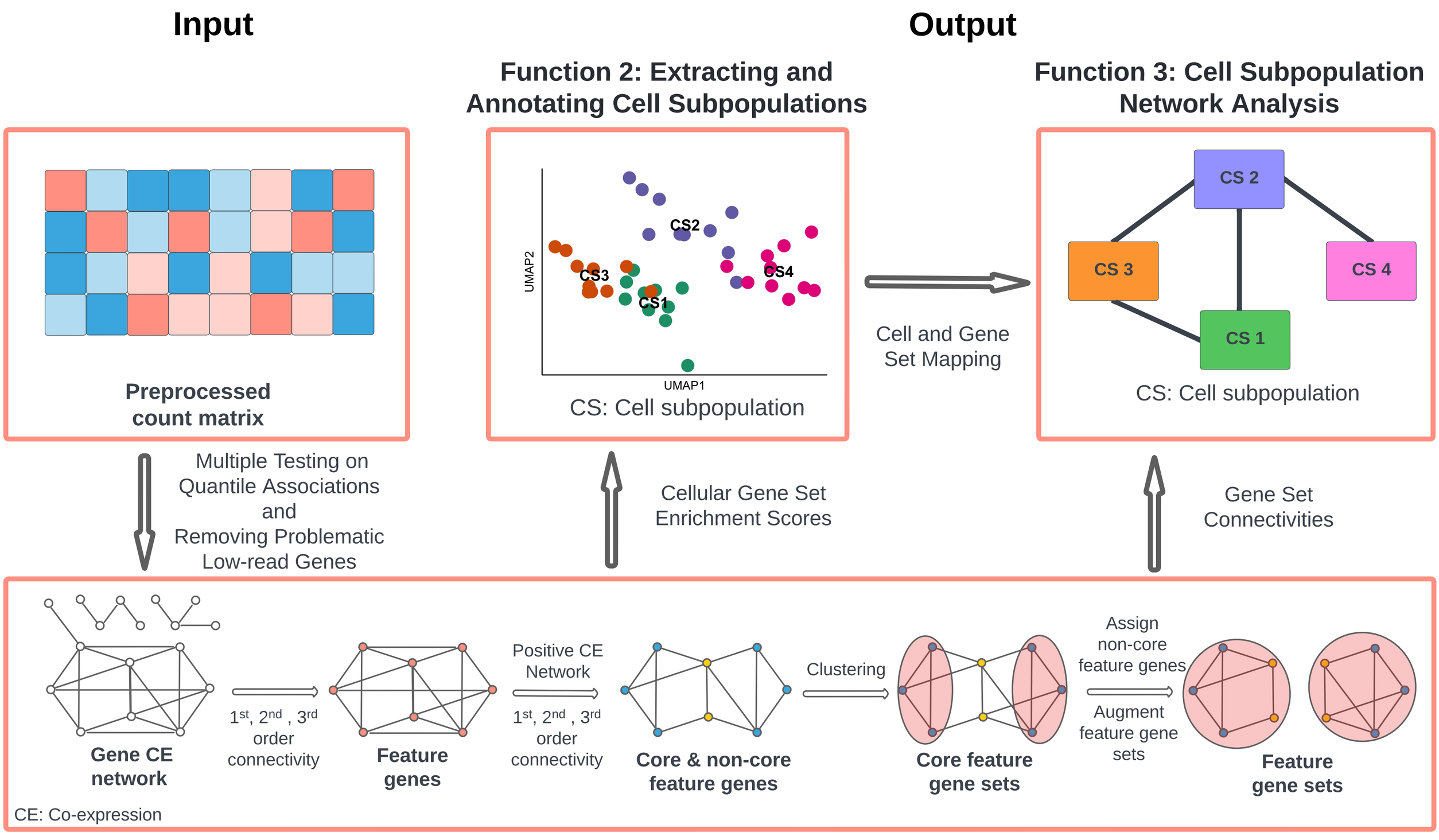
[86] Zhilin Long, Chengfang Sun, Min Tang, Yin Wang, Jiayan Ma, Jichuan Yu, Jingchao Wei, Jianzhu Ma, Bohan Wang, Qi Xie, and Jiaming Wen. Single-cell multiomics analysis reveals regulatory programs in clear cell renal cell carcinoma. *Cell Discov*, 8(1):68, Jul 2022. doi: 10.1038/s41421-022-00415-0. PMCID: PMC9296597.

[87] James R Rose, Bagdeser Akdogan-Ozdilek, Andrew R Rahmberg, Michael D Powell, Sakeenah L Hicks, Christopher D Scharer, and Jeremy M Boss. Distinct transcriptomic and epigenomic modalities underpin human memory t cell subsets and their activation potential. *Commun Biol*, 6(1):363, Apr 2023. doi: 10.1038/s42003-023-04747-9. PMCID: PMC10070634.

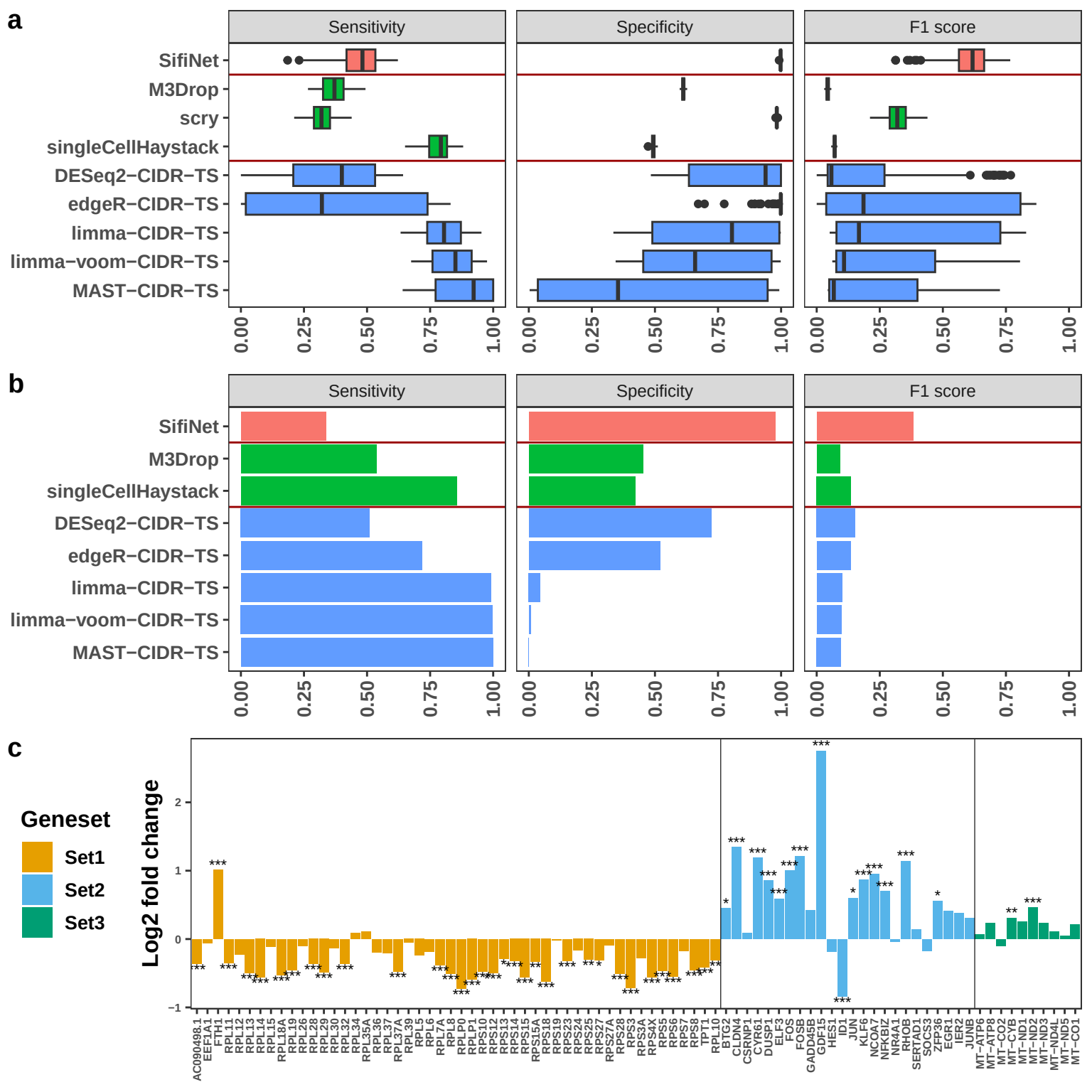
[88] Rama S Akondy, Mark Fitch, Srilatha Edupuganti, Shu Yang, Haydn T Kissick, Kelvin W Li, Ben A Youngblood, Hossam A Abdelsamed, Donald J McGuire, Kristen W Cohen, Gabriela Alexe, Shashi Nagar, Megan M McCausland, Satish Gupta, Pramila Tata, W Nicholas Haining, M Juliana McElrath, David Zhang, Bin Hu, William J Greenleaf, Jorg J Goronzy, Mark J Mulligan, Marc Hellerstein, and Rafi Ahmed. Origin and differentiation of human memory cd8 t cells after vaccination. *Nature*, 552(7685): 362–367, Dec 2017. doi: 10.1038/nature24633. PMCID: PMC6037316.

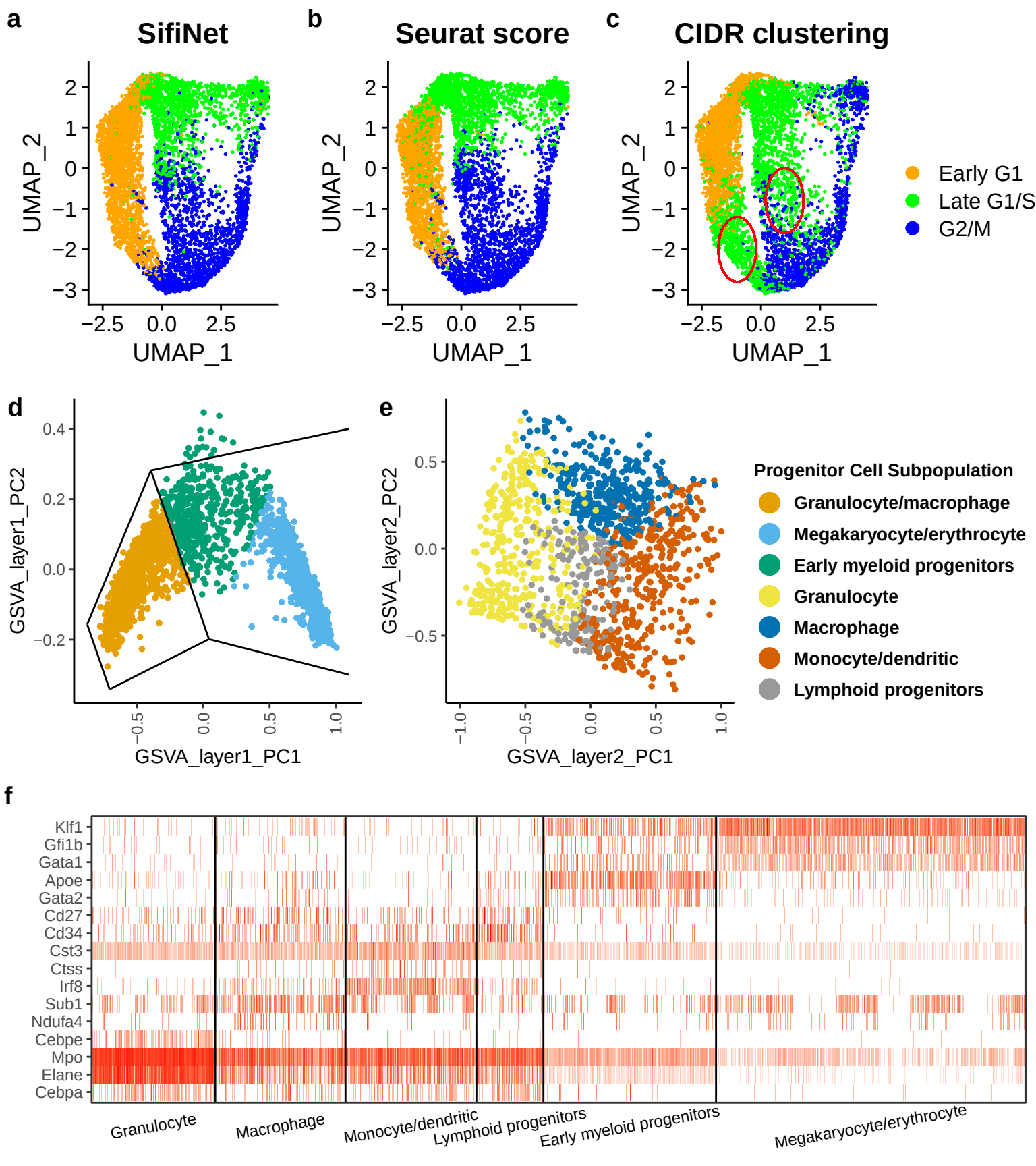
[89] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

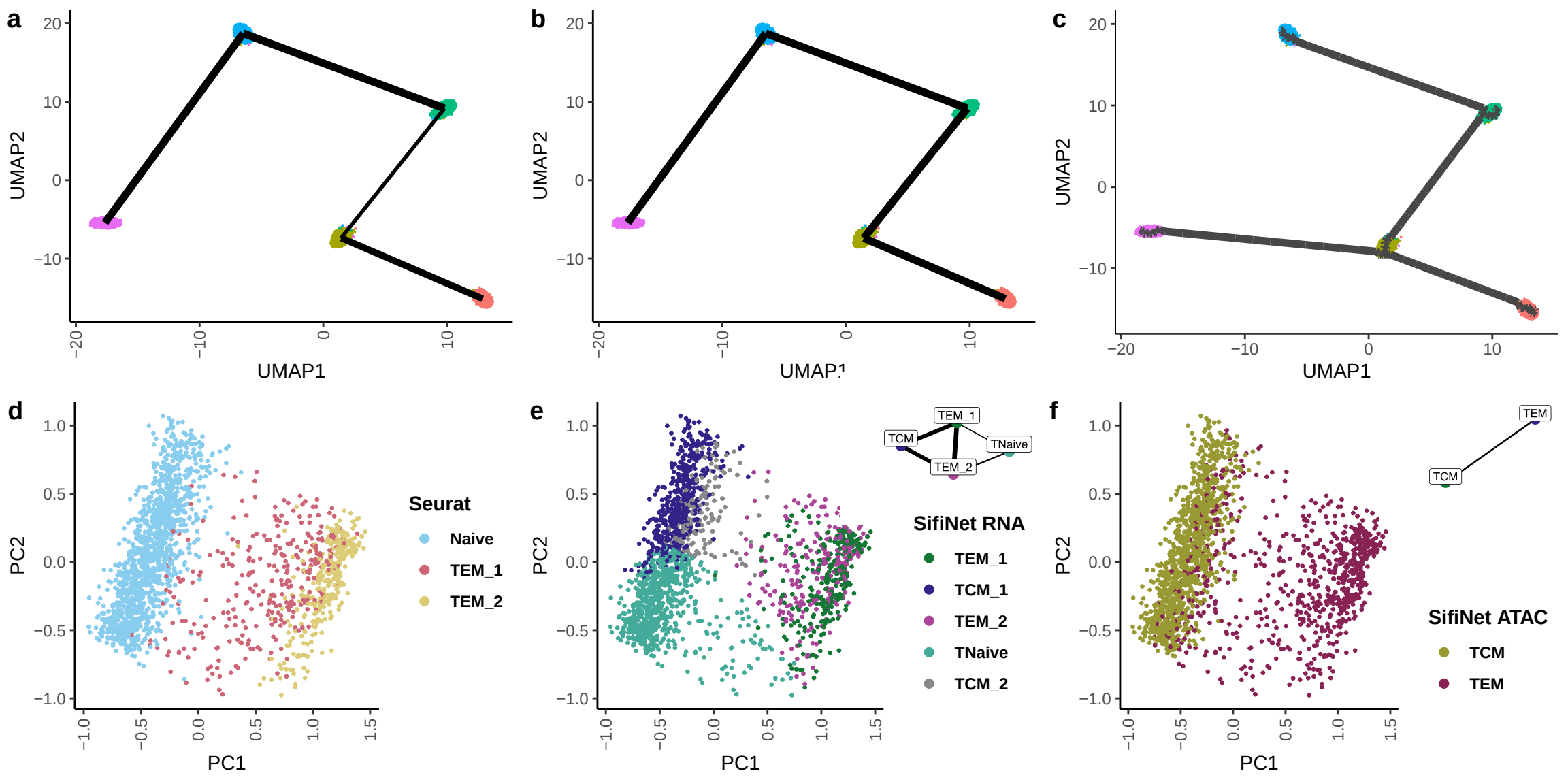
[90] Wenpin Hou and Zhicheng Ji. Reference-free and cost-effective automated cell type annotation with gpt-4 in single-cell rna-seq analysis. *bioRxiv*, Apr 2023. doi: 10.1101/2023.04.16.537094. PMCID: PMC10153208.



Function 1: Identifying Feature Gene Sets


a
Sensitivity
Specificity
F1 score
SifiNet
M3Drop
scry
singleCellHaystack
DESeq2−CIDR−TS
edgeR−CIDR−TS
limma−CIDR−TS
limma−voom−CIDR−TS
MAST−CIDR−TS
0.00
0.25
0.50
0.75
1.00
b
Sensitivity
Specificity
F1 score
SifiNet
M3Drop
singleCellHaystack
DESeq2−CIDR−TS
edgeR−CIDR−TS
limma−CIDR−TS
limma−voom−CIDR−TS
MAST−CIDR−TS
0.00
0.25
0.50
0.75
1.00
c
Geneset
Set1
Set2
Set3
Log2 fold change
−1
0
1
2
AC090498.1
EEF1A1
FTH1
RPL11
RPL12
RPL13
RPL14
RPL15
RPL18A
RPL19
RPL26
RPL28
RPL29
RPL30
RPL32
RPL34
RPL35A
RPL36
RPL37
RPL37A
RPL39
RPL5
RPL6
RPL7A
RPL8
RPLP0
RPLP1
RPS10
RPS12
RPS13
RPS14
RPS15
RPS15A
RPS18
RPS19
RPS23
RPS24
RPS25
RPS27
RPS27A
RPS28
RPS3
RPS3A
RPS4X
RPS5
RPS6
RPS7
RPS8
TPT1
RPL10
BTG2
CLDN4
CSRNP1
CYR61
DUSP1
ELF3
FOS
FOSB
GADD45B
GDF15
HES1
ID1
JUN
KLF6
NCOA7
NFKBIZ
NR4A1
RHOB
SERTAD1
SOCS3
ZFP36
EGR1
IER2
JUNB
MT−ATP6
MT−ATP8
MT−CO2
MT−CYB
MT−ND1
MT−ND2
MT−ND3
MT−ND4L
MT−ND5
MT−CO1


a
SifiNet
b
Seurat score
c
CIDR clustering
UMAP_2
UMAP_1
Early G1
Late G1/S
G2/M
d
GSVA_layer1_PC2
GSVA_layer1_PC1
e
GSVA_layer2_PC2
GSVA_layer2_PC1
Progenitor Cell Subpopulation
Granulocyte/macrophage
Megakaryocyte/erythrocyte
Early myeloid progenitors
Granulocyte
Macrophage
Monocyte/dendritic
Lymphoid progenitors
f
Klf1
Gfi1b
Gata1
Apoe
Gata2
Cd27
Cd34
Cst3
Ctss
Irf8
Sub1
Ndufa4
Cebpe
Mpo
Elane
Cebpa
Granulocyte
Macrophage
Monocyte/dendritic
Lymphoid progenitors
Early myeloid progenitors
Megakaryocyte/erythrocyte


a
b
c
UMAP2
UMAP1
d
e
f
PC2
PC1
Seurat
Naive
TEM_1
TEM_2
SifiNet RNA
TEM_1
TCM_1
TEM_2
TNaive
TCM_2
TCM
TNaive
SifiNet ATAC
TCM
TEM