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

Qi Gao, Jason Ji, Liuyang Wang, Kouros Owzar, Qi-Jing Li, Cliburn Chan, Jichun Xie

Abstract

Single-cell sequencing has provided a means of quantifying cellular omic phenotypes. Identifying cell-type-specific feature genes is a crucial aspect of understanding cellular heterogeneity. Over the past decade, many methods have been developed to identify feature genes; however, these methods either depend on dubious cell clustering or fail to provide subpopulation-specific markers. We introduce SifiNet, a robust and accurate approach for identifying marker gene sets based on gene co-expression network topology. The identified gene sets facilitate the calculation of cellular gene set enrichment scores and cell annotation, and can reveal potential transitional relationships between cell subpopulations. SifiNet outperforms state-of-the-art methods in marker gene set identification and cell-type annotation accuracy. It is applicable to both single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq) data. We have applied SifiNet to various experimental studies, successfully identifying novel gene markers, annotating cells with complex heterogeneity, and uncovering intriguing cell developmental trajectories.

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. Analyzing single-cell data enables the identification of unique gene sets specific to distinct cell subpopulations, referred to as feature gene sets.

Existing feature gene identification methods often employ a two-step approach involving cell clustering and differentially expressed gene (DEG) analysis to identify cluster-specific feature genes. However, this approach has questionable accuracy, particularly when cells exhibit complex heterogeneity, because an inaccurate initial clustering step could lead to subsequent erroneous feature gene identification. Alternatively, some methods identify feature genes by detecting highly variable genes (HVG) based on their deviations to the population gene properties, such as a gene's mean-dispersion relationship[1], dropout rates[2], and UMI count distributions[3]. However, these methods are limited in their ability to separate feature genes into gene sets for cell annotation, constraining their effectiveness in elucidating cellular heterogeneity.

We introduce SifiNet (Single-cell feature identification with Network topology), a novel cell-clustering-independent method for directly identifying feature gene sets. SifiNet is based on the observation that co-differentially-expressed genes within a cell subpopulation also exhibit co-expression patterns. Thus, SifiNet constructs a gene co-expression network and explores its topology to identify feature gene sets. It also applies to scATAC-seq data, generating a gene co-open-chromatin network and exploring network topology to identify epigenomic feature gene sets. These gene sets facilitate cellular gene set enrichment score calculation and cell annotation.

SifiNet outperforms existing state-of-the-art methods in feature gene set identification and cell annotation accuracy and can be used to investigate potential developmental lineages between cell subpopulations. We have applied SifiNet to cell cycle, senescence, myeloid progenitor, and T-cell studies, successfully identifying novel marker genes, cell subpopulations of interest, and complex heterogeneity structures among cells.

Results

SifiNet is a versatile pipeline with multiple functions, such as identifying feature gene sets, recognizing cell subpopulations, and uncovering potential developmental lineage among cell subpopulations. Its main idea is to employ gene co-expression network topology to identify feature gene sets (Fig. 1). Its main advantage lies in its high accuracy, particularly when cells display complex heterogeneity structures. For instance, some cells may be in a transitional stage, some cell subpopulations may share feature genes, and certain cell subpopulations may exhibit hierarchical subpopulation structures. SifiNet can effectively handle these complex cases.

To benchmarked SifiNet's performance, we applied it to simulated and experimental datasets, including two sets of numerical datasets (SD1, SD2) and six experimental datasets (Monoclonal, IPF1, BoneMarrow, CD8, CD4, MassiveRNA). An additional experimental dataset IPF2 was used for validation. Supplementary table 1 provides short descriptions of these datasets. The numerical datasets were generated by our customized R package GeneScape based on the R package Splatter[4]: GeneScape can simulate single-cell data with more complicated heterogeneity structures (Method).

To demonstrate the accuracy of SifiNet, we also compared it with many competing methods, including the two-step methods and the HVG identification methods. See Methods for a list of these methods and their descriptions.

SifiNet identifies feature gene sets with high accuracy.

First, we applied SifiNet and competing methods to SD1, a numerical dataset consisting of 100 simulated scRNA-seq data repetitions. Among all methods, SifiNet has the highest median F1 score of 0.62; all competing methods have much lower F1 scores (Fig. 2a). Among them, singleCellHaystack, limma-CIDR-TS, limma-voom-CIDR-TS, and MAST-CIDR-TS exhibit higher 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, SD1 datasets contain, on average, 82.89 feature genes and 3318.32 non-

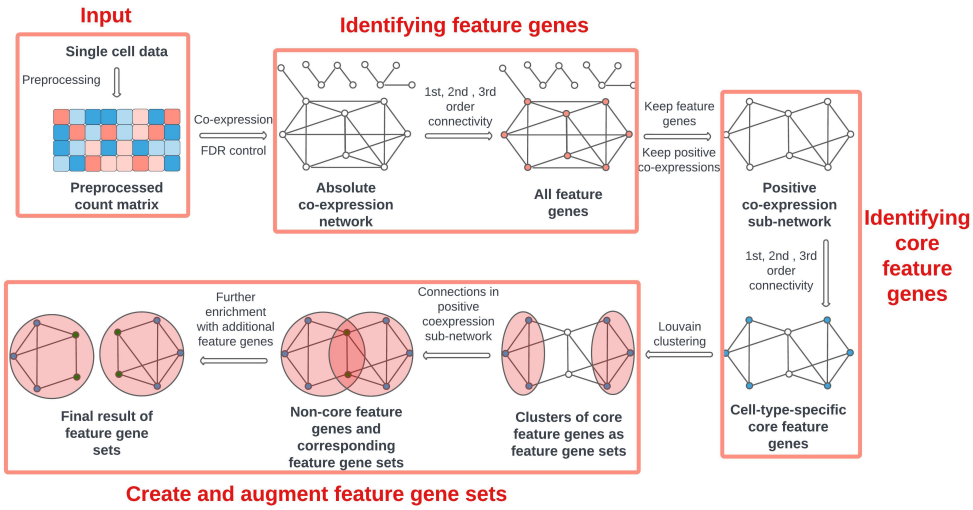

Figure 1: **SifiNet flowchart**. Starting with the preprocessed count matrix, SifiNet computes gene co-expression networks. It examines the 1st, 2nd, and 3rd order connectivity within the co-expression network to identify feature genes. SifiNet then removes negative dependence edges, focusing on the positive co-expression sub-network. Next, it investigates the 1st, 2nd, and 3rd order connectivity of the positive sub-network to pinpoint core feature genes. These core feature genes are subsequently clustered into gene sets based on their mutual connectivities. Remaining feature genes are assigned to gene sets according to their connectivity with core feature genes. The final gene sets can be utilized to construct a topology network and perform gene set enrichment analysis for cell annotation. SifiNet's details are provided in the Methods section.

feature genes. A specificity as high as 0.9 will falsely label, on average, 331.83 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.

We further investigated why the commonly-used two-step methods underperform on SD1. First, assuming that true cell type labels are known, we apply DEG methods to identify cell-type-specific feature genes. These methods are referred to as oracle-cluster methods (Methods). All oracle-cluster methods perform well on SD1, exhibiting high sensitivity and specificity (Supplementary Fig. 1). This indicates that the key factor is the accuracy of clustering. If 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, such as on SD1 (Supplementary Fig. 2), two-step methods perform poorly.

Second, we applied SifiNet to the Monoclonal dataset to identify the feature genes related to cell cycles. Monoclonal contains cloned cells, whose main difference lies in their cell cycles. SifiNet identified two feature gene sets. Set 1 contains many 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 G2/M

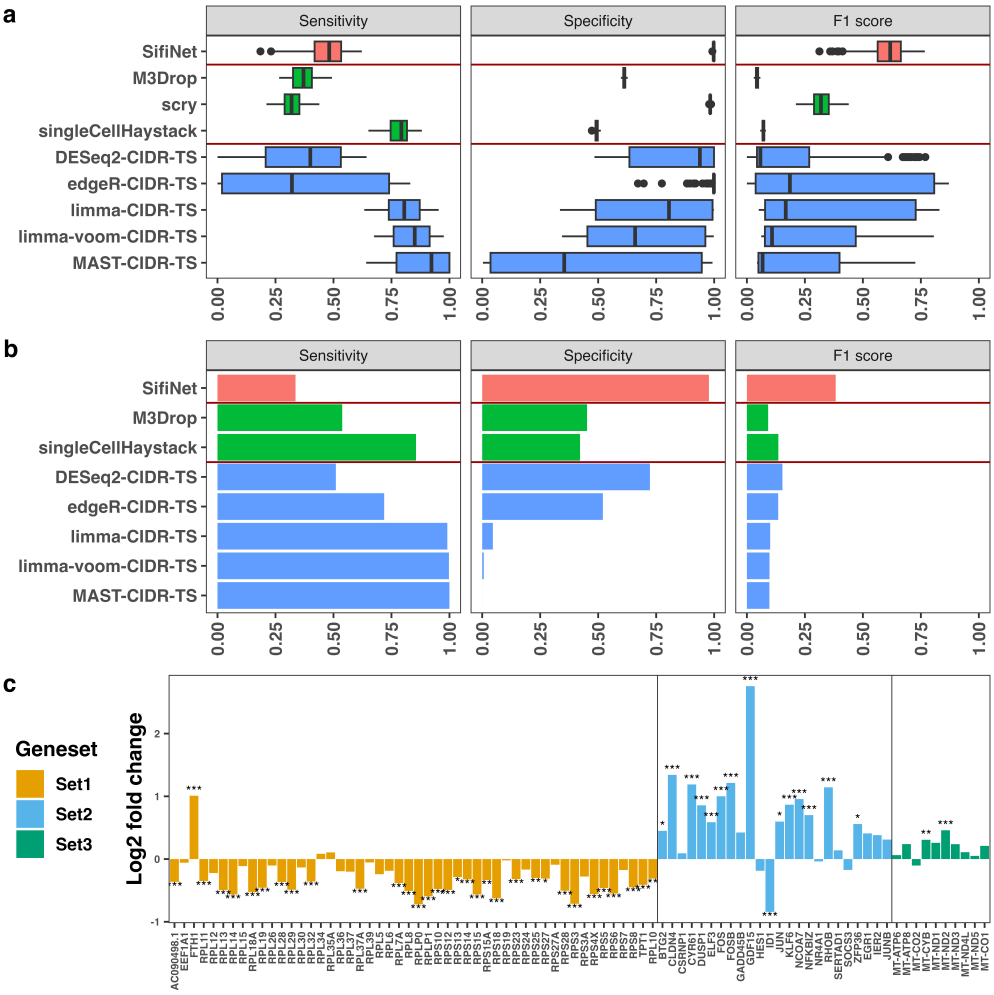

Figure 2: **Performance of SifiNet and other competing methods on feature gene identification.**. **(a)-(b)** Sensitivity, specificity, and F1 accuracy scores of various methods in datasets (a) SD1 and (b) Monoclonal. Methods were classified into 3 types: SifiNet (red), HVG identification methods (green), and two-step methods with clustering performed by CIDR at the default resolution (blue). **(c)** The log fold change (senescence versus non-senescence cells) and significance level of feature genes detected by SifiNet from IPF1 datasets in IPF2 datasets. Stars represent the significance level of the difference between senescence and non-senescence cells, *: p-value $<= 10^{-2}$, **: p-value $<= 10^{-4}$, ***: p-value $<= 10^{-6}$.

phase[5], recent studies indicate that there is a checkpoint for centrosome integrity between the G1 and S phase[6]. Thus, without HAUS genes maintaining the centrosome integrity, the cells might be unable to get into the S phase. Therefore, HAUS genes should play key roles in the G1/S phase, as suggested by SifiNet. Set 2 contains many G2 and M (G2/M) phase genes, including Tubulin and SKA, whose proteins are related to microtubules. The full list of identified genes is in Supplementary table 3.

We also applied the two-step methods to the Monoclonal dataset. We compared the SifiNet-identified feature genes and the two-step-method-identified feature genes to the benchmark gene set containing 437 benchmark cell cycle marker genes collected from the GSEA MSigDB[7, 8, 9] and Uniprot[10] databases (Method). Among the 331 SifiNet-identified feature genes, 147 of them are also benchmark genes. Several methods identified more genes than SifiNet; however, their specificities are lower (Fig. 2b). For instance, MAST-CIDR-TS has the highest sensitivity, identifying all 437 benchmark genes, but it also identified 8,134 irrelevant genes, complicating the understanding of cell cycles. Overall, SifiNet has the highest F1 score among all methods.

Third, we applied SifiNet to the IPF1 dataset to identify senescence-related feature gene sets. IPF1 contains 10,375 ciliated cells from lung tissues of Idiopathic Pulmonary Fibrosis (IPF) patients. Prior research has demonstrated a close relationship between senescence and IPF[11, 12]. Therefore, by exploring cell-type-specific feature genes, we expect to identify feature genes associated with senescence. SifiNet identified three feature gene sets (Supplementary Table 3). Set 1 includes numerous ribosomal protein genes, known to accumulate in senescent cells[13]. 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[14, 15, 16]. This family is known to be involved in proliferation, differentiation, and tumorigenesis[14], highly related to the senescence process[15]. Set 3 genes are all mitochondrial genes. Previous studies have shown that mitochondrial genes are related to the senescence-associated secretory phenotype (SASP)[17], and mitochondrial dysfunction is associated with senescence[18]. 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 421 KRT5-/KRT17+ epithelial cells and 288 proliferating epithelial cells. The KRT5-/KRT17+ cells are treated as senescent cells, and the proliferating cells are non-senescent cells. Among the 84 identified genes, 45 are significantly differentially expressed (adjusted p-values < 0.0001, Fig. 2c) between KRT5-/KRT17+ and proliferating cells. *Gdf15* has the largest log fold change (2.76, adjusted p-value= 3.89×10^{-73}), followed by *Cldn4* (1.34, adjusted p-value= 2.53×10^{-47}), and *Fosb* (1.21, adjusted p-value= 4.25×10^{-25}). Besides, 28 of the 46 identified ribosome genes and 2 out of 10 mitochondrial genes are significantly differentially expressed (Fig. 2c). The independent data analysis verified the quality of the identified feature genes.

SifiNet identifies marker-enriched cell subpopulations with high accuracy.

After SifiNet identified feature gene sets, we can calculate the gene set variation analysis (GSVA)[19] scores, which measure the enrichment level of a given gene set in each cell. Cells

with similar enrichment scores are considered to have similar expression profiles (Method).

First, we examined cells in the Monoclonal dataset to reveal their cell cycles. We calculated GSVA scores based on SifiNet-identified gene sets and clustered cells into three groups (Methods). Fig. 3a displays the Uniform Manifold Approximation and Projection (UMAP)[20] plot of SifiNet-annotated cells. To benchmark its performance, we also annotated cells according to the cell cycle scores provided by Seurat using known cell cycle markers[21] (Fig. 3b). Although SifiNet does not utilize any known marker information, the resulting cell UMAP closely resembles the benchmark UMAP, which used known marker genes. In contrast, we applied CIDR to cluster cells into three cell clusters, without using the known cell cycle markers. The clustering results are less satisfactory: notably, some G1 and G2/M phase cells are clustered into the S phase (Fig. 3c). Furthermore, CIDR, with the default resolution parameter, yielded 8 clusters for cell clustering; the results were hard to interpret. (Supplementary Fig. 3).

Second, we applied the SifiNet-GSVA pipeline to find senescent enriched cell subpopulations in the IPF1 dataset. We used SifiNet-GSVA scores to label SifiNet-senescent cells (Methods). Then, we performed the DEG analysis of a widely known key senescence marker, *Cdkn1a*[22], between these two subpopulations. The log2 fold change is 0.636, and the adjusted p-value= 1.54×10^{-34}. As a comparison, the DEG analysis of *Cdkn1a* between SenMayo-annotated senescent and non-senescent cells is less significant (log2 fold change 0.281, adjusted p-value= 0.169). This suggests that the SifiNet-senescent cell subpopulation could be more enriched with the actual senescent cells.

SifiNet reveals feature gene sets with hierarchical structures.

After we cluster cells based on the SifiNet-GSVA scores, SifiNet can be applied again to a subset of cells to explore their local heterogeneity structure. For example, we applied SifiNet to BoneMarrow, a dataset containing myeloid progenitor cells isolated from mouse bone marrow. These cells will eventually differentiate into different types of blood cells. SifiNet identified two feature gene sets in BoneMarrow, S1 and S2 (Supplementary Table 3). S1 contains 121 genes including *Cebpa*, a marker for many lineages in the granulocyte/macrophage progenitors branch[23, 24]. S2 contains 162 genes including *Gata1*, *Gfi1b*, and *Klf1* (Fig. 3f); these are markers for megakaryocyte/erythrocyte progenitors[25, 26, 27]. We computed the GSVA scores for both feature gene sets. The cells with positive S1 (S2) GSVA scores and negative S2 (S1) GSVA scores were clustered into C1 (C2); those cells with both negative S1 and S2 GSVA scores were clustered into C0. No cells have positive S1 and S2 GSVA scores, indicating that the cells were well annotated and separated by these two gene sets. By checking the DEGs of C0 with respect to other cell clusters, we found early progenitor cell markers such as *Gata2*[28] (Supplementary Fig. 4). Thus, The C0 cells are likely to be the progenitors of the C1 and C2 cells. To further explore the heterogeneity among the C1 cells, we applied SifiNet and GSVA to C1 cells only and obtained 3 feature gene sets S1a-S1c (Supplementary Table 3), and their corresponding cell subpopulations C1a-C1c; the remaining cells are grouped into C1d. S1a includes *Mpo* and *Elane* (Fig. 3f), markers for granulocyte differentiation[29]. S1b includes *Ndufa4* and *Sub1* (Fig. 3f), markers for macrophage differentiation[30, 31]. S1c includes *Irf8*, *Cst3* and *Cts* (Fig. 3f), markers for monocytes and dendritic cells[32, 33]. Thus, SifiNet helps us understand the developmental

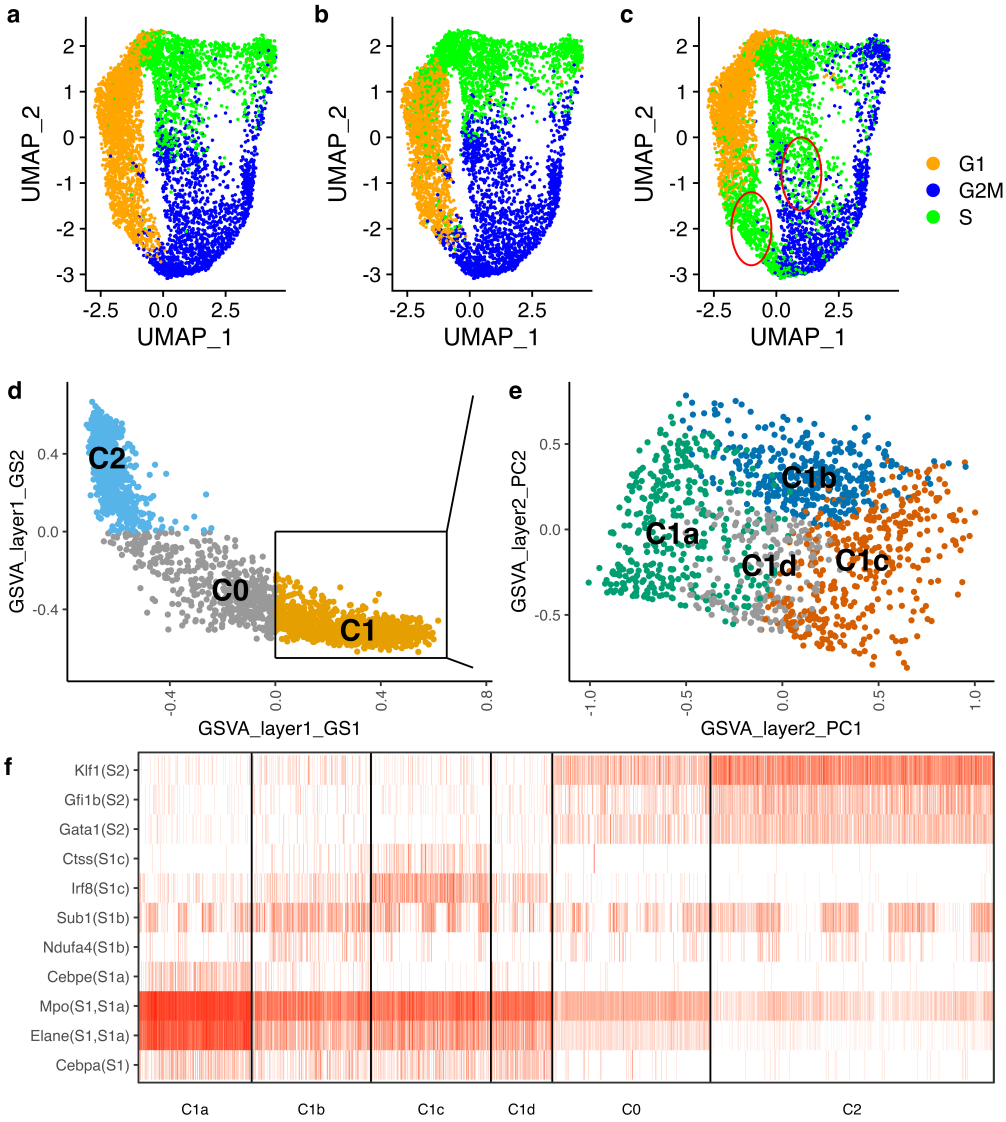

Figure 3: **(a)-(c)** UMAPs of different cell annotation methods. (a) SifiNet+ GSVA cell type annotation; (b) Seurat cell cycle annotation benchmark using cell cycle scores (with known cell cycle marker information); (c) CIDR oracle-cluster-number clustering (cluster cells into known number of clusters). Red circle in (c) shows the major differences between the UMAPs in (c) and (b). **(d)-(e)** SifiNet cell annotations on BoneMarrow dataset. (d) GSVA scores of the SifiNet-identified gene sets: x-axis is the GSVA scores of GS1, y-axis is the GSVA scores of GS2. Each point represents a cell. Three colors represent the clustering results based on the two GSVA scores. (e) After the SifiNet-GSVA pipeline was applied to the C1 subpopulation, we obtained three-dimensional gene set scores. We extracted the first two principal components (PCs) for each cell. The colors of the cells represent their clustering results based on the three-dimensional gene set scores. **(f)** Heatmap of some marker gene expressions. The bracket after the gene name contains the gene set it belongs to.

lineage of myeloid progenitor cells, even before they fully grow into the downstream cells.

SifiNet reveals gene set topology.

SifiNet calculates the gene co-expression network of all feature genes. For two gene sets, we can further calculate the proportion of the realized between-set edges over the total possible between-set edges; this proportion implies the relationship between two gene sets. Thus, we can plot a network with gene sets as nodes, and between-set relationships as edges. Moreover, if we use GSVA scores to cluster the cells into subpopulations corresponding to the gene sets, the network can also be extended to the cell subpopulations.

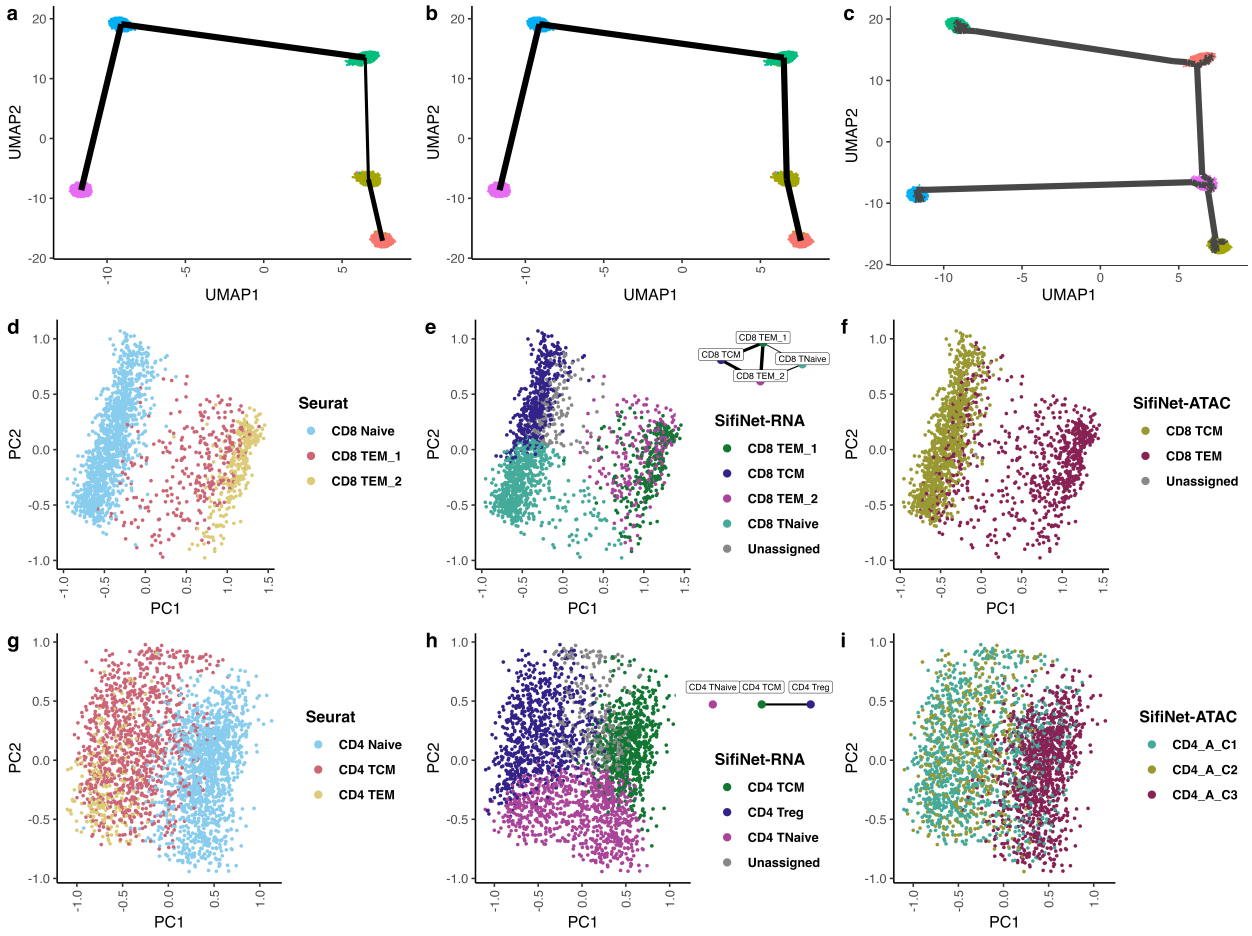

Figure 4: **(a)-(c)** UMAPs of the SD2 cells, with (a) the SifiNet-built lineage, (b) the true cell subpopulation lineage, and (c) the Monocle3-built lineage. **(d)-(f)** PC plot of the SifiNet-GSVA scores of the cells in CD8, annotated by (d) Seurat based on the multi-omics data (e) SifiNet based on the scRNA-seq data (f) SifiNet based on the scATAC-seq data. The PCs were calculated based on the GSVA scores of the SifiNet-derived gene sets based on scRNA-seq and scATAC-seq, respectively. **(g)-(i)** PC plot of the SifiNet-GSVA scores of the cells in CD4, annotated by (g) Seurat based on the multi-omics data (h) SifiNet based on the scRNA-seq data (i) SifiNet based on the scATAC-seq data. Gene set topology graphs are shown in (e) and (h).

We applied SifiNet gene set topology analysis to SD2, a simulated dataset with five cell subpopulations representing five lineage developmental stages (Fig. 4a). SifiNet identified five gene sets. We used the SifiNet-GSVA scores to separate cells into five clusters. Because the cell subpopulations are well separated, the clustering is very accurate. We then explored the gene set topology; the gene set edges were plotted on the UMAPs to represent the relationships between the corresponding cell subpopulations. Clearly, the SifiNet-built lineage (Fig. 4b) is identical to the actual lineage. As a comparison, we also applied Monocle3[20, 34, 35, 36, 37, 38] to SD2. Monocle3 learns the lineage based on the cell graph embeddings in a reduced-dimensional space. The resulting lineage (Fig. 4c) is different from the actual lineage. Thus, SifiNet outperforms Monocle 3 in SD2 to reveal cell developmental lineage.

SifiNet identified multi-omics features and their gene set topology and revealed long-lasting T cell subpopulations.

First, we applied the SifiNet-GSVA pipeline to CD8, a multi-omics (scRNA-seq+scATAC-seq) dataset. The cells were previously clustered by Seurat with the annotation CD8 Naïve cells, CD8 effector memory T cell (TEM) subpopulation 1 (TEM_1), CD8 TEM subpopulation 2 (TEM_2) (Fig. 4d). SifiNet identified four gene sets in CD8 scRNA-seq data, labeled as CD8_R_GS1 - CD8_R_GS4. CD8_R_GS1 and CD8_R_GS3 share *Gzma*, *Klrd1*, and *Klrg1*, markers for effector T cells[39, 40]. CD8_R_GS2 contains many central memory T cell (TCM) markers, including *Foxo1*, *Ccr7*, *Foxp1*, and *Tcf7*[41, 42]. CD8_R_GS4 contains 19 Rpl and Rps genes, which are highly expressed in naïve T cells[43]. Then, we calculated the GSVA scores and clustered the cells into four clusters, SifiNet-RNA CD8 TEM_1, SifiNet-RNA CD8 TCM, SifiNet-RNA CD8 TEM_2, and SifiNet-RNA CD8 TNaïve; the remaining cells were labeled as SifiNet CD8 unassigned (Fig. 4e). Seurat clustered SifiNet-RNA CD8 TCM and SifiNet-RNA CD8 TNaïve together and thus they were all labeled as CD8 Naïve. However, SifiNet successfully separated them into CD8 TCM and Naïve cells. Moreover, gene set topology analysis (Fig. 4e) showed that CD8_R_GS2 (corresponding to TCM cells) and CD8_R_GS4 (corresponding to TNaïve cells) are two end gene sets; they are connected by CD8_R_GS1 and CD8_R_GS3 (corresponding to TEM cells). This suggests that TEM cells may transit between TCM and T Naïve cells.

We then applied SifiNet to the scATAC-seq part to study the CD8 epigenomic profile. We identified two feature gene sets, CD8_A_GS1 and CD8_A_GS2. CD8_A_GS1 includes TCM markers, such as *Klf7* and *Foxp1*[42]. CD8_A_GS2 includes TEM markers, such as *Klrg1*[39]. Using GSVA scores, we further clustered the cells into SifiNet-ATAC CD8 TCM, SifiNet-ATAC CD8 TEM, and SifiNet-ATAC CD8 unassigned. Interestingly, some SifiNet-RNA CD8 TCM and SifiNet-RNA CD8 TNaïve cells are clustered into SifiNet-ATAC CD8 TEM. These cells exhibit TCM and Naïve T cell transcriptome profile; however, their epigenomic profile resembles TEMs. Recent research shows that this cell subpopulation was the most long-lasting T cells after yellow fever vaccination and thus can provide long-term protection[44].

Second, we applied the same analysis pipeline to CD4, another multi-omics dataset. CD4 cells were less-understood than CD8 cells in previous studies. Seurat clustered the CD4 cells into three clusters, CD4 Naïve, CD4 TCM, and CD4 TEM cells (Fig. 4g). SifiNet based on scRNA-seq identified three gene sets, CD4_R_GS1 - CD4_R_GS3. CD4_R_GS1 contains *Ccr7*, and *Tcf7*, markers for TCM cells[41]. CD4_R_GS2 contains *Il32*, highly expressed in

CD4 regulatory T (Treg) cells[45]. CD4_R_GS3 contains 31 Rpl and Rps genes, markers for CD4 Naïve T cells[43]. The corresponding cell subpopulations of the three gene sets are SifiNet-RNA CD4 TCM, SifiNet-RNA CD4 Treg, SifiNet-RNA CD4 T Naïve; other cells were labeled as SifiNet-RNA CD4 unassigned. We repeated the same pipeline on the CD4 scATAC-seq part (Figs. 4h, 4i). The resulting cell subpopulations look very similar to Seurat annotation, but different from the SifiNet annotation based on the cell transcriptome profiles. This implies that CD4 cells may exhibit different markers in transcriptome and epigenomic profiles. Gene set topology analysis (Fig. 4h) showed that based on the cell transcriptome profiles, CD4 TCM and CD4 Treg cells are more tightly connected. CD4 Naïve T cells are less connected with these two cell subpopulations.

SifiNet scales to million-cell datasets.

The rapid development of single-cell technology and analysis tools accelerates single-cell data collection. As more and more single-cell data become available, we tested the computation performance of SifiNet on large-scale single-cell datasets which could contain millions of cells. In general, the computation complexity of SifiNet is $O(Np^2)$, where N is the number of cells and p is the number of genes after pre-processing. Furthermore, SifiNet can be scaled to mini-batches in some key steps and thus reduces the computation working memory demand. Besides, we can further reduce the candidate feature genes by prioritizing in exploring highly variable features.

To assess SifiNet's computation feasibility, we applied SifiNet to a super-large-scale scRNA-seq datasets, MassiveRNA, containing more than 1.3 million cells and 27,998 genes. After quality filtering, MassiveRNA contains 10,448 genes and 1,287,071 cells. SifiNet took 12.18 hours and 100.75 GB peak memory to complete the analysis, even without scaling the data into mini-batches.

To learn the impact of cell numbers on time cost and peak memory usage, we down-sampled MassiveRNA datasets to 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 (Supplementary Fig. 5). When the cell number is less than 50,000, the peak memory usage was 4.1 GB, dominated by the FDR control procedure. As the cell number increases above 50,000, the log10 peak memory usage was roughly linear with the log10 cell number; the peak memory usage was reached in the co-expression calculation step (Supplementary Fig. 5). Overall, SifiNet is efficient in computation time and memory usage.

Discussion

SifiNet is a network-topology-based method to identify feature gene sets. SifiNet does not rely on cell clustering and thus will not be affected by the cell clustering accuracy. Combined with gene set enrichment analysis methods such as GSVA, SifiNet can annotate cell subpopulations corresponding to the identified gene sets. Compared with the classic two-step methods with first cell clustering and then DEG analysis, the SifiNet-GSVA pipeline provided better annotations on multiple experimental datasets. Besides, the gene set topology

analysis can reveal the lineage relationship between the cell subpopulations. The information can help researchers further understand the developmental trajectories or cell subpopulation dependencies.

As all other methods, SifiNet has its own limitations. First, SifiNet is a co-expression network-based method; thus, the quality of the estimated co-expression network is crucial to the following steps. To ensure the high-quality, we used quantile association networks to replace the commonly used correlation networks: quantile association networks have been shown to work better for the count data[46], especially for the transcripts counts. We also explored the rigorous theoretical properties of resulting co-expression networks: the mathematical theorems and proofs were provided in the Supplementary note 1 and 2. Our procedure was optimized based on the theoretical results. Second, the gene set identification procedure depends on Louvain clustering. We preset the default hyperparameter to separate the feature genes into gene sets. However, the default hyperparameter may not perform optimally for all datasets. In practice, the users can use other internal clustering evaluation metrics (such as Davies-Bouldin index[47], Silhouette coefficient[48], Calinski-Harabaz index[49]) to select the optimal gene sets. However, the optimal internal evaluation metric for SifiNet still requires further investigation.

Method

SifiNet: Identifying feature gene sets and gene set topology network

SifiNet is a computational pipeline to directly call feature gene sets without first clustering the cells. Feature genes are differentially expressed within specific cellular subpopulations. A set of these genes, exhibiting co-differential expression within the same cellular subpopulation, constitutes a feature gene set.

SifiNet rests upon a key observation: co-differentially-expressed genes show co-expression patterns, even if the expression of those genes are conditionally independent given the cell types. Mathematically, let Y_{ic} represent the feature values of gene i (e.g., UMI counts, aggregated epigenomic feature values) within cell c, and s_c is the cell library size (total number of feature counts for all genes in cell c, such as total read counts). Suppose gene i is differentially expressed in the tth cell subpopulation $\mathcal{C}_t$ with $P(c \in \mathcal{C}_t) = \pi_t$. Then, Y_{ic} follows $F_{it}(\cdot \mid s_c)$ when $c \in \mathcal{C}_t$ and follows $F_{i0}(\cdot \mid s_c)$ when $c \notin \mathcal{C}_t$, where $F_{i0} \neq F_{it}$. If two genes i and j belong to the same feature gene sets, they are both differentially expressed in a cell subpopulation $\mathcal{C}_t$. Thus, even if they are independent conditioning on cell subpopulation, they are still dependent given the cell library size.

$$
\begin{aligned}
& P(Y_{ic} \leq q_{ic}, Y_{jc} \leq q_{jc} \mid s_c) \\
&= P(Y_{ic} \leq q_{ic}, X_{jc} \leq q_{jc} \mid s_c, c \in \mathcal{C}_t) P(c \in \mathcal{C}_t) + P(Y_{ic} \leq q_{ic}, Y_{jc} \leq q_{jc} \mid s_c, c \notin \mathcal{C}_t) P(c \notin \mathcal{C}_t) \\
&= \pi_t F_{it}(q_{ic} \mid s_c) F_{jt}(q_{jc} \mid s_c) + (1-\pi_t) F_{i0}(q_{ic} \mid s_c) F_{j0}(q_{jc} \mid s_c) \\
&\neq [\pi_t F_{it}(q_{ic} \mid s_c) + (1-\pi_t) F_{i0}(q_{ic} \mid s_c)][\pi_t F_{jt}(q_{jc} \mid s_c) + (1-\pi_t) F_{j0}(q_{jc} \mid s_c)] \\
&= P(Y_{ic} \leq q_{ic} \mid s_c) P(Y_{jc} \leq q_{jc} \mid s_c)
\end{aligned}
$$

where q_{ic} is the τ_i-quantile of gene i in cell c.

Therefore, co-differential-expressions implies co-expressions (marginal dependence between two genes), although not necessarily in reverse. Nevertheless, since all feature genes exhibit mutual co-expression patterns, in a graph with all genes as nodes and co-expressions as edges, it is anticipated that these genes, over-expressed within the same cellular subpopulation, will compose a sub-graph replete with mutually connected edges, thereby revealing community patterns. SifiNet employs this attribute for the identification of feature genes. It contains the following steps.

S1. Measuring gene co-expression

After pre-processing, suppose the cell feature matrix consists of p genes and N cells. Between any two genes, we use quantile associations to measure their co-expression. Quantile associations have demonstrated superiority in characterizing gene dependence compared to Pearson correlations and rank correlations, particularly for gene count data [46]. Between gene i and gene j, we define their quantile association as follows.

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

where $\hat{q}_{ic}$ is the estimated cell specific τ_i-quantile of gene i, and $\hat{\tau}_i = \sum_{c=1}^{N} I(Y_{ic} \leq \hat{q}_{ic})/N$. By default, we used the quantile regression model $q_{ic} = \alpha_{i0} + \alpha_{i1} s_c$ to estimate $\hat{q}_{ic}$, and let $\tau_i = (1 + \hat{p}_{i0})/2$, where $\hat{p}_{i0} = \sum_{c=1}^{N} I(Y_{ic} = 0)/N$ is the proportion of zero observation of gene i. 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 used $\hat{q}_{ic} = 0$.

Under the assumption that Y_{ic} and Y_{jc} are conditionally independent given cell c, when either gene i or gene j is not a feature gene, S_{ij} asymptotically converges to $N(0, 1)$ under mild conditions. Refer to Supplementary Note 1 for the mathematical proof.

Additionally, Supplementary Note 2 employs an experimental data example to demonstrate the following aspects:

- When two genes are feature genes of the same cell subpopulation, their co-expression values follow a distribution with positive expectation.
- When two genes are feature genes of different cell subpopulations, if these cell subpopulations contain overlapping cells, the co-expression expectation may be positive; otherwise, the expectation is highly likely to be negative.

S2. Constructing gene co-expression graph

S2.1. Initiating a gene co-expression graph. We denote the initial gene co-expression graph by $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. Here, $\mathcal{V}$ denotes the vertex set, and $\mathcal{E}$ denotes the edge set, signifying gene co-expression patterns. We set $\mathcal{V} = [p] = \{1, \ldots, p\}$ as an initial value. To obtain $\mathcal{E}$, we set up $p(p-1)/2$ hypotheses: $\mathrm{H}_{\mathrm{null},ij}: \ S_{ij} \sim N(\mu_0, \sigma_0^2)$, for all $i, j \in \mathcal{V}$ and $i < j$. Although we have theoretically demonstrated that S_{ij} approaches an asymptotic $N(0, 1)$ distribution if either gene i or gene j is not a feature gene, in practice, the null distribution of S_{ij} may slightly deviate from $N(0, 1)$ due to unobserved factors. Existing multiple testing literature recommends estimating the mean and variance under the null, which has been proven to enhance testing result accuracy [50, 51]. Consequently, we adhere to this recommendation and estimate the null mean $\hat{\mu}_0$ and variance $\hat{\sigma}_0^2$ for testing [52].

We employ the following FDR control procedure to designate edges. We added (i, j) to $\mathcal{E}$ if $|S_{ij} - \hat{\mu}_0|/\hat{\sigma}_0 \geq \hat{b}$, where

$$\hat{b} = \inf\left[0 \leq b \leq b_p : \frac{p(p-1)[1 - \Phi(b)]}{\max\{1, \sum_{1 \leq i < j \leq p} I[|S_{ij} - \hat{\mu}_0|/\hat{\sigma}_0 \geq b]\}} \leq \alpha\right],$$

$b_p = 2\sqrt{\log(N \vee p)}$, Φ is the CDF of $N(0, 1)$, and α is the FDR level, which is often set at 5%. If such $\hat{b}$ does not exist, set $\hat{b} = b_p$. The FDR control method is similar to what introduced in Xie and Li [46].

S2.2. Eliminating problematic low-count genes and corresponding edges. It is possible for some low-count genes exhibit false co-expressions, as their gene count are non-zero only in a small cell subpopulation due to technical bias. Consequently, even if they are not feature genes, they may still exhibit community patterns. Thus, by default, we refer to gene i as problematic if it satisfies all the following three stringent criteria:

- Gene i must have a median read count ≤ 2.

- The proportion of gene i's positive significant edges among all its significant edges ≥ 0.9. That is, the ratio $\sum_{j=1}^{p} I((S_{ij} - \hat{\mu}_0)/\hat{\sigma}_0 \geq \hat{b}) / \sum_{j=1}^{p} I(|S_{ij} - \hat{\mu}_0|/\hat{\sigma}_0 \geq \hat{b})$ must be $\geq 90\%$.
- Among all the genes that gene i's positive edges link to, the proportion of those genes with the same median read count ≥ 0.9. In other words, let $med(Y_i)$ be the median of feature values of gene i. The ratio $\sum_{j=1}^{p} I((S_{ij} - \hat{\mu}_0)/\hat{\sigma}_0 \geq \hat{b}, med(Y_j) = med(Y_i)) / \sum_{j=1}^{p} I((S_{ij} - \hat{\mu}_0)/\hat{\sigma}_0 \geq \hat{b})$ must be $\geq 90\%$.

To mitigate the impact of those problematic low-count genes, we remove those problematic low-expressed genes from $\mathcal{V}$, as well as their corresponding edges from $\mathcal{E}$. This resulting graph is the refined co-expression graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

S3. Determining feature gene sets based on graph topology.

S3.1. Identifying feature genes. Our objective is to recognize genes that are part of communities with other genes in $\mathcal{G}$. To accomplish this, we introduce three topology measures for $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. Define $\mathcal{N}_{i1} = \{j : (i,j) \in \mathcal{E}\}$ as gene i's neighbor set.

- First-order connectivity: This measures gene i's connectivity level with other genes and is defined as $C_{1i} = |\mathcal{N}_{i1}|/(p-1)$, where $|\mathcal{S}|$ denote the cardinality of set $\mathcal{S}$.
- Second-order connectivity: This measures the connectivity level between gene i's neighbors and is defined as $C_{2i} = |\mathcal{E} \cap \{(j,k) : j,k \in \mathcal{N}_{i1}\}|/|\mathcal{N}_{i1}|(|\mathcal{N}_{i1}| - 1)$.
- Third-order connectivity: This measures the connectivity levels between gene i's 2nd-order neighbors $\mathcal{N}_{i2} = \{k : (k,j) \in \mathcal{E}, j \in \mathcal{N}_{i1}\}$ (i.e., neighbors' neighbors). Some genes may have too many 2nd-order neighbors. To improve computation efficiency, we employ a sampling strategy. For any gene i, we randomly sampled L times, each time with m genes from its 2nd-order neighbors, with the sampling probabilities proportional to the number of the gene's connecting edges with the genes in N_{i1}. In other word, the weight of gene k for calculating gene i's 3rd-order connectivity $w_{ik} \propto |\{j : (k,j) \in \mathcal{E}, j \in \mathcal{N}_{i1}\}|$. The resulting gene set at time ℓ is denoted by $\mathcal{A}_i^{(\ell)}$. Then, C_{3i} computes the connectivity levels between the genes in $\mathcal{A}_i^{(\ell)}$ as $C_{3i} = \sum_{\ell=1}^{L} |\mathcal{E} \cap \{(j,k) : j,k \in \mathcal{A}_i^{(\ell)}\}|/Lm(m-1)$.

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 non-feature genes, genes on the pathways or hub genes (Supplementary Figure 8). By default, we require $C_{1i} \geq 5/(p-1)$, $C_{2i} \geq 0.4$ and $C_{3i} \geq 0.3$. For feature genes with multiple roles, they are left out for now and will be added in S3.4.

S3.2. Identifying core feature gene sets. The core feature gene sets comprise genes that exhibit community structures with other positively co-expressed genes, as over-expressed feature genes in the same cell subpopulations tend to positively co-express, whereas those over-expressed in different cell subpopulations are likely to be negatively co-expressed (Supplementary Note 2). Thus, we first derive a positive co-expression subgraph from the original co-expression graph, denoted by $\mathcal{G}_+ = \{\mathcal{V}_+, \mathcal{E}_+\}$. We set $\mathcal{V}_+$ to be the feature genes identified in S3.1, and $\mathcal{E}_+ = \mathcal{E} \cap \{(i,j) : i,j \in \mathcal{V}_+, S_{ij} > 0\}$. Next, we calculate the 1st-, 2nd-, and 3rd-order connectivities for $\mathcal{G}_+$, denoted by $C_{+,1i}$, $C_{+,2i}$, and $C_{+,3i}$, respectively. We consider genes as core feature genes if they exhibit high values of $C_{+,2i}$ and $C_{+,3i}$ and moderate

or high values of $C_{+,1i}$ (Supplementary Figure 9). By default, we require $C_{+,1i} \geq 5/(|\mathcal{V}_+|-1)$, $C_{+,2i} \geq 0.7$, and $C_{+,3i} \geq 0.5$.

Next, we extract the subgraph that solely comprises core feature genes and their corresponding edges, denoted by $\mathcal{G}_{+,\text{core}}$. After that, we apply the Louvain algorithm to cluster these genes into K clusters, denoted by $\mathcal{V}_k$, where $k = 1, \ldots, K$. The default clustering resolution parameter is set as the same as the default value in `cluster_louvain` function in `igraph` R package. Given that core feature genes generally exhibit clear community patterns, the clustering is usually stable over a wide range of resolution parameters.

S3.3. Assigning non-core feature genes. Feature genes identified in S3.1 but not in S3.2 are referred to as non-core feature genes. These genes may have low over-expression signals or may be shared by different cell subpopulations. We assign them to the feature gene set $\mathcal{V}_k$ if their connectivity to the genes in $\mathcal{V}_k$ is high. By default, gene $i \in \mathcal{V}_+ \setminus (\cup_k \mathcal{V}_k)$ is added to $\mathcal{V}_k$ if $|\mathcal{N}_{i1} \cap \mathcal{V}_k|/|\mathcal{V}_k| \geq 40\%$. It is worth noting that different feature gene sets may have overlapping genes.

S3.4. Augmenting feature gene sets with additional genes. Certain genes may possess multiple functions, such as concurrently serving as feature genes and regulatory genes. Although these genes might have been omitted in Section 3.1, we reincorporate them during this stage. Any gene $i \in \mathcal{V} \setminus \mathcal{V}_+$ is integrated into $\mathcal{V}_k$ provided that $|\mathcal{N}_{i1} \cap \mathcal{V}_k|/|\mathcal{V}_k| \geq 90\%$. These newly incorporated genes are designated as augmented feature genes. The ultimate gene collections $\mathcal{V}_k$, $k = 1, \ldots, K$, encompass core feature genes, non-core feature genes, and augmented feature genes. Genes that are common to more than one $\mathcal{V}_k$ are identified as shared feature genes.

S4. Extracting cell subpopulations corresponding to feature gene sets

Based on $\mathcal{V}_k$, $k = 1, \ldots, K$, we calculate the cellular gene set enrichment scores $\alpha_{c,1}, \ldots, \alpha_{c,K}$ for any cell c. We use the gene set variation analysis (GSVA) scores in this study; however, other scores can also be incorporated in the future. We define the cell subpopulation corresponding to $\mathcal{V}_k$ by

$$\mathcal{C}_k = \{c : \alpha_{c,k} > 0, \text{ and } \alpha_{c,k} \text{ is the largest among all } \alpha_{c,\ell},\ \ell = 1, \ldots, K\}.$$

Those cells with all negative $\alpha_{c,k}$ will be assigned to $\mathcal{C}_{K+1}$, called the non-assigned group.

S5. Constructing feature gene set network

By considering the feature gene sets as nodes within a graph, we can construct a gene-set-level graph. In this case, $\mathcal{S} = [K]$ represents the node set, which corresponds to the K gene sets obtained from Section 3. In total, there are $K(K-1)/2$ possible edges. For the edge (k, ℓ), we define its width as:

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

The width indicates the relationship between the corresponding two gene sets. A wide edge could imply cell developmental lineage between two corresponding cell subpopulations.

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 (such as DESeq[53], edgeR[54, 55, 56], limma[57, 58, 59], limma-voom[60], and MAST[61, 62]) and the clustering methods (such as Seurat[63, 64, 65, 66], simple Louvain[67], and CIDR[68]), 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. The two-step methods include DESeq2-Seurat-TS, edgeR-Seurat-TS, limma-Seurat-TS, limma-voom-Seurat-TS, MAST-Seurat-TS, DESeq2-Louvain-TS, edgeR-Louvain-TS, limma-Louvain-TS, limma-voom-Louvain-TS, MAST-Louvain-TS, DESeq2-CIDR-TS, edgeR-CIDR-TS, limma-CIDR-TS, limma-voom-CIDR-TS, MAST-CIDR-TS. The clustering methods are applied at default resolution.

The oracle-cluster methods assume the ground truth of cell clusters, and they 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 highly variable gene (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 include in the comparisons are the following.

- scry[1, 69]: scry ranks genes by binomial deviance.
- M3Drop[2, 70]: M3Drop finds feature genes by testing the relationship between mean expression level and dropout rate (proportion of zero observation) with Michaelis-Menten function.
- singleCellHaystack[3]: singleCellHaystack detects feature genes using Kullback-Leibler divergence between observed distribution and reference distribution of each gene.

Numerical data generation and analysis

Two numerical datasets SD1 and SD2 are simulated by R package `GeneScape`: this package allows user to simulate scRNAseq data with customized differential expression structures, such as core and shared feature genes, sub-cell types, pathways and hub genes. Cells with less than 10% of the genes expressed and genes expressed in less than 10% of the cells are excluded.

SD1 contains 100 repetitions of simulated scRNA-seq data. Each data repetition includes 3 cell types with a total of 6000 cells. One of the cell types is rare (with 100 cells), sharing 60% (30 out of 50) of its feature genes with another non-rare cell type (with 4900 cells): these rare cells mimic those cells recently developed from the non-rare cell type.

In SD1, we checked the performance of various competing methods in identifying feature genes, including the oracle-cluster methods, the two-step methods, and the HVG identification methods.

The DEG methods were applied to the datasets in similar ways as described in [71]. False discovery rate (FDR) is controlled at 0.05 with Benjamini-Hochberg (BH) procedure if there are p-values available in the method. Scry does not provide gene p-values; it ranks the genes

based on the deviance scores. Thus, we selected the top p_1 genes where p_1 is the number of the actual feature genes. Supplementary Table 2 shows details of the applied methods.

Complete analysis results on SD1 datasets by using these methods are shown in Supplementary Figure 1, while partial results are displayed in Fig. 2a. Because the two-step methods with different clustering methods yield similar results and CIDR has slightly better performance (Supplementary Fig. 6), for the remaining analyses, we will only use CIDR for cell clustering in the two-step methods.

SD2 consists of 5 equal-sized cell groups, each containing 1200 cells. Among them, 2 cell types are transition cell types, and the other 3 cell types are ending cell types. Transition cell types shared feature genes with their neighboring ending cell types. The cell developmental lineage has the structure: ending-transition-ending-transition-ending. Each transition cell type further contains 3 subtypes. The feature genes across the 3 subtypes are the same but with varying fold changes: some subtypes are closer to the ending type, and others are farther away.

We applied SifiNet to SD2 to identify the feature gene sets and build the gene set topology network. Then we use GSVA to calculate gene set enrichment scores for each cell, and cluster the cells into clusters corresponding to the gene sets. Thus, the gene set topology network can be extended to cell clusters, showing the cell cluster developmental lineage.

We also applied Monocle3 to SD2. Monocle3 is a graph-based scRNA-seq trajectory analysis method. With the input of the cell clusters, it can also build a cell cluster developmental lineage.

Experimental data analysis

Monoclonal

The dataset, Monoclonal, was collected from a monoclonal murine colorectal carcinoma cell line CT26WT. The dataset details were available at [72]. We preprocessed the data in the following way. We removed the cells whose mitochondrial genes expression takes more than 5% of the total gene expression and the cells whose total number of read count is greater than 1.8 times of the mean expression levels. After that, we further removed the cells with less than 10% of the genes expressed and the genes expressed in less than 10% of the cells.

We applied two-step methods with CIDR clustering, HVG identification methods (M3Drop and singleCellHaystack), and SifiNet to Monoclonal in the similar way as we applied these methods to SD1.

The performance of these methods were evaluated using sensitivity, specificity, and F1 scores. To calculate these scores, we curated a benchmark gene set with 437 cell cycle genes from the GSEA MSigDB and Uniprot databases. First, for the GSEA MSigDB database, we collected human cell cycle related genes using 3 libraries, Reactome[73], Biocarta[74], KEGG[75, 76], and 32 datasets from single studies or pathway analyses (Supplementary table 4) with "CELLCYCLE" or "CELL_CYCLE" in their name from GSEA MSigDB C2 curated human gene sets v2023.1. Genes that were included in the related pathway in at least 1 library or at least 3 studies were selected. Next, we converted the human genes to their equivalents in mouse using `biomaRt` R package. Second, for the Uniprot database, we downloaded the Uniprot gene annotations on May 1, 2023, and selected those genes

whose "Biological process" annotations 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 were 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) [77]. IPF1 contains the pre-annotated "Ciliate" cells in IPF patients; these cells are expected to contain senescent cells. IPF2 contains the pre-annotated "KRT5-/KRT17+" epithelial cells and the "Proliferating Epithelial Cells". The KRT5-/KRT17+ cells were treated as one type of senescence cells as claimed in the original study; the proliferating cells were treated as non-senescence cells because senescence cells cannot proliferate. Then, we performed further pre-processing on both datasets to remove the cells with less than 5% of the genes expressed and the genes expressed in less than 5% of the cells. Before performing the downstream GSVA and DEG analysis, cell counts were normalized by `SCTransform` in Seurat. We used `FindMarkers` function in Seurat to perform DEG analysis: checking gene log-fold change and calculating p-values.

We first applied SifiNet to IPF1 to identify feature gene sets. We then used the independent dataset IPF2 to verify the identified gene sets. Because the cells are pre-labeled with "KRT5-/KRT17+" and "Proliferating". DEG analysis of the identified genes were performed between these two cell subpopoulations.

Also, we evaluated SifiNet's performance on senescence cell annotation without the assist of any other external senescence markers using IPF1. Because we suspected the three gene sets identified by SifiNet represent three different aspects of senescence, we call those cells with top/bottom 40% GSVA scores of all three gene sets SifiNet-senescent cells: top scores were used for up-regulated gene set (Gene Set 2), bottom scores were used for down-regulated gene sets (Gene sets 1 and 3). A total of 594 cells are annotated as SifiNet-senescent cells. We performed the DEG analysis on a well-known marker *Cdkn1a* on these two subpopulations. To demonstrate the significance of *Cdkn1a* in SifiNet-senescent cells, we also compared its significance in the senescent cells annotated using SenMayo senescence marker set (125 genes). Two of the feature genes identified by SifiNet were also in SenMayo marker set. We removed these two genes from the SenMayo marker set. Then we normalized the cellular gene UMI counts in IPF1 by `SCTransform` and calculated the cellular SenMayo-GSVA scores. After that, we chose 594 cells with the top SenMayo-GSVA scores as SenMayo-senescent cells.

BoneMarrow

BoneMarrow is collected from existing study on myeloid differentiation [28]. We focused on the pre-annotated "Unsorted myeloid" cells. We used `SCTransform` to normalize the cell UMI counts and remove the batch effect. Then, we removed the cells with fewer than 500 of the genes expressed, and the genes expressed in fewer than 2% of the cells.

SifiNet was first applied to the entire dataset, and then applied the cell subpopulation corresponding to gene set 1 (granulocyte/macrophage progenitors branch).

CD8 and CD4

CD8 and CD4 datasets are subsets of a 10x scRNA-seq and scATAC-seq multiome datasets [78]. An annotation of cell types is provided by Seurat. CD8 dataset contains cells with "CD8 Naive", "CD8 TEM_1" and "CD8 TEM_2", while CD4 dataset includes cells with "CD4 Naive", "CD4 TCM", and "CD4 TEM". In scATAC-seq data, peaks within the 5000-base-pair upstream of the transcription starting site (TSS) of genes were aggregated to represent the epigenomic features of the corresponding genes. Only genes available in both scRNAseq data and scATACseq data were kept. In scRNA-seq data, cells with total read count greater than 1.8 times of the mean total read count across all cells were removed. In scATAC-seq data, cells with total epigenomic feature count greater than 1.8 times of the mean total epigenomic feature count across all cells are removed. We further removed cells with fewer than 2% of the genes detected in either scRNA-seq or scATAC-seq, and then the genes detected in fewer than 2% of the cells in either scRNA-seq or scATAC-seq. We applied SifiNet to all the four datasets (CD8 scRNA-seq, CD8 scATAC-seq, CD4 scRNA-seq, CD4 scATAC-seq) separately and generated gene set topology networks for scRNA-seq results.

MassiveRNA

MassiveRNA dataset is a large dataset with more than 1 million cells [79]. 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. 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.

Data availability

Monoclonal dataset is available at the Genome Sequencing Archive (GSA) with accession CRA008966. IPF1 and IPF2 datasets are collected from study GSE135893. BoneMarrow dataset is collected from study GSE72857. CD8 and CD4 datasets are available at 10x geneomics website `https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k`, and are also included in SeuratData R package as pbmc.rna and pbmc.atac. Seurat annotations for CD8 and CD4 are available at 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 is used for generating numerical datasets. `GeneScape` package is available at `https://github.com/qigao1/GeneScape`.

Acknowledgement

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

References

[1] 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". In: *Genome Biol* 20.1 (Dec. 2019), p. 295. DOI: 10.1186/s13059-019-1861-6.

[2] Tallulah S Andrews and Martin Hemberg. "M3Drop: dropout-based feature selection for scRNASeq". In: *Bioinformatics* 35.16 (Aug. 2019), pp. 2865–2867. DOI: 10.1093/bioinformatics/bty1044.

[3] Alexis Vandenbon and Diego Diez. "A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data". In: *Nat Commun* 11.1 (Aug. 2020), p. 4318. DOI: 10.1038/s41467-020-17900-3.

[4] Luke Zappia, Belinda Phipson, and Alicia Oshlack. "Splatter: simulation of single-cell RNA sequencing data". In: *Genome Biol* 18.1 (Sept. 2017), p. 174. DOI: 10.1186/s13059-017-1305-0.

[5] 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". In: *Mol Cancer* 8 (Feb. 2009), p. 8. DOI: 10.1186/1476-4598-8-8.

[6] Arianne Heinrichs. "A centrosome-integrity checkpoint". In: *Nature Reviews Molecular Cell Biology* 8.2 (Feb. 2007), pp. 98–98. DOI: 10.1038/nrm2107.

[7] 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". In: *Proc Natl Acad Sci U S A* 102.43 (Oct. 2005), pp. 15545–50. DOI: 10.1073/pnas.0506580102.

[8] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. "Molecular signatures database (MSigDB) 3.0". In: *Bioinformatics* 27.12 (June 2011), pp. 1739–40. DOI: 10.1093/bioinformatics/btr260.

[9] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. "The Molecular Signatures Database (MSigDB) hallmark gene set collection". In: *Cell Syst* 1.6 (Dec. 2015), pp. 417–425. DOI: 10.1016/j.cels.2015.12.004.

[10] UniProt Consortium. "UniProt: the Universal Protein Knowledgebase in 2023". In: *Nucleic Acids Res* 51.D1 (Jan. 2023), pp. D523–D531. DOI: 10.1093/nar/gkac1052.

[11] D L Kellogg, D L Kellogg Jr, N Musi, and A M Nambiar. "Cellular Senescence in Idiopathic Pulmonary Fibrosis". In: *Curr Mol Biol Rep* 7.3 (Aug. 2021), pp. 31–40. DOI: 10.1007/s40610-021-00145-4.

[12] 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". In: *Pharmaceutics* 12.4 (Apr. 2020). DOI: 10.3390/pharmaceutics12040389.

[13] 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". In: *Nat Cell Biol* 20.7 (July 2018), pp. 789–799. DOI: 10.1038/s41556-018-0127-y.

[14] 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". In: *Cancers (Basel)* 10.4 (Mar. 2018). DOI: 10.3390/cancers10040093.

[15] 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". In: *Protein Cell* 13.10 (Oct. 2022), pp. 742–759. DOI: 10.1007/s13238-021-00894-z.

[16] Luc Rochette, Marianne Zeller, Yves Cottin, and Catherine Vergely. "Insights Into Mechanisms of GDF15 and Receptor GFRAL: Therapeutic Targets". In: *Trends Endocrinol Metab* 31.12 (Dec. 2020), pp. 939–951. DOI: 10.1016/j.tem.2020.10.004.

[17] Hélène Martini and João F Passos. "Cellular senescence: all roads lead to mitochondria". In: *FEBS J* 290.5 (Mar. 2023), pp. 1186–1202. DOI: 10.1111/febs.16361.

[18] Ruchi Kumari and Parmjit Jat. "Mechanisms of Cellular Senescence: Cell Cycle Arrest and Senescence Associated Secretory Phenotype". In: *Front Cell Dev Biol* 9 (Mar. 2021), p. 645593. DOI: 10.3389/fcell.2021.645593.

[19] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. "GSVA: gene set variation analysis for microarray and RNA-seq data". In: *BMC Bioinformatics* 14 (Jan. 2013), p. 7. DOI: 10.1186/1471-2105-14-7.

[20] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. Sept. 2020. arXiv: 1802.03426 [stat.ML].

[21] Monika S Kowalczyk, Itay Tirosh, Dirk Heckl, Tata Nageswara Rao, Atray Dixit, Brian J Haas, Rebekka K Schneider, Amy J Wagers, Benjamin L Ebert, and Aviv Regev. "Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells". In: *Genome Res* 25.12 (Dec. 2015), pp. 1860–72. DOI: 10.1101/gr.192237.115.

[22] José Alberto López-Domınénguez, Sandra Rodrıéguez-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". In: *Aging (Albany NY)* 13.10 (May 2021), pp. 13380–13392. DOI: 10.18632/aging.203110.

[23] 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". In: *Nat Immunol* 4.10 (Oct. 2003), pp. 1029–36. DOI: 10.1038/ni973.

[24] 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". In: *Immunity* 25.5 (Nov. 2006), pp. 731–44. DOI: 10.1016/j.immuni.2006.09.011.

[25] 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". In: *Immunity* 19.3 (Sept. 2003), pp. 451–62. DOI: 10.1016/s1074-7613(03)00242-5.

[26] 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". In: *PLoS One* 9.5 (May 2014), e96636. DOI: 10.1371/journal.pone.0096636.

[27] 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". In: *Stem Cells* 35.4 (Apr. 2017), pp. 886–897. DOI: 10.1002/stem.2562.

[28] 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". In: *Cell* 163.7 (Dec. 2015), pp. 1663–77. DOI: 10.1016/j.cell.2015.11.013.

[29] Bhavuk Garg, Hrishikesh M Mehta, Borwyn Wang, Ralph Kamel, Marshall S Horwitz, and Seth J Corey. "Inducible expression of a disease-associated ELANE mutation impairs granulocytic differentiation, without eliciting an unfolded protein response". In: *J Biol Chem* 295.21 (May 2020), pp. 7492–7500. DOI: 10.1074/jbc.RA120.012366.

[30] Sally A Clayton, Kalbinder K Daley, Lucy MacDonald, Erika Fernandez-Vizarra, Giovanni Bottegoni, John D O'Neil, Triin Major, Daniel Griffin, Qinqin Zhuang, Adeolu B Adewoye, Kieran Woolcock, Simon W Jones, Carl Goodyear, Aziza Elmesmari, Andrew Filer, Daniel A Tennant, Stefano Alivernini, Christopher D Buckley, Robert D S Pitceathly, Mariola Kurowska-Stolarska, and Andrew R Clark. "Inflammation causes remodeling of mitochondrial cytochrome c oxidase mediated by the bifunctional gene C15orf48". In: *Sci Adv* 7.50 (Dec. 2021), eabl5182. DOI: 10.1126/sciadv.abl5182.

[31] Rongzhong Huang, Zicheng Hu, Xiaorui Chen, Yu Cao, Hongrong Li, Hong Zhang, Yongyong Li, Liwen Liang, Yuxing Feng, Ying Wang, Wenhua Su, Zerui Kong, N D Melgiri, Lihong Jiang, Xingsheng Li, Jianlin Du, and Yunqing Chen. "The Transcription Factor SUB1 Is a Master Regulator of the Macrophage TLR Response in Atherosclerosis". In: *Adv Sci (Weinh)* 8.19 (Oct. 2021), e2004162. DOI: 10.1002/advs.202004162.

[32] Alexandar Tzankov, Fabio Facchetti, Beda Mühleisen, and Stefan Dirnhofer. "IRF8 Is a Reliable Monoblast Marker for Acute Monocytic Leukemias, But Does Not Discriminate Between Monoblasts and Plasmacytoid Dendritic Cells". In: *Am J Surg Pathol* 46.5 (May 2022), pp. 725–727. DOI: 10.1097/PAS.0000000000001874.

[33] Youjin Hu, Yixin Hu, Yuhua Xiao, Feng Wen, Shaochong Zhang, Dan Liang, Lishi Su, Yang Deng, Jiawen Luo, Jingsong Ou, Mingzhi Lu, Yanhua Hong, and Wei Chi. "Genetic landscape and autoimmunity of monocytes in developing Vogt-Koyanagi-Harada disease". In: *Proc Natl Acad Sci U S A* 117.41 (Oct. 2020), pp. 25712–25721. DOI: 10.1073/pnas.2002476117.

[34] 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". In: *Nat Biotechnol* 32.4 (Apr. 2014), pp. 381–386. DOI: 10.1038/nbt.2859.

[35] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner, and Cole Trapnell. "Reversed graph embedding resolves complex single-cell trajectories". In: *Nat Methods* 14.10 (Oct. 2017), pp. 979–982. DOI: 10.1038/nmeth.4402.

[36] 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". In: *Nature* 566.7745 (Feb. 2019), pp. 496–502. DOI: 10.1038/s41586-019-0969-x.

[37] V A Traag, L Waltman, and N J van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". In: *Sci Rep* 9.1 (Mar. 2019), p. 5233. DOI: 10.1038/s41598-019-41695-z.

[38] 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". In: *Cell* 162.1 (July 2015), pp. 184–97. DOI: 10.1016/j.cell.2015.05.047.

[39] Dietmar Herndler-Brandstetter, Harumichi Ishigame, Ryo Shinnakasu, Valerie Plajer, Carmen Stecher, Jun Zhao, Melanie Lietzenmayer, Lina Kroehling, Akiko Takumi, Kohei Kometani, Takeshi Inoue, Yuval Kluger, Susan M Kaech, Tomohiro Kurosaki, Takaharu Okada, and Richard A Flavell. "KLRG1+ Effector CD8+ T Cells Lose KLRG1, Differentiate into All Memory T Cell Lineages, and Convey Enhanced Protective Immunity". In: *Immunity* 48.4 (Apr. 2018), 716–729.e8. DOI: 10.1016/j.immuni.2018.03.015.

[40] 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". In: *Front Immunol* 12 (2021), p. 757151. DOI: 10.3389/fimmu.2021.757151.

[41] 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". In: *Immunity* 39.2 (Aug. 2013), pp. 286–97. DOI: 10.1016/j.immuni.2013.07.013.

[42] 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." eng. In: *Cell Immunol* 353 (July 2020), p. 104118. DOI: 10.1016/j.cellimm.2020.104118.

[43] 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". In: *Cell Rep* 35.6 (May 2021), p. 109118. DOI: 10.1016/j.celrep.2021.109118.

[44] 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". In: *Nature* 552.7685 (Dec. 2017), pp. 362–367. DOI: 10.1038/nature24633.

[45] Rafael de Albuquerque, Elina Komsi, Inna Starskaia, Ubaid Ullah, and Riitta Lahesmaa. "The role of Interleukin-32 in autoimmunity". In: *Scand J Immunol* 93.2 (Feb. 2021), e13012. DOI: 10.1111/sji.13012.

[46] Jichun Xie and Ruosha Li. "False discovery rate control for high dimensional networks of quantile associations conditioning on covariates". In: *J R Stat Soc Series B Stat Methodol* 80.5 (Nov. 2018), pp. 1015–1034. DOI: 10.1111/rssb.12288.

[47] David L. Davies and Donald W. Bouldin. "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (Apr. 1979), pp. 224–227. DOI: 10.1109/TPAMI.1979.4766909.

[48] Lovisa Lovmar, Annika Ahlford, Mats Jonsson, and Ann-Christine Syvänen. "Silhouette scores for assessment of SNP genotype clusters". In: *BMC Genomics* 6 (Mar. 2005), p. 35. DOI: 10.1186/1471-2164-6-35.

[49] T. Caliński and J. Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics* 3.1 (1974), pp. 1–27. DOI: 10.1080/03610927408827101. eprint: https://www.tandfonline.com/doi/pdf/10.1080/03610927408827101. URL: https://www.tandfonline.com/doi/abs/10.1080/03610927408827101.

[50] Bradley Efron. "Large-Scale Simultaneous Hypothesis Testing". In: *Journal of the American Statistical Association* 99.465 (Mar. 2004), pp. 96–104. DOI: 10.1198/016214504000000089. URL: https://doi.org/10.1198/016214504000000089.

[51] Wenguang Sun and T. Tony Cai. "Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control". In: *Journal of the American Statistical Association* 102.479 (Sept. 2007), pp. 901–912. DOI: 10.1198/016214507000000545. URL: https://doi.org/10.1198/016214507000000545.

[52] Jiashun Jin and T. Tony Cai. "Estimating the Null and the Proportion of Nonnull Effects in Large-Scale Multiple Comparisons". In: *Journal of the American Statistical Association* 102.478 (June 2007), pp. 495–506. DOI: 10.1198/016214507000000167. URL: https://doi.org/10.1198/016214507000000167.

[53] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biol* 15.12 (2014), p. 550. DOI: 10.1186/s13059-014-0550-8.

[54] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (Jan. 2010), pp. 139–40. DOI: 10.1093/bioinformatics/btp616.

[55] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation". In: *Nucleic Acids Res* 40.10 (May 2012), pp. 4288–97. DOI: 10.1093/nar/gks042.

[56] 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". In: *F1000Res* 5 (June 2016), p. 1438. DOI: 10.12688/f1000research.8987.2.

[57] 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". In: *Nucleic Acids Res* 43.7 (Apr. 2015), e47. DOI: 10.1093/nar/gkv007.

[58] 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". In: *Ann Appl Stat* 10.2 (June 2016), pp. 946–963. DOI: 10.1214/16-AOAS920.

[59] Gordon K Smyth. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments". In: *Stat Appl Genet Mol Biol* 3 (Feb. 2004), Article3. DOI: 10.2202/1544-6115.1027.

[60] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". In: *Genome Biol* 15.2 (Feb. 2014), R29. DOI: 10.1186/gb-2014-15-2-r29.

[61] 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". In: *Genome Biol* 16 (Dec. 2015), p. 278. DOI: 10.1186/s13059-015-0844-5.

[62] Andrew McDavid, Greg Finak, and Masanao Yajima. *MAST: Model-based Analysis of Single Cell Transcriptomics*. R package version 1.20.0. 2021. URL: https://github.com/RGLab/MAST/.

[63] 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". In: *Cell* 184.13 (June 2021), 3573–3587.e29. DOI: 10.1016/j.cell.2021.04.048.

[64] 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". In: *Cell* 177.7 (June 2019), 1888–1902.e21. DOI: 10.1016/j.cell.2019.05.031.

[65] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. "Integrating single-cell transcriptomic data across different conditions, technologies, and species". In: *Nat Biotechnol* 36.5 (June 2018), pp. 411–420. DOI: 10.1038/nbt.4096.

[66] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. "Spatial reconstruction of single-cell gene expression data". In: *Nat Biotechnol* 33.5 (May 2015), pp. 495–502. DOI: 10.1038/nbt.3192.

[67] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast unfolding of communities in large networks". In: 2008.10 (Oct. 2008), p. 10008. DOI: 10.1088/1742-5468/2008/10/P10008. URL: https://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

[68] Peijie Lin, Michael Troup, and Joshua W K Ho. "CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data". In: *Genome Biol* 18.1 (Mar. 2017), p. 59. DOI: 10.1186/s13059-017-1188-0.

[69] Kelly Street, F. William Townes, Davide Risso, and Stephanie Hicks. *scry: Small-Count Analysis Methods for High-Dimensional Data*. R package version 1.6.0. 2021. URL: https://bioconductor.org/packages/scry.html.

[70] Tallulah Andrews. *M3Drop: Michaelis-Menten Modelling of Dropouts in single-cell RNASeq*. R package version 1.20.0. 2021. URL: https://github.com/tallulandrews/M3Drop.

[71] Charlotte Soneson and Mark D Robinson. "Bias, robustness and scalability in single-cell differential expression analysis". In: *Nat Methods* 15.4 (Apr. 2018), pp. 255–261. DOI: 10.1038/nmeth.4612.

[72] 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". In: *Genome Biol* 23.1 (Dec. 2022), p. 269. DOI: 10.1186/s13059-022-02825-5.

[73] 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". In: *Nucleic Acids Res* 50.D1 (Jan. 2022), pp. D687–D692. DOI: 10.1093/nar/gkab1028.

[74] Darryl Nishimura. "BioCarta". In: *Biotech Software & Internet Report* 2.3 (2023/05/17 2001), pp. 117–120. DOI: 10.1089/152791601750294344. URL: https://doi.org/10.1089/152791601750294344.

[75] M Kanehisa and S Goto. "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic Acids Res* 28.1 (Jan. 2000), pp. 27–30. DOI: 10.1093/nar/28.1.27.

[76] Minoru Kanehisa. "Toward understanding the origin and evolution of cellular organisms". In: *Protein Sci* 28.11 (Nov. 2019), pp. 1947–1951. DOI: 10.1002/pro.3715.

[77] 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". In: *Sci Adv* 6.28 (July 2020), eaba1972. DOI: 10.1126/sciadv.aba1972.

[78] 10x Genomics. *PBMC from a healthy donor - granulocytes removed through cell sorting (10k)*. Single Cell Multiome ATAC + Gene Exp. Dataset by Cell Ranger ARC 1.0.0. 2020. URL: https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k.

[79] 10x Genomics. *1.3 Million Brain Cells from E18 Mice*. Single Cell Gene Expression Dataset by Cell Ranger 1.3.0. 2017. URL: https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons.