Computational and Structural Biotechnology Journal 21 (2023) 5382–5393

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

# SCALA: A complete solution for multimodal analysis of single-cell Next Generation Sequencing data

Christos Tzaferis [a], Evangelos Karatzas [b], Fotis A. Baltoumas [b], Georgios A. Pavlopoulos [b,c,*,1], George Kollias [a,c,d,**,1], Dimitris Konstantopoulos [a,***,1]

[a] Institute for Bioinnovation, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece
[b] Institute for Fundamental Biomedical Research, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece
[c] Research Institute of New Biotechnologies and Precision Medicine, National and Kapodistrian University of Athens, Greece
[d] Department of Physiology, Medical School, National and Kapodistrian University of Athens, Greece

ARTICLE INFO

*Keywords:*
Single-cell RNA sequencing analysis
Single-cell ATAC-seq analysis
Automated analysis of single-cell Next Generation Sequencing data
Integrative analysis of single-cell Next Generation Sequencing data

ABSTRACT

Analysis and interpretation of high-throughput transcriptional and chromatin accessibility data at single-cell (sc) resolution are still open challenges in the biomedical field. The existence of countless bioinformatics tools, for the different analytical steps, increases the complexity of data interpretation and the difficulty to derive biological insights. In this article, we present SCALA, a bioinformatics tool for analysis and visualization of single-cell RNA sequencing (scRNA-seq) and Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) datasets, enabling either independent or integrative analysis of the two modalities. SCALA combines standard types of analysis by integrating multiple software packages varying from quality control to the identification of distinct cell populations and cell states. Additional analysis options enable functional enrichment, cellular trajectory inference, ligand-receptor analysis, and regulatory network reconstruction. SCALA is fully parameterizable, presenting data in tabular format and producing publication-ready visualizations. The different available analysis modules can aid biomedical researchers in exploring, analyzing, and visualizing their data without any prior experience in coding. We demonstrate the functionality of SCALA through two use-cases related to TNF-driven arthritic mice, handling both scRNA-seq and scATAC-seq datasets. SCALA is developed in R, Shiny and JavaScript and is mainly available as a standalone version, while an online service of more limited capacity can be found at http://scala.pavlopouloslab.info or https://scala.fleming.gr.

## 1. Introduction

Single-cell RNA sequencing (scRNA-seq) and ATAC sequencing (scATAC-seq) are both Next Generation Sequencing (NGS) techniques that have enabled the study of the transcriptome and epigenome , respectively, at an unprecedented resolution [1–5]. Exploitation of these two modalities allows researchers to observe the heterogeneity of cell populations in more depth compared to established bulk RNA-sequencing techniques.

Since the first scRNA-seq publication [6], advances in technology and equipment have led to an exponential increase in the number of cells (from hundreds to millions) that can be simultaneously sequenced in one run. Widely used technologies that have been introduced over the past ten years include Fluidigm C1 [7], Smart-seq2 [8], Drop-seq [9] and 10x Genomics [10], whereas new protocols such as the 10x multiome and spatial transcriptomics [11] have also emerged. Both scRNA-seq and scATAC-seq techniques have been used in various experimental settings such as the investigation of different tissues, developmental timepoints, disease states and organisms. ScRNA-seq for example, has played a crucial role in the comprehensive annotation of cell types in multiple organisms (e.g., Human Cell Atlas for *Homo sapiens* [12], Tabula Muris for *Mus musculus* [13]), as well as in the identification of novel cell

* Corresponding author at: Institute for Fundamental Biomedical Research, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece.
** Corresponding author at: Institute for Bioinnovation, Biomedical Sciences Research Center "Alexander Fleming", Vari, Greece.
*** Corresponding author.
*E-mail addresses:* pavlopoulos@fleming.gr (G.A. Pavlopoulos), kollias@fleming.gr (G. Kollias), konstantopoulos@fleming.gr (D. Konstantopoulos).
[1] Present Address: Georgios A. Pavlopoulos; George Kollias, Dimitris Konstantopoulos; Biomedical Sciences Research Center "Alexander Fleming", 34 Fleming Street, Vari, 16672, Greece.

https://doi.org/10.1016/j.csbj.2023.10.032
Received 11 June 2023; Received in revised form 16 October 2023; Accepted 16 October 2023
Available online 20 October 2023

populations, sub-populations and disease states [14]. Similarly, scATAC-seq has contributed critically to determining cell types in even higher resolution, as well as the epigenetic landscape that drives cellular differentiation, by characterizing gene regulation and inferring Gene Regulatory Networks (GRNs) in several species and disease systems. Characteristic examples of scATAC-seq analysis milestones are the case of *Drosophila melanogaster* brain epigenetic profiling [15], as well as the characterization of the chromatin accessibility profiles of 30 tissue types in Human [16].

From data generation to analysis and interpretation, a thorough bioinformatics pipeline is essential. Typical steps of such analyses include: *(i)* Quality Control (QC), *(ii)* read mapping and counting, *(iii)* normalization, *(iv)* dimensionality reduction, *(v)* clustering, *(vi)* differential expression/accessibility, *(vii)* peak calling, *(viii)* functional enrichment, *(ix)* data integration, *(x)* trajectory inference, *(xi)* generation of GRNs and *(xii)* visualization of results at every step. To this end, several software applications and packages that implement the aforementioned tasks have been proposed [17]. Seurat [18], Scanpy [19], Monocle [20,21], Cicero [22], Signac [23], EpiScanpy [24], SCENIC [25], cisTopic [26] and ArchR [27] are widely used R and Python libraries, whereas software applications which come with a graphical user interface (GUI), include Scope [28], CZ CELLxGENE [29] Azimuth [18], Cerebro [30], iCellR [31], PARTEK [32] and SeuratWizard [33].

Seurat and Scanpy have been primarily used for the analysis of scRNA-seq data, offering functionalities varying from QC to population identification and integration of multiple datasets, while Signac and EpiScanpy extend Seurat's and Scanpy's functionality to process scATAC-seq data. ArchR focuses on the analysis of single-cell chromatin accessibility data by offering standard analysis steps, as well as additional powerful features like Positive Regulator identification, Transcription Factor (TF) footprinting and trajectory inference. Monocle is a scRNA-seq analysis package that offers a widely used pseudo-temporal cell ordering framework, while its extension, Cicero can be used for the analysis of scATAC-seq data. Regarding tools with a GUI, Scope offers various visualization options including a side-by-side comparative view at cluster and gene levels for datasets containing multiple samples, disease conditions, or timepoints. While this is practical for exploring clustering results, GRNs and expression patterns, the tool lacks further downstream data analysis. CZ CELLxGENE can be used for the exploration of single-cell datasets and gene expression visualization across tissues in a collection of published datasets. Although it offers some analysis options such as detection of marker genes, it doesn't currently provide the option to perform more complex analytical tasks. Azimuth focuses on the basic scRNA-seq analysis steps but lacks customization options as it mainly specializes in the characterization of the identified populations by adopting a 'reference-based mapping' approach. SeuratWizard exploits the standard steps of the analysis, while Cerebro also builds upon the initial results, allowing the user to explore additional modes such as signature scoring, cell cycle phase analysis, and trajectory inference. Finally, iCellR covers both scRNA-seq and scATAC-seq basic analyses but lacks ligand-receptor and GRN reconstruction.

In this article, we present SCALA, a holistic pipeline that integrates all the aforementioned procedures and enables biomedical researchers to get actively involved in the downstream analysis and exploration of both scRNA-seq and scATAC-seq datasets. SCALA is a fully interactive bioinformatics tool that offers access to all standard analysis modes, varying from QC and data normalization to the identification of distinct cell populations and cell states. Furthermore, SCALA supports additional analysis modes such as automatic cluster annotation, functional enrichment analysis, ligand-receptor analysis, trajectory inference, and reconstruction of GRNs. Interactive plots as well as publication-ready figures and data tables can be generated at every step of the analysis while any of the processed datasets can be exported to be further analyzed with external applications. We believe that SCALA can become a go-to tool for experimentalists who seek to analyze their scRNA-seq and scATAC-seq datasets and communicate biological findings with high resolution visualizations.

## 2. Methods

SCALA is mainly developed in R/Shiny and JavaScript. For the basic analysis of scRNA-seq data, the Seurat package is utilized, while ArchR is employed for the basic steps of scATAC-seq analysis. Furthermore, downstream applications including functional enrichment analysis, trajectory inference, GRNs reconstruction, and L-R analysis were made possible by the incorporation of the g:Profiler [34], Slingshot [35], SCENIC [25], decoupleR [36,37] and nichenetR [38] packages. The aforementioned modes of analysis are described in detail in the following paragraphs, accompanied by output screenshots of pbmc and bmmc datasets (provided in the online vignettes of Seurat and ArchR packages). In the online version, the supported dataset size is restricted to $< 2$ GB while no more than 2 CPU threads on the server are allocated per session, something that may result in slow execution. However, by downloading the standalone version of SCALA from GitHub, users can bypass both aforementioned restrictions (settings variables in server.R and ui.R files). Additionally, a docker image is available for download, enabling dataset inputs of up to 100 GB and selection of up to 100 CPU threads.

### 2.1. Input data types

SCALA is compatible with several input data types. For scRNA-seq, the primary data input consists of a unique molecular identifier (UMI) count matrix. The user can provide such a matrix by either uploading a gene (rows: features) by cell (columns: barcodes) tab-delimited data table (including row and column names) or by uploading the output of the cellranger pipeline from 10X (filtered_bc_matrix). In the latter case, the "cellranger count" output folder should contain: (i) a file named "barcodes.tsv.gz" containing only detected (filtered by cellranger count pipeline) cellular barcodes in gzip CSV format, (ii) a file named "features.tsv.gz" with features (genes) that correspond to row indices in gzip TSV format; the columns of the particular file should correspond to feature ID, feature name and feature type (Gene expression) respectively, (iii) a feature-barcode count matrix in gzip Market Exchange Format (MEX). Moreover, the user has a third option of uploading a pre-analyzed Seurat object in RDS (R saved object) format. In the case of scATAC-seq, SCALA is only compatible with arrow files in its current version. The particular file format stores all the associated data (i.e., metadata, accessible fragments, and data matrices) within a sample. Arrow files can be created by using the create_arrow_file.R helper script provided in SCALA's GitHub repository or by using the ArchR package directly. It is worth noting that in both modalities, the analysis of only human and mouse datasets is currently supported.

### 2.2. Functionality

After the input files have been uploaded, SCALA's main workflow (Fig. 1) can be utilized for both SC pipelines. The steps are: (i) QC, (ii) data normalization and scaling, (iii) variable features detection, (iv) Principal Component Analysis (PCA) dimensionality reduction, (v) Latent Semantic Indexing (LSI) dimensionality reduction, (vi) clustering, (vii) additional dimensionality reduction methods, (viii) feature inspection, (ix) markers' identification, (x) cell cycle phase analysis, (xi) functional/motif enrichment analysis, (xii) clusters' annotation, (xiii) trajectory analysis, (xiv) Ligand-Receptor (L-R) analysis, (xv) GRN analysis and (xvi) visualization of epigenome signal tracks.

### 2.3. Quality control

Identification and removal of "low quality" cells (empty, stressed, broken, or dead cells) and non-informative genes is essential for downstream analysis in SC datasets. SCALA allows the exploration of QC

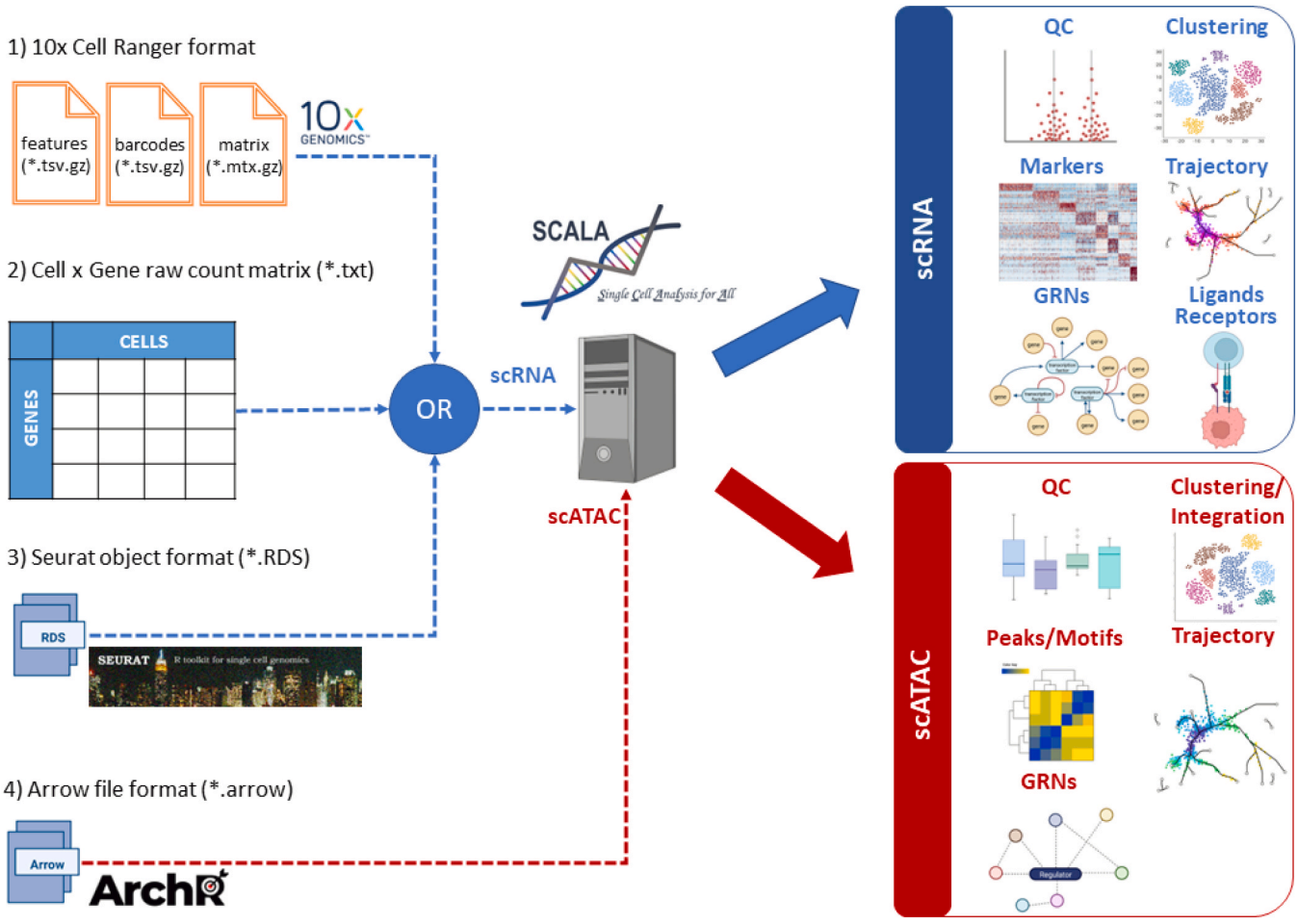


**Fig. 1.** General workflow of the SCALA pipeline. In this figure, the input files compatible with SCALA for scRNA-seq and scATAC-seq analysis are shown in the left panel. Additionally, the main functionalities and outputs, for each mode of analysis for RNA (blue box) and ATAC (red box) assays, are showcased in the right panel.

plots and filters out cell barcodes through the application of user-defined parameter thresholds. Common scRNA-seq QC criteria include *(i)* per cell number of unique features detected, *(ii)* per cell detected UMIs, and *(iii)* per cell percentage of mitochondrial content. Cells that exhibit very low numbers of *(i)* and *(ii)* are typically excluded as low-quality, while those with very high numbers are considered as putative multiplets. Barcodes with a high percentage of mitochondrial UMIs should also be excluded as low-quality/dying cells (Fig. 2A).

Similarly, typical scATAC-seq QC metrics include: *(i)* transcription start site (TSS) enrichment and *(ii)* the number of unique nuclear fragments (*log10(nFrags)*). TSS represents the chromatin accessibility signal-to-background ratio. Enrichment of ATAC-seq signal in TSS regions of expressed genes is typically high in most cell types and a classic criterion of the quality of the assay. The particular metric is calculated as the ratio between TSS enrichment relative to per-base pair 2 kb flanking regions enrichment. Furthermore, cells including too few nuclear fragments should be excluded in order to avoid the inclusion of non-interpretable data (Fig. 3A).

### 2.4. Data normalization and scaling

scRNA matrices are normalized and scaled in order to eliminate cell-depth variability biases as well as to transform the data properly before variable feature detection and dimensionality reduction. Data normalization in SCALA is applied through a global-scaling normalization method [18], where the gene count of each barcode is normalized by the total barcode counts, multiplied by 10,000, and log-transformed. Normalized values are stored in a *Seurat* object, and normalized counts are further standardized to z-scores, with column-wise mean expression equal to 0 and variance equal to 1. To mitigate the effect of unwanted sources of variation, the user can optionally provide metadata variables. In such a case, they are individually regressed against each feature, while scaling and centering is then performed on the resulting residuals.

### 2.5. Variable features detection

Using the normalized RNA data matrix, genes that exhibit the highest column-wise variation are detected. Targeted analysis of the particular subset of features aids in the identification of the underlying biological patterns in single-cell datasets. The supported methods for the identification of most variable features (supp. Fig. 1A) include three methods. These are *(i)* Variance Stabilizing Transformation (VST), *(ii)* Mean-Variance Plot selection (MVP), and *(iii)* "Dispersion". VST fits a line in the relationship of log-variance and log-mean using local polynomial regression. Consequently, standardization of feature values using the observed mean and expected variance is performed. Feature variance is finally calculated for standardized values, after clipping to a maximum. A fixed number (default = 2000) of variable features is returned. MVP uses a function to calculate average gene counts and gene dispersions. In this function, all genes are separated into 20 bins according to their average counts. Finally, dispersion z-scores are calculated for each gene group. For "Dispersion", genes with the highest dispersion values are kept. For the last two methods, a variable number of features is returned.

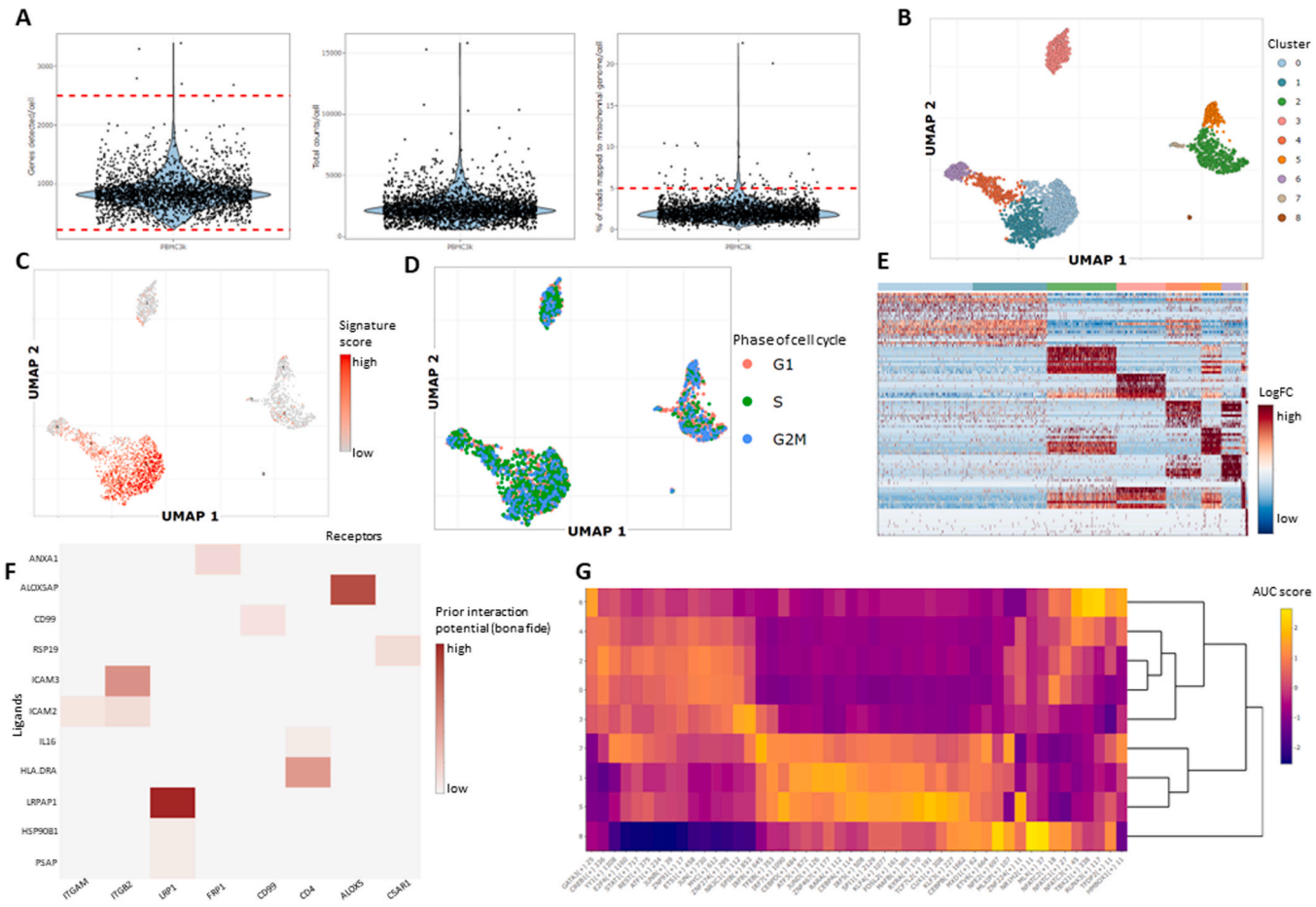


**Fig. 2.** Analysis of PBMC3k scRNA-seq dataset. (A) Violin plots depicting cell quality control measurements including the number of genes detected, the total number of reads, and the percentage of reads mapped to the mitochondrial genome. (B) Visualization of cells in UMAP space. Cells are colored according to cluster labels (clusters were identified with the Louvain algorithm). (C) Feature plot showcasing signature scores/per cell for the top marker genes of cluster 0. Color scale denotes the intensity of signature score. Red color indicates high intensity values, while grey indicates low intensity values. (D) UMAP projection showcasing the results of cell cycle phase analysis. Cells are colored according to the phase of the cell cycle they are predicted to belong to. (E) Heatmap depicting the top10 marker genes per cluster, ranked by Log2FC value. Genes are shown in y-axis and cells are shown in x-axis. Color scale denotes scaled expression values, with blue color indicating low expression and red indicating high expression. (F) Heatmap showing "bona fide" interactions between clusters 0 (ligand expressing cluster) and 2 (receptor expressing cluster). The intensity of the color represents the interaction potential score (high intensity is represented with red, while low is represented with grey). (G) Heatmap of scaled AUC scores for the top regulons per cluster. Color scale denotes z-scores of AUC values (high values represented with yellow color, while low values are represented with purple).

### 2.6. PCA dimensionality reduction

PCA is performed on the scaled values of the most variable features to determine the "dimensionality" of the dataset. The most informative Principal Components (PCs) are identified and used in the next steps of cell clustering and cluster visualization. The number of PCs that exhibit the higher variation of the scRNA matrix can be determined either automatically by applying *10-fold Singular Value Decomposition* (SVD) cross validation or manually by examining the ranking of the incremental variance of each PC (elbow plot). Additionally, loading scores for the top genes of a principal component can be plotted (Supp. Fig. 1B).

### 2.7. LSI dimensionality reduction

LSI is performed in scATAC-seq matrices, using genome-wide 500-bp tile counts [27]. Tile-counts are normalized to eliminate the cell depth bias using a constant of 10,000, followed by inverse document frequency normalization and log-transformation. During this process, the most variable features (tiles) are detected. The aforementioned process is run in an iterative manner where an LSI transformation is applied using the most accessible features (tiles). This procedure identifies lower resolution clusters that are not batch confounded. Consequently, average accessibility for each of these clusters is calculated across all features. Finally, the most variable features are identified across low-resolution clusters and are used as input for the next LSI iteration.

### 2.8. Clustering

Graph-based clustering is performed in scRNA-seq (Fig. 2B) and scATAC-seq (Fig. 3B) matrices, in order to define cell types and/or cellular states. Initially, cells are embedded in a *Shared-Nearest Neighbor* (SNN) graph structure based on Euclidean distances in the PCA/LSI space. Cells that exhibit similar gene expression/chromatin accessibility profiles are connected with edges. The newly formed graph is then partitioned into highly interconnected communities using the Louvain algorithm [39].

### 2.9. Additional dimensionality reduction methods

To visualize cells, cell clusters, and cluster relationships in 2D and 3D space, additional dimensionality reduction techniques, are applied, like uniform manifold approximation and projection (UMAP) [40] (Fig. 3C),

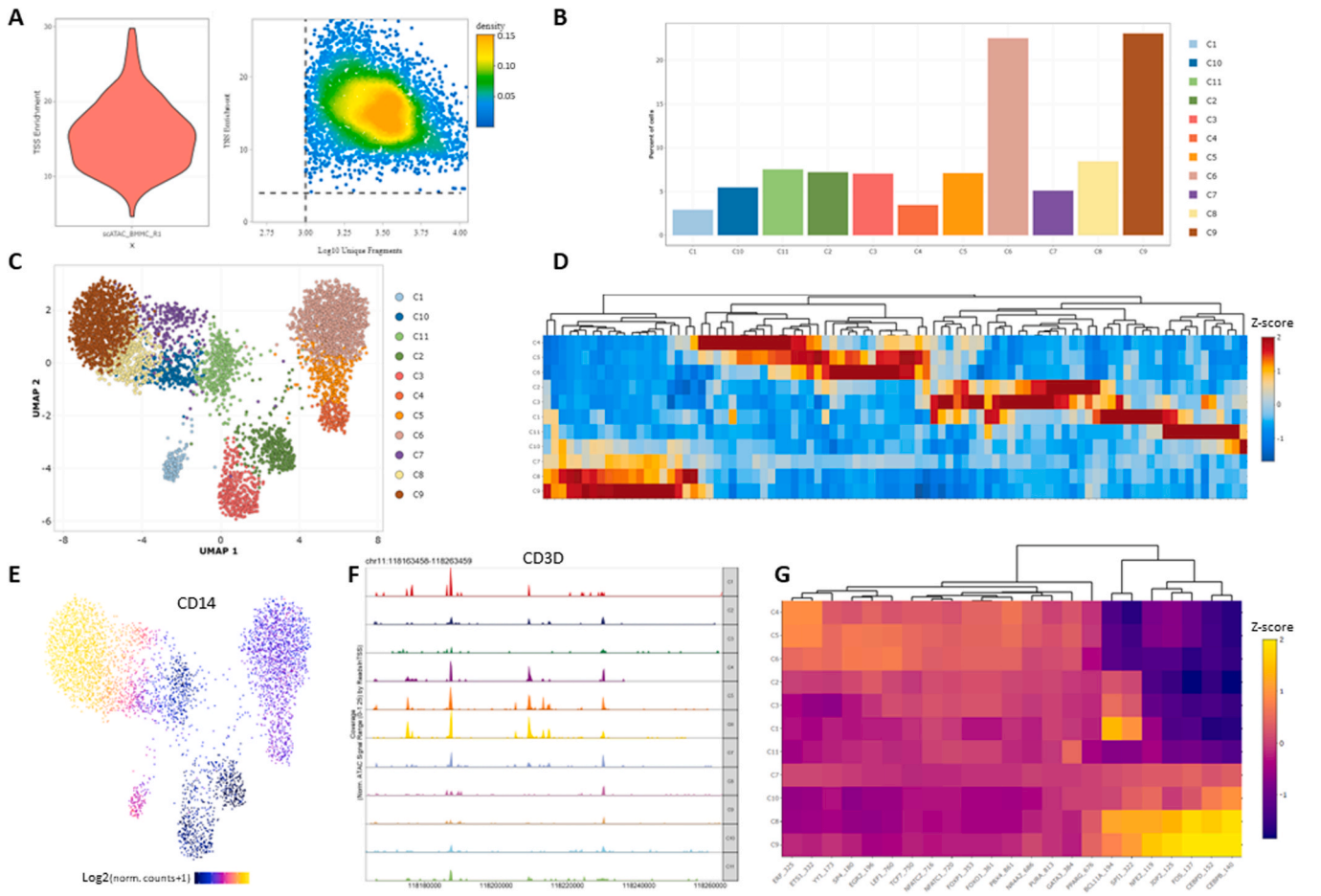


**Fig. 3.** Analysis of BMMCs scATAC-seq dataset. (A) Cell quality control plots depicting information about TSS enrichment and unique fragments measurements. (B) Bar plot showing the relative abundance of cells in each of the dataset's clusters (clusters were identified with the Louvain algorithm). (C) Projection of cells in UMAP space. Cells are colored according to cluster identity. (D) Heatmap showing z-scores of peak accessibility for the top marker peaks per cluster. Clusters are shown in y-axis, while peaks are plotted in x-axis. (E) Feature plot showcasing gene activity scores (per cell) of CD14 as a UMAP overlay. Intensity of the color denotes imputed log2 normalized expression values. (F) Genome browser tracks showing local chromatin accessibility (y-axis left panel) of CD3D gene at cluster level (y-axis right panel). (G) Heatmap displaying motif deviations z-scores of positive regulators for all clusters. Regulators are shown in x-axis, while clusters are shown in y-axis.

t-Distributed Stochastic Neighbor Embedding (tSNE) [41], diffusion maps [42] or Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE) [43]. Such visualizations uncover the underlying modularity of the datasets. Additionally, they can be utilized for feature inspection, exploration of cluster structures, and trajectory inference purposes (especially PHATE).

### 2.10. Markers' identification

Differential expression as well as differential accessibility analysis enable the identification of marker genes (Fig. 2E) and peaks (Fig. 3D) respectively and guide the cell-type and cell-state annotation/characterization of cell clusters. Differential analysis assists in the detection of key transcriptional and regulatory programs that drive pathogenicity and/or development. The analysis is performed in a cluster-specific manner, where each cluster's cells are tested against all the other cells of the dataset. The available statistical tests for scRNA-seq are: *(i)* Wilcoxon rank sum test, *(ii)* likelihood-ratio test for single-cell feature expression [44], *(iii)* standard AUC classifier, *(iv)* Student's t-test, *(v)* MAST [45] and *(vi)* DESeq2. Similarly, for scATAC-seq the tests are: *(i)* Wilcoxon rank sum test, *(ii)* Student's t-test and *(iii)* binomial test.

### 2.11. Feature inspection

Feature expression and chromatin activity can be explored by cell scatter plots (Fig. 3E) in reduced space (e.g., UMAP, tSNE, etc.), or via violin plots, heatmaps and dotplots. In scRNA-seq datasets, gene signatures can also be calculated by utilizing the UCell package and visualized as described above (Fig. 2C). Moreover, QC metrics such as total number of reads per cell, genes detected per cell, etc. can be visualized via scatter plots and violin plots at a cluster level.

### 2.12. Doublet detection

Doublet detection in scRNA-seq datasets is performed utilizing the R package DoubletFinder [46]. More specifically, artificial doublets are initially simulated and incorporated into the original data. Then cells that have a high number of artificial neighbors, in the gene expression space, are considered as potential doublets and can be removed from downstream analysis. This methodology exhibits higher accuracy in detecting doublets originating from transcriptional distinct cell types. Regarding scATAC-seq datasets, a similar approach is followed in the ArchR package, facilitating the identification of potential doublets. After the calculation of doublet enrichment measurements, the user can remove doublets by selecting a value for the filterRatio parameter (higher values lead to more cells removed as potential doublets).

### 2.13. Cell cycle phase analysis

Calculation of cell cycle phase scores is based on S, G2/M and G1

canonical markers in the scRNA-seq count matrix. In cases where cluster-specific patterns of cell cycle biases are captured, the user has the option to use the “regress out” option (in the step of scaling) in order to mitigate the cell-cycle effect. The results of the analysis can be viewed either in a scatter plot format, where cells are projected in a reduced space (PCA, UMAP, tSNE, diffusion map, PHATE) and colored according to the predicted phase of the cell cycle (Fig. 2D) or as a bar plot summarizing the percentages of cells assigned to each cell cycle phase per cluster.

### 2.14. Functional/Motif enrichment analysis

Using the previously identified marker genes and peaks, functional enrichment analysis (e.g., for pathways and Gene Ontologies (GOs)) and motif enrichment analysis can be performed for each cluster. In detail, for scRNA-seq data, up/down regulated genes from the clusters identified in previous steps are tested for enriched GO terms or KEGG pathways, using the g:Profiler package [34]. The enriched terms can be visualized in a table format accompanied by information about statistical significance and gene overlap (between the input list and the term of interest). Additionally, a bubble plot summarizing the enriched terms per database used is also available (Supp. Fig. 1C). Regarding motif enrichment analysis, marker peaks identified in previous steps are tested for enrichment of binding sites of specific transcription factors (TFs). Finally, deeper functional enrichment analysis with more informative visualization is also offered by Flame [47] external application. This can be done per cluster (one gene list), or multiple gene lists (up to 10 clusters) can be analyzed simultaneously with the help of interactive UpSet plots (Supp. Fig. 2A).

### 2.15. Cluster annotation

For automated cluster annotation, the CIPR package is utilized [48] as it contains reference datasets for human and mouse organisms, to assign cell-type identities. The end user can select a dataset to use as a reference and the type of analysis that will be employed to calculate the predictions, either by keeping all dataset genes or only the differentially expressed ones. Moreover, the user can also select the correlation metric (Pearson, Spearman) that will be used. Regarding the visualization options, a table containing all predictions per cluster is returned, as well as a dot plot that summarizes the top-5 predictions per cluster (Supp. Fig. 1D).

### 2.16. Multimodal integration analysis

In this mode of analysis, the user can upload an already processed scRNA-seq dataset to perform integration analysis with the scATAC-seq dataset, which is currently loaded in SCALA. More specifically, gene activity scores from the ATAC assay and gene expression values from the RNA assay are combined in order to align cells between datasets. The output of integration analysis results in transferring labels from scRNA clusters to cells from the scATAC dataset. The newly obtained clustering identities of the cells can be adopted in other downstream steps such as detection of marker peaks, trajectory analysis, etc.

### 2.17. Trajectory analysis

Pseudotemporal ordering of single-cells facilitates the uncovering of underlying differentiation/developmental processes which lead cells to transitions between different cellular states. In SCALA, Slingshot [35] is employed, utilizing input clustering information and dimensionality reduction coordinates for all cells of a dataset, in order to construct a Minimal Spanning Tree (MST) at the cluster level. The nodes of the tree represent the clusters while the edges represent their in-between relationships. The user can select the dimensionality reduction method that will be used for the Slingshot execution (PCA, UMAP, tSNE, diffusion map, or PHATE), as well as the initial and final states, which define the direction of the identified trajectory. The MST is always drawn in a UMAP plot, while pseudotime values are calculated per lineage and can be visualized again in UMAP space as a separate scatter plot.

### 2.18. Ligand - receptor analysis

The prediction of ligand-receptor interactions is a crucial step for deciphering cell-to-cell communication in different tissues. Inspection of communication patterns between different cell types could aid in the detection of key interactions, driving gene expression alterations (downstream of signaling pathways) in healthy and disease contexts. SCALA incorporates the analysis framework of nichenetR [38]. More specifically, after clustering the user needs to select a pair of clusters that will be used to search for L-R interactions. As a first step, overexpressed genes are calculated in each cluster. Then the reported interactions are ranked, by considering a "prior interaction potential" score that is calculated in the initial steps, when the protein-protein interaction model is constructed. A heatmap visualization summarizes all the interactions that have been detected between the two clusters of interest (Fig. 2F). L-R interactions and their respective scores are available for download in a table format including the prior interaction potential score that signifies the strength of the predicted interaction.

### 2.19. Gene regulatory network analysis

In this step, by utilizing the SCENIC workflow [25], co-expression modules of TFs (Transcription Factors) and their target genes are detected based on co-expression analysis and TF motif analysis. Area Under the Curve (AUC) scores per cell are calculated and denote the activity of a regulon, defined as a group of genes containing a TF and its target genes. Finally, average AUC values and Regulon Specificity Score (RSS) scores, which showcase the activity and specificity of regulons, can help the user visualize active regulatory networks in heatmap format and examine whether cluster-specific regulons are present in the dataset (Fig. 2G). Due to limitations in run-time in *R* environments, SCALA offers instructions so that an end-user can externally run some parts of the analysis in *Python* and then import the result files again in SCALA for visualization. An alternative option for users who do not wish to follow SCENIC analysis is the inference of TF activity levels. To achieve that we followed the approach proposed by decoupleR [36], which utilizes a curated resource [37] of interactions between TFs and their target genes. Gene regulation analysis at chromatin level aims to identify cluster-specific TFs, whose expression exhibits a high correlation with chromatin accessibility changes at genomic sites, that include their DNA binding motifs (positive regulators) (Fig. 3G).

### 2.20. Visualization of epigenome signal tracks

Chromatin accessibility tracks can be used as an alternative to feature plots (which depict gene scores in reduced space). The user can select a gene and the number of bases upstream and downstream, defining a genomic interval of interest. The inspection of the plot (Fig. 3F) through a genome browser snapshot can reveal chromatin accessibility in the gene body or upstream/downstream gene regulatory elements (promoters, enhancers, silencers, etc.).

## 3. Results

### 3.1. Analysis of two datasets for synovial fibroblasts in arthritis mouse model

To demonstrate SCALA’s functionality, we utilized two previously published datasets [49] (scRNA-seq and scATAC-seq), which were produced to investigate the single-cell transcriptome and chromatin

dynamics of Synovial Fibroblasts transitioning from homeostasis to pathology in modeled TNF-driven arthritis. For this purpose, the human TNF (*hTNFtg*) transgene overexpressing mouse model Tg197 [50] was used and compared against healthy wild type (Wt) mice. As reported previously [49], for the generation of *10x Genomics* scRNA-seq libraries (Single-cell 3' v3 reagent kits), 6667 sorted non-hematopoietic stromal cells (*Cd45-*, *Cd31-*, *Ter119-*, *Pdpn+*) were isolated from whole ankle joint synovium. These libraries were sequenced at a depth of 400 million reads, using one lane of an Illumina NextSeq 500 machine. For the second dataset, scATAC-seq libraries were generated using a similar experimental set-up, according to *10x Genomics* guidelines, profiling a total of 6679 single nuclei. In each experiment, cells were derived from three healthy mice tissues (WT, 4 weeks of age), and six diseased *hTNFtg* mice; three at an early disease stage (*hTNFtg*/4, 4 weeks of age) and three at an established pathological stage (*hTNFtg*/8, 8 weeks of age). As shown in previous publications [50,51], the Tg197 mouse model over-expresses human TNF (huTNF) transgene, leading to the development of an arthritic phenotype manifested by cartilage destruction and bone erosion, which ultimately results in loss of joint function. Here, SCALA was used to reanalyze 5903 synovial fibroblast (SFs) transcriptomes and 6046 epigenomes, originating from healthy mice (control sample) and arthritic mice at 4 and 8 weeks of age (early and established disease state).

### 3.2. Analysis using SCALA's scRNA-seq pipeline

For the scRNA-seq QC step, cells with $< 500$ features (genes) detected or having $> 10\%$ of their reads mapped to the mitochondrial genome, were excluded from further analysis. Consequently, downstream analysis of scRNA-seq was performed as follows: Most highly variable feature detection was performed by applying the mean-variance-plot (MVP) method implemented by the Seurat package, leading to the identification of 1535 variable genes. Gene counts of each cell were divided by the total cell counts, multiplied by 10,000, and natural-log transformed. Scaling of the normalized expression values was performed on all genes by utilizing the option of "regressing out" the mitochondrial reads effect.

The scaled gene-by-cell expression matrix of most variable genes was used as input to perform Principal Component Analysis (PCA). To identify the dimensionality of the dataset (most informative principal components in terms of cell heterogeneity), Singular Value Decomposition (SVD) *k*-fold cross-validation was performed using the dismo R library. This procedure determined the number of most informative principal components (25 PCs), which were used for the steps of cell clustering and non-linear dimensionality reduction analysis. Specifically, in order to identify distinct fibroblast subsets, graph-based clustering analysis was performed with Seurat's Louvain algorithm, by setting the resolution parameter to 0.6. The 25 most informative PCs were also used for non-linear dimensionality reduction analysis (tSNE and UMAP), to visualize the newly formed cell clusters in 2D/3D space.

SF clustering led to the formation of 10 SF clusters, with distinct transcriptional profiles, exhibiting homeostatic, inflammatory, and destructive characteristics in healthy and arthritic joints respectively. These characteristics were detected by performing marker gene identification analysis for each identified SF cluster. More specifically, each cluster's transcriptomes were compared against the rest of the cells' transcriptomes, through the Wilcoxon rank sum test on the normalized expression values. Genes with average log Fold Change (logFC) $> 0.25$, a percentage of expression (gene detected in a cell) $> 25\%$, and a p-value $< 0.01$ were retained.

Consequently, up-regulated genes were used as an input to perform functional enrichment analysis. In particular, GO biological processes enrichment was conducted for each SF cluster, using g:Profiler. Examination of similarities/differences between SF clusters at the level of markers and enriched terms led to the merging of two clusters, (0 and 9). The resulting nine clusters were designated as S1, S2a, S2b, S2c, S2d, S3, S4a, S4b and S5 (Fig. 4A). It should also be pointed out that the identified clusters exhibit differences in their relative abundances in healthy and diseased states. More specifically, one group of them is shrinking, while another is expanding during disease (Fig. 4B). Thy1 + clusters (S1, S2a, S2b, S2c, S3, and S5) were further annotated as "sublining". Interestingly, their transcriptional and functional characterization comprises features of tissue homeostasis preservation, except S5 which shows an immuno-regulatory role under healthy conditions. Enriched GO terms for these populations include BMP, WNT, TGFbeta, and SMAD signaling pathways, as well as response to TNF and IFN-beta/gamma. Top markers for these clusters contain *Smoc2, Thbs1, Vwa, Rgma, Dkk2, Sfrp1, Ecrg4, Osr1, Nr2f2, Klf5, Clu, Id1, Meox1, Pi16, Sema3c, Efemp1, Ccl7, Il6*, and *Notch3*. Similarly, the $Prg4^{High}$ S4a cluster was annotated as "lining" and was linked with functions that define an inflammatory-destructive profile for the particular SF subpopulation. The lining phenotype is described by markers like *Tspan15, Hbegf*, *Htra4,* and *Clic5*. Regarding the enriched biological processes we detected terms such as inflammatory response and class I antigen presentation. Finally, clusters S2d and S4b showed a mixed expression profile of *Prg4* and *Thy1* (Prg4 + Thy1 +) and thus were annotated as "intermediate" subpopulations. Marker genes such as *Fbln7, Thbs4, Cthrc1, Lrrc15, Dkk3, Mki67, Pdgfa, Birc5, Aqp1, Acta2, Cxcl5*, which were found upregulated mainly in the intermediate and lining compartments, are either previously reported as players of fibroblast pathogenicity or linked to potential pathogenic roles. Corresponding terms like regulation of immune, redox response, fibroblast proliferation, cell division, and apoptosis were found enriched in S2d and S4b. Conclusively, this group of clusters showcases a pro-inflammatory and proliferating character (Fig. 4C, Supp. Table 3).

Next, cell cycle phase analysis was performed, assigning each cell to S, G1, or G2/M phase. Interestingly, one of the three SF populations exhibiting pathogenic characteristics (S4b), showed the highest percentage of cells located in G2/M phase (Supp. Fig. 2C). This finding was also supported by cycling markers (extracted from the literature), that were specifically expressed in the S4b cluster. The mixed expression signature of *Prg4* and *Thy1* (Prg4 + Thy1 +), which characterize this "intermediate" group of cells, is thus a strong marker of disease state that is observed mainly in *hTNFtg* conditions.

Cellular trajectories were calculated for the pooled dataset using as an input to the slingshot algorithm the first 25 most informative PCs. In order to determine the clusters used as an input for the initial and final state of the trajectory, current literature [52,53] was taken into consideration as well as the results of external software applications such as scVelo [54] and CellRank [55]. The produced minimum spanning tree highlighted the existence of a pathogenic branch composed of clusters S2a, S2d, S4b, and S4a, indicating S4a as the final state and S1, S2b, S3, and S5 as potential starting points (Fig. 4D).

We next sought to study ligand-receptor interactions between the sublining and intermediate compartments with lining. By employing the nichnetR package and focusing on ligands/receptors with a percentage of expression $> 10\%$ in the clusters of interest, we identified shared and specific interactions. More particularly, we detected 157 and 152 interactions between sublining-lining and intermediate-lining respectively. In more detail, 126 of those interactions were shared, however, 26 were specific to intermediate-lining and 31 to sublining-lining. Interestingly, in the interactions of sublining and lining, we noticed pairs of ligands and receptors participating in Wnt and BMP signaling. In contrast, in the intermediate-lining interactions, we have detected pairs related to MMP13, IL-11, and RSPO2 signaling (Supp. Fig. 2D, Supp. Table 4).

As a last step in the scRNA-seq analysis pipeline, GRN analysis was performed to detect regulons that exhibit preferential activation patterns at the cluster level. That resulted in the identification of 133 regulons in total. Interestingly, distinct activation patterns were observable in the different clusters, and hierarchical clustering of the top-80 regulons revealed two groups, the first containing only sublining clusters

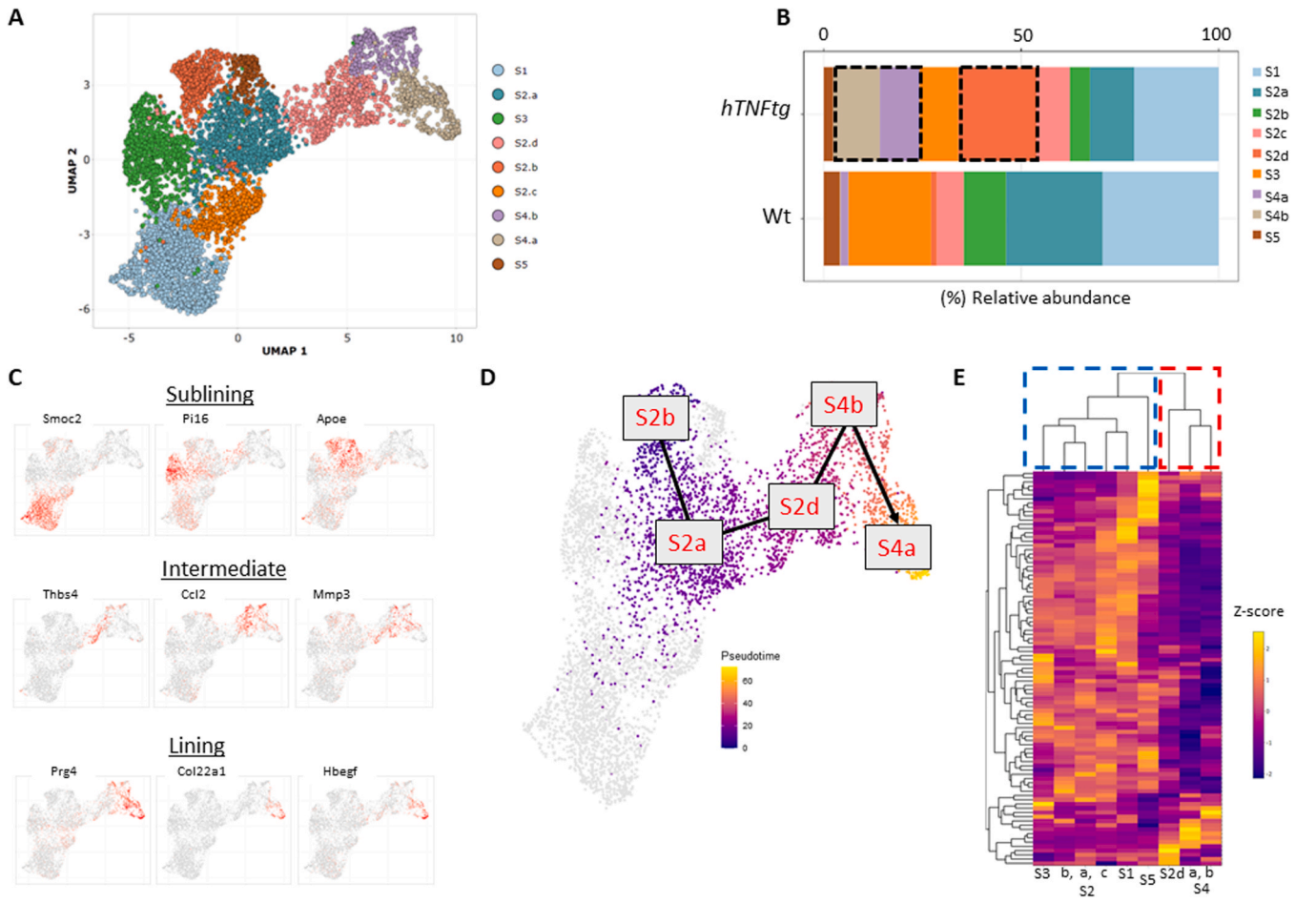


**Fig. 4.** Use case - hTNFtg scRNA-seq data analysis. (A) Graph based clustering of SFs identified 9 distinct clusters. Cells are visualized in UMAP space and are colored by cluster assignment. (B) The barplot depicts relative abundances of clusters in healthy (Wt) and disease (hTNFtg) states. The highlighted areas pinpoint the clusters that are expanded in arthritic state. (C) Feature plots showing the different gene expression patterns between the clusters of sublining (top row), intermediate (middle row), and lining (bottom row) categories. Cells are projected in the 2D UMAP space and colored by normalized gene expression. (D) One of the possible lineages (proposed by trajectory analysis) is showcased in UMAP overlay. Cells belonging to the lineage are colored according to their pseudo-time values, while cells that are not part of this lineage are colored in light gray. (E) Heatmap depicting regulon activity of top-80 regulons (z-scores of AUC values) at the cluster level. Hierarchical clustering of fibroblast subsets (using active regulons) identified two major groups (group1: sublining clusters, group2: intermediate and lining clusters).

and the second containing intermediate and lining (Fig. 4E).

### 3.3. Analysis using SCALA's scATAC-seq pipeline

Regarding the scATAC-seq data, QC was initially performed, and cells with Transcription Start Site (TSS) enrichment score $< 4$ and count-depth $< 1000$ unique nuclear fragments were removed from downstream analysis. Next, LSI was employed using a resolution of 0.6, a total of 30 dimensions, 4 iterations, and otherwise default settings. Additionally, UMAP projection was produced for the visualization of cells in 2D space. Gene activity scores were computed as the summed local accessibility of promoter-associated count-tiles in the proximity of each gene, adopting a distance-weighted accessibility model. In detail, count-tiles in the range of 100,000 bp of a gene promoter were aggregated using the following distance weight formula: $e^{(-|\frac{distance}{5000}|)} + e^{-1}$. An extra normalization step was applied (multiplication by $\frac{1}{gene\ size}$, scaled linearly from 1 to 5), in order to account for gene length biases. As a following step, the above-weighted sum was multiplied by the aggregated Tn5 insertions in each tile. Gene scores were then scaled to 10,000 counts and log2-transformed. To improve the visualization of gene activity scores, a smoothing procedure was applied using the MAGIC algorithm [56].

Similar to the RNA analysis, clustering was performed with the use of the Louvain algorithm with a resolution of 0.6. This procedure led to the identification of 8 clusters (Fig. 5A).

Afterwards, integration between the ATAC dataset and the previously analyzed RNA dataset was performed. Our goal was to achieve "label transferring" between the annotated RNA clusters and the new groups that emerged after the ATAC clustering analysis. The integration process enabled the labeling of scATAC-seq cells according to the 9 SF subpopulations occurring in RNA analysis (Fig. 5A) (differences in the software versions of Seurat and ArchR employed in SCALA, compared to the ones used in the publication containing the initial analysis of the dataset didn't let us reproduce our UMAP visualization in the exact same manner).

Following integration analysis, semi-supervised trajectory inference with ArchR (Fig. 5B), confirmed the existence of a pathogenic branch, consisting of S2a, S2d, S4b, and S4a clusters, in accordance with scRNA findings.

By utilizing the gene activity scores (calculated as described above), the Wilcoxon test was employed (Fig. 5C) in order to detect top marker features per cluster ($|$Log2FC$= \geq 0.58$ and FDR $\leq 0.05$ cut-offs were applied). Consequently, a robust merged peak set was identified across SF clusters, using MACS2 [57] by generating two pseudo-bulk replicates. Iterative overlap peak merging [58] was applied at the level of the pseudo-bulk replicates and across SF subpopulations, forming a single merged peak set of 158,713 regions with a fixed length of 500 bps.

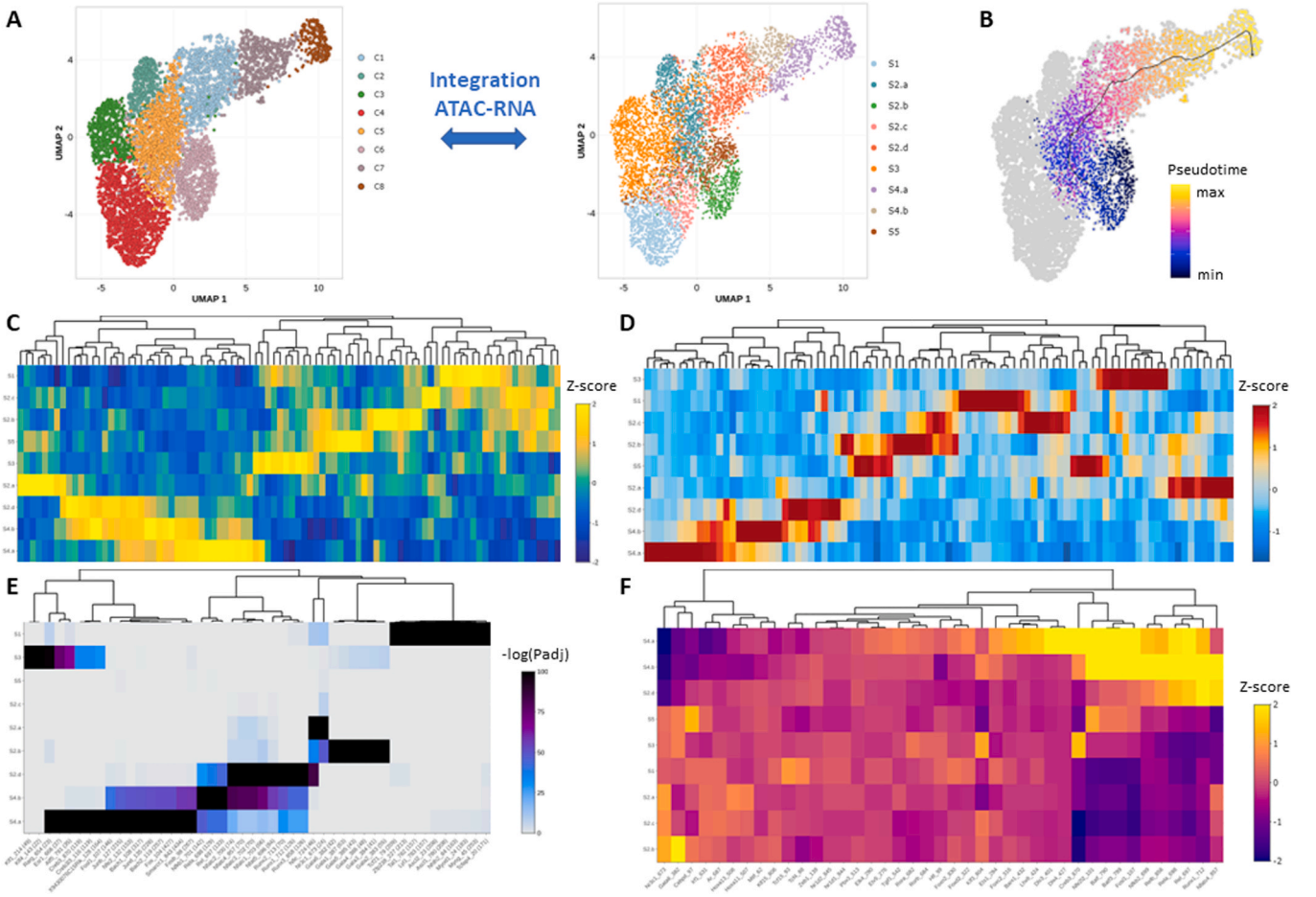


**Fig. 5.** Use case - hTNFtg scATAC-seq data analysis. (A) Integration between scRNA-seq and scATAC-seq datasets. Cluster labels from RNA analysis are transferred to ATAC. Cells are projected in UMAP space and colored according to clustering (left) or transferred labels (right). (B) Semi-supervised trajectory analysis in the ATAC dataset recapitulates the outcome of the respective analysis in RNA data. S2b was used as an initial state and S4a as a final state. (C) z-scores of gene activity values for the top-10 marker features of each cluster (after integration) are displayed in a heatmap.(D) Heatmap displaying z-scores for the accessibility of top-10 marker peaks for each cluster (after integration). (E) Motif enrichment analysis in marker peaks of each cluster. Enriched motifs of each cluster are displayed in a heatmap. Color scale denotes the significance of enrichment. (F) Gene regulation analysis identifies positive regulators for each cluster. Top regulators are displayed in a heatmap. Color scale depicts motif deviations z-scores. In panels (C-F) marker genes/peaks, enriched motifs and positive regulators are shown in x-axis, while clusters (after integration) are shown in y-axis.

Subsequently, differential accessibility analysis between cells was performed to identify cluster-specific marker peaks (|Log2FC= $\geq$ 0.58 and FDR $\leq$ 0.1 cut-offs were applied) (Fig. 5D). Consequently, marker peaks were utilized to perform motif enrichment analysis, using the CIS-BP database (|Log2FC= $\geq$ 0.58 and FDR $\leq$ 0.05 cut-offs were applied) (Fig. 5E). The three previous modes of analysis pinpointed the existence of distinct patterns of gene activity and peak/motif accessibility along clusters. Furthermore, hierarchical clustering on z-scores strengthened the categorization of clusters into three main groups namely sublining, intermediate, and lining.

Gene regulatory analysis was conducted in the ATAC assay as well. More specifically peak to gene linkages were detected using correlation analysis between enhancer peak accessibility and integrated gene expression. Moreover, TF motif accessibility was correlated with integrated TF gene expression in a cell-by-cell manner, reporting TFs with Pearson R2 $\geq$ 0.5 and p-adjusted value $\leq$ 0.05, identifying 41 "positive regulators" (Fig. 5F).

In conclusion, analysis of RNA and ATAC data of SFs in healthy and *hTNFtg* mice at single-cell resolution led to the identification of 9 subpopulations with distinct functions. Inspection of marker genes enriched functional terms, marker peaks, enriched motifs, and regulatory networks from both analyses supported the categorization of the identified clusters in 3 broad groups, namely sublining, intermediate, and lining. In the sublining group, clusters showcase gene expression and accessibility patterns related to homeostatic properties, while clusters belonging to intermediate and lining groups differ in many aspects from the previously described category, and exhibit properties related to proliferation, inflammation, and destruction of the joint. With the current use case, the complementarity of the two assays was showcased. To this end, results from the ATAC analysis were utilized in order to corroborate findings from the RNA such as cluster-specific TFs, marker genes, and trajectories. It is worth mentioning that the analysis of the use case datasets was able to reproduce the original findings of the publication and at the same time offered some alternative options for some steps of the analysis, such as functional enrichment analysis or GRNs, during these steps different tools were employed in SCALA. Finally, some concepts regarding comparisons between the three different conditions or sub-clustering of the lining population of SFs were dropped in favor of simplicity.

## 4. Discussion

SCALA is a comprehensive bioinformatics pipeline offered both as a web-server (limited capacity edition) and a stand-alone application. It performs end-to-end sc analysis, by using the current best practices of the field. It currently enables the analysis of scRNA-seq and scATAC-seq

datasets (which comprise the vast majority of the available sc data in the literature), facilitating both independent and integrative analysis of the two modalities. SCALA was employed to characterize transcriptomics and epigenetic profiles of mouse SFs in healthy and arthritic states. The different modes of analysis stratified, revealed fibroblast populations with distinct patterns of expression, functional characteristics, and regulatory networks. For comparison, we report a catalog of similar tools (e.g., pagoda2, SingleCAnalyzer [59], Bingle-seq [60], iCellR [31], cerebro [30], Is-CellR[61], SeuratWizard [33], ICARUS [62], SC1[63], alona [64], WASP [65], CHIPSTER [66], Asc-Seurat [67], GenePattern [68], PIVOT [69] and we highlight their pros and cons along with their complementarity to SCALA (Supplementary Table 1). To this end, it is worth mentioning that to our knowledge, only icellR [31] offers scATAC seq analysis whereas only six applications are available as web server applications. In addition, SCALA is among the few tools that offer L-R and GRN analysis modes. Furthermore, improvements in the user interface, such as a comparison mode enabling side-to-side visualization of dimensionality reduction plots or feature expression plots, when more than one condition is present are to be expected in the near future, alongside a mode for scRNA-seq dataset integration based on RPCA. Readers who are interested in execution times and RAM consumption for different sizes of input datasets are prompted to visit Supp. Table 2, in which a benchmarking with 8 single cell datasets is performed. As indicated in the table, users are suggested to prefer the desktop version of SCALA for larger datasets ($>$ 50,000 cells), as some of the steps cannot be completed in the online version.

## 5. Conclusions

SCALA offers a complete bioinformatics pipeline for handling scRNA and scATAC datasets, accompanied by a user friendly interface, which makes it accessible to a broad audience of biomedical community. Using state-of-the-art analysis modules and visualization, SCALA aids researchers to decipher complex biological mechanisms in an easy and convenient way. It accepts both raw and pre-processed datasets, thus giving the users the flexibility to perform their analysis from-scratch, or to visualize and reanalyze already processed data. Conclusively, we expect that due to its simplicity and its pipeline integration, SCALA will become a reference application for biomedical scientists who wish to analyze and explore their data in an interactive way.

## Funding

This project has been funded by Horizon Europe Advanced ERC project BecomingCausal (ERC-2021-ADG, ID# 101055093) to GK and by projects Single.Out (HFRI-FM17C3-3780) and BOLOGNA (HFRIFM17-1855) to G.K. and G.P. respectively, funded by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "1st Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment". We acknowledge support of this work by project pMedGR (MIS 5002802), funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014–2020) and co-financed by Greece and the European Union (European Regional Development Fund). F.A.B was also supported by The Fondation Sante and the Onassis Foundation.

## CRediT authorship contribution statement

Christos Tzaferis and Dimitris Konstantopoulos wrote most of the code. Evangelos Karatzas organized the GUI and worked on data visualization, Fotis A. Baltoumas implemented the functional enrichment analysis and worked on the back-end server and pipeline setup. Dimitris Konstantopoulos implemented the pipeline related to scATAC-seq analysis. Christos Tzaferis implemented the pipeline related to scRNA-seq analysis. Christos Tzaferis, Dimitris Konstantopoulos and George Kollias provided the datasets for the case studies. Dimitris Konstantopoulos, George Kollias and Georgios A. Pavlopoulos conceived the idea and supervised the project. All of the authors wrote parts of the manuscript.

## Declaration of Competing Interest

The authors declare that they have no competing interests.

## Acknowledgements

We thank the colleagues from Armaka, Fousteri and Kollias lab for the generation, interpretation and permission to include the single-cell data, which are presented in the section of use cases. We also thank Matthieu Lavigne for the valuable discussions and the thoughtful suggestions.

### *Authors' contributions*

CT and DK wrote most of the code. EK organized the GUI and worked on data visualization, FB implemented the functional enrichment analysis and worked on the back-end server and pipeline setup. DK implemented the pipeline related to scATAC-seq analysis. CT implemented the pipeline related to scRNA-seq analysis. CT, DK and GK provided the datasets for the case studies. DK, GK and GAP conceived the idea and supervised the project. All of the authors wrote parts of the manuscript.

### *Code availability*

**Project name**: SCALA.
**Project home page**: https://scala.fleming.gr, https://github.com/PavlopoulosLab/SCALA.
DOCKER: https://hub.docker.com/r/pavlopouloslab/scala.
**Operating system(s)**: Linux, Windows, MAC.
**Programming language**: R Shiny, JavaScript.
**Other requirements**: for the stand-alone version please visit the github page and check the section "Requirements".
**License**: GPL-3.0 license.
**Any restrictions to use by non-academics**: Not applicable.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.10.032.

## References

[1] Slovin S, Carissimo A, Panariello F, Grimaldi A, Bouché V, Gambardella G, et al. Single-Cell RNA sequencing analysis: a step-by-step overview. Methods Mol Biol 2021;2284:343–65 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/33835452/⟩.
[2] Li L, Xiong F, Wang Y, Zhang S, Gong Z, Li X, et al. What are the applications of single-cell RNA sequencing in cancer research: a systematic review. J Exp Clin Cancer Res 2021;40(1) [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/33975628/⟩.
[3] Andrews TS, Kiselev VY, McCarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat Protoc 2021;16(1). cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/33288955/⟩.
[4] Huang W, Wang D, Yao YF. Understanding the pathogenesis of infectious diseases by single-cell RNA sequencing. Micro Cell 2021;8(9):208–22 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/34527720/⟩.
[5] Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol 2019;15(6) [cited 2021 Sep 17];.
[6] Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 2009;6(5):377–82 [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/19349980/⟩.
[7] Xin Y, Kim J, Ni M, Wei Y, Okamoto H, Lee J, et al. Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. Proc Natl Acad Sci 2016;113(12):3293–8 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/26951663/⟩.
[8] Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods

2013;10(11):1096–100 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/24056875/⟩.
[9] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015;161(5):1202–14 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/26000488/⟩.
[10] Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/28091601/⟩.
[11] Moses L, Pachter L. Museum of spatial transcriptomics. Nat Methods 2022;19(5):534–46 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/35273392/⟩.
[12] Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell Atlas. Elife 2017;6 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/29206104/⟩.
[13] Schaum N, Karkanias J, Neff NF, May AP, Quake SR, Wyss-Coray T, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature 2018;562(7727):367–72 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/30283141/⟩.
[14] Chen G, Ning B, Shi T. Single-cell RNA-Seq technologies and related computational data analysis. Front Genet 2019;10(APR) [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/31024627/⟩.
[15] Janssens J, Aibar S, Taskiran II, Ismail JN, Gomez AE, Aughey G, et al. Decoding gene regulation in the fly brain. Nature 2022;601(7894):630–6 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/34987221/⟩.
[16] Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, et al. A single-cell atlas of chromatin accessibility in the human genome. Cell 2021;184(24):5985–6001.e19, ⟨https://pubmed.ncbi.nlm.nih.gov/34774128/⟩ [cited 2023 Apr 3].
[17] Zappia L, Phipson B, Oshlack A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput Biol 2018;14(6) [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/29939984/⟩.
[18] Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184(13):3573–3587.e29 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/34062119/⟩.
[19] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19(1) [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/29409532/⟩.
[20] Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32(4):381–6 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/24658644/⟩.
[21] Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 2017;14(10):979–82 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/28825705/⟩.
[22] Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, et al. Cicero predicts cis-Regulatory DNA interactions from single-cell chromatin accessibility data. Mol Cell 2018;71(5):858–871.e8 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/30078726/⟩.
[23] Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. Nat Methods 2021;18(11):1333–41 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/34725479/⟩.
[24] Danese A, Richter ML, Chaichoompu K, Fischer DS, Theis FJ, Colomé-Tatché M. EpiScanpy: integrated single-cell epigenomic analysis. Nat Commun 2021;12(1) [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/34471111/⟩.
[25] Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods 2017;14(11):1083–6.
[26] Bravo González-Blas C, Minnoye L, Papasokrati D, Aibar S, Hulselmans G, Christiaens V, et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nat Methods 2019;16(5):397–400 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/30962623/⟩.
[27] Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat Genet 2021;53(3):403–11 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/33633365/⟩.
[28] Davie, Janssens K, Koldere J, De Waegeneer D, Pech M, Kreft U, et al. A single-cell transcriptome atlas of the aging drosophila brain. Cell 2018;174(4):982–998.e20 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/29909982/⟩.
[29] Chan Zuckerberg Initiative. CZ CELLxGENE Discover. [Internet]. Available from: ⟨https://cellxgene.cziscience.com/⟩.
[30] Hillje R, Pelicci PG, Luzi L. Cerebro: interactive visualization of scRNA-seq data. Bioinformatics 2020;36(7):2311–3 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/31764967/⟩.
[31] Tang KH, Li S, Khodadadi-Jamayran A, Jen J, Han H, Guidry K, et al. Combined inhibition of SHP2 and CXCR1/2 promotes antitumor T-cell response in NSCLC. Cancer Discov 2022;12(1):47–61 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/34353854/⟩.
[32] No Title [Internet]. Available from: ⟨https://www.partek.com/single-cell-gene-expression/⟩.
[33] Yousif A, Drou N, Rowe J, Khalfan M, Gunsalus KC, Gunsalus KC. NASQAR: a web-based platform for high-throughput sequencing data analysis and visualization. BMC Bioinforma 2020;21(1) [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/32600310/⟩.
[34] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019;47(W1):W191–8.
[35] Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genom 2018;19(1):19. 477.
[36] Badia-I-Mompel P, Vélez Santiago J, Braunger J, Geiss C, Dimitrov D, Müller-Dott S, et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. Bioinforma Adv 2022;2(1) [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/36699385/⟩.
[37] Müller-Dott S, Tsirvouli E, Vázquez M, Flores ROR, Badia-i-Mompel P, Fallegger R, et al. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. bioRxiv 2023. 2023.03.30.534849. Available from: ⟨http://biorxiv.org/content/early/2023/04/01/2023.03.30.534849.abstract⟩.
[38] Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. Nat Methods 2020;17(2):159–62.
[39] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008;2008(10):P10008 [cited 2023 Apr 3], ⟨https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008⟩.
[40] Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 2018;37(1):38–47 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/30531897/⟩.
[41] Maaten LJP, van der, Hinton GE. Visualizing high-dimensional data using t-SNE. J Mach Learn Res 2008;9(nov):2579–605 [cited 2023 Apr 3], ⟨https://research.tilburguniversity.edu/en/publications/visualizing-high-dimensional-data-using-t-sne⟩.
[42] Haghverdi L, Buettner F, Theis FJ. Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics 2015;31(18):2989–98 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/26002886/⟩.
[43] Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, et al. Author correction: visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol 2020;38(1):108 [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/31896828/⟩.
[44] McDavid A, Finak G, Chattopadyay PK, Dominguez M, Lamoreaux L, Ma SS, et al. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics 2013;29(4):461–7 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/23267174/⟩.
[45] Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 2015;16(1) [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/26653891/⟩.
[46] McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst [Internet] 2019 24;8(4):329–37. e4, ⟨http://www.ncbi.nlm.nih.gov/pubmed/30954475⟩.
[47] Thanati F, Karatzas E, Baltoumas FA, Stravopodis DJ, Eliopoulos AG, Pavlopoulos GA. FLAME: a web tool for functional and literature enrichment analysis of multiple gene lists. Biology 2021;10(7) [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/34356520/⟩.
[48] Ekiz HA, Conley CJ, Stephens WZ, O'Connell RM. CIPR: a web-based R/shiny app and R package to annotate cell clusters in single cell RNA sequencing experiments. BMC Bioinforma 2020;21(1):191.
[49] Armaka M, Konstantopoulos D, Tzaferis C, Lavigne MD, Sakkou M, Liakos A, et al. Single-cell chromatin and transcriptome dynamics of Synovial Fibroblasts transitioning from homeostasis to pathology in modelled TNF-driven arthritis. bioRxiv 2021 [cited 2021 Sep 20];2021.08.27.457747. Available from, ⟨https://www.biorxiv.org/content/10.1101/2021.08.27.457747v1⟩.
[50] Keffer J, Probert L, Cazlaris H, Georgopoulos S, Kaslaris E, Kioussis D, et al. Transgenic mice expressing human tumour necrosis factor: a predictive genetic model of arthritis. EMBO J 1991;10(13):4025–31 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/1721867/⟩.
[51] Danks L, Komatsu N, Guerrini MM, Sawa S, Armaka M, Kollias G, et al. RANKL expressed on synovial fibroblasts is primarily responsible for bone erosions during joint inflammation. Ann Rheum Dis 2016;75(6):1187–95 [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/26025971/⟩.
[52] Wei K, Korsunsky I, Marshall JL, Gao A, Watts GFM, Major T, et al. Notch signalling drives synovial fibroblast identity and arthritis pathology. Nature 2020;582(7811):259–64 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/32499639/⟩.
[53] Buechler MB, Pradhan RN, Krishnamurty AT, Cox C, Calviello AK, Wang AW, et al. Cross-tissue organization of the fibroblast lineage. Nature 2021;593(7860):575–9 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/33981032/⟩.
[54] Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol 2020;38(12):1408–14 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/32747759/⟩.
[55] Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, et al. CellRank for directed single-cell fate mapping. Nat Methods 2022;19(2):159–70 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/35027767/⟩.
[56] van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering gene interactions from single-cell data using data diffusion. Cell 2018;174(3):716–729.e27 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/29961576/⟩.
[57] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9(9) [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/18798982/⟩.
[58] Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science 2018;362(6413) [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/30361341/⟩.
[59] Prieto C, Barrios D, Villaverde A. SingleCAnalyzer: interactive analysis of single Cell RNA-Seq data on the cloud. Front Bioinforma 2022;2 [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/36304292/⟩.

[60] Dimitrov D, Gu Q. BingleSeq: a user-friendly R package for bulk and single-cell RNA-Seq data analysis. PeerJ 2020;8 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/33391870/⟩.
[61] Patel MV. iS-CellR: a user-friendly tool for analyzing and visualizing single-cell RNA sequencing data. Bioinformatics 2018;34(24):4305–6 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/29982379/⟩.
[62] Jiang A, Lehnert K, You L, Snell RG. ICARUS, an interactive web server for single cell RNA-seq analysis. Nucleic Acids Res 2022;50(W1):W427–33 [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/35536286/⟩.
[63] Moussa M, Mandoiu II. SC1: A Tool for Interactive Web-Based Single-Cell RNA-Seq Data Analysis. J Comput Biol 2021;28(8):820–41 [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/34115950/⟩.
[64] Franzén O, Björkegren JLM. alona: a web server for single-cell RNA-seq analysis. Bioinformatics 2020;36(12):3910–2 [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/32324845/⟩.
[65] Hoek A, Maibach K, Özmen E, Vazquez-Armendariz AI, Mengel JP, Hain T, et al. WASP: a versatile, web-accessible single cell RNA-Seq processing platform. BMC Genom 2021;22(1) [cited 2023 Apr 3], ⟨https://pubmed.ncbi.nlm.nih.gov/33736596/⟩.
[66] Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, et al. Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genom 2011;12 [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/21999641/⟩.
[67] Pereira WJ, Almeida FM, Conde D, Balmant KM, Triozzi PM, Schmidt HW, et al. Asc-Seurat: analytical single-cell Seurat-based web application. BMC Bioinforma 2021;22(1) [cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/34794383/⟩.
[68] Mah CK, Wenzel AT, Juarez EF, Tabor T, Reich MM, Mesirov JP. An accessible, interactive genepattern notebook for analysis and exploration of single-cell transcriptomic data. F1000Research 2019 [cited 2023 Apr 3];7. Available from: ⟨https://pubmed.ncbi.nlm.nih.gov/31316748/⟩.
[69] Zhu Q, Fisher SA, Dueck H, Middleton S, Khaladkar M, Kim J. PIVOT: platform for interactive analysis and visualization of transcriptomics data. BMC Bioinforma 2018;19(1). cited 2023 Apr 3];, ⟨https://pubmed.ncbi.nlm.nih.gov/29304726/⟩.