

Oct 16, 2024

Single-cell preparation and scRNA-Seq data analysis

DOI

<https://dx.doi.org/10.17504/protocols.io.eq2lyw9qwvx9/v1>

Khalil Kass Youssef¹, Angela Nieto¹, NITIN NARWADE²

¹Instituto de Neurociencias de Alicante (CSIC-UMH); ²institute of neuroscience alicante



Khalil Kass Youssef

Instituto de Neurociencias de Alicante (CSIC-UMH)

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.eq2lyw9qwvx9/v1>

Protocol Citation: Khalil Kass Youssef, Angela Nieto, NITIN NARWADE 2024. Single-cell preparation and scRNA-Seq data analysis. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.eq2lyw9qwvx9/v1>

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: July 25, 2024



Last Modified: October 16, 2024

Protocol Integer ID: 104047

Keywords: cell rnaseq library, seq data analysis the epithelial, transcriptome, models of renal fibrosi, molecular programmes in fibrosi, scrna, renal fibrosi, cell plasticity in embryonic development, mesenchymal transition, epithelial, breast cancer

Disclaimer

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

Abstract

The Epithelial to Mesenchymal transition (EMT) triggers cell plasticity in embryonic development, adult injured tissues, and cancer. Here, we provide a step-by-step protocol to prepare single-cell RNAseq libraries from mouse models of renal fibrosis and breast cancer. We further provide a detailed workflow for different analyses used to characterise the EMT programme as assessed from single-cell transcriptomes, to reconstruct EMT trajectories, and to predict the EMT-associated molecular programmes in fibrosis and cancer.

1

A- Single-cell preparation

12 weeks old mouse males were subjected to UUO or sham surgery (CTR) and whole kidney harvested after 10 days. Mammary gland carcinomas were collected from 12 to 14 weeks old female mice. Harvest tissue were first minced manually using sterile scalpels and finely cut with a McIlwain Tissue Chopper (TED PELLA, INC). Minced samples were incubated in 2,5 ml of digestion buffer (PBS supplemented with 2.5 Wünsch units of TH Liberase/ml and 25µg/ml DNase I (Roche)) at 37°C for 45 min in an Incubator Microplate Shaker (VWR). To achieve tissue dissociation in single-cells suspensions, tumour fragments were pipetted up and down every 5 min and cellular disaggregation was evaluated under the microscope.

Disaggregated samples were neutralized with 15ml of breast medium supplemented with 25µg/ml DNase I and subsequently passed through a 70µm and 40µm filters (BD Falcon) for breast tumours, and with the medium used for MDCK cells for kidney samples (see cell culture section). Cells were spun down, and pellets resuspended in 5ml of red blood lysis ACK (Ammonium-Chloride-Potassium, BioLegend Cat#420301) buffer for 5min at RT with continuous gentle rotation. After another run of spinning, pellets were resuspended in 5ml FACS buffer/5mM EDTA and passed through a 40µm filter. 20 µl of cell preparations were mixed with an equal volume of Trypan Blue and incubated for 5min to evaluate cell number and quality control for single-cell preparation, cell viability and low debris content. Cells were further spun down (3 min at 400 G at RT) and resuspended in Dead cell removal beads (MiltenyiBiotec, Cat#130-090-101) in 100µl per 107 cells and incubated for 15min at RT with gentle mixing. During the incubation, MACS LS columns (MiltenyiBiotec) were placed on a magnet rack and conditioned by rinsing with binding buffer. After spinning for 3min at 400 G at 4°C, cells were resuspended in 3ml binding buffer and applied on a magnetic column to discard dead cells. Living cells were collected in 50ml Falcon tubes and placed on ice. Columns were rinsed 4x with 3 ml binding buffer to further collect living cells. Collected cells were spun down (5 min at 400 G and 4°C), resuspended in cold breast medium and kept on ice. 20µl of cell suspensions were mixed with 20µl Trypan Blue, incubated for 5 min to assess single-cell state. Viability can be evaluated by FACS by measuring the DAPI negative fraction. Finally, cells were adjusted to 1000 cells/µl with cold FACS buffer and directly used for GEM (Gel Bead-In-Emulsions) preparation.

B- Single-cell GEM and cDNA library preparation

Kidney samples: Three single-cell preparations were obtained from 1 SHAM and 2 UUO kidneys from 3 male mice as shown in Extended Data Fig. 4a. Individual cell

encapsulation for single-cell expression profiling was performed using 10x Chromium Controller (10x Genomics). Three libraries were generated from different samples (SHAM, UUO#1 and UUO#2). For every sample, the single-cell suspension was loaded into a Chromium Next GEM Chip G (10x Genomics) and processed following the manufacturer's instructions. Single-cell RNA-seq libraries were prepared using the Chromium Next GEM Single Cell 3' Library & Gel Bead kit v3.1 and samples were indexed using Single Index Kit T Set A (10x Genomics). Pooled libraries were then loaded on a HiSeq2500 instrument (Illumina) and sequenced to obtain 75 bp paired-end reads.

Tumour samples: Four single-cell preparations were obtained from four independent mammary gland carcinoma samples dissected from 3 female mice as shown in Extended Data Fig. 7a. Individual cell encapsulation for single-cell expression profiling was performed using 10x Chromium Controller (10x Genomics). Four libraries were generated from different samples (T1-T4). For every sample, the single-cell suspension was loaded into a Chromium Single Cell A Chip (10x Genomics) and processed following the manufacturer's instructions. Single-cell RNA-seq libraries were prepared using the Chromium Single Cell 3' Library & Gel Bead kit v2 and samples were multiplexed using Chromium i7 Multiplex kit (10x Genomics). Pooled libraries were then loaded on a HiSeq2500 instrument (Illumina) and sequenced to obtain 75 bp paired-end reads.

C- scRNA-Seq data analysis

General workflow for kidney and cancer single-cell data analysis is summarized in Supplementary table 6.

C-1- Kidney single-cell data analysis

Libraries were sequenced to obtain around 900 million reads in total (SHAM:304.07M; UUO#1: 299.8M; UUO#2: 296.76M). Quality control of sequenced reads was performed using FastQC v0.11.9 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc>) (Babraham Institute). The reads were aligned to the mouse genome assembly GRCm38 (mm10; ensembl reference annotation) and quantified using 10X Genomics Cell Ranger(Zheng et al., 2017) v5.0.0 pipeline with default parameters.

Cell quality control, filtering, and integration process: The Cell Ranger output was imported in the R v.3.6.1 (<https://www.r-project.org/>) statistical environment and converted to a Seurat(Stuart et al., 2019) (v.3.2.2) object using CreateSeuratObject function. Barcodes with total unique molecular identifier (UMI) count >10% of the 99th percentile of the expected recovered cells were selected for further analysis. In

all three libraries, mean read pairs per cell was above 19000 (SHAM: 19289; UWO#1: 22172; UWO#2: 20159). Confident mapping to exonic regions was above 50% for each library; median unique counts per cell were as follows: SHAM: 3750; UWO#1: 3124; UWO#2: 2616; and median value for detected genes per cell was SHAM: 1550; UWO#1: 1734; UWO#2 1431. The sample-specific putative doublets were predicted using Scrublet(Wolock et al., 2019) v0.2.3 and filtered out from the subsequent analysis. The expected doublet rate threshold (SHAM = 0.12, UWO#1 = 0.10, UWO#1 = 0.11) was calculated based on the total recovered cells and the doublet detection score was set to SHAM = 0.35, UWO#1 = 0.25, UWO#1 = 0.27 by manually inspecting the bimodal distribution. Next, we filtered out low-quality cells expressing high levels of mitochondrial markers (>10%) (Extended Data Fig. 4b). Finally, high-quality cells with gene detection in the range of 400-4000 were retained for downstream analysis. All subsequent analyses were performed on 25424 cells passing quality control. UMI count data from each sample was normalized following a regularized negative binomial regression using Seurat function SCTransform with default parameters. The libraries were then integrated using top 3000 high variable genes (HVGs). First, we selected features for anchors identification based on their redundant detection across samples (SelectIntegrationFeatures) and then the PrepSCTIntegration was used to ensure that all necessary Pearson residuals has been calculated. Next, we identified integration anchors by performing Canonical Correlation Analysis with FindIntegrationAnchors (default parameters and "SCT" as normalization method) and datasets were integrated based on the precomputed anchorset using IntegrateData function.

Dimensionality reduction and cluster detection: Principal Component Analysis (PCA) dimensionality reduction was performed using the RunPCA function with default parameters implemented in Seurat. A Shared Nearest Neighbor (SNN) Graph for the integrated dataset was built using FindNeighbors over top 25 Principal Component (PCs), and cell clusters were identified by a SNN modularity optimization-based clustering algorithm using FindClusters function (resolution: 0.65). Nonlinear dimension reduction with Uniform Manifold Approximation and Projection (UMAP) was performed over top 25 PCs using the RunUMAP function.

Differential gene expression testing and clusters annotation: For expression plots and differentially expressed gene testing, gene counts were Log-Normalized using NormalizeData function. Differential expression analysis was performed using logistic regression (LR) without a latent variable implemented in FindAllMarkers function. To generate the cluster specific single-cell level gene expression profile, we first sorted the differentially expressed genes by average log2 fold-change in descending order and p-adjusted values in ascending order. Then the top 20 genes from each cluster were selected and their scaled Log-Normalized expression was represented as heatmap using DoHeatmap function (Extended Data Fig. 4c). Cell types were assigned to clusters using bona fide gene markers and grouped into 6 major cell

populations (Fig. 3b and Extended Data Fig. 4d). The expression of cell type specific selected markers was represented as a dot plot using DotPlot function implemented in Seurat (Extended Data Fig. 4e).

Compositional Analysis for Kidney cell populations: To investigate the compositional changes of cell populations in SHAM and UUO we used a statistical method implemented in Cacoa R package (Petukhov et al., 2022). We run Cacoa using "runCoda" function with 1000 bootstraps. Glomerulus cell population was set as a reference cell type. Compositional data analysis (CoDA) score was represented as a boxplot along with associated P-value (Extended Data Fig. 4f).

Classification of injured-epithelial cells using supervised machine learning: Although the proximal tubules are considered the major contributor to damaged epithelial cells and drivers of tubulointerstitial fibrosis in response to injury (Chevalier, 2016), we did not in principle exclude the contribution of the other epithelial populations. Based on the superior performance of deep learning models for assigning cell types in scRNA-Seq data (Ma and Xu, 2022), we followed a similar approach to predict how each epithelial cell type could contribute to injury. Considering the hypothetical contribution of any renal epithelial cell type (9 clusters) we treated this as a multi-class classification problem that we approached using multi-layer perceptron (MLP) deep learning neural network (sklearn v.1.1.1 python library, <https://github.com/scikit-learn/scikit-learn>). To build the MLP training model, we subset the cells belonging to the epithelial clusters (8, 5, 18, 16, 15, 2, 21, 11, 23; Extended Data Fig. 4d). Next, the SCT normalised gene expression matrix for top 3000 HVGs was used as feature set (see section "Cell quality control, filtering, and integration process"). Then the MLPClassifier was used with default parameters to build a training model that we evaluated using 10-fold cross-validation (CV). Finally, the model performance in CV was evaluated using two performance measures: accuracy and Matthews Correlation Coefficient (MCC) (Extended Data Fig. 5a). To calculate the performance measures, we built a confusion matrix using `multilabel_confusion_matrix` function (sklearn package) using One vs. Rest (OvR) strategy.

		Real Class	
Predicted Class		Positive	<u>Negative</u>
	Positive	TP (<u>True Positive</u>)	FP (False Positive)
	<u>Negative</u>	FN (<u>False Negative</u>)	TN (<u>True Negative</u>)

and calculated accuracy and MCC using the following equations

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (2)$$

After validation, we rebuilt the MLP model over all cells from epithelial clusters and the resulting model was used to classify injured cells into different epithelial clusters, thus predicting the most likely closest origin of injured epithelial cells (Fig. 2d).

Proximal tubule and injured cells subset analysis: Based on the classification performed using supervised machine learning approach described above, we subset PT clusters 8, 16 and 5, each contributing to more than 10% to the injured cell population, together with the associated injured epithelial cells (cluster 0 cells originated from clusters 8, 16 and 5) (Fig. 3d). We re-computed the PCs using RunPCA functions (Seurat) with default parameters using same 3000 HVGs. Top 5 PCs were used to re-build the SNN Graph (FindNeighbors; default parameters) and UMAP. FindClusters function with a resolution of 0.1 was used for clustering, which resulted in 3 clusters (Fig. 3e).

EMT, Differentiation and inflammation score: We downloaded Hallmark_EMT gene set from MSigDB (UC San Diego and Broad Institute) and used AddModuleScore function over Log-Normalized gene expression to calculate the EMT score. To calculate the

differentiation score we used common up-regulated genes (adjusted P-value < 0.05 and average log2FC > 0.25) in PT segments S1, S2 and S3 (Ransick et al., 2019). The genes belonging to the kidney inflammatory pathways reported by Wu et. al. (Wu et al., 2020) were used to calculate the injury/inflammation score. Calculated scores were visualized over UMAP (Fig. 3f).

Trajectories inference using PAGA and RNA-Velocity:

PAGA Analysis: integrated Seurat object for proximal tubule and injured cells (see section Proximal tubule and injured cells subset analysis) was exported as a loom file using `as.loom` function and brought into python environment. We used precomputed PCs and cell embedding obtained from Seurat analysis. The neighbourhood graph was constructed with `neighbours=15` and `n_pcs=5` using `sc.pp.neighbors` function implemented in Scanpy (Wolf et al., 2018) v1.6.0 python package. The partition-based graph abstraction (PAGA) algorithm implemented in Scanpy was used to build the connectivity map (Fig. 3g upper panel).

RNA Velocity Analysis: Run10x command line utility from `velocyto` (La Manno et al., 2018) v0.17.17 package was used to calculate the sample-specific spliced and unspliced count matrices. Obtained individual loom files were merged using `combine` function provided by `loompy` package v3.0.6 (<https://linnarssonlab.org/loompy/>). To infer the directionality of the transcriptional changes in the proximal tubule and injured cells (same used for PAGA), first we performed initial gene filtering using `"score_detection_levels"` function with `min_expr_counts=20` and `min_cells_express=15`. Second, the top 3000 HVGs were selected using `score_cv_vs_mean` function. Third, HVGs were filtered based on cluster-wise spliced and unspliced expression threshold (average unspliced expression > 0.01 and average spliced expression > 0.08 in at least one of the clusters). Next, spliced and unspliced counts of filtered genes (n= 477) were normalized for the total number of cells using `_normalize_S` and `_normalize_U` functions, respectively, and used to perform PCA. The top PCs were selected based on differences between the cumulative sum of explained variance ratio (>0.002), which resulted in a total of 147 PCs. Next, we performed data imputation on 147 PCs with 500 neighbours using `"knn_imputation"` function. The RNA velocity was estimated by assuming steady-state transition with `delta=1.0`. The estimated velocity is then embedded on the regularized grid (Gaussian kernel). The step size was set to 40 and smoothness to 0.3 to visualize velocity on the top of UMAP embedding (Fig. 3g lower panel).

Pseudotime analysis for the inferred trajectory: scanpy object generated for PAGA analysis was used for the pseudotime analysis. We set a root cell (ATTCTTGAGTGCAAAT-1_2) from cluster 2 which has maximum expression of the proximal tubule marker `Slc22a12`. Then, the diffusion map (`sc.tl.diffmap`) embedding

was calculated using top 5 PCs and pseudotime values were calculated using top 5 diffusion components (sc.tl.dpt function from Scanpy).

SCENIC Analysis for regulon prediction: To understand the regulatory activity of transcription factors in our EMT trajectory we predicted the scRNA-Seq expression-based regulon using pySCENIC(Aibar et al., 2017) v0.11.0. First we exported the Seurat integrated object (see section Proximal tubule and injured cells subset analysis) with 3000 HVGs to a python compatible loom file using “as.loom” function. We constructed the gene regulatory network (GRN) between Transcription Factors (TFs) and their target genes using default parameters. Predicted GRN and scenic motif database (500bp upstream and 100bp downstream region to the TSS of target genes) for GRCm38 (mm10) was used to infer the regulatory modules and filtered using default parameters. The single-cell regulon activity was calculated using AUCell algorithm implemented in SCENIC package and average AUC score was represented as heatmap (Fig. 3h and Extended Data Fig. 6a). Additionally, single cells AUC scores were plotted over precomputed pseudotime and the local regression curve was fitted using generalized additive model implemented in mgcv R package (<https://cran.r-project.org/package=mgcv>) with splines of degree=5 (Extended Data Fig. 6b). The enriched motifs for each TF were reported based on their highest NES (Supplementary Table 7).

Trajectory-based differential expression analysis: Integrated Seurat object for proximal tubule and injured trajectory was transferred to Monocle3(Cao et al., 2019) v.0.2.2 object. The PCA and UMAP embeddings were parsed from the Seurat object along with 3000 HVGs and assigned to the Monocle3 object. We use reversed graph embedding (Monocle3 learn_graph function) to construct the principal graph from reduced dimensional space for previously inferred trajectory. Moran’s I test was used to predict the differentially expressed genes along the trajectory. Genes with significant difference (Moran’s I test $q < 0.001$) and consistent expression over trajectory (cluster-wise average log2 fold-change) were retained. The heatmap was generated using smoothed expression of the filtered genes over pseudotime. The pathway enrichment analysis for the differentially expressed genes over trajectory was performed using R based API of EnrichR(Chen et al., 2013) v3.0 against MSigDB_Hallmark_2020, KEGG_2021_Human, GO_Biological_Process_2021, GO_Cellular_Component_2021, GO_Molecular_Function_2021 databases. The selected enriched pathways and GO terms were represented as dot plot using ggplot2 v.3.3.3 (<https://cran.r-project.org/package=ggplot2>) R package.

C-2- PyMT tumours single-cell data analysis

Libraries were sequenced to obtain more than 1200 million reads in total (T1:313.0M; T2:313.55M; T3:306.81M; T4:308.46M). Quality control of sequenced reads was performed using FastQC v0.11.9 (Babraham Institute). Sequenced samples were

processed using the Cell Ranger (v.2.2.0) pipeline (10x Genomics) and aligned to the GRCm38 (mm10) mouse reference genome (Ensembl annotation v99). The genome was customized to detect reads aligned to the tdTomato-WPRE sequence from Gt(ROSA)26Sortm9(CAG-tdTomato)Hze (589 exogenous base pairs from WPRE sequence were first verified by genomic PCR sequencing and further added to the reference genome as a scaffold). Barcodes with total unique molecular identifier (UMI) count >10% of the 99th percentile of the expected recovered cells were selected for further analysis. After this filtering step, we retained a total of 36162 cells (T1:8724 cells; T2:9580 cells; T3:8434 cells; T4:9424 cells) for subsequent analyses. In all four libraries, mean read pairs per cell was above 30000 (T1:35878; T2:32730; T3:36377; T4:32731). Confident mapping to exonic regions was higher than 61% for each library; median unique counts per cell were as follows: T1: 1922; T2: 3281; T3: 2854; T4: 2589; and median detected genes per cell: T1: 842; T2: 1321; T3: 1142; T4: 1079.

Cell quality control, filtering, and integration process: Downstream analyses, such as quality control, integration, normalization, shared nearest neighbour graph-based clustering, differential expression analysis and visualization were performed using the R package Seurat (version 3.1.4) within R Statistical Computing Platform (version 3.6.3). The sample-specific putative doublets were predicted and filtered out from the subsequent analysis as described for the Kidney scRNASeq data. The expected doublet rate threshold was calculated based on the total recovered cells (T1= 0.06, T2= 0.07, T3 = 0.06, T4 = 0.07). The doublet detection score was set to T1= 0.22, T2=0.25, T3 = 0.23, T4 = 0.25 by inspecting bimodal distribution. We retained only high-quality cells based on gene detection levels (400-4000) and in which the mitochondrial percentage was less than 5% in 99.99% of cells (Extended Data Fig. 7b). 34607 cells passing quality control were further analysed. UMI count data from each sample was normalized following a regularized negative binomial regression with Seurat function SCTransform with default parameters. To integrate the four libraries, the top 3000 gene features were selected for anchors identification based on their redundant detection across samples (SelectIntegrationFeatures). PrepSCTIntegration was run to ensure that all necessary Pearson residuals had been calculated. Integration anchors were identified by performing Canonical Correlation Analysis with FindIntegrationAnchors (default parameters and "SCT" as normalization method) and finally, datasets were integrated using the pre-computed anchorset using IntegrateData function.

Dimensionality reduction and cluster detection: PCA dimensionality reduction was performed using the Seurat function RunPCA with default parameters. A Shared Nearest Neighbor (SNN) Graph for the integrated dataset was built using FindNeighbors with top 30 PCs, and cell clusters were identified by a shared nearest neighbour (SNN) modularity optimization-based clustering algorithm with FindClusters function. To explore intratumour heterogeneity, a first cell clustering with

very low resolution (0.03) was performed. This analysis led to the identification of 5 major populations (Fig. 4c and Extended data Fig. 7c-e). The largest population was composed of cells positive for tdTomato transgene, consistent with cancer cell identity. The other four populations were negative for tdTomato, indicating their stromal origin. The 5 major cell populations were annotated to cell types using bona fide gene marker genes. UMAP dimensional reduction technique was performed with top 30 PCs using the RunUMAP function.

Differential gene expression testing and clusters: For expression plots and differentially expressed gene testing, cell expression was Log-Normalized using NormalizeData function with default parameters. Differential expression analysis was performed using logistic regression gene testing with samples as latent variables (variables to test) with FindAllMarkers function. P-value adjustment was performed using Bonferroni correction based on the total number of genes in the dataset. For major populations markers heatmap, we took the logistic regression gene testing with samples as latent variables without thresholds in average log fold-change over Log-Normalized expression for the 5 major populations. A total of 3903 enriched and statistically significant genes (positive log fold-change and adj p-value < 0.1) were sorted by descending average log2 fold-change. A proportional downsampling of 12000 cells (T1: 2882 cells; T2: 3197 cells; T3: 2828 cells; T4: 3093 cells) and previous gene lists were used to plot the heatmap, showing scaled Log-Normalized expression.

Cancer cell subset for downstream analysis: To explore cancer cell heterogeneity, a second cell clustering was performed with default resolution (0.8) from SCTransform integration workflow. This analysis led to the identification of 28 clusters, with 19 clusters corresponding to the major cancer cell population. These 19 clusters were subsetted, and clusters with high proportion of ribosomal gene markers (6 clusters) were excluded. For the retained cancer cells subcluster (13 clusters) analysis, we re-run PCA using RunPCA function by setting the default parameters. Top 20 PCs were used to re-build the SNN Graph (FindNeighbors; default parameters) which led to the identification of 17 clusters (resolution = 0.6). UMAP dimensionality reduction was performed over top 20 PCs with default parameters (Fig. 4c, right panel). The Markov Affinity-based Graph Imputation of Cells (MAGIC v.3.0.0)(van Dijk et al., 2018) was applied to impute the expression of EMT-TFs encoding genes with default parameters (t=3) (Extended Data Fig. 8f).

Trajectories inference using PAGA and RNA-Velocity:

PAGA Analysis: To build a PAGA connectivity map for cancer subset we analysed the data as described in the "PAGA Analysis" section for kidney scRNA-seq data with minimum parameter tuning. The neighborhood graph was constructed over top 20

PCs using 15 neighbours and then the coarse-grained connectivity map (resolution=0.3) was build using `sc.tl.paga` function (Fig. 5a).

RNA Velocity Analysis: The spliced and unspliced count matrices were calculated for cancer subset as described above in the section “RNA Velocity analysis for the trajectory inference” in kidney scRNA-Seq data analysis. To infer the directionality of the transcriptional changes for our pre-defined EMT trajectories, we considered the subset of cells that belong to the clusters 5, 11, 10, 14, 12, 16, 13, 15 and 1. We redefined the UMAP projection using `RunUMAP` Seurat function on top 17 PCs with `min.dist=0.2`. Next, RNA velocity was inferred with minimum parameters tuning as described in section “RNA-Velocity Analysis” for kidney scRNA-Seq data. Briefly, the initial gene filtering was performed using the `velocityto` function “`score_detection_levels`” by setting the parameters `min_expr_counts=40` and `min_cells_express=30`. Then, the top 3000 high variable genes (HVGs) were selected. HVGs were further filtered based on cluster-wise spliced and unspliced expression threshold as follows: average unspliced expression > 0.01 and average spliced expression > 0.08 in at least one of the clusters. Next, spliced and unspliced counts of filtered genes (n=1231) were normalized for the total number of cells using `_normalize_S` and `_normalize_U` functions, respectively. Then, the Principal Component Analysis (PCA) was performed, and top PCs were selected based on differences between the cumulative sum of explained variance ratio (>0.002), which resulted in a total of 105 PCs. Next, we performed data imputation on 105 PCs with 500 neighbors using `knn_imputation` function. The RNA velocity was estimated by assuming steady-state transition with `delta=1.0`. The estimated velocity is then embedded on the regularized grid with a Gaussian kernel. The step size was set to 40 and smoothness to 0.8 for the velocity embedding. Additionally, we ran `Slingshot` (Street et al., 2018) v2.2.0 over top 15 PCs used for the velocity estimation. `Slingshot` uses a cluster-based Minimum Spanning Tree (MST) algorithm to stably identify global lineage structure of the data. It also fits simultaneous principal curves on predicted lineages which helps to infer the bifurcation points. Based on our previous pseudotime analysis, we run `Slingshot` in semi-supervised manner. We used cluster 5 as a starting cluster whereas cluster 1 and cluster 16 was set as terminal clusters for the lineage inference. Finally, we represented the predicted lineage as a smooth principal curve over the pre-estimated UMAP embedding (Fig. 5b).

Pseudotime analysis for the inferred trajectories: As in kidney analysis, we set a root cell (CTGATAGGTAAGAGGA_1) from cluster 5 having maximum expression of epithelial marker `Lalba` gene. Then the diffusion map (`sc.tl.diffmap`) embedding and the pseudotime (`sc.tl.dpt`) was calculated using default parameters.

SCENIC analysis: To recapitulate the regulatory programme underlying the EMT trajectories, we predicted the scRNA-Seq expression-based regulon using `pySCENIC` v0.11.0 as described in the section “SCENIC Analysis for regulon prediction” for kidney

scRNA-Seq data. The target genes from the regulatory modules with Normalized Enrichment Score (NES) below 1.75 were filtered out. The single cell level regulon activity was calculated using AUCell algorithm implemented in SCENIC package. The regulon activity was averaged by clusters and represented as a heatmap (Fig. 5c). To represent the regulatory activity of individual TFs in each trajectory (EMT-T1 and EMT-T2), we re-calculated the pseudotime separately as described earlier in the section "Trajectories and Pseudotime analysis using PAGA" using "CTGATAGGTAAGAGGA_1" as a root cell in cluster 5. We plot the single-cell level regulon activity score for the EMT-T1 and T2 over pseudotime as described in SCENIC analysis in kidney (Fig. 5d). The enriched motifs for each TF were reported based on their highest NES (Supplementary Table 7).

Trajectories-based differential expression analysis: To investigate the differentially expressed genes along the EMT trajectories we subset the cells using Seurat subset function. Then the Seurat object conversion to monocle3, graph construction and Moran's I test was performed as described in the section "Trajectory-based differential expression analysis" for kidney scRNA-Seq. The genes with significant difference over pre-defined trajectories obtained by Moran's I test ($p < 0.05$) were selected and further filtered to retain the genes with consistent expression (cluster-wise average log2 fold-change). Finally, the single-cell level gene expression heatmaps (scv.pl.heatmap function from scvelo python package) and pathway enrichment analysis was performed as described earlier.

C-3- Trunk Neural Crest scRNA-seq Data Analysis

The raw gene expression matrix for trunk neural crest scRNA-seq (Soldatov et al., 2019) was downloaded from NCBI Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/>) submitted under GEO accession GSE129114. The tSNE embedding and associated metadata for 1107 cells (Fig. 2d) was obtained from http://pklab.med.harvard.edu/ruslan/neural_crest/tSNE_main_Fig1.txt. The Seurat object was created using expression matrix and cell metadata. The connectivity map of mouse trunk NC cell population (from NT to NC-Mes as shown in Fig. 2d) was built using PAGA as described in kidney scRNA-Seq analysis and the trajectory was inferred based on the velocity analysis reported in Soldatov et al., 2019 (Soldatov et al., 2019). We performed PCA as described above and used pre-calculated tSNE embedding for visualizing PAGA connectivity map (resolution = 0.5). The pseudotime analysis was performed as described in kidney scRNA-Seq analysis by setting root cell (SS2_15_0085_F22) from Neural Tube cluster, with maximum expression for neuronal differentiation gene Hes5. Finally, the Moran's I test was performed to infer the differentially expressed genes over the defined trajectory as described for kidney scRNA-Seq analysis and the gene set enrichment analysis was performed against EMT Hallmark and BC-PING. The enrichment scores were visualized over the pre-calculated tSNE embedding (Fig. 2e).

D- Acknowledgment

This work was supported by grants MICIU RTI2018-096501-B-I00 and MCI PID2021-125682NB-I00; FEDER, UE; tAECC Scientific Foundation (FC_AECC PROYE19073NIE); Instituto de Salud Carlos III (CIBERER, CB19/07/00038); Generalitat Valenciana (Prometeo 2021/45), and the European Research Council (ERC AdG 322694), all to Angela Nieto, who also acknowledges financial support from "Centro de Excelencia Severo Ochoa" Grant CEX2021-001165-S funded by MCIN/AEI/10.13039/501100011033. Khalil Kass Youssef was holder of an EMBO Long-Term fellowship, a "Severo Ochoa Excellence Programme" Postdoctoral contract, and currently holds an investigator contract from the AECC Scientific Foundation (Ayudas AECC investigador 2022). Nitin Narwade was a holder of a contract associated with NEUcrest European Union's Horizon 2020 Research and Innovation Program under Marie Skłodowska-Curie (grant agreement No 860635, ITN NEUcrest to Angela Nieto).

Protocol references

- Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z.K., Wouters, J., Aerts, S., 2017. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. <https://doi.org/10.1038/nmeth.4463>
- Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., Trapnell, C., Shendure, J., 2019. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566, 496–502. <https://doi.org/10.1038/s41586-019-0969-x>
- Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., Ma'ayan, A., 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14, 128. <https://doi.org/10.1186/1471-2105-14-128>
- Chevalier, R.L., 2016. The proximal tubule is the primary target of injury and progression of kidney disease: role of the glomerulotubular junction. *Am. J. Physiol. Renal Physiol.* 311, F145–161. <https://doi.org/10.1152/ajprenal.00164.2016>
- La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastrioti, M.E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L.E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., Kharchenko, P.V., 2018. RNA velocity of single cells. *Nature* 560, 494–498. <https://doi.org/10.1038/s41586-018-0414-6>
- Ma, Q., Xu, D., 2022. Deep learning shapes single-cell data analysis. *Nat. Rev. Mol. Cell Biol.* 23, 303–304. <https://doi.org/10.1038/s41580-022-00466-x>
- Petukhov, V., Igolkina, A., Rydbirk, R., Mei, S., Christoffersen, L., Khodosevich, K., Kharchenko, P.V., 2022. Case-control analysis of single-cell RNA-seq studies. <https://doi.org/10.1101/2022.03.15.484475>
- Ransick, A., Lindström, N.O., Liu, J., Zhu, Q., Guo, J.-J., Alvarado, G.F., Kim, A.D., Black, H.G., Kim, J., McMahon, A.P., 2019. Single-Cell Profiling Reveals Sex, Lineage, and Regional Diversity in the Mouse Kidney. *Dev. Cell* 51, 399–413.e7. <https://doi.org/10.1016/j.devcel.2019.10.005>
- Soldatov, R., Kaucka, M., Kastrioti, M.E., Petersen, J., Chontorotzea, T., Englmaier, L., Akkuratova, N., Yang, Y., Häring, M., Dyachuk, V., Bock, C., Farlik, M., Piacentino, M.L., Boismoreau, F., Hilscher, M.M., Yokota, C., Qian, X., Nilsson, M., Bronner, M.E., Croci, L., Hsiao, W.-Y., Guertin, D.A., Brunet, J.-F., Consalez, G.G., Ernfors, P., Fried, K., Kharchenko, P.V., Adameyko, I., 2019. Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* 364, eaas9536. <https://doi.org/10.1126/science.aas9536>
- Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., Dudoit, S., 2018. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19, 477. <https://doi.org/10.1186/s12864-018-4772-0>

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R., 2019. Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888-1902.e21. <https://doi.org/10.1016/j.cell.2019.05.031>

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., Bieri, B., Mazutis, L., Wolf, G., Krishnaswamy, S., Pe'er, D., 2018. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* 174, 716-729.e27. <https://doi.org/10.1016/j.cell.2018.05.061>

Wolf, F.A., Angerer, P., Theis, F.J., 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 19, 15. <https://doi.org/10.1186/s13059-017-1382-0>

Wolock, S.L., Lopez, R., Klein, A.M., 2019. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* 8, 281-291.e9. <https://doi.org/10.1016/j.cels.2018.11.005>

Wu, H., Lai, C.-F., Chang-Panesso, M., Humphreys, B.D., 2020. Proximal Tubule Translational Profiling during Kidney Fibrosis Reveals Proinflammatory and Long Noncoding RNA Expression Patterns with Sexual Dimorphism. *J. Am. Soc. Nephrol. JASN* 31, 23-38. <https://doi.org/10.1681/ASN.2019040337>

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J., Bielas, J.H., 2017. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 14049. <https://doi.org/10.1038/ncomms14049>