Cell Plasticity Trajectories in Developmental and Pathological EMTs

Doctoral Thesis presented by

Nitin Narwade

Thesis Director:

Prof. Maria Angela Nieto Toledano

Thesis Co-Director:

Prof. Igor Adameyko

PhD Program in Neuroscience

Instituto de Neurociencias (UMH-CSIC)

Universidad Miguel Hernández de Elche

- 2024 -

Sant Joan d'Alacant, July 28th, 2024

CONVENTIONAL DOCTORAL THESIS

This Doctoral Thesis, entitled "**Cell Plasticity Trajectories in Developmental and Pathological EMTs**" is presented under the **conventional thesis format with the following publication**:

- **Two distinct Epithelial to Mesenchymal Transition Programmes Control Invasion and Inflammation in Segregated Tumour Cell Populations**
  Nature Cancer, accepted on 5th April 2024 (In Press)

  Khalil Kass Youssef, **Nitin Narwade**, Aida Arcas, Angel Marquez-Galera, Raúl Jiménez, Cristina Lopez-Blau, Hassan Fazilaty, David García-Gutierrez, Amparo Cano, Joan Galcerán, Gema Moreno-Bueno, Jose P. Lopez-Atalaya and M. Angela Nieto

Sant Joan d'Alacant, July 28th, 2024

Prof. **Maria Angela Nieto Toledano**, Director, and Prof. **Igor Adameyko**, co-director of the doctoral thesis entitled **"Cell Plasticity Trajectories in Developmental and Pathological EMTs".**

INFORMAS:

That Mr **Nitin Narwade** has carried out under our supervision the work entitled **"Cell Plasticity Trajectories in Developmental and Pathological EMTs"** in accordance with the terms and conditions defined in his/her Research Plan and in accordance with the Code of Good Practice of the Miguel Hernández University of Elche, satisfactorily fulfilling the objectives foreseen for its public defence as a doctoral thesis.

We sign for appropriate purposes

Thesis director
Prof. Maria Angela Nieto Toledano

Thesis co-director
Prof. Igor Adameyko

Sant Joan d'Alacant, July 28th, 2024

Ms. Cruz Morenilla Palao, Coordinator of the Neurosciences PhD programme at the Institute of Neurosciences in Alicante, a joint centre of the Miguel Hernández University (UMH) and the Spanish National Research Council (CSIC),

INFORMA:

That Mr. **Nitin Narwade** has carried out under the supervision of our PhD Programme the work entitled "**Cell Plasticity Trajectories in Developmental and Pathological EMTs**" in accordance with the terms and conditions defined in its Research Plan and in accordance with the Code of Good Practice of the University Miguel Hernández de Elche, fulfilling the objectives satisfactorily for its public defence as a doctoral thesis.

Which I sign for the appropriate purposes

Dra. Cruz Morenilla Palao

Coordinator of the PhD Programme in Neurosciences

E-mail: elvirap@umh.es
Website: www.in.umh.es
Tel: +34 965 919533
Fax: +34 965 919549
Av Ramón y Cajal s/n
SANT JOAN CAMPUS
03550 SANT JOAN D'ALACANT- SPAIN

Sant Joan d'Alacant, July 28th, 2024

Funding/Grant/Scholarship:

- First three years of PhD research was supported by, European Union's Marie Skłodowska-Curie Actions (MSCA) Research and Innovative Training Networks (ITN) H2020-MSCA-ITN-2019 (grant agreement No 860635 to Angela Nieto), project entitled "Training European Experts in Multiscale Studies of Neural Crest Development and Disorders: from Patient to Model Systems and Back again - NEUcrest" and another contract associated with the MCI PID2021-125682NB-I00 grant.

The research has been also funded by the following research projects:

Grants MCI PID2021-125682NB-I00; FEDER, UE; AECC 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), o Angela Nieto, who also acknowledges financial support from "Centro de Excelencia Severo Ochoa" Grant CEX2021-001165-S funded by MCIN/AEI/ 10.13039/501100011033.

To my beloved “Parents & Family…”

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to all those who have directly or indirectly supported and guided me throughout my PhD journey.

First and foremost, I am profoundly thankful to my PhD supervisor, Prof. M. Angela Nieto, for her invaluable guidance, encouragement, and insightful advice throughout this research. Her expertise and dedication have been instrumental in shaping this work and my career. Her trust, advice and life's lessons were invaluable to me and will always have an imprint on my future endeavours.

I am also deeply grateful to my PhD co-supervisor, Prof. Igor Adameyko, for his constructive feedback and insightful suggestions which greatly enhanced the quality of my work.

A special thanks goes to Dr. Khalil Kass Youssef, who has not only closely mentored me but has also imparted important life lessons. His wisdom and guidance have been a cornerstone of both my professional and personal development.

I would like to extend my sincere thanks to Raul Jimenez-Castaño for his logical insights and engaging discussions, which have significant impact on my personal and professional growth.

I am also grateful to all the other members of our lab: Angelita, Jussep, Carlos, Adrian, Sonia Vega, Teresa Gomez-Martinez, Cristina Lopez-Blau, Gemma Osuna-Tenorio, Javier Rodriguez-Baena, Marta Arumi-Planas, Pablo Ballesteros-Martinez, Francisco Cabello-Torres, Francisco Gracia-Quiles for creating a collaborative and inspiring working atmosphere. A special thanks goes to Auxi Casanova and Sonsoles Segur-Juarez for kindly assisted me with complicated administrative works.

To Joan, thank you so much for your unwavering support during difficult times and for your invaluable moral support. Your kindness and understanding have been a great source of strength.

I am immensely grateful to the NUCrest-ITN network for the funding, connections and support, which have been crucial for the successful completion of this

research. In addition, I convey my gratitude to the scientific and executive members of the Instituto de Neurociencias, especially M. Teresa García-Hedo.

Thanks to my friends who have been like a second family abroad—Sanjay Vasudevan, Doris Santiago, Moumita Chatterjee, Mahima Laxmeesha, and Pablo Ballesteros-Martinez—thank you all for your unwavering support, companionship, and for making this journey more enjoyable.

A heartfelt thank you to my childhood friend, Dr. Sandeep Karanjkar, whose inspiration and support have always motivated me to strive for excellence. Thanks to Ms. Alka Lohat for her moral support. I would like to extend my gratitude to Dr. Abhijeet Kulkarni, Dr. Roli Budhwar, Mr. Rohit Shukla, Dr. Dhiraj Dhotre, Dr. Yogesh Souche, Dr. Dattatray Mongad, Dr. Shreyash Kumbhare and Prof. Edwin Cheung, for creating a strong foundation for my research interests and helped to spark my passion for this field. Your early guidance and support have been pivotal in my academic journey.

Finally, to my parents, Pundlikrao Narwade and Vimalbai Narwade, your brave hearts and unwavering faith have been my guiding lights. Your sacrifices and support have made this journey possible. To all my family members—Dada-Vahini, Asmita Tai, Kranti Tai, and Kirti Tai and all other family members—thank you for your constant encouragement and belief in me.

INDEX

Table of Contents

ABBREVIATIONS

Abbreviations

ATAC-Seq	Assay for Transposase-Accessible Chromatin with Squencing
AUC	Area under the Curve
BC	Breast Cancer
BC-PING	Breast Cancer Pro-INvasive Genes
BM	Basement Membrane
CAFs	Cancer Associated Fibroblasts
CC	Cancer Cells
CCA	Canonical Correlation Analysis
CoDA	Compositional Data Analysis
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTC	Circulating Tumour Cells
CV	Cross Validation
DNA	DeoxyriboNucleic Acid
E	Epithelial
EC	Endothelial Cells
ECM	Extracellular Matrix
EMT	Epithelial Mesenchymal Transition
FA	Folic acid Administration
FACS	Fluorescence-Activated Cell Sorting
FBS	Foetal Bovine Serum
FC	Fold Change
FGF	Fibroblast Growth Factor
FN	False Negative
FP	False Positive
GEM	Gel Bead-In-Emulsions
GEO	Gene Expression Omnibus
GO	Gene Ontology
GRN	Gene Regulatory Networks
GSEA	Gene Set Enrichment Analysis
GSVA	Gene Set Variation Analysis
HVG	High Variable Genes
KEGG	Kyoto Encyclopedia of Genes and Genomes pathways
KNN	K-Nearest Neighbors
LC	Lymphoid Cells
LR	Logistic Regression
M	Mesenchymal
MAGIC	Mrkov Affinity-based Graph Imputation of Cells
MC	Myeloid Cells

MCC	Matthews Correlation Coefficient
MET	Mesenchymal to Epithelial Transition
miRNA	Micro-RNA
MLP	Multi-Layer Perceptron
MMTV-PyMT	Mouse Mammary Tumor Virus-Polyoma Middle Tumor-antigen
MSigDB	Molecular Signatures Database
MST	Minimum Spanning Tree
NC	Neural Crest
NC	Neural Crest
NCBI	National Center for Biotechnology Information
NES	Normalized Enrichment Score
NTN	NephroToxic-serum-induced Nephritis
OvR	One versus Rest
PAGA	PArtition-based Graph Abstraction
PC	Principal Component
PCA	Principal Component Analysis
PT	Proximal Tubule
QC	Quality Control
RNA	RiboNucleic Acid
RNA-Seq	RiboNucleic Acid Sequencing
SNN	Shared Nearest Neighbor
SVD	Singular Value Decomposition
TF	Transcription Factors
TFIDF	Term-Frequency Inverse-Document-Frequency
TGFβ	Transforming Growth Factor-β
TME	Tumour MicroEnvironment
TN	True Negative
TNBC	Triple Negative Breast Cancer
TP	True Positive
t-SNE	t-distributed Stochastic Neighbor Embedding
TSS	Transcription Start Sites
UMAP	Uniform Manifold Approximation and Projection
UMI	Unique Molecular Identifiers
UUO	Unilateral Ureteral Obstruction
WT	Wild Type

LIST OF FIGURES AND TABLES

List of figures and tables

ABSTRACT

Abstract

The Epithelial to Mesenchymal Transition (EMT) induces cell plasticity during embryonic development and tissue repair, but it also promotes tumour progression and organ degeneration. EMT is a biological process that allows a polarized epithelial cell, which normally interacts with the basement membrane via its basal surface, to undergo multiple biochemical changes that enable it to transit the spectrum between the epithelial and mesenchymal phenotypes. These changes are fundamental during embryonic development and in pathological states, particularly in cancer and degenerative diseases such as organ fibrosis.

In our research, we have analysed data from bulk and single-cell transcriptomes from cell lines, embryonic neural crest, and mouse models of renal fibrosis and breast cancer developed in the lab. We have employed comprehensive bioinformatics and computational approaches to dissect the transcriptomic landscapes, revealing the heterogeneity within the EMT processes. The embryonic EMT trajectory is characterized by a robust activation of genes associated with cell motility and invasiveness. Conversely, the adult EMT trajectory is enriched in inflammatory and immune response-related pathways.

Importantly, our data analyses have also unveiled that there is no cancer-specific EMT programme. Instead, epithelial cancer cells dedifferentiate and bifurcate into two distinct cellular trajectories in segregated populations after activating either embryonic-like (invasive) or adult-like (inflammatory and non-invasive) EMTs. Within a single tumour, these cell trajectories respectively drive dissemination or inflammation.

Our data analyses and in silico perturbation approaches predicted that two EMT transcription factors play important and distinct roles: Snail1 acting as a pioneer factor in both EMT trajectories, compatible with the known role of Snail1 in repressing epithelial genes such as E-cadherin and inducing mesenchymal genes like vimentin. Meanwhile, Prrx1 is specific and crucial for the progression of the embryonic-like invasive trajectory, facilitating the transition to invasion, first and essential step for metastatic dissemination. Additionally, we could predict that the two trajectories might be plastic and interdependent.

Functional analyses in experimental models in the lab confirm our predictions. The abrogation of the EMT invasive trajectory by deleting Prrx1 specifically in cancer cells not only significantly reduces metastatic burden but also increases the cancer cells contributing to the EMT inflammatory trajectory, enhancing the recruitment of antitumor macrophages.

In conclusion, our study underscores the importance of EMT in cancer progression and highlights its dual role in promoting metastasis and modulating immune responses. We provide a framework for future research aimed at developing targeted therapies to combat cancer more effectively. For instance, targeting the EMT process could simultaneously inhibit metastatic spread and modulate the immune microenvironment to favour antitumor responses.

Resumen

La Transición Epitelio-Mesénquima (EMT) induce la plasticidad celular durante el desarrollo embrionario y la reparación de tejidos, pero también promueve la progresión tumoral y la degeneración de órganos. Es un proceso que permite a una célula epitelial polarizada, que normalmente interactúa con la membrana basal a través de su superficie basal, experimentar múltiples cambios bioquímicos que le permiten transitar el espectro entre los fenotipos epitelial y mesenquimal. Estos cambios son fundamentales durante el desarrollo embrionario y en estados patológicos, como el cáncer y enfermedades degenerativas como la fibrosis.

En nuestra investigación, hemos analizado datos de transcriptomas de células únicas en líneas celulares, cresta neural embrionaria y modelos murinos de fibrosis renal y cáncer de mama desarrollados en el laboratorio. Hemos empleado enfoques bioinformáticos y computacionales exhaustivos para desentrañar las redes transcriptómicas, revelando la heterogeneidad dentro de los procesos de EMT. La trayectoria de EMT embrionaria se caracteriza por una robusta activación de genes asociados con la motilidad e invasividad celular. Por el contrario, la trayectoria de EMT adulta está enriquecida en vías relacionadas con la respuesta inflamatoria e inmune.

Es importante destacar que nuestros análisis de datos también han revelado que no existe un programa de EMT específico para el cáncer. En cambio, las células cancerosas epiteliales se desdiferencian y bifurcan en dos trayectorias celulares distintas en poblaciones segregadas de un mismo tumor tras activar EMT de tipo embrionario (invasivo) o de tipo adulto (inflamatorio y no invasivo). Dentro de un solo tumor, estas trayectorias celulares controlan respectivamente la diseminación o la inflamación.

Nuestros análisis de datos y ensayos de perturbación in silico predijeron que dos factores de transcripción de EMT tienen papeles importantes y distintos: Snail1 actúa como un factor pionero en ambas trayectorias de EMT, compatible con su capacidad de reprimir la expresión de genes epiteliales como la E-cadherina e inducir genes mesenquimales como vimentina. Por otra parte, Prrx1 es específico y crucial para la progresión de la trayectoria de tipo embrionario,

facilitando la transición a la invasión, primer paso esencial para la diseminación metastásica. Además, pudimos predecir que las dos trayectorias podrían ser plásticas e interdependientes.

Los análisis funcionales en modelos experimentales en el laboratorio confirman nuestras predicciones. La deleción de Snail específicamente en células cancerosas previene la progresión tumoral global y la deleción de Prrx1 inhibe la trayectoria invasiva, no solo reduciendo significativamente la carga metastásica, sino también aumentando las células cancerosas que contribuyen a la trayectoria inflamatoria, aumentando el reclutamiento de macrófagos antitumorales.

En conclusión, nuestro estudio pone de manifiesto la importancia de la EMT en la progresión del cáncer y destaca su doble papel en la promoción de la metástasis y la modulación de las respuestas inmunes. Proporcionamos un escenario para desarrollar terapias específicas para combatir el cáncer de manera más efectiva. Por ejemplo, modular la EMT podría inhibir simultáneamente la diseminación metastásica y cambiar el microambiente inmune para favorecer las respuestas antitumorales.

Chapter 1

INTRODUCTION

1.1 Epithelial to mesenchymal transition (EMT)

EMT is a highly orchestrated and dynamic biological process whereby epithelial cells undergo significant molecular changes leading to change their phenotype by losing epithelial characteristics and acquiring mesenchymal traits. It can be triggered in different physiological and pathological contexts. Epithelial cells are normally attached to a basement membrane (BM) through integrin molecules along the basal surface and with the adjacent cells through different adhesion complexes including adherens junctions, tight junctions, gap junctions, desmosomes, etc. (Figure 1) (Nieto et al., 2016). With this tight and well-defined shape and structure, epithelial cells are essentially immotile. EMT is a continuous and dynamic process in which the cells transition through series of molecular changes reaching different stable and metastable states within the E-M spectrum (Figure 1) (Nieto et al., 2016, Thiery et al., 2009).

It is a complex molecular process which involves activation of several signalling pathways, among, the most potent is the one triggered by transforming growth factor-β (TGFβ). Additionally, the ligands of receptor tyrosine kinases, WNT and NOTCH, either independently or in combination, activate the EMT effector genes (Youssef and Nieto, 2024) that lead to the repression of epithelial markers (E-cadherin, claudins, occludins, etc.) and the gain in mesenchymal markers (N-cadherin, vimentin, fibronectin, collagen, etc.). These molecular changes lead to a loss of cell-cell adhesion complexes and gain a remodelling of the cytoskeleton and the extracellular matrix (ECM) (Figure 1).

EMT primarily occurs during embryonic development, that enables epithelial cells to delaminate and migrate far away from their origin. Epithelial cells adopt migratory behaviour, and it is often a transient process by which cells progress towards the mesenchymal phenotype and revert back through a mesenchymal to epithelial transition (MET) in order to re-epithelialize and differentiate into different derivatives to fulfil their function (Nieto et al., 2016; Youssef and Nieto, 2024). In the adult, EMT is reactivated to maintain tissue homeostasis after acute damage. For instance, during would healing, epithelial cells undergo a EMT, migrate to the wound edge and help in sealing the scar (Nieto et al., 2016; Thiery et al., 2009; Youssef and Nieto, 2024). In addition to normal developmental and physiological

contexts, epithelial cells reactivate EMT programme in pathologies such as cancer and fibrosis (Figure 2).

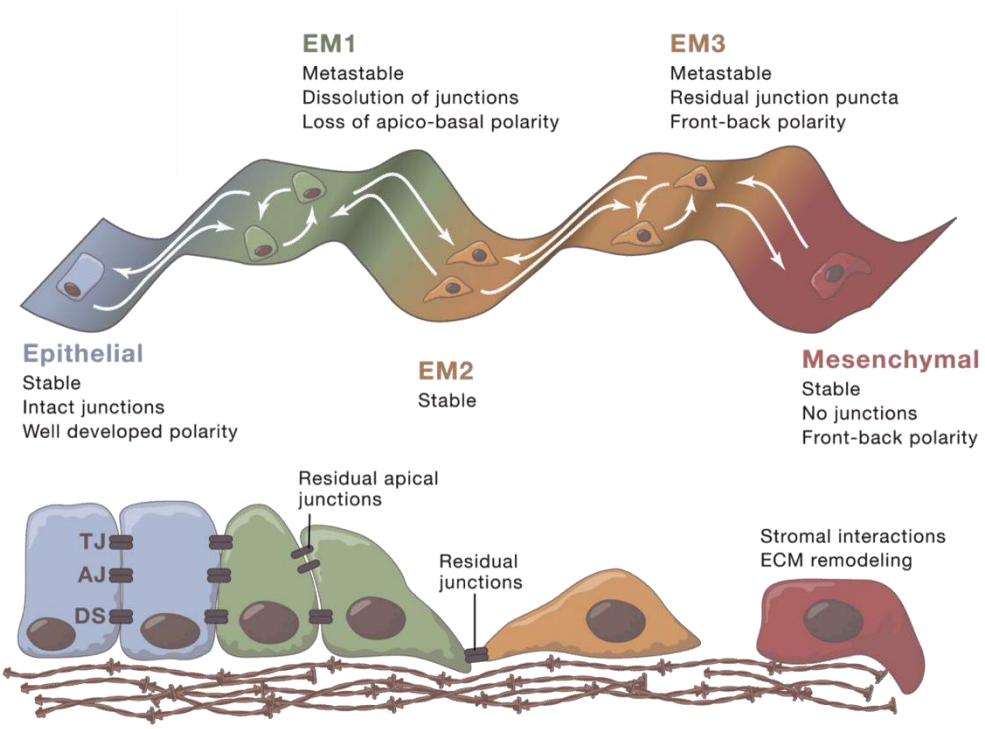


Figure 1 | Epithelial to mesenchymal transition. The cells of epithelial origin maintain their structural integrity by a combination of features such as, cell-cell adhesion, apico-basal polarity and attachment to the basement membrane, among others. These features are executed through the action of different complexes and molecules including adherens junctions, desmosomes, tight junctions, gap junctions, etc. During EMT, the epithelial cells gradually lose their epithelial properties which results in the loss of apico-basal polarity, cell to cell adhesion, and contact with basement membrane, simultaneously gaining front-back polarity, enhanced cell-matrix interactions, and invasiveness. A cell can reach multiple stable and meta-stable states within the EMT spectrum. Intermediate states are called hybrid states, result of the activation of a partial EMT. Cells maintain both epithelial and mesenchymal characteristics. EMT is a reversible process, the reverse process is called mesenchymal to epithelial transition (MET). TJ: tight junction; AJ: adherens junction; DS: desmosome. (Taken from Nieto et al., 2016).

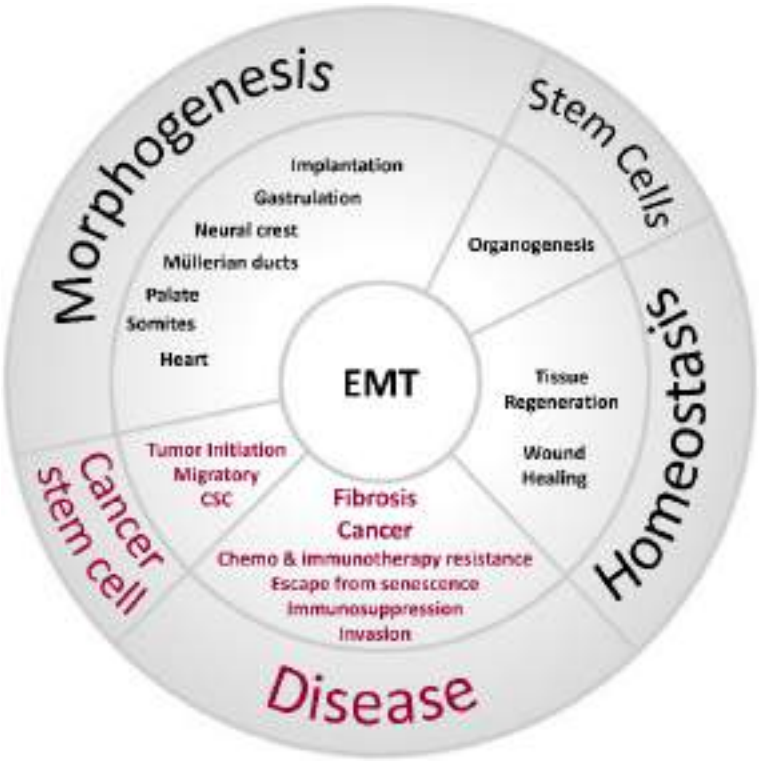


Figure 2 | Reactivation of EMT in different patho-physiological conditions. EMT plays important roles in different patho-physiological conditions. Firstly, in organogenesis (embryonic development) and later in recovering tissue homeostasis after injury. Importantly, it is reactivated in some chronic diseases such as kidney fibrosis and in cancer progression. The activation and reactivation of EMT in different conditions has commonalities, but also bear cell or tissue specificities. (Taken from Thiery et al., 2009)

EMT is a tightly regulated process and is driven by a set of transcription factors (EMT-TFs) downstream of the first inducers, different signalling pathways activated by extracellular signals. The EMT-TFs belong to SNAIL, TWIST, ZEB and PRRX1 families. Thus, a tight spatiotemporal regulation of EMT-TFs is essential, and controlled by complex gene regulatory networks (GRNs) (Nieto et al., 2016) (Figure 3). SNAIL1 was the first EMT-TF reported, and as such, the first repressor of *E-Cadherin* transcription (Batlle et al., 2000; Cano et al., 2000). SNAIL1 is a potent epithelial repressor, regulating the progressive loss of cell–

cell adhesion, apicobasal polarity and epithelial differentiation (Nieto et al., 2016) (Figure 3). Additionally, SNAIL1 can turn into a transcriptional activator promoting the expression of ECM molecules and different cytokines (Youssef and Nieto, 2024). ZEB family members can also play a dual role as a epithelial repressor, recruiting histone deacetylase (HDAC1/2 complexes) to the *E-cadherin* promoter while promoting mesenchymal activation interacting with AP-1 factors (FOSL1 and JUN) (Feldker et al., 2020; Lehmann et al., 2016). Both TWIST and PRRX are potent mesenchymal inducers, promoting invasiveness in epithelial cells (Ocaña et al., 2012a; Soldatov et al., 2019).

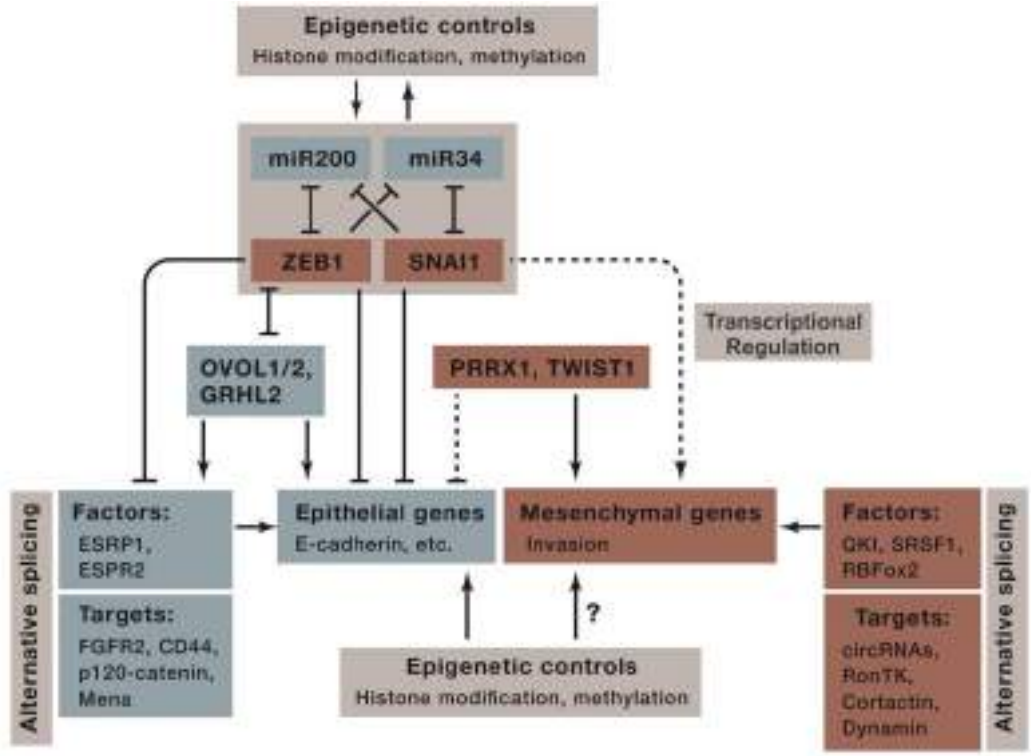


Figure 3 | Regulatory circuits in Epithelial to mesenchymal transition. EMT is a tightly regulated molecular process where spatio-temporal regulation of the expression of EMT-TFs is at utmost importance. The repression of epithelial characteristics is concomitant with the activation of mesenchymal traits. Several regulatory circuits such as direct regulation by EMT-TFs, regulation by microRNAs, alternative splicing, and epigenetic mechanisms play important roles, and alone or in combination help epithelial cells to initiate and maintain EMT program. (Taken from Nieto et al., 2016).

The regulation of the EMT-TFs is crucial to activate or to repress EMT, and the latter is necessary for cells to re-epithelialize and differentiate. Inhibiting EMT-TFs helps the cells to become more epithelial and can be mediated by various regulatory mechanisms. One direct regulatory loop involves TFs such as OVOL1/2 and GRHL2, which bind directly to the EMT-TF promoters or indirectly regulate them to protect the epithelial phenotype (Nieto et al., 2016). Additionally, several regulatory loops are mediated by post-transcriptional regulation, particularly through microRNAs (miRNAs). For example, the interaction between the *miR-200* family and ZEB proteins is a well-characterized feedback loop in EMT regulation (Gregory et al., 2008; Park et al., 2008) (Figure 3). The *miR-200 family, including miR-200a, miR-200b, miR-200c, miR-141,* and *miR-429*, directly binds to the mRNA of ZEB1 and ZEB2, leading to their degradation and inhibition of EMT progression. In turn, ZEB proteins downregulate the miR-200 family, creating a double-negative feedback loop. Another such example is the regulatory loop between miR-34 and SNAIL1 (Siemens et al., 2011) (Figure 3). miR-34 directly decreases SNAIL1 transcripts, while SNAIL1 protein repress *miR-34* expression. In our lab, we previously demonstrated an interesting gene regulatory network between two EMT-TFs, SNAIL1 and PRRX1 (Fazilaty et al., 2019). *miR-15* downregulates SNAIL1 transcript levels, and SNAIL1 directly downregulates *PRRX1* transcription, creating a double-negative feedback loop essential for selecting a specific EMT modality, either driven by Snail1 or Prrx1 (Fazilaty et al., 2019). The regulatory loops between EMT-TFs and various regulatory controls, including microRNAs, other TFs, and histone modifications, form a complex network of interactions. This network ensures the context specific and spatiotemporal regulation of EMT activation, progression, and reversion.

1.2 EMT in embryonic development

EMT is a fundamental process in embryonic development, as with the exception of the epidermis and the anterior central nervous system, the progenitors of nearly all the tissues and organs undergo at least one round of EMT (Youssef and Nieto, 2024). For instance, during kidney development, the cells undergo multiple rounds of EMT, and its reverse process, MET. Initially, the metanephric mesenchyme, a group of cells in the developing kidney, undergoes MET to form epithelial structures that will finally give rise to the nephrons, the functional units

of the kidney. Later, EMT is involved again for the development of the kidney's supportive tissue (Ho, 2014). Despite of differential activation of EMT in different tissues and in different species, the crucial role of EMT in morphogenesis is conserved throughout evolution (Thiery et al., 2009). Embryonic cells activate different degrees of EMT depending on their fate, migratory path and final destination (Youssef and Nieto, 2024). In mammals, the very first EMT occurs is during embryo implantation, in which the blastocyst adheres and invades the uterus (Acloque et al., 2009) (Figure 4). The parietal endodermal cells undergo EMT to form an extraembryonic cell population that will be incorporated into the yolk sac (Thiery et al., 2009) (Figure 4). repress E-cadherin and thereby unable to migrate (Carver et al., 2001; Nieto et al., 1994).

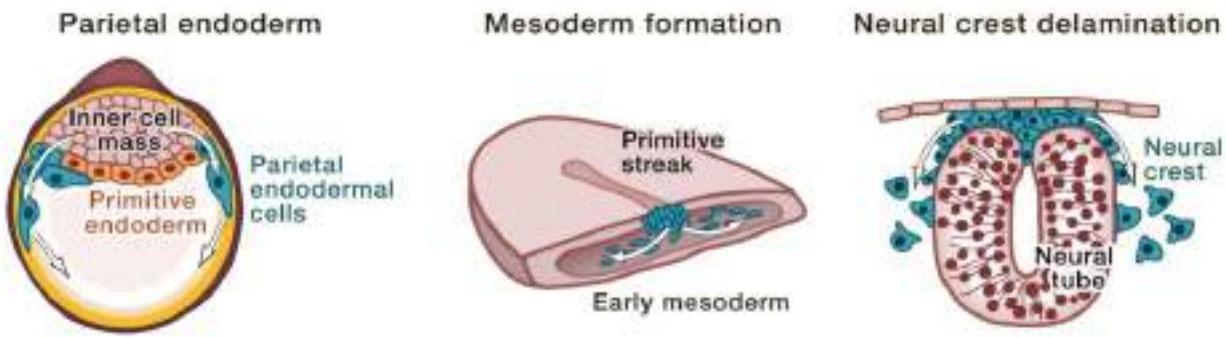


Figure 4 | Primary EMTs during embryonic development. During early mouse embryonic development, cells undergo EMT for the formation of parietal endoderm (left). After implantation, the first EMT occurs at gastrulation, when the mesodermal precursors delaminate and ingress (middle). The delamination of the neural crest cells from the neural tube during late gastrulation is another primary EMT process to delaminate and give rise to multiple derivatives including the peripheral nervous system and the craniofacial skeleton (right). (Taken from Thiery et al., 2009).

At gastrulation stages, epiblast cells undergo EMT and delaminate from the primitive streak, acquiring various phenotypes and behaviours depending on their fate. The EMT activation and progression during gastrulation relies upon activation of canonical EMT inducers such as the transforming growth factor-β (TGFβ) and fibroblast growth factor (FGF) families, WNT, and focal adhesion kinase signalling pathways, etc. that lead to the activation of the EMT-TFs (Youssef and Nieto, 2024). During vertebrate embryonic development, the lack of SNAIL results in the failure to gastrulate. Mesodermal progenitors cannot repress E-cadherin and thereby unable to migrate (Carver et al., 2001; Nieto et al., 1994).

At late gastrulation and early neurulation stages, the neural crest (NC) cells appear at the dorsal border of the neural plate and undergo delamination and migration by activating EMT. Depending on the degree of activation of the EMT programme and the interaction with the microenvironment, their migratory behaviour can vary, from individual cell migration, when NC cells reach full mesenchymal state, or collective NC migration if they activate a partial EMT programme (Piacentino et al., 2020). Upon reaching the target site, the NC cells have the ability to differentiate into a wide range of cell lineages including, neurons, glial cells, some endocrine and para-endocrine cells, cartilages, bones and all melanocytes (Acloque et al., 2009) (Figure 5).

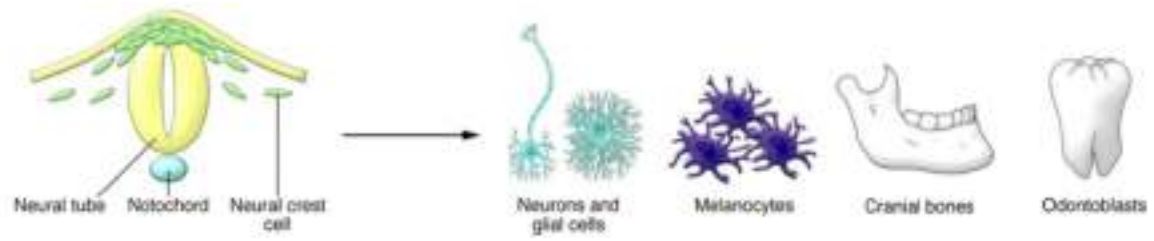


Figure 5 | Neural crest development and its derivatives. Neural crest cells are the multipotent embryonic cells. During vertebrate embryonic development, neural crest cells activate EMT to delaminate from dorsal neural tube at late gastrulation stages and adopt migratory properties. Neural crest cells can differentiate into various derivatives such as neuronal and glial cells, melanocytes, chondrocytes, osteoblasts, etc. (Taken from Acloque et al., 2009).

1.3 EMT in fibrosis

Fibrosis is a chronic degenerative disorder associated with a gradual loss in function of an organ and eventual organ failure. In fibrotic conditions, there is an accumulation of myofibroblasts, which results in secretion and deposition of an excessive amount of collagen as fibers leading to tissue stiffness and degeneration. The resident and recruited fibroblasts differentiate into myofibroblasts that can generate collagen network as well (Nieto et al., 2016) (Figure 5). Fibrosis occurs in many different organs including kidneys, liver, lung, heart, etc.

The role of EMT in organ fibrosis has long been debated, in particular the origin of the myofibroblasts, as to whether EMT could convert renal epithelial cell into myofibroblasts or not (Humphreys et al., 2010; LeBleu et al., 2013). Recent studies with lineage tracing and single-cell analyses confirmed that epithelial cells are not the origin of collagen-producing myofibroblasts in the lung, kidney or liver, however, strong evidences support the role of an irreversible EMT as a driver of fibrosis (Youssef and Nieto, 2024). As such, previous studies in the lab showed that EMT is activated in the renal epithelial cells, and TWIST1 or SNAIL1 depletion in renal epithelial cells significantly attenuates interstitial fibrosis in mouse models induced by unilateral ureteral obstruction (UUO), folic acid administration (FA), or nephrotoxic-serum-induced nephritis (NTN) (Grande et al., 2015; Lovisa et al., 2015). On the contrary, SNAIL1 and EMT reactivation promotes the fibrosis and renal failure (Boutet et al., 2006). However, despite of activating EMT renal epithelial cells do not engage into a migratory or invasive programme (Grande et al., 2015). Interestingly, these injured renal epithelial cells secrete TGFβ to promote the conversion of interstitial fibroblasts into myofibroblasts, and secrete cytokines and chemokines to promote both fibrogenesis and inflammation, the two hallmarks of fibrosis (Nieto et al., 2016; Youssef and Nieto, 2024) (Figure 6).

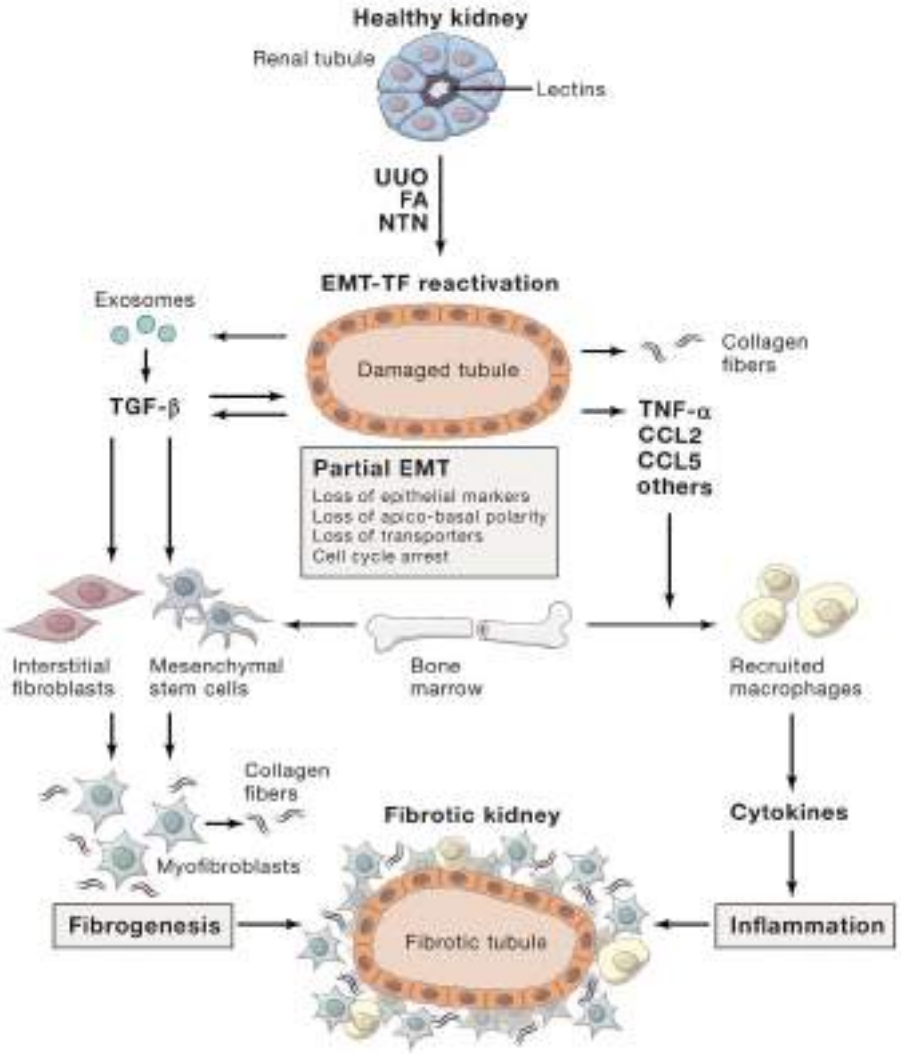


Figure 6 | Reactivation of EMT during kidney fibrosis. Renal epithelial cells reactivate EMT during fibrosis induced by unilateral ureteral obstructions (UUO), folic acid treatment (FA), or nephrotoxic serum-induced nephritis (NTN). The damaged tubular epithelial cells maintain a partial EMT state by activating EMT-TFs such as Snail1 and Twist1 (Grande et al., 2015, Lovisa et al., 2015), but do not engage in a migratory programme. Instead, they promote myofibroblast differentiation by secreting TGFβ, and modulate the immune microenvironment by secreting different cytokines and chemokines, thereby promoting fibrogenesis and inflammation respectively. (Taken from Nieto et al., 2016).

1.4 EMT in cancer

More than 80% of cancers arise from epithelial tissues, which are called carcinomas, and amongst them, more than 90% of cancer-associated deaths are due to metastasis. Metastases are secondary tumours formed in distant organs and originated from the tumour cells disseminated from the primary site. The cancer metastasis is a condition for which no or poor treatments are currently available (Hanahan and Weinberg, 2011, Riggi et al., 2018). Cancer cells hijack the embryonic EMT programme to become invasive and migratory, to invade adjacent tissues and disseminate through blood circulation finally extravasating and reaching secondary sites (Nieto et al., 2016, Lambert et al., 2017) (Figure 7). Interestingly, for metastatic colonization to occur, EMT should be reverted back through the MET process for cells to metastasise in a distant organ (Ocaña et al., 2012a; Tsai et al., 2012). Within one single tumour there is significant cell heterogeneity, not only with respect to the mutational burden being accumulate during tumour evolution but also due to the activation of EMT for cells to disseminate and (Figure 7). Recently, in several models, different degrees of epithelial plasticity have been linked to metastatic potential in primary tumours (Latil et al., 2017; Pastushenko et al., 2018), and single cell RNA-sequencing (scRNAseq) of human cancer patients has identified the existence of partial EMT states with high metastatic potential (Puram et al., 2017). Additional, scRNA-Seq based analysis revealed different EMT states during the progression in different types of cancers (Gavish et al., 2023). Furthermore, the triple-negative subtype (ER-/PR-/HER2-) of breast cancers tend to have more mesenchymal phenotype in circulating tumour cells (CTCs). However majority of CTCs exhibits an intermediate/partial EMT phenotype (Nieto et al., 2016). This diversity of circulating CTCs and primary tumours provide evidence of EMT induction during cancer progression for metatstic dissemination (Ocaña et al., 2012a; Salnikov et al., 2012; Ye et al., 2015; Yu et al., 2013). Targeting EMT can be a double-edged sword, because while it can prevent spreading of cancer cells from the primary site, it might also help already disseminated cells to revert to an epithelial state for metastatic colonization (Harper et al., 2016; Hosseini et al., 2016). This underscores the need for a deeper understanding of EMT before developing targeted cancer therapies (Nieto et al., 2016).

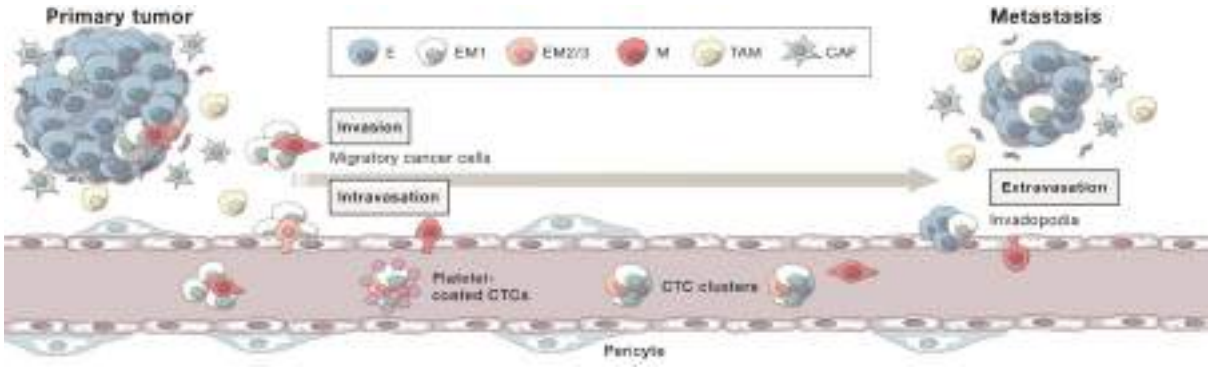


Figure 7 | Reactivation of EMT during cancer progression. Tumours of epithelial origin can reactivate EMT which help tumour cells to form distant metastasis. EMT induces invasiveness of the primary tumour cells which enables them to disseminate from primary tumour site and intravasate into the blood stream. The cells in blood stream called circulating tumour cells (CTCs) then extravasate upon reaching the target site and colonize their new destination. (Taken from Nieto et al., 2016).

1.5 Multi-omics

As described earlier, EMT is a continuous and dynamic molecular process. Numerous genes undergo changes in expression dynamics in a tightly regulated manner across the epithelial-mesenchymal spectrum, activating different pathways that lead to phenotypic changes, both in development and disease. Studying specific markers and their roles in EMT regulation can provide valuable insights into tumour evolution. However, to gain a broader understanding, there is a need to perform more global and integrative analyses. Such comprehensive studies help to identify overall changes in a particular cellular programme such as EMT and may lead to the discovery of potential new targets. Using multi-omics approaches, integrating various molecular modalities and understanding their interdependencies can help to decipher the complex regulatory networks involved in EMT in different contexts.

The basis of omics approach is to quantify the level of a particular type of molecule in biological samples to infer the patterns with respect to the sample attributes under investigation in a high throughput manner. Genomic studies measure DNA molecules, whereas epigenomic, transcriptomic, proteomic, and metabolomic studies measure the chemical states of DNA and its binding proteins, RNA, proteins, and metabolites, respectively (Yamada et al., 2021). The term bulk-*omics is used when a pull of cells is used to obtain the molecular modalities in a specific sample or tissue under study. Since the first successful libraries sequenced in mid-2000 on Roche 454 sequencer (Emrich et al., 2007) the bulk-RNASeq approach became one of the most valuable and extensively used tools in biology (Li and Wang, 2021). However, bulk-*omics analyse the average signal from a pull of a large cell populations, precluding the identification of the cellular heterogeneity within a sample. EMT is a heterogeneous and dynamic process where cells undergo a series of transitions, making bulk-*omics less suitable, except for cell line cultures where cells are homogenous. Therefore, single-cell omics is a much more suitable and informative approach for studying EMT, as it can capture the detailed changes and diversity in each single cell, allowing to capture the molecular differences in much better resolution.

1.6 Single-cell RNA Sequencing (scRNA-Seq)

As previously mentioned, the cell-averaged profiles obtained in bulk omics approach cannot uncover the heterogeneity and unique characteristics of individual cells. To tackle this problem, it is important to examine gene expression at the single-cell level. Generally, a single-cell experiment follows similar steps as a bulk RNA-Seq experiment with several tunning and adaptations. Like bulk sequencing, single-cell sequencing requires lysis, reverse transcription, amplification, and finally, high throughput sequencing. In addition, single-cell sequencing requires segregation of individual cells and cell labelling. The latter allows to identify and assign captured molecules to the cell of origin. Single-cell sequencing approaches use two major protocols: plate-based (Ramsköld et al., 2012) and microfluidic-based assays (Macosko et al., 2015), each with its own set of advantages and disadvantages. A plate-based protocol requires the manual isolation of cells into 96-well or 384-well plates, limiting the number of cells that can be sequenced in a single run. However, this method offers higher

sequencing coverage (completeness) and depth. On the other hand, a much larger number of cells can be sequenced using microfluidic-based assays, but with lower sequencing depth and coverage. In scRNA-seq, data can be generated either by sequencing full-length or only the ends (3' or 5') of the transcripts, called tag-based assay. Full-length sequencing provides comprehensive coverage and is useful for detecting isoforms, variation analysis, and detection of gene fusions. Tag-based sequencing is more cost-effective and allows for the analysis of a larger number of cells, making it ideal to study rare cell types. Given the complexity of EMT and the fact that only a fraction of cells undergoes EMT in pathological conditions like cancer, a higher number of cells could be an advantage. Therefore, we have used the 3'-end sequencing approach to obtain data from a higher number of cells, enabling a more comprehensive analysis of EMT dynamics.

Despite of its increasing use, like any other technology, scRNA-Seq has certain limitations. First, a single-cell technology allows profiling of thousands or even millions of cells, but with lower depth. Second, the downstream analysis becomes complex and more challenging due to the amount of data and its higher dimensionality, which may lead to drawing false conclusions in case if the study is not robust enough and lacks sufficient controls.

1.6.1 scRNA-Seq Data Analysis

It is essential to perform a systematic data analysis with appropriate caution to reach any conclusion from scRNA-Seq data. There are several steps such as quantification, quality control (QC), filtering, data normalisation which can greatly impact on the downstream analysis (Heumos et al., 2023; Luecken and Theis, 2019). Before, moving forward with downstream analyses, it is crucial to check the quality of sequencing reads, which mainly includes parameters such as Phred quality score, GC content, and sequencing adapter contamination. The Phred quality score indicate the measure of base quality in sequencing. Since each species has a specific range of GC content in its genome, any deviation from this range can indicate the presence of foreign sequence contamination. Generally, the modern tools implement some basic QC and filtering for the sequencing data, if not these steps need to be done before proceeding with downstream analysis.

Once the bad quality reads were removed from the raw data, next step in scRNA-Seq data analysis is the quantification where a "*gene X cell* matrix" is built by counting the reads mapped to the transcriptome/genome. In tag-based assay, as only 3' or 5' end sequences are available for quantification, it is more difficult to unambiguously map the sequences on the transcriptome/genome. However, the usage of unique molecular identifiers (UMIs), can be useful to resolve amplification biases during the quantification (Islam et al., 2014). UMIs are the unique barcodes generated for each of the transcripts in a cell. After quantification, it is important to check the reads mapping to specific genomic regions, As the sequences are from mRNA, the prediction is that the majority of the reads map to exonic regions.

1.6.2 Quality control

Several parameters have been proposed to filter out bad quality cells from scRNA-Seq data, including the number of detected genes, number of UMIs, percentage of mitochondrial genes, etc. (Ilicic et al., 2016; Luecken and Theis, 2019). A cell with a low number of detected genes, low number of UMI and high percentage of mitochondrial genes, can be considered as a dying cell (Luecken and Theis, 2019). Additionally, an unusually high number of genes and UMI count may correspond to multiplets (two or more cells tagged with the same cell barcode). The selection of appropriate values for the different QC parameters needs to be data driven. For instance, different cells may express different number of genes; a high percentage of mitochondrial genes may be expected for particular cell types, as it is proportional to the metabolic activity (Osorio and Cai, 2021). In any case, removing bad quality cells greatly improve the quality of data analysis (outlined in Figure 8) and prevent from drawing false conclusions, although the knowledge of cell type peculiarities is of very much help to decide on thresholds.

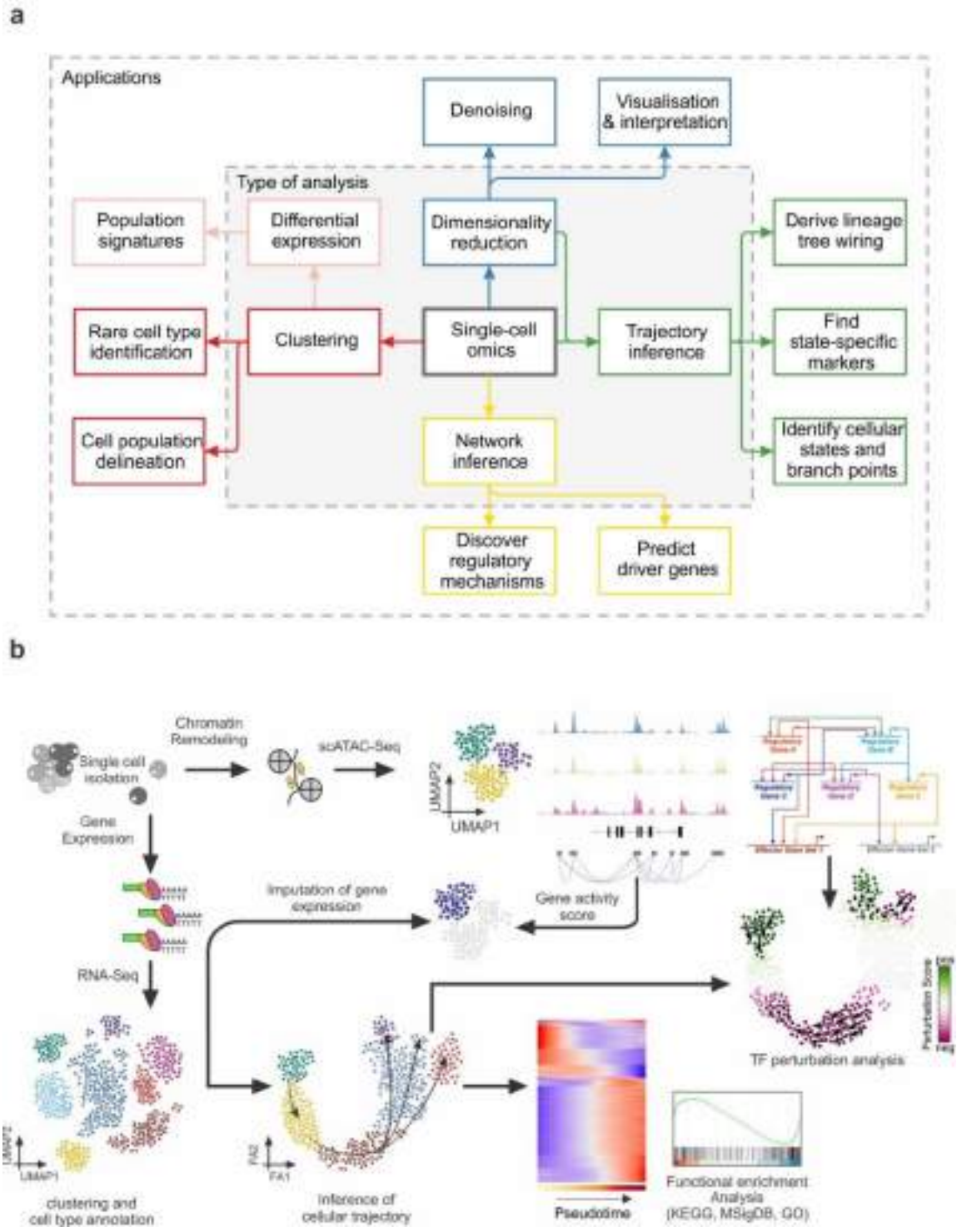


Figure 8 | single cell Omics data analysis to study cellular transitions. (a) General outline of single cell Omics data analysis and its application. **(b)** Schematic representations of the major data analysis strategies to study the cellular transitions such as EMT (Adopted and modified from multiple public sources).

1.6.3 Data Normalisation

scRNA-Seq data have significant cell-cell variation due to technical noise along with the biological confounding factors (Hafemeister and Satija, 2019). To preserve the biological differences while removing the technical noise, the data needs to be normalised. The log normalisation method with fixed size factor (usually 1e6) and pseudo-count (generally used for handling zeros in log transformation) addition is still used as default option of normalisation in popular scRNA-Seq data analysis workflows such as Seurat (Hao et al., 2021; Stuart et al., 2019) and Scanpy (Wolf et al., 2018). Additionally, Scran calculates the size factor by selecting cells with similar library sizes and uses for the normalisation (Lun et al., 2016). The normalisation methods with fixed size factor assumes constant RNA content in all the cells in a dataset (Hafemeister and Satija, 2019). Nevertheless, to address this type of questions, specialised statistical methods have been developed. SCTransform uses generalised linear models to represent the variance stabilisation without having to use uniform scaling factor (Hafemeister and Satija, 2019).

scRNA-seq datasets usually contain up to 30000 genes. However, not all of them are informative. Due to the low sequencing depth, scRNA-Seq data are sparse and zero inflated. This mainly poses two concerns in the downstream data analysis: first, the computation time and load, and second, the uninformative source of variation. To tackle these issues, the high variable genes (HVGs) can be selected based on their information index (Luecken and Theis, 2019; Townes et al., 2019). The number of HVGs should be determined by the specific biological question and objectives of the study. Typically, 1000 to 5000 HVGs are sufficient to balance computational efficiency and effectively capture cellular heterogeneity (Luecken and Theis, 2019).

1.6.4 Sample integration

Due to drop in the sequencing cost and the wide range of application of scRNA-Seq technology, multi-condition and multi-sample comparison is increasingly common. However, it imposes a challenge to integrate samples from different conditions and batches. A batch effect represents unwanted technical variation and can arise from variations in sequencing depth, protocol, sample acquisition

and handling, reagents or media and/or sampling time, etc. (Luecken et al., 2022). As detecting and removing real batch effect without losing biological context is challenging, recently, several computational algorithms have been designed to effectively address this issue (Discussed in details by Luecken et al., 2022). The joint analysis of different samples or conditions helps to answer specific questions in an easier way. For instance, integrated space of multiple samples can improve resolution of cellular heterogeneity. Checking the cellular composition (variability in number of cells in different conditions) can be misleading when performed in separate space, therefore integration needs to be done to address such issues. Additionally, the joint space will provide an option to compare gene expression between different conditions. Furthermore, data integration can be done on different modalities (scRNA-Seq and scATAC-Seq) which allows us to integrate information obtained at different levels (Stuart et al., 2019) to better understand the corresponding biological question.

1.6.5 Dimensionality reduction

scRNA-Seq assay produces high dimensional data from which we can capture thousands of genes in thousands or millions of cells, which indicates that the scRNA-Seq data suffers “curse of dimensionality”, as higher dimensional data often contains more noise and redundancy. This means, in that case, adding more information is not beneficial for the downstream analysis. As described earlier, the feature level dimensions can be reduced by selecting top HVGs. To further reduce these dimensions, specific dimensionality reduction methods are required. There are two major classes of dimensionality reduction methods i.e., linear and non-linear. Linear dimensionality reductions methods such as PCA, diffusion maps that preserve the information about data variability and can be used as a quantitative representation of the data. Non-linear dimensionality reduction methods such as UMAP, t-SNE are others, where data are subjected to different manifolds and used for the visualization (Xiang et al., 2021).

1.6.6 Clustering and cell type annotation

The main aim of scRNA-Seq assay is to identify the underlying cellular heterogeneity in a biological sample based on the gene expression patterns, and this can be achieved by dividing the cells into separate groups of highly similar

cells, called clusters. Clustering is a class of unsupervised machine learning algorithm, where a similarity index is determined using distance matrices, either on reduced dimensional space or directly on a gene expression matrix (Luecken and Theis, 2019).

The clusters obtained need to be annotated to interpret the data. Although there are several methods that have been developed for automatic cluster annotation in scRNA-Seq data (Abdelaal et al., 2019; Ianevski et al., 2022), the straightforward and more robust way to annotate detected clusters is to use the unique expression of cell-type specific genes, which can be identified using a differential gene expression approach by comparing each cell type with rest. This will provide a quantitative approach to derive cell type-specific markers. Additionally, gene signatures can be used for cluster annotation across different modalities. For instance, in the absence of multi-omics (sequencing different modalities in the same cells), annotating scATAC-Seq clusters is very difficult due to the lack of expression profiles. In such scenario, scRNA-Seq-based gene signatures derived from similar tissue can be used to annotate the clusters in scATAC-Seq data.

1.6.7 Trajectory inference and pseudotime analysis

In scRNA-Seq assay there is no spatiotemporal information due to the required cell segregation step (Trapnell et al., 2014). Thus, cellular heterogeneity explained by a discrete classification system is not sufficient to understand the continuous biological processes such as EMT or cell differentiation (Tanay and Regev, 2017). To capture the changes over such transitions sophisticated methods are required to model the dynamics of gene expression and are called trajectory inference methods. More than 70 trajectory inference tools have been already published (Saelens et al., 2019), and the basis of trajectory inference for all of them is to interpret scRNA-Seq data as snapshots of a continuous process, which helps to determine the initial and terminal points and calculate pseudotime (Figure 8), allowing to decipher the pseudo-transition of the cells. The pseudotime can be further used to better understand the molecular changes in a transition by modelling gene expression in a continuous manner.

1.6.8 scRNA-Seq based regulon prediction and *in silico* perturbation analysis

To understand the regulation of gene expression in a biological system, an additional layer of information is need. The regulation of gene expression by TFs has a direct impact on expression modulation. scRNA-Seq data hold the information of gene expression patterns from thousands or millions of samples, making it more reliable to derive statistically significant gene regulatory modules, called gene regulatory networks (GRNs) or regulons. The basis of expression based GRNs prediction is a co-expression analysis of TF and target genes, plus binding site prediction in promotor regions of target genes motifs searching (Aibar et al., 2017). These GRNs can be used to infer the regulatory impact of a TF in a specific cluster. Additionally, combining the regulatory activity with pseudotime helps to identify the dynamics of TF regulation in a cell state transition.

Recently, the advance in single-cell protocols allow a simultaneous coupling of single cell assays with the dynamic conditions generated by genetic perturbation such as those achieved with CRISPR/cas approaches to delete specific genes (Frangieh et al., 2021; Papalexi et al., 2021; Replogle et al., 2022). Experimental approaches for gene deletions in cells or animal models are robust and powerful, but also time consuming and very costly. Recently, implemented methods for *in silico* perturbation using scRNA-Seq data have become a valuable strategy to have a first readout of putative outcomes. CellOracle uses advanced machine learning algorithms to simulate the effect of perturbations on TFs or cell state transitions (Kamimoto et al., 2023). The results can be further used both to improve experimental designs and for validation purposes.

1.7 Single-cell ATAC Sequencing (scATAC-Seq)

As explained earlier, scRNA-Seq assay is useful to understand gene expression changes at single cell level. Similarly, profiling chromatin accessibility is a powerful method to understand chromatin remodelling and epigenetic changes. This technique uses the Tn5-transposase enzyme to tag and cut open chromatin regions, resulting in nucleosome-free DNA fragments (Grandi et al., 2022). Unlike scRNA-Seq, sequences are obtained not only from coding regions but from the open chromatin regions of the entire genome. The protocol for scATAC-Seq

library preparation for also differs from that used for scRNA-Seq libraries. Instead of single cell, scATAC-Seq protocol requires isolated single nuclei. Generally, the open chromatin regions are functionally active, providing space for regulatory proteins to bind and operate. Systematic analysis of ATAC-seq data can reveal important regulatory elements and transcription factors, offering an additional layer of information that can be linked to gene expression to better understand regulatory mechanisms. As EMT has a complex regulatory network (Figure 3), understanding chromatin remodelling throughout the process will help to decipher the regulatory mechanisms, and can lead to the identification of novel regulators.

The analysis of scATAC-Seq data is similar to that of scRNA-Seq with slight variation in the workflow. The first difference is that in scATAC-Seq modality the genome is sequenced rather than the transcripts. This makes scATAC-Seq data more sparse than scRNA-Seq data because the quantification values range from 0-2 (0: for closed chromatin region, 1: when one allele is open, 2: when both alleles are open), imposing challenges in downstream analysis. scATAC-Seq assays capture genomic regions and detects what are called peaks or bins. Peaks are the genomic regions derived by stitching adjacent open regions together using statistical frameworks (Zhang et al., 2008), which provides a robust quantification of chromatin accessibility.

The quality control and filtering of scATAC-Seq data differs from scRNA-Seq data, being nucleosome signal and transcription start sites (TSS) enrichment the QC parameters in the former.

Unlike gene expression, chromatin accessibility does not directly infer functional or phenotypic consequences in the cell. Therefore, it is important to associate the open chromatin regions to the corresponding genes, information that can be further used to infer the association between chromatin remodelling and gene expression. This can be done by assigning the peaks to the adjacent genes based on the distance from TSS (usually, 2kb), which allows to quantify changes in chromatin accessibility, providing so called “gene activity score” (Stuart et al., 2021). Additionally, advanced algorithms such as Cicero can be used to infer the cis-regulatory connections between open chromatin regions and target genes (Pliner et al., 2018). These connections can be used to understand the transcriptional regulation by TF by performing motif enrichment analysis, motif

foot printing or correlation analysis between TF and target genes. Additionally, the TF-targets gene list derived using scATAC-Seq is suitable to run downstream analysis such as GNR prediction and *in silico* perturbation analysis (Kamimoto et al., 2023).

Chapter 2

OBJECTIVES

Hypothesis and Objectives

The activation and reactivation of the EMT programme in various physiological and pathological conditions is pivotal to understand central biological processes ranging from embryonic pattern formation to tissue degeneration, regeneration and tumour progression towards the metastatic disease. Such different contexts impinge into the underlying molecular mechanisms driving EMT. As such, the nature and degree of EMT activation in different cell contexts add a significant amount of complexity to the process. Although EMT has been studied for decades and Prof. Nieto´s lab pioneered these studies and has contributed to its understanding along the years, it is only with the advent of new technologies, and in particular with the possibility of analysing whole genome transcriptomes in thousands of individual cells that we are now in a position to propose an unbiased, comprehensive, and comparative approach of the EMT programmes activated in different contexts they have already studied. With this as a background and the opportunity to perform state-of-the-art functional experiments in the lab, my challenging and fascinating objective as a bioinformatician has been to analyse massive transcriptomics data to improve our understanding on EMT activation, its progression and association with disease phenotypes. This, in turn, can lead in the future to the development of better therapeutic strategies to tackle devastating diseases such as cancer and fibrosis. With this notation we proposed,

General Objective:

To understand the implementation of EMT programmes, the commonalities and specificities in embryonic development, organ fibrosis and cancer progression.

Specific Objectives:

1. To define the EMT programme during Neural Crest Development
2. To define the EMT programme during Kidney Fibrosis
3. To define the EMT programme in Breast Cancer
4. Challenging EMT trajectories during breast cancer progression

Chapter 3

MATERIALS AND METHODS

3.1 In silico analysis of human cancer cell lines

Klijn Breast Cancer cell lines gene expression data was analysed for epithelial (E) and mesenchymal (M) component enrichment (Klijn et al., 2015). E and M component were obtained from merging E and M signatures in Tan et al., 2014 and Taube et al., 2010. Note that the EMT-TFs were removed from the Mesenchymal signature to avoid to biased correlations in the subsequent analyses. Singscore (Foroutan et al., 2018) was used to compute enrichment scores for E and M components. E and M enrichment values were plotted (X-axis: M score and Y-axis: E score) and K-means clustering used to partition the Breast Cancer cell lines According to the optimal number of clusters calculation, K=3.

3.2 Bulk RNA sequencing and data analysis

3.2.1 Library preparation and sequencing

RNA was extracted using illustra RNAspin Mini isolation kit from three biological replicates per condition. RNA quality check, mRNA libraries preparation (stranded) and paired-end reads (75 pb length) sequencing using Illumina HiSeq4000 platform were performed at the CNAG-CRG facility in Barcelona, Spain.

3.2.2 Data analysis

Reads were aligned to CanFam3.1 genome annotation (Ensembl v97) using STAR v2.5.3a (Dobin et al., 2013). Quality control of sequenced reads was performed using FastQC (Babraham Institute) and gene expression was quantified using RSEM v1.3.0 (Li and Dewey, 2011). We used enrichR v2.1, a R package to access the Enrichr database (Kuleshov et al., 2016) and performed general functional enrichment analysis, while the gseGO and gseKEGG functions from clusterProfiler v3.10.0 (Yu et al., 2012) were used to carry out Gene Set Enrichment Analysis (GSEA) of Gene Ontology terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2019) and Kyoto Encyclopedia of Genes and Genomes pathways (Kanehisa and Goto, 2000). The R package msigdbr v7.0.1 was used to obtain gene sets from the Molecular Signatures Database (MSigDB

v7.0; Subramanian et al., 2005) from Broad Institute. The R package GOSemSim v2.8.0 (Yu et al., 2010) was used to filter GO terms by semantic similarity, and ggplot2 v3.3.0, and enrichplot v1.6.1 were used to visualize functional enrichment results.

3.3 Animal experiments

Mice were fed *ad libitum*. Housing and experimental procedures were conducted in strict compliance with the European Community Council Directive (89/609/EEC) and the Spanish legislation. Ethical protocols were approved by the CSIC Ethical Committee and the Animal Welfare Committee of the Institute of Neurosciences. Animals for experiments were selected by genotype, and no randomization or blinding was performed. Animals were housed under SPF conditions at the RMG animal House (ES-119-002001 SEARMG).

3.3.1 Kidney fibrosis model

To genetically label renal tubular (RT) epithelial cells, we generated a mouse line with the Rosa-LSL-tdTomato reporter line Ai9/RCL-tdT (Madisen et al., 2010; kindly provided by Oscar Marin, King's College London), activated in RT cells by a Cre recombinase under the control of the kidney-specific promoter *Ksp1.3 (*Shao et al., 2002; kindly provided by Peter Igarashi, University of Minnesota). Mice were maintained in C57BL/6 background.

Male and female mice (8-12 weeks-old) were anesthetized by inhalation of Isoflurane/O2 gas mixture (5% 1L/min) in an induction chamber. Further, analgesic *Buprenorphine* (0.1mg/Kg) was administered in a single *subcutaneous injection and mice were maintained* anesthetized by inhalation of Isoflurane/O2 gas mixture (1-2% 1L/min*)* through a breathing circuit attached to a nose cone*.* Further, the abdomen was opened, and the left ureter was ligated with 6/0 non-absorbable braided silk suture thread (Lorca Marin). The abdomen was then closed with running sutures and the skin was closed with interrupted sutures. After surgery, the mice were maintained in a temperature-controlled room with a 12 hours light/dark cycle, and were reared on standard chow soaked in Buprenorphine diluted at (0.03mg/ml) in GlucoSaline (B.Braun, Germany) during

the 24h post-surgery and water *ad libitum*. Unilateral ureteral obstruction (UUO) was maintained for 1, 2 or 3 weeks.

3.3.2 Breast cancer model

Mouse experiments were carried out in MMTV-PYMT model (Guy et al., 1992) crossed with a Rosa-LSL-tdTomato reporter line (Madisen et al., 2010), expressing tdTomato upon Cre-mediated recombination. Cre recombinase is expressed under the control of *Keratin14* promoter {Tg(KRT14-cre)1Amc/J-STOCK 004782} (Dassule et al., 2000), the latter purchased from JAX MICE (The Jackson Laboratory). Mice were backcrossed in FVB background for at least 10 generations (99.9% FVB).

3.4 2D cell culture

MDCK-NBL2 and MDCK-II dog epithelial kidney cell lines were purchased from ATCC (American Type Culture Collection) and Sigma (European Collection of Authenticated Cell Culture), respectively. MDCK-NBL2 and MDCK-II cells were cultured in DMEM (Sigma) supplemented with 10% heat inactivated foetal bovine serum (FBS) (Sigma), 1% Gentamicin (Sigma) and 1% Amphotericin (Sigma). Cells were grown at 37ºC and 5% CO_2, and the media was replaced every two or three days. Cells were passaged when they reached 80-90% confluency. Cells were passed up to a maximum of 8 times.

3.5 TGFβ administration

A stock solution of human recombinant TGFβ (rH-TGFβ1) (MERQ) (SHENANDOAH) was prepared at 2μg/ml in filtered H2O supplemented with 1% BSA and 10 mM HCl, aliquoted and stored at -80ºC. All treatments in 2D cultures (5 ng/ml) started 24h after seeding cells (104 in 6 well-plates or 75x104 in 10cm culture dishes) and the medium containing TGFβ was replaced every 48h. Cells were never seeded from high confluency cultures to avoid a reduction in the response to TGFβ.

3.6 Primary tumour derived tumouroids and invasion assay

Primary tumour tumouroids were prepared and embedded in 3D collagen gel following a protocol modified from (Kevin J. Cheung et al., 2013). In summary, mammary gland carcinomas were collected from 14 weeks old female mice and first minced manually using sterile scalpels and further finely cut with a McIlwain Tissue Chopper (TED PELLA, INC). Minced samples were incubated in 2.5 ml of digestion buffer (DPBS supplemented with 2.5 Wünsch units of TH Liberase/ml and 25 µg/ml DNase I (Roche)) at 37ºC for 20 min in an Incubator Microplate Shaker (VWR). Tumour fragments were pipetted up and down every 5 min. Disaggregated samples were neutralized with 15 ml of breast medium containing DMEM:Nutrient mixture Ham F12 (N6658 Sigma) (1:1) supplemented with 10% inactivated FBS, Insulin (10µg/ml), 1% Gentamicin (Sigma) and 1% Amphotericin (Sigma) supplemented with 25 µg/ml DNaseI and spun down at low speed (30G) for 5 min. Pellets were resuspended in 5ml of breast medium and spun down again at 30G for 3 min and the supernatant carefully discarded. The rounds of spinning and resuspension were repeated 5 times before resuspending tumour fragments in Collagen gel containing Rat Collagen Solution type I (Corning) (2.5%) in DMEM (1X), $NaHCO_3$ (0.23%) and HEPES (0.1M) prepared on ice at pH 7.0-7.5. A volume of 100µl of tumour fragments and collagen mixture were added on top of previously solidified cell-free Collagen gel plated in 48-well plates and incubated at 37ºC and 5% CO_2. Breast medium was gently added after 2 h and replaced every 48h.

To analyze the cultures, tumouroid cultures were washed twice with PBS, and then fixed with PFA for 60 min at RT. Fixed organoids were washed 3X 30 min in PBS and blocked/permeabilized for 4h at RT with IF blocking buffer (IFBB+: 5% normal Goat Serum, 1% Bovine Serum Albumin, 1% TritonX-100 and 0.1% Sodium Azide in sterile PBS). Blocking solution was substituted by the primary antibody diluted in IFBB+ and incubated o/n at RT on a rocker plate. Tumouroids were washed 3x for 30 min in PBST (PBS with 1% TritonX-100) and incubated for 24 h with secondary antibodies and 4′-6-diamidino-2-phenylindole (DAPI). Finally, tumouroids were washed 3x for 60 min in PBST and mounted on Glass Bottom Microwell Dishes (MatTek) using anti-fade mounting medium (DAKO). Primary and secondary antibodies used are listed in Table 1. Tumouroids were

photographed using Leica SPEII confocal and acquired images were analyzed with ImageJ and Adobe Photoshop CS6 software programs.

3.7 Immunofluorescence (IF)

3.7.1 Cells in culture

MDCK cells were grown on coverslips in 6 well-plates, under the culture and treatment conditions described above. Cells were rapidly washed twice with PBS and fixed with PFA for 15min at RT. Cells were rinsed with PBS for at least three rounds of 10 min each and directly used for IF or stored at 4ºC in PBS+Azide 0.02% for less than one week. For IF staining, coverslips were deposited in a humidified chamber and blocked/permeabilized for 1h with IF blocking buffer (IFBB: 5% normal Goat Serum, 1% Bovine Serum Albumin and 0.2% TritonX-100 in sterile PBS). Blocking solution was substituted by the primary antibody diluted in cold IFBB and the staining chamber incubated o/n at 4ºC. Coverslips were washed 3x for 10 min in PBS and incubated for 1h with secondary antibodies and 4′-6-diamidino-2-phenylindole (DAPI). Finally, coverslips were washed 3x for 10 min in PBS and mounted on glass slides using anti-fade mounting medium (DAKO). Primary and secondary antibodies used are listed in Table 1. Cells were photographed using Leica SPEII confocal, Leica DMR or Zeiss Axio microscopes. Acquired images were analysed with ImageJ and Adobe Photoshop CS6 software programs.

3.7.2 Kidney and tumour samples

Paraffin-embedded sections were dewaxed, and protein epitopes unmasked by immersion in 95ºC preheated Citrate (pH6.0) or Tris-EDTA (pH9.0) buffer for 20 min. O.C.T or unmasked paraffin sections were washed three times in PBS for 5 min and subjected to the IF protocol described above for cell lines. For information on primary and secondary antibodies see Table 1. Pictures were acquired and analysed as described for cells in culture.

Antibodies	Source	Concentration
Primary antibodies		
IF: CDH1	BD Biosciences 610181	1:100 (cells) / 1:500 (tissue)
IF:Vimentin (Rabbit)	Abcam ab92547	1 :500
IF: ZO-1(TJP-1)	Life Technologies 339100	1 :100
IF: FN1	Abcam ab2413	1 :200C8
IF: KERATIN14	Palex 454928	1 :5000
IF: JUN	Cell Signaling mAB#9165	1 :1000
Secondary antibodies		
IF: Goat anti-Rat IgG(H+L) Alexa Fluor 488	Invitrogen A-11006	1 :500
IF: Goat anti-Rabbit IgG(H+L) Alexa Fluor 568D17	Invitrogen A-11011	1 :500
IF: Goat anti-Mouse IgG(H+L) Alexa Fluor 568	Invitrogen A-11004	1 :500
IF: Goat anti-Chicken IgG (H+L) Alexa Fluor 568	Invitrogen A-11041	1 :500
IF: Goat anti-Rabbit IgG (H+L) Alexa Fluor 647	Invitrogen A-21244	1 :500

Table 1: Antibodies used for immunofluorescence (IF)

3.8 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 tissues 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 37ºC 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) 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 RCF at RT) and resuspended in Dead cell removal beads (MiltenyiBiotec) in 100µl per 10^7 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 RCF at 4ºC, 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 RCF and 4ºC), 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.

3.9 Single-cell GEM and cDNA library preparation

3.9.1 Kidney samples

Three single-cell preparations were obtained from 1 SHAM and 2 UUO kidneys from 3 male mice. 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.

3.9.2 Tumour samples

Four single-cell preparations were obtained from four independent mammary gland carcinoma samples dissected from 3 female mice. 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.

3.10 Kidney single-cell RNA-Seq 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 (Babraham Institute). The reads were aligned to the mouse genome assembly GRCm38 (mm10; ensembl reference annotation) and quantified using 10X Genomics CellRanger v5.0.0 (Zheng et al., 2017) pipeline with default parameters.

3.10.1 Cell quality control, filtering, and integration process

The CellRanger output was imported in the R v3.6.1 (https://www.r-project.org/) statistical environment and converted to a Seurat v3.2.2 (Stuart et al., 2019) 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; UUO#1: 22172; UUO#2: 20159). Confident mapping to exonic regions was above 50% for each library; median unique counts per cell were as follows: SHAM: 3750; UUO#1: 3124; UUO#2: 2616; and median value for detected genes per cell was SHAM: 1550; UUO#1: 1734; UUO#2: 1431. The sample-specific putative doublets were predicted using Scrublet v0.2.3 (Wolock et al., 2019) and filtered out from the subsequent analysis. The expected doublet rate threshold (SHAM = 0.12; UUO#1 = 0.10; UUO#2 = 0.11) was calculated based on the total recovered cells and the doublet detection score was set to SHAM = 0.35, UUO#1 = 0.25, UUO#2 = 0.27 by manually inspecting the bimodal distribution. Next, we filtered out low-quality cells expressing high levels of mitochondrial markers (>10%; Osorio and Cai, 2021). 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 normalised 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 normalisation method) and datasets were integrated based on the precomputed anchorset using IntegrateData function.

3.10.2 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.

3.10.3 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 log2FC 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. Cell types were assigned to clusters using *bona fide* gene markers and grouped into 6 major cell populations. The expression of cell type specific selected markers was represented as a dot plot using DotPlot function implemented in Seurat.

3.10.4 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 showing even distribution in SHAM and UUO was set as a reference cell type. Compositional data analysis (CoDA) score was represented as a boxplot along with associated P-value.

3.10.5 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; Gewin, 2018), in principle, we can't exclude the possibility of contribution from other epithelial populations.

Based on the superior performance of deep learning models for assigning cell types in scRNA-Seq data (Alquicira-Hernandez et al., 2019; 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 v1.1.1 python library). To build the MLP training model, we subset the cells belonging to the epithelial clusters (8, 5, 18, 16, 15, 2, 21, 11, 23). 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). 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	Negative
	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

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.

3.10.6 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). 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.

3.10.7 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-Normalised 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., 2020 were used to calculate the injury/inflammation score. Calculated scores were visualized over UMAP using FeaturePlot function.

3.10.8 Trajectories inference using PAGA and RNA-Velocity

3.10.8.1 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 v1.6.0 (Wolf et al., 2018) python package. The partition-based graph abstraction (PAGA; Wolf et al., 2019) algorithm implemented in Scanpy was used to build the connectivity map.

3.10.8.2 RNA Velocity Analysis

Run10x command line utility from velocyto v0.17.17 (La Manno et al., 2018) 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 v3.0.6 package. To infer the directionality of the transcriptional changes in the proximal tubule and injured cells, 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 normalised 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.

3.10.9 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).

3.10.10 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 v0.11.0 (Aibar et al., 2017). 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 pySCENIC package and average AUC score was represented as heatmap. 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 with splines of degree=5. The enriched motifs for each TF were reported based on their highest NES.

3.10.11 Trajectory-based differential expression analysis

Integrated Seurat object for proximal tubule and injured trajectory was transferred to Monocle3 v0.2.2 (Cao et al., 2019) 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; Qiu et al., 2017) 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 log2FC) were retained. The min-max normalised expression for the filtered genes was used to generate the expression heatmap (scv.pl.heatmap function implemented in scVelo v0.2.2 python package; Bergen et al., 2020). The pathway enrichment analysis for the differentially expressed

genes over trajectory was performed using R based API of EnrichR v3.0 (Chen et al., 2013; Kuleshov et al., 2016) 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 v3.3.3 R package.

3.11 PyMT tumours single-cell RNA-Seq 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 CellRanger v2.2.0 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* transgene (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.

3.11.1 Cell quality control, filtering, and integration process

Downstream analyses, such as quality control, integration, normalisation, shared nearest neighbour graph-based clustering, differential expression analysis and visualization were performed using the R package Seurat v3.1.4 within R Statistical Computing Platform v3.6.3. 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 (Osorio and Cai, 2021). 34607 cells passing quality control were further analysed. UMI count data from each sample was normalised 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 normalisation method) and finally, datasets were integrated using the pre-computed anchorset using IntegrateData function.

3.11.2 Dimensionality reduction and cluster detection

PCA dimensionality reduction was performed using the Seurat function RunPCA with default parameters. A SNN Graph for the integrated dataset was built using FindNeighbors with top 30 PCs, and cell clusters were identified by a 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. 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.

3.11.3 Differential gene expression testing and clusters

For expression plots and differentially expressed gene testing, cell expression was Log-Normalised using NormaliseData 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 log2FC over Log-Normalised expression for the 5 major populations. A total of 3903 enriched and statistically significant genes (positive log2FC and adj p-value < 0.1) were sorted by descending average log2FC. A proportional down sampling 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-Normalised expression.

3.11.4 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. The Markov Affinity-based Graph Imputation of Cells (MAGIC v3.0.0; Van Dijk et al., 2018) was applied to impute the expression of EMT-TFs encoding genes with default parameters (t=3).

3.11.5 Trajectories inference using PAGA and RNA-Velocity

3.11.5.1 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 neighbors and then the coarse-grained connectivity map (resolution=0.3) was build using sc.tl.paga function.

3.11.5.2 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 tunning as described in section “RNA-Velocity Analysis” for kidney scRNA-Seq data. Briefly, the initial gene filtering was performed using the velocyto 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 normalised for the total number of cells using _normalise_S and _normalise_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 v2.2.0 (Street et al., 2018) 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 simultaneously fitting 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.

3.11.6 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.

3.11.7 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 Normalised Enrichment Score (NES) below 1.75 were filtered out. The single cell level regulon activity was calculated using AUCell algorithm implemented in pySCENIC package. The regulon activity was averaged by clusters and represented as a heatmap. 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. The enriched motifs for each TF were reported based on their highest NES.

3.11.8 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 log2FC). 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.

3.11.9 EMT analysis in breast cancer patients

To examine the putative enrichment of gene expression signatures found in each cluster of EMT-T1 and EMT-T2 trajectories across patient breast cancer subtypes, we used the GSVA v1.34.0 R package (Hänzelmann et al., 2013) to perform Gene Set Variation Analysis in breast cancer expression data obtained from Chung et al., 2017 (GEO GSE75688). R packages dplyr v1.0.3, magrittr v2.0.1 and tibble v3.1.2 were used to transform gene expression data to the required GSVA input format, and ggpubr v0.4.0 was used to generate the GSVA enrichment score boxplots.

Figure 10 | Workflow summarising important steps in scRNA-Seq data analysis in kidney fibrosis and during breast cancer progression

3.12 *In silico* Perturbation analysis

3.12.1 scATAC-Seq library prepration and sequencing

Mammary gland carcinomas were collected from 14 to 15 weeks old female mice. Harvest tissues were first minced manually using sterile scalpels and finely cut with McIlwain Tissue Chopper (TED PELLA, INC). Minced samples were transferred to a tissue homogeneizer douncer prefilled with Solution D (20mM Tris-HCL ph7.5, 0.1% Tween20, 0.25M Sucrose, 25mM KCL, 5mM MgCl2) and homogenized 8-10 times, rotating the douncer in the process. Resulting lysate was subsequently passed through a 70µm and 40µm filters (BD Falcon) and transferred to a 15ml tube. Following a 5 min centrifugation at 500rcf at 4ºC, supernatant was removed, and the pellet was resuspended in 4ml of Solution D. 2ml of Optiprep (STEMCELL) were added and the sample was centrifugated for 10mins at 1500rcf 4ºC. Pellet was resuspended in Wash Buffer (10Mm Tris-HCL ph7.5, 10Mm NaCl, 3mM MgCl2, 2% BSA, 0.1% Tween20, 1mM DTT) and passed through a 40µm filter. Sample was then transferred to a 1.5ml tube and centrifugated for 5 mins at 500 rcf 4ºC. Nucleis were resuspended in 200ul of 0.1X lysis buffer (10Mm Tris-HCL ph7.5, 10Mm NaCl, 3mM MgCl2, 2% BSA, 0.1% Tween20, 1mM DTT, 0.1% Nonidet P40, 0.01% Digitonin) and incubated on ice 5 minutes. After the incubation, 1.3 ml of Wash Buffer were added, and a 20µm filter passed the nuclei preparation. Following a 5 min centrifugation at 500rcf at 4ºC, nucleis were resuspended on 100ul of 1X Nuclei Diluted Buffer (Chromium Next GEM Single Cell ATAC Kit v2). 10ul were mixed with an equal volume of Trypan Blue and nuclei number and viability were determined with the automated cell counter Countess II (Invitrogen).

Following Chromium Next GEM Single Cell ATAC Kit v2 guidelines we prepared 2 suspensions of single nucleus from 2 different female mice, each prepared as a mix of breast cancer tumors from the same mice. The nucleus suspensions were transposed followed by individual nucleus encapsulation using 10x Chromium Controller (10x Genomics). Finally, the samples were multiplexed using Chromium Single Index Kit N. The prepared libraries were sequenced on Illumina NextSeq 2000 machine with 50bp read length.

3.12.2 PyMT tumours single-cell ATAC-Seq data analysis

Libraries were sequenced to obtain more than 517 million reads in total (T1: 263.84M; T2: 253.80M). Quality of the sequenced reads was assessed using FastQC v0.11.9 (Babraham Institute). Sequenced samples were processed using the CellRanger-atac v2.1.0 pipeline (10x Genomics) and aligned to the GRCm38 (mm10) mouse reference genome (Ensembl annotation v99). MACS2 v2.2.9 (Zhang et al., 2008) was used for the peak calling with default parameters. We detected 10030 (T1: 5400; T2: 4630) total number of cells with 221736 total no-overlapping peaks (T1: 207349; T2: 207769).

3.12.2.1 Quality control, integration, dimensionality reduction and clustering

Downstream analyses, such as quality control, filtering, integration, normalisation, clustering and visualization was performed using the R package Seurat v4.1.3 (Hao et al., 2021) and Signac v1.12.0 (Stuart et al., 2021) within R Statistical Computing Platform v3.6.3. The appropriate QC parameters were calculated for each sample separately as per the standard Signac guidelines. The sample-specific putative doublets were predicted using AMULET v1.1 (Thibodeau et al., 2021) and filtered out from the subsequent analysis. We retained only high-quality cells (n = 7723) based on number of fragments falling in peak regions (3500-35000), percentage of reads in peaks (>30%), and TSS enrichment (<10). The selected threshold was derived by manually inspecting the data distribution. To pull different samples in a single space we followed integration workflow using high quality cells. Briefly, first the two WT samples were merged using merge function provided by Seurat and then Term-Frequency Inverse-Document-Frequency (TFIDF) was calculated for the merged object. FindTopFeatures was used to select the top features with min.cutoff = 20. We performed Singular Value Decomposition (SVD) on the top selected features. This approach helps to preserve the variability in the data while reducing the feature space. Harmony (Korsunsky et al., 2019) was used to perform the integration in SVD space. The first component was not considered for any downstream analysis as it is highly correlated with the sequencing depth in scATAC-Seq data (Stuart et al., 2021). A SNN Graph was built based on the harmony components using FindNeighbors function with top 30 components, and

cells were clustered by a SNN modularity optimization-based clustering algorithm implemented in FindClusters function with resolution 0.8, this resulted in 14 major clusters. The reason of using unusually higher resolution was to separate a small number non cancer cells from the cancer cell clusters (Figure 34). The relative chromatin accessibility for all detected peaks were visualized using ComplexHeatmap R package (Gu, 2022; Gu et al., 2016).

3.12.2.2 Imputation of gene expression and cluster annotation

It is extremely difficult to associate genomic information directly to the function of the cells, therefore making it difficult to annotate the clusters in scATAC-Seq assay. To address this issue, we used label transfer approach to impute the gene expression for each single cell which was used for the cluster annotation. Briefly, first we calculated a gene activity score for high variable genes detected in our scRNA-Seq data by counting the number of fragments within 2kb upstream region of the gene. Then the canonical correlation analysis (CCA) based anchor were detected between gene expression of high variable genes from scRAN-Seq and gene activity score from scATAC-Seq assay using FindTransferAnchors function. We set scRNA-Seq data as a reference and gene activity as a query. Finally, the TransferData function was used to assign the continuous gene expression values to individual cells. The imputed gene expression values then plotted over UMAP using FeaturePlot function to visualize the specificity of markers. The genomic view of open chromatin region around selected markers was visualized using CoveragePlot function implemented in Signac package.

3.12.2.3 Cancer cell subset and gene signature enrichment

To explore cancer cell heterogeneity at the open chromatin level, we subset cancer cells based on imputed tdTomato expression (also supported by other cancer cell markers). We re-integrated the data as explained in section “Quality control, integration, dimensionality reduction and clustering” of scATAC-Seq data analysis. Then the neighbour searching was performed based on harmony components and the FindClusters functions was used to cluster the cancer cells into 14 distinct clusters (resolution = 1.4). Next, to understand the dynamics of EMT implementation and its correlation with the chromatin accessibility we used enrichment-based approach. First, we selected the top20 differential genes from

EMT trajectory clusters in our scRNA-Seq data, representing the specific set of gene for each trajectory cluster. Then the AddModuleScore function was used to calculate the score for each cluster using imputed gene expression values for top20 selected genes, to check the enrichment in our scATAC-Seq data. The enrichment score then represented as a heatmap to show the corresponding scRNA-Seq based EMT trajectory clusters in scATAC-Seq data using ComplexHeatmap R package. Additionally, the enrichment score was visualized at single cell level using FeaturePlot function. Similarly, the EMT hallmark, Inflammation and BC-PING score was calculated and represented over UMAP.

3.12.3 CellOracle simulations for the EMT-TFs perturbations

The analysed cancer cell subset was used to build the base gene regulatory network which was used in the TF perturbations. Briefly, first the detected peaks in our scATAC-Seq data were imported into python environment. The peaks were annotated to the nearest genes based on the distance form TSS region. The annotated peaks then searched for the TF motifs using TFinfo function from CellOracle v0.16.0 (Kamimoto et al., 2023) with default parameters. The Seurat object for EMT trajectory was exported to python compatible Scanpy object. We used cluster information and embedding calculated using Seurat. The CellOracle uses the same strategy as velocyto for visualizing cell state transitions. Therefore, first we performed PCA and selected top 50 PC. The KNN imputation was performed using knn_imputation function from CellOracle with 500 nearest neighbours. Next, the expression of selected EMT-TFs was set to 0.0 and run the simulations using simulate_shift function with n_propagation=3. The transition probability was estimated using estimate_transition_prob function and shift embedding was calculated using calculate_embedding_shift function. For the vector representation we used grid representation approach with n_grid=40, min_mass=12 and scale_simulation=1.5. To compare the amount of shift in the cell state from WT trajectories compared to the TF perturbations we calculated perturbation score as described in Kamimoto et al., 2023. We used the RNA velocity vectors as a reference directionality (unperturbed) of cells in our trajectory. Then the inner product was calculated between reference vectors and the TF perturbation vectors. The positive perturbation score indicates the promotion of cell state transition in the trajectory whereas the negative

perturbation score indicates the block of cell state transition in the reference trajectory.

3.13 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 was obtained from http://pklab.med.harvard.edu/ruslan/neural_crest. The Seurat object was created using expression matrix and cell metadata. The connectivity map of mouse trunk NC cell population 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. We performed PCA as described above and used pre-calculated t-SNE 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 t-SNE embedding.

Chapter 4

RESULTS

4.1 Epithelial cells activate EMT to transition towards the mesenchymal phenotype

4.1.1 Breast cancer cell lines can be distributed along the epithelial to mesenchymal spectrum

Cell lines are proven to be an excellent model to analyse phenotypes and explore the changes associated with the activation of different programmes. They are amenable to multiple studies, cost-effective and provide sufficient reproducibility (Sharma et al., 2010). Therefore, to understand the dynamics of phenotypic changes in the E-M spectrum, we analysed publicly available bulk transcriptomic data for 71 human breast cancer (BC) cells lines (Klijn et al., 2015).

The Gene set variation Analysis (GSVA) is an effective method for the analysis of single sample enrichment (Foroutan et al., 2018). The GSVA analysis of previously published epithelial (Pastushenko et al., 2018) and mesenchymal (Tan et al., 2014) gene signatures in human BC cell lines shows distinct pattern which grouped them into three major clusters, corresponding to epithelial (E), mesenchymal (M) and hybrid (E/M) phenotypes (Figure 10a). The expression analysis of epithelial and mesenchymal genes along with EMT score shows the dynamics of transcriptional changes over the E-M spectrum in the BC cell lines (Figure 10b). Additionally, we observed that the levels of expression of the EMT-TFs correlate with the transition from to the E to M phenotypes (Figure 10c).

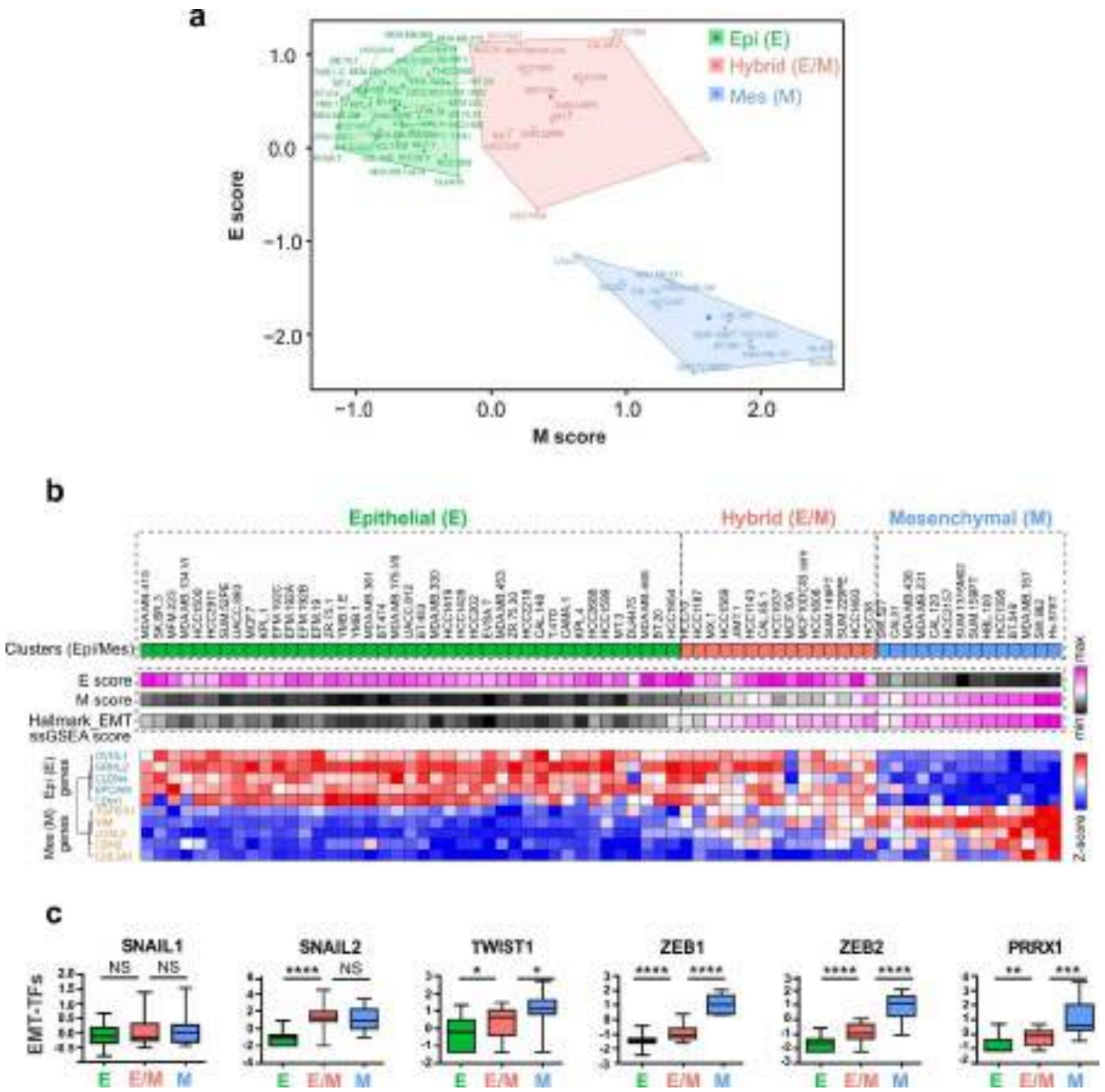


Figure 10 | Epithelial and mesenchymal gene signatures can distribute breast cancer cell lines along the EMT spectrum, identifying 3 main states: E, M and E/M. (a) Scatterplot showing clustering of 71 breast cancer cell lines (Klijn et al., 2015) based on enrichment score of epithelial and mesenchymal components. The cell lines were grouped into three different clusters using k-means clustering. **(b)** Upper panel shows epithelial (E), mesenchymal (M) and EMT hallmark (MsigDB Database) signatures enrichment score grouped the cell lines as per the clusters obtained in (a). Lower panel, relative expression of epithelial and mesenchymal markers. **(c)** Boxplots showing relative expression of EMT-TFs in different groups of cancer cell lines categorized based on epithelial and mesenchymal components.

4.1.2 Differential response of MDCK cell sublines to TGFβ treatment; construction of an invasive gene signature

The whole transcriptomic analysis of breast cancer cell lines allowed us to understand the transcriptional changes occurring over E-M spectrum. With this notion, we next decided to explore the EMT programmes activated in the two MDCK cell sublines, MDCK-II and MDCK-NBL2, responds differently to the TGFβ treatment (Figure 11a). The immunofluorescence analyses show co-occurrence of epithelial (ZO-1) and mesenchymal (FN1) markers in both MDCK cell lines after 2 days of TGFβ treatment suggesting that these cells activate partial EMT programmes. Upon continuous TGFβ treatment up to 4 days, the complete loss of epithelial and significant gain of mesenchymal markers in only NBL2 cell line suggest the activation of a full EMT programme (Figure 11b).

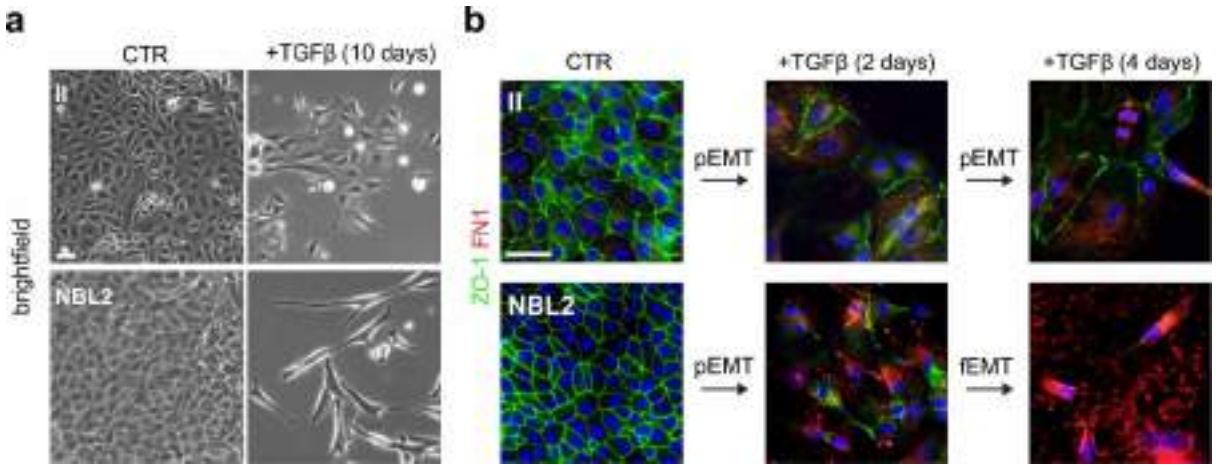


Figure 11 | MDCK cell sublines undergo EMT after treatment with TGFβ. (a) Brightfield image showing the morphology of MDCK-II and MDCK-NBL2 cells in control conditions and after 10 days of treatment with TGFβ. **(b)** Immunofluorescence analysis showing expression of the epithelial marker tight junction protein (ZO-1) and the mesenchymal marker fibronectin (FN1) at different times after TGFβ administration in the two MDCK cell lines.

Next, to understand the global transcriptional changes associated with the EMT programmes activated by these two sublines we performed bulk transcriptome analysis in control and TGFβ-treated MDCK cell lines. In our analysis, the control group shows high expression of epithelial and no expression of mesenchymal genes. After TGFβ treatment, MDCK-II cells epithelial genes are downregulated, while simultaneously activating mesenchymal genes. However, in NBL2 cells the epithelial component is completely downregulated, and mesenchymal genes are highly expressed, recapitulating global transcriptional changes associated with partial and full EMT programmes, respectively (Figure 12a).

The gene set enrichment analysis of EMT hallmarks supports the activation of an EMT pathway and global association of differentially regulated genes in these two sublines upon TGFβ treatment and EMT pathway (Figure 12b).

As we had previously described that the partial EMT activated in non-transformed adult cells was non-invasive, we wondered whether this was also the case of the partial EMT activated in MDCK-II cells. First, we selected the genes specifically upregulated in TGFβ-treated MDCK-NBL2 cells when compared to those upregulated in similarly treated TGFβ-MDCK-II cells. We identified 259 genes enriched in the NBL2 cells that interestingly, were able to segregate previously described invasive versus non-invasive cells corresponding the former to basal-like cancer cells (Neve et al., 2006) (Figure 12c). This phenotype is associated with an aggressive type of cancer, known to undergo EMT (Sarrió et al., 2008). Thus, the comparison between the transcriptional changes of two different EMT programmes implemented by these two sublines derived from same parental line, allowed us to define an EMT invasive signature that we refer to as breast cancer pro-invasion genes (BC-PINGs). The enrichment analysis of BC-PING signature shows a prototypical embryonic invasive EMT programme such as that activated during the delamination and migration of the neural crest cells (Thiery et al., 2009) (Figure 12d).

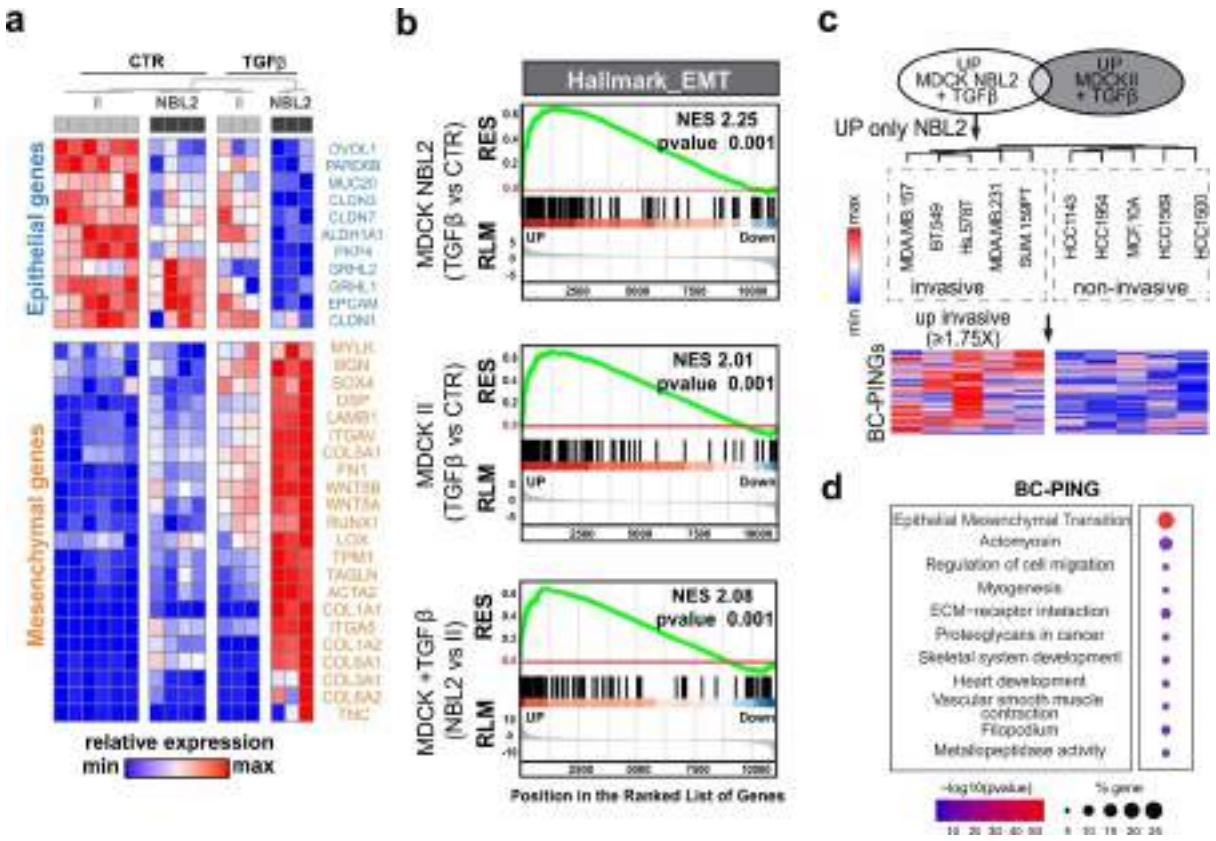


Figure 12 | Generation of a breast cancer pro-invasive gene signature after comparing two MDCK parental cell lines which respond differently to TGFβ treatment. **(a)** Heatmap with hierarchical clustering representing the relative expression of epithelial and mesenchymal genes in bulk transcriptome of MDCK-II and NBL2 cells upon TGFβ treatment. **(b)** Gene set enrichment analysis (GSEA) plots showing enrichment scores for EMT hallmark (MsigDB database) in MDCK-II and NBL2 cells upon TGFβ treatment. **(c)** Heatmap with hierarchical clustering separating invasive and non-invasive breast cancer cell lines using relative gene expression of genes specifically upregulated in NBL2 cells upon TGFβ treatment. The genes enriched ≥ 1.75X in the averaged invasive vs non-invasive basal-like breast cancer cell lines are hereinafter referred to as breast cancer pro-invasion genes (BC-PINGs). MCF10A was used as a control for non-invasive breast cancer cell lines. **(d)** Dotplot representing selected gene ontology (GO) terms significantly enriched in BC-PINGs signature.

Overall, our analysis of cell lines shows that the two MDCK sublines have a different level of EMT activation, partial and full, and that the full EMT was associated with the acquisition of invasive properties. The comparative analysis between the two EMT programmes allowed us to describe an invasive signature. With that, we decided to study EMT in different contexts, both physiological and pathological at the whole genome and single cell level, characterising the transcriptome in thousands of individual cells. With this approach, our aim was to track the changes occurring during the different transitions and to then find commonalities and specificities.

4.2 The invasive EMT programme activated during mouse trunk Neural Crest (NC) development

NC cells are multipotent embryonic cells and an excellent model to study EMT (Acloque et al., 2009; Ahlstrom and Erickson, 2009; Hay, 1958; Nieto, 2009). During vertebrate embryonic development, NC cells delaminate from the neural tube and undergo EMT to acquire migratory properties and migrate to multiple destinations to differentiate into different derivatives (Martik and Bronner, 2021) (Figure 13a).

To understand the molecular changes associated with the developmental EMT programme we used publicly available single cell RNA-Seq data (Soldatov et al., 2019). Soldatov et al. dissected the cervical region and trunk areas posterior to the otic vesicle from E9.5 $Wnt1^{Cre}/R26R^{Tomato}$ mouse embryos (Figure 13b) and captured the developmental trajectories of the neural crest. $Tomato^{+}$ NC cells were sorted and sequenced using Smart-seq2 protocol to obtain single cells transcriptional profiles.

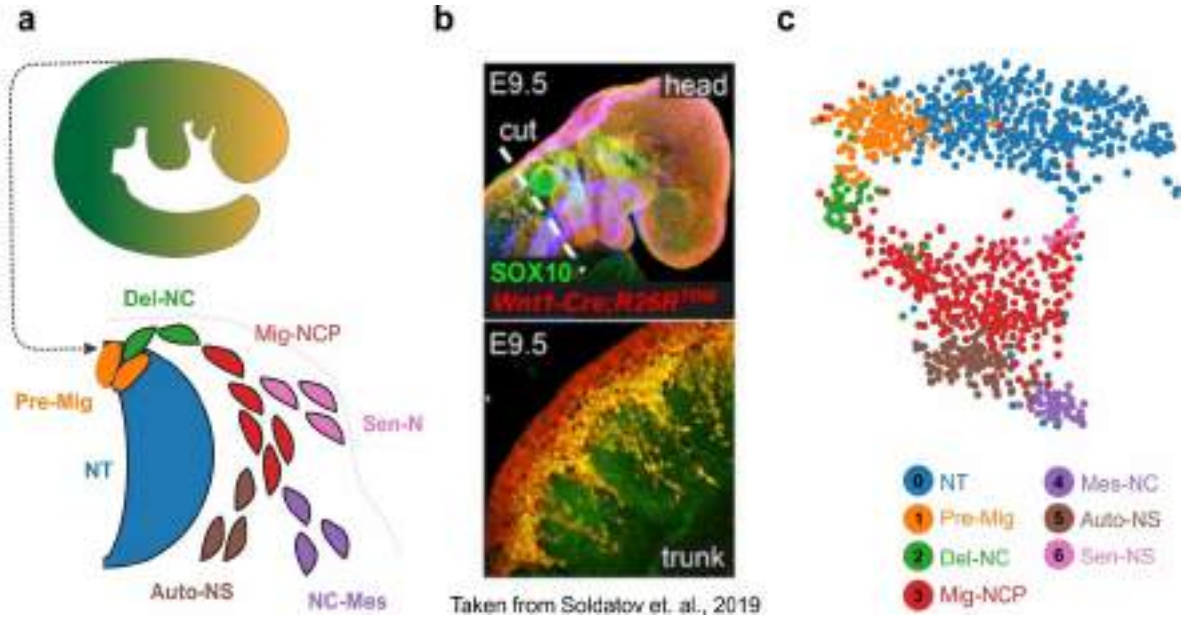


Figure 13 | Cellular heterogeneity during mouse trunk neural crest development in E9.5 mouse embryos. (a) Illustration depicting the progression of NCCs delaminating from the neural tube, adopting migratory properties and then either evolving to the mesenchymal phenotype or differentiating into neural crest derivatives, including autonomic and sensory neurons. **(b)** Immunohistochemistry of E9.5 $Wnt1^{Cre}$;$R26R^{TOM+}$ embryo, showing $Sox10^{+}$/$Wnt1^{TOM+}$ neural crest cells (NCCs) migrating in the head and trunk. The dashed line shows post-otic vesicle incision, a separation between cranial and trunk portions (taken from Soldatov et al., 2019). **(c)** t-SNE embedding showing major populations recovered using scRNA-Seq data generated for the mouse trunk neural crest development at E9.5. The 12 clusters detected in the original study were grouped into 7 major cell populations. 0: Neural Tube (NT), 1: Pre-migratory NCCs (Pre-Mig), 2: Delaminating NCCs (Del-NC), 3: Migratory Neural Crest Progenitor (Mig-NCP), 4: Mesenchymal NCCs (Mes-NC), 5: Automic Neurons (Auto-NS), 6: Sensory Neurons (Sen-NS).

The transcriptomic analysis by Soldatov et.al, revealed the cellular heterogeneity during NC development. The unbiased clustering approach divided the cells into 7 major clusters reported by Soldatov et. al. (Figure 13c), suggests a clear separation of neural tube and neural crest population in the obtained transcriptional space.

To construct unbiased connectivity map between different NC cell population during delamination, we used a sophisticated algorithm called PAGA (PArtition-based Graph Abstraction, Wolf et al., 2019). PAGA provides a framework which helps to connect transcriptionally similar cellular clusters in the form of an abstract graph. This help to understand the global topology of cellular transitions in single-cell data at the selected resolution. We employed the PAGA algorithm and obtained a connectivity map (Figure 14 a,b). RNA-velocity analysis in Soldatov et al., already proposed the directionality in the connectivity map starting form NT, premigratory NC population and delamination. Later, the NC cells become migratory progenitors and differentiate into automic and sensory neurons and cells with mesenchymal phenotype continue their migration (Figure 14c).

The major focus of Soldatov et al., was to analyse the molecular changes associated with NC differentiation into different derivatives. Our aim was to use these data to study how the EMT process was implemented. Therefore, for simplicity, we only considered the migratory NC cells and discarded those that bifurcated towards specific differentiation pathways.

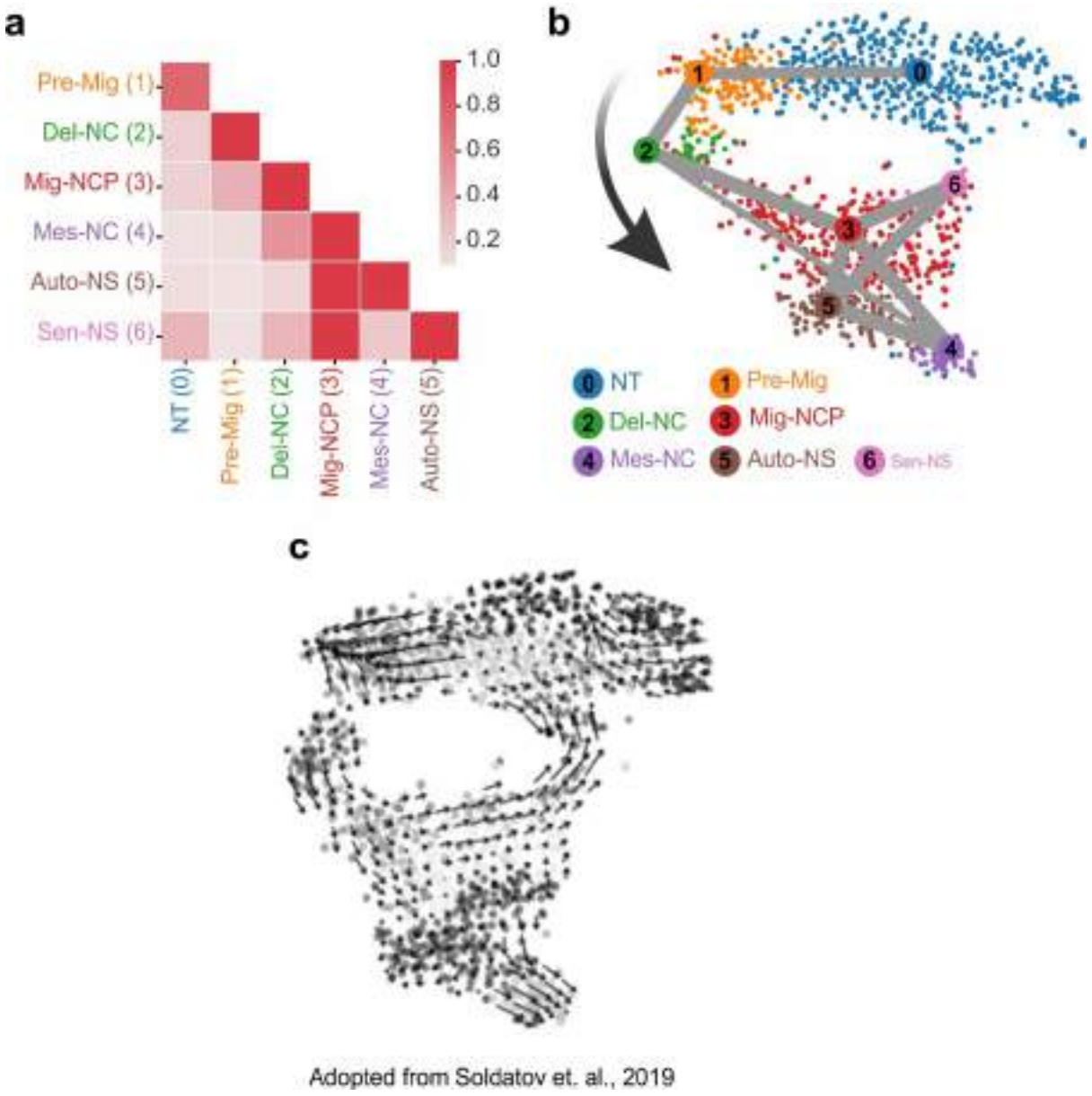


Figure 14 | NC differentiation trajectory shows the cellular transitions from the NT to different migratory routes and NC derivatives. (a) Connectivity matrix obtained by PAGA analysis showing transcription similarities between different populations during mouse trunk NC development. **(b)** PAGA connectivity map superimposed over t-SNE embedding. The nodes (circles) show different cell populations and the edges show the similarity index between them. **(c)** RNA-Velocity analysis showing the directionality of the cells during NC development (Adopted from Soldatov et al., 2019). Abbreviations as in Figure 13.

In scRNA-Seq technology, we can recover the molecular information at the level of individual cells but with a cost of losing spatio-temporal information. However, with the advancement in computation methods we can putatively reorganise the recovered cell populations in a pseudo-temporal order (Trapnell et al., 2014).

To understand the changes associated with EMT progression during NC development, we used a pseudotime approach. To calculate pseudotime, we need to select a reference cell called “root cell” which will be used as a starting point to organise all cells in a pseudotemporal order. We selected a root cell based on the highest expression of Hes5, an epithelial neural tube differentiation marker (Figure 15a). The obtained gradient of pseudo-temporal order shows that the NT cells give rise to cells with mesenchymal phenotype passing through successive intermediates states at the premigratory, delaminating and migratory states during the development of the NC (Figure 15b). As already mentioned, we excluded neuronal populations (i.e., autonomic and sensory neurons) form this point in our analysis, as their fate was committed and they were already differentiating, very likely now inhibiting EMT through the activation of the reverse programme, the Mesenchymal to Epithelial Transition (MET).

Next, to decipher the expression changes that occurred during NC delamination and migration, we clustered the genes based on their expression in the cells arranged over pseudotime. To define the gene modules, for major cell state transitions, we categorised the clustered genes into different groups based on their higher expression. This exercise resulted in the classification of cell in three major groups. Group-I, associated with the neural tube population (NT) showing high expression of genes such as *Olig3, Neurod1/4*, *Sox2*, etc. Group-II is associated with NC specification, with high expression of markers associated with pre-Migratory NC (*Zic3, Wnt4a*), delaminating NC and migratory progenitors (*Sox9/10, Est1, Nrp1/2, Snail1*). The appearance of *Snail1*, a pioneer EMT inducer and epithelial repressor (Cano et al., 2000) in this group confirms the activation of EMT and NC cell migration. Group-III is mainly associated with the mesenchymal phenotype and migratory NC progenitor profile, showing expression of *Twist1, Prrx1, Dlx1/2/3* etc. (Figure 15c). *Prrx1* and *Twist1* are the EMT transcription factors shown to help in maintaining a mesenchymal phenotype (Ocaña et al., 2012a; Soldatov et al., 2019).

NC cells delaminate from their primary location, undergo EMT to adopt migratory properties and then disseminate to different parts of embryos where they give rise to the different NC derivatives. To get an unbiased global view of EMT activation and its nature we performed gene set enrichment analysis. The concomitant enrichment of BC-PING (the invasive gene signature that we have defined using MDCK cell lines data; Figure 12c, d) with EMT hallmark confirms the activation of a graded and invasive EMT programme during NC development. The single cell level representation shows a detailed view of the activation of this invasive embryonic EMT programme (Figure 15d).

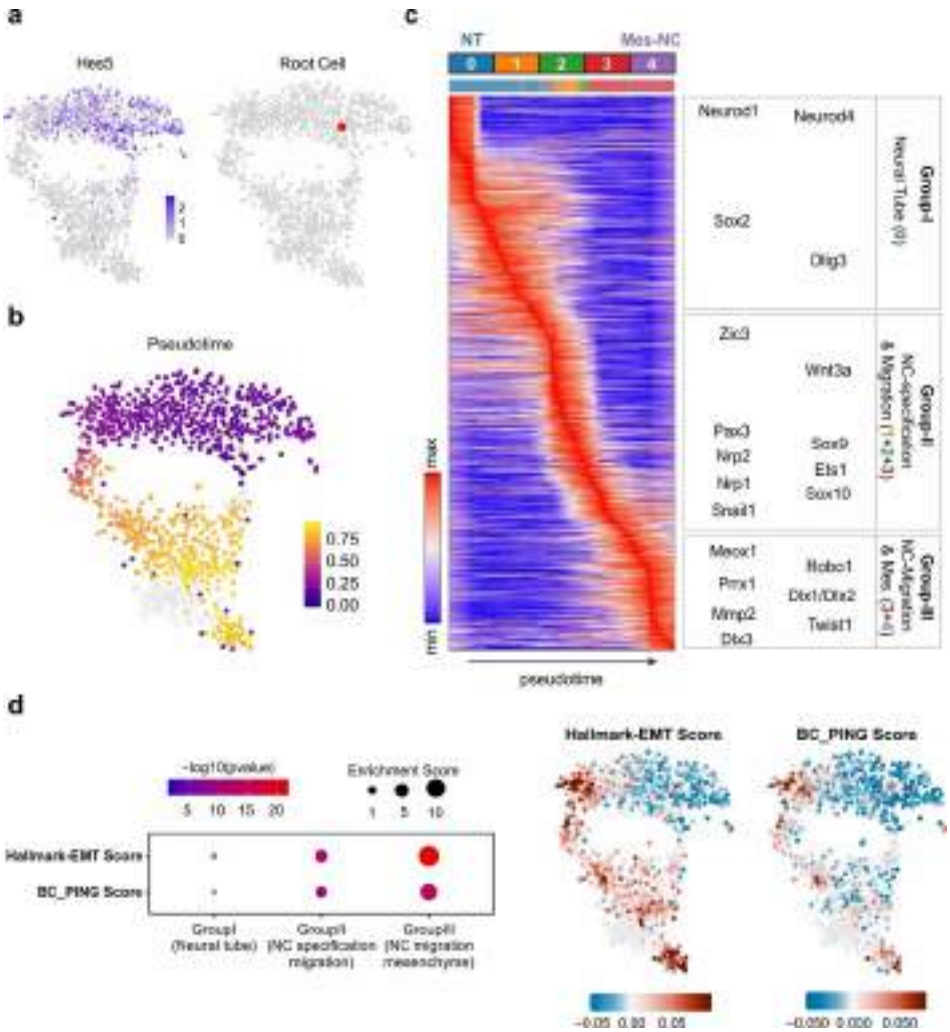


Figure 15 | Pseudotime analysis reveals the progression of cell states and gene expression dynamics during NC development. Please see legend in next page

A systematic regulation of specific transcription factors is essential for NC differentiation (Meulemans and Bronner-Fraser, 2004; Simões-Costa and Bronner, 2015). We applied SCENIC (Aibar et al., 2017) to predict an expression-based regulon of the transcription factors (TF) associated with the EMT trajectory during NC development. Figure 16 shows the regulation of gene modules that we found important to maintain the cell states. For instance, TF such as *Pax6, Olig3, Ascl1*, etc. show higher activity in the neural tube population, and are responsible for maintaining neuroepithelial identity activating the expression of neuroepithelial markers (*Hes5/6, Pax2, Dbx1*, etc.). Next, the delaminating NC population shows lower activity of these TF and higher activity of regulators such as *Mafb, Ets1, Msx1/2, Snail1*, essential for EMT activation and neural crest delamination (Bendall and Abate-Shen, 2000; Bronner and Simões-Costa, 2016; Ishii et al., 2005; Liem et al., 1997). The TFs activated in delaminating and migratory NCPs putatively regulate EMT markers such as *Pdgfra, Mmp2, Col1a2, Tfap2b* etc. Although *Snail1* is a better epithelial repressor than mesenchymal inducer (Nieto et al., 2016), our expression-based regulon prediction suggests that the *Snail1* can activate *Prrx1*.

Figure 15 | Pseudotime analysis reveals the progression of cell states and gene expression dynamics during NC development. (a) t-SNE embedding showing the relative expression of a neuroepithelial marker (Hes5) in the neural tube (left). t-SNE embedding showing selected root cells in red (right) **(b)** t-SNE embedding with pseudotime represents the pseudo-progression of cells in a differentiation trajectory. The minimum pseudotime value represents the precursor state whereas the maximum value represents an advanced state in a differentiation trajectory. **(c)** Heatmap with clustering of genes based on their expression in the cells arranged by pseudotime. The clustered genes were categorized in 3 groups: Group-I, associated with epithelial characteristics (higher expression in NT population; Group-II, with higher expression in Pre-Mig, Del-NC and Mig-NCP (neural crest specification and delamination); Group-III, associated with NC migration and mesenchymal phenotype. **(d)** Dotplot representing enrichment score of EMT hallmark (MsigDB database) and BC-PINGs in the three different groups (left. t-SNE embedding with EMT-Hallmark and BC-PINGs enrichment score at single cell level (right). Abbreviations as in Fig. 13.

Next, the high activity of TF such as *Twist1, Prrx1* or *Meox1* in the mesenchymal population indicates their described role as mesenchymal inducers required to promote and maintain mesenchymal phenotype, helping NC cells to disseminate to different target sites in the embryo. These regulons are also predicted to activate genes associated with extracellular matrix remodelling and cell migration (*Col3a1, Col1a2, Meox1/2, Bgn, Pdgfra* etc.).

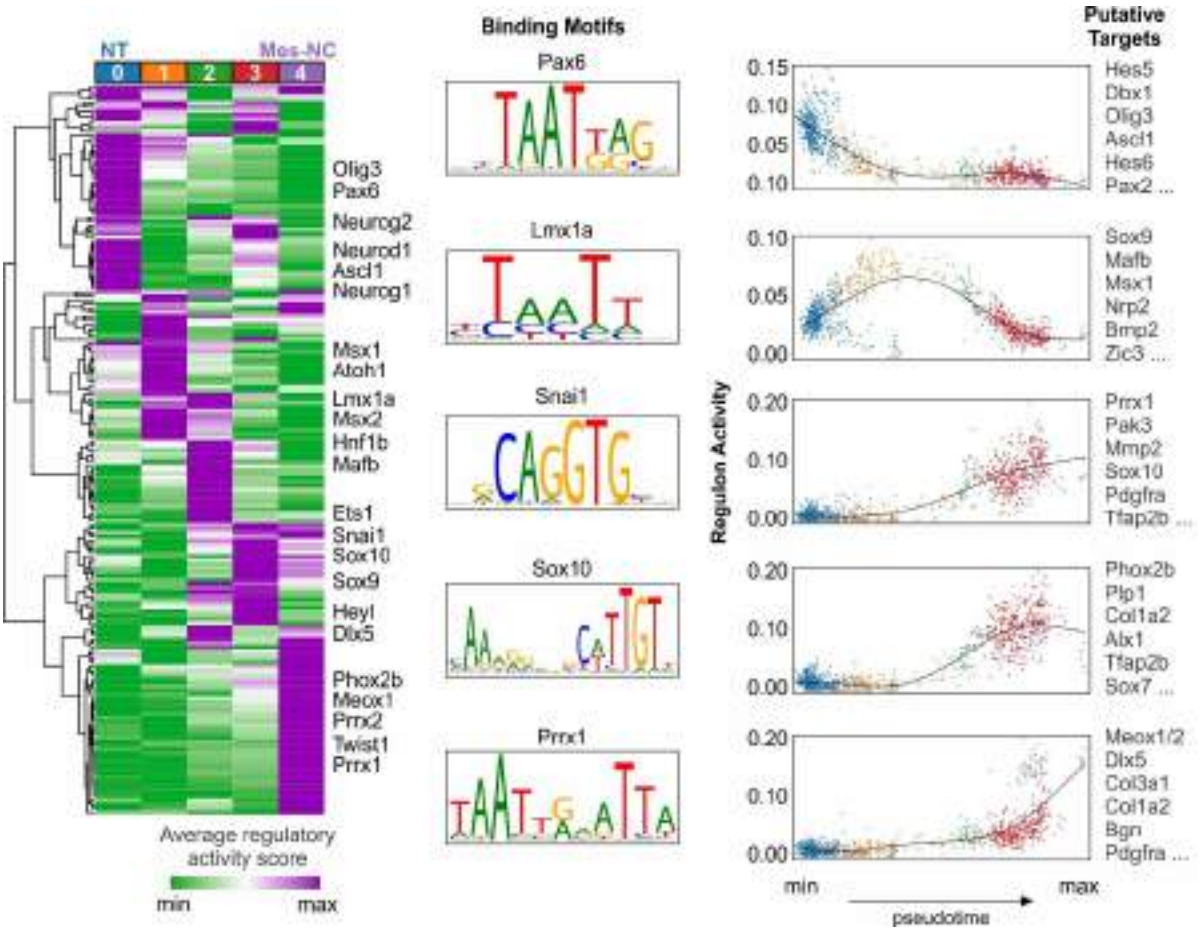


Figure 16 | Expression based regulon analysis shows the transcription factor code associated with EMT trajectory during NC development. Heatmap showing average regulatory activity of predicted regulons (left). Selected regulons show cell state specific regulatory activity represented over pseudotime along with their enriched DNA binding motif (right).

In our analysis of publicly available scRNA-Seq data of mouse trunk neural crest development, we recapitulated the changes in gene expression implemented during development. Enrichment analysis of our invasive gene signature is compatible with the activation of invasive EMT programme during NC development which is essential for the NC cells to delaminate and migrate to the different parts of embryos. Additionally, the predicted regulon analysis helped us to decipher the transcriptional code associated with the transition during NC development.

4.3 Reactivation of a partial non-invasive EMT programme in kidney fibrosis

Classically, the EMT has been associated with invasion and cell migration. However, recent reports suggest that renal tubular epithelial cells reactivate an EMT programme during renal fibrosis but do not engage in the invasive program like embryos do. Instead, these cells activate a partial EMT programme and remain attached to the tubules (Grande et al., 2015; Lovisa et al., 2015).

4.3.1 Transcriptomic analysis by single-cell RNA sequencing recovered a profile that reveals the cellular heterogeneity in control and fibrotic kidney samples

To understand the molecular changes associated with the partial EMT programme reactivated during kidney fibrosis we performed a systematic scRNA-Seq data analysis. We used a mouse model in which we induced fibrosis by unilateral ureteral obstruction (UUO). UUO induces tubular injury which leads to renal interstitial fibrosis and eventually renal failure (Chevalier, 2016). We prepared 10X libraries of two UUO and one control/SHAM samples (See methods and Figure 17a,b).

In the next generation sequencing approach, base quality is determined using a score called Phred. It indicates the likelihood of incorrect base call at given position in a sequencing read (Ewing and Green, 1998). It is crucial to investigate the base quality and remove the bases with low Phred quality score, as it greatly impacts the downstream analysis. We also need to look for the adapter contamination and remove it from raw reads. The average Phred quality score for our samples is >20. Also, we did not observe any over-represented sequences

and the adapter contamination. The detected GC content in our samples overlaps with the theoretical range of 51.24 (±7.80 SD) (Romiguier et al., 2010), indicating the absence of foreign sequence contamination in our libraries.

In our transcriptomic libraries we obtained 900 M total number of reads with 300.22 M (± 3.00 M SD) average number of reads per sample (Figure 17c). The average raw read alignment shows 91.03 % (± 1.07 SD) confident mapping on the genome, out of which 53.2% (± 1.71 SD) reads were mapped on exonic regions. We obtained 44007 total number of cells with the average of 14669 cells (± 916.03 SD) per sample. The average number of genes detected per sample was 32536.67 (± 1624.86 SD) (Figure 17c). The mean sequencing saturation point for the libraries was 0.36 (± 0.02 SD). A detailed sample wise statistics is shown in Table 2.

	SHAM	**UUO#1**	**UUO#2**	**Mean**	**std**
Estimated Number of Cells	15764	13522	14721	14669	916.03
Total Number of Genes	30239	33657	33714	32536.67	1624.86
Total Number of Reads	304.08 M	299.81 M	296.76 M	300.22M	3.00M
Mean Reads per Cell	19289	22172	20159	20540	1207.42
Median Genes per Cell	1550	1734	1431	1571.67	124.64
Median UMI Counts per Cell	3750	3124	2616	3163.33	463.79
Reads Mapped to Genome	92.4	90.9	89.8	91.03	1.07
Confident mapping on Exons	54.1	54.7	50.8	53.2	1.71
Sequencing Saturation	0.321	0.37	0.384	0.36	0.03

Table 2: Raw data and alignment statistics for scRNA-Seq libraries prepared for kidney fibrosis. Std=Standard Deviation.

Generally, in droplet-based single-cell RNA sequencing methods it is assumed that one droplet contains RNA from one cell. However, this hypothesis is not always true. There could be potential empty droplets that can be removed by CellRanger, or also, a droplet can contain RNA from multiple cells, and these are called doublets. There are certain parameters such as the number of detected genes, number of UMI, and percentage of mitochondrial genes (Ilicic et al., 2016) which can help to determine the quality of a cell. For instance, a cell with low number of detected genes, low number of UMI and high percentage of mitochondrial genes, can be considered as a dying cell. On the other hand, if a cell has a very high number of genes and count of UMI could be a doublet (Luecken and Theis, 2019). We detected and removed low quality cells based on these different QC matrices (Figure 17c-e). We detected an average of 398 (± 184.15) doublets per sample. After applying appropriate filters, explained in materials and methods, we obtained 25424 high-quality cells in three samples (8474.66 average number of cells with ± 2375.02 SD).

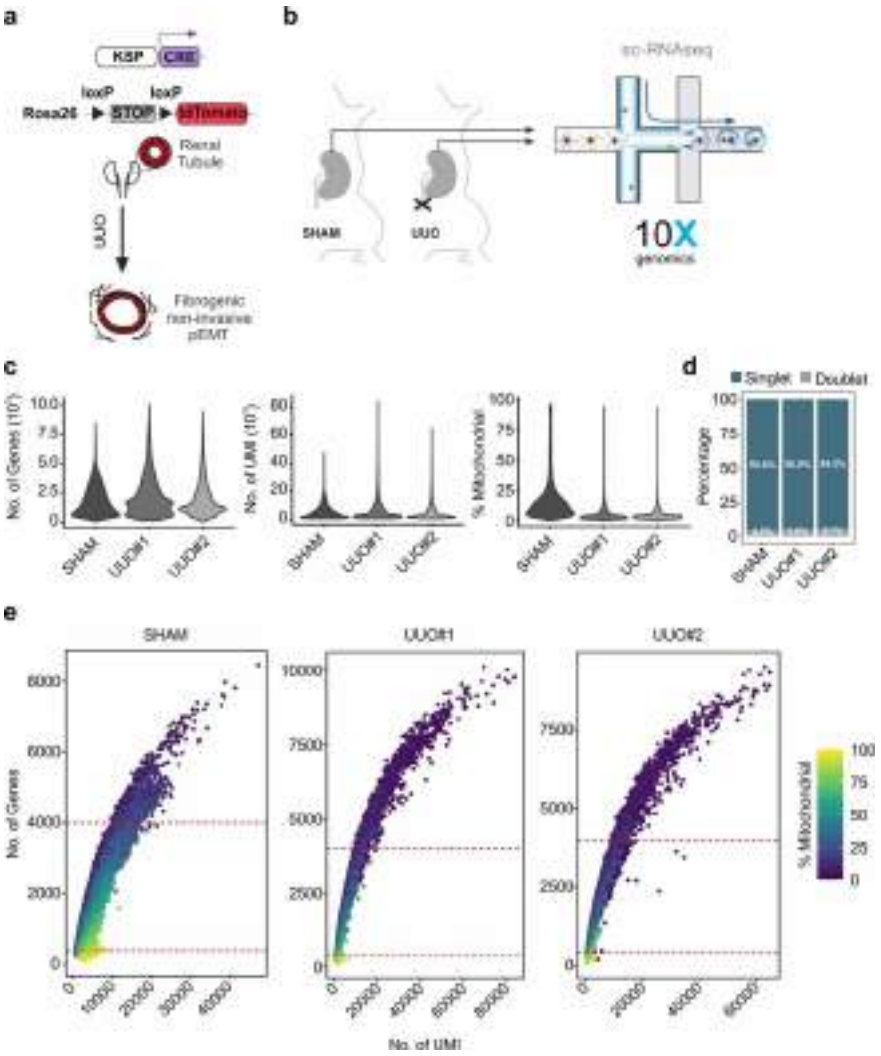


Figure 17 | Single cell transcriptome of SHAM and obstructed kidney samples recovered high quality cells during kidney fibrosis. (a) *in vivo* mouse model of Unilateral Ureteral Obstruction (UUO) induced fibrosis, including the activation of a non-invasive partial EMT (pEMT) in epithelial renal tubules (RT). **(b)** Illustration of the strategy used for sample preparation and sequencing. **(c)** Violin plot showing different QC parameters per sample, including number of cells, number of genes and percentage of mitochondrial genes. **(d)** Barplot showing the percentage of putative doublets and singlets predicted using Scrublet (Wolock et al., 2019). **(e)** Scatterplot showing different QC parameters as above to determine the quality of the captured single cells. Less genes, high numbers of UMI and high percentage of mitochondrial genes suggest dying cells, whereas a very high number of genes suggest that the detected cell could be a multiplet. Dotted lines indicate the cut-offs.

scRNA-Seq data have significant cell-cell variation due to technical noise along with biological confounding factors (Hafemeister and Satija, 2019). To preserve the biological differences while removing the technical noise, the data needs to be normalised before proceeding further with the downstream analysis. We used SCTransform implemented in Seurat for data normalisation and feature selection (Hafemeister and Satija, 2019). We selected the top 3000 high variable genes to reduce the computation and less important features. We integrated SHAM and UUO samples using the canonical correlation-based anchors detection method implemented in the Seurat package (Figure 18a). Then, the lower dimension of the integrated space was used for clustering and data representation. The clustering resulted in 26 different cellular clusters providing information on the transcriptional heterogeneity (Figure 18b).

4.3.2 Significant remodelling of cellular composition upon chronic injury (fibrotic condition) in the kidney

Next, the obtained clusters were annotated into 7 major cell populations using cell-type specific markers (Figure19 a). They comprise the epithelial component, immune components (both myeloid and lymphoid), the glomerulus, endothelial, interstitial and proliferative cells. In the epithelial component we recovered the cell populations which are structural units of kidney such as proximal tubules (PT: *Lrp2, Acsm2, Keg1*), loop of henle (LoH: *Bst1, Phgdh*), thick ascending limb (TAL: *Slc12a1, Slc5a3*), connecting tubules (CT: *Slc12a, Tmem52b*), principal cells (PC: *Aqp2, Fxyd4*), and intercalated cells (IC: *Atp6v1g3, Atp6v0d2*). Additionally, there is a cluster of injured cells close to the epithelial cells expressing injury markers such as *Vcam1, Ccl2, Ccl4* etc. We predicted the origin of these injured cells (explained in section “Contribution of epithelial clusters to injury”) and these are the part of epithelial component. The immune cell component is composed of T cells (*Trac, Ikzf2, Cd5*), B cells (*Cd79a, Cd79b*), NK cells (*Nkg7, Ctsw*), macrophages (*C1qa, Ms4a7*), monocytes (*Chil3, Hp, Emilin2*), granulocytes (*Csf3r, Cxcr2*) and dendritic cells (*Cd209a, Tnip3, Siglech, Cd300c*). We recovered two populations from glomeruli, podocytes (*Ncam1, C3*) and parietal cells (*Nphs1, Nphs2*). The endothelial compartment shows enrichment of markers of arteries and veins in both the cortical and medullar regions of the kidney (*Egfl7, Fbln5, Plvap, Ehd3*). The stromal cell population is mainly

composed of perivascular structure (*Rgs5, Acta2*) and myofibroblasts (*Col1a1, Col3a1*). We also detected a small cluster of proliferating cells expressing markers such as *Mki67, Top2a*. Overall, our single cell data of SHAM and UUO samples successfully recapitulated the cell populations present in structural and function units of the kidney, plus the injured population.

As previously reported by Conway and colleagues (Conway et al., 2020), there is remodelling of the non-epithelial component, especially of immune cells. To study the remodelling in different cell types, we performed control-case analysis (Petukhov et al., 2022). We observed a significant remodelling of the non-epithelial components, i.e. immune and interstitial cells, along with the epithelial population upon UUO (Figure19 b,c). We observed more than 7-fold increase in lymphoid and more than 12-fold increase in myeloid cells upon injury. Also, there is a significant increase in proliferating and stromal populations (>2-fold) in obstructed kidney samples compared to the control one. Finally, there is around 2-fold reduction in the epithelial cell population. This reduction shows the impact of the injury on epithelial cells.

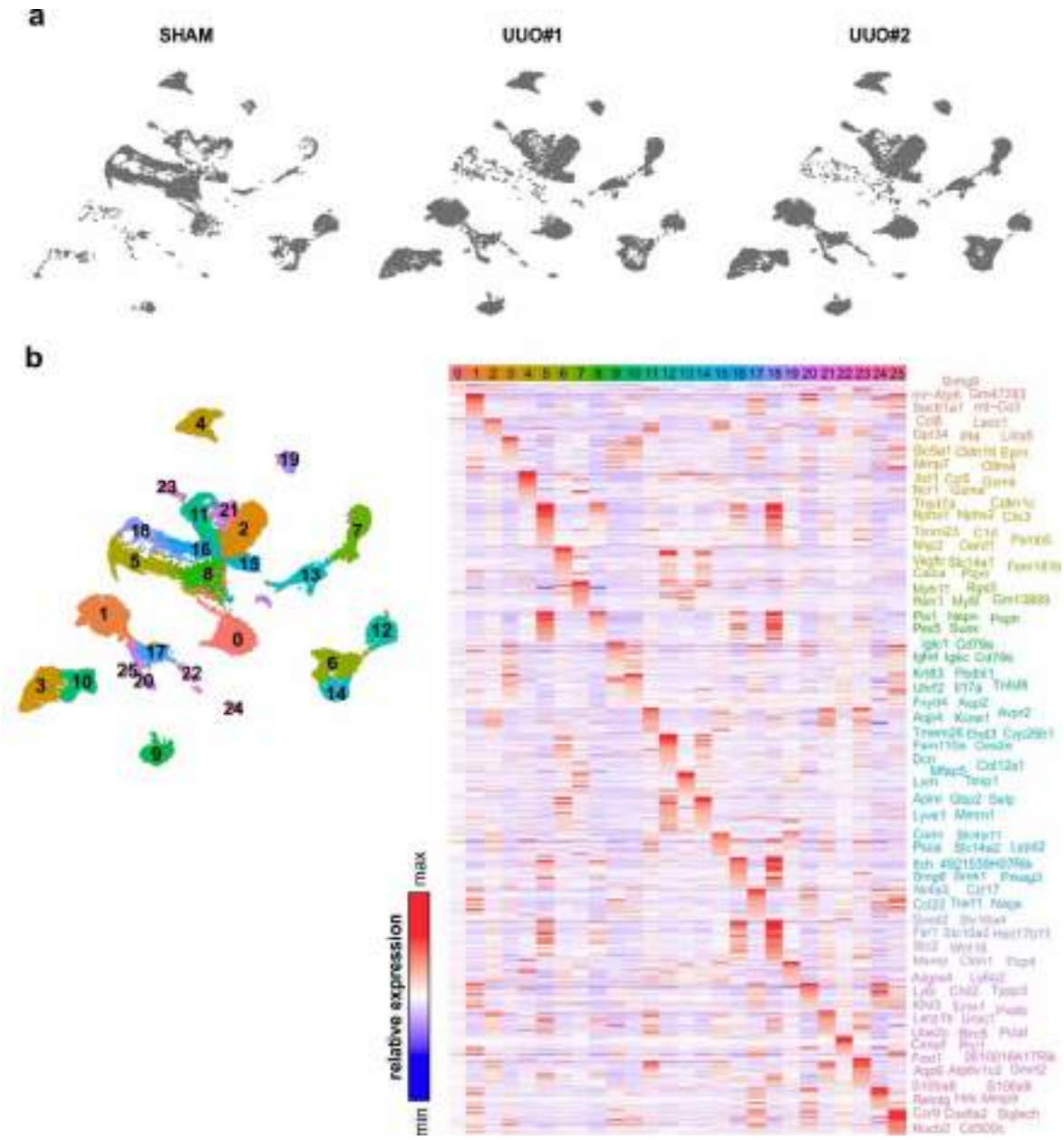


Figure 18 | Global expression profile of integrated control and obstructed kidney samples resulted in 26 cellular clusters providing information on cellular heterogeneity and global transcriptional changes. (a) Uniform Manifold Approximation and Projection (UMAP) showing integration of SHAM and UUO samples. **(b)** UMAP shows 26 distinct cellular clusters obtained using shared nearest neighbour (SNN) modularity, providing information on cellular heterogeneity in healthy and obstructed kidney samples (left). Heatmap representing average relative expression of differential expressed genes in each cluster (right; average log2FC > 0.25 and FDR < 0.05).

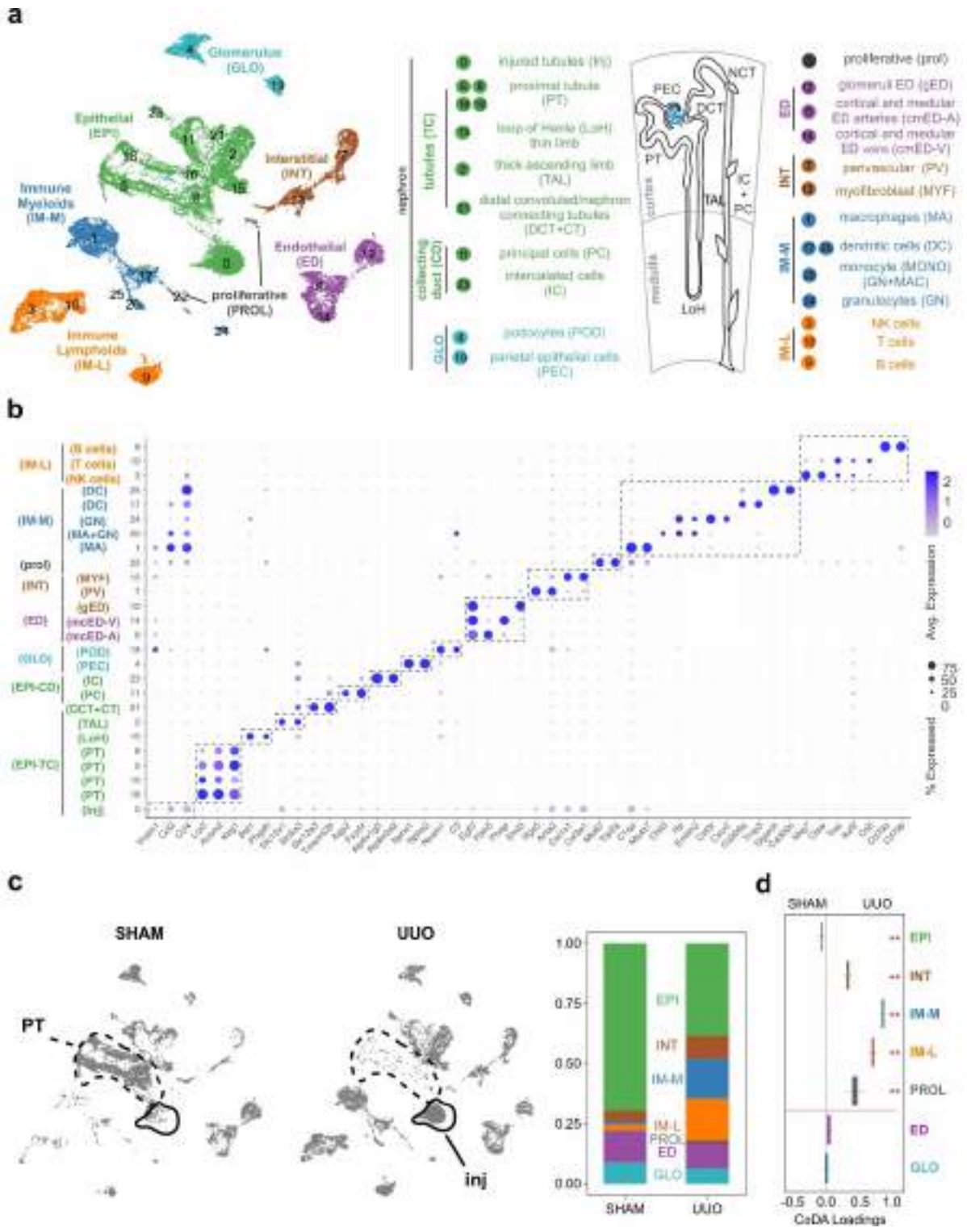


Figure 19 | UUO induced kidney fibrosis significantly remodel cellular composition compared to healthy kidney. Please see legend in next page

4.3.3 Proximal tubules are the major contributors of epithelial injury during kidney fibrosis induced by UUO

The disappearance of the proximal tubule population from the control sample and the appearance of an injured cluster upon UUO suggests that those of the proximal tubules (PT) were the most injured cells. Additionally, several studies reported that PT are the major contributors to damaged cells and driver of tubulointerstitial fibrosis in response to chronic injury (Chevalier, 2016; Gewin, 2018). However, in our scRNA-Seq data we did not want to exclude the possibility of other epithelial populations being injured. Thus, to trace the origin of injured cells during induced fibrosis, we built a deep learning (DL) multi-class classification model using expression profile of epithelial cells (excluding the injured cluster). We validated our DL model performance using Mathew Correlation Coefficient (MCC) and the percent accuracy was calculated using using a 10-fold cross validation approach. We observed that the overall average cross validation accuracy and MCC is 0.986 (± 0.008) and 0.963 (± 0.022), respectively (Figure 20a). The prediction of injured population showed that around 90% of injured cells were originated from PT (Figure 20b), compatible with the immunofluorescence of PT and injured markers after UUO (Figure 20c).

Figure 19 | UUO induced kidney fibrosis significantly remodel cellular composition compared to healthy kidney. (a) Uniform Manifold Approximation and Projection (UMAP) showing 26 clusters detected using unsupervised clustering into 7 major cell populations. Every cluster was assigned to the cell type according to their expression of *bona fide* markers (upper panel). Dotplot representing expression of the markers used to annotate the different clusters (lower panel). **(b)** UMAP representing the cellular composition in SHAM and UUO operated kidney samples (left side). The barplot shows the percentage of cell type composition changing between SHAM and UUO kidney samples. **(c)** Boxplot showing the statistically significant changes in cellular composition upon UUO induced fibrosis. We applied Cacoa (Petukhov et al., 2022), a statistical framework for the composition analysis using the glomerulus (least altered cell population) as a reference cell type and 1000 bootstrapping. A horizontal red line separates cell types passing significance threshold ($p<0.05$).

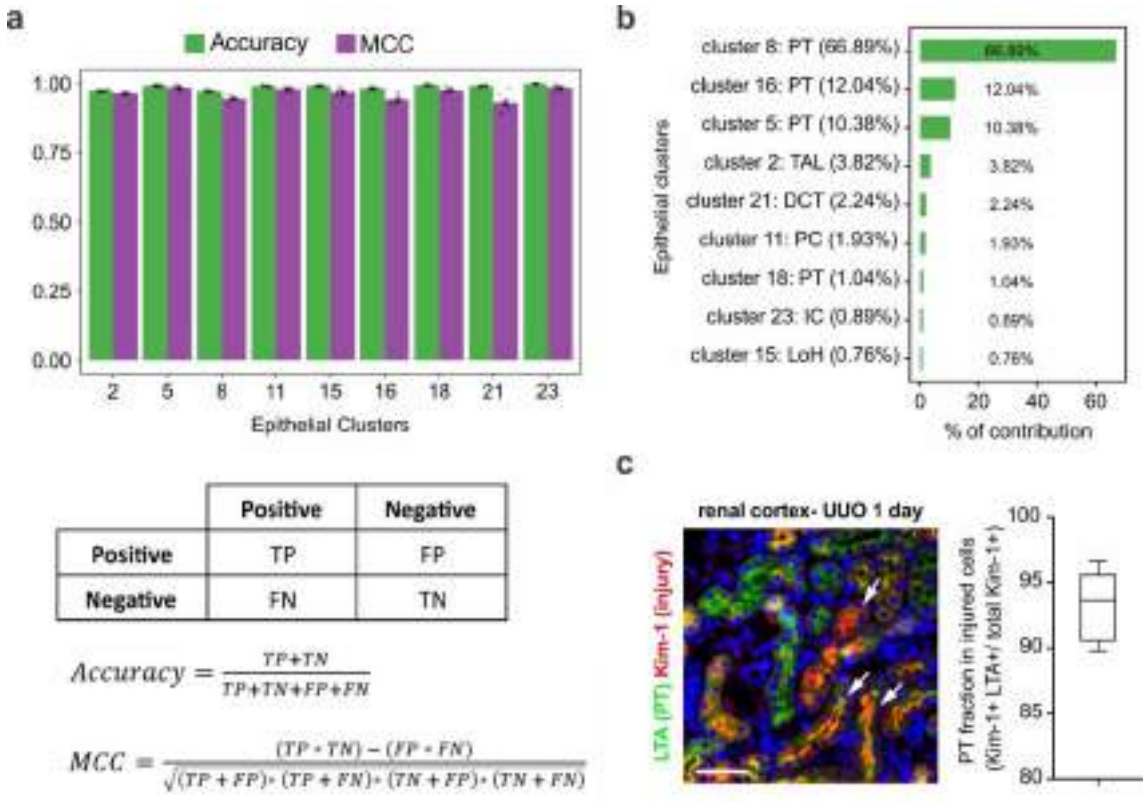


Figure 20 | The proximal tubule population of the kidney are the major contributors to UUO-induced fibrosis. (a) Bar plots showing accuracy score and normalised Matthews correlation coefficient (Chicco and Jurman, 2020) obtained by 10-fold cross validation over training dataset (details in materials & methods). The lower panel shows the formula used to calculate accuracy parameters. **(b)** Barplot showing the percentage of contributions of epithelial clusters to injury, predicted using our machine learning model. **(c)** Immunofluorescence of the injury marker Kim-1 in combination with the PT cell marker LTA. Arrows indicate Kim-1 positive cells also positive for LTA. The quantification shows the percentage of Kim-1 positive injured cortical epithelial cells also positive for LTA. Nuclei in blue. Scale bar, 50μm.

Taken all together, systematic analysis of our single cell transcriptome data of the SHAM and UUO operated kidney samples indicates that there is a significant remodelling in non-epithelial components and in the PTs, with a majority appearing as an injure cell population in kidney fibrosis.

4.3.4 Upon injury, PT cells undergo dedifferentiation upon activation of an EMT inflammatory programme

We and others have reported that there is a reactivation of non-invasive, partial EMT programme during kidney fibrosis (Grande et al., 2015; Lovisa et al., 2015). The renal tubular epithelial cells activate EMT upon injury but do not to engage in an invasive programme (as seen during embryonic development) but rather remain in the injured tubules secreting cytokines and chemokines that leads to fibrogenesis and inflammation (Grande et al., 2015; Nieto et al., 2016). However, the molecular mechanisms and the transcriptional programmes behind the activation of this adult EMT are poorly studied.

To get insight into the programme, we subset the PT cells, predicted to be the origin of the injured population (Figure 20b). We re-clustered these cells, resulting in three major clusters. Cluster-I is composed of healthy differentiating PT cells showing higher expression of differentiation markers (*Miox, Hnf4, Acy3*; Figure 21b) and as expected, with a higher differentiation score (Figure 21c). Moreover, the majority of the cells in this cluster belong to the control sample (Figure 21a, lower panel). In cluster-II, the cells are dedifferentiated. While shutting down the renal differentiation programme (Figure 21c), these cells simultaneously activate an EMT programme, concomitant with a robust inflammatory response as a response to the injury (Figure 21c). We observed a fully dedifferentiated cell state of PT in Cluster-III, which cells lack expression of differentiation markers and show activation of EMT (*Col3a1, Vim, Itgb1*), inflammatory citokines and chemokines (*Ccl2, Il1b, Cxcl2*) and injury makers (*Jun, Lcn2, Klf4*; Figure 21b). The UUO samples are the major contributors to this cluster, as shown in lower panel of Figure 21a. Additionally, Cluster-III shows a significant reduction in the differentiation score, concomitant with a significant increase in EMT and inflammation scores (Figure 21c).

The combined analysis of different markers and enrichment of different scores indicate that upon injury, healthy PT cells undergo dedifferentiation simultaneously activating an EMT programme which may help in repairing injured cells. However, due to the chronic condition, it may lead to renal failure as we described previously.

To infer the directionality of the cells in this transition we used a powerful approach called RNA-velocity which uses splicing kinetics (La Manno et al., 2018). Our RNA-velocity analysis shows that cells from the healthy PT population transition towards the injured one through the dedifferentiating PT population (Figure 22a). Finally, the reduction in *E-cadherin* (epithelial marker) and the activation of Vimentin (mesenchymal marker) in the same cells confirms the activation of an EMT programme in renal epithelial cells upon UUO (Figure 22b).

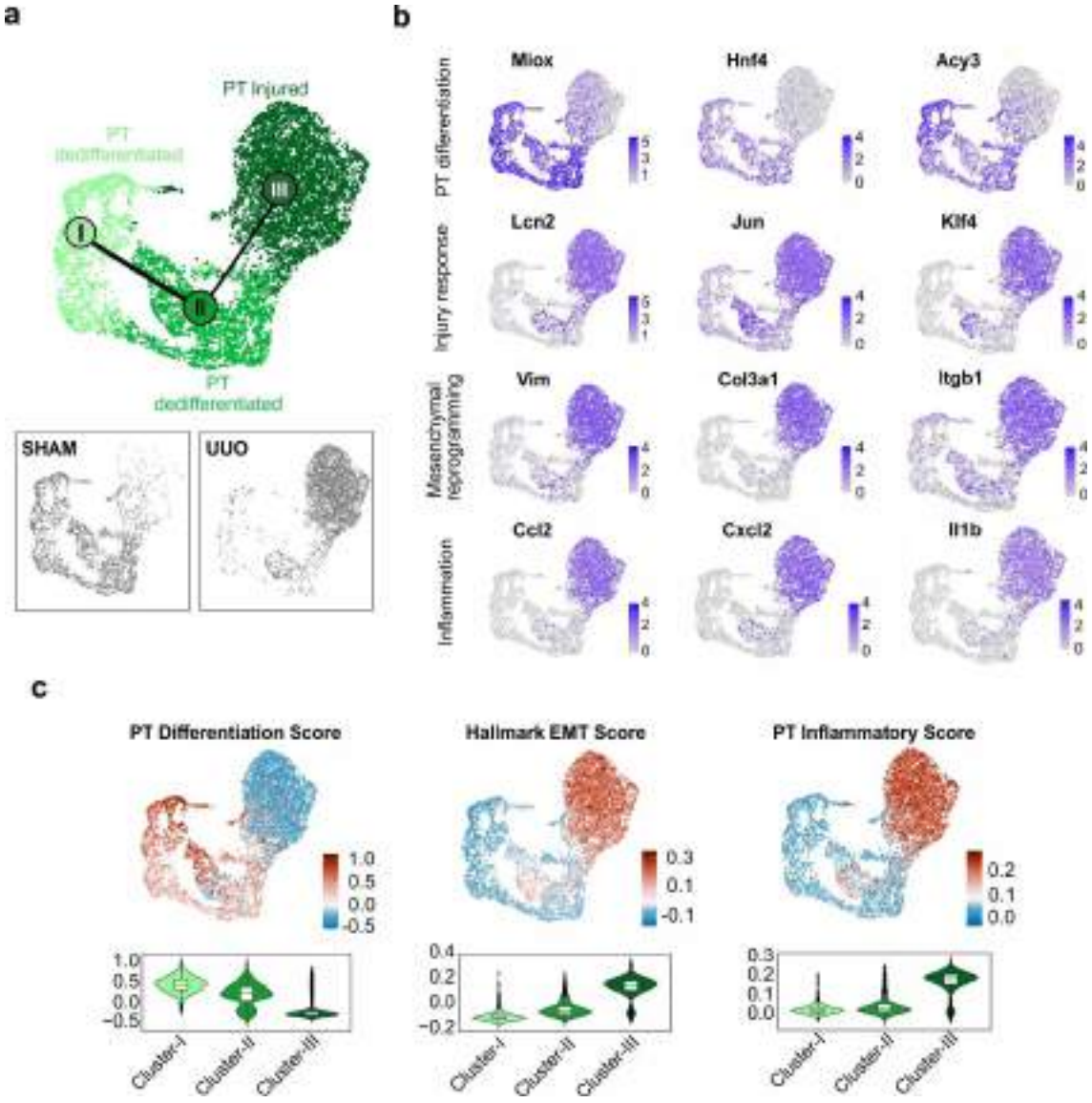


Figure 21 | PT cells undergo dedifferentiation activating an inflammatory EMT programme as a response to injury upon UUO induced fibrosis. (a) UMAP representing three major clusters of healthy and damaged PT populations. **(b)** UMAP plots showing the expression of markers of renal-specific epithelial differentiation, injury/repair, mesenchymalysation, and inflammation. **(c)** UMAP showing enrichment score for PT differentiation (Ransick et al., 2019), EMT-Hallmark (MsigDB database), and PT inflammation score (Wu et al., 2020) at single cell level.

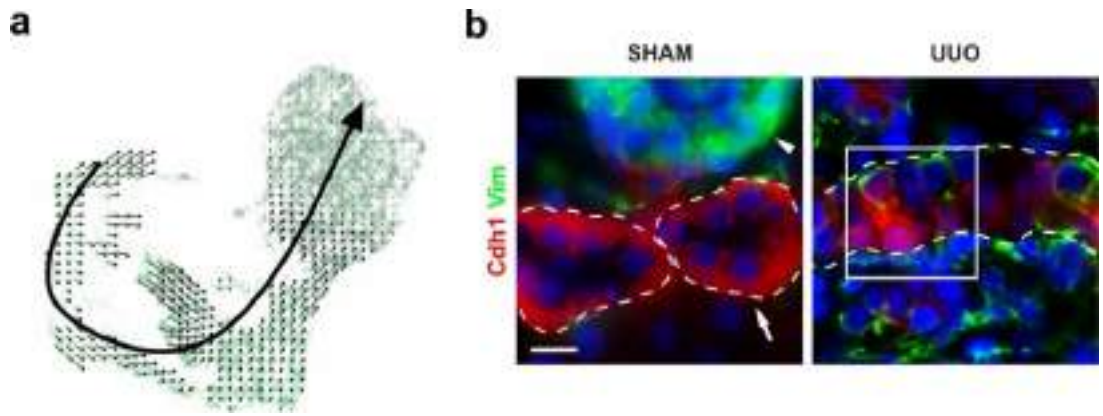


Figure 22 | PT cells undergo a partial non-invasive EMT during UUO-induced fibrosis. (a) UMAP showing the directionality of cells in the PT dedifferentiation trajectory predicted using RNA velocity. The principle curve (solid line) was fitted using Slingshot, a trajectory inference method (Street et al., 2018). **(b)** Immunofluorescence for epithelial (E-Cadherin) and mesenchymal (Vimentin) markers. Double-positive cells for epithelial and mesenchymal markers indicate the activation of a partial EMT programmme during the development of UUO-induced fibrosis. The arrow indicates renal epithelial cells and the arrowhead, a glomerulus. Scale bar, 10μm.

4.3.5 Activation of the EMT programme includes injury response and inflammatory pathways

Next, to decipher the changes in gene expression along the EMT trajectory in kidney fibrosis we used a similar strategy to that used for the analysis of the NC. After selecting a root cell based on the higher expression of PT differentiation marker (*Slc22a12*; Figure 23a) we calculated the pseudotime (Figure 23b). Then, the genes were clustered based on their expression in pseudo-temporally organised cells and categorised them into two different groups (Figure 23c). While the first group is associated with PT differentiation and show higher expression of *Miox, Hnf4a*, and *Slc34a1*, the second group is composed of genes showing higher expression in cells from Cluster-II and Cluster-III. The second group mainly shows higher expression of injury (*Lnc2, Jun/Fos, Vcam1*), inflammatory (*Ccl2/5, Notch2/3, Nfkb*) and EMT (*Tgfb1, Ecm1, Tgfbr2*) markers.

Also, we observed higher expression of metalloprotease inhibitors such as *Tim1/2* in Cluster-III compatible with the inhibition of full mesenchymal phenotype during kidney fibrosis (Pezeshkian et al., 2021). The impact of the changes in gene expression in these different groups were tested using pathway enrichment analysis. In our pathway enrichment analysis, we observed that EMT activation starts in Cluster-II and is maintained in Cluster-III, showing stronger enrichment. Additionally, the differentiation pathways are highly enriched in Cluster-I and not so in Cluster-II and III. On the other hand, Cluster-II and III show significant enrichment in pathways related to inflammatory and injury response such as interleukin signalling, interferon response, inflammatory response, cytokine-chemokine response etc.

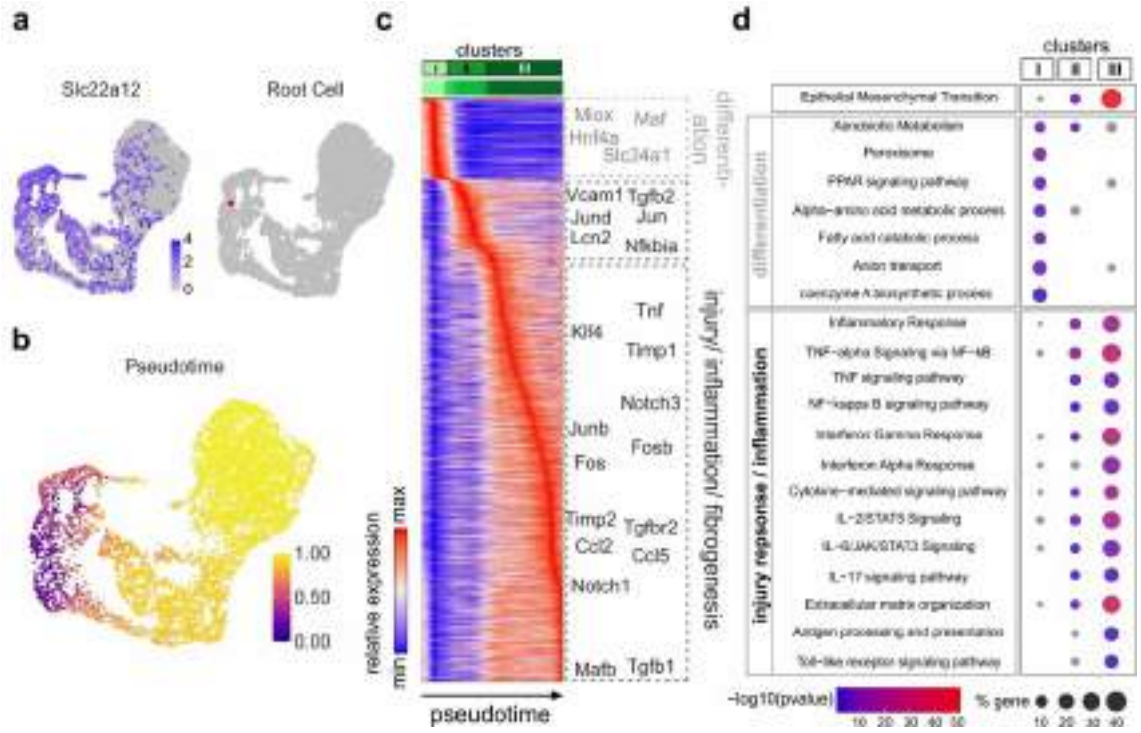


Figure 23 | Pseudotime analysis reveals molecular changes associated with dedifferentiation and injury response/ inflammatory pathways in kidney fibrosis. (a) UMAP embedding showing the relative expression of *Slc22a12*, a marker of PT differentiation (left). UMAP embedding shows the selected root cells in red (right). **(b)** UMAP embedding with pseudotime representing the pseudo progression of cells in a dedifferentiation trajectory. **(c)** Heatmap with clustering of genes based on their expression in the cells arranged over pseudotime. The clustered genes were categorised into two distinct groups. Group-I represents

genes associated with PT differentiation and a healthy PT population, showing higher expression of markers such as *Hnf4, Miox, Slc34a1* etc. Group-II comprises the activation of injury markers such as *Vcam1* and *Lcn2*, which progress further and show activation of different cytokines such as *Ccl2* and *Ccl5*. **(d)** Dotplot showing pathway enrichment analysis in the three different clusters, depicting loss of differentiation and gradual activation of an EMT programme along with pathways associated with injury response and inflammation.

Like for the NC, we predicted the regulatory activity in the EMT trajectory defined using scRNA-Seq data for the kidney fibrosis. As expected, we observed cell-state specific TF activities. Genes encoding TFs such as *Hnf4* and *Rora* show a higher regulatory activity in Cluster-I to regulate the cell differentiation by activating markers such as *Miox*, *Slc34a1, Acy3* and *Ltf* (Figure 24). As the cells engaged in the EMT trajectory dedifferentiate, *Hnf4* and *Rora* show lower or no regulatory activity in Clusters II and III (Figure 24). Meanwhile, the dedifferentiating cells activate different sets of TFs including *Fos, Jun, Runx1, Mafb* or *Irf1*in Clusters II and III. The putative regulators in these clusters are mainly associated with inflammation and immune response (Figure 24). Interestingly, we detected *Zeb1*, an EMT transcription factor as a putative regulator. We observed that *Zeb1* shows higher regulatory activity in Cluster-II, and *Zeb1* is described to play an important role in maintaining a partial EMT programme (Liao et al., 2021) in a context-specific manner. *Zeb1* is predicted here to also regulate EMT and immune genes such as *Fbln5, Cdh5, Itga4* and *Cxcr6* among others (Figure 24).

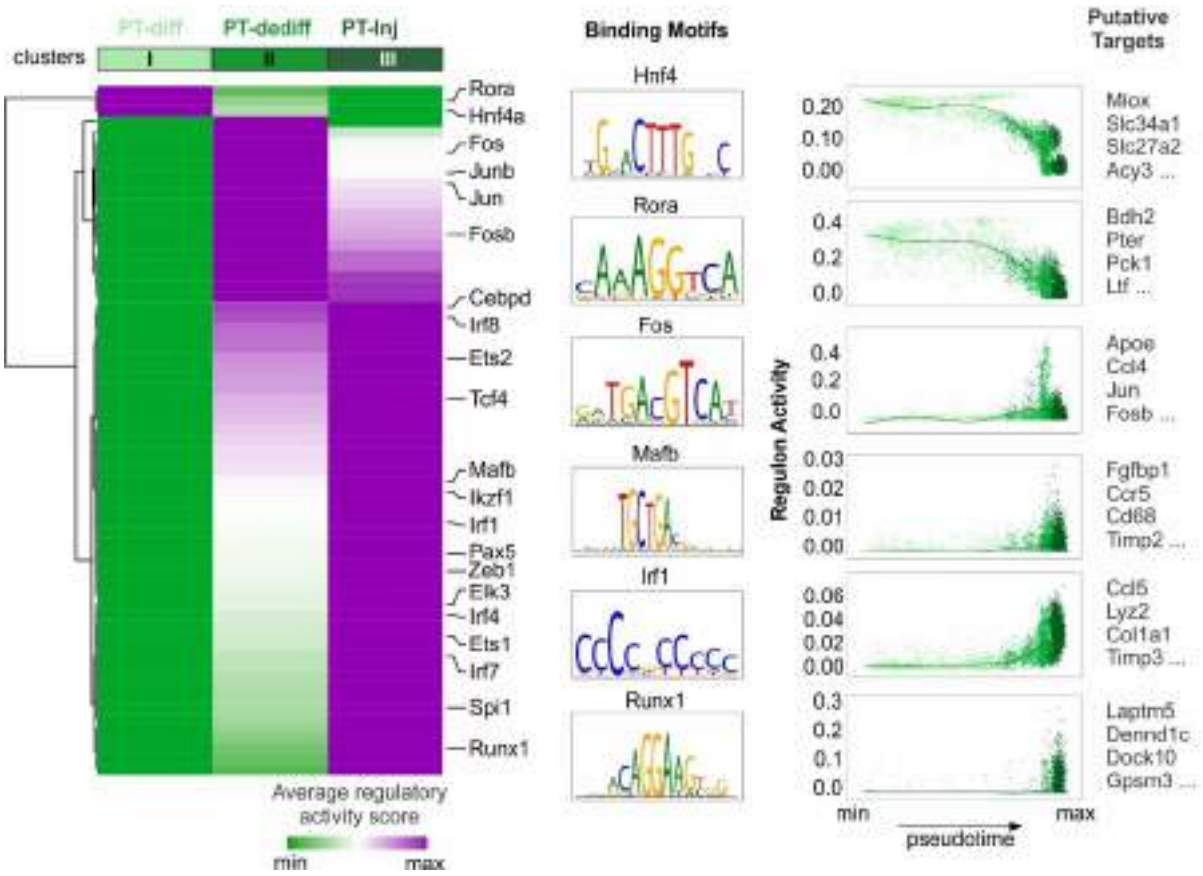


Figure 24 | Expression-based regulon analysis shows the transcription factor code associated with dedifferentiation and activation of injury/inflammatory pathways along with EMT in UUO induced fibrosis. Heatmap showing average regulatory activity of predicted regulons. Selected regulons show cell state-specific regulatory activity represented over pseudotime along with their enriched DNA binding motifs.

Overall, our analysis is fully compatible with the activation of non-invasive partial EMT programme during kidney fibrosis which concomitantly activates pathways related to inflammatory and injury responses. Furthermore, *in silico* regulon prediction indicates that gene expression changes during the partial EMT program are regulated by state-specific transcription factors.

4.4 Reactivation of two distinct EMT programmes during tumour progression in primary breast cancer

Cancer cells reactivate EMT programmes which enable them to disseminate from their primary site, intravasate into blood stream and migrate to distant organs to form metastases after extravasation (Figure 7). The reactivation of EMT in cancer cells help them to change their phenotype by providing cellular plasticity. The whole cascade of molecular pathways gets activated along with EMT activation allowing cancer cells to acquire stemness, resistance to apoptosis, immune evasion and invasive properties (Huang et al., 2022). The level of EMT activation by cancer cells is not homogenous (Nieto et al., 2016), and can be observed even at the premalignant stage (Rhim et al., 2012). Additionally, cancer stem like cells maintain a partial invasive EMT state which is essential for the metastatic colonisation (Ocaña et al., 2012a).

To capture the diversity of EMT in breast cancer, we generated scRNA-Seq libraries in a mouse model (MMTV-PyMT). These mice spontaneously develop breast carcinomas that progress to the invasive and metastatic state resembling human invasive breast cancer (Attalla et al., 2021). We tagged mammary gland progenitor cells from early embryonic stages (Van Keymeulen et al., 2015) to detect all cancer cells and discriminate them from those in the tumour microenvironment (TME) (Figure 25a,b; Youssef et al, 2024).We systematically analysed the scRNA-Seq data to study the dynamics of EMT activation, the related molecular changes, and their impact on cancer progression.

4.4.1 Single-cell transcriptomics uncovers the cellular heterogeneity of cancer cells and accessory populations during breast cancer progression

The 10X scRNA-Seq libraries were sequenced for four MMTV-PyMT WT tumours. The obtained raw reads were subjected to quality control, and we observed neither adaptor contamination nor the enrichment of overrepresented sequences. The average Phred quality score was observed to be greater than 20.

We obtained more than 1200 million reads in total with 310.45M average reads (±2.88M SD) per sample (Figure 25c). The raw read alignment reveals an average confident mapping of 91.45% (±0.46 SD) to the genome, with average

of 61.13% (±0.61 SD) mapping confidently to exonic regions per sample. The total number of cells recovered from the four different samples was 36162, averaging 9040.50 (± 475.96 SD) cells per sample. The average number of detected genes per sample is 26723.50 (± 394.92 SD; Figure 25c). The average sequence saturation level, explaining the sequencing depth, is 0.77 (± 0.05 SD). We detected an average of 366.51 (± 60.83 SD) doublets per sample (Figure 25d). Additional statistical parameters are listed in Table 3. After applying appropriate filters (see Materials and Methods), we obtained a total of 34607 high-quality cells across four WT samples, averaging 8651.75 (± 424.01 SD) cells per sample, which were used for the downstream analysis.

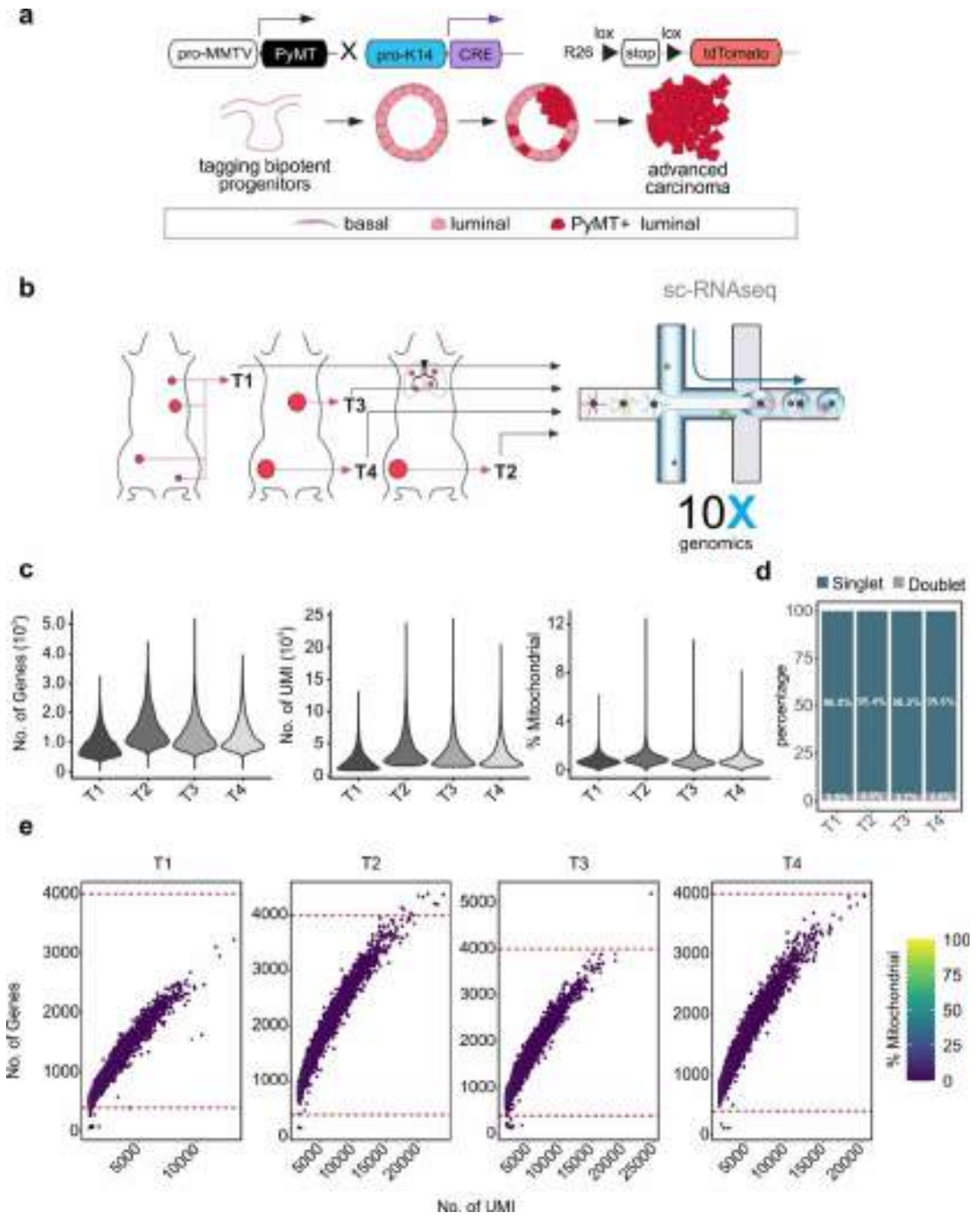


Figure 25 | Single cell transcriptome of primary breast cancer tumours in MMTV-PyMT tumours recovered high quality cells during cancer progression. (a) *in vivo* MMTV-PyMT mouse model of primary breast carcinoma. The PyMT oncogene is controlled under MMTV promoter and the cre was expressed under K14 promoter. We genetically labelled cancer using tdTomato reporter. **(b)** Illustration of strategy used for sample preparation and sequencing **(c)** Violin plot showing different QC paramters per sample: number of cells,

number of genes and percentage of mitochondrial genes. **(d)** barplot showing percentage of putative doublets and singlet predicted using scrublet (Wolock et al., 2019). **(e)** Scatterplot showing different QC paramters (number of detected genes, number of UMIs and percentage of mitochondrial genes) together to determine the captured single cell quality. Less number of genes, high numbers of UMI and high % mitochondrial genes could be dying cells whereas very high number of genes suggest the detected cell could be a multiplet. Dotted lines indicate the cut-offs.

	T1	**T2**	**T3**	**T4**	**Mean**	**std**
Estimated Number of Cells	8724	9580	8434	9424	9040.5	475.96
Total Number of Genes	26062	26791	27072	26969	26723.5	394.92
Total Number of Reads	313M	313M	306M	308M	310M	2.88M
Mean Reads per Cell	35878	32730	36377	32731	34429	1707.64
Median Genes per Cell	842	1321	1142	1079	1096	171.42
Median UMI Counts per Cell	1922	3281	2854	2589	2661.5	493.19
Reads Mapped to Genome	90.8	91.5	92.1	91.4	91.45	0.46
Confident mapping on Exons	61.8	60.8	61.6	60.3	61.13	0.61
Sequencing Saturation	0.854	0.713	0.751	0.742	0.77	0.05

Table 3: Raw data and alignment statistics for scRNA-Seq libraries prepared for BC tumour samples in MMTV-PyMT mouse model. Std=Standard Deviation

Next, to create a collective transcriptional space for the different WT samples we performed anchor-based integration as performed for the kidney scRNA-Seq data (Figure 26a). The Shared Nearest Neighbor (SNN) clustering resulted in 5 main clusters which covered cellular heterogeneity (Figure 26b). The global transcriptional changes associated with each cluster are represented as heatmap shown in Figure 26b (right panel).

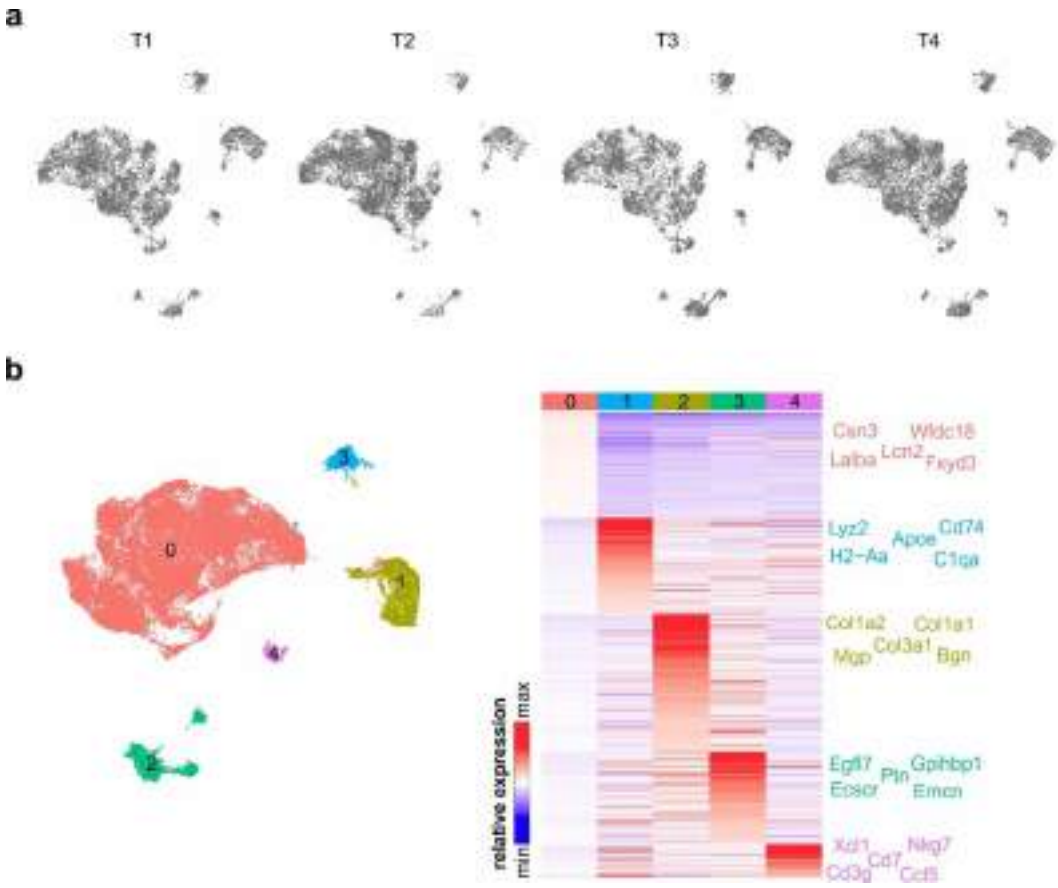


Figure 26 | Global expression profile of integrated tumour samples resulted in 5 different cellular clusters explaining cellular heterogeneity and global transcriptional changes. (a) Uniform Manifold Approximation and Projection (UMAP) showing integration of different tumour samples. **(b)** UMAP represents 5 distinct cellular clusters obtained using shared nearest neighbor (SNN) modularity explaining cellular heterogeneity during primary breast cancer progression (left). Heatmap representing average relative expression of differential expressed genes in each cluster (average log2FC > 0.25 and FDR < 0.05; right).

Next, we identified the top differentially expressed genes within each cluster based on their specificities to annotate the clusters as major populations (Figure 27). Additionally, as we genetically labelled cancer cells (CC) in our mouse model with the tdTomato reporter (Figure 26a), we could easily discriminate them from accessory populations present in the tumour microenvironment. Moreover, CC express specific markers such as *Csn3*, *Wfdc18*, *Krt18*, and *Lcn2* (Figure 27).

We also identified the cells in the tumour microenvironment, including cancer-associated fibroblasts (CAFs) by the expression of *Mfap5, Bgn, Col3a1*, and *Col1a2* (Figure 27). Angiogenesis, a hallmark of tumour progression, was evident in our scRNA-Seq analysis in a distinct cluster of endothelial cells (EC) expressing *Flt1, Cdh5,* and *Egfl1* among others (Figure 27). We also identified both myeloid (MC) and lymphoid (LC) cells expressing *bona fide* markers such as *H2-Aa, Apoe, Cd74* for myeloid cells and *Cd3g, Cd7, Nkg7* for lymphoid cells, respectively (Figure 27).

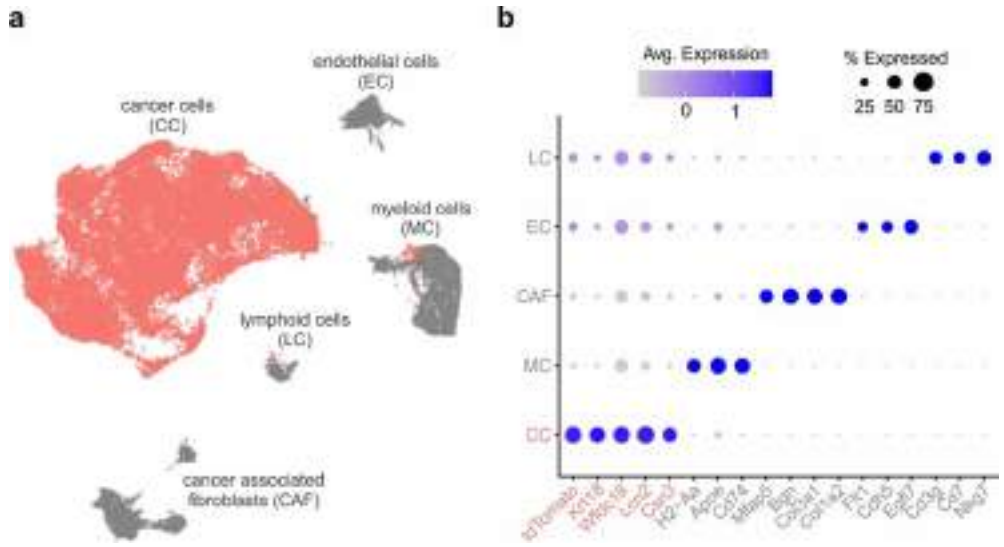


Figure 27 | Systematic analysis of scRNA-Seq data of MMTV-PyMT tumours reveals cellular heterogeneity mainly involving cancer cells and those in the tumour microenvironment (TME). (a) UMAP depicting 5 major clusters representing different cell types annotated based on *bona fide* markers (left). **(b)** Dotplot representing top5 differentially expressed genes in each clusters used for the cell type annotation in a (right).

Examples of the expression at single cell level of specific markers identifying the different populations are represented over UMAPs in Figure 28.

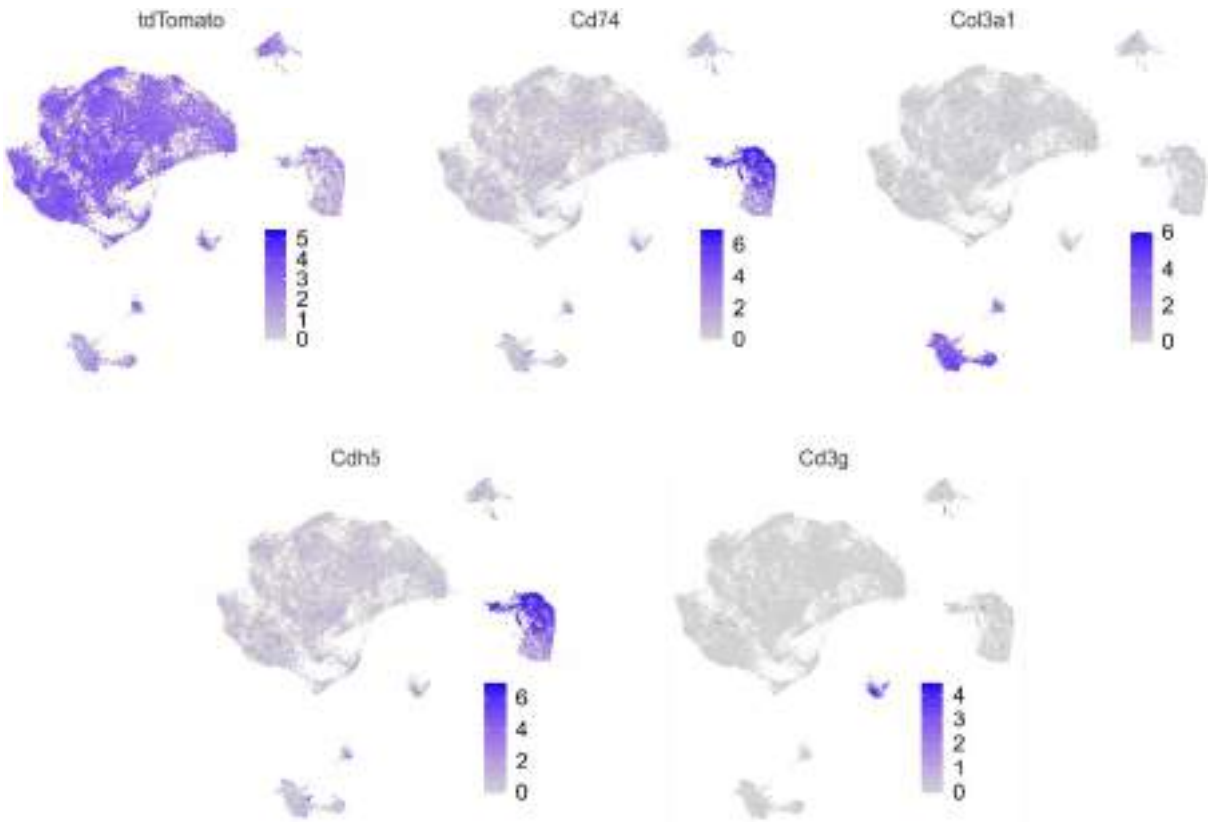


Figure 28 | UMAPs representing the relative expression of selected *bona fide* markers specific for the different cell populations.

4.4.2 Distinct differentiation and EMT states observed in during primary BC progression

To define the EMT programme and dissect the molecular changes, we subset only cancer cells from major population. All the 18621 cancer cells (average 4655.25 with ±932.05 SD per sample) were re-clustered, which resulted in 17 different clusters (Figure 29a). Relatively high and uniform expression of tdTomato in all selected cells confirms their cancer cell nature and confirms that our subset does not contain accessory cells (Figure 29b).

We did not observe sample-specific bias in the number of cells across the detected clusters, indicating the reproducibility of cancer cell states across different samples (Figure 29c).

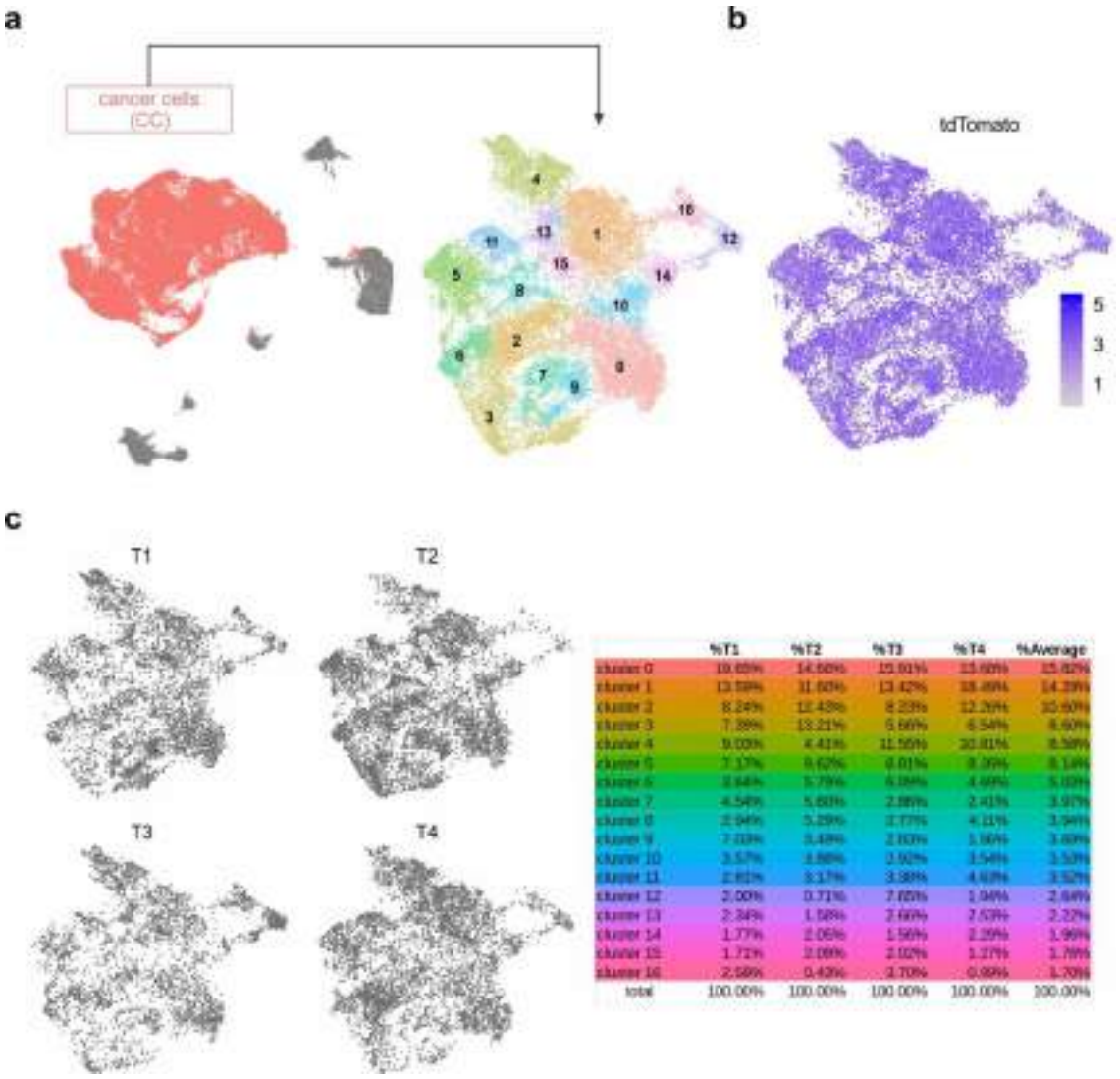


Figure 29 | Cancer cell subset of MMTV-PyMT tumours has a similar distribution in the different samples. (a) Uniform Manifold Approximation and Projection (UMAP) showing the major populations (left) and reclustering of the cancer cells. **(b)** UMAP showing the relative expression of tdTomato in the cancer cells. **(c)** UMAPs showing the cells from each individual tumor sample, and percentage of cells associated to each specific cluster in each tumor sample. Both representations all single cell distribution per sample (left) and a table showing cluster wise percentage of cells per sample to show the reproducibility.

Next, we used luminal and basal gene signatures for mammary epithelial cells (Bach et al., 2017; Pal et al., 2017; Pervolarakis et al., 2020) to identify cancer cell states based on the expression of different markers associated with differentiation states in cancer cells (Figure 30a). The MMTV-PyMT tumours have luminal origin and thus, as expected, we observed that the majority of cancer cells express genes compatible with a Luminal Alveolar (LA) phenotype (71.9%; Figure 30a).

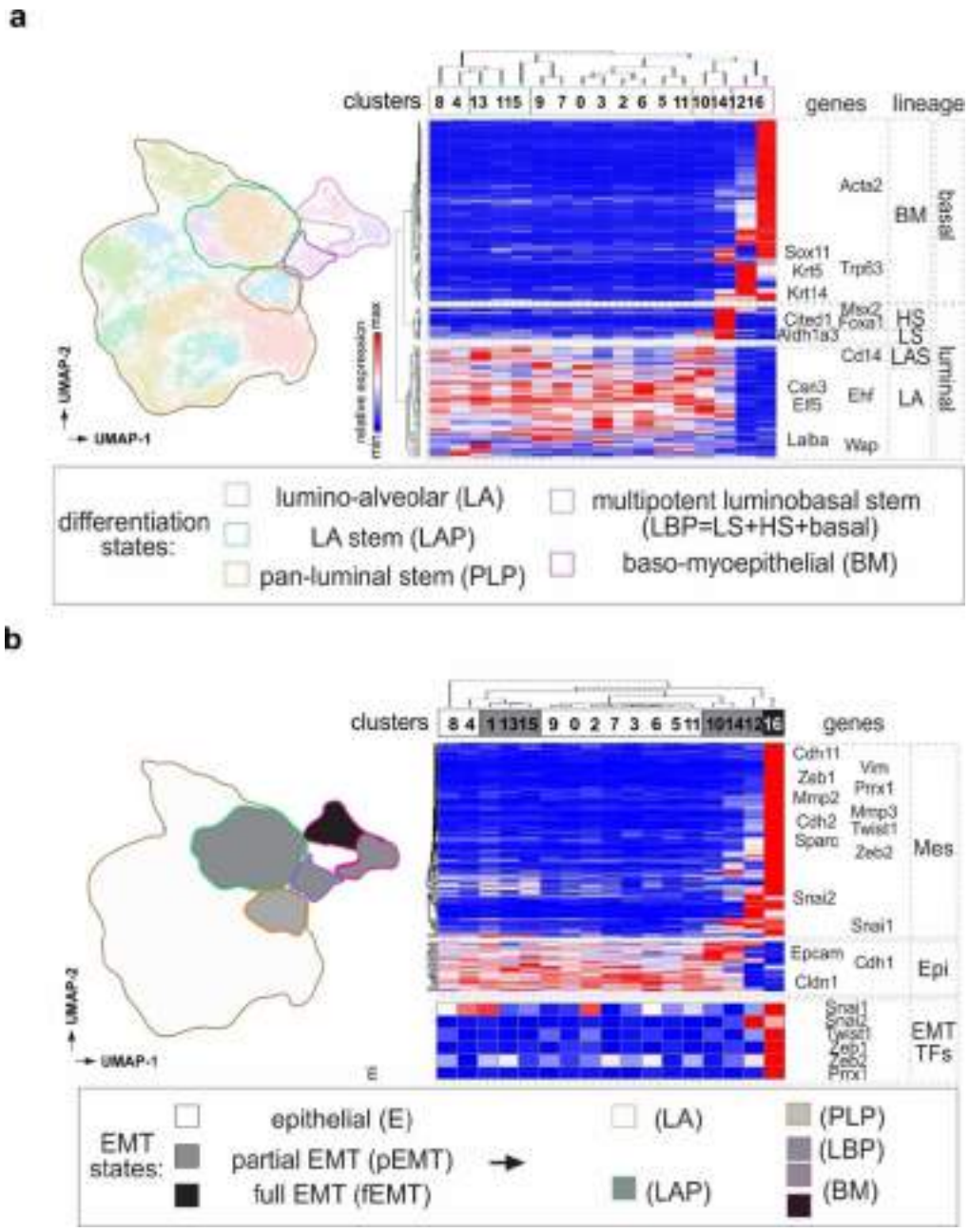


Figure 30 | The cancer cells clusters reveal different differentiation states concomitant with the activation of EMT. Legend in next page.

Interestingly, we observed that some cancer cell clusters are transcriptionally similar to different progenitor states. Clusters 1, 15, and 13 show higher expression level of the markers associated with the Luminal Alveolar Stem/Progenitor state (LAP; Figure 30a; Pal et al., 2017; Shehata et al., 2012; Yeo et al., 2020). Cluster 10 shows a transcriptional programme compatible with the Pan Luminal Stem/Progenitor state (PLP; Figure 30a; Bach et al., 2017; Charafe-Jauffret et al., 2009; Ginestier et al., 2007). We also identified a hybrid state represented by cluster 14, jointly expressing markers of PLP, luminal Hormone Sensing (HS), and Baso-Myoepithelial (BM) phenotypes (Ginestier et al., 2007; Van Keymeulen et al., 2015) equivalent to a developmental progenitor-like state (Figure 30a; Giraddi et al., 2018; Kaufman et al., 2016; Thong et al., 2020; Youssef et al., 2012). Finally, clusters 12 and 16 align with a basal transcriptional programme progressing towards an invasive phenotype (Figure 30a; Kevin J. Cheung et al., 2013).

Figure 30 | The cancer cells clusters reveal different differentiation states concomitant with the activation of EMT. (a) Hierarchical clustering of cancer cell (CC) clusters (Figure 29a) based on the expression of differentiation markers (luminal or basal/myoepithelial). Color scale represents average log2 fold change calculated for each gene in a target cluster compared to all other CC clusters. Right panel, cluster representation with associated colours according to differentiation states. BM: baso-myoepithelial genes; HS: luminal hormone-sensing genes; LS: luminal stem/progenitor genes; LAS: lumino-alveolar stem/progenitor genes; LA: lumino-alveolar genes; LAP: lumino-alveolar stem/progenitor state; PLP: pan-luminal stem/Progenitor state; LBP: lumino-basal stem/progenitor state. **(b)** Hierarchical clustering of CC based on the expression of epithelial and mesenchymal genes. The lower panel shows differentiation states colour coded as in a, and clusters with EMT activation shown in grey scale.

The observed phenotypic changes induced by reprogramming and dedifferentiation of segregated tumour cells are accompanied by the activation of the EMT programme. To further investigate the implementation of EMT and the putative correlation between states along the E/M spectrum and different dedifferentiation states, we performed similar expression-based analysis using epithelial and mesenchymal genes. We observed that the majority of cancer cells exhibit higher expression of epithelial genes which is compatible with the LA state in differentiation spectrum (Figure 30b), indicating that hose cells had not engaged into the EMT programme.

Clusters 1, 13, 15, 10 and 14 represent a partial EMT state, as cells express both epithelial and mesenchymal markers. These clusters show a progenitor-like differentiation state. In contrast, cluster12 mainly expresses mesenchymal genes with minimal expression of epithelial markers. Finally, cluster 16 shows a complete repression of epithelial genes and activation of mesenchymal markers, indicative of a full EMT state (Figure 27b). Both clusters 12 and 16 represent Baso-Myoepithelial (BM) states. The activation of EMT is also concomitant with the systematic activation of EMT-TFs as observed in cancer cell lines (Figure 10c).

Taken together, the reprogramming of tumour cells during primary breast cancer progression is concurrent with EMT activation. This provides cellular plasticity, allowing the cells to acquire invasiveness, escape the primary site, and form metastases. Our scRNA-Seq data suggests that the MMTV-PyMT model accurately represents the cellular states observed in primary tumours in the patients.

Examples of the expression at single cell level of specific markers identifying the different cancer cell states are represented over UMAPs in Figure 31.

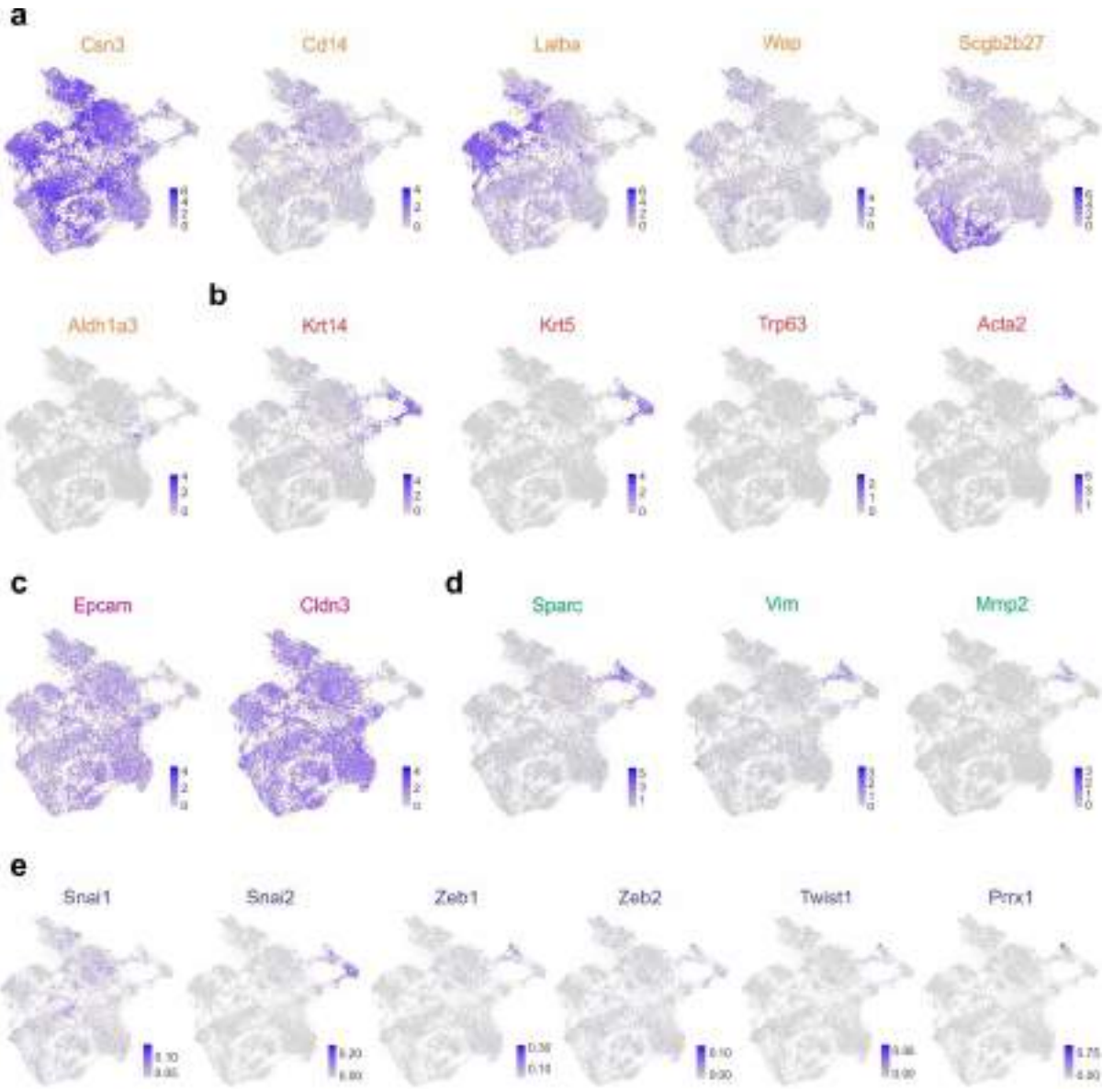


Figure 31 | UMAP representing expression of luminal (a), basal/myoepithelial (b), epithelial (c), mesenchymal markers (d), and EMT-TFs (e) in cancer cells subset at single cell level.

4.4.3 Two different EMT programmes are simultaneously activated in segregated cancer cell populations

We have analysed both differentiation and EMT status using a supervised approach (using known gene signatures). Both supervised and unsupervised approaches have their respective advantages and disadvantages in data analysis. Supervised methods are focused on specific, predefined information, making the results straightforward to interpret and easy to understand. However, these methods are limited to the predefined information, which can introduce bias into the downstream analysis and potentially overlook other relevant information. On the other hand, unsupervised methods analyse data by identifying patterns without prior knowledge. The major drawback of this approach is sometimes it generates spurious and less reliable results, making it harder to interpret. From all this, it is clear that combining supervised and unsupervised approaches can provide a more comprehensive understanding of the data. While ensuring that the analysis remains focused on key questions of interest, adding unsupervised methods helps to identify unexpected patterns and insights. Thus, we have used unsupervised approaches to decipher the connection between different CC clusters based on the change in their global gene expression pattern.

To understand the cellular transitions within the segregated tumor cells, we performed PAGA analysis similar to those done previously for the neural crest and the kidney scRNA-Seq data. In the PAGA connectivity map (Figure 32a,b), we observed that cells, progressing from the LA cells (clusters 5 and 11) bifurcate into two different paths upon dedifferentiation (differentiation states defined using supervised approach in Figure 30). One path is defined by LAP-like states whereas the other progresses towards the BM state passing through PLP- and LBP- like states (Figure 32b). In addition, our PAGA analysis shows that several other clusters (e.g. clusters 0,7 and 9) are also connected with the dedifferentiating clusters. However, these clusters belong to bulk of the tumour and do not activate EMT programme (Figure 32b).

Based on our extensive combined supervised and unsupervised analysis, we hypothesized that during BC progression a supopulation of tumour cells concomitantly undergo dedifferentiation and reactivation of EMT, resulting in different plastic cell states. To support this data-driven hypothesis of cellular

transition in the PAGA connectivity map we inferred the directionality of cells using RNA velocity approach. This analysis confirms that LA cells dedifferentiate and that they bifurcate into two different EMT programmes (Figure 32c), one composed of clusters 10, 14, 12 and 16 which progresses to give rise BM phenotype and we call EMT-T1 and another one leading to the PLP-like states and comprising clusters 13, 15 and 1, that we call EMT-T2. Additionally, the enrichment of hallmark EMT gene signature shows varying levels of EMT activation in these two different EMT trajectories (Figure 32d). Overall, our data suggests that, during primary breast cancer progression, tumour cells undergo reprogramming through dedifferentiation and concomitant reactivation of EMT at varying levels across the E/M spectrum.

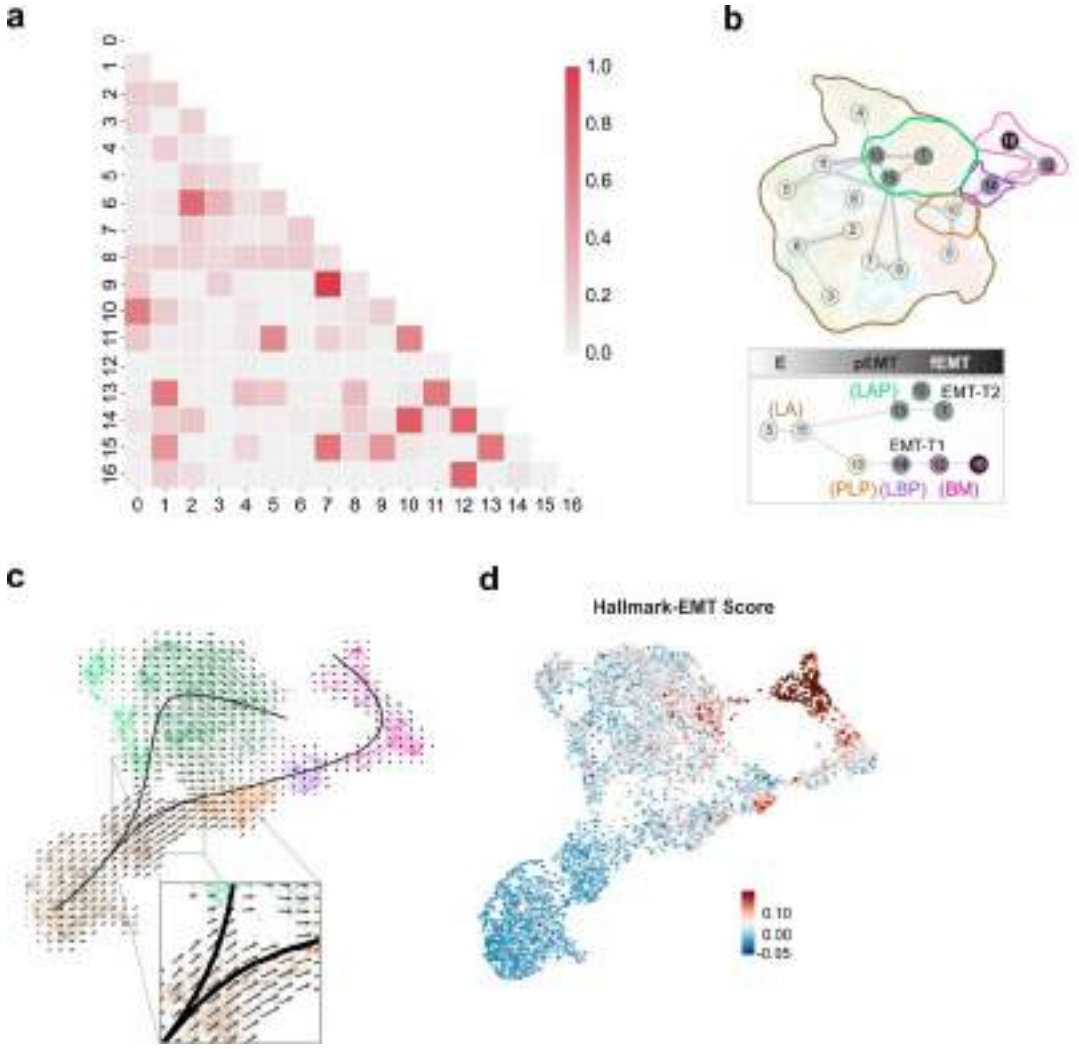


Figure 32 | The dedifferentiating tumour cells bifurcate into two distinct EMT programs during primary breast cancer progression. (a) Connectivity matrix obtained by PAGA (Partition-based graph abstraction) analysis showing transcription similarities between the different cell clusters identified. **(b)** The connectivity map of transcriptionally similar clusters was built using partition-based graph abstraction (PAGA) algorithm. The PAGA connectivity map was super imposed on UMAP (upper panel). The clusters undergoing EMT are circled with different colors and the scheme shows the bifurcation of EMT states into two distinct trajectories **(c)** Vectors showing the directionality of cells in the transition. They are calculated using the RNA velocity analysis superimposed on the UMAP. The solid lines represent the smoothed principle curve fitted using slingshot. **(d)** UMAP representing the enrichment score of EMT hallmark gene signature.

4.4.4 The two different EMT trajectories regulate dissemination

Next, to identify molecular pathways and characterise the two EMT programmes and their progression along the E/M spectrum, we also used similar approach to those in the neural crest and the kidney. First, pseudotime was calculated for the two trajectories by selecting root cells which show higher expression of Lalba, a mammary gland epithelial differentiation marker (Figure 33a-c). Next, we clustered the genes based on their expression in a pseudotemporally organized manner and categorized them into different groups for individual trajectories (Figure 33d). We observed that cells in the two EMT trajectories follow paths associated with different transcriptional programmes and phenotypic transitions.

In EMT-T1, cancer cells progressively lose lumino-alveolar differentiation genes like *Csn3, Lalba, Wap* (group 1: clusters 5 and 11) and evolve towards a stem/progenitor-like state (group 2: cluster 10) characterised by the expression of pro-stemness genes such as *Aldh1a3* and *Ndrg1* (Cao et al., 2019). Group 3 (cluster 14) is compatible with a partial EMT state, with less expression of the epithelial markers *Epcam, Cldn3* and *Cldn7* among others, activating pluripotency markers such as *Wnt9a, Bmp1, Id1, Id3,* and *Igf1* together with markers of the embryonic mammary gland and basal-like signatures while progressing towards a full EMT state, as assessed by high expression of *Vim* and *Cdh2* (Cluster 16). An invasion signature is already evident in cluster 14 (group 3), with cells expressing genes that regulate cell migration and cytoskeleton remodelling (*Tnc, Gsn, Palld, Cnn2, Tpm1, Tpm2* and *Mmp14*). The invasion signature is amplified in clusters 12 and 16 (group 4), with prominent expression of additional invasion genes including cytoskeleton regulators (*Mylk, Tagln* and *Pdpn*), guidance ligands and receptors like *Sema5a* and *Nrp2*), and microenvironmental modulators such as metalloproteinases and Lysyl oxidases (*Mmp2, Mmp3* and *LoxL1*).

In EMT-T2, the lumino-alveolar epithelial phenotype of cluster 11 progresses to a partial EMT phenotype in cells of clusters 1, 13 and 15, while still maintaining expression of epithelial genes and activating some mesenchymal genes shared with the EMT-T1 trajectory (*Sparc, Postn, S100a4*), but without progressing to full EMT. EMT-T2 is highly enriched in injury response genes (*Egr1, Jun, Junb, Fos, Fosb,* and *Lcn2*) and inflammatory regulators. These include genes encoding

secreted factors (*Spp1*), components of the TNF-α/interferon and NF-kB pathways (*Nfkbia, Ccrl2,* and *Notch2*), and inflammatory biomarkers, including serum amyloid A proteins (*Saa2, Saa1*) and Lymphocyte antigen-6 family genes (*Ly6k* and *Ly6d*). Additional enrichment for pro-inflammatory genes is observed in cluster1 (group 2), the most prominent cluster in this branch, including additional interferon regulators and downstream targets genes (e.g. Irf7, Ifitm3, Ifitm2, and *Cxcl16*). In addition, EMT-T2 is enriched for pro-fibrotic genes such as tissue inhibitors of metalloproteinases (*Timp2, Timp3* and *Timp1*). All of this indicates that, in EMT-T2, the transition to a partial EMT is concomitant with the acquisition of an inflammatory and pro-fibrotic phenotype.

Additionally, the initiation of the Hallmark EMT signature in cluster 14 concurs with the detection of *Snail2* in addition to *Snail1* and *Twist1* transcripts in cluster 12 (Figure 33e). The progression towards more advanced EMT states is coupled to an increase in the expression of *Zeb1* and *Prrx1* (Figure 33e). In contrast to EMT-T1, among the EMT-TFs, only *Snail1* is detected in the clusters belonging to EMT-T2 (Figure 33e).

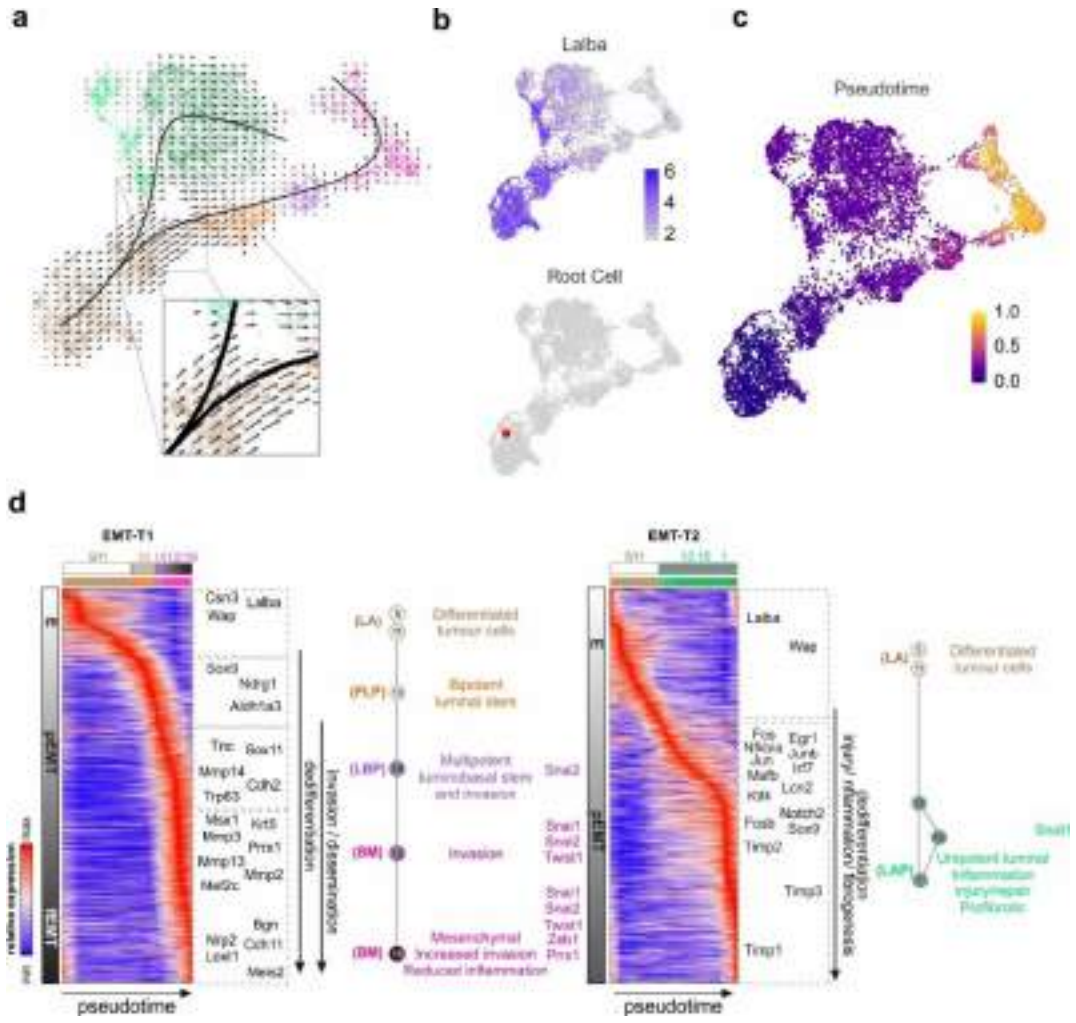


Figure 33 | The reactivation of two distinct EMT programmes during primary breast cancer progression. **(a)** UMAP showing clusters belongs to EMT trajectories. The arrows indicate the cell transition in EMT trajectory calculated using RNA- velocity analysis **(b)** UMAP embedding showing relative expression of marker (Lalba) of differentiation (left). UMAP embedding showing selected root cells in red (right). **(c)** UMAP embedding with pseudotime representing the pseudo progression of cells in the trajectories. **(d)** Heatmaps showing clustering of genes based on their expression in the cells arranged over pseudotime. The clustered genes were categorized into different groups in each trajectories based on their higher expression in the trajectory clusters. **(e)** Dotplot showing pathways and GO enrichment in the two EMT trajectories, embryonic and adult-like, respectively related to development/invasion and inflammation for the different groups in EMT-T1 and EMT-T2. **(f)** UMAP embedding representing BC-PINGs and inflammation score.

Pathway enrichment analysis in EMT-T1 is associated with cell migration and enriched in developmental pathways. EMT-T2 shows significant enrichment of pathways associated with inflammation and immune responses (Figure 34a). Pathway enrichment is accompanied by enrichment of invasion and inflammation gene signatures for EMT-T1 and EMT-T2, respectively, as shown in Figure 34b. These trajectories are reminiscent of the invasive embryonic EMT programme observed during neural crest development or with the EMT programme we have observed in the adult during fibrosis development, respectively.

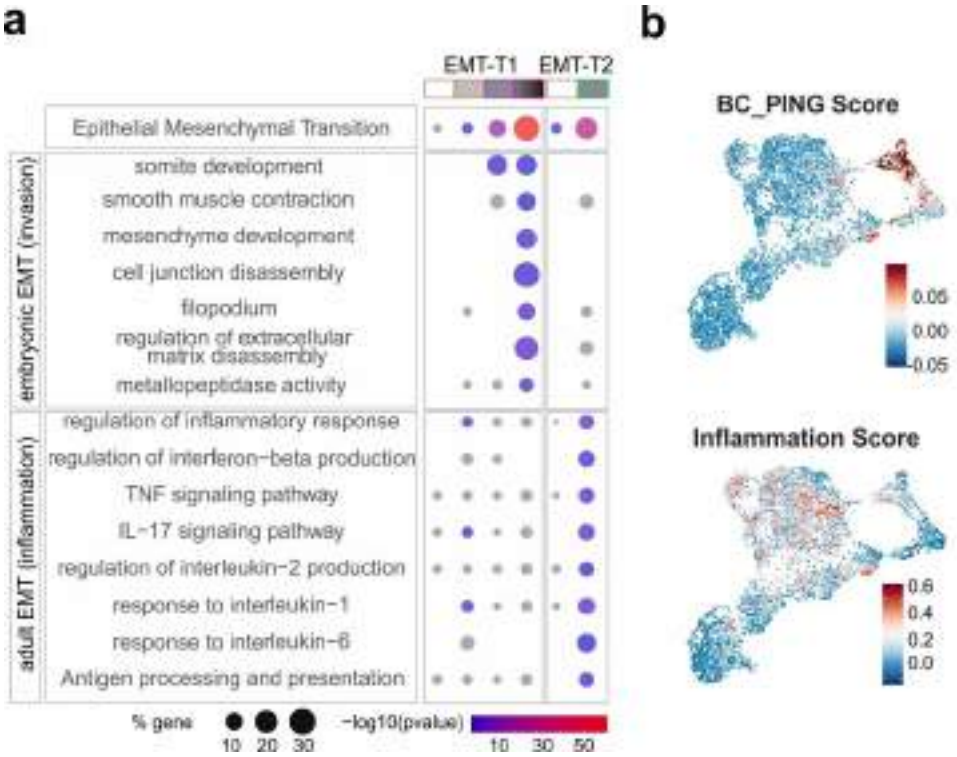


Figure 34 | The two EMT trajectories are enriched with developmental invasion pathways or adult inflammatory responses to injury. (a) Dot plot showing pathways and GO enrichment in the two EMT trajectories, embryonic-like and adult-like, respectively related to development/invasion and inflammation for the different groups in EMT-T1 and EMT-T2. **(b)** UMAP embedding representing BC-PINGs and inflammation

4.4.5 The scRNA-Seq based EMT trajectories shows spatial organization and significantly enriched in human TNBC samples

To validate the activation of the two distinct EMT programmes identified through our systematic analysis of scRNA-Seq data, *in vitro* invasion assays were performed in the lab (Figure 35a). Tumouroids were generated in 3D collagen matrices, allowing them to grow dynamically in an environment that closely mimics *in vivo* conditions. Immunostaining of the EMT-T1 specific and invasive marker keratin 14 (Krt14), shows that invasive cells are aggregated at the edge of the tumouroids, whereas the cells with higher expression of the EMT-T2 and injury marker, Jun, are located at the centre of the tumouroids (Figure 35b,c). This indicates that the two different EMT programmes segregate into different populations that are in different areas when grown *in vitro* as tumoroids.

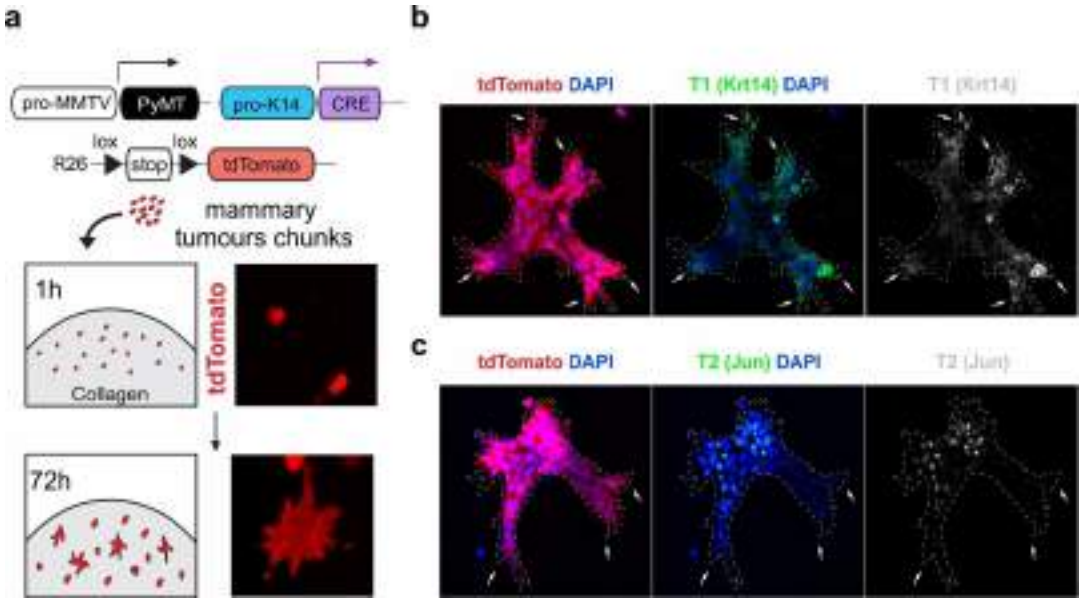


Figure 35 | scRNA-Seq based trajectories are spatially organized in segregated tumour cell population and show significant enrichment in the TNBC tumour in human BC patients. (a) illustration depicting experimental design for generating tumouroids for invasion assay. **(b)** Cells expressing Krt14, an EMT-T1 specific marker, are enriched at the invasive edges (arrows). **(c)** Cells expressing high levels of the EMT-T2 specific marker Jun are enriched in central areas of the cultured PyMT tumouroids.

Next, to check the implication of the EMT trajectories in BC progression in patients, we performed enrichment analysis using gene signature derived from each trajectory clusters (top20 differentially expressed genes). We observed that all the EMT clusters are represented in human BC samples. Interestingly, we observed a higher enrichment in TNBC samples (Figure 36; Chung et al., 2017). Thus, the reactivation of EMT programmes in segregated BC tumour cells of luminal origin makes them progress to the aggressive basal-like phenotype.

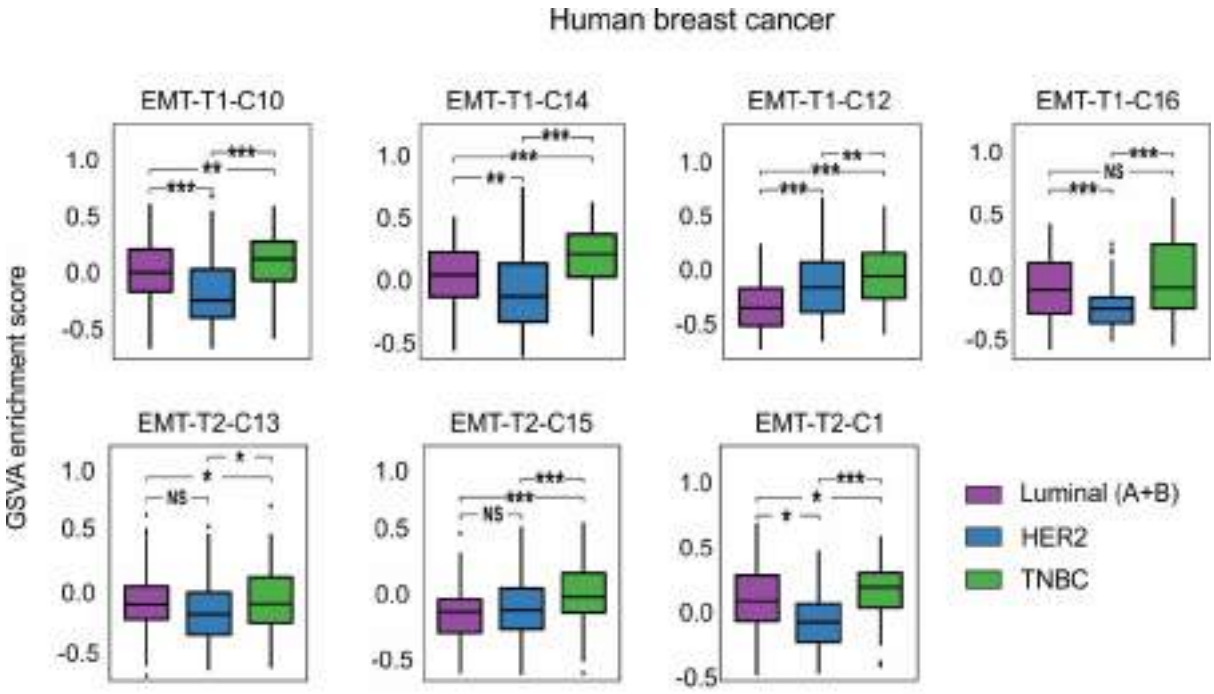


Figure 36 | Boxplot showing enrichment of EMT-T1 and EMT-T2 clusters in human breast cancer. The score represents Gene Set Variation Analysis (GSVA) calculated for EMT-T1 and EMT-T2 clusters in the different breast cancer subtypes: Luminal (A and B), HER2 and Triple Negative Breast Cancer (TNBC).

4.4.6 Identification of potential regulators in the two EMT trajectories activated during BC progression

Like for the neural crest and kidney, we performed regulon prediction to understand the putative transcriptional regulation in EMT trajectories during BC progression, and we predicted a total of 108 potential regulators. As previously described, EMT-T1 and EMT-T2 involve the activation of distinct gene modules. In the case of EMT-T1, the modules are associated with invasive properties, while in EMT-T2 they are linked to inflammation and immune response. Similarly, the EMT trajectories have differential regulatory modules, further supporting the differences between these EMT programmes.

The transcription factors (TFs) *Xbp1* and *Gata2*, known for their crucial roles in cell differentiation (Castaño et al., 2019; Onodera et al., 2016; Todd et al., 2009), exhibit higher regulatory activity in cluster 11, putatively regulating the transcription of genes encoding differentiation markers such as *Wap, Aldoa, Cldn10,* and *Etv1*. In EMT-T2 (Clusters 1, 13, and 15), we observed higher activity of genes encoding regulators like *Fos, Jun, Fosb, Runx1*, and *Mafb*, which are known to play significant roles in inflammatory and immune responses. These TFs show relatively higher activity in EMT-T2 clusters compared to EMT-T1 clusters. However, clusters 10 and 14 (early EMT-T1) exhibit a reasonable level of regulatory activity, indicating the activation of a partial EMT programme. The EMT-T2 regulators modulate the expression of genes such as *Notch, Snai1, Fos, Jun,* and *Klf4*. In contrast, the late EMT-T1 clusters (clusters 12 and 16) activate the transcription of a new set of genes encoding regulators such as *Prrx1, Msx1, Trp63, Twist1* etc., associated with invasion. Similar to the regulons predicted in the neural crest trajectory, these regulators exhibit higher activity in the full mesenchymal state, including gene modules composed of *Vim, Mmp3, Mmp14, Pdgfra, Col1a1, Msx1, Fbn2, Bgn* etc. (Figure 37).

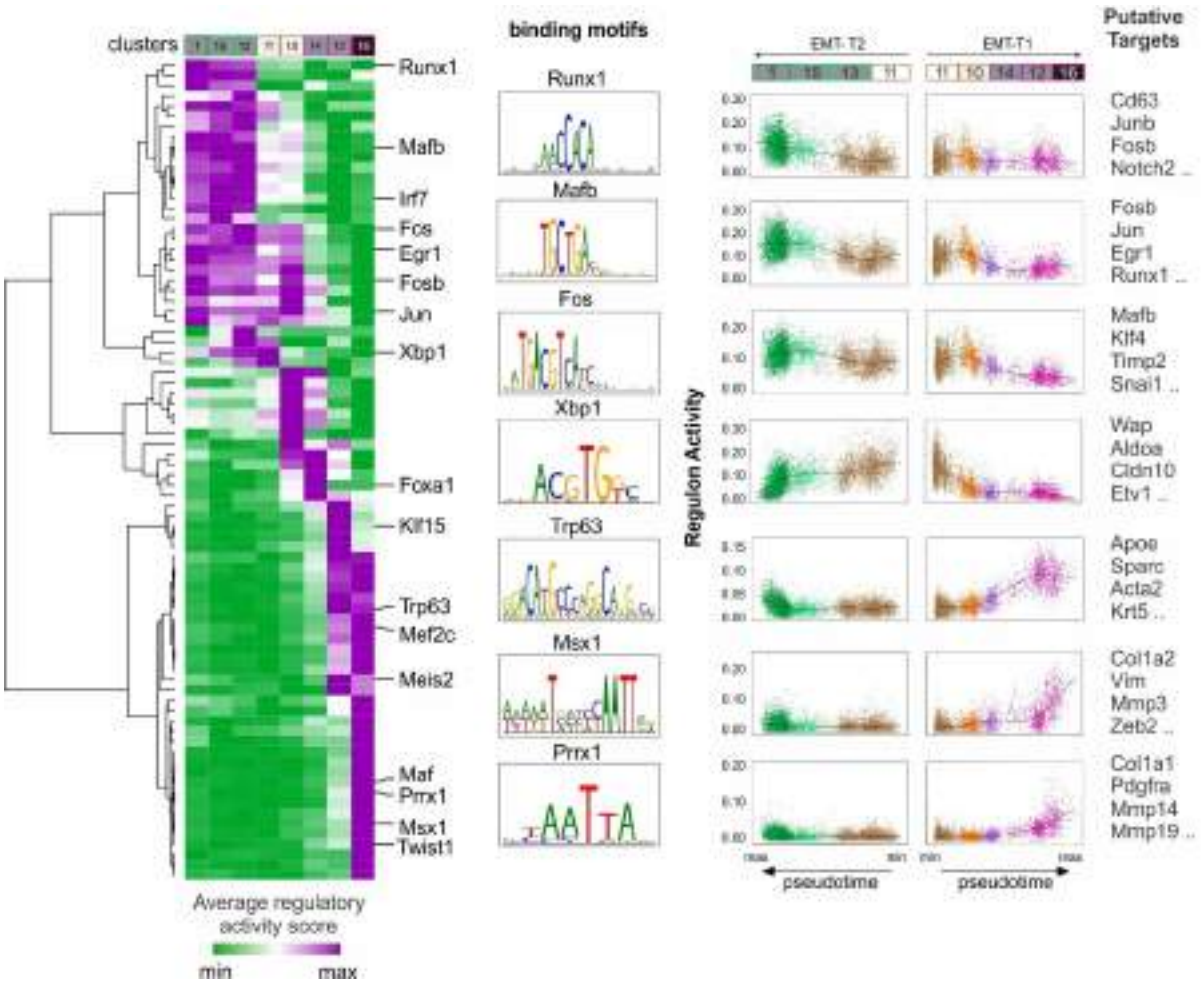


Figure 37 | Expression-based regulon analysis shows the transcription factor code associated with embryonic/invasive EMT-T1 and adult/inflammatory EMT-T2 in segregated cancer cells. Heatmap showing average regulatory activity of predicted regulons. Selected regulons show cell state-specific regulatory activity in EMT-T2 (left) and EMT-T1 (right) represented over pseudotime along with their enriched DNA binding motifs. The smooth line was fitted using a generalized additive regression model.

4.5 *In silico* perturbation analysis of the EMT-TFs shows significant remodelling of the EMT trajectories, predictive of altered BC progression

Recently, the development of advanced computational frameworks such as SCENIC+ and CellOracle (Bravo González-Blas et al., 2023; Kamimoto et al., 2023) make use of machine learning algorithms and statistical methods to predict the impact of TF perturbation in single cell RNA-Seq data. The major requirements for this type of analysis are the scRNA-Seq data (continuous trajectories are recommended) and a list of TF-target genes (Kamimoto et al., 2023). The TF-target gene list called base gene regulatory network (baseGRN) can be derived from literature survey, public bulk/single cell ChIP/ATAC-Seq data, predicted from scRNA-Seq data (other GRN methods) etc. However, for robust predictions of TF perturbations, it is recommended to use the baseGRN derived from scATAC-Seq data of the same tissue used to generate scRNA-Seq trajectories (Kamimoto et al., 2023). Therefore, we generated our own scATAC-Seq libraries for the WT tumours in the MMTV-PyMT mouse model which was used to generate scRNA-Seq libraries. These scATAC-Seq libraries were systematically analysed using state-of-the-art methods to derive a baseGRN to be further used for the *in silico* TF perturbation analysis to predict the impact on EMT-dependent BC progression.

4.5.1 scATAC-Seq profile captures chromatin accessibility and explains the cellular heterogeneity in primary BC tumours

Figure 38a,b shows the model used and the protocol to obtain libraries, similar to that shown in Figure 25 for scRNA seq experiments. We obtained 10030 cells in total for two WT samples with more than 517 million sequencing reads. On average, more than 93% (±0.65 SD) of reads show confident mapping to the genome per sample. Furthermore, there are two major approaches for scATAC-Seq data analysis, one involves binning the genome into equal chunks, and the other involves peak calling. We used the popular peak calling approach in our pipeline and obtained 207559 (±210 SD) average peaks per sample. The fragments obtained in ATAC-Seq assay have strong enrichment around transcription start sites (TSS) regions and they can be used as a QC parameter for ATAC. In our scATAC-Seq libraires, 10.11 (±0.29 SD) is an average TSS

enrichment score per sample. Additional important statistical parameters are listed in Table 4.

Before proceeding further with the downstream analysis, we removed bad quality cells based on the recommended QC parameters for the scATAC-Seq data (Stuart et al., 2021). Briefly, we used nucleosome signal, percentage of reads in peaks, number of fragments in peaks and TSS enrichment to define a bad quality cell (Figure 38c). Additionally, we detected putative doublets from our scATAC-Seq libraries and found 713 (±42 SD) average number of doublets per sample (Figure 38d) and a good TSS enrichment score (Figure 38e). Finally, upon applying filtering criterion we obtained 7723 total number of high quality cells which we used for the downstream analysis. Cell-associated barcodes are expected to have a large number of fragments per barcode and a high percentage of fragments overlapping peaks. Conversely, non-cell-associated barcodes typically exhibit a small number of fragments per barcode and a low percentage of fragments overlapping peaks. In an ideal sample, there should be a clear separation between cell barcodes and non-cell barcodes, with cell barcodes showing high fragment counts and peak overlaps, and non-cell barcodes showing low counts, located at opposite ends of the distribution (Figure 38f).

	T1	T2	Mean	std
Estimated number of cells	5400	4630	5015	385
Sequenced read pairs	263.84 M	253.8M	258.82 M	5.01M
Confidently mapped read pairs	94.60%	93.30%	93.95	0.65
Number of peaks	207349	207769	207559	210
Median high-quality fragments per cell	24781	25235	25008	227
TSS enrichment score	10.4	9.82	10.11	0.29
Fraction of high-quality fragments in cells	98.10%	97.50%	97.8	0.3
Fraction of transposition events in peaks in cells	70.60%	67.50%	69.05	1.55
Fraction of high-quality fragments overlapping TSS	35.40%	33.10%	34.25	1.15
Fraction of high-quality fragments overlapping peaks	72.60%	69.80%	71.2	1.4
Fragments in nucleosome-free regions	67.20%	63.40%	65.3	1.9
Fragments flanking a single nucleosome	30.80%	33.00%	31.9	1.1

Table 4: Raw data and alignment statistics for scATAC-Seq libraries prepared for BC tumour samples in MMTV-PyMT mouse model. Std=Standard Deviation

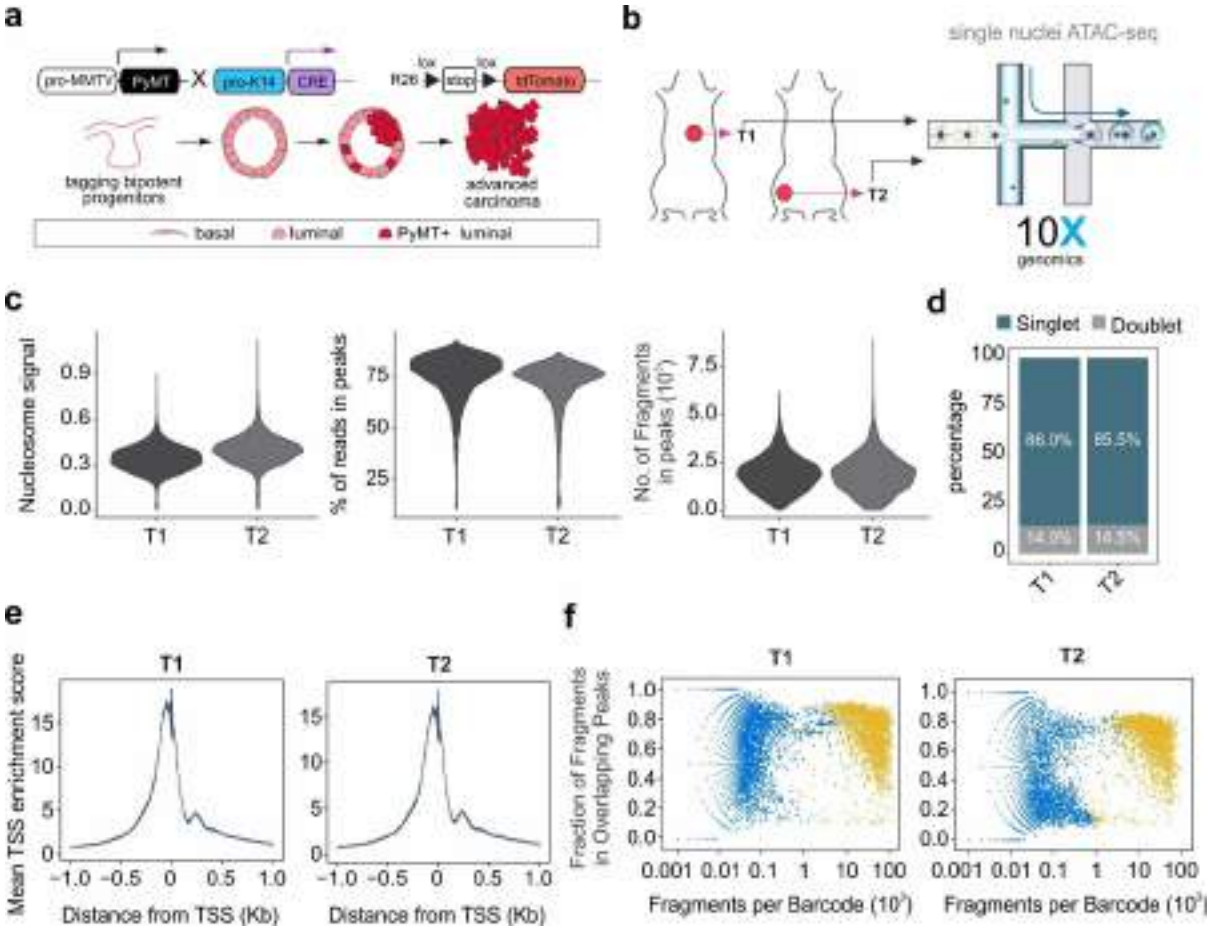


Figure 38 | High quality chromatin profile occurring during primary breast cancer progression recovered from single nuclei ATAC-Seq. (a) MMTV-PyMT mouse primary breast carcinoma model. The expression of the PyMT oncogene under the control of the *MMTV* promoter and the Cre recombinase was expressed under the control of the *K14* promoter. We genetically labelled cancer cells using a tdTomato reporter. **(b)** Illustration of the strategy used for sample preparation and sequencing. **(c)** Violin plot showing different QC parameters: nucleosome signal, percentage of reads in peak regions, and number of fragments in peak regions. **(d)** Bar plot showing the percentage of predicted doublets and singlets using AMULET (Thibodeau et al., 2021). **(e)** Line plot showing the average TSS enrichment score around TSSs. **(f)** Scatterplot showing the number of fragments per cell barcode and the percentage of fragments overlapping peaks. Yellow indicates the valid cellular barcodes, and blue indicates the rest of non-cell barcodes.

Next, to pull both WT samples in single space of chromatin accessibility, we first normalised fragment count and used this to perform an anchor-based integration (Figure 39a; details in materials and methods). The SNN-based clustering of accessible regions resulted in 14 different clusters (Figure 39b). The relative changes in chromatin accessibility detected in the clusters are shown in Figure 39c.

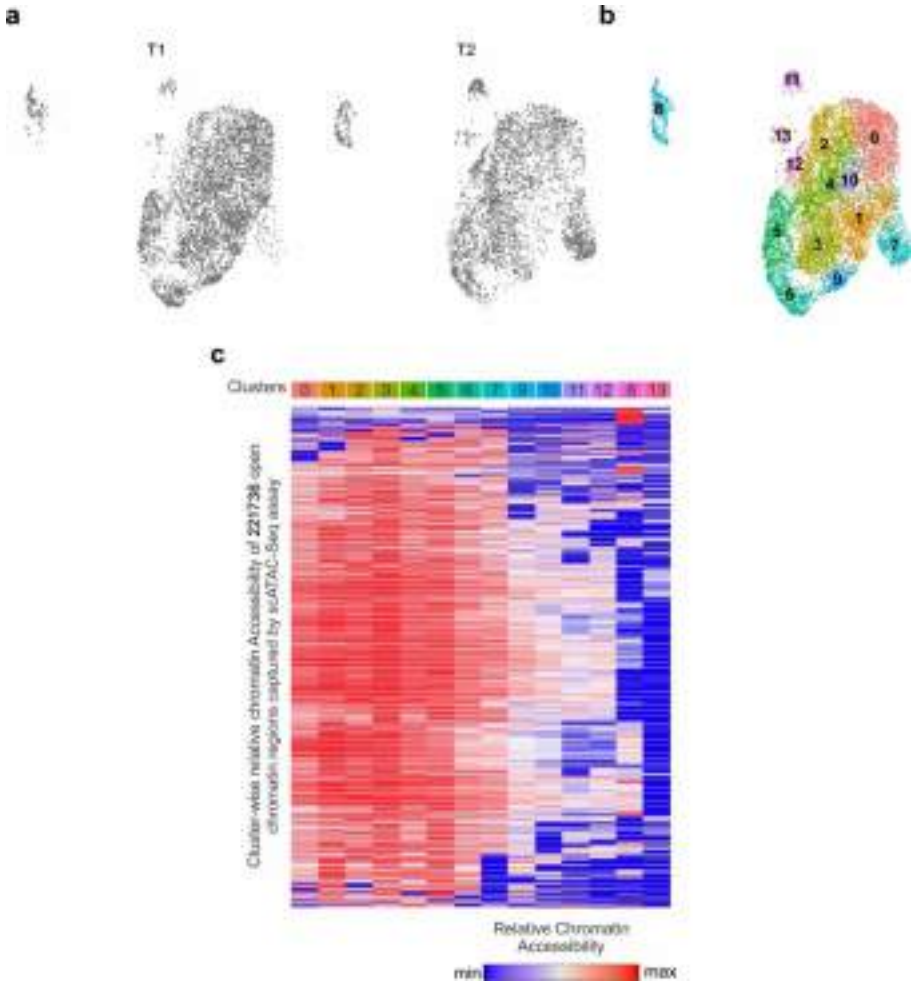


Figure 39 | Integrated single cell ATAC-Seq data explains cellular heterogeneity based on chromatin remodelling during primary breast cancer progression. (a) Uniform Manifold Approximation and Projection (UMAP) showing the integration of different tumour samples. **(b)** UMAP represents 14 distinct cellular clusters obtained using shared nearest neighbor (SNN) modularity. The analysis provides an explanation of cell heterogeneity based on chromatin architecture during primary breast cancer progression. **(c)** Heatmap representing the average relative chromatin accessibility of all open chromatin regions in each cluster.

The scATAC-Seq profile is extremely sparse, leading to zero-inflated quantification. Additionally, scATAC-Seq captures genomic locations rather than actual gene expression, making it difficult to annotate the detected clusters solely based on genomic regions. One alternative is to use the gene activity score calculated by aggregating the fragments count overlapping the gene body and a 2-kb upstream region (Stuart et al., 2021).

Since we have scRNA-Seq data for the same samples, we decided to use the more robust approach of expression imputation to annotate the detected scATAC-Seq clusters. We imputed the gene expression matrix from our scATAC-Seq profile using a label transfer approach (Stuart et al., 2021; details are given in materials and methods). The imputed expression values for the markers were then used to annotate the detected clusters (Figure 40b).

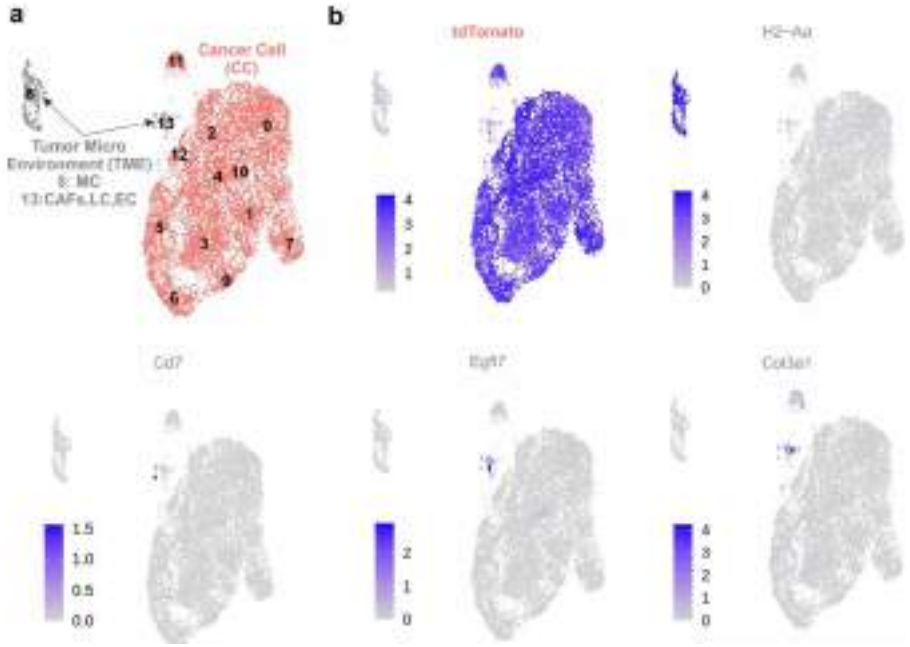


Figure 40 | scATAC-Seq profile of MMTV-PyMT tumours shows chromatin remodelling in both cancer cells and tumour micro-enviornment (TME). (a) UMAP representing the 13 major clusters found. Two of them represent TME and the rest are cancer cells. **(b)** UMAP showing imputed gene expression of the *bona fide* markers used to annotate the clusters. Gene expression is imputed using gene activity score, calculated based on the open chromatin regions around the target gene.

We further visualised the distinct patterns of chromatin accessibility for the regulatory regions of the selected markers in the different clusters (Figure 41). This marker-based approach suggests that the majority of cells in our scATAC-Seq profile were annotated as cancer cells (n = 7391). We also identified a few cells (n = 332) that belong to accessory populations including immune cells myeloid and lymphoid lineages (MC & LC), cancer associated fibroblasts (CAFs) and endothelial cells (EC) (Figures 40 and 41).

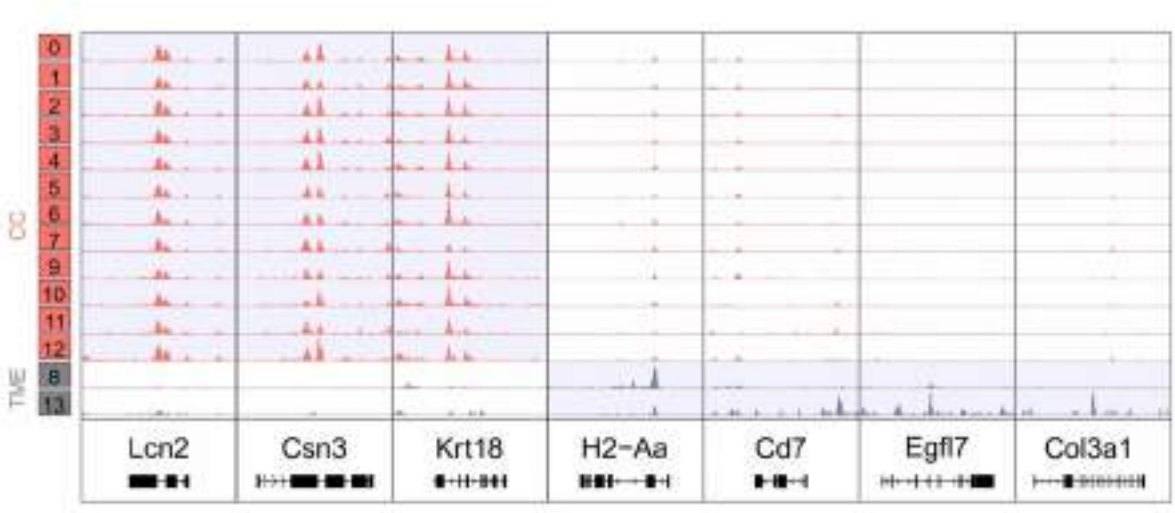


Figure 41 | Genome browser snapshot showing open chromatin regions around the selected *bona fide* markers for each annotated population.

Overall, we generated high-quality scATAC-Seq libraries for the PyMT WT tumour samples. Our systematic analysis of the scATAC-Seq data revealed distinct chromatin accessibility patterns that help to capture and understand the cellular heterogeneity during breast cancer progression.

4.5.2 scRNA-Seq-based EMT trajectories reveals unique chromatin accessibility patterns during BC progression

To understand the cellular heterogeneity explained by chromatin accessibility in segregated cancer cell populations we subset the cancer cells based on imputed tdTomato expression along with other cancer cell markers (Figure 40). The reclustering of cancer cells resulted in 14 distinct clusters (Figure 42a). Next, the gene signature-based enrichment analysis (details are given in materials and methods) enabled us to identify the EMT trajectory clusters previously found in our scRNA-Seq data (Figure 42b). Our gene signature-based analysis suggests the existence of a sequential opening and closing of the chromatin associated with the genes expressed along the EMT trajectory clusters (Figure 42b). The majority of the tumour cells in our scATAC-Seq data are of luminal phenotype, mainly belonging to clusters 5 and 11 in EMT trajectory (Figure 42b), compatible with our scRNA-Seq data (Figure 32). In the spectrum of differentiation and EMT dynamics, as the cells start dedifferentiating, we observed a lower enrichment of LA clusters, and simultaneous higher enrichment for the clusters with progenitor-like phenotype and active EMT.

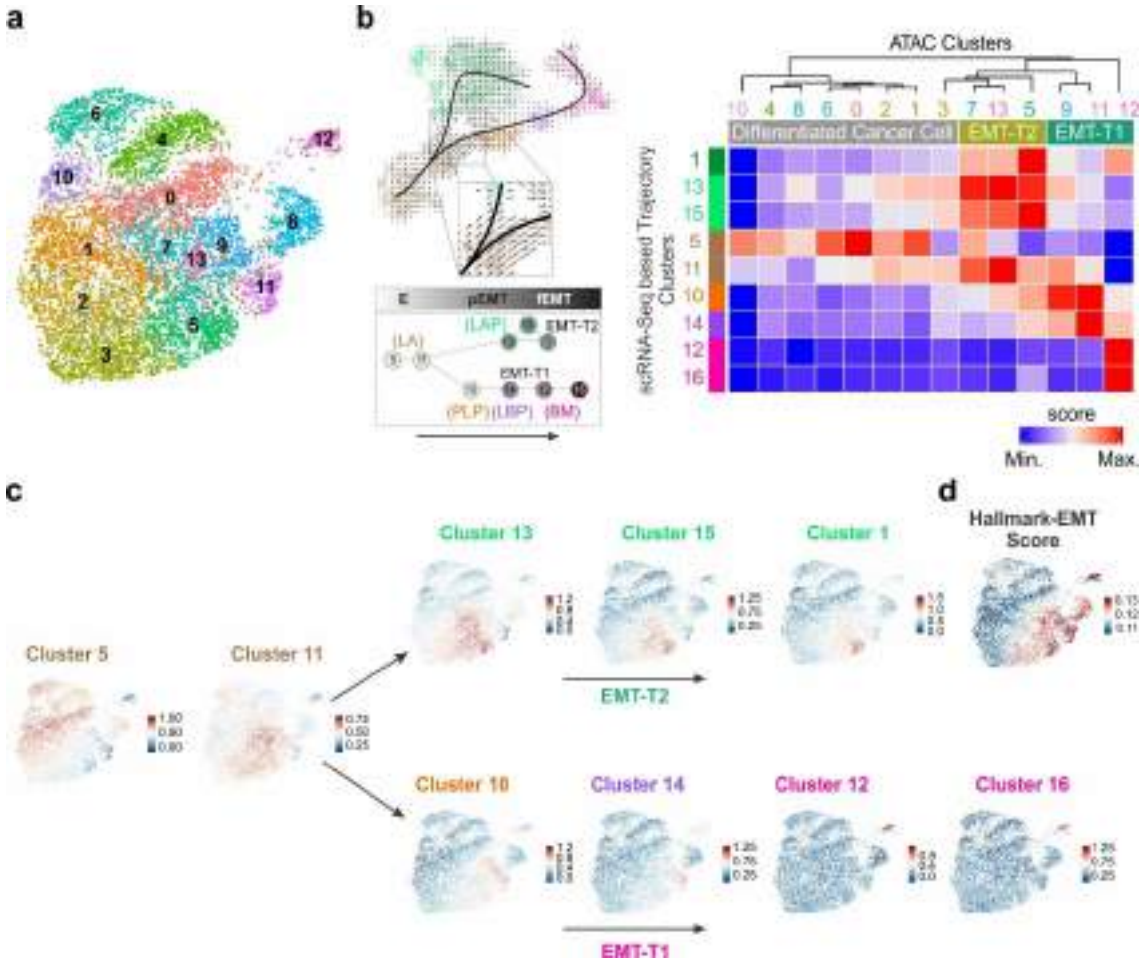


Figure 42 | Reactivation of EMT programmes (embryonic/invasive and adult/inflammatory) involves specific chromatin remodelling during breast cancer progression. (a) UMAP showing 14 distinct clusters of cancer cells obtained analysing scATAC-Seq data. **(b)** UMAP showing the reactivation of two distinct EMT programmes during breast cancer progression obtained using scRNA-Seq data (left). Heatmap showing the relative enrichment score of gene signatures for each cluster in EMT trajectories derived from scRNA-Seq data (right). Scale from blue to red indicates the lowest gene signature score (absence) to the highest enrichment score. **(c)** UMAP from scATAC-Seq data representing enrichment the score of gene signature for the clusters identified in scRNA-Seq (as plotted in heatmap). **(d)** UMAP representing enrichment score for different gene signatures: EMT-Hallmark, BC-PINGs and Inflammation.

With this approach we successfully annotated cancer cell clusters in the EMT context and described their progression based on chromatin accessibility (Figure 42c). The progression of EMT trajectory clusters shows distinct dynamics of chromatin remodeling compared to bulk of the tumor which is concomitantly enriched with EMT-Hallmark (Figure 42d). Finally, to obtain a list of potential targets of the EMT transcription factors, we first annotated all the peaks in cancer cells to the nearest TSS regions and performed TF motif enrichment. The peaks showing significant motif enrichment were assigned as a target gene to the regulator. This resulted in a baseGRN of 18933 target genes regulated by 1091 TFs which was used for the TF perturbation analysis.

Taken together, we have shown a simple yet effective approach of cell states mapping between two different modalities i.e. scRNA-Seq and scATAC-Seq. This integrated analysis helped us to correlate the dynamics of chromatin accessibility and associated gene expression changes in EMT trajectories during BC progression. Additionally, we recovered a highly reliable list of TF-target genes to construct the baseGRN for downstream analysis.

4.5.3 *In silico* perturbation analysis is a useful framework to predict the impact of TFs perturbation on well-established tumour states

The EMT-TFs belong to different families: Snail, Zeb, Twist and Prrx proteins and they play an important role in EMT initiation and progression. Specifically, Snail1 is a pioneer EMT-TF which initiates the EMT programme by repressing epithelial genes (Cano et al., 2000; Youssef et al., 2024), and Prrx1 is a potent mesenchymal inducer which levels are correlated with the progression to the full mesenchymal state. Prrx1 expression is associated with invasion (Youssef et al, 2024) and needs to be downregulated for metastatic colonisation (Ocaña et al., 2012a). Compatible with their described role in our scRNA-Seq data, we found Snail1 is expressed in both EMT trajectories (Figure 43b), whereas Prrx1 expression is restricted to the advanced EMT-T1 (Figure 43c).

CellOracle simulates global downstream shifts in gene expression using information from the TF-target gene list (baseGRN) to visualize the changes in cell state transitions. This is particularly useful for understanding how perturbing a TF can affect the cellular transition in a trajectory. To illustrate these transitions,

CellOracle calculates a perturbation score (PS). A negative PS indicates that the simulated perturbation in the expression of the corresponding TF deviates the cell state transitions in the opposite direction of the normal trajectory. Conversely, a positive PS suggests that perturbation enhances the cell state transition in the same direction as the normal trajectory.

To challenge the EMT trajectories during BC progression, first we run the simulations for Snail1. As a pioneer EMT-TF, *in silico* simulation of a Snail1 KO condition predicts a massive impact on both trajectories (EMT-T1 and EMT-T2). In simulated Snail1 KO tumours the prediction is the inhibition of the bifurcation point (cluster 11) suggesting a full inhibition of EMT-dependent tumour progression.

In a scenario where Snail1 would be inhibited in tumour cells that are already engaged in EMT-T2, this trajectory is predicted to also be inhibited. Dedifferentiation of cluster1 would not occur within the trajectory and cells would shift towards EMT-T1, increasing the size of clusters 10 and 14. Finally, if *Snail1* is inhibited in the advanced state of EMT-T1 the prediction is the inhibition of trajectory and a shift from cluster 16 to cluster 12 (Figure 43b).

On the other hand, the simulation of Prrx1 inhibition predicts impact only on advanced EMT-T1 state (cluster 16). Tumour cells could not reach the full EMT state (cluster 16) and the EMT-T1 trajectory would be truncated (Figure 43c).

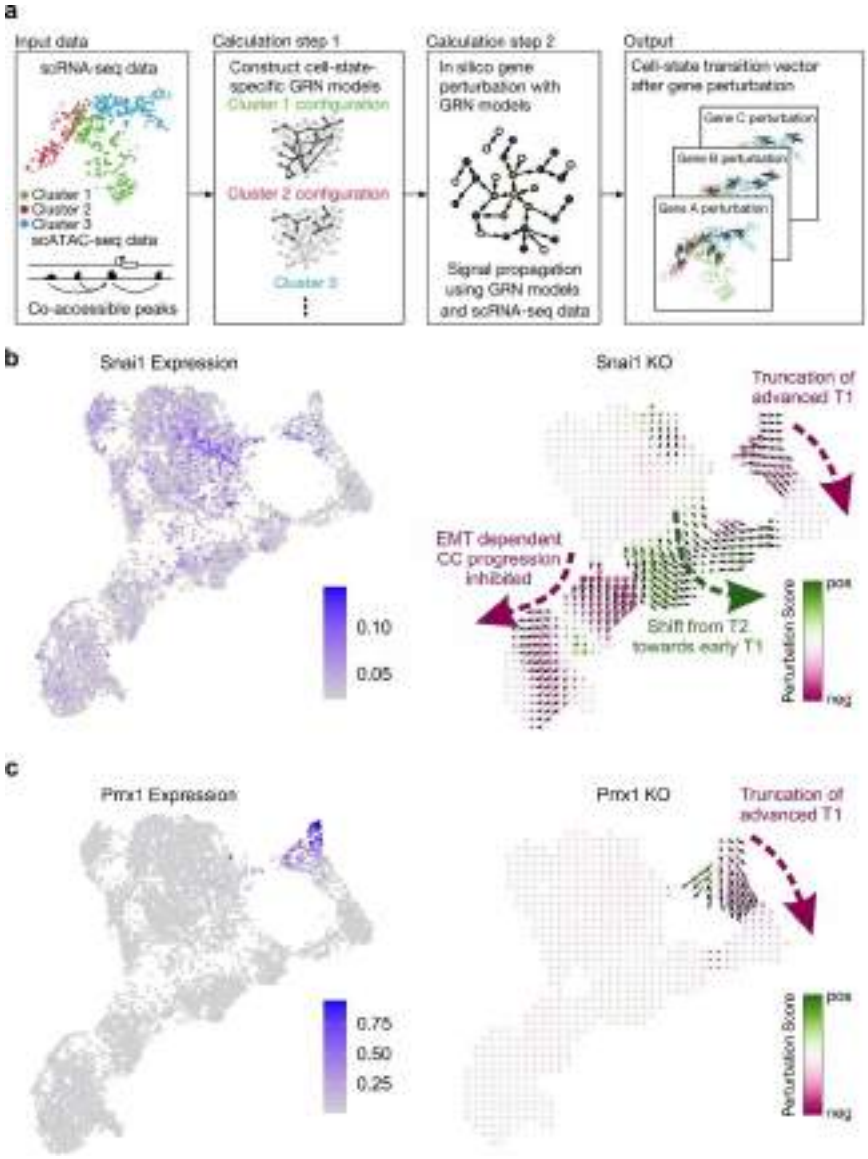


Figure 43 | *In silico* perturbation of EMT-TFs shows predicts impact on EMT-dependent progression of primary breast tumours. (a) Core algorithm of CellOracle (Kamimoto et al., 2023) for the prediction of gene regulatory networks (GRNs) and *in silico* perturbation analysis of TFs. CellOracle takes two inputs (scRNA-Seq based trajectories and TF-Target pairs) and uses machine learning algorithms to predict the perturbation of specific TF on cellular transitions. **(b)** UMAP showing MAGIC imputed expression of Snail1 in EMT trajectories (left) and *in silico* perturbation impact of *Snail1* knockout (right) on EMT trajectories. Arrows indicate the progression of altered EMT trajectories upon TF perturbation. Purple represents negative and green positive perturbation scores. **(c)** A similar analysis for Prrx1 loss.

Overall, our *in silico* perturbation analysis of EMT-TFs provides an insight on the putative state of mutant tumors including dynamics, as it can predict what would be the impact of inactivating the EMT-TF when cells are at different steps (clusters) through the trajectories. This is very useful for the design of validation experiments in mouse models (see discussion) and importantly, it can also be of help to design targeted therapies, including those derived from the identification of novel targets included in the EMT-TFs regulons.

Chapter 5

DISCUSSION

Over the past decades, the epithelial to mesenchymal transition, EMT, has been studied in multiple contexts such as morphogenesis, tissue homeostasis, wound healing, and different chronic conditions such as cancer and fibrosis (Figure 6; reviewed in Nieto et al., 2016; Thiery et al., 2009; Youssef and Nieto, 2024). The activation and reactivation of the EMT programme in different contexts shares some commonalities, but more importantly there are also specificities. These specificities can be related to the degree of EMT activation along the E-M spectrum, the activation of distinct molecular pathways, and most importantly, whether the cells are invasive or not (Nieto et al., 2016; Youssef and Nieto, 2024). During embryonic development, cells undergo an EMT, acquiring migratory and invasive properties that enable them to disseminate from the primary site and migrate to different parts of the embryo and later form different tissues and organs. In contrast, during kidney fibrosis, renal epithelial cells activate a partial EMT programme and unlike neural crest or other embryonic cells, they do not engage in invasion (Grande et al., 2015). Instead, they secrete cytokines and chemokines as part of the fibrogenic and inflammatory responses to injury, the two hallmarks of organ fibrosis progression (Grande et al., 2015; Lovisa et al., 2015). Here, through a comprehensive and parallel analysis of data obtained from single-cell transcriptomic analyses, we have deciphered the implementation of EMT programmes activated in different contexts. We have characterised the EMTs activated (i) in cells treated with TGFβ, the most potent EMT inducer, (ii) in neural crest cells delaminating from the neural tube as a paradigmatic example of embryonic EMTs, and (iii) during renal fibrosis, as an example of adult EMT activated in response to tissue damage. Additionally, we have analysed the EMT activated in cancer, in particular, that implemented during the progression of breast carcinoma in a mouse model (MMTV-PyMT) known to progress as human breast cancer of the most aggressive and deadly type (Attalla et al., 2021), the triple negative (TNBC), which still is an unmet clinical need (https://www.pharmaceutical-technology.com/sectors/healthcare/triple-negative-breast-cancer-treatment/?cf-view).

Below, I will discuss the most important findings and put into the frame of our current understanding of cell plasticity along the epithelial to mesenchymal spectrum in health and disease.

The possibility of analysing data from bulk and single cell sequencing of transcriptomes and the availability of data coming from cell lines, neural crest, renal fibrosis and breast cancer has provided an optimal framework to better understand how the EMT programme is implemented in the different contexts. Consequently, it has allowed us not only to find similarities and specificities among the different physiological and pathological settings but also to better understand the biology of the cancer cell during tumour progression. As such, we believe that only by doing this transversal study, we have been able to formulate the existence of two opposing EMT programmes in cancer, one protumour (invasive) that drives cell dissemination towards metastasis and another inflammatory that promotes antitumour responses. Finally, we have applied *in silico* perturbation analysis to predict the potential impact of targeting either of the two EMT programmes. Some predictions have been validated in the lab through state-of-the-art functional analyses in animal models.

5.1 Commonalities and specificities in the implementation of the EMT programme in different biological contexts

With respect to **commonalities** between the EMT programmes implemented in the different contexts, the most important one is the stereotyped sequential activation of EMT-TFs throughout the transition from epithelial (E) to mesenchymal (M) states. The sequence we have found for the different EMT-TF families starts with Snail, and then Twist and Zeb, and finally Prrx. Although all EMT-TFs can induce EMT in cultured cells, they have different capabilities. Snail and Zeb are potent epithelial repressors while Twist and Prrx are potent mesenchymal inducers (Youssef and Nieto, 2024). The sequence of recruitment allows to first start repressing the epithelial phenotype that allows the disruption of cell-cell adhesion (Snail), then acquisition of some mesenchymal traits that allows cell delamination plus further repression of the epithelial phenotype (Twist and Zeb), and finally, the activation of the potent mesenchymal inducer Prrx, tightly associated with the acquisition of robust invasive properties. In our data analyses we have found that this sequence is observed during the activation of EMT in epithelial cells in culture upon treatment with TGF-beta, and it is also compatible with the described phenotypes for a whole spectrum of cancer cells that we could ordered along the E-M axis. Importantly, what we have learned

from this sequential activation has allowed us to apply it to the *in vivo* models used in the lab, neural crest, renal fibrosis and cancer, which will be discussed below. Importantly, we have found it to be fully compatible with the different functions associated with the different EMT programmes implemented in each biological context.

Another commonality that we have found is cell dedifferentiation, which appears to be the first response to the activation of EMT in differentiated cells. Obviously not required during developmental EMTs, as cells are not differentiated into final cell types, it is however common to adult EMT programmes, regardless of whether the cells are transformed (cancer cells) or not (response to injury). As such, our data analyses on renal fibrosis and cancer progression highlight the loss of differentiation traits in the corresponding tissue. This dedifferentiation step is reminiscent of the lineage infidelity and plasticity described in adult skin wound healing and cancer (Gerber et al., 2018), also observed in other carcinomas (Marjanovic et al., 2020). With this, we would like to propose that this plasticity is triggered by the activation of EMT, that also occurs concomitant with cell dedifferentiation in neuroblastoma and melanoma where adrenergic cells or melanocytes, respectively, reactivate embryonic neural crest markers (Kaufman et al., 2016; van Groningen et al., 2017). Activation of EMT has also been associated with cell dedifferentiation and the emergence of repair cell states in limb, fin, and heart regeneration in axolotl and zebrafish (Youssef and Nieto, 2024). Interestingly, for repair, EMT needs to be transient, as it must be downregulated to allow redifferentiation of the new cells to occur. This has been observed in different contexts, including heart regeneration (Aharonov et al., 2020; D'Uva et al., 2015). In fact, it is known that EMT is incompatible with differentiation sates and a forced transient activation is also consistent with reinstating heart regeneration in mice (González-Iglesias and Nieto, 2020). EMT is also transiently activated in cancer, as successful metastatic colonization, involves downregulation of the EMT programme (Ocaña et al., 2012a; Tsai et al., 2012). However, during renal fibrosis, and as result of a chronic damage, EMT activation is not transient, and although triggering dedifferentiation, it progresses to degeneration and organ failure rather than repair. Thus, it is all compatible with the EMT laying at the core of somatic cell dedifferentiation, as a driver of

epimorphosis to achieve phenotypic plasticity in the adult.

Regarding **specificities**, the most important one is whether the EMT programme is associated with cell invasion and dissemination or not. We have found that even in an epithelial cell line, MDCK, we can discriminate two different EMT signatures in two different subclones, one compatible with progressing to the full mesenchymal phenotype with robust invasive properties, and another one with a reduction in the expression of epithelial markers and a modest mesenchymal activation without any sign of invasive genes being activated. This, together with the analysis of signatures in already known invasive and non-invasive breast cancer cells allowed us to define a pro-invasive signature (PINGs) that we have later applied to *in vivo* contexts. With that, we confirm the invasive nature of the embryonic trajectory and the non-invasive nature of the EMT programme activated in the adult in non-transformed cells (Figure 44; Youssef and Nieto, 2024).

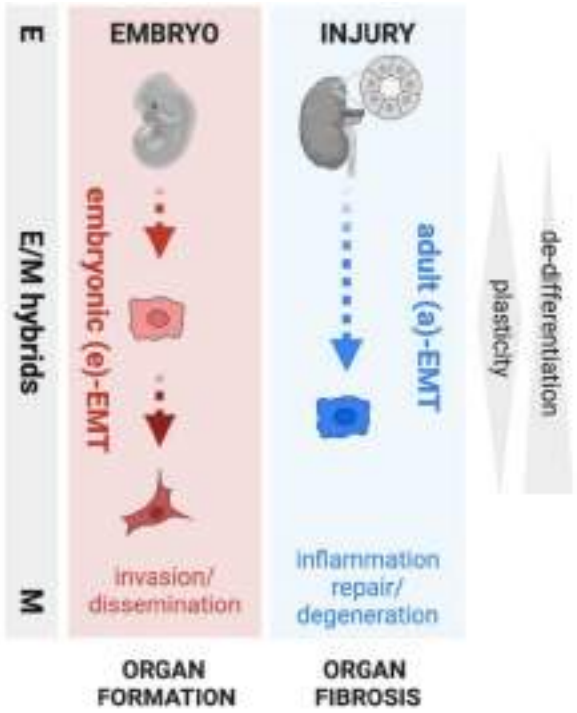


Figure 44 | The embryonic and adult EMT programmes.

The non-invasive programme of the adult is associated with an inflammatory signature, which is also specific for the adult response. This response may work as a repair mechanism is the injury in acute damage but leads to degeneration in contexts of chronic damage (Grande et al., 2015; Lovisa et al., 2015), as discussed above. The deep characterisation of embryonic and adult responses has allowed us to better interpret the implementation of EMT during cancer progression, as discussed below.

5.2 The two opposing EMT programmes implemented during tumour evolution

The EMT, as a crucial process in health and disease has been studied for decades. Importantly, while its importance was accepted by embryologists very quickly after the pioneer studies, it was a difficult concept to be acknowledged by the medical community, both in cancer and fibrosis. Nevertheless, it is now well accepted and there are numerous contributions from many labs in all the contexts we have studied. As a matter of fact, our findings in the neural crest and renal fibrosis discussed above essentially confirm and extend previous data (Grande et al., 2015; Piacentino et al., 2020; Sheng and Zhuang, 2020; Soldatov et al., 2019), although it is worth noting that in our work we specifically focus on the implementation of the EMT programme, describing the main hints of transcriptional changes and interpreting them in terms of putative associated functions along the process. Thus, we have defined the two different EMT programmes and their trajectories in neural crest and renal fibrosis, representing the response of embryonic and adult cells, respectively.

With that in mind, our main effort has been devoted to the interpretation of the EMT in cancer. Our lab was pioneer in finding the connection between embryonic development and cancer (Cano et al., 2000; Nieto et al., 1994), and as mentioned, it was not initially accepted in the cancer research field. Once studies from many different labs showed the importance of the process (discussed in Brabletz et al., 2018), the idea still was that a type of EMT different from that activated in development or fibrosis could be operating in cancer as had bene proposed 15 years ago (Kalluri and Weinberg, 2009). What we have been able to show in this work is that there is not a cancer-specific EMT programme. During

tumour progression, as in other adult contexts mentioned above, the EMT-first induces an initial dedifferentiation step that provides the required plasticity. This transition is then followed by a bifurcation, after which two alternative pathways recapitulate either the embryonic-like or the adult-like responses. This means that cancer cells hijack both developmental and adult EMT plasticity programmes normally used for cell invasion and migration, or as a response to injury, respectively, to implement cell dissemination or inflammation. This conclusion has been possible thanks to a deep analysis of the corresponding transcriptional programmes and the comparison with the transcriptomes described for the neural crest and renal fibrosis. Thus, the proposal is that the embryonic-like trajectory promotes cell dissemination and with that, tumour progression towards metastasis, while the adult-like trajectory represents a mechanism activated in response to damage, in this case induced by the oncogene. Work in the lab in state-of-the art animal models confirm these trajectories and in addition, provide spatial information, indicating that the two programmes are activated in segregated cancer cell populations (Figure 45; Youssef and Nieto, 2024).

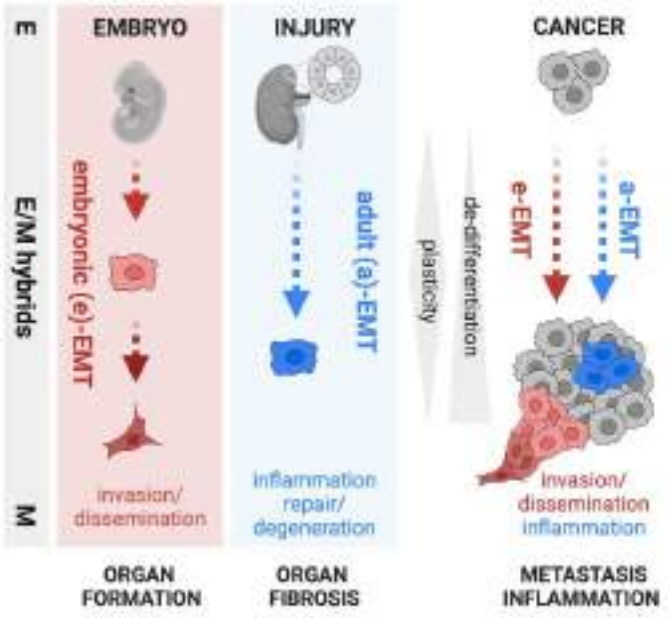


Figure 45 | Two distinct EMT programmes, reminiscent of the embryonic and the adult programmes are activated in segregated tumour populations, with signatures compatible with invasion and cell dissemination (like in embryos), and inflammation (like in organ fibrosis).

5.3 The potential value of *in silico* perturbations: efficient experimental designs and prediction of potential therapeutic interventions

In addition to the analyses of single-cell transcriptomes in different biological contexts, which lead to the results discussed above and the publication of a manuscript currently in press in Nature Cancer, in the lab we have also performed scATAc seq experiments in breast tumours. Computational analyses have allowed us to map cell states between the two modalities and integrate the data coming from both scRNA-Seq and scATAC-Seq. Thus, we have correlated the dynamics of chromatin accessibility and associated gene expression changes along the two EMT trajectories during cancer progression. Additionally, we have recovered a highly reliable list of TF-target genes and have generated a baseGRN (GRN: gene regulatory network) of close to 19,000 genes regulated by over 1000 TFs. We have used this baseGRN to *in silico* challenge the two EMT trajectories in cancer interrogating the system for the predicted impact of perturbation of two EMT-TFs. We have chosen Snail1 and Prrx1, as the pioneer factor activated in both trajectories (Snail1) and the one that is required for cells to acquire invasive properties (validated and demonstrated in the lab), and therefore specific for EMT-T1. Advanced computational frameworks can provide a reliable and quick platform to predict the perturbation effect of novel candidates on the already established phenotype (Kamimoto et al., 2023).

Our *in silico* perturbations of Snail1 expression predicted the inhibition of the bifurcation point, if Snail1 was deleted from the beginning, thus predicting a massive impact on EMT activation and suggesting a full inhibition of EMT-dependent tumour progression (Figure 43b). Importantly, this huge impact predicted has been validated in the lab through functional analyses in the mouse (Figure 46). Mice show a massive inhibition of tumour formation and growth, with those tumours that could barely develop, showing a high degree of cell differentiation, indicative of defective tumour progression.

Our *in silico* perturbations of Prrx1 expression predicted an impact only on the advanced EMT-T1 trajectory. The prediction was that tumour cells could not advance through the invasive and dissemination trajectory and therefore, that EMT-T1 trajectory would be truncated (Figure 43c). This prediction has been validated in the lab (Figure 46).

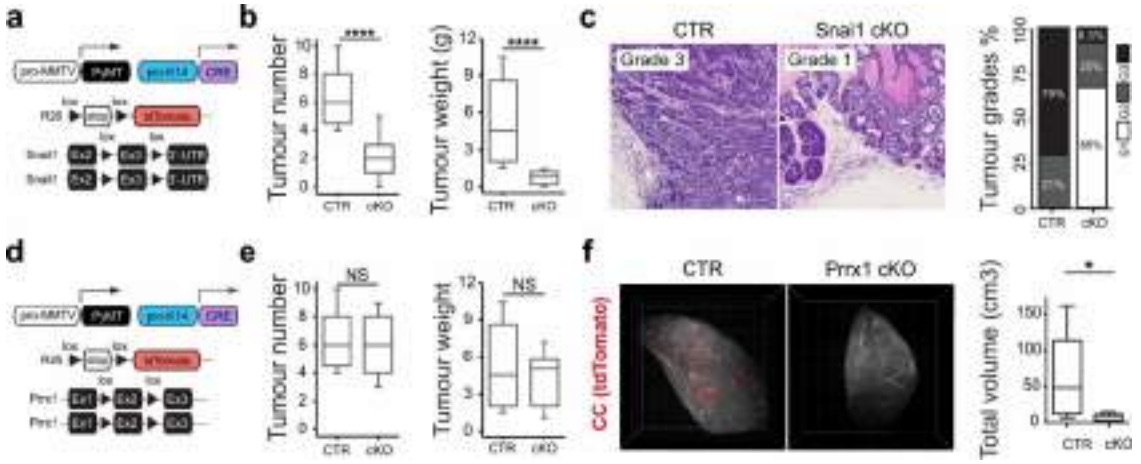


Figure 46 | *In vivo* conditional knock out (cKO) of EMT-TFs confirm the predictions obtained in *in silico* perturbation analyses for breast cancer progression. **(a)** Illustration of *Snail1* cKO mouse model. The MMTV-PyMT mouse model was crossed with mice bearing a floxed *Snail1* allele until homozygosis, leading to the inability for cancer cells to activate Snail1. Tumour cells were genetically labelled using a tdTomato reporter **(b)** Boxplot showing the number of tumours and tumour weight in WT compared to Snail1 negative tumours. **(c)** Histological sections showing tumour grade in both types of tumour samples (left). Quantification of tumour grade in the same tumours (right). **(d)** Illustration of *Prrx1* cKO mouse model following a similar strategy. **(e)** Number of tumours and tumour weight in WT compared to *Prrx1* cKO tumours. **(f)** Whole lung preparations (left) used to quantify metastatic burden in control compared to *Prrx1* cKO tumours (right). Experiments carried out by Raul Jimenez-Castaño, Khalil Kass Youssef and Joan Galcerán in the lab.

A massive impact on invasion and therefore dissemination and metastasis was predicted *in silico* and confirmed in animal models lacking Prrx1. But furthermore, we found in these tumours there is an enhanced inflammatory trajectory that was accompanied by an increased infiltration of intratumor macrophages of the antitumour type (Figure 47). This important observation indicates that the two trajectories have opposing roles, one protumoural, promoting the metastatic

cascade, and another one antitumoural, mimicking the response to injury in adult tissues.

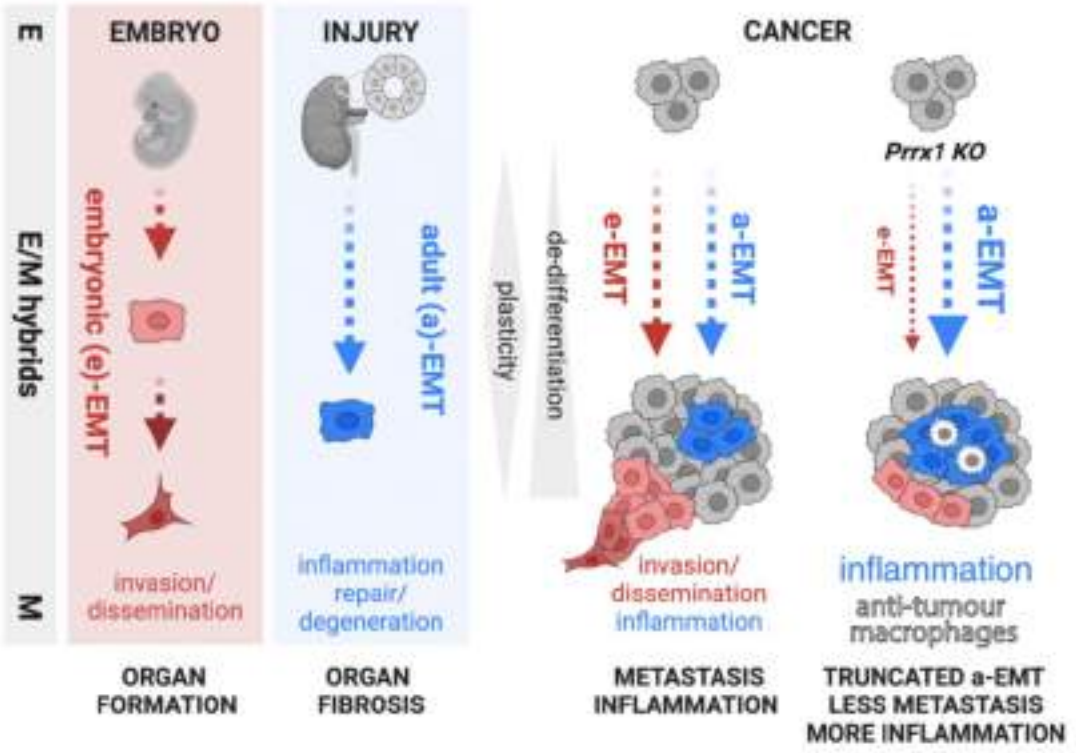


Figure 47 | The impact of Prrx1 loss in tumour progression

Finally, *in silico* perturbations can also predict the impact on tumour progression for cells located at different points along the trajectories. This is important because patients, at the time of diagnosis, usually have well developed tumours. Animal models can also be generated to delete the TFs of interest in advanced primary tumours using inducible CRE models to test the putative targets and its impact on disease progression. However, it would be difficult to generate models that interfere with intermediate states or to understand what the impact would be of targeting multiple specific molecules expressed in selected clusters. This is easily achieved *in silico* once shown that predictions provide useful information, as in our case. Furthermore, even when the generation of animal models is plausible, it is always challenging, resource-intensive and time-consuming. In this context, computational methods are invaluable, as they offer efficient and

scalable alternatives to traditional genetic models. As such, they also enable the prediction of the impact of deleting various factors simultaneously, and importantly, of putative therapeutic effects of targeting specific nodes of regulation with greater speed and accuracy. Nevertheless, confirmation with experimental models is essential, and predictions can help in selection of candidates and in planning the experiments in a more efficient manner. All this work has been possible thanks to the generation of sophisticated animal models by lab members and the existing framework in the lab, with expertise in all the required biological contexts, providing a bidirectional exercise to reach the conclusions stated below. Finally, it is also worth noting here that the different clusters identified in this work in mouse breast tumours have been identified in samples from breast cancer patients thanks to the collaboration with Dr. Gema Moreno-Bueno, at the Institute of Biomedical Research Sols-Morreale and MD Anderson Cancer Center International Foundation, in Madrid. In summary, advanced computational tools help interpreting experimental data and can bridge the gap between experimental limitations and the growing demand for targeted therapeutic strategies.

CONCLUSIONS

The availability of state-of-the-art technologies in the analysis of whole transcriptomes from thousands of individual cells has allowed us to revisit the epithelial to mesenchymal transition in different contexts thanks to the data available in the lab. We have analysed data from massive sequencing and obtained valuable information on gene signatures and specific markers associated with the acquisition of different functional properties in physiology and pathology that can be summarised in the following conclusions:

1. The EMT is not a binary process, and intermediate states can be identified along the epithelial to mesenchymal spectrum in different contexts. In the lab, after identifying different clones from a non-transformed epithelial kidney cell line that upon treatment with TGF-beta, the most potent EMT inducer, respond undergoing either a partial non-invasive or a full and invasive EMT.
2. The transition from epithelial to mesenchymal states is accompanied by the sequential recruitment of EMT transcription factors. Snail1 is a pioneer EMT inducer and Prrx1 is recruited later associated with the acquisition of the invasive phenotype.
3. The analysis of invasive and non-invasive cell lines has allowed us to identify a gene signature associated with the acquisition of invasive properties, that we call PINGS (pro-invasive gene signature).
4. The embryonic EMT programme implemented during the development of the neural crest is gradual, with cells moving from epithelial to mesenchymal states while acquiring invasive signatures.
5. The EMT implemented as a response to damage in adult tissues corresponds to a partial and non-invasive EMT, with a signature associated with fibrogenic and inflammatory pathways.
6. Dedifferentiation is a required first step to allow adult cells to respond to insults, and either repair or degenerate. We propose the EMT as a driver of somatic cell dedifferentiation to achieve phenotypic plasticity in the adult.
7. During tumour progression, cancer cells dedifferentiate and then bifurcate into two parallel and independent trajectories, EMT-T1 and EMT-T2.

8. EMT-T1 is reminiscent of the embryonic EMT trajectory, with cells transitioning towards the invasive and mesenchymal phenotypes, confirmed in the lab as the first step towards metastatic dissemination.EMT-T2 is reminiscent of the adult EMT trajectory, with cells undergoing a partial and non-invasive EMT with an inflammatory signature, confirmed in the lab to act as an antitumour mechanism.
9. *In silico* perturbation experiments applied to the two EMT trajectories in cancer were able to faithfully predict the outcome of similar genetic perturbations in mouse models. We suggest that the possibility to predict the impact of perturbations occurring at different nodes along the trajectories, can help in the design of better biological models and to anticipate the outcome of therapeutic intervention at those nodes.

La disponibilidad de tecnologías de última generación en el análisis de transcriptomas completos de miles de células individuales nos ha permitido realizar un estudio global de la transición epitelio-mesénquima gracias a los datos obtenido en el laboratorio en diferentes contextos. Hemos analizado estos datos y obtenido valiosa información sobre firmas génicas y marcadores asociados a la adquisición de propiedades funcionales en fisiología y patología que pueden resumirse en las siguientes conclusiones:

1. La EMT no es un proceso binario, y se pueden identificar estados intermedios a lo largo del espectro epitelial a mesenquimal en diferentes contextos. En el laboratorio, tras identificar diferentes clones de una línea celular renal epitelial no transformada que al tratamiento con TGF-beta, el inductor más potente de EMT, responden experimentando una EMT parcial no invasiva o una EMT completa e invasiva.
2. La transición de los estados epitelial a mesenquimal se acompaña del reclutamiento secuencial de diferentes factores de transcripción inductores de la EMT. Snail1 es un factor inductor de EMT pionero y Prrx1 es reclutado posteriormente asociado a la adquisición del fenotipo invasivo.
3. El análisis de líneas celulares invasivas y no invasivas nos ha permitido describir una firma génica asociada a la invasión, PINGS (pro-invasive gene signature).
4. El programa de EMT embrionario implementado durante el desarrollo de la cresta neural es gradual, con células que pasan de estados epiteliales a mesenquimales al tiempo que adquieren firmas invasivas.
5. La EMT adulta implementada como respuesta al daño en tejidos corresponde a una EMT parcial y no invasiva, con una firma de vías fibrogénicas e inflamatorias.
6. La desdiferenciación es un primer paso necesario para que las células adultas respondan al daño y repararse o degenerarse. Proponemos la EMT como motor de la desdiferenciación de células somáticas para conseguir plasticidad en el adulto.

7. Durante la progresión tumoral, las células cancerosas se desdiferencian y luego se bifurcan en dos trayectorias paralelas e independientes, EMT-T1 y EMT-T2.
8. EMT-T1 es semejante a la trayectoria embrionaria, con células en transición hacia los fenotipos invasivo y mesenquimático, confirmada en modelos animales en el laboratorio como el primer paso hacia la diseminación metastásica. EMT-T2 es reminiscente de la trayectoria adulta, las células experimentan una EMT parcial no invasiva con una firma inflamatoria, confirmada en el laboratorio como un mecanismo antitumoral.
9. Experimentos de perturbación *in silico* aplicados a las dos trayectorias de EMT en cáncer fueron capaces de predecir fielmente el resultado de perturbaciones genéticas similares en modelos de ratón. Con esto, sugerimos cómo la posibilidad de predecir el impacto de perturbaciones en diferentes nodos a lo largo de las trayectorias puede ayudar en el diseño de mejores modelos biológicos y también a anticipar el resultado de las intervenciones terapéuticas en esos nodos.

REFERENCES

Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, Mahfouz A. 2019. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* **20**:194. doi:10.1186/s13059-019-1795-z

Acloque H, Adams MS, Fishwick K, Bronner-Fraser M, Nieto MA. 2009. Epithelial-mesenchymal transitions: the importance of changing cell state in development and disease. *J Clin Invest* **119**:1438–1449. doi:10.1172/JCI38019

Aharonov A, Shakked A, Umansky KB, Savidor A, Genzelinakh A, Kain D, Lendengolts D, Revach O-Y, Morikawa Y, Dong J, Levin Y, Geiger B, Martin JF, Tzahor E. 2020. ERBB2 drives YAP activation and EMT-like processes during cardiac regeneration. *Nat Cell Biol* **22**:1346–1356. doi:10.1038/s41556-020-00588-4

Ahlstrom JD, Erickson CA. 2009. The neural crest epithelial-mesenchymal transition in 4D: a `tail' of multiple non-obligatory cellular mechanisms. *Development* **136**:1801–1812. doi:10.1242/dev.034785

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J, Van Den Oord J, Atak ZK, Wouters J, Aerts S. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**:1083–1086. doi:10.1038/nmeth.4463

Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE. 2019. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol* **20**:264. doi:10.1186/s13059-019-1862-5

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**:25–29. doi:10.1038/75556

Attalla S, Taifour T, Bui T, Muller W. 2021. Insights from transgenic mouse models of PyMT-induced breast cancer: recapitulating human breast cancer progression in vivo. *Oncogene* **40**:475–491. doi:10.1038/s41388-020-01560-0

Bach K, Pensa S, Grzelak M, Hadfield J, Adams DJ, Marioni JC, Khaled WT. 2017. Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat Commun* **8**:2128. doi:10.1038/s41467-017-02001-5

Batlle E, Sancho E, Francí C, Domínguez D, Monfar M, Baulida J, García De Herreros A. 2000. The transcription factor Snail is a repressor of E-cadherin gene expression in epithelial tumour cells. *Nat Cell Biol* **2**:84–89. doi:10.1038/35000034

Bendall AJ, Abate-Shen C. 2000. Roles for Msx and Dlx homeoproteins in vertebrate development. *Gene* **247**:17–31. doi:10.1016/S0378-1119(00)00081-0

Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. 2020. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* **38**:1408–1414. doi:10.1038/s41587-020-0591-3

Boutet A, De Frutos CA, Maxwell PH, Mayol MJ, Romero J, Nieto MA. 2006. Snail activation disrupts tissue homeostasis and induces fibrosis in the adult kidney. *EMBO J* **25**:5603–5613. doi:10.1038/sj.emboj.7601421

Brabletz T, Kalluri R, Nieto MA, Weinberg RA. 2018. EMT in cancer. *Nat Rev Cancer* **18**:128–134. doi:10.1038/nrc.2017.118

Bravo González-Blas C, De Winter S, Hulselmans G, Hecker N, Matetovici I, Christiaens V, Poovathingal S, Wouters J, Aibar S, Aerts S. 2023. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* **20**:1355–1367. doi:10.1038/s41592-023-01938-4

Bronner ME, Simões-Costa M. 2016. The Neural Crest Migrating into the Twenty-First CenturyCurrent Topics in Developmental Biology. Elsevier. pp. 115–134. doi:10.1016/bs.ctdb.2015.12.003

Cano A, Pérez-Moreno MA, Rodrigo I, Locascio A, Blanco MJ, Del Barrio MG, Portillo F, Nieto MA. 2000. The transcription factor Snail controls epithelial–mesenchymal transitions by repressing E-cadherin expression. *Nat Cell Biol* **2**:76–83. doi:10.1038/35000025

Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, Trapnell C, Shendure J. 2019. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**:496–502. doi:10.1038/s41586-019-0969-x

Carver EA, Jiang R, Lan Y, Oram KF, Gridley T. 2001. The Mouse Snail Gene Encodes a Key Regulator of the Epithelial-Mesenchymal Transition. *Molecular and Cellular Biology* **21**:8184–8188. doi:10.1128/MCB.21.23.8184-8188.2001

Castaño J, Aranda S, Bueno C, Calero-Nieto FJ, Mejia-Ramirez E, Mosquera JL, Blanco E, Wang X, Prieto C, Zabaleta L, Mereu E, Rovira M, Jiménez-Delgado S, Matson DR, Heyn H, Bresnick EH, Göttgens B, Di Croce L, Menendez P, Raya A, Giorgetti A. 2019. GATA2 Promotes Hematopoietic Development and Represses Cardiac Differentiation of Human Mesoderm. *Stem Cell Reports* **13**:515–529. doi:10.1016/j.stemcr.2019.07.009

Charafe-Jauffret E, Ginestier C, Iovino F, Wicinski J, Cervera N, Finetti P, Hur M-H, Diebel ME, Monville F, Dutcher J, Brown M, Viens P, Xerri L, Bertucci F, Stassi G, Dontu G, Birnbaum D, Wicha MS. 2009. Breast Cancer Cell Lines Contain Functional Cancer Stem Cells with Metastatic Capacity and a Distinct Molecular Signature. *Cancer Research* **69**:1302–1313. doi:10.1158/0008-5472.CAN-08-2741

Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**:128. doi:10.1186/1471-2105-14-128

Cheung Kevin J., Gabrielson E, Werb Z, Ewald AJ. 2013. Collective invasion in breast cancer requires a conserved basal epithelial program. *Cell* **155**:1639–1651. doi:10.1016/j.cell.2013.11.029

Cheung Kevin J., Gabrielson E, Werb Z, Ewald AJ. 2013. Collective Invasion in Breast Cancer Requires a Conserved Basal Epithelial Program. *Cell* **155**:1639–1651. doi:10.1016/j.cell.2013.11.029

Chevalier RL. 2016. The proximal tubule is the primary target of injury and progression of kidney disease: role of the glomerulotubular junction.

American Journal of Physiology-Renal Physiology **311**:F145–F161. doi:10.1152/ajprenal.00164.2016

Chicco D, Jurman G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**:6. doi:10.1186/s12864-019-6413-7

Chung W, Eum HH, Lee H-O, Lee K-M, Lee H-B, Kim K-T, Ryu HS, Kim S, Lee JE, Park YH, Kan Z, Han W, Park W-Y. 2017. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun* **8**:15081. doi:10.1038/ncomms15081

Conway BR, O'Sullivan ED, Cairns C, O'Sullivan J, Simpson DJ, Salzano A, Connor K, Ding P, Humphries D, Stewart K, Teenan O, Pius R, Henderson NC, Bénézech C, Ramachandran P, Ferenbach D, Hughes J, Chandra T, Denby L. 2020. Kidney Single-Cell Atlas Reveals Myeloid Heterogeneity in Progression and Regression of Kidney Disease. *JASN* **31**:2833–2854. doi:10.1681/ASN.2020060806

Dassule HR, Lewis P, Bei M, Maas R, McMahon AP. 2000. Sonic hedgehog regulates growth and morphogenesis of the tooth. *Development* **127**:4775–4785. doi:10.1242/dev.127.22.4775

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15–21. doi:10.1093/bioinformatics/bts635

D'Uva G, Aharonov A, Lauriola M, Kain D, Yahalom-Ronen Y, Carvalho S, Weisinger K, Bassat E, Rajchman D, Yifa O, Lysenko M, Konfino T, Hegesh J, Brenner O, Neeman M, Yarden Y, Leor J, Sarig R, Harvey RP, Tzahor E. 2015. ERBB2 triggers mammalian heart regeneration by promoting cardiomyocyte dedifferentiation and proliferation. *Nat Cell Biol* **17**:627–638. doi:10.1038/ncb3149

Emrich SJ, Barbazuk WB, Li L, Schnable PS. 2007. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res* **17**:69–73. doi:10.1101/gr.5145806

Ewing B, Green P. 1998. Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities. *Genome Res* **8**:186–194. doi:10.1101/gr.8.3.186

Fazilaty H, Rago L, Kass Youssef K, Ocaña OH, Garcia-Asencio F, Arcas A, Galceran J, Nieto MA. 2019. A gene regulatory network to control EMT programs in development and disease. *Nat Commun* **10**:5115. doi:10.1038/s41467-019-13091-8

Feldker N, Ferrazzi F, Schuhwerk H, Widholz SA, Guenther K, Frisch I, Jakob K, Kleemann J, Riegel D, Bönisch U, Lukassen S, Eccles RL, Schmidl C, Stemmler MP, Brabletz T, Brabletz S. 2020. Genome-wide cooperation of EMT transcription factor ZEB 1 with YAP and AP -1 in breast cancer. *The EMBO Journal* **39**:e103209. doi:10.15252/embj.2019103209

Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. 2018. Single sample scoring of molecular phenotypes. *BMC Bioinformatics* **19**:404. doi:10.1186/s12859-018-2435-4

Frangieh CJ, Melms JC, Thakore PI, Geiger-Schuller KR, Ho P, Luoma AM, Cleary B, Jerby-Arnon L, Malu S, Cuoco MS, Zhao M, Ager CR, Rogava M, Hovey L, Rotem A, Bernatchez C, Wucherpfennig KW, Johnson BE, Rozenblatt-Rosen O, Schadendorf D, Regev A, Izar B. 2021. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms

of cancer immune evasion. *Nat Genet* **53**:332–341. doi:10.1038/s41588-021-00779-1

Gavish A, Tyler M, Greenwald AC, Hoefflin R, Simkin D, Tschernichovsky R, Galili Darnell N, Somech E, Barbolin C, Antman T, Kovarsky D, Barrett T, Gonzalez Castro LN, Halder D, Chanoch-Myers R, Laffy J, Mints M, Wider A, Tal R, Spitzer A, Hara T, Raitses-Gurevich M, Stossel C, Golan T, Tirosh A, Suvà ML, Puram SV, Tirosh I. 2023. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**:598–606. doi:10.1038/s41586-023-06130-4

The Gene Ontology Consortium. 2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**:D330–D338. doi:10.1093/nar/gky1055

Gerber T, Murawala P, Knapp D, Masselink W, Schuez M, Hermann S, Gac-Santel M, Nowoshilow S, Kageyama J, Khattak S, Currie JD, Camp JG, Tanaka EM, Treutlein B. 2018. Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration. *Science* **362**:eaaq0681. doi:10.1126/science.aaq0681

Gewin LS. 2018. Renal fibrosis: Primacy of the proximal tubule. *Matrix Biology* **68–69**:248–262. doi:10.1016/j.matbio.2018.02.006

Ginestier C, Hur MH, Charafe-Jauffret E, Monville F, Dutcher J, Brown M, Jacquemier J, Viens P, Kleer CG, Liu S, Schott A, Hayes D, Birnbaum D, Wicha MS, Dontu G. 2007. ALDH1 Is a Marker of Normal and Malignant Human Mammary Stem Cells and a Predictor of Poor Clinical Outcome. *Cell Stem Cell* **1**:555–567. doi:10.1016/j.stem.2007.08.014

Giraddi RR, Chung C-Y, Heinz RE, Balcioglu O, Novotny M, Trejo CL, Dravis C, Hagos BM, Mehrabad EM, Rodewald LW, Hwang JY, Fan C, Lasken R, Varley KE, Perou CM, Wahl GM, Spike BT. 2018. Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. *Cell Reports* **24**:1653-1666.e7. doi:10.1016/j.celrep.2018.07.025

González-Iglesias A, Nieto MA. 2020. Proliferation and EMT trigger heart repair. *Nat Cell Biol* **22**:1291–1292. doi:10.1038/s41556-020-00594-6

Grande MT, Sánchez-Laorden B, López-Blau C, De Frutos CA, Boutet A, Arévalo M, Rowe RG, Weiss SJ, López-Novoa JM, Nieto MA. 2015. Snail1-induced partial epithelial-to-mesenchymal transition drives renal fibrosis in mice and can be targeted to reverse established disease. *Nat Med* **21**:989–997. doi:10.1038/nm.3901

Grandi FC, Modi H, Kampman L, Corces MR. 2022. Chromatin accessibility profiling by ATAC-seq. *Nat Protoc* **17**:1518–1552. doi:10.1038/s41596-022-00692-9

Gregory PA, Bert AG, Paterson EL, Barry SC, Tsykin A, Farshid G, Vadas MA, Khew-Goodall Y, Goodall GJ. 2008. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat Cell Biol* **10**:593–601. doi:10.1038/ncb1722

Gu Z. 2022. Complex heatmap visualization. *iMeta* **1**:e43. doi:10.1002/imt2.43

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**:2847–2849. doi:10.1093/bioinformatics/btw313

Guy CT, Cardiff RD, Muller WJ. 1992. Induction of mammary tumors by expression of polyomavirus middle T oncogene: a transgenic mouse

model for metastatic disease. *Mol Cell Biol* **12**:954–961. doi:10.1128/mcb.12.3.954-961.1992

Hafemeister C, Satija R. 2019. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**:296. doi:10.1186/s13059-019-1874-1

Hänzelmann S, Castelo R, Guinney J. 2013. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**:7. doi:10.1186/1471-2105-14-7

Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. 2021. Integrated analysis of multimodal single-cell data. *Cell* **184**:3573-3587.e29. doi:10.1016/j.cell.2021.04.048

Harper KL, Sosa MS, Entenberg D, Hosseini H, Cheung JF, Nobre R, Avivar-Valderas A, Nagi C, Girnius N, Davis RJ, Farias EF, Condeelis J, Klein CA, Aguirre-Ghiso JA. 2016. Mechanism of early dissemination and metastasis in Her2+ mammary cancer. *Nature* **540**:588–592. doi:10.1038/nature20609

Hay ED. 1958. The Fine Structure of Blastema Cells and Differentiating Cartilage Cells in Regenerating Limbs of *Amblystoma* Larvae. *The Journal of Cell Biology* **4**:583–592. doi:10.1083/jcb.4.5.583

Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, Lücken MD, Strobl DC, Henao J, Curion F, Single-cell Best Practices Consortium, Aliee H, Ansari M, Badia-i-Mompel P, Büttner M, Dann E, Dimitrov D, Dony L, Frishberg A, He D, Hediyeh-zadeh S, Hetzel L, Ibarra IL, Jones MG, Lotfollahi M, Martens LD, Müller CL, Nitzan M, Ostner J, Palla G, Patro R, Piran Z, Ramírez-Suástegui C, Saez-Rodriguez J, Sarkar H, Schubert B, Sikkema L, Srivastava A, Tanevski J, Virshup I, Weiler P, Schiller HB, Theis FJ. 2023. Best practices for single-cell analysis across modalities. *Nat Rev Genet* **24**:550–572. doi:10.1038/s41576-023-00586-w

Ho J. 2014. The Regulation of Apoptosis in Kidney Development: Implications for Nephron Number and Pattern? *Front Pediatr* **2**. doi:10.3389/fped.2014.00128

Hosseini H, Obradović MMS, Hoffmann M, Harper KL, Sosa MS, Werner-Klein M, Nanduri LK, Werno C, Ehrl C, Maneck M, Patwary N, Haunschild G, Gužvić M, Reimelt C, Grauvogl M, Eichner N, Weber F, Hartkopf AD, Taran F-A, Brucker SY, Fehm T, Rack B, Buchholz S, Spang R, Meister G, Aguirre-Ghiso JA, Klein CA. 2016. Early dissemination seeds metastasis in breast cancer. *Nature* **540**:552–558. doi:10.1038/nature20785

Huang Y, Hong W, Wei X. 2022. The molecular mechanisms and therapeutic strategies of EMT in tumor progression and metastasis. *J Hematol Oncol* **15**:129. doi:10.1186/s13045-022-01347-8

Humphreys BD, Lin S-L, Kobayashi A, Hudson TE, Nowlin BT, Bonventre JV, Valerius MT, McMahon AP, Duffield JS. 2010. Fate Tracing Reveals the Pericyte and Not Epithelial Origin of Myofibroblasts in Kidney Fibrosis. *The American Journal of Pathology* **176**:85–97. doi:10.2353/ajpath.2010.090517

Ianevski A, Giri AK, Aittokallio T. 2022. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* **13**:1246. doi:10.1038/s41467-022-28803-w

Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA. 2016. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **17**:29. doi:10.1186/s13059-016-0888-1

Ishii M, Han J, Yen H-Y, Sucov HM, Chai Y, Maxson RE. 2005. Combined deficiencies of *Msx1* and *Msx2* cause impaired patterning and survival of the cranial neural crest. *Development* **132**:4937–4950. doi:10.1242/dev.02072

Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* **11**:163–166. doi:10.1038/nmeth.2772

Kalluri R, Weinberg RA. 2009. The basics of epithelial-mesenchymal transition. *J Clin Invest* **119**:1420–1428. doi:10.1172/JCI39104

Kamimoto K, Stringa B, Hoffmann CM, Jindal K, Solnica-Krezel L, Morris SA. 2023. Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**:742–751. doi:10.1038/s41586-022-05688-9

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**:27–30. doi:10.1093/nar/28.1.27

Kaufman CK, Mosimann C, Fan ZP, Yang S, Thomas AJ, Ablain J, Tan JL, Fogley RD, Van Rooijen E, Hagedorn EJ, Ciarlo C, White RM, Matos DA, Puller A-C, Santoriello C, Liao EC, Young RA, Zon LI. 2016. A zebrafish melanoma model reveals emergence of neural crest identity during melanoma initiation. *Science* **351**:aad2197. doi:10.1126/science.aad2197

Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, Degenhardt J, Mayba O, Gnad F, Liu J, Pau G, Reeder J, Cao Y, Mukhyala K, Selvaraj SK, Yu M, Zynda GJ, Brauer MJ, Wu TD, Gentleman RC, Manning G, Yauch RL, Bourgon R, Stokoe D, Modrusan Z, Neve RM, De Sauvage FJ, Settleman J, Seshagiri S, Zhang Z. 2015. A comprehensive transcriptional portrait of human cancer cell lines. *Nat Biotechnol* **33**:306–312. doi:10.1038/nbt.3080

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P, Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**:1289–1296. doi:10.1038/s41592-019-0619-0

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**:W90-97. doi:10.1093/nar/gkw377

La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, Fan J, Borm LE, Liu Z, Van Bruggen D, Guo J, He X, Barker R, Sundström E, Castelo-Branco G, Cramer P, Adameyko I, Linnarsson S, Kharchenko PV. 2018. RNA velocity of single cells. *Nature* **560**:494–498. doi:10.1038/s41586-018-0414-6

Latil M, Nassar D, Beck B, Boumahdi S, Wang L, Brisebarre A, Dubois C, Nkusi E, Lenglez S, Checinska A, Vercauteren Drubbel A, Devos M, Declercq

W, Yi R, Blanpain C. 2017. Cell-Type-Specific Chromatin States Differentially Prime Squamous Cell Carcinoma Tumor-Initiating Cells for Epithelial to Mesenchymal Transition. *Cell Stem Cell* **20**:191-204.e5. doi:10.1016/j.stem.2016.10.018

LeBleu VS, Taduri G, O'Connell J, Teng Y, Cooke VG, Woda C, Sugimoto H, Kalluri R. 2013. Origin and function of myofibroblasts in kidney fibrosis. *Nat Med* **19**:1047–1053. doi:10.1038/nm.3218

Lehmann W, Mossmann D, Kleemann J, Mock K, Meisinger C, Brummer T, Herr R, Brabletz S, Stemmler MP, Brabletz T. 2016. ZEB1 turns into a transcriptional activator by interacting with YAP1 in aggressive cancer types. *Nat Commun* **7**:10498. doi:10.1038/ncomms10498

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**:323. doi:10.1186/1471-2105-12-323

Li X, Wang C-Y. 2021. From bulk, single-cell to spatial RNA sequencing. *Int J Oral Sci* **13**:36. doi:10.1038/s41368-021-00146-0

Liao C, Wang Q, An J, Long Q, Wang H, Xiang Meiling, Xiang Mingli, Zhao Y, Liu Y, Liu J, Guan X. 2021. Partial EMT in Squamous Cell Carcinoma: A Snapshot. *Int J Biol Sci* **17**:3036–3047. doi:10.7150/ijbs.61566

Liem KF, Tremml G, Jessell TM. 1997. A Role for the Roof Plate and Its Resident TGFβ-Related Proteins in Neuronal Patterning in the Dorsal Spinal Cord. *Cell* **91**:127–138. doi:10.1016/S0092-8674(01)80015-5

Lovisa S, LeBleu VS, Tampe B, Sugimoto H, Vadnagara K, Carstens JL, Wu C-C, Hagos Y, Burckhardt BC, Pentcheva-Hoang T, Nischal H, Allison JP, Zeisberg M, Kalluri R. 2015. Epithelial-to-mesenchymal transition induces cell cycle arrest and parenchymal damage in renal fibrosis. *Nat Med* **21**:998–1009. doi:10.1038/nm.3902

Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, Theis FJ. 2022. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**:41–50. doi:10.1038/s41592-021-01336-8

Luecken MD, Theis FJ. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* **15**:e8746. doi:10.15252/msb.20188746

Ma Q, Xu D. 2022. Deep learning shapes single-cell data analysis. *Nat Rev Mol Cell Biol* **23**:303–304. doi:10.1038/s41580-022-00466-x

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**:1202–1214. doi:10.1016/j.cell.2015.05.002

Madisen L, Zwingman TA, Sunkin SM, Oh SW, Zariwala HA, Gu H, Ng LL, Palmiter RD, Hawrylycz MJ, Jones AR, Lein ES, Zeng H. 2010. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat Neurosci* **13**:133–140. doi:10.1038/nn.2467

Marjanovic ND, Hofree M, Chan JE, Canner D, Wu K, Trakala M, Hartmann GG, Smith OC, Kim JY, Evans KV, Hudson A, Ashenberg O, Porter CBM, Bejnood A, Subramanian A, Pitter K, Yan Y, Delorey T, Phillips DR, Shah N, Chaudhary O, Tsankov A, Hollmann T, Rekhtman N, Massion PP, Poirier JT, Mazutis L, Li R, Lee J-H, Amon A, Rudin CM, Jacks T, Regev

A, Tammela T. 2020. Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. *Cancer Cell* **38**:229-246.e13. doi:10.1016/j.ccell.2020.06.012

Martik ML, Bronner ME. 2021. Riding the crest to get a head: neural crest evolution in vertebrates. *Nat Rev Neurosci* **22**:616–626. doi:10.1038/s41583-021-00503-2

Meulemans D, Bronner-Fraser M. 2004. Gene-Regulatory Interactions in Neural Crest Evolution and Development. *Developmental Cell* **7**:291–299. doi:10.1016/j.devcel.2004.08.007

Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe J-P, Tong F, Speed T, Spellman PT, DeVries S, Lapuk A, Wang NJ, Kuo W-L, Stilwell JL, Pinkel D, Albertson DG, Waldman FM, McCormick F, Dickson RB, Johnson MD, Lippman M, Ethier S, Gazdar A, Gray JW. 2006. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**:515–527. doi:10.1016/j.ccr.2006.10.008

Nieto MA. 2009. Epithelial-Mesenchymal Transitions in development and disease: old views and new perspectives. *Int J Dev Biol* **53**:1541–1547. doi:10.1387/ijdb.072410mn

Nieto MA, Huang RY-J, Jackson RA, Thiery JP. 2016. EMT: 2016. *Cell* **166**:21–45. doi:10.1016/j.cell.2016.06.028

Nieto MA, Sargent MG, Wilkinson DG, Cooke J. 1994. Control of Cell Behavior During Vertebrate Development by *Slug* , a Zinc Finger Gene. *Science* **264**:835–839. doi:10.1126/science.7513443

Ocaña OH, Córcoles R, Fabra A, Moreno-Bueno G, Acloque H, Vega S, Barrallo-Gimeno A, Cano A, Nieto MA. 2012a. Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer Prrx1. *Cancer Cell* **22**:709–724. doi:10.1016/j.ccr.2012.10.012

Ocaña OH, Córcoles R, Fabra A, Moreno-Bueno G, Acloque H, Vega S, Barrallo-Gimeno A, Cano A, Nieto MA. 2012b. Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer Prrx1. *Cancer Cell* **22**:709–724. doi:10.1016/j.ccr.2012.10.012

Onodera K, Fujiwara T, Onishi Y, Itoh-Nakadai A, Okitsu Y, Fukuhara N, Ishizawa K, Shimizu R, Yamamoto M, Harigae H. 2016. GATA2 regulates dendritic cell differentiation. *Blood* **128**:508–518. doi:10.1182/blood-2016-02-698118

Osorio D, Cai JJ. 2021. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics* **37**:963–967. doi:10.1093/bioinformatics/btaa751

Pal B, Chen Y, Vaillant F, Jamieson P, Gordon L, Rios AC, Wilcox S, Fu N, Liu KH, Jackling FC, Davis MJ, Lindeman GJ, Smyth GK, Visvader JE. 2017. Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. *Nat Commun* **8**:1627. doi:10.1038/s41467-017-01560-x

Papalexi E, Mimitou EP, Butler AW, Foster S, Bracken B, Mauck WM, Wessels H-H, Hao Y, Yeung BZ, Smibert P, Satija R. 2021. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat Genet* **53**:322–331. doi:10.1038/s41588-021-00778-2

Park S-M, Gaur AB, Lengyel E, Peter ME. 2008. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes Dev* **22**:894–907. doi:10.1101/gad.1640608

Pastushenko I, Brisebarre A, Sifrim A, Fioramonti M, Revenco T, Boumahdi S, Van Keymeulen A, Brown D, Moers V, Lemaire S, De Clercq S, Minguijón E, Balsat C, Sokolow Y, Dubois C, De Cock F, Scozzaro S, Sopena F, Lanas A, D'Haene N, Salmon I, Marine J-C, Voet T, Sotiropoulou PA, Blanpain C. 2018. Identification of the tumour transition states occurring during EMT. *Nature* **556**:463–468. doi:10.1038/s41586-018-0040-3

Pervolarakis N, Nguyen QH, Williams J, Gong Y, Gutierrez G, Sun P, Jhutty D, Zheng GXY, Nemec CM, Dai X, Watanabe K, Kessenbrock K. 2020. Integrated Single-Cell Transcriptomics and Chromatin Accessibility Analysis Reveals Regulators of Mammary Epithelial Cell Identity. *Cell Reports* **33**:108273. doi:10.1016/j.celrep.2020.108273

Petukhov V, Igolkina A, Rydbirk R, Mei S, Christoffersen L, Khodosevich K, Kharchenko PV. 2022. Case-control analysis of single-cell RNA-seq studies. doi:10.1101/2022.03.15.484475

Pezeshkian Z, Nobili S, Peyravian N, Shojaee B, Nazari H, Soleimani H, Asadzadeh-Aghdaei H, Ashrafian Bonab M, Nazemalhosseini-Mojarad E, Mini E. 2021. Insights into the Role of Matrix Metalloproteinases in Precancerous Conditions and in Colorectal Cancer. *Cancers* **13**:6226. doi:10.3390/cancers13246226

Piacentino ML, Li Y, Bronner ME. 2020. Epithelial-to-mesenchymal transition and different migration strategies as viewed from the neural crest. *Current Opinion in Cell Biology* **66**:43–50. doi:10.1016/j.ceb.2020.05.001

Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, Adey AC, Steemers FJ, Shendure J, Trapnell C. 2018. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular Cell* **71**:858-871.e8. doi:10.1016/j.molcel.2018.06.044

Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, Deschler DG, Varvares MA, Mylvaganam R, Rozenblatt-Rosen O, Rocco JW, Faquin WC, Lin DT, Regev A, Bernstein BE. 2017. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**:1611-1624.e24. doi:10.1016/j.cell.2017.10.044

Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. 2017. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**:979–982. doi:10.1038/nmeth.4402

Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, Schroth GP, Sandberg R. 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**:777–782. doi:10.1038/nbt.2282

Ransick A, Lindström NO, Liu J, Zhu Q, Guo J-J, Alvarado GF, Kim AD, Black HG, Kim J, McMahon AP. 2019. Single-Cell Profiling Reveals Sex, Lineage, and Regional Diversity in the Mouse Kidney. *Developmental Cell* **51**:399-413.e7. doi:10.1016/j.devcel.2019.10.005

Replogle JM, Saunders RA, Pogson AN, Hussmann JA, Lenail A, Guna A, Mascibroda L, Wagner EJ, Adelman K, Lithwick-Yanai G, Iremadze N,

Oberstrass F, Lipson D, Bonnar JL, Jost M, Norman TM, Weissman JS. 2022. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**:2559-2575.e28. doi:10.1016/j.cell.2022.05.013

Rhim AD, Mirek ET, Aiello NM, Maitra A, Bailey JM, McAllister F, Reichert M, Beatty GL, Rustgi AK, Vonderheide RH, Leach SD, Stanger BZ. 2012. EMT and Dissemination Precede Pancreatic Tumor Formation. *Cell* **148**:349–361. doi:10.1016/j.cell.2011.11.025

Romiguier J, Ranwez V, Douzery EJP, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Res* **20**:1001–1009. doi:10.1101/gr.104372.109

Saelens W, Cannoodt R, Todorov H, Saeys Y. 2019. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37**:547–554. doi:10.1038/s41587-019-0071-9

Salnikov AV, Liu L, Platen M, Gladkich J, Salnikova O, Ryschich E, Mattern J, Moldenhauer G, Werner J, Schemmer P, Büchler MW, Herr I. 2012. Hypoxia Induces EMT in Low and Highly Aggressive Pancreatic Tumor Cells but Only Cells with Cancer Stem Cell Characteristics Acquire Pronounced Migratory Potential. *PLoS ONE* **7**:e46391. doi:10.1371/journal.pone.0046391

Sarrió D, Rodriguez-Pinilla SM, Hardisson D, Cano A, Moreno-Bueno G, Palacios J. 2008. Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Res* **68**:989–997. doi:10.1158/0008-5472.CAN-07-2017

Shao X, Somlo S, Igarashi P. 2002. Epithelial-specific Cre/lox recombination in the developing kidney and genitourinary tract. *J Am Soc Nephrol* **13**:1837–1846. doi:10.1097/01.asn.0000016444.90348.50

Sharma SV, Haber DA, Settleman J. 2010. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat Rev Cancer* **10**:241–253. doi:10.1038/nrc2820

Shehata M, Teschendorff A, Sharp G, Novcic N, Russell IA, Avril S, Prater M, Eirew P, Caldas C, Watson CJ, Stingl J. 2012. Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. *Breast Cancer Res* **14**:R134. doi:10.1186/bcr3334

Sheng L, Zhuang S. 2020. New Insights Into the Role and Mechanism of Partial Epithelial-Mesenchymal Transition in Kidney Fibrosis. *Front Physiol* **11**:569322. doi:10.3389/fphys.2020.569322

Siemens H, Jackstadt R, Hünten S, Kaller M, Menssen A, Götz U, Hermeking H. 2011. miR-34 and SNAIL form a double-negative feedback loop to regulate epithelial-mesenchymal transitions. *Cell Cycle* **10**:4256–4271. doi:10.4161/cc.10.24.18552

Simões-Costa M, Bronner ME. 2015. Establishing neural crest identity: a gene regulatory recipe. *Development* **142**:242–257. doi:10.1242/dev.105445

Soldatov R, Kaucka M, Kastriti ME, Petersen J, Chontorotzea T, Englmaier L, Akkuratova N, Yang Y, Häring M, Dyachuk V, Bock C, Farlik M, Piacentino ML, Boismoreau F, Hilscher MM, Yokota C, Qian X, Nilsson M, Bronner ME, Croci L, Hsiao W-Y, Guertin DA, Brunet J-F, Consalez GG, Ernfors P, Fried K, Kharchenko PV, Adameyko I. 2019.

Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* **364**:eaas9536. doi:10.1126/science.aas9536

Street K, Risso D, Fletcher RB, 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. doi:10.1186/s12864-018-4772-0

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

Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. 2021. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**:1333–1341. doi:10.1038/s41592-021-01282-5

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**:15545–15550. doi:10.1073/pnas.0506580102

Tan TZ, Miow QH, Miki Y, Noda T, Mori S, Huang RY, Thiery JP. 2014. Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol Med* **6**:1279–1293. doi:10.15252/emmm.201404208

Tanay A, Regev A. 2017. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**:331–338. doi:10.1038/nature21350

Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J, Hartwell K, Onder TT, Gupta PB, Evans KW, Hollier BG, Ram PT, Lander ES, Rosen JM, Weinberg RA, Mani SA. 2010. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc Natl Acad Sci U S A* **107**:15449–15454. doi:10.1073/pnas.1004900107

Thibodeau A, Eroglu A, McGinnis CS, Lawlor N, Nehar-Belaid D, Kursawe R, Marches R, Conrad DN, Kuchel GA, Gartner ZJ, Banchereau J, Stitzel ML, Cicek AE, Ucar D. 2021. AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. *Genome Biol* **22**:252. doi:10.1186/s13059-021-02469-x

Thiery JP, Acloque H, Huang RYJ, Nieto MA. 2009. Epithelial-mesenchymal transitions in development and disease. *Cell* **139**:871–890. doi:10.1016/j.cell.2009.11.007

Thong T, Wang Y, Brooks MD, Lee CT, Scott C, Balzano L, Wicha MS, Colacino JA. 2020. Hybrid Stem Cell States: Insights Into the Relationship Between Mammary Development and Breast Cancer Using Single-Cell Transcriptomics. *Front Cell Dev Biol* **8**:288. doi:10.3389/fcell.2020.00288

Todd DJ, McHeyzer-Williams LJ, Kowal C, Lee A-H, Volpe BT, Diamond B, McHeyzer-Williams MG, Glimcher LH. 2009. XBP1 governs late events in plasma cell differentiation and is not required for antigen-specific memory B cell development. *Journal of Experimental Medicine* **206**:2151–2159. doi:10.1084/jem.20090738

Townes FW, Hicks SC, Aryee MJ, Irizarry RA. 2019. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol* **20**:295. doi:10.1186/s13059-019-1861-6

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. 2014. The dynamics and regulators of

cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**:381–386. doi:10.1038/nbt.2859

Tsai JH, Donaher JL, Murphy DA, Chau S, Yang J. 2012. Spatiotemporal regulation of epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. *Cancer Cell* **22**:725–736. doi:10.1016/j.ccr.2012.09.022

Van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, Bierie 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. doi:10.1016/j.cell.2018.05.061

van Groningen T, Koster J, Valentijn LJ, Zwijnenburg DA, Akogul N, Hasselt NE, Broekmans M, Haneveld F, Nowakowska NE, Bras J, van Noesel CJM, Jongejan A, van Kampen AH, Koster L, Baas F, van Dijk-Kerkhoven L, Huizer-Smit M, Lecca MC, Chan A, Lakeman A, Molenaar P, Volckmann R, Westerhout EM, Hamdi M, van Sluis PG, Ebus ME, Molenaar JJ, Tytgat GA, Westerman BA, van Nes J, Versteeg R. 2017. Neuroblastoma is composed of two super-enhancer-associated differentiation states. *Nat Genet* **49**:1261–1266. doi:10.1038/ng.3899

Van Keymeulen A, Lee MY, Ousset M, Brohée S, Rorive S, Giraddi RR, Wuidart A, Bouvencourt G, Dubois C, Salmon I, Sotiriou C, Phillips WA, Blanpain C. 2015. Reactivation of multipotency by oncogenic PIK3CA induces breast tumour heterogeneity. *Nature* **525**:119–123. doi:10.1038/nature14665

Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**:15. doi:10.1186/s13059-017-1382-0

Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, Rajewsky N, Simon L, Theis FJ. 2019. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* **20**:59. doi:10.1186/s13059-019-1663-x

Wolock SL, Lopez R, Klein AM. 2019. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**:281-291.e9. doi:10.1016/j.cels.2018.11.005

Wu H, Lai C-F, Chang-Panesso M, Humphreys BD. 2020. Proximal Tubule Translational Profiling during Kidney Fibrosis Reveals Proinflammatory and Long Noncoding RNA Expression Patterns with Sexual Dimorphism. *JASN* **31**:23–38. doi:10.1681/ASN.2019040337

Xiang R, Wang W, Yang L, Wang S, Xu C, Chen X. 2021. A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Front Genet* **12**:646936. doi:10.3389/fgene.2021.646936

Yamada R, Okada D, Wang J, Basak T, Koyama S. 2021. Interpretation of omics data analyses. *J Hum Genet* **66**:93–102. doi:10.1038/s10038-020-0763-5

Ye X, Tam WL, Shibue T, Kaygusuz Y, Reinhardt F, Eaton E, Weinberg RA. 2015. Distinct EMT programs control normal mammary stem cells and tumour-initiating cells. *Nature* **525**:256. doi:10.1038/nature14897

Yeo SK, Zhu X, Okamoto T, Hao M, Wang C, Lu P, Lu LJ, Guan J-L. 2020. Single-cell RNA-sequencing reveals distinct patterns of cell state

heterogeneity in mouse models of breast cancer. *eLife* **9**:e58810. doi:10.7554/eLife.58810

Youssef KK, Lapouge G, Bouvrée K, Rorive S, Brohée S, Appelstein O, Larsimont J-C, Sukumaran V, Van De Sande B, Pucci D, Dekoninck S, Berthe J-V, Aerts S, Salmon I, Del Marmol V, Blanpain C. 2012. Adult interfollicular tumour-initiating cells are reprogrammed into an embryonic hair follicle progenitor-like fate during basal cell carcinoma initiation. *Nat Cell Biol* **14**:1282–1294. doi:10.1038/ncb2628

Youssef KK, Nieto MA. 2024. Epithelial–mesenchymal transition in tissue repair and degeneration. *Nat Rev Mol Cell Biol*. doi:10.1038/s41580-024-00733-z

Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. 2010. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**:976–978. doi:10.1093/bioinformatics/btq064

Yu G, Wang L-G, Han Y, He Q-Y. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**:284–287. doi:10.1089/omi.2011.0118

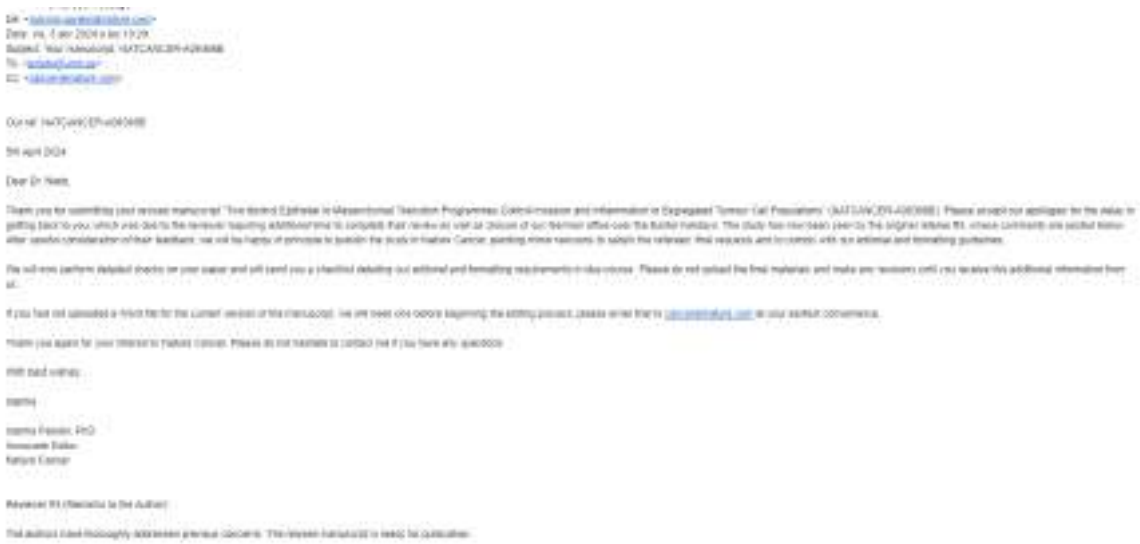
Yu M, Bardia A, Wittner BS, Stott SL, Smas ME, Ting DT, Isakoff SJ, Ciciliano JC, Wells MN, Shah AM, Concannon KF, Donaldson MC, Sequist LV, Brachtel E, Sgroi D, Baselga J, Ramaswamy S, Toner M, Haber DA, Maheswaran S. 2013. Circulating Breast Tumor Cells Exhibit Dynamic Changes in Epithelial and Mesenchymal Composition. *Science* **339**:580–584. doi:10.1126/science.1228522

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**:R137. doi:10.1186/gb-2008-9-9-r137

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**:14049. doi:10.1038/ncomms14049

ANNEX: PUBLICATION

Acceptance letter



Two distinct Epithelial to Mesenchymal Transition Programmes Control Invasion and Inflammation in Segregated Tumour Cell Populations

Khalil Kass Youssef[1], Nitin Narwade[1,*], Aida Arcas[1,6*], Angel Marquez-Galera[1,*], Raúl Jiménez-Castaño[1,*], Cristina Lopez-Blau[1], Hassan Fazilaty[1,7], David García-Gutierrez[1], Amparo Cano[2,3], Joan Galcerán[1,4], Gema Moreno-Bueno[2,3,5], Jose P. Lopez-Atalaya[1] and M. Angela Nieto[1,4,8]

1- Instituto de Neurociencias (CSIC-UMH), Sant Joan d'Alacant, Spain
2- Instituto de Investigaciones Biomédicas "Sols-Morreale" CSIC-UAM, Madrid, Spain
3- CIBERONC Centro de Investigación Biomédica en Red de Cancer, ISCIII, Spain
4- CIBERER, Centro de Investigación Biomédica en Red de Enfermedades Raras, ISCIII, Spain.
5- MD Anderson Cancer Center International Foundation, Madrid, Spain.Present address: Department of Gene Therapy and Regulation of Gene Expression, Center for Applied Medical Research, University of Navarra, Pamplona, Spain.
6- Present address: Department of Gene Therapy and Regulation of Gene Expression, Center for Applied Medical Research, University of Navarra, Pamplona, Spain.
7- Present address: Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
8- Correspondence: anieto@umh.es

*These authors contributed equally to this work

Summary

The Epithelial to Mesenchymal transition (EMT) triggers cell plasticity in embryonic development, adult injured tissues and in cancer. Combining the analysis of EMT in cell lines, embryonic neural crest and mouse models of renal fibrosis and breast cancer we find that there is not a cancer-specific EMT programme. Instead, cancer cells dedifferentiate and bifurcate into two distinct and segregated cellular trajectories after activating either embryonic-like or adult-like EMTs to, respectively, drive dissemination or inflammation. We show that Snail1 acts a pioneer factor in both EMT trajectories and Prrx1 drives the progression of the embryonic-like invasive trajectory. We also find that the two trajectories are plastic and interdependent, as the abrogation of the EMT invasive trajectory not only prevents metastasis but also enhances inflammation, increasing the recruitment of antitumour macrophages. Our data unveil an additional role for EMT in orchestrating intratumour heterogeneity, driving the distribution of functions associated with either inflammation or metastatic dissemination.

Epithelial plasticity is at the core of crucial processes including embryonic cell migration, cancer progression, organ fibrosis and tissue repair[1–4]. The epithelial to mesenchymal transition (EMT) triggers cell plasticity in all these contexts, highlighting its pleiotropy and intrinsic complexity. The EMT is not a binary process or a single programme, as it implies the generation of intermediate hybrid epithelial-mesenchymal (E/M) states that, often, never reach the full mesenchymal state[4,5]. The EMT frequently endows cells with invasive and migratory properties used by both embryonic and cancer cells to disseminate and then colonise to form tissues or metastases, respectively[1]. In other contexts such as during the progression of organ fibrosis, cells do not migrate, as they are unable to activate the invasion process[6,7]. The latter has been defined as a partial EMT, with cells showing a hybrid non-invasive E/M phenotype. However, this is not the only type of partial EMT, as during cancer progression, cells with a hybrid E/M phenotype are associated with invasion and increased metastatic potential[8–10]. The intermediate states together with the intrinsic cell plasticity and transient nature of the EMT have also contributed to the complexity in the analysis of the process[5,11]. Seminal studies have classified EMT states in cancer cell lines and animal models[8–10,12]. Our horizontal approach using both physiological and pathological models provides useful information not (only) on gene signatures, cell types or identities but mainly on biological activities, some beyond migration, invasion and metastasis. We describe two types of EMT that, reflecting the embryonic and adult cell responses occurring during embryonic development and organ fibrosis, are simultaneously implemented in primary tumours to respectively control invasion and inflammation. We also find that two EMT transcription factors (EMT-TFs) behave differently with

respect to the two trajectories. While Snail1 is activated as a pioneer factor at the base of both trajectories, Prrx1, is specific for the invasive trajectory. Compatible with this, *Snail1* mutant cancer cells can hardly develop tumours and *Prrx1* mutant tumours are barely invasive, with the progression towards metastasis highly impaired. Importantly, the truncation of the embryonic-like invasive trajectory upon Prrx1 deletion leads to an enhancement of the adult-like inflammatory trajectory, unveiling the plasticity and interdependence of the two identified EMT pathways and opening new avenues for the design of pathway-specific anti-EMT therapies.

Snail1 is a pioneer factor in EMT induction

To understand the diverse EMT phenotypes along the E-M spectrum, we first used the available whole genome transcriptome analysis of human breast cancer cells lines[13] and found three groups representing the Epithelial (E), Hybrid (E/M) and Mesenchymal (M) phenotypic states (Fig. 1a and Extended Data Fig. 1 and Supplementary Table 1), compatible with previous studies[8,12]. We found that the position in the EMT spectrum from the epithelial to the mesenchymal phenotype correlates with the level of activation of EMT-TF families (Fig. 1b), except for Snail factors, which do not show significant changes in expression between the hybrid E/M and mesenchymal (M) cell lines.

To assess the dynamics of EMT-TF activation in the progression towards the mesenchymal phenotype, we used two MDCK epithelial cell sublines and TGFβ treatment, the classical and most potent EMT inducer[14]. TGFβ signalling was robustly activated in both cell lines (Extended Data Fig. 1b-d) and both activated EMT in response to TGFβ. MDCK-II cells maintained residual cell-cell adhesion and displayed weak mesenchymal activation, therefore showing a hybrid E/M phenotype reminiscent of a stable partial EMT. In contrast, MDCK-NBL2 cells underwent a first transition to partial EMT followed by a fast and robust mesenchymal transition typical of a full EMT (Fig. 1c,d and Extended Data Fig. 1e-i). The kinetics of EMT-TFs expression relative to the onset of epithelial repression (t_{ep}) and mesenchymal activation (t_m) during the treatment with TGFβ showed that *SNAIL1* was rapidly activated in both cell lines showing a two-wave dynamics (Fig. 1e), as previously observed[15]. A first fast burst was followed by the onset of epithelial (E) repression t_{ep} (30min-8h) and the activation of early M genes, compatible with the role of Snail1 as a pioneer upstream regulator of the EMT programme[15]. As such, Snail1 interference prevents TGFβ-induced EMT in both cell lines (Extended Data Fig. 2a), attenuating the repression of epithelial markers and the activation of mesenchymal markers in NBL-2 cells (Extended Data Fig. 2b). The second wave of Snail1 coincided with the recruitment of other EMT-TFs and the enhanced regulation of different E and M markers (Fig. 1e,f). Interestingly, *TWIST1* and *PRRX1* expression was only found in the

progression of MDCK-NBL2 cells from partial to full EMT. Altogether, this is compatible with our data in cancer cell lines, revealing a conserved EMT-TF recruitment associated with EMT phenotypic states along the E to M spectrum (Fig. 1g).

Prrx1 is required for the invasive EMT phenotype

When both MDCK sublines were cultured in 3D collagen matrices, they formed polarized hollowed spheres (Fig. 2a). When treated with TGFβ, only MDCK-NBL2 cells displayed frequent protrusive events (Fig. 2a) concomitant with the loss of epithelial and the gain of mesenchymal markers (Extended Data Fig. 3a), altogether hallmarks of invasive behaviour. To identify an EMT invasive signature, we performed bulk RNA-seq and selected genes specifically upregulated in TGFβ-treated MDCK-NBL2 compared to TGFβ-treated MDCK-II cells. The genes enriched in the NBL2 cells were able to segregate basal-like[16] among all cancer cell lines, an aggressive type known to undergo EMT[17], into two clusters according to their invasive capacities (Fig. 2b). This allowed the identification of 259 genes that we refer to as breast cancer pro-invasion genes (BC-PINGs). Besides EMT, gene ontology identified developmental and invasion programmes selectively enriched in TGFβ-treated MDCK-NBL2 cells (Fig. 2c) and in the BC-PING signature (Extended Data: Fig. 3b and Supplementary Table 2), prompting us to analyse our EMT associated BC-PINGs in a prototypical embryonic invasive EMT programme, the delamination and migration of the neural crest[14].

We used single-cell transcriptomic analysis of embryonic trunk neural crest[18] to build a connectivity map (Fig. 2d) and reconstruct the transcriptional programme during delamination and migration excluding bifurcations to differentiation (clusters 5 and 6; Fig. 2d). The resulting trajectory is associated with an increase in the EMT programme (Hallmark-EMT). Although invasive properties are already present in the delaminating cells, the invasive signature (BC-PINGS) is maintained as migration proceeds towards the mesenchymal phenotype (Fig. 2e,f). Except for Zeb2, already expressed before neural crest induction[19] and not associated with invasion in some contexts[20], this trajectory concurs with a sequential activation of EMT-TFs (Fig. 2g), as we have described it in TGFβ-treated MDCK-NBL2 cells progressing towards the mesenchymal phenotype (Fig. 1). Prrx1, specifically activated at advanced EMT in all models analysed (Fig. 1b,e and Fig. 2g), can stabilize the mesenchymal phenotype in the migratory crest population as it generates ectomesenchyme derivatives, e.g. cartilage[21] or connective tissue. As high levels of PRRX1 accompany the progression to the full M phenotype in different contexts, we examined whether PRRX1 was a requirement for this transition. Knockdown of PRRX1 (si*PRRX1,* over 90% reduction of long and short isoforms) prevented the full EMT induced in MDCK-NBL2 cells by TGFβ, and reverted the EMT status to a partial EMT when administered after the mesenchymal transition was complete

(Fig. 2h and Extended Data Fig. 3c,d). When *PRRX1* was knocked down in MDCK-NBL2 cells, the mesenchymalisation observed upon TGFβ treatment was highly attenuated, and did not affect MDCK-II cells, as expected (Extended Data Fig. 3f). Bulk RNA-seq confirms that *PRRX1* knockdown (si*PRRX1*) in MDCK-NBL2 cells attenuates the repression of the epithelial programme, prevents full activation of mesenchymal genes including other EMT-TFs (Extended Data Fig. 3g), represses developmental programmes associated with cell migration (Fig. 2i), and cells lose the invasive signature (Fig. 2j, k). Thus, in the absence of *PRRX1* activation, MDCK-NBL2 cells reach an end-state in their response to TGFβ like the non-invasive partial EMT observed in MDCK-II cells. Interestingly, ectopic expression of *PRRX1* was sufficient to promote invasiveness in MDCK-II cells treated with TGFβ (Fig. 2l), supporting the role of PRRX1 in inducing the invasive EMT phenotype.

Several signalling pathways are enriched in the TGFβ-induced invasive versus non-invasive EMT, in particular focal adhesion kinase (FA), strongly dependent on PRRX1 (Extended Data Fig. 3h-i). Sublethal doses of Focal Adhesion Kinase inhibitors (FAKi) prevented the induction of a full EMT by TGFβ in NBL2 cells and induced a partial reversion (MET) if administered after cells have undergone full EMT (Extended Data Fig. 3j,k). In contrast, MDCK-II cells exposed to TGFβ showed a similar morphology and conserved EMT response irrespective of the presence of FAKi (Extended Data Fig. 3l,m). Thus, the impact of FA inhibitors is similar to that of *PRRX1* knockdown, and the activation of high PRRX1 levels and FA signalling associates with invasive EMT and promotes transition to full mesenchymal phenotype in embryos and cancer cells.

Partial and inflammatory non-invasive EMT in renal fibrosis

In contrast to the invasive (embryonic or tumoural) EMT, a non-invasive partial EMT is activated in renal fibrosis[6,7], as observed after Unilateral Ureteral Obstruction (UUO), which induces tubular injury progressively evolving into renal interstitial fibrosis and renal failure[22]. In this model, Snail1 activates EMT in the tubular cells, which dedifferentiate but do not become invasive, remaining in the damaged tubules and secreting chemokines and cytokines that promote fibrogenesis and inflammation (Fig. 3a)[6,7]. This is confirmed by the absence of red (tdTomato) cells in the stroma after genetic labelling of renal epithelial cells and UUO (Fig. 3b). E-cadherin (Cdh1) reduction, and Vimentin (Vim) activation in the same cells confirm the existence of a hybrid E/M phenotype (Fig. 3c). Residual E-cadherin and Tight-junction protein 1 help to maintain some cell-cell adhesion (Fig. 3d). All of this confirms the partial and non-invasive EMT, where epithelial cells do not become fibroblasts[6,7], in agreement with the recent demonstration that myofibroblasts derive from fibroblasts (and pericytes) in human fibrosis[23].

Following a droplet-based single-cell RNA-seq of sham-operated and obstructed kidneys, unsupervised graph-based clustering (see Methods) organized cells into 26 major cell clusters representing the different cell types shown on uniform manifold approximation and projection (UMAP) (Fig. 3e and Extended Data Fig. 4a-e), all identified by the expression of *bona fide* lineage markers[23–25]. We found a dramatic remodelling in the non-epithelial component in obstructed kidneys, with a massive increase in interstitial and immune cells, as previously observed[26]. In the epithelial component, the appearance of a cluster of injured cells is concomitant with the disappearance of *bona fide* proximal tubular (PT) cells (Fig. 3f and Extended Data Fig. 4f), identified as major contributors to the injured population (approx. 90%) (Fig. 3g and Extended Data Fig. 5a), consistent with the coexpression of PT and injury markers 1 day after UUO (Fig. 3g and Extended Data Fig. 5b,c), and with previous data[22]. The reclustering of PT (Dashed box in Fig. 3g) and PT-injured cells showed that damaged cells had activated an EMT programme concomitant with the loss of renal epithelial differentiation and the acquisition of an injury inflammatory programme[27] (Fig. 3h,i). Trajectory reconstruction using Partition-based graph abstraction (PAGA)[28], Velocity[29], and transcriptional regulatory network computation by SCENIC[30], confirmed the progression of EMT states in parallel to the increase in injury markers (Fig. 4a,b and Extended Data Fig. 6a-c). Importantly, the EMT programme, including epithelial dedifferentiation, mesenchymalysation and injury markers, significantly decreased when UUO was performed in mice bearing Snail1-deficient (*Snail1* cKO) renal epithelial cells (Fig. 4c,d), confirming the role of Snail1 in triggering tubulointerstitial inflammation and fibrosis[6,7], the activation of EMT and validating the regulatory networks predicted by SCENIC (Fig. 4b,d), during injury response in the adult kidney.

Unlike the invasive EMT found in TGFβ-treated MDCK-NBL2 cells, neural crest and breast cancer cells, the EMT programme activated in the damaged PT cells was enriched in genes associated with injury response (e.g. *Vcam1, Jun/Fos, Egr1*), inflammation (e.g. *Ccl2 and 5, Nfkbia*, *Notch1 and 3*) and fibrogenesis (e.g. *TGFβ* 1 and 2 and metalloproteinase inhibitors *Timp1 and 2*) (Fig. 4e and Extended Data Fig. 6c). This EMT inflammatory programme was confirmed by the analysis of enriched pathways (Fig. 4f) and the localization of hallmark injury markers (Fig. 4g and Extended Data Fig. 6 d,e). Genes encoding inflammatory cytokines and chemokines are expressed by damaged epithelial cells that dedifferentiate and remain in the injured tubules (Fig. 4e,g and Extended Data Fig. 6c-e). Several EMT-TFs are also activated in these damaged cells (Extended Data Fig. 6f), but the absence of Prrx1 (Fig. 4h, expression only detected in the stroma) can explain their failure to invade (see Figs. 1 and 2). Hence, PT cells acquired a stable partial EMT phenotype with residual cell-cell junctions. Interestingly, the MDCK-II cells that respond to TGFβ undergoing a non-invasive partial EMT (Fig. 2) are also enriched in pathways associated with inflammation

(Extended Data Fig. 6g). Thus, here we characterise the partial EMT programme in the epithelial cells during fibrosis as the trigger of dedifferentiation compatible with renal insufficiency and accompanied by a repair/inflammatory phenotype that secretes fibrogenic and inflammatory cytokines, influencing the stroma in a paracrine manner to promote the progression of the disease. Altogether, we describe the inflammatory EMT programme as the response to injury of adult non-transformed cells.

200

Progenitor-like EMT phenotypes in tumours

As the EMT is pathologically activated in primary tumours to favour cancer cell dissemination, we extended our studies to a widely used breast cancer model, the MMTV-PyMT[31], carcinomas that progress to the invasive and metastatic state resembling human invasive breast cancer[32]. We tagged mammary gland progenitor cells from early embryonic stages[33], to detect all cancer cells and discriminate them from those in the tumour microenvironment (TME) (Fig. 5a,b). For the transcriptomic analysis at the single-cell resolution we followed a strategy similar to that used in renal fibrosis and profiled advanced metastatic primary tumours (Extended Data Fig. 7a,b). Cells were organized into 5 major clusters (Fig. 5c, left panel) representing cancer cells (CC) and
210 associated populations, all identified by the expression of specific markers (Fig. 5c, left panel and Extended Data Fig. 7c-e). As in the kidney, we focused on the epithelial component, the cancer cells identified by the expression of td-Tomato (CC; Extended Data Fig. 8a). CC were subdivided into 17 clusters represented across the four tumour samples (Fig. 5c, right panel and Extended Data Fig. 8b), validating our experimental approach, and showing that the progression of PyMT mammary gland carcinomas is very stereotyped. Using luminal and basal gene signatures for mammary epithelial cells (MECs)[34,35], we found that as expected from the luminal origin of PyMT tumours, the majority of clusters (around 70% of the CC) had a Luminal Alveolar (LA) phenotype (Fig. 5d and Extended Data Fig. 8c). In addition, we observed clusters with transcriptomes compatible with different progenitor states, reminiscent of a Luminal Alveolar Stem/Progenitor state (LAP)[35,36]
220 (clusters 1, 13 and 15); a Pan-Luminal Stem/Progenitor state (PLP)[34,37] (cluster 10); a hybrid state combining PLP, luminal Hormone Sensing (HS) and Baso-Myoepithelial (BM) phenotype (cluster 14), compatible with a Luminobasal Bipotent Progenitor (LBP) state induced by the oncogene[38,39] and reprogramming towards a developmental progenitor-like state[40,41]. Clusters 12 and 16 acquire a basal programme (BM) compatible with the progression towards the invasive phenotype[42]. Thus, we observed a series of phenotypes associated with dedifferentiation and reprogramming towards progenitor, multipotent-like states (Fig. 5d and Extended Data Fig. 8c). This reprogramming occurs concomitantly with EMT and the progression along the E/M spectrum (Fig. 5e and Extended Data Fig. 8d-f). LA clusters show the highest level of epithelial markers, whereas partial EMT states

(clusters 1, 13 and 15) are associated with mammary gland progenitor-like lineages and stemness markers, and cluster 16 expresses high *Prrx1* levels, reminiscent of cells in a full EMT state (Fig. 5e and Extended Data Fig. 8d-f). Thus, the progression towards more dedifferentiated phenotypes concurs with the sequential recruitment of EMT-TFs, as observed in cancer cell lines and during TGFβ-induced EMT (Fig. 1), revealing the parallel progression of mammary cell dedifferentiation and EMT states.

Two distinct EMT programmes in cancer

Following again a similar strategy to that used in neural crest and the kidney, we found that cancer cell EMT clusters, rather than appearing ordered in a linear trajectory, were organized in a branched structure with two discrete paths bifurcating from the bulk of luminal cancer cell clusters at the level of cluster 11 (Fig. 6a). RNA velocity[29] analysis of individual cells was compatible with this organisation and inferred the directionality of the two trajectories (Fig. 6b). Next, we performed SCENIC analysis[30] that revealed cell state-specific transcriptional regulators across the two EMT trajectories (Fig. 6c,d). Applying pseudotime inference we reconstructed their corresponding molecular programmes (Fig. 6e). In the EMT-T1 branch, in addition to losing lumino-alveolar differentiation markers and progressing towards a mesenchymal phenotype, cancer cells acquired a pro-invasive gene profile from cluster 14 (Fig. 6e and Extended Data Fig. 9a). The recruitment of EMT factors resembles that in the full EMT response of MDCK-NBL2 cells to TGFβ, with *Snail* genes followed by *Zeb1* and *Prrx1* (Extended Data Fig. 9a), compatible with a progression to cancer cell dissemination. In contrast, in EMT-T2, the lumino-alveolar epithelial phenotype of cluster 11 progresses to the partial EMT phenotype of cell clusters 1, 13 and 15, still maintaining expression of epithelial genes and activating a limited mesenchymal programme (Fig. 6e and Extended Data Fig. 9a). *Snail1* is the only EMT-TF significantly detected in the EMT-T2 trajectory in cancer cells (Extended Data Fig. 9a), resembling the partial EMT programme observed during renal fibrosis. In relation to this, the most significant trait in EMT-T2 is the remarkable enrichment in inflammatory and pro-fibrotic genes. In sharp contrast to the highly proteolytic and invasive EMT-T1 gene signature enriched in matrix metalloproteinases (*Mmp 2, 3, 13* and *14*), EMT-T2 is enriched in metalloproteinase inhibitors (*Timp1, 2* and *3*), consistent with its non-invasive and pro-fibrotic profile (Fig. 6e and Extended Data Fig. 9a). Altogether, our single-cell transcriptomic analyses reveal that within the same tumour cancer cells progress along two different EMT trajectories both associated with dedifferentiation but with either a pro-invasive or an inflammatory phenotype. Analysis of enriched biological processes confirms the existence of trajectory-specific functions, namely invasion (EMT-T1) and inflammation (EMT-T2), respectively characteristic of embryonic and adult EMTs (Fig. 6e-g and Extended Data Fig. 9b,c).

Embryonic and adult EMTs for cancer progression

EMT-T1 and EMT-T2 markers show a non-overlapping localization, with the invasive EMT-T1 cells located at the tumour margins, and EMT-T2 cells distributed within the tumour (Fig. 7a,b). This structure was reproduced in tumouroids derived from disaggregated PyMT primary tumours cultured in 3D collagen matrices (Fig. 7c-e). Tumouroids also confirm the invasive nature of cancer cells expressing EMT-T1 markers (Fig. 7d). We next examined the distribution of EMT-T1 and T2 markers in triple negative breast cancer (TNBC) samples from human patients and observed that they are also expressed in non-overlapping populations (Fig. 7f). Furthermore, the clusters characterised in T1 and T2 trajectories can be identified in breast luminal cancer and enriched in TNBC (Fig. 7g). Thus, individual mouse and human tumours that progress to the aggressive basal-like phenotype even if they are of luminal origin[17,32,43], can bear segregated cell populations that have undergone EMT with either an embryonic-like pro-invasive phenotype or an adult progenitor phenotype with pro-inflammatory and pro-fibrotic properties.

Plasticity of invasive and inflammatory EMT trajectories in tumours

To confirm the proposed functions of the two trajectories, we decided to challenge them by deleting EMT-TFs. We first generated mice bearing PyMT tumours deficient for Snail1, activated in both T1 and T2 trajectories (Fig. 5e and Extended Data Fig. 9a). Snail1 deficiency (*Snail1 cKO,* Fig. 8a) dramatically reduced both the number of and size of tumours (Fig. 8b), compatible with the described early activation of Snail1 in luminal cells and its ability to confer tumour-initiating capacities in the PyMT model[44]. In agreement with a pioneer role in the activation of EMT, the majority of the few and small Snail1-deficient tumours were highly differentiated, in clear contrast with control tumours (CTR), undifferentiated and compatible with grade 3 (Fig. 8c). The strong impact of Snail1 loss in breast tumour development, did not allow trajectory analysis, but we revealed its regulatory role in inflammatory EMT in a human inflammatory cell line (see below). We challenged T1 trajectory generating mice bearing Prrx1-deficient tumours (Fig. 5e and Extended Data Figs. 8f and 9a). In contrast to Snail1 loss, Prrx1 deficiency (*Prrx1 cKO;* Fig. 8d and Extended Data Fig. 10a-d), did not modify the number or the size of the tumours (Fig. 8e), although they were less advanced that the controls, containing areas typical of carcinoma *in situ* (Fig. 8f).

We found cancer cells coexpressing Prrx1 and the mesenchymal marker vimentin (Vim) that are close to the tumour border (Fig. 8g), and also cells coexpressing Prrx1, the epithelial marker Epcam and vimentin (Fig. 8h). This indicates that Prrx1 is already activated in partial EMT states. As its expression induces invasive properties (Fig. 2), this state is consistent with cells with a hybrid E/M phenotype bearing invasive properties, as those shown to bear increased metastatic

potential[8–10]. We also found that the invasive areas were very much reduced in *Prrx1* mutant tumours (around 7 times) compared to control tumours (Fig. 8i and Extended Data Fig. 10e), compatible with (i) a more differentiated status in Prrx1-deficient tumours, (ii) Prrx1 localization in cancer cells at the periphery of the tumour (Fig. 8g, h) and (iii) Prrx1 association with invasiveness. As expected from a poor invasive activity, mice bearing Prrx1-deficient tumours show a dramatic reduction in lung metastatic burden (Fig. 8j). All of this corresponds to a truncated EMT-T1 trajectory, confirmed by the changes in expression of genes specific for different clusters along the trajectory (Fig. 8k). Interestingly, expression of T2-specific markers increase in *Prrx1*-deficient tumours (Fig. 8k and Extended Data Fig. 10f), including transcriptional regulators *Klf4*, *Junb, Mafb*, and markers of acute inflammation (*Saa1*, *Saa2)* and inflammatory breast cancer (*Egr1* and *Junb)*[45]. On the other hand, PRRX1 knockdown (si*PRRX1*) in TGFβ-induced invasive EMT in MDCK-NBL2 cells was sufficient to increase inflammation-associated genes (Extended Data Fig. 10g). Once confirmed the role of Prrx1 in EMT-1 and invasion, we wanted to assess the role of *Snail1* in EMT-T2, the only EMT-TF expressed in this trajectory (Extended Data Fig. 9a). In the absence of well-developed Snail1-deficient tumours, we downregulated SNAIL1 in the inflammatory breast cancer cell line SUM149PT and found a decrease in the expression of the T2-specific transcriptional regulators and inflammation markers (Fig. 8l), confirming its role in regulating inflammation as it occurs in fibrosis (Fig. 4). Finally, *Prrx1* mutant tumours, with enhanced EMT-T2, consequently show an increase in pro-inflammatory cytokines, but also a decrease in the anti-inflammatory cytokine IL-13 (Fig. 8m), and high infiltration by tumour associated macrophages (Fig. 8n; Extended Data Fig. 10h) of the pro-inflammatory anti-tumoural type (MHC-II positive; Fig. 8o; Extended Data Fig. 10i). Altogether, this confirms cell dissemination and inflammation as key functions respectively associated with EMT-T1 and T2 trajectories and reveals their interdependence in breast cancer cells and PyMT cancer cell evolution.

Discussion

The parallel analysis of the EMT programmes activated in cells treated with TGFβ, during embryonic development and adult organ fibrosis has allowed us not only to define the EMT trajectories and functions associated with each of these processes, but also to better interpret the two alternative EMT trajectories that we have found in cancer. Cancer cells respond to oncogenic activation either as embryonic-like or adult-like cells, leading to different outcomes. The former corresponds to the well-known function of EMT in invasion and dissemination, and the latter to an antitumour inflammatory injury response.

We find that EMT activation is concomitant with dedifferentiation in adult cells, both in

fibrosis and breast cancer. This dedifferentiation step is reminiscent of the lineage infidelity and plasticity described in adult skin wound healing and cancer[46], also observed in other carcinomas[47]. We propose that this plasticity is triggered by the activation of EMT, that also occurs concomitant with cell dedifferentiation in neuroblastoma and melanoma where adrenergic cells or melanocytes, respectively, reactivate embryonic neural crest markers[41,48]. Activation of EMT has also been associated with cell dedifferentiation and the emergence of repair cell states in limb, fin, and heart regeneration in axolotl and zebrafish[4]. Interestingly, for repair, EMT needs to be transient, and a forced transient activation is also consistent with reinstating heart regeneration in mice[49]. EMT is also transiently activated in cancer, as successful metastatic colonization involves downregulation of the EMT programme[50,51]. However, during renal fibrosis, EMT activation is not transient, and although triggering dedifferentiation, it progresses to degeneration and organ failure[6,7]. The EMT is also known to be transiently required at the early stages of reprogramming of adult fibroblasts to iPSCs[52]. Thus, the EMT lies at the core of somatic cell dedifferentiation, as a driver of epimorphosis to achieve phenotypic plasticity in the adult.

We have defined the two different EMT programmes and their trajectories in neural crest and renal fibrosis, representing the response of embryonic and adult cells, respectively (Fig. 8p). We also show that during tumour progression, the EMT-induced initial dedifferentiation step provides the required plasticity that is then followed by two alternative pathways that recapitulate either an embryonic-like or the adult-like response. Cancer cells hijack both developmental and adult EMT plasticity programmes normally used for cell invasion and migration, or as a response to injury, respectively, to implement cell dissemination and antitumoural inflammation. Thus, the embryonic-like trajectory promotes tumour progression towards metastasis, while the adult-like trajectory represents a defense mechanism in response to damage, in this case induced by the oncogene.

Genetic challenge of trajectory 1 by deleting Prrx1 specifically in cancer cells confirms that invasion is the functional property associated with EMT-T1 and that Prrx1 is essential for the progression towards tumour invasion and dissemination. In its absence, the invasive trajectory is truncated and metastatic burden dramatically reduced (Fig. 8p), explaining recent findings where Prrx1 expressing cells were traced as those forming metastasis in melanoma[53]. The downregulation of SNAIL1, the only EMT-TF detected in trajectory 2, confirms in a human inflammatory breast cancer cell line the predicted regulatory structure of the adult-like EMT trajectory in tumours, and reinforces inflammation as its functional property. As such, a subset of cancer cells become inflammatory-like and express genes encoding inflammatory cytokines,

including TNF-α, IL-6, CCL2 and CCL5, like renal epithelial cells during fibrosis[6,7]. Our data are compatible with these inflammatory cytokines attracting macrophages to the tumour, in particular the MHC-II positive antitumour inflammatory population, found in the proximity of the secreting EMT-T2 cancer cells. This response is exacerbated in the *Prrx1* mutant tumours, where the invasive EMT is truncated and more cells engage into the inflammatory EMT trajectory, now sufficient to convert cold tumours into hot tumours, opening avenues for the design of therapeutic approaches. The relative contribution of EMT-T1 and T2 in Prrx1-proficient and deficient tumours points to the interdependence of the two trajectories, that sharing a common origin in the breast luminal cell, can be plastic in response to tumour traits and likely, also to microenvironmental changes, including cancer cell-stromal interactions. As the two EMT programmes operate in different cells (Fig. 8p), individual tumours bear dedicated EMT populations to fulfil specific and very distinct functions, adding another layer of intratumour heterogeneity not only related to the expected different EMT phenotypes (epithelial cancer cells moving along the E to M spectrum) but also to the alluded distribution of antagonistic pro- and antitumour functions, namely dissemination and inflammation. In the latter, the EMT induces antitumour responses, but the response to injury can also lead to degeneration in chronic settings as in fibrosis. Thus, further studies are warranted to examine whether the antitumour inflammatory trajectory can also evolve to favour tumour progression.

Data availability

Bulk RNA-seq data that support the findings of this study have been deposited and are publicly accessible at the Gene Expression Omnibus (GEO) repository (GSE164488). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164488.

scRNA-seq data that support the findings of this study have been deposited and are publicly accessible at the GEO repository (GSE175412 and GSE159478), for kidney and tumours data, respectively. Links to the repository as follows:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE175412

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE159478

Single-cell RNA-seq Data for Neural Crest (Fig. 2) were downloaded from the NCBI Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) and submitted under GEO accession GSE129114. The tSNE embedding and associated metadata (Fig. 2d) were obtained from http://pklab.med.harvard.edu/ruslan/neural_crest/tSNE_main_Fig1.txt.

Raw data have been provided as one Microsoft Excel Source Data file for: Fig. 1 b, e, f; Extended Data Fig. 1c; Extended Data Fig. 2b; Fig. 2 c, e, g, i, k, l; Extended Data Fig. 3 b, c, f, h, m; Fig.4 c, d, f; Extended Data Fig. 4 f; Extended Data Fig. 6 f, g; Fig. 6 f; Extended Data Fig. 9 b, c; Fig. 8 b, c, e, f, I, j, k, l, m, n; Extended Data Fig. 10 g.

All other data supporting the findings of this study are available from the corresponding author on reasonable request.

Code availability

We have not created any custom codes or algorithms in this study. Open-source software was used to analyze data. Details of software versions are specified in the Methods and Report Summary.

Acknowledgements

We thank Berta Sanchez-Laorden for helpful discussions and suggestions all throughout the project, Sonia Vega for her help and support in managing cell lines, Diana Abad and Teresa Maria Gomez for technical support, Giovanna Expósito and Verona Villar Cerviño for the support at the imaging facility. We thank Alerie Guzman De la Fuente for helpful suggestions for macrophages heterogeneity analysis. We also thank Antonio Caler Escribano for technical help in the FACS/Omics facility. We thank the MD Anderson Foundation Biobank for providing samples (record number B.0000745, ISCIII National Biobank). This work was supported by grants MICIU RTI2018-096501-B-I00 and MCI PID2021-125682NB-I00 to MAN, RTI2018-102260-B-I00 to JPLA, and PID2022-136854OB-I00 to GMB all funded by MICIU/AEI /10.13039/501100011033 and by FEDER, UE. Funds were also provided by the AECC Scientific Foundation (FC_AECC PROYE19073NIE to MAN and PROYE19036MOR to GMB), Instituto de Salud Carlos III (CIBERONC, CB16/12/00295 to GMB and AC; CIBERER, CB19/07/00038 to MAN), Generalitat Valenciana (Prometeo 2021/45) and the European Research Council (ERC AdG 322694) to MAN, who also acknowledges financial support from Centro de Excelencia Severo Ochoa» Grant CEX2021-001165-S funded by MCIN/AEI/ 10.13039/501100011033. KKY 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). N.N held 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 MAN). R.J.C. holds a "Severo Ochoa Excellence Programme" PhD contract (PRE2020-091888).

Author contributions

KKY and MAN conceived the project, interpreted the data and wrote the manuscript. KKY performed most of the experiments and analysed the data and MAN supervised the whole project. AA performed the bulk RNA-seq analyses and helped in the *in silico* analyses. AMG and NN performed the single-cell RNA-seq analyses. RJC performed immunofluorescence stainings and analysis in metastatic lungs, kidneys and mouse tumours. CLB performed experimental animal procedures and mouse lines management. HF contributed to the *in vitro* experiments and RT-qPCR in cell lines and obstructed kidneys. DGG performed the cytokine analyses in tumours; GMB provided the breast cancer TMA, performed immunofluorescence analysis in human tumours and, together with AC, analysed and interpreted patient data. KKY and JG generated the *Prrx1* conditional mutant mouse model. JPL-A helped to design the single-cell RNA-seq experiments and supervised its analysis. MAN also ensured funding.

Competing interests

The authors declare no competing interests.

Figure legends

Fig. 1. Epithelial-mesenchymal states associate with conserved EMT-TF expression codes. a, Upper panel, 3-group clustering of 71 breast cancer (BC) cell lines[13] based on epithelial and mesenchymal component enrichments (genes listed in Supplementary Table 1 and see Methods). Middle panel, heat map showing epithelial (E), mesenchymal (M) and Hallmark_EMT (MSigDB; UC San Diego and Broad Institute) enrichment scores in the 71 BC cell lines. Lower panel, relative expression of representative E and M genes. Colour scale, relative transcript levels (log2) from lowest (dark blue) to highest (bright red). x-axis: BC cell lines; y-axis: genes. **b**, Relative expression of EMT-TFs in cells of the E (n=42), E/M (n=14) and M (n=14) groups. Boxes: medians and interquartile ranges (IQRs); whiskers: minimum and maximum values. P values were determined using an unpaired two-sided *t*-test. **c,** Brightfield images of untreated or TGFβ-treated MDCK-II

and MDCK-NBL2 cells. **d**, IF images for the epithelial tight junction protein 1 (TJP1) and mesenchymal FN1 (Fibronectin 1) markers during TGFβ treatment. Images representative of at least three biological replicates in (c) and (d). Scale bars, 20µm in **c** and **d. e**, EMT-TFs transcript levels (RT-qPCR) during TGFβ treatment. Fold change (FC) to untreated control cells (CTR). FC is represented as mean ± SEM (n⩾4 biological replicates in MDCK-II CTR and TGFβ; n⩾3 in MDCK-NBL2 CTR and TGFβ). A correction factor (*cf*) is applied to adjust the FC representation in MDCK-II cells. *cf*: ratio of gene expression in the two MDCK cell lines before TGFβ treatment (time=0). Time points on the x-axis are plotted in a non-proportional scale to better follow the early time points. **f**, RT-qPCR showing relative transcripts levels (fold change) in TGFβ treated vs untreated (CTR) NBL2 cells for epithelial (left panel) and mesenchymal (right panel) markers. FC is represented as mean ± SEM (n⩾3 biological replicates for CTR and TGFβ groups). τep and τm stand for the onset of repression or activation of epithelial or mesenchymal genes, respectively. **g**, Proposed model for the progression from the epithelial to the mesenchymal phenotype (breast cell lines and TGFβ-treated MDCK cells).

Fig. 2. EMT in developing embryos, epithelial cells and invasive cancer cells. a, Analysis of TGFβ (1 ng/ml)-induced invasiveness in MDCK cells cultured for 10 days in 3D Collagen matrices. Phalloidin staining reveals actin filaments (F-ACTIN). TGFβ was administered after spheres formation. Nuclei in blue. Scale bars, 20µm. **b**, Differentially expressed genes (DEG) upregulated in TGFβ-treated MDCK-NBL2 cells were used to cluster basal-like breast cancer cell lines. MCF10A was used as a control for non-tumourigenic and non-invasive cells. Heat map: two main clusters segregate invasive (left) from non-invasive (right) cells. Y-axis: genes enriched ≥ 1.75X in the averaged invasive vs non-invasive basal-like breast cancer cell lines, hereinafter referred to as breast cancer-pro-invasion genes (BC-PINGs) (genes listed in Supplementary Table 3). Colour scale as in Fig. 1a. **c**, Dot plot showing selected GO terms enriched in MDCK NBL2 vs MDCK II cells in response to TGFβ in 2D cultures. **d**, Top, trunk neural crest (NC) populations in E9.5 mouse embryos. Bottom, corresponding single cell t-distributed stochastic neighbour (t-SNE) embedding and connectivity map as predicted by Partition-Based Graph Abstraction (PAGA)[28] obtained using data from[18]. NT: Neural tube, Pre-M-NC: Pre-Migratory NC; Del-NC: Delaminating–NC; MP-NC: Migratory Progenitor-NC; NC-Mes: NC derived Mesenchymal cells; Sen-N: Sensory-Neurons; Auto-NS: Autonomous-Nervous System. Clusters 5 and 6 correspond to bifurcations towards neural differentiation and therefore, involve the known repression of the EMT programme and are not used in the subsequent analysis. **e,** Heatmap with transcriptional programme changes in the single-cell trajectory from NT to NC-Mes (clusters 0 to 4; genes predicted by Moran's I test

with q-values < 0.001 and ordered over pseudotime using scVelo[54]. See Supplementary Table 3 for full gene names. Right panel, dot plot showing enrichment in Hallmark_EMT and BC-PINGs in each of the indicated populations. Enrichment score=number of overlapping genes in the populations and Hallmark_EMT or BC-PINGs/number of overlaps expected by chance. Significance of enrichments assessed by hypergeometric p-value. **f,** Enrichments represented over the t-SNE embedding shown in (d). **g**, Relative expression of EMT-TFs in trunk NC migratory trajectory. Colour scale as in Fig. 2b. **h**, MDCK-NBL2 and MDCK-II cells treated with small interfering RNA for PRRX1 (si*PRRX1*) or siCTR 8 hours before treatment with TGFβ, or once cells have undergone EMT (6 days after TGFβ administration). Scale bars, 20μm. **i,** Dot plot showing selected GO terms for downregulated genes after 4 days of TGFβ treatment of MDCK-NBL2 cells in si*PRRX1* vs siCTR conditions. **j**, GSEA showing the relative enrichment for BC-PINGs in TGFβ-treated MDCK-NBL2 (invasive EMT) when compared to TGFβ- treated MDCK-II cells (non-invasive EMT response) (upper panel) and loss of the positive enrichment after si*PRRX1* treatment (lower panel). **k**, Transwell invasion assays showing the nuclei (DAPI) of MDCK-NBL2 invasive cells after 2 days of TGFβ treatment in the presence of siCTR or si*PRRX1*. Invading cells represented as mean ± SEM (n=5 biological replicates per condition). **l**, Top, representative brightfield images of TGFβ-treated MDCK-II cells transfected with empty (CTR) or PRRX1 expressing plasmids. Bottom, images and quantification of transwell invasion assays. PRRX1 expression is sufficient to induce cell scattering and invasive properties. Number of invading cells represented as mean ± SEM (n=6 biological replicates per condition). P values: determined using Fisher´s Exact Test for the dot plots (c, e and i); estimated using an empirical phenotype-based permutation test (j); determined using an unpaired two-sided *t*-test (k,l).

Fig 3. Single-cell transcriptomic analysis reveals EMT activation in renal fibrosis. a, Genetic strategy to trace renal epithelial cells (RT) (see Methods), which appear labelled in red (tdTomato). **b**, Expression of the mesenchymal marker Vimentin (Vim) in combination with tdTomato-labelled renal epithelial cells in control (SHAM) and obstructed kidneys (UUO). Dash lines surround renal tubules (RT). Arrows indicate de novo Vimentin activation in RT. Nuclei in blue. Scale bar, 5 μm. **c**, E-Cadherin (Cdh1) and Vimentin (Vim) expression. SHAM panel: the arrow indicates renal epithelial cells, and the arrowhead, a glomerulus. UUO panels: the higher magnification images (box) show E-Cadherin and Vimentin expression in single channels. Arrows: renal epithelial cells positive for both markers. Scale bar, 5 μm. **d**, Tight junction epithelial protein expression (Tjp1). Arrowheads: *puncta adhaerentia* junctions. Nuclei in blue. Scale bar, 10 μm. **e**,

Uniform manifold approximation and projection (UMAP) showing the diversity of cell types in SHAM-operated and obstructed kidneys (10 days post SHAM and UUO; total cell number n=25424). **f**, UMAP and bar plots showing the contribution of different cell populations to control (SHAM) and obstructed (UUO) kidneys. Abbreviations as in (e). PT, Proximal tubules; Inj, Injured tubules. **g**, Upper panel, origin of injured epithelial cells determined using supervised machine learning (see Methods). Dashed box contains the epithelial clusters with major contribution to injury (PT clusters 8, 16 and 5, 89.3%). Lower panel, expression of the injury marker Kim-1 in combination with the PT cell marker LTA. Arrows indicate Kim-1 positive cells also positive for LTA (proximal tubules). Box plot showing the percentage of Kim-1 positive injured cortical epithelial cells also positive for LTA (n= three mice analysed 1d after UUO, with six randomly selected cortex images quantified per kidney). Box: median and IQRs; whiskers: minimum and maximum values. Nuclei in blue. Scale bar, 50µm. **h**, Upper panel, UMAP of injured epithelial cells and clusters contributing to Injury (Dashed box in g). Lower panels, respective contribution in control and obstructed kidneys. **i**, UMAP as in h, cells coloured by the enrichment score for Hallmark_EMT, PT differentiation, and injury associated inflammation (see Methods).

Fig. 4. Non-invasive and inflammatory partial EMT programme in renal fibrosis. a, Left panel, PAGA predicted connectivity map of PT and Injured clusters (see Methods). Right panel, RNA velocity analysis (see Methods) showing the trajectory from differentiated to injured PT cells. **b,** hierarchical clustering of SCENIC computed transcription factor activities (regulons; see Extended Data Fig. 6). The regulon activity represents the mean value of AUCell score per single-cell cluster (see Extended Data Fig. 6). **c**, Expression of Cadherin-16 (Cdh16) (upper panels) and Vimentin (Vim) (lower panels) in non-obstructed (CTR) or in those after 2 weeks of obstruction from *Snail1* proficient (UUO CTR) and *Snail1* deficient kidneys (renal epithelial cell specific *Snail1* knockout mice[4]; UUO *Snail1*cKO), plus the quantification of Cdh16 expression in renal epithelial cells and of the percentage of renal epithelial cells positive for Vimentin. For Cdh16 $n \geqslant 13$ quantified fields from 3 mice for each group. For Vimentin $n \geqslant 22$ quantified fields from 3 mice for each group. **d,** qPCR showing the fold change for depicted differentiation epithelial genes and deregulated transcription factors predicted by SCENIC in kidneys similar to those described in (c). Fold change is represented as mean ± SEM ($n \geq 4$ mice per condition). **e,** changes in the transcriptional programme along the trajectory from PT to Injured cells (genes predicted by Moran's I test and ordered over pseudotime as in Fig. 1h). **f**, Dot plot showing GO terms related to differentiation or Inflammation associated with the indicated cluster groups. **g**, Expression of the injury response and TGF-β target (Krt20), and dedifferentiation (Klf4) markers in combination with a PT

differentiation marker (LTA). Lower panel, Kim-1 and Jun as markers of injury plus LTA. **h,** Prrx1 (green) and renal epithelial cells genetically traced with tdTomato (red); arrowheads indicate Prrx1 exclusive expression in interstitial cells. All markers shown in control and obstructed kidneys. Nuclei in blue. Images in (c), (g) and (h) are representative of three biological replicates per condition. Scale bar, 20μm. P values in (c) and (d) determined using an unpaired two-sided Mann-Whitney U test and Fisher´s Exact Test for the dot plots in (f).

580

Fig. 5. Concomitant dedifferentiation and EMT activation in PyMT breast cancer. a, Experimental design to generate genetically-trackable cancer cells. PyMT activation in luminal cells leads to carcinoma development with all cancer cells labelled by tdTomato but not the stroma (see Methods). **b**, Expression of the mammary epithelial cell reporter (tdTomato) and the oncogene (*PyMT*) in a tumour from a 15 week old *K14Cre;Rosa-tdTomato;MMTV-PyMT* mouse. Note that 99.9% of the tdTomato+ cells are cancer cells (PyMT+) and that all the PyMT+ cancer cells are tagged (tdTomato+). Images representative of ten tumours analysed from 5 mice. Scale bar, 25 μm. **c**, UMAPs showing the diversity of cell types in PyMT tumours (total cell number n=36091) and the clustering of the CC subset (total cell number n=19001) (see Methods). **d**, Hierarchical
590 clustering of PyMT cancer cell (CC) clusters based on the expression of cell differentiation markers (luminal or basal/myoepithelial). The x-axis represents CC clusters and y-axis the average log2 fold change in gene expression per cluster. Colour scale as in Fig. 1. Right panel, cluster representation with associated colours according to differentiation states. BM, baso-myoepithelial genes; HS, luminal hormone-sensing genes; LS, luminal stem/progenitor genes; LAS, lumino-alveolar stem/progenitor genes; LA, lumino-alveolar genes. lumino- alveolar Stem/progenitor state (LAP); pan-luminal Stem/Progenitor state (PLP); lumino-basal Stem/progenitor state (LBP). For gene names, see Supplementary Table 3. **e**, Hierarchical clustering of CC based on the expression of epithelial and mesenchymal markers. Lower panel, integration of differentiation (colour code as in c) and EMT states (grey scale).

600

Fig. 6. Two distinct EMT programmes in PyMT breast cancer. a, Left panel, Cancer cells (PyMT/tdTom+) connectivity map predicted by PAGA and represented over the UMAP embedding shown in Fig. 5c. Right panel, EMT states bifurcate into two distinct trajectories, EMT-trajectory 1 (EMT-T1) and 2 (EMT-T2). **b**, Left panel, RNA velocity analysis (see Methods). The velocities are visualized on recalculated UMAP for EMT trajectories in (a). The solid line represents smooth principal curve (see Methods) fitted over UMAP. Left lower panel, velocities of cells shown at the bifurcation point. Right lower panel, Hallmark-EMT enrichment represented over UMAP embedding. **c,** hierarchical clustering of SCENIC computed transcription factor

activities (regulons) on EMT-T1 and EMT-T2 trajectories. The regulon matrix represents the mean value of AUCell score per single-cell cluster. x-axis, cancer cells clusters. y-axis, regulons. See Supplementary Table 3 for full gene names. **d**, Binding motifs for the corresponding transcription factors and their activity (y-axis) plotted over pseudotime (x-axis) for selected examples of regulons shown in (c). The complete list of predicted regulons and their binding motifs is available in Supplementary Table 7. **e**, Expression heatmaps showing the changes in the transcriptional programmes in EMT-T1 and EMT-T2 (genes were predicted by Moran's I test and ordered over pseudotime as in Fig. 2e). **f**, Upper panel, dot plot showing GO terms associated with the two EMT trajectories, embryonic and adult-like, respectively related to development/invasion and inflammation. Lower panel, dot plot showing BC-PINGs enrichment. P- values were determined using Fisher´s Exact Test for the dot plots and BC-PINGs were P- value was determined based on the cumulative distribution function of the hypergeometric distribution. **g**, Enrichments of BC-PINGs and inflammatory score represented over UMAP embedding as in (b).

Fig. 7. The two EMT programmes are activated in segregated cancer cell populations. a, Expression of EMT-T1 markers (Krt14 and p63, upper panels) and EMT-T2 markers (Jun and Klf4, lower panels) in PyMT primary tumours. Cancer cells are identified with a PyMT antibody or by the expression of the reporter (*K14Cre;Rosa-tdTomato*). p63 is expressed in adult basal cells and can reprogram adult luminal into basal cells[33]. In the PyMT tumours, luminal cells acquire a progenitor basal-like phenotype. **b**, Non-overlapping expression of EMT-T1 marker Krt14 at tumour/stroma interface and the Jun EMT-T2 marker, enriched in more internal positions. Arrowhead (lower panel) indicates EMT-T1 cancer cells with invasive protrusions. Str: stroma. Nuclei in blue. Scale bar, 50 µm in (a) and (b). Images are representative of at least 6 primary tumours from 3 mice. **c**, PyMT tumouroids invasion assay. Primary tumours harvested from *PyMT;K14Cre;Rosa-tdTomato* mice were disaggregated into small fragments, embedded into 3D Collagen matrices (see Methods) and cultured for 3 days. Some tumouroids spontaneously invade the surrounding environment. **d**, Cells expressing Krt14, an EMT-T1 specific marker, are enriched at the invasive edges (arrows). **e**, Cells expressing high levels of the EMT-T2 specific marker Jun are enriched in central areas. Tumouroid images are representative of three independent cultures. Scale bars, 25µm. **f**, Expression of N-Cadherin (Cdh2) and Jun, EMT-T1 and EMT-T2 markers, respectively, in human TNBC. Arrowheads: cancer cells expressing either the EMT-T1 (green) or the EMT-T2 marker (magenta). Pan-Keratin (CKs) identifies cancer cells. Nuclei in blue. Scale bar, 50µm. Images are representative of 6 breast cancer sections. **g**, Enrichment in EMT-T1 and EMT-T2 clusters in human breast cancer. Gene Set Variation Analysis (GSVA) of cancer clusters from EMT-T1 and EMT-T2 (see Methods) in different breast cancer subtypes. The TNBC

subtype shows enrichment score for both EMT-T1 and EMT-T clusters. Boxes: median and IQR (interquartile range, 25th to 75th percentiles). Whiskers: highest and lowest values within 1.5 times the IQR. Outliers marked as dots. Luminal A+B group, n=3 patients; HER2, n=6 patients; TNBC, n=11 patients. P value determined using an unpaired two-sided *t*-test.

Fig. 8. Plasticity between invasive and inflammatory EMT trajectories in PyMT breast cancer. a, Design to combine conditional loss of *Snail1* and genetic tracing of PyMT cancer cells (see Methods). **b**, Analysis of primary tumour burden per mouse (mean ± SEM). CTR, n=18 and *Snail1* cKO, n=15 mice. **c**, H&E images of control (CTR) and *Snail1* cKO tumours. Tumour differentiation grade determined by mitosis rate, cellular pleomorphism and atypia (Grade 1, well; G2 moderately; G3 poorly differentiated). Quantification expressed as percentage of tumours. CTR, n=14 tumours and *Snail1* cKO, n=13 tumours from 7 mice per condition. Scale bar, 200 µm. **d-f**, Design and analysis of *Prrx1* conditional loss as shown for *Snail1* in (a-c). n=18 mice per condition (e) and CTR, n=14 tumours and *Prrx1* cKO, n=18 tumours from 7 mice per condition (f). Scale bar, 200 µm. **g**, Prrx1 and Vim co-expression in cancer and stromal cell subpopulations. Arrows: Prrx1-expressing cancer cells (red and blue) that have activated EMT (green). **h**, Prrx1 expression in hybrid epithelial/mesenchymal cancer cells identified by Epcam (blue) and Vim (green) co-expression (arrows). Str, stroma. Nuclei in blue. Scale bar, 50µm (g and h). **i**, Invasive vs total tumour areas (percentage mean ± SEM; n=5 mice per group). **j**, 3D reconstitution of whole-mounted lung lobes showing metastatic foci (tdTomato-positive) and metastatic burden quantification (n=7 mice per group). Scale bars, 1mm. **k,** qPCR fold-change expression for markers of clusters in the two EMT trajectories in Prrx1 proficient (CTR) and deficient (cKO) FACS-sorted cancer cells (n= minimum 6 tumours samples from 3 mice). **l**, Similar qPCR for EMT-T2 markers in the inflammatory breast cancer SUM149PT cell line after *SNAIL1* downregulation (n=6 biological replicates per condition). **m**, Expression of cytokines in Prrx1 cKO vs CTR tumours. **n**, Quantification of infiltrating and non-infiltrating F4/80+ cells (mean ± SEM; n=5 tumours from 3 mice per condition). Immunofluorescence in Extended Data Fig. 10h. **o**, Top, pan macrophage marker (F4/80) and Cd163+ subpopulation in CTR and *Prrx1* cKO primary tumours. In the latter, the infiltrating F4/80 macro-phages are negative for Cd163. Bottom, EMT-T2 marker Klf4 and MHC-II expression in CTR and Prrx1 cKO primary tumours. The increase in EMT-T2 is associated with an increase in MHC-II positive cells (tumour and stroma). Images representative of 5 tumours from 3 mice per condition. Nuclei in blue. Scale bar, 100 µm. **p**, EMT programmes in development, organ fibrosis and cancer. During embryonic development, the invasive EMT allows cells to disseminate and give rise to different cell types during organogenesis. In the adult, cells activate a non-invasive EMT as an inflammatory repair response to injury. This regenerative programme

can evolve towards a pro-degenerative process by promoting fibrogenesis. In cancer, both invasive and inflammatory EMTs are activated within the same tumour in distinct cell populations, with antagonic pro- and antitumour roles. In the absence of Prrx1 in cancer cells, embryonic-like EMT is truncated and the adult-like inflammatory EMT is enhanced, preventing dissemination and converting cold into hot tumours with infiltrating anti-tumour inflammatory macrophages. Red, Invasive EMT; blue, inflammatory EMT; grey, tumour bulk; white, Infiltrating macrophages. The tumour microenvironment is not shown. Created with BioRender.com under Academic License Terms with agreement number: VT24MOOYXZ. Boxes (b, e, j, k, l and n) show medians and IQRs. Whiskers: minimum and maximum values. P values determined using an unpaired two-sided Mann-Whitney U test.

Extended Data Figures legends

Extended Data Fig. 1 (related to Fig. 1). Epithelial (E) and mesenchymal (M) component analysis in breast cancer cell lines and Differential response to TGFβ in MDCK cell lines. a, E (286) genes and M (130) genes were extracted from EMT signatures described previously[12,55]. E and M genes where further used to perform k means clustering (see Methods) in the 71 breast cancer (BC) cell lines 23 (Fig. 1a). E and M genes are listed in Supplementary Table 1. **b**, Western blots showing early (1h) and sustained (4 days) SMAD2/3 activation in response to TGFβ treatment[3] in MDCK-II and MDCK-NBL2 cells. SMAD2/3 activation is measured by assessing phospho-SMAD2/3. Blots are representative of 3 biological replicates. **c**, Representative images and quantification of SMAD2/3 nuclear accumulation in MDCK-II and MDCK-NBL2 cells after four days of TGFβ treatment. MDCK-II: n= 418 cells and MDCK-NBL2: n= 222 cells each from 3 biological replicates. Boxes: medians and IQRs; whiskers: minimum and maximum values. Non-significant P>0.05. Scale bars, 20µm. **d**, Gene set enrichment analysis (GSEA) for TGF signalling (MSigDB; UC San Diego and Broad Institute) in TGFβ-treated cells (MDCK-NBL2 versus MDCK-II). NES, Normalised Enrichment Score; RES, Running Enrichment Score; RLM, Ranked List Metric. **e**, Downregulation of the epithelial marker CDH1 (E-cadherin) during TGFβ treatment. DAPI staining labels the nuclei (blue). Scale bars, 20µm. Representative images (3 independent experiments) are shown. **f**, Western-blot showing fibronectin (FN1), epithelial tight junction protein 1 (TJP1) and E-cadherin (CDH1) protein levels after four days of TFGβ treatment. **g**, CDH1 levels after 10 days of treatment. β-actin (beta-actin) is used as a housekeeping control in (f) and (g). Also in (f) and (g) blots are representative of 3 independent experiments. **h**, Gene set enrichment analysis (GSEA) for Hallmark_EMT (MSigDB; UC San Diego and Broad Institute) in TGFβ-treated cells (MDCK-NBL2 versus MDCK-II). NES, Normalised Enrichment Score; RES, Running Enrichment Score; RLM,

Ranked List Metric. **i**, Heatmap of hierarchical clustering representing the relative expression of epithelial and mesenchymal genes obtained after bulk transcriptome RNA-seq. x-axis: cell lines and treatment; y-axis: genes. The colour scale represents relative transcript levels (log2) from lowest (dark blue) to highest (bright red). In (d) and (h), P values for GSEA enrichment are estimated using an empirical phenotype-based permutation test.

Extended Data Fig. 2 (related to Fig. 1). SNAIL1 is required for the TGFβ-induced EMT response. **a**, Brightfield images of MDCK-II and MDCK-NBL2 cells treated with small interfering RNA for *SNAIL1* (si*SNAIL1*) or siCTR 24 hours before treatment with TGFβ. Images are representative of 3 independent experiments. Scale bars, 20µm. **b**, RT-qPCR showing the relative transcript levels for epithelial, mesenchymal and EMT-TFs genes after 1, 2 and 4 days of TGFβ administration in MDCK-NBL2 cells in the presence of small interfering RNA for *SNAIL1* (si*SNAIL1*, n⩾4 biological replicates for each time point) or control reagent (siCTR, n⩾4 biological replicates for each of the time points) compared to untreated cells. Data represent mean ± SEM. Consistent with the proposed pioneer role of SNAIL1, its knockdown significantly attenuates TGFβ-induced EMT including the early repression of epithelial markers, and the activation of mesenchymal markers and EMT-TFs (Fig. 1e,f). Interestingly, SNAIL1 knockdown does not prevent ZEB2 activation, which is already transcribed in the neural tube[19] before SNAIL1 induction in neural crest precursors. P values were determined using an unpaired two-sided *t*-test.

Extended Data Fig. 3 (related to Fig. 2). PRRX1 knockdown prevents TGFβ-induced invasive EMT. **a**, Analysis of epithelial (TJP1) and mesenchymal (VIM, vimentin) markers in TGFβ (1 ng/ml) treated MDCK cells cultured in Collagen matrices (as in Fig 2a). Nuclei in blue. Scale bars, 50µm. **b**, Dot plot showing selected GO terms enriched in BC-PINGs. **c**, Upper panel, relative transcript levels (RT-qPCR) for PRRX1 long (*PPRX1-L*), short (*PPRX1-S*) or both isoforms (*PRRX1*) in MDCK-NBL2 cells, either 1 or 4 days after TGFβ treatment. Note that both isoforms are activated. Boxes: medians and IQRs; whiskers: minimum and maximum values (n⩾6 biological replicates in 1d CTR and TGFβ-treated; n⩾5 biological replicates in 4d CTR and TGFβ-treated). Lower panel, *PRRX1* relative transcript levels in MDCK-NBL2 cells pre-treated with small interfering RNA for PRRX1 (si*PRRX1*) or control reagent (siCTR) followed by 4 days of TGFβ. Data represent mean ± SEM (n=6 biological replicates per condition). Asterisks indicate significant p-value in two-tailed Student's t-test. **d**, GSEA showing the reduction of Hallmark_EMT in TGFβ-treated NBL2 cells pre-treated with si*PRRX1*, and their corresponding bright- field images. NES, Normalised Enrichment Score;

RES, Running Enrichment Score; RLM, Ranked List Metric. **e**, IF images showing the expression of E-cadherin (CDH1) and α-SMA after 4 days in culture with or without TGFβ in the presence of siPRRX1 or siCTR. DAPI staining labels the nuclei in blue. Scale bars, 20µm. **f**, RT-qPCR showing the relative transcript levels for epithelial and mesenchymal genes after 4 days of TGFβ administration in MDCK-NBL2 and MDCK-II cells in the presence of small interferring RNA for PRRX1 (si*PRRX1*) and compared with those containing a control reagent (siCTR). Data represent mean ± SEM (n⩾3 biological replicates per cell line and condition). Asterisks indicate significant p-value in two-tailed Student's t-test. **g**, Hierarchical clustering analysis after bulk RNA sequencing showing the differential regulation of epithelial, mesenchymal and EMT-TFs genes after siPRRX1 treatment during TGFβ-induced EMT. The colour scale represents relative transcript levels (log2) from lowest (dark blue) to highest (bright red). **h**, Signalling pathways enriched in TGFβ-induced invasive vs non-invasive EMT in MDCK cells. **i**, GSEA showing the relative enrichment for the indicated EMT-associated signal- ling pathways for the two cell lines (upper panel) and the loss of the positive enrichment after si*PRRX1* treatment (lower panel). **j**, Brightfield images showing the Impact of pre-treatment with FAK inhibitor (FAKi) or vehicle (DMSO) in MDCK-NBL2 cells subsequently treated with TGFβ. **k**, Brightfield images showing the Impact FAKi, applied after the cells had undergone EMT. **l**, No obvious effect of FAKi can be observed in TGFβ-treated MDCK-II cells. Scale bars, 20µm in (j), (k) and (l). **m**, RT-qPCR showing the relative transcript levels for epithelial and mesenchymal genes in the MDCK-II cultures shown in (l). Data represent mean ± SEM (n=4 biological replicates in each of MDCK-II DMSO + TGFβ and FAKi + TGFβ). P values were determined using Fisher´s Exact Test for the dot plots in (b) and (h), and using an unpaired two-sided *t*-test in (c), (f) and (m). Images in (a), (e) (j), (k) and (l) are representative of 3 independent experiments.

Extended Data Fig. 4 (related to Fig. 3). Cell populations in control and obstructed kidneys revealed by single-cell RNA-seq. a, Experimental design. Three single-cell RNAseq libraries were generated from one control and two obstructed kidneys obtained from 3 mouse males. **b**, Violin plots showing gene number (detected genes), unique transcript counts and percentage of mitochondrial counts for the different 10xGenomics-based libraries. We applied filtering to remove putative cell doublets and to include only cells having number of detected genes in the range of 400-4000. 25424 cells passed this filter and were subjected to subsequent analysis, showing a mitochondrial proportion below 10% to include metabolically highly active tubular renal cells. **c**, Heatmap showing top 20 discriminative DEGs for the 26 clusters. **d**, UMAP plot showing the distribution of the 26 clusters and their assigned identities. **e**, Dot plot showing the

proportion and expression levels of markers genes that identify different cell types as in (d). Markers (y-axis), cell types (x-axis). See Supplementary Table 2 for full gene names. **f**, Cell populations changes after unilateral ureteral obstruction. Compositional data analysis (CoDA)[56] (see Methods) was used to assess the statistical relevance of changes in cell populations taking the glomerulus as a reference. Positive and negative CoDA loadings (x-axis) correspond respectively to the increases and decreases of a cell population in UUO compared to SHAM. Cell populations as in Fig. 3e. The boxplots depict the uncertainty of the loading coefficients obtained by resampling with 1000 bootstrapping. The uncertainty of the loading coefficients obtained by resampling with 1000 bootstrapping was represented using boxplot, where the boxes are IQRs split by the median (middle line) and the whiskers represent minimum and maximum loading coefficients. Corrected P values were determined using Benjamini Hochberg procedure (for more details, see Compositional Analysis for Kidney cell populations in method). The red horizontal line separates cell types passing significance threshold.

Extended Data Fig. 5 (related to Fig. 3): Origin of the injured epithelial cells in the kidney UUO model. a, 10-fold cross-validation of supervised machine learning model. Bar plots showing accuracy score and normalized Matthews correlation coefficient (normMCC = (MCC + 1) / 2)[57] obtained by n=10-fold cross validation over training dataset (see Methods for classification of injured epithelial cells using supervised machine learning model). The performance measures for each class were calculated using the One vs. Rest (OvR) method. Plots represent 10 validations, and error bars represent mean +/- SD. **b**, Expression of the injury marker Kim-1 in combination with the PT cell marker LTA in SHAM-operated and UUO 1 day after obstruction. Images are representative of kidneys obtained from 3 mice per condition. **c**, Higher power magnification showing the expression of Kim-1 and LTA in the renal cortex and medulla. Nuclei in blue. Scale bar, 50μm except for (b) where bar indicates 500μm.

Extended Data Fig. 6 (related to Fig. 4). The partial EMT programme associated with injured proximal tubules. a, Prediction of the regulatory transcriptional programme in EMT trajectory in injured renal epithelial cells. The heat map shows the hierarchical clustering of SCENIC computed transcription factor activities (regulons) for the injury trajectory. The regulon activity represents the mean value of AUCell score per single-cell cluster. See Supplementary Table 2 for full gene names. **b**, Binding motifs for the corresponding transcription factors and their activity (y-axis) plotted over pseudotime (x-axis) for selected examples of regulons shown in (b). The complete

list of predicted regulons and their binding motifs is available in Supplementary Table 6. **c**, UMAP plots showing the expression of markers of renal-specific epithelial differentiation (Myoinositol oxygenase, Miox and Hepatocyte nuclear factor 4, Hnf4), Inflammation, repair/degeneration, and mesenchymalysation. **d**, Top row, expression of the injury response marker Krt20 and the TGFβ target and dedifferentiation marker Klf4 in combination with a PT differentiation marker (LTA). Bottom row, expression of the injury response markers Kim-1 and Jun in combination with LTA. Nuclei in blue. Scale bar, 50 µm. Note the progressive loss of LTA one and two weeks after UUO and the acquisition of adult EMT in damaged proximal tubules. **e**, Time course analysis of adult EMT markers Jun and Klf4 in combination with LTA. Scale bar, 50 µm. Images in (d) and (e) are representative of those obtained from 3 mice per condition. **f**, RT-qPCR analysis of bulk kidney tissue showing relative transcript levels for EMT-TFs two weeks after UUO. Data indicate mean ± SEM (n=3 mice per condition). Asterisks indicate significant p-value in two-tailed Student's t-test. **g**, Dot plot showing the enrichment for injury response (Inflammation and pro-fibrotic GO terms) in TGFβ-treated MDCK II (non-invasive EMT) vs NBL2 (invasive EMT). P values were determined using an unpaired two-sided *t*-test (f) or Fisher´s Exact Test for the dot plot (g).

Extended Data Fig. 7 (related to Fig. 5). Analysis of cell populations in PyMT metastatic breast cancer. a, Experimental design used to prepare single-cell barcoded cDNA libraries. Four single-cell RNAseq libraries (T1-T4) were generated from n=4 independent samples obtained from three 12-14 weeks old female mice. Right panel shows the 3D reconstitution of one representative whole-mounted left lung lobe showing tdTomato-positive metastatic foci. Scale bars, 2mm. **b**, Violin plots showing gene number (detected genes), unique transcript counts and percentage of mitochondrial counts for the different 10xGenomics-based libraries of the four PyMT primary tumour samples. We removed putative cell doublets and applied stringent filtering to include only cells having number of detected genes in the range of 400-4000. The majority of cells (n=36091/36162) passed this filter and were subjected to subsequent analysis, showing a mitochondrial proportion below 2%, indicative of high-quality104. **c**, Heatmap showing discriminative genes of the five main PyMT tumour populations (see Fig. 4c). **d**, Dot-plot showing the expression levels for genes that identify the major cell types in the tumours. Symbols of cell types (y-axis) as shown in (c). **e**, UMAP visualization of cells expressing different markers for tumour cells (CC, tdTomato), myeloid cells (MC, Cd74), cancer-associated fibroblasts (CAF, Col3a1), endothelial cells (EC, Cdh5) and lymphoid cells (LC, Cd3g). See Supplementary Table 2 for all full gene names.

Extended Data Fig. 8 (related to Fig. 5). PyMT cancer cell cluster analysis. a, UMAP visualization of cancer cells (n=19001) showing expression of the tdTomato reporter. **b**, UMAP plots and table depicting the distribution of cancer cell subclusters in each single-cell RNAseq data set derived from the 4 independent tumour samples. **c**, Expression of luminal (blue) and basal/myoepithelial (red) cancer cell lineage markers on the UMAP gene expression plot. **d-f**, Distribution of expression of epithelial (d), mesenchymal markers (e), and of EMT-TFs (f) in cancer cells on the UMAP plot. Markov affinity-based graph imputation of cells (MAGIC)[58] was applied to improve EMT-Tfs representation over UMAP. See Supplementary Table 2 for full gene names.

Extended Data Fig. 9 (related to Fig. 6). Molecular characterisation of the two EMT trajectories in PyMT breast cancer. a, Expression heatmap extending the analysis shown in Fig. 5a to discriminate between EMT-T1 and EMT-T2 transcriptional programmes. Cells in the two EMT trajectories follow completely different paths associated with phenotypic transitions and EMT-TF expression codes. In EMT-T1, cancer cells progressively lose lumino-alveolar differentiation genes like *Csn3, Lalba, Wap* and evolve towards a stem/progenitor-like state (cluster 10), including expression of *Aldh1a3*, pro-stemness genes such as *Ndrg1*[59]. Cluster 14, compatible with a partial EMT status, and loosing epithelial genes such as *Epcam, Cldn3* and *Cldn7*, is followed in the pseudo time analysis by clusters 12 and 16. Cluster 14 contains pluripotency markers such as *Wnt9a, Bmp1, Id1, Id3,* and *Igf1* plus mammary gland embryonic and basal-like signatures while progressing towards a full EMT state exemplified by high expression of *Vim* and *Cdh2* (Cluster 16). An invasion program is already evident in cluster 14, with cells expressing genes that regulate cell migration and cytoskeleton remodelling (*Tnc, Gsn, Palld, Cnn2, Tpm1, Tpm2* and *Mmp14*). The invasion signature is amplified in clusters 12 and 16, with prominent expression of additional invasion genes including cytoskeleton regulators (*Mylk, Tagln* and *Pdpn*), guidance receptors (*Sema5a* and *Nrp2*) and microenvironmental modulators like metalloproteinases and Lysyl oxidases (*Mmp2, Mmp3* and *LoxL1*) in the latter. Initiation of the Hallmark_EMT signature in cluster 14 concurs with the detection of *Snail2* in addition to *Snail1* and *Twist1* in cluster 12. The progression towards more advanced EMT state is coupled to an increase in *Zeb1* and *Prrx1*. In EMT-T2, the lumino-alveolar epithelial phenotype of cluster 11 progresses to the partial EMT phenotype of cells in clusters 1, 13 and 15, still maintaining expression of epithelial genes while activating some mesenchymal genes shared with the EMT-T1 trajectory (e.g. *Sparc, Postn, S100a4*), but without progressing to full EMT. EMT2 has a remarkable enrichment in injury response genes (e.g. *Egr1, Jun, Junb, Fos, Fosb,* and *Lcn2*) and inflammatory regulators, including

secreted factors (*Spp1*), components of the TNF-α/interferon and NF-kB pathways (e.g. *Nfkbia, Ccrl2* and *Notch2*) and inflammatory biomarkers such as serum amyloid A proteins (*Saa2, Saa1*) or Lymphocyte antigen-6 family genes (*Ly6k* and *Ly6d*). Additional enrichment for pro-inflammatory genes are seen in cluster 2, the most prominent cluster in this branch, including additional interferon regulators and downstream targets genes (e.g. *Irf7, Ifitm3, Ifitm2,* and *Cxcl16*). In addition, EMT-T2 is enriched for pro-fibrotic genes as the tissue inhibitor of metalloproteinases (*Timp2, Timp3* and *Timp1*). All of this indicates that, in EMT-T2, the transition to a partial EMT is concomitant with the acquisition of an inflammatory and pro-fibrotic phenotype. In contrast to EMT-T1, among EMT-TFs, only Snail1 is detected in clusters 1, 13 and 15. Abbreviations as in Figs. 4 and 5. See Supplementary Table 2 for full gene names. **b**, Dot plot showing an extended version of the GO terms enriched in different cancer cell clusters and across EMT trajectories shown in Fig. 5b. Interestingly, common pathways are associated with the activation of an EMT programme including regulation of cell cycle and resistance to cell death[60]. Developmental pathways are associated with EMT-T1 and those related to inflammation are enriched in EMT-T2. **c**, Dot plot showing the enrichment in the BC-PINGs signature and in genes upregulated in TGFβ-treated invasive MDCK-NBL2 and non-invasive MDCK-II in cancer cells. Clusters represented as in (b). P values were determined using Fisher´s Exact Test for the dot plot (b) or based on the cumulative distribution function of the hypergeometric distribution (**c**).

Extended Data Fig. 10 (related to Fig. 8). Generation and characterisation of Prrx1 conditional mutant mice. a, Summary of the strategy used to generate an exon 2 double-floxed Prrx1 allele, Prrx1em1An. Mouse embryonic stem cells (mESC) where edited using the CRISPR/Cas9 system to endogenous Prrx1 exon2 by an exon 2 double-floxed Prrx1 cassette flanked by left and right homology arms. Green arrows show the position of the primers used to screen for the recombined alleles. Black arrows show the position of primers used for genotyping (see also Methods). **b**, Validation of the Prrx1 cKO model. Prrx1em1An/em1An zygotes were treated with TAT-CRE and at 2-cell stage implanted into pseudopregnant females. **c**, Implanted embryos were collected at E13.5, and used to generate mouse embryonic fibroblasts (MEFs) from untreated or previously TAT-CRE treated zygotes. tdTomato expression in over 90% of MEFs, indicated the high efficiency of TAT-CRE-mediated recombination (not shown). WB showing the loss of Prrx1 protein in MEFs derived from indicated genotypes. Note that TAT-CRE treated Prrx1flox/flox embryos (well number 3) have the same profile as Prrx1 homozygous mutant embryos (well number 5), detecting the Prrx2 protein. This confirms the efficacy of our strategy. The WB represents MEFs from one of three independent validation experiments. **d**, Images of the palate in newborn (P0) from untreated and TAT-CRE treated PRRX1flox/flox zygotes. Animals derived from TAT- CRE

treated zygotes show high td-Tomato recombination and a fully penetrant cleft palate phenotype, as reported in PRRX1 null mutant mice35. TAT-CRE treated zygotes, n=7; untreated, n=3. LPS: lateral palatal shelf; NS: nasal septum. **e**, Images of invasive areas (surrounded by the dashed lines) of the primary PyMT tumours defined by Pan-Laminin low/K14 positive cancer cells. At least 3 mice were analysed per genotype. Nuclei in blue. Scale bar, 100 µm. **f**, Expression of EMT-T2 specific marker Klf4 in PyMT primary tumour (n=3 mice analysed per genotype). Nuclei in blue. Scale bar, 100 µm. **g**, Upper panel, Venn diagram showing the genes enriched in TGFβ-treated MDCK-II (non-invasive EMT) vs NBL2 (invasive EMT) cells compared to those negatively regulated by PRRX1 in the latter, and the overlap (180 genes). Lower panel, dot plot showing that within the group of 180 genes, there is enrichment for those associated with inflammation and immune regulation in KEGG and GO gene datasets. Our data are compatible with Prrx1 preventing their activation in invasive cells. **h**, Expression of the pan macrophage (F4/80) marker in CTR and Prrx1 cKO primary tumours. Images are representative of 5 tumours from 3 mice per condition. **i**, Left panels, single channels corresponding to pictures shown in Fig. 8o showing the expression of the pan macrophage marker (F4/80) and of the protumour anti-inflammatory Cd163 positive subpopulation in CTR and Prrx1 cKO primary tumours. In the Prrx1 cKO tumours, the infiltrating F4/80 macrophages are negative for Cd163. Right panels, single channels corresponding to Fig. 8o showing the expression of the EMT-T2 marker Klf4 and of MHC-II in CTR and Prrx1 cKO primary tumours. Note that in the latter the increase in EMT-T2 is associated with an increase in antitumour inflammatory MHC-II positive cells in the stroma and, importantly, in the core of the tumour. n=5 tumours from 3 mice per condition. Nuclei in blue. Scale bar, 100 µm (h) and (i). P values were determined using Fisher´s Exact Test for the dot plot (g).

References

1. Nieto, M. A., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* **166**, 21–45 (2016).
2. Dongre, A. & Weinberg, R. A. New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. *Nat. Rev. Mol. Cell Biol.* **20**, 69–84 (2019).
3. Massagué, J. & Sheppard, D. TGF-β signaling in health and disease. *Cell* **186**, 4007–4037 (2023).
4. Youssef, K. K. & Nieto, M. A. Epithelial-mesenchymal transition in tissue repair and degeneration. *Nat. Rev. Mol. Cell Biol.* (2024) doi:10.1038/s41580-024-00733-z.
5. Yang, J. *et al.* Guidelines and definitions for research on epithelial-mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **21**, 341–352 (2020).
6. Grande, M. T. *et al.* Snail1-induced partial epithelial-to-mesenchymal transition drives renal fibrosis in mice and can be targeted to reverse established disease. *Nat. Med.* **21**, 989–997 (2015).
7. Lovisa, S. *et al.* Epithelial-to-mesenchymal transition induces cell cycle arrest and parenchymal damage in renal fibrosis. *Nat. Med.* **21**, 998–1009 (2015).

8. Pastushenko, I. *et al.* Identification of the tumour transition states occurring during EMT. *Nature* **556**, 463–468 (2018).
9. Kröger, C. *et al.* Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 7353–7362 (2019).
10. Simeonov, K. P. *et al.* Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. *Cancer Cell* **39**, 1150-1162.e9 (2021).
11. Nieto, M. A. Are You Interested or Afraid of Working on EMT? *Methods Mol. Biol. Clifton NJ* **2179**, 19–28 (2021).
12. Tan, T. Z. *et al.* Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.* **6**, 1279–1293 (2014).
13. Klijn, C. *et al.* A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.* **33**, 306–312 (2015).
14. Thiery, J. P., Acloque, H., Huang, R. Y. J. & Nieto, M. A. Epithelial-mesenchymal transitions in development and disease. *Cell* **139**, 871–890 (2009).
15. Zhang, J. *et al.* Pathway crosstalk enables cells to interpret TGF-β duration. *NPJ Syst. Biol. Appl.* **4**, 18 (2018).
16. Neve, R. M. *et al.* A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**, 515–527 (2006).
17. Sarrió, D. *et al.* Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Res.* **68**, 989–997 (2008).
18. Soldatov, R. *et al.* Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* **364**, eaas9536 (2019).
19. Hegarty, S. V., Sullivan, A. M. & O'Keeffe, G. W. Zeb2: A multifunctional regulator of nervous system development. *Prog. Neurobiol.* **132**, 81–95 (2015).
20. Vandamme, N. *et al.* The EMT Transcription Factor ZEB2 Promotes Proliferation of Primary and Metastatic Melanoma While Suppressing an Invasive, Mesenchymal-Like Phenotype. *Cancer Res.* **80**, 2983–2995 (2020).
21. Martin, J. F., Bradley, A. & Olson, E. N. The paired-like homeo box gene MHox is required for early events of skeletogenesis in multiple lineages. *Genes Dev.* **9**, 1237–1249 (1995).
22. Chevalier, R. L. 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 (2016).
23. Kuppe, C. *et al.* Decoding myofibroblast origins in human kidney fibrosis. *Nature* **589**, 281–286 (2021).
24. Dumas, S. J. *et al.* Single-Cell RNA Sequencing Reveals Renal Endothelium Heterogeneity and Metabolic Adaptation to Water Deprivation. *J. Am. Soc. Nephrol. JASN* **31**, 118–138 (2020).
25. Ransick, A. *et al.* Single-Cell Profiling Reveals Sex, Lineage, and Regional Diversity in the Mouse Kidney. *Dev. Cell* **51**, 399-413.e7 (2019).
26. Conway, B. R. *et al.* Kidney Single-Cell Atlas Reveals Myeloid Heterogeneity in Progression and Regression of Kidney Disease. *J. Am. Soc. Nephrol. JASN* **31**, 2833–2854 (2020).
27. Wu, H., Lai, C.-F., Chang-Panesso, M. & Humphreys, B. D. 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 (2020).
28. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
29. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
30. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
31. Guy, C. T., Cardiff, R. D. & Muller, W. J. Induction of mammary tumors by expression of polyomavirus middle T oncogene: a transgenic mouse model for metastatic disease. *Mol. Cell.*

Biol. **12**, 954–961 (1992).
32. Attalla, S., Taifour, T., Bui, T. & Muller, W. Insights from transgenic mouse models of PyMT-induced breast cancer: recapitulating human breast cancer progression in vivo. *Oncogene* **40**, 475–491 (2021).
33. Wuidart, A. *et al.* Early lineage segregation of multipotent embryonic mammary gland progenitors. *Nat. Cell Biol.* **20**, 666–676 (2018).
34. Bach, K. *et al.* Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat. Commun.* **8**, 2128 (2017).
35. Pal, B. *et al.* Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. *Nat. Commun.* **8**, 1627 (2017).
36. Shehata, M. *et al.* Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. *Breast Cancer Res. BCR* **14**, R134 (2012).
37. Ginestier, C. *et al.* ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. *Cell Stem Cell* **1**, 555–567 (2007).
38. Koren, S. *et al.* PIK3CA(H1047R) induces multipotency and multi-lineage mammary tumours. *Nature* **525**, 114–118 (2015).
39. Van Keymeulen, A. *et al.* Reactivation of multipotency by oncogenic PIK3CA induces breast tumour heterogeneity. *Nature* **525**, 119–123 (2015).
40. Youssef, K. K. *et al.* Adult interfollicular tumour-initiating cells are reprogrammed into an embryonic hair follicle progenitor-like fate during basal cell carcinoma initiation. *Nat. Cell Biol.* **14**, 1282–1294 (2012).
41. Kaufman, C. K. *et al.* A zebrafish melanoma model reveals emergence of neural crest identity during melanoma initiation. *Science* **351**, aad2197 (2016).
42. Cheung, K. J., Gabrielson, E., Werb, Z. & Ewald, A. J. Collective invasion in breast cancer requires a conserved basal epithelial program. *Cell* **155**, 1639–1651 (2013).
43. Rädler, P. D. *et al.* Highly metastatic claudin-low mammary cancers can originate from luminal epithelial cells. *Nat. Commun.* **12**, 3742 (2021).
44. Ye, X. *et al.* Distinct EMT programs control normal mammary stem cells and tumour-initiating cells. *Nature* **525**, 256 (2015).
45. Bièche, I. *et al.* Molecular Profiling of Inflammatory Breast Cancer: Identification of a Poor-Prognosis Gene Expression Signature. *Clin. Cancer Res.* **10**, 6789–6795 (2004).
46. Ge, Y. *et al.* Stem Cell Lineage Infidelity Drives Wound Repair and Cancer. *Cell* **169**, 636-650.e14 (2017).
47. Marjanovic, N. D. *et al.* Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. *Cancer Cell* **38**, 229-246.e13 (2020).
48. van Groningen, T. *et al.* Neuroblastoma is composed of two super-enhancer-associated differentiation states. *Nat. Genet.* **49**, 1261–1266 (2017).
49. González-Iglesias, A. & Nieto, M. A. Proliferation and EMT trigger heart repair. *Nat. Cell Biol.* **22**, 1291–1292 (2020).
50. Ocaña, O. H. *et al.* Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer Prrx1. *Cancer Cell* **22**, 709–724 (2012).
51. Tsai, J. H., Donaher, J. L., Murphy, D. A., Chau, S. & Yang, J. Spatiotemporal regulation of epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. *Cancer Cell* **22**, 725–736 (2012).
52. Liu, X. *et al.* Sequential introduction of reprogramming factors reveals a time-sensitive requirement for individual factors and a sequential EMT-MET mechanism for optimal reprogramming. *Nat. Cell Biol.* **15**, 829–838 (2013).
53. Karras, P. *et al.* A cellular hierarchy in melanoma uncouples growth and metastasis. *Nature* **610**, 190–198 (2022).
54. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
55. Taube, J. H. *et al.* Core epithelial-to-mesenchymal transition interactome gene-

expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 15449–15454 (2010).
56. Petukhov, V. *et al.* Case-control analysis of single-cell RNA-seq studies. 2022.03.15.484475 Preprint at https://doi.org/10.1101/2022.03.15.484475 (2022).
57. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
58. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729.e27 (2018).
59. Wang, Y. *et al.* N-myc downstream regulated gene 1(NDRG1) promotes the stem-like properties of lung cancer cells through stabilized c-Myc. *Cancer Lett.* **401**, 53–62 (2017).
60. Vega, S. *et al.* Snail blocks the cell cycle and confers resistance to cell death. *Genes Dev.* **18**, 1131–1143 (2004).

Methods

EMT analysis in breast cancer patients

EMT gene expression signatures: Enrichment of gene expression signatures found in each cluster of EMT-T1 and EMT-T2 trajectories was computed using GSVA (v.1.34.0) R/Bioconductor package[61] to perform Gene Set Variation Analysis in breast cancer expression data obtained from Chung et al. 2017 (GEO GSE75688)[62]. R packages dplyr (v.1.0.3), magrittr (v.2.0.1) and tibble (v.3.1.2) were used to transform gene expression data to the required GSVA input format, and ggpubr (v.0.4.0) was used to generate the GSVA enrichment score box-plots.

Human breast cancer tumours multiplex immunofluorescence: Triple immunofluorescence was performed on 2µm tumour sections from human TNBC (triple negative breast carcinoma) samples using BOND RX Fully Automated Research Stainer using Opal TM 7-color Automation IHC Kit (Akoya Biosciences). Opal-650, Opal-520, and Opal-570 were used to detect C-JUN, CYTOKERATINS AE1/AE3 (Pan-Cytokeratin) and N-CADHERIN antibodies, respectively. Slides were mounted with Prolong Diamond (Molecular Probes) and imaged using the Thunder imaging system (Leica). Samples were acquired from the Biobank of the Anatomy Pathology Department (record number B.0000745, ISCIII National Biobank network) of the MD Anderson Cancer Center Madrid, Spain. This study was performed following standard ethical procedures of the Spanish regulation (Ley de Investigación Orgánica Biomédica, 14 July 2007) and was approved by the ethic committees of the MD Anderson Cancer Center Madrid, Madrid, Spain.

Animal experiments

Mice were fed *ad libitum*. Housing and experimental procedures were conducted in strict compliance with the European Community Council Directive (89/609/EEC) and the Spanish

legislation. Ethical protocols were approved by the CSIC Ethical Committee and the Animal Welfare Committee of the Institute of Neurosciences. Animals for experiments were selected by genotype, and no randomization or blinding was performed. Animals were housed under SPF conditions at the RMG animal House (ES-119-002001 SEARMG).

Kidney fibrosis model: To genetically label renal tubular (RT) epithelial cells, we generated a mouse line with the Rosa-LSL-tdTomato reporter line Ai9 /RCL-tdT)[63] (kindly provided by Oscar Marin, King´s College London), activated in RT cells by a Cre recombinase under the control of the kidney-specific promoter *Ksp1.3*[64] (kindly provided by Peter Igarashi, University of Minnesota). To inactivate Snail1, we crossed Snail1$^{fl/fl}$ mice[6] with the strain bearing the Ksp1.3-Cre transgene mentioned above. Mice were maintained in C57BL/6 background. UUO procedure: Male and female mice (8-12 weeks-old) were subjected to UUO following surgery protocol in[6]. Unilateral ureteral obstruction (UUO) was maintained for 1, 2 or 3 weeks.

Breast cancer model: Mouse experiments were carried out in MMTV-PYMT model[31] crossed with a Rosa-LSL-tdTomato reporter line[63], purchased from JAX MICE (The Jackson Laboratory), expressing tdTomato upon Cre-mediated recombination. Cre recombinase is expressed under the control of *Keratin14* promoter {Tg(KRT14-cre)1Amc/J-STOCK 004782}[65]. Mice were backcrossed in FVB background for at least 10 generations (99.9% FVB).

Generation of Snail1 or Prrx1 conditional mutant tumours.

To specifically inactivate Snail1 in breast cancer cells, we generated a mouse line crossing the above-described line with a Snail1$^{f/f}$ line[6] (cKO). To specifically inactivate Prrx1 we used a similar strategy crossing the mice with a newly generated *Prrx1* conditional mutant mouse line described in the next section.

Generation of Prrx1 conditional mutant mice, Prrx1^{em1An}: Mouse embryonic stem cells (mESC) where edited using the CRISPR/Cas9 system to replace the endogenous *Prrx1* exon2 by an exon 2 double-floxed *Prrx1* cassette flanked by homology arms. We electroporated mouse embryonic stem cells (mESC) with a mix of (i) PX458 plasmid[66] (Addgene #48138) to drive the expression of the SpCas9 protein together with GFP and a guide RNA (gRNA) targeting CTGTGCTTCTTTGGGTAGAA(TGG) sequence downstream of *Prrx1* Exon-2; (ii) A linearised double stranded donor cassette containing double-floxed *Prrx1* exon-2 flanked by homology arms engineered to replace the endogenous *Prrx1* exon-2 after homologous recombination. Successfully electroplated mESC were selected assessing GFP expression and cells were expanded in culture until further recombination screening by conventional PCR and sequencing. mESC with the correct recombination of the double-floxed *Prrx1* allele were used to generate chimeric mice following conventional protocols. Chimeras with high ES contribution were backcrossed in FVB/N and C57

backgrounds to generate stable mouse colonies carrying the *Prrx1*^{em1An} allele, which were fully viable and fertile in both genetic backgrounds.

Kidney, mammary tumour and lung samples

Kidney, tumour and whole lung samples where fixed in PFA 4% o/n at RT. Prefixed kidney and tumour samples were embedded in paraffin or O.C.T ™ (Sakura) for further sectioning and collection on SuperFrost plus microscope slides.

Cell culture

2D cell culture: MDCK-NBL2 and MDCK-II cell lines were purchased from ATCC (American Type Culture Collection) and Sigma (European Collection of Authenticated Cell Culture), respectively. SUM149PT cells were purchased from Asterand. MDCK-NBL2 and MDCK-II cells were cultured in DMEM (Sigma) supplemented with 10% heat inactivated foetal bovine serum (FBS) (Sigma), 1% Gentamicin (Sigma) and 1% Amphotericin (Sigma). SUM149PT cells were cultured in Nutrient mixture Hams F12 supplemented with 5% inactivated FBS, HEPES (10mM), Insulin (5μg/ml), Hydrocortisone (1μg/ml) and antibiotics. Cells were grown at 37ºC and 5% CO_2, and the medium was replaced every two or three days. Cells were passed up to a maximum of 8 times.

3D cell culture: Collagen gel containing Bovine Collagen Solution type I (Gibco) (2.5%), Glutamax (1X), MEM (1X), $NaHCO_3$ (0.23%) and HEPES (0.1M) was prepared on ice at pH 7.0-7.5. Glass coverslips were deposited at the bottom of 24-well culture plates, covered with 100 μl of Collagen gel and cultured at 37ºC without CO_2 for 30min. After solidification, 100μl of Collagen containing 5-10x10^3 MDCK cells were added on top and incubated for 30 minutes. These 3D cultures were incubated at 37ºC and 5% CO_2. 200μl of MDCK media were added gently after 2h and replaced every 48 or 72h.

TGFβ administration, RNA interference experiments and treatment with inhibitors:

A stock solution of human recombinant TGFβ (rH-TGFβ1) (MERQ) (SHENANDOAH) was prepared at 2μg/ml. All treatments in 2D cultures (5 ng/ml) started 24h after seeding cells (10^4 in 6 well-plates or 75x10^4 in 10cm culture dishes) and the medium containing TGFβ was replaced every 48h. Cells were never seeded from high confluency cultures to avoid a reduction in the response to TGFβ. TGFβ administration (high dose 5ng/ml, low dose 0.3ng/ml) in 3D collagen cultures started after the formation of polarized MDCK spherical cysts.

For RNA interference experiments, siRNA was transfected using Lipofectamine RNAiMax (ThermoFisher Scientific) following the manufacturer's protocol. TGFβ was administered 8h after transfection and refreshed every 48h.

si*Prrx1* duplex oligonucleotides (Sigma) were prepared at 20μM. *Prrx1* siRNA (*cfa-si-PRRX1-I* sense: GAGCGCGUCUUUGAGAGAACACACU(dT)(dT)) was used at a final concentration of 10nM. BLOCK-iT™ fluorescent Oligo (20 μM) was used as RNAi control.

Hs-si-*SNAIl1* oligonucleotides were purchased as *Dicer-substrate siRNA (DsiRNAs)* duplex oligos (2nmol) directed against *SNAIL1, which* were resuspended in nuclease-free water to a final concentration of 100μM. Working stocks were prepared using the buffer provided. *Best SNAIL1 downregulation was obtained with a combination of DsiSNAIl1.13.1 and DsiSNAIl1.13.2 (*final concentration of 2,5M each) *72h after transfection. DS NC1 oligonucleotide was used as a negative control.*

For focal adhesion signalling inhibition, stock solutions of FAK Inhibitor 14 (Sigma) were prepared in DMSO at 10mM and further diluted in culture media to a final concentration of 0.2μM.

Primary tumour derived tumouroids and invasion assay: Primary tumour tumouroids were prepared and embedded in 3D collagen gels following a protocol modified from[42]. In summary, mammary gland carcinomas were collected from 14 weeks old female mice and first minced into tumour fragments and embedded in Collagen gel containing Rat Collagen Solution type I (Corning) (2.5%) in DMEM (1X), $NaHCO_3$ (0.23%) and HEPES (0.1M) prepared on ice at pH 7.0-7.5. A volume of 100μl of tumour fragments and collagen mixture was added on top of previously solidified cell-free Collagen gel plated in 48-well plates and incubated at 37ºC and 5% CO_2.

The tumouroids were washed twice with PBS, and then fixed with PFA for 60 min at RT. Fixed organoids were washed 3X 30 min in PBS and blocked/permeabilized for 4h at RT with IF blocking buffer (IFBB+: 5% normal Goat Serum, 1% Bovine Serum Albumin, 1% TritonX-100 and 0.1% Sodium Azide in sterile PBS). Blocking solution was substituted by the primary antibody diluted in IFBB+ and incubated o/n at RT on a rocker plate. Tumouroids were washed 3x for 30 min in PBST (PBS with 1% TritonX-100) and incubated for 24 h with secondary antibodies and DAPI. Finally, tumouroids were washed 3x for 60 min in PBST and mounted on Glass Bottom Microwell Dishes (MatTek) using anti-fade mounting medium (DAKO). Primary and secondary antibodies used are shown in Extended Data Table 4. Tumouroids were photographed using Leica SPEII confocal and acquired images were analyzed with ImageJ and Adobe Photoshop CS6 software programs.

Immunofluorescence (IF)

Cells in culture: MDCK cells were grown on coverslips in 6 well-plates, under the culture and treatment conditions described above. Cells were rapidly washed twice with PBS and fixed with PFA for 15min at RT, and rinsed with PBS for at least three rounds of 10 min each and directly used for IF or stored at 4ºC in PBS+Azide 0.02% for less than one week. For IF staining, coverslips were

deposited in a humidified chamber and blocked/permeabilized for 1h with IF blocking buffer (IFBB: 5% normal Goat Serum, 1% Bovine Serum Albumin and 0.2% TritonX-100 in sterile PBS). Blocking solution was substituted by the primary antibody diluted in cold IFBB and the staining chamber incubated o/n at 4ºC. Coverslips were washed 3x for 10 min in PBS and incubated for 1h with secondary antibodies and DAPI. Finally, coverslips were washed 3x for 10 min in PBS and mounted on glass slides using anti-fade mounting medium (DAKO). Primary and secondary antibodies used are shown in Extended Data Table 4. Cells were photographed using Leica SPEII confocal, Leica DMR or Zeiss Axio microscopes. Acquired images were analysed with ImageJ and Adobe Photoshop CS6 software programs.

Kidney and tumour samples: Paraffin-embedded sections were dewaxed, and protein epitopes unmasked by immersion in 95ºC preheated Citrate (pH6.0) or Tris-EDTA (pH9.0) buffer for 20 min. O.C.T or unmasked paraffin sections were washed three times in PBS for 5 min and subjected to the IF protocol described above for cell lines. For information on primary and secondary antibodies see Extended Data Table 4. Pictures were acquired and analysed as described for cells in culture.

Lungs: tdTomato+ metastasis were visualized by IF on whole lungs and cleared following the iDISCO+ protocol[67]. Images were taken using an UltraMicroscope II (LaVision BioTec). The acquired images were analysed, and 3D reconstruction was made using Vision4D (Arivis) Image Analysis Software. For the analysis of metastatic burden, 3D reconstruction was performed using Imaris software (version 9.3.1; BitPlane). "Surface" function was used for segmentation (tdTomato signal) and volumetric data extracted for the identified metastatatic objects. To avoid false-positives due to occasional secondary antibody trapping or non-specific auto-fluorescence, a tdTomato-negative lung lobule was analysed and a detection cut-off of 90.000 μm^3 was identified as minimum volume for object identification with high confidence.

Western blot

Cells were washed twice with ice-cold PBS and lysed in freshly prepared cold RIPA buffer supplemented with a protease inhibitor cocktail (Complete Mini, Roche). When necessary, cells were passed through a 25G syringe to help homogenization. Total protein extracts were quantified using BCA assay (ThermoFisher Scientific) and quality checked by Coomassie assay. Before electrophoresis, protein lysates were denatured by boiling with 6X Laemmli loading buffer at 99ºC for 10min. After electrophoresis, proteins were transferred to PVDF membranes, which were blocked for 1h at RT in 5% not-fat milk in TBST and incubated o/n at 4ºC with blocking solution containing the primary antibody. Membranes were washed 5-6 times in TBST and incubated for 45min with secondary antibody. After washing, staining was revealed with chemiluminescent

reagents (Millipore) and captured in Amersham™ Imager 680 equipment (GE Healthcare). For further information on primary and secondary antibodies see Extended Data Table 4.

Transwell cell migration assay

MDCK-II cells (non-invasive) were transfected with a plasmid carrying the coding region of the human PRRX1-L isoform. MDCK-II and MDCK-NBL2 cells transfected with an empty vector plasmid were used as negative and positive controls, respectively. Two days after transfection, cells were treated with TGFβ (5ng/ml), collected after 24h and assessed for migratory capacity using a Boyden Chamber assay. The upper chamber insert (Corning Costar Transwell) was covered with 50μl of Mouse Collagen IV (Corning, 50μg/ml) and left to dry overnight. The resulting matrix was hydrated with 25μl of H_2O before seeding the cells. $25x10^4$ MDCK cells were seeded and allowed to migrate in presence of TGFβ (5ng/ml). Nuclei of cells at the bottom of the inset were imaged 24h after seeding previous DAPI staining and automatically counted using ImageJ.

Cytokine analysis

Cytokine analysis was performed on whole tumour lysates using the proteome profiler mouse XL cytokine array (#ARY028, R&D Systems), following the manufacturer's instructions. A total of 200mg protein lysates were used per assay. Arrays membranes were imaged in an Amersham™ Imager 680 (GE Healthcare) and relative protein levels were calculated for cytokines spots using Matlab Protein Array Tool version 2.0.0.1, MATLAB Central File Exchange, Danny Allen (2022). See https://www.mathworks.com/matlabcentral/fileexchange/35128-protein-array-tool).

Total RNA extraction, cDNA synthesis and RT-qPCR

For gene expression analysis, RNA was extracted using illustra RNAspin Mini (GE Healthcare) or mirVanaTM miRNA (Ambion) Isolation Kits. Retrotranscription was performed using Maxima First Strand cDNA Synthesis kit (ThermoFisher Scientific). RT-qPCR was performed using Fast SYBR Green Mastermix in a Step One Plus machine (Applied Biosystems) according to the manufacturers' instructions. Relative RNA expression levels (relative fold change) were calculated using 2^{-ddCt} formula. Quantitative RT-qPCR primers are listed in Extended Data Table 5.

In silico analysis of human cancer cell lines

Breast Cancer cell lines gene expression data[13] were analyzed for epithelial (E) and mesenchymal (M) component enrichment. E and M component were obtained from merging E and M signatures in Refs. [12] and [55]. Note that the EMT-TFs were removed from the Mesenchymal signature to avoid

biased correlations in subsequent analyses. Singscore[68] was used to compute enrichment scores for E and M components (https://github.com/DavisLaboratory/singscore). E and M enrichment values were plotted (X-axis: M score and Y-axis: E score) and K-means clustering was used to partition the Breast Cancer cell lines According to the optimal number of clusters calculation, K=3.

Bulk RNA sequencing and data analysis

Sequencing: RNA was extracted using illustra RNAspin Mini isolation kit from three biological replicates per condition. RNA quality check, mRNA libraries preparation (stranded) and paired-end reads (75 pb length) sequencing using Illumina HiSeq4000 platform were performed at the CNAG-CRG facility in Barcelona, Spain.

Data analysis: Reads were aligned to CanFam3.1 genome annotation (Ensembl v97) using STAR (2.5.3a)[69]. Quality control of sequenced reads was performed using FastQC (Babraham Institute) and gene expression was quantified using RSEM (1.3.0)[70].

Functional enrichment analysis: We used the enrichR R package (v.2.1) to access the Enrichr database[71] (and performed general functional enrichment analysis, while the gseGO and gseKEGG functions in clusterProfiler R package (v.3.10.0)[72] were used for Gene Set Enrichment Analysis (GSEA) of Gene Ontology terms and Kyoto Encyclopedia of Genes and Genomes pathways. The R package msigdbr (7.0.1) (https://CRAN.R-project.org/package=msigdbr) was used to obtain gene sets from the Molecular Signatures Database (MSigDB) v7.0[73] from Broad Institute. The R package GOSemSim (v.2.8.0)[74] was used to filter GO terms by semantic similarity, and ggplot2 (v.3.3.0) (https://cran.r-project.org/web/packages/ggplot2/index.html), and enrichplot (v.1.6.1) (https://github.com/GuangchuangYu/enrichplot) were used to visualize functional enrichment results.

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). Protocols for dissociation and single-cell GEM preparation using 10X genomic kits and platform are available at Protocol.io:

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

Single-cell data analyses

The detailed version of this section is deposited in Protocol.io:

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

Quality control, sample integration, dimensionality reduction and clustering: The reads were aligned to the mouse genome (mm10) and gene counting was performed using the CellRanger pipeline[75] (10X Genomics). Low-quality cells were identified based on percentage of mitochondrial genes (kidney<10%; cancer< 5%), detected genes (400-4000), and putative doublets using Scrublet (https://github.com/swolock/scrublet). For integration, we used the SCTransform workflow from Seurat[76] with top 3000 highly variable genes (HVGs). A Shared Nearest Neighbor (SNN) graph and UMAP (Uniform Manifold Approximation and Projection) was built over the top PCs (kidney=25; cancer=30) and clusters were detected with resolution 0.65 and 0.03 for kidney and cancer, respectively. FindAllMarkers was used with logistic regression method to detect the differentially expressed genes.

Compositional Analysis for Kidney cell populations: To investigate cell compositional changes in SHAM and UUO, we used the runCoda function from Cacoa[56]. The compositional analysis was performed with 1000 bootstraps and glomerulus cluster was set as a reference.

Classification of injured-epithelial cells in kidney scRNA-Seq: We used a deep learning multi-class classification approach to predict the origin of injured epithelial cells. The cells from epithelial clusters (see Fig. 3e and Extended Data Fig. 4d) were subset with SCT-normalized gene expression of HVGs. The MLPClassifier from scikit-learn v1.1.1 was used to build a training model, evaluated with 10-fold cross-validation. Performance was measured using accuracy and Matthews Correlation Coefficient (MCC) calculated using a confusion matrix with a One vs. Rest strategy. After evaluation, we rebuilt the MLP model using all cells from epithelial component (excluding injured cluster) and performed the predictions for injured cells.

Proximal tubule and injured cells subset in kidney scRNA-Seq: PT clusters (see Fig. 3g) were subset, each contributing over 10% to the injured cell population, along with their associated injured epithelial cells in the prediction and re-computed the PCs using the same 3000 HVGs. Next a SNN graph and UMAP was built over top 5 PCs, followed by clustering with resolution=0.1.

EMT, Differentiation and inflammation score for PT and injured trajectory in kidney scRNA-Seq: The Hallmark_EMT from MSigDB (https://www.gsea-msigdb.org/gsea/msigdb) was used to calculate

the EMT score using AddModuleScore function. The differentiation score was calculated using common up-regulated genes from PT segments[25]. The genes belonging to the kidney inflammatory pathways reported by Wu et. al.[27] were used to calculate the injury/inflammation score.

Cancer cell subset for downstream analysis: Another round of clustering was performed with resolution=0.8 which resulted in 28 clusters, out of which 19 were cancer cell clusters (tdTomato expression). Among them, 6 clusters were identified with high proportion of ribosomal genes and excluded from the downstream analysis. Cells from the remaining 13 clusters were used to calculate the PCs followed by constructing a SNN graph and UMAP over top 20 PCs. To detect the cancer cell clusters, we performed clustering with resolution=0.6. The Markov Affinity-based Graph Imputation of Cells (MAGIC v.3.0.0)[58] was applied to impute the expression of EMT-TFs encoding genes.

Trajectories inference using PAGA and RNA-Velocity: An integrated Seurat object (proximal tubule and injured cells; cancer cells subset) was exported into a Scanpy[77] v1.6.0 compatible loom file. Using precomputed PCs and cell embeddings, we constructed a neighbourhood graph with 15 neighbours and top PCs (Kidney: 5; cancer subset: 20). A connectivity map was built using PAGA[28]. To infer the directionality of the transcriptional changes for pre-defined EMT trajectories in cancer subset, we subset the clusters: 5,11,10,14,12,16,13,15 and 1. We redefined the UMAP embedding over the top 17 PCs with min.dist=0.2. Run10x utility from the velocyto[29] v0.17.17 was used to calculate sample-specific spliced/unspliced counts. Gene filtering was applied followed by the detection of 3000. Normalized spliced/unspliced counts were used for PCA, selecting top PCs (PCs=147 for kidney; PCs=105 for cancer data) based on automatic detection of elbow point (cumulative variance ratio > 0.002). Data imputation was performed using 500 neighbours, and RNA velocity was estimated assuming a steady-state transition. Velocity was visualized over UMAP embedding using a regularized grid with a Gaussian kernel with step size of 40. Additionally, For EMT trajectories in cancer we run Slingshot v2.2.0 (https://bioconductor.org/packages/release/bioc/html/slingshot.html) over top 15 PCs obtained in RNA-Velocity analysis and fitted principal curves to predict lineages and infer bifurcation point.

Pseudotime analysis for the inferred trajectories: The root cell was set as ATTCTTGAGTGCAAAT-1_2 for the PT-injured population (maximum expression of PT marker, Slc22a12) and CTGATAGGTAAGAGGA_1 for the cancer subset (maximum expression of epithelial gene, Lalba). The diffusion map was built using the top PCs (kidney:5; cancer:15) followed by pseudotime calculation.

SCENIC Analysis for regulon prediction: pySCENIC[30] v0.11.0 was used to predicted the expression-based regulon. A co-expression matrix was constructed between Transcription Factors (TFs) and their target genes using GRNBoost2 method. The motif enrichment was performed for the target genes using scenic mm10 motif database (500bp upstream and 100bp downstream region to the TSS). For EMT trajectories in cancer, we set the normalized enrichment score to >=1.75 and filtered out the target genes. The single-cell regulatory activity was calculated using AUCell algorithm and average AUC score was used for the representation. Additionally, AUC scores for selected TFs were plotted over pseudotime, and a local regression curve was fitted using generalized additive model with splines of degree=5.

Trajectory-based differential expression analysis: Integrated Seurat objects for trajectories were converted to Monocle3[78] v.0.2.2 object with pre-computed 3000 HVGs, PCA and UMAP embeddings. Reduced dimensional space was used to construct the principal graph using reversed graph embedding. Moran's-I test was used to predict the differentially expressed genes along the trajectory. Genes with significant difference over trajectory were retained. The smoothed min-max normalized expression was represented as a heatmap. Pathway enrichment analysis was performed for the differentially expressed genes using R based EnrichR[79] API v.3.0.

Trunk Neural Crest scRNA-seq Data Analysis: The raw gene expression matrix for trunk neural crest scRNA-seq[18] was downloaded from NCBI Gene Expression Omnibus (GEO) database submitted under GEO accession GSE129114. The connectivity map was built using PAGA (resolution=0.5) as described earlier. For the pseudotime analysis SS2_15_0085_F22 was used as a root cell with maximum expression for neuronal differentiation marker, *Hes5*. Finally, the Moran's I test was performed to infer the differentially expressed genes as described above and the gene set enrichment analysis was performed for EMT Hallmark and BC-PING signature.

Cancer cell FACS sorting

Tumour cells suspensions were prepared from mammary gland carcinomas obtained from 14-15 weeks old female mice following the protocol used for cancer single cell preparation. Digestion buffer was adjusted to contain 1.0 Wünsch units of TH Liberase/ml and the incubation time extended to 75min. After red blood cells removal, cancer cells were neutralized and resuspended in 0.5ml FACS buffer with DAPI per 10^7 cells and directly sorted using BD FACSAria III flow cytometer. For each sample, 300K cancer cells were sorted at high purity following: singlet / $DAPI^{low}$ / $tdTomato^{high}$ gating strategy. Post sorting analysis was performed to verify the purity of $tdTomato^{high}$ sorted cells. Sorted cancer cells were centrifuged at 5000 rpm for 3 min and further

lysated in the lysis buffer from illustra RNAspin Mini kit (GE Healthcare). RNA extraction, cDNA synthesis and RT-qPCR were performed as described above.

Statistical analyses

All experiments were repeated at least three times and the number of independent experimental replicates is indicated for each experiment in figure legends and Source Data 2. Two technical replicates were used per experiment and averaged for each experiment. Unless stated in the legend of the figures, all statistical analyses were performed using Prism 6 (GraphPad). Statistical analyses were indicated in figure legends and in the statistical table in Source Data 1. *p*-values > 0.05 were considered not statistically significant.

References

61. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
62. Chung, W. *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* **8**, 15081 (2017).
63. Madisen, L. *et al.* A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci.* **13**, 133–140 (2010).
64. Shao, X., Somlo, S. & Igarashi, P. Epithelial-specific Cre/lox recombination in the developing kidney and genitourinary tract. *J. Am. Soc. Nephrol. JASN* **13**, 1837–1846 (2002).
65. Dassule, H. R., Lewis, P., Bei, M., Maas, R. & McMahon, A. P. Sonic hedgehog regulates growth and morphogenesis of the tooth. *Dev. Camb. Engl.* **127**, 4775–4785 (2000).
66. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
67. Renier, N. *et al.* iDISCO: a simple, rapid method to immunolabel large tissue samples for volume imaging. *Cell* **159**, 896–910 (2014).
68. Foroutan, M. *et al.* Single sample scoring of molecular phenotypes. *BMC Bioinformatics* **19**, 404 (2018).
69. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).
70. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
71. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90-97 (2016).
72. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.* **16**, 284–287 (2012).
73. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
74. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinforma. Oxf. Engl.* **26**, 976–978 (2010).
75. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
76. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
77. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data

analysis. *Genome Biol.* **19**, 15 (2018).
78. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
79. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

Main Figures

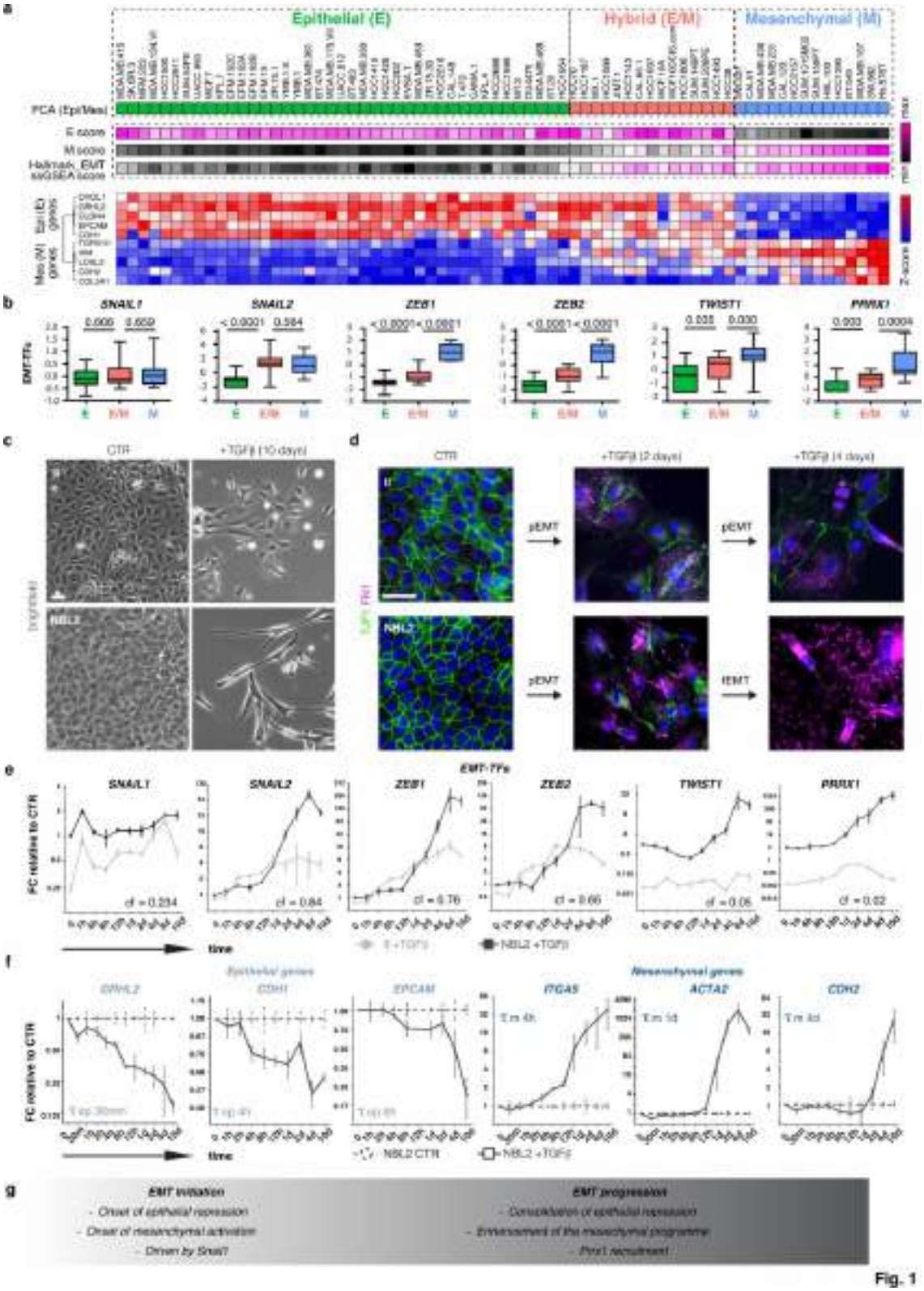

a
Epithelial (E)
Hybrid (E/M)
Mesenchymal (M)
PCA (Epi/Mes)
E score
M score
Hallmark_EMT ssGSEA score
Epi (E) genes
Mes (M) genes
Z-score
b
SNAIL1
SNAIL2
ZEB1
ZEB2
TWIST1
PRRX1
EMT-TFs
E
E/M
M
c
CTR
+TGFβ (10 days)
brightfield
NBL2
d
CTR
+TGFβ (2 days)
+TGFβ (4 days)
pEMT
lEMT
E-cad (N)
NBL2
e
SNAIL1
SNAIL2
ZEB1
ZEB2
TWIST1
PRRX1
EMT-TFs
FC relative to CTR
time
E +TGFβ
NBL2 +TGFβ
f
Epithelial genes
Mesenchymal genes
CRNN2
CDH1
EPCAM
ITGA5
ACTA2
CDH2
FC relative to CTR
time
NBL2 CTR
NBL2 +TGFβ
g
EMT initiation
- Onset of epithelial repression
- Onset of mesenchymal activation
- Driven by Snail
EMT progression
- Consolidation of epithelial repression
- Enhancement of the mesenchymal programme
- Prrx1 recruitment

Fig. 1

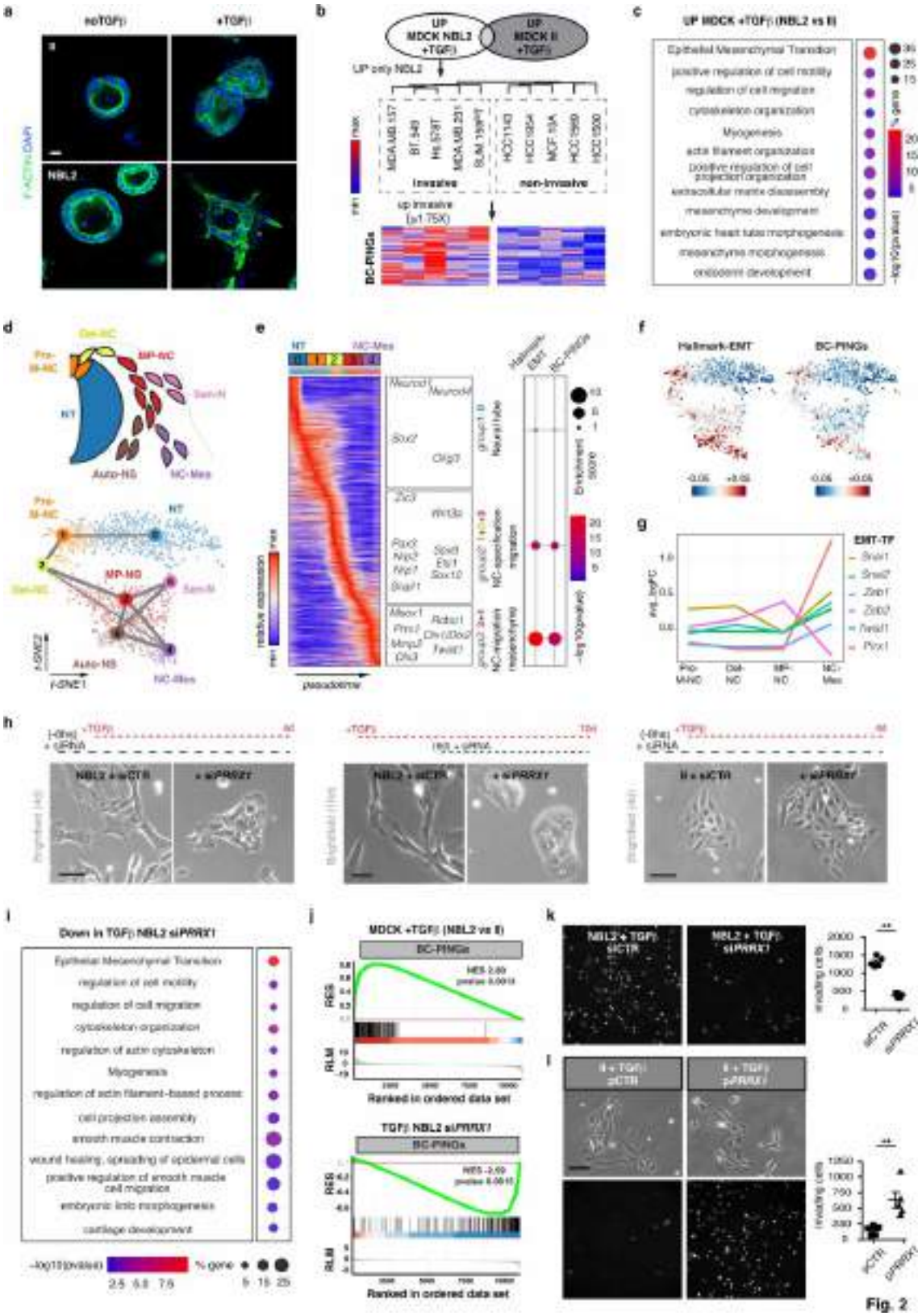

a
noTGFβ
+TGFβ
NBL2
F-ACTIN/DAPI
b
UP MDCK NBL2 +TGFβ
UP MDCK II +TGFβ
UP only NBL2
MDA.MB.157
BT 549
Hs 578T
MDA.MB.231
SUM 159PT
HCC1143
HCC1954
MCF 10A
HCC1569
HCC1500
Invasive
non-Invasive
up Invasive (p1.75X)
BC-PINGs
c
UP MDCK +TGFβ (NBL2 vs II)
Epithelial Mesenchymal Transition
positive regulation of cell motility
regulation of cell migration
cytoskeleton organization
Myogenesis
actin filament organization
positive regulation of cell projection organization
extracellular matrix disassembly
mesenchyme development
embryonic heart tube morphogenesis
mesenchyme morphogenesis
ondocrim development
% gene
-log10(padjust)
d
Pre-M-NC
MP-NC
Del-NC
NT
Sem-N
Auto-NS
NC-Mes
MP-ND
t-SNE2
t-SNE1
e
NT
NC-Mes
Neural
Neuron
Sox2
Olig3
Zic3
Wnt3a
Pax3
Nkx3
Nfz7
Snai1
Spp9
Ets1
Sox10
Msx1
Prrx1
Mmp2
Dlx3
Robo1
Dlx1/Dlx2
Fwd1
Hallmark-EMT
BC-PINGs
group1 (-)
Neural tube
group2 (+)
NC-specification
migration
group2 3+1
NC-migration
mesenchyme
Enrichment Score
-log10(padjust)
relative expression
pseudotime
f
Hallmark-EMT
BC-PINGs
g
EMT-TF
Snai1
Snai2
Zeb1
Zeb2
Twist1
Prrx1
avg. logFC
Pre-MigNC
Del-ND
MP-ND
NC-Mes
h
-TGFβ
+TGFβ
(-4hr) + siRNA
NBL2 + siCTR
+ siPRRX1
II + siCTR
i
Down in TGFβ NBL2 siPRRX1
Epithelial Mesenchymal Transition
regulation of cell motility
regulation of cell migration
cytoskeleton organization
regulation of actin cytoskeleton
Myogenesis
regulation of actin filament-based process
cell projection assembly
smooth muscle contraction
wound healing, spreading of epidermal cells
positive regulation of smooth muscle cell migration
embryonic limb morphogenesis
cartilage development
-log10(padjust)
% gene
j
MDCK +TGFβ (NBL2 vs II)
BC-PINGs
NES 2.69
pvalue 0.0010
RES
RLM
Ranked in ordered data set
TGFβ NBL2 siPRRX1
NES -2.69
pvalue 0.0010
k
NBL2 + TGFβ siCTR
NBL2 + TGFβ siPRRX1
invading cells
siCTR
siPRRX1
l
II + TGFβ siCTR
II + TGFβ siPRRX1

Fig. 2

1500

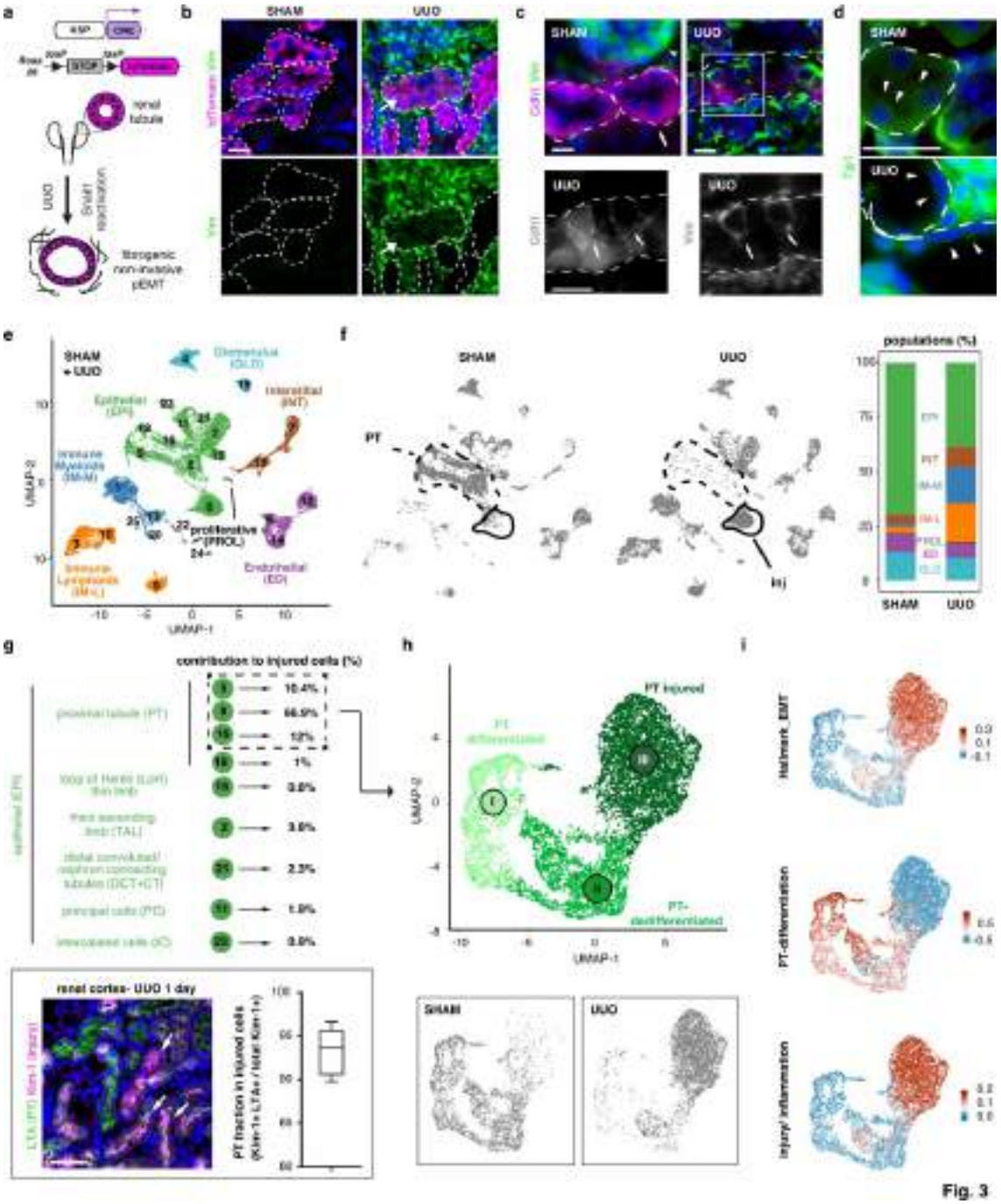

a
b
SHAM
UUO
c
d
renal tubule
UUO
Snail1 reactivation
fibrogenic non-invasive pEMT
e
SHAM + UUO
Glomerular (GLO)
Epithelial (EPI)
Interstitial (INT)
Immune Myeloid (IMM)
proliferative (PROL)
Endothelial (END)
Immune Lymphoid (IMM-L)
UMAP-1
UMAP-2
f
SHAM
UUO
PT
populations (%)
EPI
PT
IMM
IM-L
PROL
END
GLO
g
contribution to injured cells (%)
proximal tubule (PT)
loop of Henle (LOH) thin limb
thick ascending limb (TAL)
distal convoluted/ connecting tubules (DCT+CT)
principal cells (PC)
intercalated cells (IC)
10.4%
66.9%
12%
1%
3.8%
2.3%
1.9%
0.8%
renal cortex- UUO 1 day
PT fraction in injured cells (Kim-1+ LTA+ / total Kim-1+)
h
PT injured
PT differentiated
PT-dedifferentiated
UMAP-1
UMAP-2
SHAM
UUO
i
Hallmark_EMT
PT-differentiation
Injury/ inflammation

Fig. 3

1510

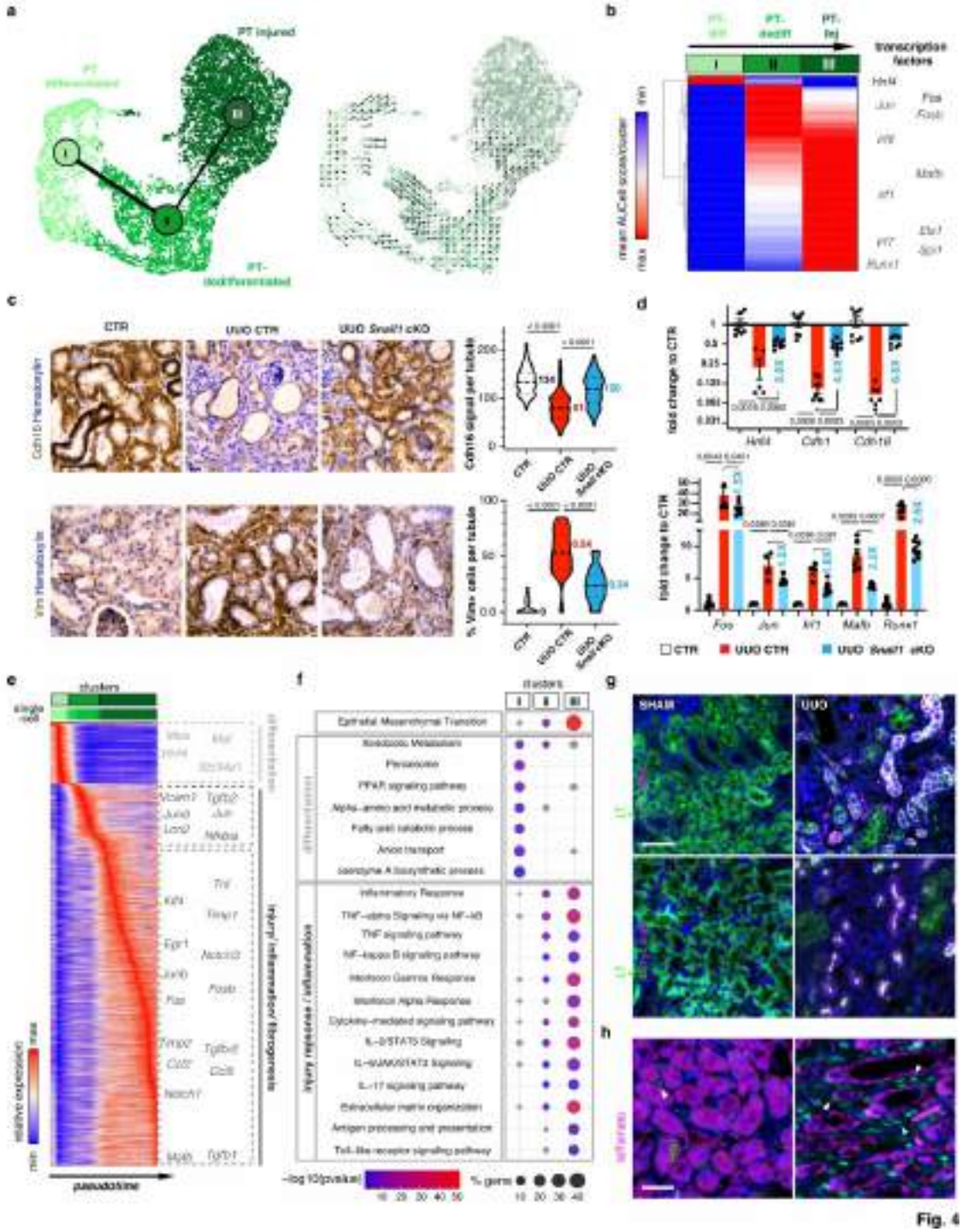


Fig. 4

1520

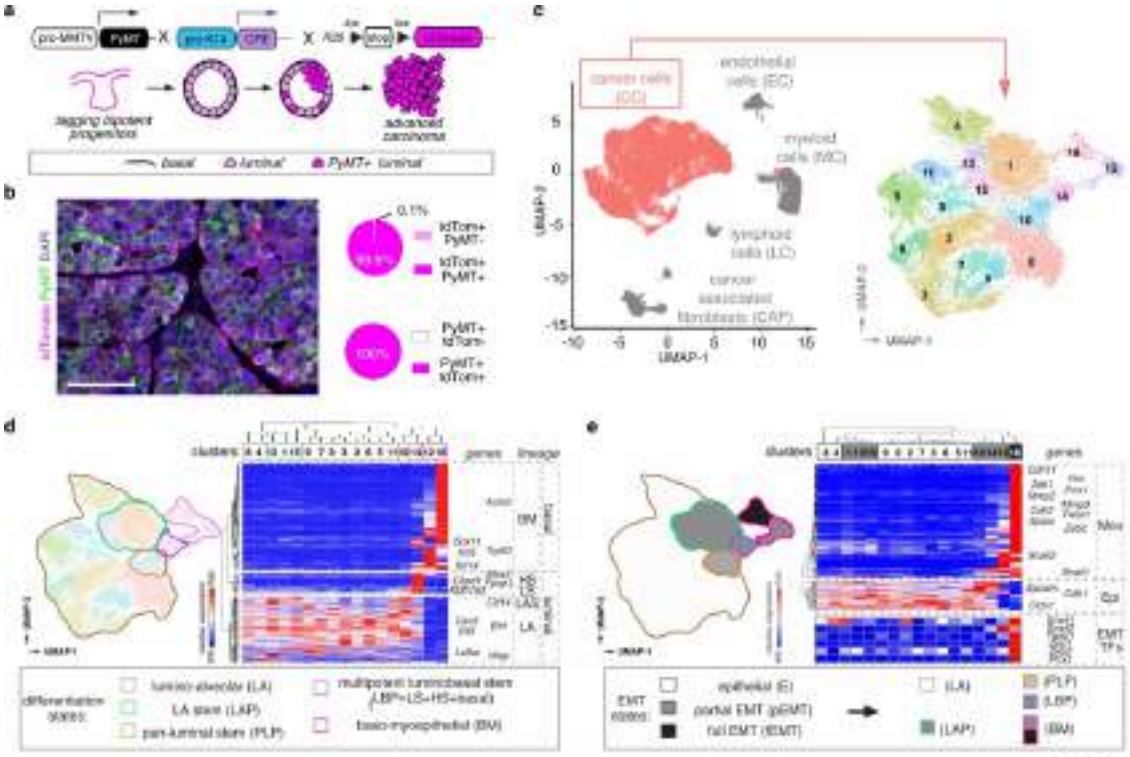


Fig. 5

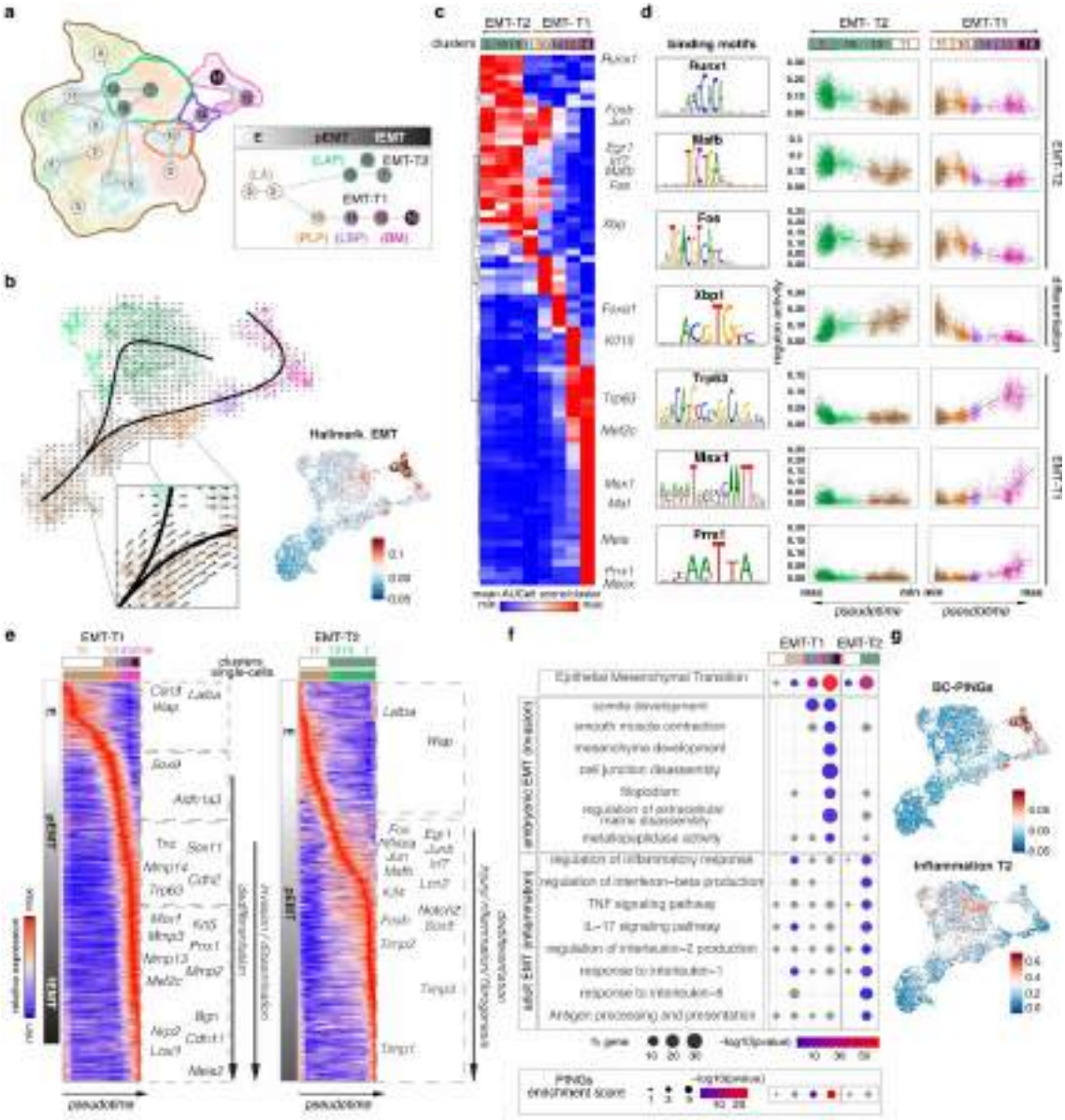


Fig. 6

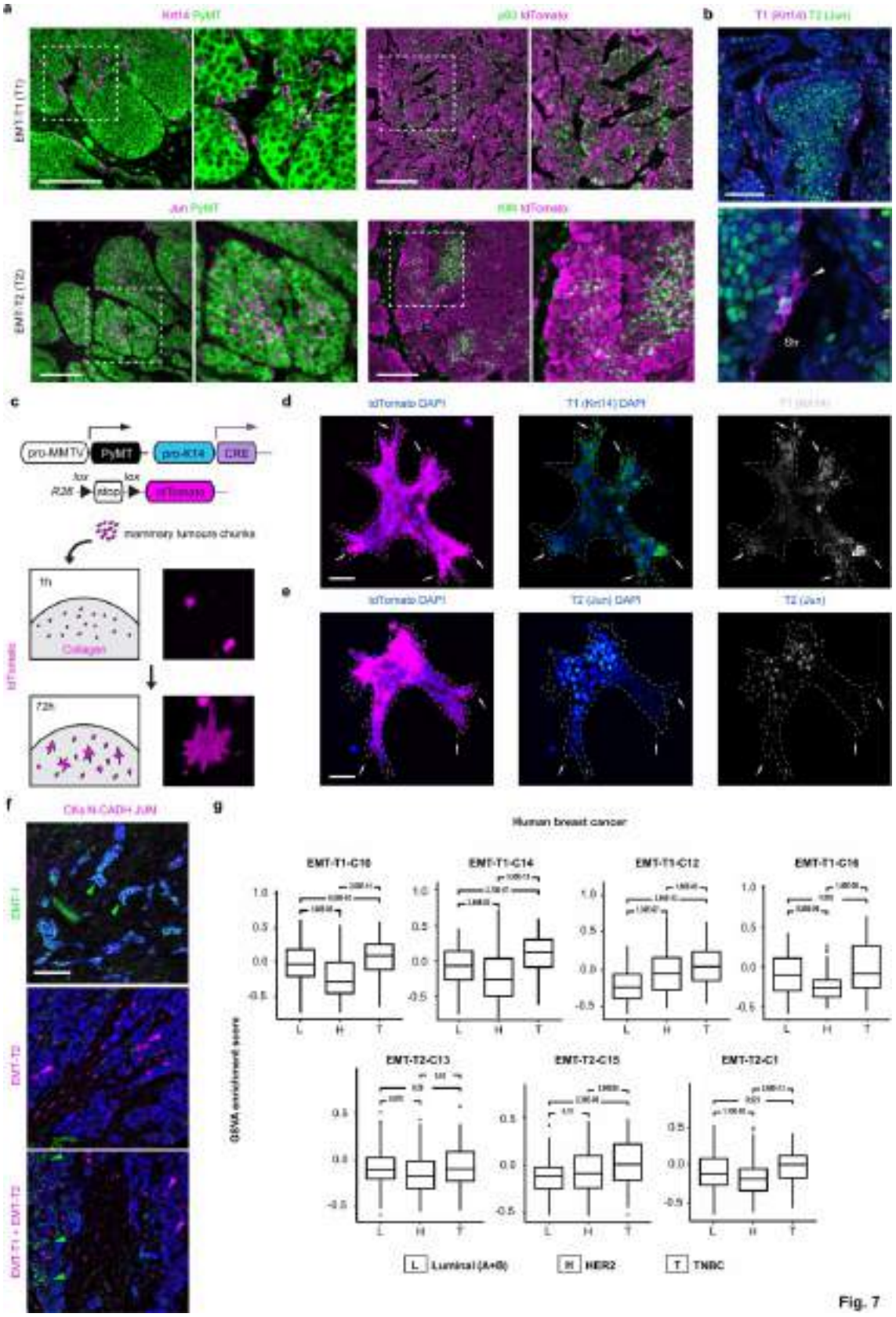

a
Krt14 PyMT
p53 tdTomato
EMT-T1 (T1)
Jun PyMT
Fos tdTomato
EMT-T2 (T2)
b
T1 (Krt14) T2 (Jun)
8h
c
pro-MMTV
PyMT
pro-Krt14
CRE
lox
R26
stop
tdTomato
mammary tumours chunks
1h
Collagen
72h
tdTomato
d
tdTomato DAPI
T1 (Krt14) DAPI
e
tdTomato DAPI
T2 (Jun) DAPI
T2 (Jun)
f
CK14 N-CADH JUN
EMT-T1
EMT-T2
EMT-T1 + EMT-T2
g
Human breast cancer
EMT-T1-C10
EMT-T1-C14
EMT-T1-C12
EMT-T1-C16
EMT-T2-C13
EMT-T2-C15
EMT-T2-C1
GSVA enrichment score
L
H
T
L Luminal (A+B)
H HER2
T TNBC

Fig. 7

1580

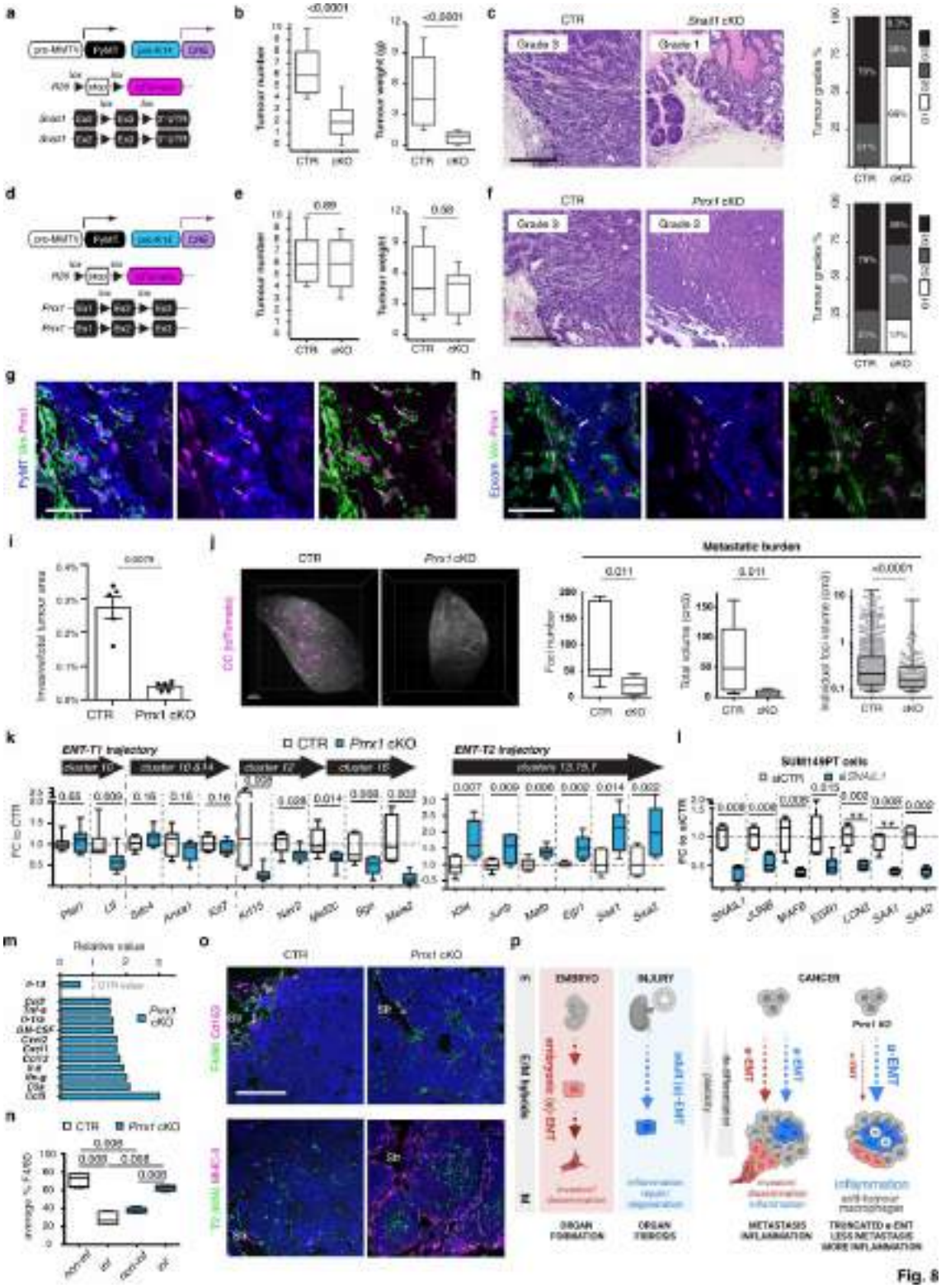
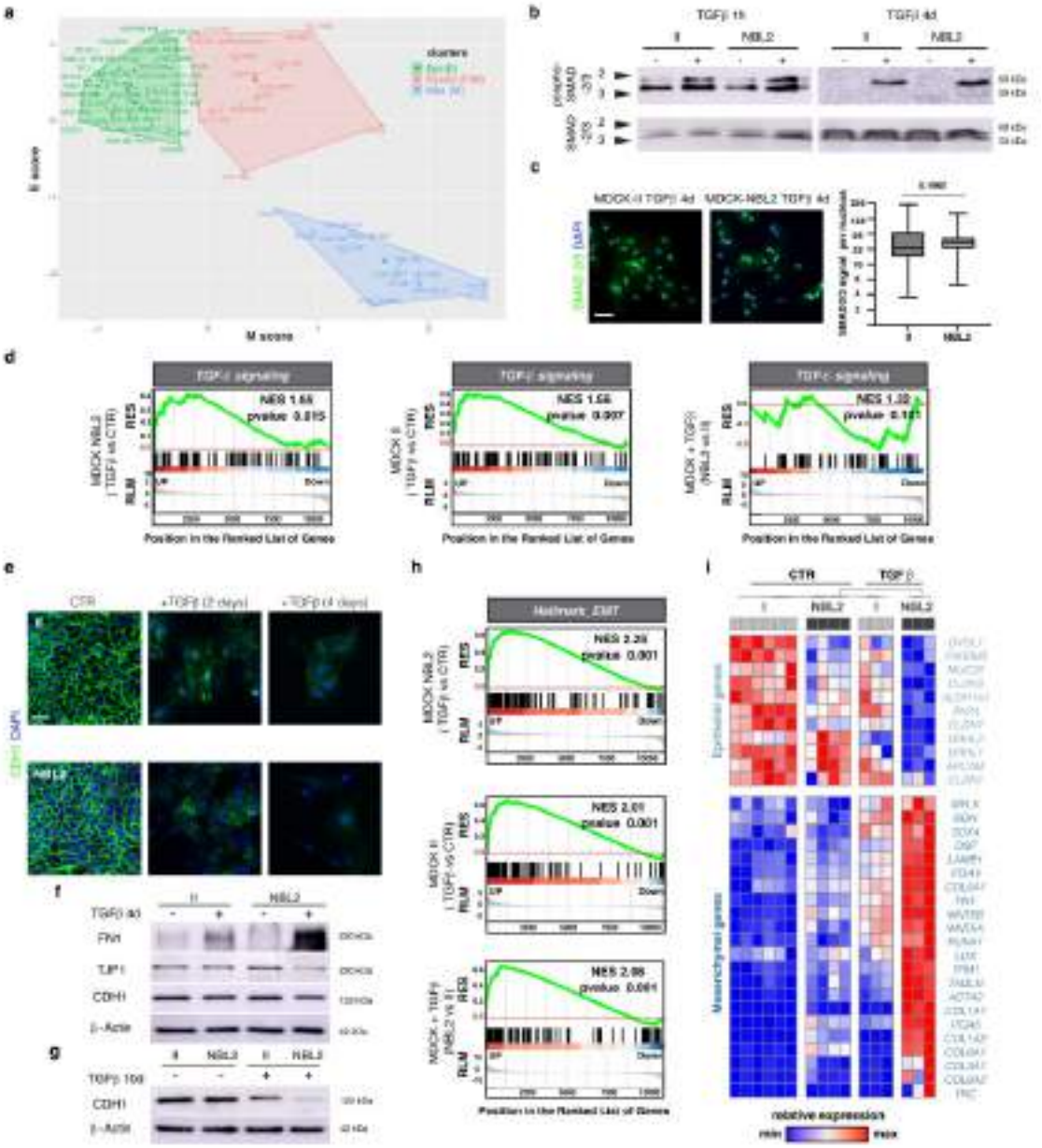


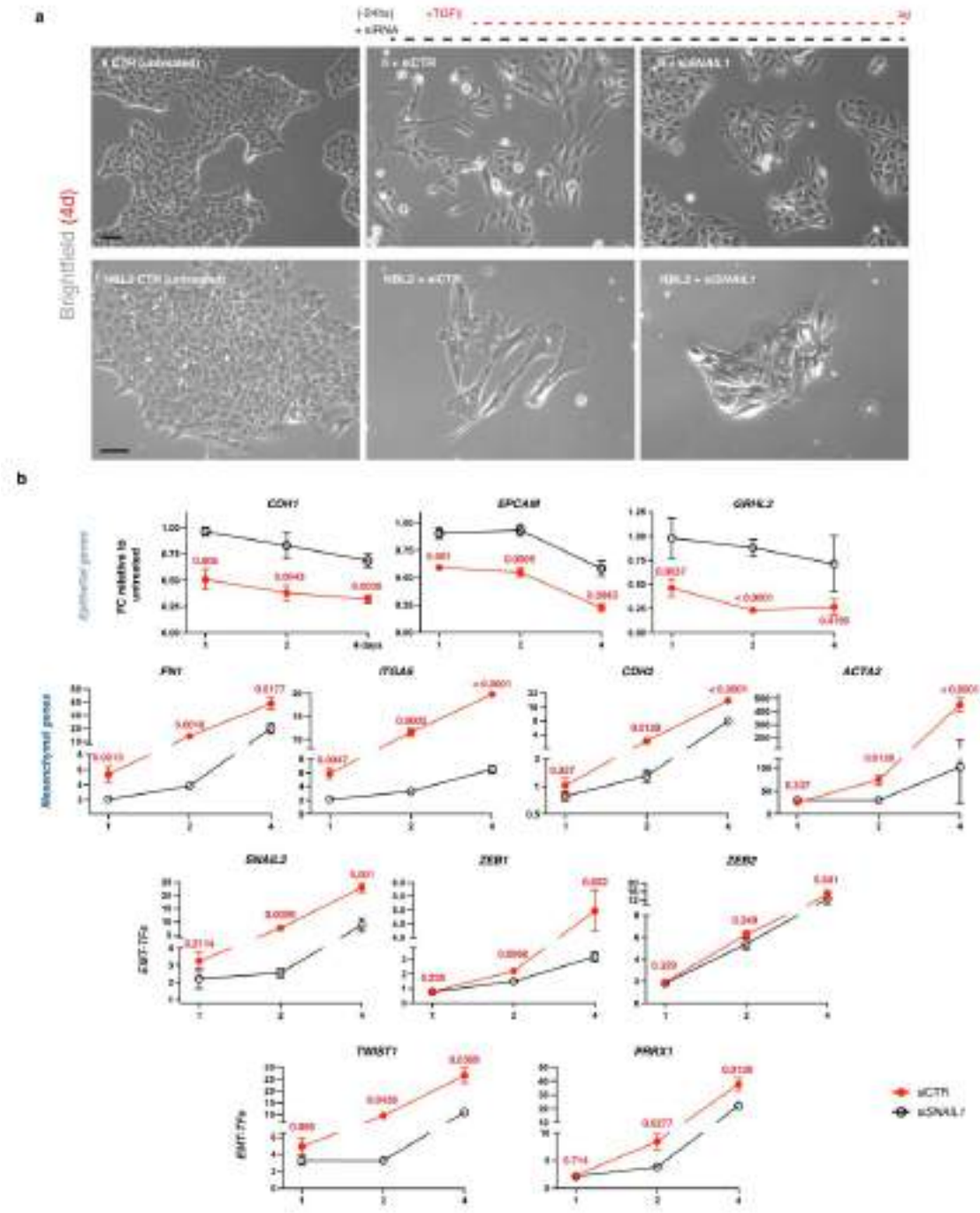
Fig. 8

Supplementary Figures

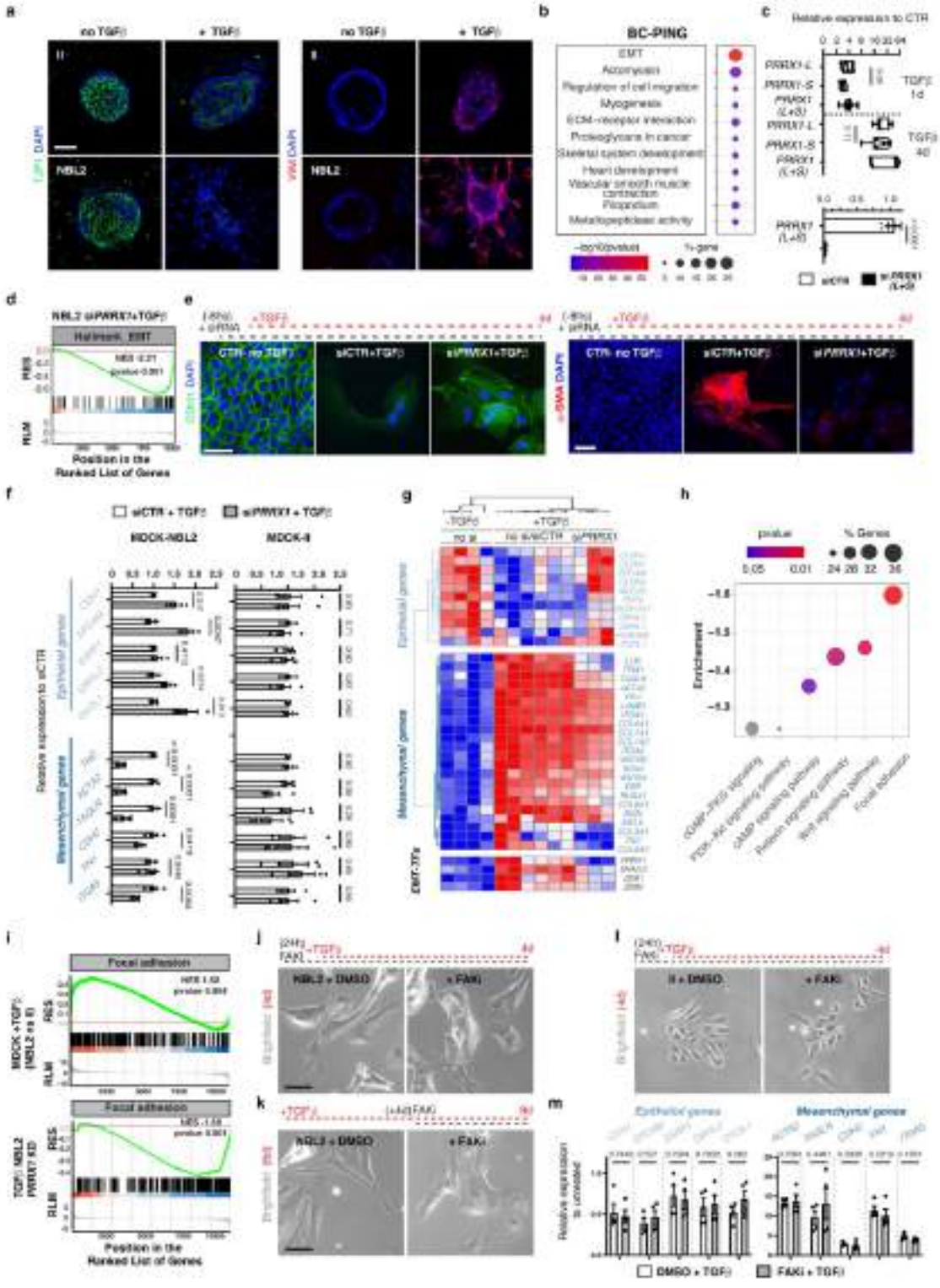


Extended Data Fig. 1 (related to Fig. 1)

1600

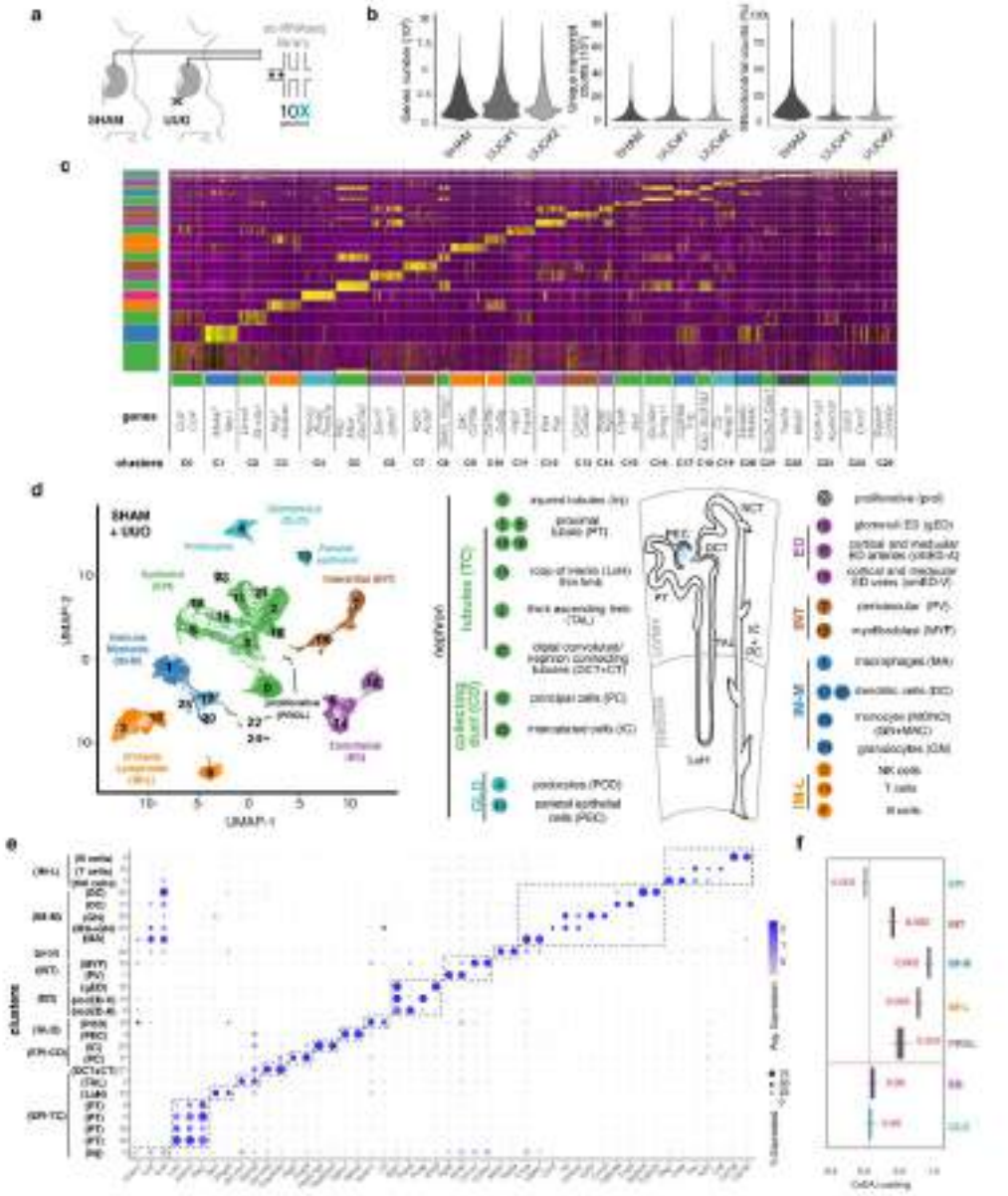


Extended Data Fig. 2 (related to Fig. 1)

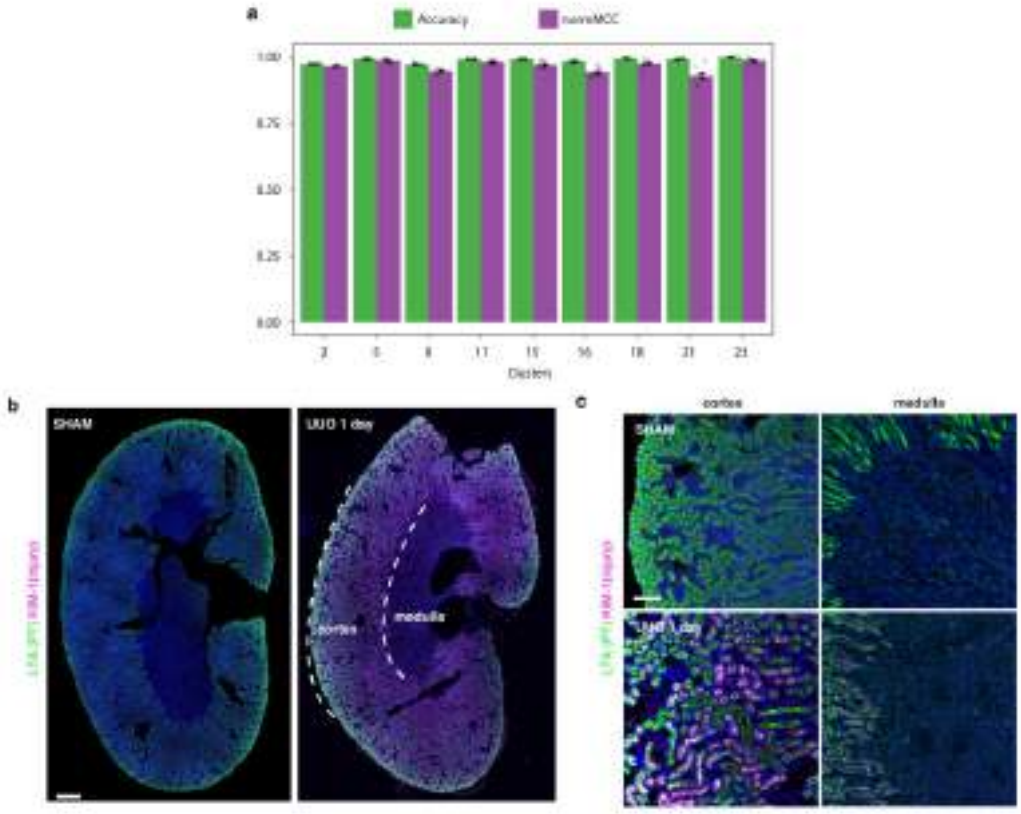


1610

Extended Data Fig. 3 (related to Fig. 2)

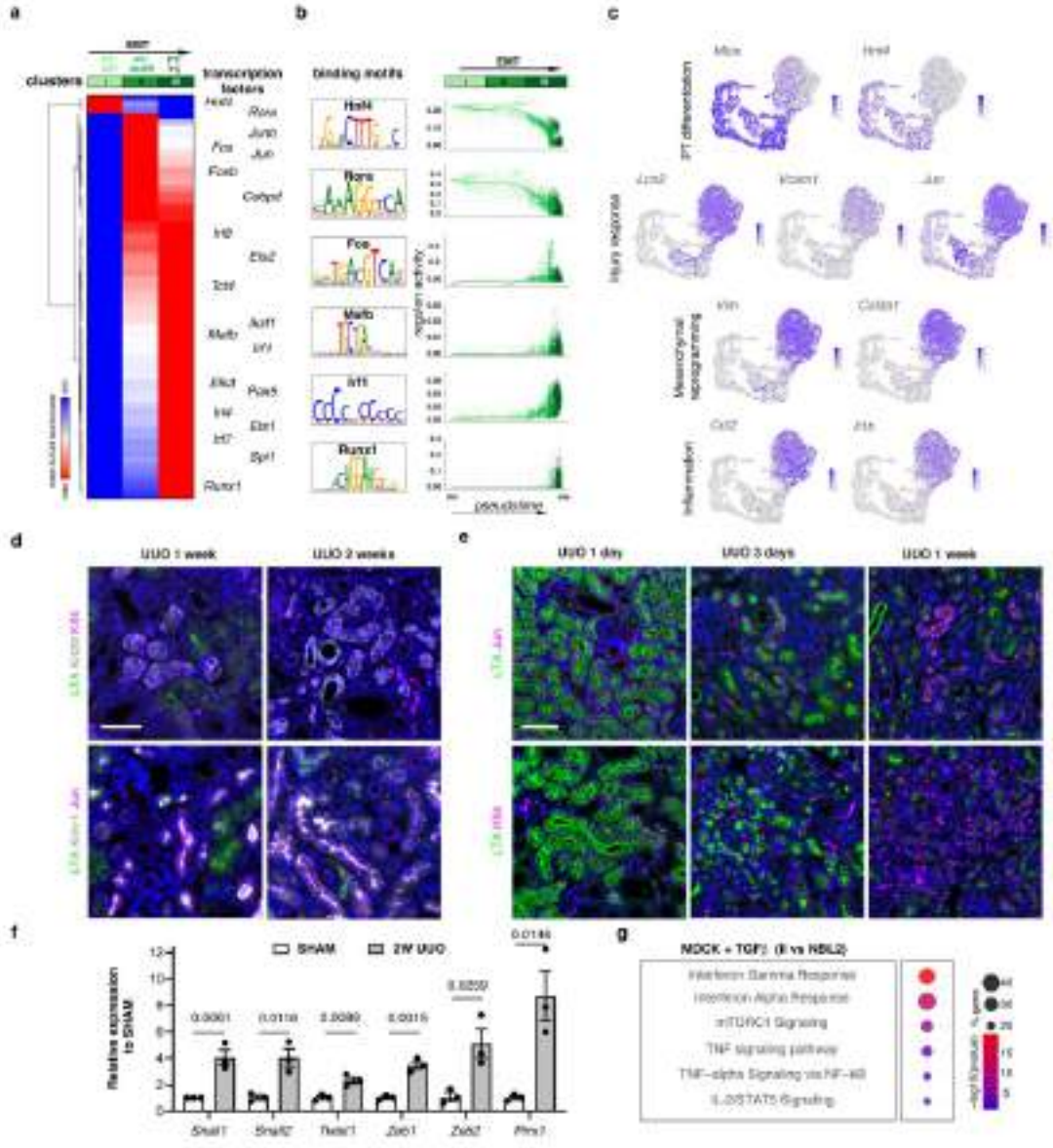


Extended Data Fig. 4 (related to Fig. 3)



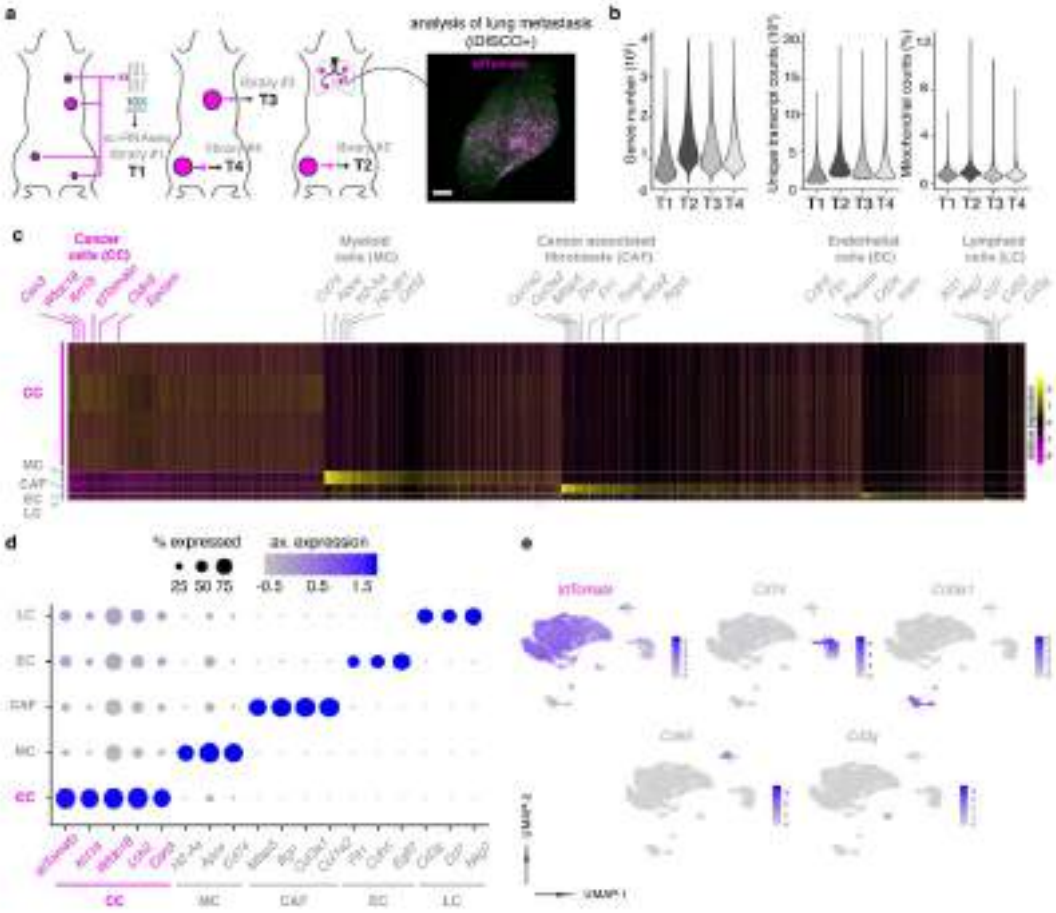
Extended Data Fig. 5 (related to Fig. 3)

1630

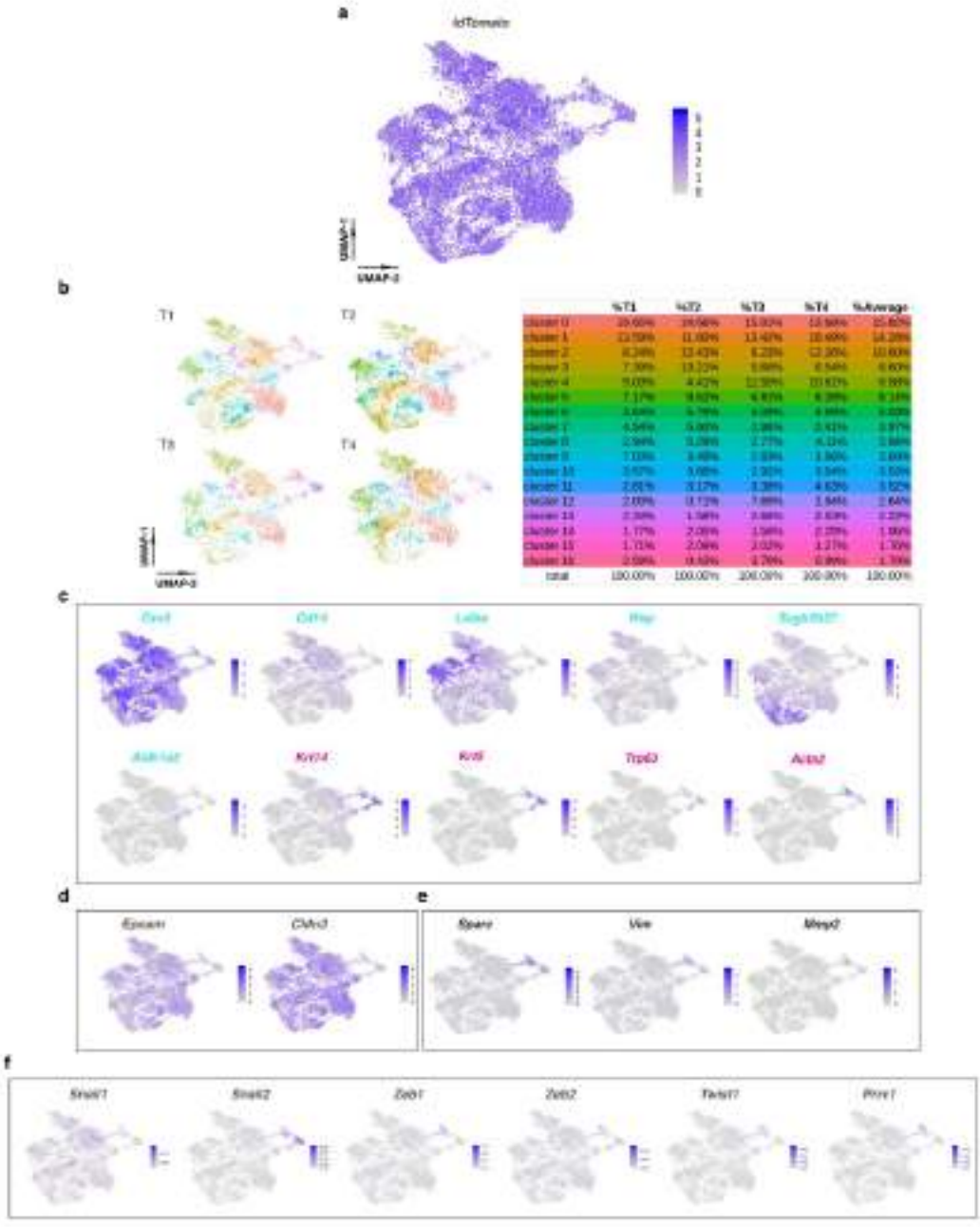


Extended Data Fig. 6 (related to Fig. 4)

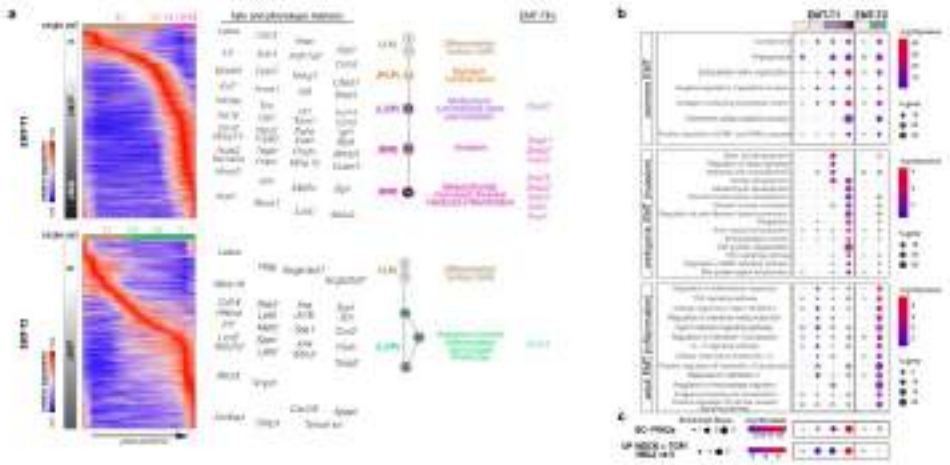
1640



Extended Data Fig. 7 (related to Fig. 5)

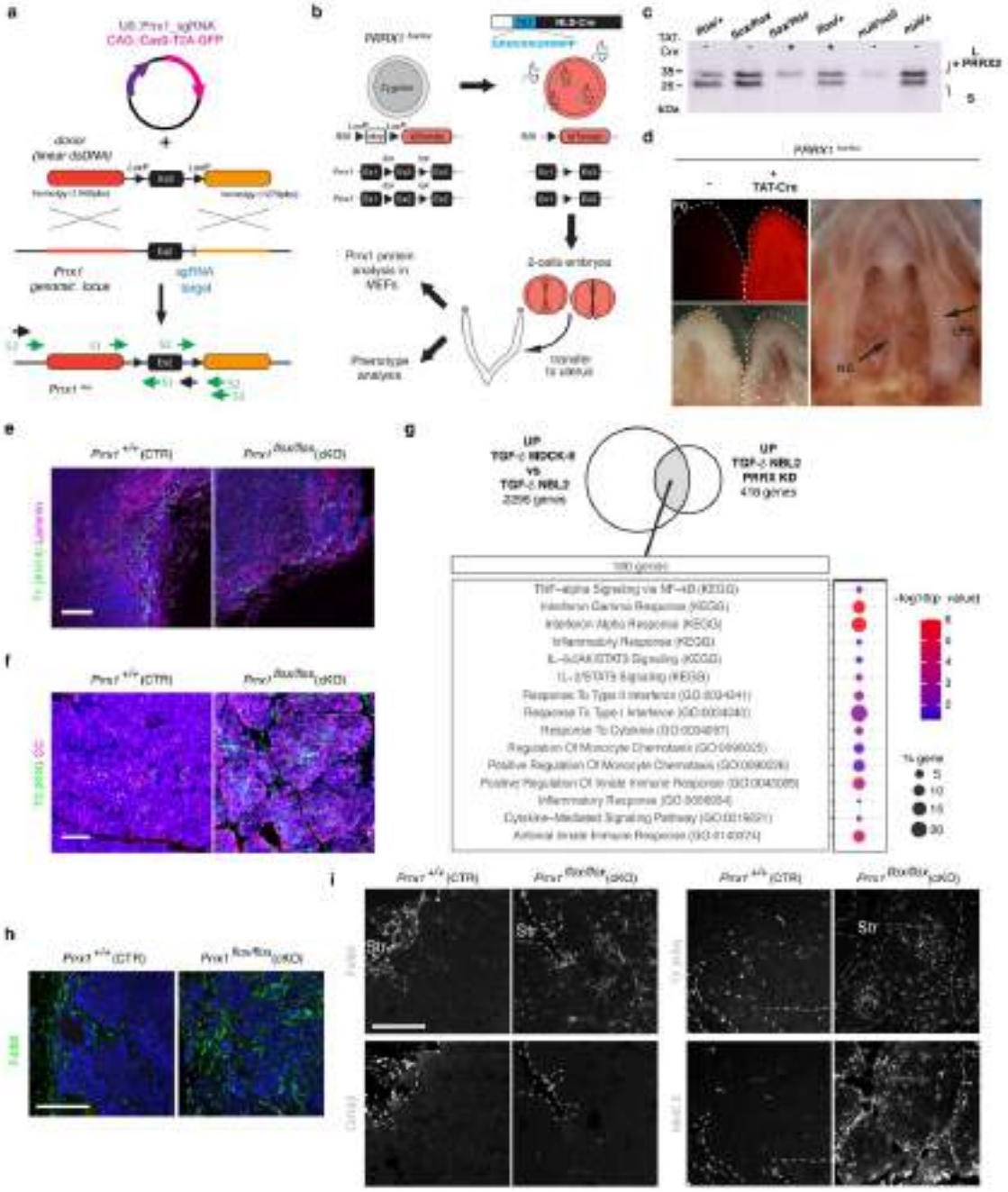


Extended Data Fig. 8 (related to Fig. 5)



1660

Extended Data Fig. 9 (related to Fig. 6)



Extended Data Fig. 10 (related to Fig. 8)

1680

Extended Data Table 1, E and M genes

Refs 18 and 89

Epithelial (E) genes		286 genes				
CDH1	*KRT15*	*DDR1*	*MALL*	*ANXA4*	*POLR3G*	*KLK8*
RAB25	*EGR2*	*VGLL1*	*EPN3*	*TNFSF13*	*IFI30*	*TMEM40*
SPINT2	*EPCAM*	*SFN*	*FOXA1*	*OCLN*	*LOC729884*	*TRIM29*
TMEM30B	*KRT19*	*TSPAN13*	*PYCARD*	*SLC9A3R1*	*TMPRSS11E*	*HBEGF*
ST14	*TACSTD2*	*PKP3*	*SLC37A1*	*XBP1*	*C1ORF116*	*ALDH1A3*
S100A14	*S100P*	*ITGB6*	*TSPAN15*	*C20ORF19*	*ALOX15B*	*RBM35B*
PRSS8	*GALNT3*	*LY75*	*HNMT*	*SNX10*	*COL17A1*	*CYP27B1*
JUP	*FXYDD3*	*MPAPK13*	*ABCC3*	*TP73L*	*RTEL1*	*IL1B*
KRT17	*SCNN1A*	*TTC39A*	*SDC1*	*BDKRB2*	*TNFRSF6B*	*NMU*
CDS1	*ESRP1*	*CEACAM1*	*TOB1*	*ANXA8*	*PTPN3*	*KRT16*
ITGB4	*CLDN7*	*DTX4*	*B3GNT3*	*ANXA8L1*	*E2F5*	*JAG2*
TMPRSS4	*ERBB3*	*ERBB2*	*TMC6*	*LOC728113*	*KRT6B*	*VSNL1*
LSR	*RBM47*	*RAB11FIP1*	*CD*	*RHBDF2*	*STAC*	*RLN2*
GRHL2	*SPINT1*	*ATP2C2*	*ADAP1*	*LOC653562*	*C6ORF105*	*CTSL2*
TSPAN1	*ELF3*	*TGFA*	*ATP1B1*	*SLC6A10P*	*GLS2*	*SYK*
SLPI	*CLDN4*	*MYO6*	*SAHNK2*	*SLC6A8*	*ANXA3*	*SAA1*
ARHGAP8	*SH3YL1*	*PTK6*	*CYB561*	*THBD*	*DST*	*EPB41L4B*
F11R	*EHF*	*OAS1*	*ERMP1*	*NEFM*	*ARHGAP25*	*RNF128*
LAD1	*LCN2*	*FBP1*	*RAB20*	*RPS6KA1*	*CLDN1*	*LEPREL1*
KRT18	*VAMP8*	*AQP3*	*MYH14*	*SMPDL3B*	*FLJ20366*	*PI3*
CDH3	*KRT8*	*CBLC*	*CAPN1*	*ABCA12*	*CYP4F11*	*CKMT1B*
MYO5C	*C1orf106*	*BSPRY*	*ALDH3B2*	*KRT14*	*CCND2*	*LOC553158*
IRF6	*DSP*	*EPS8L1*	*TRIM31*	*PRKCH*	*FGFR2*	*IGFBP2*
KCNK1	*SORL1*	*GRB7*	*ARAP2*	*ZBED2*	*ABLIM1*	*IL18*
MYO1D	*PPL*	*C4orf19*	*SSH3*	*C10ORF10*	*XDH*	*CA9*
ELMO3	*C1orf116*	*KLK6*	*ICA1*	*LRRC1*	*CAMK2B*	*CA2*
MST1R	*MAP7*	*TJP2*	*ARHGEF5*	*IL4R*	*DSG3*	*EVA1*
AP1M2	*TOX3*	*DENND2D*	*ALOX5*	*FGFBP1*	*NAIP /// OCLN*	*S100A7*
EPHA1	*GPX2*	*EPS8L2*	*TMPRSS2*	*FAT2*	*SAA1 /// SAA2*	*KLK7*
SH2D3A	*CTSH*	*IL20R1*	*MTUS1*	*WWC1*	*PRRG4*	*LGALS7*
PLS1	*GPR56*	*HES1*	*CYP4F3*	*FZD3*	*C10ORF116*	*FST*
IL1RN	*FA2H*	*ARHGDIB*	*PPFIBP2*	*SNCA*	*CST6*	*CXADR*
EXPH5	*KLF5*	*C19orf21*	*RABGAP1L*	*KIAA1815*	*NDRG1*	*RBM35A*
PERP	*AREG*	*CAMK2N1*	*PLXNB2*	*GNAL*	*S100A8*	*UCHL1*
DSC2	*SCEL*	*HPGD*	*MGST2*	*KIAA0888*	*CORO1A*	*KLK10*
BIK	*UGT1A1*	*SYNGR2*	*AR7E14P*	*KRT5*	*KLK5*	*TACSTD1*
STAP2	*MPZL2*	*C10orf116*	*EVPL*	*GJB3*	*IRX4*	*SERPINB2*
ZNF165	*AIM1*	*MANSC1*	*CD46*	*KIAA0040*	*HOOK1*	*SPRR1A*
ANK3	*OVOL2*	*POF1B*	*CNKSR1*	*CELSR2*	*ARTN*	*FGFR3*
CKMT1A	*LLGL2*	*SERINC5*	*BLNK*	*NUP62CL*	*FLJ12684*	*SPRR1B*
RHOD	*ESRP2*	*ANXA9*	*COMT*	*SERPINB1*	*SLC2A9*	

Mesenchymal (M) genes		130 genes	
VIM	*BAG2*	*CCDC92*	*CYBRD1*
ZEB1	*MXRA7*	*WNT5A*	*PPAP2B*
TUBA1A	*GFPT2*	*IGFBP3*	*PTX3*
PMP22	*RECK*	*PPM1D*	*FADS2*
CHN1	*TMEFF1*	*FILIP1L*	*BGN*
COL5A2	*PTRF*	*PDGFC*	*TSHZ1*
MYL9	*FBLN5*	*TBX3*	*ZBTB38*
FSTL1	*GREM1*	*XYLT1*	
EMP3	*COL3A1*	*FAP*	
SACS	*COL1A2*	*DPT*	
AXL	*DCN*	*STC1*	
LOXL2	*CDH2*	*KRT81*	
SPARC	*ENPP2*	*MMP1*	
FHL1	*POSTN*	*HS3ST2*	
FERMT2	*RGS4*	*LMCD1*	
TMEM158	*C5ORF13*	*N-PAC*	
CALD1	*PRRX1*	*SEPT6*	
LGALS1	*FBN1*	*NR2F1*	
MSN	*SRGN*	*SCCPDH*	
GLYR1	*SPOCK1*	*MLPH*	
MAP1B	*PRR16*	*LTBP2*	
GJA1	*DLC1*	*TPM1*	
DENND5A	*BIN1*	*ANKRD25*	
C12orf24	*RGL1*	*DDR2*	
TPM2	*IGFBP4*	*SEMA5A*	
TUBB6	*PVRL3*	*TGFB1I1*	
SPRX	*CDH11*	*PCOLCE*	
ANK2	*OLFML3*	*STARD13*	
SH2B3	*MMP2*	*NID1*	
LEPRE1	*CTGF*	*SYNC1*	
ETV1	*PLEKHC1*	*ENOX1*	
SOBP	*ROR1*	*MME*	
AKAP12	*PTGER2*	*C10ORF56*	
TGFB1L1	*TRAM2*	*LTBP1*	
SERPINE1	*TAGLN*	*NRP1*	
SOAT1	*TNFAIP6*	*THY1*	
LHFP	*CREB3L1*	*NEBL*	
CEP170	*UGDH*	*TNS3*	
POPDC3	*HAS2*	*ECM1*	
TRPC1	*DNAJB4*	*FBLN1*	
KDELC1	*CDKN2C*	*COPZ2*	

Extended Data Table 2, BC-PINGs genes

	Ratio inv/non-inv				
COL1A2	221.25	*TRAM2*	5.44	*NRP1*	3.21
COL3A1	129.28	*DKK3*	5.39	*MEDAG*	3.16
COL1A1	67.77	*HMCN1*	5.23	*GNG11*	3.12
FBN1	45.97	*MMP14*	5.17	*FAM126A*	3.10
SULF1	45.86	*AKR1B1*	5.14	*PDPN*	3.10
ACTA2	30.69	*NPTX2*	5.14	*SERPINH1*	3.09
SPARC	25.47	*FGF2*	5.13	*PDE4A*	3.05
LUM	19.96	*SEMA3D*	5.08	*LPAR1*	3.03
COL5A1	19.00	*MYOCD*	5.01	*DKK1*	3.03
ACTG2	16.38	*PRUNE2*	4.77	*CXCL8*	3.01
FLNC	14.62	*ZEB2*	4.76	*FBLIM1*	2.97
VCAN	13.98	*PAPPA*	4.76	*FSTL3*	2.95
SRGN	13.22	*ITGBL1*	4.75	*ANXA6*	2.95
TAGLN	13.18	*PRKCA*	4.53	*HSPG2*	2.94
ALPK2	12.34	*MSC*	4.43	*FAM171A1*	2.90
FN1	11.47	*COL6A1*	4.35	*PAPSS2*	2.87
ADAMTS2	10.23	*RGS4*	4.31	*MYH9*	2.86
NEXN	10.17	*NPTX1*	4.30	*INHBE*	2.86
DPYSL3	9.86	*APBA2*	4.30	*ZFPM2*	2.86
ITGA5	8.88	*PTRF*	4.13	*LPXN*	2.86
FHL1	8.25	*CPA4*	4.12	*EML1*	2.85
ADAM12	8.15	*EHD2*	4.11	*P4HA1*	2.82
CALD1	7.85	*EMP3*	4.10	*FBXL7*	2.78
GALNT5	7.80	*CLIC4*	4.09	*TFPI*	2.75
MAP1B	7.79	*EPHA5*	4.05	*NID2*	2.74
VGLL3	7.63	*FERMT2*	4.02	*FAM20C*	2.73
COL16A1	7.31	*GLIPR1*	3.90	*TPM4*	2.68
CSPG4	7.30	*ATOH8*	3.84	*GALNT10*	2.67
COL15A1	7.28	*TPM2*	3.71	*PTGFRN*	2.66
IRX2	7.28	*ADAMTS6*	3.53	*CACNA1C*	2.66
MYL9	7.02	*MSRB3*	3.46	*FAM26E*	2.62
PRRX1	6.75	*GNG12*	3.45	*ROBO1*	2.62
ZEB1	6.54	*LEPRE1*	3.44	*CALU*	2.59
ADAMTSL1	6.48	*LRP1*	3.42	*NMT2*	2.58
COL6A2	6.45	*MRC2*	3.37	*GJA1*	2.57
DLC1	6.27	*UAP1*	3.36	*ANTXR1*	2.56
DAB2	6.25	*TGFB1I1*	3.34	*KANK2*	2.55
HAS2	6.06	*GPR176*	3.34	*FAM198B*	2.54
MICAL2	6.01	*APBB1IP*	3.30	*COL18A1*	2.53
LRRC32	6.01	*ID3*	3.29	*SLC2A3*	2.53
SPON2	5.96	*DNMBP*	3.26	*TLN1*	2.52
WNT5A	5.78	*CA9*	3.25	*ZNF502*	2.52
LOX	5.60	*WNT5B*	3.22	*ARHGEF10*	2.51

TPM1	2.48	*GEM*	2.06	*VCL*	1.88
BACE1	2.47	*COLGALT1*	2.06	*PPP2CB*	1.87
SLC37A2	2.45	*XYLT1*	2.05	*HTRA1*	1.87
GDF6	2.43	*NOTCH2*	2.05	*RAB32*	1.86
HSPB6	2.42	*COPZ2*	2.05	*MTMR9*	1.86
RARB	2.41	*MYOF*	2.04	*FBLN5*	1.86
FLNA	2.40	*WIPF1*	2.03	*SMAD9*	1.85
CDC42EP3	2.37	*KIRREL*	2.03	*WLS*	1.85
ROR1	2.37	*PLEC*	2.03	*PXK*	1.85
BDKRB1	2.36	*MPRIP*	2.03	*FMNL3*	1.85
FAT4	2.35	*ALCAM*	2.02	*CHSY1*	1.84
ANTXR2	2.35	*DNAJB4*	2.02	*DZIP1*	1.83
THBD	2.34	*AGPAT5*	2.02	*UGDH*	1.83
LAMC1	2.31	*TIMP2*	2.02	*RECK*	1.82
ABCA1	2.29	*ADARB1*	2.01	*GNG2*	1.82
LEPREL1	2.28	*POLR3D*	2.01	*IGFBP7*	1.82
HBEGF	2.28	*PRICKLE2*	2.01	*PLOD2*	1.81
DACT1	2.27	*NCOR2*	2.00	*NFATC4*	1.81
PTPRM	2.27	*CNN2*	1.98	*PALD1*	1.80
MFAP2	2.26	*ADAM19*	1.98	*P4HA3*	1.80
EVC	2.26	*STX2*	1.97	*KCNMB1*	1.80
GPX3	2.25	*IL11*	1.97	*ATP10A*	1.80
SAMD4A	2.25	*TENC1*	1.96	*GRB10*	1.80
TRABD2A	2.21	*TSPAN5*	1.96	*LGALS1*	1.79
FRMD6	2.19	*SEC23A*	1.94	*NPR3*	1.79
RELN	2.19	*HECW2*	1.93	*CYTH3*	1.79
CLIP3	2.18	*ITGA1*	1.93	*NFIC*	1.79
FAT1	2.17	*VGLL2*	1.93	*NLGN2*	1.79
MOSPD1	2.17	*ZBTB47*	1.93	*ADAMTS14*	1.79
SLC4A7	2.17	*PKD1*	1.93	*MYO9B*	1.78
AP1S2	2.15	*KIF26B*	1.92	*CYBRD1*	1.78
TCF4	2.14	*ZFYVE28*	1.92	*FAM43A*	1.77
TLE4	2.14	*PEAK1*	1.92	*KRT75*	1.77
SCN9A	2.12	*CTHRC1*	1.92	*PPP3CC*	1.77
EREG	2.12	*PDLIM2*	1.91	*PDLIM4*	1.77
RAB23	2.12	*TLL1*	1.91	*RAB34*	1.77
FGFRL1	2.11	*WDR1*	1.91	*MAK16*	1.76
GNAI2	2.10	*LTBP1*	1.90	*ILK*	1.76
SLC14A1	2.10	*MAP7D1*	1.90	*FOXC2*	1.76
NCEH1	2.09	*RBPMS*	1.90	*CD82*	1.75
EDNRA	2.09	*GPC1*	1.90	*MSANTD3-TM*	1.75
LCP1	2.08	*VASN*	1.89	*DEPTOR*	1.75
CD151	2.07	*LRP12*	1.88	*DSEL*	1.75
				GTF2E2	1.75

Extended Data Table 3. Gene symbols and names

Gene Symbol	Gene Name
Acsm2	Acyl-CoA Synthetase Medium Chain Family Member 2
Acta2	Actin alpha 2, smooth muscle
Aldh1a3	Aldehyde Dehydrogenase 1 Family Member A3
Aldob	Aldolase, fructose-biphosphate B
Anxa1	Annexin A1
Apoe	Apolipoprotein E
Aqp2	Aquaporin 2
Arid5b	AT-Rich Interaction Domain 5B
Ass1	Argininosuccinate synthase 1
Atp6v0d2	ATPase H+ transporting V0 subunit D2
Atp6v1g3	ATPase H+ transporting V1 subunit G3
Axin2	Axis inhibition protein 2
Bcl6b	B-Cell CLL/Lymphoma 6, Member B
Bgn	Biglycan
Bmp4	Bone morphogenetic protein 4
Bst1	Bone marrow stromal cell antigen 1
C3	Complement 3
C1qa	Complement C1q A chain
C1qb	Complement C1q B chain
Cbx3	Chromobox Homolog 3
Ccl2	C-C motif chemokine ligand 2
Ccl5	C-C motif chemokine ligand 5
Ccr2	C-C motif chemokine receptor 2
Ccr9	C-C motif chemokine receptor 9
Ccrl2	C-C motif chemokine receptor like 2
Cd3g	T-cell surface glycoprotein CD3 gamma chain
Cd5	Cluster of differentiation 5
Cd7	T-cell antigen CD7
Cd14	Cluster of differentiation 14
Cd34	Hematopoietic Progenitor Cell Antigen CD34
Cd52	CAMPATH-1 antigen
Cd55	Cluster of differentiation 55
Cd74	H-2 class II histocompatibility antigen gamma chain

Cd79a	Cd79a molecule, immunoglobulin-associated alpha
Cd79b	Cd79a molecule, immunoglobulin-associated beta
Cd209a	Cluster of differentiation 209a
Cd300c	Cluster of differentiation 300c
Cdh1	E-cadherin
Cdh2	N-cadherin
Cdh5	VE-cadherin
Cdh11	Cadherin-11
Cebpd	CCAAT Enhancer Binding Protein Delata
Chil3	Chitinase-like 3
Ciita	Class II major histocompatibility complex transactivator
Cited1	Cbp/P300-Interacting Transactivator 1
Cldn3	Claudin 3
Cldn4	Claudin 4
Cldn7	Claudin 7
Cnn2	Calponin
Col1a1	Collagen Type I Alpha 1 Chain
Col1a2	Collagen Type I Alpha 2 Chain
Col3a1	Collagen Type III Alpha 1 Chain
Col4a1	Collagen Type IV Alpha 1 Chain
Csn3	Casein Kappa
Csf3r	Colony Stimulating Factor 3
Ctsw	Cathepsin W
Cxcl2	C-X-C Motif Chemokine Ligand 2
Cxcl16	C-X-C Motif Chemokine Ligand 16
Cxcr2	C-X-C Motif Chemokine Receptor 2
Cx3cr1	C-X3-C Motif Chemokine Receptor 1
Dcn	Decorin
Dlx1	Distal-Less Homeobox 1
Dlx2	Distal-Less Homeobox *2*
Dlx3	Distal-Less Homeobox 3
Egfl7	Epidermal growth factor-like protein
Egr1	Early Growth Response 1
Ehd3	EH Domain Containing 3
Ehf	ETS homologous factor

Elf5	E74-like factor 5
Elk3	ETS Transcription Factor ELK3
Elovl7	Elovl fatty acid elongase 7
Emilin2	Elastin Microfibril Interfacer 2
Epcam	Epithelial cell adhesion molecule
Ephb1	Ephrin type-B receptor 1
Ets1	ETS Proto-Oncogene 1, Transcription Factor
Ets2	ETS Proto-Oncogene 2, Transcription Factor
Fbln5	Fibulin 5
Flt1	VEGF receptor 1
Fn1	Fibronectin
Fos	Fos proto-oncogene, AP-1 transcription factor subunit
Fosb	FosB proto-oncogene, AP-1 transcription factor subunit
Foxa1	Forkhead Box A1
Fxyd4	Fxyd domain containing ion transport regulator 4
Gata2	GATA Binding Protein 2
Gli3	GLI Family Zinc Finger 3
Gsn	Gelsolin
H2-Aa	H-2 class II histocompatibility antigen, A-B alpha chain
H2-ab1	H-2 class II histocompatibility antigen, A beta chain
Hnf4a	Hepatocyte nuclear factor 4 alpha
Hp	Haptoglobin
Icam	Intercellular adhesion molecule 1
Id1	Inhibitor of DNA binding 1, HLH protein
Id3	Inhibitor of DNA binding 3, HLH protein
Ifitm2	Interferon induced transmembrane protein 2
Ifitm3	Interferon induced transmembrane protein 3
Igkc	Immunoglobulin kappa constant
Il1b	Interleukin 1 beta
Ikzf1	IKAROS Family Zinc Finger 1
Ikzf2	IKAROS Family Zinc Finger 2
Il6st	Interleukin 6 cytokine family signal transducer
Inhba	Inhibin subunit beta a
Irf1	Interferon regulatory protein 1
Irf7	Interferon regulatory protein 7

Irf8	Interferon regulatory protein 8
Jun	Jun proto-oncogene, AP-1 transcription factor subunit
Junb	JunB proto-oncogene, AP-1 transcription factor subunit
Jund	JunD proto-oncogene, AP-1 transcription factor subunit
Keg1	kidney-expressed gene 1
Klf4	Kruppel like factor 4
Klf15	Kruppel like factor 15
Krt5	Keratin 5
Krt7	Keratin 7
Krt14	Keratin 14
Krt15	Keratin 15
Krt18	Keratin 18
Lalba	Alpha-Lactalbumin
Lcn2	Lipocalin 2
Lox	Lysyl oxidase
LoxL1	Lysyl oxidase like 1
LoxL2	Lysyl oxidase like 2
Lrp2	LDL Receptor Related protein 2
Ltf	Lactotransferin
Ly6d	Lymphocyte antigen 6 family member D
Ly6e	Lymphocyte antigen 6 family member E
Maf	MAF BZIP transcription factor
Mafb	MAF BZIP transcription factor B
Meis2	Meis Homeobox
Meox1	Mesenchyme Homeobox 1
Mfap2	Microfibrillar-associated protein 2
Mfap5	Microfibril associated protein 5
Miox	Myo-inositol oxygenase
Mki67	Marker of proliferation Ki-67
Mmp2	Metallopeptidases 2
Mmp3	Metallopeptidases 3
Mmp13	Metallopeptidases 14
Mmp14	Metallopeptidases 14
Ms4a4b	Membrane spanning 4-Domains Subfamily A Member 4b

Ms4a4c	Membrane spanning 4-Domains Subfamily A Member 4c
Ms4a6c	Membrane spanning 4-Domains Subfamily A Member 6c
Ms4a7	Membrane spanning 4-Domains A7
Msx1	Msh Homeobox 1
Msx2	Msh Homeobox 2
Mylk	Myosin Light Chain Kinase
Nav2	Neuron Navigator
Ncam1	Neural Cell Adhesion Molecule 1
Ndrg1	N-Myc Downstream Regulated 1
Neurod1	Neuronal Differentiation 1
Neurod4	Neuronal Differentiation 4
Nfib	Nuclear Factor I B
Nfkbia	NFKB Inhibitor Alpha
Nkg7	Natural killer cell protein 7
Notch1	Notch homolog 1
Notch2	Notch homolog 2
Notch3	Notch homolog 3
Nox4	NADPH oxidase 4
Nphs1	Nphs1 adhesion molecule, Nephrin
Nphs2	Nphs2 stomatin family member, Podocin
Nr2f1	Nuclear Receptor Subfamily 2 Group F Member 1
Nr2f2	Nuclear Receptor Subfamily 2 Group F Member 1
Nrp1	Neuropilin2
Nrp2	Neuropilin 2
Olig3	Oligodendrocyte transcription factor 3
Osm	Oncostatin M
Palld	Palladin
Pax3	Paired Box 3
Pax5	Paired Box 5
Pcalf	Pcna clamp associated factor
Pdpn	Podoplanin
Pdgfra	Platlet derived growth factor receptor alpha
Pdgfrb	Platlet derived growth factor receptor beta
Pecam	Platelet and endothelial cell adhesion molecule 1
Phgdh	Phosphoglycerate Dehydrogenase

Plp1	Proteolipid protein 1
Plet1	Placenta Expressed Transcript 1
Plvap	Plasmalema Vesicle Associated protein
Postn	Periostin
Ppp1r1a	Protein Phosphatase 1 Regulatory Inhibitor Subunit 1A
Prrx1	Paired Related Homeobox 1
Rgs5	Regulator of G protein signalling 5
Robo1	Roundabout Guidance receptor 1
Rora	RAR Related Orphan Receptor A
Runx1	RUNX Family Transcription Factor 1
S100a4	S100 calcium binding protein A4
S100a8	S100 calcium binding protein A8
S100a9	S100 calcium binding protein A9
Saa1	Serum amyloid a1
Saa3	Serum amyloid a1
Scgb1b27	Secretoglobulin family 1B member 27
Scgb2b27	Secretoglobulin family 2B member 27
Scn7a	Sodium voltage-gated channel alpha subunit 7
Sdc4	Syndecan 4
Siglech	Sialic acid-binding ig-loke lectin H
Slc5a3	Solute carrier family 5 member 3
Slc12a1	Solute carrier family 12 member 1
Slc12a3	Solute carrier family 12 member 3
Slc34a4	Solute carrier family 34 member 1
Slc34a1	Solute carrier family 4 member 4
Snail1	Zinc finger protein SNAI1
Snail2	Zinc finger protein SNAI2
Sostdc1	Sclerostin Domain Containing 1
Sox2	SRY-Box Transcription Factor 2
Sox9	SRY-Box Transcription Factor 9
Sox10	SRY-Box Transcription Factor 10
Sox11	SRY-box Transcription Factor 11
Sox18	SRY-box Transcription Factor 18
Sparc	Secreted protein acidic and cysteine rich
Spi1	Spi-1 Proto-Oncogene

Spp1	Secreted phosphoprotein 1
Srebf1	Sterol Regulatory Element Binding Transcription factor 1
Tagln	Transgelin
Tcf4	Transcription Factor 4
Tgfb1	Transformation growth factor beta 1
Tgfb2	Transformation growth factor beta 2
Tgfb3	Transformation growth factor beta 3
Tgfbr2	Transformation growth factor beta receptor 2
Timp1	Tissue Inhibitor metallopeptidase inhibitors 1
Timp2	Tissue Inhibitor metallopeptidase inhibitors 2
Timp3	Tissue Inhibitor metallopeptidase inhibitors 3
Tmem52b	Transmembrane Protein 52 beta
Tmem72	Transmembrane Protein 72
Tnc	Tenascin C
Tnf	Tumor necrosis factor
Tnfrsf11b	TNF receptor superfamily member 11b
Tnip3	Tnfaip3 interacting protein 3
Top2a	DNA topoisomerase II alpha
Tpm1	Tropomyosin 1
Tpm2	Tropomyosin 2
Trac	T Cell Receptor Alpha Constant
Tnip3	TNFAIP3 Interacting protein 3
Trp63	Tumor protein 63
Twist1	Twist Family BHLH Transcription Factor 1
Vcam1	vascular cell adhesion molecule 1
Vcan	Versican
Vim	Vimentin
Wap	Whey acid protein
Wfdc18	WAP four-disulfide core domain protein 18
Wnt3a	Wnt Family Member 3A
Xbp1	X-Box Binding Protein 1
Xcl1	X-C Motif Chemokine Ligand 1
Zeb1	Zinc finger E-box binding homeobox 1
Zeb2	Zinc finger E-box binding homeobox 2
Zic3	Zic Family Member 3

Extended Data Table 4. Antibodies

Primary antibodies

IF (TNBC sample): CYTOKERATIN AE1/AE3 (CKs)	Dako GA053	Ready to use
IF : JUN	Cell Signaling mAB#9165	1:500 (Human sample)/1:1000 (Mouse sample)
IF : PRRX1	Kind gift from Prof. TANAKA (IMP Vienna BioCenter)	1 :100
IF: CDH1	BD Biosciences 610181	1:100 (cells) / 1:500 (tissue)
IF: α-SMA	SIGMA A2547	1 :2000
IF: ZO-1(TJP-1)	Life Technologies 339100	1:100
IF: FN1	Abcam ab2413	1:200
IF : F-actin	Phalloidin-TRITC	1:200
IF: LTA Biotinylated	Vector Lab B-1325	1:500
IF: tdTomato (RFP)	Fisher scientific RF5R	1:500
IF:Vimentin (Rabbit)	Abcam ab92547	1:500
IF:Vimentin (Goat)	SantaCruz SC7557	1:500
IHC:Vimentin (Goat)	SantaCruz SC7557	1:200
IHC: Cadherin 16	Abcam ab212243	1:200
IF (iDISCO+): tdTomato (RFP)	ROCKLAND 600-401-379	1:2000
IF: PyMT	Life Technologies MA146061	1:500
IF: Trp63	GenTex GTX102425	1:1000
IF:KLF4	Abcam ab214666	1 :200
IF:KERATIN20	Pro-gen GP-K20	1:500
IF:KIM-1	R&D AF1817-SP	1:200
IF:KERATIN14	Palex 454928	1:5000
IF:F4/80	BioRad clone Cl:A3-1	1:200
IF:Cd163	ab182422	1:200
IF:MHC-II	ThermoFisher Clone M5/114.15.2	1:200
WB: SMAD2/3	Cell Signaling mAB#8685	1:1000
WB: pSMAD2(Ser465/467) /pSMAD3(Ser423/425)	Cell Signaling mAB#8828	1:1000
WB: CDH1	BD 610181	1:500
WB: ZO-1(TJP-1)	Life Technologies 339100	1 :500

WB: FN1	Abcam ab2413	1:500
WB: β -Actin	Abcam/ab8227	1:10,000

Secondary antibodies

IF: Goat anti-Rabbit IgG(H+L) Alexa Fluor 488	Invitrogen A-11008	1:500
IF: Goat anti-Mouse IgG(H+L) Alexa Fluor 488	Invitrogen A-11001	1:500
IF: Goat anti-Rat IgG(H+L) Alexa Fluor 488	Invitrogen A-11006	1:500
IF: Goat anti-Rabbit IgG(H+L) Alexa Fluor 568	Invitrogen A-11011	1:500
IF: Goat anti-Mouse IgG(H+L) Alexa Fluor 568	Invitrogen A-11004	1:500
IF: Goat anti-Rat IgG(H+L) Alexa Fluor 568	Invitrogen A-11077	1:500
IF: Goat anti-Chicken IgG (H+L) Alexa Fluor 568	Invitrogen A-11041	1:500
IF: Donkey anti-Goat IgG (H+L) Alexa Fluor 594	Invitrogen A-11058	1:500
IF: Goat anti-Guinea Pig IgG (H+L) Alexa Fluor 647	Invitrogen A-21450	1:500
IF: Donkey anti-Rabbit IgG (H+L) Alexa Fluor 647	Invitrogen A-31573	1:500
IF: Goat anti-Chicken IgG (H+L) Alexa Fluor 647	Invitrogen A-32933	1:500
IF: Goat anti-Rabbit IgG (H+L) Alexa Fluor 647	Invitrogen A-21244	1:500
IF: Goat anti-Mouse IgG(H+L) Alexa Fluor 647	Invitrogen A-21235	1:500
IF: Goat anti-Rat IgG(H+L) Alexa Fluor 647	Invitrogen A-21247	1:500
IF: Streptavidin Alexa Fluor 488 Conjugate	Invitrogen S11223	1:1000
IF: Streptavidin Alexa Fluor 568 Conjugate	Invitrogen S11226	1:1000
IF: Streptavidin Alexa Fluor 647 Conjugate	Invitrogen S32357	1:1000
WB: Goat Anti Rabbit IgG Peroxidase Conjugate	SIGMA A0545	1:3000
WB: Goat anti-Mouse IgG(H+L) HRP Conjugate	Bio-Rad #1706516	1:3000

(IF, immunofluorescence; IHC, immunohistochemistry WB, western blot)

Extended Data Table 5. Oligonucleotides

Primers for qRT-PCR

	FORWARD (5' to 3')	REVERSE
Cfa-RS17	CAAGATCGCAGGCTATGTGA	CCTCGATGATCTCCTGATCC
Cfa-SNAIL1	AACTGCAAATACTGCAACAAGGAATAC	CAGGAAAACGGCTTCTCACCG
Cfa-SNAIL2	CGTTTTCCAGACCCTGGTTA	TGACCTGTCTGCAAATGCTC
Cfa-ZEB1	TGGTCATGATGACAGTGGAACA	GTTCTGTACGCAAAGGTGTAACTG
Cfa-ZEB2	ATGACACTATGGGGCCTGAA	ATCGCGTTCCTCCAGTTTTC
Cfa-TWIST1	GCCGGAGACCTAGATGTCATT	CACGCCCTGTTTCTTTGAAT
Cfa-PRRX1	CGCGTCTTTGAGAGAACACAC	GCATGGCTCTCTCATTCCTTC
Cfa-CDH1	CACCTCCTGTTGGTGTGTTTATTA	CGATCTCCATTGGGTCTTCAAC
Cfa-EPCAM	CAGAAAGCTCAGAATGATGTGG	GCATTGAAAATTCAGGTGGTTT
Cfa-GRHL2	CCATAGCACCTACCTCAAAGATG	GGCTTCCAGAGTGTACTGAAATG
Cfa-OVOL1	GAGATCTACGTGCCAGTCAGC	TGAGGGTTCAGCCACTGACG
Cfa-ESRP1	GCATTGCAGAGGCACAAACA	TCCTTGGAGAGAAACTGGGC
Cfa-FN1	GTTACCGTGGGCAACTCTGT	CAATGGCATAATGGGAAACC
Cfa-ITGA5	GGGAGGACTGCAGAGAGATG	GGGTCCAGGGAGAAGTTGAG
Cfa-ACTA2	CCCAGACATCAGGGAGTGAT	TCGGGTACTTCAAGGTCAGG
Cfa-CDH2	GTAGAGGCTTCGGGTGAAATC	ATACACCATGCCGTCTTCATC
Cfa-TAGLN	ACCCCAACTGGTTTATGAAGAAAG	CTGTTGCTGCCCATCTGTAG
Cfa-TNC	CCCGGGCAAGAGTATGAGATC	CAGTGGTGTCCGTGACATCT
Mm-TBP	CCTTGTACCCTTCACCAATGAC	ACAGCCAAGATTCACGGTAGA
Mm-Snail1	CAGCTGCTTCGAGCCATAGA	TGAGGGAGGTAGGGAAGTGG
Mm- Snail2	ACACATTAGAACTCACACTGGGG	AGCAGTTTTTGCACTGGTATTTCT
Mm- Twist1	CAGGCCGGAGACCTAGATG	AGATTTATTTTTAGTTATCCAGCTCCAG
Mm-Zeb1	CTGCTCCCTGTGCAGTTACA	GTGCACTTGAACTTGCGGTT
Mm- Zeb2	AATCCCAGGAGGAAAAACGTGG	AAGGCTATCATCTTCAGCAATGT
Mm- Prrx1	TCAGCAGGACAATGACCAGT	TGCGAGATCTTCTCGAACAA
Mm- Hnf4	GAGGAGCGTGAGGAAGAACC	CCGGAAGCACTTCTTAAGCC
Mm-Cdh1	Grande et al., 2015	Grande et al., 2015
Mm-Cdh16	Grande et al., 2015	Grande et al., 2015
Mm- Fos	TACTACCATTCCCCAGCCGA	GCTGTCACCGTGGGGATAAA
Mm- Jun	TTCCTCCAGTCCGAGAGCG	TGAGAAGGTCCGAGTTCTTGG
Mm- Irf1	ACCCTGGCTAGAGATGCAGA	CGGAACAGACAGGCATCCTT
Mm- Runx1	CGGCCATGAAGAACCAGGTA	TGGTAGGTGGCAACTTGTGG
Mm- Plet1	TCTCTGGTGGGGATGTAACCT	CTGCTTTCAGGATCACGGCA
Mm- Ltf	GTCAAGAAATCCTCCACCCG	CGAACATAGTGCCACCATCA
Mm- Sdc4	CCCAGGGCAGCAACATCTTT	CAGCAGCAGGATCAGGAAAAC
Mm- Anxa1	CGGAAAGCCTTGCTTGCTCTT	TCTCCAGCTTCATACAAAGCCC
Mm- Krt7	CGCTCTATCCAGAGGCTGC	CTTGATTGCCAGCTCCCCTT
Mm- Krt15	ATGGAAGAGATCCGGGACAAAA	GCAGGGTCAGCTCATTCTCATAC

Mm- Nav2	CCTCATCAAGGACCTCCAGC	GCTGCCAGGAAATTCAAGCA
Mm- Meis2	GAACTTTCAGGCTCCTCCACA	GGTTGCGTCATCGTGGTCTC
Mm- Postn	TTCGTGGCAGCACCTTCAAA	TTTTGGTTCCACGACTTTGGT
Mm- Bgn	CTGGCCTCCCAGATCTCAAG	ATGCCCACCTTGGTGATGTT
Mm- Klf4	GGGAGAAGACACTGCGTCC	TGGGGGAAGTCGCTTCATGT
Mm- Junb	AGGCAGCTACTTTTCGGGTC	TTGCTGTTGGGGACGATCAA
Mm- Mafb	CTTCTCCCAGCTTCAGTCCG	GTAGTTGCTCGCCATCCAGT
Mm-Mef2c	TGAGCGTAACAGACAGGTGAC	ACAGCACGCTCAGCTCATAA
Mm-Egr1	GCCGAGCGAACAACCCTAT	TCGTTTGGCTGGGATAACTCG
Mm-Saa1	TTCACGAGGCTTTCCAAGGG	CCTGAAAGGCCTCTCTTCCA
Mm-Saa2	GAGTCTGGGCTGCTGAGAAA	CATGGTGTCCTCGTGTCCTC
Hs-MAFB	ACGTGAAGAAGGAGCCACTG	TGTGTCTTCTGTTCGGTCGG
Hs-SNAIL1	GACCCCAATCGGAAGCCTAA	AGGGCTGCTGGAAGGTAAAC
Hs-EGR1	ACCTGACCGCAGAGTCTTTT	GTTTGGCTGGGGTAACTGGT
Hs-SAA1	CTGCCAAAAGGGGACCTGG	CCGCACCATGGCCAAAGAAT
Hs-SAA2	GCTCCTTGGTCCTGAGTGTC	GTCCCGAGCCCCATCAAAA
Hs-JUND	ATCGACATGGACACGCAGG	CCAGCTCCGTGTTCTGACTC

Primers for screening and genotyping

	FORWARD (5' to 3')	REVERSE
Prrx1^{em1An}–S1	CGTCTTTGGTGAAATGCAGA	GGCCAAGAATCCTCCTCAGT
Prrx1^{em1An}–S2	TGGAGCTCGTACTGATGTGG	CCTTCCCTTCCCTACAGCAT
Prrx1^{em1An}–S3	TCCCAAACATAAATGTAGAGCA	AAGGAGATAGCCTTCCCTTCC
Prrx1^{em1An}– genotyping	TTTCTTTCCCCGTCTTTGGT	AAGGAGATAGCCTTCCCTTCC

Cfa, Canis familiaris

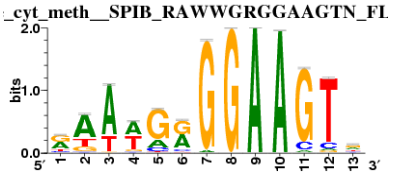
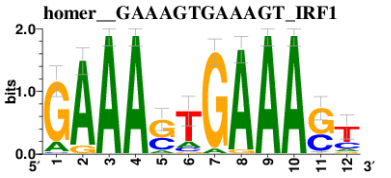
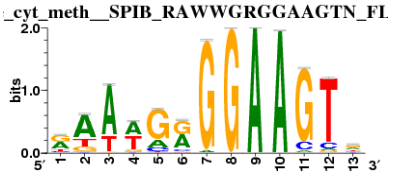
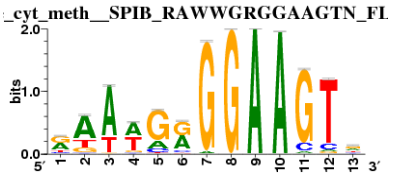
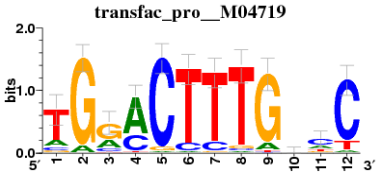
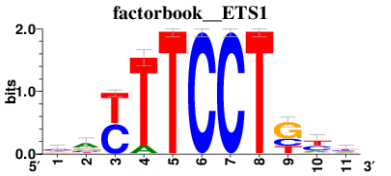
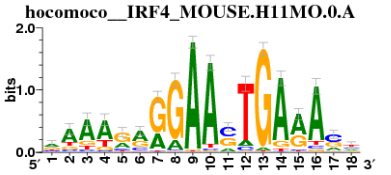
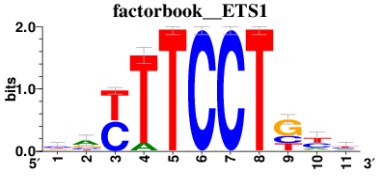
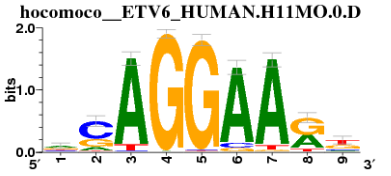
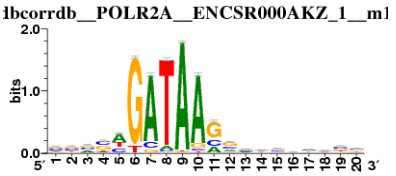
Mm, Mus musculus

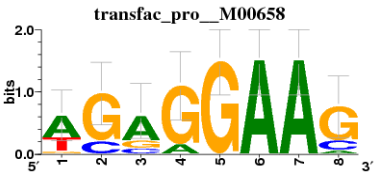
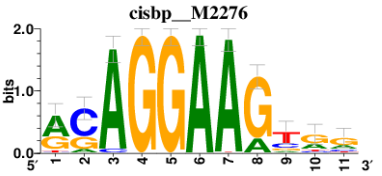
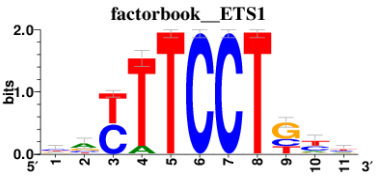
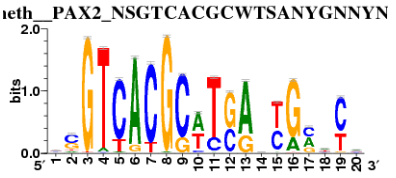
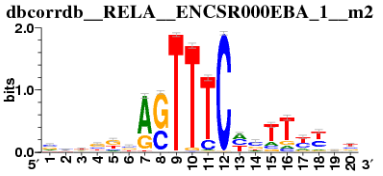
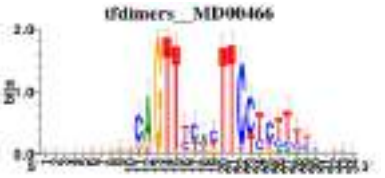
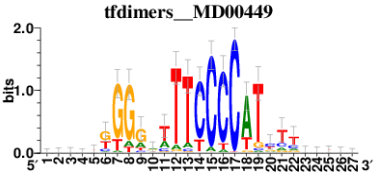
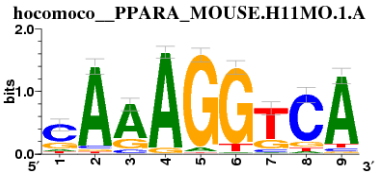
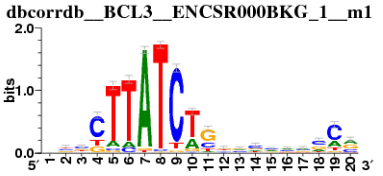
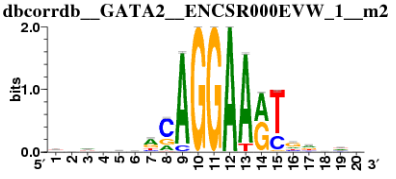
Hs, Homo sapiens

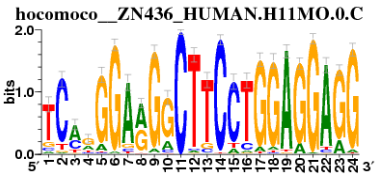
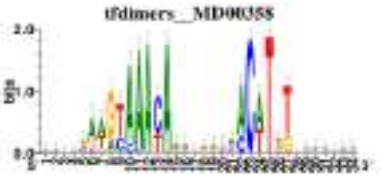
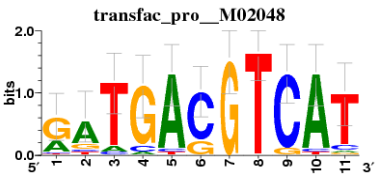
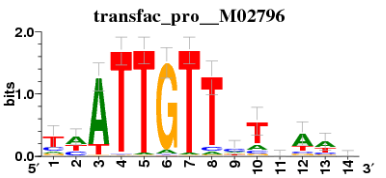
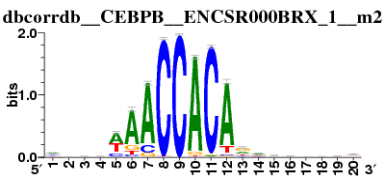
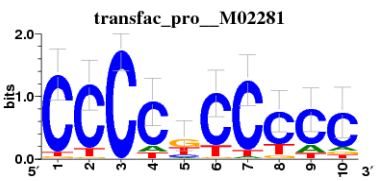
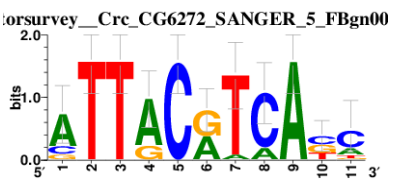
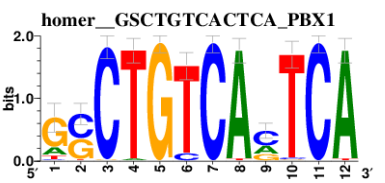
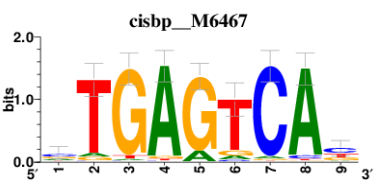
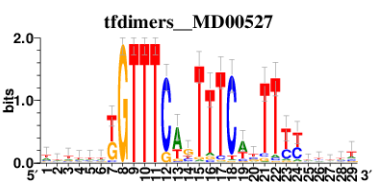
Extended Data Table 6, workflow in kidney and cancer single-cell analyses

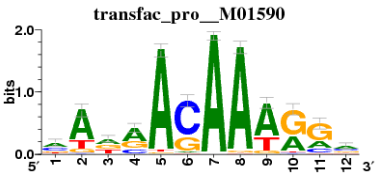
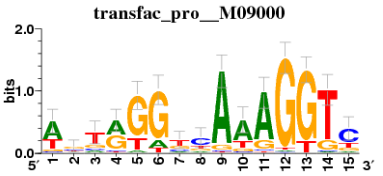
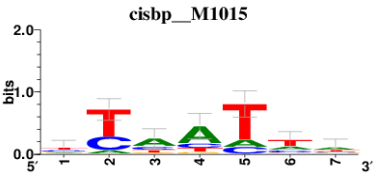
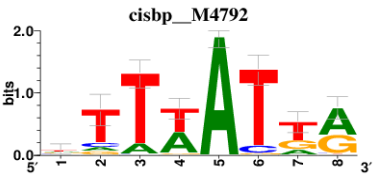
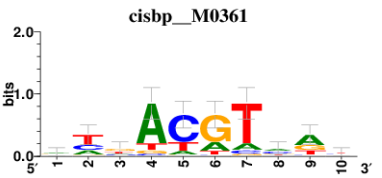
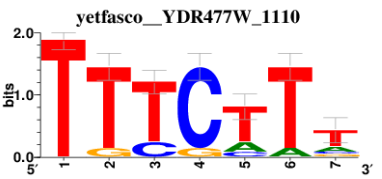
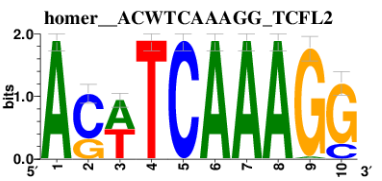
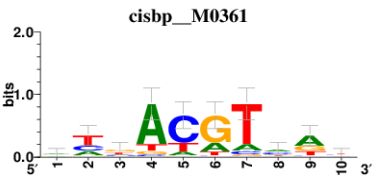
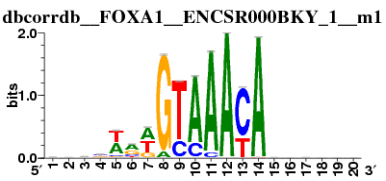
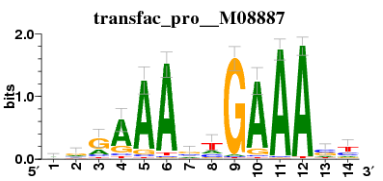
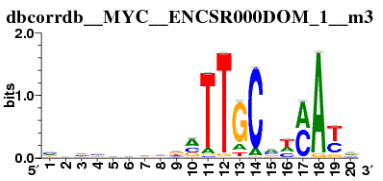
	Intersititial renal fibrosis	Metastatic breast cancer
Step1 Quality control, filtering and integration	- Demultiplexing and alignment (Cellranger) - Removal of low quality cells and doublets	- Demultiplexing and alignment (Cellranger) - Removal of low quality cells and doublets
Step2 Dimensionality reduction, clusters detection and annotation	Identification of cell populations: - Epithelial cells - Endothelial and stromal - Immune cells	Identification of cell populations: - Cancer cells - Tumour microenvironment: CAFs, endothelial and immune cells
Step3 Subsetting cells of interest	Subsetting: - Epithelial injured cells - Epithelial cells at the origin of injury	Subsetting: - Proliferative and high ribosomial cells removal
Step4 EMT trajectory inference	- Connectivity map for epithelial and injured populations using PAGA - Reconstruction of EMT trajectory using Velocity - EMT trajectory transcriptional regulation (regulon) using SCENIC	- Connectivity map for epithelial and injured populations using PAGA - Reconstruction of EMT trajectories using Velocity - EMT trajectory transcriptional regulation (regulon) using SCENIC
Step5 EMT molecular programme	- Trajectory based differential gene expression (Moran's I test) - Gene ontology and pathways enrichment	- Trajectory based differential gene expression (Moran's I test) - Gene ontology and pathways enrichment

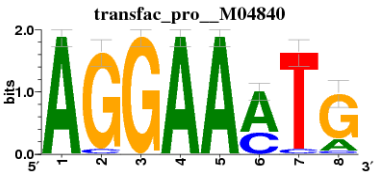
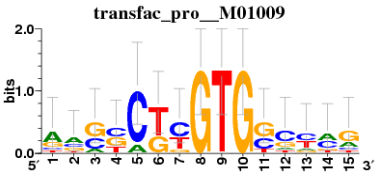
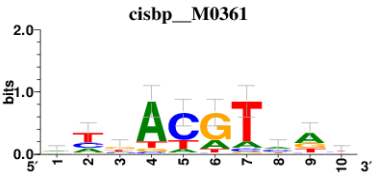
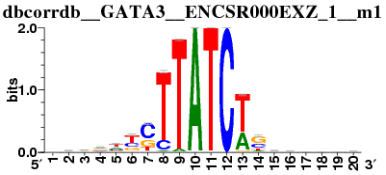
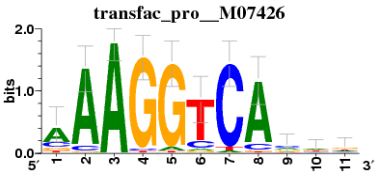
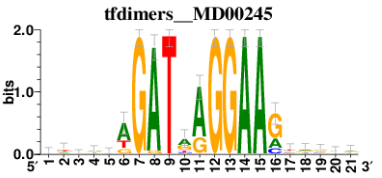
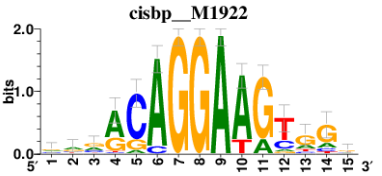
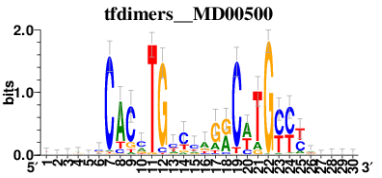
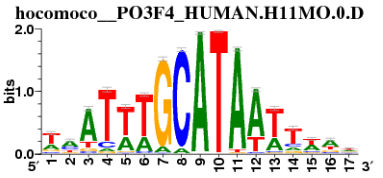
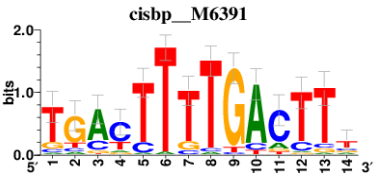
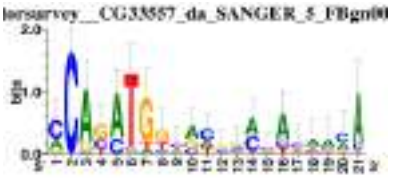
Extended Data Table 7, Predicted regulons and their binding motifs

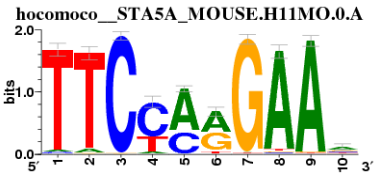
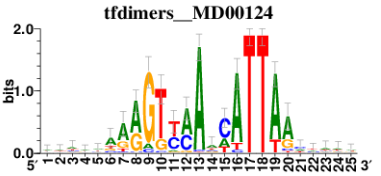
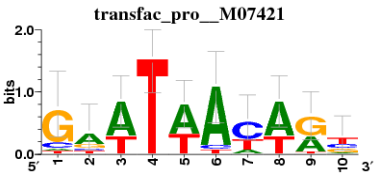
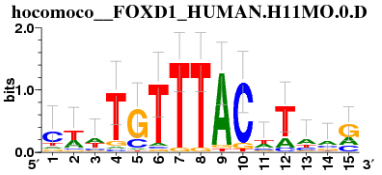
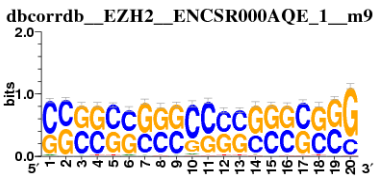
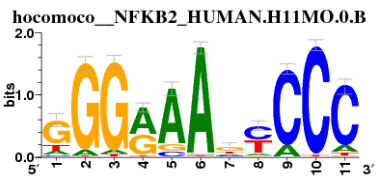
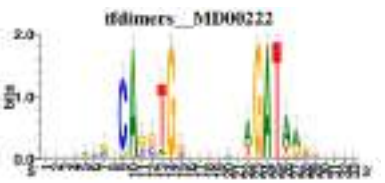
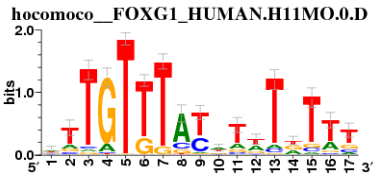
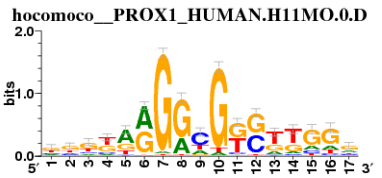
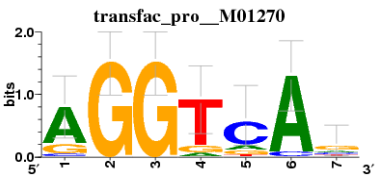
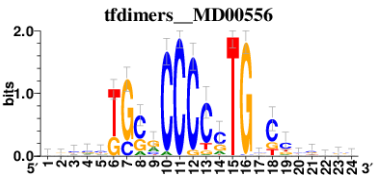
Datasets	TF Name	AUC	NES	Logo
Kidney	Spi1	0.214	10.572	
Kidney	Irf7	0.228	10.416	
Kidney	Irf8	0.266	10.085	
Kidney	Bcl11a	0.16	9.047	
Kidney	Hnf4a	0.138	8.589	
Kidney	Erg	0.208	7.884	
Kidney	Ikzf1	0.179	7.525	
Kidney	Ets1	0.115	7.139	
Kidney	Fli1	0.138	5.985	
Kidney	Gata3	0.125	5.849	

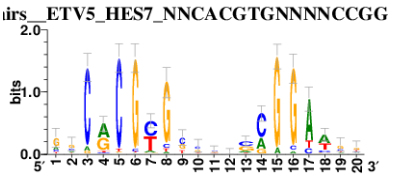
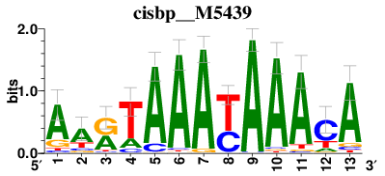
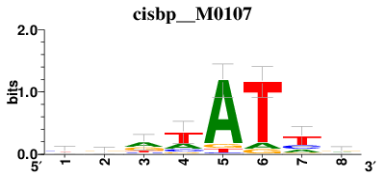
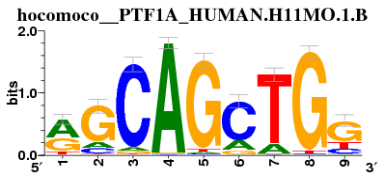
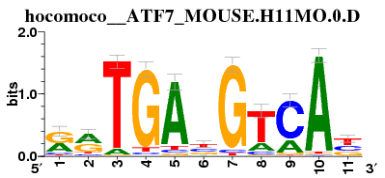
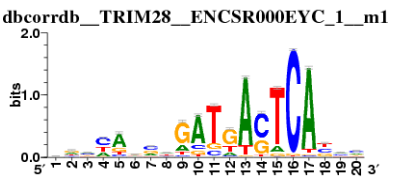
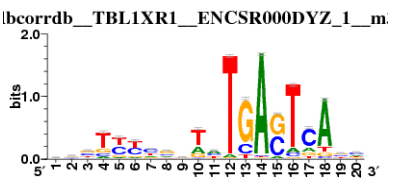
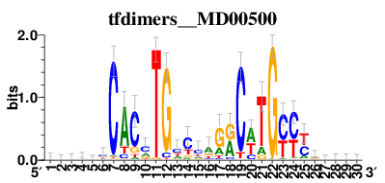
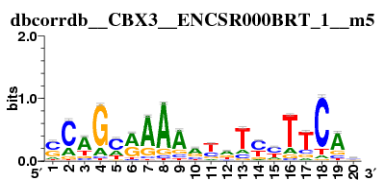
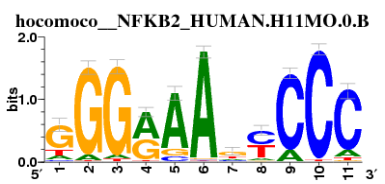
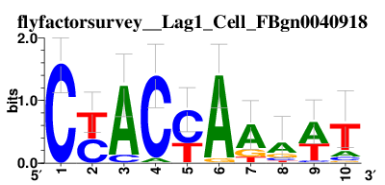
Kidney	Spib	0.175	5.773	
Kidney	Etv3	0.12	5.53	
Kidney	Ets2	0.136	5.523	
Kidney	Runx3	0.108	5.508	
Kidney	Pax5	0.161	5.413	
Kidney	Irf4	0.114	5.365	
Kidney	Myb	0.123	5.325	
Kidney	Rel	0.151	5.12	
Kidney	Rora	0.117	4.964	
Kidney	Gata2	0.144	4.926	
Kidney	Elk3	0.165	4.867	

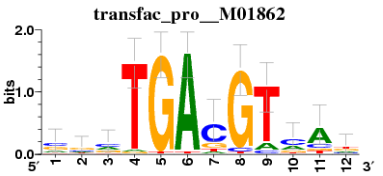
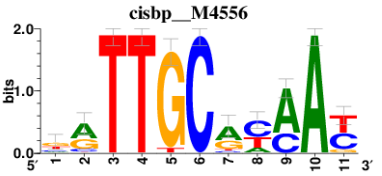
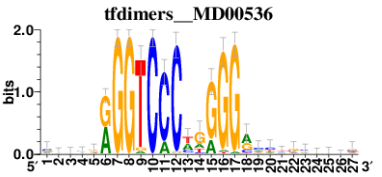
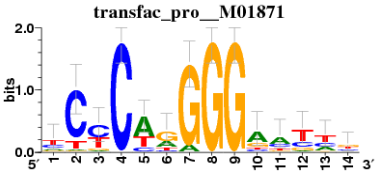
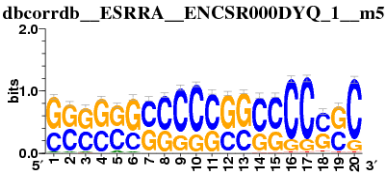
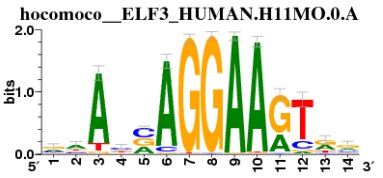
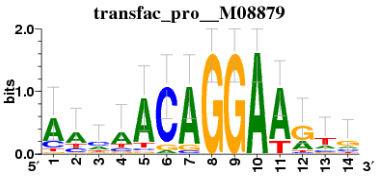
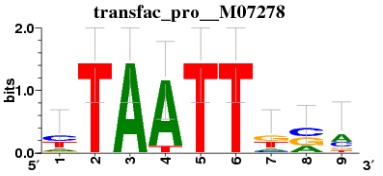
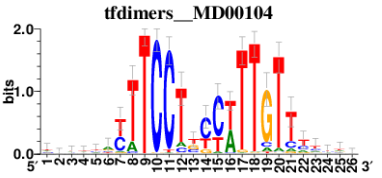
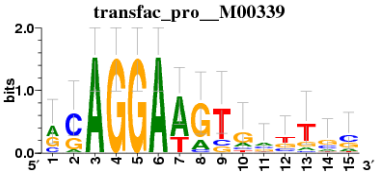
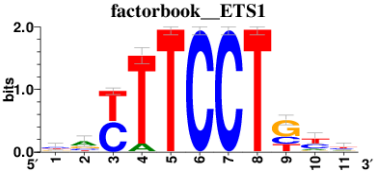
Kidney	Gm29609	0.121	4.598	
Kidney	Foxa1	0.151	4.472	
Kidney	Wt1	0.158	4.462	
Kidney	Fos	0.213	4.39	
Kidney	Sox18	0.102	4.318	
Kidney	Runx2	0.145	4.308	
Kidney	Irf1	0.151	4.222	
Kidney	Creb5	0.169	4.15	
Kidney	Pbx1	0.132	4.137	
Kidney	Mafb	0.086	4.129	
Kidney	Zeb1	0.159	4.044	

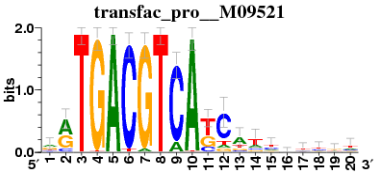
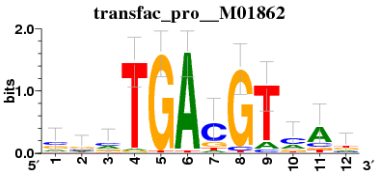
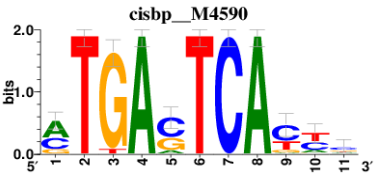
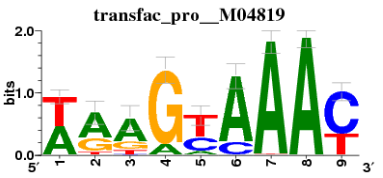
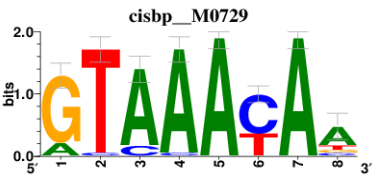
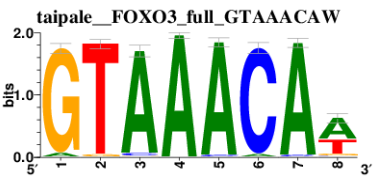
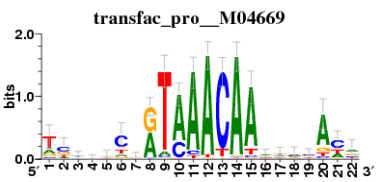
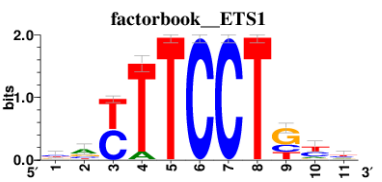
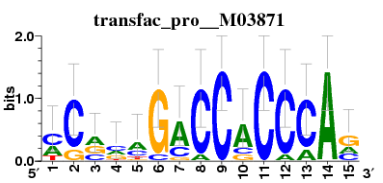
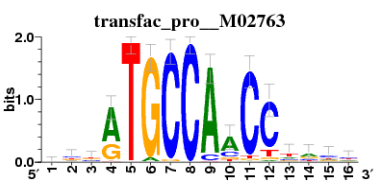
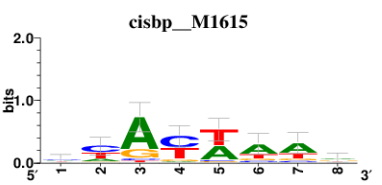
Kidney	Sox11	0.102	4.015	
Kidney	Rxrg	0.181	4.006	
Kidney	Lhx1	0.137	3.949	
Kidney	Hoxb9	0.118	3.909	
Kidney	Jun	0.137	3.893	
Kidney	Nuak1	0.109	3.866	
Kidney	Tcf4	0.158	3.852	
Kidney	Fosb	0.22	3.846	
Kidney	Foxp1	0.116	3.82	
Kidney	Prdm1	0.085	3.738	
Kidney	Cebpd	0.142	3.721	

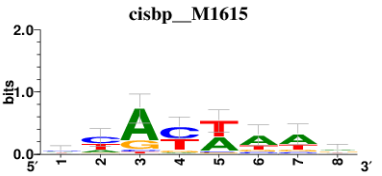
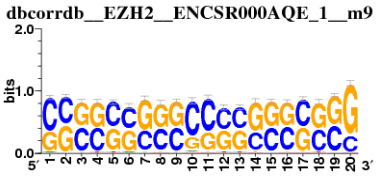
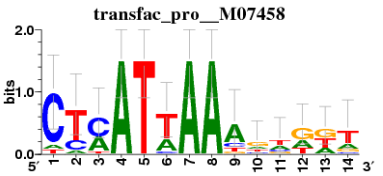
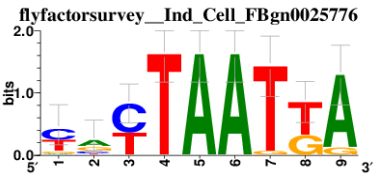
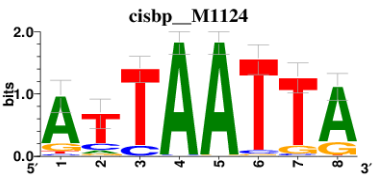
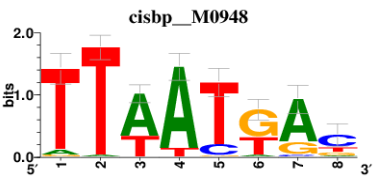
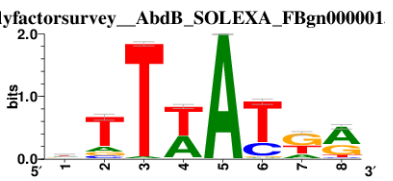
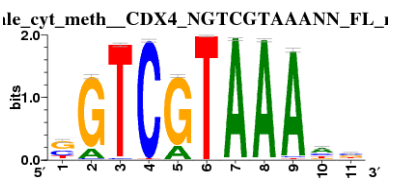
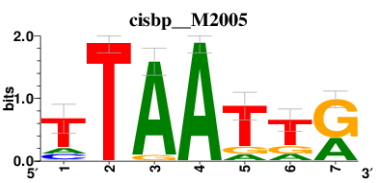
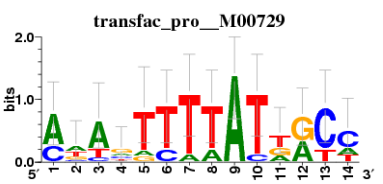
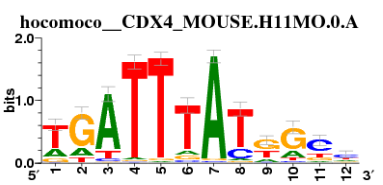
Kidney	Mef2c	0.116	3.649	
Kidney	Hes1	0.128	3.636	
Kidney	Junb	0.157	3.636	
Kidney	Gata6	0.116	3.525	
Kidney	Nr4a2	0.134	3.487	
Kidney	Gata5	0.089	3.421	
Kidney	Runx1	0.118	3.413	
Kidney	Bhlhe40	0.127	3.389	
Kidney	Pou3f3	0.124	3.329	
Kidney	Nr2e3	0.119	3.324	
Kidney	Scx	0.116	3.32	

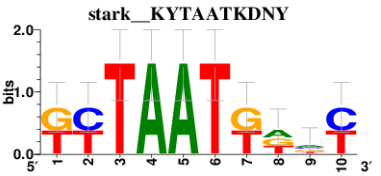
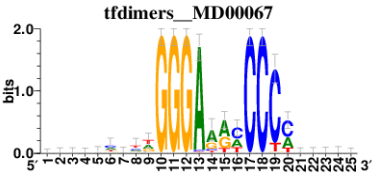
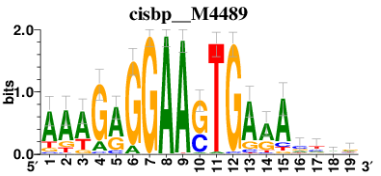
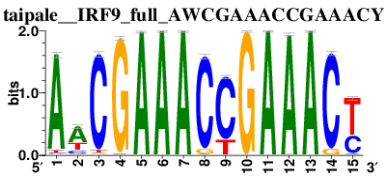
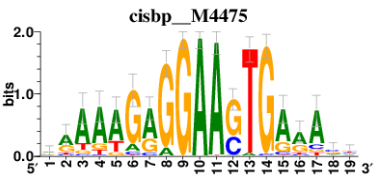
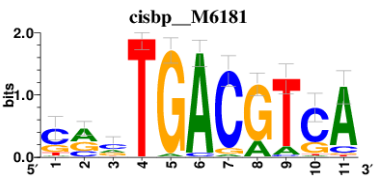
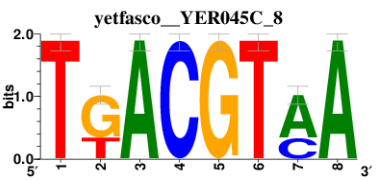
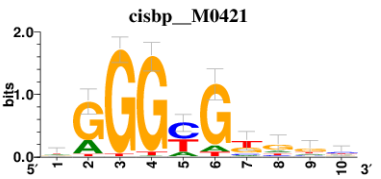
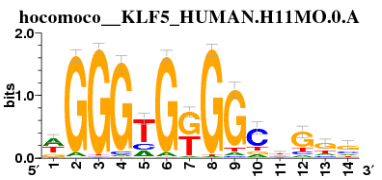
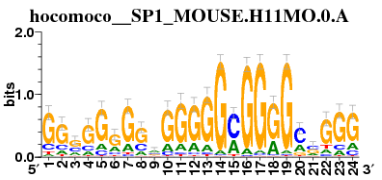
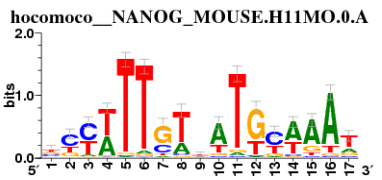
Kidney	Stat4	0.076	3.319	
Kidney	Nr2f2	0.089	3.283	
Kidney	Hoxc8	0.095	3.256	
Kidney	Foxc1	0.098	3.242	
Kidney	Jund	0.144	3.202	
Kidney	Cd59b	0.093	3.166	
Kidney	Mxd3	0.1	3.153	
Kidney	Foxp2	0.126	3.144	
Kidney	Prox1	0.13	3.042	
Kidney	Pparg	0.08	3.042	
Kidney	Tal2	0.083	3.038	

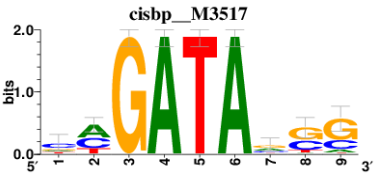
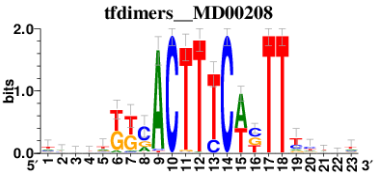
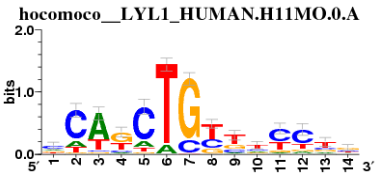
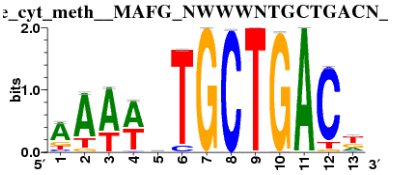
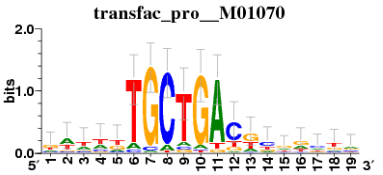
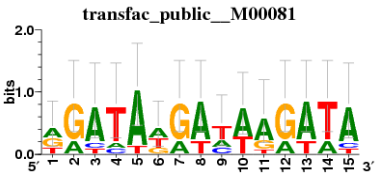
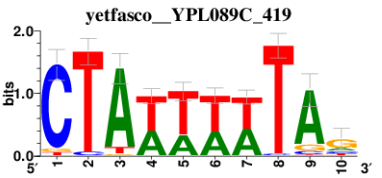
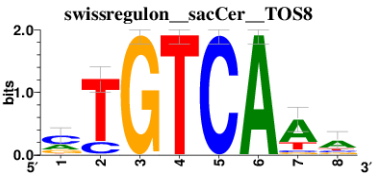
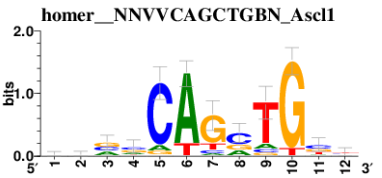
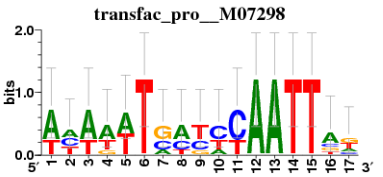
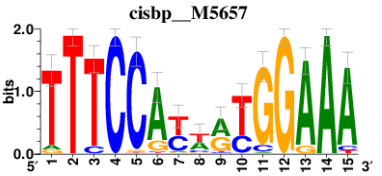
Kidney	Hes7	0.084	3.027	
Kidney	Foxc2	0.097	3.003	
Cancer	Arid5b	0.131	3.724	
Cancer	Ascl2	0.107	3.068	
Cancer	Atf3	0.21	4.73	
Cancer	Batf	0.082	2.163	
Cancer	Batf3	0.101	2.812	
Cancer	Bhlhe41	0.079	2.053	
Cancer	Cbx3	0.155	3.323	
Cancer	Cd59a	0.067	2.401	
Cancer	Cers4	0.096	2.502	

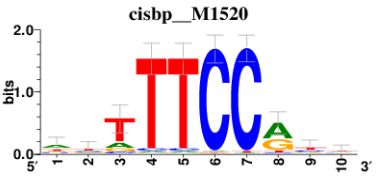
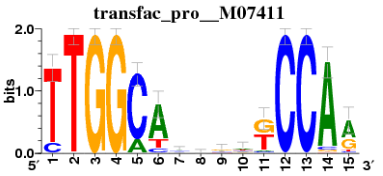
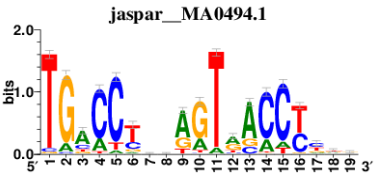
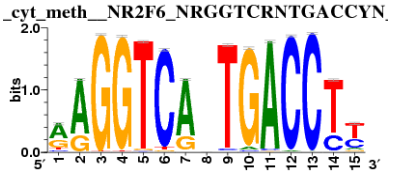
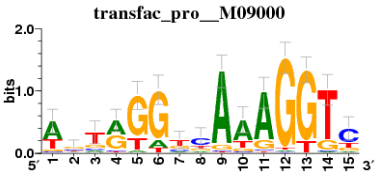
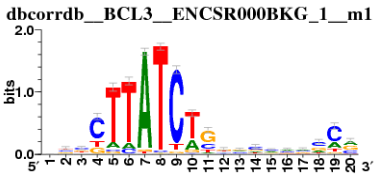
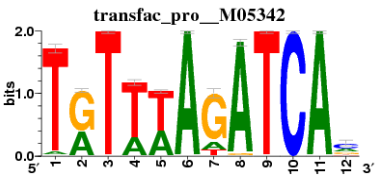
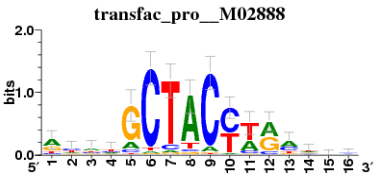
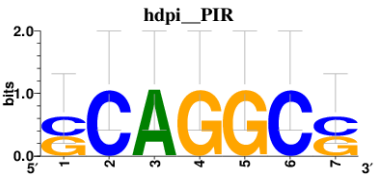
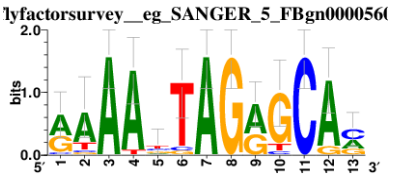
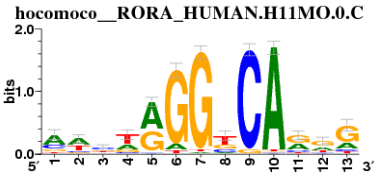
Cancer	Creb5	0.198	5.266	
Cancer	Ddit3	0.186	5.425	
Cancer	Ebf1	0.097	1.893	
Cancer	Ebf2	0.117	3.13	
Cancer	Egr1	0.158	3.487	
Cancer	Elf4	0.151	5.587	
Cancer	Elk3	0.207	4.993	
Cancer	En1	0.133	3.38	
Cancer	Erg	0.091	3.833	
Cancer	Ets1	0.162	6.059	
Cancer	Fli1	0.161	7.103	

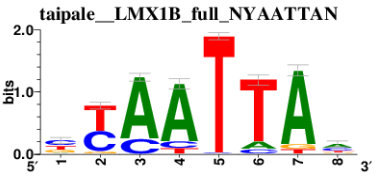
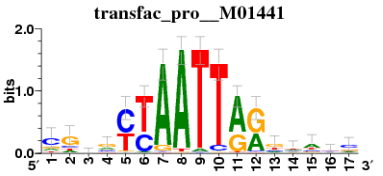
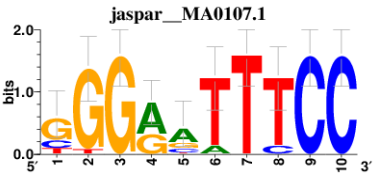
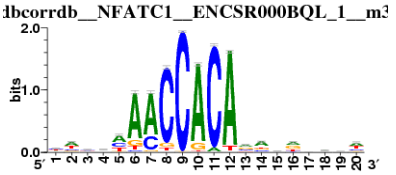
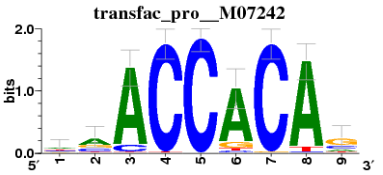
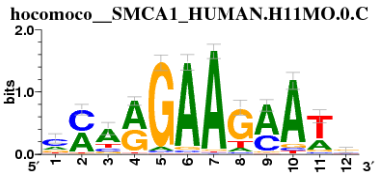
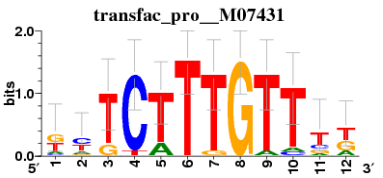
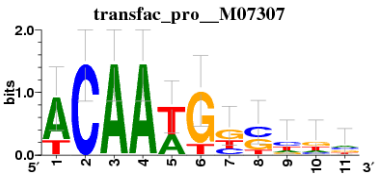
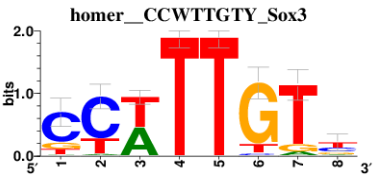
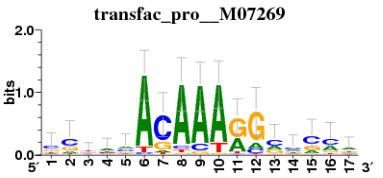
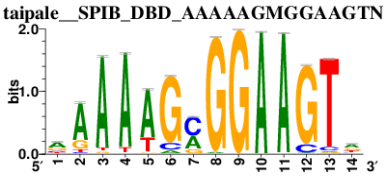
Cancer	Fos	0.193	3.802	
Cancer	Fosb	0.178	3.969	
Cancer	Fosl1	0.167	4.493	
Cancer	Foxa1	0.116	3.51	
Cancer	Foxi1	0.113	3.954	
Cancer	Foxp2	0.115	2.952	
Cancer	Foxq1	0.097	2.901	
Cancer	Gata2	0.142	5.896	
Cancer	Gli3	0.092	2.445	
Cancer	Hic1	0.137	2.997	
Cancer	Hmgb1	0.113	4.162	

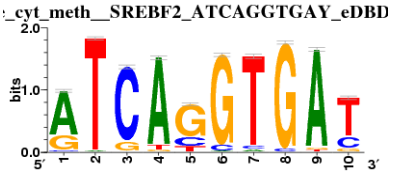
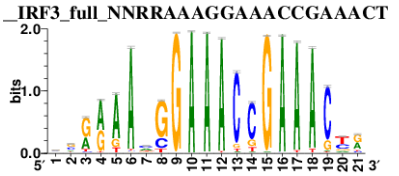
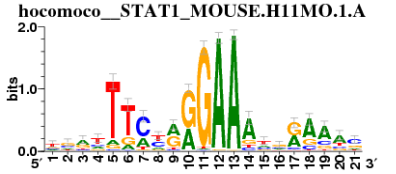
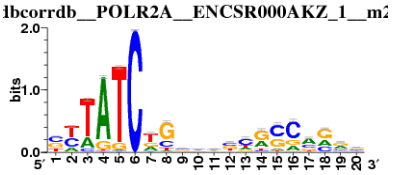
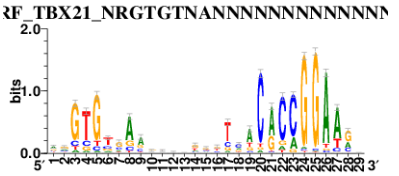
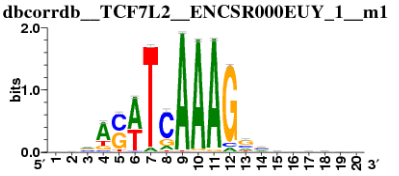
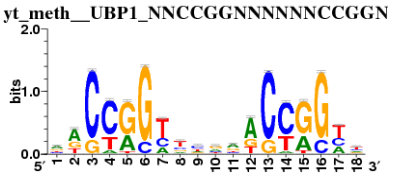
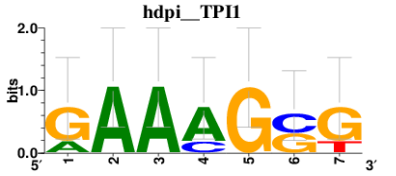
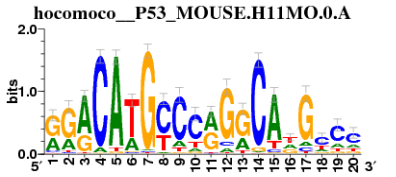
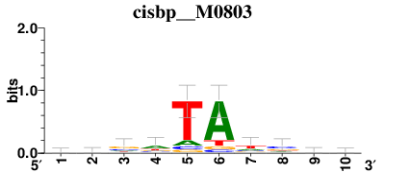
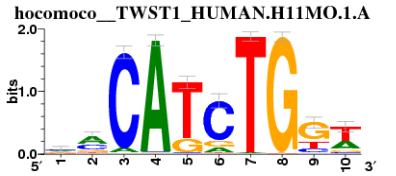
Cancer	Hmgb2	0.073	2.582	
Cancer	Hmgn3	0.067	2.368	
Cancer	Hoxa7	0.107	2.985	
Cancer	Hoxb2	0.206	4.283	
Cancer	Hoxb4	0.15	3.126	
Cancer	Hoxb5	0.211	4.453	
Cancer	Hoxb6	0.149	4.397	
Cancer	Hoxb9	0.13	2.644	
Cancer	Hoxc6	0.115	2.564	
Cancer	Hoxc8	0.092	1.924	
Cancer	Hoxc9	0.102	2.099	

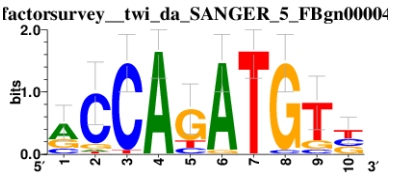
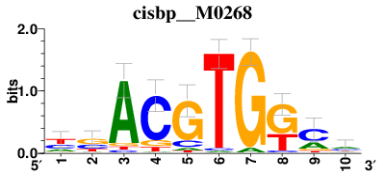
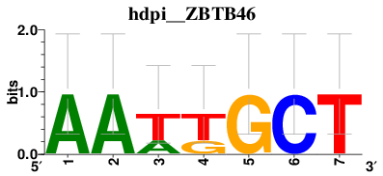
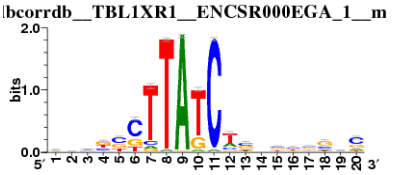
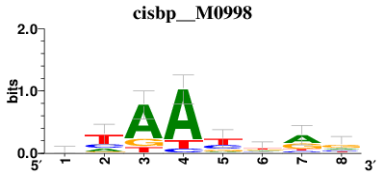
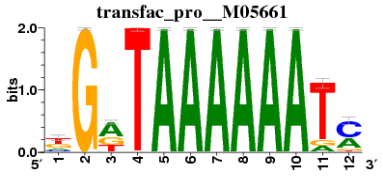
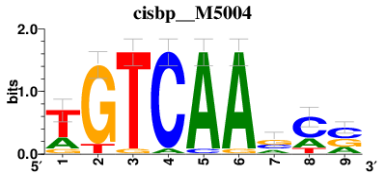
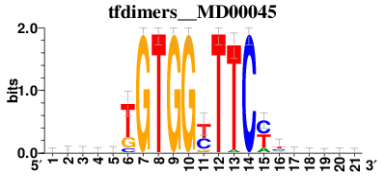
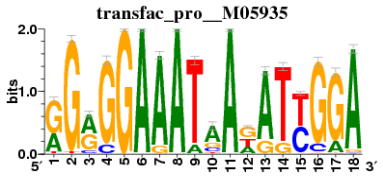
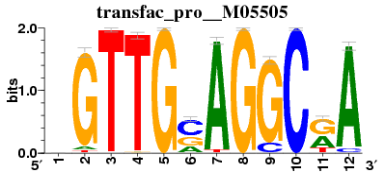
Cancer	Hoxd8	0.159	2.81	
Cancer	Ikzf2	0.134	2.643	
Cancer	Irf5	0.203	7.623	
Cancer	Irf7	0.423	19.275	
Cancer	Irf8	0.217	10.855	
Cancer	Jun	0.223	4.226	
Cancer	Junb	0.23	4.324	
Cancer	Klf15	0.161	4.473	
Cancer	Klf4	0.102	2.902	
Cancer	Klf6	0.136	3.113	
Cancer	Lef1	0.118	3.075	

Cancer	Lmo2	0.14	2.575	
Cancer	Ltf	0.052	1.759	
Cancer	Lyl1	0.063	2.29	
Cancer	Maf	0.101	2.813	
Cancer	Mafb	0.121	3.851	
Cancer	Mecom	0.101	2.79	
Cancer	Mef2c	0.163	3.468	
Cancer	Meis2	0.094	2.063	
Cancer	Msc	0.111	4.034	
Cancer	Msx1	0.083	2.662	
Cancer	Nfatc1	0.115	3.582	

Cancer	Nfatc4	0.1	2.396	
Cancer	Nfib	0.09	2.424	
Cancer	Nr1h3	0.136	2.932	
Cancer	Nr1i2	0.078	2.546	
Cancer	Nr2f1	0.172	3.297	
Cancer	Nr2f2	0.148	3.018	
Cancer	Nr4a1	0.111	1.851	
Cancer	Osr2	0.076	1.77	
Cancer	Pir	0.098	2.732	
Cancer	Ppara	0.143	3.563	
Cancer	Pparg	0.103	3.032	

Cancer	Prrx1	0.096	2.466	
Cancer	Prrx2	0.099	2.006	
Cancer	Rel	0.199	5.451	
Cancer	Runx1	0.123	3.355	
Cancer	Runx3	0.179	7.138	
Cancer	Smarca1	0.093	2.136	
Cancer	Sox11	0.085	2.212	
Cancer	Sox18	0.133	3.599	
Cancer	Sox4	0.095	1.909	
Cancer	Sox9	0.075	2.168	
Cancer	Spi1	0.277	10.821	

Cancer	Srebf1	0.133	4.105	
Cancer	Stat1	0.344	16.667	
Cancer	Stat4	0.069	2.302	
Cancer	Tal1	0.117	2.674	
Cancer	Tbx21	0.094	2.669	
Cancer	Tcf4	0.144	4.065	
Cancer	Tfcp2l1	0.103	3.482	
Cancer	Tpi1	0.087	2.234	
Cancer	Trp63	0.088	3.108	
Cancer	Trps1	0.126	2.848	
Cancer	Twist1	0.095	2.408	

Cancer	Twist2	0.102	2.214	
Cancer	Xbp1	0.096	3.773	
Cancer	Zbtb46	0.088	1.968	
Cancer	Zeb1	0.105	2.491	
Cancer	Zfhx3	0.093	2.547	
Cancer	Zfp105	0.09	2.337	
Cancer	Zfp275	0.093	3.363	
Cancer	Zfp354c	0.114	3.363	
Cancer	Zfp382	0.07	1.942	
Cancer	Zfp661	0.11	3.328	
Cancer	Zfp963	0.087	2.147