UNIVERSITY OF CALIFORNIA,
IRVINE

Inference of Cell Fate Transition from Single-Cell Transcriptomic Data

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematics

by

Yutong Sha

Dissertation Committee:
Professor Qing Nie, Chair
Professor Jun Allard
Professor Long Chen

2022

DEDICATION

To

my boyfriend Yuchi Qiu,

my father Hao Sha,

my grandfather Fengshan Sha,

and my mother Qin Mao

for their love, support and unfaltering belief in me

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

It is my fortune to have many people generously helping me out all the time. It is impossible to complete my doctor of philosophy degree without their help and supports.

I would like to express my most sincere gratitude to my committee chair, Professor Qing Nie, for his continuous support of my PhD study and related research, for his patience, immense knowledge, and belief in me. His visionary insights and enthusiasm for research inspire me to continue my academic path. I would not have been able to complete this dissertation without his constant guidance and encouragement.

I would like to express my gratitude to my committee members, Professor Jun Allard and Professor Long Chen, for their insightful comments, advising, and encouragement.

I would like to thank the Department of Mathematics, Center for Complex Biological System and Center for Multiscale Cell Fate Research for their help and cares for my lives and studies.

I thank Physical biology for permission to include Chapter 1 of my dissertation, Nucleic acids research for permission to include Chapter 2 and Chapter 3 of my dissertation, and Frontiers in genetics for permission to include Chapter 4 of my dissertation.

I would like to thank all former and current members of Nie lab. In addition, I want to take this opportunity to express my deepest appreciation to Dr. Shuxiong Wang and Dr. Peijie Zhou, for inspiring me and helping me deepen my existing understanding of the computational mathematics that serve as the background for my dissertation.

Special thanks go to Professor Xing Dai for stimulating discussions, which deepened my existing understanding of the biological components.

Last but not least, I would like to thank my parents and grandparents for always believing in me and supporting me morally and financially during my pursuit of a higher education. I also want to express my appreciation for my partner Yuchi Qiu and my friends Chao Chen, Lili Yan and Xiaowen Zhu for being my constant companions. Their love and support gave me the strength and courage to go through the tough time.

VITA

Yutong Sha

EDUCATION

Doctor of Philosophy in Mathematics **2022**
University of California, Irvine

Master of Science in Mathematics **2018**
University of California, Irvine

Master of Art in Mathematics **2016**
University of Wisconsin-Madison

Bachelor of Science in Mathematics **2015**
Nanjing University

EXPERIENCE

Research Assistant **2017–2022**
University of California, Irvine

Teaching Assistant **2017–2021**
University of California, Irvine

AWARD

NSF-Simons Center for Multiscale Cell Fate Research
Interdisciplinary Opportunity Award **2018**

PUBLICATIONS

- Y. Sha, D. Haensel, G. Guitierrez, H. Du, X. Dai, Q. Nie. Intermediate Cell States in Epithelial-to-Mesenchymal Transition. Physical Biology, 18:16(2):021001, 2019
- Y. Sha, S. Wang, P. Zhou, Q. Nie. Inference and multiscale model of Epithelial-toMesenchymal transition via single-cell transcriptomic data. Nucleic Acids Research, 48(17), Pages 9505-9530, 2020
- Y. Sha, S. Wang, F. Bocci, P. Zhou, Q. Nie. Inference of Intercellular Communications and Multilayer Gene-Regulations of Epithelial–Mesenchymal Transition From Single-Cell Transcriptomic Data. Frontiers in genetics, 11, 1700, 2021
- M. Mehrabi, T. A. Morris, Z. Cang, C. H. Nguyen, Y. Sha, M. N. Asad, N. Khachikyan et al. A Study of Gene Expression, Structure, and Contractility of iPSC-Derived Cardiac Myocytes from a Family with Heart Disease due to LMNA Mutation. Annals of biomedical engineering 49, no. 12 3524-3539, 2021

ABSTRACT OF THE DISSERTATION

Inference of Cell Fate Transition from Single-Cell Transcriptomic Data

By

Yutong Sha

Doctor of Philosophy in Mathematics

University of California, Irvine, 2022

Professor Qing Nie, Chair

Rapid growth of single-cell technologies provides unprecedented opportunities for close scrutinizing of heterogeneous cell states. However, detecting cell fate transition especially inferring the intermediate cell states (ICS) and transition cells from single-cell transcriptomic data remains challenging. In this dissertation, we focus on the epithelial-to-mesenchymal transition (EMT) as an example of cell fate transition. In Chapter 1, we introduce the existence and plausible biological roles of ICS in EMT. In Chapter 2, we present QuanTC, a method to infer cell fate transition, and a single-cell stochastic model of EMT which provides as a benchmark for QuanTC. In Chapter 3, we further apply QuanTC to single-cell transcriptomic datasets. We analyze the dynamical properties of inferred ICS based on a cell population model. In Chapter 4, we study the cellular crosstalk and the underlying gene regulatory dynamics along EMT from cancer cell lines with different inducing factors and find that the induced EMTs are context-specific. In Chapter 5, we combine deep learning with unbalanced optimal transport to model the temporal dynamics of time series single-cell transcriptomic data.

Chapter 1

Intermediate Cell States in Epithelial-to-Mesenchymal Transition

This chapter is a reprint of the material as it appears in [138]. The co-authors listed in this publication directed and supervised research which forms the basis for this chapter.

1.1 Background

The transition from epithelial to mesenchymal cells, as well as the reverse (mesenchymal-to-epithelial transition or MET), are highly dynamic processes implicated in various biological processes. EMT is well studied in the context of embryogenesis, organ fibrosis, and cancer metastasis, which highlight the categorical subtypes of EMT [77]. As our understanding of how EMT is regulated expands, it has become clear that EMT exists on a spectrum or continuum with various possible cell states existing between epithelial and mesenchymal phenotypes [118]. Cells do not necessarily exist in "pure" epithelial or mesenchymal states, but instead can be in intermediate cell states (ICS) or hybrid states that possess character-

istics of both epithelial and mesenchymal cells [118]. Context and tissue specific functions of EMT exist, but a number of key features and phenotypes are shared amongst the different subtypes of EMT [77, 118].

EMT involves the transient reduction and sometimes full loss of adhesion between cells. Hallmark changes involve alterations in cytoskeleton architecture, changes in cell-cell and cell-matrix adhesions, and loss of apical-basal polarities [92]. In the past, EMT identification had been focused on the expression of a few markers such as E-cadherin (E-cad); however, as our understanding of EMT has progressed, an expanded view of EMT identification encompasses multiple layers of molecular changes including EMT transcription factors (EMT-TFs) and their gene regulatory networks that govern the transition [118]. A number of key EMT-TFs such as Snail/Slug and Zeb1/2 have been identified in promoting the mesenchymal phenotype while others TFs such as Ovol1/2 and Grhl2 have been shown to suppress the mesenchymal phenotype thus promoting the epithelial phenotype [118, 92]. EMT regulation also occurs at the microRNA, long noncoding RNA, chromatin, and post-translational levels [118].

1.2 EMT in normal and diseased epithelial tissues

The development and maintenance of epithelial tissues is largely driven by epithelial stem cells [15]. The dynamic activities of these stem cells such as proliferation and differentiation vary depending on the stage of the tissue such as during development, adult regeneration, or in pathological conditions such as cancer [15]. Increasing evidence implicates EMT (and its regulation) as being another cellular dynamic component that can promote stem cell function [118]. In general, stem cell function in adult epithelial tissues is largely focused on the homeostatic maintenance of the tissue where cells that are differentiated or lost need to be replaced [15]. In line with the notion that EMT exists in a spectrum of states and

degrees, its role can vary depending on the epithelial tissue, stage of development, whether the tissue is undergoing regeneration or repair, and the pathological condition [118]. Below we highlight several instances of EMT function and regulation in committed epithelial tissues in different contexts such as normal development, adult regeneration and repair, and briefly discuss in the context of cancer (readers are referred to other comprehensive reviews on this topic, e.g. [118]).

The normal and cancerous mammary epithelium has been utilized as a model to study the potential role and regulation of EMT in epithelial stem cell function. Experiments using immortalized human mammary epithelial cells provided *in vitro* evidence that ectopic expression of EMT-TFs Snail and Twist leads to acquisition of stem-like activities [108]. A positive correlation between heightened EMT gene signature and stem cell fate has also been noted for normal and cancerous mammary epithelial cell types in culture or directly isolated from the human or mouse tissue [112, 158]. These data generated much interest in EMT as they suggest a link between EMT and the gain of stem-like features.

Further work in the mammary model *in vivo* found that Snai2 (Slug) is the major EMT-TF expressed in mouse mammary basal cells known to contain multipotent stem cells, and that ectopic expression of Snai2 leads to enhanced stem-like features [114]. Moreover, knockout or knockdown of Snai2 compromises mammary epithelial development and/or the ability of primary mammary epithelial cells to regenerate a mammary tree [114, 48]. Zeb1 has also been found to be expressed in normal mouse and human mammary basal cells [115], with expression particularly enriched in the Procr+ stem cell subset [158]. However, its functional significance remains to be elucidated. Interestingly, using a transgenic mammary tumor model, Snail- but not Snai2- expressing cells appeared in the early hyperplastic lesions as well as more high-grade carcinomas [173]. These cells lack E-cad expression and begin to express other EMT-TFs such as Zeb1, suggesting that Snail (but not Snai2) is responsible for governing the EMT program in cancer progression [173]. These observations highlight

the notion that different EMT-TFs can have different, context-specific functions even in the same tissue, and the exact underlying molecular and cellular mechanisms may differ. Thus, we emphasize again the importance to now expand our view of EMT beyond a simple binary, linear or universally identical process with the end goal of generating mesenchymal cells. EMT can be thought of as a historical term that is redirected to describe the diverse and complex variant forms associated with epithelial-mesenchymal plasticity. Specifically, EMT may be considered a navigation through a rugged, highly nonlinear multidimensional landscape of different axes that cumulatively define EMT [76, 136]. On this landscape, cell states other than epithelial and mesenchymal cells often exist, exhibiting mixed (or hybrid) features of epithelial and mesenchymal states. Such cell states, termed as Intermediate Cell States (ICSs) in this paper, may play important roles in regulating transitions between epithelial cells and mesenchymal cells.

Growing evidence also points to the importance of regulating EMT during physiological epithelial development and regeneration. Within the mammary epithelium, suppression of EMT by Elf5 and Ovol2 TFs appears to be an integral component of its normal development and regeneration [31, 164]. Loss of Ovol2 in the mammary epithelium results in an up-regulation of a large number of EMT/mesenchymal markers such as vimentin (Vim) and EMT-TFs such as Zeb1, as well as *in vivo* morphological transformation reminiscent of EMT [164]. Importantly, many of these EMT genes are direct targets of Ovol2's transcriptional repressor activity and depletion of Zeb1 rescues the regenerative defect caused by Ovol2 deficiency [164], underscoring an EMT-centric function of Ovol2 in the mammary gland. Interestingly and adding to the clinical importance of EMT regulation, incidence of metastasis-free survival increases in breast cancer patients with low levels of Ovol2 [164].

Transcriptional inhibition of EMT by Ovol2 and its homolog, Ovol1, is also critically important for normal skin epithelial development during embryogenesis. Loss of both Ovol2 and Ovol1 leads to defective epidermal and hair follicle morphogenesis [96]. Similar to the

observations in the mammary gland, loss of Ovol leads to up-regulated expression of EMT structural markers and EMT-TFs, as well as EMT-like phenotypes such as reduced adhesion between, and aberrant migration of, embryonic epidermal cells [96]. In adult skin, loss of Ovol2 alone results in defective wound healing [51] , a process that has been proposed to involve partial EMT of wound peripheral epidermal cells so they can efficiently migrate to close the wound [118, 50, 5]. Ovol2-deficient epidermal and hair follicle stem cells migrate faster than their normal counterparts, but with significantly reduced directionality [51] – defects that are near-completely rescued when EMT-TF Zeb1 is simultaneously lost. On the other hand, loss of EMT-TF Snai2 compromises epidermal migration after wounding, and results in a thin epidermis and transient delay of hair growth during normal development [5, 141, 60]. Together these data suggest that a delicate balance of EMT regulation exists to maintain both epithelial-like and mesenchymal-like states and this ability to toggle between the two states might be critical for skin epithelial development and regeneration. The study of EMT in cancer has largely been focused on its role in promoting invasion and metastasis with an involvement in chemoresistance recently demonstrated [118]; however many questions remain unanswered. Conflicting literature exists and the extent of EMT importance in patients' clinical outcomes remains unclear [118]. The roles of partial EMT and intermediate states in the context of cancer is still not well understood likely due to the transient nature of these events [118, 30]. The variation of different EMT-TFs' expression together with the diverse cell types in the tissues of origin adds to the complexity. Below we delve deeper into the idea of intermediate states and discuss relevant existing data in the contexts of both normal and diseased cells.

1.3 Existence of ICS in EMT

During EMT, cells exist on a reversible spectrum, and can toggle between ICSs, which lie between epithelial and mesenchymal states. Several lines of evidence using *in vitro*, *in vivo*, and computational modeling suggest the existence of ICS. *In vitro* models that have focused on using various cell lines are largely focused on the dual expression of epithelial and mesenchymal markers for identification of ICS. Glomerular parietal epithelial cells (GPECs) from adult mouse kidney adopt an intermediate phenotype expressing both epithelial and mesenchymal markers [144]. The MCF10A mammary epithelial cell line exists in an ICS with expression of both epithelial (E-cad) and mesenchymal (VIM) markers as compared to other well characterized breast cancer cell lines [55]. Cancerous cell lines have also been shown to exhibit ICS. Some cells from the human non-small-cell lung cancer cell line A549 are in an intermediate EMT state in culture based on co-expression of VIM and SNAI2 [4]. ITGB4+ triple-negative breast cancer stem cell-enriched cells reside in an intermediate state between "pure" epithelial and mesenchymal cells [14]. Research from a number of groups indicates that individual carcinoma cells, rather than all cells at a population level, can exist in ICSs. Distinguished by the widely used cell-surface markers resolving epithelial and mesenchymal cells, single CD24+/CD44+ cells in tumorigenic human mammary epithelial cells are ICS cells [57]. H1975 lung cancer cells co-stained for both VIM and E-cad display an ICS phenotype at a single-cell level [74]. Circulating tumor cells from human breast cancer cell lines captured by a microfluidic device from blood exhibit both epithelial (MCF7 and SKBR3) and mesenchymal (MDA-MB-231) characteristics, indicating existence of potential ICSs [174]. The classical view of the metastatic cascade depicts a sequence of events including dissemination, invasion, intravasation, and eventually extravasation, and ICS cells expressing both epithelial and mesenchymal features can be observed at these different stages; however, dissecting the exact role that EMT plays at these stages *in vivo* has proven to be difficult [118]. Findings generated by modeling and computational approaches also support the idea

of ICS. Modeling of the microRNA (miR)-based chimeric modules showed the miR-200/Zeb module functions as a ternary switch, allowing an intermediate phenotype in addition to the epithelial and mesenchymal phenotypes [105]. Random circuit perturbation (RACIPE), interrogating the robust dynamical behavior of a gene regulatory network, has identified two different types of intermediate phenotypes when applying to a proposed 22-gene network of EMT [56]. A topographic map using the combination of numerical simulations of a Boolean network model and the analysis of bulk and single-cell gene expression data revealed multiple ICSs, separating stable epithelial and mesenchymal states [43]. Energy-landscape based methods like scEpath could also be used to identify ICS [68] and the landscape of the free energy changes during the EMT of lung cancer cells suggests a stable ICS [176]. Previous studies predicted that multiple ICSs may exist between epithelial and mesenchymal states [146], but the number of intermediate phenotypes is still a matter of debate [46]. Recent experiments have provided evidence for one, two and three intermediate phenotypes from TGF-β–induced EMT in MCF10A cells [179], ovarian carcinoma (OC) cell lines [58], and circulating tumor cells in blood [174], respectively. While theoretical analysis helps to predict the number of ICSs, modeling studies have revealed that complex EMT regulatory networks govern the existence of multiple ICSs [107]. Modeling of miRNA-based regulation composed of a tristable circuit miR-200/Zeb driven by the monostable module miR-34/Snail allowed one intermediate EMT phenotype [105]. Ovol can modulate cellular plasticity by restricting EMT, driving MET, expanding the existence of the ICS and turning both EMT and MET into two-step processes based on the mechanism-based mathematical model coupling Ovol with the core regulatory network miR-200/Zeb [62]. Later, two distinct intermediate phenotypes in EMT dynamics were predicted and experimentally validated, and were shown to be modulated by Ovol2 and the Ovol2-Zeb1 mutual inhibition circuit [55]. Recently, analysis of a mathematical model that integrates expression data with the reported Tcf21-Slug interactions reveals one stable and two metastable ICSs in EMT [155].

1.4 Roles of ICS

1.4.1 Stemness

Recent theoretical and experimental studies have suggested that cells in ICSs can be more stem-like than both "pure" epithelial and mesenchymal cells. For example, trophoblast stem cells isolated from the conceptuses of MAP3K4 kinase-inactive mouse (TSKI4) cells are found to be trapped in an ICS and exhibit properties of both EMT and stemness [1]. Other cells at ICS states may also acquire stemness. For example, co-culture of cells from E and M states actually enhances mammosphere formation (an assay for stemness), and the isolation of the CD24+/CD44+ hybrid E/M cell state leads to enhanced stemness [57]. In another example, more than 90% of CD24+ cells from kidney GPECs show co-expression of surface markers of renal progenitors CD24 and cadherin-11, suggesting that these cells have acquired stem cell-like properties [144]. Mathematical modeling and computational analysis also predict that the ICS confers stemness potential. A model on three-way switch LIN28/let-7 circuit and a model on EMT-STEM-Notch coupled circuit both revealed a high likelihood of an ICS phenotype (when compared with either epithelial or mesenchymal states) in gaining stemness [72, 19]. Moreover, the ICS location in the low-dimensional gene expression space can shift towards either end of the EMT spectrum, and a crosstalk between EMT and stem cell regulatory modules in conjunction with Notch signaling creates a window of opportunity for stemness [19]. If one could reduce the size of the window or shift the window away from the mesenchymal end of the EMT space, the tumor would be less aggressive in growth or invasiveness [116]. ICS importance has also been analyzed in other systems not directly related to EMT [107]. For example, based on a five-node stochastic gene regulatory network controlling cellular stemness, ICS is found to significantly decrease the barrier of potential landscape and the minimal action value along the transition path to promote the differentiation process [175]. Clearly, additional experimental studies are needed to clarify the

mechanistic link between ICS and stemness, the relationship of which may be correlative rather than casual in some cases.

1.4.2 Collective migration

The ability of cells in ICS to migrate in a collective manner has been implicated during embryonic development, wound healing and the formation of tumor cell clusters [5, 74, 110, 131]. Cells undergoing collective migration display some migratory characteristics reminiscent of mesenchymal cells while maintain epithelial characteristics, implicating a connection to the ICS. For example, collective migration of H1975 lung cancer cells *in vitro* is associated with an ICS phenotype, and both are impaired after depletion of EMT-inhibiting TFs [74]. Cross-talks between Tcf21 and Slug have been identified to mediate phenotypic and migration plasticity in high-grade serous ovarian adenocarcinoma, and ICSs were found to be important in collective cell migration [155]. As discussed above, our recent studies highlight the importance of balancing the functions of EMT-inhibiting TF Ovol2 and EMT-TF Zeb1 in achieving directional cell migration of skin epithelial cells to support tissue repair and regeneration [51].

1.4.3 Drug resistance

The ICSs in EMT have great clinical relevance as they have been associated with drug resistance. The triple-negative breast cancer cells, which contain a number of intermediate E/M cells in primary tumor [174], exhibit de novo resistance to current standard therapies such as anthracyclines and taxanes [3, 70]. However, the underlying signaling pathways and molecular mechanisms of the interplay between ICS and drug resistance remain largely elusive [70].

1.4.4 Metastasis

The overarching role of EMT in metastasis has been extensively studied, however the precise role of ICS remains unclear. During the metastatic cascade, cells in primary tumor adopt mesenchymal-like features and are then able to leave the primary tumor and colonize in other locations, such as in mouse breast or pancreatic cancer models where Snail-positive or "EMTing" cells are sometimes fully detached from epithelial islands while losing E-cad expression [132]. The EMT-TF Zeb1 has been shown to be important for promoting a spectra of tumor types from mesenchymal, mixed (possibly at ICS), and epithelial as its loss leads to a confined epithelial state in pancreatic cancer [87]. *In vivo* live cell imaging coupled with gene expression analysis indicates that cells disseminated from primary tumors exhibit low expression of E-cad and enhanced expression of mesenchymal markers [9], suggesting that those cells have characteristics of ICSs. Interestingly, in a pancreatic cancer model, deletion of Twist or Snail does not cause any alterations in invasiveness [182], indicating cells (possibly ICS) that lose some of the well-studied mesenchymal markers may possess mesenchymal-like features. Along the same lines, in a lung metastases model, although EMT occurs within the primary epithelial tumor, the initial lung metastases are mainly derived from non-EMT tumor cells [42], and EMT cells do play a major role in recurrent lung metastasis after chemotherapy [42]. Beyond the gain of mesenchymal-like features during primary tumor dissemination, dynamic EMT gene expression is also observed in circulating breast tumor cells [174]. Together these data suggest that during EMT there are cells other than mesenchymal cells showing mesenchymal-like features with tissue and context specific roles during metastasis. Further analysis on such cells and their connection with ICS is important for better understanding of metastasis.

1.4.5 Speculated roles of ICS: controlling noise and dynamical robustness

Mathematical modeling and computational analysis have shown that ICS has the ability to control noise and increase the robustness of EMT dynamics. Increased number of intermediate phenotypes in the EMT system can better attenuate the overall fluctuations of the cell population in terms of phenotypic compositions, thereby stabilizing a heterogeneous cell population on the EMT spectrum, via a dynamic ODE modeling of the population of each cell phenotype [145]. The existence of ICS can also allow noise attenuation while maintaining the mean of the signal [129]. In another example using a regulatory circuit of miR-200, Zeb, and Snail, the ICS is observed to increase the plasticity of cell fate and the robustness of EMT dynamics [100]. With ICS, the EMT process can be carried out through transition first from the epithelial state to a ICS temporarily, and making further transition then from the ICS to the mesenchymal state or directly going from the epithelial state to the mesenchymal state. Moreover, cells at ICS can go back to epithelial state at other times, depending on the signal inputs for flexibility [100].

1.4.6 Stability

The ICS of EMT has been considered "metastable" [146, 97], reflecting the flexibility of these cells to undergo or reverse the EMT process [133]. In angiogenesis, hybrid tip/stalk phenotype, resulting from higher production levels of Jagged, relates with poorly perfused and chaotic angiogenesis based on the theoretical framework for Notch-Delta-Jagged-VEGF signaling [16]. But some evidence show that ICS could be stable. Experimentally, H1975 lung cancer cells can display a stable ICS over two months in culture [74]. In addition, the landscape of the free energy changes during EMT of the lung cancer cells shows a stable intermediate state [176]. Not only in EMT, ICS of immune system's T-helper cells are also

observed and are stable, having two canonical subtypes [59]. Interestingly, computational modeling that considers the mutual inhibitory loops between several miRNAs and EMT-TFs indicates that such networks are capable of generating additional stable ICSs [118, 55, 74] and have identified multiple phenotypic stability factors, such as Ovol, GRHL2, miR-145 [74] and NUMB, that can stabilize an intermediate E/M phenotype [20]. What's more, ICSs can be stabilized due to the increase of gene expression noises [56, 85].

1.5 Rising issues and challenges

Recently, single-cell RNA sequencing (scRNA-Seq) has emerged as a powerful means to dissect the heterogeneity in normal and diseased epithelial tissues. The ability to sample the transcriptome and quantify gene expression changes at single-cell resolution has yielded unprecedented insights into the physiology of epithelial systems by aiding in the discovery of rare cell types, shedding light on the dynamics of lineage differentiation, and providing comprehensive gene expression profiles of diverse cell types [115, 94]. scRNA-Seq has also begun to be employed to better understand the transition states that occur during EMT in cancer. In a recent study, scRNA-Seq revealed the presence of a partial EMT-like state in human squamous cell carcinoma (SCC) samples, implicating the existence of ICS *in vivo*. Another study utilized mouse models of SCC and mammary tumorigenesis for single-cell surface marker/RNA-seq analyses, and identified subpopulations of tumor cells corresponding to different degrees of EMT, with some displaying hybrid phenotypes that likely represent multiple distinct ICSs *in vivo* [121]. It is anticipated that single-cell approaches will also reveal EMT-related heterogeneity and enable the detection of ICSs in various normal epithelial tissues. Indeed, sequencing of 1,916 single cells from eight different organs from E9.5 to E11.5 embryos led to the identification of epithelial-mesenchymal ICSs in all epithelial tissues sampled, including intestine, liver, lung, and skin [39]. In addition to characteriz-

ing ICSs, single-cell approaches have been used to suggest a role for EMT and/or ICS in development/differentiation and disease. Using an experimental pipeline that incorporates single-cell qPCR analysis, it was shown that human embryonic stem cells (hESCs) go through a stepwise differentiation process into hepatocytes by undergoing sequential EMT-MET with an obligatory ICS, implicating a potential role for EMT/ICS in hESC differentiation into a definitive endodermal fate [101]. The SCC cells of the distinct ICSs discussed above display different clonogenic and differentiation potentials, as well as distinct invasive and plastic properties, providing experimental evidence that the different transition states that arise from EMT progression have different biological functions [121]. It has become increasingly clear that ICSs during EMT are important biological entities and must be considered whenever EMT is studied. The exact roles of EMT, and particularly of the constituent ICSs in different developmental and disease contexts need to be investigated using a combination of experimental and computational approaches. It would be interesting to computationally model the overall dynamics of the cellular states in a continuum fashion to capture their relatively unstable properties rather than studying multiple statuses in between the main stable states. Developing sophisticated computational tools that facilitate the identification and characterization of various ICSs – cell states that are more transient and may be different from other major cell states - represents a major challenge as well as opportunity in the EMT field.

Chapter 2

Inference of Transition Cells via Single-cell Transcriptomic Data

This chapter is a reprint of the material as it appears in [140]. The co-authors listed in this publication directed and supervised research which forms the basis for this chapter.

2.1 Background

Rapid growth of single-cell transcriptomic data provides unprecedented opportunities for close scrutinizing of dynamical cellular processes. Through investigating epithelial-to-mesenchymal transition (EMT), we develop an integrative tool that combines unsupervised learning of single-cell transcriptomic data and multiscale mathematical modeling to analyze transitions during cell fate decision. Our approach allows identification of individual cells making transition between all cell states, and inference of genes that drive transitions. Multiscale extractions of single-cell scale outputs naturally reveal intermediate cell states (ICS) and ICS-regulated transition trajectories, producing emergent population-scale models to be explored

for design principles. Tested on the newly designed single-cell gene regulatory network model, our unsupervised learning method faithfully captures cell plasticity and transition trajectory.

2.2 Introduction

The epithelial-to-mesenchymal transition (EMT) is an important process observed in many biological systems, including embryogenesis, wound healing and malignant progression [118]. Recently, several lines of *in vitro* and *in vivo* evidence, along with computational modeling, suggest that cells undergoing EMT is not a simple binary switch, and during the transition some cells exhibit mixed features of both epithelial and mesenchymal features [118, 138]. Those cells characterized as intermediate cell state (ICS) have been implicated in the potential roles of stemness, collective migration, drug resistance, metastasis, and noise control [118, 179, 58].

Key gene regulatory elements of EMT, such as EMT-suppressing microRNAs and EMT-promoting transcriptional factors, have been used for modeling and experimental analysis of ICS. Existence of multi-stable states of the modeled gene regulatory networks has been used to imply existence of ICS [105, 149, 55]. Few regulators have been found to be critical in formation of ICS, such as a transcriptional factor Ovol for regulating growth and Notch signaling for cell-cell communications [55, 142, 17], and few others have been suggested in stabilizing ICS [62, 74, 69].

Are the cells in ICS showing strong variability or tightly controlled? Single-cell RNA sequencing (scRNA-seq) technology provides unprecedented opportunities to explore cellular heterogeneity, distinct cell states, marker genes and the accompanying functions [82, 157, 135]. Expression levels of epithelial and mesenchymal markers and transcription factors of ICS have been recently analyzed in EMT at single-cell resolution [120]. EMT scoring metrics

have been developed by applying the best-fit model obtained from a previously-developed iterative statistical procedure to quantify EMT status of cells in different cell lines [44, 61, 64]. More recently, a topographic map underlying EMT has been constructed to explore ICS for its phenotypic plasticity [43].

One major challenge is to analyze temporal dynamics of cells in EMT from the snapshot transcriptomic data. Pseudo-temporal ordering (pseudotime) of cells in scRNA-seq data provides trajectories of cells that may recapitulate transition between cell states. However, such approach is usually dependent on the cell-embedding in the low-dimensional space via dimension reduction or structured graphs [185, 47, 143]. Recently, the single-cell method SOUP allows classification of both pure and intermediate cells by constructing the cell-cell similarity matrix and estimating a membership matrix [185]. Robust tools to quantify the transition trajectories and detect driving genes in EMT are still in need.

What are the transitional properties of cells near or at ICS? Is ICS simply another stable cell state between epithelial and mesenchymal states? Can we construct and quantify the transition paths in EMT? Here we first develop an unsupervised learning method (QuanTC) to infer and quantify transitional property of individual cells in scRNA-seq data. Then we validate QuanTC against our EMT multiscale single-cell model, which combines several previously published gene regulatory networks.

2.3 Method details

2.3.1 Overview of QuanTC

QuanTC takes the scRNA-seq data matrix as input to construct a cell-cell similarity matrix using a consensus clustering method (Figure 2.1) [82]. Via non-negative matrix factorization

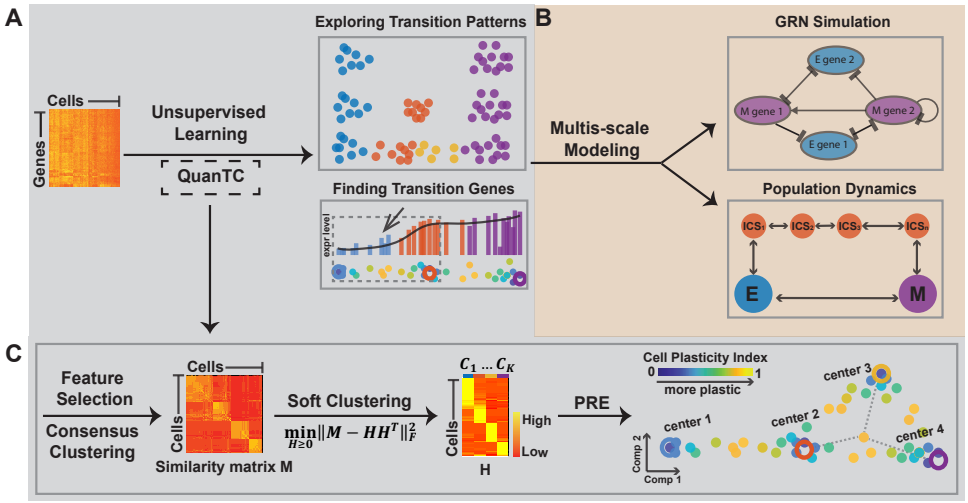

Figure 2.1: Outline of key components of the approach in analyzing transition cells and ICS. (A) Input single-cell transcriptomic datasets to an unsupervised learning method (QuanTC) to explore the transition cells, transition genes and other transition properties. (B) Develop multi-scale agent-based of gene regulatory network and cell-population dynamics models to validate and test outputs from QuanTC. (C) Overview of QuanTC: 1) feature selection and consensus clustering, 2) calculation of cell-to-cell similarity matrix, 3) computing cell-to-cluster matrix via NMF, and 4) using probabilistic regularized embedding (PRE) for two-dimensional visualization: Each solid circle represents one cell, colored by the value of Cell Plasticity Index (CPI) that quantifies the transition capability of each cell, and each larger circle represents the center of a stable cell subpopulation.

[88], a method of soft clustering, QuanTC then calculates the probabilities of a given cell belonging to the identified clusters (Figure 2.1C). To detect transition cells (TC), the cell-to-cluster probabilities are next used to measure the plasticity of each cell, i.e. the extent to which the cell may change its cluster identity. To better visualize cells in transition, we project cells to a low-dimensional space based on a probabilistic regularized embedding (PRE) (Figure 2.1C). The transition trajectories are then inferred by summing the cluster-to-cluster transition probabilities that are calculated from cell-to-cluster probabilities and TC between clusters. The clusters in the middle of the transition trajectories are denoted as ICS. The transition genes and marker genes of clusters are obtained through factorizing the gene expression matrix as product of cell-to-cluster probabilities and likelihoods of genes uniquely marking each cluster.

2.3.2 Feature selection and consensus matrix construction

We start by removing the low-expressed cells (expressed $< 5\%$ of the total number of genes), and the rare and ubiquitous genes that are either expressed in less than 10% of cells or expressed with low variance (< 0.005) among all cells (Figure 2.1C). Then we fit expressions of each gene with a Gaussian mixture model consisting of three distributions and use the weights and means of the model to choose the most informative (bimodal distributed) genes. We remove the rarely expressed genes for which the components of the mixture models with mean 0 accounting for more than 90% weights. To select the bimodal distributed genes, we rank the remaining genes according to two criteria. We first sort the difference between means of the top two components in descending order. Then we sort the difference between weights of the top two components in ascending order. By aggregating the ranks of the two orders, we select the top 3000 informative genes for further analysis.

QuanTC computes a cell-to-cell similarity matrix, M, through the cluster-based similarity

partitioning algorithm to estimate the similarity between cells. A binary matrix is constructed for each clustering outcome such that two cells are classified within one cluster, the corresponding value in the binary matrix is one, otherwise zero. A cell-to-cell similarity matrix M is calculated as the mean of the binary matrices constructed from clustering, leading to a symmetric non-negative matrix.

2.3.3 Quantifying transition cells via cell plasticity index (CPI)

Through symmetric non-negative matrix factorization [88, 89, 187, 22], the cell-cell similarity matrix M is decomposed into a product of a non-negative low-rank matrix H and its transpose (n is the number of cells, k is the number of clusters) (Figure 2.1C):

$$\min_{H \geq 0} \left\| M - HH^T \right\|_F^2, H \in R^{n \times k} \tag{2.1}$$

Each column of H represents a cluster and each row of H corresponds to the relative weights of a cell belonging to all the clusters. In other words, H contains the clustering information of cells: the largest element in each row showing the cluster identity of the corresponding cell and the likelihood of a cell belonging to each cluster. The number of clusters k is estimated by analyzing the largest gap of the sorted eigenvalues of symmetric normalized graph Laplacian (Figure B.1A).

By normalizing each row of H, we obtain a probability-like matrix $P = [p_{ij}]$ where p_{ij} represents the probability of cell i belonging to cluster j. QuanTC uses an entropy approach to characterize the degree of plasticity of each cell through a Cell Plasticity Index (CPI) (for cell i) defined as (Figure 2.1C):

$$\mathrm{CPI}_i = -\frac{1}{\log k} \sum_{j=1}^{k} P_{ij} \log \left(P_{ij} \right). \tag{2.2}$$

A cell undergoing the transition between clusters has higher entropy in contrast to cells located in one well-defined cluster. A higher value of CPI for a cell implies the cell is more plastic, making transition between clusters.

2.3.4 Visualization of transition trajectories

In order to faithfully capture both transition trajectories and discrete cell states, the cells are visualized through a probabilistic regularized embedding (PRE) approach using a probability-like matrix P in a low-dimensional space (Figure 2.1C). We first calculate the cluster-cluster relationship from $H^T H$, where each row of H denotes to what extend the cells belonging to each cluster while each row of H^T defines a distribution of weights over all cells in the cluster. The locations of cluster centers a_j in the two-dimensional space are then computed via the projection of the cluster-cluster relationship [86]. The projection of cells x is achieved by aligning each cell to the cluster centers based on the probabilities while keeping cells separate from each other through the following constraint:

$$\min_X \sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij} \left\| x_i - a_j \right\|_2^2 - \frac{\lambda_1}{n} \sum_{i=1}^{n} \sum_{l=1}^{n} \left\| x_i - x_l \right\|_2^2 . \tag{2.3}$$

The cluster with possible transitions to all the other clusters, which shows strong potential of high plasticity, is considered as a candidate for an ICS. The potential transition trajectory among clusters are then inferred via selecting one of the non-ICS (e.g. epithelial cells) as the initial cluster and ordering the clusters according to transitions. Two clusters are considered as neighbor if there are TC between them. By aligning cells along the potential cluster transition via the probability matrix P, QuanTC detects the transition trajectories. A cell i is aligned between cluster k and j if the two largest elements of ith row of the probability matrix P are located at kth and jth columns. The cells aligned from cluster k to j are then ordered in ascending CPI with the largest element at kth location and in

descending CPI with the largest element at jth location. The starting cell is selected as the cell with the largest probability belonging to the chosen initial cluster. In the method, multiple transition trajectories might exist, and the probabilities of occurrence of different transition trajectories are calculated by the percentage of cells included in each trajectory over the entire cell population size.

Furthermore, QuanTC calculates its own pseudotime of cells in each transition trajectory. A cell's pseudotime value is calculated as the Euclidean distance in PRE from the starting cell. In order to make the pseudotime value comparable for cells from different trajectories, we scale the range of pseudotime values between neighboring clusters to obtain a global pseudotime value of each cell by using the minimum value along all possible transition trajectories.

2.3.5 Finding cluster marker genes and the transition genes that mark transition

In order to identify the marker genes of clusters, we calculate the probabilities for each gene to uniquely mark a cluster. This is achieved by minimizing the difference between the submatrix D_s, containing cells from one inferred transition trajectory of the original feature selected gene expression D, and the submatrix H_s, with such cells of the factorized matrix H (m is the number of genes):

$$\min_{\bar{H},W} \left\| D_s - \bar{H}W \right\|_F^2 - \lambda_2 \operatorname{Tr}\left(\bar{H}^T H_s\right), \bar{H} \in R_+^{n\times k}, W \in R_+^{k\times m}. \tag{2.4}$$

The optimization solution leads to a gene-cluster matrix W to ensure that the factor matrix $\bar{H}$ is similar to H_s derived from the consensus similarity matrix. Then the gene-cluster matrix W can be used to infer transition genes and marker genes. Each column of W,

after normalization, describes likelihoods for the corresponding gene to uniquely mark the clusters. Each row of W, describes how well the genes delineate the corresponding cluster. The marker genes of cluster j are the genes with the largest values located at jth row of the column-normalized W. The marker genes of a specific cluster are then ordered based on their corresponding elements in row j of the column-normalized W. The difference of the top two elements of each gene is chosen to be greater than a given value (default value is 0.03) to ensure that the gene is differentially expressed in cluster j. The default value of λ_2 is 10, and how W depends on the parameter is investigated, showing robustness of the method (Figure B.1B).

In order to uncover genes that mark the transition, that is, the genes varying most among the transition (Figure 2.1A), we select the marker genes of the two clusters involved in the transition and calculate the Spearman's rank correlation coefficient between gene expression and the order of cells by CPI undergoing transition. Genes with absolute value of Spearman's rank correlation coefficient above a specified threshold (default value is 0.64) are considered as transition genes for the transition of the two clusters. A positive coefficient implies the gene expression levels of aligned cells show increasing changes while the negative coefficient implies decreasing in gene expressions during transition.

2.4 Multiscale agent-based single-cell model based on gene regulatory network

A multiscale model is constructed to track the gene expression values in each cell using an EMT regulatory circuit of genes [55] that are stochastic in time. 18 ordinary differential equations are used to describe the expression levels over time based on a previous study [55]. With certain parameters, the circuit has four distinct stable steady states. Each cell

is located at one of the four steady states or makes transition towards those steady states. The transition between different steady states may be caused by external signals or induced by stochastic influences over time. In the model, we make the following assumption:

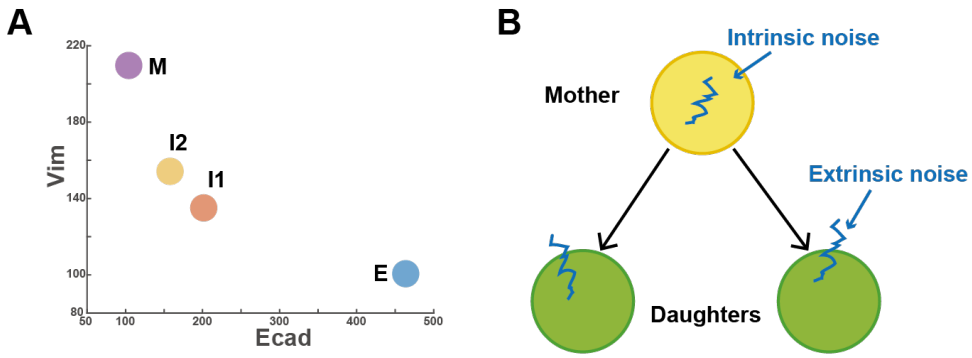

Figure 2.2: Modeling illustration. (A) Relationship between the four stable steady states and the expression levels of the epithelial marker (Ecad) and mesenchymal marker (Vim) in the model. Each dot represents a stable steady state. (B) Illustration of individual cells and cell division: the cell state transition may be caused by the intrinsic noise in gene regulatory dynamics or stochastic effects in cell divisions.

1. The initial population is composed of 200 cells: 50 epithelial cells (E), 50 first intermediate cells (I1), 50 second intermediate cells (I2) and 50 mesenchymal cells (M).

2. All cells divide at a normally-distributed rate $\sim \mathcal{N}(700, 200)$ ($\mathcal{N}$ refers to a normal distribution). The time unit in the model is hour and the parameter values of the model are chosen based on a previous study [55]. Every time a cell divides, its expression levels of all the EMT factors are used as initial conditions to its daughter cells. The gene expression levels of each cell are compared to the expression levels of different stable steady states in the EMT spectrum to determine the cell's phenotype. The E state is characterized by high Ecad expression, and M state is characterized by high Vim expression. I1 and I2 states are characterized by both relatively high Ecad and Vim expression while I1 corresponds to stronger Ecad expression and I2 corresponds to stronger Vim expression among the stable steady states (Figure 2.2A). The cells not at any steady states are considered as TC.

3. Stochastic effects are integrated into our model by adding two types of noise (Figure 2.2B). (a) We first perturb the expression levels of the mother cell upon its division into two daughter cells:

$$noise_{div} = I_{expr}^{mother} * \mathcal{N}(0, 0.7)$$
$$I_{expr}^{daughter1} = I_{expr}^{mother} + noise_{div}$$
$$I_{expr}^{daughter2} = I_{expr}^{mother} - noise_{div}$$

In this case, the noise added at the division is the expression levels of mother cell multiplied by a normally-distributed rate. The perturbed expressions serve as the initial conditions for the daughter cells. (b) The multiplicative noise is applied to the parameters in the EMT model:

$$dI_{expr} = f\left(I_{expr}\right) dt + \sigma I_{expr} dW_t$$

The function f represents the EMT regulatory circuit dynamics and W stands for the Wiener process with $\mathbb{E}W_t = 0$ and $\mathbb{E}W_t W_s = \min(t, s)$. σ represents the noise amplitude with default value 0.01. We use Euler-Maruyama scheme to numerically solve the system.

4. The number of times a cell can divide is described by a discrete uniform distribution $\sim \mathcal{U}(2, 7)$ with an equal probability chosen from a natural number between 2 and 7. Once the cell cannot divide any more, the cell dies at a normally-distributed rate $\mathcal{N}(1000, 100)$.

The multiscale model is simulated over a time span of five cell division cycles.

2.5 Results

Our study consists of two major components: a) unsupervised learning of scRNA-seq data and b) modeling EMT dynamics (Figure 2.1). To scrutinize the transition of cells, we first propose QuanTC (Figure 2.1C, Materials and Methods), a method to quantify the transitional status of individual cells and identify the transition genes that mark the transition process and the marker genes that distinguish different cell states. The QuanTC is then validated on a multiscale agent-based stochastic model based on a core EMT gene regulatory network (Figure 2.1B).

QuanTC faithfully captures cell plasticity and transition trajectory in simulated datasets

To test capability of QuanTC in capturing transition cells and intermediate cell states, we first constructed a multiscale single-cell model using a core EMT/MET gene regulatory network (Figure 2.3A) [105, 55, 62, 145, 107]. The new agent-based model dynamically describes the expression levels of genes featured in the regulatory circuit within individual cells, and explicitly includes cell division to track the individual cells. The cell state transition may be caused by the external signal (TGF-β) or stochastic effects in cell division and/or gene regulatory dynamics (Figure 2.2B). The single-cell model outputs a group of single cells along with the expression values of the 18 modeled regulatory components at each temporal point (Materials and Methods) to mimic an EMT scRNA-seq dataset.

One typical model simulation exhibits four distinct stable steady states corresponding to four cell phenotypes: epithelial state (E), two intermediate cell states (I1 and I2) and mesenchymal state (M) (Figure 2.3B). The intermediate state closer to the E is denoted as I1, and the one closer to the M as I2. The cells that have not reached any of the steady states are considered as transition cells (TC). In this simulated system, initially each state con-

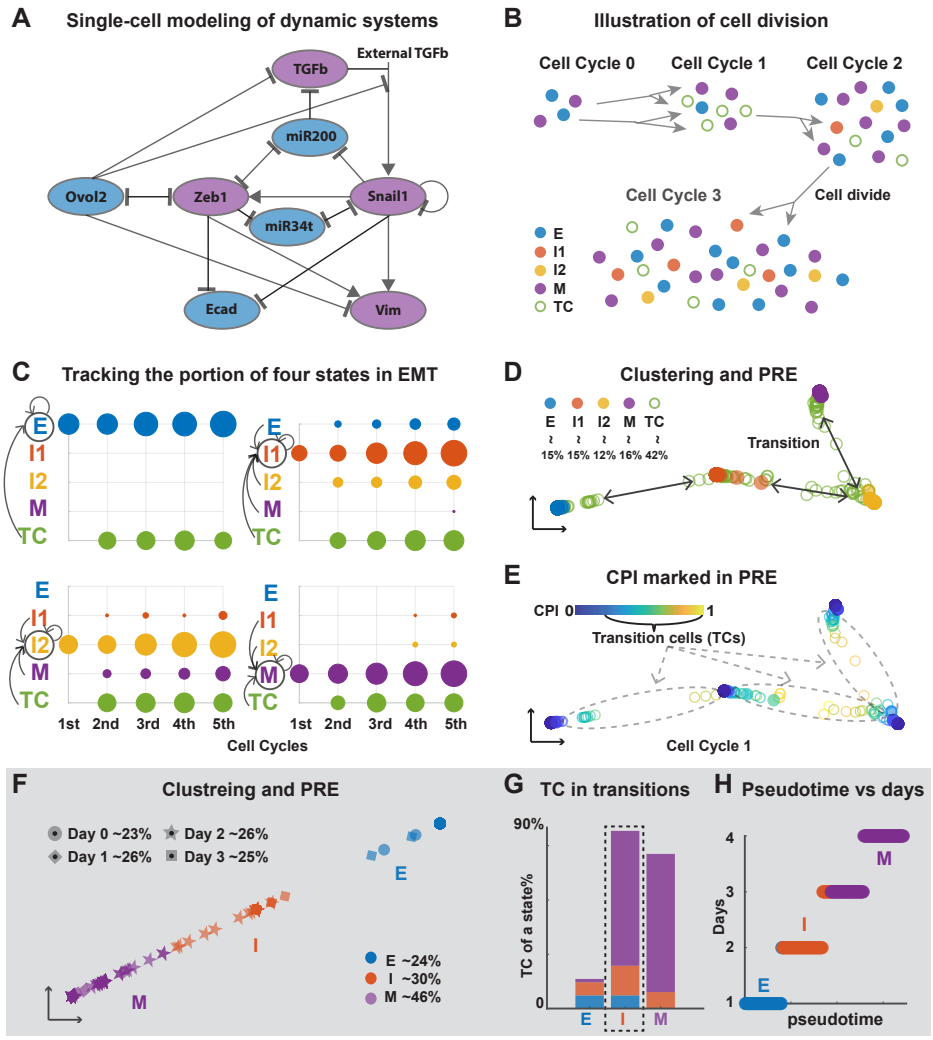

Figure 2.3: Testing QuanTC on simulated EMT datasets and a qPCR dataset for hepatic differentiation of hESCs. (A) The EMT gene regulatory network used in the multi-scale agent-based model; blue: epithelial promoting factor; purple: mesenchymal promoting factor. (B) Illustration of the modeling output: each cell colored by its true state labels. (C) A simulation dataset: the proportion of each state induced by the previous cell states at the end of each cell cycle. The size of the dot is proportional to the number of cells, and the color denotes the cell states of the mother cell. The arrows represent the occurred state transitions and the circle represents the state of the daughter cell. It shows the transition dynamics of each state. (D-E) PRE visualization of each cell at the end of first cell cycle (a circle) colored by its true state from the model (D) and the calculated CPI value (E). The percentage for each cell type is the percentage of a given cell type over the entire cell population size. (F) Clustering and PRE visualization of the qPCR dataset. Each dot represents one cell colored by the identified state, and its shape represents its real time. (G) Percentage of TC in each state relative to the total number of TC with colors consistent with (F). Dashed box: the intermediate cell state. (H) Comparison of the inferred pseudotime and the day collected in the experiment of each cell. The parameters are provided in Table B.1.

sists of 50 cells and after five cell cycles the system grows to 2030 cells. To detect possible transitions between the different states, the cells at the end of each cell cycle were tracked back to the previous cell cycle to identify their mother cells (Figure 2.3C and Figure 2.4A)). For example, E cells were found to come from TC whereas M cells came from TC with few from I1 and I2. The observed transitions among the four states indicate that TC have the strongest capability to give rise to all different EMT subpopulations with the cells in ICS next in such transition capability. Interestingly, E and M cells show less potential to make transitions directly (Figure 2.3C and Figure 2.4A).

The simulation dataset provides the true label of each cell and its transition details. Applying QuanTC to the data collected at the end of the first cell cycle, we identified four cell states and TC between them (Figure 2.3D), Materials and Methods). Principal component analysis (PCA) was unable to separate different states at the end of later cell cycles let alone detecting the potential transitions between states (Figure 2.4D-E). To quantify the transition capability, we computed cell plasticity index (CPI) of all cells (Figure 2.3E) and found that the TC marked using modeling data have relatively high CPI values while cells closer to the primary states have lower CPI values. More TC with higher CPI values were found to be around the two ICS (Figure 2.4B, D-E), suggesting high transition potential of ICS.

The transition genes that mark the transition processes between states, and the marker genes of identified states were uncovered using QuanTC (Figure 2.4. Ecad and ZEB, along with other genes sharing the similar expression behavior, were found to be marker genes of E and M cells. As for ICS cells, while no clear state marker genes were identified, multiple transition genes are highly expressed due to their strong potential to make transitions (Figure 2.4C).

Through cell state identification, estimating cell plasticity, and inferring marker and transition genes, QuanTC recapitulates the observed states and their transitions in the single-cell model that can be explicitly delineated.

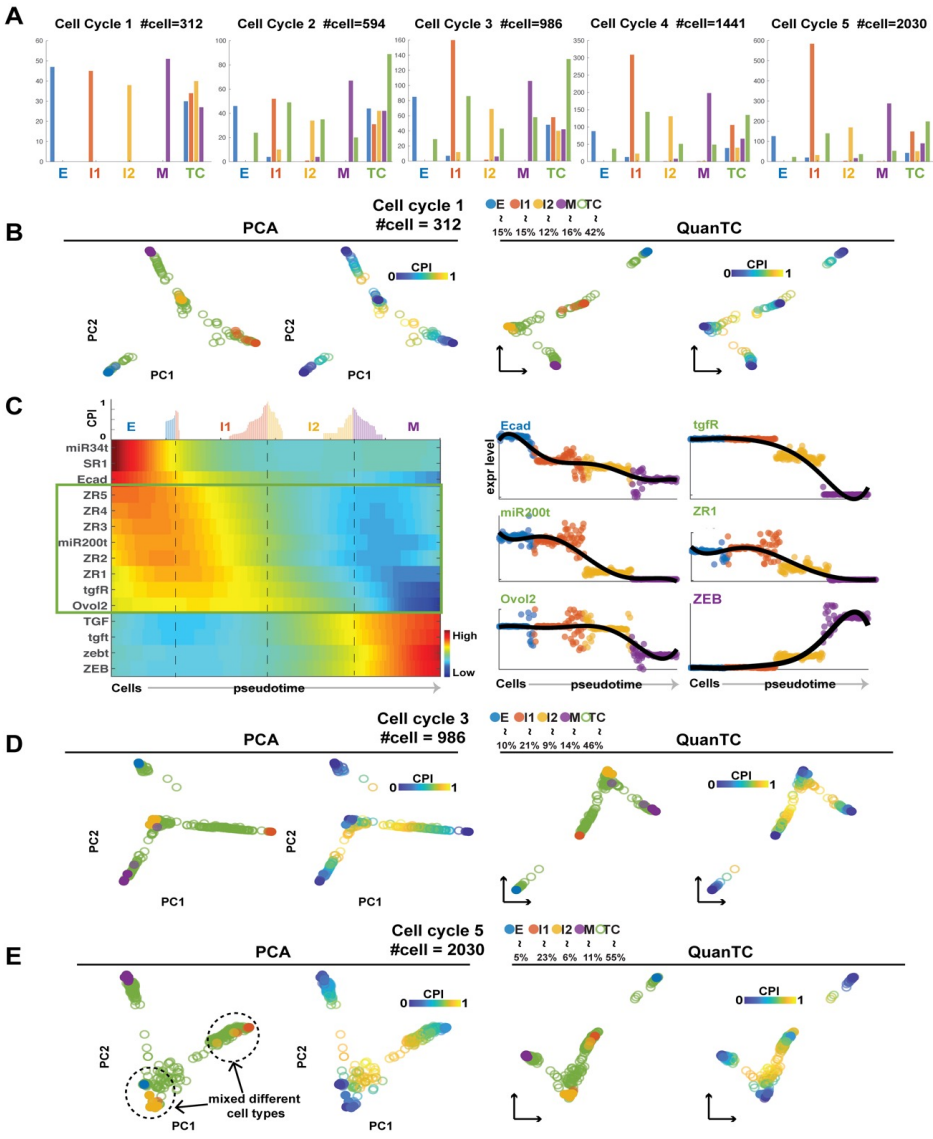

Figure 2.4: The distribution of the cell population at the end of cell cycles.

Figure 2.4 *(continued)*: (A) Histogram of the number of cell population at the end of each cycle. The color denotes the mother cell states. The x-labels represent the states of the daughter cell. (B-C) Simulated EMT/MET datasets at the end of first cell cycle. The percentage for each cell type is the percentage of a given cell type over the entire cell population size. (B) PCA and PRE visualization of the cells with each cell (a circle) colored by its true state (left) and the calculated CPI value (right). (C) Heat map of normalized expression of marker genes and transition genes (left). Columns represent cells ordered along the transition trajectory and rows represent genes. Coloring represents the normalized expression of each gene. Transition genes are marked in the box. Top: CPI values of each cell along the transition trajectory. Expression levels of top marker genes and transition genes with cells ordered along the most probable transition trajectories (right). Solid lines, smoothed expression curves for each gene in the transition trajectory. (D-E) PCA and PRE visualization of the cells with each cell (a circle) colored by its true state (left) and the calculated CPI value (right) from simulated EMT/MET datasets at the end of third (D) and fifth cell cycle (E).

2.6 Discussion

Compared with direct clustering [82, 185] and pseudotime analysis [126, 152, 165] for scRNA-seq data, the unsupervised learning algorithm QuanTC can simultaneously detect the intermediate cell states, and construct transition trajectories via quantifying the cell plasticity. An attractive feature of QuanTC is its soft clustering approach to identify cells in mixed states or undergoing transition between states, a ubiquitous property in many cell fate systems. The projection of cells in PRE marked by CPI for transitions offers a parsimonious and meaningful alternative to analyzing a large number of discrete cell states.

Unlike other methods that can only infer marker genes for cell subpopulations, such as a recent random coefficient matrix-based regularization method on identifying transition cells [183], QuanTC can uncover key genes that mark the state transitions. The projection of cells in PRE marked by CPI for transition processes offers a parsimonious and meaningful alternative to analyzing a large number of discrete cell types.

A multiscale agent-based model of EMT gene regulatory network has been developed to generate simulation data with the ground truth, allowing easy validation of our unsuper-

vised learning method QuanTC. Previous models were mainly focused on the regulation mechanisms of EMT by ODEs with feedback control to identify important agents that are responsible for initiating or suppressing EMT [179, 105, 149, 55]. In those models, cell activities or states defined by changes in gene expressions are confined within each individual cell. We have extended the modeling of EMT to a heterogeneous population of cells, while still incorporating gene regulatory networks, offering a convenient framework to explore cell proliferation by monitoring the changes in gene expressions prompted by interactions between various EMT agents, which is important for cancer studies [120, 79, 81]. Our model explicitly incorporates stochastic effects caused by each cell division [153, 66] that may affect cell fates. Our model can also easily incorporate different assumptions on proliferative dynamics of each cell state. For example, we have analyzed a case in which the I1 cells are assumed to be non-proliferative (Figure A.1) to investigate ICS under cell cycle arrest during EMT [104, 168].

In our study, more efficient algorithms to explore cell-cell similarities will likely improve QuanTC significantly in its speed and ability to learn transition trajectories. The agent-based multiscale model can be further improved by adding new interactions between genes and cell-cell communications over time, and the inclusion of other cell types, such as immune cells, may gain further insights into the functional role of ICS.

Chapter 3

Application of QuanTC to Single-cell Transcriptomic Data

This chapter is a reprint of the material as it appears in [140]. The co-authors listed in this publication directed and supervised research which forms the basis for this chapter.

3.1 Background

Applying our unsupervised learning method, QuanTC, to twelve published single-cell EMT datasets in cancer and embryogenesis, we uncover the roles of intermediate cell states (ICS) on adaptation, noise attenuation, and transition efficiency in EMT, and reveal their trade-off relations. Overall, our unsupervised learning method is applicable to general single-cell transcriptomic datasets, and our integrative approach at single-cell resolution may be adopted for other cell fate transition systems beyond EMT.

3.2 Introduction

What are the functional advantages of ICS in state transitions? Cell population modeling suggests the increased number of ICS attenuates the fluctuations in cell numbers during transition [145] in addition to help maintain the mean of signal response [129]. Experimental and modeling analysis shows ICS can also facilitate the robustness of population dynamics [49]. Signal adaptation has been found to tightly constrain gene regulations [106], and however, could be important as a "survival strategy" in growth and migration of cells [10]. At the level of gene regulations, achieving robustness and signal adaptation, which both are important to cell fate transition, often require different, sometimes competitive, gene regulations [124]. Comparisons of ICS across different EMT systems remain a major open problem [71].

Here we apply QuanTC to twelve published EMT transcriptomic datasets in cancer and embryogenesis. By inspecting transition cells, ICS, and their relationship with epithelial and mesenchymal states, we construct the ICS-regulated EMT trajectories. We then compare the inferred transition trajectories, which are different between cancer and embryogenesis, with another method based on critical transition theory, and re-constructed core gene regulatory circuits for the published datasets to analyze the similarity and consistency in state transition.

To further investigate the inferred trajectories shared by various EMT systems, we develop and analyze cell transition models by defining and measuring three metrics emergent from EMT cell population dynamics. Differences between inferred EMT trajectories and their integrations with scRNA-seq data are then analyzed. Our integrative approach, which fuses unsupervised learning of gene expression data at single-cell resolution along with principle-guided cell population model, provides multiscale effective connections between genes and cells in analyzing complex cell fate decision that involves ICS, multiple trajectories, and

genes that mark transitions.

3.3 Materials and methods

3.3.1 Quantification and statistical analysis

hESCs data

The single-cell qPCR data [101] was performed with 48 selected genes during a sequential EMT-MET from days 0 to 21. We start with 345 cells from day 0 to day 3. Based on the cell-cell similarity matrix resulting from consensus clustering [82], we use the largest gap of consecutive eigenvalues of symmetric normalized graph Laplacian to infer the number of cluster $k = 3$. The initial cluster chosen to be the start of transition trajectory because of including day 0 (epithelial) cells.

SCC data

We apply QuanTC to the SCC dataset [121] including 382 cells. After removing the low-expressed cells (expressed $< 5\%$ of the total number of genes), 361 cells remain for further analysis. After feature selection, we use top 3000 genes for consensus clustering and inference of marker genes and transition genes. The cluster having the smallest number of TC around (i.e. low transition taking place) is considered as the start or the end of the transition trajectory. The initial cluster is named as E state based on the high expression levels of Epcam. Other clusters are named based on the inferred transition trajectories compared with the E-I1-I2-M spectrum in EMT. The cell-cycle phase of each cell is determined based on the computed cell cycle scores provided in Seurat [150, 25].

Mouse embryonic development data

This scRNA-Seq data [39] includes cells from skin (155 cells), lung (176 cells), liver (123 cells), and intestine (173 cells) during E9.5 to E11.5. After removing the low-expressed cells (expressed $< 5\%$ of the total number of genes), 155 skin cells, 176 lung cells, 123 liver cells and 173 intestine cells remain for future analysis as in SCC data.

HNSCC data

This dataset [123] has $\sim 6,000$ single cells from 18 head and neck squamous cell carcinoma (HNSCC) patients. We focus on six tumors from which the largest numbers of malignant cell transcriptomes and cells involved in EMT were acquired. The six tumors include patient 5 (132 tumor cells), patient 6 (123 tumor cells), patient 17 (330 tumor cells), patient 18 (140 tumor cells), patient 25 (209 tumor cells) and patient 28 (138 tumor cells). For each patient, we first use all the tumor cells, based on the selected features by QuanTC, for clustering. Similar to the original study [123], we remove the clusters having high expression levels of the cell cycle and stress markers because those cells are known not involved in EMT. For the remaining tumor cells mostly similar to epithelial cells, we add 20 fibroblast cells to each dataset to act as a reference of mesenchymal cells. We then apply QuanTC to the mixed datasets of each patient. We notice that all the six datasets have four clusters including two ICS. The raw and filtered datasets are available on the package website (https://github.com/yutongo/QuanTC).

Mouse hematopoietic progenitors data

This scRNA-Seq data [54] includes 2018 cells. After removing the low-expressed cells (expressed $< 5\%$ of the total number of genes), 1957 cells remain for further analysis. Twelve

clusters are identified by QuanTC (Figure B.2A). The cells with high CPI values (> 0.34) are considered as TC (Figure B.2B). Cluster C6, C7 and C12 are considered as non-ICS or a potential start or end of the transition trajectories because fewer TC exist in or around them (Figure B.2C) with weak capability of making transition. B cells and plasmacytoid dendritic cells (pDC) share a common progenitor (42). Cluster C6, C7 are B cells and pDC, respectively, based on the high expressions of the known marker genes (*Ebf1, Irf8 and Siglech*). Based on the relative number of TC between clusters (Figure B.2D), the transition trajectories C5-C8-C7 and C5-C11-C6 indicate that B cells (C6) and pDC (C7) share a common progenitor C5. The transition trajectories inferred by QuanTC are consistent with the previous findings [54]. QuanTC identifies the maker genes and transition genes involved in the two transition trajectories (Figure B.2E). When ordering cells in the transition trajectories, the known lineage markers increase along the pseudotime (Figure B.2F).

Gene Ontology enrichment

The Gene Ontology enrichment analysis [6, 36, 109] is performed on the top 100 markers genes (Table S2 in [140]) of each ICS selected by QuanTC.

Comparison with Monocle 3

Monocle 3 [29] is applied to the simulation and SCC datasets (Figure B.3). While Monocle 3 separates Epcam+ tumor cells from Epcam- tumor cells in SCC dataset, it is unable to obtain the known epithelial to mesenchymal lineage (Figure B.3A). However, if only using the top 3000 genes selected by QuanTC (Figure B.3B), Monocle 3 is able to capture the previously observed epithelial to mesenchymal lineage, suggesting usefulness of QuanTC in feature selection. For the simulation dataset, Monocle 3 separates different cell states, however, it cannot identify TC, consequently cannot obtain the transitions between clusters

(Figure B.3C).

3.3.2 Dynamical system modeling of transition trajectories and three dynamic quantities

To reduce the parameter complexity and increase model accountability, we simplify the model to incorporate only three dimensionless parameters α,β and γ (Figure 3.1). For easy comparison, the direct transition rate (DTR) from E to M state is used as a base for comparison (set to one). The parameter α represents the dimensionless cell-state transition rate from M state directly to the E state (i.e. the reverse DTR). We assume that $\alpha > 1$ to guarantee that E state is more stable at equilibrium when there is no induced EMT by extrinsic signal. It also incorporates the effects of other possible M-to-E transitions (MET) that might not be revealed by the trajectories in EMT datasets. The parameter γ depicts the forward transition rate between adjacent cell states along the ICS-regulated transition path, also denoted as the indirect transition rate (IDR) of EMT. We use $\beta\gamma$ to represent the reverse cell-state transition rates along the indirect EMT routes with ICS. Based on the inferred transition paths (Results), we assume that $\gamma \gg 1$ and $\beta \ll 1$ such that EMT is mainly carried out through the ICS-regulated trajectories, and the rate of EMT is significantly larger than the reverse MET along these trajectories.

Then the prescribed ordinary differential equations (ODEs) that describe the population fraction change of epithelial $E(t)$, mesenchymal $M(t)$ and ICS $I_k(t)(k = 1, 2, .., N)$ can be derived.

$$\frac{dE}{dt} = \alpha M + \beta\gamma I_1 - (1+\gamma)E \tag{3.1}$$

$$\frac{dI_1}{dt} = \gamma E + \beta\gamma I_2 - \gamma(1+\beta)I_1 \tag{3.2}$$

$$\frac{dI_k}{dt} = \gamma I_{k-1} + \beta\gamma I_{k+1} - \gamma(1+\beta)I_k, 2 \leq k \leq N-1 \tag{3.3}$$

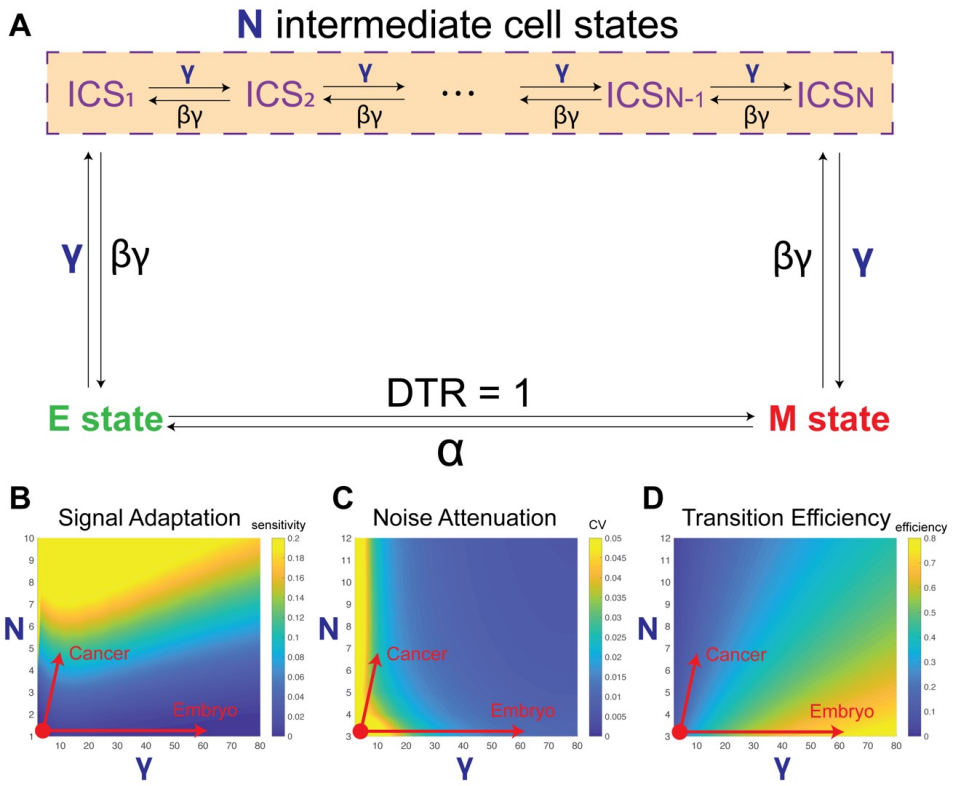

Figure 3.1: Mechanism and results of EMT population model. (A) The state-transition structure of population model and associated parameters. The model focuses on two major possible routes of EMT 1) the direct transition from E to M state, with rate DTR normalized as 1 and inverse transition rate α. 2) the indirect EMT transition mediated by N ICS, with the forward transition rate (also denoted as the indirect transition rate, ITR) γ and backward transition rate $\beta\gamma$. (B-D) The dependence of signal adaptation, noise attenuation and transition efficiency measures over the space of key parameter N and γ. We fix other parameters $\alpha = 10$ and $\beta = 0.01$ in B-D. (B) The dependence of signal adaptation sensitivity on N and γ. The colors represent the value of sensitivity. The arrows indicate the corresponding transition structures in cancer (increase of both N and γ) and embryo (increase only in γ) respectively. (C) The dependence of noise attenuation property on N and γ. The colors represent the CV of output M population dynamics. (D) The dependence of transition efficiency on N and γ. The colors represent the value of efficiency.

$$\frac{\mathrm{dI}_N}{dt} = \gamma I_{N-1} + \beta\gamma M - \gamma(1+\beta) I_N \tag{3.4}$$

$$\frac{dM}{dt} = E + \gamma I_N - (\alpha + \beta\gamma) M \tag{3.5}$$

The initial conditions of ODEs are set as $E(0) = 1$, $M(0) = I_k(0) = 0$ to assume only E cells initially. To tackle the stiffness problem introduced by large N or γ, we called ODE15s solver in Matlab to evolve the dynamical systems.

To study noise attenuation, we add the persistent white noise term to epithelial dynamics, equation 3.1 to simulate the extrinsic fluctuation, i.e. we modify the dynamics as stochastic differential equation (SDE)

$$d\tilde{E}(t) = \left[\alpha\tilde{M}(t) + \beta\gamma\tilde{I}_1(t) - (1+\gamma)\tilde{E}(t)\right] dt + \sigma dW_t \tag{3.6}$$

where W_t is the standard Wiener process with $\mathbb{E}W_t = 0$ and $\mathbb{E}W_t = 0$ and σ represents the noise amplitude, which is set as 1 in our simulation. We use Euler-Maruyama scheme to simulate system described by equation 3.2-3.6.

The mesenchymal population fraction $M(t)$ potentially measures how the EMT process adapts or responds to extrinsic signals or fluctuations, as well as the efficiency of transition from epithelial to mesenchymal cells. To quantify the three properties, in a model with N intermediate states we define adaptation sensitivity (AS), noise attenuation (NA) and transition efficiency (TE) as

$$AS_N = \frac{\max_t M(t) - M(+\infty)}{\max_t M(t)}$$
$$NA_N = \frac{\mathrm{std}[\tilde{\mathbf{M}}(t)]}{\mathrm{mean}[\tilde{\mathbf{M}}(t)]}, TE_N = M(+\infty)$$

where $\tilde{\mathbf{M}}(t)$ denotes the mesenchymal population in the stochastic ODEs. The reliance of AS_N, NA_N and TE_N on N and γ are investigated to study different EMT lineage structures and role of ICS in population-survival. We explore the AS, NA and TE as the functions of

key parameters N and γ (Figure 3.1B-D). From the single-cell data analysis, the embryonic EMT is associated with an increase of γ, while in cancer EMT there is a simultaneous increase of N and γ.

Roles of ICS in adaptation

When the ICS does not exist in the system, the dynamics of M population can be solved explicitly as $M(t) = \frac{1}{1+\alpha}\left(1 - e^{-(1+\alpha)t}\right)$, which is a monotonic function of time. Therefore, the adaptation sensitivity is zero in the two-state system. Generally, in the linear system equation 3.1-3.5 with N ICS, the solution can be expressed as $M(\mathrm{t}) = C_0 + \sum_{k=1}^{N+1} C_k e^{\lambda_k t}, \mathrm{Re}\,(\lambda_k) < 0$. When the eigen-values $\lambda_k t$ are real and C_k have different signs, there could exist local maximums of $M(t)$ trajectory, resulting in the non-zero adaptation sensitivity. Meanwhile, if the eigen-values $\lambda_k t$ are complex, we even can have the oscillatory trajectory of $M(t)$ before it reaches stationary state. Through numerical simulation, we validate that the adaptation sensitivity will increase with N when keeping other parameters as constant (Figure 3.1B).

3.4 Results

By applying QuanTC to twelve published single-cell datasets during embryogenesis or cancer, we reveal the common cell lineage structures mediated by the ICS. We finally model such cell lineages (Figure 2.1B) to investigate similarity and difference of identified cell lineages in terms of signal adaptation, noise attenuation and EMT transition.

3.4.1 A near synchronous EMT though one ICS during embryonic stem cell differentiation

Previous studies revealed a global epithelial–mesenchymal–epithelial transition during the hepatic differentiation of human embryonic stem cells (hESCs) [102]. Recently, a single-cell qPCR analysis with 48 selected genes was performed to study this process [101]. In this dataset, cells from day 0 are all epithelial cells in a pluripotent state while cells at day 3 are definitive endoderm (DE) cells in a typical mesenchymal-like status. Cells from day 0 to day 3 are found to follow a near synchronous EMT.

We applied QuanTC to the dataset of 345 cells from day 0 to day 3, identifying three clusters (Figure 2.3F). Two clusters are E (high expression of pluripotent marker gene *SOX2*) and M (high expression of DE marker genes *FOXA2* and *GATA6*) whereas the other expresses both epithelial marker gene *CDH1* and DE marker gene *FOXA2* (Figure B.4), named as intermediate state I.

Next we quantified the transition dynamics of EMT in embryonic stem cell differentiation using QuanTC. We found that the cells located around the overlapping space between clusters have higher CPI values, while cells closer to cluster centers have lower CPI value (Figure B.4A). More TC with higher CPI values locate around the identified state I, suggesting that the I state has high potential to make transitions to both E and M (Figure 2.3G). The transition trajectory from E to M via I state includes 99.7% of total cells, indicating that the ICS-mediated path dominates the cell transitions during EMT.

The cells in early pseudotime were found to be the same ones in early real time (Figure 2.3H), suggesting the transition from day 0 to day 3 follows a near synchronous EMT, a result consistent with the experimental observations on differentiation of hESCs to hepatic lineage [101].

Novel transition genes and marker genes of the three states were identified (Figure B.4B-C). *MIXL1*, the marker of DE, is identified as a transition gene from E-I, because its expression level increases gradually during E-I transition (Figure B.4D). Two pluripotency markers, *POU5F1* and *NANOG*, and other genes sharing similar expression profiles are transition genes of I-M because of the observed gradual decrease from I to M.

For this dataset, QuanTC not only captures the synchronous EMT but also detects ICS that express both E and M markers. The ICS identified by QuanTC shows strong transition dynamics and ICS-regulated path dominates the cell transitions during EMT.

3.4.2 Multiple ICS found in mouse skin tumor dataset

To study epithelial-to-mesenchymal transition in cancer [118, 122], we applied QuanTC to a skin squamous cell carcinoma (SCC) dataset, in which multiple tumor subpopulations associated with different EMT stages were identified, and some of them displayed hybrid phenotypes that likely represent multiple distinct ICS *in vivo* [121]. This dataset of 382 cells on skin tumors contains FACS-isolated epithelial YFP+Epcam+ tumor cells, which are relatively homogeneous, and mesenchymal-like YFP+Epcam- tumor cells, which are more heterogeneous [121].

Four clusters were identified by QuanTC, showing two clusters are clearly E and quasi-mesenchymal (QM) states (Figure 3.2A and Figure B.5, B.6) and the two other clusters, labeled as I1 and I2, express both epithelial marker gene Dsp and mesenchymal marker gene *Vim*. Nearly all epithelial YFP+Epcam+ cells were found in the E state while most mesenchymal-like cells were clustered into I1, I2, or the QM state. The remaining mesenchymal-like cells were clustered into E but closer to I1, similar to the I1 cells. The overall cell distributions in four different states are very much consistent with the previous observed Epcam+ and Epcam- cells in their levels of heterogeneity [121].

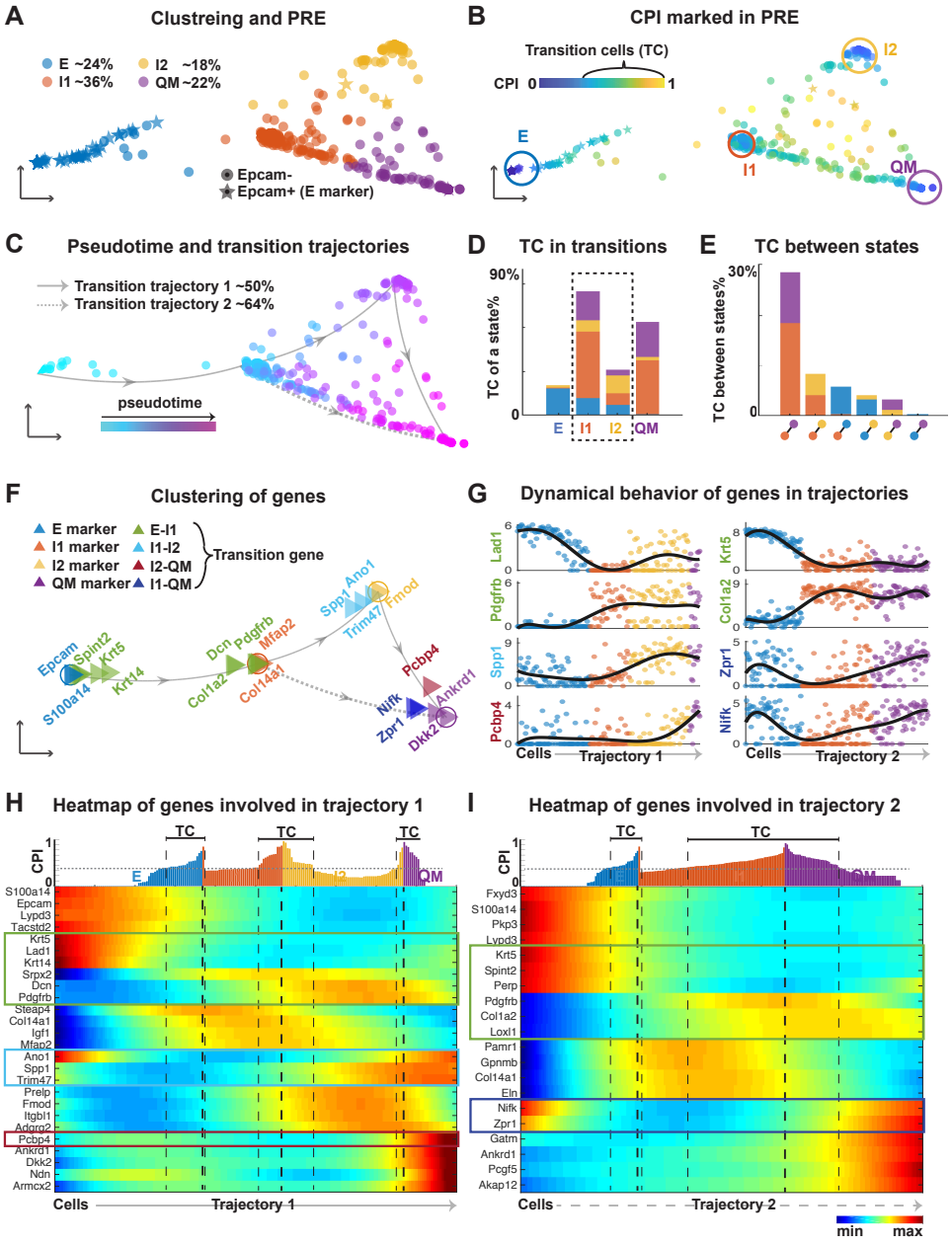

Figure 3.2: Analyzing EMT in mouse skin squamous cell carcinoma (SCC) dataset using QuanTC. (A-C) Visualization of cells via PRE.

Figure 3.2 *(continued)*: (A) Each star or solid circle colored by the corresponding cell state represents one of the 67 epithelial YFP+Epcam+ and 292 mesenchymal-like YFP+Epcam- tumor cells. (B) Identification of TC. Each dot is colored by its CPI value. The cells outside circles with relatively high CPI values are considered as TC. The parameters are given in Table B.1. (C) Transition trajectory inference. Arrowed solid and dashed lines show two main transition trajectories, with cells colored based on their pseudotime. (D) Percentage of TC associated with each state relative to the total number of TC. (E) Percentage of TC between two states relative to the total number of cells. (F) Visualization of marker genes and transition genes between states. Each triangle represents a gene colored by its type and arrowed lines indicate the transition direction of EMT. (G) Expression levels of top transition genes with cells ordered along the two most probable transition trajectories. Solid lines, smoothed expression curves for each gene in the transition trajectory. (H-I) Heat map of normalized expression of marker genes and transition genes. Columns represent cells ordered along the transition trajectory and rows represent genes. Coloring represents the normalized expression value of each gene. Transition genes are marked in the box. Top: CPI values of each cell along the transition trajectory.

Novel transition trajectories from E to QM were revealed according to the locations of TC (Figure 3.2B). There are two main transition trajectories: E-I1-I2-QM and E-I1-QM, which consist of 94% of cells (Figure 3.2C). This suggests the two most probable transition trajectories from E to QM both pass through ICS. The I1 and I2 states, consisting of TC from all the other states around them (Figure 3.2D), show strong capability of making transition – a nature property of cells in intermediate cell state. The transition between I1 and QM was found to have most TC (almost 30% TC in total) followed by the transition between I1 and I2 (Figure 3.2E).

The identified marker genes of E (Figure 3.2F-I) have a broad agreement with known markers of epithelial cells [119] (Figure B.6), with their levels of transition genes varying significantly during transition. For example, *Lad1* decreases gradually and *Pdgfrb* increase gradually as E cells transition to I1.

Using QuanTC we identified new marker genes for ICS, with some of them shown to have special functions in EMT via separating ICS from the mesenchymal-like states. For example, *Igf1* and *Mfap2*, differentially expressed in I1 state, have been shown to induce EMT in

hepatocellular carcinoma and in gastric cancer cells respectively [181, 159]. As a result, ICS can be identified not only via co-expression of epithelial and mesenchymal markers but also through specific ICS markers.

The two ICS, I1 and I2 states, are indeed distinct cell states based on the Gene Ontology enrichment analysis of the top marker genes of I1 and I2 states. Both I1 and I2 states share similar biological processes including cell migration and cell motility (mesenchymal features), in addition to proliferation and cell-to-cell communications (Table S2 in [140]). The ability of regulating cell communication and signaling is uniquely found for ICS. I1 state not only has all the biological processes included in I2 state but also has the unique biological processes related to cell adhesion that shares with the epithelial cells. This suggests that the cells in ICS display hybrid epithelial/mesenchymal features [74] as well as communicates with other cells through cell signaling [17, 178].

3.4.3 EMT via ICS during mouse embryonic development

scRNA-seq datasets were collected for four organs and tissues of E9.5 to E11.5 mouse embryos: skin (155 cells), lung (176 cells), liver (123 cells), and intestine (173 cells) [39]. Applying QuanTC to the four datasets, three clusters were observed for each dataset (Figure 3.3). Based on the known cluster labels of epithelial and mesenchymal cells [39] and the marker genes inferred by QuanTC, two clusters are clearly E and M cells (Figure 3.3 and Figure B.7, B.8). The remaining cluster is located between E and M, with more TC of higher CPI values around it, showing clear characteristics of ICS. The cells close to the I state matches the known labels well, exhibiting mixture of features of epithelial and mesenchymal cells [39].

In the four datasets, $> 86\%$ cells were found to be involved in the newly discovered E-I-M transition trajectory, suggesting most cells undergoing EMT via the intermediate cell state instead of direct transition from E to M (Figure B.7, B.8A, G). Except for skin having only

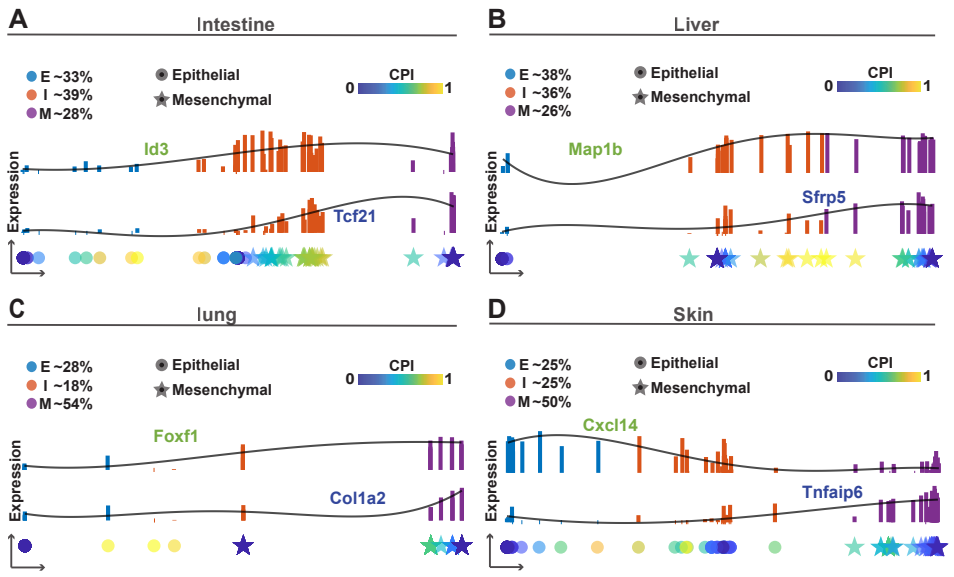

Figure 3.3: Comparison analysis of EMT during organogenesis in intestine, liver, lung and skin. (A-D) Top: the expression levels of E-I transition genes (green) and I-M transition genes (blue) along the E-I-M transition colored by inferred state of cells. Solid lines are smoothed expression curves for each gene in the transition trajectory. Bottom: Cells are ordered along a line according to their pseudotime values. Each dot represents a single cell shaped by the cell states previously identified in the original study on the corresponding dataset and colored by the CPI value. The parameters are given in Table B.1.

a few more TC in E-I than I-M transition, the other three have significantly more TC in I-M transition than E-I transition (Figure B.7, B.8D, J). This observation suggests that I and M states are potentially more similar to each other whereas E could be a distinct state.

Gene Ontology enrichment analysis of the top marker genes (Table S2 in [140]) indicates that the ICS from intestine and liver share several biological processes, including cellular component movement, cell motility and cell migration (mesenchymal features), cell adhesion (epithelial features), regulation of signal transduction and cell communication. The ICS from lung and skin relate to the mesenchymal and epithelial cell differentiation. Interestingly, the transition genes inferred from the four organs or tissues are quite different (Figure B.7, B.8), indicating that genes regulating EMT may vary under different conditions at different developmental stages.

3.4.4 Comparisons with another state transition method and inference of gene regulatory networks

To further investigate the transition in EMT and validate QuanTC, we next used a previously developed state transition index I_c to predict transitions based on a different method that uses correlated information between cells and genes [111]. The index I_c serves as an early warning signal of a critical transition that coincides with lineage commitment [111]. By evaluating I_c for all five datasets, we found nearly all TC identified via QuanTC admit higher I_c than the cells in the stable states (Figure 3.4A), consistent with the observation that TC are the cells involved in the transition process. The relatively low cell-cell correlation and high gene-gene correlation (Figure B.9A) during state transitions correspond to the idea that the state transition involves a decrease of cell-cell correlation and concomitant increase of gene-gene correlation. One exception happens for the E-I trajectory in lung, partly due to a very small number of TC cells (only 3 cells) identified between E and I.

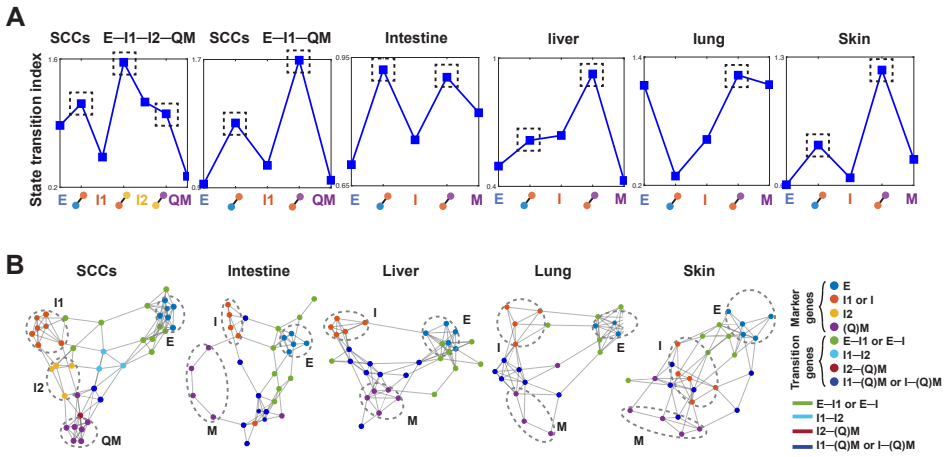

Figure 3.4: State transition index and gene regulatory networks for five EMT datasets and their comparisons with QuanTC outputs. (A) State transition index of relatively stable cells in each state and the TC between states. Dashed box: TC with high value of state transition index. (B) Gene regulatory networks of top marker genes and transition genes using the PIDC algorithm from the SCC and mouse embryonic development datasets (the top $\sim 80\%$ of edges are shown). The parameters are given in Table B.1. Each dot represents a gene colored by its type. Each large dashed circle labels marker genes of a particular cell state. Graph edges indicate the top interactions and the length of the edge is inversely proportional to the interaction strength between genes.

To investigate how transition genes may regulate state marker genes in EMT, we inferred gene regulatory networks of both state marker genes and transition genes via the PIDC algorithm [32]. The inferred markers of different states were projected into lower-dimensional space, with top genes marked by their states or transition trajectories and the edge length, which is inversely proportional to the interaction strength between genes (Figure 3.4B and Figure B.9B). Two genes that are close to each other with a short edge indicate a strong regulatory interaction, in contrast to genes located away from each other with a longer edge between them.

For example, in the SCC dataset, E markers are mostly linked to I1 markers through E-I1 transition genes, and marker genes of I1 and I2 are linked directly or via I1-I2 transition genes, showing a gene regulatory circuit consistent with the inferred trajectory and CPI values using QuanTC (Figure 3.2B-C). In addition, marker genes of I2 and QM are linked directly or via I2-QM transition genes along with an edge linking markers of I1 and QM to I1-QM transition genes nearby, suggesting that E-I1-QM is another transition trajectory, consistent with the two previously inferred trajectories (Figure 3.2C). Interestingly, markers of E have longer edges linking to other marker genes, suggesting the relative dissimilarity of E to I1, I2 and Q, consistent with our findings directly using QuanTC (Figure 3.2). Similar structures in gene regulatory networks were seen among the intestine, liver and lung. In particular, marker genes of E, I and M form distinct groups and markers of E and I are linked directly or via E-I transition genes, while markers of I and M are linked directly or via I-M transition genes. Interestingly, for skin, different markers are much less separated compared to other three embryonic development systems, except for markers of E, suggesting the transitions and the genes regulating the transition in developing skin could be more intermingled and complicated.

3.4.5 Dynamical properties of inferred ICS-regulated EMT trajectories

To explore the dynamics of the inferred transition trajectories, we developed a cell population model that contains multiple ICS and only relies on three effective dimensionless parameters (Materials and Methods, Figure 3.1A). Subsequently, three emergent quantities were then defined to measure the EMT population dynamics (Figure 3.5A, Materials and Methods): 1) sensitivity of signal adaptation, 2) coefficient of variance (CV) to quantify noise attenuation and 3) the efficiency of population transition from epithelial to mesenchymal states. We then investigated how the existence of ICS, as well as the transitions via ICS, affect the robustness and efficacy of EMT dynamics using these three quantities.

The signal adaptation property is demonstrated by the reset of output level after the response to stimulus in cell populations (Figure 3.5A). In cancer EMT, adaptation with high sensitivity permits the transient peak of the massive release of malignant mesenchymal population, forming the effective metastasis strategy under the immune regulation. In the two-state system with only pure epithelial or mesenchymal states, we rigorously proved that no adaptation is allowed (Materials and Methods). The modeling results suggest that both the increase in ICS number and the moderate increase in indirect transition rate (ITR) via the ICS (Figure 3.1B, Materials and Methods) can increase the adaptation sensitivity (Figure 3.5C), however, further increase in ITR (over a certain threshold) can instead decrease the sensitivity. Interestingly, the increase in ICS number may result in the oscillatory adaptation of cell population dynamics, i.e. the M population goes through multiple peaks before reaching a steady level (Figure 3.5B). This potentially provides a "hide-and-seek" strategy for metastatic mesenchymal cells battling with immune systems in cancer.

The noise attenuation property depicts the system's capability to reduce fluctuations in population dynamics. Both the increase in ICS number and ITR help reduce the CV of

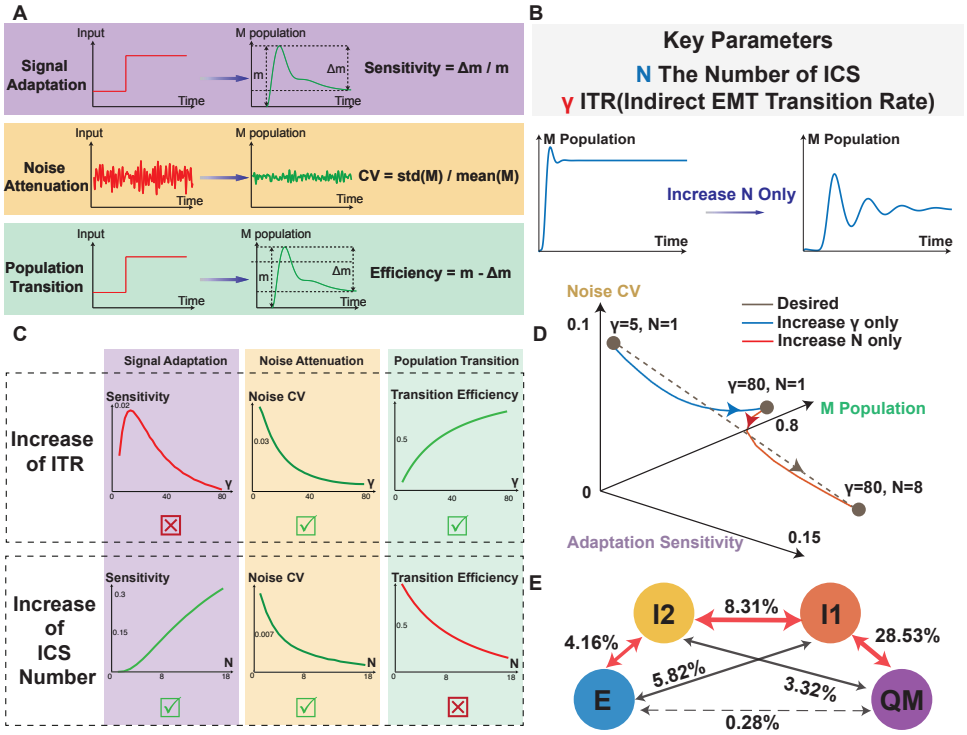

Figure 3.5: Dynamical properties of inferred ICS-regulated EMT trajectories. (A) The definitions and measurements of three quantities – adaptation, noise attenuation and population transition properties of cell population dynamics. (B) The key parameters of model including ICS number N and ITR gamma (see also Materials and Methods, Figure 3.1). Increase of ICS number N can result in the multiple peaks in M population trajectory, forming the oscillatory adaptation. (C) Effect of tuning N and gamma on the three quantities (see also Figure 3.1). (top row) Changes in three quantities by fixing N=2 and tuning gamma from 5 to 80. The increase in ITR gamma lowers the noise coefficient of variance (CV) of output M population, and increases the transition efficiency from E to M. The signal adaptation sensitivity is not a monotonic function of gamma, which reaches the peak before a certain threshold and declines afterwards with further increase in gamma. (bottom row) Change of three quantities by fixing gamma and tuning N from 1 to 18. The increase in N improves adaptation sensitivity and noise attenuation, however reducing the value of transition efficiency. (D) Tuning parameter gamma and N separately cannot achieve all the desired properties (i.e. simultaneous increase of adaptation sensitivity, noise attenuation and EMT efficiency, indicated by brown dashed line). The desired properties can be achieved by increasing ITR gamma (blue line, increase gamma from 5 to 80 and fix N as 1) first and increasing N subsequently (red line, increase N from 1 to 8 and fix gamma as 80). (E) EMT trajectories inferred from SCC dataset, with node colors consistent with Figure 3.2. Other inferred trajectories are shown in Figure B.9 and Figure B.10. The arrow represents potential transition between states, and number represents the percentage of TC. The red arrows indicate the major transition trajectory mediated by ICS, and the dashed arrow refers to the direct transition route from E to QM state.

M population trajectories (Figure 3.5C), stabilizing the dynamics in population transition. The property of population transition is quantified by the final fraction of M population that originates from pure E population. The increase in ITR results in boosting of population transition efficiency in EMT, while the increase in ICS number reduces such efficiency.

The trade-off between adaptation sensitivity and transition efficiency were observed in EMT (Figure 3.5C-D). Although larger ICS numbers may increase adaptation sensitivity, it also impairs the effective transition toward M state (Figure 3.5C). On the other hand, increasing ITR can boost efficiency while the overly-large value results in a decrease in adaptation sensitivity. Hence, an increase in one parameter only, either ICS number or ITR, fails to optimize all the properties simultaneously (Figure 3.5D). The transition trajectories may need a combined increase in both ICS and ITR to achieve the desired property, as seen in the inferred SCC transition trajectories (Figure 3.5E).

The derived relationships between three emergent quantities and EMT population parameters shed light on our findings obtained from single-cell EMT data mining. Based on the percentages of TC between state transitions among all the cells involved in EMT, we quantified the EMT trajectories in twelve single-cell datasets by QuanTC (Figure 3.5E and Figure B.9C, B.10B), which include six additional head and neck squamous cell carcinoma (HNSCC) datasets (Materials and Methods, Figure B.10). For all the investigated mouse and human datasets from both normal and tumor tissues, we found that the majority of transitions involve ICS while the direct transition between epithelial and mesenchymal states is relatively rare (FFigure B.9C, B.10B). This corresponds to the increase in ITR in the model, resulting in the strengthening of noise attenuation property (Figure 3.5C), as well as enhancement of adaptation sensitivity (provided that increase in ITR does not over-exceed the observed threshold in Figure 3.5C). Besides, compared to only one ICS involved in EMT in embryo, cancer EMT has more numbers of ICS. Therefore, in cancer EMT the adaptation sensitivity of population dynamics is further enforced by the presence of multiple ICS, with

sacrifice of E-to-M transition efficiency. In comparison, in embryogenesis EMT fewer ICS and the large ITR flux can lead to higher E-to-M transition efficiency, however, at the cost of lower sensitivity of population dynamics adaptation.

3.5 Discussion

By unsupervised learning of transition trajectories in twelve EMT single-cell datasets and multiscale mathematical modeling, we have analyzed transition cells and dynamics of EMT that highlights the transition trajectories mediated by ICS. By investigating several emergent dynamic quantities of describing transitions, we have suggested that the inferred transition trajectories not only attenuate the noise, but also enhance the signal adaptation in EMT. Modeling analysis has indicated cancer EMT trajectories strengthen the signal adaptation, whereas trajectories in embryogenesis EMT is in favor of effective population transition toward mesenchymal states.

To compare with other methods, we have applied the popular pseudotime inference method Monocle 3 to the simulation datasets and SCC datasets (Materials and Methods, Figure B.3). While Monocle 3 correctly depicts the overall progression of epithelial-mesenchymal transition, it lacks the resolution to distinguish transition cells from other stable cells. In addition, the trajectories inferred by Monocle 3 strongly depends on input gene selections. Interestingly, the features selected by QuanTC could improve the consistency of trajectory inference by Monocle 3 in SCC dataset (Figure B.3), suggesting usefulness and its broader application of the feature selection function in QuanTC.

QuanTC is adaptive to the downstream analysis of other soft clustering methods and is applicable to systems beyond EMT. For instance, we applied QuanTC to a single-cell RNA-seq dataset of 2,000 mouse hematopoietic progenitors (Materials and Methods, Figure B.2). We

found two prominent non-ICS, i.e. plasmacytoid dendritic cells (pDCs) and B cells, exactly corresponding to the target states identified in the original study [54]. The transition cells along the trajectory indicates that pDCs and B cells share the same progenitors, consistent with the findings based on the FateID inference [54].

Interesting trade-offs among signal adaptation, noise attenuation and effective transition have been observed in modeling analysis. Consistent with previous findings [145], the increase in ICS number during EMT attenuates fluctuations; in addition, boosting the transitions via ICS (i.e. ITR) also plays the similar role in noise buffering. The concept of adaptation sensitivity, previously mainly used for signal transductions [106, 124], was introduced in this study to quantify the transient, adaptive dynamics in EMT populations. Such transient property were previously reported in breast cancer cell lines [49], and theoretically studied in the context of non-equilibrium statistical physics. Interestingly, the increase of ITR alone cannot improve adaptation persistently, and the robust adaptation in population dynamics requires both large ITR and multiple ICS, a result consistent with the learned single-cell trajectories in SCC. We reason that the transient peaks in highly-adaptive trajectories ensure adequate release of mesenchymal cells, with the short-lasting times impeding immune systems to efficiently capture and respond timely to metastasis. It is very interesting to note that ICS in EMT are associated with poor prognosis of cancer treatment according to clinical studies [120]– our findings between ICS number and adaptation may serve as the potential explanation from cell population dynamics.

Overall, our integrative approach provides an initial attempt to bridge single-cell data mining and multiscale modeling to investigate transitions and role of intermediate cell states in EMT.

Chapter 4

Inference of Intercellular Communications and Multilayer Gene-Regulations of Epithelial–Mesenchymal Transition From Single-Cell Transcriptomic Data

This chapter is a reprint of the material as it appears in [139]. The co-authors listed in this publication directed and supervised research which forms the basis for this chapter.

4.1 Background

Epithelial to mesenchymal transition (EMT) plays an important role in many biological processes during development and cancer. The advent of single-cell transcriptome sequencing

techniques allows the dissection of dynamical details underlying EMT with unprecedented resolution. Despite several single-cell data analysis on EMT, how cell communicates and regulates dynamics along the EMT trajectory remains elusive. Using single-cell transcriptomic datasets, here we infer the cell-cell communications and the multilayer gene-gene regulation networks to analyze and visualize the complex cellular crosstalk and the underlying gene regulatory dynamics along EMT. Combining with trajectory analysis, our approach reveals the existence of multiple intermediate cell states (ICS) with hybrid epithelial and mesenchymal features. Analysis on the time-series datasets from cancer cell lines with different inducing factors show that the induced EMT are context-specific: the EMT induced by TGFB1 is synchronous while the EMT induced by EGF and TNF are asynchronous, and the response of TGF-β pathway in terms of gene expression regulations are heterogeneous under different treatments or among various cell states. Meanwhile, network topology analysis suggests that the ICS during EMT serve as the signaling in cellular communication under different conditions. Interestingly, our analysis of a mouse skin squamous cell carcinoma (SCC) dataset also suggests regardless of the significant discrepancy in concrete genes between *in vitro* and *in vivo* EMT systems, the ICS play dominant role in the TGF-β signaling crosstalk. Overall, our approach reveals the multiscale mechanisms coupling cell-cell communications and gene-gene regulations responsible for complex cell-state transitions.

4.2 Introduction

Epithelial to mesenchymal transition (EMT) is a biological process where epithelial cells lose cell-cell adhesion and gain some mesenchymal traits of migration and invasion [75, 77]. EMT not only occurs widely during normal embryonic development, organ fibrosis and wound healing, but also plays an important role in tumor progression with metastasis [118, 91].

Recent studies have underscored that EMT is not a binary process, but instead exists on

a spectrum with various hybrid states ranging from epithelial to mesenchymal phenotypes [118]. Cells undergoing EMT can display mixed epithelial and mesenchymal features, and are considered in the intermediate cell states (ICS) [70, 138, 61]. In the context of cancer progression, these ICS have been proposed as the main drivers of metastasis due to their ability to undergo collective cell migration as highly metastatic multicellular clusters [70]. Therefore, understanding the features and role of ICS during EMT could potentially unlock novel clinical strategies. With the unprecedented opportunities brought by single-cell RNA sequencing (scRNA-seq), the existence of multiple ICS and their transcriptomic profiles has been observed and analyzed via pseudotemporal ordering or energy landscapes [126, 121, 68, 99, 34, 2]. Very recently, specially designed methods have also been proposed to infer EMT trajectories or transition paths from the single-cell transcriptomic [140] or imaging data [162]. The integrative analysis combining unsupervised learning of single-cell transcriptomic data and computational modeling of EMT in cancer and embryogenesis successfully uncovered the novel roles of ICS on adaption, noise attenuation, and transition efficiency [140]. While these methods have provided insights into the dynamics of EMT from a single cell perspective, the role of intercellular communication in EMT remains largely unknown.

Indeed, EMT is not necessarily a cell autonomous process. Cells secrete, and in turn respond to various growth and differentiation signaling factors secreted by other cells in the extracellular environment, including transforming growth factor β (TGF-β), WNT and Notch proteins [169, 113, 17, 19]. Among them, the well-characterized TGF-β pathway has received much attention as a major inducer of EMT during embryogenesis, cancer progression and fibrosis [169, 166]. The TGF-β pathway can also cross-talk with other pathways such as WNT and SHH [168], forming the complex response of signaling. In addition, signaling in cell-cell communications have also been found important in the formation and regulation of ICS (e.g. through Notch pathway) [21]. This intercellular communication has been shown to play significant roles in regulating gene expression dynamics within individual cells, through analysis of scRNA-seq datasets from several development and cancer systems

[123, 27, 177, 90, 160]. Computational methods have been developed to infer cell-cell communication networks based on ligand-receptor interactions [161, 26, 163, 67], and elucidate how cell-cell communications propagate to downstream target genes through transcription factors [24]. While methods have been developed to infer EMT gene regulatory network from RNA-seq single cell data [130], the role of cell-cell communications on gene regulation dynamics along EMT trajectory is poorly understood.

Through both experimental and mathematical modelling studies, the key circuits of EMT involving few epithelial/mesenchymal markers, transcription factors and signaling molecules have been summarized [168, 170, 154, 41, 55, 100, 79]. Due to different roles of nodes, the circuits can be modelled as a multilayer network [83] with hierarchical structures [24]. In the multilayer network, cells communicate with each other and the environment via signal transduction pathways (Layer 1), which directly targets the downstream factors or genes (Layer 2) that subsequently regulate the expression of marker genes of various cell states (Layer 3). In addition, there may be dynamical changes of network structure during EMT, where the temporal (or pseudo-temporal) information constitutes another independent dimension of the layer sets. The complex interactions among nodes may exist within the same layers or across different layers, in controlling EMT. Here we study the time-series scRNA-seq datasets of OVCA420 cancer cell line exposed to various EMT-inducing factors [37]. We first delineate the underlying transition details at individual cell resolution with a recently developed method, QuanTC. For the cancer cell lines undergoing EMT under three different treatments, we quantify the ICS-regulated trajectories and detect the driver genes in EMT for each case, respectively. While cells undergo TGFb1-driven EMT in a highly synchronized fashion, EMT guided by EGF and TNF is asynchronous. Next, we develop a multilayer network approach to infer and visualize the hierarchical interactions that combine cell-cell communications through the TGF-β pathway, signal transductions and gene regulatory networks from single-cell transcriptomic data. After trajectory inference, we then utilize the multilayer network approach to decipher the role of TGF-β pathway in regulating EMT

dynamics with different inducing factors. We also compare the results of *in vitro* cancer cell lines with further analysis of *in vivo* mouse skin squamous cell carcinoma (SCC) dataset [121].

4.3 Materials and methods

4.3.1 scRNA-Seq data clustering and transition trajectory reconstruction

QuanTC was used to perform clustering and transition trajectory reconstruction. QuanTC can simultaneously detect the ICS and construct transition trajectories via quantifying the cell plasticity index (CPI) [140]. The cells with higher CPI values are considered to be transitioning between clusters and are identified as transition cells (TC). Via non-negative matrix factorization, QuanTC calculates the probabilities of a given cell belonging to the identified clusters. Cells are projected to a low-dimensional space based on a probabilistic regularized embedding. The transition trajectories are then inferred by summing the cluster-to-cluster transition probabilities that are calculated from cell-to-cluster probabilities and TC between clusters. The transition genes and marker genes of clusters are obtained through factorizing the gene expression matrix as product of cell-to-cluster probabilities and likelihoods of genes uniquely marking each cluster. In the first step of QuanTC, we applied two additional considerations when choosing the number of identified clusters. First, we know from the original experiment that cells undergo EMT (i.e. there is at least one E state and one M state); furthermore, given that we seek to study intermediate cell states during EMT, we search for at least 3 total states.

Preprocessing

Single cells with less than 95% expressed genes among all detected genes were considered as low-quality cells and were filtered. Top 3000 bimodal distributed genes were selected by QuanTC with default parameters to do downstream analysis.

Clustering

A total of 3000 selected genes and 558 cells of OVCA420 induced by TGFB1, 1137 cells of OVCA420 induced by EGF, 1856 cells of OVCA420 induced by TNF from day 0 to day 7 were retained for clustering. Consensus clustering via SC3 [82] was performed on the expression matrix to capture the cell-cell similarity. The clusters were defined based symmetric non-negative factorization as wrapped in QuanTC.

Transition trajectory

The beginning and end of EMT transition trajectory, E/M states, were identified based on the percentage of TC around each cluster. The parameters to choose TC were given in Table C.1. The clusters with fewer TC around were considered as E/M states while the rest clusters were considered as ICS along EMT. The E/M states between the two clusters were then identified based on the canonical epithelial and mesenchymal marker genes. The potential transition trajectory was inferred according to the TC between clusters using "traj" function wrapped in QuanTC. The pseudotime value of each cell was then computed by QuanTC based on the two most probable trajectories.

EMT marker genes

The marker genes and transition genes were defined using "markers" function wrapped in QuanTC.

GO analysis

The analysis of gene ontology biological processes was performed by Metascape [184] on the top 50 markers genes of each ICS selected by QuanTC.

4.3.2 Qualitatively characterizing cell-cell communications

SoptSC [161] was used on the datasets without gene filtering to calculate the probability matrix of signals being passed between cells and clusters. Signaling probabilities between cells are defined based on weighted co-expression of signaling pathway activity in sender-receiver cell pairs. With the input of ligand-receptor pairs and target genes (upregulated or downregulated in response to pathway activation), SoptSC computes signaling probabilities between sender cells (expressing ligands) and receiver cells (expressing receptors and exhibiting differential target genes activity). Intuitively, given a ligand-receptor pair for a specific signaling pathway, if the ligand is highly expressed in cell i, the cognate receptor is highly expressed in cell j and the target gene activity in cell j suggests that the signaling pathway may have been activated in this cell, then there is a chance that communication occurred between these two cells. The signaling passed from cell i to j for a given ligand-receptor pair is quantified by the signaling probability $P_{i,j}$. For a set of ligand-receptor pairs, SoptSC considers the consensus signaling probabilities between cells by taking the average over all signaling probability matrices. The signaling probability passed from cluster u to cluster v is then given by $P_{u,v} = \frac{\sum_{i \in C_u, j \in C_v} P_{i,j}}{|C_u||C_v|}$ with $|C_u|$ representing the number of cells in cluster u.

The list of ligands, receptors and target genes were retrieved from previous studies [169, 166, 67] and are given in Table 4.1.

Ligand	Receptor	Target genes (up)	Target genes (down)
TGFB1	TGFBR1	FN1 VTN CDH2 COL1A1 COL1A2 MMP2 MMP3 MMP9 TWIST1 TWIST2 IDS ZEB1 ZEB2 SPARC ITGA5 ITGB3 NCAM VIM ACTA2 PLAU DAB2 HIC5 TGFB1I1 HMGA2	OCLN CRB3 ESR1 CD34 CDH1 DSP CLDN1 CLDN2 CLDN3 CLDN4 CLDN5 CLDN6 CLDN7 CLDN8 CLDN9 CLDN10 CLDN11 CLDN12 CLDN13 CLDN14 CLDN15 CLDN16 CLDN17 CLDN18 CLDN19 CLDN20 CLDN21 CLDN22 CLDN23 PKP1 PKP2 PKP3 CK5 CK14 CK8 CK18
TGFB2	TGFBR1		
TGFB3	TGFBR1		
TGFB1	TGFBR2		
TGFB2	TGFBR2		
TGFB3	TGFBR2		
TGFB1	ACVR1		
TGFB2	ACVR1		
TGFB3	ACVR1		
TGFB1	ACVR1B		
TGFB2	ACVR1B		
TGFB3	ACVR1B		
TGFB1	ACVR1C		
TGFB2	ACVR1C		
TGFB3	ACVR1C		

Table 4.1: TGFB pathway used for generating cell-to-cell signaling networks and cluster-to-cluster signaling networks

4.3.3 Measuring node centrality

The centrality of a node (cluster) in cellular communication network is used to quantify its importance in the signaling. We used strength, closeness and pagerank as metrics to measure node centrality. All these centralities were calculated with the package igraph 1.2.4 [38].

Strength is one of the basic measures of centrality: it is measured by summing up the edge weights of the adjacent edges for a given node. Our inferred cluster-cluster communication networks are directed so we calculated in-strength (incoming edges) and out-strength (out-going edges). Closeness of a given node is defined by the inverse of the average length of the shortest path to/from all the other nodes. In-closeness measures the path to the node while out-closeness measures the paths from the node. We used the normalized values to avoid biases based on the network size. Pagerank is proportional to the average time spent

at a given node during all random walks. In the cluster-cluster communication networks, the clusters with high pagerank can be seen as the signaling hub.

4.3.4 Multilayer regulations of EMT

We utilized the multilayer network framework [83] to analyze and visualize the changes of complex hierarchical signaling and gene expression regulations in EMT across multiple scales.

Mathematically, the multilayer network can be expressed as the $M = (V_M, E_M, V, \boldsymbol{L})$. Here, V denotes sets of all nodes in the network (as in the regular case), and $\boldsymbol{L} = \{L_a\}_{a=1}^{d}$ denotes d aspects of the network layers, with each aspect $L_a = \{L_a^i\}_{i=1}^{k_a}$ contains k_a elementary layers. Denotes $\times$ as the Cartesian product of sets, then the node-layer tuple set $V_M \subseteq V \times L_1 \times \cdots \times L_d$ represents all the feasible node-layer combinations in which a node is present in the corresponding layers. The edges set $E_M \subseteq V_M \times V_M$ denotes the weighted links across nodes and layers.

In our context, the nodes set V not only contains cell states $S = \bigcup_{k=1}^{N_c} S_k$ along the EMT trajectories with N_c denoting the number of cell states but also contains target genes T of specified signal transduction pathway and marker genes A of each cell state. The layers $\boldsymbol{L} = \{L_H, L_C\}$ have two aspects: The hierarchy aspect $L_H = \{L_H^1, L_H^2, L_H^3\}$ represents the elementary layers of cell-cell communication L_H^1, target genes L_H^2 and marker genes L_H^3 respectively, and the cell states aspect $L_C = \left\{L_C^k\right\}_{k=1}^{N_c}$ represents the EMT stages of E state, ICS and M state ordered by pseudotime of QuanTC, since we are interested in constructing cell-state specific regulatory relations. For simplicity, we denote the node-layer tuples in EMT as $V_M = \{(S, L_H^1, \cdot), (T, L_H^2, \cdot), (A, L_H^3, \cdot)\} \subseteq V \times L_H \times L_C$, representing the hierarchical regulation structures at different stages. For instance, (A, L_H^3, L_C^1) denotes the marker genes analyzed in the E state, while (T, L_H^2, L_C^2) represents the target genes considered in the first ICS. We next specify how the edges E_M are constructed based on the

V_M.

The edges within layer $(S, L_H^1, \cdot)$

The first layer L_H^1 in hierarchy aspect displays the cluster-cluster interactions of intercellular communication, where the aligned nodes showing the different EMT states/clusters. Using the notations above, $\left(S, L_H^1, L_C^k\right)$ only contains one node for each k, representing the cell state S_k. The weights for the directed edges to connect (S, L_H^1, L_C^i) and $\left(S, L_H^1, L_C^j\right)$ are the cluster-cluster interactions between state S_i and state S_j computed by SoptSC above threshold 0.7.

The edges within layer $(T, L_H^2, \cdot)$

The second layer L_H^2 demonstrates the state-specific interactions among target genes at different stages. The target genes T are the intersection of the list of target genes and the top 3000 selected informative genes. Given the stage L_C^k, the weighted edges between target gene pair $\left(T_X, L_H^2, L_C^k\right)$ and $\left(T_Y, L_H^2, L_C^k\right)$ were constructed by PIDC algorithm [32] using partial information decomposition, only with the cells in cluster S_k. The input to PIDC is an expression matrix with cells from S_k, and the confidence of an edge between a pair of genes is given by $c = F_X\left(U_{X,Y}\right) + F_Y\left(U_{X,Y}\right)$ where $F_X(U)$ is the cumulative distribution function of all the proportional unique contribution scores involving gene X. The top 30% weights were used to embed the inferred network in $\left(T, L_H^2, L_C^k\right)$ using "graph" function in Matlab based on spectral layout [86]. The weights were normalized with max 2 to be comparable with other datasets.

The edges within layer $(A, L_H^3, \cdot)$

The third layer L_H^3 demonstrates the state-specific interactions among marker genes at different stages. The marker genes selected were the identical for $\left(A, L_H^3, L_C^k\right)$ with respect to the choice of k , which represent the union of top five marker genes in each cluster inferred by QuanTC. The edges between marker genes are state-specific for each cell-state layer L_C^k, using the same strategy as for the target genes described above.

The Edges Connecting Layer $(S, L_H^1, \cdot)$ **and** $(T, L_H^2, \cdot)$

These edges quantify the expression of target genes within different states during EMT. The weights for the edges between $\left(S, L_H^1, L_C^k\right)$ and $\left(T, L_H^2, L_C^k\right)$ are the mean expression levels of target genes within cell state S_k, and top 20% weights were shown.

The Edges Connecting Layer $(T, L_H^2, \cdot)$ **and** $(A, L_H^3, \cdot)$

These edges display the regulatory interactions from target genes to marker genes within different states during EMT. The weights for the edges between $\left(T, L_H^2, L_C^k\right)$ and $\left(A, L_H^3, L_C^k\right)$ were inferred by PIDC within cell state S_k, and top 1.5% weights were shown.

4.4 Results

4.4.1 Synchronous EMT with two ICS induced by TGFB1

We analyzed the published datasets [37] with ovarian OVCA420 cancer cell line capable of undergoing EMT. This cell line, which normally shows an epithelial morphology, was exposed to known EMT-inducing factors: TGFB1, EGF and TNF, respectively, to promote

EMT. We used the samples collected at five distinct time points from day 0 to day 7 after the treatment. To compare the process of EMT under three treatments, we used QuanTC [140] to perform the clustering and transition trajectory reconstruction. QuanTC estimates the optimal number of clusters by analyzing the sorted eigenvalues of symmetric normalized graph Laplacian (Figure C.1A). Four clusters were identified in EMT induced by TGFB1 (Figure 4.1A). A first cluster (C3) was mostly composed by cell subpopulations collected at day 0 and 8 hours after induction (Figure 4.1B) and expressed relatively high levels of epithelial markers CDH1 (Figure C.1B). Conversely, a second cluster (C2) consisted of cells collected at day 3 and day 7 (Figure 4.1A-B) and expressed relatively high levels of mesenchymal markers FN1 and SNAI2 (Figure C.1C). Furthermore, cells in these clusters had a low Cell Plasticity Index (CPI). CPI employs an entropy-based approach to estimate cell plasticity, so that a higher index implies a higher probability of transition between clusters (see Methods). Based on the CPI values, QuanTC predicted that clusters C2 and C3 have lower percentages of transition cells (TC) (Figure 4.1C-D), thus suggesting that they are the beginning or end of the trajectory. Based on these observations, we identified cluster C3 as the E state and cluster C2 as the M state.

After choosing the E state, C3, as the beginning of the transition, QuanTC computed the most probable transition trajectory, C3-C4-C1-C2, consisting of 67% of the total cell population (Figure 4.1E). The cluster C4 and C1 were thus identified as intermediate cell state (ICS) I1 and I2, respectively. The marker genes of each state and the transition genes marking the transition between states along the transition trajectory were inferred by QuanTC (Figure C.1D). To characterize the two ICS, I1 and I2, we performed a Gene Ontology biological processes analysis [36] of the top 50 marker genes of each state (Figure C.1E). Both ICS shared similar biological processes including signaling and localization. Furthermore, I2 also related to adhesion and locomotion. This suggested that the cells in ICS displayed both epithelial and mesenchymal features and communications with other cells through cell signaling.

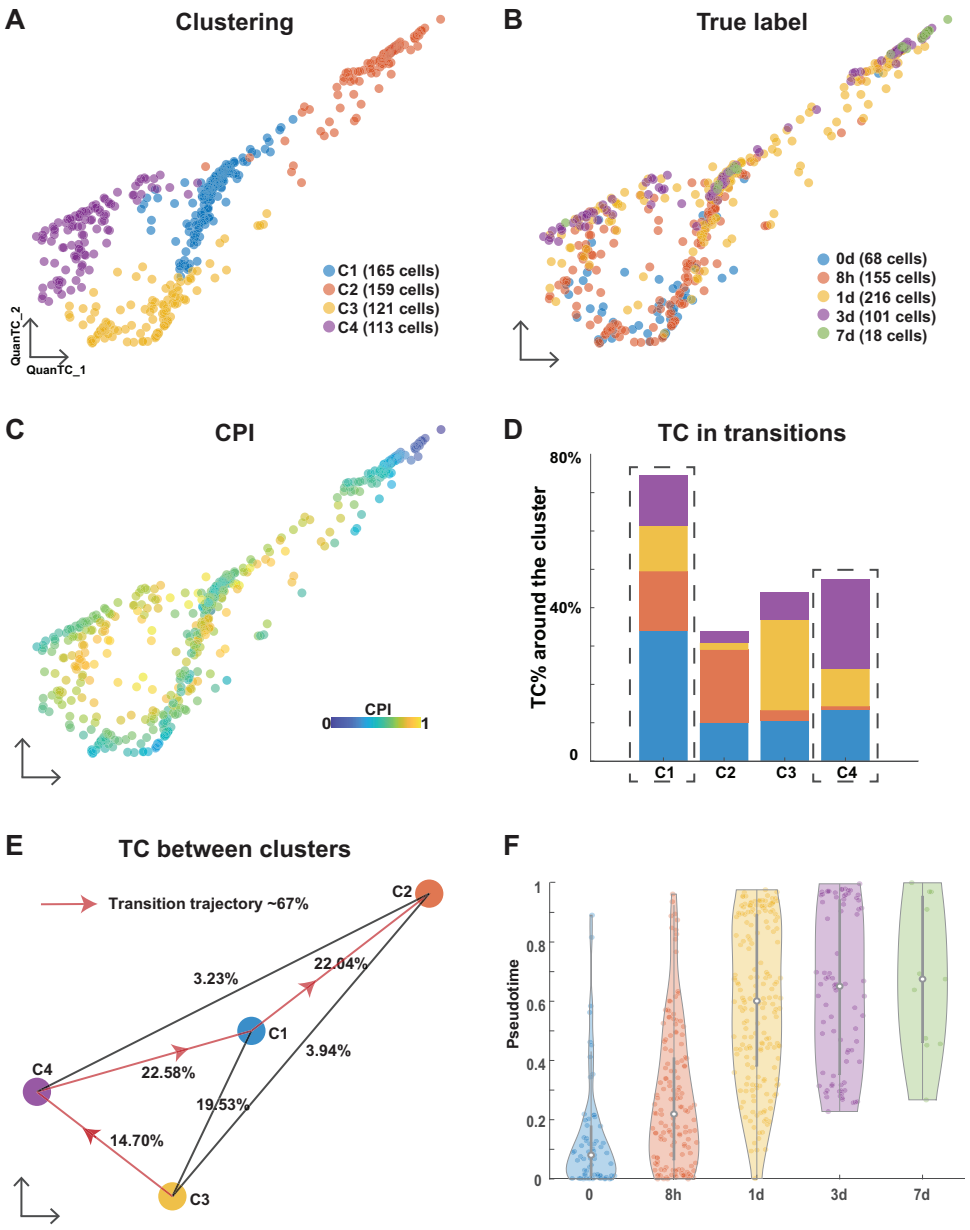

Figure 4.1: Analyzing OVCA420 cancer cell line undergoing EMT induced by TGFB1 using QuanTC. (A-C) Visualization of cells in the two dimensional space by QuanTC. Each circle represents one cell colored by clustering (A), the collection time of the samples after the treatment (B) and CPI values (C). (D) Percentage of TC associated with each cluster relative to the total number of TC. The dashed box covers the ICS having more TC around. The parameters to choose TC are given in Table C.1. (E) Visualization of cluster centers with color consistent with (A). Each percentage on the line show the percentage of TC between two clusters relative to the total number of cells. Arrowed solid line shows the main transition trajectory. (F) Violin plot of pseudotime value of each cell vs the collection time points. Each dot represents a cell colored by collection time points. The circle displays the mean and vertical line shows the interquartile ranges.

Finally, we inspected the population dynamics during TGFB1-driven EMT by considering the pseudotime distribution. Pseudotime quantifies the position of a given cell along the transition trajectory predicted by QuanTC, and therefore does not necessarily correlate with the experiment's physical time. In this time series, however, most cells at t=0 days were characterized by a low pseudotime (i.e. they were positioned toward the beginning of the transition trajectory), while cells at later time points exhibited progressively higher pseudotime values (Figure 4.1F). In other words, OVCA420 cells started from the E state and progressively transitioned throughout the 7 days of EMT induced by TGFB1 in a nearly synchronous fashion.

4.4.2 Asynchronous EMT induced by EGF and TNF

Applying QuanTC to the OVCA420 dataset where EMT was induced by EGF, four clusters were also identified based on the biggest eigenvalue gap after the first two eigenvalues since we want to investigate the ICS during EMT (Figure C.2A, 4.2A). Differently from TGFB1-driven EMT, however, cells collected at different time points co-localized within the same clusters and no group of cells at any given time point dominated any cluster (Figure 4.2B). Based on the CPI values, the two clusters (C2 and C3) were considered as the E and M states based on the fewer TC around them (Figure 4.2C-D). Specifically, C2 was then identified as the E state according to the relatively high expression levels of epithelial markers CDH1 (Figure C.2B), and C3 was identified as the M state because of higher expressions of mesenchymal markers FOXC2 and SNAI2 (Figure C.2C).

The most probable transition trajectory was inferred after choosing cluster C2 as the starting state (Figure 4.2E). The two remaining clusters (C1 and C4) between E and M along the transition trajectory had more TC around them and were identified as I1 and I2, respectively. According to the Gene Ontology analysis of the top marker genes (Figure C.2D), the I2 state

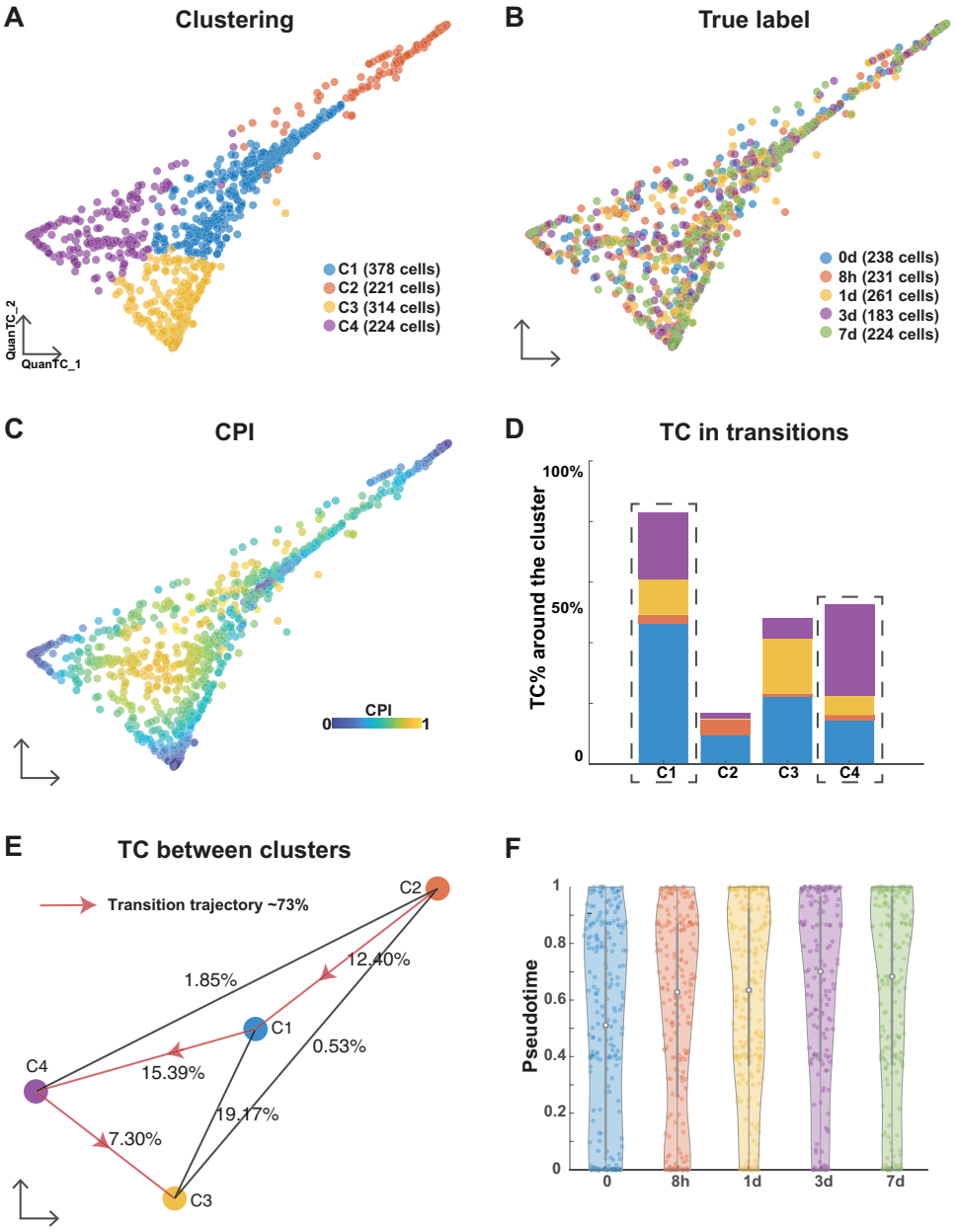

Figure 4.2: Analyzing OVCA420 cancer cell line undergoing EMT induced by EGF using QuanTC. (A-C) Visualization of cells. Each circle represents one cell colored by clustering (A), the collection time of the samples after the treatment (B) and CPI values (C). (D) Percentage of TC associated with each cluster relative to the total number of TC. The dashed box covers the ICS having more TC around. The parameters to choose TC are given in Table C.1. (E) Visualization of cluster centers with color consistent with (A). Each percentage on the line show the percentage of TC between two clusters relative to the total number of cells. Arrowed solid line shows the main transition trajectory. (F) Violin plot of pseudotime value of each cell vs the collection time points. Each dot represents a cell colored by collection time points. The circle displays the mean and vertical line shows the interquartile ranges.

displayed biological processes including adhesion, locomotion and signaling showing mixed feature of both epithelial and mesenchymal cells (Figure C.2E).

The average pseudotime values slightly increased along collection time points, hence demonstrating that the EGF stimulus induces an EMT response. Compared to TGFB1-driven EMT, however, pseudotime distribution within each time point had a high variance, thus indicating that the EMT induced by EGF was more asynchronous (Figure 4.2F). We applied a similar analysis to EMT induced by TNF, and also identified four clusters with two ICS (Figure C.3A, 4.33A). Similar to the case of EGF induction, cells collected at different time points were mixed up in different clusters (Figure 4.3B). After selecting cluster C3 as the E state based on fewer TC around (Figure 4.3C-D) and expression levels of canonical epithelial and mesenchymal marker genes (Figure C.3B-C), the most probable transition trajectories were revealed (Figure 4.3E). Based on the Gene Ontology analysis of the top marker genes (Figure C.3D), the two ICS were different states (Figure C.3E). The I1 state were related to signaling and locomotion indicating the communications with other cells and sharing mesenchymal features.

Similar to EMT induced by EGF, the average pseudotime values slightly increased across time points with high variance within each time point, thus suggesting the heterogeneity of cells undergoing EMT (Figure 4.3F). Therefore, EMT induced by TNF was also found to be an asynchronous process.

4.4.3 Context-specific cellular communications with underlying gene regulations in TGF-β signaling

TGF-β is a strong promoter of EMT [52]. TGF-β ligands are not exclusively provided as an external EMT-inducing signal, but can also be secreted by cells, thus raising the possibility of cell-cell communication and EMT driven by intercellular signaling. In order to determine the

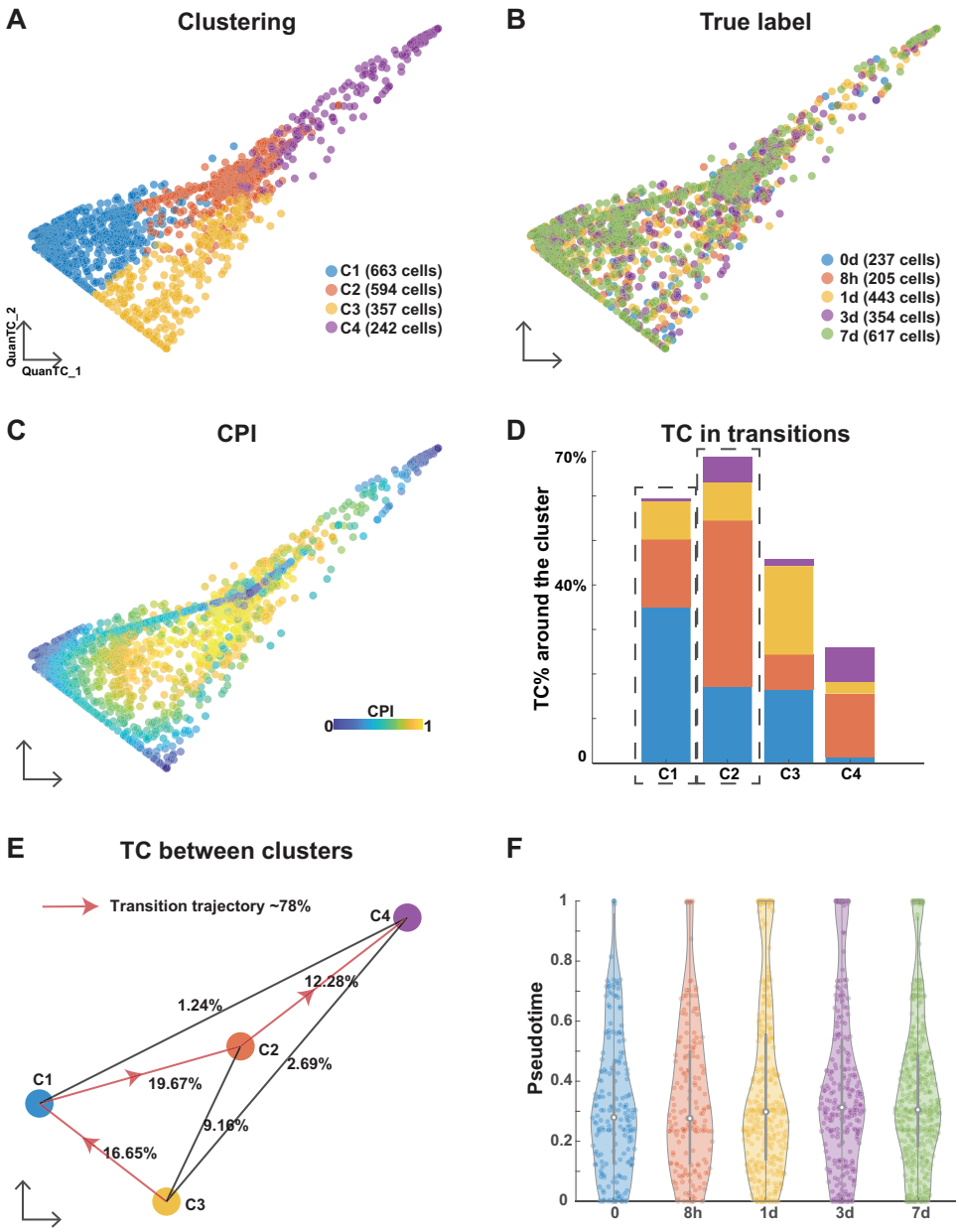

Figure 4.3: Analyzing OVCA420 cancer cell line undergoing EMT induced by TNF using QuanTC. (A-C) Visualization of cells. Each circle represents one cell colored by clustering (A), the collection time of the samples after the treatment (B) and CPI values (C). (D) Percentage of TC associated with each cluster relative to the total number of TC. The dashed box covers the ICS having more TC around. The parameters to choose TC are given in Table C.1. (E) Visualization of cluster centers with color consistent with (A). Each percentage on the line show the percentage of TC between two clusters relative to the total number of cells. Arrowed solid line shows the main transition trajectory. (F) Violin plot of pseudotime value of each cell vs the collection time points. Each dot represents a cell colored by collection time points. The circle displays the mean and vertical line shows the interquartile ranges.

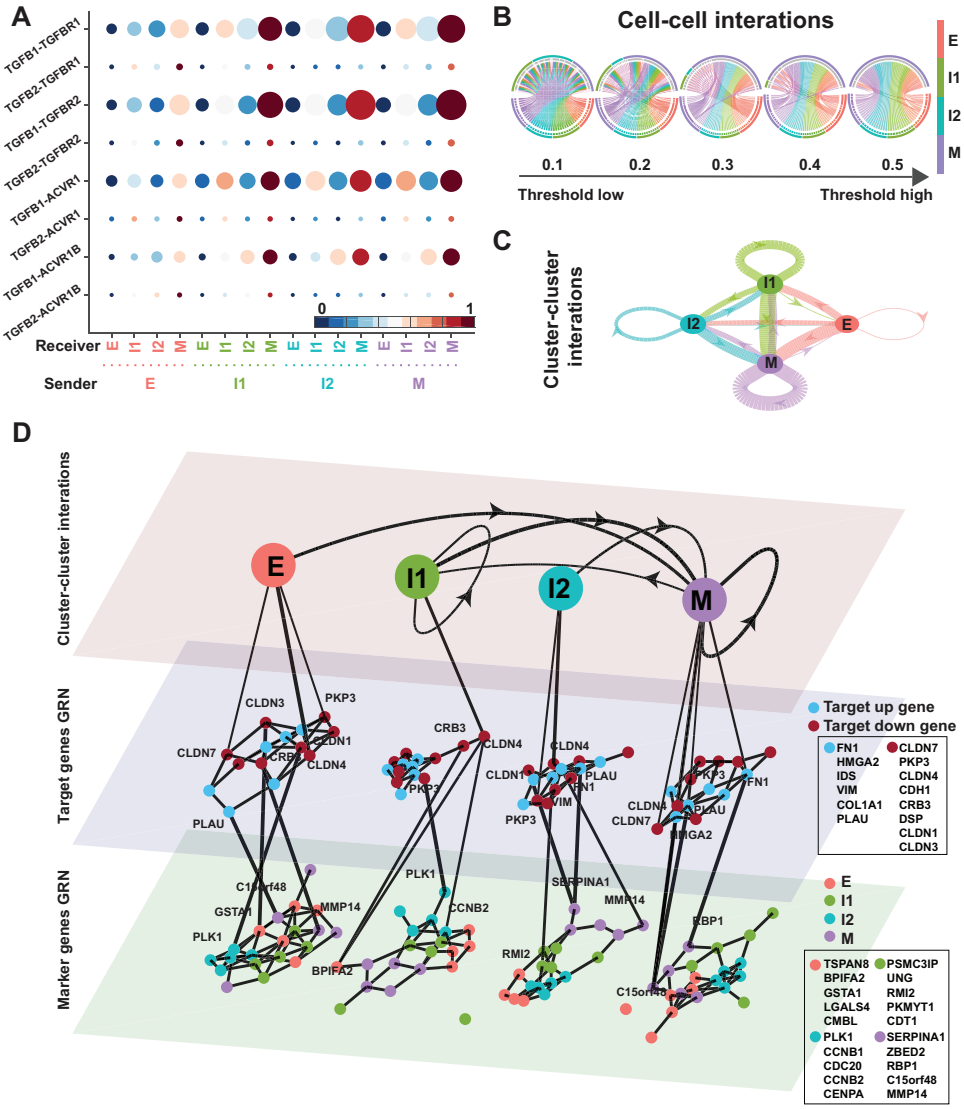

Figure 4.4: TGFB pathway on OVCA420 cancer cell line undergoing EMT induced by TGFB1.

Figure 4.4 *(continued)*: (A) Visualization of signaling probability scores of Ligand-Receptor pairs and their downstream signaling components. Dot size represents the number of averaged cells with nonzero probability scores between clusters. Dot color represents the signaling probability scores. (B) Circos plot of intercellular network on the top ten ligand-producing and top ten receptor-bearing cells from every cluster. The upper hemisphere of the plot shows receptor-bearing cells. The chords of the plot are colored by the ligand-producing cells in the lower hemisphere. The directed edges from the lower hemisphere to the upper hemisphere represent the probabilities of signaling between cells. The probabilities of signaling between cells above the thresholds are presented. (C) Intercluster network. The widths of edges are proportional to the signaling probability scores between clusters. The directed edges are colored by the ligand-producing clusters. (D) Multilayer network. The first layer shows the intercluster network as in (C) but with higher signaling probabilities greater than 0.5. Second and third layers show gene regulatory networks of target genes and top marker genes of clusters respectively using the PIDC algorithm. The target up (down) genes are the up-regulated (down-regulated) target genes of TGF-β signaling. Each dot represents a gene colored by its type. Graph edges indicate the top interactions and the length of the edge is inversely proportional to the interaction strength between genes. The link between first and second layer indicates the target gene are higher expressed within the cluster. The link between second and third layer indicates the strong interaction strength between target and marker genes. The widths of links between layers are proportional to the interaction strength. The ligands, receptors and target genes are given in Table 4.1.

possible role of TGF-β signaling in EMT, we assembled *in silico* ligand-receptor interaction pairs to explore the crosstalk between ICS and E/M states. We applied SoptSC [161] to the expression matrix with inferred states, and calculated the signaling probability of each ligand-receptor pair and their downstream targets between pairs of cells. Finally, averaging these pairwise signaling probabilities within each EMT state provides a snapshot of how cells tend to communicate based on their degree of EMT progression (Figure 4.4A-C).

In Figure 4.4B, the directed edges from lower hemisphere to upper hemisphere were inferred between cells where a high probability of signaling was predicted according to the expressions of ligands in a 'sender' (lower hemisphere in the figure) cell and the appropriate expressions of cognate receptors and target genes in a 'receiver' cell (upper hemisphere in the figure). The large proportion of M state behaving as 'receiver' with high signaling probabilities suggests that the M state played a dominant role as receiver in TGF-β signaling. All the four states

behaved as 'sender' in TGF-β signaling.

The cluster-cluster signaling network was then constructed based on the average cell-cell signaling within each cluster (Figure 4.4C). We used strength, closeness and pagerank as metrics to measure node centrality in the signaling network so that we can quantify the centralities of states in TGF-β signaling. Strength is defined as the sum over weights of the adjacent edges for a given node. Closeness of a node is the inverse of the average length of the shortest path to/from all the other nodes. Pagerank is proportional to the average time spent at a given node during all random walks; therefore, we interpret a high pagerank score as an indication that a node serves as a signaling hub in the network. The pagerank centrality of I1 and M were higher, thus showing the signaling hub potential (Table C.2). The I1 and M states had higher in-strength and lower in-closeness indicating that they behaved more like receivers (Table C.2).

To explore the change of the gene regulatory networks (GRN) underlying TGF-β signaling with respect to EMT progress, we applied PIDC [32], an algorithm using partial information decomposition to identify GRN, to the gene expression matrix of target genes and marker genes inferred by QuanTC within each state. In the dataset induced by TGFB1, the first layer of the multilayer network showed the cluster-cluster interactions as in Figure 4.4C but with only higher signaling probabilities greater than 0.5 (Figure 4.4D, top layer). The widths of the directed lines were proportional to the signaling probabilities. The central and bottom layers displayed the GRN of target genes and marker genes within each state respectively. The interactions between genes within each state were shown by the edges with lengths inversely proportional to the correlations between genes.

Based on the average correlations between target genes of TGF-β signaling and marker genes (Figure C.1F), both the up-regulated target genes and down-regulated target genes had stronger interactions with marker genes within E and M states. The up-regulated target genes always had largest correlations with marker genes of M stats while the down-regulated

target genes had relatively larger correlations with E marker within only E and M states.

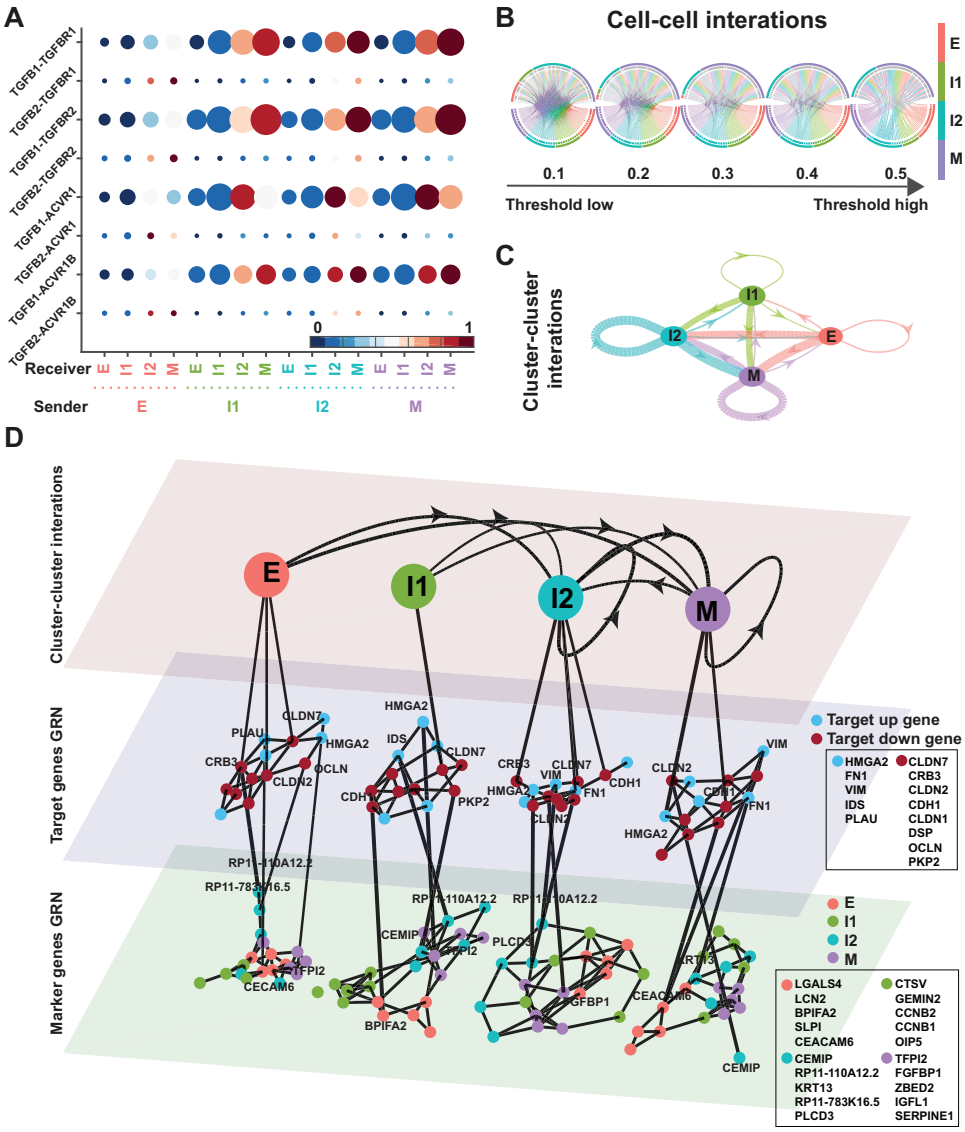

Figure 4.5: TGFB pathway on OVCA420 cancer cell line undergoing EMT induced by EGF.

In the dataset of EMT induced by EGF, the average TGF-β signaling probabilities suggests that I2 and M states played important roles as receivers while all four states shared similar importance as senders (Figure 4.5A-C). Compared to EMT induced by TGFB1, the pagerank centrality of I2, instead of I1, and M states were higher (Table C.2).

In the multilayer network, the highly varied target genes were quite similar to EMT induced

Figure 4.5 *(continued)*: (A) Visualization of signaling probability scores of Ligand-Receptor pairs and their downstream signaling components. Dot size represents the number of averaged cells with nonzero probability scores between clusters. Dot color represents the signaling probability scores. Dot color represents the signaling probability scores. (B) Circos plot of intercellular network on the top ten ligand-producing and top ten receptor-bearing cells from every cluster. The upper hemisphere of the plot shows receptor-bearing cells. The chords of the plot are colored by the ligand-producing cells in the lower hemisphere. The directed edges from the lower hemisphere to the upper hemisphere represent the probabilities of signaling between cells. The probabilities of signaling between cells above the thresholds are presented. (C) Intercluster network. The widths of edges are proportional to the signaling probability scores between clusters. The directed edges are colored by the ligand-producing clusters. (D) Multilayer network. The first layer shows the intercluster network as in (C) but with higher signaling probabilities greater than 0.5. Second and third layers show gene regulatory networks of target genes and top marker genes of clusters respectively using the PIDC algorithm. The target up (down) genes are the up-regulated (down-regulated) target genes of TGF-β signaling. Each dot represents a gene colored by its type. Graph edges indicate the top interactions and the length of the edge is inversely proportional to the interaction strength between genes. The link between first and second layer indicates the target gene are higher expressed within the cluster. The link between second and third layer indicates the strong interaction strength between target and marker genes. The widths of links between layers are proportional to the interaction strength. The ligands, receptors and target genes are given in Table 4.1.

by TGFB1 (Figure 4.4, 4.5D). The up-regulated target genes were the same except missing COL1A1 and five out of the eight down-regulated target genes were the same as in Fig. 4D. However, the top five marker genes of each state varied between the two treatments. Only LGALS4, BPIFA2 and ZBED2 shared marker genes of E and M states. CCNB1 and CCNB2, used to be I2 markers, were I1 markers for EMT induced by EGF.

The average correlations between target genes and marker genes were stronger within the I1 state (Figure C.2F). The up-regulated target genes did not always have largest correlations with marker genes of M state but still with relatively large correlations. The down-regulated target genes had stronger correlations with E markers except in the M state.

In the dataset of EMT induced by TNF, the different EMT states seemed have similar importance as sender in TGF-β signaling (Figure 4.6A-C). The E and M states behaved as the main receivers. The M state had higher pagerank value showing the potential of signaling hub (Table C.2).

In the multilayer network, the varied up-regulated target genes were the subset of the genes in EMT induced by EGF except having CLDN3, and the down-regulated target gene were the subset of those genes in EMT induced by TGFB1 (Figure 4.4, 4.5, 4.6D). More than half of the marker genes of E, I1 and M states were the same as in EMT induced by EGF suggesting the similarity of the EMT under the two treatments.

The target genes and marker genes had higher correlations within the I2 state (Figure C.3F). The up-regulated target genes always had relatively large correlations with marker genes of M state. The down-regulated target genes had stronger correlations with E markers except in the I2 state.

Overall, the M state and part of the ICS behaved as the signaling hub in the TGF-β signaling of EMT under three different treatments (Figure 4.4, 4.5, reffig:c6). The M state was the main receiver in OVCA420 under three treatments with lowest in-closeness (Table C.2).

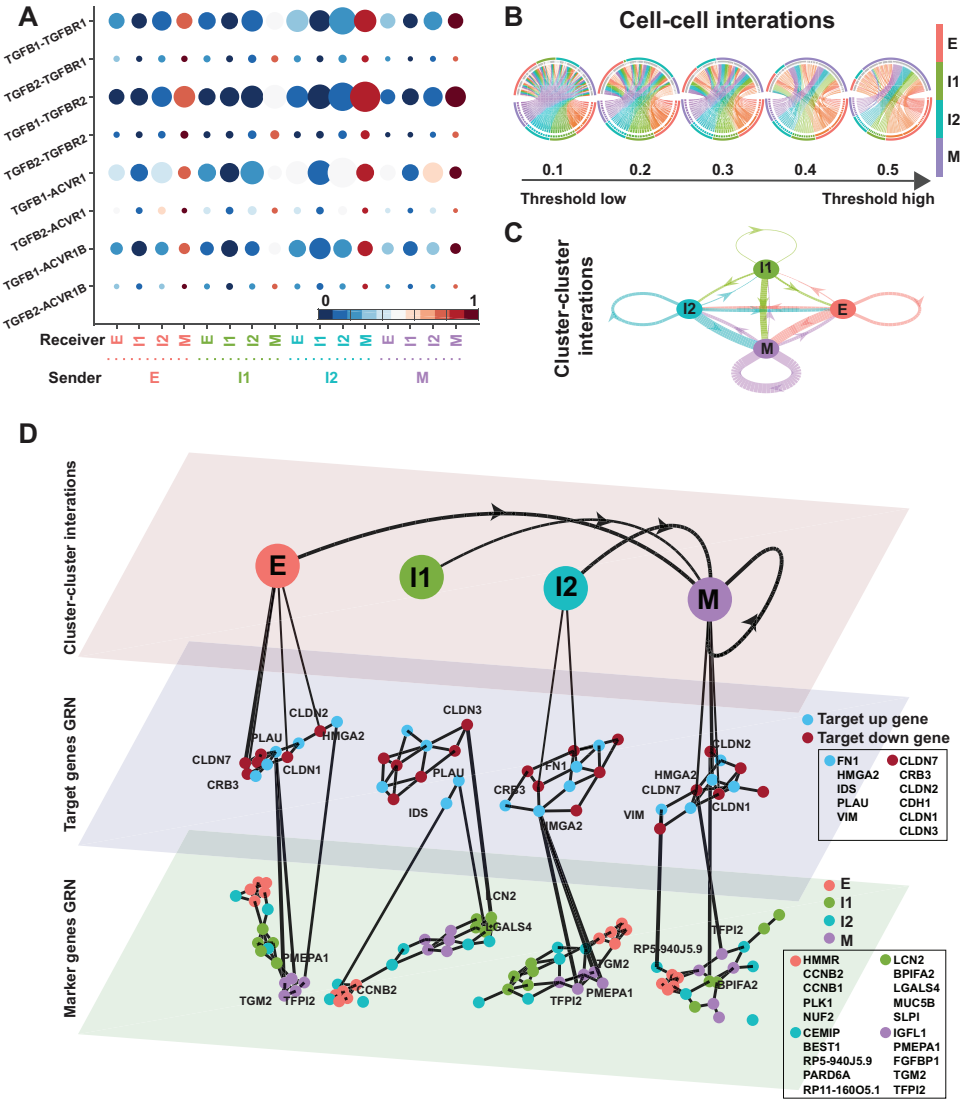

Figure 4.6: TGFB pathway on OVCA420 cancer cell line undergoing EMT induced by TNF.

Figure 4.6 *(continued)*: (A) Visualization of signaling probability scores of Ligand-Receptor pairs and their downstream signaling components. Dot size represents the number of averaged cells with nonzero probability scores between clusters. Dot color represents the signaling probability scores. (B) Circos plot of intercellular network on the top ten ligand-producing and top ten receptor-bearing cells from every cluster. The upper hemisphere of the plot shows receptor-bearing cells. The chords of the plot are colored by the ligand-producing cells in the lower hemisphere. The directed edges from the lower hemisphere to the upper hemisphere represent the probabilities of signaling between cells. The probabilities of signaling between cells above the thresholds are presented. (C) Intercluster network. The widths of edges are proportional to the signaling probability scores between clusters. The directed edges are colored by the ligand-producing clusters. (D) Multilayer network. The first layer shows the intercluster network as in (C) but with higher signaling probabilities greater than 0.5. Second and third layers show gene regulatory networks of target genes and top marker genes of clusters respectively using the PIDC algorithm. The target up (down) genes are the up-regulated (down-regulated) target genes of TGF-β signaling. Each dot represents a gene colored by its type. Graph edges indicate the top interactions and the length of the edge is inversely proportional to the interaction strength between genes. The link between first and second layer indicates the target gene are higher expressed within the cluster. The link between second and third layer indicates the strong interaction strength between target and marker genes. The widths of links between layers are proportional to the interaction strength. The ligands, receptors and target genes are given in Table 4.1.

While the underlying GRN changed between different treatments and along EMT progress. Besides, the top marker genes of different EMT states were quite different among the EMT induced by different treatments, all suggesting the context-specific regulation of GRN during EMT.

4.4.4 Dominant role of ICS *in vivo* during TGF-β signaling

Finally, we compare the results obtained for OVCA420 cells with *in vivo* data from a skin squamous cell carcinoma (SCC) mouse model to seek whether the defining traits of EMT dynamics are conserved or context-specific. In the original study, a total of six distinct cell populations were identified based on differential expression of cell surface markers (CD106, CD61, and CD51), including four transition states [121].

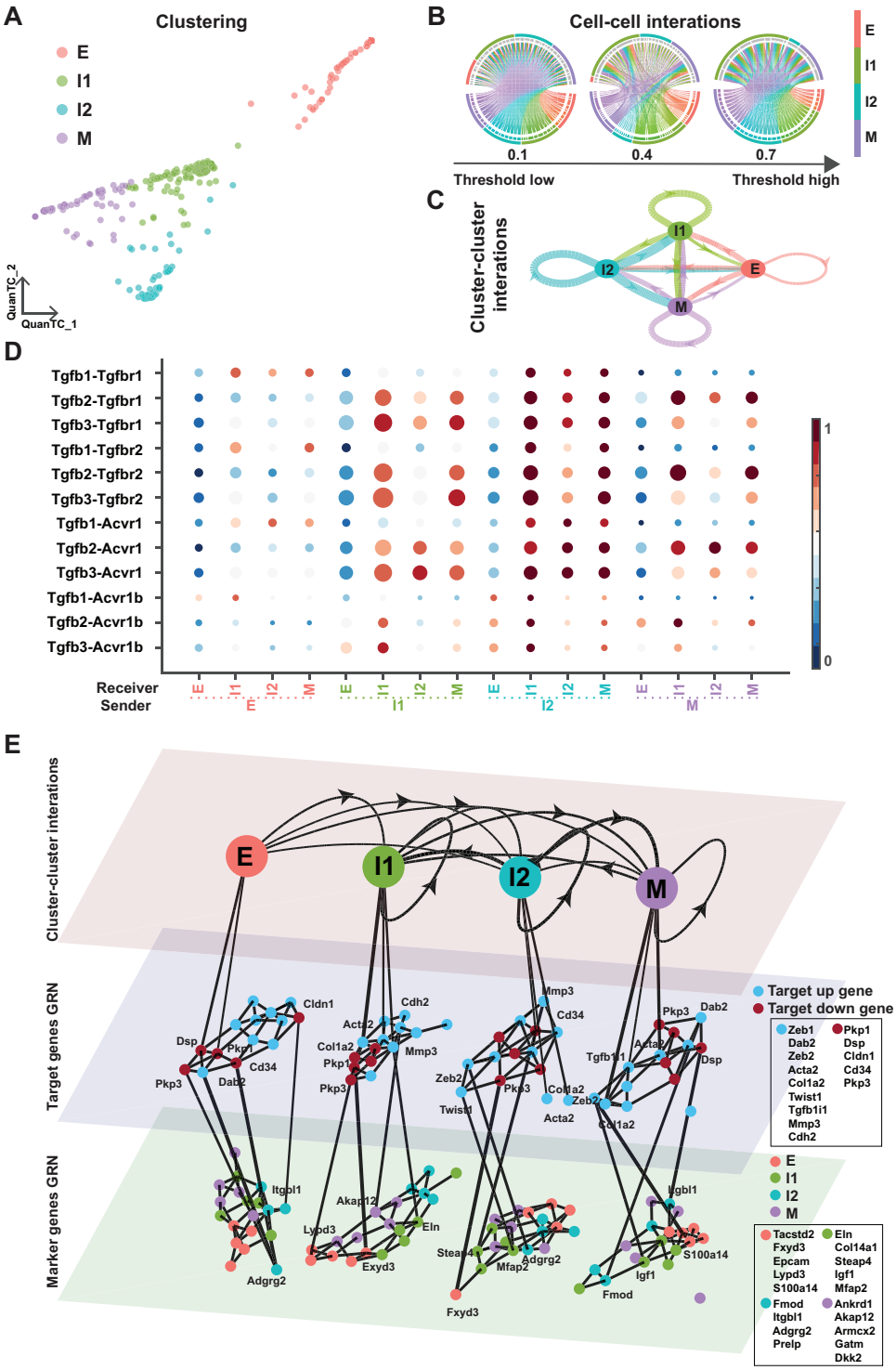

Figure 4.7: TGF-β pathway on EMT in SCC dataset.

Figure 4.7 *(continued)*: (A) Visualization of cells using QuanTC. Each circle represents a cell colored by corresponding cell state. (B) Circos plot of intercellular network on the top ten ligand-producing and top ten receptor-bearing cells from every cluster. The upper hemisphere of the plot shows receptor-bearing cells. The chords of the plot are colored by the ligand-producing cells in the lower hemisphere. The directed edges from the lower hemisphere to the upper hemisphere represent the probabilities of signaling between cells. The probabilities of signaling between cells above the thresholds are presented. (C) Intercluster network. The widths of edges are proportional to the signaling probability scores between clusters. The directed edges are colored by the ligand-producing clusters. (D) Visualization of signaling probability scores of Ligand-Receptor pairs and their downstream signaling components. Dot size represents the number of averaged cells with nonzero probability scores between clusters. Dot color represents the signaling probability scores. (E) Multilayer network. The first layer shows the intercluster network as in (C) but with higher signaling probabilities greater than 0.5. Second and third layers show gene regulatory networks of target genes and top marker genes of clusters respectively using the PIDC algorithm. The target up (down) genes are the up-regulated (down-regulated) target genes of TGF-β signaling. Each dot represents a gene colored by its type. Graph edges indicate the top interactions and the length of the edge is inversely proportional to the interaction strength between genes. The link between first and second layer indicates the target gene are higher expressed within the cluster. The link between second and third layer indicates the strong interaction strength between target and marker genes. The widths of links between layers are proportional to the interaction strength. The ligands, receptors and target genes are given in Table 4.1.

In our previous work [140], we identified a total of four EMT states (Figure C.4A, 4.7A) when applying QuanTC unsupervised clustering [121]. There were two ICS displaying biological processes including cell-cell adhesion and cell migration indicating hybrid epithelial/mesenchymal features (Figure C.4B).

Compared to the OVCA420 cancer cell line undergoing EMT, the ICS in SCC had higher probabilities of signaling and played the even more dominant role of cell-cell and cluster-cluster interactions during TGF-β signaling (Figure 4.7B-D). The ICS, especially the I1 state, had higher Pagerank scores and served as the signaling hub (Table C.2). Both two ICS had lower out-closeness score, indicating that they played the dominant role as the sender in TGF-β signaling. While the M state had by far the higher pagerank score in the three OVCA420 datasets, the pagerank score of the M state in SCC was comparable to those of I1 and I2. Consistently, in the original study, the mesenchymal SCC exhibited a "quasi-mesenchymal" phenotype, which was more similar to intermediate state, instead of a fully mesenchymal phenotype [121].

The highly varied target genes and marker genes of each states shared no similarity to the OVCA420 cancer line (Figure 4.7E). The target genes had strong associations with inferred marker genes within E and I1 states (Figure C.4C). It suggests that EMT varies both between mouse versus human, and *in vitro* versus *in vivo*.

4.5 Discussion

In this study, we have developed an approach combing unsupervised learning, multivariate information theory and multilayer network approach to uncover the complex cellular crosstalk and the underlying gene regulatory relationship of EMT from scRNA-seq data.

We started with trajectory reconstruction on the time-series datasets of an OVCA420 can-

cer cell line undergoing EMT induced by three different external signal (TGFB1, EGF, TNF) and uncovered the existence of multiple intermediate cell states (ICS) displaying hybrid epithelial and mesenchymal features. Analysis of single cell RNA-sequencing previously demonstrated that EMT induction by TGFB1, EGF and TNF is carried by context-specific signaling pathways [37]. Here, we show striking differences in the EMT population dynamics as well. While EMT induced by TGFB1 is synchronous, EGF and TNF induce asynchronous transitions because cells collected at different time points spread all over different clusters. These differences at the cell population level could be explained by the signaling complexity and modularity in response to different EMT inducers. TNF can activate NF-kB signaling, which in turn crosstalks with several transduction pathways and induces responses to inflammation [53]. TNF-NF-kB signaling has also been proposed as a stability factor for hybrid E/M phenotypes, thus potentially resisting a complete EMT in TNF-induced EMT [18]. Similarly, EGF regulation of EMT is not direct, but rather relies on several intermediate signaling steps that could hamper a synchronized transition [78]. Certainly, future efforts focusing on integrating high-throughput data analysis with *in silico* modeling of the underlying regulatory circuitry will help validate or falsify these hypotheses.

To clarify how cells in different EMT states contribute to cell-cell signaling, we subsequently constructed multilayer networks displaying the TGF-β signaling communication between cells in different EMT states and the underlying GRN that regulates EMT at different EMT stages. We found that ICS serve as signaling hubs of cell-cell communication, as well as the context-specific response of TGF-β under different treatments. In other words, cells in intermediate EMT states can send and receive inputs from other cells through TGF-β signaling, potentially inducing EMT in their neighbors. Therefore, both cell autonomous TGFB1 induction and intercellular TGFB signaling could contribute to EMT. Future experiments controlling conditional knock-outs of TGFB ligands could validate this prediction and quantify the role played by cell-cell communication in EMT. These observations also raise an interesting parallel with Notch signaling, another master regulator of cell-cell communica-

tion [23]. Signaling through the Notch-Jagged pathway between cancer cells in intermediate EMT states has been proposed as a mechanism that (i) stabilizes intermediate EMT states and (ii) further induces 'partial EMT' in other cells [69, 20]. Our analysis on *in vivo* dataset also suggests that ICS plays the more dominant role in the TGF-β signaling communication.

The core gene circuits for EMT are known to involve multiple molecular components and interactions [170, 63, 148], providing mechanisms of the EMT transition process [73]. Recent time-series scRNA-seq data suggests that EMT are indeed highly context-specific [37], calling for the need of inferring EMT regulation circuits from a data-driven approach [130, 147]. Previous works have constructed the GRN of EMT based on the combination of prior knowledge, transcription factor predictions and model validations from single-cell datasets [130]. Here we have incorporated the intercellular communications in the context of analyzing transition cells and Intermediate Cell States to inspect the dynamical change of regulation interactions along the EMT spectrum.

Our analysis reveals that ICS plays the crucial role in not only interchanging information with both pure epithelial and mesenchymal states, but also communicating with other cells in ICS during EMT. Previously, the role of ICS has been studied for tumor metastasis [70], and analyzed through the emergent dynamical properties such as signal adaptation, noise attenuation and population transition [138, 45, 145]. Taken together, the EMT cell lineage models with ICS-mediated feedback through cell-cell communications [93, 103] could be further developed to explore the nonlinear effects on different cell populations [65].

The integrative analysis here is a general approach and can be applied to other cell-state transition processes beyond EMT. In particular, the multiplayer gene regulatory and intercellular network provides a multiscale framework to simultaneously explore the cellular communications, the underlying gene regulations and dynamics of GRN along transitions. By incorporating additional layers of different transduction elements beyond TGF-β [67] and associated transcription factors, one can investigate the more complex regulation processes,

such as signal crosstalk and corporation of multiple pathways [168]. In addition, the inclusion of spatial information layer may also facilitate the accuracy of intercellular communication analysis [28].

Overall, our study provides an initial attempt to investigate the multiscale interactions of intercellular communications and gene expression regulations during the dynamical process of cell-fate determination.

Chapter 5

Dynamic Unbalanced Optimal Transport Network for Modeling Cellular Dynamics

5.1 Introduction

The advanced single-cell technologies such as scRNA-seq [84] provide great opportunities for dissection of gene expression at single-cell resolution. However, the lineage relationships and gene expression dynamics of individual cell is untraceable since cells are killed during single-cell sequencing. Experimental lineage tracing approaches can be combined with scRNA-seq but are mostly limited to in vitro applications [7, 156, 40, 8, 125].

Reconstructing the cell fate transitions thus rely heavily on the computational approaches. One extensively used approach is to order cells along differentiation trajectories based on the assumption that developmentally related cells tend to share similarities in gene expressions [167, 126, 134, 152]. Recently, RNA-velocity uses the spliced-to-unspliced mRNA ratio to

get the cell transition direction [13]. The aforementioned approaches were mainly designed for one dataset and largely omitted the temporal information from experiments by integrating multiple datasets together for downstream analysis. With the outbreaks of single-cell RNA-seq technology, time series data become largely available. However, integrating the temporal datasets remains challenging. In addition, inferring cellular dynamics from the high-dimensional gene expression space, including growth and death, remain elusive.

Conventional mechanism-driven mathematical modeling serves as a critical tool in studying cell fate transition. In particular, stochasticity [128, 127, 117] and growth [186] are two major factors that drive cell fate transition, that have been extensively studied. The potential integration of modeling and data provides new opportunities in recovering the cell fate dynamics. Recently, optimal transport (OT) has been used to link the time series data. OT was introduced by Monge in 1871 and formally formulated by Kantorovich in 1942 [80]. OT has served as an approach in finding the transport from two distributions that require minimal transport cost. Kantorovich formulation has been applied to infer the correspondence of scRNA-seq measured at different time points [137, 171, 180]. However, the time dependency of multiple time points is missing in these works because of the pairwise interpolation of two consecutive time points. Moreover, the traditional OT method requires mass conservation which is not appropriate in cellular dynamics.

In this work, we propose an unbalanced optimal transport to model the temporal dynamics of gene expression as a dynamical system. We assume that cells collected at any time are drawn from a distribution in gene expression space. Our method is capable of inferring the temporal changes of those distributions as well as division/death rates of cells. We also use deep learning methods to provide a mesh-free solver which can be extended to high dimensional unpaired time series snapshots.

5.2 Method

5.2.1 Dynamic optimal transport

Dynamic optimal transport introduced by Benamou and Brenier [12] models the transport in a continuum sense utilizing the fluid dynamic framework. They consider a smooth and time-dependent density $\rho(x,t) \geq 0$ and velocity fields $v(x,t) \in \mathbb{R}^d$ satisfying the continuity equation

$$\partial_t \rho + \nabla \cdot (v\rho) = 0 \tag{5.1}$$

for all $t \in [0,T]$, and it transfers ρ_0 to ρ_T with a continuum sense:

$$\rho(\cdot,0) = \rho_0, \rho(\cdot,T) = \rho_T. \tag{5.2}$$

In Kantorovich formulation, the optimal transport function attains the Wasserstein distance. The dynamic optimal transport uses an objective function which is equivalent to the square of Wasserstein distance in the case of $p = 2$ [12]:

$$\inf_{(\rho,v)} T \int_0^T \int_{\mathbb{R}^d} |v(x,t)|^2 \rho(x,t) dx dt = W\left(\rho_0, \rho_T\right)_2^2 . \tag{5.3}$$

Various numerical solvers can be used for the dynamic optimal transport problem [12]. To handle high-dimensional problems, several deep learning-based methods have been introduced [151, 172].

5.2.2 Unbalanced dynamic optimal transport

A major constraint of optimal transport is that they are restricted to measures of equal total mass. The mass conservation is not an appropriate approach in modeling biological systems

that involve growth (mass creation) and death (mass destruction). The unbalanced optimal transport has drawn increasingly interests in connecting two densities with different masses. The unbalanced optimal transport introduces a source term $g(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ in the continuity equation 5.1,

$$\begin{aligned} \partial_t \rho + \nabla \cdot (v\rho) &= g\rho \\ \rho(\cdot, 0) = \rho_0, \rho(\cdot, T) &= \rho_T \end{aligned} \tag{5.4}$$

Chizat et.al picked the objective function in a form of dynamical mixture of Wasserstein distance W_2 and Fisher-Rao distance, called Fisher-Rao over Hellinger [35]:

$$T \int_0^T \int_{\mathbb{R}^d} \left(|v(x,t)|^2 + |g(x,t)|^2 \right) \rho(x,t) dx dt \tag{5.5}$$

to measure the influence of the source term and then induced a convex variational problem. Later, Lee et.al presented fast numerical algorithms for L^1 and L^2 unbalanced dynamic optimal transport [98].

Traditional numerical solvers may become computationally inefficient for the high-dimensional problem. Alternatively, we present our framework in solving the unbalanced dynamic optimal transport problem 5.4 focusing exclusively on the transport cost 5.5 using deep learning models [33, 189, 188]. This framework allows us to access the dynamic of high-dimensional gene regulation.

Here, we present our framework to solve the unbalanced optimal transport in high dimensions. For simplification, we use two time point snapshots, ρ_0 to ρ_T, as an example for explanation. To computationally accelerate high-dimensional integral, we derive an equivalent form of Fisher-Rao over Hellinger:

$$\begin{aligned} & T \int_0^T \int_{\mathbb{R}^d} \left(|v(x,t)|^2 + |g(x,t)|^2 \right) \rho(x,t) dx dt \\ =& T \mathbb{E}_{x_0 \sim \rho_0} \int_0^T \left(|v(x,t)|^2 + |g(x,t)|^2 \right) e^{\int_0^t g(x,s) ds} dt \end{aligned} \tag{5.6}$$

The derivation of 5.6 relies on Theorem 5.1 by taking $f(x,t) = |v(x,t)|^2 + |g(x,t)|^2$:

Theorem 5.1. *If smooth density* $\rho(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}^+$*, velocity field* $v(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ *and growth rate* $g(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ *satisfying*

$$\begin{cases} \partial_t \rho(x,t) + \nabla \cdot (v(x,t)\rho(x,t)) = g(x,t)\rho(x,t) \\ \rho(x,0) = \rho_0(x) \end{cases}$$

for all $0 \le t \le T$ *where* $\begin{cases} \dfrac{dx(t)}{dt} = v(x,t) \\ x(0) = x_0 \end{cases}$*, then for measurable function* $f(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$*, we have*

$$\int_0^T \int_{\mathbb{R}^d} f(x,t)\rho(x,t)dxdt = \mathbb{E}_{x_0 \sim \rho_0} \int_0^T f(x,t) e^{\int_0^t g(x,s)ds} dt$$

Proof. Let $\sigma(x_0,t) = x(t)$, then by Jacobi's formula

$$\begin{aligned} \frac{\partial}{\partial t}\left|\frac{\partial \sigma}{\partial x_0}\right| &= \mathrm{Tr}\left(\mathrm{adj}\left(\frac{\partial \sigma}{\partial x_0}\right)\frac{\partial}{\partial t}\frac{\partial \sigma}{\partial x_0}\right) = \mathrm{Tr}\left(\mathrm{adj}\left(\frac{\partial \sigma}{\partial x_0}\right)\frac{\partial\left(\frac{\partial \sigma}{\partial t}\right)}{\partial x_0}\right) \\ &= \mathrm{Tr}\left(\mathrm{adj}\left(\frac{\partial \sigma}{\partial x_0}\right)\frac{\partial v}{\partial x}\frac{\partial \sigma}{\partial x_0}\right) = \mathrm{Tr}\left(\mathrm{adj}\left(\frac{\partial \sigma}{\partial x_0}\right)\frac{\partial \sigma}{\partial x_0}\nabla v\right) = \left|\frac{\partial \sigma}{\partial x_0}\right| \nabla \cdot v \end{aligned}$$

$$\begin{aligned} \frac{d}{dt}\left(\rho\left|\frac{\partial \sigma}{\partial x_0}\right|\right) &= \frac{d\rho}{dt}\left|\frac{\partial \sigma}{\partial x_0}\right| + \rho\frac{d}{dt}\left(\left|\frac{\partial \sigma}{\partial x_0}\right|\right) = \left(\frac{\partial \rho}{\partial x}v + \frac{\partial \rho}{\partial t}\right)\left|\frac{\partial \sigma}{\partial x_0}\right| + \rho\left|\frac{\partial \sigma}{\partial x_0}\right|\nabla \cdot v \\ &= (\nabla\rho \cdot v + g\rho - \nabla \cdot (v\rho))\left|\frac{\partial \sigma}{\partial x_0}\right| + \rho\left|\frac{\partial \sigma}{\partial x_0}\right|\nabla \cdot v = g\rho\left|\frac{\partial \sigma}{\partial x_0}\right| \end{aligned}$$

Let $\rho \left| \frac{\partial \sigma}{\partial x_0} \right| = M(t)$, then $M(t) = M(0) e^{\int_0^t g(x,s)ds}$

$$\begin{aligned}
\int_{\mathbb{R}^d} f(y,t)\rho(y,t)dy &= \int_{\mathbb{R}^d} f\left(\sigma\left(x_0,t\right),t\right) \rho\left(\sigma\left(x_0,t\right),t\right) \left| \frac{\partial \sigma}{\partial x_0} \right| dx_0 \\
&= \int_{\mathbb{R}^d} f\left(\sigma\left(x_0,t\right),t\right) \rho\left(\sigma\left(x_0,0\right),0\right) e^{\int_0^t g(x,s)ds} dx_0 \\
&= \mathbb{E}_{x_0 \sim \rho_0} f(x,t) e^{\int_0^t g(x,s)ds}
\end{aligned}$$

□

Note that we assume the characteristic curves do not intersect in the theorem.

To efficiently find the optimal transport, we also relax the constraint of estimated density at $t = T$ in 5.4 with the replacement of L^2 norm of the difference as a penalization term [11]. Then the optimization problem is

$$\begin{aligned}
&\inf_{\rho,v,g} T \int_0^T \int_{\mathbb{R}^d} \left(|v(x,t)|^2 + |g(x,t)|^2 \right) \rho(x,t) dx dt + \lambda_d \frac{1}{N^T} \sum_{j=1}^{N^T} \left[\tilde{\rho}_T \left(x_{\mathrm{T}}^{(j)} \right) - \rho_T \left(x_{\mathrm{T}}^{(j)} \right) \right]^2 \\
&= \inf_{\rho,v,g} T \mathbb{E}_{x_0 \sim \rho_0} \int_0^T \left(|v(x,t)|^2 + |g(x,t)|^2 \right) e^{\int_0^t g(x,s)ds} dt + \lambda_d \frac{1}{N^T} \sum_{j=1}^{N^T} \left[\tilde{\rho}_T \left(x_{\mathrm{T}}^{(j)} \right) - \rho_T \left(x_{\mathrm{T}}^{(j)} \right) \right]^2
\end{aligned} \tag{5.7}$$

where $\tilde{\rho}_T$ denotes the numerically estimated density at $t = T$ and the integral above computes using an ODE solver. KL divergence is a commonly used distance to measure the difference between two probability distributions. However, when applying KL divergence or the equivalent maximum likelihood training, the estimated density $\tilde{\rho}_T$ could be unnormalized.

We consider the scenario where the available data are the discrete observations:

$$(t_1, D^1), (t_2, D^2), \cdots, (t_T, D^T).$$

The data $D^i = \left\{ d_{t_i}^{(j)} \right\}_{j=1}^{N^i} \in \mathbb{R}^{N^i \times d}$ is a set of independent and identically distributed samples drawn from the particle distributions at t_i. This sampling captures both the randomness in

the measurement process and the random selection from the population. If no prior information about the mass is given, we consider the number of points observed N^i is proportional to the mass at that time point. We generate the input density ρ_{t_i} via a Gaussian mixture model where the mean of each component is a sample point, and the covariance matrix is $\Sigma = \sigma I \in \mathbb{R}^{d\times d}$. The density ρ_{t_i} is then normalized and times the total mass at t_i.

5.2.3 Data-derived regularization

We have shown how to perform unbalanced dynamic optimal transport in high dimensions. When it comes to the single-cell datasets, we consider the spatial variable x is on gene expression space $\mathbb{R}^d$, and $\rho(x,t) = \rho_t(x) : \mathbb{R}^d \times [0,T] \to \mathbb{R}^+$ is the continuum density function interpolates between densities ρ_0 and ρ_T given from the input data. The advection term $v(x,t) = \frac{dx}{dt} : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ is the velocity field at the gene expression space. The term $g(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ describes the mass change, which models the cell growth/death. The smooth and time-dependent density $\rho(x,t)$, velocity fields $v(x,t)$ and growth rate $g(x,t)$ satisfy 5.4. We assume that cells can move continuously through a real-valued d dimensional space.

In addition, we may have prior knowledge of cellular systems characterized by growth/death, i.e. stem cells/terminally differentiated cells, rather than transport and local velocity arrows, i.e. RNA-velocity [13]. We can add such corresponding regularizations as follows.

Growth/death rate regularization

In some cellular systems, we have the prior knowledge of growth/death rate. For example, the stem cells ($x \in S$) are capable of dividing and renewing themselves ($g \geq 0$) for long periods while terminally differentiated cells ($x \in T$) keratinize and eventually die ($g \leq 0$).

We can use them to regularize the growth rate at t_i:

$$L_{growth,t_i} = \sum_{i=1}^{T} \tanh\left(\mathop{\mathbb{E}}_{x\sim\rho_{t_i}, x\in S} g(x)\right) - \tanh\left(\mathop{\mathbb{E}}_{x\sim\rho_{t_i}, x\in T} g(x)\right) \tag{5.8}$$

Velocity regularization

In scRNA-seq data, RNA-velocity [13] provides the velocity estimation $\hat{v}$ of cellular dynamics so that we can use them to regularize the direction of flow at t_i:

$$L_{velocity,t_i} = \frac{1}{N^i} \sum_{j=1}^{N^i} \frac{v\left(x_{t_i}^{(j)}, t_i\right) \cdot \hat{v}\left(x_{t_i}^{(j)}, t_i\right)}{\left\|v\left(x_{t_i}^{(j)}, t_i\right)\right\| \left\|\hat{v}\left(x_{t_i}^{(j)}, t_i\right)\right\|} \tag{5.9}$$

5.3 Deep learning-based solver for OT in high dimension

Lemma 5.1. *If density* $\rho(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}^+$*, velocity field* $v(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ *and growth rate* $g(x,t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ *satisfying* $\begin{cases} \partial_t \rho(x,t) + \nabla \cdot (v(x,t)\rho(x,t)) = g(x,t)\rho(x,t) \\ \rho(x,0) = \rho_0(x) \end{cases}$

for all $0 \le t \le T$ *where* $\begin{cases} \dfrac{dx(t)}{dt} = v(x,t) \\ x(0) = x_0 \end{cases}$*, we have* $\frac{d(\ln \rho)}{dt} = g - \nabla \cdot v$

Proof. Let $\sigma(x_0, t) = x(t)$

Then $\rho(x,t) = \rho(\sigma(x_0,t)) = \tilde{\rho}(x_0,t)$

$\frac{\partial \rho}{\partial t} = g\rho - \nabla \cdot (v\rho) = g\rho - \nabla\rho \cdot v - \rho\nabla \cdot v$

$\frac{d\rho}{dt} = \frac{d\tilde{\rho}}{dt} = \nabla\rho \cdot \frac{dx}{dt} + \frac{\partial \rho}{\partial t} = \nabla\rho \cdot v + \frac{\partial \rho}{\partial t} = \nabla\rho \cdot v + g\rho - \nabla\rho \cdot v - \rho\nabla \cdot v = g\rho - \rho\nabla \cdot v$

So that $\frac{d(\ln \rho)}{dt} = g - \nabla \cdot v$ $\square$

We have two neural networks and each one consists of four fully connected layers of 16 nodes with tanh activations. The two networks take as input the samples and time and output the derivative of samples with respect to time $v = \frac{dx}{dt}$ and growth rate g of the sample at that point respectively. Based on Lemma 5.1, we then have $\frac{d(\ln \rho)}{dt} = g - \nabla \cdot v$. During each training iteration, we start at the final two consecutive time points, integrate to the one time point earlier, and continue till the initial time t_1. For two consecutive time points t_i, t_{i+1}, we draw samples $x_{t_{i+1}} \sim \rho_{t_{i+1}}$ in the following way: we collect 1000 random input samples and perturb the spatial coordinate of each point with Gaussian noise $\mathcal{N}(0, \sigma I)$. The covariance matrix σI is consistent with the Gaussian mixture model. Taking the integral backward along the trajectory satisfying $v = \frac{dx}{dt}$ with the initial $x_{t_{i+1}}$, we have

$$\hat{x}_{t_i} = x_{t_{i+1}} + \int_{t_{i+1}}^{t_i} v(x,t)dt \tag{5.10}$$

We can then estimate

$$\ln \tilde{\rho}_{t_{i+1}} \left(x_{t_{i+1}}\right) = \ln \rho_{t_i} \left(\hat{x}_{t_i}\right) - \int_{t_{i+1}}^{t_i} \frac{d \ln \tilde{\rho}}{dt} dt \tag{5.11}$$

where $\tilde{\rho}$ denotes the estimated density at t_{i+1}. These regularizations are summarized in the following single loss function:

$$\begin{aligned} L_{t_i, t_{i+1}} = (t_{i+1} - t_i)\, \mathbb{E}_{x_0 \sim \rho_{t_i}} & \int_{t_i}^{t_{i+1}} \left(|v(x,t)|^2 + |g(x,t)|^2\right) e^{\int_{t_i}^{t} g(x(s),s)ds} dt \\ & + \lambda_d \frac{1}{N^{i+1}} \sum_{j=1}^{N^{i+1}} \left[\tilde{\rho}_{t_{i+1}} \left(x_{t_{i+1}}^{(j)}\right) - \rho_{t_{i+1}} \left(x_{t_{i+1}}^{(j)}\right)\right]^2 \\ & - \lambda_g \left[\tanh\left(\mathop{\mathbb{E}}_{x \sim \rho_{t_i}, x \in S} g(x)\right) - \tanh\left(\mathop{\mathbb{E}}_{x \sim \rho_{t_i}, x \in T} g(x)\right)\right] \\ & - \lambda_v \sum_{j=1}^{N^i} \frac{v\left(x_{t_i}^{(j)}, t_i\right) \cdot \hat{v}\left(x_{t_i}^{(j)}, t_i\right)}{\left\|v\left(x_{t_i}^{(j)}, t_i\right)\right\| \left\|\hat{v}\left(x_{t_i}^{(j)}, t_i\right)\right\|} \end{aligned} \tag{5.12}$$

where the integral above is computed using an ODE solver. Here we use a memory efficient and reverse accurate integrator for Neural ODEs in the training process [188]. Algorithm 1 shows the framework of our method.

Algorithm 1

Require: A series of snapshots $(t_1, D^1), (t_2, D^2), \cdots, \left(t_T, D^T\right)$ where $D^i = \left\{d_{t_i}^{(j)}\right\}_{j=1}^{N^i} \in \mathbb{R}^{N^i \times d}$. If mass M^i is not provided, $M^i = \frac{N^i}{N^1}$

Ensure: Neuron networks: $(x,t) \to NN1 \to v(x,t)$ and $(x,t) \to NN2 \to g(x,t)$

Preprocessing: Using Gaussian mixture model to generate density ρ_{t_i} from snapshot D^i

$$\rho_{t_i}(x) = \frac{M^i}{N^i} \sum_{j=1}^{N^i} \frac{\exp\left(-\frac{1}{2}\left(x - d_{t_i}^{(j)}\right)^T \Sigma^{-1}\left(x - d_{t_i}^{(j)}\right)\right)}{\sqrt{(2\pi)^d |\Sigma|}}, \Sigma = \sigma I \in \mathbb{R}^{d \times d}$$

for *epoch* from 1 to *Epochs* **do**

$Loss = 0$

for i from $T-1$ to 1 **do**

$x_{t_{i+1}} \sim \rho_{t_{i+1}}, x_{t_{i+1}} = \left(x_{t_{i+1}}^{(1)}, x_{t_{i+1}}^{(2)}, \cdots, x_{t_{i+1}}^{(\text{batch})}\right)$ ▷ i.i.d sampling

Integrating backward from t_{i+1} to t_i

$$\begin{cases} \dfrac{dx}{dt} = v(x,t) \\ \dfrac{d(\mathrm{z}(x,t))}{dt} = g(x,t) - \nabla \cdot v(x,t) \end{cases}, \begin{cases} x\left(t_{i+1}\right) = x_{t_{i+1}} \\ z\left(t_{i+1}\right) = 0 \end{cases}$$ ▷ Estimate $x\left(t_i\right) = \hat{x}_{t_i}$

$z_{t_i} = \int_{t_{i+1}}^{t_i} (g(x,t) - \nabla \cdot v(x,t))dt = \int_{t_{i+1}}^{t_i} \frac{d(\ln \rho(x,t))}{dt} dt$ ▷ Intermediate variable z_{t_i}

$\ln \tilde{\rho}_{t_{i+1}}\left(x_{t_{i+1}}\right) = \ln \rho_{t_i}\left(\hat{x}_{t_i}\right) - z_{t_i}$ ▷ Estimate $\tilde{\rho}_{t_{i+1}}$

$L_{t_i, t_{i+1}} =$

$(t_{i+1} - t_i)\, \mathbb{E}_{x_0 \sim \rho_{t_i}} \int_{t_i}^{t_{i+1}} \left(|v(x,t)|^2 + |g(x,t)|^2\right) e^{\int_{t_i}^{t} g(x,s)ds} dt$ ▷ Compute transport cost

$+ \frac{\lambda_d}{\text{batch}} \sum_{j=1}^{\text{batch}} \left[\tilde{\rho}_{t_{i+1}}\left(x_{t_{i+1}}^{(j)}\right) - \rho_{t_{i+1}}\left(x_{t_{i+1}}^{(j)}\right)\right]^2$ ▷ Compute error of estimation

$Loss = Loss + L_{t_i, t_{i+1}}$

end for

Update $NN1$ and $NN2$ using the Adam algorithm by minimizing the *Loss*

end for

5.4 Results

5.4.1 Simulated data from a stochastic model

We build a simulation for stochastic movement of particles in gene expression space. We assume there is an attractor $z = [0, 0]$. The resulting velocity v is given by the negative gradient of the potential: $v(x) = -\nabla\psi(x)$, where the potential function is $\psi(x) = 5\|x - z\|^2$. Simulated particles are initially isotropically distributed from a ring where the radius and angle (r, θ) are uniformly and independently distributed in $[2, 2.5] \times (0, 2\pi]$. Particles then evolve following the drift-diffusion dynamics $dx_t = v(x_t)\,dt + 0.1dB_t$ where dB_t denotes the increments of a Brownian motion. Particles located outside the circle of radius 0.8 of the origin are capable of the ability to divide with 1% probability. Every time a particle divides, its current state x_t will be the initial state for the two new trajectories.

Initially, we pick $N = 500$ particles and run the simulations. The stochastic differential equation is solved by Euler-Maruyama method using time step $\Delta t = 0.001$. At time $t = 0, 0.1, 0.2$, we take observed positions of particle as input data and the input densities are generated by Gaussian mixture model with variance $\sigma = 0.03$.

Our deep learning model successfully captures the dynamics in population level connecting the three input densities (Figure 5.1). In the stochastic model, the particles depart from a ring and move toward to the center of the ring. The velocity decreases as a particle closer to the center. In the deep learning model, the velocity is approximated by a neural network. Rather than the population level, our deep learning model also captures the individual transition path of each particle where the approximated velocity shows great fit to the ground truth in both direction and magnitude (Figure 5.1A-B). Moreover, only particles located outside the circle of radius 0.8 of the origin have the ability for division, indicating the positive growth rate ($g > 0$). Our deep learning model also captures the correct signs of

the growth rate (Figure 5.1A-B).

5.4.2 Epithelial-to-mesenchymal transition (EMT) scRNA-seq data

We apply our method to the time series scRNA-seq which comprise of TGFB1 induced EMT from A549 cancer cell line [37]. This data consists of cells collected at five timepoints (Figure 5.2A). For data at each time point, the same number of cells were initially cultured and treated with the same conditions. As a result, the number of cells collected at different time points is highly correlated with the mass, and we take number of cells as the input mass.

The cells at day 0 show epithelial morphology in culture. Under the treatment of TGFB1, cells exhibit morphological changes which is consistent with EMT during 7 days. The single-cell data was collected at 5 time points (0d, 8h, 1d, 3d, 7d) independently. We take the UMAP embeddings of these datasets at two-dimensional space as the input. The input densities are constructed by Gaussian mixture model with $\sigma = 0.02$. From these datasets, cell fate transition can be observed in the projected two-dimensional UMAP space where cells roughly move in one direction from lower right corner to upper left corner (Figure 5.2A). Our deep learning model first captures the dynamics of cell densities in the projected space (Figure 5.2C). The velocity approximately captures directions of the cell fate transition (Figure 5.2D). Moreover, we reversely track the individual cell fate trajectory at day 7 (Figure 5.2B). The multiple trajectories roughly recover the cell densities at each day. Indeed, our model can track cell fate transition in a continuous manner over a large time scale. Strikingly, our estimation of growth rate is negative at day 7 when cells have become mesenchymal cells and is positive when cells are undergoing EMT (day 0 – day3) which matches the stemness properties of partial EMT [19, 95].

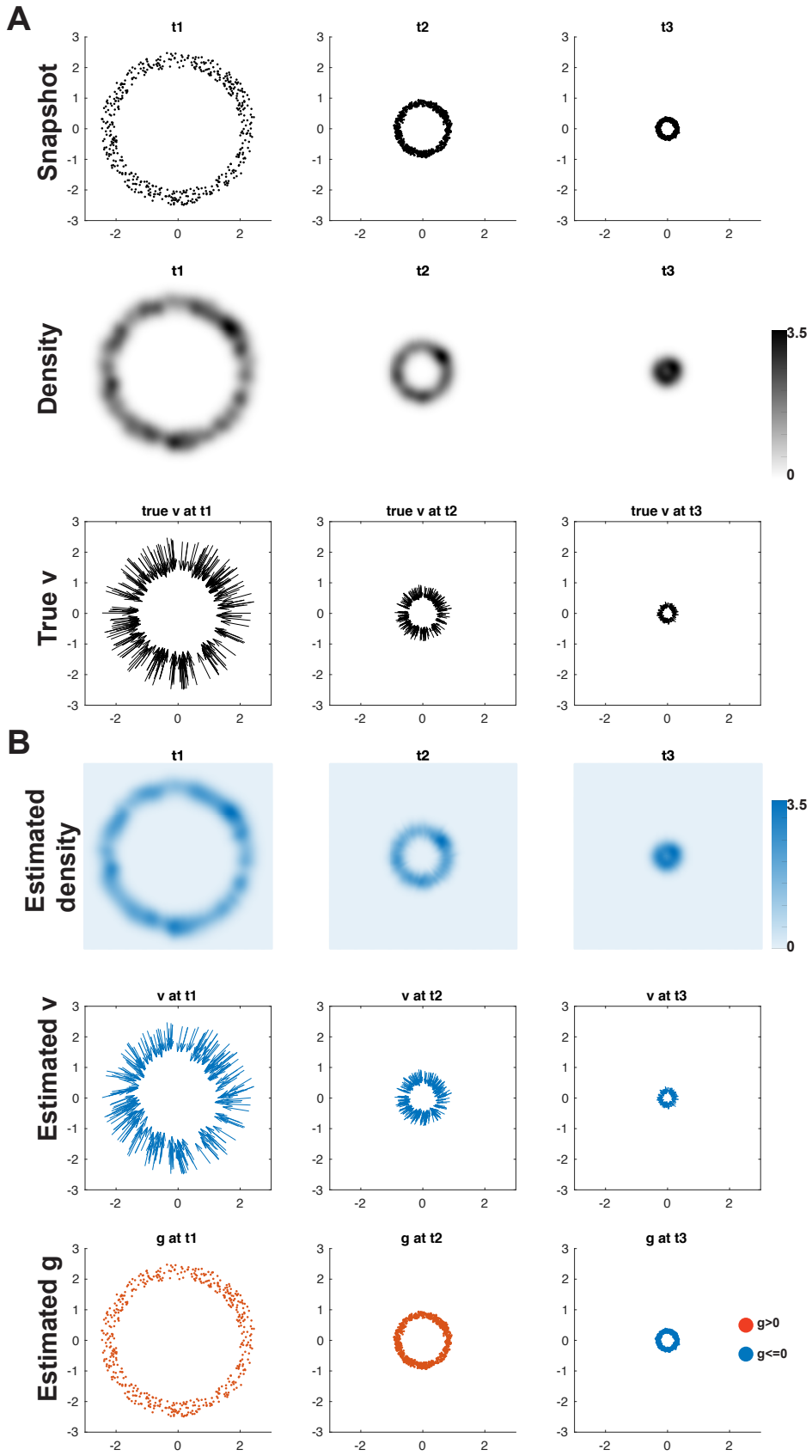

Figure 5.1: Simulated data and inferred dynamics.

Figure 5.1 *(continued)*: (A) Snapshots of particles, the input densities ρ_{t_i} and the true velocities $v(x) = -\nabla\psi(x)$ of 200 random selected given particles at each time point. (B) Estimated densities, velocities v and growth rate g at the same subset of particles.

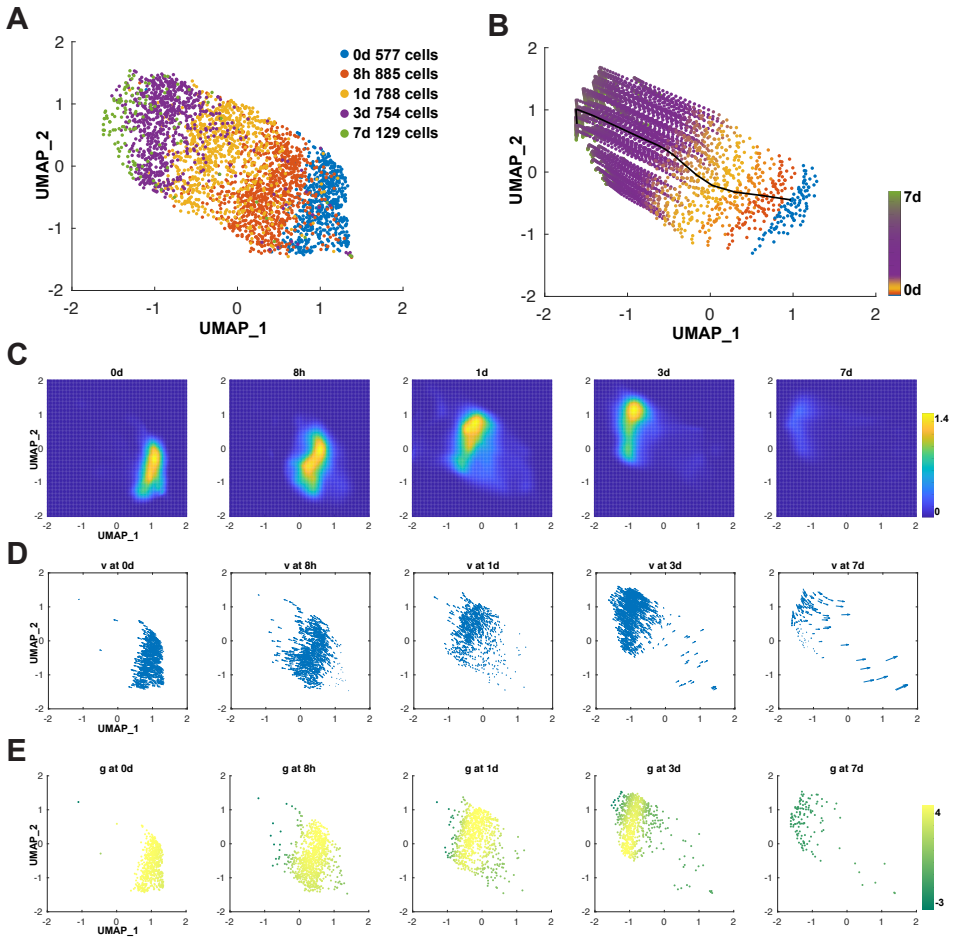

Figure 5.2: EMT scRNA-seq and inferred dynamics shown on the UMAP embedding. (A) Five time point scRNA-seq data. (B) Trajectory of 100 sampled cells from day 7. The black line links one cell trajectory. (C-E) Estimated densities, velocities v and growth rate g at the observed cells.

5.5 Discussion

We present a method using deep learning to generate a mesh free solver for computing the unbalanced dynamical optimal transport. Our method is capable of inferring the temporal dynamics from densities with mass changes due to cellular division and death. We demonstrate the efficacy of our method first on a benchmark simulated data. Then we apply it to a time-series scRNA-seq data to infer individual large time scale cell fate transition and cellular dynamics. Although the method is only applied to problem at two-dimensional space, it can be efficiently applied to high-dimension. Inferring the trajectories at projected space to gene space would be interesting to study the temporal dynamics of gene regulatory networks.

Bibliography

[1] A. N. Abell, N. V. Jordan, W. Huang, A. Prat, A. A. Midland, N. L. Johnson, D. A. Granger, P. A. Mieczkowski, C. M. Perou, S. M. Gomez, et al. Map3k4/cbp-regulated h2b acetylation controls epithelial-mesenchymal transition in trophoblast stem cells. *Cell stem cell*, 8(5):525–537, 2011.

[2] S. An, L. Ma, and L. Wan. Tsee: an elastic embedding method to visualize the dynamic gene expression patterns of time series single-cell rna sequencing data. *BMC genomics*, 20(2):77–92, 2019.

[3] F. Andre and C. Zielinski. Optimal strategies for the treatment of metastatic triple-negative breast cancer with currently approved agents. *Annals of oncology*, 23:vi46–vi51, 2012.

[4] F. Andriani, G. Bertolini, F. Facchinetti, E. Baldoli, M. Moro, P. Casalini, R. Caserini, M. Milione, G. Leone, G. Pelosi, et al. Conversion to stem-cell state in response to microenvironmental cues is regulated by balance between epithelial and mesenchymal features in lung cancer cells. *Molecular oncology*, 10(2):253–271, 2016.

[5] V. Arnoux, C. Côme, D. F. Kusewitt, L. G. Hudson, and P. Savagner. Cutaneous wound reepithelialization. In *Rise and fall of epithelial phenotype*, pages 111–134. Springer, 2005.

[6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[7] C. S. Baron and A. van Oudenaarden. Unravelling cellular relationships during development and regeneration using genetic lineage tracing. *Nature reviews molecular cell biology*, 20(12):753–765, 2019.

[8] N. Battich, J. Beumer, B. de Barbanson, L. Krenning, C. S. Baron, M. E. Tanenbaum, H. Clevers, and A. van Oudenaarden. Sequencing metabolically labeled transcripts in single cells reveals mrna turnover strategies. *Science*, 367(6482):1151–1156, 2020.

[9] E. Beerling, D. Seinstra, E. de Wit, L. Kester, D. van der Velden, C. Maynard, R. Schäfer, P. van Diest, E. Voest, A. van Oudenaarden, et al. Plasticity between epithelial and mesenchymal states unlinks EMT from metastasis-enhancing stem cell capacity. *Cell reports*, 14(10):2281–2288, 2016.

[10] E. Ben-Jacob, D. S. Coffey, and H. Levine. Bacterial survival strategies suggest rethinking cancer cooperativity. *Trends in microbiology*, 20(9):403–410, 2012.

[11] J.-D. Benamou. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(5):851–868, 2003.

[12] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[13] V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology*, 38(12):1408–1414, 2020.

[14] B. Bierie, S. E. Pierce, C. Kroeger, D. G. Stover, D. R. Pattabiraman, P. Thiru, J. L. Donaher, F. Reinhardt, C. L. Chaffer, Z. Keckesova, et al. Integrin-β4 identifies cancer stem cell-enriched populations of partially mesenchymal carcinoma cells. *Proceedings of the National Academy of Sciences*, 114(12):E2337–E2346, 2017.

[15] C. Blanpain and E. Fuchs. Plasticity of epithelial stem cells in tissue regeneration. *Science*, 344(6189):1242281, 2014.

[16] M. Boareto, M. K. Jolly, E. Ben-Jacob, and J. N. Onuchic. Jagged mediates differences in normal and tumor angiogenesis by affecting tip-stalk fate decision. *Proceedings of the National Academy of Sciences*, 112(29):E3836–E3844, 2015.

[17] M. Boareto, M. K. Jolly, A. Goldman, M. Pietilä, S. A. Mani, S. Sengupta, E. Ben-Jacob, H. Levine, and J. Onuchic. Notch-jagged signalling can give rise to clusters of cells exhibiting a hybrid epithelial/mesenchymal phenotype. *Journal of the Royal Society Interface*, 13(118):20151106, 2016.

[18] F. Bocci, L. Gearhart-Serna, M. Boareto, M. Ribeiro, E. Ben-Jacob, G. R. Devi, H. Levine, J. N. Onuchic, and M. K. Jolly. Toward understanding cancer stem cell heterogeneity in the tumor microenvironment. *Proceedings of the National Academy of Sciences*, 116(1):148–157, 2019.

[19] F. Bocci, M. K. Jolly, J. T. George, H. Levine, and J. N. Onuchic. A mechanism-based computational model to capture the interconnections among epithelial-mesenchymal transition, cancer stem cells and notch-jagged signaling. *Oncotarget*, 9(52):29906, 2018.

[20] F. Bocci, M. K. Jolly, S. C. Tripathi, M. Aguilar, S. M. Hanash, H. Levine, and J. N. Onuchic. Numb prevents a complete epithelial–mesenchymal transition by modulating notch signalling. *Journal of the Royal Society Interface*, 14(136):20170512, 2017.

[21] F. Bocci, J. N. Onuchic, and M. K. Jolly. Understanding the principles of pattern formation driven by notch signaling by integrating experiments and theoretical models. *Frontiers in Physiology*, page 929, 2020.

[22] C. Boutsidis and E. Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern recognition*, 41(4):1350–1362, 2008.

[23] S. J. Bray. Notch signalling in context. *Nature reviews Molecular cell biology*, 17(11):722–735, 2016.

[24] R. Browaeys, W. Saelens, and Y. Saeys. Nichenet: modeling intercellular communication by linking ligands to target genes. *Nature Methods*, 17(2):159–162, 2020.

[25] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature-Biotechnology*, 36(5):411–420, 2018.

[26] S. Cabello-Aguilar, M. Alame, F. Kon-Sun-Tack, C. Fau, M. Lacroix, and J. Colinge. Singlecellsignalr: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Research*, 48(10):e55–e55, 2020.

[27] J. G. Camp, K. Sekine, T. Gerber, H. Loeffler-Wirth, H. Binder, M. Gac, S. Kanton, J. Kageyama, G. Damm, D. Seehofer, et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature*, 546(7659):533–538, 2017.

[28] Z. Cang and Q. Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature Communications*, 11(1):1–13, 2020.

[29] J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.

[30] C. L. Chaffer, B. P. San Juan, E. Lim, and R. A. Weinberg. EMT, cell plasticity and metastasis. *Cancer and Metastasis Reviews*, 35(4):645–654, 2016.

[31] R. Chakrabarti, J. Hwang, M. Andres Blanco, Y. Wei, M. Lukačišin, R.-A. Romano, K. Smalley, S. Liu, Q. Yang, T. Ibrahim, et al. Elf5 inhibits the epithelial-mesenchymal transition in mammary gland development and breast cancer metastasis by transcriptionally repressing snail2. *Nature cell biology*, 14(11):1212–1222, 2012.

[32] T. E. Chan, M. P. Stumpf, and A. C. Babtie. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, 5(3):251–267, 2017.

[33] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[34] Z. Chen, S. An, X. Bai, F. Gong, L. Ma, and L. Wan. Densitypath: an algorithm to visualize and reconstruct cell state-transition path on density landscape for single-cell rna sequencing data. *Bioinformatics*, 35(15):2593–2601, 2019.

[35] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. An interpolating distance between optimal transport and fisher–rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

[36] G. O. Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Research*, 47(D1):D330–D338, 2019.

[37] D. P. Cook and B. C. Vanderhyden. Context specificity of the EMT transcriptional response. *Nature Communications*, 11(1):1–9, 2020.

[38] G. Csardi, T. Nepusz, et al. The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9, 2006.

[39] J. Dong, Y. Hu, X. Fan, X. Wu, Y. Mao, B. Hu, H. Guo, L. Wen, and F. Tang. Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. *Genome biology*, 19(1):1–20, 2018.

[40] F. Erhard, M. A. Baptista, T. Krammer, T. Hennig, M. Lange, P. Arampatzi, C. S. Jürges, F. J. Theis, A.-E. Saliba, and L. Dölken. scslam-seq reveals core features of transcription dynamics in single cells. *Nature*, 571(7765):419–423, 2019.

[41] H. Fazilaty, L. Rago, K. Kass Youssef, O. H. Ocaña, F. Garcia-Asencio, A. Arcas, J. Galceran, and M. A. Nieto. A gene regulatory network to control EMT programs in development and disease. *Nature Communications*, 10(1):1–16, 2019.

[42] K. R. Fischer, A. Durrans, S. Lee, J. Sheng, F. Li, S. T. Wong, H. Choi, T. El Rayes, S. Ryu, J. Troeger, et al. Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. *Nature*, 527(7579):472–476, 2015.

[43] F. Font-Clos, S. Zapperi, and C. A. La Porta. Topography of epithelial–mesenchymal plasticity. *Proceedings of the National Academy of Sciences*, 115(23):5902–5907, 2018.

[44] J. T. George, M. K. Jolly, S. Xu, J. A. Somarelli, and H. Levine. Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric. *Cancer research*, 77(22):6415–6428, 2017.

[45] H. Goetz, J. R. Melendez-Alvarez, L. Chen, and X.-J. Tian. A plausible accelerating function of intermediate states in cancer metastasis. *PLoS Computational Biology*, 16(3):e1007682, 2020.

[46] A. D. Grigore, M. K. Jolly, D. Jia, M. C. Farach-Carson, and H. Levine. Tumor budding: the name is EMT. partial EMT. *Journal of clinical medicine*, 5(5):51, 2016.

[47] M. Guo, E. L. Bao, M. Wagner, J. A. Whitsett, and Y. Xu. Slice: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Research*, 45(7):e54–e54, 2017.

[48] W. Guo, Z. Keckesova, J. L. Donaher, T. Shibue, V. Tischler, F. Reinhardt, S. Itzkovitz, A. Noske, U. Zürrer-Härdi, G. Bell, et al. Slug and sox9 cooperatively determine the mammary stem cell state. *Cell*, 148(5):1015–1028, 2012.

[49] P. B. Gupta, C. M. Fillmore, G. Jiang, S. D. Shapira, K. Tao, C. Kuperwasser, and E. S. Lander. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*, 146(4):633–644, 2011.

[50] D. Haensel and X. Dai. Epithelial-to-mesenchymal transition in cutaneous wound healing: Where we are and where we are heading. *Developmental dynamics*, 247(3):473–480, 2018.

[51] D. Haensel, P. Sun, A. L. MacLean, X. Ma, Y. Zhou, M. P. Stemmler, S. Brabletz, G. Berx, M. V. Plikus, Q. Nie, et al. An ovol2-zeb1 transcriptional circuit regulates epithelial directional migration and proliferation. *EMBO reports*, 20(1):e46273, 2019.

[52] Y. Hao, D. Baker, and P. Ten Dijke. TGF-β-mediated epithelial-mesenchymal transition and cancer metastasis. *International journal of molecular sciences*, 20(11):2767, 2019.

[53] M. S. Hayden and S. Ghosh. Regulation of nf-κB by tnf family cytokines. In *Seminars in immunology*, volume 26, pages 253–266. Elsevier, 2014.

[54] J. S. Herman, D. Grün, et al. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nature Methods*, 15(5):379–386, 2018.

[55] T. Hong, K. Watanabe, C. H. Ta, A. Villarreal-Ponce, Q. Nie, and X. Dai. An ovol2-zeb1 mutual inhibitory circuit governs bidirectional and multi-step transition between epithelial and mesenchymal states. *PLoS Computational Biology*, 11(11):e1004569, 2015.

[56] B. Huang, M. Lu, D. Jia, E. Ben-Jacob, H. Levine, and J. N. Onuchic. Interrogating the topological robustness of gene regulatory circuits by randomization. *PLoS computational biology*, 13(3):e1005456, 2017.

[57] H.-M. Huang, D. S.-T. Chang, and P.-C. Wu. The association between near work activities and myopia in children—a systematic review and meta-analysis. *PloS one*, 10(10):e0140419, 2015.

[58] R. Y. Huang, M. Wong, T. Tan, K. Kuay, A. Ng, V. Chung, Y. Chu, N. Matsumura, H. Lai, Y. Lee, et al. An EMT spectrum defines an anoikis-resistant and spheroidogenic intermediate mesenchymal state that is sensitive to e-cadherin restoration by a src-kinase inhibitor, saracatinib (AZD0530). *Cell death & disease*, 4(11):e915–e915, 2013.

[59] S. Huang. Hybrid t-helper cells: stabilizing the moderate center in a polarized system. *PLoS biology*, 11(8):e1001632, 2013.

[60] L. G. Hudson, K. M. Newkirk, H. L. Chandler, C. Choi, S. L. Fossey, A. E. Parent, and D. F. Kusewitt. Cutaneous wound reepithelialization is compromised in mice lacking functional slug (snai2). *Journal of dermatological science*, 56(1):19–26, 2009.

[61] D. Jia, J. T. George, S. C. Tripathi, D. L. Kundnani, M. Lu, S. M. Hanash, J. N. Onuchic, M. K. Jolly, and H. Levine. Testing the gene expression classification of the EMT spectrum. *Physical biology*, 16(2):025002, 2019.

[62] D. Jia, M. K. Jolly, M. Boareto, P. Parsana, S. M. Mooney, K. J. Pienta, H. Levine, and E. Ben-Jacob. Ovol guides the epithelial-hybrid-mesenchymal transition. *Oncotarget*, 6(17):15436, 2015.

[63] D. Jia, M. K. Jolly, S. C. Tripathi, P. Den Hollander, B. Huang, M. Lu, M. Celiktas, E. Ramirez-Peña, E. Ben-Jacob, J. N. Onuchic, et al. Distinguishing mechanisms underlying EMT tristability. *Cancer convergence*, 1(1):1–19, 2017.

[64] D. Jia, X. Li, F. Bocci, S. Tripathi, Y. Deng, M. K. Jolly, J. N. Onuchic, and H. Levine. Quantifying cancer epithelial-mesenchymal plasticity and its association with stemness and immune response. *Journal of clinical medicine*, 8(5):725, 2019.

[65] W. Jia, A. Deshmukh, S. A. Mani, M. K. Jolly, and H. Levine. A possible role for epigenetic feedback regulation in the dynamics of the epithelial–mesenchymal transition (EMT). *Physical biology*, 16(6):066004, 2019.

[66] W. Jia, S. Tripathi, P. Chakraborty, A. Chedere, A. Rangarajan, H. Levine, and M. K. Jolly. Epigenetic feedback and stochastic partitioning during cell division can drive resistance to EMT. *Oncotarget*, 11(27):2611, 2020.

[67] S. Jin, C. F. Guerrero-Juarez, L. Zhang, I. Chang, R. Ramos, C.-H. Kuan, P. Myung, M. V. Plikus, and Q. Nie. Inference and analysis of cell-cell communication using cellchat. *Nature Communications*, 12(1):1–20, 2021.

[68] S. Jin, A. L. MacLean, T. Peng, and Q. Nie. scepath: energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data. *Bioinformatics*, 34(12):2077–2086, 2018.

[69] M. K. Jolly, M. Boareto, B. G. Debeb, N. Aceto, M. C. Farach-Carson, W. A. Woodward, and H. Levine. Inflammatory breast cancer: a model for investigating cluster-based dissemination. *NPJ Breast Cancer*, 3(1):1–8, 2017.

[70] M. K. Jolly, M. Boareto, B. Huang, D. Jia, M. Lu, E. Ben-Jacob, J. N. Onuchic, and H. Levine. Implications of the hybrid epithelial/mesenchymal phenotype in metastasis. *Frontiers in oncology*, 5:155, 2015.

[71] M. K. Jolly and T. Celià-Terrassa. Dynamics of phenotypic heterogeneity associated with EMT and stemness during cancer progression. *Journal of Clinical Medicine*, 8(10):1542, 2019.

[72] M. K. Jolly, B. Huang, M. Lu, S. A. Mani, H. Levine, and E. Ben-Jacob. Towards elucidating the connection between epithelial-mesenchymal transitions and stemness. *Journal of The Royal Society Interface*, 11(101):20140962, 2014.

[73] M. K. Jolly and H. Levine. Computational systems biology of epithelial-hybrid-mesenchymal transitions. *Current Opinion in Systems Biology*, 3:1–6, 2017.

[74] M. K. Jolly, S. C. Tripathi, D. Jia, S. M. Mooney, M. Celiktas, S. M. Hanash, S. A. Mani, K. J. Pienta, E. Ben-Jacob, and H. Levine. Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget*, 7(19):27067, 2016.

[75] M. K. Jolly, C. Ward, M. S. Eapen, S. Myers, O. Hallgren, H. Levine, and S. S. Sohal. Epithelial–mesenchymal transition, a spectrum of states: Role in lung development, homeostasis, and disease. *Developmental dynamics*, 247(3):346–358, 2018.

[76] M. K. Jolly, K. E. Ware, S. Gilja, J. A. Somarelli, and H. Levine. EMT and MET: necessary or permissive for metastasis? *Molecular oncology*, 11(7):755–769, 2017.

[77] R. Kalluri, R. A. Weinberg, et al. The basics of epithelial-mesenchymal transition. *The Journal of clinical investigation*, 119(6):1420–1428, 2009.

[78] H.-W. Kang, M. Crawford, M. Fabbri, G. Nuovo, M. Garofalo, S. P. Nana-Sinkam, and A. Friedman. A mathematical model for microrna in lung cancer. *PLoS One*, 8(1):e53663, 2013.

[79] X. Kang, J. Wang, and C. Li. Exposing the underlying relationship of cancer metastasis to metabolism and epithelial-mesenchymal transitions. *Iscience*, 21:754–772, 2019.

[80] L. V. Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

[81] L. G. Karacosta, B. Anchang, N. Ignatiadis, S. C. Kimmey, J. A. Benson, J. B. Shrager, R. Tibshirani, S. C. Bendall, and S. K. Plevritis. Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell resolution. *Nature Communications*, 10(1):1–15, 2019.

[82] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, et al. Sc3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14(5):483–486, 2017.

[83] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

[84] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[85] V. Kohar and M. Lu. Role of noise and parametric variation in the dynamics of gene regulatory circuits. *NPJ systems biology and applications*, 4(1):1–11, 2018.

[86] Y. Koren. Drawing graphs by eigenvectors: theory and practice. *Computers & Mathematics with Applications*, 49(11-12):1867–1888, 2005.

[87] A. M. Krebs, J. Mitschke, M. Lasierra Losada, O. Schmalhofer, M. Boerries, H. Busch, M. Boettcher, D. Mougiakakos, W. Reichardt, P. Bronsert, et al. The EMT-activator zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic cancer. *Nature cell biology*, 19(5):518–529, 2017.

[88] D. Kuang, C. Ding, and H. Park. Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 2012 SIAM international conference on data mining*, pages 106–117. SIAM, 2012.

[89] D. Kuang, S. Yun, and H. Park. Symnmf: nonnegative low-rank approximation of a similarity matrix for graph clustering. *Journal of Global Optimization*, 62(3):545–574, 2015.

[90] M. P. Kumar, J. Du, G. Lagoudas, Y. Jiao, A. Sawyer, D. C. Drummond, D. A. Lauffenburger, and A. Raue. Analysis of single-cell RNA-seq identifies cell-cell communication associated with tumor characteristics. *Cell Reports*, 25(6):1458–1468, 2018.

[91] A. W. Lambert, D. R. Pattabiraman, and R. A. Weinberg. Emerging biological principles of metastasis. *Cell*, 168(4):670–691, 2017.

[92] S. Lamouille, J. Xu, and R. Derynck. Molecular mechanisms of epithelial-mesenchymal transition. *Nature reviews Molecular cell biology*, 15(3):178–196, 2014.

[93] A. D. Lander, K. K. Gokoffski, F. Y. M. Wan, Q. Nie, and A. L. Calof. Cell lineages and the logic of proliferative control. *PLoS Biology*, 7(1):e1000015, 2009.

[94] D. A. Lawson, N. R. Bhakta, K. Kessenbrock, K. D. Prummel, Y. Yu, K. Takai, A. Zhou, H. Eyob, S. Balakrishnan, C.-Y. Wang, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature*, 526(7571):131–135, 2015.

[95] A. Lecharpentier, P. Vielh, P. Perez-Moreno, D. Planchard, J. Soria, and F. Farace. Detection of circulating tumour cells with a hybrid (epithelial/mesenchymal) phenotype in patients with metastatic non-small cell lung cancer. *British journal of cancer*, 105(9):1338–1341, 2011.

[96] B. Lee, A. Villarreal-Ponce, M. Fallahi, J. Ovadia, P. Sun, Q.-C. Yu, S. Ito, S. Sinha, Q. Nie, and X. Dai. Transcriptional mechanisms link epithelial plasticity to adhesion and differentiation of epidermal progenitor cells. *Developmental cell*, 29(1):47–58, 2014.

[97] J. M. Lee, S. Dedhar, R. Kalluri, and E. W. Thompson. The epithelial–mesenchymal transition: new insights in signaling, development, and disease. *The Journal of cell biology*, 172(7):973–981, 2006.

[98] W. Lee, R. Lai, W. Li, and S. Osher. Generalized unnormalized optimal transport and its fast algorithms. *Journal of Computational Physics*, 436:110041, 2021.

[99] C. Li and G. Balazsi. A landscape view on the interplay between EMT and cancer metastasis. *NPJ systems biology and applications*, 4(1):1–9, 2018.

[100] C. Li, T. Hong, and Q. Nie. Quantifying the landscape and kinetic paths for epithelial–mesenchymal transition from a core circuit. *Physical Chemistry Chemical Physics*, 18(27):17949–17956, 2016.

[101] Q. Li, A. P. Hutchins, Y. Chen, S. Li, Y. Shan, B. Liao, D. Zheng, X. Shi, Y. Li, W.-Y. Chan, et al. A sequential EMT-MET mechanism drives the differentiation of human embryonic stem cells towards hepatocytes. *Nature Communications*, 8(1):1–12, 2017.

[102] X. Liu, H. Sun, J. Qi, L. Wang, S. He, J. Liu, C. Feng, C. Chen, W. Li, Y. Guo, et al. Sequential introduction of reprogramming factors reveals a time-sensitive requirement for individual factors and a sequential EMT–MET mechanism for optimal reprogramming. *Nature Cell Biology*, 15(7):829–838, 2013.

[103] W.-C. Lo, C.-S. Chou, K. K. Gokoffski, F. Y.-M. Wan, A. D. Lander, A. L. Calof, and Q. Nie. Feedback regulation in multistage cell lineages. *Mathematical biosciences and engineering: MBE*, 6(1):59, 2009.

[104] S. Lovisa, V. S. LeBleu, B. Tampe, H. Sugimoto, K. Vadnagara, J. L. Carstens, C.-C. Wu, Y. Hagos, B. C. Burckhardt, T. Pentcheva-Hoang, et al. Epithelial-to-mesenchymal transition induces cell cycle arrest and parenchymal damage in renal fibrosis. *Nature Medicine*, 21(9):998–1009, 2015.

[105] M. Lu, M. K. Jolly, H. Levine, J. N. Onuchic, and E. Ben-Jacob. MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination. *Proceedings of the National Academy of Sciences*, 110(45):18144–18149, 2013.

[106] W. Ma, A. Trusina, H. El-Samad, W. A. Lim, and C. Tang. Defining network topologies that can achieve biochemical adaptation. *Cell*, 138(4):760–773, 2009.

[107] A. L. MacLean, T. Hong, and Q. Nie. Exploring intermediate cell states through the lens of single cells. *Current Opinion in Systems Biology*, 9:32–41, 2018.

[108] S. A. Mani, W. Guo, M.-J. Liao, E. N. Eaton, A. Ayyanan, A. Y. Zhou, M. Brooks, F. Reinhard, C. C. Zhang, M. Shipitsin, et al. The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell*, 133(4):704–715, 2008.

[109] H. Mi, A. Muruganujan, D. Ebert, X. Huang, and P. D. Thomas. Panther version 14: more genomes, a new panther go-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, 47(D1):D419–D426, 2019.

[110] D. S. Micalizzi, S. M. Farabaugh, and H. L. Ford. Epithelial-mesenchymal transition in cancer: parallels between normal development and tumor progression. *Journal of mammary gland biology and neoplasia*, 15(2):117–134, 2010.

[111] M. Mojtahedi, A. Skupin, J. Zhou, I. G. Castaño, R. Y. Leong-Quong, H. Chang, K. Trachana, A. Giuliani, and S. Huang. Cell fate decision as high-dimensional critical state transition. *PLoS Biology*, 14(12):e2000640, 2016.

[112] A.-P. Morel, M. Lièvre, C. Thomas, G. Hinkal, S. Ansieau, and A. Puisieux. Generation of breast cancer stem cells through epithelial-mesenchymal transition. *PloS one*, 3(8):e2888, 2008.

[113] A. Moustakas and C.-H. Heldin. Signaling networks guiding epithelial–mesenchymal transitions during embryogenesis and cancer progression. *Cancer science*, 98(10):1512–1520, 2007.

[114] M. Nassour, Y. Idoux-Gillet, A. Selmi, C. Côme, M.-L. M. Faraldo, M.-A. Deugnier, and P. Savagner. Slug controls stem/progenitor cell growth dynamics during mammary gland morphogenesis. *PloS one*, 7(12):e53498, 2012.

[115] Q. H. Nguyen, N. Pervolarakis, K. Blake, D. Ma, R. T. Davis, N. James, A. T. Phung, E. Willey, R. Kumar, E. Jabart, et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nature communications*, 9(1):1–12, 2018.

[116] Q. Nie. Stem cells: a window of opportunity in low-dimensional EMT space. *Oncotarget*, 9(61):31790, 2018.

[117] Q. Nie, L. Qiao, Y. Qiu, L. Zhang, and W. Zhao. Noise control and utility: From regulatory network to spatial patterning. *Science China Mathematics*, 63(3):425–440, 2020.

[118] M. A. Nieto, R. Y.-J. Huang, R. A. Jackson, and J. P. Thiery. EMT: 2016. *Cell*, 166(1):21–45, 2016.

[119] P. Parsana, S. R. Amend, J. Hernandez, K. J. Pienta, and A. Battle. Identifying global expression patterns and key regulators in epithelial to mesenchymal transition through multi-study integration. *BMC cancer*, 17(1):1–14, 2017.

[120] I. Pastushenko and C. Blanpain. EMT transition states during tumor progression and metastasis. *Trends in cell biology*, 29(3):212–226, 2019.

[121] I. Pastushenko, A. Brisebarre, A. Sifrim, M. Fioramonti, T. Revenco, S. Boumahdi, A. Van Keymeulen, D. Brown, V. Moers, S. Lemaire, et al. Identification of the tumour transition states occurring during EMT. *Nature*, 556(7702):463–468, 2018.

[122] A. Puisieux, T. Brabletz, and J. Caramel. Oncogenic roles of EMT-inducing transcription factors. *Nature Cell Biology*, 16(6):488–494, 2014.

[123] S. V. Puram, I. Tirosh, A. S. Parikh, A. P. Patel, K. Yizhak, S. Gillespie, C. Rodman, C. L. Luo, E. A. Mroz, K. S. Emerick, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171(7):1611–1624, 2017.

[124] L. Qiao, W. Zhao, C. Tang, Q. Nie, and L. Zhang. Network topologies that can achieve dual function of adaptation and noise attenuation. *Cell systems*, 9(3):271–285, 2019.

[125] Q. Qiu, P. Hu, X. Qiu, K. W. Govek, P. G. Cámara, and H. Wu. Massively parallel and time-resolved rna sequencing in single cells with scnt-seq. *Nature methods*, 17(10):991–1001, 2020.

[126] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979–982, 2017.

[127] Y. Qiu, W. Chen, and Q. Nie. Stochastic dynamics of cell lineage in tissue homeostasis. *Discrete and continuous dynamical systems. Series B*, 24(8):3971, 2019.

[128] Y. Qiu, L. Fung, T. F. Schilling, and Q. Nie. Multiple morphogens and rapid elongation promote segmental patterning during development. *PLoS computational biology*, 17(6):e1009077, 2021.

[129] C. Rackauckas, T. Schilling, and Q. Nie. Mean-independent noise control of cell fates via intermediate states. *Iscience*, 3:11–20, 2018.

[130] D. Ramirez, V. Kohar, and M. Lu. Toward modeling context-specific EMT regulatory networks using temporal single cell RNA-seq data. *Frontiers in molecular biosciences*, 7:54, 2020.

[131] C. Revenu and D. Gilmour. EMT 2.0: shaping epithelia through collective migration. *Current opinion in genetics & development*, 19(4):338–342, 2009.

[132] A. D. Rhim, E. T. Mirek, N. M. Aiello, A. Maitra, J. M. Bailey, F. McAllister, M. Reichert, G. L. Beatty, A. K. Rustgi, R. H. Vonderheide, et al. EMT and dissemination precede pancreatic tumor formation. *Cell*, 148(1-2):349–361, 2012.

[133] L. Rosanò, F. Spinella, V. Di Castro, S. Decandia, M. R. Nicotra, P. G. Natali, and A. Bagnato. Endothelin-1 is required during epithelial to mesenchymal transition in ovarian cancer progression. *Experimental Biology and Medicine*, 231(6):1128–1131, 2006.

[134] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[135] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015.

[136] L. R. Saunders and D. R. McClay. Sub-circuits of a gene regulatory network control a developmental epithelial-mesenchymal transition. *Development*, 141(7):1503–1513, 2014.

[137] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[138] Y. Sha, D. Haensel, G. Gutierrez, H. Du, X. Dai, and Q. Nie. Intermediate cell states in epithelial-to-mesenchymal transition. *Physical biology*, 16(2):021001, 2019.

[139] Y. Sha, S. Wang, F. Bocci, P. Zhou, and Q. Nie. Inference of intercellular communications and multilayer gene-regulations of epithelial–mesenchymal transition from single-cell transcriptomic data. *Frontiers in genetics*, page 1700, 2021.

[140] Y. Sha, S. Wang, P. Zhou, and Q. Nie. Inference and multiscale model of epithelial-to-mesenchymal transition via single-cell transcriptomic data. *Nucleic Acids Research*, 48(17):9505–9520, 2020.

[141] S. H. Shirley, L. G. Hudson, J. He, and D. F. Kusewitt. The skinny on slug. *Molecular carcinogenesis*, 49(10):851–861, 2010.

[142] S. N. Steinway, J. G. Zañudo, W. Ding, C. B. Rountree, D. J. Feith, T. P. Loughran, and R. Albert. Network modeling of TGFβ signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint sonic hedgehog and wnt pathway activation. *Cancer Research*, 74(21):5963–5977, 2014.

[143] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):1–16, 2018.

[144] G. Swetha, V. Chandra, S. Phadnis, and R. Bhonde. Glomerular parietal epithelial cells of adult murine kidney undergo EMT to generate cells with traits of renal progenitors. *Journal of cellular and molecular medicine*, 15(2):396, 2011.

[145] C. H. Ta, Q. Nie, and T. Hong. Controlling stochasticity in epithelial-mesenchymal transition through multiple intermediate cellular states. *Discrete and continuous dynamical systems. Series B*, 21(7):2275, 2016.

[146] W. L. Tam and R. A. Weinberg. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nature medicine*, 19(11):1438–1449, 2013.

[147] H. Tanaka and S. Ogishima. Network biology approach to epithelial–mesenchymal transition in cancer metastasis: three stage theory. *Journal of molecular cell biology*, 7(3):253–266, 2015.

[148] X.-J. Tian, M. V. Ferro, and H. Goetz. Modeling ncRNA-mediated circuits in cell fate decision. *Computational Biology of Non-Coding RNA*, pages 411–426, 2019.

[149] X.-J. Tian, H. Zhang, and J. Xing. Coupled reversible and irreversible bistable switches underlying TGFβ-induced epithelial to mesenchymal transition. *Biophysical journal*, 105(4):1079–1089, 2013.

[150] I. Tirosh, B. Izar, S. M. Prakadan, M. H. Wadsworth, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, 2016.

[151] A. Tong, J. Huang, G. Wolf, D. Van Dijk, and S. Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International Conference on Machine Learning*, pages 9526–9536. PMLR, 2020.

[152] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014.

[153] S. Tripathi, P. Chakraborty, H. Levine, and M. K. Jolly. A mechanism for epithelial-mesenchymal heterogeneity in a population of cancer cells. *PLoS Computational Biology*, 16(2):e1007619, 2020.

[154] S. Tripathi, H. Levine, and M. K. Jolly. The physics of cellular decision making during epithelial–mesenchymal transition. *Annual Review of Biophysics*, 49:1–18, 2020.

[155] S. S. Varankar, M. More, A. Abraham, K. Pansare, B. Kumar, N. J. Narayanan, M. K. Jolly, A. M. Mali, and S. A. Bapat. Functional balance between tcf21–slug defines cellular plasticity and migratory modalities in high grade serous ovarian cancer cell lines. *Carcinogenesis*, 41(4):515–526, 2020.

[156] D. E. Wagner and A. M. Klein. Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics*, 21(7):410–427, 2020.

[157] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature Methods*, 14(4):414–416, 2017.

[158] D. Wang, C. Cai, X. Dong, Q. C. Yu, X.-O. Zhang, L. Yang, and Y. A. Zeng. Identification of multipotent mammary stem cells by protein C receptor expression. *Nature*, 517(7532):81–84, 2015.

[159] J.-K. Wang, W.-J. Wang, H.-Y. Cai, B.-B. Du, P. Mai, L.-J. Zhang, W. Ma, Y.-G. Hu, S.-F. Feng, and G.-Y. Miao. Mfap2 promotes epithelial–mesenchymal transition in gastric cancer cells by activating TGF-β/smad2/3 signaling pathway. *OncoTargets and therapy*, 11:4001, 2018.

[160] S. Wang, M. L. Drummond, C. F. Guerrero-Juarez, E. Tarapore, A. L. MacLean, A. R. Stabell, S. C. Wu, G. Gutierrez, B. T. That, C. A. Benavente, et al. Single cell transcriptomics of human epidermis identifies basal stem cell transition states. *Nature Communications*, 11(1):1–14, 2020.

[161] S. Wang, M. Karikomi, A. L. MacLean, and Q. Nie. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Research*, 47(11):e66–e66, 2019.

[162] W. Wang, D. Douglas, J. Zhang, S. Kumari, M. S. Enuameh, Y. Dai, C. T. Wallace, S. C. Watkins, W. Shu, and J. Xing. Live-cell imaging and analysis reveal cell phenotypic transition dynamics inherently missing in snapshot data. *Science advances*, 6(36):eaba9319, 2020.

[163] Y. Wang, R. Wang, S. Zhang, S. Song, C. Jiang, G. Han, M. Wang, J. Ajani, A. Futreal, and L. Wang. iTALK: an R package to characterize and illustrate intercellular communication. *BioRxiv*, page 507871, 2019.

[164] K. Watanabe, A. Villarreal-Ponce, P. Sun, M. L. Salmans, M. Fallahi, B. Andersen, and X. Dai. Mammary morphogenesis and regeneration require the inhibition of EMT at

terminal end buds by ovol2 transcriptional repressor. *Developmental cell*, 29(1):59–74, 2014.

[165] J. D. Welch, A. J. Hartemink, and J. F. Prins. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome biology*, 17(1):1–15, 2016.

[166] M. K. Wendt, T. M. Allington, and W. P. Schiemann. Mechanisms of the epithelial–mesenchymal transition by TGF-β. *Future Oncology*, 5(8):1145–1168, 2009.

[167] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20(1):1–9, 2019.

[168] J. Xing and X.-J. Tian. Investigating epithelial-to-mesenchymal transition with integrated computational and experimental approaches. *Physical biology*, 16(3):031001, 2019.

[169] J. Xu, S. Lamouille, and R. Derynck. TGF-β-induced epithelial to mesenchymal transition. *Cell research*, 19(2):156–172, 2009.

[170] J. Yang, P. Antin, G. Berx, C. Blanpain, T. Brabletz, M. Bronner, K. Campbell, A. Cano, J. Casanova, G. Christofori, et al. Guidelines and definitions for research on epithelial–mesenchymal transition. *Nature reviews Molecular cell biology*, 21(6):341–352, 2020.

[171] K. D. Yang and C. Uhler. Scalable unbalanced optimal transport using generative adversarial networks. *arXiv preprint arXiv:1810.11447*, 2018.

[172] L. Yang and G. E. Karniadakis. Potential flow generator with l_2 optimal transport regularity for generative models. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[173] X. Ye, W. L. Tam, T. Shibue, Y. Kaygusuz, F. Reinhardt, E. Ng Eaton, and R. A. Weinberg. Distinct EMT programs control normal mammary stem cells and tumour-initiating cells. *Nature*, 525(7568):256–260, 2015.

[174] M. Yu, A. Bardia, B. S. Wittner, S. L. Stott, M. E. Smas, D. T. Ting, S. J. Isakoff, J. C. Ciciliano, M. N. Wells, A. M. Shah, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *science*, 339(6119):580–584, 2013.

[175] P. Yu, Q. Nie, C. Tang, and L. Zhang. Nanog induced intermediate state in regulating stem cell differentiation and reprogramming. *BMC systems biology*, 12(1):1–13, 2018.

[176] S. Zadran, R. Arumugam, H. Herschman, M. E. Phelps, and R. Levine. Surprisal analysis characterizes the free energy time course of cancer cells undergoing epithelial-to-mesenchymal transition. *Proceedings of the National Academy of Sciences*, 111(36):13235–13240, 2014.

[177] J. A. Zepp, W. J. Zacharias, D. B. Frank, C. A. Cavanaugh, S. Zhou, M. P. Morley, and E. E. Morrisey. Distinct mesenchymal lineages and niches promote epithelial self-renewal and myofibrogenesis in the lung. *Cell*, 170(6):1134–1148, 2017.

[178] J. Zhang, X.-J. Tian, and J. Xing. Signal transduction pathways of EMT induced by TGF-β, SHH, and WNT and their crosstalks. *Journal of clinical medicine*, 5(4):41, 2016.

[179] J. Zhang, X.-J. Tian, H. Zhang, Y. Teng, R. Li, F. Bai, S. Elankumaran, and J. Xing. TGF-β–induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Science signaling*, 7(345):ra91–ra91, 2014.

[180] S. Zhang, A. Afanassiev, L. Greenstreet, T. Matsumoto, and G. Schiebinger. Optimal transport analysis reveals trajectories in steady-state systems. *PLoS computational biology*, 17(12):e1009466, 2021.

[181] C. Zhao, Q. Wang, B. Wang, Q. Sun, Z. He, J. Hong, F. Kuehn, E. Liu, and Z. Zhang. Igf-1 induces the epithelial-mesenchymal transition via stat5 in hepatocellular carcinoma. *Oncotarget*, 8(67):111922, 2017.

[182] X. Zheng, J. L. Carstens, J. Kim, M. Scheible, J. Kaye, H. Sugimoto, C.-C. Wu, V. S. LeBleu, and R. Kalluri. Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. *Nature*, 527(7579):525–530, 2015.

[183] X. Zheng, S. Jin, Q. Nie, and X. Zou. scRCMF: Identification of cell subpopulations and transition states from single-cell transcriptomes. *IEEE Transactions on Biomedical Engineering*, 67(5):1418–1428, 2019.

[184] Y. Zhou, B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, C. Benner, and S. K. Chanda. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, 10(1):1–10, 2019.

[185] L. Zhu, J. Lei, L. Klei, B. Devlin, and K. Roeder. Semisoft clustering of single-cell data. *Proceedings of the National Academy of Sciences*, 116(2):466–471, 2019.

[186] Y. Zhu, Y. Qiu, W. Chen, Q. Nie, and A. D. Lander. Scaling a dpp morphogen gradient through feedback control of receptors and co-receptors. *Developmental cell*, 53(6):724–739, 2020.

[187] Z. Zhu, X. Li, K. Liu, and Q. Li. Dropping symmetry for fast symmetric nonnegative matrix factorization. *Advances in Neural Information Processing Systems*, 31, 2018.

[188] J. Zhuang, N. Dvornek, X. Li, S. Tatikonda, X. Papademetris, and J. Duncan. Adaptive checkpoint adjoint method for gradient estimation in neural ode. In *International Conference on Machine Learning*, pages 11639–11649. PMLR, 2020.

[189] J. Zhuang, N. C. Dvornek, S. Tatikonda, and J. S. Duncan. Mali: A memory efficient and reverse accurate integrator for neural odes. *arXiv preprint arXiv:2102.04668*, 2021.

Appendix A

Additional file for Chapter 2

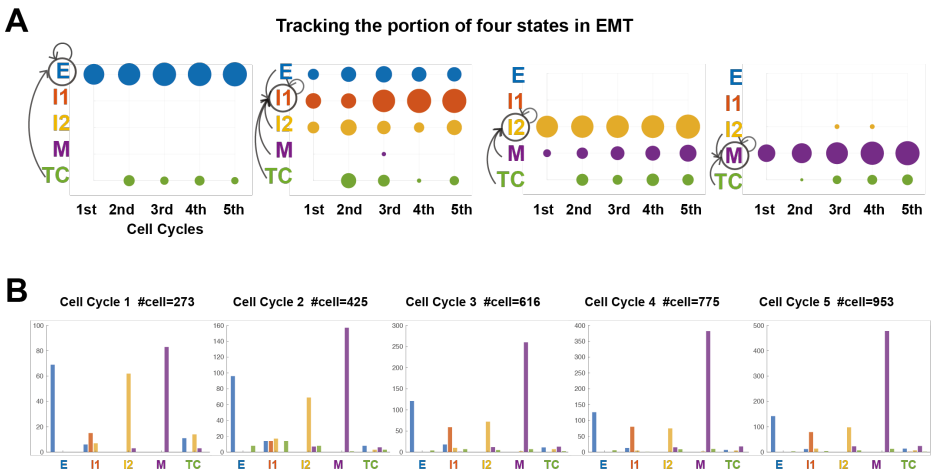

Figure A.1: Simulation of the model when I1 state is non-proliferative. (A) A simulation dataset: the proportion of each state induced by the previous cell states at the end of each cell cycle. The size of the dot is proportional to the number of cells, and the color denotes the cell states of the mother cell. The arrows represent the occurred state transitions and the circle represents the state of the daughter cell. (B) Histogram of the number of cell population at the end of each cycle. The color denotes the mother cell states. The x-labels represent the states of the daughter cell.

Appendix B

Additional file for Chapter 3

	hESC	SCC	Intestine	Liver	Lung	Skin
Number top marker genes	4	4	5	5	4	5
Number top transition genes	4	6	8	8	6	6
Percentage of top edges	69%	84%	86%	80%	74%	80%
Thresholds of CPI to select TC	0.34	0.34	0.34	0.2	0.1	0.2

Table B.1: Thresholds of CPI values and to select top genes and edges in PIDC

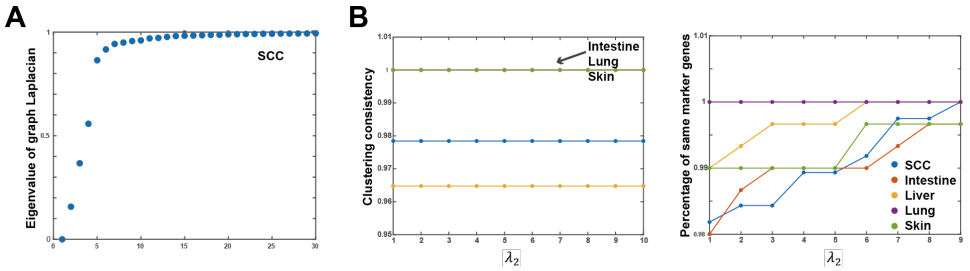

Figure B.1: Number of clusters predicted based on the graph Laplacian and robustness of finding marker genes with varied λ_2. (A) The first 30 sorted eigenvalues of the graph Laplacian of the constructed consensus matrix M for the SCC dataset. (B) Clustering accuracy and robustness of identifying marker genes when λ_2 varies from 1 to 10 in the SCC and mouse embryonic development datasets. Left: the consistency of clustering results based on the new decomposed $\overline{H}$ compared to the pre-inferred clusters. Right: the percentage of finding the same top 100 marker genes for each cluster compared to the marker genes identified when λ_2.

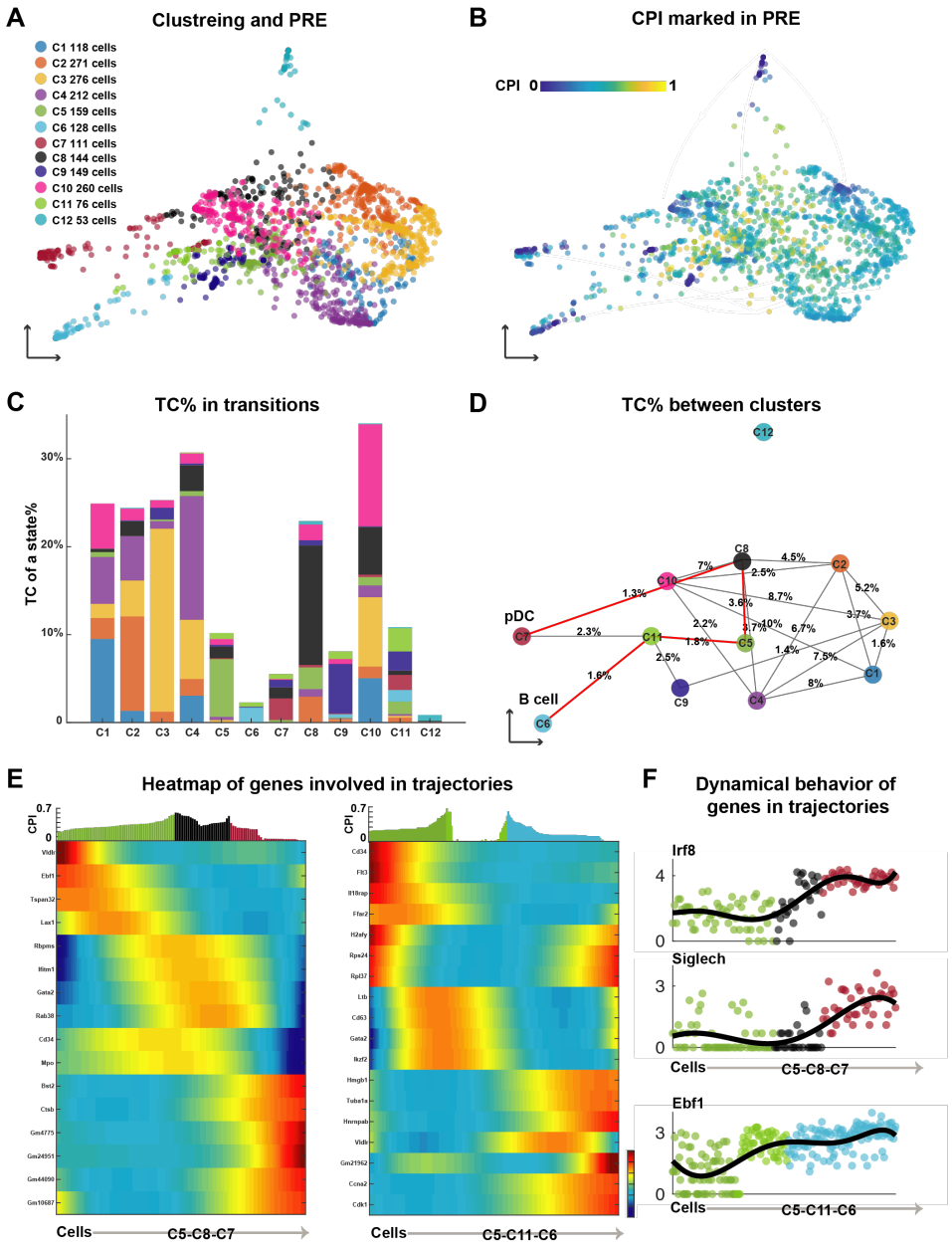

Figure B.2: Analyzing the dataset of mouse hematopoietic progenitors using QuanTC. (A-B) Visualization of cells via PRE. (A) Clustering result of 1957 cells after preprocess of the data. Each dot represents a cell and is colored by its cluster. (B) Each dot is colored by its CPI value. (C) Percentage of TC associated with each cluster relative to the total number of TC. (D) Percentage of TC between clusters relative to the entire cell population size. The lines show the potential transitions between clusters and the dots located at the cluster centers show the different clusters. Red lines show the potential transition trajectories related to clusters C6 and C7. (E) Heatmap of normalized expression of marker genes and transition genes. Columns represent cells ordered along the transition trajectory and rows represent genes. Coloring represents the normalized expression value of each gene. Top: CPI values of each cell along the transition trajectory. (F) Expression levels of known lineage markers with cells ordered along the transition trajectories. Solid lines, smoothed expression curves for each gene in the transition trajectory.

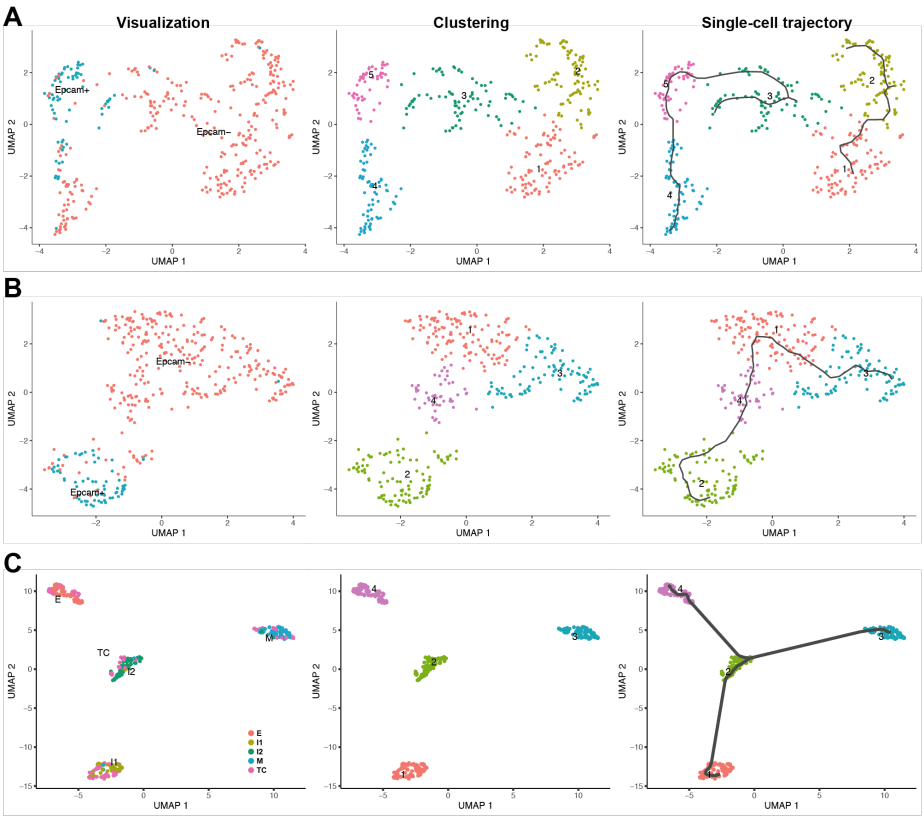

Figure B.3: Analyzing SCC and simulation datasets using Monocle 3. (A) Dimensionality reduction of scRNA-seq data using UMAP coloring by the surface marker (left) and by the clusters identified by Monocle 3 (middle, right) using the SCC dataset. Dots represent single cells. Right: Pseudotime-ordering trajectory of scRNA-seq data using Monocle 3. (B) Dimensionality reduction of scRNA-seq data using UMAP coloring by the surface marker (left) and by the clusters identified by Monocle 3 (middle, right) using the feature selected (top 3000 genes) SCC dataset from QuanTC. Dots represent single cells. Right: Pseudotime-ordering trajectory of scRNA-seq data using Monocle 3. (C) Dimensionality reduction of first cell cycle simulation dataset using UMAP coloring by the known cell types (left) and by the clusters identified by Monocle 3 (middle, right). Dots represent single cells. Right: Pseudotime-ordering trajectory of the data using Monocle 3.

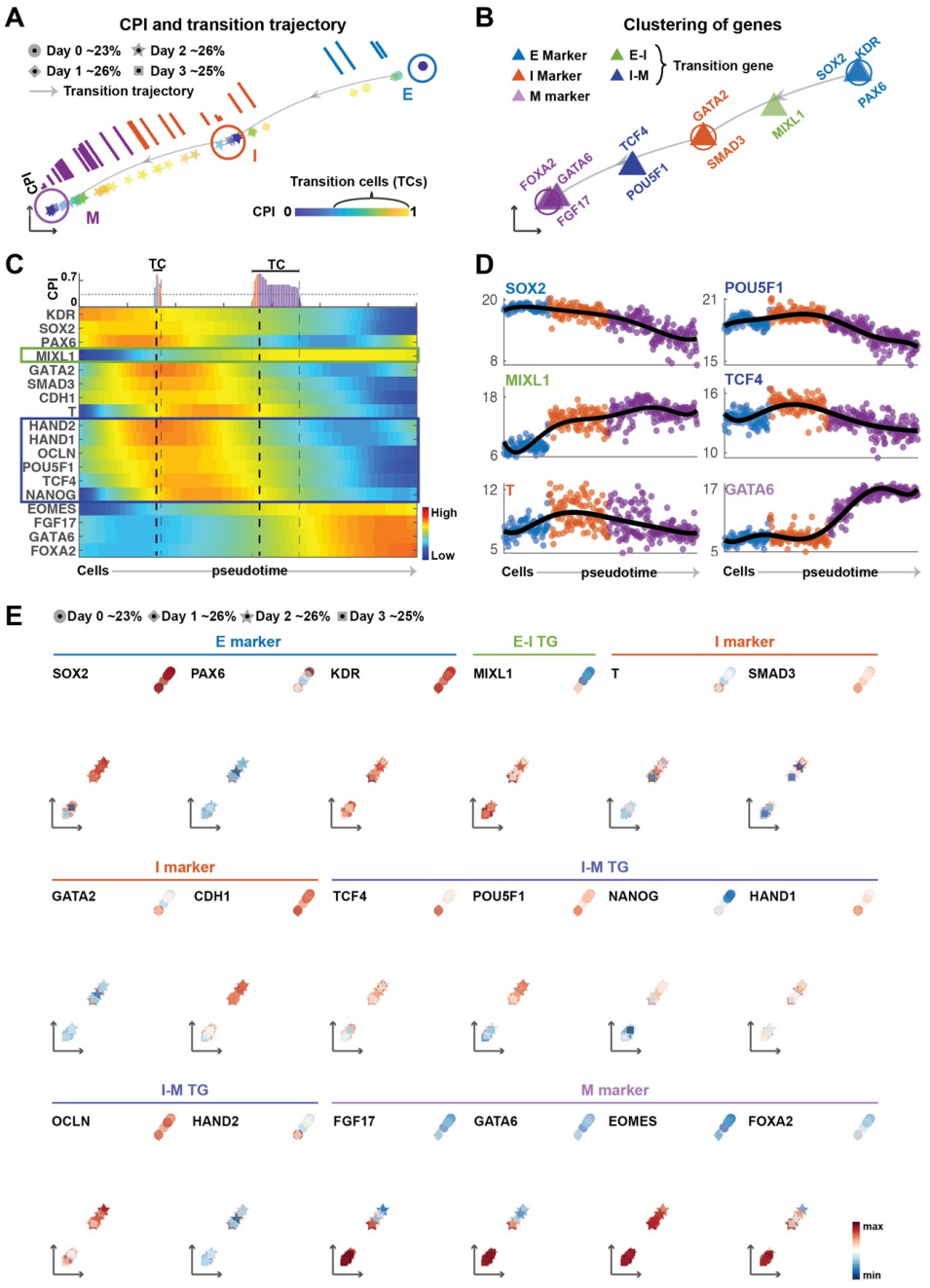

Figure B.4: Analyzing EMT during hepatic differentiation of hESCs using QuanTC.

Figure B.4 *(continued)*: (A) Transition trajectory inference. Each dot represents one cell colored by the value of CPI and its shape represents the time when the data collected given in the original study. The percentage for each cell type is the percentage of a given cell type over the entire cell population size. The cells surrounded by larger circles with relatively low CPI are considered as stable cells. The remaining cells with higher CPI are considered as TC. Arrows indicate the transition direction of EMT. Top, CPI of the cells colored by identified states. (B) Visualization of top marker genes and transition genes between states. Each triangle represents a gene colored by its type. Arrows indicate the transition direction of EMT. (C) Heat map of normalized expression of marker genes and transition genes. Columns represent cells ordered along the transition trajectory and rows represent genes. Coloring represents the normalized expression of each gene. Transition genes are marked in the box. Top: CPI values of each cell along the transition trajectory. (D) Expression levels of top transition genes with cells ordered along the most probable transition trajectories. Solid lines, smoothed expression curves for each gene in the transition trajectory. (E) Dimensionality reduction of the data using QuanTC coloring for top marker genes and transition genes. Every dot represents a single cell shaped by its real time and the color scale represents the normalized expression of the respective genes.

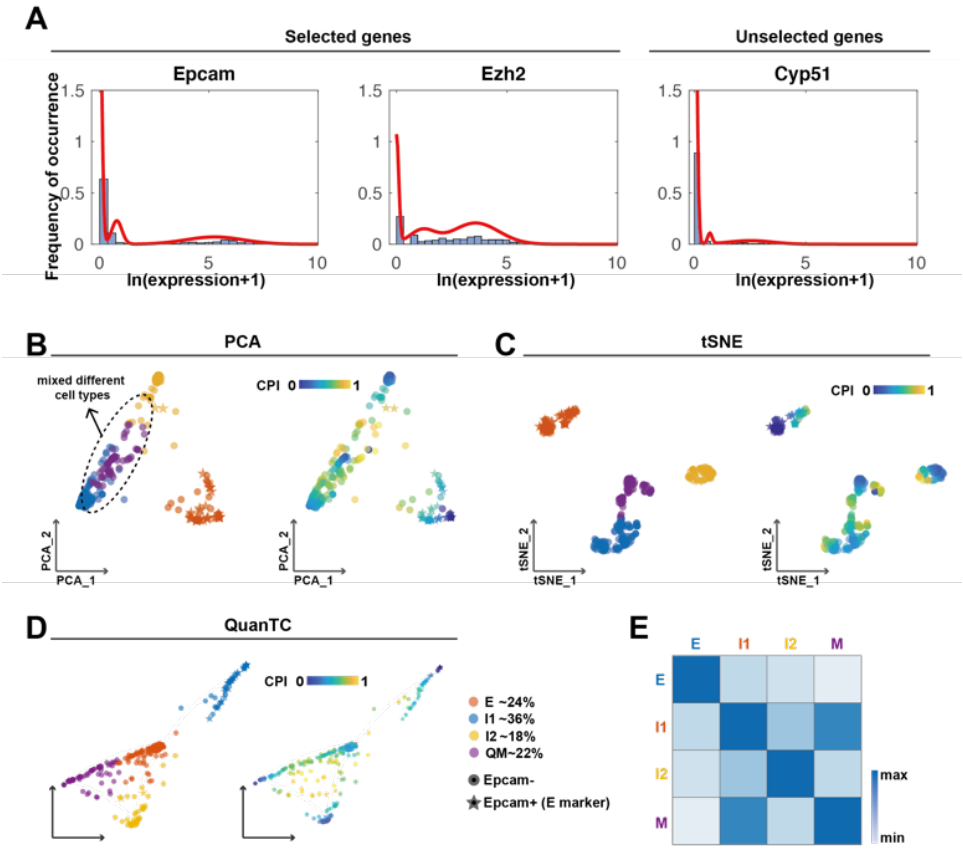

Figure B.5: Analyzing SCC dataset using QuanTC. (A) Histograms of the expression levels of selected informative genes and unselected genes. The red curve represents the fitted Gaussian mixture model. (B-D) Dimensionality reduction of the scRNA-seq data using PCA, tSNE and PRE visualization. Every dot represents a single cell colored by the clustering inferred by QuanTC (left) and CPI value (right) and its shape represents the FACS sorting criteria (Epcam+ or Epcam-). The percentage for each cell type is the percentage of a given cell type over the entire cell population size. (E) Similarities between states with color showing the value of similarities.

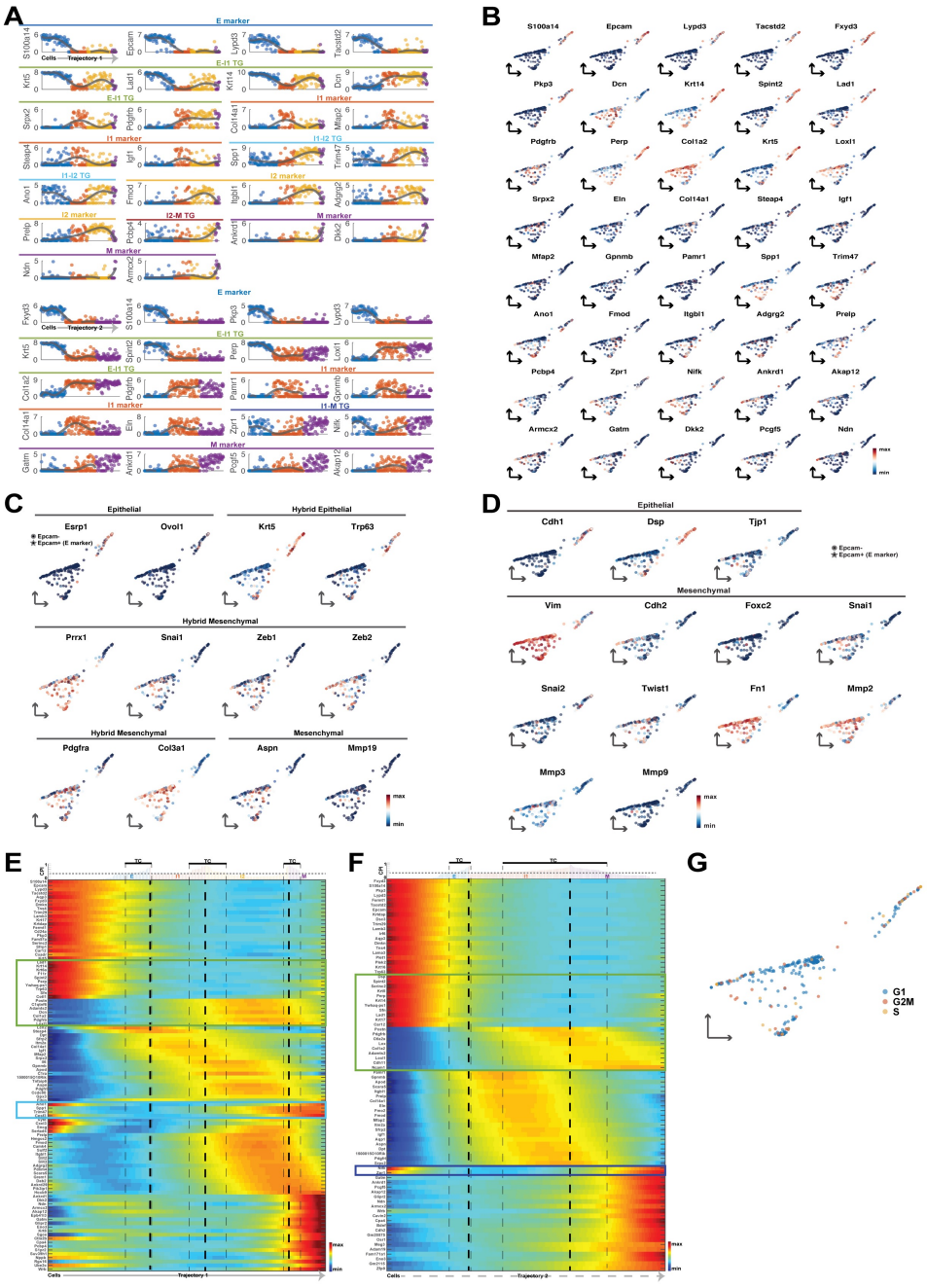

Figure B.6: Expression levels of top transition genes and marker genes involved in two main transition trajectories.

Figure B.6 *(continued)*: (A) Expression levels of top transition genes and marker genes with cells ordered along transition trajectory 1 and transition trajectory 2. Solid lines, smoothed expression curves for each gene in the transition trajectory. (B-D) Dimensionality reduction of the data using QuanTC coloring for top marker genes and transition genes. Every dot represents a single cell shaped by the FACS sorting criteria (Epcam+ or Epcam-) and the color scale represents the normalized expression of the respective genes. (C) Dimensionality reduction of the data using QuanTC coloring for pure epithelial genes, hybrid epithelial genes, hybrid mesenchymal genes and pure mesenchymal genes previously identified in the original study. (D) Dimensionality reduction of SCC data using QuanTC coloring for known epithelial genes and mesenchymal genes. (E-F) Heat map of normalized expression of top 20 marker genes and top 20 transition genes. Columns represent cells ordered along the transition trajectory 1 (E) and trajectory 2 (F) with rows represent genes. Coloring represents the normalized expression of each gene. Transition genes are marked in the box. Top: CPI values of each cell along the transition trajectories. (G) Dimensionality reduction of the data using QuanTC coloring for cell-cycle phase based on computed cell cycle scores. Every dot represents a single cell.

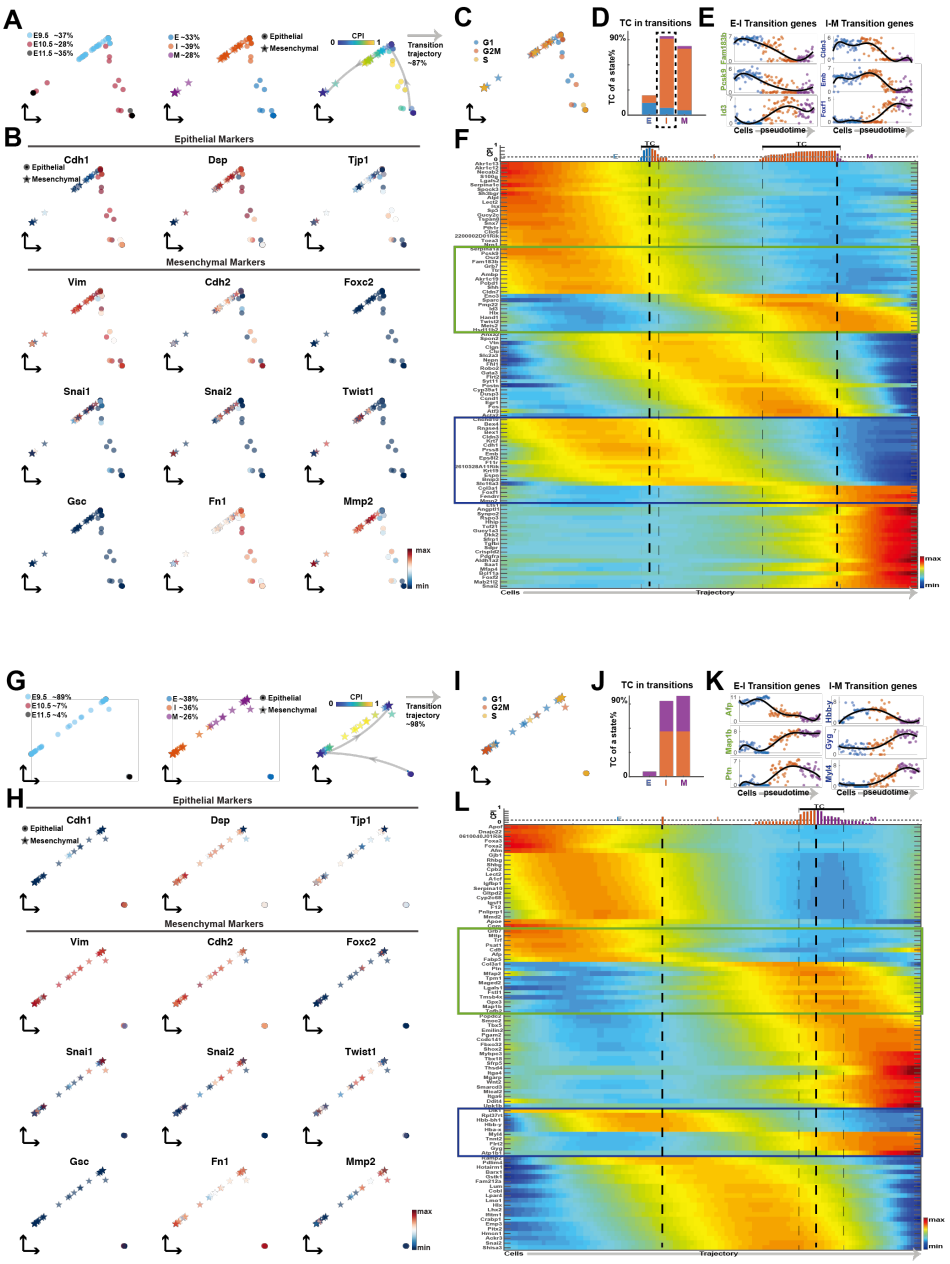

Figure B.7: Analysis of EMT during organogenesis in Intestine (top) and liver (bottom). (A) Visualization of cells via PRE. Each dot represents a single cell colored by the time of the tissue in mouse embryos from the original study on the corresponding dataset (left), clustering result from QuanTC (middle) and the value of CPI (right). The shape of each dot represents the cell states previously identified in the original study on the corresponding dataset. The thresholds to select TC are given in Table B.1. Arrowed solid line shows the main transition trajectory. The percentage for each cell type is the percentage of a given cell type over the entire cell population size.

Figure B.7 *(continued)*: (B) Dimensionality reduction of the data coloring for known epithelial genes and mesenchymal genes. Every dot represents a single cell and its shape represents the cell states previously identified in the original study on the corresponding dataset. The color scale represents the normalized expression of the respective genes. (C) Dimensionality reduction of the data coloring for cell-cycle phase based on computed cell cycle scores. Every dot represents a single cell and its shape represents the cell states previously identified in the original study on the corresponding dataset. (D) Percentage of TC associated with each state relative to the total number of TC. (E) Expression levels of top transition genes with cells ordered along the most probable transition trajectory. Solid lines, smoothed expression curves for each gene in the transition trajectory. (F) Heat map of normalized expression of top 20 marker genes and top 20 transition genes. Columns represent cells ordered along the transition trajectory 2 and rows represent genes. Coloring represents the normalized expression of each gene. Transition genes are marked in the box. Top: CPI values of each cell along the transition trajectory. (G) Visualization of cells via PRE. Each dot represents a single cell colored by the time of the tissue in mouse embryos from the original study on the corresponding dataset (left), clustering result from QuanTC (middle) and the value of CPI (right). The shape of each dot represents the cell states previously identified in the original study on the corresponding dataset. The thresholds to select TC are given in Table B.1. Arrowed solid line shows the main transition trajectory. (H) Dimensionality reduction of the data coloring for known epithelial genes and mesenchymal genes. Every dot represents a single cell and its shape represents the cell states previously identified in the original study on the corresponding dataset. The color scale represents the normalized expression of the respective genes. (I) Dimensionality reduction of the data coloring for cell-cycle phase based on computed cell cycle scores. Every dot represents a single cell and its shape represents the cell states previously identified in the original study on the corresponding dataset. (J) Percentage of TC associated with each state relative to the total number of TC. (K) Expression levels of top transition genes with cells ordered along the most probable transition trajectory. Solid lines, smoothed expression curves for each gene in the transition trajectory. (L) Heat map of normalized expression of top 20 marker genes and top 20 transition genes. Columns represent cells ordered along the transition trajectory 2 and rows represent genes. Coloring represents the normalized expression of each gene. Transition genes are marked in the box. Top: CPI values of each cell along the transition trajectory.

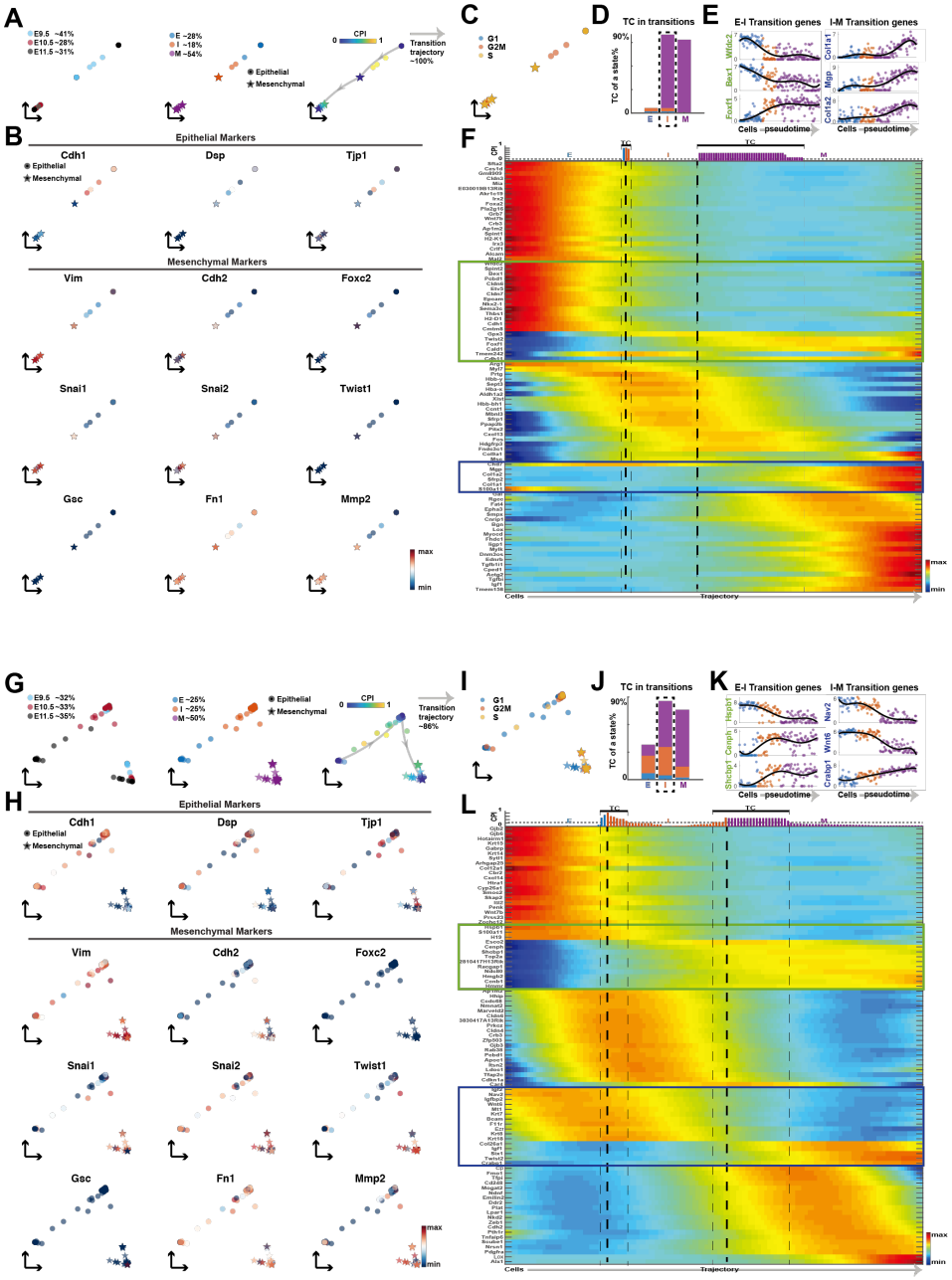

Figure B.8: Analysis of EMT during organogenesis in lung (top) and skin (bottom). (A) Visualization of cells via PRE. Each dot represents a single cell colored by the time of the tissue in mouse embryos from the original study on the corresponding dataset (left), clustering result from QuanTC (middle) and the value of CPI (right). The shape of each dot represents the cell states previously identified in the original study on the corresponding dataset. The thresholds to select TC are given in Table B.1. Arrowed solid line shows the main transition trajectory. The percentage for each cell type is the percentage of a given cell type over the entire cell population size.

Figure B.8 *(continued)*: (B) Dimensionality reduction of the data coloring for known epithelial genes and mesenchymal genes. Every dot represents a single cell and its shape represents the cell states previously identified in the original study on the corresponding dataset. The color scale represents the normalized expression of the respective genes. (C) Dimensionality reduction of the data coloring for cell-cycle phase based on computed cell cycle scores. Every dot represents a single cell and its shape represents the cell states previously identified in the original study on the corresponding dataset. (D) Percentage of TC associated each state relative to the total number of TC. (E) Expression levels of top transition genes with cells ordered along the most probable transition trajectory. Solid lines, smoothed expression curves for each gene in the transition trajectory. (F) Heat map of normalized expression of top 20 marker genes and top 20 transition genes. Columns represent cells ordered along the transition trajectory 2 and rows represent genes. Coloring represents the normalized expression of each gene. Transition genes are marked in the box. Top: CPI values of each cell along the transition trajectory. (G) Visualization of cells via PRE. Each dot represents a single cell colored by the time of the tissue in mouse embryos from the original study on the corresponding dataset (left), clustering result from QuanTC (middle) and the value of CPI (right). The shape of each dot represents the cell states previously identified in the original study on the corresponding dataset. The thresholds to select TC are given in Table B.1. Arrowed solid line shows the main transition trajectory. (H) Dimensionality reduction of the data coloring for known epithelial genes and mesenchymal genes. Every dot represents a single cell and its shape represents the cell states previously identified in the original study on the corresponding dataset. The color scale represents the normalized expression of the respective genes. (I) Dimensionality reduction of the data coloring for cell-cycle phase based on computed cell cycle scores. Every dot represents a single cell and its shape represents the cell states previously identified in the original study on the corresponding dataset. (J) Percentage of TC associated with each state relative to the total number of TC. (K) Expression levels of top transition genes with cells ordered along the most probable transition trajectory. Solid lines, smoothed expression curves for each gene in the transition trajectory. (L) Heat map of normalized expression of top 20 marker genes and top 20 transition genes. Columns represent cells ordered along the transition trajectory 2 and rows represent genes. Coloring represents the normalized expression of each gene. Transition genes are marked in the box. Top: CPI values of each cell along the transition trajectory.

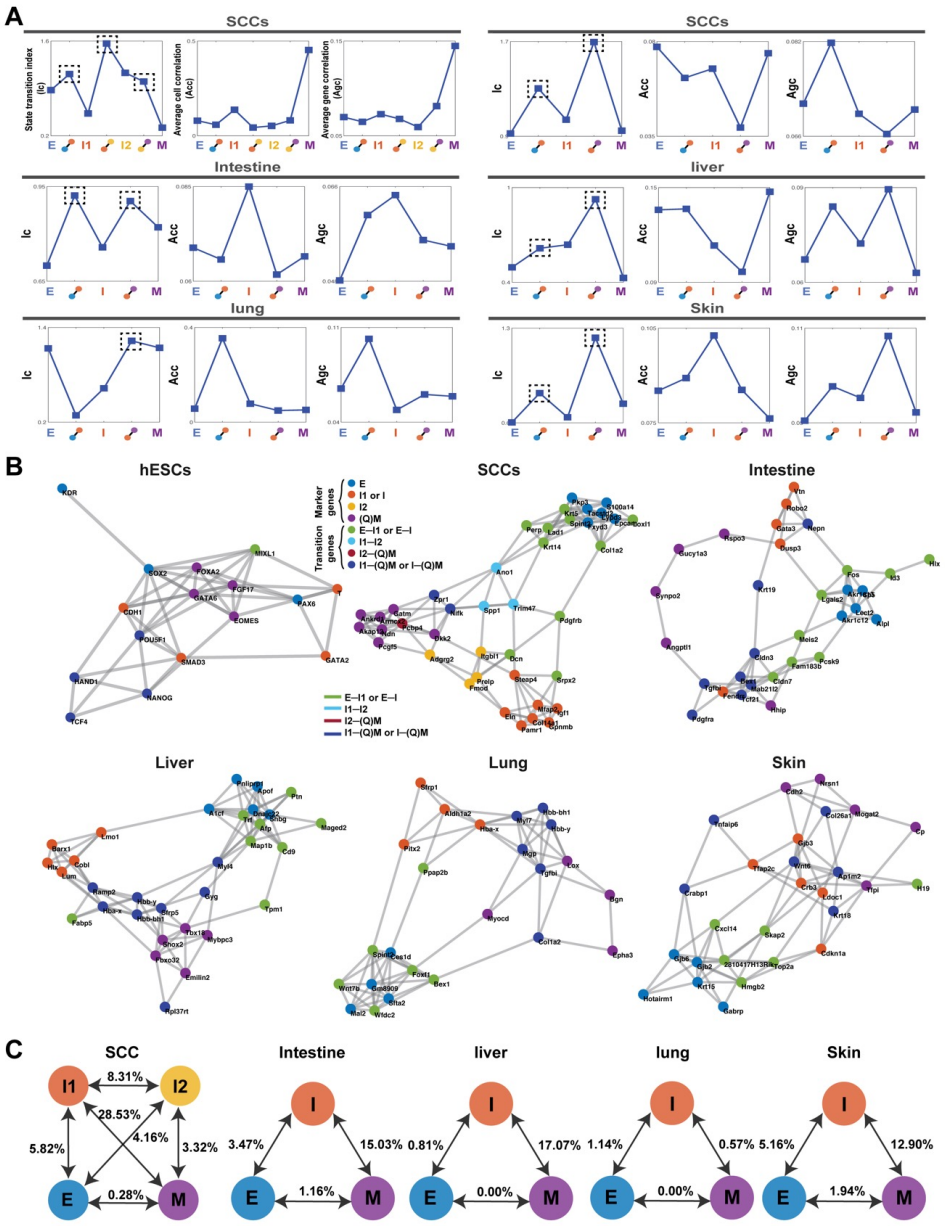

Figure B.9: State transition index and gene regulatory networks for the EMT datasets and their comparisons with QuanTC outputs. (A) State transition index of relatively stable cells in each state and the TC between states. Dashed box: TC with high value of state transition index. (B) Gene regulatory networks of top marker genes and transition genes using the PIDC algorithm from the datasets (the top $\sim 80\%$ of edges are shown). The thresholds to select top genes and edges are given in Table B.1. Each dot represents a gene colored by its type. Graph edges indicate the top interactions and the length of the edge is inversely proportional to the interaction strength between genes. (C) EMT cell lineage inferred from datasets, with node colors consistent with previous figures. The arrow represents potential transition between states, and number represents the percentage of TC among total number of cells.

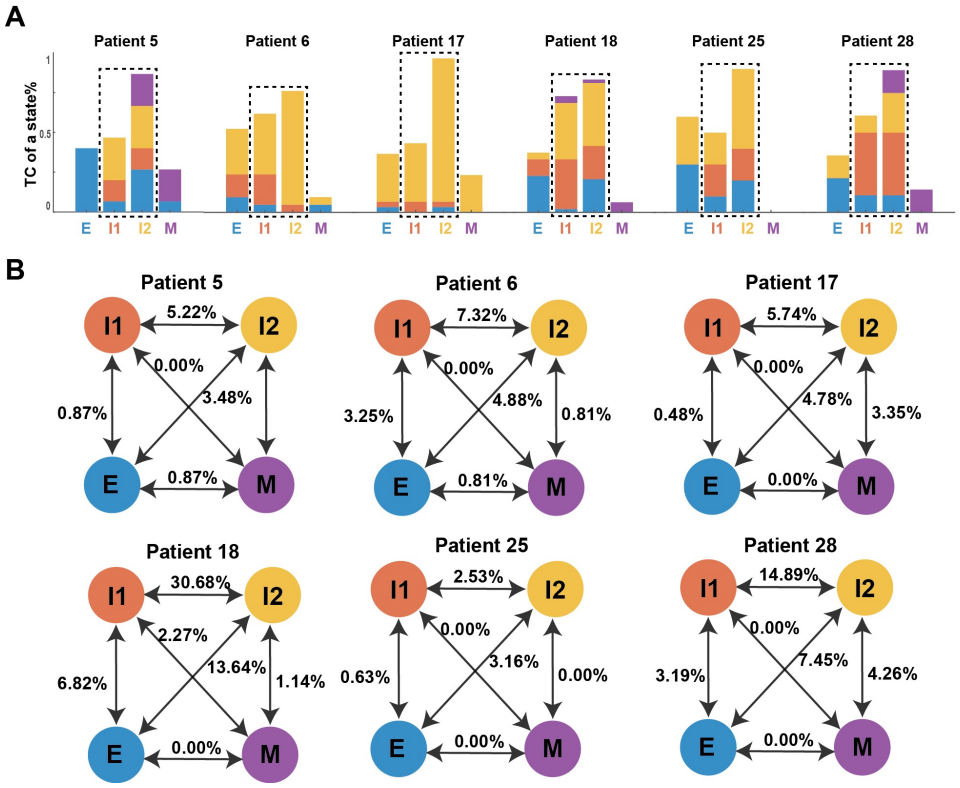

Figure B.10: (A) Percentage of TC in each state relative to the total number of TC from the six patients in HNSCC dataset. (B) EMT cell lineage inferred from datasets, with node colors consistent with previous figures. The arrow represents potential transition between states, and number represents the percentage of TC among total number of cells.

Appendix C

Additional file for Chapter 4

	OVCA420_TGFB1	OVCA420_EGF	OVCA420_TNF
Thresholds of CPI to select TC	0.4	0.55	0.4

Table C.1: Thresholds of CPI values

		E	I1	I2	M
OVCA420_TGFB1	In-strength	0.1776162	1.9648623	1.2557214	3.4196665
	Out-strength	1.689800	1.993284	1.283902	1.850880
	In-closeness	22.565594	2.157533	2.940819	1.204396
	Out-closeness	1.823565	2.114316	2.861763	3.253483
	Pagerank	0.05957221	0.28235406	0.19392289	0.46415084
OVCA420_EGF	In-strength	0.5808987	0.6194308	3.0256635	3.3411261
	Out-strength	2.019406	1.465237	2.266148	1.816328
	In-closeness	7.040529	6.006635	1.415538	1.180988
	Out-closeness	1.608914	2.230063	2.206174	2.954360
	Pagerank	0.1027799	0.1071867	0.3774523	0.4125810
OVCA420_TNF	In-strength	1.1865127	0.3591376	1.1528117	3.6014175
	Out-strength	1.692995	1.192109	1.666992	1.747783
	In-closeness	3.458899	10.294901	3.538624	1.153217
	Out-closeness	2.183708	2.668141	2.202695	4.011858
	Pagerank	0.19736113	0.08600868	0.19305720	0.52357299
SCC	In-strength	1.506968	3.005718	2.501364	2.870468
	Out-strength	1.969614	2.411808	3.287782	2.215314
	In-closeness	2.488402	1.320666	1.796591	1.347659
	Out-closeness	1.798303	1.788193	1.221375	1.909698
	Pagerank	0.1670528	0.2960197	0.2525167	0.2844108

Table C.2: Measuring node centrality

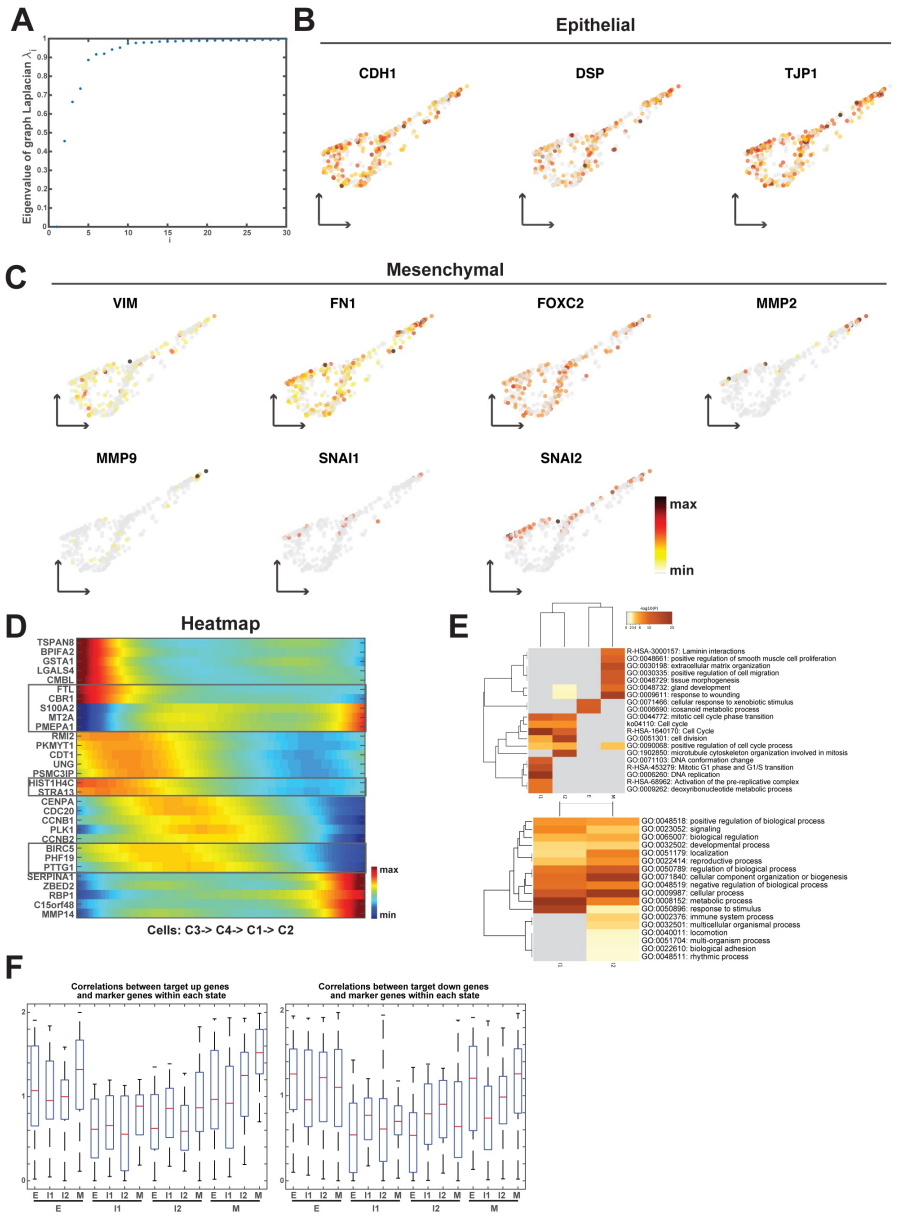

Figure C.1: OVCA420 cancer cell line undergoing EMT induced by TGFB1. (A) The first 30 sorted eigenvalues of the graph Laplacian of the cell-cell similarity matrix from consensus clustering. (B-C) Dimensionality reduction of the dataset by QuanTC coloring for known epithelial genes (B) and mesenchymal genes (C). (D) Heatmap of normalized expression of marker genes and transition genes. Columns represent cells ordered along the transition trajectory and rows represent genes. Coloring represents the normalized expression value of each gene. (E) The top-level gene ontology biological processes analyzed by Metascape of the marker genes of all cell states and ICS respectively. (F) Boxplot of the correlations between target genes and marker genes from Fig. 4D within each state. The central red mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points.

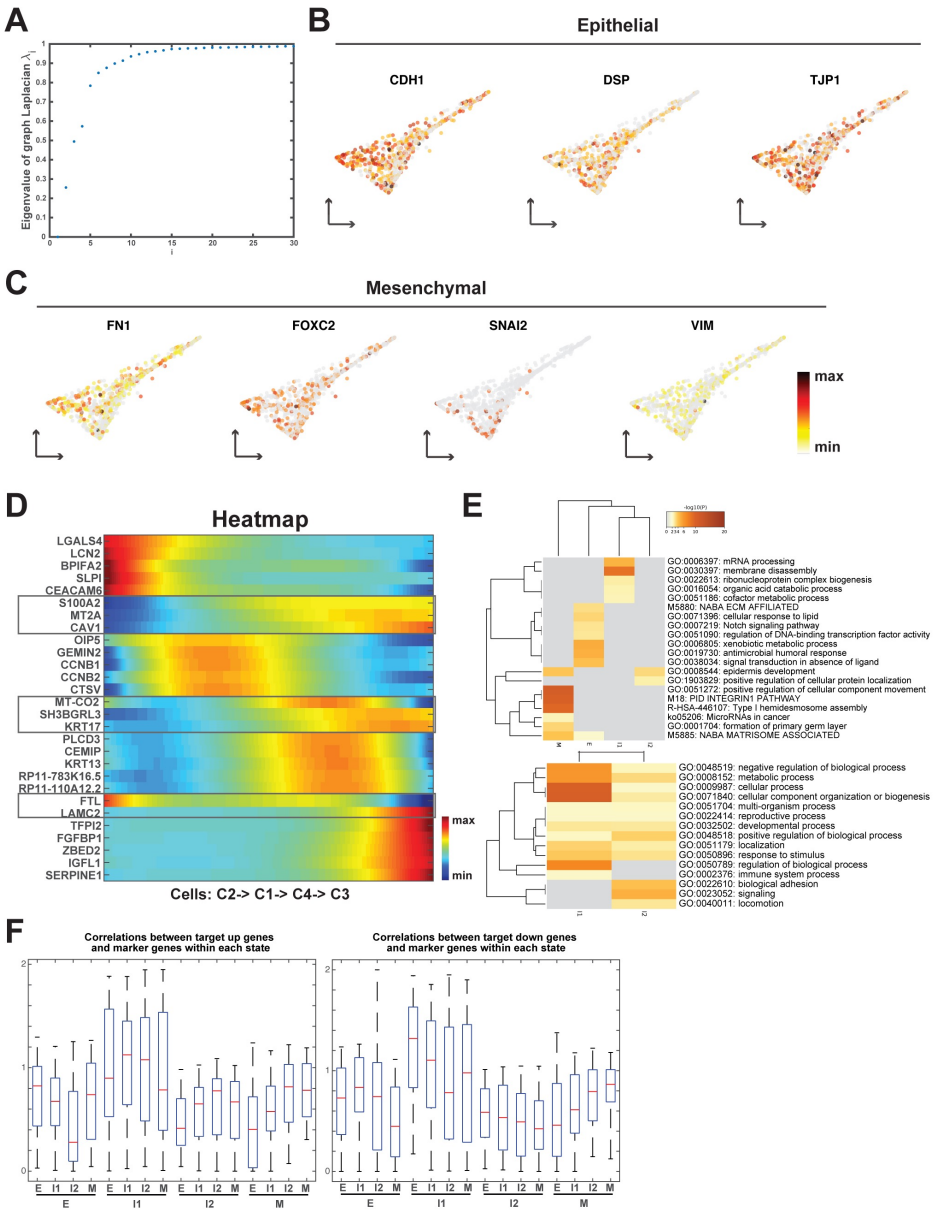

Figure C.2: OVCA420 cancer cell line undergoing EMT induced by EGF. (A) The first 30 sorted eigenvalues of the graph Laplacian of the cell-cell similarity matrix from consensus clustering. (B-C) Dimensionality reduction of the dataset by QuanTC coloring for known epithelial genes (B) and mesenchymal genes (C). (D) Heatmap of normalized expression of marker genes and transition genes. Columns represent cells ordered along the transition trajectory and rows represent genes. Coloring represents the normalized expression value of each gene. (E) The top-level gene ontology biological processes analyzed by Metascape of the marker genes of all cell states and ICS respectively. (F) Boxplot of the correlations between target genes and marker genes from Fig. 4D within each state. The central red mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points.

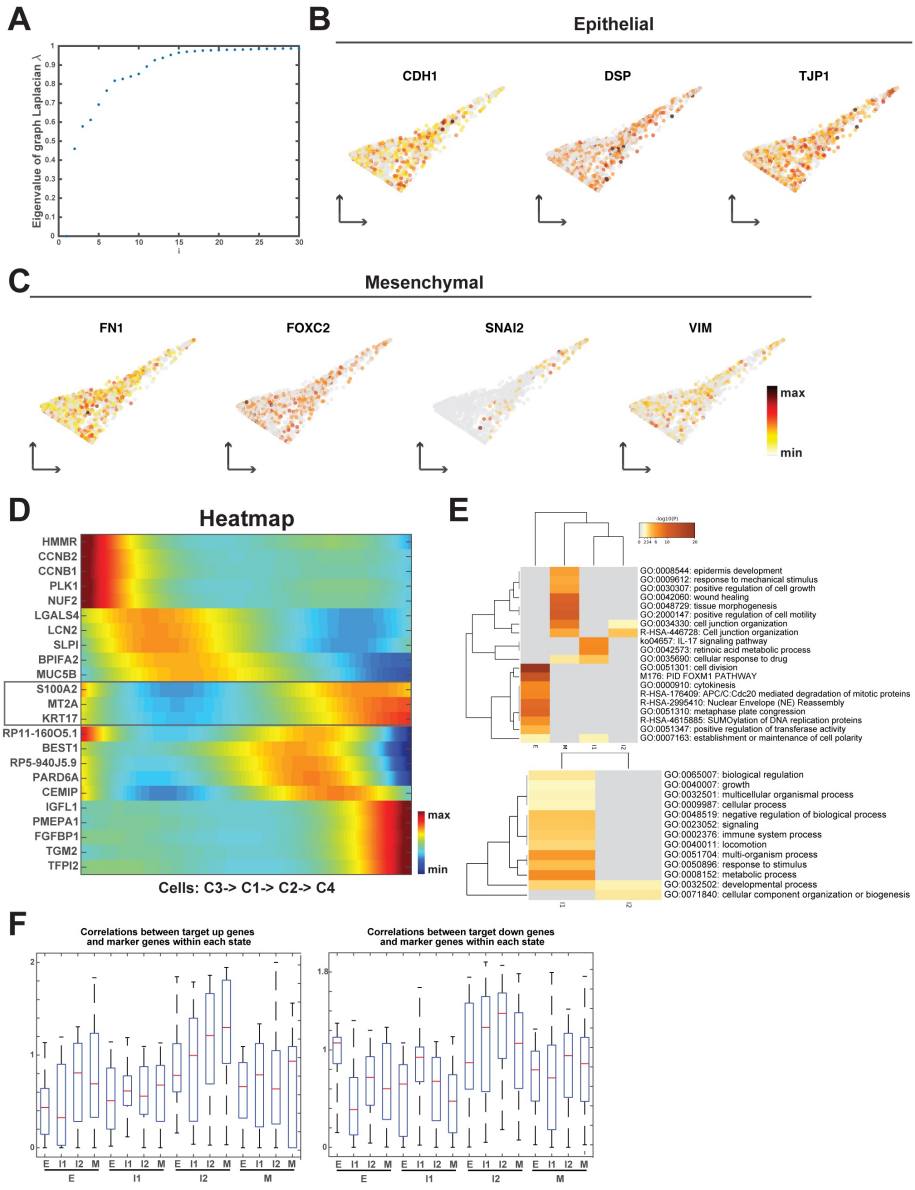

Figure C.3: OVCA420 cancer cell line undergoing EMT induced by TNF. (A) The first 30 sorted eigenvalues of the graph Laplacian of the cell-cell similarity matrix from consensus clustering. (B-C) Dimensionality reduction of the dataset by QuanTC coloring for known epithelial genes (B) and mesenchymal genes (C). (D) Heatmap of normalized expression of marker genes and transition genes. Columns represent cells ordered along the transition trajectory and rows represent genes. Coloring represents the normalized expression value of each gene. (E) The top-level gene ontology biological processes analyzed by Metascape of the marker genes of all cell states and ICS respectively. (F) Boxplot of the correlations between target genes and marker genes from Fig. 4D within each state. The central red mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points.

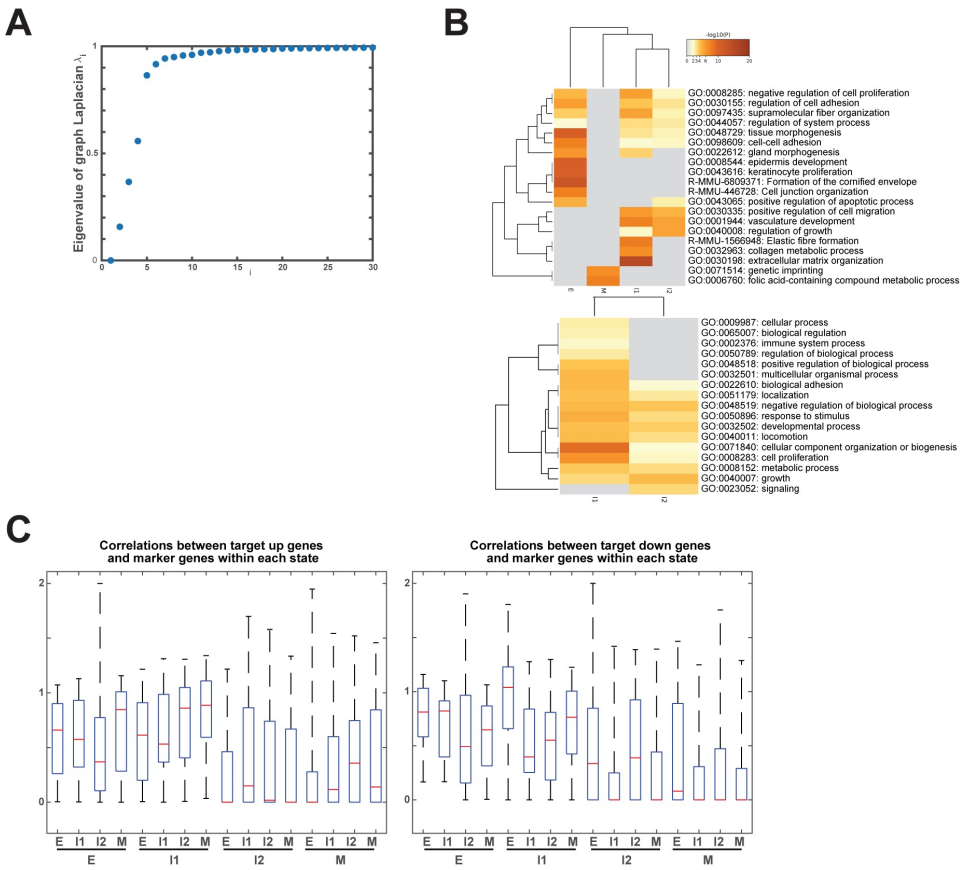

Figure C.4: (A) The first 30 sorted eigenvalues of the graph Laplacian of the cell-cell similarity matrix from consensus clustering. (B) The top-level gene ontology biological processes analyzed by Metascape of the marker genes of all cell states and ICS respectively. (C) Boxplot of the correlations between target genes and marker genes from Fig. 4D within each state. The central red mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points.