QUANTITATIVE STUDY ON REGULATORY MECHANISMS OF CELL PHENOTYPE TRANSITION

by

Jingyu Zhang

B.S., Shandong University, 2006

M.S., Ocean University of China, 2010

M.S., Virginia Polytechnology and State University, 2013

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment of

the requirements for the degree of

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

By

Jingyu Zhang

It was defended on

June 6, 2018

and approved by

Ipsita Banerjee, PhD, Associate Professor, Department of Bioengineering

Robin E.C. Lee, PhD, Assistant Professor, Department of Computational and Systems Biology

Jian Ma, PhD, Associate Professor, CMU Department of Computational Biology

Dissertation Director: Jianhua Xing, PhD, Associate Professor, Department of Computational and Systems Biology

QUANTITATIVE STUDY ON REGULATORY MECHANISMS OF CELL PHENOTYPE TRANSITION

Jingyu Zhang, PhD

University of Pittsburgh, 2018

Cells in a multicellular organism share the same set of genome and can assume multiple phenotypes. Uncovering mechanisms of regulating cell phenotype changes has become an important and active research area. This dissertation presents a collection of my combined computational and experimental efforts on studying cell phenotype transitions.

Chapter I gives a literature overview on cell phenotype conversion and regulation. One can identify four generic modules that function coordinately to regulate cell phenotypes. The whole system forms a highly interconnected network and involves a large number of molecular species for epigenetic, transcriptional and translational regulations.

Chapter II addresses how a cell interprets temporal and strength information of signals and makes cell fate decision. I performed an integrated quantitative and computational analysis on how extracellular TGF-β signal is transmitted intracellularly to activate SNAIL1 expression. I demonstrated how quantitative information of TGF-β is distributed through upstream divergent pathways then crosstalk at various places and converge on to SNAIL1. This crosstalk network interprets the duration of TGF-β signal and is robust against stochastic fluctuations.

Chapter III and IV focus on co-regulation of multiple genes that orchestrate cell functionand phenotype changes. In eukaryotic cells, the expression level of a gene is determined by both transcription factors and the local environment, such as histone modifications and three-dimensional chromosome structure. I used EMT in human cell line and neural cell differentiation process in mouse as model systems, respectively. Through performing combined analysis of data of gene expression, epigenetic modification and chromosome conformation, I examined how the local environment and transcriptional factor regulation is coupled. I discovered that genes co-regulated by a common transcription factor (TF) has tendency to be close both sequentially and spatially. The local spatial organization bridged by TFs is cell type specific. Reorganization of DNA local conformation has impact on gene co-regulation during cell phenotype transition.

The final chapter gives a brief summary of the conclusion from my computational and experimental researches in this dissertation. In this chapter, I also introduced the future work in investigating the relationship between chromosome conformation and gene regulation during cell type transition.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to first thank my advisor, Dr. Jianhua Xing. Without his support and guidance for all these years, this dissertation would not have been possible. I would like to extend my sincere gratitude to my committee members: Dr. Robin Lee, Dr. Jian Ma, and Dr. Ipsita Banerjee, for their suggestions and encouragements. I am grateful to Dr. Fan Bai and members of his team, Dr. Ruoyan Li and Ms. Hengyu Chen, in Peking University for all their support.

My research was supported by the NIGMS-DMS joint mathematical biology program (DMS-1545771 and DMS-1462049), and PA Department of Health (SAP 4100062224). I would like to acknowledge the NIH supported microscopy resources in the Center for Biologic Imaging at University of Pittsburgh.

I am indebted to my respected colleagues: Dr. Xiao-jun Tian, Dr. Yi-jiun Chen, Dr. Weikang Wang, and Mr. David Taft. Their relentless helped my projects go smoothly.

I would like to express my gratitude to my cousin, my friends, and my roommates for helping to keep my inner peace.

Finally, most of all, I would like to thank my family, especially my parents and my husband.

1.0 INTRODUCTION

1.1 CELL PHENOTYPE TRANSITION PLAYS KEY ROLE IN LIFE OF MULTICELLULAR ORGANISMS

Multicellular organisms contain cell types that share the same set of genome but are phenotypically different. Start from the beginning of the multicellular organisms, transitions among different cell types are essential in organism development, growth, life cycle, senescence and death. The regulation mechanisms of cell phenotype transition have been an active research topic for long time[1,2], and abundant information has been accumulated. Basically, a cell type transition is under precise control in magnitude, both temporally and spatially[3-5]. Signals that trigger cell type transition can be various, including cytokines[6], chemicals, temperature[7], or other stresses[8,9]. After receiving signals, cells switch on a first wave of responses to interpret the types, duration and strength of signals[10,11]. These responses can be turned on within seconds and prime cells for further reactions. In the next several steps, the information from stimulating signals is relayed downstream through a signal transduction network and alternate multiple aspects of cells, such as primary and secondary metabolisms, cell cycles, cell shapes and movements. Eventually, the changes are further reinforced and the new phenotype is fixed. Fig. 1 schematically summarizes the overall cellular process of sensing, relaying, and responding to stimulating signals.

Cell phenotype transition is in general not a one-zero process, but contains multiple intermediate states. These intermediate states can exist stably, until additional cue is received for further change. For example, in nervous system development, cyclopamine induces embryonic stem cells (ESCs) to neural progenitor cells (NPC), a stable cell phenotype. Further signals can induce NPC cells to fully differentiated cortical neural cells[12]. Epithelial to mesenchymal transition (EMT), a process that we will discuss in detail later, is another example that involves intermediate states. Our previous computational and experimental studies revealed how an EMT regulatory network leads to different states[13]. Recently the potential roles of intermediate EMT states on cancer metastasis and fibrosis development have received great attention[14].

Events of a cell phenotype transition processes span a broad range of time scales and involve a large number of molecular species. For example, at the early signal receiving and transduction stage, post-translational modifications can take place in seconds. While approaching the final stage of cell type transition, it normally takes days or longer for establishing epigenetic modification and chromosome structural reorganization. The complexity of a cell phenotype transition process requires combinations of traditional molecular cell biology and quantitative systems biology approaches, which is the focus of studies presented in this dissertation.

In the remaining part of Introduction, I will first discuss the signal crosstalk in TGF-β induced EMT. A key part of cell phenotype transitions is switch of transcriptional activities of groups of genes. Therefore in the second part I will provide general discussions on layered regulation of gene expressions.

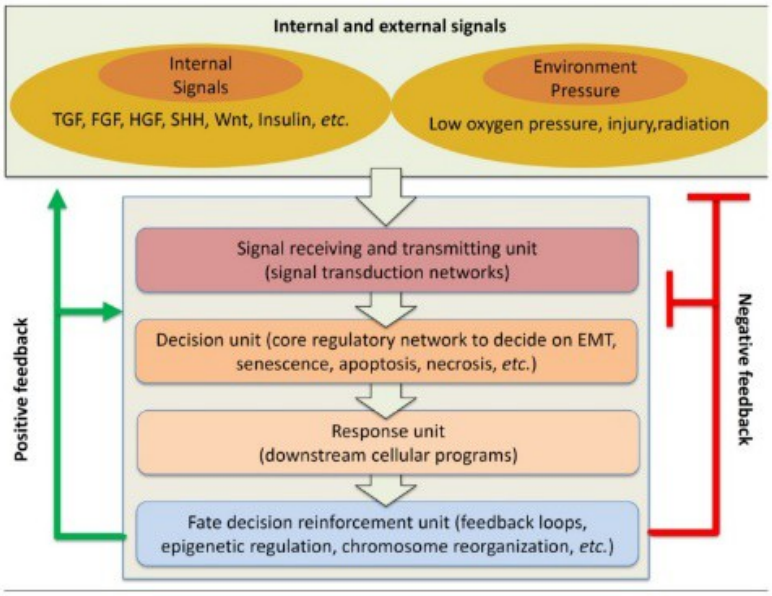

Figure 1 Schematic of the overall signal reception, transduction, and response process of a generic cell phenotype transition process. Intracellular and extracellular signals are relayed via four basic units and regulated by positive and negative feedback loops.

1.2 SIGNAL TRANSDUCTION PATHWAYS OF EMT INDUCED BY TGF-B, SHH, AND WNT AND THEIR CROSSTALKS

A well-documented example of cell phenotype transition under different signal regulation is epithelial to mesenchymal transition (EMT). Epithelial cells are differentiated cells characterized by uniform cell shape, apical-basal polarity, strong cell-cell adherent junctions and cell-matrix hemidesmosomes, and limited mobility. Based on these characteristics, epithelial cells normally form single-cell-layered tubes or sheets to cover body, organs, or compose glands [15,16]. On the contrary, mesenchymal cells have front-back polarity and loose cell attachment. They typically have much higher mobility, which is closely related to their regeneration function[17,18]. A majorfunction of mesenchymal cells is to compose and give rise to other types of cells, such as cells in lymphatic, circulatory, or some connective tissues[19,20].

In 1968, Hay noticed that during chicken embryotic development epithelial cells undergo differentiation and dedifferentiation several times, as well as migrate a relatively long distance within the body. All of these processes require inter-conversion between epithelial and mesenchymal cell phenotypes, called epithelial-to-mesenchymal transition (EMT) and its reverse process, mesenchymal-to-epithelial transition (MET)[21]. Subsequent studies showed that EMT and MET are fundamental in amniotes' gastrulation and neural crest formation, and generation of body patterns. Specifically, during mammalian embryotic development, several rounds of EMT companied by MET take place, which are generally assigned as primary, secondary and tertiary EMT based on the developmental stages[22]. The primary EMT occurs in early embryonic development, such as parietal endoderm formation, mesoderm formation, and neural crest delamination. Signal molecules that initialize and regulate the primary EMT include the transforming growth factor-β (TGF-β) superfamily (e.g. BMP, Nodal, *et. al.*), the WNT

family[23,24] and the fibroblast growth factor (FGF) family[25]. Two members of the SNAIL family, SNAIL1 and SNAIL2, are important during human early embryotic development for repressing E-cadherin and weakening cell-cell junction[26]. Mesodermal cells generated from the primary EMT later undergo MET and form the secondary epithelial structures, such as the notochord and the somites. The transit epithelial structures go through the secondary EMT and give rise to more differentiated structures, such as endocardial progenitors and connective tissues. Some of the epithelial or mesenchymal structures from the secondary EMT-MET process complete their final differentiation and maintain their phenotypes, provided there is no additional abnormal inducing signal. Others continue on a tertiary EMT (and MET), again controlled by multiple signaling pathways, and eventually give rise to complex organs, such as lung[27] and heart[28].

In wound healing, the initial input signals are from injury. Transcription factors such as ERK, SLUG, SNAIL[29-31] are activated to promote conversion of epithelial cells to a partial EMT state. Partial EMT cells also have loose cell-to-cell connection, which allow them to migrate to wounding site. This process is reversible, so the partial EMT cells return back to the epithelial phenotype after the injury site has been healed and the EMT triggering signal has been withdrawn.

Studies also support that EMT and MET take place during metastasis. Some hypothesis regards cancer as an over healing wound[32] or an abnormal development process[33], while there are features specific for cancer progression. Invasion and metastasis are decisive steps in cancer progression and the major cause for cancer-related mortality[34]. At cellular level, EMT or partial EMT leaves cells loosely connected to others, and enables them to depart from the primary location and migrate along the circulatory system to a secondary location, where the migratory cells go through MET to epithelial cells again, proliferate and form a secondary tumor[35].

Though EMT in breast cancer was first observed in 1890s, it did not attract attention in the carcinoma biology community along almost the whole past century. Only during the past decades many crucial signaling pathways and core regulatory elements that induce or contribute to EMT have been uncovered[22,36]. Like in embryotic development, metastasis in cancer progression can also be induced by various signaling molecules or cytokines[37], such as proteins in the TGF-β superfamily[38], hedgehog (HH) family, WNT family[39], and interleukin (IL) family[40], *etc.*. Stressful microenvironments, such as hypoxia[41] or free radicals[42], also trigger EMT[43].

1.2.1 Multiple signal transduction pathway and their crosstalks

A plethora of stimuli activates multiple signal transduction pathways, which then converge to a core regulatory network composed by transcription factors (such as SNAIL1/2, ZEB1/2, TWIST) and miRNAs (such as miR34 and miR200 families)[44]. The latter further interact with other regulatory elements to instruct a cell to choose of the several possible cell fates. For example, SNAIL1 can bind to P53, a major regulator protein that induces senescence or apoptosis, and trap the free P53 in cytosol. Thus, SNAIL1 inhibits the choice of senescence or apoptosis[45]. The above flow-of-information is not unidirectional, but at every stage there exist negative and positive feedbacks to previous stages, and form a closed network.

The TGF-β, sonic hedgehog (SHH) and WNT pathways are three well-studied signal transduction pathways that can induce EMT. Below we will give a review on how these three pathways participate in signal processing during EMT, and how the knowledge has been applied on carcinoma treatment. Especially we hope to provide a perspective on the signaling network based on systems biology approaches, and insights on biomedical interventions of EMT.

1.2.1.1 TGF-β pathway TGF- β is a type of secretive protein that affects both the cell secreting the protein (autocrine) and its neighboring cells (paracrine). The proteins are secreted as inactive precursors, and are cleaved proteolytically by the latency-associated peptide (LAP) before they can combine with the receptors on the cell surface[46]. In human cancer cells, three TGF-β isoforms (TGF-β1, TGF-β2 and TGF-β3) have been discovered, which share over 70% homological sequence. There are two types of transmembrane TGF-β receptors (TGFBR). Generally, the type II TGFBR (TGFBR-II) molecules recognize and bind to TGF-β. Next, the complex recruits other TGFBRs and forms a complex with a stoichiometry of two molecules of TGFBR-I, two molecules of TGFBR-II and one molecules of TGF-β. Formation of the complex activates the phosphorylation function of the intracellular part of TGFBRs to relay the TGF-β signals downstream of the pathway[47].

Paradoxically, TGF-β functions as both tumor repressor and oncogene. More specifically, in normal cells or even some pre-cancer cells, TGF-β promotes proliferation arrest and thus represses tumor growth. However, in advanced malignant carcinoma cells, TGF-β promotes EMT and tumor metastasis. These seemingly contradictory functions come from the two parallel TGF-β pathways[48,49].

SMAD-dependent TGF-β pathway: The more canonical TGF-β pathway depends on activationand deactivation of a SMAD family. In this pathway, signaling is cooperatively regulated by three types of SMADs in this family. The receptor-regulated SMADs (R-SMADs), such as SMAD2 and SMAD3, receive signals transmitted from the membrane signal receptors, and are phosphorylated[50]. Then two molecules of phosphorylated R-SMADs come together and recruit a common-mediator SMAD (co-SMAD) to form a trimmer, which can be transported from cytosol into nucleus[51-53]. In nucleus, the complex binds to DNA-sequence-specific

transcription factors, individually or with other co-activators, to active transcription of target genes[54], such as snail1[55,56], snail2[57,58], and other oncogenes. The inhibitory SMADs (I-SMAD), such as SMAD6 and SMAD7[59], have opposite function from R-SMADs and co-SMAD. They negatively regulate the activity of R-SMADs and co-SMADs through interfering phosphorylation of R-SMAD[60], or competing with them[61], thus compose a negative feedback loop at the early step of TGF-β induced SMAD-dependent pathway[62,63].

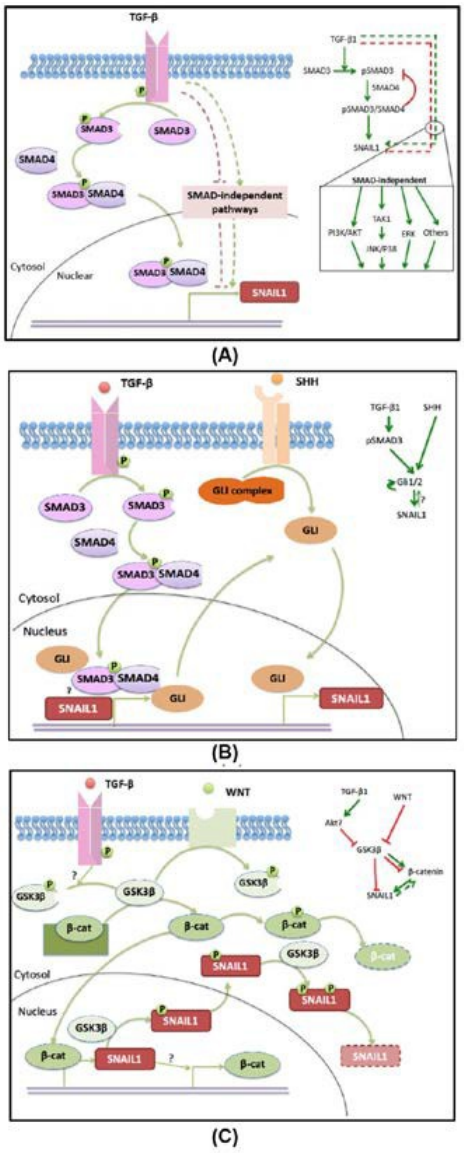

Figure 2 Crosstalk among the TGF-β, SHH, and WNT signaling pathways converging to the core regulation unit. **(A)** TGF-β pathway (proteins in pink). **(B)** SHH pathway (proteins in orange). **(C)** WNT pathway **(**proteins in green) In the inserted regulation networks, point arrows represent activation, blunt ones represent inhibition, and dashed lines represent indirect links.

TGF-β-induced SMAD activation has been widely considered as a tumor promotion event, especially in highly malignant cancer cells. On the other hand, many reports documented that TGF-β-induced activation of SMADs also suppresses tumor formation and development by blocking cell cycle and arresting cell growth[64] in breast cancer cell lines[65], liver cancer cells[66], and normal epithelial cells[67]. Mutations on smad2 and smad4 have been reported in colorectal cancer[68], pancreatic ductal adenocarcinoma[69], or hepatocellular cancer[70]. These mutations implicate the potential antitumor function of SMADs[71,72]. Moreover, depending on cell types, TGF-β induced SMAD signals can also induce apoptosis as a safety mechanism to prevent transformed cell from EMT or metastasis[73,74].

The diverse roles of SMAD activation come from how SMAD2/3 and SMAD4 heterotrimeric complex perform its function. Some of the targeted genes can be activated by SMAD trimer only at a low basal level. More efficient activation of these genes requires binding of the trimer to other sequencing-specific activators[75]. For example, the SMAD2/3/4 complex recruits CBP/P300 as the co-activator to activate p15[76] or p21[77], which inhibit cells from progression of cell cycles[78]. On the other hand, if the heterotrimer recruits EMT promoting transcription factors as the co-activator, such as TWISTs, it can up-regulate oncogenes expression[79]. This cofactor-dependent gene expression pattern explains why TGF-β functions differently in cells of different type and cell stage, and on the existence of other stimuli.

SMAD independent pathway: Other than the SMAD-dependent pathway, TGF-β receptors also relay the signals through a group of additional signal proteins, such as PI3K/AKT, MEK/ERK1/2[80], RHO-A and JNK/P38[81]. The SMAD-independent pathways are more complicated than the SMAD-dependent pathway, since crosstalks among the signaling proteins form a more intricate molecule-molecule interaction network. For instance, TGF-β can induce

AKT phosphorylation and activate phosphatidylinositol 3-kinase (PI3K) rapidly, which possibly contributes to SMAD2-induced EMT[82]. Studies on keratinocyte cells show that the PI3K/AKT pathway helps to complete the TGF-β induced SMAD-dependent EMT[83]. Furthermore, PI3K/AKT antagonize TGF-β-induced apoptosis and growth arrest[84,85], and bias TGF-β treated cells to undergo EMT. Similarly, the TGF-β induced ERK/MAPK pathway contributes to EMT induction, since ERK is required for removing cell adheren junctions to increase cell mobility. TGF-β also activates RHO-like GTPases in the RHO pathway, which have multiple functions in cytoskeletal organization, apoptosis, and EMT[86], as well as the RHO-A-dependent signaling pathway, to promote mesenchymal characteristics in epithelial cells through inducing stress fiber formation[81,87]. Another TGF-β induced SMAD-independent pathway is the JNK/P38 pathway. JNK itself can phosphorylate R-SMAD directly, thus turns on the EMT program. Both JNK and P38 can synergistically work with SMADs to promote TGF-β induced apoptosis[81,88].

Put all of the above together, at pro-oncogenic stage, TGF-β has opposing effects on tumor development. SMAD-independent pathways mostly inhibit tumor suppressors[89], while the SMAD-dependent pathway promotes cell cycle arrest or apoptosis to stop tumor growth at the pre-tumor stage. In malignant cells, when the amount of cofactors of oncogenes is much higher than that for apoptosis or growth arrest, TGF-β functions more as an EMT-inducer through both SMAD-dependent and SMAD-independent pathways. Therefore in cancer progression, the relative balance between SMAD-dependent and -independent signal transduction likely plays a critical role on determining cell fates[81].

Drugs targeted to TGF-β pathway: Based on its cancer-promotion function, TGF-β is a potential drug target for clinic therapy. Moreover, giving the fact that TGF-β pathway involves in DNA damage repairment[90], inhibition of TGF-β signals may enhance the efficacy of

radiotherapy and chemotherapeutic drugs[91]. Popular clinical treatments that are targeted to blocking TGF-β signals include trapping TGF-β ligands, blocking the receptor kinases signaling, using antisense oligunucleotides to decrease the translation of TGF-β protein, and using peptide aptamers to block the transduction pathway[92]. For example, to trap TGF-β molecules from binding to TGFBRs, 1D11[93] and GC-1008 (Fresolimumab) have been generated as TGF-β neutralization monoclonal antibodies and used in treatment of melanoma[94] and glioblastoma[95]. AP12009 (Trabedersen) is designed to inhibit TGF-β2 expression and has been used to treat pancreatic cancer[96]. Small chemical molecules are developed to block TGF-β signaling by inhibiting the phosphorylation function of TGFBRs, among them SB431542 is widely used to inhibit TGFBR-I in breast cancer therapy[97,98]. Development of peptide aptamer drugs is a new direction that is still at its early stage. Some peptide aptamers target R-SMADs or co-SMAD has been discovered and tested in cell lines[99]. Since they target specific proteins, using peptide aptamers might be a better clinic therapy strategy considering the dual roles of TGF-β. By blocking only one sub-pathway under TGF-β signals, one can possibly suppress the tumor-promotion function of TGF-β without removing its anti-tumor benefits.

Therapeutic blockage of TGF-β signaling is tricky due to the pleiotropic effects of TGF-β on tumor progress. As a secretive protein, controlling the TGF-β signaling microenvironmen near carcinoma is as important as controlling the intracellular TGF-β signaling pathway. Moreover, TGF-β signal has important regulatory functions on normal cell physiology. Blocking of TGF-β pathway completely is detrimental to normal cells and thus not recommended. Therefore, while significant progress has been made on developing drugs that target the TGF-β signaling pathways, clinically these drugs should be used with caution and perhaps only in certain cancer types.

1.2.1.2 SHH pathway engages in EMT and crosstalks to TGF-β In addition to the TGF-β pathway, the HH pathway has been reported to induce EMT in lymphatic and gastric tumors[100], pancreatic cancer[101], breast cancer[102], *etc.*, individually or cooperated with other pathways.

Like the TGF-β pathway, the HH pathway starts from the secretive hedgehog proteins, a family of glycoproteins. The precursors of HH proteins undergo many steps of post-translational modification and cleavage before maturation, which are secreted as oligomers or soluble multimers, and can diffuse over various distances between tissues in body before being removed[103]. Three mammalian HH proteins have been identified in the HH family recently, sonic hedgehog (SHH), Indian hedgehog (IHH) and desert hedgehog (DHH). While these three HH proteins share some redundant functions, each of them also has evolutionally specified roles. For example, sonic hedgehog (SHH), the most common and understood one, is crucial in embryo development, cancer progression, and body patterning[104,105].

The off and on states of SHH pathway: The canonical SHH pathway can be generally divided into three parts, the signal reception elements, the signal transmission elements, and downstream transcription factors. In the absence of SHH, the pathway is at an 'off' state. HH-Patched protein (PTC), which is the transmembrane receptor element, binds to smoothened protein (SMO) to inhibit SMO activation. In the cytosol, activated protein kinase A (PKA) binds to other kinases including glycogen synthase kinase 3β (GSK3β) and other factors to phosphorate glioma- associated oncogene homologs (GLIs). In the absence of HH signals, GLIs have only low basal level expression, and the proteins assume a repressor form. That is, GLIs repress the expression of their target genes. When SHH is present, the pathway switches to an

"on" state. SMO proteins are released and phosphorylated to promote the activation of GLIs[106-108]. Some GLI proteins (e.g. GLI2) are truncated at the carboxy-terminal by the proteasome and turn to the activator form[109,110], which then activate the expression of their target genes.

Regulation of GLI proteins and crosstalk to the TGF-β pathway: GLI proteins are the major transcription factors in the SHH pathway. A high level of GLI proteins indicates activation of the SHH pathway[111]. Three GLI proteins have been identified in mammals, GLI1, GLI2 and GLI3. Though they share long homologue sequence and also similar DNA-binding sequence, they play quite different roles in development, EMT, and cancer promotion. Besides the DNA binding domain, GLI1 has only the activator domain and can be activated by SHH. GLI2 and GLI3 have both the repressor domain and the activator domain. However, in most contexts, SHH activates GLI2[112], while it is unclear SHH activates or represses GLI3[113].

Activated GLI1 and GLI2 can directly promote the expression of a group of genes by physically binding to their promoter region, including oncogenes and genes that are involved in the EMT process[114], such as bmi1[115], nanog[116], snail1[117,118]. Based on the fact that expression of GLI1 can be regulated by the E-box[119], positive feedback loops may exist between GLI1 and its target transcription factors that contain E-box at the promoter region of their genes, such as SNAIL1. Furthermore, GLI proteins can also be up-regulated by SMAD proteins[120,121]. Actually, the TGF-β/SMAD/GLI2 axis has been suggested to be essential for cancer metastasis[122]. Consequently, the SHH pathway and the TGF-β pathway crosstalk to each other and coordinately induce EMT. GLI proteins are also involved in several positive or negative feedback loops within the SHH signaling pathway. For example, the activated form of GLI2 can directly bind to the promoter region of gli1 to up-regulate GLI1 protein expression, while GLI1

can also induce GLI2 expression directly or indirectly, so the two form a positive feedback loop[112,123]. On the other hand, GLI1 induces PTC, and PTC inhibits GLIs to form a negative feedback loop[112].

Clinical observation and interventions of SHH signaling pathway in cancer: There are clinical reports on abnormal activation of the SHH signaling pathway in different types of cancer. For example, in thyroid cancer, SHH is expressed in 64% of PTC tissues but only in 17% of non- cancerous tissues, and GLI1 is expressed in 48% and 9% of these two different tissues, respectively[124]. Activation of the SHH pathway is related to promotion of the EMT process in lung cancer cell line[125], renal cell cancer[126], and gastric cancer[100]. Based on these observations, blocking SHH signaling is a popular strategy in cancer therapy. Indeed inhibition of SHH signaling can reduce the proliferation rate of non-small-cell-lung-cancer cells significantly[127]. Similar phenomenon has also been observed in breast cancer[128].

A basic strategy of intervening the SHH pathway is blocking the SHH receptor or other major players downstream in this pathway[113]. In pancreatic cancer therapy, combination of two SMO inhibitors, gemcitabine and cyclopamine, completely abrogated pancreatic cancer cells metastasis while also significantly reduced the size of primary tumor[129]. Cyclopamine, vismodegib, or other SMO inhibitors have been used widely in clinic for medulloblastoma[130], ovarian cancer[131], and pancreatic cancer[132] treatment. Given the importance of GLI1/2 in the SHH pathway on promoting EMT and metastasis, blockade of GLI1/2 is a candidate for cancer treatment. For example, small chemical molecules, GANT58 and/or GANT61, which block GLI1/2 function, arrest tumor growth in prostate cancer cell[133]. Compared to blocking the

upstream regulators in the SHH pathway, an advantage of targeting GLI proteins is that GLI proteins serve as a signaling hub of multiple pathways that are activated in cancer cells, such as the TGF-β, WNT, and SHH pathways.

1.2.1.3 WNT pathway in cancer progress and EMT The WNT pathway is another signaling pathway that crosstalks to the TGF-β pathway and promotes EMT. Two major converging elements between the WNT and TGF-β pathways are the tumor repressor GSK3β, and the activator β-CATENIN (Fig. 2C).

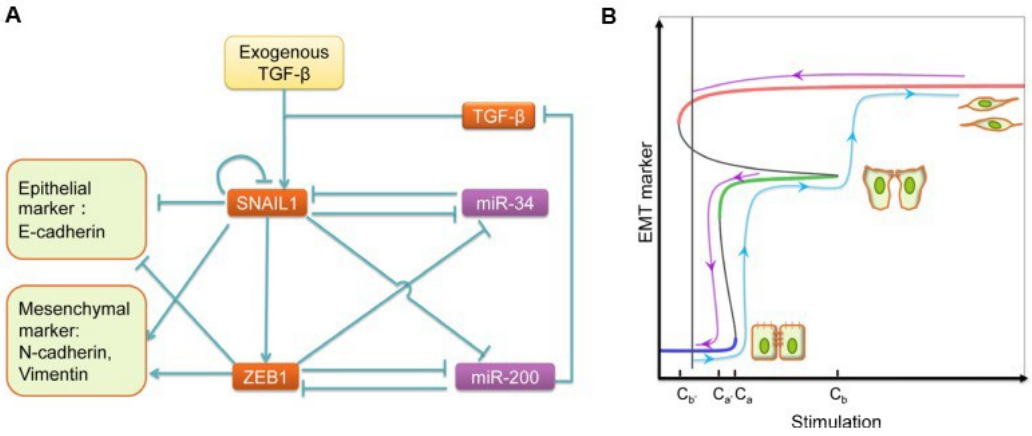

Figure 3 Systems biology studies on core EMT network **(A)** Core regulatory network of TGF-β induced EMT revealed by experimental studies. Point arrows represent activation, and blunt ones represent inhibition; and **(B)** mathematically-predicted bifurcation diagram. Also shown are the corresponding dose-response (D-R) curves that are more familiar to experimentalists. Notice that the D-R curves are different for cells starting from different phenotypes and treated with increasing (blue curve) and decreasing (purple curves) exogenous TGF-β, respectively. This history-dependent hysteresis is a signature of bistable dynamics. The predicted bifurcation diagram has been experimentally confirmed in MCF10A cells. Adapted from[142,144].

WNT is a large signal family, and the WNT proteins are highly conserved from fruit fly to human[134]. In *homo* species, 19 discovered WNTs compose a very intricate network, which is essential in development, stress response, or cancer. Abnormal activation or mutations in the WNT pathway has been reported in many cancers, such as intestinal neoplasms, breast cancer, prostate cancer, and lung cancer[135].

The canonical WNT pathway starts from reception of signal molecules on cell membrane. GSK3β is a major downstream regulator of the receptor. Without WNT signal, GSK3β keeps its active form, which can phosphorylate its target proteins (e.g. β-CATENIN[136]) for further degradation. When the WNT signaling pathway is activated, GSK3β is phosphorylated to an inactive form. Thus, functional β-CATENIN is accumulated in cytosol and further is transported into nucleus. In nucleus, together with TCF/LEF, β-CATENIN binds to the promoter region of target genes, such as snail1, and activates their transcription[23,137]. Furthermore, SNAIL1 can also form a positive feedback with β-CATENIN by interacting with the β-CATENIN physically[138] or increase the amount of free β-CATENIN indirectly through EMT process[139]. Inactivation of GSK3β can also increase SNAIL1 expression directly following two steps: in nucleus, it is phosphorylated by GSK3β; then SNAIL1 can be transported from nucleus to cytosol, where it can be phosphorylated again by GSK3β for final degradation[140].

The WNT pathway affects and is affected by several signaling pathways, including SHH and TGF-β. GSK3β affects GLI proteins both positively and negatively. On one side, GSK3β phosphorylates GLI proteins for degradation[141]. On the other side, GSK3β phosphorylates SUFU, a scarf protein for GLI proteins, and releases free GLI proteins[142]. GSK3β also stabilizes GLI mRNA indirectly, leading to increase of the amount of GLI proteins[143]. Subject to TGF-β treatment, the WNT pathway can be activated by SMAD-independent pathways. For instance, in

human lung fibroblast cells, TGF-β1 can inactivate GSK3β by activating the mitrogen-activated protein kinase (MAPK) pathway and phosphorylating ERKs[144]. GSK3β can be inhibited by the ARK pathway, and the latter can be activated by TGF-β1 in some cell lines. In addition, GSK3β negatively affects the TGF-β pathway by phosphorylating SMAD3 with its cooperator, AXIN, and triggers its ubiquitination and degradation when TGF-β is absent[145].

Clinical observation and interventions of the WNT signaling pathway in cancer: Based on the well documented close relationship between WNT signaling aberrance and cancers, intensive efforts have been put on designing drugs that specific target to the WNT pathway[146]. However, no drug has been approved for clinical usage yet[147]. Special caution has to be taken since the WNT pathway has important functions in almost every aspect of mammalian cells, such as proliferation and regeneration.

Beside target-specific small molecules, a group of widespread-used drugs or compounds have been proven to help in cancer treatment as they also block the WNT pathway. For instance, Aspirin affects and blocks the WNT pathway as a non-steroidal anti-inflammatory drug (NSAID) at many levels, such as facilitates β-CATENIN degradation[136]. Vitamins, such as retinoids, vitamin D, *etc.*, show effects in colorectal cancer and breast cancer probably through interacting with β-CATENIN and TCFs[146].

1.2.2 Systems biology in signaling crosstalk and drug discovery

As we discussed above, the crosstalk network among TGF-β, SHH, and WNT signaling pathways is complex. In addition, some of the signals, such as TGF-β, have opposite roles as both cancer repressor and promoter, depending on cell types and cancer stages. Similarly, some of the regulators can both turn on ‘off’ and ‘on’ their target genes. For example, GLI proteins can

both negatively and positively regulate expression of themselves. Another typical example is GSK3β, which can covalently modify both oncogene proteins (e.g. SNAIL1) and tumor repressors (SUFU) for degradation[142]. Furthermore, all of these signaling pathways have essential roles in both normal cell life cycle and in cancer development.

Due to the above-mentioned molecular biology complexity, a naive drug-design strategy based on simply blocking certain pathway likely has serious side effect to normal cells and patients. Cancer therapy and anti-tumor drug discovery are thus difficult, time consuming, and face three basic challenges.

1. Which of the proteins/regulators to target for clinic and commercial consideration?
2. How to select chemicals that are suitable for therapy from the gigantic data pool?
3. How to design treatments that target to specific population of tumors in patients?

Currently computational and system biology studies become indispensible on revealing the molecular mechanism and addressing the three challenges. These studies engage widely in current molecular biology, and provide systematic and integrative perspective on understanding the biological implications underlying individual experimental results. In the next we will use a series of recent studies of us to illustrate how we can advance our understanding of the EMT regulatory mechanism through combined mathematical modeling and experimental studies.

With increasing reports on new signals and pathways leading to EMT, one might have the impression that EMT can be induced easily. Actually EMT is tightly regulated at multiple levels, and pathological EMT is a rare event in a healthy body. Furthermore, EMT is not a 'to be or not to be' question. Instead, EMT proceeds through a wide spectrum of intermediate states generally referred as the partial EMT state[148,149]. Tian *et al.*[150] mathematically analyzed the core EMT regulatory network (Fig. 3a), and proposed a sequential two-step mechanism, as

summarized in Fig. 3b. A SNAIL1/miR34 double negative feedback loop and a ZEB1/miR200 feedback loop form two binary switches. In epithelial cells, both miR34 and miR200 are highly expressed, while SNAIL1 and ZEB1 express only at basal levels. With an intermediate concentration of TGF-β (exceeding a threshold value C_a), miR34 is degraded and the level of Snail1 increases. Snail 1 further inhibits transcription of miR34[151], partially upregulates epithelial markers such as E-cadherin and downregulate mesenchymal makers such as Vimentin and N-cadherin. At this stage, the ZEB1/miR200 switch has not been reverted, and cells exist in a partial EMT state. Experimentally, if now one reduces the exogenous TGF-β to a lower level (below a threshold value $C_{a'} < C_a$), cells returns to the epithelial state. That is, the Epithelial-to-partial EMT transition is reversible under TGF-β treatment. When the exogenous TGF-β level exceeds a second threshold ($C_b > C_a$), miR200 is degraded and the level of ZEB1/2 increases. ZEB further inhibits transcription of miR200[151], works together with SNAIL1 and other factors to upregulate epithelial markers and downregulate mesenchymal makers, so cells undergo a full EMT. At this stage cells express autocrine TGF-β, which maintains cells in the mesenchymal state even when the exogenous TGF-β is removed. That is, the full EMT is irreversible under TGF-β treatment. Subsequently, Zhang *et al.* performed quantitative experimental studies using the human mammary MCF10A cell line, and confirmed all model predictions. Therefore, the studies reveal how different cell phenotypes emerge out of the interactions among the transcription factors and microRNAs, and provide clues for biomedical intervention to regulate the conversion between them.

The above-mentioned model clearly does not provide a complete picture on EMT but rather should be considered as a starting point. It only considers a small core network without explicitly considering many other key EMT players such as TWIST. Actually, there are many

more positive feedback loops formed by various regulating elements. These feedback loops can also form multiple stable switches, which may function either in synergy or in sequence to give rise to possibly a combinatorial number of partial EMT states, consistent with the notion of a quasi- continuum EMT spectrum. Further experimental studies can also analyze whether the revealed two-step mechanism is general for different cell types and cell lines, and if not (, which is very likely), what are the differences and common themes. While there are many possible directions for expanding the modeling efforts, below we discuss three of them.

First, the model can be systematically expanded to include other involved molecular species. Using a more coarse-grained Boolean network modeling framework, recently Steinway *et al.* studied a more complex network of EMT including crosstalks among the TGF-β, SHH, and WNT pathways, and tested model predictions experimentally[150]. Expansion of an ordinary differential equation based model like ours can more faithfully describe the temporal and steady state dynamics of the system.

Second, the field awaits further methodology developments on incorporating high throughput data into detailed dynamics modeling. The past decade has observed an explosion of accumulation of "omics" data sets, such as transcriptomics, epigenomics, proteomics, and metabolomics. Typically a high throughput data set provides a global view of a system or process under study, although at relatively low resolution both in temporal features and in data quality compared to a more focused study like the ones discussed above. Bioinformatics tools have been widely used to analyze the omics data. For example, Nam, *et al.* used PATHOME, analgorithm based on entire pathway information, to connect the WNT and AMPK pathways at

HNF4a-WNT5A in gastric carcinogenesis, and further predict WNT5A as a suitable therapy target[152]. Given the importance of including dynamics into pharmaceutical development[153,154], the challenge is how to combine these global and focused levels of studies.

Third, we may observe more examples of integrating computational and systems biology approaches in new drug discovery, especially in screening druggable structure of targeted proteins and selecting drug candidates[155]. Computational structure biology has already been used widely in drug candidate screening. For example, Baken *et al.* analyzed a pharmacophore model (PM), and selected seven compounds for further experimental screening out of over two millions of candidates[156]. The procedure can be more efficient and effective by placing drug discovery in the context of network dynamics.

In summary, EMT is a complex process and many pathways crosstalk extensively to initialize and regulate EMT. Therefore integrated computational and experimental approaches are necessary to tackle the molecular and cellular regulation mechanisms, and optimize biomedical intervention strategies.

1.3 GENE REGULATION IS MUTUALLY DEPENDENT ON CHROMOSOME STRUCTURE DYNAMICS DURING CELL TYPE TRANSITION

In multicellular organisms, cells continuously receive internal or environment signals, such as signals from development, senescence, or in response to stress, illness, *et.al.*. Cells must adjust accordingly in response to these signals to survive or to perform the specific function for the organisms[157-160]. The intrinsic part of cell responses is tuning gene expression. Genes are

information written in the form of DNA sequences. To maintain the stability of DNA, in eukaryotic cells, DNA chains form double helix and bind with structure proteins (*e.g.* histones) to form chromatins. Thus, regulation of gene expression can be achieved at distinct levels[161]. Binding of regulatory factors, such as transcription factors (TFs) to specific DNA site regulates transcription activities of individual genes. In contrast, alternation of the spatial organization and epigenetic modification patterns of an extended region of a chromosome can have impacts on a large number of genes. The two scales of regulation are mutually dependent on each other. A large-scale modification starts from the alteration of single gene expression level and proceeds through successive steps, which may involve more genes. On the other hand, regulation on a targeted gene is controlled by both specific regulators and large-scale chromosome modifications. Therefore, it is important to understand the coupling mechanisms of gene regulation at distinct hierarchies. In this section, we will first review gene regulation through chromosome conformation, epigenetic modifications and TF binding. Then, we will brief introduce current technology progressions that help us to investigate gene regulation mechanisms.

1.3.1 Regulation of gene expression takes place at three distinct layers

In eukaryotic cells, DNA double helix binds to structure proteins, such as histones, to for stable chromatin structures. First, 147 base pairs wrap with eight histone proteins to form an 11 nm-diameter core nucleosome structure. The DNA thus assumes a bead-on-a-string structure, which then undergoes chemical modification, recruits additional structural proteins, and further folds into three-dimensional hierarchical chromosome structures. The three-dimensional DNA packing

affects its local accessibility to regulatory elements such as TFs. In this section, we will review two levels of regulation on chromosome accessibility, and gene-TF reactions that interplay with DNA spatial structure and dynamics.

1.3.1.1 Conformational change of chromosomes coordinates global gene expression

Transformation of chromosome spatial structure is both the reason and consequence of gene regulation during cellular transition. Condensed chromosome regions prevent binding of regulatory factors or RNA polymerase physically, while low packing level generally indicates a higher local accessibility. Thus, an effective way of regulating multiple spatially colocalized genes is controlling the chromosome density of an extended area. A/B compartment is a typical example of this type of gene regulation[162].

A/B compartments are computationally deduced from Hi-C data, which reflect packing density of relatively large lengths of chromosomes[162]. DNA within an A compartment is more loosely packed compared to that in B compartment. In other word, genes on A compartment have high possibility to bind with TFs and RNA polymerases. Loci within the same compartment tend to be closer spatially than the loci in different compartments[162,163]. A/B compartments are conserved under most mild stimulation[164,165], but switches between A-B compartments are observed in cell type transition[165] and coordinates large group of genes expression[166]. Euchromatin and heterochromatin are similar to A/B compartments, but are directly observed under optical microscope after staining nuclear regions with aniline dyes. Compared to heterochromatin, euchromatin is lighter stained, which indicates a lower density[167,168]. Heterochromatin is clustered at nucleolus and periphery of a nucleus, while euchromatin fills the remaining space in the nucleus. Switches between euchromatin and heterochromatin were

observed in many cell transition processes, such as in differentiation from ESC to neural cell[169]. Conversions between A/B compartments or euchromatin/heterochromatin have impact on multiple genes contained in the regions. They coordinate and maintain expression pattern changes of cell type specific genes.

Topologically associating domains (TADs) are smaller chromosome structure units. TAD structures are maintained by DNA and structure proteins, such as cohesion, CTCF, and so on[170]. DNA segments within one TAD have higher frequency to interact with each other than DNA segments belonging to different TADs [171]. Structural TADs are relatively conserved among cells[172]. Destroy of conserved TAD boundary is associated to many severe diseases, such as polydactyly[173]. Some TADs are more flexible that can be changed during cell phenotype transition, such as by the promoter-enhancer-TFs structure, or by eRNA-promoter-enhancer structure[174,175]. Several TADs may also form a metaTAD structure, whose structure changes are even more obvious in cell type transition[176].

1.3.1.2 Epigenetic modification affects local accessibility of chromosomes Cells with the same genetic materials assume different cell types largely because of epigenetic modification[177,178]. Two types of epigenetic modification are typically considered, DNA methylation and histone modification. Regulation of both of them involves collaboration among groups of enzymes for reading, writing, and erasing the covalent modification marks[179]. Epigenetic modifications change local DNA structure and thus play a significant role in regulating a group of co-localized genes.

DNA methylation: DNA methylation is catalyzed by DNA methyltransferases, and is highly associated with inactive genes[180]. For example, DNA methylation on CpG islands in promoter regions is related to low gene expression level[181,182]. However, the importance of methylation at promoter regions on gene inactivation is questioned[182,183]. A major function of DNA methylation is to condense local chromatin. Regions of DNA with elevated level of methylation become more condensed and thus less accessible. Also, methylated DNA recruits proteins that help on maintaining tight chromatin structure[181,184].

Presence of DNA methylation affects histone modification. It provides recognizable sites for histone modification writers[185]. In turn, histone modification marks also have impact on DNA methylation[183,186]. For example, some of them (e.g., H3K4me1) are pioneers for gene activation by losing DNA structure from methylation. We will introduce them in detail below.

H3K9: Different levels of methylation at lysine 9 on histone 3 have distinct functions. Monomethylation (H3K9me1) is often found at promoter regions of active genes. Functions of H3K9me2/3 have been overlooked previously, since its distribution is mostly in 'gene deserts'[187]. H3K9me2 is a prominent character of X chromosome X inactivation[188]. One of the two copies of X chromosome in female cell has to be inactivated mainly through epigenetic modification to keep the gene product balance[189-191]. H3K9me3 is mainly associated with heterochromatin regions[187,192]. Especially during cell type transition, H3K9me3 regulates formation of heterochromatin and thus prevents access of TFs and other regulators. Genes that have to be inactive in the new cell type are therefore shielded[192]. Acetylation of H3K9 (H3K9ac) at promoter regions indicates gene activation. The most important function of H3K9ac is to promote transition from transcription initiation to elongation[193].

H3K4: All kinds of modification at lysine 4 on histone 3 are marks of active genes.

Monomethylation of H3K4 (H3K4me1) is considered as the pioneer histone mark for enhancer activation[194]. Collaborating with DNA demethylation, the function of H3K4me1 is to 'open' the tight chromatin region. H3K4me1 recruits epigenetic modification and other DNA regulatory proteins and labels local chromatin regions for active regulators of genes[195,196], so the DNA sequence is available for regulator binding or further histone modification[197]. Recent studies showed that H3K4me1 also participates in gene activation maintenance by fine-tuning gene transcriptional activities[198-200]. Trimethylation (H3K4me3) is another major epigenetic modification at the same site. It can be found widely distributing at promoter region near transcription start site (TSS) and is associated to gene activation[201]. Regulation and function of H3K4me3 is important for cell development, since it recruits additional proteins to open the chromatin and/or directly start the preinitiation of transcription[202]. Moreover, combination of H3K4me3 and H3K27me2, another repressive mark, H3K27me2, on promoters of bivalent genes are important for stem cells to keep pluripotency and flexibility[203].

H3K27: Methylation and acetylation on H3K27 have opposite effects. As mentioned before, H3K27me2 is a repressive mark that can be found in many bivalent gene[203]. H3K27me3 is tightly associated to promoter regions of inactive genes with PRC2 repression function[204]. Since one lysine can not be modified by both methylation and acetylation simultaneously, H3K27ac is considered an antagonizing modification to H3K27me2/3. Similar to other acetylation modification, H3K27ac is associated with chromatin open state and gene activity. H3K27ac can be found on both promoter and enhancer r egions of active genes[205].

Table 1 Summary of histone modification characters.

Site	Modification	Function to gene	Distribution
H3K9	me1	Active	Promoter
	me2/3	Repressive	Non-gene region
	ac	Active	Promoter
H3K4	me1	Active	Enhancer
	me3	Active	TSS
	ac	Active	Promoter
H3K27	me2	Repressive	Bivalent genes' promoters
	me3	Repressive	Promoters
	ac	Active	Enhancer, proximal and distal TSS

Transcription factors determine expression levels of individual genes: The above two mechanisms are involved in relatively large regions on chromosomes, thus affect many linearly or spatially clustered genes. For an individual gene, initiation of its transcription is determined by binding between promoter and transcription initiation complex (TIC). The latter is composed of TFs and RNA polymerase. TF binding to the promoter changes the 3D DNA structure near the TSS region and thus prepares the region for RNA polymerase binding. TFs can also bind to enhancer regions of the targeted gene and drag enhancers to the promoter region. The chromatin structure near the TSS is then changed, which significantly affects the transcription efficiency of the targeted gene[206].

TFs can be further classified as active or repressive TFs. If a gene needs to be activated in response to cell type transition and the chromatin sites are prepared for docking TFs, an active TF is recruited and binds to the target DNA sequence to promote gene expression. Some TFs bind to the targeted sequence transiently to initiate the transcription process. After other factors take over the role for gene activation, these TFs dissociate from the DNA sequence. Some active

TFs can stay on the targeted DNA sites until the genes need not to be transcriptionally active. SNAIL1 is such an example that activates many EMT related genes, such as ZEB1 and MMP9, by binding to their promoter regions directly[207]. In contrast, repressive TFs are involved to silence some genes that are originally active. A repressive TF functions by recognizing and binding to a target DNA site to change the 3D conformation of the local chromatin, and prevent RNA polymerases from binding to the DNA.

Besides directly changing the binding affinity between DNA and RNA polymerase, many TFs, such as SNAIL1, also regulate individual genes by changing the epigenetic modification marks near TSS sites. By binding to the E-box (5′-CACCTG-3’), SNAIL1 can recruit LSD1 and CoREST as well as other histone modification writers, such as HDAC1/2 and PRC2. These enzymes can remove acetylation modification from H3/H4 or add trimethylation modification to H3K27, respectively[208,209]. By modifying the histone epigenetic marks, SNAIL1 represses many epithelial genes, such as *cdh1*[210,211].

Regulation of gene expression normally requires cooperation of several different TFs. For example, in SNAIL1-induced ZEB1 and MMP9 activation, EGR-1 and SP-1 are involved as the co-regulator[207]. The genes change their transcription state only when all TFs are recruited. As mentioned previously, enhancers also play a key role in gene regulation. One gene can have multiple enhancers, with some of the distal enhancers located as far as hundreds of kb away from the TSS. For gene activation, a TF binds to its specific enhancer and changes the 3D structure of DNA near the gene TSS site. The enhancers, promoter, and the TSS form a hub-like structure bridged by TFs. During cell type transition formation of this hub is initially dynamic and transient, then this hub structure is reinforced by some mechanisms to maintain the cell fate. For

example, if the targeted gene is constitutively expressed in the new cell type, in some situation, the TF is not required any more. Instead, the enhancer itself produces enhancer RNAs that knits the promoter – enhancer regions together stably.

1.3.2 High-throughput technique and data pool at a glance

Numerous biochemical technology have been developed to elucidate the mechanisms underlying gene regulation, such as chromosome conformation capture (3C) to detect long-range chromosome interaction[212], chromatin immunoprecipitation (ChIP) to observe DNA-protein interaction[213], methylated DNA immunoprecipitation (MeDIP) to observe DNA methylation status[183], reverse transcription PCR and qPCR to detect gene expression level. However, each of them can only reveal limited information on specific aspects of chromosome structures and gene regulation. A complete scenario is still kept in black. With technology improvement, especially new-generation sequencing technique, large-scale, high-throughput observation came to utility. For example, RNA-seq combines reverse transcription and sequencing, which allows detecting genome-wise gene expression levels. Similarly, combination of traditional biochemistry methods with next-generation sequencing technology makes it much easier and unbiased to observe events taking place in cells. In this section, I will have a brief introduction about the current technologies that are performed in detecting chromatin interaction, histone modification and TF-gene regulation, as well as advances in computational approaches.

Technology in chromatin interaction: Studying interactions among chromatin loci is one of the most active research fields currently. From the locus-locus interactions, we can deduce information related to gene expression, such as TAD structures, A/B compartments, *et. al*. The 4C technique combines the 3C technology with chip/microarray technology, and can generate

high-throughput datasets on the whole genome[214]. Later, with sequencing technology, chromosome conformation capture carbon copy (5C) was developed[215]. However, this method asks for oligonucleotide pairs matched to every ligation site from 3C, which is an astronomical number. Thus, this method is limited from entire genome investigation. Hi-C[162] is a more powerful method that uses blunt-end adaptor instead of designed oligonucleotides, which makes it possible to capture chromatin interaction genome-wide. Currently, Hi-C technology is the most popular one to detect chromatin interactions. Analyses of Hi-C results lead to identification of TAD structures, which elucidate the spatial relationship among genes. Computational algorithms have been developed to extract from Hi-C data, information such as TF binding sites.

Current technology in epigenetic modifications: Epigenetic modification is another active research area in understanding cell development, differentiation, or other cell fate decision. ChIP-seq, which mates traditional ChIP technique with sequencing techniques, is powerful tool to investigate targeted epigenetic modification marks genome-wide. For example, ChIP-seq can directly read the histone-modified DNA sequences and quantify local epigenetic mark levels[216]. Similarly, using methylation DNA specific antibodies, IP-seq technique can reveal the DNA methylation status, such as mCpG[217], MBP[218].

Current technology in gene accessibility and TF: ChIP-seq, which combines ChIP with sequencing and read the output peaks, is used to detect transcription factors binding sequences or binding states[219]. CHIP-seq can reveal cell type specific TF binding sites, and predict tissue or cell type specific enhancers, promoters, and other factors[220].

Accessible promoter region is a prerequisite for TF-DNA binding. HiC data or histone ChIP-seq data allows indirect speculation on DNA accessibility. Some direct methods are also often used. Two of them, assay for transposes accessible chromatin (ATAC)-seq and DNase1-

seq are the most popular techniques currently. Both of them are based on activities of DNA cutting enzymes. ATGC-seq uses a modified hyperactive transposase to cut the exposed DNA[221], while DNaseI-seq use DNAaseI to cut the open chromatin regions[222]. Researchers normally choose one of them based on their experimental design: ATGC-seq requires a small amount of fresh samples. DNase1-seq use large number of fixed samples but can get higher resolution.

Computational technology improvement: Improvement of biochemistry or sequencing technology offers only generation of the high-throughput data. Advances in computational methods to read, store, interpret, calculate, and comprehensive analyze the big data are equally important. The data mass for seq data is easily to reach TB, and analysis of the data has tremendous requirements on memory and resource. Progresses in computer science, informatics and improvement in data mining algorithms make efficient analysis of high-throughput data possible[223,224]. Development of machine learning algorithms further adds predictive power. For example, one can use Hi-C and DNase1 input data to train the code and predict possible TF binding sites in other cells.

2.0 PATHWAY CROSSTALK ENABLES CELLS TO INTERPRET TGF-β DURATION

The detection and transmission of the temporal quality of intracellular and extracellular signals is an essential cellular mechanism. It remains largely unexplored how cells interpret the duration information of a stimulus. In this paper, we performed an integrated quantitative and computational analysis on TGF-β induced activation of SNAIL1, a key transcription factor that regulates several subsequent cell fate decisions such as apoptosis and epithelial-to-mesenchymal transition. We demonstrate that crosstalk among multiple TGF-β activated pathways forms a relay from SMAD to GLI1 that initializes and maintains SNAILl expression, respectively. SNAIL1 functions as a key integrator of information from TGF-β signaling distributed through upstream divergent pathways. The intertwined network serves as a temporal checkpoint, so that cells can generate a transient or sustained expression of SNAIL1 depending on TGF-β duration. Furthermore, we observed that TGF-β treatment leads to an unexpected accumulation of GSK3 molecules in an enzymatically active tyrosine phosphorylation form in Golgi apparatus and ER, followed by accumulation of GSK3 molecules in an enzymatically inhibitive serine phosphorylation in the nucleus. Subsequent model analysis and inhibition experiments revealed that the initial localized increase of GSK3 enzymatic activity couples to the positive feedback loop of the substrate Gli1 to form a network motif with multi-objective functions. That is, the

motif is robust against stochastic fluctuations, and has a narrow distribution of response time that is insensitive to initial conditions. Specifically, for TGF-β signaling, the motif ensures a smoothelay from SMAD to GLI1 on regulating SNAIL1 expression

2.1 INTRODUCTION

Cells live in a state of constant environmental flux and must reliably receive, decode, integrate and transmit information from extracellular signals such that response is appropriate[157,225-227]. Dysregulation of signal transduction networks leads to inappropriate transmission of signaling information, which may ultimately lead to pathologies such as cancer. Therefore, a central problem in systems biology has been to untangle how quantitative information of cellular signals is encoded and decoded. In general cells respond to one or more properties of a stimulus, such as its identity, strength, rate of change, duration and even its temporal profile[228-234]. There are extensive studies on the dose-response curves to reveal how cells respond differentially to a signal with different strength. In comparison, how cells respond to the temporal code of signals is less studies, and its physiological relevance receives much attention recently since most extracellular signals exist only transiently and cellular responses show dependence on signal duration[235-239].

Transforming growth factor-β (TGF-β) is a secreted protein that regulates both transient and persistent cellular processes such as proliferation, morphogenesis, homeostasis, differentiation, and the epithelial-to-mesenchymal transition (EMT)[37,47,49,75,240]. Because it plays

essential roles in developmental and normal physiological process, and its dysregulation is related to cancer, fibrosis, inflammation, Alzheimer's disease and many other diseases, the TGF-β signaling pathway has been probed extensively as a potential pharmaceutical target[241,242].

Several quantitative studies have expanded our knowledge on how the TGF-β-SMAD signaling pathway transmits the duration and strength information of the signal[243-247].

TGF-β can activate both SMAD -dependent and multiple SMAD -independent pathways, which then converge onto some downstream signaling elements. It is unclear how cells transmit and integrate information of the TGF-β signaling distributed among these parallel pathways. Addressing this question requires studies beyond the TGF-β/SMAD axis as in earlier work, where quantifying SMAD proteins serves as the fundamental readout[243-245]. Here we focused on expression of SNAIL1, which is such a downstream target and plays a key role in regulating a number of subsequent processes. Our results confirmed that crosstalk between the SMAD-dependent and independent pathways is key for cells to decode and transmit temporal and contextual information from TGF-β. We posit that the mechanism may be a central mechanistic signal transduction process as many signaling pathways share the network structure.

2.2 RESULTS

2.2.1 Network analysis reveals a highly connected TGF-β signaling network

Through integrating the existing literature, we reconstructed an intertwined TGF-β-SNAIL1 network formed with SMAD-dependent and -independent pathways (Supplementary Fig. 10). For further studies we then identified a coarse-grained network composed of a list of key molecular species (Fig. 4, and Supporting text for details). Along the canonical SMAD pathway,

TGF-β leads to phosphorylation of SMAD2 and/or SMAD3 (pSMAD2/3), followed by nuclear entry after recruiting SMAD4 and forming the complex. The complex acts as a direct transcription factor for multiple downstream genes including SNAIL1 and I-SMAD[62,243]. I-SMAD functions as an inhibitor of pSMAD2/3, thus closes a negative feedback loop. TGF-β also activates GLI1, a key component of the Hedgehog pathway, both through transcriptional activation by pSMAD2/3, and through suppressing the enzymatic activity of glycogen synthase kinase 3 (GSK3). The latter is constitutively active on facilitating GLI1 and SNAIL1 protein degradation in untreated epithelial cells[122,248], thus suppressing GSK3 is expected to lead to GLI1 and SNAIL1 protein accumulation. Other non-SMAD signaling pathways, such as MAPK, ERK, et.al., may also impact on SNAIL1 expression but to a less extent[62,249]. We represented them as 'others' in the model without further explicit treatment within the period of TGF-β treatment studied here. Therefore, the network integrates multiple feed-forward loops that converge at the regulation of SNAIL1 transcription. In the following sections, we will examine the functional roles of individual pathways in the network using several human cell lines.

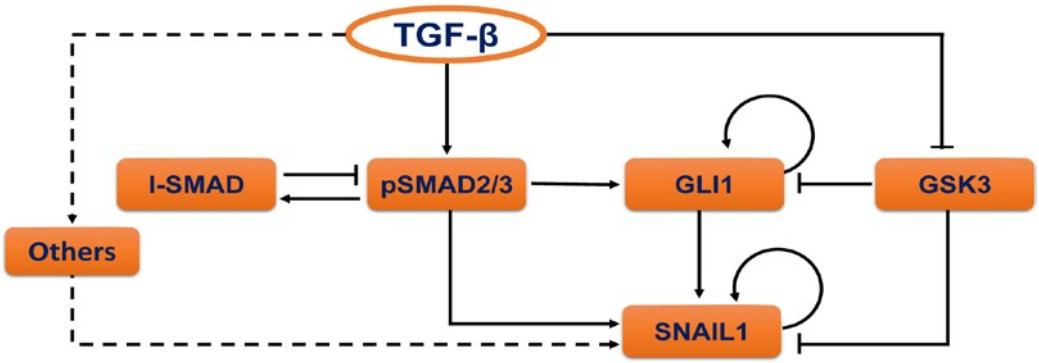

Figure 4 TGF-β induced signaling crosstalk network converges to SNAIL1.Reconstructed literature-based pathway crosstalk for TGF-β induced SNAIL1 expression. The node "Others" refer remaining SNAIL1 activation pathways that have minor contributions to the time window under study and thus are not explicitly treated.

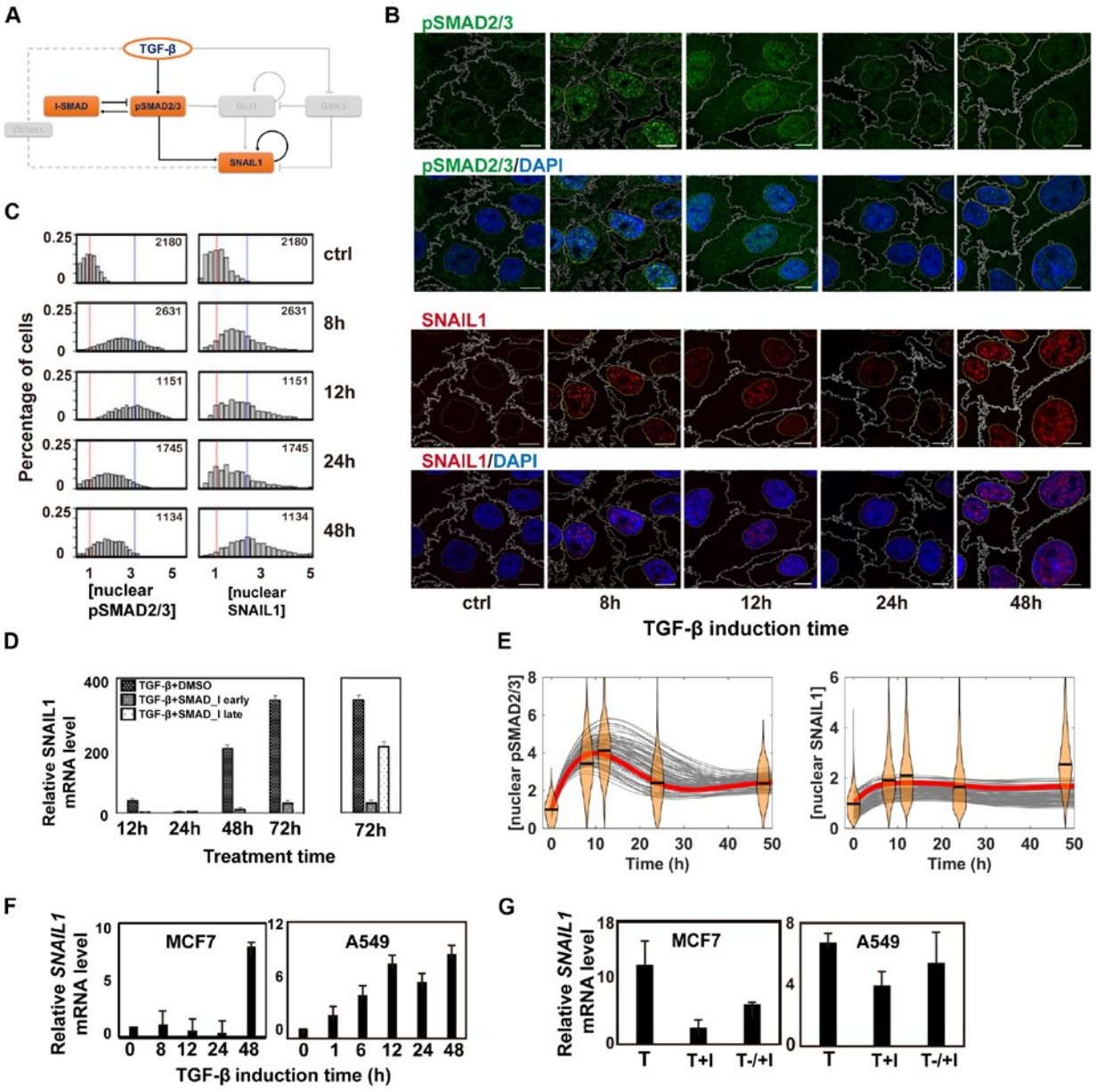

Figure 5 The SMAD proteins induce the first wave of SNAIL1. Canonical SMAD-dependent pathway for TGF-β activation of SNAIL1 highlighted from the network in Fig. 4. Two-color immunofluorescence (IF) images of pSMAD2/3 and SNAIL1 of MCF10A cells induced by 4 ng/ml TGF-β1 at various time points. The scale bar is 10 μm and is the same for other IF images in this paper. **(C)** Distributions of nuclear pSMAD2/3 and SNAIL1 concentrations quantified from the IF images. Red vertical lines indicate the mean value of the distributions at time 0, and blue vertical lines represent that at 12 h (for pSMAD2/3) or at 48 h (for SNAIL1), respectively. The number marked in each figure panel is the number of randomly selected cells used for the analysis. Throughout the paper we

report fold changes of concentration and amount relative to the mean basal value of the corresponding quantity. **(D)** Effects of early (added together with TGF-β) and late (48 h after adding TGF-β) pSMAD inhibition on the *SNAIL1* mRNA level in MCF10A cells. **(E)** Thorough parameter space search confirmed that with the model in panel A one can fit the pSMAD2/3 dynamics, but not the two-wave SNAIL1 dynamics. The experimental data are shown as violin plots with the medians given by black bars. Solid curves are computational results with parameter sets sampled from the Monte Carlo search, and the red curves are the best-fit results. **(F)** Fold change of *SNAIL1* mRNA levels in MCF7 and A549 cells measured with quantitative RT-PCR after TGF-β1 treatment. **(G)** Fold change of *SNAIL1* mRNA levels measured with quantitative RT-PCR at 72 h after TGF-β1 (T) treatment. For early inhibition (T+I) the inhibitor was added at the time of starting TGF-β1 treatment. For late inhibition (T-/+I) the inhibitor was added 48 h (for MCF7) and 24 h (for A549) after starting TGF-β1 treatment, respectively. The inhibition results were compared to the TGF-β treatment (T) result at the same time point.

2.2.2 The canonical TGF-β/SMAD pathway initializes a transient wave of SNAIL1 expression.

First we examined the TGF-β/SMAD/SNAIL1 pathway (Fig. 5A) by treating human MCF10A cells with recombinant human TGF-β1, and performing multicolor immunofluorescence (IF) using antibodies directed against pSMAD2/3, SNAIL1. As expected from the pSMAD/I-SMAD negative feedback loop, pSMAD2/3 proteins accumulated in the nucleus transiently, peaking at around 12 hours after TGF-β1 treatment, followed by a decrease by 24 hours (Fig. 5B & C). We confirmed the transient pSMAD2/3 dynamics by sampling 1100-2600 cells at each time point (Fig. 5C). The result is also consistent with reports in the literature[243,245,250]. Nuclear SNAIL1 concentration rose concurrently with pSMAD2/3 (Fig. 5B & C), then there was a transient dip at 24 hours, followed by another increase then a persistent elevation for one week[13].

Next, we investigated the function of phosphorylated SMAD2/3 on promoting snail1 transcription during TGF-β treatment. In addition to adding TGF-β, we treated MCF10A cells with an inhibitor LY2109761, which prevents SMAD2/3 phosphorylation through inhibiting TGF-β receptor kinase activity (Fig. 5D). Without the inhibitor, the *SNAIL1* mRNA showed the two-wave dynamics consistent to that of the protein. When the inhibitor was added concurrently with TGF-β treatment, the *SNAIL1* mRNA level was reduced to ~ 9% of that of the control experiment (without inhibitor) by day 3. This result is consistent with previous observation that directly blocking SMAD2/3 phosphorylation or pSMAD activation at the early stage of TGF-β treatment depletes snail1 expression significantly (1-4), and indicates that indeed pSMAD2/3 are required for SNAIL1 initial activation. However, the *SNAIL1* mRNA level remained ~ 70% when the inhibitor was added 48 hours after initiation of TGF-β treatment (when nuclear pSMAD2/3 concentration has dropped to a minimum). Furthermore we constructed a mathematical model that contains only the TGF-β/SMAD/SNAIL1 pathway, and performed a thorough parameter space search using a multi-configuration Monte Carlo algorithm (Supplementary Fig. 11). The search revealed regions of the parameter space that quantitatively reproduced the transient pSMAD2/3 dynamics, but not the two-wave dynamics of SNAIL1 expression (Fig. 5E). This computational result further confirmed that pSMAD2/3 is less essential for the second wave of SNAIL1.

Furthermore, this SNAIL1 dynamics is not cell type specific as equivalent two-wave dynamics were seen for *SNAIL1* mRNA in MCF7 and A549 cells (Fig. 5F). Similar to that of MCF10A, it is more effective on inhibiting *SNAIL1* mRNA by adding LY2109761 together with TGF-β than later (Fig. 5G). The impact of SMAD phosphorylation inhibitor on A549 is less than that on MCF10A or MCF7 at either early or late inhibition, which could be due to the higher

level of EMT-related factors in A549[251]. In total, these results reveal that pSMAD2/3 is essential for the early phase of SNAIL1 activation, but is less important for the secondary phase elevation and persistence of SNAIL1 expression/localization.

2.2.3 GLI1 contributes to activating the second wave of SNAIL1

The regulatory network suggests that GLI1 may be responsible for the second wave of SNAIL1 (Fig. 6A). To test this hypothesis, we performed microscopy studies of SNAIL1-GLI1 using MCF10A cells. The distribution of SNAIL1 found in this study (Supplementary Fig. 12A) was consistent with those from the pSMAD2/3-SNAIL1 studies. Elevated and sustained expression of GLI1 under TGF-β treatment (Fig. 6B & C) was clearly evident. More interestingly GLI1 also showed an unexpected multi-phasic dynamic. Around 8 hours after TGF-β treatment, cytosolic GLI1 concentration started to increase. At 12 hours when SMAD activities decreased toward basal levels there was a clear accumulation of GLI1 in the nucleus, which continued to increase through day 2. Notably, at this time point cells expressing a high level of nuclear SNAIL1 consistently showed high nuclear GLI1 concentrations (Supplementary Fig. 12A). Expanding the mathematical model of the network to Fig. 5A also reproduced the temporal dynamics of pSAMD2/3 and SNAIL1 (Supplementary Fig. 12B), supporting the role of GLI1 as the activator of the second wave of SNAIL1.

If GLI1 is mainly involved only in the later maintenance of SNAIL1 expression, it is reasonable to predict that inhibiting GLI1 activity, either at the onset of or at some subsequent time after TGF-β treatment, would have minimal effect on the pSMAD2/3 induced initial wave of SNAIL1 expression. However, GLI1 inhibition would severely reduce the second wave of SNAIL1 expression. Indeed this was what observed experimentally. When GLI1 inhibitor

GANT61 was added together with TGF-β at the beginning of the experiment, the *SNAIL1* mRNA level was reduced to be 55% (at 12 h and 24 h), 12% (at 48 h) and 7% (at 72 h) compared to that without inhibition at the corresponding time points (Fig. 6D). In another experiment adding the inhibitor 48 hours after TGF-β treatment also reduced the mRNA level measured at 72 h to be 25% (Fig. 6E). These results are qualitatively different from those with the SMAD inhibitor (Fig. 5D).

To confirm that GLI1 activation is not restricted to the MCF10A cell line, we also examined MCF7 and A549 cells with TGF-β treatment. We observed similar increased and sustained GLI1 expression, albeit with initial slight downregulation before 12 h, possibly due to cell line specific activation of some GLI1 inhibition pathways (Fig. 6F). Furthermore, early and late GLI1 inhibition lead to a reduction of the *SNAIL1* mRNA level to be 13% and 22% for MCF7 cells, and to a less extent of 57% and 66% for A549 cells, respectively (Fig. 6G). Additionally, increased GLI1 expression after TGF-β treatment has been found for multiple liver cancer cell lines[252]. *In toto* these results support the role of GLI1 as a signaling relay from pSMAD2/3 to SNAIL1.

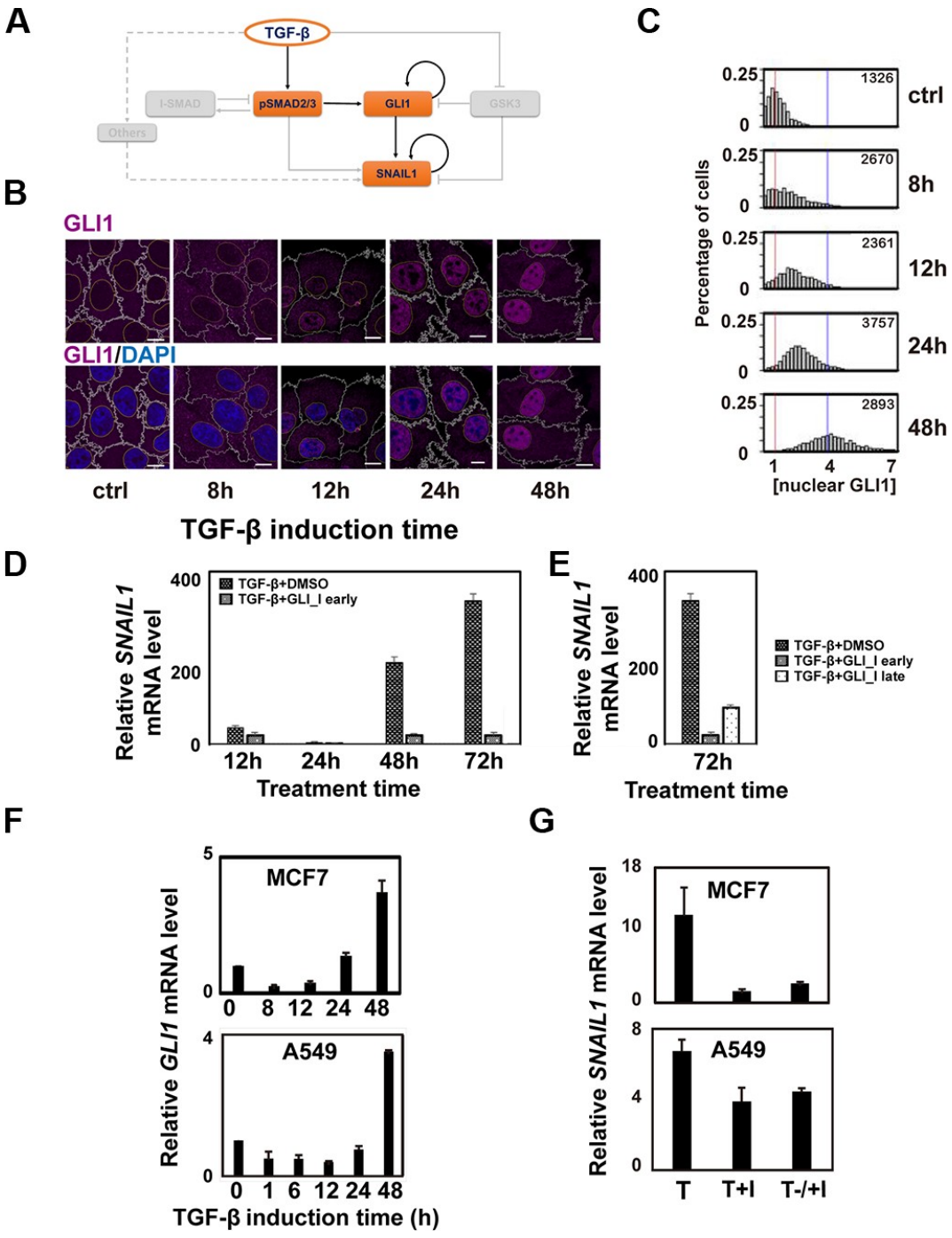

Figure 6 GLI1 is a major contributor to activate the second wave of SNAIL1 expression **(A)** TGF-β activates the GLI1/SNAIL1 module partly through pSMAD2/3. **(B)** IF images on protein levels of GLI1 (in the free form). Red and blue vertical lines indicate the mean values of the distributions at time 0 and at 48 h, respectively. **(C)** Distributions of nuclear GLI1 concentrations quantified from the IF images. **(D)** Experimental alidation of the

results for early (added together with TGF-β) GLI1 inhibition on the *SNAIL1* mRNA level in MCF10A cells. **(E)** Experimental validation of the results for late (48 h after adding TGF-β) GLI1 inhibition on the *SNAIL1* mRNA level in MCF10A cells. **(F)** Fold change of *GLI1* mRNA levels measured with quantitative RT-PCR at different time points after combined TGF-β1 treatment in MCF7 or A549 cells. **(G)** Fold change of *SNAIL1* mRNA levels measured with quantitative RT-PCR at 72 h after combined TGF-β1 and GLI1 inhibitor GANT61 treatment in MCF7 or A549 cells. For early inhibition (T+I) the inhibitor was added at the time of starting TGF-β1 treatment. For late inhibition (T-/+I) the inhibitor was added 48 h (for MCF7) and 24 h (for A549) after starting TGF-β1 treatment, respectively. TGF-β treatment group (T) is shown as a positive control.

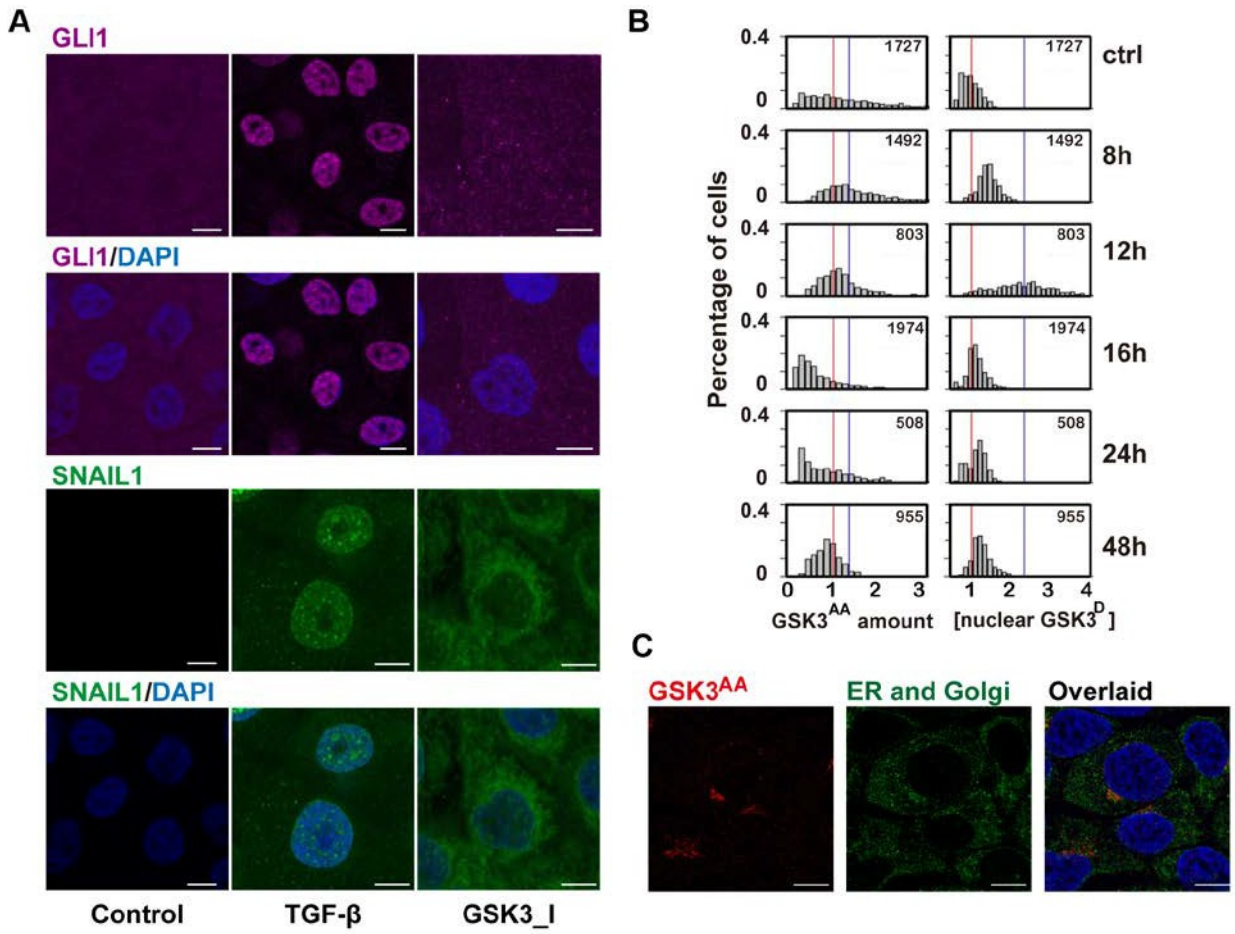

Figure 7 TGF-β induced temporal switch between active and inhibitive phosphorylation forms of GSK3 proteins **(A)** IF images showed that inhibiting GSK3 enzymatic activity alone increased SNAIL1 accumulation but did not recapitulate TGF-β induced GLI1 nuclear translocation. **(B)** Quantification of the IF images of MCF10A cells at different time points after TGF-β treatment. Red vertical lines indicate the mean value of the distributions at time 0, and blue vertical lines represent that at 8 h (for $GSK3^{AA}$) or at 12 h (for $GSK3^{D}$), respectively. **(C)** IF images showing $GSK3^{AA}$ localization at the endoplasmic reticulum center (ERC).

2.2.4 GSK3 in a phosphorylation form with augmented enzymatic activity accumulates at endoplasmic reticulum and Golgi apparatus.

Next, we hypothesized that GSK3 is fundamental to the observed multi-phasic GLI1 dynamic (Fig. 6B). Most published studies suggest that GSK3 is constitutively active in untreated cells, facilitating degradation of SNAIL1 and GLI1; TGF-β treatment leads to GSK3 phosphorylation and inactivation, which leads to an accumulation of SNAIL1 and GLI1[141,253]. Initially we tested whether the above mechanism is sufficient to explain the multi-phasic GLI1 dynamics. We treated MCF10A cells in the absence of TGF-β with a GSK3 activity inhibitor. Given the above mechanism, one should expect the GSK3 inhibitor to promote both GLI1 and SNAIL1. In our experiment, SNAIL1 did increase in the nucleus and even more in the cytoplasm due to inhibition of GSK3-dependent SNAIL1 degradation, but there was no noticeable change in GLI1 expression in either nucleus or cytoplasm (Fig. 7A), suggesting additional signaling mechanisms may be involved. Besides the inhibitory serine phosphorylation (S21 in GSK-3α and S9 in GSK-3β), previous studies showed that tyrosine (Y279 in GSK-3α and Y216 in GSK-3β) phosphorylation leads to augmented enzymatic activity of GSK3[254]. As a convenience when discussing the three forms of GSK3, we refer the enzymatically active unphosphorylated form and the more active tyrosine phosphorylated form as "$GSK3^{A}$" and "$GSK3^{AA}$", respectively, and the inactive serine phosphorylation form as "$GSK3^{D}$". Also we reserve "GSK3" for the total GSK3. As expected, microscopy studies showed an increased concentration of $GSK3^{D}$ peaking around 12 hours after TGF-β treatment (Fig. 7B, Supplementary Fig. 13A). Large cell-to-cell variations in the concentration of $GSK3^{D}$ were observed, however, the abundance of cytosolic and nuclear $GSK3^{D}$ were essentially equivalent (the expression ratio was close to one) for cells

without TGF- β treatment (Supplementary Fig. 13B). This observation corroborates earlier report that the serine phosphorylation does not affect GSK3 nuclear location[255]. TGF-β treatment led to transient deviation of this ratio from equivalence, reflecting additional active and dynamic regulation of GSK3 including covalent modification, location and protein stability. Specifically prior to inhibitory serine phosphorylation we observed transient $GSK3^{AA}$ accumulation in the perinuclear region peaking at eight hours (Fig. 7B, Supplementary Fig. 13A). Close examination of higher magnification confocal images revealed that the $GSK3^{AA}$ formed clusters in the endoplasmic reticulum (ER) and Golgi apparatus, but not associated with actin filaments (Fig.7C). Given that a function of active GSK3 is to modify target proteins post-translationally, our observation suggests an unreported role for $GSK3^{AA}$ accumulating at the ER and Golgi apparatus as to modify newly synthesized proteins before their release to the cytosol. Specifically previous studies showed that in mammalian cells a scaffold protein SUFU binds to GLI to form an inhibitory complex; SUFU phosphorylation by GSK3β prevents the complex formation, and exposes the GLI1 nuclear localization sequence[142]. This mechanism explains the observed increase of free GLI1 in the cytosol followed by nuclear translocation (Fig. 6B). Since the two phosphorylation forms, $GSK3^{AA}$ and $GSK3^{D}$, coexist within single cells at defined time points, we performed co-immunoprecipitation and found that the probability of having the two GSK3 phosphorylation forms in one molecule was rare (Supplementary Fig. 13C).

Contrary to our observation that TGF-β regulates $GSK3^{AA}$ dynamics, other studies posit that $GSK3^{AA}$ is not regulated by external cues[256]. To resolve this paradox, we measured the relative amount of different GSK3 forms through silver staining (Fig. 13D). Among the three forms, the overall percentage of $GSK3^{D}$ increased from a basal level of 37% to 65% at 12 h after TGF-β treatment. In contrast, only a small fraction of GSK3 molecules assumed the $GSK3^{AA}$

form and its overall abundance was stable over time (from ~10% basal level to ~13% at 8 h then back to ~10% at 12 h after TGF-β treatment). Essentially $GSK3^{AA}$ did not change in abundance but did change in localizations (homing to the ER and Golgi apparatus) to form a high local concentration, which imbue an important role in TGF-β signaling.

2.2.5 A temporal and compartment switch from active to inhibitory GSK3 phosphorylation smoothens the SMAD-GLI1 relay and reduces cell-to-cell heterogeneity on GLI1 activation.

Based on the above results, we constructed an expanded network for TGF-β induced SNAIL1 expression (Fig. 8A), which integrates a role for GSK3 and its temporal change of enzymatic activities in the cytosol and nucleus (Supplementary Fig. 13E). The model reproduces the multiphasic dynamics of GLI1 as well as that of pSMAD2/3 and SNAIL1 (Fig. 14A).

To understand the function of the early nuclear accumulation of GLI1 induced by $GSK3^{AA}$, it is important to recognize that GLI1 has a positive-feedback loop, and this network motif (Fig. 8B, left panel without the part in green) has characteristic sigmoidal shaped temporal dynamics, with the substrate concentration increasing slowly at first then accelerating with time until it approaches saturation (Fig. 8B, right panel, red curve). The response time, t_R, defined as the time taken to reach a target concentration value $[X]_R$, is highly sensitive to initial substrate concentration $[X]_0$: in fact a slight increase in the initial concentration, $\Delta[X]$, can significantly shorten the response time (Fig. 8B, right panel, blue curve). In contrast, one can accelerate the response time with an expanded network (Fig. 8B including the green part) that the signal triggers a fast conversion of the substrate from a preformed inhibitory form (X_I) to active form

X, effectively a boost of $[X]_0$ to $[X]_0 + \Delta[X]$. For a fixed $\Delta[X]$ a greater acceleration is seen in cells with lower initial concentrations (Fig. 8B, right panel, inlet figure). Consequently despite variations of their initial concentration $[X]_0$, most cells within a population can reach $[X]_R$ by a targeted time point t_T in a series of temporally regulated events such as cell differentiation and immune response. Indeed, many examples of this modified feedback loop motif exist. Figure S2.5B gives some examples involving members of intrinsically disordered proteins and inhibitors of DNA binding proteins, β-catenin and the STING motif for immune responses. In the present scenario the accelerated GLI1 dynamic ensures sufficient accumulation of GLI1 before nuclear pSMAD2/3 level decreases, essentially analogous to a relay race when the first runner can only release the baton after the second runner has grabbed it. Later when the GLI1 and SNAIL1 concentrations start to increase, the $GSK3^A \rightarrow GSK3^D$ conversion became necessary to reduce the rates of their degradation catalyzed by active GSK3. Interestingly, this conversion takes place concurrently with maximal concentration of nuclear pSMAD2/3, which activates GLI1 and SNAIL1 transcription. Furthermore, the small initial concentration boost does not affect another major function of the positive feedback loop, which is to robustly buffer temporal and strength fluctuations of signals (Supplementary Fig. 14C)[257].

To test the functional roles of GSK3 suggested above, we performed a series of GSK3 activity inhibition experiments. First, we pretreated MCF10A cells with GSK3 inhibitor SB216763, washed out the inhibitor then added TGF-β1 (Supplementary Fig. 14D). We predicted that the treatment would slow down GLI1 nuclear accumulation, and at later times decrease the overall increase of GLI1 and SNAIL1 compared to cells without GSK3 inhibitor. Indeed this was observed (Fig. 8C, TGF-β +/- GSK3_I). More interestingly, the scatter plots (Fig. 8D) show the distributions with and without the inhibitor are similar in cells with high

GLI1, but in the presence of the inhibitor there is a population of non-responsive cells with low GLI1 and SNAIL1. This observation is consistent with model predictions that the GSK3-induced boost of initial GLI1 concentration leads to acceleration in the GLI1 and SNAIL1 dynamics, and this boost is more evident for cells with lower level of initial nuclear GLI1 (Fig. 8E). In a separate experiment (Supplementary Fig. 14E), we did not wash out GSK3 inhibitor while adding TGF-β. In this case the inhibitor had opposite effects on GLI1 and SNAIL1 protein concentrations: it slowed down the initial release and translocation of GLI1 needed to accelerate the GLI1 accumulation, but also decreased GLI1 and SNAIL1 degradation that becomes pre-eminent when the proteins were present at high levels. Compared to the samples grown in the absence of the GSK3 inhibitor, we also observed slower and more scattered GLI1 nuclear accumulation and SNAIL1 increase on day 2, but by day 3 the overall levels of GLI1 and SNAIL1 were actually higher than the case without the inhibitor (Fig. 8D, TGF-β + GSK3_I).

2.2.6 The SMAD-GLI1 relay increases the network information capacity and leads to differential response to TGF-β duration

Our results show that TGF-β1 signaling is effected through pSMAD2/3 directly with fast pulsed dynamics concurrently with a relay through GLI1 which has a much slower dynamics. The signaling ported by these two channels converges on SNAIL1 with a resultant two-wave expression pattern. To further dissect the potential functional interactions between these two pathways, we performed mathematical modeling and predicted that the two distinct dynamics allows cells to respond to TGF-β differentially depending on stimulus duration (Fig. 9A). Short pulses of TGF-β only activate pSMAD2/3 and the first wave of transient SNAIL1 expression. When the signal duration is longer than a defined threshold value, activation of GLI1 will lead to

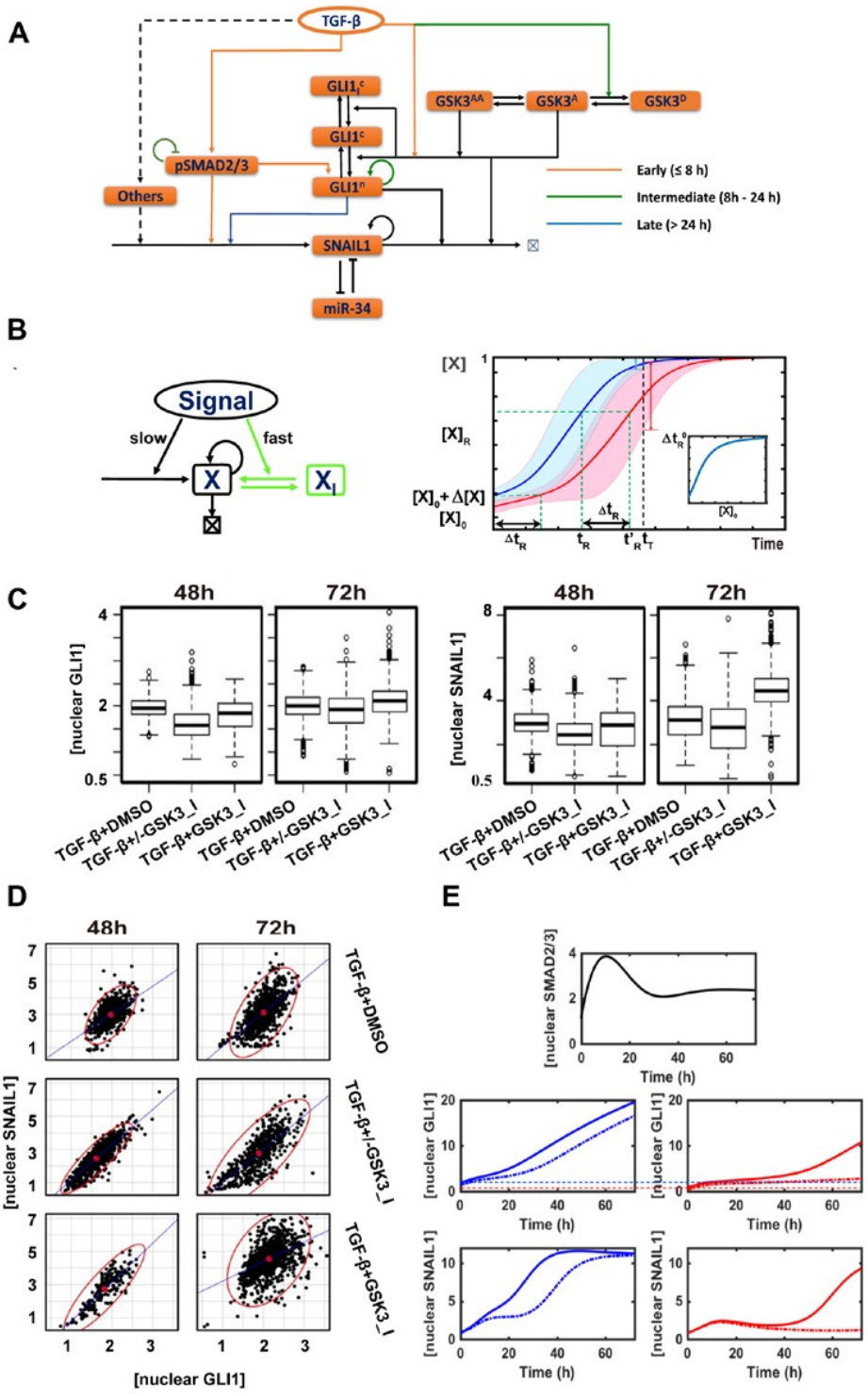

Figure 8 The GSK3 phosphorylation switch smoothens the SMAD-GLI1 relay.**(A)** Proposed expanded network for TGF-β induced SNAIL1 expression. **(B)** *left*: Schematic of a generic positive feedback loop network. Also shown in green is an additional reservoir of the molecules in inactive form (X_I) that can convert quickly into the active form

(X) upon stimulation. *Right*: The response time t_R is sensitive to the initial concentration, $[X]_0$ v.s. $[X]_0 + \Delta[X]_0$. The inlet figure shows the dependence of Δt_R on $[X]_0$ with $\Delta[X]_0$ fixed. **(C)** Box plots of GSK3 inhibition experimental data. **(D)** Scattered plots of GSK3 inhibition experimental data. Red points are the center of the scattered plots and each ellipse encloses 97.5% of the data points. Both were drawn with the R package, *car::data.ellipse*. **(E)** Computational simulation of SMAD, SNAIL1 and GLI1 behavior with (solid line) or without (dotted line) initial boosting in cells with high basal GLI1 level (left panels) or low basal GLI1 (right panels).

the observed second wave of SNAIL1 expression. We confirmed the predictions with MCF10A cells (Fig. 9B). Both TGF-β1 pulses with duration of two hours and eight hours activated pSMAD2/3 and the first wave of SNAIL1 expressions. However, only the eight-hour but not the two-hour pulse activated sustained GLI1 and the second wave of SNAIL1 expression, similar to those with continuous TGF-β1 treatment.

Clearly cellular responses have different temporal profiles depending on the TGF-β duration, and one can use the information theory to quantify their information content[232,233]. In this study we utilized a more intuitive understanding of network function from an information encoding viewpoint. Consider the pSMAD complex, which has three coarse-grained states, High (H), Medium (M), and Low (L), and each of GLI1 and SNAIL1 has two states, H and L (Fig. 9C). Then one can use three 4-element states, (L, L; L, L), (H, L; L, L), (H, M; L, H) to roughly describe the case without TGF-β and the two hour and eight hour pulse results in Fig. 9B, where each number in a state represents in the order the 12 h and 48 h concentrations of pSMAD2/3 and GLI1, respectively. The three states are part of a temporally ordered state space, and encode information of TGF-β duration roughly as not detectable, short, and long. The same information is encoded by the SNAIL1 dynamics as (L, L), (H, L), and (H, H), reflecting SNAIL1 as an information integrator of the two converging pathways.

Further modeling suggests that components in the network function cooperatively to encode the TGF-β information (Supplementary Fig. 15B). Increasing or decreasing the nuclear GSK3 enzymatic activity tunes the system to generate the second SNAIL1 wave with a higher or lower threshold of TGF-β duration, respectively, while changing the cytosol GSK3 enzymatic activity has the opposite effect. Upregulation of GLI1, or downregulation of I-SMAD, both of which have been observed in various cancer cells, also decrease the threshold for generating the second SNAIL1 wave. Therefore cells of different types can share the same network structure, but fine-tune their context-dependent responses by varying some dynamic parameters, and for a specific type of cells dysregulation of any of the signaling network components may lead to misinterpretation of the quantitative information of TGF-β signal.

We have shown that the SNAIL1 dynamics is TGF-β1 duration dependent. To further confirm that cells respond differentially to TGF-β1 with different duration, we measured the mRNA levels of another four genes, all of which respond to TGF-β1 (Supplementary Fig. 15C)[258,259]. Gene *FN1* codes for the cell motility related protein fibronectin. Its expression is activated even by the 2-h TGF-β1 pulse, and increases with longer TGF-β1. Gene *CTGF*, whose product is an extracellular matrix protein and related to cell motility, is activated at similar extent by both 2-h and 8-h TGF-β1 pulses, and its expression level increases by additional 14 folds with continuous TGF-β1 treatment. Expressions of genes *MMP2 and CLDN4*, coding proteins related to mesenchymal extracellular hallmark and cell migration, increase only slightly (less than two folds) with either 2h or 8h TGF-β1 pulse, compared to the more significant change under continuous TGF-β1 treatment. Therefore, these downstream genes also show differential expression patterns depending on TGF-β1 duration, and cells activate different response programs correspondingly.

2.3 DISCUSSION

TGF-β is a multifunctional cytokine that can induce a plethora of different and mutually exclusive cellular responses. A significant open question is how cells interpret various features of the signal and make the cell fate decision. TGF-β can activate a number of pathways interconnected with multiple crosstalk points. Our studies reveal that this interconnection is essential such that components of the network can function coordinately and appropriately to interpret the temporal (time and duration) information from TGF-β.

2.3.1 pSMADs are major inducers for the first wave of SNAIL1 expression.

The two-wave dynamic of TGF-β-induced SNAIL1 expression has been observed in several cellular systems[260,261], supporting the underlying relay mechanism discovered in this work. The first wave is fundamentally induced by pSMAD2/3, as evidenced from our SMAD inhibition experiments, and similarity between the dynamics of pSMAD2/3 and the first wave of SNAIL1. SNAIL1 may act as cofactor of pSMADs to induce other early response genes[79]. At later times the nuclear concentrations of pSMAD2/3 decrease though continue to contribute to SNAIL1 activation at a lower level.

2.3.2 GSK3 fine-tunes the threshold of the GLI1 checkpoint and synchronizes responses of a population of cells

The functional switch from pSMAD2/3 to GLI1 relays information from TGF-β signaling beyond the initial induction of SNAIL1, and this relay is facilitated by a second relay from the active to the inactive phosphorylation form of GSK3 proteins. Active regulation of the abundance and nuclear location of $GSK3^{AA}$ form has been observed in neurons[263]. In contrast to

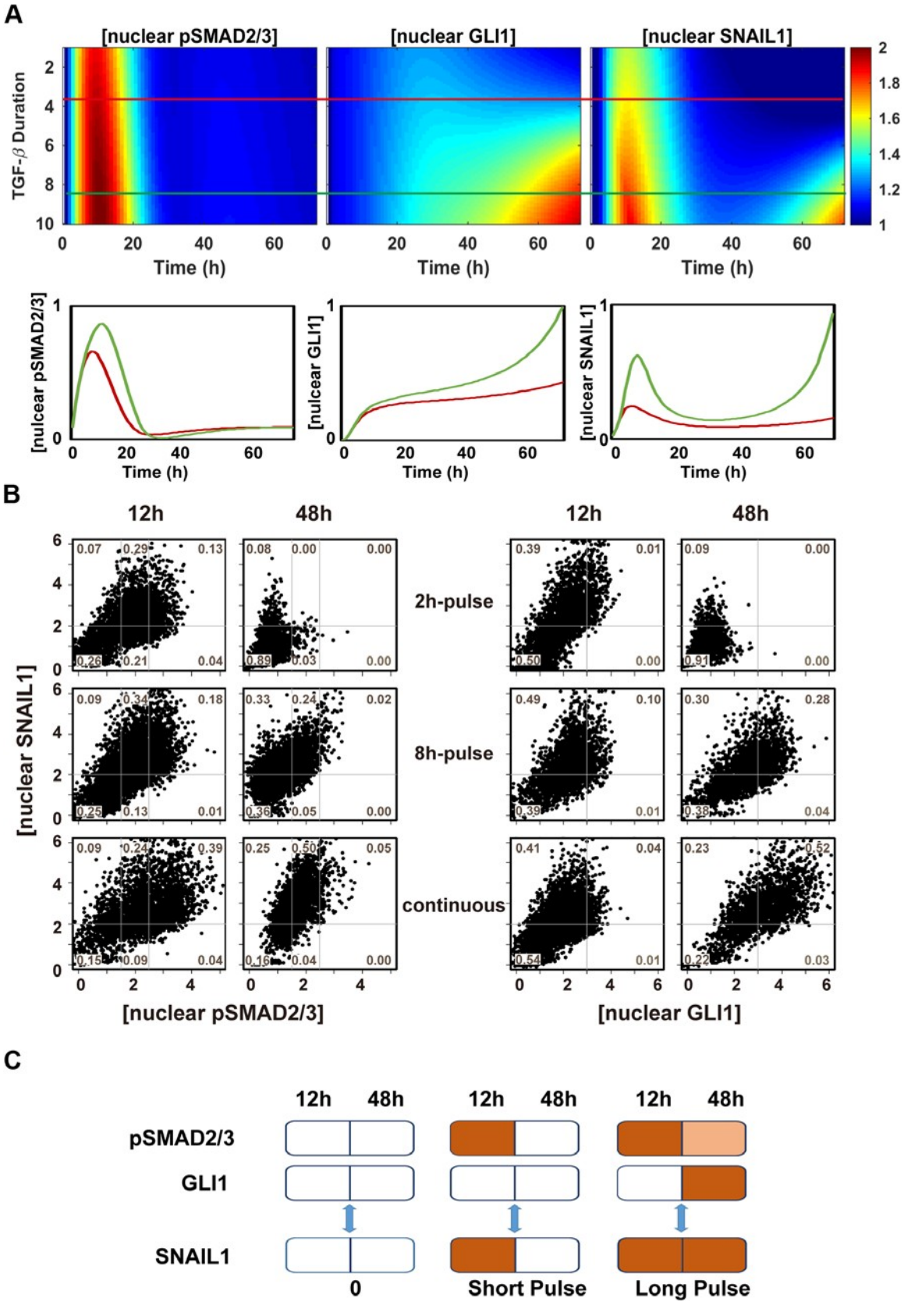
A
[nuclear pSMAD2/3]
[nuclear GLI1]
[nuclear SNAIL1]
TGF-β Duration
Time (h)
[nulcear pSMAD2/3]
[nulcear GLI1]
[nulcear SNAIL1]
B
12h
48h
2h-pulse
8h-pulse
continuous
[nuclear SNAIL1]
[nuclear pSMAD2/3]
[nuclear GLI1]
C
pSMAD2/3
GLI1
SNAIL1
0
Short Pulse
Long Pulse

Figure 9 The TGF-β-SNAIL1 network permits detection of TGF-β duration and differential responses. (A) Model predictions that the network generates one or two waves of SNAIL1 depending on TGF-β duration. The red line overlaid on the heatmap is a sampling time of the short-time TGF-β induction. The green line represents the long-time TGF-β treatment. (B) Single cell protein concentrations quantified from IF images of cells under pulsed and continuous TGF-β treatments. The solid lines divide the space into coarse-grained states with respect to the corresponding mean values without TGF-β treatments (= 1). (C) Schematics of how cells encode information of TGF-β duration through a temporally ordered state space.

these earlier reports we observed an accumulation of $GSK3^{AA}$ in the ER and Golgi apparatus. Mechanistically this may be caused by redistribution of cytosolic $GSK3^{AA}$, or a simple accumulation of *de novo* synthesized and phosphorylated GSK3 proteins. The overall consequence is an increase in local GSK3 enzymatic activity, which forms part of the GSK3 switch that smooths the pSMAD2/3-GLI1 transition and the duration threshold of TGF-β pulse that generates the second wave of SNAIL1.

This seemingly simple process, which accelerates the response time through transient and minor increases in the initial concentration of a molecular species subject to positive feedback control, may have profound biological functions. Positive feedback loops are ubiquitous in cellular regulation, with a major function to filter both the strength and temporal fluctuations of stimulating signals and to prevent inadvertent cell fate change. This network, however, may have an inherently slow response time, and the response is highly sensitive to the initial concentration of the substrate that lead to large cell-to-cell variation of temporal dynamics. This variation and slow dynamic may be problematic for processes such as neural crest formation and wound healing where precise and synchronized temporal control is crucial for generating collective responses of multiple cells. The expanded network shown in Fig. 8B allows transient increase of

the initial substrate concentration, and solves the seemingly incompatible requirements for the simple positive feedback motif on robustness against fluctuations as well as fast and synchronized responses. It assures that despite a possible broad distribution of basal expression levels of the protein, cells are activated within a designated period of time at the presence of persistent activation signal, without sacrificing the filtering function of the feedback loop.

2.3.3 Cells use TOSS formed by a composite network to increase information transfer capacity.

Cells constantly encounter TGF-β signals with different strengths and duration, and must respond accordingly. It is well documented that biological networks reliably transmit information about the extracellular environment despite intrinsic and extrinsic noise in a subtle and functional way. However, quantitative analyses using information theory reveal that the dynamic of each individual readout is quite coarse with one or few bits[232,233]. This is a paradox. However, our results suggest that cells use multiple readouts to generate a TOSS with an expanded capacity to encode signal information and generate a far more subtle response system. For example, the SMAD motif has a refractory period due to the negative feedback loop and thus can accurately encode the duration information of TGF-β only within a limited temporal range. The GLI1 motif encodes information of longer TGF-β duration, which then saturates. This TOSS may be further expanded, such that the SNAIL1 motif itself possibly encodes information of longer TGF-β duration and relays to other transcription factors such as TWIST and ZEB, and leads to stepwise transition from the epithelial to the mesenchymal phenotype depending on the TGF-β duration[13].

Therefore although each motif has limited information coding capacity, a combination of motifs can code and transmit detailed signaling information. This is analogous to the design of a computer composed of many binary logic gates.

As with other signaling process, TGF-β signaling is context dependent, and the dynamic and regulatory network vary between cell types[244,264]. For the three cell lines we examined our results identify GLI1 as a major relaying factor for the TGF-β signaling. The inhibition experiments show that other possible peripheral links have minor contributions to SNAIL1 activation, while their weights may grow at time later than we examined. Consequently the present work has focused on the early event of TGF-β activation of SNAIL1, which is within 72 hs for MCF10A cells. Nevertheless, the relay mechanism and the corresponding network structure identified here can be general for transmitting quantitative information of TGF-β and other signals. It is typical that an extracellular signal is transmitted through a canonical pathway with negative feedbacks and multiple non-canonical pathways, and these pathways crosstalk at multiple points, and Supplementary Fig. 15D gives some examples including IL-12, DNA double strand breaking, and LPS. Therefore, the mechanism revealed in this work is likely beyond TGF-β signaling.

2.3.4 Network temporal dynamics is a key for effective pharmaceutical intervention.

Upregulation of GLI1, and GSK3 and the responsive SMAD family has been reported in pathological tissues of fibrosis[265] and cancer[262], and all three have been considered as potential drug targets. The present study emphasizes that in cell signaling timing is fundamental for function. The same network structure may generate drastically different dynamics with different

parameters, as observed for different cell types. Consequently, effective biomedical intervention needs to take into account the network dynamics. We have already demonstrated that adding the inhibitors at different stages of TGF-β induction can be either effective versus not effective on reducing SNAIL1 (by inhibiting pSMAD2/3), both effective (by inhibiting GLI1), and even opposite (by inhibiting GSK3). Actually, one may even exploit this dynamic specificity for precisely targeting certain group of cells while reducing undesired side effects on other cell types.

In summary through integrated quantitative measurements and mathematical modeling we provided a mechanistic explanation for how cells read TGF-β duration. Several uncovered specific mechanisms, such as expanding information transmission capacity through signal relaying, and reducing response times of positive feedback loops by increasing initial protein concentrations, may be general design principles for signal transduction.

2.4 METHODS

Cell Culture

MCF10A cells were purchased from the American Type Culture Collection (ATCC) and were cultured in DMEM/F12 (1:1) medium (Gibco) with 5% horse serum (Gibco), 100 µg/ml of human epidermal growth factor (PeproTech), 10 mg/ml of insulin (Sigma), 10 mg/ml of hydrocortisone (Sigma), 0.5 mg/ml of cholera toxin (Sigma), and 1x penicillin-streptomycin (Gibco). MCF7 cells were purchased from ATCC and cultured in EMEM medium (Gibco) with

10% FBS (Gibco), 10 mg/ml of insulin, and 1x penicillin-streptomycin. A549 cells were purchased from ATCC and were cultured in F12 medium (Corning) with 10% FBS and 1x penicillin-streptomycin. All cells were incubated at 37 °C with 5% CO_2.

TGF-β induce and inhibitor treatment

Cells for TGF-β induction and inhibitor treatment were seeded at ~60-70% confluence without serum starvation. For TGF-β treatment, 4 ng/ml human recombinant TGF-β1 (Cell signaling) was added into culture medium directly. For inhibition experiment, 4 μM of LY2109761 (Selleckchem), 20 μM of GANT61 (Selleckchem), and 10 μM of SB216763 (Selleckchem) were used to inhibit SMAD, GLI and GSK3, respectively. The medium was changed every day during treatment to keep the reagent concentration constantly. For reproducibility, we used cells within 10^{th} -15^{th} generations, same patches of reagents, serum, and tried to perform each group of experiments (e.g., those in Fig. 5C) together.

Immunofluorescence Microscopy and Data Analysis

Cells were seeded on four-well glass-bottom petri dishes at ~60% confluence overnight and treated with reagents (TGF-β1 and/or inhibitors). Three independent experiments were performed in every treatment. At designated time points, cells were harvested and stained with specific antibodies following procedure modified from the protocols at the Center of Biological Imaging (CBI) in the University of Pittsburgh. In general, cells were washed with DPBS for five minutes for three times followed by 4% formaldehyde fixation for 10 minutes at room temperature. Cells were then washed three times with PBS for five minutes every time. PBS with 0.1% TritonX-100 (PBS_Triton) was used for penetration. BSA of 2% in PBST was used for blocking before staining with antibodies. The first antibodies, anti-pSMAD2/3 (Santa Cruz), anit-SNAIL1 (Cell signaling), anti-GLI1 (Santa Cruz) were diluted by PBST with 1% BSA.

Samples were incubated with the first antibodies at 4°C overnight. Then cells were washed three times with 10 minutes for each before being incubated with the secondary antibodies, anti-mouse Alexa Fluor 647 (Abcam), anti-rabbit Alexa Fluor 647 (Abcam), anti-gaot Alexa Fluor 647 (Abcam), anti-mouse Alexa Fluor 555 (Abcam), anti-rabbit Alexa Fluor 555 (Abcam), anti-gaot Alexa Fluor 555 (Abcam), anti-mouse Alexa Fluor 488 (Abcam), anti-rabbit Alexa Fluor 488 (Abcam), anti-gaot Alexa Fluor 488 (Abcam), for one hour at room temperature. After antibody incubation, cells were washed with PBS_Triton for five minutes and stained with DAPI (Fisher) for 10 minutes at room temperature. Cells were washed with PBS_Triton for five minutes three times and stored in PBS for imaging. Photos were taken with Nikon A1 confocal microscopy at CBI. The microscope was controlled by the build-in software, Nikon NIS Elementary. All photos, except the photo for $GSK3^{AA}$ subcellular localization, were taken with plan fluor 40x DIC M/N2 oil objective with 0.75 numerical aperture and 0.72 mm working distance. The scan field were chosen randomly all over the glass-bottom area. For identifying the $GSK3^{AA}$ subcellular localization, plan apo λ 100x oil objective with 1.45 NA and 0.13 mm WD was used. The 3D model of $GSK3^{AA}$ overlapped with ERC and DAPI were reconstructed from 25 of z-stack images in 11.6 μm and videos were produced also by NIS Element software. To minimize photobleaching, an object field was firstly chosen by fast scan, then the photos were taken at 2014 × 2014 pixel or 4096 × 4096 pixel resolution, for generation large data or for photo presentation, respectively. CellProfiller was used for cell segregation and initial imaging analysis as what described in Carpenter et.al. [266].

Image Correction. To keep identical background through all images, background correction was performed before further image processing. For each image fluorescent intensities in space without cells were used as local background. Photos that have obviously uneven illumination and

background fluorescence were removed from further processing. Otherwise the mean background fluorescent intensity was obtained through averaging over the whole image, and was deducted uniformly from the image.

Image Segmentation. Cell number and position were determined by nuclear recognition with DAPI. The global strategy was used to identify the nuclear shape, and the Otsu algorithm was used for further calculation. Clumped objectives were identified by shape and divided by intensity. Next, using the shrank nuclear shape as seed, cell shape was identified by the Watershed algorithm. For identifying the clusters of $GSK3^{AA}$ formed around a nucleus, the nuclear shape was shrank manually by 3 pixels and used as a new seed to grown the boundary with the watershed method until reaching background intensity level. All parameters were optimized through an iterative process of automatic segmentation and manual inspection.

Image quantification. Averaged fluorescence density and integrated fluorescence intensity were calculated automatically with CellProfiler. The amount of the $GSK3^{AA}$ form was quantified as the sum of intensities of pixels belonging to the cluster formed around a nucleus. Concentrations of all other proteins were quantified by the average pixel intensity within the nucleus or cytosol region of a cell. Next, the quantified results were examined manually, and those cells with either cell area, nuclear area, or fluorescent intensity beyond five folds of the 95% confidence range of samples from a given treatment were discarded, which account for less than 1% of the cells analyzed. Immunofluorescence data were further processed and plots were generated using customized R codes and Matlab codes.

Quantitative PCR

Cells were seeded in 12-well plastic bottom cell culture plates and treated as described above. Three parallel experiments were performed in every treatment. Total RNA was isolated with the

TRIZOL RNA isolation kit (Fisher), and mRNA was reversely transcribed with the RNAscript II kit (ABI). The stem-loop method was used for microRNA reverse transcription. The qPCR system was prepared with the SYBR green qPCR kit with designed primers and performed on StepOnePlus real-time PCR (ABI).

Immunoprecipitation and silver staining

Immunoprecipitation was performed with SureBeads magnetic beads (Bio-Rad) following a protocol modified from the one provided by the manufacture. We washed beads with PBS with 0.1% Tween 20 (PBS_Tween) for three times, then harvested cells by RIPA (Thermo) with proteinase and phosphatase inhibitor (Roche). Samples were pre-cleaned with 100 µl of suspended Protein G per 450 µl of lysis mixture. Antibodies targeting GSK3 (Cell Signaling), $GSK3^{AA}$ (Santa Cruz), and $GSK3^{D}$ (Santa Cruz) were added into every 100 µl of bead mixture respectively. The mixture was rotated at 4 °C for 3 hours. Beads that were conjugated with antibodies were washed with PBS_Tween. An amount of 100 µl of pre-cleaned lysis buffer was added into conjugated beads and rotated at 4 °C overnight. Targeted proteins were eluded from beads by incubating with 40 µl 1x Laemmli buffer with SDS at 70 °C for 10 minutes. For the samples an amount of 5 µl was used for western blot assay, and an amount of 30 µl was loaded for SDS-PAGE (Bio-Rad) and followed by silver staining (Fisher).

Network reconstruction and coarse-graining

The full network from TGF-β1 to SNAIL1 (Supplementary Fig. 10) was generated with IPA (Qiagen®). Specifically, all downstream regulators of TGF-β1 and upstream regulators of SNAIL1 in human, mice and rat were searched and added to the network. Then, direct or indirect relationships between every pair of regulators were searched and added to the network. After obtaining the whole network, regulators that have been reported to be activated later than

SNAIL1 were removed. Examination of the network reveals that the network can be further organized into three groups: the TGF-β-SMAD-SNAIL canonical pathway, the TGF-β-GSK3-β-catenin pathway that has the most number of links, and others. We further noticed that GLI1 is a central connector of TGF-β, SMAD, GSK3 and SNAIL1. We performed western blot and IF studies on β-CATENIN and found that neither its concentration nor its location changes significantly on day 3, therefore we removed β-CATENIN from the network. In addition, previous studies report that the SMAD-GLI axis plays important role in TGF-β induced EMT [122]. Therefore we further grouped the network as the SMAD module, the GLI module, and the GSK3 module, as well as the remaining ones that we referred as "Others", and reached the network shown in Fig. 5A. Those molecular species not explicitly specified in Fig. 5A either have their effects implicitly included in the links, for example the link from TGF-β to GSK3, or are included in the links of "Others". This treatment is justified since our various inhibition experiments indeed showed that the three factors we identified affect SNAIL1 expression the most. These "other" species may contribute to snail1 activation at a time later than what considered in this work. Therefore we emphasize the network in Fig. 5A is valid only within the time window we examined, i.e., within three days after TGF-β1 treatment for MCF10A cells.

2.5 SUPPLEMENTARY MATERIALS AND FIGURES

2.5.1 Mathematical modeling

Canonical TGF-β/pSMAD2/3/SNAILl pathway (Fig. 5A)

We used the following ordinary differential equation (ODE) model in Fig. 5 and Fig. 11.

TGF-β/SMAD2/3 module

$$[Smad]' = (kp_{smad0} + kp_{smad} * TGF) * \frac{Smad_{all} - [Smad]}{Jp_{smad0} + (Smad_{all} - [Smad])} \frac{1}{1 + \frac{Smad_I}{Jdp_{smad1}}} - dp_{smad} * \frac{[Smad]}{Jdp_{smad} + [Smad]}, \quad (1)$$

$$[Smad_I]' = k_{SmadI} * [Smad] - kd_{SmadI} * [Smad_I], \quad (2)$$

where [Smad] and [$Smad_I$] are the concentrations of pSMAD2/3 and inhibitory SMAD, respectively.

SNAIL-miR-34 module

It is expanded from our previous model[13] by considering transcription activation of SNAIL1 by pSMAD2/3 and TGF-β, and degradation of SNAIL1.

$$[snail]'_n = k0_{snail} + k_{snail0} * \frac{[Smad]^2}{J^2_{snail0} + [Smad]^2} \frac{1}{1 + \frac{[SNAIL]}{J_{snail2}}} - kd_{snail} * [snail] - kd_{SR1} * [SR1], \quad (3)$$

$$[miR34]'_n = k0_{34} + \frac{k_{34}}{1 + \left(\frac{SNAIL}{J1_{34}}\right)^2} - kd_{34} * [miR34] - (1 - \lambda_s) * kd_{SR1} * [SR1], \quad (4)$$

$$[SNAIL]' = k_{SNAIL} * [snail] - kd_{SNAIL} * [SNAIL], \quad (5)$$

$$[miR34] = [miR34]_t - [SR1], \quad (6)$$

$$[snail] = [snail]_t - [SR1], \quad (7)$$

$$[SR1] = Ks * [snail] * [miR34], \quad (8)$$

where [snail], [miR34], [SNAIL], [snail], [SR1] are the concentrations of total *SNAIL1* mRNA, miR-34, SNAIL1 protein, free *SNAIL1* mRNA and miR-34-*SNAIL1* mRNA complex, respectively.

Canonical TGF-β/SMAD/SNAILl pathway with GLI1 (Fig. 6, Fig. 12)

Taking into account the GLI1 self-activation and GLI1 mediated expression of *SNAIL1* mRNA, we added another ODE for GLI1 and revised the ODE of *SNAIL1* mRNA.

$$[GLI]_n' = k_{gli0} + k_{gli1} * \frac{[Smad]^2}{J_{gli1}^2 + [Smad]^2} + k_{gli2} * \frac{[GLI]_n{}^4}{[GLI]n^4 + J_{gli2}^4} - d_{gli} * [GLI]_n \quad (9)$$

$$[snail]_t' = k0_{snail} + \left(k_{snail0} * \frac{[Smad]^2}{J_{snail0}^2 + [Smad]^2} + k_{snail1} * \frac{[GLI]_n{}^4}{[GLI]_n{}^4 + J_{snail1}^4} \right) \frac{1}{1 + \frac{[SNAIL]}{J_{snail2}}} - kd_{snail} * [snail] - kd_{SR1} * [SR1] \quad (10)$$

where $[GLI]_n$ is the concentration of nuclear GLI1. We used this ODE model to generate results in Fig. 6 and Fig. 12.

Model for the GSK3/GLI module (Fig. 8A)

Since the process involves many steps and a detailed model would require many parameters to determine, instead we used two phenomenological time-dependent functions to qualitatively mimic the dynamics of the enzyme activities of cytosol GSK3 and nuclear GSK3 we experimentally measured (shown in Fig. 7B),

$$A_{GSK}^{C}(t) = k_{GSKc} * TGF * \left(1 - \exp\left(-\frac{t}{a1}\right)\right) * \exp\left(-\frac{t-b1}{a1}\right) \quad (11)$$

$$A_{GSK}^{n}(t) = 1 - k_{GSKn} * TGF * \left(1 - \exp\left(-\frac{t}{a2}\right)\right) * \left(\exp\left(-\frac{t-b2}{a2}\right)\right) \quad (12)$$

Figure S13E shows the relative enzymatic activity. We played with different choices of the parameters, and found the results are insensitive to the choices provided there is an early pulsate increase of cytosol GSK3 enzymatic activity followed by a decrease of nuclear GSK3 enzymatic activity.

Furthermore, the basal pool of cytosol GLI1 is considered, which is by sequestered in the cytosol by SUFU but could translocate to the nuclear after SUFU is inactivated by the cytosol enzyme GSK3 activity. We used a revised ODE of nuclear GLI1 concentration derived with the quasi-equilibrium approximation (see below)

$$[GLI]_t' \approx k_{gli0}+k_{gli1}*\frac{[Smad]^2}{J_{gli1}^2+[Smad]^2} + k_{gli2}*\frac{[GLI]_n^4}{[GLI]_n^4+J_{gli2}^4} - d_{gli}*[GLI]_n*A_{GSK}^n \quad (13)$$

Derivation of GLI ODE

We assumed the quasi-equilibrium approximation for the GLI nuclear and cytosol shuttling, the GSK3 regulated binding/unbinding between Sufu and GLI in the cytosol, and obtained the following equations,

$$K2*[GLI]_c*[Sufu]=(K1+A_{GSK}^C)*[GLIsufu] \quad (14)$$

$$[GLIsufu]=Sufu_{max}-[Sufu] \quad (15)$$

Thus we have

$$[GLIsufu]=Sufu_{max}-(K1+A_{GSK}^C)*\frac{[GLIsufu]}{K2*[GLI]_c} \quad (16)$$

That is,

$$[GLIsufu]=\frac{1}{1+\frac{(K1+A_{GSK}^C)}{K2*[GLI]_c}}Sufu_{max} \quad (17)$$

Also we have

$$[GLI]_c = K3*[GLI]_n, \tag{18}$$

thus

$$[GLIsufu] = \frac{K3*K2*[GLI]_n}{K3*K2*[GLI]_n+(K1+A_{GSK}^C)} Sufu_{max} \tag{19}$$

The total level of GLI1 is the sum of the three forms, GLISuFu, GLIc and GLIn,

$$[GLI]_t = [GLI]_c+[GLI]_n+[GLIsufu] = \left(K3*[GLI]_n+[GLI]_n+\frac{K2*K3*A_{GSK}^n}{K3*K2*[GLI]_n+(K1+A_{GSK}^C)} Sufu_{max}\right). \tag{20}$$

Thus, we obtained the relation among $[GLI]_n$, $[GLI]_t$ and A_{GSK}^C

$$[GLI]_n = f\left(A_{GSK}^C, [GLI]_t\right), \tag{21}$$

The total concentration of GLI1 is given by,

$$[GLI]_t' = k_{gli0}+k_{gli1}*\frac{[Smad]^2}{J_{gli1}^2+[Smad]^2} + k_{gli2}*\frac{[GLI]_n{}^4}{[GLI]_n{}^4+J_{gli2}^4} - d_{gli}*[GLI]_c*A_{GSK}^C - d_{gli}*[GLI]_n*A_{GSK}^n \tag{22}$$

Given that our data shows that $[GLI]_c$ is low throughout the process, we neglected the degradation term of $[GLI]_c$

$$[GLI]_t' \approx k_{gli0}+k_{gli1}*\frac{[Smad]^2}{J_{gli1}^2+[Smad]^2} + k_{gli2}*\frac{[GLI]_n{}^4}{[GLI]_n{}^4+J_{gli2}^4} - d_{gli}*[GLI]_n*A_{GSK}^n \tag{23}$$

<u>TGF-β pulse</u>

Since TGF-β1 can enter to cells through endocytosis, washing the extracellular TGF-β1 does not stop the signaling immediately. Therefore, we modeled the effective TGF-β1 concentration by the following equation,

$$[TGF](t) = TGF_0*\exp\left(-d_{tgf}*\left(t-TGF_{Duration}\right)*Heaviside\left(t-TGF_{Duration}\right) \right). \tag{24}$$

Full model

By considering all the modules, the full model is as following,

$$[Smad]' = \left(kp_{smad0} + kp_{smad} * [TGF]\right) * \frac{Smad_{all} - [Smad]}{Jp_{smad0} + \left(Smad_{all} - [Smad]\right)} \frac{1}{1 + \frac{[Smad_I]}{Jdp_{smad1}}} - dp_{smad} * \frac{[Smad]}{Jdp_{smad} + [Smad]}$$

(25)

$$[Smad_I]' = k_{SmadI} * [Smad] - kd_{SmadI} * [Smad_I] \quad (26)$$

$$[GLI]_t' = k_{gli0} + k_{gli1} * \frac{[Smad]^2}{J_{gli1}^2 + [Smad]^2} + k_{gli2} * \frac{[GLI]_n^4}{[GLI]_n^4 + J_{gli2}^4} - d_{gli} * [GLI]_c * A_{GSK}^C - d_{gli} * [GLI]_n * A_{GSK}^n \quad (27)$$

$$[snail]_t' = k0_{snail} + \left(k_{snail0} * \frac{[Smad]^2}{J_{snail0}^2 + [Smad]^2} + k_{snail1} * \frac{[GLI]_n^4}{[GLI]_n^4 + J_{snail1}^4}\right) \frac{1}{1 + \frac{[SNAIL]}{J_{snail2}}} - kd_{snail} * [snail] - kd_{SR1} * [SR1]$$

(28)

$$[miR34]_n' = k0_{34} + \frac{k_{34}}{1 + \left(\frac{SNAIL}{J1_{34}}\right)^2} - kd_{34} * [miR34] - \left(1 - \lambda_s\right) * kd_{SR1} * [SR1] \quad (29)$$

$$[SNAIL]' = k_{SNAIL} * [snail] - kd_{SNAIL} * [SNAIL] * A_{GSK}^n \quad (30)$$

$$A_{GSK}^C(t) = k_{GSKc} * [TGF] * \left(1 - \exp\left(-\frac{t}{a1}\right)\right) * \exp\left(-\frac{t - b1}{a1}\right) \quad (31)$$

$$A_{GSK}^n(t) = 1 - k_{GSKn} * [TGF] * \left(1 - \exp\left(-\frac{t}{a2}\right)\right) * \left(\exp\left(-\frac{t - b2}{a2}\right)\right), \quad (32)$$

$$[GLI]_n = f([GSK]_c, [GLI]_t) \quad (33)$$

$$[miR34] = [miR34]_t - [SR1] \quad (34)$$

$$[snail] = [snail]_t - [SR1] \quad (35)$$

$$[SR1] = Ks * [snail] * [miR34] \quad (36)$$

We used this ODE model to generate results in Fig. 9A and Supplementary Fig. 15B. In the above equations we chose a Hill coefficient of 2 for Smad and SNAIL1 based on their dimeric binding. We used a value of 4 for GLI1 for sufficient nonlinearity. To keep the model consistent, we used 4 in all our equations regard to GLI1.

Parameter space searching

Step 1: Calculate single cell distributions of experimental observables. We calculated histograms of the distributions from the single cell experimental data. Suppose that we have *N* observables measured in *M* time points, we have an $N \times M$ dimensional distribution of the data. Since we used fixed cells and we had no information on the temporal correlation, we treated the distributions from different time points as independent, *i.e.*, $P=\prod_{i=1}^{M} P_i$.

Step 2: Define pseudo-potentials from the parameterized distribution. We defined a pseudo-scalar-potential function $U(x_1,x_2,\ldots,x_M)=-T_{eff}(\ln P-\ln P_{max})$. The constant T_{eff} is an effective temperature, which we chose $T_{eff}=1$. The constant term $\ln P_{max}$ sets the potential to be zero at the peak position of the distribution, and does not affect the parameter space search results. This pseudo-potential is just an auxiliary scalar function for the following application of the Metropolis algorithm. If a mathematical model can faithfully describe the system dynamics, with given initial conditionals and non-adjustable parameter set of ζ, we should be able to find distributions of the parameter set λ (to take into account cell-to-cell heterogeneity), and generate the corresponding distributions of (x_1, x_2, … x_M) reproduce *U*. That is, for a specific set of λ, $x_i=x_i(x_0; \lambda, \zeta)$, i=1,…M, and $U(x_1,x_2,\ldots,x_M)\equiv V(\lambda)$. Unlike *U*, the function form of *V* can be very complex, but fortunately we do not need to know its explicit function form to perform the following Metropolis sampling.

Step 3: Obtain model parameter distributions that reproduce the distributions of experimental observables. Now it is clear why we define the pseudo-potential. We performed Monte Carlo random walks along the pseudo-potential V in the λ space using the Metropolis algorithm, just as how the algorithm is typically applied along real physical potentials. At each step with a set of λ, we generated a trial move $\lambda' = \lambda + \delta\lambda$. We propagated the ODEs to obtain $V(\lambda)$ and $V(\lambda')$, then use the Metropolis criteria to decide whether to accept the new move. If $V(\lambda') \leq V(\lambda)$ accpet this step and update the parameter set $\lambda = \lambda'$. If $V(\lambda') \geq V(\lambda)$, accept this step with a probability $\exp(-(V(\lambda')-V(\lambda))/T)$ with $T = 1$.

In our model, there is no feedback between the SMAD2/3 module and the SNAIL1/miR-34 module, thus we used a two-step to search the parameter space for the TGF-β/SMAD2/3 module,

1. Search the parameter space (nine parameters) in the SMAD2/3 module;
2. Search the parameter space (six parameters) for the SNAIL1/miR-34 module based on the 50 samples of good-fit parameter set of the SMAD2/3 module from step 1.

In step 2 some of the parameters in the SNAIL1/miR-34 module were fixed and used as a well-trained parameter set from our previous work [13]. Instead only six new parameters that connect the module SMAD2/3 and module SNAIL1/miR-34 were considered in the parameter space searching.

When the GLI1 module was included, we again used the fact that there is no feedback between the SMAD2/3 module and the GLI1 module, and used a three-step searching procedure to reduce the computational efforts,

1. Search the parameter space (nine parameters) for the SMAD2/3 module;

2. Search the parameter space (seven parameters) for the GLI1 module;

3. Search the parameter space (six parameters) for the SNAIL1/miR-34 module based on the 50 samples of good-fit parameter set of the SMAD2/3 module the GLI1 module from step 1-2.

Parameter change in various over-expression/down-expression or over-active/down-active conditions (Supplementary Fig. 15)

To produce the results in Fig. 15B, a 1.2-fold change of k_{gli0} is used in the case of GLI1 over-expression, a 0.8-fold change of k_{smadi} in the case of I-SMAD down-regulation. There is 0.8-fold change of k_{gskn} in the case of over-active cytosol GSK3, 1.2-fold change of k_{gskc} in the case of under-active cytosol GSK3. Similarly, there is 1.2-fold change of k_{gskc} in the case of over-active nuclear GSK3, and 0.5-fold change of k_{gskc} in the case of under-active nuclear GSK3.

2.5.2 Supplementary figures

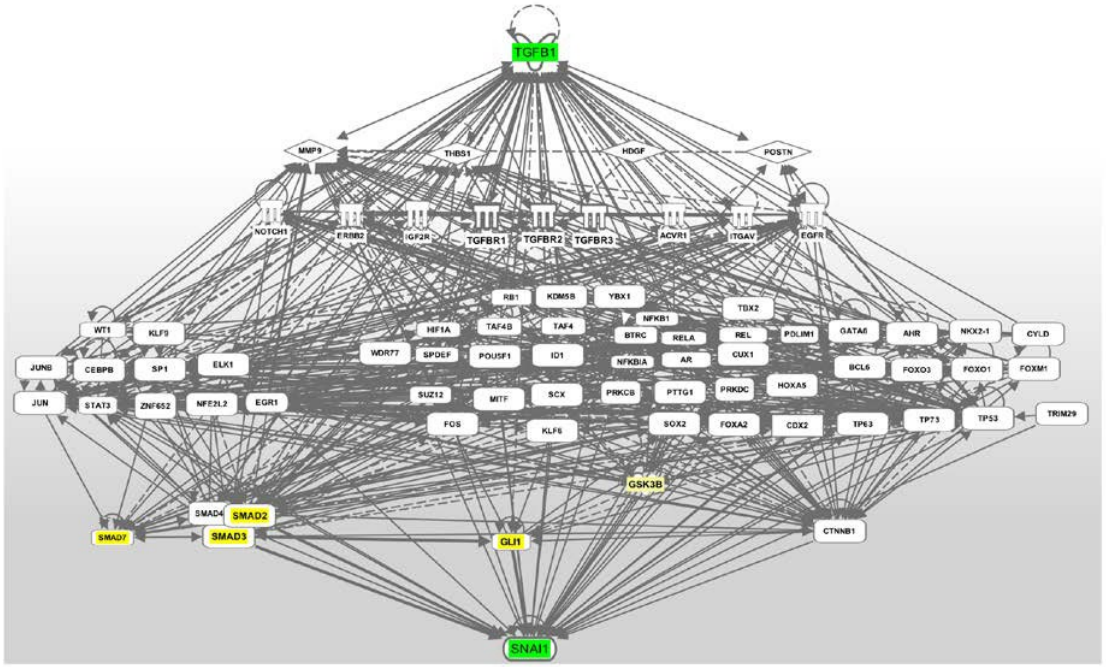

Figure 10 Network of TGF-β activating SNAIL1 reconstructed with IPA.

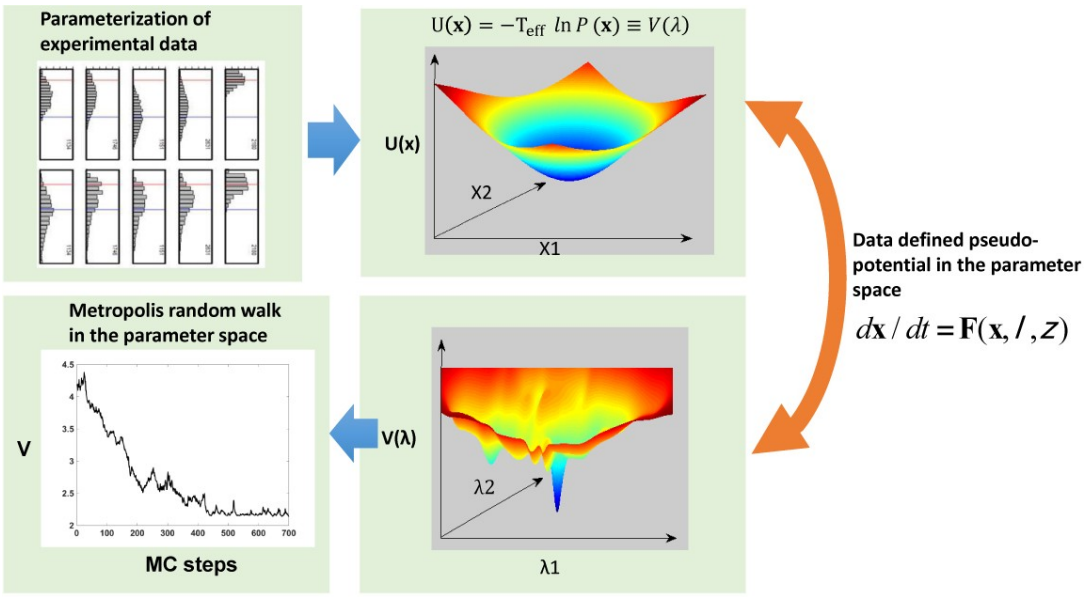

Figure 11 Schematic of the parameter space search approach.1) Calculate single cell distributions of experimental observables. 2) Define pseudo-potentials from the parameterized distribution. 3) Obtain model parameter distributions that reproduce the distributions of experimental observables. (See parameter space searching in the SI for more detail).

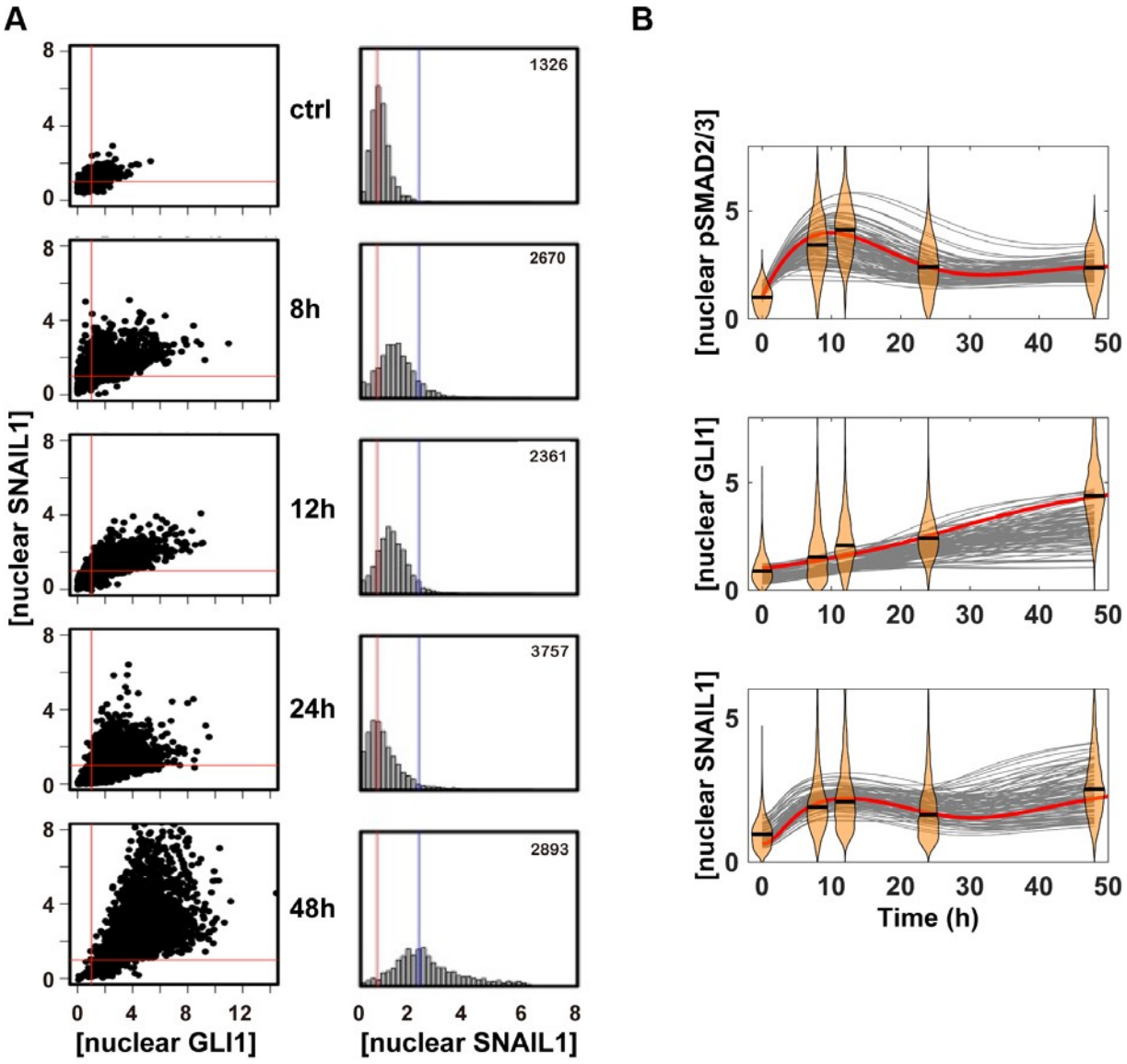

Figure 12 Supplemental results showing GLI1 contributes to the second wave of SNAIL1. **(A)** Scatted plot of measured nuclear GLI1 and SNAIL1 concentrations and the corresponding histogram representation for [nuclear SNAIL1]. The same sets of data of Fig. 6C are used. **(B)** The model of Fig. 5A with GLI1 reproduces the observed pSMAD2/3-SNAIL1 dynamics. To fit the SNAIL1 dynamics the exact temporal profile of GLI1 is not important except the requirement of its activation after 24 h.

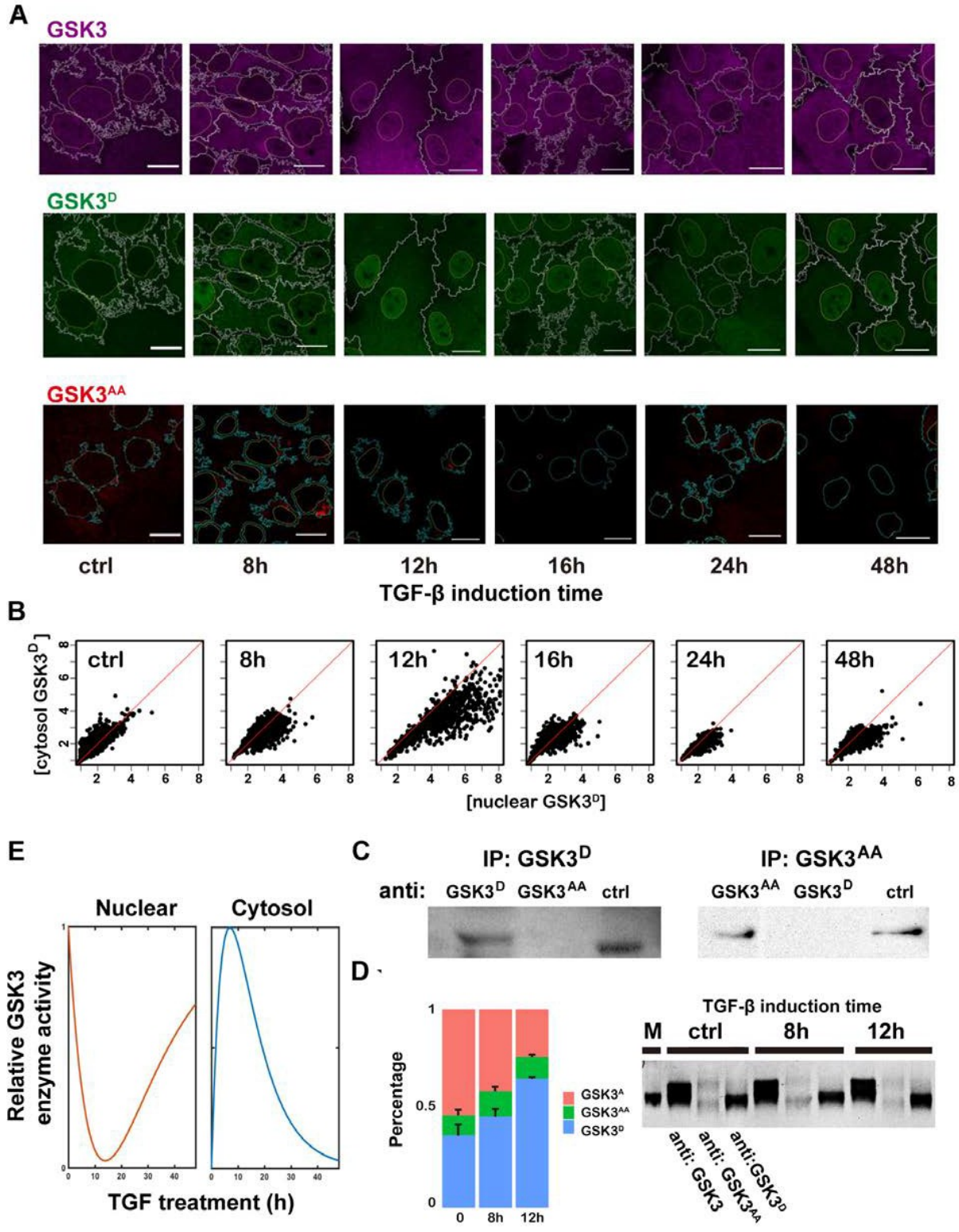
A
GSK3
GSK3D
GSK3AA
ctrl
8h
12h
16h
24h
48h
TGF-β induction time
B
[cytosol GSK3D]
[nuclear GSK3D]
E
Nuclear
Cytosol
Relative GSK3 enzyme activity
TGF treatment (h)
C
IP: GSK3D
IP: GSK3AA
anti:
D
Percentage
GSK3A
GSK3AA
GSK3D
TGF-β induction time
M
anti: GSK3
anti: GSK3AA
anti:GSK3D

Figure 13 Supplemental results showing temporal switch between two phosphorylation forms of GSK3. **(A)** IF images showing the temporal switch between two phosphorylation forms of GSK3. **(B)** Scattered plots showing correlation between nuclear and cytosol concentrations of $GSK3^{D}$. **(C)** Immunoprecipitation studies showing two phosphorylation forms do not coexist. MCF10A cells were treated with TGF-β for 8 hours and the total proteins were harvested by RIPA. $GSK3^{D}$ and $GSK3^{AA}$ antibody were used for immunoprecipitation, respectively. $GSK3^{D}$ and $GSK3^{AA}$ proteins were used for western blot. Ctrl is sample that did not undergo immunoprecipitation. **(D)** Silver staining measurement of the relative amount of different GSK3 forms. The right figure shows a representative of three independent replicates. M refers to the marker with mass as 50 kd. **(E)** Relative GSK3 enzymatic activities in the cytosol and nucleus during TGF-β treatment.

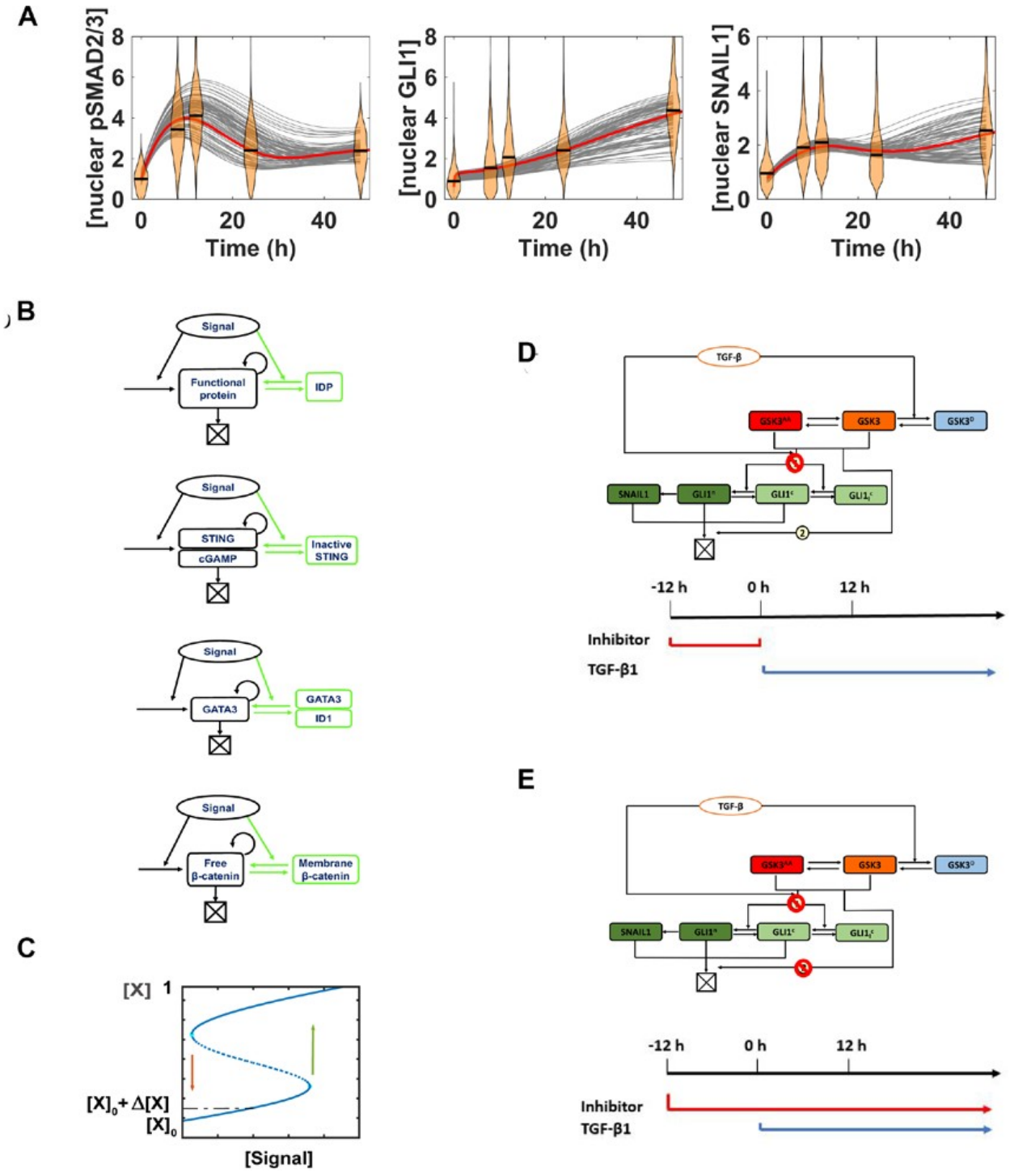

Figure 14 Supplemental results of the full model. **(A)** The model of Fig. 8A reproduces the observed GLI1 as well as pSMAD2/3-SNAIL1 dynamics. **(B)** Examples of regulatory factors having positive feedback loop and reservoir of molecules in inactive form that can be activated by another stimulus. IDPs refer to intrinsically disordered

proteins, and some of them are transcription factors, which change into folded form and have higher DNA binding affinity upon binding of cofactors or posttranslational modification. ID1 is a member of the family of inhibitors of DNA binding proteins. **(C)** Bifurcation diagram showing that the initial concentration boost is small compared to the concentration jump associated with external signal induced switch of cell states. **(D-E)** Schematics of the early and full GSK3 inhibition experiments.

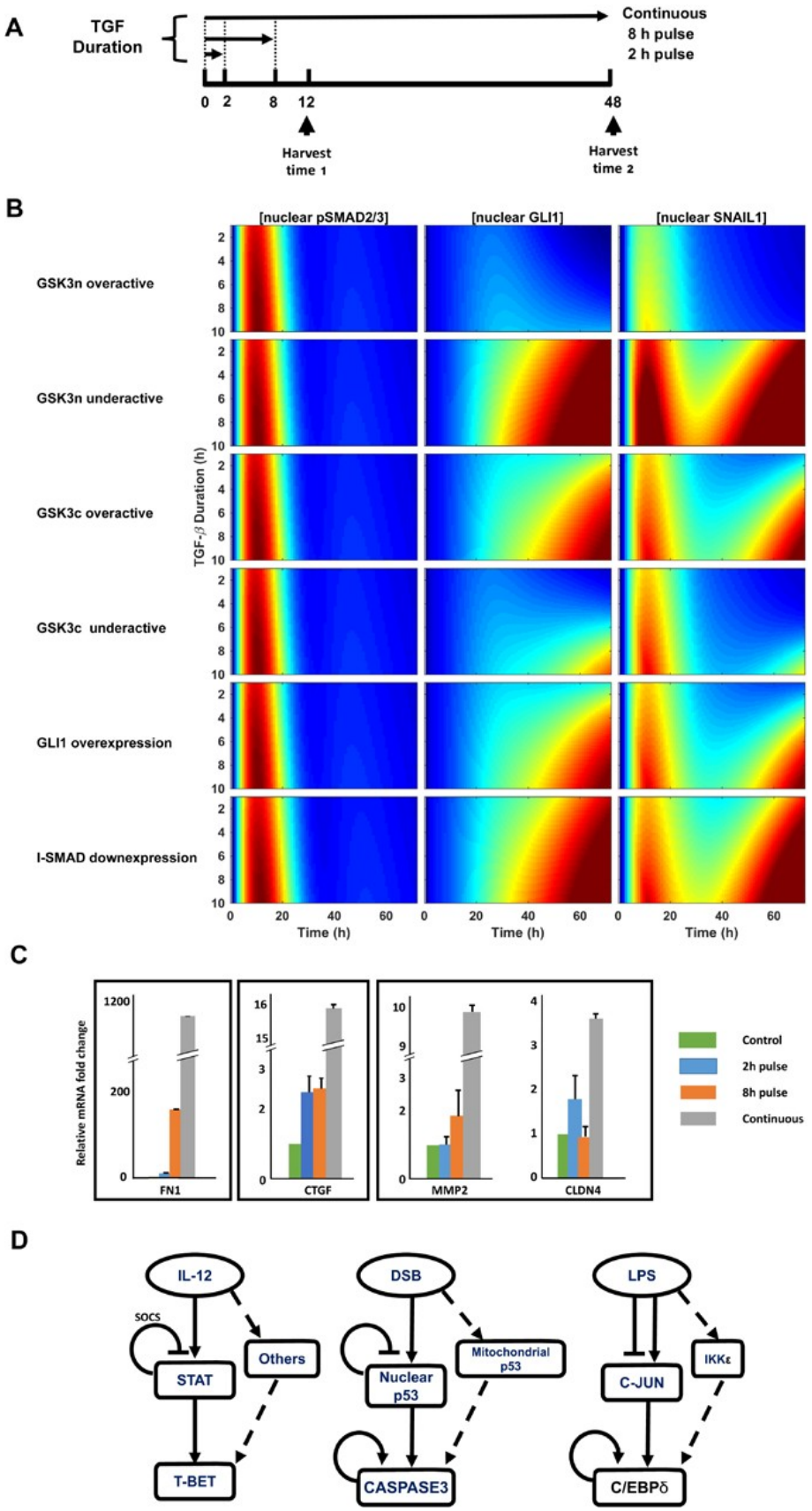
A
TGF Duration
Continuous
8 h pulse
2 h pulse
0 2 8 12 48
Harvest time 1
Harvest time 2
B
[nuclear pSMAD2/3]
[nuclear GLI1]
[nuclear SNAIL1]
GSK3n overactive
GSK3n underactive
GSK3c overactive
GSK3c underactive
GLI1 overexpression
I-SMAD downexpression
TGF-β Duration (h)
Time (h)
C
Relative mRNA fold change
FN1
CTGF
MMP2
CLDN4
Control
2h pulse
8h pulse
Continuous
D
IL-12
SOCS
STAT
Others
T-BET
DSB
Nuclear p53
Mitochondrial p53
CASPASE3
LPS
C-JUN
IKKε
C/EBPδ

Figure 15 Supplemental results that cells can detect signal duration. **(A)** Schematic of TGF-β pulse experiments for Fig. 9B and 6C. **(B)** Supplemental model results of pulsed TGF-β1 treatments with various mutations. **(C)** The mRNA levels of selected TGF-β activated genes at day 3 after different durations of TGF-β1 treatments. **(D)** Examples of other signaling transduction pathways that share similar motif structure as that of TGF-β signaling, including IL-12, DNA double strand breaking, and LPS.

3.0 SPATIAL CLUSTERING AND COMMON REGULATORY ELEMENTS CORRELATE WITH TGF-B INDUCED CONCERTED GENE EXPRESSION

Cellular responses to surrounding cues require temporally concerted transcriptional regulation of multiple genes. A single-input-module motif with one transcription factor regulating multiple target genes can generate the coordinated gene expression. In eukaryotic cells, transcriptional activity of a gene is affected by not only transcription factors but also the ambient DNA condition of the gene, characterized by epigenetic modifications and the three-dimensional chromosome structure. To examine how the local gene environment and transcriptional factor regulation are coupled, we performed a combined analysis of time-course RNA-seq data of TGF-β treated MCF10A cells and corresponding epigenome and Hi-C data. Using the Dynamic Regulatory Events Miner (DREM), we first clustered differentially expressed genes based on gene expression profiles and associated transcription factors. Next we defined a set of linear and radial distribution functions as used in statistical physics to measure distributions of the genes within a cluster both along the genome sequence and spatially. Remarkably genes within each cluster, i.e., having similar temporal gene expression pattern and sharing common upstream transcription factors, have significantly higher tendency to be spatially close in the three-dimensional structure, compared to those belonging to different clusters. Specifically, we identified local spatial organization of over a hundred of AP1-regulated genes. We propose that competition between hetero- and homo-dimeric AP1 fine-tunes the chromosome structure and

leads to differential expression of target genes. Overall our computational results are consistent with a model that transcriptional factors actively orchestrate spatial clustering of a group of genes as well as fine structure of individual genomic loci, and future studies can test whether this spatial chromosome organization contributes to concerted gene expression.

3.1 INTRODUCTION

A cell continuously receives signals from its local extracellular environment and adjusts cellular programs accordingly such as cell proliferation, motility and metabolism[22]. Typically, regulating a cellular process requires expression change of a group of genes in a temporally coordinated manner[267]. It is an important question how such a coordinated regulation is achieved.

A mechanism of such regulation is through specific structures of interaction networks of transcription factors (TFs). TFs bind to specific DNA sites and regulate transcription activities of targeted genes. One TF can regulate multiple target genes to form a so-called single input module (SIM, or fan- out). This SIM network motif appears in high frequency to coordinate the expression of genes with related functions such as those in bacterial metabolic pathways[268]. Gene regulation in eukaryotic cells is more complex since the three-dimensional structure of DNA has more profound impact on gene transcription than that in prokaryotic cells. For instance, the nucleosome structure with high packing level limits gene accessibility[269]. Furthermore, epigenetic modifications can strongly influence gene transcription. It is not fully understood how these different regulation mechanisms may collectively regulate a group of related genes.

To investigate the coupling of gene regulation mechanisms at different levels, here we used transcriptional response to TGF-β as a model system. The TGF-β family is crucial for regulating a complex signal transduction network in embryonic and fetal development, and is involved in multiple physiological and pathological processes such as cancer progression and wound healing[75]. The signaling event starts from membrane embedded TGF-β receptors, which capture active TGF-β molecules from extracellular environment[270]. The TGF-β signal is then transmitted into cells through a signal transduction network and triggers a cascade of cellular responses. The latter is achieved through temporally coordinated expression changes of a groups of genes with related functions such as those involved in cell proliferation, metabolism, and motility[271]. TGF-β also induces global reprogramming of cell epigenome[272], which reinforces cellular responses for committed cell phenotype transition.

A main aim of the present work is to analyze how multiple levels of regulation lead to concerted expression of groups of genes. For this purpose, we analyzed the temporal gene expression profiles of TGF-β1 treated human MCF10A cells in the context of histone modification patterns and chromosome structures derived from Hi-C data. Our analyses reveal that genes co-regulated by a common transcription factor have tendency to be spatially close, even if they are not linearly close on a chromosome. The results suggest correlation among transcription factors, chromosome structure, and gene expression that requires further study.

3.2 RESULTS

3.2.1 Gene expression change reflects cell phenotype transition in response to TGF-β

We used human mammary MCF10A cells, a non-tumorigenic breast epithelial cell line, as the major *in vitro* model in this study. Cells were induced with 4 ng/ml TGF-β1 for 12 hours, 2, 3, 5, 8, 12, and 21 days. Untreated MCF10A cells show typical epithelial morphology with tight cell-to-cell adherence. With TGF-β1 treatment, we observed progressive morphological changes, indicating cell transformation from the epithelial to mesenchymal phenotype. From day 2 to day 5, cells start to show loosened intercellular adherence. After day 5, some cells expand cell size and show polarity. With further TGF-β treatment, more cells acquire a spindle-like shape. On day 21, only a small fraction of cells still remain morphologically unchanged (Fig. 20A) and most cells have undergone epithelial-to-mesenchymal transition (EMT).

Next we performed RNA-seq studies to uncover changes of gene expression accompanying EMT. At each time point, we harvested samples and extracted RNA for RNA-seq analysis. The RNA-seq results reveal that about 33% human genes were differentially expressed (DE) upon TGF-β treatment. Principal component analysis (PCA) over these ~7000 DE genes show that gene expression profiles of samples from different time points are better separated than those of replicate samples from the same time point (Fig. 20B), as expected. The global transcriptome change over time as seen in the PCA space is consistent with the gradual morphology change of cells over time and the previous report that TGF-β-induced EMT proceeds through intermediate states[13].

3.2.2 Genes classes based on both expression pattern and upstream regulators have more common characters

To further examine the temporal patterns and functions of the DE genes, we performed hierarchical clustering analysis, which divides the DE genes into seven classes based on their expression level changes (16A)[273]. Expressions of ~1,700 genes decrease, and those of ~2,000 genes increase with time. Another two classes show transient up and transient down dynamics, respectively. The remaining three classes have dynamic features similar to those of the transient up and transient down classes, but with temporal shift. Gene ontological (GO) analysis (Fig. 21) reveals that genes in each class typically are involved in multiple cellular processes. For example, genes in the decreasing class are related RNA polymerase I activity and snoRNA binding. These two classes of genes are related to RNA metabolic process, including ribosomal RNA production, modification, and binding to regulatory factors. This observation is consistent with previous reports that under TGF-β treatment cells are under growth arrest until they finish EMT[274].

Histone modifications affect gene expression[275]. To investigate the relationship between histone modification and gene expression pattern, we integrated H3K4me3 and H3K4ac data based on genome- wide ChIP-seq analysis from Messier et.al.[276] to the RNA-seq data. Both H3K4me3 and H3K4ac are histone modification marks that are associated with active or poised genes[277]. H3K4ac enriches on active genes at both the early and late stages of cancer progression and H3K4me3 enriches only on genes related to late cancer stages[276]. We used the distributions of H3K4me3 and H3K4ac modification on all human genes as a control, and examined the marks in each clustering gene class. The results in 16B show that all gene classes have elevated H3K4me3 and H3K4ac compared to the control, and there is no apparent difference on histone

modification patterns between different classes. Each gene class also has a broad and even bimodal distribution. That is, genes within a hierarchical clustering class do not share common histone modification patterns. Given that histone modification patterns correlate with local chromosome structures[179], these results suggest that genes from the same clustering class have heterogeneous local chromosome environments.

Next, we adopted a different clustering scheme, the Dynamic Regulatory Events Miner (DREM), which cluster genes by combining gene expression time series with additional pre-established transcription networks[278]. Figure 2A shows clustering results analyzed with DREM2 based on a Hidden Markov Model (HMM)[279]. At each conjunction node, genes are assigned to different branches based on their expression trend and the up-stream regulators (transcription factors on this node). Genes from an upstream branch can become key regulators at subsequent nodes[278,279]. It reveals a hierarchy of gene regulation during the process of TGF-β-induced phenotype change. With DREM2 the DE genes are clustered prominently into 46 branches with 19 nodes at the conjunction sites and 25 end classes. For clarity, we call the latter HMM classes to distinguish them from the expression-only clustering classes.

Compared to the hierarchical clustering classes, HMM classes show finer dynamic patterns and GO enrichment information (Table 2). For example, genes in the first seven classes all have increased expression, but differ in the detailed temporal profiles. Genes in Class C1 increase to high levels already on day 2. Genes relating to metalloendopeptidase activity are enriched in this class by over 17 folds in respect to the reference genes. Four of the matrix metalloproteinases (*mmp*s), *mmp2*/*7*/*11*/*13*, are in this class. These four MMPs degrade components of extracellular matrix proteins such as gelatin, fibronectin, and laminin, and mediate biological activities including migration, mammary epithelial cell apoptosis, and

EMT[280]. Heparin binding genes are another type of highly enriched genes. These genes, such as *periostin* (*postn*), *fibronectin* (*fn1*), are also related to matrix or cell membrane formation and thus affect cell migration and adhesion. Another class of early activation genes, Class C2, is also enriched with genes related to cell matrix and membrane structure. Among them five of *pcdh* family members, including *pcdh7*/*a4*/*b9*/*b10*/*b13*, are integral membrane proteins and have function as cell-cell recognition and adhesion. Genes within each HMM class show narrower distributions of histone modification patterns (17B) than those of the expression-only classes (16B), meaning those genes have more similar patterns of these histone marks. Therefore, DREM2 analysis reveals that genes with related functions are often regulated by common TFs, and have similar dynamic profiles.

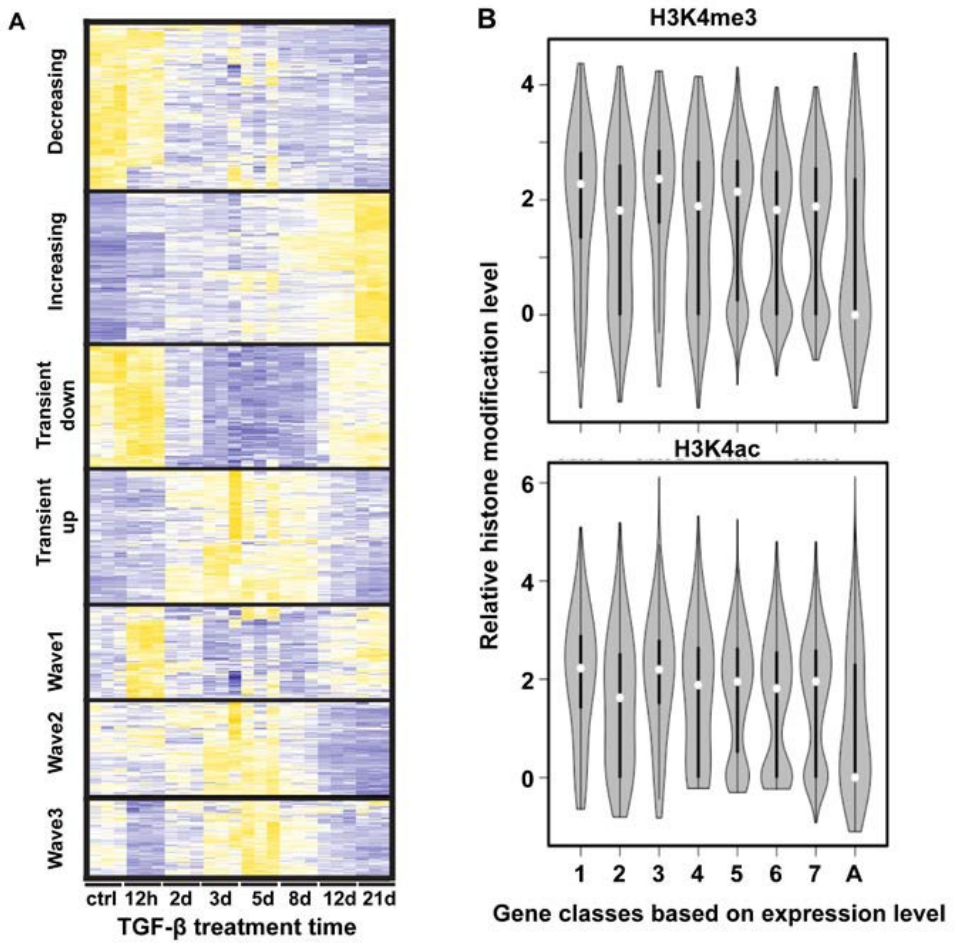

Figure 16 TGF-β induced gene expression change shows distinct temporal patterns. **(A)** Hierarchical-clusters of genes based on gene expression patterns during TGF-β treatment. **(B)** Violin plots of indicated histone modification level distribution sampled through genes belonging to individual expression pattern clusters. The numbers 1-7 on the x-axis follow the order of classes in panel A. The control group 'A' is sampled through all genes.

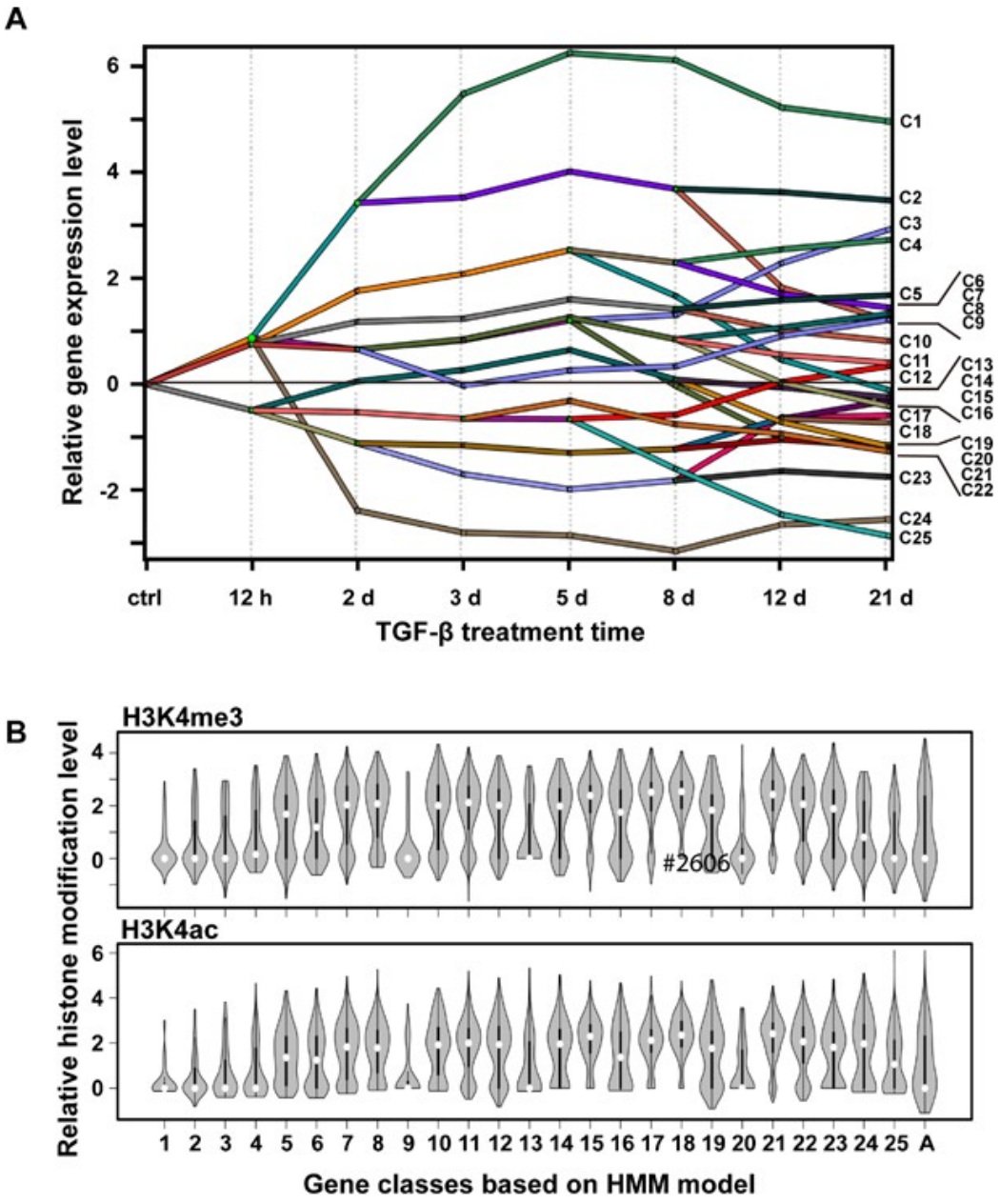

Figure 17 Genes clustered by both expression patterns and key transcription factors show correlation between patterns of expression and histone modification. **(A)** Dynamic regulatory map obtained through DREM2 analysis. **(B)** Violin plots of indicated histone modification level distribution sampled through genes belonging to individual DREM2 clusters.

3.2.3 Genes sharing common regulators have higher probability to be spatially close

As mentioned in Introduction, local DNA environment of a gene affects its transcriptional. Then it is natural to ask whether genes sharing similar expression patterns and common regulatory factors, as identified by DREM2 analysis, may be arranged to be spatially close to share similar local DNA environment. To test this idea, we first examined gene arrangement along the linear DNA sequences. We divided the whole human genomic DNA into bins with a resolution of one million base pairs, a typical size of a topologically associated domain. Then we matched genes to the relevant bins based on their linear chromosomal positions. Statistical analysis of the genes spreading along the chromosomes shows that genes are not evenly distributed along the DNA sequences (18A). Most bins have less than ten genes, among them one third are gene-free. Around 3% bins (less than 100 bins) contain 17% of the overall human genes. This uneven distribution is slightly more profound for the DE genes under TGF-β treatment: DE genes reside in less than half of the bins and 17% of them are enriched in 2.5% bins.

To further examine the gene distribution along chromosomes, we defined an averaged linear distribution function (see supplementary materials for additional details). It measures how the density of genes belonging to a specific HMM cluster changes in respect to the transcription starting site (TSS) of a tagged gene, also a member of the cluster. Specifically, for a given DE gene belonging to HMM class α, we divided sequences flanking to its TSS into bins with a size of 125 kb on both sides (Fig. 22A, left panel, r = 125 kb), and calculated fraction of class α genes within each bin, then averaged over taking every gene in cluster α as the tagged one. For statistical comparison, we also calculated other two distribution functions as control that represent the fractions of all human and DE genes, respectively. Genes of over half of HMM clusters do not show clustering significantly higher than DE genes and all genes do. The upper

left panel of 18B shows such an example, Class C23. Only five classes show small but significant accumulation of genes from the same HMM class within the first pair of r = 125 kb bins compared to controls, and one of them (C24) is shown in 18B upper right panel.

Next, we investigated spatial arrangement of the DE genes using a set of available Hi-C data of MCF10A cells[172]. Similar to what used in statistical mechanics[281], we defined a set of radial distribution functions that measure the average radial density of genes cluster relative to a tagged gene (Fig. 22A, right). Again genes in Class C24 tend to spatially close (18B bottom right), likely due to their proximity along the linear sequence. Notably, genes in Class C23 also show significant tendency of spatial colocalization. The average relative density of C23 genes within the first shell is more than two folds of that of all genes. That means, some C23 genes not close along the linear sequence come close together spatially. Spatial clustering of genes from a HMM class can also be visualized from a two-dimensional plot of bins on chromosome 14 based on bin-bin contact frequencies from the Hi-C data (Fig. 18C). While C24 genes mainly reside in a number of isolated bins, most C23 genes reside in three bins that are spatially close. Further examinations reveal significant gene spatial clustering for all HMM classes compared to that of the controls (Fig. 18D), and the spatial clustering mainly takes place within each HMM cluster (Fig. 22B). These results indicate that it is a general phenomenon that genes sharing a common upstream regulator have higher tendency of spatial clustering.

We also examined how the genes within the first shell of a tagged gene are distributed along the chromosome sequence (Fig. 22C). While a large contribution comes from genes that are already close along the chromosome sequences, some gene elements as far as ~ 50M bp apart are brought together spatially. These observations are also consistent with other reports that transcription factors play a key role of pulling distal chromosome elements together[174,282].

3.2.4 Temporal switching between AP1 hetero- and homo-dimers fine tunes local chromosome structures and leads to different expression patterns of downstream genes.

Next we examined how co-localization of co-regulated genes influences gene regulation. We focused on a specific bin on chromosome 14 (Fig. 19A, red), which contains 12 genes, and five of them are down-regulated with TGF-β treatment. One of them is gene *fos* that encodes transcription factor FOS. The FOS protein functions by forming heterodimer AP1 with another transcription factor, JUN. AP1 binds to DNA and recruit other factors to bind to transcription initiation complexes of targeted genes and increase their expression activities[283]. The other four genes in the bin have AP1 binding enhancers, and DREM2 analysis identified FOS and JUN as the key regulators for them (Fig. 22D). Therefore a common local chromosome environment and a common regulator AP1 likely contribute to the similar expression pattern of these genes. In later discussions we call this bin as the "*fos*-bin".

To examine whether there is a larger spatial cluster of genes sharing a common regulator FOS, we expanded the analysis to genes in bins that are close to the fos-bin based on Hi-C results (with bin-bin distance < 8 a.u.) (Fig. 19A), which contain ~ 100 DE genes. The histone modification patterns among these groups do not show significant differences (Fig. 23A). Transcription factor binding site analysis revealed that most of them contain the core AP1 binding site 5'-TGAG/CTCA-3' (Fig. 19B). Interesting, the temporal expression patterns of these genes can be divided into four classes (Fig. 19C), decreasing as fos does (Fig. 23B, left), increasing similar to jun that is not within this expanded bin cluster (Fig. 23B, right), and transient down and up. The latter show similar initial dynamics as fos and jun.

Then the question is how these genes subject to the same local DNA environment and co-regulated by AP1 can have distinct and even opposite temporal patterns. It turns out that AP1 exist in two forms, FOS-JUN (FJ) heterodimers, and JUN-JUN (JJ) homodimers[284]. These two forms share the same core DNA binding site but with biased peripheral sequences, and the FJ dimer has higher DNA binding affinity than the JJ dimer does[285]. FJ dimer and JJ dimer also have distinct 3D protein complex structures[284], and the lists of their target genes only partially overlap. A plausible working model that is consistent with known information is sketched in Fig. 19D. Within a large-scale 3D chromosome structure formed by these genes, both FJ and JJ forms further compete for binding and bridging their respective target enhancers and promoters to form transient enhancer-promoter complexes for gene transcription. In untreated MCF10A cells the FJ form dominates. TGF-β treatment leads to decrease of FOS, and increase of JUN, with the dominating form of AP1 shifted to the JJ form. Consequently, transcriptions of genes regulated by FJ only decrease, and those of genes regulated either by JJ only or by both forms increase with time.

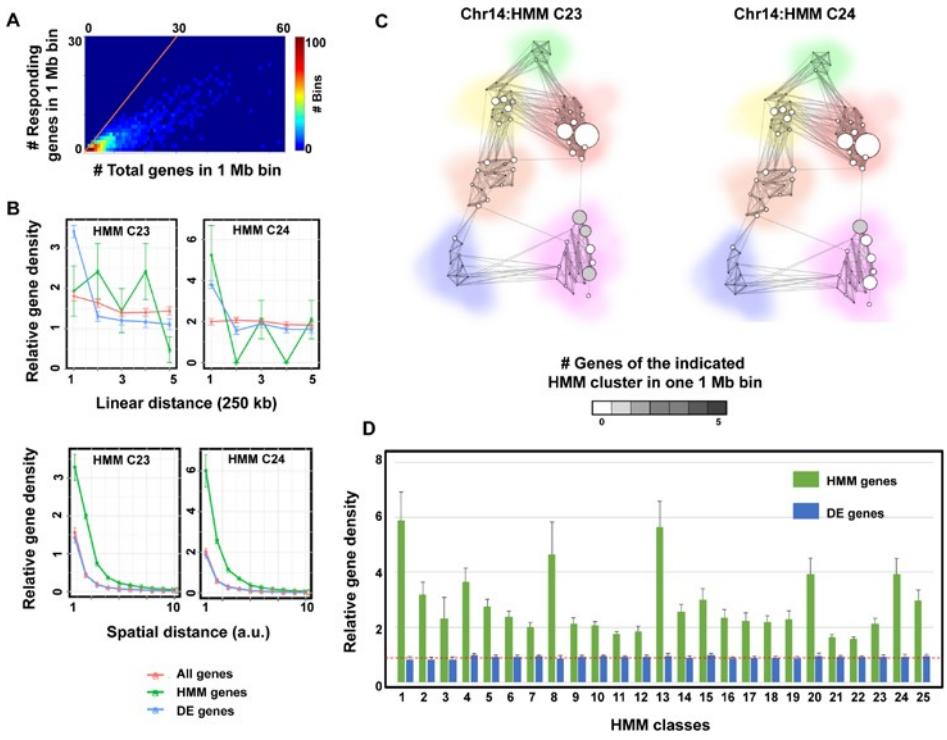

Figure 18 A fraction of genes with similar expression pattern and controlled by the same up-stream egulators tend to co-localize in the 3D chromosome structure**. (A)** Heat map shows numbers of 1 Mb bins containing given number of genes and TGF-β responding genes. The orange line highlight bins that all genes within these bins are response to TGF-β treatment. **(B)** Linear and radial distribution functions of TGF-β-responding genes within two HMM classes. We calculated the distribution of genes at three levels: all genes used samples of all available genes. HMM genes are genes within indicated HMM group. DE genes are genes that shown different expression during TGF-β treatment. For spatial distance 1 a.u. ≈ 60 nm. **(C)** Spatial clustering of TGF-β-responding genes belonging to two representative HMM classes, respectively. Every circle indicates a 1-Mb bin. The size of a circle scales to the total number of genes in this bin. The gray level in a circle scales to the number of genes in the indicated class. The two-dimensional spatial arrangement of bins within one chromosome was calculated by a fast-greedy algorithm based on the contact frequencies between every pair of bins from Hi-C data. The line width between two circles is proportional to the contact frequencies between the two corresponding bins. Background color indicates clustering results from fast-greedy algorithm. **(D)** The relative gene density of all HMM classes within the first shell of radial distribution.

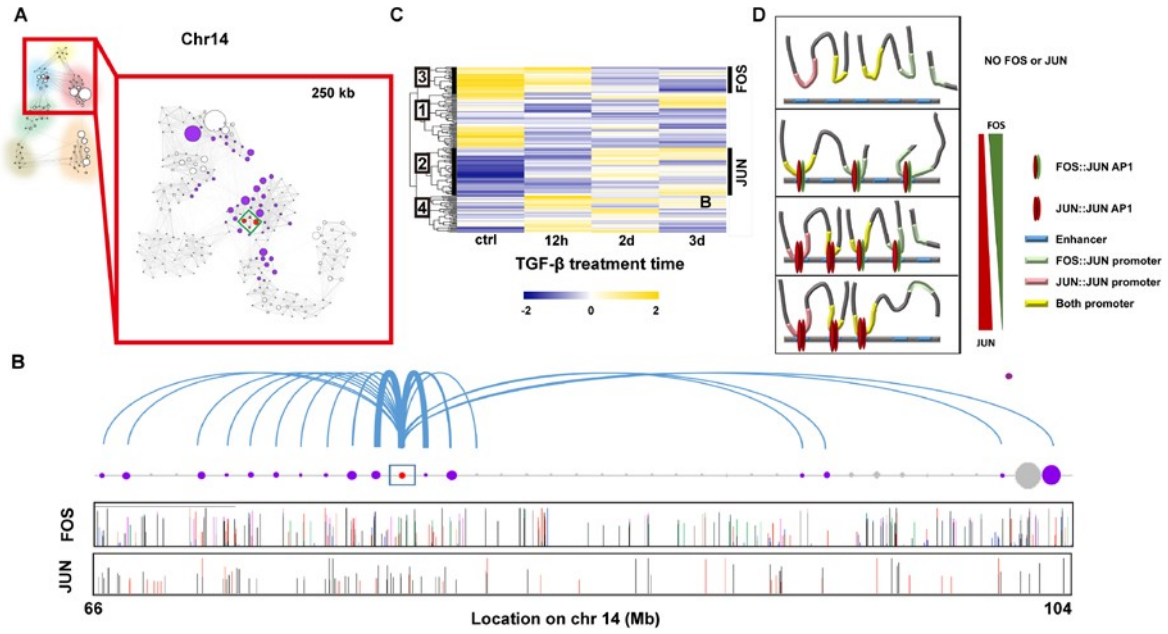

Figure 19 Temporal switch between hetero- (FJ) and homo-dimer (JJ) forms of AP1 regulates downstream group of genes and modifies local DNA spatial structure. **(A)** Spatial structure of chromosome 14. The red box delimits the bins (purple) that have strong correlation with the *fos*-bin. **(B)** Combination of visualizing chromatin interaction map and FOS/JUN binding site data near the fos bin. Purple circles are the 1-Mb bins that are spatially close to the fos-bin from Hi-C data as shown in panel A. Blue curves on the top indicate Hi-C correlation values between two bins. The size of a circle indicates the total number of genes in the bin. The predicted FOS and JUN binding sites are shown in bars. Colors of bars represent data from different cell lines and the height of a bar reflects the normalized peak. **(C)** Clustering of DE genes in the bins that defined in A. **(D)** Proposed model of chromosome structure change fine tuned by the two forms of AP1.

3.3 DISCUSSION

Cellular responses typically require change of transcriptional activities of groups of genes in a temporally coordinated manner. Through analyzing available transcriptome, epigenome, and Hi-C data of MCF10A cells in response to TGF-β treatment, we revealed spatial clustering of genes co-regulated by common factors. This co-localization may be a consequence of chromosome reorganization orchestrated by the common transcription factors. Reversely, this co-localization may also contribute to concerted gene regulation in eukaryotic cells, which can be viewed a further refinement of a strategy with the SIM network motif first noticed in prokaryotic cells. Spatial co-localization may facilitate simultaneous regulation of local chromosome environment of these genes, such as DNA methylation and histone modification patterns, and related local chromosome compaction, all of which affect gene expression activities.

Our analysis reveals two related mechanisms to achieve the spatial clustering of genes subject to common regulators. Some genes reside closely in the genomic sequence and thus spatially as well. Genes that are linearly separated can also become adjacent through forming three-dimensional structures. DNA binding factors such as transcription factors and transcription initiation complexes can drag targeted chromosome regions together to form enhancer-promoter hub structures[286]. These structures are cell type specific and more dynamic. Consequently, we predict that under long time TGF-β treatment MCF10A cells undergo EMT with accompanying large spatial rearrangement of chromatins, so genes in the upregulated HMM classes (e.g., C1 and C2) may show higher tendency of spatial clustering after TGF-β treatment than that without treatment as examined here. Hi-C measurements of TGF-β treated MCF10A cells in the mesenchymal phenotype may test this prediction.

While our analyses only provide correlation between common upstream transcription factor (and associated similar temporal gene expression pattern) and gene spatial localization, extensive studies have revealed that transcription factors actively orchestrate chromosome structure organization[282,287,288]. In this work we further propose that within a group of spatially clustered genes, trans-regulatory elements such as TFs can further fine-tune chromosome structures and transcriptional activities of individual genes that are spatially close. We identified such an example of two forms of AP1 regulating their target genes. The FJ and JJ form of AP1 complexes regulate multiple cellular processes such as cell growth, proliferation, apoptosis and EMT through acting on different sets of target genes. Proteins in FOS family and JUN family can bind to form AP1. The mRNA levels of both fos and another FOS family element fosl1 were down-regulated starting from 12 hours after TGF-β treatment, while those of another two members fosb and fosl2 did not change significantly. Similarly, genes in the JUN family members, junb and jund, showed similar expression pattern as jun (Fig. 23C). Based on gene expression patterns we propose that a temporal switch from FJ to JJ forms may adjust the local chromosome 3D structure and target gene transcriptions. The mechanism can be experimentally tested through in situ chromosome structure studies.

In summary, based on integrated analysis of transcriptome, epigenome and chromosome 3D structural information we propose a mechanism for concerted regulation of a group of genes. That is, it can be achieved through sharing a common trans regulatory element and physical colocalization of the target genes.

3.4 MATERIALS AND METHODS

Cell culture

MCF10A cells were purchased from the American Type Culture Collection (ATCC) and were cultured in DMEM/F12 (1:1) medium (Gibco) with 5% horse serum (Gibco), 100 μg/ml of human epidermal growth factor (PeproTech), 10 mg/ml of insulin (Sigma), 10 mg/ml of hydrocortisone (Sigma), 0.5 mg/ml of cholera toxin (Sigma), and 1x penicillin-streptomycin (Gibco). Cells were incubated at 37 °C with 5% CO_2. We induced the cells with 4 ng/ml human recombinant TGF-β1 (Cell signaling). Medium was changed every the other day.

RNA extraction and library preparation

Total RNA was isolated from the cell pellets with RNA extraction kit (Qiagen, Cat No. 74104) All the RNA products were confirmed with high quality (RQN score = 10.0) using the Fragment AnalyzerTM platform (Advanced Analytical Technologies, Inc). Libraries were prepared using NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, Cat No. E7530L) according to the manufacturer's instructions. Briefly, mRNA was first isolated from total RNA with oligo d(T)25 beads (all volumes were halved except for washing step, NEB, Cat No. E7490S), and then purified mRNA was denatured and melted into small fragments. Next, mRNA product was subjected to random priming and extension for reverse transcription. After that, double-stranded cDNA was end-repaired, dA-tailed, adaptor ligated and then amplified after 12 cycles of PCR. Purification and quality control were the last step for library preparation. The final quality-ensured libraries were pooled and sequenced on the Illumina HiSeq 4000 instrument for 150 bp paired-end sequencing.

RNA-seq data processing

Paired-end cleaned reads were aligned to human reference genome hg19 (UCSC) using TopHat (v 2.1.1) with default parameters. The BAM files of mapped reads were further used to annotate transcripts and calculate the FPKM values using the Cufflinks, Cuffquant, Cuffnorm suite[289]. DE genes were identified between any two time points with the criteri(A) fold change >2 or < 0.5 and FDR < 0.05. The FPKM values of genes from RNA-seq dataset were further cleanup by R. hierarchical-cluster of genes was performed by R package (pheatmap). DREM2 cluster was performed with the DREM2 software[279].

Chromosome structure establishment and distribution function calculations

MCF10A Hi-C results were downloaded online (GEO:GSE66733). Chromosome structure was established by an R package (igraph). Clustering of bins was obtained with the fast-greedy algorithm. Details of calculating the linear and radial distributions are provided in supplementary materials.

FOS/JUN binding site analysis

ChIP-seq of FOS binding results of Hela, K562 and MCF10A; JUN binding ChIP-seq results of hESCh1 and hUVEC were downloaded from ENCODE. 19B (bottom) was created by customized R code.

Linear distribution function

For a tagged gene belonging to class α, we divided the flanking sequences into bins with a size of Δl base pairs, and the *i*-th pair of bins $[(-i-1)\Delta l, -i\,\Delta l]$ and $[\,i\,\Delta l,(i+1)\Delta l]$, i = 0, 1, etc. In the *i*-th

pair of bins, there are $n_{i\alpha}$ genes belonging to the same class as the tagged gene. For the 0-th pair of bins counting of the genes should exclude the tagged gene. The linear correlation is calculated as $\sigma_{i\alpha}^{L}=\frac{\langle n_{i\alpha}+n_{-i\alpha}\rangle_{\alpha}}{N_{\alpha}-1}$, where i = 0, 1, 2, etc. N_α is total number of genes belonging to class α, and the average $<n_{i\alpha}+n_{-i\alpha}>_{\alpha}$ is performed over all genes belonging to class α as the tagged gene.

As a control, we calculated

$$\sigma_{i}^{La}=\frac{\langle n_{i}+n_{-i}\rangle_{\alpha}}{N-1}. \quad (37)$$

where n_i is the number of genes in the *i*-th bin, and *N* is total number of human genes, and

$$\sigma_{i}^{Ld}=\frac{\langle n_{di}+n_{-di}\rangle_{\alpha}}{N_{d}-1}. \quad (38)$$

where n_{di} is the number of DE genes in the *i*-th bin, and N_d is total number of DE genes,

If there were no class-specific gene clustering, we would expect that

$$\sigma_{i\alpha}^{L}=\frac{\langle n_{i\alpha}+n_{-i\alpha}\rangle_{\alpha}}{N_{\alpha}-1}=\frac{\langle n_{i}+n_{-i}\rangle_{\alpha}}{N_{\alpha}-1}\frac{N_{\alpha}-1}{N}=\sigma_{i}^{La}=\sigma_{i}^{Ld} \quad (39)$$

within statistical errors.

Instead we observed significant differences between the two for some HMM classes.

Spatial distribution function

Here we borrowed the idea of radial distribution function from statistical mechanics. We divided each chromosome into bins with a size of a 250 kb. A tagged gene that belongs to class α resides in a bin that we refer as the tagged bin. We set to analyze the spatial distance between the tagged bin and another bin containing a specific gene, for which we used the Shrec3D algorithm[290] to convert the contact frequency between two bins from hi-C data to a spatial distance. We also

divided the sphere centered at the tagged bin into shells with a width Δr (Fig. 22A), and defined the spatial correlation function between this bin and another one as,

$$\sigma^{R}_{\alpha\alpha}(i)=\left\langle \frac{n_{\alpha i}}{\frac{(N_\alpha-1)}{V}\frac{4}{3}\pi\left[((i+1)\Delta r)^3-(i\Delta r)^3\right]}\right\rangle_\alpha, \quad (40)$$

$$\sigma^{R}_{\alpha\beta}(i)=\left\langle \frac{n_{\beta i}}{\frac{N_\beta}{V}\frac{4}{3}\pi\left[((i+1)\Delta r)^3-(i\Delta r)^3\right]}\right\rangle_\alpha,\ i = 0,1,\text{ etc}, \quad (41)$$

where $n_{\alpha i}$ is the number of genes belonging to class α within a spherical shell (iΔr, (i+1)Δr), with that of the first shell excluding the tagged gene, $n_{\beta i}$ is the number of genes belonging to class β within a spherical shell (iΔr, (i+1)Δr), N_α and N_β are the total numbers of genes belonging to class α and β, respectively, V is the nucleus volume and we choose the unit so that $V = 1$, and the average is the same as above.

Similarly, we defined controls as,

$$\sigma^{Ra}(i)=\left\langle \frac{n_i}{\frac{(N-1)}{V}\frac{4}{3}\pi\left[((i+1)\Delta r)^3-(i\Delta r)^3\right]}\right\rangle_\alpha, \quad (42)$$

$$\sigma^{Rd}(i)=\left\langle \frac{n_{di}}{\frac{(N_d-1)}{V}\frac{4}{3}\pi\left[((i+1)\Delta r)^3-(i\Delta r)^3\right]}\right\rangle_\alpha, \quad (43)$$

where n_i is the number of genes within the i-th shell, and N is total number of human genes, n_{di} is the number of DE genes within the i-th shell, and N_d is total number of human genes. Again the tagged gene is excluded for counting n_0 and n_{d0}.

With similar derivation as for the linear distribution case, if there were no class-specific gene clustering, we would expect that within statistical errors,

$$\sigma^{R}_{\alpha\alpha}(i)=\sigma^{R}_{\alpha\beta}(i)=\sigma^{Ra}(i)=\sigma^{Rd}(i), \quad (44)$$

3.5 SUPPLEMENTARY TABLES AND FIGURES

Table 2 Gene ontology of DREM2 classes.

Class index	Enriched functions	Fold changes
Class 1	Metalloendopeptidase activity	16
	Heparin binding	15
Class 2	Calcium ion binding	4
Class 3	NA	--
Class 4	NA	--
Class 5	NA	--
Class 6	Protein kinase activity	3
	Enzyme binding	3
Class 7	NA	--
Class 8	Proteoglycan binding	17
Class 9	Metal ion binding	2
Class 10	DNA binding transcription factor activity	2
	DNA binding	2
	Metal ion binding	2
	G-protein coupled receptor activity	0.13
Class 11	NA	--
Class 12	NA	--
Class 13	NA	--
Class 14	NA	--
Class 15	Cargo receptor activity	13
Class 16	NA	--
Class 17	Signaling receptor activity	0.7
Class 18	Catalytic activity, acting on a tRNA	5
	Structural constituent of ribosome	5
	Catalytic activity, acting on DNA	5
	RNA binding	3
Class 19	Single-stranded DNA-dependent ATPase activity	40
	Histone kinase activity	33
	ATP-dependent microtubule motor activity, plus-	21
	Cyclin-dependent protein serine/threonine kinase	19
Class 20	Thiamine pyrophosphate binding	86
	RNA binding	3
	Protein binding	1.34
Class 21	Aminoacyl-tRNA ligase activity	13
	Structural constituent of ribosome	8
	RNA binding	4
	Adenyl ribonucleotide binding	2

Table 2 (continued)

Class 22	RNA polymerase I activity	23
	Proton-transporting ATP synthase activity,	23
	Oxidoreductase activity, acting on the CH-CH	12
	snoRNA binding	12
Class 23	Ligase activity, forming carbon-carbon bonds	50
	Coenzyme binding	5
Class 24	NA	--
Class 25	NA	--

NA: no significant enrichment

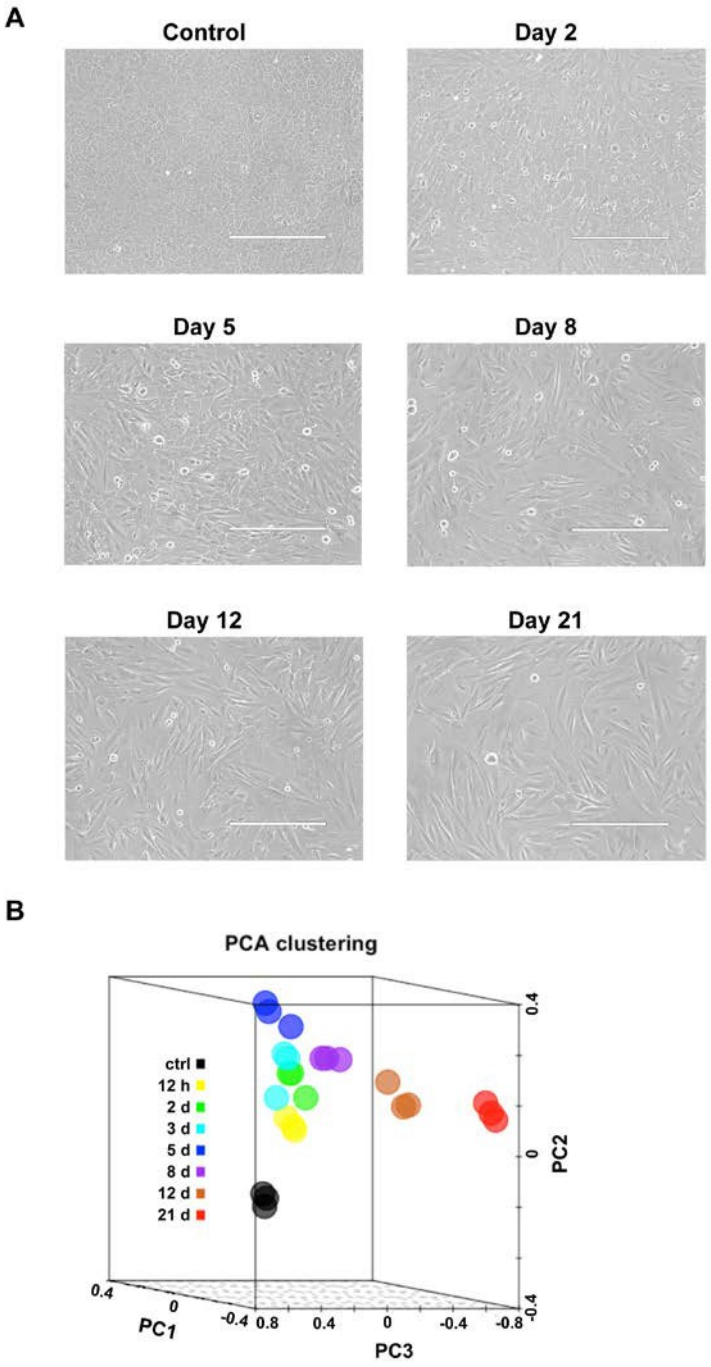

Figure 20 MCF10A cells respond to TGF-β treatment. **(A)** MCF10A cells show morphological change in responses to 4 ng/ml TGF-β at different time points. **(B)** PCA clustering reveals distinct gene expression patterns over time.

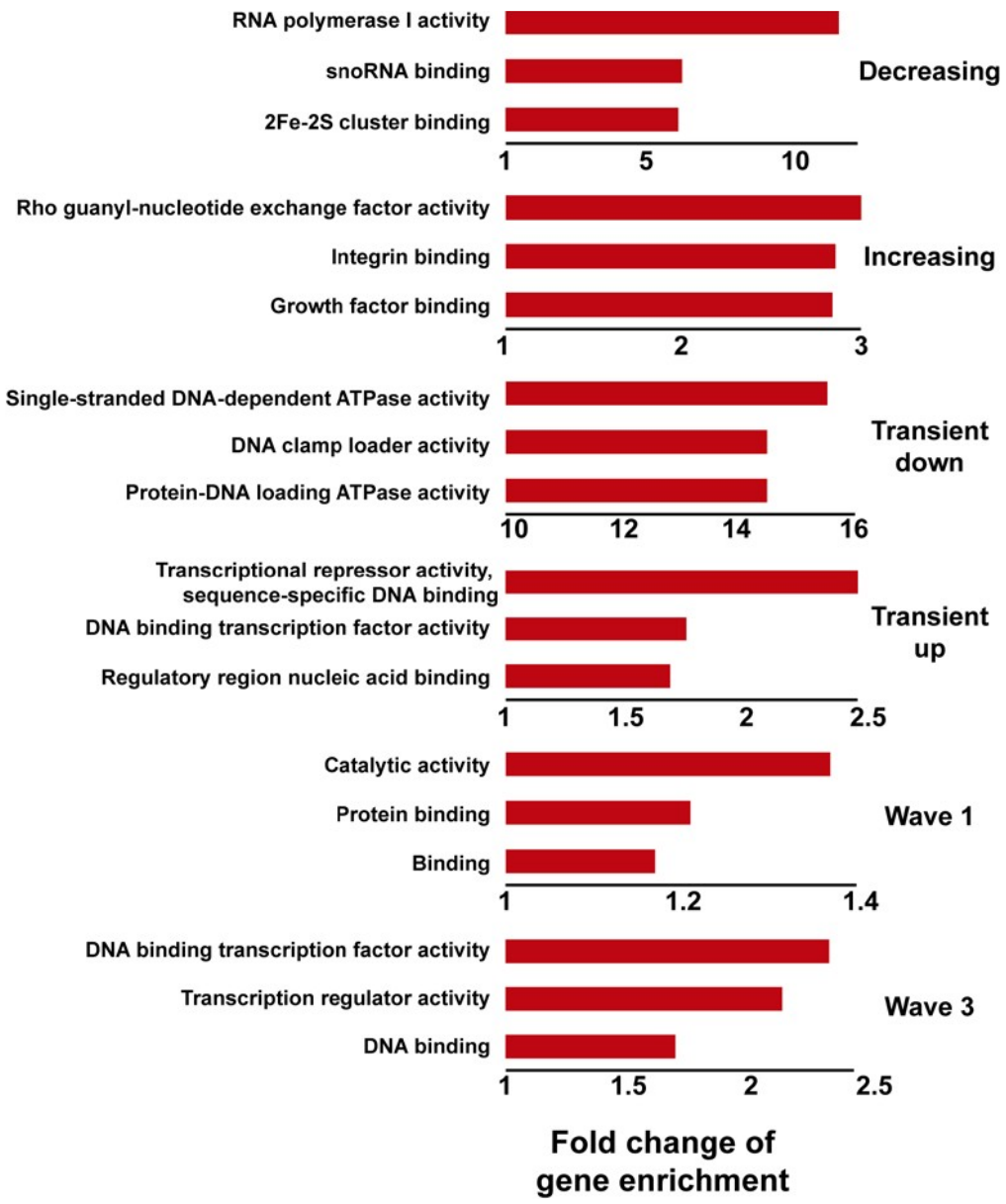

Figure 21 Gene ontology (GO) analysis of hierarchical-clustering classes.

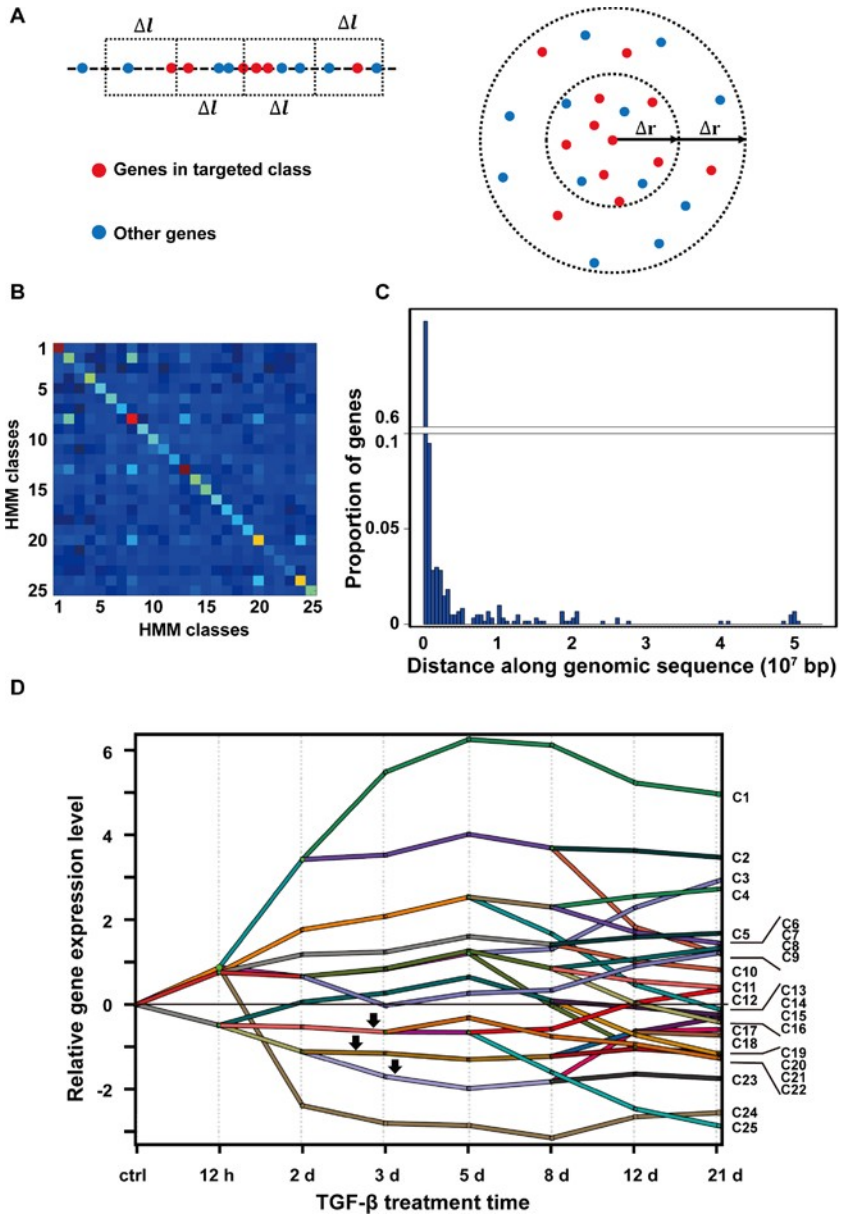

Figure 22 Genes in HMM class that are regulated by FOS. **(A)** Cartoon illustration of the linear and radial distribution functions. **(B)** Average density of genes within the same HMM class (diagonal) to the targeted gene or other HMM classes (off-diagonal) in the first shell of the radial distribution. **(C)** Distribution of distances along genomic sequence between a tagged gene and genes of the same HMM class in the first shell. **(D)** HMM classes (indicated by black arrows) containing genes that are regulated by FOS.

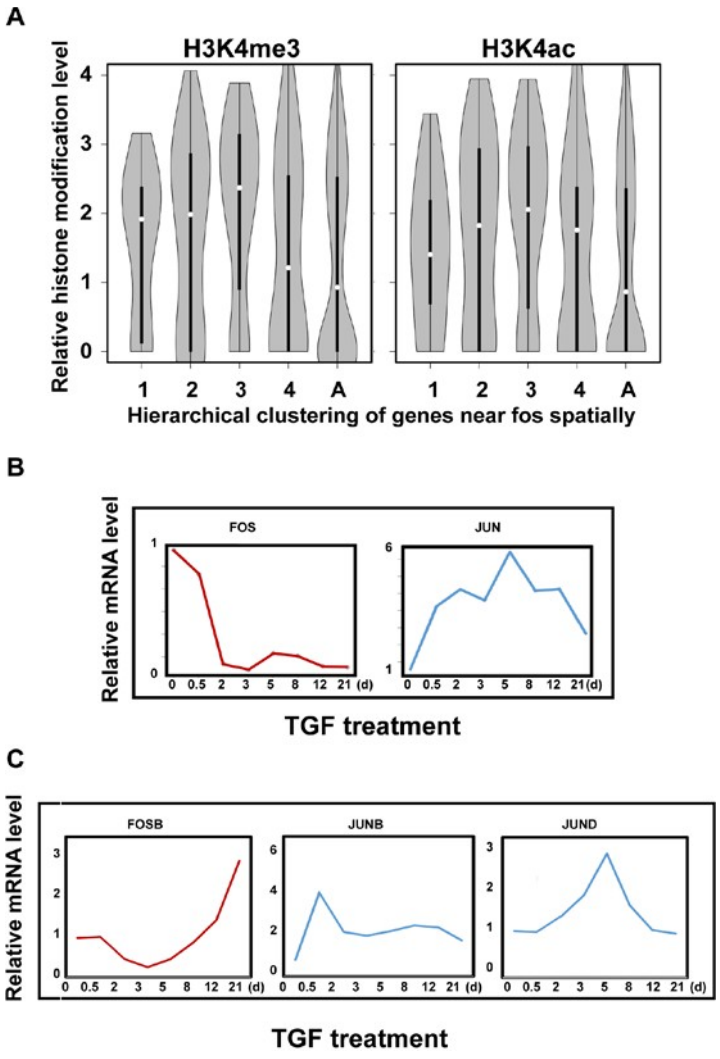

Figure 23 Gene expression is regulated by transcription factor and local chromosome environment. **(A)** Histone modification level in every expression cluster. The numbers on x-axis indicate the clusters in Fig. 4B, and "A" refers to all genes in the region highlighted in 19A. **(B)** mRNA expression levels of fos and jun in MCF10A cell line. **(C)** Relative mRNA level of FOS-family genes and JUN-family genes in MCF10A cell line.

4.0 CELLS REORGANIZE CHROMOSOME STRUCTURE FOR COORDINATED HISTONE MODIFICATION AND GENE EXPRESSION DURING CELL DIFFERENTIATION

A cell type transition process requires well-orchestrated global changes of gene expressions. It is intriguing how temporally coordinated regulation of multiple genes is achieved despite existence of large extrinsic and intrinsic stochasticity. One such major source of stochasticity is transcriptional bursting, where even under constant levels of trans-regulatory elements genes remain transcriptionally active for a period, then become inactive for a more extended period before switching back to be active. One expects that such bursting dynamics may destroy temporal coordination of genes that are required for specific cellular functions. We integrated datasets of gene expression, histone modification, and chromosome conformation during mouse nervous system development, and identified that genes having related functions and regulated by common TFs tend to cluster spatially and share similar histone modification patterns. The observations lead to a model that genes in proximity synchronize their transcriptions by synchronizing their inherently stochastic switching between histone modification states with different transcriptional activities. Analysis of single cell RNAseq data confirmed such correlated fluctuations.

4.1 INTRODUCTION

During mammalian embryonic development, a totipotent fertilized egg differentiates step by step into differentiated functional cells with unique transcriptome profiles and chromosome structures that are cell type and tissue specific[291]. For example, in neuron development one can isolate pluripotent embryonic stem cells (ESC), which can be induced to differentiate into neural progenitor cells (NPC). The later further differentiates to neural cells that perform nervous system specific function[292,293]. Each cell type transition is accompanied with a tightly temporally regulated global reprogramming of gene expressions, since coordination of multiple proteins is often necessary for carrying specific biological processes. Unbalanced protein levels may impede completion of a process and ineffective usage of those proteins in high expression. Similar coordinated gene expression is also required in differentiated cells for processes such as cell cycle progression and cell motility change.

From an engineering perspective precise control of waves of coordinated and balanced gene expressions is a highly nontrivial task. While systems biology analysis has identified a single input module (SIM, or fan-out) network motif, in which one transcription factor (TF) regulates and thus synchronizes expressions of multiple gene targets with related functions, this coordination in principle can be jeopardized by the involved stochastic processes such as transcription and translation in cells (Fig. 24A). This difficulty arises especially in eukaryotic cells, where local chromosome structures and epigenetic modification patterns critically affect gene promoter activities together with TFs. Specifically, single molecule studies revealed bursts of gene expressions, indicating that a gene stochastically switches between two or more modes of different transcriptional activities. While transcriptional bursting has been observed in both

prokaryotic and eukaryotic cells, the latter typically has longer inactive period in the order of hours[294,295-302]. This long period is comparable to the time scale of transcription, translation, and protein half-lives, and makes it a potential major contributor to temporal fluctuations of protein levels (Fig. 24B). Existence of transcriptional burst is a major source of intrinsic noises. Genes with uncorrelated bursting can have temporally uncorrelated expressions even if they are regulated by a common TF, therefore additional regulation mechanism is needed.

Several studies reveal that local genomic environments such as chromosome structures and histone modification patterns contribute to transcriptional burst through modulating chromosome accessibility and binding affinities of gene regulatory elements[303]. Notably several histone marks affect transcriptional bursting size and frequency[301,302,304-307]. Consider a simple case that a histone exists in one of two states, with and without a specific mark[308-310]. Histones with proximity interact with each other, which lead to a collection of nucleosomes of a gene promoter or gene body existing in a collective state of either dominated by histones with or without the marks[13,36,150,311]. These two collective states have different gene transcriptional activities, and can stochastically interchange that contribute to transcriptional bursting (Fig. 24C). Then hypothetically a mechanism that correlates epigenetic state switches of two genes can synchronize expression fluctuations of these genes.

To investigate the above hypothesis, we combined data of RNA-seq, histone modification ChIP-seq, and Hi-C from mouse nervous system development[312,313]. Development of murine ESCs (mESC) to NPCs then neuron cortical neural (CN) cells is an experimentally well-characterized process. Through integrated interpretation of different types of data, we found that cells reorganize chromosome structures dynamically to posit some commonly regulated genes

with related functions spatially close. Our model analysis predicts such spatial colocalization results in correlated histone modification state switching and gene expression, and the latter was subsequently confirmed through analyzing single cell RNAseq data.

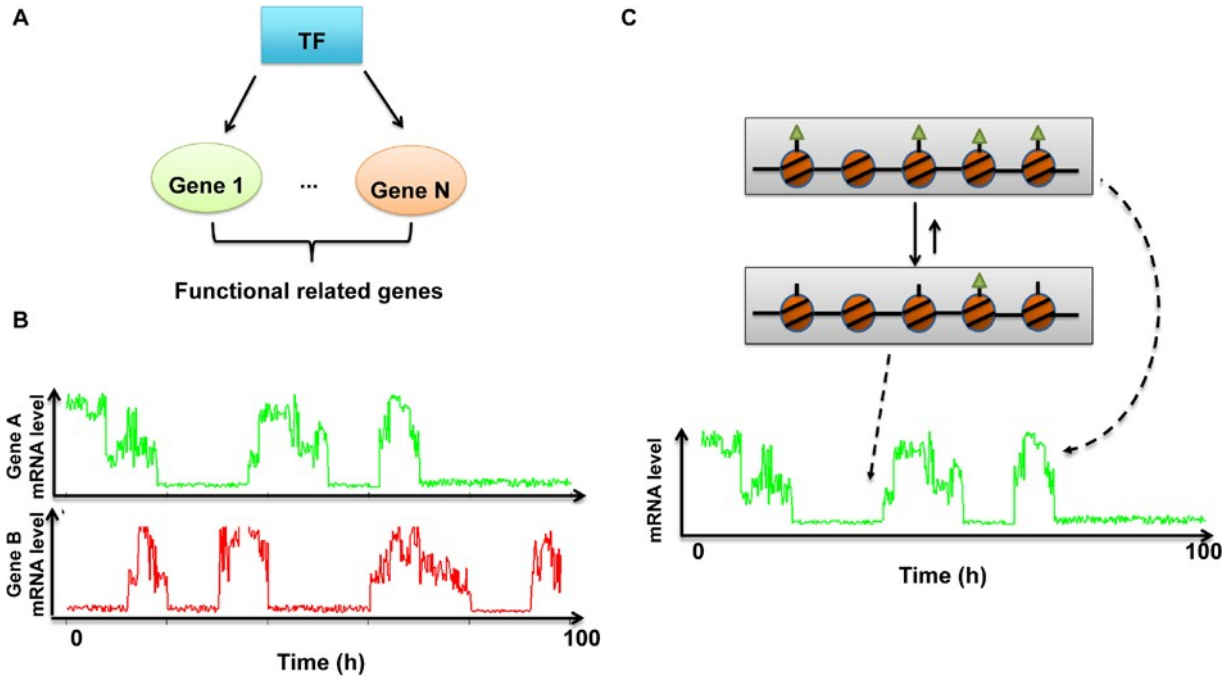

Figure 24 Gene regulation and bursting expression.

(A) Schematic of the SIM module in gene regulation. **(B)** Asynchronized bursting expression of two genes. **(C)** Transcription activities of a gene is regulated by collective histone modifications.

4.2 RESULTS

4.2.1 Genes with related functions tend to be co-regulated by common TFs during cell differentiation

We first performed Dynamic Regulatory Events Miner (version 2.03, DREM2) analysis on available RNAseq data[313], which combines time series clustering with pre-established transcription networks to classify genes using a Hidden Markov Model (HMM)[278,279]. The analysis clustered over 12,000 differentiated genes (DE) out of ~24,000 overall genes into seven final classes (Fig. 25A, PA3-PA9), and identified a small number of master TFs that regulate their expression and form SIM motif structures (e.g. in Fig. 25B). Table 3 shows the first ten TFs that play key rules in each splitting pathway. The score split indicates the probability of having a greater value than the number of genes assigned to be regulated by targeted TF. A lower score split associates to more significance. Split % is the percentage of the number of genes regulated by targeted TF out of the parent path. For example, REST is the top key TF in pathway 4 that regulates proteins and miRNAs with functions related to neuron development etc., and a number of TFs such as SP1, HDAC1 that are required in nervous system development and cell fate decision. GO analysis also shows that genes within one class have related functions (Fig. 29). For example, genes that relate to nervous system development and neuron are enriched in PA4 and PA5 classes.

Table 3 Key TFs of DREM2 classes.

TF	Score Split*	% Split**
PA1		
REST	1.27E-19	60.2
EGR1	3.46E-15	58.43
EGR3	3.40E-12	57.04
EGR4	9.90E-12	61.69
TEAD2	1.14E-11	61.85
EGR2	1.28E-11	56.81
ZFP219	2.46E-11	61.93
POU2F1	1.13E-10	53.61
WT1	1.86E-10	61.01
HIC1	3.45E-09	57.81
PA2		
PITX2	3.41E-36	66.86
ELK4	7.96E-19	59.83
RB1	5.56E-12	54
E2F5	9.26E-12	53.79
E2F2	9.26E-12	53.79
TFDP2	3.09E-11	52.57
NFYB	9.44E-10	48.44
HLF	8.93E-08	48.83
NFYA	1.28E-07	46.69
E2F4	5.80E-07	45.71
PA3		
PPARA	3.62E-06	16.57
TLX1	2.72E-05	17.92
TMEM37	3.70E-05	15.57
PGR	3.70E-05	15.57
ERF	5.54E-05	17.68
NR3C1	7.88E-05	15.13
PIK3R3	1.11E-04	17.4
ERG 1	1.11E-04	17.4
PSMD12	1.11E-04	17.4
FSCN1	1.11E-04	17.4
PA4		
REST	2.08E-49	50.52
POU2F1	3.99E-23	37.27
MEF2A	2.27E-21	38.15
LHX3	2.56E-14	46.96
POU2F2	3.33E-13	41.39
SLC22A1	3.33E-13	41.39
VSX2	2.51E-12	45.03
ONECUT2	4.72E-11	44.18
CUX1	7.75E-11	34.25
POU3F2	2.73E-10	35.67
PA5		
ZBTB7A	2.15E-22	51.53
TRP53	3.70E-19	41.61
SMAD4	8.74E-18	42.86
SMAD1	1.86E-16	45.19
ZIC2	2.43E-15	47.8
USF2	4.52E-15	40.95
ESCO1	4.50E-15	48.06
CTF1	4.50E-15	48.06
SMAD3	1.60E-14	41.56
EP300	1.64E-14	47.12
PA6		
PITX2	4.07E-21	87.5
ELK1	1.23E-14	62.5
ELK4	9.36E-13	72.99
NRF1	1.37E-08	59.55
NFYA	8.14E-07	56.07
NFYB	1.18E-06	56.84
YY1	1.57E-06	55.72
GABPA	7.55E-06	56.19
FLI1	2.23E-04	55.74
CUZD1	2.23E-04	55.74
PA7		
SMAD3	4.72E-30	34.17
SMAD7	6.73E-30	42.05
SMAD5	6.73E-30	42.05
SMAD6	6.73E-30	42.05

Table 3 (continued)

SMAD2	6.73E-30	42.05
SMAD1	1.64E-29	36.92
SFPI1	2.52E-26	33.56
SMAD4	3.67E-26	30.46
KLF12	3.25E-25	39.78
ZIC2	4.05E-25	39.27
PA8		
SRF	3.71E-12	59.46
MEF2A	2.99E-11	58.14
ZFP238 1	1.34E-10	70.72
ELK4	5.36E-10	65.37
FOXD3	1.23E-07	61.52
NKX3-1	2.66E-07	60.44
EGR2	2.27E-06	58.29

FOXF1A	3.37E-06	62.5
CREB1	1.19E-05	54.24
ATF6 1	1.68E-05	57.05
PA9		
RB1	1.85E-10	59.21
E2F5	2.39E-09	58.03
E2F2	2.39E-09	58.03
NFYB	2.97E-05	49.73
NFYA	4.18E-05	49.11
TFDP2	3.40E-03	48.82
E2F7	4.09E-03	49.62
OTX1	0.046	47.55
OTX2	0.046	47.55
TFDP1	0.172	43.28

* The score split indicates the probability of having a greater value than the number of genes assigned to be regulated by targeted TF.

** Split % is the percentage of the number of genes regulated by targeted TF out of the parent path.

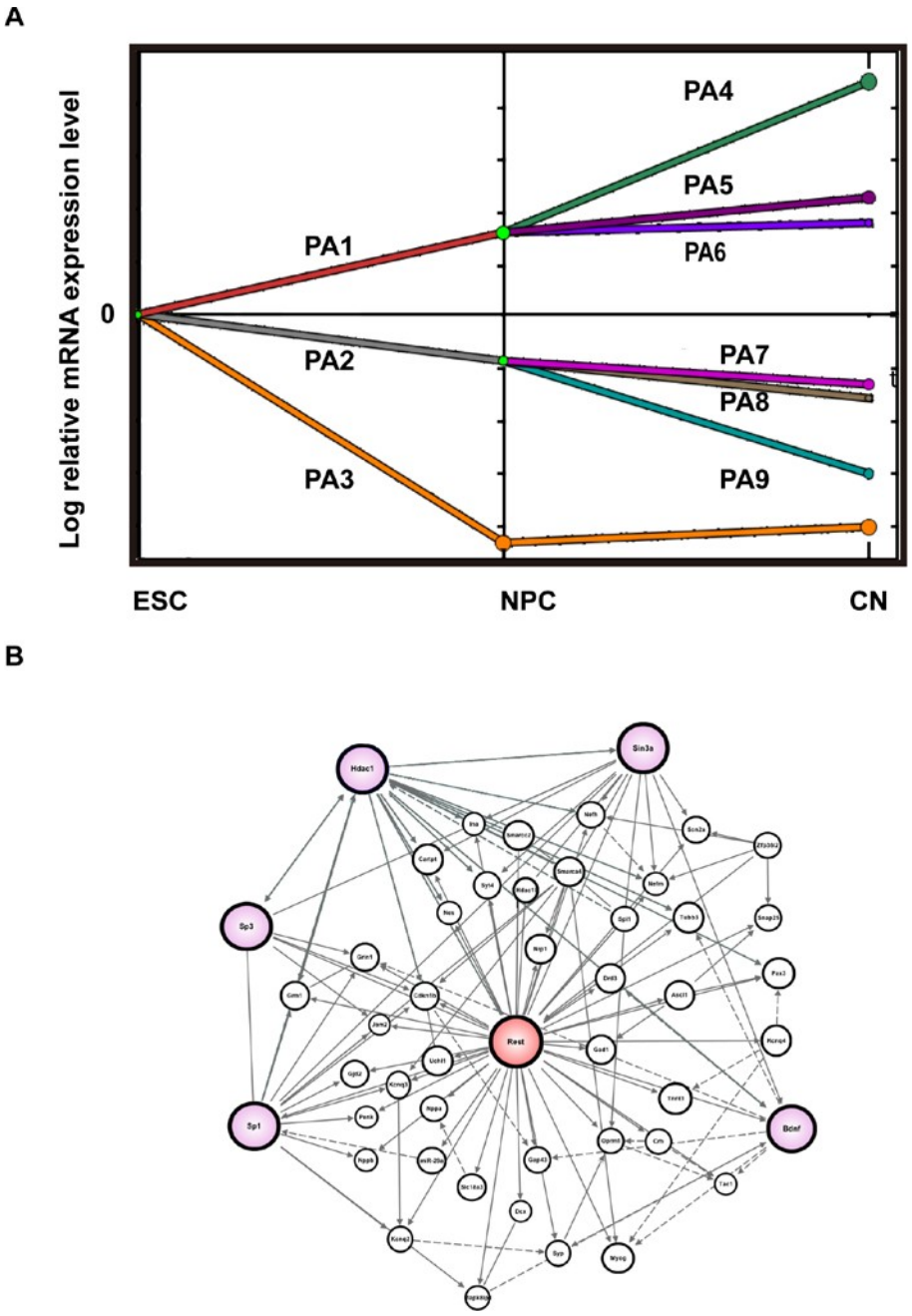

Figure 25 DREM2 analysis clusters DE genes based on gene expression and involved regulatory TFs. **(A)** Gene clustering based on the DREM2 algorithm. **(B)** Regulation network describing regulation of REST on targeted genes.

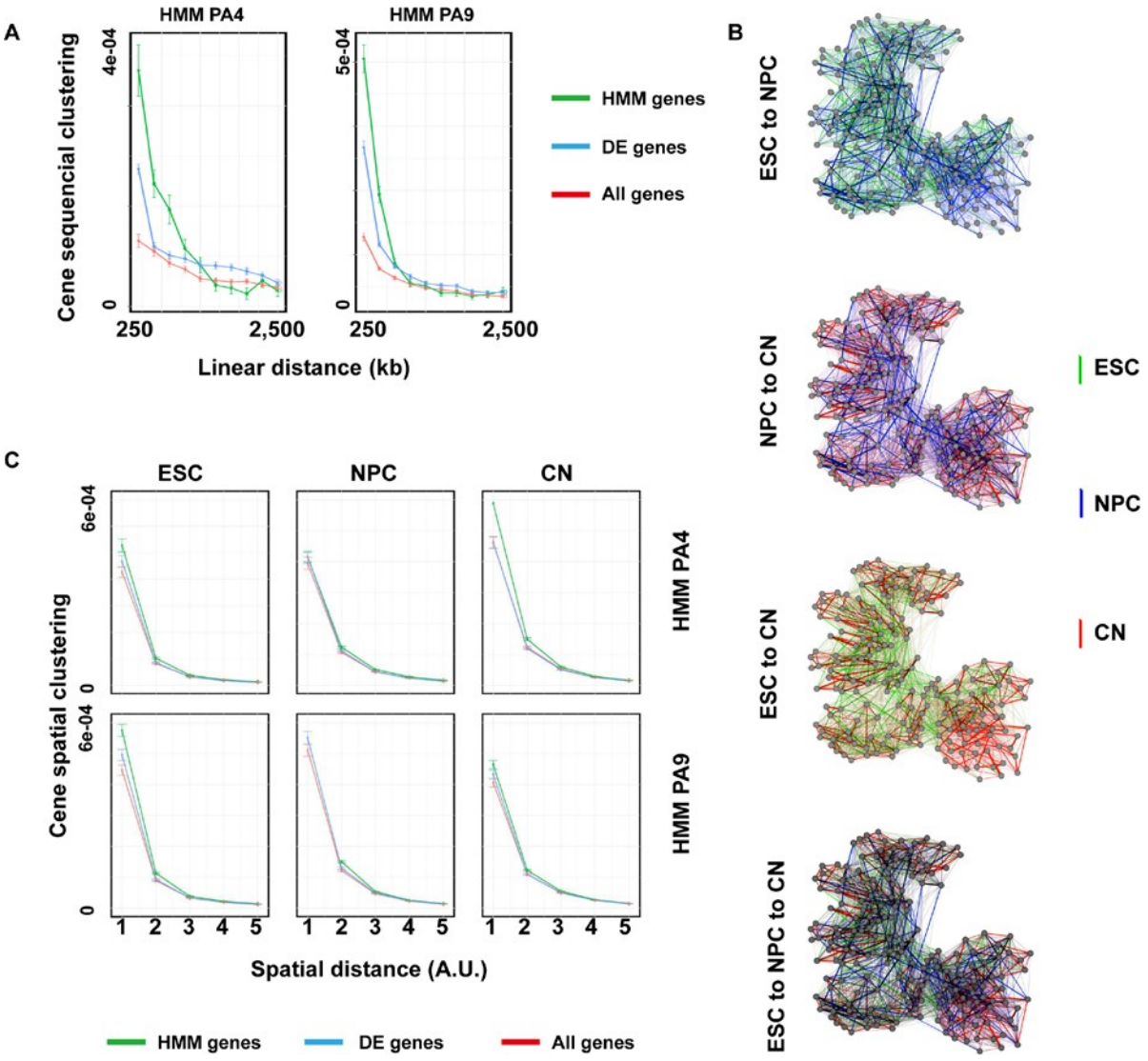

Figure 26 Relationship between gene expression and chromosome localization. **(A)** Distributions of genes grouped by DREM2 clustering along the linear DNA sequence. **(B)** Spatial rearrangement of chromosome 1 structure during differentiation of ESC to NPC then CN cells. Each node in the plot represents an 1-Mb genome bin. The line width between each pair of nodes is proportional to the Hi-C contact frequency. Fast- greedy algorithm was used to calculate clustering of the nodes. **(C)** Radial distributions of genes belonging to each DREM 2 class. Each bin represents a 50-kb genome bin. Roughly 1 a.u ≈ 60 nm

4.2.2 Co-expressed genes tend to cluster spatially.

To investigate the relation between gene expression and local chromosome environment, we analyzed the spatial arrangements of genes based on the HMM clustering. First, under 250 kb resolution, genes in the same DREM2 class tend to cluster along the DNA sequence compared to reference genes (Fig. 26A). A finer 50-kb bin analysis shows that the clustering actually mainly take place within the first 100 kb. These results indicate that there is certain build-in tendency in the DNA sequence to cluster genes with related functions and common regulatory elements together.

Consider that DNA is a three-dimensional entity, it is natural to ask whether gene clustering is beyond the linear sequence. Indeed recent studies uncovered several basic features of chromosome structures, such as topologically associated domains (TADs) and larger meta-TAD structures. Enhancers, super-enhancers, and other long-distance *cis*-regulatory elements work cooperatively with trans-regulatory elements such as TFs and lnRNAs orchestrate these three-dimensional structures[174], and drag DNA segments far from each other along the linear sequence spatially close. Hi-C data reveals large spatial reorganization of the chromosome three-dimensional structure during neuron cell differentiation (Fig. 30A). To better visualize such reorganization, we designed a two-dimensional representation, and figure 4.3B shows example plots of chromosome 1. In these plots, each circle represents an 1-Mb bin, with widths of the edges being proportional to the Hi-C contact frequency between two corresponding bins. We arranged the bins in the plot through a fast-greedy clustering based on Hi-C contact frequencies of ESC cells. The clustering analysis shows that the bins spatially segregate into several clusters, with some bins separated by 100 Mbs or more reside closely in the three-dimensional space. Next to compare structures of different cell types, we fixed the arrangement of the bins on the

plots, using only the edges to reflect changes of the contact frequencies. From ESC to NPC, there is a large concerted reorganization of the last one-third part of chromosome 1, reflecting a major cell type change from multipotent stem cells to a specific cell lineage. We observed genes related to nervous system development, such as *brinp2/3, astn1*, *et.al.*, in that region. Similarly there are scattered chromatin changes from NPC to CN. Overlaying the 2D structure of all three cells reveals some strong links that exist uniquely in NPC cells. That is, these DNA conformation structure appear transiently during differentiation from ESC to NPC, then disassemble during differentiation from NPC to CN.

To quantify the relationship between gene spatial clustering and gene co-regulation in neuron cell differentiation, we performed a radial distribution function analysis on genes within each DREM2 class. Again, we divided chromosomes into 50 kb bins and align genes into bins based on their location. Then, we used Hi-C data with 50 kb resolution and calculated the relative spatial distance between every pair of bins on the same chromosomes in the three types of cells using the Shrec3D algorithm[290]. Next, we calculated the radial distribution of genes in each HMM cluster, which reflects the extent of co-localization of genes both linearly close and separated. First, we observed that compared to all genes or all differentiated genes, genes in specific HMM class have higher tendency to cluster spatially. For example, in ESC cells, genes in HMM class PA4 or PA9 have higher density in the first spatial bin ($\leq$ 60 nm) compare to references genes or all genes (Fig. 26, left column). Moreover, if we compare gene densities of every HMM class near one targeted genes, we reached similar conclusion that genes within the same HMM class as the targeted gene have higher density around this targeted genes than genes in other HMM (Fig. 30B). These observations indicate the spatial localization of genes correlate with their regulation. Next, we performed the same analysis on other cell types during ESC

differentiation, and observed similar pattern of gene clustering (Fig. 26C, Fig. 30C). The comparison also reveals cell type specific chromosome reorganization. For example, In HMM class PA4, genes are more spatially clustered in CN cells than in others; while genes in HMM PA9 have higher tendency of clustering in ESC cells. Interestingly, comparing gene density of every HMM class in the initial ESC and final CN cells, genes in the sub-branches of PA2 (HMM PA7-PA9) and PA3, which are down-regulated during neuron development, tend to have higher density in ESC than CN. In contrast, genes in PA4 and PA6, which are up-regulated during cell differentiation cluster more in CN than ESC (Fig. 30C). This scenario indicates that active genes cluster more spatially.

Taken together, we classified genes into different DREM2 class based on both their expression levels and key regulators. A radial distribution analysis shows close relationship between the DREM2 class of genes and their spatial co-localization. The latter changes during cell type transition.

4.2.3 Co-localized and co-regulated genes also have stronger correlation on histone modification patterns.

Previous studies showed that genes expression levels are related to local epigenetic modifications on chromosomes[179]. Neural cell differentiation is accompanied with genome-wise reprogramming of epigenetic modifications. Therefore we set to examine the relationship between gene spatial distribution and histone modifications by focusing on a specific hot region identified from the above combined DREM2 and Hi-C data analysis. Genes in this chromosome

chromosome region need to be in the same DREM2 class and thus functional related. The region should also be in proximity with some distal chromosome regions so we can analyze whether and how genes from these different regions are co-regulated.

The region we chose is on chromosome 1, which we named $region_{act}$. All four genes within $region_{act}$ have increased expression during neural cell differentiation, and three of them are required in the nervous system. For example, *astn1*, which highly expresses in cortex and frontal lobe, relates to neuronal migration[314]. BMP/retinoic acid inducible neural specific 2 (*brinp2*) encodes the important neural cell specific proteins[315]. During neural cell development there is clear spatial reorganization between $region_{act}$ and several other bins. Figure 27A shows correlation among bins near $region_{act}$. Strong interactions exist between $region_{act}$ and one of its neighbors in ESC, which become much weaker in CN cells. Instead, interactions between $region_{act}$ and two bins nearby on each side appear significantly stronger in CNs. That is, genes in $region_{act}$. form stronger contacts to both bins during the transition from ESC to CN.

Next, we analyzed the epigenetic features in $region_{act}$ and regions clustered to it during cell type transition. We examined three typical histone modification marks, H3K4me1, H3K27ac, and H3K27me3, which are strongly associated to gene activity. The histone modification data of ESC is obtained from Santos lab. Those from CN are obtained from Bonev *et.al.*[313]. We aligned the histone modification ChIP-seq location to chromosomes. Then we compared ranks of local histone modification levels among three cell types. Figure 27B combines information of gene expression, DNA dynamics, and histone modification levels. The red curves on the top indicate that the two underlying regions become physically closer to each other with the transition from ESC to CN. The green curve indicates that the two regions are physically closer in ESC than in CN cells. Under the curves, seven colors are used to distinguish

genes in different DREM2 class. For convenience of inspection genes in PA4, the class all genes in region$_{act}$ belong to, are singled out and re-drawn underneath. Interestingly, three regions under the red curves have more similar pattern of change in H3K4me1 and H3K27ac. All of them acquire more active histone modifications in CN cells. The two regions under green curve have opposite changes of histone modification level. Analysis on the differences of histone modification between ESC and CN shows that genes physically close tend to have similar histone modification patterns.

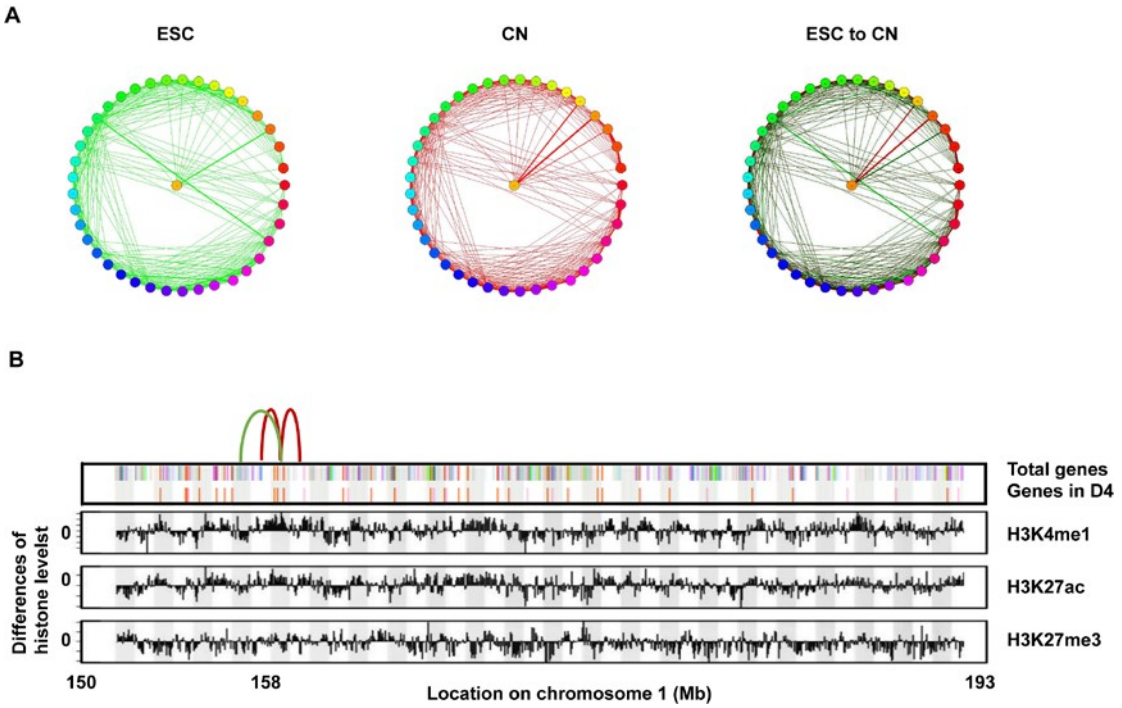

Figure 27 Commonly regulated genes tend to cluster spatially and have similar histone modification**(A)** Contact frequencies of selected regions on chromosome 1 based on high resolution Hi-C map. The lines in green, blue, and red are correlation between pairs of bin in ESC and CN samples, respectively. When overlaid, the overlapping lines turn to black. The width of a line represents the correlation between bins. The center bin is the targeted spatial bin. **(B)** Combination of visualizing chromatin interaction map of targeted region and changes of histone modification during neuron system development. Red curves on top are the regions have significant stronger interaction in CN cells. Green curve links two regions that have closer distance in ESC. Colors of bars immediately under curves represent genes in different HMM class. Height of histone marks is the differences of histone modification level from ESC to CN (histone level in CN – histone level in ESC).

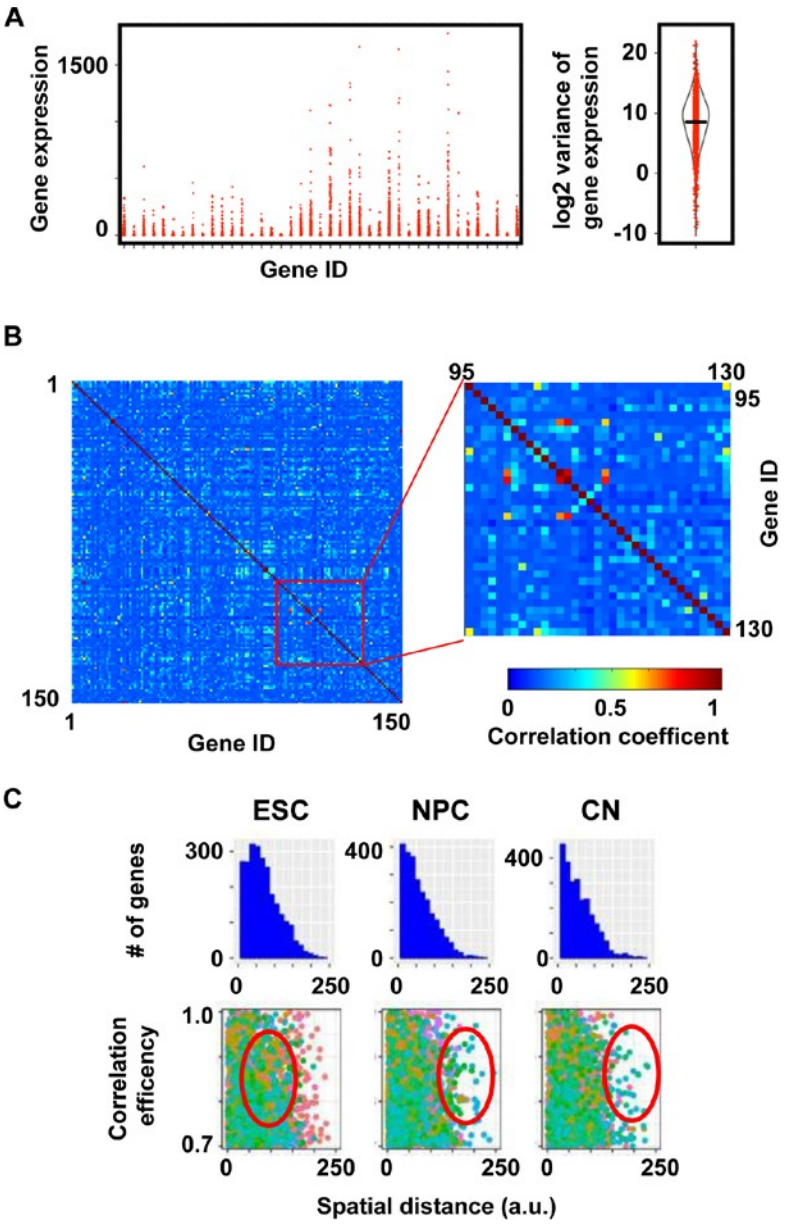

Figure 28 Transcriptions of functionally related genes in proximity are correlated.**(A)** Variances of genes among single cells are normally large. Left: Representative genes in regions near $region_{act}$.Right: variances of all D4 genes from single cell data. **(B)** Correlation coefficient of genes in chromosome 1 (left) and zoomed in figure of targeted map (right). **(C)** Histograms of distance distribution between pair of genes with high correlation coefficient (> 0.7) in different cell types (top). Relationship between correlation coefficient and spatial distance of pair of genes in different cell types (bottom)

4.2.4 Co-localized genes have both similar histone modification patterns and more synchronized transcriptional bursting

The observation that colocalized genes tend to have similar histone modification patterns can be explained from cooperative histone modification reading-and-writing mechanism discussed in previous studies.[311,316-318] That is, a nucleosome with certain histone mark may facilitate recruitment of the corresponding enzymes to neighboring nucleosomes and adding the same mark. Consequently, genes that are spatially close tend to have similar histone modification patterns. More importantly, the stochastic switching dynamics of the histone modification states of the genes tend to be synchronized. Since histone modification state switching of the gene promoter or gene body is a major contributor to transcriptional bursting, the model predicts that the bursting dynamics of these genes is better synchronized.

To test this prediction, we analyzed the expressions of 153 DREM2 PA4 class genes on chromosome 1 using existing single cell RNAseq data of 434 mouse brain cells [312]. All the genes show large cell-to-cell variations of readings of the transcripts, typically spanning several decades (Fig. 28A). Each pair of the genes shows positive correlations, which is not surprising since these genes are regulated by common TFs (Fig. 28B). Noticeably a number of pairs, including some well-separated (5 - 35 Mbs) along the DNA sequence, show significantly higher (> 0.7) correlations. That is, these pairs of genes tend to have either correlated high or low expression in individual cells. Comparing these genes and Hi-C data reveals that they are spatially close.

To further analyze the relationship between the correlation coefficient and spatial distance of genes in genome-wide, we picked all pairs of genes that have correlation coefficient higher than 0.7 and plot their spatial distances (Fig. 28C). The results show obvious changes of

the spatial distance distribution accompanied with differentiation of ESC to CN. Generally speaking the chromosomes become increasingly more compact from ESC to CN (Fig. 28C, top). However, some pairs of genes have larger distance in CN than in ESC. For example, pairs of genes in red circles are further from each other during cell differentiation. Therefore the single cell RNAseq data corroborates the prediction that spatially localized genes can synchronize their transcriptional bursting.

4.3 DISCUSSION

An intriguing observation is that biological systems can achieve remarkable performance, often to the limit set by physical laws, despite large environmental noises and intrinsic stochasticity of processes involved. Studies at single molecule and single cell levels have demonstrated transcriptional bursting, i.e., the stochastic activation and inactivation of gene promoters that lead to discontinuous and bursting synthesis of mRNAs. The universal existence of transcriptional bursting suggests that this existence is inherent and largely unavoidable for the transcriptional process especially in eukaryotic cells, and contributes to cell heterogeneity. This observation raises a serious question on how cells minimize effects of this type of stochasticity with its time scale comparable to the transcriptional and translational processes.

In this work we discovered a simple and effective strategy cells use to coordinate expressions of multiple genes, i.e., synchronizing instead of reducing transcriptional bursting of different genes. Mechanistically it is achieved by physically clustering genes with related functions and regulated by common TFs. Due to cooperativity of histone mark reading-and-writing among spatially close nucleosomes, two genes in proximity synchronize their stochastic

switch of histone modification states. Since the latter is a major contributor of transcriptional bursting, transcriptional events of the two genes are temporally correlated, albeit stochasticity of each event may not be much affected.

Our proposed mechanism has a number of testable predictions. First we predict higher temporal expression correlation between two functionally related genes in physical proximity compared to that between two that are also functionally related with a common regulator but resides distally in the nucleus. Analysis of single cell RNAseq data have confirmed such increased expression correlation. The temporal correlation can also be tested directly with live cell imaging using two-color transcription reporters incorporated at the endogenous sites of selected pairs of genes[302]. We expect to observe transcriptional bursting as in existing live cell studies in the field, but with the two signals showing strong temporal correlation. Recent advances of CRISPR-based gene editing can facilitate such study[319]. Another prediction is related to the histone modification patterns of these two groups of functionally related and co-regulated genes. We predict that in both cases genes show cell-to-cell heterogeneity of levels of some histone modification marks, reflecting existence of multiple states; but those genes in proximity show higher correlation than those in distance do. Experimentally one can again generate a cell line with reporter sequence being inserted to a specific gene, sort cells with high and low reporter signals and for each group perform CHIP-PCR analysis against histone marks under study for the tagged and other genes in the same DREM2 class. Using results with unsorted cells as references, for genes that are in proximity with the tagged gene we expect to observe significant correlation of histone mark levels between them, while for genes that are distal to the tagged gene there is no significant sign of additional correlation with respect to the reference.

In this study we focused on how spatial clustering affects histone modifications and its implications on transcriptional bursting and coordinated gene expression. It is possible that spatial clustering also affects other aspects of gene regulation such as enhancer-promoter interactions, which may also influence transcriptional bursting[320]. Clustering of genes regulated by a common TF also facilitates the TF to search for the targets. That is after it dissociates from one binding site, it can easily locate another one nearby. This so-called serial ligand rebinding mechanism has been discussed in the context of bacterial chemotaxis[321]. Since the typical target searching time for a TF is much shorter than the period between two transcriptional bursting events, we do not expect such mechanism contribute much to synchronize transcriptional events of two proximal genes.

4.4 SUPPLEMENTARY MATERIALS

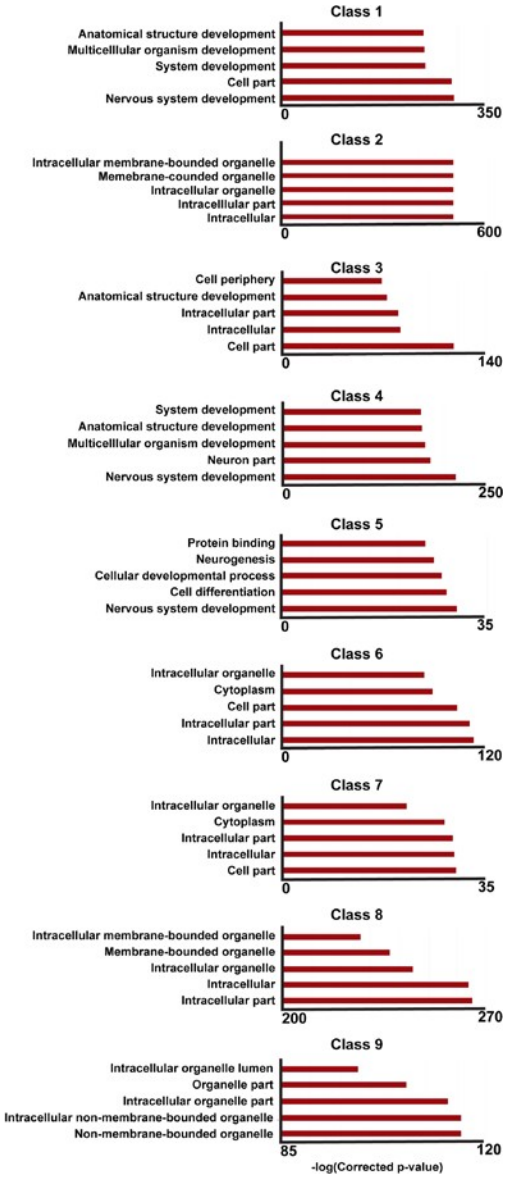

Figure 29 GO analysis of genes in every HMM classes.

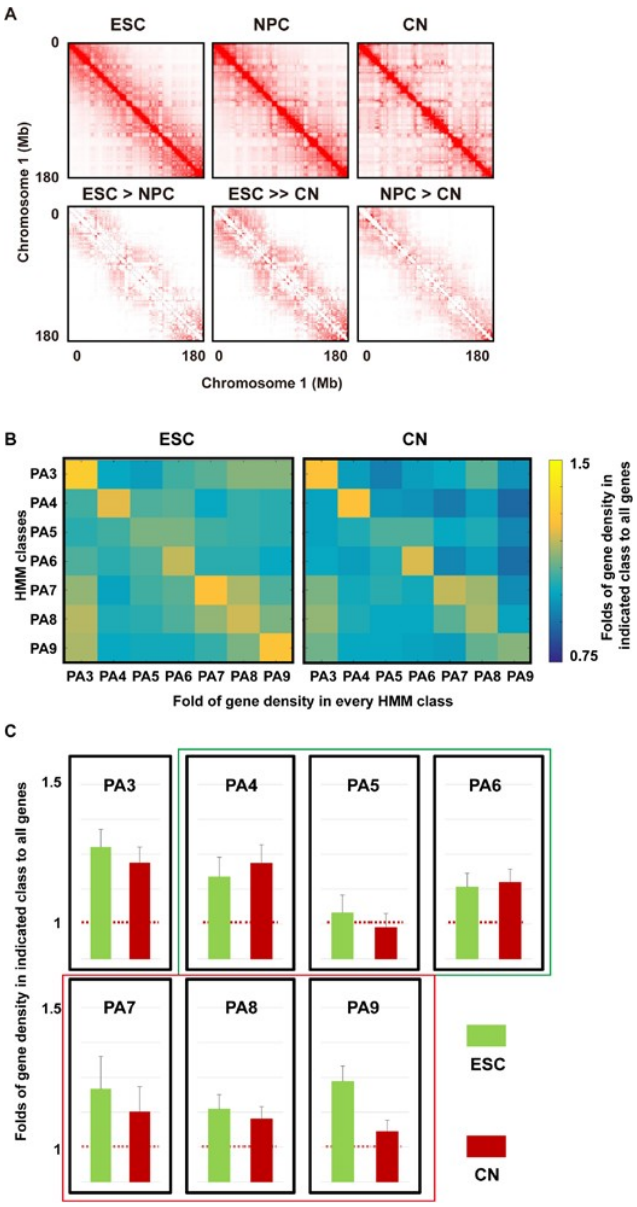

Figure 30 Dynamic change of chromosome structure during cell differentiation. **(A)** Top: Heatmaps from 250 kb-resolution HiC data of ESC, NPC, and CN cells, respectively. Bottom: Differences of the correlation between pair of bins. **(B)** Average density of genes within the same HMM class (diagonal) of the targeted gene or other HMM classes (off-diagonal) in the first shell of the radial distribution in ESC and CN. **(C)** Changes of average density of genes compared to all genes' density in the first shell of every HMM class from ESC to CN.

5.0 CONCLUSION

5.1 SUMMARY

In multicellular organisms, development starts from differentiation of totipotential stem cells step-by-step to functional cells. Cells need to maintain stable cell phenotypes, and be phenotypic plastic for development and in response to stress, where cells often need to choose one fate among others and undergo phenotype transition[22,197]. In response to signals, groups of cells are required to transit in a coordinated manner[47,322]. Cell fate decision and cell phenotype transition involve changes of almost every aspect of cell physiology, such as cell shape, cell movement, cell cycle, primary and secondary metabolism. These functional changes require coordinated synthesis and degradation of collaborative proteins, such as structure proteins, regulated proteins (TFs), and enzymes. Thus, coordinate regulation of multiple proteins is a key for cell phenotype transition.

In this dissertation, I presented mechanistic studies on gene regulation during cell phenotype transition at different levels. First, I investigated functions of the core TF network for cells to interpret the external signals. In this project, I used TGF-β induced EMT in MCF10A cells as the model system. The results show a nested TF network that can detect duration of external signals. Thus, cells remain as epithelial cells in response to short pulse signals, but have

flexibility to acquire mesenchymal features if the stimulation is strong and continuous. The nested network also helps in synchronizing cell responses. Thus, despite inherent cell-cell heterogeneity cells can respond together to fulfill specific functional requirements.

Next, to understand balanced expression levels of multiple gene regulation can be achieved from different regulatory levels, I used TGF-β induced EMT in MCF10A cells and cortical neuron differentiation from ESC in mouse as model systems, and analyzed combined data of gene expression (RNA-seq), DNA conformation (HiC), and histone modification (ChIP-seq). Our analysis showed that besides TFs binding or local histone modification levels, cells cluster some functionally related genes spatially together to coordinate their expression correlation. Specifically, single molecule and single cell studies reveal that most eukaryotic genes show stochastic transcriptional bursting, which is a major reason for heterogeneous gene expression among isogenic individual cells. By clustering spatially, functionally related genes share more common histone modification features compared to genes regulated by the same TF but reside at distant locations, and have more synchronized transcriptional bursting and correlated expression fluctuations.

5.2 FUTURE PERSPECTIVE

Stochasticity and determination are two important aspects for cell phenotype transition[4,323,324]. Stochasticity can lead to heterogeneity among a group of isogenic cells. This heterogeneity is concealed with experimental measurements at bulk levels, in which the observed quantities are population averaged. Thus the information within single cells is simply overlooked. Currently, with technological advances, one can extract quantitative information from single cells. In the

near future I expect that to see more exciting single cell studies on cell phenotypic transitions. Below I just briefly discuss a number of possible directions.

In my published single cell studies on EMT[13,325], we worked with fixed cells. While it is convenient to obtain a large number of samples, temporal correlation of an observable is missing with fixed cells. A complementary strategy is to track individual cells over time with life cell imaging. Fluorescence labeling of target proteins is often used in time-lapse imaging, but generating the labeled cell lines can be time consuming and technically challenging. Development of CRISPR-based gene editing techniques greatly changes such situation, and allows efficient inserting fluorescence protein sequences into the endogenous sites of targeted genes. I was involved in developing an efficient procedure of generating the knockin DNA constructs and cell lines[319]. The Xing lab is applying the technique for EMT studies. The lab is also using the technique to generate cell lines for MS2-based live cell imaging of mRNAs. For processes like EMT where cells undergo large morphological change, even tracing cell morphology over time with bright-field imaging can reveal much important information. I was involved in the project of developing an efficient deep learning algorithm for cell segmentation[326], which makes image analysis more convenient.

In Chapter III and IV I presented my analyses on chromosome structure and gene regulation. All the chromosome structure data are obtained from fixed cells. Recently several live cell chromosome labeling techniques have been developed. Notably the CRISPR system has been repurposed to label specific chromosome loci with catalytically deactivated Cas9 and RNA aptamers[327-335]. The Xing lab has improved this CRISPR-dCas9 chromosome labeling system, and I am collaborating with other group members to apply the technique to test the AP1 dynamics predicted in Chapter III.

Meanwhile, there also continuous improvements on single cell techniques with fixed cells. For example, the efficiency of high throughout single cell RNA-seq analysis has been dramatically improved while the cost being greatly reduced. Several recent studies reconstructed differentiation processes from single cell RNAseq measurements of tens of thousands of cells[336,337]. Similar, other traditional biochemical technologies also turn to single cell levels, such as single cell PCR, single cell ChIP-seq, and single cell Hi-C. All of those technologies are quickly advancing biomedical and basic life science researches.

BIBLIOGRAPHY

1 Linsenmayer, T. F., Toole, B. P. & Trelstad, R. L. Temporal and spatial transitions in collagen types during embryonic chick limb development. *Dev Biol* **35**, 232-239 (1973).

2 Hay, E. D. An overview of epithelio-mesenchymal transformation. *Cells Tissues Organs* **154**, 8-20 (1995).

3 Aird, W. Spatial and temporal dynamics of the endothelium. *J Thromb Haemost* **3**, 1392-1406 (2005).

4 Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* **12**, 323 (2012).

5 Mikhailov, K. V. *et al.* The origin of Metazoa: a transition from temporal to spatial cell differentiation. *Bioessays* **31**, 758-768 (2009).

6 Zavadil, J. & Böttinger, E. P. TGF-β and epithelial-to-mesenchymal transitions. *Oncogene* **24**, 5764 (2005).

7 Ibañez, C. *et al.* Ambient temperature and genotype differentially affect developmental and phenotypic plasticity in Arabidopsis thaliana. *BMC Plant Biol* **17**, 114 (2017).

8 Salabei, J. K. *et al.* PDGF-mediated autophagy regulates vascular smooth muscle cell phenotype and resistance to oxidative stress. *Biochem J* **451**, 375-388 (2013).

9 Kajita, M., McClinic, K. N. & Wade, P. A. Aberrant expression of the transcription factors snail and slug alters the response to genotoxic stress. *Mol Cell Biol* **24**, 7559-7566 (2004).

10 Wrana, J. L., Attisano, L., Wieser, R., Ventura, F. & Massagué, J. Mechanism of activation of the TGF-β receptor. *Nature* **370**, 341 (1994).

11 Sagner, A. & Briscoe, J. Morphogen interpretation: concentration, time, competence, and signaling dynamics. *Wiley Interdiscip Rev Dev Biol* **6**, e271 (2017).

12 Gaspard, N. *et al.* An intrinsic mechanism of corticogenesis from embryonic stem cells. *Nature* **455**, 351 (2008).

13 Zhang, J. *et al.* TGF-β–induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Sci Signal* **7**, ra91-ra91 (2014).

14 Nieto, M. A., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* **166**, 21-45 (2016).

15 Debnath, J. & Brugge, J. S. Modelling glandular epithelial cancers in three-dimensional cultures. *Nat Rev Cancer* **5**, 675 (2005).

16 Mather, J. P. & Roberts, P. E. *Introduction to cell and tissue culture: theory and technique*. (Springer Science & Business Media, 1998)

17 Mackenzie, T. C. & Flake, A. W. Human mesenchymal stem cells persist, demonstrate site- specific multipotential differentiation, and are present in sites of wound healing and tissue regeneration after transplantation into fetal sheep. *Blood Cells Mol Dis* **27**, 601-604 (2001).

18 Kidd, S. *et al.* Direct evidence of mesenchymal stem cell tropism for tumor and wounding microenvironments using in vivo bioluminescent imaging. *Stem cells* **27**, 2614-2623 (2009).

19 Hoogduijn, M. J. & Dor, F. J. Mesenchymal stem cells in transplantation and tissue regeneration. *Front Immunol* **2**, 84 (2011).

20 Hematti, P. & Keating, A. *Mesenchymal stromal cells: biology and clinical applications*. (Springer Science & Business Media, 2013).

21 Hay, E. D. Organization and fine structure of epithelium and mesenchyme in the developing chick embryo. *Epithelial-mesenchymal interactions* **2**, 31-35 (1968).

22 Thiery, J. P., Acloque, H., Huang, R. Y. J. & Nieto, M. A. Epithelial-Mesenchymal Transitions in Development and Disease. *Cell* **139**, 871-890 (2009).

23 ten Berge, D. *et al.* Wnt signaling mediates self-organization and axis formation in embryoid bodies. *Cell stem cell* **3**, 508-518 (2008).

24 Lindsley, R. C., Gill, J. G., Kyba, M., Murphy, T. L. & Murphy, K. M. Canonical Wnt signaling is required for development of embryonic stem cell-derived mesoderm. *Development* **133**, 3787- 3796 (2006).

25 Böttcher, R. T. & Niehrs, C. Fibroblast growth factor signaling during early vertebrate development. *Endocr Rev* **26**, 63-77 (2004).

26 Eastham, A. M. *et al.* Epithelial-mesenchymal transition events during human embryonic stem cell differentiation. *Cancer Res* **67**, 11254-11262 (2007).

27 Kasai, H., Allen, J. T., Mason, R. M., Kamimura, T. & Zhang, Z. TGF-β1 induces human alveolar epithelial to mesenchymal cell transition (EMT). *Respir Res* **6**, 56 (2005).

28 Boyer, A. S. *et al.* TGFβ2 and TGFβ3 have separate and sequential activities during epithelial– mesenchymal cell transformation in the embryonic heart. *Dev Biol* **208**, 530-545 (1999).

29 Arnoux, V., Nassour, M., L'Helgoualc'h, A., Hipskind, R. A. & Savagner, P. Erk5 controls Slug expression and keratinocyte activation during wound healing. *Mol Biol Cell* **19**, 4738-4749 (2008).

30 Carretero, M. *et al.* In vitro and in vivo wound healing-promoting activities of human cathelicidin LL-37. *J Invest Dermatol* **128**, 223-236 (2008).

31 Yan, C. *et al.* Epithelial to mesenchymal transition in human skin wound healing is induced by tumor necrosis factor-α through bone morphogenic protein-2. *Am J Pathol* **176**, 2247-2258 (2010).

32 Schäfer, M. & Werner, S. Cancer as an overhealing wound: an old hypothesis revisited. *Nat Rev Mol Cell Biol* **9**, 628 (2008).

33 Micalizzi, D. S., Farabaugh, S. M. & Ford, H. L. Epithelial-mesenchymal transition in cancer: parallels between normal development and tumor progression. *J Mammary Gland Biol Neoplasia* **15**, 117-134 (2010).

34 Weinberg, R. *The biology of cancer*. (Garland science, 2013).

35 Gupta, G. P. & Massagué, J. Cancer metastasis: building a framework. *Cell* **127**, 679-695 (2006).

36 Nieto, M. A. The ins and outs of the epithelial to mesenchymal transition in health and disease. *Annu Rev Cell Dev Biol* **27**, 347-376 (2011).

37 Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol* **15**, 178-196 (2014).

38 Yingling, J. M., Blanchard, K. L. & Sawyer, J. S. Development of TGF-β signalling inhibitors for cancer therapy. *Nat Rev Drug Discov* **3**, 1011 (2004).

39 Holland, J. D., Klaus, A., Garratt, A. N. & Birchmeier, W. Wnt signaling in stem and cancer stem cells. *Curr Opin Cell Biol* **25**, 254-264 (2013).

40 Sullivan, N. *et al.* Interleukin-6 induces an epithelial–mesenchymal transition phenotype in human breast cancer cells. *Oncogene* **28**, 2940 (2009).

41 Allen, M. & Louise Jones, J. Jekyll and Hyde: the role of the microenvironment on the progression of cancer. *J Pathol* **223**, 163-177 (2011).

42 Radisky, D. C. *et al.* Rac1b and reactive oxygen species mediate MMP-3-induced EMT and genomic instability. *Nature* **436**, 123 (2005).

43 Giannoni, E., Parri, M. & Chiarugi, P. EMT and oxidative stress: a bidirectional interplay affecting tumor malignancy. *Antioxid Redox Signal* **16**, 1248-1263 (2012).

44 Montserrat, N. *et al.* Epithelial to mesenchymal transition in early stage endometrioid endometrial carcinoma. *Hum Pathol* **43**, 632-643 (2012).

45 Lee, S.-H. *et al.* Blocking of p53-Snail binding, promoted by oncogenic K-Ras, recovers p53 expression and function. *Neoplasia* **11**, 22-31 (2009).

46 Munger, J. S. *et al.* Latent transforming growth factor-β: structural features and mechanisms of activation. *Kidney Int* **51**, 1376-1382 (1997).

47 Wu, M. Y. & Hill, C. S. TGF-β superfamily signaling in embryonic development and homeostasis. *Dev Cell* **16**, 329-343 (2009).

48 Zi, Z., Chapnick, D. A. & Liu, X. Dynamics of TGF-β/Smad signaling. *FEBS letters* **586**, 1921-1928 (2012).

49 Massagué, J. How cells read TGF-β signals. *Nat Rev Mol Cell Biol* **1**, 169-178 (2000).

50 Heldin, C.-H., Miyazono, K. & Ten Dijke, P. TGF-β signalling from cell membrane to nucleus through SMAD proteins. *Nature* **390**, 465 (1997).

51 Wu, J.-W., Fairman, R., Penry, J. & Shi, Y. Formation of a stable heterodimer between Smad2 and Smad4. *J Biol Chem* **276**, 20688-20694 (2001).

52 Chacko, B. M. *et al.* The L3 loop and C-terminal phosphorylation jointly define Smad protein trimerization. *Nat Struct Mol Biol* **8**, 248 (2001).

53 Chacko, B. M. *et al.* Structural basis of heteromeric Smad protein assembly in TGF-β signaling. *Mol cell* **15**, 813-823 (2004).

54 Massagué, J., Seoane, J. & Wotton, D. Smad transcription factors. *Genes Dev* **19**, 2783-2810 (2005).

55 Thuault, S. *et al.* HMGA2 and Smads co-regulate SNAIL1 expression during induction of epithelial-to-mesenchymal transition. *J Biol Chem* **283**, 33437-33446 (2008).

56 Vincent, T. *et al.* A SNAIL1–SMAD3/4 transcriptional repressor complex promotes TGF-β mediated epithelial–mesenchymal transition. *Nat Cell Biol* **11**, 943 (2009).

57 Brandl, M. *et al.* IKKα controls canonical TGFβ–SMAD signaling to regulate genes expressing SNAIL and SLUG during EMT in Panc1 cells. *J Cell Sci* **123**, 4231-4239 (2010).

58 Choi, J., Park, S. Y. & Joo, C.-K. Transforming growth factor-β1 represses E-cadherin production via Slug expression in lens epithelial cells. *Invest Ophthalmol Vis Sci* **48**, 2708-2718 (2007).

59 Nakao, A. *et al.* Identification of Smad7, a TGFβ-inducible antagonist of TGF-β signalling. *Nature* **389**, 631 (1997).

60 Wang, H., Song, K., Krebs, T. L., Yang, J. & Danielpour, D. Smad7 is inactivated through a direct physical interaction with the LIM protein Hic-5/ARA55. *Oncogene* **27**, 6791 (2008).

61 Singh, P., Srinivasan, R., Wig, J. D. & Radotra, B. D. A study of Smad4, Smad6 and Smad7 in surgically resected samples of pancreatic ductal adenocarcinoma and their correlation with clinicopathological parameters and patient survival. *BMC Res Notes* **4**, 560 (2011).

62 Xu, J., Lamouille, S. & Derynck, R. TGF-β-induced epithelial to mesenchymal transition. *Cell Res* **19**, 156-172 (2009).

63 Derynck, R. & Zhang, Y. E. Smad-dependent and Smad-independent pathways in TGF-β family signalling. *Nature* **425**, 577 (2003).

64 Derynck, R., Akhurst, R. J. & Balmain, A. TGF-β signaling in tumor suppression and cancer progression. *Nat Genet* **29**, 117 (2001).

65 Kamaraju, A. K. & Roberts, A. B. Role of Rho/ROCK and p38 MAP kinase pathways in transforming growth factor-β-mediated Smad-dependent growth inhibition of human breast carcinoma cells in vivo. *J Biol Chem* **280**, 1024-1036 (2005).

66 Kanamaru, C., Yasuda, H. & Fujita, T. Involvement of Smad proteins in TGF-β and activin A- induced apoptosis and growth inhibition of liver cells. *Hepatol Res* **23**, 211-219 (2002).

67 Liu, X. *et al.* Transforming growth factor β-induced phosphorylation of Smad3 is required for growth inhibition and transcriptional induction in epithelial cells. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 10669-10674 (1997).

68 Fleming, N. I. *et al.* SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. *Cancer Res* **73**, 725-735 (2013).

69 Ungefroren, H. *et al.* Differential roles of Smad2 and Smad3 in the regulation of TGF-β1-mediated growth inhibition and cell migration in pancreatic ductal adenocarcinoma cells: control by Rac1. *Mol cancer* **10**, 67 (2011).

70 Yakicier, M., Irmak, M., Romano, A., Kew, M. & Ozturk, M. Smad2 and Smad4 gene mutations in hepatocellular carcinoma. *Oncogene* **18**, 4879 (1999).

71 Riggins, G. J., Kinzler, K. W., Vogelstein, B. & Thiagalingam, S. Frequency of Smad gene mutations in human cancers. *Cancer Res* **57**, 2578-2580 (1997).

72 Miyaki, M. & Kuroki, T. Role of Smad4 (DPC4) inactivation in human cancer. *Biochem Biophys Res Commun* **306**, 799-804 (2003).

73 Chen, H.-S., Bai, M.-H., Zhang, T., Li, G.-D. & Liu, M. Ellagic acid induces cell cycle arrest and apoptosis through TGF-β/Smad3 signaling pathway in human breast cancer MCF-7 cells. *Int J Oncol* **46**, 1730-1738 (2015).

74 Cheng, J.-C., Auersperg, N. & Leung, P. C. TGF-beta induces serous borderline ovarian tumor cell invasion by activating EMT but triggers apoptosis in low-grade serous ovarian carcinoma cells. *PloS One* **7**, e42436 (2012).

75 Massagué, J. TGFβ signalling in context. *Nat Rev Mol Cell Biol* **13**, 616-630 (2012).

76 Ho, J. *et al.* Activin induces hepatocyte cell growth arrest through induction of the cyclin-dependent kinase inhibitor p15INK4B and Sp1. *Cell Signal* **16**, 693-701 (2004).

77 Kim, Y. K. *et al.* Cooperation of H2O2-mediated ERK activation with Smad pathway in TGF-β1 induction of p21WAF1/Cip1. *Cell Signal* **18**, 236-243 (2006).

78 Shi, Y. & Massagué, J. Mechanisms of TGF-β signaling from cell membrane to the nucleus. *Cell* **113**, 685-700 (2003).

79 Fuxe, J., Vincent, T. & Garcia de Herreros, A. Transcriptional crosstalk between TGFβ and stem cell pathways in tumor cell invasion: role of EMT promoting Smad complexes. *Cell cycle* **9**, 2363- 2374 (2010).

80 Chen, X.-F. *et al.* Transforming growth factor-β1 induces epithelial-to-mesenchymal transition in human lung cancer cells via PI3K/Akt and MEK/Erk1/2 signaling pathways. *Mol Biol Rep* **39**, 3549-3556 (2012).

81 Zhang, Y. E. Non-Smad pathways in TGF-β signaling. *Cell Res* **19**, 128 (2009).

82 Bakin, A. V., Tomlinson, A. K., Bhowmick, N. A., Moses, H. L. & Arteaga, C. L. Phosphatidylinositol 3-kinase function is required for transforming growth factor β-mediated epithelial to mesenchymal transition and cell migration. *J Biol Chem* **275**, 36803-36810 (2000).

83 Lamouille, S. & Derynck, R. Cell size and invasion in TGF-β–induced epithelial to mesenchymal transition is regulated by activation of the mTOR pathway. *J Cell Biol* **178**, 437-451 (2007).

84 Shin, I., Bakin, A. V., Rodeck, U., Brunet, A. & Arteaga, C. L. Transforming growth factor β enhances epithelial cell survival via Akt-dependent regulation of FKHRL1. *Mol Biol Cell* **12**, 3328- 3339 (2001).

85 Song, K., Wang, H., Krebs, T. L. & Danielpour, D. Novel roles of Akt and mTOR in suppressing TGFβ/ALK5-mediated Smad3 activation. *EMBO J* **25**, 58-69 (2006).

86 Masszi, A. *et al.* Central role for Rho in TGF-β1-induced α-smooth muscle actin expression during epithelial-mesenchymal transition. *Am J Physiol Renal Physiol* **284**, F911-F924 (2003).

87 Mu, Y., Gudey, S. K. & Landström, M. Non-Smad signaling pathways. *Cell Tissue Res* **347**, 11-20 (2012).

88 Yamamura, Y., Hua, X., Bergelson, S. & Lodish, H. F. Critical role of Smads and AP-1 complex in transforming growth factor-β-dependent apoptosis. *J Biol Chem* **275**, 36295-36302 (2000).

89 Gal, A. *et al.* Sustained TGFβ exposure suppresses Smad and non-Smad signalling in mammary epithelial cells, leading to EMT and inhibition of growth arrest and apoptosis. *Oncogene* **27**, 1218 (2008).

90 Datto, M. B. *et al.* Transforming growth factor beta induces the cyclin-dependent kinase inhibitor p21 through a p53-independent mechanism. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 5545-5549 (1995).

91 Barcellos-Hoff, M. H. & Akhurst, R. J. Transforming growth factor-β in breast cancer: too much, too late. *Breast Cancer Res* **11**, 202 (2009).

92 Arteaga, C. L. Inhibition of TGFβ signaling in cancer therapy. *Curr Opin Genet Dev* **16**, 30-37 (2006).

93 Terabe, M. *et al.* Synergistic enhancement of CD8+ T cell–mediated tumor vaccine efficacy by an anti–transforming growth factor-β monoclonal antibody. *Clin Cancer Res* **15**, 6560-6569 (2009).

94 Morris, J. C. *et al.* Phase I study of GC1008 (fresolimumab): a human anti-transforming growth factor-beta (TGFβ) monoclonal antibody in patients with advanced malignant melanoma or renal cell carcinoma. *PloS One* **9**, e90353 (2014).

95 Han, J., Alvarez-Breckenridge, C. A., Wang, Q.-E. & Yu, J. TGF-β signaling and its targeting for glioma treatment. *Am J Cancer Res* **5**, 945 (2015).

96 Schlingensiepen, K. H. *et al.* Transforming growth factor-beta 2 gene silencing with trabedersen (AP 12009) in pancreatic cancer. *Cancer Sci* **102**, 1193-1200 (2011).

97 Tanaka, H. *et al.* Transforming growth factor β signaling inhibitor, SB-431542, induces maturation of dendritic cells and enhances anti-tumor activity. *Oncol Rep* **24**, 1637-1643 (2010).

98 Wendt, M. K., Smith, J. A. & Schiemann, W. P. Transforming growth factor-β-induced epithelial– mesenchymal transition facilitates epidermal growth factor-dependent breast cancer progression. *Oncogene* **29**, 6485 (2010).

99 Zhao, B. M. & Hoffmann, F. M. Inhibition of Transforming Growth Factor-β1–induced Signaling and Epithelial-to-Mesenchymal Transition by the Smad-binding Peptide Aptamer Trx-SARA. *Mol Biol Cell* **17**, 3819-3831 (2006).

100 Yoo, Y. A. *et al.* Sonic hedgehog pathway promotes metastasis and lymphangiogenesis via activation of Akt, EMT, and MMP-9 pathway in gastric cancer. *Cancer Res* **71**, 7061-7070 (2011).

101 Xu, X. *et al.* Genome-wide screening reveals an EMT molecular network mediated by Sonic hedgehog-Gli1 signaling in pancreatic cancer cells. *PLoS One* **7**, e43119 (2012).

102 Islam, S. *et al.* Sonic hedgehog (Shh) signaling promotes tumorigenicity and stemness via activation of epithelial-to-mesenchymal transition (EMT) in bladder cancer. *Mol Carcinog* **55**, 537-551 (2016).

103 Chen, M.-H., Li, Y.-J., Kawakami, T., Xu, S.-M. & Chuang, P.-T. Palmitoylation is required for the production of a soluble multimeric Hedgehog protein complex and long-range signaling in vertebrates. *Genes Dev* **18**, 641-659 (2004).

104 Jiang, J. & Hui, C.-c. Hedgehog signaling in development and cancer. *Dev Cell* **15**, 801-812 (2008).

105 Takebe, N., Harris, P. J., Warren, R. Q. & Ivy, S. P. Targeting cancer stem cells by inhibiting Wnt, Notch, and Hedgehog pathways. *Nat Rev Clin Oncol* **8**, 97 (2011).

106 Stone, D. M. *et al.* The tumour-suppressor gene patched encodes a candidate receptor for Sonic hedgehog. *Nature* **384**, 129 (1996).

107 Huangfu, D. & Anderson, K. V. Signaling from Smo to Ci/Gli: conservation and divergence of Hedgehog pathways from Drosophila to vertebrates. *Development* **133**, 3-14 (2006).

108 Murone, M., Rosenthal, A. & de Sauvage, F. J. Sonic hedgehog signaling by the patched–smoothened receptor complex. *Curr Biol* **9**, 76-84 (1999).

109 Roessler, E. *et al.* A previously unidentified amino-terminal domain regulates transcriptional activity of wild-type and disease-associated human GLI2. *Hum Mol Genet* **14**, 2181-2188 (2005).

110 Sasaki, H., Nishizaki, Y., Hui, C.-c., Nakafuku, M. & Kondoh, H. Regulation of Gli2 and Gli3 activities by an amino-terminal repression domain: implication of Gli2 and Gli3 as primary mediators of Shh signaling. *Development* **126**, 3915-3924 (1999).

111 Varjosalo, M. & Taipale, J. Hedgehog: functions and mechanisms. *Genes Dev* **22**, 2454-2472 (2008).

112 Regl, G. *et al.* Human GLI2 and GLI1 are part of a positive feedback mechanism in Basal Cell Carcinoma. *Oncogene* **21**, 5529 (2002).

113 Amakye, D., Jagani, Z. & Dorsch, M. Unraveling the therapeutic potential of the Hedgehog pathway in cancer. *Nat Med* **19**, 1410 (2013).

114 Katoh, Y. & Katoh, M. Hedgehog target genes: mechanisms of carcinogenesis induced by aberrant hedgehog signaling activation. *Curr Mol Med* **9**, 873-886 (2009).

115 Wang, X. *et al.* Sonic hedgehog regulates Bmi1 in human medulloblastoma brain tumor-initiating cells. *Oncogene* **31**, 187 (2012).

116 Zbinden, M. *et al.* NANOG regulates glioma stem cells and is essential in vivo acting in a cross-functional network with GLI1 and p53. *EMBO J* **29**, 2659-2674 (2010).

117 Li, X. *et al.* Snail induction is an early response to Gli1 that determines the efficiency of epithelial transformation. *Oncogene* **25**, 609 (2006).

118 Li, X., Deng, W., Lobo-Ruppert, S. & Ruppert, J. Gli1 acts through Snail and E-cadherin to promote nuclear signaling by β-catenin. *Oncogene* **26**, 4489 (2007).

119 Villavicencio, E. H. *et al.* Cooperative E-box regulation of human GLI1 by TWIST and USF. *genesis* **32**, 247-258 (2002).

120 Dennler, S. *et al.* Induction of sonic hedgehog mediators by transforming growth factor-β: Smad3-dependent activation of Gli2 and Gli1 expression in vitro and in vivo. *Cancer Res* **67**, 6981-6986 (2007).

121 Dennler, S., André, J., Verrecchia, F. & Mauviel, A. Cloning of the human GLI2 promoter transcriptional activation by transforming growth factor-β via Smad3/β-catenin cooperation. *J Biol Chem* **284**, 31523-31531 (2009).

122 Javelaud, D. *et al.* TGF-β/SMAD/GLI2 signaling axis in cancer progression and metastasis. *Cancer Res* **71**, 5606-5610 (2011).

123 Yang, L., Xie, G., Fan, Q. & Xie, J. Activation of the hedgehog-signaling pathway in human cancer and the clinical implications. *Oncogene* **29**, 469 (2010).

124 Bian, X.-h. *et al.* Expression and clinical significance of Shh/Gli-1 in papillary thyroid carcinoma. *Tumour Biol* **35**, 10523-10528 (2014).

125 Yue, D. *et al.* Hedgehog/Gli promotes epithelial-mesenchymal transition in lung squamous cell carcinomas. *J Exp Clin Cancer Res* **33**, 34 (2014).

126 Behnsawy, H. M. *et al.* Possible role of sonic hedgehog and epithelial-mesenchymal transition in renal cell cancer progression. *Korean J Urol* **54**, 547-554 (2013).

127 Bermudez, O., Hennen, E., Koch, I., Lindner, M. & Eickelberg, O. Gli1 mediates lung cancer cell proliferation and Sonic Hedgehog-dependent mesenchymal cell activation. *PloS One* **8**, e63226 (2013).

128 i Altaba, A. R., Sánchez, P. & Dahmane, N. Gli and hedgehog in cancer: tumours, embryos and stem cells. *Nat Rev Cancer* **2**, 361 (2002).

129 Feldmann, G. *et al.* Blockade of hedgehog signaling inhibits pancreatic cancer invasion and metastases: a new paradigm for combination therapy in solid cancers. *Cancer Res* **67**, 2187-2196 (2007).

130 LoRusso, P. M. *et al.* Phase I trial of hedgehog pathway inhibitor vismodegib (GDC-0449) in patients with refractory, locally advanced or metastatic solid tumors. *Clin Cancer Res* **17**, 2502- 2511 (2011).

131 Kaye, S. B. *et al.* A phase II, randomized, placebo-controlled study of vismodegib as maintenance therapy in patients with ovarian cancer in second or third complete remission. *Clin Cancer Res*, clincanres. 1796.2012 (2012).

132 Singh, B. N., Fu, J., Srivastava, R. K. & Shankar, S. Hedgehog signaling antagonist GDC-0449 (Vismodegib) inhibits pancreatic cancer stem cell characteristics: molecular mechanisms. *PloS One* **6**, e27306 (2011).

133 Lauth, M., Bergström, Å., Shimokawa, T. & Toftgård, R. Inhibition of GLI-mediated transcription and tumor cell growth by small-molecule antagonists. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 8455-8460 (2007).

134 Peifer, M. & Polakis, P. Wnt signaling in oncogenesis and embryogenesis--a look outside the nucleus. *Science* **287**, 1606-1609 (2000).

135 Zhou, B. P. & Hung, M.-C. Wnt, hedgehog, and snail: sister pathways that control by GSK-3beta and beta-Trcp in the regulation of metastasis. *Cell Cycle* **4**, 772-776 (2005).

136 Fodde, R. & Brabletz, T. Wnt/β-catenin signaling in cancer stemness and malignant behavior. *Curr Opin Cell Biol* **19**, 150-158 (2007).

137 Easwaran, V., Pishvaian, M. & Byers, S. Cross-regulation of β-catenin–LEF/TCF and retinoid signaling pathways. *Curr Biol* **9**, 1415-1419 (1999).

138 Stemmer, V., De Craene, B., Berx, G. & Behrens, J. Snail promotes Wnt target gene expression and interacts with β-catenin. *Oncogene* **27**, 5075 (2008).

139 Orsulic, S., Huber, O., Aberle, H., Arnold, S. & Kemler, R. E-cadherin binding prevents beta- catenin nuclear localization and beta-catenin/LEF-1-mediated transactivation. *J Cell Sci* **112**, 1237-1245 (1999).

140 Yook, J. I. *et al.* A Wnt–Axin2–GSK3β cascade regulates Snail1 activity in breast cancer cells. *Nat Cell Biol* **8**, 1398 (2006).

141 Mizuarai, S., Kawagishi, A. & Kotani, H. Inhibition of p70S6K2 down-regulates Hedgehog/GLI pathway in non-small cell lung cancer cell lines. *Mol Cancer* **8**, 44 (2009).

142 Takenaka, K., Kise, Y. & Miki, H. GSK3β positively regulates Hedgehog signaling through Sufu in mammalian cells. *Biochem Bioph Res Co* **353**, 501-508 (2007).

143 Noubissi, F. K. *et al.* Wnt signaling stimulates transcriptional outcome of the Hedgehog pathway by stabilizing GLI1 mRNA. *Cancer Res* **69**, 8572-8578 (2009).

144 Caraci, F. *et al.* TGF-β1 targets the GSK-3β/β-catenin pathway via ERK activation in the transition of human lung fibroblasts into myofibroblasts. *Pharmacol Res* **57**, 274-282 (2008).

145 Guo, X. *et al.* Axin and GSK3-β control Smad3 protein stability and modulate TGF-β signaling. *Genes Dev* **22**, 106-120 (2008).

146 Takahashi-Yanaga, F. & Kahn, M. Targeting Wnt signaling: can we safely eradicate cancer stem cells? *Clin Cancer Res* **16**, 3153-3162 (2010).

147 Kahn, M. Can we safely target the WNT pathway? *Nat Rev Drug Discov* **13**, 513 (2014).

148 Revenu, C. & Gilmour, D. EMT 2.0: shaping epithelia through collective migration. *Curr Opin Genet Dev* **19**, 338-342 (2009).

149 Tam, W. L. & Weinberg, R. A. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nat Med* **19**, 1438 (2013).

150 Tian, X.-J., Zhang, H. & Xing, J. Coupled Reversible and Irreversible Bistable Switches Underlying TGFβ-induced Epithelial to Mesenchymal Transition. *Biophys J* **105**, 1079-1089 (2013).

151 Siemens, H. *et al.* miR-34 and SNAIL form a double-negative feedback loop to regulate epithelial- mesenchymal transitions. *Cell cycle* **10**, 4256-4271 (2011).

152 Nam, S. *et al.* PATHOME: an algorithm for accurately detecting differentially expressed subpathways. *Oncogene* **33**, 4941 (2014).

153 Behar, M., Barken, D., Werner, S. L. & Hoffmann, A. The dynamics of signaling as a pharmacological target. *Cell* **155**, 448-461 (2013).

154 Lee, M. J. *et al.* Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell* **149**, 780-794 (2012).

155 Kreeger, P. K. & Lauffenburger, D. A. Cancer systems biology: a network modeling perspective. *Carcinogenesis* **31**, 2-8 (2009).

156 Bakan, A. *et al.* Inhibition of peroxidase activity of cytochrome c: De novo compound discovery and validation. *Mol Pharmacol* **88**, 421-427 (2015).

157 Kholodenko, B. N. Cell-signalling dynamics in time and space. *Nat Rev Mol Cell Biol* **7**, 165-176 (2006).

158 Mishina, Y. & Snider, T. N. Neural crest cell signaling pathways critical to cranial bone development and pathology. *Exp Cell Res* **325**, 138-147 (2014).

159 Lee, J. M., Dedhar, S., Kalluri, R. & Thompson, E. W. The epithelial–mesenchymal transition: new insights in signaling, development, and disease. *J Cell biol* **172**, 973-981 (2006).

160 Campisi, J. Senescent cells, tumor suppression, and organismal aging: good citizens, bad neighbors. *Cell* **120**, 513-522 (2005).

161 Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292 (2001).

162 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* **326**, 289-293 (2009).

163 Xie, W. J. *et al.* Structural modeling of chromatin integrates genome features and reveals chromosome folding principle. *Sci Rep* **7**, 2818 (2017).

164 Criscione, S. W. *et al.* Reorganization of chromosome architecture in replicative cellular senescence. *Sci Adv* **2**, e1500882 (2016).

165 Flyamer, I. M. *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to- zygote transition. *Nature* **544**, 110 (2017).

166 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331 (2015).

167 Jost, K. L., Bertulat, B. & Cardoso, M. C. Heterochromatin and gene positioning: inside, outside, any side? *Chromosoma* **121**, 555-563 (2012).

168 Sarma, N. P. & Natarajan, A. Identification of heterochromatic regions in the chromosomes of rye. *Hereditas* **74**, 233-238 (1973).

169 Le Gros, M. A. *et al.* Soft X-ray tomography reveals gradual chromatin compaction and reorganization during neurogenesis in vivo. *Cell Rep* **17**, 2125-2136 (2016).

170 Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat Rev Mol Cell Biol* **16**, 245 (2015).

171 Dekker, J. & Heard, E. Structural and functional diversity of topologically associating domains. *FEBS letters* **589**, 2877-2884 (2015).

172 Barutcu, A. R. *et al.* Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol* **16**, 214 (2015).

173 Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110 (2016).

174 Novo, C. L. *et al.* Long-Range Enhancer Interactions Are Prevalent in Mouse Embryonic Stem Cells and Are Reorganized upon Pluripotent State Transition. *Cell Rep* **22**, 2615-2627 (2018).

175 Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **172** (2018).

176 Fraser, J. *et al.* Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol* **11**, 852 (2015).

177 Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**, 425 (2007).

178 Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33**, 245 (2003).

179 Boland, M. J., Nazor, K. L. & Loring, J. F. Epigenetic regulation of pluripotency and differentiation. *Circ Res* **115**, 311-324 (2014).

180 Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-257 (1999).

181 Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**, 6-21 (2002).

182 Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* **39**, 457 (2007).

183 Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**, 853 (2005).

184 Razin, A. CpG methylation, chromatin structure and gene silencing a three-way connection. *EMBO J* **17**, 4905-4908 (1998).

185 Hashimshony, T., Zhang, J., Keshet, I., Bustin, M. & Cedar, H. The role of DNA methylation in setting up chromatin structure during development. *Nat Genet* **34**, 187 (2003).

186 Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253 (2010).

187 Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705 (2007).

188 Rosenfeld, J. A. *et al.* Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics* **10**, 143 (2009).

189 Miniou, P. *et al.* Abnormal methylation pattern in constitutive and facultative (X inactive chromosome) heterochromatin of ICF patients. *Hum Mol Genet* **3**, 2093-2102 (1994).

190 Jeppesen, P. & Turner, B. M. The inactive X chromosome in female mammals is distinguished by a lack of histone H4 acetylation, a cytogenetic marker for gene expression. *Cell* **74**, 281-289 (1993).

191 Costanzi, C. & Pehrson, J. R. Histone macroH2A1 is concentrated in the inactive X chromosome of female mammals. *Nature* **393**, 599 (1998).

192 Becker, J. S., Nicetto, D. & Zaret, K. S. H3K9me3-dependent heterochromatin: barrier to cell fate changes. *Trends Genet* **32**, 29-41 (2016).

193 Gates, L. A. *et al.* Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. *J Biol Chem* **292**, 14456-14472 (2017).

194 Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279 (2011).

195 Choukrallah, M.-A., Song, S., Rolink, A. G., Burger, L. & Matthias, P. Enhancer repertoires are reshaped independently of early priming and heterochromatin dynamics during B cell differentiation. *Nat Commun* **6**, 8324 (2015).

196 Zhang, T., Cooper, S. & Brockdorff, N. The interplay of histone modifications–writers that read. *EMBO Rep* **16**, 1467-1481 (2015).

197 Atlasi, Y. & Stunnenberg, H. G. The interplay of epigenetic marks during stem cell differentiation and development. *Nat Rev Genet* **18**, 643 (2017).

198 Dorighi, K. M. *et al.* Mll3 and Mll4 facilitate enhancer RNA synthesis and transcription from promoters independently of H3K4 monomethylation. *Mol cell* **66**, 568-576. e564 (2017).

199 Local, A. *et al.* Identification of H3K4me1-associated proteins at mammalian enhancers. *Nat Genet* **50**, 73 (2018).

200 Rickels, R. *et al.* Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nat Genet* **49**, 1647 (2017).

201 Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77-88 (2007).

202 Lauberth, S. M. *et al.* H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell* **152**, 1021-1036 (2013).

203 Grandy, R. A. *et al.* Genome-wide studies reveal that H3K4me3 modification in bivalent genes is dynamically regulated during the pluripotent cell cycle and stabilized upon differentiation. *Mol Cell Biol* **36**, 615-627 (2016).

204 Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343 (2011).

205 Tie, F. *et al.* CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development* **136**, 3131-3141 (2009).

206 Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16**, 167 (2015).

207 Wu, W.-S. *et al.* Snail collaborates with EGR-1 and SP-1 to directly activate transcription of MMP 9 and ZEB1. *Sci Rep* **7**, 17753 (2017).

208 Christofori, G. Snail1 links transcriptional control with epigenetic regulation. *EMBO J* **29**, 1787- 1789 (2010).

209 Peinado, H., Ballestar, E., Esteller, M. & Cano, A. Snail mediates E-cadherin repression by the recruitment of the Sin3A/histone deacetylase 1 (HDAC1)/HDAC2 complex. *Mol Cell Biol* **24**, 306- 319 (2004).

210 Harney, A. S., Meade, T. J. & LaBonne, C. Targeted inactivation of Snail family EMT regulatory factors by a Co (III)-Ebox conjugate. *PLoS One* **7**, e32318 (2012).

211 Batlle, E. *et al.* The transcription factor snail is a repressor of E-cadherin gene expression in epithelial tumour cells. *Nat Cell Biol* **2**, 84 (2000).

212 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311 (2002).

213 Jackson, V. Studies on histone organization in the nucleosome using formaldehyde as a reversible cross-linking agent. *Cell* **15**, 945-954 (1978).

214 Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nat Genet* **38**, 1348 (2006).

215 Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* **16**, 1299-1309 (2006).

216 O'Geen, H., Echipare, L. & Farnham, P. J. in *Epigenetics Protocols* 265-286 (Springer, 2011).

217 Down, T. A. *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* **26**, 779 (2008).

218 Komatsu, Y. *et al.* Global analysis of DNA methylation in early-stage liver fibrosis. *BMC Med Genomics* **5**, 5 (2012).

219 Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669 (2009).

220 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854 (2009).

221 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213 (2013).

222 Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322 (2008).

223 Cooper, S. *et al.* Predicting protein structures with a multiplayer online game. *Nature* **466**, 756 (2010).

224 Kawrykow, A. *et al.* Phylo: a citizen science approach for improving multiple sequence alignment. *PloS One* **7**, e31362 (2012).

225 Albeck, J. G., Mills, G. B. & Brugge, J. S. Frequency-modulated pulses of ERK activity transmit quantitative proliferation signals. *Mol Cell* **49**, 249-26 (2013).

226 English, J. G. *et al.* MAPK feedback encodes a switch and timer for tunable stress adaptation in yeast. *Sci Signal* **8**, ra5 (2015).

227 Fu, Y. *et al.* Reciprocal encoding of signal intensity and duration in a glucose-sensing circuit. *Cell* **156**, 1084-1095 (2014).

228 Behar, M. & Hoffmann, A. Understanding the temporal codes of intra-cellular signals. *Curr Opin Genet Dev* **20**, 684-693 (2010).

229 Murphy, L. O., Smith, S., Chen, R.-H., Fingar, D. C. & Blenis, J. Molecular interpretation of ERK signal duration by immediate early gene products. *Nat Cell Biol* **4**, 556-564 (2002).

230 Locasale, J. W. Signal duration and the time scale dependence of signal integration in biochemical pathways. *BMC Syst Biol* **2**, 108 (2008).

231 Purvis, Jeremy E. & Lahav, G. Encoding and Decoding Cellular Information through Signaling Dynamics. *Cell* **152**, 945-956 (2013).

232 Selimkhanov, J. *et al.* Accurate information transmission through dynamic biochemical signaling networks. *Science* **346**, 1370-1373 (2014).

233 Cheong, R., Rhee, A., Wang, C. J., Nemenman, I. & Levchenko, A. Information transduction capacity of noisy biochemical signaling networks. *Science* **334**, 354-358 (2011).

234 Kellogg, R. A., Tian, C., Lipniacki, T., Quake, S. R. & Tay, S. Digital signaling decouples activation probability and population heterogeneity. *eLife* **4**, e08931 (2015).

235 Rallis, A., Moore, C. & Ng, J. Signal strength and signal duration define two distinct aspects of JNK-regulated axon stability. *Dev Biol* **339**, 65-77 (2010).

236 Macagno, A., Napolitani, G., Lanzavecchia, A. & Sallusto, F. Duration, combination and timing: the signal integration model of dendritic cell activation. *Trends Immunol* **28**, 227-233 (2007).

237 Marshall, C. J. Specificity of receptor tyrosine kinase signaling: Transient versus sustained extracellular signal-regulated kinase activation. *Cell* **80**, 179-185 (1995).

238 Santos, S. D. M., Verveer, P. J. & Bastiaens, P. I. H. Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nat Cell Biol* **9**, 324-330 (2007).

239 Sasagawa, S., Ozaki, Y.-i., Fujita, K. & Kuroda, S. Prediction and validation of the distinct dynamics of transient and sustained ERK activation. *Nat Cell Biol* **7**, 365-373 (2005).

240 Borthwick, L. A. & Wynn, T. A. IL-13 and TGF-β1: core mediators of fibrosis. *Curr Pathobiol Rep* **3** (2015).

241 Fabregat, I., Fernando, J., Mainez, J. & Sancho, P. TGF-β signaling in cancer treatment. *Curr Pharm Des* **20**, 2934-2947 (2014).

242 Colak, S. & ten Dijke, P. Targeting TGF-β signaling in cancer. *Trends Cancer* **3**, 56-71 (2017). Zi, Z. *et al.* Quantitative analysis of transient and sustained transforming growth factor-β signaling dynamics. *Mol Syst Biol* **7**, 492 (2011).

243 Warmflash, A. *et al.* Dynamics of TGF-β signaling reveal adaptive and pulsatile behaviors reflected in the nuclear localization of transcription factor Smad4. *Proc Nat Acad Scis USA* **109**, E1947-1956 (2012).

244 Vizán, P. *et al.* Controlling long-term signaling: receptor dynamics determine attenuation and refractory behavior of the TGF-β pathway. *Sci Signal* **6**, ra106 (2013).

245 Sorre, B., Warmflash, A., Brivanlou, Ali H. & Siggia, Eric D. Encoding of Temporal Signals by the TGF-β Pathway and Implications for Embryonic Patterning. *Dev Cell* **30**, 334-342 (2014).

246 Frick, C. L., Yarka, C., Nunns, H. & Goentoro, L. Sensing relative signal in the Tgf-β/Smad pathway. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E2975- E2982 (2017).

247 Dennler, S. *et al.* Induction of sonic hedgehog mediators by transforming growth factor-β: Smad3-dependent activation of Gli2 and Gli1 expression in vitro and in vivo. *Cancer Res* **67**, 6981-6986 (2007).

248 Zhang, J., Tian, X.-J. & Xing, J. Signal transduction pathways of EMT induced by TGF-β, SHH, and WNT and their crosstalks. *J Clin Med Res* **5**, 41 (2016).

249 Schmierer, B., Tournier, A. L., Bates, P. A. & Hill, C. S. Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system. *Proceedings of the National Academy of Sciences of the United States of America* **105** (2008).

250 Yuki, K., Yoshida, Y., Inagaki, R., Hiai, H. & Noda, M. E-cadherin–downregulation and RECK- upregulation are coupled in the non-malignant epithelial cell line MCF10A but not in multiple carcinoma-derived cell lines. *Sci Rep* **4** (2014).

251 Steinway, S. N. *et al.* Network modeling of TGFβ signaling in hepatocellular carcinoma epithelial- to-mesenchymal transition reveals joint sonic hedgehog and Wnt pathway activation. *Cancer Res* **74**, 5963-5977 (2014).

252 Schlessinger, K. & Hall, A. GSK-3β sets Snail's pace. *Nat Cell Biol* **6**, 913-915 (2004).

253 Hughes, K., Nikolakaki, E., Plyte, S. E., Totty, N. F. & Woodgett, J. R. Modulation of the glycogen synthase kinase-3 family by tyrosine phosphorylation. *EMBO J* **12**, 803-808 (1993).

254 Meares, G. P. & Jope, R. S. Resolution of the nuclear localization mechanism of glycogen synthase kinase-3: functional effects in apoptosis. *J Biol Chem* **282**, 16989-17001 (2007).

255 Cole, A., Frame, S. & Cohen, P. Further evidence that the tyrosine phosphorylation of glycogen synthase kinase-3 (GSK3) in mammalian cells is an autophosphorylation event. *Biochem J* **377**, 249 (2004).

256 Hornung, G. & Barkai, N. Noise propagation and signaling sensitivity in biological networks: a role for positive feedback. *PLoS Comput Biol* **4**, e8 (2008).

257 Liu, Y. *et al.* Transforming growth factor-β (TGF-β)-mediated connective tissue growth factor (CTGF) expression in hepatic stellate cells requires Stat3 signaling activation. *J Biol Chem* **288**, 30708-30719 (2013).

258 Kaufhold, S. & Bonavida, B. Central role of Snail1 in the regulation of EMT and resistance in cancer: a target for therapeutic intervention. *J Exp Clin Cancer Res* **33**, 62 (2014).

259 Shirakihara, T., Saitoh, M. & Miyazono, K. Differential regulation of epithelial and mesenchymal markers by δEF1 proteins in epithelial–mesenchymal transition induced by TGF-β. *Mol Biol Cell* **18**, 3533-3544 (2007).

260 Aomatsu, K. *et al.* TGF-β induces sustained upregulation of SNAI1 and SNAI2 through Smad and non-Smad pathways in a human corneal epithelial cell line. *Invest Ophthalmol Vis Sci* **52**, 2437- 2443 (2011).

261 Aberger, F. & Ruiz, I. A. A. Context-dependent signal integration by the GLI code: the oncogenic load, pathways, modifiers and implications for cancer therapy. *Semin Cell Dev Biol* **33**, 93-104 (2014).

262 Doble, B. W. & Woodgett, J. R. GSK-3: tricks of the trade for a multi-tasking kinase. *J Cell Sci* **116**, 1175-1186 (2003).

263 Kretschmer, A. *et al.* Differential regulation of TGF-β signaling through Smad2, Smad3 and Smad4. *Oncogene* **22**, 6748-6763 (2003).

264 Fabian, S. L. *et al.* Hedgehog-Gli pathway activation during kidney fibrosis. *Am J Pathol* **180** (2012).

265 Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol* **7**, R100, doi:0.1186/gb-2006-7-10-r100 (2006).

266 Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646-674 (2011).

267 Zaslaver, A. *et al.* Just-in-time transcription program in metabolic pathways. *Nat Genet* **36**, 486 (2004).

268 Phillips, T. Regulation of transcription and gene expression in eukaryotes. *Nature Education* **1**, 199 (2008).

269 Ruscetti, F. W., Akel, S. & Bartelmez, S. H. Autocrine transforming growth factor-β regulation of hematopoiesis: many outcomes that depend on the context. *Oncogene* **24**, 5751-5763 (2005).

270 Strasen, J. *et al.* Cell-specific responses to the cytokine TGFβ are determined by variability in protein levels. *Mol Syst Biol* **14**, e7733 (2018).

271 Meng, X.-m., Nikolic-Paterson, D. J. & Lan, H. Y. TGF-β: the master regulator of fibrosis. *Nat Rev Nephrol* **12**, 325 (2016).

272 Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single- cell RNA-seq. *Nature* **509**, 371 (2014).

273 Heldin, C.-H., Landström, M. & Moustakas, A. Mechanism of TGF-β signaling to growth arrest, apoptosis, and epithelial–mesenchymal transition. *Curr Opin Cell Biol* **21**, 166-176 (2009).

274 Ke, X.-S. *et al.* Global profiling of histone and DNA methylation reveals epigenetic-based regulation of gene expression during epithelial to mesenchymal transition in prostate cells. *BMC Genomics* **11**, 669 (2010).

275 Messier, T. L. *et al.* Histone H3 lysine 4 acetylation and methylation dynamics define breast cancer subtypes. *Oncotarget* **7**, 5094-5109 (2016).

276 Kimura, H. Histone modifications for human epigenome analysis. *J Hum Genet* **58**, 439 (2013). Bar-Joseph, Z., Gitter, A. & Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat Rev Genet* **13**, 552-564 (2012).

277 Schulz, M. H. *et al.* DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Syst Biol* **6**, 104 (2012).

278 Nagase, H., Visse, R. & Murphy, G. Structure and function of matrix metalloproteinases and TIMPs. *Cardivovasc Res* **69**, 562-573 (2006).

279 Chandler, D. Introduction to modern statistical mechanics. *Introduction to Modern Statistical Mechanics*, 288 (1987).

280 Phanstiel, D. H. *et al.* Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Mol Cell* **67**, 1037-1048.e1036 (2017)

281 Liu, Y. *et al.* AP-1 blockade in breast cancer cells causes cell cycle arrest by suppressing G1 cyclin expression and reducing cyclin-dependent kinase activity. *Oncogene* **23** (2004).

282 Kerppola, T. K. & Curran, T. Fos-Jun heterodimers and jun homodimers bend DNA in opposite orientations: Implications for transcription factor cooperativity. *Cell* **66**, 317-326,

283 Shaulian, E. & Karin, M. AP-1 in cell proliferation and survival. *Oncogene* **20** (2001).

284 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319 (2013).

285 Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12**, 283 (2011).

286 Krijger, P. H. L. *et al.* Cell-of-origin-specific 3D genome structure acquired during somatic cell reprogramming. *Cell Stem Cell* **18**, 597-610 (2016).

287 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562 (2012).

288 Lesne, A., Riposo, J., Roger, P., Cournac, A. & Mozziconacci, J. 3D genome reconstruction from chromosomal contacts. *Nat Methods* **11**, 1141 (2014).

289 Weissman, I. L. Stem cells. *cell* **100**, 157-168 (2000).

290 Murry, C. E. & Keller, G. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. *Cell* **132**, 661-680 (2008).

291 Bibel, M. *et al.* Differentiation of mouse embryonic stem cells into a defined neuronal lineage. *Nat Neurosci* **7**, 1003 (2004).

292 Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025-1036 (2005).

293 Pedraza, J. M. & Paulsson, J. Effects of molecular memory and bursting on fluctuations in gene expression. *Science* **319**, 339-343 (2008).

294 Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* **440**, 358 (2006).

295 Newman, J. R. *et al.* Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**, 840 (2006).

296 Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nat Genet* **38**, 636 (2006).

297 So, L.-h. *et al.* General properties of transcriptional time series in Escherichia coli. *Nat Genet* **43**, 554 (2011).

298 Blake, W. J., Kærn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422**, 633 (2003).

299 Dar, R. D. *et al.* Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17454-17459 (2012).

300 Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472-474 (2011).

301 Nicolas, D., Phillips, N. E. & Naef, F. What shapes eukaryotic transcriptional bursting? *Mol Biosyst* **13**, 1280-1290 (2017).

302 Harper, C. V. *et al.* Dynamic analysis of stochastic transcription cycles. *PLoS Biol* **9**, e1000607 (2011).

303 Zoller, B., Nicolas, D., Molina, N. & Naef, F. Structure of silent transcription intervals and noise characteristics of mammalian genes. *Mol Syst Biol* **11**, 823 (2015).

304 Viñuelas, J. *et al.* Quantifying the contribution of chromatin dynamics to stochastic gene expression reveals long, locus-dependent periods between transcriptional bursts. *BMC Biol* **11**, 15 (2013).

305 Muramoto, T., Müller, I., Thomas, G., Melvin, A. & Chubb, J. R. Methylation of H3K4 Is required for inheritance of active transcriptional states. *Curr Biol* **20**, 397-406 (2010).

306 Wu, S. *et al.* Independent regulation of gene expression level and noise by histone modifications. *PLoS Comput Biol* **13**, e1005585 (2017).

307 Román, A.-C. *et al.* Histone H4 acetylation regulates behavioral inter-individual variability in zebrafish. *Genome Biol* **19**, 55 (2018).

308 Weinberger, L. *et al.* Expression noise and acetylation profiles distinguish HDAC functions. *Mol cell* **47**, 193-202 (2012).

309 Dodd, I. B., Micheelsen, M. A., Sneppen, K. & Thon, G. Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell* **129**, 813-822 (2007).

310 Dong, J. *et al.* Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. *Genome Biol* **19**, 31 (2018).

311 Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572.e524 (2017).

312 Fink, J. M. Astrotactin (ASTN), a gene for glial-guided neuronal migration, maps to human chromosome 1q25. 2. *Genomics* **49**, 202-205 (1997).

313 Kawano, H. *et al.* Identification and characterization of novel developmentally regulated neural- specific proteins, BRINP family. *Mol Brain Res* **125**, 60-75 (2004).

314 Zhang, H., Tian, X.-J., Mukhopadhyay, A., Kim, K. & Xing, J. Statistical mechanics model for the dynamics of collective epigenetic histone modification. *Phys Rev Lett* **112**, 068101 (2014).

315 Angel, A., Song, J., Dean, C. & Howard, M. A Polycomb-based switch underlying quantitative epigenetic memory. *Nature* **476**, 105-108 (2011).

316 Tian, X.-J., Zhang, H., Sannerud, J. & Xing, J. Achieving diverse and monoallelic olfactory receptor selection through dual-objective optimization design. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E2889-E2898 (2016).

317 Chen, Y.-j. *et al.* Rapid, modular, and cost-effective generation of donor DNA constructs for CRISPR-based gene knock-in. *Under Review* (2018).

318 Bartman, C. R., Hsu, S. C., Hsiung, C. C.-S., Raj, A. & Blobel, G. A. Enhancer regulation of transcriptional bursting parameters revealed by forced chromatin looping. *Mol cell* **62**, 237-247 (2016).

319 Andrews, S. S. Serial rebinding of ligands to clustered receptors as exemplified by bacterial chemotaxis. *Phys Biol* **2**, 111 (2005).

320 Tay, S. *et al.* Single-cell NF-κB dynamics reveal digital activation and analogue information processing. *Nature* **466**, 267 (2010).

321 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415 (2013).

322 Ge, H., Qian, H. & Xie, X. S. Stochastic phenotype transition of a single cell in an intermediate region of gene state switching. *Phys Rev Lett* **114**, 078101 (2015).

323 Zhang, J. *et al.* Pathway crosstalk enables cells to interpret TGF-β duration. *NPJ Syst Biol Appl* **4**, 18 (2018).

324 Wang, W. *et al.* Learn to estimate outline of cells with deep convolution neural networks and watershed. *Submitted* (2018).

325 Chen, B. *et al.* Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell* **155**, 1479-1491 (2013).

326 Ma, H. *et al.* Multicolor CRISPR labeling of chromosomal loci in human cells. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 3002-3007 (2015).

327 Chen, B., Guan, J. & Huang, B. Imaging Specific Genomic DNA in Living Cells. *Annu Rev Biophys* **45**, 1-23 (2016).

328 Fu, Y. *et al.* CRISPR-dCas9 and sgRNA scaffolds enable dual-colour live imaging of satellite sequences and repeat-enriched individual loci. *Nat Commun* **7**, 11707 (2016).

329 Shao, S. *et al.* Long-term dual-color tracking of genomic loci by modified sgRNAs of the CRISPR/Cas9 system. *Nucleic Acids Res* **44** (2016).

330 Qin, P. *et al.* Live cell imaging of low- and non-repetitive chromosome loci using CRISPR-Cas9. *Nat Commun* **8**, 14725 (2017).

331 Gu, B. *et al.* Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science* **359**, 1050 (2018).

332 Maass, P. G., Barutcu, A. R., Weiner, C. L. & Rinn, J. L. Inter-chromosomal Contact Properties in Live-Cell Imaging and in Hi-C. *Mol Cell* **69**, 1039-1045.e1033 (2018).

333 Ma, H. *et al.* Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. *Nat Biotechnol* **34**, 528 (2016).

334 Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, eaar4362 (2018).

335 Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, eaar5780 (2018).