npj | Systems Biology and Applications

www.nature.com/npjsba

ARTICLE OPEN

Combinatorial perturbation analysis reveals divergent regulations of mesenchymal genes during epithelial-to-mesenchymal transition

Kazuhide Watanabe[1], Nicholas Panchy[2,3], Shuhei Noguchi[1], Harukazu Suzuki[1] and Tian Hong[2,3]

Epithelial-to-mesenchymal transition (EMT), a fundamental transdifferentiation process in development, produces diverse phenotypes in different physiological or pathological conditions. Many genes involved in EMT have been identified to date, but mechanisms contributing to the phenotypic diversity and those governing the coupling between the dynamics of epithelial (E) genes and that of the mesenchymal (M) genes are unclear. In this study, we employed combinatorial perturbations to mammary epithelial cells to induce a series of EMT phenotypes by manipulating two essential EMT-inducing elements, namely TGF-β and ZEB1. By measuring transcriptional changes in more than 700 E-genes and M-genes, we discovered that the M-genes exhibit a significant diversity in their dependency to these regulatory elements and identified three groups of M-genes that are controlled by different regulatory circuits. Notably, functional differences were detected among the M-gene clusters in motility regulation and in survival of breast cancer patients. We computationally predicted and experimentally confirmed that the reciprocity and reversibility of EMT are jointly regulated by ZEB1. Our integrative analysis reveals the key roles of ZEB1 in coordinating the dynamics of a large number of genes during EMT, and it provides new insights into the mechanisms for the diversity of EMT phenotypes.

npj Systems Biology and Applications (2019)5:1 ; https://doi.org/10.1038/s41540-019-0097-0

INTRODUCTION

In epithelial-to-mesenchymal transition (EMT), a developmental program essential for morphogenic processes in embryogenesis and crucial in pathogenesis of malignant tumors, the cell-state transition occurs between two major states, epithelial (E) and mesenchymal (M), with well-characterized morphological features.[1–3] In the classical EMT process during development, transition from E to M state is unidirectional and two phenotypes are mutually exclusive. However, in cancer cells, phenotypes induced by EMT can be diverse with multiple intermediate or metastable states[4–7] and the transition between E and M states is often bidirectional.[2,4] Thus, E and M phenotypes are regulated primarily in a reciprocal fashion during classical EMT, but context-dependent regulations of E/M phenotypes were observed in nonconical or partial EMT. It is therefore challenging to obtain a comprehensive picture of molecular regulation, especially of reciprocity of E and M phenotypes.

EMT is characterized by a number of effecter genes each of which contributes to defining E or M phenotypes. Transcriptional profiling has been used to systematically measure such molecular phenotypes of EMT in a quantitative manner.[8,9] In these studies, several hundreds of EMT-related genes were selected through meta-analysis and manual curation using the expressional as well as functional characterization. For example, Tan et al. applied machine learning to obtain a list of signature genes, which can precisely predict aggressiveness of cancer cells.[8] Such molecular approaches contributed to understanding correlation of EMT and disease phenotypes.[8,9] Notably, EMT phenotypes with diverse transcriptional profiles has been observed in various pathological conditions.[8,9] However, understanding regulatory mechanisms of wide variety of EMT signature genes, particularly the coordination of these genes during EMT, requires mechanistic studies in appropriate model systems.

Among a myriad of EMT-regulating factors discovered to date, TGF-β has been shown to be a potent EMT-promoting signal,[10] and ZEB1 is an EMT-inducing transcription factor, that not only functions as a regulator for EMT program but is also involved in tumorigenesis.[11] Although TGF-β induces ZEB1 expression,[12] it is not clear whether ZEB1 can serve as an indispensable master regulator for TGF-β-induced EMT among other master EMT-TFs including SNAIL/SLUG and TWIST families,[2] and whether TGF-β and ZEB1 are in a linear axis that controls the entire EMT program. In addition, TGF-β has been shown to play paradoxical (tumor-initiating and tumor-suppressing) roles in cancer progression.[13,14] Similarly, a poised chromatin configuration of ZEB1 promoter was shown to be tumorigenic, and ZEB1 can be both tumor-promoting and pro-apoptotic factors.[14–16] These observations suggest that there are complex transcriptional programs activated by these two factors. However, the regulatory networks connecting these two factors to diverse transcriptional activities are not clear at the transcriptomic level.

In this study, we employed combinatorial perturbations to TGF-β and ZEB1, and created a series of EMT states in mammary epithelial cells. Using cells at these states of EMT, we applied transcriptomics, machine learning, mathematical modeling, and

[1]RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan; [2]Department of Biochemistry & Cellular and Molecular Biology, The University of Tennessee, Knoxville, Knoxville, TN 37996, USA and [3]National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996, USA
Correspondence: Kazuhide Watanabe (kazuhide.watanabe@riken.jp) or Tian Hong (hongtian@utk.edu)
These authors contributed equally: Kazuhide Watanabe, Nicholas Panchy

Received: 19 February 2019 Accepted: 28 May 2019
Published online: 14 June 2019

live-cell imaging analyses to examine how the coordinated transition between E and M states are regulated. We identified three groups of M-genes that can be distinguished by their responsiveness to TGF-β and ZEB1 pathways, and demonstrated the distinct biological impacts of the three M clusters in breast cancer patient survival and cell motility regulation. Surprisingly, high expressions of a cluster of M-genes that are strictly dependent on ZEB1 have significant association with good prognosis in breast cancer patients. Furthermore, using a mathematical model, we show that the reciprocity of EMT is synergistically controlled by TGF-β and ZEB1, and that the loss of this reciprocity transitions leads to partial EMT state with increased reversibility, which reduces the robustness of the destination state. Our results provide a holistic view of regulations of diverse mesenchymal genes during EMT, and they elucidate the mechanisms by which the cells ensure the coupling between E-gene and M-gene expressions in the transition. The classification of M-genes that we developed can be useful for the understanding of the diversity of EMT that is observed in various physiological and pathological conditions.

RESULTS

Divergence of mesenchymal genes upon perturbations of TGF-β and ZEB1

To dissect the molecular events involved in switching E and M phenotypes during TGF-β-induced EMT, we first generated ZEB1 knockout (KO) clones of MCF10A cells using CRISPR/Cas9 genome-editing technology (Supplementary Fig. 1, see the section "Materials and methods" for details). The KO cells did not show any detectable phenotypes in the basal culture condition (Supplementary Fig. 2). TGF-β treatment induced suppression of a representative E marker E-cadherin (E-cad, encoded by *CDH1* gene) and activation of a representative M marker Vimentin (VIM) in wild type (WT) cells (Supplementary Fig. 2), confirming that the E-genes and M-genes are reciprocally regulated during EMT. However, TGF-β failed to downregulate E-cad in KO cells, while VIM was still upregulated to the similar extent with the WT cells (Supplementary Fig. 2). These results suggest that while ZEB1 is a potent EMT-inducing transcription factor, its expression is dispensable for the induction of some M-genes.

To obtain a comprehensive view of the relative contribution of ZEB1 and TGF-β to EMT expression, we compared the MCF10A cells under four treatment conditions (TGF-β treated, TGF-β treated and ZEB1 KO, ZEB1 overexpressed, ZEB1 overexpressed and TGF-β inhibited) and their respective control conditions (eight conditions in total, Fig. 1a and Table 1). ZEB1 overexpression and TGF-β-signaling inhibition were performed by using doxycycline (DOX)-inducible system and TGF-β type1 receptor kinase inhibitor SB-431542, respectively. We examined the transcriptomes of the cells under these eight conditions with cap analysis of gene expression (CAGE), a highly sensitive and quantitative transcriptome assay which detects activities of transcription start site (TSS).[17,18] We also defined eight contrast conditions (described in Table 2) for the purpose of calculating log fold-change (logFC) to quantify differential expression under different regulatory regimes. We used a list of EMT genes curated from two sources: a set of 416 E-genes and M-genes annotated by Tan et al.[8] (see the section "Materials and methods") and additional 319 EMT genes without explicit E-genes or M-genes annotation.[9] Overall, 60.6% of annotated EMT genes, exhibited significant differential expression ($q < 0.05$, see the section "Materials and methods" for details) under at least one of the eight contrast conditions, compared to the rest of the genome, where only 10.0% were differentially expressed. The difference represents a significant enrichment of differently expressed genes in the annotated EMT set (Fisher's exact test, $p < 2.2e{-}16$), indicating that our set of EMT genes is more responsive to manipulation of TGF-β and ZEB1 than the genome in general.

We next explored expression differences between genes associated with epithelial phenotypes (E-genes) and those associated with mesenchymal phenotypes (M-genes). As expected, E-genes have lower expression under TGF-β treatment or ZEB1 overexpression conditions, and higher expression when these factors are inhibited or knocked out, while M-genes show the inverse pattern (Fig. 1b and Supplementary Fig. 3), though there is obvious variation in the strength and level of response of each class of EMT genes to TGF-β or ZEB1 gain and loss. If we only consider cases of significant differential expression, 89.1% of E-genes were down-regulated by TGF-β treatment (TGF-β vs. WT) or ZEB1 overexpression (DOX vs. DMSO), while 89.7% of M-genes were up-regulated in response to TGF-β treatment or ZEB1 overexpression. This is consistent with TGF-β/ZEB1 induction being responsible for a shift from epithelial phenotype to mesenchymal phenotype. As expected, the mean expression levels of E-genes and M-genes showed a negative correlation among the eight conditions. However, the overexpression of ZEB1 had much stronger effect on E-genes than on M-genes (Fig. 1b, the blue dots compared to the red dots), suggesting the primary role of ZEB1 in inhibiting E genes.

We next considered the response to TGF-β and ZEB1 independently. E-genes and M-genes exhibit different patterns of differential expression to TGF-β and ZEB1. Among E-genes, there is a large degree of overlap between genes, which are down-regulated in response to both TGF-β and ZEB1 (67.1%), while almost twice as many M-genes are differentially expressed in response to TGF-β compared ZEB1, such that differential expression of most M-genes (64.4%) is specific to one factor or the other (Fig. 1c). This difference in expression between E-genes and M-genes is robust to variations in the *q*-value and logFC thresholds that we used to define differential expression (Supplementary Fig. 4). The overlap for E-genes and that for M-genes become similar only with very large fold-change thresholds ($logFC > 2$) and even with this high threshold, E and M genes remain distinct as more E genes are differentially expressed in response to ZEB1, while more M genes are differentially expressed in response to TGF-β. In addition, we observed a difference between E-genes and M-genes in terms of the dependence of one EMT factor on the other one: there is an 80% overlap in E-genes that were both differentially expressed in response to ZEB1 induction (DMSO+DOX vs. DMSO) and in response to ZEB1 induction in the absence of TGF-β (DOX+SB vs. SB), and a 55% overlap between E-genes differentially expressed in response to TGF-β (WT+TGF-β vs. WT) and in response to TGF-β in the absence of ZEB1 (TGF-β+dZEB vs. dZEB) (Fig. 1d). This suggests that most E-genes are differentially expressed in response to both TGF-β and ZEB1, and the response to ZEB1 alone was stronger than that to TGF-β. Comparably, in M-genes these overlaps are only 64% for ZEB1-induced differential expressions, and 43% for TGF-β-induced differential expressions (Fig. 1d). We also observed a distinction between E-genes and M-genes across EMT-inducing factors: among E-genes that were differentially expressed in response to ZEB1 without TGF-β, 64% were differentially expressed in response to TGF-β with ZEB1, and 42% were differentially expressed in response to TGF-β without ZEB1. The corresponding values for M-genes were only 36% and 22%, respectively. While the overlap between TGF-β and ZEB1 expressions, both with and without the other factor, represents the largest differences in the amount of overlapping differential expression, E-genes are differentially expressed more uniformly than M-genes across all but a few comparisons (Fig. 1d), an observation which is also robust to stricter definitions of differential expression (Supplementary Fig. 4). Therefore, we conclude that M-genes exhibiting a divergent response to different combinations of ZEB1 and TGF-β input, while E-genes are more uniformly responsive to the

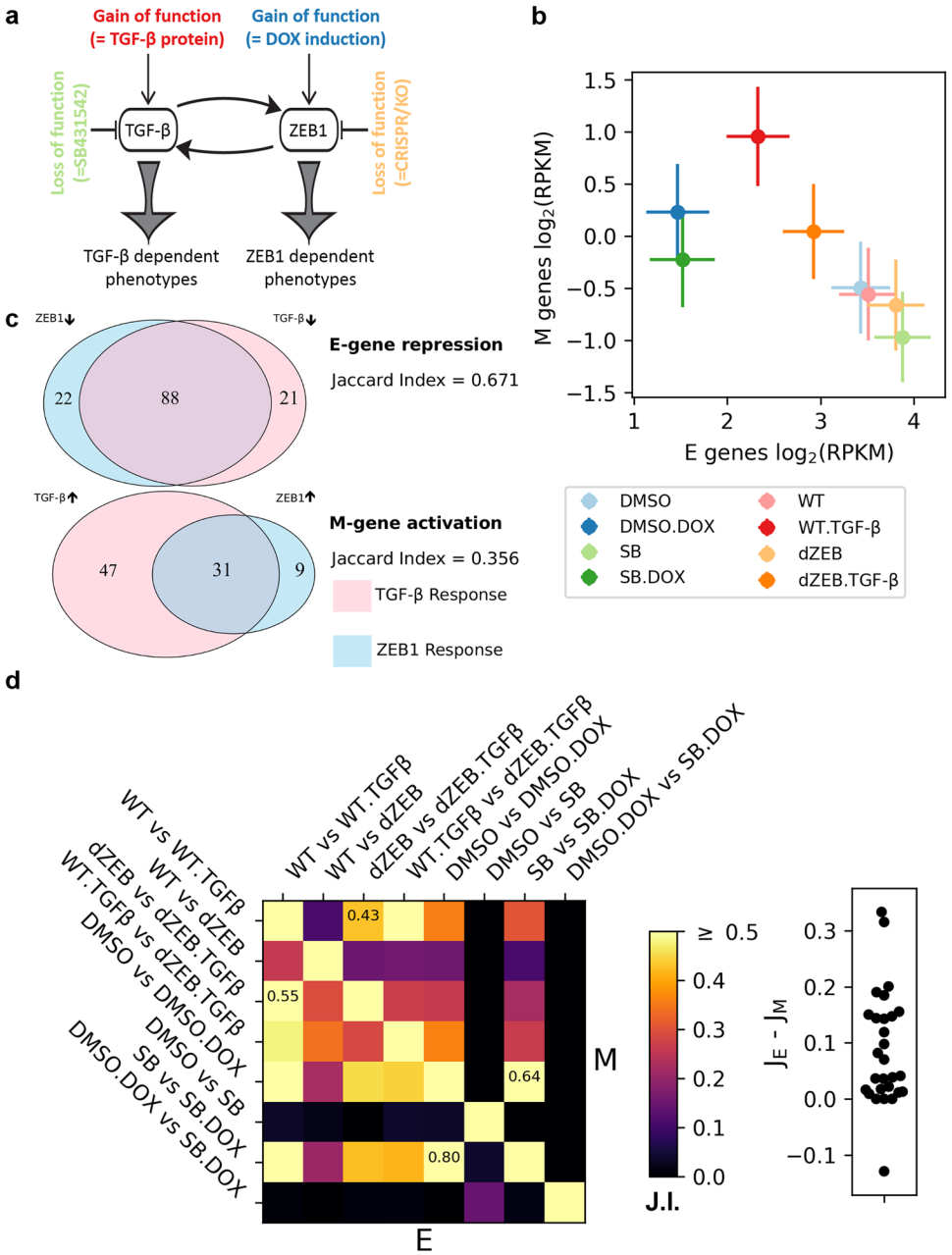


Fig. 1 Quantification of EMT gene expression in response to TGF-β and ZEB1. **a** Illustration of perturbations of TGF-β and ZEB1 conditions in this study. Each colored perturbation has a control condition. **b** Mean expression levels of E-genes and M-genes of eight conditions. Vertical and horizontal bars show standard error of the means for all annotated E or M genes. The colors are matched to **a**. **c** Venn diagrams showing the overlap in E (top) and M (bottom) genes, which show significant expression differences in response to TGF-β or ZEB1 treatment relative to their respective control conditions. TGF-β response is in pink while ZEB1 response is in light blue. The number E-genes and M-genes responding to TGF-β or ZEB1 uniquely, as well as those responding to both are listed in the respective part of the Venn diagram. The type of response, activation, or repression, is indicated by directional arrows (up-arrow: activation, down-arrow: repression). The overlap was quantified using the Jaccard index, which is the number of genes differentially expressed by both TGF-β and ZEB1 divided by the total number of differentially expressed genes. **d** Jaccard indices of E-genes and M-genes for each pair of conditions. Heatmap shows all the Jaccard indices. Lower triangular entries: E-genes. Upper triangular entries: M genes. The Jaccard index for the conditions which differ most between E-genes and M-genes are shown in those cells. Swarm plot shows the differences between the Jaccard indices of E-genes and those of the M-genes for each of the 28 pairs of conditions ($p < 0.001$ for single value t-test with a null distribution centered at 0)

TGF-β-ZEB1 axis (Fig. 1d). Furthermore, given both the divergence between TGF-β-induced and ZEB1-induced differential expression among M-genes, as well as the fact that M-genes are less frequently differentially expressed by either TGF-β treatment or ZEB1 induction independently, we hypothesized the existence of multiple M-gene regulatory modules under the control of TGF-β and ZEB1.

Three distinct types of regulatory circuits connect TGF-β and ZEB1 to M-genes

In an attempt to integrate all our expression profiles into single analysis, we first applied hierarchical clustering to logFC expression data (Supplementary Fig. 5). While the resulting tree largely separates annotated E-genes and M-genes into two main clusters, respectively, 19.5% of annotated EMT genes were incorrectly

Table 1. Combinatorial perturbation conditions

Expression condition	Description
WT	Control
TGF-β	TGF-β treatment
dZEB	CRISPR-mediated ZEB1 knockout
TGF-β+dZEB	TGF-β treatment and ZEB1 knockout
DMSO	Control for doxycycline-induction of ZEB1 and SB431542
DOX	Doxycycline-induction of ZEB1
SB	TGF-β inhibition by TGF-β receptor kinase inhibitor SB431542
DOX+SB	ZEB1 over-expression and TGF-β inhibition

Table 2. Contrast between expression conditions

Expression contrasts	Description
TGF-β vs. WT	Response to TGF-β induction
DOX vs. DMSO	Response to ZEB1 induction
SB vs. DMSO	Response to loss of TGF-β
dZEB vs. WT	Response to loss of ZEB1
TGF-β + dZEB vs. dZEB	Response to TGF-β induction in the absence of ZEB1
TGF-β vs. TGF-β + dZEB	Response to TGF-β induction in the presence of ZEB1
DOX + SB vs. SB	Response to ZEB1 induction in the absence of TGF-β
DOX vs. DOX + SB	Response to ZEB1 induction in the presence of TGF-β

classified (19.2% E-genes and 19.8% M-genes). Although some of this error may be the result of misannotated EMT genes, the difficulty in correctly separating E-genes and M-genes may in part be due to the high-dimensionality of our data set, as well as the assumption that all conditions are equally important to the distinction between E-genes and M-genes.

To address this issue, we used a semi-supervised approach to classify the E-genes and M-genes, and to identify major classes of M-genes subsequently. We first applied a self-organizing maps (SOM) algorithm (see the section "Materials and methods") to map EMT genes onto a 10 × 10 grid based on their logFC of expression under the eight contrast conditions. This dimensionality reduction clearly separated E-genes and M-genes (Fig. 2a), with 87 of the 100 nodes dominated by one type of EMT gene (E:M ≥ 2 or M:E ≥ 2), and 65 consisting exclusively of E-genes or M-genes. Of the remaining 13 nodes, 5 contain a mixture of annotated E-genes and M-genes, while the remaining nodes lack EMT genes with E or M annotations (black nodes, Fig. 2a), though this accounts for only 15 of the 319 (4.8%) EMT genes without an E or M annotation. Overall, 95.5% of EMT genes fall into a node with more than a 2:1 ratio of E/M or M/E genes. Using this cutoff, 7.9% of annotated EMT genes were incorrectly classified, and 87.6% of annotated EMT genes were correctly classified. Based on this classification, we further propagated E and M annotations to non-annotated EMT genes in nodes with predominant E-gene or M-gene annotations. Importantly, after updating our definition of E-genes and M-genes, we observe the same overall pattern of response to different combinations of TGF-β and ZEB1 expression, though there is slight reduction in responsiveness overall (Supplementary Table 1).

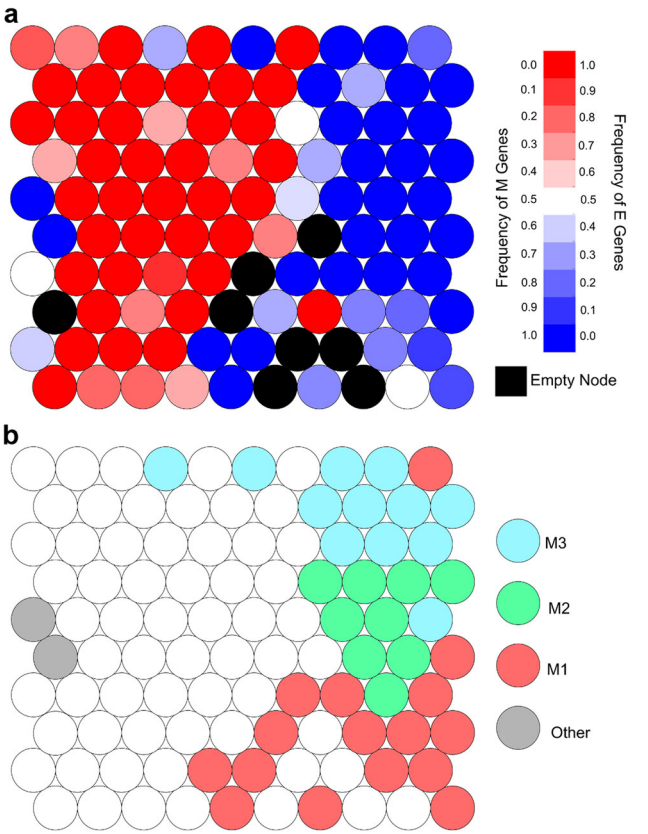


Fig. 2 Self-organizing maps (SOMs) for E-genes and M-genes and for clustering M-genes. **a** SOM nodes by frequency of E-genes and M-genes using a visual representation of the final map of EMT genes onto a 10-by-10 grid by SOM. The color of each node indicates the frequency of E-genes (darker-red) and M-genes (darker-blue) in each node. Nodes that are colored black are empty. **b** Clustering of SOM nodes with predominantly M-gene membership using a visual representation of the final map of EMT genes onto a 10-by-10 grid by SOM. Each of the 38 genes which have a 2:1 or greater ratio of M-genes to E-genes is colored according to the cluster of M-genes it was assigned to by hierarchical clustering (red = M1, green = M2, blue = M3, gray = other)

Table 3. Size and membership of the three M-genes expression clusters

Cluster #	Total genes	M-genes	E-genes	Notable genes
1	180	92 (48.9%)	16 (6.8%)	SNAI1
2	80	39 (20.7%)	1 (0.4%)	FN1, VIM
3	77	31 (16.5%)	4 (1.7%)	TWIST1, TWIST2

Based on the diversity in M-gene regulation, we used nodes with predominant M-gene annotation for the further exploration of M-gene regulatory modules. We selected 38 nodes from our SOM grid with predominantly M-genes and applied hierarchical clustering, and cut the resulting tree to generate four clusters (Fig. 2b and Supplementary Fig. 6; see the section "Materials and methods"). Of the resulting clusters, three ('M1', 'M2', and 'M3' in Fig. 2b) contain at least 70 genes overall and more than 30 annotated M-genes, but the smallest ('Other' in Fig. 2b) contained only 11 genes total and was therefore deemed too small for further analysis. The M1, M2, and M3 clusters together cover 86.2% of annotated M-genes while including only 9.4% of E-genes (see Table 3). To visualize expression of our M-gene clusters, we

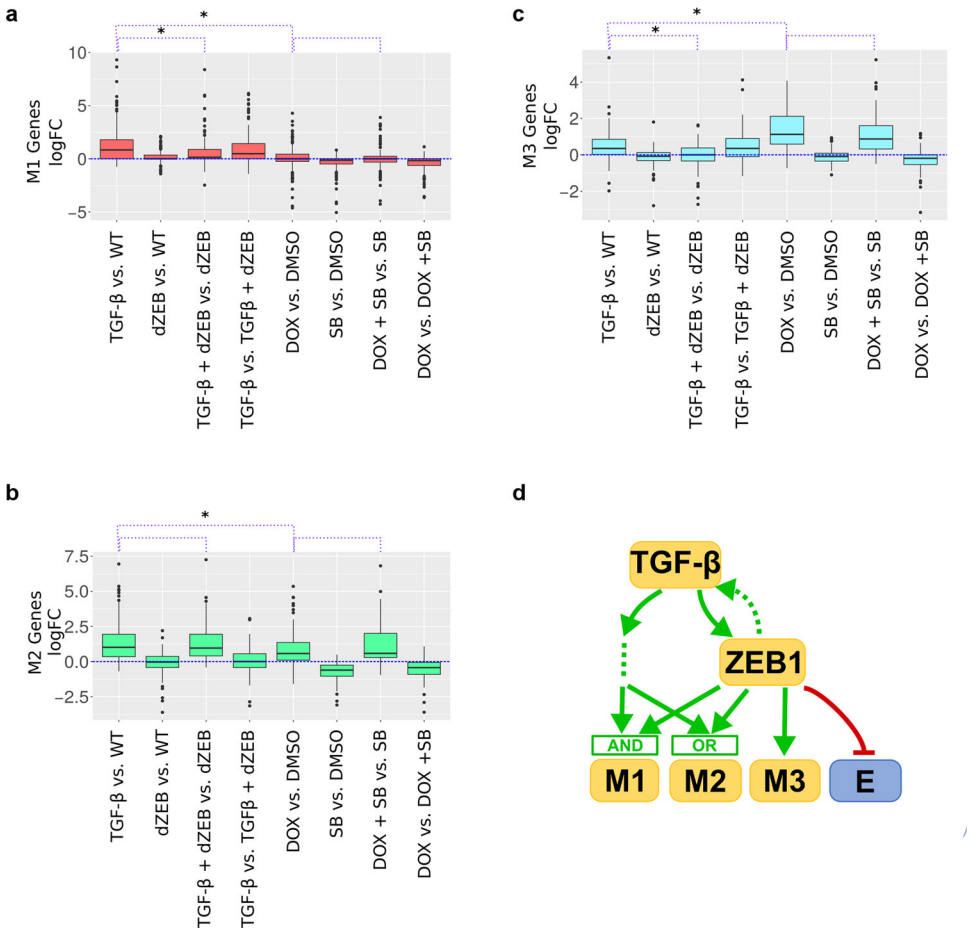


Fig. 3 Expression of M-gene clusters under TGF-β and ZEB1 regulation. **a** Boxplot of log fold-change of expression of M1 genes in the eight contrast conditions (see Table 2). The colored region (red) indicates the inter-quartile range of expression while whiskers extend 1.5 times this range on either side. Outliers are indicated by black dots. The purple dotted lines above the plot indicate comparisons between the expression under conditions, specifically, TGF-β vs. WT to DOX vs. DMSO, TGF-β vs. WT to TGF-β+dZEB vs. dZEB and DOX vs. DMSO to DOX+SB vs. SB. A * indicates that distribution of expression is significantly different based on the Mann–Whitney *U*-test at an alpha of 0.05. **b** Similar to **a** but for M2 genes, with a green colored region. **c** Similar to **a**, but for M3 genes, with a blue colored region. **d** A model of EMT-gene regulation based on the expression patterns of M-genes clusters in **a**–**c**. Green arrows indicate activation while red arrows indicate repression

performed principal component analysis (PCA) with the logFC expression data (Supplementary Fig. 7). We found that each M-gene cluster has a distinct pattern of expression, and we inferred the key perturbations that gave rise to the differences among the M-gene clusters by examining the loading (*R*) of individual principal components (PCs). In particular, PC1, which separates M1 and M3, is correlated with TGF-β response ($R = 0.363$ for WT +TGF-β vs. WT and 0.264 for dZEB+TGF-β vs. dZEB) and anticorrelated with ZEB1 response ($R = -0.518$ DMSO+DOX vs. DMSO and −0.541 for SB+DOX vs. SB), suggesting that M1 genes are more responsive to TGF-β and M3 genes are more responsive to ZEB1. Similarly, PC2 is correlated with ZEB1-independent TGF-β response ($R = 0.653$ dZEB+TGF-β vs. dZEB), suggesting M2 genes have a stronger response to TGF-β independent of ZEB1 than M1 and M3 genes do. While visualizing expression data in key directions is useful,[19] we were unable to distinguish our three clusters of M-genes with PCA alone (compare the labeled plots to the unlabeled plots (Supplementary Fig. 7, upper triangle).

We next performed further characterization of each group of M-genes. The M1 cluster covers almost half (48.9%) of M-genes, including the regulator SNAI1, fibroblast growth factor (FGF2) and fibroblast growth factor receptors (FGFR2), heat-shock proteins CRYAB and HSBP2, as well as several genes associated with tumor growth and migration including MMP9,[20] CDH11,[21] FOXC1,[22] and RAC1.[23] However, 25 of the M-genes in this cluster (13.3% of total M-genes) belonged to a single node comprised of genes that were not expressed across all data sets, and so were excluded from subsequent analysis (doing this also excluded 23% of the E-genes initially included in our M-gene clusters). These M1 genes were up-regulated in response to TGF-β treatment more than ZEB1 induction (Mann–Whitney *U*-test, $p < 2.2e{-}16$), but the response of M1 genes to TGF-β was significantly reduced when ZEB1 was absent (median logFC changed from 0.83 to 0.14, $p = 1.10e{-}10$, Mann–Whitney *U*-test) (Fig. 3a). Therefore, M1 genes are upregulated by TGF-β, and the responses depend on ZEB1. However, neither of the factors has a strong influence on these genes alone. The next largest cluster M2 (20.7% of M-genes) includes the structural proteins FN1 and VIM, which are both associated with the mesenchymal cell phenotype.[24,25] Like M1 genes, M2 genes had greater response to TGF-β than to ZEB1 (Mann–Whitney *U*-test, $p = 0.02$) (Fig. 3b), but the difference in median response was much smaller for M2 genes (TGF-β = 1.01, ZEB1 = 0.57) than for M1 genes (TGF-β = 0.83, ZEB1 = 0.0, $p = 2.56{-}e15$, Mann–Whitney *U*-test). Furthermore, the response of M2 genes to TGF-β was not affected by the absence of ZEB1 (Mann–Whitney *U*-test, $p = 0.69$) nor was its response to ZEB1 affected by the inhibition of TGF-β (Mann–Whitney *U*-test, $p = 0.33$) (Fig. 3b). As such, while TGF-β had a larger impact on M2

expression, both TGF-β and ZEB1 regulated M2 genes independent of one another. The M3 cluster (16.5% of M-genes) contains the transcription factors TWIST1 and TWIST2, growth factors HFG and FGFR1, two Mitogen-activated protein kinases, MAP3K3 and MAPK7, which are over-expressed in tumors,[26] and the Insulin-like growth factor-binding IFGBP3, which has complicated relationship with cancer progression depending on cancer type.[27–29] M3 genes responded to ZEB1 more than TGF-β (Mann–Whitney *U*-test, $p = 2.32e{-}7$) with no significant difference in ZEB1 response in the absence of TGF-β (Mann–Whitney *U*-test, $p = 0.09$) (Fig. 3c). M3 genes responded to TGF-β induction (median logFC = 0.35). This response was lost when ZEB1 was absent (median logFC = 0) and this change in response was significantly different (Mann–Whitney *U*-test, $p = 7.18e{-}6$). Therefore, we conclude that M3 genes are regulated by ZEB1 independent of TGF-β. To further confirm the differential expression patterns among the M-gene clusters, we performed RT-PCR for representative genes in each cluster and the results were consistent with our CAGE experiments and the clustering analysis (Supplementary Fig. 8). The clustering information for all 735 EMT genes that we analyzed is listed in Supplementary Table 2. In general, the three largest clusters obtained from our analysis show distinct patterns of regulation by TGF-β and ZEB1, in contrast to E-genes which primarily respond to ZEB1 directly (Supplementary Fig. 9). We summarized our findings in an illustrative model shown in Fig. 3d, which is largely reflective of the differences in expression suggested by our PCA analysis. In this model, M1 genes are regulated by TGF-β and ZEB1 via an AND logic gate. This AND-gate can only be turned on by TGF-β but not by ZEB1, possibly because TGF-β can activate ZEB1 completely, but ZEB1 can only partially activate TGF-β signaling (Fig. 3d, dashed arrow). The activation of TGF-β signaling by ZEB1 is supported by previous findings that show the mechanisms and importance of the mutual activation between TGF-β and ZEB1.[30–32] In particular, ZEB1 activates SMAD proteins which serve as key mediators of TGF-β signaling.[31,32] In contrast to the M1 gene cluster, M2 genes are regulated by the two factors via an OR-gate, M3 genes are regulated by ZEB1 but not ZEB1-independent TGF-β pathway, and E-genes are assumed to be controlled by ZEB1. The latter assumption is based on the observation that 82% of E-genes were downregulated by the expression of ZEB1 alone. Note that the arrows in this simplified network diagram (Fig. 3d) do not represent direct molecular interactions, but the diagram establishes causal, rather than correlative, relationships between the two core EMT factors and other M-genes because of the controlled perturbations that we performed.

We tested if the M-gene clusters are differentially regulated by other EMT inducers using a set of previously published microarray data[33] (see Supplementary Method 1, Supplementary Table S3). We found that M2 and M3 genes have a significantly larger response to EMT inducing factor Gsc than M1 genes do (Supplementary Figs. 10 and 11). Future work involving systematic and controlled perturbations of EMT factors will be required to elucidate the regulation of M-genes by other core EMT transcription factors.

M-gene clusters have distinct biological functions

To investigate the functional difference between the three M-gene groups, we retrieved human GO annotation from Gene Ontology Consortium and identified significantly enriched terms within each group of M-genes compared to the genome overall using Fisher's exact test ($p < 0.05$, see "Materials and methods"). M1 and M2 genes both had 85 significantly enriched terms while M3 genes had only 29. However, the majority of these terms are uniquely enriched to one group of M-genes, with only six terms found in all groups and 67, 66, 18 appearing exclusively in M1, M2, and M3 genes, respectively (Fig. 4a). In comparison, there are 169 GO terms are uniquely enriched among E-genes (Supplementary Table 4). However, when we categorized E genes into three groups by the responsiveness to ZEB1 and TGF-β (as in Fig. 1c), neither E genes responsive to ZEB1 or to TGF-β alone were enriched for any GO terms, while those responsive to both were enriched for 20 terms, primarily related to cell–cell junctions and adhesion (Supplementary Table 5). Among the set of unique genes in each group, we identified several sets of GO terms related to the same overarching function, which we have exhibited in Fig. 4a. In particular, we highlight the contrast between M1 and M2 genes: although both groups are enriched for term related to apoptosis, M1 genes are enriched for several terms related to cell motility, adhesion, and proliferation, while M2 genes are enriched for terms relating to the negative regulation of the same processes. Additionally, M2 are enriched for a wide variety development-associated GO terms not found in M1, including circulatory system, nervous system, embryonic, and other organs. This indicates a distinction between classes of M-genes: M1 genes are associated with the mobility and proliferative qualities associated with EMT, and M2 genes are related to development functions.

In addition to normal cellular functions, EMT is also strongly implicated in cancer progression, especially in breast cancers.[34,35] To check whether the three clusters of M-genes have differential roles in prognosis of breast cancer patients, we performed survival analysis for all the EMT genes annotated in this and earlier studies[36] (see the section "Materials and methods" for details). We found that there is no significant difference between patients with low and high expressions of E, M1, and M2 genes in terms of the median survival months (Fig. 4b), indicating the overall complexity of EMT's role in cancer progression. However, high expression of M3 genes is significantly associated with better prognosis compared to low expression of the corresponding genes in the breast cancer patients (Fig. 4b, see Supplementary Table 2 for the full list of median survival months for the EMT genes). The Kaplan–Meier plots for three representative genes are shown in Fig. 4b. Among these genes, SPARCL1 was identified as a tumor suppressor gene.[37] ZCCHC24 is strongly correlated with sensitivity to drug treatment.[38] MFAP4 is downregulated in several types of cancer and it was recently suggested to be a marker for developing therapies against cancer.[39] These results suggest that the high expression of the genes that are strictly dependent on ZEB1 pathway may play protective roles in cancer progression, or they can be served as markers for improved prognosis. To exclude the possibility that the association between better prognosis and high expression of M3 genes was driven by a few outlier genes, we calculated the percentages of genes of which higher expressions are significantly associated better prognosis for each cluster. Consistent with the boxplot and the *t*-test (Fig. 4b), M3 cluster has higher percentage of such genes (45.7%) than any other cluster does (M1: 33.6%, M2: 22.5%, E: 30.7%). These results are consistent with several recent findings that challenge the simple association between mesenchymal state and the invasiveness of cancer: higher expression of certain mesenchymal genes is associated with better prognosis, or inversely, higher expression of certain epithelial genes is associated with worse survival.[8,40,41] For instance, the E-gene GRHL2 (Supplementary Table 2) was shown to correlate with poor survival across all subtypes of breast cancer.[42]

We next performed gene set enrichment analysis to further explore the significance of the M-gene clusters in cancer settings. We found that these clusters are differentially expressed across a variety of cancer types (see Supplementary Method 2), suggesting that co-expression of M-gene cluster occurs in cancer. M3 genes specifically are uniquely enriched among genes differentially expressed in luminal A breast cancer and papillary thyroid cancer (Supplementary Tables 6 and 7). However, the set of differentially expressed genes tends to be small (between 4% and 28% of the cluster), so such relationship between M-gene clusters and cancer is likely driven by sub-clusters of cancer-related genes. We performed further analysis on the expression of the M-gene clusters in mesenchymal-like cancer cell lines (Supplementary

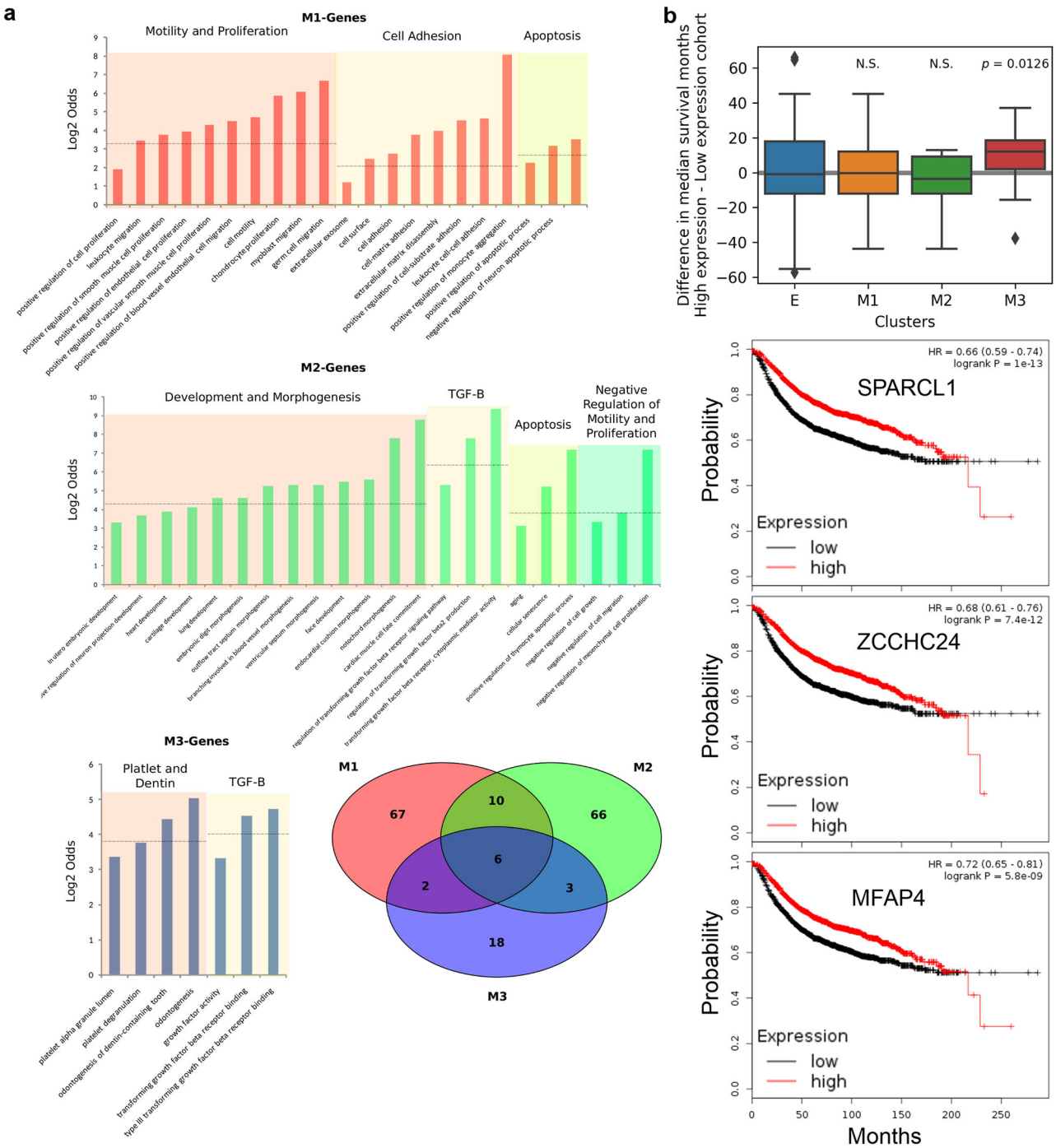
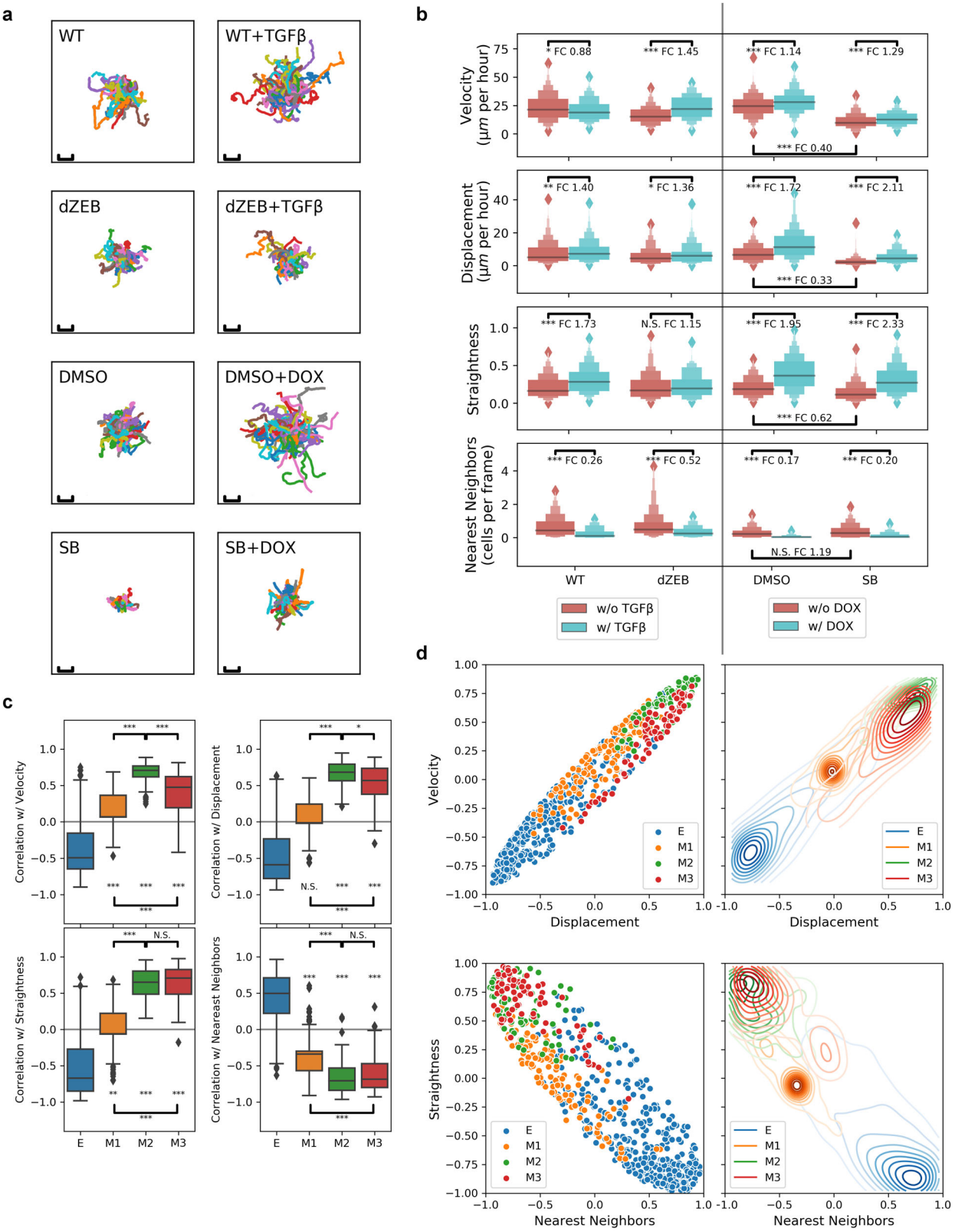


Fig. 4 Functional annotations and survival analysis of M clusters. **a** Bar graphs show groups of related GO terms associated with each cluster of M-genes. Venn diagram shows the overlapped GO terms ($p < 0.05$, Fisher's exact test, for selection of GO terms) among the three M clusters. **b** Top: Differences between high expression and low expression cohorts in terms of median survival months for breast cancer patients. Each group represents one type (cluster) of genes. The colored region indicates the inter-quartile range of expression, while whiskers extend 1.5 times this range on either side. Outliers are indicated by black dots. Single-value *t*-test was performed with each group of median survival months. Bottom: Kaplan–Meier plots for three representative M3 genes. Survival months and Kaplan–Meier plots were obtained from KM-Plotter.[36]

Method 3). We found that the mean expressions of the three M-gene clusters are significantly different in these cancer cell lines: M2 has the highest expression level and M1 has the lowest expression level (Supplementary Fig. 12). In addition, M3 genes have stronger within-cluster correlations (M3 genes vs. M3 genes) than between-cluster correlations (M3 genes vs. M1/M2 genes) in these cells (Supplementary Figs. 12 and 13). These results suggest the significance of the M-gene clusters in cancer cells.

Differential cell movement patterns regulated by TGF-β and ZEB1

To explore the relation between gene clusters and cellular phenotypes controlled by TGF-β and ZEB1, we collected imaging



data for MCF10A cells under the eight conditions listed in Table 1 and analyzed the movement patterns of the cells for each condition (Fig. 5a). We used four metrics to quantify the cell movement for each trajectory: the mean instantaneous velocity, the displacement scaled by duration, the straightness index, and the lifetime-averaged number of nearest neighbors (see the section "Materials and methods"). Overexpression of ZEB1 significantly increased the velocity of the cells, whereas the influence of TGF-β treatment on velocity is less significant (Fig. 5b, top panel). Nonetheless, inhibition of TGF-β signaling had the

Fig. 5 Differential cell movement patterns regulated by ZEB1-dependent and ZEB1-independent pathways. **a** Cell movement trajectories when TGF-β signaling and/or ZEB1 expression is perturbed under eight conditions. Hundred cells were randomly selected for each condition. Each trajectory was centered at its starting position. Scale bar represents a length of 100 μm. **b** Distributions of four metrics (instantaneous velocity, mean displacement normalized by duration of trajectory, straightness index of the movement, and number of nearest neighbors) shown in letter-value plots for cell trajectories under eight conditions. Statistical significance was obtained using Mann–Whitney *U*-test. FC fold-change. **c** Distributions of Spearman correlations coefficients (as a distance measurement) between gene expression and movement metrics for four type of genes (E, M1, M2, and M3) across eight conditions. The colored region indicates the inter-quartile range of expression while whiskers extend 1.5 times this range on either side. Outliers are indicated by black dots. **d** Scatter plots showing pairwise relationships between correlation coefficients of gene expression and different movement metrics. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, N.S.: not significant ($p > 0.05$), *t*-test. Significant mark at each box indicates the *p* value for testing if the mean of the group is significantly different from 0. Significant mark at each horizontal bar indicates the *p* value for testing if two groups of values are significantly different

most prominent negative effect on the velocity of the movement. In contrast, increasing TGF-β or ZEB1 signal had significant positive effect on displacement, which quantifies the overall migration efficiency of the cells. ZEB1 overexpression was the most potent condition to increase displacement, while the presence of TGF-β-signaling pathway was also essential for the increase (Fig. 5b, panel 2). ZEB1 alone positively influenced the straightness of the cell movement, even in the absence of the TGF-β-signaling pathway, whereas the influence of TGF-β on straightness depends on the presence of ZEB1 (Fig. 5b, panel 3). ZEB1 also had a predominant role in reducing the number of neighboring cells during the lifetime of the trajectories (Fig. 5b, bottom panel), suggesting that its positive regulation of the straightness of the movement is partially via the reduction of cell-to-cell contact. These results demonstrate the key role of ZEB1 in regulating the straightness of the cell movement, which is correlated with the overall efficiency of cell migration.

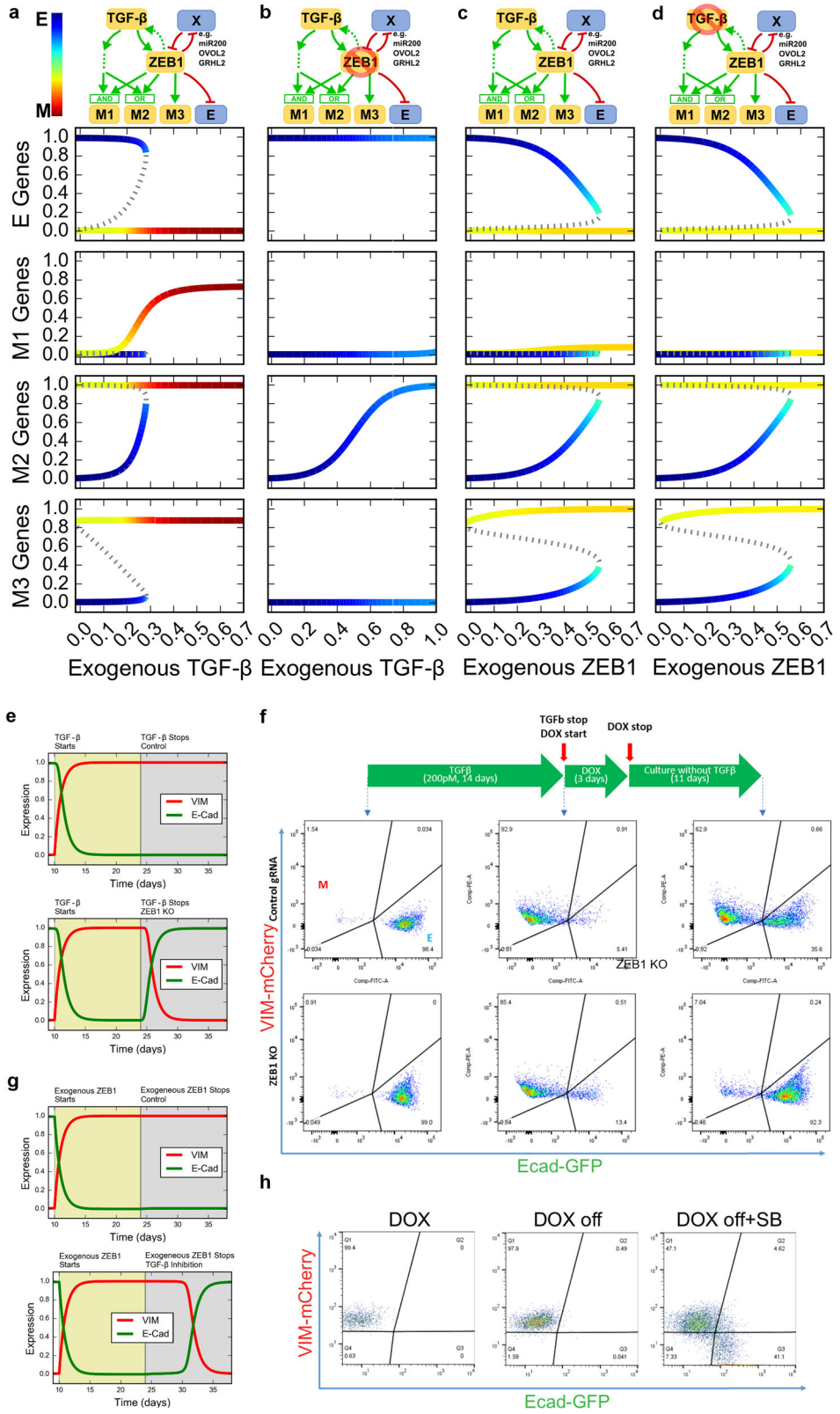
We next asked how the expression of M-gene clusters is correlated with the movement patterns. We calculated the Spearman correlation coefficients between the gene expression levels across the eight conditions and values of each of the movement metrics described above under the same conditions. These correlation coefficients serve as distance measurements between the gene expression pattern of each gene and movement pattern. For example, a positive coefficient between the expression of a gene and displacement means that the higher expression of that gene is correlated with the higher displacement. Among the three M-gene clusters, the expression of M2 genes has the strongest correlation with velocity, displacement, straightness, and nearest neighbors (all four metrics of movements). Compared with M2 genes, the expression of M3 genes has weaker, but still significantly positive correlation with velocity, and comparable correlations with all other three metrics. This suggests that M2 and M3 genes have similar contributions to the overall movement patterns. We asked under which specific conditions M2 gene expression shows better correlation with the velocity than the expression of other M-genes does, and we found that when cells were treated with TGF-β in the absence of ZEB1, velocity was significantly increased, and this is the condition under which M3 genes, but not the other two groups of genes, were significantly (>2-fold) upregulated (Supplementary Fig. 14). In contrast to M2 and M3 genes, the expression of M1 genes is not significantly correlated with the displacement, and its correlations with other movement metrics are much weaker than that of the expression of M2 and M3 genes. These weak correlations are consistent with the differential sensitivities of cell movement patterns and M1 gene expression to EMT signals: the movement of the cells is sensitive to perturbations to either TGF-β or ZEB1, whereas M1 genes can only be upregulated when both signals are present. Nonetheless, the significant correlation between the expression of M1 genes and some movement patterns is consistent with the GO analysis (Fig. 4a).

A mathematical model for ZEB1-TGF-β transcriptional network reveals the role of ZEB1 in controlling reciprocity and reversibility of EMT

To gain more insights into the roles of ZEB1 in controlling EMT, we built a mathematical model to describe the gene regulatory network for EMT in response to TGF-β signal based on the three groups of M-genes and one E-gene group (Fig. 6a). In particular, M1 genes are activated by TGF-β and ZEB1 via an AND logic gate, M3 genes are directly activated through ZEB1-dependent pathway but not ZEB1-independent pathway, and M2 genes are influenced by both ZEB1-dependent and ZEB1-independent pathway via an OR logic gate. We assumed that the E-genes are primarily controlled by ZEB1-dependent pathway, as suggested by the expression analysis of E genes in earlier sections (Supplementary Fig. 9). In addition, we included some known feedback loops involving ZEB1 that were established in earlier studies, including a positive feedback loop between ZEB1 and TGF-β,[30–32] and a representative mutual inhibition loop formed by ZEB1 and another factor (e.g. miR200, OVOL2, and GRHL2.[4,43,44] Note that all of these typical E-genes were correctly classified as E-genes with our algorithm. See Supplementary Table 2). In our classification analysis of M-genes, ZEB1 belongs to M2 cluster (Supplementary Table 2), which is activated by upregulating ZEB1 itself. This is consistent with the positive feedback loops described in Fig. 6a, and this self-activating feedback may be mediated by other genes, such as ESRP1 and HAS2, as well.[45,46]

We performed bifurcation analysis with the model by varying two parameters: the strength of external TGF-β signal, which represents physiological inducer of EMT and the concentrations of the exogenous ZEB1 which was controlled experimentally in this study. With the increasing strength of external TGF-β, the production of E-genes was turned off, and the production of all types of M-genes was turned on (Fig. 6a). The 'flipping' from E-on-M-off state to E-off-M-on state shows a reciprocal regulation of E-genes and M-genes under normal conditions. In addition, the model suggests that the TGF-β-induced EMT is irreversible once the cells have committed to the complete M state, and this is consistent with earlier mathematical models and experimental observations.[6,47] Note that in this study we define reversibility of EMT as the ability for the system to return to the E state upon the complete withdrawal of EMT signal TGF-β or ZEB1 (transition from red/yellow branch to blue branch with EMT signal decreasing to zero in Fig. 6a–d). It is possible to examine the reversibility of EMT upon partial withdrawal of EMT signal but this property is more accurately described as hysteresis instead of (ir)reversibility.[48] We next blocked the production of ZEB1 in the model and performed similar bifurcation analysis. In the absence of ZEB1, the TGF-β signaling only triggered the transition into a partial EMT state, in which only M2 was upregulated, whereas the ZEB1-dependent M1 and M3 genes were not responsive (Fig. 6b). In addition, the model predicts that TGF-β-induced EMT becomes reversible upon the loss of ZEB1 (Fig. 6b, e).

Exogenous expression of ZEB1-triggered reciprocal regulation of E-genes and some M-genes: with increasing ZEB1 production



rate, E-genes were downregulated whereas M2 and M3 genes were upregulated (Fig. 6c). However, exogenous ZEB1 did not activate M1 genes because of the absence of exogenous TGF-β. Nonetheless, the model suggests that ZEB1 can trigger irreversible EMT (Fig. 6c). In addition, the model predicts that the inhibition of TGF-β can reduce the irreversibility (Fig. 6d, g).

To validate the predictions in terms of the reversibility (Fig. 6e), we first verified that TGF-β-induced EMT is irreversible in most WT cells at least for 10 days without continuous exposure to TGF-β using an EMT reporter system (see the section "Materials and method", Fig. 6f). In contrast, the EMT phenotype induced by TGF-β was reversed by inducing ZEB1 deletion using an inducible

Fig. 6 Mathematical modeling of EMT under control of TGF-β and ZEB1. **a**–**d** Top diagrams: influence diagrams for gene regulatory networks under four conditions. Lower panels: bifurcation diagram for four types of genes (E, M1, M2, and M3) with respect to exogenous TGF-β and ZEB1 expression under two conditions. Solid curves represent stable steady states. Dashed curves represent unstable steady states. Color gradient represents the position in the EMT spectrum, which is calculated by adding the expression of the three M nodes and subtracting the expression of the E node. **a** Normal condition. Induced by TGF-β. **b** ZEB1 KO. Induced by TGF-β. **c** Normal condition. Induced by ZEB1. **d** TGF-β-inhibited condition. Induced by ZEB1. **e** Simulation for expression of VIM and E-cad upon treatment and withdrawal of TGF-β. Top: control. Bottom: ZEB1 KO after TGF-β withdrawal. **f** Expression of VIM and E-cad analyzed by FACS upon treatment and withdrawal of TGF-β. Top: control. Bottom: ZEB1 knockout induced by DOX. Cells were treated with TGF-β for 2 weeks and then subject to TGF-β withdrawal, and (for the experiment group) to ZEB1 knockout. **g** Simulations for expression of VIM and E-cad upon treatment and withdrawal of exogenous ZEB1 expression followed by inhibition of TGF-β signaling. **h** Expression of VIM and E-cad analyzed by FACS upon induction and withdrawal of exogenous ZEB1 by DOX. Exogenous ZEB1 was induced for 1 week and then subject to the withdrawal of the induction signal for 2 weeks, and (for the experiment group) to the inhibition of TGF-β signal by SB431542 (SB) for the same period

genome-editing system with DOX-inducible Cas9 and constitutive expression of ZEB1-targeting gRNA (Fig. 6f). Furthermore, transient induction of ZEB1 expression triggered irreversible EMT (Fig. 6h, left and middle panels), which is consistent with our modeling analysis (Fig. 6g, top panel). Interestingly, the irreversibility of the transition is even more robust than that of the TGF-β-induced EMT (Fig. 6h, middle panel, Fig. 6f, top right panel) 10 days after signal withdrawal. Treatment with TGF-β inhibitor upon withdrawal of exogenous ZEB1 expression caused a partial reversal of the EMT (Fig. 6g, bottom panel, Fig. 6h, right panel). These results suggest that the endogenous ZEB1 with its feedback regulation is essential for maintenance of irreversible EMT phenotype in mammary epithelial cells as reported for other cell types.[30] Together with the results shown in earlier sections, our data indicate that the irreversibility and the reciprocity of EMT are both regulated by ZEB1, and these two properties may be closely related.

DISCUSSION

EMT is involved in many biological processes, but a remarkable diversity of EMT phenotypes has been observed in both distinct and similar pathological conditions.[8,9,49] This diversity may be attributed to the metastable states that exist between terminal E and M states.[2,4–6] To describe these multiple states, a linear lineage progression model with coupled changes of E-gene and M-gene expression was widely used.[2,4,49] In this study, we identified the key roles of ZEB1 in regulations of the coupling between E-genes and M-genes. In addition, our findings raised the possibility that genetic or microenvironmental perturbations on ZEB1 activity may result in decoupling of E and M phenotypes, which may contribute to the diverse 'hybrid' EMT populations observed previously.[49] Furthermore, our analysis of the expression of EMT genes in response to ZEB1 and TGF-β not only indicates that M-genes exhibit a greater diversity of responses, but also suggests that they can be classified into no less than three sub-clusters. M-gene clusters have distinct patterns of expression in response to different perturbations of ZEB1 and TGF-β, and they can also be separated based on their functions, with M2 exhibiting development-related function, while M1 genes are involved in cell motility, growth, and adhesion. Furthermore, M3 genes have a significant association with the prognosis of breast cancer, indicating that genes that are controlled by ZEB1 but not other TGF-β-dependent pathways likely play a protective role. Although ZEB1 was shown to promote tumor initiation and metastasis,[50] the importance of the poised ZEB1 promoter status in cancer progression[15] indicates that there is a subset of ZEB1-induced genes that may be inversely correlated with tumorigenesis. This may account for the differential roles of multiple EMT states in progression of cancer.[49] The existence of these functionally distinct sub-groups of M-genes further elucidates the connection between decoupling and hybrid EMT phenotypes because the disruption of normal regulation may not affect all processes involved in EMT equally. Our study focused on the diversity of M-genes instead of E-genes, because M-genes are more diverse than E-genes in terms of their responsiveness to EMT-inducing factors (Fig. 1). However, it is possible that E-genes exhibit significantly albeit more weakly divergent responses, and they may be diversely regulated by factors other than ZEB1 and TGF-β. Future work is warranted to test these possibilities.

Our model elucidates the key roles of ZEB1 in regulating the reciprocity of EMT. During EMT, ZEB1 is directly responsible for the downregulation of most E-genes and the upregulation of a group of M3 genes. Effectively, ZEB1 ensures the coupling between the loss of the E phenotypes and the gain of M phenotypes. Conversely, the absence of ZEB1 blocks the ability of the cells to transition to terminal M state, and it decouples the dynamics of E-genes and M-genes. In addition, a ZEB1-independent pathway that activates M1 genes is also required for the complete transition to the terminal M state. Taken together, our model provides a mechanistic view of the EMT transcriptional network controlled by ZEB1 and TGF-β at the transcriptomic level. Our results also demonstrated the importance of ZEB1 in controlling the reversibility of EMT. This is consistent with a very recent study, which showed the essential roles of ZEB1-miR200 feedback loop in the hysteresis of EMT, and that the varied reversibility of EMT can influence the metastatic potentials of cancer cells.[48] Our results further suggest that ZEB1 serves as the hub for coordinating the reciprocity of E-genes and M-genes, and this coupling is closely related to the variable reversibility of EMT. In our mathematical model, we took the simple assumption that ZEB1 forms a positive feedback (double-negative) loop with another E-gene. In fact, ZEB1 may form positive feedback (double-positive) loops with other M-genes, such as SNAIL or TWIST as well, and these feedback loops may also contribute to the irreversibility of EMT. With the positive feedback loops involving ZEB1, E-genes and M-genes are reciprocally regulated, and their dynamics are always inversely correlated when the extracellular signal is varied (Supplementary Fig. 15). This reciprocity is also essential for the irreversibility of the EMT. Without the feedback loops (e.g. loss of ZEB1), the reciprocity of E-genes and M-genes is compromised, and EMT becomes more reversible because some of the feedback loops requires the downregulation of E-genes (Supplementary Fig. 15).

Controversial findings have been reported for roles of EMT in cancer metastasis.[51–54] Accumulating evidence supports that sequential induction of EMT and its reverse process mesenchymal-to-epithelial transition (MET) allows cancer cells enter into the systemic circulation and subsequently colonize at distant organs.[2,55] On the other hand, other studies suggested that cancer metastasis is often caused by circulating clusters of epithelial tumor cells and that EMT is dispensable for this process.[52,53] These conflicting findings suggest two modes of spreading processes of cancer cells: migratory and invasive properties of individual cancer cells that are related to EMT phenotypes and collective cell migration which occurs without losing epithelial integrity. The latter type of collective migration may not fit with the classical linear lineage progression definition of EMT. In fact, recent theoretical work suggests the importance of

hybrid EMT states for circulating tumor cell clusters.[56] Our findings suggest that E and M phenotypes are not necessarily regulated in a reciprocal manner and that EMT process can be flexibly reversed in these cells. It has been suggested that such plastic states are related to cancer stem cell phenotypes.[7,15,57] In addition, our finding on the causal influence of ZEB1 on key EMT genes, such as SNAIL and TWIST1/2, provides a possible explanation for the differential effects of knocking out ZEB1, SNAIL, and TWIST in mouse model of pancreatic cancer.[16,53] ZEB1 was considered a gene downstream of SNAIL and TWIST,[58,59] and our results show that ZEB1 can be a key regulator broadly influencing M-genes, including SNAIL and TWIST. Furthermore, several reports suggested that sequential activation of EMT and MET promotes reprograming or differentiation of cell lineages including iPS, neurons, or hepatocyte lineages.[60,61] Our study provides a possible molecular basis for such plastic transition of cellular state and manipulating the balance between the two inhibitory networks may be useful to develop new treatment for diseases or novel cell conversion methods.

Previous mathematical models and experiments showed the existence of intermediate EMT states in silico, in vitro, and in vivo.[4–6,49] In contrast, the model and the experiment presented in this study focus on the diversity of M-genes in terms of their connectivity with TGF-β and ZEB1, as well as the role of ZEB1 in controlling reversibility and reciprocity of EMT. However, the model does not describe the intricate feedback loops in the EMT transcriptional network. These feedback loops were shown to be critical for the formation of intermediate cell states.[4,47,62] Therefore, the current model has limited predictive power in terms of the detailed transitions involving intermediate EMT states. In particular, the bifurcation point shown in Fig. 6a–d may be decomposed into several consecutive switches that govern the critical transitions of different E-genes or M-genes. Future work is needed to integrate different elements of the EMT control circuits into a unified model to analyze the system in a more comprehensive manner. In addition, previous models predicted discrete states between E and M phenotypes,[4–6,32,47] whereas our model and another EMT model by Celià-Terrassa et al.[48] suggest that continuous EMT phenotypes may be observed with increasing EMT signals. Distinguishing these two scenarios requires future experiments. Nonetheless, our model and experiment demonstrated the existence of hybrid EMT state at which both E-genes and M-genes are upregulated. It is unclear whether the 'intermediate' states can be distinguished from the 'hybrid' states under physiological conditions, and whether the intermediate phenotypes observed in vivo are transitional cells ready to commit to the destination states, or those unable to make the complete transition due to genetic or environmental perturbations and likely to be reverted to the initial states. More modeling work and single-cell experiments are warranted to address these questions, and the insights into these problems will help elucidating the functional roles of the intermediate or hybrid EMT states.[63]

Our study provides a holistic view of the roles of ZEB1 and its interplay with TGF-β signaling at transcriptomic level. We found the key roles of ZEB1 in regulating both the reciprocity and reversibility of EMT, and we identified three classes of mesenchymal genes that are controlled by three different types of regulatory circuits downstream of TGF-β and ZEB1. These results shed light into the complex molecular mechanisms for regulating EMT, and they are useful for the understanding of the diversity of EMT phenotypes observed in many physiological and pathological conditions.

MATERIALS AND METHODS

Cell lines

MCF10A cells (ATCC) were grown in DMEM/F12(1:1) medium with 5% horse serum, epidermal growth factor (10 ng/mL), cholera toxin (100 ng/mL), and insulin (0.023 IU/mL). For TGF-β treatment, cells were incubated with titrated concentrations of human TGF-β1 protein (R&D systems) in the complete culture medium. The culture medium was replaced daily, and cells were passaged right before reaching full confluency.

CRISPR/Cas9-mediated ZEB1 deletion

CRISPR/Cas9-mediated genome editing of ZEB1 locus was performed using lentiviral gRNA expression system with lentiGuide-Puro (a gift from Feng Zhang Addgene plasmid #52963) and lentiCas9-Blast (a gift from Feng Zhang Addgene plasmid #52962). For inducible Cas9 expression and genome editing, Lenti-X™ Tet-One™ Inducible Expression System (Clontech) was used. Production of lentiviruses was carried out as previously described.[64] Two gRNA sequences were used to delete ZEB1 expression; TGAAGACAAACTG-CATATTG (tgg: PAM sequence) and CAGACCAGACAGTGTTACCA (ggg: PAM sequence), and the following gRNA sequences were used as controls: ACCAGGATGGGCACCACCC and GGCCAAACGTGCCCTGACGG. For ZEB1 KO clones, two clones (clone 2 and clone 5) were established showing complete absence of ZEB1 protein upon TGF-β stimulation (Supplementary Fig. 2). As the morphology, proliferation rates, and gene expression patterns are similar, we chose to use clone 5 for the following studies. This clone contained homozygous 370-bp deletion in Intron1–Exon 2 boundary which results in exon skipping and frameshift (Supplementary Fig. 1).

Inducible expression of ZEB1 protein

Puromycin-resistant Tet-based inducible cDNA expression system (pSLIK-Puro) was engineered by replacing the hygromycin-resistant gene in pSLIK-Hygro vector (a gift from Iain Fraser, Addgene plasmid # 25737) with a puromycin-resistant gene obtained from lentiGuide-Puro by PCR amplification. Mouse Zeb1 cDNA was digested from the pHIV-ZsGreen-Zeb1 lentiviral construct[64] and cloned into the pSLIK-Puro vector. Production and infection of lentivirus were carried out as described above. Induction of the transgene was performed by DOX treatment at a concentration of 500 ng/mL.

EMT reporter and flow cytometry

The EMT reporter was engineered by removing the puromycin-resistant gene from a TCGP-Puro lentiviral EMT reporter system, which contains E-cad promoter-driven eGFP and VIM promoter-driven mCherry (a kind gift from Kiyotsugu Yoshikawa). The EMT reporter-expressing MCF10A clone was established by infection and FACS sorting of eGFP-positive cells, followed by serial dilution cloning. Several clones were screened by flow cytometry and clones that showed most distinguishable FACS profiles before and after TGF-β treatment was selected for the downstream analyses. Flow cytometry was performed on a BD FACSAria equipped with FACS DiVa6.0 software operating. Cell clusters and doublets were electronically gated out. Positive and negative gates for eGFP and mCherry fluorescence were determined using untreated and ZEB1-induced MCF10A cells as controls.

RT-PCR

Total RNA was isolated using the TRIzol Reagent (Invitrogen) followed by cleaning up and RNase-free DNaseI treatment using the RNeasy mini kit (QIAGEN). cDNA was prepared using Retroscript Kit (Applied Biosystems) according to manufacturer's instructions. Real-time PCR was performed using using a StepOnePlus™ real-time PCR system (Thermo Fisher), with SYBR Premix Ex Taq™ II (Takara). Comparative analysis was performed between the genes of interest normalized by the house keeping genes *GAPDH* and *ACTB*. The primer sequences used in this study are described in Supplementary Table 8.

Cap analysis of gene expression

CAGE libraries were prepared as previously described.[17] Briefly, 3 µg of total RNA from each sample were subjected to reverse transcription, using SuperScript III Reverse Transcriptase (Thermo Fisher) with random primers. The 5-end cap structure was biotinylated by sequential oxidation with $NaIO_4$ and biotinylation with biotin hydrazide (Vector Laboratories, USA). After RNase I treatment (Promega, USA), the biotinylated cap structure was captured with streptavidin-coated magnetic beads (Thermo Fisher). After ligation of 5′ and 3′ adaptors, second-strand cDNA was synthesized with DeepVent (exo−) DNA polymerase (New England BioLabs, USA). The double-stranded cDNA was treated with exonuclease I (New England

BioLabs) and purified. The resulting CAGE libraries were sequenced using single-end reads of 50 bp on the Illumina HiSeq 2500 (Illumina, USA). The extracted CAGE tags were then mapped to the human genome (hg38). After filtering low-quality reads, mapped CAGE tags were counted regarding FANTOM5-CAT TSSs,[18] providing a unit of CAGE tag start site. The tags per million (tpm) were calculated for each TSS peak and computed as gene expression levels for multiple TSS peaks associated with a single gene. Differentially expressed genes were identified using QL *F*-test implemented by EdgeR, followed by false discovery rate (FDR) control with Benjamini–Hochberg method.

Clustering EMT genes with a semi-supervised approach

For the unsupervised step, two approaches were used to cluster EMT genes based on the logFC of expression under the eight contrast conditions described in Table 2. logFC data was row-scaled to adjust for the large difference in average expression across EMT genes. Both approaches were implemented in the R programming language. In the first approach, a matrix of distances between annotated E-genes and M-genes was calculated based on logFC expression data using the 'dist' function with the Euclidean method. The 'hclust' function was then used to generate the dendrogram seen in Supplementary Fig. 5. In the second approach, we use the 'som' function, which is part of the kohonen package,[65] to map EMT genes onto a 10×10 grid of hexagonal cells. SOM were performed using the 'som' function. The full data set was presented to the network 1000 times during the learning process and the learning rate was set to decline linearly from 0.05 to 0.01 over the course of the learning process. From the final grid, we selected nodes which predominantly (2:1) consisted of M-genes (this supervised method is essentially a *k*-nearest-neighbors algorithm). We then clustered these selected nodes using the hierarchical clustering method described above, replacing logFC data with the codebook vectors describing each node. The 'cutree' function was used to derive clusters from the resulting dendrogram and we selected four as optimal numbers of clusters using the within cluster sums of squares error using the elbow criterion (Supplementary Fig. 6).

Identifying enriched GO annotations in M-genes

GO annotations were obtained from Gene Ontology Consortium.[66] The significance of enrichment of individual terms in each of the M-gene clusters was evaluated using Fisher's exact test, and multiple test correction was implemented using the Benjamini–Hochberg correction.

Survival analysis

Survival analysis was performed with KM-plotter.[36] Patients of all breast cancer types were selected. Out of 735 EMT genes, probes for 720 genes were identified and analyzed. Medium survival months for high expression and low expression cohorts, as well as the *p* values for the significance of their differences, were obtained from the website. We corrected the *p* values using Benjamini–Hochberg procedure with FDR of 0.05. Genes without significant difference in survival months were discarded for subsequent analysis (assuming zero difference for these genes produced very similar results). Differences in survival months for genes in each M clusters were aggregated, and *t*-tests were performed to compare means of the differences to 0.

Statistical tests and boxplots

Unless otherwise indicated, all *p* values were obtained with two-sided *t*-test assuming unequal variances. Other tests include two-sided Fisher's exact test for count data and two-sided Mann–Whitney *U*-test for continuous numerical data with distributions far from normal. In all boxplots, center lines indicate median values, box heights indicate the inter-quartile range of data, whiskers extend 1.5 times this range on either side, and outliers are indicated by black dots.

Mathematical modeling

We used a gene regulatory network that is simplified from earlier models.[4–6] We incorporated the effector E-genes and three types of M-genes in the network, and their regulations by TGF-β and ZEB1 are based on the analysis from this study (Figs. 2 and 3). To describe the system mathematically, we used a generic form of ordinary differential equations (ODEs) suitable for describing both gene expression and molecular interaction networks.[67–71] Each ODE system in the model has the form:

$$\begin{aligned} dX_i/dt &= \gamma_i(F(\sigma_i W_i) - X_i) \\ F(\sigma W) &= 1/(1+e^{-\sigma W}) \\ W_i &= \left(\omega_i^0 + \sum_j^N \omega_{j\to i} X_j\right) \\ i &= 1, \ldots, N \end{aligned} \quad (1)$$

Here, X_i is the activity or concentration of protein *i*. On a time scale $1/\gamma_i$, $X_i(t)$ relaxes toward a value determined by the sigmoidal function, *F*, which has a steepness set by σ. The basal value of *F*, in the absence of any influencing factors, is determined by ω_i^0. The coefficients $\omega_{j\to i}$ determine the influence of protein *j* on protein *i*. *N* is the total number of proteins in the network. All variables and parameters are dimensionless. One time unit in the simulations corresponds to approximately 1 day.

To model the AND logic gate on M1 genes regulated by TGF-β and ZEB1, we assumed that the total influences of TGF-β $(\omega_{\mathrm{TGF}\beta\to\mathrm{M1}}\mathrm{TGF}\beta)$ on M1 and ZEB1 $(\omega_{\mathrm{ZEB1}\to\mathrm{M1}}\mathrm{ZEB1})$ on M1 are both saturated at values less than $-\omega_{\mathrm{M1}}^0$, so that these signals do not activate M1 genes alone.

Steady-state analysis was performed by varying the parameters representing exogenous TGF-β or ZEB1 production rate. The total value of the state variables representing the M-genes (M1, M2, M3) was used to quantify the M phenotype, and the value of the 'E-genes' was used to quantify the E phenotype. The difference between the phenotype was used to determine the position of the state in the EMT spectrum.

Parameter values were selected to fit to the observation that TGF-β-induced EMT is irreversible under normal conditions. These values are listed in Supplementary Table 9.

Imaging analysis

MCF10A cells were transduced with nuclear RFP-expressing lentivirus (LV-RFP, a gift from Elaine Fuchs, Addgene plasmid #26001) and treated under conditions listed in Table 1. Cell movement dynamics in 2D culture was monitored and recorded by IncuCyte® for 40 h. Using binarized images, cells were identified and tracked by a Fiji package TrackMate. Cell trajectories longer than 24 frames (6 h) were used for analysis. Movement analysis was performed using similar metrics described in an earlier study.[72] Instantaneous velocity was computed as $v_\tau = \sqrt{v_x^2 + v_y^2}$, where $v_x = (x_\tau - x_{\tau-1})/(t_\tau - t_{\tau-1})$. Here, x_τ is the *x* coordinate at time τ, and $t_\tau - t_{\tau-1}$ is the time inverval between frames (15 min). The mean instantaneous velocity was calculated for each cell, and they were aggregated and compared across condtions. Scaled displacement was calculated with $\mathrm{SD} = |x(t_{\mathrm{end}}) - x(t_{\mathrm{start}})|/(t_{\mathrm{end}} - t_{\mathrm{start}})$, where $x(t_{\mathrm{start}})$ and $x(t_{\mathrm{end}})$ are the initial and final positional vectors of the trajectory, respectively. Straightness index was calculated with

$$\mathrm{SI} = (x_i(t_{\mathrm{end}}) - x_i(t_{\mathrm{start}}))/\sum_{\tau=t_{\mathrm{start}}+1}^{t_{\mathrm{end}}} (x_i(\tau) - x_i(\tau-1)) \quad (2)$$

or the ratio of the distance between the initial and final positions for each cell to the integrated distance traveled. The mean number of nearest neighbors was computed for each cell in each frame by counting the other cells within a 30 μm search radius. This value was then divided by the total number of frames of each trajectory. These summary statistics of the trajectories were compared between pairs of conditions listed in Table 2. Mann–Whitney *U*-test was performed to obtain statisitcal significance.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Sequencing data generated for this work have been deposited in the NCBI Gene Expression Omnibus (GEO) under accession number GSE124843. Computer code to reproduce the results of self-organizing maps, clustering and mathematical modeling is available upon request.

ACKNOWLEDGEMENTS

The authors would like to thank Ms. Hajime Nishimura and Ms. Mami Kishima for their technical assistance in preparing this article and Dr. Kiyotsugu Yoshikawa for kindly providing the TCGP reporter construct. The authors thank Ms. Esha Dutta for the assistance with survival analysis. This work was partially supported by the startup funds from The University of Tennessee, Knoxville to TH, and by Naito Memorial Foundation, Grant-in-Aid for Scientific Research (KAKENHI) on Innovative Areas, "Cellular Diversity" Grant Number JP18H05106 and KAKENHI Grant Number JP16KK0165 to K.W. This work was partially supported by a research grant from the Ministry of Education, Culture, Sport, Science and Technology of Japan for the RIKEN Center for Integrative Medical Sciences. Funding for open access to this research was provided by University of Tennessee's Open Publishing Support Fund.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: K.W., T.H. Performed the experiments and modeling: K.W., N.P., T.H. Analyzed the data: K.W., N.P., S.N., H.S., T.H. Wrote the manuscript: K.W., N.P., T.H. K.W. and N.P. are co-first authors.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Systems Biology and Applications* website (https://doi.org/10.1038/s41540-019-0097-0).

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial–mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **15**, 178 (2014).
2. Nieto, M. A., Huang, R. Y., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* **166**, 21–45 (2016).
3. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* **119**, 1420–1428 (2009).
4. Hong, T. et al. An Ovol2-Zeb1 mutual inhibitory circuit governs bidirectional and multi-step transition between epithelial and mesenchymal states. *PLoS Comput. Biol.* **11**, e1004569 (2015).
5. Lu, M., Jolly, M. K., Levine, H., Onuchic, J. N. & Ben-Jacob, E. MicroRNA-based regulation of epithelial-hybrid-mesenchymal fate determination. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1318192110 (2013).
6. 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).
7. Grosse-Wilde, A. et al. Stemness of the hybrid epithelial/mesenchymal state in breast cancer and its association with poor survival. *PLoS One* **10**, e0126522 (2015).
8. Tan, T. Z. et al. Epithelial–mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.* **6**, 1279–1293 (2014).
9. Zhao, M., Kong, L., Liu, Y. & Qu, H. dbEMT: an epithelial–mesenchymal transition associated gene resource. *Sci. Rep.* **5**, 11459 (2015).
10. Zavadil, J. & Böttinger, E. P. TGF-β and epithelial-to-mesenchymal transitions. *Oncogene* **24**, 5764 (2005).
11. Sanchez-Tillo, E. et al. ZEB1 represses E-cadherin and induces an EMT by recruiting the SWI/SNF chromatin-remodeling protein BRG1. *Oncogene* **29**, 3490 (2010).
12. Joseph, J. V. et al. TGF-β is an inducer of ZEB1-dependent mesenchymal transdifferentiation in glioblastoma that is associated with tumor invasion. *Cell Death Dis.* **5**, e1443 (2014).
13. Morrison, C. D., Parvani, J. G. & Schiemann, W. P. The relevance of the TGF-β Paradox to EMT-MET programs. *Cancer Lett.* **341**, 30–40 (2013).
14. David, C. J. et al. TGF-β tumor suppression through a lethal EMT. *Cell* **164**, 1015–1030 (2016).
15. Chaffer, C. L. et al. Poised chromatin at the ZEB1 promoter enables breast cancer cell plasticity and enhances tumorigenicity. *Cell* **154**, 61–74 (2013).
16. Krebs, A. M. et al. The EMT-activator Zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic cancer. *Nat. Cell Biol.* **19**, 518 (2017).
17. Kodzius, R. et al. CAGE: cap analysis of gene expression. *Nat. Methods* **3**, 211 (2006).
18. Forrest, A. R. R. et al. A promoter-level mammalian expression atlas. *Nature* **507**, 462 (2014).
19. Clark, N. R. et al. The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinforma.* **15**, 79 (2014).
20. Farina, A. R. & Mackay, A. R. Gelatinase B/MMP-9 in tumour pathogenesis and progression. *Cancers* **6**, 240–296 (2014).
21. Assefnia, S. et al. Cadherin-11 in poor prognosis malignancies and rheumatoid arthritis: common target, common therapies. *Oncotarget* **5**, 1458–1474 (2014).
22. Han, B. et al. FOXC1 activates smoothened-independent hedgehog signaling in basal-like breast cancer. *Cell Rep.* **13**, 1046–1058 (2015).
23. Stallings-Mann, M. L. et al. Matrix metalloproteinase induction of Rac1b, a key effector of lung cancer progression. *Sci. Transl. Med.* **4**, 142ra195 (2012).
24. Shinde, A. et al. Autocrine fibronectin inhibits breast cancer metastasis. *Mol. Cancer Res.* https://doi.org/10.1158/1541-7786.mcr-18-0151 (2018).
25. Mendez, M. G., Kojima, S.-I. & Goldman, R. D. Vimentin induces changes in cell shape, motility, and adhesion during the epithelial to mesenchymal transition. *FASEB J.* **24**, 1838–1851 (2010).
26. Gilmore, P. M. et al. BRCA1 interacts with and is required for paclitaxel-induced activation of mitogen-activated protein kinase kinase kinase 3. *Cancer Res.* **64**, 4148–4154 (2004).
27. Chang, Y. S. et al. Correlation between insulin-like growth factor-binding protein-3 promoter methylation and prognosis of patients with stage I non-small cell lung cancer. *Clin. Cancer Res.* **8**, 3669–3675 (2002).
28. Hanafusa, T. et al. Reduced expression of insulin-like growth factor binding protein-3 and its promoter hypermethylation in human hepatocellular carcinoma. *Cancer Lett.* **176**, 149–158 (2002).
29. Xue, A., Scarlett, C. J., Jackson, C. J., Allen, B. J. & Smith, R. C. Prognostic significance of growth factors and the urokinase-type plasminogen activator system in pancreatic ductal adenocarcinoma. *Pancreas* **36**, 160–167 (2008).
30. Gregory, P.A. et al. An autocrine TGF-beta/ZEB/miR-200 signaling network regulates establishment and maintenance of epithelial-mesenchymal transition. *Mol. Biol. Cell* **22**, 1686–1698 (2011).
31. Steinway, S. N. et al. Network modeling of TGFbeta signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint sonic hedgehog and Wnt pathway activation. *Cancer Res.* **74**, 5963–5977 (2014).
32. Steinway, S. N. et al. Combinatorial interventions inhibit TGFbeta-driven epithelial-to-mesenchymal transition and support hybrid cellular phenotypes. *NPJ Syst. Biol. Appl.* **1**, 15014 (2015).
33. Taube, J. H. et al. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc. Natl Acad. Sci. USA* **107**, 15449–15454 (2010).
34. Gao, D., Vahdat, L. T., Wong, S., Chang, J. C. & Mittal, V. Microenvironmental regulation of epithelial–mesenchymal transitions in cancer. *Cancer Res.* **72**, 4883–4889 (2012).
35. Yu, M. et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science* **339**, 580–584 (2013).
36. Lánczky, A. et al. miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. *Breast Cancer Res. Treat.* **160**, 439–446 (2016).
37. Li, T. et al. Associations of tumor suppressor SPARCL1 with cancer progression and prognosis. *Oncol. Lett.* **14**, 2603–2610 (2017).
38. Lin, X. et al. HEXIM1 as a robust pharmacodynamic marker for monitoring target engagement of BET family bromodomain inhibitors in tumors and surrogate tissues. *Mol. Cancer Ther.* **16**, 388–396 (2017).
39. Yang, J. et al. Integrated analysis of microfibrillar-associated proteins reveals MFAP4 as a novel biomarker in human cancers. *Epigenomics* **11**, 1635–1651 (2018).
40. George, J. T., Jolly, M. K., Xu, S., Somarelli, J. A. & Levine, H. Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric. *Cancer Res.* **77**, 6415–6428 (2017).
41. Jolly, M. K. et al. Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget* **7**, 27067 (2016).
42. Mooney, S. M. et al. The GRHL2/ZEB feedback loop—a key axis in the regulation of EMT in breast cancer. *J. Cell. Biochem.* **118**, 2559–2570 (2017).
43. Cieply, B., Farris, J., Denvir, J., Ford, H. L. & Frisch, S. M. Epithelial–mesenchymal transition and tumor suppression are controlled by a reciprocal feedback loop between ZEB1 and Grainyhead-like-2. *Cancer Res.* **73**, 6299–6309 (2013).
44. Bracken, C. P. et al. A double-negative feedback loop between ZEB1-SIP1 and the microRNA-200 family regulates epithelial-mesenchymal transition. *Cancer Res.* **68**, 7846–7854 (2008).
45. Preca, B.-T. et al. A novel ZEB1/HAS2 positive feedback loop promotes EMT in breast cancer. *Oncotarget* **8**, 11530 (2017).
46. Jolly, M. K. et al. Interconnected feedback loops among ESRP1, HAS2, and CD44 regulate epithelial–mesenchymal plasticity in cancer. *APL Bioeng.* **2**, 031908 (2018).

47. 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).
48. Celià-Terrassa, T. et al. Hysteresis control of epithelial-mesenchymal transition dynamics conveys a distinct program with enhanced metastatic ability. *Nat. Commun.* **9**, 5005 (2018).
49. Pastushenko, I. et al. Identification of the tumour transition states occurring during EMT. *Nature* **556**, 463–468 (2018).
50. Wellner, U. et al. The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs. *Nat. Cell Biol.* **11**, 1487 (2009).
51. Prieto-García, E., Díaz-García, C. V., García-Ruiz, I. & Agulló-Ortuño, M. T. Epithelial-to-mesenchymal transition in tumor progression. *Med. Oncol.* **34**, 122 (2017).
52. Fischer, K. R. et al. Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. *Nature* **527**, 472–476 (2015).
53. Zheng, X. et al. Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. *Nature* **527**, 525–530 (2015).
54. Hardy, S. D., Shinde, A., Wang, W.-H., Wendt, M. K. & Geahlen, R. L. Regulation of epithelial-mesenchymal transition and metastasis by TGF-β, P-bodies, and autophagy. *Oncotarget* **8**, 103302 (2017).
55. Nieto, M. A. Epithelial plasticity: a common theme in embryonic and cancer cells. *Science* **342**, 1234850–1234850 (2013).
56. Bocci, F., Kumar Jolly, M. & Onuchic, J. N. A biophysical model of epithelial–mesenchymal transition uncovers the frequency and size distribution of circulating tumor cell clusters across cancer types. *bioRxiv*, 563049, https://doi.org/10.1101/563049 (2019).
57. Jolly, M. K. et al. Coupling the modules of EMT and stemness: a tunable 'stemness window'model. *Oncotarget* **6**, 25161 (2015).
58. Guaita, S. et al. Snail induction of epithelial to mesenchymal transition in tumor cells is accompanied by MUC1 repression and ZEB1 expression. *J. Biol. Chem.* **277**, 39209–39216 (2002).
59. Dave, N. et al. Functional cooperation between Snail1 and twist in the regulation of ZEB1 expression during epithelial to mesenchymal transition. *J. Biol. Chem.* **286**, 12024–12032 (2011).
60. Li, R. et al. A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. *Cell Stem Cell* **7**, 51–63 (2010).
61. Zeisberg, M. et al. Fibroblasts derive from hepatocytes in liver fibrosis via epithelial to mesenchymal transition. *J. Biol. Chem.* **282**, 23337–23347 (2007).
62. Ye, Y., Kang, X., Bailey, J., Li, C. & Hong, T. An enriched network motif family regulates multistep cell fate transitions with restricted reversibility. *PLoS Comput. Biol.* **15**, e1006855 (2019).
63. Ta, C. H., Nie, Q. & Hong, T. Controlling stochasticity in epithelial-mesenchymal transition through multiple intermediate cellular states. *Discret. Contin. Dyn. Syst.-Ser. B* **21**, https://doi.org/10.3934/dcdsb.2016047 (2016).
64. Watanabe, K. et al. Mammary morphogenesis and regeneration require the inhibition of EMT at terminal end buds by Ovol2 transcriptional repressor. *Dev. Cell* **29**, 59–74 (2014).
65. Wehrens, R. & Buydens, L. M. C. Self- and super-organizing maps in R: the kohonen package. *J. Stat. Softw.* **21**, https://doi.org/10.18637/jss.v021.i05 (2007).
66. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
67. Mjolsness, E., Sharp, D. H. & Reinitz, J. A connectionist model of development. *J. Theor. Biol.* **152**, 429–453 (1991).
68. Li, C.-J. et al. MicroRNA filters Hox temporal transcription noise to confer boundary formation in the spinal cord. *Nat. Commun.* **8**, 14685 (2017).
69. Hong, T., Oguz, C. & Tyson, J. J. A mathematical framework for understanding four-dimensional heterogeneous differentiation of CD4+ T cells. *Bull. Math. Biol.*, 1–19, https://doi.org/10.1007/s11538-015-0076-6 (2015).
70. Hong, T., Xing, J., Li, L. & Tyson, J. J. A mathematical model for the reciprocal differentiation of T helper 17 cells and induced regulatory T cells. *PLoS Comput. Biol.* **7**, e1002122–e1002122 (2011).
71. Hong, T., Xing, J., Li, L. & Tyson, J. J. A simple theoretical framework for understanding heterogeneous differentiation of CD4+ T cells. *BMC Syst. Biol.* **6**, 66–66 (2012).
72. Wong, I. Y. et al. Collective and individual migration following the epithelial–mesenchymal transition. *Nat. Mater.* **13**, 1063–1071 (2014).