npj | systems biology and applications

Article

Published in partnership with the Systems Biology Institute

https://doi.org/10.1038/s41540-025-00512-2

Modeling the dynamics of EMT reveals genes associated with pan-cancer intermediate states and plasticity

Check for updates

MeiLu McDermott[1], Riddhee Mehta[1], Evanthia T. Roussos Torres[2] & Adam L. MacLean[1] ✉

Epithelial–mesenchymal transition (EMT) is a cell state transition co-opted by cancer that drives metastasis via stable intermediate states. Here we study EMT dynamics to identify marker genes of highly metastatic intermediate cells via mathematical modeling with single-cell RNA sequencing (scRNA-seq) data. Across multiple tumor types and stimuli, we identified genes consistently upregulated in EMT intermediate states, many previously unrecognized as EMT markers. Bayesian parameter inference of a simple EMT mathematical model revealed tumor-specific transition rates, providing a framework to quantify EMT progression. Consensus analysis of differential expression, RNA velocity, and model-derived dynamics highlighted *SFN* and *NRG1* as key regulators of intermediate EMT. Independent validation confirmed *SFN* as an intermediate state marker. Our approach integrates modeling and inference to identify genes associated with EMT dynamics, offering biomarkers and therapeutic targets to modulate tumor-promoting cell state transitions driven by EMT.

Cell state transitions are phenotypic changes in the state of a cell, primarily driven by transcriptional programs. Such phenotypic transitions underlie development, regeneration, and cancer. Our ability to interrogate cell state transitions and their consequences has dramatically increased with advances in single-cell genomics[1]. We can dissect the timing of key events as cells change state[2] and identify transient or intermediate states[3]. Efforts to produce a comprehensive catalog of cell states are underway[4], yet large gaps in our understanding remain: both regarding cell states and even more regarding the transitions they undertake. We do not have satisfactory explanations of what are the initiating factors of a cell state transition, nor what is the relationship between the dynamics of cell phenotypic change and the transcriptional dynamics acting within the cell.

The epithelial-to-mesenchymal transition (EMT), during which epithelial cells become mesenchymal or mesenchymal-like[5], is an exemplary cell state transition. EMT is necessary during development and wound healing and is co-opted by cancer, where it is a crucial component of metastasis. Understanding EMT is thus imperative to slowing or preventing metastasis, the leading cause of death from cancer[6]. Classical conceptions of EMT characterize a binary process, with cells being either completely epithelial or mesenchymal[5]. However, experimental and theoretical studies have demonstrated the existence of EMT intermediate states[7–12]. Pan-cancer studies of intermediate EMT states have revealed insight into transcriptomic signatures underlying EMT transformation[13]. The intermediate state displays partial EMT phenotypes, with characteristics of both epithelial and mesenchymal states, and may also be called partial EMT, hybrid EMT, or an E/M state[14]. EMT intermediate states are closely tied with the concept of epithelial–mesenchymal plasticity (EMP): dynamic, bidirectional transitions through multiple EMT states.

EMT intermediate states are found in both non-malignant EMT and cancer[14,15]. The relevance of targeting these states is compelling: EMT intermediate states have been associated with circulating tumor cells[16–18] and metastasis[19], perhaps even more potently than mesenchymal cells alone[20,21]. We focus here on stable EMT intermediate states: biologically, this refers to cells in a state that can be isolated and persist under sufficient conditions; mathematically, stability is defined via the Lyapunov exponents of a dynamical system[22]. EMT intermediate states have been described as "metastable" in the literature, which in this case refers to stable cell states with small basins of attraction. EMT intermediate state cells may be hard to observe in part due to their rarity (small population sizes or small basins of attraction) or their location (existing at the margins rather than throughout a tissue[23]), although they are not necessarily a minority of cells in a sample.

Mathematical models of EMT have predicted and identified intermediate states, using transcriptional networks that can successfully capture both the steady states of the system and its dynamic properties[9,24–28]. These transcription models of EMT, typically regulated by transforming growth factor-beta (TGF-β), primarily focus on a core network with transcription

[1]Department of Quantitative and Computational Biology, Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA, USA. [2]Department of Medicine, Division of Medical Oncology, Keck School of Medicine, Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA. ✉e-mail: macleana@usc.edu

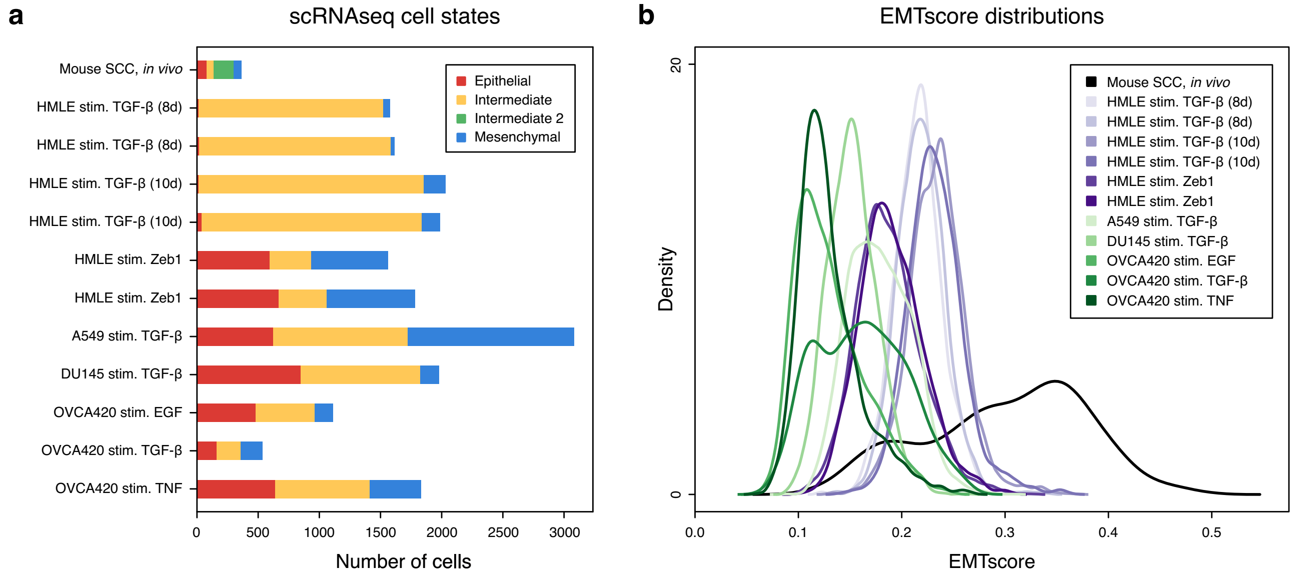


Fig. 1 | Quantification and distribution of EMT states across scRNA-seq cell lines and stimuli. **a** Cell count of EMT states per cancer sample, including intermediate states. Cell states were identified via clustering and gene expression. **b** Kernel density estimates of the EMTscore distributions for each scRNA-seq sample. Single-cell EMT scores were assigned via Hallmark EMT genes from MSigDB.

factors ZEB, SNAIL, and OVOL, and micro-RNAs miR200 and miR34. Although greater attention has been paid to the transcriptional dynamics, there has also been mathematical modeling of the cell population dynamics during EMT, as reviewed in ref. 29.

Integrating single-cell genomics with mathematical models offers means to infer dynamic properties from high-dimensional systems[30,31]. EMT, with its relatively straightforward trajectory (non-branching, non-cyclical), lends itself well to analysis via trajectory inference (pseudotime)[32], albeit not taking into account the spatial components of the cell fate decisions which can be decidedly more complex[33]. Trajectory inference coupled with mathematical modeling has led to insight into the initiation and timing of EMT[34]. Despite limitations in inferring Markovian cell dynamics from single-cell data[35], experimental methodologies such as metabolic labeling[36] or lineage tracing[21] can overcome these challenges. Here, we take an alternative approach to inferring the population dynamics model directly from data[35,37], and (in keeping with the observation that cell state transition dynamics are non-Markovian[38]) we propose a population model of EMT cell state transitions a priori. We subsequently learn rates of cell state transition for each individual sample via Bayesian parameter inference of the cell dynamics over pseudotime.

Here we use single-cell RNA sequencing (scRNA-seq) data to fit mathematical models of EMT population dynamics across various tumor types and stimuli. Parameter inference across these different conditions reveals shared and distinct properties of the routes of EMT. We identify shared genes associated with EMT intermediate states across tumor types via differential expression and differential RNA velocity analyses. By comparing intermediate state genes with inferred EMT parameters, we identify genes associated with EMT dynamics—that is, genes that speed up or slow down EMT. We confirm top predictions by an independent analysis of EMT in a new cell type, demonstrating how these methods offer novel means to identify biomarkers or potential targets during cell state transitions.

Results

Single-cell analysis of EMT across cancer types & stimuli identifies a spectrum of EMT states

To characterize trajectories across a spectrum of EMT, we studied twelve scRNA-seq datasets across five cancer types. Cells were processed and clustered to identify cell states. We found evidence for three cell states in each of the in vitro cell populations and four states in the in vivo mouse skin squamous cell carcinoma (SCC) sample (Supplementary Figs. 1 and 2). Silhouette scores broadly support the selected clustering resolutions, balancing cluster quality and number of states (Supplementary Fig. 3). Clusters were labeled based on EMT markers from the literature, including Hallmark EMT genes from the Molecular Signatures Database (MSigDB)[39] and epithelial cell genes from PanglaoDB[40]. Distinct clusters representing epithelial and mesenchymal cell types were identified in each dataset, although the relative sizes of these clusters varied widely (Fig. 1a). In all datasets, at least one cluster expressing combinations of epithelial and mesenchymal marker genes was identified as an intermediate state. Certain samples from ref. 41 that did not exhibit a clear EMT were excluded from further analyses (Supplementary Figs. 4 and 5). This is in agreement with ref. 41, who also found that certain conditions did not permit a full EMT within the experimental timeframe.

EMT scores were assigned to single cells across all datasets (Fig. 1b). Single cells were each assigned an EMTscore via UCell[42] using MSigDB Hallmark EMT genes. Each sample exhibited a range of EMTscore, reproducible by replicate and varying considerably by cell type and stimulus. Notably, not only the variance but also the start and end points vary by cell type, highlighting differences not only in EMT but also in the "epithelial-ness" of different cell types. Samples excluded from analysis due to lack of/incomplete EMT, as identified by marker gene expression, exhibited little to no variation in EMTscore (Supplementary Fig. 5), confirming the lack of cell state transition under the tested conditions. The in vivo EMT in mouse SCC exhibited the largest range of EMTscore by a wide margin, highlighting the increase in heterogeneity among single cells during a spontaneous, unstimulated, environment-dependent EMT. Since an additional intermediate state was identified in this dataset, in line with previous work[43,44], the data suggest that both the number of attractor states and the size of their basins of attraction are larger for cells in their natural environment than cell line-derived models stimulated in vitro.

Shared marker genes of intermediate EMT states are associated with extracellular function

EMT can proceed along many paths[11], and both cell/treatment-specific and consensus EMT pathways are important to study in different contexts. Here, we focus on the shared properties of EMT cell state transitions. To study intermediate state gene expression across an EMT spectrum, we performed differential gene expression and differential RNA velocity analysis across

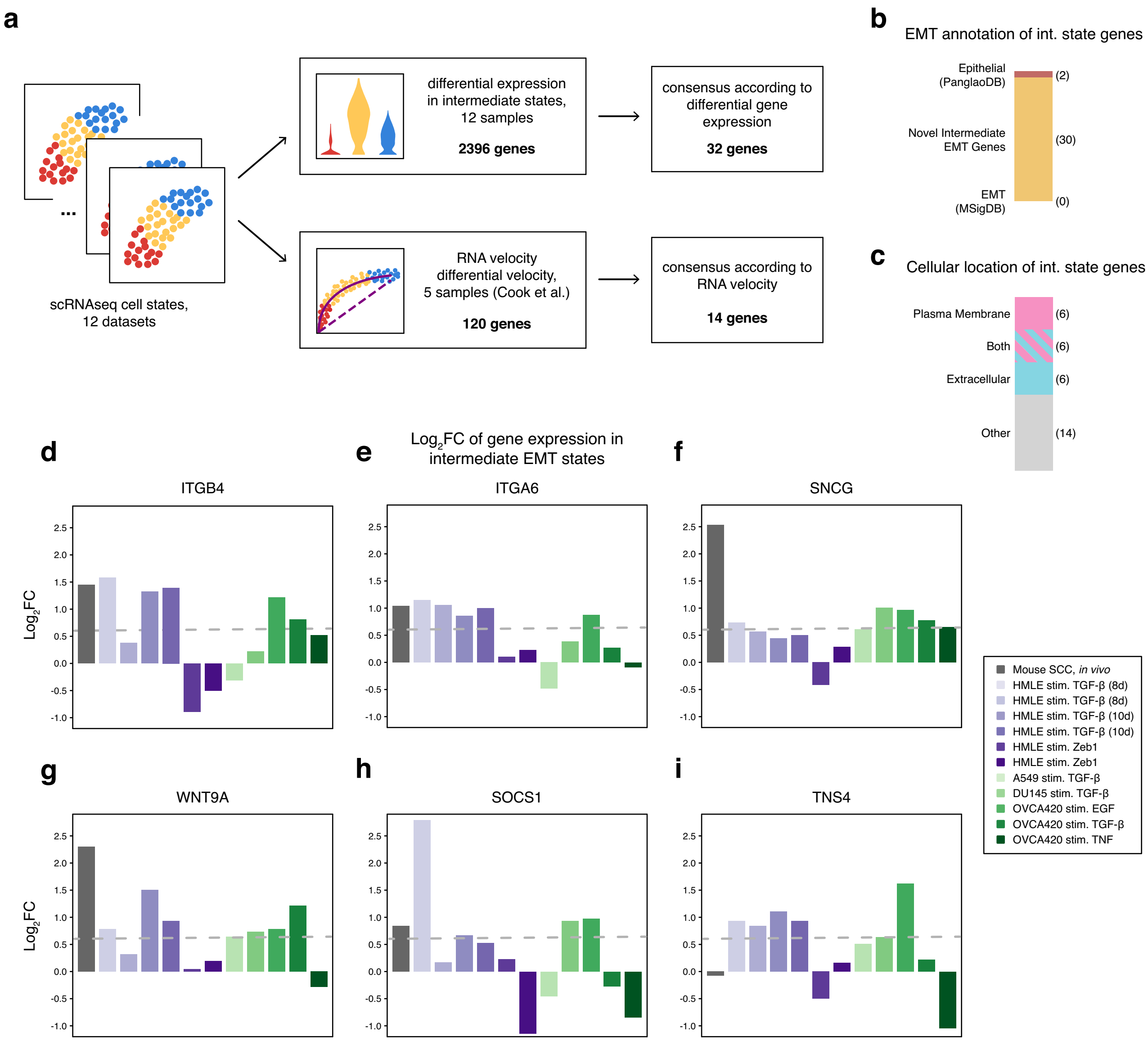


Fig. 2 | Identification of pan-cancer intermediate EMT state genes. a Data analysis pipeline: pan-cancer intermediate EMT marker genes were identified using differential expression and compared against genes differentially regulated via RNA velocity. **b** Annotation of predicted intermediate EMT genes by canonical epithelial/mesenchymal gene sets. **c** Annotation of predicted intermediate EMT genes by cellular location. **d**–**i** Comparison of expression of predicted intermediate EMT genes by log$_2$FC (in the intermediate state) across samples. The dashed line represents 1.5-fold change (log$_2$FC >0.58).

intermediate states in different cell populations (Fig. 2a). We identified differentially expressed genes for intermediate states in each sample (2396 genes total) and examined shared intermediate state-specific genes, defined as those upregulated in an intermediate state relative to epithelial/mesenchymal states. Using a log$_2$ fold change (log$_2$FC) threshold of +0.58 (1.5-fold change) in at least five samples, we identified 32 genes shared among EMT intermediate states (Supplementary Fig. 6 and Supplementary Data 1). No single gene is universally upregulated across intermediate EMT states; notably, the same holds for epithelial and mesenchymal states across all datasets. This observed heterogeneity is consistent with previous findings that canonical EMT genes[45] and empirically derived EMT gene sets[46] exhibit substantial variability, underscoring the complex & context-dependent nature of EMT.

Among the 32 genes shared across EMT intermediate states, most were absent from canonical EMT or epithelial gene sets (Fig. 2b). Two predicted intermediate state genes, *ITGB4* and *SFN*, are annotated as epithelial genes in PanglaoDB[40], although the literature on these genes is complicated: integrin *β*4 (*ITGB4*) (Fig. 2d) was initially identified in epithelial cells and tumors[47] but has also been linked to promoting EMT in hepatocellular and pancreatic carcinoma[48,49]. *ITGB4* pairs with another intermediate state gene, integrin *α*6 (*ITGA6*) (Fig. 2e), to form the *α*6*β*4 complex, which is implicated in promoting EMT characteristics in hepatocellular carcinoma cells[50]. Stratifin (*SFN*) (Supplementary Fig. 6) is annotated as epithelial (named for its role in the stratification of epithelial cells[51]) but is also linked to cell migration and EMT markers in cervical and hepatocellular carcinoma[52–54]. The apparent contradictory roles of both *ITGB4* and *SFN* as marking for both epithelial and mesenchymal states can be reconciled if these genes are in fact markers of an intermediate EMT state, as predicted by our analysis.

A majority of predicted intermediate EMT marker genes encode proteins localized in the extracellular space, on the plasma membrane or as secreted signaling factors (Fig. 2c). Gamma-synuclein (*SNCG*), upregulated across multiple cell lines (Fig. 2f), is found in the extracellular exosome. It plays a role in suppressing mesenchymal markers including *CDH2* (N-cadherin) and *VIM*[55] while promoting cancer cell migration[56–58]. Other notable upregulated genes include *WNT9A*, *IL4R*, and *IL6R*. Wnt-9a (*WNT9A*) (Fig. 2g) is a secreted protein in the canonical Wnt/*β*-catenin signaling pathway that is implicated in partial EMT by mediating cell adhesion[59]. *IL4R* and *IL6R* (Supplementary Fig. 6) are interleukin cell surface receptors, with their cytokines IL4, IL13, and IL6 associated with EMT promotion[60–62]. Interestingly, *SOCS1* (suppressor of cytokine signaling 1) (Fig. 2h) is a negative regulator of IL6 yet conversely has been found to promote EMT[63], highlighting the bidirectional signaling at play during the establishment of intermediate EMT states.

Tensin 4 (*TNS4*) (Fig. 2i) is involved in focal adhesion & integrin interaction and promotes EMT and cell motility[64,65]. Tubulointerstitial nephritis antigen-like 1 (*TINAGL1*) (Supplementary Fig. 6) encodes another secreted protein that binds directly to certain integrins, and it is found to both promote and inhibit metastasis in different cancers in vivo[66,67]. Both *TNS4* and *TINAGL1* interact with epidermal growth factor receptor *EGFR*, yet their effects are contradictory: *TNS4* reduces *EGFR* degradation[68], while *TINAGL1* binds directly to *EGFR* and suppresses *EGFR* signaling[66]. These opposing interactions may again reflect the dynamic balance necessary to sustain the intermediate EMT state.

Overall, many genes associated with the intermediate EMT state exhibit conflicting roles in the literature, including *ITGB4*, *SFN*, *IL4R* and *IL6R* with *SOCS1*, and *TINAGL1* with *TNS4*. These genes can contribute both to the promotion and inhibition of EMT as well as the balance between epithelial and mesenchymal states. This duality underscores the dynamic nature of EMT and the importance of intermediate states. Gene set enrichment analysis (GSEA) of intermediate state genes identified enriched pathways (Supplementary Data 2), though most represented general cellular processes or were supported by only 2–3 genes. GSEA also revealed enrichment of transcription factor binding sites, particularly for AP-1, suggesting a regulatory role during transitions through EMT intermediate states.

The tumor microenvironment (TME) likely influences EMT-driven cell state transitions, as indicated by greater variability in intermediate states and EMT scores in vivo compared to in vitro (Fig. 1). We identified genes marking EMT intermediate states across all datasets (TME + non-TME) and compared them to those shared only among in vitro datasets (non-TME). Including in vivo data revealed 32 differentially expressed genes in intermediate states (Fig. 2), whereas excluding it reduced this number to 22 (Supplementary Fig. 7), hinting at the complexity the TME introduces. However, gene ontology and pathway enrichment analysis via PantherDB did not find any significant terms that distinguished TME from non-TME intermediate EMT genes, suggesting subtle regulatory influences.

Differential regulation via RNA velocity reveals EM plasticity genes in EMT intermediate states

To investigate dynamically regulated genes during EMT, we performed differential RNA velocity across EMT cell states[69,70]. Fourteen genes had differential velocity (DV) in the intermediate state in at least three of the five[41] samples, which includes cells from lung, prostate, and ovarian tumors (Supplementary Fig. 8). Of the 14 DV genes, all but one encode proteins located extracellularly or in the plasma membrane (Fig. 3a). Several of these genes are involved in focal adhesion, including integrins *ITGA2* and *ITGB4*, laminins *LAMC2* and *LAMB3*, collagen *COL4A2*, and plasma membrane caveolae component *CAV1*. Eleven of the DV genes have annotated signal peptide sequences, underscoring their designation as secretory/membrane proteins[71,72].

DV genes showed greater overlap with canonical EMT gene sets than the intermediate state marker genes we identified (Fig. 3b). This is expected, as genes actively upregulated in EMT intermediate states are more likely to overlap with mesenchymal markers. Epithelial–mesenchymal plasticity (EMP), i.e., bidirectional cell state transitions between epithelial and mesenchymal phenotypes[46], is also characteristic of the DV genes identified. This overlap supports EMP conceptually: capricious cells require dynamic changes in gene expression to change state.

Comparison of DV genes across samples revealed a variety of responses: some genes were shared across different cell types and conditions, while others were specific to certain conditions. Genes upregulated regardless of cell type or stimulus included *LAMC2* (Fig. 3c), *FRMD6*, and *SERPINE1* (Supplementary Fig. 8). In contrast, and perhaps unsurprisingly, TGF-β-induced protein *TGFBI* (Fig. 3d) was upregulated in various cell types only when stimulated by TGF-β. A similar pattern was observed for *COL4A2* (Supplementary Fig. 8). Genes upregulated by multiple stimuli in one cell type, human ovarian OVCA420 cells, included *ITGB4* (Fig. 3e), *CAV1*, *HMGA2*, *F3*, and *LAMB3* (Supplementary Fig. 8). Overall, RNA velocity analysis elucidates gene regulation during EMT. Most differentially regulated genes are specific to a stimulus or cell line; fewer are conserved across conditions. There is substantial overlap between actively regulated genes during EMT and those linked to EMP, highlighting the role of dynamic transitions between cell states during EMT.

Mathematical modeling & parameter inference quantifies EMT population dynamics

Gene expression is not static: life arises from dynamics. To study the dynamics of EMT in more depth, we developed a mathematical model describing cell state transitions during EMT (Fig. 4a). The model is characterized by rate parameters for transitions between epithelial (E), intermediate (I), and mesenchymal (M) states, such as $E \rightarrow I$ at rate k_1 (Fig. 4b and Supplementary Fig. 9A). These rate parameters were fit to scRNA-seq data, characterizing cell state transitions during EMT across pseudotime. Multiple pseudotime trajectories were calculated for each sample, rooted by different epithelial cells, to estimate the mean & variance in pseudotime based on root node selection. Cell state proportions across pseudotime, representing cell population dynamics during EMT, were fitted to the model.

EMT dynamics for each dataset were fit using Bayesian parameter inference (Fig. 4c, Supplementary Fig. 9, and Supplementary Table 2). Differences in EMT dynamics were observed across different datasets, both by cell type and by stimulus. For instance, the intermediate state persisted longer in HMLE cells compared to A549 or OVCA420 cells. Analysis of the parameter posterior distributions for each fitted EMT trajectory revealed similarities and differences in EMT dynamics (Fig. 4d). Dividing the posterior space into three approximate regions: $k_1 \approx k_2$ (similar transition rates across EMT); $k_1 > k_2$ (faster transition rates for $E \rightarrow I$ than $I \rightarrow M$); and $k_1 < k_2$ (faster transition rates for $I \rightarrow M$ than $E \rightarrow I$) highlights how both cell type and stimulus can strongly impact EMT dynamics. For example, OVCA420 cells exhibited $k_1 > k_2$ dynamics regardless of stimulus, where $k_1 > k_2$ implies a larger/more stable intermediate state. In contrast, HMLE cells exhibited $k_1 > k_2$ dynamics for TGF-β stimulation but $k_1 < k_2$ for ZEB1 stimulation, indicating that the persistence/stability of the HMLE intermediate state depends on the stimulating factor.

An inverse proportion relationship is evident across cell types/stimuli and within a sample; this concordance is notable since more generally different types of parameter covariation can exist[30]. This analysis highlights how EMT intermediate persistence and stability depend on the intrinsic properties of the EMT experiment, with different carcinomas exhibiting greater or lesser sensitivity to EMT-inducing factors and thus affecting EMT progression.

For samples from ref. 41 with biological timepoints, we compared inferred pseudotime with experimental time (Supplementary Fig. 10). Pseudotime reconstructs latent dynamics by inferring a trajectory that does not necessarily align with discrete biological sampling; for instance, ref. 41 measured cell states at five timepoints, whereas we infer trajectories using 15 pseudotime points. In most samples, the absence of clear state transitions over experimental time precluded model fitting, underscoring the utility of pseudotime in capturing cell state dynamics. However, in two cases (A549 and OCA420 stimulated with TGF-β), cell state transitions occurred directly over experimental time, allowing us to fit a mathematical model to these trajectories. The lower temporal resolution of experimental sampling (two timepoints capturing the intermediate state) compared to pseudotime (four to five points) limits the precision of the intermediate state dynamics, highlighting the strength of pseudotime analysis in revealing the cell state transition dynamics that may not be observed by sparse experimental sampling.

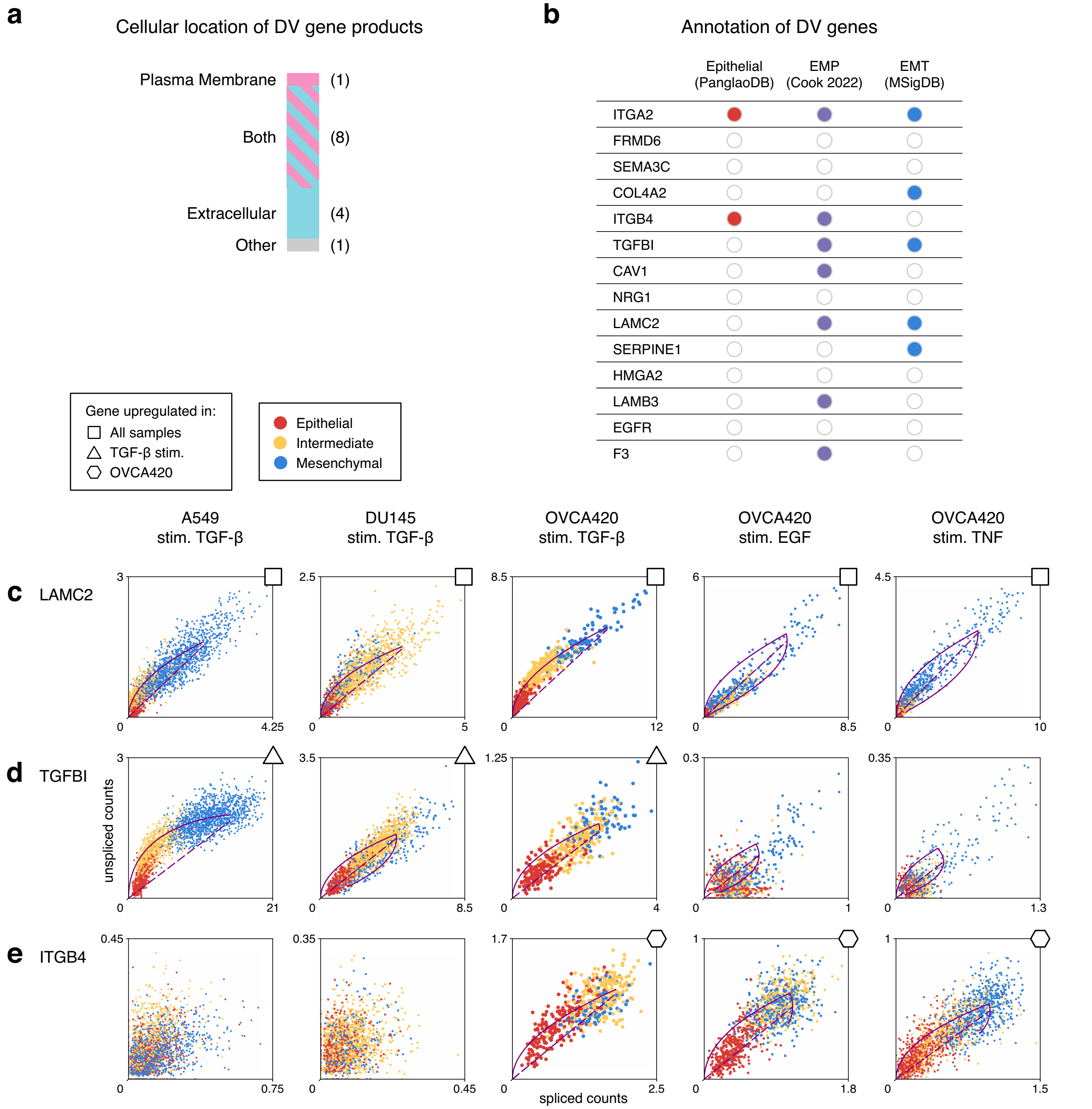
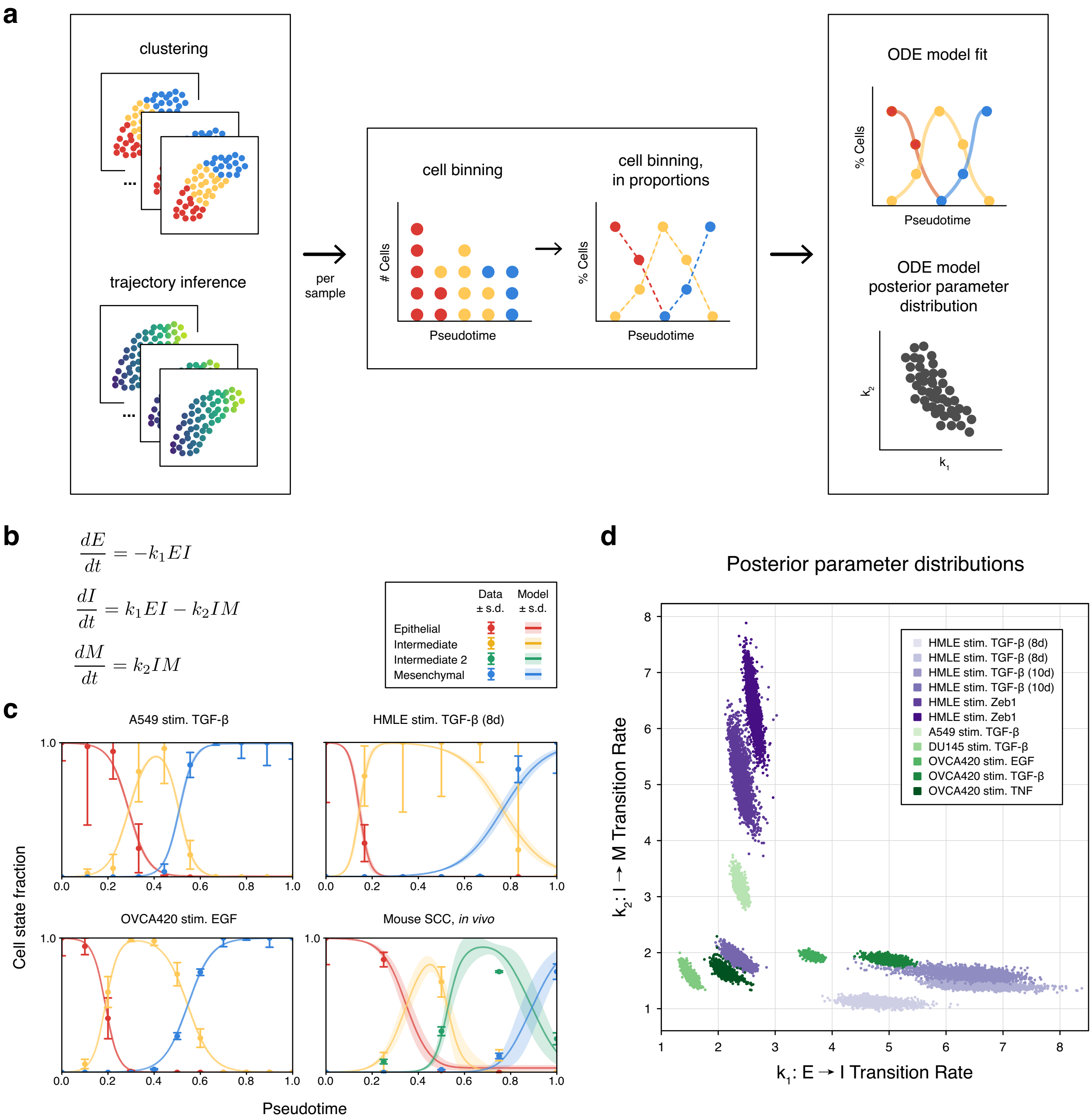


Fig. 3 | Differentially regulated intermediate EMT state genes identified via RNA velocity. a Annotation of cellular locations for genes differentially regulated in the intermediate state, identified through differential velocity (DV) analysis. **b** Annotation of the EMT properties of DV genes by comparison with three EMT marker gene sources. EMP: epithelial–mesenchymal plasticity. **c** Examples of DV genes upregulated in intermediate EMT states across different conditions. Solid line represents the dynamical model fit; dashed line represents the inferred steady state. *LAMC2* is upregulated across samples from different cell lines & stimuli. **d** *ITGB4* is upregulated only with a specific stimulus: TGF-β. **e** *ITGB4* is upregulated only in a specific tumor type: ovarian cells OVCA420.

Consensus analysis predicts that *SFN* and *NRG1* influence intermediate cell state dynamics during EMT

To identify genes influencing intermediate EMT dynamics, we studied associations between intermediate EMT genes and fitted parameters of the mathematical model. A gene's positive correlation with k_1 indicates faster transition $E \longrightarrow I$, while a negative correlation with k_2 means a slower transition $I \longrightarrow M$; either correlation suggests that the gene is associated with a more persistent intermediate state. Genes with significant Spearman's correlation were compared with differential expression and differential velocity genes in intermediate states, and those supported by multiple lines of evidence were consolidated into a consensus gene list of 14 genes (Supplementary Table 3 and Supplementary Fig. 11). The majority of intermediate EMT dynamics genes were located at the plasma membrane or in the extracellular region (Fig. 5a). Of the 14 predicted intermediate EMT dynamics genes, three were identified in a prior EMP study[46] (Fig. 5b), consistent with the conceptual overlap between intermediate EMT dynamics and EMP. Notably, there is no overlap between intermediate EMT dynamics genes and those from hallmark EMT (mesenchymal) genes, demonstrating that our proposed gene set is novel and distinct from previous EMT gene sets.



$$\frac{dE}{dt} = -k_1 EI$$

$$\frac{dI}{dt} = k_1 EI - k_2 IM$$

$$\frac{dM}{dt} = k_2 IM$$

Fig. 4 | Mathematical model of EMT dynamics fitted to scRNA-seq data. **a** Workflow to infer dynamic EMT transition rates from scRNA-seq data. For each sample, clustering and trajectory inference information was processed to quantify cell states over pseudotime. A mathematical model was then fit to each sample to infer parameter posterior distributions. **b** Mathematical model representing transitions from epithelial (*E*) to intermediate (*I*) to mesenchymal (*M*) state cells. k_1 is the transition rate $E \rightarrow I$; k_2 is the transition rate $I \rightarrow M$. Additional intermediate states can be seamlessly added (Supplementary Fig. 9A). **c** Model fits the following parameter inference: data vs. trajectory simulations, with simulation parameters sampled from the posterior of each model. **d** Posterior parameter distributions of the model for each fitted sample.

Two predicted EMT dynamics genes had the strongest support (three lines of evidence each; Supplementary Table 3). *NRG1* was the only gene identified in all three analyses, while *SFN* was the only gene with intermediate EMT differential expression and significant correlations with both k_1 and k_2 transition rates. Stratifin (*SFN*) was positively correlated with k_1 ($E \rightarrow I$) and negatively correlated with k_2 ($I \rightarrow M$) across cancer samples (Fig. 5c), suggesting that it stabilizes the intermediate EMT state. Although RNA velocity for *SFN* was not captured due to insufficient counts, it was differentially expressed in intermediate states. Neuregulin 1 (*NRG1*) was negatively correlated with k_2 ($I \rightarrow M$), suggesting it slows the exit from the intermediate state (Fig. 5d), and *NRG1* was also significant in intermediate EMT differential expression and velocity (Supplementary Fig. 8).

Consensus gene analysis predicts that *SFN* promotes transitions from an epithelial state to the metastatic intermediate EMT state. This prediction helps to reconcile literature, which reports both epithelial and pro-EMT roles for *SFN*. Named for its expression in stratified epithelial cells[51], *SFN* can be secreted and is found in extracellular vesicles[73]. Recombinant *SFN* treatment has been shown to significantly enhance extracellular matrix degradation in human dermal fibroblasts in vitro[74]. Despite its epithelial association, *SFN* knockdown in in vitro models has led to reduced

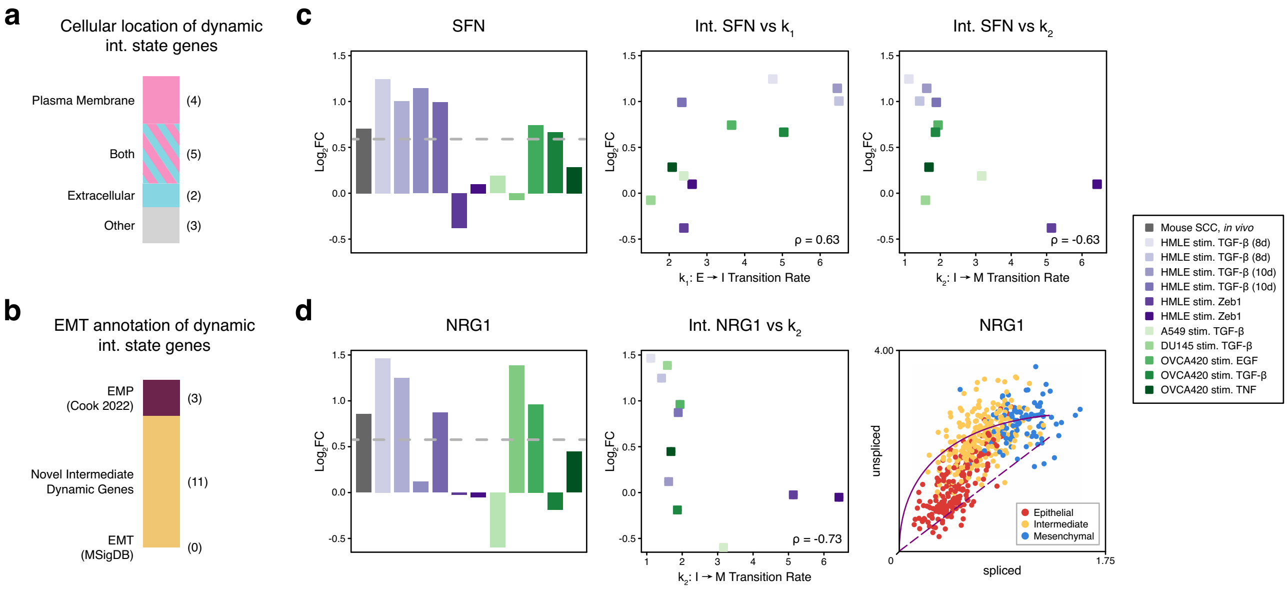


Fig. 5 | Genes influencing EMT intermediate state dynamics are identified via consensus analysis of expression, regulation, and parameter inference. **a** Annotation of predicted dynamic intermediate state genes by cellular location. **b** Annotation of predicted dynamic intermediate state genes by EMT marker gene sources. **c** *SFN* is a predicted dynamic intermediate state gene, with its differential expression in multiple EMT intermediate states and significant correlations with model parameters k_1 and k_2. **d** *NRG1* is another predicted dynamic intermediate state gene, with its differential expression in multiple EMT intermediate states, significant negative correlation with model parameter k_2, and differential regulation in EMT intermediate states via RNA velocity.

mesenchymal marker expression in cervical cancer cells[52] and decreased cell migration in other carcinomas[53,54,75]. In vivo, *SFN* knockdown suppressed tumor formation and metastasis in lung adenocarcinoma models[76]. Clinically, *SFN* is linked to poor prognosis, including advanced tumor stages in lung adenocarcinoma and hepatocellular carcinoma[53,77], as well as lower survival rates in pancreatic ductal adenocarcinoma[78] and head and neck squamous cell carcinoma[79]. Our findings suggest that *SFN* promotes intermediate EMT dynamics, potentially explaining its dual role in epithelial cells while facilitating EMT.

Consensus gene analysis also identified *NRG1* as playing a pivotal role in intermediate EMT state dynamics, as the sole gene that was significant in intermediate expression, regulation, and modeled dynamics. A member of the epidermal growth factor (EGF) family[71], *NRG1* activates *ERBB2* (*HER2*) and *ERBB3* (*HER3*)[80]. *NRG1* isoforms can be found in the plasma membrane or secreted[81], and it binds integrins including *ITGA6*:*ITGB4* and *ITGAV*:*ITGB3*[82]. In vivo, *NRG1* suppression reduces tumor growth and metastasis in hepatocellular carcinoma[83]. Clinically, *NRG1* overexpression correlated with poor outcomes, including lymph node metastasis, in gastric cancer[84]. Notably, *NRG1* has been found to promote partial EMT in cultured patient *HER2*-positive breast cancer[85]. While *NRG1* has been mostly described to drive EMT in epithelial cells, *NRG1* stimulation on mesenchymal cells that already underwent EMT has been shown to instead induce epithelial gene expression in esophageal adenocarcinoma[86]. Taken together, our analyses along with literature suggest that *NRG1* is a marker of highly plastic intermediate state cells during EMT.

SFN is a marker of intermediate state EMT in independently analyzed MCF10A cells

To assess predicted intermediate EMT genes, we analyzed a dataset of EMT under different experimental conditions and in a different cell line: the dose-dependent TGF-β stimulation of MCF10A breast cells[87]. Similar to previous analyses, scRNA-seq data was clustered, and canonical markers were used to identify epithelial, intermediate, and mesenchymal states (Fig. 6a). Differential expression by cell state showed strong agreement with our predictions, with 11 of the top 25 intermediate state genes in this sample overlapping with our predicted intermediate EMT genes (Fig. 6b), notably including *SFN*. These results highlight that shared EMT intermediate state features can be found across diverse biological and experimental conditions, with independent evidence corroborating one of the top genes associated with intermediate EMT.

To assess EMT dynamics in these MCF10A cells, we applied the mathematical model using the same analytical pipeline (Fig. 6c). While dose-dependent EMT does not follow a true temporal progression, single-cell heterogeneity across TGF-β doses was evident. We used a pseudotime axis to represent a continuum of EMT states, capturing EMT transitions with different TGF-β doses. The posterior parameter distribution lies in the region where $k_1 \geq k_2$, consistent with EMT dynamics induced by TGF-β in other cell types (Fig. 6d). Across different cancer types, we see that mammary (MCF10A and HMLE) and ovarian (OVCA420) cells stimulated with TGF-β generally exhibit $k_1 > k_2$ dynamics, favoring stabilization of the intermediate state. In contrast, lung (A549) and prostate (DU145) cells stimulated with TGF-β show balanced rates of entry and exit from the intermediate state, with $k_1 \approx k_2$. The similarity in transition dynamics between mammary and ovarian cells is notable, given the shared genetic and microenvironmental factors during oncogenesis and tumor progression[88].

Discussion

Here, we characterized intermediate EMT states and identified genes involved in dynamic transitions between states. Multiple lines of evidence suggest EMT intermediate states are the most cancer stem-like and exhibit the highest metastatic potential[43,89–92]. Our analysis predicted intermediate state genes in agreement with recent work, such as *ITGB4* and *LAMB3*[93], as well as novel EMT intermediate genes, such as *SFN* and *NRG1*. While there are many paths of EMT, our comparison across different cell types and stimuli revealed common markers for intermediate states and highlighted the role many of these genes have in extracellular remodeling.

EMT is heterogeneous[45]. Multiple transcription factors can initiate EMT[14] and act in complex and nonlinear ways, both alone[94] or in combination[95]. Future work could shed more light on EMT intermediate state transitions by broadening the scope of EMT-inducing factors[96]. Additional factors contributing to EMT complexity, including subtypes and intermediate states, are hysteresis during the reverse mesenchymal-epithelial transition, differences in cell types or stimuli, and state transitions driven by intrinsic or extrinsic noise. Whereas EMT is most frequently

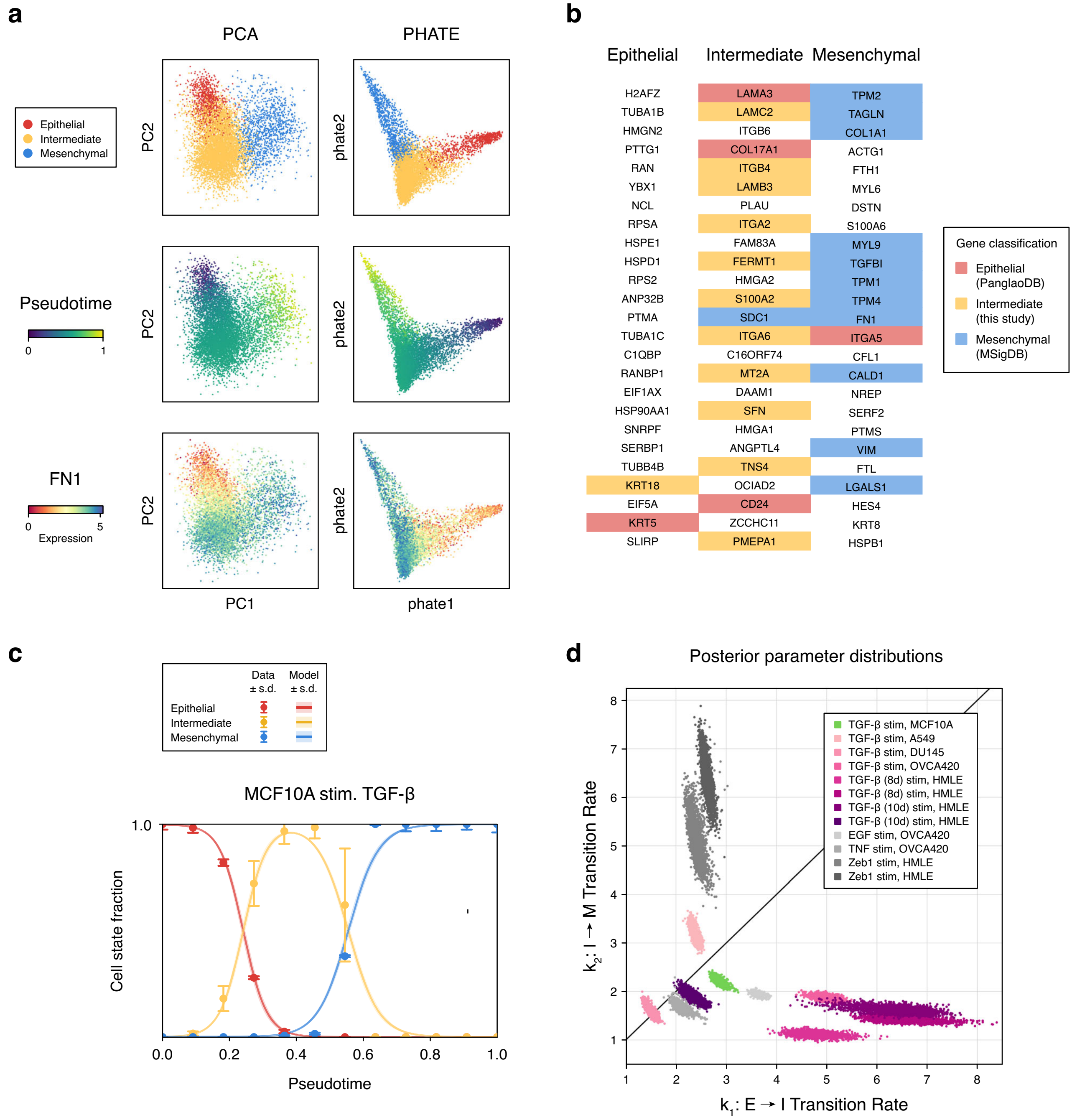


Fig. 6 | *SFN* is identified as a marker of intermediate EMT in an analysis of dose-dependent EMT. a MCF10A cells were analyzed separately and exhibit a linear trajectory across EMT states. **b** For each MCF10A cell state, the top 25 differentially expressed genes colored by gene set annotation. **c** Model fit following parameter inference: data vs. trajectory simulations, with simulation parameters sampled from the posterior distribution. **d** Comparison of posterior parameter distributions, with the MCF10A sample highlighted in green. Other distributions are replicated from Fig. 4d, shown in different colors here to highlight stimulation by TGF-*β*.

modeled via gene regulatory networks, here we modeled the population dynamics to study cell state heterogeneity and its effects on EMT path variation. In doing so, we assumed a monostable landscape, whereas in reality multiple stable steady states exist[9]. Some of the gene expression heterogeneity underlying these multiple states is likely collapsed by this approach, but in doing so we can identify consensus genes marking for properties of EMT states across different conditions. Our model can be adapted in the future to consider multiple intermediate states and more complex (e.g., convergent/divergent) EMT paths.

Summarizing complex data across conditions to find consensus requires simplifying assumptions. To compare gene expression across datasets, we used log-fold changes and rank-based comparisons, similar to other recent work[97]. Doing so relies on the accurate quantification of cell states, which is not guaranteed, and can obscure single-cell resolution information by taking pseudo-bulk measurements. While we sought to standardize data analysis pipelines as far as possible, scRNA-seq data analysis relies on certain parameter choices. While clustering cells, we sought fewer clusters (lower resolution) where supported, to reduce overfitting cell states. Clustering-based cell state definitions differ between studies: ref. 43 identified four EMT states, whereas ref. 98 later identified five in the same dataset. Similarly, ref. 99 identified five states in the ref. 41 dataset, aligning with experimental timepoints. Each approach reveals distinct aspects of

EMT; we take the perspective of using pseudotime to infer EMT dynamics over a 3–4 cell state space across the EMT spectrum. Trajectory inference relies on accurate choice of root cells and the sufficiency of the similarity metric used. RNA velocity analysis is limited by the ratio of spliced to unspliced counts, typically around 75–85% spliced to 15–25% unspliced[69]. This abundance limitation affects genes with low or no unspliced counts, such as *SFN* in our study, where RNA velocity analysis could not be performed due to a lack of unspliced counts. This abundance limitation could be addressed by experimental methods targeting dynamics, such as RNA metabolic labeling[36]. We applied a standardized pipeline to each dataset while preserving dataset-specific parameters. Although integration could reveal shared EMT features, differences in experimental design, single-cell platforms, and study conditions complicate the distinction between biological variation and technical batch effects. To maintain dataset-specific nuances, we prioritized a comparative analysis of differential expression and dynamic model parameters, though future studies incorporating batch correction may help determine whether intermediate EMT states share a universal transcriptional signature.

Mathematical modeling and parameter inference with single-cell data allow us to investigate the genes and pathways associated with dynamic transitions between states rather than the cell states themselves—transitions which are strongly relevant to epithelial–mesenchymal plasticity[100]. EMP, exemplary of cell state plasticity, has been shown to play decisive roles in tumorigenesis and cancer progression[101,102]. This property can assist tumors in developing powerful "generalist" phenotypes as they evolve[103]. The mathematical model with which we study EMT population dynamics is phenomenological: capturing the rates of entry/exit between EMT states without transcriptional information or feedback signaling. It does not incorporate additional complexities such as reverse transitions or stochasticity. We have used external information from the biological properties of EMT to construct our mathematical model, and not obtained it purely from the cell dynamics observed in the data[35]. Nonetheless, to compare relative transition rates, a simple three-compartment model seems reasonable to describe most conditions analyzed and fits both the inferred cell states (clusters) and the pseudotemporal dynamics during EMT. Incorporating cell proliferation and death could refine EMT modeling, but doing so presents challenges in parameter identifiability. Our approach prioritized a parsimonious model, using normalized cell proportions to implicitly account for differences in cell numbers, though this does not explicitly capture variations in cell survival. Future extensions integrating proliferation and death rates, potentially constrained by lineage tracing or live-cell imaging, could provide a more comprehensive understanding of EMT dynamics and intermediate state persistence. Additional future work could include combining cell population dynamics with a transcriptional EMT network[9] to investigate the role of cell–cell communication[104] on the population dynamics of EMT—though additional data may be required for the transcriptional dynamics of such a model to avoid double dipping[105].

Canonical EMT states are defined by morphological features: epithelial cells adhere to each other with apical–basal polarity; mesenchymal cells are spindle-shaped, migratory, and lack cell–cell adhesion[106]. These morphological/adhesive properties cannot be fully captured by sequencing data alone. Moreover, multiple EMT gene lists (typically focusing on mesenchymal traits) have been proposed, with varying levels of agreement[45,46,89,107,108]. This variability in consensus genes also applies to epithelial genes, which can show tissue-specific heterogeneity. No single gene list can do justice to the heterogeneous paths of EMT, yet as we have shown, distinctive dynamic properties of EMT intermediate states can be captured by marker genes. Although our study focused on cancer-related EMT, the in vitro stimuli such as TGF-β also apply to healthy EMT, suggesting potential relevance beyond cancer. Nevertheless, the heterogeneity observed among tumor cell lines underscores the need to investigate EMT in wound healing and tissue regeneration to determine whether the identified marker genes and intermediate states are conserved in the context of non-malignant EMT.

Genes predicted here as candidate markers of intermediate state EMT genes may serve as biomarkers of cells likely to metastasize and could be tested as predictors of clinical progression. In addition, such genes may mark for high-risk tumor cells prone to metastasis or recurrence, given the high metastatic potential of EMT intermediate state cells[16,21,92,109]. More broadly, this study has shed new light on the plasticity of the EMT landscape and how it shapes the cell state transitions underlying cancer metastasis.

Methods

scRNA-seq data sources

In this study, we conducted an integrated analysis of several single-cell RNA sequencing (scRNA-seq) datasets in the public domain. We included datasets from Pastushenko et al.[43] (GEO accession GSE110357); van Dijk et al.[110] (GSE114397); Cook and Vanderhyden[41] (GSE147405); and Panchy et al.[87] (GSE213753). For data from Cook and Vanderhyden[41], samples collected after the removal of the EMT stimulus were not included. For data from Panchy et al.[87], unstimulated cells were not included.

scRNA-seq sequence alignment

Data from ref. 41 were re-aligned to obtain spliced and unspliced read counts for RNA velocity analysis below. Raw sequence files (accession SRP253729) were downloaded from the NIH Sequence Read Archive using the SRA Toolkit[111] and converted from SRA to FASTQ files using `fasterq-dump`. Python package cutadapt was used to trim the barcode sequences to 26 base pairs[112]. The splici (spliced+intron) index was constructed using the GRCh38 human reference genome with Python package salmon[113]. Sequence pseudoalignment was performed with salmon alevin-fry. Barcode demultiplexing was carried out using the R package MULTIseq[114]. Contaminant cells in the OVCA420 samples were removed as noted by the original authors.

scRNA-seq data preprocessing and normalization

All scRNA-seq data were processed and analyzed using Scanpy[115]. Cells with fewer than 200 genes and genes expressed in fewer than three cells were filtered out. Cells with high mitochondrial percentages or disproportionately high total read counts were excluded based on dataset-specific cutoffs (Supp. Table 1). In HMLE samples stimulated with TGF-β, cells with disproportionately high ribosomal percentages were filtered out (<1% of cells). Counts were normalized to 10,000 and $log(x+1)$ transformed. Batch correction for samples from[41] was performed using ComBat in Scanpy[116]. Cell cycle effects, which significantly impacted clustering by EMT state identity, were regressed out[117,118], similar to the original analyses. Additional preprocessing included regressing out total counts and percent mitochondrial counts per cell, scaling counts to uniform variance, and selecting highly variable genes for downstream analysis.

Cell clustering and scoring by EMT status

Principal component analysis (PCA) was performed, and the top 15 components were used to construct a nearest-neighbor graph. Based on this graph, cell clustering was conducted using the Leiden algorithm[119] with dataset-specific Leiden resolutions (Supplementary Table 1). Silhouette scores were computed for Leiden resolutions ranging from 0.3 to 1.0 in 0.05 increments using `silhouette_score` from scikit-learn[120]. Differentially expressed genes for each cell cluster were identified using the Wilcoxon rank-sum test with Benjamini–Hochberg correction. Cell clusters were visualized in two dimensions using UMAP and PHATE[121,122].

To infer the EMT status of single cells based on a set of EMT marker genes, an "EMTscore" was created using the UCell scoring method[42] with the Hallmark EMT gene set from the Molecular Signatures Database (MSigDB)[39,123]. UCell calculates single-cell gene expression scores from a gene set using a rank-based approach, which we found to effectively quantify EMT across disparate tumor types and experimental conditions. Genes were input into UCell as filtered and normalized counts.

Identifying shared EMT intermediate state genes

Genes were included in the intermediate state analysis if they were differentially expressed (DE) in an intermediate state with a Benjamini–Hochberg adjusted Wilcoxon rank-sum P value of $P < 0.01$, up to a maximum of 500 genes per sample. To account for the complexities of comparing gene expression across different datasets and conditions (e.g., batch effects, instrumentation, sequencing depth), we calculated $\log_2$ fold change ($\log_2$FC) values of intermediate state genes in Scanpy, following the notation of Moses et al.[124]:

$$\begin{aligned}\log_2 \text{FC of gene } g = {} & \log_2\left(\exp\left(\tfrac{1}{n_1}\sum_{i\in G_1} Y_{ig}\right) - 1 + \epsilon\right)\\ & -\log_2\left(\exp\left(\tfrac{1}{n_2}\sum_{i\in G_2} Y_{ig}\right) - 1 + \epsilon\right)\end{aligned} \tag{1}$$

where G_1 is the focal group of cells of size n_1, G_2 is the comparison group with n_2 cells, and Y_{ig} denotes the log-normalized counts of gene g in cell i. The pseudocount $\epsilon = 10^{-9}$ is added to avoid division by zero[124]. Genes were selected as intermediate state-associated if they met the following criteria: (i) a $\log_2$FC ≥0.58 (1.5-fold change) in at least five samples, and (ii) at least two of these samples were from experiments not performed on HMLE cells. Gene set enrichment analysis (GSEA) was performed on identified intermediate state genes[123].

Trajectory inference and EMT subpopulation dynamics

Diffusion pseudotime (DPT) was used for trajectory analysis[125]. Root nodes were chosen as the epithelial cells with extreme coordinates on a diffusion map. Pseudotime was calculated five times with different epithelial root nodes, and the median values were assigned to each cell, with the standard deviation indicating pseudotime variation. This approach minimized the impact of root node selection on pseudotime calculation. Pseudotime values range from 0 (epithelial) to 1 (mesenchymal). This range was divided into 15 bins (12 for Pastushenko et al.[43] due to fewer cells), and cell counts were calculated for each cluster (epithelial, intermediate, and mesenchymal) for each bin. The counts per bin were converted into cell population proportions.

RNA velocity analysis

RNA velocity analysis was conducted in Python using the package scVelo in dynamical mode on highly variable genes[70]. Each sample was analyzed individually. Differential velocity (DV) was assessed using the `rank_dynamical_genes` function on clusters. Genes with a DV score above 0.25 were retained as DV genes. To ensure monotonic transitions, genes with Spearman correlation coefficients below 0.5 were excluded. In addition, DV genes with poor dynamical model fits were filtered out. Ultimately, we retained DV genes that were upregulated in the majority of cancer samples, designating them as shared upregulated velocity genes across EMT.

A mathematical model of EMT dynamics

We developed a mathematical model of the dynamics of EMT described by ordinary differential equations (ODEs). Specifically, we sought to describe the cell state transitions during EMT, from the epithelial (E) to intermediate (I) state or states, and then to the mesenchymal (M) state. While EMT systems may also exhibit direct transitions ($E \to M$) and reverse transitions, our data specifically investigate forward EMT and do not exhibit strong evidence for direct transitions.

The population dynamics of E, I, and M are described by:

$$\begin{aligned}\frac{dE}{dt} &= -k_1 EI\\ \frac{dI}{dt} &= k_1 EI - k_2 IM\\ \frac{dM}{dt} &= k_2 IM\end{aligned} \tag{2}$$

where k_1 denotes the transition rate from E to I, and k_2 denotes the transition rate from I to M. We consider second-order transitions, meaning both the initial and final states influence the transition rate to the final state. In cases where more than one intermediate state exists, the model can be extended using the same framework (Supplementary Fig. 9A).

Parameter inference of cell population dynamics over pseudotime

We sought to infer the rates of EMT using Bayesian parameter inference with the Turing.jl package in Julia[126–128]. The input data for each model consists of the cell state dynamics over pseudotime. To focus on relevant dynamics, we excluded periods where all cells remained in the epithelial state. Timepoints along pseudotime were normalized to a range of $t \in [0, 10]$, facilitating direct comparison of EMT trajectories across samples. For each sample with one intermediate state, we fit three parameters: k_1, k_2, and the observational noise parameter σ. For the in vivo sample with two intermediate states, we fit four parameters: k_1, k_2, k_3, and σ.

Letting f represent the numerical solution to the ODE model and y_0 the initial conditions, we performed parameter inference as follows:

$$\begin{aligned}\theta_{k_i} &\sim \mathcal{N}(4,1)\\ \sigma &\sim \text{Inv}-\text{Gamma}\,(3,1)\\ \hat{y}(t) &= f(y_0, t; \theta)\\ y(t) &\sim \mathcal{N}(\hat{y}(t), \sigma)\end{aligned} \tag{3}$$

where $\theta = (\theta_{k_i}, \sigma)$ gives the prior parameter distribution, and $y(t)$ defines the likelihood function in terms of ODE model simulation ($\hat{y}(t)$) for transition rate parameters θ_{k_i} and noise parameter σ.

The posterior parameter distribution was estimated via Markov chain Monte Carlo (MCMC) simulations using the No-U-Turn Sampler (NUTS)[129]. MCMC chains were each run for 1000 iterations following 250 warmup iterations to ensure convergence. Fitted trajectories were visualized by solving the model using 300 joint parameter sets of k_n, randomly selected from the posterior distribution for each sample, and plotting the mean and standard deviation of the resulting trajectories.

Comparative analysis of EMT intermediate state-associated genes

We identified genes associated with EMT transition rates by analyzing correlations between model-inferred posterior parameters and gene expression. For each transition rate parameter k_n, we used its maximum a posteriori value for each sample and examined pairwise correlations with the $\log_2$FC expression of 145 genes, each present in at least 5 samples with an intermediate state $\log_2\text{FC} \geq 0.2$. Genes with a Spearman's rank correlation coefficient of $\rho > |0.6|$ ($P < 0.05$) were considered associated with, and potential influencers of, transitions into or out of EMT states.

To specifically identify genes linked to EMT intermediate state dynamics, we focused on genes positively correlated with k_1 (faster $E \to I$) and negatively correlated with k_2 (slower $I \to M$). Genes meeting both correlation criteria were included, as well as those showing either correlation as well as differential expression or differential velocity in the intermediate state. Cellular location annotations were performed using DAVID[72,130] and PANTHER[131].

Data availability

All scRNA-seq data used in this study are publicly available on the Gene Expression Omnibus (GEO). We included datasets from Pastushenko et al.[43] (accession GSE110357); van Dijk et al.[110] (GSE114397); Cook and Vanderhyden[41] (GSE147405); and Panchy et al.[87] (GSE213753). Raw sequence files from Cook and Vanderhyden[41] were downloaded from the NIH Sequence Read Archive (accession SRP253729). Supplementary Tables and Figures are provided as a separate document. Supplementary datasets provided are: gene expression plots of all 32 identified EMT intermediate state genes (Supplementary Data 1, available at: https://github.com/maclean-lab/dynamicEMT-genes under 3–DE genes) and GSEA results (Supplementary Data 2).

Code availability

All code and data analysis associated with this study are released under an MIT license, available on GitHub: https://github.com/maclean-lab/dynamicEMT-genes.

Received: 12 November 2024; Accepted: 28 March 2025;
Published online: 10 April 2025

References

1. Rood, J. E., Maartens, A., Hupalowska, A., Teichmann, S. A. & Regev, A. Impact of the Human Cell Atlas on medicine. *Nat. Med.* **28**, 2486–2496 (2022).
2. Qiu, C. et al. A single-cell time-lapse of mouse prenatal development from gastrula to birth. *Nature* **626**, 1084–1093 (2024).
3. MacLean, A. L., Hong, T. & Nie, Q. Exploring intermediate cell states through the lens of single cells. *Curr. Opin. Syst. Biol.* **9**, 32–41 (2018).
4. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
5. Thiery, J. P. Epithelial–mesenchymal transitions in development and pathologies. *Curr. Opin. Cell Biol.* **15**, 740–746 (2003).
6. Dillekås, H., Rogers, M. S. & Straume, O. Are 90% of deaths from cancer caused by metastases? *Cancer Med.* **8**, 5574–5576 (2019).
7. 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).
8. Xing, J. & Tian, X.-J. Investigating epithelial-to-mesenchymal transition with integrated computational and experimental approaches. *Phys. Biol.* **16**, 031001 (2019).
9. 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).
10. Sha, Y. et al. Intermediate cell states in epithelial-to-mesenchymal transition. *Phys. Biol.* **16**, 021001 (2019).
11. Hong, T. & Xing, J. Data- and theory-driven approaches for understanding paths of epithelial–mesenchymal transition. *Genesis* **62**, e23591 (2024).
12. Deshmukh, A. P. et al. Identification of EMT signaling cross-talk and gene regulatory networks by single-cell RNA sequencing. *Proc. Natl. Acad. Sci. USA* **118**, e2102050118 (2021).
13. Tagliazucchi, G. M., Wiecek, A. J., Withnell, E. & Secrier, M. Genomic and microenvironmental heterogeneity shaping epithelial-to-mesenchymal trajectories in cancer. *Nat. Commun.* **14**, 789 (2023).
14. Nieto, M. A., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* **166**, 21–45 (2016).
15. Shaw, T. J. & Martin, P. Wound repair: a showcase for cell plasticity and migration. *Curr. Opin. Cell Biol.* **42**, 29–37 (2016).
16. Yu, M. et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science* **339**, 580–584 (2013).
17. Ting, D. T. et al. Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* **8**, 1905–1918 (2014).
18. Ruscetti, M., Quach, B., Dadashian, E. L., Mulholland, D. J. & Wu, H. Tracking and functional characterization of epithelial-mesenchymal transition and mesenchymal tumor cells during prostate cancer metastasis. *Cancer Res.* **75**, 2749–2759 (2015).
19. Hendrix, M. J., Seftor, E. A., Seftor, R. E. & Trevor, K. T. Experimental co-expression of vimentin and keratin intermediate filaments in human breast cancer cells results in phenotypic interconversion and increased invasive behavior. *Am. J. Pathol.* **150**, 483–495 (1997).
20. Jolly, M. K. et al. Implications of the hybrid epithelial/mesenchymal phenotype in metastasis. *Front. Oncol.* **5**, 155 (2015).
21. Simeonov, K. P. et al. Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. *Cancer Cell* **39**, 1150–1162.e9 (2021).
22. Jost, J. *Dynamical Systems: Examples of Complex Behaviour*. (Springer, 2005).
23. Leggett, S. E., Hruska, A. M., Guo, M. & Wong, I. Y. The epithelial-mesenchymal transition and the cytoskeleton in bioengineered systems. *Cell Commun. Signal.* **19**, 32 (2021).
24. Chaffer, C. L., San Juan, B. P., Lim, E. & Weinberg, R. A. EMT, cell plasticity and metastasis. *Cancer Metastasis Rev.* **35**, 645–654 (2016).
25. Medici, D., Hay, E. D. & Olsen, B. R. Snail and slug promote epithelial-mesenchymal transition through β-catenin–T-cell factor-4-dependent expression of transforming growth factor-β3. *Mol. Biol. Cell* **19**, 4875–4887 (2008).
26. 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).
27. Zhang, J. et al. TGF-β-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Sci. Signal.* **7**, ra91 (2014).
28. 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* **110**, 18144–18149 (2013).
29. Tripathi, S., Xing, J., Levine, H. & Jolly, M. K. Mathematical modeling of plasticity and heterogeneity in EMT. In *The Epithelial-to Mesenchymal Transition: Methods and Protocols*, Methods in Molecular Biology (eds Campbell, K. & Theveneau, E.) 385–413 (Springer Protocols, 2021).
30. Wu, X., Wollman, R. & MacLean, A. L. Single-cell Ca^{2+} parameter inference reveals how transcriptional states inform dynamic cell responses. *J. R. Soc. Interface* **20**, 20230172 (2023).
31. Cho, H., Kuo, Y.-H. & Rockne, R. C. Comparison of cell state models derived from single-cell RNA sequencing data: graph versus multi-dimensional space. *Math. Biosci. Eng.* **19**, 8505–8536 (2022).
32. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
33. MacLean, A. L. et al. Single cell phenotyping reveals heterogeneity among hematopoietic stem cells following infection. *Stem Cells* **35**, 2292–2304 (2017).
34. Sha, Y., Wang, S., Zhou, P. & Nie, Q. Inference and multiscale model of epithelial-to-mesenchymal transition via single-cell transcriptomic data. *Nucleic Acids Res.* **48**, 9505–9520 (2020).
35. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci. USA* **115**, E2467–E2476 (2018).
36. Qiu, X. et al. Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690–711.e45 (2022).
37. Fischer, D. S. et al. Inferring population dynamics from single-cell RNA-sequencing time series data. *Nat. Biotechnol.* **37**, 461–468 (2019).
38. Stumpf, P. S. et al. Stem cell differentiation as a non-Markov stochastic process. *Cell Syst.* **5**, 268–282.e7 (2017).
39. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
40. O.Franzén, Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database: J. Biol. Databases Cur.* **2019**, baz046 (2019).
41. Cook, D. P. & Vanderhyden, B. C. Context specificity of the EMT transcriptional response. *Nat. Commun.* **11**, 2142 (2020).
42. Andreatta, M. & Carmona, S. J. UCell: Robust and scalable single-cell gene signature scoring. *Comput. Struct. Biotechnol. J.* **19**, 3796–3798 (2021).
43. Pastushenko, I. et al. Identification of the tumour transition states occurring during EMT. *Nature* **556**, 463–468 (2018).
44. Sha, Y., Wang, S., Bocci, F., Zhou, P. & Nie, Q. Inference of intercellular communications and multilayer gene-regulations of epithelial–mesenchymal transition from single-cell transcriptomic data. *Front. Genet.* **11**, 604585 (2021).

45. Yang, J. et al. Guidelines and definitions for research on epithelial–mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **21**, 341–352 (2020).
46. Cook, D. P. & Vanderhyden, B. C. Transcriptional census of epithelial-mesenchymal plasticity in cancer. *Sci. Adv.* **8**, eabi7640 (2022).
47. Biffo, S. et al. Isolation of a novel β4 integrin-binding protein (p27BBP) highly expressed in epithelial cells. *J. Biol. Chem.* **272**, 30314–30321 (1997).
48. Li, X.-L. et al. Integrin β4 promotes cell invasion and epithelial-mesenchymal transition through the modulation of Slug expression in hepatocellular carcinoma. *Sci. Rep.* **7**, 40464 (2017).
49. Masugi, Y. et al. Upregulation of integrin β4 promotes epithelial–mesenchymal transition and is a novel prognostic marker in pancreatic ductal adenocarcinoma. *Lab. Investig.* **95**, 308–319 (2015).
50. Zheng, G. et al. Integrin alpha 6 is upregulated and drives hepatocellular carcinoma progression through integrin α6β4 complex. *Int. J. Cancer* **151**, 930–943 (2022).
51. Leffers, H. et al. Molecular cloning and expression of the transformation sensitive epithelial marker stratifin. A member of a protein family that has been involved in the protein kinase C signalling pathway. *J. Mol. Biol.* **231**, 982–998 (1993).
52. Hu, Y. et al. LINC01128 expedites cervical cancer progression by regulating miR-383-5p/SFN axis. *BMC Cancer* **19**, 1157 (2019).
53. Ye, S.-P. et al. Stratifin promotes hepatocellular carcinoma progression by modulating the Wnt/β-catenin pathway. *Int. J. Genomics* **2023**, 9731675 (2023).
54. Zhao, X., Wang, E., Xu, H. & Zhang, L. Stratifin promotes the growth and proliferation of hepatocellular carcinoma. *Tissue Cell* **82**, 102080 (2023).
55. Ni, M., Zhao, Y. & Wang, X. Suppression of synuclein gamma inhibits the movability of endometrial carcinoma cells by PI3K/AKT/ERK signaling pathway. *Genes Genomics* **43**, 633–641 (2021).
56. Takemura, Y. et al. Gamma-synuclein is a novel prognostic marker that promotes tumor cell migration in biliary tract carcinoma. *Cancer Med.* **10**, 5599–5613 (2021).
57. Liu, J. et al. Gamma synuclein promotes cancer metastasis through the MKK3/6-p38MAPK cascade. *Int. J. Biol. Sci.* **18**, 3167–3177 (2022).
58. Zhuang, Q., Liu, C., Qu, L. & Shou, C. Synuclein-γ promotes migration of MCF7 breast cancer cells by activating extracellular-signal regulated kinase pathway and breaking cell-cell junctions. *Mol. Med. Rep.* **12**, 3795–3800 (2015).
59. Basu, S., Cheriyamundath, S. & Ben-Ze'ev, A. Cell–cell adhesion: linking Wnt/β-catenin signaling with partial EMT and stemness traits in tumorigenesis. *F1000Research* **7**, 1488 (2018).
60. Sun, Q. et al. Interleukin-6 promotes epithelial-mesenchymal transition and cell invasion through integrin β6 upregulation in colorectal cancer. *Oxid. Med. Cell. Longev.* **2020**, 8032187 (2020).
61. Chen, J. et al. E2F1/SP3/STAT6 axis is required for IL-4-induced epithelial-mesenchymal transition of colorectal cancer cells. *Int. J. Oncol.* **53**, 567–578 (2018).
62. Cao, H. et al. IL-13/STAT6 signaling plays a critical role in the epithelial-mesenchymal transition of colorectal cancer cells. *Oncotarget* **7**, 61183–61198 (2016).
63. Berzaghi, R. et al. SOCS1 favors the epithelial-mesenchymal transition in melanoma, promotes tumor progression and prevents antitumor immunity by PD-L1 expression. *Sci. Rep.* **7**, 40585 (2017).
64. Thorpe, H., Asiri, A., Akhlaq, M. & Ilyas, M. Cten promotes epithelial-mesenchymal transition through the post-transcriptional stabilization of Snail. *Mol. Carcinogenesis* **56**, 2601–2609 (2017).
65. Katz, M. et al. A reciprocal tensin-3–cten switch mediates EGF-driven mammary cell migration. *Nat. Cell Biol.* **9**, 961–969 (2007).
66. Shen, M. et al. Tinagl1 suppresses triple-negative breast cancer progression and metastasis by simultaneously inhibiting integrin/FAK and EGFR signaling. *Cancer Cell* **35**, 64–80.e7 (2019).
67. Shan, Z.-G. et al. Upregulation of tubulointerstitial nephritis antigen like 1 promotes gastric cancer growth and metastasis by regulating multiple matrix metallopeptidase expression. *J. Gastroenterol. Hepatol.* **36**, 196–203 (2021).
68. Hong, S.-Y., Shih, Y.-P., Li, T., Carraway, K. L. & Lo, S. H. CTEN prolongs signaling by EGFR through reducing its ligand-induced degradation. *Cancer Res.* **73**, 5266–5276 (2013).
69. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
70. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
71. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
72. Sherman, B. T. et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, gkac194 (2022).
73. Hou, W., Pan, M., Xiao, Y. & Ge, W. Serum extracellular vesicle stratifin is a biomarker of perineural invasion in patients with colorectal cancer and predicts worse prognosis. *Front. Oncol.* **12**, 912584 (2022).
74. Ghaffari, A. et al. Fibroblast extracellular matrix gene expression in response to keratinocyte-releasable stratifin. *J. Cell. Biochem.* **98**, 383–393 (2006).
75. Kim, J. Y. et al. Stratifin (SFN) regulates lung cancer progression via nucleating the Vps34-BECN1-TRAF6 complex for autophagy induction. *Clin. Transl. Med.* **12**, e896 (2022).
76. Shiba-Ishii, A. et al. Stratifin accelerates progression of lung adenocarcinoma at an early stage. *Mol. Cancer* **14**, 142 (2015).
77. Kim, Y. et al. Stratifin regulates stabilization of receptor tyrosine kinases via interaction with ubiquitin-specific protease 8 in lung adenocarcinoma. *Oncogene* **37**, 5387–5402 (2018).
78. Robin, F. et al. Molecular profiling of stroma highlights stratifin as a novel biomarker of poor prognosis in pancreatic ductal adenocarcinoma. *Br. J. Cancer* **123**, 72–80 (2020).
79. Chung, C. H. et al. Gene expression profiles identify epithelial-to-mesenchymal transition and activation of nuclear factor-κB signaling as characteristics of a high-risk head and neck squamous cell carcinoma. *Cancer Res.* **66**, 8210–8218 (2006).
80. Miano, C. et al. NRG1/ERBB3/ERBB2 axis triggers anchorage-independent growth of basal-like/triple-negative breast cancer cells. *Cancers* **14**, 1603 (2022).
81. Esper, R. M., Pankonin, M. S. & Loeb, J. A. Neuregulins: versatile growth and differentiation factors in nervous system development and human disease. *Brain Res. Rev.* **51**, 161–175 (2006).
82. Ieguchi, K. et al. Direct binding of the EGF-like domain of neuregulin-1 to integrins (αvβ3 and α6β4) is involved in neuregulin-1/ErbB signaling. *J. Biol. Chem.* **285**, 31388–31398 (2010).
83. Shi, D.-M. et al. miR-296-5p suppresses EMT of hepatocellular carcinoma via attenuating NRG1/ERBB2/ERBB3 signaling. *J. Exp. Clin. Cancer Res.* **37**, 294 (2018).
84. Yun, S. et al. Clinical significance of overexpression of NRG1 and its receptors, HER3 and HER4, in gastric cancer patients. *Gastric Cancer* **21**, 225–236 (2018).
85. Guardia, C. et al. Preclinical and clinical characterization of fibroblast-derived neuregulin-1 on trastuzumab and pertuzumab activity in HER2-positive breast cancer. *Clin. Cancer Res.* **27**, 5096–5108 (2021).
86. Ebbing, E. A. et al. Esophageal adenocarcinoma cells and xenograft tumors exposed to Erb-b2 receptor tyrosine kinase 2 and 3 inhibitors activate transforming growth factor beta signaling, which induces

epithelial to mesenchymal transition. *Gastroenterology* **153**, 63–76 (2017).

87. Panchy, N., Watanabe, K., Takahashi, M., Willems, A. & Hong, T. Comparative single-cell transcriptomes of dose and time dependent epithelial–mesenchymal spectrums. *NAR Genomics Bioinforma.* **4**, lqac072 (2022).
88. Roskelley, C. D. & Bissell, M. J. The dominance of the microenvironment in breast and ovarian cancer. *Semin. Cancer Biol.* **12**, 97–104 (2002).
89. Puram, S. V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624.e24 (2017).
90. Aiello, N. M. et al. EMT subtype influences epithelial plasticity and mode of cell migration. *Dev. Cell* **45**, 681–695.e4 (2018).
91. Zhang, Y. et al. Genome-wide CRISPR screen identifies PRC2 and KMT2D-COMPASS as regulators of distinct EMT trajectories that contribute differentially to metastasis. *Nat. Cell Biol.* **24**, 554–564 (2022).
92. Brown, M. S. et al. Phenotypic heterogeneity driven by plasticity of the intermediate EMT state governs disease progression and metastasis in breast cancer. *Sci. Adv.* **8**, eabj8002 (2022).
93. Cheng, Y.-C. et al. Reconstruction of single-cell lineage trajectories and identification of diversity in fates during the epithelial-to-mesenchymal transition. *Proc. Natl. Acad. Sci. USA* **121**, e2406842121 (2024).
94. Hartmann, L. et al. Transcriptional regulators ensuring specific gene expression and decision making at high TGFβ doses. *Life Sci. Alliance* **8**, e202402859 (2024).
95. Sanford, E. M., Emert, B. L., Coté, A. & Raj, A. Gene regulation gravitates toward either addition or multiplication when combining the effects of two signals. *eLife* **9**, e59388 (2020).
96. Barbeau, M. C., Brown, B. A., Adair, S. J., Bauer, T. W. & Lazzara, M. J. ERK plays a conserved dominant role in pancreas cancer cell EMT heterogeneity driven by diverse growth factors and chemotherapies. Preprint at https://www.biorxiv.org/content/biorxiv/early/2025/02/09/2025.02.08.637251.full.pdf (2025).
97. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
98. Zhou, P., Wang, S., Li, T. & Nie, Q. Dissecting transition cells from single-cell transcriptome data through multiscale stochastic dynamics. *Nat. Commun.* **12**, 5609 (2021).
99. Pan, S., Withnell, E. & Secrier, M. Classifying epithelial-mesenchymal transition states in single cell cancer data using large language models. Preprint at https://www.biorxiv.org/content/10.1101/2024.08.16.608311v1 (2024).
100. Williams, E. D., Gao, D., Redfern, A. & Thompson, E. W. Controversies around epithelial–mesenchymal plasticity in cancer metastasis. *Nat. Rev. Cancer* **19**, 716–732 (2019).
101. Househam, J. et al. Phenotypic plasticity and genetic control in colorectal cancer evolution. *Nature* **611**, 744–753 (2022).
102. Street, K., Siegmund, K. & Shibata, D. Epigenetic conservation infers that colorectal cancer progenitors retain the phenotypic plasticity of normal colon. Preprint at https://www.researchsquare.com/article/rs-2609517 (2023).
103. Cook, D. P. & Wrana, J. L. A specialist-generalist framework for epithelial-mesenchymal plasticity in cancer. *Trends Cancer* **8**, 358–368 (2022).
104. Rommelfanger, M. K. & MacLean, A. L. A single-cell resolved cell-cell communication model explains lineage commitment in hematopoiesis. *Development* **148**, dev199779 (2021).
105. Neufeld, A., Gao, L. L., Popp, J., Battle, A. & Witten, D. Inference after latent variable estimation for single-cell RNA sequencing data. *Biostatistics* **25**, 270–287 (2024).
106. Thiery, J. P., Acloque, H., Huang, R. Y. J. & Nieto, M. A. Epithelial-mesenchymal transitions in development and disease. *Cell* **139**, 871–890 (2009).
107. 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).
108. Kinker, G. S. et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).
109. Lüönd, F. et al. Distinct contributions of partial and full EMT to breast cancer malignancy. *Dev. Cell* **56**, 3203–3221.e11 (2021).
110. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e27 (2018).
111. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
112. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
113. Srivastava, A., Malik, L., Smith, T., Sudbery, I. & Patro, R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol.* **20**, 65 (2019).
114. McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
115. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
116. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
117. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
118. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
119. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
120. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
121. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2020).
122. Moon, K. R. et al. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**, 1482–1492 (2019).
123. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
124. Moses, L. et al. Voyager: exploratory single-cell genomics data analysis with geospatial statistics. Preprint at https://www.biorxiv.org/content/10.1101/2023.07.20.549945v2.full.pdf (2023).
125. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
126. Ge, H., Xu, K. & Ghahramani, Z. Turing: a language for flexible probabilistic inference. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 1682–1690 (PMLR, 2018).
127. Rackauckas, C. & Nie, Q. DifferentialEquations.jl—a performant and feature-rich ecosystem for solving differential equations in Julia. *J. Open Res. Softw.* **5**, 15 (2017).
128. Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. Julia: a fresh approach to numerical computing. *SIAM Rev.* **59**, 65–98 (2017).
129. Hoffman, M. D. & Gelman, A. The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **31**, 1593–1623 (2014).
130. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
131. Thomas, P. D. et al. PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci.* **31**, 8–22 (2022).

Acknowledgements

This work was supported by the National Institutes of Health (R35GM143019) and the National Science Foundation (DMS2045327) (to A.L.M.).

Author contributions

M.M.: conceptualization, software, methodology, investigation, formal analysis, writing—original draft, writing—reviewing and editing. R.M.: investigation, writing—reviewing and editing. E.T.R.T.: investigation, supervision, writing—reviewing & editing. A.L.M.: conceptualization, software, methodology, investigation, funding acquisition, supervision, writing—original draft, writing—reviewing and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-025-00512-2.

Correspondence and requests for materials should be addressed to Adam L. MacLean.

Reprints and permissions information is available at http://www.nature.com/reprints