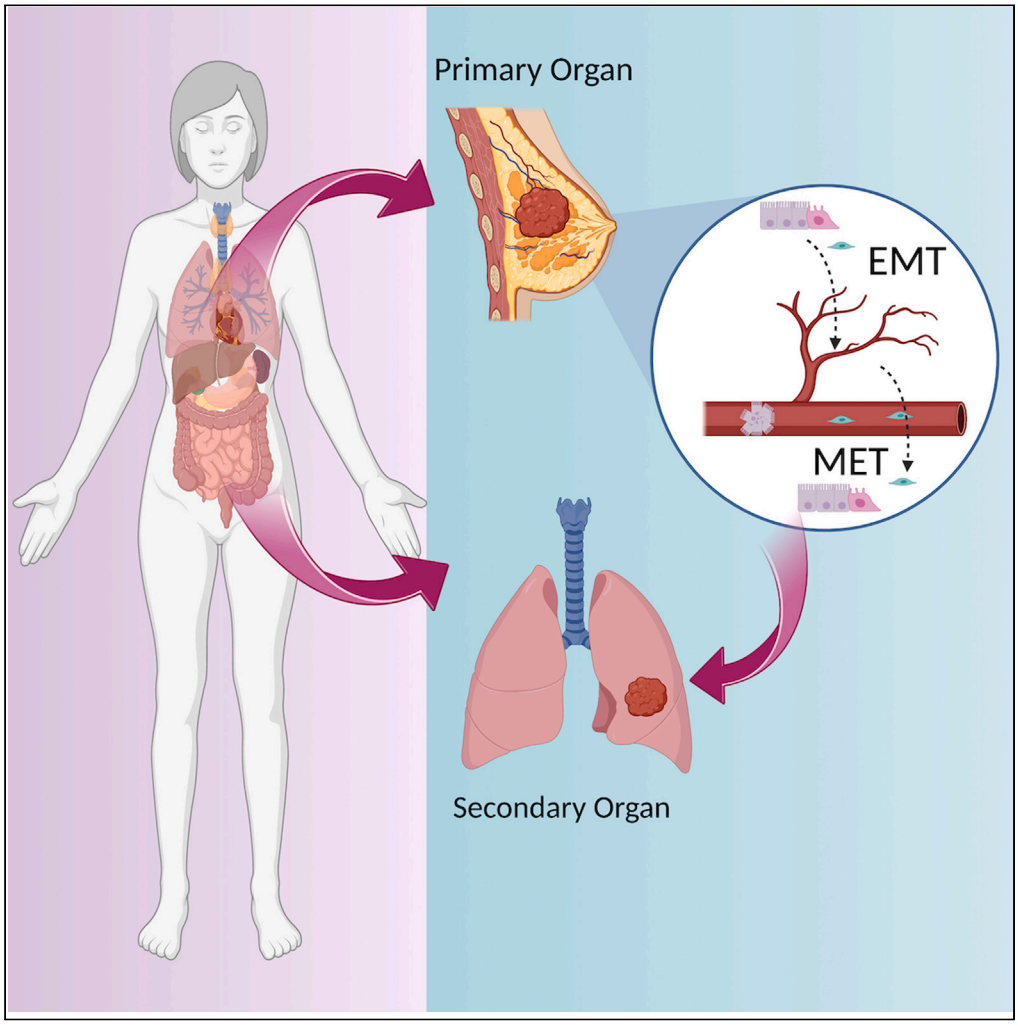
iScience

CellPress
OPEN ACCESS

Article

Population dynamics of EMT elucidates the timing and distribution of phenotypic intra-tumoral heterogeneity



Annice Najafi,
Mohit K. Jolly,
Jason T. George

mkjolly@iisc.ac.in (M.K.J.)
jason.george@tamu.edu (J.T.G.)

Highlights

A fully stochastic model elucidates the population dynamics of EMT

A data-driven pipeline infers EMT trajectories from single-cell RNA seq

Cell cycle scoring and GSEA reveal cell line-dependent patterns of EMT induction

Najafi et al., iScience 26, 106964
July 21, 2023
https://doi.org/10.1016/j.isci.2023.106964

iScience

CellPress
OPEN ACCESS

Article

Population dynamics of EMT elucidates the timing and distribution of phenotypic intra-tumoral heterogeneity

Annice Najafi,[1] Mohit K. Jolly,[2,*] and Jason T. George[1,3,4,*]

SUMMARY

The Epithelial-to-Mesenchymal Transition (EMT) is a hallmark of cancer metastasis and morbidity. EMT is a non-binary process, and cells can be stably arrested en route to EMT in an intermediate hybrid state associated with enhanced tumor aggressiveness and worse patient outcomes. Understanding EMT progression in detail will provide fundamental insights into the mechanisms underlying metastasis. Despite increasingly available single-cell RNA sequencing (scRNA-seq) data that enable in-depth analyses of EMT at the single-cell resolution, current inferential approaches are limited to bulk microarray data. There is thus a great need for computational frameworks to systematically infer and predict the timing and distribution of EMT-related states at single-cell resolution. Here, we develop a computational framework for reliable inference and prediction of EMT-related trajectories from scRNA-seq data. Our model can be utilized across a variety of applications to predict the timing and distribution of EMT from single-cell sequencing data.

INTRODUCTION

The Epithelial-Mesenchymal Transition (EMT) is a reversible dynamic cellular process involving the transformation of membrane-bound epithelial cells to motile mesenchymal cells. Although EMT was initially discovered in the setting of embryogenesis,[1] its importance in wound healing and the migration of cells to distant organs during metastasis is now appreciated and these topics have become active areas of research.[2,3]

EMT is driven by biomechanical and biochemical signals that lead to the loss of apical-basal polarity and cell-cell junctions. These structural disturbances in organization arise because of the down-regulation of epithelial biomarkers and up-regulation of mesenchymal biomarkers.[4,5] Multiple EMT-associated transcription factors orchestrate EMT in an elaborate gene regulatory circuit.[4,6,7] As a result, EMT is not only affected by genetic alterations but is also driven by epigenetic and post-transcriptional modifications.[8,9] Thus, this process and its phenotypic implications are imperfectly understood by studying whole genome sequencing or allelotyping alone, particularly evident in studies attempting to identify mutant genes underlying metastasis[10]

Studying EMT and its role in cancer progression is further complicated by the presence of non-cancer cells in the Tumor Microenvironment (TME), such as cancer-associated fibroblasts[11,12] and tumor-associated macrophages,[13] which interact with and are in close proximity to tumor cells, leading to difficulties in distinguishing tumor cells en route to EMT from the TME.[14] Consequently, there is an ongoing debate on the exact contributions of EMT in cancer metastasis[15]: Mesenchymal cells are known to escape from the adaptive immune system, actively present in the tumor stroma.[16,17] On the other hand, epithelial cells can also break away from the tumor and travel collectively as circulating tumor cells (CTC) such that their epithelial marker expression patterns remain intact. Although some studies report cells going through partial EMT, others report that the leaders at the edge of the clusters display promiscuous gene expression patterns whereas the inner layers remain more uniform.[18,19]

EMT conceptualization has benefited significantly from the theoretical prediction of a hybrid intermediate state whose properties fall on a spectrum between epithelial and mesenchymal phenotypes.[20] This

[1]Department of Biomedical Engineering, Texas A&M University, College Station, TX 77843, USA

[2]Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore 560012, India

[3]Intercollegiate School of Engineering Medicine, Texas A&M University, Houston, TX 77030, USA

[4]Lead contact

*Correspondence: mkjolly@iisc.ac.in (M.K.J.), jason.george@tamu.edu (J.T.G.)

https://doi.org/10.1016/j.isci.2023.106964

phenotypic spectrum permits fine-tuned adaptations to environmental cues through cellular plasticity that can manifest as enhanced tumor aggressiveness in the clinic. Consequently, recent investigations have focused on the partial EMT state and its subsequent transition as the main culprit behind metastasis.[21,22] Moreover, it is now appreciated that this hybrid state is non-transient and reinforced by phenotypic stability factors (PSF), such as GRHL2 and NFATc, which enhance tumor-initiation.[23,24]

TGFβ is a critical EMT inducer, both *in-vitro* and *in-vivo*, that acts through transcriptional regulation resulting in the down-regulation of E-cadherin (E-cad) and further up-regulation of TGFβ in a positive feedback loop.[7] TGFβ EMT induction *in-vitro* is known to be reversible whereby TGFβ withdrawal results in a reverse MET. Notably, not all cells are responsive to the EMT-inducing signal resulting in considerable phenotypic intra-tumoral heterogeneity and coexistence amongst multiple states[25] (Figure 1B). This has been reinforced by experimental observations supporting the existence of phenotypic transitions amongst EMT states and model-driven predictions of phenotypic heterogeneity generated by noisy cell division.[26,27] Furthermore, the environmental cues that govern these cell state transitions are highly context-specific.[28,29] This in turn highlights the importance of computational frameworks that interrogate context-specific EMT in the absence of cell-division.

Prior computational frameworks have successfully inferred epithelial, hybrid, and mesenchymal states from transcriptomic data.[30–33] However, these methods were all optimized to explain microarray data, and are often ill-equipped to perform well on more modern approaches. Thus, there is a significant need to identify and predict context-specific EMT-related trajectories at transient and steady states from next-generation sequencing data.

Here, we provide a dual theoretical and data-driven framework, COMET (Cell line-specific Optimization Method of EMT Trajectories) for understanding the stochastic progression of EMT at stationary populations (Figure 1C). We track the dynamics of these systems, which are treated with an EMT-induction factor, and show that COMET can successfully predict the timing and distribution of EMT states. Next, we accurately infer the three epithelial, hybrid, and mesenchymal trajectories from single-cell RNA sequencing (scRNA-seq) data using COMET. We show that COMET explains early and late transition dynamics of context-specific EMT and relates the timing and equilibrium distribution to systemic noise through EMT induction exposure time. In addition, we show that COMET reveals tumor subtype plays a more significant role than induction factors with respect to the dynamics of time-course EMT data which can give us insights into the patterns and extent of context-specific metastasis.

Model development

EMT population dynamics modeled with a three-state continuous-time Markov chain

We consider a non-dividing population of N cells, each of which belonging to one of three states; Epithelial (E), Hybrid (H), and Mesenchymal (M). We denote by $\pi_k(t)$ the relative abundance of population k ($k \in \{E,H,M\}$). For each time-point t, we have:

$$\pi_E(t) + \pi_H(t) + \pi_M(t) = 1. \qquad \text{(Equation 1)}$$

We model this process as a Continuous-Time Markov Chain (CTMC).[34] Letting $P(t)$ denote the probability transition matrix of this Markov chain, this process evolves temporally via Equation 2:

$$P(t) = e^{tG} = \sum_{j=0}^{\infty} \frac{(tG)^n}{n!}, \qquad \text{(Equation 2)}$$

where G is the infinitesimal generator matrix representing the rates of transitions between the states of the system, given by Equation 3:

$$G = \begin{pmatrix} -\mu_E & \mu_E & 0 \\ \lambda_E & -(\lambda_E + \lambda_M) & \lambda_M \\ 0 & \mu_M & -\mu_M \end{pmatrix}. \qquad \text{(Equation 3)}$$

The rates μ_k (resp. λ_k) represent the transition rates out of (resp. into) state k ($k \in \{E,M\}$). It can be shown that this system permits a unique equilibrium distribution, which is represented by $\pi_\infty = [\pi_{E,\infty}, \pi_{H,\infty}, \pi_{M,\infty}]$ and satisfies

$$\pi_\infty P(t) = \pi_\infty ; \pi_\infty G = 0. \qquad \text{(Equation 4)}$$

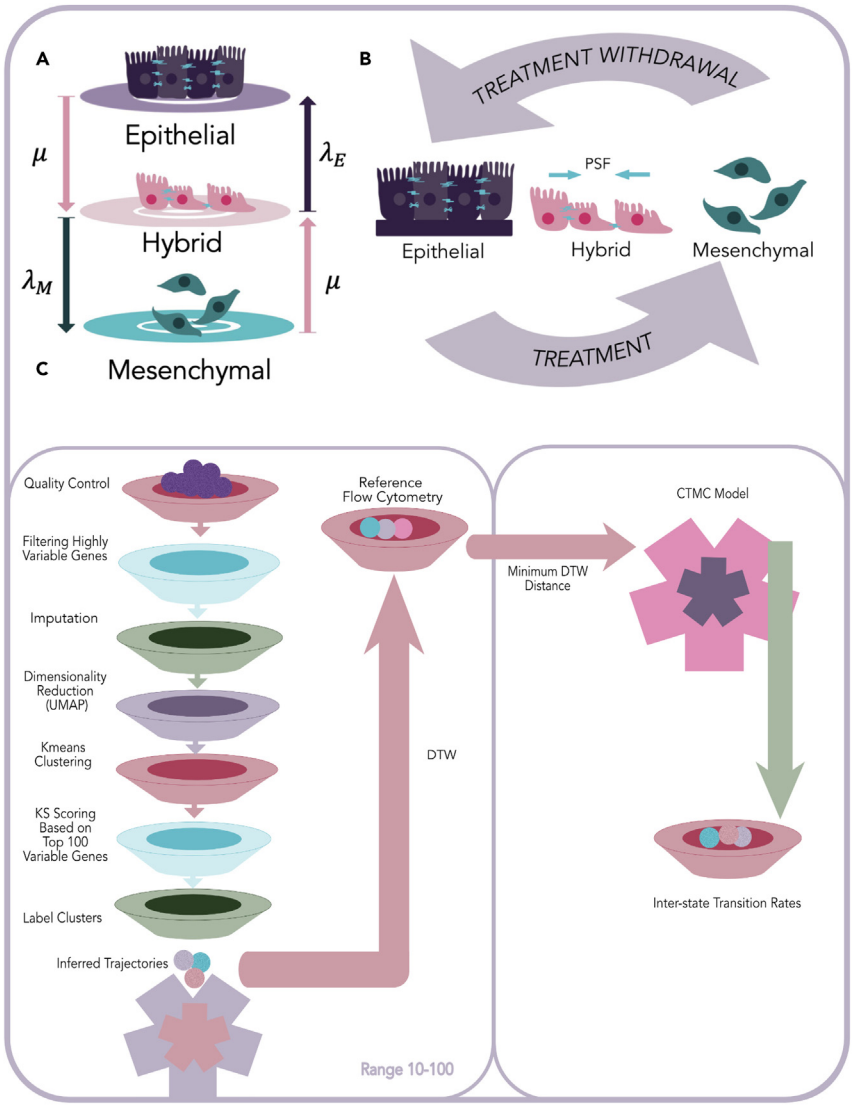


Figure 1. Stochastic Phenotypic Transition Model

(A) Illustration of the continuous-time three-state model and the four allowable transitions with their corresponding rates. Phenotypic stability factors enhance the stability of the intermediate state, and their corresponding transition rates (μ) are assumed to be symmetric.

(B) Illustration of cell populations undergoing division-independent induced EMT, characterized by epithelial cells that lose their cell-cell junctions and apical-basal polarity. On treatment withdrawal cells regain their epithelial characteristics through MET. PSFs stabilize the hybrid state by making the transition rates into the hybrid state symmetric.

(C) Illustrates the flow of data through the dual data-driven and theoretic model of COMET. The data-driven pipeline on the left is run for a specific number of top variable genes (from 10 to 100 in increments of 5). First, the raw read count matrix is quality controlled and filtered for the top 10 variable genes. The matrix is MAGIC imputed and reduced in dimensionality via UMAP. This is followed by K-means clustering, and every cluster is labeled through the KS method. The inferred trajectories are then aligned to the trajectories of the flow cytometry data through DTW alignment, and this process - enclosed within the lavender box - is repeated for a different number of variable genes in the filtration step from 10 to 100 in increments of 5. The trajectories with the minimum DTW distance are then fitted to the CTMC model, and the interstate transition rates are captured through this process.

The stationary distribution in this case can be solved using Equations. (3)-(4) and is given by

$$\pi_\infty = \left(\frac{\varphi_E}{1+\varphi_E+\varphi_M} \quad \frac{1}{1+\varphi_E+\varphi_M} \quad \frac{\varphi_M}{1+\varphi_E+\varphi_M} \right), \qquad \text{(Equation 5)}$$

where $\varphi_i \equiv \lambda_i/\mu_i$.

Phenotypic stability factors augment transitions into the hybrid state

PSFs have been previously reported[23,24] and act to stabilize the *H* phenotype. Here, we assume that their effect on EMT symmetrically enhances transitions into the *H* state ($\mu_E = \mu_M = \mu$ as shown in Figure 1A). In this case, we can explicitly derive the transition probability matrix, $P(t)$, through eigenvalue decomposition of the generator matrix, *G* and Equation 2 (See supplemental information for detailed solution). $P(t)$ can be represented as a function of the stationary distribution and given by (Equation 6) where $k_1 = -\eta_1 = \mu$, $k_2 = -\eta_2 = \lambda_E + \lambda_M + \mu$, $k_3 = k_2 - k_1 = \lambda_E + \lambda_M$:

$$P = \begin{pmatrix} \pi_{\infty,E} + e^{-k_1 t}\left[1 - \frac{\lambda_E}{k_3}\left(1 - \pi_{\infty,H} e^{-k_3 t}\right)\right] & \pi_{\infty,H}\left(1 - e^{-k_2 t}\right) & \pi_{\infty,M} - e^{-k_1 t}\left[\pi_{\infty,M} + \frac{\lambda_M}{k_3}\pi_{\infty,H}\left(1 - e^{-k_3 t}\right)\right] \\ \pi_{\infty,E}\left(1 - e^{-k_2 t}\right) & \pi_{\infty,H} + e^{-k_2 t}\left(1 - \pi_{\infty,H}\right) & \pi_{\infty,M}\left(1 - e^{-k_2 t}\right) \\ \pi_{\infty,E} - e^{-k_1 t}\left[\pi_{\infty,E} + \frac{\lambda_E}{k_3}\pi_{\infty,H}\left(1 - e^{-k_3 t}\right)\right] & \pi_{\infty,H}\left(1 - e^{-k_2 t}\right) & \pi_{\infty,M} + e^{-k_1 t}\left[1 - \frac{\lambda_M}{k_3}\left(1 - \pi_{\infty,H} e^{-k_3 t}\right)\right] \end{pmatrix}. \quad \text{(Equation 6)}$$

The temporal evolution described above can be extended by adding a spatial dimension. The dynamics of transitions in this case can be tracked by assuming a 1-dimensional signaling factor (such as TGFβ) inversely influences the rates of transitions into the epithelial, and mesenchymal states (λ_E, and λ_M) through a sigmoidal function. We show in the supplemental information that in this case the hybrid state is constant in space (See Figure S1).

In addition, the environment may be subject to discrete temporal changes. We let the times associated with each transition consist of an ordered set $0 = T_0 < T_1 < \cdots < T_i < \ldots T_M$ (T_M denotes the terminal time), and denote by $\Delta T_j \equiv T_j - T_{j-1}$. We also denote by $\lambda_{E,i}$, $\lambda_{M,i}$, and μ_i the rates associated on the interval $T_i \leq t_i \leq T_{i+1}$. Then, for the stochastic matrix P_i corresponding to t_i we have via the Chapman-Kolmogorov Equation that,

$$P(t) = \left(\prod_{j=1}^{i} P_j(\Delta T_j)\right) P_i(t - T_i), \text{for } T_i \leq t \leq T_{i+1} \quad \text{(Equation 7)}$$

Our approach therefore can capably handle any description of temporal dynamics that may be divided into finitely many time-homogeneous regimes.

Symmetric noise tunes the temporal evolution of EMT trajectories

Empirical evidence supports a relationship between longer exposure time to the EMT stimulating signal and noisier gene expression patterns resulting from the stochasticity of the underlying biological mechanisms (intrinsic noise).[35] In addition, cells are subject to stochasticity in the stimulating signal resulting from longer exposure to the inducing factor (extrinsic noise).[36] We account for both sources of noise using a *noise parameter* α that symmetrically scales the three transition rates such that $\lambda_i' = \alpha\lambda_i, \mu' = \alpha\mu$. The α parameter influences the temporal evolution of the process without affecting the stationary distribution given in Equation 5.

RESULTS

TGFβ time-course flow cytometry data reveals early and late EMT dynamics

Our model was trained on tri-replicate flow cytometry data from Jia et al.[37] Cells were treated with 5 ng/mL of TGFβ for a variable amount of time, after which treatment was withdrawn. This was repeated where withdrawal occurred for 3 biological replicates starting from day 3 to day 27 in increments of 3 days. The data was classified using a dual-color reporter system which tracked the expression of two well-established EMT biomarkers, E-cad (RFP) and ZEB1 (GFP). Although there is current debate on the role of ZEB1 in EMT induction for the MCF10A cell line,[38] many studies have reported ZEB1 as a reliable biomarker in this setting.[39–42] Furthermore, we note that the general pattern of increase in the *H* and *M* fractions following EMT induction with TGFβ is in agreement with previous studies.[43] Through this reporter system four cases of RFP+GFP- (*E*), RFP+GFP+ (*H*), RFP-GFP+ (*M*), and RFP-GFP- were identified. To isolate the three EMT-related trajectories, we normalized the three EMT-related cell fractions by the sum of *E*, *H*, and *M* cells.

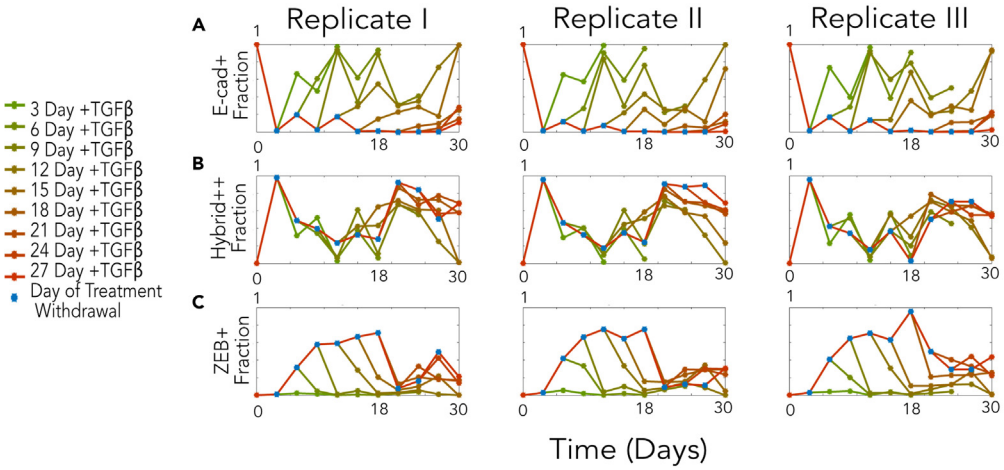


Figure 2. Time-course Flow Cytometry Data

(A–C) The experimental fraction of cells in (A) *E*, (B) *H*, and (C) *M* phenotypic states in time. The original flow cytometry data was normalized by omitting RFP-GFP- cells for each replicate.[44] Data shows a transient increase in the fraction of cells in the *H* state in the short-term that is followed by a transition into the *M* state long-term.

In reconstructing the temporal evolution of the three replicates via the above processed signals, we noted a pattern of short-term phenotypic stability followed by a long-term transition to a steady state under continued EMT induction (Figure 2). More specifically, the temporal EMT distribution was characterized by an initial predominance of the *E* fraction that declined and transiently transitioned into an *H* phenotype as shown in Figures 2A and 2B. This *H* phenotype then slowly transitioned into an *M* phenotype resulting in the stable co-existence of *H* and *M* cells (Figures 2B and 2C). Surprisingly, the *H* and *M* trajectories of the data switched at day 18. However, in absence of any specific underlying biological reason, we considered this switch as an empirical artifact and discarded it for any training purposes. Motivated by these findings, we proceeded to fit a dual-regime CTMC to the time-course data.

After calculating the mean of the trajectories over the three replicates, the first Markov chain was fitted from day 0 to the day where the *H* trajectory peaked (day 3). The system evolved via a probability transition matrix with parameters determined from an assumed hidden stationary distribution at day 3. We hypothesized that on reaching that threshold the system switches to a different regime until relaxation to a steady state. The second regime then began at day 3 and relaxed to steady state with the distribution at day 18 assumed to be the stationary distribution. This regime was normalized in time relative to the first regime, which was selected arbitrarily (we later discuss the approximation of steady-state using day 18 values).

Increases in the relaxation time to equilibrium distributions observed during the second regime were accounted for by applying the noise parameter discussed above to optimally fit the theoretic trajectories to empirical data in this regime (Figure 3A). Without loss of generality, we first set one of the transition rates (λ_E) to α and solved for the others by equating Equation 5 to the empirical stationary distribution. Optimization was performed using the *fminsearch* function in MATLAB to find the optimal α minimizing the total squared error between the three theoretic and corresponding empirical trajectories[45] (Figure 3B). For further validation, we also applied a non-linear unconstrained optimization method, *fminunc*, that utilizes gradient descent and confirmed agreement in optimal parameters (See Figure S2 in supplemental information). Collectively, we find that our two-regime approach performs well at characterizing empirical phenotypic fractions.

COMET's CTMC framework predicts enhanced transition rates en route to MET as a result of longer TGFβ exposure

The same procedure was repeated to minimize the Mean Squared Error (MSE) between the three theoretic and corresponding empirical trajectories following TGFβ treatment withdrawal. Here, we assumed that phenotypic reversion would ultimately recapitulate the initial, pre-TGFβ distributions. Using this approach, we found that the α values correlate directly with TGFβ treatment exposure (Figure 3C). In addition, we

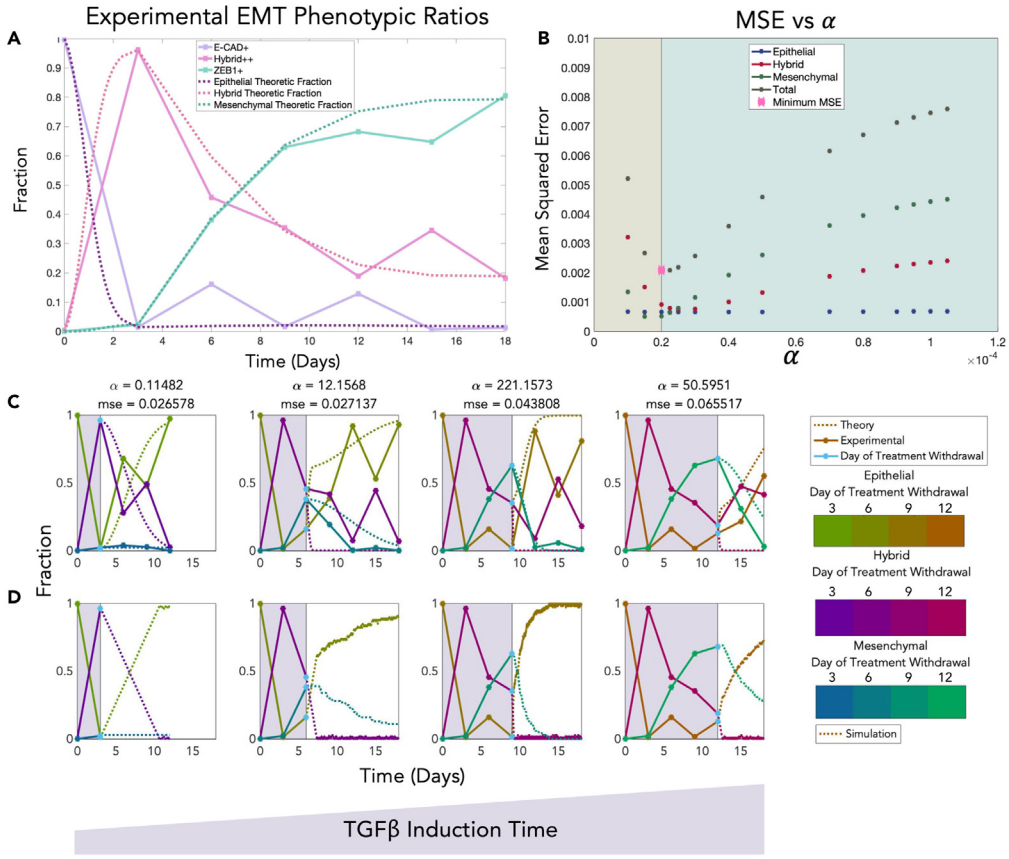


Figure 3. Stochastic Modeling Framework Applied to Time-course EMT Data

(A) The two-regime Markov chain (dotted) fits the non-monotonic phenotypic composition of EMT induction observed empirically (solid, averaged over 3 biological replicates).

(B) Numerical optimization is performed to fit the time-course data in part (A) by identifying the best-fit α.

(C) Experimental and theoretical phenotypic trajectories for EMT through time as a function of TGFβ treatment induction explained by increasing values of fitted α.

(D) This panel shows how stochastic Gillespie simulations generate distributions that are in large agreement with the experimental results. Gillespie simulations were performed over sufficiently large number of iterations to ensure steady state was reached (10000). The resulting values were then trimmed and normalized based on their inter-arrival times. The smallest inter-arrival times belonged to the third case (treatment withdrawal at day 9). This panel shows only one run of the simulations.

simulated the random arrival of the four possible transitions between the three states as shown in Figure 3D, using Gillespie algorithm[46] and found great consistency between simulations and theoretic fractions.

COMET infers the three EMT-related trajectories from scRNA-seq data

To evaluate our theoretical CTMC model on additional empirical data and to extend our framework into an analytical tool, we developed a data-driven pipeline that enables inference of EMT-related trajectories from scRNA-seq data. Time-dependent data of four cell lines (A549, DU145, MCF7, and OVCA420) treated with three EMT induction factors (10 ng/mL of TGFβ, 10 ng/mL of TNF, and 30 ng/mL EGF),[47] and dose-dependent steady state data of MCF10A cell line treated with various doses of TGFβ[48] were used to infer the three EMT-related trajectories.

We applied initial quality control (which included filtering based on the number of expressed house-keeping genes, total expressed genes, and mitochondrial percentage), followed by library size normalization and data filtration using previously reported EMT-related genes.[30] Next, we extracted the top 100 most variable genes using Seurat[49] in R across each available combination of cell line and induction

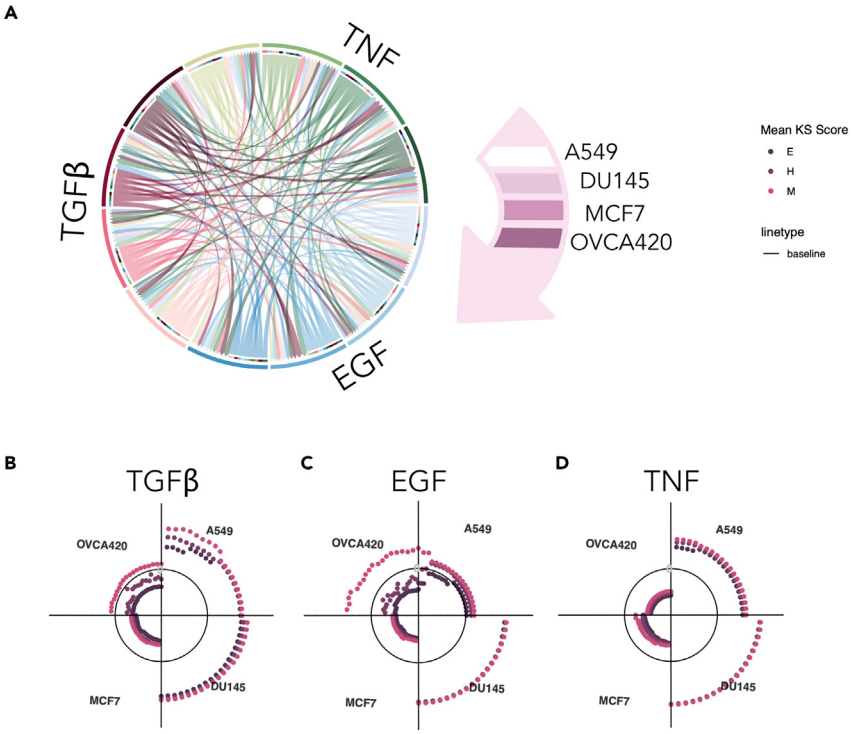


Figure 4. Data-driven Pipeline Scoring of Time-dependent Data

(A) Shows the Circos plot for the top 100 highly variable genes. The plot demonstrates that highly variable EMT-associated genes are shared among cell lines rather than across treatments. The length of the incoming arrows to every subsection of the plot indicates the number of genes that are common between two cases. Each colored slice is related to a specific induction factor (clockwise lighter to darker: TGFβ, EGF, and TNF).

(B–D) Each panel slice represents the mean KS score over 10 runs of the algorithm when using 10 to 100 highly variable genes (increasing clockwise in increments of 5). As shown in the plots, the mean KS scores appear to follow the same pattern across cell lines and do not depend on the EMT induction factors.

factor.[47] We found that the top variable genes were largely shared among cell lines rather than EMT induction factors (Figure 4A).

We then filtered the data further by identifying the 10 most highly variable genes, followed by imputation via MAGIC.[50] This was followed by dimensionality reduction via UMAP[51] and K-means clustering (The UMAP plots for one run of the algorithm - with data filtered for the optimal number of EMT genes - for the four cell lines is shown in Figures S3–S6 in supplemental information). We then labeled the three clusters using a previously developed EMT metric, the Kolmogorov-Smirnov (KS) based method.[31] KS scoring was performed on every cell based on their MAGIC imputed genetic profile with respect to the top 100 variable genes included (The top 100 variable EMT genes for every cell line and induction factor is reported in Table S1). This decision was made to ensure enough genes exist for KS scoring. Lastly, to associate the three clusters with *E*, *H*, and *M* phenotypes, we took the average of the KS scores for every cluster and sorted the clusters such that low, intermediate, and high KS scores corresponded with *E*, *H*, and *M* states, respectively.

From the fraction of cells in every cluster we inferred the three EMT-related trajectories and repeated the same procedure as before, this time varying the number of included variable genes starting from the top 15 to the top 100 in increments of 5. Owing to the stochasticity of our algorithm, we repeated this procedure 10 times for each cutoff and found the mean KS scores (Figures 4B–4D).

Consistent with the original study and other investigations of this data,[47,52] the pattern of the mean KS scores was found to be highly dependent on the cell line than EMT induction factor (See Tables S2–S4 for TGFβ, EGF, and TNF cases respectively.). This finding, in light of the conservation of variable genes within

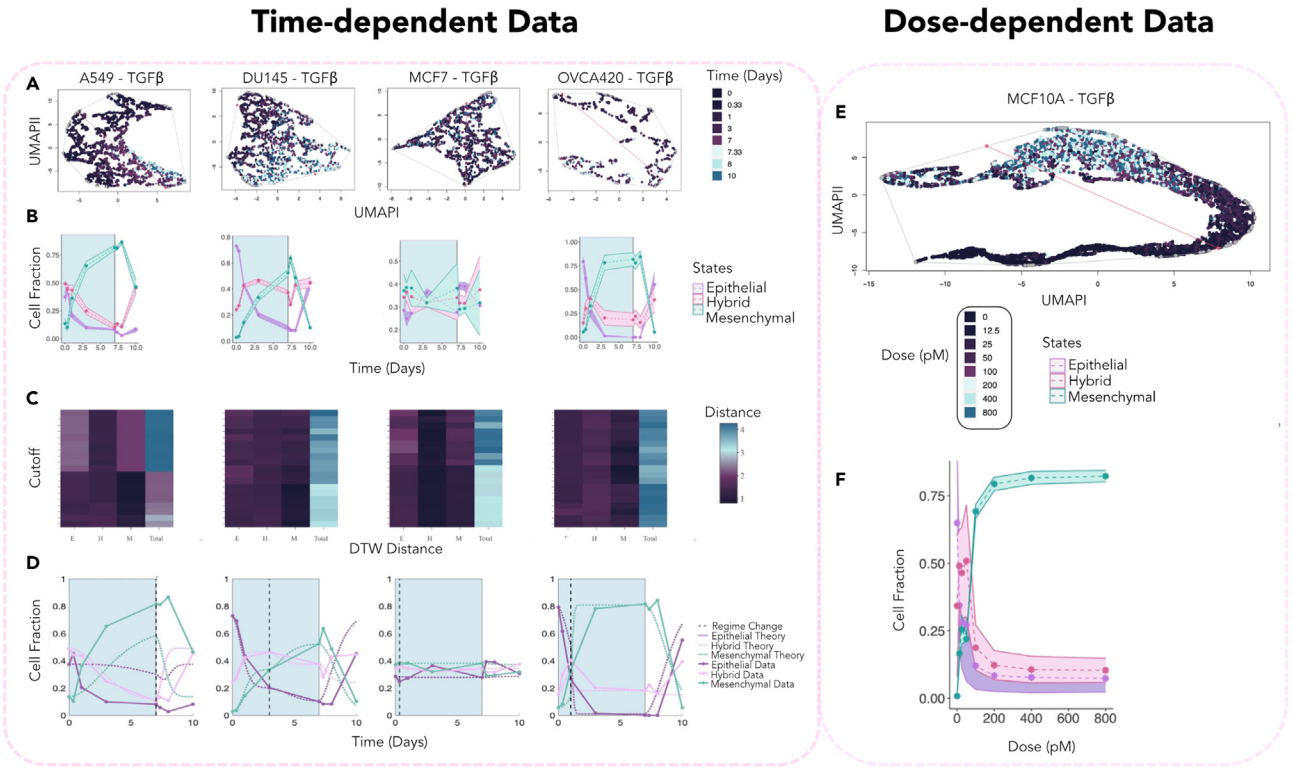


Figure 5. COMET Predicts the Timing and Distribution of Context-specific EMT

Figures (A), (B), (C), and (D) are time-dependent data of four cell lines treated with 10 ng/mL of TGFβ for seven days which underwent MET following treatment withdrawal for three days. Figures (E) and (F) represent the steady state information of the MCF10A cell line treated with various doses of TGFβ.

(A) Shows the UMAP plots for the four cell lines treated with TGFβ. Archetypal analysis with two archetypes[54] was performed on data and is shown in the figure. The figures demonstrate a clear transition from the *E* archetype to the *M* archetype for the three cases of A549, DU145, and OVCA420.

(B) This panel shows the inferred time-course trajectories with confidence intervals for 10 runs of the algorithm. The blue region depicts the duration of treatment.

(C) Heatmaps show the DTW distances resulting from the DTW alignment of the three EMT trajectories inferred using the pipeline and flow cytometry data. The lowest total DTW distance was used to choose the cutoff of highly variable genes which resulted in the time-course trajectories in the panel above.

(D) The stochastic model was fitted to the mean trajectories of the 10 runs for the optimal cutoff.

(E) The UMAP plot shows a transition from the *E* to *M* phenotypes as a function of treatment dose (gene cutoff 45).

(F) Plot illustrates the inferred trajectories from the data-driven pipeline for the dose-dependent data. The specific transition rates for each case are reported in Figure S23.

cell lines experiencing distinct induction treatment, further substantiates our prior finding that phenotypic transitions are dominated largely by the particular cell line, and not the particular induction factor, involved in undergoing an EMT. Our approach predicts that A549 and DU145 cell lines exhibit higher mean KS scores (A549-TGFβ: 0.501, DU145-TGFβ: 0.559, A549-EGF: 0.059, DU145-EGF: 0.596, A549-TNF: 0.405, and DU145-TNF: 0.648) compared to OVCA420 and MCF7 (OVCA420-TGFβ: − 0.110, MCF7-TGFβ: − 0.276, OVCA420-EGF: − 0.074, MCF7-EGF: − 0.340, OVCA420-TNF: − 0.374, and MCF7-TNF: − 0.259) over a majority of cutoffs (Figures 4B–4D), which is consistent with previously reported findings of MCF7[30] and OVCA420[53] exhibiting *E* characteristics. Moreover, prior work identified a higher percentage of hybrid cells for DU145 and coexistence of *E*, *H*, and *M* fractions for A549,[30] which can also be appreciated by constructing our inferred trajectories for day 0 (Figure 5B). Intriguingly, this suggests that our approach reliably quantifies EMT phenotypic composition at the single-cell resolution from next generation sequencing data.

Next, to account for the variability in different runs of the algorithm, we calculated confidence intervals for every trajectory (See Figures S7–S18 in supplemental information for inferred trajectories with confidence intervals. Figures 5A, and 5B show the UMAP and inferred trajectories with confidence intervals for the optimal cutoffs of highly variable EMT genes respectively). Occasionally, we observed that the inclusion of 5 additional genes drastically changes the EMT trajectories (Figures S7–S18). This is likely because of

an abrupt change in the number of resolvable *H* states based on the genes included which could not be detected by our three-state model. Our pipeline inferred ambiguous and highly variable trajectories for the majority of the EGF and TNF cohort (all inferred trajectories for these cases are reported in Figures S11–S18). We note that from the phase contrast images reported by Cook and Vanderhyden,[47] it is indeed unclear whether cells went through EMT in these cases. As a result, we proceeded with our subsequent analysis performed on TGFβ-treated cell lines only.

We hypothesized that EMT in this additional dataset would temporally evolve in a similar fashion as the flow cytometry data of Jia et al.[37] and could thus be explained by our framework. Toward this end, we optimized the number of EMT-related genes that result in mean trajectories most similar to those of the originally considered flow cytometry data by utilizing Dynamic Time Warping (DTW) alignment. DTW alignment optimizes the alignment of time-course trajectories by tweaking the time axis recursively.[55] As a result, the DTW distance can be used to measure similarities in dynamics of transitions by ignoring the context-specific timing of events.

We performed DTW alignment of the scRNA-seq trajectories with the corresponding trajectories of the mean signal of the flow cytometry data. The resulting heatmap which illustrates the DTW distances for every cutoff of highly variable genes is shown in Figure 5C. The cutoff with the lowest sum of DTW distances of the three trajectories from the flow cytometry data was considered for fitting the stochastic model to time-dependent data. Although for the optimal cases we observed cells to be on a spectrum from epithelial to mesenchymal phenotypes on the UMAP plot (Figures S3–S6), we noticed that the measured EMT spectrum visualized by UMAP plots became less discernible as we increased or decreased the number of included highly variable genes (Figures S3–S6). To quantify this phenomenon, we fit a minimum spanning tree (MST) to the UMAP plot to measure the pairwise distances between cells in Figure S19 and showed that the maximum edge of the tree is minimized in the neighborhood of the optimal cutoff of variable gene (the minima fall between a cutoff of 40–55 for all TGFβ cases; Figure S19 in supplemental information).

We next proceeded to test our results on an independent dataset featuring dose-dependent EMT induction.[48] Because the optimal cutoff of highly variable genes fell uniformly between 40 and 55, we considered a cutoff of the top 45 highly variable genes for this analysis (Figure 5E depicts the dimension reduction plot for the dose-dependent data where cells treated with low doses of TGFβ at steady state move from one cluster in the neighborhood of an *E* archetype to an *M* phenotype as the dose of treatment increases). The resulting optimal trajectories for the time-dependent and dose-dependent data with confidence intervals are shown in Figures 5B and 5F respectively. Although our 45-gene approach exhibited some differences between the predicted and simulated phenotypic fractions as a function of dose reported by previous studies, our overall trends were in general agreement[56] (Figure 5F). From the steady state fractions obtained from the dose-dependent data, it is evident that the distribution of fractions for the flow cytometry data at day 18 resembles the equilibrium state fractions closely (roughly 113.6 pM, see supplemental information for dimensionality analysis).

COMET CTMC predicts time dynamics of context-specific EMT

To further evaluate the dynamical effects of various induction factors on a cell-line-specific basis, we next fitted COMET's CTMC model to time-dependent data of each of the four cell lines treated with TGFβ. We then fitted Markov chains from day 0 to the day where the *H* state peaked with the timescale normalized based on the flow cytometry data. Given our prior analysis on empirical time-course proportions, we performed this normalization assuming that TGFβ induction, with a different dose on a distinct cell line, would follow similar dynamics en route to EMT. However, the stationary distribution of both regimes was determined based on the time-dependent data because of the discrepancies in cell line type and dose of TGFβ as well as lack of available steady state information. For one of the cell lines, A549, the *H* state was maximized at day 0 thus we only fitted a single Markov chain to the data, assuming this case starts within the second regime. This assumption is supported by the fact that CEACAM6, an inducer of EMT,[57] is the most variably expressed gene at day 0 for A549 across all treatment cases (Figures S20–S22 in supplemental information). For all cases, the second regime of the Markov chain was fitted from the day when the hybrid state peaked to the day of treatment withdrawal (day 7).

Next, we proceeded to apply the noise parameter, α, to explain changes in the temporal evolution of the Markov chain. We then fitted a separate Markov chain following treatment withdrawal from day 7 to the end

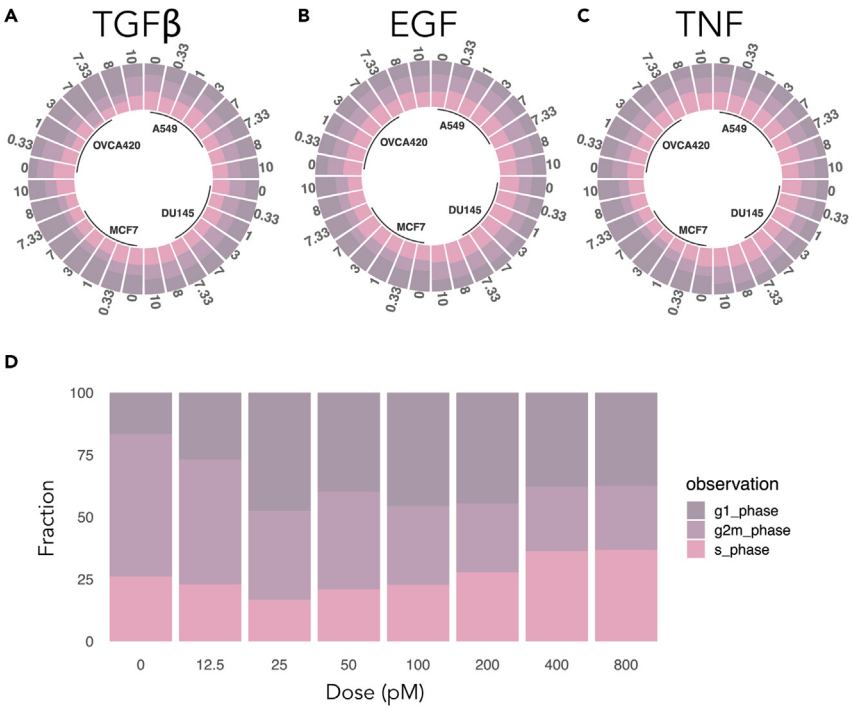


Figure 6. Time-dependent and Dose-dependent Data Cell Cycle Scoring

(A–C) Illustrate the cell cycle scores for every cell line of the time-dependent data. Similar to the KS scores, cell cycle scores appear to be more cell line dependent than induction factor specific.

(D) Shows the cell cycle fractions for the dose-dependent data. Cell cycle fractions at steady state are reported as a function of TGFβ dose. Cell cycle information starts with a high percentage of cells in the G2M phase which declines, followed by a transient increase in the G1 phase and an increase in S phase long-term.

of the experiment (day 10) assuming the trajectories revert back to the initial distribution at day 0. Similarly, the optimal fit was obtained using the α parameter. The final results are shown in Figure 5D. Because our optimization depends on the MSE of three trajectories, occasionally an optimized theoretic curve only fits one trajectory well while poorly fitting the other two (A549 of Figure 5D). In these cases, a similar MSE can be obtained through a larger or smaller α parameter that optimally fits the other two trajectories. Overall, we observed reasonable consistency in our inferred trajectories and the CTMC theoretic trajectories (See Figure S23 in supplemental information for the exact inter-state transition rates for the second regime and following treatment withdrawal for the TGFβ cases).

Cell cycle scoring of the time-dependent data reveals cell line-dependent cell cycle fractions

To further evaluate the utility of COMET, we next interrogated the growth rates of cells by extracting cell cycle fractions through Seurat normalization and performing cell cycle scoring.[49] Consistent with previous reports,[58] we found cell cycle stages to be similar across different cell lines for the time-dependent data as shown in Figures 6A–6C.

We assessed the trends in G1 phase of cell cycle stages of time-dependent data during TGFβ treatment using Kendall Tau statistic[59] and found that in general the G1 stage positively trends for TGFβ treated cells during treatment (day 0–7) as shown in Figure 6A, consistent with previous reports.[60] However, this trend was stronger for the two cell lines with lower mean KS scores (MCF7: 0.6, and OVCA420: 0.8) than DU145 and A549 (DU145: − 0.2, and A549: 0.4) (Figure 6A). This is consistent with previous reports of normal epithelial cells arresting at G1 phase following TGFβ EMT induction whereas those with high cell proliferation fail at cell-cycle arrest resulting from genomic instabilities and mitotic defects.[61]

We also discovered a higher proportion of cells in S and G2M phases that happened to coincide with the two cell lines having higher KS scores (DU145, and A549) as they progressed through EMT. Furthermore, we

found the fraction of cells in G1 phase to be relatively high for the MCF7 cell line compared to the other cell lines at baseline day. Surprisingly, in cell cycle scoring the dose-dependent data we observed the fraction of cells in the S phase increasing as a function of TGFβ dose (Figure 6D). Consistent with studies reporting an increase in the fraction of cells in S phase as a result of a decrease in p21 expression while cells go through EMT,[62] COMET predicted an increase in the fraction of *M* cells as a function of dose for the optimal cutoff (Figure 5F). However, as reported in Figure S24 in supplemental information, for a cutoff of 100 highly variable genes COMET predicted a higher percentage of cells in the hybrid state for the dose-dependent data.

GSEA resolves enrichment in cancer-specific hallmarks for each of the EMT-related phenotypes across distinct cell lines

Lastly, to further test the ability of COMET to accurately infer the three EMT states from time-dependent scRNA-seq data, we performed Gene Set Enrichment Analysis (GSEA)[63] on every run of the pipeline for the optimal number of highly variable genes. We start our analysis with the TGFβ signaling pathway because we assume cells would be up-regulated for the TGFβ signaling pathway when undergoing EMT and down-regulated following treatment withdrawal with the intensity of these patterns dependent on the rate and fraction of cells undergoing each process.

For instance, as shown in Figure 5D, a large fraction of *E* cells in the A549 sample appear to have undergone EMT before TGFβ treatment, appearing stationary post-treatment. This would suggest that the *E* cluster may be significantly enriched for TGFβ signaling, which we in fact confirm (Figure 7C). Similarly, the dynamics of the *M* proportion appear to indicate a rapid decline following treatment withdrawal and indeed we find that the TGFβ signaling pathway is depleted in this case. Lastly, the *H* trajectory of the A549 cell line appears to have a mixed response where TGFβ withdrawal results in more noticeable effects than induction.

A similar pattern is observed for DU145. In this case, the majority of the *E* cells reside in the pre-treatment regime thus significant up-regulation of the TGFβ signaling pathway occurs for the *E* cluster. Although the *M* fraction almost reverts back to its initial fraction at a shorter period. Thus, TGFβ signaling pathway is down-regulated for the *M* trajectory. On the other hand, the *H* fraction is almost stabilized across time and because the other cells are transitioning into the *H* state, a larger fraction of hybrid cells is up-regulated for TGFβ pre-treatment.

In contrast, for the MCF7 case, the three trajectories appear to stably coexist and thus we do not observe either up-regulation or down-regulation for any of the clusters. As mentioned previously, we are unsure whether this cell line underwent EMT in the original data.[47] Lastly, for the OVCA420 cell line almost all of the three trajectories revert back to their initial fraction and the pattern of EMT progression and MET is fairly symmetric. The GSEA analysis faithfully captures this observation, with enrichment scores approaching zero with small deviance for the *E* and *H* trajectories (as they do not fully revert back to the initial day fractions).

We then analyzed the EMT hallmark cases and noted their similarity to the TGFβ signaling pathway enrichment scores. However, issues arise during GSEA analysis of EMT-related clusters because of the concomitant existence of both *E* and *M* biomarkers in the gene set. As a result, net zero enrichment may be reported when, for example, the up-regulation of *E* markers happens concurrently with the down-regulation of *M* markers (Figure 7A).

We observed a transition from up-regulation to down-regulation from *E* to *M* clusters for the P53 pathway gene set across all cell lines (Figure 7D). This collective pattern is consistent with previous reports that highlight the role of the P53 pathway in regulating EMT by binding to the miR-200c promoter.[64] Surprisingly, this contrasted with our observation of different patterns of up-regulation or down-regulation of the angiogenesis hallmarks across different cell lines. Although angiogenesis followed the same pattern of the EMT gene set in the case of A549 (Figures 7A and 7B), it followed the exact opposite pattern for DU145 (Figure 7B). Although the tumor stroma is known to influence patterns of vascularization for different tumor types, our results suggest tumor sub-type-specific patterns of angiogenesis in the absence of tumor stroma and under the influence of the same signaling factor.[65]

From the gene set enrichment analysis of cell lines for the notch signaling pathway, we observed a pattern of up-regulation to down-regulation only for the OVCA420 case. For DU145 and MCF7, we did not observe

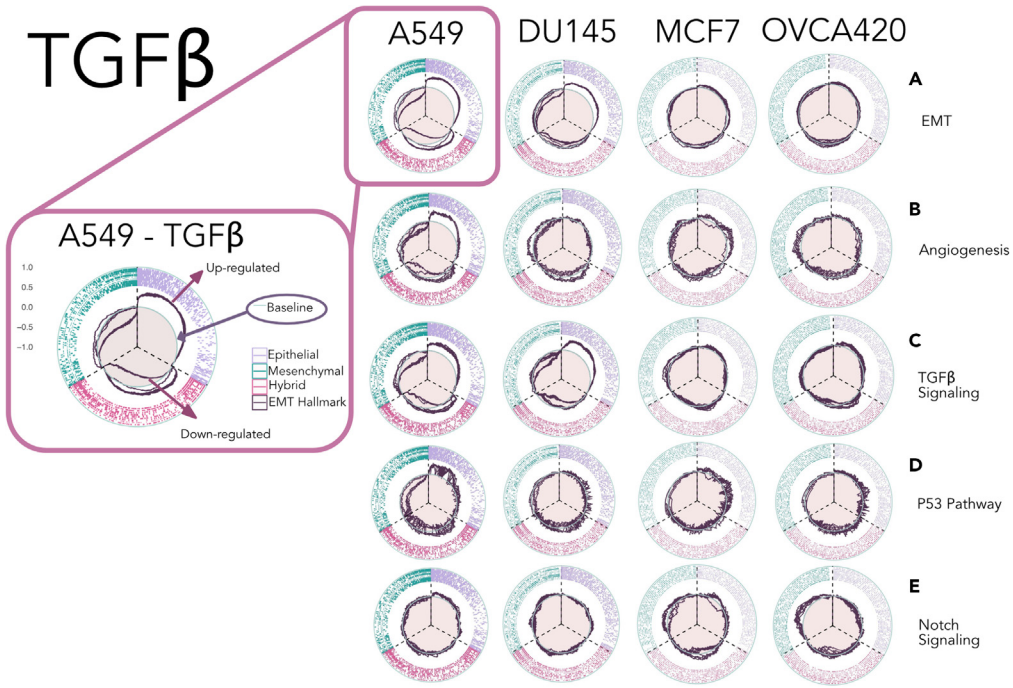


Figure 7. GSEA Plots of Relevant Hallmarks for Each Phenotypic State

GSEA plots are overlapped for 10 runs of the algorithm. The baseline separates the up-regulated and down-regulated genes. Every one-third of the circular plot bound by dashed lines represents one phenotypic state (Lines falling inside the baseline circle demonstrate an abundance of down-regulated hallmark genes and lines falling outside the circle show an abundance of up-regulated hallmark genes).

(A–E) Shows cluster-specific GSEA plot for the EMT hallmarks, it is evident from the plot that the EMT hallmark gene set is more variable across states for DU145 and A549. An overall up-regulation trend for the EMT gene set transitions to a down-regulation trend for the mesenchymal state. GSEA plots for the three phenotypic states are given for (B) angiogenesis, (C) TGFβ signaling, (D) P53 pathway, and (E) notch signaling.

significant up-regulation or down-regulation of the notch signaling pathway. However, for A549 we observed a slight up-regulation of the notch signaling pathway. This result is inconsistent with previous reports of TGFβ directly resulting in the up-regulation of notch ligand after EMT induction. However, the cell lines used in this study are different and from Figure 7E, our results suggest tumor type dependent pattern for notch signaling activation.[66]

As shown in Figures 7A–7D, the patterns of GSEA up-regulation or down-regulation are more pronounced across the four gene sets of EMT, angiogenesis, TGFβ signaling, and P53 pathway in the cell lines with higher KS scores (A549, and DU145) versus those with lower KS scores (OVCA420, and MCF7).

Furthermore, we note a similar pattern of GSEA enrichment for clusters across several different hallmarks for the same cell line. We evaluated the similarities between gene sets via pairwise Jaccard distance and the number of shared genes across gene sets fell below 10%. This result suggests extensive cross talks between these processes and the existence of gene regulatory networks controlling the concurrent up-regulation or down-regulation of gene sets across cell lines. Collectively, enrichment analysis further validates COMET's characterization of scRNA-seq signatures of each of the three relevant phenotypes, and can be more generally used to investigate pathways that are actively involved in maintaining each state.

DISCUSSION

Despite the vast body of EMT-related studies, our understanding of EMT and its specific role in metastasis is far from complete. Owing to the multifaceted role of transcriptional regulation in driving EMT,[67] interrogating metastasis at the genetic level has often yielded in inaccurate findings,[10] and there is a larger need for tools to infer the dynamics of EMT at the phenotypic level.

Here, we presented an analytic tool, COMET, for reliably inferring and predicting EMT trajectories from scRNA-seq data. We applied COMET to time-course data of cell lines treated with TGFβ, and showed that our method was able to recapitulate the findings of previous reports at the single-cell level.[30] Reliable identification of EMT-related phenotypic fractions from increasingly available scRNA-seq data will enable us to better understand EMT and its pattern of progression in cancer.

Although there are at present three well-established EMT metrics that are widely used,[30–32] each are trained on microarray data with one metric inferring the three E, H, and M phenotypes using only CDH1, VIM, and CLDN7.[30] As a result, applying these methods to scRNA-seq data presents unique challenges. Unlike microarray data, the normalization of scRNA-seq data requires a different approach, and many of the housekeeping genes relied on for normalization in previous studies are filtered out. Furthermore, during our analysis we found that several known EMT biomarkers, such as CDH1, VIM, and CLDN7, were not among the most variable genes in the data we analyzed. To better elucidate the difficulties of applying these metrics to scRNA-seq data, we added CDH1, VIM, and CLDN7 to the list of EMT genes post-filtration and re-ran the pipeline. We associated the expression values of these three biomarkers with the clustering results from a previous run of the pipeline, in which they were excluded from the list of EMT genes. As demonstrated in Figure S25 in supplemental information, although the three clusters appear to be well-separated, it is nearly impossible to resolve the data based on the expression of these three biomarkers alone. Moreover, applying the KS method to microarray data also presents its own obstacles. The thresholds imposed on KS scores for the identification of phenotypes often fail because of the differences in the normalization and data processing steps of scRNA-seq data. Therefore, to infer EMT-related states from time course scRNA-seq data, COMET utilizes a hybrid cluster-based KS scoring pipeline.

On reliable identification of EMT related phenotypes, COMET feeds the inferred states to a continuous-time Markov chain for the characterization of EMT dynamics in a stationary population of cells. Markov chains have been widely used to model phenotypic transitions in plastic populations.[68,69] Previous models have employed both discrete-time and continuous-time Markov chains to describe transitions in heterogeneous populations, with successful applications in various biological phenomena.[70–72] Our model assumed no influence from cell division or genetics, which may contribute to tumor heterogeneity, but we also expect that their effects are negligible within the short time frame of the *in-vitro* experiments considered here.

Using COMET, we showed that our inferred trajectories collectively followed a pattern of a short-term monotonic increase in the *H* state followed by a transition to the *M* phenotype leading to the stable coexistence of the two phenotypes. This pattern was exceptionally not observed for the MCF7 data among the TGFβ-treated cohort. We note that based on the phase contrast images of the MCF7 cell line reported by Cook and Vanderhyden,[47] cells do not display fully epithelial characteristics at day 0 and their progression through EMT is subsequently unclear. As a result, this heterogeneity may confound the ability of our pipeline to resolve the EMT trajectories where the gene expression signature does not undergo phenotypic transitions.[73] However, this observation was constricted to the time length of the study. We note that for the scRNA-seq data, cell lines were only treated with the EMT induction factor for 7 days and as the original paper states it is unclear whether cells reached steady state distribution at day 7.[47]

Adding to this complexity is the fact that we observed a switch between the *H* and *M* trajectories at day 18 from the flow cytometry data of Jia et al.[37] which was acquired over a longer period of 30 days. Although we discarded data from days 18 − 30, assuming the switch is an empirical artifact, this temporal in-homogeneity can be theoretically modeled via a separate CTMC regime. However, the lack of available gene expression data for this period limited our ability to investigate the underlying reasons behind this empirical switch through other concurrent processes such as cell cycle scoring or GSEA. Furthermore, this pattern of a switch between *H* and *M* phenotypes was not consistent with our dose-dependent steady-state distributions obtained from the data of Panchy et al.[48] Nonetheless, our analytic tool, COMET, provides researchers with the ability of reliably investigating this switch further on longer time-course scRNA-seq data in the future.

Our approach, although useful for robustly characterizing EMT at the single-cell level, is not without limitation. Phenotypically heterogeneous cells can exhibit gene expression profiles that cooperate or interfere with overall signal detection,[74] and this is further compounded by noisy cellular division.[26] Although we

have assumed in our analysis that populations under study were non-dividing, cell cycle scoring suggested the variable presence of division signatures across available experimental contexts as shown in Figure 6. However, despite this simplifying assumption, COMET performed exceptionally well in cases of even higher proliferating cell lines such as DU145 which has an abundance of cells in S phase (See Figure 6A). This result may suggest that the dominating contribution to EMT heterogeneity in the datasets considered here are driven by stochastic transitions rather than by cell-division.

Furthermore, there is a lack of consensus in the literature on the number of *H* phenotypes[75] with their estimation possibly dependent on the number of biomarkers considered for EMT classification. Additional studies have also suggested that EMT is a non-Markovian process because of the existence of several microstates within macrostates.[76] In our model, we specifically optimized the number of EMT biomarkers based on their ability to resolve the data into three states depending on their sample-specific gene expression variation. Using a rigorous validation procedure, we showed that the number of highly variable genes, selected as representative of EMT and commonly chosen arbitrarily in computational studies, can drastically change the results of the analysis. Although our general findings on multiple independent time- and dose-dependent datasets[47,48] were consistent with previous reports, because of the lack of predefined sample-specific *E*, *H*, and *M* fractions from scRNA-seq data, subsequent analysis on additional EMT-specific datasets will further test COMET's predictive accuracy. Our modeling approach, given sufficient availability of additional detailed data, could be used in an identical manner to account for and study multiple intermediate phenotypes in EMT.

Using COMET, our data-driven pipeline captured the phenotypic spectrum observed following dimension reduction of EMT-related genes and found through archetypal analysis with two archetypes (*E* and *M*) that the *H* state with an intermediate KS score is always spatially intermediary to *E* and *M*. This observation is in broad agreement with the recent specialist-generalist frameworks devised for the EMT process whereby the *E* and *M* populations are predicted to be optimized for one task at hand, whereas the *H* population confers fitness advantages owing to better collective performance at multiple tasks.[77]

In addition, previous models included the possibility of directly transitioning from an epithelial state into a mesenchymal state during EMT,[78] and many more report that cells transition into an epithelial state without ever visiting the intermediate state during MET.[79,80] In our approach, we decided to assume all transitions proceed through an intermediate state as cells are shown to retrace the EMT footsteps through a continuum of positions in the space of EMT-related genes while going through MET. However, we note that this reverse transition happens much faster than EMT, thus it may be more difficult to observe this phenomenon empirically.

One of the main assumptions of our model was the symmetric effects of PSFs on the transition rates into the *H* state. This is supported to an extent by the symmetric trajectories and consistency between predicted theoretic and inferred trajectories in Figure 5D. However, some evidence has suggested that the stronger presence of PSFs delays the transition into the *H* state in addition to the longer mean residence time within the *H* state.[24] We demonstrate through Gillespie simulation in Figure S26 that the short-term stability of the EMT-related trajectories can be explained by relaxing this assumption and the entire process can be modeled via a single CTMC (See supplemental information for full solution). However, we emphasize that a single-regime model as an alternative approach would introduce another free parameter, which would require further investigation. We also note that to gain a deeper understanding of the extent of regime changes during early versus late EMT induction, it is important to examine the constitutive relationship between the transitions and the time homogeneity of the Markov chain. This topic warrants further exploration in future work, and empirical analyses will be necessary to support or refute any hypotheses.

Our analysis showed concurrent up-regulation and down-regulation of the angiogenesis, TGFβ signaling, and P53 pathway hallmarks consistent with the EMT hallmark. The low number of common genes between these sets suggests the possible existence of a strong gene regulatory network controlling these processes. Future work inferring the gene regulatory networks governing context-specific EMT would provide additional insights into the underlying mechanisms that govern the phenotypic transition rates.

Lastly, to further extend our framework to *in-vivo* data at the tumor setting, we need to account for the difficulties of distinguishing stem-like cells from the elements of the TME and detecting them in circulation

in-vivo.[15] Adding to this complexity is the fact that cells can also break away from the primary tumor and travel collectively as CTCs while retaining their epithelial characteristics. This phenomenon, also known as the Unjamming Transition (UJT), may happen separately from EMT and contribute to metastasis.[81] As a result, although our simplistic model could infer and predict EMT trajectories at the single-cell resolution, future work needs to consider the complex interactions of tumor cells with the TME to reliably understand the progression of EMT *in vivo*.

In conclusion, we introduced COMET, an EMT trajectory inferential and predictive analytic tool. Utilizing COMET, we further resolve the context-specific nature of EMT and argue that the observed pattern of progression is highly dependent on the tumor subtype. Here, we specifically applied COMET to EMT data. However, we anticipate that our general framework is widely applicable for studying phenotypic transitions in other biological processes with intermediate phenotypes.

Limitations of the study

Here, we presented a dual data-driven theoretical framework to infer EMT-related trajectories and predict the timing and distribution of consequent phenotypic intra-tumoral heterogeneity from scRNA-seq data. We confirmed the validity of our results through cell cycle scoring and gene set enrichment analysis. However, we note that our pipeline was only applied on cell line data and as mentioned in the manuscript, there is a large debate on the role of EMT in cancer metastasis and future investigation is required to assess the applicability of our method more thoroughly on tumor data in the presence of stroma.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Model development
  - Flow cytometry data
  - Gillespie simulation
  - Gene set enrichment analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106964.

ACKNOWLEDGMENTS

We would like to thank Abhijeet Deshmukh and Sendurai A Mani from UT MD Anderson Cancer Center for providing the raw flow cytometry data used in this study.[37] We would also like to thank David P Cook for insightful suggestions and comments which we believe improved the quality of our manuscript.

J.T.G. was supported by the Cancer Prevention and Research Institute of Texas (RR210080). J.T.G. is a CPRIT Scholars in Cancer Research.

AUTHOR CONTRIBUTIONS

A.N. performed the research, created the computational pipeline and developed and refined the methodology. A.N., J.T.G., and M.K.J. analyzed and interpreted the data. A.N., J.T.G., and M.K.J. wrote the paper. J.T.G. conceived of the research. J.T.G. and M.K.J. designed the research. All authors have read and edited the manuscript.

DECLARATION OF INTERESTS

Authors report no competing interests.

Received: December 23, 2022
Revised: March 24, 2023
Accepted: May 22, 2023
Published: June 5, 2023

REFERENCES

1. Trelstad, R.L., Hay, E.D., and Revel, J.D. (1967). Cell contact during early morphogenesis in the chick embryo. Dev. Biol. *16*, 78–106. https://doi.org/10.1016/0012-1606(67)90018-8.

2. Haensel, D., and Dai, X. (2018). Epithelial-to-mesenchymal transition in cutaneous wound healing: where we are and where we are heading. Dev. Dynam. *247*, 473–480. https://doi.org/10.1002/dvdy.24561.

3. Mittal, V. (2018). Epithelial mesenchymal transition in tumor metastasis. Annu. Rev. Pathol. *13*, 395–412. https://doi.org/10.1146/annurev-pathol-020117-043854.

4. Shibue, T., and Weinberg, R.A. (2017). Emt, cscs, and drug resistance: the mechanistic link and clinical implications. Nat. Rev. Clin. Oncol. *14*, 611–629. https://doi.org/10.1038/nrclinonc.2017.44.

5. Kalluri, R., Weinberg, R.A., et al. (2009). The basics of epithelial-mesenchymal transition. J. Clin. Invest. *119*, 1420–1428. https://doi.org/10.1172/JCI39104.

6. Espinoza, I., and Miele, L. (2013). Deadly crosstalk: notch signaling at the intersection of emt and cancer stem cells. Cancer Lett. *341*, 41–45. https://doi.org/10.1016/j.canlet.2013.08.027.

7. Heldin, C.H., Vanlandewijck, M., and Moustakas, A. (2012). Regulation of emt by tgfβ in cancer. FEBS Lett. *586*, 1959–1970. https://doi.org/10.1016/j.febslet.2012.02.037.

8. Tam, W.L., and Weinberg, R.A. (2013). The epigenetics of epithelial-mesenchymal plasticity in cancer. Nat. Med. *19*, 1438–1449. https://doi.org/10.1038/nm.3336.

9. Nieto, M.A. (2009). Epithelial-mesenchymal transitions in development and disease: old views and new perspectives. Int. J. Dev. Biol. *53*, 1541–1547. https://doi.org/10.1387/ijdb.072410mn.

10. Vogelstein, B., Fearon, E.R., Kern, S.E., Hamilton, S.R., Preisinger, A.C., Nakamura, Y., and White, R. (1989). Allelotype of colorectal carcinomas. Science *244*, 207–211. https://doi.org/10.1126/science.2565047.

11. Yu, Y., Xiao, C.H., Tan, L.D., Wang, Q.S., Li, X.Q., and Feng, Y.M. (2014). Cancer-associated fibroblasts induce epithelial–mesenchymal transition of breast cancer cells through paracrine tgf-β signalling. Br. J. Cancer *110*, 724–732. https://doi.org/10.1038/bjc.2013.768.

12. Pistore, C., Giannoni, E., Colangelo, T., Rizzo, F., Magnani, E., Muccillo, L., Giurato, G., Mancini, M., Rizzo, S., Riccardi, M., et al. (2017). Dna methylation variations are required for epithelial-to-mesenchymal transition induced by cancer-associated fibroblasts in prostate cancer cells. Oncogene *36*, 5551–5566. https://doi.org/10.1038/onc.2017.159.

13. Vakili-Ghartavol, R., Mombeiny, R., Salmaninejad, A., Sorkhabadi, S.M.R., Faridi-Majidi, R., Jaafari, M.R., and Mirzaei, H. (2018). Tumor-associated macrophages and epithelial–mesenchymal transition in cancer: nanotechnology comes into view. J. Cell. Physiol. *233*, 9223–9236. https://doi.org/10.1002/jcp.27027.

14. Gao, D., Joshi, N., Choi, H., Ryu, S., Hahn, M., Catena, R., Sadik, H., Argani, P., Wagner, P., Vahdat, L.T., et al. (2012). Myeloid progenitor cells in the premetastatic lung promote metastases by inducing mesenchymal to epithelial transitionmyeloid progenitors promote metastatic outgrowth. Cancer Res. *72*, 1384–1394. https://doi.org/10.1158/0008-5472.CAN-11-2905.

15. Fischer, K.R., Durrans, A., Lee, S., Sheng, J., Li, F., Wong, S.T.C., Choi, H., El Rayes, T., Ryu, S., Troeger, J., et al. (2015). Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. Nature *527*, 472–476. https://doi.org/10.1038/nature15748.

16. Terry, S., Savagner, P., Ortiz-Cuaran, S., Mahjoubi, L., Saintigny, P., Thiery, J.P., and Chouaib, S. (2017). New insights into the role of emt in tumor immune escape. Mol. Oncol. *11*, 824–846. https://doi.org/10.1002/1878-0261.12093.

17. Tripathi, S.C., Peters, H.L., Taguchi, A., Katayama, H., Wang, H., Momin, A., Jolly, M.K., Celiktas, M., Rodriguez-Canales, J., Liu, H., et al. (2016). Immunoproteasome deficiency is a feature of non-small cell lung cancer with a mesenchymal phenotype and is associated with a poor outcome. Proc. Natl. Acad. Sci. USA *113*, E1555–E1564. https://doi.org/10.1073/pnas.1521812113.

18. Yu, M., Bardia, A., Wittner, B.S., Stott, S.L., Smas, M.E., Ting, D.T., Isakoff, S.J., Ciciliano, J.C., Wells, M.N., Shah, A.M., et al. (2013). Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. science *339*, 580–584. https://doi.org/10.1126/science.1228522.

19. Cheung, K.J., and Ewald, A.J. (2016). A collective route to metastasis: seeding by tumor cell clusters. Science *352*, 167–169. https://doi.org/10.1126/science.aaf6546.

20. Lu, M., Jolly, M.K., Levine, H., Onuchic, J.N., and Ben-Jacob, E. (2013). Microrna-based regulation of epithelial–hybrid–mesenchymal fate determination. Proc. Natl. Acad. Sci. USA *110*, 18144–18149. https://doi.org/10.1073/pnas.1318192110.

21. Jolly, M.K., Boareto, M., Huang, B., Jia, D., Lu, M., Ben-Jacob, E., Onuchic, J.N., and Levine, H. (2015). Implications of the hybrid epithelial/mesenchymal phenotype in metastasis. Front. Oncol. *5*, 155. https://doi.org/10.3389/fonc.2015.00155.

22. Bierie, B., Pierce, S.E., Kroeger, C., Stover, D.G., Pattabiraman, D.R., Thiru, P., Liu Donaher, J., Reinhardt, F., Chaffer, C.L., Keckesova, Z., and Weinberg, R.A. (2017). Integrin-β4 identifies cancer stem cell-enriched populations of partially mesenchymal carcinoma cells. Proc. Natl. Acad. Sci. USA *114*, E2337–E2346. https://doi.org/10.1073/pnas.1618298114.

23. Pal, A., Barrett, T.F., Paolini, R., Parikh, A., and Puram, S.V. (2021). Partial emt in head and neck cancer biology: a spectrum instead of a switch. Oncogene *40*, 5049–5065. https://doi.org/10.1038/s41388-021-01868-5.

24. Subbalakshmi, A.R., Kundnani, D., Biswas, K., Ghosh, A., Hanash, S.M., Tripathi, S.C., and Jolly, M.K. (2020). Nfatc acts as a non-canonical phenotypic stability factor for a hybrid epithelial/mesenchymal phenotype. Front. Oncol. *10*, 553342. https://doi.org/10.3389/fonc.2020.553342.

25. Xu, J., Lamouille, S., and Derynck, R. (2009). Tgf-β-induced epithelial to mesenchymal transition. Cell Res. *19*, 156–172. https://doi.org/10.1038/cr.2009.5.

26. Jain, P., Bhatia, S., Thompson, E.W., and Jolly, M.K. (2022). Population dynamics of epithelial-mesenchymal heterogeneity in cancer cells. Biomolecules *12*, 348. https://doi.org/10.3390/biom12030348.

27. Tripathi, S., Chakraborty, P., Levine, H., and Jolly, M.K. (2020). A mechanism for epithelial-mesenchymal heterogeneity in a population of cancer cells. PLoS Comput. Biol. *16*, e1007619. https://doi.org/10.1371/journal.pcbi.1007619.

28. Roche, J. (2018). The epithelial-to-mesenchymal transition in cancer. Cancers *10*, 52. https://doi.org/10.3390/cancers10020052.

29. Vasaikar, S.V., Deshmukh, A.P., den Hollander, P., Addanki, S., Kuburich, N.A., Kudaravalli, S., Joseph, R., Chang, J.T., Soundararajan, R., and Mani, S.A. (2021). Emtome: a resource for pan-cancer analysis of epithelial-mesenchymal transition genes and signatures. Br. J. Cancer *124*, 259–269. https://doi.org/10.1038/s41416-020-01178-9.

30. George, J.T., Jolly, M.K., Xu, S., Somarelli, J.A., and Levine, H. (2017). Survival outcomes in cancer patients predicted by a partial emt

gene expression scoring metricpartial emt gene expression scoring metric. Cancer Res. *77*, 6415–6428. https://doi.org/10.1158/0008-5472.CAN-16-3521.

31. Tan, T.Z., Miow, Q.H., Miki, Y., Noda, T., Mori, S., Huang, R.Y.J., and Thiery, J.P. (2014). Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. EMBO Mol. Med. *6*, 1279–1293. https://doi.org/10.15252/emmm.201404208.

32. Byers, L.A., Diao, L., Wang, J., Saintigny, P., Girard, L., Peyton, M., Shen, L., Fan, Y., Giri, U., Tumula, P.K., et al. (2013). An epithelial–mesenchymal transition gene signature predicts resistance to egfr and pi3k inhibitors and identifies axl as a therapeutic target for overcoming egfr inhibitor resistanceemt predicts egfr and pi3k inhibitor resistance in nsclc. Clin. Cancer Res. *19*, 279–290. https://doi.org/10.1158/1078-0432.CCR-12-1558.

33. Chakraborty, P., George, J.T., Tripathi, S., Levine, H., and Jolly, M.K. (2020). Comparative study of transcriptomics-based scoring metrics for the epithelial-hybrid-mesenchymal spectrum. Front. Bioeng. Biotechnol. *8*, 220. https://doi.org/10.3389/fbioe.2020.00220.

34. Karlin, S., and Taylor, H.E. (1981). A Second Course in Stochastic Processes (Elsevier).

35. Zhao, X., Hu, J., Li, Y., and Guo, M. (2021). Volumetric compression develops noise-driven single-cell heterogeneity. Proc. Natl. Acad. Sci. USA *118*, e2110550118. https://doi.org/10.1073/pnas.2110550118.

36. Lei, X., Tian, W., Zhu, H., Chen, T., and Ao, P. (2015). Biological sources of intrinsic and extrinsic noise in ci expression of lysogenic phage lambda. Sci. Rep. *5*, 13597. https://doi.org/10.1038/srep13597.

37. Jia, W., Deshmukh, A., Mani, S.A., Jolly, M.K., and Levine, H. (2019). A possible role for epigenetic feedback regulation in the dynamics of the epithelial–mesenchymal transition (emt). Phys. Biol. *16*, 066004. https://doi.org/10.1088/1478-3975/ab34df.

38. Antón-García, P., Haghighi, E.B., Rose, K., Vladimirov, G., Boerries, M., and Hecht, A. (2023). Tgfβ1-induced emt in the mcf10a mammary epithelial cell line model is executed independently of snail1 and zeb1 but relies on junb-coordinated transcriptional regulation. Cancers *15*, 558. https://doi.org/10.3390/cancers15020558.

39. Koh, M., Woo, Y., Valiathan, R.R., Jung, H.Y., Park, S.Y., Kim, Y.N., Kim, H.R.C., Fridman, R., and Moon, A. (2015). Discoidin domain receptor 1 is a novel transcriptional target of zeb 1 in breast epithelial cells undergoing h-r as-induced epithelial to mesenchymal transition. Int. J. Cancer *136*, E508–E520. https://doi.org/10.1002/ijc.29154.

40. Zhang, J., Tian, X.J., Zhang, H., Teng, Y., Li, R., Bai, F., Elankumaran, S., and Xing, J. (2014). Tgf-β–induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. Sci. Signal. *7*, ra91. https://doi.org/10.1126/scisignal.2005304.

41. Watanabe, K., Panchy, N., Noguchi, S., Suzuki, H., and Hong, T. (2019). Combinatorial perturbation analysis reveals divergent regulations of mesenchymal genes during epithelial-to-mesenchymal transition. NPJ Syst. Biol. Appl. *5*, 21. https://doi.org/10.1038/s41540-019-0097-0.

42. Han, Y., Villarreal-Ponce, A., Gutierrez, G., Nguyen, Q., Sun, P., Wu, T., Sui, B., Berx, G., Brabletz, T., Kessenbrock, K., et al. (2022). Coordinate control of basal epithelial cell fate and stem cell maintenance by core emt transcription factor zeb1. Cell Rep. *38*, 110240. https://doi.org/10.1016/j.celrep.2021.110240.

43. Wagner, J., Masek, M., Jacobs, A., Soneson, C., Sivapatham, S., Damond, N., de Souza, N., Robinson, M.D., and Bodenmiller, B. (2022). Mass cytometric and transcriptomic profiling of epithelial-mesenchymal transitions in human mammary cell lines. Sci. Data *9*, 44. https://doi.org/10.1038/s41597-022-01137-4.

44. Deshmukh, A.P., Vasaikar, S.V., Tomczak, K., Tripathi, S., Den Hollander, P., Arslan, E., Chakraborty, P., Soundararajan, R., Jolly, M.K., Rai, K., et al. (2021). Identification of emt signaling cross-talk and gene regulatory networks by single-cell rna sequencing. Proc. Natl. Acad. Sci. USA *118*, e2102050118. https://doi.org/10.1073/pnas.2102050118.

45. Lagarias, J.C., Reeds, J.A., Wright, M.H., and Wright, P.E. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. SIAM J. Optim. *9*, 112–147. https://doi.org/10.1137/S1052623496303470.

46. Gillespie, D.T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comput. Phys. *22*, 403–434. https://doi.org/10.1016/0021-9991(76)90041-3.

47. Cook, D.P., and Vanderhyden, B.C. (2020). Context specificity of the emt transcriptional response. Nat. Commun. *11*, 2142–2149. https://doi.org/10.1038/s41467-020-16066-2.

48. Panchy, N., Watanabe, K., Takahashi, M., Willems, A., and Hong, T. (2022). Comparative single-cell transcriptomes of dose and time dependent epithelial–mesenchymal spectrums. NAR Genom. Bioinform. *4*, lqac072. https://doi.org/10.1093/nargab/lqac072.

49. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502. https://doi.org/10.1038/nbt.3192.

50. Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. Cell *174*, 716–729.e27. https://doi.org/10.1016/j.cell.2018.05.061.

51. McInnes, L., Healy, J., and Melville, J. (2018). Umap: uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.

52. Sahoo, S., Nayak, S.P., Hari, K., Purkait, P., Mandal, S., Kishore, A., Levine, H., and Jolly, M.K. (2021). Immunosuppressive traits of the hybrid epithelial/mesenchymal phenotype. Front. Immunol. *12*, 797261. https://doi.org/10.3389/fimmu.2021.797261.

53. Sundararajan, V., Tan, M., Zea Tan, T., Pang, Q.Y., Ye, J., Chung, V.Y., and Huang, R.Y.J. (2020). Snai1-driven sequential emt changes attributed by selective chromatin enrichment of rad21 and grhl2. Cancers *12*, 1140. https://doi.org/10.3390/cancers12051140.

54. Eugster, M.J.A., and Leisch, F. (2009). From spider-man to hero-archetypal analysis in r. J. Stat. Softw. *30*. https://doi.org/10.18637/jss.v030.i08.

55. Senin, P. (2008). Dynamic Time Warping Algorithm Review, *855* (Information and Computer Science Department University of Hawaii at Manoa Honolulu), p. 40.

56. Mendez, M.J., Hoffman, M.J., Cherry, E.M., Lemmon, C.A., and Weinberg, S.H. (2022). A data-assimilation approach to predict population dynamics during epithelial-mesenchymal transition. Biophys. J. *121*, 3061–3080. https://doi.org/10.1016/j.bpj.2022.07.014.

57. Chen, J., Li, Q., An, Y., Lv, N., Xue, X., Wei, J., Jiang, K., Wu, J., Gao, W., Qian, Z., et al. (2013). Ceacam6 induces epithelial-mesenchymal transition and mediates invasion and metastasis in pancreatic cancer. Int. J. Oncol. *43*, 877–885. https://doi.org/10.3892/ijo.2013.2015.

58. Xing, J., and Tian, X.J. (2019). Investigating epithelial-to-mesenchymal transition with integrated computational and experimental approaches. Phys. Biol. *16*, 031001. https://doi.org/10.1088/1478-3975/ab0032.

59. Kendall, M.G. (1938). A new measure of rank correlation. Biometrika *30*, 81–93. https://doi.org/10.2307/2332226.

60. Takahashi, K., Podyma-Inoue, K.A., Saito, M., Sakakitani, S., Sugauchi, A., Iida, K., Iwabuchi, S., Koinuma, D., Kurioka, K., Konishi, T., et al. (2022). Tgf-β generates a population of cancer cells residing in g1 phase with high motility and metastatic potential via krtap2-3. Cell Rep. *40*, 111411. https://doi.org/10.1016/j.celrep.2022.111411.

61. Comaills, V., Kabeche, L., Morris, R., Buisson, R., Yu, M., Madden, M.W., LiCausi, J.A., Boukhali, M., Tajima, K., Pan, S., et al. (2016). Genomic instability is induced by persistent proliferation of cells undergoing epithelial-to-mesenchymal transition. Cell Rep. *17*, 2632–2647. https://doi.org/10.1016/j.celrep.2016.11.022.

62. Vega, S., Morales, A.V., Ocaña, O.H., Valdés, F., Fabregat, I., and Nieto, M.A. (2004). Snail blocks the cell cycle and confers resistance to cell death. Genes Dev. *18*, 1131–1143. https://doi.org/10.1101/gad.294104.

63. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based

approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

64. Chang, C.J., Chao, C.H., Xia, W., Yang, J.Y., Xiong, Y., Li, C.W., Yu, W.H., Rehman, S.K., Hsu, J.L., Lee, H.H., et al. (2011). p53 regulates epithelial–mesenchymal transition and stem cell properties through modulating mirnas. Nat. Cell Biol. *13*, 317–323. https://doi.org/10.1038/ncb2173.

65. Lugano, R., Ramachandran, M., and Dimberg, A. (2020). Tumor angiogenesis: causes, consequences, challenges and opportunities. Cell. Mol. Life Sci. *77*, 1745–1770. https://doi.org/10.1007/s00018-019-03351-7.

66. Zavadil, J., Cermak, L., Soto-Nieves, N., and Böttinger, E.P. (2004). Integration of tgf-β/smad and jagged1/notch signalling in epithelial-to-mesenchymal transition. EMBO J. *23*, 1155–1165. https://doi.org/10.1038/sj.emboj.7600069.

67. Teng, Y., Zeisberg, M., and Kalluri, R. (2007). Transcriptional regulation of epithelial-mesenchymal transition. J. Clin. Invest. *117*, 304–306. https://doi.org/10.1172/JCI31200.

68. Kim, S.H., Ichikawa, K., Koshiishi, I., Utsumi, H., Chen, Y., Bittner, M., and Suh, E.B. (2002). Can Markov chain models mimic biological regulation? Water Sci. Technol. *46*, 337–341. https://doi.org/10.1142/S0218339002000676.

69. Jagannathan, N.S., Ihsan, M.O., Kin, X.X., Welsch, R.E., Clément, M.V., and Tucker-Kellogg, L. (2020). Transcompp: understanding phenotypic plasticity by estimating Markov transition rates for cell state transitions. Bioinformatics *36*, 2813–2820. https://doi.org/10.1093/bioinformatics/btaa021.

70. Gupta, P.B., Fillmore, C.M., Jiang, G., Shapira, S.D., Tao, K., Kuperwasser, C., and Lander, E.S. (2011). Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. Cell *146*, 633–644. https://doi.org/10.1016/j.cell.2011.07.026.

71. Su, Y., Wei, W., Robert, L., Xue, M., Tsoi, J., Garcia-Diaz, A., Homet Moreno, B., Kim, J., Ng, R.H., Lee, J.W., et al. (2017). Single-cell analysis resolves the cell state transition and signaling dynamics associated with melanoma drug-induced resistance. Proc. Natl. Acad. Sci. USA *114*, 13679–13684. https://doi.org/10.1073/pnas.1712064115.

72. Vipparthi, K., Hari, K., Chakraborty, P., Ghosh, S., Patel, A.K., Ghosh, A., Biswas, N.K., Sharan, R., Arun, P., Jolly, M.K., and Singh, S. (2022). Emergence of hybrid states of stem-like cancer cells correlates with poor prognosis in oral cancer. iScience *25*, 104317. https://doi.org/10.1016/j.isci.2022.104317.

73. Nugoli, M., Chuchana, P., Vendrell, J., Orsetti, B., Ursule, L., Nguyen, C., Birnbaum, D., Douzery, E.J.P., Cohen, P., and Theillet, C. (2003). Genetic variability in mcf-7 sublines: evidence of rapid genomic and rna expression profile modifications. BMC Cancer *3*, 1–12. https://doi.org/10.1186/1471-2407-3-13.

74. Axelrod, R., Axelrod, D.E., and Pienta, K.J. (2006). Evolution of cooperation among tumor cells. Proc. Natl. Acad. Sci. USA *103*, 13474–13479. https://doi.org/10.1073/pnas.0606053103.

75. Jolly, M.K., Mani, S.A., and Levine, H. (2018). Hybrid epithelial/mesenchymal phenotype (s): the 'fittest'for metastasis? Biochim. Biophys. Acta Rev. Canc *1870*, 151–157. https://doi.org/10.1016/j.bbcan.2018.07.001.

76. Goetz, H., Melendez-Alvarez, J.R., Chen, L., and Tian, X.J. (2020). A plausible accelerating function of intermediate states in cancer metastasis. PLoS Comput. Biol. *16*, e1007682. https://doi.org/10.1371/journal.pcbi.1007682.

77. Cook, D.P., and Wrana, J.L. (2022). A specialist-generalist framework for epithelial-mesenchymal plasticity in cancer. Trends Cancer *8*, 358–368. https://doi.org/10.1016/j.trecan.2022.01.014.

78. Joo, J.I., Zhou, J.X., Huang, S., and Cho, K.H. (2018). Determining relative dynamic stability of cell states using boolean network model. Sci. Rep. *8*, 12077. https://doi.org/10.1038/s41598-018-30544-0.

79. Li, R., Liang, J., Ni, S., Zhou, T., Qing, X., Li, H., He, W., Chen, J., Li, F., Zhuang, Q., et al. (2010). A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. Cell Stem Cell *7*, 51–63. https://doi.org/10.1016/j.stem.2010.04.014.

80. Samavarchi-Tehrani, P., Golipour, A., David, L., Sung, H.K., Beyer, T.A., Datti, A., Woltjen, K., Nagy, A., and Wrana, J.L. (2010). Functional genomics reveals a bmp-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. Cell Stem Cell *7*, 64–77. https://doi.org/10.1016/j.stem.2010.04.015.

81. Mitchel, J.A., Das, A., O'Sullivan, M.J., Stancil, I.T., DeCamp, S.J., Koehler, S., Ocaña, O.H., Butler, J.P., Fredberg, J.J., Nieto, M.A., et al. (2020). In primary airway epithelial cells, the unjamming transition is distinct from the epithelial-to-mesenchymal transition. Nat. Commun. *11*, 5053. https://doi.org/10.1038/s41467-020-18841-7.

82. Sergushichev, A., Korotkevich, G., Sukhov, V., and Artyomov, M. (2019). Fast gene set enrichment analysis. Preprint at bioRxiv. https://doi.org/10.1101/060012.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Analyzed dataset	Cook & Vanderhyden, 2020[47]	GEO:GSE147405
Analyzed dataset	Panchy et al., 2022[48]	GEO:GSE213753
Software and algorithms		
MAGIC	Van Dijk et al., 2018[50]	https://github.com/KrishnaswamyLab/MAGIC
Dynamic Time Warping (DTW) Alignment	Senin, 2008[55]	https://www.mathworks.com/help/signal/ref/dtw.html
Gene Set Enrichment Analysis (GSEA)	Subramanian et al., 2005[63]	https://www.gsea-msigdb.org/gsea/index.jsp
Seurat	Satija et al., 2015[49]	https://satijalab.org/seurat/
Gillespie's Algorithm	Gillespie, 1976[46]	–
Kendall Tau Statistic	Kendall, 1938[59]	https://finnstats.com/index.php/2021/06/10/kendalls-rank-correlation-in-r-correlation-test/

RESOURCE AVAILABILITY

Lead contact

For additional resources or inquiries please kindly contact the corresponding authors, Dr. Mohit K. Jolly (Mohit K. Jolly) and Dr. Jason T. George (Jason T. George).

Materials availability

This paper did not include newly generated reagents.

Data and code availability

- The data used in this paper are publicly available.
- All scripts used in the manuscript are uploaded and available at https://github.com/TAMUGeorgeGroup/Stochastic_EMT_2023.
- For further details needed to re-examine the data and methodology presented in this paper, please kindly request the authors for additional information.

METHOD DETAILS

Model development

We consider a population of fixed size consisting of N total cells divided into 3 sub-populations Epithelial (E), Hybrid (H), and Mesenchymal (M), where H is intermediary to E and M states. The dynamics of these transitions will be modeled as a continuous-time Markov process, so that the inter-arrival times for each event are exponentially distributed. Transition rates to (resp. from) the i^{th} state to the next step will be denoted as λ_i (resp. μ_i) for $i \in \{E, M\}$. In general, these transition rates are governed by the cellular environment and as such may vary due to the result of a particular signaling effect w, which we assume fixed for a given cell type, as well as the effects of cell spatial location of the EMT phenotype, so that $\lambda_i = \lambda_i(x, y, z; w)$, $\mu_i = \mu_i(x, y, z; w)$. In the results to follow we consider a time inhomogeneous process, which assumes that dynamics are described at some environmental state that is fixed in time. Toward this end, we let the state space consist of the ordered set E, H, M, $E < H < M$. Let $X(t)$ be the random variable denoting the particular state of the population at time t. Let $P(t)$ denote the transition matrix of the process and let $\pi(t)$ denote the row vector detailing the distribution of states at time t. We may express the evolution of this process via the generator matrix G so that

$$P(t) = e^{tG} = \sum_{j=0}^{\infty} \frac{(tG)^n}{n!}, \qquad \text{(Equation 8)}$$

where the infinitesimal generator matrix, G, is given by

$$G = \begin{pmatrix} -\mu_E & \mu_E & 0 \\ \lambda_E & -(\lambda_E+\lambda_M) & \lambda_M \\ 0 & \mu_M & -\mu_M \end{pmatrix}. \quad \text{(Equation 9)}$$

The Kolmogorov forward equation for this process is

$$\frac{dP}{dt} = P(t)G. \quad \text{(Equation 10)}$$

Which can be written as:

$$p'_{ij}(t) = \sum_{k \in S} p_{ik}(t) g_{kj} \quad \text{(Equation 11)}$$

for all i,j $\in S$

Let $\pi = (\pi_E, \pi_H, \pi_M)$ denote the stationary distribution of the process corresponding to

$$\pi P(t) = \pi; \pi G = 0. \quad \text{(Equation 12)}$$

Together with the probability constraint, Equation. (12), we can solve the following system of equations

$$-\mu_E\pi_E + \lambda_E\pi_H = 0, \lambda_M\pi_H - \mu_M\pi_M = 0, \pi_E + \pi_H + \pi_M = 1. \quad \text{(Equation 13)}$$

This system admits a unique solution for the stationary distribution, given by

$$\pi = (\pi_E \quad \pi_H \quad \pi_M) = \left(\frac{\varphi_E}{1+\varphi_E+\varphi_M} \quad \frac{1}{1+\varphi_E+\varphi_M} \quad \frac{\varphi_M}{1+\varphi_E+\varphi_M}\right), \text{where} \varphi_i \equiv \lambda_i \Big/ \mu_i. \quad \text{(Equation 14)}$$

The eigenvalues of the transition probability matrix can be found by solving the characteristic equation (Equation 15):

$$\det(G - \eta I) = 0. \quad \text{(Equation 15)}$$

By definition, $\eta_0 = 0$ is one root that corresponds to $v_0 = (1 \quad 1 \quad 1)^T$. The remaining are found to be:

$$\eta_{1,2} = \frac{1}{2}\left[-(\lambda_E + \lambda_M + \mu_E + \mu_M) \pm \sqrt{(\lambda_E+\lambda_M+\mu_E - \mu_M)^2 - 4\lambda_M(\mu_E - \mu_M)}\right]. \quad \text{(Equation 16)}$$

From Equation 16, it is clear that $\eta_1, \eta_2 < 0$, and each can be used to solve for a corresponding eigenvector.

Considering that PSFs stabilize the hybrid phenotype in a symmetric fashion, then we may take $\mu_E = \mu_M \equiv \mu$. In this case, the generator becomes

$$G = \begin{pmatrix} -\mu & \mu & 0 \\ \lambda_E & -(\lambda_E+\lambda_M) & \lambda_M \\ 0 & \mu & -\mu \end{pmatrix} \quad \text{(Equation 17)}$$

and Equation 16 simplifies considerably to

$$\eta_{1,2} = \frac{1}{2}\left[-(\lambda_E + \lambda_M + 2\mu) \pm \sqrt{(\lambda_E+\lambda_M)^2}\right], \quad \text{(Equation 18)}$$

giving

$$\eta_1 = -\mu; \eta_2 = -(\lambda_E + \lambda_M + \mu). \quad \text{(Equation 19)}$$

$$\eta_{1,2} = \frac{1}{2}\left[-(\lambda_E + \lambda_M + 2\mu) \pm \sqrt{(\lambda_E+\lambda_M)^2}\right], \quad \text{(Equation 20)}$$

The corresponding eigenvectors solve

$$\begin{pmatrix} 0 & \mu & 0 \\ \lambda_E & -(\lambda_E+\lambda_M-\mu) & \lambda_M \\ 0 & \mu & 0 \end{pmatrix} v_1 = 0; \begin{pmatrix} \lambda_E+\lambda_M & \mu & 0 \\ \lambda_E & \mu & \lambda_M \\ 0 & \mu & \lambda_E+\lambda_M \end{pmatrix} v_2 = 0. \quad \text{(Equation 21)}$$

Thus, we may take the diagonalization matrix, $Q = (v_0 \quad v_1 \quad v_2)$, as

$$Q = \begin{pmatrix} 1 & \lambda_M & -\mu \\ 1 & 0 & \lambda_E + \lambda_M \\ 1 & -\lambda_E & -\mu \end{pmatrix} \quad \text{(Equation 22)}$$

Note that $\nu \equiv \det Q$ may be calculated by expansion along the first row of Q:

$$\nu = = 1\begin{vmatrix} 0 & \lambda_E + \lambda_M \\ -\lambda_E & -\mu \end{vmatrix} - \lambda_M\begin{vmatrix} 1 & \lambda_E + \lambda_M \\ 1 & -\mu \end{vmatrix} + (-\mu)\begin{vmatrix} 1 & 0 \\ 1 & -\lambda_E \end{vmatrix} \quad \text{(Equation 23)}$$

$$= (\lambda_E + \lambda_M)(\lambda_E + \lambda_M + \mu) > 0. \quad \text{(Equation 24)}$$

Q is therefore invertible. Its inverse may be calculated as

$$Q^{-1} = \frac{1}{\nu}\begin{pmatrix} \lambda_E(\lambda_E + \lambda_M) & \mu(\lambda_E + \lambda_M) & \lambda_M(\lambda_E + \lambda_M) \\ \lambda_E + \lambda_M + \mu & 0 & -(\lambda_E + \lambda_M + \mu) \\ -\lambda_E & (\lambda_E + \lambda_M) & -\lambda_M \end{pmatrix}. \quad \text{(Equation 25)}$$

Q and Q^{-1} can be used to diagonalize the infinitesimal generator via:

$$G = Q\Lambda Q^{-1}, \quad \text{(Equation 26)}$$

where

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\mu & 0 \\ 0 & 0 & -(\lambda_E + \lambda_M + \mu) \end{pmatrix}. \quad \text{(Equation 27)}$$

Clearly,

$$G^n = Q\Lambda^n Q^{-1}. \quad \text{(Equation 28)}$$

Therefore, we have

$$P(t) = \sum_{j=0}^{\infty} \frac{t^n}{n!} Q\Lambda^n Q^{-1} = Qe^{t\Lambda}Q^{-1}. \quad \text{(Equation 29)}$$

Considering

$$k_1 = -\eta_1 = \mu \quad \text{(Equation 30)}$$

$$k_2 = -\eta_2 = \lambda_E + \lambda_M + \mu$$

$$k_3 = k_2 - k_1 = \lambda_E + \lambda_M,$$

Now if we evaluate Equation. (29), and express the result in terms of the stationary distribution from Equation. (14), we may express the exact distribution by

$$P = \begin{pmatrix} \pi_E + e^{-k_1 t}\left[1 - \frac{\lambda_E}{k_3}\left(1 - \pi_H e^{-k_3 t}\right)\right] & \pi_H\left(1 - e^{-k_2 t}\right) & \pi_M - e^{-k_1 t}\left[\pi_M + \frac{\lambda_M}{k_3}\pi_H\left(1 - e^{-k_3 t}\right)\right] \\ \pi_E\left(1 - e^{-k_2 t}\right) & \pi_H + e^{-k_2 t}(1 - \pi_H) & \pi_M\left(1 - e^{-k_2 t}\right) \\ \pi_E - e^{-k_1 t}\left[\pi_E + \frac{\lambda_E}{k_3}\pi_H\left(1 - e^{-k_3 t}\right)\right] & \pi_H\left(1 - e^{-k_2 t}\right) & \pi_M + e^{-k_1 t}\left[1 - \frac{\lambda_M}{k_3}\left(1 - \pi_H e^{-k_3 t}\right)\right]. \end{pmatrix}$$

(Equation 31)

12.2. Flow cytometry data

The flow cytometry data was acquired through a Z-Cad dual sensor which was inserted into an MCF10A cell line.[37] This sensor consisted of two components; a green fluorescent protein (GFP) reporter and a red fluorescent protein (RFP) reporter. While the former was regulated by ZEB, the latter was regulated through E-cad. Throughout the time-course experiment, a consistent cell density was maintained for each passage (5000 cells/cm ^{2}). We note that during the experiment, less proliferative cells (cell confluence 50%) had a

tendency of switching to a mesenchymal state. However, their contribution to the unexpected rise of cells with a *H* phenotype after day 18 was unclear. The fraction of cells in E, H, and M states were found by counting the number of RFP+GFP-, RFP+GFP+, and RFP-GFP+ cells.[37] Since the RFP-GFP- cells were excluded from counting, the three fractions were normalized for inferring E, H, and M trajectories.

Gillespie simulation

Gillespie simulations were performed in Matlab 2021. We simulated the arrival of four exponential random variables with parameters representing the transition rates of the CTMC. At discrete time steps, the system updated with the fastest arriving transition events. There were four transition events from *E* to *H*, *H* to *M*, *M* to *H*, and *H* to *E*. The rates of transitions were drawn from an exponential distribution. We ran simulations until sufficiently enough steps had passed (10000 iterations) and the system had reached steady state. For Figure 3D, the timescale was normalized and trimmed based on the case with the lowest timescale (Treatment withdrawal at day 9). Approximate number of iterations needed to recapitulate Figure 3D are: 98, 733, 4000, and 154.

Gene set enrichment analysis

GSEA was performed using the *fgsea* package[82] in R 4.2.1. Cells in every cluster were quality controlled and library size normalized. Wilcoxon test was then performed on genes and the resulting rankings were feeded to *fgsea*. All of the reference genes used in this analysis where from the MSigDB database. GSEA was then repeated 10 times for every cluster and the results were overlapped and attached to the results across all clusters in a circular fashion.

QUANTIFICATION AND STATISTICAL ANALYSIS

The Graphical abstract was created with Biorender. All other graphical illustrations were created with Affinity Designer. The mathematical model and simulation were performed using Matlab 2021. The data-driven pipeline, data processing, and data analysis were performed in R 4.2.1.