PNAS RESEARCH ARTICLE BIOPHYSICS AND COMPUTATIONAL BIOLOGY APPLIED MATHEMATICS

Reconstruction of single-cell lineage trajectories and identification of diversity in fates during the epithelial-to-mesenchymal transition

Yu-Chen Cheng[a,b,c,d], Yun Zhang[e], Shubham Tripathi[f], B. V. Harshavardhan[g], Mohit Kumar Jolly[h], Geoffrey Schiebinger[i], Herbert Levine[j,k,1], Thomas O. McDonald[a,b,c,d], and Franziska Michor[a,b,c,d,l,m,1]

Affiliations are included on p. 10.

Contributed by Herbert Levine; received April 10, 2024; accepted June 25, 2024; reviewed by Yibin Kang and Qing Nie

Exploring the complexity of the epithelial-to-mesenchymal transition (EMT) unveils a diversity of potential cell fates; however, the exact timing and mechanisms by which early cell states diverge into distinct EMT trajectories remain unclear. Studying these EMT trajectories through single-cell RNA sequencing is challenging due to the necessity of sacrificing cells for each measurement. In this study, we employed optimal-transport analysis to reconstruct the past trajectories of different cell fates during TGF-beta-induced EMT in the MCF10A cell line. Our analysis revealed three distinct trajectories leading to low EMT, partial EMT, and high EMT states. Cells along the partial EMT trajectory showed substantial variations in the EMT signature and exhibited pronounced stemness. Throughout this EMT trajectory, we observed a consistent downregulation of the *EED* and *EZH2* genes. This finding was validated by recent inhibitor screens of EMT regulators and CRISPR screen studies. Moreover, we applied our analysis of early-phase differential gene expression to gene sets associated with stemness and proliferation, pinpointing *ITGB4*, *LAMA3*, and *LAMB3* as genes differentially expressed in the initial stages of the partial versus high EMT trajectories. We also found that *CENPF*, *CKS1B*, and *MKI67* showed significant upregulation in the high EMT trajectory. While the first group of genes aligns with findings from previous studies, our work uniquely pinpoints the precise timing of these upregulations. Finally, the identification of the latter group of genes sheds light on potential cell cycle targets for modulating EMT trajectories.

EMT | cell fate | scRNA-seq

Significance

In our study, optimal-transport analysis was used to infer cell-to-cell connections from scRNAseq data, allowing us to predict cell linkages and overcome limitations of sequencing such as the need to sacrifice cells for each measurement. This approach led us to identify diverse EMT responses under uniform treatment, a significant advancement over previous studies limited by the static nature of scRNAseq data. Our analysis identified a broad set of genes involved in the EMT process, uncovering insights such as the upregulation of cell cycle genes in cells predisposed to a high EMT state and the enhancement of cell adhesion marker genes in cells veering toward a partial EMT state.

The epithelial–mesenchymal transition (EMT) is a pivotal process underpinning a range of biological phenomena from embryonic development and wound healing to tumor metastasis (1–5). During EMT, epithelial cells lose their apical–basal polarity and adhesion to other cells and acquire mesenchymal traits such as invasiveness and migratory capabilities (3–5). At the molecular level, this process is accompanied by the downregulation of epithelial markers such as E-cadherin (*CDH1*) and a concurrent upregulation of mesenchymal markers like N-cadherin (*CDH2*), vimentin (*VIM*), and fibronectin (*FN*) (6, 7). Importantly, EMT is not merely a binary transition from an epithelial (E) to a mesenchymal (M) state. Recent findings redefine EMT as a continuum, with cells capable of occupying intermediate states, often referred to as "partial" EMT (8, 9). Progression along this spectrum is tightly regulated by a set of key transcription factors, including members of the Snail, Zeb, and Twist families (10, 11). The expression and activities of these transcriptional factors are governed by a complex network of several epigenetic regulators and signaling pathways, encompassing TGF-beta, Wnt, EGF, FGF, PI3K/Akt/mTOR, IL-6/JAK/STAT3, and NOTCH (5, 12–16).

Cells in a syngeneic, phenotypically homogeneous population have been observed to adopt distinct fates upon treatment with an EMT inducer (17, 18). However, the intricate mechanisms that drive early cell states to branch into unique EMT trajectories are yet to be fully understood. The idea of divergent trajectories, through a developmental Waddington landscape (19), is well accepted in stem cell biology (20). Given the close association between EMT and stemness (21, 22), we aimed to investigate whether the heterogeneous response to EMT inducers extends beyond mere temporal variations and involves multiple distinct trajectories. To this end, we analyzed previously published time series scRNAseq data from MCF10A cells treated with TGF-beta (18) (Fig. 1*A*).

While scRNAseq data offer a wealth of insights into the heterogeneity of cellular states (23, 24), the inherent need to sacrifice cells at each time point precludes the ability

Reviewers: Y.K., Princeton University; and Q.N., University of California Irvine.

Competing interest statement: F.M. is a co-founder of and has equity in Harbinger Health, has equity in Zephyr AI, and serves as a consultant for both companies. She is also on the board of directors of Exscientia Plc. F.M. declares that none of these relationships are directly or indirectly related to the content of this manuscript. One of the authors (H.L.) was co-author on a review paper (in 2021) with one of the referees (Y.K.), and a different author (M.K.J.) was co-author on a different review paper (in 2023) with the other referee (Q.N.). All other authors declare no conflicts.

[1]To whom correspondence may be addressed. Email: h.levine@northeastern.edu or michor@jimmy.harvard.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2406842121/-/DCSupplemental.

Published August 2, 2024.

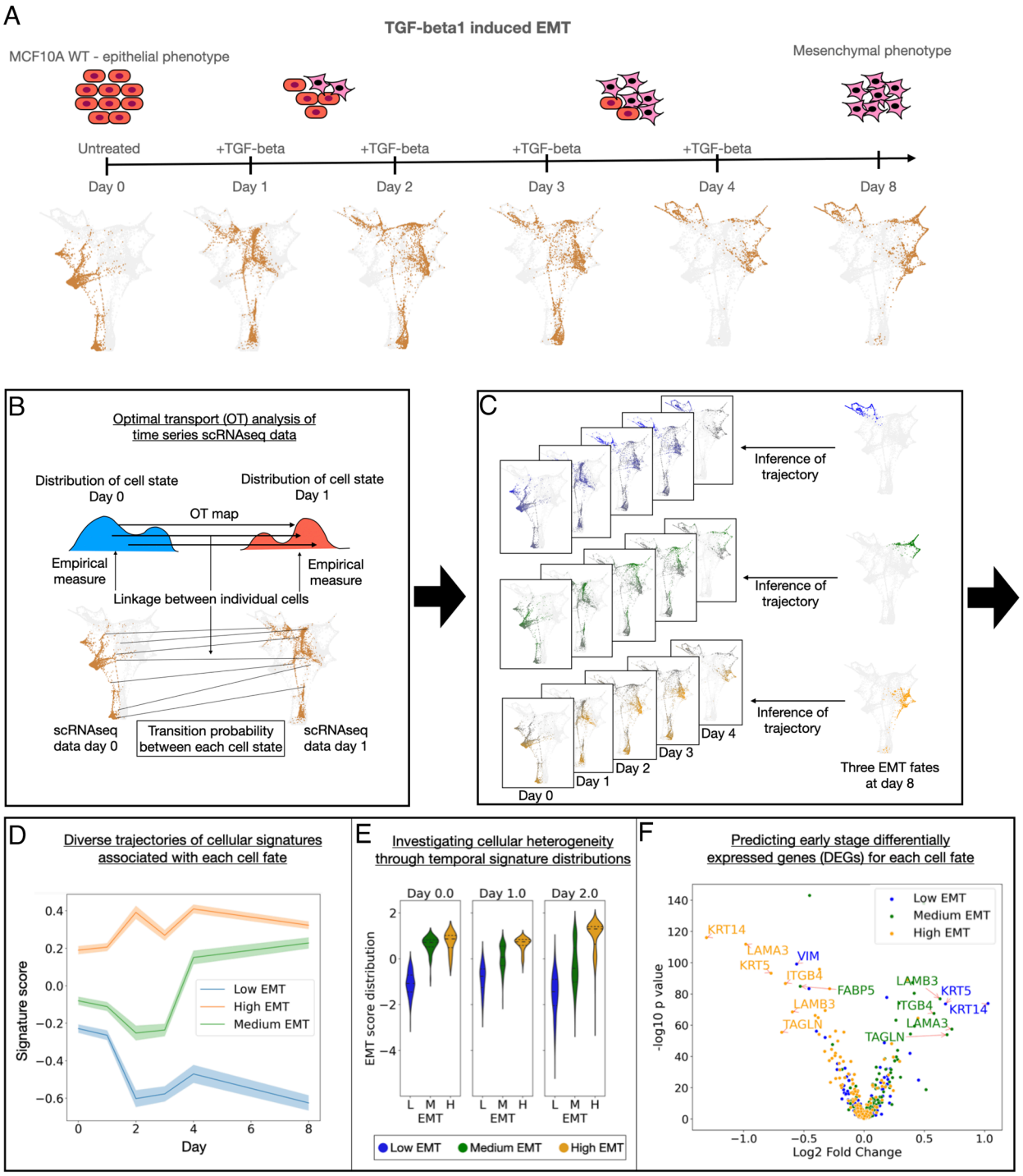

Fig. 1. Optimal transport analysis of scRNAseq time course data of the EMT. (*A*) Schematic representation of the experiments performed to induce EMT in MCF10A cells, accompanied by a time series of scRNA-seq data visualized using force-directed layout embedding (FLE). Cells are depicted as gray dots, with brown dots highlighting the cells for each day. This figure was adapted from Deshmukh et al. (17) (*B–F*) General framework of optimal-transport analysis of EMT single-cell RNA sequencing data: (*B*) OT analysis was employed to identify the transition probability of cell-to-cell connections. (*C*) From the entire consecutive time-series scRNAseq data, transition probabilities were integrated to determine the likelihood of each early cell state acting as an ancestor for the three fate subpopulations. (*D–F*) Three downstream analyses: (*D*) reconstruction of diverse cellular signature trajectories, (*E*) exploration of cellular heterogeneity across these trajectories, and (*F*) differential analysis of early gene expression in ancestral cells associated with each distinct cell fate.

to trace individual cell lineages over time. This restriction poses a challenge to reconstructing trajectories from time-series scRNAseq data (25). To address this challenge, we employed a method based on OT analysis (26, 27), known as Waddington OT (WOT) (28). This method stands in contrast to other widely used trajectory tools such as pseudotime analysis, which infers a temporal sequence within a cell population but cannot deduce direct cell-to-cell transitions (29, 30). Another method, RNA velocity, utilizes additional information from unspliced and spliced RNA to predict the direction of movement across RNAseq space of individual cells (31, 32). This method deepens our insight into the velocity field and short-term cellular changes.

However, the applications of the RNA velocity method have sometimes been found to lack precision and can yield ambiguous results, particularly due to assumptions of constant kinetic rate parameters (33). To address these challenges, Qiu et al. developed,

a method that precisely infers the vector field from time-resolved, metabolically labeled scRNA-seq data (34). In contrast, our study utilizes conventional, daily-collected scRNA-seq data. We employ WOT specifically for its ability to analyze direct cell-to-cell transitions within scRNA-seq data at discrete, predetermined time points. This approach avoids the complexities and potential noise associated with velocity field inference, ensuring that our analysis remains precise and directly interpretable.

Utilizing the WOT technique, we reconstructed lineage trajectories at single-cell resolution using the time series scRNAseq data from MCF10A cells undergoing EMT stimulated by TGF-beta (18), enabling identification of diverse trajectories leading to distinct EMT fates. In this study, we extend previous EMT research by not only examining state heterogeneity at various time points within a single EMT process but also by uncovering the diversity of EMT responses as unique, distinct processes under the same treatment. We delved into the roles of stemness, proliferation, and cellular hypoxic response signatures. While these signatures have known associations with EMT (5, 21, 35), their variations across different EMT trajectories have not been extensively explored.

Furthermore, our trajectory analysis at the single gene level enabled us to predict differentially expressed genes (DEGs) in the early phases of each fate. Early gene expression changes linked to a specific fate were then partially validated through methods such as inhibitor screens of EMT regulators and CRISPR-associated gene knockout screens (1, 15, 16, 36), highlighting the robustness of our predictions. We then included a wider array of genes implicated in EMT regulation but not yet fully examined. This approach led to several insights, notably that cell cycle-related genes are up-regulated in the ancestors of cells entering the high EMT state. Additionally, we found that genes linked to cell surface markers that play a critical role in cell–matrix and cell–cell adhesion are markedly up-regulated in the ancestors of cells transitioning to the partial EMT state. An overview of the general framework is provided in Fig. 1 *B*–*F*.

Results

Uncovering Three Distinct EMT Trajectories via Optimal Transport Analysis. Given that scRNAseq data cannot be obtained from individual cells at multiple time points of lineage tracing experiments, due to the assay's destructive nature, we set out to computationally infer likely ancestor cell states for different EMT fates. In the study by Deshmukh et al. (18), an immortalized human mammary epithelial cell line, MCF10A, was treated with TGF-beta for 1, 2, 3, 4, or 8 d (Fig. 1*A*), and scRNAseq data were obtained from populations sacrificed at each time point. Through cluster analysis of the scRNAseq data at day 8, we identified three subpopulations representing three significantly different cell fates (*Materials and Methods*). These fates were categorized as low, medium, and high EMT by utilizing the 76GS and KS scoring metrics to compute the average EMT scores (37–40), for each subpopulation (Fig. 2*A*, day 8). For instance, using the 76GS method, we derived average EMT scores of –0.63, 0.23, and 0.32 for the low, medium, and high EMT categories, respectively, with

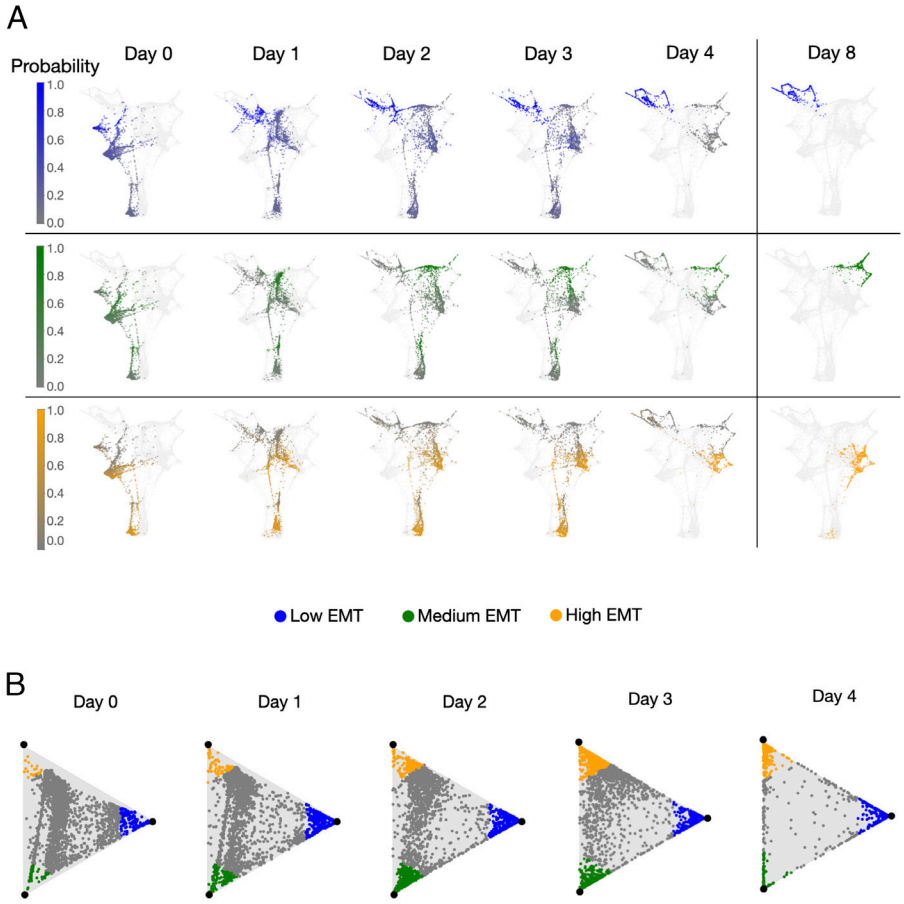

Fig. 2. Optimal transport recovers diverse trajectories of EMT. (*A*) The colormap presents the inferred ATF distributions, showcasing the probability of early cell states (from day 0 to day 4) serving as ancestors for the three fate subpopulations by day 8. (*B*) Barycentric coordinate projection visualizes ATF distributions. For each time point, every individual cell is associated with a three-dimensional probability vector, as determined by that specific time point's ATF distributions of the three fates (each column in *A*). This vector is then mapped onto an equilateral triangle (*SI Appendix*, S3). A position at one of the triangle's vertices indicates a 100% commitment of the cell state to the corresponding fate.

significant p-values (*t*-test, $P < 0.005$) for each pairwise comparison (*SI Appendix*, Table S1, day 8). Use of the KS method yielded consistent results (*SI Appendix*, Table S2, day 8).

To infer the trajectories of individual cell states across the sequential scRNAseq dataset, we utilized WOT (28). The dataset consists of six distinct batches, each sourced from a uniformly mixed single culture of around 10,000 cells. This setup provides a consistent starting point for each batch before the application of TGF-beta, allowing us to assume uniform initial conditions across the batches. Leveraging this baseline, the WOT method predicts a unique transition probability (i.e., the likelihood that one cellular state is the ancestor and the other the descendant) between two adjacent scRNAseq time points. The WOT approach assumes that cellular states navigate the gene expression space using the shortest overall distance (*SI Appendix*, Fig. S1 and *Materials and Methods*). By multiplying the inferred transition probabilities from initial to subsequent time points within our scRNAseq data series, we computed the probability of each early cell state, termed "ancestors", transitioning into a final cell state at day 8 or "fate" (Fig. 2*A*). We refer to these transition probabilities as "ancestor-to-fate (ATF) distributions." To validate our inferred distributions, we followed an approach of omitting data of a specific time point, designated as test data, and comparing our estimated cell state distribution to the actual data of this time point. The results showed minimal deviations between predictions and actual data, confirming our predictions' reliability when contrasted with other intrinsic cellular variations and unbiased interpolations (*SI Appendix*, S1). Note that in the main text, both the inference and the validation of ATF distributions were confined to the first 30 PCA dimensions of the gene expression space, as validated in the original WOT paper to accurately predict cell states in the test data at held-out time points (37). Additionally, to broaden our analysis, we expanded the dimensional range up to 3,000 and repeated our analysis for comparative purposes. Our results demonstrated consistency in all main conclusions of the inferred ATF distributions across various dimensionalities (*SI Appendix*, S2).

To identify cellular origins leading to various fates, we categorized cells with over 75% probability of transitioning to specific fates as "top ancestors" (*SI Appendix*, S3). Notably, prior to treatment, the percentage of top ancestors for the low EMT fate constituted double the combined percentage of the other two fates (5.52% vs. 2.85% at day 0). By the second day of treatment, the proportions of top ancestors across all three fates converged, with values of 13.57%, 12.35%, and 13.31% for low, medium, and high EMT (Fig. 2*B* and *SI Appendix*, Fig. S2 and Table S3). This temporal shift in proportions indicates a delayed inclination toward the medium and high EMT fates, induced by TGF-beta. Additionally, cells falling below the probability threshold for any EMT fate were classified as "undetermined ancestors." With ongoing TGF-beta treatment, the portion of the undetermined ancestors decreased sharply from 91.62% on day 0 to 15.95% on day 4 (*SI Appendix*, Fig. S2 and Table S3). This trend may suggest that initially, a high percentage of undetermined ancestors indicates a high level of cell plasticity before treatment; however, following treatment initiation, this plasticity might reduce as more cells advance toward predetermined fates. Consistently, these interpretations are supported across alternative probability thresholds for defining top ancestors, ranging from 75% to 90% (*SI Appendix*, S4).

Upon inspection of the full trajectories, we observed that the ancestors of the three fates were dispersed without clear boundaries, unlike the three distinct, well-outlined regions for the three fates seen at day 8 (Fig. 2*A* and *SI Appendix*, Fig. S3). This observation, combined with the profound reduction in the percentage of undetermined ancestors posttreatment (75% decrease, Fig. 2*B*) suggests that over time, cells exhibit decreased plasticity and increasingly tend toward more determined states. This trend indicates a divergence in EMT phenotypes. To quantify this divergence, we computed the total variation-distance (41) between the cell state distributions of each pair of trajectories at every time point (*Materials and Methods*). Our analysis revealed a marked divergence between every pair among the three trajectories: by day 8, the distance had increased 2.67 times from its day 0 measurement for both the low vs. medium and low vs. high EMT trajectories, and 2.17 times for the high vs. medium EMT trajectory. Notably, this divergence was most pronounced before day 4, accounting for 90% of the total increase. The divergence then leveled off, with only a 10% increase observed afterward, indicating that the divergence between trajectories increases most significantly during early stages of TGF-beta treatment (*SI Appendix*, Fig. S4).

Deciphering Unique Gene Signatures Across Trajectories of Distinct EMT States. To trace the EMT characteristics of the three fate subpopulations to their origins, we first determined the EMT score for each cellular state, from day 0 to day 8, using the 76GS and KS EMT scoring methods (Fig. 3*A* and *Materials and Methods*). For each time point, we integrated the EMT scores across all cell states, each weighted by their likelihood of being the ancestor for a particular fate subpopulation as determined by the ATF distributions (*SI Appendix*, S5 and Fig. S5). This approach unveiled three distinct trajectories, each showing unique average EMT score trajectories with nonoverlapping 95% CI, throughout the course of TGF-beta treatment (Fig. 3*B* and *SI Appendix*, Fig. S6*A*). The clear separation into low, medium, and high EMT trajectories was consistently observed using both the 76GS and KS EMT scoring methods (*SI Appendix*, Fig. S6*A* and *Materials and Methods*).

Notably, the separation of trajectories was observed even before the initiation of TGF-beta treatment on day 0, implying that early EMT hallmarks could predestine cellular EMT fates (Fig. 3*B*). In light of this finding, we limited our WOT analysis to the gene expression space encompassing genes associated with the EMT signaling pathway (42, 43), and repeated the computation of ATF distributions and gene signature dynamics across the three trajectories. We found that the three trajectories remained profoundly divergent, similar to the ATF computations using the full gene space. However, when examining early time point (before day 3), we observed that the ancestral cell populations were less separable when analyzed using the EMT gene set compared to the full gene set (*SI Appendix*, S6).

Furthermore, we analyzed stemness, hypoxia response, and proliferation signatures among cells belonging to the three fates. For each cell, we computed those signatures using single-sample gene set enrichment analysis, ssGSEA (Fig. 3*A* and *Materials and Methods*). All trajectories showed an over 1.9 z-score increase in both stemness and hypoxia response (Fig. 3*B* and *SI Appendix*, Fig. S6*A*). This trend aligns with prior research that links hypoxia to enhanced stemness in EMT (44, 45). Like the EMT signature, these three trajectories stood out with their nonoverlapping 95% CI when characterized by these two signatures (Fig. 3*B* and *SI Appendix*, Fig. S6*A*). Of particular interest was that by day 8, the medium EMT trajectory exhibited the highest levels of stemness and hypoxia response, with enrichment scores of 1.7 for both. In comparison, the low and high EMT trajectories displayed scores of 1.4 and 1.2, respectively (Fig. 3*B* and *SI Appendix*, Fig. S6*A*). These findings resonate with earlier studies identifying an intermediate EMT stage characterized by heightened stemness and a pronounced response to hypoxia (4, 35, 46).

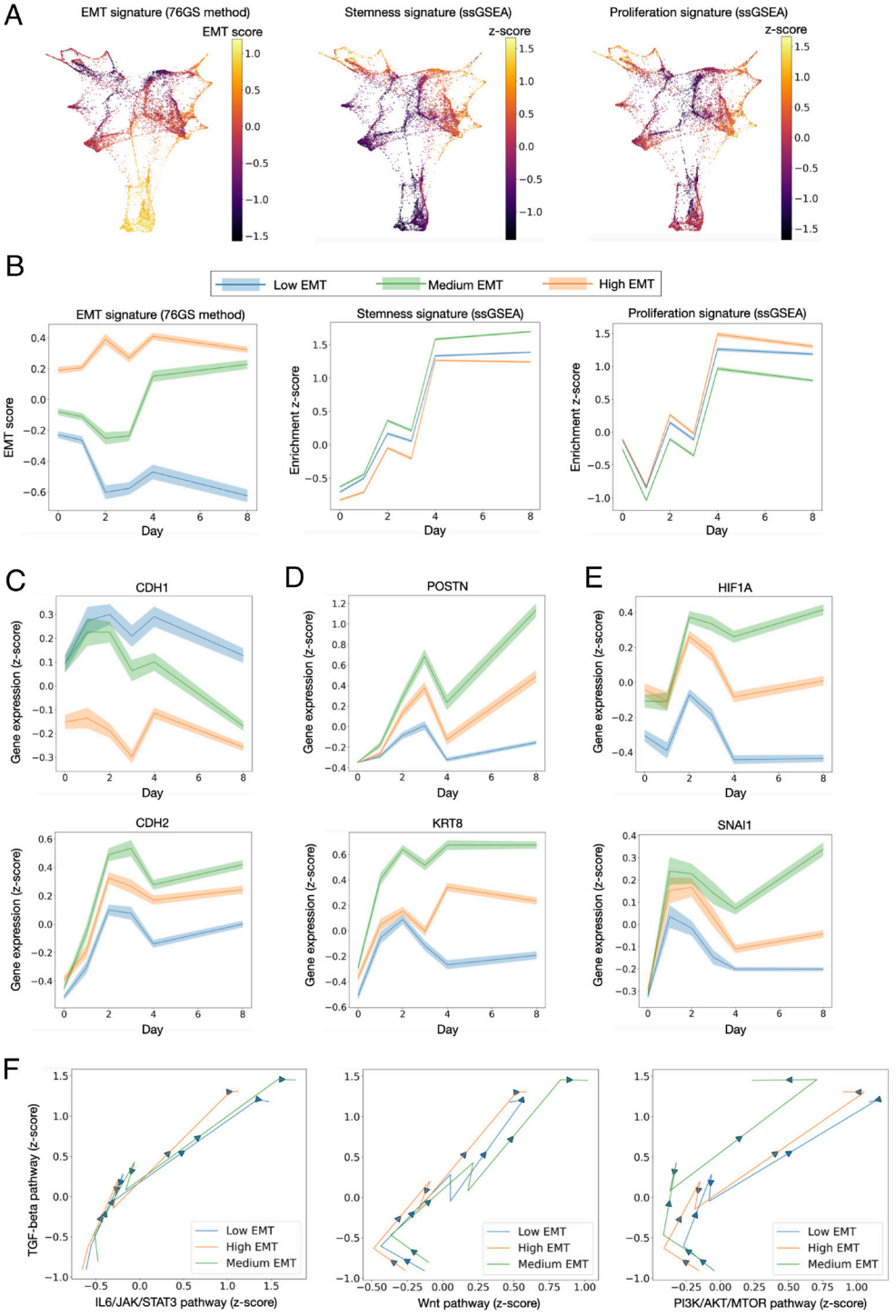

Fig. 3. OT analysis reveals unique cellular signatures across distinct EMT trajectories. (*A*) Color maps illustrate the EMT signature score (using the 76GS method), stemness signature score (via ssGSEA), and proliferation signature score (via ssGSEA) for all cellular states gathered from day 0 to day 8. (*B*) The panels depict the time progression of average cellular signature scores (left to right: EMT, stemness, and proliferation) across the three distinct EMT trajectories. Shaded regions denote the 95% CI. (*C*–*E*) Temporal evolution of mean gene expression across the three EMT trajectories. Shaded regions denote the 95% CI (*C*) for *CDH1* and *CDH2* genes, (*D*) for *POSTN* and *KRT8* genes, and (*E*) for *HIF1A* and *SNAI1* genes. (*F*) Two-dimensional plots illustrate the time-course progression of average cellular signature scores for paired signaling pathways. Lines connect daily average scores for each signature pair, with arrows highlighting the directional flow of time.

Upon analyzing the proliferation signature trajectories, we noted enrichment z-score declines from day 0 to day 1 (low EMT: −0.12 to −0.84, medium EMT: −0.26 to −1.04, high EMT: −0.11 to −0.8, Fig. 3*B*). A similar trend was observed in the G2M checkpoint and mitotic spindle hallmarks (*SI Appendix*, Fig. S6*A*). This decrease reflects the known role of TGF-beta in inhibiting cell division (47–49). From day 1 through day 8, cells regain their proliferative capacity, evidenced by enrichment score of proliferation rebounds of 2.0 for low, 1.8 for medium, and 2.1 for high EMT (Fig. 3*B* and *SI Appendix*, Fig. S6*A*). Based on these changes, we concluded that the medium EMT trajectory was distinctive, exhibiting the most pronounced decline and the least recovery in proliferation signatures. This unique trend in the medium EMT cells corresponds with their pronounced response to the TGF-beta inducer, evident by TGFBI (a TGF-beta-induced gene) showing more elevated expression in this trajectory than in the other two (*SI Appendix*, Fig. S7).

We then further analyzed the dynamics of individual genes pivotal to EMT, such as *CDH1* and *CDH2*. We found that the medium EMT trajectory initially displays high *CDH1* expression that diminished toward the end of treatment, shifting from a z-score of 0.09 on day 0 to −0.17 on day 8 (Fig. 3*C*). This downward trend aligns with previous findings indicating that *CDH1* downregulation triggers partial EMT (50). Conversely, *CDH2* expression notably increased in the medium EMT trajectory, diverging from the patterns seen in the high and low EMT trajectories (Fig. 3*C*). Furthermore, these gene expression patterns are markedly distinct, underscored by their nonoverlapping 95% CI. This finding aligns with a previous study showing elevated expression of *CDH2* in partial EMT using the same cell line and treatment type (36). Beyond *CDH2*, Zhang et al. (36) highlighted elevated expression of *POSTN* and *KRT8* expressions as indicators of the partial EMT phase. In our study, the medium EMT trajectory mirrored this finding, with *POSTN* and *KRT8* expression levels surpassing those in the high and low trajectories (Fig. 3*D*). Additionally, we detected a pronounced rise in *HIF-1A* and *Snail* expression within the medium trajectory compared to the others (Fig. 3*E*). This finding further supports the classification of the medium EMT as partial EMT, given the known roles of these genes in hypoxia and partial EMT fates (35, 51).

To investigate whether TGF-beta treatment correlates with other essential EMT-related signaling pathways, we further conducted pairwise comparisons of various cellular signatures over time (*SI Appendix*, Fig. S8 and *Materials and Methods*). Across all trajectories, we found positive correlations between TGF-beta signaling and the IL6-JAK-STAT3, Wnt, and PI3K-AKT-mTOR pathways, with Pearson correlation coefficients ranging across trajectories from 0.95 to 0.97, 0.92 to 0.96, and 0.76 to 0.92, respectively (Fig. 3*F*). These observations are consistent with previous findings regarding the concurrent regulation of these pathways throughout the EMT process (9, 52, 53). Particularly, the intricate interplay between TGF-beta and PI3K signaling pathways, which includes both antagonistic and cooperative interactions, has been discussed previously (9). In our study, while the TGF-beta pathway activity increased from day 0 to day 8 across all three EMT trajectories, the PI3K pathway interestingly showed a decline in the partial EMT trajectory by the end of treatment. In contrast, the enrichment scores for the other two trajectories remained relatively stable (Fig. 3*F*). With PI3K signaling recognized as a prominent driver of cell growth and proliferation (54), this observed decline aligns with the lower proliferation scores and G2M checkpoint pathway activity levels noted along the partial EMT trajectory (Fig. 3*B* and *SI Appendix*, Fig. S6*A*).

Note that the signature trajectories calculated in this section represent only the mean scores for cells on a specific path. The nonoverlapping CI clearly confirm the distinct separations of these mean dynamics. Indeed, variations in these signature scores exist within the entire cell population, as illustrated in *SI Appendix*, Fig. S6*B*. In the next section, we further expand our analysis to encompass the full distribution of scores. Furthermore, to explore the potential variations in lineage trajectories across different EMT models, we applied the WOT method to an additional dataset (17). Our findings confirm that variations in lineage trajectories indeed exist across different cell lines, even under the same EMT inducer, TGF-beta (*SI Appendix*, S7).

Unveiling Increased EMT Heterogeneity Within the Partial EMT Trajectory. To deepen our understanding of cellular heterogeneity across EMT trajectories, we studied the temporal evolution of EMT signature distributions along the three identified paths. We employed several methodologies to evaluate the within-trajectory distributions. First, we incorporated chronological sequences of triangle plots (Fig. 2*B*) with time-ordered individual cellular EMT signature scores (Fig. 4 *A* and *B* and *SI Appendix*, Fig. S9). This integration elucidated the relationship between ancestral cell EMT states and their potential to transition into a specific fate (*Materials and Methods*). The triangle plots demonstrate that the top ancestors, showing over 75% commitment to the high/low EMT fate, consistently exhibited high/low EMT signatures during the initial stages of the treatment process (Fig. 4*B*). Conversely, for the top ancestors of the partial EMT fate, EMT scores were notably heterogeneous, encompassing the full spectrum from low to high EMT cell types (Fig. 4*B* *SI Appendix*, Fig. S9). This pattern is discernible throughout days 0-3 (Fig. 4*B* *SI Appendix*, Fig. S9), suggesting that this early phase of the partial EMT trajectory displays a greater degree of variability in EMT expression scores compared to the early phases of the other EMT trajectories.

To explore the variability within the three EMT trajectories, we assessed the distributions of EMT, stemness, proliferation, and hypoxia scores among the top ancestors of the three identified EMT fates (*SI Appendix*, S8). We found that one distinguishing feature of the partial EMT trajectory was its broad variation in the EMT signature, paired with consistent stemness, proliferation, and hypoxia signatures (Fig. 4 *C* and *D* and *SI Appendix*, Fig. S10). We used Levene's test for equality of variances to determine whether any population had a significantly different variance from the others. For instance, the top ancestors of the partial EMT trajectory exhibited a more pronounced variance in the EMT scores compared to those of the high EMT trajectory (Levene's test, *P*-value < 1e-10 in days 1 to 8, Fig. 4*C* and *SI Appendix*, Table S1). In contrast, the partial EMT trajectory exhibited a stemness score variance similar to, or even less than, that of the high EMT trajectory (Levene's test, $P > 0.05$ on days 1, 4, and 8). On days when significant differences did occur (Levene's test, $P < 1e\text{-}5$ on days 2 and 3), the variances were more significant in the high than the partial EMT trajectory (Fig. 4*D* and *SI Appendix*, Table S4).

To further characterize the extent of heterogeneity within the partial EMT trajectory, we calculated pairwise cell state distances (55) (*SI Appendix*, S9), focusing on the differences between the partial and high EMT trajectories. The low EMT trajectory was excluded due to the high number of outliers (for details, see *SI Appendix*, Table S5). To determine whether the high variability was uniquely tied to the EMT signature, we computed cell state distances across three gene expression spaces: the full gene set, the EMT signature gene set, and genes differentially expressed between the partial and high EMT fates (*SI Appendix*, S10 and Table S6). Our findings reveal that, within the EMT gene expression space, variability in cell states was substantially greater in the partial EMT trajectory compared to the high EMT trajectory, as supported by statistically significant differences (*t* test, $P < 1e\text{-}9$) with fold changes of 1.11, 1.09, 1.10, and 1.09 for days 1 through 4, respectively (Fig. 4*E* and *SI Appendix*, Table S7). Conversely, during this period, these differences were not significant when cell state heterogeneity was analyzed using either the full gene set or the DEG set (Fig. 4*E* and *SI Appendix*, Table S7). This specific variability of the EMT score in the partial EMT trajectory aligns with prior research suggesting a lack of association between core EMT transcription factors and the partial EMT state (56).

To further explore the interplay between EMT and stemness signatures, we examined the joint distributions of these signatures at various time points (*SI Appendix*, S8). Our analysis revealed that the three trajectories during days 2 to 8 occupied different regions in the two-dimensional EMT and stemness score space. Specifically, the EMT signature predominantly distinguished

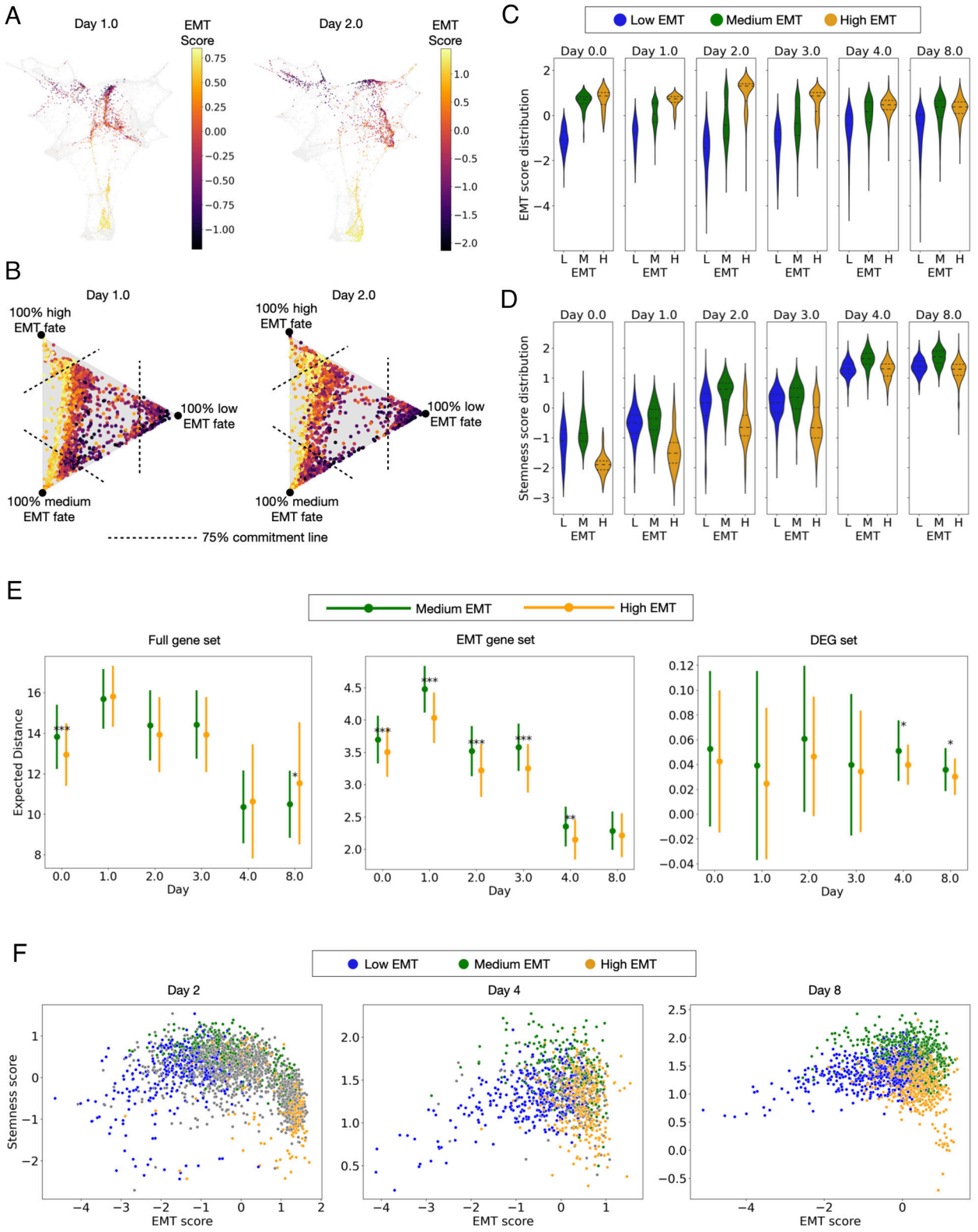

Fig. 4. Tracing cellular signature variations across three EMT trajectories. (*A*) EMT signature scores for cell states from days 1 and 2 (for the complete time course see *SI Appendix*, Fig. S6). (*B*) EMT signature scores from (*A*) are paired with ATF distributions and plotted within a triangle using barycentric coordinates. As in Fig. 2*C*, a point's location represents its ATF distribution. Concurrently, the color map showcases the EMT score. Dashed lines demarcate a 75% commitment to the fate linked to the corresponding triangle vertex. (*C* and *D*) Violin plots depict the distribution of each cellular signature score for the top ancestors of each fate: (*C*) for EMT score (via 76GS method) and (*D*) for stemness score (via ssGSEA). (*E*) The error bar plots depict the mean of weighted pairwise distances in cellular transcriptomics (indicated at the center of each bar), and the SD errors of these pairwise distances (symbolized by the length of the error bars). Significance levels are denoted by asterisks: one star for α = 1e-4, two stars for α = 1e-8, and three stars for α = 1e-12. (*F*) Scatter plots display paired cellular signature scores for days 2, 4, and 8. Color codes designate the top ancestors for each trajectory.

between the low and high EMT trajectories, whereas a pronounced stemness signature demarcated the partial EMT trajectory from the other two (Fig. 4*F*). Additionally, within the low EMT subset, a consistent positive correlation between EMT and stemness signatures was observed from days 1 to 8 (Pearson coefficients ranging from 0.22 to 0.44). In contrast, the high EMT subset presented a negative correlation between EMT and stemness signatures (Pearson coefficients ranging from −0.3 to −0.6) (*SI Appendix*, Table S8). This analysis reveals that cells with marked EMT signatures, whether extremely low or high, display reduced stemness. This trend is in line with earlier research suggesting that cells moving toward a distinctly differentiated state, whether closer to a pure E or M state along the EMT continuum, tend to exhibit less stemness (22).

Leveraging CRISPR Screening for Validation of Key Early Predicted Genes in EMT. To validate our identified trajectories, we compared our findings with a recent study that reported a substantial induction of the partial EMT fate following TGF-beta treatment in a background of PRC2 dysfunction, which was conducted across various epithelial cell lines including HMLER and MCF10A cells (36). The 2D gene expression maps showed that the levels of *EED* and *EZH2*—key constituents of PRC2—were notably diminished in areas aligning with the high-probability regions for the partial EMT trajectory (Fig. 5*A* and *SI Appendix*, Fig. S11*A*). To validate this observation, we quantified the expressions of these genes across the three trajectories. Both genes exhibited distinct average expression trends, each distinctly demarcated by nonoverlapping 95% CI. Importantly, there was a noticeable decline in *EED* and *EZH2* expressions, predominantly within the partial EMT trajectory (Fig. 5*B* and *SI Appendix*, Fig. S11*B*). Concurrently, within the top ancestors of the partial EMT, there was a discernible contraction in the distribution of *EED* expression, marked by a decrease in the number of cells exhibiting high gene expression, which is evidenced by a shift in the mean of the distribution (*SI Appendix*, Fig. S11*C* and Table S9). Similar patterns were observed for the *EZH2* gene (*SI Appendix*, Fig. S11*C* and Table S10).

In line with these findings, the CRISPR screen study revealed that knocking out *EED* and *EZH2* promotes a partial EMT state with increased stemness (36). This study was performed using the HMLER cell line, which, like MCF10A, is an immortalized human mammary epithelial cell line and exhibits similar changes in gene expression during TGF-beta-induced EMT as the MCF10A cell line (36, 57). Therefore, we curated an EMT-related gene list from both the time course data (18) and the CRISPR screen study (36) (*Materials and Methods*). Two mesenchymal states were identified in the CRISPR study: C1-sgEED-Mes (partial EMT with *EED* gene knockout) retained some epithelial traits, while C1-sgKMT2D-Mes (high EMT with *KMT2D* knockout) lacked them. We then examined the differential expression of the curated gene set between the partial and high EMT fates in our dataset on day 8, and between the C1-sgEED-Mes and C1-sgKMT2D-Mes cells in the CRISPR screen data (*SI Appendix*, S10 and Fig. S12). Remarkably, three out of the top four ranked genes—*TGFBI*, *POSTN*, and *KRT8*—were significantly up-regulated in the

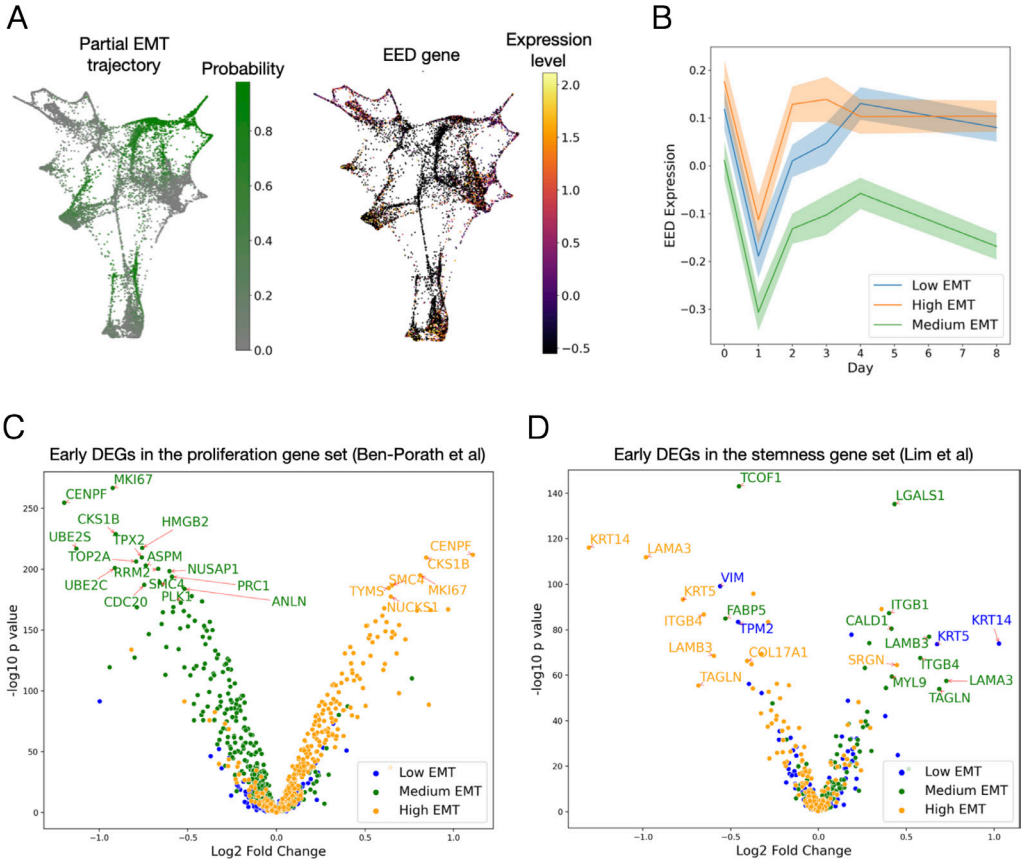

Fig. 5. Early predictors of EMT fate through early DEG analysis and CRISPR knock-out screening. (*A* and *B*) *EED* gene expression analysis: (*A*) color maps display the ATF distributions for the partial EMT trajectory alongside the expression levels of the *EED* gene across all cellular states from day 0 to day 8. In the trajectory map, the color gradient signifies probability, while in the gene expression map, it indicates the level of gene expression. (*B*) Line plots illustrate the average dynamics of *EED* gene expression over. Shaded regions denote the 95% CI. (*C* and *D*) Early DEGs of proliferation-related genes (*C*) and stemness-related genes (*D*). Distinct color codes showcase the differential gene expressions in cell states from a specific trajectory when contrasted with the combined cell states of the remaining two trajectories.

partial EMT state in both datasets (*t* test, $P < 1e\text{-}10$; fold changes in *SI Appendix*, Table S11). This analysis further supports our characterization of the partial and high EMT fates within our dataset.

We then evaluated early differential gene expression patterns between each pair of cellular states in our data, weighted according to their ancestral distributions on day 2. We compared our findings with a standard differential gene expression analysis conducted on two groups of CRISPR knockout epithelial cells, C1-sgEED-Epi and C1-sgKMT2D-Epi (*SI Appendix*, S10 and Fig. S12). These two groups of epithelial cells were the ancestral cells for their respective EMT fates: the C1-sgEED-Mes and C1-sgKMT2D-Mes cells, respectively (36). Notably, four of the top five ranked genes overlapped between our and the CRISPR study—*TGFBI*, *KRT8*, and *CDH1* were significantly up-regulated, while *PHF19* was significantly down-regulated in the partial EMT trajectory (*t* test, $P < 1e\text{-}9$; fold changes are in *SI Appendix*, Table S12). The concordance observed between our predictions and the results from the CRISPR screen study partially validates our inference approach of predicting early key genes in EMT.

Our methodology leverages the inherent heterogeneity of cellular states that culminate in diverse cell fates under a single EMT inducer. This approach enables the identification of crucial early-stage genes that govern specific cell destinies, effectively circumventing the necessity for extensive preexisting biological knowledge when selecting a specific EMT inducer or applying CRISPR to knock out a specific gene for a corresponding cell fate. We further applied our early DEG analysis to two comprehensive gene sets—the ones that we employed to delineate stemness and proliferation patterns (58, 59). Our results highlighted that in the early phase of TGF-beta-induced EMT, genes such as *CENPF*, *CKS1B*, and *MKI67* were significantly up-regulated in the ancestors of the high EMT state on day 2 (*t* test, $P < 1e\text{-}10$ and fold changes 2.16, 1.80, and 1.76, respectively) (Fig. 5*C*). Similar patterns were observed on days 1 and 3 (*SI Appendix*, Fig. S13). In contrast, in the early cellular states of the partial EMT state, genes like *LAMA3*, *LAMB3*, and *ITGB4* were prominently expressed on day 2 (*t* test, $P < 1e\text{-}10$ and fold changes 1.66, 1.55, and 1.50, respectively) (Fig. 5*D*). Again, similar trends were observed on days 1 and 3 (*SI Appendix*, Fig. S13). Our findings concerning the increased expression of *LAMA3*, *LAMB3*, and *ITGB4* align with prior research that identified their role in demarcating cancer stem cell-enriched populations in a partially mesenchymal state (60, 61). Our methodology provides insights by enabling the identification of DEGs across a temporal spectrum (*SI Appendix*, Fig. S7). For instance, our time-resolved analysis reveals that the differential expression of *ITGB4* in the partial EMT state, when compared to the high EMT state, is more pronounced during the early stages than it is in the later phases of EMT. When comparing day 2 to day 8, the fold changes were 1.50 vs. 1.09, respectively. These findings potentially underscore crucial moments for timely interventions to influence the direction of EMT evolution.

Discussion

In this study, we utilized WOT to infer EMT trajectories as a data-driven model; however, the underlying mechanisms remain unidentified. Existing work in the field employs computational single-cell approaches to model EMT, utilizing mechanistic methods such as bifurcation and stability analysis from dynamical systems theory. These methods illustrate varying EMT responses to TGF-beta, corroborating our findings from data-driven models and further elucidating the underlying mechanisms of these diverse responses (62, 63). Moreover, another mechanistic approach involves constructing gene regulatory network circuits through a combination of transcriptomics data and network modeling. This approach helps identify the context-specific activity dynamics of common EMT transcription factors (64), and the activity dynamics of common EMT transcription factors in varying contexts (65). Therefore, future work should consider integrating these mechanistic models with OT analysis to enhance the predictions and uncover the underlying mechanisms driving these predictions.

Furthermore, WOT is based on optimal transport theory, which assumes that cells traverse the gene expression space via the shortest overall distance (26, 27). This foundational assumption serves as an unbiased starting point for cell state transitions (28). Future refinements could integrate prior knowledge of specific gene expression changes, adjusting gene distances based on this knowledge. This approach would allow us to leverage WOT more adaptively, inferring unknown system parts from existing biological understanding. Additionally, WOT employs an unbalanced optimal transport method, accommodating the effects of cell proliferation and death in the transport of cell states. However, the estimation of cell proliferation and death depends on our selection of gene sets from the literature. A recently published tool, TIGON (66), addresses this limitation by simultaneously reconstructing dynamic trajectories and population growth directly from the data.

Despite these caveats, the use of WOT has uncovered several insights into individual EMT trajectories. These insights, when integrated with existing EMT research, can offer a more comprehensive view of the EMT landscape. First, we found that the low EMT trajectory is determined early on, within a day of treatment. This result suggests that the initial state of these cells renders them resistant to TGF-beta, providing insights into two prior studies on EMT resistance: one study identified a subpopulation of epithelial cells with similar capabilities to receive and process TGF-beta signals but exhibited a notably weaker downstream response compared to more sensitive cell populations (36). Another study revealed that sustained *EPCAM* expression acts as a marker for epithelial clones in metastatic breast cancer that resist EMT induction, a trait shaped by the interplay between human *ZEB1* and its target, *GRHL2* (67).

Additionally, we observed that the expression of the *EED* and *EZH2* genes was down-regulated from day 0 to day 1 following TGF-beta treatment in the MCF10A cell line (Fig. 5*B* and *SI Appendix*, Fig. S11*B*). Although there is no established mechanism for this effect, we hypothesize that the TGF-beta-induced cytostatic effect is associated with decreased expression of *PRC2* components. *PRC2* components, particularly *EZH2*, are well-documented targets of cell cycle transcriptional regulation, which is up-regulated in proliferating stem cells and cancer cells (68, 69). Consequently, from day 0 to day 1, MCF10A cells show sensitivity to the TGF-beta-induced reduction in cell proliferation (Fig. 3*B* and *SI Appendix*, Fig. S6*A*), likely contributing to the reduced expression of *EED* and *EZH2*. Furthermore, after day 1, the dynamics of *EED* and *EZH2* expression diverged across the three trajectories. As shown previously (36), either a stable or transient loss of *PRC2* function is sufficient to activate an EMT trajectory and generate a partial mesenchymal cell state. As depicted in Fig. 5*B* and *SI Appendix*, Fig. S11*B*, after day 1, *EED* and *EZH2* maintained a low expression level in the partial EMT trajectory, suggesting a functional reduction of *PRC2*, which aligns with the previous findings (36). In contrast, the low and high EMT trajectories showed a restoration of *EED* and *EZH2* levels to pretreatment levels, indicating that *PRC2* remains functional.

Last, leveraging the heterogeneity of cellular responses to TGF-beta-induced EMT, our method effectively pinpoints early differentially expressed genes across distinct EMT trajectories from a broad set of candidates. For instance, we distinguished *ITGB4*,

LAMA3, and *LAMB3* due to their pronounced differential expressions in the early stages of the partial versus high EMT trajectories. As previously highlighted, *ITGB4* serves as an integrin subunit that interacts with specific matrix proteins, while *LAMB3* and *LAMA3* engage with different integrin subunits than does *ITGB4* (70, 71). Future validation of our findings could employ cell surface markers encoded by these genes to isolate early-phase cells and observe their responses under a consistent TGF-beta treatment timeline.

Materials and Methods

scRNA-seq Data Analysis. The single-cell RNA-seq datasets analyzed here were obtained from published studies (18, 36). For the dataset from Deshmukh et al., we used the processed sequencing data made available by the authors; the raw sequencing reads are available from the NCBI Sequence Read Archive (72) (BioProject Accession No. PRJNA698642). From the dataset from Zhang et al., processed single-cell RNA-seq profiles of HMLER cells subjected to EED/EZH2 knockout were downloaded from the Gene Expression Omnibus (GEO) database (73) (GEO Accession No. GSE158115). The procedures for quality control, data normalization, batch correction, and other steps for this dataset were performed as in the original paper (36). Detailed descriptions of the scRNA-seq data dimensionality reduction and clustering analysis are available in *SI Appendix*, S11.

Inferring Trajectories with WOT. We employed the WOT (28) method to analyze cell state transition probabilities over time in the scRNA-seq data, using normalized expression matrices and day annotations. Cell growth rates were determined using a logistic function based on cells' proliferation and apoptosis signatures from MsigDB gene sets (42, 43). These rates were then incorporated into an unbalanced transport optimization to model transitions over consecutive days, with parameters previously validated (28). This methodology enabled the prediction of transition maps following TGF-beta treatment and facilitated the computation of ATF distributions, which quantify the likelihood of each cell differentiating into specific fate subpopulations at early time points (*SI Appendix*, S12).

Assessing EMT Scores: The 76GS and KS Methods. EMT scores were calculated using two distinct methodologies, each employing different gene sets and metrics. The consistency between these methods has been verified through a comparative study involving multiple individual samples (40). In the 76GS method (37, 39), we computed the EMT score as a weighted sum of the expression levels of 76 EMT-related genes. The weight assigned to each gene was determined by its correlation with the *CDH1* (E-cadherin) expression level. The scores were subsequently adjusted such that the mean is 0. As a result, a negative score signifies that a cell's EMT state is closer to the epithelial (E) state than the mesenchymal (M) state. We then rescaled the scores by taking their negatives, thus aligning the direction of the scores with the progression from the E to M state. The second method, known as the KS method, was initially established based on a comparison between the cumulative distribution functions (CDFs) of the E and M signatures (38). According to this method, the EMT score is computed as the maximum difference between the two CDFs, i.e., the CDF of the M signature minus the CDF of the E signature. Therefore, a positive score for a sample indicates its closeness to the M state, and vice versa.

Computation of Cellular Signature Scores by ssGSEA. For determining the expression level of the stemness signature, we adopted gene sets from Lim et al. (58), specifically designed to distinguish between stemness and mature cell signatures by investigating mammary stem and luminal cells. For the hypoxia response signature, we employed gene sets from MSigDB (42, 43). The proliferation signature was determined using a gene set from Ben-Porath et al. (59) This set was compiled by merging three distinct gene groups: those that are functionally involved in proliferation, those with cyclical expression within the cell cycle, and those that were instrumental in the clustering of proliferative subpopulations within human breast tumor expression data. Additionally, we calculated the proliferation signature using two other gene sets associated with specific proliferation signaling pathways, G2M and mitotic spindle, from MSigDB (43). Details of the gene sets used for stemness, proliferation, hypoxia, and G2M signatures are provided in *SI Appendix*, Tables S13–S17. For further signaling pathway analysis, we investigated the gene sets of TGF-beta, PI3K-AKT-mTOR, Wnt, and IL6-JAK-STAT3 hallmarks from MSigDB (42, 43). We performed single-sample Gene Set Enrichment Analysis (ssGSEA) on all gene sets using GSEAPY (v1.0.4), a Python package[8686] (74). The enrichment scores for gene sets were transformed into z-scores, with adjustments made by shifting the mean and normalizing by the SD.

Data, Materials, and Software Availability. All code used to process data and generate figures is available on a public GitHub repository at https://github.com/Michorlab/OT-EMT (75). Previously published data were used for this work (72, 73).

ACKNOWLEDGMENTS. We would like to thank the Michor lab, Dr. Petra den Hollander, Dr. Sahand Hormoz, and Dr. Sendurai Mani for helpful discussions and comments. We gratefully acknowledge support of the Ludwig Center at Harvard and the Dana-Farber Cancer Institute's Center for Cancer Evolution.

Author affiliations: [a]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215; [b]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02215; [c]Center for Cancer Evolution, Dana-Farber Cancer Institute, Boston, MA 02215; [d]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138; [e]State Key Laboratory of Molecular Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China; [f]Yale Center for Systems and Engineering Immunology and Department of Immunobiology, Yale School of Medicine, New Haven, CT 06510; [g]Interdisciplinary Mathematics Initiative, Indian Institute of Science, Bangalore 560012, India; [h]Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore 560012, India; [i]Department of Mathematics, University of British Columbia, Vancouver, BC V6T 1Z2, Canada; [j]Center for Theoretical Biological Physics, Northeastern University, Boston, MA 02115; [k]Department of Physics, Northeastern University, Boston, MA 02115; [l]The Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02138; and [m]The Ludwig Center at Harvard, Boston, MA 02115

Author contributions: Y.-C.C., S.T., G.S., H.L., and F.M. designed research; Y.-C.C., Y.Z., S.T., H.L., T.O.M., and F.M. performed research; Y.-C.C., Y.Z., B.V.H., M.K.J., G.S., H.L., T.O.M., and F.M. analyzed data; and Y.-C.C., Y.Z., S.T., M.K.J., G.S., H.L., T.O.M., and F.M. wrote the paper.

1. E. D. Hay, An overview of epithelio-mesenchymal transformation. *Acta Anat. (Basel)* **154**, 8–20 (1995).
2. Y. Kang, J. Massagué, Epithelial-mesenchymal transitions: Twist in development and metastasis. *Cell* **118**, 277–279 (2004).
3. J. P. Thiery, J. P. Sleeman, Complex networks orchestrate epithelial–mesenchymal transitions. *Nat. Rev. Mol. Cell Biol.* **7**, 131–142 (2006).
4. M. A. Nieto, R. Y.-J. Huang, R. A. Jackson, J. P. Thiery, EMT: 2016. *Cell* **166**, 21–45 (2016).
5. A. Dongre, R. A. Weinberg, New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. *Nat. Rev. Mol. Cell Biol.* **20**, 69–84 (2019).
6. M. K. Wendt, M. A. Taylor, B. J. Schiemann, W. P. Schiemann, Down-regulation of epithelial cadherin is required to initiate metastatic outgrowth of breast cancer. *MBoC* **22**, 2423–2435 (2011).
7. Y. Wang, B. P. Zhou, Epithelial-mesenchymal transition–A hallmark of breast cancer metastasis. *Cancer Hallm.* **1**, 38–49 (2013).
8. M. Lu, M. K. Jolly, H. Levine, J. N. Onuchic, E. Ben-Jacob, MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 18144–18149 (2013).
9. L. Zhang, F. Zhou, P. ten Dijke, Signaling interplay between transforming growth factor-β receptor and PI3K/AKT pathways in cancer. *Trends Biochem. Sci.* **38**, 612–620 (2013).
10. J. Roche, The epithelial-to-mesenchymal transition in cancer. *Cancers* **10**, 52 (2018).
11. S. V. Vasaikar *et al.*, EMTome: A resource for pan-cancer analysis of epithelial-mesenchymal transition genes and signatures. *Br. J. Cancer* **124**, 259–269 (2021).
12. J. Taipale, P. A. Beachy, The Hedgehog and Wnt signalling pathways in cancer. *Nature* **411**, 349–354 (2001).
13. Y. Katoh, M. Katoh, FGFR2-related pathogenesis and FGFR2-targeted therapeutics (Review). *Int. J. Mol. Med.* **23**, 307–311 (2009).
14. A.-E.A. Moustafa, A. Achkhar, A. Yasmeen, EGF-receptor signaling and epithelial–mesenchymal transition in human carcinomas. *FBS* **4**, 671–684 (2012).
15. I. Espinoza, L. Miele, Deadly crosstalk: Notch signaling at the intersection of EMT and cancer stem cells. *Cancer Lett.* **341**, 41–45 (2013).
16. C.-H. Heldin, M. Vanlandewijck, A. Moustakas, Regulation of EMT by TGFβ in cancer. *FEBS Lett.* **586**, 1959–1970 (2012).
17. D. P. Cook, B. C. Vanderhyden, Context specificity of the EMT transcriptional response. *Nat. Commun.* **11**, 2142 (2020).
18. A. P. Deshmukh *et al.*, Identification of EMT signaling cross-talk and gene regulatory networks by single-cell RNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2102050118 (2021).
19. C. H. Waddington, *The Strategy of the Genes* (Routledge, 2014).
20. J. E. Ferrell, Bistability, bifurcations, and Waddington's epigenetic landscape. *Curr. Biol.* **22**, R458–R466 (2012).

21. S. A. Mani *et al.*, The epithelial–mesenchymal transition generates cells with properties of stem cells. *Cell* **133**, 704–715 (2008).
22. M. K. Jolly *et al.*, Towards elucidating the connection between epithelial–mesenchymal transitions and stemness. *J. R Soc. Interface* **11**, 20140962 (2014).
23. L. Mazutis *et al.*, Single-cell analysis and sorting using droplet-based microfluidics. *Nat. Protoc.* **8**, 870–891 (2013).
24. A. Tanay, A. Regev, Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
25. D. E. Wagner, A. M. Klein, Lineage tracing meets single-cell omics: Opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).
26. C. Villani, *Optimal Transport: Old and New* (Springer, Berlin/Heidelberg, Germany, 2016).
27. L. Chizat, G. Peyré, B. Schmitzer, F.-X. Vialard, Scaling algorithms for unbalanced optimal transport problems. *Math. Comput.* **87**, 2563–2609 (2018).
28. G. Schiebinger *et al.*, Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943.e22 (2019).
29. L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, F. J. Theis, Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
30. W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
31. G. Gorin, V. Svensson, L. Pachter, Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biol.* **21**, 39 (2020).
32. V. Bergen, M. Lange, S. Peidli, F. A. Wolf, F. J. Theis, Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
33. V. Bergen, R. A. Soldatov, P. V. Kharchenko, F. J. Theis, RNA velocity–Current challenges and future perspectives. *Mol. Syst. Biol.* **17**, e10282 (2021).
34. X. Qiu *et al.*, Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690–711.e45 (2022).
35. K. Lundgren, B. Nordenskjöld, G. Landberg, Hypoxia, snail and incomplete epithelial–mesenchymal transition in breast cancer. *Br. J. Cancer* **101**, 1769–1781 (2009).
36. Y. Zhang *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).
37. L. A. Byers *et al.*, An epithelial–mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin. Cancer Res.* **19**, 279–290 (2013).
38. T. Z. Tan *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).
39. C. C. Guo *et al.*, Dysregulation of EMT drives the progression to clinically aggressive sarcomatoid bladder cancer. *Cell Rep.* **27**, 1781–1793.e4 (2019).
40. P. Chakraborty, J. T. George, S. Tripathi, H. Levine, M. K. Jolly, Comparative study of transcriptomics-based scoring metrics for the epithelial-hybrid-mesenchymal spectrum. *Front. Bioeng. Biotechnol.* **8**, 220 (2020).
41. L. Ambrosio, N. Gigli, G. Savaré, "Gradient Flows" in *Metric Spaces and in the Space of Probability Measures*, M. Struwe, Ed. (Birkhäuser, Basel, Switzerland, 2008).
42. A. Subramanian *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005).
43. A. Liberzon *et al.*, Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
44. B. Bao *et al.*, The biological kinship of hypoxia with CSC and EMT and their relationship with deregulated expression of miRNAs and tumor aggressiveness. *Biochim. Biophys. Acta (BBA): Rev. Cancer* **1826**, 272–296 (2012).
45. A. Emami Nejad *et al.*, The role of hypoxia in the tumor microenvironment and development of cancer stem cell: A novel approach to developing treatment. *Cancer Cell Int.* **21**, 62 (2021).
46. X. Lu, Y. Kang, Hypoxia and hypoxia-inducible factors: Master regulators of metastasis. *Clin. Cancer Res.* **16**, 5928–5935 (2010).
47. J. Massagué, S. W. Blain, R. S. Lo, TGFβ signaling in growth control, cancer, and heritable disorders. *Cell* **103**, 295–309 (2000).
48. J. Donovan, J. Slingerland, Transforming growth factor-β and breast cancer: Cell cycle arrest by transforming growth factor-β and its disruption in cancer. *Breast Cancer Res.* **2**, 116 (2000).
49. Y. Zhang, P. B. Alexander, X.-F. Wang, TGF-β family signaling in the control of cell proliferation and survival. *Cold Spring Harb. Perspect. Biol.* **9**, a022145 (2017).
50. C. E. Aban *et al.*, Downregulation of E-cadherin in pluripotent stem cells triggers partial EMT. *Sci. Rep.* **11**, 2048 (2021).
51. V. Aggarwal *et al.*, P4HA2: A link between tumor–intrinsic hypoxia, partial EMT and collective migration. *Adv. Cancer Biol.: Metastasis* **5**, 100057 (2022).
52. Z. Yao *et al.*, TGF-β IL-6 axis mediates selective and adaptive mechanisms of resistance to molecular targeted therapy in lung cancer. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 15535–15540 (2010).
53. A. Akhmetshina *et al.*, Activation of canonical Wnt signalling is required for TGF-β-mediated fibrosis. *Nat. Commun.* **3**, 735 (2012).
54. C. L. Carpenter, L. C. Cantley, Phosphoinositide 3-kinase and the regulation of cell growth. *Biochim. Biophys. Acta (BBA): Rev. Cancer* **1288**, M11–M16 (1996).
55. K. Hinohara *et al.*, KDM5 histone demethylase activity links cellular transcriptomic heterogeneity to therapeutic resistance. *Cancer Cell* **34**, 939–953.e9 (2018).
56. M. Tyler, I. Tirosh, Decoupling epithelial–mesenchymal transitions from stromal profiles by integrative expression analysis. *Nat. Commun.* **12**, 2592 (2021).
57. B. Elenbaas *et al.*, Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. *Genes Dev.* **15**, 50–65 (2001).
58. E. Lim *et al.*, Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast Cancer Res.* **12**, R21 (2010).
59. I. Ben-Porath *et al.*, An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nat. Genet.* **40**, 499–507 (2008).
60. B. Bierie *et al.*, Integrin-β4 identifies cancer stem cell-enriched populations of partially mesenchymal carcinoma cells. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E2337–E2346 (2017).
61. S. V. Puram *et al.*, Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624.e24 (2017).
62. W. Wang, D. Poe, Y. Yang, T. Hyatt, J. Xing, Epithelial-to-mesenchymal transition proceeds through directional destabilization of multidimensional attractor. *eLife* **11**, e74866 (2022).
63. M. Barcenas, F. Bocci, Q. Nie, Tipping points in epithelial-mesenchymal lineages from single-cell transcriptomics data. *Biophys. J.* 10.1016/j.bpj.2024.03.021 (2024).
64. D. Ramirez, V. Kohar, M. Lu, Toward modeling context-specific EMT regulatory networks using temporal single cell RNA-Seq data. *Front. Mol. Biosci.* **7**, 54 (2020).
65. Y. Sha, S. Wang, F. Bocci, P. Zhou, Q. Nie, Inference of intercellular communications and multilayer gene-regulations of epithelial–mesenchymal transition from single-cell transcriptomic data. *Front. Genet.* **11**, 604585 (2021).
66. Y. Sha, Y. Qiu, P. Zhou, Q. Nie, Reconstructing growth and dynamic trajectories from single-cell transcriptomics data. *Nat. Mach. Intell.* **6**, 25–39 (2024).
67. M. Saini *et al.*, Resistance to mesenchymal reprogramming sustains clonal propagation in metastatic breast cancer. *Cell Rep.* **42**, 112533 (2023).
68. A. P. Bracken *et al.*, EZH2 is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer. *EMBO J.* **22**, 5323–5335 (2003).
69. Y. Ma, K. Kanakousaki, L. Buttitta, How the cell cycle impacts chromatin architecture and influences cell fate. *Front. Genet.* **6**, 19 (2015).
70. F. G. Giancotti, G. Tarone, Positional control of cell fate through joint integrin/receptor protein kinase signaling. *Annu. Rev. Cell Dev. Biol.* **19**, 173–206 (2003).
71. S. Spaderna *et al.*, A transient, EMT-linked loss of basement membranes indicates metastasis and poor survival in colorectal cancer. *Gastroenterology* **131**, 830–840 (2006).
72. R. Leinonen, H. Sugawara, M. Shumway, On behalf of the International Nucleotide Sequence Database Collaboration, The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
73. T. Barrett *et al.*, NCBI GEO: Archive for functional genomics data sets–Update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
74. Z. Fang, X. Liu, G. Peltz, GSEApy: A comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023).
75. Y.-C. Cheng, OT-EMT. GitHub. https://github.com/Michorlab/OT-EMT. Deposited 14 September 2023.