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

Yu-Chen Cheng [1,2,3,4], Yun Zhang [5], Shubham Tripathi [6], Harshavardhan BV [7], Mohit Kumar Jolly [8], Geoffrey Schiebinger [9], Herbert Levine [10,11], Thomas O. McDonald [1,2,3,4], and Franziska Michor [1,2,3,4,12,13,*]

1. Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA
2. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
3. Center for Cancer Evolution, Dana-Farber Cancer Institute, Boston, MA, USA
4. Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA
5. 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
6. Department of Immunology, Yale University, New Haven, CT, USA
7. Interdisciplinary Mathematics Initiative, Indian Institute of Science, Bangalore, 560012, India
8. Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore, 560012, India
9. Department of Mathematics, University of British Columbia, Vancouver, BC, Canada
10. Center for Theoretical Biological Physics, Northeastern University, Boston, MA, USA
11. Department of Physics, Northeastern University, Boston, MA, USA
12. The Broad Institute of MIT and Harvard, Cambridge, MA, USA
13. The Ludwig Center at Harvard, Boston, MA, USA

*Author for correspondence. Email: michor@jimmy.harvard.edu.

Abstract

The epithelial-to-mesenchymal transition (EMT) gives rise to a variety of different cell states, yet it remains unclear whether these distinct states represent varying degrees of progression along a single trajectory or serve as endpoints of multiple potential trajectories that individual cells might pursue. Understanding the intricacies of these diverse EMT trajectories and cell fates, especially using lineage tracing at the single cell level using single cell RNA sequencing (scRNAseq) data, remains challenging. In this study, we employed optimal-transport analysis to reconstruct the most probable past trajectories of different cell fates during TGF-beta-induced EMT in the MCF10A cell line. Our analysis revealed three distinct temporal processes: those resulting in a low EMT, partial EMT, and high EMT state. The partial EMT trajectory exhibited substantial variations in the EMT signature and displayed unique cellular characteristics, including high stemness, low proliferation scores, and a potent cellular response to hypoxia. Along the partial EMT trajectory, we observed downregulation of the EED and EZH2 genes together with early high expression of TGFBI, KRT8, and CDH1—a finding that was validated by recent CRISPR-associated knock-out screening studies. Through our early differential gene expression analysis, we comprehensively explored genes linked to proliferation and stemness, identifying ITGB4, LAMA3, and LAMB3 as differentially expressed in the early stages of the partial versus high EMT trajectory. Our analysis thus elucidates novel potential targets for modulating EMT trajectories.

Introduction

The epithelial-mesenchymal transition (EMT) is a pivotal process underpinning a range of biological phenomena, from embryonic development and wound healing to tumor metastasis[1–4]. During EMT, epithelial cells lose apical-basal polarity and adhesion to other cells and acquire mesenchymal traits such as invasiveness and migratory capabilities[2–4]. 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)[5,6] . 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[7,8]. Progression along this spectrum is tightly regulated by a set of key transcription factors, including the Snail, Zeb, and Twist families[9,10]. 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, IL6-JAK-STA3, and NOTCH[4,11–15].

Cells in a syngeneic, phenotypically homogeneous population have been observed to adopt distinct states upon treatment with an EMT inducer[16,17]. However, it is unclear whether these distinct EMT states correspond to different degrees of progression along the same overall trajectory or whether they are end points of distinct trajectories that individual cells can embark upon. The idea of divergent trajectories, through a developmental Waddington landscape[18], is well-accepted in stem cell biology[19]. Given the close association between EMT and stemness[20,21], 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[17] (Fig 1A).

While scRNAseq data offers a wealth of insights into the heterogeneity of cellular states[22,23], the inherent need to sacrifice cells at each time point precludes the ability to trace individual lineages over time. This restriction poses a challenge to reconstructing varied trajectories from time-series scRNAseq data[24]. To address this challenge, we employed a method based on optimal transport (OT) analysis[25,26], known as Waddington OT (WOT)[27]. Utilizing this technique, we reconstructed lineage trajectories at single-cell resolution using the time series scRNAseq data in Deshmukh *et al*[17], enabling identification of diverse trajectories leading to distinct EMT fates. We then investigated various cellular signatures along the predicted trajectories of cells with distinct fates after TGF-beta treatment. An overview of the general framework is provided in Fig 1B-F. We focused on stemness, proliferation, and cellular hypoxic response signatures since, while these signatures have known associations with EMT[4,20,28], their variations across different EMT pathways have not been extensively explored. In addition, our trajectory analysis at the single gene level enabled us to predict differentially expressed genes (DEGs) in the earliest phases of each fate. Early gene expression linked to a specific fate was then partially validated through methods such as inhibitor screens of EMT regulators and CRISPR-associated gene knockout screens[29], highlighting the robustness of our predictions. Our approach helps identify early key genes that determine different EMT outcomes.

Results

Reconstruction of Probabilistic Trajectories of EMT using Optimal Transport Analysis

Given that scRNAseq data currently cannot be connected to lineage tracing experiments, we set out to computationally infer likely ancestor cell states for different EMT fates. In the study by Deshmukh *et al.*[17], an immortalized human mammary epithelial cell line, MCF10A, was treated with TGF-beta for 1, 2, 3, 4, or 8 days (Fig 1A) and scRNAseq data was 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 (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[30–33], for each subpopulation (Fig 2A, 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 significant p-values (*t*-test, $p < 0.005$) for each pairwise comparison (Fig 3B and Table S4, day 8). Use of the KS method yielded consistent outcomes (Fig S5).

To deduce the trajectories of individual cell states across sequential scRNAseq datasets, we utilized WOT[27]. This 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. This prediction is based on the assumption that cellular states navigate the gene expression space using the shortest overall distance (Fig S1 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 2A). 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 (Methods).

To trace cellular origins leading to various fates, we categorized cells with over 75% probability of transitioning to specific fates as 'top ancestors' (Methods). Cells below this threshold were labeled as ‘undetermined ancestors’. With ongoing TGF-beta treatment, the portion of the undetermined ancestors decreased sharply from 91.62% on day 0 to 15.92% by day 4. 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 2BC). This temporal shift in proportions indicates a delayed inclination towards the medium and high EMT fates, induced by TGF-beta. Our findings were robust when alternative probability thresholds, from 75% to 90%, for defining top ancestors were applied (Methods and Supplementary Appendix S1).

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 seen at day 8 for the three fates (Fig 2A and Fig S2). This observation suggests a divergence in EMT phenotypes. To quantify this divergence, we computed the distance between the cell state distributions of each pair of trajectories at every time point (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 (Fig S3B).

Deciphering Unique Cellular and 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 3A, Fig S5, 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 (Fig S4 and Methods). This approach unveiled three distinct trajectories, each showing unique average EMT score trajectories with non-overlapping 95% confidence intervals,

throughout the course of TGF-beta treatment (Fig 3B and Fig S5). The clear separation into low, medium, and high EMT trajectories was consistently observed using both the 76GS and KS EMT scoring methods (Fig S5 and Methods).

Furthermore, we analyzed stemness, proliferation signatures, and associated hallmarks of hypoxia response and the G2M checkpoint among cells belonging to the three fates. For each cell, we computed those signatures using single-sample gene set enrichment analysis (Fig 3A and Methods). In both low and high EMT trajectories, stemness enrichment scores increased by 2.1 as determined by the z-score difference between day 8 and day 0. For the medium EMT trajectory, the enrichment z-scores increased by 2.3. Similarly, hypoxia scores increased by 2.0 for low, 2.1 for medium, and 1.9 for high EMT trajectories. This trend aligns with prior research that links hypoxia to enhanced stemness in EMT[34,35]. Like the EMT signature, these three trajectories stood out with their non-overlapping 95% confidence intervals when characterized by these two signatures (Fig 3B and Fig S5). Of particular interest, 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 3B and Fig S5). These findings resonate with earlier studies identifying an intermediate EMT stage characterized by heightened stemness and a pronounced response to hypoxia[3,28].

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 3B). These z-score shifts indicate notable decreases relative to the average enrichment score. A similar trend was observed in the G2M checkpoint hallmark (Fig S5). This decrease reflects the known role of TGF-beta in inhibiting cell division[36–38]. From day 1 through day 8, cells regain their proliferative capacity, evidenced by enrichment score rebounds of 2.0 for low, 1.8 for medium, and 2.1 for high EMT (Fig 3B and Fig S5). 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 (Fig S7).

We then further analyzed the dynamics of individual genes pivotal to EMT, such as CDH1 (E-cadherin) and CDH2 (N-cadherin). 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 3C). This downward trend aligns with previous findings indicating that CDH1 downregulation triggers partial EMT[39]. Conversely, CDH2 showed its most pronounced expression rise in the medium EMT trajectory, with an increase of 0.87 in z-score, in contrast to increments of 0.63 and 0.51 in the high and low EMT trajectories, respectively. Furthermore, these gene expression patterns are markedly distinct, underscored by their non-overlapping 95% confidence intervals (Fig 3C). This finding aligns with a previous study showing elevated expression of CDH2 in partial EMT using the same cell line and treatment type[29]. Beyond CDH2, Zhang *et al.*[29] highlighted elevated expression of POSTN and KRT8 expressions as indicators of the partial EMT phase. Consistent with their findings, our medium EMT trajectory displayed higher increases in POSTN and KRT8 expression levels, with z-score increases of 1.48 and 0.96 from day 0 to day 8, respectively. In comparison, the high and low EMT trajectories showed more modest z-score gains of 0.84 and 0.6 for POSTN, and 0.12 and 0.3 for KRT8, respectively (Fig 3D). Lastly, the medium EMT trajectory demonstrated a notable increase in HIF-1A and Snail expression levels, with z-score increases from day 0 to day 8 of 0.94 and 0.66, respectively. When compared to the high and low EMT trajectories, these genes only showed modest z-score increases of 0.25 and 0.05 for HIF-1A, and 0.12 and -0.13 for Snail, respectively (Fig 3E). This finding further supports the classification of the medium EMT trajectory as partial EMT, given the known roles of these genes in hypoxia and partial EMT fates[28,40].

To investigate whether TGF-beta treatment correlates with other essential EMT-related signaling pathways, we further characterized the pairwise dynamics of several cellular signatures over time (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 3F). These observations resonate with previous findings regarding the concurrent regulation of these pathways throughout the EMT process[8,41,42]. The intricate interplay between TGF-beta and PI3K signaling pathways, which includes both antagonistic and cooperative interactions, has been discussed previously[8]. Each of the three EMT trajectories demonstrated increased activity in the TGF-beta pathway from day 0 to day 8, with enrichment scores rising to 2.1, 2.3, and 2.2 for the low, partial and high EMT trajectories, respectively. In contrast, the PI3K pathway experienced a notable decrease in the partial EMT trajectory towards the end of the treatment window, dropping from an enrichment score of 0.7 on day 4 to 0.2 on day 8. Meanwhile, the enrichment scores for the other two trajectories remained relatively stable, decreasing from 1.2 to 1.1 for the low EMT and 1.1 to 0.9 for the high EMT in the same period (Fig 3F). With PI3K signaling recognized as a prominent driver of cell growth and proliferation[43], this observed decline aligns with the lower proliferation scores and G2M checkpoint pathway activity levels noted along the partial EMT trajectory (Fig. 3B and Fig. S5).

Analyzing Temporal Signature Distributions to Unravel Heterogeneity in EMT Trajectories

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 2B) with time-ordered individual cellular EMT signature scores (Fig 4A and Fig S8). This integration elucidated the relationship between ancestral cell EMT states and their potential to transition into a specific fate (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. 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 4B and Fig S8). This pattern is discernible throughout day 3 (Fig 4B and Fig S8), suggesting that this early phase of partial EMT trajectory displays a greater degree of variability in EMT score expressions compared to the early phases of the other EMT trajectories.

To further explore the significant heterogeneity within the partial EMT trajectory, we calculated pairwise cell state distances[44] (Methods), 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 refer to Table S5). To determine if 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 (Table S1 and Methods). Notably, during the four days of treatment, cell state variability defined by the average pairwise cell distances in the EMT gene expression space was markedly higher in the partial EMT compared to the high EMT (*t*-test, $p < 1e-9$; fold changes were 1.11, 1.09, 1.10 for days 1 through 3, respectively) (Fig 4E and Table S2). However, within the same period, such differences were less evident when cell state heterogeneity was determined using the other two gene spaces; in these cases, p-values were generally above 0.05. Exceptions were only found on days 2 and 3 within the full gene set, but they demonstrated minor fold changes of 1.03 and 1.04, respectively. 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 partial EMT[45].

Beyond pairwise cell state distances, we also explored cellular heterogeneity by analyzing EMT and stemness score distributions among the top ancestors of the three identified EMT fates (Methods). We

found that one distinguishing feature of the partial EMT trajectory was its broad variation in the EMT signature, paired with a consistent stemness signature. The top ancestors of the partial EMT trajectory exhibited a more pronounced EMT score variation than those of the high EMT fate (Levine test, $p < 1e-10$ for days 1-8) (Fig 4C, Table S4 and Methods). In contrast, the partial EMT trajectory exhibited a level of variation in stemness scores similar to, or even less than, that of the high EMT trajectory (Levine test, $p > 0.05$ on days 1, 4, and 8). On days when significant differences did occur (Levine test, $p < 1e-5$ on days 2 and 3), the variations were more significant in the high than the partial EMT trajectory (Fig 4D and Table S6).

To further explore the interplay between EMT and stemness signatures, we examined the joint distributions of these signatures at various time points (Methods). 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 between the low and high EMT trajectories, whereas a pronounced stemness signature demarcated the partial EMT trajectory from the other two (Fig 4F). 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) (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 towards a distinctly differentiated state, whether closer to a pure E or M state, tend to exhibit less stemness[21].

Predicting and Validating Key Early Genes in EMT

To validate our identified trajectories, we compared our findings with a recent study that reported a substantial induction of a 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[29]. 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 5A). 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 non-overlapping 95% confidence intervals. Importantly, there was a noticeable decline in EED and EZH2 expressions, predominantly within the partial EMT trajectory (Fig 5B). For instance, the decline for EED was quantified as a 0.18 z-score decrease in its expression level, contrasted with 0.07 in the high EMT and 0.04 in the low EMT trajectories. This decline in expression was accentuated in the later phases of TGF-beta treatment, registering a 0.11 z-score drop between days 4 and 8. 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 from -0.07 to -0.16 (Fig 5C and Table S9). Similar patterns were observed for the EZH2 gene (Fig 5A-C). These findings align with Zhang *et al.*[29], where the combination of TGF-beta with either the EED inhibitor (EED226) or the EZH2 inhibitor (tazemetostat) induces a partial EMT fate.

In line with these findings, the CRISPR screen study revealed that knocking out EED genes promotes a partial EMT state with increased stemness[29]. 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[29,46]. Therefore, we curated an EMT-related gene list from both the primary study that generated our data[17] and the CRISPR screen study[29] (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 (Fig S11). Remarkably, three out of the top four ranked genes—TGFBI, POSTN, and KRT8—were significantly upregulated in the partial EMT state in both datasets (*t*-test, $p < 1e-10$; fold changes in Table S11). This analysis further supports our characterization of the partial and high EMT fates within our dataset.

We then conducted an early phase gene expression analysis on our dataset, evaluating differential gene expression between each pair of cellular states, weighted according to their ancestral distributions one day after TGF-beta treatment. 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 cells (Fig S11 and Methods). 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[29]. Notably, four of the top five ranked genes overlapped between our study and the CRISPR-based study—TGFBI, KRT8, and CDH1 were significantly upregulated, while PHF19 was significantly downregulated in partial EMT (*t*-test, $p < 1e-9$; fold changes are in Fig 5D-F). 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 pre-existing biological knowledge when selecting a specific EMT inducer or applying CRISPR to knock out a specific gene for a corresponding cell fate. We applied our early DEGs analysis to two comprehensive gene sets – the ones that we employed to delineate stemness and proliferation patterns[47,48]. Our results highlighted that in the early phase of TGF-beta induced EMT, genes such as CENPF, CKS1B, and MKI67 were significantly upregulated in the ancestors of the high EMT state on day 2 (*t*-test, $p < 1e-10$ and fold changes 2.16, 1.80, and 1.76, respectively) (Fig 5G). Similar patterns were observed on days 1 and 3 (Fig S12). 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, $p < 1e-10$ and fold changes 1.66, 1.55, and 1.50, respectively) (Fig 5H). Again, similar trends were observed on days 1 and 3 (Fig S12 for details). Specifically, our finding concerning the increased expression of ITGB4 aligns with prior research that identified its role in demarcating cancer stem cell-enriched populations in a partially mesenchymal state[49]. Our methodology advances the field by enabling the identification of DEGs across a temporal spectrum. This 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 (Fig S7). For instance, when comparing day 2 to day 8, the fold changes were 1.50 vs 1.09, respectively (Fig 5H). These findings potentially underscore crucial moments for timely interventions to influence the direction of EMT evolution.

Discussion

ScRNAseq data provides deep insights into the intricate heterogeneity of cellular states[22,23]. However, a key limitation concerns the lack of ability to track individual lineages arising from the need to sacrifice cells at each data collection point, complicating the reconstruction of diverse trajectories from time series scRNAseq datasets. To circumvent this limitation, we used WOT to reconstruct lineages at the single cell level from EMT time series scRNAseq data[17], allowing us to identify three separate trajectories culminating in distinct EMT outcomes. Of the three trajectories, the low-EMT trajectory, characterized by slight alterations in the EMT score and minimal changes in E-cadherin and N-cadherin expression, represents cells with a weak TGF-beta-induced EMT response. The partial EMT trajectory, aside from its medium EMT score, is marked by a low proliferation signature alongside elevated stemness. In addition,

its marked variability in EMT signatures underscores its inherent EMT heterogeneity, suggesting a spectrum of phenotypes rather than a single partial EMT state. In our early DEG analysis, top-tier genes like TGFBI, KRT8, and CDH1 were highlighted for their increased expression early in the partial EMT phase, aligning with findings from an independent CRISPR study. Expanding our DEG analysis to larger gene sets related to stemness and proliferation, we unveiled a novel set of pivotal early-stage genes, of which ITGB4 was particularly noteworthy, that distinguish the partial EMT from the high EMT trajectories.

Through our analysis, we have uncovered several novel findings that differentiate 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 novel 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[29]. 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[50]. Furthermore, TGF-beta is well-known as a promoter of mesenchymal gene expression and a suppressor of cell proliferation[36–38]. However, as far as we are aware, the pronounced reaction to TGF-beta within the partial EMT trajectory, leading to a significant reduction in proliferation and a moderate increase in mesenchymal markers, has not been emphasized in earlier research, thereby indicating a promising area for future exploration. Lastly, 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 partial versus high EMT trajectories. As previously highlighted, ITGB4 serves as an integrin subunit that interacts with specific matrix proteins, while LAMB3 and LAMA3 appear to engage with different integrin subunits than does ITGB4[51,52]. Taken together with our predictions, these findings imply that the extracellular matrix may have a pivotal role in steering cells toward a partial EMT fate, a notion deserving of further study.

However, in this study, we focused on gene expression alterations following EMT induction through scRNA seq data analysis. Such a scope, however, might not capture the complete EMT landscape. Variabilities in gene expression can arise from intrinsic noise[53], extracellular environmental changes[54,55], and differing epigenetic states[56]. The overall approach described in this study, when applied to suitable time series datasets, could help disentangle the role of the aforementioned mechanisms in driving differing responses to EMT induction and in driving heterogeneity in various cell fate transition processes[57]. For instance, using this approach on spatial transcriptomics from tumors, paired with latent time inferences, can elucidate the tumor environment's influence on EMT heterogeneity[58]. Furthermore, the same approach may be applied to datasets that jointly profile RNA expression and chromatin accessibility[59] at different time points to probe the contribution of epigenetic states in driving distinct transcriptional responses[50], and whether epigenetic changes precede gene expression changes during a cell fate transition.

Additionally, WOT is based on optimal transport theory, which assumes that cells traverse the gene expression space via the shortest overall distance[25,26]. This foundational assumption serves as an unbiased starting point for cell state transitions[27]. 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. Another limitation of the WOT approach is its requirement of time-series scRNAseq data[27,60], a need which is restricted by the cost associated with conducting multiple time-point

experiments. To overcome this limitation, WOT could potentially be extended to a pool of scRNAseq data without knowledge of the timeline. This advancement would entail inferring the latent time variable prior to using WOT. Given the prevalent availability of EMT data of this type[61,62], this extension could lead to the effective identification of key genes that demonstrate significant expression across various cell lines when exposed to different EMT inducers.

In summary, we uncovered three distinct trajectories leading to TGF-beta induced EMT. Most interestingly, the partial trajectory is characterized by an intermediate EMT score, elevated stemness, pronounced hypoxia response, and decreased proliferation. The reduced expression levels of the EED and EZH2 genes in this trajectory are consistent with the findings of a CRISPR study. Finally, we identified key genes upregulated in the early phases of cells on this trajectory, which had not been previously linked to early EMT regulation. Our approach represents a computational method to circumvent the need to lineage-trace in single cell experiments, and sheds light onto the diverse trajectories leading to different EMT fates.

Methods

scRNA-Seq Data Analysis

The single cell RNA-seq datasets analyzed here were obtained from published studies[17,29]. 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[63] (BioProject accession 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[64] (GEO accession GSE158115). The procedures for quality control, data normalization, batch correction, and other steps for this dataset were performed as in the original paper[29].

We used two methods to reduce the dimensionality of the single cell RNA-seq datasets. Following the approach in[29], we performed Uniform Manifold Approximation and Projection (UMAP) dimensional reduction using Seurat v3.61[65] for the HMLER dataset. For the MCF10A dataset, we followed the approach in Deshmukh *et al.*[17] and used force-directed layout embedding (FLE)[66], a type of graph visualization algorithm, to generate a nearest-neighbor graph built from the single cell gene expression data. In the FLE visualization, each cell is portrayed as a node and is linked to its 'k' closest neighboring cells. The application of FLE allows for an intuitive visualization of cellular relationships, as it induces an attraction between nodes (cells) connected by an edge and a repulsion between nodes that lack an interconnecting edge. The FLE visualization was carried out using the Pegasus tool [Pegasus 1.8.0]. MCF10A cells on day 8 were divided into three distinct subpopulations (shown in Fig. 2A, day 8: blue, green, and orange points) by applying the k-means clustering algorithm (k = 3) to the FLE visualization coordinates of the day 8 subset. The choice of k = 3 was primarily informed by our existing biological understanding of the three EMT states—epithelial, partial, and mesenchymal—referenced in earlier studies[61]. To further substantiate this choice, we investigated the results of k values ranging from 2 to 6, using the Silhouette score as a metric to evaluate the clustering performance. The Silhouette score ranges from -1 to +1, where a higher score suggests that objects in the same cluster are closer to each other than to objects in neighboring clusters[67]. Among these, k = 3 resulted in the second-highest Silhouette score of 0.65, only slightly behind k = 4, which had a score of 0.67 (Fig S13). However, k = 3 produced balanced cluster sizes of 589, 746, and 556 cells, while k = 4 resulted in a notably smaller cluster with just 32 cells alongside clusters of 589, 721, and 549 cells. Taking both the Silhouette score and cell cluster sizes into account, we selected k=3 as the most appropriate value.

Inferring trajectories with Waddington-OT

We employed Waddington-OT (WOT)[27] to calculate the transition probabilities between cells across successive time points as done previously[27]. Specifically, our input data consisted of the normalized gene expression matrix alongside the day annotations from the preprocessed scRNAseq data for cell collection. By default in WOT, cellular states—defined by the gene expression matrix—are assumed to be uniformly distributed on each day, though adjusted by growth rates as delineated in Schiebinger *et al.*[27]. To calculate the growth rate, we used WOT's inherent logistic function, which transforms the proliferation (or apoptosis) signature score into the birth (or death) rate for each cell. Subsequently, the net growth rate was derived from the difference between the birth and death rate. In line with WOT, we used the G2M hallmark and the apoptosis hallmark gene sets from MsigDB[68,69] to compute the proliferation and apoptosis signature scores, respectively.

Furthermore, the cell state distributions, covering two consecutive days, were then used as the marginal distributions in our unbalanced transport optimization problem[26,27]. The cost function in this optimization task is given by the squared Euclidean distance. By default in WOT, we conducted the unbalanced transport optimization with parameters ε=0.05, $\lambda 1$=1, and $\lambda 2$=50, incorporating a single iteration of growth rate learning[27]. The rationale and stability of these parameters are discussed in the WOT method and have been validated in the quantification and statistical analysis section of the original paper[27].

We then combined the OT maps predicted from every pair of adjacent time points , drawing from the time-series scRNA-seq data. This integration enabled us to generate a comprehensive probability map that traces each cell's likelihood of being an ancestor at earlier time points and transitioning to one of the terminal cell states following TGF-beta treatment. Building on this complete map, we summed up all probabilistic trajectories at the single-cell resolution that culminate in different fate subpopulations. As a result, we calculated the ancestor-to-fate (ATF) distributions for each early time point, specifying the likelihood of each ancestor differentiation into distinct fate subgroups.

Visualizing divergence of fates with barycentric coordinates

For visual representation of ATF distributions, we employed the tmap_model.fates() function from WOT[27]. Within this visualization approach, cell states are represented by a three-dimensional vector (p_1, p_2, p_3) adhering to the condition $p_1 + p_2 + p_3 = 1$. This vector characterizes the ATF distribution of a cell state transitioning into the orange, green, and blue fates. Notably, by day 8, all cells within a particular cluster show full commitment to a specific fate, as evidenced by a value of 1 in the corresponding position; for instance, cells within the orange cluster are denoted by the vector (1, 0, 0). We visualized these three-dimensional probability vectors by projecting them onto a two-dimensional triangle using barycentric coordinate representation. In this coordinate system, each vertex of the triangle is associated with a specific fate, and the cells are positioned based on their relative transition probabilities. We thus designated (a, b, c) as the vertices of the triangle on the 2D plane. The projection of the vector (p_1, p_2, p_3) onto this triangle utilized the convex combination of the vertices' locations (a, b, c) with the coefficients (p_1, p_2, p_3), resulting in the location given by $p_1*a + p_2*b + p_3*c$. Given that $p_1 + p_2 + p_3 = 1$, each probability vector corresponds to a point within the triangle. Cells situated at the triangle's center, represented by p = (1/3, 1/3, 1/3), denote undetermined states, suggesting an equal likelihood of differentiating into any of the fates. Conversely, cells closer to the vertices indicate a strong commitment to transition into the fate associated with the corresponding vertex.

We adopted a 75% commitment as a threshold for transitioning into a fate, given its conventional use as an upper quantile indicative of high probability. For example, for the orange fate, this threshold is symbolized as p= (0.75, p_2, p_3) and is represented as a line delineating a region within the upper right corner with the vertex (1,0,0). Within this area, all cells display over 75% commitment to transition into the blue fate subpopulation by day 8. This series of barycentric plots dynamically portrays the temporal

evolution of major ancestral contributions, providing an intuitive visualization of our predictions of cell fate trajectories over time. To ensure robustness of our findings, we investigated the effects of adjusting the threshold value upwards to 90% in 3% increments by reanalyzing our data for each threshold. We set 90% as the highest threshold due to the scarcity of cells beyond this point, particularly in early time points. For instance, at a 91% threshold on day 0, there were only 2 top ancestor cells for the medium EMT fate and 1 top ancestor cell for the high EMT fate. In Supplement S1, we present the outcomes corresponding to these varied thresholds, emphasizing that our conclusions remain consistent irrespective of the specific threshold utilized.

Validating trajectories with geodesic interpolation

To assess the reliability of the WOT-inferred cellular trajectories, we employed a cross-validation approach as outlined in the "validation of geodesic interpolation" section in the WOT methods[27]. In this approach, for every test data time point, its adjacent time points were used as training data, creating a series of consecutive time point triplets. Utilizing WOT's built-in function, wot.graphics.plot_ot_validation_summary_stats, we then computed the Wasserstein 2-distance between distributions[70]. This distance measures the least amount of "cost" required to transform one distribution into another, with "cost" denoted by the squared Euclidean distance of moving a unit mass within the distribution. We thus obtained the deviation, using the Wasserstein 2-distance, between the inferred and the actual midpoint test data. This deviation is depicted by the blue curve in Fig S3C. As per WOT's default settings, our results were benchmarked against three null models: 1) the Wasserstein distance between the test data and the data from the preceding day (depicted by green curves); 2) the Wasserstein distance between the test data and the data from the subsequent day (illustrated by purple curves); and 3) the Wasserstein distance between the test data and the estimated data from the interpolation based on a random map (highlighted by orange curves). In the random map model, WOT assumed all cell state transitions to be equally probable. Compared to these null models, our inferred distributions consistently exhibited the smallest deviations at most time points. For instance, the WOT-interpolated deviation values for days 1-4 are 8.16, 6.5, 6.85, and 6.85, which are smaller than the random coupling interpolation values of 8.85, 7.78, 7.08, and 7.4 for the same days (Fig S3C). Only on day 4 did our interpolation's deviation surpass that of the day 4 to day 8 data comparison (6.85 vs 5.47). This outcome may be due to the near completion of cell transition by day 4 (Fig 1A).

Assessing EMT Scores: The 76 GS 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[33]. In the first method, the 76 GS method[30,32], 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[31]. 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.*[47], specifically designed to distinguish between stemness and mature cell signatures by investigating mammary stem and luminal cells. Additionally, we employed the proliferation gene set from Ben-Porath *et al.*[48]. 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. For signaling pathway analysis, we investigated the gene sets of hypoxia, G2M, mitotic spindle, TGF-beta, PI3K-AKT-mTOR, Wnt, and IL6-JAK-STAT3 hallmarks from MSigDB[68,69]. We performed single-sample Gene Set Enrichment Analysis (ssGSEA) on all gene sets using GSEAPY (v1.0.4), a Python package[71]. The enrichment scores for gene sets were transformed into z-scores, with adjustments made by shifting the mean and normalizing by the standard deviation.

Inference of average expression levels of cellular/genetic signatures in the diverse EMT trajectories

To determine average signature expressions across different trajectories, we employed the update_trends_vis() function from WOT, as detailed in Schiebinger *et al*[27]. Specifically, this function requires a single trajectory, identified as described in the "Inferring trajectories with Waddington-OT" section, along with signature scores for each individual cell. These scores were determined either by calculating the cellular signature scores as described above, or by using gene expression scores extracted from individual levels within the gene expression matrix, as elaborated in the “scRNA-Seq Data Analysis” section. The update_trends_vis() function subsequently computes the average scores for each distinct EMT trajectory at every time point by calculating the cumulative score for each cell state weighted by the probability of that cell state serving as an ancestor for each of the three fates. By iterating this computation for data from each time point, this method reconstructs the varied trajectories of score dynamics, with each path pointing towards a unique fate (Fig 3A and Fig S4-6). To elucidate the correlations between different signatures, we paired various score trajectories and projected them onto a 2D plane (Fig 3C and Fig S7B).

Quantitative analysis and visualization of top ancestor signature distributions

For each top ancestor group, we visualized the distribution of cellular signatures using violin plots created with the seaborn tool[72]. Additionally, we used Seaborn's scatter plots to study the joint distributions of various cellular signature pairs within each of the three top ancestor groups and the undetermined ancestors (Fig 4D). Utilizing the pandas library[73], we computed a comprehensive set of relevant statistics, including means, variances, and t-tests for both the means and variances, based on the cellular signature scores of the top ancestor sets (Tables S4, S6, and S7). In our assessment of variances for the EMT signature score, we considered both the inclusion and exclusion of outliers. These outliers were identified as cells that positioned outside the 1st to 99th percentile range. Additionally, we employed the Series.corr function from pandas to determine Pearson correlations from pairwise cellular signature scores within each top ancestor group (Table S8).

Cellular transcriptomic heterogeneity measured by pairwise distances

To examine the variability within cellular transcriptomes, we computed the pairwise distance between each cell state based on the normalized RNA expression. Adapting the methodology from Hinohara *et al*[44], we obtained the mean pairwise distance by summing these distances and dividing by the total number of cell state pairs. However, diverging from Hinohara et al’s approach, this distance was weighted by multiplying the probabilities of the cellular states in each pair, guided by ATF distributions. This approach yields distinct expected pairwise distances that correspond to the specific ATF distribution for each fate at every time point. These pairwise distances and their associated probabilistic weights are also used to calculate the variances.

These heterogeneity computations were carried out based on different RNA expression spaces defined by the gene sets we used to define heterogeneity. We used the full gene set, provided by all genes in the scRNAseq dataset, the EMT gene set, and the the DEG set identified from comparisons between the high and medium EMT fate populations (Methods, differential expression analysis). For consistency in dimensionality, given that we used 72 genes in the EMT gene set (the original set had 76 genes but only 72 overlapped with the genes used in our scRNAseq data), we performed Principal Component Analysis (PCA) to select the 72 principal components for the distance in space of the full gene set. We also selected the 72 top-ranked genes in the differentially expressed gene sets (Table S1). As a validation, we also calculated the pairwise distance in the original dimension (15,000) of the full gene set to ensure consistency with the distance measured in the 72-dimensional space derived from PCA (Fig S9 and Table S3).

Differential expression analysis

In order to identify differentially expressed genes (DEGs) between the partial EMT and high EMT fates, we compared the expression level of each gene between these two cell populations using the stats.ttest_ind() function from scipy[74] to conduct a t-test between the groups, producing p-values and t-statistics. Additionally, we calculated the fold change between the mean gene expression levels from one group to the other. Drawing from the principle of dimensional consistency highlighted in the previous section, we curated two distinct gene sets, each with the top 36 genes. This amounts to a total of 72 genes in the DEG set, consistent with the size of the EMT gene set. These genes are significantly differential by their heightened expression levels in either the medium EMT or high EMT states (Table S1). We followed two criteria for gene selection: (1) selecting the smallest 36 p-values with a fold change greater than 100, and (2) choosing the largest 36 fold-changes with p-values smaller than 1e-7 (Table S1). By applying these two criteria, we obtained two sets of genes (Table S1) for our downstream analysis. Importantly, all conclusions drawn from both sets of genes were consistent. The main text presents the results derived from the first gene set, while detailed results from the second gene set can be found in Supplementary Fig S9 and Table S2. For inferring early DEGs between the partial EMT and complete EMT trajectories, we employed the wot.tmap.diff_exp() function from the WOT package[27]. This function calculates the differential expression of each gene across two cell sets, which includes all cell states collected at a specific time point. The differentiating factor between these two cell sets are the weights assigned to their cell states, which were determined based on their probabilities of transitioning into the corresponding fates. Notably, by day 8, when the cell states are determined by their fates, these weights take on a binary value (either 1 or 0), aligning with the conventional approach to identifying DEGs. For the dataset from the CRISPR knockout study, where all cell states are definitively determined and have been preprocessed by Seurat v3.61[65], the findMarkers() function from the Seurat package in R was used.

Code availability

All code used to process data and generate figures is available on a public Github repository at https://github.com/Michorlab/OT-EMT.

Acknowledgements

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.

Competing interests

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.

References

1. Hay, E. D. An overview of epithelio-mesenchymal transformation. *Acta Anat. (Basel)* **154**, 8–20 (1995).
2. Thiery, J. P. & Sleeman, J. P. Complex networks orchestrate epithelial–mesenchymal transitions. *Nat. Rev. Mol. Cell Biol.* **7**, 131–142 (2006).
3. Nieto, M. A., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* **166**, 21–45 (2016).
4. Dongre, A. & Weinberg, R. A. New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. *Nat. Rev. Mol. Cell Biol.* **20**, 69–84 (2019).
5. Wendt, M. K., Taylor, M. A., Schiemann, B. J. & Schiemann, W. P. Down-regulation of epithelial cadherin is required to initiate metastatic outgrowth of breast cancer. *Mol. Biol. Cell* **22**, 2423–2435 (2011).
6. Wang, Y. & Zhou, B. P. Epithelial-mesenchymal Transition---A Hallmark of Breast Cancer Metastasis. *Cancer Hallm.* **1**, 38–49 (2013).
7. Lu, M., Jolly, M. K., Levine, H., Onuchic, J. N. & Ben-Jacob, E. MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination. *Proc. Natl. Acad. Sci.* **110**, 18144–18149 (2013).
8. Zhang, L., Zhou, F. & ten Dijke, P. Signaling interplay between transforming growth factor-β receptor and PI3K/AKT pathways in cancer. *Trends Biochem. Sci.* **38**, 612–620 (2013).
9. Roche, J. The Epithelial-to-Mesenchymal Transition in Cancer. *Cancers* **10**, 52 (2018).
10. Vasaikar, S. V. *et al.* EMTome: a resource for pan-cancer analysis of epithelial-mesenchymal transition genes and signatures. *Br. J. Cancer* **124**, 259–269 (2021).
11. Taipale, J. & Beachy, P. A. The Hedgehog and Wnt signalling pathways in cancer. *Nature* **411**, 349–354 (2001).

12. Katoh, Y. & Katoh, M. FGFR2-related pathogenesis and FGFR2-targeted therapeutics (Review). *Int. J. Mol. Med.* **23**, 307–311 (2009).

13. Moustafa, A.-E. A., Achkhar, A. & Yasmeen, A. EGF-receptor signaling and epithelial-mesenchymal transition in human carcinomas. *Front. Biosci.-Sch.* **4**, 671–684 (2012).

14. Espinoza, I. & Miele, L. Deadly crosstalk: Notch signaling at the intersection of EMT and cancer stem cells. *Cancer Lett.* **341**, 41–45 (2013).

15. Heldin, C.-H., Vanlandewijck, M. & Moustakas, A. Regulation of EMT by TGFβ in cancer. *FEBS Lett.* **586**, 1959–1970 (2012).

16. Cook, D. P. & Vanderhyden, B. C. Context specificity of the EMT transcriptional response. *Nat. Commun.* **11**, 2142 (2020).

17. Deshmukh, A. P. *et al.* Identification of EMT signaling cross-talk and gene regulatory networks by single-cell RNA sequencing. *Proc. Natl. Acad. Sci.* **118**, e2102050118 (2021).

18. Waddington, C. H. *The Strategy of the Genes*. (Routledge, 2014).

19. Ferrell, J. E. Bistability, Bifurcations, and Waddington's Epigenetic Landscape. *Curr. Biol.* **22**, R458–R466 (2012).

20. Mani, S. A. *et al.* The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. *Cell* **133**, 704–715 (2008).

21. Jolly, M. K. *et al.* Towards elucidating the connection between epithelial–mesenchymal transitions and stemness. *J. R. Soc. Interface* **11**, 20140962 (2014).

22. Mazutis, L. *et al.* Single-cell analysis and sorting using droplet-based microfluidics. *Nat. Protoc.* **8**, 870–891 (2013).

23. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).

24. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).

25. Villani, C. *Optimal Transport: Old and New*. (Springer Berlin Heidelberg, 2016).

26. Chizat, L., Peyré, G., Schmitzer, B. & Vialard, F.-X. Scaling algorithms for unbalanced optimal transport problems. *Math. Comput.* **87**, 2563–2609 (2018).

27. Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 928-943.e22 (2019).

28. Lundgren, K., Nordenskjöld, B. & Landberg, G. Hypoxia, Snail and incomplete epithelial–mesenchymal transition in breast cancer. *Br. J. Cancer* **101**, 1769–1781 (2009).

29. Zhang, Y. *et al.* Genome-wide CRISPR screen identifies PRC2 and KMT2D-COMPASS as regulators of distinct EMT trajectories that contribute differentially to metastasis. *Nat. Cell Biol.* **24**, 554–564 (2022).

30. Byers, L. A. *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).

31. Tan, T. Z. *et al.* Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.* **6**, 1279–1293 (2014).

32. Guo, C. C. *et al.* Dysregulation of EMT Drives the Progression to Clinically Aggressive Sarcomatoid Bladder Cancer. *Cell Rep.* **27**, 1781-1793.e4 (2019).

33. Chakraborty, P., George, J. T., Tripathi, S., Levine, H. & Jolly, M. K. Comparative Study of Transcriptomics-Based Scoring Metrics for the Epithelial-Hybrid-Mesenchymal Spectrum. *Front. Bioeng. Biotechnol.* **8**, (2020).

34. Bao, B. *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).

35. Emami Nejad, A. *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).

36. Massagué, J., Blain, S. W. & Lo, R. S. TGFβ Signaling in Growth Control, Cancer, and Heritable Disorders. *Cell* **103**, 295–309 (2000).

37. Donovan, J. & Slingerland, J. 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).

38. Zhang, Y., Alexander, P. B. & Wang, X.-F. TGF-β Family Signaling in the Control of Cell Proliferation and Survival. *Cold Spring Harb. Perspect. Biol.* **9**, a022145 (2017).

39. Aban, C. E. *et al.* Downregulation of E-cadherin in pluripotent stem cells triggers partial EMT. *Sci. Rep.* **11**, 2048 (2021).

40. Aggarwal, V. *et al.* P4HA2: A link between tumor-intrinsic hypoxia, partial EMT and collective migration. *Adv. Cancer Biol. - Metastasis* **5**, 100057 (2022).

41. Yao, Z. *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.* **107**, 15535–15540 (2010).

42. Akhmetshina, A. *et al.* Activation of canonical Wnt signalling is required for TGF-β-mediated fibrosis. *Nat. Commun.* **3**, 735 (2012).

43. Carpenter, C. L. & Cantley, L. C. Phosphoinositide 3-kinase and the regulation of cell growth. *Biochim. Biophys. Acta BBA - Rev. Cancer* **1288**, M11–M16 (1996).

44. Hinohara, K. *et al.* KDM5 Histone Demethylase Activity Links Cellular Transcriptomic Heterogeneity to Therapeutic Resistance. *Cancer Cell* **34**, 939-953.e9 (2018).

45. Tyler, M. & Tirosh, I. Decoupling epithelial-mesenchymal transitions from stromal profiles by integrative expression analysis. *Nat. Commun.* **12**, 2592 (2021).

46. Elenbaas, B. *et al.* Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. *Genes Dev.* **15**, 50–65 (2001).

47. Lim, E. *et al.* Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast Cancer Res.* **12**, R21 (2010).

48. Ben-Porath, I. *et al.* An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nat. Genet.* **40**, 499–507 (2008).

49. Bierie, B. *et al.* Integrin-β4 identifies cancer stem cell-enriched populations of partially mesenchymal carcinoma cells. *Proc. Natl. Acad. Sci.* **114**, E2337–E2346 (2017).

50. Saini, M. *et al.* Resistance to mesenchymal reprogramming sustains clonal propagation in metastatic breast cancer. *Cell Rep.* **42**, 112533 (2023).

51. Giancotti, F. G. & Tarone, G. Positional Control of Cell Fate Through Joint Integrin/Receptor Protein Kinase Signaling. *Annu. Rev. Cell Dev. Biol.* **19**, 173–206 (2003).

52. Spaderna, S. *et al.* A Transient, EMT-Linked Loss of Basement Membranes Indicates Metastasis and Poor Survival in Colorectal Cancer. *Gastroenterology* **131**, 830–840 (2006).

53. Raser, J. M. & O'Shea, E. K. Noise in Gene Expression: Origins, Consequences, and Control. *Science* **309**, 2010–2013 (2005).

54. Schmalhofer, O., Brabletz, S. & Brabletz, T. E-cadherin, β-catenin, and ZEB1 in malignant progression of cancer. *Cancer Metastasis Rev.* **28**, 151–166 (2009).

55. Bocci, F. *et al.* Toward understanding cancer stem cell heterogeneity in the tumor microenvironment. *Proc. Natl. Acad. Sci.* **116**, 148–157 (2019).

56. Feinberg, A. P. & Levchenko, A. Epigenetics as a mediator of plasticity in cancer. *Science* **379**, eaaw3835 (2023).

57. Francesconi, M. *et al.* Single cell RNA-seq identifies the origins of heterogeneity in efficient cell transdifferentiation and reprogramming. *eLife* **8**, e41627 (2019).

58. Deng, Y. *et al.* Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* **609**, 375–383 (2022).

59. Ranzoni, A. M. *et al.* Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. *Cell Stem Cell* **28**, 472-487.e7 (2021).

60. Massri, A. J. *et al.* Developmental single-cell transcriptomics in the Lytechinus variegatus sea urchin embryo. *Development* **148**, dev198614 (2021).

61. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).

62. McFaline-Figueroa, J. L. *et al.* A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nat. Genet.* **51**, 1389–1398 (2019).

63. Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).

64. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).

65. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

66. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE* **9**, e98679 (2014).

67. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

68. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

69. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

70. Ambrosio, L., Gigli, N. & Savare, G. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. (Birkhäuser, 2008).

71. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023).

72. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

73. McKinney, W. Data Structures for Statistical Computing in Python. in 56–61 (2010). doi:10.25080/Majora-92bf1922-00a.

74. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

Figures

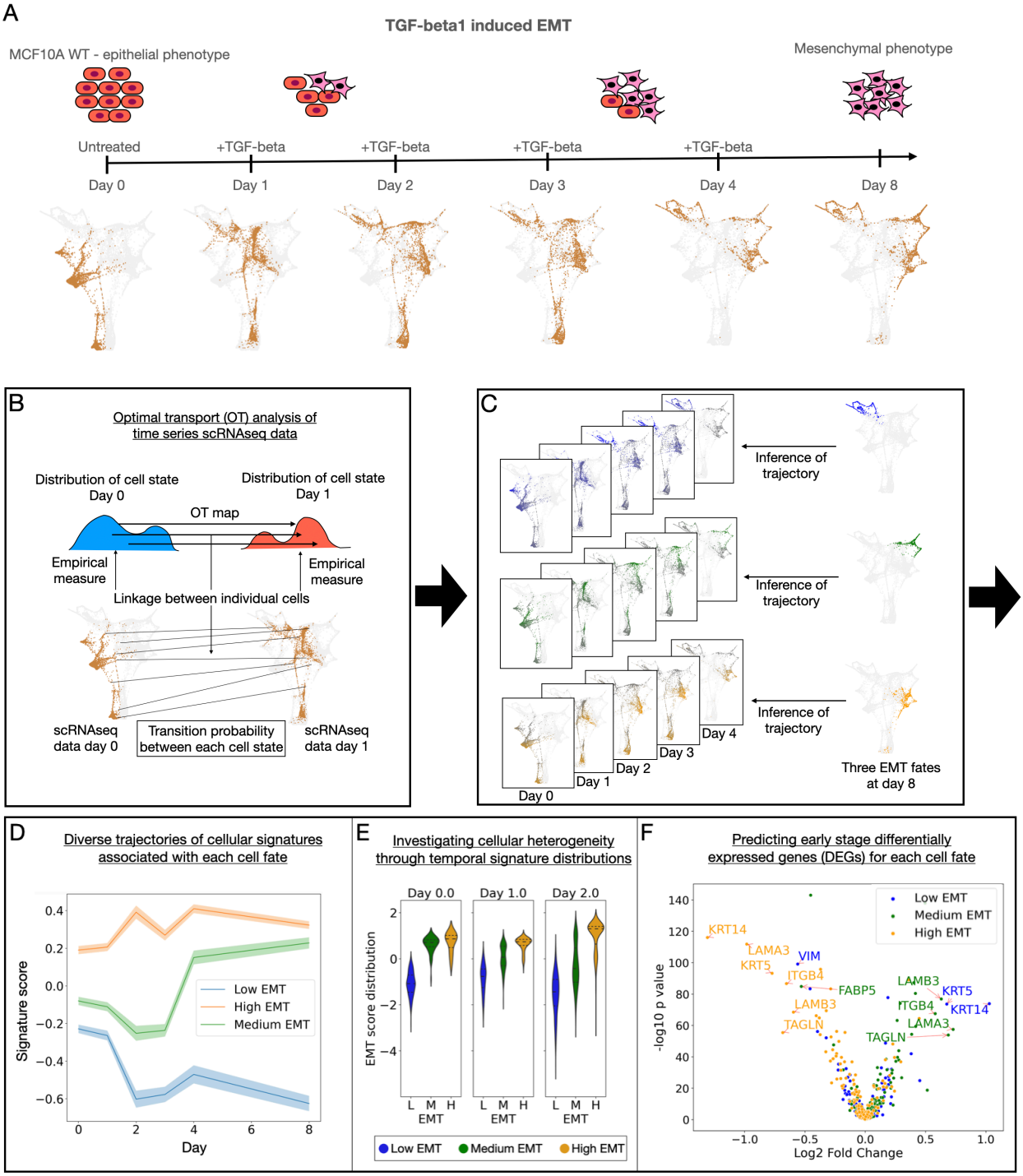


Figure 1. Optimal transport analysis of scRNAseq time course data of the epithelial mesenchymal transition. **(A)** Schematic representation of the experiments performed to induce EMT in MCF10A cells, accompanied by a time series of single-cell RNA sequencing data visualized using force-directed layout embedding (FLE). 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 between adjacent time points within the scRNAseq

datasets. **(C)** From the entire consecutive time-series scRNAseq data, transition probabilities, inferred by OT, were integrated to determine the likelihood of each early cell state acting as an ancestor for the three fate subpopulations. **(D-F)** Three downstream trajectory inference analyses are depicted from left to right: **(D)** reconstruction of diverse cellular signature trajectories, **(E)** exploration of cellular heterogeneity across these diverse trajectories, and **(F)** differential analysis of early gene expression in ancestral cells associated with each distinct cell fate.

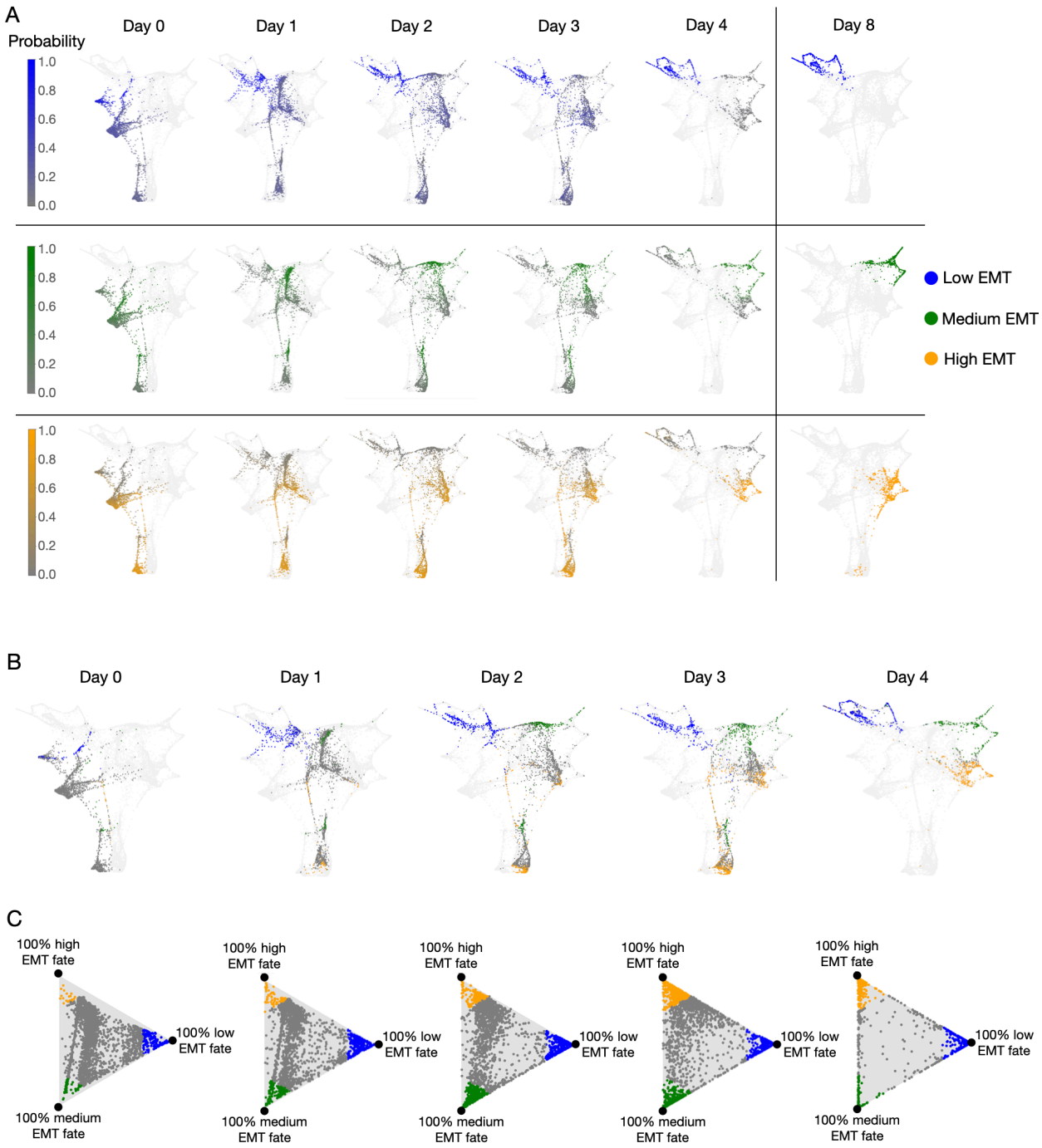


Figure 2. Optimal transport recovers diverse trajectories of EMT. (A) The colormap presents the inferred ATF distributions, showcasing the probability of early cell states (from day 0 to day 4) serving as ancestors for the three fate subpopulations by day 8. **(B)** Early cell state classification via ATF distributions: blue cells have over 75% probability of becoming cells with the low EMT fate; green cells have over 75% probability of becoming the medium EMT fate; orange cells have over 75% probability of becoming the high EMT fate; and gray cells are not highly committed to any of the fates. **(C)** Barycentric coordinate projection visualizes ATF distributions: this coordinate system represents the three-dimensional simplex within a two-dimensional space. 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 (Methods). A position at one of the triangle's vertices indicates a 100% commitment of the cell state to the corresponding fate.

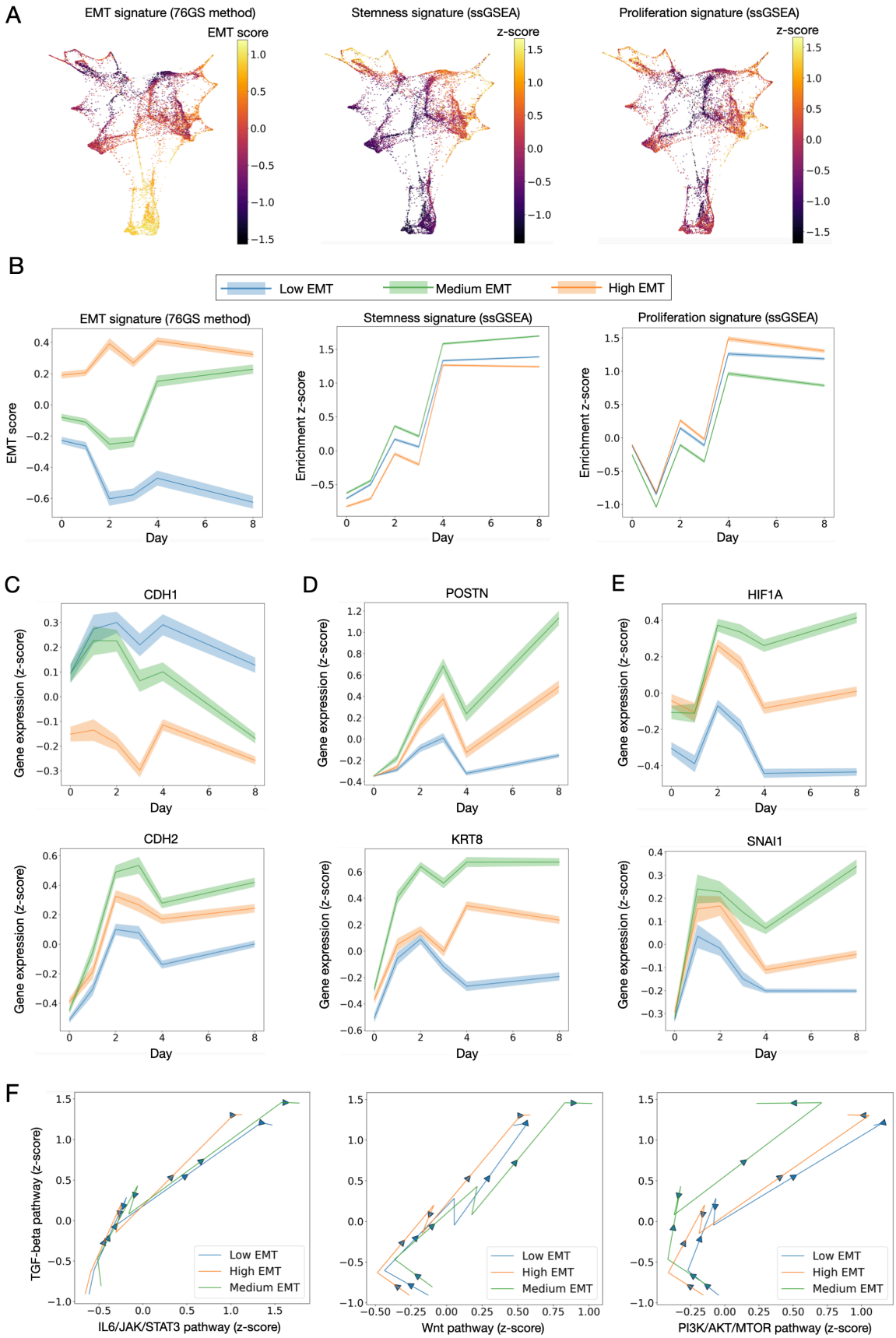

A
EMT signature (76GS method)
EMT score
Stemness signature (ssGSEA)
z-score
Proliferation signature (ssGSEA)
z-score
B
Low EMT
Medium EMT
High EMT
EMT signature (76GS method)
EMT score
Day
Stemness signature (ssGSEA)
Enrichment z-score
Day
Proliferation signature (ssGSEA)
Enrichment z-score
Day
C
CDH1
Gene expression (z-score)
Day
D
POSTN
Gene expression (z-score)
Day
E
HIF1A
Gene expression (z-score)
Day
CDH2
Gene expression (z-score)
Day
KRT8
Gene expression (z-score)
Day
SNAI1
Gene expression (z-score)
Day
F
TGF-beta pathway (z-score)
IL6/JAK/STAT3 pathway (z-score)
Low EMT
High EMT
Medium EMT
Wnt pathway (z-score)
Low EMT
High EMT
Medium EMT
PI3K/AKT/MTOR pathway (z-score)
Low EMT
High EMT
Medium EMT

Figure 3. OT analysis reveals unique cellular signatures across distinct EMT trajectories. (A) **(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. The p-values, generated via paired-sample *t*-tests for each pairwise comparison at individual time points, are all < 0.005. Shaded regions denote the 95% confidence intervals. **(C-E)** Temporal evolution of mean gene expression, from day 0 to day 8, across the three EMT trajectories. Shaded regions denote the 95% confidence intervals (C) for CDH1and CDH2 genes, (D) for HIF1A and SNAI1 genes, and (E) for POSTN and FN8 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.

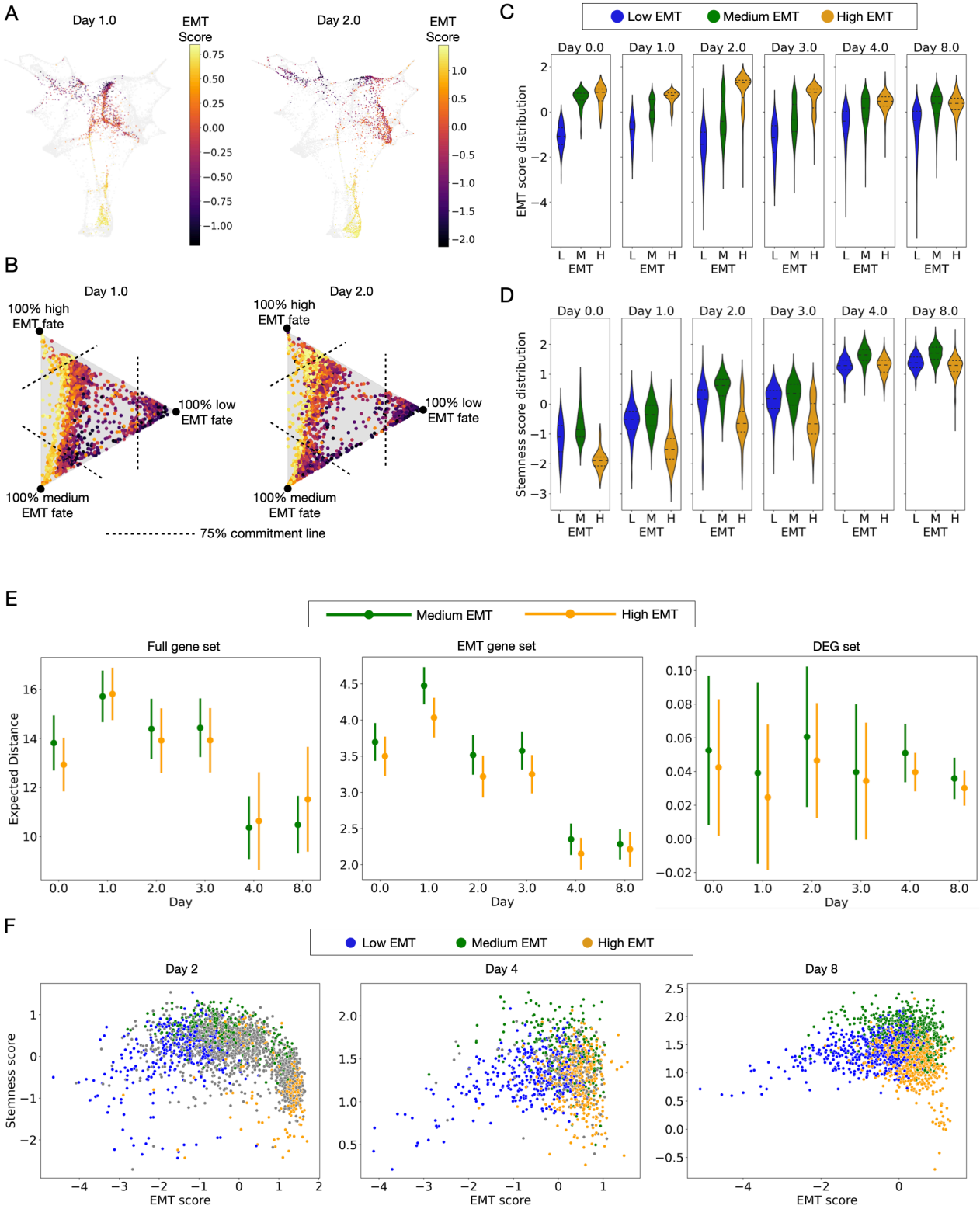
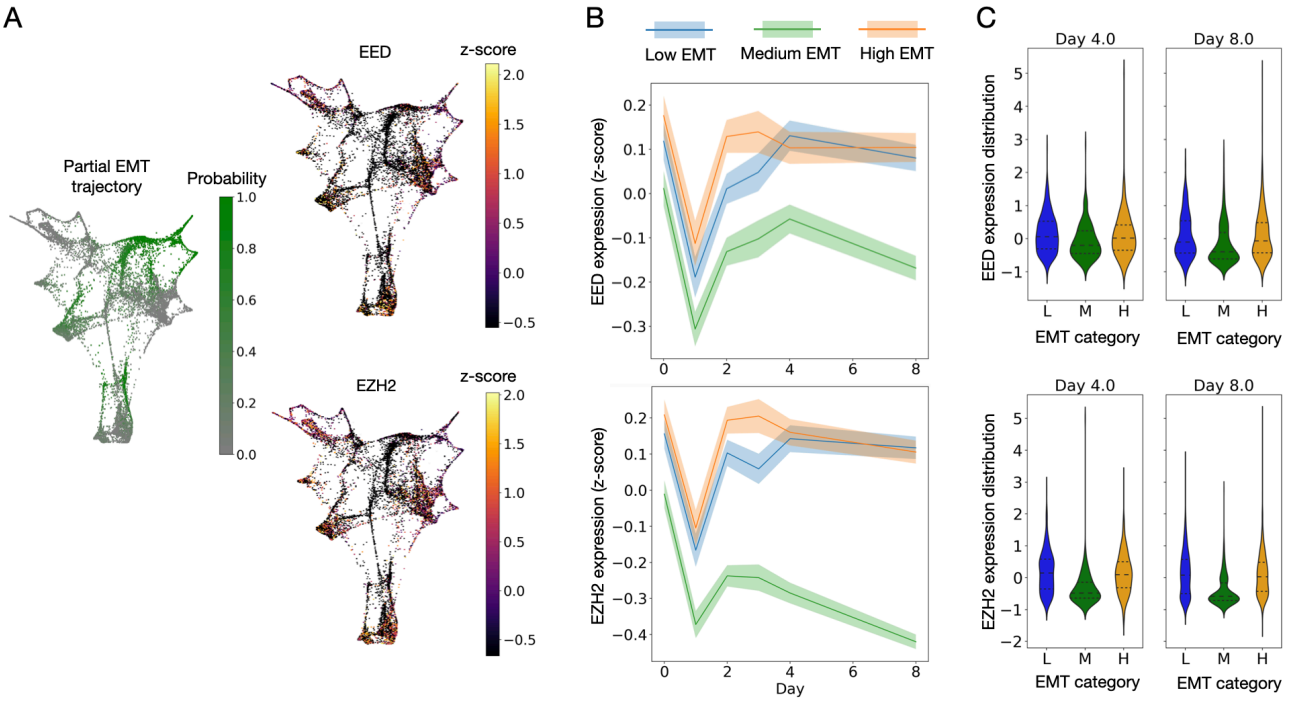


Figure 4. Tracing cellular signature variations across three EMT trajectories. (A) EMT signature scores for cell states from days 1 and 2 (for the complete time course see Supplementary Fig 8). **(B)** EMT signature scores from (A) are paired with ATF distributions and plotted within a triangle using

barycentric coordinates. As in Fig 2C, a point's location represents its ATF distribution, indicating its likelihood of differentiating into the associated fate. Concurrently, the color map showcases the EMT score. Dashed lines demarcate a 75% commitment to the fate linked to the corresponding triangle vertex. **(C-D)** Violin plots depict the distribution of each cellular signature score for the top ancestors of each fate, color-coded in blue for low, green for medium, and orange for high EMT trajectories (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 standard deviation errors of these pairwise distances (symbolized by the length of the error bars). The low EMT trajectory was not included due to the high number of outliers, as explained in the main text and Supplementary S1. These distances are derived from the gene expression space defined by the pertinent gene sets: complete gene set, EMT gene set, and DEGs. **(F)** Scatter plots display paired cellular signature scores for days 2, 4, and 8. Color codes designate the top ancestors for each trajectory.



D

DEGs of partial EMT / high EMT (day 2)

Up-regulation

	P-value	Fold change
TGFBI	7.66e -125	1.53
KRT8	1.24e -90	1.41
DSP	2.53e -74	1.37
CDH1	3.20e -61	1.15
POSTN	3.59e -05	1.07
CDH2	2.00e -09	1.06
KMT2C	2.71e -05	1.03
EPCAM	1.67e -05	1.03
SNAI1	3.44e -01	1.01
TWIST1	7.56e -05	0.97
ZEB1	1.07e -11	0.96
EZH2	2.84e -59	0.88
SUZ12	3.97e -43	0.87
PHF19	3.43e -192	0.66

Down-regulation

E

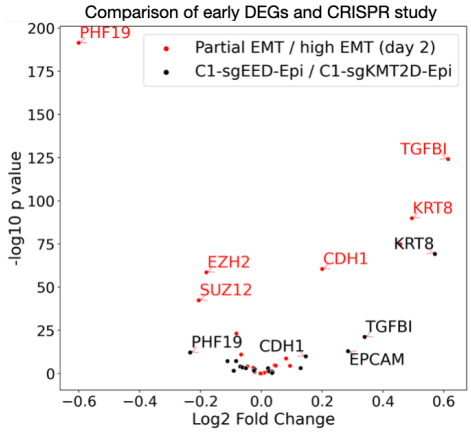
DEGs of C1-sgEED-Epi / C1-sgKMT2D-Epi

Up-regulation

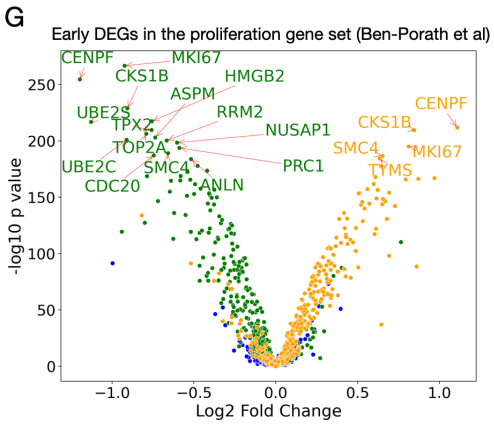
	P-value	Fold change
KRT8	6.13e -70	1.49
TGFBI	6.66e -22	1.27
EPCAM	1.22e -13	1.29
CDH1	1.19e -10	1.11
DYP30	7.49e -04	1.09
SNAI1	8.86e -04	1.02
DSP	2.51e -02	0.94
EZH1	1.93e -02	0.98
WDR5	7.89e -04	0.98
KMT2C	5.64e -04	0.97
SUZ12	2.70e -04	0.96
JARID2	6.81e -08	0.93
TWIST1	6.30e -08	0.94
PHF19	6.61e -13	0.85

Down-regulation

F



G



H

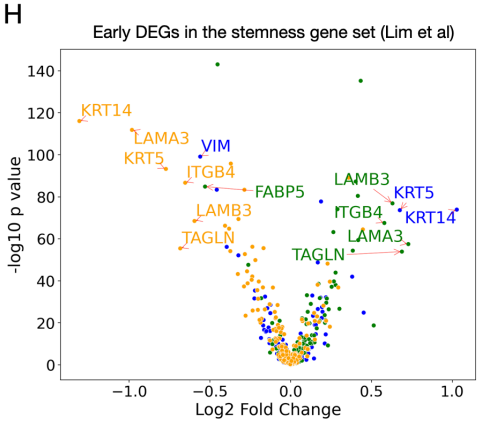


Figure 5. Early predictors of EMT fate through early DEGs analysis and CRISPR knock-out screening. (A-C) EED and EZH2 gene expression analysis. **(A)** Color maps display the ATF distributions for the partial EMT trajectory alongside the expression levels of the EED/EZH2 genes across all cellular states 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/EZH2 gene expression over. Shaded regions denote the 95% confidence intervals. **(C)** Violin plots depict day-to-day distribution of expression levels for EED/EZH2 gene among the top ancestor cells. **(D-E)** Two tables presenting the DEGs of EMT-related genes for two distinct datasets. Table (D) depicts our predicted DEGs at day 2 between the partial EMT and high EMT trajectories derived from the Deshmukh *et al*. dataset. Table (E) illustrates the DEGs in epithelial cells with EED knock-out (C1-sgEEG-Epi) compared to KMT2D knock-out cells (C1-sgKMT2D-Epi) using data from the Zhang *et al*. CRISPR study. **(F)** Volcano plot of the differential gene expression analysis conducted on the two datasets, as outlined in tables (D) and (E). Red dots indicate our early phase gene expression predictions, while black dots represent results from the CRISPR knockout study. **(G-H)** Early DEGs of proliferation related genes (G) and stemness related genes (H). 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. As an example, green indicates genes with differential expression in the partial EMT trajectory compared to the joint low and high EMT trajectories. While the shown instance is based on day 2 data, comprehensive results spanning days 1-3 can be found in Supplementary Fig S12.