*Published online 21 September 2022*

*NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 3* **1**
*https://doi.org/10.1093/nargab/lqac072*

# Comparative single-cell transcriptomes of dose and time dependent epithelial–mesenchymal spectrums

**Nicholas Panchy[1,†], Kazuhide Watanabe[2,*,†], Masataka Takahashi[2], Andrew Willems[3] and Tian Hong[1,4,*]**

[1]Department of Biochemistry & Cellular and Molecular Biology. The University of Tennessee, Knoxville, Knoxville, TN 37996, USA, [2]RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan, [3]School of Genome Science and Technology, The University of Tennessee, Knoxville, Knoxville, TN 37916, USA and [4]National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996, USA

Received June 19, 2022; Revised August 17, 2022; Editorial Decision August 29, 2022; Accepted August 31, 2022

## ABSTRACT

**Epithelial–mesenchymal transition (EMT) is a cellular process involved in development and disease progression. Intermediate EMT states were observed in tumors and fibrotic tissues, but previous *in vitro* studies focused on time-dependent responses with single doses of signals; it was unclear whether single-cell transcriptomes support stable intermediates observed in diseases. Here, we performed single-cell RNA-sequencing with human mammary epithelial cells treated with multiple doses of TGF-β. We found that dose-dependent EMT harbors multiple intermediate states at nearly steady state. Comparisons of dose- and time-dependent EMT transcriptomes revealed that the dose-dependent data enable higher sensitivity to detect genes associated with EMT. We identified cell clusters unique to time-dependent EMT, reflecting cells *en route* to stable states. Combining dose- and time-dependent cell clusters gave rise to accurate prognosis for cancer patients. Our transcriptomic data and analyses uncover a stable EMT continuum at the single-cell resolution, and complementary information of two types of single-cell experiments.**

## INTRODUCTION

Epithelial-mesenchymal transition (EMT) is a cellular process in which epithelial (E) cells undergo fate switches towards mesenchymal (M) types. This process renders the loss of apical-basal polarity and the gain of migratory properties. EMT plays crucial roles in development and disease progressions such as metastasis and fibrosis (1,2). EMT is not a binary process. In tumor cells, for example, intermediate (partial) EMT states were observed (3–5), and it was suggested that there is an association between intermediate EMT states and metastatic potentials (6,7). Interestingly, intermediate EMT states can also be observed *in vitro* with epithelial cell lines treated with EMT signals, such as TGF-β (3,8), and these *in vitro* experiments provide useful insights into molecular programs underlying partial EMT (9). For example, experiments with genetically perturbed cells have suggested that interconnected feedback loops in gene regulatory networks can generate multiple intermediate EMT states (10). Additionally, mathematical models postulated stability of these states arising from intricate gene regulatory networks (10–12).

At the fundamental level, intermediate EMT states can be understood as either cell states *en route* to M-like states, or those stable states induced by weak (low-dose) EMT signals in the microenvironment. Recent single-cell transcriptomic studies showed that the time-dependent EMT programs contain intermediate states that delineate a continuum-like EMT spectrum (13–15). However, it is unclear whether stable cell states in EMT program induced by multiple levels of signals support a continuum or a discrete EMT spectrum. While previous dose-dependent single-cell experiments with two EMT markers (E-cadherin for E, Vimentin for M) support the existence of intermediate EMT states (8,10), much less is known about the transcriptomic profiles of the dose-dependent EMT spectrum.

In this work, we performed single-cell RNA-sequencing (scRNA-seq) using human mammary epithelial (MCF10A) cells treated with multiple concentrations of TGF-β. We found that the dose-dependent EMT program is a continuum containing multiple intermediate states that are stable after two-week treatment of TGF-β. We performed integrated analyses with our dataset and a recent time-dependent scRNA-seq dataset for the same cell line and

*To whom correspondence should be addressed. Tel: +1 865 974 3089; Fax: +1 865 974 6306; Email: hongtian@utk.edu
Correspondence may also be addressed to Kazuhide Watanabe. Tel: +81 45 503 9222; Fax: +81 45 503 9216; Email: kazuhide.watanabe@riken.jp
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

EMT inducer (13) (Figure 1A). We found that the dose-dependent EMT spectrum has a stronger anti-correlation of E and M transcriptional programs than the time-dependent spectrum. While both spectrums show strong cell-to-cell variability and continuum-like patterns, the dose-dependent dataset has higher separability in terms of the groups of cells with neighboring labels (similar doses vs. similar time points). These differences enable higher sensitivity for the dose-dependent model to detect non-canonical EMT genes that are associated with the core EMT programs in terms of the expression pattern. Furthermore, the time-dependent dataset contains unique cell clusters at E-low region in the transcriptomic space, which correspond to *en route* cell states that do not appear at steady state. We found that signature genes in both dose- and time-enriched clusters are useful for prognostic predictions of cancer patients. Our analyses revealed key differences between dose- and time-dependent EMT programs in terms of the underlying dynamical processes, and showed the widespread existence of stable EMT continuum under multiple assumptions that may be relevant to physiological and pathological conditions.

## MATERIALS AND METHODS

### Cell culture

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

### Single-cell RNA-sequencing

MCF10A cells were first labelled with Perturb-seq vectors without sgRNA expression using guide barcodes (GBCs) that were originally used to identify sgRNAs (16). Barcoded MCF10A cells were then treated with different dosages of TGF-β for 14 days and single cells were prepared and mixed at a concentration of ∼1000 cell/μl. Transcriptome library generation was performed following the Chromium Single Cell 3′ Reagents Kits v2 (following the CG00052 Rev B. user guide) where we target 10 000 cells per sample for capture. GBC library was generated from a fraction (5 ng) of amplified whole transcriptome by dial-out PCR method according to a previous publication (16). Both libraries were mixed at 9:1 ratio and sequenced by paired-end sequencing (26 bp Read 1 and 98 bp Read 2) with a single sample index (8 bp) on the Illumina HiSeq 2500. Generated FASTQ files were aligned utilizing 10× Genomics Cell Ranger 2.1.0. Each library was aligned to an indexed hg38 genome using Cell Ranger Count. The cell barcode (CBC)-GBC table was generated from the GBC library and used to identify the treatment groups. We first performed doublet exclusion with standard 10× pipeline based on RNA amounts, and then excluded cell barcodes that were assigned with multiple GBCs (12.7% of all cells assigned with GBCs) (16). Nonetheless, we examined the 2402 unlabeled dose samples in the dose data set, which are roughly evenly distributed across UMAP and E- and M-score space (Supplemental Figure S1A) (see the method for computing the scores below). We also compared the average E- and M-scores of labeled and unbaled samples in each nearest neighbor cluster and found them to be similar (Supplemental Figure S1B) and excluding them from calculating E and M-scores showed little affect ($R > 0.999$ with original scores). These results suggest that the unlabeled cells are merely missing annotation rather than contamination or doublets, which are consistent with prior observations of GBC labeled data (16) and thus were retained for processing and scoreing. Additionally, following the identification of multiple putative end states in both dose and integrate data, we also looked at potential homotypic doublets with individual treatment samples. To do this, we used DoubletFinder (17,18). Using 2.4% for time data which targeted 3000 cells per treatment, and 8.0% for dose data which targeted 10 000 cells adjusted for the proportion of each labeled dose treatment in the dataset to estimate the homotypic rate, we found 90 putative homotypic doublets in dose data, which are broadly distributed in integrated data (Supplemental Figure S2A), and 332 in the time data, which showed some clustering around the edge of one cluster (Supplemental Figure S2B). However, excluding these clustered time doublets did not affect the pattern of separation of treatment samples, anti-correlation of E- and M-scores, or enrichment of time and dose samples in quartile clusters. Based on these observations, we chose not to filter these samples.

### Data processing and integration

Sequencing data for time-course single-cell data from Deshmukh *et al.* (13) was obtained from National Center for Biotechnology Information Sequence Read Archive (BioProject ID: PRJNA698642) and mapped using to the same human genome assembly as our dose (CRXh38.84) data using Cell Ranger (19). Aligned sequences we processed using the Seurat package (version 4.1) in R (20). Gene names between experiments were correlated using the HGNChelper package (21), using the suggested gene symbol for each gene except when it would create a duplicate reference. Genes were filtered from individual runs if they did not appear in three or more cells. We then filtered each dataset for cells with fewer than 500 features or more than three median absolute deviations beyond the median number of features of the sample set (i.e. the long time, short time, and dose datasets). We additionally eliminated any sample with a fraction of mitochondrial reads that was >0.2. We then integrated all time and dose data following the procedure used in Deshmukh et al. to best preserve the relationship between time samples observed in the study (13). Briefly, the integration involves identifying the most variable genes in each dataset, defining 'anchor' samples between datasets using canonical correlation analysis, correcting the expression of related anchor samples and finally propagating this correction to other samples based on the similarity to the anchors. We normalized and calculated cell cycles scores for each dataset independently prior to integration and used the 15 000 most variable genes to identify anchors. We applied the same top 15 000 variable genes filter to uninte-

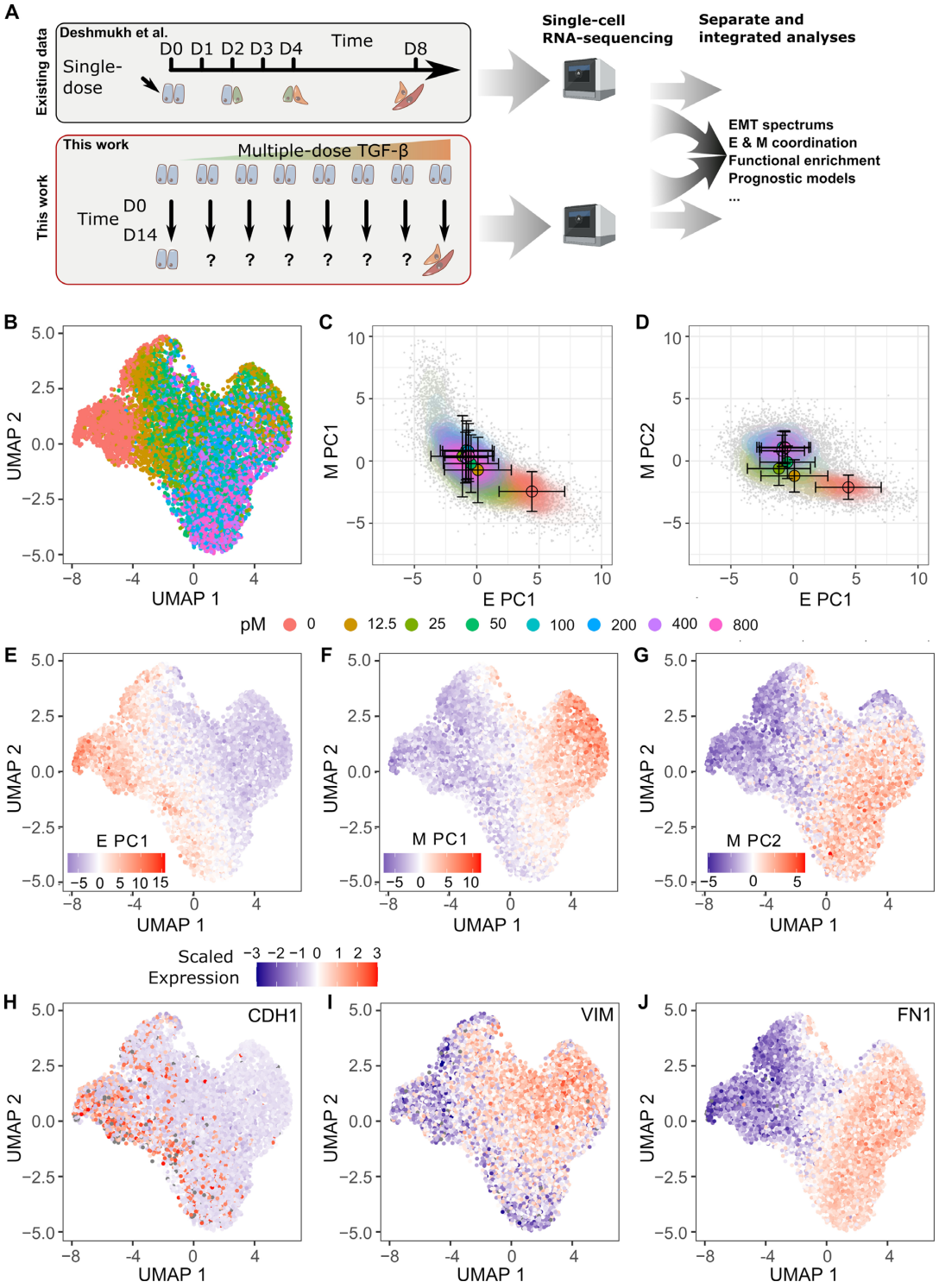


**Figure 1.** Analysis overview and progression of dose-dependent EMT at single-cell level. (**A**) A schematic of our analysis in this study. The analyses involve the existing time-course data (top) contain MCF10A cells at different time points following a common TGF-β treatment (about 200 pM; Deshmukh *et al.* (13)) as well as the dose-dependent data (bottom) containing MCF10A cells treated with different dosage levels of TGF-β after a fixed time period representing near-steady-state. Gene expression of cells from both experiments were measured using single-cell RNA-sequencing and were subsequently used individually and integrated for downstream analyses. (**B**) Projection of dose treatment single-cell expression data using UMAP. The color of individual points indicates the dose of TGF-β treatment from 0 pM (red) to 800 pM (pink). (**C, D**) Contour plots of gene set scores of E (x-axis) and M (y-axis) genes using M PC1 (C) and M PC2 (D) for M-scores. Color indicates the dose of TGF-β as in (B). Circles indicate the mean E- and M-score of samples from each dose point and the associated error bars show the standard deviation (**E–G**) Overlay of nnPCA-derived scores from the first E (E PC1, E), first M (M PC1, F) and second M (M PC2, G) principal components. The color of individuals points indicates the score from low (blue) to high (red). (**H–J**) Overlay of the scaled expression of EMT marker genes CDH1 (an epithelial marker, H), VIM (a mesenchymal marker, I), and FN1 (a highly expressed mesenchymal gene, J). The color of individual points indicates the *Z*-score of expression of each gene from low (blue) to high (red).

grated dose data for calculating UMAP, nnPCA scores and clustering. Finally, each dataset was scaled across genes and the expression was corrected for the total read count, percent of mitochondrial reads, and cell-cycle phases using S and G2M scores. The importance of correction for cell-cycle phase can be seen by contrasting the UMAP and nearest neighbor clustering of dose data with and without cell-cycle score correction (Supplemental Figure S3).

### Projection of single cell data in reduced dimensional space

Projection of single cell data was done using the scaled expression values for both unintegrated and integrated data. For this process we employed Uniform Manifold Approximation and Projection (UMAP) and non-negative Principal Component Analysis (nnPCA). For UMAP, we first performed principal component analysis (PCA). We then used the first 15 principal components to construct the map using the RunUMAP function from Seurat with defaults parameters beyond specifying the PCA input and size. UMAP was primarily used to assess the continuity of both unintegrated and integrated datasets. Average $Z$-score was determined using the gsva package (22). nnPCA-based scores were computed by performing non-negative principal component analysis on subsets of the scaled dataset defined using epithelial (E) and mesenchymal (M) genes identified by Tan *et al.* (23) and a list of ten additional EMT genes (GRHL1, GRHL3, OVOL1, FOXC2, ELF5, TWIST2, TCF3, GSC, KLF8 and SNAI1). We used the nsprcomp function from the R package of the same name with the option nneg = TRUE, ncomp = 25, em_maxiter = 10000, em_tol = 0.00001 and a consistent seed (set.seed (5)) to identify the top twenty five components for each subset (24). While the first twenty-five components were calculated, as the function greedily optimizes the variance explained by each component in order, we initially considered only the top E and M components and added the second M component when clustering suggested the possibility of multiple paths of progression. Further details about our nnPCA scoring approach can be found in Panchy *et al.* (25). In addition to E and M axes, we analyzed the relationship between EMT progression and PD-1 ligation. The latter was represented by the gene set M4534 from the Molecular Signature Database (26).

### Signature genes in EMT spectrum

To examine the distribution of time and dose samples, we divided E- and M-score space into a $4 \times 4$ grid based on the 25th, 50th and 75th percentile of E- and M-scores. To obtain results comparable to progression plots in Figure 2, enrichment only considered labeled time and dosage samples. Odds of the enrichment of time and dose samples in each segment of the $4 \times 4$ grid of E- and M-score space were calculated using the Fisher's exact test implemented in R (fisher.test). Differentially expressed marker genes for each segment were identified using the FindAllMarkers function from Seurat. As recommended, for FindAllMarkers we used unintegrated, normalized counts as integration introduces non-independence between samples. In FindAllMarkers, we used the logistic regression framework ('LR') so we could use the number reads per cell, percent mitochondrial genes and cell cycles scores as latent variables as well as the experiment of origin (dose, time long, and time short) to account for batch effect. We also manually corrected $P$-values to account for comparing all possible genes (i.e. those with non-zero expression in at least one sample including retesting EMT genes discussed below) across all clusters, as FindAllMarkers does not account for testing multiple clusters. We used the Bonferroni Correction as the standard to FindAllMarkers and applied the same cutoffs as the standard FindAllMarker approach (adjusted $P$-value $< 0.05$, absolute value average $\log_2$ fold-change $> 0.25$). The same approach was also used to compare VIM, FN1, EPCAM and FN1 across all samples, and between time and dose samples in the middle 50% of E and M scores, except we use the options 'logfc.threshold = 0, min.pt = 0, return.thresh = 1' to return all results. Additionally, batch correction was excluded for comparison of time and dose data. Finally, because EMT genes were used to construct cell clusters, raising a possible concern of a bias due to feature reuse, we iteratively removed each EMT gene from the list, recalculated EMT scores, and looked at differential expression for the removed gene in a clustering context agnostic to its influence. Overall, E and M scores were robust to the loss of any individual genes ($R > 0.99$), but cluster membership vary for marginal samples around the borders of clusters, so we filtered our EMT markers genes for significant differential expression in clusters agnostic to it. This led to the removal of 12, 14 and 33 EMT marker genes from all, time and dose clusters respectively, representing 1.7%, 1.8% and 6.7% of EMT marker genes. Notably we did not apply any average log fold-change threshold to gene agnostic results as a primary concern was biasing $P$-values due to feature reuse.

### GO enrichment

GO enrichment of identified genes set was performed using the clusterProfiler package in R (27), using a background of all genes in the integrated data, and significance was assessed using the Benjamini–Hochberg adjusted $P$-value with a cutoff of 0.05. Otherwise, we used default parameters which filters overly narrow (<10 genes) and overly board (>500 genes) sets as well as terms which do not overlap with the gene set of interest. Each combination of gene set and ontology (Biological Process, Cellular Component, and Molecular Function) were tested independently.

### Prognostic models

TCGA bulk RNA-seq data and associated meta-data were obtained from TCGAbiolinks (28). Of the 33 available datasets (cancer types), we selected those with at least ten patients in both the survivor and non-survivor groups, which gave rise to the 27 cancer types listed in Figure 5B. For each dataset, FPKM expression values were extracted from all non-normal tissue samples, filtered for duplicate samples from the same aliquot, and normalized to log plus 1 TPM values. We then calculated overall survival as our survival metric, which is defined as the time from date of diagnosis to the date of death (for non-survivors) or date

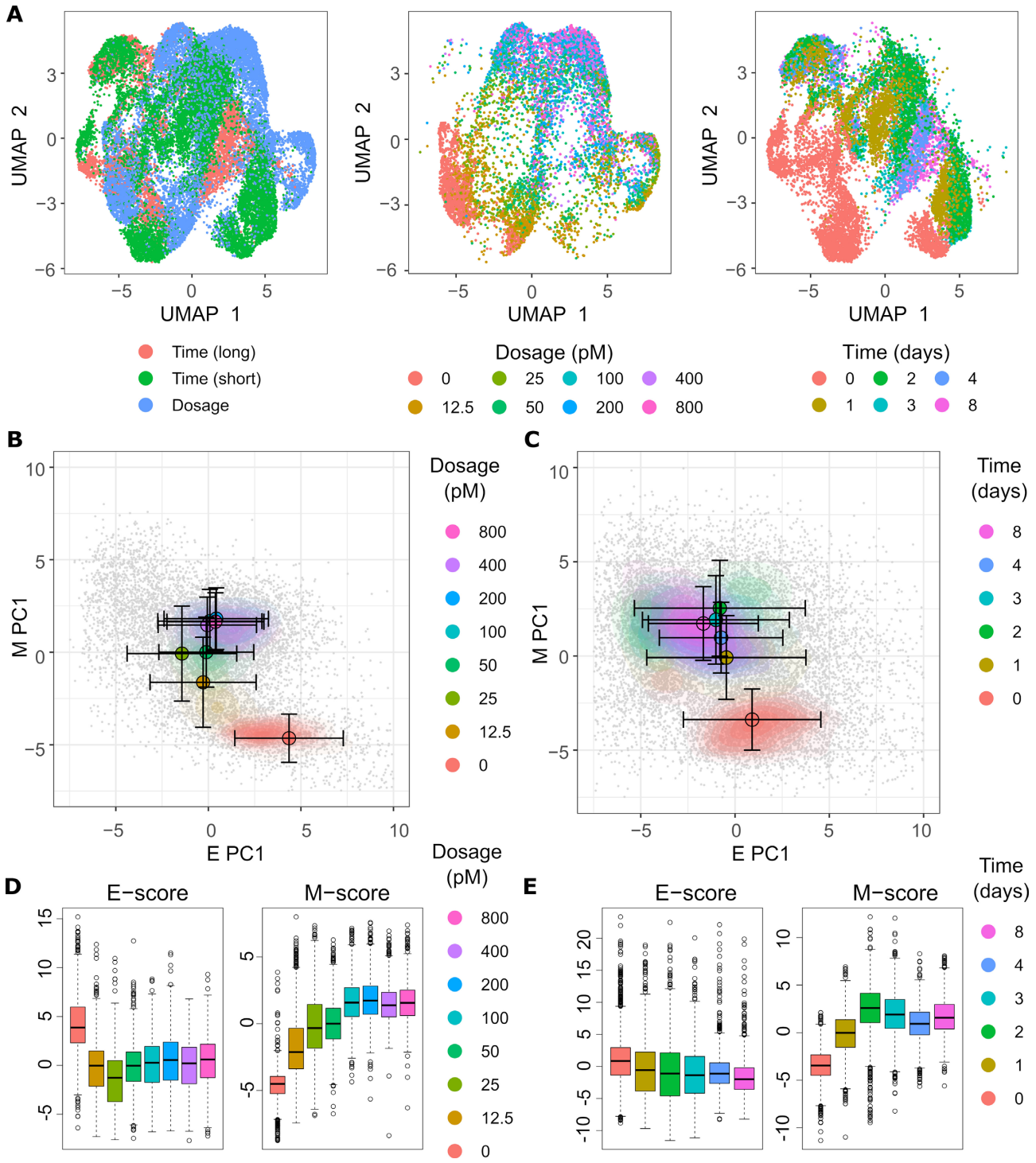


**Figure 2.** Continuity of integrated single cell dose and time data in low-dimensional projections. (**A**) Projection of integrated dose and time data using UMAP. Each panel uses the same underlying UMAP for samples, but the coloring of points has different meaning. Left panel: color indicates the origin of the sample from the Days 0, 4 and 8 of the time experiment (red), the Days 0, 1, 2 and 3 of the time experiment (green), or the dose experiment (blue). Middle panel: color indicates the treatment dose of samples from the dose experiment; time samples are masked. Right panel: color indicates the time of treatment for samples from the time experiment, dose samples are masked. (**B, C**) Contour plots of gene set scores of E (x-axis) and M (y-axis) genes using nnPCA for dose (B) and time (C) samples from integrate data. Color indicates the dose of TGF-β treatment from 0 pM (red) to 800 pM (pink) for dose data and time of treatment from 0 days (red) to 8 days (pink) for time data. Circles indicate the mean E- and M-score of samples from each dose point and the associated error bars show the standard deviation. (**D**) Boxplots show the distribution of E (left) and M (right) scores across different dose treatments from integrated data. Color indicates the dose of TGF-β as in (B). (**E**) Boxplots show the distribution of E (left) and M (right) scores across different time treatments from integrated data. Color indicates the time of treatment as in (C).

of last follow up (for survivors), the latter being right censored in our model. Samples with missing or non-positive values for survival time were dropped as the modelling approach does not tolerate these values. For features, we used the set of up-regulated marker genes (a fold change in expression > the 75th percentile of all significant marker genes across all segments at an adjusted $P$-value < 0.05) in each segment. Each combination of cancer expression data and gene set was then used to construct Cox proportional hazard models using glmnet implemented in R. A glmnet function employed Lasso regularization (alpha = 1). The function optimizes the lambda parameter which influences the strength of regularization. For each lambda value, a 10-fold cross validation was performed, and the performance was measured by $C$-index, which is the proportion of concordant pairs to total pairs in the dataset, i.e. the proportion of all possible samples pairs where increased model hazard corresponds to reduced survival. For each model, we used the default range and interval of the tested lambda values determined by glmnet. In theory, the stronger Lasso regularization should reduce the number of active predictors (genes) in the model, but we do not actively seek the minimized predictors. Instead, we report the model which maximizes the average cross-fold $C$-index. This measure is roughly analogous to AUC–ROC in a censored data context such as survival.

### Clustering from shared-nearest-neighbor graphs

To confirm that the associations we found in E- and M-score space reflect the full dataset, we examined sample associations across the 15 000 most variable expressed genes in the integrated and dose only dataset using the clustering based on shared-nearest-neighbor graphs via Seurat. After scaling and correcting single cell data, principal component analysis was done using RunPCA with 15 components (npcs = 15) and then passed to FindNeighbors with options reduction = 'pca', dims = 1:15. Finally, FindCluster was run with resolution = 0.5 and default parameters and seeding otherwise. Comparisons between clusters we done using the Jaccard package in R which implements tests from (29).

## RESULTS

### A single-cell transcriptomic landscape of dose-dependent EMT reveals a continuum-like spectrum

To characterize the transcriptomic spectrum with multiple levels of EMT signals, we performed dose-dependent induction of EMT with MCF10A cells and analyzed cells at a near-steady-state time point (14 days after TGF-β treatment) using scRNA-seq (Figure 1A, red box). Transcriptomic profiles of 8876 cells with dosage annotation were identified after a standard filtering process and each condition of TGF-β concentration (dose) yielded more than 800 cells. We found that cells treated with various concentrations of TGF-β showed a continuous spectrum when visualized in the low-dimensional Uniform Manifold Approximation and Projection (UMAP) space (Figure 1B). To visualize the transcriptomic variability with interpretable, functional space, we used a recently developed projection method based on nonnegative principal component analysis (nnPCA) (25). Previously identified epithelial-associated genes (E-genes) and mesenchymal-associated genes (M-genes), of which 203 E- and 137 M-genes were present in the processed dataset, were used to construct low-dimensional space using the first principal component of E scores and either the first (M PC1, Figure 1C) or the second (M PC2, Figure 1D) principal component of M-scores (23). We observed a progression of MCF10A cells from E-high-M-low state to E-low-M-high state with increasing concentrations of TGF-β. The effect of the progression was saturated as higher concentration of TGF-β treatment overlap in both E- and M-scores (Figure 1C-D), such that our scores and significantly separated treatments of 100 pM and below, but not above (Table 1). Furthermore, we observed the continuous progression of key E-genes (e.g. CDH1) and M-genes (e.g. VIM and FIN1) expression (Figure 1E–G). The continuity of the transition and the saturation of the progression were similar to time-dependent EMT in MCF10A cells reported recently (13). Additionally, we found that the expression of VIM and FN1 followed the M PC1 and M PC2 scores, respectively, when mapped to UMAP space, and both regions are distinct from the high E-score/EPCAM region on the opposite end of UMAP space (Figure 1H–J). Furthermore, these regions are reflected in the nearest neighbor clusters of dose samples (Supplemental Figure S4A), which also showed a progression from E-high M-low to E-low M-high in the E- and M-score space (Supplemental Figure S4B). The divergence of the mesenchymal-like cell states represented by these two clusters appears to be driven by a difference in high (≥100 pM) dose cells (Cluster 0) and predominantly lower dose cells in G1 (Cluster 2), which may have undergone cell cycle arrest (30) (Supplemental Figure S4C-D). Notably, neither dimensionality reductions nor nearest neighbor clusters depend on dose labels, yet all approaches showed a progression of the expression of EMT genes with multiple possible end states. Overall, our results show that the near-steady-state EMT program of MCF10A cells has a continuum-like spectrum that is independent of projection methods and sample labels.

**Table 1.** Common language effective size of separation of dosage groups based on EMT scores

| Comparison | E PC1 | M PC1 | M PC2 |
|---|---|---|---|
| **0 versus 12.5 pM** | 0.879[a] | 0.700[a] | 0.706[a] |
| **12.5 versus 25 pM** | 0.638[a] | 0.601[a] | 0.624[a] |
| **25 versus 50 pM** | 0.403 | 0.474 | 0.609[a] |
| **50 versus 100 pM** | 0.551[a] | 0.641[a] | 0.743[a] |
| **100 versus 200 pM** | 0.474 | 0.495 | 0.498 |
| **200 versus 400 pM** | 0.528 | 0.425 | 0.438 |
| **400 versus 800 pM** | 0.461 | 0.522 | 0.554[a] |

[a]Indicates statistically significant separation based on Mann–Whitney $U$-test ($P$ < 0.05 after Bonferroni correction).

### Dose-dependent states show less variation and are more separable than time-dependent states in the E/M space

We next compared the dose-dependent EMT and time-dependent EMT at the single-cell resolution. We first integrated our dose-dependent dataset (abbreviated as dose

dataset) with the previously published time-dependent dataset (abbreviated as time dataset) using the same approach as in Deshmukh *et al.* (13) (see Materials and Methods). Because the time-dependent dataset contains two sets of experiment (short-time and long-time conditions), the integration involved three experiments. While comparison can be performed with unintegrated datasets, the integration is helpful as it reduces the effect of experimental batches not associated with the meaningful biological differences and increases the similarity between common cell states. As expected, we found that the integrated data contains cells from all experiments distributed in one region more continuously (Figure 2A) compared to the same analysis of unintegrated data (Supplemental Figure S5). Furthermore, the untreated control in the dose-dependent dataset and the Time-0 control in the time-dependent dataset were located in similar regions in the expression space (Figure 2A, middle and right panels). These results suggest that the datasets can be compared in a reasonably uniform framework.

The data integration resulted in some alteration of the gene expression values and coverage (207 E and 144 M genes), but the progression of EMT in the dose-dependent manner was preserved. In fact, comparing dose samples between dose only and integrated data, the first E and M components in each data set were highly correlated ($R = 0.84$ and 0.81, respectively). We also found a strong correlation between dose M PC1 and integrated M PC2 ($R = 0.78$), as well as between dose M PC2 and integrated M PC1 ($R = 0.79$), but only moderate correlation between dose M PC2 and integrated M PC2 ($R = 0.21$). As such, it is unsurprising that when we applied dose PCs to the dose samples in integrated data, we saw a similar distribution of scores (Supplemental Figure S6) and vice versa for integrated PCs applied to dose data (Supplemental Figure S7), particularly between dose M PC1 and integrated M PC2 as well as integrated M PC1 and dose M PC2. Likewise, we observed a similar correspondence between nearest neighbor clusters defined using each data set (Supplemental Figure S8); in particular, dose Cluster 0 had a significant overlap with dose samples in the union of integrated Clusters 0 and 3 (Jaccard index = 0.74, $P$-value < 1e−5 based on bootstrap testing), while dose Cluster 2 had significant overlap with dose samples from the union of Clusters 1 and 10 (Jaccard index = 0.76, $P$-value < 1e−5 based on bootstrap testing). Notably, Cluster 1 is predominantly time samples, suggesting that the region identified by integrated M PC2 scores is not simply an artifact of the dose data.

For subsequent definition of E- and M-score space, we chose to use M PC1 as it represents the greatest variation in M-genes across our data and is more correlated with time ($R = 0.48$) and dose ($R = 0.42$) progression, than M2 ($R = 0.16$ for time and $R = 0.06$ for dose). In this E- and M-score space, the time-dependent data points were distributed more broadly compared to the dose-dependent data (Figure 2A-C). Nonetheless, the continuous E-to-M progression was observed (Figure 2C and E). Note that the broader distribution of cells in the time-dependent dataset was also observed with unintegrated data (Supplemental Figure S9). This broader distribution of time-dependent data likely reflects distinct transient trajectories during the response to the EMT signal, and the difference in these trajectories may be caused by different initial conditions at unperturbed state of MCF10A cells as well as temporal fluctuations. Dosage data maintained a similar pattern of separability as the unintegrated data with significant differentiation between conditions up to the putative saturation point between 100, though we could no longer separate 25 and 50 pM treatments as we could with dose data alone. Time samples, however, show greater separability between the earliest (0–2 days) and latest (4 versus 8 days) time points, but not interior points, which includes 3 versus 4 days, the breakpoint between the two time data batches (Table 2).

**Table 2.** Common language effect size of separation of time and dosage groups based on EMT scores of integrated data

| Comparison | E PC1 | M PC1 | M PC2 |
|---|---|---|---|
| **Time** | | | |
| **0 versus 1 days** | 0.605[a] | 0.885[a] | 0.524[a] |
| **1 versus 2 days** | 0.530[a] | 0.797[a] | 0.683[a] |
| **2 versus 3 days** | 0.502 | 0.418 | 0.445 |
| **3 versus 4 days** | 0.469 | 0.371 | 0.478 |
| **4 versus 8 days** | 0.593[a] | 0.605[a] | 0.569[a] |
| **Dosage** | | | |
| **0 versus 12.5 pM** | 0.887[a] | 0.892[a] | 0.833[a] |
| **12.5 versus 25 pM** | 0.616[a] | 0.685[a] | 0.603[a] |
| **25 versus 50 pM** | 0.363 | 0.528 | 0.420 |
| **50 versus 100 pM** | 0.472 | 0.753[a] | 0.516 |
| **100 versus 200 pM** | 0.468 | 0.524 | 0.478 |
| **200 versus 400 pM** | 0.546[a] | 0.434 | 0.513 |
| **400 versus 800 pM** | 0.455 | 0.533[a] | 0.474 |

[a]Indicates statistically significant separation based on Mann–Whitney $U$-test ($P < 0.05$ after Bonferroni correction).

In addition to the difference in the width of the distribution, the dose-dependent dataset had stronger anti-correlation between E and M scores ($R = -0.495$) than the time dataset ($R = -0.307$) (Figure 2B and C). We hypothesized that the stronger coordination between E and M transcriptional programs can facilitate the discovery of EMT-associated genes within our integrated dataset that were not classically considered EMT genes in previous studies (23,31). Indeed, with the same threshold of Pearson correlation coefficient (+/−0.25), the dose-dependent dataset revealed greater numbers of genes that had association (correlation with M-scores and anti-correlation with $E$-scores or vice-versa) with the overall expression of E and M genes (167) compared to the time-dependent dataset (41) (Figure 3A and B). The genes that showed association with E-program in the dose-dependent dataset but not in the time-dependent dataset were enriched with GO terms such as keratinization, keratinocyte differentiation and epidermis development, while M-correlated genes included those associated with integrin, chemokine binding, and neuronal components like the dendrite and protosynapse (Supplemental Tables S1 and S2) while there were too few (0 E and 2 M) found only in the time data to test. Notably, TGF-β has been shown to be involved in keratinocyte growth arrest (32), which is consistent with keratinocyte differentiation being associated with the E-program. We then focused on specific genes there were specifically correlated in with EMT progression in dose data, but not time data. FARP1 was among the genes most highly correlated

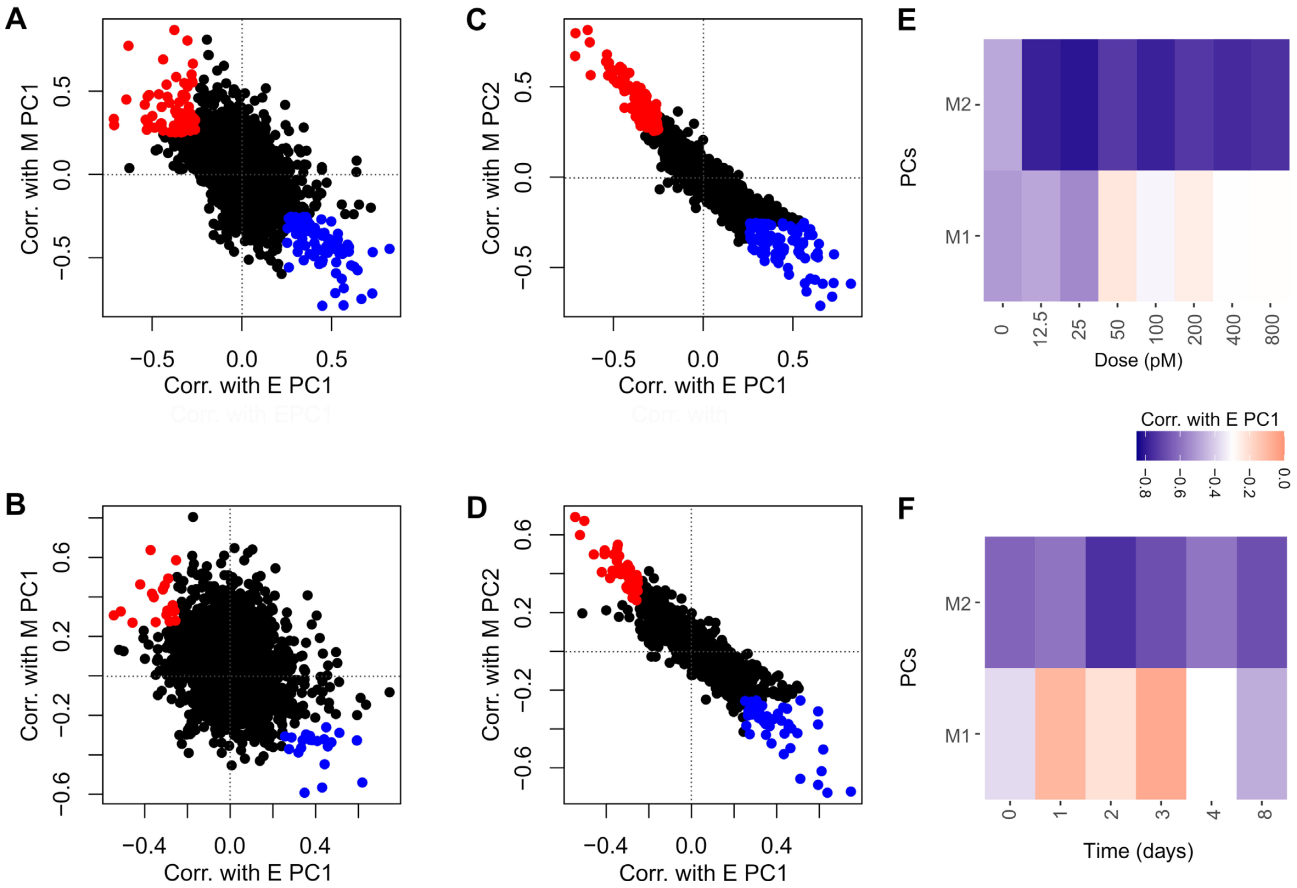


**Figure 3.** Correlation of non-classical EMT genes with E and M scores in integrated data. (**A, B**) Correlation of non-EMT gene expression with sample E-scores (x-axis) and M Pc1-scores (y-axis) in dose (A) and time (B) samples from the integrated dataset. Each point represents one of the 15 000 most variable genes from the integrated single cell dataset, excluding those genes used to define the E- and M- scores. Blue points indicate genes which are anti-correlated with EMT progression ($R > 0.25$ for E-scores, $< -0.25$ for M-scores), while red points indicate genes which are correlated with EMT progression ($R < -0.25$ for E-scores, $>0.25$ for M-scores). (**C, D**) Correlation of non-EMT gene expression with sample E-scores (x-axis) and M PC2-scores (y-axis) in dose (C) and time (D) samples from the integrated dataset. Labels and colors are defined as in (A, B). (**E, F**) The correlation between E-scores with M PC1 (bottom) and M PC2 (top) across different treatments in dose (E) and time (F) data. The color of each cell indicates the strength of anti-correlation from weak (red) to strong (blue).

M-scores scores ($R = 0.60$) in dose data, but not in time data ($R = 0.18$). FARP1 was recently shown to be important for cancer cell motility and associated with poor prognosis (33). Similarly, ESM1 was highly correlated with M-scores in dose ($R = 0.47$) but not time ($R = 0.10$), and it contributes to the metastasis in colorectal cancer via NK-κB activation (34). In contrast, we saw comparable anti-correlation between E-scores and M PC2 in dose ($R = -0.764$) and time ($R = -0.666$), with 212 anti-correlated genes in dose and 102 in time using the 0.25 threshold. Among these genes, those unique to dose data showed similar, but fewer significantly overrepresented GO terms (Supplemental Tables S3 and S4). M-program genes from time data were enriched in a variety of metabolic processes, including those related to fatty acids, alcohol, and steroids, while E-program genes from time data were found in actin filament and cell-cell junction terms, but this represents only two or three of seven genes (Supplemental Tables S5 and S6). We found two genes correlated with M PC2 in time data, but not dose, SLC20A1 and HPCAL1, which are involved in Wnt signaling in tumor cells, though in this context SLC20A1 showed antagonism to mesenchymal markers without affect epithelial markers, suggesting a possible atypical/intermediate EMT state (35,36). Finally, we considered anti-correlation between E and M-scores at individual treatment points in dose (Figure 3E) and time (Figure 3F). For M PC2, values were relatively similar across treatments and datasets, consistent with the higher overall anti-correlation for this principal component. Dose and time data show more distinctions for M PC1: dose data have the strongest anti-correlation for low dose (<50 pM) after which it saturates, while anti-correlations in time samples are strong at the end points (0 and 8 days), but weak for the intermediate samples. This would suggest dose treatment achieves the greatest anti-correlation prior to saturation, while time treatments are most anti-correlated at end points of no or protracted treatment. However, care must be taken with interpreting these results as the presence of multiple path or end points of EMT suggested by our data, indicates these results may be influenced by Simpson's paradox compared to the overall correlation scores which intentionally consider the variation across multiple treatment

populations. Note that the correlation analyses (Figure 3A and B) were performed in the absence of dose and time treatment labels, so the correlations were primarily driven by the intrinsic EMT continuum. Our results suggest that compared to the time-dependent data, the dose-dependent scRNA-seq data may provide higher sensitivity to detect non-classical EMT genes that are coordinated by the core EMT module.

### *En route* cell clusters unique to time dependent EMT program

Based on the distinct distributions of cells between time- and dose-dependent EMT programs (Figure 4A), we wondered whether our comparative study can reveal expression regions containing cells that are *en route* to the M state rather than (partially) stabilized at intermediate EMT attractors. To simplify the representation of the expression profiles in the EMT spectrum, we focused on a $4 \times 4$ grid in the E- and M-score space. We found that a region with low scores of one (E or M) program, and low-to-medium scores of the other program (M or E) is enriched with cells in the time-dependent EMT program (Figure 4B, lower left). This result reveals a transient EMT path that may primarily involve the relatively low expression of both E and M genes. Nonetheless, the enrichment of time data points in the regions of high E-gene activity and M-gene activity suggested a possible alternative transient EMT path (Figure 4B, right) (37,38). We next considered the relationship between these transient paths, particularly the E-low and M-low region, with our alternative M-score. To do this, we considered the nearest neighbor clusters identified in the integrated data which coincide with high expression in M PC2 and found that Cluster 1 was broadly distributed across the lowest E-quartile and is predominantly time samples (Supplemental Figure S10A). Cluster 10, which was mostly dose samples, is also in the lowest E quartile (Supplemental Figure S10B), but was more concentrated in the highest M cluster (Fisher's Exact Test, odds = 1.73, $P$-value = 6.9e-12), consistent with transient progression through the region. Notably, both clusters have high M PC2 expression in the integrated data (compare Supplemental Figure S6 and Figure S8), suggesting that while there is transience in the highest variance of component of M-genes, there is a signature of expression for these sample that has been identified by the second M principal component. Finally, we looked at differentially expressed genes across all sixteen segments of the $4 \times 4$ grid and focused first on the known EMT marker genes CDH1, EPCAM, VIM and FN1. Interestingly, while the transient path crossing the E-low-M-low region is generally consistent with the profiles of the E marker CDH1 and EPCAM as well as M markers VIM and FN1 (Figure 4C–F), a distinct sequence of M gene activations was observed in the hypothetical E-to-M path: for example, VIM was activated before FN1 in this path (Figure 4E and F). This observation is consistent with our earlier transcriptomic data showing significant diversity of M-genes in response to EMT signals (31). We also considered the differential expression of these genes between time and dose samples in the middle 50% of E and M scores. We found a significant difference (adjusted $P$-value $< 0.05$) in the expression of VIM, FN1 and EPCAM, with FN1 favoring dose samples (average $\log_2$ fold change = 2.14) while EPCAM (average $\log_2$ fold change = 0.31) and VIM (average $\log_2$ fold change = 1.29) favor time samples. However, the fold change of these genes in the central clusters was low compared to clusters around the extrema (Figure 4B). As such, while this result is consistent with the expected expression differences associated with transient EMT, it is difficult to infer a specific biological meaning to this difference as the underlying expression values are moderate.

### Signature genes from both dose- and time-dependent datasets contribute to better prognostic models

To examine the roles of the signature genes at various locations of the EMT spectrum in prognosis, we again focused on the sixteen segments of the $4 \times 4$ grid of the EMT space containing both time and dose data. First, we expanded our search for differentially expressed genes to include all possible (non-zero expression) genes and selected those in the top 75th percentile of fold change values after filtering for significance (adjusted $P$-value $< 0.05$) and minimum average $\log_2$ fold-change threshold (absolute value of 0.25). We then obtained 27 diverse cancer datasets with tumor RNA-seq and patient data from The Cancer Genome Atlas (TCGA). We used the signature genes of each segment to construct a Lasso penalized Cox model for predicting the survival outcomes of cancer patients for each of the 27 cancer types. Interestingly, segments with high average prognostic performance measured in C-index were primarily found in the lowest E-score quartile (Figure 5A, left edge) and highest M-score quartile (Figure 5A, top edge). We next scaled performance index across cancer types to account for dataset level variance in model construction, (Figure 5B). We found that, in general, E-low intermediate states (green) and M-high intermediates (light blue) both outperformed other (blue) segments, apart from (3,4) and (1,1), with similar average performance to the extreme M state (yellow), although there was a large degree of variance across cancers and on the high end of the performance spectrum. The superior prognostic performance of the extreme M-state and certain intermediate states was also observed in the recent study based on time-dependent EMT alone (13), and our analysis revealed the characteristics of these high-performing groups in the 2-dimensional EMT spectrum. We also considered the predictive power of signature genes defined using only time and only dose samples. We found that signature genes from time samples alone had the highest C-index in the lowest E-quartile (Figure 5C), while signature genes from dose samples alone had high C-index only in the high M end of the lowest E-quartile (Figure 5D), reflecting difference in distribution of Clusters 1 and 8 (see Supplemental Figure S10). This suggests that gene differentially enriched in time and dose samples contribute to the predictive power of lower E-quartile sample. However, time and dose samples also showed unique prognostic values, as high E and high M signatures segments had high C-index only in time data, but high E low M signature genes had high C-index only in dose data. This suggests that there may be further differences in expression between transient and steady state EMT that were not highlighted by our

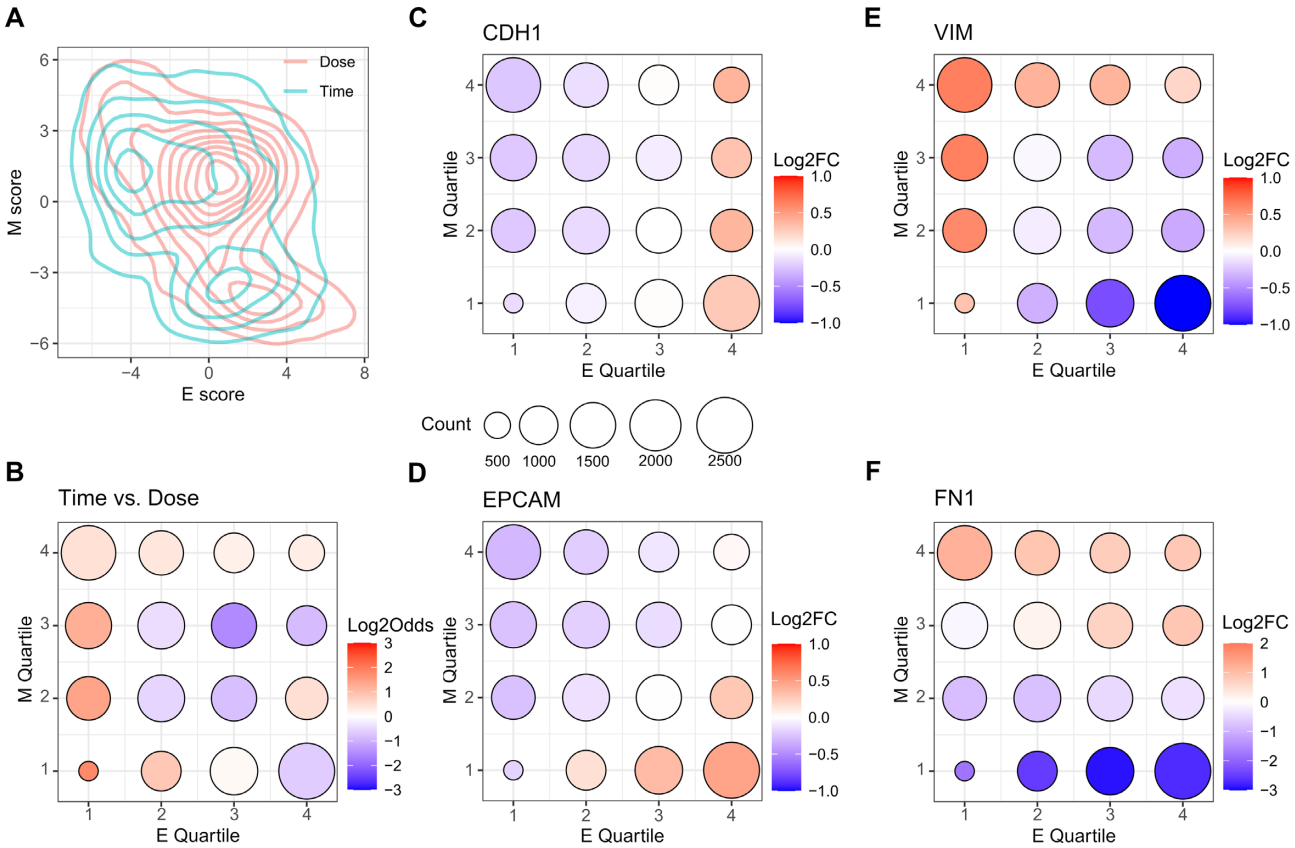


**Figure 4.** Enrichment of dose and time samples in E- and M-score space. (**A**) A contour map showing the distributions of time (blue) and dose (red) samples in E- and M-score space. (**B**) Bubble chart showing the enrichment of time samples in different segments of E- and M-score space. Each point represents a segment of E- and M-score space defined by a particular quartile of E-score (x-axis) and M-score (y-axis). The size of the point corresponds to the total number samples in the segment and the color of the point represent the log-odds of time sample enrichment from low (blue) to high (red) with a log-odds of 0 (white) indicating balanced representation. (**C–F**) Bubble charts showing the differential expression of key EMT genes, CDH1 (C), EPCAM (D) VIM (E) and FN1 (F) in different segments of E- and M-score space. Point positions and size are defined as in (B). The color of the point represents the log fold change of expression of the gene from low (blue) to high (red) with a log fold change of 0 (white) indicating no-change relative to other samples. All fold change values are shown regardless of significance.

M-scores because they do not represent broad variance across both data sets. Comparably, the moderate E- and M-score clusters are weak predictors in all cases, likely related to the few signature genes found there. As such, further study of the extrema of EMT expression in both transient and steady state conditions may be important for identifying prognostic predictors.

We next asked whether the comparable contributions of dose- and time-dependent EMT to the prognostic models can be partially explained by their involvement in immune evasion activity (39,40). We used a list for genes previously known to be upregulated by PD-1 ligation (26), as a gene set representative of immune evasion, and we found that progressions of both dose- and time- dependent EMT had moderately positive correlations with the expression of the gene set ($R \approx 0.3$). This result was expected because some M-genes such as VIM and LGALS1 are in this PD-1 upregulation set.

## DISCUSSION

Time dependent EMT processes have been extensively studied with single-cell transcriptomics in recent years (3,13,14). While these studies provided substantial information about EMT progression in multiple contexts, the connection between the intermediate cell states observed in these experiments and cell attractors was elusive. Using near-steady-state single-cell transcriptomic profiling, we showed that the EMT spectrum can be described as a stable continuum under multiple levels of EMT signals. This information complements earlier studies with single-dose time course data, and it provides stronger evidence supporting the existence of multiple intermediate EMT states that widely exist in tumors and metastatic cells (5). Furthermore, by comparing the time- and dose- dependent single-cell data, we identified groups of cells that are exclusively *en route* to M state, which shows the possibility that some cells can transiently deactivate the epithelial program before activating the mesenchymal program. Nonetheless, it is likely that many other *en route* cells are also close to the stable intermediate EMT states, and these states have intermediate levels of both E and M genes. Furthermore, given the heterogeneous microenvironment of cells under physiological conditions, the cells states may be determined by both the multi-level EMT signals and the time-dependent stages of the EMT process, so both dose dependent and time dependent *in vitro* data

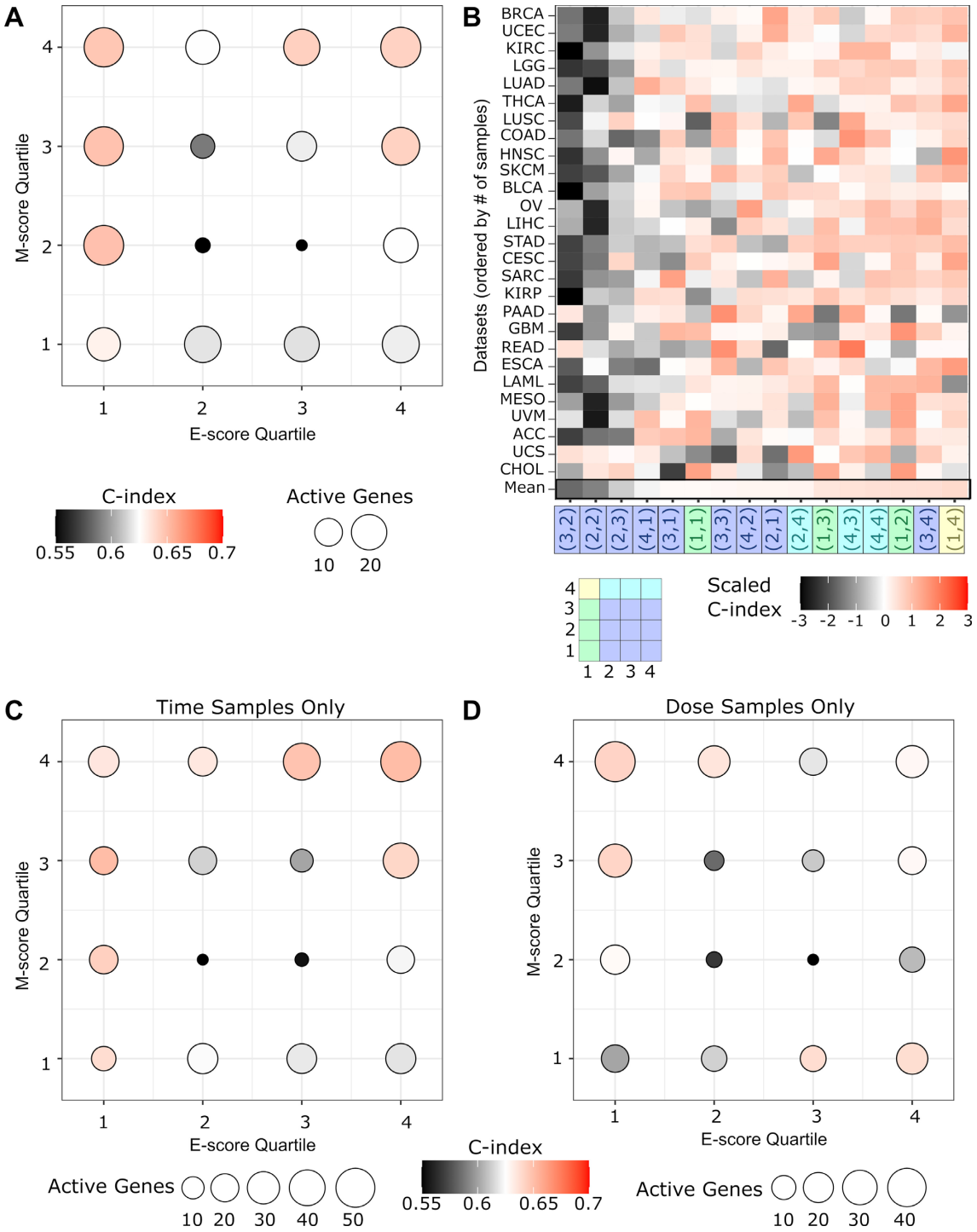


**Figure 5.** Performance of prognostic models. (**A**) Bubble chart shows the average performance of Cox hazard models built using the up-regulated genes from different segments of the E- and M-score space. Each point represents a segment defined by a particular quartile of E-score (x-axis) and M-score (y-axis). The size of the point corresponds to the average number predictor genes across all models built using the up-regulated genes from that segment. Color represents the C-index averaged across all models of cancer datasets, where C-index is a measure of the correspondence between the models predicted risk and survivorship, similar to the area under the receiver operating curve (AUC-ROC). Red represents higher C-index (better performance). Black represents lower C-index (poorer) performance. White represents the approximate median of scores (0.635). (**B**) Heatmap of individual model performance across all combinations of segment-gene-derived models (x-axis) and cancer datasets (y-axis). The color of individual cells represents the C-index of the model scaled against all models of that cancer datasets to account for data-set-level differences in model performance. Segments are ordered across the x-axis by the increased average scaled C-index from left to right, while cancer datasets are ordered by increasing number of samples from bottom to top. The shading of segment coordinates along the x-axis distinguishes between the extreme M state (yellow), E-low intermediates (green), M-high intermediates (light blue), and other segments (dark blue). (**C-D**) Bubble charts show the average performance Cox hazard models built using the up-regulated genes identified using only time (C) or dose (D) samples from different segments of E- and M-score space. Point position, size, and color as defined as in (A).

can contribute to the understanding of the EMT program *in vivo*.

Previous mathematical models have provided mechanistic insights into the gene network structures supporting multiple intermediate EMT states (10–12,41–44). While interconnected positive feedback loops involving a few genes can govern discrete intermediate states, the continuum-like, wide distributions of cells in the EMT spectrum even under the near-steady-state condition suggest that the discrete cells states may only partially explain the stable phenotypical heterogeneity. At least two mechanisms may explain the gap between the existing theories the observed continuum: realistic gene regulatory networks may contain much more positive feedback loops than those described by existing models (45), and these loops can support many intermediate states; dynamical and reversible cell-state transitions at stationary phase, such as those driven by transcriptional noise and slow-timescale oscillations (46), can give rise to cells that are far away from point attractors in the gene expression space. While it is possible that many unknown transcriptional and post-transcriptional circuits may contribute to the numbers and dynamical properties of the EMT attractors, we cannot exclude the possibility that the technical noise in scRNA-seq experiments may mask the effects of discrete attractors. Nonetheless, existing models have provided a strong theoretical foundation for stable intermediate EMT states, and future development of the theories and quantitative experiments will be important to the understanding of EMT continuum.

EMT involves dramatic phenotypic changes of cells. It is therefore expected that transcriptome-wide alteration is induced during the multi-stage transition. This suggests that classically defined core EMT genes may not be sufficient to provide a holistic view of the EMT program. We showed that the dose-dependent single-cell transcriptomes can be useful to identify genes that show expression patterns highly correlated with the core EMT genes at the single-cell level. The dose-dependent scRNA-seq experiment can therefore provide crucial links of the core EMT networks to the rest of the transcriptomes, and some of these connections may not be revealed by time-course scRNA-seq experiments due to the large numbers of cells that transiently activate parts of the transcriptional program. Nonetheless, the two types of scRNA-seq experiments contain complementary information, and we suggest that both can be used in future studies to reveal cellular programming that determines cell-to-cell variabilities in cell populations.

## DATA AVAILABILITY

All code is available at GitHub repository: https://github.com/panchyni/Time_and_Dose.

Sequencing data for dose-dependent EMT in MCF10A cells are deposited in the GEO database (GSE213753).

## ETHICS DECLARATION

All experimental work was performed in compliance with protocols approved by RIKEN Center for Integrative Medical Sciences. No human, animal subject, tissue, gamete, or stem cell was involved in this study.

## SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

## ACKNOWLEDGEMENTS

*Author contributions*: Designed research: K.W. and T.H. Performed experiments: K.W. and M.T. Analyzed data: N.P., K.W., A.W. and T.H. Wrote manuscript: N.P., K.W. and T.H. All authors read and approved the manuscript.

## FUNDING

National Institutes of Health [R01GM140462 to T.H.]; Grant-in-Aid for Scientific Research (KAKENHI) on Innovative Areas ‘Cellular Diversity’ [JP18H05106 to K.W.].

*Conflict of interest statement*. None declared.

## REFERENCES

1. Mani,S.a., Guo,W., Liao,M.-J., Eaton,E.N., Ayyanan,A., Zhou,A.Y., Brooks,M., Reinhard,F., Zhang,C.C., Shipitsin,M. *et al.* (2008) The epithelial–mesenchymal transition generates cells with properties of stem cells. *Cell*, **133**, 704–715.
2. Thiery,J.P., Acloque,H., Huang,R.Y.J. and Nieto,M.A. (2009) Epithelial-mesenchymal transitions in development and disease. *Cell*, **139**, 871–890.
3. Karacosta,L.G., Anchang,B., Ignatiadis,N., Kimmey,S.C., Benson,J.A., Shrager,J.B., Tibshirani,R., Bendall,S.C. and Plevritis,S.K. (2019) Mapping lung cancer epithelial–mesenchymal transition states and trajectories with single-cell resolution. *Nat. Commun.*, **10**, 5587.
4. Pastushenko,I. and Blanpain,C. (2018) EMT transition states during tumor progression and metastasis. *Trends Cell Biol.*, **29**, 212–226.
5. Pastushenko,I., Brisebarre,A., Sifrim,A., Fioramonti,M., Revenco,T., Boumahdi,S., Van Keymeulen,A., Brown,D., Moers,V. and Lemaire,S. (2018) Identification of the tumour transition states occurring during EMT. *Nature*, **556**, 463–468.
6. Kröger,C., Afeyan,A., Mraz,J., Eaton,E.N., Reinhardt,F., Khodor,Y.L., Thiru,P., Bierie,B., Ye,X. and Burge,C.B. (2019) Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 7353–7362.
7. Grosse-Wilde,A., Fouquier d’Herouel,A., McIntosh,E., Ertaylan,G., Skupin,A., Kuestner,R.E., Sol,A., Walters,K.A. and Huang,S. (2015) Stemness of the hybrid epithelial/mesenchymal state in breast cancer and its association with poor survival. *PLoS One*, **10**, e0126522.
8. Zhang,J., Tian,X.J., Zhang,H., Teng,Y., Li,R., Bai,F., Elankumaran,S. and Xing,J. (2014) TGF-β -induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Sci. Signal.*, **7**, ra91.
9. Wang,X. and Thiery,J.P. (2021) Harnessing carcinoma cell plasticity mediated by TGF-β signaling. *Cancers (Basel)*, **13**, 3397.
10. Hong,T., Watanabe,K., Ta,C.H., Villarreal-Ponce,A., Nie,Q. and Dai,X. (2015) An ovol2-zeb1 mutual inhibitory circuit governs bidirectional and Multi-step transition between epithelial and mesenchymal states. *PLoS Comput. Biol.*, **11**, e1004569.
11. Huang,B., Lu,M., Jia,D., Ben-Jacob,E., Levine,H. and Onuchic,J.N. (2017) Interrogating the topological robustness of gene regulatory circuits by randomization. *PLoS Comput. Biol.*, **13**, e1005456.
12. Subbalakshmi,A.R., Sahoo,S., Biswas,K. and Jolly,M.K. (2022) A computational systems biology approach identifies SLUG as a mediator of partial epithelial–mesenchymal transition (EMT). *Cells Tissues Organs*, **211**, 105–118.
13. Deshmukh,A.P., Vasaikar,S.V., Tomczak,K., Tripathi,S., Den Hollander,P., Arslan,E., Chakraborty,P., Soundararajan,R., Jolly,M.K. and Rai,K. (2021) Identification of EMT signaling cross-talk and gene regulatory networks by single-cell RNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, 118.

14. Cook,D.P. and Vanderhyden,B.C. (2020) Context specificity of the EMT transcriptional response. *Nat. Commun.*, **11**, 2142.
15. Ramirez,D., Kohar,V. and Lu,M. (2020) Toward modeling context-specific EMT regulatory networks using temporal single cell RNA-Seq data. *Front. Mol. Biosci.*, **7**, 54.
16. Dixit,A., Parnas,O., Li,B., Chen,J., Fulco,C.P., Jerby-Arnon,L., Marjanovic,N.D., Dionne,D., Burks,T. and Raychowdhury,R. (2016) Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, **167**, 1853–1866.
17. Xi,N.M. and Li,J.J. (2021) Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. *Cell Syst.*, **12**, 176–194.
18. McGinnis,C.S., Murrow,L.M. and Gartner,Z.J. (2019) DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.*, **8**, 329–337.
19. Zheng,G.X.Y., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P. and Zhu,J. (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.
20. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck Iii,W.M., Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C. and Zager,M. (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
21. Oh,S., Abdelnabi,J., Al-Dulaimi,R., Aggarwal,A., Ramos,M., Davis,S., Riester,M. and Waldron,L. (2020) HGNChelper: identification and correction of invalid gene symbols for human and mouse. *F1000Research*, **9**, 1493.
22. Hänzelmann,S., Castelo,R. and Guinney,J. (2013) GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinf.*, **14**, 7.
23. Tan,T.Z., Miow,Q.H., Miki,Y., Noda,T., Mori,S., Huang,R.Y.J. and Thiery,J.P. (2014) Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.*, **6**, 1279–1293.
24. Sigg,C.D. and Buhmann,J.M. (2008) *Proceedings of the 25th international conference on Machine learning*. pp. 960–967.
25. Panchy,N., Watanabe,K. and Hong,T. (2021) Interpretable, scalable, and transferrable functional projection of large-scale transcriptome data using constrained matrix decomposition. *Frontiers in Genetics*, **12**, 1555.
26. Quigley,M., Pereyra,F., Nilsson,B., Porichis,F., Fonseca,C., Eichbaum,Q., Julg,B., Jesneck,J.L., Brosnahan,K. and Imam,S. (2010) Transcriptional analysis of HIV-specific CD8+ t cells shows that PD-1 inhibits t cell function by upregulating BATF. *Nat. Med.*, **16**, 1147–1151.
27. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W. and Zhan,L. (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation*, **2**, 100141.
28. Colaprico,A., Silva,T.C., Olsen,C., Garofano,L., Cava,C., Garolini,D., Sabedot,T.S., Malta,T.M., Pagnotta,S.M. and Castiglioni,I. (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71.
29. Chung,N.C., Miasojedow,B., Startek,M. and Gambin,A. (2019) Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. *BMC Bioinf.*, **20**, 644.
30. Lee,K.-Y. and Bae,S.-C. (2002) TGF-β-dependent cell growth arrest and apoptosis. *BMB Rep.*, **35**, 47–53.
31. Watanabe,K., Panchy,N., Noguchi,S., Suzuki,H. and Hong,T. (2019) Combinatorial perturbation analysis reveals divergent regulations of mesenchymal genes during epithelial-to-mesenchymal transition. *NPJ Syst. Biol. Appl.*, **5**, 21.
32. Freedberg,I.M., Tomic-Canic,M., Komine,M. and Blumenberg,M. (2001) Keratins and the keratinocyte activation cycle. *J. Invest. Dermatol.*, **116**, 633–640.
33. Hirano,T., Shinsato,Y., Tanabe,K., Higa,N., Kamil,M., Kawahara,K., Yamamoto,M., Minami,K., Shimokawa,M. and Arigami,T. (2020) FARP1 boosts CDC42 activity from integrin αvβ5 signaling and correlates with poor prognosis of advanced gastric cancer. *Oncogenesis*, **9**, 13.
34. Kang,Y.H., Ji,N.Y., Han,S.R., Lee,C.I., Kim,J.W., Yeom,Y.I., Kim,Y.H., Chun,H.K., Kim,J.W. and Chung,J.W. (2012) ESM-1 regulates cell growth and metastatic process through activation of NF-κB in colorectal cancer. *Cell. Signal.*, **24**, 1940–1949.
35. Zhang,D., Liu,X., Xu,X., Xu,J., Yi,Z., Shan,B. and Liu,B. (2019) HPCAL 1 promotes glioblastoma proliferation via activation of Wnt/β-catenin signalling pathway. *J. Cell. Mol. Med.*, **23**, 3108–3117.
36. Li,J., Dong,W., Li,Z., Wang,H., Gao,H. and Zhang,Y. (2019) Impact of SLC20A1 on the Wnt/β-catenin signaling pathway in somatotroph adenomas. *Mol. Med. Report.*, **20**, 3276–3284.
37. Wang,W., Poe,D., Yang,Y., Hyatt,T. and Xing,J. (2022) Epithelial-to-mesenchymal transition proceeds through directional destabilization of multidimensional attractor. *Elife*, **11**, e74866.
38. Panchy,N., Azeredo-Tseng,C., Luo,M., Randall,N. and Hong,T. (2020) Integrative transcriptomic analysis reveals a multiphasic epithelial–mesenchymal spectrum in cancer and non-tumorigenic cells. *Front. Oncol.*, **9**, 1479.
39. Sahoo,S., Nayak,S.P., Hari,K., Purkait,P., Mandal,S., Kishore,A., Levine,H. and Jolly,M.K. (2021) Immunosuppressive traits of the hybrid epithelial/mesenchymal phenotype. *Front. Immunol.*, **12**, 797261.
40. Dongre,A., Rashidian,M., Eaton,E.N., Reinhardt,F., Thiru,P., Zagorulya,M., Nepal,S., Banaz,T., Martner,A. and Spranger,S. (2021) Direct and indirect regulators of epithelial–mesenchymal transition–mediated immunosuppression in breast CarcinomasEMT and resistance to checkpoint blockade immunotherapy. *Cancer Discov.*, **11**, 1286–1305.
41. Font-Clos,F., Zapperi,S. and La Porta,C.A.M. (2018) Topography of epithelial–mesenchymal plasticity. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 5902.
42. Lu,M., Jolly,M.K., Levine,H., Onuchic,J.N. and Ben-Jacob,E. (2013) MicroRNA-based regulation of epithelial-hybrid-mesenchymal fate determination. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 18144–18149.
43. Jolly,M.K., Preca,B.-T., Tripathi,S.C., Jia,D., George,J.T., Hanash,S.M., Brabletz,T., Stemmler,M.P., Maurer,J. and Levine,H. (2018) Interconnected feedback loops among ESRP1, HAS2, and CD44 regulate epithelial–mesenchymal plasticity in cancer. *APL Bioeng.*, **2**, 031908.
44. Tian,X.-J., Zhang,H. and Xing,J. (2013) Coupled reversible and irreversible bistable switches underlying TGFβ-induced epithelial to mesenchymal transition. *Biophys. J.*, **105**, 1079–1089.
45. Nordick,B. and Hong,T. (2021) Identification, visualization, statistical analysis and mathematical modeling of high-feedback loops in gene regulatory networks. *BMC Bioinf.*, **22**, 481.
46. Nordick,B., Yu,P.Y., Liao,G. and Hong,T. (2022) Nonmodular oscillator and switch based on RNA decay drive regeneration of multimodal gene expression. *Nucleic Acids Res.*, **50**, 3693.