Testing the Gene Expression Classification of the EMT Spectrum

Dongya Jia[1,2,*], Jason T. George[1,3,7,*], Satyendra C. Tripathi[8], Deepali L Kundnani[9], Mingyang Lu[10], Samir M. Hanash[8,9], José N. Onuchic[1,4,5,6], Mohit Kumar Jolly[1,11,#], Herbert Levine[1,3,5,6,#]

[1]Center for Theoretical Biological Physics, [2]Program in Systems, Synthetic and Physical Biology [3]Department of Bioengineering, [4]Department of Chemistry, [5]Department of Biosciences, [6]Department of Physics and Astronomy, Rice University, Houston, TX 77005;
[7]Medical Scientist Training Program, Baylor College of Medicine, Houston, TX 77030;
[8]Department of Clinical Cancer Prevention, [9]Red and Charline McCombs Institute for the Early Detection and Treatment of Cancer, University of Texas MD Anderson Cancer Center, Houston, TX, USA
[10]The Jackson Laboratory, Bar Harbor, ME.
[11]Current Address: Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore 560012, India.

[*]These authors contributed equally.

[#]To whom correspondence should be addressed:
Mohit Kumar Jolly (mkjolly@iisc.ac.in) and Herbert Levine (herbert.levine@rice.edu)

Abstract:

The epithelial-mesenchymal transition (EMT) plays a central role in cancer metastasis and drug resistance – two persistent clinical challenges. Epithelial cells can undergo a partial or full EMT, attaining either a hybrid epithelial/mesenchymal (E/M) or mesenchymal phenotype, respectively. Recent studies have emphasized that hybrid E/M cells may be more aggressive than their mesenchymal counterparts. However, mechanisms driving hybrid E/M phenotypes remain largely elusive. Here, to better characterize the hybrid E/M phenotype(s) and tumor aggressiveness, we integrate two computational methods – (a) RACIPE – to identify the robust gene expression patterns emerging from the dynamics of a given gene regulatory network, and (b) EMT scoring metric - to calculate the probability that a given gene expression profile displays a hybrid E/M phenotype. We apply the EMT scoring metric to RACIPE-generated gene expression data generated from a core EMT regulatory network and classify the gene expression profiles into relevant categories (epithelial, hybrid E/M, mesenchymal). This categorization is broadly consistent with hierarchical clustering readouts of RACIPE-generated gene expression data. We show that the EMT scoring metric can be used to distinguish between samples composed of exclusively hybrid E/M cells and those containing mixtures of epithelial and mesenchymal subpopulations using the RACIPE-generated gene expression data.

Keywords: epithelial-mesenchymal transition; EMT; random circuit perturbation; RACIPE; EMT scoring metric; hybrid epithelial/mesenchymal; hybrid E/M; phenotypic stability factor; PSF

Introduction

The epithelial-mesenchymal transition (EMT) is a trans-differentiation developmental program by which epithelial (E) cells weaken their adhesive bonds with neighbors as well as apico-basal polarity, while concomitantly gaining mesenchymal (M) traits of migration and/or invasion [1,2]. Cancer cells undergoing EMT typically have enhanced metastatic abilities [2,3], elevated resistance against chemotherapy [4,5] and can evade the immune system [6], thus accelerating malignancy. In the developmental context of mammary morphogenesis and the physiological context of wound healing, cells rarely undergo a complete EMT. Instead, they exhibit a transient hybrid epithelial/mesenchymal (E/M) phenotype that shows a mixture of traits - epithelial traits of cell-cell adhesion, and mesenchymal traits of migration [3]. Thus, the hybrid E/M phenotype facilitates collective cell migration, instead of individual cell migration - a characteristic of cells having undergone a complete EMT. This concept has recently gained momentum in the context of cancer metastasis, wherein it has been observed that malignant cells can maintain a hybrid E/M phenotype and migrate collectively, eventually appearing in the blood as clusters of circulating tumor cells (CTCs) [7–10]. Moreover, these CTC clusters, often five to eight cells in size, contribute much more than their proportional share to metastases [11], thus behaving as 'chief instigators' of metastasis. Cells in these clusters often display a hybrid E/M phenotype [9,10,12], and can resist cell death in circulation by evading immune attacks and providing survival signals to one another through their cell-cell contacts [12]. Therefore, hybrid E/M cells may occupy the 'metastatic sweet spot' [13,14].

The hybrid E/M phenotype has been tacitly assumed to be 'metastable' or transient [15,16], but recent computational studies suggest that the hybrid E/M phenotype can be a stable cell phenotype, particularly in the presence of 'phenotypic stability factors' (PSFs) such as the transcription factors GRHL2, OVOL2, NUMB and ΔNp63α [10,17–22]. Resonating with predictions of these computational models, non-small cell lung cancer (NSCLC) cells H1975 have been shown to stably (i.e. over multiple passages) co-express epithelial and mesenchymal markers - E-cadherin and vimentin - at a single-cell level *in vitro* [19]. Consistent with predictions from mathematical models, knockdown of GRHL2, OVOL2 or NUMB in H1975 cells results in a complete EMT [18,19,22]. Overall, these studies have supported the relevance of a stable hybrid E/M phenotype and PSFs along with their implications for both mammary morphogenesis and cancer metastasis [12].

EMT is orchestrated by complex gene regulatory networks (GRNs) that involve microRNAs (miRs), transcription factors (TFs), alternative splicing mechanisms, and epigenetic modifications [23,24]. Particularly, members of the miR-200 and the miR-34 family can maintain or induce epithelial features, and TFs such as ZEB1/2, SNAI1/2 and FOXC2 can trigger the transition toward a mesenchymal phenotype.

2

Moreover, these regulators exhibit extensive crosstalk, forming multiple feedback loops [23]. Mathematical modeling approaches have been used to analyze the dynamics of EMT regulatory networks [17,25–31]. However, most of these studies consider interactions among only a small number of genes regulating EMT, because a quantitative simulation of the GRN dynamics is difficult owing to insufficient knowledge of kinetic parameter values, particularly when the network size is large.

Recently, we developed a computational method, random circuit perturbation (RACIPE), to interrogate the robust dynamical features of GRNs without the need for precise parameter values [32,33]. RACIPE generates an unbiased ensemble of mathematical models based on the topological information (i.e. genes and interactions) of a GRN. Each model is characterized by a unique set of parameters that are randomly sampled within ranges of values that are biologically relevant, and then numerically solved. The stable state solutions of the ensemble models are collected and analyzed, through which the most significant gene expression patterns associated with a specific GRN are extracted. In this framework, variability among different sets of parameters generated by RACIPE captures both the heterogeneity among individual cells and the effects of different input signals to the GRN. Thus, the RACIPE-generated stable state solutions can, in principle, represent the gene expression patterns of diverse individual cell phenotypes in multiple contexts. Here, we apply RACIPE to EMT regulatory networks to generate *in silico* gene expression data of different EMT phenotypes.

Separately, the lack of a rigorous quantitative definition of a hybrid E/M phenotype motivated us to develop a statistical learning method, the EMT scoring metric [34]. This approach assigns a probability of membership in the hybrid E/M category, given the gene expression data of a sample. It also assigns each sample a corresponding EMT score on a scale from 0 (fully epithelial) to 2 (fully mesenchymal) with an intermediate score of 1 representing maximally hybrid E/M signatures. This approach, when applied to the NCI-60 human tumor cell line gene expression dataset as the training set, identified claudin-7 (CLDN7) and the ratio of vimentin to E-cadherin (VIM/CDH1) as the best-fit pair of predictors to categorize the EMT phenotypes. The algorithm has been accurate in characterizing the EMT status of multiple cancer cell lines and has elucidated a subtype-specific association between the EMT status and patient survival.

Here, we use RACIPE to test the ability of our EMT scoring metric to quantify the EMT phenotypes. We first apply RACIPE on a proposed core EMT regulatory network, containing 26 genes and 100 regulatory links (**Fig. 1a**) and use the hierarchical clustering analysis (HCA) to classify the RACIPE-generated gene expression data into four main clusters corresponding to one epithelial, one mesenchymal, and two hybrid E/M (E/M I and II) phenotypes. We show that the EMT scoring metric can accurately recapitulate the EMT status of RACIPE-generated hybrid E/M samples. We next construct *in silico* mixtures of RACIPE-

generated E, M and E/M II samples in various ratios. We present a proof-of-principle for distinguishing samples containing individual hybrid E/M cells from those containing mixtures of E and M cells by applying an extension of our EMT scoring metric. We demonstrate that the EMT metric approach can accurately identify the proportions of the different subpopulations for all the various combinations of mixtures studied here. This new method is enabled by having samples generated from RACIPE projected on to the EMT scoring metric space. Finally we validate the proposed method comparing its prediction to data from multiple non-small cell lung cancer (NSCLC) cell lines.

Materials and Methods

Generation of EMT gene expression data by RACIPE

The EMT regulatory network (**Fig. 1a**) we consider here contains 26 gene products and 100 regulatory links, the majority of which are transcriptional in nature. There are also microRNA-mediated translational regulation. All links in the network considered here are modeled by shifted Hill functions [25]. RACIPE takes the topological information of the EMT regulatory network as the input and generates multiple mathematical models.

The general rate equation representing the temporal dynamics for a single gene is of the form:

$$dA/dt = g_A * \prod_{i=1}^{n} H_i^{S-} * \prod_{j=1}^{m} H_j^{S+} - k_A * A,$$

where A represents the expression level of gene A (the level of mRNA A), g_A and k_A represent the innate production and degradation rates of mRNA A, $\prod_{i=0}^{n} H_i^{S-}$ and $\prod_{j=0}^{m} H_j^{S+}$ represent the inhibitory regulation and excitatory regulation of gene A by other genes, respectively; the number of inhibitory links is n and the number of excitatory links is m. H^{S-} and H^{S+} are negative and positive shifted Hill functions of the form:

$$H^{S-}\left(B, h_{B,A}^0, \lambda_{B,A}, n_{B,A}\right) = \frac{1 + \lambda_{B,A}\left(\frac{B}{h_{B,A}^0}\right)}{1 + \left(\frac{B}{h_{B,A}^0}\right)},$$

and

$$H^{S+}\left(C, h_{C,A}^0, \lambda_{C,A}, n_{C,A}\right) = \frac{1 + \lambda_{C,A}\left(\frac{C}{h_{C,A}^0}\right)}{1 + \left(\frac{C}{h_{C,A}^0}\right)},$$

where $\lambda_{B,A} < 1$ is the fold-change representing the inhibition of gene A expression by the protein encoded by gene B, $\lambda_{C,A} > 1$ is the fold-change representing the activation of gene A by the protein encoded by gene C, B represents the levels of mRNA B, and C represents the levels of mRNA C. $h_{B,A}^0$ and $h_{C,A}^0$ are thresholds in the shifted Hill function.

The temporal dynamics of all 26 nodes of the EMT regulatory network are characterized by the above functional form. In the RACIPE framework, mathematical model refers to a specific choice of parameters, and many models are created. We typically use 10,000 to 100,000 models. For each model, the parameters are randomly sampled within biologically reasonable ranges. The sampled parameters satisfy the 'half-functional' rule [32] so that each regulatory link has an equal chance (~50%) to be active or inactive. For generating each mathematical model, 200 initial conditions are used to numerically integrate the dynamical equations using Euler's Method and we keep all simulations for which the system reaches steady state on the chosen integration interval. Depending on the specific parameters, a single model may give rise to one or more stable steady state solutions dependent upon the model's initial conditions. The stable steady state solutions from all models are analyzed via the EMT scoring metric.

EMT scoring metric

In this approach, statistical models are created to resolve training samples into their *a priori* known categories. EMT scoring is performed by applying the best-fit model obtained from an iterative statistical procedure developed previously to relevant gene expression samples [34]. This best fit model was found to maximally resolve the NCI-60 cell line gene expression dataset (GSE5846) based on E, M, or hybrid E/M phenotype using ordinal multinomial logistic regression. The ($\log_2$-normalized) input predictors include the ratio of vimentin and E-cadherin (VIM/CDH1) and claudin 7 (CLDN7). Model output for a given sample, S, is given by

$$O(S) = [P_E(S), P_H(S), P_M(S)], \text{where } P_E(S) + P_H(S) + P_M(S) = 1.$$

This ordered triplet is interpreted as the probability that sample S falls into either of the following three categories, epithelial (E), hybrid E/M (H) and mesenchymal (M), $P_E(S), P_H(S), P_M(S)$. Ordinal regression places state H intermediary to E and M, and the EMT metric, μ, assigns a numeric value on the interval [0, 2] with the ordering $E < H < M$. In particular,

$$\mu(S) = \begin{cases} P_H(S), & P_E > P_M; \\ 2 - P_H, & P_E < P_M; \\ 1, & P_E = P_M. \end{cases}$$

The metric represents epithelial samples in the range $0 \leq \mu < 0.5$, hybrid for $0.5 \leq \mu \leq 1.5$, and mesenchymal for $1.5 < \mu \leq 2$.

Generating the mixtures of E, M and hybrid E/M II samples

We generate mixtures of E and M samples with three replicates of five different ratios of E to M: [50%, 50%], [20%, 80%], [80%, 20%], [40%, 60%] and [60%, 40%] (n=20,000 per ratio per replicate). For example, 10^4 E samples are randomly sampled from RACIPE-generated E samples and 10^4 M samples are randomly sampled from RACIPE-generated M samples and their mixture has a proportion of E and M - [50%, 50%]. The mean gene expression of the mixture is used to estimate the proportions of E and M by the EMT metric. In a similar way, mixtures of E and hybrid E/M II samples and those of M and hybrid E/M II samples with three replicates of five different proportions - [50%, 50%], [60%, 40%], [40%, 60%], [20%, 80%] and [80%, 20%] - are generated respectively (n=15,000 per ratio per replicate). Mixtures of E, M and hybrid E/M II samples with three replicates of five different proportions - [25%, 25%, 50%], [30%, 30%, 40%], [20%, 20%, 60%], [10%, 10%, 80%] and [40%, 40%, 20%] - are generated respectively (n=15,000 per ratio per replicate).

Predictions of the mixture proportions

Samples are depicted by their location in the two-dimensional EMT predictor space. This space is spanned by VIM/CDH1 and CLDN7 expression levels partitioned into three regions based on which one of $\{P_E(S), P_H(S), P_M(S)\}$ is maximal. For example, when investigating mixtures of E and M populations, predictions of the average proportion of E and M cells are obtained by projecting the coordinates of a sample (x_S, y_S) to the curve adjoining the location of the mean E signature at (x_E, y_E), and mean M signature at (x_M, y_M) (x and y correspond to mean CLDN7 and (mean VIM)/(mean CDH1) levels, respectively). The proportion of E and M for sample (x_S, y_S) is estimated by finding the least-squares projection of the sample onto the curve spanned by the convex combination of E and M signatures. The mixture proportions of E and hybrid E/M II populations and the mixture proportions of M and hybrid EM II populations are predicted in an identical way. Mixtures of equal parts E and M with variable proportions of hybrid E/M II samples were analyzed in a similar manner, instead using convex signatures of E/M II and the midpoint on the convex curve adjoining E and M.

Cell culture and immunofluorescence

NSCLC cells were cultured in RPMI 1640 medium containing 10% fetal bovine serum and 1% penicillin/streptomycin cocktail (Thermo Fisher Scientific, Waltham, MA). For immunofluorescence, cells were fixed in 4% paraformaldehyde, permeabilized in 0.2% Triton X-100, and then stained with anti-CDH1 (1:100; Abcam) and anti-vimentin (1:100; Cell Signaling Technology). The primary antibodies were then detected with Alexa conjugated secondary antibodies (Life technologies). Nuclei were visualized by co-staining with DAPI.

Results

### Construction of the regulatory network of EMT

Based on an extensive literature search, we construct a core gene regulatory network of EMT including nine microRNAs (miR-141, miR-200a,b,c, miR-34a, miR-101, miR-30c, miR-9, miR-205), ten EMT-inducing transcription factors (EMT-TFs) (ZEB1, ZEB2, FOXC2, SNAI1, SNAI2, TWIST1, TWIST2, GSC, KLF8, TCF3), one EMT-inducing signal (TGF-β), three PSFs (OVOL2, GRHL2 and ΔNp63α), CLDN7 and the canonical epithelial and mesenchymal markers CDH1 and VIM respectively. The network of EMT presented here is an extended version of the one used in our previous study [32]. Specifically, three proposed PSFs, OVOL2, GRHL2 and ΔNp63α, together with CLDN7 are included due to their identified roles in epithelial-mesenchymal plasticity (See **SI Section 1** and **Table S1** for more information).

RACIPE identifies the gene expression patterns of multiple EMT phenotypes

We apply RACIPE to the core gene regulatory network above (**Fig. 1a**) and collect the gene expression data from 10,000 different sets of parameters (10,000 distinct RACIPE models). Hierarchical clustering analysis, used to analyze the gene expression patterns of RACIPE-generated data, reveals four large clusters (**Fig. 1b**). The first cluster is characterized by high expression of epithelial-associated genes or microRNAs, such as miR-200, miR-34, CLDN7 and CDH1, and low expression of mesenchymal-associated genes, such as ZEB1, FOXC2, SNAIL and VIM, thus identified as an epithelial signature. The second cluster shows low expression of epithelial-associated genes or microRNAs and high expression of mesenchymal-associated genes and therefore represents a mesenchymal signature. The remaining two clusters have co-expression of epithelial-associated and mesenchymal-associated genes to varying extents, and so these two clusters are characterized as hybrid E/M I and II signatures. Notably, the PSFs GRHL2, OVOL2 and ΔNp63α are expressed at high levels in hybrid E/M I samples, but not in hybrid E/M II samples. In contrast, miR-101 and GSC are expressed at high levels in hybrid II samples while not in hybrid E/M I samples. The

7

distinct gene expression profiles of hybrid E/M I and hybrid E/M II shown here emphasize the heterogeneity of one or more hybrid E/M phenotype(s) that cells may attain during EMT [35–37].
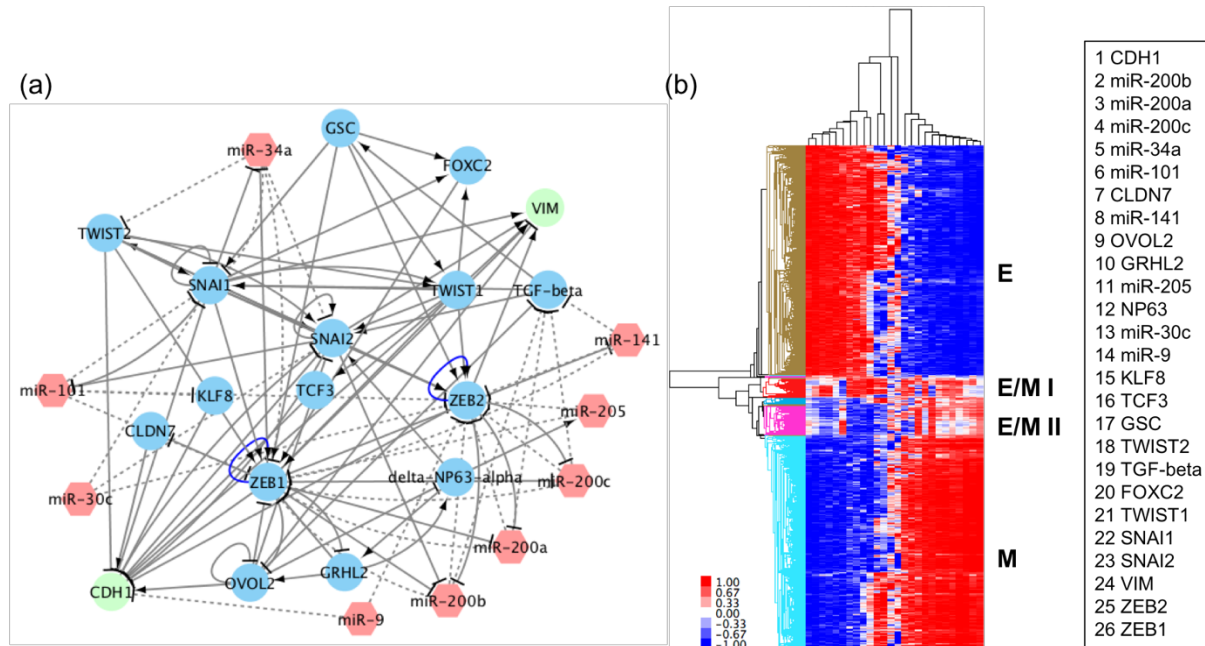

Figure 1. RACIPE identified robust gene expression patterns of EMT. (a) A proposed core EMT regulatory network. The EMT regulatory network contains 9 microRNAs (red hexagons) and 17 non-microRNAs (circles). Solid arrows represent excitatory regulation, mostly due to transcriptional activators. Solid bar-headed arrows represent inhibitory regulation, mostly due to transcriptional inhibitors. Dotted bar-headed arrows represent microRNA-mediated regulations. (b) Hierarchical clustering analysis of the RACIPE-generated gene expression data from 10,000 RACIPE models. Each row represents one sample, i.e., one RACIPE-generated stable steady state, and each column represents one gene. The names of genes corresponding to each column from left to right are listed.

The EMT scoring metric captures the hybrid E/M I and II signatures generated by RACIPE

The heterogeneity of RACIPE-generated aggregate output may be appreciated by projection of these samples onto the two-dimensional EMT metric predictor space (**Fig. 2a**). This space is divided into three colored areas: blue (mesenchymal), dark green (hybrid E/M), and red (epithelial). Samples belong to a particular category with high probability if they localize far from the boundary between two regions. We observe high variability within each phenotypic cluster when projecting RACIPE-generated data on this

predictor space, which is unsurprising given that solutions are generated from an ensemble of randomized parameters. Thus, we focus our attention to the average (mean) values of individual genes for all the samples in a given category, referred to as the mean signatures. The mean signatures of hybrid E/M I and E/M II samples are appropriately binned via the EMT metric – the probability that hybrid E/M I is categorized as hybrid E/M is 0.76 with intermediate EMT score of $\mu = 0.76$. and the probability of hybrid categorization for hybrid E/M II is 0.88 with intermediate EMT score of $\mu = 1.12$. The probability of hybrid categorization may be visually appreciated in **Fig. 2b**, wherein samples with larger probabilities of belonging to hybrid E/M localize inside the green region further from the interface between other groups. Our findings highlight the consistency between RACIPE-generated gene expression samples for hybrid E/M phenotypes and their expected location on the EMT spectrum as generated by our predictive metric. The mean signature of mesenchymal samples is also captured by the EMT metric with a 0.71 probability of being binned as mesenchymal, and a high EMT score of $\mu = 1.71$. (**Fig. 2**b). However, the mean signature of epithelial samples has a slightly higher probability of being binned as hybrid E/M ($P_H = 0.59$) as compared to epithelial ($P_E = 0.41$) with an EMT score of $\mu = 0.59$. This deviation is likely to be due to the underlying high variance of gene expression data generated by RACIPE. Nonetheless, the EMT metric has shown a consistent characterization results of the RACIPE-generated samples for the hybrid E/M I, II and M categories. We also find that the mean hybrid E/M I signature is close to that of epithelial samples, while the mean hybrid E/M II signature localizes closer to that of mesenchymal samples in the predictor space (**Fig. 2b**), which is consistent with the clustering results observed in **Fig. 1b**. Interestingly, the hybrid E/M II samples have similar mean VIM/CDH1 levels as that of epithelial samples and similar mean CLDN7 levels as that of mesenchymal samples, thus sharing features of both epithelial and mesenchymal signatures (Fig. 2b). This observation is reminiscent of a recent study showing that overexpression of miR-200 in mesenchymal subpopulation of HMLE cells induced only a partial EMT - the mesenchymal score of these cells did not change upon miR-200 overexpression, but the epithelial score increased [38].
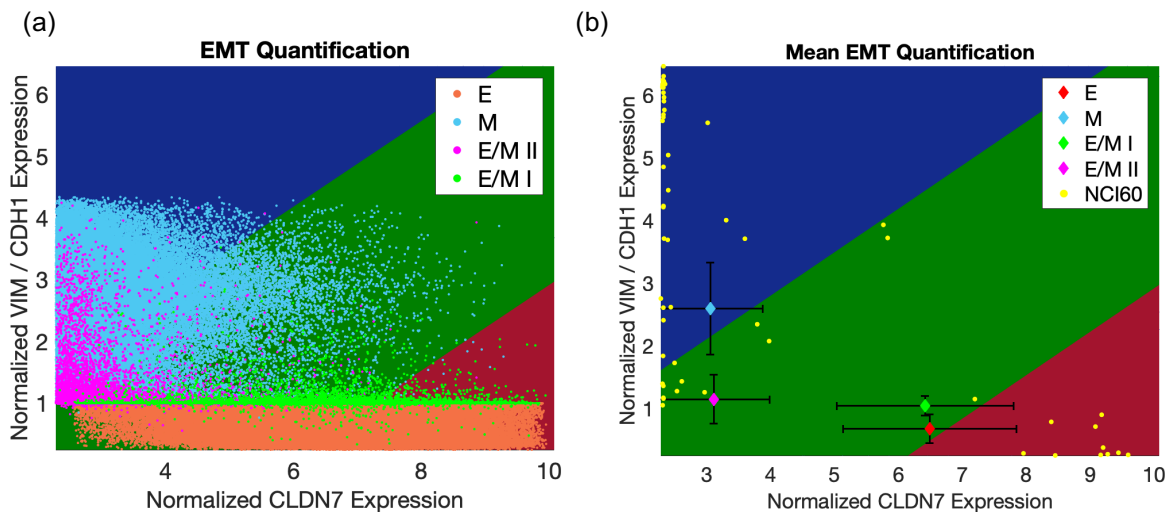
Figure 2. Characterization of the RACIPE-generated samples by the EMT metric. (a) Population level projection of RACIPE generated data onto EMT metric predictor space. VIM, CDH1, and CLDN7 expression values for each sample generated from RACIPE are used to plot various populations categorized by RACIPE (red region of EMT metric predictor space corresponding to E, dark green corresponding to hybrid E/M I and hybrid E/M II, and blue corresponding to M; salmon identifying RACIPE-generated E samples, light green identifying hybrid E/M I samples, magenta identifying hybrid E/M II samples and sky blue identifying M samples). (b) Predicted mean expression signature for each RACIPE population cluster. Each colored diamond represents the mean expression of RACIPE-predicted hybrid E/M I (light green), hybrid E/M II (magenta), E (salmon), and M (sky blue) samples, plotted alongside the NCI-60 training dataset (yellow). Black error bars represent standard deviations along each predictor dimension.

Using the EMT scoring metric to distinguish between samples comprised of individual hybrid E/M cells and samples that are mixtures of E and M cells

After establishing this initial success in predicting the population phenotype, we evaluated the ability of the EMT scoring metric to distinguish between populations consisting mostly of individual hybrid E/M cells versus samples that are mixtures of E and M cells. Unmixed samples of epithelial (EMT score $0 \leq \mu < 0.5$) or unmixed mesenchymal (EMT score $1.5 < \mu \leq 2$) cells are easily characterized by the EMT metric, which localizes their positions on the EMT predictor space to the red and blue regions, respectively **(Fig. 3)**. Categorizing a hybrid E/M (EMT score $0.5 \leq \mu \leq 1.5$) sample is less straightforward as there are multiple possibilities that explain such a prediction. The sample may: a) contain individual hybrid E/M cells, b) contain a mixture of epithelial and pure mesenchymal cells, or c) some combination of epithelial, mesenchymal, and hybrid E/M cells. As shown in **Fig. 4**, mixtures of RACIPE-generated epithelial and

mesenchymal samples may be well-approximated by a convex combination curve in the two-dimensional predictor space, hereafter referred to as a 'mixture curve', connecting the positions of mean signatures of pure epithelial and pure mesenchymal RACIPE-generated samples. Interestingly, the position of mean RACIPE-generated hybrid E/M signatures, particularly E/M II samples, concentrates further away from the mixture curve. Moreover, the higher percentage of the individual hybrid E/M II cells in the sample, the further the position of the sample from the mixture curve in the predictor space (Fig. 3). These results indicate that the EMT scoring metric may be used to predict the proportions of different subpopulations of samples that are mixtures of different EMT phenotypes.
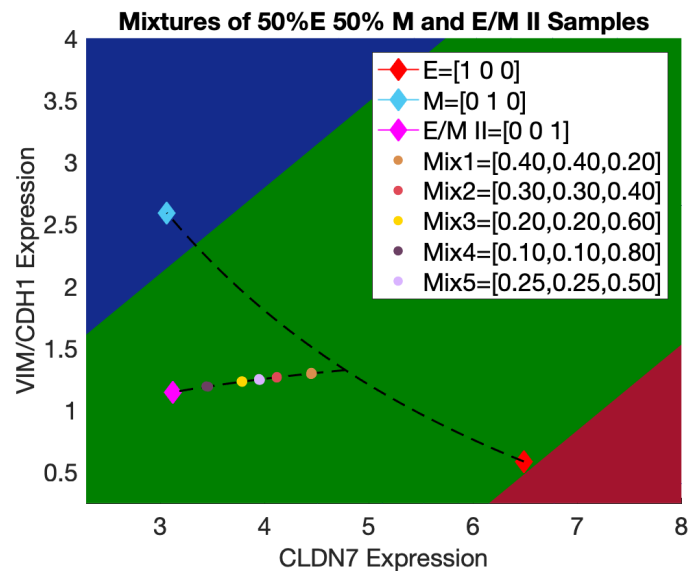
Figure 3. Use of the EMT metric to characterize the mixtures of E, M and hybrid E/M samples where the proportion of E is equal to that of M in each mixture. Predicted mixture proportions contained in each mixture {E, M, E/M II} are reported in the figure legends for each sample and predicted based on their projection to the convex mixture curve adjoining E/M II and the midpoint of E and M.

The EMT scoring metric can accurately resolve mixtures of pairs of EMT phenotypes

As the EMT metric has shown its power in distinguishing between individual hybrid E/M cells and mixtures of E and M cells, we evaluated the ability of the EMT scoring metric to resolve the proportions of two different EMT phenotypes mixed in varying ratios. Using RACIPE, we generated three pairs of *in silico* mixtures – a) mixtures of E and M samples, referred to as {E, M}, b) mixtures of E and hybrid E/M II samples, referred to as {E, E/M II} and c) mixtures of M and hybrid E/M II samples, referred to as {M, EM II} with five different ratios (see **Materials and Methods** for details). We used hybrid E/M II for generating mixtures because the mean expression signature of hybrid E/M II exhibits similar CLDN7 expression as that of a M phenotype and similar VIM/CDH1 level as that of an E phenotype, and is thus more likely to be a canonical representative 'hybrid' E/M phenotype. For every fixed ratio of a given pair of phenotypes, three independent replicates were generated, and they showed highly reproducible behavior (**Fig. 4**). The EMT metric-estimated proportions of E and M in {E, M} are [49%, 51%], [40%, 60%], [59%, 41%], [20%, 80%], [79%, 21%] with respective mean absolute errors of [0.44%, 0.45%, 0.55%, 0.86%, 1.37%] across replicates. These estimates are very close to the true proportions of E and M that are [50%, 50%], [40%, 60%], [60%, 40%], [20%, 80%], [80%, 20%], respectively (**Fig. 4a, d**). Similarly, the EMT metric-estimated proportions of E and hybrid E/M II in {E, E/M II} are converge to their true values of [50%, 50%], [60%, 40%], [40%, 60%], [20%, 80%], [80%, 20%] with respective mean errors of [0.26%, 0.25%, 0.23%, 0.04%, 0.11%] (Fig. 3b, d). Finally, the EMT metric-estimated proportions of M and hybrid E/M II in {M, E/M II} are [50%, 40%], [60%, 40%], [40%, 60%], [20%, 80%], [80%, 20%] across replicates with respective mean errors of [0.26%, 0.08%, 0.18%, 0.11%, 0.51%], again converging to their true values (Fig. 3c, d). Together, these results indicate that the EMT metric can effectively deconvolve mixtures of multiple EMT phenotypes, and thereby accurately estimate their proportions in a given mixture containing two different EMT phenotypes.
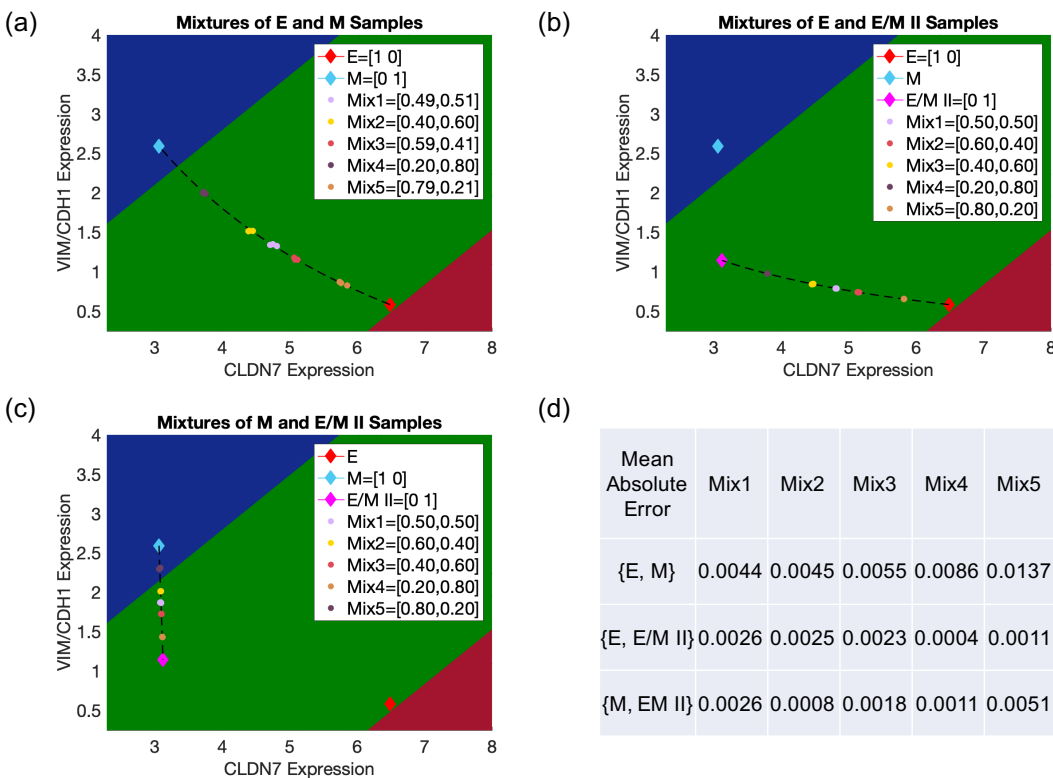
Figure 4. Use of the EMT scoring metric to characterize the mixtures of two different EMT phenotypes. Predicted mixture proportions contained in each mixture {E, M} (a), {E, E/M II} (b) and {M, E/M II} (c) are reported in the figure legends for each sample and predicted based on their projection to the curve of convex combinations for each non-mixed sample pair. In all cases, all three replicates for one mixture are tightly distributed and therefore have significant overlap. (d) The mean absolute error of predicted mixture proportions. The true mixture proportions of {E, M} are [50%, 50%] (Mix1) [40%, 60%] (Mix2) [60%, 40%] (Mix3) [20%, 80%] (Mix4) and [80%, 20%] (Mix5). The true mixture proportions of {E, E/M II} and {M, E/M II} are [50%, 50%] (Mix1) [60%, 40%] (Mix2) [40%, 60%] (Mix3) [20%, 80%] (Mix4) and [80%, 20%] (Mix5).

Motivated by the initial observation that the mean RACIPE-generated epithelial signature was binned as hybrid E/M, we further analyzed whether a mixture model constructed on samples binned as epithelial by the EMT metric would lead to significantly different results. We performed sample purification and isolated the purified epithelial samples that exhibit high expression of multiple E markers including CDH1, GRHL2, OVOL2, ΔNp63α and CLDN7. All of these players are known to be key regulators of epithelial phenotype [39–42]. Similarly, we performed sample purification to isolate the purified mesenchymal samples that have high expression of multiple M markers, including VIM, ZEB1, ZEB2, TGF-β and FOXC2. These

13

markers are experimentally known as key inducers of the mesenchymal phenotype [26,43]. This step is the *in silico* analogue to flow cytometric sorting of E (and M) samples for a more robust characterization of these two cell states, as captured by the increased accuracy of EMT metric in categorizing these samples (**Tables S2, S3**). The purified epithelial and mesenchymal samples are accurately predicted via EMT metric ($P_E = 0.82, \mu = 0.18$ and $P_M = 0.88, \mu = 1.88$, respectively), and visually fall well within their respective domains of predictor space (**Fig. S1, See SI sections 2-3 for more details**). Mixtures containing purified E and M samples for five different ratios were synthesized and analyzed by the EMT metric. Similar to the case with 'unpurified' samples (i.e. using all E samples), the EMT metric accurately predicts the proportions of the mixtures of 'purified' E and M samples (**Fig. S2**). In short, the sample purification analysis here was only to offer alternative reference points for E and M signatures which might be more representative of real biological samples.

The EMT scoring metric can accurately resolve mixtures of hybrid E/M, E and M phenotypes when E and M phenotypes have the same proportions

As the first step to evaluate the ability of the EMT metric in resolving mixtures of E, hybrid E/M and M phenotypes, we use the RACIPE-generated E, hybrid E/M II and M gene expression data to synthesize mixtures of all three phenotypes, referred to as {E, M, E/M II}. We keep the proportions of E and M samples the same and vary the proportions of the hybrid E/M II samples when preparing {E, M, E/M II}. Mixtures with five different ratios were randomly selected in triplicate and were highly reproducible and resolvable (Fig. 4). The EMT metric-estimated proportions of E, M and E/M II in {E, M, EM II} are [25.5%, 25.5%, 49%], [30%, 30%, 40%], [20%, 20%, 60%], [10%, 10%, 80%], [40%, 40%, 20%] with respective mean errors of [0.32%, 0.2%, 0.19%, 0.37%, 0.22%] across replicates which are very close to the true proportions of E, M and E/M II: [25%, 25%, 50%], [30%, 30%, 40%], [20%, 20%, 60%], [10%, 10%, 80%], [40%, 40%, 20%], respectively (**Fig. 3**). These results lead to the hypothesis that samples dominated by the hybrid E/M II phenotype tend to localize far from the 'mixture curve', while samples locating close to the 'mixture curve' are possible mixtures of E and M cells.

Using the EMT scoring metric to characterize the EMT proportions of NSCLC cell lines

To test the hypothesis that samples dominated by the hybrid E/M phenotype localize far from the 'mixture curve', we performed experiments in multiple NSCLC cell lines that have been categorized as epithelial, mesenchymal, or hybrid E/M based on bulk proteomic measurements (GSE63882) [36]. The EMT scoring metric characterizes H1568 cell line as largely epithelial, H1792 cell line as largely mesenchymal, and the H920, H1944, H2228 and HCC2279 cell lines as hybrid E/M (**Fig. 5a**). Among the hybrid E/M cell lines, H920 and H2228 locate far from the 'mixture curve' in the EMT metric predictor space, thus being

14

predicted to contain a large proportion of the hybrid E/M cells, while H1944 and HCC2279 cell lines lie on/close to the curve, which means that they are possible mixtures of E and M cells.

To validate this prediction, we conduct single-cell immunofluorescence experiments to analyze the presence of canonical epithelial and mesenchymal markers - E-cadherin and vimentin - in these cell lines. H1568 cells largely stain for E-cadherin, with only a few cells staining for vimentin, thus depicting a predominantly epithelial phenotype. Conversely, H1792 cells express only vimentin, and very low E-cadherin, exhibiting a predominantly mesenchymal phenotype (**Fig. 5b**). Most cells in H920 and H2228 tend to co-express E-cadherin and vimentin and some cells express only E-cadherin, thus demonstrating a mixture of hybrid E/M and epithelial cells, with a hybrid E/M phenotype being the dominant subpopulation (**Fig. 5b**). The H1944 cell line contains cells expressing either only E-cadherin or only vimentin thus being a mixture of E and M cells. This is consistent with the above EMT score finding. On the other hand, most individual cells in HCC2279 cell line co-express E-cadherin and vimentin, thus being hybrid E/M cells. Results of H1944 and HCC2279 suggest that a cell line lying on/close to the mixture curve might be a mixture of E and M but also might be a single-cell hybrid that just happens to lie on that line. Put together, this *in vitro* characterization of these six NSCLC cell lines offer a first proof-of-principle validation of our predictions about their EMT-ness, estimated by their relative distance to the 'mixture curve' in the predictor space of the EMT scoring metric.
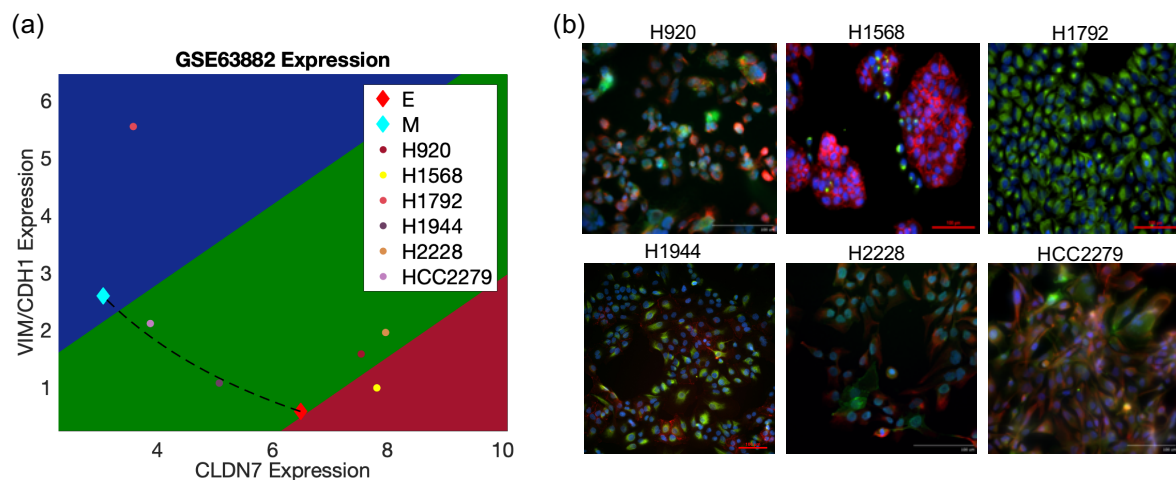
Figure 5. Characterizing the EMT phenotypes. (a) EMT expression signatures for H920, H1568, H1792, H1944, H2228, and HCC2279 are projected into EMT predictor space relative to the mixture curve between RACIPE-generated E and M samples. (b) Expression of E-cadherin (red) and vimentin (green) examined by immunofluorescence staining. Scale bar 100 μm.

15

Discussion

The existence and characterization of one or more hybrid epithelial/mesenchymal (E/M) phenotypes accessible during the epithelial-mesenchymal transition has generated recent attention due to their association with high metastatic potential and cancer stemness [35,44–46]. Increasing evidence suggests that cancer cells may acquire a spectrum of hybrid E/M phenotypes, characterized by varying extents of epithelial and mesenchymal traits [35]. Thus, a rigorous characterization of the EMT status of tumors, especially the probability that a given sample is hybrid E/M-like, may contain significant prognostic value and may hold avenues for better targeted therapeutic developments [47].

Recent years have seen a surge in various computational and experimental approaches to quantitatively characterize various aspects of EMT, including characterization of hybrid E/M phenotypes [17–21,25,26, 27–29,35,36,44,45]. Here, we test our EMT scoring metric by applying it to the *in silico* gene expression data generated by RACIPE. We show that the hybrid E/M II phenotype is effectively distinguished from mixtures of E and M samples, based on their different localizations in the predictor space generated by the EMT scoring metric. As compared to another EMT scoring method that requires data of the entire transcriptome [48], our approach presented here can quantify a sample's 'EMT-ness' using a very small set of predictors with reasonable categorization results.

Our results presented here may suffer from multiple limitations. First, the EMT regulatory network constructed here, although based on extensive literature search, is in no way complete. For the purpose of computational cost, we focused on a conserved network implicated in EMT in multiple contexts. However, there may be crucial connections that we have inadvertently missed. Second, we simulated both transcriptional factors and microRNA-dependent mechanisms via the shifted Hill function formulation, while their detailed dynamics may be quite different [25]. That means we must determine the EMT status of RACIPE models only by the "expression" data as opposed to using protein concentration and/or actual cell shape and behaviors. It is worth noting that EMT is a multi-dimensional process including changes in not only gene expression, but also in many biophysical properties such as cell polarity [49]. Thus, future studies should identify biophysical features that, when combined with gene expression patterns, might provide further refinement in resolving hybrid E/M phenotypes. Third, the training dataset used for EMT scoring metric are bulk-cell gene expression profiles of 59 cancer cell lines, and 37 out of the 59 are mesenchymal [50]. Further detailed experimental characterization of the EMT phenotype at a single-cell level [51] can help expand the training dataset and perhaps increase the sensitivity, specificity and accuracy in quantifying the EMT spectrum. Despite these weaknesses, we feel that we have succeeded in showing

16

broad consistency between the RACIPE and the EMT scoring metric and also generated a test for the existence of single cell hybrids.

The ability to fully distinguish between mixtures of E and M cells and purely hybrid E/M samples remains a relevant challenge as both of these share a common hybrid gene expression pattern on the population level. By projecting the mixtures and purely hybrid E/M cells into the predictor space of the EMT metric, we find that the mixtures of epithelial and mesenchymal tend to cluster along the convex curve adjoining purely epithelial and purely mesenchymal while samples containing purely hybrid E/M cells are not necessarily restricted to the mixture curve between unmixed epithelial and mesenchymal samples, as shown in **Figs. 3** and **4**. Based on this observation, we speculate that hybrid E/M cells and mixtures of epithelial and mesenchymal cells may often be distinguishable. Gene expression data of more cell lines with a well-characterized EMT status, i.e. purely epithelial, mesenchymal, hybrid E/M or mixtures of well-known proportions of these phenotypes, can be used to test this hypothesis.

References

[1]   Kalluri R and Weinberg R A 2009 The basics of epithelial-mesenchymal transition *J. Clin. Invest.* **119** 1420–8

[2]   Jolly M K, Ward C, Eapen M S, Myers S, Hallgren O, Levine H and Sohal S S 2018 Epithelial-mesenchymal transition, a spectrum of states: Role in lung development, homeostasis, and disease *Dev. Dyn.* **247** 346–58

[3]   Micalizzi D S, Farabaugh S M and Ford H L 2010 Epithelial-mesenchymal transition in cancer: parallels between normal development and tumor progression. *J. Mammary Gland Biol. Neoplasia* **15** 117–34

[4]   Fischer K R, Durrans A, Lee S, Sheng J, Li F, Wong S T C, Choi H, El Rayes T, Ryu S, Troeger J, Schwabe R F, Vahdat L T, Altorki N K, Mittal V and Gao D 2015 Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance *Nature* **527** 472–6

[5]   Zheng X, Carstens J L, Kim J, Scheible M, Kaye J, Sugimoto H, Wu C-C, LeBleu V S and Kalluri R 2015 Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer *Nature* **527** 525–30

[6]   Tripathi S C, Peters H L, Taguchi A, Katayama H, Wang H, Momin A, Jolly M K, Celiktas M, Rodriguez-Canales J, Liu H, Behrens C, Wistuba I I, Ben-Jacob E, Levine H, Molldrem J J, Hanash S M and Ostrin E J 2016 Immunoproteasome deficiency is a feature of non-small cell lung cancer with a mesenchymal phenotype and is associated with a poor outcome *Proc. Natl. Acad. Sci.* **113** E1555-64

[7]   Cheung K J and Ewald A J 2016 A collective route to metastasis: Seeding by tumor cell clusters. *Science* **352** 167–9

[8]   Aceto N, Bardia A, Miyamoto D T, Donaldson M C, Wittner B S, Spencer J A, Yu M, Pely A, Engstrom A, Zhu H, Brannigan B W, Kapur R, Stott S L, Shioda T, Ramaswamy S, Ting D T, Lin C P, Toner M, Haber D A and Maheswaran S 2014 Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis *Cell* **158** 1110–22

[9]   Sarioglu A F, Aceto N, Kojic N, Donaldson M C, Zeinali M, Hamza B, Engstrom A, Zhu H, Sundaresan T K, Miyamoto D T, Luo X, Bardia A, Wittner B S, Ramaswamy S, Shioda T, Ting D T, Stott S L, Kapur R, Maheswaran S, Haber D A and Toner M 2015 A microfluidic device for label-free, physical capture of circulating tumor cell clusters *Nat. Methods* **12** 685–91

[10] Jolly M K, Boareto M, Debeb B G, Aceto N, Farach-Carson M C, Woodward W A and Levine H 2017 Inflammatory Breast Cancer: a model for investigating cluster-based dissemination *NPJ Breast Cancer* **3** 21

[11] Cheung K J, Padmanaban V, Silvestri V, Schipper K, Cohen J D, Fairchild A N, Gorin M A, Verdone J E, Pienta K J, Bader J S and Ewald A J 2016 Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters *Proc. Natl. Acad. Sci.* **113** E854–E863

[12] Jolly M, Boareto M, Huang B, Jia D, Lu M, Ben-Jacob E, Onuchic J N and Levine H 2015 Implications of the hybrid epithelial/mesenchymal phenotype in metastasis *Front. Oncol.* **5** 155

[13] Shibue T and Weinberg R A 2017 EMT, CSCs, and drug resistance: the mechanistic link and clinical implications *Nat. Rev. Clin. Oncol.* **14** 611–29

[14] Jolly M K, Mani S A and Levine H 2018 Hybrid epithelial/mesenchymal phenotype(s): The 'fittest' for metastasis? *Biochim. Biophys. Acta BBA - Rev. Cancer* **1870** 151–7

[15] Lee J M, Dedhar S, Kalluri R and Thompson E W 2006 The epithelial–mesenchymal transition: new insights in signaling, development, and disease *J. Cell Biol.* **172** 973–81

[16] Jordan N V, Johnson G L and Abell A N 2011 Tracking the intermediate stages of epithelial-mesenchymal transition in epithelial stem cells and cancer *Cell Cycle* **10** 2865–73

[17] 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

[18] Bocci F, Jolly M K, Tripathi S C, Aguilar M, Hanash S M, Levine H and Onuchic J N 2017 Numb prevents a complete epithelial-mesenchymal transition by modulating Notch signalling *J. R. Soc. Interface* **14** 136

[19] Jolly M K, Tripathi S C, Jia D, Mooney S M, Celiktas M, Hanash S M, Mani S A, Pienta K J, Ben-Jacob E, Levine H, Jolly M K, Tripathi S C, Jia D, Mooney S M, Celiktas M, Hanash S M, Mani S A, Pienta K J, Ben-Jacob E and Levine H 2016 Stability of the hybrid epithelial/mesenchymal phenotype *Oncotarget* **7** 27067–84

[20] Dang T T, Esparza M A, Maine E A, Westcott J M and Pearson G W 2015 ΔNp63α promotes breast cancer cell motility through the selective activation of components of the epithelial-to-mesenchymal transition program *Cancer Res.* **75** 3925–35

[21] Jia D, Jolly M K, Boareto M, Parsana P, Mooney S M, Pienta K J, Levine H and Ben-Jacob E 2015 OVOL guides the epithelial-hybrid-mesenchymal transition *Oncotarget* **6** 15436–48

[22] Bocci F, Tripathi S C, George J T, Casabar J, Mercedes S V, Wong P, Hanash S, Levine H, Onuchic J and Jolly M 2018 NRF2 activates a partial Epithelial-Mesenchymal Transition and is maximally present in a hybrid Epithelial/Mesenchymal phenotype *biorxiv* 390237

[23] De Craene B and Berx G 2013 Regulatory networks defining EMT during cancer initiation and progression. *Nat. Rev. Cancer* **13** 97–110

[24] Bhatia S, Monkman J, Toh A K L, Nagaraj S H and Thompson E W 2017 Targeting epithelial–mesenchymal plasticity in cancer: clinical and preclinical advances in therapy and monitoring *Biochem. J.* **474** 3269–306

[25] 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.* **110** 18144–9

[26] Jia D, Jolly M K, Tripathi S C, Hollander P Den, Huang B, Lu M, Celiktas M, Ramirez-Pena E, Ben-Jacob E, Onuchic J N, Hanash S M, Mani S A and Levine H 2017 Distinguishing Mechanisms Underlying EMT Tristability *Cancer Converg.* **1** 2

[27] Steinway S N, Zañudo J G T, Michel P J, Feith D J, Loughran T P and Albert R 2015 Combinatorial interventions inhibit TGFβ-driven epithelial-to-mesenchymal transition and support hybrid cellular phenotypes *Npj Syst. Biol. Appl.* **1** 15014

[28] 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** 1079–89

[29] 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

[30] Font-Clos F, Zapperi S and Porta C A M L 2018 Topography of epithelial–mesenchymal plasticity *Proc. Natl. Acad. Sci.* **115** 5902–7

[31] Steinway S N, Gomez Tejeda Zañudo J, Ding W, Rountree C B, Feith D J, Loughran T P and Albert R 2014 Network modeling of TGFβ signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint Sonic hedgehog and Wnt pathway activation. *Cancer Res.* **74** 5963–77

[32] 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

[33] Huang B, Jia D, Feng J, Levine H, Onuchic J N and Lu M 2018 RACIPE: a computational tool for modeling gene regulatory circuits using randomization *BMC Syst. Biol.* **12** 74

[34] George J T, Jolly M K, Xu S, Somarelli J A and Levine H 2017 Survival Outcomes in Cancer Patients Predicted by a Partial EMT Gene Expression Scoring Metric *Cancer Res.* **77** 6415–28

[35] Pastushenko I, Brisebarre A, Sifrim A, Fioramonti M, Revenco T, Boumahdi S, Keymeulen A V, Brown D, Moers V, Lemaire S, Clercq S D, Minguijón E, Balsat C, Sokolow Y, Dubois C, Cock F D, Scozzaro S, Sopena F, Lanas A, D'Haene N, Salmon I, Marine J-C, Voet T, Sotiropoulou P A and Blanpain C 2018 Identification of the tumour transition states occurring during EMT *Nature* **556** 463–8

[36] Schliekelman M J, Taguchi A, Zhu J, Dai X, Rodriguez J, Celiktas M, Zhang Q, Chin A, Wong C-H, Wang H, McFerrin L, Selamat S A, Yang C, Kroh E M, Garg K S, Behrens C, Gazdar A F, Laird-Offringa I A, Tewari M, Wistuba I I, Thiery J P and Hanash S M 2015 Molecular Portraits of Epithelial, Mesenchymal, and Hybrid States in Lung Adenocarcinoma and Their Relevance to Survival *Cancer Res.* **75** 1789–800

[37] Anon Functional Balance between TCF21-Slug defines phenotypic plasticity and sub-classes in high-grade serous ovarian cancer | bioRxiv

[38] Cursons J, Pillman K A, Scheer K G, Gregory P A, Foroutan M, Hediyeh-Zadeh S, Toubia J, Crampin E J, Goodall G J, Bracken C P and Davis M J 2018 Combinatorial Targeting by MicroRNAs Co-ordinates Post-transcriptional Control of EMT *Cell Syst.* **7** 77-91.e7

[39] Roca H, Hernandez J, Weidner S, McEachin R C, Fuller D, Sud S, Schumann T, Wilkinson J E, Zaslavsky A, Li H, Maher C A, Daignault-Newton S, Healy P N and Pienta K J 2013 Transcription Factors OVOL1 and OVOL2 Induce the Mesenchymal to Epithelial Transition in Human Cancer *PLoS ONE* **8** e76773

[40] Mooney S M, Talebian V, Jolly M K, Jia D, Gromala M, Levine H and McConkey B J 2017 The GRHL2/ZEB Feedback Loop-A Key Axis in the Regulation of EMT in Breast Cancer *J. Cell. Biochem.* **118** 2559–70

[41] Bhat A A, Pope J L, Smith J J, Ahmad R, Chen X, Washington M K, Beauchamp R D, Singh A B and Dhawan P 2015 Claudin-7 expression induces mesenchymal to epithelial transformation (MET) to inhibit colon tumorigenesis *Oncogene* **34** 4570–80

[42] Tran M N, Choi W, Matthew F, Navai N, Lee I C, Nitti G, Wen S, Flores E R, Siefker-radtke A, Czerniak B, Dinney C, Barton M, David J, Wszolek M F and Mcconkey D J 2013 The p63 Protein Isoform ΔNp63α Inhibits Epithelial-Mesenchymal Transition in Human Bladder Cancer Cells: Role of miR-205 *J. Biol. Chem.* **288** 3275–88

[43] Taube J H, Herschkowitz J I, Komurov K, Zhou A Y, Gupta S, Yang J, Hartwell K, Onder T T, Gupta P B, Evans K W, Hollier B G, Ram P T, Lander E S, Rosen J M, Weinberg R A and Mani S A 2010 Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc. Natl. Acad. Sci. U. S. A.* **107** 15449–54

[44] Grosse-Wilde A, Fouquier d' Herouei A, McIntosh E, Ertaylan G, Skupin A, Kuestner R E, del 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

[45] Bierie B, Pierce S E, Kroeger C, Stover D G, Pattabiraman D R, Thiru P, Liu Donaher J, Reinhardt F, Chaffer C L, Keckesova Z and Weinberg R A 2017 Integrin-β4 identifies cancer stem cell-enriched populations of partially mesenchymal carcinoma cells *Proc. Natl. Acad. Sci.* **114** E2337-2346

[46] Bocci F, Jolly M K, George J T, Levine H, Onuchic J N, Bocci F, Jolly M K, George J T, Levine H and Onuchic J N 2018 A mechanism-based computational model to capture the interconnections among epithelial-mesenchymal transition, cancer stem cells and Notch-Jagged signaling *Oncotarget* **9** 29906–20

[47] Biddle A, Gammon L, Liang X, Costea D E and Mackenzie I C 2016 Phenotypic Plasticity Determines Cancer Stem Cell Therapeutic Resistance in Oral Squamous Cell Carcinoma *EBioMedicine* **4** 138–45

[48] Tan T Z, Miow Q H, Miki Y, Noda T, Mori S, Huang R Y 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** 1279–93

[49] Jolly M K, Ware K E, Gilja S, Somarelli J A and Levine H 2017 EMT and MET : necessary or permissive for metastasis ? *Mol. Oncol.* **11** 755–69

[50] Park S-M M, Gaur A B, Lengyel E and Peter M E 2008 The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes Dev.* **22** 894–907

[51] Jolly M K, Somarelli J A, Sheth M, Biddle A, Tripathi S C, Armstrong A J, Hanash S M, Bapat S A, Rangarajan A and Levine H 2018 Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas *Pharmacol. Ther.* In Press.