Published in final edited form as:
Cancer Res. 2017 November 15; 77(22): 6415–6428. doi:10.1158/0008-5472.CAN-16-3521.

Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric

Jason T. George[1,2,4,*], **Mohit Kumar Jolly**[1,*], **Shengnan Xu**[5], **Jason A. Somarelli**[5], and **Herbert Levine**[1,2,3,†]

[1]Center for Theoretical Biological Physics, Rice University, 6100 Main Street, Houston, TX 77005

[2]Department of Bioengineering, Rice University, 6100 Main Street, Houston, TX 77005

[3]Department of Physics and Astronomy, Rice University, 6100 Main Street, Houston, TX 77005

[4]Medical Scientist Training Program, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX, 77030

[5]Duke Cancer Institute & Department of Medicine, Duke University Medical Center, Durham, NC 27708

Abstract

Metastasis is a significant contributor to morbidity and mortality for many cancer patients and remains a major obstacle for effective treatment. In many tissue types, metastasis is fueled by the epithelial-to-mesenchymal transition (EMT) - a dynamic process characterized by phenotypic and morphologic changes concomitant with increased migratory and invasive potential. Recent experimental and theoretical evidence suggests that cells can be stably halted *en route* to EMT in a hybrid E/M phenotype. Cells in this phenotype tend to move collectively, forming clusters of circulating-tumor-cells that are key tumor-initiating agents. Here, we developed an inferential model built on the gene expression of multiple cancer subtypes to devise an EMT metric that characterizes the degree to which a given cell line exhibits hybrid E/M features. Our model identified drivers and fine-tuners of epithelial-mesenchymal plasticity and recapitulated the behavior observed in multiple *in vitro* experiments across cancer types. We also predicted and experimentally validated the hybrid E/M status of certain cancer cell lines, including DU145 and A549. Lastly, we demonstrated the relevance of predicted EMT scores to patient survival and observed that the role of the hybrid E/M phenotype in characterizing tumor aggressiveness is tissue- and subtype-specific. Our algorithm is a promising tool to quantify the EMT spectrum, to investigate the correlation of EMT score with cancer treatment response and survival, and to provide an important metric for systematic clinical risk stratification and treatment.

Keywords

Hybrid E/M; computational biology; partial EMT; machine learning; gene expression profiling

[†]Corresponding Author: Herbert Levine (herbert.levine@rice.edu).
[*]These authors contributed equally.

The authors declare no potential conflicts of interest.

Major Findings

We develop an iterative method that ranks candidate gene products based on their ability to resolve NCI-60 cohort samples with regard to their respective EMT status, and construct a metric that quantifies the EMT spectrum. We validate model predictions by correctly recapitulating multiple *in vitro* experiments containing samples with well-established EMT status. We then demonstrate the utility of our metric by identifying certain hybrid E/M cell lines, followed by experimental validation via immunofluorescence and single-cell analysis. Lastly, we demonstrate the relevance of EMT-state predictions to cancer progression across multiple cancer types by comparing differences in patient survival among the three predicted categories (E, E/M, M).

Quick Guide to Equations and Assumptions

Equations

The approach outlined in Materials and Methods effectively creates many statistical models based on combinations of predictors selected from a large pool of EMT-relevant genes. These models are all created using ordinal multinomial logistic regression (MLR). MLR allows output predictions to categorize more than two (in this case three) distinct groups. Ordinal regression is employed to indicate the order structure between groups, whereby the hybrid E/M state is appropriately placed intermediary to E & M. Each model, m, may be represented either by its regression coefficients, $\beta = (\alpha_1, \alpha_2, \beta_1, \beta_2)$, or by its collection of output classifiers, $\tilde{\pi}(m)$. In this way, the output of model m for sample s is indicated by $\left\{\hat{\pi}^{(m)}_{s,1}, \hat{\pi}^{(m)}_{s,2}\right\}$ where $\hat{\pi}^{(m)}_{s,k}$ is model m's best assessment that sample s belongs to one of the groups from 1 to k (k ranges from 1 to 2, and $\hat{\pi}^{(m)}_{s,3}=1$). Predictions from each model may be compared with known observations in the training set to produce a deviance, D. The best fit model may be identified by selecting the model with maximal log-likelihood. This is equivalent to minimizing D, given by

$$D(m)=2\sum_{j=1}^{N}\sum_{k=1}^{3}Y_{j,k}(\log Y_{j,k}-\log\hat{\pi}^{(m)}_{j,k}) \qquad (1)$$

Here, N represents the number of samples in the training set (~60), j the index for each sample, k the index for each of the three categories, $Y_{j,k}$ the observable categories, $\hat{\pi}^{(m)}_{j,k}$ the fitted, cumulative distribution value for the j^{th} observation, and log $Y_{j,k}$ the maximal attainable log-likelihood value.

By minimizing over all combinations of predictors, we may generate a model that best classifies a given training set into 1 of 3 ordered ($E < E/M < M$) categories using two predictors. The relationship between regression coefficients (α_1, α_2, β_1, β_2) is given by

$$\log\left(\frac{\hat{\pi}_{j,k}}{1-\hat{\pi}_{j,k}}\right)=\alpha_k-(\beta_1 X_{j,1}+\beta_2 X_{j,2}), \qquad (2)$$

defined for $k = 1,2$, where $X_{j,1}$ and $X_{j,2}$ represent the j^{th} sample values for predictors 1 and 2, respectively. In this context, the cumulative probabilities may be given for each category k (belonging to one of $\{E,E/M,M\}$) by:

$$\hat{\pi}_{j,k}=\mathbb{P}(Y_j \le k)=\begin{cases} \frac{e^{\alpha_1-(\beta_1 X_{j,1}+\beta_2 X_{j,2})}}{1+e^{\alpha_1-(\beta_1 X_{j,1}+\beta_2 X_{j,2})}}, & k=1; \\ \frac{e^{\alpha_2-(\beta_1 X_{j,1}+\beta_2 X_{j,2})}}{1+e^{\alpha_2-(\beta_1 X_{j,1}+\beta_2 X_{j,2})}}, & k=2; \\ 1, & k=3. \end{cases} \qquad (3)$$

This provides an explicit representation for the categorical probabilities as:

$$\mathbb{P}(Y_j=n)=\begin{cases} \frac{e^{\alpha_1-(\beta_1 X_{j,1}+\beta_2 X_{j,2})}}{1+e^{\alpha_1-(\beta_1 X_{j,1}+\beta_2 X_{j,2})}}, & n=E; \\ \frac{e^{\alpha_2-(\beta_1 X_{j,1}+\beta_2 X_{j,2})}}{1+e^{\alpha_2-(\beta_1 X_{j,1}+\beta_2 X_{j,2})}}-\frac{e^{\alpha_1-(\beta_1 X_{j,1}+\beta_2 X_{j,2})}}{1+e^{\alpha_1-(\beta_1 X_{j,1}+\beta_2 X_{j,2})}}, & n=E/M; \\ 1-\frac{e^{\alpha_2-(\beta_1 X_{j,1}+\beta_2 X_{j,2})}}{1+e^{\alpha_2-(\beta_1 X_{j,1}+\beta_2 X_{j,2})}}, & n=M. \end{cases} \qquad (4)$$

As stated above, ordinal MLR places order structure on categories consistent with the belief that the hybrid E/M cells fall in a region between E and M. Using this characterization, we propose the EMT metric, μ, defined in relation to the probability of obtaining a hybrid, P_H (Equation 5). P_H is calculated by Equation 4 with $n = E/M$ (P_E with $n = E$, P_M with $n = M$), and μ may take values in [0,2], with the value $\mu = 0$ interpreted as a purely E signature, $\mu =$ 2a purely M signature, and $\mu =$ 1a maximally hybrid E/M signature.

$$\mu(Y_i)=\begin{cases} P_H, & P_E>P_M; \\ 2-P_H, & P_E<P_M; \\ 1, & P_E=P_M. \end{cases} \qquad (5)$$

In working with large datasets, we may characterize the distribution of EMT scores for a given cancer subtype. This is graphically represented by plotting a histogram of the sample partitioned across [0, 2] into 20 equally-spaced bins, from which an empirical probability density can be approximated by spline interpolation of the histogram.

Assumptions

The model assumes that the major features of EMT may be characterized in a general sense by gene expression signatures. Ordinal logistic regression requires that an order structure exist among the categories to be predicted. In this case, E/M is intermediate to E and M. Additionally, the model assumes a proportional response that is the same for each category with regard to changes in predictor levels. Model normalization assumes that systematic differences across experimental setups and gene expression platforms can be captured by comparing the relative levels of a small collection (~20) of gene products that the model predicts to be least correlated with respect to EMT. Lastly, our extension of the model to primary tissue samples assumes that differences other than those accounted for in the normalization step between the training and test sets are minimal.

Body Text

Introduction

Epithelial-to-Mesenchymal Transition (EMT) is a critical phenomenon during tumor progression that can drive metastasis, tumor-initiation potential, resistance to anoikis, refractory response to chemotherapy, and immune system evasion [1–3]. Accumulating evidence in cell lines, primary tumors, mouse models, and circulating tumor cells (CTCs) across multiple tumor types has indicated that EMT is not an all-or-none process, but rather that cells can exhibit a mix of epithelial and mesenchymal traits such as (a) co-expression of epithelial (CDH1, EpCAM) and mesenchymal (VIM, CDH2, Zeb1, SNAI2) markers, and (b) collective cell migration by giving rise to clusters of CTCs [1, 4–7]. The enhanced metastatic potential of these clusters as compared to individually migrating ones, a poor prognosis associated with co-expression of epithelial and mesenchymal markers instead of solely mesenchymal markers, and a predominance of such hybrid epithelial/mesenchymal (E/M) cells in highly aggressive cancers such as melanomas and triple negative breast cancer (TNBC) strongly argue for a hybrid E/M phenotype to be construed as a hallmark of cancer aggressiveness [1, 5–10].

Despite its paramount importance in driving tumor progression, a hybrid E/M phenotype remains poorly characterized largely due to a lack of quantitative gene expression data at different time points during EMT or its reverse Mesenchymal-to-Epithelial Transition (MET). Moreover, the hybrid E/M phenotype has been tacitly assumed to be metastable or transient [11]. Recent studies, however, have challenged this assumption by demonstrating that a hybrid E/M phenotype can be stably maintained *in vitro* at a single-cell level, especially under the influence of factors such as GRHL2 and OVOL2 that contribute to the stability of a hybrid E/M phenotype [12–14]. These factors are referred to as ‘phenotypic stability factors’ (PSFs), and their elevated expression, indicative of a stable hybrid E/M levels, are associated with worse patient survival [12].

Here, we devise an iterative statistical model built upon the gene expression profiles from multiple cancer subtypes that can quantitatively predict where a given sample lies on the EMT spectrum. The model can categorize the NCI-60 cohort of cell lines into epithelial, mesenchymal and hybrid E/M phenotypes with high specificity, sensitivity and accuracy, while only using a small set of predictors. Furthermore, it validates the relevance of PSFs in stabilizing the hybrid E/M phenotype, captures the different EMT score for various conditions such as EMT induction and multiple isogenic subpopulations, and can correlate EMT status with clinical outcome across different tumor types. This statistical model illustrates common molecular features associated with EMT across multiple contexts and tissue types, and will be crucial to further our understanding of a hybrid E/M phenotype in tumor progression.

Materials and Methods

In order to develop a quantification of EMT that incorporates the E/M phenotype, iterative multinomial logistic regression (MLR) in two dimensions is applied to the NCI-60 training set to find the pair of predictors (i.e. genes) best able to resolve each phenotype. The output of the model is modified to create an EMT-metric by which additional samples may be characterized (Equations 4,5). All datasets were obtained from the National Center for Biotechnology Information Gene Expression Onmibus (GEO) portal and identified by their GEO ID, unless otherwise noted. Model construction and predictions were performed using MATLab R2015b, along with its Curve Fitting Toolbox™ and Statistics and Machine Learning Toolbox™. Additional explanations, supporting information for the model, and a complete list of experimental procedures may be found in 'Quick Guide To Equations and Assumptions' section as well as 'Supplementary Information' and 'Supplementary Data'.

Training set classification—We primarily require that the model represent a generalized characterization of EMT, which can then be applied to a number of tissue types. Consequently, the training set must contain a broad collection of cancer subtypes. The NCI-60 cohort of cell lines (GSE5846) is selected as the training set because of its diverse collection of cancer types. Additionally, previous empirical investigations using VIM and CDH1 protein markers have categorized this data into E, M, and E/M categories [15], which are used as the observable categories.

Feature selection—A list of EMT-relevant candidate genes is complied from the literature and employed as the space of possible EMT-predictors, significantly reducing the high dimensional input space of all possible gene products [16–20] (Supplementary Data). This restriction helps to mitigate over-fitting by partially eliminating sources of variability extraneous to the problem at hand. A list of these features, along with simple combinations (for example, the ratio of two canonical epithelial and mesenchymal genes such as E-cadherin and vimentin - CDH1/VIM, or that of a canonical mesenchymal gene and a typical 'phenotypic stability factor' for a hybrid E/M phenotype [12] - GRHL2/VIM) for a subset of these genes are utilized as the set of candidate predictors in the training of NCI-60 data (see Supplementary Data). We limit our extension of ratios to a subset as finding the top two predictors out of relevant transcripts and their ratios would be computationally infeasible. Over-fitting may also occur by incorporating a large number of predictors, thereby reducing model predictive power [21, 22]. This risk was minimized by only considering up to two candidate predictors in combination. Although it is computationally infeasible to find the best three predictors in combination, we characterize the change in sensitivity and specificity when adding the next-best predictor individually to the top 50 predictor pairs.

Selected candidate predictors are ranked by ordering the list of all combinations of candidate genes according to their ability to fit the training set (Table 1). Better candidate predictor combinations are characterized by lower deviance (D) scores, which are calculated via MATLab's built-in 'mnrfit.m' function (Equation 1). Minimizing D corresponds to a higher maximum likelihood estimate, which gives a better overall fit to the training data. The best predictor combination is obtained by selecting candidate predictors with the lowest value of D. Although only two predictors are ultimately used for sample classification, the procedure also orders candidate predictors based on their individual ability to resolve EMT.

Model construction—The model is constructed using supervised machine learning on the NCI-60 training data. MLR is applied to each pair of potential predictors. MLR is employed as it is an effective tool in handling categorical data with a continuous input (e.g. gene expression data). An explicit description of the intermediate state (as opposed to description relative to the distance between E and M extremes) was one of the main advantages of our approach. Ordinal regression is assumed, with E<E/M<M, since the E/M phenotype is known to share features of both E and M and it seems reasonable to suppose that E/M cells exist in a state that is intermediate to both E and M. In order to ensure that the ultimate model indeed characterizes the training data, deviances are calculated for 10^6 similar statistical models with two predictors randomly chosen out of the same EMT-relevant feature selection pool.

Cross-validation—The result of applying MLR on the predictor combination, (X_1, X_2), is a set of regression coefficients, β_i, which can be used to predict the EMT-status of unknown samples (Equations 2–3). Leave-one-out analysis was employed in order to characterize the predictive capability of the model and ensure that the algorithm was not significantly over-fitting the training data. In this step, statistical regression is constructed identically as before, but this time using all but one sample in the NCI-60 training set. The regression is then applied to predict the category of the withheld sample. Sensitivities and specificities are estimated by repeating this procedure, withholding a different sample each time.

Normalization—Systematic differences in expression values as a result of different experiments and cross-platform analysis lead to variability in gene expression that may significantly affect predictions using the model trained on NCI-60. Normalization is performed prior to each analysis in order to make a more appropriate comparison between the model regression coefficients and new samples. Toward this end, MLR is performed on the training set as before, this time using individual genes only. This is iterated for every gene product available, now a much larger collection of genes than the set used for model construction. The output of this step is a list of gene products based on their individual ability to resolve {E, E/M, M} phenotypes in training data. This list is sorted to prioritize genes least capable of resolving categories. The top genes are those most agnostic to EMT status and play a similar role in our analysis to housekeeping genes used for establishing baseline expression profiles. In order to prevent over reliance on a single normalizer, the 20 lowest-ranked gene products that show non-saturated signals in the training set are selected as normalizers. Once selected, expression values for each of these genes in the training set are averaged together. Similarly, the expression values for the same genes are averaged in the test (NCI-60) set. The systematic difference in average expression of these normalizers is applied uniformly to all genes in the test set as follows: Average gene expression values for this collection create a background expression profile for both the training set (E_{train}) and the test set (E_{test}). The net differences in background expression, $\Delta \equiv E_{test} - E_{train}$, is subtracted from each expression values in the test set for fair predictions (for example, if there is no difference in background expression then $\Delta = 0$ and no net correction is required). As stated above, the role of these genes is similar to utilizing the housekeeping genes as relative measures of consistent expression. Here, however, these gene products have been shown to remain consistent regardless of EMT status.

Occasionally, gene signatures exist that fall far outside the domain of reasonable expression levels post-normalization. The model can still assign an EMT score to such samples, but the validity of such predictions becomes questionable. To filter anomalous data, samples designated as outliers are withheld from EMT metric assignment. Outliers are samples which fall outside of range (greater than 5-fold on either axis, when compared to the total range of NCI-60 data) not only for the top predictor (X_1, X_2), but also for the next two top predictors as well. This is a generous range relative to allowable maximum and minimum-fold values seen across all training set samples.

EMT metric—The mRNA expression values ($\log_2$-normalized) for the predictors identified in the feature selection step are used as input to the model. The output for each sample is an ordered triple, (P_E, P_H, P_M), that may be interpreted as the probability of falling into each phenotype. Categorical predictions are made by binning samples based on the type with maximal probability. In order to provide quantitative estimates of EMT, samples are given a score, μ, ranging from 0 (pure E) to 2 (pure M), with a score of 1 indicating a maximal hybrid E/M phenotype (Equation 5). In particular, $0 < \mu < 0.5$ corresponds to an epithelial prediction, $0.5 \leq \mu \leq 1.5$ to a hybrid E/M prediction, and $1.5 < \mu < 2$ to a mesenchymal prediction.

Cell line validation and prediction—Gene expression profiles of EMT-relevant cell lines and experimental treatments are analyzed to evaluate the consistency between the model output and established empirical observations. In each of these cases, the EMT score, μ, is used in predictions. The predictive algorithm was applied to samples for previously reported EMT status in order to compare EMT categorization with known results. Additional predictions were made on datasets with unknown EMT state. Lastly, the model was applied to large sample TCGA datasets with available gene expression signatures to provide a distribution for the extent of EMT in multiple cancer subtypes. The results were normalized to represent empirical probability density functions, and the relevant histograms were smoothed using cubic spline interpolation.

Survival analysis—EMT scores are generated for various patient primary tumor samples containing both gene expression and survival metrics. Observed survival distributions are graphically displayed for all three categories using Kaplan-Meier plots, and significant differences in survival metrics among each category were pairwise assessed using the log-rank test at significance level $\alpha = 0.05$.

Cell lines and culture conditions—All cell lines were obtained from the Duke University Cell Culture Facility Shared Resource in 2017, which regularly performs cell line authentication by short tandem repeat typing. Cells were cultured in Dulbeccos Modified

Eagles Medium (DMEM) supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin and incubated at 37°C with 5% CO_2.

RNA extraction, reverse transcription, and qPCR—Total RNA was isolated from cultured cells plated in 24-well format at a density of 50,000 cells/well using the Zymo Quick RNA MiniPrep kit. Reverse transcription reactions were comprised of 250–500 ng of total RNA, 200 ng of random hexamer primers, 1X IMPROMII reverse transcriptase buffer, 10 M dNTPs, 3.75 mM $MgCl_2$, 0.1 l RNasin, and 1 l of IMPROMII reverse transcriptase in a total volume of 20 l. Following RT, cDNAs were diluted 1:5 with nuclease-free H_2O, and quantitative polymerase chain reactions (qPCRs) were prepared using 2 l of diluted cDNA, 5 l of SYBR master mix (Kapa), and 60 nM of each primer in a 10:l reaction volume. All qPCRs were performed in a ViiA-7 Real-Time PCR System (Applied Biosystems). Primer sequences are listed in Supplementary Data. All experiments were performed in triplicate and repeated on separate days. Data were graphed in Microsoft Excel and analyzed in JMP Pro 13 using analysis of variance with Tukey's post-hoc correction. Any p-value< 0.05 was considered statistically significant.

Western blotting and immunofluorescence staining—To prepare cells for western blots, cells were plated at 300,000 cells/well in 6-well format. The next day, cells were lysed in ice-cold 1X radio-immunoprecipitation assay buffer supplemented with 1X Halt Protease and Phosphatase Inhibitor Cocktail (ThermoFisher). Cells were incubated for 15 minutes on a rocking platform at 4°C, and lysates were clarified by centrifugation at high speed in a benchtop centrifuge at 4°C. A total of 10 g from each lysate was boiled in 1X sodium dodecyl sulfate loading buffer, and proteins were separated in 4–15% MiniPROTEAN TGX Pre-cast Gels (BioRad) at 200V. Subsequent to transfer onto nitrocellulose, membranes were blocked in StartingBlock PBS blocking buffer for 1 hour at room temperature on a rocking platform, incubated overnight in the presence of primary antibodies diluted in StartingBlock PBS buffer, washed two times for five minutes each with PBS, incubated 1 hour at room temperature in a 1:20,000 dilution of Licor anti-mouse 800 and Licor anti-rabbit 680 diluted in StartingBlock PBS buffer, washed two times for five minutes each with PBS and imaged using the Licor Odyssey imaging system. For immunofluorescence staining, cells were plated at 50,000 cells/well in 24-well format and allowed to grow for 48 hours prior to fixing to allow re-establishment of E-cadherin at cell membranes. Cells were then fixed in 4% paraformaldehyde for 15 minutes, permeabilized in phosphate buffered saline (PBS)+0.2% Triton X-100 for 30 minutes at room temperature, blocked for 30 minutes in 5% bovine serum albumin (BSA) in PBS at room temperature, and incubated in the presence of a 1:1,000 dilution of anti-vimentin primary antibody diluted in 5% BSA in PBS overnight at 4°C. The next day, wells were washed with PBS, and incubated in the presence of a 1:2,000 dilution of anti-mouse AlexaFluor 488 secondary antibody and 1:2,000 dilution of Hoechst dye for one hour at room temperature in the dark. Next, cells were washed in PBS and incubated with 1 g of anti-E-cadherin antibody conjugated to AlexaFluor 647 anti-mouse IgG2a diluted in 5% BSA in PBS for one hour at room temperature in the dark. Wells were washed in PBS, and fluorescence images were captured using an Olympus IX 71 epifluorescence microscope with a DP70 digital camera and processed with CellSens software (Olympus). The following antibodies and dilutions were used: mouse anti-E-cadherin (BD Biosciences; cat. #610181), mouse anti-vimentin (ABD Serotec; cat. #MCA862), anti-Zeb1 (Santa Cruz; cat. #sc-25388), and rabbit anti-GAPDH (Santa Cruz; cat. #sc-25778).

ImageStream and flow cytometry analysis—Cell lines were analyzed by ImageStream and flow cytometry at the Duke Cancer Institute Flow Cytometry Shared Resource. MCF-7 and 143B cells were used as controls to create a compensation matrix for the ImageStream analysis. The following antibodies were used: mouse IgG2a isotype control antibody (Life Technologies; cat. #MG2A00), mouse IgG1 isotype control antibody (Life Technologies; cat. #MG100), Zenon Alexa Fluor 488 mouse IgG1 labeling kit (Thermo; cat. #Z25002), Zenon Alexa Fluor 647 mouse IgG1 labeling kit (Thermo; cat. #Z25108), mouse anti-E-cadherin (BD Biosciences; cat. #610181), and mouse anti-vimentin (ABD Serotec; cat. #MCA862).

Results

The model identifies both the drivers and fine-tuners of epithelial plasticity—The output of this data-driven approach results in a model which, when supplied with an appropriate training set and list of relevant predictor genes, generates predictions of the hybrid E/M phenotype for individual cell lines and patient samples by identifying a subset of predictors that can best fit the NCI-60 training set (Figure 1). NCI-60 cell lines have previously been categorized as epithelial, mesenchymal, or hybrid E/M based on the ratio of protein levels of CDH1/VIM [15]. Our model calculates how well each two-set combination of roughly 480 predictors (461 genes, 22 ratios of two genes; see Supplementary Data) can fit the training set.

The top 5% of candidate predictors that are best able to individually resolve the training set classification groups into E, hybrid E/M, and M represent the ability of individual genes to characterize EMT (Table 1A). Not surprisingly, this list contains canonical epithelial and mesenchymal markers such as CDH1 (E-cadherin) and VIM (Vimentin) respectively. Importantly, it also contains phenotypic stability factors (PSFs) – the factors that can stabilize a hybrid E/M phenotype by acting as molecular brakes, thereby preventing them from undergoing a full EMT, such as GRHL2, OVOL1, and OVOL2 (Table 1A) [12, 13, 23]. Overexpression of one or more of these PSFs can drive a MET, whereas their knockdown can induce a full EMT as observed in breast and prostate cancer cells [12, 24, 25]. Similar observations have been reported for another element in this list - Claudin 7 (CLDN7), a crucial component of tight junctions, thereby illustrating the ability of the statistical model in identifying the drivers as well as fine-tuners of epithelial plasticity [26].

Another top candidate listed is the vesicle protein Rab25, a member of Rab11 family that regulates E-cadherin turnover rate and whose levels are modulated by GRHL2 as well as Zeb1 - a key transcription factor that drives EMT [25, 27]. Furthermore, CDH3 (P-cadherin), a proposed marker of hybrid E/M phenotype [28], also appears in the list of top 5% EMT-relevant genes (Table 1A). An identical analysis ranked in an opposite manner on the entire NCI-60 transcriptome reveals gene products *least correlated* with EMT state, the results of which bear no resemblance to known EMT pathways (Table 1B).

Model feature selection is determined by the top pair of candidate predictors that can best resolve E, hybrid E/M, and M phenotypes, and results in the identification of CLDN7 (X_1) with VIM/CDH1 (X_2) (Table 1C). The best-fit model we ultimately utilized is completely described by $\beta = [-7.87, 0.0413, 1.36, -1.96]$ (See Equation 2). However, all top 10 combinations fit training data with near-equal ability (Table 1C). The frequent presence of PSFs such as OVOL1, OVOL2, and/or GRHL2 in this list of top 10 2-predictor combinations further reinforces our confidence in the ability of the model to resolve samples into three categories: E, hybrid E/M, and M. The top pair, CLDN7 and VIM/CDH1, performs well with respect to making leave-one-out predictions, which suggests that the risk of model over-fitting is minimal (Table 1D). On the other hand, this top pair performs significantly better than only VIM/CDH1, clearly illustrating the role of CLDN7 in resolving these three phenotypes.

Model sensitivity and specificity shows consistent performance with one exception - sensitivity for the hybrid E/M phenotype. This exception is a manifestation of lower resolution (more overlap) between E/M and M groups relative to that between E and E/M groups in the available training data (Figure 2B). We expect that the variability in E/M and M groups could be further resolved with additional samples (currently 11 E/M, 11 E, and 37 M samples in the NCI-60 cohort, as categorized based on the ratio of protein levels of CDH1 and VIM) [15].

The deviance, D (Equation 1), of 10^6 randomly constructed models from the EMT-relevant feature selection pool was found to be $D = 90.54 \pm 14.74$. The deviance of the best predictor combination, $D = 26.78$ falls well outside this range, indicating that significant improvements in describing the data can be made by applying our feature selection approach even when compared to the output generated by an average, EMT-relevant pair of predictors (Figure S1A). Lastly, the results of adding an additional predictor to the top 50 2-predictor combinations does not result in significant changes in leave-one-out sensitivity and specificity (Figure S1B). This observation does not rule out the possibility that a new 3-predictor combination may outperform the best 2-predictor combination. However, given our computational limitations and reservations for model over-fitting, we are satisfied with using the two most relevant predictors in combination to quantify EMT.

Normalization with respect EMT-independent gene signatures accounts for tissue-specific differences—The top two-predictor (CLDN7, VIM/CDH1) model can be visualized in three dimensions where the x- (resp. y-) axis represents $\log_2$CLDN7 (resp. $\log_2$VIM/$\log_2$CDH1) expression levels. For each data point, three related outputs provide an estimate of the probability that a sample has phenotype, E (Equation 4, $n = 1$), E/M (Equation 4, $n = 2$), and M (Equation 4, $n = 3$) (Figure 2A). Projections of each probability into the x–y plane reveal the relevant range for which each phenotype resides (Figure 2B). The representation of EMT status as the maximal predicted probability state can be appreciated by projecting Equation 4 (n=1,2,3). Overlaying NCI-60 data reveals that a majority of training samples fall within their expected range (Figure 2C). A prototypical demonstration of normalization is provided for cell lines composed of CD44+/CD24− and CD44−/CD24+ human mammary epithelial cells (GSE15192) (Figure 2D). Here, pre-normalized (purple) and post-normalized (pink) samples are plotted alongside NCI-60 training set samples (black). In this case, normalization provided significant shift in several mesenchymal samples originally classified as E/M, and several hybrid E/M samples originally classified as epithelial. Additional illustrations of normalization are given in Figure S1.

The model captures known phenotypes for multiple cancer types *in vitro*—Our algorithm was able to recapitulate the known phenotypes for multiple *in vitro* studies across various cancers. For instance, ectopic expression of EMT-inducing transcription factor SNAIL in an epithelial breast cancer cell line MCF-7 was predicted to drive a full EMT (GSE58252) [29] (Table 2A), and subpopulations of epithelial prostate cancer cells PC3 exhibiting enhanced transendothelial migration were predicted to be more mesenchymal (GSE14405). TEM4-18 cells, negative for E-cadherin and displaying nuclear staining for Zeb1 [30], were predicted to be mesenchymal, whereas TEM2-5, with relatively higher levels of cell-adhesion molecules as compared to TEM4-18 [30], were predicted to be hybrid E/M (Table 2A). Similarly, PC-3/Mc cells, a subpopulation of PC-3 cells that co-expressed CD24 and CD44 [31] (a signature of hybrid E/M [9]), were predicted to be hybrid E/M, and PC-3/S cells, being enriched in mesenchymal gene expression [31], were predicted as mesenchymal (Table 2A) (GSE24868). Higher tumor-initiation potential and an active self-renewal program in PC-3/Mc further reinforce the hypothesis that cells in a hybrid E/M state, instead of those frozen in a mesenchymal state, are most likely to be more stem-like [1, 32, 33]. Furthermore, multiple Ewing sarcoma (GSE70826) (Table 2A) and osteosarcoma (GSE70414, GSE55957) (Table S2) datasets predicted to be mesenchymal, and the epithelial and mesenchymal sub-populations of HMLE cells (GSE28681) (Table 2A) had significantly different EMT scores. The algorithm also predicts that short-term treatment of cells with EMT or MET inducers is usually not suffcient to drive a transition (GSE7868, GSE17708, GSE59771 and GSE53603) (Table S2).

We also calculated EMT scores for *in vivo* mouse model of pancreatic cancer, KPC, both in control cases and when specific EMT-inducing transcription factors were genetically knocked out (KO). Tumors from both KPC control mice, and the KO-Twist or KO-Snail KPC mice (GSE66981; [34]) were predicted as hybrid E/M, but cell lines established from those with KO-Zeb1 KPC mice (GSE87472; [35]) were categorized as almost purely epithelial (Table 2C). Further, our algorithm accurately recapitulated the experimental observation that an EMT was not induced in epithelial cells from Zeb1-KO mice upon TGF treatment [35]. Together, these results reinforce a key role of Zeb1 in mediating EMT [27, 36].

Cell lines predicted as hybrid E/M tend to co-express epithelial and mesenchymal markers—Next, we ran our model for transcriptomes of multiple cell lines, including SW480 and SW620 (both colorectal cancer), DU145 (prostate cancer), and A549, H1975, H460, and H1650 (all non-small-cell lung cancer) (GSE36821, GSE15392, GSE10843). SW480, H460 and H1650 were predicted to be epithelial, whereas H1975, DU145, SW620, and A549 were predicted to be hybrid E/M (Table 2B). Consistent with their predicted phenotypes, H1975 cells have been shown to stably co-express E-cadherin and vimentin at a single-cell level [12], while H460 and H1650 cells have been previously categorized as epithelial-like based on proteomic measurements [37].

To better understand the predicted hybrid E/M cell lines, we first quantified the levels of known EMT master regulators of qRT-PCR. We also included epithelial MCF-7 cells and mesenchymal 143B osteosarcoma cells for comparison. Relative to the strongly epithelial MCF-7 cells, the hybrid E/M cell lines consistently expressed elevated levels of Zeb1 and Snail and were more similar in expression of Zeb1 and Snail to the mesenchymal 143B cells (Figure S2A). Interestingly, the SW480 cells, which were predicted to be epithelial, also resembled hybrid cells in their expression of Zeb1 and Snail (Figure S2A). Similarly, the hybrid E/M lines had undetectable levels of the transcription factor GRHL2, while SW480, predicted to be epithelial, expressed low levels of GRHL2 compared to MCF-7 (Figure S2B–C). E-cadherin levels were also substantially lower in the hybrid E/M lines and SW480 when compared to MCF-7 at both the mRNA (Figure S2B–C) and protein (Figure 3A) levels, with variable levels of vimentin protein (Figure 3A). Together, these results confirm that the cell lines predicted as hybrid co-express epithelial and mesenchymal biomarkers at intermediate levels compared to strongly epithelial or strongly mesenchymal cell lines.

All of the datasets above contain gene expression on an ensemble level instead of single-cell gene expression data. Therefore, a hybrid E/M signature may be predicted either because they truly contain hybrid E/M cell co-expressing epithelial and mesenchymal markers (as shown for H1975), or because they are comprised of sub-populations of epithelial and mesenchymal phenotypes. In order to further investigate these cell lines at a quantitative and single-cell level, we performed two-color flow cytometry for DU145, A549, SW620, and SW480 cells, which were predicted to be epithelial, but co-expressed CDH1 and VIM. We also included MCF-7 cells as a control for cells predicted to be epithelial. While the MCF-7 cells were 86–98% CDH1^{high}/VIM^{low}, all other lines had three distinct sub-populations of epithelial-like (CDH1^{high}/VIM^{low}), hybrid E/M (CDH1^{high}/VIM^{high}), and mesenchymal-like (CDH1^{low}/VIM^{high}) cells (Figure 3B–D). An experimental quantification of each sample's EMT score, μ_{exp}, was estimated by weighting the given categorical scores (E=0, E/M=1, M=2) by the observed proportion of E-cadherin and vimentin expressed: $\mu_{exp} = 0 \cdot [\%CDH1^{+}/VIM^{-} \text{ cells}] + 1 \cdot [\%CDH1^{+}/VIM^{+} \text{cells}] + 2 \cdot [\%CDH1^{-}/VIM^{+} \text{cells}]$. This was compared to theoretical predictions of EMT scores using Equation 5 (Figure 3C, S3). We then used two-color staining for CDH1 and VIM on the ImageStream, which combines flow cytometry with single-cell imaging. Using this instrument, we were able to clearly identify three distinct sub-populations of cells in all four cell lines DU145, A549, SW480, SW620, including CDH1^{high}/VIM^{low}, CDH1^{high}/VIM^{high}, and CDH1^{low}/VIM^{high} (Figure 3E). These results not only highlight the extent of phenotypic heterogeneity in the cell lines studied above, but also offer a potential reason for why SW480 cells were predicted to be epithelial; in cell lines that are admixtures of different phenotypes, a context-dependent enrichment of one phenotype is unsurprising.

Next, we performed immunofluorescence staining for CDH1 and VIM in A549, DU145, SW620, and SW480 cells. Consistent with the predictions of the model, DU145 cells expressed clear co-staining of membrane-localized CDH1 and VIM in numerous cells (Figure 4A). On the other hand, A549 cells were predominantly CDH1-low and VIM-positive, with distinct clusters of CDH1$^+$/VIM$^-$ cells (Figure 4B). Like the DU145 cells, SW480 cells also contained a population of cells with co-expression of CDH1 and VIM (Figure 4C); however, a subset of SW480 cells possessed CDH1$^+$/VIM$^-$ cell clusters (Figure 4C). The SW620s displayed a patchier distribution of membrane CDH1 positivity and strong VIM expression, with a small subpopulation of cells that co-express CDH1 and VIM (Figure 4D).

Together, our quantitative analysis at the single-cell level revealed that the cells predicted to be hybrid can contain subsets of epithelial-like, hybrid E/M, and mesenchymal-like cells.

Association between EMT status and survival is tissue and subtype-specific— Kaplan-Meier survival analysis reveals statistically significant ($p < 0.05$ at significance level $\alpha = 0.05$) differences between epithelial and non-epithelial signatures for multiple breast cancer datasets. In a majority of cases (Figure 5A–E), patients exhibiting a more epithelial phenotype had poorer survival as compared to those displaying a partial- or full-EMT signature (GSE17705, GSE1456, GSE45255, GSE5327, GSE6532). Although statistically significant, some of these cases - especially 5A (Hazard Ratio=0.760), 5B (HR=0.614), and 5E (HR=0.625) - do not show dramatic separation in clinical parameters. However, in a cohort with a larger percentage of basal-like breast cancer, patients with a hybrid E/M phenotype demonstrate significant reductions in disease-free survival when compared to patients with an epithelial signature (Figure 5F). This result is consistent with independent attempts at describing subtype-specific differences in correlations between EMT status and survival in which the authors described a scenario wherein the epithelial phenotype was prognostic for worse survival in some cancer types and better survival in others [38]. Therefore, a higher EMT score need not always correlate with poor survival, at least in multiple subtypes of breast cancer. Such a correlation may also be confounded by heterogeneous factors such as molecular subtype (ER+ samples in GSE17705 and ER- samples in GSE1456 and GSE5327) and varied prior therapy regimens (tamoxifen treatment for patients in GSE17705, GSE1456 and GSE6532, and neoadjuvant taxane-anthracycline chemotherapy for patients in GSE25066) that may alter cell EMT status [39].

In lung cancer (GSE31210), patients categorized as hybrid E/M phenotype had significantly lower relapse-free (HR=1.942) and overall survival (HR=1.391) as compared to those binned for epithelial phenotype, with a relatively wider separation in clinical parameters (Figure 5G–H). Ovarian cancer patient datasets for which there were statistically significant differences in overall survival revealed mixed results. In one case (GSE63885), hybrid E/M samples demonstrated improved overall survival, while in another (GSE26712), hybrid E/M signatures were significantly more aggressive (Figure 5I–J). These differences in ovarian cancer may possibly be the result of different therapy regimens. No treatment information could be found for patients in GSE26712, while GSE63885 represents a collection of patients post-first-line chemotherapy.

To assess the significance of the role of CLDN7 in this EMT-survival association, we plotted Kaplan-Meier curves for the same datasets mentioned above for two cases: a) using median levels of CDH1/VIM to resolve patients into two groups, CDH1/VIM^{high} and CDH1/VIM^{low} (Figure S4), and, b) using CDH1 and VIM as the two predictors in our statistical model (Figure S5). In either case, the significant correlation observed by using CDH1/VIM and CLDN7 as the predictors was lost in 8 or more of 10 cases evaluated. This difference reinforces our earlier analysis that {CDH1/VIM, CLDN7} predictor set can resolve the multi-dimensional gene expression landscape onto an EMT axis much more accurately than {CDH1/VIM} or {CDH1, VIM}.

EMT spectrum for TCGA datasets—Next, we ran our model on multiple TCGA datasets [40–46] and observed a wide spectrum of EMT states for multiple cancer types. Breast (BCA) and lung (LCA) cancer samples displayed an epithelial phenotype predominantly, and most sarcoma samples were categorized as mesenchymal. Notably, pancreatic adenocarcinoma (PDAC) and renal clear cell carcinoma (RCC) samples were enriched for a hybrid E/M phenotype (Figure S6A), reminiscent of co-expression of epithelial and mesenchymal markers *in vivo* in PDAC and *in vitro* in RCC cell lines [1]. Lastly, we investigated the correlation of EMT scores with metastatic potential in these TCGA datasets. Breast cancer samples that exhibited metastasis were either categorized as epithelial or hybrid E/M (Figure S6B), reinforcing the concept that a complete EMT need not occur for metastatic dissemination [47].

Discussion

We have applied iterated regression trained on the NCI-60 dataset in order to create an inferential statistical model of the EMT spectrum. Our model relates gene expression patterns for a small collection of EMT-relevant transcripts to the proclivity of a sample for one of three categories - E, hybrid E/M, and M. Advantages of this approach include an explicit quantitative description of the intermediate, hybrid E/M state, as well as a simple and relatively affordable diagnostic tool that may be used in assessing the EMT status of human tissue samples. Characterizing the hybrid E/M phenotype(s) is a crucial step toward addressing recent controversies in the literature. In particular, several recent studies have questioned the indispensable role of at least a complete EMT and MET in metastatic progression [34,47,48]. This model is therefore valuable in investigating systematically the role of hybrid E/M phenotype(s) in the metastatic cascade and can help us appreciate a more nuanced view of cellular plasticity.

Working within our computational limits, we find that CLDN7 and VIM/CDH1 constitute the best pair of predictors to fit the NCI-60 training set, and maintain predictive value in in categorizing the NCI-60 cell lines via leave-one-out analysis. CDH1 and VIM are canonical markers of epithelial and mesenchymal states respectively, whereas CLDN7 (Claudin 7) may be crucial in maintaining the hybrid E/M phenotype. This proposed role of CLDN7 is based on observations made for other ‘phenotypic stability factors’ for a hybrid E/M phenotype such as GRHL2 and OVOL2 [12, 13, 24, 25]. Therefore, our model identifies representative features from E, hybrid E/M, and M phenotypes, and is therefore able to recapitulate the observed role of drivers as well as fine-tuners of cellular plasticity.

The identification of CDH1/VIM as one of the two elements constituting the top predictor set may appear as ‘circular reasoning,’ but as highlighted both via agreement to training data and patient survival data, having CLDN7 as another member in the top predictor set enables a much better resolution of the expression signature landscape on EMT axis. We validated our approach by comparing model predictions against samples whose phenotypes are known *a priori*, both across tissue types and across different experimental conditions such as isogenic subpopulations and treatment with EMT-inducing signals for different durations. We also predicted a hybrid E/M status of multiple cell lines and later validated that they may contain either subpopulations of epithelial and mesenchymal cells (A549) or cells co-expressing epithelial and mesenchymal markers (DU145). When applying our model to TCGA datasets, we similarly observed a wide distribution of phenotypes in multiple cancer types. Particularly, renal cell carcinoma and PDAC samples were predominantly predicted to be hybrid E/M, but these observations are inconclusive on whether these samples contain hybrid E/M cells. Future studies focusing on single-cell gene expression analysis will be fundamental in order to dissect cellular heterogeneity and investigate underlying reasons for high aggressiveness of a hybrid E/M phenotype, due to cooperating epithelial or mesenchymal subpopulations and/or enhanced drug resistance of ‘double positive’ cells co-expressing epithelial and mesenchymal markers [1].

While multiple previous studies have associated EMT with poor survival [16, 17, 49], our results are consistent with prior observations [38] and suggest that such correlation can be highly tissue- and subtype-specific, even after normalizing the data to minimize the effect of external factors such as platform-specific variations. Of particular interest is the observation that breast cancer patients with lower EMT scores had better overall and progression-free survival, except when investigating a dataset enriched in basal-like breast cancer. These apparent contradictions may result from a combination of factors such as different therapeutic treatments driving phenotypic transitions [39, 50], and methods of generating EMT-specific signature used to classify patients for survival analysis [9]. Prior work has relied on inferring characteristics of the intermediate E/M phenotype by interpolating between known behavior for E and M states [9,38]. In contrast to other large gene expression analyses that correlate EMT with survival, our model is trained directly on known hybrid E/M samples in addition to E and M. Moreover, it provides a continuous, explicit quantification of all three regimes on the EMT spectrum. This allows for a quantification of the aggregate signature at the population level, as well as a probabilistic interpretation of EMT category on the single-cell level.

In conclusion, we develop an algorithm to quantify the extent of EMT, independent of cancer type, that can be used to systematically investigate the role of intermediate or hybrid epithelial/mesenchymal phenotype(s) in multiple hallmarks of tumor progression, such as invasion and metastasis, angiogenesis, resistance to apoptosis, and resistance to multiple therapies. This metric, based on gene expression, has the potential to be integrated with proteomics and metabolomics data among others, and offers an EMT score that can objectively characterize the EMT status of both *in vitro* samples as well as *in vivo* xenografted tissue and patient tissue samples.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial Support: J.T. George was supported by the National Cancer Institute of the National Institutes of Health (F30CA213878). M.K. Jolly was supported by a training fellowship from the Gulf Coast Consortia on the Computational Cancer Biology Training Program (CPRIT Grant No. RP170593). J.A. Somarelli acknowledges support from the Duke Cancer Institute, the Duke Genitourinary Oncology Laboratory, the Department of Orthopaedics, and the Triangle Center for Evolutionary Medicine (TriCEM). H. Levine was supported by Cancer Prevention and Research Institute of Texas Scholars program (R1111), National Science Foundation (NSF) Center for Theoretical Biological Physics (NSF PHY-1427654, NSF PHY-1605817, and NSF DMS-1361411).

We would like to thank Kenneth J. Pienta and Princy Parsana for fruitful discussion on EMT.

References

1. Jolly MK, Boareto M, Huang B, Jia D, Lu M, Ben-Jacob E, et al. Implications of the hybrid epithelial/mesenchymal phenotype in metastasis. Front Oncol. 2015; 5:155. [PubMed: 26258068]
2. Tripathi SC, Peters HL, Taguchi A, Katayama H, Wang H, Momin A, et al. 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 U S A. 2016; 113:155564.
3. Huang RY-J, Wong MK, Tan TZ, Kuay KT, Ng aHC, Chung VY, et al. An EMT spectrum defines an anoikis-resistant and spheroidogenic intermediate mesenchymal state that is sensitive to e-cadherin restoration by a src-kinase inhibitor, saracatinib (AZD0530). Cell Death Dis. 2013; 4:e915. [PubMed: 24201814]
4. Nieto MA, Huang RY, Jackson RA, Thiery JP. EMT: 2016. Cell. 2016; 166:2145.
5. Andriani F, Bertolini G, Facchinetti F, Baldoli E, Moro M, Casalini P, et al. Conversion to stem-cell state in response to microenvironmental cues is regulated by balance between epithelial and mesenchymal features in lung cancer cells. Mol Oncol. 2016; 10:25371.
6. Yu M, Bardia A, Wittner BS, Stott SL, Smas ME, Ting DT, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. Science. 2013; 339:5804.
7. Aceto N, Toner M, Maheswaran S, Haber DA. En Route to Metastasis: Circulating Tumor Cell Clusters and Epithelial-to-Mesenchymal Transition. Trends in Cancer. 2015; 1:4452.
8. Cheung KJ, Ewald AJ. A collective route to metastasis: Seeding by tumor cell clusters. Science. 2016; 352:1679.
9. Grosse-Wilde A, Fouquier d Herouei A, McIntosh E, Ertaylan G, Skupin A, Kuestner RE, et al. Stemness of the hybrid epithelial/mesenchymal state in breast cancer and its association with poor survival. PLoS One. 2015; 10:e0126522. [PubMed: 26020648]
10. Jolly MK, Boareto M, Debeb BG, et al. Inflammatory breast cancer: a model for investigating cluster-based dissemination. NPJ Breast Cancer. 2017; 3:21. [PubMed: 28649661]
11. Savagner P. The epithelial-mesenchymal transition (EMT) phenomenon. Ann Oncol. 2010; 21(Suppl 7):vii8992.
12. Jolly MK, Tripathi SC, Jia D, Mooney SM, Celiktas M, Hanash SM, et al. Stability of the hybrid epithelial/mesenchymal phentoype. Oncotarget. 2016; 7:2706784.
13. Jia D, Jolly MK, Boareto M, Parsana P, Mooney SM, Pienta KJ, et al. OVOL guides the epithelial-hybrid-mesenchymal transition. Oncotarget. 2015; 6:1543648.
14. Hong T, Watanabe K, Ta CH, Villarreal-Ponce A, Nie Q, Dai X. An Ovol2-Zeb1 mutual inhibitory circuit governs bidirectional and multi-step transition between epithelial and mesenchymal states. PLoS Computational Biology. 2015; 11(11):e1004569. [PubMed: 26554584]
15. Park S-MM, Gaur AB, Lengyel E, Peter ME. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. Genes Dev. 2008; 22:894907.
16. Taube JH, Herschkowitz JI, Komurov K, et al. 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. 2010; 107:15449–54. [PubMed: 20713713]
17. Loboda A, Nebozhyn MV, Watters JW, et al. EMT is the dominant program in human colon cancer. BMC Med Genomics. 2011; 4:9. [PubMed: 21251323]

18. van't Veer LJ, Dai H, Vijver Hvd, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415:530–36. [PubMed: 11823860]

19. Ben-Porath I, Thompson MW, Carey VJ, et al. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nature Genetics. 2008; 40:499–507. [PubMed: 18443585]

20. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol. 2011; 7:e1002240. [PubMed: 22028643]

21. Bellman, RE., Dreyfus, SE. Applied Dynamic Programming. Princeton, N.J: Princeton University Press; 1962.

22. Hughes G. On the mean accuracy of statistical pattern recognizers. IEEE Trans Inf Theory. 1968; 14:55–63.

23. Watanabe K, Villarreal-Ponce A, Sun P, et al. Mammary morphogenesis and regeneration require the inhibition of EMT at terminal end buds by Ovol2 transcriptional repressor. Dev Cell. 2014; 29:59–74. [PubMed: 24735879]

24. Roca H, Hernandez J, Weidner S, et al. Transcription factors OVOL1 and OVOL2 induce the mesenchymal to epithelial transition in human cancer. PloS one. 2013; 8:e76773. [PubMed: 24124593]

25. Xiang X, Deng Z, Zhuang X, et al. Grhl2 determines the epithelial phenotype of breast cancers and promotes tumor progression. PloS one. 2012; 7:e50781. [PubMed: 23284647]

26. Bhat AA, Pope JL, Smith JJ, et al. Claudin-7 expression induces mesenchymal to epithelial transformation (MET) to inhibit colon tumorigenesis. Oncogene. 2015; 34:4570–80. [PubMed: 25500541]

27. Lu M, Jolly MK, Levine H, Onuchic JN, Ben-Jacob E. MicroRNA-based regulation of epithelial - hybrid - mesenchymal fate determination. Proc Natl Acad Sci U S A. 2013; 110:18144–9. [PubMed: 24154725]

28. Ribeiro AS, Paredes J. P-cadherin linking breast cancer stem cells and invasion: a promising marker to identify an intermediate/metastable EMT state. Front Oncol. 2015:4. [PubMed: 25674540]

29. McGrail DJ, Mezencev R, Kieu QMN, McDonald JF, Dawson MR. SNAIL-induced epithelial-to-mesenchymal transition produces concerted biophysical changes from altered cytoskeletal gene expression. FASEB J. 2015; 29:12809.

30. Drake JM, Strohbehn G, Bair TB, Moreland JG, Henry MD. ZEB1 enhances transendothelial migration and represses the epithelial phenotype of prostate cancer cells. Mol Biol Cell. 2009; 20:220717.

31. Celiá-Terrassa T, Meca-Cortés Ó, Mateo F, De Paz AM, Rubio N, Arnal-Estapé A, et al. Epithelial-mesenchymal transition can suppress major attributes of human epithelial tumor-initiating cells. J Clin Invest. 2012; 122:184968.

32. Ombrato L, Malanchi I. The EMT Universe: Space between Cancer Cell Dissemination and Metastasis Initiation. Crit Rev Oncog. 2014; 19:34961.

33. Shibue T, Weinberg RA. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. Nat Rev Clin Oncol. 2017

34. Zheng X, Carstens JL, Kim J, et al. Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. Nature. 2015; 527:525–30. [PubMed: 26560028]

35. Krebs AM, Mitschke J, Losada ML, et al. The EMT-activator Zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic cancer. Nature Cell Biology. 2017; 19:518. [PubMed: 28414315]

36. Somarelli JA, Shetler S, Jolly MK, et al. Mesenchymal-Epithelial Transition in Sarcomas Is Controlled by the Combinatorial Expression of MicroRNA 200s and GRHL2. Mol Cell Biol. 2016; 36:2503–13. [PubMed: 27402864]

37. Schliekelman MJ, Taguchi A, Zhu J, Dai X, Rodriguez J, Celiktas M, et al. Molecular portraits of epithelial, mesenchymal and hybrid states in lung adenocarcinoma and their relevance to survival. Cancer Res. 2015; 75:1789800.

38. Tan TZ, Miow QH, Miki Y, et al. Epithelial mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. EMBO Mol Med. 2014; 6:1279–93. [PubMed: 25214461]

39. Goldman A, Majumder B, Dhawan A, et al. Temporally sequenced anticancer drugs overcome adaptive resistance by targeting a vulnerable chemotherapy-induced phenotypic transition. NatCommun. 2015; 6:6139.

40. Koboldt DC, Fulton RS, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490:61–70. [PubMed: 23000897]

41. Ciriello G, Gatza ML, Beck AH, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015; 163:506–19. [PubMed: 26451490]

42. Muzny DM, Bainbridge MN, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–37. [PubMed: 22810696]

43. Grasso CS, Wu Y, Robinson DR, et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature. 2012; 487:239–43. [PubMed: 22722839]

44. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013; 499:43–49. [PubMed: 23792563]

45. Hammerman PS, Lawrence MS, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–25. [PubMed: 22960745]

46. Silver SJ, Lash A, Lau C, et al. Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. Nature Genet. 2010; 42:715–21. [PubMed: 20601955]

47. Jolly MK, Ware KE, Gilja S, et al. EMT and MET: necessary or permissive for metastasis? Mol Oncol. 2017

48. Fischer KR, Durrans A, Lee S, et al. Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. Nature. 2015; 527:472–6. [PubMed: 26560033]

49. Byers LA, Diao L, Wang J, et al. An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. Clin Cancer Res. 2013; 19:279–90. [PubMed: 23091115]

50. Boareto M, Jolly MK, Goldman A, et al. Notch-Jagged signaling can give rise to clusters of cells exhibiting a hybrid epithelial/mesenchymal phenotype. J R Soc Interface. 2016; 13:20151106. [PubMed: 27170649]

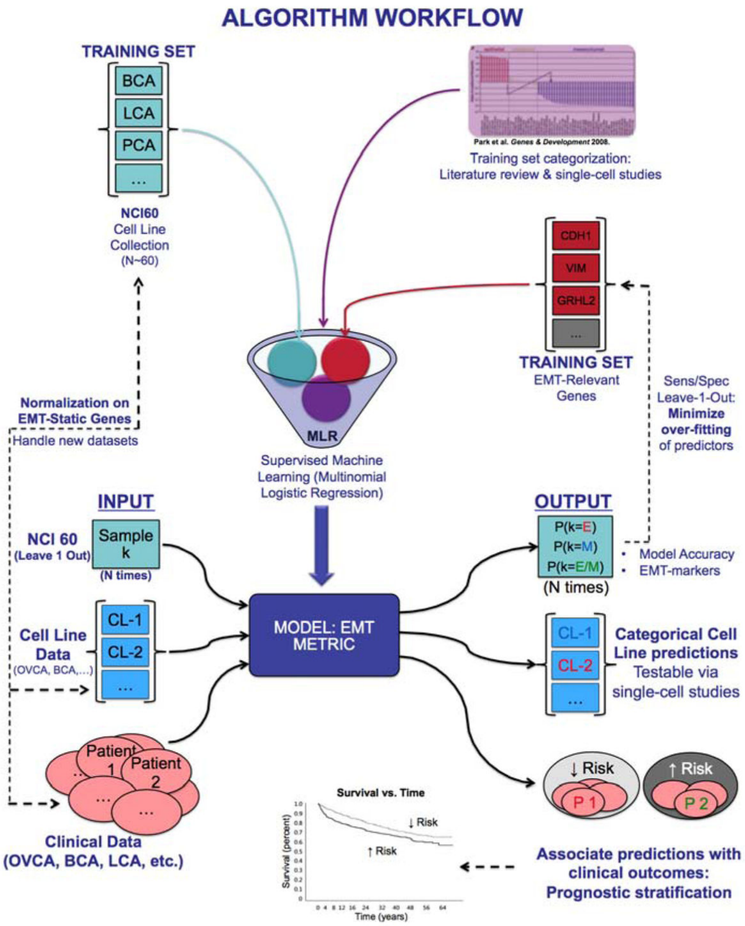


Figure 1. Schematic illustration of model construction and prediction

Input elements relevant to model construction include a NCI-60 training data (teal boxes), *a priori* training set categorization (purple), and a list of candidate predictors (maroon). Model construction is used in the leave-one out characterization of predictors and construction of normalizers, to predict categories of EMT-relevant cell lines, and to categorize patient primary tumor samples for risk stratification (bottom half).

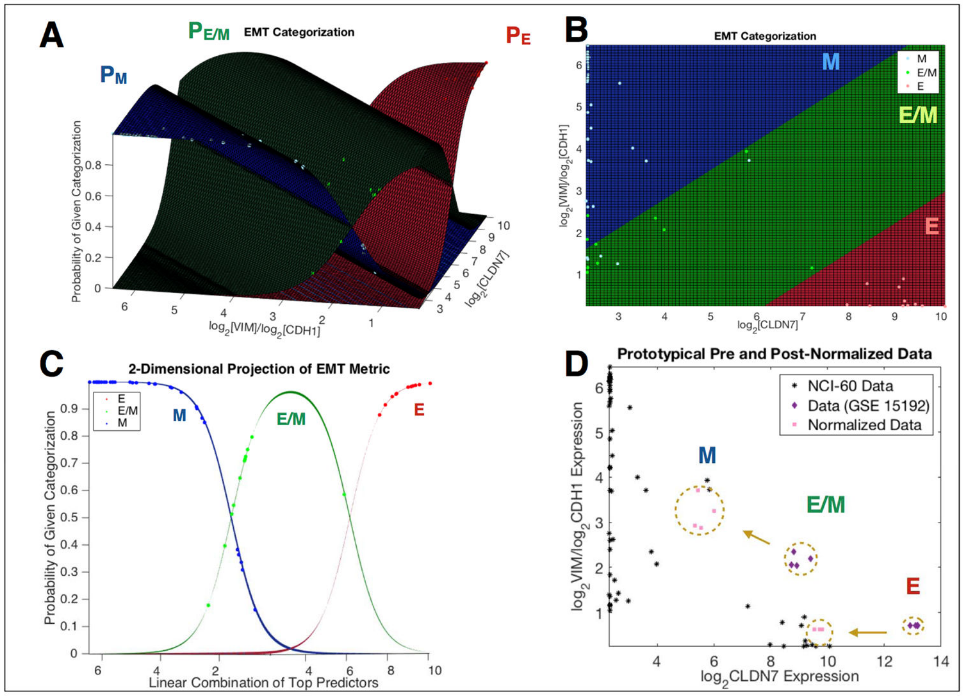
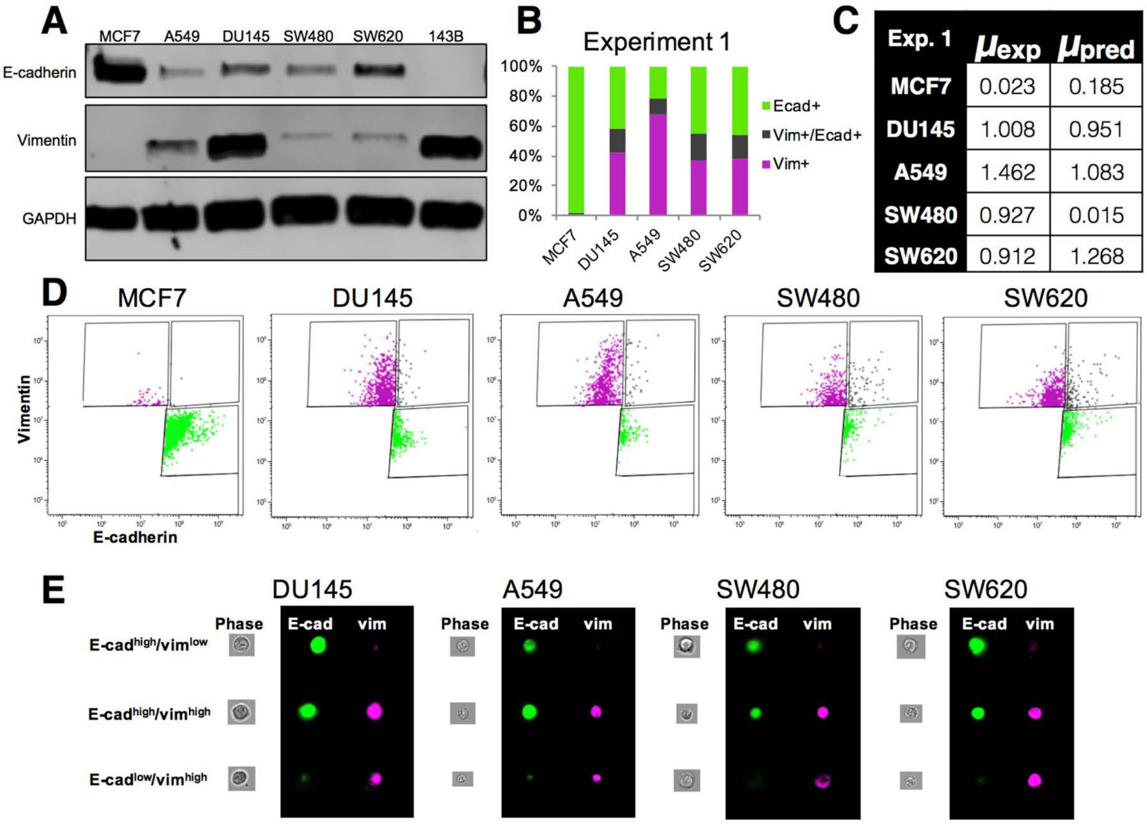


Figure 2. Model representation

(A) 3-dimensional view of model constructed using top predictors; (B) Model viewed from overhead representing various regions of predictor space that define E, E/M, and M categories; (C) 2-dimensional model projection of model for use in defining the EMT metric, μ, described by equation 5; (D) Prototypical example of pre- vs post-normalization comparisons in an immortalized human mammary epithelial cell line (GSE15192).



Exp. 1	μ_{exp}	μ_{pred}
MCF7	0.023	0.185
DU145	1.008	0.951
A549	1.462	1.083
SW480	0.927	0.015
SW620	0.912	1.268

Figure 3. Western blot, ImageStream and flow cytometry analysis of epithelial-like, hybrid, and mesenchymal-like cells

(A) Western blot analysis of CDH1 and VIM reveals cell lines predicted to be hybrid E/M display co-expression of CDH1 and VIM. MCF-7 and 143B are included as known epithelial and mesenchymal lines, respectively; (B) Quantification of relative proportions of epithelial-like, hybrid, and mesenchymal-like cells in DU145, A549, SW480, and SW620 cells compared to epithelial MCF-7 cells, for the data presented in Figure 3D; (C) Comparison of experimentally-observed EMT score for DU145, A549, SW480, and SW620 cells (μ_{exp}) and theoretical prediction of EMT score via Equation 5 (μ_{pred}); (D) Flow cytometry analysis of CDH1^{high}/VIM^{low} (green), CDH1^{high}/VIM^{high} (gray), and CDH1^{low}/VIM^{high} (magenta) sub-populations; (E) ImageStream analysis using two-color staining of E-cadherin and vimentin reveals the presence of distinct subpopulations of epithelial-like, hybrid, and mesenchymal-like cells in DU145, A549, SW480, and SW620 cells.

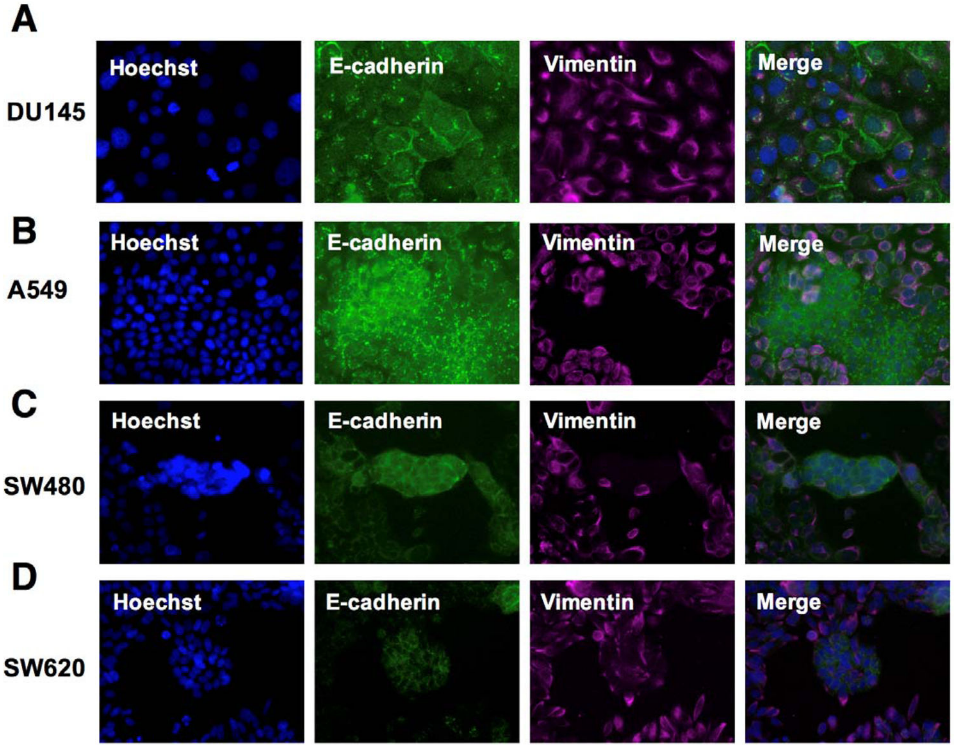


Figure 4. Validation of the hybrid E/M state reveals distinct subpopulations of epithelial-like, hybrid, and mesenchymal-like cells

(A) DU145 cell line contains cells that co-express membrane CDH1 and VIM; (B) A549 cells predicted to be hybrid E/M contain sub-populations of CDH1^{high}/VIM^{low} and CDH1^{low}/VIM^{high} cells, along with cells that co-express both CDH1 and VIM; (C) SW480 cells, predicted to be epithelial, have all three subpopulations of cell types; (D) SW620 cells are comprised predominantly of CDH1low/VIM^{high} cells, with nests of cells that display upregulated CDH1 and reduced levels of VIM.

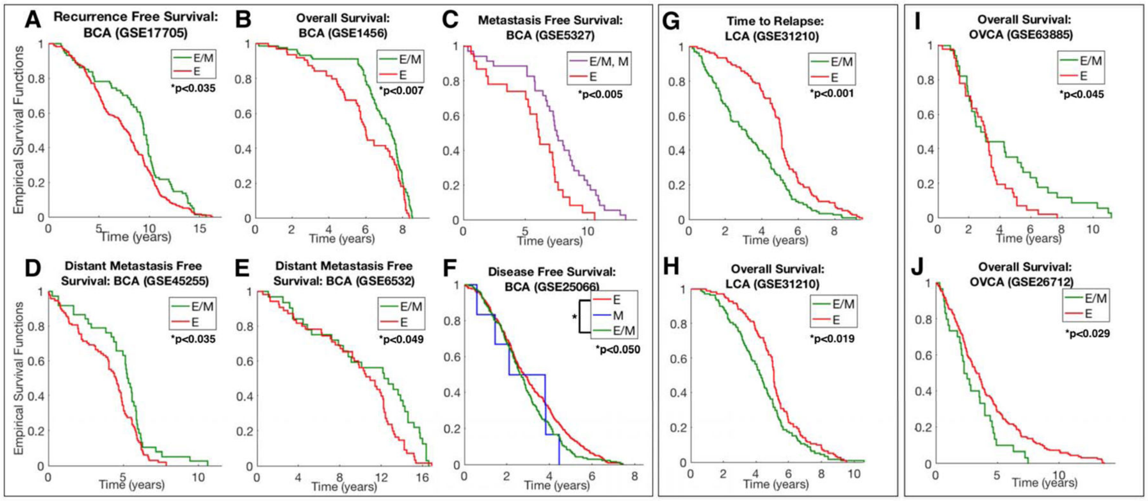


Figure 5. Correlation between EMT status and clinical survival metrics

Kaplan-Meier survival analysis is performed in order to compare statistically assess differences in survival and tumor aggressiveness between tumors predicted to be E, E/M, and M. This was performed for a variety of breast cancer (A–F), lung (G), and ovarian (H) primary tumor samples with Hazard Ratios and 95% confidence intervals: (A) HR=0.760 95%CI=(0.593, 0.974); (B) HR=0.614 95%CI=(0.593, 0.974); (C) HR=0.408 with 95%CI=(0.219, 0.761); (D) HR=0.667 with 95%CI=(0.466, 0.955); (E) HR=0.625 with 95%CI=(0.402, 0.971); (F) HR=0.818 with 95%CI=(0.673, 0.995); (G) HR=1.942 with 95%CI=(1.472, 2.561); (H) HR=1.391 with 95%CI=(1.066, 1.815); (I) HR=0.590 with 95%CI=(0.363,0.959); (J) HR=1.736 with 95%CI=(1.088, 2.771).

Table 1

Iterative regression output

(A) Candidate genes are ranked individually by their deviance, and the top 5% are illustrated to provide a list of the most resolvable EMT genes. Predictors involving EMT stability factors are identified; (B) Gene products that show the weakest correlation to training set categories are identified as normalizers, used to for cross-data comparison; (C) The top 10 optimal predictor combinations are ranked according to their deviance; (D) Prognostic outputs of leave-one-out analysis on the top predictor set, {CDH1/VIM, CLDN7} are provided.

A. Top 5% EMT-relevant genes	
Predictor	**Deviance**[†]
🟫	37.61
🟫	45.96
🟫	46.74
🟫	49.74
🟫	50.60
🟫	51.47
🟫	51.50
🟫	51.85
🟫	52.12
🟫	52.48
🟫	53.47
🟫	53.91
🟫	54.27
🟫	55.75
🟫	57.12
🟫	57.50
🟫	57.64
🟫	58.44
🟫	59.07
🟫	59.34
🟫	59.39
🟫	59.50
🟫	59.79

B. EMT-Normalizer
Normalizer
🟪
🟪
🟪
🟪

B. EMT-Normalizer
Normalizer
🟥
🟥
🟥
🟥
🟥
🟥
🟥
🟥
🟥
🟥
🟥
🟥
🟥
🟥
🟥

C. Top 10 2-Predictor Combinations			
Rank	**Predictor 1**	**Predictor 2**	**Deviance**[†]
1)	🟩	🟩	26.78
2)	🟩	🟩	27.73
3)	🟩	🟩	28.27
4)	🟩	🟩	28.31
5)	🟩	🟩	28.31
6)	🟩	🟩	28.48
7)	🟩	🟩	28.56
8)	🟩	🟩	28.63
9)	🟩	🟩	28.86
10)	🟩	🟩	29.08

D. Leave-One-Out Analysis: CLDN7, VIM/CDH1		
Categor	**Sensitivity**	**Specificity**
E	100%	98%
E/M	55%	90%
M	86%	82%
Diagnostic Accuracy: 83%		

[*] Single predictor sets containing EMT-stability factors OVOL1 or GRHL2

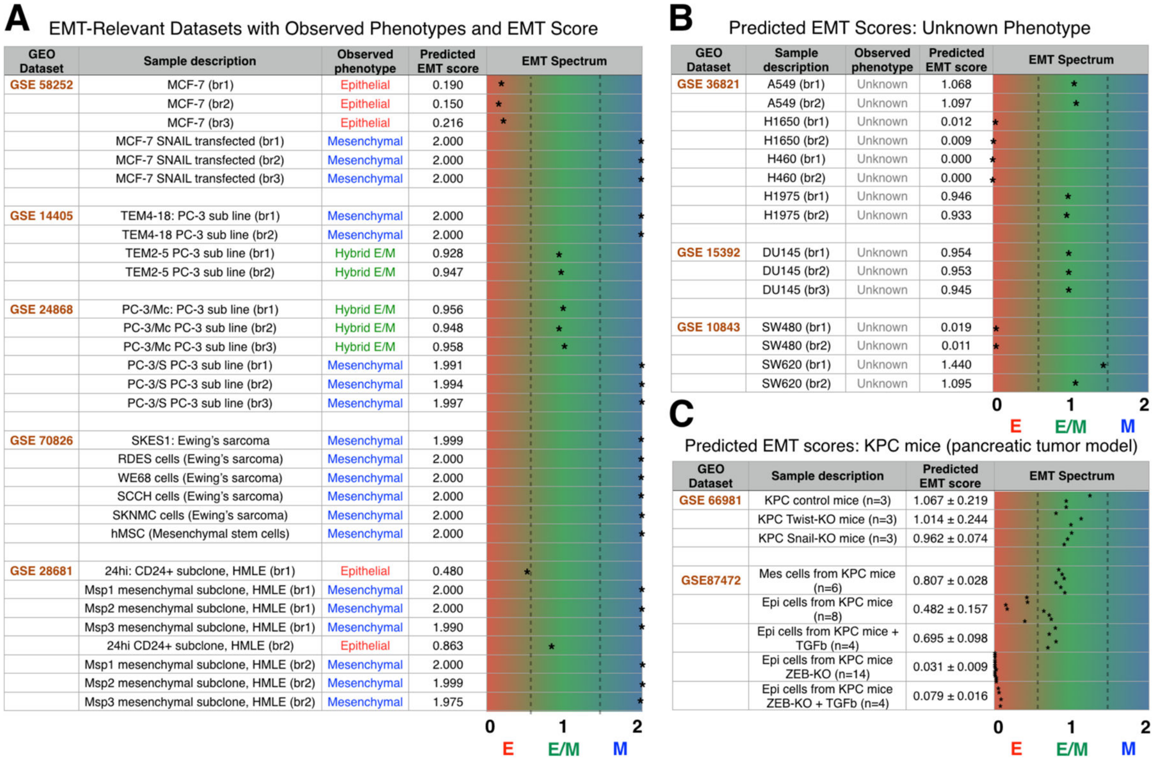
[†] Deviance, D, as defined in Equation 1.

[††] Top predictors (X_1,X_2) used in model construction

Table 2

Model predictions on relevant *in vitro* experimental datasets.

(A) Model predictions on datasets across multiple cancer types: GSE58252 - MCF-7 cells treated with SNAIL, GSE14405 - PC-3 sublines generated through transendothelial migration, GSE24868 - sublines of PC-3 with different EMT status and tumor-initiation potential, GSE70826 - sarcoma cell lines, and GSE28681 - epithelial and mesenchymal subpopulations of HMLE cells. Observed phenotype denotes the *a priori* known EMT status (red for E, green for hybrid E/M and blue for M), and the EMT spectrum plots a sample's EMT score, μ, as defined in Equation 5 (μ<0.5 corresponds to E, 0.05 < μ< 1.5 corresponds to E/M, and μ> 1.5 corresponds to M); (B) Same as A but applied to datasets with *a priori* unknown EMT status: GSE36821 - NSCLC lung cancer datasets, GSE15392 - DU145 dataset, and GSE10843 - dataset for SW480 and SW620 populations; (C) Same as B but for genetically engineered mouse models of pancreatic tumors (KPC mice).



A EMT-Relevant Datasets with Observed Phenotypes and EMT Score

GEO Dataset	Sample description	Observed phenotype	Predicted EMT score	EMT Spectrum
GSE 58252	MCF-7 (br1)	Epithelial	0.190	
	MCF-7 (br2)	Epithelial	0.150	
	MCF-7 (br3)	Epithelial	0.216	
	MCF-7 SNAIL transfected (br1)	Mesenchymal	2.000	
	MCF-7 SNAIL transfected (br2)	Mesenchymal	2.000	
	MCF-7 SNAIL transfected (br3)	Mesenchymal	2.000	
GSE 14405	TEM4-18: PC-3 sub line (br1)	Mesenchymal	2.000	
	TEM4-18 PC-3 sub line (br2)	Mesenchymal	2.000	
	TEM2-5 PC-3 sub line (br1)	Hybrid E/M	0.928	
	TEM2-5 PC-3 sub line (br2)	Hybrid E/M	0.947	
GSE 24868	PC-3/Mc: PC-3 sub line (br1)	Hybrid E/M	0.956	
	PC-3/Mc PC-3 sub line (br2)	Hybrid E/M	0.948	
	PC-3/Mc PC-3 sub line (br3)	Hybrid E/M	0.958	
	PC-3/S PC-3 sub line (br1)	Mesenchymal	1.991	
	PC-3/S PC-3 sub line (br2)	Mesenchymal	1.994	
	PC-3/S PC-3 sub line (br3)	Mesenchymal	1.997	
GSE 70826	SKES1: Ewing's sarcoma	Mesenchymal	1.999	
	RDES cells (Ewing's sarcoma)	Mesenchymal	2.000	
	WE68 cells (Ewing's sarcoma)	Mesenchymal	2.000	
	SCCH cells (Ewing's sarcoma)	Mesenchymal	2.000	
	SKNMC cells (Ewing's sarcoma)	Mesenchymal	2.000	
	hMSC (Mesenchymal stem cells)	Mesenchymal	2.000	
GSE 28681	24hi: CD24+ subclone, HMLE (br1)	Epithelial	0.480	
	Msp1 mesenchymal subclone, HMLE (br1)	Mesenchymal	2.000	
	Msp2 mesenchymal subclone, HMLE (br1)	Mesenchymal	2.000	
	Msp3 mesenchymal subclone, HMLE (br1)	Mesenchymal	1.990	
	24hi CD24+ subclone, HMLE (br2)	Epithelial	0.863	
	Msp1 mesenchymal subclone, HMLE (br2)	Mesenchymal	2.000	
	Msp2 mesenchymal subclone, HMLE (br2)	Mesenchymal	1.999	
	Msp3 mesenchymal subclone, HMLE (br2)	Mesenchymal	1.975	

B Predicted EMT Scores: Unknown Phenotype

GEO Dataset	Sample description	Observed phenotype	Predicted EMT score	EMT Spectrum
GSE 36821	A549 (br1)	Unknown	1.068	
	A549 (br2)	Unknown	1.097	
	H1650 (br1)	Unknown	0.012	
	H1650 (br2)	Unknown	0.009	
	H460 (br1)	Unknown	0.000	
	H460 (br2)	Unknown	0.000	
	H1975 (br1)	Unknown	0.946	
	H1975 (br2)	Unknown	0.933	
GSE 15392	DU145 (br1)	Unknown	0.954	
	DU145 (br2)	Unknown	0.953	
	DU145 (br3)	Unknown	0.945	
GSE 10843	SW480 (br1)	Unknown	0.019	
	SW480 (br2)	Unknown	0.011	
	SW620 (br1)	Unknown	1.440	
	SW620 (br2)	Unknown	1.095	

C Predicted EMT scores: KPC mice (pancreatic tumor model)

GEO Dataset	Sample description	Predicted EMT score	EMT Spectrum
GSE 66981	KPC control mice (n=3)	1.067 ± 0.219	
	KPC Twist-KO mice (n=3)	1.014 ± 0.244	
	KPC Snail-KO mice (n=3)	0.962 ± 0.074	
GSE87472	Mes cells from KPC mice (n=6)	0.807 ± 0.028	
	Epi cells from KPC mice (n=8)	0.482 ± 0.157	
	Epi cells from KPC mice + TGFb (n=4)	0.695 ± 0.098	
	Epi cells from KPC mice ZEB-KO (n=14)	0.031 ± 0.009	
	Epi cells from KPC mice ZEB-KO + TGFb (n=4)	0.079 ± 0.016	