Integrated Systems and Technologies: *Mathematical Oncology*

Cancer Research

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]

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. Finally, 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. *Cancer Res; 77(22); 6415–28. ©2017 AACR.*

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. Finally, 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).

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 (CTC) 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 (i) coexpression of epithelial (CDH1, EPCAM) and mesenchymal (VIM, CDH2, ZEB1, SNAI2) markers, and (ii) collective cell migration by giving rise to clusters of CTCs (1, 4–7). The enhanced metastatic potential of these clusters as compared with that of individually migrating ones, a poor prognosis associated with coexpression 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

[1]Center for Theoretical Biological Physics, Rice University, Houston, Texas. [2]Department of Bioengineering, Rice University, Houston, Texas. [3]Department of Physics and Astronomy, Rice University, Houston, Texas. [4]Medical Scientist Training Program, Baylor College of Medicine, Houston, Texas. [5]Duke Cancer Institute & Department of Medicine, Duke University Medical Center, Durham, North Carolina.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

J.T. George and M.K. Jolly contributed equally to this article.

Corresponding Author: Herbert Levine, Rice University, Houston, TX 77030. Phone: 713-348-8121; E-mail:herbert.levine@rice.edu

doi: 10.1158/0008-5472.CAN-16-3521

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 and 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, $\hat{\pi}^{(m)}$. In this way, the output of model m for sample s is indicated by $\{\hat{\pi}_{s,1}^{(m)}, \hat{\pi}_{s,2}^{(m)}\}$, where $\hat{\pi}_{s,k}^{(m)}$ 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 $\hat{\pi}_{s,3}^{(m)} = 1$). 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

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

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, $\hat{\pi}_{j,k}^{(m)}$ the fitted, cumulative distribution value for the jth 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 $(\alpha_1, \alpha_2, \beta_1, \beta_2)$ is given by

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

defined for $k = 1,\ 2$, where $X_{j,1}$ and $X_{j,2}$ represent the jth 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:

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

This provides an explicit representation for the categorical probabilities as:

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

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 (Eq. E). P_H is calculated by Eq. D 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, $\mu = 2$ a purely M signature, and $\mu = 1$ a maximally hybrid E/M signature.

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

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. In addition, 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 ($\sim$ 20) of gene products that the model predicts to be least correlated with respect to EMT. Finally, 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.

to as "phenotypic stability factors" (PSF), and their elevated expression, indicative of a stable hybrid E/M state, can be 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

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., gene products) 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 (Eqs. D and E). 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. In addition, 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 compiled from the literature and employed as the space of possible EMT predictors, significantly reducing the high dimensional input space of all possible gene products (Supplementary Data; refs. 16–20). This restriction helps to mitigate overfitting 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 markers such as E-cadherin and vimentin—CDH1/VIM, or that of a canonical mesenchymal gene and a typical "PSF" 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. Overfitting 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 not feasible 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 (Eq. A). 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 a 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, as 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. 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 (Eqs. B and C). Leave-one-out analysis was employed to characterize the predictive capability of the model and ensure that the algorithm was not significantly overfitting 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 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. To prevent over reliance on a single normalizer, the 20 lowest-ranked gene products that show nonsaturated 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 value 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, $E_{test} - E_{train}$, is subtracted from each expression value in the test set for fair predictions (for example, if there is no difference in background expression, then 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 that fall outside of range (greater than 5-fold on either axis, when compared with 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. 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 (Eq. 5). In particular, $0 \leq \mu < 0.5$ corresponds to an epithelial prediction, $0.5 \leq \mu \leq 1.5$ to a hybrid E/M prediction, and $1.5 < \mu \leq 2$ 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 with previously reported EMT status to compare EMT categorization with known results. Additional predictions were made on datasets with unknown EMT state. Finally, the model was applied to large sample The Cancer Genome Atlas (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 DMEM supplemented with 10% FBS 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, 1× IMPROMII reverse transcriptase buffer, 10 mol/L dNTPs, 3.75 mmol/L $MgCl_2$, 0.1 L RNasin, and 1 L of IMPROMII reverse transcriptase in a total volume of 20 L. Following reverse transcription, cDNAs were diluted 1:5 with nuclease-free H_2O, and qPCRs were prepared using 2 L of diluted cDNA, 5 L of SYBR Master Mix (Kapa Biosystems), and 60 nmol/L of each primer in a 10:1 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 *post hoc* correction. Any $P < 0.05$ was considered statistically significant.

Western blotting and immunofluorescence staining

To prepare cells for Western blots, cells were plated at 3×10^5 cells/well in 6-well format. The next day, cells were lysed in ice-cold 1× radio-immunoprecipitation assay buffer supplemented with 1× Halt Protease and Phosphatase Inhibitor Cocktail (Thermo Fisher Scientific). 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 1× SDS loading buffer, and proteins were separated in 4%–15% MiniPROTEAN TGX Precast Gels (Bio-Rad) 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 5 minutes each with PBS, incubated 1 hour at room temperature in a 1:20,000 dilution of LI-COR anti-mouse 800 and LI-COR anti-rabbit 680 diluted in StartingBlock PBS buffer, washed two times for 5 minutes each with PBS, and imaged using the LI-COR Odyssey imaging system. For immunofluorescence staining, cells were plated at 5×10^4 cells/well in 24-well format and allowed to grow for 48 hours prior to fixing to allow reestablishment of E-cadherin at cell membranes. Cells were then fixed in 4% paraformaldehyde for 15 minutes, permeabilized in PBS+0.2% Triton X-100 for 30 minutes at room temperature, blocked for 30 minutes in 5% 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 1 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; catalog no. 610181), mouse anti-vimentin (ABD Serotec; catalog no. MCA862), anti-Zeb1 (Santa Cruz Biotechnology; catalog no. sc-25388), and rabbit anti-GAPDH (Santa Cruz Biotechnology; catalog no. 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; catalog no. MG2A00), mouse IgG1 isotype control antibody (Life Technologies; catalog no. MG100), Zenon AlexaFluor 488 mouse IgG1 labeling kit (Thermo Scientific; catalog #Z25002), Zenon Alexa Fluor 647 mouse IgG1 labeling kit (Thermo Scientific; catalog no. Z25108), mouse anti-E-cadherin (BD Biosciences; catalog no. 610181), and mouse anti-vimentin (ABD Serotec; catalog no. 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 (Fig. 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 1). Not surprisingly, this list contains canonical epithelial and mesenchymal markers such as CDH1 (E-cadherin) and VIM (vimentin) respectively. Importantly, it also contains 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 1; refs. 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 to identify 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 1). 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 2).

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 3). The best-fit model we ultimately utilized is completely described by $\beta = [-7.87,\ 0.0413,\ 1.36, -1.96]$ (see Eq. B). However, all top 10 combinations fit training data with near-equal ability (Table 3). The frequent presence of PSFs such as OVOL1, OVOL2, and/or GRHL2 in this list of top 10 two-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 4). 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 (see Supplementary Data).

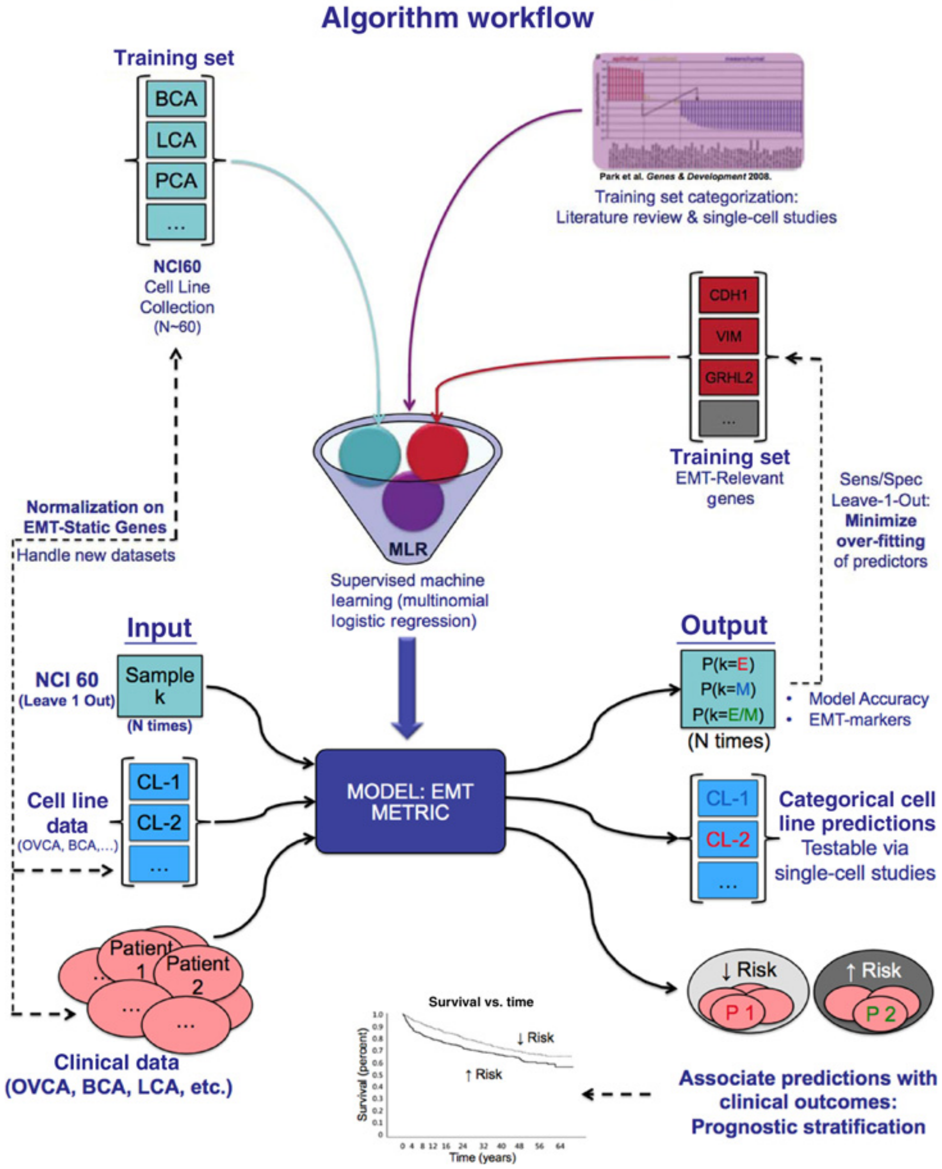


Figure 1.

Schematic illustration of model construction and prediction. Input elements relevant to model construction include 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).

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 (Fig. 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 on the basis of the ratio of protein levels of CDH1 and VIM; ref. 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 with the output generated by an average, EMT-relevant pair of predictors (Supplementary Table S1A). Finally, the addition of another predictor to the top 50 two-predictor combinations does not result in significant changes in leave-one-out sensitivity and specificity (Supplementary Table S1B). This observation does not rule out the possibility that a new three-predictor combination may outperform the best two-predictor combination. However, given our computational limitations and reservations for model overfitting, we are satisfied with using the two most relevant predictors in combination to quantify EMT.

Normalization with respect to 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 (Eq. D, $n = 1$), E/M (Eq. D, $n = 2$), and M (Eq. D, $n = 3$; Fig. 2A). Projections of each probability into the x-y plane reveal the relevant range for

Table 1. Iterative regression output (top 5% of EMT-relevant genes)

Predictor	Deviance[a]
CDH1/VIM	37.61
OVOL2/VIM[b]	45.96
VIM/CDH1	46.74
TMEM125	49.74
VIM/GRHL2[b]	50.60
GRHL2[b]	51.47
GRHL2/VIM	51.50
VIM/OVOL2[b]	51.85
RAB25	52.12
CLDN7	52.48
BICDL2	53.47
IRF6	53.91
TMC4	54.27
CDH3/VIM	55.75
VIM/OVOL1[b]	57.12
VIM	57.50
OVOL1/VIM[b]	57.64
C1ORF210	58.44
MARVELD3	59.07
CDS1	59.34
BSPRY	59.39
CDH1	59.50
ANAX9	59.79

NOTE: 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.

[a] Deviance, *D*, as defined in Eq. A.

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

Table 2. EMT-Normalizer

Normalizer
SLC25A42
SNX13
TAF4B
CDK2
MBNL1
NEURL1B
ANG
PPFIBP1
PACSIN1
LRRTM1
TMEM182
CSMD1
ZNF503-AS2
CCNF
DIRC1
MBTPS2
RNF150
RC3H2
UBE3C

NOTE: Gene products that show the weakest correlation to training set categories are identified as normalizers, used for cross-data comparison.

Table 3. Top 10 two-predictor combinations

Rank	Predictor 1	Predictor 2	Deviance[a]
1)	CLDN7[b]	VIM/CDH1[b]	26.78
2)	VIM/GRHL2	OVOL1/CDH1	27.73
3)	VIM/CDH1	VIM/GRHL2	28.27
4)	GRHL2	VIM/CDH1	28.31
5)	ST3GAL2	VIM/CDH1	28.31
6)	VIM/CDH1	GRHL2/CDH1	28.48
7)	VIM/CDH1	OVOL2/VIM	28.56
8)	GRHL2/CDH1	VIM/GRHL2	28.63
9)	VIM/CDH1	GRHL2/VIM	28.86
10)	OVOL1	VIM/CDH1	29.08

NOTE: The top 10 optimal predictor combinations are ranked according to their deviance.

[a] Deviance, *D*, as defined in Eq. A.

[b] Top predictors (X1,X2) used in model construction.

Table 4. Leave-one-out analysis: CLDN7, VIM/CDH1

Category	Sensitivity	Specificity
E	100%	98%
E/M	55%	90%
M	86%	82%
Diagnostic accuracy: 83%		

NOTE: Prognostic outputs of leave-one-out analysis on the top predictor set {CDH1/VIM, CLDN7} are provided.

which each phenotype resides (Fig. 2B). The representation of EMT status as the maximal predicted probability state can be appreciated by projecting Eq. D ($n = 1, 2, 3$). Overlaying NCI-60 data reveals that a majority of training samples fall within their expected range (Fig. 2C). A prototypical demonstration of normalization is provided for cell lines composed of CD44^{+}/CD24^{-} and CD44^{-}/CD24^{+} human mammary epithelial cells (GSE15192; Fig. 2D). Here, prenormalized (purple) and postnormalized (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 Supplementary Fig. 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; Table 5A; ref. 29), 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 with TEM4-18 (30), were predicted to be hybrid E/M (Table 5A). Similarly, PC-3/Mc cells, a subpopulation of PC-3 cells that coexpressed CD24 and CD44 (ref. 31; a signature of hybrid E/M; ref. 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 5A; 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 5A) and osteosarcoma (GSE70414, GSE55957; Supplementary Table S2) datasets were predicted to be mesenchymal, and the epithelial and mesenchymal subpopulations of HMLE cells (GSE28681; Table 5A) had significantly different EMT scores. The algorithm also predicts that short-term treatment of cells with EMT or MET inducers is usually not sufficient to drive a transition (GSE7868, GSE17708, GSE59771, and GSE53603; Supplementary 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; ref. 34) were predicted as hybrid E/M, but cell lines established from those with KO-*Zeb1* KPC mice (GSE87472; ref. 35) were categorized as almost purely epithelial

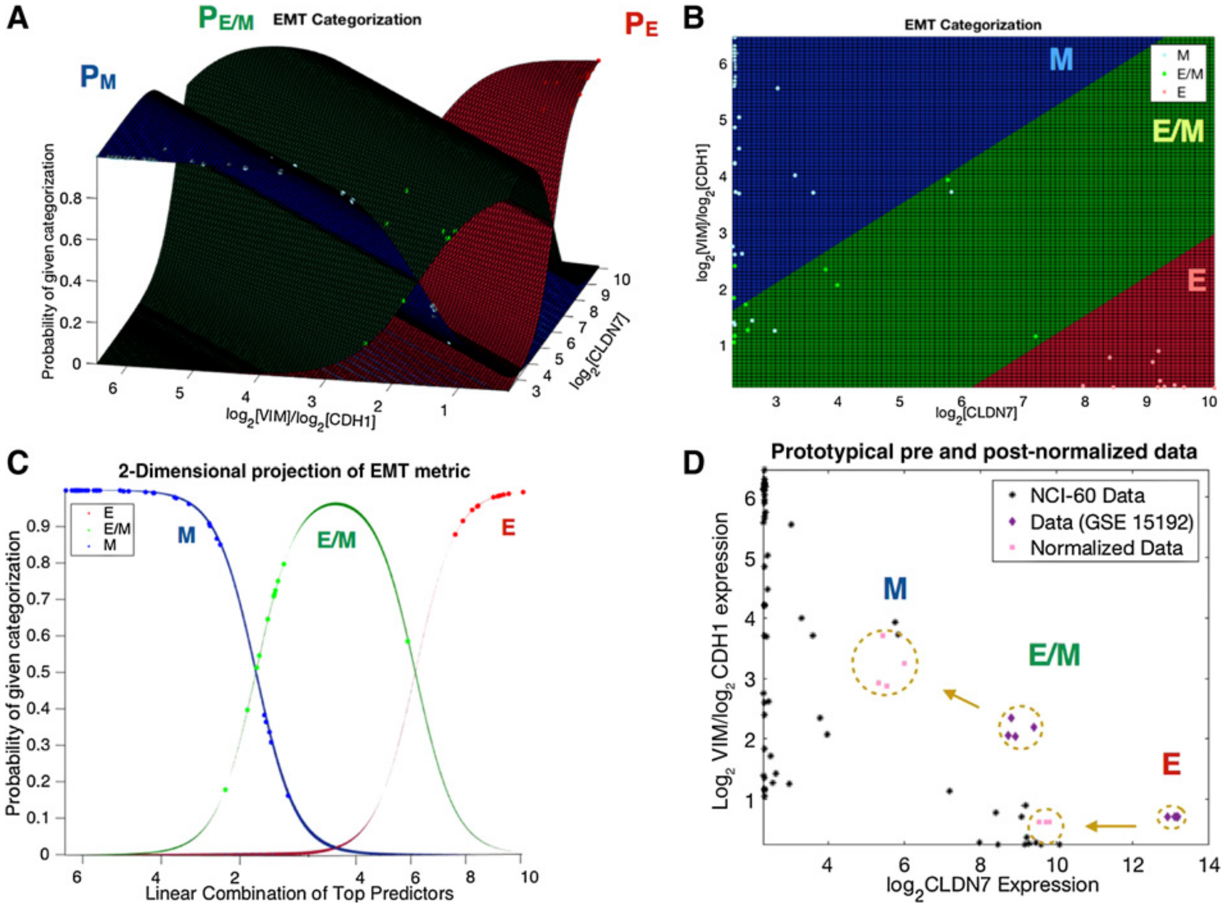


Figure 2.

Model representation. **A,** Three-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,** Two-dimensional model projection of model for use in defining the EMT metric, μ, described by Eq. E. **D,** Prototypical example of pre- versus postnormalization comparisons in an immortalized human mammary epithelial cell line (GSE15192).

(Table 5C). Furthermore, 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 coexpress epithelial and mesenchymal markers

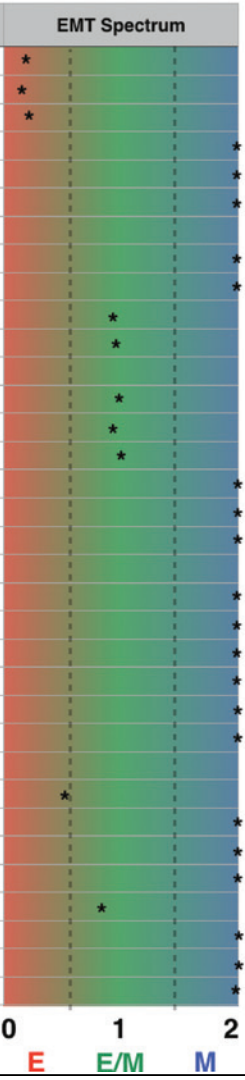
Next, we ran our model for transcriptomes of multiple cell lines, including SW480 and SW620 (both colorectal cancer), DU145 (prostate cancer), and A549, H1975, H460, and H1650 (all non–small cell lung cancer; GSE36821, GSE15392, GSE10843). SW480, H460, and H1650 were predicted to be epithelial, whereas H1975, DU145, SW620, and A549 were predicted to be hybrid E/M (Table 5B). Consistent with their predicted phenotypes, H1975 cells have been shown to stably coexpress E-cadherin and vimentin at a single-cell level (12), whereas 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 (Supplementary Fig. S2A). Interestingly, the SW480 cells, which were predicted to be epithelial, also resembled hybrid cells in their expression of ZEB1 and SNAIL (Supplementary Fig. 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 with MCF-7 (Supplementary Fig. S2B and S2C). E-cadherin levels were also substantially lower in the hybrid E/M lines and SW480 when compared with MCF-7 at both the mRNA (Supplementary Fig. S2B and S2C) and protein (Fig. 3A) levels, with variable levels of vimentin protein (Fig. 3A). Together, these results confirm that the cell lines predicted as hybrid coexpress epithelial and mesenchymal biomarkers at intermediate levels compared with strongly epithelial or strongly mesenchymal cell lines.

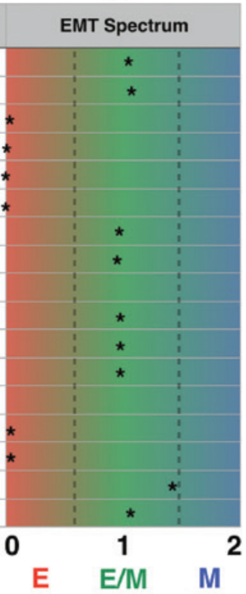
All of the datasets above contain gene expression on an ensemble level instead of single-cell gene expression data. Therefore, a hybrid E/M signature may be predicted either because they truly contain hybrid E/M cell coexpressing epithelial and mesenchymal

Table 5. Model predictions on relevant *in vitro* experimental datasets

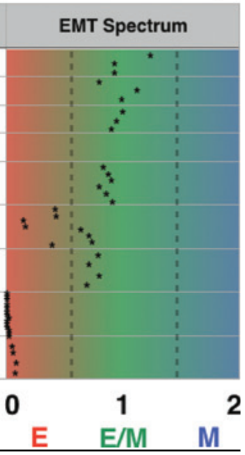
A EMT-Relevant Datasets with Observed Phenotypes and EMT Score

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	

NOTE: **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 Eq. E (μ<0.5 corresponds to E, $0.5<\mu<1.5$ 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).

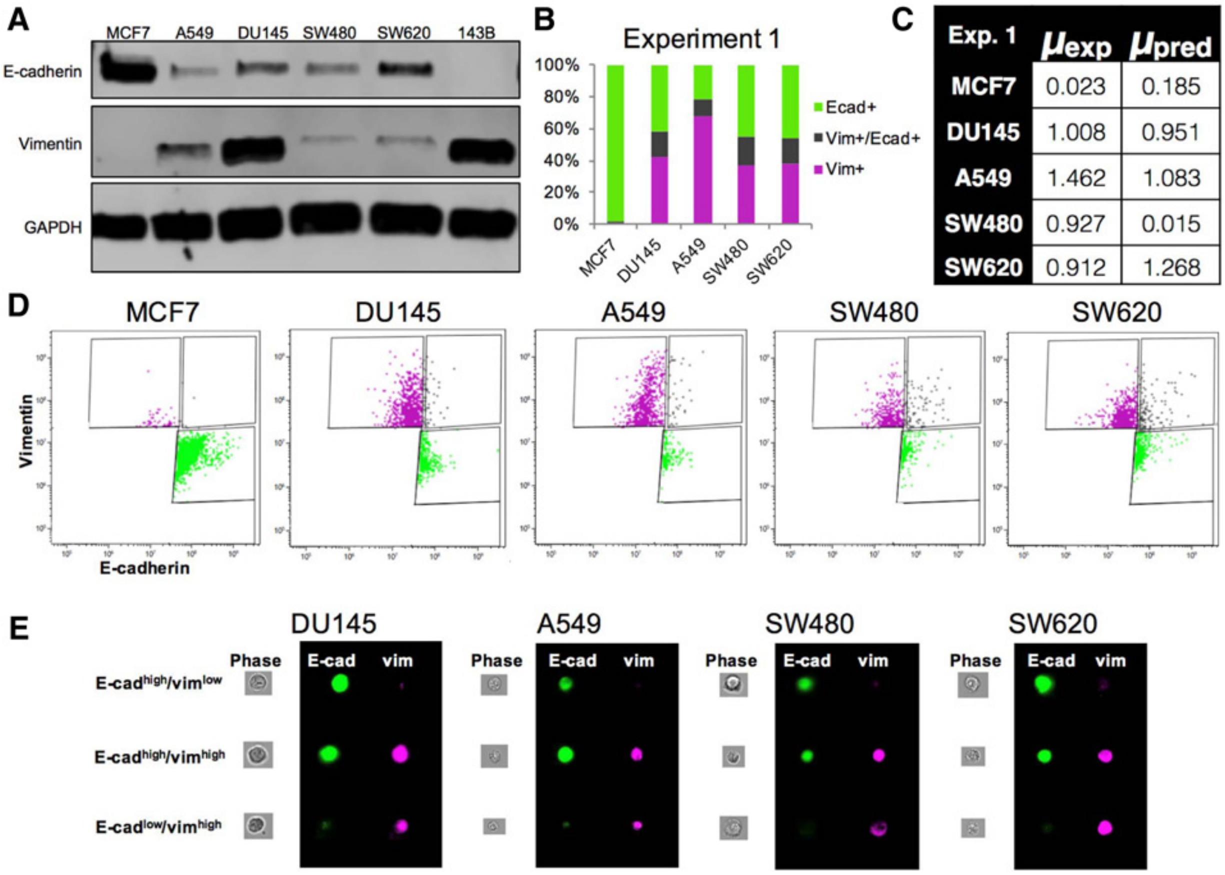


Figure 3.

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 coexpression 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 with epithelial MCF-7 cells for the data presented in **D**. **C,** Comparison of experimentally observed EMT score for DU145, A549, SW480, and SW620 cells (μ_{exp}) and theoretical prediction of EMT score via Eq. E (μ_{pred}). **D,** Flow cytometry analysis of CDH1^{high}/VIM^{low} (green), CDH1^{high}/VIM^{high} (gray), and CDH1^{low}/VIM^{high} (magenta) subpopulations. **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.

markers (as shown for H1975), or because they are comprised of subpopulations of epithelial and mesenchymal phenotypes. 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 coexpressed 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 subpopulations of epithelial-like (CDH1^{high}/VIM^{low}), hybrid E/M (CDH1^{high}/VIM^{high}), and mesenchymal-like (CDH1^{low}/VIM^{high}) cells (Fig. 3B–D). An experimental quantification of each sample's EMT score, u_{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: $u_{exp} = 0\cdot[\%\mathrm{CDH1}^{+}/\mathrm{VIM}^{-}\ \mathrm{cells}]+1\cdot[\%\mathrm{CDH1}^{+}/\mathrm{VIM}^{+}\mathrm{cells}]+2\cdot[\%\mathrm{CDH1}^{-}/\mathrm{VIM}^{+}\mathrm{cells}]$. This was compared with theoretical predictions of EMT scores using Eq. E (Fig. 3C; Supplementary Fig. 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 subpopulations 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} (Fig. 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 costaining of membrane-localized CDH1 and VIM in numerous cells (Fig. 4A). On the other hand, A549 cells were predominantly CDH1-low and VIM-positive, with distinct clusters of CDH1^{+}/VIM^{-} cells (Fig. 4B). Like the DU145 cells, SW480 cells also contained a population of cells with coexpression of CDH1 and VIM (Fig. 4C); however, a subset of SW480 cells possessed CDH1^{+}/VIM^{-} cell clusters (Fig. 4C). The SW620s displayed a patchier distribution of membrane CDH1 positivity and strong VIM expression, with a small subpopulation of cells that coexpress CDH1 and VIM (Fig. 4D).

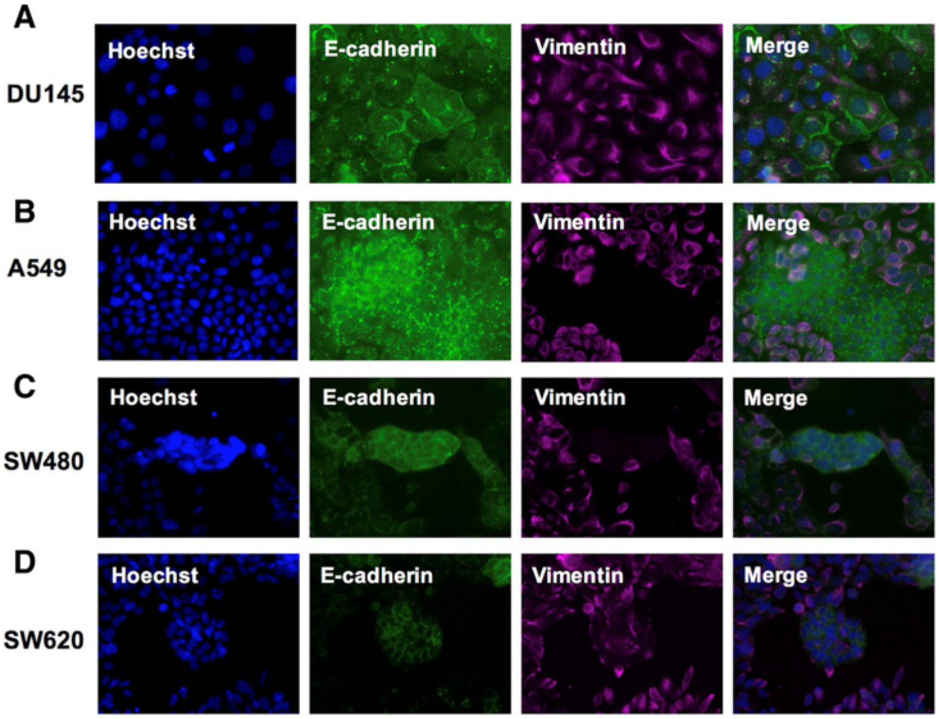


Figure 4.
Validation of the hybrid E/M state reveals distinct subpopulations of epithelial-like, hybrid, and mesenchymal-like cells. **A,** DU145 cell line contains cells that coexpress membrane CDH1 and VIM. **B,** A549 cells predicted to be hybrid E/M contain subpopulations of CDH1^{high}/VIM^{low} and CDH1^{low}/VIM^{high} cells, along with cells that coexpress 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.

Together, our quantitative analysis at the single-cell level revealed that the cell lines 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 nonepithelial signatures for multiple breast cancer datasets. In a majority of cases (Fig. 5A–E), patients exhibiting a more epithelial phenotype had poorer survival as compared with those displaying a partial or full EMT signature (GSE17705, GSE1456, GSE45255, GSE5327, GSE6532). Although statistically significant, some of these cases—especially Fig. 5A (HR = 0.760), Fig. 5B (HR = 0.614), and Fig. 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 with patients with an epithelial signature (Fig. 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 with those binned for epithelial phenotype, with a relatively wider separation in clinical parameters (Fig. 5G and 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 (Fig. 5I and 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: (i) using median levels of CDH1/VIM to resolve patients into two groups, CDH1/VIM^{high} and CDH1/VIM^{low} (Supplementary Fig. S4), and, (ii) using CDH1 and VIM as the two predictors in our statistical model (Supplementary Fig. 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 multidimensional 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 and lung cancer samples displayed an epithelial phenotype predominantly, and most sarcoma samples were categorized as mesenchymal. Notably, pancreatic

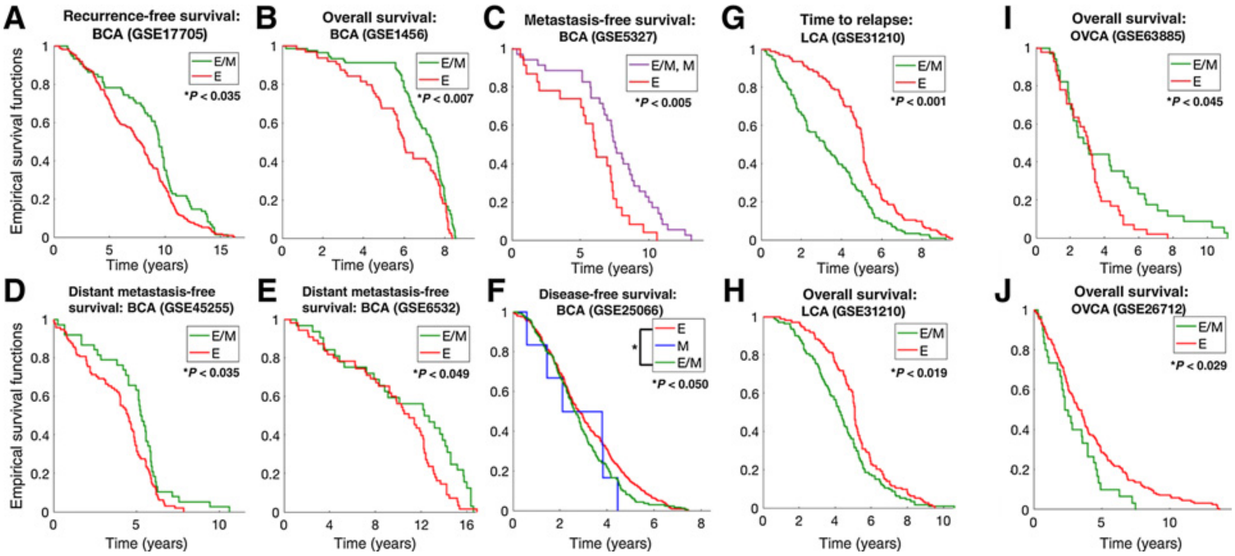


Figure 5.

Correlation between EMT status and clinical survival metrics. Kaplan–Meier survival analysis is performed 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 HRs 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.

adenocarcinoma (PDAC) and renal clear cell carcinoma (RCC) samples were enriched for a hybrid E/M phenotype (Supplementary Fig. S6A), reminiscent of coexpression of epithelial and mesenchymal markers *in vivo* in PDAC and *in vitro* in RCC cell lines (1). Finally, 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 (Supplementary Fig. 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 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 coexpressing 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 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 coexpressing 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.

Disclosure of Potential Conflicts of Interest

J. Somarelli reports receiving a commercial research grant from Arvinas. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: J.T. George, M.K. Jolly, H. Levine
Development of methodology: J.T. George, M.K. Jolly, H. Levine
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J. Xu, J. Somarelli
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): J.T. George, M.K. Jolly, J. Xu, J. Somarelli, H. Levine
Writing, review, and/or revision of the manuscript: J.T. George, M.K. Jolly, J. Somarelli, H. Levine
Study supervision: J.T. George

Acknowledgments

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

Grant Support

J.T. George was supported by the National Cancer Institute of the NIH (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). 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). 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).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 30, 2016; revised June 12, 2017; accepted September 19, 2017; published OnlineFirst September 25, 2017.

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.
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:E-1555–64.
3. Huang RY-J, Wong MK, Tan TZ, Kuay KT, Ng a HC, 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.
4. Nieto MA, Huang RY, Jackson RA, Thiery JP. EMT: 2016. Cell 2016;166:21–45.
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;10:253–71.
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;339:580–4.
7. Aceto N, Toner M, Maheswaran S, Haber DA. En route to metastasis: circulating tumor cell clusters and epithelial-to-mesenchymal transition. Trends Cancer 2015;1:44–52.
8. Cheung KJ, Ewald AJ. A collective route to metastasis: Seeding by tumor cell clusters. Science 2016;352:167–9.
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.
10. Jolly MK, Boareto M, Debeb BG, Aceto N, Farach-Carson MC, Woodward WA, et al. Inflammatory breast cancer: a model for investigating cluster-based dissemination. NPJ Breast Cancer 2017;3:21.
11. Savagner P. The epithelial-mesenchymal transition (EMT) phenomenon. Ann Oncol 2010;21Suppl 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;7:27067–84.
13. Jia D, Jolly MK, Boareto M, Parsana P, Mooney SM, Pienta KJ, et al. OVOL guides the epithelial-hybrid-mesenchymal transition. Oncotarget 2015;6:15436–48.
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 Comput Biol 2015;11:e1004569.
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;22:894–907.
16. Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J, 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.
17. Loboda A, Nebozhyn MV, Watters JW, Buser CA, Shaw PM, Huang PS, et al. EMT is the dominant program in human colon cancer. BMC Med Genomics 2011;4:9.

18. van't Veer LJ, Dai H, Vijver Hvd, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.
19. Ben-Porath I, Thompson MW, Carey VJ, Ge R, Bell GW, Regev A, et al. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nature Genetics 2008;40:499–507.
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.
21. Bellman RE, Dreyfus SE. Applied dynamic programming. Princeton, NJ: 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, Salmans ML, Fallahi M, Andersen B, 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.
24. Roca H, Hernandez J, Weidner S, McEachin RC, Fuller D, Sud S, et al. Transcription factors OVOL1 and OVOL2 induce the mesenchymal to epithelial transition in human cancer. PLoS One 2013;8:e76773.
25. Xiang X, Deng Z, Zhuang X, Ju S, Mu J, Jiang H, et al. Grhl2 determines the epithelial phenotype of breast cancers and promotes tumor progression. PLoS One 2012;7:e50781.
26. Bhat AA, Pope JL, Smith JJ, Ahmad R, Chen X, Washington MK, et al. Claudin-7 expression induces mesenchymal to epithelial transformation (MET) to inhibit colon tumorigenesis. Oncogene 2015;34:4570–80.
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.
28. Ribeiro AS, eParedes J. P-cadherin linking breast cancer stem cells and invasion: a promising marker to identify an intermediate/metastable EMT state. Front Oncol 2015;4:371. doi: 10.3389/fonc.2014.00371.
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; 29:1280–9.
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;20:2207–17.
31. Celiá-Terrassa T, Meca-Cort'es 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;122:1849–68.
32. Ombrato L, Malanchi I. The EMT universe: space between cancer cell dissemination and metastasis initiation. Crit Rev Oncog 2014;19:349–61.
33. Shibue T, Weinberg RA. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. Nat Rev Clin Oncol 2017;14;611–29.
34. Zheng X, Carstens JL, Kim J, Scheible M, Kaye J, Sugimoto H, et al. Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. Nature 2015;527:525–30.
35. Krebs AM, Mitschke J, Losada ML, Schmalhofer O, Boerries M, Busch H, et al. The EMT-activator Zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic cancer. Nat Cell Biol 2017;19:518.
36. Somarelli JA, Shetler S, Jolly MK, Wang X, Bartholf Dewitt S, Hish AJ, 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.
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: 1789–800.
38. Tan TZ, Miow QH, Miki Y, Noda T, Mori S, Huang RY, 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.
39. Goldman A, Majumder B, Dhawan A, Ravi S, Goldman D, Kohandel M, et al. Temporally sequenced anticancer drugs overcome adaptive resistance by targeting a vulnerable chemotherapy-induced phenotypic transition. Nat Commun 2015;6:6139.
40. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.
41. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell 2015;163:506–19.
42. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–7.
43. Grasso CS, Wu Y, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature 2012;487:239–43.
44. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 2013;499:43–9.
45. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489:519–25.
46. Silver SJ, Lash A, Lau C, et al. Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. Nat Genet 2010;42:715–21.
47. Jolly MK, Ware KE, Gilja S, Somarelli JA, Levine H. EMT and MET: necessary or permissive for metastasis? Mol Oncol 2017;11:755–69.
48. Fischer KR, Durrans A, Lee S, Sheng J, Li F, Wong ST, et al. Epithelial-to-mesenchymal transition is not required for lung metastasis but contributes to chemoresistance. Nature 2015;527:472–6.
49. Byers LA, Diao L, Wang J, Saintigny P, Girard L, Peyton M, 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.
50. Boareto M, Jolly MK, Goldman A, Pietilä M, Mani SA, Sengupta S, 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.

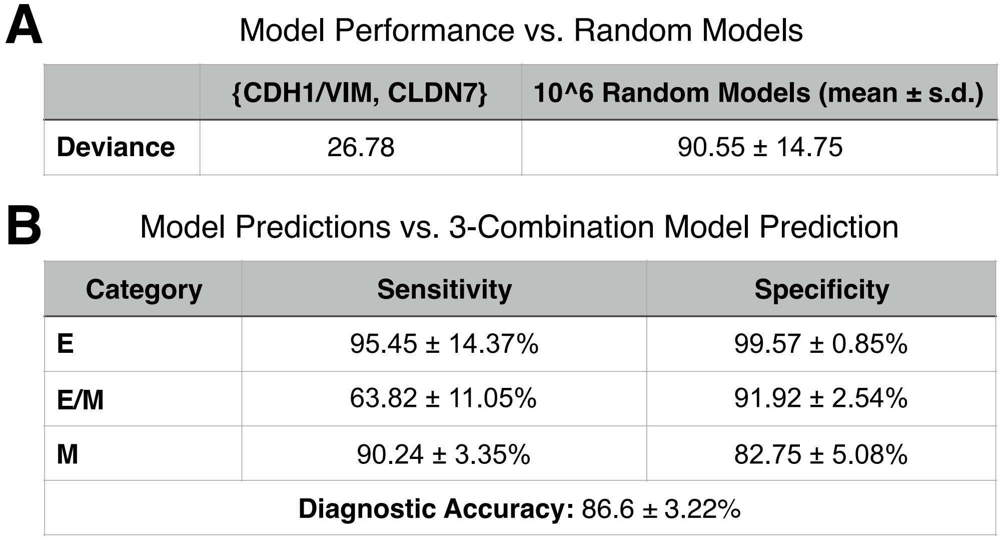
Supplementary Information: Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric

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

September 11, 2017

[1]Center for Theoretical Biological Physics, [2]Deparment of Bioengineering, [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. [*]These authors contributed equally. [†]Corresponding Author: Herbert Levine (herbert.levine@rice.edu).

Supplementary Figures and Tables



A

Model Performance vs. Random Models

	{CDH1/VIM, CLDN7}	10^6 Random Models (mean ± s.d.)
Deviance	26.78	90.55 ± 14.75

B

Model Predictions vs. 3-Combination Model Prediction

Category	Sensitivity	Specificity
E	95.45 ± 14.37%	99.57 ± 0.85%
E/M	63.82 ± 11.05%	91.92 ± 2.54%
M	90.24 ± 3.35%	82.75 ± 5.08%
Diagnostic Accuracy: 86.6 ± 3.22%		

Table S1: {CDH1/VIM, CLDN7} vs. Other Models.

(A) The goodness of fit for the {CDH1/VIM,CLDN7} model is compared to the mean ± s.d. for that of 10^6 randomly generated models. Better fit is reflected in lower deviance values, indicating significant improvements by using the generated model; (B) Mean and standard deviation values for sensitivity and specificity are provided for models that include an additional (third) best predictor in combination with the best pair for the top 50-combination predictors. There is no statistically significant difference between any of the categories and the top 2-combination predictor selected for analysis, and so for simplicity and to avoid over-fitting, we proceed to characterize EMT using the model built on CLDN7 and VIM/CDH1.

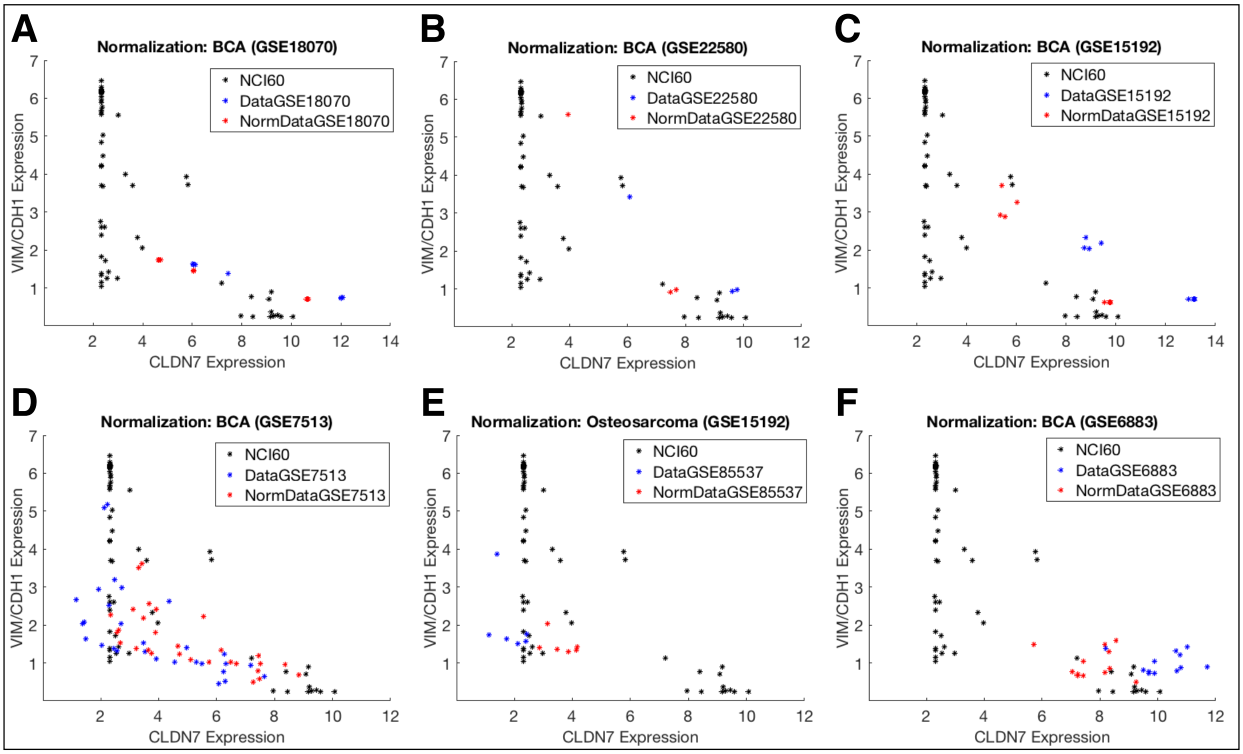


Figure S1: Additional examples of normalization.

Further examples of normalization are provided for (A) MCF10A and transformed MCF10ATk.cl2 and MCF10CA1h mammary epithelial cell lines (GSE18070); (B) Type I K5+/K19- and Type II K5+/K19+ immortalized human mammary epithelial cells (GSE22580); (C) Normal and malignant CD44+/CD24- and CD44-/CD24+ breast epithelial MCF-10A cells (GSE15192); (D) Core biopsies of primary human CD44+/CD24-, CD24+, and CD44-/CD24+ breast tumors (GSE7513); (E) MCF-10A CD44+/CD24- and CD44-/CD24+ breast epithelial cell lines (GSE15192), and; (F) CD44+/CD24- tumorigenic breast-cancer cells and normal breast epithelium.

Additional EMT Score Calculations

GEO Dataset	Sample description	Observed phenotype	Predicted EMT score	EMT Spectrum
GSE 70414	MG63	Mesenchymal	2.000	*
	Saos	Mesenchymal	2.000	*
	HOS	Mesenchymal	2.000	*
	NY	Mesenchymal	2.000	*
	Hu09	Mesenchymal	2.000	*
	hMSC	Mesenchymal	2.000	*
GSE 55957	ZOS osteosarcoma	Mesenchymal	1.685	*
	ZOSM osteosarcoma	Mesenchymal	1.841	*
GSE 7868	LNCaP expression at 0 hr (n=3)	Epithelial	0.014 ± 0.005	*
	LNCaP expression at 4 hr (n=3)	Epithelial	0.016 ± 0.002	*
	LNCaP expression at 16 hr (n=3)	Epithelial	0.014 ± 0.002	*
GSE 17708	A549 untreated (n=3)	Hybrid E/M	0.955 ± 0.002	*
	A549 TGFB1 0.5 hr (n=3)	Hybrid E/M	0.958 ± 0.004	*
	A549 TGFB1 1 hr (n=3)	Hybrid E/M	0.956 ± 0.002	*
	A549 TGFB1 2 hr (n=2)	Hybrid E/M	0.954 ± 0.003	*
	A549 TGFB1 4 hr (n=3)	Hybrid E/M	0.957 ± 0.003	*
	A549 TGFB1 8 hr (n=3)	Hybrid E/M	0.961 ± 0.002	*
	A549 TGFB1 16 hr (n=3)	Hybrid E/M	1.040 ± 0.002	*
	A549 TGFB1 24 hr (n=3)	Hybrid E/M	1.046 ± 0.004	*
	A549 TGFB1 72 hr (n=3)	Hybrid E/M	1.049 ± 0.006	*
GSE 59771	LSTGFBR2-Ctrl (n=2)	Epithelial	0.019 ± 0.002	*
	LSTGFBR2-Ctrl (n=2)	Epithelial	0.017 ± 0.002	*
GSE 53603	Vehicle 6 hr (n=2)	Hybrid E/M	0.886 ± 0.057	*
	SAHA 6 hr (n=2)	Hybrid E/M	0.865 ± 0.035	*
GSE 53603	Vehicle 24 hr (n=3)	Hybrid E/M	0.717 ± 0.042	*
	SAHA 24 hr (n=2)	Hybrid E/M	0.935 ± 0.008	*

0 1 2

E E/M M

Table S2: Additional EMT score categorization.

Model predictions on datasets across multiple cancer types: GSE70414-osteosarcoma and GSE 55957-osteosarcoma cell lines, GSE7868-LNCaP cells treated with DHT for 0, 4, 16 hr, GSE17708-time-course TGFb treatment of A549 for 0, 0.5, 1, 2, 4, 8, 16, 24, and 72 h, GSE59771-CRC cell line LS174T with restored TGFBR2 expression (LS) treated with TGFB for 16 hr, GSE53603-SKOV3 cells treated with vehicle or SAHA. 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 ($\mu < 0.5$ corresponds to E, $0.5 < \mu < 1.5$ corresponds to E/M, and $\mu > 1.5$ corresponds to M).

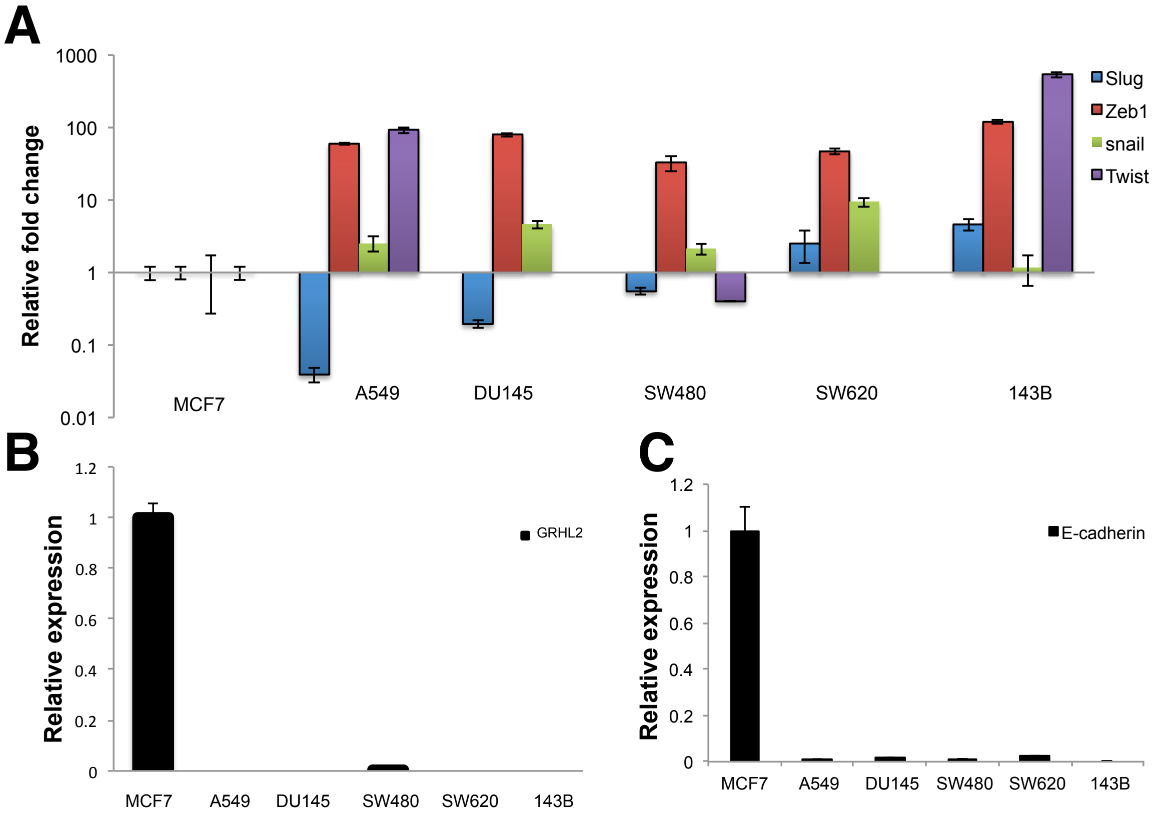
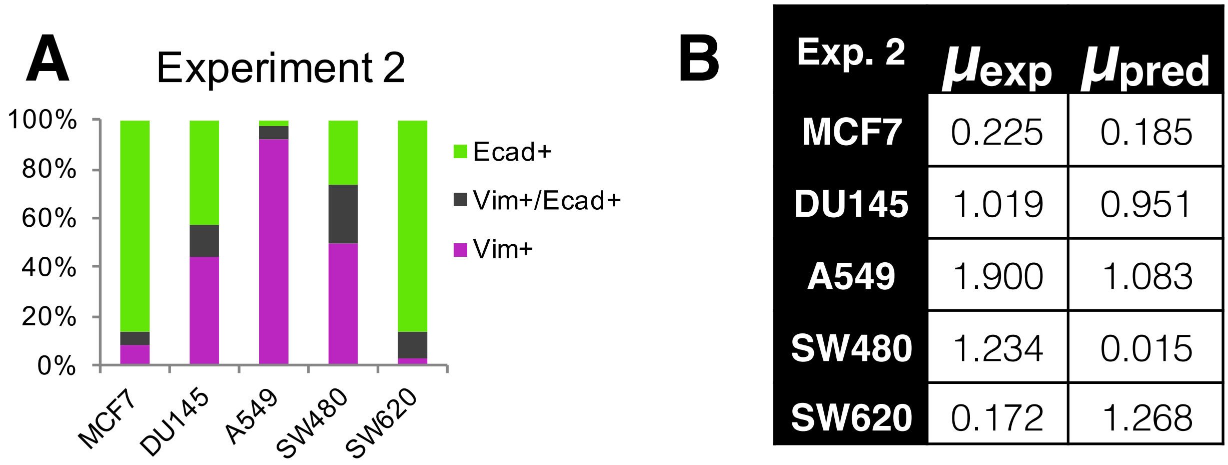


Figure S2: Levels of canonical epithelial and mesenchymal markers in multiple cell lines.

(A) RT-qPCR of EMT transcription factors Snail, Slug, Zeb1, and Twist indicate that cell lines predicted to be hybrid express higher levels of Zeb1 and Snail than the strongly epithelial cell line, MCF-7. 143B cells are included as a mesenchymal cell line control; (B) All hybrid lines have no detectable GRHL2, while the SW480 cells, predicted to be epithelial express a relatively low level of GRHL2 compared to epithelial MCF-7 cells; (C) E-cadherin is downregulated in hybrid E/M lines compared to epithelial MCF-7 cells.



Exp. 2	μ_{exp}	μ_{pred}
MCF7	0.225	0.185
DU145	1.019	0.951
A549	1.900	1.083
SW480	1.234	0.015
SW620	0.172	1.268

Figure S3: Flow cytometric quantification of epithelial-like, hybrid, and mesenchymal-like cells.

(A) Second experimental 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 (Figure 3); (B) 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}).

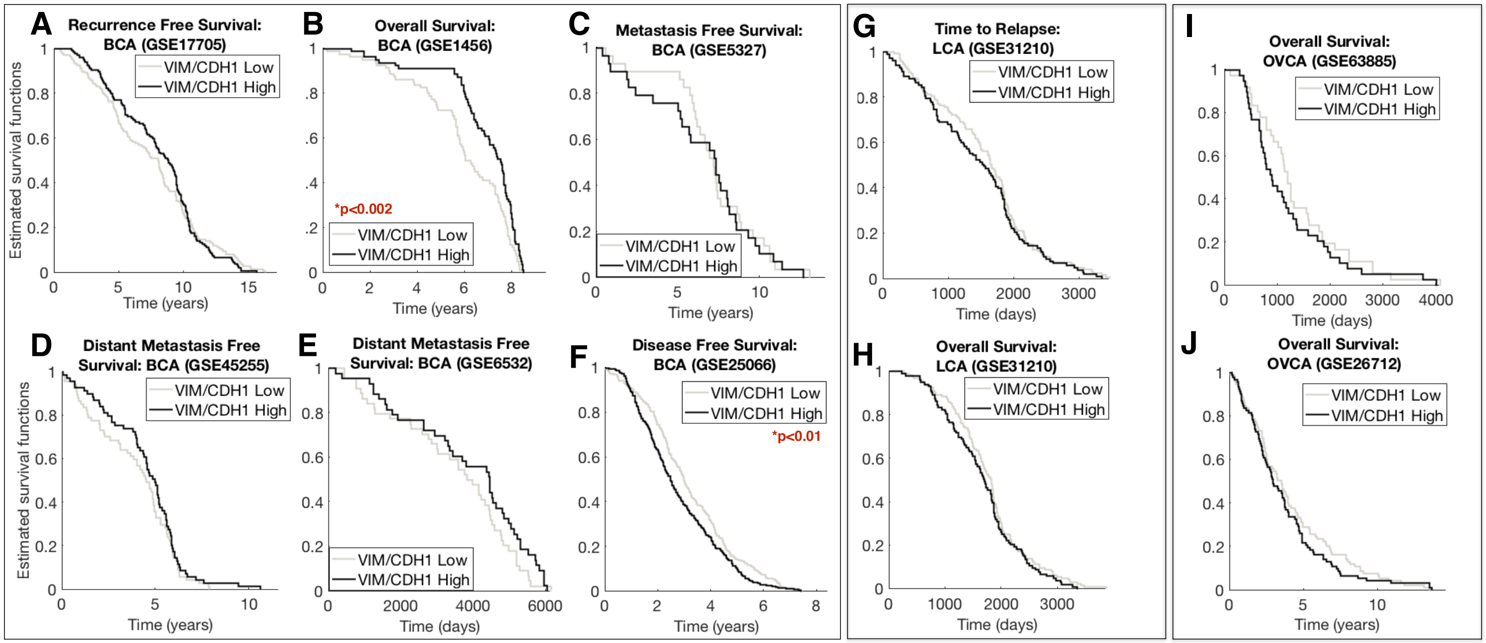

A
Recurrence Free Survival:
BCA (GSE17705)
VIM/CDH1 Low
VIM/CDH1 High
Estimated survival functions
1
0.8
0.6
0.4
0.2
0
0
5
10
15
Time (years)
B
Overall Survival:
BCA (GSE1456)
*p<0.002
VIM/CDH1 Low
VIM/CDH1 High
1
.8
.6
.4
.2
0
0
2
4
6
8
Time (years)
C
Metastasis Free Survival:
BCA (GSE5327)
VIM/CDH1 Low
VIM/CDH1 High
1
0.8
0.6
0.4
0.2
0
0
5
10
Time (years)
D
Distant Metastasis Free
Survival: BCA (GSE45255)
VIM/CDH1 Low
VIM/CDH1 High
Estimated survival functions
1
0.8
0.6
0.4
0.2
0
0
5
10
Time (years)
E
Distant Metastasis Free
Survival: BCA (GSE6532)
VIM/CDH1 Low
VIM/CDH1 High
1
0.8
0.6
0.4
0.2
0
0
2000
4000
6000
Time (days)
F
Disease Free Survival:
BCA (GSE25066)
VIM/CDH1 Low
VIM/CDH1 High
*p<0.01
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
Time (years)
G
Time to Relapse:
LCA (GSE31210)
VIM/CDH1 Low
VIM/CDH1 High
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
Time (days)
H
Overall Survival:
LCA (GSE31210)
VIM/CDH1 Low
VIM/CDH1 High
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
Time (days)
I
Overall Survival:
OVCA (GSE63885)
VIM/CDH1 Low
VIM/CDH1 High
1
0.8
0.6
0.4
0.2
0
0
1000
2000
3000
4000
Time (days)
J
Overall Survival:
OVCA (GSE26712)
VIM/CDH1 Low
VIM/CDH1 High
1
0.8
0.6
0.4
0.2
0
0
5
10
Time (years)

Figure S4: Survival Analysis distinguishing groups via median CDH1/VIM.

Kaplan-Meier survival analysis for the same datasets shown in Figure 5, but when patients are categorized into VIM/CDH1$^{\mathrm{low}}$ or VIM/CDH1$^{\mathrm{high}}$ classes based on median expression instead of being categorized via the statistical model using {CDH1/VIM, CLDN7} as the predictor set. 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=0.997 95%CI=(0.792, 1.255); (B) HR=1.561 95%CI=(1.129, 2.160); (C) HR=0.925 with 95%CI=(0.549, 1.560); (D) HR=1.205 with 95%CI=(0.855, 1.697); (E) HR=1.349 with 95%CI=(0.874, 2.084); (F) HR=0.782 with 95%CI=(0.656, 0.933); (G) HR=0.860 with 95%CI=(0.659, 1.122); (H) HR=0.895 with 95%CI=(0.687, 1.166); (I) HR=0.776 with 95%CI=(0.491, 1.228); (J) HR=0.889 with 95%CI=(0.663, 1.193).

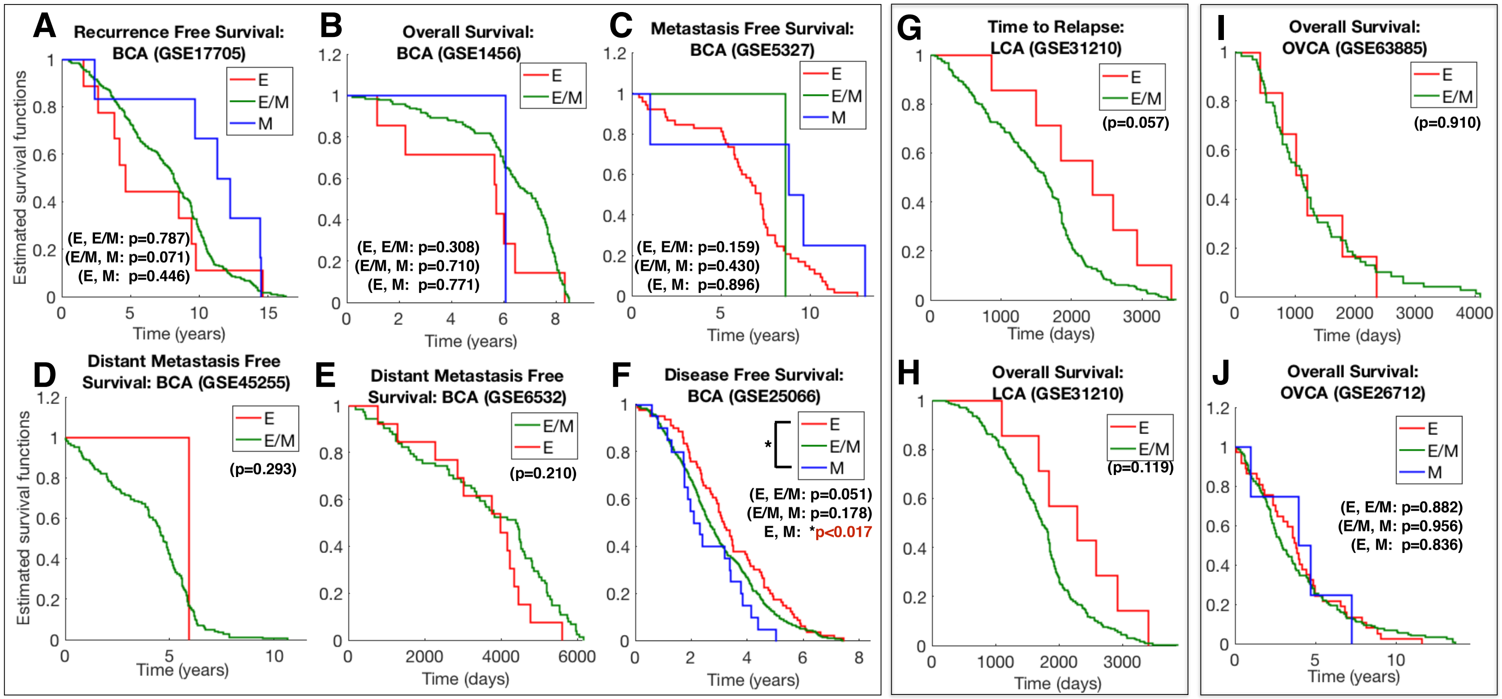

A
Recurrence Free Survival:
BCA (GSE17705)
Estimated survival functions
E
E/M
M
(E, E/M: p=0.787)
(E/M, M: p=0.071)
(E, M: p=0.446)
0 0.2 0.4 0.6 0.8 1
0 5 10 15
Time (years)
B
Overall Survival:
BCA (GSE1456)
E
E/M
(E, E/M: p=0.308)
(E/M, M: p=0.710)
(E, M: p=0.771)
0 0.2 0.4 0.6 0.8 1 1.2
0 2 4 6 8
Time (years)
C
Metastasis Free Survival:
BCA (GSE5327)
E
E/M
M
(E, E/M: p=0.159)
(E/M, M: p=0.430)
(E, M: p=0.896)
0 0.2 0.4 0.6 0.8 1 1.2
0 5 10
Time (years)
D
Distant Metastasis Free
Survival: BCA (GSE45255)
Estimated survival functions
E
E/M
(p=0.293)
0 0.2 0.4 0.6 0.8 1 1.2
0 5 10
Time (years)
E
Distant Metastasis Free
Survival: BCA (GSE6532)
E/M
E
(p=0.210)
0 0.2 0.4 0.6 0.8 1
0 2000 4000 6000
Time (days)
F
Disease Free Survival:
BCA (GSE25066)
*
E
E/M
M
(E, E/M: p=0.051)
(E/M, M: p=0.178)
E, M: *p<0.017
0 0.2 0.4 0.6 0.8 1
0 2 4 6 8
Time (years)
G
Time to Relapse:
LCA (GSE31210)
E
E/M
(p=0.057)
0 0.2 0.4 0.6 0.8 1
0 1000 2000 3000
Time (days)
H
Overall Survival:
LCA (GSE31210)
E
E/M
(p=0.119)
0 0.2 0.4 0.6 0.8 1
0 1000 2000 3000
Time (days)
I
Overall Survival:
OVCA (GSE63885)
E
E/M
(p=0.910)
0 0.2 0.4 0.6 0.8 1
0 1000 2000 3000 4000
Time (days)
J
Overall Survival:
OVCA (GSE26712)
E
E/M
M
(E, E/M: p=0.882)
(E/M, M: p=0.956)
(E, M: p=0.836)
0 0.2 0.4 0.6 0.8 1 1.2
0 5 10
Time (years)

Figure S5: Survival Analysis for model using only {CDH1, VIM} as predictors.

Kaplan-Meier survival analysis for the same datasets shown in Figure 5, but when patients are categorized into E, E/M, M using CDH1, VIM as the predictor set in our statistical model instead of using CDH1/VIM, CLDN7 as shown in Figure 5. 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) E vs. E/M - HR=1.181 95%CI=(0.576, 2.421), E/M vs. M - HR=1.764 with 95%CI=(0.997, 3.123), E vs. M - HR=1.793 with 95%CI=(0.598, 5.373); (B) E vs. E/M - HR=1.865 with 95%CI=(0.711, 4.893), E/M vs. M - HR=0.812 with 95%CI=(0.094, 7.124), E vs. M - HR=1.935 with 95%CI=(0.335, 11.180); (C) E vs. E/M - HR=0.508 with 95%CI=(0.224, 1.154), E/M vs. M - HR=1.994 with 95%CI=(0.585, 6.802), E vs. M - HR=0.362 with 95%CI=(0.017, 7.723); (D) HR=0.474 with 95%CI=(0.158, 1.423); (E) HR=1.671 with 95%CI=(0.828 ,3.373); (F) E vs. E/M - HR=0.795 with 95%CI=(0.635, 0.995), E/M vs. M - HR=0.672 with 95%CI=(0.397, 1.137), E vs. M - HR=0.449 with 95%CI=(0.242, 0.832); (G) HR=0.566 with 95%CI=(0.328, 0.977); (H) HR=0.609 with 95%CI=(0.345, 1.078); (I) HR=1.047 with 95%CI=(0.445, 2.460); (J) E vs. E/M - HR=0.957 with 95%CI=(0.668, 1.371), E/M vs. M - HR=1.096 with 95%CI=(0.422, 2.845), E vs. M - HR=1.030 with 95%CI=(0.368, 2.885).

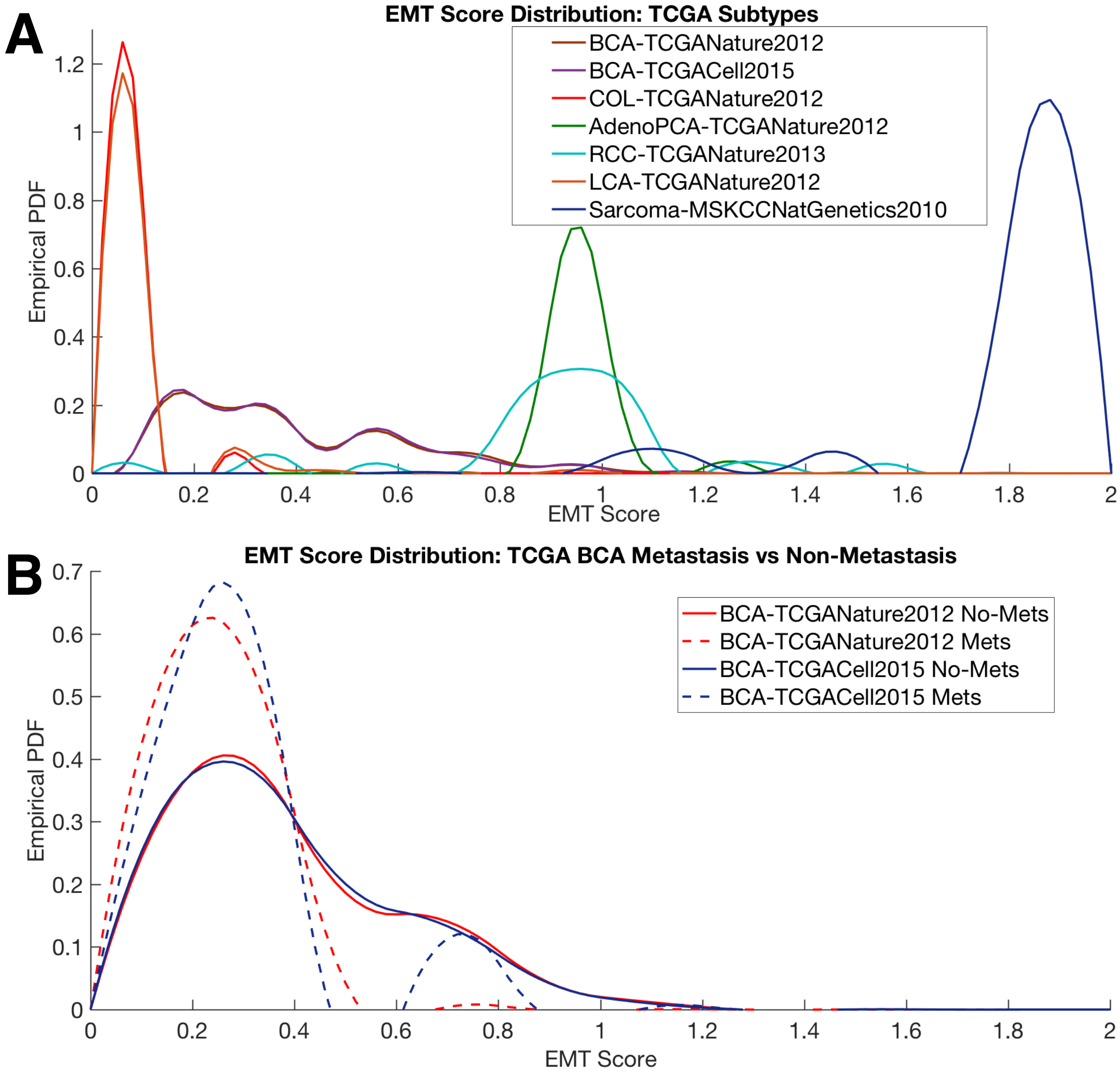


Figure S6: EMT spectrum distributions for large datasets.

(A) Distributions of EMT score for samples in multiple TCGA datasets belonging to different cancer types; (B) Spectrum of EMT score distributions for segregated metastatic and non-metastatic TCGA breast cancer samples.