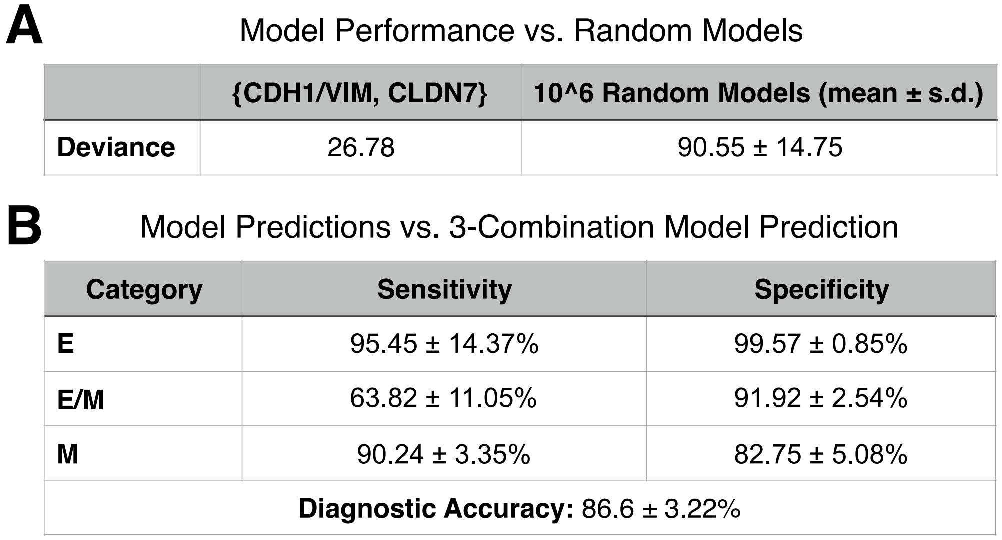
Supplementary Information: Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric

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.

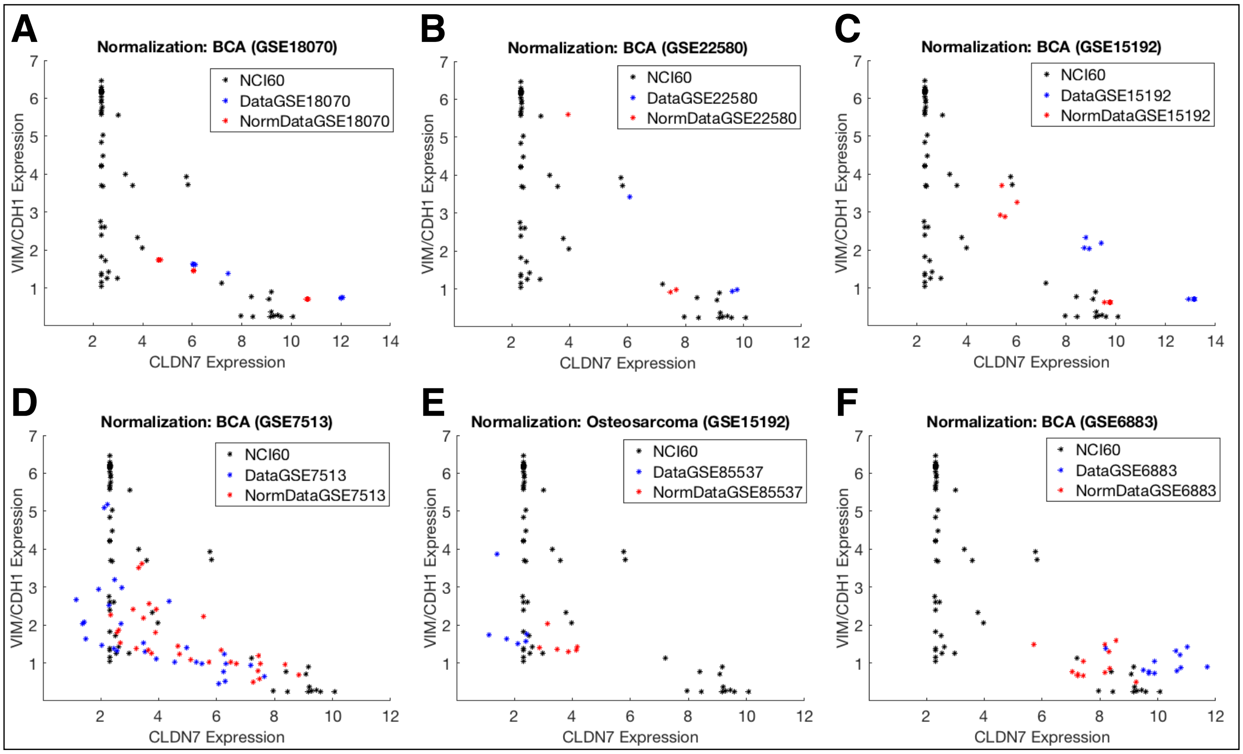


Figure S1: Additional examples of normalization.

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).

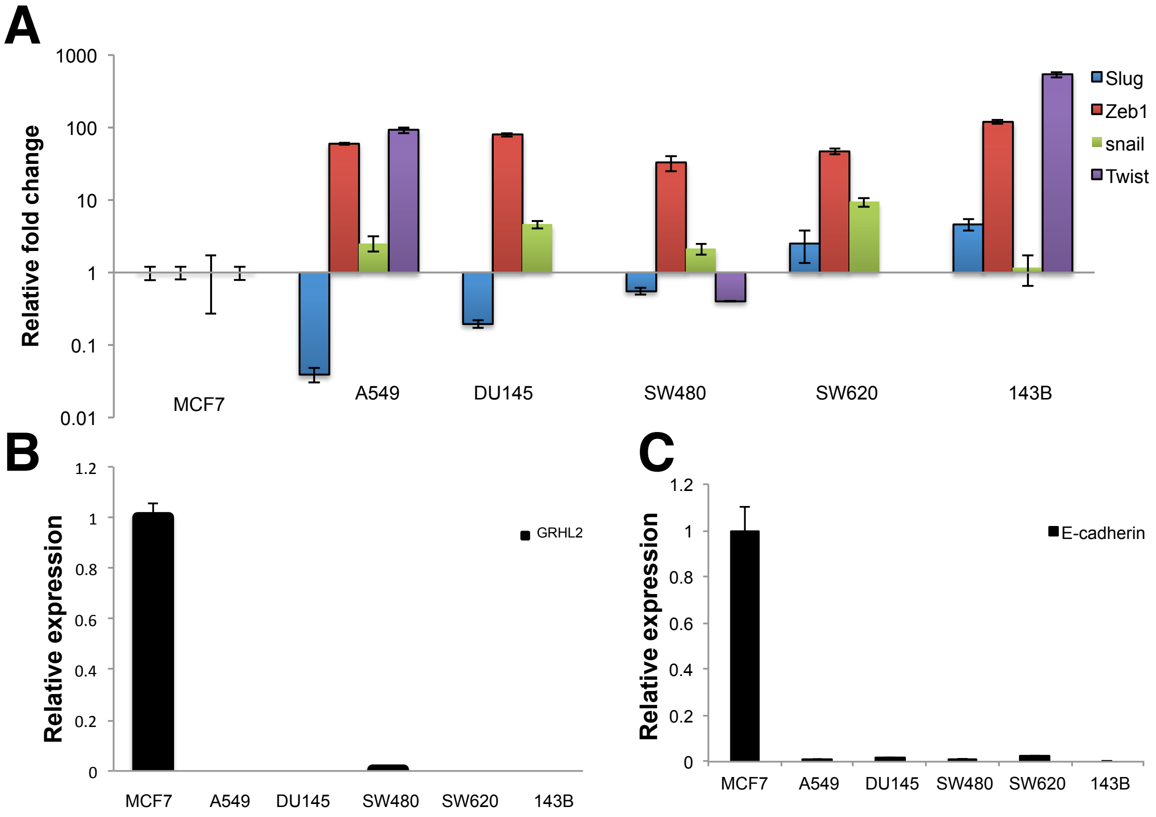
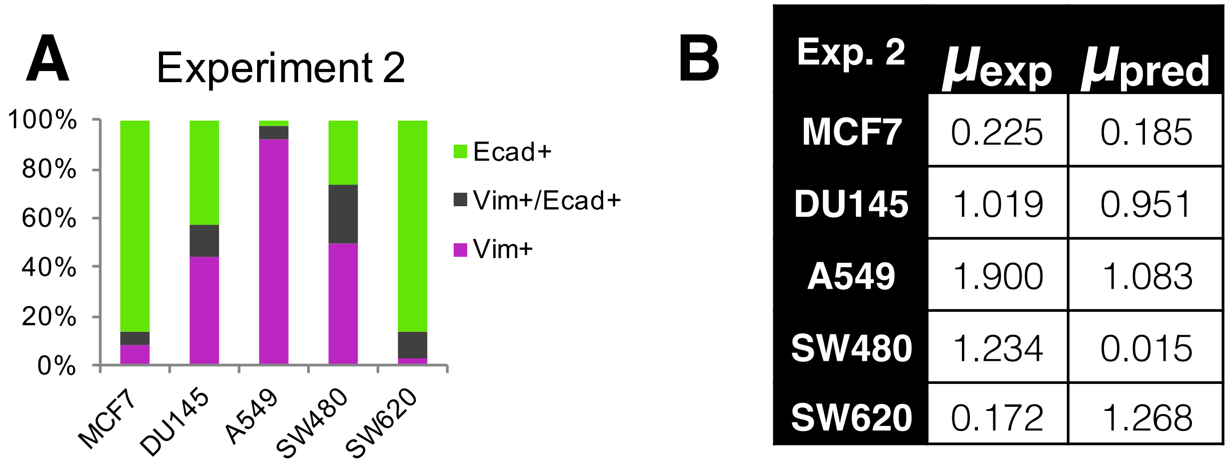


Figure S2: Levels of canonical epithelial and mesenchymal markers in multiple cell lines.

(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}).

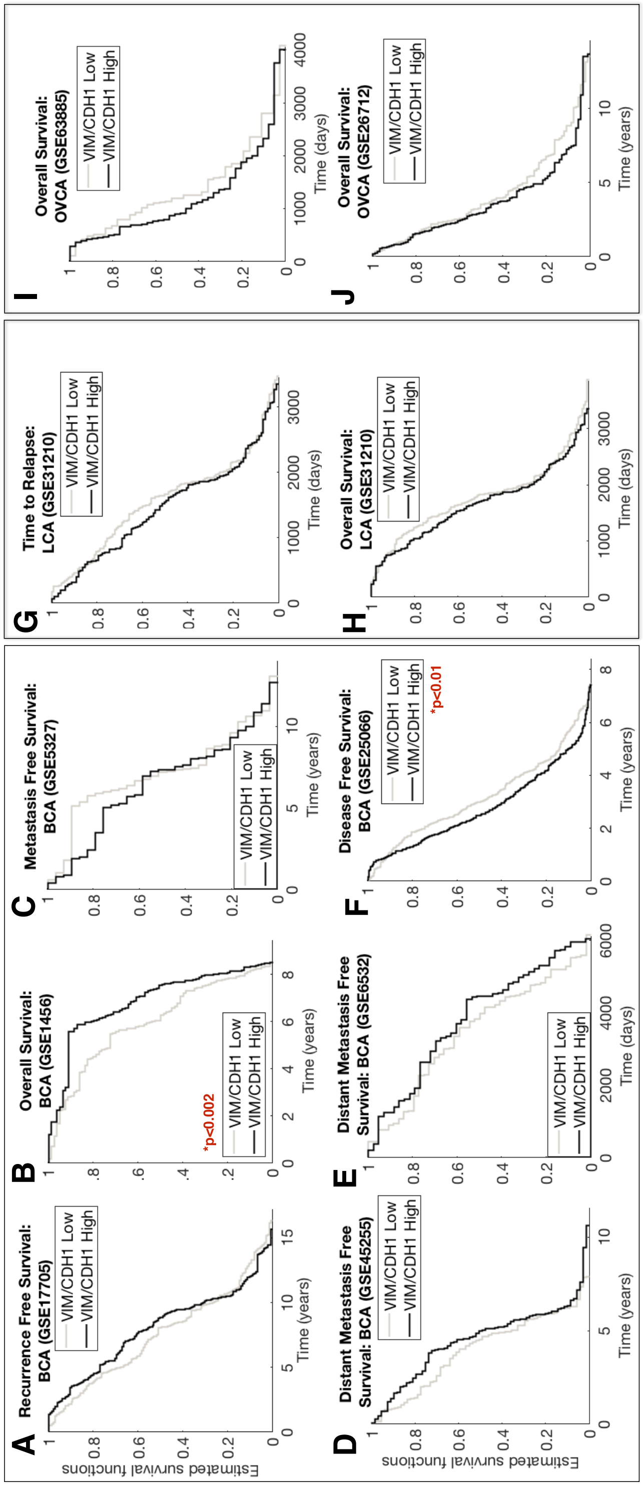

A
Recurrence Free Survival: BCA (GSE17705)
VIM/CDH1 Low
VIM/CDH1 High
Estimated survival functions
Time (years)
B
Overall Survival: BCA (GSE1456)
*p<0.002
VIM/CDH1 Low
VIM/CDH1 High
Time (years)
C
Metastasis Free Survival: BCA (GSE5327)
VIM/CDH1 Low
VIM/CDH1 High
Time (years)
D
Distant Metastasis Free Survival: BCA (GSE45255)
VIM/CDH1 Low
VIM/CDH1 High
Estimated survival functions
Time (years)
E
Distant Metastasis Free Survival: BCA (GSE6532)
VIM/CDH1 Low
VIM/CDH1 High
Time (days)
F
Disease Free Survival: BCA (GSE25066)
VIM/CDH1 Low
VIM/CDH1 High
*p<0.01
Time (years)
G
Time to Relapse: LCA (GSE31210)
VIM/CDH1 Low
VIM/CDH1 High
Time (days)
H
Overall Survival: LCA (GSE31210)
VIM/CDH1 Low
VIM/CDH1 High
Time (days)
I
Overall Survival: OVCA (GSE63885)
VIM/CDH1 Low
VIM/CDH1 High
Time (days)
J
Overall Survival: OVCA (GSE26712)
VIM/CDH1 Low
VIM/CDH1 High
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$^{\text{low}}$ or VIM/CDH1$^{\text{high}}$ 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).

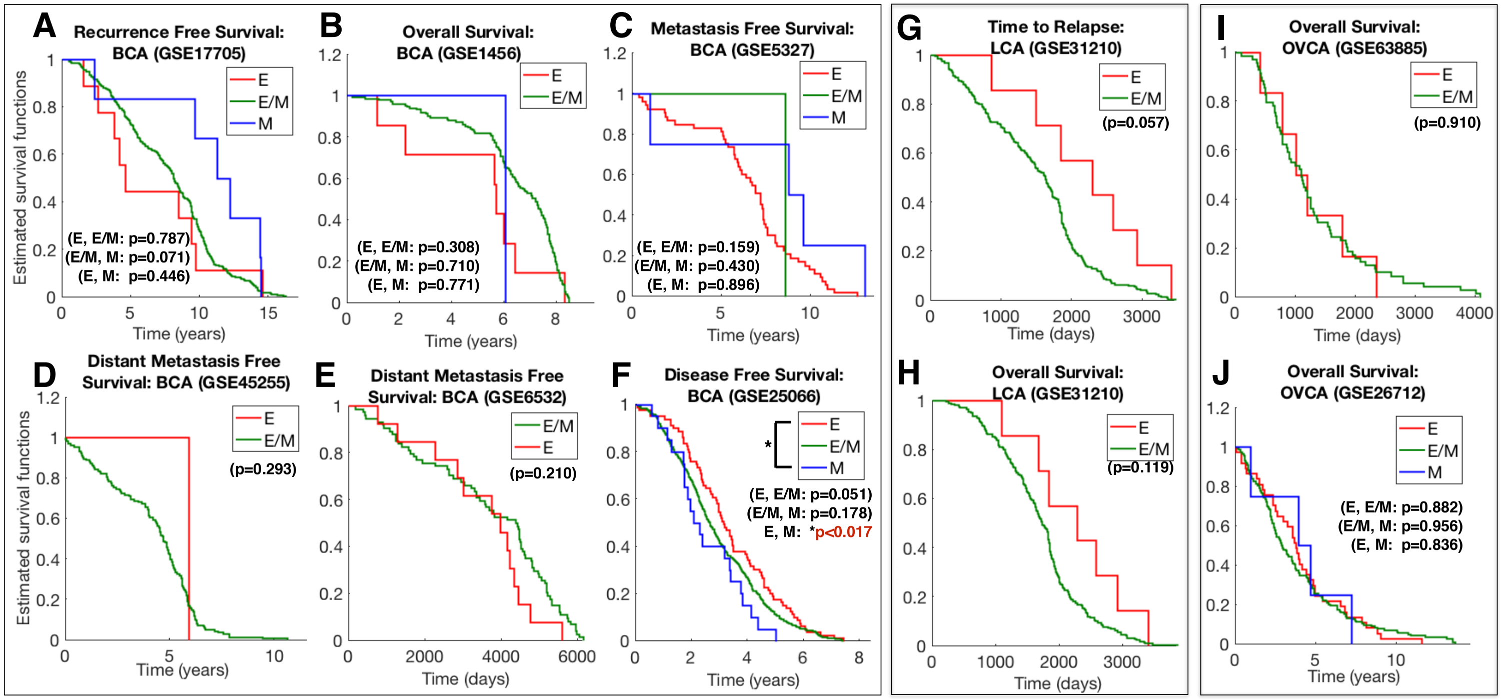

A
Recurrence Free Survival:
BCA (GSE17705)
Estimated survival functions
E
E/M
M
(E, E/M: p=0.787)
(E/M, M: p=0.071)
(E, M: p=0.446)
0
0.2
0.4
0.6
0.8
1
0
5
10
15
Time (years)
B
Overall Survival:
BCA (GSE1456)
E
E/M
(E, E/M: p=0.308)
(E/M, M: p=0.710)
(E, M: p=0.771)
0
0.2
0.4
0.6
0.8
1
1.2
0
2
4
6
8
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).

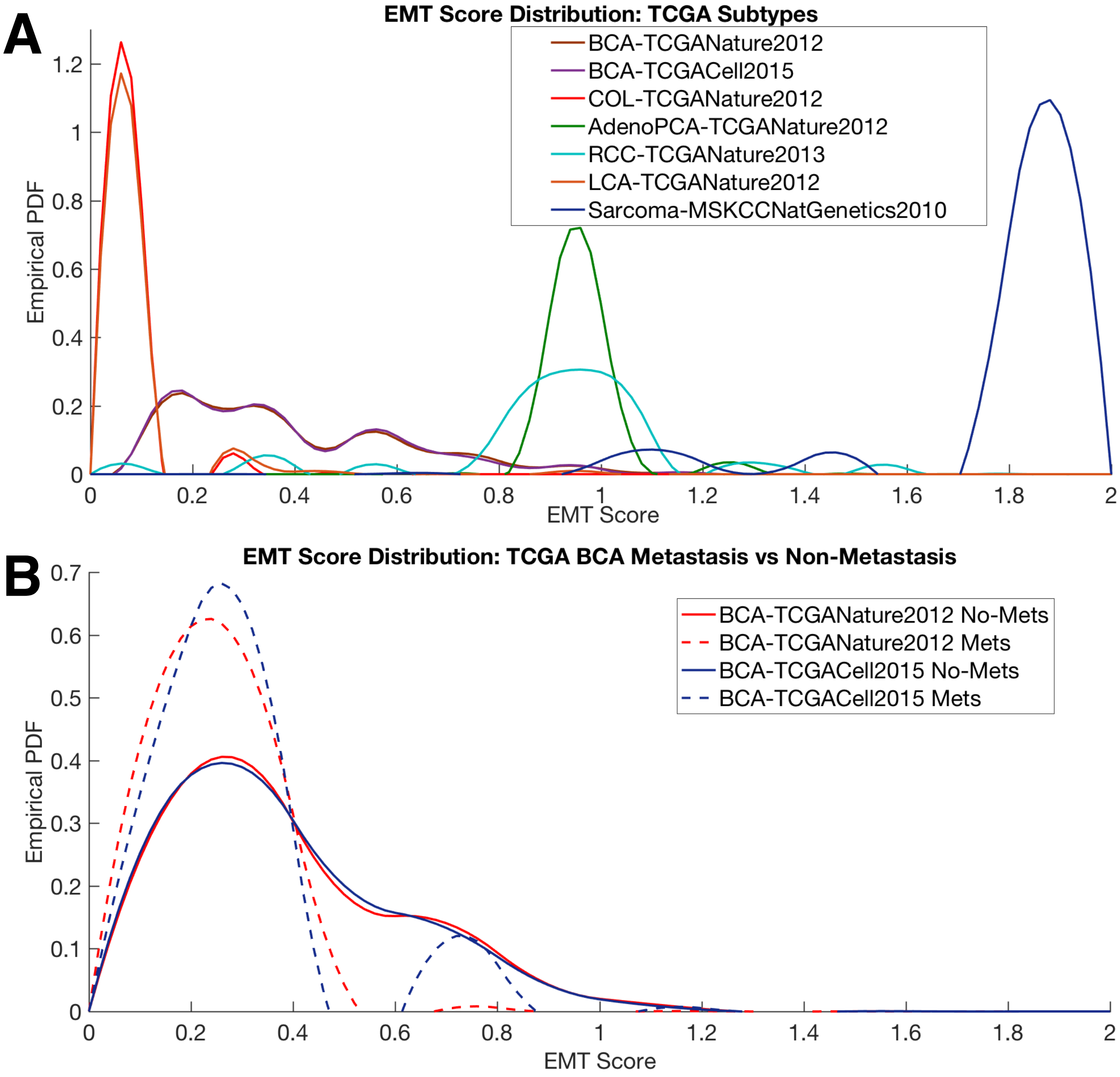


Figure S6: EMT spectrum distributions for large datasets.

(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.