An epithelial-mesenchymal-amoeboid transition gene signature reveals molecular subtypes of breast cancer progression and metastasis

Amin Emad[1], Tania Ray[2], Tor W. Jensen[3], Meera Parat[4], Rachael Natrajan[5], Saurabh Sinha[1,4], Partha S. Ray[2,6]

[1] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, USA

[2] Onconostic Technologies, Inc., Champaign, Illinois, USA

[3] Illinois Health Sciences Institute, University of Illinois at Urbana-Champaign, USA

[4] Department of Computer Science, University of Illinois at Urbana-Champaign, USA

[5] The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK

[6] HCG Oncology Cancer Centre, Surgical Oncology, Kolkata, India

Corresponding Authors:

Partha S. Ray

Onconostic Technologies, Inc.

60 Hazelwood Drive, Suite 208

Champaign, IL, USA 61820

Phone: (+1) 908-625-6906

Email: partha.ray@onconostictechnologies.com

Saurabh Sinha

2122 Siebel Center

201 N. Goodwin Ave

Urbana, IL, USA 61801

Phone: (+1) 217-333-3233

Email: sinhas@illinois.ed

ABSTRACT

Cancer cells within a tumor are known to display varying degrees of metastatic propensity but the molecular basis underlying such heterogeneity remains unclear. We analyzed genome-wide gene expression data obtained from primary tumors of lymph node-negative breast cancer patients using a novel metastasis biology-based Epithelial-Mesenchymal-Amoeboid Transition (EMAT) gene signature, and identified molecularly distinct subtypes associated with distinct prognostic profiles. EMAT subtype status improved prognosis accuracy of clinical parameters and statistically outperformed traditional breast cancer intrinsic subtypes. Additionally, analysis of 3D spheroids from an in vitro isogenic model of breast cancer progression reveals that EMAT subtypes display progression from premalignant to malignant and pre-invasive to invasive cancer. EMAT classification is a biologically informed method to assess metastasis risk in early stage, lymph node-negative breast cancer patients and appears to molecularly reconcile conflicting paradigms of breast cancer metastasis.

INTRODUCTION

Metastasis accounts for nearly 90% of cancer related mortality, and cancer cells within a tumor are known to possess different metastatic potentials[1]. However, the molecular basis underlying the observed heterogeneity in metastatic proclivity remains unclear and a suitable molecular classification is lacking. Intrinsic molecular subtypes of breast cancer have been associated with distinct metastatic predilections for one organ or the other, but are not necessarily associated with an increased metastatic propensity per se[2]. For instance, an intrinsic subtype that displays a higher rate of brain metastasis does not necessarily mean all patients, or even the majority of patients diagnosed with that subtype of cancer will go on to manifest with metastatic disease in the brain. Clearly other factors, independent of and in addition to those that determine the intrinsic molecular subtype of the cancer, influence its invasive potential and metastatic propensity.

Although implicated in cancer progression and metastasis, the clinical significance of processes like epithelial-to-mesenchymal transition (EMT) and mesenchymal-to-amoeboid transition (MAT) remains to be fully appreciated. EMT, a cellular transformation process that plays a key role in embryonic development, is widely considered to be one such factor influencing metastasis. Cancer cells derepress the normally silenced EMT molecular program, acquiring malignant properties that enable them to invade tissues surrounding their site of origin thereby effectively spreading and colonizing distant sites[3,4]. Likewise, MAT is another process that plays an important role in embryonic development and is similarly reawakened by cancers during the invasion-metastasis cascade[5,6].

Since the EMT process is exploited by cancer cells progressing to metastasis, there have been several attempts to subtype patient tumors based on an EMT signature, but these have not been successful in demonstrating a discernible difference in associated breast cancer prognosis[7-9]. Additionally, for cancer cells that have already transitioned through EMT but are facing microenvironmental (e.g. hypoxia) or xenobiotic (e.g. chemotherapy) stress, MAT may be an effective adaptive response to bypass the stress[10]. Indeed a recent report of effectively thwarting metastatic spread through simultaneous targeting of both mesenchymal and amoeboid motility in an animal model of cancer progression appears to support this notion[11]. We thus hypothesized that the true clinical and prognostic significance of EMT as a driving process in cancer progression towards distant metastasis cannot be fully appreciated unless it is considered in the context of being complemented by the conditional occurrence of MAT as well. Only when both processes are considered to coexist and undergo plastic interchange triggered by environmental pressures can the clinical significance of both be recognized and prognostic impact demonstrated. We therefore sought to develop a more inclusive gene expression signature that accurately captures EMT, MAT, and the variable dynamic co-occurrence of both the processes in the

same tumor. In this study our aims were to i) elucidate molecular subtypes in a primary tumor based on an EMT-MAT continuum that captures the heterogeneity of metastatic propensity and ii) to more comprehensively define biologically informed subtypes predictive of breast cancer metastasis and survival.

We constructed a gene signature (henceforth referred to as the EMAT signature) by combining a previously reported signature of EMT[7], derived from gene expression data of multiple distinct EMT-inducing cell perturbation experiments, with a newly generated MAT signature, derived following identical methodology to minimize derivation bias. We utilized this new EMAT gene signature to probe lymph node-negative human primary breast cancer gene expression datasets and were able to delineate informative subtypes of breast cancer that possess distinct molecular features, morphology/motility phenotypes and metastatic propensities. While EMT or MAT signatures on their own could not identify clusters of distinct prognosis, consideration of EMAT subtypes significantly improved survival prediction of standard clinical parameters and significantly outperformed prognosis accuracy of other known subtypes of breast cancer. Focusing on lymph node-negative samples allowed us to identify clusters that can predict clinical outcome at an early clinical stage of cancer. In addition, since cancer cells may utilize different mechanisms to promote metastasis before and after lymph node invasion, we studied lymph node-negative and positive samples separately and found the clinical characteristics of EMAT-defined subtypes to be largely similar, with some key differences. Furthermore, we identified transcription factors that may play key regulatory roles in establishing the newly identified EMAT subtypes and may be used as biomarkers for these subtypes. Finally, the identified subtypes were validated in a comparable but independent dataset and their enrichment was examined in an established and accepted *in vitro* cell line model of breast cancer progression, further confirming that EMAT subtypes reflect the progressive cellular transition from benign to malignant and from non-invasive to invasive

cancer. Our successful demonstration of prognostic stratification of breast cancer patients based on an EMAT signature opens up exciting avenues for future research into the mechanisms of the invasion/metastasis cascade and for potential clinical use of EMAT subtype status in guiding patient care.

RESULTS

An EMAT gene signature predicts clinical outcome in breast cancer patients

We constructed an "EMAT" gene signature comprising a list of 253 previously reported EMT-related genes[7] and 138 newly derived MAT-related genes obtained through analysis of publicly available gene expression data from multiple distinct MAT-inducing cell perturbation experiments[12] (see Methods and Supplementary Table S1). The analyzed dataset comprised mRNA gene expression profiles and clinical data corresponding to 562 lymph node-negative (LN) breast cancer primary tumors from the METABRIC study[13]. (We used lymph node negativity as a criterion to ensure that the samples were obtained during early clinical stages of the invasion-metastasis cascade.) We performed hierarchical clustering of samples (see Methods), represented by expression levels of EMAT genes (Figure 1A), evaluated resulting partitions with n = 3, 4, or 5 clusters, and found that n = 4 yields the best grouping of samples based on average silhouette scores (data not shown)[14]. This clustering provided a clear separation of the Kaplan-Meier survival curves (p = 2.42E-4, log rank test, Figure 1C), even though the clustering procedure was not privy to survival data. In addition, univariable and multivariable Cox regression analysis (when considering clinical parameters) showed that these clusters provide a statistically significant prognostic value (Supplementary Table S2). We refer to the four clusters, in decreasing order of their 10-year disease-specific survival (DSS) probabilities, as EMAT1, EMAT2, EMAT3 and EMAT4. Interestingly, clusters EMAT2 and EMAT3 have very similar Kaplan-Meier survival curves (Figure 1C), even though their gene expression profiles are distinct (indicated by orange bar in Figure 1A).

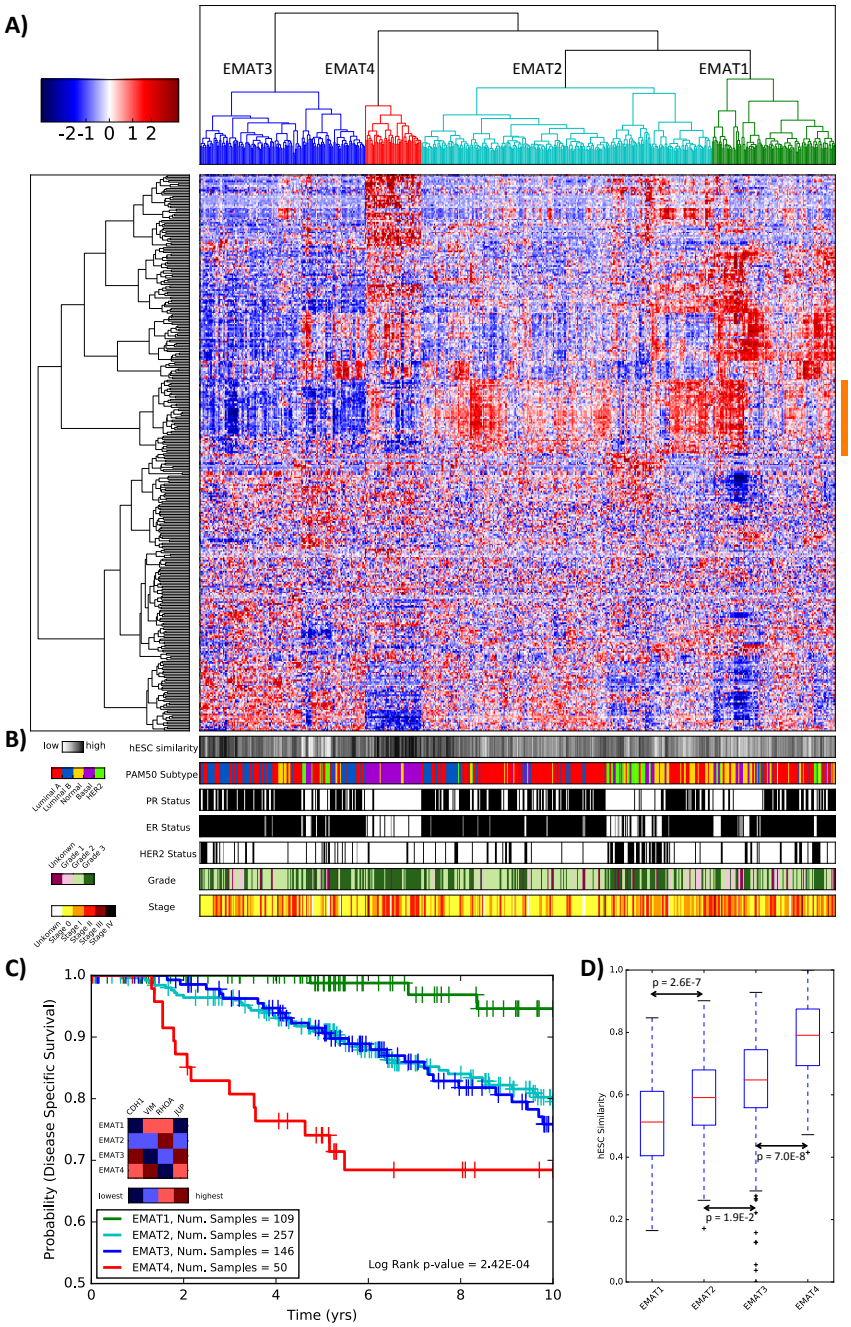

Figure 1: EMAT clusters and their characteristics. (A) EMAT clusters based on lymph node-negative METABRIC samples obtained using hierarchical clustering. The heatmap shows the normalized expression of EMAT genes (rows) in each sample (columns). Sample dendogram colors are chosen to match those of Kaplan-Meier plots in Figure 1C. (B) Characterization of samples based on similarity to hESC, PAM50 subtypes, ER, PR and HER2 status, stage, and grade. Spearman's rank correlation, scaled between 0 and 1 using min-max normalization, is used as the measure of similarity of samples to hESC, in which 0 and 1 represent least similar and most similar, respectively. (C) Kaplan-Meier plots corresponding to n = 4 clusters. The heatmap shows the relative ranking of the average expression of four biomarkers in each cluster compared to other clusters. (D) The box plots show the distribution of hESC similarity of the samples in each cluster. The similarity is defined as the Spearman's rank correlation (scaled between 0 and 1) between expression profiles of H1 hESC lines and each sample. The p-values (calculated using a one-sided t-test) show how significant the differences between two adjacent EMAT clusters are with respect to their similarity to hESC. The significance p-value for the cluster with the least similarity to hESC (EMAT1) and the cluster with the most similarity to hESC (EMAT4) is p = 1.7E-23.

To characterize each cluster obtained above, we examined the expression of CDH1, VIM, RHOA, and JUP, well-accepted biomarkers of Epithelial (E), Mesenchymal (M), Amoeboid (A), and collective cell migration morphologies in breast cancer[15], respectively. We assessed the expression of JUP as a collective cell migration marker because while EMT and MAT usually manifest as single cell migration, both epithelial and amoeboid collective cell migration have also been observed to contribute to the metastatic spread of cancer cells[16]. EMAT4, the cluster with the lowest DSS probability, showed M-like characteristics (VIM was over-expressed in this cluster compared to the other three clusters, $p = 4.8E-8$, unpaired two-tailed t-test). Cluster EMAT3, had the lowest average expression of VIM ($p = 3.7E-79$) and the highest average expression of CDH1 ($p = 1.3 E-3$) and JUP ($p = 8.3E-6$), suggesting that early onset of epithelial collective cell migration may be manifested in this group of patients. EMAT2 had the highest average expression of RHOA ($p = 3.8E-3$), suggesting A-like characteristics. Finally, EMAT1 had high expressions of VIM and RHOA but low expression of CDH1 and JUP, with the average expression of JUP being smallest ($p = 3.9E-15$) in this cluster. These results indicate existence of clusters having hybrid characteristics rather than discrete E-, M- and A- subtypes and emphasize the advantage of using the EMAT signature over using only E, M, or A biomarkers to distinguish groups of patient tumors associated with distinct prognosis. It is worth mentioning that genes most differentially expressed in each cluster (Bonferroni adjusted $p < 0.01$, t-test) included both EMT and MAT genes for all clusters (Supplementary Table S3), suggesting the involvement of both signatures in identification of these clusters.

Our finding that an EMAT signature allows grouping of patients with significantly different DSS is particularly interesting in light of previous reports where EMT gene signatures alone failed to do so[7,8]. We confirmed these previous observations in our dataset by clustering the tumor samples based on the expression of the EMT signature (subset of the EMAT signature above) using hierarchical clustering into two clusters, following the method of previous studies. We noted that biomarkers of Epithelial (CDH1)

and Mesenchymal (VIM) morphology were differentially expressed between these two clusters, with VIM being over-expressed in EMT1 ($p = 1.8E-69$) and CDH1 over-expressed in EMT2 ($p = 1.4E-2$). In spite of this distinction of biomarker expression, Kaplan-Meier analysis (Supplementary Figure S1A) did not show a significant difference between the clinical outcomes of these two groups ($p = 0.28$). We also clustered samples based on the expression of MAT genes, and noted differential expression of known biomarkers of Mesenchymal (VIM, $p = 2.7E-7$) and Amoeboid (RHOA, $p = 6.4E-4$) morphology. However, Kaplan-Meier analysis (Supplementary Figure S1B) did not show a significant difference between the clinical outcomes of these two groups ($p = 0.56$). It would thus appear that while both EMT and MAT processes have been demonstrated to play contributory roles in the biology of metastasis, when considered individually they cannot completely capture the diversity that exists in the metastatic propensity of breast tumors, as evidenced by their lack of individual prognostic value.

EMAT clusters provide prognostic information not present in clinical parameters or intrinsic subtypes

Next, we examined whether the EMAT clusters are associated with clinical parameters or previously described breast cancer subtypes (Figure 1B). Of note is the visually apparent enrichment of EMAT4, the cluster with worst survival, with triple negative (negative for ER, PR, HER2) and basal-like (PAM50 subtype[17], purple) patients. In addition, most HER2 positive patients (green) appear in EMAT2. Similarity to human embryonic stem cells (hESC) has been previously shown to be associated with elevated metastasis risk[18]. Figure 1D shows that EMAT clusters display worsening prognosis proportionate to their degree of similarity to hESC, consistent with this theory. This was despite the fact that derivation of the EMAT gene signature was not designed to intentionally enrich for stem cell traits. This provides further evidence in support of the EMAT clusters representing a progressive transition from less stem-like to more stem-like cell states, and less invasive to more invasive modes of cancer.

To assess the quantitative significance of the above associations, we computed the enrichment p-value (hypergeometric test) of each EMAT cluster with respect to tumor size, PAM50 subtypes, and receptor status (Figure 2). While the presence of small tumors in clusters with good prognosis is expected, the enrichment of EMAT1 (the cluster with the best prognosis) in large tumors is particularly interesting, as it suggests that large tumors do not necessarily result in poor survival in the absence of necessary metastatic mechanisms[19]. Figure 2B shows that EMAT1 was moderately enriched in the 'normal-like' PAM50 subtype (p = 5.59E-3), EMAT2 in Luminal A (p = 1.84e-06), EMAT3 in Luminal B (7.34E-12), and EMAT4 in basal-like subtype (p = 7.73E-41). Finally, EMAT3 was enriched in ER positive (p = 1.76E-06) and PR positive (p = 2E-06) samples, while EMAT4 was enriched in ER negative (p = 2.64E-28) and PR negative (p = 4.18E-18) samples (see Table 1 for a summary).

Table 1: A summary of the characteristics of the EMAT clusters obtained using lymph node-negative breast cancer patients from the METABRIC study. In this table, P stands for positive and N for negative. EMAT1, the cluster with the best prognosis, is the least similar to hESC, is enriched in normal-like PAM50 subtype of breast cancer and large tumors. This cluster shows M-like and A-like characteristics. EMAT2, the cluster with a relatively good prognosis, has little similarity to hESC, is enriched in HER2 and Luminal A subtypes and small tumors. This cluster shows A-like characteristics. EMAT3, the cluster with a relatively poor prognosis, has a high degree of similarity to hESC, is enriched in Luminal B samples and medium and large tumors. This cluster shows E-like and collective cell migration characteristics. EMAT4, the cluster with the worst prognosis, shows the highest degree of similarity to hESC, is enriched in the basal-like subtype of breast cancer and medium-sized tumors. This cluster shows a mix of E-like, M-like and collective cell migration characteristics.

Cluster	Prognosis	hESC Similarity	PAM50 Enrichment	Tumor Size	ER	PR	HER2	CDH1	VIM	RHOA	JUP
EMAT1	Best	Lowest	Normal-like	Large	Mix	Mix	N	Low	High	High	Low
EMAT2	Good	Low	Luminal A /Her2	Small	P	P	Mix	Low	Low	High	Low
EMAT3	Poor	High	Luminal B	Medium /Large	P	P	N	High	Low	Low	High
EMAT4	Worst	Highest	Basal	Medium	N	N	N	High	High	Low	High

Given the observation that EMAT clusters are enriched in intrinsic subtypes defined based on PAM50 and receptor status, we sought to determine whether EMAT clustering could provide prognostic information beyond what is captured by the aforementioned subtypes. To this end, we trained a Cox regression model on the LN samples from the METABRIC dataset using four types of predictors: 1) clinical parameters, including age and tumor size, 2) clinical parameters and receptor status, 3) clinical

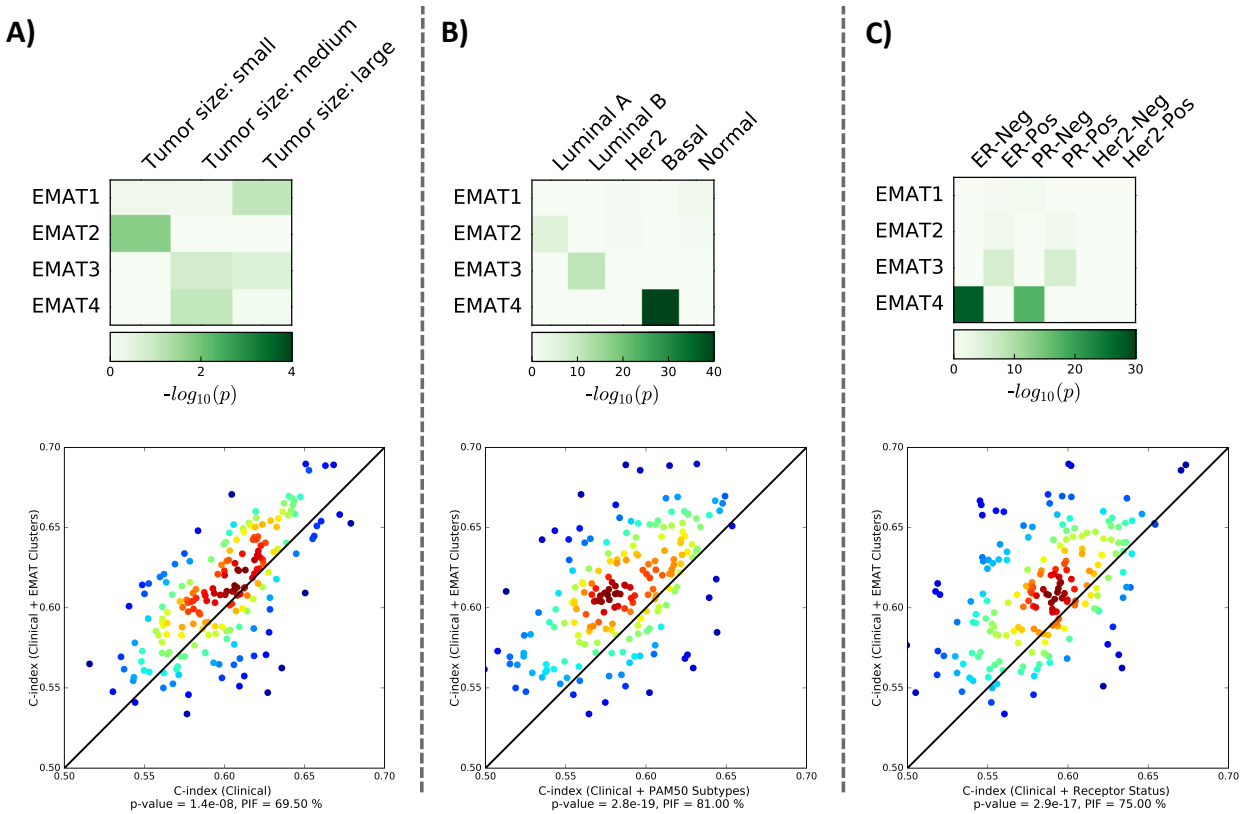

Figure 2: Enrichment of EMAT clusters in other breast cancer subtypes and systematic comparison of their prognostic power using cross-validation. The heatmaps show the -log10 (p-value) of enrichment of EMAT clusters in different subtypes or clinical parameters (using a hypergeometric test). The scatter plots compare the performance (measured in C-index) of Cox regression predictions using EMAT cluster status and clinical parameters (y-axis) versus other predictors (x-axis). Here, half of the samples were randomly selected as the training set, and grouped into n = 4 clusters using the EMAT signature. A Cox regression model was trained on these samples, using clinical variables as well as EMAT cluster status. In parallel, Cox regression models were trained on the same samples using the three other types of predictors. Each trained model was then used to estimate the expected survival of the remaining samples (i.e., test samples). In this step, EMAT cluster status was assigned to test samples using a centroid-based classifier trained on the training samples. The p-values were calculated using a one-sided Wilcoxon signed rank test and represent the significance of the improvement obtained using EMAT cluster status and clinical parameters as compared to other predictors. The measure PIF shows the percent of times in which EMAT + clinical parameters provided a more accurate prediction compared to the baseline. (A) The heatmap shows enrichment of EMAT clusters by samples of different tumor sizes. The scatter plot shows performance of Cox regression predictions using EMAT + clinical parameters versus clinical parameters alone. (B) The heatmap shows enrichment of EMAT clusters by samples of different PAM50 molecular subtypes. The scatter plot shows performance of Cox regression predictions using EMAT + clinical parameters versus PAM50 subtypes + clinical parameters. (C) The heatmap shows enrichment of EMAT clusters by samples of different receptor status. The scatter plot shows performance of Cox regression predictions using EMAT + clinical parameters versus receptor status + clinical parameters.

parameters and PAM50 subtype status, and 4) clinical parameters and EMAT cluster status. The Cox regression model trained on clinical and EMAT cluster status had the smallest p-value using a likelihood ratio test (p = 1.4E-5 compared to p = 3.9E-3 for clinical parameters and receptor status, p = 5.7E-3 for clinical parameters and PAM50 subtypes, and p = 6.9E-4 for clinical parameters only). This was a

promising result, but not an entirely convincing comparison since EMAT cluster status was defined based on samples used in the Cox regression test, while the other subtypes were defined *a priori*.

To rigorously compare the predictive ability of the above classes of predictors while removing the effect of the varying number of predicting features, and also to test the generalizability of these models on unseen data, we next used a cross-validation framework. In this framework, the samples were randomly divided in two groups of (almost) equal size and a Cox regression model trained on one half was used to estimate the expected survival on the other half. This process was repeated 200 times, each time using a distinct random partition of data. The Cox regression model using EMAT labels along with clinical parameters provided the best predictions (Figure 2 bottom panels), evaluated using a one-sided Wilcoxon signed rank test on paired C-index values of the compared methods as well as another measure called percentage of improved folds (PIF)[20] defined as percent of the partitions in which one class of features outperforms another class. These results indicate that although some of the EMAT clusters share characteristics with known intrinsic subtypes of breast cancer, they are better predictors of patient survival outcome than PAM50 or receptor-based subtypes.

Cross-dataset analysis shows similar survival behavior of the EMAT clusters

We next evaluated the prognostic power of the EMAT clusters on an independent dataset. To this end, we developed a subtyping procedure using a centroid-based classifier, trained on the gene expression profiles and EMAT cluster status of the METABRIC LN samples, to assign EMAT "subtypes" to any new dataset (see Methods). We obtained gene expression and clinical data corresponding to lymph node-negative breast cancer primary tumors from[21] (GEO accession number: GSE11121) and determined their EMAT subtypes. One advantage of this dataset is that it contains distant metastasis-free survival (DMFS) information, which allows us to specifically study the ability of EMAT clusters to predict the occurrence of metastasis. As can be seen in Figure 3A, the separation among Kaplan-Meier curves for the subtypes

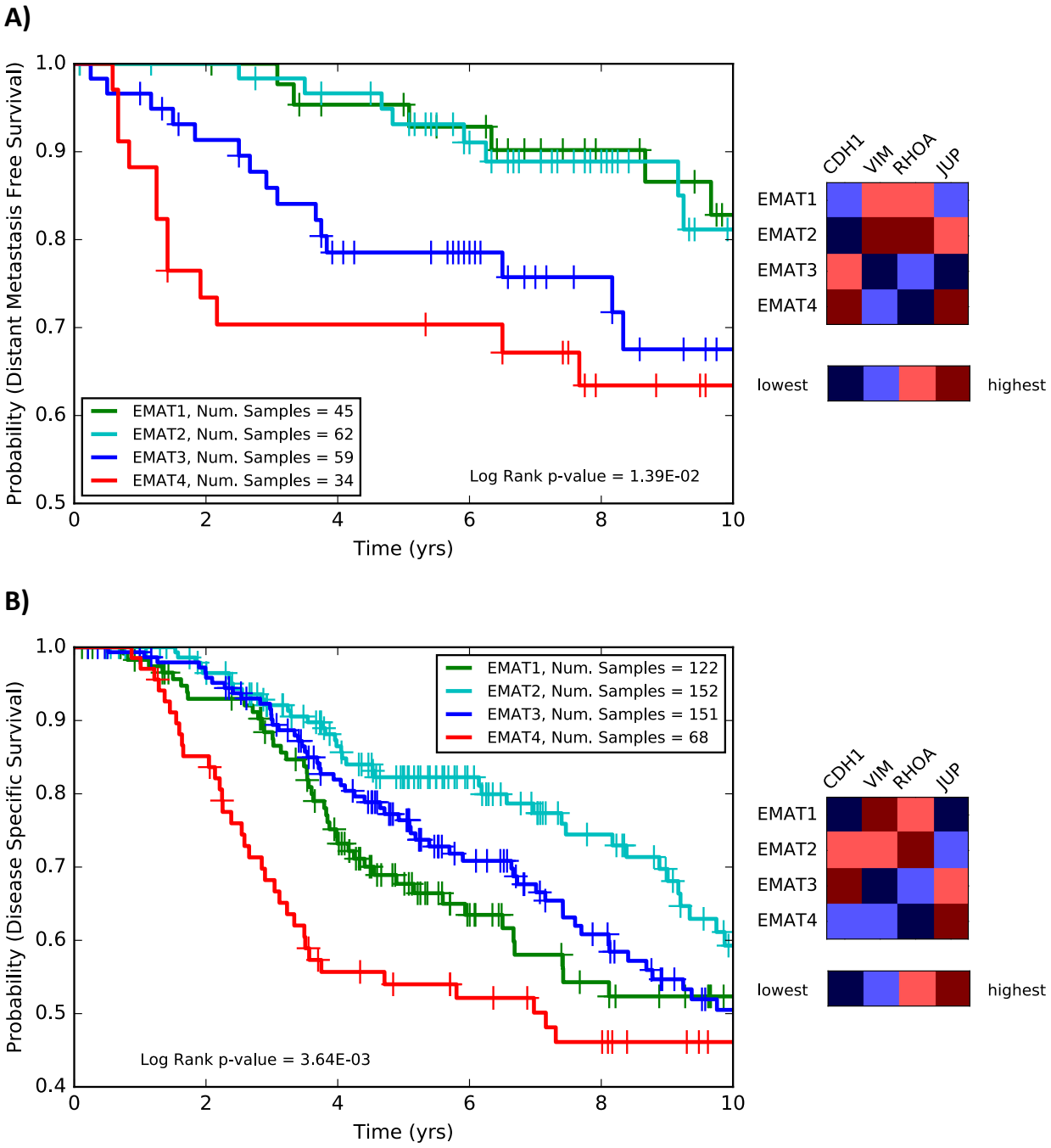

Figure 3: Cross-dataset analysis. A centroid-based classifier trained on LN METABRIC samples is used to assign EMAT subtype labels to each sample. (A) The Kaplan-Meier survival plots and biomarker status for EMAT subtypes of LN breast cancer samples from the GSE11121 dataset. Expression of the four biomarkers generally follows a similar trend as in Figure 1C; however, VIM is no longer highly expressed in EMAT4, but instead has the highest expression in EMAT2. In addition, expression of JUP is lowest in EMAT3, while it was highest in this cluster in Figure 1C. (B) The Kaplan-Meier survival plots and biomarker status for EMAT clusters of LP breast cancer samples from the METABRIC dataset. The expression of biomarkers is largely consistent with LN samples. Two exceptions are the higher expression of VIM and CDH1 in EMAT2 and lower expression of VIM in EMAT4 in LP samples compared to LN samples in Figure 1C.

is statistically significant (p = 1.39E-2, log rank test). In addition univariable and multivariable Cox regression analysis (when considering clinical parameters) showed that these clusters provide a statistically significant prognostic value in this independent dataset as well (Supplementary Table S4). It is interesting to note that the survival behavior of EMAT subtypes remains largely similar to their behavior in the LN METABRIC dataset, with EMAT4 having the worst survival, EMAT3 the second worst survival and EMAT1 and EMAT2 having the best survival probabilities. (We also tried a 5-nearest neighbor classifier for EMAT subtype status assignment which resulted in clusters with survival behaviors better matching those of Figure 1C; see Supplementary Figure S2.) Expression of the four biomarkers in GSE11121 samples assigned to each EMAT subtype generally follows a similar trend as in Figure 1C. These results show a high concordance between the characteristics of the EMAT subtypes in two independent studies.

Up until now, by focusing on the LN samples, our analysis presumably reflected the metastatic mechanisms active prior to lymph node invasion. However, samples with similar gene expression patterns may have different clinical outcome conditioned on their lymph node status, as cells that have already acquired invasive characteristics may rely on different mechanisms to promote metastasis. To evaluate this, we obtained clinical and gene expression data corresponding to 493 lymph node-positive (LP) samples from the METABRIC study, and used our subtyping procedure to assign EMAT subtype designations to each sample. Kaplan-Meier analysis (Figure 3B) shows a significant distinction among survival probabilities of each EMAT subtype in LP patients (p = 3.64E-3). In addition, univariable and multivariable Cox regression analysis (when considering clinical parameters) showed that these subtypes provide a statistically significant prognostic value (Supplementary Table S2). Similar analyses on all samples (LN and LP together) also showed that the EMAT subtypes provide significant prognostic value in addition to lymph node status (Supplementary Table S2).

The main difference between the results obtained using the LP and LN samples is in the survival characteristics of EMAT1: while this subtype had the best survival probability in LN patients, it exhibits poor prognosis (second worst) in LP samples. However, the relative behaviors of EMAT2, EMAT3, and EMAT4 samples (Figure 3B) mimic those of their LN counterparts, in terms of their relative survival probabilities (Figure 1C) and biomarker characteristics (Figure 1C, inset). In summary, our analysis suggests that the clinical and molecular characteristics of the subtypes are preserved between LN and LP samples, with the exception of EMAT1: the EMAT1 phenotype when manifested prior to lymph node invasion, represents a less aggressive mode of cancer, but at later stages the same phenotype represents an aggressive mode, although biomarker expressions, enrichment in breast cancer subtypes, and resemblance to hESCs was similar in both LN and LP groups (Supplementary Figure S3). In the absence of nodal involvement, EMAT1 tumors appear to lack metastatic propensity, while the metastatic propensity of EMAT2, EMAT3 or EMAT4 tumors is manifested even in lymph node-negative patients. Ostensibly, this suggests that in tumors classified under EMAT1 subtype, nodal involvement is not only predictive of an increased risk of metastatic dissemination, but in this case is also likely to be a necessary prerequisite for distant metastasis.

Identification of transcription factors associated with EMAT clusters

We evaluated the expression of 1,338 transcription factors (TFs) present in the METABRIC dataset. For this purpose, we used a t-test (corrected for multiple hypothesis testing) to identify TFs that are differentially expressed in one cluster compared to others, in LN samples (Supplementary Table S5). Next, for each cluster we identified two TFs that were most over-expressed and under-expressed (total of 8 TFs). These TFs include PPP1R13L and MNDA (EMAT1), ETV7 and TSHZ3 (EMAT2), SLUG and AFF3 (EMAT3), and FOXA1 and FOXC1 (EMAT4), which were under-expressed and over-expressed in each

cluster, respectively, indicating a link with progression and invasion mechanisms (see Discussion). Hierarchical clustering of LN samples from the METABRIC dataset using the eight identified TFs had a very high concordance with clusters obtained using the full EMAT gene signature. In addition, Kaplan-Meier analysis using the clusters based on these TFs demonstrated a significant separation of survival curves (Figure 4). These results further support the key role of these TFs in potentially regulating the distinct expression profiles of EMAT subtypes and lend support to their clinical utility as potential biomarkers of the identified EMAT clusters.

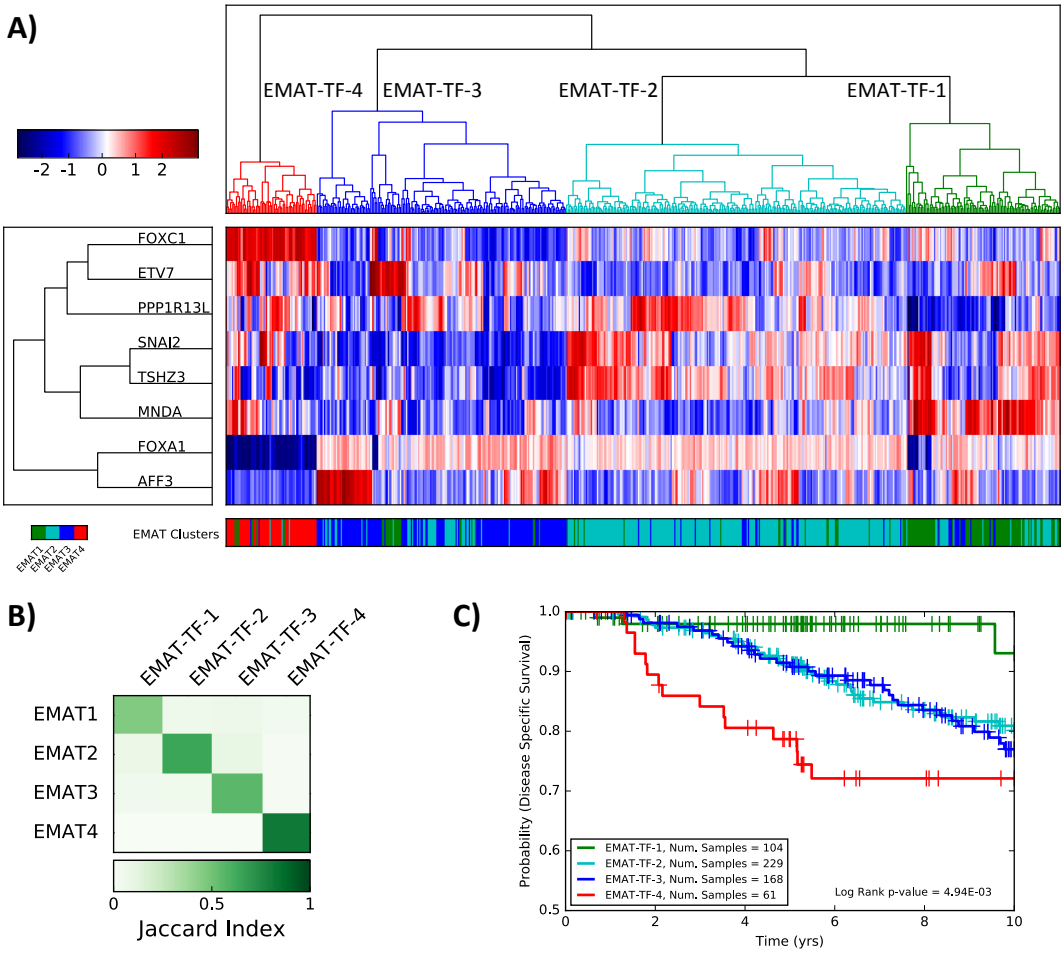

Figure 4: Analysis of clusters obtained using eight TFs most under-expressed or over-expressed in each EMAT cluster. (A) Hierarchical clustering based on the expression of the eight identified TFs is used to cluster samples into four groups. The color bar at the bottom shows the true EMAT cluster label of each sample. (B) Concordance of clusters obtained using eight TFs with EMAT clusters based on the Jaccard index. (C) Kaplan-Meier survival plots for clusters obtained using the TFs.

EMAT subtypes reflect a progressive acquisition of malignancy and invasiveness

We next assessed the EMAT signature in the well-characterized isogenic MCF10 model of breast cancer progression which was produced by consecutive *in vivo* implantation of HRAS transformed MCF10A cells in immune-deficient mice followed by *in vitro* cell culture[22]. RNA-Seq data corresponding to these seven cell lines grown in three-dimensional (3D) spheroid culture systems[23] were analyzed. For our analysis, we considered the cell lines in three groups – normal, benign and premalignant cells (MCF10A, MCF10AneoT, MCF10AT1), pre-invasive *in situ* carcinoma cells (MCF10DCIS.com), and invasive carcinoma cells (MCF10CA1a, MCF10CA1d, and MCF10CA1h) – and assessed their similarity to LN METABRIC EMAT clusters. For this purpose, we used the expression of the TFs identified above, except for MNDA and TSHZ3 which were not expressed in the cell lines. Figure 5 shows similarity between the centroids of each of the three cell line groups and those of the EMAT clusters, measured by pairwise Spearman's rank correlation. Normal and pre-malignant cells were most similar to EMAT2, a cluster with relatively good survival compared to other clusters in both LN and LP samples. The non-invasive *in situ* cell line is most similar to EMAT3 cluster and the group containing invasive carcinoma cells is most similar to EMAT4, the cluster with the worst survival. Notably, the group containing invasive cell lines also has a high similarity to EMAT1 cluster. This supports the suggestion we made above, in the context of LP samples (Figure 3B), that the EMAT1 phenotype when manifested in later clinical stages of the invasion-metastasis cascade, i.e. post nodal involvement, denotes aggressive biology. Overall, these results support the hypothesis that the EMAT clusters identified *in vivo* represent a gradual progression of cancer cell transition from non-invasive to fully invasive states capable of metastasis. In addition, the observed association between the EMAT clusters with cell states of cancer progression in this model made intuitive sense with the progressively worsening prognosis of such EMAT clusters, as was observed in Figures 1C, 3A and 3B.

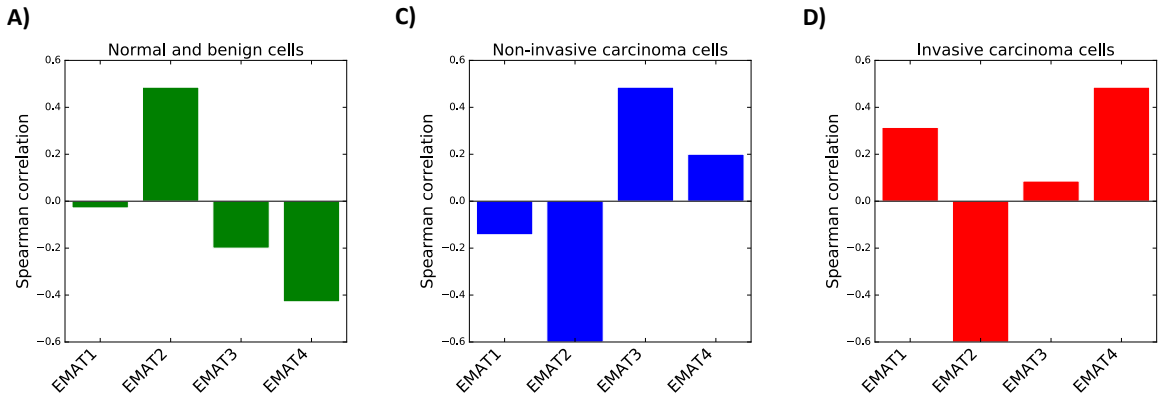

Figure 5: Analysis of MCF10 cell line series. (A) Spearman's rank correlation of the centroid of normal and benign cells and EMAT clusters using expression of identified TFs. (B) Spearman's rank correlation of the non-invasive carcinoma cells and EMAT clusters using expression of identified TFs. (C) Spearman's rank correlation of the centroid of invasive carcinoma cells and EMAT clusters using expression of identified TFs.

DISCUSSION

Prior attempts at utilizing EMT gene expression signatures of primary tumors to predict future metastasis have largely approached the problem as being able to define either one of two possible binary states, thereby treating metastasis as an "all-or-nothing" phenomenon[7,8]. In these studies, it was assumed that EMT exists in either a turned "ON" or turned "OFF" state: this we thought might be too simplistic to discern the true survival impact of a complex tightly regulated biologic program like EMT, that most likely exists along an entire spectrum ranging from depressed states to elevated states. In our attempt to derive a gene expression signature based on metastasis biology that could classify this complex and dynamic process, we postulated that the heterogeneity that exists in the metastatic propensity of tumors might be better captured by considering the existence of multiple groups or clusters of patient tumors, each with characteristic molecular features and distinct prognostic profiles.

Of note, the clinical significance of EMT in driving metastasis has been questioned in recent publications that demonstrate the transient nature of EMT[24] and dispute the necessity of EMT occurrence in mediating metastasis[25-27]. While the findings in such studies have been challenged to be inconclusive[28,29],

there is another possible explanation for the reported findings without upending the potential significance of EMT in metastasis. One could speculate that when the cancer cell was proceeding along EMT, an increase in the entropy of an EMT path of cellular evolution could have resulted due to the cell being subjected to either microenvironmental pressures (e.g., hypoxia) or xenobiotic exposure (e.g., chemotherapy or EMT-targeted biologic therapy). A cancer cell of epithelial origin that had already undergone EMT when faced with such a situation could then undergo plastic transformation by a process such as MAT and adopt an alternate amoeboid motility program allowing it to proceed to metastasis via a lower entropy alternative route as recently reported[10]. Indeed simultaneous targeting of both mesenchymal and amoeboid motility in an animal model of cancer progression effectively arrests metastatic spread[11]. Only when both processes are considered together to contribute to multiple distinct molecular states of progressively increasing metastatic propensity does the prognostic value of such distinct phenotypes become demonstrable.

In this study, we utilized an EMAT gene expression signature and clustered breast cancer patients into four different groups with distinct prognosis, morphology, migratory behavior and molecular characteristics. Our results revealed the existence of subtypes of hybrid characteristics rather than discrete E-, M-, and A-like clusters, emphasizing the advantage of using the EMAT signature over using only E, M, or A biomarkers, to distinguish groups of patient tumors associated with distinct prognosis. We showed that EMAT subtypes have varying similarity to hESC, EMAT subtypes with worse prognosis being more similar to hESC. Our findings suggest that during the invasion–metastasis cascade in cancer, both EMT and MAT contribute to an EMAT continuum of reverse evolution from a multicellular differentiated state towards a more primitive unicellular dedifferentiated state resembling embryonic stem cells[30]. In addition, using *in vitro* cell line analysis, we showed that EMAT subtypes reflect a gradual and progressive acquisition of malignancy and invasiveness. The expression patterns of biomarkers of

different modes of cell migration were generally consistent across different datasets. While in some subtypes one distinct morphology/motility mode was manifested, for other subtypes a hybrid of several morphologies was observed. Collectively, these results support the notion that several mechanisms are in effect simultaneously to promote metastasis, even among groups of individuals that share a similar transcriptomic profile. In addition, we showed that collective cell migration besides single cell migration is also a very important factor in determining clinical outcome, as individuals that showed this mode of cell migration consistently exhibited poor clinical outcome.

In order to uncover the regulatory mechanisms underlying the discovered EMAT clusters, we identified transcriptional regulators characteristic of each EMAT cluster. In addition to confirming previously known factors, new factors involved in metastasis biology were also revealed. A known inhibitor of TP53, PPP1R13L is involved in regulation of apoptosis through NF-kappa-B and TP53 proteins and has been associated with lymph node metastasis in endometrial endometrioid adenocarcinoma. In addition, the over-expression of this gene has been shown to enhance tumorigenesis and invasion through both TP53-dependent and TP53-independent mechanisms[31]. MNDA has been shown to be upregulated in tumorigenic breast cell lines compared to control[32], and has been suggested as a master regulator in primary breast cancer[33]. In a recent study, this TF has been suggested to be a tumor suppressor and novel therapeutic target in human osteosarcoma[34]. ETV7 is a member of the ETS family, known to be involved in cancer initiation, progression and metastasis[35]. This TF has been shown to inhibit cell migration and invasion *in vitro* and suppress metastasis *in vivo* in nasopharyngeal carcinoma[36]. TSHZ3, a member of the T-shirt zinc-finger family, has been suggested as a candidate tumor suppressor in breast and prostate cancers and its expression is downregulated in breast cancer cell lines[37]. SLUG is a well-known EMT transcription factor[7], which represses E-cadherin (CDH1) expression in breast carcinoma. The role of this TF in metastasis is well documented in many cancers[38,39]. AFF3 has been shown to

mediate the oncogenic effects of β-catenin in adrenocortical carcinoma (ACC) whose silencing has been associated with a decrease in cell proliferation and an increase in apoptosis[40]. Finally, FOXC1 and FOXA1, the two TFs representing the cluster with worst prognosis (EMAT4), belong to the forkhead family of transcription factors. The roles of FOXC1 as an EMT driver and the opposing role of FOXA1 as an EMT suppressor are well established[41-45].

Our understanding of breast cancer progression and metastasis has largely been defined by three paradigms proposed over the past century. The first explanation conforms to the anatomic Tumor-Node-Metastasis model on which cancer staging is based. It proposed that breast cancer progression is an orderly process of contiguous spread, from primary tumor site, to regional lymph nodes via lymphatics, on to distant metastatic sites, and thus the anatomically precise removal of any and all demonstrable loco-regional disease, in the form of radical mastectomy, would favorably alter clinical outcome. This has come to be known as Halsted's paradigm[46]. The second explanation, referred to as Fisher's paradigm[47], was triggered by results of clinical trials proving the non-superiority of radical mastectomy, compared to lumpectomy and radiation. It proposed that breast cancer cells can spread and metastasize in a non-contiguous manner, wherein tumor biology trumps the anatomical extent of disease and small tumors are only a local manifestation of disease that has already disseminated. Hence their surgical removal would not favorably alter clinical outcome, as systemic disease already exists at the time of diagnosis. By the same token, nodal involvement is simply a clinical biomarker of distant disease already present and does not signify that adoption of the lymphatic route is mandatory in the path to distant metastasis. A third hypothesis accepts the existence of both of the above paradigms, and considers breast cancers to display a heterogeneous spectrum of metastatic proclivity, ranging from loco-regionally confined disease to systemically disseminated disease when first detected. This has come to be known as Hellman's paradigm[48,49] and provides a rational explanation for why loco-regional

treatment of some breast cancers with surgery and radiation is effective in favorably impacting clinical outcome, while seemingly ineffective at preventing distant metastasis in other cases.

In our study, the characteristics of EMAT2, EMAT3 and EMAT4 subtypes in LN and LP patients (comprising the majority of cases) are consistent with the Fisher's paradigm. However, the characteristics of the EMAT1 subtype (comprising of a smaller but significant number of cases) can be better explained in light of the Halstedian paradigm. EMAT1 had the best prognostic outcome in LN samples but exhibited poor prognosis in LP samples despite preservation of molecular features. Since the molecular characteristics of the EMAT1 subtype suggest non-aggressive biology, in LN as well as in LP samples, it cannot alone be the cause of the observed increase in metastatic propensity in LP patients. Instead, additional unique factors acquired in the involved lymph node microenvironment of LP patients may be the cause of such an effect, as suggested by Hellman[49]: "Lymph node involvement is of prognostic importance not only because it indicates a more malignant tumor biology, but also because persistent disease in the lymph nodes can be the source of distant disease." Just as we have come to realize that all primary breast cancers are not the same, so too we must realize that all nodal involvement may not be the same: while some tumor laden lymph nodes are a clinical biomarker of more aggressive disease, whose extirpation will not impact clinical outcome (more consistent with Fisher's paradigm), still others may actually be sources of future metastatic spread, if left untreated (more consistent with Halsted's paradigm). The EMAT model that we propose provides an effective starting molecular framework in which both conflicting paradigms find support. In addition, the fact that three molecularly distinct clusters predict heightened metastatic propensity in LN patients, supports the existence of a heterogeneous spectrum of metastatic proclivity within breast cancers, supporting the Hellman's paradigm. Thus the EMAT model, in addition to revealing molecularly distinct subtypes of

cancer metastasis, also appears to molecularly reconcile the relative contribution of three proposed paradigms governing breast cancer metastasis.

Taken together, our analysis for the first time identifies EMAT molecular subtypes of breast cancer progression and metastasis and provides a mathematically robust reasoning for the observed variance in metastatic propensity. While this advances our understanding of the heterogeneity and complexity of metastasis, delineation of the detailed molecular mechanisms underlying each EMAT subtype merits further investigation. Also, it still remains unclear whether the identified EMAT subtypes reflect the metastatic propensity of adenocarcinomas (originating from epithelial tissues) alone, or has an overlap with processes involved in driving metastatic propensity in cancers of mesenchymal origin as well.

METHODS

Derivation of MAT and EMAT gene lists

We obtained the list of 253 EMT-related genes from[7]. This list had been derived in the original publication by analyzing gene expression data obtained from 5 distinct and separate EMT-inducing cell perturbation experiments to identify genes up- or down-regulated at least 2-fold in at least 3 experimental groups relative to control cells. Following identical methodology to minimize derivation bias, we derived a new MAT signature by analyzing gene expression data obtained from 4 distinct and separate MAT-inducing cell perturbation experiments[12] to identify 138 genes up- or down-regulated at least 1.5-fold in at least 2 experimental groups relative to control cells. We then combined both the EMT and MAT gene signatures above to create the 388-member EMAT gene expression signature.

Data Collection

We downloaded gene expression and clinical data corresponding to the METABRIC study from OASIS (http://oasis-genomics.org/) and supplemental material of[13], respectively. In this dataset, 1055 samples (562 LN and 493 LP) had gene expression, survival information, and lymph node status. Gene expression and clinical information for cross-dataset analysis was downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE11121, which contains 200 LN breast cancer samples. Gene expression profile of H1 hESC lines[50] were downloaded from GEO under the accession number GSE54186. In all datasets, the probe intensities were log2 transformed and Z normalized prior to analysis.

Hierarchical clustering and survival analysis

Hierarchical clustering of samples was performed using the python module SciPy, using Ward's variance minimization algorithm. Survival analysis (Kaplan-Meier and Cox regression) was performed using lifelines python module (doi: 10.5281/zenodo.815943) and Survival R package (https://CRAN.R-project.org/package=survival). Variables included in the multivariate analysis are age, tumor size, and subtype status based on receptor profile, PAM50 centroid or EMAT centroid. All tests were two-sided, and P values of less than .05 were considered statistically significant.

Cross validation framework to evaluate survival predictive ability of different predictors

In order to compare the predictive ability of different classes of features (Figure 2), we used a cross-validation framework, in which half of the samples (patients) were randomly selected as the training set and the other half were used as a test set. A Cox regression model was trained on the training set using one of the four options for predictor features described before, and the model was used to estimate the expected survival of each sample in the test set. Then, the estimated survival times were compared to

the observed survival times using the concordance index (C-index). This process was repeated 200 times, each time using a distinct random partition of data. To determine the EMAT cluster assignment of the test samples, we first used hierarchical clustering to cluster samples in the training set into four clusters. Then we trained a centroid-based classifier on the training set and predicted the cluster labels for the test set.

Cluster silhouette score calculations

The silhouette score[14] is a measure of the similarity of a sample to its own cluster compared to other clusters. A higher silhouette score averaged over all samples indicates a better separation of samples into clusters. We calculated this score for n = 3, 4, 5 clusters using Scikit-learn python module (http://scikit-learn.org). We used three metrics, the Euclidean, correlation and cosine distances. Two of these metrics showed that n = 4 clusters generate the highest average silhouette score, which we used for our analysis.

The KNN and centroid-based classifiers

We implemented a KNN (only used to generate Supplementary Figure S2) and a centroid-based classifier for our analyses. In order to reduce the batch effects due to cross-dataset analysis and relax the normality assumption for gene expression values, we used the Spearman's rank correlation as the measure of the similarity of samples in these classifiers. In the centroid-based classifier, for each cluster a centroid was calculated as a vector in which each entry corresponds to the mean expression of a gene across the samples of the cluster. The cluster whose centroid had the highest Spearman's rank correlation with the expression profile of a test sample was selected as its label.

Acknowledgements

We gratefully acknowledge and thank our friend and colleague Dr. Peter Friedl for his critical review of the manuscript and invaluable suggestions for improvement.

Disclosure of Potential Conflicts of Interest

The authors declare the following conflicts of interest: AE, none; TR, employee, Onconostic Technologies, Inc.; TWJ, consultant, Onconostic Technologies, Inc.; MP, none; RN, none; SS, none; PSR, patents related to FOXC1 and FOXA1 in cancer, stock ownership in and consultant, Onconostic Technologies, Inc.

Disclaimer

The study funders had no role in the design of the study, the collection, analysis, or interpretation of the data, the writing of the manuscript, nor the decision to submit the manuscript for publication.

Authors' Contributions

PSR and TR conceptualized and designed the study. AE, SS and PSR developed the methodology. AE, TR, TWJ, and PSR performed data acquisition. AE, TR, TWJ, MP, RN, SS and PSR analyzed and interpreted the data. AE, TR and PSR wrote the manuscript, which was edited by all the authors who approved the final version. SS and PSR provided study supervision. All authors are guarantors of the integrity of the data collection and interpretation.

Funding sources

This work was supported by research grants/contracts from the National Cancer Institute at the National Institutes of Health 261201300028C-0-0-1 (PSR, TR), the grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (SS), Carle Foundation Translational Cancer Research Fund (PSR), and a Career Development Award from the Warren H. and Clara Cole Society (PSR).

REFERENCES:

1 Fidler, I. J. & Kripke, M. L. Metastasis results from preexisting variant cells within a malignant tumor. *Science* **197**, 893-895 (1977).

2 Smid, M. *et al.* Subtypes of breast cancer show preferential site of relapse. *Cancer research* **68**, 3108-3114 (2008).

3 Thompson, E. W. *et al.* Association of increased basement membrane invasiveness with absence of estrogen receptor and expression of vimentin in human breast cancer cell lines. *J Cell Physiol* **150**, 534-544, doi:10.1002/jcp.1041500314 (1992).

4 Yang, J. & Weinberg, R. A. Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. *Developmental cell* **14**, 818-829 (2008).

5 Wolf, K. *et al.* Compensation mechanism in tumor cell migration: mesenchymal-amoeboid transition after blocking of pericellular proteolysis. *J Cell Biol* **160**, 267-277, doi:10.1083/jcb.200209006 (2003).

6 Wolf, K. & Friedl, P. Molecular mechanisms of cancer cell invasion and plasticity. *Br J Dermatol* **154 Suppl 1**, 11-15, doi:10.1111/j.1365-2133.2006.07231.x (2006).

7 Taube, J. H. *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* **107**, 15449-15454, doi:10.1073/pnas.1004900107 (2010).

8 Tan, T. Z. *et al.* Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO molecular medicine* **6**, 1279-1293 (2014).

9 Marsan, M. *et al.* A core invasiveness gene signature reflects epithelial-to-mesenchymal transition but not metastatic potential in breast cancer cell lines and tissue samples. *PloS one* **9**, e89262 (2014).

10 Lehmann, S. *et al.* Hypoxia induces a HIF-1-dependent transition from collective-to-amoeboid dissemination in epithelial cancer cells. *Current Biology* **27**, 392-400 (2017).

11 Jones, B. C. *et al.* Dual Targeting of Mesenchymal and Amoeboid Motility Hinders Metastatic Behavior. *Mol Cancer Res* **15**, 670-682, doi:10.1158/1541-7786.MCR-16-0411 (2017).

12 Taddei, M. L. *et al.* Mesenchymal to amoeboid transition is associated with stem-like features of melanoma cells. *Cell Communication and Signaling* **12**, 24 (2014).

13 Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-352, doi:10.1038/nature10983 (2012).

14 Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53-65 (1987).

15 Weber, G. F., Bjerke, M. A. & DeSimone, D. W. A mechanoresponsive cadherin-keratin complex directs polarized protrusive behavior and collective cell migration. *Dev Cell* **22**, 104-115, doi:10.1016/j.devcel.2011.10.013 (2012).

16 Friedl, P. & Wolf, K. Tube travel: the role of proteases in individual and collective cancer cell invasion. *Cancer Res* **68**, 7247-7249, doi:10.1158/0008-5472.CAN-08-0784 (2008).

17 Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160-1167, doi:10.1200/JCO.2008.18.1370 (2009).

18 Ben-Porath, I. *et al.* An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nature genetics* **40**, 499-507 (2008).

19 Comen, E. A., Norton, L. & Massague, J. Breast cancer tumor size, nodal status, and prognosis: biology trumps anatomy. *J Clin Oncol* **29**, 2610-2612, doi:10.1200/JCO.2011.36.1873 (2011).

20 Emad, A., Cairns, J., Kalari, K. R., Wang, L. & Sinha, S. Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance. *Genome Biology* **18**, 153 (2017).
21 Schmidt, M. *et al.* The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer research* **68**, 5405-5413 (2008).
22 Santner, S. J. *et al.* Malignant MCF10CA1 cell lines derived from premalignant human breast epithelial MCF10AT cells. *Breast Cancer Res Treat* **65**, 101-110 (2001).
23 Maguire, S. L. *et al.* Three-dimensional modelling identifies novel genetic dependencies associated with breast cancer progression in the isogenic MCF10 model. *J Pathol* **240**, 315-328, doi:10.1002/path.4778 (2016).
24 Beerling, E. *et al.* Plasticity between epithelial and mesenchymal states unlinks EMT from metastasis-enhancing stem cell capacity. *Cell reports* **14**, 2281-2288 (2016).
25 Zheng, X. *et al.* Epithelial-to-mesenchymal transition is dispensable for metastasis but induces chemoresistance in pancreatic cancer. *Nature* **527**, 525-530, doi:10.1038/nature16064

http://www.nature.com/nature/journal/v527/n7579/abs/nature16064.html#supplementary-information (2015).
26 Bill, R. & Christofori, G. The relevance of EMT in breast cancer metastasis: Correlation or causality? *FEBS Lett* **589**, 1577-1587, doi:10.1016/j.febslet.2015.05.002 (2015).
27 Diepenbruck, M. & Christofori, G. Epithelial–mesenchymal transition (EMT) and metastasis: yes, no, maybe? *Current opinion in cell biology* **43**, 7-13 (2016).
28 Ye, X. *et al.* Upholding a role for EMT in breast cancer metastasis. *Nature* **547**, E1-E3, doi:10.1038/nature22816 (2017).
29 Aiello, N. M. *et al.* Upholding a role for EMT in pancreatic cancer metastasis. *Nature* **547**, E7 (2017).
30 Chen, H., Lin, F., Xing, K. & He, X. The reverse evolution from multicellularity to unicellularity during carcinogenesis. *Nat Commun* **6**, 6367, doi:10.1038/ncomms7367 (2015).
31 Laska, M. J. *et al.* Enforced expression of PPP1R13L increases tumorigenesis and invasion through p53-dependent and p53-independent mechanisms. *Mol Carcinog* **48**, 832-842, doi:10.1002/mc.20528 (2009).
32 Roy, D., Calaf, G. & Hei, T. K. Profiling of differentially expressed genes induced by high linear energy transfer radiation in breast epithelial cells. *Mol Carcinog* **31**, 192-203 (2001).
33 Baca-Lopez, K., Mayorga, M., Hidalgo-Miranda, A., Gutierrez-Najera, N. & Hernandez-Lemus, E. The role of master regulators in the metabolic/transcriptional coupling in breast carcinomas. *PLoS One* **7**, e42678, doi:10.1371/journal.pone.0042678 (2012).
34 Sun, C., Liu, C., Dong, J., Li, D. & Li, W. Effects of the myeloid cell nuclear differentiation antigen on the proliferation, apoptosis and migration of osteosarcoma cells. *Oncol Lett* **7**, 815-819, doi:10.3892/ol.2014.1811 (2014).
35 Sizemore, G. M., Pitarresi, J. R., Balakrishnan, S. & Ostrowski, M. C. The ETS family of oncogenic transcription factors in solid tumours. *Nat Rev Cancer* **17**, 337-351, doi:10.1038/nrc.2017.20 (2017).
36 Sang, Y. *et al.* TEL2 suppresses metastasis by down-regulating SERPINE1 in nasopharyngeal carcinoma. *Oncotarget* **6**, 29240-29253, doi:10.18632/oncotarget.5074 (2015).
37 Yamamoto, M., Cid, E., Bru, S. & Yamamoto, F. Rare and frequent promoter methylation, respectively, of TSHZ2 and 3 genes that are both downregulated in expression in breast and prostate cancers. *PLoS One* **6**, e17149, doi:10.1371/journal.pone.0017149 (2011).
38 Luanpitpong, S. *et al.* SLUG is required for SOX9 stabilization and functions to promote cancer stem cells and metastasis in human lung carcinoma. *Oncogene* **35**, 2824 (2016).

39 Zheng, M. *et al.* Snail and Slug collaborate on EMT and tumor metastasis through miR-101-mediated EZH2 axis in oral tongue squamous cell carcinoma. *Oncotarget* **6**, 6794 (2015).
40 Lefèvre, L. *et al.* Combined transcriptome studies identify aff3 as a mediator of the oncogenic effects of β-catenin in adrenocortical carcinoma. *Oncogenesis* **4**, e161 (2015).
41 Ray, P. S. *et al.* FOXC1 is a potential prognostic biomarker with functional significance in basal-like breast cancer. *Cancer research* **70**, 3870-3876 (2010).
42 Ray, P. S. *et al.* Abstract P1-07-30: FOXC1/FOXA1 transcriptional balance in breast cancer: From acquisition of mesenchymal and stem cell traits to occult lymph node independent breast cancer metastasis. *Cancer Research* **75**, P1-07-30 (2015).
43 Jensen, T. W. *et al.* Diagnosis of basal-like breast cancer using a FOXC1-based assay. *Journal of the National Cancer Institute* **107**, djv148 (2015).
44 Xu, Y. *et al.* Twist1 promotes breast cancer invasion and metastasis by silencing Foxa1 expression. *Oncogene* **36**, 1157-1166, doi:10.1038/onc.2016.286 (2017).
45 Sizemore, S. T. & Keri, R. A. The forkhead box transcription factor FOXC1 promotes breast cancer invasion by inducing matrix metalloprotease 7 (MMP7) expression. *J Biol Chem* **287**, 24631-24640, doi:10.1074/jbc.M112.375865 (2012).
46 Halsted, W. S. I. The results of operations for the cure of cancer of the breast performed at the Johns Hopkins Hospital from June, 1889, to January, 1894. *Annals of surgery* **20**, 497 (1894).
47 Fisher, B. *et al.* Five-year results of a randomized clinical trial comparing total mastectomy and segmental mastectomy with or without radiation in the treatment of breast cancer. *New England Journal of Medicine* **312**, 665-673 (1985).
48 Hellman, S. & Harris, J. R. The appropriate breast cancer paradigm. *Cancer Res* **47**, 339-342 (1987).
49 Hellman, S. Karnofsky Memorial Lecture. Natural history of small breast cancers. *Journal of Clinical Oncology* **12**, 2229-2234 (1994).
50 Kim, J. J. *et al.* Discovery of consensus gene signature and intermodular connectivity defining self-renewal of human embryonic stem cells. *Stem Cells* **32**, 1468-1479, doi:10.1002/stem.1675 (2014).