SCIENTIFIC REPORTS

OPEN

A novel embryonic plasticity gene signature that predicts metastatic competence and clinical outcome

Received: 05 January 2015
Accepted: 02 June 2015
Published: 30 June 2015

Rama Soundararajan[1], Anurag N. Paranjape[1], Valentin Barsan[1], Jeffrey T. Chang[4] & Sendurai A. Mani[1,2,3]

Currently, very few prognosticators accurately predict metastasis in cancer patients. In order to complete the metastatic cascade and successfully colonize distant sites, carcinoma cells undergo dynamic epithelial-mesenchymal-transition (EMT) and its reversal, mesenchymal-epithelial-transition (MET). While EMT-centric signatures correlate with response to therapy, they are unable to predict metastatic outcome. One reason is due to the wide range of transient phenotypes required for a tumor cell to disseminate and recreate a similar histology at distant sites. Since such dynamic cellular processes are also seen during embryo development (epithelial-like epiblast cells undergo transient EMT to generate the mesoderm, which eventually redifferentiates into epithelial tissues by MET), we sought to utilize this unique and highly conserved property of cellular plasticity to predict metastasis. Here we present the identification of a novel prognostic gene expression signature derived from mouse embryonic day 6.5 that is representative of extensive cellular plasticity, and predicts metastatic competence in human breast tumor cells. This signature may thus complement conventional clinical parameters to offer accurate prediction for outcome among multiple classes of breast cancer patients.

Much like the multi-step process of tumor progression to metastasis, mammalian embryo development encompasses an intricate series of morphogenetic events, and involves spatial and temporal coordination of multiple cell types as they proliferate, migrate, and differentiate to form complex higher-order organs and organ systems. The spatially-confined and sessile epithelial cells form tight cell-cell contacts in various organs, thus furnishing a critical barrier necessary to maintain a regulated environment. In contrast, the malleability offered by the motile and invasive mesenchymal cell type is what permits cell migration for tissue- and organ development. While these two cell states are quite distinct, they are also highly dynamic and interchangeable during embryo development, as well as during tumor progression to metastasis. A complex cellular reprogramming event called epithelial-to-mesenchymal-transition (or EMT) facilitates the conversion of differentiated epithelial cells (expressing surface E-cadherin) into loosely organized, highly migratory and invasive mesenchymal cells (lacking E-cadherin expression, among other profound alterations)[1]. Recent studies have established a strong molecular link between EMT and stem-cells, and further, suggested that EMT confers differentiated cells with stem-cell properties[2,3]. Because of the adaptive malleability that cancer cells acquire through the activation of EMT, many transcription factors (TF) capable of regulating EMT and their associated signaling pathways, have been proposed for the identification and classification of tumors that metastasize[4–6]. Supporting the importance of EMT in disease, the EMT inducer Snail has been shown to predict recurrence in breast cancer[7].

[1]Department of Translational Molecular Pathology. [2]Metastasis Research Center. [3]Center for Stem Cells and Developmental Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas. [4]Department of Integrative Biology and Pharmacology, School of Medicine; School of Biomedical Informatics; The University of Texas Health Sciences Center at Houston, Houston, Texas. Correspondence and requests for materials should be addressed to J.T.C. (email: Jeffrey.T.Chang@uth.tmc.edu) or S.A.M. (email: mani@mdanderson.org)

Unfortunately, in many cases, gene expression profiles of EMT genes do not predict outcomes. For example, we previously generated an EMT-specific signature by over-expressing Snail or Twist or Goosecoid or TGFβ1 in breast epithelial cells[8], but discovered that this signature was incapable of identifying patients who could potentially develop metastasis or tumor relapse (Suppl Figs 1,2 – column 'EMT'). This suggests that EMT does not completely explain clinical outcomes. Other factors are also important. A solution to this conundrum came in a series of experiments that showed that mesenchymal cells that had undergone EMT are unable to form macrometastases[9,10]. The subsequent differentiation of these mesenchymal/stem-like cells, with restoration of epithelial features (the reverse process referred to as mesenchymal-to-epithelial-transition, MET), is what critically determines the formation of multiple tissues and organs during embryo development. It is becoming increasingly evident that the pathophysiological course of tumor cell invasion to metastasis is dependent *not* on the sole possession of particular EMT or MET or stem-cell traits, but instead on the innate flexibility of cancer cells in being able to dynamically switch between these various states, alter cellular morphology and function[1,11,12]. That is, metastatic cells must possess 'plasticity'.

The past decade has witnessed the development of a number of gene expression signatures related to embryo development, cell migration, stemness or EMT, tested for use in cancer patient outcome predictions[13–22]. These comprise a wide variety of cell types of embryonic origin, lines that exhibit stem-cell features, cells which display classic EMT, and include different methods of EMT/stemness/pluripotency induction, or microarray/retrospective RT-PCR analyses of a defined set of cancer-related genes from patient samples. While these resources all have their own merit, they have either not been tested for metastasis predictions, or are unfortunately quite limited in their potential to accurately predict the metastatic recurrence of tumors or in their extended applicability across a broad spectrum of cancer subtypes.

Human breast cancer includes a highly diverse set of diseases, and is classified into at least 6 distinct clinically-relevant molecular subgroups - luminal A, luminal B, HER2+/ER-, basal-like, normal breast-like, and the most recently recognized claudin-low[23,24]. It is clear that at the molecular level, each breast cancer subtype displays distinct gene expression profiles, and exhibits significant differences in response to therapy and patient outcome[25,26]. Importantly, cancer cells with EMT- and stem-cell (CSC) properties are present in varying degrees among these different tumor subtypes, and may therefore not be adequately represented, in order to facilitate identification in gene signatures that are uniquely enriched for EMT/CSC traits. This is reiterated by the inability of our previously published EMT-signature to predict metastatic outcome[8]. This signature is, however, capable of identifying claudin-low tumors that are enriched with cells harboring mesenchymal and stem-cell traits, suggesting that EMT-centric signatures can only identify a "static cellular state" and not the "dynamic EMT/MET process".

We therefore hypothesized that gene expression signatures that correlate with the time when the embryo undergoes extensive yet defined spatiotemporal cellular dynamics, that encapsulates a spectrum of changes that *together* permit cellular plasticity, would be superior to those rely on the expression of individual EMT/MET/stem-cell markers. In this manuscript, we document the discovery and characterization of one such unique gene expression signature derived from mouse embryo development (E6.5), that can forecast breast cancer patient survival rate. We further describe a novel score that can be used to quantify the extent of cellular plasticity in the context of breast tumor progression, and demonstrate that this score, which is based on the E6.5 signature, can accurately predict the metastatic ability of tumor cells.

Results

Derivation of a physiologically-relevant embryonic gene expression signature that reflects cellular plasticity. The parallels observed in cellular behavior during early embryo development and tumor metastasis support the hypothesis that metastatic competence in cancer cells is dictated by cellular plasticity, which can be modeled by examining similar gene expression patterns during embryogenesis. We therefore sought to isolate a specific developmental stage demonstrative of high cellular plasticity, and utilize its gene expression profile as a unique tool to predict tumor metastasis.

Among the various stages of murine embryonic development, the onset of gastrulation occurs at E6.5 and represents high cellular plasticity[11,27]. Gastrulation begins with the establishment of the 'primitive-streak', which provides a path for future cell migration. Cells from the epiblast migrate into the core of the embryo *via* this 'primitive-streak', in a process termed ingression, which involves EMT. The event of EMT bestows early epithelial cells of the epiblast with the flexibility associated with a more invasive and motile phenotype, required for movement along, and eventually away from, the 'primitive-streak' to define the trilaminar disc comprised of three germ layers (ectoderm, mesoderm and endoderm) that eventually contribute to the formation of the various organs[11,27]. This process of trilaminar cell allocation is completed by E8.0. E6.5 of the developing mouse embryo is thus representative of extensive cellular plasticity, wherein epithelial cells are programmed to traverse the streak. We therefore analyzed the expression of genes at E6.5 of murine embryonic development, using the published database at the BioGPS website[28], and thus arrived at a gene signature, which is representative of cellular plasticity. We further reasoned that as a corollary, a "differentiated" expression signature identified from "adult" tissues, which are not expressed in the embryo, should serve as the corresponding negative correlate representative of decreased cellular plasticity. Figure 1 illustrates the source and step-wise derivation of

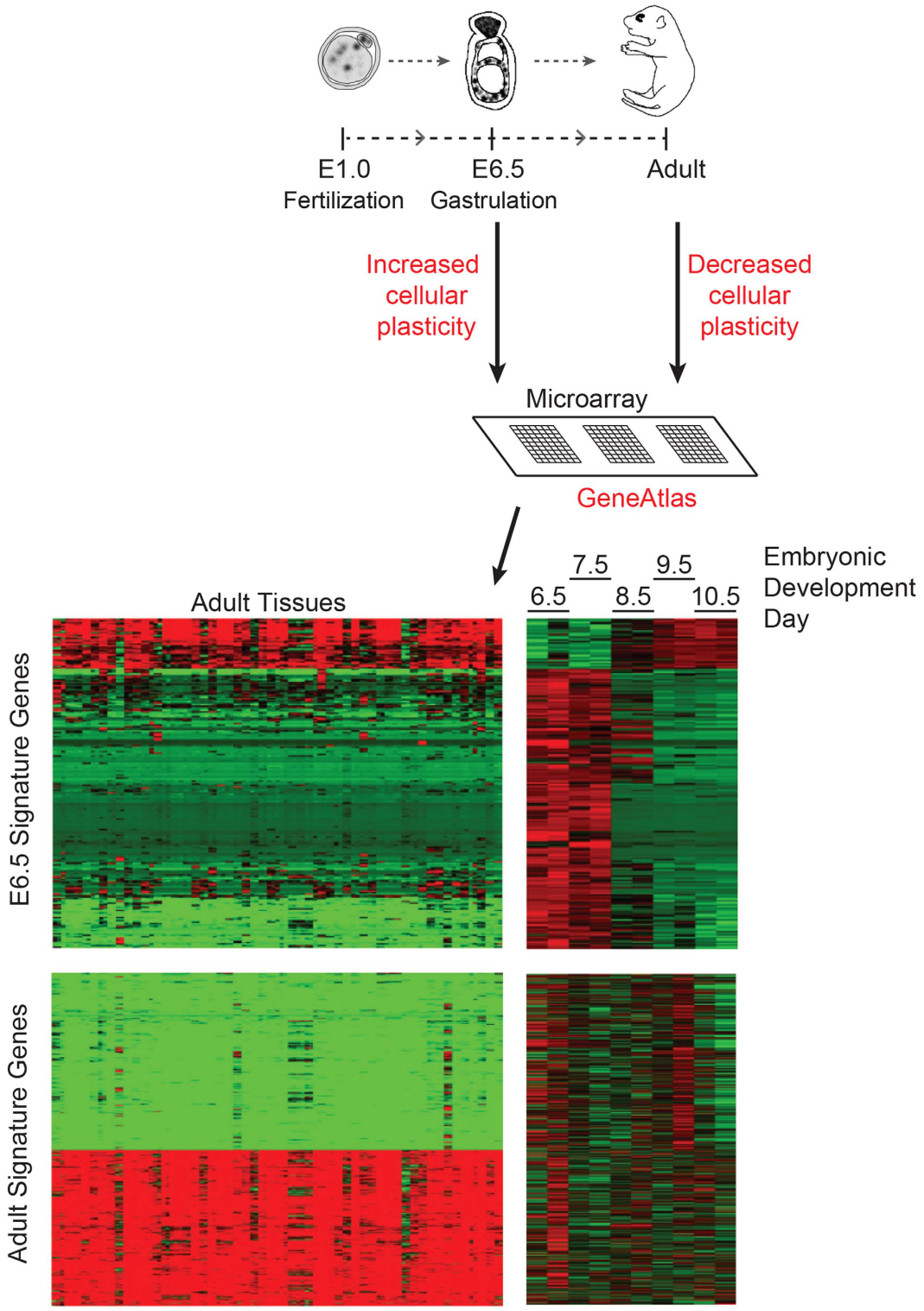


Figure 1. Derivation of a stage-specific gene expression signature from mouse embryonic development, as well as from adult murine tissues. A three-step schematic illustrating the generation of gene signatures specifically from murine E6.5- and adult tissues using the published Gene Atlas microarray database[28]. The heat maps show the expression profile of genes in these signatures. The left column represents the expression of the genes in adult tissues and the right column shows expression at various stages of mouse embryonic development (E6.5–E10.5), with green indicating lower gene expression and red representing higher relative expression. Each row represents a gene, and each column, a specific tissue.

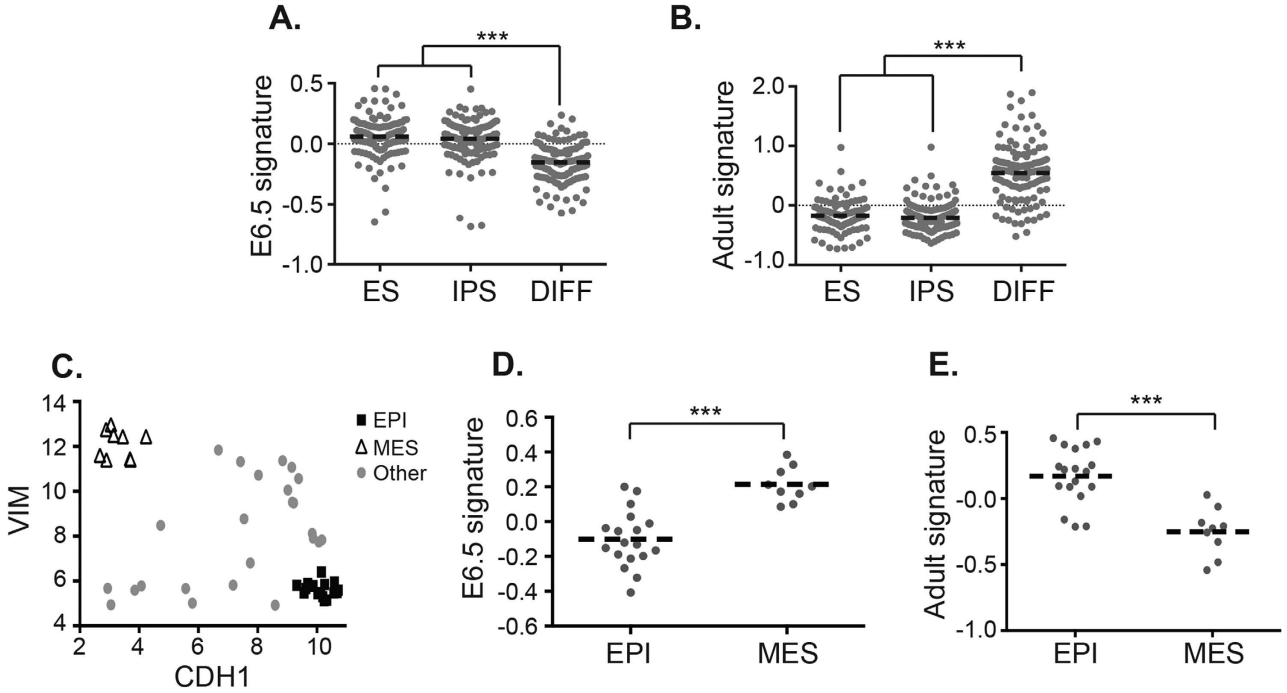


Figure 2. The newly identified E6.5 gene signature recognizes features representative of the mesenchymal phenotype, embryonic stem-cells (ES), as well as induced pluripotent stem-cells (IPS). Scatter plots **A** and **B** show that the E6.5 signature identifies with ES- and IPS- gene expression profiles procured from publicly available cell line databases, whereas the adult signature (negative correlate) recognizes the profile of the differentiated phenotype. Each dot represents a distinct cell line. Next, numerous cancer cell lines from the E-TABM-157 database were segregated as epithelial (EPI) or mesenchymal (MES), based on their expression levels of E-cadherin (CDH1) and Vimentin (VIM), and the "strongly epithelial (squares)" and "strongly mesenchymal (triangles)" sources alone (**C**) were used to generate comparative gene expression profiles for scoring the newly identified signature. Scatter plot **D** demonstrates that the E6.5- signature distinctly correlates with the EMT archetype, strongly recognizing mesenchymal profiles of breast cancer cell lines, while the adult signature does not (**E**). ***$p < 0.001$, ES- Embryonic Stem cell signature, IPS- Induced pluripotent stem cell signature, DIFF- Differentiated cell signature.

the E6.5- and adult gene expression signatures. The complete list of genes constituting both signatures is provided in Suppl Table-1.

Characterization of the E6.5 plasticity gene signature. Because of the significantly increased expression of EMT- and stem cell-related markers at this stage of embryonic development, we investigated the ability of the newly identified signature to recognize features characteristic of stem cells, as well as cells displaying the classic EMT phenotype. Figure 2A demonstrates the correlation of the E6.5 signature with embryonic stem cell- and induced pluripotent stem cell-gene expression profiles obtained from publicly available gene expression databases. In contrast, the adult signature (the intended negative correlate) recognizes the profile of the differentiated phenotype (Fig. 2B). We further confirmed the potential of the E6.5 signature for recognition of the EMT archetype in a comprehensive cancer cell line database (E-TABM-157) (Fig. 2C–E).

A quantitative score to predict the metastatic behavior of breast tumor cells. The genes defining the E6.5 signature exhibit remarkable functional diversity ranging from cellular maintenance, tissue morphology and embryonic/organismal development to lipid metabolism, free radical scavenging, and cancer metastasis (Suppl Table-2), the sum total of which, we reasoned, should collectively exemplify cellular plasticity. We next compared various human breast tumor cell lines of established metastatic capacity in animal models, and scored them based on their expression of the E6.5 signature (Fig. 3). Interestingly, a positive E6.5 score consistently correlated with a strong propensity for distant metastasis, while a negative score suggested lack of ability to metastasize (Fig. 3, Suppl Fig. 3A). As expected, and in contrast, a positive adult score correlated with an increased propensity for distant metastasis (Fig. 3, Suppl Fig. 3B).

Cell line	E6.5 Score	Adult Score	Distant Metastasis Strong/Poor (S/P)	Reference (PMID#)
MCF7	-0.133	0.084	P	**20228842, 23118918**
BT-474	-0.097	0.062	P	**23118918**
T47D	-0.052	0.077	P	**16982776**
SK-BR-3	0.064	0.116	S	**24805814, 25284724, 20498642**
BT-549	0.175	-0.168	S	**20595686**
MDA-MB-231	0.233	-0.097	S	**10516850, 23118918**
MDA-MB-436	0.240	-0.123	S	**23118918**

Figure 3. A "plasticity" score conferred by the E6.5 signature accurately predicts the metastatic competence of breast tumor cells. Various human breast cancer cell lines of known distant metastatic capacity *in vivo* were scored for their concordance with the E6.5- or adult-signature, and assigned a 'plasticity' score. A positive E6.5-score correlated strongly with the propensity for distant metastasis, and a negative score, poorly. In contrast, a positive adult-score correlated negatively with the propensity for distant metastasis (also refer to Suppl Fig. 3 for a quantitative representation of assessment of metastatic competence).

Established EMT factors fail in their ability to reliably predict distant metastasis or tumor recurrence. We next tested the predictive power of selected well-recognized EMT- or stem cell-related genes, in foretelling distant metastasis and tumor recurrence among breast cancer patients, in four independent publicly available databases [GSE20685, GSE6532_GPL96, GSE7390, GSE11121 for distant metastasis-free survival (DMFS) analyses, and GSE1456, GSE12276, GSE4922_UPPSALA, and GSE21653 for recurrence-free survival (RFS) analyses]. Representative DMFS- and RFS analyses using 6 key EMT- or stem cell-related genes (CDH1, PRRX1, TCF8, SNAI2, VIM and TWIST1), from one cohort, are shown in Fig. 4A,B respectively. Cumulative analyses from all four cohorts (displayed as 4 distinct colored dots, each dot representing one cohort) for these 6 genes, are presented in Fig. 4C,D. A more detailed analysis querying the relevance of 39 well-recognized EMT- or stem cell-related genes (in all 4 cohorts), is presented in Suppl Fig. 8.

As clearly discernable in Fig. 4 (and Suppl Fig. 8), most classic EMT- or stem cell-related markers considered individually, are poorly prognostic. They are mostly either non-predictive, or predict the opposite result. Further, their performance is inconsistent across the datasets. E-cadherin ("purple" dots) is a good example. E-cadherin is perhaps the most commonly used determinant of the EMT phenotype. E-cadherin is an epithelial cell adhesion molecule, and its loss has been reported to be a prerequisite for the onset of EMT and tumor cell migration[29]. However, our results demonstrate its poor prognostic power for determination of DMFS (Fig. 4A-CDH1; Fig. 4C-"purple" dots) and RFS (Fig. 4B-CDH1; Fig. 4D-"purple" dots), and further, indicate its inconsistency when tested in multiple patient datasets (represented as 4 distinct "purple" dots, Fig. 4C,D).

The E6.5 gene signature has prognostic value in predicting clinical outcome. Since the identified gene signature from E6.5 is reflective of cellular plasticity, a property critical for carcinoma metastasis, we next queried if the net pattern of gene expression at this particular stage of the developing mammalian embryo, can be used for prognosis of DMFS and RFS, in all four datasets. Representative graphs from one cohort are presented in Fig. 5A,D (Individual survival graphs for DMFS and RFS, across all cohorts tested, are provided in Suppl Figs 1 and 2 respectively – "E6.5"). We observed that the E6.5 signature was capable of accurately stratifying patients according to risk for distant metastasis as well as for tumor recurrence (Fig. 5C,F; each database is represented as a distinct shape in each plot).

In contrast, and as expected, the adult signature (triangles) representative of decreased cellular plasticity, served as the negative correlate for the clinical prediction. Patients exhibiting a "high" score for the "adult" signature clearly demonstrated better clinical outcomes, in independent patient cohorts (represented as 4 separate triangles) (Fig. 5B,C,E,F, Suppl Figs 1 and 2 – "Adult").

Notably, the datasets used for DMFS and RFS analyses comprised heterogeneous groups of breast cancer patients that were assigned to various molecular sub-categories. The specific tumor subtypes included in each study are highlighted in Fig. 5G. These data suggest that the newly identified E6.5 signature captures a distinct gene expression pattern that is independently indicative of tumor recurrence and/or distant metastasis among a wide array of breast cancer patients - luminal A, -B, -C, ER+/−, PR+/−, Her2+/−, basal and grade I, II or III tumors.

Discussion

EMT is often heralded as a significant participant in tumor progression, and many studies have provided morphological evidence for EMT at the invasive fronts of human tumors[30,31]. Interestingly however, the EMT status of the primary tumor does not independently predict post-operative recurrence or disease-free survival in cancer patients[18,32]. Although EMT-endowed migratory and invasive properties contribute to the dissemination of carcinoma cells, in many instances, these features *alone*, are not sufficient for carcinoma cells to complete the cascade leading to eventual macrometastatic colonization.

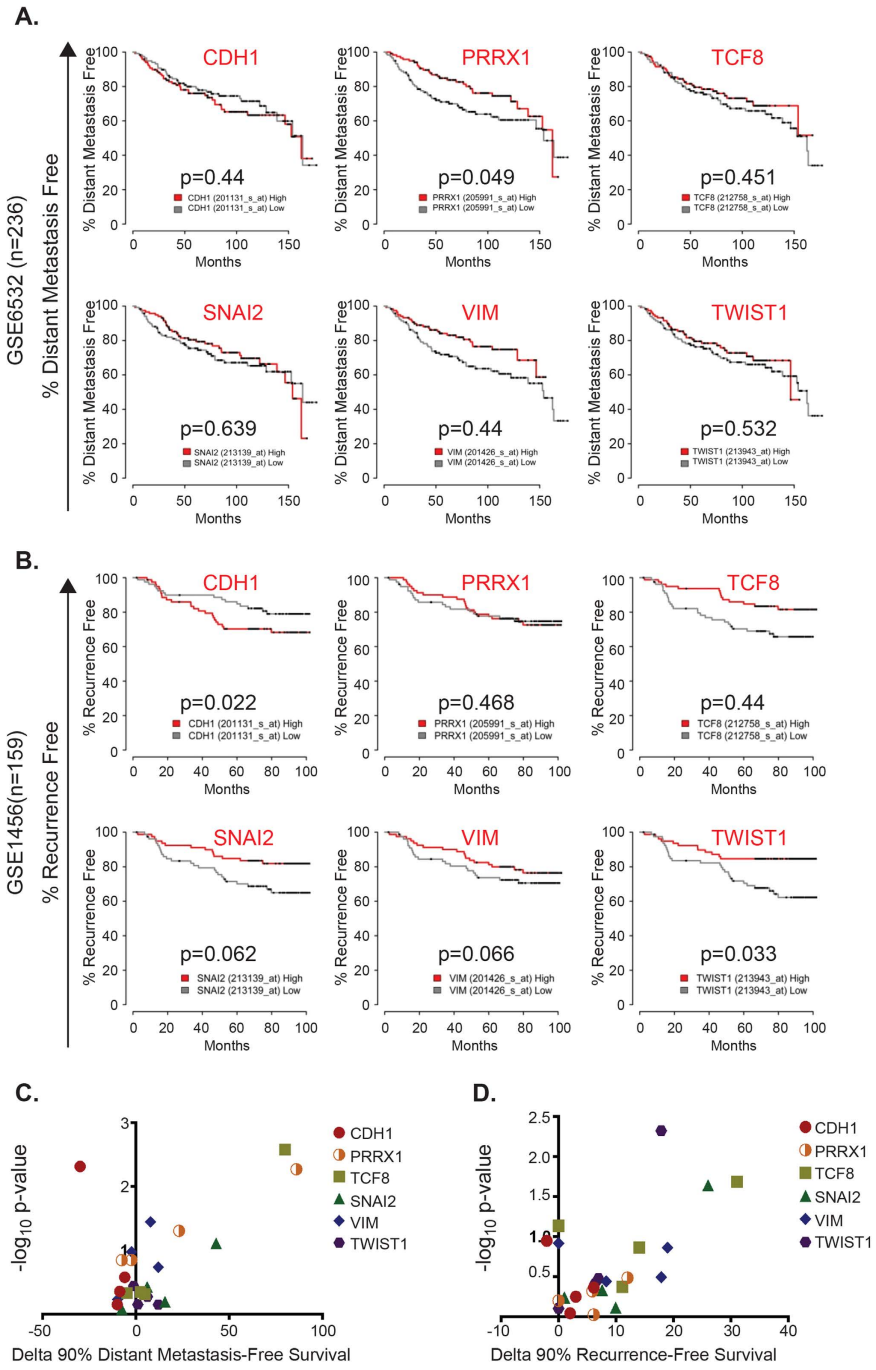
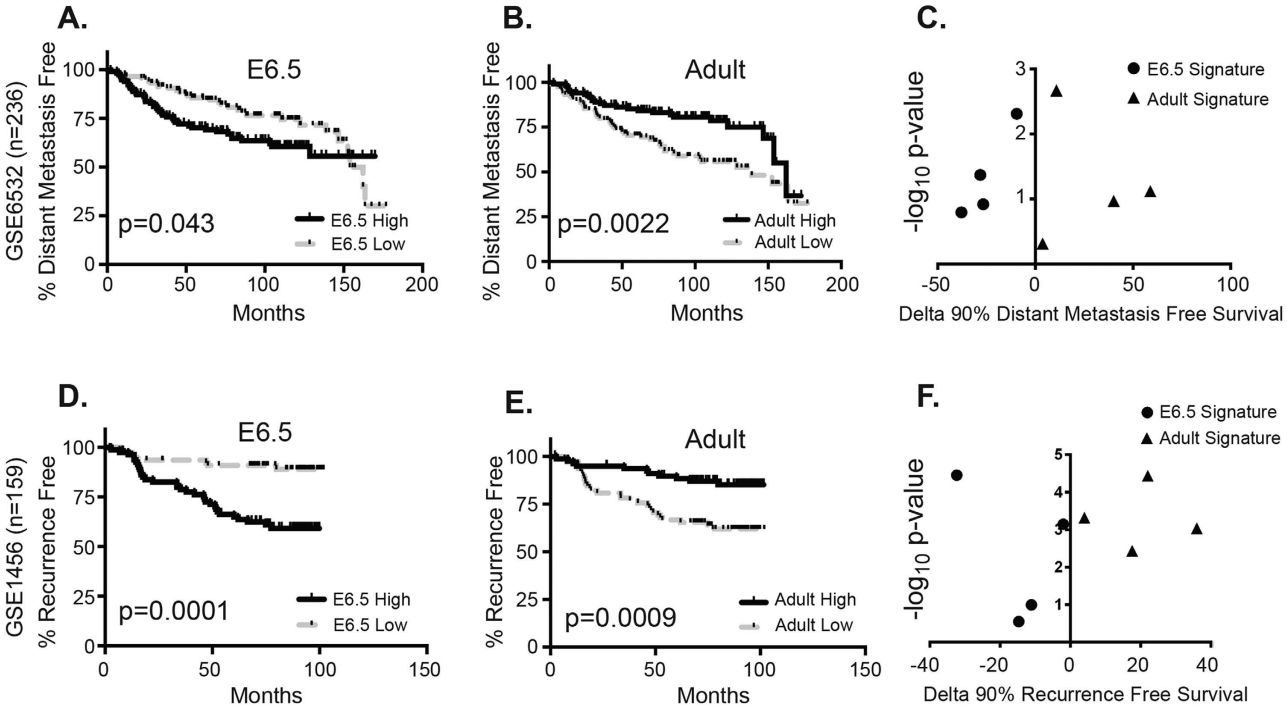


Figure 4. Classic EMT-/stem cell-related markers are poor predictors of distant metastasis-free survival (DMFS) as well as recurrence-free survival (RFS) in breast cancer patients. Patient cohorts representing various subtypes of breast cancer were discretized into high- or low-expression groups, based on their concordance with expression of well-established EMT-/stem-cell markers popularly tested for risk assessment (full gene-list provided in Suppl Fig. 8A; also refer to compiled Kaplan-Meier prediction data in Suppl Figs 8B, C). Panel **A** shows DMFS curves for 6 well-known EMT-/stem cell-related factors, based on data analyzed from one patient cohort (GSE6532). Cumulative data from survival curves of all four patient cohorts analyzed (represented as 4 distinct colored dots; also refer to individual Kaplan-Meier plots in Suppl Figs 4, 5) were aligned on a scatter plot, and their performance in foretelling risk for distant metastasis was tested in C. Panel B shows respective representative RFS curves based on data analyzed from one patient cohort (GSE1456). Cumulative data from survival curves of all four patient cohorts analyzed (represented as 4 distinct colored dots; also refer to individual Kaplan-Meier plots in Suppl Fig. 6, 7) were aligned on a scatter plot, and their performance in foretelling risk for tumor recurrence was tested in D. CDH1: E-cadherin; PRRX1: Paired mesoderm homeobox 1; TCF8: Zinc finger E-box-binding homeobox 1 or Zeb1; SNAI2: Snail2 or Slug; VIM: Vimentin; TWIST1: Twist1.



G.

PMID	RFS Data Sets	Salient patient tumor features
16280042	GSE1456	-All breast cancer patients undergoing surgery between 1994-1996 operated at the Karolinska Hospital
		-Includes ER+ and ER- cases, Grade I, II and III tumors
19421193	GSE12276	-Breast cancer patients with known site of relapse
		-MSK18 cohort includes ER-/PR-/Basal subtypes; NKI-295 cohort includes ER+ and ER- cases, Grade I, II and III tumors
		-Study includes Basal, Luminal A, Luminal B, ERBB2, Normal-like,TBRS+ and TBRS- cases
17079448	GSE4922_UPPSALA	-Primary invasive breast tumors
		-Includes ER+, ER-, PR+, PR- cases
20490655	GSE21653	-Distinct invasive early breast adenocarcinomas
		-Includes Medullary Breast Carcinomas displaying basal profile
	DMFS Data Sets	
21501481	GSE20685	-Cancer tissue from every third patient diagnosed and treated between 1991-2004 at KFSYSCC
		-Includes ER+, ER-, PR+, PR-, Her2+, Her2- cases
17401012	GSE6532_GPL96	-Includes breast cancer Grade I, II and III tumors
		-Data compared to Sorlie dataset that includes Basal-like, ERBB2, Luminal A, B, C and normal-like subtypes
17545524	GSE7390	-Node-negative tumors
		-Includes ER+ and ER- cases
18593943	GSE11121	-Lymph node-negative breast cancer patients
		-Includes ER+, ER-, PR+ and PR- cases
		-Analyzed gene clusters represented luminal, ERBB2, normal-like geneontology

Figure 5. The E6.5 gene expression signature reflecting the plasticity phenotype, predicts distant metastasis-free survival (DMFS) as well as recurrence-free survival (RFS) in breast cancer patients. Patients cohorts representing various subtypes of breast cancer were discretized into high- or low-expression groups, based on their concordance with E6.5, or adult signatures. **A** and **B** show respective representative DMFS curves based on data analyzed from one patient cohort (GSE6532). Cumulative data from survival curves of all four patient cohorts analyzed (represented as 4 distinct dots; also refer to individual Kaplan-Meier plots in Suppl Fig. 1) were aligned on a scatter plot, and their performance in foretelling risk for distant metastasis was validated in **C**. **D** and **E** show respective representative RFS curves based on data analyzed from one patient cohort (GSE1456). Cumulative data from survival curves of all four patient cohorts analyzed (represented as 4 distinct dots; also refer to individual Kaplan-Meier plots in Suppl Fig. 2) were aligned on a scatter plot, and their performance in foretelling risk for tumor recurrence was validated in **F**. As expected, the adult signature (triangles) correlates with better survival in both analyses. The identity of various publicly available patient databases used for these analyses is listed in **G**, along with the included breast tumor types in each dataset. ER: Estrogen receptor; PR: Progesterone receptor; ERBB2/Her2: Human epidermal growth factor receptor 2; TBRS: TGFβ-responsive signature.

Recent work has highlighted the necessity for down-regulating potent EMT-inducers such as Snail, Prrx1 or Twist, in order to allow for distant metastasis[9,10,33]. Thus constitutive/sustained/unregulated EMT actually *suppresses* metastasis. In support of this notion, Ocana *et al.*, found no clear correlation between the expression of Twist or Snail in primary tumors and relapse-free survival in breast cancer patients[9]. Moreover, carcinoma macrometastases at distant sites share an epithelial histopathology similar to that of the primary tumor at the site of origin[10,34,35]. While this appears paradoxical, and in a way, questions the participation and relevance of EMT in cancer progression, the following outlook lends a plausible explanation, accounting for both sets of observations[9,10,33,36,37] - EMT is a transitory phenomenon employed by carcinoma cells during their multistep progression towards metastasis. Such migratory cancer cells harboring increased expression and function of classical EMT-TFs invade through the extracellular matrix, enter the blood and lymphatic vessels, and thereby spread to new locations within the body. While EMT- and stem cell-related traits are prominent players in the initial stages of metastasis, loss of these mesenchymal/stem-cell attributes and simultaneous retrieval of epithelial features (namely, MET) is what permits the establishment of metastatic colonies at outlying sites. These distant tumors are therefore still epithelial (hence, similar to the primary tumors); however, they have additionally recovered the proliferative state necessary for subsistence.

This implies that cells that are able to successfully chart the EMT-MET course (presumably in response to various intrinsic and extrinsic factors) have a higher propensity to metastasize. In other words, it is the plasticity of the carcinoma cell that dictates whether or not it can metastasize. Implicit in this notion, is the fact that gene expression signatures that are solely based on EMT (in other words, signatures representing features required for completing just one-half of the journey to metastasis), therefore cannot reliably predict distant colonization. We therefore reasoned that plasticity-bestowing traits have to be necessarily factored into this equation, in order to be able to foretell metastatic competence.

Interestingly, although whole-tumor gene expression profiles are quite distinct across the various molecular subtypes of human breast cancer, the stem-cell populations isolated from these tumors (that are considered to be the harbingers of metastases), exhibit high conservation in gene expression patterns, that is independent of the sub-classification[38]. Importantly, these cancer stem-cells were shown to readily transition between epithelial and mesenchymal states[38]. This further emphasized that a common biological phenomenon such as "cellular plasticity" (that can be conserved across the various subtypes) would better serve as the basis for the development of gene expression signatures aimed at impartially addressing all breast tumors.

The E6.5 signature presented in this manuscript in unique in that it denotes a net gene expression pattern indicative of "cellular plasticity" (regardless of the expression states of individual EMT/stem-cell/pluripotency markers), and thus is capable of reliably predicting DMFS and RFS in breast cancer patients, better than some of the most well-recognized EMT/stem-cell genes considered individually (Figs 4 and 5). While the E6.5 signature does indeed recognize features characteristic of stem-cells and the EMT phenotype, included in this signature list are various other protein kinases, proliferation modulators and regulators of lipid metabolism, among others (Suppl Tables-1, 2), the sum total of which is predictive of tumor spread. This discovery complements recent work involving other cancers (lung) that demonstrate that acquisition of ectopic embryonic gene expression profiles characterizes tumors of the aggressive phenotype[39]. Indeed, assignment of a score based on concordance with the E6.5 signature (which we have termed the "plasticity score") to some of the most popular human cell lines used for breast cancer research (Fig. 3), or even prostate cancer research for that matter (Suppl Fig. 9), appears to separate them based on their ability to invade and metastasize *in vivo*, with a striking degree of accuracy.

Prognostic value of the E6.5 embryonic gene signature. Histopathological grading is still the most commonly used clinical prognostic indicator for breast cancer patients, and the vital benchmark for the appraisal of therapeutic options. This evaluation is, however, based on anatomical information, or the binary presence/absence of a few select markers. Existing biomarkers are limited in their application because breast cancer is a heterogeneous disease, and patients often exhibit drastic differences in response to therapy despite similarities in histological types, grade and stage. For example, while analysis of E-cadherin expression may be helpful as an aid to the histopathological sub-classification of breast tumors[40], our study demonstrates that its practical utility as a prognostic biomarker is questionable because reproducibility across multiple datasets is an apparent problem (Fig. 4). Although many studies suggest a direct association between reduced E-cadherin expression and high histopathologic grade, its correlation with metastatic progression is far less consistent[41–43]. Further, E-cadherin expression has been clearly shown to have no prognostic role in the intermediate stage II cases of breast cancer[44] (which represent 30–60% of all cases, and are also the major source of inter-observer discrepancy among pathologists, making treatment decisions difficult), thus limiting its potential use as a common predictive factor applicable to all breast tumors.

Two well-characterized gene expression signatures for breast cancer that have been translated to the clinic are the Mammaprint (a 70-gene signature[45,46]) and the Oncotype, (a 21-gene signature[14]). While their reproducibility has been tested in multiple independent patient cohorts across the world, their use is unfortunately limited to patients with ER-positive disease. Similarly, the 241-gene signature capable of isolating patients at risk for late relapse, identified by Mittempergher *et al.*[47], is designed for ER-positive/HER2-negative patients. The prognostic power of the 186 gene-based "invasiveness

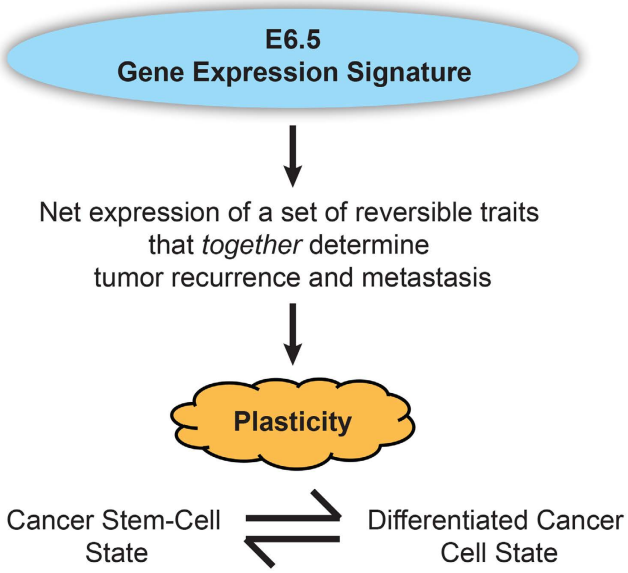


Figure 6. Study summary schematic. The newly identified E6.5 gene expression signature derived from a specific stage of the developing murine embryo, which is representative of high cellular plasticity (and reflective of the potential to readily switch between distinct cell states), encapsulates a spectrum of changes that, *collectively*, recognizes the ability of breast tumors to relapse and/or metastasize to distant organs.

gene signature" developed by Liu *et al.*[48] (derived from genes differentially-expressed in tumorigenic $CD44^{hi}/CD24^{lo}$ breast cancer cells), which is associated with increased risk of metastasis and early death, is significant only in patients with ER-positive and intermediate grade tumors. On the other hand, a recently described 41-gene signature derived from breast cancer stem-cells, or a 31-gene signature derived from tumor-initiating stem-like cells that demonstrated good prognostic significance with a strong capability of predicting distant metastasis-free survival and overall-survival, were only able to perform well in ER-negative breast cancer patients[17,49].

In contrast, our results indicate that the E6.5 gene signature, that addresses a common biologic property underlying tumor spread among most sub-classifications of breast cancer, is able to serve as a reliable prognostic predictor of distant metastasis and poor clinical outcome among a broad range of patients (Fig. 5).

In conclusion, we document the identification and potential application of a unique gene expression signature derived from the E6.5 stage of mouse embryonic development (Fig. 6), as a reliable prognostic indicator that may guide treatment options in breast cancer patients based on estimated survival benefit, as well as aid in monitoring of disease progression. This signature is unique in that it encapsulates a distinct biologic property that appears to dictate the metastatic competence of tumor cells, and because its prognostic power is applicable to a wide range of breast cancer patients. Further, the E6.5 score can be applied to predict the metastatic capacity of breast tumor cells *in vivo*. Directed prospective studies in the future will determine if this signature can indeed complement classic prognostic factors to improve patient outcome.

Methods

Derivation of the E6.5- and adult gene expression signatures. We downloaded the GeneAtlas gene expression data, as well as the accompanying annotations from the BioGPS website[28], and converted the mouse Entrez Gene IDs to human using the Homologene database[50]. We then excluded the samples from cell lines, those that had been stimulated, or those from undeveloped tissue (excluding tissues from embryonic day 6.5 to 10.5), and logged expression values with base 2.

To generate the embryonic day 6.5 signature, we compared the expression at E6.5 and E7.5 against those at E9.5, E10.5, and adult tissues. We omitted day E8.5 because, based on the expression patterns, it appeared to be a transition period between the earlier and later time points. We selected the genes that were differentially expressed with a 5-fold change and false discovery rate <0.05, as determined by an empirical Bayes approach[51]. This identified 190 probes that were expressed significantly higher in embryonic days 6.5 and 7.5, and 34 probes higher in day E9.5, E10.5, and adult tissues. For the differentiated signature, we compared the expression in adult tissues against those at E6.5, E7.5, as well as E9.5 and E10.5. We selected the genes for this signature using the same strategy, resulting in 370 probes higher in adult tissues, and 420 probes higher in embryonic days 6.5, 7.5, 9.5, and 10.5. Significantly differentially

up-regulated ("Up") and down-regulated ("Down") genes were combined in each set to develop a unique stage-specific signature. Complete gene list is presented in Suppl Table-1.

To score the signatures on a gene expression data set, we first processed it. If an Entrez Gene ID matched multiple probes on the array, we selected the probe with the highest variance across the data set. Then, for each gene, we normalized it so that the mean and variance would be 0 and 1, respectively. To calculate the score for a signature, we averaged the expression values for the genes in the signature after applying a weight of 1 for the up-regulated genes and −1 for the down-regulated ones. This yielded a signature score for each sample in the data set.

Patient gene expression- and clinical data. We obtained the raw CEL files, together with patients' characteristics, from the following public databases - GSE20685[52], GSE6532_GPL96[53], GSE7390[54], GSE11121[55], GSE1456[56], GSE12276[57], GSE4922_UPPSALA[58], and GSE21653[59]. We then preprocessed the CEL files with RMA[60].

Survival analysis. Outcome data included recurrence-free survival (RFS) and distant metastasis-free survival (DMFS) analyses. Kaplan-Meier plots were used for the survival analyses. To quantify the differences in outcome between two groups of patients, we compared the differences in time before 10% of the patients acquired the outcome (either RFS or DMFS), because of the relatively low frequency of these events in cancer data sets. We calculated the p-values using a log rank test.

Gene network enrichment analyses. QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity) was used to identify enriched functional gene networks among differentially regulated transcripts in the E6.5 signature. The full gene list (Suppl Table-1) that resulted from the significance analysis of the microarrays was used for the IPA analysis. P-values were calculated by IPA using a right-tailed Fisher Exact test (a cutoff of $p < 0.05$ was considered significant).

References

1. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *Journal of Clinical Investigation* **119,** 1420–1428 (2009).
2. Mani, S. A. *et al.* The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. *Cell* **133,** 704–715 (2008).
3. Eastham, A. M. *et al.* Epithelial-mesenchymal transition events during human embryonic stem cell differentiation. *Cancer Research* **67,** 11254–11262 (2007).
4. Zheng, H. & Kang, Y. Multilayer control of the EMT master regulators. *Oncogene* **33,** 1755–1763 (2014).
5. Craene, B. D. & Berx, G. Regulatory networks defining EMT during cancer initiation and progression. *Nature Reviews Cancer* **13,** 97–110 (2013).
6. Thiery, J. P. & Sleeman, J. P. Complex networks orchestrate epithelial-mesenchymal transitions. *Nature reviews Molecular cell biology* **7,** 131–142 (2006).
7. Moody, S. E. *et al.* The transcriptional repressor Snail promotes mammary tumor recurrence. *Cancer Cell* **8,** 197–209 (2005).
8. Taube, J. H. *et al.* Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 15449–15454 (2010).
9. Ocaña, O. *et al.* Metastatic Colonization Requires the Repression of the Epithelial-Mesenchymal Transition Inducer Prrx1. *Cancer Cell* **22,** 709–724 (2012).
10. Tsai, J. H., Donaher, J. L., Murphy, D. A., Chau, S. & Yang, J. Spatiotemporal regulation of epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. *Cancer Cell* **22,** 725–736 (2012).
11. Thiery, J. P., Acloque, H., Huang, R. Y. J. & Nieto, M. A. Epithelial-Mesenchymal Transitions in Development and Disease. *Cell* **139,** 871–890 (2009).
12. Thiery, J. P. & Lim, C. T. Tumor dissemination: An EMT affair. *Cancer Cell* **23,** 272–273 (2013).
13. Al-Dhfyan, A. Embryonic signature in breast cancers; Pluripotency roots of cancer stem cells. *Saudi pharmaceutical journal : SPJ : the official publication of the Saudi Pharmaceutical Society* **21,** 229–232 (2013).
14. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England journal of medicine* **351,** 2817–2826 (2004).
15. Palmer, N. P., Schmid, P. R., Berger, B. & Kohane I. S. A gene expression profile of stem cell pluripotentiality and differentiation is conserved across diverse solid and hematopoietic cancers. *Genome biology* **13,** R71 (2012).
16. Hassan, K. A., Chen, G., Kalemkerian, G. P., Wicha, M. S. & Beer, D. G. An embryonic stem cell-like signature identifies poorly differentiated lung adenocarcinoma but not squamous cell carcinoma. *Clin Cancer Res* **15,** 6386–6390 (2009).
17. Yin, Z. Q. *et al.* A 41-gene signature derived from breast cancer stem cells as a predictor of survival. *Journal of experimental & clinical cancer research : CR* **33,** 49 (2014).
18. Tan, T. Z. *et al.* Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol Med* **6,** 1279–1293 (2014).
19. Karlsson, J. *et al.* Clear cell sarcoma of the kidney demonstrates an embryonic signature indicative of a primitive nephrogenic origin. *Genes, chromosomes & cancer* **53,** 381–391 (2014).
20. Zvelebil, M. *et al.* Embryonic mammary signature subsets are activated in Brca1-/- and basal-like breast cancers. *Breast Cancer Res* **15,** R25 (2013).
21. Kim, J. & Orkin, S. H. Embryonic stem cell-specific signatures in cancer: insights into genomic regulatory networks and implications for medicine. *Genome medicine* **3,** 75 (2011).
22. Patsialou, A. *et al.* Selective gene-expression profiling of migratory tumor cells *in vivo* predicts clinical outcome in breast cancer patients. *Breast Cancer Res* **14,** R139 (2012).
23. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).
24. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* **12,** R68 (2010).
25. Prat, A. & Perou, C. M. Deconstructing the molecular portraits of breast cancer. *Mol Oncol* **5,** 5–23 (2011).
26. Baird, R. D. & Caldas, C. Genetic heterogeneity in breast cancer: the road to personalized medicine? *BMC Med* **11,** 151 (2013).

27. Hay, E. D. Role of cell-matrix contacts in cell migration and epithelial-mesenchymal transformation. *Cell differentiation and development : the official journal of the International Society of Developmental Biologists* **32,** 367–375 (1990).
28. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome biology* **10,** R130 (2009).
29. Yilmaz, M. & Christofori, G. Mechanisms of motility in metastasizing cells. *Molecular cancer research : MCR* **8,** 629–642 (2010).
30. Trimboli, A. J. *et al.* Direct evidence for epithelial-mesenchymal transitions in breast cancer. *Cancer Res* **68,** 937–945 (2008).
31. Brabletz, T. *et al.* Variable beta-catenin expression in colorectal cancers indicates tumor progression driven by the tumor environment. *Proc Natl Acad Sci U S A* **98,** 10356–10361 (2001).
32. Chikaishi, Y., Uramoto, H. & Tanaka, F. The EMT status in the primary tumor does not predict postoperative recurrence or disease-free survival in lung adenocarcinoma. *Anticancer Res* **31,** 4451–4456 (2011).
33. Celià-Terrassa, T. *et al.* Epithelial-mesenchymal transition can suppress major attributes of human epithelial tumor-initiating cells. *Journal of Clinical Investigation* **122,** 1849–1868 (2012).
34. Gunasinghe NPAD, Wells A., Thompson, E. W. & Hugo, H. J. Mesenchymal-epithelial transition (MET) as a mechanism for metastatic colonisation in breast cancer. *Cancer and Metastasis Reviews* **31,** 469–478 (2012).
35. Chaffer, C. L. *et al.* Mesenchymal-to-epithelial transition facilitates bladder cancer metastasis: role of fibroblast growth factor receptor-2. *Cancer Res* **66,** 11271–11278 (2006).
36. Brabletz, T. To differentiate or not-routes towards metastasis. *Nature Reviews Cancer* **12,** 425–436 (2012).
37. Brabletz, T. EMT and MET in Metastasis: Where Are the Cancer Stem Cells? *Cancer Cell* **22,** 699–701 (2012).
38. Liu, S. *et al.* Breast cancer stem cells transition between epithelial and mesenchymal states reflective of their normal counterparts. *Stem cell reports* **2,** 78–91 (2014).
39. Rousseaux, S. *et al.* Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci Transl Med* **5,** 186ra166 (2013).
40. Singhai, R. *et al.* E-Cadherin as a diagnostic biomarker in breast cancer. *North American journal of medical sciences* **3,** 227–233 (2011).
41. Gamallo, C. *et al.* Correlation of E-cadherin expression with differentiation grade and histological type in breast carcinoma. *The American journal of pathology* **142,** 987–993 (1993).
42. Moll, R., Mitze, M., Frixen, U. H. & Birchmeier, W. Differential loss of E-cadherin expression in infiltrating ductal and lobular breast carcinomas. *The American journal of pathology* **143,** 1731–1742 (1993).
43. Siitonen, S. M. *et al.* Reduced E-cadherin expression is associated with invasiveness and unfavorable prognosis in breast cancer. *Am J Clin Pathol* **105,** 394–402 (1996).
44. Swiatoniowski, G. *et al.* E-cadherin and fibronectin expressions have no prognostic role in stage II ductal breast cancer. *Anticancer Res* **25,** 2879–2883 (2005).
45. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–536 (2002).
46. Glas, A. M. *et al.* Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics* **7,** 278 (2006).
47. Mittempergher, L. *et al.* A gene signature for late distant metastasis in breast cancer identifies a potential mechanism of late recurrences. *Mol Oncol* **7,** 987–999 (2013).
48. Liu, R. *et al.* The prognostic role of a gene signature from tumorigenic breast-cancer cells. *The New England journal of medicine* **356,** 217–226 (2007).
49. Leth-Larsen, R. *et al.* Functional heterogeneity within the CD44 high human breast cancer stem cell-like compartment reveals a gene signature predictive of distant metastasis. *Mol Med* **18,** 1109–1121 (2012).
50. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **42,** D7–17 (2014).
51. Smyth, G. K. limma: Linear Models for Microarray Data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (ed^(eds Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A., Dudoit, S). Springer (2005).
52. Kao, K. J., Chang, K. M., Hsu, H. C. & Huang, A. T. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC Cancer* **11,** 143 (2011).
53. Loi, S. *et al.* Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **25,** 1239–1246 (2007).
54. Desmedt, C. *et al.* Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res* **13,** 3207–3214 (2007).
55. Schmidt, M. *et al.* The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res* **68,** 5405–5413 (2008).
56. Pawitan, Y. *et al.* Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res* **7,** R953–964 (2005).
57. Bos, P. D. *et al.* Genes that mediate breast cancer metastasis to the brain. *Nature* **459,** 1005–1009 (2009).
58. Ivshina, A. V. *et al.* Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res* **66,** 10292–10301 (2006).
59. Sabatier, R. *et al.* A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Res Treat* **126,** 407–420 (2011).
60. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31,** e15 (2003).

Acknowledgments

This work is supported by NIH/NCI CA155243-01 (SAM). JTC is supported by NIH R00LM009837.

Author Contributions

R.S. coordinated the research project, performed data analyses, and wrote the manuscript A.N.P. performed data analyses, and wrote the manuscript V.B. performed data analyses J.T.C. designed the research project, performed data analyses, and wrote the manuscript S.A.M. developed the concept, designed the research project, wrote the manuscript, and supervised the research project

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors of this manuscript are inventors of a patent application in part based on findings described in this manuscript.

How to cite this article: Soundararajan, R. *et al.* A novel embryonic plasticity gene signature that predicts metastatic competence and clinical outcome. *Sci. Rep.* **5**, 11766; doi: 10.1038/srep11766 (2015).