Breast Cancer Res Treat
DOI 10.1007/s10549-011-1619-7

PRECLINICAL STUDY

Genomic analysis identifies unique signatures predictive of brain, lung, and liver relapse

J. Chuck Harrell · Aleix Prat · Joel S. Parker · Cheng Fan · Xiaping He · Lisa Carey · Carey Anders · Matthew Ewend · Charles M. Perou

Received: 27 May 2011 / Accepted: 28 May 2011

Abstract The ability to predict metastatic potential could be of great clinical importance, however, it is uncertain if predicting metastasis to specific vital organs is feasible. As a first step in evaluating metastatic predictions, we analyzed multiple primary tumors and metastasis pairs and determined that >90% of 298 gene expression signatures were found to be similarly expressed between matched pairs of tumors and metastases; therefore, primary tumors may be a good predictor of metastatic propensity. Next, using a dataset of >1,000 human breast tumor gene expression microarrays we determined that HER2-enriched subtype tumors aggressively spread to the liver, while basal-like and claudin-low subtypes colonize the brain and lung. Correspondingly, brain and lung metastasis signatures, along with embryonic stem cell, tumor initiating cell, and hypoxia signatures, were also strongly expressed in the basal-like and claudin-low tumors. Interestingly, low "Differentiation Scores," or high expression of the aforementioned signatures, further predicted for brain and lung metastases. In total, these data identify that depending upon the organ of relapse, a combination of gene expression signatures most accurately predicts metastatic behavior.

Keywords Breast cancer · Gene expression · Intrinsic subtype · Metastasis · Microarray

Electronic supplementary material The online version of this article (doi:10.1007/s10549-011-1619-7) contains supplementary material, which is available to authorized users.

J. C. Harrell · A. Prat · J. S. Parker · C. Fan · X. He · C. M. Perou (✉)
Lineberger Comprehensive Cancer Center, University of North Carolina, 450 West Drive, CB7295, Chapel Hill, NC 27599, USA
e-mail: cperou@med.unc.edu

J. C. Harrell · A. Prat · J. S. Parker · C. Fan · X. He · C. M. Perou
Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA

J. C. Harrell · A. Prat · J. S. Parker · C. Fan · C. M. Perou
Department of Pathology & Laboratory Medicine, University of North Carolina, Chapel Hill, NC 27599, USA

A. Prat
Department of Medicine, Universitat Autonoma de Barcelona, Passeig Vall d'Hebron 119, 08035 Barcelona, Spain

L. Carey · C. Anders
Division of Hematology/Oncology, Department of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA

M. Ewend
Department of Neurosurgery, University of North Carolina, Chapel Hill, NC 27599, USA

Introduction

The vast majority of deaths due to breast cancer for nearly half a million people annually worldwide are due to distant metastases in the lung, liver, and brain [1]. Numerous studies have focused on breast cancer metastases and how they might differ from primary breast tumors; however, controversy remains regarding (A) the predisposition of specific classes of breast tumors to spread to distant sites and (B) the degree of similarity between primary breast tumors and their associated metastases.

Estrogen receptor (ER) status is known to be associated with breast cancer relapse in specific organs [2]. In 2008, this organ selectivity was refined by contrasting relapse patterns in 344 patients who had their tumors genomically subtyped as luminal A or B, HER2-enriched, basal-like, or

Published online: 14 June 2011

normal-like [3]. In general, bone metastases were associated with the luminal subtypes, whereas basal-like and HER2-enriched tumors were significantly associated with brain and lung relapse. Similar results were also observed in an immunohistochemical-based study on 3,726 patients [4]. Recently, a new breast cancer subtype was identified, named claudin-low [5–7]. This subtype exhibits aggressive characteristics including expression of mesenchymal markers and low expression of genes involved in tight junctions and cell–cell adhesion. The lack of epithelial cell features and expression of mesenchymal traits is reminiscent of features associated with breast stem cells [8]. Since breast cancer stem cells are relatively resistant to both chemotherapy and radiation [9, 10], and because metastases frequently progress despite treatment, it is important to determine if these claudin-low/mesenchymal cells are associated with metastatic potential.

To better understand the biology driving breast cancer metastases, 1,319 human gene expression microarrays from primary tumors, metastases, and cancer cell lines were analyzed here. Tumors and their associated metastases, on average, were much more similar to each other than they were different. By including the recently defined claudin-low subtype we extend previous findings [3, 4] and better define the metastatic predilections of each intrinsic subtype. Increasingly "undifferentiated" breast cancer cells [as quantitatively measured by a Differentiation Score predictor (DS)] tend to express stem cell signatures and preferentially metastasize to the brain and lung. These results identify that breast cancer intrinsic subtype is maintained throughout disease progression, and that a combination of several genomic signatures can add prognostic value and therefore direct where disease monitoring should be focused.

Results

Genetic similarity among tumors and metastases

Previously, we examined the genome-wide gene expression profiles of five primary breast tumor/matched metastatic pairs and noted an overall high degree of similarity within a pair [11]. To further examine the degree of relatedness of breast tumors and their metastases, we performed correlation analysis using thousands of genes, and hundreds of predefined gene expression signatures/modules [12] incorporating a large set of tumors and paired metastases. Intraclass correlation (ICC) values were determined between pairs of samples using multiple classification/grouping methods: (1) different pieces of the same primary tumor ("intrinsic pairs"), (2) tumors and their matched metastases [all metastases, or further separated into either lymph node (LN) or distant], (3) tumors and their matched metachronous metastases, (4) sets of synchronous metastases from the same patient, (5) tumors from different patients grouped by intrinsic subtype, and (6) metastases from different patients (Fig. 1a). On average when using all expressed genes, there was high concordance between two pieces of the same primary tumor (ICC = 0.9 [0.89–0.91]), while pairs of tumors and their metastases exhibit lower concordance values (0.82 [0.8–0.83]). As observed by the metachronously paired tumor-metastasis samples, gene expression did not change substantially over time. The autopsy patient data (0.72 [0.68–0.75]) suggest that normal organ RNA may be the variable most responsible for the decreased similarity between tumor and metastasis pairs. This hypothesis was supported by increased ICC values of 20 matched pairs of laser-captured tumors and LN metastases [13] (0.9 [0.85–0.94]).

Individual gene measurements can be fraught with "noise." Thus, to further test the relationship between tumors and metastases, ICC values were identified using a compendium of 298 different gene expression signatures/modules [12], where each module is a summary measure of tens to hundreds of genes. The overall ICC values were higher than individual genes (thus showing greater robustness for gene signatures) and the breast tumor–metastasis pairs showed high conservation of pathways (Fig. 1b). The signatures with the most variability between tumors and matched metastases were associated with extracellular matrix (ECM) proteins. These genes may be microenvironment-induced or may be due to different amounts of fibroblasts found in tumors as compared to metastases (Supplemental Table 1).

Association of subtypes and sites of metastasis

Since the majority of genes maintain their RNA expression levels when growing as either primary tumors in the breast or as metastases, we sought to determine if the different intrinsic subtypes showed a predilection for metastasis to specific organs using genomic data arising from primary tumors only. Therefore, we combined four public microarray datasets with Distance Weighted Discrimination [14], providing 855 tumors with documented first site of relapse (Supplemental Table 2) [15–18]. Principal components analysis found that the overall variation of gene expression was due to the biology of the tumors, and not by cohort/source or microarray platform (Supplemental Fig. 1). Status for ER, progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) was recorded for 852, 537, and 499 tumors, respectively, and of the 482 tumors with defined status for all three markers, 110 were triple negative (TN); Kaplan–Meier analyses for site of relapse with these markers are shown in Supplemental Fig. 2. For all sites of relapse, ER/PR negativity was

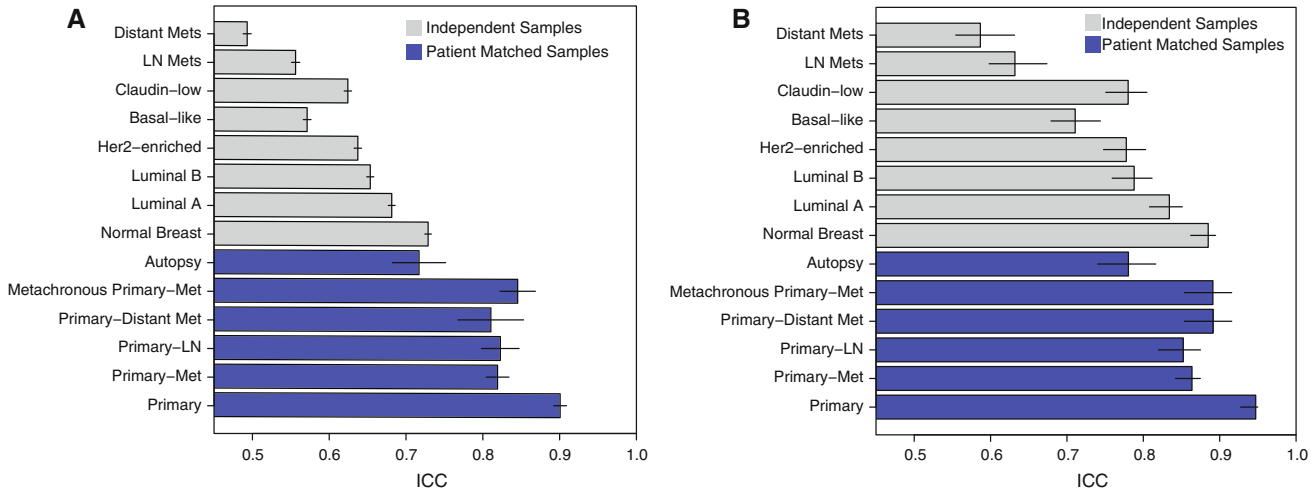


Fig. 1 Genomic similarity of breast tumors and metastases. Microarrays were performed on 265 primary tumors and 85 metastases and the overall similarity was measured by intra-class correlation (ICC), with estimates plotted showing 95% confidence intervals. **a** Using all variably expressed genes, gene expression concordance values were measured in matched samples from the same patient; primary tumors split in 2 ($n = 40$), tumor-metastasis pairs ($n = 34$), tumor–LN metastasis pairs ($n = 24$), tumor-distant metastasis ($n = 10$), autopsy patient metastases from multiple organs within the same patient ($n = 33$), metachronous tumor–metastasis pairs ($n = 10$), or from independent patient samples; normal breast ($n = 17$), luminal A tumors ($n = 86$), luminal B tumors ($n = 50$), HER2-enriched tumors ($n = 25$), basal-like tumors ($n = 44$), claudin-low tumors ($n = 45$), LN metastases ($n = 21$), and distant metastases ($n = 45$). **b** ICC of 298 gene expression signatures/modules [12] using the same samples and pairing used in (**a**)

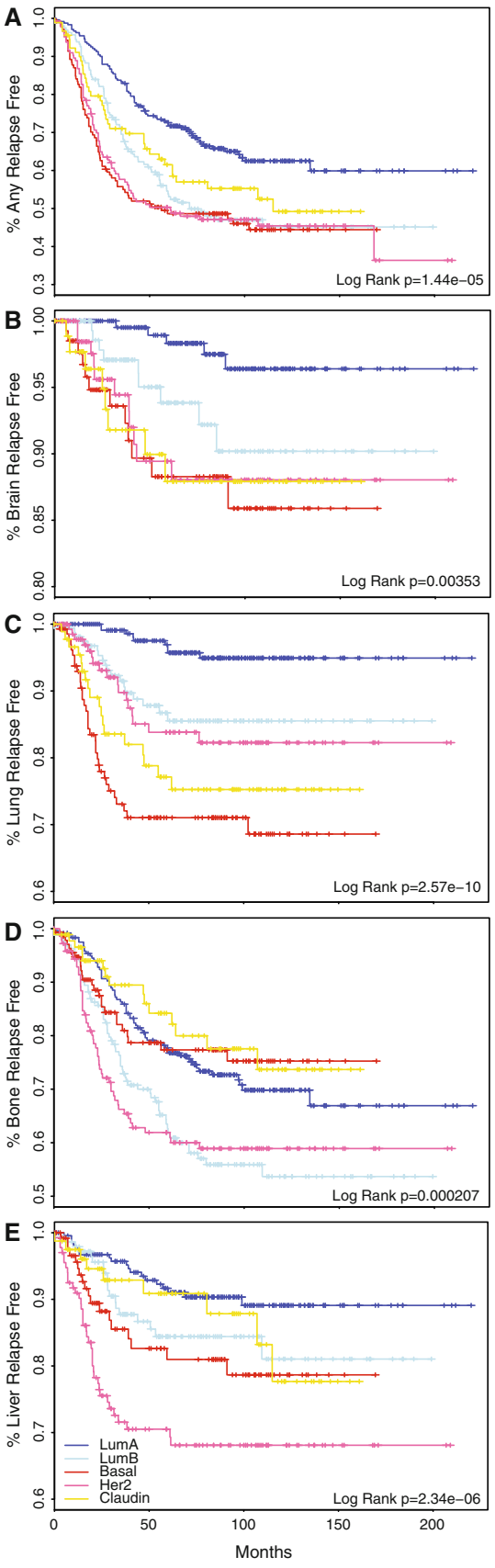
associated with increased metastases, except for bone, in which both ER+ and ER− tumors recurred. Clinical HER2+ and TN status were associated with liver and brain/lung relapse, respectively.

Next, each tumor's intrinsic subtype was calculated for this combined data set using the PAM50 [19] and the claudin-low subtype predictors [6] (Supplemental Table 3). Of the 855 tumors, 76 were identified as normal breast-like, and since this tumor classification is reflective of mostly normal breast tissue [19], these tumors/samples were excluded from further analyses, leaving a dataset of 779 tumors. Based on the site of first relapse data for liver, lung, brain, and bone, Kaplan–Meier plots were generated, and we determined that intrinsic subtype was correlated with site of relapse (Fig. 2, Supplemental Fig. 3). Compared to luminal A, basal-like and HER2-enriched tumors showed the highest hazard ratio (HR) of relapse to any site (basal-like vs. luminal A hazard ratio [HR] 2.1, $P < 0.0001$; HER2-enriched vs. luminal A HR 2.0, $P < 0.0001$) followed by luminal B (HR 1.69, $P < 0.001$) and claudin-low (HR 1.47, $P = 0.051$) tumors. Important findings included: (1) bone metastasis was the most common—regardless of subtype (Table 1), (2) brain relapse occurred most frequently in non-luminal samples, (3) liver relapse was associated with HER2-enriched tumors, and (4) lung relapse occurred often within the claudin-low and basal-like subtypes. In all analyses, luminal B tumors were more metastatic than luminal A tumors, thus providing a useful stratification within ER+ tumors.

Undifferentiated tumors and brain metastases

In 2009, Bos et al. [16] utilized two human breast cancer cell lines, CN34 and variants of the MDA-MB-231 human breast cancer cell line (a claudin-low cell line[6]), along with gene expression data from human breast tumors, to identify 17 genes whose expression correlated with brain relapse (BrMS). Given the clear associations observed for the intrinsic subtypes and sites of metastases, we hypothesized that the BrMS would correlate with basal-like and/or claudin-low subtypes. ANOVA from two different datasets supported this hypothesis (Fig. 3a, b). A lung metastasis signature (LMS) [20] is also associated with intrinsic subtype (Fig. 3c, d).

Recently, a genomic method to quantify breast epithelial cell differentiation status, known as the Differentiation Score (DS) predictor [6] was developed. This predictor is based on the genomic signatures of FACS purified populations of mammary stem cells, luminal progenitors, and mature luminal cells of the normal human breast [8]. The scoring of the DS predictor is based on the premise that mammary stem cells are the least differentiated cells in the breast and they give rise to luminal progenitors, which then produce mature luminal cells; for the DS, higher scores represent greater differentiation along this axis that starts with the mammary stem cell signature and culminates in mature ER+ luminal cells. In this spectrum, claudin-low tumors are the least differentiated, followed by basal-like, HER2-enriched, and ending with luminal B and A tumors



◀**Fig. 2** Association of breast cancer subtype with site of first relapse. Shown are Kaplan–Meier plots and log rank tests of first site of relapse in each breast tumor subtype in the 779 tumor dataset. If a patient showed two or more simultaneous sites of relapse, then this patient was counted as being site of first relapse for both. Organ of first relapse; **a** any, **b** brain, **c** lung, **d** bone, **e** liver

[6]. Since claudin-low and basal-like tumors were associated with brain relapse, we postulated that the more undifferentiated a tumor is on this axis, the more likely it would be to metastasize to the brain. To test this hypothesis, gene expression data from parental and organ-tropic (brain, lung, and bone) MDA-MB-231 cell lines were obtained from the Gene Expression Omnibus, and their DS calculated and plotted on the DS axis (Fig. 4a). Shown on the same scale are the 779 breast tumor dataset (Fig. 4b), cancer cell lines of various tissue origins (NCI60) [21] (Fig. 4c), and the MDA-MB-231 series [16, 20, 22] (Fig. 4d). Overall, claudin-low and luminal breast cancer cells lines show the same relative differences in differentiation status as is seen in primary tumors. Importantly, the MDA-MB-231 cells from the NCI60 and Massagué studies showed nearly identical DS, and the brain-tropic MDA-MB-231 cells were significantly less differentiated than the parental cell line.

To identify other features shared between low DS tumors and brain metastasis, we analyzed the NCI60 [21] cell line series. Interestingly, DS were found to be similar in claudin-low breast cancer cell lines, central nervous system (CNS), and melanoma cell lines, a tumor type known to aggressively spread to the brain[23] (Fig. 4c). To identify genes that mediate cerebral colonization, significance analysis of microarrays (SAM) was performed on the NCI60 data set by comparing these three cancer cell line types versus the rest. Two-hundred and sixty-five genes were identified as being highly expressed (FDR = 0%) in claudin-low, CNS and melanoma cell lines; Ingenuity Systems Pathway Analysis found that "cellular movement" was the top biological function associated with these genes (Supplemental Fig. 4).

The triple-negative SUM149PT breast tumor-derived cell line contains two distinct populations of breast cancer cells[24], which can be separated by FACS to yield one population with basal-like and another with claudin-low-like features and a lower DS [6]. To test if lower DS correlates with increased migration, we fluorescence-activated cell sorted (FACS) the SUM149PT cell line into $CD49f^{+}/Epcam^{-/low}$ and $CD49f^{+/high}/Epcam^{+}$ subpopulations, performed Boyden chamber migration assays, and determined that the less differentiated (i.e., lower DS) SUM149PT $CD49f^{+}/Epcam^{-/low}$ cells were significantly ($P < 0.001$) more migratory than the more differentiated $Epcam^{+}$ population (Supplemental Fig. 5).

Table 1 Site of first relapse of the 779 tumors from each cohort according to intrinsic subtype

Cohort	Subtype	# of tumors	% that relapsed	Site of first relapse (%)				
				Brain	Lung	Bone	Liver	LN
EMC192	Basal	40	90.0	8.3	41.7	30.6	19.4	NA
	Claudin-low	23	73.9	17.6	41.2	35.3	17.6	NA
	HER2	32	100.0	9.4	18.8	62.5	59.4	NA
	Luminal A	57	89.5	2.0	7.8	76.5	31.4	NA
	Luminal B	31	90.3	3.6	17.9	71.4	14.3	NA
EMC286	Basal	45	37.8	23.5	47.1	41.2	17.6	NA
	Claudin-low	32	28.1	22.2	33.3	44.4	22.2	NA
	HER2	54	38.9	9.5	14.3	76.2	28.6	NA
	Luminal A	72	22.2	0.0	18.8	87.5	0.0	NA
	Luminal B	49	46.9	8.7	34.8	87.0	13.0	NA
MSK82	Basal	17	29.4	20.0	100.0	40.0	NA	NA
	Claudin-low	10	50.0	20.0	100.0	40.0	NA	NA
	HER2	10	20.0	50.0	50.0	50.0	NA	NA
	Luminal A	23	30.4	14.3	14.3	57.1	NA	NA
	Luminal B	16	18.8	0.0	0.0	100.0	NA	NA
NKI295	Basal	38	36.8	28.6	42.9	35.7	57.1	42.9
	Claudin-low	25	28.0	28.6	42.9	42.9	57.1	0.0
	HER2	48	43.8	23.8	33.3	71.4	57.1	28.6
	Luminal A	91	11.0	30.0	10.0	70.0	30.0	30.0
	Luminal B	66	40.9	22.2	18.5	74.1	44.4	25.9
Combined	Basal	140	51.4	16.7	47.2	34.7	28.1	42.9
	Claudin-low	90	42.2	21.1	47.4	39.5	23.1	0.0
	HER2	144	52.8	14.5	22.4	68.4	59.7	28.6
	Luminal A	243	34.6	6.0	10.7	76.2	23.5	30.0
	Luminal B	162	50.0	11.1	22.2	77.8	22.6	25.9
	Any subtype	779	45.1	12.8	27.4	62.7	29.1	27.8

Several tumors had multiple first relapses: basal-like 22/69, claudin-low 12/44, HER2-enriched 35/64, luminal A 19/88, luminal B 31/86, and these were thus counted as being sites of first relapse for each site

Differentiation Scores and metastasis

We next sought to better understand the information that DS provides for predicting site of metastasis. Since there is a range of differentiation within each intrinsic subtype (Fig. 4b), we tested if the least differentiated basal-like/claudin-low tumors were more metastatic than the more differentiated basal-like/claudin-low tumors. Kaplan–Meier analysis and log-rank tests determined that the least differentiated half of these tumor subtypes were associated with significantly more relapse to brain ($P = 2E{-}03$, log rank-test) and lung ($P = 2.4E{-}02$). This same approach applied within luminal and HER2-enriched tumors found no association of DS with bone or liver relapse, thus this association appears specific for brain and lung relapses, although it should be noted that the least differentiated luminal and HER2-enriched tumors do not have low overall DS.

To visualize the information that DS and intrinsic subtypes provide for predicting site of metastasis, we plotted the DS of the 779 tumors versus the HR for each site of metastasis (Fig. 5a). The tumors were then ordered based on DS and all genes (11,068) hierarchical clustered (Fig. 5b). Interestingly, tumors with the lowest DS have a much higher HR for brain and lung metastases, and this risk drops off quickly as differentiation increases. Importantly, this analysis identified a subset of tumors within the largely ER− claudin-low and basal-like tumors that aggressively metastasize.

Stem cell signatures correlate with brain and lung metastases

Several studies have shown an association of stem cell characteristics and metastatic proclivity [25–27]. Therefore, the 855 tumor dataset was used to test if several

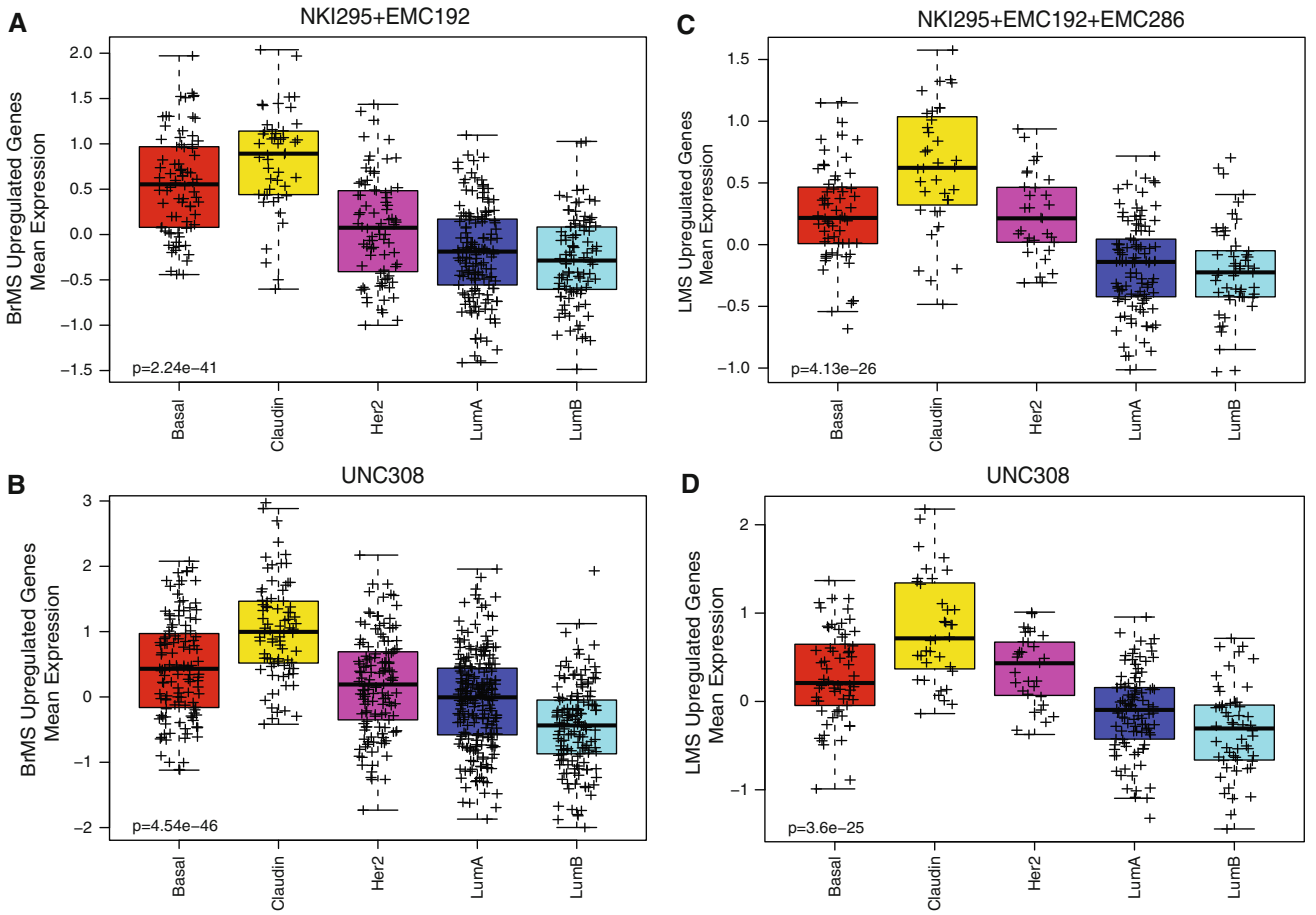


Fig. 3 Association of the brain (BrMS) and lung (LMS) cell line-based metastasis signatures with intrinsic subtype. *Box*-and-*whisker* plots are shown for each signature on multiple breast tumor microarray data sets according to intrinsic subtype. *P* values were calculated with ANOVA. Shown are the same data sets used for the testing of the BrMS (**a**) or LMS (**c**) signatures, as well as an independent UNC dataset (**b**, **d**)

previously published stem cell signatures contained within our set of 298 gene modules [12] were associated with site of relapse. Univariate Cox proportional hazards models identified that many of the signatures with the strongest associations for brain (and lung) relapse were either expressed in normal brain and/or have been identified as essential components of embryonic stem cells and tumor initiating cells [26, 27] (Supplemental Table 4). Of the 13 embryonic stem cell signatures analyzed in Ben-Porath et al. [27], all were significantly associated with relapse to brain/lung, 11 with LN metastasis, 10 with liver, and 5 with bone. Nearly all the signatures that predicted for brain relapse correlated with low DS, and those not strongly correlated with DS were correlated with proliferation. Some of these signatures further identified subsets of basal-like and claudin-low tumors most likely to metastasize to the brain (log-rank test: PRC2_targets; $P = 0.0090$, MM_WapINT3; $P = 0.0001$). Thus, ES cell signatures, DS, and proliferation appear to be strong predictors of CNS and lung metastases, and in general, the signatures most predominant for brain/lung relapse were weakly expressed in tumors that spread to the bone.

Univariate and multivariable survival analyses

The ability to predict the presence and/or location of a tumor recurrence could influence the location and frequency of radiographic surveillance for patients with a history of breast cancer. Therefore, we sought to identify the most informative signature, or combination of signatures that predicts metastasis to specific sites. First, we performed univariate survival analyses for multiple signatures, including the many described above and our previously published VEGF/hypoxia signature [28]. As shown in Table 2A, all signatures tested were highly prognostic overall and, interestingly, both BrMS and LMS signatures predicted lung and brain relapse, providing evidence that metastases to these two organs utilize similar genetic

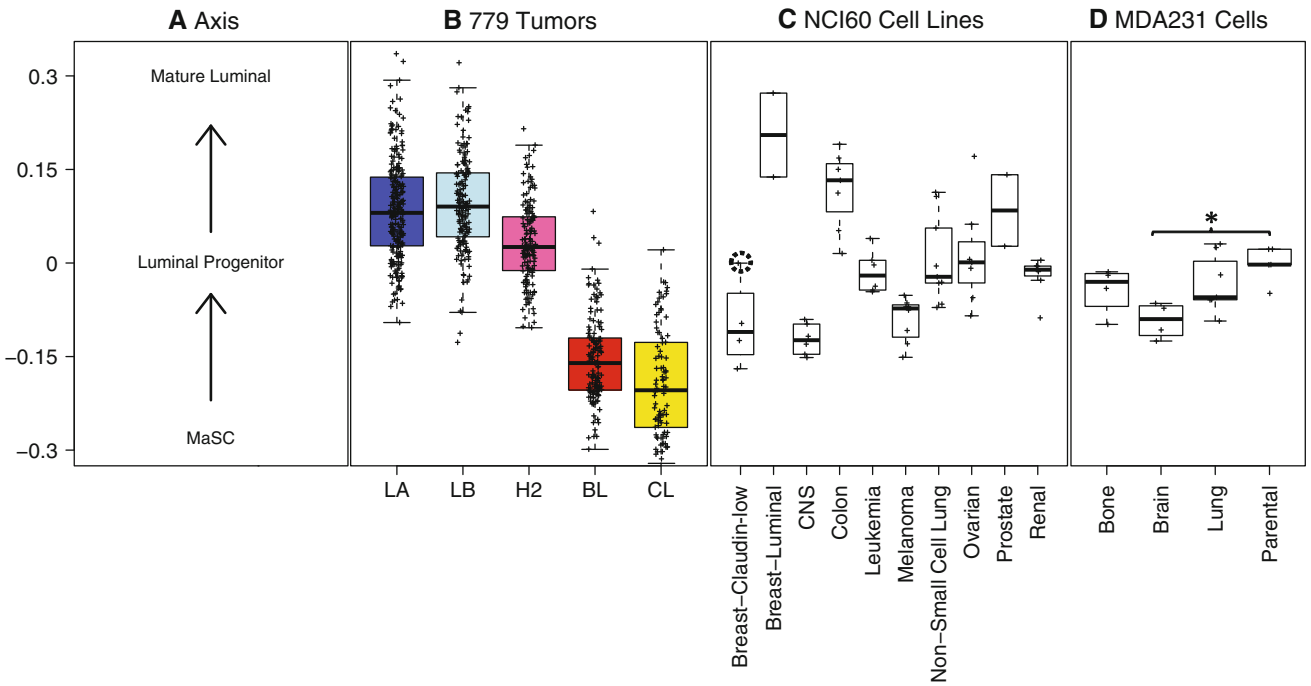


Fig. 4 Differentiation Score analysis of the 779 human breast tumors, NCI60 cell lines, and MDA-MB-231 cell lines. **a** Differentiation axis diagram based on FACS fractions Lim et al. [8], which is described in Prat et al. [6]. **b** *Box*-and-*whisker* plots of the distributions of scores from the 779 tumor dataset according to intrinsic subtype. **c** NCI60 cancer cell lines gene DS values [21], with the breast cancer cell lines divided into claudin-low (*dashed circle* value for MDA-MB-231) or luminal cell lines. **d** MDA-MB-231 parental, lung-tropic, brain-tropic, and bone-tropic cell lines from the studies of Massagué and colleagues. The *asterisk* indicates statistical significance difference in DS between parental and brain-tropic lines (T test $P = 0.002$)

mechanisms. Second, we performed multivariate analysis using the backward stepwise procedure and observed that subtype information (i.e., subtype calls or risk of relapse categories based on subtype [ROR-S]) was selected in each evaluation (Table 2B). For liver relapse, specifically, knowing the subtype call instead of the ROR-S risk category was found particularly informative; indeed, the risk of liver relapse of the HER2-enriched subtype was 4.0 times higher compared to the luminal A subtype despite that the HER2 status (as determined by gene expression) was also included. In addition to intrinsic subtype information, other signatures were found statistically significant in the various MVA final models, such as the upregulated genes of the BrMS in brain relapse, or the VEGF/hypoxia signature and the downregulated genes of the LMS in lung relapse. Interestingly, the BrMS and VEGF/hypoxia-signature were found highly correlated with DS (Pearson $= -0.68$), and correspondingly, the BrMS, DS, and VEGF/hypoxia-signature identify a subset of basal-like/claudin-low tumors that spread to the brain ($P < 0.05$). Thus, when each metastatic site is individually examined, a unique combination of signatures is chosen that includes intrinsic subtype (individual subtype or ROR-S) as well as another signature or two, ultimately resulting in the optimal set of variables for predicting relapse to that organ.

Discussion

Metastases are the main cause of death for breast cancer patients and predicting a tumor's likelihood to spread, and organ of relapse, is clinically important information. Analysis of 265 breast tumors and 85 metastases found that a breast tumor's overall gene expression phenotype is largely maintained in its metastases. The gene expression differences that do occur may be due to a combination of different amounts of epithelial/stromal cells (Fig. 1, Supplemental Table 1), and/or clonal expansion of a more aggressive subclone of a tumor [4, 29]. The microenvironment also effects gene expression and response to therapeutics [30], therefore, targeting the host organ cells, vascular cells, as well as tumor cell specific targets may be the best approach to inhibit disease progression [31]. This overall similarity, however, does suggest that important information about metastatic potential can be revealed by studying primary tumors.

Basal-like and claudin-low breast cancers both exhibit a high probability to metastasize to the brain and lung while HER2-enriched subtype tumors preferentially colonize the liver (Fig. 2; Table 1). The basal-like and claudin-low tumor types are genomically related [6], exhibit similar treatment response characteristics, and as shown here, have

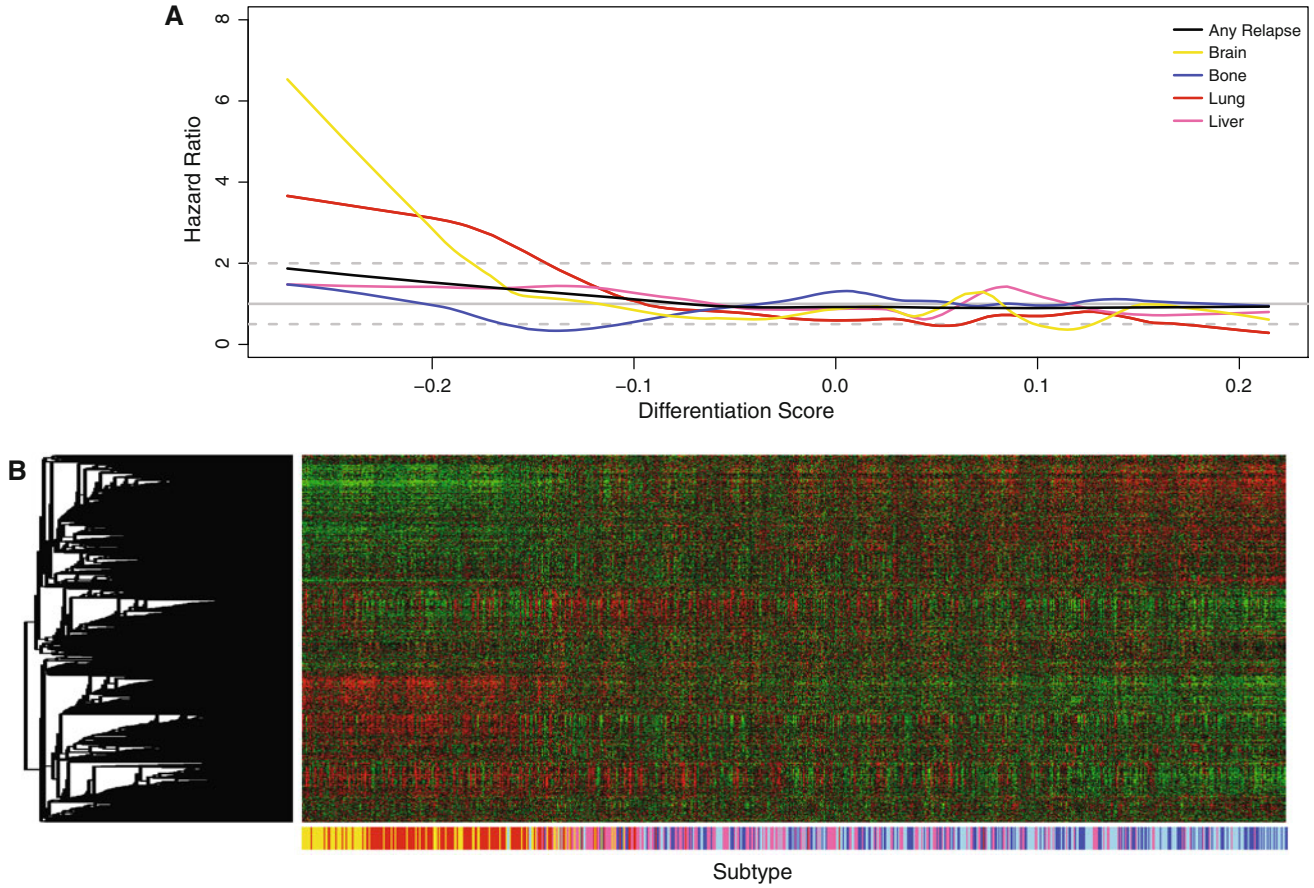


Fig. 5 Relationship of Differentiation Score, breast cancer subtype, and likelihood of site of metastasis. 779 tumors with known first site of relapse were ordered based on low to high DS. **a** Hazard ratios for each site of metastasis were estimated by grouping a sliding window of 50 samples with consecutive DS and contrasting against those outside the window. Estimates were then smoothed with Lowess prior to plotting. **b** Hierarchical clustering of all genes. Below the dendogram is a colored bar identifying the intrinsic subtype of each tumor (*yellow* claudin-low, *red* basal-like, *pink* HER2-enriched, *dark blue* luminal A, *light blue* luminal B)

similar metastasis patterns. The $CD49f^{+}/Epcam^{-/low}$ fraction of the SUM149PT cell line (which is enriched for claudin-low tumor features) was significantly more migratory than the more differentiated basal-like component cells. Interestingly over time, the SUM149PT cells with claudin-low characteristics asymmetrically divide into two distinct populations of more (i.e., basal-like) and less-differentiated cells, whereas the more differentiated fraction produces similarly differentiated cells [6]. Since the less-differentiated claudin-low-like cells contain higher levels of genes that facilitate cellular movement (Supplemental Figs. 4, 5), we hypothesize that these cells may initiate the metastatic cascade; after seeding a host organ, they asymmetrically divide, spawning both more and less differentiated cells. Precisely why these cells show predilection for the brain and lung requires further investigation, however, the cell line studies of Massagué and colleagues using the claudin-low MDA-MB-231 cells are providing for some initial candidates. These studies have shown that the cells that are relatively more capable of spreading to the CNS express genes that function to increase cellular extravasation and blood brain barrier penetration [16], while also upregulating glycolytic pathways and increasing vascularization [28].

Our re-analyses of the data presented by Bos et al. [16] find that the DS of brain-tropic breast cancer cells is significantly lower than the parental cell line (Fig. 4); correspondingly, low DS was also found to associate with brain relapse in patients (Fig. 5). While basal-like and claudin-low breast tumors can relapse in bone, recurrence in vital organs, such as the brain and lung is more symptomatic. Thus, first site of recorded relapse for basal-like and claudin-low tumors is typically not bone. DS, however, is not the only factor that determines metastagenicity. For example, luminal A and B tumors have similar DS, yet luminal B tumors are much more likely to relapse. Perhaps all luminal tumors can effectively seed certain organs, however, the faster proliferation rate inherent to luminal B tumors accounts for the differential relapse frequency. Correspondingly, 58% of luminal B tumors present with

Table 2 Breast cancer metastasis-free survival (A) univariate and (B) multivariable analyses among all patients

Variable	All		Brain		Lung		Bone		Liver	
	Hazard ratio	*P* value	Hazard ratio	*P* value	Hazard ratio	*P* value	Hazard ratio	*P* value	Hazard ratio	*P* value
A										
Intrinsic subtype										
LumB vs. LumA	1.7	7.6E−04	3.2	3.6E−02	3.4	2.6E−03	1.7	2.3E−03	1.7	8.9E−02
Basal-like vs. LumA	2.1	1.7E−05	5.8	9.3E−04	9.1	4.3E−09	0.9	7.8E−01	2.3	1.2E−02
HER2-enriched vs. LumA	2.0	1.4E−05	5.0	3.0E−03	4.1	5.7E−04	1.8	1.9E−03	4.1	5.6E−07
Claudin-low vs. LumA	1.5	5.1E−02	5.4	3.2E−03	6.6	3.9E−06	0.8	3.8E−01	1.6	2.8E−01
ROR-S										
Continuous variable	1.4	5.0E−09	1.9	2.4E−04	2.3	1.1E−09	1.3	1.7E−03	1.5	1.1E−04
Intermediate vs. low	1.5	5.5E−03	2.2	1.5E−01	3.1	5.9E−03	1.4	6.4E−02	1.6	1.1E−01
High vs. low	2.2	2.3E−08	5.3	5.5E−04	6.9	2.5E−07	1.7	3.2E−03	2.7	2.8E−04
Differentiation Score	0.9	1.4E−02	0.6	1.5E−04	0.5	4.2E−10	1.2	1.9E−02	0.9	4.2E−01
VEGF-13 hypoxia										
Continuous variable	1.3	6.0E−08	2.1	6.7E−08	2.2	0.0E + 00	1.1	2.9E−01	1.3	2.0E−02
Intermediate vs. low	1.0	8.7E−01	1.2	7.5E−01	2.9	8.4E−02	1.0	8.5E−01	0.9	6.2E−01
High vs. low	1.7	1.3E−03	5.0	8.0E−03	11.0	4.8E−05	1.4	1.3E−01	1.7	9.6E−02
Proliferation Index										
Continuous variable	1.3	1.2E−06	1.7	1.8E−03	1.8	3.1E−07	1.2	5.4E−03	1.3	1.1E−02
High vs. low	1.5	2.5E−04	2.5	6.0E−03	2.5	4.7E−05	1.3	5.9E−02	1.4	1.0E−01
BrMS (up genes)	1.2	2.4E−04	2.2	9.8E−07	2.0	2.3E−10	1.0	9.7E−01	1.2	1.2E−01
BrMS (down genes)	0.8	2.4E−03	0.6	8.9E−04	0.6	1.0E−07	1.0	7.9E−01	0.9	3.8E−01
LMS (up genes)	1.1	2.3E−02	1.6	2.2E−03	1.5	1.3E−05	1.0	7.7E−01	1.2	5.4E−02
LMS (down genes)	0.9	2.5E−01	0.8	2.8E−01	0.8	4.3E−02	0.9	2.8E−01	1.2	1.9E−01
BoMS (up genes)	1.1	3.6E−01	1.2	2.7E−01	1.3	5.5E−03	1.0	8.8E−01	1.1	2.0E−01
BoMS (down genes)	0.9	2.6E−01	1.0	8.3E−01	0.9	4.4E−01	0.9	1.5E−01	1.1	3.1E−01
ER (+ vs. −)	0.6	5.2E−05	0.4	4.4E−04	0.2	2.8E−11	1.0	7.9E−01	0.5	2.8E−04
HER2 (+ vs. −)	1.3	6.2E−02	1.6	1.5E−01	1.3	2.8E−01	1.3	1.1E−01	2.3	7.8E−05

Table 2 continued

Variable	All		Brain		Lung		Bone		Liver	
	Hazard ratio	*P* value	Hazard ratio	*P* value	Hazard ratio	*P* value	Hazard ratio	*P* value	Hazard ratio	*P* value
B										
Intrinsic subtype										
LumB vs. LumA									2.5	1.7E−02
Basal-like vs. LumA									1.4	4.5E−10
HER2-enriche d vs.									4.0	1.3E−04
LumA claudin-low vs. LumA									1.0	9.7E−01
ROR-S										
Intermediate vs. low	1.7	4.0E−03	1.8	3.1E−01	2.3	5.1E−02	1.4	5.4E−02		
High vs. low	2.5	6.2E−05	3.3	1.7E−02	3.2	5.6E−03	2.0	2.9E−04		
Differentiation Score	1.1	9.6E−02					1.8	8.1E−08		
VEGF-13 hypoxia										
Intermediate vs. low	0.9	6.5E−01			2.3	1.7E−01	1.0	9.4E−01	0.6	1.1E−01
High vs. low	1.5	4.0E−02			5.4	6.8E−03	1.9	1.2E−02	1.1	7.8E−01
Proliferation Index										
High vs. low	0.8	1.5E−01							0.7	8.5E−02
BrMS (up genes)	–	–	2.0	3.2E−05	–	–	–	–	–	–
BrMS (down genes)	–	–			–	–	–	–	–	–
LMS (up genes)	–	–	–	–			–	–	–	–
LMS (down genes)	–	–	–	–	0.8	3.8E−02	–	–	–	–
BoMS (up genes)	–	–	–	–	–	–	1.2	3.7E−02	–	–
BoMS (down genes)	–	–	–	–	–	–			–	–
ER (+ vs. −)					0.6	8.3E−02			0.6	3.5E−02
HER2 (+ vs. −)			1.8	8.9E−02						

Signatures without risk or group categories were evaluated as continuous variables. Metastatic-specific signatures (BrMS, LMS and BoMS) were evaluated in their specific site of relapse context. Since HER2 clinical status was available for a limited number of patients, we used a gene expression surrogate based on the mRNA levels of the HER2 gene, using the top 20% rank order highest expressers as the cutoff value for calling a tumor "HER2-positive" as previously described [12]. For variable selection in the multivariable analyses, the backward stepwise procedure was used

multiple organs as first relapse events, compared to only 21% from luminal A.

After observing the metastasis patterns of the less-differentiated basal-like and claudin-low breast tumors, it was not surprising that the BrMS and LMS signatures associate with subtype and DS. The BoMS was not strongly expressed in any subtype, a finding which may reflect the fact that bone was the most common site of metastasis in our study. These findings complement analyses by Culhane and Quackenbush [32] who found that a different lung metastasis signature [33] was a surrogate for the basal-like subtype. This does not argue, however, that these signatures are not biologically important. In fact, the BrMS identifies some of the least differentiated tumors within the claudin-low and basal-like subtypes and these data support continued investigation of select genes within the BrMS as targeting these genes, along with others that function to increase cellular differentiation, may serve to slow metastatic progression.

To gain a mechanistic understanding for site-specific tumor colonization, we tested a compendium of 298 expression signatures as individual predictors of site of relapse. These analyses showed enrichment for stem cell signatures in brain/lung relapse (Supplemental Table 4). The majority of these signatures provide information that is encoded within DS; however, some of the signatures further divide ER-negative tumors into two distinct groups that are more or less likely to metastasize to the brain/lung. As an example, one such signature is the MM_WapINT3, which is a signature derived from a transgenic mouse mammary tumor model that over-expresses Notch4 and aggressively spreads to the lung [34]. This is a clinically relevant finding in that half of patients with advanced triple negative breast cancer relapse within the brain [35], and survival following CNS relapse is less than 4 months [36], regardless of receipt of systemic therapy.

Overall, the results from Table 2 reveal shared and unique features predicting relapse to distinct sites. For example, intrinsic subtype (as represented by individual subtypes or the ROR-S score) make every final MVA model, but then each site of relapse shows individual characteristics. For brain, the BrMS signature and HER2 status add important information, while for lung the VEGF/hypoxia and LMS signature add information, for bone the DS score was valuable, and for liver, most information was carried by the HER2-enriched subtype; thus for the most accurate site of metastasis predictions, multiple signatures and/or clinical variables are needed. Our ability to predict patients at the highest risk for CNS relapse may impact the manner in which we approach CNS screening and future prevention strategies. The data presented herein provides clinically useful information that could be used to identify patients most likely to experience site-specific breast cancer relapse.

Materials and methods

Human breast tumor microarray datasets

Two distinct microarray data sets were studied here. The first was based upon Agilent Technologies DNA microarrays taken from Prat et al., with 42 new additional metastasis samples profiled here using identical protocols as previously described [6, 19, 37]. All human tumor and normal tissue samples were collected using IRB-approved protocols and all microarray and patient clinical data are available at UNC Microarray Database (https://genome.unc.edu) and have been deposited in the Gene Expression Omnibus (GEO) under the accession number GSE26338. The probes/genes for these analyses were filtered by requiring the Lowess normalized intensity values in both sample and control to be >10. All probes for each gene were averaged. The normalized log2 ratios (Cy5sample/Cy3 control) of probes mapping to the same gene (Entrez ID as defined by the manufacturer) were averaged to generate independent expression estimates.

The second data consisted of a combined microarray data set of four studies taken from the public domain. We utilized the microarray as presented in the following breast cancer datasets: GSE2034, GSE12276, GSE2603, and the NKI295 (*microarray-pubs.stanford.edu/wound_NKI/Clinical_Data_Supplement.xls*). The clinical data from these patients was obtained from previous studies [16, 38]. NCI60 cell line microarray data was obtained from http://genome-www.stanford.edu/nci60/. Additional microarrays from the GEO for the MDA-MB-231 cells were downloaded from GSE12237 and GSE2603. Probes in these external sets were assigned to Entrez Gene identifiers and replicate gene names were collapsed to the median. The data from the four tumor datasets were then combined using Distance Weighted Discrimination [14] to remove the systematic biases present in different microarray datasets. In all datasets, samples were standardized to zero mean and unit variances before other analyses were performed.

Microarray data processing

Samples in the final normalized data were assigned to the five subtypes (luminal A, luminal B, Her2-enriched, basal-like, and normal-like) using the PAM50 classifier [19]. Assignment of claudin-low and DS were performed according to the protocol described in Prat et al. [6]. 298 gene expression modules first characterized in Fan et al. [12] were applied to both data sets and expression estimates obtained for each tumor in each data set; the gene list corresponding to each module was summarized to the mean expression within each sample, or the principal component, or according to a predetermined algorithm. Testing for differential expression of the modules between

primary and metastatic pairs from the same individual was performed with the SAM [39] two class paired test.

Statistics and survival analyses

The intra-class correlation (ICC) [40] was utilized to estimate concordance within specific groups of samples. For groups of paired samples, the ICC was calculated for each pair and then summarized by the mean ICC for each group of interest. ICC values for groups of unpaired samples were estimated from all samples in the group. ICC estimates were performed identically for the set of modules or the set of all genes. All ICC estimates were generated using the R package "irr." Principal components analyses were performed in R. Categorical survival analyses were performed using a log-rank test and visualized with Kaplan–Meier plots. Box-and-whisker plots were used to observe the relationship of the intrinsic subtypes with the organ-specific metastasis signatures and were performed in R. Univariate and multivariable Cox proportional hazard analyses were used to estimate HR and determine the significance of the intrinsic subtypes and gene signatures. Subtypes and DS were compared along with ER status and published signatures using time to first relapse (for each site) as the end point. To visualize the association of DS with Subtype and site of metastasis HR for each site of metastasis were identified by using a sliding window of 50 samples with consecutive DS, and the HR was calculated by contrasting the samples in the window versus those outside the window. HR estimates were smoothed across DS with Lowess.

Functional analysis of gene sets

261 genes that were differently expressed in the three undifferentiated NCI 60 cell lines (as compared to the rest) were uploaded into Ingenuity Systems Pathway Analysis (www.ingenuity.com) based on their Entrez gene identification number.

Boyden chamber migration assays

SUM149PT cells were fluorescence associated cell sorted after immunolabeling with CD49f and Epcam as previously described [6]. $CD49f^{+/high}/Epcam^{+}$ and $CD49f^{+}/Epcam^{-/low}$ cells were plated in 0% FBS Boyden chambers with 8.0μm pores and chemoattracted to 0.5% FBS for 24 h. Migrated cells were stained with crystal violet, and then solubilized and read at 470 nm.

Acknowledgments This study was supported by funds from the NCI Breast SPORE program (P50-CA58223), by RO1-CA138255, by the Breast Cancer Research Foundation, the V Foundation for Cancer Research, and a 2008 Department of Defense Era of Hope Postdoc Award (BC085270) to JCH. We thank Olga Karginova (UNC) for FACS and Xiang Zhang (MSKCC) for the 855 tumor database information. A. Prat is affiliated to the Internal Medicine PhD program of the Autonomous University of Barcelona, Spain.

Conflict of interest Dr. Perou and Parker are inventors on a patent application for intrinsic subtyping, and Dr. Perou has equity interests in University Genomics and Bioclassifier LLC.

References

1. Parkin DM, Pisani P, Ferlay J (1999) Estimates of the worldwide incidence of 25 major cancers in 1990. Int J Cancer 80(6): 827–841
2. Maki DD, Grossman RI (2000) Patterns of disease spread in metastatic breast carcinoma: influence of estrogen and progesterone status. Am J Neuroradiol 21:1064–1066
3. Smid M, Wang Y, Zhang Y et al (2008) Subtypes of breast cancer show preferential site of relapse. Cancer Res 68(9):3108–3114
4. Kennecke H, Yerushalmi R, Woods R et al (2010) Metastatic behavior of breast cancer subtypes. J Clin Oncol 28(20): 3271–3277
5. Herschkowitz JI, Simin K, Weigman VJ et al (2007) Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome Biol 8(5):R76
6. Prat A, Parker JS, Karginova O et al (2010) Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res 12(5):R68
7. Prat A, Perou CM (2011) Deconstructing the molecular portraits of breast cancer. Mol Oncol 5(1):5–23
8. Lim E, Vaillant F, Wu D et al (2009) Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat Med 15(8):907–913
9. Li X, Lewis MT, Huang J et al (2008) Intrinsic resistance of tumorigenic breast cancer cells to chemotherapy. J Natl Cancer Inst 100(9):672–679
10. Phillips TM, McBride WH, Pajonk F (2006) The response of CD24(−/low)/CD44+ breast cancer-initiating cells to radiation. J Natl Cancer Inst 98(24):1777–1785
11. Weigelt B, Hu Z, He X et al (2005) Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. Cancer Res 65(20):9155–9158
12. Fan C, Prat A, Parker JS et al (2011) Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. BMC Med Genomics 4(1):3
13. Ellsworth RE, Seebach J, Field LA et al (2009) A gene expression signature that defines breast cancer metastases. Clin Exp Metastasis 26(3):205–213
14. Benito M, Parker J, Du Q et al (2004) Adjustment of systematic microarray data biases. Bioinformatics 20(1):105–114
15. Wang Y, Klijn JG, Zhang Y et al (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365(9460):671–679
16. Bos PD, Zhang XH, Nadal C et al (2009) Genes that mediate breast cancer metastasis to the brain. Nature 459(7249): 1005–1009
17. Minn AJ, Gupta GP, Siegel PM et al (2005) Genes that mediate breast cancer metastasis to lung. Nature 436(7050):518–524
18. van de Vijver MJ, He YD, van't Veer LJ et al (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347(25):1999–2009
19. Parker JS, Mullins M, Cheang MC et al (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 27(8):1160–1167

20. Minn AJ, Gupta GP, Padua D et al (2007) Lung metastasis genes couple breast tumor size and metastatic spread. Proc Natl Acad Sci USA 104(16):6740–6745
21. Ross DT, Scherf U, Eisen MB et al (2000) Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 24(3):227–235
22. Kang Y, Siegel PM, Shu W et al (2003) A multigenic program mediating breast cancer metastasis to bone. Cancer Cell 3(6):537–549
23. Amer MH, Al-Sarraf M, Baker LH et al (1978) Malignant melanoma and central nervous system metastases: incidence, diagnosis, treatment and survival. Cancer 42(2):660–668
24. Fillmore CM, Kuperwasser C (2008) Human breast cancer cell lines contain stem-like cells that self-renew, give rise to phenotypically diverse progeny and survive chemotherapy. Breast Cancer Res 10(2):R25
25. Creighton CJ, Li X, Landis M et al (2009) Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. Proc Natl Acad Sci USA 106(33): 13820–13825
26. Shipitsin M, Campbell LL, Argani P et al (2007) Molecular definition of breast tumor heterogeneity. Cancer Cell 11(3): 259–273
27. Ben-Porath I, Thomson MW, Carey VJ et al (2008) An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet 40(5):499–507
28. Hu Z, Fan C, Livasy C et al (2009) A compact VEGF signature associated with distant metastases, poor outcomes. BMC Med 7:9
29. Hoefnagel LD, van de Vijver MJ, van Slooten HJ et al (2010) Receptor conversion in distant breast cancer metastases. Breast Cancer Res 12(5):R75
30. Weigelt B, Lo AT, Park CC et al (2010) HER2 signaling pathway activation and response of breast cancer cells to HER2-targeting agents is dependent strongly on the 3D microenvironment. Breast Cancer Res Treat 122(1):35–43
31. Kuwai T, Nakamura T, Sasaki T et al (2008) Targeting the EGFR, VEGFR, and PDGFR on colon cancer cells and stromal cells is required for therapy. Clin Exp Metastasis 25(4):477–489
32. Culhane AC, Quackenbush J (2009) Confounding effects in "a six-gene signature predicting breast cancer lung metastasis". Cancer Res 69(18):7480–7485
33. Landemaine T, Jackson A, Bellahcene A et al (2008) A six-gene signature predicting breast cancer lung metastasis. Cancer Res 68(15):6092–6099
34. Gallahan D, Jhappan C, Robinson G et al (1996) Expression of a truncated Int3 gene in developing secretory mammary epithelium specifically retards lobular differentiation resulting in tumorigenesis. Cancer Res 56(8):1775–1785
35. Lin NU, Claus E, Sohl J et al (2008) Sites of distant recurrence and clinical outcomes in patients with metastatic triple-negative breast cancer: high incidence of central nervous system metastases. Cancer 113(10):2638–2645
36. Niwinska A, Murawska M, Pogoda K (2009) Breast cancer brain metastases: differences in survival depending on biological subtype, RPA RTOG prognostic class and systemic treatment after whole-brain radiotherapy (WBRT). Ann Oncol 21(5):942–948
37. Hu Z, Troester M, Perou CM (2005) High reproducibility using sodium hydroxide-stripped long oligonucleotide DNA microarrays. Biotechniques 38(1):121–124
38. Zhang XH, Wang Q, Gerald W et al (2009) Latent bone metastasis in breast cancer tied to Src-dependent survival signals. Cancer Cell 16(1):67–78
39. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98(9):5116–5121
40. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86(2):420–428