SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
BIOINFORMATICS
COMPARATIVE GENOMICS
CANCER MODELS
CANCER GENOMICS

Received
14 September 2011

Accepted
2 January 2012

Published
17 January 2012

Correspondence and requests for materials should be addressed to J.A.H. (hassell@mcmaster.ca)

A Gene Signature for Predicting Outcome in Patients with Basal-like Breast Cancer

Robin M. Hallett[1], Anna Dvorkin-Gheva[1], Anita Bane[2] & John A. Hassell[1]

[1]Department of Biochemistry and Biomedical Sciences, Centre for Functional Genomics, McMaster University, 1200 Main Street West, Hamilton, Ontario, Canada L8N 3Z5, [2]Department of Oncology, McMaster University, Juravinski Cancer Centre, 699 Concession St East, Hamilton, Ontario, Canada, L8V 5C2.

Basal-like breast cancer is a molecular subtype of breast cancer with a poor prognosis. Follow-up studies of long-term outcome in these patients, demonstrates they can be separated into two clinical groups: those who succumb to their disease within the first 5 years and those expected to show excellent long term survival. Currently available clinical/histopathological variables as well as molecular signatures show little capacity to identify basal breast cancer patients with either a high or low risk of disease relapse. Using data derived from 85 basal-like breast cancer patients, we identified a 14-gene signature, which we subsequently validated on an additional 49 basal breast cancer patient set. The ability to distinguish between these two sub-groups of basal breast cancer patients at the time of initial diagnosis would permit tailoring aggressive therapeutic regimens to those patients with a poor prognosis and conversely avoid such therapy in low risk patients.

Traditionally a number of tumor characteristics have been used to determine the prognosis of breast cancer patients. Such factors include tumor size, grade, hormone receptor status, HER2 status, lympho-vascular space invasion and lymph node involvement[1,2]. More recently whole genome analysis technology (gene expression profiling) has been added to the armamentarium of experimental techniques, thus providing a new molecular classification for breast cancer and contributing to the development of a number of prognostic multi-gene assays including a 21-gene, 70-gene, 76-gene, 77-gene genomic grade profile, wound response signature and others[3–9]. One of these assays that is commercially available is Oncotype DX®, a 21-gene quantitative (q)RT-PCR assay, which evaluates expression of 16 genes identified to be of prognostic importance as well as 5 house-keeping genes[3]. Oncotype DX® predicts the risk of distant recurrence in Estrogen Receptor (ER) positive breast cancers and their responsiveness to CMF (Cyclophosphamide, Methotrexate and 5-Fluorouracil) chemotherapy[10]. MammaPrint®, a commercially available microarray evaluates the expression of 70 genes using RNA extracted from fresh frozen tumor samples. This assay distinguishes patients that have a good prognosis (no relapse within 5 years) from those that have a poor prognosis (relapse within 5 years)[11]. Indeed, large clinical trials, such as TAILORx [Trial Assigning Individualized Options for Treatment] and MINDACT [Microarray In Node Negative and 1-3 positive lymph node Disease may Avoid Chemotherapy] are ongoing to evaluate the use of both Oncotype DX® and MammaPrint® in clinical practice.

The term basal-like breast cancer (BLBC) originated in 2000 from gene expression profiling experiments conducted on invasive breast cancers by Perou and colleagues at Stanford University[12–14]. Using hierarchical clustering these investigators identified a new molecular taxonomy for breast cancer based on the relative expression of the ~500 genes, known as the 'intrinsic' gene set. These investigators discovered that breast cancers could be classified into five molecular subgroups. Two of these are ER positive whereas three are ER negative. The ER positive subgroups, termed Luminal A and Luminal B, were identified based on their relative expression of the ER gene, ER regulated genes and other genes expressed by normal breast 'luminal' cells. The ER negative subgroups were termed HER2 overexpressing (ERBB2+), normal breast-like and BLBC. The HER2 overexpressing subgroup was characterized by the overexpression of the *HER-2* and other genes on the 17q amplicon, such as *GRB7*. The normal breast-like subgroup expresses genes characteristic of adipose tissue suggesting that this subgroup may be a technical artifact resulting from low tumor cellularity. Lastly, the basal-like subgroup represents a distinct class of tumors characterized by the lack of expression of ER, PR and HER2 and the high expression of cytokeratins (CK) 5, and/or CK 17 (amongst other genes), characteristic of the basal/myoepithelial

cell layer of the normal breast epithelium. As gene expression studies continued to evolve, new molecular subtypes of breast cancer continued to be discovered; for example in 2007 the claudin low subtype was identified[15].

Most importantly the initial gene expression profiling experiments demonstrated that BLBCs together with the HER2 overexpressing subtype were associated with a particularly poor prognosis. By comparison, patients with Luminal A type tumors displayed an excellent prognosis[13,14]. However, on closer examination these studies additionally demonstrated that the prognosis of patients with BLBCs is highly time dependent. Some patients with BLBCs experience particularly poor survival in the first 3–5 years following diagnosis, but for others their mortality wanes such that at 10 years post diagnosis these patients have a better survival than those with luminal-type (ER+) tumors[16–19]. This suggests that patients with BLBCs can be separated into two clinically distinct groups: those likely to experience a recurrence and to succumb to their disease in the first 3–5 years after diagnosis, and those expected to show excellent long term survival.

Whereas several multi-gene signatures exist to predict breast cancer patient prognosis, their prognostic values appears to be mostly derived from their capacity to measure expression of genes associated with proliferation[20,21]. Because BLBCs are generally highly proliferative, the existing prognostic signatures fail to identify a subset of BLBC with good prognosis[22]. Some recent work has focused on identifying multi-gene predictors of outcome in triple negative (ER-, PR-, HER2-) and hormone receptor negative breast cancer[21–26]. However, a robust method of distinguishing between BLBCs with good and poor outcome has yet to be developed. To the latter end, we have begun optimizing such a method and report here the identification of a 14-gene signature that is associated with patient outcome in BLBCs.

Results

Compiling multiple gene expression profiles of basal breast tumors. To identify genes whose expression might be associated with the clinical outcome of BLBC patients, we compiled a large collection of human breast tumor gene expression data for which clinical data was also available (n=995). Hierarchical clustering using the 'intrinsic' gene set revealed that many of these tumors (n=547) clustered into the previously described molecular subtypes[12–14] (Fig. 1a). Importantly, survival analysis using Kaplan-Meier survival curves revealed distinct differences in clinical outcome among the patients with tumors of different molecular subtypes. As observed previously patients with tumors of the basal-like, ERBB2, claudin-low and luminal B subtypes experienced the poorest 10-year survival, whereas patients with luminal A or normal-like tumors experienced the best 10-year survival[13] (Fig. 1b). Interestingly, the 10-year survival rate of patients with basal-like tumors was approximately 60% and very few BLBC patient mortalities occurred after this time (Fig. 1b). The latter findings are consistent with previous observations that the prognosis of BLBC patients is time dependent, where these patients are at highest risk for relapse during the first 5 years post diagnosis and experience a very low risk for relapse 10-years post diagnosis[16–18].

Importantly, the BLBC tumor cohort comprised 134 patients with clinical follow-up data, thus providing a fairly large number of basal tumors to identify a genomic predictor that could be used to guide prognosis for patients with basal-like breast tumors.

Training signature. To develop a genomic predictor that could be used to identify BLBC patients who were likely to have either good or poor survival outcomes, we first divided the 134 patient BLBC cohort into a 85 patient training set and a 49 patient validation set. We used binary regression probabilistic models for feature selection to identify genes that had the best prognostic performance among the gene expression profiles derived from the 85 BLBCs of the training set[27]. For these analysis, <5 year DFS was taken to indicate poor outcome, whereas >5 year DFS was taken to indicate good outcome. Previous studies have shown that the vast majority of disease recurrence among BLBC patients occurs within the first 5 years[16–18]. Starting with a single probe set signature, we iteratively generated signatures by gradually adding probe sets and tested the resulting signature using leave-one-out cross-validation. In this fashion we generated multiple signatures comprising n probe sets, where n = 1,2,3…,50 (Fig. 2A). For each discrete value of n, this technique assigned a probability to every patient within the training set that indicated the likelihood of a patient experiencing disease relapse. To establish a probability cut-point, where patients with higher probability are assigned into the poor prognosis category and patients with lower probability are assigned into the good prognosis category, we used a previously described tertile method[28]. In this fashion, good prognosis was assigned to patients whose probability score fell in the lowest 1/3 of all probability scores, whereas poor prognosis was assigned to patients whose score fell into the higher 2/3 of probability scores. Indeed, these approximate proportions have been observed in several gene expression based breast cancer prognostication studies[4,7,29,30]. We therefore took this approach as a relatively non-biased and simple means to divide patients into predicted good and poor outcome groups. To determine which n-element signature had optimal performance we compared the relative risk of relapse for each signature (Fig. 2B, red line: relative risk, black line: LOWESS (LOcally WEighted Scatterplot Smoothing) curve fitted to relative risk data, n=14 identifies optimal signature length). In this fashion we identified a 14-probe-set (each gene represented by 1 probe set,

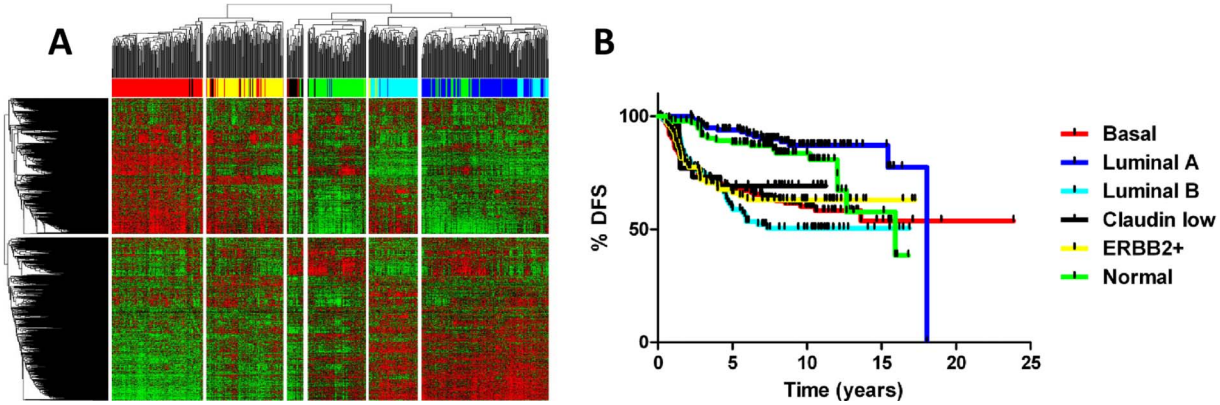

Figure 1 | Human breast tumors cluster into 6 distinct molecular subtypes of breast cancer with differences in patient survival. (A) Hierarchical clustering of 547 breast tumors using the 'intrinsic' gene set separates tumors into the 6 molecular subtypes of human breast tumors. (B) Kaplan-Meier survival analysis of patients comprising each of the molecular subtypes.

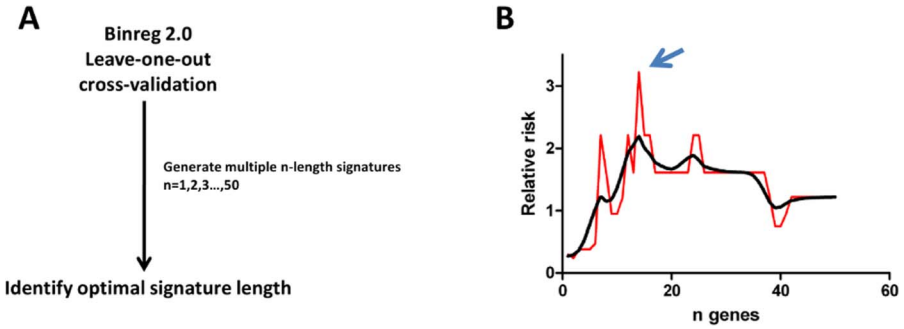

Figure 2 | 14 probe sets optimally separate patients into good and poor survival groups. (A) Experimental strategy to identify an optimal signature to separate patients with BLBC into high and low risk groups. (B) Comparison of relative risk between leave-one-out cross-validation predicted high and low risk groups for n length signatures (n=1,2,3…,50). 14 probe sets produces maximal risk separation between high and low risk groups (blue arrow).

henceforth called Basal 14 signature) signature, which optimally separated patients into good and poor outcome groups (Table 1).

Assessment of Signature Performance. Validation of a gene signature using an independent data set is a more accurate measurement of its prognostic value than using cross-validation on a training data set. Therefore, we tested our Basal-14 signature on an independent cohort of patients with BLBC (n=49). To learn whether the probability of disease relapse predicted by the Basal-14 signature could be used as a continuous predictor of disease relapse, we calculated the proportion of patients who had experienced disease relapse while increasing the cut-off (decreasing stringency) for assigning a patient into the good outcome group. Indeed, the proportion of patients experiencing disease relapse increased in an approximate linear fashion as the probability assigned for disease relapse by the Basal-14 signature increased (Fig. 3A). To assess the predictive accuracy of the Basal-14 signature, we completed receiver-operator characteristic (ROC) curve analysis. In this fashion, an AUC (Area Under Curve) value of 0.5 indicates predictive performance which is no better than chance, whereas values greater than 0.5 indicate true predictive capacity. The Basal-14 signature produced an AUC that was statistically significantly higher than 0.5 (AUC: 0.76, p=0.003, Fig. 3B). Taken together, these data demonstrate the capacity for the Basal-14 signature to identify BLBC patients at high risk for disease relapse. To visualize survival differences between groups of patients that were predicted to have either high or low risk for disease relapse, we stratified patients from the validation cohort into good and poor outcome groups using tertiles, and completed Kaplan-Meier survival analysis. Patients whose predicted probability for disease relapse fell within the lowest tertile of predicted probabilities were stratified into the good outcome group, whereas those whose predicted probabilities fell within the upper two tertiles were stratified into the poor outcome group. The Kaplan-Meier estimate for the proportion of patients in the low-risk group who did not experience a disease relapse at 5 years (94%) was significantly greater than the proportion in the poor outcome category (48%) (Table 2, Fig. 3C, HR: 4.7 [CI95: 1.8–12.3], p = 0.0017). Because our overarching objective was to identify patients who could be spared aggressive chemotherapy, we also tested the capacity of our signature to predict the outcome of patients who had not received adjuvant chemotherapy. In this fashion, we were able to test the relationship between the Basal-14 signature and the natural progression of BLBCs without having adjuvant chemotherapy as a potentially confounding variable. 26 patients within the 49 patient validation cohort met this criterion (patients from GSE7390 & GSE2034). We re-tested the predictive capacity of the Basal-14 signature on these 26 chemotherapy naïve patients and observed a statistically significant difference in the survival of patients who were predicted to have either good or poor outcome (Fig. 3D, HR: 4.4 [CI95: 1.1–16.7], p = 0.03, Table 3). The proportion of patients in the chemotherapy naïve validation cohort who were predicted to have good survival and were free of disease at 5 years was 100%, whereas among those patients who were predicted to have poor survival, only 50% were disease free after 5 years. Taken together, these findings demonstrate the capacity of our gene signature to identify patients who have excellent long-term survival even when patients did not receive aggressive adjuvant chemotherapy.

Comparison of the Basal-14 signature with other multigene predictors. Previous studies have reported that many published multigene predictors fail to accurately identify high and low risk patients among patients with ER-negative breast cancer[22,24]. As the majority of BLBCs are ER-negative, we sought to test whether

Table 1 | Features comprising the optimal 14-gene signature

Correlation	Affymtetrix Probe	Description
+	201022_s_at	destrin (actin depolymerizing factor), **DSTN**
+	203072_at	myosin IE, **MYO1E**
+	208089_s_at	tudor domain containing 3, **TDRD3**
+	204338_s_at	regulator of G-protein signaling 4, **RGS4**
+	220719_at	hypothetical protein FLJ13769, **FLJ13769**
+	212039_x_at	ribosomal protein L3, **RPL3**
+	211073_x_at	ribosomal protein L3, **RPL3**
+	201217_x_at	ribosomal protein L3, **RPL3**
+	208538_at	acidic (leucine-rich) nuclear phosphoprotein 32 family, member C, **ANP32C**
–	217434_at	melanocortin 2 receptor (adrenocorticotropic hormone), **MC2R**
–	216143_at	MRNA; cDNA DKFZp434L092 (from clone DKFZp434L092), –
–	221306_at	G protein-coupled receptor 27, **GPR27**
–	204544_at	Hermansky-Pudlak syndrome 5, **HPS5**
–	208885_at	lymphocyte cytosolic protein 1 (L-plastin), **LCP1**

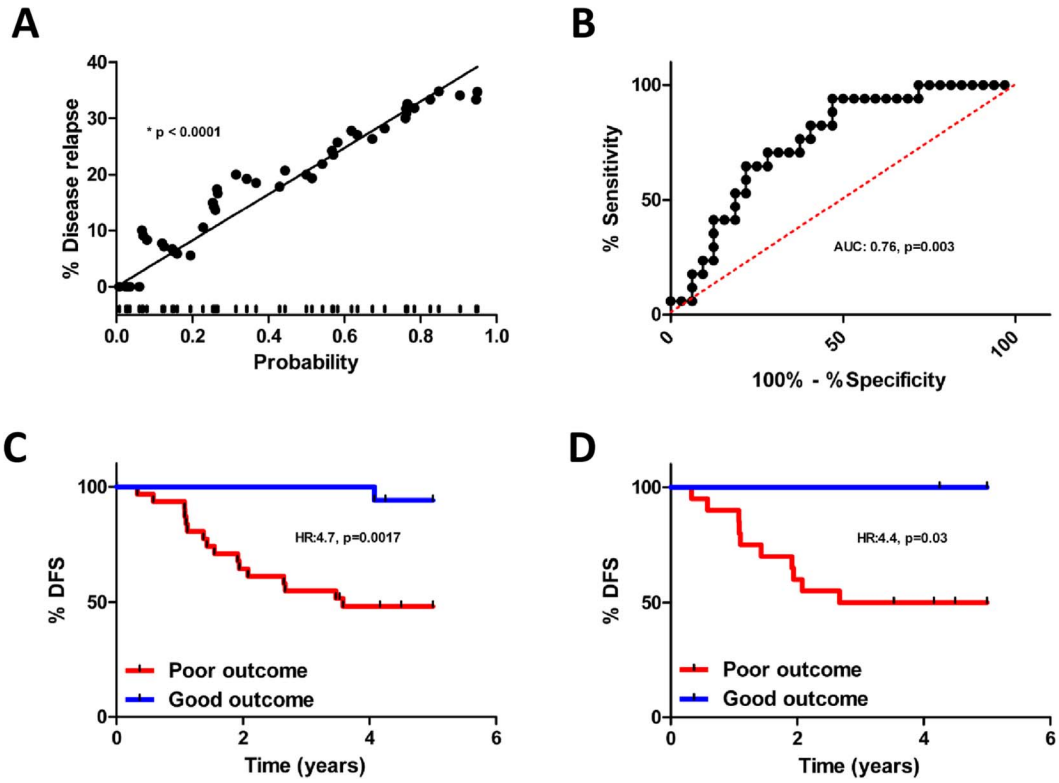

Figure 3 | The Basal 14 signature accurately predicts outcome in independent patients with BLBC. (A) Rug plot (distribution of predicted probabilities) of proportion of patients experiencing disease relapse increases linearly with probability of relapse predicted by Basal 14 signature. (B) ROC curve to assess the accuracy of the Basal 14 signature in the validation cohort (AUC: 0.76, p = 0.003). (C) Kaplan-Meier survival analysis with the validation (HR: 4.7, [CI95: 1.8–12.3], p = 0.0017, Log-rank test). (D) Kaplan-Meier survival analysis with chemotherapy naïve patients (HR: 4.4, [CI95: 1.1–16.7], p = 0.03, Log-rank test).

multiple previously described multigene predictors were prognostic in the context of BLBC. To this end, we measured the association of the Genomic Grade Index[5], NKI-70 signature[31], Recurrence score[3], CSR/Wound response signature[6], Triple-negative signature[22], MS-14 signature[32], as well as the Basal-14 signature in the 49 patient validation cohort by calculating a signature index and completing either Kaplan-Meier survival analysis using tertiles to dichotomize the validation cohort into good and poor outcome groups, or generating ROC curves. Interestingly, other than the Basal-14 signature (Fig. 4A, HR: 4.3 [CI95: 1.6–11.4], p = 0.0032) none of the other signatures identified patient groups with statistically significant differences in survival (Kaplan-Meier: Fig. 4A-F. ROC: supplementary figure. 1A–F). These data suggest that the prognostic capacity of previously reported multigene outcome predictors may be diminished in patients with BLBC. However, it should be noted that the tertile method used to separate patients into good and poor outcome groups may be non-optimal for these signatures. Interestingly, the triple negative signature trended towards significance in the Kaplan-Meier analysis (Fig. 4F, HR: 2.0 [CI95: 0.8–5.4], p = 0.15) and was statistically significant in the ROC curve analysis (Supplementary fig. 1G, AUC: 0.7, p = 0.02). This is likely because the triple negative signature was developed with triple negative breast tumors, which comprises a sub-group that overlaps with the basal-like molecular subtype. Together, these findings underscore the need for prognostic multigene signatures, such as the Basal 14 signature, for guiding therapy choice for breast cancer patients.

Performance of Basal-14 signature in other molecular subtypes of breast cancer. Previous studies have demonstrated that biological processes that can be linked to breast cancer patient outcome vary among the different molecular subtypes of breast cancer[21]. In this regard, we sought to test whether the Basal-14 signature could be used to identify high and low risk patients among the other molecular subtypes of breast cancer, or whether its capacity to stratify patients into high and low risks groups was limited to patients with BLBCs. The Basal-14 signature showed no capacity to identify patients at high and low risk for disease relapse among the luminal A (HR: 1.3, p=n.s.), luminal B (HR: 1.2, p=n.s.), claudin low (HR: 1.0, p=n.s.) and normal (HR: 0.4, p=n.s.) molecular subtypes of breast cancer (Fig. 5A–D). Unexpectedly, the Basal-14 signature was also prognostic in the ERBB2 molecular subtype (HR: 2.8 [CI95: 1.3–6.5], p = 0.01). Interestingly, a previously reported prognostic gene signature developed using Her2-positive tumors was also found to be prognostic in BLBCs[33]. These data suggest that similar biological

Table 2 | Survival characteristics of the 49 patient validation cohort

Validation cohort (n=49)			
Risk Category	# Patients	% Patients	% Disease free survival (5 yr)
Low	16	33	94
High	33	67	48

Table 3 | Survival characteristics of the 26 patient chemo-naïve validation cohort

Chemo-naïve validation cohort (n=26)			
Risk Category	# Patients	% Patients	% Disease free survival (5 yr)
Low	6	23	100
High	20	77	50

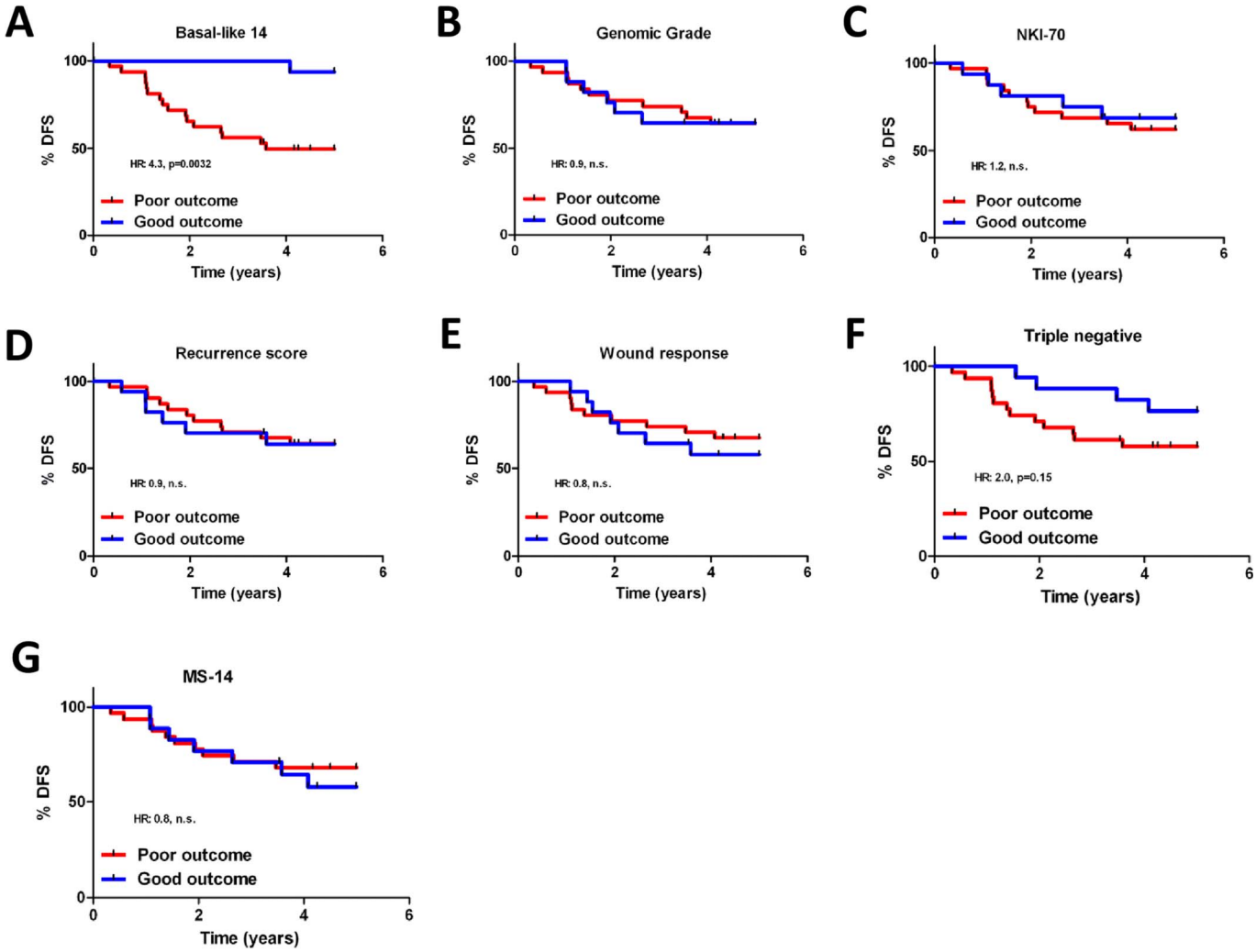

Figure 4 | Other reported prognostic signatures fail to predict patient outcome in the context of BLBC. We calculated a signature index for the (A) Basal 14, (B) Genomic Grade Index, (C) NKI-70, (D) Recurrence Score, (E) CSR/Wound response, (F) Triple Negative and (G) MS-14 signatures. Only the Basal 14 signature was prognostic in the validation cohort of BLBC patients HR: 4.7 [CI95: 1.8–12.3], p = 0.0017, Log-rank test). Although, the Triple negative signature did trend to significance (HR: 2.0 [CI95: 0.8–5.4], p = 0.15, Log-rank test).

processes may govern patient outcome in both the basal-like and ERBB2 molecular subtypes of breast cancer. Taken with our previous findings, it appears that transcripts whose expression may be informative for patient prognosis vary between the different molecular subtypes of breast cancer. For example, it appears that signatures that are prognostic in ER-positive breast tumors, such as the Reccurrence score (OncotypeDX®) and the Genomic Grade Index, fail to stratify BLBCs into good and poor outcome groups, whereas the Basal-14 signature is prognostic in basal-like and ERBB2-overexpressing breast cancer, but fails to identify patients in the ER-positive luminal subtypes of breast cancer.

Discussion

Few, if any, clinical variables show prognostic capacity in the context of BLBC. Therefore, we sought to identify a genomic predictor of patient outcome for patients with BLBC. In the present study, we identified a 14 probe set signature, which we named the Basal 14 signature. We tested the Basal 14 signature on an independent validation cohort of BLBC patients and were able to accurately stratify patients into good and poor outcome groups. Importantly, the difference in risk for disease relapse for patients who were predicted to have either good or poor outcome was both relatively large and statistically significant. Because it was unclear whether the Basal 14 signature was related to the natural progression of BLBCs, tumor response to therapy, or both, we also tested the Basal 14 signature on a smaller group of patients who did not received treatment with adjuvant chemotherapy. In this fashion, we were able to confirm a relationship, albeit in a small number of patients, between the Basal 14 signature and patient survival in chemotherapy naïve patients. Notably, previous reports suggest that immune-based signatures predict response to chemotherapy in triple negative breast cancer patients, suggesting that the Basal 14 signature might also measure treatment response[21,34]. The relationship between the Basal 14 signature and response to chemotherapy was not examined in this study. Another possibility is that the Basal 14 signature is associated with histological subtypes of BLBC with known good prognosis, such as the medullary subtype[35,36]. However, the frequency of medullary breast tumors is exceptionally low (2%), suggesting that the Basal 14 signature would also need to identify good prognosis non-medullary BLBCs the achieve the level of accuracy described here. In total, the capacity of the Basal 14 signature to identify BLBC patients with good prognosis is likely multi-factorial, and many additional possibilities remain unexplored.

Interestingly, the Basal 14 signature comprised multiple genes with known roles in cancer. For example, destrin (DSTN) is one of three mammalian actin depolymerisation factors (ADFs). These

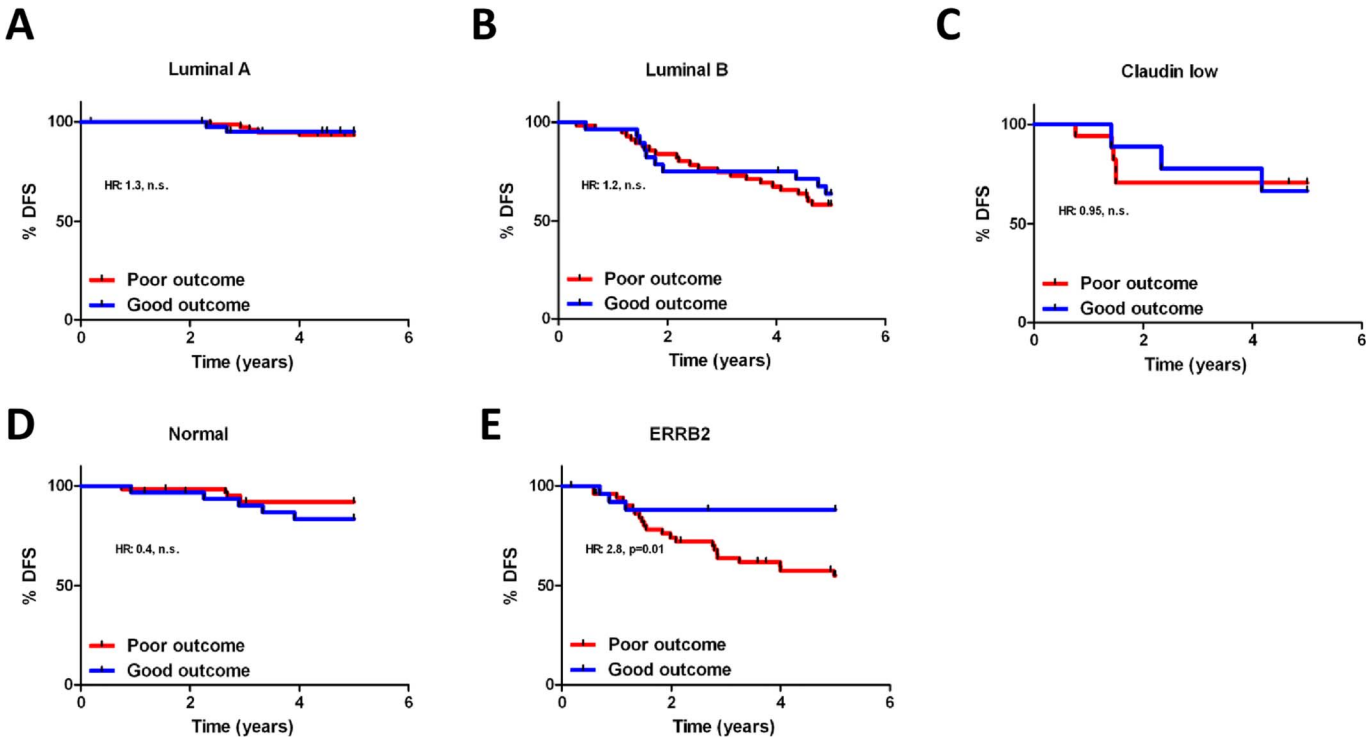

Figure 5 | Basal 14 signature is prognostic in the basal and ERBB2 molecular subtypes of breast cancer. Prognostic capacity of the Basal 14 signature was evaluated in the (A) luminal A, (B) luminal B, (C) claudin low, (D) Normal, and (F) ERBB2 molecular subtypes of breast cancer. Notably, the Basal 14 signature was prognostic in patients with the ERBB2 molecular subtype of breast cancer (HR: 2.8 [CI95: 1.3–6.5], p = 0.01, Log-rank test).

proteins are fundamental for multiple cellular processes such as cell survival, cytokinesis, as well as cell migration and chemotaxis[37], and have been linked as a major determinant of metastasis in cancer patients[38,39]. Tudor domain containing protein 3 (TDRD3) has previously been linked to outcome in patients with ER-negative breast tumors[40], and while being relatively poorly characterized, is thought to play a role in the regulation of cytoplasmic stress granules[41]. Regulator of G-protein signaling (RGS4) has also been linked to patient outcome in patients with triple negative tumors[22]. Notably, RGS4 appears to be a key negative regulator of breast cancer cell migration and invasion[42]. It is therefore somewhat surprising that high levels of RGS4 transcripts are associated with poor outcome. However, it appears that RGS4 function is heavily regulated post-translationally by proteosomal degradation, suggesting that a negative feedback loop occurs where high levels of RGS4 transcripts indicate low levels of RGS4 protein[42]. Interestingly, proteasome inhibitors are being explored as possible means for cancer therapy[43,44]. In this regard, BLBC patients may represent a cancer sub-group that might benefit from such a therapeutic approach. Three of the probe sets comprising the Basal 14 signature bind to transcripts that encode ribosomal protein L3 (RPL3). While it seems likely that this gene is involved in mRNA translation, implying that BLBCs with high levels of protein synthesis are associated with poor patient outcome, the role of RPL3 in cancer is uncharacterized. The genes representing transcripts whose expression was related to good survival are largely uncharacterized in regards to roles in tumor cell biology. Lymphocyte cytosolic protein 1 (LCP1), which is likely expressed by tumor infiltrating lymphocytes, might represent a readout of the extent of tumor lymphocyte infiltrate. This suggests that patient outcome may be influenced by host immune response, where infiltrating immune cells, such as lymphocytes, within a tumor indicate a good prognosis. Indeed, similar observations have been made by multiple other groups in the context of ER negative breast tumors[22,24]. Taken together, these data highlight the diverse biology of the genes comprising the Basal 14 signature and provide a scientific rationale for new lines of research aimed at developing BLBC specific therapies.

Several issues remain to be addressed for the Basal 14 signature to be a useful clinical tool. Our conclusions are based on the analysis of retrospective data, which limits its clinical value. Moreover, the validation cohort we used to test the predictive accuracy of the Basal 14 signature was relatively small. Finally, many of the patients in our data-set had incomplete clinical data, making it impossible to learn whether the Basal 14 signature was independently prognostic in the context of other additional factors such as patient age, tumor size, tumor grade, etc. However, it is important to note that previous reports suggest that factors such as tumor size, tumor grade, extent of vascular invasion, and patient age show little relationship to patient outcome in the context of BLBC especially in lymph-node negative patients[45,46]. Indeed, the only standard clinical variable that is consistently prognostic in BLBC appears to be nodal status[45,47]. Interestingly, we found that the Genomic Grade Index, a genomic based measurement of tumor grade showed no capacity to stratify BLBC patients into good and poor outcome groups. Subsequent validation of the Basal 14 signature will need to be completed in larger cohorts of patients that include such multivariate analyses. In this regard, a major focus of our research is the optimization of the Basal 14 signature for use on breast tumor tissue that is routinely available after surgery, such as formalin fixed paraffin embedded tumor blocks.

No rigorously validated assay exists to guide prognosis of patients specifically with BLBC. Indeed, the data we present here suggests that the possibility of developing such a test exists. Future experiments will aim to extend these findings in additional retrospective cohorts of patients with BLBCs and ultimately in a prospective based clinical trial aimed at sparing low risk BLBCs patients from detrimental and unnecessary adjuvant chemotherapy.

Methods

We used a four-step approach to complete proof-of-principle experiments to show that gene-expression signatures can be identified and used to classify patients with BLBCs into good and poor outcome groups.

1. We assembled a large cohort of 995 breast tumor gene expression profiles for which clinical follow-up data was available.
2. We classified each tumor on the basis of its 'intrinsic' molecular subtype from which we generated a new dataset consisting of only BLBCs.
3. We used a subset of BLBCs to iteratively identify several prognostic gene signatures, and used cross-validation to identify the optimal signature for patient outcome classification.
4. We validated our optimized signature prospectively on an independent subset of basal breast tumors with accompanying gene expression profiles and clinical follow-up data.

Collecting Microarray Data. We analyzed the gene expression profiles of 5 independent external datasets, obtained using Affymetrix HG-U133A GeneChips arrays, which have been deposited in the Gene Expression Omnibus (GEO); accession numbers GSE1456, GSE2034, GSE3494, GSE6532, and GSE7390. Together these datasets provided expression profiles of 1,077 human breast tumor samples. All gene expression profiles were normalized with frozen Robust Multi-Array Analysis (fRMA), a procedure that allows one to pre-process microarrays individually or in small batches and to then combine the data into a single comparable dataset for further analyses[48]. To remove batch effect from the combined dataset, we used the ComBat method, which uses an Empirical Bayes method to adjust for potential batch effects in the dataset[49] (http://genepattern.broadinstitute.org), and computed Pearson correlation coefficients for pair-wise comparisons of samples using 68 house-keeping probe sets; only samples exhibiting correlations higher than 0.95 with at least half of the dataset were selected for further classification. The latter filtering method yielded a dataset comprising 995 human breast tumor samples.

Tumor Classification. Each of the selected 995 samples described above, were classified as basal-like, HER2+, Luminal A, Luminal B, claudin-low or normal-like by assigning it to a cluster representing the subtype to which it had the highest Pearson correlation[12,13,15]. The correlation was computed using the subset of 1,500 averaged and median-centered 'intrinsic' genes[50] common to both our dataset (Affymetrix Human Genome U133A Array) and the dataset used by Parker *et al.* (Stanford Microarray). For robustness, only tumors exhibiting a correlation higher than 0.3 with any of the molecular subtypes were used for further analysis. This led to the classification of 137 breast tumors into the basal-like molecular subtype yielding a group of 134 tumors with useable clinical follow-up data. We randomly separated the 134 patients with basal breast tumors; approximately 2/3 (n=85) were taken for signature training purposes (training set), whereas and the remaining 1/3 (n=49) was used as an independent validation set.

Binary regression. Identification of the prognostic signature was completed using the Bayesian binary regression algorithm BinReg ver2.0. The binary regression software (BinReg2.0) was downloaded from http://web.duke.edu/~dinbarry/BINREG/ and was used as a MATLAB plug-in[27]. In most cases, we used disease free survival (DFS) as the relevant clinical variable, however, in some cases only distant metastasis free survival (DMFS) was available within a patient's clinical annotation. In these cases we counted DMFS as DFS. We used 5 years DFS as the clinical endpoint for these studies.

Assessing signature performance. Survival differences between predicted good and poor outcome groups were evaluated with Kaplan-Meier survival curves and a log-rank test for significance. Many standard prognostic clinical variables (node, grad, size, age…, etc) were unavailable in the GEO files associated with the patients used in this study, thus a limitation of this study is that we were not able to test the capacity of the Basal 14 signature to remain prognostic in the context of standard prognostic factors.

Comparison of the Basal 14 signature with other genomic based predictors. We tested multiple additional prognostic signatures on the 49 patient validation cohort: Genomic Grade[5], NKI-70[31], Recurrence score[3], Wound response[6], Triple negative[22] and MS-14[32]. For cross platform comparisons with other gene signatures, signature elements were mapped by Unigene IDs to Affymetrix HG-U133A GeneChip arrays for testing in the 49 patient validation set. The expression values for each gene were transformed such that the mean was 0 and the standard deviation was 1. A signature index was calculated for each patient as follows:

$$\frac{\sum_{i\in P} x_i}{n_P} - \frac{\sum_{i\in N} x_j}{n_N}$$

Where x is the transformed expression, n is the number of genes that could be mapped to the Affymetrix HG-U133 arrays, P is the set of probes with reported positive correlation to poor outcome, and N is the set of probes with reported positive correlation to good outcome. For each signature, patients were divided into high and low signature index groups using tertiles[28].

1. Hayes, D. F., Trock, B. & Harris, A. L. Assessing the clinical impact of prognostic factors: when is "statistically significant" clinically useful? *Breast Cancer Res Treat* **52**, 305–319 (1998).
2. 1997 update of recommendations for the use of tumor markers in breast and colorectal cancer. Adopted on November 7, 1997 by the American Society of Clinical Oncology. *J Clin Oncol* **16**, 793–795 (1998).
3. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* **351**, 2817–2826 (2004).
4. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679 (2005).
5. Sotiriou, C. *et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* **98**, 262–272 (2006).
6. Chang, H. Y. *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci U S A* **102**, 3738–3743 (2005).
7. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
8. Hallett, R. M., Dvorkin, A., Gabardo, C. M. & Hassell, J. A. An algorithm to discover gene signatures with predictive potential. *J Exp Clin Cancer Res* **29**, 120 (2010).
9. Hallett, R. M. & Hassell, J. A. E2F1 and KIAA0191 expression predicts breast cancer patient survival. *BMC Res Notes* **4**, 95 (2011).
10. Paik, S. *et al.* Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol* **24**, 3726–3734 (2006).
11. Bogaerts, J. *et al.* Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. *Nat Clin Pract Oncol* **3**, 540–551 (2006).
12. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
13. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869–10874 (2001).
14. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100**, 8418–8423 (2003).
15. Herschkowitz, J. I. *et al.* Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol* **8**, R76 (2007).
16. Cheang, M. C. *et al.* Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clin Cancer Res* **14**, 1368–1376 (2008).
17. Mulligan, A. M., Pinnaduwage, D., Bull, S. B., O'Malley, F. P. & Andrulis, I. L. Prognostic effect of basal-like breast cancers is time dependent: evidence from tissue microarray studies on a lymph node-negative cohort. *Clin Cancer Res* **14**, 4168–4174 (2008).
18. Dent, R. *et al.* Triple-negative breast cancer: clinical features and patterns of recurrence. *Clin Cancer Res* **13**, 4429–4434 (2007).
19. Blows, F. M. *et al.* Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS Med* **7**, e1000279.
20. Wirapati, P. *et al.* Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res* **10**, R65 (2008).
21. Desmedt, C. *et al.* Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin Cancer Res* **14**, 5158–5165 (2008).
22. Yau, C. *et al.* A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer. *Breast Cancer Res* **12**, R85.
23. Sabatier, R. *et al.* Kinome expression profiling and prognosis of basal breast cancers. *Mol Cancer* **10**, 86.
24. Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O. & Caldas, C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biol* **8**, R157 (2007).
25. Kreike, B. *et al.* Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas. *Breast Cancer Res* **9**, R65 (2007).
26. Rody, A. *et al.* T-cell metagene predicts a favorable prognosis in estrogen receptor-negative and HER2-positive breast cancers. *Breast Cancer Res* **11**, R15 (2009).
27. West, M. *et al.* Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A* **98**, 11462–11467 (2001).
28. Haibe-Kains, B., Desmedt, C., Sotiriou, C. & Bontempi, G. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics* **24**, 2200–2208 (2008).
29. Buyse, M. *et al.* Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* **98**, 1183–1192 (2006).
30. 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).
31. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999–2009 (2002).
32. Tutt, A. *et al.* Risk estimation of distant metastasis in node-negative, estrogen receptor-positive breast cancer patients using an RT-PCR based prognostic expression signature. *BMC Cancer* **8**, 339 (2008).

33. Staaf, J. *et al*. Identification of subtypes in human epidermal growth factor receptor 2--positive breast cancer reveals a gene signature prognostic of outcome. *J Clin Oncol* **28**, 1813–1820.
34. Hess, K. R. *et al*. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol* **24**, 4236–4244 (2006).
35. Ridolfi, R. L., Rosen, P. P., Port, A., Kinne, D. & Mike, V. Medullary carcinoma of the breast: a clinicopathologic study with 10 year follow-up. *Cancer* **40**, 1365–1385 (1977).
36. Vincent-Salomon, A. *et al*. Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity. *Breast Cancer Res* **9**, R24 (2007).
37. Van Troys, M. *et al*. Ins and outs of ADF/cofilin activity and regulation. *Eur J Cell Biol* **87**, 649–667 (2008).
38. Wang, W., Eddy, R. & Condeelis, J. The cofilin pathway in breast cancer invasion and metastasis. *Nat Rev Cancer* **7**, 429–440 (2007).
39. Wang, W. *et al*. Identification and testing of a gene expression signature of invasive carcinoma cells within primary mammary tumors. *Cancer Res* **64**, 8585–8594 (2004).
40. Nagahata, T. *et al*. Expression profiling to predict postoperative prognosis for estrogen receptor-negative breast cancers by analysis of 25,344 genes on a cDNA microarray. *Cancer Sci* **95**, 218–225 (2004).
41. Goulet, I., Boisvenue, S., Mokas, S., Mazroui, R. & Cote, J. TDRD3, a novel Tudor domain-containing protein, localizes to cytoplasmic stress granules. *Hum Mol Genet* **17**, 3055–3074 (2008).
42. Xie, Y. *et al*. Breast cancer migration and invasion depend on proteasome degradation of regulator of G-protein signaling 4. *Cancer Res* **69**, 5743–5751 (2009).
43. Orlowski, R. Z. & Kuhn, D. J. Proteasome inhibitors in cancer therapy: lessons from the first decade. *Clin Cancer Res* **14**, 1649–1657 (2008).
44. Voorhees, P. M. & Orlowski, R. Z. The proteasome and proteasome inhibitors in cancer therapy. *Annu Rev Pharmacol Toxicol* **46**, 189–213 (2006).
45. Rakha, E. A. *et al*. Prognostic markers in triple-negative breast cancer. *Cancer* **109**, 25–32 (2007).
46. Hudis, C. A. & Gianni, L. Triple-negative breast cancer: an unmet medical need. *Oncologist* **16 Suppl 1**, 1–11.
47. Hernandez-Aya, L. F. *et al*. Nodal status and clinical outcomes in a large cohort of patients with triple-negative breast cancer. *J Clin Oncol* **29**, 2628–2634.
48. McCall, M. N., Bolstad, B. M. & Irizarry, R. A. Frozen robust multiarray analysis (fRMA). *Biostatistics* **11**, 242–253.
49. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
50. Parker, J. S. *et al*. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160–1167 (2009).

Acknowledgement

This work was generously supported by grants from the Canadian Stem Cell Network, the Ontario Institute for Cancer Research and the Canadian Breast Cancer Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

RMH, ADG, AB, JAH, conception of project. RMH & ADG performed research. RMH, ADG, AB, JAH, wrote manuscript.

Additional information

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

Competing financial interests: The authors have filed a provisional patent application on using the described gene signature for prognosis of breast cancer patients.

How to cite this article: Hallett, R.M., Dvorkin-Gheva, A., Bane, A. & Hassell, J.A. A Gene Signature for Predicting Outcome in Patients with Basal-like Breast Cancer. *Sci. Rep.* **2**, 227; DOI:10.1038/srep00227 (2012).