ARTICLE OPEN

An assessment of prognostic immunity markers in breast cancer

Benlong Yang[1,2,3], Jeff Chou[4], Yaozhong Tao[5], Dengbin Wu[6], Xinhong Wu[7], Xueqing Li[8], Yan Li[5], Yiwei Chu[9], Feng Tang[10], Yanxia Shi[11], Linlin Ma[12], Tong Zhou[12], William Kaufmann[13], Lisa A Carey[5,14,15], Jiong Wu[1,2,3] and Zhiyuan Hu[5]

Tumor-infiltrating lymphocytes (TIL) and immunity gene signatures have been reported to be significantly prognostic in breast cancer but have not yet been applied for calculation of risk of recurrence in clinical assays. A compact set of 17 immunity genes was derived herein from an Affymetrix-derived gene expression dataset including 1951 patients (AFFY1951). The 17 immunity genes demonstrated significant prognostic stratification of estrogen receptor (ER)-negative breast cancer patients with high proliferation gene expression. Further analysis of blood and breast cancer single-cell RNA-seq datasets revealed that the 17 immunity genes were derived from TIL that were inactive in the blood and became active in tumor tissue. Expression of the 17 immunity genes was significantly ($p < 2.2\text{E-}16$, $n = 91$) correlated with TILs percentage on H&E in triple negative breast cancer. To demonstrate the impact of tumor immunity genes on prognosis, we built a Cox model to incorporate breast cancer subtypes, proliferation score and immunity score (72 gene panel) with significant prediction of outcomes ($p < 0.0001$, $n = 1951$). The 72 gene panel and its risk evaluation model were validated in two other published gene expression datasets including Illumina beads array data METABRIC ($p < 0.0001$, $n = 1997$) and whole transcriptomic mRNA-seq data TCGA ($p = 0.00019$, $n = 996$) and in our own targeted RNA-seq data TARGETSEQ ($p < 0.0001$, $n = 303$). Further examination of the 72 gene panel in single cell RNA-seq of tumors demonstrated tumor heterogeneity with more than two subtypes observed in each tumor. In conclusion, immunity gene expression was an important parameter for prognosis and should be incorporated into current multi-gene assays to improve assessment of risk of distant metastasis in breast cancer.

npj Breast Cancer (2018)4:35 ; doi:10.1038/s41523-018-0088-0

INTRODUCTION

Metastasis is the main cause of mortality for breast cancer patients. Factors such as cell cycle deregulation, stromal microenvironment, proteases, endothelial cells, myoepithelial cells and immunity status within a tumor can drive metastasis.[1] Targeted inhibition of immune checkpoint function by antibodies against PD-1,[2,3] PD-L1[4,5] and CTLA4[6] has revealed active anti-tumor, T cell-mediated immunity. Tumor-infiltrating lymphocytes (TIL) have been well-reported to play critical roles in response to chemotherapy and prognosis in breast cancer, specifically in triple-negative and HER2-positive breast cancers, with a survival benefit being seen in patients having >50% lymphocyte-predominant tumors.[7–9] Immunity-related gene classifiers have also been reported to stratify prognosis in immune-benefit-enabled tumors comprised mostly of Basal-like, HER2-enhanced (HER2E), and Luminal B tumors.[10] Quantitative assessment of anti-tumor immunity and responsiveness to immunotherapy represents an important new avenue of breast cancer research.

Gene expression profiles of primary tumors are highly predictive of distant metastasis[11–15] in breast cancer and the genomic portrait is maintained between the primary tumor and its metastases.[1,16,17] As the genetic and epigenetic properties of a primary tumor define its fate and capability to develop metastasis, the expression signatures of the primary tumor are prognostic and predict a patient's outcome. Three multi-gene expression assays, PCR-based Oncotype DX (Genomic Health Inc., Redwood City, CA, USA),[15,18] microarray-based MammaPrint (Agendia Inc., Amsterdam, Netherlands),[13,19] and nanostring-based PAM50 Prosigna Assay (NanoString Technologies Inc., Seattle, WA, USA),[20–23] have been widely used in clinical practice to determine the risk of recurrence in patients with breast cancer. Genes monitored in these assays mainly include drivers of cell proliferation, hormone receptors, HER2 and basal cytokeratins. The PAM50 expression assay with integration of breast cancer subtype and proliferation score in risk assessment was shown to provide better prognostic information in ER-positive, node-negative patients than Oncotype DX.[24] Agreement between risk classifications based on Oncotype DX and PAM50 was as low as 54%, demonstrating substantial differences between the molecular classifiers in patient risk stratification.[25] However, none of the current multi-gene

[1]Department of Breast Surgery, Shanghai Cancer Center, Shanghai, China; [2]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China; [3]Collaborative Innovation Center for Cancer Medicine, Shanghai, China; [4]Department of Biostatistics, Wake Forest Baptist Medical Center, Winston-Salem, NC, USA; [5]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [6]Department of Oncology, An-Steel Group Hospital, Anshan, Liaoning, China; [7]Department of Breast Surgery, Hubei Cancer Hospital, Huazhong University of Science and Technology, Wuhan, Hubei, China; [8]Department of Thyroid and Breast Surgery at the Fifth People's Hospital, Fudan University, Shanghai, China; [9]Department of Immunology, Fudan University, Shanghai, China; [10]Department of Pathology, Huashan Hospital, Fudan University, Shanghai, China; [11]Department of Medicine, Sun Yat-Sen University Cancer Center, Guangzhou, China; [12]Shanghai Precision Diagnostics Co. Ltd., Shanghai, China; [13]Asystbio Laboratories LLC, Chapel Hill, NC, USA; [14]North Carolina Cancer Hospital, Chapel Hill, NC, USA and [15]Division of Hematology-Oncology UNC School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Correspondence: Jiong Wu (wujiong1122@vip.sina.com) or Zhiyuan Hu (zhiyuanh@ad.unc.edu)

These authors contributed equally: Benlong Yang, Jeff Chou

Received: 19 April 2018 Revised: 20 September 2018 Accepted: 25 September 2018
Published online: 29 October 2018

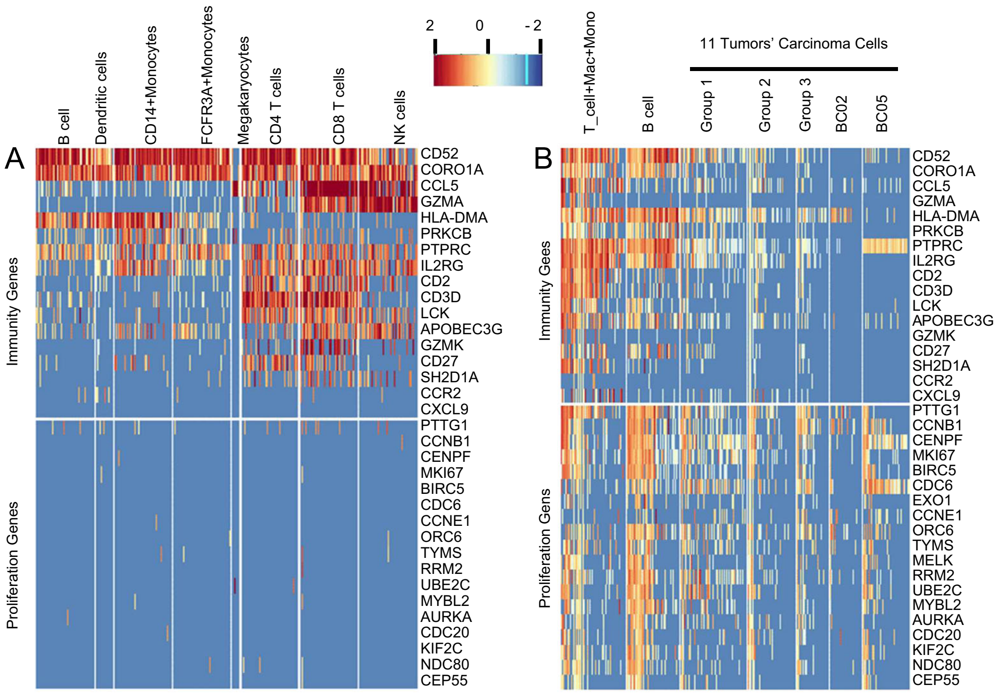

Fig. 1 RNA-seq gene expression of 17 immunity genes and 19 proliferation genes in published PBMC single-cell dataset (Macosko et al. Cell 2015)[26] and breast cancer solid tumor single cells (Chung et al. Nature Communications 2017).[27] (A) Expression of immunity and proliferation genes in different PBMC cell types including B cells, CD4 T cells, CD8 T cells, NK cells, Monocytes and dendritic cells. (B) Expression of immunity and proliferation genes of single cells including breast tumors' immune single cells labeled as T_cell + Mac + M & B_cell (Mac = macrophages, M = monocytes) and carcinoma single cells groups 1 to 3 (mixed carcinoma single cells from different tumors), BC02 and BC05 (carcinoma single cells from each individual tumor)

expression assays have included the very important and prognosis-related immunity genes. To improve accuracy in evaluating risk of distant metastasis of breast cancer we created a new model that added immunity genes based on gene expression profiling.

RESULTS

Prognostic immunity and proliferation genes in breast cancer

We analyzed 1951 Affymetrics gene expression profiles (AFFY1951) from 14 breast cancer cohorts with median follow-up of 7.12 years, median time-to-distant metastasis (DM) of 2.7 years and 481 DM events. There were 20% ER-negative, 69% ER-positive, and 11% unknown ER status in the AFFY1951 training data set. HER2 status for the 14 published cohorts was not provided (Supplementary Table 1). Two highly significant biological categories, immune response ($p < 0.001$) and cell cycle ($p < 0.001$) were identified with 119 and 71 genes in each category respectively (Supplementary Table 2). The Cox coefficient and the magnitude of change associated with distant metastasis-free survival (DMFS) were used for selection of immunity and cell cycle genes for further analysis. The top-ranked 17 immunity genes and 19 cell cycle genes were used for calculation of immunity and proliferation scores.

Single cell RNA-seq analysis of peripheral blood mononuclear cells (PBMC) using a publicly available dataset[26] indicated that 15 of the 17 immunity genes, excepting CCR2 and CXCL9, were expressed in at least one of eight different types of immune cells. However, none of the 19 cell cycle genes were expressed, indicating that the immune cells are not proliferating in the blood (Fig. 1a). We further analyzed gene expression in single cells including both immune cells and tumor cells isolated from solid breast tumor tissues.[27] The results showed that expression of the 17 immunity genes was evident in the two immune cell groups but sporadically or not in the five tumor carcinoma cell groups (Fig. 1b). A portion of the total immune cells in the 11 tumors (about 20% T cell group and 60% B cells) highly expressed the 19 proliferation genes while less than 20%, on average, of the carcinoma cells expressed the proliferation genes in this single cell RNA-seq data[27] (Fig. 1b).

We next compared the 17 immunity genes with other immunity signature modules. The 17 immunity genes were representatives of 119 immunity genes with high correlation (Pearson's correlation = 0.87, 95%CI: 0.86-0.88, $p < 0.0001$, $n = 1951$) derived by EPIG in the AFFY1951 dataset (Supplementary Figure 1A). We compared the 17-gene immunity signature with 500 other published immunity signatures (data not shown). The most highly correlated published immunity signatures were shown, as an example, in the TCGA breast cancer RNA-seq dataset (Supplementary Figure 1B). Correlation analysis revealed CD4 and CD8 T cell signatures, B cell signature, LCK signature, NK cell signature, Miller's immune signature, the UNC immune signature, and Cluadin-low upregulated gene signature were all significantly positively correlated with the 17-gene immunity signature with correlations from 0.77 to 0.94, but negatively (lowest −0.51) correlated with the Claudin-low down-regulated gene signature (Supplementary Figure 1B). We also noticed that the 17-gene immunity signature was the most compact significant gene signature and therefore suitable for further analysis for clinical application.

Patients were first divided into two groups based on either immunity scores or proliferation scores. In general, patients who were in the immunity-strong group (istrong) had better prognosis (DMFS) than those in the immunity-weak group (iweak), and patients who were in the high proliferation group had worse prognosis than those in the low proliferation group in AFFY1951

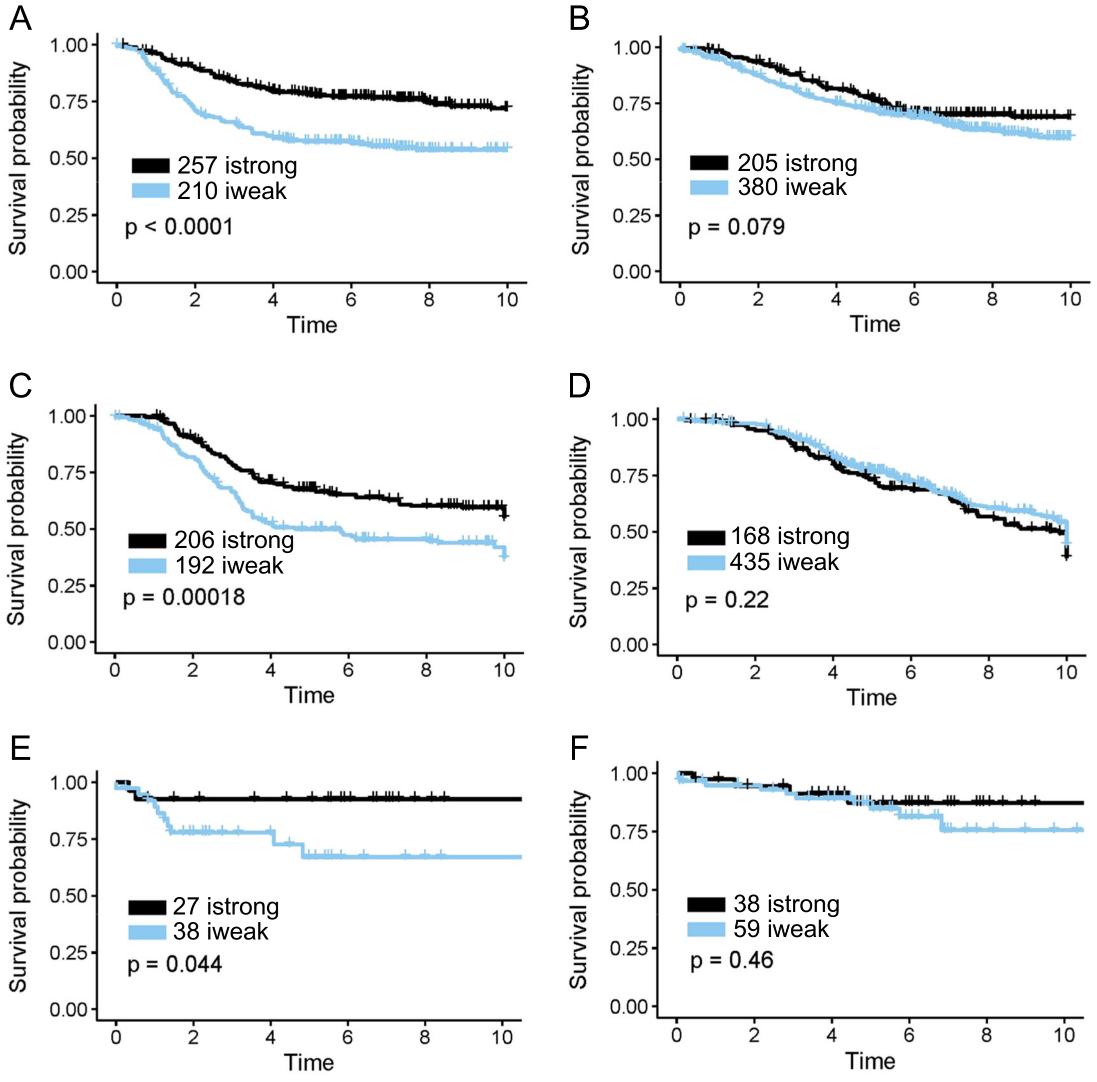

Fig. 2 Survival plots of Immunity Score in different patient groups identified by proliferation and ER status in the AFFY1951 training dataset, two test datasets METABRIC and TARGETSEQ. Immunity Score demonstrated strongest outcome prediction in patients who were ER-negative and proliferation high in AFFY1951 (A) ($p < 0.0001$, $n = 467$), METABRIC (C) ($p = 0.00018$, $n = 398$) and TARGETSEQ (E) ($p = 0.044$, $n = 65$), but was insignificant in ER-positive and proliferation high patients in AFFY1951 (B) ($p = 0.079$, $n = 585$), METABRIC (D) ($p = 0.22$, $n = 603$) and TARGETSEQ (F) ($p = 0.46$, $n = 97$). High proliferation groups had proliferation scores no less than 50 and low proliferation groups had proliferation scores less than 50. Survival analysis of ER-negative or ER-positive and low proliferation patients were demonstrated in Supplementary Figure 1

(Supplementary Figure 2). However, further analysis of the training dataset AFFY1951 showed that immunity score was prognostic of DMFS only in patients who were proliferation-high and ER-negative, and patients in the istrong group displayed a significantly better outcome (Fig. 2a, $p < 0.0001$). Immunity score had no significant effect on DMFS in all other patients who were either proliferation-high & ER-positive (Fig. 2b) or proliferation-low regardless of ER status (Supplementary Figure 3A and B). The same results were replicated in two independent validation datasets, the publicly available gene expression dataset "METABRIC"[28] of fresh-frozen breast tissues and our targeted RNA-seq dataset "TARGETSEQ" of breast cancer FFPE tissues, in proliferation-high and ER-negative (Fig. 2c,e) or ER-positive groups (Fig. 2d,f), and proliferation-low groups (Supplementary Figure 3C, D and E). Proliferation gene expression and ER status demonstrated significant impact on the prognostic value of immunity genes in breast cancer.

The "Immunity-enhanced" group and immunity score in evaluation of risk of distant metastasis

To further evaluate the significance of immunity genes and proliferation genes in prognosis, a 72-gene test panel, including the 17 immunity genes, 19 proliferation genes, 11 Basal genes, 14 ER genes, 3 HER2 genes, 2 invasion genes, and 6 housekeeper genes (Supplementary Table 3) was applied for subtype and immunity-adjusted risk of distant metastasis (iRDM) analysis. The five PAM50 breast cancer subtypes Luminal A (LumA), Luminal B (LumB), Basal-like (Basal), HER2-Enriched (HER2E) and Normal-like (Normal)[16,29,30] were recaptured by the iRDM analysis. Interestingly, an additional group termed "Immunity-enhanced" (Immuno) was identified (Supplementary Figure 4A. The new group accounting for about 18% of tumors (Table 1) demonstrated high expression of the 17 immunity genes and low or sporadic expression of the other breast cancer biomarker genes. Comparison of heatmaps sorted by expression of immunity genes in both

Table 1. Comparison of subtype classification between iRDM and PAM50 in four breast cancer datasets

	Method	Subtypes	Basal	HER2E	Immuno	Lum A	Lum B	Normal	Mixed*	Total
AFFY1951	iRDM	counts	310	211	342	441	351	209	87	1951
		percent	16%	11%	18%	23%	18%	11%	4%	100%
	PAM50	counts	361	244	NA	490	368	232	256	1951
		percent	19%	13%	NA	25%	19%	12%	13%	100%
METABRIC	iRDM	counts	298	208	383	425	387	199	97	1997
		percent	15%	10%	19%	21%	19%	10%	5%	100%
	PAM50	counts	339	255	NA	428	410	253	312	1997
		percent	17%	13%	NA	21%	21%	13%	16%	100%
TCGA	iRDM	counts	187	113	171	194	244	181	50	1140
		percent	16%	10%	15%	17%	21%	16%	4%	100%
	PAM50	counts	207	140	NA	262	256	171	104	1140
		percent	18%	12%	NA	23%	22%	15%	9%	100%
TARGETSEQ	iRDM	counts	112	49	80	120	55	41	26	483
		percent	23%	10%	16%	25%	11%	9%	5%	100%
	PAM50	counts	128	64	NA	134	67	33	57	483
		percent	26%	13%	NA	28%	14%	7%	12%	100%

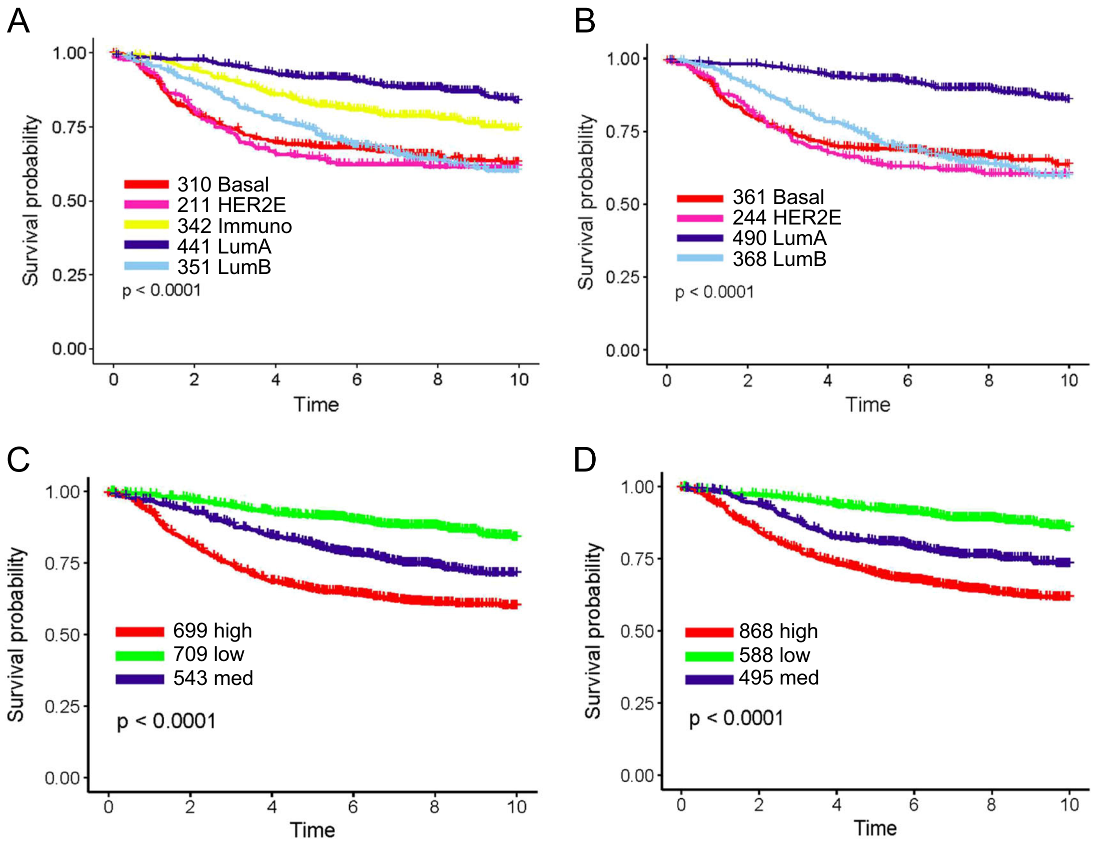

Fig. 3 Comparison of survival analysis of iRDM and PAM50 in AFFY1951 breast cancer training dataset. Subtypes and risk groups are color-coded: Basal-like (Red), HER2E (Hot Pink), Immuno (Yellow), Luminal A (Dark Blue), Luminal B (Sky Blue), Normal-like (Green); low (Green), med (Dark Blue), and high (Red) risks. Kaplan-Meier plots were used to show Distant Metastasis-Free Survival (DMFS) by subtypes for iRDM (A) ($p <$ 0.0001, $n = 1655$) and PAM50 (B) ($p < 0.0001$, $n = 1463$) and risk groups for iRDM (C) ($p < 0.0001$, $n = 1951$) and PAM50 (D) ($p < 0.0001$, $n = 1951$)

iRDM and PAM50 subtypes (Supplementary Figure 4A and B) using the AFFY1951 dataset showed that high immunity gene expression was present within each molecular subtype but not well-correlated with any of the other molecular markers. Excluding Normal-like and Mixed samples, iRDM subtypes demonstrated significant outcome prediction in the AFFY1951 dataset (Fig. 3a), very similar to the result of the PAM50 analysis (Fig. 3b) except for an additional Immunity-enhanced group that represented an intermediate outcome (Fig. 3a), worse than LumA but better than LumB, Basal and HER2E subtypes.

The samples in the iRDM Immunity-enhanced group were further classified using the PAM50 algorithm. In the AFFY1951

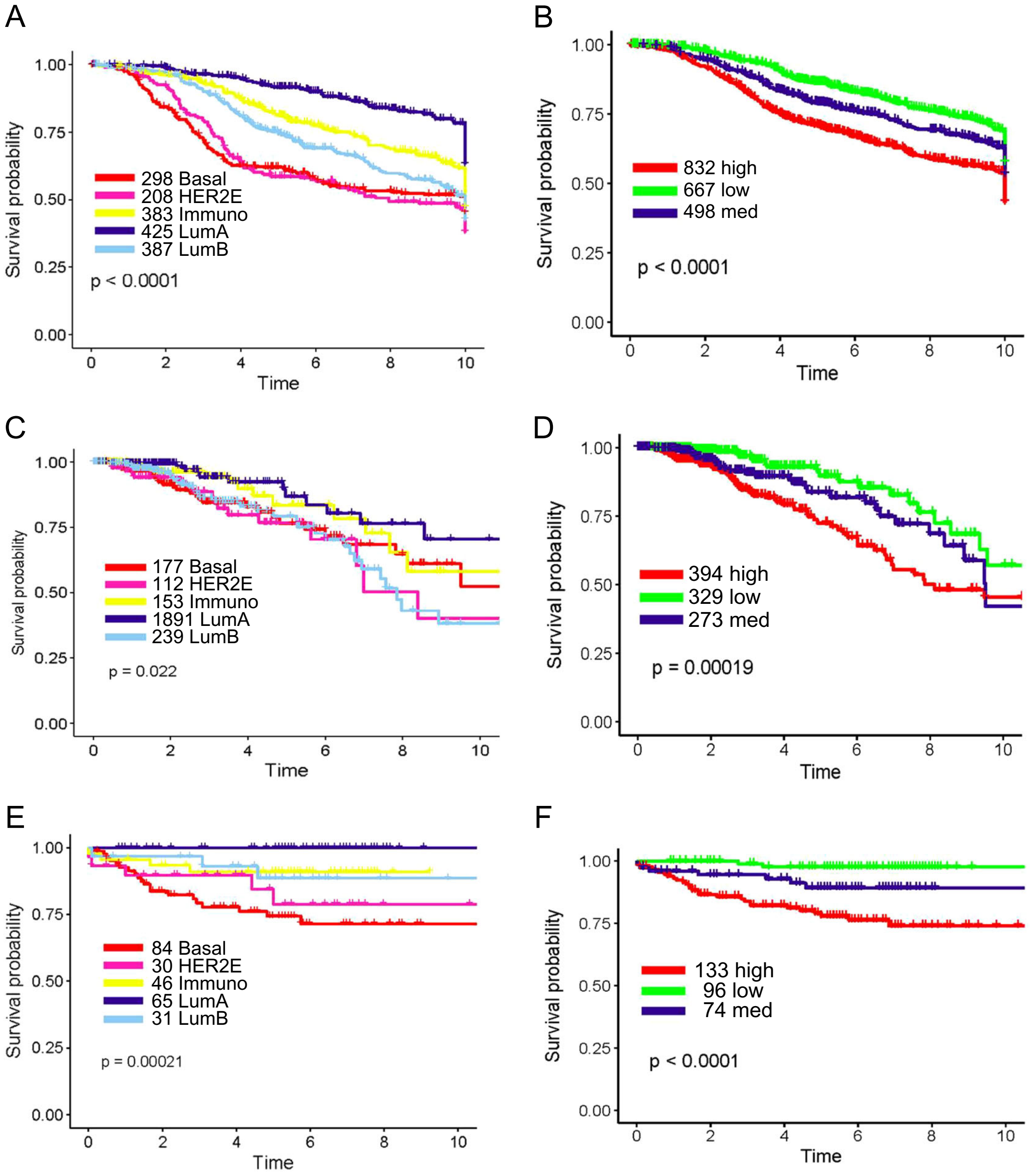

Fig. 4 Validation of iRDM subtype and risk survival analysis in three independent test datasets. Survival plots of iRDM subtypes in METABRIC (A) ($p < 0.0001$, $n = 1997$), TCGA (C) ($p = 0.022$, $n = 996$), TARGETSEQ (E) ($p = 0.00021$, $n = 303$) and survival plots of risk groups (high, low, med) in METABRIC (B) ($p < 0.0001$, $n = 1997$), TCGA (D) ($p = 0.00019$, $n = 996$), TARGETSEQ (E) ($p < 0.0001$, $n = 303$) were shown. Subtypes and risk groups are color-coded: Basal-like (Red), HER2E (Hot Pink), Immuno (Yellow), Luminal A (Dark Blue), Luminal B (Sky Blue), Normal-like (Green); low (Green), med (Dark Blue), and high (Red) risks

training dataset there were 342 high-immunity tumors which were further separated into 14% Basal, 14% HER2E, 15% LumA, 27% LumB, 3% Normal, and 27% Mixed (Supplementary Table 4). Samples classified as Mixed displayed <95% confidence for assignment to a subtype. As shown in the heatmap (Supplementary Figure 5), expression of the 17 immunity genes was high in the Immunity-enhanced group compared to other PAM50 subtyping genes which were expressed at lower levels if expressed at all.

Similar to the calculated risk of recurrence (ROR) score using PAM50, the iRDM score was calculated using an immunity score in addition to subtype and proliferation scores to adjust the risk of distant metastasis. The iRDM score was calculated by two equations depending on ER and proliferation status, as described in the Materials and Methods section. As seen with the PAM50 assay (Fig. 3d), iRDM also divided patients into three risk groups with low, intermediate (med) and high risk of distant metastasis. The three risk groups demonstrated significant ($p < 0.0001$, $n = 1951$) outcome prediction in the training dataset AFFY1951 (Fig. 3c). Considering the impact of immunity genes on DMFS, iRDM significantly adjusted more patients into the lower risk category (Fig. 3c,d).

In both PAM50 and iRDM algorithms in this study, samples with low confidence (confidence < 0.95) were classified into a "Mixed" group (Table 1). A Mixed sample was not treated as a subtype as its gene expression pattern did not correlate well with any defined subtype. The percentage of Mixed samples was significantly reduced using the iRDM assay when compared with PAM50 (Table 1). The Immunity-enhanced group accounted for about 40% of the Mixed group defined by PAM50 (Data not shown).

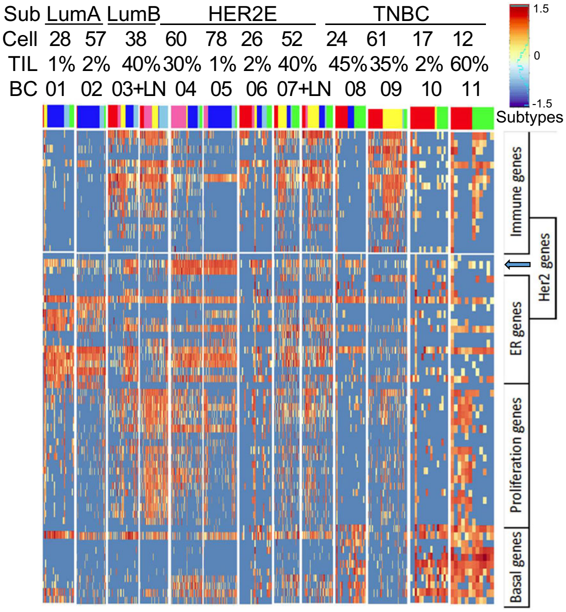

Fig. 5 Heatmap of iRDM subtypes analyzed in 549 single cells from 11 primary breast tumors and two lymph node metastases. All tumors showed two or more iRDM subtypes. Sub, subtype; Cell, single cell number; TIL, percentage of tumor-infiltrating lymphocytes. Color-coded individual cell subtypes: Basal-like (Red), HER2E (Hot Pink), Immuno (Yellow), Luminal A (Dark Blue), Luminal B (Sky Blue), Normal-like (Green), Mixed (Black)

The same results were observed in the three validation datasets, METABRIC, TCGA and TARGETSEQ (Fig. 4, Supplementary Figure 4) (Table 1). Overall survival (OS) was used as the outcome endpoint for the TCGA dataset and the results were slightly less significant in survival analysis compared to the other two datasets where DMFS was used.

One of the characteristics of the Claudin-low subtype (CLOW) of triple-negative breast cancers is high expression of immunity genes.[31] To see if Immunity-enhanced and CLOW subtypes were the same, we monitored CLOW subtype tumors using expression of the top 80 CLOW signature genes.[31] Less than 30% of the Immunity-enhanced tumors were classified as CLOW in the AFFY1951 dataset (Supplementary Figure 6A) although expression profiles of CLOW and Not-CLOW subgroups were very similar with a Pearson correlation of 0.97 (95% CI: 0.95–0.98, $p < 0.00001$). There was no prognostic difference between the CLOW and Not-CLOW subgroups (Supplementary Figure 6D) within the iRDM-defined Immunity-enhanced group. Similar results were observed in the METABRIC data with a Pearson correlation of 0.82 (95% CI: 0.74–0.88, $p < 0.00001$) and the TCGA data with a Pearson correlation of 0.92 (95% CI: 0.88–0.95, $p < 0.00001$) (Supplementary Figure 6) as two validation datasets. As the pattern of immunity gene expression in the CLOW tumors was the same as in the Not-CLOW tumors but of increased intensity, the CLOW subtype tumors appear to represent Immunity-enhanced tumors with the greatest level of immunity gene expression.

The immunity gene signature included B and T lymphocyte transcripts (Fig. 1) indicating a population of lymphocytes was present within breast cancers. To compare the gene expression-based immunity score with pathologist-assessed TIL, we retrieved 91H&E slides from triple-negative tumors with corresponding immunity scores (Supplementary Table 5). The average percentages of TIL in each tumor were determined by an experienced, licensed pathologist using an internationally recommended method[8] (see also: www.tilsinbreastcancer.org). TIL percentages and immunity scores were significantly correlated (Pearson's correlation = 0.75, 95%CI: 0.64 to 0.83, p value < 2.2E-16). As expected both TIL and immunity score predicted DMFS in the 91 patients (Supplementary Figure 7A & B). Patients with high TIL (>50% infiltrating lymphocytes) and high immunity scores had the best DMFS (Supplementary Figure 7C). Only two patients had high TIL and a low immunity score, and these patients are not shown. Among patients with low TIL (< 50% infiltrating lymphocytes), the immunity score influenced outcome, with a high immunity score portending better DMFS than a low immunity score (Supplementary Figure 7C).

Tumor heterogeneity within breast cancer subtype

Single-cell expression of the 72 genes was also analyzed in 11 tumor samples containing 12 to 78 single cells.[27] The subtype of each bulk tumor was defined by immunohistochemistry (IHC) in the original paper, identifying 2 LumA, 1 LumB, 4 HER2E and 4 triple-negative breast cancers (TNBC). We used iRDM algorithm to identify each cell's subtype (Supplementary Figure 8). Single cell analysis indicated that tumor cells in each tumor displayed expression of at least two subtypes detected by iRDM (Fig. 5), showing heterogeneity of gene expression within a tumor. The correlation between the percentage of TIL estimated in the tumor and immunity gene expression was of intermediate strength. In most cases, tumors with a high percentage of TIL had higher expression of immunity genes and tumors with a low percentage of TIL had lower or no expression of immunity genes. However, there was a tumor with high TIL and low immunity gene expression (BC08) and a tumor with low TIL and high immunity gene expression (BC06). Immunity gene expression did not appear to coincide with expression of basal cytokeratins or signals from HER2 (ERBB2) and steroid receptors, implying that the expression of immunity genes was not correlated with breast cancer subtypes (Fig. 5).

DISCUSSION

Recent studies have identified genes that influence anti-tumor immunity.[3–6,10,32,33] In this paper, we identified 17 co-expressed immunity genes that, as a group in gene ontology analysis, play an important role in immune response and its regulation. The Immunity-enhanced group identified in this analysis consisted in part of previously unclassified tumors with relatively low expression of the PAM50 intrinsic subtype marker genes. It also included the Claudin-low subtype of triple-negative breast cancers. We defined Immunity-enhanced patients as a new prognostic group instead of a subtype due to subsequent data analysis in single cells showing that the majority of immunity gene expression was contributed by infiltrating immune cells in tumors. The iRDM algorithm improved the accuracy for breast cancer classification by significantly reducing the frequency of unclassified Mixed tumors in four independent datasets. Overexpression of the 17 immunity genes was found to be predictive of a good or better prognosis, meaning a lower risk of cancer recurrence, metastasis or death. This observation supports others' prognostic analyses of breast cancer and ovarian cancer.[10,34] In the current study, one important finding was immunity genes play a significant role in breast cancer prognosis only in patients whose tumors are estrogen-negative and highly proliferative, accounting for about 24% of all patients in our training dataset. This was corroborated by gene expression analysis of tumor single cells showing that expression of proliferation genes can sometimes be contributed by active TIL instead of carcinoma cells. To estimate more precisely the risk of distant metastasis for these patients, it is necessary to incorporate an immunity score, besides the intrinsic subtype and proliferation scores as used in the PAM50 assay, into the equation for calculation of risk.

The Immunity-enhanced group in breast cancer was identified by adding a set of 17 co-expressed immunity genes into the PAM50 marker genes. Interestingly, Immunity-enhanced group tumors had high expression of genes that were up-regulated in Claudin-low subtype tumors. We considered whether the Claudin-low subtype was similar to the Immunity-enhanced group as both have high immunity gene expression and low or no intrinsic subtyping gene expression. Only one third of tumors in the Immunity-enhanced group expressed the Claudin-low signature genes above the cutoff threshold for the Claudin-low subtyping algorithm.[31] However, the Claudin-low and not-Claudin-low subtypes within the Immunity-enhanced group were highly correlated and had the same prognosis. Thus, we conclude that Claudin-low subtype tumors represent Immunity-enhanced group tumors with the greatest levels of immunity gene expression. It is notable that anti-tumor immunity within CLOW tumors appears to include immune-suppressive T lymphocytes.[35]

For tumors with high TIL and low expression of HER2 (ERBB2), ER, PR, basal keratins and proliferation genes, the PAM50 classification spreads them across five subtypes but predominantly in the uncertain Mixed category. The presence of the immune cells does not appear to affect expression of the other PAM50 markers as, for the most part, there was no correlation between high immunity gene expression and any of the other marker genes (Fig. 1B and Supplementary Figure 4). When the PAM50 algorithm was used to cluster genes including the 17 immunity genes, there was a clear gradation of immunity in all subtypes and generally without gradation of the other markers within each subtype. This suggests that Immunity-enhanced tumors are not a subtype within a PAM50-defined subtype, but a distinct group of breast cancers, similar to the Immunity-enhanced groups in melanoma and ovarian cancer.[36]

The lack of a strong correlation between immunity gene expression and the other molecular markers indicates that the low expression of PAM50 markers in the Immunity-enhanced group was not due to dilution of tumor mRNA with infiltrating lymphocytes. The major reason why immunity genes were not observed in the intrinsic subtype analysis is the algorithm filters out variable genes, such as the highly variable immunity genes, within a subtype. There appears to be a distinct subset of breast cancers with low expression of the PAM50 markers and high expression of immunity genes.

The Immunity-enhanced group of breast cancer with relative low expression of ER, PR, HER2, basal cytokeratins and proliferation drivers might be less responsive to treatments with anti-estrogen, Herceptin and general chemotherapies, but may benefit more from the immunotherapies. The immunity score could be a companion diagnostic maker in addition to PDL-1 expression for immune-checkpoint inhibitors. A clinical trial that stratifies patients based on subtypes including the Immunity-enhanced group may distinguish patients with high anti-tumor immunity from those with low anti-tumor immunity and provide more precise design of trials testing the efficacy of immune-checkpoint inhibitors.

Several clinical trials have demonstrated that number of TIL are prognostic in breast cancers.[8,9,37,38] Pathologic TIL on H&E slides were significantly correlated with immunity scores based on our analysis of 91 triple-negative breast cancers. Our analysis of single-cell RNA-seq data indicated that expression of the immunity genes was contributed mainly by immune cells, not carcinoma cells, suggesting that expression levels of immunity genes may reflect the number of TIL. Single-cell analysis also indicated that only a portion of immune cells, not all TIL, were proliferative in some solid tumors. We speculate that only the proliferative TIL are active in anti-tumor immunity and this may explain why some patients having high TIL had poor prognosis in clinical studies.[33,34]

Tumor heterogeneity was also studied by analyzing expression of the 72 genes in single tumor cells. Of the 11 tumors analyzed in the current study, each tumor had at least two intrinsic subtypes. This observation needs further validation by conducting large-scale single-cell RNA-seq of solid tumors using advanced technologies such as 10x Genomics to evaluate its value in clinical design of adjuvant therapies for breast cancer patients.

In summary, a set of immunity genes was extracted through analysis of a large dataset of breast cancers. High expression of the immunity genes identified an Immunity-enhanced group and indicated a better prognosis in ER-negative and high-proliferation breast cancers. Single-cell sequencing provides a useful tool for mechanistic studies of tumor immunity and heterogeneity with more studies needed to evaluate its clinical value.

METHODS

Patients

This study included 225 anonymous patients from a multi-center study in Shanghai and 250 patients from the University of North Carolina at Chapel Hill (UNC-CH). 225 breast tumor FFPE blocks were obtained from patients hospitalized and receiving modified radical mastectomy or lumpectomy in Shanghai Cancer Hospital. The patient-anonymous 250 FFPE tissues were obtained from UNC Hospital and were part of a molecular epidemiology study, LCCC-9830. The study was approved by the two independent institutional review boards (IRB) at the Shanghai Tumor Institute at Fudan University and UNC at Chapel Hill.

Microarray data mining and analysis

Affymetrix microarray data sets for 2034 patients from fourteen breast cancer cohort studies were retrieved from GEO: GSE1112, GSE12093, GSE1456, GSE2034, GSE2603, GSE3494, GSE4922, GSE5327, GSE6532, GSE7378, GSE7390, GSE8193, GSE9195, and ArrayExpress|E-TABM-158.[39,40] The downloaded individual CEL files were first processed by Robust Multi-chip Average[41] and then merged into one dataset of 2034 expression profiles which were further batch-corrected using Combat[42] with subtype as covariate. An unsupervised analysis of the 2034 expression profiles using the pattern-recognition algorithm EPIG[43,44] was performed to identify sets of co-expressed genes. Two co-expressed gene clusters with significant enrichment of gene ontology categories “Immune Response” and “Cell Cycle” were identified (Supplementary Table 2). 119 immunity and 71 cell cycle genes were consistently selected in 1000 iterations by EPIG, in which 80% of the 2034 expression profiles were randomly selected in each iteration. Gene numbers were subsequently reduced by selecting the top-weighted EPIG values, by correlation with the 119 immunity genes ($R > 0.87$, Figure 8), and by using gene ontology to eliminate duplicate functional genes. A compact 17 immunity gene signature was generated containing *APOBEC3G, CCL5, CCR2, CD2, CD27, CD3D, CD52, CORO1A, CXCL9, GZMA, GZMK, HLA-DMA, IL2RG, LCK, PRKCB, PTPRC*, and *SH2D1A*. We used the same methods to identify 19 proliferation genes including *AURKA, BIRC5, CCNB1, CCNE1, CDC20, CDC6, CENPF, CEP55, EXO1, MKI67, KIF2C, MELK, MYBL2, NDC80, ORC6, PTTG1, RRM2, TYMS, and UBE2C*. 1951 patients of the 2034 patients had follow-up and clinical data which were combined using the same established method as described[45] and this group was named “AFFY1951”.

In addition to the AFFY1951 training dataset, we also assembled three test datasets on different platforms including Illumina beads arrays of 1997 fresh-frozen breast tumors (METABRIC),[28] RNA-seq of 1140 fresh-frozen breast tumors (TCGA)[46] and targeted RNA-seq of breast cancer FFPE tissues from 225 samples from Shanghai Cancer Hospital and 258 samples (250 patients with 8 duplicate samples) from UNC-CH (TARGETSEQ, GSE113863).

Calculation of Proliferation and Immunity Scores

A proliferation score was calculated by averaging expression levels of the afore-mentioned 19 proliferation genes as “unscaled proliferation score” in a sample and then scaled between 0 and 100 using the formula: $38 \times$ (unscaled proliferation score $+ 1.35$). Proliferation-high were those samples with proliferation score larger or equal to 50 while proliferation-low were the others with score less than 50.

An Immunity score was calculated by averaging gene expression values of the above 17 immunity genes as “unscaled immunity score” and then scaled between 0 and 100 for each sample using the formula: $30\times$ (unscaled immunity score $+ 1.4$). For Immunity score group classification, the patients were divided into two groups, “iweak” and “istrong”, based on their Immunity score values using the cut-off value of 42 that was derived from the combined data using X-tile.[47]

Breast Cancer Molecular Subtyping

A panel of 72 genes, consisting of the 17 immunity genes, 19 proliferation genes, 11 Basal genes, 14 ER genes, 3 HER2E genes, 2 invasion genes, and 6 housekeeper genes was formed for analysis of breast cancer subtypes (Supplementary Table 3). Ten-fold CV included different statistical predictors including PAM,[48] a k-Nearest Neighbor Classifier (KNN) with either Euclidean distance or one-minus-Spearman-correlation as the distance function and a Class Nearest Centroid (CNC) metric with either Euclidean distance or one-minus Spearman-correlation as the distance function. A sample was assigned the subtype corresponding to the highest one-minus Spearman-correlation value among the six values versus centroids for iRDM subtypes: Basal, HER2E, Immuno, LumA, LumB, and Normal based on Single Sample Predictor algorithms for subtyping breast cancer.[23,29,49]

Confidence intervals for each subtype identification were calculated[50] and a subtype with confidence lower than 95% was called "Mixed". The formula for confidence calculation is: Subtyping "confidence" = 1 - Spearman's test p value.

Survival plots were done using R package Survminer (downloaded from Bioconductor "RTCGA") which provided censored survival curves. In addition, Univariate Kaplan-Meier survival analysis was performed for validation using WINSTAT for EXCEL® (R. Fitch Software, Lehigh Valley, Pa.).

The iRDM prediction model

Factors included in the model to optimize an outcome predictor were molecular subtype, proliferation score and immunity score that were calculated based on expression profiles of the 72-gene panel. We slightly modified established algorithm[23] for iRDM and used Cox models for iRDM score calculation in which Distant Metastasis-Free Survival Time (DMFS) was used with patient follow-up for up to 10 years. A subset of 404 patients was selected as a training dataset through ranking of correlation to the centroids with a cutoff at 0.7. Coefficients for each subtype were calculated using this Cox model and used as constant factors for subtype Spearman correlations, proliferation score, and immunity score. The immunity-stratified Risk of Distance Metastasis (iRDM) was calculated using two formulas:

A. For the proliferation-high and ER-negative group only:

$$\begin{aligned}\text{Unscaled iRDM score} = {} & -(0.02\times \text{Basal}) + 0.16\times \text{HER2E} + (-0.34\times \text{Immuno})\\ & + 0.07\times \text{LumA} + 0.08\times \text{LumB} + 0.09\times \text{Proliferation Score}\\ & + (-0.40\times \text{Immunity Score})\end{aligned}$$

B. For all other groups except A:

$$\begin{aligned}\text{Unscaled iRDM score} = {} & 0.40\times \text{Basal} + 0.48\times \text{HER2E} + (-0.06\times \text{Immuno})\\ & + (-0.46\times \text{LumA}) + 0.19\times \text{LumB} + 0.24\\ & \times \text{Proliferation Score} + (-0.08\times \text{Immunity Score})\end{aligned}$$

The unscaled iRDM score was further scaled to values spanning 0 to 100 by the formula:

$$\text{iRDM score} = 90\times \text{Unscaled iRDM score} + 50$$

Patients were categorized into three groups, low, intermediate (med) and high risk, according to iRDM scores (range 0–100) with cutoffs at 33 and 50 optimized by X tile.[47]

Targeted RNA Expression by RNA-seq

For breast tumor FFPE tissues RNA extraction was routinely performed using Roche FFPE RNA extraction kit according to the manufacturer's protocol. The Illumina TruSeq Targeted RNA expression kit was used to build libraries of the targeted 72-genes. To synthesize cDNA, 200 to 800 nanograms of purified FFPE RNA in a total volume of 3 µl was mixed with 4.0 µl RCS1, 2.0 µl ProtoScrip II Reverse Transcriptase, 1.0 µl 10 mM DTT at 42 °C for 30 min and 94 °C for 10 min. The cDNA was hybridized with custom oligo pools in a thermal cycler programed to gradually decrease temperature from 70 °C to 30 °C in 30 min. The RNA/Oligo hybrid products were washed, extended and ligated. The ligated DNA was amplified by DNA polymerase on the thermal cycler with 35 PCR cycles of 98 °C for 30 s, 62 °C for 30 s and 72 °C for 60 s. The PCR products were purified with AMPure XP beads and eluted in 15 µl of buffer, measured using Agilent Bioanalyzer2100 and DNA1000 chips, pooled with equal amounts of DNA from each sample's library, and finally diluted to 4 nM, denatured, and loaded to NextSeq 500 according to the manufacturer's protocol. Illumina Casava1.7 software was used for basecalling and sequencing data were demultiplexed with Illumina bcl2fastq2 software to generate one fastq file per sample. To ensure sequencing data integrity of libraries derived from FFPE RNA tissues, only samples with total reads larger than 10000 and missing genes less than 30% of all 72 genes were further processed in the validation study. Single read 1 sequence in each fastq file were mapped to known targeted regions of human genomes to generate raw counts using R package ShortRead. Raw counts of all samples were normalized by the size of the transcripts and by the size of the library and then calculated for CPM per sample as a gene expression matrix using R package edgeR from Bioconductor and finally log based 2 transformed and imputed by KNN method. Gene expression data were further median-centered and column-standardized. We also used Distance Weighted Discrimination DWD[51] to combine 225 Targeted RNA samples from Shanghai Cancer Hospital and 258 Targeted RNA samples including 8 technical repeats from UNC-CH to create a test dataset denoted "TARGETSEQ" (GSE113863) in which 303 patients had DMFS time and event (Supplementary Table 4).

Histopathological TIL-assessment

The TILs were evaluated by a trained pathologist using an internally recommended method.[8] See also www.tilsinbreastcancer.org for a freely available training tool for the assessment of TILs by pathologists on HE-slides.

Study approval and consent to participate

All studies were carried out according to institutional guidelines, and with appropriate informed consent from participants. Institutional ethics committees of the clinical centers where samples were collected reviewed and approved all protocols. The Institutional Review Board of the Shanghai Cancer Center at Fudan University and University of North Carolina separately approved procurement and handling of the human materials. All data were analyzed anonymously.

DATA AVAILABILITY

Original primary sequencing data and processed data for Targeted RNA-seq data TARGETSEQ created in this study were submitted to GEO sequencing data (GSE113863) at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113863. Published data for generating training data AFFFY1951 were retrieved from GEO: GSE1112, GSE12093, GSE1456, GSE2034, GSE2603, GSE3494, GSE4922, GSE5327, GSE6532, GSE7378, GSE7390, GSE8193, GSE9195, and ArrayExpress|E-TABM-15839,40. METABRIC gene expression data are available at the European Genome-Phenome Archive http://www.ebi.ac.uk/ega/ which is hosted by the European Bioinformatics Institute, under accession number EGAS00000000083. TCGA breast cancer sequencing data are available in CGHub (https://cghub.ucsc.edu/) and sample lists, data matrices and supporting data can be found at (http://tcga-data.nci.nih.gov/docs/publications/brca_2012/). A RNA-seq dataset of 2,700 Peripheral Blood Mononuclear Cells (PBMC) single cells is freely available from 10X Genomics that were sequenced on the Illumina NextSeq 500 and the raw data can be found on http://satijalab.org/seurat/pbmc3k_tutorial.html. TPM counts of eleven solid breast tumors' single cell RNA-seq can be downloaded from GEO (GSE75688) https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75688.

ACKNOWLEDGEMENTS

We thank Dr. Charles M. Perou (UNC Lineberger Comprehensive Cancer Center) for providing FFPE tissues of breast tumors to perform RNA extraction and Targeted RNA-seq in the Rapid Adoption Molecular Laboratory in Lineberger Comprehensive Cancer Center at University of North Carolina at Chapel Hill. We thank Shanghai Major Biomedical Laboratory for providing RNA-seq services on Illumina Nexseq500. Jiong Wu and Benlong Yang were supported by Chinese Natural Science Foundation (81472456 and 81272924) and Shanghai Municipal Science and Technology Commission International Joint Research Foundation (15410724100) awarded to Jiong Wu. Benlong Yang was also supported in part by a grant awarded by the Shanghai Science and Technology Committee Yang Fan project (15YF1402300). Lisa A. Carey's grant for 9830 projects was supported by Breast Cancer Research Foundation (BCRF). Zhiyuan Hu was supported by The University Cancer Research Fund (UCRF), a State of North Carolina initiative to fund innovative cancer research, administered at Lineberger Comprehensive Cancer Center in University of North Carolina at Chapel Hill.

AUTHOR CONTRIBUTIONS

B.Y., J.W., and Z.H. conceived and supervised the study and obtained funding. B.Y., J.C., T. Z., and Z.H. conceived and performed the statistical and bioinformatics analyses. B.Y., J.C., Y.T., D.W., X.W., X.L., Y.L., Y.C., Y.S., T.Z., W.K., L.C., J.W., and Z.H. contributed materials and discussion. FT evaluated TILs on H&E slides. LM retrieved and collected Chinese Han patients' clinical data. L.C. provided deidentified US patients' survival clinical data. B.Y.

and Y.T. performed the experiments. B.Y., J.C., Y.C., T.Z., W.K., J.W., and Z.H. wrote the manuscript with input from all authors. All authors approved the final manuscript.

ADDITIONAL INFORMATION

Supplementary Information accompanies the paper on the *npj Breast Cancer* website (https://doi.org/10.1038/s41523-018-0088-0).

Competing interests: The Authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Hu, Z. et al. A compact VEGF signature associated with distant metastases and poor outcomes. *BMC Med* **7**, 9 (2009).
2. Lesokhin, A. M. et al. Nivolumab in Patients With Relapsed or Refractory Hematologic Malignancy: Preliminary Results of a Phase Ib Study. *J. Clin. Oncol.* **34**, 2698–2704 (2016).
3. Tumeh, P. C. et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568–571 (2014).
4. Herbst, R. S. et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature* **515**, 563–567 (2014).
5. Powles, T. et al. MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer. *Nature* **515**, 558–562 (2014).
6. Snyder, A. et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med* **371**, 2189–2199 (2015).
7. Stanton, S. E. & Disis, M. L. Clinical significance of tumor-infiltrating lymphocytes in breast cancer. *J. Immunother. Cancer* **4**, 59 (2016).
8. Salgado, R. et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Ann. Oncol.* **26**, 259–271 (2015).
9. Denkert, C. et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol.* **19**, 40–50 (2018).
10. Miller, L. D. et al. Immunogenic Subtypes of Breast Cancer Delineated by Gene Classifiers of Immune Responsiveness. *Cancer Immunol. Res* **4**, 600–610 (2016).
11. Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. A molecular signature of metastasis in primary solid tumors. *Nat. Genet* **33**, 49–54 (2003).
12. 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).
13. van 't Veer, L. J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
14. Wang, Y. et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679 (2005).
15. 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).
16. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
17. Weigelt, B. et al. Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer Res* **65**, 9155–9158 (2005).
18. Sparano, J. A. & Paik, S. Development of the 21-gene assay and its application in clinical practice and clinical trials. *J. Clin. Oncol.* **26**, 721–728 (2008).
19. Slodkowska, E. A. & Ross, J. S. MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Rev. Mol. Diagn.* **9**, 417–422 (2009).
20. Filipits, M. et al. The PAM50 risk-of-recurrence score predicts risk for late distant recurrence after endocrine therapy in postmenopausal women with endocrine-responsive early breast cancer. *Clin. Cancer Res* **20**, 1298–1305 (2014).
21. Prat, A. et al. Research-based PAM50 subtype predictor identifies higher responses and improved survival outcomes in HER2-positive breast cancer in the NOAH study. *Clin. Cancer Res* **20**, 511–521 (2014).
22. Dieci, M.V. et al. Integrated evaluation of PAM50 subtypes and immune modulation of pCR in HER2-positive breast cancer patients treated with chemotherapy and HER2-targeted agents in the CherLOB trial. *Ann Oncol* (2016).
23. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
24. Dowsett, M. et al. Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *J. Clin. Oncol.* **31**, 2783–2790 (2009).
25. Alvarado, M. D. et al. A Prospective Comparison of the 21-Gene Recurrence Score and the PAM50-Based Prosigna in Estrogen Receptor-Positive Early-Stage Breast Cancer. *Adv. Ther.* **32**, 1237–1247 (2015).
26. Macosko, E. Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
27. Chung, W. et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications* **8**, 15081 (2017).
28. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
29. Hu, Z. et al. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genom.* **7**, 96 (2006).
30. Sorlie, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* **100**, 8418–8423 (2003).
31. Prat, A. et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* **12**, R68 (2010).
32. Alistar, A. et al. Dual roles for immune metagenes in breast cancer prognosis and therapy prediction. *Genome Med* **6**, 80 (2014).
33. Wolchok, J. D. & Chan, T. A. Cancer: Antitumour immunity gets a boost. *Nature* **515**, 496–498 (2014).
34. Iglesia, M. D. et al. Prognostic B-cell signatures using mRNA-seq in patients with subtype-specific breast and ovarian cancer. *Clin. Cancer Res* **20**, 3818–3829 (2014).
35. Taylor, N. A. et al. Treg depletion potentiates checkpoint inhibition in claudin-low breast cancer. *J. Clin. Invest* **127**, 3472–3483 (2017).
36. Iglesia, M.D. et al. Genomic Analysis of Immune Cell Infiltrates Across 11 Tumor Types. *J Natl Cancer Inst* 108, https://doi.org/10.1093/jnci/djw144 (2016).
37. Mao, Y. et al. The Prognostic Value of Tumor-Infiltrating Lymphocytes in Breast Cancer: A Systematic Review and Meta-Analysis. *PLoS One* **11**, e0152500 (2016).
38. Yue, Q. et al. The prognostic value of Foxp3 + tumor-infiltrating lymphocytes in patients with glioblastoma. *J. Neurooncol* **116**, 251–259 (2013).
39. Loi, S. et al. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genom.* **9**, 239 (2008).
40. Loi, S. et al. PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor-positive breast cancer. *Proc. Natl. Acad. Sci. USA* **107**, 10208–10213 (2010).
41. Irizarry, R. A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
42. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
43. Chou, J. W., Zhou, T., Kaufmann, W. K., Paules, R. S. & Bushel, P. R. Extracting gene expression patterns and identifying co-expressed genes from microarray data reveals biologically responsive processes. *BMC Bioinforma.* **8**, 427 (2007).
44. Zhou, T. et al. Profiles of global gene expression in ionizing-radiation-damaged human diploid fibroblasts reveal synchronization behind the G1 checkpoint in a G0-like state of quiescence. *Environ. Health Perspect.* **114**, 553–559 (2006).
45. Nagalla, S. et al. Interactions between immunity, proliferation and molecular subtype in breast cancer prognosis. *Genome Biol.* **14**, R34 (2013).
46. TCGA. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
47. Camp, R. L., Dolled-Filhart, M. & Rimm, D. L. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin. Cancer Res* **10**, 7252–7259 (2004).
48. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 6567–6572 (2002).
49. Fan, C. et al. Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med* **355**, 560–569 (2006).
50. Newcombe, R. G. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat. Med* **17**, 857–872 (1998).
51. Benito, M. et al. Adjustment of systematic microarray data biases. *Bioinformatics* **20**, 105–114 (2004).