Tumor Cell Intrinsic And Extrinsic Features Predicts Prognosis In Estrogen Receptor Positive Breast Cancer

Kevin Yao
Texas A&M University College Station: Texas A&M University
Evelien Schaafsma
Dartmouth Medical School: Dartmouth College Geisel School of Medicine
Baoyi Zhang
Rice University
Chao Cheng (chao.cheng@bcm.edu)
Baylor College of Medicine https://orcid.org/0000-0002-5002-3417

Research article

Keywords: Genomic aberration, immune infiltration, breast cancer, prognostic prediction

Posted Date: August 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-829187/v1

Version of Record: A version of this preprint was published at PLOS Computational Biology on March 9th, 2022. See the published version at https://doi.org/10.1371/journal.pcbi.1009495.

Tumor cell intrinsic and extrinsic features predicts prognosis in estrogen receptor positive breast cancer

Running title: Intrinsic/extrinsic features predict breast cancer prognosis

Kevin Yao[1], Evelien Schaafsma[2,3], Baoyi Zhang[4], Chao Cheng[5,6,7]*

1. Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA.

2. Department of Molecular and Systems Biology, Dartmouth College, Lebanon, NH, USA.

3. Department of Biomedical Data Science, The Geisel School of Medicine at Dartmouth College, Lebanon, NH, USA.

4. Department of Chemical and Biomolecular Engineering, Rice University, Houston, Texas, USA

5. Department of Medicine, Baylor College of Medicine, Houston, TX 77030, USA

6. Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA

7. Institute for Clinical and Transcriptional Research, Baylor College of Medicine, Houston, TX 77030, USA

*Corresponding author

Email: chao.cheng@bcm.edu

Phone: 713-798-3332

Abstract:

Background: Although estrogen-receptor-positive (ER+) breast cancer is generally associated with favorable prognosis, clinical outcome varies substantially among patients. Genomic assays have been developed and applied to predict patient prognosis for personalized treatment.

Methods: We hypothesize that the recurrence risk of ER+ breast cancer patients is determined by both genomic mutations intrinsic to tumor cells and extrinsic immunological features in the tumor microenvironment. Based on the Cancer Genome Atlas (TCGA) breast cancer data, we identified the 72 most common genomic aberrations (including gene mutations and indels) in ER+ breast cancer and defined sample-specific scores that systematically characterized the deregulated pathways intrinsic to tumor cells. To further consider tumor cell extrinsic features, we calculated immune infiltration scores for six major immune cell types.

Results: Many individual intrinsic features are predictive of patient prognosis in ER+ breast cancer, and some of them achieved comparable accuracy with the Oncotype DX assay. In addition, statistical learning models that integrated these features predicts the recurrence risk of patients with significantly better performance than the Oncotype DX assay.

Conclusions: As a proof-of-concept, our study indicates the great potential of genomic and immunological features in prognostic prediction for improving breast cancer precision medicine. The framework introduced in this work can be readily applied to other cancers.

Keywords: Genomic aberration, immune infiltration, breast cancer, prognostic prediction

Background

Breast cancer is the leading cause of cancer in women worldwide. In 2021, it is estimated that 284,200 new patients will be diagnosed and 44,130 will die from breast cancer in the USA [1]. Breast cancer patients are often grouped by estrogen receptor (ER) and progesterone receptor (PR) status. Approximately 80% of breast cancers are estrogen receptor-positive (ER+) [2], and these patients show better response to endocrine therapy and have favorable prognosis as compared to ER-negative breast cancer [3]. Regardless, there exists a wide disparity in prognosis within ER+ breast cancer patients. A subtype of ER+ breast cancer patients with an insensitivity to endocrine therapy has been reported, involving complex interactions between the human epidermal growth factor receptor-2 (HER2), ER, and other signaling pathways [2, 4]. In fact, the PAM50 gene expression test for intrinsic cancer subtyping has been shown to provide greater prognostic information than immunohistochemistry (IHC) for ER status or clinical variables [5]. For example, some ER+ patients with Luminal A breast cancer will live for over 10 years without experiencing breast cancer recurrence when treated with adjuvant tamoxifen despite exhibiting high grade, lymph node invasion, and overall higher recurrence risk, while other ER+ patients in the Basal subtype all relapse within 5 years [5]. This variance in prognosis even within the ER+ breast cancer subtype has motivated significant efforts to develop gene signatures to predict clinical outcomes and provide personalized treatment.

Indeed, a number of gene signatures have been developed to predict prognosis and stratify patients. The Oncotype DX Breast Recurrence Score has been the most widely validated gene signature in predicting prognosis for ER+ breast cancer patients, considering metrics such as its clinical utility and its impact on decision making [6]. The assay assigns a score from 0-100 (Onco-score) based on a breast cancer biopsy which classifies a patient as having a low (Onco-score: 0-17), intermediate (Onco-score: 18-30), or high (Onco-score: 31-100) risk for distal metastasis. The analysis is based on 21 genes with specific functions in tumor proliferation or invasion, HER2, and hormone receptor [7]. These 21 genes were selected based on their high association between gene expression levels and patient recurrence.

Patients classified as high risk by the Oncotype DX Assay have been shown to experience a significantly lower risk for distal recurrence when treated with chemotherapy, whereas patients in the low or intermediate risk category experience little benefit from chemotherapy [8]. Interestingly, for patients treated with tamoxifen, a form of endocrine therapy, patients classified with a low or intermediate risk experience a significantly improved distal metastasis-free survival (DMFS) rate whereas patients in the high-risk category have a smaller benefit [9]. Another gene signature, MammaPrint, has also been well-studied and has shown efficacy in predicting both ER+ and ER- breast cancer prognosis. It is based on 70 genes that were chosen using a "leave-one-out" strategy from 231 genes that were significantly associated with patient prognosis [10]. The MammaPrint signature classifies patients into good or poor signature groups based on their risk for distal metastasis. Similar to Oncotype DX, patients with high risk experience significant benefit from chemotherapy, whereas those in the low risk category do not [11]. It is evident that prognostic prediction and the discovery of risk groups via genomic assays can guide therapeutic decisions.

A number of cancer driver genes, including oncogenes and tumor suppressors, are frequently altered through somatic mutations and copy number variations (CNVs). Such genomic events are responsible for the initiation, progression and metastasis of tumors [12, 13]. For example, the *TP53* oncogene has functions in tumor suppression and DNA repair and is the most commonly mutated gene in cancer: depending on cancer type, up to 50% of cancer cases have a somatic mutation in the *TP53* gene[14]. Specifically, 25% of breast cancer patients have somatic *TP53* mutations [14]. Mutations in driver genes leads to aberrant activation (oncogenes) or inactivation (tumor suppression), which in turn deregulate downstream oncogenic pathways. However, genomic aberrations such as somatic mutations and CNVs are often only weakly associated with prognosis [15-18]. As an example, the prognostic value of the *TP53* gene mutation is inconsistent and sometimes controversial in breast cancer [19, 20]. Numerous different mechanisms can deactivate the p53 pathway besides *TP53* mutations, such as hypermethylation, CNV, or mutation of other genes in the p53 pathway [21], convoluting the impact of

TP53 mutations on oncogenic pathways. Gene signatures that recapitulate the downstream pathways of p53 mutations have been proposed as better prognostic markers [22-25]. Signatures for other genes have also been developed to predict prognosis, such as *PIK3CA* [26, 27], *BRAF* [26], *KRAS* [26], *TMPRSS-ERG* [28], etc. However, the prognostic value of signatures for all common genomic alterations have not been investigated in a systematic manner.

Genomic aberration of driver genes occurring in cancer cells represent a set of tumor-intrinsic features. In addition to genomic changes intrinsic to cancerous cells, patient prognosis is also determined by immunological features in the tumor microenvironment (TME). Tumor-infiltrating immune cells interact with the TME through a complex process known as cancer immunoediting, which involves both immunologic clearance of cancer cells and the promotion of non-immunogenic cancer clones that edits the overall immunogenicity of the tumor [29]. Somatic mutations in cancer cells can be presented as neoantigens that can be recognized by T cells to trigger an immune response. Thus, an increase of CD8+ and CD4+ T-cell infiltration in the TME is correlated with better prognosis across cancer types, including colorectal [30], lung [31], and breast cancers [32]. However, the presence of regulatory T cells serves to facilitate tumor escape and expansion, worsening prognosis [33]. Although less commonly studied, infiltrating B cells in the tumor microenvironment have also been correlated with good prognosis in melanoma [34] and multiple subtypes of breast cancers [35]. Natural Killer cells, despite functioning as tumor cell killers, have also been correlated with advanced disease and may facilitate cancer growth and poor prognosis [36, 37]. Lastly, increased macrophagic infiltration has been correlated with favorable clinical outcomes [38]. These various effects of immune infiltration on prognosis highlight the importance of such data in prognosis prediction models.

In this study, we aim to systematically investigate the prognostic value of tumor-intrinsic and -extrinsic features in ER+ breast cancer. Particularly, we define gene signatures for 72 genomic aberrations, including somatic mutations in *TP53*, *PIK3CA*, *CDH1*, and *GATA3*, which comprise tumor-intrinsic features. We also infer the immune infiltration of 6 immune cells, which comprises tumor-

extrinsic features. We then construct predictive models that integrate these genomic features (Sig model), immunological features (Imm model), and a combination of both (Sig+Imm model). We compare the predictive power of these models with Oncotype DX scores as a reference with or without incorporating established clinical variables (e.g. age, tumor stage, etc.). Our results indicate that many of our individual signatures and immune cell infiltration scores are prognostic when used as solo predictors. We also find that an optimized model that integrates both intrinsic and extrinsic features achieves a significantly higher prediction accuracy than Oncotype DX. Our study provides a generic framework to define integrative models based on genomic data to combine intrinsic and extrinsic features for predicting clinical outcomes.

Methods:

Datasets used in this study

Level 3 processed RNA sequencing (RNA-seq) data for 1097 breast cancer patients was downloaded from The Cancer Genome Atlas (TCGA) via FireHose (http://gdac.broadinstitute.org/). The data was normalized in the RNA-seq by Expectation-Maximization (RSEM) format. The processed somatic mutation data and copy number variation (CNV) segments were downloaded as Mutation Annotation Format (MAF) and segmented copy number alterations (sCNA) files, respectively, from TCGA using FireHose.

Gene expression profiles of other datasets were obtained as follows. The first dataset is from Curtis et al. (METABRIC)[39] and was downloaded from the European Genome Phenome Archive with accession ID EGAS00000000083. This dataset consists of 1992 breast cancer patients in two cohorts, the discovery cohort (997 patients), and the validation cohort (995 patients). 1508 patients were ER+. The data was measured using the Illumina HT-12 v3 platform (Illumina_Human_WG-v3). This dataset included the mutation status of the *TP53* gene, determined by examining exons 2-11 for mutations and scoring them using Mutation Surveyor software. It also included *HER2* amplification status (whether the patient was

experiencing *HER2* gain or *HER2* neutral), which was determined based on their empirical expression distributions using MCLUST. An additional seven datasets were downloaded from the Gene Expression Omnibus (GEO) under accession numbers GSE41994, GSE101780, GSE3494, GSE22093, GSE22358, GSE22597, and GSE47561. GSE41994 contains *PIK3CA* mutation status and gene expression data from 95 ER+ breast cancer patients [40]. GSE101780 contains *GATA3* mutation status and gene expression data from 13 ER+ breast cancer samples [41]. GSE3494 [23] and GSE22093 [42] contain information on 251 and 103 distinct breast cancer samples, respectively, and also contain *TP53* mutation status. GSE22358 [43] and GSE22597 [44] contain data on *HER2* amplification status of 158 and 82 breast cancer samples, respectively. The Ur-Rehman dataset, GSE47561, combines 10 other breast cancer datasets and contains a total of 1570 samples, 856 of which are ER+ [45]. The expression was measured by the Affymetrix microarray platform.

Identification of frequently mutated or amplified/deleted genes in ER+ breast cancer

MAF somatic mutation data from TCGA contained the number of nonsynonymous mutations in each gene. We considered all genes with mutations in which there was at least one nonsynonymous point mutation in at least 10 percent of all ER+ samples.

CNV data from TCGA contained genomic segments that significantly deviated from diploid (as in normal tissues) and their copy numbers in each tumor samples. Based on these data, copy numbers of genes were calculated by referring to the Ensemble human genome annotation file, which contains the genomic localization of genes. We $\log_2$ transformed these numbers to obtain the fold change of each gene's copy number. Genes that showed amplifications with a fold increase greater than $\log_2(\frac{2.8}{2})$ or deletions with a fold decrease less than $\log_2(\frac{1.3}{2})$ were selected for further investigation. A list of genes commonly exhibiting mutations and CNVs based on our approach are shown in Table 1, along with the type of genomic aberration occurring in that gene.

Definition of weighted gene signatures for genomic events

For each recurrent genomic event (gene mutations, amplifications and deletions), we defined a weighted gene signature based on TCGA ER+ breast cancer RNA-seq data. The expression levels of genes (originally represented as RSEM) were adjusted based on the formula $\log_2(\text{RSEM} + 1)$ to avoid extreme values. First, the status of each genomic aberration j in Table 1 ($X_j = 1$ if there is an aberration, 0 if no aberration) is used in a multivariate linear model (Eq. 1) as predictive variables, and the log2(RSEM+1) expression of gene i in the gene expression profile is the response variable (Y_i) (Eq. 1).

$$Y_i = \beta_{i,0} + \beta_{i,1}X_1 + \cdots + \beta_{i,j}X_j + \cdots + \beta_{i,72}X_{72} \qquad \text{Equation 1}$$

For each genomic aberration j in Table 1, the linear coefficients ($\beta_{i,j}$) and p value ($p_{i,j}$) for each gene i in the gene expression profile is calculated. Two sets of weights are then calculated for each combination of genomic aberration j and gene i in the gene expression profile, $w_{i,j}^{+}$ and $w_{i,j}^{-}$. If the expression of gene i in the gene expression profile is positively correlated with the presence of genomic event j ($\beta_{i,j} > 0$), then $w_{i,j}^{+} = -\log p_{i,j}$ and $w_{i,j}^{-} = 0$. For negative correlations ($\beta_{i,j} < 0$), then $w_{i,j}^{-} = -\log p_{i,j}$ and $w_{i,j}^{+} = 0$. The weights are then normalized by capping the maximum weight at 10 and dividing by the range to transform the weights to a decimal between zero and one. The result of this definition of $w_{i,j}^{+}$ and $w_{i,j}^{-}$ is the following: comparing patients with and without genomic aberration j, if gene m is more upregulated than gene n, then $w_{m,j}^{+} > w_{n,j}^{+}$ and $w_{m,j}^{-} = 0$. Similarly, if gene m is more downregulated than gene n, then $w_{m,j}^{-} < w_{n,j}^{-}$ and $w_{m,j}^{+} = 0$.

The above gene signatures take co-occurrences and mutual exclusiveness among different genomic features into consideration. We also defined univariate gene signatures that characterize the regulation of genes for individual genomic aberrations without considering inter-feature dependencies. A similar

procedure was applied except that in the model only a single genomic aberration was used as the independent variable (Eq. 2).

$$Y_i = \beta_{i,0} + \beta_{i,j} X_j \qquad \text{Equation 2}$$

Calculation of tumor-intrinsic genomic aberration scores

For each genomic aberration, we calculated a score that describes its pathway activity from gene expression data using the univariate and multivariate weighted gene signatures described in the previous section. To do so, we applied a rank-based statistical method called Binding Association with Sorted Expression (BASE) [46], which examines the expression of signature genes in each tumor sample to calculate a sample-specific score. Details on signature score calculation has been described previously [28, 47]. Each patient in the Curtis and Ur-Rehman datasets was given a score for each of the 78 genomic aberrations, using both the univariate and the multivariate weights. A higher score correlates with a more deficient pathway due to a greater propensity for occurrences of the genomic aberration.

The scores based on the multivariate weights are used for our Random Forest or Cox proportional hazards models because they may benefit from the inter-dependent adjustments in the multivariate gene signature. On the other hand, scores defined by the univariate weights are used to show that our scores recapitulate driver genomic aberrations by themselves, as these scores more closely correlate with the aberration status when scores for the other genomic aberrations are not considered.

Calculation of tumor-extrinsic immune infiltration scores

To calculate the six immune cell infiltration scores given the expression profile of a patient, we utilized a method described previously [48], which includes first defining immune cell-specific reference gene expression profiles and then examining immune-specific gene expression in the patient gene expression profile. Essentially, a high score indicates higher infiltration of the corresponding immune cell type in the tumor sample, while a low score indicates low infiltration level. In this study, we focused on six major

immune cell types that have previously been reported to have potential prognostic value, including naïve B (NavB), memory B (MemB), CD4+ T, CD8+ T, NK cells and monocytes.

Calculation of Oncotype DX classes and scores

We used the genefu R package to calculate the Oncotype DX risk category (Onco-class) and score for a given breast cancer gene expression dataset [49]. Specifically, the function "oncotypedx" was used. The Oncotype DX score (denoted as Onco-score in the Figs.) is a number from 0 to 100, where higher scores correlate with a higher risk for distal metastasis. The Onco-class is directly calculated from the Onco-score (0-17 is low risk, 18-30 is intermediate risk, and 31-100 is high risk).

Construction of Random Forest models to classify ER+ breast cancers into good versus poor prognostic groups

In both the Curtis and Ur-Rehman datasets, the data for prognosis was provided as time-to-event. For the Curtis dataset, the event was defined as disease-specific death. The Ur-Rehman dataset contained data on the time to recurrence and to distal metastasis. To maintain consistency between the datasets, we used "distal metastasis of disease" as the definition of an event for the Ur-Rehman dataset, since distal metastasis is a better indicator of poor prognosis and likely death due to disease.

To convert the time data to a classification problem, we defined a patient as having good prognosis when an event did not occur within 10 years of follow-up. Poor prognosis is defined as the incidence of an event within 10 years. Patients censored before 10 years are not included in the analysis. In the Curtis dataset, we counted 387 samples as having good prognosis and 293 samples as having poor prognosis. In the Ur-Rehman dataset, 168 samples were counted as having good prognosis and 177 samples were considered to have poor prognosis. We next built a Random Forest model using the R package "randomForest" to classify good vs poor prognosis with various combinations of features. We used a 10-fold cross-validation method, where the data is divided in tenths, and one tenth is used as validation while the other nine-tenths are used to train the Random Forest model. The tenths are cycled such that each

tenth serves as the validation set once, and the area under the curve (AUC) scores from all the validations are pooled to generate the aggregate AUC score.

We also trained the optimized random forest model in one dataset and used that predefined model to predict prognosis in external validation datasets. The "predict" function in the "randomForest" library was used to extract the probability of a good prognosis for each patient in the validation datasets. These probabilities are then used to generate the receiver operating characteristic (ROC) plot and calculate the AUC scores using the R package "ROCR."

Construction of Cox regression models for predicting patient survival and recurrence risk

To determine the performance of individual signatures and immune scores, we fit a univariate cox proportional hazards model for all 78 scores using the "coxph" function from the "survival" package in R and extracted the p-value and hazard ratios. To adjust for clinical variables, we included age, stage, lymph node status, grade, and size as covariates in addition to the signature or immune cell scores in a multivariate cox proportional hazards model. To plot Kaplan-Meier (KM) survival curves, we used the median score as the cutoff to dichotomize low-score and high-score patients and utilized the R function "survfit" to plot the curve. The log-rank test is applied to calculate the p-value for the null hypothesis that no difference in survival exists between the two patient groups.

We also calculated the C-index (CI) to quantify the performance of our integrative, optimized Cox proportional hazards model compared to Cox proportional hazards models with just clinical variables or Onco+Clin as predictive features. The CI is calculated as the proportion of all comparable patient pairs with predicted risks concordant with their survival time (for example, a patient pair is concordant if the patient with higher risk has a shorter survival time). Note that patients are only comparable if both patients experience the event, or if one patient experiences the event before the other patient is censored.

To calculate the recurrence risk of patients in a validation dataset with a Cox proportional hazards model fitted in the training dataset, we used the "predict" function from the "survival" package in R to extract

the predicted patient risks in the validation dataset. We then ranked the patients by their risk score and separated patients into three roughly equally sized groups to define the low, intermediate, and high risk groups. The KM method and log-rank test is again used to determine the performance of the model in defining risk groups.

Optimization of classification and regression models

To determine the optimal combination of predictive features for the RF model, we performed backward selection by successive removal of features with the lowest importance as determined using the RF relative importance function. The AUC score is calculated after each addition/removal by a 10-fold cross-validation using both Curtis discovery and validation datasets. We then picked the predictive features with the highest AUC and the least number of features as our optimized model.

To similarly optimize the Cox proportional hazards model, we used the R function "My.stepwise.coxph" in the R package "My.stepwise" on the Curtis discovery dataset, with and without clinical features. This function performs a stepwise feature selection using the coxph model. The features chosen in the optimized model are shown in Table 1 as "Cox optimized model."

Results

Overview of this study

Cancer is a genetic disease, in which mutations of oncogenes or tumor suppressors lead to the aberrant activation or inactivation of specific oncogenic pathways and eventually deregulated gene expression (see examples shown in Fig. 1a). Previous studies have shown that gene signatures modeling deregulated pathways better predict prognosis than the corresponding gene mutations. In this study, we develop a statistical framework to systematically investigate a comprehensive list of genomic events of genes frequently observed in ER+ breast cancer, including 4 mutations, 53 amplifications, and 15 deletions (Table 1). By using gene-specific multivariate regressions, we modeled the combinatorial effect of these genomic events in regulating gene expression in TCGA ER+ breast cancer samples. Based on these

models, we defined a gene signature for each of these genomic events. These gene signatures can be used to calculate sample-specific scores that indicate downstream pathways associated with the corresponding genomic events. We rationalized that these signatures will provide a systematic characterization of deregulated pathways intrinsic to tumor cells and are thus informative for predicting patient prognosis in ER+ breast cancer.

To further include tumor-extrinsic features, we considered the infiltration levels of six major types of immune cells in the tumor microenvironment, which can also be calculated based on gene expression profiles of tumor samples. We then developed prediction models to integrate these intrinsic and extrinsic features as well as established clinical factors to predict prognosis in ER+ breast cancer (Fig. 1b). Specifically, classification models based on Random Forest were constructed to classify patients with good versus poor prognosis; and Cox proportional hazards regression were constructed to predict prognostic risk of patients and support the Random Forest results. Using this framework, we can systematically characterize potential tumor intrinsic and extrinsic features fully based on transcriptomic data for prognostic prediction.

Gene signature scores recapitulate the mutation status of driver genes

Based on the TCGA data, we defined a total of 72 gene signatures to characterize all frequently occurring genomic aberrations in ER+ breast cancer. Patients with higher gene signature scores have expression profiles with a greater degree of similarity to the expression profile of patients with the genomic aberration. To investigate the interdependence between different gene signatures, we calculated pairwise Spearman correlation coefficients (SCCs). Most of these signatures were weakly correlated (Fig. 2a), indicating that they capture different downstream signaling outputs.

To examine whether these gene signatures recapitulate their genomic events, we applied them to independent ER+ breast cancer expression datasets that provided gene mutation status. As shown in Fig. 2B, the *TP53* mutation (TP53_mut) signature scores were significantly higher in *TP53* mutant samples

than wild-type samples (p=1e-12). Similarly, for another three genomic aberrations (*HER2* amplification, *PIK3CA* mutation, and *GATA3* mutation), patients with the aberration displayed significantly higher signature scores than wild-type patients (Fig. 2c-e). To quantify the accuracy of signature scores in classifying mutant versus wild-type samples, we determined the ROC curves for the TP53_mut (Fig. 2f) and ERBB2_amp (Fig. 2g) signatures in multiple ER+ breast cancer datasets (*TP53*: GSE3494 and GSE22093, *ERBB2*: GSE 22358 and GSE22597). High AUC scores are observed (AUC>0.8), indicating that our gene signature scores are able to recapitulate information on the aberration status of driver genes in validation datasets

Since these signatures were defined based on ER+ breast cancer data, we expected that they might only be effective in ER+ breast cancer. Indeed, when applied to ER- breast cancer, these signatures poorly discriminated mutant versus wild-type samples as exemplified by the *TP53* and *ERBB2* signatures (Supplementary Fig. 1).

Association of gene signatures and immune cell infiltration with patient prognosis

We then analyzed the distribution of individual signature scores based on patient prognosis, reasoning that individual signature scores must be differently distributed to be able to predict prognosis. In the Curtis data, we stratified ER+ breast cancer patients into a good (alive after 10 years) and poor (death due to disease before 10 years) prognosis group with 387 and 293 patients, respectively. For each of the 72 gene signatures, we examined the association with patient prognosis and found that 52 gene signatures were prognostic (Supplementary Table 1). For example, TP53_mut signature scores were significantly higher in patients with poor prognosis as compared to those with good prognosis (p=7e-14, Fig. 3a).This is consistent with previous reports that mutations in the *TP53* gene and disruption of the p53 pathway were correlated with poorer prognosis [50-52]. Similarly, the TNFRSF17_amp signature scores were significantly different between groups with higher values in patients with good prognosis (p = 2e-13, Fig.

3b). Such an association has not been previously reported in ER+ breast cancer, but its prognostic value has been studied for ER- breast cancer [53-55]. *TNFRSF17* is an immune response gene with overexpression correlating with better prognosis [53].

In additional to these gene signatures, we also investigated the prognostic value of infiltration levels of six types of immune cells, including Naïve B, Memory B, CD4+ T, CD8+ T, NK, and Monocytes (see Supplementary Table 2). Based on the inferred infiltration scores, we observed a correlation between MemB cell infiltration and patient prognosis in ER+ breast cancer: patients with good prognosis had significantly higher MemB cell infiltration scores than those with poor prognosis ($p = 8e\text{-}9$, Fig. 3c).

For all gene signatures and infiltrating immune cell types, we calculated AUC scores to quantify the ability to classify good versus poor prognosis groups in ER+ breast cancer (Supplementary Tables 1 and 2). Specifically, the AUC scores for TP53_mut, TNFRSF17_amp and MemB cells were 0.665, 0.663 and 0.626, respectively, which was comparable to the accuracy achieved by the widely used Onco-score (AUC = 0.668) (Fig. 3d). All together, our results indicated that both intrinsic gene signatures and extrinsic infiltrating immune cells were informative for predicting patient prognosis in ER+ breast cancer.

Integrative models for classifying good and poor prognosis patient groups

After showing the clinical significance and prognostic power of our individual signature scores, we constructed RF models with different combinations of scores to classify patient prognosis. We then determined their performance by 10-fold cross-validation in the Curtis data (Fig. 3e). First, we constructed three RF models that integrated the 72 gene signatures (Sig), the six immune cell types (Imm) and a combination of both features (Sig+Imm). The Sig model achieved an accuracy of AUC=0.707, which was higher than the accuracy achieved by the Onco-score (AUC=0.668). In contrast, the Imm model had a relatively low AUC (AUC=0.595). In addition, the Sig+Imm model had similar accuracy as the Sig model, suggesting that adding immunological features does not further increase the performance

of the Sig model. In fact, breast cancer has been reported as immune cold and the overall response rate of ER+ breast cancer patients to immunotherapy is about 12% [56, 57]. Second, we further incorporated several clinical factors (age, tumor size, grade, stage and lymph node status) to construct the Sig+Clin, Imm+Clin, Sig+Imm+Clin models. The AUC scores of these models (0.811, 0.808 and 0.817, respectively) were higher than the accuracy of the Clin model (AUC=0.764), which was solely based on clinical factors. Moreover, the accuracies of these models were also higher than the Onco+Clin model (AUC=0.792), suggesting that both genomic features and immunological features provide additional prognostic value that surpass the predictive ability of the Oncotype DX assay. Of note, incorporating the Onco-score (the Sig+Imm+Onco+Clin model) did not further improve the classification accuracy, suggesting that information provided by the Onco-score is captured by the genomic and/or immunological features. Thus, our gene signatures and immune infiltration scores may encompass information contained in the Oncotype DX assay and provide additional prognostic prediction potential.

To identify features that contributed most to the prediction ability, we examined the relative importance of all features included in the Sig+Imm+Clin RF model. As shown in Fig. 3f, the top 10 most important Sig or Imm features are CCND1_amp, ABL2_amp, COX6C_amp, TNFRSF17_amp, MYC_amp, MemB, NTRK1_amp, TP53_mut, CLTC_amp, and ELK4_amp. A complete list of the relative importance for all features can be found in Supplementary Table 3. Out of the most predictive features, many have been studied previously as prognostic factors in breast cancer, such as CCND1_amp [58], ABL2_amp [59], TP53_mut [60], TNFRSF17_amp [53-55], and memory B infiltration [61]. However, there are other important gene aberrations like COX6C_amp that have not been previously reported as prognostic. Our models thus shed light on the roles of various unexplored mutations/CNVs in breast cancer prognosis.

An optimized model outperforms Oncotype DX scores for prognosis classification

We optimized the Sig+Imm+Clin model by iteratively removing the least important features from the model and then recalculating the relative importance of the remaining features (Fig. 4a). Eventually, we obtained an optimized model with a total of 22 (including 18 Sig, 1 Imm, and 3 Clin features) predictive features (Table 1). The accuracy of this optimized model was AUC=0.841 according to cross-validation results in the Curtis data, which was much higher the Onco+Clin model (AUC=0.791) and the Clin model (AUC=0.763) (Fig. 4b). In addition, when trained in the Curtis discovery dataset and evaluated in the Curtis validation data, the optimized model achieved an AUC of 0.837. Conversely, an AUC 0f 0.798 was obtained when the training and test datasets were swapped (Fig. 4c).

To demonstrate the model's ability to perform in clinical situations, the optimized model was trained in a training dataset and then applied to an independent test dataset for predicting prognosis. When the optimized model was trained in the Curtis data and tested in the Ur-Rehman data, we observed an accuracy of AUC=0.708, and vice versa, an accuracy of AUC=0.737 (Fig. 4d). This performance was significantly higher than that of the Onco+Clin model, which achieved AUC=0.594 and AUC=0.671 for Curtis-to-Rehman and Rehman-to-Curtis predictions, respectively (Fig. 4e). Altogether, our results indicated that the optimized model achieved consistently higher performance than the Oncotype DX assay.

Integrative models for predicting prognostic risk based on Cox regression

We constructed Cox proportional hazards models to further validate the prognostic values of the gene signatures and infiltrating immune cells. We applied univariate and multivariate Cox regression models to investigate the association between these individual features and patient disease-free survival with or without including clinical variables. Using the gene signature scores or the immune infiltration scores as continuous variables, we found 50 features (46 gene signatures and 4 immune cells) that were significantly associated with patient survival without considering clinical variables (Fig 5a and Supplementary Table 4). After adjusting for clinical variables, 36 features (34 gene signatures and 2 immune cells) were significantly associated with patient survival (Fig 5b and Supplementary Table 4).

These results indicated that many of gene signatures alone can stratify ER+ breast cancer patients into subgroups with different prognosis. For example, patients with high TP53_mut score showed significantly shorter survival time than those with low score (p=2e-16, Fig 5c). In contrast, patients with high TNFRSF17_amp (Fig. 5d) or MemB infiltration scores (Fig. 5e) had significantly prolonged survival. For comparison, we showed the survival curves of patients dichotomized based on the Onco-score (p=5e-12, Fig. 5f), which exhibited a lesser or comparable significance as compared to our signatures. Our results indicated that many gene signatures could give rise to better or comparable prognostic stratification than Onco-score in ER+ breast cancer.

Following this, we integrated all 78 signature scores and immune infiltration scores using multivariate Cox regression models and performed feature selection to obtain an optimized model with 16 variables. The selected features are listed in Table 1. We applied this optimized model to predict patient prognostic risk, which achieved a fairly high performance with a CI of 0.758 based on cross-validation in the Curtis data. Of note, a similar model based on clinical factors achieved a CI of 0.701, and the CI increased to 0.718 if Oncotype scores are further incorporated with clinical variables.

An optimized Cox regression model for prognostic risk prediction

To demonstrate the clinical utility of the Cox-optimized model, we trained the model in the Curtis discovery dataset and then applied it to predict patient risk score in the Curtis validation dataset. Based on the predicted risk scores, we stratified patients into high-, intermediate- and low-risk groups of equal size. As shown in Fig. 6a, the optimized model was overall able to separate patients into different risk groups (p = 1e-23). In particular, patients in the high risk category had a hazard 3.196 times that of the intermediate risk category (p = 3e-11), and patients in the intermediate risk category had a hazard 2.523 times that of the low risk category (p = 0.001). We also considered the predictive power of model after adjusting for the contributions of the clinical features by removing clinical features from the model (Fig. 6b). The overall significance of the risk groups decreased moderately (p=1e-09). Patients in the high risk group had a hazard 2.138 times that of the intermediate risk category (p = 3e-5), and patients in the

intermediate risk category had a hazard 1.616 times that of the low risk category (p = 0.05). The performance of our models in predicting prognostic risk were then compared to that of the three Onco-classes for patients in the Curtis validation dataset (Fig. 6c). Overall, the Onco-classes slightly underperformed the optimized model adjusted for clinical features (p = 5e-8). Patients classified as high risk had a hazard 1.972 times that of patients in the intermediate risk category (p = 3e-4), and those classified as intermediate risk had a hazard 1.784 times that of the low risk category (p = 0.04). Although the ability to distinguish low risk from intermediate risk patients was similar between Onco-class and our clinical variable-adjusted optimized model, our model was able to more significantly define a higher risk group from the intermediate risk group (p=3e-5 for our model vs p=3e-4 for Onco-class).

To investigate whether the optimized model has the potential to improve the Oncotype DX assay, we used our optimized model's risk predictions to further dichotomize patients in each Oncotype DX class. In each of the Onco high (Fig. 6d), Onco intermediate (Fig. 6e), and Onco low (Fig. 6f) classes, our optimized model's risk prediction was able to separate patients into two statistically significant risk groups (Onco high: p = 3e-9, Onco int: p = 0.001, and Onco low: p = 0.05). To identify the prognostic power of our signature and immune scores alone, we repeated this test with the clinical variables removed (Supplementary Fig. 2). We found that the Onco high and Onco intermediate risk classes were significantly separated (p = 0.009 and p = 0.004, respectively) and the Onco low risk class was moderately separated (p = 0.2). These results show the potential of our models to be integrated in conjunction with Oncotype DX to provide more precise risk predictions.

Discussion

In this study, we defined gene signatures for a comprehensive list of gene aberrations frequently observed in ER+ breast cancer to recapitulate their downstream regulatory pathways. These signatures were then used to calculate sample-specific scores based on tumor gene expression profiles. These scores represented a collection of tumor-intrinsic features that project gene expression to cancer-related pathway activities. The prognostic values of these signatures were validated by both Random Forest classification

and Cox regression models. The performance of these prediction models that integrated gene signatures and immune cell infiltration scores were evaluated with or without including clinical variables. Our results indicated that these features have the great potential to further improve the prediction accuracy achieved by the Oncotype DX assay.

A number of genomic aberrations, such as *TP53* mutation and *ERBB2* amplification, have been frequently observed in breast cancer [62]. These genomic mutations lead to the deregulation of specific downstream pathways and confer selection advantage to tumor cells at certain stage of cancer development and progression. Essentially, it is the downstream pathways that drive tumorigenesis and determine clinical outcomes. The frequently observed gene mutations represent the most likely but not the only mechanism that deregulate the corresponding pathways. For example, it has been shown the p53 pathway can be inactivated not only by *TP53* mutation but also by alternative mechanisms like hypermethylation or mutation of other genes in the p53 pathway [21]. Therefore, the gene signatures defined by the proposed method provide a collection of candidate features that are prognostic. Importantly, a tumor sample may harbor multiple driver genomic mutations. Some of the driver genomic mutations are correlated, presenting together or mutually exclusive. As such, we defined gene signatures for 72 frequent gene mutation events in a systematic manner, where all signatures are calculated together in a multivariate linear model. The resulting gene signatures takes into the correlations between different genomic events and the cross-talking in their downstream pathways, and therefore are expected to better predict prognosis when combined using integrative models. We have also defined gene signatures for the same set of genomic aberrations but in a separate fashion, where the gene signatures are defined based on only their value in a univariate linear model (see Method). Such a definition most directly correlates with the aberration status of the driver gene, but the lack of consideration for the interdependence between signatures makes them unfit for multivariate prediction models. Indeed, integrative prediction models based on these signatures resulted in worse prediction accuracy compared with those simultaneously defined signatures.

In addition to the gene signatures designed for charactering tumor-intrinsic features, we also include immune infiltration scores to capture tumor-extrinsic features. However, our results indicated that the prognostic values contributed by these immunological features, with the exception of MemB, were relatively low compared with gene signatures and clinical factors. This might be explained by the fact that breast cancer is relatively immune cold compared to other cancers. Immune cytolytic activity, mutation burden, and neoepitopes load are often correlated with immune response, yet breast cancer patients generally have relatively moderate levels of these factors [17, 63]. Indeed, the overall response rate of ER+, HER2 negative breast cancer to pembrolizumab, an antibody targeting programmed cell death-1/programmed death ligand-1, is only about 12% [56]. Specific infiltrating immune cell subsets associated with patient prognosis are also low in ER positive breast cancer. Such immune cells include CD8+ T and CD4+ T, which have been extensively reported as good prognostic markers [64, 65], whereas regulatory T cells are known to be poor prognostic markers [66]. While triple negative breast cancer most commonly exhibit high infiltrations of these tumor-infiltrating lymphocytes (TILs), ER+ patients generally have lower TIL infiltration levels [67], making infiltration levels difficult to be accurately inferred by immune deconvolution methods from gene expression data. In fact, TIL levels have been found to not correlate with overall survival rate for ER+ breast cancer patients [68]. Future work may investigate the prognostic potential of these tumor extrinsic immune cell infiltration scores in prognostic prediction models for immune hot cancers like melanoma, lung cancer, and acute lymphoblastic leukemia.

Our analysis showed that the clinical factors alone can achieve relative high accuracy in prognostic prediction. The prognosis of ER+ breast cancer is largely determined by proliferation rate of tumors [69], which can be captured at least partially by the clinical variables, tumor size and stage. Nevertheless, the prognostic prediction accuracy can be further improved when the gene signatures and immune infiltration scores are used (the Sig+Clin and the Imm+Clin models). In particular, the Imm+Clin model showed an improved prediction accuracy than the clinical model, even though immunological features alone had

fairly poor performance. Overall, these results indicate that additional biomarkers developed from genomic, molecular or immunological characterization of tumors can further improve prognostic prediction in ER+ breast cancer.

Conclusions

In conclusion, we have proposed a framework to systematically extract both tumor-intrinsic and extrinsic features from gene expression data for integrative prediction of prognosis in ER+ breast cancer. Using this framework, we assessed the prognostic values contributed by different categories of features as well as by different genomic aberration events. This framework can be readily applied to all cancer types for improving precision medicine.

List of Abbreviations

Abbreviations:

AUC: area under the curve

CI: C-index

CNV: copy number variation

DMFS: distal metastasis-free survival

ER: estrogen receptor

HER2: human epidermal growth factor receptor-2

KM: Kaplan-Meier

MAF: Mutation Annotation Format

PR: progesterone receptor

SCC: Spearman correlation coefficient

sCNA: segmented copy number alterations

TCGA: The Cancer Genome Atlas

TME: tumor microenvironment

TIL: tumor-infiltrating lymphocyte

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All datasets included in this study are publicly available at The Cancer Genome Atlas (TCGA) via FireHose (http://gdac.broadinstitute.org/), the European Genome Phenome Archive with accession ID EGAS00000000083, and the Gene Expression Omnibus (GEO) under accession numbers GSE41994, GSE101780, GSE3494, GSE22093, GSE22358, GSE22597, and GSE47561.

Competing interests

The authors declare that they have no competing interests

Funding

This work is supported by the Cancer Prevention Research Institute of Texas (CPRIT) (RR180061 to CC) and the National Cancer Institute of the National Institutes of Health (1R21CA227996 to CC). CC is a CPRIT Scholar in Cancer Research.

Authors' contributions

CC is responsible for the study concept and design. CC and KY are responsible for the acquisition and verification of underlying data. CC and KY are responsible for the analysis and interpretation of data. CC, KY, ES, and BZ are responsible for the drafting and revising of the manuscript. CC is responsible for the study supervision. All authors have approved the final version of the manuscript.

Acknowledgements

Not applicable.

References

[1] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. CA: A Cancer Journal for Clinicians. 2021;71(1):7-33.
[2] Lumachi F, Santeufemia DA, Basso SM. Current medical treatment of estrogen receptor-positive breast cancer. World journal of biological chemistry. 2015;6(3):231.
[3] Henry N, Shah P, Haider I, Freer P, Jagsi R, Sabel M. Chapter 88: cancer of the breast. Abeloff's clinical oncology, 6th edn Elsevier, Philadelphia, PA. 2020.
[4] Jerusalem G, Bachelot T, Barrios C, Neven P, Di Leo A, Janni W, et al. A new era of improving progression-free survival with dual blockade in postmenopausal HR+, HER2− advanced breast cancer. Cancer treatment reviews. 2015;41(2):94-104.
[5] Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor–positive breast cancer. Clinical cancer research. 2010;16(21):5222-32.
[6] Markopoulos C, van de Velde C, Zarca D, Ozmen V, Masetti R. Clinical evidence supporting genomic tests in early breast cancer: Do all genomic tests provide the same information? European Journal of Surgical Oncology (EJSO). 2017;43(5):909-20.
[7] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New England Journal of Medicine. 2004;351(27):2817-26.
[8] Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol. 2006;24(23):3726-34.
[9] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. Expression of the 21 genes in the Recurrence Score assay and tamoxifen clinical benefit in the NSABP study B-14 of node negative, estrogen receptor positive breast cancer. Journal of Clinical Oncology. 2005;23(16_suppl):510-.
[10] Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. nature. 2002;415(6871):530-6.
[11] Knauer M, Mook S, Rutgers EJ, Bender RA, Hauptmann M, Van de Vijver MJ, et al. The predictive value of the 70-gene signature for adjuvant chemotherapy in early breast cancer. Breast cancer research and treatment. 2010;120(3):655-61.
[12] Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010;463(7278):191-6.
[13] Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007;446(7132):153-8.
[14] Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harbor perspectives in biology. 2010;2(1):a001008.
[15] Lipsyc M, Yaeger R. Impact of somatic mutations on patterns of metastasis in colorectal cancer. Journal of gastrointestinal oncology. 2015;6(6):645.
[16] Zhang S, Xu Y, Hui X, Yang F, Hu Y, Shao J, et al. Improvement in prediction of prostate cancer prognosis with somatic mutational signatures. Journal of Cancer. 2017;8(16):3261.
[17] Haricharan S, Bainbridge MN, Scheet P, Brown PH. Somatic mutation load of estrogen receptor-positive breast tumors predicts overall survival: an analysis of genome sequence data. Breast cancer research and treatment. 2014;146(1):211-20.
[18] Griffith OL, Spies NC, Anurag M, Griffith M, Luo J, Tu D, et al. The prognostic effects of somatic mutations in ER-positive breast cancer. Nature communications. 2018;9(1):1-16.

[19] Archer S, Eliopoulos A, Spandidos D, Barnes D, Ellis I, Blamey R, et al. Expression of ras p21, p53 and c-erb B-2 in advanced breast cancer and response to first line hormonal therapy. British journal of cancer. 1995;72(5):1259-66.
[20] Rozan S, Vincent‐Salomon A, Zafrani B, Validire P, De Cremoux P, Bernoux A, et al. No significant predictive value of c‐erbB‐2 or p53 expression regarding sensitivity to primary chemotherapy or radiotherapy in breast cancer. International Journal of Cancer. 1998;79(1):27-33.
[21] Joerger AC, Fersht AR. The p53 Pathway: Origins, Inactivation in Cancer, and Emerging Therapeutic Approaches. Annu Rev Biochem. 2016;85:375-404.
[22] Zhao Y, Varn FS, Cai G, Xiao F, Amos CI, Cheng C. A P53-deficiency gene signature predicts recurrence risk of patients with early-stage lung adenocarcinoma. Cancer Epidemiology and Prevention Biomarkers. 2018;27(1):86-95.
[23] Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proceedings of the National Academy of Sciences. 2005;102(38):13550-5.
[24] Xu F, Lin H, He P, He L, Chen J, Lin L, et al. A TP53-associated gene signature for prediction of prognosis and therapeutic responses in lung squamous cell carcinoma. Oncoimmunology. 2020;9(1):1731943.
[25] Coutant C, Rouzier R, Qi Y, Lehmann-Che J, Bianchini G, Iwamoto T, et al. Distinct p53 gene signatures are needed to predict prognosis and response to chemotherapy in ER-positive and ER-negative breast cancers. Clinical Cancer Research. 2011;17(8):2591-601.
[26] Tian S, Simon I, Moreno V, Roepman P, Tabernero J, Snel M, et al. A combined oncogenic pathway signature of BRAF, KRAS and PI3KCA mutation improves colorectal cancer classification and cetuximab treatment prediction. Gut. 2013;62(4):540-9.
[27] Loi S, Michiels S, Baselga J, Bartlett JM, Singhal SK, Sabine VS, et al. PIK3CA genotype and a PIK3CA mutation-related gene signature and response to everolimus and letrozole in estrogen receptor positive breast cancer. PloS one. 2013;8(1):e53292.
[28] Zhou E, Zhang B, Zhu K, Schaafsma E, Kumar RD, Cheng C. A TMPRSS2-ERG gene signature predicts prognosis of patients with prostate adenocarcinoma. Clin Transl Med. 2020;10(8):e216.
[29] Schreiber RD, Old LJ, Smyth MJ. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. Science. 2011;331(6024):1565-70.
[30] Dahlin AM, Henriksson ML, Van Guelpen B, Stenling R, Öberg Å, Rutegård J, et al. Colorectal cancer prognosis depends on T-cell infiltration and molecular characteristics of the tumor. Modern Pathology. 2011;24(5):671-82.
[31] Hiraoka K, Miyamoto M, Cho Y, Suzuoki M, Oshikiri T, Nakakubo Y, et al. Concurrent infiltration by CD8+ T cells and CD4+ T cells is a favourable prognostic factor in non-small-cell lung carcinoma. British journal of cancer. 2006;94(2):275-80.
[32] Ali H, Provenzano E, Dawson S-J, Blows F, Liu B, Shah M, et al. Association between CD8+ T-cell infiltration and breast cancer survival in 12 439 patients. Annals of oncology. 2014;25(8):1536-43.
[33] Marshall EA, Ng KW, Kung SH, Conway EM, Martinez VD, Halvorsen EC, et al. Emerging roles of T helper 17 and regulatory T cells in lung cancer progression and metastasis. Molecular cancer. 2016;15(1):1-15.
[34] Ladányi A, Kiss J, Mohos A, Somlai B, Liszkay G, Gilde K, et al. Prognostic impact of B-cell density in cutaneous melanoma. Cancer Immunology, Immunotherapy. 2011;60(12):1729-38.
[35] Iglesia MD, Vincent BG, Parker JS, Hoadley KA, Carey LA, Perou CM, et al. Prognostic B-cell signatures using mRNA-seq in patients with subtype-specific breast and ovarian cancer. Clinical Cancer Research. 2014;20(14):3818-29.

[36] Vgenopoulou S, Lazaris AC, Markopoulos C, Boltetsou E, Kyriakou V, Kavantzas N, et al. Immunohistochemical evaluation of immune response in invasive ductal breast cancer of not-otherwise-specified type. The Breast. 2003;12(3):172-8.
[37] Triki H, Charfi S, Bouzidi L, Kridis WB, Daoud J, Chaabane K, et al. CD155 expression in human breast cancer: clinical significance and relevance to natural killer cell infiltration. Life sciences. 2019;231:116543.
[38] Leek RD, Lewis CE, Whitehouse R, Greenall M, Clarke J, Harris AL. Association of macrophage infiltration with angiogenesis and prognosis in invasive breast carcinoma. Cancer research. 1996;56(20):4625-9.
[39] Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346-52.
[40] Jansen MP, Knijnenburg T, Reijm EA, Simon I, Kerkhoven R, Droog M, et al. Hallmarks of aromatase inhibitor drug resistance revealed by epigenetic profiling in breast cancer. Cancer research. 2013;73(22):6632-41.
[41] Gustin JP, Miller J, Farag M, Rosen DM, Thomas M, Scharpf RB, et al. GATA3 frameshift mutation promotes tumor growth in human luminal breast cancer cells and induces transcriptional changes seen in primary GATA3 mutant breast cancers. Oncotarget. 2017;8(61):103415.
[42] Iwamoto T, Bianchini G, Booser D, Qi Y, Coutant C, Ya-Hui Shiang C, et al. Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer. Journal of the National Cancer Institute. 2011;103(3):264-72.
[43] Glück S, Ross JS, Royce M, McKenna EF, Perou CM, Avisar E, et al. TP53 genomics predict higher clinical and pathologic tumor response in operable early-stage breast cancer treated with docetaxel-capecitabine±trastuzumab. Breast cancer research and treatment. 2012;132(3):781-91.
[44] Iwamoto T, Bianchini G, Qi Y, Cristofanilli M, Lucci A, Woodward WA, et al. Different gene expressions are associated with the different molecular subtypes of inflammatory breast cancer. Breast cancer research and treatment. 2011;125(3):785-95.
[45] Ur-Rehman S, Gao Q, Mitsopoulos C, Zvelebil M. ROCK: a resource for integrative breast cancer data analysis. Breast cancer research and treatment. 2013;139(3):907-21.
[46] Cheng C, Yan X, Sun F, Li LM. Inferring activity changes of transcription factors by binding association with sorted expression profiles. BMC Bioinformatics. 2007;8:452.
[47] Zhao Y, Varn FS, Cai G, Xiao F, Amos CI, Cheng C. A P53-Deficiency Gene Signature Predicts Recurrence Risk of Patients with Early-Stage Lung Adenocarcinoma. Cancer Epidemiol Biomarkers Prev. 2018;27(1):86-95.
[48] Varn FS, Wang Y, Mullins DW, Fiering S, Cheng C. Systematic pan-cancer analysis reveals immune cell interactions in the tumor microenvironment. Cancer research. 2017;77(6):1271-82.
[49] Gendoo DM, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics. 2016;32(7):1097-9.
[50] Done SJ, Eskandarian S, Bull S, Redston M, Andrulis IL. p53 missense mutations in microdissected high-grade ductal carcinoma in situ of the breast. Journal of the National Cancer Institute. 2001;93(9):700-4.
[51] Gasco M, Shami S, Crook T. The p53 pathway in breast cancer. Breast cancer research. 2002;4(2):1-7.
[52] Pharoah P, Day N, Caldas C. Somatic mutations in the p53 gene and prognosis in breast cancer: a meta-analysis. British journal of cancer. 1999;80(12):1968-73.
[53] Criscitiello C, Azim Jr H, Schouten P, Linn S, Sotiriou C. Understanding the biology of triple-negative breast cancer. Annals of oncology. 2012;23:vi13-vi8.

[54] Sninsky JJ, Christopherson C, Lagier R, Chang M, Kwok S, Tandon V, et al. Multiplex TaqMan assays for a 7-gene prognostic immune response score to differentiate risk among women with ER-negative breast cancer. American Society of Clinical Oncology; 2012.
[55] Yau C, Sninsky J, Kwok S, Wang A, Degnim A, Ingle JN, et al. An optimized five-gene multi-platform predictor of hormone receptor negative and triple negative breast cancer metastatic risk. Breast Cancer Research. 2013;15(5):1-14.
[56] Rugo HS, Delord J-P, Im S-A, Ott PA, Piha-Paul SA, Bedard PL, et al. Safety and antitumor activity of pembrolizumab in patients with estrogen receptor–positive/human epidermal growth factor receptor 2–negative advanced breast cancer. Clinical Cancer Research. 2018;24(12):2804-11.
[57] Vonderheide RH, Domchek SM, Clark AS. Immunotherapy for breast cancer: what are we missing? : AACR; 2017.
[58] Roy PG, Pratt N, Purdie CA, Baker L, Ashfield A, Quinlan P, et al. High CCND1 amplification identifies a group of poor prognosis women with estrogen receptor positive breast cancer. International journal of cancer. 2010;127(2):355-60.
[59] Meirson T, Genna A, Lukic N, Makhnii T, Alter J, Sharma VP, et al. Targeting invadopodia-mediated breast cancer metastasis by using ABL kinase inhibitors. Oncotarget. 2018;9(31):22158.
[60] Olivier M, Langer A, Carrieri P, Bergh J, Klaar S, Eyfjord J, et al. The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer. Clinical cancer research. 2006;12(4):1157-67.
[61] Ali HR, Chlon L, Pharoah PD, Markowetz F, Caldas C. Patterns of immune infiltration in breast cancer and their clinical implications: a gene-expression-based retrospective study. PLoS medicine. 2016;13(12):e1002194.
[62] Zhang G, Wang Y, Chen B, Guo L, Cao L, Ren C, et al. Characterization of frequently mutated cancer genes in Chinese breast tumors: a comparison of Chinese and TCGA cohorts. Annals of translational medicine. 2019;7(8).
[63] Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015;160(1-2):48-61.
[64] Savas P, Salgado R, Denkert C, Sotiriou C, Darcy PK, Smyth MJ, et al. Clinical relevance of host immunity in breast cancer: from TILs to the clinic. Nature reviews Clinical oncology. 2016;13(4):228.
[65] Gu-Trantien C, Loi S, Garaud S, Equeter C, Libin M, De Wind A, et al. CD4+ follicular helper T cell infiltration predicts breast cancer survival. The Journal of clinical investigation. 2013;123(7):2873-92.
[66] Wang Y, Sun J, Zheng R, Shao Q, Gao W, Song B, et al. Regulatory T cells are an important prognostic factor in breast cancer: a systematic review and meta-analysis. Neoplasma. 2016;63(5):789-98.
[67] Stanton SE, Disis ML. Clinical significance of tumor-infiltrating lymphocytes in breast cancer. Journal for immunotherapy of cancer. 2016;4(1):1-7.
[68] Krishnamurti U, Wetherilt CS, Yang J, Peng L, Li X. Tumor-infiltrating lymphocytes are significantly associated with better overall survival and disease-free survival in triple-negative but not estrogen receptor–positive breast cancers. Human pathology. 2017;64:7-12.
[69] Beresford MJ, Wilson GD, Makris A. Measuring proliferation in breast cancer: practicalities and applications. Breast Cancer Research. 2006;8(6):1-11.

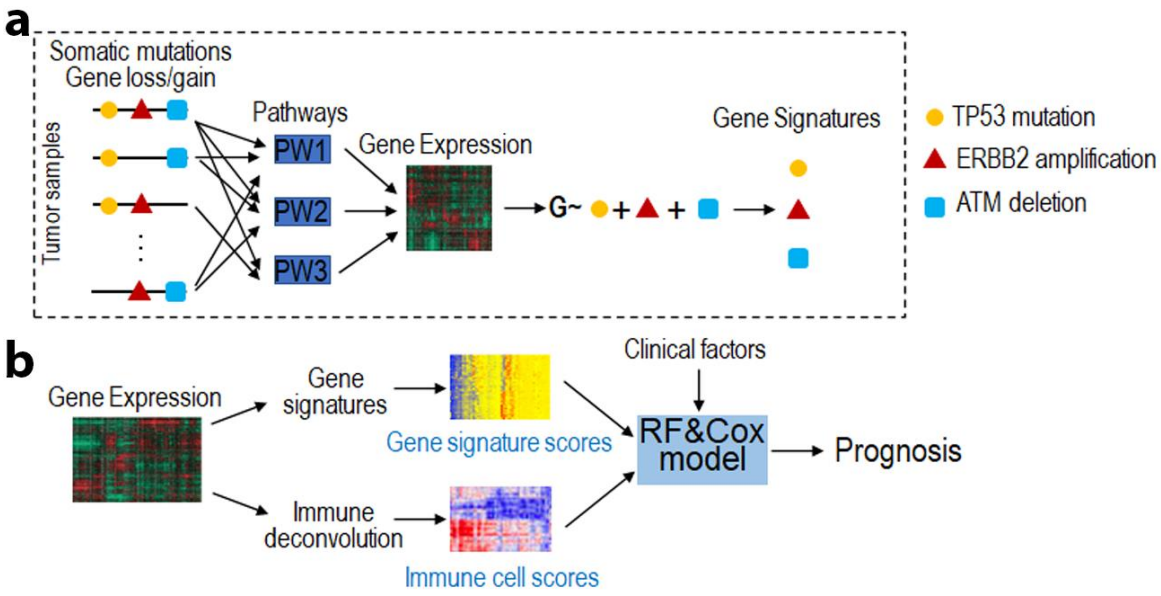

Fig. 1: Schematic diagram of this study. **a** Definition of gene signatures to recapitulate the pathways underlying driver genomic aberrations. Here we use three genomic aberrations (*TP53* mutation, *ERBB2* amplification and *ATM* deletion) as examples. In ER+ breast cancer, we defined a total of 72 gene signatures, each for a specific genomic aberration. **b** To predict patient prognosis in ER+ breast cancer, we constructed prediction models to integrate the 72 gene signatures (intrinsic features), 6 types of infiltrating immune cells (extrinsic features), and clinical factors (e.g., age, tumor stage). Gene signature scores and immune cell scores were calculated based on gene expression of tumor samples. Random Forest models were used to classify good versus poor prognosis, and Cox regression models were used to predict prognostic risk scores.

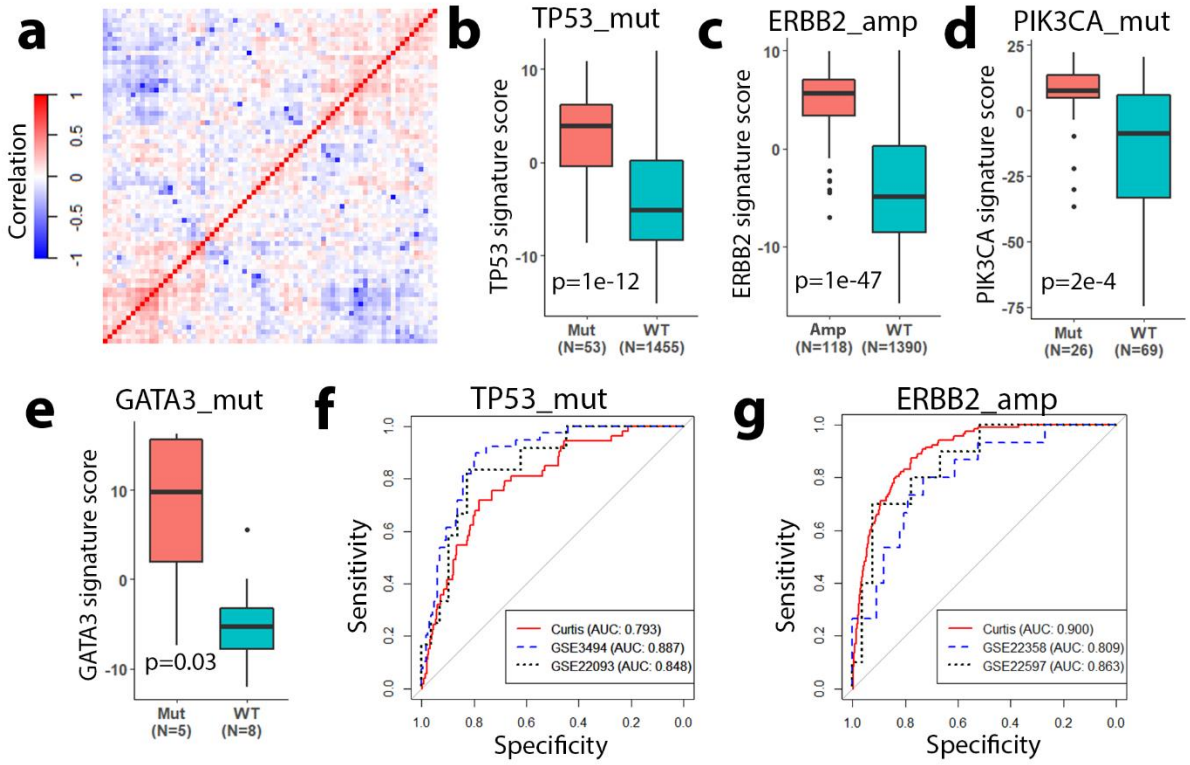

Fig. 2. Gene signatures recapitulate the downstream pathways of mutated driver genes. **a** Spearman correlation coefficients (Correlation) between different gene signatures defined based on the TCGA ER+ breast cancer data. Signature scores can distinguish ER+ breast cancer samples with TP53_mut (**b**), ERBB2_amp (**c**), PIK3CA_mut (**d**), and GATA3_mut (**e**) from samples without the aberrations. **b, c** were based on the Curtis data; (**d**) was based on GSE41994; and (**e**) was based on GSE101780. ROC curves showing that TP53_mut (**f**) and ERBB2_amp (**g**) signature scores can predict the mutation status of their respective driver genomic aberration.

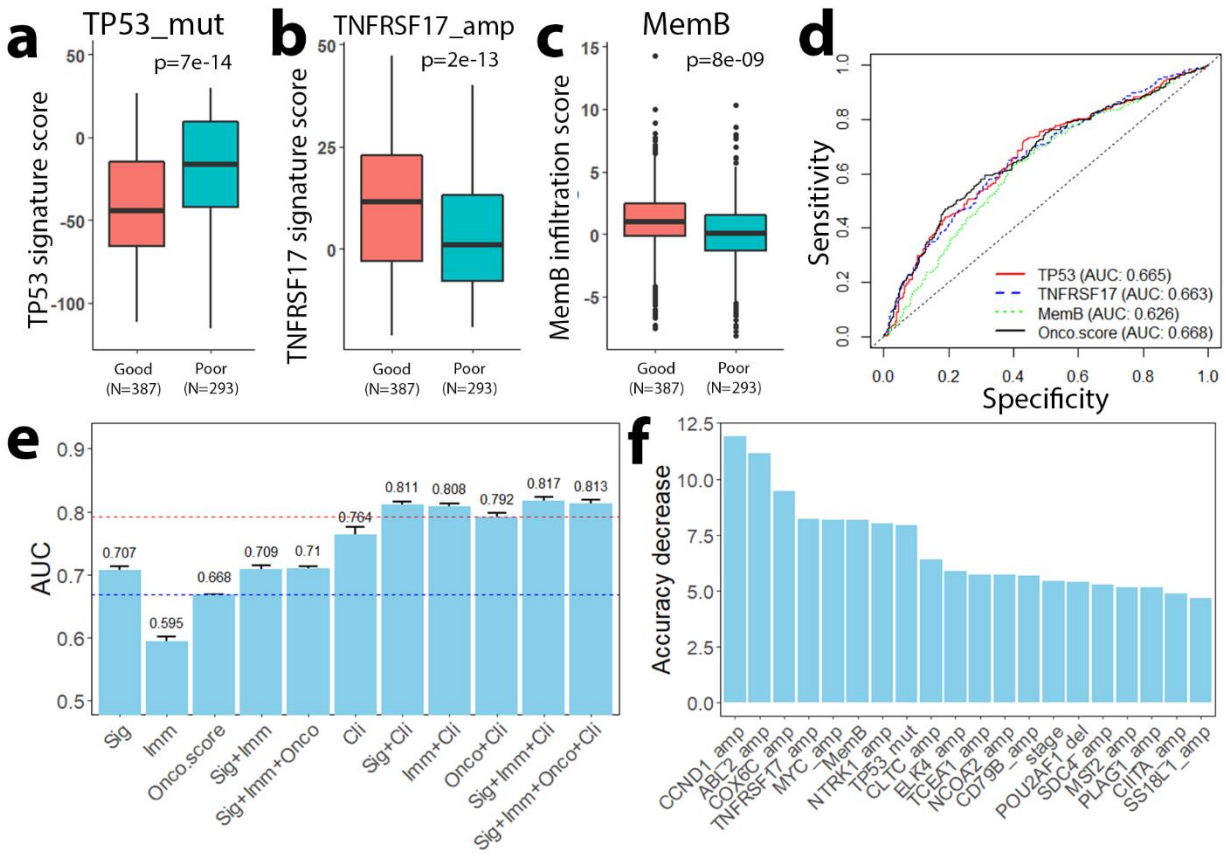

Fig. 3. Gene signatures and immune infiltration scores predict patient prognosis. a-c Signature scores for TP53_mut (**a**), TNFRSF17_amp (**b**), and MemB (**c**) distinguishes patients with good and poor prognosis. **d** ROC curves showing that TP53_mut, TNFRSF17_amp and MemB infiltration score predicts prognosis at a comparable level to Onco-score. **e** AUC scores of random forest models with different combinations of predictive features. Our Sig+Imm model performs with higher AUC scores than the Onco-score models with and without clinical features. **f** Relative importance of the top 20 most important genomic aberration and immune infiltration features.

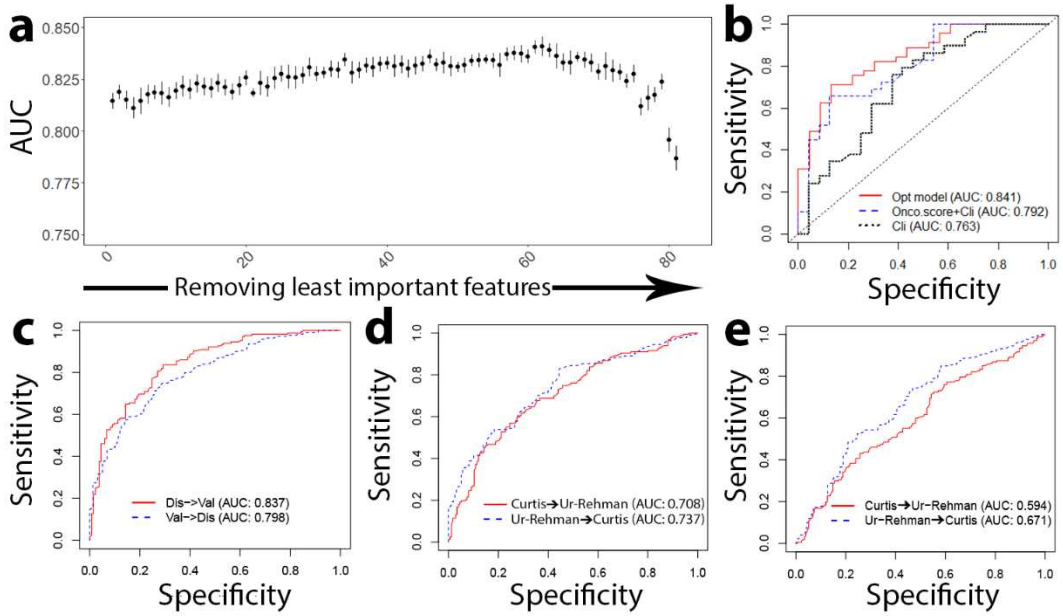

Fig. 4. Optimized model outperforms Oncotype DX risk scores for prognostic prediction. a Results from backward selection to find an optimized set of features—AUC score of the model plotted as a function of the number of features removed. The optimized model is chosen as the highest AUC score at the smallest number of features. **b** ROC curves of the performance of our optimized model as compared to the Onco-score + Cli model and just the Cli model, showing that our optimized model overperforms both Oncotype DX and clinical features. **c** ROC curves of our optimized model when trained in Curtis discovery and validated in Curtis validation, and vise versa. **d** ROC curves of our optimized model when trained in Curtis discovery and validated in the test dataset, the Ur-Rehman dataset, and vise versa. **e** ROC curves of the Onco+Cli model trained and validated in the same way as (**d**), showing decreased performance compared to our optimized model.

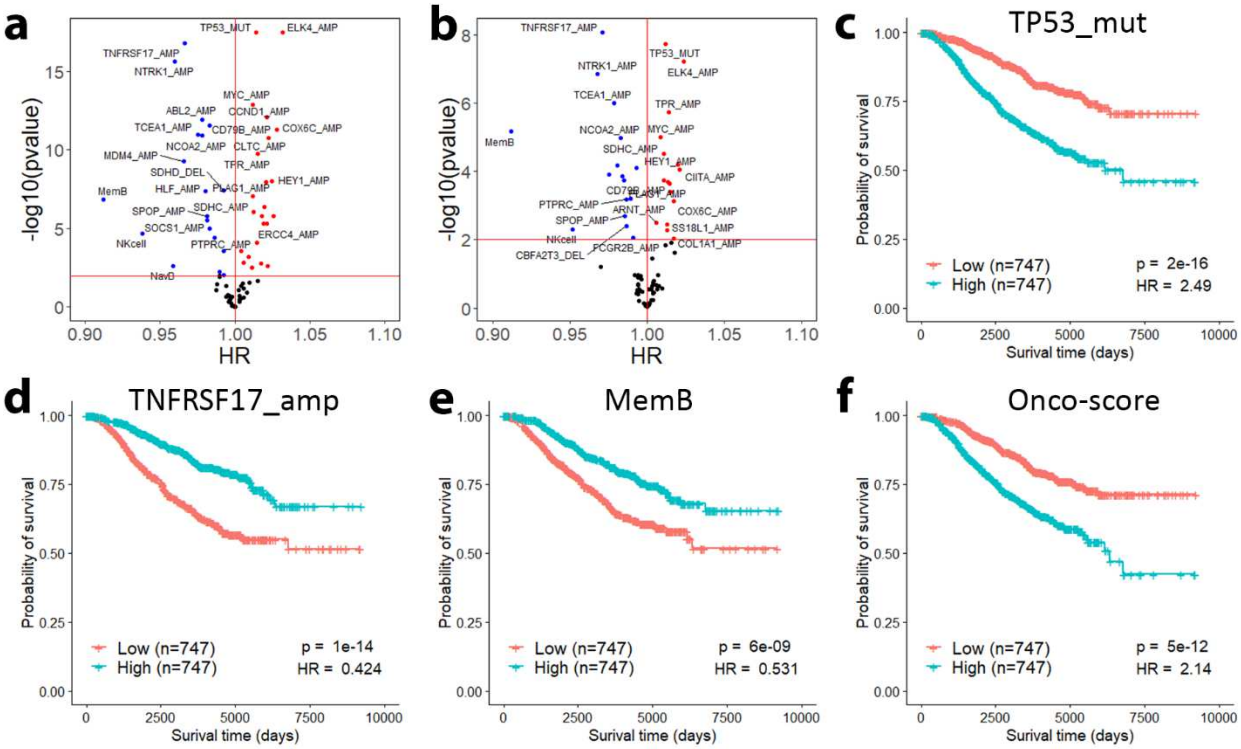

Fig. 5. Individual signature and immune infiltration scores can identify prognostic patient groups. Signature and immune infiltration scores fitted in a univariate Cox proportional hazards model without clinical adjustment (**a**) and with clinical adjustment (**b**) are associated with survival time. Patients are significantly dichotomized by their TP53_mut (**c**), MemB (**d**), and TNFRSF17_amp (**e**) score. **f** Risk groups dichotomized by the median Onco-score have lower or comparable significance to some of our signature and immune infiltration scores.

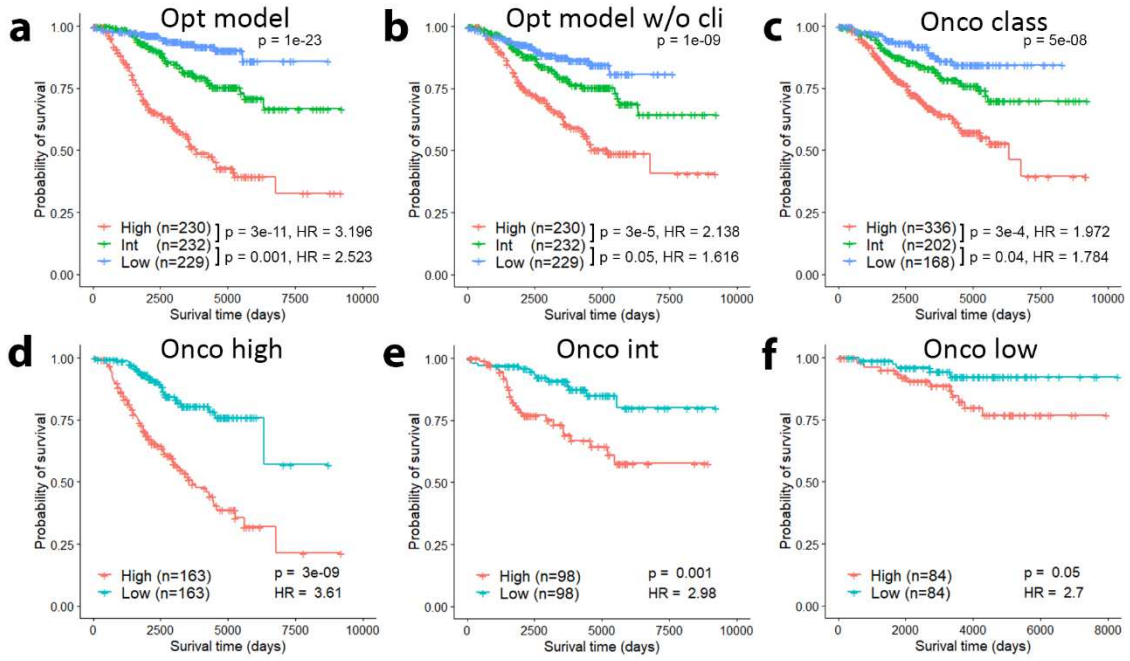

Fig. 6. Optimized Cox regression model for prognostic risk prediction. a Patients in the Curtis validation dataset are significantly grouped by their risk as predicted by the optimized model trained in the Curtis discovery dataset. **b** Patients are still significantly dichotomized by their risk when clinical variables are removed from the optimized model. **c** Onco class achieves slightly lower performance than our optimized model without clinical information in grouping patient risk categories. **d-f** Our optimized model is able to further stratify the Onco high (**d**), intermediate (**e**), and low (**f**) risk classes.

Table 1. Summary of features in each model discussed in the manuscript.

Feature Type	Features
Clinical	Age, Stage, Lymph node status, Grade, Size
Signature (somatic mutation)	*CDH1, GATA3, PIK3CA, TP53*
Signature (gene gain)	*ABL2, ARNT, BRIP1, CCND1, CD79B, CDK12, CIITA, CLTC, COL1A1, COX6C, CREBBP, DDX5, ELK4, ERBB2, ERCC4, EXT1, FCGR2B, FGFR1, FH, FUS, GNAS, H3F3A, HEY1, HLF, HOOK3, IL21R, MDM4, MSI2, MUC1, MYC, MYH11, NCOA2, NDRG1, NTRK1, PALB2, PBX1, PLAG1, PRKAR1A, PTPRC, RAD21, RNF43, SDC4, SDHC, SOCS1, SPOP, SRSF2, SS18L1, TCEA1, TNFRSF17, TPM3, TPR, TSC2, WHSC1L1*
Signature (gene loss)	*ARHGEF12, ATM, BIRC3, CBFA2T3, CDH1, CYLD, FLI1, HERPUD1, MAF, MAP2K4, PCM1, PCSK7, POU2AF1, SDHD, WRN*
Immunological	Naïve B, Memory B, CD4+ T, CD8+ T, NK, Monocytes
Random Forest optimized model	Size, lymph_nodes_positive, age_at_diagnosis, TP53_mut, ABL2_amp, CCND1_amp, CD79B_amp, CIITA_amp, CLTC_amp, COX6C_amp, ELK4_amp, FH_amp, MSI2_amp, MYC_amp, NCOA2_amp, NTRK1_amp, PLAG1_amp, SDC4_amp, SS18L1_amp, TCEA1_amp, TNFRSF17_amp, MemB
Cox optimized model	Size, lymph nodes status, age at diagnosis, ELK4_amp, CCND1_amp, NTRK1_amp, CREBBP_amp, MAP2K4_del, PCM1_del, SDHD_del, PCSK7_del, MemB, SOCS1_amp, DDX5_amp, SDHC_amp, MSI2_amp

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.docx