EBioMedicine 46 (2019) 150–159

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

EBioMedicine
Published by THE LANCET

Research paper

A 23 gene–based molecular prognostic score precisely predicts overall survival of breast cancer patients

Hideyuki Shimizu, Keiichi I. Nakayama *

Department of Molecular and Cellular Biology, Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka, Japan

ARTICLE INFO

Article history:
Received 25 May 2019
Received in revised form 16 July 2019
Accepted 17 July 2019
Available online 26 July 2019

Keywords:
Breast cancer
Prognosis
Scoring system
AI
Personalized medicine

ABSTRACT

Background: Although many prognosis-predicting molecular scores for breast cancer have been developed, they are applicable to only limited disease subtypes. We aimed to develop a novel prognostic score that is applicable to a wider range of breast cancer patients.
Methods: We initially examined The Cancer Genome Atlas breast cancer cohort to identify potential prognosis-related genes. We then performed a meta-analysis of 36 international breast cancer cohorts to validate such genes. We trained artificial intelligence models (random forest and neural network) to predict prognosis precisely, and we finally validated our prediction with the log-rank test.
Findings: We identified a comprehensive list of 184 prognosis-related genes, most of which have been not extensively studied to date. We then established a universal molecular prognostic score (mPS) that relies on the expression status of only 23 of these genes. The mPS system is almost universally applicable to breast cancer patients (log-rank $P < 0.05$) in a manner independent of platform (microarray or RNA sequencing).
Interpretation: The mPS system is simple and cost-effective to apply and yet is able to reveal previously unrecognized heterogeneity among patient subpopulations in a platform-independent manner. The combination of mPS and clinical stage stratifies prognosis even more precisely and should prove of value for avoidance of overtreatment. In addition, the prognosis-related genes uncovered in this study are potential drug targets.
Fund: This work was supported by KAKENHI grants from the Ministry of Education, Culture, Sports, Science, and Technology of Japan to H.S. (19K20403) and to K.I·N (18H05215).

1. Introduction

Cancer is the leading cause of death in developed countries, with breast cancer being one of the most prevalent cancer types in women [1]. Given the long latency and relatively young age of onset for breast cancer, the ability to predict prognosis would be of great value with regard to selection of the optimal therapy for each patient and avoidance of overtreatment [2].

Methods to better stratify individuals at high risk for breast cancer development have been a focus of research interest for more than a decade [3]. Patients have been categorized on the basis of clinical information, with sorting based on TNM (tumour, node, metastasis) stage and the Nottingham Prognostic Index (NPI) [4] having been the most widely accepted clinical classification systems for breast cancer. Although these systems have proved to be of use, overall prognosis can differ markedly even for patients at the same clinical stage [5–7].

Recent technological advances have allowed the development of various molecular prognostic indicators, some of which are recommended in American Society of Clinical Oncology guidelines. The Oncotype Dx 21-gene recurrence score (RS) is the best-validated prognostic assay for breast cancer and estimates the risk of recurrence within 10 years after diagnosis [8–10]. Other useful tools including MammaPrint [11] have been summarized in a recent review [12]. However, these tools are not necessarily universal, given that they are restricted to specific platforms and to subsets of patients based on criteria such as hormone receptor, menopause and nodal status [13]. In addition, none of the tests developed to date are sufficiently fine-tuned to predict overall survival (OS). These limitations are attributable, at least in part, to the fact that no complete atlas of prognosis-related genes has been available, with only limited numbers of genes having been extensively investigated in this regard [14]. This situation highlights the need for unbiased comprehensive approaches to unveil and list all prognosis-related molecules, with a next generation of molecular profiles being anticipated as a result of the application of large-scale sequencing to tumour genomes and transcriptomes [15–17].

We have now developed a novel framework for the prediction of the prognosis of breast cancer patients (Fig. 1). We first examined all protein-coding genes for their relation to OS in breast cancer patients. We then validated 184 prognosis-related genes by meta-analysis of

* Corresponding author.
E-mail address: nakayak1@bioreg.kyushu-u.ac.jp (K.I. Nakayama).

https://doi.org/10.1016/j.ebiom.2019.07.046

Research in context

Evidence before this study

Cancer is the leading cause of death in developed countries, with methods to better stratify susceptible individuals being actively pursued. Recent technological advances have allowed us to develop various molecular prognostic indicators for cancer. However, such indicators for breast cancer are not universal, given that they are restricted to specific platforms and subsets of patients base on criteria such as hormone receptor, menopause and nodal status.

Added value of this study

We integrated statistical and artificial intelligence (AI)–based methods to develop mPS, a universal molecular prognostic score that is able to precisely predict overall survival (OS) and disease free survival of breast cancer patients on the basis of the binary expression status of only 23 genes.

Implications of all the available evidence

We have revealed all OS-related genes for breast cancer, with these genes being potential drug targets. We also developed an AI-based prognosis-prediction score that is applicable to almost all subsets of breast cancer patients. We anticipate that this unbiased approach will not only facilitate appropriate treatment selection for breast cancer patients but also provide molecular insight into the complex nature of this disease.

one of the largest breast cancer cohorts ever assembled. We next applied artificial intelligence (AI)–based methods to develop mPS, a universal molecular prognostic score that is able to precisely predict OS and disease free survival (DFS) of breast cancer patients on the basis of the binary expression status of only 23 genes. Unlike existing tools, mPS was found to be applicable to almost all breast cancer subtypes. We also show that mPS can stratify patients even at the same clinical stage, emphasizing the importance of the combination of mPS with conventional staging systems.

2. Materials and methods

2.1. Study design and cohorts

We performed a retrospective integrated analysis of 40 independent breast cancer cohorts, all published previously. The initial analysis was conducted with The Cancer Genome Atlas (TCGA) breast cancer cohort (discovery cohort) given that this is the best-characterized cohort available. We then performed a meta-analysis (random effects model) to validate the identified prognosis-related genes in a large combined multicenter validation cohort consisting of 36 international breast cancer data sets (Supplementary Table S1) that include 5696 patients with early-stage (IA, IIA, IIB) breast cancer (Fig. 1, Step 1), as previously described [18].

For establishment of the molecular prognostic score (mPS), we adopted another breast cancer data set, the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) breast cancer cohort [19,20]. We used half of the METABRIC cohort as the source of a training set (METABRIC training cohort) for AI-based machine learning (Fig. 1, Step 2) and neural network methods (Fig. 1, Step 3) [21] to develop mPS.

We then validated mPS with the other half of the METABRIC cohort (METABRIC test cohort). We also used two independent breast cancer cohorts (the microarray-based public data set GSE86166 [22] and the RNA-sequencing–based ongoing data set GSE96058 [23]) for further validation of mPS.

Inclusion criteria and clinicopathologic information for the various cohorts are provided in the original papers. Integrative Cluster for the METABRIC cohort and the 12-chemokine gene expression score for GSE86166 were calculated by the providers and included in the public data sets [19,22].

2.2. Gene list

For comprehensive analysis of all protein-coding genes, we obtained a complete list of human genes from the HUGO Gene Nomenclature Committee (HGNC).

2.3. Identification of 184 prognosis-related genes

We downloaded public data from cBioPortal with the CGDS-R package and Web APIs as well as from GEO (https://www.ncbi.nlm.nih.gov/geo). For all human protein-coding genes, we first examined the potential utility of each gene as a prognostic marker with the TCGA breast cancer discovery cohort, and we then validated potential markers by meta-analysis with the 36 international breast cancer cohorts. We adopted a preprocessing pipeline previously described [24]: For Affymetrix data, we applied the MAS5 method [25] for normalization before $\log_2$ conversion for preprocessing, whereas non-Affymetrix data were downloaded as they were deposited in the public databases. For each cohort, we stratified the patients into two groups (high or low expression level for a particular gene) and calculated an integrated hazard ratio (HR) by meta-analysis (Fig. 1, Step 1). We defined a prognosis-related gene as a gene whose 95% confidence interval (CI) for the HR does not cross 1 after meta-analysis.

2.4. Generation and validation of the mPS scoring system

The combining of several machine learning approaches, so-called "ensemble learning," has been shown to improve prediction performance. In particular, combination of an AI-based machine learning algorithm known as random forest with a neural network was found to be effective in many machine learning tasks, including those with transcriptome data [26]. We therefore applied these two approaches to build the mPS system (Supplementary Fig. S1). For the development and subsequent validation of mPS, we used the METABRIC cohort.

We first applied data from half of the METABRIC cohort (training set, $n = 952$) to a random forest classifier. Expression levels of the 184 newly identified prognosis-related genes (designated X) and the survival status [designated t, alive (0) or deceased (1)] at 10 years after diagnosis for each patient in the METABRIC training set ($n = 952$) were thus entered into the random forest classifier. We generated this model with the use of the Python-based scikit-learn library and with default parameters with the exception of n_estimators ($= 500$) and max_depth ($= 10$). After stratified 10-fold cross validation (CV $= 10$), we selected 23 genes on the basis of feature importance values (cutoff $= 0.0075$). These 23 genes could account for OS of the patients, with 13 and 10 genes being associated with a poor OS if their expression level is higher or lower than the median, respectively.

The expression status (X) of the 23 genes was first transformed to "Gene_Score" (S) on the basis of the expression level (above or below the median) and integrated HR for each gene with the following step function:

Gene_Score matrix		Gene_Expression	
		Low	High
Integrated HR	<1	1	0
	>1	0	1

We then built and trained a dense neural network system (Supplementary Fig. S1). In each hidden node, we exploited ReLU (rectified

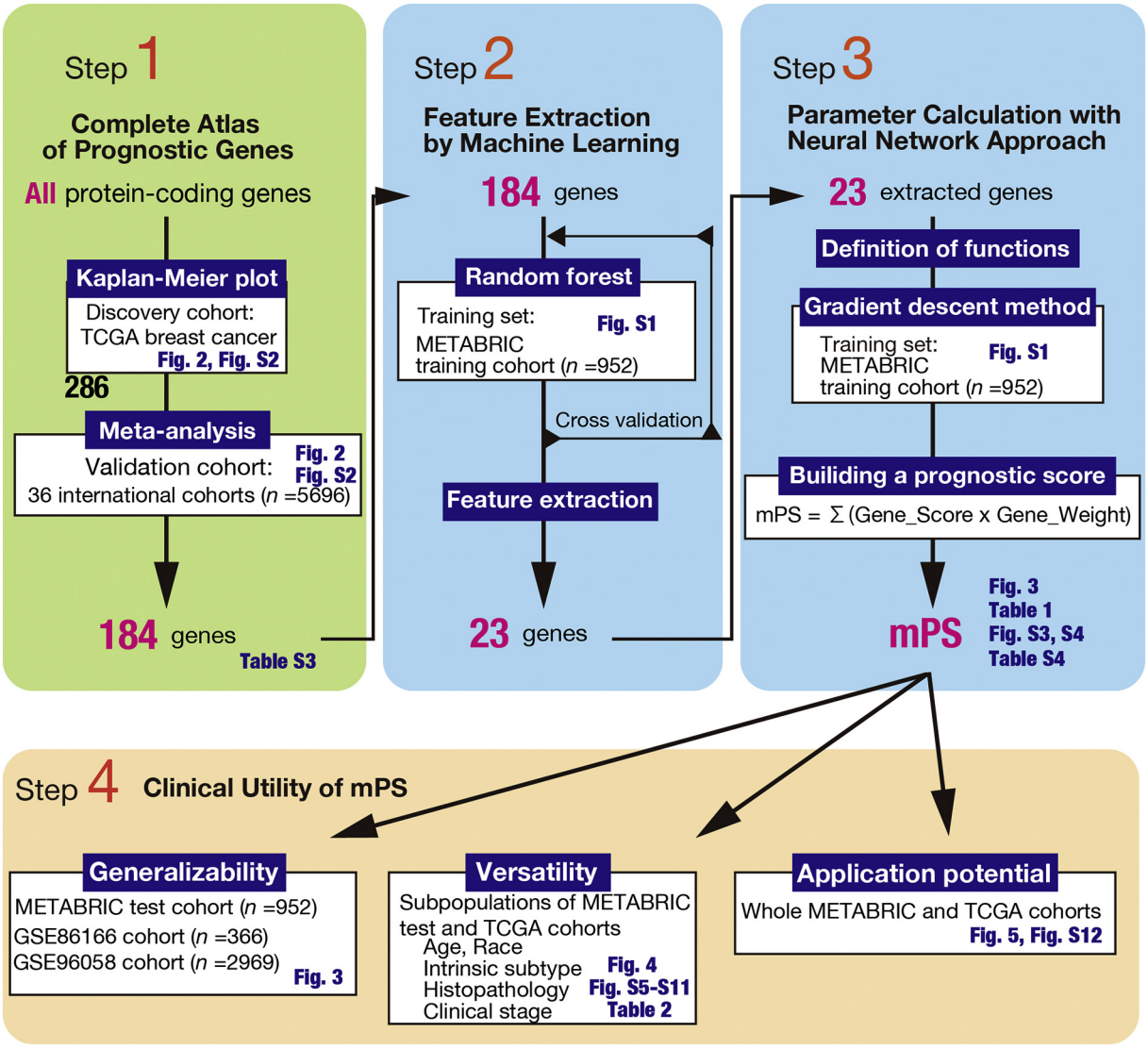


Fig. 1. Study overview. All protein-coding genes were tested for their potential as prognosis-related genes with the use of the TCGA breast cancer cohort and 36 independent multicenter data sets (Step 1). Machine learning with the random forest approach reduced the number of validated genes to 23 (Step 2). A versatile prognostic score, designated mPS, was established with the use of a neural network approach (Step 3). Finally, the utility of mPS was validated in various settings (Step 4).

linear unit) as an activation function. In the output layer, we created two nodes (a_1 and a_2, for alive and deceased, respectively). We applied a softmax function to each node, and designated y_2 (probability of death; that is, the a_2 node) as Y. We utilized cross entropy error as a loss function (E) and optimized the value of each weight with Adam method (learning rate, 0·001; epochs, 1000). After the training, we used the weights of the nodes ("Gene_Weight") to calculate mPS (summation of Gene_Score x Gene_Weight for all 23 genes). We used the Python-based Keras library for this neural network training.

For validation, we used the other half of the METABRIC cohort (METABRIC test set) and the independent cohorts GSE86166 and GSE96058. Within each cohort, we converted expression level to binary status (above or below the median), which was then converted to Gene_Score by the above-mentioned step function.

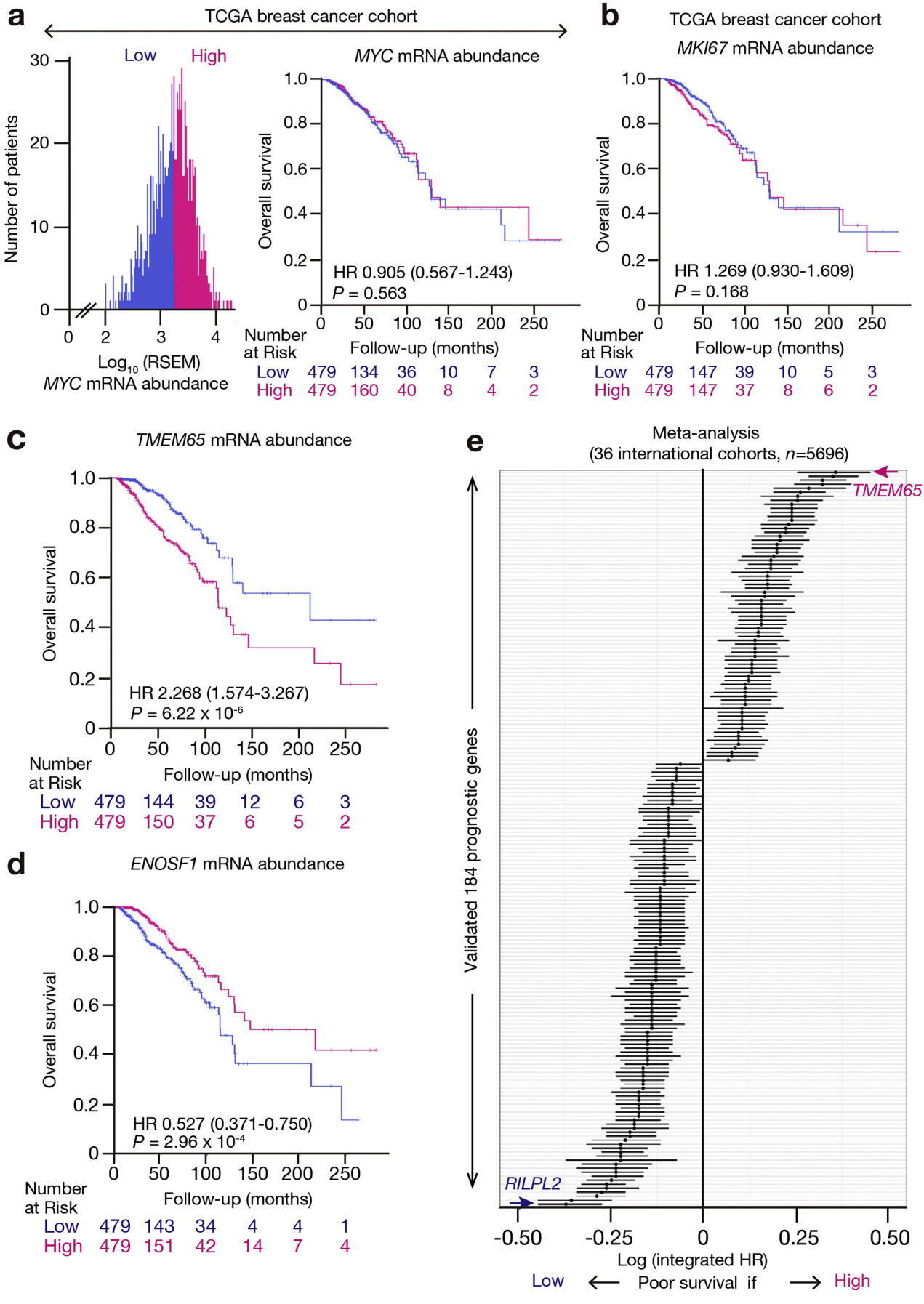
The cutoff criterion (median value) was study specific and calculated for each cohort independently.

2.5. Statistics

Kaplan-Meier plots were constructed with the use of R (survival package). The median value was used as the cutoff between low and high expression levels of each gene. For mPS validation, we truncated the survival data at 10 years unless indicated otherwise. We computed OS from the date of diagnosis to the date of death from any cause. For most of the data (Figs. 3–5 and Supplementary Figs. S5–S11), survival outcomes were compared with the log-rank test. For the survival analysis shown in Fig. 2 and Table 2, the HR and its 95% CI were calculated by Cox regression analysis after proper evaluation of the assumptions of the Cox regression models with the use of the survival package. Statistical significance was determined at a two-sided *P* value of 0·05, with the exception of the TCGA discovery cohort, for which we adopted 0·01 as the cutoff criterion.

2.6. Data availability

All the data analyzed in this study are open to the public and can be downloaded from cBioPortal (http://www.cbioportal.org) and GEO (https://www.ncbi.nlm.nih.gov/geo). A Web-based tool we created in this study is freely available at our github page (https://hideyukishimizu.github.io/mPS_breast).


a
TCGA breast cancer cohort
Low
High
Number of patients
Log10 (RSEM)
MYC mRNA abundance
MYC mRNA abundance
Overall survival
HR 0.905 (0.567-1.243)
P = 0.563
Follow-up (months)
Number at Risk
Low 479 134 36 10 7 3
High 479 160 40 8 4 2
b
TCGA breast cancer cohort
MKI67 mRNA abundance
Overall survival
HR 1.269 (0.930-1.609)
P = 0.168
Follow-up (months)
Number at Risk
Low 479 147 39 10 5 3
High 479 147 37 8 6 2
c
TMEM65 mRNA abundance
Overall survival
HR 2.268 (1.574-3.267)
P = 6.22 x 10-6
Follow-up (months)
Number at Risk
Low 479 144 39 12 6 3
High 479 150 37 6 5 2
d
ENOSF1 mRNA abundance
Overall survival
HR 0.527 (0.371-0.750)
P = 2.96 x 10-4
Follow-up (months)
Number at Risk
Low 479 143 34 4 4 1
High 479 151 42 14 7 4
e
Meta-analysis
(36 international cohorts, n=5696)
TMEM65
RILPL2
Validated 184 prognostic genes
Log (integrated HR)
Low ← Poor survival if → High

3. Results

3.1. Limitation of hypothesis-driven approaches

Since the initial discovery of the cancer-causing *Src* gene, tremendous advances have been made in the field of oncogenes. Given that *MYC* plays many important roles related to cancer development [27], we hypothesized that the expression of *MYC* might be associated with OS in cancer patients. Breast cancer patients in the TCGA cohort [28] were divided into two groups (low and high expression level) on the basis of the median mRNA abundance for *MYC*, and the difference in survival outcome between the two groups was assessed. Unexpectedly, OS did not differ between the two groups (Fig. 2a). Similarly, the mRNA level for *MKI67*, which is frequently examined in clinicopathologic studies [29], was not related to OS in the TCGA breast cancer cohort (Fig. 2b).

3.2. Computational elucidation of all prognosis-related genes

We therefore examined the relation between mRNA abundance for all protein-coding genes and OS with the TCGA breast cancer data set as a discovery cohort. Although the expression of most (18,894) genes was not associated with OS, a high expression level of 117 genes including *TMEM65* (Fig. 2c) and *PGK1* (Supplementary Fig. S2a) was related to reduced survival. Conversely, the OS of patients with a low expression level of 169 genes including *ENOSF1* (Fig. 2d) and *BEND5* (Supplementary Fig. S2b) was poorer than that of those with a corresponding high level of expression. We thus identified a total of 286 OS-related genes in the TCGA discovery cohort, with the complete list of these genes and their log-rank *P* values being provided in Supplementary Table S2.

We next subjected these 286 potential prognostic genes identified with the TCGA discovery cohort to validation by meta-analysis of combined multicentre breast cancer cohorts (Supplementary Table S1) [18], revealing that 184 of these genes were also prognosis-related genes in the validation data set (Fig. 2e). *TMEM65* and *RILPL2* were the most promising prognosis-related genes, with the highest and lowest HRs, respectively, in this multicentre validation cohort (Supplementary Fig. S2, c and d). It is of note that these two genes were not adopted by the MammaPrint or Oncotype Dx 21-gene RS systems. Indeed, most of the validated prognosis-related genes have not been well characterized to date with regard to their relation to basic or clinical oncology, with *TMEM65* and *RILPL2* apparently not having been studied at all in this field. These results thus revealed the promise of our computer-based comprehensive approach to uncovering previously uncharacterized, yet important genes in breast cancer. The complete list of validated prognosis-related genes with their estimated HRs is provided in Supplementary Table S3.

3.3. AI-based development of a molecular prognostic score

We examined whether these 184 newly identified prognosis-related genes might suffice to predict the survival rate of breast cancer patients at 10 years. We used a third breast cancer data set, METABRIC [19,20], to build a molecular prognostic score system as described in detail in the Materials and methods section and Supplementary Fig. S1.

In brief, we first applied data from half of the METABRIC cohort (training set, $n = 952$) to a machine learning algorithm known as a random forest classifier and thereby selected 23 genes. We optimized the weight for each gene with a neural network algorithm. We thus built a molecular prognostic score (mPS) that is calculated by summation of "Gene_Score" × "Gene_Weight" for all 23 genes, with the potential value ranging from 0 to 50 (Table 1). Two examples of actual mPS calculations are presented (Supplementary Fig. S3). For the METABRIC training cohort, the mean of mPS was 24·22 (interquartile range [IQR] of 15·56–33·60), and its distribution pattern is shown in Supplementary Fig. S4. The characteristics of mPS groups based on assignment to six bins are summarized for the METABRIC training cohort in Supplementary Table S4. The mPS system is well correlated with pathological tumour grade, clinical TNM stage, and the NPI.

Table 1
The 23 genes necessary and sufficient for calculation of mPS. For genes in red, patients with a high level of expression (above the median) are assigned a score of 1. Conversely, for genes in blue, patients with a low level of expression (below the median) are assigned a score of 1. None of the 23 genes are included in existing indicators of relapse-free survival such as Oncotype and MammaPrint.

Symbol	Gene ID	Full name	Score (high)	Score (low)	Weight
FOXM1	2305	Forkhead box M1	1	0	3.424
CPT1A	1374	Carnitine palmitoyltransferase 1A	1	0	3.399
GARS	2617	Glycyl-tRNA synthetase	1	0	2.539
MARS	4141	Methionyl-tRNA synthetase	1	0	2.312
UTP23	84,294	UTP23, small subunit processome component	1	0	2.311
ANLN	54,443	Anillin actin binding protein	1	0	2.225
HMGB3	3149	High mobility group box 3	1	0	2.202
ATP5F1B	506	ATP synthase F1 subunit beta	1	0	1.934
APOOL	139,322	Apolipoprotein O like	1	0	1.754
CYB561	1534	Cytochrome b561	1	0	1.594
GRHL2	79,977	Grainyhead like transcription factor 2	1	0	1.526
ESRP1	54,845	Epithelial splicing regulatory protein 1	1	0	1.485
EZR	7430	Ezrin	1	0	1.372
RBBP8	5932	RB binding protein 8, endonuclease	0	1	3.095
CIRBP	1153	Cold inducible RNA binding protein	0	1	3.083
PTGER3	5733	Prostaglandin E receptor 3	0	1	2.802
LAMA3	3909	Laminin subunit alpha 3	0	1	2.601
OARD1	221,443	O-acyl-ADP-ribose deacylase 1	0	1	2.008
ANKRD29	147,463	Ankyrin repeat domain 29	0	1	1.886
EGR3	1960	Early growth response 3	0	1	1.836
DIRAS3	9077	DIRAS family GTPase 3	0	1	1.821
MITD1	129,531	Microtubule interacting and trafficking domain containing 1	0	1	1.425
LAMB3	3914	Laminin subunit beta 3	0	1	1.366
			mPS = Σ (Gene_Score x Gene_Weight)		

3.4. mPS stratifies prognosis of independent cohorts

To study whether mPS can stratify prognosis not only in the METABRIC training cohort but also in other independent breast cancer cohorts, we first examined the other half of the METABRIC data set (METABRIC test cohort). We found that mPS stratifies prognosis in this test cohort (Fig. 3a). The mPS shows superiority to PAM50 classification, which is widely used in clinical settings, with regard to the stratification of prognosis (Fig. 3, a and b). It is also of note that mPS has two advantages over Integrative Cluster, which was originally proposed by the provider of the METABRIC data set [19]. The stratification based on mPS is thus more significant than that based on Integrative Cluster (Fig. 3, a and c), and mPS is less expensive to apply than Integrative Cluster, for which whole-genome sequence analysis is required.

Fig. 2. Identification of all prognosis-related genes in the TCGA breast cancer cohort and validation in 36 independent cohorts. (a) Distribution of *MYC* expression level (RSEM) among patients in the TCGA breast cancer cohort (left), and Kaplan-Meier curves of OS for these patients based on a *MYC* expression level higher or lower than the median (right). The HR, its 95% CI, and the log-rank *P* value are shown. (b) Kaplan-Meier curves of OS for the TCGA cohort based on *MKI67* expression level. (c and d) Kaplan-Meier curves of OS for the TCGA cohort based on *TMEM65* (c) and *ENOSF1* (d) expression levels, respectively. The complete list of OS-related genes in this TCGA discovery cohort is provided in Supplementary Table S2. (e) Logarithm of the integrated HR for all 184 prognosis-related genes in the validation data sets. The complete list of these genes identified by meta-analysis is provided in Supplementary Table S3.

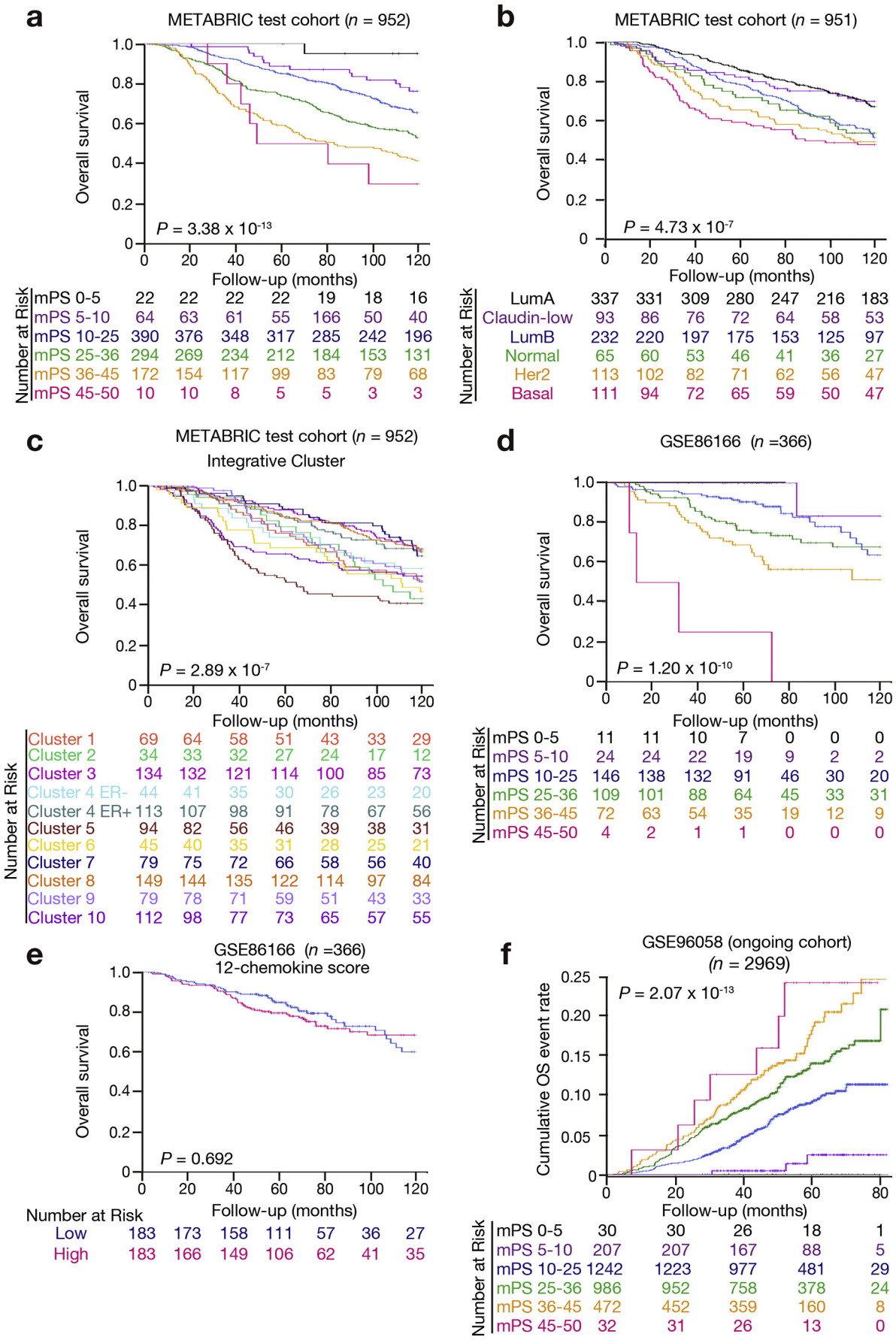


Fig. 3. mPS precisely stratifies prognosis of breast cancer patients. (a) Kaplan-Meier curves of OS according to mPS for the METABRIC test cohort. (b) Kaplan-Meier curves of OS according to PAM50 classification for the METABRIC test cohort. We omitted one patient whose PAM50 classification was not available. (c) Kaplan-Meier curves of OS for the METABRIC test cohort according to Integrative Cluster, which was proposed by the provider of the METABRIC data set. (d) Kaplan-Meier curves of OS according to mPS for the public data set GSE86166. (e) Kaplan-Meier curves of OS for the GSE86166 data set according to the 12-chemokine gene expression score proposed by the provider of the data set. (f) Kaplan-Meier curves of OS according to mPS for the public data set GSE96058.

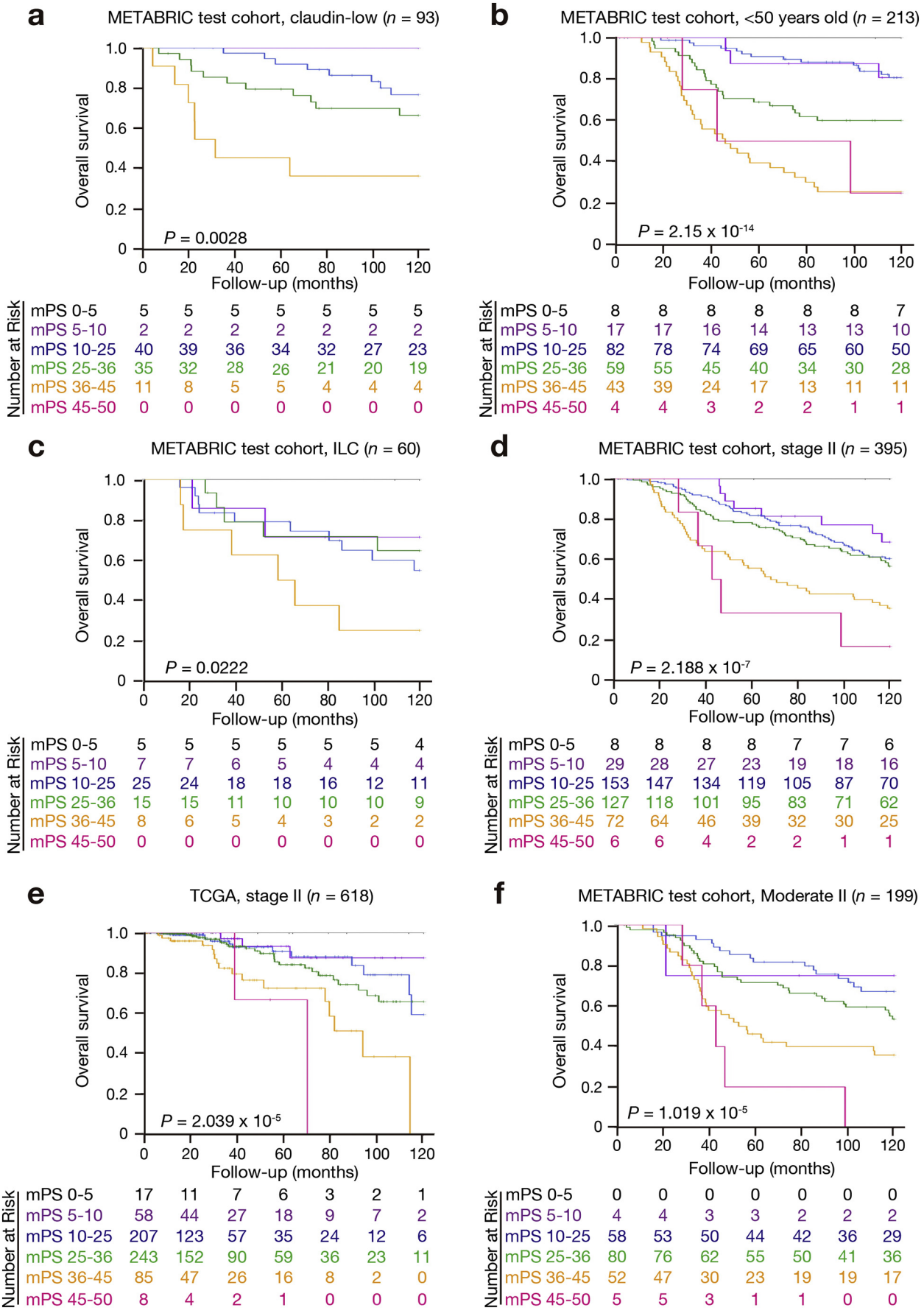


Fig. 4. mPS is applicable to most breast cancer subsets. (a) Kaplan-Meier curves according to mPS for OS of patients with claudin-low tumours in the METABRIC test cohort. (b) Kaplan-Meier curves according to mPS for OS of premenopausal patients (<50 years of age) in the METABRIC test cohort. (c) Kaplan-Meier curves according to mPS for OS of patients in the METABRIC test cohort with ILC. (d and e) Kaplan-Meier curves according to mPS for OS of patients in the METABRIC test (d) and TCGA (e) cohorts at clinical TNM stage II. (f) Kaplan-Meier curves according to mPS for OS of patients in the Moderate II cluster of the NPI in the METABRIC test cohort.

Table 2
Univariate and multivariate analyses of OS in the TCGA cohort. The HR relative to the indicated reference (ref) value, its 95% CI, and *P* value (those of <0.05 are indicated in bold) for the Cox hazard model are shown.

	Univariate			Multivariate		
	Hazard ratio	95% CI	*P*	Hazard ratio	95% CI	*P*
Age						
<50	1 (ref)			1 (ref)		
50–70	1.11	0.737–1.681	0.610	1.02	0.655–1.587	0.931
>70	2.59	1.656–4.049	<0.001	2.47	1.541–3.972	<0.001
Gender						
Female	1 (ref)			1 (ref)		
Male	0.84	0.117–6.000	0.859	0.72	0.099–5.185	0.741
Stage						
1	1 (ref)			1 (ref)		
2	1.91	1.025–3.550	0.042	2.00	1.069–3.732	0.030
3	3.95	2.087–7.491	<0.001	3.8	1.992–7.243	<0.001
4	15.75	7.274–34.088	<0.001	10.64	4.824–23.453	<0.001
mPS						
<10	1 (ref)			1 (ref)		
10–36	2.62	1.210–5.648	0.015	2.36	1.086–5.116	0.030
>36	7.76	3.498–17.229	<0.001	5.45	2.423–12.245	<0.001

Many existing prognostic indicators are able to predict prognosis on the basis on only one specific platform or pipeline. We overcame this limitation by changing continuous gene expression values (which may vary depending on method) to discrete values (high or low relative to the median), rendering mPS independent of platform. For demonstration purposes, we analyzed another breast cancer data set, GSE86166, in which transcriptome profiling was performed by microarray analysis [22]. We found that mPS stratifies OS into different bins in the same way as in the METABRIC test cohort (Fig. 3d), showing that mPS is applicable to both RNA-sequencing–based (METABRIC) and microarray-based (GSE86166) data sets. In addition, the mPS system is superior to the 12-chemokine gene expression score [22] proposed by the provider of the GSE86166 data set (Fig. 3, d and e).

We also analyzed GSE96058, an ongoing cohort in Sweden, in which nearly 3000 breast cancer patients are followed up for up to 7 years [23]. Our mPS system also stratifies these patients into different bins, although the event (death) rate is relatively low in this cohort, likely because of the shorter follow-up time (Fig. 3f).

Furthermore, mPS stratifies not only OS but also DFS in both the RNA sequencing–based TCGA breast cancer cohort (Supplementary Fig. S5a) and the microarray-based cohort GSE86166 (Supplementary Fig. S5b).

Together, these various lines of evidence show that mPS allows the precise stratification of prognosis into distinct groups in a platform-independent manner, demonstrating its general applicability. Given that we developed mPS computationally, these results indicate that this system concisely reflects transcriptome alterations necessary for tumour progression and that it will therefore be of value not only for clinical oncologists but also for basic biomedical researchers.

3.5. mPS is applicable to most breast cancer subsets

We next investigated the utility of mPS for various subtypes of breast cancer. Application of the mPS system to each of the PAM50 intrinsic subtypes revealed that not only estrogen receptor–positive (lumA/lumB) patients (Supplementary Fig. S6a) but also patients with HER2-enriched (Supplementary Fig. S6b), claudin-low (Fig. 4a), or normal-like (Supplementary Fig. S6c) subtypes are well stratified. Although the mPS system could not stratify the prognosis of patients with basal-like tumours into six groups, likely as a result of the malignant nature of these tumours (most such patients had an mPS of >25), OS tended to be better in mPS-low (<25) patients than in mPS-high (>25) patients (Supplementary Fig. S6d).

The mPS system precisely predicted OS not only of patients in their 50s and 60s (Supplementary Fig. S7) but also of younger (Fig. 4b) patients, showing that mPS is applicable to patients of various ages regardless of menopausal status.

Most breast cancer specimens are classified pathologically as invasive ductal carcinoma (IDC), and we found that the mPS system is able to stratify the prognosis of IDC patients (Supplementary Fig. S8a). This system also clarifies distinct subpopulations of invasive lobular carcinoma (ILC) (Fig. 4c), the second most frequent histological subtype of breast cancer, as well as of the mixed IDC and ILC (MDLC) subtype (Supplementary Fig. S8b). Although ILC and MDLC differ histopathologically from IDC [30], the mPS system could thus be applied to all three major pathological subtypes of breast cancer.

We also found that mPS system is applicable not only for Caucasian (Supplementary Fig. S9a), but also Black or African American (Supplementary Fig. S9b), and Asian (Supplementary Fig. S9c) patients in the TCGA cohort, demonstrating that mPS is not race specific.

3.6. mPS is suitable even for patients at the same clinical stage

We further examined whether mPS is also applicable to well-established TNM tumour stages determined from clinical information. The mPS system revealed that stage I patients in the METABRIC test cohort ($n = 246$) are heterogeneous, with >90% of individuals with an mPS of <10 surviving for >10 years whereas only ~70% of patients with an mPS of >25 survived this long (Supplementary Fig. S10a). This trend was more prominent for the stage II patients ($n = 395$), with those with an mPS of <5 showing excellent prognosis and those with an mPS of >45 having the worst prognosis (Fig. 4d). The mPS system also stratified OS of patients at the same clinical stage in the TCGA cohort (Fig. 4e). Even for stage III patients in the METABRIC test cohort ($n = 59$), mPS-low (<25) individuals showed a better prognosis than did their mPS-high counterparts (Supplementary Fig. S10b).

We next evaluated the relation of mPS to NPI, which is calculated on the basis of the size of the primary tumour, the number of involved lymph nodes, and the tumour grade [31]. We found that each NPI group was still substantially heterogeneous with regard to mPS. For example, the Moderate II group was further divided by mPS, with the prognosis of mPS-low patients being much better than that of mPS-high patients (Fig. 4f). Similar results were obtained for the other NPI groups (Supplementary Fig. S11).

We also performed univariate and multivariate analyses (Cox proportional hazards model) with the TCGA cohort and found that mPS stratifies prognosis independently of age, gender, and tumour stage (Table 2). We thus conclude that the mPS system further stratifies patients even at the same clinical stage.

3.7. Combination of mPS with clinical stage

Finally, we propose an integrated classification system that is based on the combination of mPS and clinical stage and which consists of seven classes (Fig. 5a). Given that patients with distant metastasis at diagnosis (stage IV) are generally inoperable and their mPS therefore cannot be estimated, they are categorized as class F-II. This integrated score revealed that patients of class A, which mostly comprise individuals without distant metastasis and with an mPS of <5, survive longer than other patients, regardless of clinical stage (Fig. 5b). Moreover, for patients of class F-I, which largely comprise individuals with an mPS of >45, OS is poor even for those at stage I. This seven-class system thus precisely stratifies OS of breast cancer patients.

Evaluation of the predictive value of mPS revealed that OS of patients of classes A and B in the METABRIC cohort was not affected by cytotoxic chemotherapy (Fig. 5c), suggesting that such patients should not be subjected to such treatment so as to avoid possible adverse events. For more severe cases (classes C to F-II), however, patients who received cytotoxic chemotherapy showed a poorer prognosis compared with those who did not in this data set (Supplementary Fig. S12a), probably because patients with faster disease progression are more likely to receive

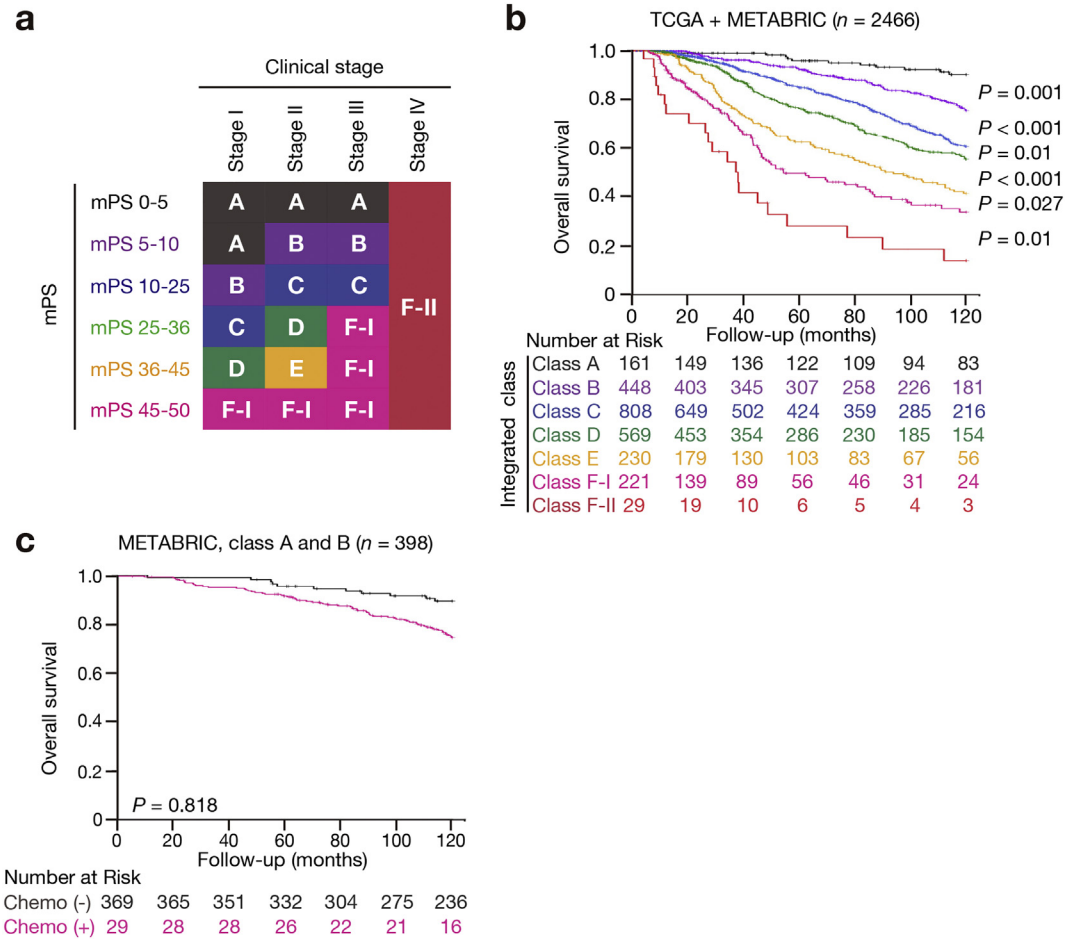


Fig. 5. Combination of mPS with clinical stage for facilitation of treatment selection. (a) Proposed classification of breast cancer patients on the basis of both mPS and clinical stage. (b) Kaplan-Meier curves according to class based on mPS and clinical stage for OS in patients of the combined TCGA and METABRIC cohorts. (c) Kaplan-Meier curves for patients of class A or B in the METABRIC cohort according to whether they received cytotoxic chemotherapy or not. Only patients with available stage information are included.

chemotherapy (Supplementary Fig. S12b). Collectively, these data suggest that mPS is also informative with regard to the avoidance of overtreatment in certain classes of patients.

4. Discussion

In this study, we have delineated a complete atlas of prognosis-related genes for breast cancer and developed a computational framework and new prognostic prediction score designated mPS that is applicable to almost all subsets of breast cancer patients. The mPS system is simple and cost-effective to apply and yet is able to reveal previously unrecognized heterogeneity among patient subpopulations in a platform-independent manner. We also provide a Web-based tool (https://hideyukishimizu.github.io/mPS_breast) that allows clinicians to estimate prognosis of patients and select an optimal therapy.

Unfortunately, we were unable to compare the performance of mPS with that of other representative tools including MammaPrint and Oncotype Dx 21–gene RS, given that these are commercial products and the formulas for their calculation have not been disclosed. However, we demonstrated the superiority of mPS relative to PAM50 classification (Fig. 3, a and b), which is routinely applied in hospitals, as well as to two recently proposed prognostic indicators [19,22] with their own data sets (Fig. 3, a and c–e). Our method is likely to outperform previous scores because mPS stratifies patients at the same clinical stage (Fig. 4, d–f) as well as those with estrogen receptor–negative subtypes of breast cancer (Fig. 4a), unlike existing methods.

There are numerous protocols for preservation of tumour samples, RNA extraction, and analysis of expression status, which hindered us from establishing one universal cutoff for each of the 23 genes in the present study. We aimed to build a "platform-independent" score that can be calculated from data obtained by any method once the necessary protocols and distribution patterns obtained with these protocols are established. Comparison of these protocols and development of a robust and precise method to examine the expression levels of the 23 genes, followed by the performance of pilot studies to test the distribution patterns, are remaining challenges that must be addressed before mPS can be applied in the clinical setting.

Other limitations of our study include the fact that all analyses were performed in a retrospective manner. Although the total number of patients analyzed ($n = 11{,}893$), including the ongoing cohort GSE96058 [23], is among the largest of those previously examined, prospective studies will be needed to validate our findings.

The best-characterized gene among the 23 prognosis-related genes identified in the present study is *FOXM1*. A PubMed search for "FOXM1 breast cancer" identified ~180 papers. The FOXM1 protein functions as a transcriptional activator. It is phosphorylated in M phase of the cell cycle and up-regulates the expression of several proliferation-related genes including those for cyclin B1 and Skp2, the latter of which plays an essential role in cell cycle progression by mediating the ubiquitin-dependent degradation of the cyclin-dependent kinase inhibitors p21, p27, and p57 [32]. The prognostic value of FOXM1 for solid tumours as identified by meta-analysis is also documented in a recent review [33]. In contrast, most of the 23 prognosis-related genes (*GARS*, *UTP23*, *HMGB3*, *ATP5F1B*, *CYB561*, *EZR*, *CIRBP*, *PTGER3*, *LAMA3*, *OARD1*, *ANKRD29*, *MITD1*, and *LAMB3*) have not been studied in relation to breast cancer, given that PubMed searches for "*GENE*

breast cancer" identified fewer than 10 publications for each gene, with there being no published papers at all for five of these genes (*UTP23*, *CYB561*, *OARD1*, *ANKRD29I*, and *MITD1*). Both basic and clinical studies will be necessary for further elucidation of the fundamental mechanisms responsible for the effects of the 23 genes on which mPS is based and for the development of novel drugs to prolong OS of breast cancer patients.

We expect that application of mPS will not only facilitate selection of therapeutic strategies on the basis of the precise prediction of personal prognosis, but also contribute to further understanding of the basic biology of breast cancer and thereby inform the development of new therapeutic approaches.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2019.07.046.

Funding sources

This work was supported by a KAKENHI grant from the Ministry of Education, Culture, Sports, Science, and Technology of Japan to H.S. (19K20403) and K.I.N (18H05215). The funding agency played no role in this study.

Author contributions

H.S. designed the study, analyzed the data, and developed mPS. K.I.N. coordinated the study and wrote the manuscript. Both authors read and approved the final version of the manuscript.

Declaration of Competing Interest

The authors declare no potential conflicts of interest.

Acknowledgments

We thank S. Miyano for comments on computational and statistical methods; K. Mimori for advice on the clinical aspects of breast cancer; S. Fujinuma, Y. Yamauchi, and other laboratory members for discussion; and A. Ohta for help with preparation of the manuscript.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018;68:7–30.
[2] Kwa M, Makris A, Esteva FJ. Clinical utility of gene-expression signatures in early stage breast cancer. Nat Rev Clin Oncol 2017;14:595–610.
[3] Esteva FJ, Sahin AA, Cristofanilli M, Arun B, Hortobagyi GN. Molecular prognostic factors for breast cancer metastasis and survival. Semin Radiat Oncol 2002;12:319–28.
[4] Haybittle JL, Blamey RW, Elston CW, et al. A prognostic index in primary breast cancer. Br J Cancer 1982;45:361–6.
[5] Carter CL, Allen C, Henson DE. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. Cancer 1989;63:181–7.
[6] Anderson WF, Rosenberg PS, Prat A, Perou CM, Sherman ME. How many etiological subtypes of breast cancer: two, three, four, or more? J Natl Cancer Inst 2014;106 [pii: dju165].
[7] Yeo SK, Guan JL. Breast Cancer: multiple subtypes within a tumor? Trends Cancer 2017;3:753–60.
[8] Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–26.
[9] Paik S, Tang G, Shak S, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol 2006;24:3726–34.
[10] Sparano JA, Gray RJ, Makower DF, et al. Prospective validation of a 21-gene expression assay in breast cancer. N Engl J Med 2015;373:2005–14.
[11] van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.
[12] Nicolini A, Ferrari P, Duffy MJ. Prognostic and predictive biomarkers in breast cancer: past, present and future. Semin Cancer Biol 2018;52:56–73.
[13] Krop I, Ismaila N, Andre F, et al. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: american society of clinical oncology clinical practice guideline focused update. J Clin Oncol 2017;35:2838–47.
[14] Sun X, Pittard WS, Xu T, et al. Omicseq: a web-based search engine for exploring omics datasets. Nucleic Acids Res 2017;45:W445–52.
[15] Banerji S, Cibulskis K, Rangel-Escareno C, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature 2012;486:405–9.
[16] Ciriello G, Gatza ML, Beck AH, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell 2015;163:506–19.
[17] Gupta A, Mutebi M, Bardia A. Gene-expression-based predictors for breast cancer. Ann Surg Oncol 2015;22:3418–32.
[18] Abdel-Fatah TMA, Agarwal D, Liu DX, et al. SPAG5 as a prognostic biomarker and chemotherapy sensitivity predictor in breast cancer: a retrospective, integrated genomic, transcriptomic, and protein analysis. Lancet Oncol 2016;17:1004–18.
[19] Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486:346–52.
[20] Pereira B, Chin SF, Rueda OM, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat Commun 2016;7:11479.
[21] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17.
[22] Prabhakaran S, Rizk VT, Ma Z, et al. Evaluation of invasive breast cancer samples using a 12-chemokine gene expression score: correlation with clinical outcomes. Breast Cancer Res 2017;19:71.
[23] Brueffer C, Vallon-Christersson J, Grabau D, et al. Clinical value of RNA sequencing–based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter Sweden cancerome analysis network—breast initiative. JCO Precis Oncol 2018. https://doi.org/10.1200/PO.17.00135.
[24] Jezequel P, Campone M, Gouraud W, et al. bc-GenExMiner: an easy-to-use online platform for gene prognostic analyses in breast cancer. Breast Cancer Res Treat 2012;131:765–75.
[25] Hubbell E, Liu WM, Mei R. Robust estimators for expression analysis. Bioinformatics 2002;18:1585–92.
[26] Kong Y, Yu T. A deep neural network model using random forest to extract feature representation for gene expression data classification. Sci Rep 2018;8:16477.
[27] Dang CV. A time for MYC: metabolism and therapy. Cold Spring Harb Symp Quant Biol 2016;81:79–83.
[28] The TCGA Network. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.
[29] Pathmanathan N, Balleine RL. Ki67 and proliferation in breast cancer. J Clin Pathol 2013;66:512–6.
[30] Pestalozzi BC, Zahrieh D, Mallon E, et al. Distinct clinical and prognostic features of infiltrating lobular carcinoma of the breast: combined results of 15 international breast cancer study group clinical trials. J Clin Oncol 2008;26:3006–14.
[31] Rakha EA, Soria D, Green AR, et al. Nottingham prognostic index plus (NPI+): a modern clinical decision making tool in breast cancer. Br J Cancer 2014;110:1688–97.
[32] Koo CY, Muir KW, Lam EW. FOXM1: from cancer initiation to progression and treatment. Biochim Biophys Acta 2012;1819:28–37.
[33] Li L, Wu D, Yu Q, Li L, Wu P. Prognostic value of FOXM1 in solid tumors: a systematic review and meta-analysis. Oncotarget 2017;8:32298–308.