Breast Cancer Research and Treatment (2019) 175:691–699
https://doi.org/10.1007/s10549-019-05177-0

EPIDEMIOLOGY

Integrating of genomic and transcriptomic profiles for the prognostic assessment of breast cancer

Chengxiao Yu[1,2,3] · Na Qin[1,2,3] · Zhening Pu[1,2,3,5] · Ci Song[1,2,3] · Cheng Wang[1,2,3,4] · Jiaping Chen[1,2,3] · Juncheng Dai[1,2,3] · Hongxia Ma[1,2,3] · Tao Jiang[1,2,3] · Yue Jiang[1,2,3]

Received: 19 December 2018 / Accepted: 19 February 2019 / Published online: 13 March 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Purpose To evaluate the prognostic effect of the integration of genomic and transcriptomic profiles in breast cancer.

Methods Eight hundred and ten samples from the Cancer Genome Atlas (TCGA) data sets were randomly divided into the training set (540 subjects) and validation set (270 subjects). We first selected single-nucleotide polymorphism (SNPs) and genes associated with breast cancer prognosis in the training set to construct the prognostic prediction model, and then replicated the prediction efficiency in the validation set.

Results Four SNPs and three genes associated with the prognosis of breast cancer in the training set were included in the prognostic model. Patients were divided into the high-risk group and low-risk group based on the four SNPs and three genes signature-based genetic prognostic index. High-risk patients showed a significant worse overall survival [Hazard Ratio (HR) 9.43, 95% confidence interval (CI) 3.81–23.33, $P < 0.001$] than the low-risk group. Compared to the model constructed with only gene expression, the C statistics for the signature-based genetic prognostic index [area under curves (AUC) = 0.79, 95% CI 0.72–0.86] showed a significant increase ($P < 0.001$). Additionally, we further replicated the prognostic prediction model in the validation set as patients in the high-risk group also showed a significantly worse overall survival (HR 4.55, 95% CI 1.50–13.88, $P < 0.001$), and the C statistics for the signature-based genetic prognostic index was 0.76 (95% CI 0.65–0.86). The following time-dependent ROC revealed that the mean of AUCs were 0.839 and 0.748 in the training set and the validation set, respectively.

Conclusions Our findings suggested that integrating genomic and transcriptomic profiles could greatly improve the predictive efficiency of the prognosis of breast cancer patients.

Keywords Breast cancer · Prognostic model · Genomic · Transcriptomic

Introduction

Breast cancer is the most diagnosed cancer and the leading cause of cancer death among females, accounting for 25% of all cancer cases and 15% of all cancer deaths in the world [1]. The 5-year survival rate of breast cancer ranged from

Chengxiao Yu and Na Qin contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10549-019-05177-0) contains supplementary material, which is available to authorized users.

✉ Tao Jiang
tao.chiang0923@njmu.edu.cn

✉ Yue Jiang
jiangyue@njmu.edu.cn

[1] Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China

[2] State Key Laboratory of Reproductive Medicine, Nanjing Medical University, Nanjing 211166, China

[3] Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center of Cancer Medicine, Nanjing Medical University, Nanjing 211166, China

[4] Department of Bioinformatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166, China

[5] Center of Clinical Research, Wuxi Institute of Translational Medicine, Wuxi People's Hospital of Nanjing Medical University, Wuxi 214000, China

30 to 90%, while early-stage patients have a higher 5-year survival rate [2]. It has been reported that some demographic and clinicopathological characteristics of patients, such as age and pathological stage, have an impact on the prognosis of breast cancer [2, 3]. However, these clinical features only partially explain the survival status of breast cancer, suggesting that other risk factors also contribute to the prognosis of breast cancer [4–6].

Traditional epidemiology studies have evaluated the prognostic factors for breast cancer, and established the prognostic effect of clinicopathological features, single-nucleotide polymorphisms (SNPs), and gene expression on the outcome of breast cancer [7–11]. Wang CQ et al. reported that carriers of G allele of rs3801004 were more likely to progress to stage III/IV and lymph node metastasis [12]. Phan NN et al. discovered that higher expression levels of *CDCA3, CDCA5*, and *CDCA8* in breast cancer patients dramatically reduced patient survival time [13]. Predictive models constructed with pathological stage can also predict the prognosis of breast cancer [14]. However, the predictive efficiency of these factors appears to be limited when only single variable was included [12, 13, 15–18]. As the development of breast cancer is a multistage process involving the dysregulation of many factors, and increasing evidence demonstrated that integrating multiomics data sets can be effective in exploring the development of diseases [19]. Thus, we hypothesized that integrating genomic and transcriptomic profiles may contribute the prediction of the prognosis of breast cancer.

In this study, we comprehensively evaluated the prognostic effect of SNPs and genes expression in 810 breast cancer subjects from the Cancer Genome Atlas (TCGA). We developed a prognostic predictive index for breast cancer in the training set and the effectiveness of this model was replicated in the validation sample set. Our findings suggested that the integration of genomic and transcriptomic data can improve the predictive efficiency of the prognosis of breast cancer.

Methods

Study populations

The relevant samples are based on the TCGA Breast Cancer (BRCA) data set. We selected 1098 subjects from the TCGA breast cancer cohort with baseline information (survival state, follow-up time, age, gender, and pathological stage) in the study. The baseline information of these samples was downloaded from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). The genotyping array data (1012 subjects) and gene expression quantification data (1058 subjects) were also obtained from the TCGA database. A total of 810 TCGA female breast cancer samples with both baseline information (follow-up time ≥ 30 d), RNA-Seq data, and Genotyping Array data available were included in the study. The 810 samples were randomly divided into the training set (540 subjects) and the validation set (270 subjects) at the ratio of 2:1. The demographic description of the samples is shown in Table 1. Tumors were classified by immunohistochemical staining (IHC) according to St. Gallen subtypes as follows: luminal A or luminal B HER2− (ER + and/or PR+, HER2−), luminal B HER2+ (ER + and/or PR + and HER2+), HER2-neu non-luminal (ER−, PR −, and HER2+), and basal-like (ER−, PR −, and HER2−) [20].

Quality control and imputation

The samples included in this study were genotyped with the Affymetrix Genome-Wide Human SNP Array 6.0. The raw data were retrieved from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). We further removed unqualified samples with (1) call rate < 95%; (2) gender discordance; (3) duplicates or probable relatives; (4) an extreme heterozygosity rate; or (5) outliers according to principal component analysis (PCA). For post-QC data, we phased haplotypes with Shapeit v2 (http://www.shapeit.fr/, Phasing step) and imputed using IMPUTE2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html, Imputation step). The 1000 Genomes Project (Phase III integrated variant set release, across 2504 samples) was set as the reference. Poorly imputed SNPs (Imputed info < 0.4) were excluded from the analysis.

Table 1 Clinical characteristics of the patients

	TCGA training set	TCGA validation set	*P*
No. of patients	540	270	(–)
Age, years, mean (SD)	58.01 (13.10)	58.49 (12.93)	0.625[a]
Median survivor follow-up, months	17.15	17.28	0.293[a]
Stage, no. (%)			1[b]
I	95 (17.8)	48 (18.0)	
II	302 (56.4)	150 (56.2)	
III	121 (22.6)	60 (22.5)	
IV	9 (1.7)	5 (1.9)	
V	8 (1.5)	4 (1.5)	

[a]Student's *t* test

[b]Fisher's exact test

Selection of survival-related SNPs

We first performed a post-imputation QC procedure, and removed SNPs with the following criteria: (1) imputation quality INFO < 0.9; (2) derivation from Hardy–Weinberg equilibrium ($P < 0.05$); and (3) minor allele frequency (MAF) < 0.05. A total of 4474,182 SNPs remained. Then we performed functional annotation, and 2448 SNPs with potential functions were retained: (1) Combined Annotation Dependent Depletion (CADD) score > 20; (2) located in the promoter and enhancer regions, or the DNaseI Hypersensitivity Clusters (DHS) regions. The functional annotation data were downloaded from the Encyclopedia of DNA Elements (ENCODE) project, including the DHS in 125 cell types. The definition of promoter and enhancer were retrieved from the Functional Annotation of The Mammalian Genome (FANTOM) project (http://fantom.gsc.riken.jp/). Univariate Cox proportional hazards regression model was further performed in the training set to evaluate the association of 2448 included SNPs with the overall survival of breast cancer. 107 significant SNPs ($P < 0.05$) were included in further analysis. Finally, we utilized the Least Absolute Shrinkage and Selection Operator (LASSO) penalized Cox regression [21, 22] to define the promising prognosis-related SNPs and four SNPs remained (Supplementary Fig. 1).

Selection of survival-related genes

A total of 19,265 unique genes with expression abundance data (normalized read counts) from TCGA were included in the analysis, and the data were retrieved from the Firehose (http://gdac.broadinstitute.org/runs/stddata__2014_07_15/data/BRCA/20140715/). 459 mutational cancer driver genes from the IntOGen-mutations platform (https://www.intogen.org/search) and 36 breast cancer-related genes extracted from recent literature reports [23–29] were included. We first filtered out genes with normalized read counts (RC) greater than or equal to 5 in less than half of the ($1058 \times 1/2 = 529$) breast cancer samples. Of the above 495 genes, only 482 genes were sufficiently expressed in the samples. Univariate Cox proportional hazards regression was used to screen candidate genes ($P < 0.05$), and 61 candidate genes were included in the following analysis. Then, LASSO penalized Cox regression [21, 22] was used to define the promising prognostic-related genes and three genes remained (Supplementary Fig. 1).

The construction and validation of the prognostic prediction model

We created the prognostic index to predict the survival status of breast cancer by summing the expression of three survival-related genes, the genotype of four survival-related SNPs, and traditional prognostic-related factors with the following formula:

$$\text{Total prognostic index} = \sum_{i=1}^{3} \beta_i \text{Gene}_i + \sum_{j=1}^{4} \beta_j \text{SNP}_j + \beta_{\text{age}} \text{Age} + \beta_{\text{stage}} \text{Stage},$$

$$\text{Genetic prognostic index} = \sum_{i=1}^{3} \beta_i \text{Gene}_i + \sum_{j=1}^{4} \beta_j \text{SNP}_j,$$

where β_i is the estimated regression coefficient (beta) of the ith Gene; Gene_i is the expression of the gene; β_j is the estimated regression coefficient (beta) of the jth SNP; SNP_j is the dosage of the SNP (coded as 0, 1 or 2 for wild-type homozygous, heterozygous, or homozygous); β_{age} is the estimated regression coefficient (beta) of the age; and β_{stage} is the estimated regression coefficient (beta) of the clinical stage.

Statistical analyses

Fisher's exact test was used to evaluate the difference of categorical variables. Student's t test was used for continuous variables if equal-variance was assumed. Hazard Ratios (HRs) and 95% confidence intervals (CIs) of Cox regression model Cox regression model were used to evaluate the prognosis of breast cancer with adjustment for age (as continuous covariable) and clinical stage (as categorical covariable). LASSO regression, a shrinkage and variable selection method for regression models [21, 22, 30], was used to reduce dimensions and select the final prediction factors. All the statistical analyses were two-sided with 0.05 as the significant level and were performed with R software (Version 3.3.2). Kaplan–Meier analysis with log-rank test for difference was performed in GraphPad Prism 7. Heatmap and scatter diagram were generated in Microsoft Excel 2013 with the size of the order genetic prognostic index (gene&snp) as a horizontal coordinate, coded as 0, 1, or 2 for the gene expression value trisection from low to high. The Receiver Operating Characteristic (ROC) and C statistics were used to evaluate the discrimination of prediction models with "pROC" R package. Time-dependent ROC was calculated at different time nodes using "survivalROC" R package. All these ROC analyses were conducted using R software (version 3.3.2).

Results

General description of the study population

The flow chart was shown in Supplementary Fig. 1. 810 breast cancer samples were randomly divided into the

training set (540 subjects) and validation set (270 subjects). The demographics of these sets were well balanced (Table 1). The distribution of age was comparable in both two sample sets (training set = 58.01, validation set = 58.49; $P = 0.625$), as well as the median survival time (training set = 17.15 months, validation set = 17.28 months; $P = 0.293$).

Construction of the prognostic index

In the training set, 109 SNPs and 61 genes were statistically significantly ($P < 0.05$) associated with the prognosis of breast cancer. Four SNPs and three genes were selected with the systematically dimensional reduction strategy, including the protein-coding genes *PGR, ROBO2,* and *WNT5A*; and SNPs rs6568703, rs6669563, rs11630197, and rs7700810. We constructed the genetic prognostic index with the regression coefficients derived from the Cox regression model, and divided 540 samples into high- and low-risk groups based on the median of genetic prognostic index (Fig. 1). As shown in Fig. 1a, six patients died in the low-risk group and 33 patients died in the high-risk group ($P < 0.001$). The three prognostic genes were significantly down-regulated in the high-risk group compared with the low-risk group

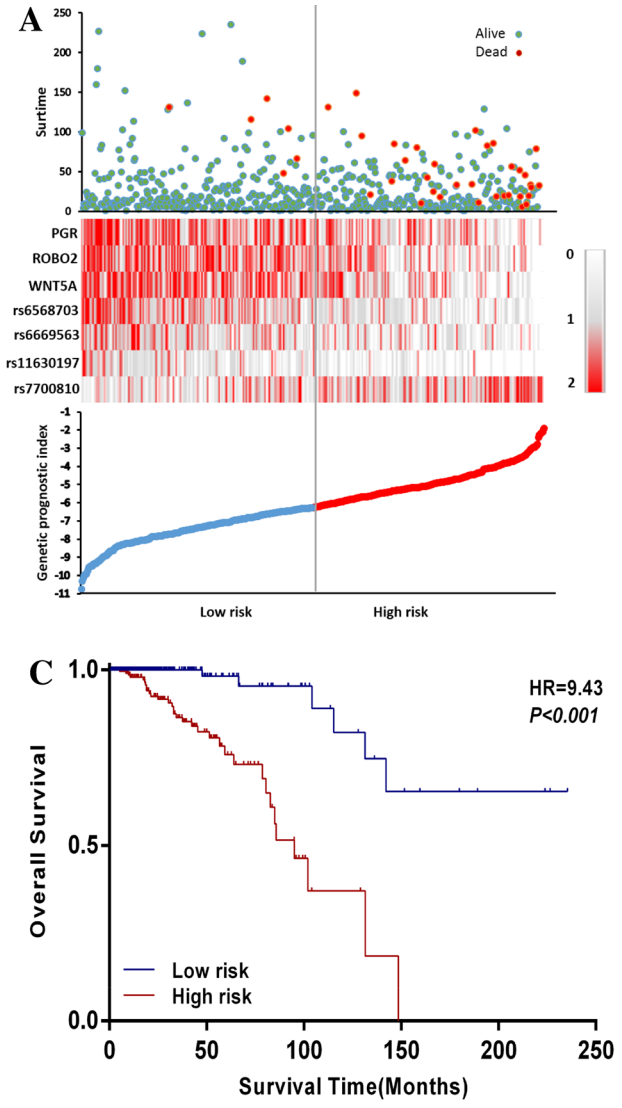

Fig. 1 Four-SNP and three-gene prognostic signature biomarker characteristics in the training set. **a** The genetic prognostic index for all patients in the training set is plotted in ascending order and marked as low risk (blue) or high risk (red), as divided by the threshold (vertical gray line). The SNPs were coded as 0, 1, or 2 for wild-type homozygous, heterozygous, or homozygous. Coded as 0, 1, or 2 for the signature genes expression value from low to high. **b** The three signature genes were statistically significantly under expressed in the high-risk group compared with the low-risk group. ***$P < 0.001$, *LR* low risk, and *HR* high risk. **c** Kaplan–Meier curves of overall survival in the training set stratified by four-SNP and three-gene prognostic signature in high and low risk

($P < 0.001$) (Fig. 1b). Patients in the high-risk group had a significantly worse overall survival (HR 9.43, 95% CI 3.81–23.33, $P < 0.001$) compared to the low-risk group in the training set (Fig. 1c). The prognostic signature risk prediction statistically significant in the training set for overall survival (HR 2.11, 95% CI 1.66–2.67, $P < 0.001$) (Table 2). The HR for breast cancer was examined by genetic prognostic index in different subgroups (Supplementary Tables 1, 3).

Validation of the prognostic index

We further tested the prognostic index constructed with four SNPs and three genes in the validation set. With the same prognostic index threshold (the median of genetic prognostic index) used in the training set, 4 patients died in the low-risk group and 14 patients died in the high-risk group ($P < 0.001$) (Supplementary Fig. 2a). As shown in Supplementary Fig. 2b, patients in the high-risk group had a significantly worse overall survival compared to those in the low-risk group (HR 4.55, 95% CI 1.50–13.88, $P < 0.001$). The prognostic signature risk prediction statistically significantly in the validation set for overall survival (HR 2.36, 95% CI 1.56–3.57, $P < 0.001$) (Table 3; Supplementary Tables 2, 3).

Evaluation of the prognostic index

To further evaluate the accuracy of the prognostic model, we evaluated the area under curves (AUCs) by C-statistic in the training set. The AUC was 0.67(95% CI 0.59–0.75) when only three genes were included, and 0.76(95% CI 0.68–0.83) when only four SNPs were included, and 0.79 (95% CI 0.72–0.86) when combining the DNA level data and RNA level data together. Additionally, the AUC greatly increased (AUC = 0.83, 95% CI 0.77–0.89) when the clinical information (age and pathological stage) was further added (Fig. 2). Similar results were achieved in the validation set (3 genes, AUC = 0.71, 95% CI 0.58–0.84; 4 SNPs, AUC = 0.67, 95% CI 0.54–0.80; 3 genes & 4 SNPs, AUC = 0.76, 95% CI 0.65–0.86; 3 genes & 4 SNPs & clinical information: AUC = 0.82, 95% CI 0.71–0.93) (Supplementary Fig. 3). Time-dependent Receiver Operating Characteristic (ROC) was calculated to evaluate the discrimination ability of the genetic predictive variables at different time nodes (12, 24, 36, 48, 60, 72, 84, 96, 108, 120, 132, 144, 156, 168, 180, 192, 204, 216, 228, 240 months) (Fig. 3); the mean of AUCs (standard deviation, SD) were 0.839 (0.025) and 0.748 (0.052) for the training set and the validation set, respectively.

Table 2 Cox proportional hazards models in the TCGA training set

Factor	Univariate		Multivariable	
	HR (95% CI)[a]	P^a	HR (95% CI)[b]	P^b
Stage	2.32 (1.75, 3.07)	4.82E−09	1.84 (1.40, 2.41)	1.08E−05
Age	1.02 (0.99, 1.04)	1.25E−01	1.01 (0.99, 1.04)	2.39E−01
Genetic prognostic index	2.21 (1.77, 2.76)	3.45E−12	2.11 (1.66, 2.67)	7.65E−10

CI confidence interval, *HR* hazard ratio

[a]Results in univariable cox proportional hazard regression model

[b]Age, stage, and genetic prognostic index were adjusted

Table 3 Cox proportional hazards models in the TCGA validation set

Factor	Univariate		Multivariable	
	HR (95% CI)[a]	P^a	HR (95% CI)[b]	P^b
Stage	1.72 (1.17, 2.53)	6.20E−03	1.72 (1.10, 2.68)	1.68E−02
Age	1.04 (0.99, 1.08)	9.35E−02	1.04 (0.99, 1.09)	8.41E−02
Genetic prognostic index	2.16 (1.50, 3.12)	3.30E−05	2.36 (1.56, 3.57)	4.94E−05

[a]Results in univariable cox proportional hazard regression model

[b]Age, stage, and genetic prognostic index were adjusted

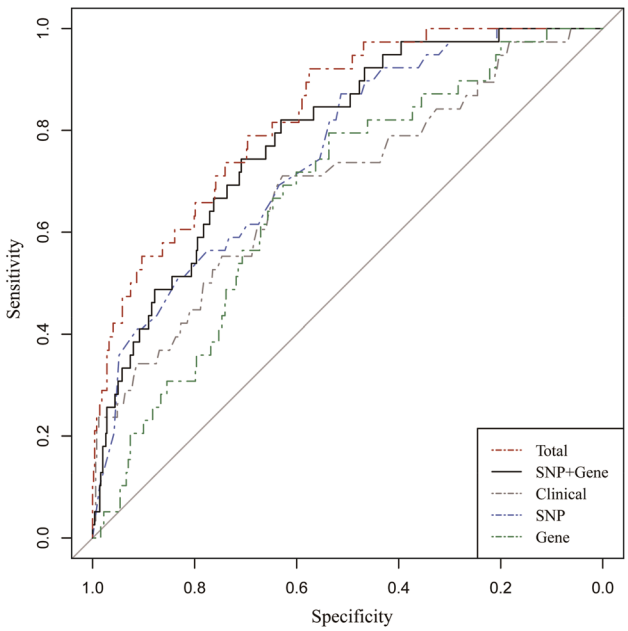

Fig. 2 Receiver operating characteristic (ROC) was created at different variable groups of genomic, transcriptomic, genomic and transcriptomic, clinical variables (age, pathological staging) in the training set

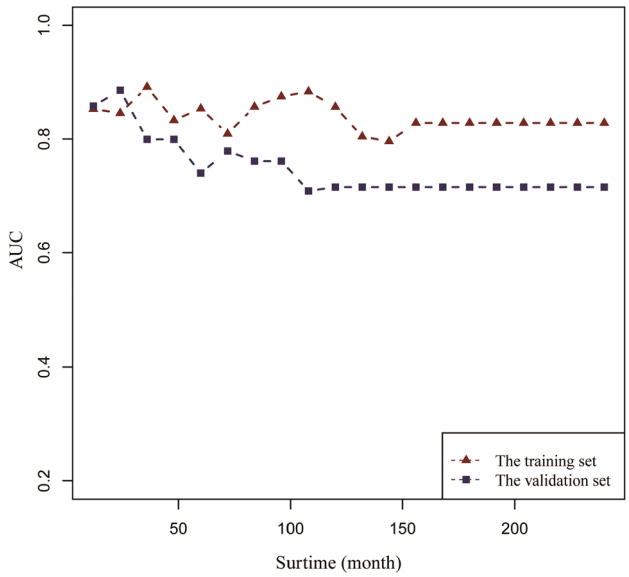

Fig. 3 The time-dependent ROC evaluates the discrimination ability of the genetic predictive variables at different time nodes (12, 24, 36, 48, 60, 72, 84, 96, 108, 120, 132, 144, 156, 168, 180, 192, 204, 216, 228, 240 months)

Discussion

Breast cancer is a major cause of cancer mortality in women worldwide, leading to 521,900 deaths per year [1]. Recently, some studies have found that a better prediction model can be constructed if we combined genetic variables into traditional approaches [31–33]. For example, Weissfeld et al. constructed a lung cancer risk prediction model and found that the AUC enlarged from 0.717 to 0.725 when adding GWAS susceptibility regions to an age and smoking risk factor-only model [33]. However, most of these studies only utilized the genotype information to construct prognostic prediction models, and the prediction performance was limited. In order to improve the predictive efficiency of breast cancer prognosis, we conducted the prognostic assessment model for breast cancer patients with both genotype and expression data. Three genes (*PGR, ROBO2*, and *WNT5A*) and four SNPs (rs6568703, rs6669563, rs11630197, and rs7700810) were integrated in our breast cancer prognosis model and the predictive efficiency was replicated in an independent sample set, suggesting that integrating genomic and transcriptomic profiles can provide more reliable information for breast cancer prognosis study.

In the TCGA BRCA training set, we detected three genes associated with the prognostic of breast cancer, and the higher expression of *PGR* (HR 0.91, 95% CI 0.84–0.99, $P=0.04$), *ROBO2* (HR 0.84, 95% CI 0.74–0.97, $P=0.016$), and *WNT5A* (HR 0.70, 95% CI 0.57–0.84, $P<0.001$) can decrease the mortality risk (Supplementary Fig. 4a). *PGR* encodes a member of the steroid receptor superfamily and the encoded protein mediates the physiological effects of progesterone, which plays a central role in reproductive events associated with the establishment and maintenance of pregnancy. Du X et al. measured *ESR1, PGR*, and *ERBB2* mRNA levels in 294 breast cancer patients, and their results suggested that the detection of *ESR1*/*PGR*/*ERBB2* mRNA levels can serve as a better approach for predicting the prognosis [34]. The protein encoded by *ROBO2* belongs to the *ROBO* family, part of the immunoglobulin superfamily. The encoded protein is a transmembrane receptor for the slit homolog 2 protein and functions in axon guidance and cell migration. Multiple lines of evidence indicate that axon guidance genes are involved in cancer development. Crucial regulators of axon guidance *ROBO1* and *ROBO2* are considered as potential tumor suppressor genes [35]. Some studies reported that the expression of *ROBO2* was associated with cancer development, such as pancreatic cancer, prostate cancer, gastric cancer, and colorectal cancer [35–37]. The *WNT* gene family consists of structurally related genes which encode secreted signaling proteins. The protein encoded by *WNT5A* has been implicated in oncogenesis and several developmental processes, including regulation of cell fate and patterning during embryogenesis. *WNT5A* is a representative ligand that activates the β-catenin-independent pathway in the Wnt signaling which can stimulate cell migration by regulating focal adhesion complexes. Additionally, *WNT5A* is also associated with the cancer migration, such as breast cancer, lung cancer, and gastric cancer [38–41].

The A allele of rs6568703 (HR 0.35, 95% CI 0.20–0.61, $P<0.001$), the A allele of rs6669563 (HR 0.44, 95% CI 0.26–0.74, $P=0.002$), the A allele of rs11630197 (HR 0.33, 95% CI 0.17–0.63, $P<0.001$), and the C allele of rs7700810 (HR 1.89, 95% CI 1.19–2.98, $P=0.007$) showed significant effects on the progress of breast cancer in the training sets (Supplementary Fig. 4b). SNP rs6669563 was located in the intron of *SPOCD1* gene at 1p35.2, and was predicted to be deleterious by SIFT (Score: 0.03). SNP rs11630197 was located in the exon of *SAXO2* gene at 15q25.2, and was predicted to be deleterious by SIFT (Score: 0) and probably damaging by PolyPhen (Score: 0.998).

In our study, we constructed BRCA prognosis models with RNA and DNA level predictors after systematic screening for the prognosis predictors using cox-hazard ratio model and LASSO regression. In training set, the AUC was only 0.67 for the RNA level predictors, and 0.76 for the DNA level predictors. When combining the DNA level data and RNA level data, the AUC significantly increased to 0.79 ($P<0.001$) compared with the model containing only RNA level predictors. The result was further replicated in the validation set. The following time-dependent ROC further proved the good prediction ability of our final model. There are also some models that predict the prognosis of breast cancer with pathological stage [14] or gene expression

data [42], etc. But the predictive efficiency of these factors appears to be limited when only single variable was included. For example, Wang L et al. constructed prognostic assessment model for breast cancer which contained only gene expression factor with an AUC 0.75 [43], much lower than our prediction model.

As we know, this study has some meaningful strengths. First, we included both genomic and transcriptomic data to construct the prognostic risk model for breast cancer, which greatly increased the prediction efficiency. Additionally, the stability of the prognostic model was further replicated in the validation set. However, as the samples included in this study are almost Europeans, whether the model can be applicable in other populations is still unknown. Further external validation and prospective cohort study is needed to evaluate the extensive ability of our model.

Conclusions

Generally, this is a meaningful attempt on the construction of prognostic model of breast cancer using genomic and transcriptomic data, and the results showed good prediction ability. Further independent dataset and population investigation are warranted to evaluate the predictive efficiency of our model.

Acknowledgements We thank the study participants and research staff for their contributions and commitment to this study.

Author contributions YJ and TJ conceived the project; CXY analyzed the data and drafted the manuscript; NQ modify the manuscript; ZNP, CS, and CW contributed to the interpretation of the results; JPC reviewed the manuscript; JCD and HXM supervised the research. All authors read and approved the final manuscript.

Funding This work was supported by Science Fund for Creative Research Groups of the National Natural Science Foundation of China (81521004), Cheung Kong Scholars Programme of China, the Priority Academic Program for the Development of Jiangsu Higher Education Institutions (Public Health and Preventive Medicine), and Top-notch Academic Programs Project of Jiangsu Higher Education Institutions (PPZY2015A067).

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A (2015) Global cancer statistics, 2012. CA Cancer J Clin 65(2):87–108. https://doi.org/10.3322/caac.21262
2. DeSantis CE, Ma J, Goding Sauer A, Newman LA, Jemal A (2017) Breast cancer statistics, 2017, racial disparity in mortality by state. CA Cancer J Clin 67(6):439–448. https://doi.org/10.3322/caac.21412
3. Terranova CO, Protani MM, Reeves MM (2018) Overall dietary intake and prognosis after breast cancer: a systematic review. Nutr Cancer 70(2):153–163. https://doi.org/10.1080/01635581.2018.1412478
4. Azzato EM, Tyrer J, Fasching PA, Beckmann MW, Ekici AB, Schulz-Wendtland R, Bojesen SE, Nordestgaard BG, Flyger H, Milne RL, Arias JI, Menendez P, Benitez J, Chang-Claude J, Hein R, Wang-Gohrke S, Nevanlinna H, Heikkinen T, Aittomaki K, Blomqvist C, Margolin S, Mannermaa A, Kosma VM, Kataja V, Kathleen Cuningham Foundation Consortium for Research into Familial Breast C, Beesley J, Chen X, Chenevix-Trench G, Couch FJ, Olson JE, Fredericksen ZS, Wang X, Giles GG, Severi G, Baglietto L, Southey MC, Devilee P, Tollenaar RA, Seynaeve C, Garcia-Closas M, Lissowska J, Sherman ME, Bolton KL, Hall P, Czene K, Cox A, Brock IW, Elliott GC, Reed MW, Greenberg D, Anton-Culver H, Ziogas A, Humphreys M, Easton DF, Caporaso NE, Pharoah PD (2010) Association between a germline OCA2 polymorphism at chromosome 15q13.1 and estrogen receptor-negative breast cancer survival. J Natl Cancer Inst 102(9):650–662. https://doi.org/10.1093/jnci/djq057
5. Wu C, Xu B, Yuan P, Miao X, Liu Y, Guan Y, Yu D, Xu J, Zhang T, Shen H, Wu T, Lin D (2010) Genome-wide interrogation identifies YAP1 variants associated with survival of small-cell lung cancer patients. Cancer Res 70(23):9721–9729. https://doi.org/10.1158/0008-5472.CAN-10-1493
6. Wu X, Ye Y, Rosell R, Amos CI, Stewart DJ, Hildebrandt MA, Roth JA, Minna JD, Gu J, Lin J, Buch SC, Nukui T, Ramirez Serrano JL, Taron M, Cassidy A, Lu C, Chang JY, Lippman SM, Hong WK, Spitz MR, Romkes M, Yang P (2011) Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy. J Natl Cancer Inst 103(10):817–825. https://doi.org/10.1093/jnci/djr075
7. Kadalayil L, Khan S, Nevanlinna H, Fasching PA, Couch FJ, Hopper JL, Liu J, Maishman T, Durcan L, Gerty S, Blomqvist C, Rack B, Janni W, Collins A, Eccles D, Tapper W (2017) Germline variation in ADAMTSL1 is associated with prognosis following breast cancer treatment in young women. Nat Commun 8(1):1632. https://doi.org/10.1038/s41467-017-01775-y
8. Qiu Q, Su Y, Zheng Y, Cai H, Wu S, Lu W, Zheng W, Shu XO, Cai Q (2015) Increased pSmad2 expression and cytoplasmic predominant presence of TGF-betaRII in breast cancer tissue are associated with poor prognosis: results from the Shanghai Breast Cancer Study. Breast Cancer Res Treat 149(2):467–477. https://doi.org/10.1007/s10549-014-3251-9
9. Saunderson EA, Stepper P, Gomm JJ, Hoa L, Morgan A, Allen MD, Jones JL, Gribben JG, Jurkowski TP, Ficz G (2017) Hit-and-run epigenetic editing prevents senescence entry in primary breast cells from healthy donors. Nat Commun 8(1):1450. https://doi.org/10.1038/s41467-017-01078-2
10. Thakur SS, Li H, Chan AMY, Tudor R, Bigras G, Morris D, Enwere EK, Yang H (2018) The use of automated Ki67 analysis to predict Oncotype DX risk-of-recurrence categories in early-stage breast cancer. PLoS ONE 13(1):e0188983. https://doi.org/10.1371/journal.pone.0188983
11. Narayan HK, Finkelman B, French B, Plappert T, Hyman D, Smith AM, Margulies KB, Ky B (2017) Detailed echocardiographic phenotyping in breast cancer patients: associations with ejection fraction decline, recovery, and heart failure symptoms over 3 years of follow-up. Circulation 135(15):1397–1412. https://doi.org/10.1161/CIRCULATIONAHA.116.023463
12. Wang CQ, Tang CH, Wang Y, Jin L, Wang Q, Li X, Hu GN, Huang BF, Zhao YM, Su CM (2017) FSCN1 gene polymorphisms: biomarkers for the development and progression of breast

cancer. Sci Rep 7(1):15887. https://doi.org/10.1038/s41598-017-16196-6
13. Phan NN, Wang CY, Li KL, Chen CF, Chiao CC, Yu HG, Huang PL, Lin YC (2018) Distinct expression of CDCA3, CDCA5, and CDCA8 leads to shorter relapse free survival in breast cancer patient. Oncotarget 9(6):6977–6992. https://doi.org/10.18632/oncotarget.24059
14. Lai J, Wang H, Peng J, Chen P, Pan Z (2018) Establishment and external validation of a prognostic model for predicting disease-free survival and risk stratification in breast cancer patients treated with neoadjuvant chemotherapy. Cancer Manage Res 10:2347–2356. https://doi.org/10.2147/CMAR.S171129
15. Curtit E, Pivot X, Henriques J, Paget-Bailly S, Fumoleau P, Rios M, Bonnefoi H, Bachelot T, Soulie P, Jouannaud C, Bourgeois H, Petit T, Tennevet I, Assouline D, Mathieu MC, Jacquin JP, Lavau-Denes S, Darut-Jouve A, Ferrero JM, Tarpin C, Levy C, Delecroix V, Trillet-Lenoir V, Cojocarasu O, Meunier J, Pierga JY, Kerbrat P, Faure-Mercier C, Blanche H, Sahbatou M, Boland A, Bacq D, Besse C, Thomas G, Deleuze JF, Pauporte I, Romieu G, Cox DG (2017) Assessment of the prognostic role of a 94-single nucleotide polymorphisms risk score in early breast cancer in the SIGNAL/PHARE prospective cohort: no correlation with clinico-pathological characteristics and outcomes. Breast Cancer Res 19(1):98. https://doi.org/10.1186/s13058-017-0888-4
16. Lan B, Ma F, Zhai X, Li Q, Chen S, Wang J, Fan Y, Luo Y, Cai R, Yuan P, Zhang P, Li Q, Xu B (2018) The relationship between the CYP2D6 polymorphisms and tamoxifen efficacy in adjuvant endocrine therapy of breast cancer patients in Chinese Han population. Int J Cancer 143(1):184–189. https://doi.org/10.1002/ijc.31291
17. Rudolph M, Sizemore ST, Lu Y, Teng KY, Basree MM, Reinbolt R, Timmers CD, Leone G, Ostrowski MC, Majumder S, Ramaswamy B (2018) A hedgehog pathway-dependent gene signature is associated with poor clinical outcomes in Luminal A breast cancer. Breast Cancer Res Treat 169(3):457–467. https://doi.org/10.1007/s10549-018-4718-x
18. Gibbs LD, Vishwanatha JK (2018) Prognostic impact of AnxA1 and AnxA2 gene expression in triple-negative breast cancer. Oncotarget 9(2):2697–2704. https://doi.org/10.18632/oncotarget.23627
19. Hasin Y, Seldin M, Lusis A (2017) Multi-omics approaches to disease. Genome Biol 18(1):83. https://doi.org/10.1186/s13059-017-1215-1
20. Vasconcelos I, Hussainzada A, Berger S, Fietze E, Linke J, Siedentopf F, Schoenegg W (2016) The St. Gallen surrogate classification for breast cancer subtypes successfully predicts tumor presenting features, nodal involvement, recurrence patterns and disease free survival. Breast 29:181–185. https://doi.org/10.1016/j.breast.2016.07.016
21. Pineda S, Real FX, Kogevinas M, Carrato A, Chanock SJ, Malats N, Van Steen K (2015) Integration analysis of three omics data using penalized regression methods: an application to bladder cancer. PLoS Genet 11(12):e1005689. https://doi.org/10.1371/journal.pgen.1005689
22. Lu Y, Zhou Y, Qu W, Deng M, Zhang C (2011) A Lasso regression model for the construction of microRNA-target regulatory networks. Bioinformatics 27(17):2406–2413. https://doi.org/10.1093/bioinformatics/btr410
23. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC, Van Loo P, Ju YS, Smid M, Brinkman AB, Morganella S, Aure MR, Lingjaerde OC, Langerod A, Ringner M, Ahn SM, Boyault S, Brock JE, Broeks A, Butler A, Desmedt C, Dirix L, Dronov S, Fatima A, Foekens JA, Gerstung M, Hooijer GK, Jang SJ, Jones DR, Kim HY, King TA, Krishnamurthy S, Lee HJ, Lee JY, Li Y, McLaren S, Menzies A, Mustonen V, O'Meara S, Pauporte I, Pivot X, Purdie CA, Raine K, Ramakrishnan K, Rodriguez-Gonzalez FG, Romieu G, Sieuwerts AM, Simpson PT, Shepherd R, Stebbings L, Stefansson OA, Teague J, Tommasi S, Treilleux I, Van den Eynden GG, Vermeulen P, Vincent-Salomon A, Yates L, Caldas C, van't Veer L, Tutt A, Knappskog S, Tan BK, Jonkers J, Borg A, Ueno NT, Sotiriou C, Viari A, Futreal PA, Campbell PJ, Span PN, Van Laere S, Lakhani SR, Eyfjord JE, Thompson AM, Birney E, Stunnenberg HG, van de Vijver MJ, Martens JW, Borresen-Dale AL, Richardson AL, Kong G, Thomas G, Stratton MR (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534(7605):47–54. https://doi.org/10.1038/nature17676
24. Polak P, Kim J, Braunstein LZ, Karlic R, Haradhavala NJ, Tiao G, Rosebrock D, Livitz D, Kubler K, Mouw KW, Kamburov A, Maruvka YE, Leshchiner I, Lander ES, Golub TR, Zick A, Orthwein A, Lawrence MS, Batra RN, Caldas C, Haber DA, Laird PW, Shen H, Ellisen LW, D'Andrea AD, Chanock SJ, Foulkes WD, Getz G (2017) A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nat Genet 49(10):1476–1486. https://doi.org/10.1038/ng.3934
25. Lefebvre C, Bachelot T, Filleron T, Pedrero M, Campone M, Soria JC, Massard C, Levy C, Arnedos M, Lacroix-Triki M, Garrabey J, Boursin Y, Deloger M, Fu Y, Commo F, Scott V, Lacroix L, Dieci MV, Kamal M, Dieras V, Goncalves A, Ferrerro JM, Romieu G, Vanlemmens L, Mouret Reynier MA, Thery JC, Le Du F, Guiu S, Dalenc F, Clapisson G, Bonnefoi H, Jimenez M, Le Tourneau C, Andre F (2016) Mutational profile of metastatic breast cancers: a retrospective analysis. PLoS Med 13(12):e1002201. https://doi.org/10.1371/journal.pmed.1002201
26. Jiang T, Shi W, Wali VB, Pongor LS, Li C, Lau R, Gyorffy B, Lifton RP, Symmans WF, Pusztai L, Hatzis C (2016) Predictors of chemosensitivity in triple negative breast cancer: an integrated genomic analysis. PLoS Med 13(12):e1002193. https://doi.org/10.1371/journal.pmed.1002193
27. Goodarzi H, Liu X, Nguyen HC, Zhang S, Fish L, Tavazoie SF (2015) Endogenous tRNA-derived fragments suppress breast cancer progression via YBX1 displacement. Cell 161(4):790–802. https://doi.org/10.1016/j.cell.2015.02.053
28. Dethlefsen C, Hansen LS, Lillelund C, Andersen C, Gehl J, Christensen JF, Pedersen BK, Hojman P (2017) Exercise-induced catecholamines activate the hippo tumor suppressor pathway to reduce risks of breast cancer development. Cancer Res 77(18):4894–4904. https://doi.org/10.1158/0008-5472.CAN-16-3125
29. Surveying Breast Cancer's Genomic Landscape (2016) Cancer discovery 6(7):OF2. https://doi.org/10.1158/2159-8290.CD-NB2016-064
30. Ranstam J, Cook JA (2018) LASSO regression. Br J Surg 105(10):1348. https://doi.org/10.1002/bjs.10895
31. Li H, Yang L, Zhao X, Wang J, Qian J, Chen H, Fan W, Liu H, Jin L, Wang W, Lu D (2012) Prediction of lung cancer risk in a Chinese population using a multifactorial genetic model. BMC Med Genet 13:118. https://doi.org/10.1186/1471-2350-13-118
32. Spitz MR, Amos CI, Land S, Wu X, Dong Q, Wenzlaff AS, Schwartz AG (2013) Role of selected genetic variants in lung cancer risk in African Americans. J Thorac Oncol 8(4):391–397. https://doi.org/10.1097/JTO.0b013e318283da29
33. Weissfeld JL, Lin Y, Lin HM, Kurland BF, Wilson DO, Fuhrman CR, Pennathur A, Romkes M, Nukui T, Yuan JM, Siegfried JM, Diergaarde B (2015) Lung cancer risk prediction using common SNPs located in GWAS-identified susceptibility regions. J Thorac Oncol 10(11):1538–1545. https://doi.org/10.1097/JTO.0000000000000666
34. Du X, Li XQ, Li L, Xu YY, Feng YM (2013) The detection of ESR1/PGR/ERBB2 mRNA levels by RT-QPCR: a better approach for subtyping breast cancer and predicting prognosis. Breast

Cancer Res Treat 138(1):59–67. https://doi.org/10.1007/s10549-013-2432-2
35. Je EM, Gwak M, Oh H, Choi MR, Choi YJ, Lee SH, Yoo NJ (2013) Frameshift mutations of axon guidance genes ROBO1 and ROBO2 in gastric and colorectal cancers with microsatellite instability. Pathology 45(7):645–650. https://doi.org/10.1097/PAT.0000000000000007
36. Waddell N, Pajic M, Patch AM, Chang DK, Kassahn KS, Bailey P, Johns AL, Miller D, Nones K, Quek K, Quinn MC, Robertson AJ, Fadlullah MZ, Bruxner TJ, Christ AN, Harliwong I, Idrisoglu S, Manning S, Nourse C, Nourbakhsh E, Wani S, Wilson PJ, Markham E, Cloonan N, Anderson MJ, Fink JL, Holmes O, Kazakoff SH, Leonard C, Newell F, Poudel B, Song S, Taylor D, Waddell N, Wood S, Xu Q, Wu J, Pinese M, Cowley MJ, Lee HC, Jones MD, Nagrial AM, Humphris J, Chantrill LA, Chin V, Steinmann AM, Mawson A, Humphrey ES, Colvin EK, Chou A, Scarlett CJ, Pinho AV, Giry-Laterriere M, Rooman I, Samra JS, Kench JG, Pettitt JA, Merrett ND, Toon C, Epari K, Nguyen NQ, Barbour A, Zeps N, Jamieson NB, Graham JS, Niclou SP, Bjerkvig R, Grutzmann R, Aust D, Hruban RH, Maitra A, Iacobuzio-Donahue CA, Wolfgang CL, Morgan RA, Lawlor RT, Corbo V, Bassi C, Falconi M, Zamboni G, Tortora G, Tempero MA, Australian Pancreatic Cancer Genome I, Gill AJ, Eshleman JR, Pilarsky C, Scarpa A, Musgrove EA, Pearson JV, Biankin AV, Grimmond SM (2015) Whole genomes redefine the mutational landscape of pancreatic cancer. Nature 518(7540):495–501. https://doi.org/10.1038/nature14169
37. Choi YJ, Yoo NJ, Lee SH (2014) Down-regulation of ROBO2 expression in prostate cancers. Pathol Oncol Research 20(3):517–519. https://doi.org/10.1007/s12253-013-9722-1
38. Prasad CP, Sodergren K, Andersson T (2017) Reduced production and uptake of lactate are essential for the ability of WNT5A signaling to inhibit breast cancer cell migration and invasion. Oncotarget 8(42):71471–71488. https://doi.org/10.18632/oncotarget.17277
39. Han B, Zhou B, Qu Y, Gao B, Xu Y, Chung S, Tanaka H, Yang W, Giuliano AE, Cui X (2018) FOXC1-induced non-canonical WNT5A-MMP7 signaling regulates invasiveness in triple-negative breast cancer. Oncogene 37(10):1399–1408. https://doi.org/10.1038/s41388-017-0021-2
40. Wang B, Tang Z, Gong H, Zhu L, Liu X (2017) Wnt5a promotes epithelial-to-mesenchymal transition and metastasis in non-small-cell lung cancer. Biosci Rep 37 (6). https://doi.org/10.1042/BSR20171092
41. Kurayoshi M, Oue N, Yamamoto H, Kishida M, Inoue A, Asahara T, Yasui W, Kikuchi A (2006) Expression of Wnt-5a is correlated with aggressiveness of gastric cancer by stimulating cell migration and invasion. Cancer Res 66(21):10439–10448. https://doi.org/10.1158/0008-5472.CAN-06-2359
42. Huang S, Yee C, Ching T, Yu H, Garmire LX (2014) A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. PLoS Comput Biol 10(9):e1003851. https://doi.org/10.1371/journal.pcbi.1003851
43. Wang L, Yao L, Zheng YZ, Xu Q, Liu XP, Hu X, Wang P, Shao ZM (2015) Expression of autophagy-related proteins ATG5 and FIP200 predicts favorable disease-free survival in patients with breast cancer. Biochem Biophys Res Commun 458(4):816–822. https://doi.org/10.1016/j.bbrc.2015.02.037

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