The Construction and Analysis of a Novel Four-Long Non-Coding RNA Signature Predicting Survival of Breast Cancer Patients

Chengdong Qin
Cancer Hospital of the University of Chinese Academy of Sciences（Zhejiang Cancer Hospital）
Weiliang Feng
Cancer Hospital of the University of Chinese Academy of Sciences（Zhejiang Cancer Hospital）
Xingfei Yu
Cancer Hospital of the University of Chinese Academy of Sciences（Zhejiang Cancer Hospital）
Chenlu Liang
Cancer Hospital of the University of Chinese Academy of Sciences（Zhejiang Cancer Hospital）
Yuqin Ding
Cancer Hospital of the University of Chinese Academy of Sciences（Zhejiang Cancer Hospital）
Haojun Xuan
Cancer Hospital of the University of Chinese Academy of Sciences（Zhejiang Cancer Hospital）
Jiejie Hu
Cancer Hospital of the University of Chinese Academy of Sciences（Zhejiang Cancer Hospital）
Yang Hongjian
Cancer Hospital of the University of Chinese Academy of Sciences（Zhejiang Cancer Hospital）
Yang Yu（yuyangkaiyu@163.com）
Cancer Hospital of the University of Chinese Academy of Sciences（Zhejiang Cancer Hospital）

Research Article

Keywords: Breast cancer, long non-coding RNA, bioinformatics, prognosis, signature

Posted Date: February 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-196427/v1

Abstract

Background

As the critical regulators for tumorigenesis and progression, long-noncoding RNAs (lncRNAs) are becoming novel prognostic biomarkers for tumor patients. By the levels of lncRNAs expression, the patients with breast carcinoma may be divided into subgroups with different risk scores. Nevertheless, there is limited evidence to evaluate the role of lncRNAs in the prognosis of breast carcinoma. The present study aimed to construct lncRNA signatures for prognostic analysis and assist clinicians in choosing optimal therapies.

Methods

Abnormal expression profiles of breast cancer-associated lncRNAs were analyzed based on the TCGA datasets. Univariate and multivariate Cox regression analysis was used to build a prognostic risk signature according to the lncRNAs expression. The prognostic ability of this signature was verified in various subgroups. Functional enrichment analysis was employed to reveal the potential roles of these predictive lncRNAs in cancer-related biological processes and pathways.

Results

Compared with normal breast tissues, the differential analysis demonstrated that 286 lncRNAs were abnormally expressed in breast carcinoma. A four-lncRNA signature (RP1-193H18.2, AL022341.3, WDR86-AS1, LINC00511) was found to be closely related to the prognosis of breast carcinoma. The four-lncRNA signature could also qualify the magnitude of treatment benefits for different breast cancer subtypes. Additionally, it was an independent risk factor out of other clinicopathological parameters based on the multivariate Cox analysis. We also uncovered that the four predictive lncRNAs are involved in multiple cellular progression and pathways of breast cancer.

Conclusions

The four-lncRNA signature could be an essential reference for prognostic prediction and making therapeutic strategies.

Background

Breast carcinoma is a heterogeneous illness characterized by its various manifestations, morphologies and behaviors[1]. Currently, clinical characteristics, histomorphological properties and immunohistochemical markers are the primary references for the treatment decisions of breast carcinoma. However, response variability to certain types of treatment and chemoresistance are still

puzzling clinical staff. According to various investigations, the genetic heterogeneity of breast carcinoma is a crucial impediment in the achievement of satisfactory results[2]. Gene expression is varied due to the interaction among countless regulators, such as chromatin regulators, transcription factors and non-coding RNAs (ncRNAs)[3]. Over the past few decades, the exploitation of genome-wide transcriptomic and epigenetic screening approaches has resulted in developing predictive tools, allowing breast cancer prognosis and monitoring therapy efficacy[4]. However, up to 75% of patients diagnosed with early breast cancer finally relapse and develop metastatic disease[5]. Meanwhile, a large proportion of patients with breast cancer are overtreated, suffering from the toxic effects of adjuvant treatments without deriving benefits [6]. Therefore, the accuracy of prognostic prediction and the availability of systemic treatment options can be prompted by further exploring the genome pathogenesis of breast carcinoma.

As a major type of ncRNAs, long non-coding RNAs (lncRNAs) refer to RNA transcripts longer than 200 nucleotides [7]. LncRNAs are usually expressed in an illness-, organism- or growth phase-specific manner leading these molecules to compelling therapy targets and pointing toward particular functions for lncRNAs in development and diseases[8]. With the significance of lncRNAs in neoplasm uncovered, the contributions of lncRNAs to the progression of breast cancer have been identified. Accumulating studies have exhibited the critical functions of lncRNAs that support the hallmarks of breast cancer. This repertoire includes maintaining genomic instability, sustaining proliferative signaling, inducing invasion and metastasis, evading growth suppressors or possessing stem cell properties[9, 10]. Previous investigations also demonstrated that lncRNAs specifically expressed or silenced in human cancer could play an essential role in these cancer entities and therefore make them as ideal candidates for breast cancer diagnosis[8]. Such as, FAM83H-AS1 and lncRNA-ATB were found to be overexpressed in patients' sera and could serve as noninvasive tools for the diagnosis of breast carcinoma[11]. Besides, specific lncRNAs could interact with estrogen receptor (ER), progesterone receptor (PR) and human-epidermal growth factor receptor type 2 (HER-2) and be used as biomarkers for different subclasses of breast carcinoma [12]. However, there are still many challenges that should be addressed before their application in the clinic. Thus, further exploring the involvement of lncRNAs in breast cancer may facilitate this progression.

To systematically distinguish prognosis-related lncRNAs that are involved in breast carcinoma, 577 patients from The Cancer Genome Atlas (TCGA) were employed. Dependence on the sample splitting method and Cox regression analysis, a four-lncRNA signature with effective survival risk stratification was successfully established. Besides, it could also be used to assess the degree of benefit from chemotherapy and hormonotherapy. The functional enrichment analysis suggested that the four lncRNAs also participate in multiple biological processes and pathways of breast carcinoma. In short, the present study confirmed the underlying roles of lncRNAs in the prognosis, evaluation of treatment benefit and pathogenesis of breast carcinoma.

Methods

The lncRNA expression patterns and clinicopathological data of breast carcinoma patients were derived from the TCGA database (https://portal.gdc.cancer.gov/). After filtering the information, a total of 577 patients with adequate clinical information (including age, survival status, survival time, AJCC stage, PR status, HER2 level, ER status, treatment measures, radiotherapy and subtypes) were enrolled in the present project. The 577 patients were further randomly subdivided into the training (n=289) and testing (n=288) sets by applied the R package “survival”.

Acquisition of lncRNA expression profile for breast cancer patients

The RNA-seq data of breast cancer was acquired from the TCGA data portal. A total of 15,878 lncRNAs were obtained after adding annotation using the human genome (Ensemble database v72 assembly). The reads per kilobase of exon model per million mapped reads (RPKM) was employed to normalize the expressed values of lncRNAs and mRNAs, and RPKM $\geq$ 0.1 was used as the threshold for adding lncRNAs into the lncRNA expression profiles. Finally, a total of 5,884 lncRNAs were retained for further analysis. Subsequently, the R package ‘DESeq2’[13] was applied to screen out the differently lncRNAs by defining the cut-off value of log2 fold change and *P*-value < 0.05 as the threshold.

Construction of the prognostic lncRNA signature

In the training set, the univariate Cox regression analysis was employed to evaluate the association between the abundantly expressed lncRNAs and the OS of breast cancer patients, which was achieved by performing the R package ‘survival’ (http://cran.r-project.org/package=survival). Those lncRNAs were considered as candidate variables if the *p*-value was less than 0.001. Subsequently, the Random Survival Forest (RSF) method[14] was used to select a smaller number of lncRNAs further to construct the Cox regression analyzed models. The risk index was calculated based on the expression of lncRNAs and coefficients obtained from the multivariate Cox model. The risk-score formula was presented above. Relying on the risk-score formula, the breast cancer patients in the training set were classified into a high-risk and a low-risk subgroup. Then, Kaplan-Meier survival analyses and log-rank test analyses were performed to reveal the differences in patients’ survival time between these two subgroups, and the interactions with treatment therapies. The two-sided log-rank test compared differences in survival times between the low-risk and high-risk groups in each set, a *P*-value < 0.05 was statistically significant. Besides, multivariate Cox regression analysis was also carried out to evaluate whether the four-lncRNA signature was an independent factor for patients’ survival. Meanwhile, to assess the sensitivity and specificity of the survival prediction based on the risk score, the time-dependent receiver operating characteristic (ROC) curve was performed using the R package ‘timeROC’ (version: 0.4)[15] to calculate the area under the curve (AUC).

Functional enrichment analyses

To reveal the functional implications of the four prognosis-related lncRNAs, Pearson correlation coefficients were carried out to reveal the correlation between the prognostic lncRNAs and PCGs, and the genes with a Pearson's correlation coefficient > 0.40 were deemed as lncRNA-associated PCGs.

Thereafter, GO and KEGG pathway enrichment analyses were performed for the lncRNA-associated PCGs using the clusterProfiler package (version 3.12.0)[16], and *P*-values <0.05 were treated as the threshold for GO and KEGG pathway enrichment analyses. The enrichment results were visualized using the ggplot2 package in R language[17].

Results

Identification and selection of prognostic-related lncRNAs

Compared with 36 normal specimens, 69 downregulated and 217 upregulated lncRNAs were discovered in 578 breast carcinoma specimens by performing the DESeq2 package[13] (Fig. 1a, b). In all, 577 samples with whole follow-up data were randomized to the training or testing set. For the training set, the lncRNAs were subjected to the univariate Cox regression model. Thirty-two lncRNAs were screened out that dramatically correlated with overall survival (OS, $P < 0.001$) among the 286 differentially expressed lncRNAs (Supplementary Table1).

Construction of a lncRNA-based prognostic prediction system and validation in the training group

Stepwise random survival forests analysis and the multivariate Cox regression model was further used to screen for the best prognostic assessment indexes in the 32 candidate lncRNAs. Based on the calculation results, a final four lncRNAs presented with an independent statistically significant association with survival prognosis (Fig. 2a). Three of them (RP1-193H18.2, AL022341.3, WDR86-AS1) had negative coefficients, representing an inverse relationship between the expression of lncRNAs with survival. The positive coefficients for the remaining one lncRNA (LINC00511) indicated a positive correlation between lncRNA expression with survival.

Dependence on the levels of four lncRNAs, a risk scoring equation weighted by their regression coefficients for breast cancer patients' survival prediction was constructed as below: risk score= (-0.858 × expression level of RP1-193H18.2) + (-0.684 × expression level of AL022341.3) + (-0.720 × expression level of WDR86-AS1) + (0.466 × expression level of LINC00511). In the training subset, the risk scoring equation was performed to calculate the risk scores for all patients, and the median value of risk scores was regarded as a threshold to divide the set into a high-risk (n=144) and a low-risk group (n=145). The Kaplan Meier curve confirmed that the prognosis of the high-risk subset was prominently worse than that of the low-risk subset. The median survival times for the high-risk and the low-risk groups were 10.85 and 17.27 months, respectively (*P*-value=2.38e-04, Fig. 2b). Furthermore, the high-risk group's 3- and 6-year survival rate was 81.4% and 75.1%, whereas the corresponding survival rates were 100% and 91.7%, respectively, in the low-risk group. We applied the time-dependent receiver operating characteristic (ROC) curves to evaluate the four-lncRNA signature's prognostic accuracy. AUCs of the four-lncRNA signature were 0.78, 0.82 and 0.80 at 1-, 3- and 5-year survival times, respectively, which indicated excellent performance in predicting the prognosis (Fig. 2c). The risk score distribution and survival status for every patient were plotted as a separate dot in the diagram (Fig. 2d). Patients with high-risk score had more significant mortality than those with low-risk score. A heat map demonstrated the expression pattern of

these four lncRNAs in the training set, and the expression pattern was clustered depend on the risk score (Fig. 2e). Among the four lncRNAs, LINC00511 displayed a positive coefficient derived for the multivariate Cox regression model, indicating that LINC00511could be a risk predictor, as its overexpression signified a shorter OS time of patients. However, the other three lncRNAs, including RP1-193H18.2, AL022341.3, and WDR86-AS1, which are negative coefficients, were observed in the multivariate Cox regression model. As their expression levels were higher for the low-risk subset *vs*. high-risk subset, these three lncRNAs could be protective factors.

Verification of the ability of the four-lncRNA signature to predict the prognosis in the testing set

We further estimated whether the four-lncRNA signature maintains its prognostic value in the testing subset. In conformity with the same algorithm used in the training subset, every patient's risk score in the testing subset was computed and subdivided into the low-risk (n=128) and high-risk subset(n=160) by the same threshold point used in the training set. The Kaplan-Meier analysis demonstrated that the high-risk group gets a worse survival time than that of the low-risk group in the testing subset (11.52 months vs. 14.62 months; *P* value= 0.0058; Fig. 3a). The 3- and 6-year survival rates were 86.2% and 79.2% in the high-risk group, 98.4% and 91.9% in the low-risk group. The AUC score at 1, 3 and 5 years also indicated that the four-lncRNA signature could maintain excellent predictive accuracy in the testing subset (Fig. 3b). Additionally, Fig. 3c presents the risk score and survival status for each patient in the testing subset. Not surprisingly, the high-risk lncRNA had the tendency to be upregulated in patients with a high-risk score. By contrast, the protective lncRNAs were highly expressed in patients with a low-risk score (Fig. 3d).

Correlation between the four-lncRNA signature and standard clinicopathologic characteristics

To illustrate the four-lncRNA signature's clinical relevance in breast carcinoma, all the patients were divided into a high-risk and low-risk group in accordance with the median risk score obtained from the training subset. Given this criterion, associations between the four-lncRNA signature and clinicopathologic characteristics of breast carcinoma were evaluated. The findings demonstrated that the four-lncRNA signature has strong affinities with PR status, ER status, treatment therapies and the subtypes of breast cancer (table1). As the four-lncRNA signature is associated with breast carcinoma subtypes, the relationship between the four-lncRNA signature with the prognosis of luminal -type, Her2-type and triple negative-type patients was estimated by the Kaplan-Meier analysis. The findings exhibited that luminal-type patients in the high-risk subset had shorter OS than those in the low-risk subset, the 3- and 6-year OS of the high-risk subset was 88.7% and 79.0%, whereas the corresponding OS was 100% and 94.7% in the low-risk group, respectively (Fig. 4a). However, in the Her2-type and triple negative-type breast carcinoma, no statistical difference in OS was observed between the high-risk and low-risk subset (Fig. 4b, c).

Effect of chemotherapy for patient groups defined by the four-lncRNA signature

In light of the strong associations between the four-lncRNA signature and treatment therapies, we assessed whether it could be utilized to estimate chemosensitivity in patients with breast carcinoma. The

patients were categorized into four main subgroups, depending on whether to accept chemotherapy or not and the risk grades of the four-lncRNA signature. For the subgroup with low signature risk, the 3- and 6-year OS of the patients who received no chemotherapy were 100% and 100%, as compared with the 3- and 6-year OS were 99.0% and 90.5% among the patients who received chemotherapy (*P*=0.369) (Fig. 4d). Besides, for the subgroup with high signature risk, we observed a remarkable difference between the no-chemotherapy and the chemotherapy group concerning OS (*P*=0.030), the 3- and 6-year OS was 73.6% *vs.* 88.3% and 58.3% *vs.* 85.5%, respectively (Fig. 4d). We further validate whether the four-lncRNA signature has a similar role in the subtypes of breast carcinoma. For the luminal type, no statistical difference in OS was identified between the no-chemotherapy and chemotherapy group in patients with high or low signature-risk (Supplementary Fig. 1a-c). We further divided the luminal type into ER+/ PR+ and ER+/ PR- subsets. For the ER+/ PR+ subset, the results confirmed that OS of chemotherapy vs. no chemotherapy was not different for patients with high signature-risk (Fig. 4e); however, the OS of chemotherapy *vs*. no chemotherapy was much longer for patients with high signature-risk in the ER+/ PR- subset (*P*=0.001) (Fig. 4f). The triple-negative type, with or without chemotherapy did not affect the OS of the low-risk subgroup (Supplementary Fig. 1d). By contrast, chemotherapy could prolong the survival time of the high-risk subgroup (Fig. 4g). Thus, the four-lncRNA signature could act as a predictor for chemotherapy benefit.

Effect of hormonotherapy for patient groups defined by the four-lncRNA signature

Since the four-lncRNA signature has a close relationship with hormone receptors and treatment therapies, we also assessed whether it could be applied to direct the uses of hormonotherapy in the luminal type of breast carcinoma. The outcomes of Kaplan-Meier analysis exhibited that the application of hormonotherapy does not affect the survival time of the low-risk subgroup (Fig. 5a); on the contrary, the application of hormonotherapy could conspicuously prolong the OS time of the high-risk subgroup (Fig. 5b). Moreover, in the subset of luminal type without chemotherapy, the OS rate of the high-risk subgroup was worse than the low-risk subgroup (Fig. 5c). In the high-risk group of the luminal type without chemotherapy, the results demonstrated that these patients' OS rate was conspicuously increased by applying hormonotherapy (Fig. 5d). Herein, the four-lncRNA signature could be a reference tool when counseling patients about hormonotherapy options.

Independence of the four-lncRNA signature and other clinical characteristics

To verify the independence of the four-lncRNA signature from other clinicopathological features containing age, AJCC stage, progesterone receptor, HER2 level, estrogen receptor, treatment therapy, radiotherapy and subtypes, the Univariate and Multivariate Cox regression analysis was performed. Univariate Cox regression exhibited that the four-lncRNA signature, AJCC stage, progesterone receptor, estrogen receptor, treatment therapy, radiotherapy and subtype could efficiently predict the outcomes of patients with breast carcinoma (table2). As the multivariate Cox regression analysis demonstrated, confirmed factors with independent prognostic significance for breast carcinoma patients were four-lncRNA signature and AJCC stage (table2). To explore whether the four lncRNA signature could maintain

its prognostic capacity at the same AJCC stage, stratified analysis was employed. In the light of their AJCC stage, patients were categorized into two stratums: the early-stage (stage I/II, n=429) and late-stage (stage III/IV, n=148). Dependence on the threshold point for the training set, the four-lncRNA signature was utilized to further divide breast carcinoma patients into high-risk and low-risk subgroups within each stratum. Kaplan-Meier plots exhibited that the survival rate of the high-risk group was significantly poorer than that of the low-risk group (Fig. 6a-c). Taken together, the four-lncRNA signature was an independent clinical prognostic biomarker for patients with breast carcinoma.

Functional characteristics of four prognostic lncRNAs

For the purpose of revealing the precise mechanism of the four lncRNAs in the tumorigenesis of breast carcinoma, functional category enrichment analysis was applied. Because lncRNAs can act as cis-regulators to modulate their neighboring polycomb group genes (PCGs), Pearson correlation analysis was applied to calculate the correlations between the four lncRNAs and PCGs. The results exhibited that 1,446 genes are closely related to at least one of the four lncRNAs (Pearson's correlation coefficient 0.4 P<0.05). Functional enrichment analysis suggested that lncRNA correlated PCGs were mainly enriched in 362 gene ontology (GO) terms and 17 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (P<0.005); these GO terms were further clustered into different functional categories (Fig. 7a, b). Not only the distribution but also the expression changes of these genes in functional categories (top 10) and KEGG pathways (top 10) were displayed in Fig. 7c and d. Collectively, the four prognostic lncRNAs could be crucial regulatory factors for breast carcinoma-related signal pathways by interacting with PCGs.

Discussion

Breast cancer is a kind of tumor with very complex biological behavior due to its molecular heterogeneity [18]. Traditionally, breast cancer has been primarily stratified according to the AJCC staging system, hormone receptor, Her-2 and Ki-67[19, 20]. However, patients' variable clinical endpoints in the same subgroup reflect that conventional predictors are insufficient to predict the outcomes of breast carcinoma patients precisely. Thus, it is desperate to seek innovative predictors, especially the genetic hallmarks, to more accurately classify breast carcinoma, which may promote the precision of individualized treatment, and enhance the prediction of survival rate and disease recurrence after comprehensive treatment.

Recent technological advancements in gene expression profiling have shed light on numerous diagnostic parameters and help us to differentiate between various subsets of breast carcinoma, which would enable the recommendation of personalized systemic therapies for the particular breast carcinoma subtypes. Currently, multi-gene signatures have been extensively studied to provide prognostic information for breast cancer patients[21]. Multiple signatures have been approved by the FDA for clinical use or recommended by ASCO and NCCN guidelines to assist clinicians in making therapeutic strategies[21]. Whereas their clinical applicability was confined due to their high cost. In recent years, most of the emphasis has been placed on the involvement of Protein Coding Genes (PCGs) in oncogenesis, and the majority of the signatures were composed by PCGs. It should be noted that PCGs

account for only 2% of all transcribed genes in eukaryotes[22], so there is a limitation of fitting in survival modeling of PCGs-based signatures, characterized by the reduced accuracy of predictors when applied to independent cohorts. With the booming advancement of genome and transcriptome sequencing technology, an abundance of non-coding RNAs have been considered as crucial biomarkers in the diagnosis, treatment, and prognosis of breast cancer, and their status in tumors is not inferior to that of PCGs[23, 24]. LncRNAs, a novel and crucial component of non-coding RNAs, emerged as an up-dated layer of cancer-associated procedures[25, 26]. Thus, lncRNAs have great potential to become noninvasive hallmarks for the diagnosis or prognosis of breast carcinoma.

The Cancer Genome Atlas (TCGA) is an open-access database containing complete genomic information derived from 33 cancer types by high-throughput sequencing technology[27]. Such diverse data provide an excellent opportunity to further solve the problems related to tumor heterogeneity. In this project, comprehensive research was carried out to explore lncRNA expression profiles and corresponding clinical information of a large set of breast carcinoma patients acquired from the TCGA database. A four-lncRNA prognostic signature was identified by performing univariate Cox regression analysis and multivariate Cox regression analysis in the training subset, stratifying patients into distinct risk subgroups with a statistically significant difference in the prognosis. Time-dependent ROC curve analysis also displayed the excellent predictive capability of the four-lncRNA signature. Furthermore, the predictive performance of the four-lncRNA signature was well verified in the testing set, which conformed the satisfactory repeatability of the four-lncRNA signature. Above all, our project provides innovative insight into constructing an original prognostic model based on the multi-marker signature for patients with breast carcinoma. As only four members of lncRNA construct the signature, it could be a cheap as well as accurate molecular test and suitable in the clinic prognostication.

For optimal therapeutic strategy making in breast cancer patients, precisely predicting the outcome and sensitivity to treatment therapy will be substantially helpful. Clinicopathological variables are traditional predictive factors that have been employed to achieve this goal. For example, Rouzier developed prediction models that relied on clinical and pathologic characteristics to estimate the probability of pCR and metastasis-free survival of breast carcinoma patients[28]. Nevertheless, these traditional factors are still insufficient for personalized treatment decisions. With the development of the sequencing technology, several multigene panels, including the Oncotype DX (21 genes) and the Amsterdam 70-gene signature (70 genes), have been shown to provide additional prognosis value beyond that provided by clinicopathological factors[29]. However, these signatures are limited to HR+/HER2 − breast cancer but no other subtypes of breast cancer. Recently, prognostic models depended on lncRNAs had been manufactured to predict prognosis in patients with breast carcinoma[30, 31]. Even so, most of the lncRNA signatures are focus on the prognosis of patients, rarely could guide clinical treatment decision-making. This project created a prognostic signature that relied on four lncRNAs expression, which demonstrated an accurate forecast performance. Moreover,strong correlations were found between the four-lncRNA signature with hormone receptors, cancer subtypes and treatment therapies, which indicated that the four-lncRNA signature was not just about predicting prognosis. It might have the potential to guide treatment. Chemotherapies are vital weapons for us to restrain cancer recurrence and progression.

However, the heterogeneous clinical endpoints of patients in the same subtype treated by chemotherapy reflect that part of the patients were undertreated, while others were suffering from the toxicity of over-treatment. Thus, novel signatures are wanted to improve the precision of chemotherapy application. Our investigations verified that for high-risk patients, especially in ER+/PR- and triple-negative subsets, the magnitude of the chemotherapy benefit was greater than that of low-risk patients. The clinical implications of these results for breast cancer patients are relatively straightforward. For the ER+/PR- subset with a low-risk score, the anticipated benefit of adding chemotherapy to hormonotherapy may not exceed hormonotherapy alone. Patients with triple-negative breast carcinoma in the low-risk subgroup got the opportunity not to undergo chemotherapy or decrease the dose and cycles of chemotherapy. For the high-risk subgroup, chemotherapy combined with hormonotherapy may be a critical strategy to improve the prognosis of ER+/PR- subset; and the dose-dense or multiagent chemotherapy regimens could be the primary choice for the treatment of triple-negative breast cancer.

Hormonotherapy is the mainstay of adjuvant treatment for the luminal type of breast cancer. Traditionally, the decision and strategies for hormonotherapy in the treatment of hormone-receptor-positive breast cancer are based on pathological features and clinical trials. However, there is still a lack of trustworthy markers that predict who will benefit from extended adjuvant endocrine therapy, or define a subset of patients with luminal A-type tumors can be safely treated with hormonotherapy alone[32]. Our results showed that the four-lncRNA signature could also be used to make the strategies of hormonotherapy. Dependence on the four-lncRNA signature, the patients with a high-risk score, especially those without chemotherapy, should be treated by hormonotherapy, and combination therapy, including aromatase inhibitors (AI) or tamoxifen plus ovarian function suppression (OFS), might be the primary choices. Patients with a low-risk score might be safely treated with simple hormonotherapy like tamoxifen alone, and the time span of hormonotherapy might not need to be prolonged. In brief, the four-lncRNA signature could be used to guide the formation of systematic treatment strategies for breast carcinoma.

Up to now, multiple gene signatures have been developed to predict breast carcinoma prognosis and responses to treatments beyond what can be accomplished by traditional factors. However, many gene signatures were constructed by selecting genes whose expression levels are related to clinical outcomes without any concern for gene functions. The clinical outcomes are the external appearance of gene functions. We assumed that the lncRNAs in this signature are functionally associated with breast cancer, either directly or indirectly. Increasing studies suggest that lncRNAs can serve as master regulators for gene expression at the levels of chromosome remodeling and transcriptional and posttranscriptional control[33, 34]. Therefore, the PCGs which were closely linked with lncRNAs provided a feasible way to expound the underlying function of lncRNAs[35–37]. For this investigation, 1,446 PCGs were found to have interspecific relations with at least one of the four lncRNAs. Besides, these genes were also found to be enriched in 362 GO terms and 17 KEGG pathways. All the GO terms and KEGG pathways are essential physiological events for cell survival, and their deregulation is frequently observed in oncogenesis [38, 39]. So, the four lncRNAs could be regarded as the key nodes for PCGs to regulate signal pathways. Meanwhile, functional enrichment analyses demonstrated that PCGs have something to do with

immunization and inflammation. According to the evidence that the immunity system and inflammation participate in malignant transformation [40, 41], and lncRNAs acted as the signal transducers that are transmitted between immune and tumor cells to provoke chemoresistance[42]. These findings affirmed that the four prognostic lncRNAs might be involved in tumor immunity. They reflected the genomic changes in patients on chemotherapy, which may help us understand the potential pathogenesis of resistance. Our study provided preliminary insights into the immune-associated roles of the four prognostic lncRNAs in breast carcinoma. Further functional annotation and experimental validations are needed to illustrate how the four lncRNAs implicate the biological processes and define the chemosensitivity of breast carcinoma.

There are still several deficiencies in this study that need to be remedied. First, our investigation's sample size was finite. Large-scale cohort studies are required in order to further assess the predictive value of this four-lncRNA signature and its guiding significance for planning individual comprehensive strategies. Furthermore, multicenter prospective researches are also necessary to validate the prognostic stability of the four-lncRNA signature. Second, the discovery of the targets of lncRNAs plays a vital role in revealing the specific functions of lncRNAs. Although our study displayed the co-expressed genes and the enriched pathways of the four lncRNAs, the mechanism of how lncRNAs interact with these genes warrants a more in-depth study. Identifying the targets of lncRNAs will be of considerable importance in enhancing our understanding of lncRNA-mediated pathways and broadening our views of lncRNA functions.

Conclusions

To summarize, a four-lncRNA signature was successfully constructed to predict the prognosis of breast carcinoma patients. And this signature was proven to be reproducible and robust, and its predictive ability was demonstrated to be independent of other clinicopathological variables and certified to be reliable and stable. Additionally, it could also predict the magnitude of chemotherapy and hormonotherapy benefits. Hence, the four-lncRNAs signature has potential clinical implications as a useful predictive tool for guiding the formation of synthetical therapies in patients with breast carcinoma.

Abbreviations

LncRNAs
long-noncoding RNAs; ncRNAs:non-coding RNAs; ER:estrogen receptor; PR:progesterone receptor; HER-2:human-epidermal growth factor receptor type 2; TCGA:The Cancer Genome Atlas; RPKM:The reads per kilobase of exon model per million mapped reads; RSF:Random Survival Forest; ROC:the time-dependent receiver operating characteristic; AUC:the area under the curve; OS:overall survival; PCGs:Protein Coding Genes; GO:gene ontology; KEGG:Kyoto Encyclopedia of Genes and Genomes; OFS:ovarian function suppression.

Declarations

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Authors' contributions

QC and FW contributed conception and design of the study, and performed most experimental work. YX and LC carried out some experiments. DY, XH and HJ conducted the statistical analysis and interpretation of data. YH and YY conceived and designed the study and prepared the manuscript. All the authors read and approved the final manuscript.

Acknowledgments

This work benefited from the database of TCGA. We are grateful for the access to the resources and the efforts of the staff to expand and improve the databases.

Ethics approval and consent to participate

Not applicable.

Availability of data and materials

The data comes from TCGA database, which is a public open platform (https://portal.gdc.cancer.gov/).

Funding

This study was supported by the National Natural Science Foundation of China (No. 81702371), the Key Research and Development Program of Zhejiang Province (LY18H160033, Y19H160046), the Science and Technology Project of Zhejiang Provincial Department of Health (No. 2020KY062). This study was also supported by the Science and Technology Project of Zhejiang Cancer Hospital (QN201808).

References

1. Sachs N, de Ligt J, Kopper O, Gogola E, Bounova G, Weeber F, Balgobind AV, Wind K, Gracanin A, Begthel H, Korving J, van Boxtel R, Duarte AA, Lelieveld D, van Hoeck A, Ernst RF, Blokzijl F, Nijman IJ, Hoogstraat M, van de Ven M, Egan DA, Zinzalla V, Moll J, Boj SF, Voest EE, Wessels L, van Diest PJ, Rottenberg S, Vries R, Cuppen E, Clevers H. A Living Biobank of Breast Cancer Organoids Captures Disease Heterogeneity. Cell. 2018;172:373 – 86.e10.
2. Haynes B, Sarma A, Nangia-Makker P, Shekhar MP. Breast cancer complexity: implications of intratumoral heterogeneity in clinical management. Cancer Metastasis Rev. 2017;36:547–55.

3. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013;152:1237–51.
4. Napieralski R, Brünner N, Mengele K, Schmitt M. Emerging biomarkers in breast cancer care. Biomark Med. 2010;4:505–22.
5. Rakha EA. Pitfalls in outcome prediction of breast cancer. J Clin Pathol. 2013;66:458–64.
6. van der Hoeven JJ. [70-Gene signature as an aid to treatment decisions in early-stage breast cancer]. Ned Tijdschr Geneeskd. 2017;161:D1369.
7. Tsagakis I, Douka K, Birds I, Aspden JL. Long non-coding RNAs in development and disease: conservation to mechanisms. J Pathol. 2020;250:480–95.
8. Gutschner T, Diederichs S. The hallmarks of cancer: a long non-coding RNA point of view. RNA Biol. 2012;9:703–19.
9. Chi Y, Wang D, Wang J, Yu W, Yang J. Long Non-Coding RNA in the Pathogenesis of Cancers. Cells. 2019;8.
10. Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. Cancer Discov. 2011;1:391–407.
11. El-Ashmawy NE, Hussien FZ, El-Feky OA, Hamouda SM, Al-Ashmawy GM. Serum LncRNA-ATB and FAM83H-AS1 as diagnostic/prognostic non-invasive biomarkers for breast cancer. Life Sci. 2020;259:118193.
12. Yousefi H, Maheronnaghsh M, Molaei F, Mashouri L, Reza Aref A, Momeny M, Alahari SK. Long noncoding RNAs and exosomal lncRNAs: classification, and mechanisms in breast cancer metastasis and drug resistance. Oncogene. 2020;39:953–74.
13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
14. Sun J, Yu H, Zhong G, Dong J, Zhang S, Yu H. Random Shapley Forests: Cooperative Game-Based Random Forests With Consistency. IEEE Trans Cybern. 2020.
15. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med. 2013;32:5381–97.
16. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.
17. Ito K, Murphy D. Application of ggplot2 to Pharmacometric Graphics. CPT Pharmacometrics Syst Pharmacol. 2013;2:e79.
18. Ellsworth RE, Blackburn HL, Shriver CD, Soon-Shiong P, Ellsworth DL. Molecular heterogeneity in breast cancer: State of the science and implications for patient care. Semin Cell Dev Biol. 2017;64:65–72.
19. Mittendorf EA, Bartlett J, Lichtensztajn DL, Chandarlapaty S. Incorporating Biology Into Breast Cancer Staging: American Joint Committee on Cancer, Eighth Edition, Revisions and Beyond. Am Soc

Clin Oncol Educ Book. 2018;38:38–46.

20. Orucevic A, Chen J, McLoughlin JM, Heidel RE, Panella T, Bell J. Is the TNM staging system for breast cancer still relevant in the era of biomarkers and emerging personalized medicine for breast cancer - an institution's 10-year experience. Breast J. 2015;21:147–54.
21. Huang S, Murphy L, Xu W. Genes and functions from breast cancer signatures. BMC Cancer. 2018;18:473.
22. Hangauer MJ, Vaughn IW, McManus MT. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. 2013;9:e1003569.
23. Bautista RR, Gómez AO, Miranda AH, Dehesa AZ, Villarreal-Garza C, Ávila-Moreno F, Arrieta O. Correction to: Long non-coding RNAs: implications in targeted diagnoses, prognosis, and improved therapeutic strategies in human non- and triple-negative breast cancer. Clin Epigenetics. 2018;10:106.
24. Huang QY, Liu GF, Qian XL, Tang LB, Huang QY, Xiong LX. Long Non-Coding RNA: Dual Effects on Breast Cancer Metastasis and Clinical Applications. Cancers (Basel). 2019;11.
25. Bhan A, Soleimani M, Mandal SS. Long Noncoding RNA and Cancer: A New Paradigm. Cancer Res. 2017;77:3965–81.
26. Kang C, Liu Z. Global identification and analysis of long non-coding RNAs in diploid strawberry Fragaria vesca during flower and fruit development. BMC Genomics. 2015;16:815.
27. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19:A68-77.
28. Rouzier R, Pusztai L, Delaloge S, Gonzalez-Angulo AM, Andre F, Hess KR, Buzdar AU, Garbay JR, Spielmann M, Mathieu MC, Symmans WF, Wagner P, Atallah D, Valero V, Berry DA, Hortobagyi GN. Nomograms to predict pathologic complete response and metastasis-free survival after preoperative chemotherapy for breast cancer. J Clin Oncol. 2005;23:8331–9.
29. Kwon MJ. Emerging immune gene signatures as prognostic or predictive biomarkers in breast cancer. Arch Pharm Res. 2019;42:947–61.
30. Sun M, Wu D, Zhou K, Li H, Gong X, Wei Q, Du M, Lei P, Zha J, Zhu H, Gu X, Huang D. An eight-lncRNA signature predicts survival of breast cancer patients: a comprehensive study based on weighted gene co-expression network analysis and competing endogenous RNA network. Breast Cancer Res Treat. 2019;175:59–75.
31. Tang J, Ren J, Cui Q, Zhang D, Kong D, Liao X, Lu M, Gong Y, Wu G. A prognostic 10-lncRNA expression signature for predicting the risk of tumour recurrence in breast cancer patients. J Cell Mol Med. 2019;23:6775–84.
32. Mathew A, Davidson NE. Adjuvant endocrine therapy for premenopausal women with hormone-responsive breast cancer. Breast. 2015;24 Suppl 2:S120-5.
33. Lin C, Yang L. Long Noncoding RNA in Cancer: Wiring Signaling Circuitry. Trends Cell Biol. 2018;28:287–301.

34. Sun Q, Hao Q, Prasanth KV. Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression. Trends Genet. 2018;34:142–57.
35. Balas MM, Johnson AM. Exploring the mechanisms behind long noncoding RNAs and cancer. Noncoding RNA Res. 2018;3:108–17.
36. Marchese FP, Raimondi I, Huarte M. The multidimensional mechanisms of long noncoding RNA function. Genome Biol. 2017;18:206.
37. Sun W, Yang Y, Xu C, Guo J. Regulatory mechanisms of long noncoding RNAs on gene expression in cancers. Cancer Genet. 2017;216–217:105 – 10.
38. Chen L, Zhang YH, Lu G, Huang T, Cai YD. Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways. Artif Intell Med. 2017;76:27–36.
39. Xing Z, Chu C, Chen L, Kong X. The use of Gene Ontology terms and KEGG pathways for analysis and prediction of oncogenes. Biochim Biophys Acta. 2016;1860:2725–34.
40. Gonzalez H, Hagerling C, Werb Z. Roles of the immune system in cancer: from tumor initiation to metastatic progression. Genes Dev. 2018;32:1267–84.
41. Upadhyay S, Sharma N, Gupta KB, Dhiman M. Role of immune system in tumor progression and carcinogenesis. J Cell Biochem. 2018;119:5028–42.
42. Chen F, Chen J, Yang L, Liu J, Zhang X, Zhang Y, Tu Q, Yin D, Lin D, Wong PP, Huang D, Xing Y, Zhao J, Li M, Liu Q, Su F, Su S, Song E. Extracellular vesicle-packaged HIF-1α-stabilizing lncRNA from tumour-associated macrophages regulates aerobic glycolysis of breast cancer cells. Nat Cell Biol. 2019;21:498–510.

Tables

Table 1 Correlations of the four-lncRNA signature with the clinicopathological features of breast cancer

	four-lncRNA signature		
Characteristics	low risk(n=305)	high risk(n=272)	*P*-value
age			0.055
≤50	102	71	
>50	203	201	
AJCC stage			
□	54	45	0.587
□	180	150	
□	65	71	
□	6	6	
Progesterone receptor			
Negative	68	118	**<0.0001**
Positive	237	154	
HER2 level			
0	29	30	0.23
1+	131	119	
2+	109	79	
3+	36	44	
Estrogen receptor			
Negative	44	84	**<0.0001**
Positive	261	188	
Treatment therapy			
No treatment	45	60	**<0.0001**
Hormonotherapy	67	48	
Chemotherapy	45	72	
Hormonotherapy+Chemotherapy	122	60	
Targeted therapy+Chemotherapy	7	12	
Hormonotherapy+Chemotherapy+Targeted therapy	17	20	
Targeted therapy+Hormonotherapy	2	0	

Radiotherapy			
No	130	135	0.092
Yes	175	137	
Subtype			
Luminal type	265	190	**<0.0001**
Her2 type	10	17	
Triple negative type	30	65	

Table 2 Univariate and multivariate Cox regression analysis in the entire set

	Univariable model			Multivariable model		
Variables	HR	95% CI of HR	*P*-value	HR	95% CI of HR	*P*-value
Four-lncRNA signature	1.028	1.021-1.034	**<0.001**	1.023	1.016-1.031	**<0.001**
Age	2.288	0.916-5.716	0.076			
AJCC stage	2.571	1.594-4.148	**<0.001**	2.507	1.483-4.237	**0.001**
Progesterone receptor	0.272	0.124-0.595	**0.001**	0.272	0.124-1.288	0.124
HER2 level	1.200	0.784-1.837	0.402			
Estrogen receptor	0.344	0.161-0.736	**0.006**	0.765	0.051-11.547	0.847
Treatment therapy	0.678	0.531-0.865	**0.002**	0.819	0.621-1.082	0.160
Radiotherapy	0.334	0.150-0.744	**0.007**	0.491	0.200-1.203	0.120
Subtype	1.837	1.239-2.723	**0.002**	1.038	0.242-4.457	0.960

Figures

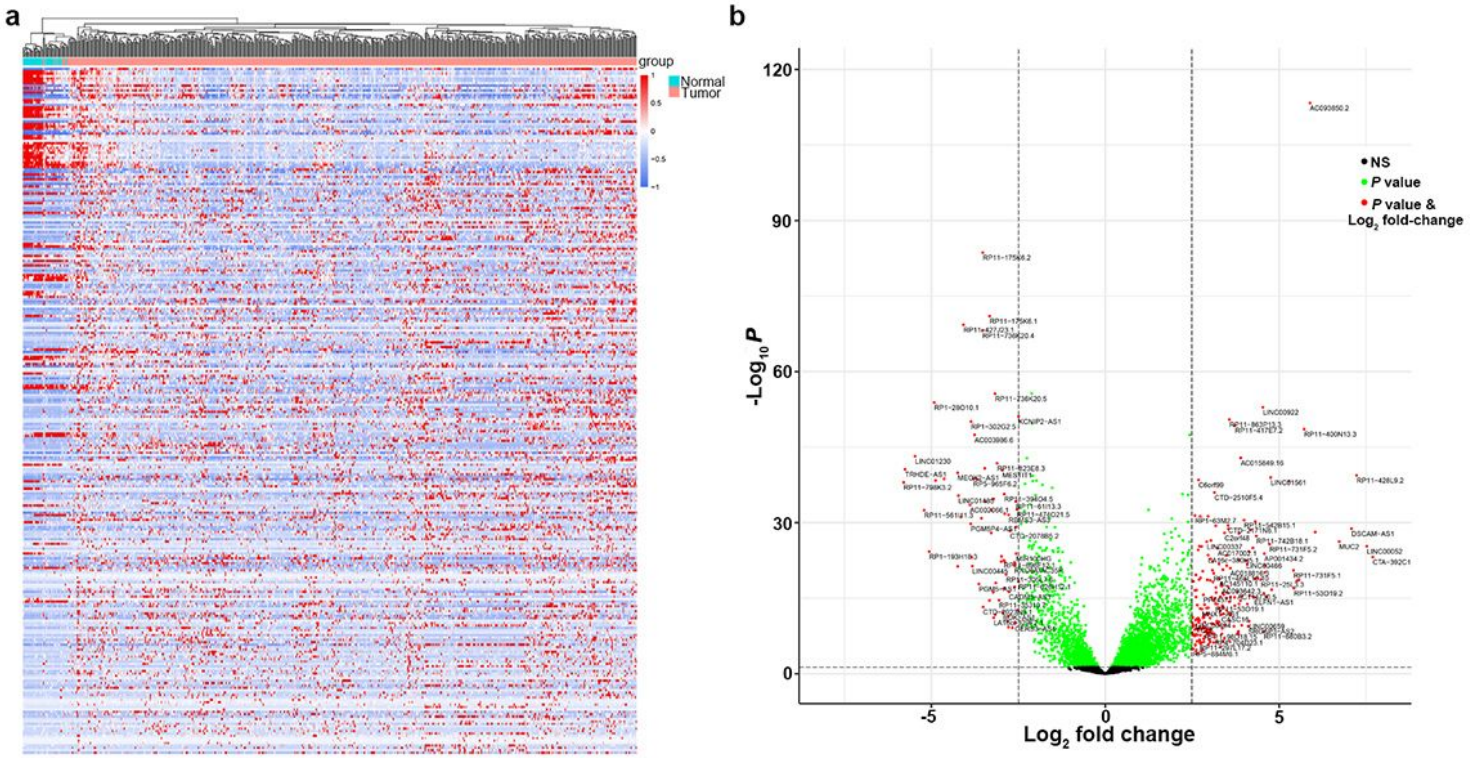

Figure 1

The differentially expressed long non-coding RNAs between breast cancer tissues and peritumoral normal breast tissues. a The unsupervised hierarchical clustering of heatmap exhibits a distinguishable lncRNA expression profile between the breast cancer tissues and normal tissues. b Volcano plot showed the distribution of novel differential lncRNAs. Log2 fold changes are displayed on the X-axis, vertical dotted lines: Log2 fold changes ≥ 2 or ≤ -2. The Y-axis indicates the −log10 P-value, smaller P-value have a greater −log10 P-value, and lncRNAs with a -log10 P-value greater than 1.3 (corresponding to a P-value less than 0.05) are above the dashed horizontal line.

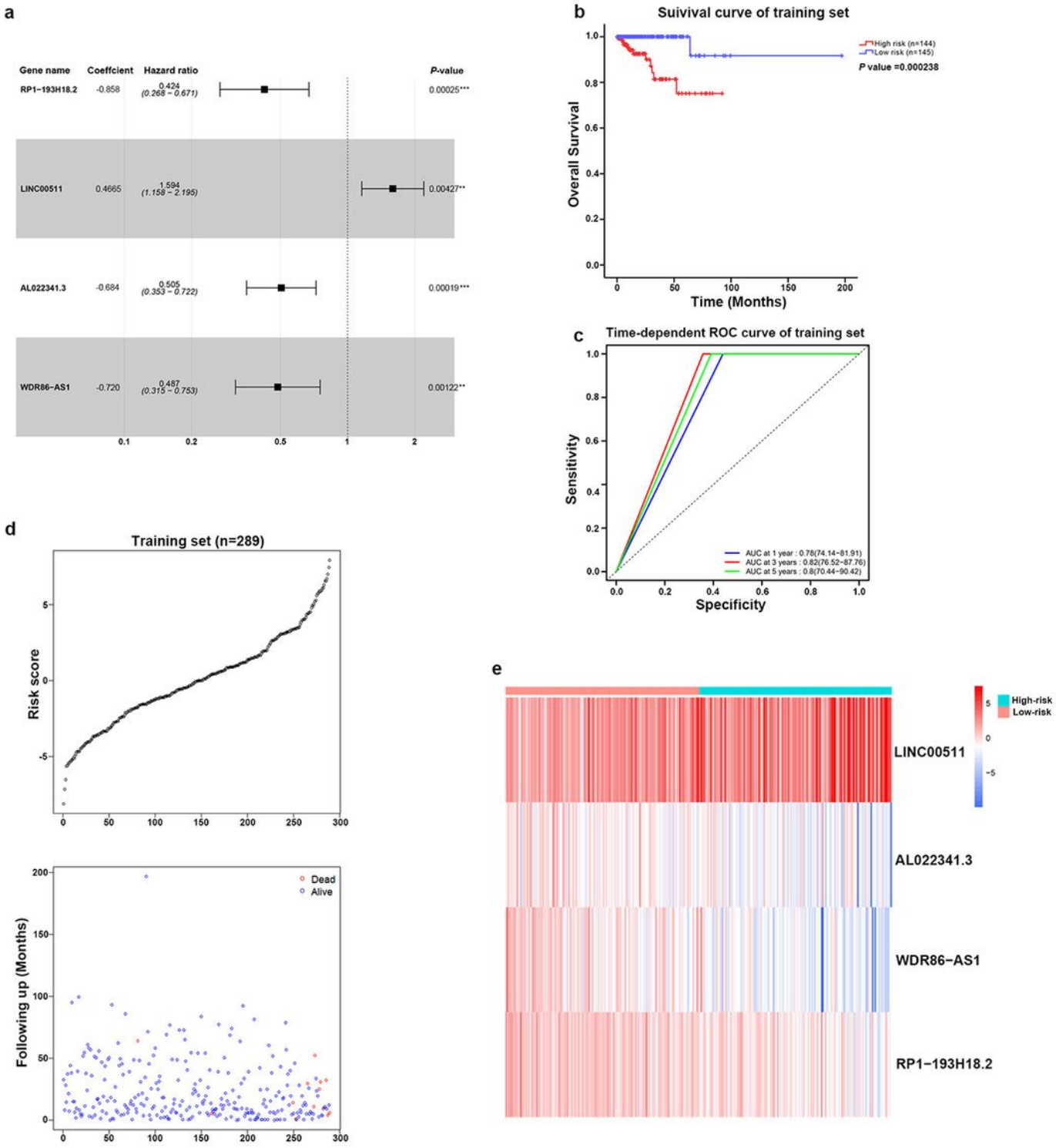

Figure 2

Recognition and performance evaluation of the four-lncRNA signature in the training set. a The forest plot presents the coefficient, hazard ratio and P-value of the four prognosis-associated lncRNAs. b Kaplan-Meier survival curve analysis for the overall survival of breast cancer patients with high or low risk based on the four-lncRNA signature in the training dataset. The P-value represents the differences among the two curves from the results of two-sided log-rank tests. c Time-dependent ROC curve analysis of the four-

lncRNA signature in the training set. d The distribution of four-lncRNA-based risk score and patients' survival status in the training set. e Heatmap of the four-lncRNA expression profiles in the high-risk and low-risk subgroups for the training set.

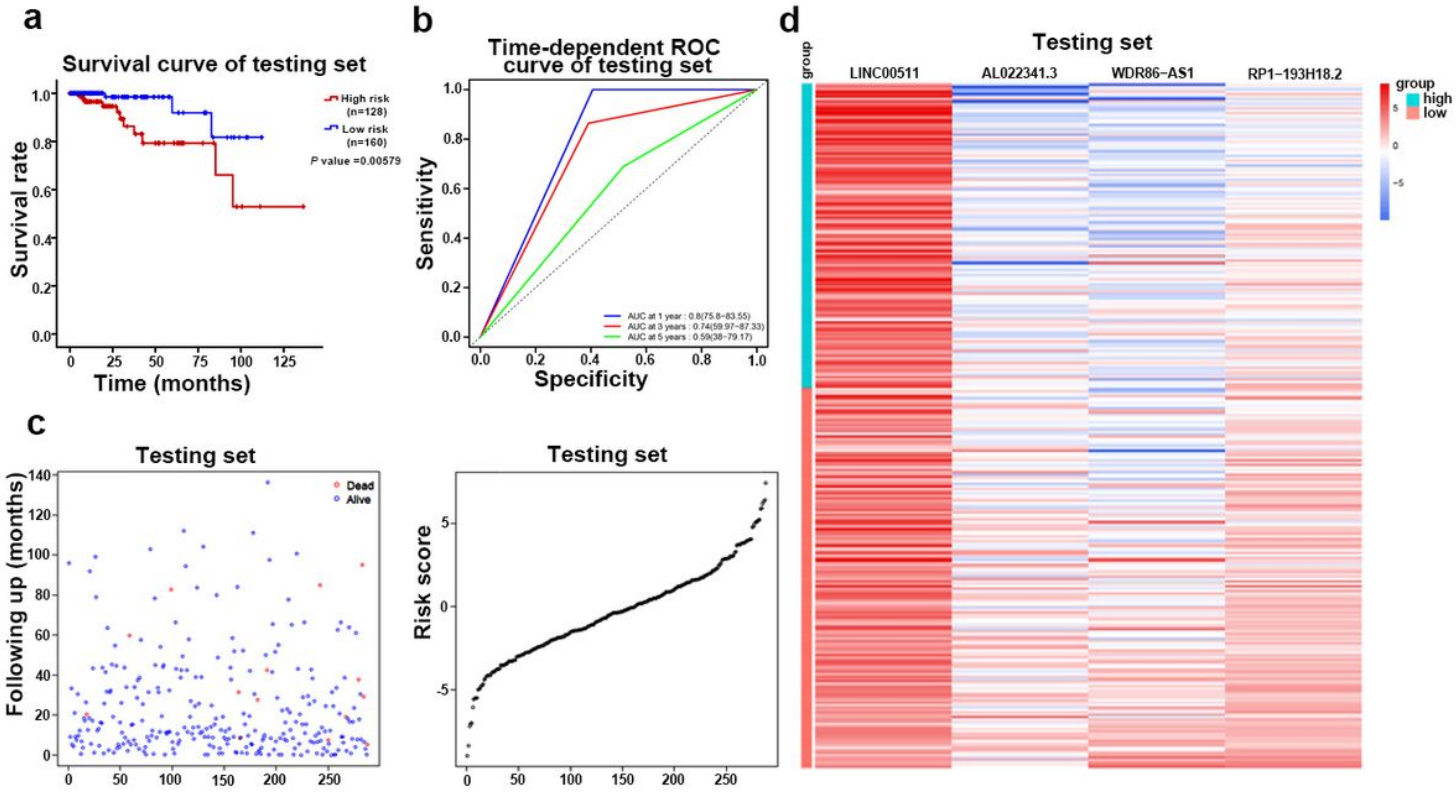

Figure 3

Validation of the prognostic value of the four-lncRNA signature for the breast cancer patients in the testing set. a Kaplan-Meier survival curve analysis for the overall survival of breast cancer patients with high or low risk based on the four-lncRNA signature in the testing set. The differences between the two curves were determined by the two-sided log-rank test. b Time-dependent ROC curve analysis of the four-lncRNA signature in the testing set. c The distribution of four-lncRNA-based risk score and patients' survival status in the testing set. d Heatmap of the four-lncRNA expression profiles in the high-risk and low-risk subgroups for the testing set.

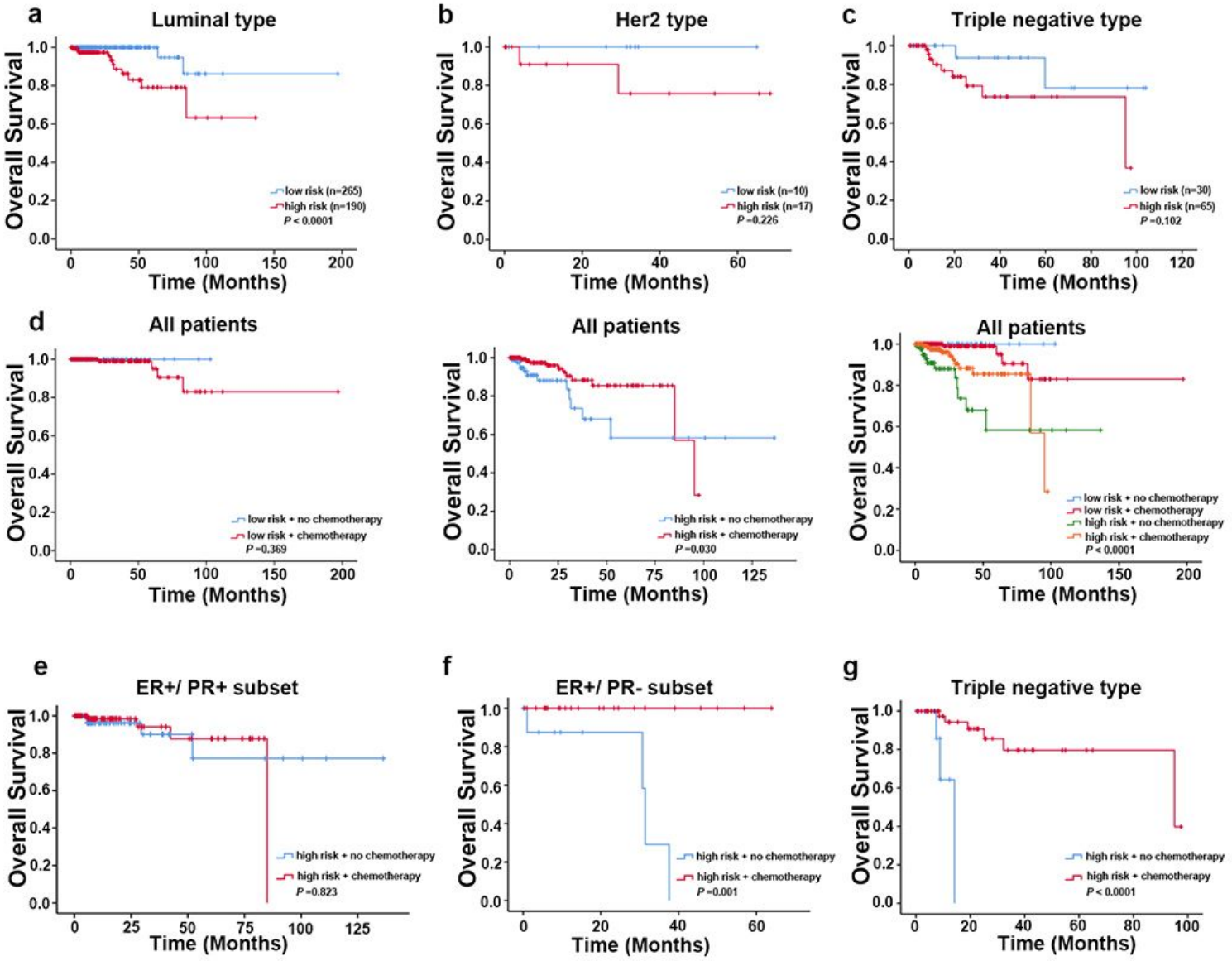

Figure 4

Kaplan–Meier plot for subtypes of breast cancer. a-c Kaplan–Meier survival curves of OS between high-risk and low-risk patients of luminal type, Her2 type and triple-negative type. d The impact of chemotherapy treatments on the OS rate of breast cancer patients in different risk subgroups defined by the four-lncRNA signature. e-g The impact of chemotherapy treatments on the OS rate of the high-risk patients in the ER+/PR+ subset, ER+/PR- subset and triple-negative type breast cancer. The P-value represents the differences among the two curves from the results of two-sided log-rank tests.

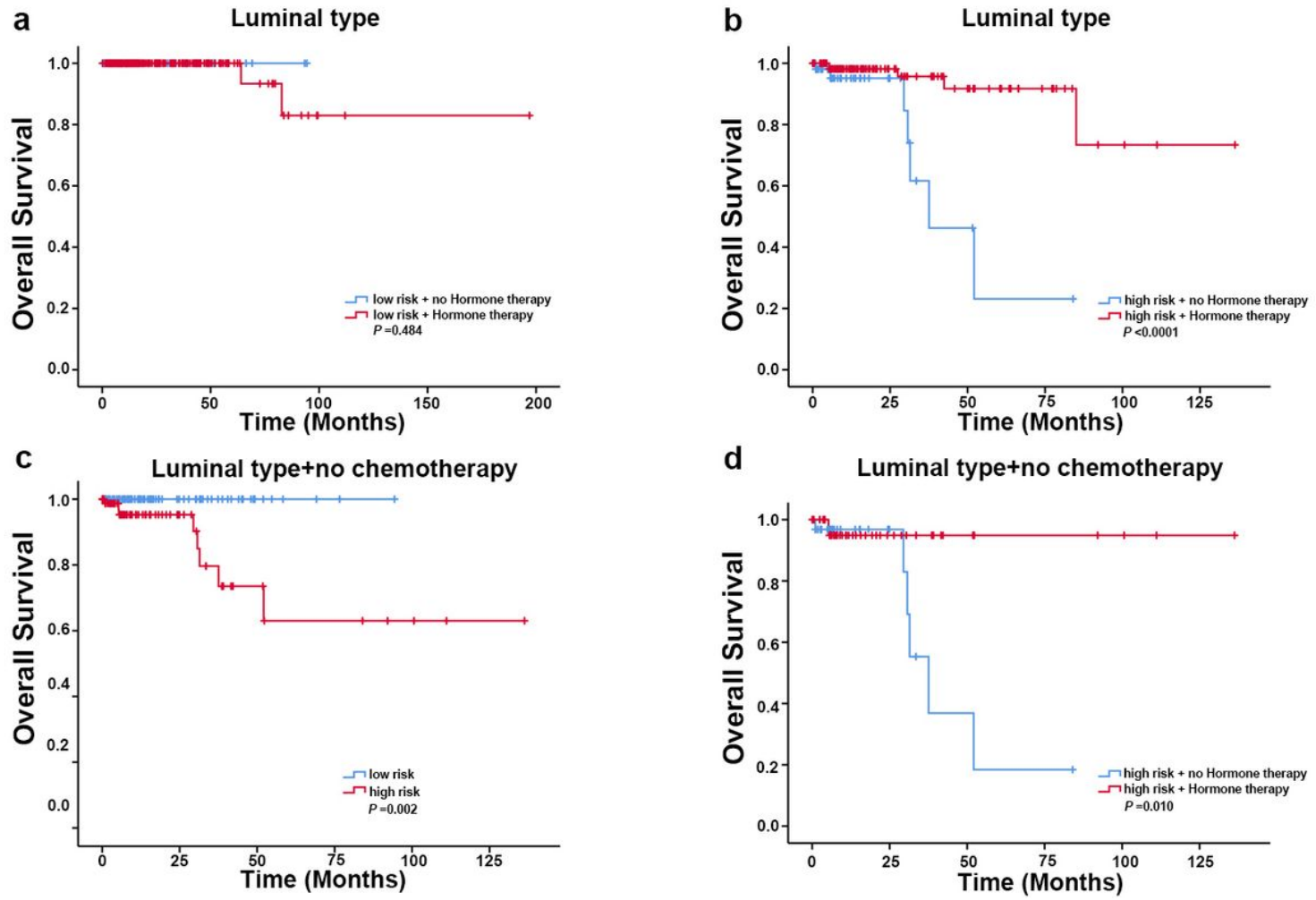

Figure 5

Performance evaluation of the four-lncRNA signature for OS of patients in luminal subtype treated by different strategies. a-b The impact of hormonotherapy treatments on the OS rate of patients with luminal breast cancer in different risk subgroups defined by the four-lncRNA signature. c Kaplan–Meier survival curves of OS between high-risk and low-risk patients without chemotherapy in the luminal subgroup. d The impact of hormonotherapy treatments on the OS rate of high-risk patients without chemotherapy in the luminal subgroup. The P-value represents the differences among the two curves from the results of two-sided log-rank tests.

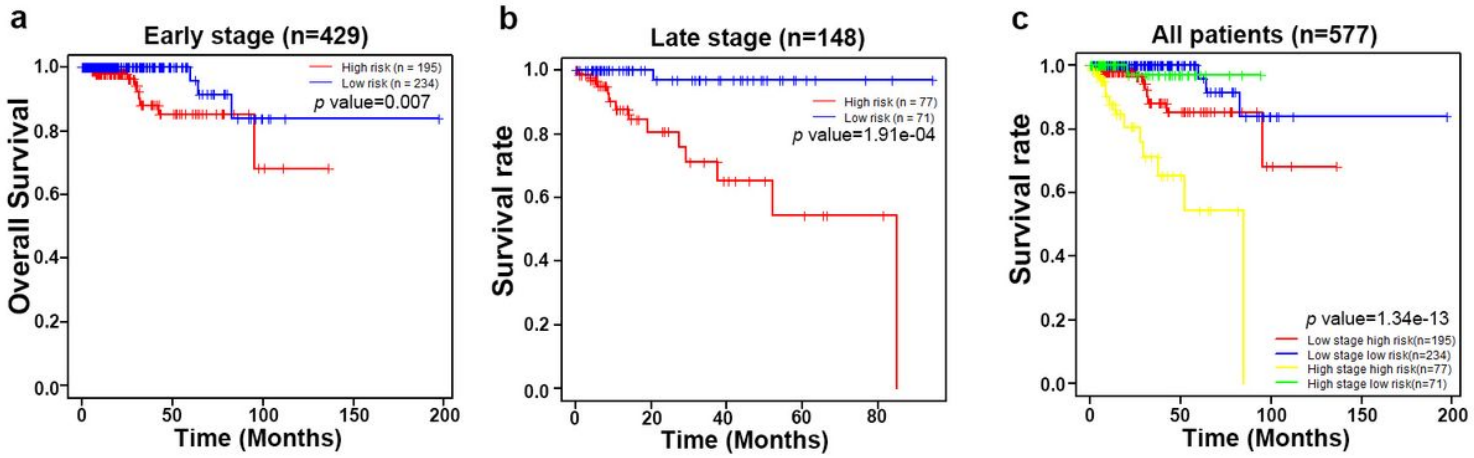

Figure 6

Stratification analyses of all patients adjusted to the AJCC stage using the four-lncRNA signature. a Kaplan-Meier analysis of the early-stage patients' overall survival in the high-risk and low-risk subgroups. b Kaplan-Meier analysis of the late-stage patients' overall survival in the high-risk and low-risk subgroups. c The Kaplan-Meier plot of the entire patients with breast cancer. The P-value represents the differences among the two curves from the results of two-sided log-rank tests.

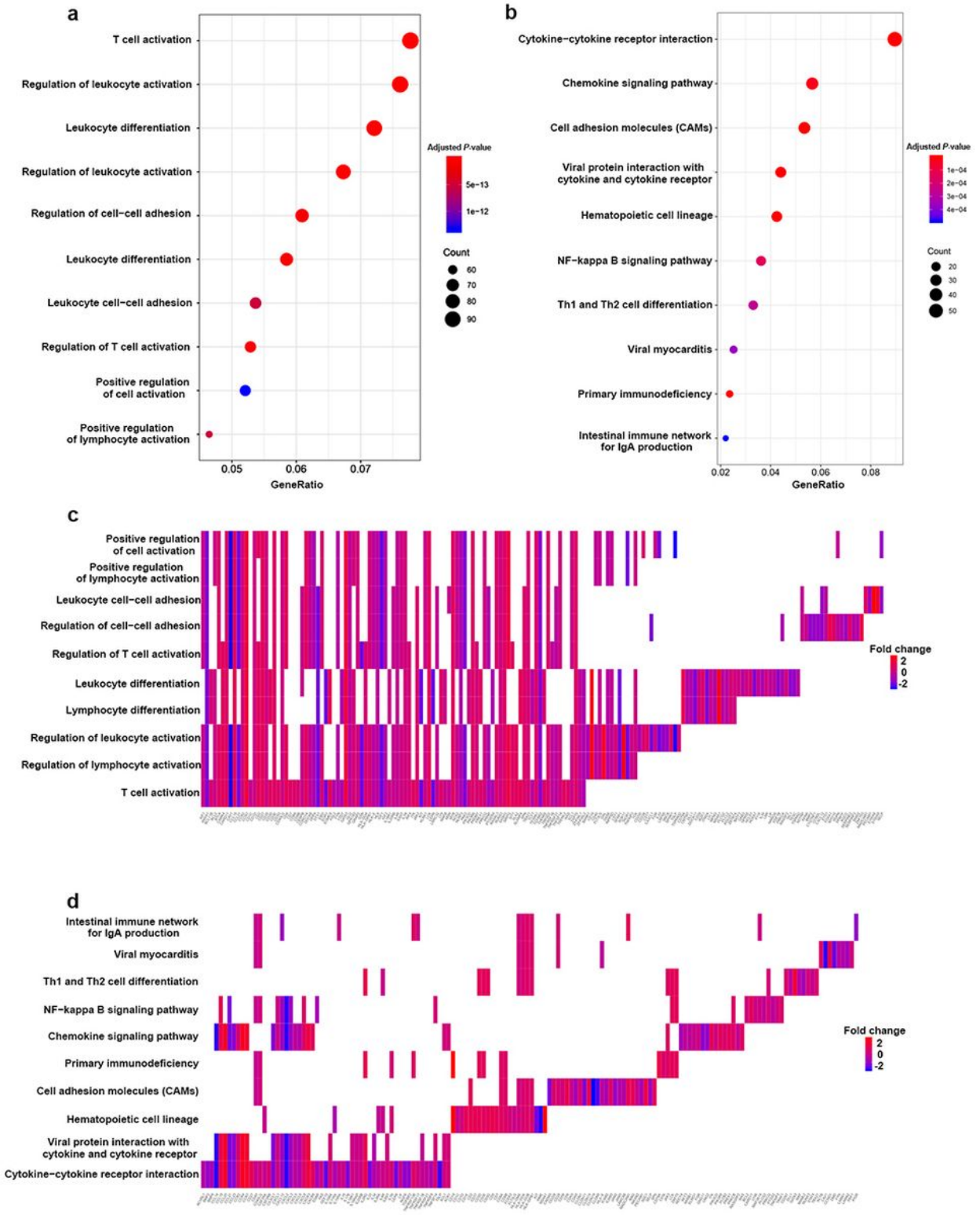

Figure 7

Functional enrichment analysis of the PCGs co-expressed with the four prognostic long non-coding RNAs. a, b The bubble charts visualize the results of GO term and KEGG pathway enrichment analysis. The node size represents the number of genes, and the color intensity represents the adjusted P-value of enrichment analysis. c, d Heatmaps displays the expression changes of PCGs in functional categories and KEGG pathways. The color intensity represents the fold change of PCGs.

Supplementary Files

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

- SupplementaryFiguresandlegends.docx
- Supplementarytable1.docx