A novel metabolic gene signature-based nomogram to predict overall survival in breast cancer

Xi Sun[1#], Zhi-Rui Zhou[2#], Yan Fang[1], Shuning Ding[1], Shuangshuang Lu[1], Zheng Wang[1], Hui Wang[1], Xiaosong Chen[1], Kunwei Shen[1]

[1]Department of General Surgery, Comprehensive Breast Health Center, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [2]Radiation Oncology Center, Huashan Hospital, Shanghai Medical College, Fudan University, Shanghai, China

Contributions: (I) Conception and design: X Sun; (II) Administrative support: X Chen, K Shen; (III) Provision of study materials or patients: S Lu, Z Wang; (IV) Collection and assembly of data: Y Fang, S Ding; (V) Data analysis and interpretation: X Sun, ZR Zhou; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors

[#]These authors contributed equally to this work.

Correspondence to: Kunwei Shen; Xiaosong Chen. No. 197, Rui Jin Er Road, Shanghai 200025, China.
Email: kwshen@medmail.com.cn; chenxiaosong0156@hotmail.com.

Background: Breast cancer risk prediction is often based on clinicopathological characteristics despite the high heterogeneity derived from gene expression. Metabolic alteration is a hallmark of cancer, and thus, the integration of a metabolic signature with clinical parameters is necessary to predict disease outcomes in breast cancers.

Methods: Metabolic genes were downloaded from the Gene Set Enrichment Analysis (GSEA) dataset. Genes with statistical significance in the univariate analysis were applied in the least absolute shrinkage and selection operator (LASSO) analysis to build a gene signature in the GSE20685 dataset. Clinicopathological characteristics and risk scores with prognostic significance were incorporated into the nomogram to predict the overall survival (OS) of patients. The Cancer Genome Atlas (TCGA) and GSE866166 datasets were used as the validation datasets. Time-dependent receiver operating characteristic (tROC) curves and calibration plots were used to assess the accuracy and discrimination of the model.

Results: A 55-gene metabolic gene signature (MGS) was constructed, and was significantly related to OS both in the discovery (P<0.001) and validation (P<0.001) datasets. The MGS was an independent prognostic factor and could divide patients into high- and low-risk groups regardless of their different prediction analysis of microarray 50 (PAM50) subtypes. Time-dependent ROC curves indicated that the risk scores based on the MGS [area under the ROC curve (AUC): 0.931] were superior to the those based on the American Joint Committee on Cancer (AJCC) stage (AUC: 0.781) and PAM50 (AUC: 0.675). A nomogram based on the AJCC stage and risk score could predict OS, and the calibration curves showed good agreement to the actual outcome, indicating that the nomogram may have practical utility. Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) analysis indicated that this MGS was primarily enriched in amino acid pathways.

Conclusions: Our results demonstrated that the MGS was superior to existing risk predictors such as PAM50 and AJCC stage. By combining clinical factors (AJCC stage) and the MGS, a nomogram was constructed and showed good predictive ability for OS in breast cancer.

Keywords: Metabolism; gene signature; breast cancer; survival prediction

Page 2 of 14

Sun et al. MGS and survival in breast cancer

Introduction

Breast cancer is the most common cancer in females, accounting for 30% of new cancer diagnoses (1). It is the second leading cause of cancer death in females worldwide, responsible for nearly 15% of cancer-related fatalities. Despite the dramatic improvement in breast cancer prognosis due to advances in early diagnosis and treatment over the past decades, the high incidence and mortality of breast cancer still poses a significant threat to human health (2).

Growing evidence has shown that breast cancer is a heterogeneous disease. Patients with similar clinicopathological characteristics may differ in clinical outcomes because of different gene expression patterns (3). In recent years, the advent of high-throughput platforms has created great opportunities for researchers to explore the distinct molecular aberrations among tumors (4). On the basis of conventional clinicopathological features, individual gene signatures provide complementary information to predict disease prognosis (5,6). Among these gene signatures, a 50-gene classifier called prediction analysis of microarray 50 (PAM50) helps classify breast cancers into five intrinsic subtypes with distinct clinical outcomes: luminal A, luminal B, human epidermal growth factor receptor 2 (HER2)-overexpressed, basal-like, and normal-like tumors (7). However, the PAM50 signature alone cannot explain the discrepancies between the prognoses in patients with the same subtype, possibly because PAM50 primarily includes proliferation-associated genes (8). As a result, gene signatures involving diverse biological processes are needed to more precisely predict patient survival.

The relationship between metabolism and cancer is multifaceted and bidirectional. Despite the biological significance of metabolism in cancer, metabolic gene signature (MGS) has not been implemented to predict survival in breast cancer patients. For one, aberrant metabolism is a hallmark of cancer, and is involved in diverse aspects of cancer including tumorigenesis, proliferation, and metastasis (9,10). Energy and biomass (nucleotides, amino acids, and lipids) produced via metabolic processes are necessary to satisfy the excessive proliferation of cancer cells and to adapt to environmental stress (11). Metabolites are essential in coordinating gene expression and nutrient utilization (12). Under these circumstances, metabolic pathways are rewritten to adapt to changes in cancer progression. For another, dysregulation of oncogenes and tumor suppressors could result in impaired regulation of metabolic pathways (13-15), and metabolic phenotype is now used as a treatment target and can provide clues regarding cancer treatment (10).

Many metabolic pathways support the survival of cancer (16-19). Based on the complexity and diversity of tumor metabolism, genes from various metabolic pathways were selected to develop an MGS to predict breast cancer prognosis. In this study, metabolic genes with statistical significance in univariate Cox hazard analysis were incorporated into the least absolute shrinkage and selection operator (LASSO) model to calculate the risk score for each patient. Time-dependent receiver operating characteristic (tROC) curve analysis was applied to compare the prediction accuracy of the risk scores and clinicopathological parameters. A nomogram was established based on the risk score and the American Joint Committee on Cancer (AJCC) staging, and calibration plots were constructed in both the discovery and validation datasets.

Our study aimed to elucidate the metabolic pathways contributing to cancer and explore the value of MGS in breast cancer risk stratification. We found that MGS could be used as an effective prognostic predictor and provides a rationale for the individual management of breast cancer patients. We present the following article in accordance with the TRIPOD reporting checklist (available at http://dx.doi.org/10.21037/atm-20-4813).

Methods

Datasets and preprocessing

Datasets from the Gene Expression Omnibus (GEO) database and The Cancer Genome Atlas (TCGA) dataset between 2011 and 2017 were reviewed. Datasets were included if they reported overall survival (OS) and clinicopathological parameters. GSE20685 (platform GPL570), comprising 327 patients, was selected as the discovery dataset, from which messenger RNA (mRNA) expression data and related clinical data for breast cancer were downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo/). Probes were transformed to corresponding Entrez gene names referring to the annotation files.

For the validation dataset, the GSE86166 dataset and related clinical data were downloaded from the GEO dataset. A total of 366 patients were included in this dataset. Another validation dataset was obtained from TCGA database comprising 1,039 patients. Normalized fragments per kilobase of exon model per million reads mapped (FPKM) data and related clinical data were downloaded

from TCGA (https://portal.gdc.cancer.gov/). The metabolic signature was downloaded from the Gene Set Enrichment Analysis (GSEA) dataset. Pathways including glucose and lipid metabolism, amino acid metabolism, and nucleotide metabolism were examined in our analysis, and metabolic genes in both datasets were extracted for further analysis. Batches were removed with the R package "sva". The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Construction of the risk score

Univariate Cox regression was conducted to assess the association between gene expression levels and the OS of patients. Genes with a P value <0.05 were included in further analyses. Genes with a hazard ratio (HR) >1 indicated a poor prognosis, while genes with an HR <1 denoted a good prognosis. The LASSO model, which uses a L1-norm to penalize the weight of the model parameters, was then applied to remove genes of high correlation (20,21). A risk score formula was established by including individual normalized gene expression values weighted by their LASSO Cox coefficients as follows: $\sum_1^i$ Coefficient $(mRNA_i)*Expression (mRNA_i)$. The R package "glmnet" in R 3.5.2 was used to conduct the LASSO analysis. The median value of the risk scores in the discovery dataset was used as the cutoff to divide patients into high- and low-risk groups in both the discovery and validation datasets. In the discovery dataset, the cutoff value from the tROC curve was also tested in different subtypes. Survival was compared between the high- and low-risk groups using the Kaplan-Meier (K-M) survival curve with log-rank tests.

Construction of the nomogram

To evaluate whether the risk score could serve as an independent prognostic indicator for breast cancer patients, univariate and multivariate Cox regression was used in both the discovery and validation datasets. There were 327, 716, and 341 patients with clinicopathological parameters in GSE20685, TCGA, and GSE86166 datasets, respectively. Parameters available in all datasets such as age, intrinsic subtype, and the AJCC tumor-node-metastasis (TNM) staging system (AJCC stage) were included in the analysis. The subtypes of patients were calculated using the "genefu" package in R. AJCC stage was statistically significant in univariate analysis and was assessed in further analysis. A nomogram was constructed to visualize the multivariate Cox

regression. Variables with a P value <0.05 in the multivariate Cox regression were incorporated into the nomogram using the R package "rms" to predict 3-year OS rates.

Statistical analysis

The predictive accuracy of independent prognostic parameters including AJCC stage, subtypes, and risk score was calculated using the tROC curve. The discriminatory ability of different parameters was evaluated by the area under the ROC curve (AUC). The "survivalROC" package in R software was used to plot the tROC curve and calculate the AUC. The calibration plot was used to visualize the performance of the nomogram in all datasets. Functional annotation analysis, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) analysis, was performed on the Database for Annotation, Visualization and Integrated Discovery (DAVID) website (https://david.ncifcrf.gov/). Heatmaps of the LASSO analysis genes were plotted using the "heatmap" R package. A P value <0.05 was considered statistically significant.

Results

Screening of genes related to OS and construction of the MGS

The flowchart of the screening process for prognostic metabolic genes is presented in *Figure 1*. First, metabolic genes were extracted from the GSEA database and a total of 863 genes were included. The genes were further validated as metabolic genes via KEGG (*Figure 2A*) and GO (*Figure 2B*) analysis. In the KEGG analysis, the top five pathways were purine metabolism, pyrimidine metabolism, carbon metabolism, inositol phosphate metabolism, and glycerophospholipid metabolism (*Figure 2A*). In the GO analysis, the top five pathways were oxidation-reduction process, metabolic process, xenobiotic metabolic process, phosphatidylinositol biosynthetic process, and inositol phosphate metabolic process (*Figure 2B*). The GSE20685 dataset was used as the discovery cohort, while the GSE86166 and TCGA datasets were used as the validation cohorts. In the GSE20685 dataset, a total of 161 metabolic genes were statistically significant (P<0.001) in the univariate Cox regression analysis and were selected as candidate genes for further analysis (Table S1). To reduce the high correlation between the genes, LASSO regression analysis was conducted to identify robust markers

Page 4 of 14

Sun et al. MGS and survival in breast cancer
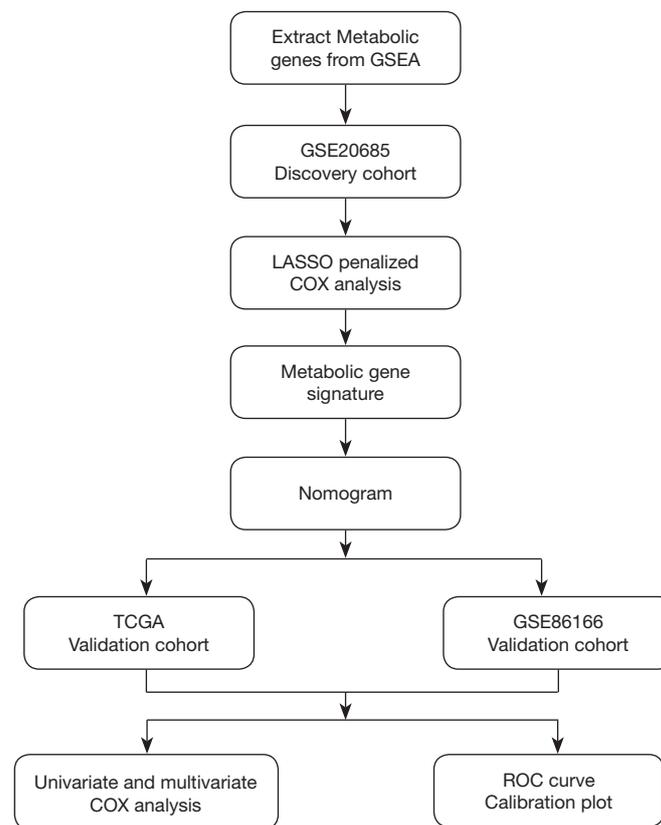
Figure 1 Flow chart of the process used to select target genes included in the analysis. The GSE20625 dataset was used as a discovery dataset. LASSO regression analysis was used to identify the MGS. A nomogram was constructed using clinical parameters and the MGS. ROC curves and calibration plots were used to assess the accuracy and discrimination of the nomogram. LASSO, least absolute shrinkage and selection operator; ROC, receiver operating characteristic; MGS, metabolic gene signature.

(*Figure 2C*), and 55 genes were finally included to construct the MGS (*Figure 2D*). The coefficients of the genes are listed in Table S2. As shown in *Figure 3*, patients with a high-risk score exhibited significantly worse OS compared to those with a low-risk score in the discovery dataset (P<0.001, *Figure 3A*), which was confirmed in TCGA (P<0.001, *Figure 3B*) and the GSE86166 (P=0.032, *Figure 3C*) validation cohorts.

Assessment of the MGS

The OS of the low-risk group was significantly higher than that of the high-risk group in the discovery cohort (*Figure 4A,B*). The median OS was 9.48 years in the high-risk group, while it was still unreached in the low-risk group. Furthermore, in TCGA dataset, the OS of the high-risk group was also worse than that of the low-risk group (*Figure 4C,D*). The median survival time for the high- and

low-risk groups was 11.4 and 17.69 years, respectively. In the GSE86166 validation dataset cohort, the median survival time for high- and low-risk groups was 12.6 and 14.4 years, respectively (*Figure 4E,F*).

Detailed information about clinical characteristics is described in *Table 1*. In the discovery cohort, prognosis was correlated with risk score, AJCC stage, and subtype in the univariate analysis (*Figure 5A*); these factors were included in the multivariate analysis. Risk score [P<0.001, HR =2.275, 95% confidence interval (CI): 1.411, 3.668] and AJCC stage (P<0.01, HR =5.050, 95% CI: 3.910, 6.524) were identified as independent prognostic predictors in the multivariate analysis (*Figure 5B*). In TCGA cohort, age (P<0.001, HR =1.036, 95% CI: 1.019, 1.054), AJCC stage (P=0.010, HR =1.825, 95% CI: 1.157, 2.877), and risk score (P<0.001, HR =1.247, 95% CI: 1.115, 1.393) were identified as unfavorable prognostic factors and independent prognostic predictors in the multivariate analysis
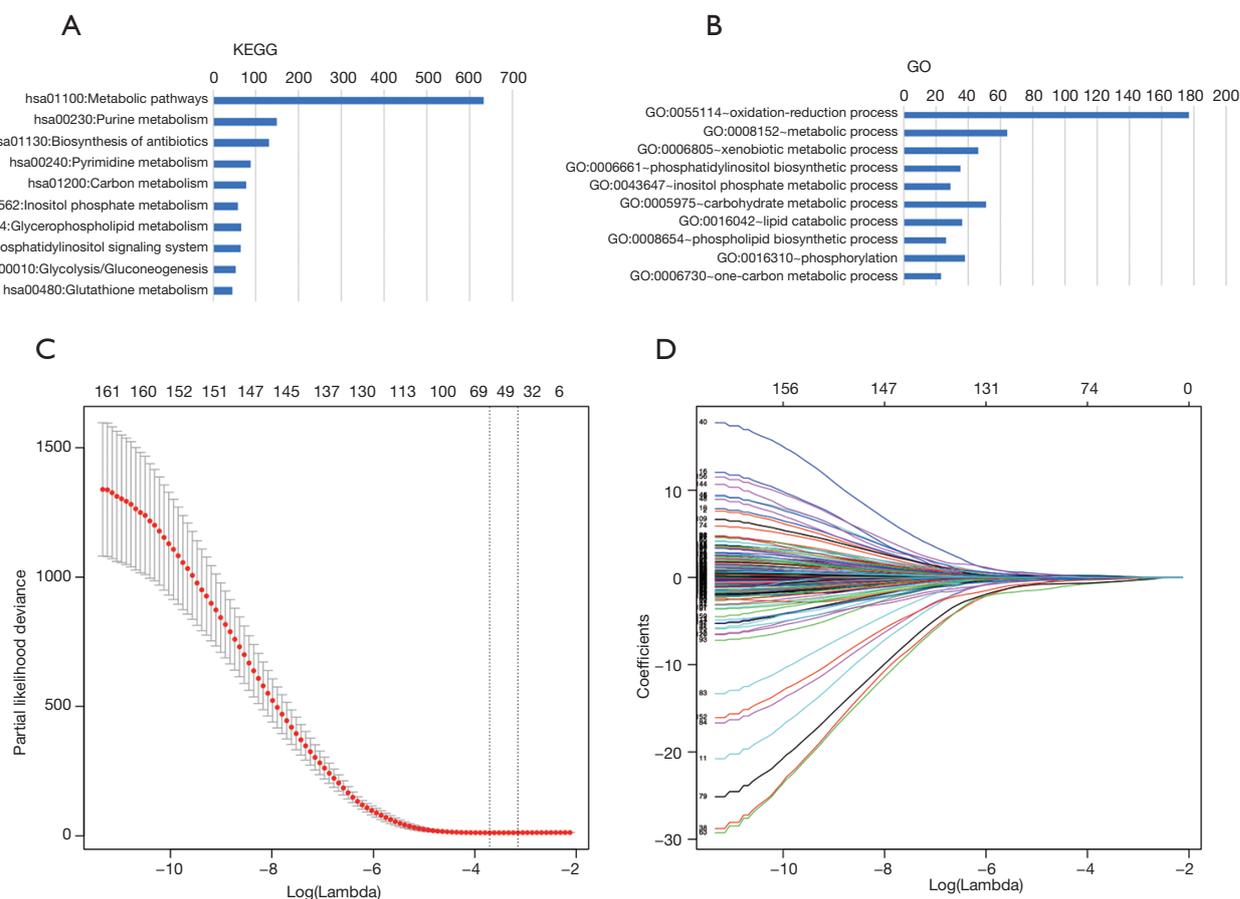
Figure 2 LASSO Cox regression analysis. (A) KEGG analysis of metabolic genes extracted from the GSEA dataset. X-lab is the number of genes; (B) GO analysis of metabolic genes extracted from the GSEA dataset. X-lab is the number of genes; (C) parameter selection in the LASSO model; (D) LASSO coefficient of the genes of prognostic value. KEGG, Kyoto Encyclopedia of Genes and Genomes; GSEA, Gene Set Enrichment Analysis; GO, Gene Ontology; LASSO, least absolute shrinkage and selection operator.

(*Figure 5C,D*). In the GSE86188 dataset, risk score (P=0.025, HR =1.229, 95% CI: 1.026, 1.473) and AJCC stage (P=0.001, HR =2.159, 95% CI: 1.363, 3.418) were independent prognostic predictors as identified through multivariate analysis (*Figure 5E,F*).

To evaluate whether MGS could predict survival regardless of intrinsic subtypes, K-M curves were plotted in different subtypes in the discovery dataset. The outcome of high-risk patients was significantly worse than that of low-risk patients (*Figure 6A*, P<0.001; *Figure 6B*, P<0.001; *Figure 6C*, P=0.007) regardless of their different cutoff values (Figure S1). The median survival time for the low-risk group was unreached in all three subtypes, which indicated that this MGS could stratify patients into different risk groups regardless of intrinsic subtypes.

Indeed, the histological subtype of breast cancer has

a notable effect on the prognosis. Thus, we performed subgroup analysis according to the histological subtype and found that there was no significant interaction between the MGS and the histological subtype (interaction P=0.36). In invasive ductal carcinoma (Figure S2A), a total of 462 patients (60.5%) were classified into the high-risk group, and 302 patients (39.5%) were classified into the low-risk group (P=0.002). In invasive lobular carcinoma (Figure S2B), a total of 86 patients (42.6%) were classified into the high-risk group and 116 patients (57.4%) were classified into the low-risk group (P=0.0004). In both invasive ductal carcinoma and invasive lobular carcinoma, patients with a high-risk score exhibited worse OS compared with patients with a low-risk score. This indicates that the MGS was a robust gene signature regardless of the pathological types in breast cancer. To assess whether the
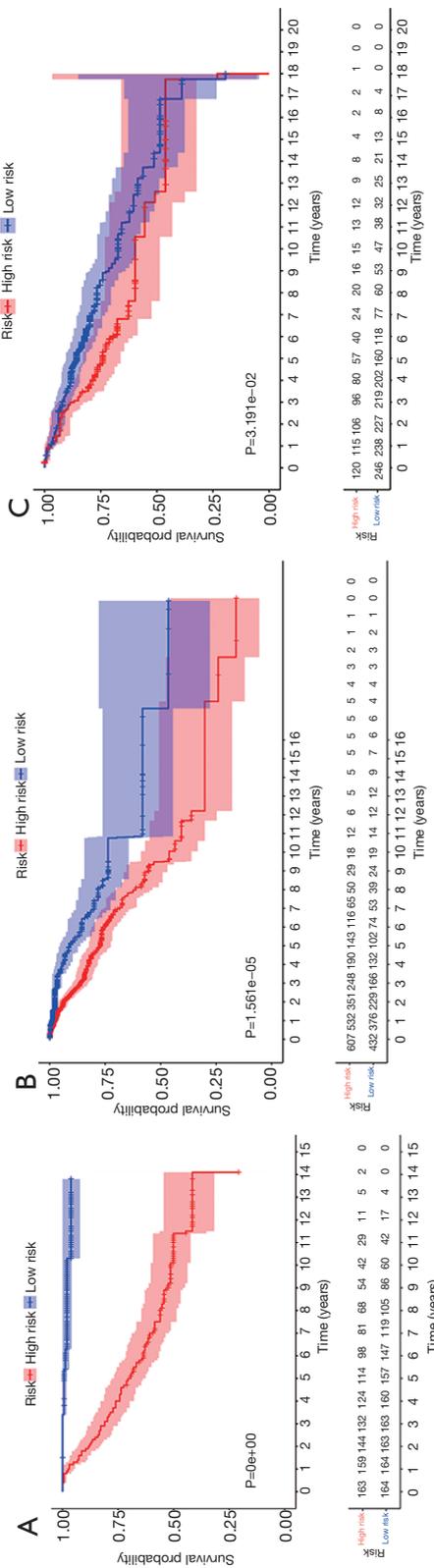
Page 6 of 14

Sun et al. MGS and survival in breast cancer

performance of the MGS was different in predicting the prognosis of males versus females, we applied our signature to lung cancer from TCGA database. A total of 938 patients with OS were included in this analysis. High-risk patients showed worse prognosis compared to low-risk patients in the entire dataset (P=0.026, Figure S3A). We then stratified patients according to sex and found that MGS could stratify female patients into high- and low-risk groups (P=0.045), while there was no difference between the high- and low-risk groups in males (P=0.23, Figure S3B,C). Thus, we cannot conclude that this method is more advantageous in predicting the prognosis of males compared to females; more studies are needed to evaluate whether there is a difference in the MGS for predicting the prognosis in different sexes.
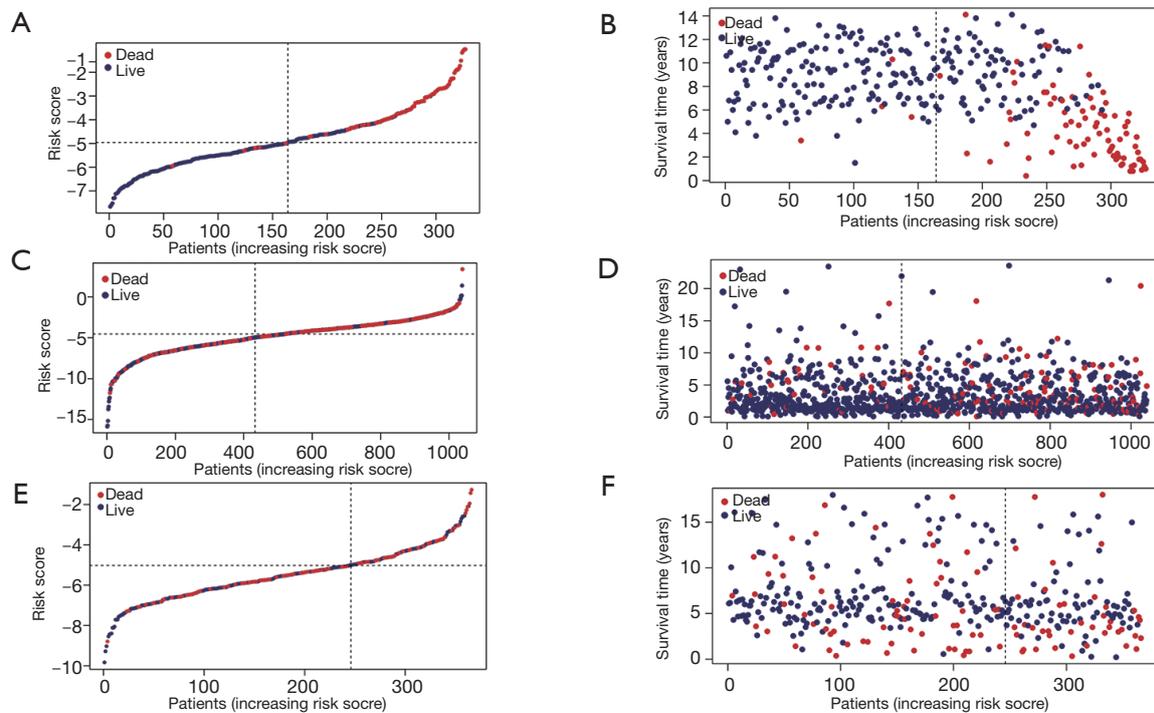
The tROC curves were plotted to compare the prediction accuracy of the risk score, AJCC stage, and subtype on 3-year OS. In the discovery dataset, the AUC of the risk score was equal to 0.931. The AUC values of AJCC stage and subtype were 0.781 and 0.675, respectively (*Figure 7A*). In TCGA dataset, the AUC of risk score was 0.704, which was higher than that of subtype (AUC: 0.629) and AJCC stage (AUC: 0.601; *Figure 7B*). In the GSE86166 dataset, the AUC values of the risk score, AJCC stage, and subtype were 0.676, 0.667, and 0.615, respectively (*Figure 7C*).

Construction of MGS-based model

In our analysis, statistically significant clinicopathological and genetic factors in both datasets (AJCC stage and risk score) were included to construct a nomogram (*Figure 8A*). The concordance index (C-index) of the nomogram was 0.91 (95% CI: 0.85–0.98), which indicated that this model had high prediction accuracy. The presentation of the calibration plot for patient survival prediction in the discovery and validation datasets demonstrated that the nomogram-predicted outcome had good agreement with the actual outcome (*Figure 8B,C,D*).

KEGG and GO pathway analysis of the MGS

KEGG and GO analyses were performed to further evaluate the function of the genes in the MGS (*Figure 9*). In the KEGG analysis, the top three pathways were histidine metabolism, arginine and proline metabolism, and glutathione metabolism (*Figure 9A*), indicating that our signature was associated with amino acid metabolism.

Figure 3 Kaplan-Meier plot of the discovery and validation cohorts based on the metabolic gene signature (MGS). (A) Kaplan-Meier plot of the discovery cohort; (B) Kaplan-Meier plot of the test cohort; (C) Kaplan-Meier plot of the TCGA validation cohort; (C) Kaplan-Meier plot of the GSE86166 validation dataset. TCGA, The Cancer Genome Atlas.

Figure 4 Risk plot of the discovery and validation cohorts. (A) Risk score distribution of patients in the prognostic model in the discovery cohort; (B) relationship between the survival time and risk score rank in the discovery cohort; (C) risk score distribution of patients in the prognostic model in the GSE20685 validation dataset; (D) relationship between the survival time and risk score rank in the GSE20685 validation cohort; (E) risk score distribution of patients in the prognostic model in the GSE86166 validation dataset; (F) relationship between the survival time and risk score rank in the GSE86166 validation cohort.

In the GO analysis, the top three pathways were oxidation-reduction process pathway, xenobiotic metabolic process, and one-carbon metabolic process (*Figure 9B*).

Discussion

The high incidence and mortality of breast cancer poses a significant threat to human health. Traditionally, clinicopathological factors have been common and important considerations for survival prediction of breast cancer patients (22). However, patients with the same clinical factors may differ in clinical outcomes due to diverse gene expression patterns (23,24). The conceptual changes in tumor heterogeneity have provided rationale for risk stratification and treatment decision-making based on gene expression profiles (25). In our study, an MGS was constructed and validated in different datasets. It was shown to be superior to the commonly used clinicopathological factors and could better predict the OS of breast cancer patients.

Cancer metabolism has experienced renewed interest in the past decade, with increasing evidence implying a vital role of metabolism in cancer tumorigenicity and malignancy (26). The application of a metabolic signature has been reported to predict prognosis in different cancers, such as thyroid and hepatocellular cancers; however, no metabolic signatures have been applied to predict outcomes in breast cancer. Metabolism pathways are complex, and one fundamental feature of cancer metabolism is the ability to obtain nutrients and biomass under nutritional stress (27).

To fully understand the impact of metabolism on prognosis, genes covering a variety of metabolic pathways including glucose and lipid metabolism, amino acid metabolism, and nucleotide metabolism were all involved in our analysis. After selection by LASSO analysis, amino acid metabolism pathways including histidine metabolism, arginine and proline metabolism, and glutathione metabolism were found to be largely enriched in our MGS, which indicated that amino acid metabolism played an important role in prognosis. Amino acids are essential

Page 8 of 14

Sun et al. MGS and survival in breast cancer

Table 1 Clinicopathological characteristics of breast cancer patients from the training and validation sets

Variables	Training set (n=327), GSE20685 (%)	Validation set (n=716), TCGA (%)	Validation set (n=341), GSE86166 (%)
Age			
Median [range]	46 [24–84]	58 [26–89]	–
≤55 years	252 (77.1)	316 (44.1)	–
>55 years	75 (22.9)	400 (55.9)	–
Stage			
I + II	216 (66.1)	556 (77.7)	272 (79.8)
III + IV	111 (33.9)	160 (22.3)	69 (20.2)
Subtype			
Lumina A	116 (35.5)	125 (17.5)	133 (39.0)
Lumina B	73 (22.3)	405 (56.6)	91 (26.7)
HER2	74 (22.6)	67 (9.4)	55 (16.1)
Basal	45 (13.8)	109 (15.2)	54 (15.8)
Normal	16 (4.9)	10 (1.4)	8 (2.3)
T category			
T1	101 (30.9)	185 (25.8)	–
T2	188 (57.5)	428 (59.8)	–
T3	26 (8.0)	81 (11.3)	–
T4	12 (3.7)	22 (3.1)	–
Node status			
N0	137 (41.9)	350 (48.9)	–
N1	87 (26.6)	253 (35.3)	–
N2	63 (19.3)	66 (9.2)	–
N3	40 (12.2)	47 (6.6)	–
Metastasis			
M0	319 (97.6)	705 (98.5)	–
M1	8 (2.4)	11 (1.5)	–

nutrients for breast cancer proliferation and are involved in various cancer pathways (28). Histidine was reported to be associated with chemotherapy sensitivity in breast cancer (29), arginine plays an essential role in maintaining mammalian target of rapamycin (mTOR) pathway activation (30,31), while proline, which is catalyzed by two different enzymes (pyrroline-5-carboxylate reductase and proline dehydrogenase), can sustain intracellular nucleotide levels. The inhibition of proline biosynthesis in cancer cells was also found to impair tumorigenic

potential and metastasis inhibition (32,33). Glutathione is the most abundant antioxidant in cells; it maintains cellular redox homeostasis and has a bidirectional effect on cancer progression. It is vital to the detoxification of carcinogens; however, excess glutathione promotes tumor progression (34). Incorporating these different metabolic pathways rather than selecting a particular metabolic pathway makes our prediction model more robust.

Our study highlighted the role of amino acid metabolism in breast cancer prognosis prediction. The application
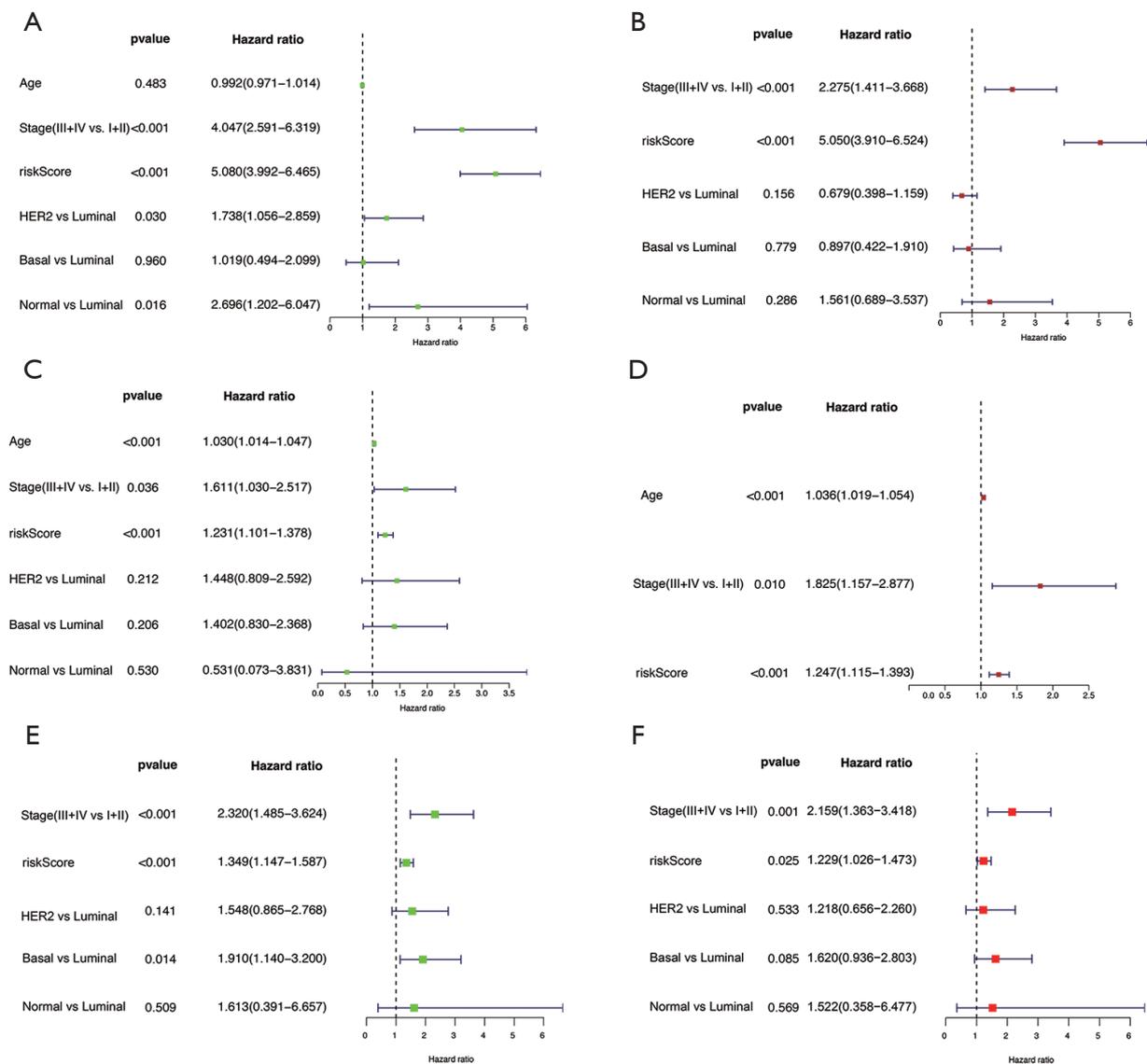
Figure 5 Univariate and multivariate Cox regression analyses of the discovery and validation cohorts. (A) Univariate Cox regression analysis in the discovery cohort; (B) significant parameters in the univariate Cox regression analysis were included in the multivariate Cox regression analysis in the discovery cohort; (C) univariate Cox regression analysis in TCGA validation cohort; (D) significant parameters in the univariate Cox regression analysis were included in the multivariate Cox regression analysis in TCGA validation cohort; (E) univariate Cox regression analysis in the GSE86166 validation cohort; (F) significant parameters in the univariate Cox regression analysis were included in the multivariate Cox regression analysis in the GSE86166 validation cohort.

of an MGS in other types of tumors needs to be further explored. Moreover, it is also well established that there may be marked differences in metabolism between the sexes (35). Thus, the accuracy of an MGS in predicting the OS among patients of different sexes in other tumors needs to be further investigated.

In our study, OS was selected as the end point to evaluate clinical outcomes. LASSO regression analysis, an innovative shrinkage and selection method for regression analysis, converted a high-dimensional predictor into a

Page 10 of 14

Sun et al. MGS and survival in breast cancer
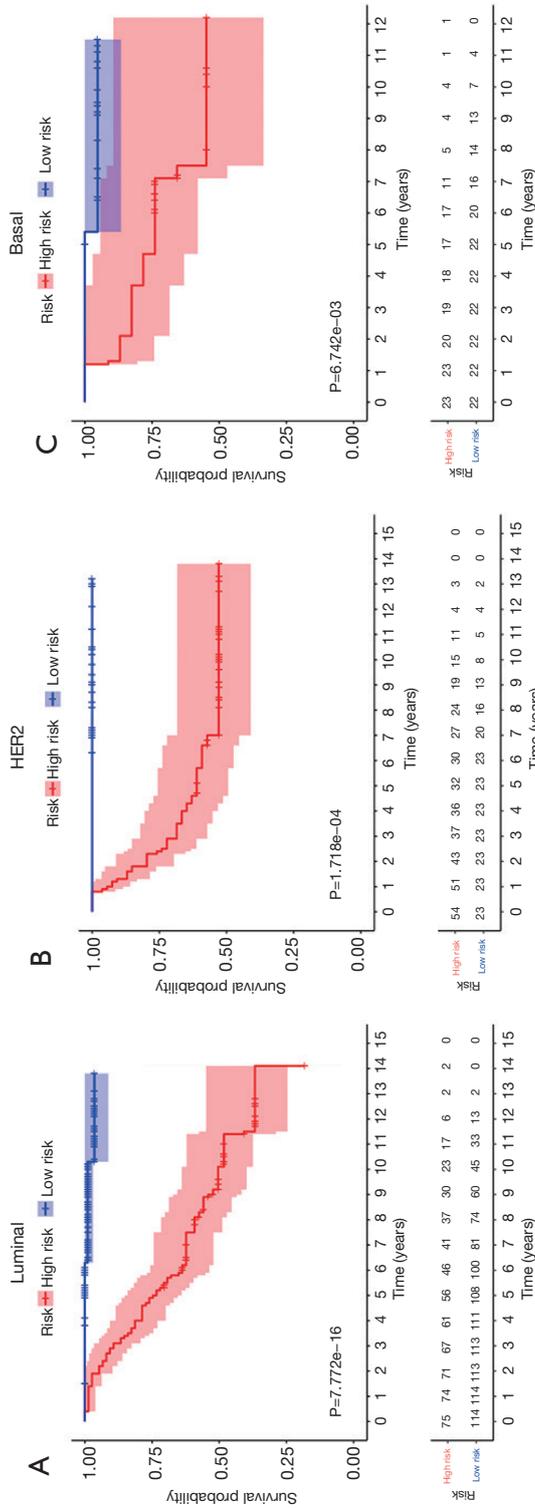
Figure 6 Kaplan-Meier plot of the risk score in different subtypes. Kaplan-Meier plot of the risk score in the luminal (A), HER2 (B), and basal (C) subtypes in the discovery dataset.
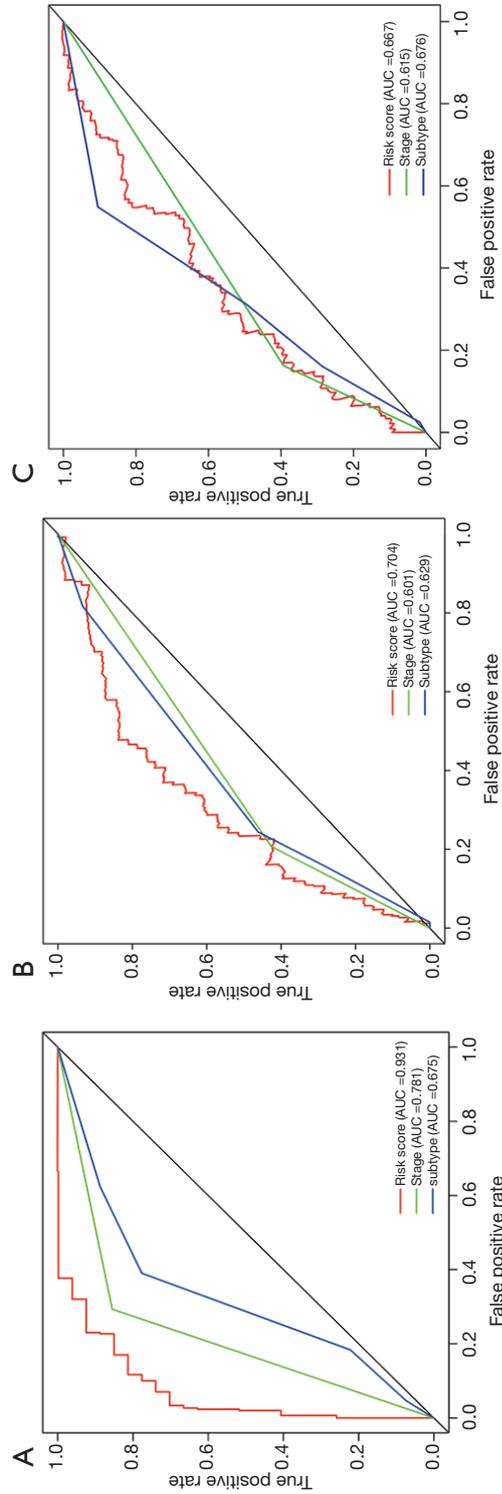
Figure 7 Prediction accuracy based on different factors. (A) Time-dependent ROC curve for OS of the discovery dataset; (B) time-dependent ROC curve for OS of TCGA validation dataset; (C) time-dependent ROC curve for OS of the GSE86166 validation dataset. OS, overall survival; ROC, receiver operating characteristic; TCGA, The Cancer Genome Atlas.
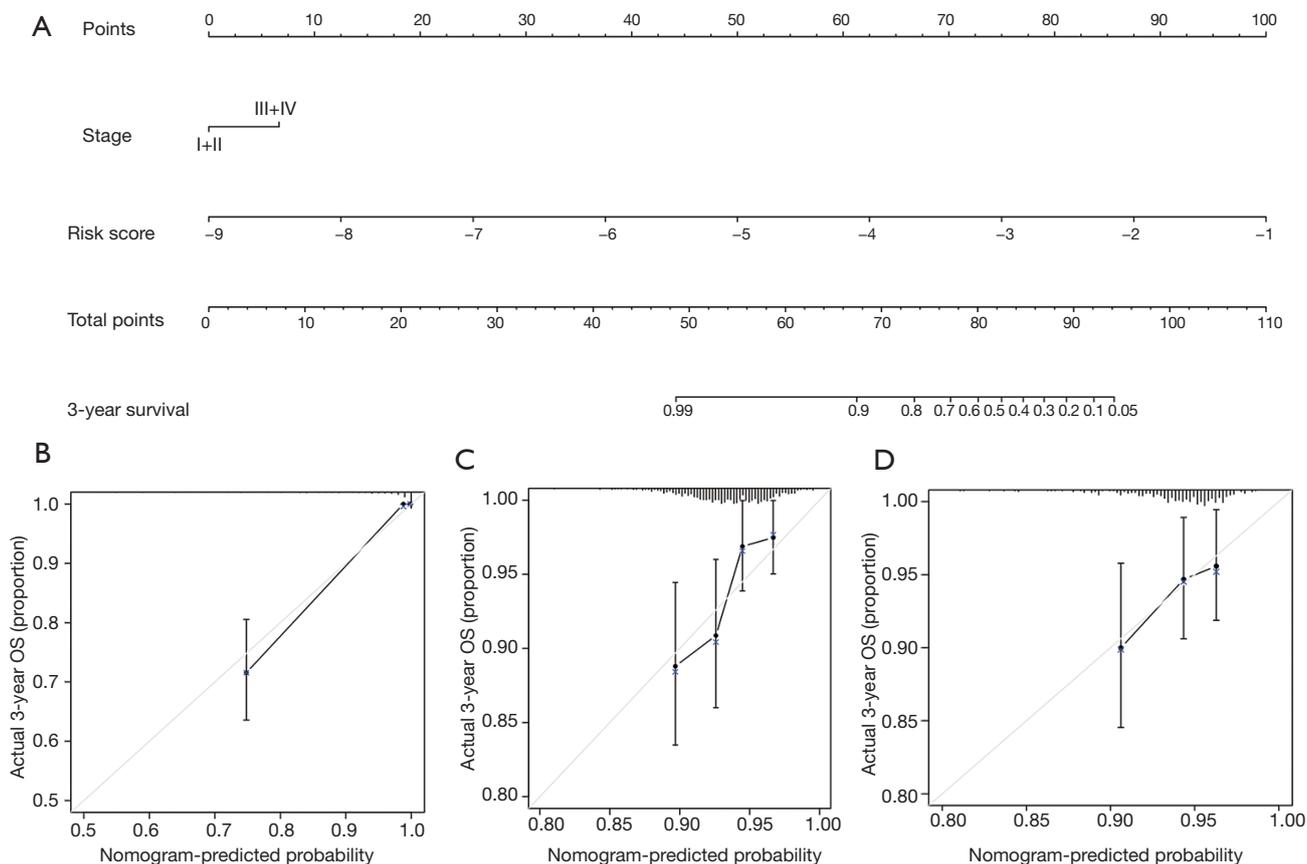
Figure 8 Nomogram for the discovery cohort (verified in the validation cohort). (A) Nomogram to predict 3-year OS in the discovery set; (B) calibration plot for the discovery dataset; (C) calibration plot for TCGA validation dataset; (D) calibration plot for the GSE86166 validation dataset. OS, overall survival; TCGA, The Cancer Genome Atlas.
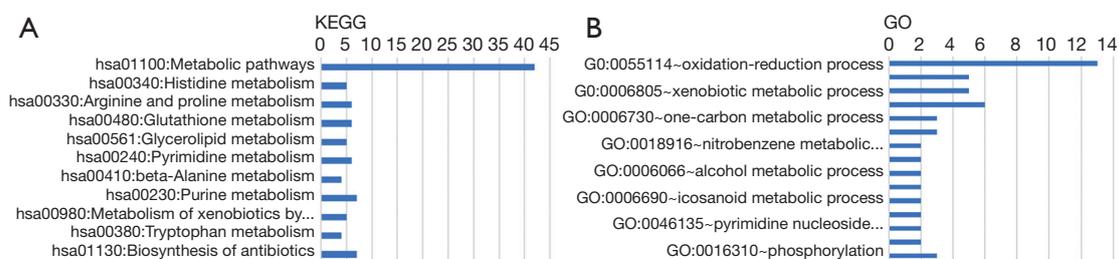
Figure 9 Functional annotation analysis of the metabolic gene signature. (A) KEGG analysis of the metabolic gene signature. X-lab is the number of genes; (B) GO analysis of the metabolic gene signature. X-lab is the number of genes. KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology.

low-dimensional predictor (36,37). The combination of univariate analysis and LASSO Cox regression, which has been used widely in numerous studies, was conducted to screen genes indicating either poor or good prognosis and

to construct a robust gene signature (38,39). In our analysis, a 55-gene metabolic signature was developed to predict the OS of breast cancer patients. According to the tROC curves, the MGS showed a better prediction accuracy than

Page 12 of 14

Sun et al. MGS and survival in breast cancer

did AJCC stage or PAM50.

Intrinsic subtypes are related to patient survival and can guide therapeutic strategies. Among patients with different PAM50 subtypes, our MGS could further classify these patients into different risk groups. Patients with low-risk scores had a better prognosis than patients with high-risk scores, indicating that the MGS can further guide individualized therapy regardless of molecular subtypes. Chemotherapy may be prescribed for patients with high-risk MGSs. Moreover, an appropriate target agent may be designed to improve the outcomes of these patients. Nomograms have been widely used to assess prognostic outcomes in cancer patients (40). A nomogram can incorporate statistical predictive models into a single numerical estimate of the probability of long-term OS for an individual patient. In our study, the metabolic signature and AJCC stage were identified as independent prognostic factors and were incorporated into the nomogram to predict the survival of patients. The high C-index of our nomogram indicated good accuracy in predicting the outcome of patients. Furthermore, calibration plots demonstrated that the nomogram had good discrimination in both the validation and discovery cohorts. However, more validation sets are needed to verify our signature.

Conclusions

Our results demonstrate that a risk score based on an MGS is superior to that based on AJCC stage or PAM50 subtypes in predicting the prognosis of breast cancer. This will provide evidence for further risk classification in guiding individualized management of breast cancer patients.

Acknowledgments

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist Available at http://dx.doi.org/10.21037/atm-20-4813

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-20-4813). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020;70:7-30.
2. Davies C, Godwin J, Gray R, et al. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet 2011;378:771-84.
3. Walker RA, Jones JL, Chappell S, et al. Molecular pathology of breast cancer and its application to clinical management. Cancer Metastasis Rev 1997;16:5-27.
4. Benjamin DI, Cravatt BF, Nomura DK. Global profiling strategies for mapping dysregulated metabolic pathways in cancer. Cell Metab 2012;16:565-77.
5. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817-26.
6. Varga Z, Sinn P, Seidman AD. Summary of head-to-head comparisons of patient risk classifications by the 21-gene Recurrence Score(R) (RS) assay and other genomic assays for early breast cancer. Int J Cancer 2019;145:882-93.
7. Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J

Clin Oncol 2009;27:1160-7.

8. Dai X, Li T, Bai Z, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res 2015;5:2929-43.

9. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646-74.

10. Vander Heiden MG, DeBerardinis RJ. Understanding the Intersections between Metabolism and Cancer Biology. Cell 2017;168:657-69.

11. Sivanand S, Vander Heiden MG. Emerging Roles for Branched-Chain Amino Acid Metabolism in Cancer. Cancer Cell 2020;37:147-56.

12. Dang CV. Links between metabolism and cancer. Genes Dev 2012;26:877-90.

13. White E. Exploiting the bad eating habits of Ras-driven cancers. Genes Dev 2013;27:2065-71.

14. Howell JJ, Ricoult SJ, Ben-Sahra I, et al. A growing role for mTOR in promoting anabolic metabolism. Biochem Soc Trans 2013;41:906-12.

15. Kim IK, Rao G, Zhao X, et al. Mutant GTF2I induces cell transformation and metabolic alterations in thymic epithelial cells. Cell Death Differ 2020;27:2263-79.

16. Fridman A, Saha A, Chan A, et al. Cell cycle regulation of purine synthesis by phosphoribosyl pyrophosphate and inorganic phosphate. Biochem J 2013;454:91-9.

17. Bahreyni A, Samani SS, Rahmani F, et al. Role of adenosine signaling in the pathogenesis of breast cancer. J Cell Physiol 2018;233:1836-43.

18. Wang X, Yang K, Wu Q, et al. Targeting pyrimidine synthesis accentuates molecular therapy response in glioblastoma stem cells. Sci Transl Med 2019;11:eaau4972.

19. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science 2009;324:1029-33.

20. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997;16:385-95.

21. Zhou ZR, Wang WW, Li Y, et al. In-depth mining of clinical data: the construction of clinical prediction model with R. Ann Transl Med 2019;7:796.

22. Giuliano AE, Connolly JL, Edge SB, et al. Breast Cancer-Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA Cancer J Clin 2017;67:290-303.

23. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. N Engl J Med 2009;360:790-800.

24. Beitsch PD, Whitworth PW, Hughes K, et al. Underdiagnosis of Hereditary Breast Cancer: Are Genetic Testing Guidelines a Tool or an Obstacle? J Clin Oncol 2019;37:453-60.

25. Iwamoto T, Pusztai L. Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data? Genome Med 2010;2:81.

26. Boroughs LK, DeBerardinis RJ. Metabolic pathways promoting cancer cell survival and growth. Nat Cell Biol 2015;17:351-9.

27. Pavlova NN, Thompson CB. The Emerging Hallmarks of Cancer Metabolism. Cell Metab 2016;23:27-47.

28. Bhutia YD, Babu E, Ramachandran S, et al. Amino Acid transporters in cancer and their relevance to "glutamine addiction": novel targets for the design of a new class of anticancer drugs. Cancer Res 2015;75:1782-8.

29. Kanarek N, Keys HR, Cantor JR, et al. Histidine catabolism is a major determinant of methotrexate sensitivity. Nature 2018;559:632-6.

30. Wise DR, Thompson CB. Glutamine addiction: a new therapeutic target in cancer. Trends Biochem Sci 2010;35:427-33.

31. Locasale JW. Serine, glycine and one-carbon units: cancer metabolism in full circle. Nat Rev Cancer 2013;13:572-83.

32. Elia I, Broekaert D, Christen S, et al. Proline metabolism supports metastasis formation and could be inhibited to selectively target metastasizing cancer cells. Nat Commun 2017;8:15267.

33. Liu W, Hancock CN, Fischer JW, et al. Proline biosynthesis augments tumor cell growth and aerobic glycolysis: involvement of pyridine nucleotides. Sci Rep 2015;5:17206.

34. Bansal A, Simon MC. Glutathione metabolism in cancer progression and treatment resistance. J Cell Biol 2018;217:2291-8.

35. Tarnopolsky MA. Gender differences in metabolism; nutrition and supplements. J Sci Med Sport 2000;3:287-98.

36. Ushkaryov YA, Lelianova V, Vysokov NV. Catching Latrophilin With Lasso: A Universal Mechanism for Axonal Attraction and Synapse Formation. Front Neurosci 2019;13:257.

37. Ternes N, Rotolo F, Michiels S. Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models. Stat Med 2016;35:2561-73.

38. Wu M, Li X, Zhang T, et al. Identification of a Nine-Gene Signature and Establishment of a Prognostic Nomogram Predicting Overall Survival of Pancreatic Cancer. Front Oncol 2019;9:996.

39. Liu GM, Xie WX, Zhang CY, et al. Identification of

Page 14 of 14

Sun et al. MGS and survival in breast cancer

a four-gene metabolic signature predicting overall survival for hepatocellular carcinoma. J Cell Physiol 2020;235:1624-36.

40. Iasonos A, Schrag D, Raj GV, et al. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol 2008;26:1364-70.

(English Language Editor: A. Kassem; Quality Control Editor: J. Gray)

Table S1 Univariate cox analysis of genes in discovery dataset

ID	HR	HR.95L	HR.95H	P value
GMPR2	0.93928678	0.89133232	0.98982124	0.01914916
HMGCL	0.88829962	0.81878939	0.96371085	0.00438443
LPCAT2	1.12104977	1.02322958	1.22822151	0.01416989
ACSL1	1.02251285	1.00481533	1.04052208	0.01244644
POLR3GL	0.92258178	0.87554178	0.97214909	0.00254594
TAT	0.98671978	0.97509192	0.9984863	0.02707577
LAP3	0.96268921	0.9351051	0.99108702	0.0103606
NAT1	0.98885541	0.98116516	0.99660594	0.00490108
RDH11	1.05997457	1.00245807	1.12079111	0.04073542
OCRL	1.10241284	1.03560898	1.173526	0.00223553
CYP3A5	1.54701459	1.1033392	2.16910097	0.01139845
CNDP1	1.56878361	1.13940445	2.15997228	0.00578351
GSTM1	0.97351446	0.95791165	0.98937141	0.00112921
ACMSD	0.8281429	0.7151563	0.9589801	0.01174814
ASS1	1.01915058	1.0058305	1.03264704	0.00471235
ADH1A	0.68270431	0.49575891	0.94014481	0.01938552
ACADM	0.94144329	0.89812247	0.98685367	0.01205405
GSTM2	0.83311382	0.74841643	0.92739631	0.0008441
GUK1	1.04289432	1.00087553	1.08667714	0.04531997
PGM2L1	1.08428191	1.006178	1.16844857	0.0338852
NME5	0.86181536	0.78063704	0.95143541	0.00321654
CA9	1.04080906	1.01858663	1.06351632	0.00028082
PAFAH2	0.90134425	0.82804583	0.98113103	0.01638844
PLCG1	1.09016127	1.01760919	1.16788606	0.01402029
GLUD1	0.96724806	0.93647506	0.99903229	0.04352309
GNE	1.09174118	1.04707314	1.13831476	3.82E-05
TYRP1	1.05100977	1.00192209	1.10250243	0.04148504
AKR1C1	1.05313592	1.00777427	1.10053938	0.02118301
MTR	0.89726251	0.81262252	0.99071832	0.03199886
TPH1	0.94129425	0.89803004	0.98664278	0.01173185
GSTM3	0.97200339	0.95755327	0.98667157	0.00020256
PFKFB4	1.19741193	1.0139887	1.41401509	0.03369381
NMRK1	0.86923592	0.77560326	0.97417215	0.01595455
IMPA1	1.10914793	1.00462244	1.22454874	0.04023925
B4GALT2	1.05323052	1.0089475	1.09945714	0.01796139
ENTPD4	0.90257535	0.83825275	0.9718337	0.0065802
RDH10	1.01945397	1.00298738	1.03619091	0.02039572
CYP4A11	0.57608177	0.36329863	0.9134915	0.01904603
ADCY3	0.93409429	0.8747008	0.99752069	0.04194922
AGPAT4	1.32870351	1.04785103	1.68483206	0.01898843
POLR2G	0.96103293	0.92431615	0.99920822	0.04551999
PGM3	1.08695954	1.01249577	1.16689974	0.02128234
PSAT1	1.01840917	1.00160456	1.03549572	0.03164828
FPGT	0.84686858	0.75705006	0.94734342	0.00366545
RRM2B	1.06295583	1.0216939	1.10588415	0.00250756
NME7	0.7991865	0.6623959	0.96422557	0.01926806
PPOX	0.87457314	0.79530887	0.96173726	0.00569536
TBXAS1	0.82702663	0.72513526	0.94323513	0.0046384
SULT2B1	1.07518182	1.02523826	1.12755833	0.00281701
WARS2	0.89953354	0.83310456	0.97125933	0.00683072
AHCYL1	0.92908279	0.89476288	0.96471909	0.00012798
PIP4K2C	1.05155799	1.01085041	1.0939049	0.01257042
GAD1	0.87162412	0.7731162	0.9826836	0.02473992
ALOX5	0.92421569	0.87646165	0.97457161	0.00359656
ACOX1	1.08637668	1.0285938	1.14740561	0.00296868
COMT	1.03408292	1.0013122	1.06792616	0.04137281

Table S1 (*continued*)

ID	HR	HR.95L	HR.95H	P value
CHKB	0.61422262	0.44790989	0.84228869	0.00248394
SEPHS1	1.0465633	1.00901643	1.08550733	0.01462702
RDH16	1.12765272	1.00698942	1.26277459	0.03747166
NADSYN1	1.13588603	1.03221106	1.24997409	0.00907576
PNMT	1.00885599	1.00228806	1.01546697	0.0081505
POLD2	1.03756878	1.00255989	1.07380016	0.03520842
POLR2F	0.53889535	0.31789905	0.91352331	0.02168498
SGMS2	1.18190018	1.07738306	1.29655651	0.00040353
GSTT2	0.60859857	0.44134599	0.83923321	0.00245356
PYGB	1.03373864	1.00962457	1.05842866	0.00586311
IDO2	0.68875583	0.49249252	0.96323208	0.02934149
P4HA3	1.15348742	1.02460328	1.29858381	0.01817544
GALC	0.89102211	0.81770824	0.97090915	0.00844214
POLR2D	1.08685517	1.00016986	1.18105355	0.04953368
HAL	1.41455666	1.01713756	1.96725656	0.03930873
CHDH	0.92205973	0.86712027	0.98048007	0.00962845
GSTK1	0.94276428	0.9104742	0.97619953	0.00091754
CYP2J2	1.09368377	1.02862383	1.16285871	0.00421149
GNPDA1	0.936275	0.8766975	0.9999012	0.04965669
UPP1	1.14691292	1.05162324	1.25083699	0.00195268
ALDH2	0.97580175	0.95355556	0.99856694	0.03735676
SHMT2	1.05177226	1.01563809	1.08919202	0.00465612
CA7	0.55668819	0.37807641	0.81968018	0.00300493
FADS2	1.01561905	1.00052305	1.03094283	0.04251918
PLA2G2D	0.92605432	0.87532534	0.97972328	0.00752575
GLA	0.97667728	0.95677598	0.99699254	0.02465809
GYS2	0.72838824	0.55367255	0.95823683	0.02352314
NMNAT1	0.68889869	0.53956853	0.87955724	0.00279455
PFKP	1.04277858	1.01294843	1.07348721	0.00467287
FMO5	0.93111393	0.88864685	0.97561045	0.00272946
ABAT	0.94607714	0.91610165	0.97703346	0.00073991
ACER3	1.06462488	1.01366287	1.11814901	0.01234269
SPTLC2	1.0451376	1.00708258	1.08463063	0.01965323
GGCT	1.02864477	1.0029446	1.05500349	0.0286895
POLR2L	0.97794573	0.95739463	0.99893798	0.03958794
AK1	1.14543884	1.00394101	1.30687972	0.04354646
GSTM4	0.83595974	0.7772085	0.89915214	1.44E-06
GGT7	0.94951452	0.90433644	0.99694956	0.03727001
GUCY1A2	1.13844323	1.00380616	1.29113871	0.04347455
EPHX2	0.92294631	0.89037385	0.95671036	1.22E-05
NT5M	0.89598463	0.81512708	0.98486295	0.02284318
QPRT	1.04158053	1.00796645	1.07631559	0.01493098
PYCR1	1.03384775	1.01045857	1.05777833	0.0043569
RRM2	1.02672081	1.0001274	1.05402134	0.04889816
FBP1	0.98529453	0.97119271	0.9996011	0.04398756
ACSL5	0.9452098	0.90088727	0.99171295	0.02147281
SDS	1.06552834	1.00912736	1.12508162	0.02217183
PNPLA4	0.94247207	0.90373661	0.98286778	0.00565766
ALDH9A1	0.95800646	0.92365978	0.99363034	0.02127924
HMGCS1	1.04521207	1.01419307	1.07717979	0.00401677
PI4KB	0.94254539	0.88889393	0.99943513	0.04783173
AADAT	1.15029906	1.05426189	1.25508467	0.00164443
NOS3	1.24595442	1.06364633	1.45950996	0.00644152
AKR1B10	1.03216298	1.01163834	1.05310404	0.00200771
HADHB	0.92655013	0.88007659	0.97547777	0.00366547
ACY3	1.02498342	1.00795563	1.04229887	0.00388855

Table S1 (*continued*)

Table S1 (*continued*)

ID	HR	HR.95L	HR.95H	P value
ACYP1	0.79887278	0.68973981	0.92527313	0.00273286
APRT	1.03749206	1.01014782	1.06557649	0.00691599
HMGCS2	0.99105154	0.9826809	0.99949348	0.03779833
NUDT2	0.96662197	0.93463537	0.99970328	0.04801353
ACP5	0.98170868	0.96755126	0.99607326	0.0127446
G6PD	1.03562466	1.01666399	1.05493894	0.00020487
DNMT3B	1.19506933	1.0823914	1.31947712	0.00042045
BST1	0.83968975	0.72129957	0.97751185	0.02424086
CP	1.01042422	1.00069363	1.02024942	0.03569337
UGCG	0.97992958	0.9615458	0.99866485	0.03588462
AGXT	1.10230048	1.01999549	1.19124679	0.0138935
AMDHD1	0.83009727	0.70453085	0.97804302	0.02606257
GK	0.73427975	0.60138476	0.8965421	0.00242898
ENTPD5	0.96623398	0.93866183	0.99461603	0.02004828
UGDH	0.98241867	0.96552116	0.99961191	0.04508987
ACADSB	0.97188792	0.95574417	0.98830436	0.00084825
AOC2	1.43067831	1.10338094	1.85506233	0.00688751
AOC1	1.05276271	1.01370481	1.0933255	0.00768439
CTPS1	1.04789201	1.00190846	1.09598603	0.04102864
GNPNAT1	1.04398732	1.00604596	1.08335959	0.0226613
PNP	1.05394073	1.01174451	1.09789681	0.01173422
TSTA3	1.03149967	1.00764714	1.05591683	0.0093724
PTGDS	0.98448256	0.97101637	0.99813549	0.02604484
ACACB	0.96584968	0.93633514	0.99629454	0.02820576
GLUL	0.95632979	0.93520721	0.97792944	8.91E-05
ECI2	0.9511449	0.92298804	0.98016072	0.00108712
NMNAT3	0.90536339	0.83132083	0.98600063	0.02238173
AFMID	0.97273679	0.94826937	0.99783553	0.03344735
ALDH3A2	0.9649469	0.93303706	0.99794805	0.03755476
DGKA	0.83668554	0.73863889	0.94774685	0.00504891
ASAH1	0.96723393	0.94307976	0.99200673	0.00982467
CHST13	0.53275815	0.28411808	0.99899044	0.04963305
CAT	0.96910367	0.94085917	0.99819608	0.0375616
ALDH1A1	0.97702672	0.95897878	0.99541432	0.01456095
PTDSS1	1.04241309	1.00624335	1.07988297	0.02114443
AHCYL2	0.89157111	0.82179687	0.96726949	0.00577418
ACSM1	0.98046187	0.96453315	0.99665364	0.01822286
BDH2	0.93218544	0.87372685	0.99455533	0.033571
MAT2A	0.95816582	0.92048538	0.99738872	0.03682496
AGMAT	0.77934155	0.61212978	0.99222954	0.04304507
ADCY1	0.95712783	0.92801714	0.98715168	0.0054266
ADO	1.08818765	1.00905737	1.17352333	0.02823238
CHPT1	0.95003371	0.92659765	0.97406254	5.77E-05
NME6	1.20998713	1.02187239	1.43273159	0.02704048
ETNK1	0.94352226	0.89969099	0.98948892	0.01660481
MBOAT1	0.96294432	0.9403508	0.98608069	0.0018265
PDE6B	0.81705428	0.71034595	0.93979236	0.00466083
AGPAT2	1.02753505	1.00371354	1.05192192	0.02322649
LCT	1.47339319	1.0488791	2.06972139	0.02540505

Table S2 Coefficient of genes in metabolic prognostic table

Gene	Coefficient
LPCAT2	0.04095005
TAT	−0.0021002
OCRL	0.06942362
CNDP1	0.00608969
ADH1A	−0.0472956
GNE	0.06236155
AKR1C1	0.01225844
GSTM3	−0.0210316
CYP4A11	−0.0464556
AGPAT4	0.14475202
RRM2B	0.0147272
PPOX	−0.0389313
TBXAS1	−0.0096578
SULT2B1	0.05279369
WARS2	−0.0208845
CHKB	−0.0175868
NADSYN1	0.0820863
POLR2F	−0.6217342
SGMS2	0.05449231
GSTT2	−0.0555444
PYGB	0.00726802
HAL	0.26174843
CYP2J2	0.02452044
UPP1	0.05409782
ALDH2	−0.0167485
CA7	−0.5635596
PLA2G2D	−0.0239303
GYS2	−0.1563698
SPTLC2	0.00659571
POLR2L	−0.0184509
AK1	0.06426064
GSTM4	−0.0715992
GGT7	−0.007687
NT5M	−0.0130065
ALDH9A1	−0.0088336
NOS3	0.15157501
ACY3	0.01008885
ACYP1	−0.0814231
G6PD	0.00299901
BST1	−0.150993
AGXT	0.02307994
GK	−0.1001818
AOC1	0.01578934
TSTA3	0.00435417
GLUL	−0.0118676
AFMID	−0.0133855
DGKA	−0.1688997
CHST13	−0.0798251
AHCYL2	−0.0564555
MAT2A	−0.0198193
AGMAT	−0.2480762
ADCY1	−0.0022782
ADO	0.00781095
NME6	0.07458559
AOC1	0.01578934
TSTA3	0.00435417

Table S2 (*continued*)

Table S2 (*continued*)

Gene	Coefficient
GLUL	−0.0118676
AFMID	−0.0133855
DGKA	−0.1688997
CHST13	−0.0798251
AHCYL2	−0.0564555
MAT2A	−0.0198193
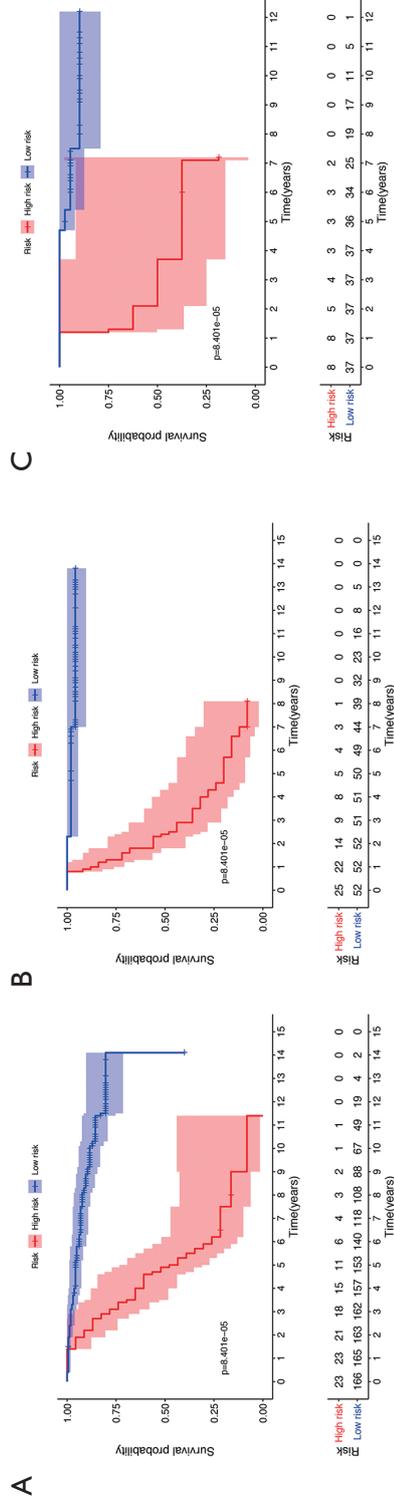
Figure S1 Kaplan-Meier plot of the risk score in different subtypes using a cutoff value from the ROC curve. Kaplan-Meier plot of the risk score in the luminal (A), HER2 (B), and basal (C) types in the discovery dataset. ROC, receiver operating characteristic.
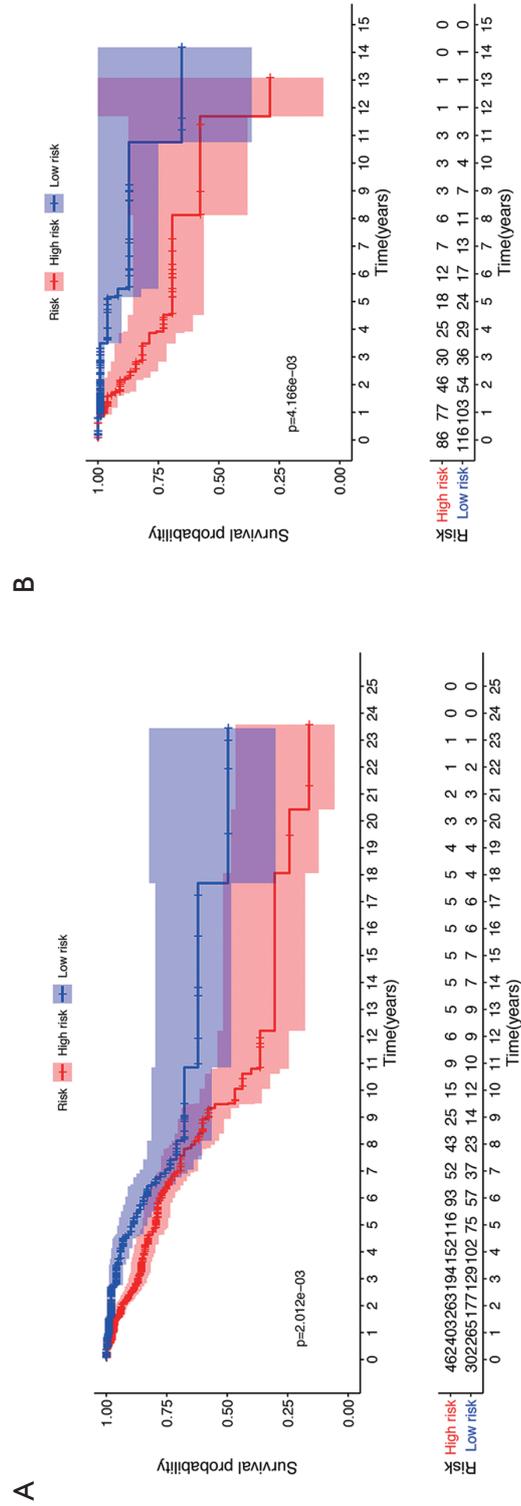
Figure S2 Kaplan-Meier plot of the risk scores in different pathological types. Kaplan-Meier plot of the risk scores in invasive ductal carcinoma (A) and invasive lobular carcinoma (B) in TCGA dataset. TCGA, The Cancer Genome Atlas.
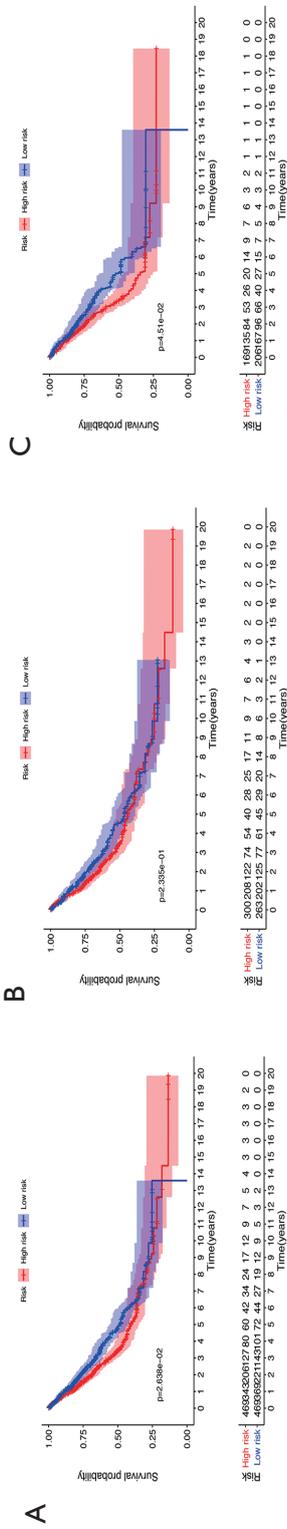
Figure S3 Kaplan–Meier plot of risk scores in lung cancer patients. Kaplan–Meier plot of the risk score in the entire dataset (A), males (B), and females (C).