Research Paper

Identification of a 15-pseudogene based prognostic signature for predicting survival and antitumor immune response in breast cancer

Liqiang Tan[1,2,*], **Xiaofang He**[3,4,*], **Guoping Shen**[3]

[1]Department of Medical Bioinformatics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China
[2]Center for Stem Cell Biology and Tissue Engineering, Key Laboratory for Stem Cells and Tissue Engineering, Ministry of Education, Sun Yat-sen University, Guangzhou 510080, China
[3]Department of Radiation Oncology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510080, China
[4]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA
*Equal contribution

Correspondence to: Guoping Shen; **email:** shenguop@mail.sysu.edu.cn

Keywords: pseudogene, breast cancer, survival, risk score, antitumor immune response
Received: March 3, 2020 **Accepted:** July 7, 2020 **Published:** December 16, 2020

ABSTRACT

Pseudogenes are noncoding RNAs that have been revealed to play critical roles in oncogenesis and tumor progression. However, their functional roles have not been comprehensively clarified in breast cancer. Here, we systematically analyzed the RNA sequencing data of 13931 pseudogenes in 775 breast cancer patients from The Cancer Genome Atlas dataset, and ultimately identified 15 prognostic pseudogenes by univariate Cox proportional hazard regression. A risk score model was constructed based on the prognostic pseudogenes via LASSO analysis and dichotomized patients into low- and high-risk subgroups. Patients in the high-risk group had a significantly shorter overall survival than those in the low-risk group. The prognostic value of these 15 pseudogenes and the risk score model were further validated in the European Genome-Phenome Archive dataset. Furthermore, we performed consensus clustering of the 15 prognostic pseudogenes and found that their expression pattern was significantly associated with tumor malignancy and host antitumor immune response, in terms of infiltrating immune cell compositions, antigen presenting genes expression, cytolytic activity and T-cell exhausted markers. This study indicated that these 15 prognostic pseudogenes were significantly correlated with tumor malignancy and host antitumor immune response in breast cancer, and might serve as potential targets for immunotherapy.

INTRODUCTION

Breast cancer is the most common malignant tumor and the second leading cause of death for females globally. Nowadays, the robust predictive factors for prognosis of breast cancer patients are two clinical features-tumor size and lymph node status at the time of detection [1]. Carcinomas with large tumor size or lymph node metastasis are usually associated with poor survival outcomes. However, breast cancers are well known as highly heterogenous neoplasms and driven by complex signaling pathways [2], which in part accounts for the fact that different therapeutic responses and then different survival outcomes can be observed even in patients diagnosed with the same breast cancer molecular subtype [3] and TNM stage. Therefore, looking for additional promising prognostic biomarkers, especially in the intrinsic molecular level [4, 5], is imperative so as to identify high-risk subgroups and make precise therapeutic strategies.

Nowadays, the standard of care for primary breast cancer is surgery, followed by chemotherapy, endocrine therapy, radiotherapy and targeted therapy on the basis of molecular subtypes and TNM stage. While in recent years, immunotherapy is emerging as a novel treatment modality due to the promising therapeutic effect of selective immune checkpoint inhibitors in combination with other strategies [6], especially monoclonal antibodies against programmed death 1 (PD-1), programmed death-ligand 1 (PD-L1) and cytotoxic T lymphocyte-associated protein-4 axes (CTLA-4) [7]. As of September, 2018, the number of registered trials that are open to breast cancer patients, which assess novel approaches by harnessing the immune system, has reached up to 285 [8–10]. At the current stage, the expression level of PD-L1 in tumor tissue is a commonly-used predictive marker for immune response [11, 12]. However, the predictive results were not satisfactory, which indicates that immune modulation is a complicated process and requires much more functional predictors [13–15]. Therefore, it is essential to identify robust biological predictive markers of immune response when conducting clinical trials of immunotherapy.

Pseudogenes are non-coding homologs of protein-coding genes, which are often caused by accumulation of multiple mutations within genes, and their products are nonfunctional [16]. Pseudogenes were once labeled as “genetic fossils” because of lack of protein-coding ability or cellular gene expression. However, due to the development of high-throughput sequencing technologies, pseudogenes have been revealed to participant in various biological functions by regulating their parental transcripts, acting as competitive endogenous RNAs (ceRNA) [17–19]. What’s more, the significance of pseudogenes in gene regulation has also been highlighted in tumorigenesis and tumor progression recently [20, 21], which was largely attributed to the finding that PTEN pseudogene 1 could upregulate his parental gene PTEN, a well-known tumor suppressor, via ceRNA mechanism and thus played a pivotal role in tumorigenesis in breast cancer [22]. However, until now, it has not been comprehensively clarified the prognostic effect of pseudogenes in patients with breast cancer, and their potential roles in host antitumor immune response remain largely unexplored.

Based on the concerns mentioned above, we systematically analyzed the RNA sequencing data of pseudogenes in 775 patients with breast cancer from The Cancer Genome Atlas (TCGA) dataset and eventually identified 15 prognostic indicators, which were further validated using the European Genome-Phenome Archive (EGA) dataset. A risk score model was constructed based on the prognostic pseudogenes, and their expression pattern was functionally annotated by Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses and Gene Set Enrichment Analysis (GSEA). Besides, we also investigated the association between the prognostic pseudogenes and the host antitumor immune response, in terms of tumor-infiltrating immune cell compositions, antigen presenting genes expression, immunomodulator genes expression and cytolytic activity, so as to provide potential predictive markers to immunotherapy in breast cancer.

RESULTS

Identification of 15 prognostic pseudogenes

Altogether, a whole list of 13931 pseudogenes were obtained from Vega databases and psiCube databases, of which 308 pseudogenes were available in TCGA datasets and thus included in the subsequent analyses. As results of univariate Cox proportional hazard regression indicated, a total of 15 pseudogenes were ultimately identified to be significantly associated with survival outcomes in TCGA dataset (Figure 1A), which was further verified in EGA dataset (Supplementary Figure 1).

Construction of a risk signature based on the prognostic pseudogenes

To improve the predictive effect of pseudogenes in the clinical outcomes of breast cancer, we applied the least absolute shrinkage and selection operator (LASSO) Cox regression algorithm to the 15 prognostic pseudogenes and constructed a risk signature based on the minimum criteria using TCGA data as the training set (Figure 1B, 1C) and EGA data as the validation set. The coefficients of the 15 pseudogenes were listed in Supplementary Table 1. The risk score was calculated according to survival risk score model formula. Then, the breast cancer patients were dichotomized into low or high-risk groups according to the median risk score. Results indicated that patients in high-risk group displayed significantly shorter overall survival than those in low-risk group (TCGA dataset, median overall survival 8.94 years vs. 10.85 years, log-rank test, $p = 0.0025$, Figure 1D; EGA dataset, median overall survival 10.79 years vs. 12.77 years, log-rank test, $p = 0.0313$, Figure 1E). The ROC curves showed that the risk score was good to predict survival rates with an AUC value of 0.769 in the training set (Figure 1F) and 0.778 in the validation set (Figure 1G). In addition,

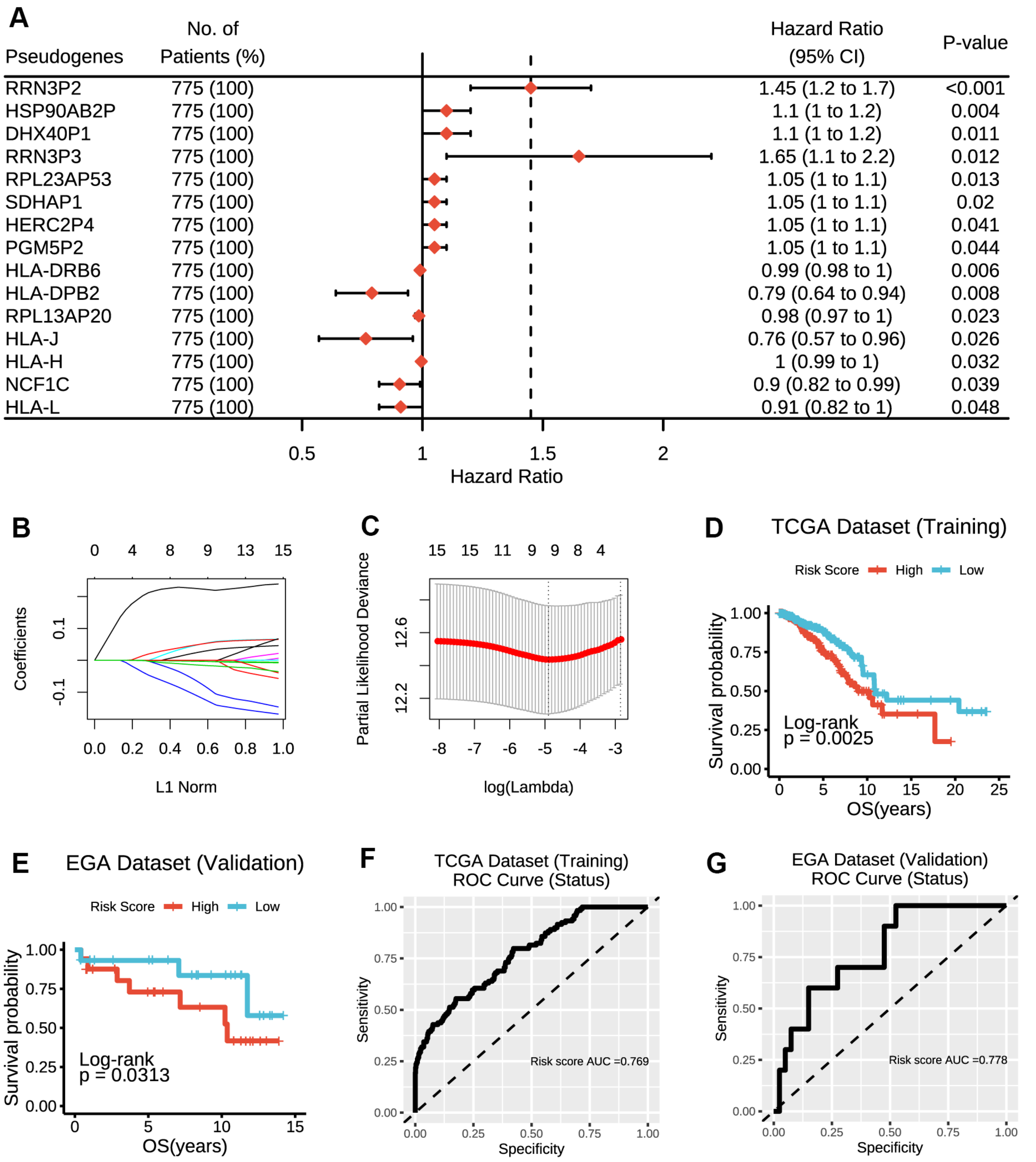


Figure 1. Construction of the risk score model based on prognostic pseudogenes. (**A**) The hazard ratios (HR), 95% confidence intervals (CI) calculated by univariate Cox proportional hazard regression of 15 prognostic pseudogenes using TCGA data. (**B**) LASSO coefficient profiles of 15 prognostic pseudogenes. (**C**) Ten-time cross-validation for tuning parameter selection in the LASSO model of 15 prognostic pseudogenes. (**D**) The breast cancer patients from TCGA dataset in high-risk group displayed significantly shorter overall survival than those in low-risk group (p = 0.0025). (**E**) The breast cancer patients from EGA dataset in high-risk group displayed significantly shorter overall survival than those in low-risk group (p = 0.0313). (**F**) The ROC curve and AUC for the risk score model in TCGA dataset. (**G**) The ROC curve and AUC for the risk score model in EGA dataset.

RRN3P2 and HLA-DRB6 were found to have significant associations with overall survivals. Patients with high expression of RRN3P2 had significantly shorter survival than those with low expression (median overall survival 8.94 years vs. 11.69 years, log-rank test, p = 0.0088, Supplementary Figure 2A), indicating that high expression of RRN3P2 might correlate with high malignancy. On the contrary, patients with high expression of HLA-DRB6 had significantly longer survival than those with low expression (median overall survival 20.42 years vs. 10.24 years, log-rank test, p = 0.014, Supplementary Figure 2B). Besides, RPL23AP53, HLA-DRB6, RPL13AP20, NCF1C and HLA-L were found to have significant associations with overall survivals in EGA dataset (Supplementary Figure 3A–3E).

Expressions of prognostic pseudogenes significantly associated with different clinicopathological features and survival outcomes

The distribution of the risk scores, overall survival, and corresponding pseudogene expression profiles in TCGA dataset were demonstrated in Figure 2A. Heatmap indicated that NCF1C, HLA-DRB6, HLA-DPB2, HLA-J, HLA-H, HLA-L and RPL13AP20 displayed high expressions in the low-risk group, and thus were categorized as tumor-suppressor pseudogenes in the current study. On the other hand, the remaining pseudogenes (PGM5P2, HERC2P4, HSP90AB2P, DHX40P1, RRN3P3, RRN3P2, SDHAP1 and RPL23AP53) displayed high expressions in the high-risk group and thus were categorized as oncogene pseudogenes in the current study. Besides, we also found that the risk score and prognostic pseudogenes were closely related to different clinicopathological features of breast cancer patients. The low-risk group was significantly associated with ER status (p = 8e-08), PR status (p = 6e-04), more basal-like molecular subtype (p = 4e-06) and lower lymph node stage (p = 0.037) compared with high-risk group (Supplementary Table 2). Basal-like subtype had significantly higher expressions of NCF1C, HLA-H, RPL13AP20 RRN3P3 and SDHAP1, but lower expressions of PGM5P2, HSP90AB2P and DHX40P1 than other subtypes (Figure 2B). In addition, patients with lymph node metastasis had significantly higher expressions of PGM5P2, HERC2P4 and RRN3P2 but lower expressions of HLA-H and RPL13AP20 than those without lymph node metastasis (Figure 2C). There were no significant differences in the expressions of the 15 prognostic pseudogenes between patients with or without distant metastasis (Figure 2D).

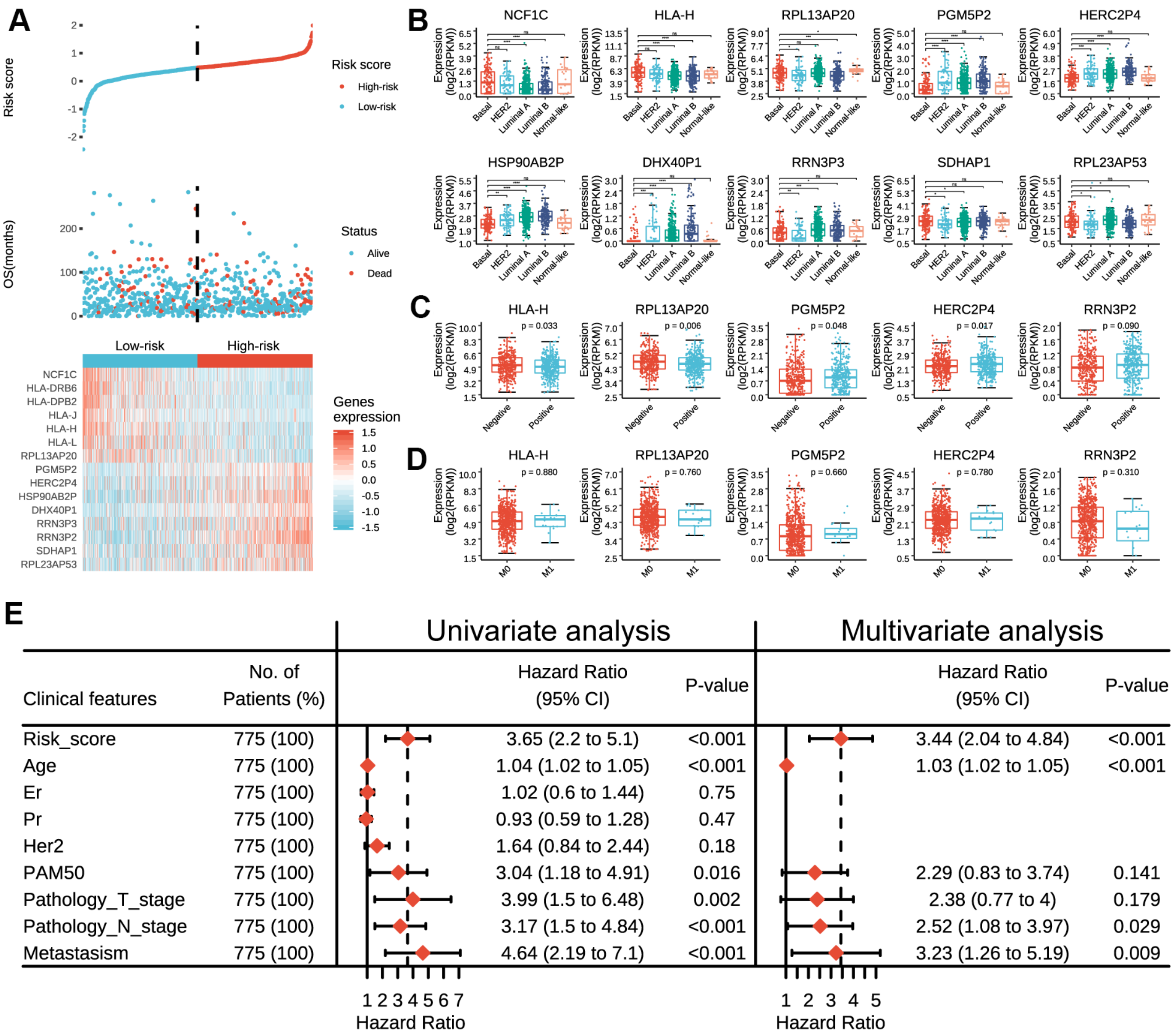
Furthermore, univariate Cox regression analysis demonstrated that risk score, age, PAM50, pathology T stage, pathology N stage and metastasis status were all correlated with the overall survival. When including these factors in the multivariate Cox regression, we found that risk score (p < 0.001), age (p < 0.001), pathology N stage (p = 0.029) and metastasis status (p = 0.009) remained significantly associated with the clinical outcome (Figure 2E), which indicated that the risk score derived from these 15 pseudogenes could independently predict prognosis in breast cancer patients.

Consensus clustering of prognostic pseudogenes identified two clusters highly consistent with that of the risk score

Considering the large amounts of prognostic pseudogenes, we adopted dimensionality reduction analysis through consensus clustering of the 15 prognostic pseudogenes in the subsequent analysis. According to the expression similarity of pseudogenes, k = 2 seemed to be the optimal selection when clustering stability increased from k = 2 to 10 in the TCGA dataset (Figure 3A–3C). Therefore, we divided the 775 breast cancer patients into two subgroups by making 2 as the k value, namely, P1 (Patients subgroup 1) and P2 (Patients subgroup 2). Kaplan-Meier analysis revealed that patients in P2 subgroup had significantly longer overall survival than those in P1 subgroup (median overall survival 17.69 years vs. 10.24 years, log-rank test, p = 0.045, Figure 3D). The expression pattern of the 15 prognostic pseudogenes across P1 and P2 subgroups was displayed in Figure 3E. Results indicated that P1 subgroup had lower expressions of tumor-suppressor pseudogenes and higher expressions of oncogene pseudogenes, while P2 subgroup showed the opposite trends. What's more, compared with P1, P2 subgroup had significantly higher expressions of 6 tumor-suppressor pseudogenes (NCF1C, p < 2e-16; HLA-DRB6, p < 2e-16; HLA-DPB2, p = 4e-12; HLA-J, p < 2e-16; HLA-H, p < 2e-16; HLA-L, p < 2e-16), and significantly lower expressions of 6 oncogene pseudogenes (PGM5P2, p = 0.002; HERC2P4, p = 0.003; HSP90AB2P, p = 6e-07; DHX40P1, p = 7e-04; RRN3P2, p = 0.014; RPL23AP53, p = 2e-04) (Figure 3F). In addition, we found that the P1 and P2 subgroup were also significantly associated with the clinicopathological features. P2 subgroup was significantly associated with ER status (p = 0.028), more basal-like molecular subtype (p = 0.002) and lower lymph node stage (p = 0.047) compared with P1 (Supplementary Table 3). These findings were highly consistent with those of the risk score mentioned above, which indicated that the expression pattern of prognostic pseudogenes was significantly associated with tumor survival.

Expression pattern of prognostic pseudogenes was closely associated with malignancy of breast cancer

To better illuminate the association between prognostic pseudogenes and malignancy of breast cancer, we identified the differentially expressed genes between P1 and P2 subgroups and annotated their functions using GO, KEGG pathway analysis and GSEA. GO pathway analyses revealed that upregulated genes in P2 were significantly enriched in tumor-related biological processes and pathways (Figure 4A), including regulation of JAK-STAT cascade, pattern recognition receptor signaling pathway and type I interferon signaling pathway. KEGG pathway analysis indicated that upregulated genes in P2 were enriched in TNF, JAK-STAT, IL-17, B cell receptor, Chemokine,



Clinical features	No. of Patients (%)	Univariate Hazard Ratio (95% CI)	Univariate P-value	Multivariate Hazard Ratio (95% CI)	Multivariate P-value
Risk_score	775 (100)	3.65 (2.2 to 5.1)	<0.001	3.44 (2.04 to 4.84)	<0.001
Age	775 (100)	1.04 (1.02 to 1.05)	<0.001	1.03 (1.02 to 1.05)	<0.001
Er	775 (100)	1.02 (0.6 to 1.44)	0.75		
Pr	775 (100)	0.93 (0.59 to 1.28)	0.47		
Her2	775 (100)	1.64 (0.84 to 2.44)	0.18		
PAM50	775 (100)	3.04 (1.18 to 4.91)	0.016	2.29 (0.83 to 3.74)	0.141
Pathology_T_stage	775 (100)	3.99 (1.5 to 6.48)	0.002	2.38 (0.77 to 4)	0.179
Pathology_N_stage	775 (100)	3.17 (1.5 to 4.84)	<0.001	2.52 (1.08 to 3.97)	0.029
Metastasism	775 (100)	4.64 (2.19 to 7.1)	<0.001	3.23 (1.26 to 5.19)	0.009

Figure 2. Expressions of prognostic pseudogenes in breast cancer by different clinicopathological features in TCGA dataset. (**A**) The distribution of risk score, vital status and the expression pattern of 15 prognostic pseudogenes in 775 breast cancer patients. The risk scores are arranged in ascending order from left to right. (**B**) Expression levels of NCF1C, HLA-DRB6, HLA-DPB2, HLA-J, HLA-H, HLA-L, RPL13AP20, PGM5P2, HERC2P4, HSP90AB2P, DHX40P1, RRN3P3, RRN3P2, SDHAP1 and RPL23AP53 across different breast cancer subtypes. (**C**) Expression levels of HLA-H, RPL13AP20, PGM5P2, HERC2P4 and RRN3P2 in patients with or without lymph node metastasis. (**D**) Expression levels of HLA-H, RPL13AP20, PGM5P2, HERC2P4 and RRN3P23 in patients with or without distant metastasis. (**E**) Univariate and multivariate Cox regression analyses of the association between clinicopathological factors (including the risk score) and overall survival of breast cancer patients. ns denotes no significance, *** denotes P < 0.001 and **** denotes P < 0.0001.

NF-kappa B, T cell receptor signaling pathway and PD-L1expression and PD-1 checkpoint pathway in cancer (Figure 4A). Furthermore, “METASTASIS”, “SMAD”, “SIGNALING_BY_WNT_IN_CANCER” and “PI3KCI_AKT” were strikingly enriched in P1 subgroup indicating by GSEA (Figure 4B), while the hallmarks of “INTERFERON GAMMA RESPONSE”, “IL2 STAT5 SIGNALING”, “IL6 JAK STAT3 SIGNALING”, and “TNF SIGNALING” were remarkably enriched in P2 subgroup (Figure 4C). All these results partially clarified the mechanism underlying the prognostic effect of pseudogenes in breast cancer.

Expression pattern of prognostic pseudogenes was significantly correlated with antitumor immune response

To investigate the correlation between the expression pattern of pseudogenes and antitumor immune response in breast cancer, we assessed the immune cell infiltration using CIBERSORT, and estimated the expressions of antigen presentation genes, cytolytic genes and immunomodulator genes in tumor tissues between P1 and P2 subgroups.

As summarized in Figure 5A, P2 subgroup had significantly higher number of tumor-infiltrating CD8$^+$ T cells, CD4$^+$ T cells, helper T cells, activated natural killer cells and lower number of M2 macrophage than P1, suggesting an enhanced immunosurveillance in P2 subgroup. Of note, the regulatory T cell, a well-known member of suppressor T cell population, displayed a significantly higher fraction in P2 than in P1 subgroup.

As for antigen presenting genes, we found that P2 had dramatically higher expressions of HLA-A, HLA-B, HLA-C, TAP1 and B2M than P1 (Figure 5B), which are main regulatory genes for human MHC class I cell surface receptors and thus activate cytotoxic T cells. Besides, P2 subgroup was also associated with higher expressions of GZMA and RPRF (Figure 5C), two

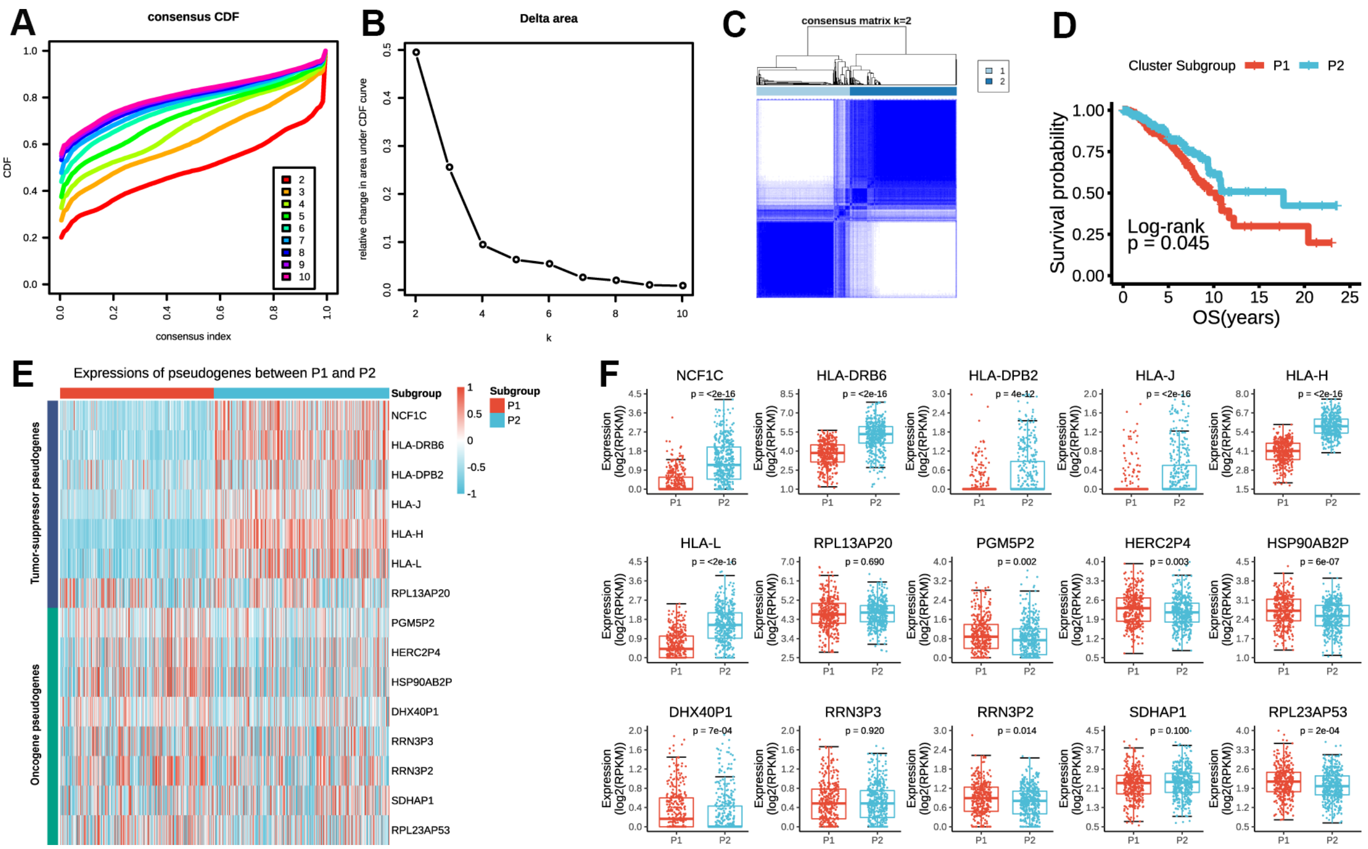


Figure 3. P1/2 subgroups identified by consensus clustering of the 15 prognostic pseudogenes in TCGA dataset. (**A**) Consensus clustering cumulative distribution function (CDF) for k = 2 to 10. (**B**) Relative change in area under CDF curve for k = 2 to 10. (**C**) Consensus clustering of 775 breast cancers with k = 2. (**D**) Kaplan-Meier overall survival (OS) curves for patients in P1 and P2 subgroups. (**E**) The expression pattern of prognostic pseudogenes between P1 and P2 subgroups. (**F**) Expression levels of NCF1C, HLA-DRB6, HLA-DPB2, HLA-J, HLA-H, HLA-L, RPL13AP20, PGM5P2, HERC2P4, HSP90AB2P, DHX40P1, RRN3P3, RRN3P2, SDHAP1 and RPL23AP53 between P1 and P2 subgroups.

important regulatory genes for cytolytic activity. These findings partially accounted for the above results that tumors in P2 subgroup had stronger immunogenicity and therefore presented higher numbers of active immune cell infiltrations.

However, in terms of immunomodulator genes, P2 subgroup was significantly associated with higher expressions of PD-1, PD-L1, PD-L2, LAG3, TIM3, CTLA-4, CCR4 and TIGIT than P1 subgroup (Figure 5D), all of which are key genes of T-cell exhaustion markers. Besides, the expressions of CD27 and ICOS were also significantly higher in P2 than P1 subgroup. Therefore, it indicated that prognostic pseudogenes played a critical role in host antitumor response and might serve as potential targets for immunotherapy.

Pseudogene-miRNA-target gene regulatory networks

To elucidate the underlying mechanism how pseudogenes regulated anti-tumor immune response, we built a pseudogene-miRNA-target gene regulatory network. Potential miRNAs binding to the 15 pseudogenes were identified using the dreamBase database (Supplementary Table 4). Pearson correlation analysis was used to calculate expression correlations between each pseudogene and its miRNA target genes. Target genes with | r | ≥ 0.3 and P < 0.05 were picked up (Supplementary Table 5). Ultimately, 4 tumor-suppressor pseudogenes (HLA-J, HLA-H, HLA-L and RPL13AP20) together with 13 microRNAs and 19 targeted genes, and 5 oncogene pseudogenes

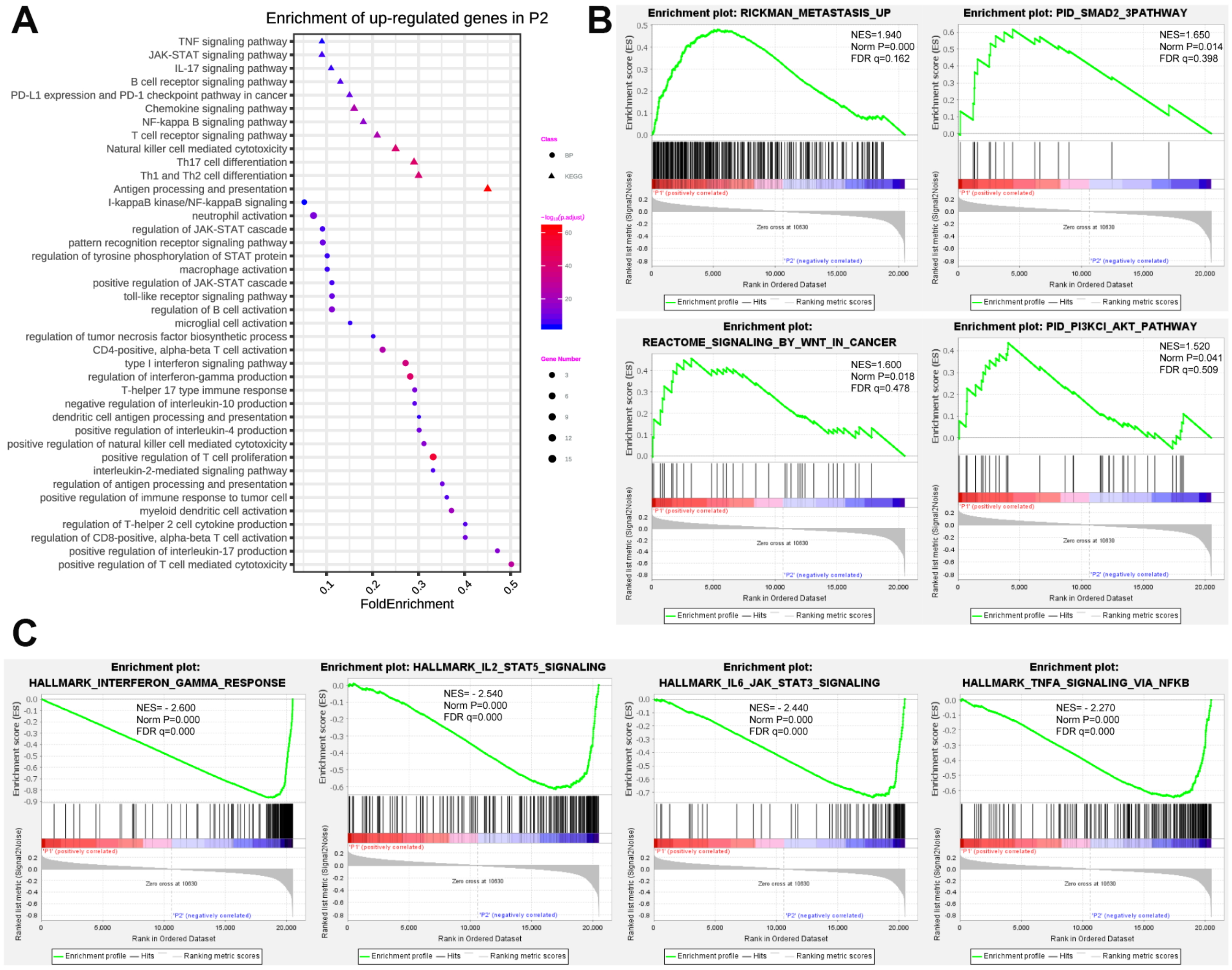


Figure 4. Functional annotation of differentially expressed genes in P1/P2 subgroups. (**A**) Functional annotation of up-regulated genes in P2 subgroup compared with P1 by using GO in terms of biological process (BP) and KEGG signaling pathway. (**B**) GSEA revealed that up-regulated genes in P1 subgroup were enriched for hallmarks of malignant tumors. (**C**) GSEA revealed that up-regulated genes in P2 subgroup were enriched for hallmarks of antitumor immune response.

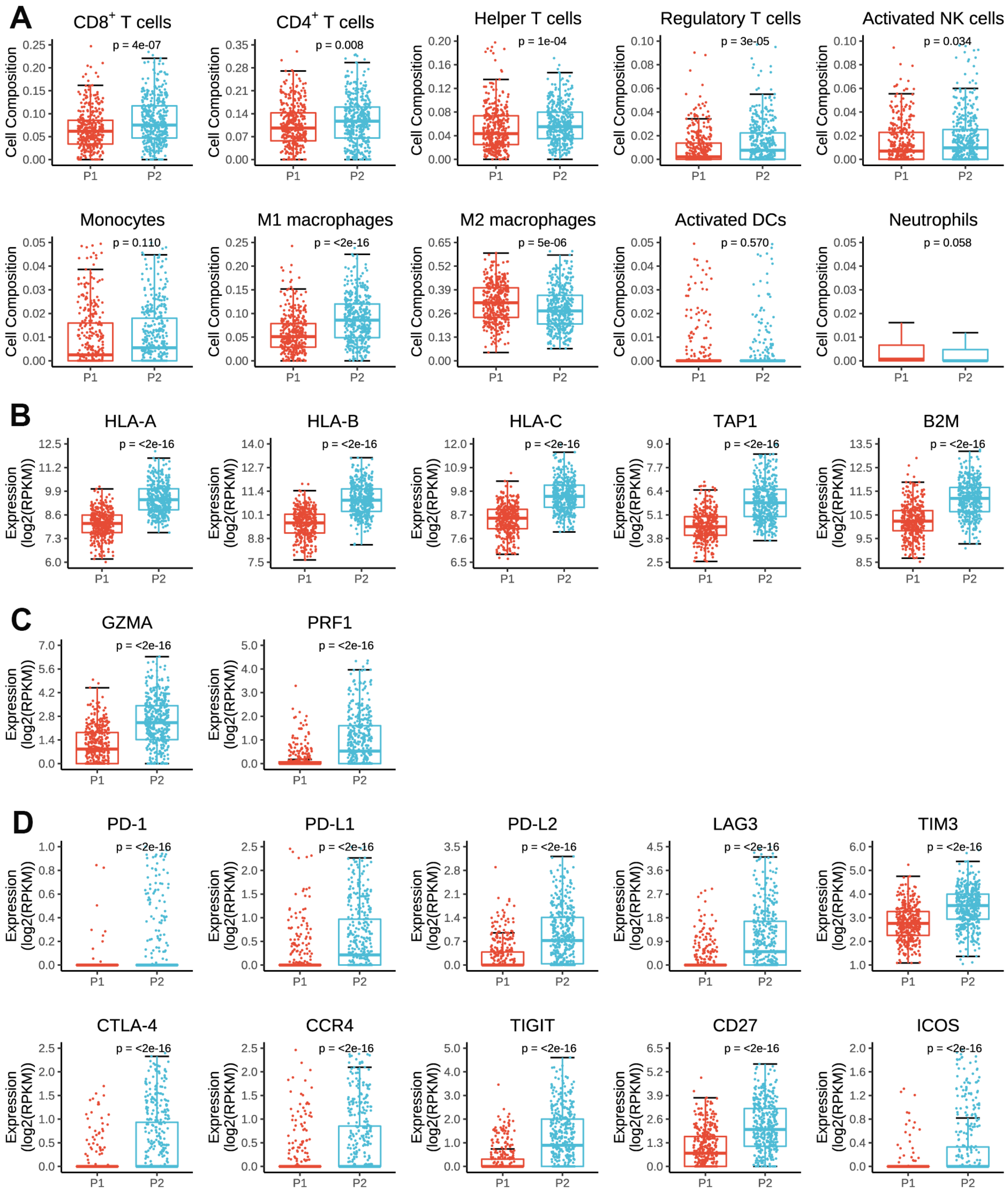


Figure 5. Immune cell infiltration and expressions of antigen presenting genes, immune cytolysis genes and immunomodulator genes in tumor tissues by P1/P2 subgroups. (**A**) Comparisons of cell composition fraction of CD8+ T cells, CD4+ T cells, helper T cells, regulatory T cells, activated natural killer (NK) cells, monocytes, M1 macrophages, M2 macrophages, activated dendritic cells (activated DCs) and neutrophils between P1 and P2 subgroups. (**B**) Expressions of HLA-A, HLA-B, HLA-C, TAP1 and B2M between P1 and P2 subgroups. (**C**) Expressions of GZMA and PRF1 between P1 and P2 subgroups. (**D**) Expressions of PD-1, PD-L1, PD-L2, LAG3, TIM3, CTLA-4, CCR4, TIGIT, CD27 and ICOS between P1 and P2 subgroups.

(HSP90AB2P, DHX40P1, RRN3P2, SDHAP1 and RPL23AP53) together with 35 microRNAs and 43 targeted genes, were used to construct the pseudogene-miRNA-target gene regulatory networks and visualized using Cytoscape (Figure 6). As results indicated, pseudogene HLA-L upregulated the expression of PD-L1 by competitively binding hsa-miR-124-3p, which explained the higher expression of PD-L1 in P2 subgroup. Pseudogene HLA-H, acting as decoy of has-miR-140-3p, upregulated the expression of CD38 and then upregulated the infiltrations of many immune cells (including CD4$^+$ T, CD8$^+$ T, B lymphocytes and natural killer cells) by signal transduction and calcium signaling (Figure 6A). Other pseudogenes played regulatory roles in signaling pathways involving cell proliferation, oncogenic transformation, cell survival, cell migration, and intracellular protein trafficking as ceRNAs (Figure 6B). The pseudogene-miRNA-target gene regulatory networks partially clarified the mechanism how pseudogenes participated in regulating the antitumor immune response in breast cancer.

DISCUSSION

In the current study, 15 pseudogenes were identified as promising prognostic indicators for breast cancer by univariate Cox regression analysis and classified into tumor-suppressor pseudogenes (NCF1C, HLA-DRB6, HLA-DPB2, HLA-J, HLA-H, HLA-L, RPL13AP20) and oncogene pseudogenes (PGM5P2, HERC2P4, HSP90AB2P, DHX40P1, RRN3P3, RRN3P2, SDHAP1, RPL23AP53) based on their different effects in clinical outcomes using TCGA dataset. Then a risk score model was constructed based on the 15 prognostic pseudogenes, and was found good in predicting prognosis in breast cancer. The prognostic value for these 15 pseudogenes and the risk score signature was further validated in EGA dataset. What's more, we also found that the expression pattern of these 15 prognostic pseudogenes was significantly associated with antitumor immune response in terms of tumor-infiltrating immune cell compositions, antigen presenting genes expression, immunomodulator genes expression and cytolytic activity. Pseudogene-miRNA-target gene regulatory networks were further performed to elucidate the underlying mechanisms. To the best of our knowledge, this is the first study to systemically clarify the prognostic value of pseudogenes in breast cancer, and their regulatory roles in host antitumor immune response.

Pseudogenes, belonging to the non-coding RNA family, are pervasively transcribed in the genome [23]. The noncoding transcripts range from 100 bp to approximately 100 kilobases (kb) in length and lack significant open reading frames, which once mislead people to consider pseudogenes as "genetic fossils". However, recent evidence suggests that pseudogenes can play important regulatory functions in diverse human diseases [24]. They were found to contain miRNA-binding elements and thus increase their parental transcripts by acting as competitive endogenous RNAs (ceRNA) [25, 26]. This significant finding worked as a strong cornerstone for studying the biological roles of pseudogenes in cancer.

Although currently there are no published studies concerning the prognostic effects of pseudogenes in breast cancer, previous studies have indicated the crucial roles of pseudogenes in tumorigenesis, tumor development and progression of other malignant tumors. For instant, pseudogene PTENP1 could suppress the progression of clear-cell renal cell carcinoma by functioning as a ceRNA [27]. Pseudogenes PKMP3, AC027612.4, HILS1, RP5-1132H15.3 and HSPB1P1 were identified as prognostic predictors for lower-grade gliomas [28]. In addition, pseudogenes ANXA2P2, EEF1A1P9, FER1L4, HILS1, and RAET1K were found to be significantly correlated with glioma survival [29]. What' more, pseudogene RNA5SP141 was able to strongly enhance the RIG-I-mediated antiviral immunity response to herpes simplex virus 1 [30]. In the current study, we identified 15 prognostic pseudogenes that significantly associated with clinical outcomes in breast cancer. They were further classified into two functional subgroups, tumor-suppressor pseudogenes (NCF1C, HLA-DRB6, HLA-DPB2, HLA-J, HLA-H, HLA-L, RPL13AP20) and oncogene pseudogenes (PGM5P2, HERC2P4, HSP90AB2P, DHX40P1, RRN3P3, RRN3P2, SDHAP1, RPL23AP53) based on their different effects in clinical outcomes. Then we constructed a risk score model based on the 15 prognostic pseudogenes by LASSO Cox regression, which was found good in predicting prognosis in breast cancer. All in all, our study provides promising prognostic predictors for breast cancer patients, which can better execute the principle of precise medicine.

To the best of our knowledge, this is the first study concerning the correlation between pseudogenes and host antitumor immune response in breast cancer. Surprisingly, our study revealed that the expression pattern of the 15 prognostic pseudogenes was significantly associated with active tumor-infiltrating CD8$^+$ T cells, CD4$^+$ T cells, helper T cells and activated natural killer cells, as well as the expressions of HLA-A, HLA-B, HLA-C, TAP1, B2M, GZMA and RPRF. What's more, T cell exhausted markers including PD-1, PD-L1, PD-L2, LAG3, TIM3, CTLA-4, CCR4 and TIGIT were also significantly associated with the expression pattern of pseudogenes. In addition, pseudogene-miRNA-target gene regulatory

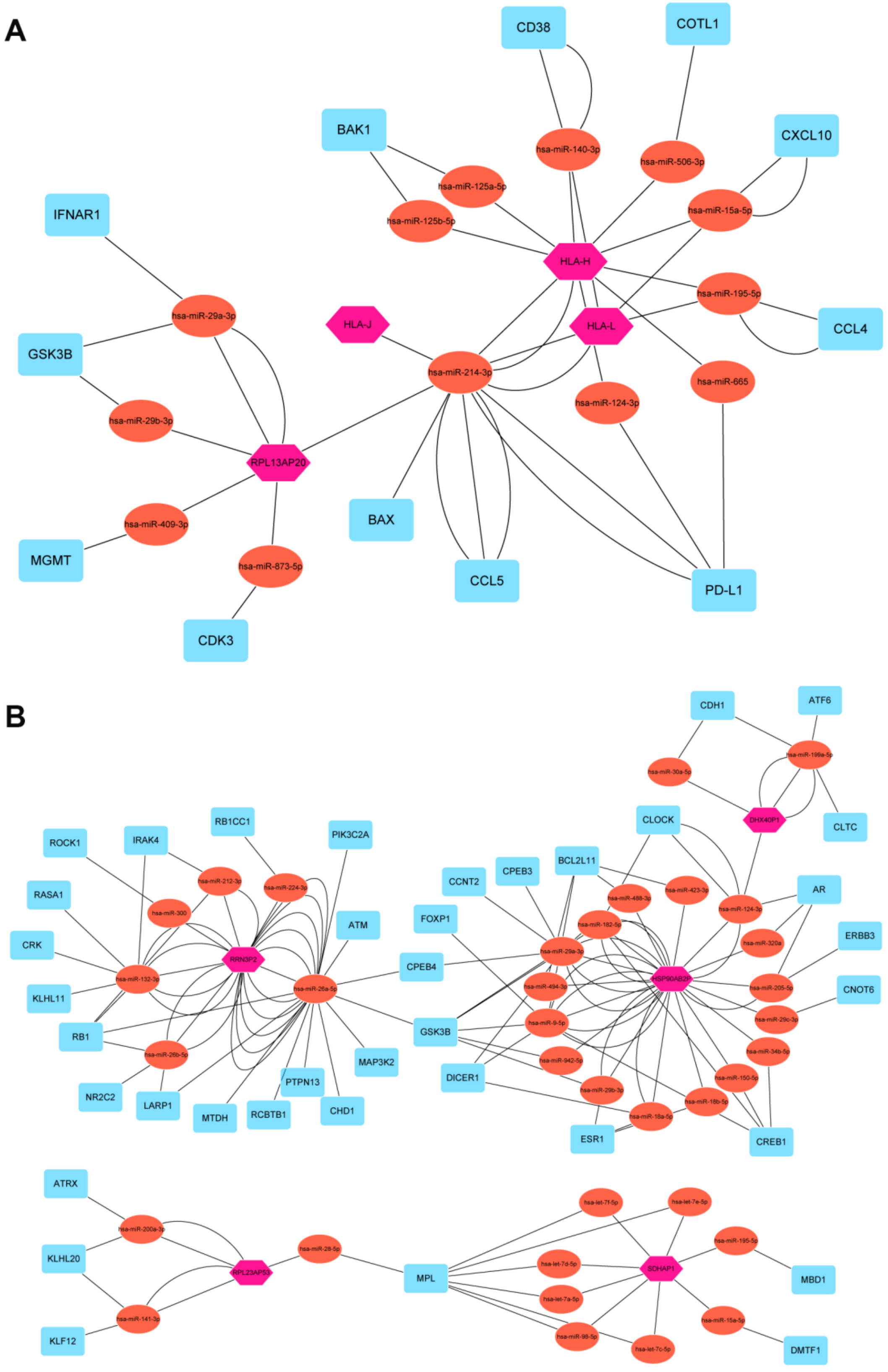


Figure 6. Pseudogene-miRNA-target gene regulatory networks. Nine pseudogenes together with binding miRNAs and target genes with |r| ≥ 0.3 and P < 0.05 were used to construct the pseudogene-miRNA-target gene regulatory networks by subgroups of tumor-suppressor pseudogenes (**A**) and oncogene pseudogenes (**B**). Pink hexagons represented pseudogenes, which are located at the cores of the networks. Tomato ellipses and blue round rectangles stand for binding miRNAs and target genes, respectively.

networks were further performed to elucidate the underlying mechanisms and demonstrated 4 tumor-suppressor pseudogenes (HLA-J, HLA-H, HLA-L and RPL13AP20) together with 13 microRNAs and 19 targeted genes, and 5 oncogene pseudogenes (HSP90AB2P, DHX40P1, RRN3P2, SDHAP1 and RPL23AP53) together with 35 microRNAs and 43 targeted genes as main regulatory factors. This large network could provide robust evidence for the further study about the biological roles of pseudogenes in host antitumor immune response in breast cancer.

One limitation of this study needs to be taken into consideration. All the analyses in the current study were based on the bioinformatics tools, therefore, further experimental validation is warranted to confirm the results of our study.

To sum up, we identified 15 prognostic pseudogenes and demonstrated that their expression pattern was significantly correlated with the clinicopathological features, survival outcomes and expressions of immunomodulator genes in breast cancer. This current study provided comprehensive evidence for further study of pseudogenes in breast cancer, and shed new light on the epigenetic regulation of antitumor immune response.

MATERIALS AND METHODS

Data sources

Genome and transcript sequences and annotation were obtained from the human genome (GRCh37), version 19 (Ensembl 74) (https://www.encodeproject.org/). A list of pseudogenes was collected from Vega databases (http://vega.archive.ensembl.org/index.html) and psiCube databases according to online pseudogene posted data (http://pseudogene.org/) [31, 32]. The breast cancer gene expression data and corresponding clinical information were obtained from TCGA data portal (http://firebrowse.org/) and EGA dataset (https://ega-archive.org/) by access number (EGAS00001001908). Altogether, 775 samples from TCGA and 50 samples from EGA with pseudogene expression data and corresponding clinical data were included. Immune cell fraction data were downloaded through CIBERSORT (https://cibersort.stanford.edu/) [33, 34]. The antigen presenting genes and immunomodulator genes were obtained from TCIA (https://tcia.at/home) [35]. miRNAs binding to pseudogenes were extracted from the dreamBase database (http://rna.sysu.edu.cn/dreamBase/index.php) [36]. miRNA target genes were identified using the miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/php/index.php) [37].

Screening for prognostic pseudogenes by cox proportional hazard regression analysis

Since most of the pseudogenes were not expressed, we first excluded those with the expression values (RPKM) less than 1. Then, univariate Cox proportional hazard regression was performed to screen for the candidate pseudogenes closely associated with overall survival. After these two steps, 15 pseudogenes were identified significantly associated with survival outcomes in TCGA dataset ($P < 0.05$) and further validated in EGA dataset.

Construction of the risk score model

Based on LASSO Cox regression algorithm [38], a L1-penalized regression on the strength of the highest lambda value selected by means of a 1,000 cross-validations (‘1-se’ lambda) was conducted to further identify the regression coefficients of the 15 prognostic pseudogenes. Then a survival risk score model was established by the LASSO coefficients (β) as follows:

$$\text{Risk score} = \sum\nolimits_{i=1}^{n} Geneexpri \times \beta i.$$

The breast cancer patients were divided into low or high-risk groups based on the median risk score. The receiver operating characteristic (ROC) curve and area under the curve (AUC) were conducted to estimate the prediction accuracy of the risk score model. Each prognostic pseudogene was dichotomized into low or high expression level, with cut-off value defining as the median expression value. Kaplan-Meier plots and Log-rank test were utilized to evaluate and compare the survival rate between subgroups. All the analyses mentioned above were performed using TCGA data as the training set and EGA data as the validation set. Univariate and multivariate Cox regression analyses were carried out to determine the prognostic value of the risk score and various clinical characteristics.

Consensus clustering analysis

To investigate the functional roles of pseudogenes in breast cancer, we clustered the patients into different subgroups by the R package “ConsensusClusterPlus” (50 iterations, resample rate of 80%, and Pearson correlation) based on the expression levels of the prognostic pseudogenes in TCGA dataset [39].

Functional analysis of the prognostic pseudogenes

To better understand the association between prognostic pseudogenes and malignancy of breast cancer, GO pathway analysis, KEGG analysis and GSEA [40] were carried out to functionally annotate genes that differentially expressed in different subgroups by using the R package "clusterProfiler" [41].

Immune cell infiltration, immune response and immune cytolysis

CIBERSORT [33], a bioinformatic deconvolution algorithm to calculate immune cell composition from their gene expression profiles, was used to assess tumor-infiltrating cell compositions in diverse tumors [42]. The immune cell fractions, expressions of antigen presenting genes, immunomodulator genes [43] and immune cytolysis genes [44] were compared in different subgroups by Wilcoxon signed-rank test.

Pseudogene-miRNA-target gene regulatory networks

miRNAs binding to prognostic pseudogenes were obtained from the dreamBase database [36]. miRNA target genes with at least one strong experimental method (reporter assay or western blot) were extracted by the miRTarBase [37]. Pearson analysis was conducted to calculate expression correlation between pseudogenes and miRNA target genes. Target genes conforming to $| r | \geq 0.3$ and $P < 0.05$ were selected and applied to construct pseudogene-miRNA-target gene regulatory networks using Cytoscape 3.7.1.

Statistical analysis

One-way ANOVA and t test were carried out to compare the expression levels of prognostic pseudogenes in different subgroups differentiated by lymph node status, molecular subtypes and distant metastasis status. Chi-square test was used to evaluate the differences of clinicopathological characteristics between subgroups identified by consensus clustering of pseudogenes. All statistical analyses were performed using R software (http://www.r-project.org/) and Bioconductor (http://bioconductor.org/). A two-sided p value < 0.05 was considered statistically significant in all analyses.

Abbreviations

PD-1: programmed cell death 1; PD-L1: programmed cell death 1 ligand 1; PD-L2: programmed cell death 1 ligand 2; CTLA-4: cytotoxic T-lymphocyte-associated protein 4; ceRNA: competitive endogenous RNA; TCGA: The Cancer Genome Atlas; EGA: European Genome-Phenome Archive; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; GSEA: Gene Set Enrichment Analysis; LASSO: the least absolute shrinkage and selection operator; ROC: the receiver operating characteristic curve; AUC: area under the curve; NCF1C: neutrophil cytosolic factor 1 pseudogene; HLA: human leukocyte antigen; RPL13AP20: L13P family of ribosomal proteins pseudogene 20; PGM5P2: Phosphoglucomutase 5 Pseudogene 2; HERC2P4: HECT And RLD Domain Containing E3 Ubiquitin Protein Ligase 2 pseudogene 4; HSP90AB2P: Heat Shock Protein 90 Alpha Family Class B Member 2 Pseudogene; DHX40P1: DEAH-Box Helicase 40 pseudogene 1; RRN3: RRN3 homolog, RNA polymerase I transcription factor; SDHAP1: Succinate Dehydrogenase Complex Flavoprotein Subunit A Pseudogene 1; RPL23AP53: L23P family of ribosomal proteins pseudogene 53; P1: patient subgroup 1; P2: patient subgroup 2; GZMA: Granzyme A; PRF1: Perforin 1; LAG3: lymphocyte-activation gene 3; TIM3: T-cell immunoglobulin and mucin-domain containing-3; CCR4: C–C chemokine receptor type 4; TIGIT: T cell immunoreceptor with Ig and ITIM domains; ICOS: inducible T-cell costimulatory.

AUTHOR CONTRIBUTIONS

LT, XH and GS conceived of the study. LT and XH carried out its design, analyzed and interpreted the data. LT and XH wrote and edited the paper. GS revised the paper. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors gratefully acknowledge contributions from the TCGA Network, EGA database, psiCube databases, CIBERSORT networks, TCIA, dreamBase database, miRTarBase database and Cytoscape.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

FUNDING

This work was supported by National Natural Science Foundation of China (81201747), Natural Science Foundation of Guangdong Province, China (S2012040006323, 2014A030313023), and the open Foundation of State Key Laboratory of Oncology in South China (HN2013-07). The funding sources had no involvements in the study design, collection, analysis and interpretation of data, writing of the article or decision to submit the article for publication.

REFERENCES

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer. 2010; 127:2893–917.

https://doi.org/10.1002/ijc.25516 PMID:21351269

2. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, McLellan M, Yau C, Kandoth C, Bowlby R, Shen H, Hayat S, et al, and TCGA Research Network. Comprehensive molecular portraits of invasive lobular breast cancer. Cell. 2015; 163:506–19. https://doi.org/10.1016/j.cell.2015.09.033 PMID:26451490

3. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490:61–70. https://doi.org/10.1038/nature11412 PMID:23000897

4. Goldhirsch A, Wood WC, Coates AS, Gelber RD, Thürlimann B, Senn HJ, and Panel members. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St. Gallen international expert consensus on the primary therapy of early breast cancer 2011. Ann Oncol. 2011; 22:1736–47. https://doi.org/10.1093/annonc/mdr304 PMID:21709140

5. Hancock BA, Chen YH, Solzak JP, Ahmad MN, Wedge DC, Brinza D, Scafe C, Veitch J, Gottimukkala R, Short W, Atale RV, Ivan M, Badve SS, et al. Profiling molecular regulators of recurrence in chemorefractory triple-negative breast cancers. Breast Cancer Res. 2019; 21:87. https://doi.org/10.1186/s13058-019-1171-7 PMID:31383035

6. Li X, Shao C, Shi Y, Han W. Lessons learned from the blockade of immune checkpoints in cancer immunotherapy. J Hematol Oncol. 2018; 11:31. https://doi.org/10.1186/s13045-018-0578-4 PMID:29482595

7. Li Y, Li F, Jiang F, Lv X, Zhang R, Lu A, Zhang G. A mini-review for cancer immunotherapy: molecular understanding of PD-1/PD-L1 pathway & translational blockade of immune checkpoints. Int J Mol Sci. 2016; 17:1151. https://doi.org/10.3390/ijms17071151 PMID:27438833

8. Bu X, Yao Y, Li X. Immune checkpoint blockade in breast cancer therapy. Adv Exp Med Biol. 2017; 1026:383–402. https://doi.org/10.1007/978-981-10-6020-5_18 PMID:29282694

9. Hu ZI, Ho AY, McArthur HL. Combined Radiation Therapy and Immune Checkpoint Blockade Therapy for Breast Cancer. Int J Radiat Oncol Biol Phys. 2017; 99:153–164. https://doi.org/10.1016/j.ijrobp.2017.05.029 PMID:28816141

10. Jiang P, Gao W, Ma T, Wang R, Piao Y, Dong X, Wang P, Zhang X, Liu Y, Su W, Xiang R, Zhang J, Li N. CD137 promotes bone metastasis of breast cancer by enhancing the migration and osteoclast differentiation of monocytes/macrophages. Theranostics. 2019; 9:2950–66. https://doi.org/10.7150/thno.29617 PMID:31244935

11. Hargadon KM, Johnson CE, Williams CJ. Immune checkpoint blockade therapy for cancer: an overview of FDA-approved immune checkpoint inhibitors. Int Immunopharmacol. 2018; 62:29–39. https://doi.org/10.1016/j.intimp.2018.06.001 PMID:29990692

12. Khan Z, Hammer C, Guardino E, Chandler GS, Albert ML. Mechanisms of immune-related adverse events associated with immune checkpoint blockade: using germline genetics to develop a personalized approach. Genome Med. 2019; 11:39. https://doi.org/10.1186/s13073-019-0652-8 PMID:31221204

13. Postow MA, Callahan MK, Wolchok JD. Immune checkpoint blockade in cancer therapy. J Clin Oncol. 2015; 33:1974–82. https://doi.org/10.1200/JCO.2014.59.4358 PMID:25605845

14. Postow MA, Sidlow R, Hellmann MD. Immune-related adverse events associated with immune checkpoint blockade. N Engl J Med. 2018; 378:158–68. https://doi.org/10.1056/NEJMra1703481 PMID:29320654

15. Topalian SL, Taube JM, Anders RA, Pardoll DM. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. Nat Rev Cancer. 2016; 16:275–87. https://doi.org/10.1038/nrc.2016.36 PMID:27079802

16. Hu X, Yang L, Mo YY. Role of pseudogenes in tumorigenesis. Cancers (Basel). 2018; 10:256. https://doi.org/10.3390/cancers10080256 PMID:30071685

17. Glenfield C, McLysaght A. Pseudogenes provide evolutionary evidence for the competitive endogenous RNA hypothesis. Mol Biol Evol. 2018; 35:2886–99. https://doi.org/10.1093/molbev/msy183 PMID:30252115

18. Johnson TS, Li S, Franz E, Huang Z, Dan Li S, Campbell MJ, Huang K, Zhang Y. PseudoFuN: deriving functional potentials of pseudogenes from integrative relationships with genes and microRNAs across 32 cancers. Gigascience. 2019; 8:giz046. https://doi.org/10.1093/gigascience/giz046 PMID:31029062

19. Xiao-Jie L, Ai-Mei G, Li-Juan J, Jiang X. Pseudogene in cancer: real functions and promising signature. J Med Genet. 2015; 52:17–24.
https://doi.org/10.1136/jmedgenet-2014-102785
PMID:25391452

20. Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu YM, Cao X, Asangani IA, Kothari V, Prensner JR, Lonigro RJ, Iyer MK, Barrette T, Shanmugam A, et al. Expressed pseudogenes in the transcriptional landscape of human cancers. Cell. 2012; 149:1622–34.
https://doi.org/10.1016/j.cell.2012.04.041
PMID:22726445

21. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010; 465:1033–38.
https://doi.org/10.1038/nature09144 PMID:20577206

22. Johnsson P, Ackley A, Vidarsdottir L, Lui WO, Corcoran M, Grandér D, Morris KV. A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. Nat Struct Mol Biol. 2013; 20:440–46.
https://doi.org/10.1038/nsmb.2516 PMID:23435381

23. Irizar H, Muñoz-Culla M, Sáenz-Cuesta M, Osorio-Querejeta I, Sepúlveda L, Castillo-Triviño T, Prada A, Lopez de Munain A, Olascoaga J, Otaegui D. Identification of ncRNAs as potential therapeutic targets in multiple sclerosis through differential ncRNA - mRNA network analysis. BMC Genomics. 2015; 16:250.
https://doi.org/10.1186/s12864-015-1396-5
PMID:25880556

24. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell. 2011; 43:904–14.
https://doi.org/10.1016/j.molcel.2011.08.018
PMID:21925379

25. An Y, Furber KL, Ji S. Pseudogenes regulate parental gene expression via ceRNA network. J Cell Mol Med. 2017; 21:185–92.
https://doi.org/10.1111/jcmm.12952 PMID:27561207

26. Park JY, Lee JE, Park JB, Yoo H, Lee SH, Kim JH. Roles of long non-coding RNAs on tumorigenesis and glioma development. Brain Tumor Res Treat. 2014; 2:1–6.
https://doi.org/10.14791/btrt.2014.2.1.1
PMID:24926466

27. Liu B, Ma T, Li Q, Wang S, Sun W, Li W, Liu J, Guo Y. Identification of a lncRNA-associated competing endogenous RNA-regulated network in clear cell renal cell carcinoma. Mol Med Rep. 2019; 20:485–94.
https://doi.org/10.3892/mmr.2019.10290
PMID:31180525

28. Liu B, Liu J, Liu K, Huang H, Li Y, Hu X, Wang K, Cao H, Cheng Q. A prognostic signature of five pseudogenes for predicting lower-grade gliomas. Biomed Pharmacother. 2019; 117:109116.
https://doi.org/10.1016/j.biopha.2019.109116
PMID:31247469

29. Wang Y, Liu X, Guan G, Xiao Z, Zhao W, Zhuang M. Identification of a five-pseudogene signature for predicting survival and its ceRNA network in glioma. Front Oncol. 2019; 9:1059.
https://doi.org/10.3389/fonc.2019.01059
PMID:31681595

30. Chiang JJ, Sparrer KM, van Gent M, Lässig C, Huang T, Osterrieder N, Hopfner KP, Gack MU. Viral unmasking of cellular 5S rRNA pseudogene transcripts induces RIG-I-mediated immunity. Nat Immunol. 2018; 19:53–62.
https://doi.org/10.1038/s41590-017-0005-y
PMID:29180807

31. Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, Cayting P, Harrrison P, Gerstein M. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res. 2007; 35:D55–60.
https://doi.org/10.1093/nar/gkl851 PMID:17099229

32. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, Reymond A, Hubbard TJ, Harrow J, Gerstein MB. The GENCODE pseudogene resource. Genome Biol. 2012; 13:R51.
https://doi.org/10.1186/gb-2012-13-9-r51
PMID:22951037

33. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015; 12:453–57.
https://doi.org/10.1038/nmeth.3337 PMID:25822800

34. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. Methods Mol Biol. 2018; 1711:243–59.
https://doi.org/10.1007/978-1-4939-7493-1_12
PMID:29344893

35. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, Hackl H, Trajanoski Z. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep. 2017; 18:248–62.
https://doi.org/10.1016/j.celrep.2016.12.019
PMID:28052254

36. Zheng LL, Zhou KR, Liu S, Zhang DY, Wang ZL, Chen ZR, Yang JH, Qu LH. dreamBase: DNA modification, RNA

regulation and protein binding of expressed pseudogenes in human health and disease. Nucleic Acids Res. 2018; 46:D85–91. https://doi.org/10.1093/nar/gkx972 PMID:29059382

37. Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, Huang WC, Sun TH, Tu SJ, Lee WH, Chiew MY, Tai CS, Wei TY, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. 2018; 46:D296–302. https://doi.org/10.1093/nar/gkx1067 PMID:29126174

38. Bøvelstad HM, Nygård S, Størvold HL, Aldrin M, Borgan Ø, Frigessi A, Lingjaerde OC. Predicting survival from microarray data—a comparative study. Bioinformatics. 2007; 23:2080–87. https://doi.org/10.1093/bioinformatics/btm305 PMID:17553857

39. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010; 26:1572–73. https://doi.org/10.1093/bioinformatics/btq170 PMID:20427518

40. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102:15545–50. https://doi.org/10.1073/pnas.0506580102 PMID:16199517

41. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012; 16:284–87. https://doi.org/10.1089/omi.2011.0118 PMID:22455463

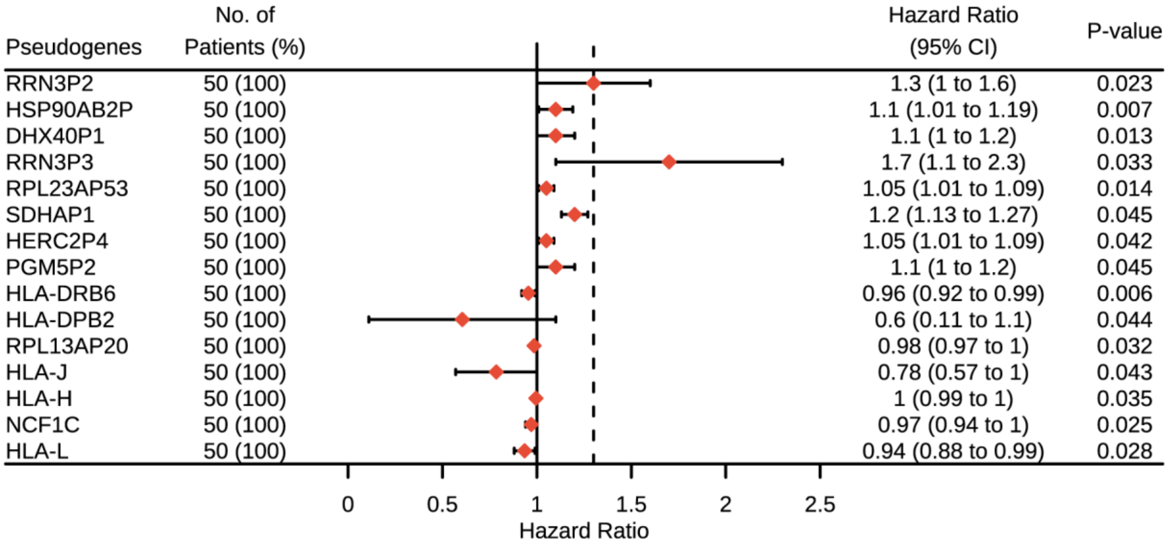
42. Ali HR, Chlon L, Pharoah PD, Markowetz F, Caldas C. Patterns of immune infiltration in breast cancer and their clinical implications: a gene-expression-based retrospective study. PLoS Med. 2016; 13:e1002194. https://doi.org/10.1371/journal.pmed.1002194 PMID:27959923

43. McDonald KA, Kawaguchi T, Qi Q, Peng X, Asaoka M, Young J, Opyrchal M, Yan L, Patnaik S, Otsuji E, Takabe K. Tumor heterogeneity correlates with less immune response and worse survival in breast cancer patients. Ann Surg Oncol. 2019; 26:2191–99. https://doi.org/10.1245/s10434-019-07338-3 PMID:30963401

44. Narayanan S, Kawaguchi T, Yan L, Peng X, Qi Q, Takabe K. Cytolytic activity score to assess anticancer immunity in colorectal cancer. Ann Surg Oncol. 2018; 25:2323–31. https://doi.org/10.1245/s10434-018-6506-6 PMID:29770915

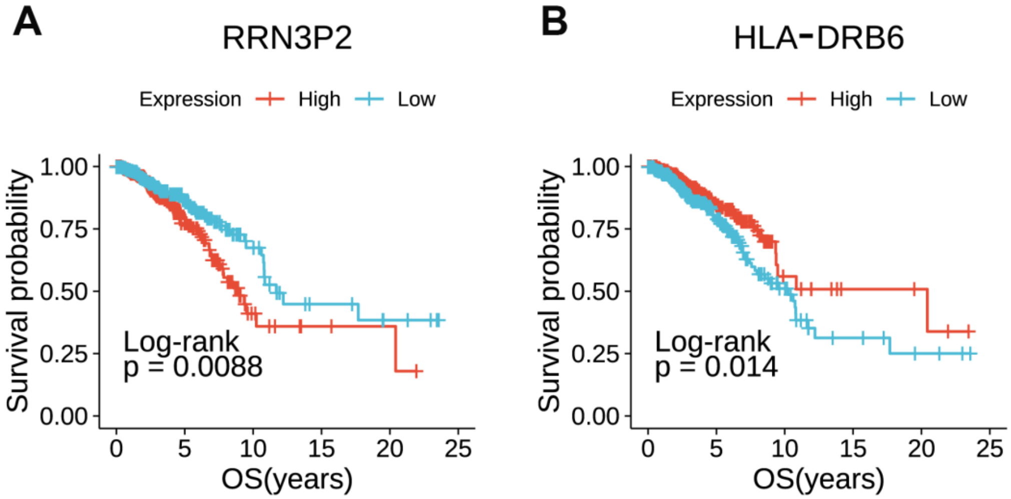
SUPPLEMENTARY MATERIALS

Supplementary Figures

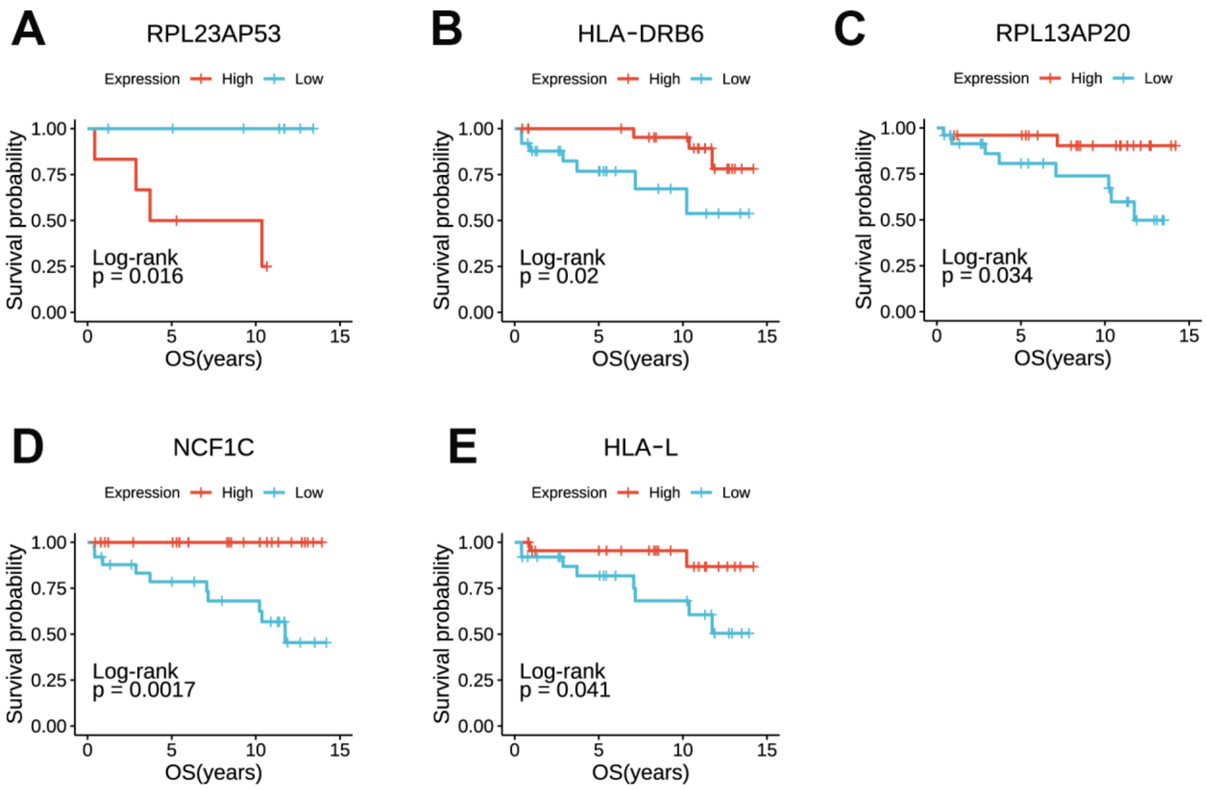


Pseudogenes	No. of Patients (%)	Hazard Ratio (95% CI)	P-value
RRN3P2	50 (100)	1.3 (1 to 1.6)	0.023
HSP90AB2P	50 (100)	1.1 (1.01 to 1.19)	0.007
DHX40P1	50 (100)	1.1 (1 to 1.2)	0.013
RRN3P3	50 (100)	1.7 (1.1 to 2.3)	0.033
RPL23AP53	50 (100)	1.05 (1.01 to 1.09)	0.014
SDHAP1	50 (100)	1.2 (1.13 to 1.27)	0.045
HERC2P4	50 (100)	1.05 (1.01 to 1.09)	0.042
PGM5P2	50 (100)	1.1 (1 to 1.2)	0.045
HLA-DRB6	50 (100)	0.96 (0.92 to 0.99)	0.006
HLA-DPB2	50 (100)	0.6 (0.11 to 1.1)	0.044
RPL13AP20	50 (100)	0.98 (0.97 to 1)	0.032
HLA-J	50 (100)	0.78 (0.57 to 1)	0.043
HLA-H	50 (100)	1 (0.99 to 1)	0.035
NCF1C	50 (100)	0.97 (0.94 to 1)	0.025
HLA-L	50 (100)	0.94 (0.88 to 0.99)	0.028

Supplementary Figure 1. Validation of the 15 prognostic pseudogenes in EGA dataset. The hazard ratios (HR), 95% confidence intervals (CI) calculated by univariate Cox proportional hazard regression of the 15 prognostic pseudogenes in EGA dataset.



Supplementary Figure 2. Correlation between the expression levels of RRN3P2 (**A**) and HLA-DRB6 (**B**) and overall survival in patients with breast cancer from TCGA dataset.



Supplementary Figure 3. Correlation between the expression levels of RPL23AP53 (**A**), HLA-DRB6 (**B**), RPL13AP20 (**C**), NCF1C (**D**) and HLA-L (**E**) and overall survival in patients with breast cancer from EGA dataset.

Supplementary Tables

Supplementary Table 1. The coefficients of the 15 prognostic pseudogenes by LASSO.

Pseudogenes	Coefficients
NCF1C	0.000000000
HLA-DRB6	-0.002846803
HLA-DRB2	-0.104956422
HLA-J	-0.057654172
HLA-H	0.000000000
HLA-L	-0.013823392
RPL13AP20	-0.001763734
PGM5P2	0.000000000
HERC2P4	0.000000000
HSP90AB2P	0.000000000
DHX40P1	0.044932918
RRN3P3	0.000000000
RRN3P2	0.226104532
SDHAP1	0.046539494
RPL23AP53	0.021710806

Supplementary Table 2. Clinicopathological features stratified by high-risk and low-risk subgroups.

Variable	High-risk, n (%)	Low-risk, n (%)	***P***
No. of Patients	387(49.94)	388(50.06)	-
Age at diagnosis, years			0.603
≤ 50	116(14.97)	124(16.00)	
> 50	271(34.96)	264(34.06)	
ER status			8e-08
Negative	53(6.84)	118(15.23)	
Positive	311(40.13)	257(33.16)	
Unknown	23(2.97)	15(1.68)	
PR status			6e-04
Negative	97(12.52)	145(18.71)	
Positive	265(34.19)	228(29.42)	
Unknown	25(3.22)	15(1.94)	
HER-2 status			0.904
Negative	269(34.71)	272(35.10)	
Positive	65(8.39)	67(8.65)	
Unknown	53(6.84)	49(6.32)	
Molecular subtypes			4e-06
Normal-like	42(5.42)	89(11.48)	
Luminal A	26(3.35)	39(5.03)	
Luminal B	206(26.58)	166(21.42)	
HER2 positive	102(13.16)	75(9.68)	
Basal-like	6(0.77)	16(2.06)	
Unknown	5(0.64)	3(0.39)	
T stage			0.688
T1	105(13.55)	100(12.90)	
T2	223(28.77)	235(30.32)	
T3	41(5.29)	42(5.42)	
T4	17(2.19)	10(1.29)	
Unknown	1(0.13)	1(0.13)	
Lymph node stage			0.037
N0	168(21.68)	193(24.90)	
N1	137(17.68)	127(16.39)	
N2	52(6.71)	40(5.16)	
N3	18(2.32)	25(3.23)	
Unknown	12(1.55)	3(0.39)	
Distance metastasis status			0.248
Negative	347(44.77)	354(45.68)	
Positive	6(0.77)	10(1.29)	
Unknown	34(4.39)	24(3.10)	
Vital status			0.070
Alive	318(41.03)	338(43.61)	
Dead	69(8.90)	50(6.45)	

Abbreviations: ER: estrogen receptor; PR: progesterone receptor; HER-2: human epidermal growth receptor-2.

* Evaluated by Chi-square test.

Supplementary Table 3. Clinicopathological features stratified by P1 and P2 subgroups.

Variable	P1, n (%)	P2, n (%)	*P*
No. of Patients	362(46.71)	413(53.29)	-
Age at diagnosis, years			0.709
≤ 50	115(14.84)	125(16.12)	
> 50	247(31.87)	288(37.16)	
ER status			0.028
Negative	66(8.52)	105(13.55)	
Positive	275(35.48)	293(37.81)	
Unknown	21(2.71)	15(1.94)	
PR status			0.413
Negative	107(13.81)	135(17.42)	
Positive	233(30.06)	260(33.55)	
Unknown	22(2.84)	18(2.32)	
HER-2 status			0.358
Negative	250(32.25)	291(37.55)	
Positive	58(7.48)	74(9.55)	
Unknown	54(6.97)	48(6.19)	
Molecular subtypes			0.002
Normal-like	47(6.06)	84(10.84)	
Luminal A	26(3.35)	39(5.03)	
Luminal B	180(23.23)	192(24.77)	
HER2 positive	95(12.25)	82(10.58)	
Basal-like	7(0.90)	15(1.94)	
Unknown	7(0.90)	1(0.13)	
T stage			0.075
T1	93(12.00)	112(14.45)	
T2	211(27.23)	247(31.87)	
T3	37(4.77)	46(5.94)	
T4	20(2.58)	7(0.90)	
Unknown	1(0.13)	1(0.13)	
Lymph node stage			0.047
N0	153(19.74)	208(26.84)	
N1	132(17.03)	132(17.03)	
N2	48(6.19)	44(5.68)	
N3	18(2.32)	25(3.23)	
Unknown	11(1.42)	4(0.52)	
Distance metastasis status			0.441
Negative	325(41.94)	376(48.52)	
Positive	6(0.77)	10(1.29)	
Unknown	31(4.00)	27(3.48)	
Vital status			0.560
Alive	303(39.1)	353(45.5)	
Dead	59(7.61)	60(7.74)	

Abbreviations: P1: patient subgroup 1; P2: patient subgroup 2; ER: estrogen receptor; PR: progesterone receptor; HER-2: human epidermal growth receptor-2.

* Evaluated by Chi-square test.

Supplementary Table 4. Potential miRNAs binding to the 15 prognostic pseudogenes identified by dreamBase.

Pseudogene	miRNA
NCF1C	0
HLA-DRB6	0
HLA-DRB2	0
HLA-J	hsa-miR-1193; hsa-miR-140-3p; hsa-miR-15a-5p; hsa-miR-15b-5p; hsa-miR-16-5p; hsa-miR-195-5p; hsa-miR-214-3p; hsa-miR-2278; hsa-miR-3619-5p; hsa-miR-3918; hsa-miR-424-5p; hsa-miR-4428; hsa-miR-4726-5p; hsa-miR-497-5p; hsa-miR-589-5p; hsa-miR-6838-5p; hsa-miR-761
HLA-H	hsa-miR-124-3p; hsa-miR-125a-5p; hsa-miR-125b-5p; hsa-miR-1343-3p; hsa-miR-140-3p; hsa-miR-143-3p; hsa-miR-15a-5p; hsa-miR-15b-5p; hsa-miR-16-5p; hsa-miR-195-5p; hsa-miR-214-3p; hsa-miR-22-3p; hsa-miR-2278; hsa-miR-296-5p; hsa-miR-3127-5p; hsa-miR-3184-5p; hsa-miR-3200-5p; hsa-miR-323a-3p; hsa-miR-323b-3p; hsa-miR-3605-3p; hsa-miR-3619-5p; hsa-miR-362-5p; hsa-miR-380-3p; hsa-miR-3918; hsa-miR-423-5p; hsa-miR-424-5p; hsa-miR-4319; hsa-miR-4640-5p; hsa-miR-4726-5p; hsa-miR-4770; hsa-miR-497-5p; hsa-miR-500b-5p; hsa-miR-506-3p; hsa-miR-514a-5p; hsa-miR-532-3p; hsa-miR-605-3p; hsa-miR-6088; hsa-miR-665; hsa-miR-6746-3p; hsa-miR-6783-3p; hsa-miR-6838-5p; hsa-miR-744-5p; hsa-miR-761; hsa-miR-766-5p
HLA-L	hsa-miR-140-3p; hsa-miR-15a-5p; hsa-miR-15b-5p; hsa-miR-16-5p; hsa-miR-195-5p; hsa-miR-214-3p; hsa-miR-2278; hsa-miR-296-5p; hsa-miR-3127-5p; hsa-miR-335-5p; hsa-miR-3619-5p; hsa-miR-370-3p; hsa-miR-380-3p; hsa-miR-3918; hsa-miR-424-5p; hsa-miR-4428; hsa-miR-4726-5p; hsa-miR-497-5p; hsa-miR-500b-5p; hsa-miR-6838-5p; hsa-miR-6893-3p; hsa-miR-761
RPL13AP20	hsa-miR-1224-5p; hsa-miR-193a-5p; hsa-miR-214-3p; hsa-miR-296-3p; hsa-miR-29a-3p; hsa-miR-29b-3p; hsa-miR-29c-3p; hsa-miR-3619-5p; hsa-miR-3681-5p; hsa-miR-409-3p; hsa-miR-452-5p; hsa-miR-4664-3p; hsa-miR-4676-3p; hsa-miR-486-5p; hsa-miR-526b-5p; hsa-miR-532-5p; hsa-miR-6512-3p; hsa-miR-665; hsa-miR-6849-5p; hsa-miR-761; hsa-miR-766-5p; hsa-miR-873-5p; hsa-miR-892c-3p
PGM5P2	hsa-miR-328-3p
HERC2P4	hsa-miR-146a-5p; hsa-miR-146b-5p; hsa-miR-181a-5p; hsa-miR-181b-5p; hsa-miR-181c-5p; hsa-miR-181d-5p; hsa-miR-205-5p; hsa-miR-4262; hsa-miR-7153-5p
HSP90AB2P	hsa-miR-124-3p; hsa-miR-1252-5p; hsa-miR-128-3p; hsa-miR-142-5p; hsa-miR-144-5p; hsa-miR-150-5p; hsa-miR-182-5p; hsa-miR-18a-5p; hsa-miR-18b-5p; hsa-miR-205-5p; hsa-miR-2115-3p; hsa-miR-216a-3p; hsa-miR-2682-5p; hsa-miR-29a-3p; hsa-miR-29b-3p; hsa-miR-29c-3p; hsa-miR-3187-3p; hsa-miR-320a; hsa-miR-320b; hsa-miR-320c; hsa-miR-320d; hsa-miR-345-3p; hsa-miR-34b-5p; hsa-miR-365a-3p; hsa-miR-365b-3p; hsa-miR-3681-3p; hsa-miR-376a-3p; hsa-miR-376b-3p; hsa-miR-376c-3p; hsa-miR-380-3p; hsa-miR-423-3p; hsa-miR-4429; hsa-miR-449c-5p; hsa-miR-4735-3p; hsa-miR-4761-5p; hsa-miR-4766-3p; hsa-miR-4766-5p; hsa-miR-488-3p; hsa-miR-494-3p; hsa-miR-496; hsa-miR-506-3p; hsa-miR-514a-5p; hsa-miR-515-5p; hsa-miR-519e-5p; hsa-miR-5590-3p; hsa-miR-577; hsa-miR-616-3p; hsa-miR-670-3p; hsa-miR-670-5p; hsa-miR-766-5p; hsa-miR-874-3p; hsa-miR-9-5p; hsa-miR-942-5p; hsa-miR-944
DHX40P1	hsa-miR-124-3p; hsa-miR-1271-5p; hsa-miR-1306-5p; hsa-miR-186-5p; hsa-miR-199a-5p; hsa-miR-199b-5p; hsa-miR-22-3p; hsa-miR-30a-5p; hsa-miR-30b-5p; hsa-miR-30c-5p; hsa-miR-30d-5p; hsa-miR-30e-5p; hsa-miR-33a-5p; hsa-miR-33b-5p; hsa-miR-3612; hsa-miR-506-3p; hsa-miR-545-5p; hsa-miR-625-3p; hsa-miR-628-5p; hsa-miR-650; hsa-miR-7151-5p; hsa-miR-96-5p
RRN3P3	0
RRN3P2	hsa-miR-1297; hsa-miR-132-3p; hsa-miR-191-5p; hsa-miR-212-3p; hsa-miR-224-3p; hsa-miR-26a-5p; hsa-miR-26b-5p; hsa-miR-300; hsa-miR-381-3p; hsa-miR-4465;

	hsa-miR-4524a-5p; hsa-miR-4524b-5p; hsa-miR-522-3p; hsa-miR-532-5p
SDHAP1	hsa-let-7a-5p; hsa-let-7b-5p; hsa-let-7c-5p; hsa-let-7d-5p; hsa-let-7e-5p; hsa-let-7f-5p; hsa-let-7g-5p; hsa-let-7i-5p; hsa-miR-105-5p; hsa-miR-1249-3p; hsa-miR-1301-3p; hsa-miR-136-5p; hsa-miR-15a-5p; hsa-miR-15b-5p; hsa-miR-16-5p; hsa-miR-195-5p; hsa-miR-216a-5p; hsa-miR-2355-5p; hsa-miR-2681-3p; hsa-miR-3150a-3p; hsa-miR-3529-5p; hsa-miR-361-3p; hsa-miR-3622b-5p; hsa-miR-379-5p; hsa-miR-424-5p; hsa-miR-4458; hsa-miR-4500; hsa-miR-4731-5p; hsa-miR-4761-3p; hsa-miR-485-5p; hsa-miR-491-5p; hsa-miR-495-3p; hsa-miR-497-5p; hsa-miR-5047; hsa-miR-505-3p; hsa-miR-516b-5p; hsa-miR-542-3p; hsa-miR-543; hsa-miR-545-3p; hsa-miR-5688; hsa-miR-5691; hsa-miR-6763-5p; hsa-miR-6805-3p; hsa-miR-6838-5p; hsa-miR-6884-5p; hsa-miR-7853-5p; hsa-miR-98-5p
RPL23AP53	hsa-miR-1343-3p; hsa-miR-141-3p; hsa-miR-200a-3p; hsa-miR-214-5p; hsa-miR-28-5p; hsa-miR-3139; hsa-miR-376a-3p; hsa-miR-376b-3p; hsa-miR-6783-3p; hsa-miR-708-5p

Supplementary Table 5. miRNA targeted genes correlated with their pseudogenes at | r | ≥ 0.3 and P < 0.05.

Pseudogene	miRNA targeted genes
NCF1C	0
HLA-DRB6	0
HLA-DRB2	0
HLA-J	CCL5
HLA-H	CD274; BAK1; CD38; CXCL10; CCL4; CCL5; COTL1
HLA-L	CD38; CXCL10; CCL4; CCL5; CD274
RPL13AP20	BAX; GSK3B; IFNAR1; GSK3B; MGMT; CDK3
PGM5P2	0
HERC2P4	0
HSP90AB2P	AR; BCL2L11; CCNT2; CLOCK; CNOT6; CPEB3; CPEB4; CREB1; DICER1; ERBB3; ESR1; FOXP1; GSK3B
DHX40P1	CLOCK; ATF6; CDH1; CLTC
RRN3P3	0
RRN3P2	ATM; CHD1; CPEB4; CRK; GSK3B; IRAK4; KLHL11; LARP1; MAP3K2; MTDH; NR2C2; PIK3C2A; PTPN13; RASA1; RB1; RB1CC1; RCBTB1; ROCK1
SDHAP1	MPL; DMTF1; MBD1; MPL
RPL23AP53	KLF12; KLHL20; ATRX; MPL