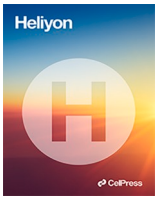
Heliyon 10 (2024) e25640

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Integrating single-cell and bulk expression data to identify and analyze cancer prognosis-related genes

Shengbao Bao, Yaxin Fan, Yichao Mei, Junxiang Gao *

Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, 430070, China

ARTICLE INFO

Keywords:
Single-cell
Biomarker
Prognosis model
Polygenic risk score

ABSTRACT

Compared with traditional evaluation methods of cancer prognosis based on tissue samples, single-cell sequencing technology can provide information on cell type heterogeneity for predicting biomarkers related to cancer prognosis. Therefore, the bulk and single-cell expression profiles of breast cancer and normal cells were comprehensively analyzed to identify malignant and non-malignant markers and construct a reliable prognosis model. We first screened highly reliable differentially expressed genes from bulk expression profiles of multiple breast cancer tissues and normal tissues, and inferred genes related to cell malignancy from single-cell data. Then we identified eight critical genes related to breast cancer to conduct Cox regression analysis, calculate polygenic risk score (PRS), and verify the predictive ability of PRS in two data groups. The results show that PRS can divide breast cancer patients into high-risk group and low-risk group. PRS is related to the overall survival time and relapse-free interval and is a prognosis factor independent of conventional clinicopathological characteristics. Breast cancer is usually regarded as a cancer with a relatively good prognosis. In order to further explore whether this workflow can be applied to cancer with poor prognosis, we selected lung cancer for a comparative study. The results show that this workflow can also build a reasonable prognosis model for lung cancer. This study provides new insight and practical source code for further research on cancer biomarkers and drug targets. It also provides basis for survival prediction, treatment response prediction, and personalized treatment.

1. Introduction

Cancer prognosis models can predict the development trend and prognosis results based on patient biomarkers, clinical features, and other information to help doctors better assess the probability of future recurrence, death, disability, or complications of patients, thereby formulating more personalized treatment plans, improving treatment effectiveness and life quality of patients. Constructing a cancer prognosis model requires collecting and integrating multiple data types. In addition to clinical data such as age, gender, and medical history collected during diagnosis and treatment, molecular-level information such as DNA, RNA, or protein extracted from patient tissue samples should also be used. Many studies have identified cancer-related genes using bulk RNA-seq [1,2] and detected differentially expressed genes (DEGs) by comparing their expression profiles in normal and cancer tissues. These DEGs may be closely related to the occurrence, development, and prognosis of cancer and can, therefore, be used to construct cancer prognosis models.

Compared to bulk transcriptome sequencing of a tissue, the recently developed single-cell transcriptome can sequence individual

* Corresponding author.
E-mail address: gao200@mail.hzau.edu.cn (J. Gao).

https://doi.org/10.1016/j.heliyon.2024.e25640
Received 27 September 2023; Received in revised form 3 January 2024; Accepted 31 January 2024
Available online 10 February 2024

cells, thus detecting gene expression heterogeneity at single-cell resolution [3,4] and more cell subpopulations and biological differences [5]. Single-cell transcriptome can also detect rare cell subpopulations [6], which may have different gene expression patterns and biological characteristics, and these subpopulations may be masked in bulk RNA-seq. Therefore, the single-cell transcriptome can provide additional information and play an essential role in identifying cancer-related genes. For example, Zhou et al. comprehensively analyzed gene regulatory networks for the intrinsic subtypes of triple-negative breast cancer patients using scRNA-seq. The subtypes of the malignant cells were assigned based on the PAM50 model. The authors constructed gene regulatory networks by integrating gene co-expression and enrichment of transcription-binding motifs and identified the critical genes based on the centrality metrics of genes [7]. The prognosis of different cancers varies greatly; breast cancer and lung cancer can represent two prognosis types. Breast cancer is generally considered to have a better prognosis, and the 5-year survival rate can reach more than 80% [8]. Compared with breast cancer, lung cancer has a higher degree of malignancy and a significantly worse prognosis. More than half of patients will die within one year after diagnosis of lung cancer, and the five-year survival rate is only 17.8% [9]. Therefore, breast cancer and lung cancer can represent two types of prognosis.

To mine potential targets for clinical diagnosis and treatment from the expression profile data of breast cancer and lung cancer, our research explored whether single-cell transcriptome and bulk expression data can be effectively integrated to identify prognostic-related genes of breast cancer and lung cancer. In this paper, we identified differentially expressed genes from breast cancer and normal tissues based on bulk expression profiles and predicted genes related to cell malignancy based on single-cell transcriptome data. Then, we integrated the two groups of genes to obtain eight critical genes related to breast cancer and validated the performance of the polygenic risk score. We also conducted extensive research on lung cancer with poor prognosis. Our study provides two paradigms for fully mining bulk and single-cell transcriptomic data to construct prognosis models. The source code has been uploaded to the GitHub and can help subsequent researchers use this workflow to mine other cancer data.

2. Materials and methods

2.1. Bulk and single-cell expression datasets

The datasets and workflow chart in this study are shown in Fig. 1. The expression datasets GSE65194, GSE93601 [10], GSE109169 [11], and GSE202203 [12] are all from Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/), where the first three are microarray data and the last is high-throughput sequencing data. The dataset GSE65194 contains 11 normal breast tissues and 167 breast cancer tissues or cell lines. The dataset GSE93601 contains 602 tumor and 508 tumor-adjacent tissues from patients diagnosed with invasive breast cancer. The dataset GSE109169 contains 25 normal breast tissues and 25 breast cancer tissues. GSE202203 contains 3207 breast cancer tissues, 2913 of which have clinical information and were selected for analysis. The dataset GDC TCGA Breast Cancer is from UCSC Xena (https://xenabrowser.net/datapages/), which contains bulk RNA-seq data of 1104 breast cancer tissues and 113 normal tissues, as well as clinical data of 1099 breast cancer samples. The single-cell expression profile data GSM5956094 [13] of primary breast cancer also comes from the GEO database (Table S1).

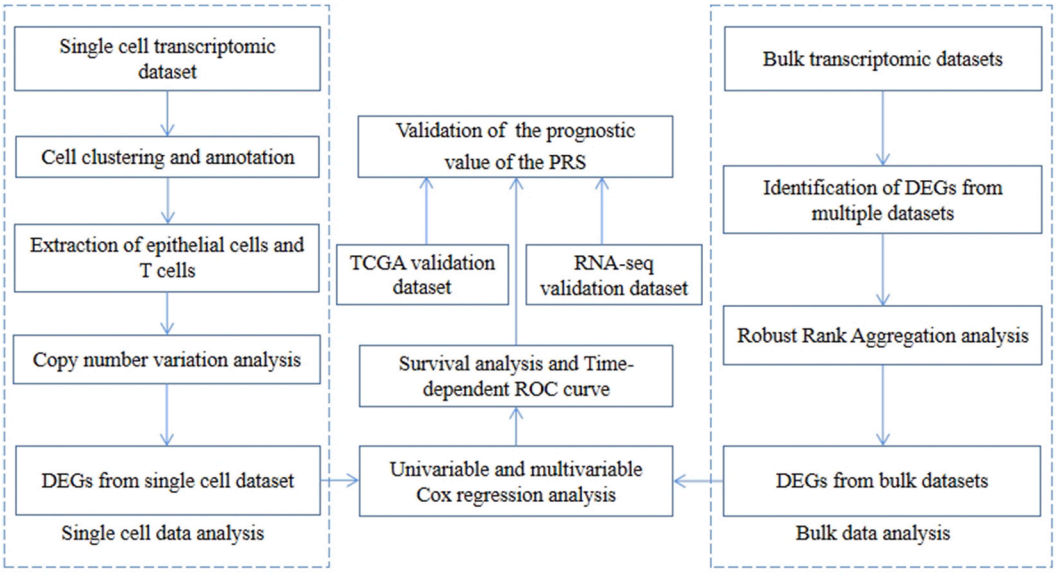


Fig. 1. The workflow of the study.

2.2. Identification of up-regulated and down-regulated genes from bulk expression profile

Firstly, the gene names of datasets GSE65194, GSE93601, and GSE109169 were converted to gene symbol form. If several probes detect a gene, its expression is defined as the average value of the genes calculated on all probes. Finally, the limma package (v3.50.3) [14] was used for differential expression analysis. The DEGs in the GDC TCGA Break Cancer RNA-seq dataset were identified using the DESeq2 software package (v1.34.0) [15]. Based on a criterion of $p_{adj} < 0.05$, the "subset" function was used to filter the genes up-regulated and down-regulated in the datasets GSE65194, GSE93601, GSE109169, and GDC TCGA Breast Cancer. Subsequently, up-regulated and down-regulated genes were ranked based on log2 (Fold Change) in a single dataset. Then, the RobustRankAggreg package (v1.2.1) [16] was used to merge the up-regulation and down-regulation gene sets from the four datasets mentioned above and sort them by p-value to extract the top 100 significantly up-regulated genes and the top 100 significantly down-regulated genes.

2.3. Processing and analysis of single-cell data

Breast cancer single-cell expression profiling data GSM5956094 was imported and pre-processed using the R package "Seurat" (v4.3.0.1) [17]. The "CreateSeuratObject" function created a Seurat object; the "min.cells" parameter was set to 3, and the "min.features" parameter was set to 500. Then, the "subset" function was used to filter out low-quality cells meeting any of the following conditions: (1) the number of expressed genes is greater than or equal to 5000; (2) unique molecular identifiers (UMIs) mapping to mitochondrial genes is greater than or equal to 15%; (3) UMIs mapping to red blood cell genes is greater than or equal to 1%; and (4) the number of UMI is greater than or equal to 40,000.

The data remaining after filtering were normalized using the "SCTransform" function, and then principal component analysis was performed using the "RunPCA" function. The first 33 principal components that could explain most of the variance were used for subsequent analyses. The "FindNeighbors" and "FindClusters" functions were then used for cell clustering analysis with the resolution parameter of 0.2. Furthermore, uniform manifold approximation and projection (UMAP) dimensionality reduction was conducted and visualized using the "RunUMAP" function.

Then, R package "DoubletFinder" (v2.0.3) was used to remove inferred doublets, using default parameters [18]. The "FindNeighbors" and "FindClusters" functions were used again for cell clustering analysis of the remaining cells.

The R package "celldex" (v1.11.1) was used to obtain the built-in dataset "HumanPrimaryCellAtlasData" [19], which was subsequently used as the reference dataset for cell annotation. R package "SingleR" (v1.8.1) was used for cell annotation to infer the cell type of each cell [19]. Cluster names were named as the most common cell type in each cluster. Then, the expression profiles of T cells and epithelial cells were extracted. The "SCTransform" function standardized and normalized the extracted expression profiles. Finally, the "FindNeighbors" and the "FindClusters" functions were used for cell clustering with the resolution parameter 1.0.

Copy number variation in epithelial cells was inferred using R package "inferCNV" (v1.10.1) using T cells as a reference [20]. The cell cluster with the highest copy number variation score was considered the most malignant. In contrast, the cell cluster with the lowest copy number variation score was considered the weakest malignant. Then, we used the "FindMarkers" function to obtain the DEGs between the two clusters, with the following parameter settings: min.pct = 0.25, logfc.threshold = 0.5, test.use = "roc".

2.4. Construction of prognosis model

Merging 100 up-regulated and 100 down-regulated genes (Table S2) obtained from the bulk expression profiles with 185 DEGs (Table S3) obtained from single-cell data, 354 genes remained after removing genes that do not exist in the validation dataset GDC TCGA Breast Cancer and GSE202203. After loading the standardized GDC TCGA Breast Cancer data, only the expression profiles of these 354 genes were retained, and the expression matrix and survival information were merged. After merging, 70% of the data was used as the training set and 30% as the validation set.

Univariate Cox regression analysis was performed on the training set to identify genes significantly associated with survival ($p < 0.05$). Then, these genes were used for multivariate Cox regression to create a Polygenic Risk Score (PRS):

$$PRS = B \cdot X^T \tag{1}$$

where $B = [b_1, b_2, b_3, \ldots b_n]$, $X = [x_1, x_2, x_3, \ldots x_n]$. "$x_1, x_2, x_3, \ldots x_n$" represents the expression value of the genes in the multivariate Cox regression analysis, and "$b_1, b_2, b_3, \ldots b_n$" is the corresponding estimated regression coefficient.

2.5. Validation of prognosis model

A polygenic risk prediction model was used to calculate the PRS for each sample in the TCGA validation set and GSE202203 validation set, and patients were categorized into high-risk and low-risk types based on the median PRS. Survival analyses were performed using the R package "survival" (v3.2-13) [21]. Time-dependent receiver operating characteristic curve (tROC) analyses were performed using the R package "survivalROC" (v1.0.3.1) [22].

3. Result

3.1. Identification of highly reliable up-regulated and down-regulated genes in breast cancer by integrating multiple bulk expression profiles

We first identified breast cancer-related up-regulated and down-regulated genes (Fig. 2A–B) from four bulk expression datasets. In the GDC TCGA Breast Cancer dataset, 27,800 DEGs were identified, of which 17,101 genes were significantly up-regulated, and 10,699 genes were significantly down-regulated in tumor samples. In the dataset GSE109169, 7494 DEGs were identified, of which 4091 genes were significantly up-regulated, and 3403 genes were significantly down-regulated. In the dataset GSE93601, 9568 DEGs were identified, of which 4492 genes were significantly up-regulated, and 5076 genes were significantly down-regulated. In the dataset GSE65194, 13,005 DEGs were identified, of which 4798 genes were significantly up-regulated, and 8207 genes were significantly down-regulated. Notably, the up-regulated (Fig. 2A) and down-regulated genes (Fig. 2B) varied substantially across the four datasets, making it necessary to integrate different datasets. After integrating with the "aggregateRanks" function, a total of 19,894 up-regulated genes and 18,235 down-regulated genes were obtained. Then, genes with $p \geq 0.05$ were filtered out, resulting in highly reliable 1231 up-regulated genes and 1325 down-regulated genes from the bulk expression profiles.

3.2. Heterogeneity analysis of breast cancer cells based on single-cell RNA-seq data

The single-cell dataset GSE5956094 contains 13,434 cells, and 9256 high-quality cells were screened using the quality control steps described in the Methods section, including 25,953 expressed genes (Fig. S1).

After dimensionality reduction and clustering, 9256 cells were divided into 10 clusters, which were annotated as nine cell types using the "SingleR" method, and the main cell types of breast tissue were included in them (Fig. 2C), including epithelial cells, chondrocytes, macrophages, T cells, and endothelial cells. The results show that breast cancer cells are highly heterogeneous, and cell types have different gene expression patterns; cell type-specific information can be extracted from single-cell expression profiles.

The annotation results of epithelial and T cells were validated using marker genes of epithelial cells and T cells. Epithelial marker genes *CDH1*, *EPCAM*, *ESR1*, *KRT18*, and *KRT19* are highly expressed in annotated epithelial cells. In contrast, T cell marker genes *CD2*, *LCK*, *CD247*, *CD96*, and *IL7R* are highly expressed in annotated T cells (Fig. 2D). These results indicate that the annotation results of T

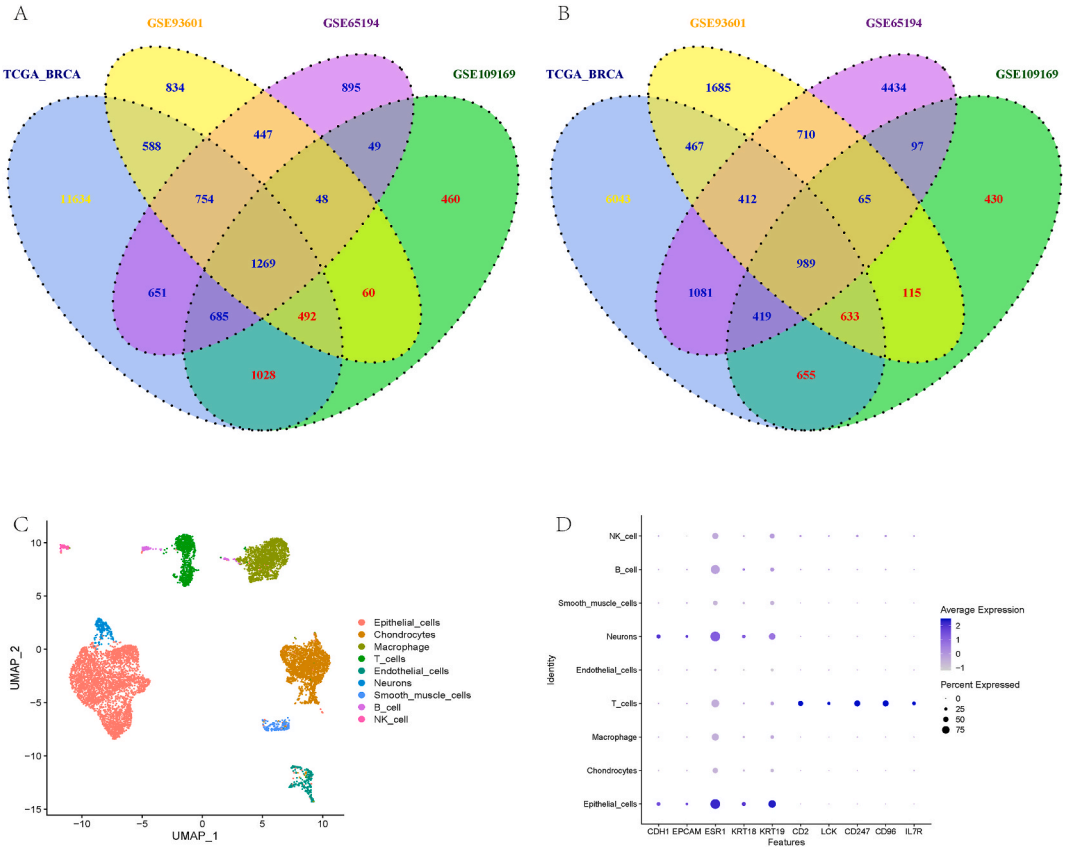


Fig. 2. (A) Venn plot of overlap up-regulated genes in the four datasets. (B) Venn plot of overlap down-regulated genes in the four datasets. (C) UMAP shows cell heterogeneity of breast cancer. (D) Distribution of epithelial and T cell marker genes.

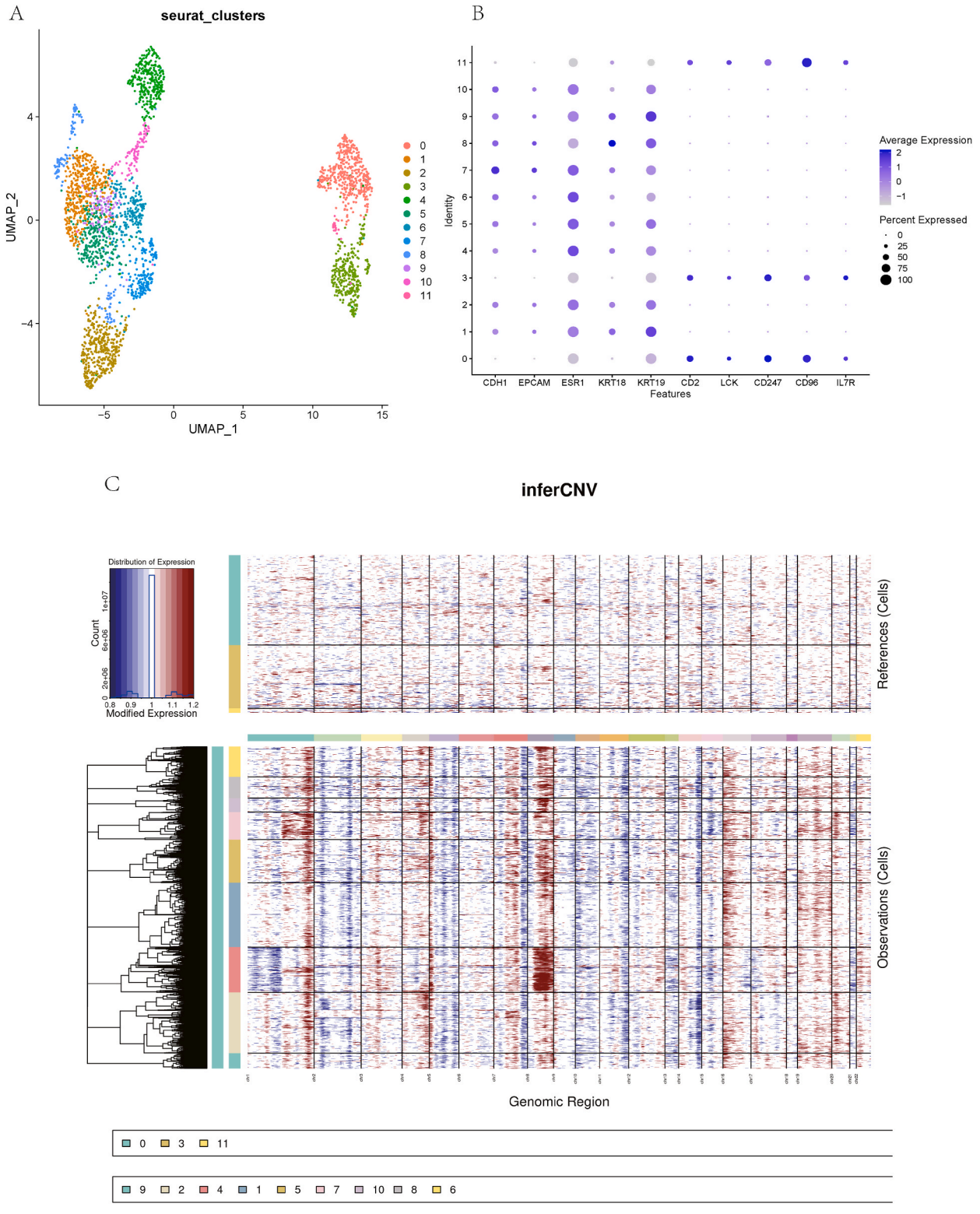
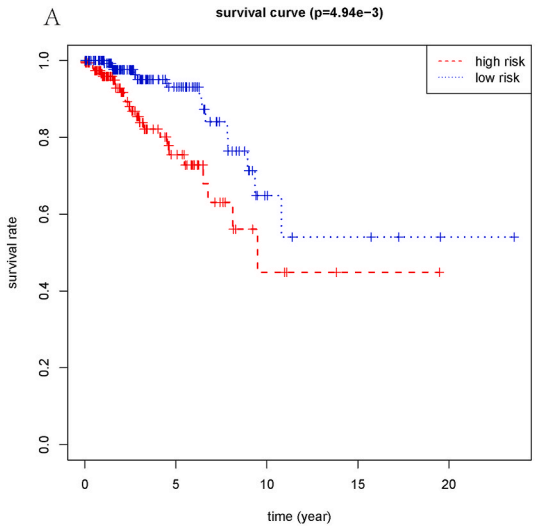
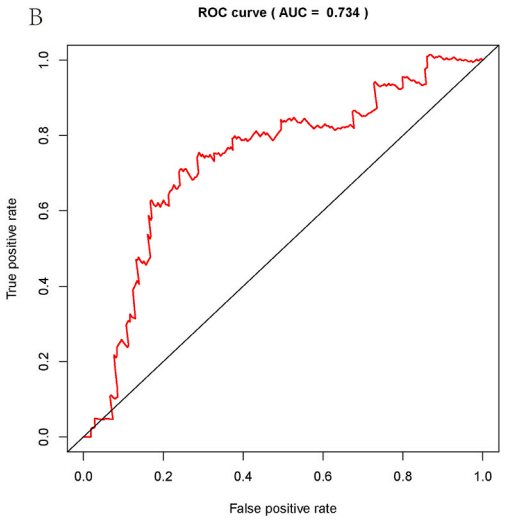
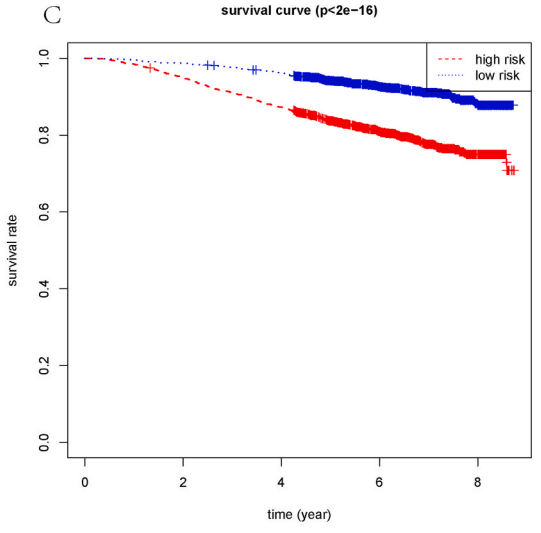
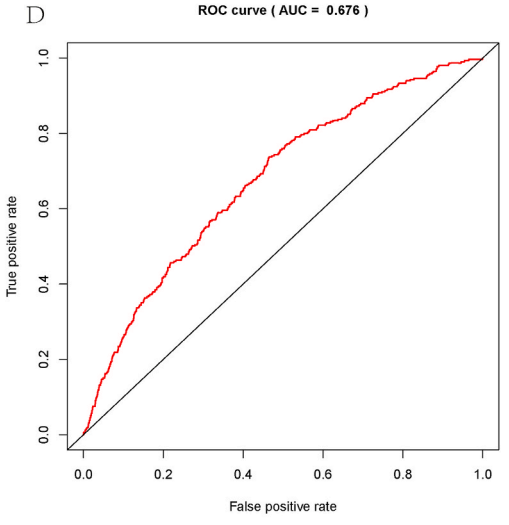


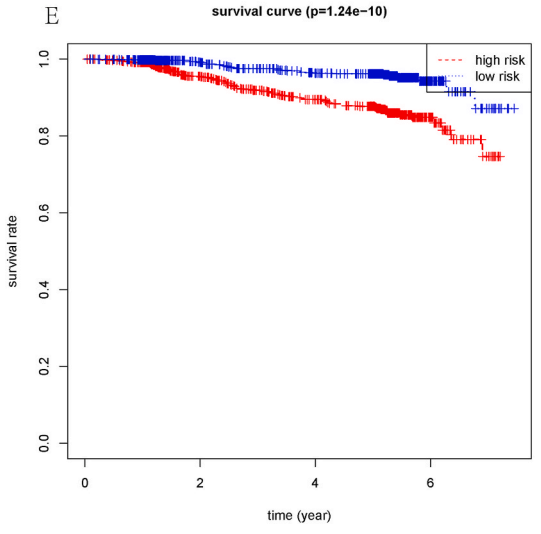
Fig. 3. (A) Epithelial and T cells are divided into 12 clusters. (B) Dot plot of epithelial and T cell marker genes shows the identity of different clusters. (C) Distribution of copy number variation among different clusters.

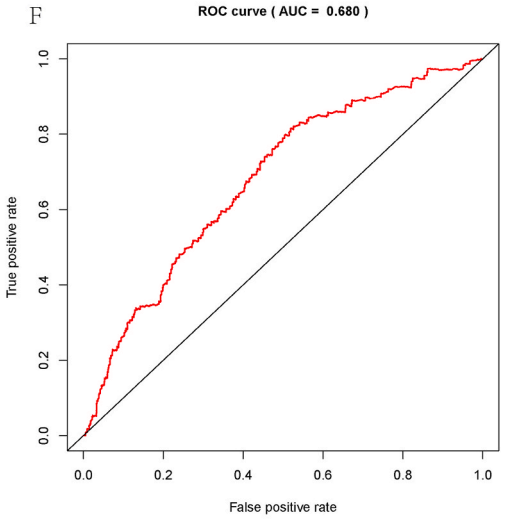












(caption on next page)

Fig. 4. Validation of PRS in breast cancer. (A) Survival curve of TCGA dataset. (B) tROC curve of TCGA dataset. (C) Survival curve of OS of GSE202203 dataset. (D) tROC curve of OS of GSE202203 dataset. (E) Survival curve of RFI of GSE202203 dataset. (F) tROC curve of RFI of GSE202203 dataset.

cells and epithelial cells are accurate. The identified T cells would be used to infer copy number variation in the next step.

3.3. Inferring genes related to malignancy based on copy number variation

Compared with other cell types, canceration of immune cells is relatively rare, so this study used T cells in immune cells as a reference to infer the gene copy number variation in a single cell. We first selected T cells and epithelial cells for cell re-clustering. The "Clustree" algorithm was used to calculate the clustering results of different resolutions to find the appropriate resolution (Fig. S2). T cells and epithelial cells were divided into 12 clusters (Fig. 3A). According to markers, cluster 0, cluster 3, and cluster 11 were T cells, and the remaining 9 clusters were epithelial cells (Fig. 3B).

Further, we used the R package "inferCNV" to infer the copy number variation in each T cell and epithelial cell. The cell cluster with the highest copy number variation score was regarded as the cell cluster with the highest degree of malignancy, and the cell cluster with the lowest copy number variation score was regarded as the cell cluster with the lowest degree of malignancy, thereby identifying the DEGs of the two clusters. According to the results of "inferCNV", it can be seen that the amplification and deletion of copy number in epithelial cells are significantly more than those in T cells and are unevenly distributed on different chromosomes (Fig. 3C). Copy number amplification mainly occurred on chromosomes 1, 4, 8, 16, 19 and 20. Copy number deletions mainly occurred on chromosomes 1, 2, 5, 10, 11, 13, 14 and 15. The copy number variation in each cluster of epithelial cells showed different patterns. Clusters 4 and 1 have the highest and lowest copy number variation scores, respectively. After identifying the DEGs between the two clusters and excluding mitochondrial genes, a total of 185 DEGs were obtained, of which 27 genes were highly expressed in cluster 1 with a low malignant degree and 158 genes were highly expressed in cluster 4 with a high malignant degree. We inferred that these genes are closely related to the malignancy of cells.

3.4. Prediction of breast cancer prognosis based on polygenic risk score

By combining 200 genes from bulk expression profiles and 185 genes from single-cell expression profile, 385 genes were obtained. After removing the genes that do not exist in the two validation datasets, GDC TCGA Breast Cancer and GSE202203, 354 DEGs remained. Univariate Cox analysis showed that there were 26 genes whose p-value was less than 0.05. Multivariate Cox analysis and stepwise regression analysis were performed using these 26 genes to obtain a polygenic risk score containing eight genes, namely *PIGR*, *S100B*, *LEF1*, *ZNF385B*, *WWOX*, *RYR2*, *SLC19A2*, and *HIPK2*. The corresponding regression coefficients constitute vector ***B***:

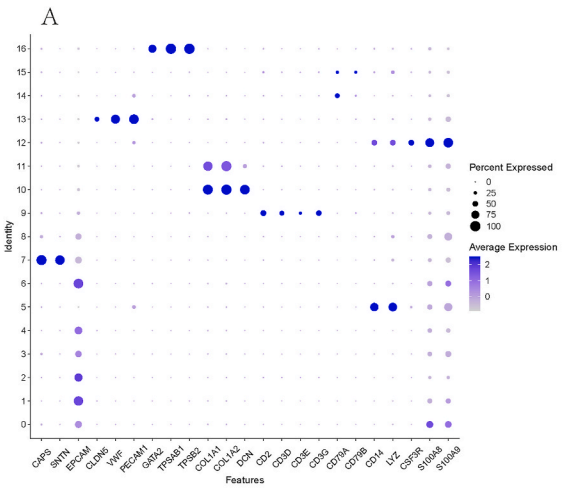
$$B = [-0.0853, -0.1597, -0.1437, -0.1158, 0.2534, 0.2117, -0.2748, -0.2025] \quad (2)$$

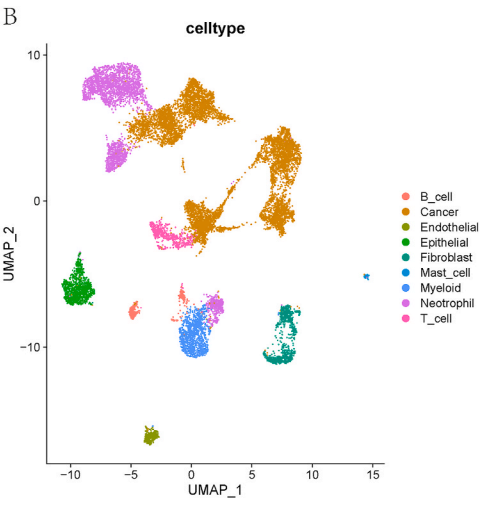
In univariate Cox regression analysis, the factors with Hazard Ratio (HR) greater than 1 are the disease risk factors, and the people exposed to these factors are more likely to have positive events (recurrence, death). On the contrary, the factors with HR values of less than 1 are the protective factors of the disease, and the people exposed to these factors are less likely to have positive events. In multivariate Cox regression analysis, the higher the PRS value, the greater the risk to the patient, so the high expression of genes with positive regression coefficients in PRS is usually associated with poor prognosis, while the low expression of genes with negative regression coefficients in PRS is usually associated with poor prognosis. We found that the risk factors in univariate risk regression have positive regression coefficients in multivariate Cox regression. For example, the regression coefficients of *RYR2* (HR = 1.16) and *WWOX* (HR = 1.23) with HR values greater than 1 in PRS are 0.2117 and 0.2534, respectively. The protective factors in univariate Cox regression analysis have negative regression coefficients in multivariate Cox regression. For example, the regression coefficients of *S100B* (HR = 0.889), *PIGR* (HR = 0.881), *ZNF385B* (HR = 0.877), *SLC19A2* (HR = 0.818), *LEF1* (HR = 0.799), and *HIPK2* (HR = 0.784) with HR values less than 1 in PRS are −0.1597, −0.0853, −0.1158, −0.2748, −0.1437, and −0.2025, respectively. These results indicate that the eight genes involved in PRS exhibited consistency in univariate and multivariate regression analyses.

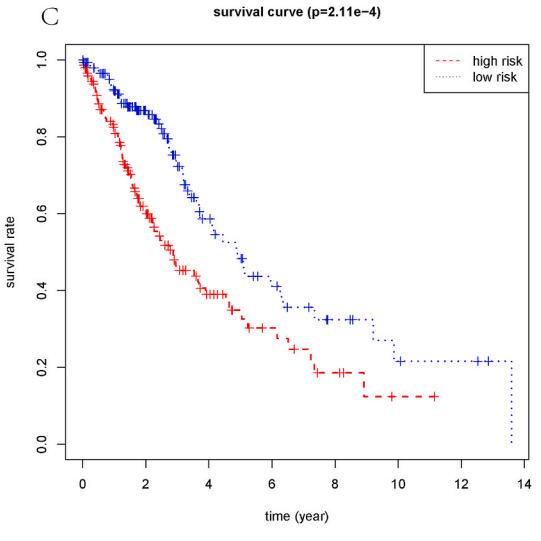
The eight genes in PRS, *PIGR*, and *S100B* were identified in bulk expression profiles and were significantly down-regulated in breast cancer samples. *LEF1*, *ZNF385B*, *WWOX*, *RYR2*, *SLC19A2*, and *HIPK2* were obtained from single-cell expression profiles, and they were significantly up-regulated in highly malignant cell clusters. The results indicate that both the bulk expression profiles and single-cell expression profiles provide information for the construction of the PRS model, and the analysis and processing of these two types of data mentioned in the Materials and methods section are essential components of the model.

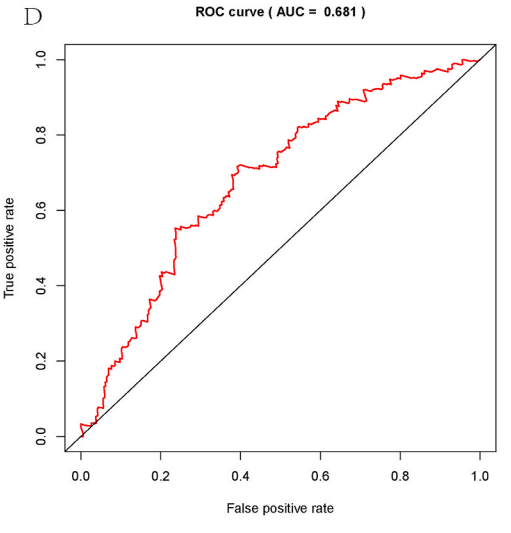
We used the validation set from TCGA to verify the prognosis model and calculated that the median PRS was 0.93, the p-value of the survival curve was 4.94×10^{-3} (Fig. 4A), and the AUC value of tROC at five years was 0.734 (Fig. 4B). Using another validation dataset GSE202203, we calculated the survival curve and tROC curve based on overall survival (OS) and relapse-free interval (RFI), respectively (Fig. 4C–F). These results further demonstrate the effectiveness of the prognosis model.

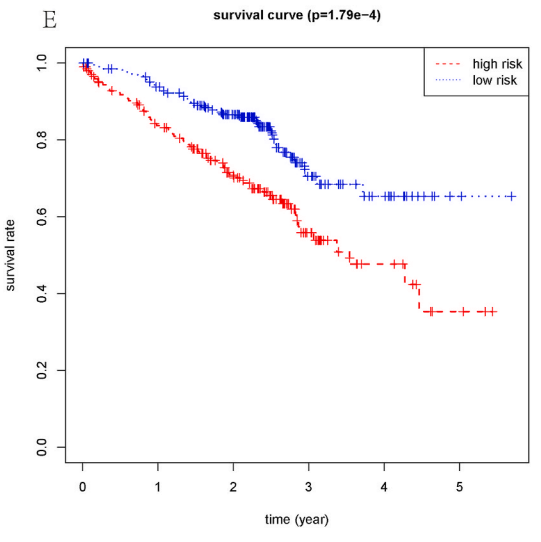
Next, we investigated at the molecular level whether these genes are involved in the malignant transformation of breast cancer. *PIGR* means polymeric immunoglobulin receptor and its high expression is associated with an increased five-year breast cancer survival rate. Sun et al. used the breast cancer cell line MCF7 to conduct a transwell migration assay to study the association between *PIGR* and cancer cell migration. The assay showed that inhibiting the expression of *PIGR* will significantly enhance the migration of MCF7 breast cancer cells, so it is inferred that *PIGR* plays an anti-tumor role by inhibiting the migration of MCF7 breast cancer cells. In

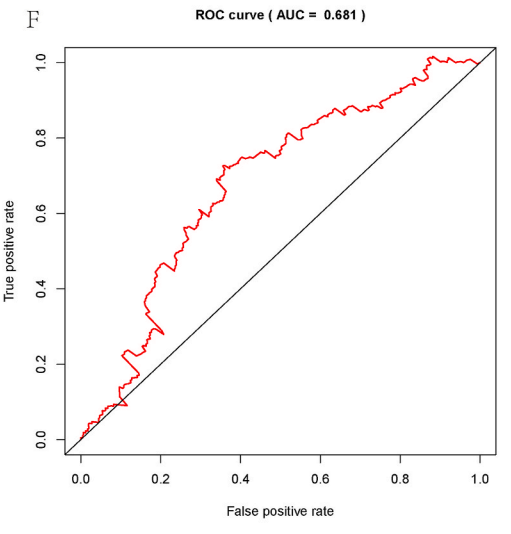












(caption on next page)

Fig. 5. Prognosis model and survival curves of lung cancer. (A) A dot plot of marker genes. (B) UMAP of lung cancer cells. (C) Survival curve of TCGA dataset. (D) tROC curve of TCGA dataset. (E) Survival curve of GSE72094 dataset. (F) tROC curve of GSE72094 dataset.

addition, the GO Term annotations and KEGG pathways of the co-expressed genes of PIGR mainly involve immune responses. In breast cancer, various immune-related molecules are positively correlated with the level of *PIGR*. These results indicate that *PIGR* may also enhance tumor immunity in breast cancer to fight tumors and improve the prognosis of breast cancer patients [23]. S100B is a calmodulin that can be highly expressed in certain tumors, such as gliomas and malignant melanoma. Yen et al. used the Transwell migration assay to study the effect of *S100B* on the migration rate of breast cancer cells [24]. This experiment used recombinant *S100B* protein to treat breast cancer cell lines MDA-MB-231 and Hs578T. The results showed that the migration rate of these two types of breast cancer cell lines was significantly reduced. Therefore, *S100B* may exert anti-tumor effects by inhibiting tumor cell migration. *ZNF385B* is considered a potential transcription inhibitor. Studies have shown that *ZNF385B* can activate *Caspase-8* and *Caspase*-3 by upregulating *PERP* and *FAS/CD95*, affecting tumor suppressor *p53* and mediating cell apoptosis. So, *ZNF385B* may reduce tumor cell formation and metastasis through transcriptional inhibition [25]. Intermediate filament protein vimentin is overexpressed in a variety of epithelial cancers, including breast cancer, and is associated with tumor growth and metastasis. Researchers transfected *HIPK2* into MDA-MB-231 breast cancer cells and proved that overexpression of *HIPK2* can down-regulate vimentin expression and inhibit breast cancer cell invasion [26]. Matrix metalloproteinase-7 (MMP-7) is a small proteolytic enzyme associated with the invasion and metastasis of various cancers. Cyclin D1 is a cell cycle regulatory protein family member, and its abnormal expression can change the cell cycle, stimulating normal cells to transform into cancer cells. Lymph enhancer binding factor-1 (*LEF-1*) is associated with cyclin D1 and MMP-7 gene regulation. Bucan et al. used siRNA to reduce *LEF-1* expression, leading to the downregulation of cyclin D1 and MMP-7 expression, respectively. The results indicate that *LEF-1* mediates the transcription of cyclin D1 and MMP-7, thereby regulating the proliferation of cancer cells [27]. Among the eight inferred critical genes, there is limited conclusive evidence linking specific genes to breast cancer, such as *SLC19A2*, member 2 of solid carrier family 19, and a thiamine transporter. Further research is needed to determine whether they may become potential drug targets or markers of malignant transformation.

3.5. Prognosis model construction and polygenic risk score for lung cancer

The prognosis of different cancers is significantly different. Breast cancer is usually regarded as one of the cancers with a relatively good prognosis because it can be diagnosed at an early stage, and there are many treatment methods, such as surgery, radiotherapy, chemotherapy, and endocrine therapy. Therefore, we contrived to explore whether the workflow in this study can be applied to cancer with a poor prognosis. Suppose this workflow can be generalized to more bulk and single-cell cancer datasets. In that case, it can fully mine these data to identify molecular markers, signaling pathways, and cell types related to disease progression and prognosis. Lung cancer is usually diagnosed later, and its treatment methods are relatively limited, resulting in a generally poor prognosis. Therefore, we chose lung cancer for a comparative study.

Single-cell data from six lung cancer samples were obtained from the GEO database, including GSM4453581, GSM4453589, GSM4453593, GSM4453594, GSM4453615, and GSM4453616 [28]. We used R package harmony (v0.1.1) [29] to remove batch effects and annotated cells according to marker genes as epithelial cells (*CAPS*, *SNTN*), endothelial cells (*CLDN5*, *VWF*, *PECAM1*), T cells (*CD2*, *CD3D*, *CD3E*, *CD3G*), B cells (*CD79A*, *CD79B*), neutrophils (*CSF3R*, *S100A8*, *S100A9*), myeloid cells (*CD14*, *LYZ*), fibroblasts (*COL1A1*, *COL1A2*, *DCN*), mast cells (*GATA2*, *TPSAB1*, *TPSB2*). Cancer cells were negative for regular epithelial cell markers but positive for *EPCAM* (Fig. 5A–B). Using the "FindMarkers" function to analyze the DEGs between epithelial cells and cancer cells, we obtained 1475 genes from single-cell datasets.

Bulk expression data were obtained from two databases, the GEO and UCSC Xena. We downloaded datasets GSE30219, GSE81089, and GSE151103 from GEO database [30], which contain 205 normal tissue samples and 680 lung cancer tissue samples. The data from UCSC Xena were GDC TCGA Lung Adenocarcinoma (LUAD) and GDC TCGA Lung Squamous Cell Carcinoma (LUSC). Merge the two datasets to form a new one, TCGA-LUNG, which includes 108 normal tissues and 1027 tumor tissues. Difference analysis was performed on normal and lung cancer tissues in the above four datasets to obtain significantly up-regulated and down-regulated genes in each dataset.

The validation datasets of the model were served by TCGA-LUNG from UCSC Xena and GSE72094 from GEO because they have clinical information. Similar to the analysis of breast cancer, we used the RobustRankAggreg software package (v1.2.1) to extract 100 up-regulated genes and 100 down-regulated genes (Table S4) from the lung cancer bulk datasets, screened 98 genes related to malignancy from the single-cell datasets (Table S5), merged these genes and removed the genes that did not exist in the datasets TCGA-LUNG and GSE72094, and finally obtained 277 genes.

Then, univariate Cox regression analysis was performed, and the genes were filtered out if their p-values were equal or more than 0.02 in univariate Cox regression analysis. Subsequently, multivariate Cox and stepwise regression analyses were performed to construct a PRS containing six genes: *FAM83A*, *TFAP2A*, *KLK6*, *NEIL3*, *TRHDE*, and *CAPS*. The regression coefficients are the value of the corresponding elements in vector B':

$$B' = [0.0648, 0.0702, 0.0360, 0.1065, 0.1340, -0.0931] \tag{3}$$

Among the six genes, *FAM83A*, *TFAP2A*, *KLK6*, *NEIL3*, and *TRHDE* were identified in bulk expression profiles, and *CAPS* was identified in single-cell expression profile, which indicates that similar to breast cancer analysis, bulk expression profile data and

single-cell expression profile data both contribute to the construction of prognosis model. Previous studies have shown that high expression of *FAM83A* is related to poor prognosis of lung adenocarcinoma patients [31], lung cancer patients with high expression of *TFAP2A* have poor prognosis [32], high expression of *KLK6* is related to low survival rate [33], and high expression of *NEIL3* is related to poor prognosis [34,35]. In brief, most of the genes in our study are supported by existing studies.

The constructed PRS was validated using TCGA-LUNG and GSE72094. In the validation dataset of TCGA-LUNG, the p-value of the survival curve is 2.11×10^{-4} (Fig. 5C). The AUC value at three years is 0.681 (Fig. 5D). In the dataset GSE72094, the p-value of the survival curve is 1.79×10^{-4} (Fig. 5E). The AUC value at five years is 0.681 (Fig. 5F). In summary, the workflow shows good generalization ability on both breast and lung cancers datasets, and thus has the potential to mine other cancer data more extensively.

There have been many experimental studies on the mechanisms by which these six essential genes we identified play a role in lung cancer. Zheng et al. overexpressed *FAM83A* and used cell proliferation, colony formation, and invasion assays to detect the proliferation and invasion of lung cancer cells. The results indicate that overexpression of *FAM83A* increases the activity of β-catenin, target genes of Wnt signaling pathways, and Epithelial-Mesenchymal Transition (EMT). In contrast, the Hippo pathway's activity is downregulated, indicating that *FAM83A* promotes cancer cell proliferation and invasion by regulating the Wnt and Hippo signaling pathways and EMT processes [36]. LncRNA SLC2A1-AS1 is significantly overexpressed in lung adenocarcinoma (LUAD) and is closely associated with overall patient survival. Cui et al. confirmed through ChIP-PCR and RT-qPCR experiments that *TFAP2A* can directly bind to the promoter region of the SLC2A1-AS1 coding gene. Knockout of *TFAP2A* significantly inhibited the transcription of SLC2A1-AS1 in LUAD cells. The author further demonstrated through a colony formation assay that downregulation of SLC2A1-AS1 substantially inhibits cancer cell proliferation. These results indicate that the upregulation of SLC2A1-AS1 mediated by *TFAP2A* promotes cancer cell proliferation in lung squamous cell carcinoma (LUSC) [37]. Experiments have shown that the mechanism of action of the *KLK6* gene is that it promotes the proliferation of Non-Small Cell Lung Cancer (NSCLC) cells and restricts their apoptosis via an activation cascade initiated by Protease-Activated Receptor 2 (*PAR2*) and involving the ligand-dependent transactivation of Epidermal Growth Factor Receptor (*EGFR*) [38]. Colony formation and *CCK8* assays showed that knocking *NEIL3* inhibited NSCLC cell proliferation. Transwell and wound healing assays indicated that knocking down *NEIL3* inhibits the invasion and migration ability of NSCLC cells. In lung cancer, the PI3K/AKT/mTOR signaling pathway is a critical regulatory pathway leading to malignant phenotype and drug resistance. Gene Set Enrichment Analysis shows abnormal activation of the PI3K/AKT/mTOR pathway, G2/M checkpoint, and E2F target in *NEIL3* patients. In contrast, Western blot shows that *NEIL3* can partially activate the PI3K/AKT/mTOR signaling pathway, which may be why *NEIL3* promotes cancer [39].

4. Discussion

Our study integrated single-cell and bulk expression profiles to identify and analyze cancer prognosis-related genes, constructed a cancer prognosis model, and confirmed the method's effectivenessin two types of cancer. In previous studies, breast cancer prognosis models often used factors such as lymph node status, patient age, tumor size, cancer grade, and estrogen receptor status as prognosis indicators. However, these pieces of clinical information are not in a unified numerical form, so some arbitrary factors will inevitably be introduced, which poses obstacles to quantitative calculations. Our model alleviates some of the limitations of traditional methods and, combined with routine clinical pathological evaluations, can more objectively predict patients' survival time and quality of life. Our study has a specific biological basis for identifying prognosis-related genes and calculating PRS from gene expression profiles. There are highly complex gene regulatory networks within organisms. Both endogenous and exogenous stimuli and changes can affect the expression of different genes in various cells within an organism. During the process of canceration, the expression patterns of genes may change, with some genes up-regulated and some genes down-regulated. These changes in gene expression patterns contain rich information and reflect the state of patients. Therefore, the essential genes related to carcinogenesis can be used to construct a model to predict the prognosis by measuring the expression values of different genes in patients. Our prognosis model comprehensively utilizes the advantages of bulk and single-cell expression profiles. To avoid the inaccuracy of prognosis models in some populations, we use DEGs with significant changes in multiple datasets as input factors for model construction. At the same time, because the bulk expression profiles can only reflect the overall expression level of the sequenced tissue, ignoring cell heterogeneity, some crucial DEGs may be masked. We also identified two clusters of cells with high and low malignancy based on copy number variation from the single-cell dataset. We extracted the DEGs between the two clusters, supplementing the gene set to construct the prognosis model. In the breast cancer prognosis model, six genes are from the single-cell dataset, proving that combining bulk and single-cell expression profiles for analysis is reasonable.

Using our cancer prognosis model, doctors can infer the prognosis of breast cancer patients and adjust the nursing and prognosis strategies for patients according to the risk assessment of patients to improve the treatment effect and improve the survival time and life quality of patients. At the same time, the eight genes used to construct PRS can be used as candidate biomarkers in breast cancer research, guiding the design of anti-cancer drugs, predicting treatment responses, and providing a basis for survival prediction and personalized treatment. The workflow and code of this study can be easily extended to other cancers, providing convenience for researchers unfamiliar with programming.

The following research work can be carried out from two aspects. Firstly, categorize the different subtypes of cancer, such as small cell lung cancer (SCLC), ductal papillary carcinoma (DCIS), and invasive ductal carcinoma (IDC), and construct prognostic models using expression data from different subtypes of cancer to aid in more targeted diagnosis and treatment. Secondly, integrating multimodal data, such as spatial omics data and single-cell chromatin accessibility data, complements information from different modalities to enhance the predictive performance of prognostic models.

Funding

This project was supported by the National Natural Science Foundation of China (31871269).

Data available

Data will be made available on request. All the code we used can be found and downloaded from: https://github.com/Shengbao-Bao/prognosis.

CRediT authorship contribution statement

Shengbao Bao: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation. **Yaxin Fan:** Validation, Data curation. **Yichao Mei:** Visualization, Validation, Software. **Junxiang Gao:** Writing – review & editing, Supervision, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e25640.

References

[1] S. Shukla, J.R. Evans, R. Malik, et al., Development of a RNA-Seq based prognostic signature in lung adenocarcinoma, JNCI, Journal of the National Cancer Institute 109 (2017) djw200.
[2] G. Liu, X. Zeng, B. Wu, et al., RNA-Seq analysis of peripheral blood mononuclear cells reveals unique transcriptional signatures associated with radiotherapy response of nasopharyngeal carcinoma and prognosis of head and neck cancer, Cancer biology & therapy 21 (2020) 139–146.
[3] G. Chen, B. Ning, T. Shi, Single-cell RNA-seq technologies and related computational data analysis, Frontiers in genetics 10 (2019) 317.
[4] S. Slovin, A. Carissimo, F. Panariello, et al., Single-cell RNA sequencing analysis: a step-by-step overview, RNA Bioinformatics (2021) 343–365.
[5] M. Karaayvaz, S. Cristea, S.M. Gillespie, et al., Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq, Nature communications 9 (2018) 3588.
[6] A. Jindal, P. Gupta, Jayadeva, et al., Discovery of rare cells from voluminous single cell expression data, Nature communications 9 (2018) 4719.
[7] S. Zhou, Y.-e. Huang, H. Liu, et al., Single-cell RNA-seq dissects the intratumoral heterogeneity of triple-negative breast cancer based on gene regulatory networks, Molecular Therapy-Nucleic Acids 23 (2021) 682–690.
[8] A. Gakwaya, J. Kigula-Mugambe, A. Kavuma, et al., Cancer of the breast: 5-year survival in a tertiary hospital in Uganda, British journal of cancer 99 (2008) 63–67.
[9] C. Zappa, S.A. Mousa, Non-small cell lung cancer: current treatment and future advances, Translational lung cancer research 5 (2016) 288.
[10] J. Wang, X. Zhang, A.H. Beck, et al., Alcohol consumption and risk of breast cancer by tumor receptor expression, Hormones and Cancer 6 (2015) 237–246.
[11] J.-W. Chang, W.-H. Kuo, C.-M. Lin, et al., Wild-type p53 upregulates an early onset breast cancer-associated gene GAS7 to suppress metastasis via GAS7-CYFIP1-mediated signaling pathway, Oncogene 37 (2018) 4137–4150.
[12] H. Dalal, M. Dahlgren, S. Gladchuk, et al., Clinical associations of ESR2 (estrogen receptor beta) expression across thousands of primary breast tumors, Scientific Reports 12 (2022) 4696.
[13] S.-Q. Liu, Z.-J. Gao, J. Wu, et al., Single-cell and spatially resolved analysis uncovers cell heterogeneity of breast cancer, Journal of Hematology & Oncology 15 (2022) 19.
[14] M.E. Ritchie, B. Phipson, D. Wu, et al., Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic acids research 43 (2015) e47–e47.
[15] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome biology 15 (2014) 1–21.
[16] R. Kolde, S. Laur, P. Adler, et al., Robust rank aggregation for gene list integration and meta-analysis, Bioinformatics 28 (2012) 573–580.
[17] Y. Hao, S. Hao, E. Andersen-Nissen, et al., Integrated analysis of multimodal single-cell data, Cell 184 (2021) 3573–3587, e3529.
[18] C.S. McGinnis, L.M. Murrow, Z.J. Gartner, DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors, Cell systems 8 (2019) 329–337, e324.
[19] D. Aran, A.P. Looney, L. Liu, et al., Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage, Nature immunology 20 (2019) 163–172.
[20] A.P. Patel, I. Tirosh, J.J. Trombetta, et al., Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma, Science 344 (2014) 1396–1401.
[21] T. Therneau, T. Lumley, R Survival Package, R Core Team, 2013.
[22] P.J. Heagerty, Y. Zheng, Survival model predictive accuracy and ROC curves, Biometrics 61 (2005) 92–105.
[23] Z. Sun, B. Tan, K. Dong, et al., PIGR Predicts Good Clinical Outcomes and Plays a Tumor Suppressor Role in the Development of Breast Cancer via Enhancing Tumor Immunity, 2023.
[24] M.-C. Yen, Y.-C. Huang, J.-Y. Kan, et al., S100B expression in breast cancer as a predictive marker for cancer metastasis, International journal of oncology 52 (2018) 433–440.
[25] N. Yan, C. Liu, F. Tian, et al., Downregulated mRNA expression of ZNF385B is an independent predictor of breast cancer, International Journal of Genomics 2021 (2021).
[26] C. Nodale, M. Sheffer, J. Jacob-Hirsch, et al., HIPK2 downregulates vimentin and inhibits breast cancer cell invasion, Cancer biology & therapy 13 (2012) 198–205.
[27] V. Bucan, K. Mandel, C. Bertram, et al., LEF-1 regulates proliferation and MMP-7 transcription in breast cancer cells, Genes to Cells 17 (2012) 559–567.

[28] F. Wu, J. Fan, Y. He, et al., Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer, Nature communications 12 (2021) 2540.
[29] I. Korsunsky, N. Millard, J. Fan, et al., Fast, sensitive and accurate integration of single-cell data with Harmony, Nature methods 16 (2019) 1289–1296.
[30] S.A. Willis-Owen, C. Domingo-Sabugo, E. Starren, et al., Y disruption, autosomal hypomethylation and poor male lung cancer survival, Scientific Reports 11 (2021) 12453.
[31] J. Yu, M. Hou, T. Pei, FAM83A is a prognosis signature and potential oncogene of lung adenocarcinoma, DNA and Cell Biology 39 (2020) 890–899.
[32] Y. Xiong, Y. Feng, J. Zhao, et al., TFAP2A potentiates lung adenocarcinoma metastasis by a novel miR-16 family/TFAP2A/PSG9/TGF-β signaling pathway, Cell Death & Disease 12 (2021) 352.
[33] H.V.h. Nathalie, P. Chris, G. Serge, et al., High kallikrein-related peptidase 6 in non-small cell lung cancer cells: an indicator of tumour proliferation and poor prognosis, Journal of cellular and molecular medicine 13 (2009) 4014–4022.
[34] O.T. Tran, S. Tadesse, C. Chu, et al., Overexpression of NEIL3 associated with altered genome and poor survival in selected types of human cancer, Tumor Biology 42 (2020) 1010428320918404.
[35] C. Zhao, J. Liu, H. Zhou, et al., Construction of NEIL3 as a Prognostic Biomarker and Co-expressed Prognostic Signature in Lung Adenocarcinoma, 2020.
[36] Y.-W. Zheng, Z.-H. Li, L. Lei, et al., FAM83A promotes lung cancer progression by regulating the Wnt and Hippo signaling pathways and indicates poor prognosis, Frontiers in Oncology 10 (2020) 180.
[37] Y. Cui, C. Zhang, S. Ma, et al., TFAP2A-induced SLC2A1-AS1 promotes cancer cell proliferation, Biological chemistry 402 (2021) 717–727.
[38] N. Michel, N. Heuzé-Vourc'h, E. Lavergne, et al., Growth and survival of lung cancer cells: regulation by kallikrein-related peptidase 6 via activation of proteinase-activated receptor 2 and the epidermal growth factor receptor, Biological Chemistry 395 (2014) 1015–1025.
[39] H. Huang, Q. Hua, NEIL3 mediates lung cancer progression and modulates PI3K/AKT/mTOR signaling: a potential therapeutic target, International Journal of Genomics 2022 (2022).