Development and validation of a prognostic model based on single-cell RNA-seq in Wilms Tumor of children

Tian-Qu He
Hunan Children's Hospital
Yao-Wang Zhao
Hunan Children's Hospital
Yu Liu
Hunan Children's Hospital
Lei Tu
Hunan Children's Hospital
Feng Ning
Hunan Children's Hospital
Jun He (hjys804808@163.com)
Hunan Children's Hospital

Research Article

Keywords: single-cell RNA-seq, Wilms tumor, malignant cells, prognostic Biomarkers

Posted Date: May 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1668982/v1

Abstract

Objective: To analyze the heterogeneity between different cell types in Wilms Tumor (WT) tissue of children, and identify the differentially expressed genes (DEGs) of malignant tumor cells, thereby establishing a prognostic model.

Methods: The single-cell sequencing data of pediatric WT tissues were downloaded from the public database. Data filtration and normalization, principal component analysis (PCA), and TSNE cluster analysis were performed using the Seurat package of R. Cells were divided into different clusters, malignant tumor cells were extracted, and DEGs were obtained. Then, the pseudo-time trajectory analysis was performed. Prognostic biomarkers were determined by univariate COX regression analysis, multivariate COX regression analysis, and LASSO regression analysis. Kaplan-Meier survival analysis and receiver operator characteristic (ROC) curve analysis were performed on the prognostic biomarkers. Combined with the prognostic biomarkers and clinical characteristics, a nomogram was generated to predict the WT prognosis of children. The prognostic power of the prognostic model was validated in the external datasets.

Results: Cells in the WT tissue were divided into 10 clusters. 3 prognostic biomarkers that affected the survival time of patients were screened from 215 DEGs in malignant tumor cells, and a nomogram was constructed using the 3 genes and clinical characteristics. The AUC values of 3- and 5-year disease-free survival were 0.756 and 0.734, respectively. In the external validation dataset, the AUC value of this nomogram model was 0.826.

Conclusion: Based on single-cell RNA-seq, we recognized cell clusters in the WT tissue of children, identified prognostic biomarkers in malignant tumor cells, and established a comprehensive prognostic model. Our findings might provide new ideas and methods for the diagnosis and treatment of WT

Introduction

Wilms Tumor (WT) is the most common kidney cancer in children, and the fourth most common childhood cancer [1]. It is genetically heterogeneous [2]. WT affects 1 in 10,000 children. Its histology resembles that of a developing kidney, and is usually triphasic with blastemal, stromal, and epithelial components [3]. WT is a malignant solid tumor, mainly composed of stroma, germ and epithelium, including epithelial tissue, muscle tissue, connective tissue, skeletal tissue and nerve tissue. Abdominal mass is its main clinical symptom. When the mass is small, there are no obvious symptoms, so that WT is easy to be ignored. When the mass is found, WT may have developed to an advanced stage. However, in the early stage of WT, distant metastasis such as lung, liver, and brain can also occur, which seriously affects the life and health of children. Therefore, it is necessary to conduct in-depth research on the factors associated with the occurrence and development of WT.

At present, genomic and transcriptomic studies on WT tissues mainly use bulk profiling techniques to investigate the effects of genes [4, 5], miRNAs [6], and lncRNAs [7, 8] on tumor tissues and patient

survival time. However, these studies typically rely on data from the bulk profiling, limiting their ability to accurately capture tumor heterogeneity. Therefore, an in-depth understanding of cellular heterogeneity and the interaction between WT cells and their microenvironment may lead to the development of new therapeutic approaches for treating WT.

Single-cell RNA sequencing (scRNA-seq) analysis has emerged as a powerful tool for revealing cellular diversity and cell-to-cell communication at single-cell resolution. Recently, scRNA-seq has been applied to dissect the complex tumor and immune landscapes of some cancers, including glioblastoma, breast cancer, lung cancer, head and neck cancer, pancreatic ductal adenocarcinoma, and liver cancer. This technique improves our understanding of cellular heterogeneity and facilitates the screening of promising molecular targets to guide antitumor therapy. However, tumor heterogeneity and the interaction between malignant cells and normal cells at single-cell resolution in human WT remains poorly understood.

Based on scRNA-seq, this present paper analyzed the data at the single cell level, which solved the problem that tissue samples cannot obtain information on the heterogeneity among different cells. In this paper, 10 clusters were identified in the single-cell sample data of WT patients, and the differences in gene expression between the selected malignant tumor cells and other cells were analyzed, and the molecular functions, biological processes, cellular components, and signal pathways involved were determined by Gene Ontology (GO) function enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. Prognostic biomarkers were determined by univariate COX regression analysis, multivariate COX regression analysis, and LASSO regression analysis. Kaplan-Meier survival analysis and receiver operator characteristic (ROC) curve analysis were performed on the prognostic biomarkers, and a nomogram was finally constructed and and calibrated. Finally, the accuracy of these results was validated in other datasets. The details were reported as follows.

1. Materials And Methods

1.1. Data collection and pre-processing

The training dataset was derived from the published single-cell sequencing data [9], in which 3 pediatric WT samples, 5 adult clear cell (ccRCC) or papillary renal cell carcinoma (pRCC) samples, 2 healthy fetal samples, and 1 healthy adolescent were analyzed. The data of WT patients were extracted as the research object, and a total of 2967 single-cell sequencing data of WT patients were obtained. Then the Seurat package in the R software was used to pre-process the data, and the "CreateSeuratObject" function was used to build a Seurat object.

The survival data and gene expression data were obtained from the WT project in the TARGET database (https://ocg.cancer.gov/). The TARGET database is a childhood tumor database designed to use a comprehensive genomic approach to identify molecular changes in the occurrence and development of difficult-to-treat childhood cancers, including acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), kidney tumors (KT), neuroblastoma (NBL), and osteosarcoma (OS). The data of 124 WT

patients were selected as the experimental data set in this study. The GSE73209 and GSE11024 data sets were downloaded from the GEO data platform (https://www.ncbi.nlm.nih.gov/geo), and were used as the validation data sets.

1.2 Quality control, data filtering, and normalization

The “PercentageFeatureset” function was used to calculate the percentage of mitochondrial genes, and the "FilterCells" function was used to filter the Seurat object, with the retention criteria of 200 < features < 2500, and the percentage of mitochondrial genes < 5%. The "NormalizeData" function was used to normalize the data with the *normalization.method* = "*LogNormalize*".

1.3 Identification of genes with highly variable expression between cells

Some genes, whose expression varies widely between cells, are called hypervariable genes. The "FindVariableGenes" function was used to calculate hypervariable genes, and the "vst" method was used to select the top 2000 hypervariable genes. The data were further scaled using the "ScaleData" function with the *features* = "*all.genes*".

1.4 Principal component analysis (PCA)

The "RunPCA" function was used to perform PCA on highly variable genes, the "ScoreJackStraw" function was used to score each principal component (PC), the "JackStrawPlot" function was used to visualize the score of each PC, and the "EbowPlot" function was further used for visual analysis of each principal component. The dimension of PCs for cluster analysis was selected accordingly. Then, the “FindClusters” function was used to perform cell clustering analysis, and the parameter resolution was set as 0.4.

1.5 T-distributed stochastic neighbor embedding (TSNE) cluster classification analysis

The TSNE algorithm was used for dimensionality reduction, and cluster classification analysis. Then, the "TSNEPlot" function was used to visualize cell clusters. The SingleR package in the R language was used to annotate the cell clusters by checking the expression of common cell grouping marker genes in tumor tissue, the results of cell clusters annotation were checked and confirmed, and the cell clusters were finally determined comprehensively.

1.6 Pseudo-time trajectory analysis

Investigation of the differentiation trajectories and corresponding genes in various cell populations may shed light on the molecular mechanisms underlying cancer development. Pseudo-time trajectory analysis of single-cell data was performed using the monocle package to determine differentiation trends and differentiation paths of cell clusters.

1.7 Identification of differentially expressed genes (DEGs) in malignant tumor cells

The "FindMarkers" function was used to determine the DEGs between malignant tumor cells and other cells, with the screening criteria of |logFC|>1 and adjPval < 0.05. The obtained DEGs were subjected to enrichment analysis to determine the possible changes in gene functions and signaling pathways caused by these DEGs. Gene Ontology (GO) functional enrichment analysis (including the molecular function (MF), biological processes (BP), and cellular component (CC)), and Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathway enrichment analysis were performed using the Metascape online tool.

1.8 Univariate COX regression analysis

In order to preliminarily screen out the genes associated with WT prognosis, univariate Cox regression analysis was performed on the DEGs to calculate the hazard ratio (HR) and 95% confidence interval of the candidate genes using the "survival" package of the R language. P < 0.05 was considered statistically significant.

1.9 LASSO regression and multivariate COX regression analysis

In order to further select variables and avoid the over-fitting of genes obtained by univariate Cox regression analysis, the "glmet" package of R was used to perform LASSO regression analysis on the DEGs screened by univariate Cox regression. In this present study, 10-fold cross-validation was used to determine the value of λ during the model construction, taking the λ with the smallest partial likelihood deviation as the optimal λ. The selected genes were then subjected to multivariate Cox regression analysis, and finally genes with the statistical significance were selected to establish a prognosis prediction model, and a risk score calculation formula was obtained.

The risk score of each case was calculated using this following formula, Risk score = ExpGENE1 × β1 + ExpGENE2 × β2 +…+ ExpGENEn × βn, where “Exp” and “β” represent the corresponding gene expression level, and the regression coefficient from the multivariate Cox analysis, respectively [10].

1.10 ROC curve analysis

In order to evaluate the predictive power of the prognostic model, this study firstly performed KM survival analysis on the high-risk and low-risk groups; secondly, the ROC curves of the 3-year and 5-year disease-free survival (DFS) of WT patients were drawn, and the area under the curve (AUC) values of the 3-year and 5-year DFS were calculated by the "survival" and "timeROC" packages. When the AUC value is less than 0.5, the accuracy of the model is not significant; when the AUC value is greater than 0.7, the accuracy of the model is moderate; when the AUC value is greater than 0.9, the accuracy of the model is quite high.

1.11 Nomogram construction and calibration

Combining prognostic genes and clinical features, a nomogram was generated to predict prognosis of WT patients. The nomogram was constructed according to the regression coefficients obtained from the multivariate COX regression analysis using the “rms” package of R. The nomogram prediction probabilities against the observed rates was visualized by drawing a calibration curve, and the predictive power of the nomogram were evaluated by the ROC curve. The proportional hazard assumption was tested by Kaplan–Meier curves.

1.12 Validation of external dataset

GSE73209 and GSE11024 from GEO were selected to verify the predictive ability of this nomogram model. The GSE73209 contains 32 WT tissue samples and 6 normal tissue samples; the GSE11024 contains 27 WT tissue samples and 12 normal tissue samples.

1.13 Statistical Analysis

Kaplan–Meier methodology was used to compare the differences of DFS among different groups, and Kaplan–Meier curves were generated. Univariate and multivariate Cox analyses, together with LASSO regression were performed to screen independent predictors of DFS. All statistical analyses were performed using R software (v4.0.3). P value less than 0.05 was considered statistically significant.

2. Results

The flow chart and principal findings of the comprehensive analysis are shown in the Fig. 1.

2.1. Quality control of the single-cell data

The quality control chart is shown in Fig. 2A, where the range of detected gene numbers and the sequencing count of each cell are illustrated. We accordingly retained cells with a percentage of mitochondrial sequencing count < 5% and 200 < features < 2500. Finally, 2695 high-quality cells were retained.

In addition, statistics on the data found that there was a significant correlation between features and counts (Pearson's r = 0.85), while there was no correlation between counts and the percentage of mitochondrial genes (Pearson's r = -0.16), as shown in Fig. 2B and 2C

2.2 Identification of highly variable genes

The top 2000 highly variable genes were screened out after analysis, and the top 10 were: *TPSAB1*, *TPSB2*, *COL1A1*, *LUM*, *COL1A2*, *COL3A1*, *C1QC*, *HBG1*, *JCHAIN*, and *HLA-DQB2*, as shown in Fig. 2D.

2.3. PCA

Furthermore, we used the principal component analysis (PCA) method and screened out the significantly associated genes in each component. The top 30 significantly associated genes are shown by dot plot in

the Fig. 2E and 2F. The heatmap of the top 10 genes in top 15 principal components is shown in Fig. 2G.

The PCA, a linear dimensionality reduction method, was utilized to identify significantly available dimensions of data sets with estimated P value. The JackStraw Plot and Elbow plot showed tha 10 PCs were appropriate (Fig. 2H and 2I).

2.4. TSNE cluster classification analysis

The TSNE cluster classification analysis was performed to divide cells into 10 clusters, as shown in Fig. 3A. The SingleR package was used to annotate the cells of each cluster by combining the gene expression of several common cell types, and finally the cell type of each cluster was determined, as shown in Fig. 3B. The most significant marker genes in each cluster were shown in Fig. 3C, and the top 10 DEGs in each cluster were shown in Fig. 3D

2.5. Pseudo-time trajectory analysis

Pseudo-time trajectory analysis was performed by the “monocle” package, as shown in Fig. 3E, which suggested that the differentiation of cells was mainly divided into three directions. Cluster3 and Cluster7 were on the same branch, and Cluster7 was located at the end of this branch, which implied that Cluster7 was more developed and mature tumor cells.

2.6. Analysis of DEGs in malignant tumor cells

Malignant tumor cells were the focus of this present research. Cluster3 and cluster7 were identified as malignant tumor cells. The “FindMarkers” function in the Seurat package was used to find the DEGs between malignant cells and other cells, and finally 215 DEGs were obtained.

The results of GO enrichment analysis showed that DEGs were mainly enriched in positive regulation of cell migration, blood vessel development and other functions that may be related to carcinogenesis (Fig. 3F). The results of KEGG signaling pathway enrichment analysis showed that DEGs were mainly enriched in signaling pathways that may be related to carcinogenesis, such as apoptosis, transcriptional misregulation in cancer, and antigen processing and presentation (Fig. 3G).

2.7. Identification of prognosis- associated *genes*

Based on the above DEGs, and the corresponding clinical and gene expression information from TARGET, 13 genes associated with the prognosis of WT were preliminarily identified by univariate Cox regression analysis, and their P values were all less than 0.05, including *BTG1, CXCR4, DNAJA1, EIF3M, FBXO21, NES, NPNT, PLTP, PRDX1, PTGDS, PTPRO, TAGLN*, and *TUBB*. LASSO regression analysis was further performed on these 13 genes, and finally 12 genes were determined to be included in this study, namely: *BTG1, CXCR4, DNAJA1, EIF3M, FBXO21, NES, NPNT, PLTP, PRDX1, PTGDS, PTPRO*, and *TAGLN* (Fig. 4A and 4B). Subsequently, multivariate COX regression analysis was performed on these 12 genes, the results of which showed that *DNAJA1, NES*, and *TAGLN* were significantly associated with patient

survival time (P < 0.05) (see Table 1). Kaplan–Meier survival analysis was performed on DNAJA1, NES, and TAGLN, and it was found that the P values of the three genes were all less than 0.05, indicating that these three genes were significantly associated with the survival of patients, and might be used as prognostic marker genes (Fig. 4C-4E). The prognostic risk score of each WT patient was calculated based on these 3 genes. The calculation formula was Risk score = Exp_{DNAJA1}×0.9166 + Exp_{TAGLN}×0.5986 + Exp_{NES}×(-1.0138). All patients were divided into high-risk and low-risk groups according to the median risk score, and a prognostic model was constructed. To further evaluate the predictive power of the prognostic model, Kaplan–Meier survival analysis was performed, the results of which showed that the high-risk group had a 3-year survival probability of 27%, a 5-year survival probability of 27%, and the low-risk group had a 3-year survival probability of 70% and a 5-year survival probability of 63%. DFS time in the low-risk group was longer than that in the high-risk group (P < 0.05, Fig. 4F). The AUC value of the 3-year DFS was 0.733, and the AUC value of the 5-year DFS was 0.709 (see Fig. 4G).

Table 1
multivariate Cox regression analyses of prognostic genes

gene	coef	exp (coef)	95%CI		P
DNAJA1	0.9166	2.5007	1.3897	4.4999	0.0022
TAGLN	0.5986	1.8195	1.0229	3.2365	0.0416
NES	-1.0138	0.3629	0.1988	0.6624	0.0009

2.8. Identification of prognosis-associated clinical characteristics

Univariate Cox regression analysis was performed using DFS as the dependent variable, and age, gender and race as covariates. The results showed that gender could be used as an independent prognostic factor (P = 0.0013), and the prognosis of male patients was worse than that of female patients, as shown in Table 2. In addition, Kaplan–Meier survival analysis was performed on the clinical characteristics, and it was also found that gender was significantly associated with patient survival, as shown in Fig. 4H.

Table 2
multivariate Cox regression analysis of clinical characteristics

Clinical characteristics	Multivariate Cox regression analysis				
	coef	exp(coef)	95%CI		P
Age(> = 5)	-0.477	0.620	0.334	1.151	0.13
Gender(Male)	0.937	2.556	1.445	4.522	0.0013
Race(White)	-0.024	0.976	0.526	1.811	0.939

2.9. Nomogram construction

Gender, *DNAJA1*, *NES*, and *TAGLN* were used to construct the nomogram (see Fig. 5A); the calibration curve suggested that the model had good predictive power for 3- and 5-year survival (see Fig. 5B). The AUC value for the 3-year DFS was 0.756, and the AUC value for the 5-year DFS was 0.734 (see Fig. 5C). In addition, the Kaplan–Meier survival curves of *DNAJA1*, *NES*, *TAGLN*, and Gender suggested that the proportional hazard assumption was valid (see Fig. 4C-4E, 4H).

2.10. Validation of external dataset

The ROC curve also showed that the AUC value of this nomogram model was 0.826 (Fig. 5D), suggesting a moderate prediction value.

Discussion

Wilms Tumor or nephroblastoma is the most common renal tumor in children, and is associated with different congenital anomalies and syndromes [11]. WT is the most common primary malignant tumor of the urinary system in children, with an incidence of about 1 in 10,000 [12], accounting for more than 90% of renal malignant tumors in children [13]. There is significant heterogeneity between malignant tumor cells and other cells in the WT tissue, and single-cell sequencing technology can study the differences between different types of cells at the level of single-cell resolution. The previous researches [14–16] on the construction of prognostic nomogram for WT was based on clinical information database or based on bulk seq sequencing technology, while no prognostic prediction model for malignant tumor cells has been reported. To our knowledge, this present study is the first to screen out malignant tumor cells based on single-cell sequencing technology to construct the prognostic nomogram.

In the previous studies on the nomogram prognostic model of WT, the predictive accuracy ranged from 0.656 to 0.879 [14–16]. Tang *et al.* [16] constructed nomograms to forecast overall survival and cancer-specific survival (CSS) of children with WT based on Surveillance, Epidemiology, and End Results (SEER) database, where five predictors were included, such as age, tumor laterality, size, stage, and surgery, the AUC values for 3- and 5-year overall survival were 0.659 and 0.656, and the AUC values for 3- and 5-year

CSS were both 0.677. Pan *et al.* [15] constructed a nomogram to predict the cancer-specific survival (CSS) of WT patients based on SEER database, where age, the number of examined LNs, SEER stage, and tumor size were included, the AUC values for 3- and 5-year CSS were 0.755 and 0.749, respectively. He *et al.* [14] screened four autophagy-related genes based on the TCGA database, and a nomogram was constructed together with first event, stage and histology. Due to the introduction of more variables, the AUC values for 3-, and 5-year survival based on this nomogram were 0.879, and 0.856, respectively. These previous studies did not consider the effect of heterogeneity between malignant tumor cells and other cells on the results, and thus they have certain limitations. In our present study, based on the malignant tumor cells, the AUC values for 3- and 5-year DFS were 0.756, and 0.734, respectively

In this present paper, single-cell transcriptome analysis was used to group cells based on single-cell sequencing data. Malignant cells were screened out, and compared with other cell data, the DEGs were obtained. GO enrichment analysis and KEGG pathway enrichment analysis were performed on these DEGs. Prognostic biomarkers were determined by univariate COX regression analysis, multivariate COX regression analysis, and LASSO regression analysis. Kaplan–Meier survival analysis and ROC analysis were performed on these prognostic biomarkers, and finally a prognosis model and nomogram were constructed, and the results were calibrated. The results of functional enrichment analysis and signaling pathway enrichment analysis of 215 DEGs showed that malignant tumor cells were significantly enriched in functions or pathways such as blood vessel development, positive regulation of cell migration, and transcriptional misregulation in cancer. As we know, transcriptional misregulation can result in cell canceration, and cancerous cells can promote tumor development by promoting angiogenesis, cell migration and other processes.

In this present study, we identified 3 prognostic biomarkers of malignant tumor cells in the WT tissue, including *DNAJA1*, *TAGLN*, and *NES*.

DNAJA1, a member of J-domain containing proteins or heat shock protein 40, is evidenced to prevent unfolded mutp53 from proteasomal degradation, which is associated with several cancers [17, 18]. It was found that *DNAJA1* promoted tumor metastasis by accumulating unfolded mutp53 [19]. *TAGLN* is a regulator of the actin cytoskeleton, and affects the survival, migration, and apoptosis of various cancer cells to varying degrees [20]. Overexpression of *TAGLN* is associated with cell infiltration, which in turn promotes tumor metastasis [21, 22]. *NES* encodes a member of the intermediate filament protein family. At present, there are few studies on it, and it can be used as a new potential gene for in-depth research.

However, there are still some limitations in our study. This research is developed only based on the public databases, and lack of the validation form in vivo or in vitro experiments. Therefore, it is necessary to further confirm our findings by in vivo or in vitro experiments in the future.

In conclusion, this study detected the biomarkers of malignant tumor cells in the WT tissue, developed a novel nomogram to predictive the WT prognosis based on the candidate genes and clinical information, and validated it in other datasets. It might provide useful help for the diagnosis and treatment of WT in the future.

Declarations

Consent for publication: N/A

Acknowledgements: N/A

Funding: Scientific Research Project of Hunan Provincial Health Commission (No. 202204053869); National Key Clinical Specialty Construction Project - Pediatric Surgery of Hunan Children's Hospital (XWYF [2022] No. 2)

Availability of data and materials: The datasets generated and/or analysed during the current study are available in the TARGET database repository, (https://ocg.cancer.gov/)

Authors' contributions

(1) Tian-Qu He, Jun He, conceiving and designing the study;

(2) Yu Liu, Feng Ning, collecting the data;

(3) Tian-Qu He, Lei Tu, analyzing and interpreting the data;

(4) Tian-Qu He, writing the manuscript;

(5) Yao-Wang Zhao, Jun He, providing critical revisions that are important for the intellectual content;

(6) Tian-Qu He, Yao-Wang Zhao, Yu Liu, Lei Tu, Feng Ning, Jun He, approving the final version of the manuscript.

Conflict of interest statement

The authors have no ethical, legal and financial conflicts related to the article.

References

1. Cone EB, Dalton SS, Van Noord M, Tracy ET, Rice HE, Routh JC. Biomarkers for Wilms Tumor: A Systematic Review. The Journal of urology. 2016;196(5):1530-1535. 10.1016/j.juro.2016.05.100.
2. Mahamdallie S, Yost S, Poyastro-Pearson E, Holt E, Zachariou A, Seal S, Elliott A, Clarke M, Warren-Perry M, Hanks S, Anderson J, Bomken S, Cole T, Farah R, Furtwaengler R, Glaser A, Grundy R, Hayden J, Lowis S, Millot F, Nicholson J, Ronghe M, Skeen J, Williams D, Yeomanson D, Ruark E, Rahman N. Identification of new Wilms tumour predisposition genes: an exome sequencing study. The Lancet Child & adolescent health. 2019;3(5):322-331. 10.1016/s2352-4642(19)30018-5.
3. Gadd S, Huff V, Huang CC, Ruteshouser EC, Dome JS, Grundy PE, Breslow N, Jennings L, Green DM, Beckwith JB, Perlman EJ. Clinically relevant subsets identified by gene expression patterns support a

revised ontogenic model of Wilms tumor: a Children's Oncology Group Study. Neoplasia (New York, NY). 2012;14(8):742-756. 10.1593/neo.12714.

4. Gadd S, Huff V, Walz AL, Ooms A, Armstrong AE, Gerhard DS, Smith MA, Auvil JMG, Meerzaman D, Chen QR, Hsu CH, Yan C, Nguyen C, Hu Y, Hermida LC, Davidsen T, Gesuwan P, Ma Y, Zong Z, Mungall AJ, Moore RA, Marra MA, Dome JS, Mullighan CG, Ma J, Wheeler DA, Hampton OA, Ross N, Gastier-Foster JM, Arold ST, Perlman EJ. A Children's Oncology Group and TARGET initiative exploring the genetic landscape of Wilms tumor. Nature genetics. 2017;49(10):1487-1494. 10.1038/ng.3940.
5. Phelps HM, Pierce JM, Murphy AJ, Correa H, Qian J, Massion PP, Lovvorn HN, 3rd. FXR1 expression domain in Wilms tumor. Journal of pediatric surgery. 2019;54(6):1198-1205. 10.1016/j.jpedsurg.2019.02.030.
6. Zhang L, Gao X, Zhou X, Qin Z, Wang Y, Li R, Tang M, Wang W, Zhang W. Identification of key genes and microRNAs involved in kidney Wilms tumor by integrated bioinformatics analysis. Experimental and therapeutic medicine. 2019;18(4):2554-2564. 10.3892/etm.2019.7870.
7. Lin XD, Wu YP, Chen SH, Sun XL, Ke ZB, Chen DN, Li XD, Lin YZ, Wei Y, Zheng QS, Xu N, Xue XY. Identification of a five-mRNA signature as a novel potential prognostic biomarker in pediatric Wilms tumor. Molecular genetics & genomic medicine. 2020;8(1):e1032. 10.1002/mgg3.1032.
8. Zheng H, Li BH, Liu C, Jia L, Liu FT. Comprehensive Analysis of lncRNA-Mediated ceRNA Crosstalk and Identification of Prognostic Biomarkers in Wilms' Tumor. BioMed research international. 2020;2020(4951692. 10.1155/2020/4951692.
9. Young MD, Mitchell TJ, Vieira Braga FA, Tran MGB, Stewart BJ, Ferdinand JR, Collord G, Botting RA, Popescu DM, Loudon KW, Vento-Tormo R, Stephenson E, Cagan A, Farndon SJ, Del Castillo Velasco-Herrera M, Guzzo C, Richoz N, Mamanova L, Aho T, Armitage JN, Riddick ACP, Mushtaq I, Farrell S, Rampling D, Nicholson J, Filby A, Burge J, Lisgo S, Maxwell PH, Lindsay S, Warren AY, Stewart GD, Sebire N, Coleman N, Haniffa M, Teichmann SA, Clatworthy M, Behjati S. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science (New York, NY). 2018;361(6402):594-599. 10.1126/science.aat1699.
10. Wang Z, Guo X, Gao L, Wang Y, Ma W, Xing B. Glioblastoma cell differentiation trajectory predicts the immunotherapy response and overall survival of patients. Aging. 2020;12(18):18297-18321. 10.18632/aging.103695.
11. Martínez CH, Dave S, Izawa J. Wilms' tumor. Advances in experimental medicine and biology. 2010;685(196-209.
12. Charlton J, Pavasovic V, Pritchard-Jones K. Biomarkers to detect Wilms tumors in pediatric patients: where are we now? Future oncology (London, England). 2015;11(15):2221-2234. 10.2217/fon.15.136.
13. Stokes CL, Stokes WA, Kalapurakal JA, Paulino AC, Cost NG, Cost CR, Garrington TP, Greffe BS, Roach JP, Bruny JL, Liu AK. Timing of Radiation Therapy in Pediatric Wilms Tumor: A Report From the National Cancer Database. International journal of radiation oncology, biology, physics. 2018;101(2):453-461. 10.1016/j.ijrobp.2018.01.110.

14. He L, Wang X, Jin Y, Xu W, Lyu J, Guan Y, Wu J, Han S, Liu G. A Prognostic Nomogram for Predicting Overall Survival in Pediatric Wilms Tumor Based on an Autophagy-related Gene Signature. Combinatorial chemistry & high throughput screening. 2021;10.2174/1386207324666210826143727.
15. Pan Z, You H, Bu Q, Feng X, Zhao F, Li Y, Lyu J. Development and validation of a nomogram for predicting cancer-specific survival in patients with Wilms' tumor. Journal of Cancer. 2019;10(21):5299-5305. 10.7150/jca.32741.
16. Tang F, Zhang H, Lu Z, Wang J, He C, He Z. Prognostic Factors and Nomograms to Predict Overall and Cancer-Specific Survival for Children with Wilms' Tumor. Disease markers. 2019;2019(1092769. 10.1155/2019/1092769.
17. Ileri FC, Acun T. High expression of DNAJA1 (HDJ2) predicts unfavorable survival outcomes in breast cancer. Biomarkers in medicine. 2021;15(12):941-950. 10.2217/bmm-2020-0728.
18. Yang S, Ren X, Liang Y, Yan Y, Zhou Y, Hu J, Wang Z, Song F, Wang F, Liao W, Liao W, Ding Y, Liang L. KNK437 restricts the growth and metastasis of colorectal cancer via targeting DNAJA1/CDC45 axis. Oncogene. 2020;39(2):249-261. 10.1038/s41388-019-0978-0.
19. Kaida A, Yamamoto S, Parrales A, Young ED, Ranjan A, Alalem MA, Morita KI, Oikawa Y, Harada H, Ikeda T, Thomas SM, Diaz FJ, Iwakuma T. DNAJA1 promotes cancer metastasis through interaction with mutant p53. Oncogene. 2021;40(31):5013-5025. 10.1038/s41388-021-01921-3.
20. Zhou Y, Bian S, Zhou X, Cui Y, Wang W, Wen L, Guo L, Fu W, Tang F. Single-Cell Multiomics Sequencing Reveals Prevalent Genomic Alterations in Tumor Stromal Cells of Human Colorectal Cancer. Cancer cell. 2020;38(6):818-828.e815. 10.1016/j.ccell.2020.09.015.
21. Yu B, Chen X, Li J, Qu Y, Su L, Peng Y, Huang J, Yan J, Yu Y, Gu Q, Zhu Z, Liu B. Stromal fibroblasts in the microenvironment of gastric carcinomas promote tumor metastasis via upregulating TAGLN expression. BMC cell biology. 2013;14(17. 10.1186/1471-2121-14-17.
22. Liu Y, Wu J, Huang W, Weng S, Wang B, Chen Y, Wang H. Development and validation of a hypoxia-immune-based microenvironment gene signature for risk stratification in gastric cancer. Journal of translational medicine. 2020;18(1):201. 10.1186/s12967-020-02366-0.

Figures

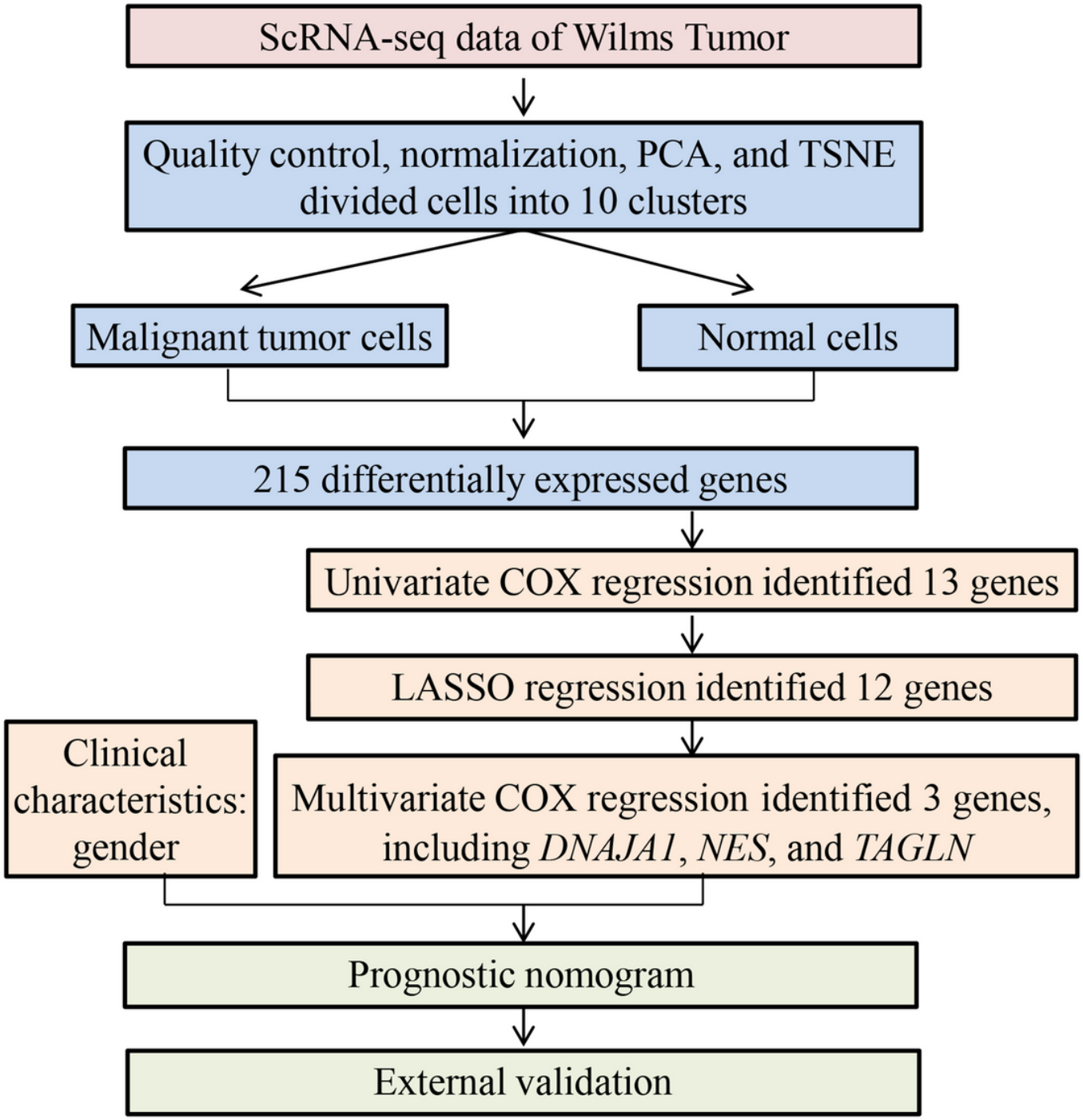

Figure 1

The flow chart of this study

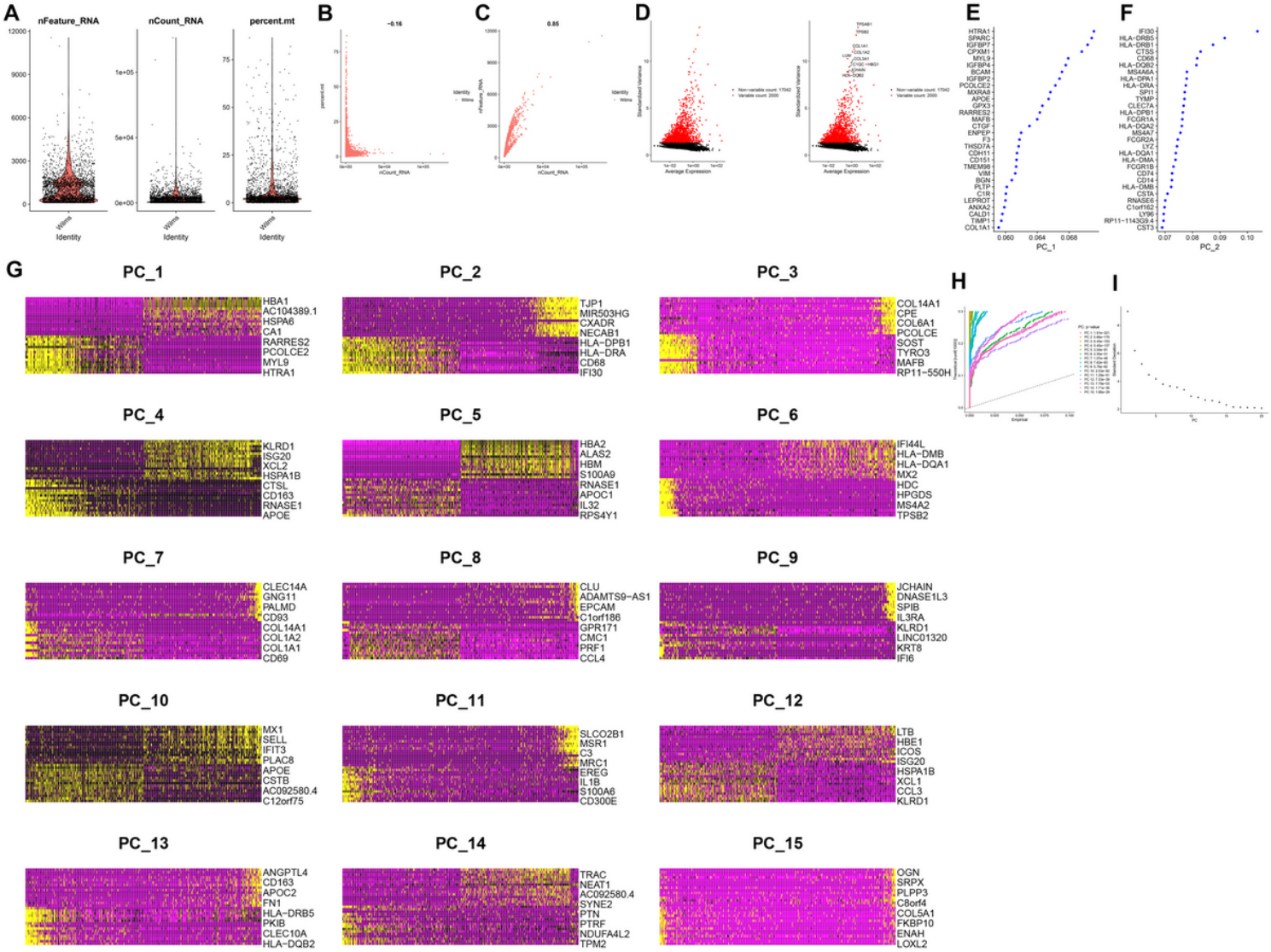

Figure 2

Characterization of single cell sequencing from Wilms tumor

(A) Quality control of single cell RNA sequencing for Wilms tumor. The Y axes represent RNA numbers, RNA counts and percentage of mitochondrial for each cell respectively. We filtered out the cells with poor quality and analyzed the detected gene counts and sequencing depth.

(B and C)The relationship between the percentage of mitochondrial genes and the mRNA reads, together with the relationship between the amount of mRNA and the reads of mRNA

(D) We calculate a subset of features that exhibit high cell-to-cell variation in the dataset. Red dots mean the 2000 variable genes. The top 10 gene names are labeled out.

(E and F) The top 30 significantly correlated genes in top 2 principal components.

(G) The heatmap of top 10 genes in top 15 principal components.

(H and I) The JackStrawPlot and ElbowPlot of principal components, which were used to identify the significantly available dimensions of data sets with estimated P value and Elbow.

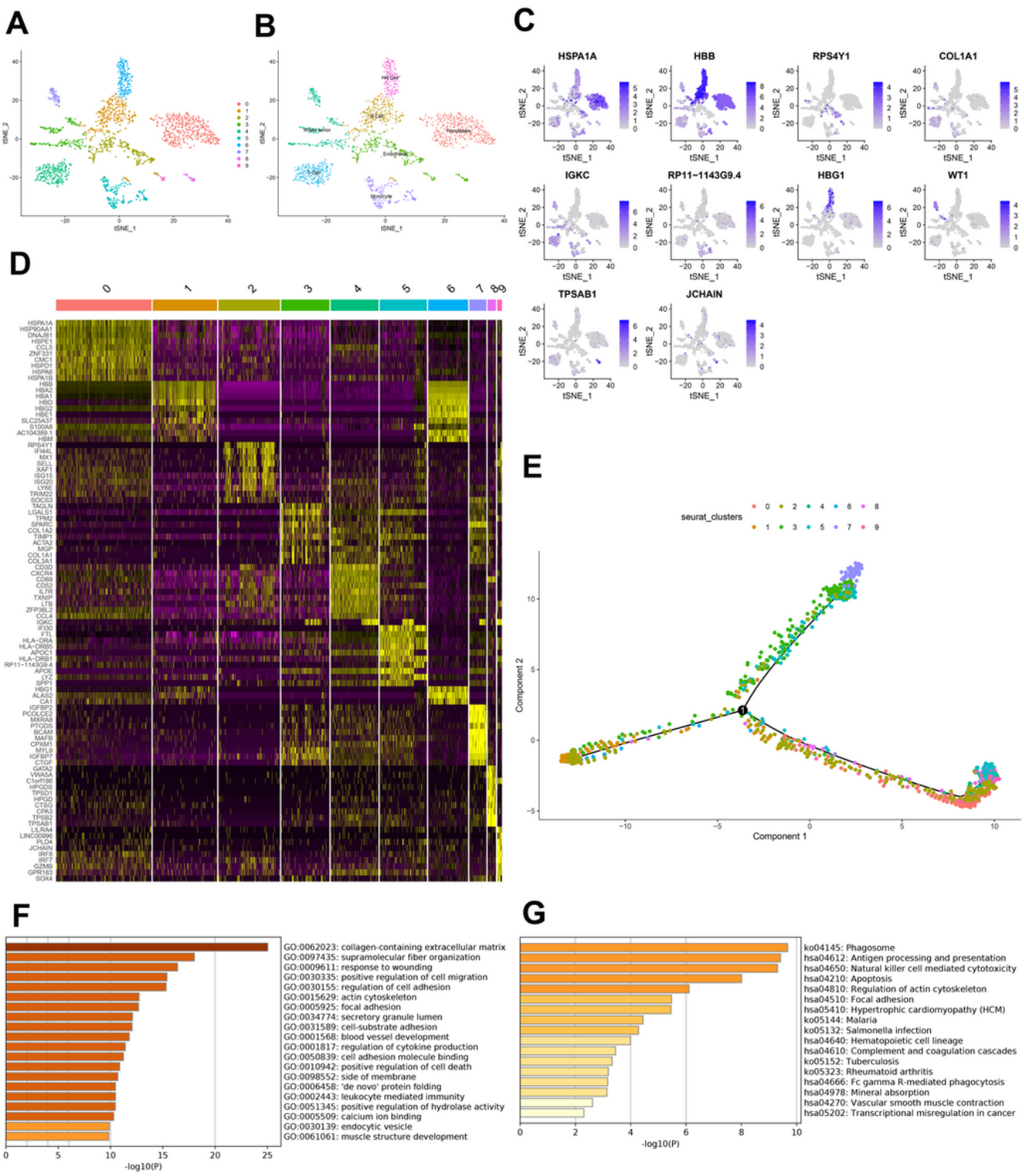

Figure 3

Clusters of single-cell RNA sequencing

(A) Based on available significant components, we conducted TSNE algorithm to learn the underlying manifold of the data in order to place similar cells together in low-dimensional space.

(B) Cells were annotated into 7 cell types: Wilms tumor, T cell, NK cell, B cell, Monocyte, Endothelial and Fibroblasts.

(C) Feather plot of most significant gene in each cluster

(D) Top 10 differentially expressed genes in each group

(E) Pseudo-time trajectory of each cluster using the Monocle algorithm

(F) GO enrichment analysis of DEGs

(G) KEGG enrichment analysis of DEGs

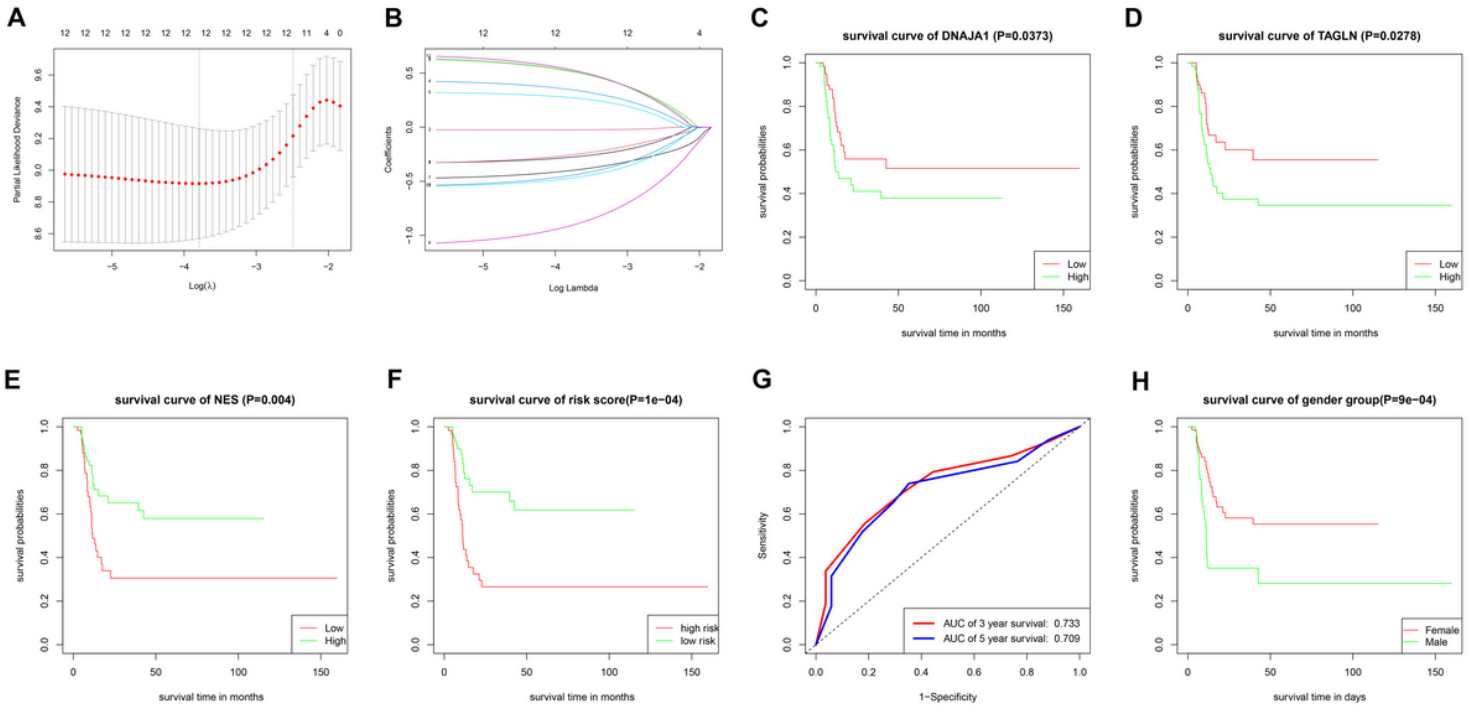

Figure 4

The prognosis-related genes were screened with univariate and multivariate Cox regression analysis, and LASSO regression analysis

(A) Optimal parameter (lambda) selection in the LASSO model used ten-fold cross-validation via minimum criteria.

(B) The coefficient profile plot was produced against the log (lambda) sequence.

(C-E) Kaplan-Meier plot of 3 prognosis-related genes: DNAJA1, TAGLN, and NES.

(F) Kaplan-Meier plot of risk score.

(G) ROC curve of risk score.

(F) Kaplan-Meier plot of gender.

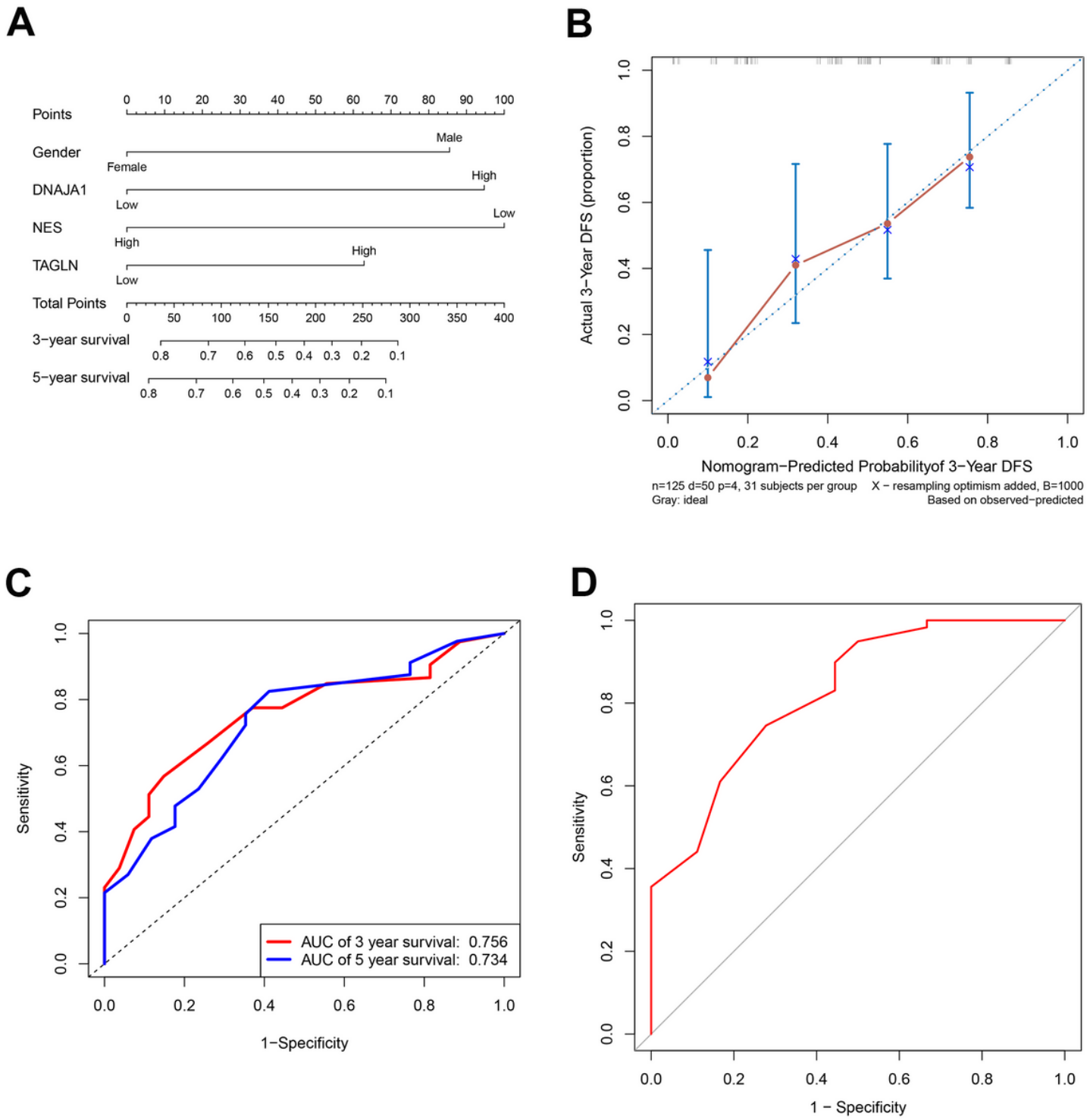

Figure 5

Construction and validation of the prognostic nomogram

(A) Nomogram model to predict the prognosis of Wilms tumor patients

(B) Calibration plots of the prognostic nomogram

(C) ROC curve of the prognostic nomogram model

(D) The validation of prognostic nomogram model though ROC curve in other dataset.