Iran J Public Health, Vol. 50, No.11, Nov 2021, pp.2238-2245

Original Article

Identification of Hub Genes in Pancreatic Ductal Adenocarcinoma Using Bioinformatics Analysis

Congcong Wang [1,2], Jianping Guo [2,3], Xiaoyang Zhao [4], Jia Jia [4], Wenting Xu [4], Peng Wan [5], *Changgang Sun [6,7]

1. *Clinical Medical College, Cheeloo College of Medicine, Shandong University, Jinan 250100, Shandong, China*
2. *Department of Oncology, Zibo Maternal and Children Hospital, Zibo 255000, Shandong, China*
3. *Shandong Qianfoshan Hospital, Cheeloo College of Medicine, Shandong University, Jinan 250014, Shandong, China*
4. *Department of Oncology Surgery, 4th People's Hospital of Zibo, Zibo 255000, Shandong, China*
5. *Department of Gastroenterology, Zibo Central Hospital, Zibo 255000, Shandong, China*
6. *Department of Oncology, Weifang Traditional Chinese Hospital, Weifang 261053, Shandong, China*
7. *Department of Oncology, Affiliated Hospital of Weifang Medical University, Weifang 261053, Shandong, China*

***Corresponding Author:** Email: sjxj7847@163.com

(Received 17 Jan 2021; accepted 14 Mar 2021)

Abstract

Background: To address the biomarkers that correlated with the prognosis of patients with PDCA using bioinformatics analysis.

Methods: The raw data of genes were obtained from the Gene Expression Omnibus. We screened differently expressed genes (DEGs) by Rstudio. Database for Annotation,Visualization and Intergrated Discovery was used to investigate their biological function by Gene Ontology(GO) and Kyoto Encyclopedia of Genes (KEGG) analysis. Protein-protein interaction of these DEGs were analyzed based on the Search Tool for the Retrieval of Interacting Genes database (STRING) and visualized by Cytoscape. Genes calculated by Cyto-Hubba with degree >10 were identified as hub genes. Then, the identified hub genes were verified by UALCAN online analysis tool to evaluate the prognostic value in PDCA.

Results: Three expression profiles (GSE15471, GSE16515 and GSE32676) were downloaded from GEO database. The three sets of DEGs exhibited an intersection consisting of 223 genes (214 upregulated DEGs and 9 downregulated DEGs). GO analysis showed that the 223 DEGs were significantly enriched in extracellular exosome, plasma membrane and extracellular space. ECM-receptor interaction, PI3K-Akt signaling pathway and Focal adhesion were the most significantly enriched pathway according to KEGG analysis. By combining the results of Cytohubba, 30 hub genes with a high degree of connectivity were picked out. Finally, we candidated 3 biomarkers by UALCAN online survival analysis, including CEP55, ANLN and PRC1.

Conclusion: we identified CEP55, ANLN and PRC1 may be the potential biomarkers and therapeutic targets of PDCA, which used for prognostic assessment and scheme selection.

Keywords: Bioinformatics analysis; Differently expressed genes; Hub genes; Pancreatic ductal adenocarcinoma

Introduction

Pancreatic ductal adenocarcinoma (PDCA), one of the most frequent digestive tumors in the world, is a devastating malignant disease with more aggressive in clinical behaviors. An estimated 53670 new cases of morbidity and 43090 related deaths occurred in the United States alone in 2017 (1). Nearly half of patients were asymptomatic until the disease develops to a distant stage, and most of those people lacks the optimal period for effective systemic therapy. Consequently, it was urgent and necessary for us to explore novel therapeutic targets for PDCA patients.

Plentiful clinical and experiment research of PDCA has finally lead to the identification of sensitive and effective biomarkers. These findings provided a good foundation to analyze key genes associated with PDCA that may act as diagnostic, prognostic or therapeutic targets. As a highly heterogeneous and comprehensive tumor, PDCA might result from different biological behaviors .Identification of different expression genes (DEGs) in PDCA varied from experimental conditions, individual difference, and any other aspects.Therefore,taking these aspects into conditions, only then we can screen additional co-expressed genes associated with PDCA.

Gene Expression Omnibus (GEO) database just provided the opportunity for the bioinformatics mining of gene expression profiles in various cancers (2). In this study, we extracted a set of DEGs from 3 gene expression datasets based on the same platform, which are potentially involved in tumorigenesis and progression. Hub genes with with close relationship to the PDCA pathogenic system were screened. Finally, we used UALCAN, an online tool based on The Cancer Genome Atlas (TCGA) datasets, to screen the expression level of the hub gene associated with tumor expression and OS.

Methods

Gene expression data

A total of three gene expression datasets were obtained from the Gene Expression Omnibus (GEO, *http://www.ncbbi.nlm.nih.gov/geo/*) online database (3). GSE15471, GSE16515 and GSE32676 downloaded from GEO were used to identify different expression genes (DEGs) between PDCA and normal tissues, with 100 primary tumor samples and 62 normal samples. All these data were based on the Agilent GPL570 platform (Affymetrix Human Genome U133 plus 2.0 Array; Agilent Technologies, Santa Clara, CA, USA). All of the raw data were freely available online, and these study did not involve any experiment on animals or cell lines performed by any of the authors.

The study was approved by the Ethics Committee of Zibo Maternal and Children Hospital.

Data preprocessing

The raw probe-level data were pre-processed by Affy package of Rstudio, an integrated development environment for R community, used for background correction and normalization of the data.

Identification of DEGs

The Limma package in Bioconductor was used to screen DEGs in PDCA tissues compared normal pancreatic tissues. DEGs were calculated using Limma and impute package of Rstudio.A threshold criteria of $|\log2FC| \geq 1$ and $P<0.05$, DEGs considered significant.Venny online tool (*http://bioinfogp.cnb.csic.es/tools/venny/index.html*), a scientific service of Spanish National Biotechnology Centre (CNB) was used to analyze the overlapping DEGs by veen diagram in three database.

Functional enrichment analysis of DEGs

To determine the functions of the overlapping DEGs, an enrichment analysis was performed on Gene ontology (GO) and Kyoto Encyclopedia of Gene and Genomes (KEGG). GO is a major bioinformatic tool for gene annotation that uses a highly structured vocabulary contains three main

categories: biological processes (BP), cellular components (CC) and molecular functions (MF).KEGG is a database aimed to associate related genes by pathway (4). The Database for Annotation, Visualization and Integrated Discovery (DAVID) (Version 6.8 , *http://davaid.ncifcrf.gov*) is a reliable program for a comprehensive set of functional annotation, enable investigators to understand the biological meaning behind large lists of genes or proteins (5). Go annotation and KEGG pathway enrichment analyses of DEGs was performed by DAVID online tools. The cutoff criteria for pathway screening and significant functionality was set $P<0.05$ as thresholds.

PPI network construction

STRING (*http://string-db.org*;version10.5) was utilized for functional interaction analysis to construct a protein-protein interaction (PPI) network (6). Confidence scores >0.7 were considered statistically significant. Genes calculated by CytoHubba(a plugin in Cytoscape) with a high degree were selected as hub genes. The network of PPI was visualized by Cytoscape software (v3.6.1).

Survival analysis of hub genes

UALCAN, an online tool utilized for facilitating tumor hub gene expression and survival analyses (*http://ualcan.path.uab.edu/*).To estimate the effects of hub genes expression levels based on clinic pathological data in the Cancer Genome Atlas (TCGA) pancreatic ductal adenocarcinoma datasets. Survival analysis was performed by Kaplan-Meier method, and the log-rank test was carried out. $P<0.05$ was selected as cutoff value.

Results

Identification of DEGs

A total of 3867 DEGs were detected in the datasets of GSE15471, GSE16515 and GSE32676, after pre-recession of raw data. Only 223 genes were common to all PDCA samples analyzed; 847 genes were common between 2 sets of DEGs; and 1504 genes were unique (Fig. 1). Among DEGs combined 3 sets, a total of 2573 genes, of which 2205 were upregulated and 369 were downregulated in PDCA tissues compared with normal pancreatic tissues, suggests a high heterogeneity. The combined 3 sets of DEGs, which 214 were upregulated and 9 were downregulated, were regarded as PDCA-related DEGs for further analysis.

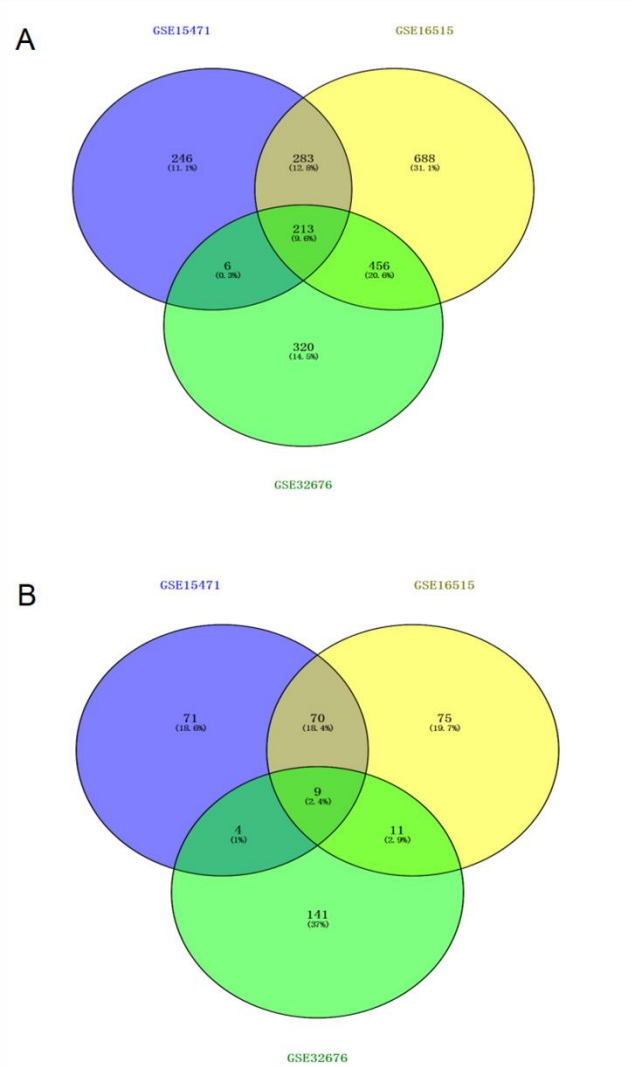

Fig. 1: Identification of overlapping DEGs. (A) Veen diagram of 2212 over-lapping upregulated genes in GSE15471, GSE16515 and GSE32676; (B) Veen diagram of 381 over-lapping upregulated genes in same datasets

Functional enrichment analysis of DEGs

After performing Go analysis of overlapping 223 DEGs with DAVID online, the DEGs were classified into three groups: Cellular component, molecular function and biological process groups. As shown in cellular component group, the common DEGs are significantly enriched in the extracellular matrix organization, collagen catabolic process and cell migration (Fig. 2A). In terms of molecular function, the enriched Go terms were mainly in calcium ion binding, protease binding, integrin binding and cysteine-type endopeptidase inhibitor activity (Fig. 2B). In addition, biological process analysis also revealed that the DEGs were significantly enriched in extracellular exosome, extracellular space, extracellular matrix and proteinaceous extracellular matrix (Fig. 2C).

KEGG pathway enrichment analysis of common 223 DEGs was also conducted by DAVID online. KEGG analysis of the DEGs were displayed in ECM-receptor interaction, amoebiasis, protein digestion and absorption and focal adhesion (Fig. 2 D).

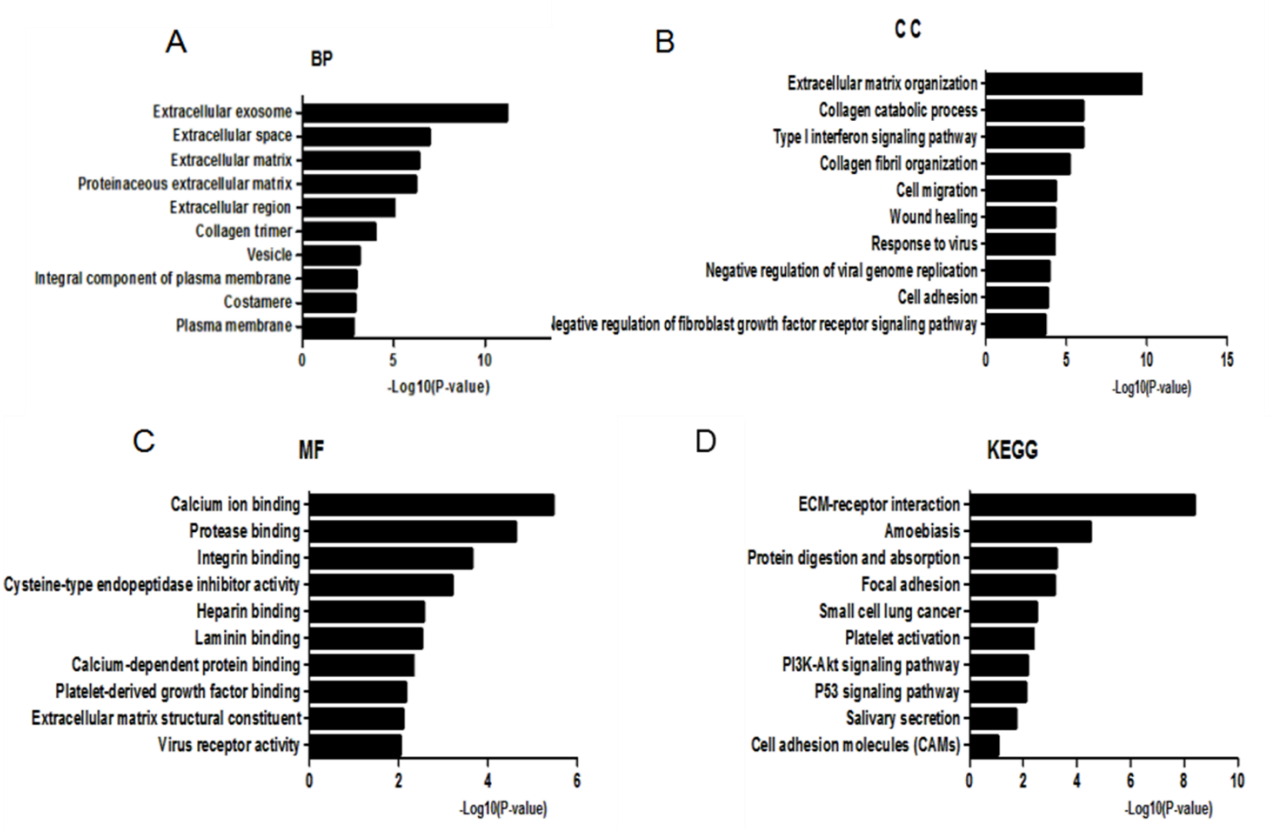

Fig. 2: GO and KEGG analysis of the overlapped DEGs.Black bars represent the number of DEGs.Here only show the top 10:(A)biological processes(BP);(B) cellular components(CC);(C) molecular functions(MF);(D)Kyoto Encyclopaedia of Gene and Genomes(KEGG)

PPI network construction

Based on the information of the overlapping 223 DEGs obtained from STRING online database, we constructed a PPI network diagram by Cytoscape software and calculated the degree of each gene by CytoHubba (Fig. 3).

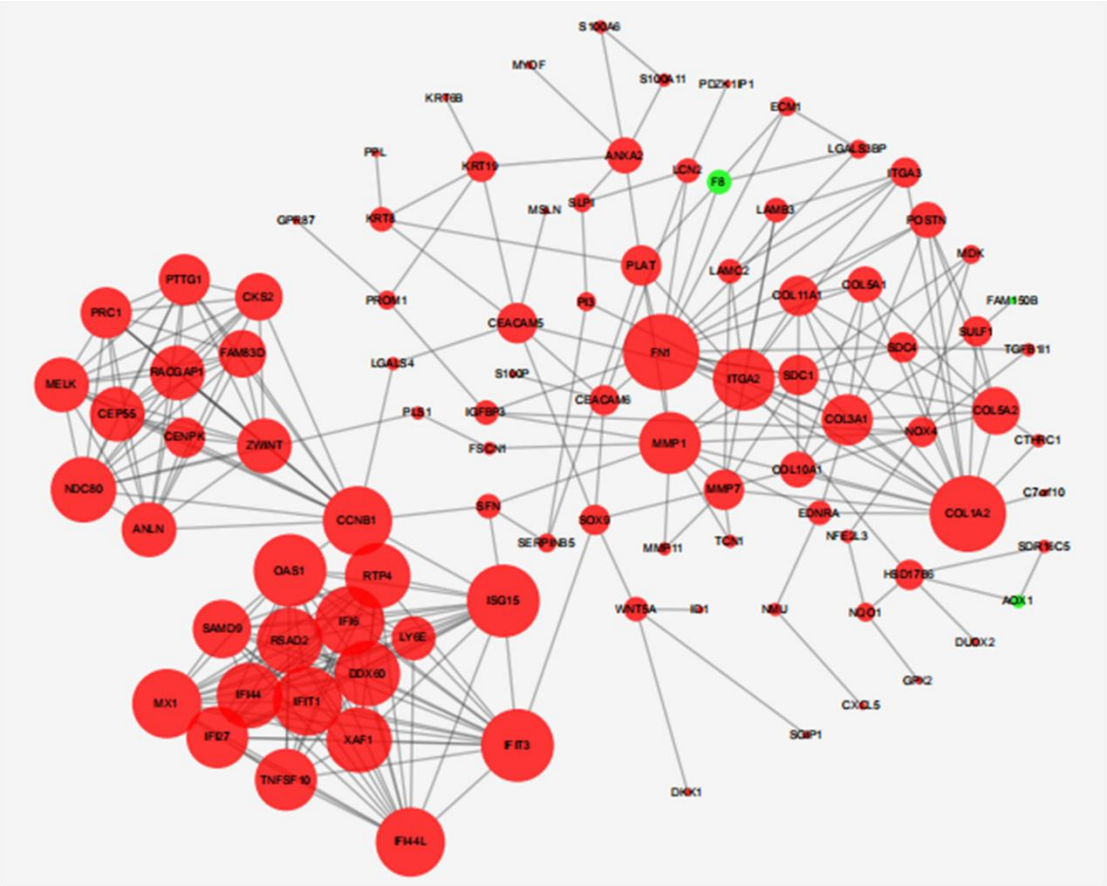

Fig. 3: Protein-Protein interaction network of DEGs constructed using weighted gene-co-expression network analysis in pancreatic ductal adenocarcinoma, and visualized using Cytoscape software. Red represents an expression level above the mean; Green represents an expression level below the mean

Survival analysis of hub genes

Among the 30 hub genes, all genes were identified from TCGA PAAD (Pancreatic adenocarcinoma) database on UALCAN.5 hub genes (*IFI6, MMP1, CEP55, ANLN* and *PRC1*) showed significantly higher expressions levels in primary tumor tissues than normal tissues. Furthermore, lower expression of 15 hub genes (*OAS1, IFIT3, CCNB1, IFIT1, IFI44L, DDX60, NDC80, TNFSF10, SAMD9, IWINT, RACGAP1, CEP55, NLN, MELK, PRC1* and *PTTG1*) were associated with longer OS of PAAD patients by survival analysis. Based on UALCAN analysis, 3 genes (*CEP55, ANLN* and *PRC1*) finally showed significantly correlated with OS in PDCA patients (Fig. 4).

Discussion

Pancreatic ductal adenocarcinoma (PDCA) is a heterogeneous disease, not a small proportion of PDCA patients are diagnosed of advanced stage, which lacks the effective and thoroughgoing measures before detection (7). Although numerous clinical and basic research had been conducted, no markedly improved had happened in overall incidence and survival rate over the past decades. At the same time, it was acknowledged that key genes associated as diagnostic, prognostic or therapeutic biomarkers would differ from experimental conditions and any other factors taken these aspects into consideration. Hence, it is necessary and crucial to find the reliable biomarkers for early detection.

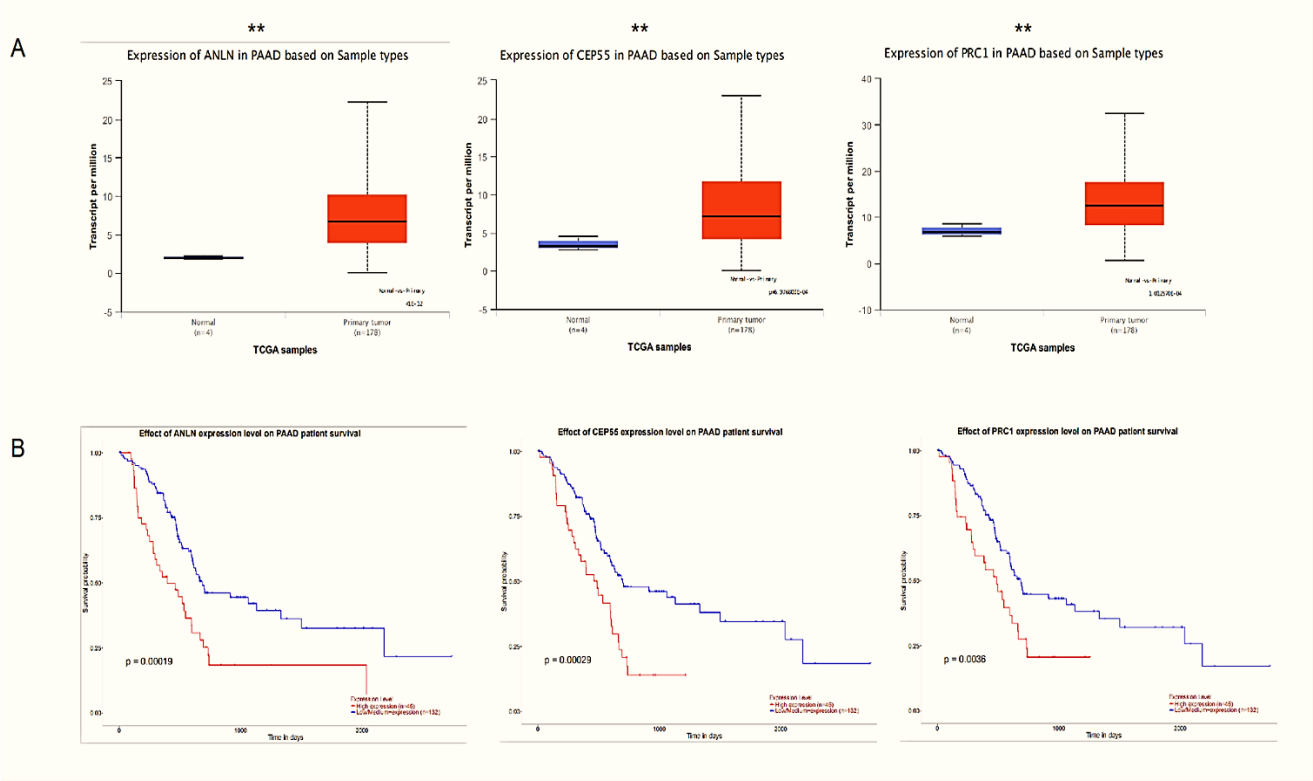

Fig. 4: Validation of the altered expression and Kaplan-Meier survival curves of CEP55, ANLN and PRC1. (A)Boxplots showing the expression of CEP55, ANLN and PRC1 in normal controls (n=4) and PDCA tissues (n=178) of TCGA samples (**means P<0.01). (B)Kaplan-Meier survival curves according to CEP55, ANLN and PRC1expression (**means P<0.05). PDCA, pancreatic ductal adenocarcinoma; TCGA, The Cancer Genome Atlas

In the present study, bioinformatics analysis have been aimed for finding new therapeutic and diagnosis markers for various tumors (2). However, compared with our study, their study only analyzed a profile, and only used the module method to select genes with a high degree of connectivity. In addition, their targeted genes were validated only via the Kaplan-Meier plotter database. Our study integrated three profiles datasets from the same platform by bioinformatics methods: 223 DEGs were screened, consisting 214 upregulated and 9 downregulated genes. Combined the results of gene expression and protein-protein expression analysis on publicly available databases for the identification of the potential genes correlated with PDCA.The results of functional enrichment analysis of GO-BP terms were closely related to extracellular exosome, extracellular space, extracellular matrix and proteinaceous extracellular matrix. Pathway enrichment analysis of the overlapping DEGs were enriched in ECM-receptor interaction, amoebiasis, protein digestion, absorption and focal adhesion. Furthermore, we verified the key genes by UALCAN, a reliable online tools, thus increasing the reliability of our results. We predicted 3 genes including CEP55, ANLN and PRC1 finally. All of these genes were upregulated in PDCA, which overexpression was related to unfavorable prognosis of patients.

Previous studies have reported some of these genes. Centrosomal protein 55(CEP55) was a microtubule-bundling protein that participants in cell mitosis, overexpressed in several solid tumors, which promotes the growth and invasion of cancer cells. CEP55 activated the activity of NF-κB signaling and promoted pancreatic cancer cells aggressiveness (8). In addition, increasing evidence showed that CEP55 has an oncogenic

role and its overexpression correlates markedly with tumor stage, aggressiveness, and poor prognosis across multiple tumor types, such as gastric carcinoma, breast cancer, and ovarian carcinoma (9-12). Moreover, in the group of patients with higher CEP55 expression levels, the poorer overall survival rate and median survival time were reported (13).

Except for CEP55, actin binding protein anillin (ANLN) is a conserved protein implicated in cytoskeletal dynamics, and it is a ubiquitously expressed protein required for cytokinesis (14). Overexpressed ANLN was reported in several cancers and elevated expression appears to be involved in the metastatic potential of human cancers (15-17). In non-small cell lung cancer (NSCLC), nuclear localization of ANLN was associated with poor survival of patients with NSCLC (18). Likewise, detection of nuclear ANLN was significantly associated with decreased breast cancer survival and recurrence-free survival (19). Present immunohistochemical assessment of ANLN protein expression showed that ANLN was localized in cell nuclei in PDAC cells (20). PRC1, also known as polycomb repressor complex 1, is directly involved in acinar gene regulation by inducing the progress of carcinogenesis in PDCA (21). PRC1, the identification of genetic mutations, it is of major importance to elucidate epigenetic alterations. This will increase the knowledge of pancreatic carcinogenesis and open new fields for therapeutic interventions.

According to our functional enrichment analysis results, CEP55, ANLN and PRC1 were involved in several pathways compactly related to PDCA pathogenesis such as plasma membrane, mitotic cytokinesis and cell-cell adherence junction. In addition, CEP55, ANLN and PRC1 were overexpressed in PDCA compared with normal pancreatic tissues, and overexpression of these genes was significantly correlated with unfavorable clinical prognosis in those patients. The results of our study were consistent with other study (8, 20, 21). However, the mechanism of these genes in PDCA is still not clear and further study is needed.

Our bioinformatics analysis identified 223 DEGs between PDCA and normal pancreatic tissues based on gene expression datasets obtained from the GEO database. Among them, hub genes might be the core genes of pancreatic cancer, including CEP55, ANLN and PRC1.All of them were upregulated in PDCA and associated with unfavorable clinical outcome in these patients, all of them are unfavorable prognostic factor. Further molecular biological study in vivo and in vitro are also needed to confirm the results in PDCA of our research.

Ethical considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (grant numbers: 81672974) and the Development Foundation of Zibo Maternal and Child Health Hospital.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Siegel RL, Miller KD, Jemal A (2017). Cancer statistics, 2017. *CA Cancer J Clin*, 67(1): 7-30.
2. Jiang P, Liu XS (2015). Big data mining yields novel insights on cancer. *Nat Genet*, 47(2): 103-104.
3. Barrett T, Suzek TO, Troup DB, et al (2005). NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res*, 33(Database issue): D562–D566.
4. Dennis GJ, Sherman BT, Hosack DA, et al (2003). DAVID: Datebase for Annotation,

Visualization, and Integrated Discovery. *Genome Biol*, 4(5): P3.
5. Szklarczyk D, Morris JH, Cook H, et al (2017). The STRING database in 2017: Quality-cantrolled protein-protein association networks, made broadly accessible. *Nucleic Acids Res*, 45(D1): D362-D368.
6. Kanehisa M, Goto S, Sato Y, et al (2012). KEGG for integration and interpretation of large scale molecular data sets. *Nucleic Acids Res*, 40(Database issue): D109–D114.
7. Tang Y, Zhang Z, Tang Y, et al (2018). Identification of potential target genes in pancreatic ductal adenocarcinoma by bioinformatics analysis. *Oncol Lett*, 16(2): 2453-2461.
8. Peng T, Zhou W, Guo F, et al (2017). Centrosomal protein 55 activates NF-*κ*B signalling and promotes pancreatic cancer cells aggressiveness. *Sci Rep*, 7(1): 5925.
9. Tao J, Zhi X, Tian Y, et al (2014). CEP55 contributes to human gastric carcinoma by regulating cell proliferation. *Tumour Biol*, 35(5): 4389–4399.
10. Wang Y, Jin T, Dai X, et al (2016). Lentivirus-mediated knockdown of CEP55 suppresses cell proliferation of breast cancer cells. *Biosci Trends*, 10(1): 67–73.
11. Zhang W, Niu C, He W, et al (2016). Upregulation of centrosomal protein 55 is associated with unfavorable prognosis and tumor invasion in epithelial ovarian carcinoma. *Tumour Biol*, 37(5): 6239–6254.
12. Jeffery J, Sinha D, Srihari S, et al (2016). Beyond cytokinesis: the emerging roles of CEP55 in tumorigenesis. *Oncogene*, 35(6): 683–690.
13. Fabbro M, Zhou BB, Takahashi M, et al (2005). Cdk1/Erk2- and Plk1-dependent phosphorylation of a centrosome protein, Cep55, is required for its recruitment to midbody and cytokinesis. *Dev Cell*, 9(4): 477–488.
14. Oegema K, Savoian MS, Mitchison TJ, et al (2000). Functional analysis of a human homologue of the Drosophila actin binding protein anillin suggests a role in cytokinesis. *J Cell Biol*, 150(3): 539-552.
15. Hall PA, Todd CB, Hyland PL, et al (2005). The septin-binding protein anillin is overexpressed in diverse human tumors. *Clin Cancer Res*, 11(19 Pt 1): 6780-6786.
16. Liang PI, Chen WT, Li CF, et al (2015). Subcellular localisation of anillin is associated with different survival outcomes in upper urinary tract urothelial carcinoma. *J Clin Pathol*, 68(12): 1026-1032.
17. Wang S, Mo Y, Midorikawa K, et al (2015). The potent tumor suppressor miR-497 inhibits cancer phenotypes in nasopharyngeal carcinoma by targeting ANLN and HSPA4L. *Oncotarget*, 6(34): 35893–35907.
18. Suzuki C, Daigo Y, Ishikawa N, et al (2005). ANLN plays a critical role in human lung carcinogenesis through the activation of RHOA and by involvement in the phosphoinositide 3-kinase/AKT pathway. *Cancer Res*, 65(24): 11314-11325.
19. Magnusson K, Gremel G, Ryden L, et al (2016). ANLN is a prognostic biomarker independent of Ki-67 and essential for cell cycle progression in primary breast cancer. *BMC Cancer*, 16(1): 904.
20. Idichi T, Seki N, Kurahara H, et al (2017). Regulation of actin-binding protein ANLN by antitumor miR-217 inhibits cancer cell aggressiveness in pancreatic ductal adenocarcinoma. *Oncotarget*, 8(32): 53180-53193.
21. Benitz S, Regel I, Reinhard T, et al (2016). Polycomb repressor complex 1 promotes gene silencing through H2AK119 mono-ubiquitination in acinar-to-ductal metaplasia and pancreatic cancer cells. *Oncotarget*, 7(10): 11424-11433.