Identification and characterization of differentially expressed genes in cervical cancer: insights from transcriptomic analysis

Haojue Wang1,2*, Xianyi Lu2, Ling Fang3, Qifeng Shi4, XiaoJun Yang1*

1The First Affiliated Hospital of Soochow University, Suzhou, Jiangsu, 215006, China
2Department of Obstetrics and Gynecology, Wuxi Xishan People's Hospital, (Wuxi Branch of Zhongda Hospital Southeast University), 214015, Wuxi, Jiangsu, China
3Department of Dermatology, Wuxi Xishan People's Hospital, (Wuxi Branch of Zhongda Hospital Southeast University), Wuxi, Jiangsu, 214105, China
4Department of Pathology, Wuxi Xishan People's Hospital, (Wuxi Branch of Zhongda Hospital Southeast University), Wuxi, Jiangsu, 214105, China

*These authors contributed equally to this work.

ABSTRACT
Cervical cancer is a significant global health burden, necessitating a comprehensive understanding of its underlying molecular mechanisms to improve diagnostic and therapeutic strategies. In this study, we conducted an in-depth bioinformatics analysis of cervical cancer using a high-throughput microarray dataset, GSE9750. Through robust screening and selection, we identified 1633 differentially expressed genes (DEGs) associated with cervical cancer. Enrichment analysis revealed crucial pathways and processes, such as DNA replication, cell cycle, and epithelial cell differentiation, implicated in cancer development. Additionally, we discovered key genes, including NEK2, AURKA, FOXM1, CDCA8, and CDC25A, linked to these pathways, which also showed significant differences in expression levels between various clinical characteristics. Our findings shed light on potential molecular targets for therapeutic interventions and contribute to the growing body of knowledge in cervical cancer research. This integrative bioinformatics approach serves as a valuable resource for future studies aiming to unravel the intricate molecular landscape of cervical cancer.

DOI: http://dx.doi.org/10.14715/cmb/2023.69.10.40

Copyright: © 2023 by the C.M.B. Association. All rights reserved.

Introduction

The outlook for cervical cancer is influenced by various elements, such as the stage of the cancer at the time of diagnosis (1). Per the National Cancer Institute, the 5-year relative survival rate for early-stage cervical cancer is 91-92% (2). However, if diagnosed after spreading to nearby tissues, organs, or lymph nodes, the rate decreases to 57-59% (3). The FIGO stage (including the extent of disease spread and lymph node involvement) is considered the most important prognostic factor in cervical cancer survival (4). While early-stage cervical cancer treatment can achieve a clinical cure, advanced cervical cancer has a poor prognosis. Identifying Hub genes related to cervical cancer progression is valuable for early diagnosis, discovering new treatments, and improving patient survival and prognosis. There have been several recent studies that have used bioinformatics analysis and gene expression profiling to identify key genes and pathways involved in the development of cervical cancer (5-9). For instance, a BMC Cancer study utilized bioinformatics analysis to discover key genes and pathways involved in cervical cancer progression (10). The study identified 476 DEGs enriched in 22 biological processes, 16 cellular components, and 9 molecular functions. Another study found 7 modules with 76 hub genes, with enrichment analysis revealing increased genes related to cell cycle, DNA replication, and p53 signaling (11). The study. While several studies have shed light on potential molecular targets for therapeutic interventions and contribute to the growing body of knowledge in cervical cancer research, this integrative bioinformatics approach serves as a valuable resource for future studies aiming to unravel the intricate molecular landscape of cervical cancer.

Materials and Methods

Data Acquisition
Transcriptomic data and clinical information of cervical squamous cell carcinoma (CESC) patients were obtained from the Gene Expression Omnibus (GEO) database. The
dataset GSE9750, which includes data from 66 samples, including 33 primary tumors, 9 cell lines, and 24 normal cervical epithelium samples, was utilized for the analysis.

**Differential Gene Expression Screening**

GEO dataset GSE9750 was taken as the study subject. The criteria utilized for screening DEGs involved setting a log$_2$(FC) $> 1$ and adjusted p $< 0.05$. The results were illustrated through a Venn diagram to depict the overlap between different conditions, and gene expression patterns were visualized using the R software package "heatmap". Additionally, volcano plots representing the significance of differential expression were generated using the R package "volcano".

**Functional Enrichment Analysis of differentially expressed genes (DEGs)**

Firstly, the DEGs identified in the previous step were subjected to Gene Ontology (GO) analysis, which involved annotating the genes based on their biological processes, molecular functions, and cellular components. The GO analysis was carried out using online platforms Database for Annotation, Visualization, and Integrated Discovery (DAVID). Subsequently, pathway enrichment analysis was performed utilizing pathway databases Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome. The statistical significance of enrichment was assessed using hypergeometric tests to identify pathways that were overrepresented among the DEGs.

**Protein-Protein Interaction (PPI) Network Analysis**

DEGs were mapped onto existing PPI databases STRING to establish a comprehensive network of physical protein interactions. Subsequently, Cytoscape was employed to visualize and analyze the complex network. The generated PPI network allowed for the identification of densely connected regions and clusters of interacting proteins. A heatmap was utilized to illustrate the expression patterns of hub genes in different samples or conditions. The molecular Complex Detection (MCODE) algorithm was used to identify densely connected subnetworks with distinct colors. Top hub genes were identified from ranking using the Diffusion State Distance Matrix (DMNC) method, which is based on their network centralities.

**Gene Set Enrichment Analysis (GSEA)**

The functional significance of DEGs was analyzed using GSEA (version 4.1.0), with pathway exploration conducted using the MSIGDB database (http://software.broadinstitute.org/gsea/msigdb) (12). For the enrichment analysis, the standard weighted enrichment method was utilized. A random combination was performed 1000 times, and an FDR of less than 0.25, an absolute NES value greater than 1, and an NOM p-value of less than 0.05 were deemed to signify significant enrichment.

**LASSO Analysis and Prognostic Gene Signature Development**

To develop a robust prognostic gene signature, LASSO (Least Absolute Shrinkage and Selection Operator) analysis was employed as a feature selection method. The "glmnet" package in R was employed to perform LASSO regression. To identify the most suitable lambda value, k-fold cross-validation was conducted using the cv.glmnet function. The "timeROC" package was used to evaluate the predictive accuracy of the signature by constructing receiver operating characteristic (ROC) curves. Risk Score $= \sum (n, i) X_i \times Y_i$ (X: coefficient, Y: expression level).

**The Prognostic Values of the Oxidative Stress-Related DEGs**

To identify genes that play a role in the prognosis of cervical cancer patients, we searched The Cancer Genome Atlas (https://portal.gdc.cancer.gov/) for relevant prognostic and clinicopathological data. Using the R packages survminer and survival, we screened for genes related to prognosis. Kaplan-Meier survival analysis was then employed to assess the impact of these genes on overall survival (OS) in cervical cancer patients. A p-value of less than 0.05 was used as the criterion for significance.

**Verification of Prognosis-Related DEGs Expression**

To verify the protein expression of DEGs in normal and tumor tissues, we analyzed data from TCGA cervical cancer dataset. This dataset provides comprehensive genomic and clinical data for a large cohort of cervical cancer patients, including gene expression profiles and corresponding clinical information. The gene expression data were generated using microarray technology, and the clinical information included T stage, M stage, histologic grade, histologic type, overall survival (OS) status, clinical stage, and menopause status. Student’s t-tests and ANOVA were utilized to compare the expression levels.

**Results**

**Transcriptomic Profiling and Sample Quality Assessment in Cervical Cancer**

A transcriptomic analysis was performed using the GSE9750 dataset to investigate differential gene expression in cervical cancer. A total of 12,549 genes from 66 samples were included in the analysis. The PCA plot displayed a clear separation of points between the reference group (blue) and the test group (red), indicating distinct differences based on the analyzed variables (Figure 1A). This observation confirmed reliable data quality and effective sample stratification. In volcano plot, applying a threshold of $|\log_2$ (Fold Change) $| > 1$ and p.adj $< 0.05$, a total of 1633 genes were filtered and considered for subsequent analysis (Figure. 1B). The expression patterns of the top 40 DEGs were examined using a heatmap (Figure. 1C). Furthermore, the box and whisker plot displayed a consistent and comparable distribution of normalized values between the sample groups, with overlapping boxes and similar median values indicating an effective normalization process that adjusted gene expression levels across samples, resulting in a reliable and standardized dataset for further analysis (Figure.1D).

**Detection and Enrichment Analysis of Prognosis-Related Genes**

The Volcano plot displayed the differential gene expression results for 33, 148 diagnosis-related genes (Figure. 2A), of which 3, 488 genes were highlighted and ranked based on their differential expressions (Figure. 2B). The Venn plot intersection revealed the overlap and uniqueness of gene sets between GSE9750 and prognosis-related genes, with 1, 290 genes unique to GSE9750, 1, 562 genes...
gesting their involvement in DNA synthesis and keratinization, respectively (Figure 3A-E). Additionally, "REACTOME DNA REPLICATION" showed moderate enrichment (NES=2.696). A ridge plot visualized the enrichment scores for various cell cycle-related terms, illustrating their relative significance (Figure 3F).

PPI Network Analysis and Hub Gene Identification in Cervical Cancer

An enrichment heatmap was generated showcasing the ranking of terms related to biological processes and pathways. Based on these findings, the majority of proteins were found to be involved in the following processes: mitotic cell cycle, cell cycle regulation, transitions in the cell cycle phases, and positive modulation of cell cycle progression (Figure 4A). PPI network comprised 193 nodes (proteins) and 2530 edges (protein interactions). The network was visualized with different colors, indicating distinct clusters (MCODE1 to MCODE11) within the network (Figure 4B). The PPI network was then enriched unique to prognosis, and 343 genes shared between both datasets (Figure 2C). The bar plot illustrated the enrichment analysis results, presenting the ranking of terms and pathways based on their significance represented by -log10 (P value). The top-ranked terms and pathways included mitotic nuclear division (BP), nuclear division (BP), organelle fission (BP), Cell cycle (KEGG), and chromosomal region (CC) (Figure 2D). The gene-concept network analysis highlighted the most prominent concepts based on gene counts. Among them, organelle fission demonstrates the highest gene count with 43 genes, followed closely by nuclear division (42 genes), mitotic nuclear division (35 genes), chromosomal region (29 genes), and spindle (28 genes) (Figure 2E). The dot plot presented the GeneRatio rankings for the top concepts. The concepts were listed in descending order as follows: organelle fission, nuclear division, Cell cycle, mitotic nuclear division, and chromosomal region (Figure 2F). By digging comprehensive insights into the analysis results, the Z-score ranking reveals the following top-ranked concepts, listed in descending order: nuclear division, chromosomal region, organelle fission, spindle, and mitotic nuclear division (Figure 2G-2I).

Cell Cycle-Related Pathways Enriched in GSEA Analysis

Enrichment analysis using GSEA revealed several cell cycle-related pathways with significant enrichment scores. The pathways "REACTOME_CELL_CYCLE" and "REACTOME_CELL_CYCLE_MITOTIC" exhibited high enrichment scores (NES=4.178 and NES=4.021, respectively), indicating their crucial involvement in cell cycle progression. "REACTOME_CELL_CYCLE_CHECKPOINTS" demonstrated a significant enrichment score (NES=3.199), emphasizing its role in regulating the progression of the cell cycle. The pathway "REACTOME_MITOTIC_G1 PHASE_AND_G1_S_TRANSITION" displayed moderate enrichment (NES=2.859), while "REACTOME_SYNTHESIS_OF_DNA" and "REACTOMEFORMATION_OF_THE_CORNIFIED_ENVELOPE" exhibited distinct enrichment scores, suggesting their involvement in DNA synthesis and keratinization, respectively (Figure 3A-E). Additionally, "REACTOME DNA REPLICATION" showed moderate enrichment (NES=2.696). A ridge plot visualized the enrichment scores for various cell cycle-related terms, illustrating their relative significance (Figure 3F).
with MCODE analysis, with each node representing a protein, and the different colors representing distinct clusters or subnetworks of proteins based on their functional associations (Figure 4C). Among them, specific GO terms were color-coded to represent distinct biological processes, such as mitotic cell cycle, cell cycle phase transition, and positive regulation of cell cycle process (Figure. 4D). Hub genes were identified within PPI network using DMNC method and Cytoscape software. The top 10 hub genes in the network, namely HJURP, CDC25C, NCAPH, OIP5, CDC25A, DTL, DLGAP5, CENPF, ASPM, and CDCA3, were identified as key regulators in cervical cancer (Figure. 4E).

**Differential Expression Patterns of DEGs across Various Clinical Parameters and Outcomes in Cervical Cancer Patients.**

Notably, distinct expression patterns were observed for C19orf53 and DMPK concerning the T stage (Figure. 6A), with C19orf53 showing higher expression in T1 compared to T2, T3, and T4, while DMPK exhibited the opposite trend. Regarding the M stage (Figure. 6B), C19orf53 and ILVBL displayed higher expression levels in M1 compared to M0 and MX. Histologic grade analysis (Figure. 6C) indicated lower expression of SYDE1, TLE6, and C19orf53 in G1 and G2 grades compared to G3 and G4.
grades. Moreover, histologic type analysis (Figure. 6D) revealed differences in the expression of SYDE1, TLE6, RASAL3, and DMPK between patients who were alive and those who had passed away (Figure. 6E). Additionally, the clinical stage (Figure. 6F) showed higher expression in Stage I compared to Stage II, Stage III, and Stage IV. Lastly, menopause status analysis (Figure. 6G) indicated differential expression in postmenopausal and perimenopausal patients for ILVBL, PPP6R1, and TLE6. These findings underscored the potential prognostic and diagnostic significance of the identified DEGs in cervical cancer.

Discussion

Cervical cancer, a common type of cancer among women, is characterized by complex interactions between various factors and genetic variations that contribute to its multi-stage development (13). Gaining insight into the molecular mechanisms behind cervical cancer is essential for improving treatment and early detection. Using microarray technology, a high-throughput method (14), we examined gene expression profiles in cervical cancer using the original dataset GSE9750, identifying 12, 549 differentially expressed genes (DEGs) between cancerous cervical tissue and normal samples. Among these, 1633 DEGs with fold changes exceeding 1 were selected for further comprehensive bioinformatics analysis, providing valuable insights into potential molecular targets for future research and therapeutic interventions.

The enrichment analysis highlighted significant terms and pathways, including mitotic nuclear division, nuclear division, organelle fission, Cell cycle, and chromosomal region, with key genes such as EDN3, NEK2, AURKA, and CDCA8. These findings align with existing research, where AURKA's highest expression during G2/M phase of the cell cycle and its role in regulating Polo Like Kinase 1 (PLK1) have been well-documented in cervical cancer (15). Moreover, CDCA8's significant negative correlation with 5-year overall survival further supports its relevance as a prognostic marker (16, 17).

GSEA analysis yielded significant enrichment of pathways such as cell cycle, DNA replication, and epithelial cell differentiation, which are pivotal in cancer development and underscore their potential as therapeutic targets (18). Notably, NEK2, AURKA, FOXM1, CDCA8, and CDC25A were identified as key genes associated with these pathways, aligning with previous research findings. For instance, NEK2, a serine/threonine kinase involved in mitosis, has been linked to chromosome instability, tumor progression, and metastasis in cervical cancer (19). Additionally, the downregulation of CDC25A in ME180 and C33A cells has been shown to impede cell proliferation, arrest cell cycle progression, and induce cell apoptosis (20).

The construction of a PPI network gained insights into the functional interactions among the identified DEGs. Through this network analysis, we identified critical subnetworks and hub genes that appear to play central roles in the network. These hub genes, such as HJURP, CDC25C, NCAPH, OIP5, CDC25A, DTL, DLGAP5, CENPF, ASPM and CDCA3, have been associated with cell division, DNA replication, and cell cycle regulation, indicating their importance in cervical carcinogenesis. Despite what is mentioned above (20), it has been evident that changes in the expression of CDC25C are closely related to tumorigenesis and tumor development and can be used as a potential target for cancer treatment (21). All these findings solidify our results.

LASSO regression analysis was conducted to identify a prognostic gene signature based on the DEGs. This gene signature, including RETN, DMPK, TLE6, and others, was able to accurately predict the overall survival of cervical cancer patients. Additionally, the gene signature showed significant differences in expression levels between various clinical characteristics strictly related to prognosis, such as T stage, M stage, histological grade, histological type, overall survival, clinical stage, and menopause status (1, 22, 23), further confirming its potential clinical relevance. These findings deepen our understanding of the molecular heterogeneity in cervical cancer and may help tailor personalized treatment strategies.

Overall, our integrative bioinformatics analysis has provided valuable insights into the molecular landscape of cervical cancer. The identified DEGs, enriched pathways, hub genes, and prognostic gene signature hold great promise for further research and potential clinical applications. However, further experimental validation and functional studies are necessary to confirm the biological significance of these findings and translate them into clinical practice. Our study lays the groundwork for future investigations and highlights the importance of bioinformatics approaches in deciphering complex diseases like cervical cancer.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

XJ Y designed the study, HJ W, and DJ X analyzed data, and wrote the manuscript. XY L, L F, CJ C, and QF S contributed to the manuscript modification. All authors read and approved the final submitted manuscript.

Funding

This work is supported by the scientific research project plan of the Wuxi Municipal Health Commission (NO. M202007); Wuxi Science and Technology Plan Project (NO.N2020X018); Top Talent Support Program for young and middle-aged people of Wuxi Health Committee(NO. bj2020111); National Nature Science Foundation of China(NO.81971335).

Data Availability Statement

The data in this article can be obtained from the corresponding author under reasonable circumstances.

References

3. Bourgioti CK. Chatoupis, and L.A. Moulopoulos, Current