Identification of hub genes in cervical cancer using weighted gene co-expression network analysis

Xiao Feng Lv (✉ 2456010821@qq.com)  
Third Affiliated Hospital of Zhengzhou University  https://orcid.org/0000-0002-2766-7855

Ruyue Gong  
The Third Affiliated Hospital of Zhengzhou University

Xiao Han  
Zhengzhou University Third Hospital and Henan Province Women and Children's Hospital

Shihong Cui  
Zhengzhou University Third Hospital and Henan Province Women and Children's Hospital

Research article

Keywords: cervical cancer, hub genes, weighted gene co-expression network analysis, bioinformatics, translational medicine

DOI: https://doi.org/10.21203/rs.3.rs-41524/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background

Cervical cancer ranks second among malignancies in females around the world. Due to the elevated incidence and mortality of this malignancy, deciphering its pathogenesis and identifying related biomarkers is urgently required.

Methods

First, raw cervical squamous cell carcinoma (CESC) data in GSE63514 were downloaded from the Gene Expression Omnibus (GEO) database. Then, weighted Correlation Network Analysis (WGCNA) was performed to build a co-expression network. Next, comprehensive bioinformatics was performed to determine hub genes, and assess the associated functional annotation, prognostic signature, tumor immunity, DNA mismatch repair, methylation mechanism, candidate molecular drugs, and gene mutations.

Results

From the key module, ALOX12B, KRT78, RHOD and ZNF750 were selected for validation. K-M plots indicated that these genes had good diagnostic and prognostic values in CESC. Moreover, mutations in these hub genes resulted in the downregulation of most immune genes in CESC. On the other hand, most of the four core genes were negatively correlated with DNA mismatch genes. In addition, we found that RHOD and ZNF750 had decreased methylation in the disease state, while ALOX12B and KRT78 showed no significant differences. Meanwhile, GSVA revealed that most core genes had associations with P53 signaling and the hypoxia signaling pathway.

Conclusion

WGCNA could identify groups of genes significantly associated with cervical cancer prognosis. These findings provide new insights into CESC pathogenesis, and identify ALOX12B, KRT78, RHOD and ZNF750 as candidate biomarkers for CESC diagnosis and prognosis.

Background

Cervical cancer represents a major disease affecting women more than any other gynecological tumor, with about 560,847 new cases and 311,365 deaths worldwide each year [1]. It also constitutes an important cause of cancer-related mortality in women of low-and middle-income nations [2, 3]. The stratified squamous epithelium and columnar epithelium's transitional area of the cervix are the most common sites of cervical cancer; cervical squamous cell carcinoma (CESC) easily occurs in this area,
comprising > 80% of all cervical cancer diagnoses [4]. In contrast, cervical tumors are predominantly adenocarcinomas. Other cervical tumors, such as adenosquamous, small cell, neuroendocrine, serous papillary, and clear cell carcinomas are less commonly detected histological subgroups. Cervical cancer is rarely detected at an early stage due to limited diagnostic technologies, especially in poor nations [5, 6]. Therefore, screening for biomarkers that predict the early progression of cervical cancer is critical for disease management.

The TCGA database empowers the application of the genome analysis technology in cancer genome changes, as well as large-scale genome sequencing. As the largest cancer gene information database, TCGA comprehensively comprises many types of cancer and reflects multiple omics data, including mRNA and miRNA data, copy number variations, DNA methylation, and SNPs [7]. In view of the comprehensiveness of TCGA database information, it would be helpful for the early management of cervical cancer to describe a molecular map for early prevention by exploring the detailed molecular networks of TCGA genes associated with progression in this malignancy. In the early stage of this study, through the GEO database, WGCNA assessment was performed for building a co-expression network to identify genes related to cervical cancer development. Then, comprehensive bioinformatics was performed to further investigate the functions, pathways, regulatory mechanisms, and candidate drugs associated with the screened genes, for exploring the specific mechanisms of key genes in cervical cancer development, providing new insights into early cervical cancer detection.

**Methods**

1. **WGCNA analysis**

The Series Matrix File data GSE63514 was retrieved from the NCBI GEO public database. A total of 128 CESC transcriptional data were extracted, and used for building a WGCNA co-expression network, discussing the differences in molecular mechanism of CESC progression. By constructing a weighted gene co-expression network, co-expression gene modules were found, and the association of gene network with phenotype was examined, also exploring the core genes of the network. The WGCNA-R package was utilized to build the co-expression network of various genes in the GSE63514 data set, and the algorithm was employed to screen the first 5000 genes with variations for further analysis, in which the soft threshold was set to 10 [8]. This was followed by weighted adjacency matrix transformation into a topological overlap measure (TOM) matrix for estimating its connectivity feature in the network. Distinct cluster tree branches depict various gene modules, represented by distinct colors. According to their weighted correlation coefficients, genes were grouped by expression pattern into modules.

2. **TCGA data retrieval**

The TCGA database (https://portal.gdc.cancer.gov/) as the biggest cancer gene information database, includes mRNA and miRNA expression data, copy number variations, DNA methylation, SNPs and others. Raw SNP data in TCGA are not available to the public. We downloaded the processed raw mRNA
expression data of CESC. A total of 309 specimens were collected, including 3 normal and 306 cancer specimens.

3. Function annotation in gene ontology (GO) and Kyoto encyclopedia of genes and genomes (KEGG)

The R package "clusterProfiler" was used to annotate the functions of mutated genes to fully assess the functional relevance of the mutated genes [9]. The GO [10] and KEGG [11] databases were employed for assessing the associated functional categories. GO and KEGG enrichment pathways with p or q below 0.05 were deemed significant.

4. Construction of a hub gene-based prognostic signature

RNA sequencing (RNAseq) data obtained on the Illumina HiSeq RNA-Seq platform and associated follow-up patient data were obtained from TCGA (https://cancergenome.nih.gov/). The R software was used to draw K-M survival curves. Prognosis-associated genes were considered with P < 0.05.

5. Co-expression analysis

Co-expression of ALOX12B, KRT78, RHOD and ZNF750 was analyzed by the R software, in which the filter condition of the correlation coefficient was 0.3 with p < 0.001. Genes with the most significant expression were screened. Then, the "corrplot" and "circlize" packages were used to plot correlation analysis curves.

6. Drug sensitivity analysis

In this study, the sensitivities of key hub genes to the cancer therapeutics response portal (CTRP) were analyzed with GSCALite (http://bioinfo.life.hust.edu.cn/web/GSCALite/), which integrates cancer genomics data of 33 cancer types from TCGA, drug response information from the Genomics of Drug Sensitivity in Cancer (GDSC), CTRP, and noncancerous tissue information from the Genotype-Tissue Expression database (GTEx).

7. Relationships between pivotal genes and immune cells

The TIMER database (https://cistrome.shinyapps.io/timer/) detects immune cell infiltration in cancer tissues through RNA-seq expression profiling. In this study, the TIMER database was employed to assess the associated of hub genes with immune cell amounts and to compare the infiltration levels between tumors with different somatic copy number changes of hub genes.

8. Methylation of hub genes in CESC

DNA methylation represents a critical epigenetic modification for genomic modulation in higher organisms. The DiseaseMeth database is one of the most important worldwide methylation databases for human diseases, comprising over 14,000 entries from a wide range of sources and 175 high-volume datasets of experimental information. The latest edition combines genetically-centered methylation data from 72 human ailments from various technologies and platforms. The integrated gene methylation data of the diseased and normal samples of poultry are analyzed across datasets and could be used to study
the gene-disease relationships of differentially methylated genes. The methylation of key genes over CESC was studied by applying the poultry database. According to DNA mismatch repair related genes, correlation analysis of core genes was performed with the R software, and $P<0.05$ was considered significant.

9. Gene set variation analysis (GSVA)

GSVA constitutes a nonparametric and unsupervised technique for assessing gene accumulation in the transcriptome. By comprehensively scoring the gene set of interest, GSVA converts gene level changes into pathway level changes, and then assesses the biological function of the sample. In this study, gene sets were downloaded from the Molecular signatures database (v7.0 version), and the GSVA algorithm was used to comprehensively score each gene set to evaluate the potential changes in biological function of different samples.

10. Statistical analysis

R version 3.6 was employed for data analysis. Two-sided $p<0.05$ indicated statistical significance.

Results

1. Data Source and Processing

The workflow is shown in Fig. 1. Gene expression profiling was retrieved from GEO (https://www.ncbi.nlm.nih.gov/geo). WGCNA of CESC and noncancerous specimens was carried out. Co-expression networks for CESC and normal conditions were built, detecting gene modules to identify hub genes and compare expression patterns between CESC and noncancerous specimens. Next, comprehensive bioinformatics was performed for hub gene selection, assessing the functional annotation, prognostic signature, tumor immunity, DNA mismatch repair, methylation mechanisms, candidate molecular drugs, and gene mutations of the hub genes.

2. WGCNA and identification of key modules

There were 128 transcriptome data groups in GSE63514, including the normal control group (n = 24), CIN1 group (n = 14), CIN2 group (n = 22), CIN3 group (n = 40), and tumor group (n = 28). All specimens were included in the co-expression network. A soft threshold was obtained via the function "SFT powerEstimate" and set to 10. Next, gene modules were determined using TOM matrix. A total of 12 gene modules were detected in this analysis. As shown in Fig. 2, modules were black (n = 149), blue (n = 838), brown (n = 564), green (n = 317), green-yellow (n = 37), grey (n = 1104), magenta (n = 88), pink (n = 104), purple (n = 74), red (n = 189), turquoise (n = 1213), and yellow (n = 323) colored. In further analysis between modules and traits, the green module showed the highest phenotypic association with cervical cancer progression. Therefore, the green module was selected for subsequent validation.

3. GO term and KEGG pathway enrichment analyses
All genes in the green module were assessed by GO and KEGG analyses. As shown in Fig. 3, the green module was mainly enriched with epidermis development, peptide cross-linking, establishment of skin barrier, fatty acid derivative metabolic process, desmosome, serine-type peptidase activity, phosphatidylcholine acyl-chain remodeling and other signaling mechanisms. Among them, epithelial migration and fatty acid metabolism were related to cancer progression. Therefore, it was hypothesized that the influence of the green module on cervical cancer progression is tightly associated with the above pathways. The interactions among the modules are shown in the PPI diagram.

4. Survival analysis in relation to hub genes

Next, the relationships between genes in the green module with the prognosis of cervical cancer patients in the TCGA data set were assessed, screening core genes associated with cervical cancer progression. Our study showed that (Fig. 4) four genes had associations with cervical cancer prognosis, namely ALOX12B (p = 0.03854), KRT78 (p = 0.03178), RHOD (p = 0.02513), and ZNF750 (p = 0.04405). Further exploration of the potential roles of the above key genes in cervical cancer patients and the molecular mapping of early prevention of cervical cancer would contribute to early screening of this malignancy.

5. Co-expression analysis of hub genes

Then, co-expression of ALOX12B, KRT78, RHOD and ZNF750 was assessed. Genes with the most significant expression were screened (Fig. 5).

6. Related small molecule drugs screening

The GSCALite database was used to assess the drug sensitivities of hub genes and provide evidence for hub gene-based targeted therapy. The results are shown in Fig. 6, providing insights into drug targeted therapy.

7. Associations of hub genes’ expression with immune infiltration in CESC

The tumor microenvironment is mainly composed of tumor related fibroblasts, immune cells, the extracellular matrix, various growth factors, inflammatory factors, specific physical and chemical characteristics and cancer cells themselves. The tumor microenvironment significantly affects tumor diagnosis, survival outcome and treatment sensitivity. By analyzing the associations of core genes with tumor immune infiltration, the potential molecular mechanisms of core genes influencing the progression of cervical cancer were discussed. As shown in Fig. 7, ALOX12B expression was highly associated with macrophage infiltration in cervical cancer (r=-0.2, p = 8.05e-04); ZNF750 was associated with CD4 + T cell infiltration in cervical cancer (r = 0.217, p = 2.73e - 04) and dendritic cell infiltration (r = 0.249, p = 2.83e - 05). In addition, this study further explored the associations of core gene mutations with immune
infiltration. As shown in Fig. 8, mutations of the core genes could reduce the infiltration levels of most immune cells in cervical cancer.

8. DNA mismatch repair and methylation analyses of hub genes

The DNA mismatch repair system is a repair mechanism following cell replication, which plays a role in maintaining the fidelity of DNA replication and controlling genomic variations. Defects in mismatch repair function may lead to genetic instability or microsatellite instability, which may easily lead to the occurrence of tumors. In this study, the associations of core genes with 5 common DNA mismatch genes were analyzed, and the results showed that most of the 4 core genes were negatively correlated with DNA mismatch genes, and all were closely related to EPCAM gene expression (Fig. 9). It was hypothesized that EPCAM may participate in CESC tumor progression through the DNA mismatch repair system. In addition, the apparent modification mechanisms of core genes were studied in the DiseaseMeth database, and the results showed that the degrees of methylation of RHOD and ZNF750 were decreased, while those of ALOX12B and KRT78 showed no significant differences (Fig. 10). However, since there were only three normal samples, the results of methylation analysis still need to be further verified.

9. GSVA reveals close associations of hub genes with immune response in CESC

We assessed the mutations of core genes in cervical cancer via the cBioportal database, and selected the TCGA-PanCancer data set. The results (Fig. 11) showed that 60 patients (22%) had the four genes mutated; mutation rates were ZNF750 (9%), RHOD (7%), KRT78 (5%), and ALOX12B (4%), respectively, but the mutations had no significant effects on the mRNA expression of the four core genes. We then investigated the specific signaling pathways involved in the four core genes to explore the potential molecular mechanisms by which core genes influence tumor progression. GSVA results showed that the major pathways involved in the high expression of ALOX12B were ESTROGEN_RESPONSE_LATE, KRAS_SIGNALING, APICAL_SURFACE, HYPOXIA, and P53_PATHWAY. High expression of KRT78 was involved in P53_PATHWAY, CHOLESTEROL_HOMEOSTASIS, KRAS_SIGNALING, HYPOXIA, APOPTOSIS, etc. High RHOD expression was involved in P53_PATHWAY, CHOLESTEROL_HOMEOSTASIS, TNFA_SIGNALING_VIA_NFKB, APICAL_JUNCTION, and COAGULATION, and high expression of ZNF750 mainly contributed to P53_PATHWAY, APOPTOSIS, ESTROGEN_RESPONSE_LATE, and HEME_METABOLISM, among others. Thus, GSVA data indicated that most core genes had associations with P53 signaling and the hypoxia pathway (Fig. 12). These findings suggested that cervical cancer progression was related to P53 and hypoxia pathway regulation. Similarly, many immune-related pathways, such as the IL2 and interferon response pathways, were also enriched in the high expression groups of these core genes, suggesting their potential contributions to CESC-associated immune response.

Discussion
The early stage of cervical cancer has no obvious symptoms, and most patients are in the advanced stage when they see a doctor with symptoms such as vaginal discharge and bleeding [12]. Nowadays, human papillomavirus (HPV) infection is considered a necessary condition for cervical cancer, followed by multiple sexual partners, smoking, early sexual life, sexually transmitted diseases, low economic status, oral contraceptives and immunosuppression [13, 14]. Despite intense screening and various treatment options such as surgery, radiotherapy and chemotherapy, prognosis in CESC remains poor due to high risk of metastasis and recurrence. Additionally, mounting evidence suggests that altered expression of many genes contributes to CESC pathogenesis [15–17]. Therefore, we expect to identify some sensitive and novel biomarkers for diagnosis and prognosis in CESC.

In the complex tumor microenvironment, conventional single-gene assessments are not capable of revealing the complex signal transduction network of tumor-related genes. Meanwhile, WGCNA uses data of thousands of genes to determine interesting gene modules, performing association analyses with various phenotypes. WGCNA results are largely available [18, 19]. Here, gene expression profiling in GSE63514 (28 CESC and 24 noncancerous specimens) was assessed by WGCNA. As shown above, the green module had a significant correlation with CESC. Next, GO and KEGG analyses were performed for this module, and oncogenic signaling pathways were assessed to functionally characterize green module genes.

KEGG enrichment analysis demonstrated that green module genes were enriched in epidermis development, peptide cross-linking, establishment of skin barrier, fatty acid derivative metabolic process, desmosome, serine-type peptidase activity and phosphatidylcholine acyl-chain remodeling signaling pathways. In addition, they contributed to many biological processes, including cell adhesion, proliferation, differentiation and fatty acid metabolism [20–22], which may be related to the fact that cervical squamous cell carcinoma is mostly advanced at diagnosis. The majority of tumours show abnormally induced lipid metabolism enabling them to biosynthesize, elongate and desaturate fatty acids for increased growth [23–25]. Several reports have shown that lipid metabolism in cancer cells and immune cells in the tumor microenvironment highly contributes to immunosuppression, and antitumor immunity could be enhanced by targeting such metabolic pathways [22, 26]. This study suggested that genes in the green module might affect cervical cancer occurrence and progression by interfering with fatty acid metabolism.

Tumor immunotherapy is an anticancer therapy with the goal of activating the immune system in the hope that its own immune function would inhibit tumor tissues [27]. At present, tumor immunotherapy has been shown to have strong antitumor activities in some tumor types, including melanoma and non-small cell lung cancer, and tumor immunotherapy drugs are clinically applied [28–30]. This study confirmed that ALOX12B, KRT78, RHOD, and ZNF750, which are related to tumor progression, affected the prognosis of CESC patients. As shown above, ALOX12B expression was highly associated with macrophage infiltration in cervical cancer, and ZNF750 was correlated with CD4+ T cell infiltration and dendritic cell infiltration, suggesting that these two factors may influence the immune response of patients by controlling immune infiltration.
Next, to investigate the biological functions of hub genes, GSVA was performed. Both P53 and hypoxia pathways have important roles in tumor survival and recurrence [31–33]. The above results showed that these four core genes were mostly related to the P53 and hypoxia signaling pathways. Similarly, many immune-related pathways, such as the IL2 and interferon response pathways, were also enriched in the high-expression groups of these hub genes, suggesting their potential contributions to immune response in CESC.

To our knowledge, this is the first study assessing the relationship between RHOD and CESC. Interestingly, the results showed that RHOD is sensitive to most CESC drugs, especially Dasatinib (a tyrosine kinase inhibitor used to treat leukemia) [34, 35], BI – 2536 (an efficacious Polo-like kinase inhibitor inducing apoptosis in multiple human cancers) [36, 37], and clofarabine (second-generation nucleoside analog potently inhibiting leukemia) [38, 39]. These findings suggest that RHOD may be a new target for developing CESC drugs. However, the clinical significance of RHOD in CESC has not been reported to date, and more research is needed for confirmation.

In conclusion, four hub genes (ALOX12B, KRT78, RHOD, and ZNF750) were shown to be closely related to CESC survival by combining WGCNA, PPI, and other bioinformatics tools. Tumor microenvironment and drug sensitivity analyses examined the potential of immunotherapy and targeted therapy, and GSVA further demonstrated their significant impacts on CESC progression. More detailed studies are needed in the future to fully reveal their roles in CESC pathogenesis and usefulness as diagnostic and/or prognostic markers.

**Conclusions**

To sum up, the present study used various bioinformatics analysis tools to identify four novel hub genes, which may serve key roles in the progression and prognosis of cervical squamous cell carcinoma. Meanwhile, the green module might affect cervical cancer occurrence and progression by interfering with fatty acid metabolism. The ALOX12B and ZNF750 may influence the immune response of patients by controlling immune infiltration. And RHOD be novel potential drug targets by CTRP analysis. However, the lack of in vivo and in vitro experiments is a limitation of the present study, further molecular biological experiments are required to confirm the present findings, and confirm the role and function of these hub genes in CESC.

**Abbreviations**


**Declarations**
Availability of data and materials

The following information was supplied regarding data availability: Data is available at NCBI GEO: GSE63514.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This project was funded by the Science and Technology Department of Henan (China) (Grant no. 161100311100), the National Natural Science Foundation of China (Grant no. 81702967) and the Henan Medical Science and Technique Foundation (Grant no. LHGJ20190353).

Authors’ contributions

Xiaofeng Lv conceived and designed the experiments, performed the experiments, analyzed the data, contributed materials/analysis tools, prepared figures, authored or reviewed drafts of the paper, approved the final draft.

Acknowledgements

We thank all members of our study team for their whole-hearted cooperation and all included participants for their wonderful cooperation.

References


Figures
Figure 1

Study workflow. GEO, Gene Expression Omnibus; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; WGCNA, Weighted Gene Co-expression Network Analysis; GSVA, Gene Set Variation Analysis; TCGA, The Cancer Genome Atlas.
Figure 2

Determination of essential modules correlated with clinical traits in the TCGA dataset by WGCNA. (A) Sample clustering was carried out for outlier detection. The top and bottom images show a gene dendrogram and gene modules with distinct colors, respectively. (B) Soft-thresholding power analysis was performed to determine the scale-free fit index of network topology. (C) Scatter plot depicting eigengenes in the green module. (D) Associations of modules with traits (upper numbers in cells, correlation coefficients for a given module in the trait; lower numbers, p-values). The green module showed most correlation with cancer traits. (E) Hierarchical cluster analysis was performed for detecting co-expression clusters with respective color assignments. Various colors represent distinct modules in the built gene co-expression network by WGCNA. (F) Heatmap depicting the Topological Overlap Matrix (TOM) of genes assessed by weighted co-expression network analysis. Light and red colors represent low and high overlaps, respectively.
Figure 3

Significantly enriched biological processes and KEGG pathways of genes in the green module in CESC.
Figure 4

Associations of ALOX12B, KRT78, RHOD, and ZNF750 expression with disease-free survival time. (A) ALOX12B. (B) KRT78. (C) RHOD. (D) ZNF750. Red line, specimens with highly expressed genes; blue line, specimens with lowly expressed genes.
Figure 5

Chord plots indicating the relationships between each hub gene and its 10 related genes. (A) ALOX12B. (B) KRT78. (C) RHOD. (D) ZNF750.
Figure 6

CESC-related drugs in GSCALite database analysis.
Figure 7

Associations of hub genes with immune infiltration in CESC. (A) ALOX12B. (B) KRT78. (C) RHOD. (D) ZNF750. Each dot represents a sample in the dataset.
Figure 8

Associations of core gene mutations with immune infiltration. (A) ALOX12B. (B) KRT78. (C) RHOD. (D) ZNF750. *p<0.05, **p<0.01, ***p<0.001.
Figure 9

Associations of DNA mismatch repair genes with hub genes.*p<0.05, **p<0.01, ***p<0.001.
Figure 10

Methylation levels of hub genes. (A) ALOX12B. (B) KRT78. (C) RHOD. (D) ZNF750.
Figure 11

Mutations of core genes in cervical cancer detected via the cBioportal database.
Figure 12

Signaling pathways involving the four core genes. (A) ALOX12B. (B) KRT78. (C) RHOD. (D) ZNF750.