1. Introduction

Cervical cancer (CC) was the fourth most common malignant tumor in women as in 2018 approximately 570,000 new cases were reported and 311,000 deaths occurred. China and India account for the one-third of the global burden of CC [1]. Screening and vaccination have decreased incidence of CC in developed countries [2, 3]; however, in less developed countries, the incidence is increasing [4, 5]. About 69,000 new cases and 38,000 deaths of CC were reported in Association of Southeast Asian Nations (ASEAN) countries in 2020 [1]. In addition, there is an increase in incidence of CC in Western countries, accounting for 28% of the cases worldwide [6–8].

Cervical intraepithelial neoplasia gradually transforms into CC within 8 years and it is divided into low- and high-grade dysplastic lesions [9, 10]. The primary treatments for CC include surgery, radiotherapy, and chemotherapy [11]. Surgery cannot be performed for CC detected in late stages, while, radiotherapy and chemotherapy are burdensome for these patients. Chemotherapy can treat numerous advanced cancers but causes toxicity, drug resistance, and tumor recurrence related to drug-resistant cells [12–14]. It is reported that early CC treatment has a 5-year survival rate of 92% [15]; however, the survival rate of patients with advanced CC is only 17.3% [16]. The global incidence rate of CC is tremendously high. The survival rate of patients with CC who are effectively treated in the early stage is substantially higher than that of patients with advanced CC. Therefore, it is necessary to identify diagnostic markers for early CC treatment.

Pathological diagnosis is the gold standard for CC detection. However, in the early stages, before the beginning of morphological changes of cells, merely pathological examination cannot help make the diagnosis. The CC screening program in China includes human papillomavirus detection and thin-prep
cytologic test. The results of these tests are not necessarily abnormal in early CC. Therefore, if specific genetic alterations can be identified in the early stages of CC and targeted for treatment, the prognosis of CC can be enhanced. In this study, gene expression in normal and CC samples was analyzed using different databases, to screen suitable candidate genes for the early diagnosis of CC. Furthermore, the correlation between candidate gene expression and prognosis was verified.

2. Materials and Methods

2.1 Data acquisition and analysis

The research design and pathway followed in this study are summarized in Fig. 1. Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) was used as source of Microarray expression datasets for CC (GSE7803 and GSE9750) [17] (Table 1). GEO2R was used to screen the differentially expressed genes (DEGs) by comparing gene expression levels in normal subjects and patients with CC. In each dataset, the cut-off value was set to a $|\log \text{fold-change}| > 1$, and the adjusted $p$-value was $< 0.05$.

2.2 GO and KEGG enrichment analysis

The Entrez Gene ID of the DEGs were converted through org.Hs.eg.db package of R software. The clusterProfiler package was used for Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the DEGs. Results with $p < 0.05$ and $q < 0.2$ were considered as statistically significant.

2.3 Building the PPI network and screening hub genes

DEGs were rescreened at a $|\log \text{FC}| > 2$ and $p < 0.05$, and the DEGs obtained from the GSE9750 and GSE8703, datasets were considered as the intersection. The protein-protein interaction (PPI) network of the DEGs was constructed by using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (https://string-db.org/). The PPI network was structured using Cytoscape 3.7.2, and the hub genes were distinguished by the degree method using the cytoHubba plug-in. The DEGs were classified based on scores. Finally, the top three genes were considered to be hub genes for CC.

<table>
<thead>
<tr>
<th>GEO series</th>
<th>Normal</th>
<th>Tumor</th>
<th>Platform</th>
<th>Year</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE7803</td>
<td>10</td>
<td>21</td>
<td>GPL96</td>
<td>2007</td>
<td>Zhai Y</td>
</tr>
<tr>
<td>GSE9750</td>
<td>24</td>
<td>33</td>
<td>GPL96</td>
<td>2008</td>
<td>Murty VV</td>
</tr>
</tbody>
</table>

2. Materials and Methods

2.1 Data acquisition and analysis

The research design and pathway followed in this study are summarized in Fig. 1. Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) was used as source of Microarray expression datasets for CC (GSE7803 and GSE9750) [17] (Table 1). GEO2R was used to screen the differentially expressed genes (DEGs) by comparing gene expression levels in normal subjects and patients with CC. In each dataset, the cut-off value was set to a $|\log \text{fold-change}| > 1$, and the adjusted $p$-value was $< 0.05$.

2.2 GO and KEGG enrichment analysis

The Entrez Gene ID of the DEGs were converted through org.Hs.eg.db package of R software. The clusterProfiler package was used for Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the DEGs. Results with $p < 0.05$ and $q < 0.2$ were considered as statistically significant.

2.3 Building the PPI network and screening hub genes

DEGs were rescreened at a $|\log \text{FC}| > 2$ and $p < 0.05$, and the DEGs obtained from the GSE9750 and GSE8703, datasets were considered as the intersection. The protein-protein interaction (PPI) network of the DEGs was constructed by using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (https://string-db.org/). The PPI network was structured using Cytoscape 3.7.2, and the hub genes were distinguished by the degree method using the cytoHubba plug-in. The DEGs were classified based on scores. Finally, the top three genes were considered to be hub genes for CC.

<table>
<thead>
<tr>
<th>GEO series</th>
<th>Normal</th>
<th>Tumor</th>
<th>Platform</th>
<th>Year</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE7803</td>
<td>10</td>
<td>21</td>
<td>GPL96</td>
<td>2007</td>
<td>Zhai Y</td>
</tr>
<tr>
<td>GSE9750</td>
<td>24</td>
<td>33</td>
<td>GPL96</td>
<td>2008</td>
<td>Murty VV</td>
</tr>
</tbody>
</table>
2.4 ROC curve and survival analysis in GEPIA

Hub genes expression was analyzed using Gene Expression Profile Interaction Analysis (GEPIA) database (http://geopia.cancer-pku.cn) with Student’s t-test. Statistical significance was set at \( p < 0.05 \). Statistical package pROC was used to plot receiver operating curves (ROC). The ggplot2 package of R was used to assess the diagnostic ability of central genes for CC by calculating area under ROC curve (AUC). These hub genes were used as candidate genes for CC diagnosis. The candidate genes for CC were subjected to survival analysis using the GEPIA database.

2.5 Analysis via UCSC XENA

The University of California, Santa Cruz, (UCSC)-XENA (https://xenabrowser.net/datapages/) clinical/phenotype data was used to analyze candidate genes. The RNA-sequencing data of 26 tumors and the corresponding normal tissues were downloaded from the UCSC-XENA database and further processed into the transcripts per million reads format using the Toil process [18]. After log2 transformation, the Mann-Whitney U test was used to analyze and compare the expression of candidate genes in tumors and the corresponding normal tissues. Values with \( p < 0.05 \) were considered as statistically significant.

2.6 Verification of candidate genes in the HPA

Human Protein Atlas (HPA) database enables the mapping of proteins in almost all major human tissues and organs [19], including those different between normal and CC tissues (\( p < 0.05 \)); the candidate genes were verified using this data.

2.7 Statistical analysis

The DEGs of normal and CC tissues in the GSE7803 and GSE9750 datasets were evaluated using GEO2R, R software (version 3.6.3, R Foundation for Statistical Computing), and the Student’s t-test. Expression of candidate genes in tumor and normal tissues was analyzed and compared using the Mann-Whitney U test. Statistical significance was set at \( p < 0.05 \).

3. Results

3.1 Identification of DEGs

The GSE7803 and GSE9750 datasets were accessed from the GEO on January 7th, 2022, and gene expression in normal and CC tissues was analyzed. Genes with \(|\text{logFC}| > 1\) and adjusted \( p < 0.05 \) were considered statistically significant. Finally, the results of differential expression analysis showed that there were 481 upregulated and 559 downregulated DEGs in the GSE7803 dataset, and 1213 upregulated and 457 downregulated DEGs in the GSE9750 dataset (Fig. 2).

3.2 Biological functional analysis

To comprehend the biological functions of the DEGs and CC pathogenesis, GO and KEGG enrichment were analyzed. In the KEGG pathway enrichment analysis, cell cycle, DNA replication, p53 signaling pathways, cellular senescence, viral protein interaction with cytokine and cytokine receptor, Kaposi sarcoma-associated herpes virus infection, rheumatoid arthritis, oocyte meiosis, prostate cancer, and apoptosis were enriched in the GSE7803 dataset (Fig. 3A). In the GSE9750 dataset, the enriched pathways were the interleukin-17 (IL-17) signaling pathway, amoebiasis, cell cycle, p53 signaling pathway, DNA replication, proteoglycans in cancer, cytokine-cytokine receptor interaction, bladder cancer, tight junction, protein digestion and absorption (Fig. 3B). The DEGs of both datasets namely GSE7803 and GSE9750 play their roles in the cell cycle, DNA replication, and the p53 signaling pathway. The GO term enrichment of DEGs in the GSE7803 dataset suggested that DNA replication, mitotic nuclear division, spindle, cornified envelope, mitotic spindle, chromosomal region, single-stranded DNA-dependent Adenosine triphosphate (ATP)-dependent DNA helicase activity, single-stranded DNA-dependent ATPase activity, ATP-dependent DNA helicase activity, and ATP-dependent helicase activity were important GO terms (Fig. 3C). The DEGs in the GSE9705 dataset were enriched in epidermis development, epidermal cell differentiation, cornification, cornified envelope, cell-cell junction, desmosome, basolateral plasma membrane, structural constituent of the epidermis, endopeptidase inhibitor activity, serine-type endopeptidase activity, and structural constituent of the cytoskeleton (Fig. 3D). The DEGs of both GSE7803 and GSE9750 were enriched in the “cornified envelope”.

3.3 PPI network and hub genes

To further narrow down genes with considerable variations, DEGs were screened using \(|\text{logFC}| > 2\). The Venn diagram (http://bioinformatics.psb.ugent.be/webtools/Venn/) revealed 107 overlapping DEGs between both datasets (Fig. 4A). A PPI network of these DEGs was constructed using Cytoscape 3.7.2 software (Fig. 4B), and the degree method in the cytoHubba plug-in was used to identify the top 10 genes ranked by score (Table 2). This score indicates the strength of the relationship between the genes and the PPI network. A higher score indicates stronger connectivity. The top three genes, Estrogen receptor 1 (ESR1), fibronectin 1 (FN1), and Aurora kinase A (AURKA), were considered hub genes.

To validate the expression of hub genes in normal and CC tissues, Gene Expression Profiling Interactive Analysis (GEPIA) was used. AURKA was significantly up-regulated (Fig. 5A) and ESR1 was significantly down-regulated (Fig. 5B) in CC (\( p < 0.05 \)). The expression of FN1 (Fig. 5C) in CC was not significant. The diagnostic value of the hub genes for AURKA, ESR1 and FN1 were 0.999, 0.983, and 0.651, respectively (Fig. 5D–F). On the basis of these results, ESR1 and AURKA were identified as candidate genes for CC diagnosis. Survival analysis suggested that expression of the candidate genes was not correlated with overall survival (OS) in CC (Fig. 6A,B).
**FIGURE 2.** Identification of the differentially expressed genes (DEGs) in normal subjects and patients with cervical cancer (CC). Volcano plot of DEGs in (A) GSE7803 and (B) GSE9750 datasets. Red and blue colors indicate up- and down-regulated DEGs, respectively.

**FIGURE 3.** Functional enrichment of the differentially expressed genes (DEGs) in cervical cancer (CC) samples. (A, C) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Gene Ontology (GO) analysis of DEGs in the GSE7803 dataset in CC samples. (B, D) KEGG pathway and GO analysis of DEGs in the GSE9750 dataset in CC samples. BP: Biological Process, CC: Cellular Component, MF: Molecular Function, ATP: Adenosine triphosphate.
3.4 Expression of candidate genes in cancers

To further probe the relationship between the candidate genes and cancer of various types, expression of the candidate genes in various types of cancers and normal tissues was analyzed using UCSC-XENA. The results indicated that the expression of AURKA was significantly higher expression in several cancers \( (p < 0.05) \), including ovarian, breast, and cervical cancers, compared with that in the normal tissues while its expression was not prominent in pheochromocytomas, paragangliomas, and tenosynovial giant cell tumors (Fig. 7A). ESR1 was significantly overexpressed in normal tissues compared with that in 23 cancer types, except glioblastoma, head and neck squamous cell carcinoma, and stomach adenocarcinoma (Fig. 7B).

3.5 Expression of candidate genes in CC

To further investigate the expression of candidate genes (AURKA and ESR1) in HPA (full form), immunohistochemistry data obtained was evaluated in normal and CC tissues. It was observed that he expression of AURKA and ESR1 was significantly up-regulated and down-regulated, respectively, in CC tissues compared with that of normal tissues (Fig. 8). This result validated differential gene analysis (DEGs).

4. Discussion

For early diagnosis of CC, identification of biomarkers by the development of sequencing and molecular biology technique is favorable. The DEGs were identified by analyzing the GSE7803 and GSE9750 datasets in the GEO database. Three hub genes, AURKA, ESR1, and FN1, were identified while...
Figure 5. Expression of hub genes in normal and cervical cancer (CC) samples in Gene Expression Profiling Interactive Analysis (GEPIA) database. Expression of (A) Aurora kinase A (AURKA), (B) Estrogen receptor 1 (ESR1), and (C) fibronectin 1 (FN1). *p < 0.01; red and grey colors indicate cancerous tissue and normal tissue, respectively. The receiver operating characteristic curve (ROC) with the area under the curve (AUC) plots for (D) AURKA, (E) ESR1, and (F) FN1. TPM: Transcripts per million reads, T: tumour, N: normal, CESC: cervical squamous cell carcinoma, TPR: true positive rate, FPR: false positive rate, CI: Confidence interval.
**AURKA** and **ESR1** were candidate genes and also proteins encoded by both genes were highly specific for early diagnosis of CC. **AURKA** and **ESR1** had no significant correlation with OS in CC patients.

**AURK** is a member of the serine/threonine kinase family, which includes **AURKA**, **AURKB**, and **AURKC** [20, 21]. **AURKA** is on chromosome 20q13 [22] and plays a vital role, in cell functions such as centrosome maturation, cytokinesis, and spindle formation [23]. In the cell cycle, it regulates the G2 to M phases [24] and **AURKA** promotes tumor growth and cell survival by regulating human double minute 2 (HDM2)—induced ubiquitination and inhibiting p53 [25]. It allows cancer cells to escape apoptosis by inhibiting p73, an important member of the p53 family [26].

The size, stage, and metastasis of CC are closely linked with Hsa_circ_0075341, obtained by regulating the mir-149-5p/AURKA axis [27]. The 5' leading sequence of **AURKA** contains an internal ribosomal entry site, which increases the internal ribosomal entry site activity and its expression is pronounced in early stages of cancer [28]. During the tumor cycle, **AURKA** forms a positive feedback loop with the upregulated potassium channel tetramerization domain 12 (kctd12) promoting G2/M conversion and tumorigenesis [29]. The **AURKA** increases the migration and invasion abilities of tumor cells significantly [30] and is related with tumor stage and regional lymph node and distant metastases [31]. The **AURKA** also leads to an imbalance between cell division and apoptosis and participates in the upregulation of the p53 and breast invasive carcinoma 1 (BRCA1) pathways, the induction of telomerase activity through c-Myc, leading towards tumorigenesis [32]. Blocking **AURKA** expression can inhibit tumor proliferation and migration. Hence **AURKA** inhibitors are effective against CC, but its role in CC is not fully understood, and requires further research [33]. Low expression of **AURKA** is beneficial in patients with CC undergoing radiotherapy [34]; while high expression is related to relapse-free survival and OS of CC [34]. However, our findings are different; therefore, further research is required. The **AURKA** is overexpressed in many cancers, including gastric, esophageal [35], ovarian [36], hepatocellular carcinoma [32], and breast cancer [37]. Many cancers can be predicted due to **AURKA** much earlier [38].

**ESR1** encodes estrogen receptor α (ERα) and acts as a ligand-mediated transcription factor to activate multiple hormone target genes [39]. **ESR1** expression was inhibited by miR-944 as it enhances the migration and invasion of CC cells. The overexpression of miR-944 reversed the effect of **ESR1** in CC cells [40]. Drugs promoting the expression of **ESR1** can inhibit the proliferation of CC cells. **ESR1** is expressed in CC stromal cells and its expression in early tumor stromal cells is higher than that in late tumor stromal cells [41]. **ESR1** deletion plays an essential role in the progression and invasion of CC [42].

**ESR1** expression is upregulated in a various gynecological tumors while it is underexpressed in CC [43]. **ESR1** is significantly overexpressed in ovarian cancer and is a protective factor for the survival of patients with ovarian cancer [44].

**Septin 9 (SEPT9)** and sex-determining region Y (SRY)-box transcription factor 14 (SOX14) are also reported biomarkers of CC [45, 46], but their specificity for diagnosis is lower than that of **AURKA** and **ESR1**. Analysis of CC mRNA expression in the TCGA database and the GSE9750 dataset suggested that replication factor C subunit 5 (RFC5) is a useful biomarker for CC [47]. Using different datasets can lead to variable results, therefore, further experiments will verify the findings of the bioinformatics analysis. **ESR1** expression is significantly down-regulated in CC, and lower **ESR1** expression is inversely correlated with the OS of patients with CC in the HPA [48]. This is not consistent with our results and warrants further research.

**FIGURE 6.** Correlation of (A) Aurora kinase A (AURKA) and (B) Estrogen receptor 1 (ESR1) expression with overall survival (OS) in patients with cervical cancer (CC) in the GEPIA database. HR: Hazard ratio.
research.

The increase in AURKA expression and decrease in ESR1 expression can be useful for the clinical assessment of CC. It will improve prognosis and provide more treatment options for patients. The changes in gene expression can be detected before formation of low-grade squamous intraepithelial lesions (LSILs) which are precancerous lesions of CC. Otherwise it is quite difficult to find abnormal cervical cells through the pathological examination before the development of LSIL. Therefore, our findings indicate that AURKA and ESR1 may be useful as biomarkers for the early diagnosis of CC and can act as targets for new clinical approaches. As two datasets were used to screen for biomarkers in this study hence results have limitations. In future studies, more samples will be collected for experimental validation and evaluation of the carcinogenic mechanisms of these two genes.

5. Conclusions

AURKA and ESR1 are potential biomarkers identified through bioinformatics analysis, for early CC diagnosis and act as effective targets for treating CC.

AUTHOR CONTRIBUTIONS

YZ and GW—designed and directed the study; HC and XC—collected data and analyzed; HC, XC and XY—drafted the manuscript.

REFERENCES


