Identification of key genes and pathways in Ewing’s sarcoma using bioinformatics analysis

Chen Yan1*, Yehua Wang2*, Qian Wang3*, Xiao Feng2, Long Wang2, Zan Bu1, Bo Lu1, Jian Jiang2

1Department of Orthopaedics, Xuzhou Medical University, Xuzhou, Jiangsu Province, China; 2Department of Orthopaedics, the Affiliated Hospital of Xuzhou Medical University, Xuzhou, Jiangsu Province, China; 3Department of Rheumatology and Immunology, the Affiliated Hospital of Xuzhou Medical University, Xuzhou, Jiangsu Province, China.

*Chen Yan, Yehua Wang and Qian Wang contributed equally to this work.

Summary

Purpose: Ewing’s sarcoma (ES) is a highly aggressive malignant bone and soft tissue neoplasm. The purpose of our study was to identify candidate biomarkers of ES and uncover their potential molecular mechanisms.

Methods: The gene expression profiles of GSE45544 and GSE73166 were downloaded from Gene Expression Omnibus (GEO) database. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes pathway (KEGG) enrichment analyses were performed based on the Database for Annotation, Visualization and Integrated Discovery (DAVID) database, protein-protein interaction (PPI) network was constructed, and hub genes of the differentially expressed genes (DEGs) were identified by Cytoscape software.

Results: 586 DEGs in total were identified in ES, and the top up-regulated genes were extremely potent as ES biomarkers. GO function analysis data showed that these DEGs were significantly enriched in DNA translation, nucleus, and protein binding. KEGG pathway analysis revealed that DEGs were enriched in the pathways involved in cancer, HTLV-I infection, viral carcinogenesis, ribosome, and cell cycle. PPI network showed that most of the hub genes were for ribosomal proteins, associated in some way with the biological process of ES.

Conclusions: In conclusion, we identified the DEGs and hub genes which could promote our comprehension of detailed mechanisms involved in the development of ES.

Key words: bioinformatics, biomarker, differentially expressed genes, Ewing’s sarcoma, pathway

Introduction

ES is a highly aggressive malignant bone and soft tissue neoplasm, which predominately influences children, adolescents and young adults. It is the second primary malignant bone tumor in young people, with common occurrence in the soft tissues of adults [1]. ES can arise in any bone of the body, and about 25% of the patients are diagnosed post-metastasis [2]. These patients, diagnosed with organ metastases have a < 50% disease-free survival (DFS) time after 5 years of follow-up [3,4]. Until now, the treatment of ES was mainly depended on eradication of the sarcoma; patients with a localized tumor experienced improved outcomes compared to those with metastasis over decades [5,6]. Besides, patients with localized tumor, who experienced initial remission, could suffer a relapse of ES months or years later. Unfortunately, therapy for relapsed ES is usually ineffective. Till to date, there
is no reliable biomarker for early detection of ES in the clinic. Therefore, it is urgent to discover new biomarkers to screen ES at the earliest possible time. In addition, ES is a genetically complex and heterogeneous disease; its molecular mechanism and pathogenesis are still not clearly understood. Recently, the key genes associated with ES could be screened by microarray-based gene expression profiling technology. This enabled the investigation of molecular mechanisms involved in ES.

Microarray-based gene expression profiling could provide data on DEGs in the studied samples. By comparing gene expression profiles of ES tumor samples with those of normal (control) tissue samples, candidate biomarkers, functional in human ES tumors but not in control tissues, could be identified.

At first, we downloaded the original data (GSE45544 and GSE73166) from the GEO (http://www.ncbi.nlm.nih.gov/geo/). Gene expression profiles of sarcoma in patients with ES were compared with those in normal human tissues to identify the DEGs. These DEGs were subjected to GO, KEGG pathway enrichment analysis, and PPI network analysis. Based on the analyses of DEGs in biological functions and pathways, we aimed to identify potential candidate biomarkers to obtain deeper insights into the mechanisms of ES.

**Methods**

**Microarray data preparation**

GSE45544 and GSE73166, based on GPL6244 platform (Affymetrix Human Gene 1.0 ST Array), were used in this study. The gene expression profiles were downloaded from GEO database.

**DEG identification**

The raw data files used for this analysis included the CEL files (Affymetrix platform). Statistical software R (version 3.4.0, https://www.r-project.org/) and packages of Bioconductor (http://www.bioconductor.org/) were applied for significance analysis of DEGs between ES samples and normal tissue samples. Significant DEGs were selected based on “limma” package of Bioconductor. P<0.05 was considered statistically significant.

**GO and KEGG pathway enrichment analysis**

GO analysis is a common method for annotating genes [7] and KEGG analysis is a popular method for functional analyses of genes [8,9]. Biological analysis of DEGs was performed by GO enrichment, containing biological process (BP), cellular component (CC), and molecular function (MF). To analyze the effect of signaling pathway of these DEGs on function, KEGG pathway analysis was executed by the Database for Annotation, Visualization and Integrated Discovery (DAVID) online tool (https://david.ncifcrf.gov/) [10]. P<0.05 was considered statistically significant.

**PPI network and module construction**

Search Tool for the Retrieval of Interacting Genes (STRING) database is a very useful online tool to evaluate PPI information, providing systematic data regarding cellular processes [11]. In order to evaluate the interactive network relationships of the DEGs, we integrated and mapped them to STRING database; only a confidence score of 0.7 was selected as the cut-off criterion. Subsequently, PPI networks were constructed using the Cytoscape software [12]. In order to screen the clusters of original PPI network, module analysis by Molecular Complex Detection (MCODE) was carried out in Cytoscape software. The criteria for the network module screening were set beyond the number of nodes and MCODE scores (Degree Cutoff: 2, Node Score Cutoff: 0.2, K-Core: 2). Moreover, the function annotation and pathway enrichment analysis for DEGs were performed by DAVID software in the modules with a threshold of p<0.05.

**Results**

**ES-related DEG identification**

After quality evaluation of raw microarray data, we removed 6 cell line samples and 10 patient

<table>
<thead>
<tr>
<th>Probe set</th>
<th>Gene symbol</th>
<th>Fold change</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up-regulated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8111524</td>
<td>UGT3A2</td>
<td>15.39</td>
<td>3.33E-07</td>
</tr>
<tr>
<td>7908204</td>
<td>HMCN1</td>
<td>15.16</td>
<td>1.86E-09</td>
</tr>
<tr>
<td>8067869</td>
<td>RBM11</td>
<td>14.72</td>
<td>7.45E-08</td>
</tr>
<tr>
<td>8102200</td>
<td>DKK2</td>
<td>12.66</td>
<td>1.60E-08</td>
</tr>
<tr>
<td>7938329</td>
<td>SNORA23</td>
<td>12.26</td>
<td>6.74E-10</td>
</tr>
<tr>
<td>8096176</td>
<td>PTPN13</td>
<td>12.14</td>
<td>4.58E-09</td>
</tr>
<tr>
<td>8045688</td>
<td>TNFAIP6</td>
<td>12.11</td>
<td>1.22E-07</td>
</tr>
<tr>
<td>8095251</td>
<td>LIP1</td>
<td>11.43</td>
<td>1.97E-07</td>
</tr>
<tr>
<td>8021245</td>
<td>DCC</td>
<td>10.77</td>
<td>1.56E-07</td>
</tr>
<tr>
<td>8046536</td>
<td>HOXD10</td>
<td>10.35</td>
<td>7.00E-07</td>
</tr>
<tr>
<td>Down-regulated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7907160</td>
<td>ATP1B1</td>
<td>-5.76</td>
<td>2.17E-07</td>
</tr>
<tr>
<td>8149927</td>
<td>CLU</td>
<td>-5.32</td>
<td>3.24E-06</td>
</tr>
<tr>
<td>8172204</td>
<td>MAOB</td>
<td>-5.31</td>
<td>1.64E-06</td>
</tr>
<tr>
<td>7935188</td>
<td>SORBS1</td>
<td>-5.25</td>
<td>3.19E-06</td>
</tr>
<tr>
<td>7944667</td>
<td>SORL1</td>
<td>-5.11</td>
<td>1.51E-08</td>
</tr>
<tr>
<td>8090780</td>
<td>SYNO2</td>
<td>-4.94</td>
<td>5.61E-05</td>
</tr>
<tr>
<td>7905592</td>
<td>KIAA1524</td>
<td>-4.88</td>
<td>2.48E-04</td>
</tr>
<tr>
<td>7988414</td>
<td>GATM</td>
<td>-4.84</td>
<td>2.08E-05</td>
</tr>
<tr>
<td>8045438</td>
<td>IGKV2D-28</td>
<td>-4.45</td>
<td>2.64E-04</td>
</tr>
<tr>
<td>7985256</td>
<td>CKMT1B</td>
<td>-5.91</td>
<td>1.81E-04</td>
</tr>
</tbody>
</table>

Table 1. The most significantly up-regulated and down-regulated DEGs (top ten, ES versus normal tissue)
samples with osteosarcoma, based on the GPL6244 platform. The total number of samples analyzed was 22 each of ES samples and control tissue samples. The GSE45544 dataset contained 14 patient and 22 normal tissue samples; GSE73166 dataset contained 8 patient tissue samples only. For statistical analysis, software R was used for preprocessing and differential gene expression analysis of microarray data. Using the p<0.01 and fold change >2.0 criteria, a total of 586 genes were identified from the analyses of GSE45544 and GSE73166, out of which, 447 were up-regulated and 139 down-regulated; the heat map is shown in Figure 1. The expression levels of the top 10 overexpressed and down-regulated DEGs are displayed in Table 1.

**GO term analysis**

We uploaded all DEGs to the online software DAVID to identify altered biological functions. GO analysis results showed that enriched terms of the DEGs were significantly focused on biological processes (BP), including DNA templated transcription, negative regulation of RNA polymerase II promoter-driven transcription, cell division, rRNA processing, and covalent chromatin modification (Figure 2A). In cell component (CC) part, the DEGs were enriched in nucleus, nucleoplasm, cytoplasm, and membranes (Figure 2B). The molecular function (MF) analysis further showed that these DEGs were mainly enriched in protein binding, poly (A) binding, DNA binding, and nucleic acid binding (Figure 2C).

![Color Key](image1.png)

![Figure 1](image2.png)

**Figure 1.** Heat map of differentially expressed genes (DEGs) of Ewing’s sarcoma (ES) with fold change more than twice. Red, up-regulated DEGs; green, down-regulated DEGs; Purple, ES sample; Orange, normal sample.
DEGs in Ewing’s sarcoma

KEGG pathway analysis

The most significantly enriched pathways of DEGs, as obtained from KEGG analysis are listed in Table 2. The DEGs were mainly enriched in the pathways associated with cancer, HTLV-I infection, viral carcinogenesis, ribosome, and cell cycle.

Hub genes and PPI network module identification

STRING (a database of known and predicted protein interactions) is a method generally used to predict protein-protein interactions among the DEGs. The screened DEGs were uploaded to STRING to gain PPI data. In PPI networks, 9 nodal genes, including RPS27, RPS2, RPS24, RPL23A, RPS19, RPL7, WDR12, RPL18A and NSA2 showed strong association with other nodal genes, suggesting higher degree for hubs (Figure 3A). These hub genes might have crucial effects on the development of ES.

Network module analysis

A significant module, obtained from PPI network using Cytoscape by default criteria, was further followed up with functional enrichment analysis (Figure 3B). The module chosen had 20 nodes with a MCODE score of 19. The GO function and KEGG pathway analysis of these hub genes are shown in Table 3. The module was mainly associated with translation, ribosome, and structural constituent of ribosome. This module was related
to homologous recombination during KEGG pathway analysis.

**Discussion**

In the present study we identified significant DEGs by comparing human ES tissue with normal tissues through reanalysis of GSE45544 and GSE73166, downloaded from GEO database. To further analyze the roles of these DEGs, bioinformatics analyses including GO and KEGG pathway were performed using DAVID database, PPI network, and module construction by Cytoscape software.

A series of DEGs, namely **HMCN1**, **RBM11**, **DKK2**, **SNORA23**, **PTPN13** and **TNFAIP6** exhibited more than remarkable 12-fold change in the ES group than in the control group. Thus, we hypothesized that these DEGs might have the potential to be used as biomarkers for diagnosing ES, but only after further experimental confirmation in the future. **HMCN1** was reported to be significantly up-regulated in the ES and fibromatosis samples [13]. Though **RBM11** (RNA binding motif 11) has no such report, another RNA binding motif 3 was reported to participate in cancer. **RBM3** was overexpressed in esophageal and gastric adenocarcinoma, which might independently predict tumor recurrence [14]. **Dickkopf-2** (**DKK2**) could increase tumor vascular density and perfusion, and coverage of blood vessels by pericytes [15]. **SNORA23** (small nucleolar noncoding RNA 23) was overexpressed in human pancreatic ductal adenocarcinoma and correlated with patients' survival [16]. **PTPN13** is a new candidate tumor-suppressing gene in human cancer, negatively correlated with tumor grade and stage [17,18]. As already known, **TNF alpha** takes part in cancer development [19,20]. Serum **TNFAIP6** (TNF alpha-induced protein 6) was reported as a potential biomarker for Crohn's disease and ulcerative colitis [21]. Taken together, these DEGs are potent biomarkers to diagnose ES, provided further validations are performed.
We also analyzed the biological functions of the DEGs based on GO annotation in DAVID database. At first, the majority of the DEGs were found enriched in DNA-templated transcription and regulation of transcription from RNA polymerase II promoter by BP function analysis. The results indicated that ES occurrence may begin with an erroneous DNA transcription. Other enriched categories containing those associated with cell division, chromatin modification, and rRNA processing were also indicated. In CC category, our data revealed nucleus, nucleoplasm, and cytoplasm to be the most significant items. The results showed association of DEGs with DNA reproduction and their possible participation in various biological processes of ES. In MF ontology, the most enriched item was protein binding, followed by DNA-, RNA-, and nucleic acid binding.

From the PPI network, a number of hub proteins have been identified, namely RPS19, RPS24, RPS27, RPL2, RPL7, RPL18A, RPL23A, WDR12 and NSA2, of which, most are ribosomal proteins. A full ribosome contains one large (60S) and one small (40S) subunit. It regulates the major step of gene expression by providing the site for translation. Ribosome biogenesis should be highly coordinated owing to the high accuracy and speed of the protein synthesis process. Hence, any aberration at any step might lead to abnormal cellular phenotype [22]. Recently, ribosomal proteins were related to diverse extra-ribosomal functions, including those associated with DNA repair, cell cycle, cell proliferation, differentiation, and malignant transformation. Dysfunction and impairment of ribosomal proteins could lead to malignant tumor or hematological and cardiovascular diseases [23-26]. RPS27 is correlated with gastric cancer metastasis [27]. RPS19 deficiency could lead to macrocytic anemia and bone marrow failure in mouse model [28]. RPL6 may be a potential prognostic biomarker for patients with gastric cancer [29]. The loss of RPL29 occurs during regulation of angiogenesis [30] by reducing micro-vessel formation stimulated by vascular endothelial growth factor [31]. Modules established from PPI network showed that 80% (16/20) complements were ribosomal proteins. This implies that the hub ribosomal proteins in the module may play important roles in ES.

### Table 3. GO and KEGG pathway analyses of the module from PPI network

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>p value</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP ontology</td>
<td>Translation</td>
<td>6.35E-14</td>
<td>RPS27, RPS19, RPL18A, RPL13A, RPL6, RPL3, RPL23A, RPL10A, RPL12, RPS2, RPL59, RPS24</td>
</tr>
<tr>
<td></td>
<td>Ribosomal large subunit assembly</td>
<td>1.40E-05</td>
<td>RPL6, RPL3, RPL23A, RPL12</td>
</tr>
<tr>
<td></td>
<td>Regulation of translational initiation</td>
<td>5.34E-04</td>
<td>EIF3E, EIF4A2, EIF4A1</td>
</tr>
<tr>
<td></td>
<td>Cytoplasmic translation</td>
<td>0.001</td>
<td>RPL7, RPL6, RPLP0</td>
</tr>
<tr>
<td>CC ontology</td>
<td>Ribosome</td>
<td>1.66E-14</td>
<td>RPS27, RPL18A, RPL7, RPL6, RPLP0, RPL3, RPLP2, RPL12, RPL59, RPS24</td>
</tr>
<tr>
<td></td>
<td>Cytosolic large ribosomal subunit</td>
<td>1.45E-10</td>
<td>RPL7, RPL13A, RPL6, RPLP0, RPL3, RPL23A, RPL10A, RPL12</td>
</tr>
<tr>
<td></td>
<td>Nucleolus</td>
<td>6.07E-04</td>
<td>RPS19, RPL7, RPL23A, RPL10A, SRP72, NSA2</td>
</tr>
<tr>
<td></td>
<td>Membrane</td>
<td>0.005</td>
<td>RPS19, RPL7, EIF3E, RPLP0, EIF4A1, RPL10A</td>
</tr>
<tr>
<td></td>
<td>Focal adhesion</td>
<td>0.006</td>
<td>RPS19, RPL7, RPLP0, RPL10A</td>
</tr>
<tr>
<td>MF ontology</td>
<td>Structural constituent of ribosome</td>
<td>6.77E-19</td>
<td>RPL2, RPL23A, RPS2, RPS39, RPS27, RPS19, RPL7, RPL18A, RPL15A, RPL6, RPLP0, RPL3, RPL12, RPL10A, RPS24</td>
</tr>
<tr>
<td></td>
<td>Poly(A) RNA binding</td>
<td>1.76E-06</td>
<td>RPS19, RPL7, EIF3E, EIF4A2, RPLP0, EIF4A1, RPL23A, RPL10A, SRP72, NSA2</td>
</tr>
<tr>
<td></td>
<td>Translation initiation factor activity</td>
<td>0.007</td>
<td>EIF3E, EIF4A2, EIF4A1</td>
</tr>
<tr>
<td>KEGG</td>
<td>Homologous recombination</td>
<td>0.02</td>
<td>NBN, RAD50</td>
</tr>
</tbody>
</table>

BP: biological process, CC: cellular component, MF: molecular function
The biological function of these module genes was further analyzed by GO annotation. In BP ontology, these module genes were enriched in gene translation and ribosomal large subunit assembly. In CC ontology, the main items were ribosome, cytosolic large ribosomal subunit, and nucleolus. In MF ontology, they focused on the structural constituents of ribosome. As the functions were mostly associated with ribosomal proteins, hence, we considered that the process of ES was correlated to the dysfunction of ribosomal proteins.

Pathway analysis of DEGs revealed 5 pathways that were mainly enriched: pathways in cancer, HTLV-1 (human T-cell leukemia virus type 1) infection, viral carcinogenesis, ribosome, and cell cycle. Obviously, general pathways involved in cancer would affect ES. Most hub genes/proteins, discussed above, are ribosomal genes/proteins, hence implying their role in ES by ribosome-specific pathway. HTLV-1 infection could result in adult T-cell leukemia/lymphoma that could group with viral carcinogenesis. Statistically, approximately 12% of human tumors, across the world, are caused by oncoviral infection. Oncoviral replication and persistence in human host were related to the activation of cancer-causing pathways. It was also reported that RPL18A interacts with nuclear/nucleolar protein to form a complex that retains viral mRNA in HTLV-1 infected cells [32]. Hence, it is comprehensible as to why HTLV-1 infection...
DEGs in Ewing’s sarcoma

and viral carcinogenesis influence the occurrence of ES.

Reports also showed that ribosomal proteins could serve as cancer biomarkers. RPL23A, RPL27, and RPL30 are increased in hepatocellular carcinoma, while RPL6 and RPL15 are up-regulated in gastric cancer [29,33]. RPL6 [29] and RPL23 [34] were suggested as candidate prognostic indicators of cancer and RPS2 could serve as therapeutic molecular target in prostate cancer [35]. Recombinant RPL23a and RPL31 also possessed anticancer activities [36,37]. Overexpressed RPL6 was associated with drug resistance and its reduced expression level could revert the resistance [38]. The 9 hub genes in PPI network were overexpressed (doubled approximately) in ES than in normal tissues, which may be considered as a bunch of proteins participating in ES development. Based on these results, we speculate that these hub genes could be recognized as targets for the diagnosis or prediction of ES development, if validated by further investigations.

Conclusion

In summary, our data provide a comprehensive bioinformatics analysis of DEGs, which may be involved in the development of ES. Moreover, it also predicts a set of targets for future investigation into the molecular mechanisms and biomarkers of ES. However, further experiments are recommended to confirm the functions of the identified genes and validate our speculation.

Conflict of interests

The authors declare no conflict of interests.

References

15. Park H, Jung HY, Choi HJ et al. Distinct roles of DKK1 and DKK2 in tumor angiogenesis. Angiogenesis 2014;17:221-34.


