Original Article
Identification of differentially-expressed genes in lung squamous cell carcinoma and correlation levels with prognosis through integrated bioinformatics analysis

Licui Zhang1*, Chen Zhong2*, Yang Gu1, Yajing Ma1, Xinliang Ming3, Xin Su1, Min Liu1

1Department of Clinical Laboratory, The First Affiliated Hospital of The Medical College, Shihezi University, Shihezi 832000, Xinjiang Autonomous Region, PR China; 2Department of Gastrointestinal & Thyroid Surgery, The First Affiliated Hospital of The Medical College, Shihezi University, Shihezi 832000, Xinjiang Autonomous Region, PR China; 3Department of Genetic Diagnosis Center, Zhongnan Hospital of Wuhan University, Wuhan 430000, PR China. *Equal contributors.

Received June 30, 2019; Accepted September 3, 2019; Epub November 15, 2019; Published November 30, 2019

Abstract: Background: The aim of the current study was to screen differentially-expressed genes (DEGs) relevant to cancer progression and prognosis of squamous cell lung carcinoma (SqCLC). Methods: DEGs mRNA expression data of SqCLC was screened from the Oncomine database. This data was further analyzed by comparing tumor tissues to normal tissues. Prognostic values of DEGs relevant to SqCLC were investigated using the “Kaplan-Meier Plotter” (KM plotter) database. Bioinformation for included genes was analyzed by gene GO and KEGG enrichment, aiming to explain the potential roles of identified genes in SqCLC. Protein-protein interaction (PPI) of the genes was evaluated using the STRING database. Results: Four independent microarray datasets relevant to SqCLC were identified in the Oncomine database, with the top 10 consistently upregulated and top 10 consistently downregulated genes included in the present analysis. Significant differences of overall survival (OS) were correlated with SMC4, HIST2H2AA3, GMPS, CKS1B, POLR2H, PDCD10, PLOD2, DVL3, C-type CLEC3B, TNNC1, FAM107A, FYR, MEF2C, SLIT3, CX3CR1, C17orf91, LIM, and LIMCH1 (all P < 0.05). Possible protein-protein interaction analysis of the top 20 dysregulated genes showed that proteins of SMC4, POLR2H, and NCBP2 in upregulated genes and TNNC1 and MEF2C in downregulated genes interacted with more than 5 other proteins. This may play an important role in the development of SqCLC. Conclusion: SMC4, POLR2H, TNNC1, and MEF2C genes were dysregulated in SqCLC. Thus, they may play an essential role in the development of SqCLC, as biomarkers for patient prognosis.

Keywords: Squamous cell lung carcinoma, data mining, prognosis, Oncomine database

Introduction
Lung cancer is one of the most diagnosed malignant tumors, clinically, and the leading cause of carcinoma related deaths [1]. Epidemiology studies have indicated that lung cancer ranks in the top 3 for cancer incidence rates in the U.S [2, 3]. Therefore, lung cancer has become a big problem for humans, placing a heavy burden on the government [4]. Squamous cell lung carcinoma (SqCLC) is one of the major subtypes of non-small cell lung cancer (NSCLC), accounting for 40% of all diagnosed lung carcinomas [5, 6]. Prognosis of SqCLC remains poor. SqCLC is difficult to manage, as it is not sensitive to chemotherapy and radiation [7, 8]. Differentially-expressed genes (DEGs) are common in malignant tumors. They play an essential role in cancer cell development. Liu et al. [9] evaluated genes that correlated with lung adenocarcinoma progression and prognosis using Oncomine and Cancer Genome Atlas databases. They found eighty DEGs between cancer tissues and corresponding normal lung tissues. Results suggested that AGER and CCNB1 can be used as potential biomarkers for lung adenocarcinoma diagnosis and prognosis. Dysregulated expression of cancer related genes is common in SqCLC, playing an important role in development and prognosis [10]. However, identification of dysregulated genes and analysis of the correlation and relationship with SqCLC patients prognosis remain difficult in
Identification of differentially-expressed genes in lung squamous cell carcinoma

Table 1. General information of included data sets

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Method</th>
<th>Sample size</th>
<th>Genes measured</th>
<th>Data type</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bhattacharjee</td>
<td>Human Genome U95A-Av2 Array</td>
<td>203</td>
<td>8,603</td>
<td>mRNA</td>
<td>2001</td>
</tr>
<tr>
<td>Hou</td>
<td>Human Genome U133 Plus 2.0</td>
<td>156</td>
<td>19,574</td>
<td>mRNA</td>
<td>2010</td>
</tr>
<tr>
<td>Talbot</td>
<td>Human Genome U95A-Av2 Array</td>
<td>93</td>
<td>8,603</td>
<td>mRNA</td>
<td>2005</td>
</tr>
<tr>
<td>Yamagata</td>
<td>Platform not pre-defined</td>
<td>31</td>
<td>2,509</td>
<td>mRNA</td>
<td>2003</td>
</tr>
</tbody>
</table>

many datasets [11]. The present study provides a method of screening dysregulated genes relevant to SqCLC, examining biological function and correlation levels with patient prognosis.

**Material and methods**

**Database selection**

The Oncomine database (https://www.oncomine.org/) provides differential expression analyses. It compares most major types of cancer with respective normal tissues, as well as a variety of cancer subtypes, providing clinical-based and pathology-based analyses. Dysregulated genes between cancer and normal tissues of SqCLC patients were first screened through the Oncomine database (https://www.oncomine.org/). The top 120 dysregulated genes were included and 20 constantly upregulated or downregulated genes were further analyzed for prognosis (Table 1). Biological information of the included 20 gene were enriched through ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). Protein-protein interaction (PPI) of the 20 constantly dysregulated genes were investigated using the STRING database.

**Data extraction from Oncomine database and analysis**

Using the Oncomine database, the current study screened all dysregulated genes between cancer tissues and normal tissues of SqCLC patients. Data screen restrictions was as follows: (1) Cancer type: squamous cell lung carcinoma; (2) Tissue comparison: cancer tissue versus normal tissue; (3) Data type: mRNA; (4) mRNA fold change: more than two fold; and (5) Significance: P < 0.001. The top 120 dysregulated genes (60 upregulated and 60 downregulated) were included in the biological information analysis via GO and KEGG enrichment. Of the included 120 genes, 20 constantly dysregulated genes (10 upregulated and 10 downregulated genes) were selected for survival and PPI analysis.

**Survival analysis by Kaplan-Meier plotter**

Prognosis (overall survival, OS) significance of the 20 constantly dysregulated genes of SqCLC patients was analyzed using the Kaplan-Meier Plotter database. According to the median mRNA expression level, the data was divided into two groups. These included a high expression and low expression group for each of the included 20 genes.

**Bioinformatics analyses of the databases**

Online analysis software DAVID (https://david.ncifcrf.gov/) was used to enrich the biological information of the top 120 genes screened in the Oncomine database. Two-tailed P < 0.05 was set as the cut-off in enrichment of GO and KEGG.

**Results**

**Dysregulated genes for SqCLC**

Four independent microarray datasets, relevant to SqCLC, were identified in the Oncomine database [12-15]. The current study identified the 120 most dysregulated genes, comparing tumor tissues to normal tissues of SqCLC patients (P < 0.05). Included patients had comparable demographic characteristics. Regarding the 120 dysregulated genes, 60 were upregulated (Figure 1) and 60 were downregulated (Figure 2). Of the 120 most dysregulated genes, 10 consistently upregulated and 10 consistently downregulated genes were selected as objective genes for further analysis.

**Functional enrichment analyses**

The 20 dysregulated genes contained three enriched gene ontology categories, including BP: Biological process, CC: Cellular component,
Identification of differentially-expressed genes in lung squamous cell carcinoma

and MF: molecular function. For the 10 upregulated genes, enriched biological processes were mainly related to positive regulation of transcription from RNA polymerase II promoter, negative regulation of the apoptotic process, activation of protein kinase activity, and protein homotetramerization (Figure 3A). In the CC category, the membrane, cytosol, nucleoplasm, nucleus, and cytoplasm were mainly enriched in the cellular component of the 10 upregulated genes (Figure 3B). Regarding MF, ATP binding, chromatin binding, transcription regulatory region DNA binding, histone binding, and ubiquitin conjugating enzyme activity were the top 4 enrichments (Figure 3C). Ten upregulated genes were enriched in cell cycle, purine metabolism, pyrimidine metabolism, and drug metabolism-other enzymes (Figure 3D).

Regarding the 10 downregulated genes, BP enrichment included negative regulation of gene expression, angiogenesis, positive regulation of angiogenesis, and positive regulation of GTPase activity (Figure 4A). For the cellular

Figure 1. The 60 most upregulated genes, comparing tumor tissues to normal tissues of SqCLC patients screened in the Oncomine database.

Figure 2. The 60 most downregulated genes, comparing tumor tissues to normal tissues of SqCLC patients screened in the Oncomine database.
Identification of differentially-expressed genes in lung squamous cell carcinoma

Figure 3. Go and KEGG analysis of the 10 upregulated genes (A: Biological process; B: Cellular component; C: Molecular function; D: KEGG).

Figure 4. Go and KEGG analysis of the 10 downregulated genes. Y-axis demonstrates the gene function and pathway enrichment and X-axis indicates counts. (A: Biological process; B: Cellular component; C: Molecular function; D: KEGG).
Identification of differentially-expressed genes in lung squamous cell carcinoma

- **ACTL6A**
- **SMC4**
- **HIST2H2AA3**
- **GMPS**
- **CKS1B**
- **POLR2H**
- **NCBP2**
- **PDCD10**
- **PLOD2**
- **DVL3**
- **CLEC3B**
- **TNNC1**
Identification of differentially-expressed genes in lung squamous cell carcinoma

Figure 5. Survival curve comparing low and high expression of the 20 genes.
Identification of differentially-expressed genes in lung squamous cell carcinoma

component, two items were enriched, cell leading edge and cell surface (Figure 4B). Binding, protein binding, metal ion binding, and molecular transducer activity were enriched in the category of molecular function enrichment (Figure 4C). Concerning KEGG analysis, only axon guidance was included (Figure 4D).

Potential roles of genes in SqCLC progression

The prognostic significance of mRNA expression levels of the 20 genes and prognosis of lung cancer patients was analyzed via online software. Differences in OS between upregulated and down regulated genes are shown in Figure 5. OS was correlated with SMC4, HIST2H2AA3, GMPS, CKS1B, POLR2H, PDCD10, PLOD2, DVL3, CLEC3B, TNNC1, FAM107A, FYR, MEF2C, SLIT3, CX3CR1, C17orf91, and LIMCH1 (all p < 0.05).

Protein-protein interaction of the 20 dysregulated genes

The possible protein-protein interaction of the top 20 dysregulated genes was analyzed using the STRING database. Results revealed that proteins of SMC4, POLR2H, and NCBP2 in upregulated genes (Figure 6) and TNNC1 and MEF2C in downregulated genes (Figure 7) interacted with more than 5 other proteins.

Discussion

In recent years, bioinformation analysis has developed quickly. Thus, more and more microarray and sequence datasets can be used in open databases, including Oncomine, GEO (http://www.ncbi.nlm.nih.gov/geo), TGCA (http://www.tcga.org/), and Kaplan-Meier Plotter [16, 17]. Relevant data, such as gene expression and clinical information (disease type, age, gender, survival), can be freely downloaded or analyzed. This data may be further used for clinical practices or deep data mining.

In the Oncomine database, gene expression data can be screened and mined according to the needs of researchers [18, 19]. Moreover, researchers can discover genes highly ranked by over-expression or DNA copy gain in medulloblastoma clinical specimens. This database allows researchers to explore interest genes in the largest normal tissue panel and survey a growing collection of TCGA gene expression and DNA copy number datasets [17, 20-22].

Long et al. [23] identified DEGs and enriched pathways in lung carcinoma using bioinformatics methods. In their study, the author screened GEO databases and selected GSE19804 as the study data. They identified DEGs and relevant pathways. They also provided survival data.

The present study screened the Oncomine database, finding the top 120 dysregulated genes relevant to SqCLC. Of the 120 included genes, 60 upregulated genes were enriched in the biological functions of positive regulation of transcription from RNA polymerase II promoter, negative regulation of apoptotic process, activation of protein kinase activity, protein homotetramerization with the KEGG pathway enrichment of cell cycle, purine metabolism, pyrimidine metabolism, and drug metabolism-other enzymes. Regarding the top 60
Identification of differentially-expressed genes in lung squamous cell carcinoma

C17orf91, and LIMCH1 genes, high expression of mRNA was correlated with good prognosis. Results suggest that the above gene expression levels can be used as biomarkers for SqCLC patient prognosis. The current study also screened PubMed databases, but did not discover any studies that used the above genes coding protein as biomarkers for SqCLC prognosis. Therefore, present findings suggest that the survival biomarkers of the above genes may be applied to mRNA levels. PPI analysis of the top 20 dysregulated genes demonstrated that SMC4, POLR2H, NCBP2, TNNC1, and MEF2C interacted with more than 5 other proteins. Results suggest that these 5 genes may play important roles in SqCLC development.

Conclusion

In summary, the current study screened the Oncomine database. A total of 4 microarray datasets were included and analyzed in the present work. After deep data mining, it was found that SMC4, POLR2H, TNNC1, and MEF2C genes were dysregulated in SqCLC. Thus, they may play essential roles in SqCLC development. They may be used as biomarkers of SqCLC prognosis. However, present results require further validation in future studies.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Min Liu, Department of Clinical Laboratory, The First Affiliated Hospital of The Medical College, Shihze University, No. 107 Beier Road, Shihze 832008, Xinjiang Autonomous Region, PR China. E-mail: jm37wc@163.com

References

Identification of differentially-expressed genes in lung squamous cell carcinoma


