Critical genes in head and neck squamous cell carcinoma revealed by bioinformatic analysis of gene expression data

B. Wang¹, T. Wang², X.L. Cao³ and Y. Li³

¹Department of Otorhinolaryngology, Hangzhou Obstetrics & Gynecology Hospital, Hangzhou, Zhejiang Province, China
²Shandong Traffic Hospital, Jinan, Shandong Province, China
³Department of Otorhinolaryngology, Hangzhou First People's Hospital, Hangzhou, Zhejiang Province, China

Corresponding authors: B. Wang / Y. Li
E-mail: yongajdshjajhsf@163.com

Received September 22, 2015
Accepted November 2, 2015
Published December 21, 2015
DOI http://dx.doi.org/10.4238/2015.December.21.10

ABSTRACT. In this study, bioinformatic analysis of gene expression data of head and neck squamous cell carcinoma (HNSCC) was performed to identify critical genes. Gene expression data of HNSCC were downloaded from the Cancer Genome Atlas (TCGA) and differentially expressed genes were determined through significance analysis of microarrays. Protein-protein interaction networks were constructed and used to identify hub genes. Functional enrichment analysis was performed with DAVID. Relevant microRNAs, transcription factors, and small molecule drugs were predicted by the Fisher exact test. Survival analysis was performed with the Kaplan-Meier plot from a package for survival analysis in R. In the five groups of HNSCC patients, a total of 5946 DEGs were identified in group 1, 4575 DEGs in group 2, 5580 DEGs in group 3, 8017 DEGs in group 4, and 5469 DEGs in group 5. DEGs in the cell cycle and immune response were significantly over-represented. Five PPI networks were constructed from...
which hub genes were acquired, such as minichromosome maintenance complex component 7 (MCM7), MCM2, decorin (DCN), retinoblastoma 1 (RB1), and tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein gamma (YWHAG). No significant difference in survival was observed among the 5 groups; however, a significant difference existed between two combined groups (groups 1, 3, and 5 vs groups 2 and 4). Our study revealed critical genes in HNSCC, which could supplement the knowledge about the pathogenesis of HNSCC and provide clues for future therapy development.

**Key words:** Head and neck squamous cell carcinoma; Survival curves; Differentially expressed genes; Protein-protein interaction network; Hub genes

**INTRODUCTION**

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common type of cancer worldwide, accounting for about 5% of all malignancies. Risk factors include nicotine and alcohol consumption and human papilloma virus (HPV) infection (Zur Hausen, 2002). Previous studies have identified molecular mechanisms contributing to the pathogenesis of HNSCC. HPV oncoproteins E6 and E7 are involved in tumorigenesis and are known to induce degradation of the tumor suppressors p53 and pRB, respectively (Pannone et al., 2011). Lack of p53 mutation (Westra et al., 2008) and p16 proteinover expression (Kumar et al., 2008) are also linked to HPV-related HNSCC. Increased epidermal growth factor receptor gene copy number is also associated with poor prognosis in HNSCC (Chung et al., 2006).

MicroRNAs (miRNAs) are implicated in the development of HNSCC. Nohata et al. (2011) found that tumor suppressive miRNA-375 regulates oncogene astrocyte elevated gene-1/metadherin (AEG-1/MTDH) in HNSCC (Nohata, 2011). MiR-133αdirectly regulates glutathione S-transferase P1 (GSTP1), a protein that suppresses cell apoptosis in HNSCC (Mutallip et al., 2011). Transcription factors (TFs) such as nuclear factor-κB (NF-κB), activator protein-1 (AP-1), and nuclear factor-interleukin-6 (NF-IL6) are constitutively active in human HNSCC cell lines (Ondrey et al., 1999). Inhibition of NF-κB can attenuate resistance of HNSCC to TNF-α caspase-mediated cell death (Duffey et al., 2000), suggesting the critical role of NF-κB. Sex determining region Y-box 2 (SOX2) (Freier et al., 2010) and protein C-ets-1 (Horvath et al., 2005) are also involved in the pathogenesis of HNSCC.

Despite these findings, the 5-year survival rate for HNSCC patients remains relatively unchanged at 50%. Therefore, more studies are necessary to better understand the pathogenesis of HNSCC and provide more efficient therapies. In the present study, we analyzed existing gene expression data with the latest bioinformatic tools with the aim of identifying key genes, miRNAs, and TFs in HNSCC.

**MATERIAL AND METHODS**

**Gene expression data**

Gene expression data were downloaded from the Cancer Genome Atlas (TCGA) using
TCGA-Assembler and package TCC was used for data normalization. A total of 541 samples were collected with 10629 genes, including 498 HNSCC samples and 43 normal controls. Missing values were filled with “1”, and then log2 transformation was applied. HNSCC patients were divided into groups using non-negative matrix transformation (NMF) (Rossini et al., 1994) based upon the cophenetic correlation coefficient (Farris, 1969). The K value was tested to generate a maximum cophenetic correlation coefficient.

**Screening of differentially expressed genes (DEGs)**

Significance analysis of microarrays (SAM) (Larsson et al., 2005) was adopted in differential analysis, which can reduce the false-positive rate (FDR) in the multiple testing. P < 0.05 and log (fold change) <1 were set as the cut-off to identify DEGs.

**Construction of protein-protein interaction networks**

Protein-protein interaction (PPI) networks were constructed with information from the human protein reference database (HPRD) (Keshava Prasad et al., 2009). Interactions with a Pearson correlation coefficient >0.5 were retained in the PPI networks. Gene regulatory networks were also constructed with information from cGRNB (http://www.scbit.org/cgmb/) (Tu et al., 2009; Yu et al., 2012), including both miRNA- and TF-target gene regulatory relationships. Relevant small molecules drugs were predicted using DrugBank (http://www.drugbank.ca/).

**Functional enrichment analysis**

Gene Ontology (GO)-enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway-enrichment analysis were performed with the database for annotation, visualization, and integrated discovery (DAVID) (Dennis et al., 2003).

**Screening of relevant miRNAs, TFs and drugs**

Using the databases described above, target genes of certain miRNAs, TFs, and drugs can be obtained; they were defined as $M_i$. The DEGs for each group of HNSCC patients were defined as $N$. Fisher’s exact test (Liao et al., 2006) was used to test the significance of the overlap between $M_i$ and $N$. $P < 0.05$ was set as the cut-off to identify relevant miRNAs, TFs and drugs for each group of DEGs.

The algorithm was as follows:

$$
p = \frac{(a+b)^2}{a^2 + c^2} / \frac{a}{a+c}
$$

where $a$ is the number of genes they belong to DEGs and the targets of miRNAs, TFs and drugs; $b$ is the number of genes they belong to the targets of miRNAs, TFs and drugs, but not DEGs; $c$ is the number of genes they belong to DEGs, but not the targets of miRNAs, TFs and drugs; $d$ is the number of genes they neither belong to DEGs, nor the targets of miRNAs.
Survival analysis

Survival analysis was performed for each group of HNSCC patients with a Kaplan-Meier plot using package survival of R.

RESULTS

HNSCC groups

The box plot of the pre-treated data is shown in Figure 1. A good performance of normalization was obtained. The coefficient of variation was calculated for each gene and the top 1500 genes were selected for grouping. The K value was set to “4” to generate the highest cophenetic correlation coefficient (Figure 2). Finally, HNSCC patients were divided into 5 groups. Group 1 had 100 samples, group 2 had 168 samples, group 3 had 106 samples, group 4 had 76 samples, and group 5 had 47 samples.

Figure 1. Box plot of the normalized gene expression data. The abscissa contains 100 samples randomly selected from all 498 HNSCC samples and the ordinate represents the gene expression levels of the genome. Good performance of normalization was achieved.
DEGs and over-represented functions

Package samr of R was used to identify DEGs in each group. A total of 5946 DEGs were identified in group 1, 4575 DEGs in group 2, 5580 DEGs in group 3, 8017 DEGs in group 4, and 5469 DEGs in group 5 (Figure 3). DEGs included in more than three groups were selected for further analysis. Functional enrichment analysis of the overlapping 6504 DEGs revealed that cell cycle and immune response were significantly over-represented (Table 1). In addition, a heat map was generated with the expression levels of the top 500 DEGs (sorted by P value) (Figure 4). Different expression patterns existed among the 5 groups.
**Table 1.** Top 15 pathways over-represented in more than 2 groups of differentially expressed genes.

<table>
<thead>
<tr>
<th>Source</th>
<th>Name</th>
<th>P value (Bonferroni corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>REACTOME</td>
<td>Cell cycle, Mitotic</td>
<td>1.10E-20</td>
</tr>
<tr>
<td>REACTOME</td>
<td>Cell cycle</td>
<td>4.24E-12</td>
</tr>
<tr>
<td>KEGG</td>
<td>Huntington’s disease</td>
<td>1.74E-11</td>
</tr>
<tr>
<td>REACTOME</td>
<td>M-phase</td>
<td>5.76E-10</td>
</tr>
<tr>
<td>REACTOME</td>
<td>Mitochondrial protein import</td>
<td>2.89E-09</td>
</tr>
<tr>
<td>REACTOME</td>
<td>Mitotic G1-G1/S phases</td>
<td>1.09E-08</td>
</tr>
<tr>
<td>KEGG</td>
<td>Ubiquitin-mediated proteolysis</td>
<td>1.17E-08</td>
</tr>
<tr>
<td>REACTOME</td>
<td>Cytokine signaling in immune system</td>
<td>1.41E-08</td>
</tr>
<tr>
<td>REACTOME</td>
<td>Mitotic metaphase and anaphase</td>
<td>2.75E-08</td>
</tr>
<tr>
<td>REACTOME</td>
<td>Mitotic anaphase</td>
<td>4.86E-08</td>
</tr>
<tr>
<td>REACTOME</td>
<td>S-phase</td>
<td>1.14E-07</td>
</tr>
<tr>
<td>REACTOME</td>
<td>Membrane trafficking</td>
<td>1.47E-07</td>
</tr>
<tr>
<td>KEGG</td>
<td>Non-alcoholic fatty liver disease (NAFLD)</td>
<td>2.03E-07</td>
</tr>
<tr>
<td>KEGG</td>
<td>HTLV-I infection</td>
<td>4.41E-07</td>
</tr>
<tr>
<td>REACTOME</td>
<td>Class I MHC-mediated antigen processing &amp; presentation</td>
<td>4.41E-07</td>
</tr>
</tbody>
</table>

**Figure 4.** Heat map of gene expression based upon the top 500 differentially expressed genes.

**Relevant miRNAs, TFs, and drugs**

Relevant miRNAs, TFs, and drugs were identified, with the number of them in all five groups as follows: miRNA, 33 (for group 1), 22 (for group 2), 52 (for group 3), 331 (for group 4), and 52 (for group 5); TF, 63 (for group 1), 40 (for group 2), 46 (for group 3), 58 (for group 4) and 55 (for group 5); and drugs 1 (Alendronate, for group 1), 0 (for group 2), 0 (for group 2), 0 (for group 4), and...
2 (Bortezomib, Flavin adenine dinucleotide, for group 5). The top 5 regulators (sorted by P value) are listed in Table 2. Sp1 transcription factor (SP1) and DAN domain family member 5 (DAND5) were common regulators in several groups while some TFs were unique to certain groups such as transcription factor activating protein 2α (AP-2α) for group 2 and E2F transcription factor (E2F) for group 5. Groups 1, 3, and 5 shared the same top 5 TFs, indicating high similarity between the groups. No common miRNAs and small molecule drugs were observed.

<table>
<thead>
<tr>
<th>miRNAs</th>
<th>Transcription factors</th>
<th>Drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1 miR-520a-5p, miR-1304,</td>
<td>SP1, DAND5, PSG1, ELK1,</td>
<td>Alendronate</td>
</tr>
<tr>
<td>miR-124, miR-1203, miR-583</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 2 miR-421, miR-569,</td>
<td>SP1, DAND5, PSG1, ELK1,</td>
<td>N/A</td>
</tr>
<tr>
<td>miR-33b, miR-520a-5p, miR-596</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 3 miR-613, miR-339-5p,</td>
<td>DAND5, PSG1, SP1, ELK1,</td>
<td>N/A</td>
</tr>
<tr>
<td>miR-605, miR-1251, miR-922</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 4 miR-548d-5p, miR-1283,</td>
<td>DAND5, PSG1, SP1, ELK1,</td>
<td>N/A</td>
</tr>
<tr>
<td>miR-548a-5, miR-548b-5p,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 5 miR-34b, miR-33-3p,</td>
<td>DAND5, PSG1, SP1, ELK1,</td>
<td>Bortezomib, Flavin adenine</td>
</tr>
<tr>
<td>miR-671-5p, miR-548c-3p</td>
<td></td>
<td>dinucleotide</td>
</tr>
</tbody>
</table>

**Table 2.** Top 5 relevant microRNAs, transcription factors, and drugs in each group.

PPI networks

One PPI network was constructed for each of the five groups: network 1 (for group 1) contained 520 interactions and 511 genes; network 2 (for group 2) contained 409 interactions and 394 genes; network 3 (for group 3) included 454 interactions and 485 genes; network 4 (for group 4) included 852 interactions and 893 genes; and network 5 (for group 5) contained 1416 interactions and 1209 genes. The degree was calculated for each node and the top 10 hub genes are listed in Table 3 for each PPI network. Some group-specific hub genes were identified, including minichromosome maintenance complex component 7 (MCM7), MCM2, decorin (DCN), retinoblastoma 1 (RB1), and tyrosine 3-monoxygenase/tryptophan 5-monoxygenase activation protein gamma (YWHAG).

<table>
<thead>
<tr>
<th>Genes</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDK1</td>
<td>21</td>
</tr>
<tr>
<td>PCNA</td>
<td>12</td>
</tr>
<tr>
<td>FN1</td>
<td>12</td>
</tr>
<tr>
<td>CDK2</td>
<td>12</td>
</tr>
<tr>
<td>FYN</td>
<td>12</td>
</tr>
<tr>
<td>EP300</td>
<td>11</td>
</tr>
<tr>
<td>MMP2</td>
<td>10</td>
</tr>
<tr>
<td>MCM7</td>
<td>9</td>
</tr>
<tr>
<td>COL1A1</td>
<td>9</td>
</tr>
<tr>
<td>CD4</td>
<td>8</td>
</tr>
</tbody>
</table>

**Table 3.** Top 10 hub genes in each protein-protein interaction network.

Survival analysis results

Information regarding survival time was collected from the TCGA and then analyzed by the Kaplan-Meier method to study the difference in survival time among the 5 groups. A total of 161 patients out of 497 had died, that is, 30, 62, 34, 24, and 11 patients from groups 1 to 5, respectively. No significant difference was observed among the 5 groups (P = 0.104). As shown...
in Figure 5A, the survival curves of groups 1, 3, and 5 were similar while those of groups 2 and 4 were also similar. This was in accordance with the TFs identified between the groups. Groups 1, 3, and 5 were then combined as m-group 1, while groups 2 and 4 were combined as m-group 2. A significant difference (P = 0.00793) in survival time between the two groups was then revealed by the log-rank test (Figure 5B).

**Figure 5.** Survival curves for A. all 5 groups and B. the two combined groups (m-group 1 = groups 1, 3, and 5 and m-group 2 = groups 2 and 4).

**DISCUSSION**

In the present study, a total of 541 HNSCC samples were collected, which were further divided into 5 groups. A total of 5946 DEGs were identified in group 1, 4575 DEGs in group 2, 5580 DEGs in group 3, 8017 DEGs in group 4, and 5469 DEGs in group 5. Functional enrichment analysis of the overlapping 6504 DEGs revealed that targets involved in cell cycle and immune response were significantly over-represented. Relevant miRNAs, TFs, and drugs were predicted by Fisher’s exact test. To further identify critical genes in the development of HNSCC, PPI networks were constructed and hub genes were identified. In addition, survival analysis was performed for the 5 groups of HNSCC patients.

According to the functional enrichment analysis results, genes in cell cycle and immune response were significantly over-represented in the DEGs. Cell cycle and immune response are closely associated with tumorigenesis, which is confirmed by our analysis. We identified several relevant miRNAs and TFs, several of which have been shown to be involved in HNSCC. MiR-124 acts as a tumor suppressor gene by inhibiting expression of integrin beta-1 (Perez-Sayans et al., 2012) and it has been found to be downregulated in HNSCC (Nohata et al., 2013). Aberrant overexpression of miR-421 downregulates the expression of ataxia-telangiectasia mutated protein (ATM) and causes clinically manifested tumor radio sensitivity (Mansour et al., 2013). It has been reported that paired box (PAX) is inactivated by cancer-specific promoter methylation (Guerrero-Preston et al., 2014). Printz et al. (2014) indicated that HNSCC patients with greater PAX5 methylation and p53 mutations have worse overall survival. Zhang et al. (2000) found that
overexpression of E2F transcription factor 1 (E2F-1) plays a positive role in cell cycle re-entry from quiescence and is associated with increased invasiveness in vivo. However, Kwong et al. (2003) reported that overexpression of E2F-1 is associated with increased disease-free survival in squamous cell carcinoma of the anterior tongue. In addition, ETS domain-containing protein 1 (ELK1) is involved in the antiangiogenic and antitumor activities of insulin-like growth factor binding protein-3 (IGFBP-3) (Kim et al., 2011). Further research on these miRNAs and TFs may provide important clues to uncover the molecular mechanisms underlying HNSCC.

Critical genes were also revealed by PPI network analysis. E1A binding protein p300 (EP300) is a histone acetyltransferase that regulates transcription via chromatin remodeling and is important in the processes of cell proliferation and differentiation. Alteration in EP300 expression is much more prominent in HNSCC lines (Martin et al., 2014) and we speculate it might play a role in the development of HNSCC. Matrix metalloproteinase 2 (MMP2) is capable of cleaving components of the extracellular matrix and is thus involved in signal transduction. Previous studies have demonstrated that high levels of MMPs are expressed during HNSCC in vivo and that inhibition of these enzymes in vitro and in mouse models decreases invasion and metastasis (Rosenthal and Matrisian, 2006). It has been reported that genetic polymorphisms in the promoters of MMP2 and TIMP metalloproteinase inhibitor 2 (TIMP2) may be associated with the development and aggressiveness of HNSCC (O-Charoenrat and Khantapura, 2006). Osteonectin (SPARC) is a cysteine-rich acidic matrix-associated protein that is associated with tumor suppression but has also been correlated with metastasis. Chin et al. (2005) suggest that SPARC is an independent prognostic marker in HNSCC. Differential alterations of the genes in the CDKN2A-CCND1-CDK4-RB pathway are associated with the development of HNSCC (Tripathi Bhar et al., 2003). Hypermethylation of RB1 has also been reported in HNSCC (De Carvalho et al., 2007). We speculate that other hub genes might be potential biomarkers or play important roles in the development of HNSCC.

Overall, a number of critical miRNAs, TFs, and genes were identified in HNSCC through bioinformatic analysis of gene expression data. These findings could improve the understanding about the pathogenesis of HNSCC and thus benefit therapy development.

Conflicts of interest

The authors declare no conflict of interest.

REFERENCES


