Mining Proteins Associated with Oral Squamous Cell Carcinoma in Complex Networks

Ying Liu¹,²&, Chuan-Xia Liu¹&, Zhong-Ting Wu¹&, Lin Ge¹,³*, Hong-Mei Zhou³*

Abstract

The purpose of this study was to construct a protein-protein interaction (PPI) network related to oral squamous cell carcinoma (OSCC). Each protein was ranked and those most associated with OSCC were mined within the network. First, OSCC-related genes were retrieved from the Online Mendelian Inheritance in Man (OMIM) database. Then they were mapped to their protein identifiers and a seed set of proteins was built. The seed proteins were expanded using the nearest neighbor expansion method to construct a PPI network through the Online Predicated Human Interaction Database (OPHID). The network was verified to be statistically significant, the score of each protein was evaluated by algorithm, then the OSCC-related proteins were ranked. 38 OSCC related seed proteins were expanded to 750 protein pairs. A protein-protein interaction network was then constructed and the 30 top-ranked proteins listed. The four highest-scoring seed proteins were SMAD4, CTNNB1, HRAS, NOTCH1, and four non-seed proteins P53, EP300, SMAD3, SRC were mined using the nearest neighbor expansion method. The methods shown here may facilitate the discovery of important OSCC proteins and guide medical researchers in further pertinent studies.

Keywords: Protein-protein interaction - nearest neighbor expansion - oral squamous cell carcinoma

Introduction

Oral squamous cell carcinoma (OSCC) is a major healthcare problem. It includes approximately 90% of oral malignancies and accounts for more than 300,000 of newly diagnosed cancers every year. Although significant progress has been made in cancer treatment, the death rate associated with OSCC remains unchanged, and the overall 5-year survival rate is estimated at about 50% (Choi and Myers, 2008; Pasini et al., 2012). Identifying high risk factors may facilitate early diagnosis, treatment and lower the incidence of OSCC. Proteins are the final executants of physical functions, and play the key role in the development of cancer. Traditional research methods only focus on individual proteins. However, a better understanding of protein-protein interactions is crucial to investigating their roles in cancer development and identifying potential drug targets for use in clinical applications (Bonetta, 2010). Researchers need a network that can describe a large number of protein interactions clearly and explain the mutual influences on structures and functions. With the help of high-throughput screening technologies and computational models, information can be integrated and PPI networks can be constructed. The PPI networks might help researchers determine the best candidates for assessing disease risks and identify therapeutic targets. The purpose of the present work was to construct the OSCC-related PPI network and mine the important OSCC proteins using specific bioinformatic tools and theories.

Materials and Methods

Collection of OSCC related genes and proteins

The content search for “oral squamous cell carcinoma” was performed in OMIM, produced a list of OSCC-related 42 genes records (Oti et al., 2011). The list was subjected to the search tool in HUGO Gene Nomenclature Committee (HGNC) in order to identify the exact identifiers. HGNC stores all confirmed human genes and each gene receives exactly one unique standard gene identity (Seal et al., 2011). The content search tool in HUGO Gene Nomenclature Committee (HGNC) in order to identify the exact identifiers.

HGNC stores all confirmed human genes and each gene receives exactly one unique standard gene identity (Seal et al., 2011). The content search tool in HUGO Gene Nomenclature Committee (HGNC) in order to identify the exact identifiers.

Expansion of OSCC related protein-protein interactions

I. First neighbor of the seed proteins were found using
From two columns of data covering the 750 pairs, 1908 protein interaction pairs were acquired using the total number proteins of sub-network and entire network. The specificity of this network was tested to prove that all these proteins were interacted really rather than randomly. The same number of protein pairs were selected randomly for 1000 times to calculate the p-value and to generate distribution of the index of aggregation for further calculation (Wagner and Fell, 2001; Maslov and Sneppen, 2002). Finally, we verified if the degree centrality distribution of all the proteins obey the power law.

**Evaluation of the contributions of each protein**

The role of a protein in the network can be qualitatively evaluated. The ability to connect with other protein partners with high specificity reflects the contribution of a protein to the network which was calculated using the following formula

$$S_i = 2 \times \ln(t(i) \times 0.9) - \ln(t(i)) \quad \text{Eq.1}$$

where $t(i)$ indicates the number of connections of a given protein i. 0.9 is the fixed coefficient, which has been verified for protein interactions through real human protein interaction experiments. These interactions are assigned a high confidence score of 0.9 (Chen, 2006).

## Results

**OSCC-related genes and proteins**

42 gene records were collected from the OMIM database. 38 seed proteins were retrieved from HGNC (Table 1). 1908 protein interaction pairs were acquired and only 750 PPI pairs were accepted. The details are available in the supplementary material.

**Visualization of PPI network**

From two columns of data covering the 750 pairs, Pajek produced the graph shown in Figure 1. The entire network contained several clusters of different scales, in which the number of involved proteins ranged from 2 to 626. The largest sub-network contained 626 proteins and

### Table 1. Forty-two OSCC-related genes and mapped to their protein identifiers

| Num. | Input | UniCode | Approv | HGNC ID | Location |
|------|-------|---------|--------|---------|----------|
| 1    | TNFRSF10B | O14763 | TNFRSF10B | HGNC:11905 | 8p22-p21 |
| 2    | PTEN   | P04084 | PTEN   | HGNC:9588 | 10q23   |
| 3    | ING1   | Q9UK53 | ING1   | HGNC:6062 | 13q34   |
| 4    | TGFBR2 | P37173 | TGFBR2 | HGNC:11773 | 3p22   |
| 5    | DLEC1  | Q9Y238 | DLEC1  | HGNC:2809 | 3p13   |
| 6    | LZTS1  | Q9Y250 | LZTS1  | HGNC:13861 | 8p22 |
| 7    | Dec1   | Q9P2X7 | Dec1   | HGNC:23658 | 9q32   |
| 8    | RNF5   | Q9Y252 | RNF5   | HGNC:10669 | 13q12   |
| 9    | WWOX   | Q9NZC7 | WWOX   | HGNC:12799 | 16q23.3-q24.1 |
| 10   | CDK2NA | P42771 | CDK2NA | HGNC:1787 | 9p21 |
| 11   | NOTCH1 | P46531 | NOTCH1 | HGNC:7881 | 9q34.3 |
| 12   | SMAD4  | Q34855 | SMAD4  | HGNC:6770 | 18q11.1 |
| 13   | CTNNB1 | P35222 | CTNNB1 | HGNC:2514 | 3p21 |
| 14   | ORA0V1 | Q8WV07 | ORA0V1 | HGNC:17589 | 11q13.2 |
| 15   | SHH    | Q15465 | SHH    | HGNC:10848 | 7q36 |
| 16   | STK11  | Q15831 | STK11  | HGNC:11389 | 19p13.3 |
| 17   | FHT    | P49789 | FHIT   | HGNC:3701 | 3p14.2 |
| 18   | XPC    | Q01831 | XPC    | HGNC:12816 | 3p25 |
| 19   | COL7A1 | Q02388 | COL7A1 | HGNC:2214 | 3p21.1 |
| 20   | MMP1   | P03966 | MMP1   | HGNC:7155 | 11q13.1-q22 |
| 21   | DCK1   | Q90832 | DCK1   | HGNC:2890 | Xq28 |
| 22   | GJB2   | P29033 | GJB2   | HGNC:4284 | 13q11.1-q12 |
| 23   | XPA    | P23025 | XPA    | HGNC:12814 | 4q23.3 |
| 24   | ENOSF1 | P16781 | ENOSF1 | HGNC:3065 | 18q11.32 |
| 25   | KRT5   | P13647 | KRT5   | HGNC:6442 | 12q13.13 |
| 26   | HRAS   | P10112 | HRAS   | HGNC:5173 | 11p15.5 |
| 27   | CEP55  | Q53EZA | CEP55  | HGNC:1161 | 10q24.1 |
| 28   | SERPINB13 | Q9URV | SERPINB13 | HGNC:8944 | 11q21.3-q22 |
| 29   | WRAP53 | Q9BU4R | WRAP53 | HGNC:2522 | 17p13.1 |
| 30   | CTSC   | P53634 | CTSC   | HGNC:2528 | 17q12-q21 |
| 31   | KRT14  | P02533 | KRT14  | HGNC:6416 | 17q12-q21 |
| 32   | CSDM1  | Q60977 | CSDM1  | HGNC:14626 | 9p22.2 |
| 33   | IPK2   | Q9UI9H | IPK2   | HGNC:17313 | 3p21.31 |
| 34   | LIN7C  | Q9NU9P | LIN7C  | HGNC:17769 | 11p14 |
| 35   | NOL3A3 | Q9NEP3 | NOL3A3 | HGNC:14378 | 15q14.1-q15 |
| 36   | CXL14  | Q95715 | CXL14  | HGNC:10640 | 5q31 |
| 37   | GPRC5A | Q8NF35 | GPRC5A | HGNC:9836 | 12p13.1-q23 |
| 38   | GJB6   | O95452 | GJB6   | HGNC:4288 | 13q12 |
| 39   | MSSE   | Matches | MSSE | HGNC:7377 | 9q22.32 |
| 40   | CMM    | Matches | CMM | HGNC:2124 | 1p36 |
| 41   | TOC    | Matches | TOC | HGNC:11981 | 17q25.1 |
| 42   | TERC   | Matches | TERC | HGNC:11727 | 3q26.2 |

---

42 OSCC-related genes were retrieved from OMIM database and then be mapped to HGNC database to determine their protein identifiers. Four genes were not matched to their identifiers.
the index of aggregation was 93.43%. Repeating random selection of the same number of protein pairs for 1000 times made index of aggregation greater than 93.43% for only 5 times. It indicated that the p-value was 0.005 and the OSCC-related PPI was statistically significant and only 5 times made index of aggregation greater than 93.43% for 1000 times under random selection of the same number of protein pairs. These proteins were directly from the OMIM. It is indicated that they had already been verified in previous studies. Four proteins (P53, EP300, SMAD3, SRC) were not included in the seed set, so, they were not initially retrieved from the OMIM.

Degree centrality is defined as the number of links of each node was plotted and fitted, producing a curve consistent with the power law. (B) Logarithmic transformation applied to the X axle transformed the curve into a line. This indicates that the fitting curve is consistent with the power law (Ning et al., 2010). Therefore, it suggested that the distribution of degree centraliy obeys the power law. Maslov had certified that the protein interaction network was consistent with the power law distribution model (Maslov and Sneppen, 2002). If some interaction network was consistent with the power law, then the proteins connected with each other biologically rather than randomly.

We evaluated the contribution of each protein in the network. YES and NO indicate whether or not the protein belonged to the seed set. The distribution of the index of aggregation was 93.43%. Repeating random selection of the same number of protein pairs for 1000 times made index of aggregation greater than 93.43% for only 5 times. It indicated that the p-value was 0.005 and the OSCC-related PPI was statistically significant and specific. The distribution of the index of aggregations is shown in Figure 2.

Degree centrality is defined as the number of links incident upon a node that is widely used in network analysis. There are two charts show information about degree centrality of the OSCC-related network (Figure 3). In Figure 3A, degree centrality are shown by the X-axis, whereas the Y-axis shows the counts of the correspond proteins. In Figure 3B, X-axis and Y-axis are transformed by log function, the curve fitting result prove that the distribution of degree centraliy obeys the power law. Maslov had certified that the protein interaction network was consistent with the power law distribution model (Maslov and Sneppen, 2002). If some proteins had connected randomly, the degree centrality distribution of the network would not have obeyed the power law (Ning et al., 2010). Therefore, it suggested that the proteins connected with each other biologically rather than randomly.

### Table 2. Top Thirty Rank-ordered OSCC-related Proteins

| Index | Score  | Pro. Name | Unip Id | Int Pairs | Seed |
|-------|--------|-----------|---------|-----------|------|
| 1     | 4.75221| SMAD4     | P13485  | 181       | YES  |
| 2     | 4.550925| CTNNB1   | P35222  | 148       | YES  |
| 3     | 3.787819| HRAS     | P01112  | 69        | YES  |
| 4     | 3.360375| NOTCH1   | P46531  | 45        | YES  |
| 5     | 3.080073| TGFBRII | P37173  | 34        | YES  |
| 6     | 2.95491| WWOX     | Q9mxz7  | 30        | YES  |
| 7     | 2.849555| PTEN     | P60484  | 27        | YES  |
| 8     | 2.731767| ING1     | P15831  | 24        | YES  |
| 9     | 2.731767| STK1    | P15831  | 24        | YES  |
| 10    | 2.326302| MMP1     | P303956 | 16        | YES  |
| 11    | 2.19277| CEAP5    | Q53ec4  | 14        | YES  |
| 12    | 2.19277| COL7A1   | P02388  | 14        | YES  |
| 13    | 2.118662| CDKN2A   | P42771  | 13        | YES  |
| 14    | 1.951608| XPC      | P15831  | 11        | YES  |
| 15    | 1.856298| XPA     | P23025  | 10        | YES  |
| 16    | 1.750937| KRT5     | P13647  | 9         | YES  |
| 17    | 1.633154| TNFRSF10B| O14763  | 8         | YES  |
| 18    | 1.633154| SERPINB13| P9uv8   | 8         | YES  |
| 19    | 1.163151| LIN7C     | P15831  | 5         | YES  |
| 20    | 1.163151| DKC1     | O60832  | 5         | YES  |
| 21    | 1.163151| FHT      | P94789  | 5         | YES  |
| 22    | 1.163151| SHH      | Q15465  | 5         | YES  |
| 23    | 0.940007| KRT14    | O2533   | 4         | YES  |
| 24    | 0.940007| P53      | P04637  | 4         | NO   |
| 25    | 0.940007| EP300    | Q09472  | 4         | NO   |
| 26    | 0.940007| SMAD3    | P12931  | 4         | NO   |
| 27    | 0.940007| SRC      | Q9uv8   | 4         | NO   |
| 28    | 0.652325| CTSC     | P53634  | 3         | YES  |
| 29    | 0.652325| LATS1    | P9y250  | 3         | YES  |
| 30    | 0.652325| NR3C4    | P10275  | 3         | YES  |

We evaluated the contribution of each proteins in the network. YES and NO indicate whether or not the protein belonged to the seed set.
Furthermore, it contributes to adherens junctions through family and plays an important role in Wnt signaling. al., 2013). CTNNB1 (β-catenin) belongs to the armadillo family and is called co-Smad. SMAD4 and the R-SMADs play a critical role in the suppression of transcriptional responses of TGF-β signaling pathway. In this way, SMAD4 plays a critical role in the suppression of carcinogenesis and maintenance of tissue homeostasis. The not activation of the SRC gene, this protein phosphorylates specific tyrosine residues of other proteins and the activation of P53 acts as tumor suppressor and the activation of P53 can initiate responses such as DNA repair, differentiation, senescence and the inhibition of angiogenesis (Mroz and Rocco, 2010; Pasini et al., 2012). EP300, a transcriptional coactivator, promotes maturation and differentiation of cells and prevents the growth of cancer. Studies suggest that EP300 mutations contribute to the development of colon cancer, breast cancer and OSCC. It may also help predict cancer prognosis (Gayther et al., 2000). SMAD3, a mediator of TGF-β signaling pathway, can combine with SMAD4 to activate the pathway. SMAD3 may have a bidirectional function in cancer development (Han and Wan, 2011). SRC is a proto-oncogene tyrosine-protein kinase encoded by the SRC gene, this protein phosphorylates specific tyrosine residues of other proteins and the activation promotes angiogenesis, proliferation and invasion of cancer (Cheng et al., 2011).

In summary, this work describes the construction of a protein-protein interaction network of OSCC. The Four highest-scoring proteins SMAD4, CTNNB1, HRAS, NOTCH1 were identified, and four non-seed proteins P53, EP300, SMAD3 and SRC were mined using the nearest neighbor expansion method. These proteins affect the development and metastasis of OSCC through regulation of transcriptional responses, differentiation, angiogenesis, proliferation, and apoptotic programs. The present study may help researchers identify crucial targets for the prevention and treatment of OSCC and guide medical research toward further pertinent study.

Discussion

The present study screened the proteins which come from the human protein interaction experiments database. This screening method could diminish the influence of the interference factors and uncertain factors and could help to evaluate the contribution of proteins more accurately. The proteins were integrated and analyzed to construct the OSCC-related protein interaction network, which contributes to more comprehensive and systematic research. The nearest neighbor expansion method not only validated existing OSCC protein targets but also mined ones absent in the initial seed set of OSCC protein targets. The specificity and the reliability of the PPI network were tested to be fine. The important candidates for assessing OSCC risk and therapeutic targets were mined. The recommended research method may also help to screen other target molecules for further study of OSCC.

The four highest-scoring proteins (SMAD4, CTNNB1, HRAS, NOTCH1) were proposed as the most important candidates for assessing OSCC risks and therapeutic targets. And they had been confirmed to play an important role in the occurrence and development of OSCC. SMAD4 protein plays the role of common-mediator in the Smad family and is called co-Smad. SMAD4 and the R-SMADs complex can target DNA binding proteins to promote transcriptional responses of TGF-β signaling pathway. In this way, SMAD4 plays a critical role in the suppression of carcinogenesis and maintenance of tissue homeostasis. The loss of expression may promote the development and metastasis of OSCC (Yang and Yang, 2010; Xia et al., 2013). CTNNB1 (β-catenin) belongs to the armadillo family and plays an important role in Wnt signaling. Furthermore, it contributes to adherens junctions through protein-protein binding and regulates E-cadherin-mediated cell-cell adhesion. Abnormal expression of CTNNB1 can impact on oral cancer cell behavior (Duan et al., 2006; Leel et al., 2010). HRAS, a GTPase, has been proven to be a proto-oncogene and overaction drives the cells to uncontrolled division and thus carcinogenesis. The variant ‘C’ allele of the H-RAS T81C was founded to be associated with higher risk of oral cancer (Murugan et al., 2009; Jayaraman et al., 2012). NOTCH1, a transmembrane protein with repeated extracellular EGF domains and the NOTCH domains, works in multiple processes such as differentiation, proliferation and apoptosis. Overactivated Notch1 signaling facilitates tumor recurrence and drug resistance of cancer stem cell and cancer stem-like cells. However, activated NOTCH1 can increase the expression of p21WAF1/CIP1 and P53 and trigger down-regulate Wnt/β-catenin signaling, which can induce OSCC cells apoptosis and cell cycle arrest (Duan et al., 2006; Ravindran and Devaraj, 2012).

Four proteins, P53, EP300, SMAD3 and SRC, were mined using the nearest neighbor expansion method. However, these proteins were not included in the seed set, they were all found to interact with important seed proteins (Table 3). In this way, it was indirectly proven that they might play an important role in OSCC. These proteins meritare further research. P53 acts as tumor suppressor and the activation of P53 can initiate responses such as DNA repair, differentiation, senescence and the inhibition of angiogenesis (Mroz and Rocco, 2010; Pasini et al., 2012). EP300, a transcriptional coactivator, promotes maturation and differentiation of cells and prevents the growth of cancer. Studies suggest that EP300 mutations contribute to the development of colon cancer, breast cancer and OSCC. It may also help predict cancer prognosis (Gayther et al., 2000). SMAD3, a mediator of TGF-β signaling pathway, can combine with SMAD4 to activate the pathway. SMAD3 may have a bidirectional function in cancer development (Han and Wan, 2011).

In summary, this work describes the construction of a protein-protein interaction network of OSCC. The Four highest-scoring proteins SMAD4, CTNNB1, HRAS, NOTCH1 were identified, and four non-seed proteins P53, EP300, SMAD3 and SRC were mined using the nearest neighbor expansion method. These proteins affect the development and metastasis of OSCC through regulation of transcriptional responses, differentiation, angiogenesis, proliferation, and apoptotic programs. The present study may help researchers identify crucial targets for the prevention and treatment of OSCC and guide medical research toward further pertinent study.

Acknowledgements

This study was supported by grants from the National Natural Science Foundation of China (No. 81172581, No. 81102063 and No. 81272962). The author(s) declare that they have no competing interests.
References

Batagelj V, Mrvar A (2011). Pajek program for analysis and visualization of large networks. Reference Manual, University of Ljubljana, Slovenia.

Bonetta L. (2010). Protein-protein interactions: interactome under construction. Nature, 468, 851-4.

Chen JY, Shen C, Sivachenko AY (2006). Mining alzheimer disease relevant proteins from integrated protein interactome data. Pac Symp Biocomput, 11, 367-78.

Cheng SJ, Kok SH, Lee JJ (2011). Significant association of SRC protein expression with the progression, recurrence, and prognosis of oral squamous cell carcinoma in taiwan. Head Neck, 34, 1340-5.

Choi S, Myers JN (2008). Molecular pathogenesis of oral squamous cell carcinoma: implications for therapy. J Dent Res, 87, 14-32.

Duan L, Yao J, Wu XX, Fan MW (2006). Growth suppression induced by Notch1 activation involves Wnt-β-catenin down-regulation in human tongue carcinoma cells. Biol Cell, 98, 479-90.

Gayther SA, Batley SJ, Linger L, et al (2000). Mutations truncating the EP300 acetylase in human cancers. Nat Genet, 24, 300-3.

Han GW, Wan XJ (2011). Roles of TGFβ signaling smads in squamous cell carcinoma. Cell Biosci, 2011, 41, 1-8.

Jayaraman B, Valiathan GM, Jayakumar K, et al (2012). Lack of mutation in p53 and H-ras genes in phenytoin induced gingival overgrowth suggests its non cancerous nature. Asian Pac J Cancer Prev, 13, 5535-8.

Lee1 CH, Hung HW, Hung PH, Shieh YS (2010). Epidermal growth factor receptor regulates β-catenin location, stability, and transcriptional activity in oral cancer. Mol Cancer, 9, 1-12.

Maslov S, Sneppen K (2002). Specificity and stability in topology of protein networks. Science, 296, 910-3.

Mroz EA, Rocco JW (2010). Functional P53 status as a biomarker for chemotherapy response in oral-cavity cancer. J Clin Oncol, 28, 715-7.

Murugan AK, Hong NT, Cuc TT, et al (2009). Detection of two novel mutations and relatively high incidence of H-RAS mutations in vietnamese oral cancer. Oral Oncol, 45, 161-6.

Oti M, Ballouz S, Wouters MA (2011). Web tools for the prioritization of candidate disease genes. Methods Mol Biol, 760, 189-206.

Pasini FS, Maistro S, Smitcovsky I (2012). Four-gene expression model predictive of lymph node metastases in oral squamous cell carcinoma. Acta Oncol, 51, 77-85.

Ravindran G, Devaraj H (2012). Aberrant expression of β-catenin and its association with ANp63, Notch-1, and clinicopathological factors in oral squamous cell carcinoma. Clin Oral Investig, 16, 1275-88.

Seal RL, Gordon SM, Lush MJ, et al (2011). Genenames Org: The HGNC resources in 2011. Nucleic Acids Res, 39, 514-9.

Wagner A, Fell DA (2001). The small world inside large metabolic networks. Proceedings of the royal society of London. Int J Biol Sci, 268, 1803-10.

Xia RH, Song XM, Wang XJ, et al (2013). The combination of SMAD4 expression and histological grade of dysplasia is a better predictor for the malignant transformation of oral leukoplakia. PLoS One, 8, 1-6.

Yang G, Yang X (2010). Smad4-mediated TGF-β signaling in tumorigenesis. Int J Biol Sci, 6, 1-8.