Combined genomic expressions as a diagnostic factor for oral squamous cell carcinoma

Ki-Yeol Kim a, Xianglan Zhang b, In-Ho Cha a,c,⁎

a Oral Cancer Research Institute, College of Dentistry, Yonsei University, Seoul 120-752, Republic of Korea
b Department of Pathology, Yanbian University Medical College, Yanji City, Jilin Province, China
c Department of Oral and Maxillofacial Surgery, College of Dentistry, Yonsei University, Seoul 120-752, Republic of Korea

Abstract

Trends in genetics are transforming in order to identify differential coexpressions of correlated gene expression rather than the significant individual gene. Moreover, it is known that a combined biomarker pattern improves the discrimination of a specific cancer. The identification of the combined biomarker is also necessary for the early detection of invasive oral squamous cell carcinoma (OSCC). To identify the combined biomarker that could improve the discrimination of OSCC, we explored an appropriate number of genes in a combined gene set in order to attain the highest level of accuracy. After detecting a significant gene set, including the pre-defined number of genes, a combined expression was identified using the weights of genes in a gene set. We used the Principal Component Analysis (PCA) for the weight calculation. In this process, we used three public microarray datasets. One dataset was used for identifying the combined biomarker, and the other two datasets were used for validation. The discrimination accuracy was measured by the out-of-bag (OOB) error. There was no relation between the significance and the discrimination accuracy in each individual gene. The identified gene set included both significant and insignificant genes. One of the most significant gene sets in the classification of normal and OSCC included MMP1, SOCS3 and ACOX1. Furthermore, in the case of oral dysplasia and OSCC discrimination, two combined biomarkers were identified. The combined expression revealed good performance in the validation datasets. The combined genomic expression achieved better performance in the discrimination of different conditions than a single significant gene. Therefore, it could be expected that accurate diagnosis for cancer could be possible with a combined biomarker.

© 2013 Published by Elsevier Inc.

1. Introduction

It is known that about 20% of oral dysplasia undergoes malignant transformation to OSCC [1–3] and the local recurrence rate in OSCC patients, with histologic positive tumor margins, is as high as 70% to 80%. It has been reported that the probability of the recurrence is 30% to 40% even in patients with negative margins [4]. This indicated that the histologic examination alone is inadequate in predicting the progression to cancer and its recurrence [3–5]. Therefore, it is necessary to identify better ways to predict which patients with oral dysplasia will develop OSCC and which patients treated for OSCC will develop recurrence, such that high risk patients can be more rigorously treated.

Gene expression profiling has been widely used for cancer research. Studies with gene expressions are typically performed by comparing the gene expression levels between diseased and healthy tissues. They are normally conducted by testing the statistical significant changes in the mean level of expression of each individual gene [6–8]. With various statistical techniques, gene expression patterns have been explored in many types of cancer. Additionally, most microarray analyses on cancer have focused on the comparison of tumor and normal tissues; moreover, genes have been treated individually and the interaction among them has not been much considered [9]. Therefore, much of the information contained in gene expression datasets could be ignored by dealing with genes individually, although the differential expression approaches have been very successful.

Biologically known genes are often not differently expressed in diseases because the mutations in the coding region can affect the function of the gene without affecting its expression level [8]. This indicates that a statistically insignificant gene, which is not differently expressed between two different phenotypic groups, can be biologically meaningful and thus, it is not reliable to detect a biomarker by significance of an individual gene. Further, genetic trends have been changing from differential expression to differential coexpression, and from differential coexpression to differential networking [8,10]. Hence, a set-wise differential coexpressed analysis can be useful for understanding the biological process [11].

Previous studies have reported that a combined biomarker pattern improved the discrimination of disease as well as the sensitivity...
and specificity for cancer diagnosis compared with a single gene [12,13]. However, these studies considered only the genes which were known to be related with a disease, when combined. However, Chen et al. (2008) identified discriminative gene sets using the statistical method and showed that the combined gene set achieved high accuracy in predicting the different phenotypes [3]. They also considered only the significant individual genes for identifying a combined gene set. However, the individually insignificant genes may be influential as a member of a gene set. Hence, the insignificant genes should be considered in the identification of a significant gene set, since the combined gene set including these insignificant genes could be more discriminative and more biologically meaningful compared to a gene set with only individually significant genes. Therefore, it is important to identify a coexpressed gene set in addition to the common differential mean expression testing. In this study, we investigated the discriminative gene sets between OSCC and normal, OSCC and oral dysplasia, and also identified the combined gene expressions from the selected gene sets. The combined gene expressions were evaluated in publicly available two microarray datasets.

2. Results

2.1. The association of the significance and prediction accuracy in a single gene

We explored the association of the significance and the prediction accuracy of an individual gene (Figs. 1A and B). We used OOB error rate as a measure of prediction accuracy. Here, we calculated the significance and OOB error rates in classifying normal and OSCC patients. Figs. 1A and B showed that there was no association of significance or a prediction accuracy of a single gene. It indicated that all of the significant genes are not discriminative. Therefore, it is necessary to identify a gene set, including significant and insignificant genes, in order to maximize the prediction accuracy.

2.2. The association of the significance and prediction accuracy in a gene set

In order to explore the association of significance and OOB error rates of genes in a gene set, we selected gene sets with 3 genes 10,000 times repeatedly and plotted out their relation of genes in a gene set (Fig. 1C). Fig. 1C illustrated that insignificant genes can achieve high prediction accuracy and low OOB error rates (gray rectangle in Fig. 1C). Therefore, it indicated that the combination of genes, including such insignificant genes, can be more appropriate as a biomarker; a combined biomarker.

2.3. Appropriate number of genes to be combined for a biomarker

To identify a combined expression of genes, it is necessary to decide how many genes should be combined. We simulated gene sets with size 1 to 5, 10,000 times repeatedly, and explored the association of the OOB error rates and the number of genes (Fig. 2).

When gene set size was 1, the OOB error rates were same to those in Fig. 1A. The OOB error rates were decreased as the size of the gene set was increased (Fig. 2). We decided to use 3 as the gene set size, since
The selected genes by the OOB error rate were identified by their individual predictability. Moreover, the selected genes by significance were identified by independent two samples t-test. We used whole probes in this stage, which are the not corrected duplicated gene IDs. MMP1 and FAP were the most predictable and the most significant genes in classifying normal and OSCC groups (Supplementary Table 1). PTPRJ and EVA1A were the most predictable and the most significant genes in classifying oral dysplasia and OSCC groups (Supplementary Table 2). We identified the combined biomarkers with these two seed genes.

2.5. Identified gene sets and the validation of the combined expressions

The significant gene sets were identified with seed genes, and the weights of each gene were calculated by PCA. The sensitivity, specificity, and accuracy of each gene set were summarized in Table 1.

After identifying a significant gene set with three genes, the combined expression was calculated using weights of genes from PCA. The gene set with MMP1, RUNX2 and MTERFD2 had the most discriminative power to separate OSCC from normal (accuracy = 0.981). Further, the gene set with EVA1A, NEK6 and KLHL8 had the most discriminative power to separate oral dysplasia from OSCC (accuracy = 0.924). In the case of classification of OSCC and oral dysplasia, oral dysplasia groups tended to allocate to OSCC. Therefore, the specificity was low despite the high accuracy.

We compared the performance of the identified gene sets with that of the previous study (Table 2).

We compared the performance of two combined expressions identified by Chen et al. [3], with our three combined expressions, in two datasets, GSE30784 and GSE9844. In both datasets, the three combined expressions, which were identified in our study, demonstrated better performance compared to the previous study, particularly in GSE9844.

![Image](image-url)
We explored the three combined expressions in GSE6791 [14] (Fig. 3). The combined expressions were clearly differentiated between normal and cancer groups. The combined expressions of MMP1, SOCS3, ACOX1 and FAP, MTAP and C10orf128 were up-regulated in cancer. Moreover, the combined expressions of MMP1, RUNX2 and MTERFD2 were down-regulated in cancer compared to the normal group. The combined expressions in cancer patient groups were distributed more extensively compared to those in the normal group.

We explored the two combined expressions in classifying oral dysplasia and cancer in GSE30784 (Fig. 4). The significance and predictability in individual gene were not related in oral dysplasia and cancer classification. Further, predictability of most genes was about 50%, even when their significances were strong (data not shown). However, the predictability was increased as two more genes were combined (Fig. 4).

The OOB error rates, which were shown in Fig. 4, were the mean values of OOB error rates calculated from 100 times oral dysplasia and cancer classification. The reason for the 100 repetitions is that the size of these two groups was very unbalanced; we randomly sampled 17 cancer cases 100 times for a more accurate comparison. Fig. 4 illustrated that the combined expressions were more predictive compared to single genes. It indicated that the combined expressions could be more appropriate as a biomarker.

3. Conclusion and discussion

There can be two approaches to the analysis of gene expression data collected in microarray studies. The first identifies genes that show significantly different expressions between conditions. The second is to identify the patterns of coexpressed gene expressions [11,15]. Identifying an individually significant gene expression might even lead to incorrect conclusions about the involvement of particular pathways in disease conditions [8,16]. By the way, different coexpression provides information that would be missed using the classical methods, which focus on the identification of differently expressed genes, and may be engendered by different biological mechanisms [15]. Therefore, it is vital to perform differential coexpression analyses in addition to the common differential mean expression test.

Previous studies have reported various methods for the identification of coexpressed genes and gene network [6,8,10,11,15]; further, some studies reported that the combined biomarkers improved the predictability of specific cancers [3,12,13].

In this study, we detected coexpressed genes and identified the combined biomarker of these coexpressed genes using PCA. Nine coexpressed gene sets were identified, which were highly effective in the classification of normal and OSCC, and 2 gene sets for the classification of oral dysplasia and cancer. Two of the most significant gene sets in the classification of normal and OSCC included MMP1, SOCS3,
Matrix metalloproteinas 1 (MMP1) was reported as a biomarker related with cancers [17,18], particularly with oral cancer [19–21]. SOCS3 was reported to be related with cancer cell growth [22] and an endogenous inhibitor of pathologic angiogenesis [23]. ACOX1 has not been much studied; however, it can play a role as a member of a gene set. The decreased expression of RUNX2 was observed on colorectal cancer [24]. And it was reported that the increased expression of RUNX2 modulated the expression of apoptosis-associated factors, specifically Bcl-2 [25].

In the case of oral dysplasia and OSCC discrimination, two combined biomarkers were identified. The gene set with EVA1A, NEK6 and KLHL8 had the most discriminative power to separate oral dysplasia from OSCC (accuracy = 0.924). The gene set with PTMRJ, NEK6 and SLC44A1 achieved 91.3% accuracy in the classification of oral dysplasia from OSCC.

It was reported that PTMRJ has been associated with the protective effects for breast cancer risk; however, there was no significant increase in the risk for colorectal cancer with variants or haplotypes in PTMRJ [26]. NEK6 is known to play a critical role in mitotic cell cycle progression [27–28]; the dysregulation of NEK6 expression plays a key role in oncogenesis [28]. Recent studies have shown that NEK6 is upregulated in various cancers [27] and NEK6 is involved in the pathogenesis of hepatocellular carcinoma, and it may be a favorable independent prognostic parameter for hepatocellular carcinoma [28]. Lee et al. (2010) suggested that the downregulation of NEK6 is required for the onset of p53-induced cellular senescence and imply a possible role of NEK6 in tumorigenesis [27].

Some genes in a combined gene set were already reported to be related with cancer progression in previous studies, whereas some genes were not. However, it is possible that these unreported genes have been playing a role as a member in a combined gene set. Therefore, the functional study of a combined gene set including genes, which are not still known, would be meaningful.

The performance of the three combined expressions identified in this study was compared with previous study in two public microarray datasets, GSE30784 and GSE9844. In both datasets, the three combined expressions identified in this study showed better performance compared to those of the previous study, particularly in GSE9844. The limitation of this study is that the identified biomarkers were evaluated using independent two different microarray data sets, not biologically validated. Therefore, the biological validation of the identified biomarkers would be our further work in addition to the implementation of prognosis prediction system including the biomarkers.

For diagnosis of the specific disease, the significant biomarker has been commonly used. However, the performance of diagnosis can be improved by the combined biomarker including insignificant genes, which are not significantly differently expressed between diseased and normal groups. It is why the insignificant genes can play an important role in a combined biomarker by combining with significant genes. And, it could be possible to investigate the still unknown biological pathways of the identified combined biomarkers.

In conclusion, the combined expression was more predictive compared to individual significant genes, which implies that it could be a biomarker for more accurate diagnosis of cancer.

4. Materials and method

4.1. Data preparation

We used three publicly available expression datasets in this study. The first dataset was published by Chen et al. (GSE30784) [3], which was utilized for detecting significant gene sets. This dataset consisted of 45 normal controls, 17 oral dysplasias and 167 oral squamous cell carcinomas. Two further datasets were used for the validation of the combined gene expression. These were cDNA microarray datasets published by Ye et al. (GSE9844) [29] and Pyeon et al. (GSE6791) [14]. The datasets are summarized in Table 3.

4.2. Statistical method

In order to decide the appropriate number of genes in a gene set, we compared the out of bag (OOB) error rates of a gene set with a different size. To calculate OOB error rates, we used Random Forest algorithm (RF) [30], with the following steps.

(1) Generate n datasets of bootstrap samples \{B1, B2,..., Bn\} by allowing repetition of the same sample.
(2) Use each sample \(B_k\) to construct a Tree classifier \(T_k\) to predict those samples that are not in \(B_k\), called out-of-bag (OOB) samples. These predictions are called out-of-bag estimators.
(3) Final prediction is the average of out-of-bag estimators over all bootstrap samples and we get average of them which is overall classification error (OOB error).

OOB error rates were calculated 10,000 times with a resampled gene set with the same size for minimizing the selection bias. The gene set with a minimum OOB error was identified as a significantly discriminative gene set. The process of selecting a significant gene set was summarized in Fig. 5, where the appropriate gene set size is supposed to be 3.

We identified the combined expression of genes in a gene set using the Principal Component Analysis (PCA) [31,32]. PCA is a simple non-parametric statistical method of extracting relevant information from complex data sets involving multiple variables [31], and can be used in relatively small data sets. PCA uses an orthogonal transformation in order to convert correlated variables into uncorrelated variables, known as principal components. The shortcoming of PCA is that the methods for finding the principal components have trouble with high dimensional data or large numbers of data points. However, we will apply PCA to the dataset which contains pre-identified number of genes and several tens of data points. Principal component is represented in the form of linear combinations of the original variables, and it is calculated according to the following formula for each case; we call it the combined expression of genes, or combined biomarker.

Combined expression = \(w_1g_1 + w_2g_2 + w_3g_3\), where \(w_1, w_2,\) and \(w_3\) are the weights of genes and \(g_1, g_2,\) and \(g_3\) are gene expressions.

Table 3

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N stage</td>
<td>Female F 67 F 9 F 48</td>
<td>Male F 54.1</td>
<td>Male F 54.1</td>
</tr>
<tr>
<td>Gender</td>
<td>Normal 45</td>
<td>Normal 12</td>
<td>Normal 12</td>
</tr>
<tr>
<td>Age</td>
<td>Mean 56.40 (std 12.22)</td>
<td>Median 57</td>
<td>Mean 54.1</td>
</tr>
<tr>
<td>Cancer</td>
<td>OSCC 26</td>
<td>Cervical cancer (CC) 20</td>
<td>Cervical cancer (CC) 20</td>
</tr>
<tr>
<td>Dysplasia</td>
<td>Normal 17</td>
<td>Normal 12</td>
<td>Normal 12</td>
</tr>
<tr>
<td>Platform</td>
<td>GPL570</td>
<td>GPL570</td>
<td>GPL570</td>
</tr>
<tr>
<td>(HG-133_Plus_2)</td>
<td>[HG-133_Plus_2]</td>
<td>[HG-133_Plus_2]</td>
<td>[HG-133_Plus_2]</td>
</tr>
<tr>
<td>U133_Plus_2</td>
<td>Affymetrix Human</td>
<td>Affymetrix Human</td>
<td>Affymetrix Human</td>
</tr>
<tr>
<td>Human</td>
<td>Genome U133</td>
<td>Genome U133</td>
<td>Genome U133</td>
</tr>
<tr>
<td>Array</td>
<td>Plus 2.0 Array</td>
<td>Plus 2.0 Array</td>
<td>Plus 2.0 Array</td>
</tr>
<tr>
<td>N probes</td>
<td>54,675 probes</td>
<td>54,675 probes</td>
<td>54,675 probes</td>
</tr>
</tbody>
</table>

* std: standard deviation.
The effectiveness of the combined expression was measured by sensitivity, specificity and accuracy. The statistical analysis was performed using R (version 2.13.0) [33]. Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2013.11.007.

Acknowledgments

This work was supported by the Priority Research Centers Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (2009-0094027).

References