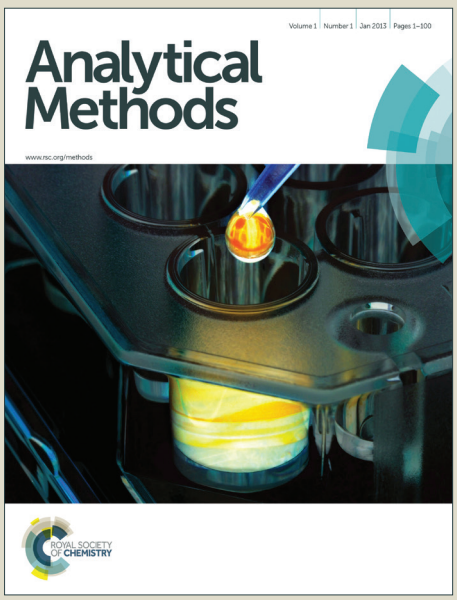
# Analytical Methods

Accepted Manuscript





ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

# Characteristic wavenumbers of Raman spectra reveal the molecular mechanisms of oral leukoplakia and can help to improve the performance of diagnostic models

Liqiu Huang[1], Runyu Jing[1], Yongning Yang[1], Xuemei Pu[1], Menglong Li[1], Zhining Wen[1*], Yi Li[2*]

[1]College of Chemistry, Sichuan University, Chengdu 610064, China

[2]State Key Laboratory of Oral Disease, West China Hospital of Stomatology, Sichuan University, Chengdu 610041, China

*** Correspondence:**

Zhining Wen, College of Chemistry, Sichuan University, Chengdu 610064, China.

Tel: +86-28-85412138

Fax: +86-28-85412356

Email: w_zhining@163.com

Yi Li, State Key Laboratory of Oral Disease, West China Hospital of Stomatology, Sichuan University, Chengdu 610041, China

Tel: +86-28-85501440

Email: liyi1012@163.com

**Abstract**

The correct diagnosis and the prompt treatment of oral leukoplakia (OLK) can efficiently prevent OLK from undergoing malignant transformation to oral squamous cell carcinoma (OSCC). However, the diagnostic model in distinguishing normal mucosa from low-grade dysplasia as well as high-grade dysplasia from OSCC could not be better established in previous study. In this study, the characteristic wavenumbers in the Raman spectra were firstly identified by the variable selection methods. Then, the intensities at these wavenumbers were used to classify the biopsies. As results, the accuracies achieved by using the intensities at the characteristic wavenumbers were 70.5% and 94.0% for the classification of normal *vs.* low-grade dysplasia and high-grade dysplasia *vs.* OSCC, respectively, which were greater than those (accuracy = 65.4% and 88.0%, respectively) using all the intensities in the Raman spectra. Our results suggested constructing the diagnostic models with the intensities at the characteristic wavenumbers can improve the identification of the different lesions of oral mucosa. Moreover, most of the Raman intensities for predicting normal *vs.* low-grade dysplasia indicated the transformation from normal mucosa to low-grade dysplasia was associated with the changes in the contents of the lipids, while most of intensities for predicting high-grade dysplasia *vs.* OSCC indicated that the transformation from high-grade dysplasia to OSCC was associated with the changes in the contents of proteins and nucleic acids. Our findings can be helpful for diagnosing the various grades of OLK with dysplasia and understanding the molecular mechanisms of the potential malignant transformation of oral leukoplakia.

**Keywords**: Oral leukoplakia; Squamous cell carcinoma; Near-infrared Raman spectroscopy; Random forest; Logistic regression

# 1. Introduction

Oral leukoplakia (OLK) is defined as "A white plaques of questionable risk having excluded (other) known diseases or disorders that carry no increased risk for cancer"[1] and is one of the most common disease of oral mucosa. Moreover, it is well known that OLK is the most common precancerous lesion of the oral mucosa with a higher tendency of malignant transformation to oral squamous cell carcinoma (OSCC)[2-5]. The development of OLK in potential malignant transformation is through a multistep process followed by varying grades of oral dysplasia[6, 7]. The presence of oral dysplasia may be a significant predictor for malignant transformation of oral leukoplakia. In addition, there is a general agreement that high-grade dysplasia have significantly higher malignant incidence than low-grade dysplasia[8, 9]. Therefore, the correct detection of various grades of oral dysplasia is a high priority for reducing malignant transformation.

However, the histological features of various grades of dysplasia are not always significantly different, and the exact mechanism of malignant transformation is still unknown. The low-grade dysplasia in oral leukoplakia cannot be easily diagnosed because there is no significant difference in pathological manifestation between low-grade dysplasia and normal mucosa. In the same time, high-grade dysplasia has a strong tendency of malignant transformation and the pathological manifestations in the tissue are similar to those in the OSCC. The histological investigation of high-grade dysplasia in oral leukoplakia could be easily misdiagnosed as the carcinoma lesion[10]. So, it is an important issue to accurately diagnose the grades of oral dysplasia. Nowadays, researchers and pathologists are eager to discover the exact mechanisms of malignant transformation in OLK and hope to build diagnostic models of OLK and OSCC effectively.

The gene expression signatures have the functional relevance to cancers. In recent years, several researches have been carried out for revealing the prognostic markers and building diagnostic models of OLK and OSCC based on the gene expression data. Mario *et al.* measured the expression levels of ATP6V1C1 in OSCC patients and the healthy persons, and used this gene as a prognostic marker to discriminate the OSCC and normal mucosa[11]. Chang *et al.* investigated the pathogenetic implications of miR-211 in oral carcinogenesis[12]. They found the high expression level of miR-211 was associated with the most advanced nodal metastasis, vascular invasion, and poor treatment outcomes of OSCC. Tang *et al.* evaluated the changes of expression levels of six well-documented long non‑coding RNAs

(lncRNAs) that associated with cancer in saliva samples obtained from OSCC patients and suggested that lncRNAs in saliva can be used as potential diagnostic markers for OSCC diagnosis[13]. Although the gene biomarkers were continuously discovered for diagnosing the OSCC, it still lacks a diagnostic model to discriminate the grades of dysplasia in OLK.

The occurrence of malignant transformation is usually caused by the changes in the contents of biomolecules, such as nucleic acids, proteins, lipids and carbohydrates. These changes can provide an opportunity for spectrometer to capture the characteristics involved in the pathological manifestation of biological samples. The vibrational spectroscopy techniques can capture the molecular fingerprint of specific molecular structures and conformations of biomolecules in tissue samples. For the last few years, Raman spectroscopy and NIR spectroscopy have been successfully used in cancer and pre-cancer researches, such as the diagnoses of premalignant and malignant tumor in epithelial tissues[14-16], stomach[17-19], brain[20], oral[21-23] and skin[24, 25]. Meanwhile, various algorithms in chemometrics, such as uninformative variable elimination (UVE), Monte Carlo based UVE (MC-UVE), randomization test (RT), Bayesian variable selection, variable importance projection (VIP), locally linear embedding （LLE）[26-32], have been widely applied to the measured Raman and NIR spectra for selecting characteristic wavenumbers and building the diagnostic models. In our previous works[33], we established diagnostic models by using the Raman spectra generated by Fourier transform near-infrared (FT-NIR) Raman spectrometer. The diagnostic models performed well in discriminating normal mucosa from OLK and OSCC. However, the normal versus the low-grade dysplasia as well as the high-grade dysplasia versus OSCC cannot be accurately classified because of the high similarity of the Raman spectra of the biopsies in these two compared groups. In current study, for the purpose of improving the model performance, we firstly identified the characteristic wavenumbers, for which the Raman intensities were significantly different between the compared biopsies. Then, the intensities at the characteristic wavenumbers were used as features to construct the predictive models. Two variable selection methods, namely, ReliefF and OneR, were used to evaluate the importance of the spectral wavenumbers, and two classification algorithms, namely, Random Forest and Logistic Regression, were used to classify the two groups of compared tissues. As results, depending on the characteristic wavenumbers, we can not only discriminate the normal mucosa from the low-grade dysplasia with accuracy of 70.5% for testing set, but also discriminate high-grade dysplasia from OSCC with accuracy of 94.0% for testing set. Moreover, we found the transformation from the normal tissues to the

low-grade dysplasia was associated with the changes in lipids. The intensity at spectral wavenumber ranged from 300 to 600 $cm^{-1}$ of normal mucosa was higher than those of low-grade dysplasia, which indicated the lipids may mainly dominate in normal oral mucosa tissues. The transformation from the high-grade dysplasia to the OSCC was associated with the changes in collagens and nucleic acids.

## 2. Materials and methods

### 2.1 Tissue samples

Tissue samples were collected from the West China Hospital, Sichuan University. 63 patients clinically diagnosed with OLK with dysplasia or OSCC participated in the present research. The detailed information of patients was shown in Table 1. Diagnoses were carried out by experienced pathologists according to the 2005 World Health Organization (WHO) histological classification[34]. All patients did not receive any treatment before biopsy and were without a history of drug abuse or systemic diseases. All patients preoperatively signed an informed consent and permitted the use of the tissues for research. Our study followed the Declaration of Helsinki protocols and was approved by our Institutional Review Board.

The normal samples of twenty-three patients were obtained from the surgical margin in the tumor surgery, or from the excess mucosa in the trauma or orthognathic surgery, which were confirmed by experienced pathologists. All samples were fixed by 10% formalin and embedded in paraffin. Five parallel 5 μm formalin-fixed paraffin preserved (FFPP) sections were cut from each block using a microtome, and one of them was selected randomly to be mounted on glass slides, dewaxed and stained with hematoxyline-eosin (HE) as the reference section for pathological verification. Another 10 μm FFPP sections were cut from each block using a microtome, and one of them was selected randomly to be mounted on custom $CaF_2$ chips, dewaxed and air-dried as the Raman spectral sections for Raman spectral investigations. All the tissue sections were characterized by the pathologist from the Department of Pathology, West China Hospital of Stomatology, Sichuan University. More details about the methods of dealing with tissue samples have been published previously[33].

### 2.2 Raman spectrometer

A Nicolet Nexus 670 Raman spectrometer (Thermo Nicolet Co., USA) was used with a Nd:YAG laser performing at 1064 nm as a excitation light source. The Nd:YAG laser power detecting the sample was maintained at 1000 mW and the spectrum ranged from 98 to 2000 $cm^{-1}$, which involved a total of 494 spectral wavenumbers. The spectrometer resolution was 8 $cm^{-1}$. The spectra were recorded with 256 scans. Baseline correction was executed by OMNIC for Raman 6.0 software (Nicolet). Finally, a total of 128 spectra were obtained from different biopsies by spectroscopic examination, including 45

spectra from normal mucosa tissue sections, 33 spectra from low-grade dysplasia tissue sections, 31 spectra from high-grade dysplasia sample tissue sections and 19 spectra from OSCC tissue sections (Table 1).

### 2.3 Feature extraction algorithms

In this study, the quality of the features (the Raman intensities at 494 spectral wavenumbers) was evaluated by two feature evaluators, namely OneR and ReliefF. This kind of evaluators will give a score to each of the features, which indicates how well the predictive model discriminate the samples by using this feature. All the features were ranked by the scores in descending order and the top *n* features were used to construct the predictive models. The feature evaluation was conducted in WEKA[35] environment in version 3.6.8 with the packages named "ReliefFAttributeEval" and "OneRAttributeEval".

#### 2.3.1 OneR

OneR[36] can evaluate the importance of a feature by inspecting the prediction error, which is obtained by classifying the samples with this feature and counting the number of incorrectly classified samples. The smaller the error is, the more important the feature is. All the features will be tested and ranked by the errors in ascending order. In our study, we used the top *n* features (*n*=1, 2, 3, ..., the number of features) to construct the predictive models in training procedure and chose the best subset of features to build the models in validation procedure.

#### 2.3.2 ReliefF

Relief[37], which was developed by Kira and Rendel, is proved to be an efficient feature evaluator. A basic idea of the original Relief algorithm is to estimate the quality of features according to how well their values can discriminate the differences between the samples. Kononenko *et al.* improved Relief algorithm and developed ReliefF[38, 39]. Given a dataset of *m* samples ($S = \{s_1, s_2, \ldots, s_m\}$) and *n* features ($A = \{a_1, a_2, \ldots, a_n\}$), ReliefF separately searches for *k* nearest neighbours (*k* was set to 10 in current study) for each of the samples from the same class ($H_j$ denotes the selected neighbours in the same class, where $j = 1, 2, \ldots, k$) and from the different class ($M_j$ denotes the neighbours in different class). The initial weights $W(A)_0$ are set to zero for all the features. The weight $W[a]$ for the *l*th feature will be updated by

using the following equations:

$$W(a_l) = W(a_l) - \sum_{j=1}^{k} diff(s_{i,a_l}, H_{j,a_l}) / (m \times k) + \sum_{j=1}^{k} diff(s_{i,a_l}, M_{j,a_l}) / (m \times k) \quad (1)$$

Function $diff(s_{i,a_l}, H_{j,a_l})$ calculates the difference of the *l*th feature between the *i*th sample and the *j*th neighbour in the same class. Function $diff(s_{i,a_l}, M_{j,a_l})$ calculates the difference of the *l*th feature between the *i*th sample and the *j*th neighbour in the different class. They are defined as:

$$diff(s_{i,a_l}, H_{j,a_l}) = \frac{| a_{s_i,l} - a_{h_j,l} |}{max(A) - min(A)} \quad (2)$$

$$diff(s_{i,a_l}, M_{j,a_l}) = \frac{| a_{s_i,l} - a_{m_j,l} |}{max(A) - min(A)} \quad (3)$$

Where the *max(A)* is the largest value of features and the *min(A)* is the smallest value of features. $a_{s_i,l}$ is the value of the *l*th feature for the *i*th sample. $a_{h_i,l}$ is the value of the *l*th feature for the *j*th sample in the same class. $a_{m_j,l}$ is the value of the *l*th feature for the *j*th sample in the different class. All the processes have been repeated for *m* times. Eventually, the weights *W(a)* for all the features were calculated and applied to determining the importance of the features. The larger the weight of a feature is, the more important the feature is.

Likewise, ReliefF were used to estimate the quality of a total of 494 spectral wavenumbers (features) in current study and generate the ranked lists for the spectral wavenumbers according to their importance. We chose the top *n* features, with which the models can achieved the best prediction results in training procedure, to construct the diagnostic models. ReliefF and OneR were employed in WEKA environment in version 3.6.8 with the packages named "ReliefFAttributeEval" and "OneRAttributeEval", respectively.

**2.4 Predictive model construction**

In our study, two classification algorithms, namely Random Forest and Logistic Regression, were used to build the diagnostic models. The model construction was conducted by using WEKA 3.6.8. For the purpose of selecting the suitable classification algorithms for the two compared groups, we firstly

used both two algorithms to discriminate the grades of oral dysplasia with the full spectrum as features. The predictive models were validated by leave-one-out cross-validation (LOOCV). Then, we chose a suitable variable selection method for the classification of normal *vs.* low-grade dysplasia and high-grade dysplasia *vs.* OSCC to identify the characteristic wavenumbers, for which the intensities were used as features to build the models. In this procedure, the models were validated by 5-fold cross-validation. Finally, the diagnostic models were constructed by using the optimal variable selection method and classification algorithm, and validated by LOOCV. The prediction accuracy (*ACC*), sensitivity (*SEN*), specificity (*SPE*) and Mathew's correlation coefficient (*MCC*) were considered as the performance metrics and calculated via:

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \tag{6}$$

$$SEN = \frac{TP}{TP + FN} \tag{7}$$

$$SPE = \frac{TN}{TN + FP} \tag{8}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{9}$$

where *TP, FP, TN* and *FN* denote the number of true-positive, false-positive, true-negative and false-negative, respectively.

### 2.4.1 Random Forest

The random forest algorithm[40] is an ensemble classification method, which is widely applied to classification owing to its robustness against overfitting and good tolerance to outliers and noise. The key idea of the random forest is to improve the performance of the ensemble classification via majority voting to perform prediction. *Ntree* and *Mtry* are two parameters in the random forest algorithm, which determines the number of individual trees and the number of features that are randomly selected for each of the trees, respectively. We set *Ntree* at 10 and *Mtry* at $\log_2 N+1$ by default in WEKA to grow the ensemble trees, where *N* is the number of features in the dataset. Each of the trees is grown to the largest extent possible, without pruning according to the CART methodology[41, 42].

### 2.4.2 Logistic Regression

The logistic regression algorithm[43] is described as follows:

$$f\left(t\right) = P(Y = 1 \mid x) = 1 \Big/ \left(1 + e^{-t}\right) = e^{t} \Big/ \left(1 + e^{t}\right) \quad (4)$$

$$t = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \quad (5)$$

Where $f(t)$ is the probability of an event occurring and varies from 0 to 1. In current study, $Y$ is a binary variable representing the positive sample (defined as 1) or the negative sample (defined as 0). $t$ is the linear combination of features. $b_0$ denotes the intercept for model. $\{b_1, \ldots, b_n\}$ are the partial regression coefficients. $\{x_1, \ldots, x_n\}$ are the independent spectral features. $b_0$ and a series of regression coefficients can be estimated by using the training set.

# 3. Results

### 3.1 Subtracted mean spectra

A total of 128 spectra were collected from the tissue samples, including 45 Raman spectra from normal mucosa, 33 from low-grade dysplasia, 31 Raman spectra from high-grade dysplasia, and 19 from OSCC. The predictive models were constructed to discriminate the normal mucosa from the low-grade dysplasia tissue (normal *vs.* low-grade dysplasia) and discriminate the high-grade dysplasia tissue from OSCC (high-grade dysplasia *vs.* OSCC). Fig. 1 showed the Raman spectra of four types of tissues in these two compared groups. The mean spectra of four types of tissues were calculated and the subtracted mean spectra of the two compared groups were shown in Fig. 2. The main differences in the subtracted spectra between the normal mucosa and the low-grade dysplasia were located in the region ranged from 300 to 600 $cm^{-1}$ (Fig. 2A), while the differences between the high-grade dysplasia and OSCC were located in the region ranged from 700 to 1100 $cm^{-1}$ (Fig. 2B).

### 3.2 Constructing diagnostic models for tissue classification

For the comparisons of normal *vs.* low-grade dysplasia and high-grade dysplasia *vs.* OSCC, The diagnostic models were separately constructed by using random forest and logistic regression with all the Raman intensities, and validated by leave-one-out cross-validation (LOOCV). The prediction results were listed in Table 2. For the classification of normal *vs.* low-grade dysplasia, the random forest algorithm performed better (accuracy = 65.4%) than the logistic regression algorithm (accuracy = 48.7%), while for the classification of high-grade dysplasia *vs.* OSCC, the logistic regression algorithm is better (accuracy = 88.0% and 74.0% for the logistic regression algorithm and the random forest algorithm, respectively). Therefore, in the subsequent analysis, we used the random forest algorithm to discriminate normal mucosa from low-grade dysplasia and used the logistic regression algorithm for the discrimination of high-grade dysplasia *vs.* OSCC.

In order to improve the performance of models, two variable selection methods, namely OneR and ReliefF, were involved in our study to evaluate the importance of the features and only the intensities at the characteristic wavenumbers were used to build the models. For each of the comparison groups, both two variable selection methods were applied to identifying the characteristic wavenumbers and the prediction results with the intensities at the characteristic wavenumbers were listed in Table 3. When

discriminating the normal mucosa from the low-grade dysplasia, constructing the models with the features selected by OneR performed better (accuracy = 82.5%) than that with the features selected by ReliefF (accuracy = 72.4%). With regard to the classification of high-grade dysplasia *vs.* OSCC, it is better to construct the models with the features selected by ReliefF (accuracy = 94.7% and 89.3% for ReliefF and OneR, respectively). Consequently, we build the diagnostic model for the classification of normal *vs.* low-grade dysplasia by using the random forest algorithm combined with OneR, while build the model for the classification of high-grade dysplasia *vs.* OSCC by using the logistic regression algorithm combined with ReliefF. The models were validated by using LOOCV. The prediction results achieved by the diagnostic models were listed in Table 4. Compared with the prediction results achieved by the diagnostic models that were constructed with all the Raman intensities (Table 2), the accuracies achieved by the models constructed with the intensities at the characteristic wavenumbers, which were 70.5% and 94.0% for the classification of normal *vs.* low-grade dysplasia and high-grade dysplasia *vs.* OSCC, respectively, were higher than those listed in Table 2.

For the purpose of investigating the characteristic wavenumbers that were suitable for model construction, we counted the frequency for each of the wavenumbers, for which the intensity was involved in the cross-validation procedures as feature for models construction (Fig. 3). For the classification of normal *vs.* low-grade dysplasia, the wavenumbers with the frequency > 60 were mainly located in the region ranged from 300 to 600 $cm^{-1}$ (Fig. 3A). For the classification of high-grade dysplasia *vs.* OSCC, the wavenumbers with the frequency > 45 were mainly located in the region ranged from 700 to 1100 $cm^{-1}$ (Fig. 3B).

## 4. Discussion

In the diagnosis of oral leukoplakia, it is important to grade oral lesions accurately for a proper treatment strategy. The near-infrared Raman spectra could reflect the vibrational modes of functional groups of the biomolecules in biological samples, and could be used for the diagnosis of oral diseases. In the previous study, our diagnostic models achieved the satisfied results in discriminating the normal mucosa from OLK and OSCC when using all the Raman intensities as features. However, the performance of the diagnostic model in discriminating the normal mucosa from the low-grade dysplasia in OLK as well as the high-grade dysplasia in OLK from OSCC is poor. In current study, we tried to improve the predictive models by using the Raman intensities at the characteristic wavenumbers as features. The prediction results showed that the diagnostic models constructed with intensities at the characteristic wavenumbers performed better than those constructed with all the intensities. The accuracies achieved by our improved models were 70.5% and 94.0% (Table 4) for the classification of normal *vs.* low-grade dysplasia and high-grade dysplasia *vs.* OSCC, respectively, which were higher than those achieved by the models constructed with all the intensities (accuracy = 65.4% and 88.0% for the classification of normal *vs.* low-grade dysplasia and high-grade dysplasia *vs.* OSCC, respectively). Our results indicated that it can efficiently improve the performance of the models by using the intensities at the characteristic wavenumbers as features.

In the subsequent analysis, we counted the frequency for each of the wavenumbers, for which the intensity was selected as feature to build the models in the LOOCV procedures. The results showed that most of the selected wavenumbers with high frequency (frequency > 60) were located in the region ranged from 300 to 600 $cm^{-1}$ (Fig. 3A) when discriminating the normal tissue from the low-grade dysplasia, while the selected wavenumbers (frequency > 45) were mainly located in the region ranged from 700 to 1100 $cm^{-1}$ (Fig. 3B) when discriminating the high-grade dysplasia from OSCC. It can also be seen that the characteristic wavenumbers identified by the variable selection algorithms for the two compared groups were located in the peaks of the subtracted mean spectra (Fig. 2).

For the classification of normal *vs.* low-grade dysplasia, the characteristic wavenumbers fell in the region ranged from 300 to 600 $cm^{-1}$, which was attributed to C-C bending vibration within aliphatic chains. The characteristic wavenumber of 1309 $cm^{-1}$ corresponded to the $CH_3/CH_2$ wagging, twisting or blending mode of lipids[44], and the characteristic wavenumber of 1452 $cm^{-1}$ corresponded to $CH_2$

blending vibration of lipids (Fig. 2A)[45]. Therefore, it can be inferred that, to a certain extent, biomaterial changes in the tissues were usually associated with the content of lipid when normal tissue transformed into low-grade dysplasia, in which the lipid would help less in cell's proliferation. For the classification of high-grade dysplasia *vs.* OSCC, the characteristic wavenumbers fell in the regions ranged from 800 to 990 $cm^{-1}$and from 1010 to 1100 $cm^{-1}$, which were associated with the changes of the content of collagen. In addition, the peaks in the region ranged from 1080 to 1100 $cm^{-1}$ were dominated by the contributions of lipids, nucleic acids, proteins and carbohydrates[45]. The characteristic wavenumbers at 727, 1032, and 1124 $cm^{-1}$ corresponded to C-C stretching, twisting and bending of collagen[46-48]. The wavenumbers at 727, 731, 746, 766, 1055, 1078, 1080, 1120 and 1336 $cm^{-1}$ corresponded to Uracil, Cytosine and Thymine ring breathing mode of DNA/RNA, O-P-O backbone symmetric stretching and $CH_2$ rocking of nucleic acids[45, 46, 49]. The wavenumbers at 1070, 1078, and 1124 $cm^{-1}$ were the C-C, C-O skeletal transconformation of acyl backbone in lipid[44, 50, 51]. The wavenumber at 1082 $cm^{-1}$ was the characteristic carbohydrate corresponded to C-C stretching of glycogen and at 1105 $cm^{-1}$ was Carbohydrates peak for solutions[52] (Table 5). It can be inferred that during the malignant transformation from high-grade dysplasia into OSCC, the genetic materials have been activated and the corresponding molecular signaling pathways have been impacted, which result in the changes in the content of proteins in OSCC. Note that the accuracy for discriminating the normal tissues from low-grade dysplasia was lower than that for the classification of high-grade dysplasia vs. OSCC. The main reason is that the contents of biomolecules changed slightly when the transformation occurred from the normal tissues to the low-grade dysplasia. In order to further improve the predictive accuracy, more information is needed for the model construction.

## 5. Conclusions

In this study, we developed the predictive models for diagnosing the various grades of oral leukoplakia by using the intensities at the characteristic wavenumbers as features. Our results suggested that it can efficiently improve the performance of the predictive models in the classifications of the normal *vs.* low-grade dysplasia as well as high-grade dysplasia *vs.* OSCC by using the intensities at the characteristic wavenumbers instead of all the intensities as features. Moreover, the characteristic wavenumbers revealed the potential mechanisms of the transformations from the normal tissue to the low-grade dysplasia and from the high-grade dysplasia to OSCC at the molecular levels. Our findings can be helpful for understanding the molecular mechanisms of potential malignant transformation of oral leukoplakia.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This work was supported by the National Nature Science Foundation of China (Nos. 21205085，81001209 and U1230121).

REFERENCE:


1. S. Warnakulasuriya, N. W. Johnson and I. Van Der Waal, *J. Oral Pathol. Med.*, 2007, **36**, 575-580.
2. S. S. Jr, M. Gorsky and F. L. D. Ms, *Cancer*, 1984, **53**, 563-568.
3. A. Mashberg, *Oral Oncology*, 2000, **36**, 253-255.
4. M. Lopez, J. Aguirre, N. Cuevas, M. Anzola, J. Videgain, J. Aguirregaviria and M . Martínez de Pancorbo, *European Journal of Cancer*, 2003, **39**, 2306-2309.
5. F. Galle, G. Colella, V. Di Onofrio, R. Rossiello, I. F. Angelillo and G. Liguori, *New Microbiol.*, 2013, **36**, 283-288.
6. B. W. Neville and T. A. Day, *CA: A Cancer Journal for Clinicians*, 2002, **52**, 195-215.
7. I. van der Waal, *Oral Oncology*, 2009, **45**, 317-323.
8. H. M. Mehanna, T. Rattay, J. Smith and C. C. McConkey, *Head & Neck*, 2009, **31**, 1600-1609.
9. S. Warnakulasuriya, T. Kovacevic, P. Madden, V. H. Coupland, M. Sperandio, E. Odell and H. Møller, *J. Oral Pathol. Med.*, 2011, **40**, 677-683.
10. D. M. Walker, G. Boey and L. McDonald, *Pathology*, 2003, **35**, 376-383.
11. M. Pérez-Sayáns, M. D. Reboiras-López, J. M. Somoza-Martín, F. Barros-Angueira, P. G. Diz, J. M. G. Rey and A. García-García, *Cancer biology & therapy*, 2010, **9**, 1057-1064.
12. K.-W. Chang, C.-J. Liu, T.-H. Chu, H.-W. Cheng, P.-S. Hung, W.-Y. Hu and S.-C. Lin, *Journal of Dental Research*, 2008, **87**, 1063-1068.
13. H. Tang, Z. Wu, J. Zhang and B. Su, *Molecular medicine reports*, 2013, **7**, 761-766.
14. M. D. Morris and P. Matousek, *Emerging Raman applications and techniques in biomedical and pharmaceutical fields*, Springer, 2010.
15. N. Stone, C. Kendall, N. Shepherd, P. Crow and H. Barr, *Journal of Raman Spectroscopy*, 2002, **33**, 564-573.
16. C. Bielecki, T. W. Bocklitz, M. Schmitt, C. Krafft, C. Marquardt, A. Gharbi, T. Knösel, A. Stallmach and J. Popp, *Journal of biomedical optics*, 2012, **17**, 0760301-0760308.
17. F. M. Lyng, E. Ó. Faolán, J. Conroy, A. Meade, P. Knief, B. Duffy, M. Hunter, J. Byrne, P. Kelehan and H. J. Byrne, *Experimental and molecular pathology*, 2007, **82**, 121-129.
18. S. Teh, W. Zheng, K. Ho, M. Teh, K. Yeoh and Z. Huang, *British journal of cancer*, 2008, **98**, 457-465.
19. S. Teh, W. Zheng, K. Ho, M. Teh, K. Yeoh and Z. Huang, *British Journal of Surgery*, 2010, **97**, 550-557.
20. L. M. Fullwood, G. Clemens, D. Griffiths, K. Ashton, T. P. Dawson, R. W. Lea, C. Davis, F. Bonnier, H. J. Byrne and M. J. Baker, *Analytical Methods*, 2014, **6**, 3948-3961.
21. A. P. Oliveira, R. A. Bitar, L. Silveira Jr, R. A. Zângaro and A. A. Martin, *Photomedicine and Laser Therapy*, 2006, **24**, 348-353.
22. K. Guze, M. Short, H. Zeng, M. Lerman and S. Sonis, *Journal of Raman Spectroscopy*, 2011, **42**, 1232-1239.
23. S. Singh, A. Deshmukh, P. Chaturvedi and C. M. Krishna, *Journal of biomedical optics*, 2012, **17**, 1050021-1050029.
24. C. A. Lieber, S. K. Majumder, D. L. Ellis, D. D. Billheimer and A. Mahadevan‐Jansen, *Lasers in surgery and medicine*, 2008, **40**, 461-467.
25. H. Lui, J. Zhao, D. McLean and H. Zeng, *Cancer Research*, 2012, **72**, 2491-2500.
26. W. Cai, Y. Li and X. Shao, *Chemometrics and Intelligent Laboratory Systems*, 2008, **90**, 188-194.
27. L. Yan-kun, *Analytical Methods*, 2012, **4**, 254-258.

28. D. Wu, X. Chen, X. Zhu, X. Guan and G. Wu, *Analytical Methods*, 2011, **3**, 1790-1796.
29. H. Xu, Z. Liu, W. Cai and X. Shao, *Chemometrics and Intelligent Laboratory Systems*, 2009, **97**, 189-193.
30. T. Chen and E. Martin, *Analytica chimica acta*, 2009, **631**, 13-21.
31. M. A. Hedegaard, K. L. Cloyd, C.-M. Horejs and M. M. Stevens, *Analyst*, 2014, **139**, 4629-4633.
32. N. Qi, Z. Zhang, Y. Xiang and P. d. B. Harrington, *Analytica chimica acta*, 2012, **724**, 12-19.
33. Y. Li, Z. N. Wen, L. J. Li, M. L. Li, N. Gao and Y. Z. Guo, *Journal of Raman Spectroscopy*, 2010, **41**, 142-147.
34. D. M. Parkin, F. Bray, J. Ferlay and P. Pisani, *International journal of cancer*, 2001, **94**, 153-156.
35. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, *ACM SIGKDD Explorations Newsletter*, 2009, **11**, 10-18.
36. R. C. Holte, *Machine learning*, 1993, **11**, 63-90.
37. K. Kira and L. A. Rendell, Proceedings of the ninth international workshop on Machine learning, 1992.
38. I. Kononenko, Machine Learning: ECML-94, 1994.
39. M. Robnik-Šikonja and I. Kononenko, *Machine learning*, 2003, **53**, 23-69.
40. L. Breiman, *Machine learning*, 2001, **45**, 5-32.
41. L. B. J. F. R. Olshen and C. J. Stone, *Wadsworth International Group*, 1984.
42. R. Duda, P. Hart and D. Stork, *New York*, 2001.
43. B. G. Tabachnick, L. S. Fidell and S. J. Osterlind, 2001.
44. W. T. Cheng, M. T. Liu, H. N. Liu and S. Y. Lin, *Microscopy research and technique*, 2005, **68**, 75-79.
45. P. Andrade, R. Bitar, K. Yassoyama, H. Martinho, A. Santo, P. Bruno and A. Martin, *Analytical and bioanalytical chemistry*, 2007, **387**, 1643-1648.
46. A. Nijssen, T. C. B. Schut, F. Heule, P. J. Caspers, D. P. Hayes, M. H. Neumann and G. J. Puppels, *Journal of Investigative Dermatology*, 2002, **119**, 64-69.
47. C. M. Krishna, G. Sockalingum, J. Kurien, L. Rao, L. Venteo, M. Pluot, M. Manfait and V. Kartha, *Applied spectroscopy*, 2004, **58**, 1128-1135.
48. R. Malini, K. Venkatakrishna, J. Kurien, K. M Pai, L. Rao, V. Kartha and C. M. Krishna, *Biopolymers*, 2006, **81**, 179-193.
49. A. W. Auner, R. E. Kast, R. Rabah, J. M. Poulik and M. D. Klein, *Pediatric surgery international*, 2013, **29**, 129-140.
50. L. Silveira, S. Sathaiah, R. A. Zângaro, M. T. Pacheco, M. C. Chavantes and C. A. Pasqualucci, *Lasers in surgery and medicine*, 2002, **30**, 290-297.
51. N. Stone, C. Kendall, J. Smith, P. Crow and H. Barr, *Faraday discussions*, 2004, **126**, 141-157.
52. Z. Movasaghi, S. Rehman and I. U. Rehman, *Applied Spectroscopy Reviews*, 2007, **42**, 493-541.

# Figures

**Figure 1 - The Raman spectra of four subclasses of normal tissue, the low-grade dysplasia, the high-grade dysplasia and OSCC.** The comparison of Raman spectra between A) normal tissues and the low-grade dysplasia tissues, and B) The comparison of Raman spectra between the high-grade dysplasia tissues and OSCC tissues. For clarity, we randomly picked 10 spectra for each of the subclasses.

**Figure 2 - The subtracted mean spectra and the characteristic wavenumbers selected for the diagnostic model construction.** The subtracted mean spectra of A) the normal *vs.* the low-grade dysplasia, and B) The high-grade dysplasia *vs.* OSCC. The black sold line is the subtracted spectra and the red circles indicate the position of the characteristic wavenumbers.

**Figure 3 - The frequencies of the wavenumbers selected as features in the model construction procedures.** The frequencies of the wavenumbers selected as features in classification of A) the normal *vs.* the low-grade dysplasia, and B) the high-grade dysplasia *vs.* OSCC. The radius of the circle indicates the frequency and the bars indicate the frequencies of the wavenumbers selected in the predictive models. The red bars indicates the frequencies that are above the given threshold, which were set to 60 and 45 for the classification of normal *vs.* low-grade dysplasia and high-grade dysplasia *vs.* OSCC, respectively.

# Tables

**Table 1. The detailed of patients and the number of Raman spectra**

| Case information | | Normal | Low-grade dysplasia | High-grade dysplasia | OSCC |
|---|---|---|---|---|---|
| All patients | | 23 | 16 | 14 | 10 |
| Age | | Range 28~54 years | Range 31~54 years | Range 32~56 years | Range 29~55 years |
| | | Median 41 years | Median 43 years | Median 41.5 yaers | Median 42 years |
| Gender | Male | 11 | 7 | 6 | 5 |
| | Female | 12 | 9 | 8 | 5 |
| Primary site | Tongue | 10 | 8 | 7 | 6 |
| | Bucca | 13 | 8 | 7 | 4 |
| Raman spectra | | 45 | 33 | 31 | 19 |

**Table 2. The prediction results of normal *vs.* low-grade dysplasia and high-grade dysplasia *vs.* OSCC with all the intensities in Raman spectra as features.**

| Groups | Algorithms | SEN | SPE | ACC | MCC |
|---|---|---|---|---|---|
| Normal *vs.* Low-grade dysplasia | Random Forest | 80.0% | 45.5% | 65.4% | 0.273 |
| | Logistic Regression | 57.8% | 36.4% | 48.7% | 0.0591 |
| High-grade dysplasia *vs.* OSCC | Random Forest | 80.6% | 63.2% | 74.0% | 0.443 |
| | Logistic Regression | 93.5% | 79.0% | 88.0% | 0.743 |

**Table 3. Prediction results of normal *vs.* low-grade dysplasia and high-grade dysplasia *vs.* OSCC when using the intensities at the characteristic wavenumbers identified by two variable selection methods.**

| Groups | Algorithms | SEN | SPE | ACC | MCC |
|---|---|---|---|---|---|
| Normal *vs.* Low-grade dysplasia | OneR_Random Forest | 92.7% | 75.6% | 82.5% | 0.675 |
| | ReliefF__Random Forest | 80.0% | 73.3% | 72.4% | 0.472 |
| High-grade dysplasia *vs.* OSCC | OneR_Logistic Regression | 95.0% | 80.0% | 89.3% | 0.771 |
| | ReliefF_Logistic Regression | 100% | 85.0% | 94.7% | 0.887 |

**Table 4. Prediction results of normal *vs.* low-grade dysplasia and high-grade dysplasia *vs.* OSCC by using the optimal variable selection method and classification algorithm.**

| Groups | Algorithms | SEN | SPE | ACC | MCC |
|---|---|---|---|---|---|
| Normal *vs.* Low-grade dysplasia | OneR_Random Forest | 80.0% | 57.6% | 70.5% | 0.387 |
| High-grade dysplasia *vs.* OSCC | ReliefF_Logistic Regression | 96.8% | 89.5% | 94.0% | 0.872 |

**Table 5. Raman wavenumbers and their assignments**

| Wavenumber in $cm^{-1}$ | Peak assignment |
|---|---|
| 727 | C-C stretching of collagen, Adenine |
| 731 | $CH_2$ rocking of Adenine |
| 746 | Thymine (ring breathing mode of DNA/RNA bases), $CH_2$ rocking of phenylalanine |
| 766 | Pyrimidine ring breathing mode |
| 1032 | C-C bending modes of collagen & phospholipids |
| 1055 | nucleic acids |
| 1059 | Phospholipids/phosphatidylcholine |
| 1070 | Triglycerides (fatty acids) |
| 1078 | C-C or C-O stretching mode of phospholipids, C-C or O-P-O stretching of nucleic acid |
| 1082 | C-C stretching mode of glycogen |
| 1086 | C-N stretching of proteins, lipids; C-C, C-O stretching of phospholipids |
| 1090 | Symmetric phosphate stretching vibrations |
| 1105 | Carbohydrates peak for solutions |
| 1120 | the strong C-O band of RNA |
| 1124 | C-C skeletal of acyl backbone in lipid (trans conformation) |
| 1309 | $CH_3/CH_2$ wagging, twisting or blending mode of lipids or collagen |
| ≈1450 | $\delta(CH_2)$ of Lipids, carbohydrates and proteins |

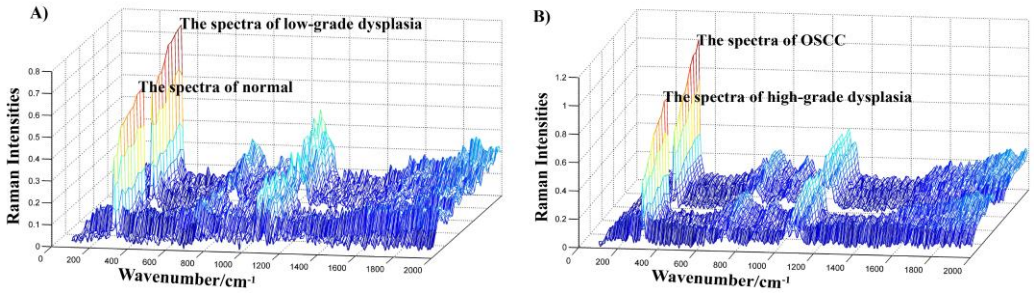



**Fig.1. The Raman spectra of four subclasses of normal tissue, the low-grade dysplasia, the high-grade dysplasia and OSCC.** The comparison of Raman spectra between A) normal tissues and the low-grade dysplasia tissues, and B) The comparison of Raman spectra between the high-grade dysplasia tissues and OSCC tissues. For clarity, we randomly picked 10 spectra for each of the subclasses.

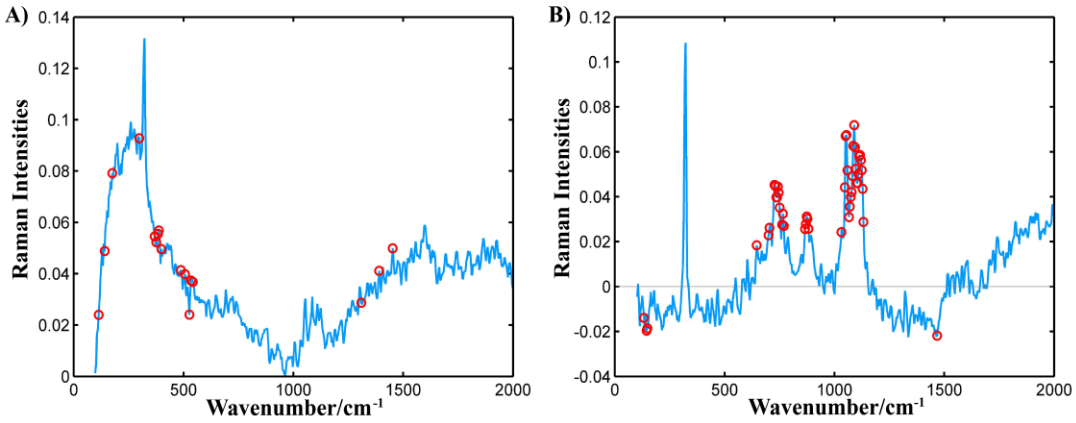



Fig.2. **The subtracted mean spectra and the characteristic wavenumbers selected for the diagnostic model construction.** The subtracted mean spectra of A) the normal vs. the low-grade dysplasia, and B) The high-grade dysplasia vs. OSCC. The black sold line is the subtracted spectra and the red circles indicate the position of the characteristic wavenumbers.

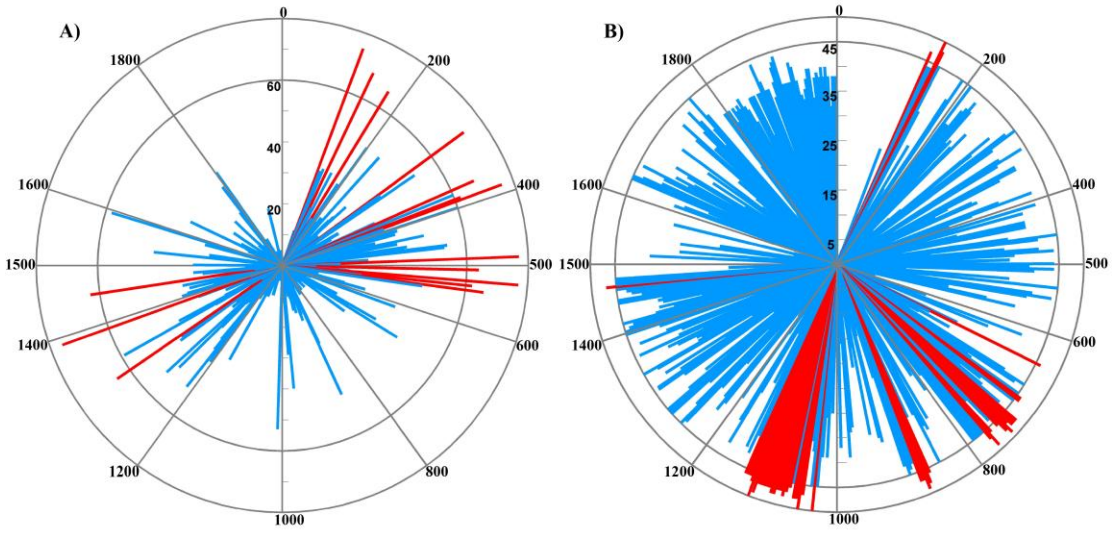



**Fig.3. The frequencies of the wavenumbers selected as features in the model construction procedures.** The frequencies of the wavenumbers selected as features in classification of A) the normal vs. the low-grade dysplasia, and B) the high-grade dysplasia vs. OSCC. The radius of the circle indicates the frequency and the bars indicate the frequencies of the wavenumbers selected in the predictive models. The red bars indicates the frequencies that are above the given threshold, which were set to 60 and 45 for the classification of normal vs. low-grade dysplasia and high-grade dysplasia vs. OSCC, respectively.