Clinical Oral Investigations (2019) 23:3021–3031
https://doi.org/10.1007/s00784-018-2714-5

ORIGINAL ARTICLE

In vivo Raman spectroscopic characteristics of different sites of the oral mucosa in healthy volunteers

Luis Felipe C. S. Carvalho[1,2,3] · Marcelo Saito Nogueira[4] · Tanmoy Bhattacharjee[5] · Lazaro P. M. Neto[6] · Lucas Daun[7] · Thiago O. Mendes[6] · Ramu Rajasekaran[7] · Maurílio Chagas[1] · Airton A. Martin[6] · Luis Eduardo S. Soares[1]

Received: 18 May 2018 / Accepted: 17 October 2018 / Published online: 7 November 2018

Abstract

Objectives Investigate the biochemistry of in vivo healthy oral tissues through Raman spectroscopy. We aimed to characterize the biochemical features of healthy condition in oral subsites (buccal mucosa, lip, tongue, and gingiva) of healthy subjects. More specifically, we investigated Raman spectral characteristics and biochemical content of in vivo healthy tissues on Brazilian population. This characterization can be used to better define normal tissue and improve the detection of oral premalignant conditions in future studies.

Materials and methods For spectroscopic analysis a Raman spectrometer (Kaiser Optical Systems imaging spectrograph Holospec, f / 1.8i-NIR) coupled with a laser 785 nm, 60 mW was used. Raman measurements were obtained by means of an optical fiber (EMVision fiber optic probe) coupled between the laser and the spectrometer. Three spectra per site were acquired from the lip, buccal mucosa, tongue, and gingiva of ten healthy volunteers. This resulted in 30 spectra per oral sub-site and in total 120 spectra.

Results We report detailed biochemical information on these subsites and their relative composition based on deconvolution studies of their spectra. Finally, we also report classification efficiency of 61, 83, 41, and 93% for buccal, gingiva, lip, and tongue respectively after applying multivariate statistical tools.

Conclusions We quantitated the contribution of various biochemicals in terms of percentage, and this will enable comparison not only across anatomical sites but also across studies. Raman spectroscopy can rapidly probe tissue biochemistry of healthy oral regions. Moreover, the study suggests the possibility of using Raman spectroscopy combined with signal processing and multivariate analysis methods to differentiate the oral sites in healthy conditions and compare with pathological conditions in future studies.

Clinical relevance The spectral characterization of the healthy condition of oral tissues by a noninvasive, label-free, and real-time analytical techniques is important to create a spectral reference for future diagnosis of pathological conditions.

Keywords In vivo · Oral pathology · Clinical

Luis Felipe C. S. Carvalho, Marcelo Saito Nogueira and Tanmoy Bhattacharjee contributed equally to this work.

✉ Luis Felipe C. S. Carvalho
luisfelipecarvalho@hotmail.com

✉ Marcelo Saito Nogueira
marcelosaitonogueira@gmail.com

[1] Laboratory of Dentistry and Applied Materials, Univap/Instituto de Pesquisa e Desenvolvimento, Avenida Shishima Hifumi 2911, São José dos Campos, SP CEP: 12244-000, Brazil

[2] Faculdade De Odontologia, Da Universidade De Taubaté (Unitau), Rua dos Operarios, 53, Taubate, SP CEP 12020-270, Brazil

[3] Centro Universitário Braz Cubas, Mogi das Cruzes, São Paulo, Brazil

[4] Tyndall National Institute, Lee Maltings Complex, Dyke Parade, Cork T12 R5CP, Ireland

[5] Laboratory of Nanosensors, Univap/Instituto de Pesquisa e Desenvolvimento, Avenida Shishima Hifumi 2911, São José dos Campos, SP CEP:12244-000, Brazil

[6] Biomedical Engineering innovation Center-Biomedical Vibrational Spectroscopy Group, Universidade Brasil-UnBr, Rua Carolina Fonseca 235, Itaquera, São Paulo/SP 08230-030, Brazil

[7] Univap/Instituto de Pesquisa e Desenvolvimento, Avenida Shishima Hifumi 2911, São José dos Campos, SP CEP:12244-000, Brazil

Introduction

The search for new methods of oral lesion diagnosis has resulted in the development of research aimed at detecting changes in their initial phase. Many cases are diagnosed at a later stage, making treatment more difficult and expensive, compromising the prognosis. Histopathological examination, performed through tissue analysis, obtained by excisional or incisional biopsy, is still considered by many authors to be the gold standard of diagnosis. In addition to histopathological examination, additional tests may be required at the time of the diagnostic process, to assist in the elaboration of diagnostic hypotheses, including imaging (radiography, computed tomography, magnetic resonance imaging) and laboratory tests (serological, biochemical, microbiological) among others [1].

Thus, other tools to aid in the diagnosis of various types of lesions should be investigated. Considering the noninvasive diagnostic techniques under development, Raman spectroscopy stands out. Raman spectroscopy, considered a noninvasive analytical, can be very useful in the early diagnosis of various lesions, as alterations in the molecular composition of pathological tissues can be detected by the spectral reflections [2–4]. The term "optical biopsy" is widely used because spectroscopy analyzes the tissue for its optical properties and can provide additional information of the assessed tissue and thus assist in the diagnostic process. Research has shown that Raman spectroscopy is able to detect spectral tissue changes and provide biochemical information in breast and skin among others types of cancer [5].

Raman spectroscopy has been shown to quickly and reliably identify oral malignant condition with approximately 90–95% classification efficiency [3, 6–12]. These studies also highlight the large biochemical difference between the cancer and normal tissue/cells. However, as comparison moves towards conditions preceding full malignancy, such as premalignant [13–17], contralateral, and cancer field effects [11, 18], the accuracy of tissue classification (i.e., ability to identify tissue types) steadily decreases. This is expected, since as one studies pre-cancer conditions, the biochemical differences decrease and start to become more similar to healthy tissues. These differences decrease further when taking into consideration clinically/histopathologically undetectable pre-cancer conditions, i.e., conditions where no morphological changes can be observed, although tissue biochemistry is altered. Undetectable conditions were previously studied in hamster models [19, 20].

Thus, it becomes increasingly important to rigorously define controls and thresholds of their biochemical composition. Sahu et al. [18] have previously reported spectral characteristics of healthy oral sites. However, given the importance of defining healthy, we believe that studies on healthy population are relevant. Moreover, we report extensive data on biochemistry of each healthy sub-site. With this in mind, the present Raman spectroscopy study was carried out in healthy patients to rapidly demonstrate the biochemistry associated with healthy oral tissues and differences in specific biochemical components.

Materials and methods

Ethics statement

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The study was approved by Research Ethics Committee of Universidade do Vale do Paraíba (UNIVAP) via Plataforma Brasil—Brazil (number 1132237-2015).

Raman instrument

For spectroscopic analysis a Raman spectrometer (Kaiser Optical Systems imaging spectrograph Holospec, f/1.8i-NIR) coupled with a laser 785 nm, 60 mW was used. Raman measurements were obtained by means of an optical fiber (EMVision fiber optic probe) coupled between the laser and the spectrometer. The Raman scattered light was collected by the same fiber through dichroic mirror gold and finally focused on the entrance aperture of the spectrometer through a holographic notch filter. Laser excitation energy interacts with the molecules in the tissue, promoting light scattering from the vibrational modes. Raman scattered light was captured by the system and converted into spectra. The acquisition time was performed with an interaction of 40 s for each spectrum (20 × 2 s). The signal was then collected by a CCD detector (Andor-IDUs 420 series) whose quantum efficiency is around 95%. The use of laser excitation in the infrared region largely removes the intrinsic fluorescence of biological tissues, facilitating the collection of Raman signals. This whole system is connected to a computer that receives the spectrometer information and turns them into spectra.

Subject information

Three spectra per site were acquired from the lip, buccal mucosa, tongue, and gingiva (Fig. 6) of ten healthy volunteers. This resulted in 30 spectra per oral sub-site and in total 120 spectra.

Data analysis

For mean spectrum, the spectra were corrected by subtracting a polynomial of order five, and then vector was normalized.

Savitsky-Golay filter (5th order, frame size 7) was used to smooth the spectra. These spectra were then deconvoluted using OriginPro 8.5 software based on peaks obtained by a second derivatization. Raw spectra were first derivatized and vector normalized for Principal Component Analysis (PCA), preprocessing method to reduce the number of spectral parameters by generating a new set of independent features ordered by the largest variability in our dataset, and subsequent employment of linear discriminant analysis (LDA) as a classifier. LDA followed by leave-one-out cross-validation was first employed to classify normal tissues from the buccal mucosa, gingiva, lip, and tongue sites with few parameters. Then, accuracy improvements were evaluated by considering two types of spectrum normalization (by the area under the curve or by its intensity maximum) and four other classifiers (K-nearest neighbors, unpruned C4.5 decision tree or J48, random forest, and multilayer perceptron). These improvements were investigated by using only three principal components. The spectral range 900–1800 cm^{-1} was used for all analyses.

Results and discussion

Characterization of tissues biochemical content

Figure 1 shows the mean spectra of the four oral subsites and have features similar to those reported earlier [18, 21]. Amide I, amide III, and lipid bands (1443, 1745) are more prominent in the buccal and lip compared to the gingiva and tongue. Gingiva has a sharp phosphate peak (960 cm^{-1}) indicative of contribution from teeth and bone. Anatomically, the buccal and lip have more lipids and matches the features observed in mean spectra (Fig. 2).

To further explore these differences, spectral deconvolution and curve fitting were performed. It is well known that the mean spectrum is a combination of signals from several biochemical components of the tissue. To delineate contribution from each component, the spectrum can be deconvoluted—that is, contribution of each signal can be separated. To achieve this, second derivative of the spectrum is obtained, and peaks apparent in the modified spectrum are noted. Then, in a deconvolution software (OriginPro in this case), the information is provided. The software uses the information to fit Gaussian curves to each peak. If the fitting is correct (χ^2 values are in the range of 10^{-5} or less, and R^2 values are $\sim = 1$), areas under these fitted curve can be considered as contribution of the biochemical component associated with the peak to the spectrum. By calculating the percentage area of each peak, contribution of all the biochemical to the sample spectrum can be acquired. These can then be compared across samples to derive information on chemical variation of each sample.

Figure 3 shows the deconvolution of different spectral regions for each subsite, while Table 1 lists all the biochemical features and their contribution in terms of area. The percent contribution of each chemical was calculated by dividing the area under the peak for that chemical by the total areas of all biochemical in that mean spectrum. While it would be more convenient to have a single column listing the vibrations, separate columns have been used since the peaks vary for each sub-site. From Table 1, it can be seen that buccal mucosa

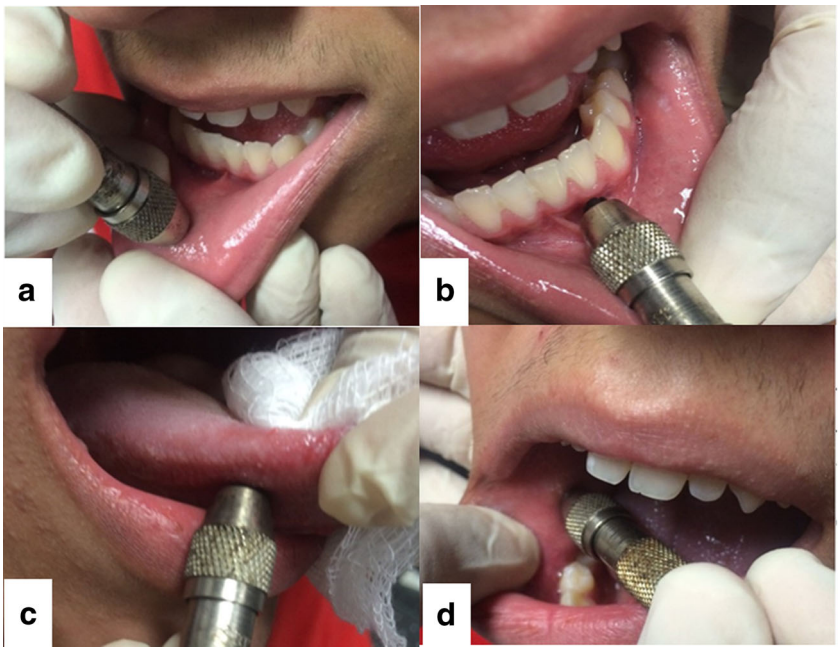

Fig. 1 Sites of spectra acquisition. **a** Lip, **b** gingiva, **c** tongue, **d** buccal mucosa

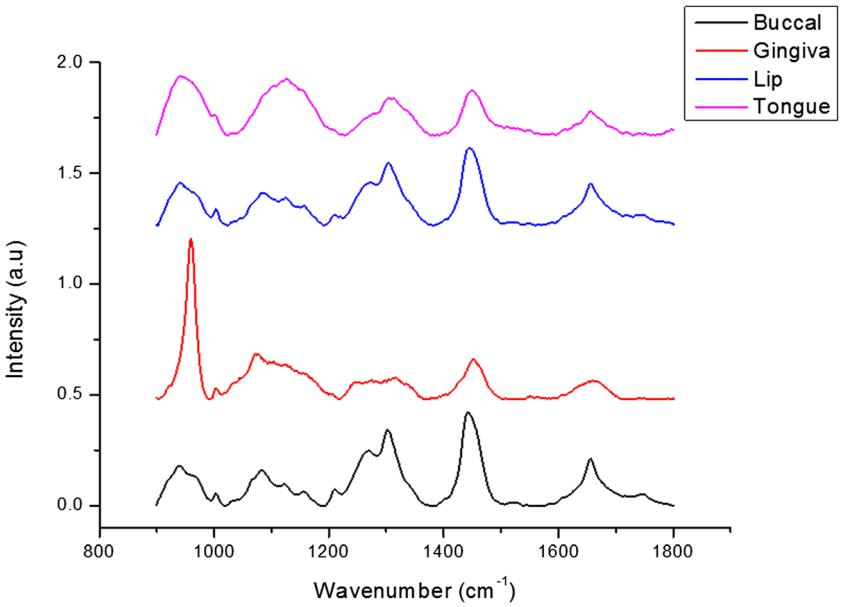

Fig. 2 Mean spectra of the subsites. Different features can be observed along almost all spectral regions in the gingiva. The lip and buccal have the highest spectral similarity

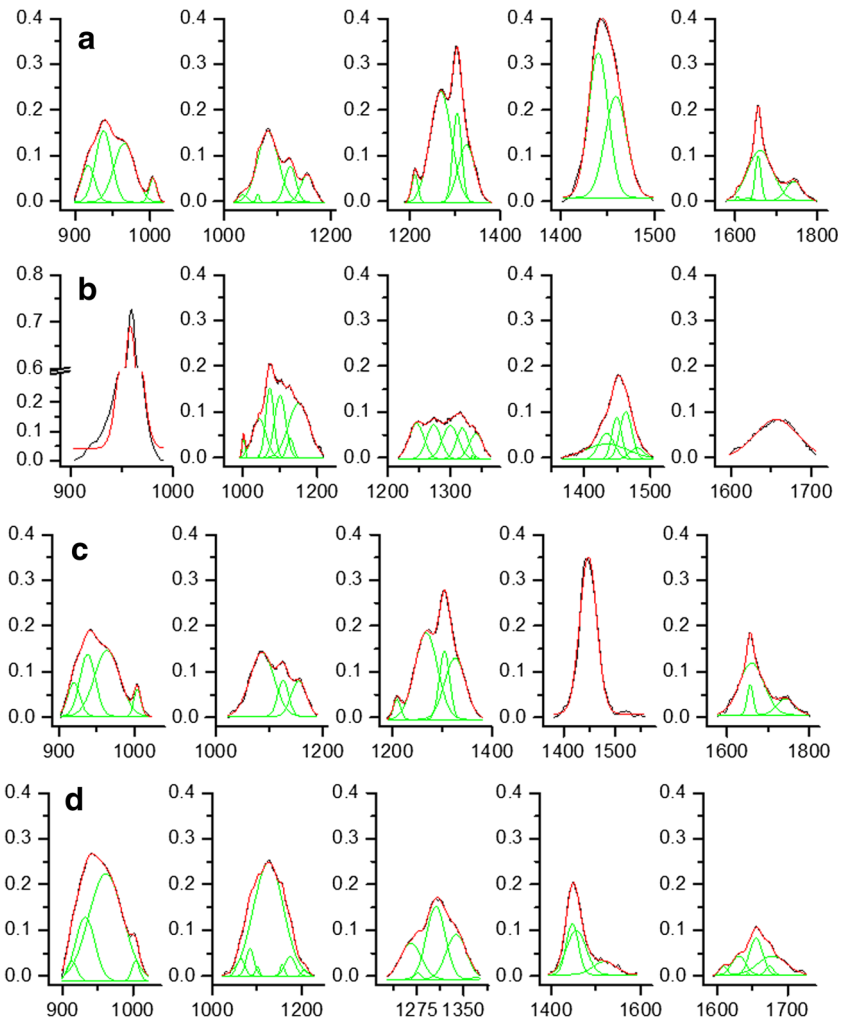

Fig. 3 Deconvolution of mean spectra of oral subsites. Red curves show measured spectra bands and green curves are the deconvolved Gaussian bands to fit measured spectra and extract tissue biochemical content

Table 1 Vibrational modes, assignment, and area of each spectral components

Buccal			Gingiva			Lip			Tongue		
Assignment	Center	Area	Assignment	Center	Area	Assignment	Center	Area	Assignment	Center	Area
Proline/Hydroxyproline	917	1.82	Phosphate	959	14.77	Proline/hydroxyproline	919	1.50	Glucose	913	0.79
	938	4.23	Phenylalanine	1003	0.43	Proline	938	3.53	Alpha helix protein	932	4.79
Hydroxyapatite	966	4.48	Phosphate	1045	4.01	Protein	964	6.08	Phosphate	961	14.27
	1004	0.64	Phosphate	1045	0.04		1004	0.69		1004	0.61
Collagen	1036	0.34	Carbonate	1073	4.20	C-C	1087	7.88		1064	0.86
Protein random confo.	1064	0.12	C-C	1100	5.14	Protein/lipid	1126	2.02	Nucleic acid phosphate	1084	1.31
Collagen carbohydrate residue	1083	7.53	C-N	1127	0.97	Proteins	1155	2.79		1101	0.22
C-C in lipid	1124	2.03	Glycogen	1150	8.41	Tryptophan and phenylalanine	1210	0.81	C-C in lipid	1125	20.51
	1156	1.81	Nucleic acid	1201	2.98	Phospholipid	1268	10.86	Proteins	1156	0.31
Tryptophan and phenylalanine	1210	1.10	Amide III	1246	2.32		1305	3.88	Nucleic acid/amino acid	1173	1.22
Collagen/lipid	1268	14.42	Ch rocking	1273	2.14	Purine	1326	7.07	Collagen content difference	1206	0.23
Collagen/lipid/nucleic acid	1304	4.86	Lipids	1300	2.29	Collagen/lipid	1448	14.33	Collagen amide III	1265	3.41
Collagen	1325	6.05	Guanine	1319	1.63		1657	1.20	Amide III	1276	0.19
	1418	4.04	Collagen/nucleic acid	1336	0.05	Amide I proteins/lipids	1660	8.59		1306	6.04
CH2 deformation lipid	1440	8.39	Nucleic acid	1341	1.51	Lipid	1743	1.85	Tryptophan	1338	4.01
Nucleic acid	1459	6.47		1434	1.67				Nucleic acid/lipid	1369	0.09
Phenylalanine/tyrosine	1606	0.09	CH2 deformation lipid	1433	2.78				Protein/ lipid	1448	3.76
Amide I	1631	0.19		1449	1.69				Nucleic acid	1456	5.28
	1656	1.75	Carbohydrates	1463	2.40					1519	1.88
Amide I proteins/ lipids	1661	8.44		1480	0.69				Nucleic acid	1609	0.25
Lipids	1743	2.10	Amide I/lipids	1657	5.70					1631	1.03
									Amide I	1644	0.02
									Amide I (collagen)/lipids	1655	1.88
										1674	0.26
										1678	2.12

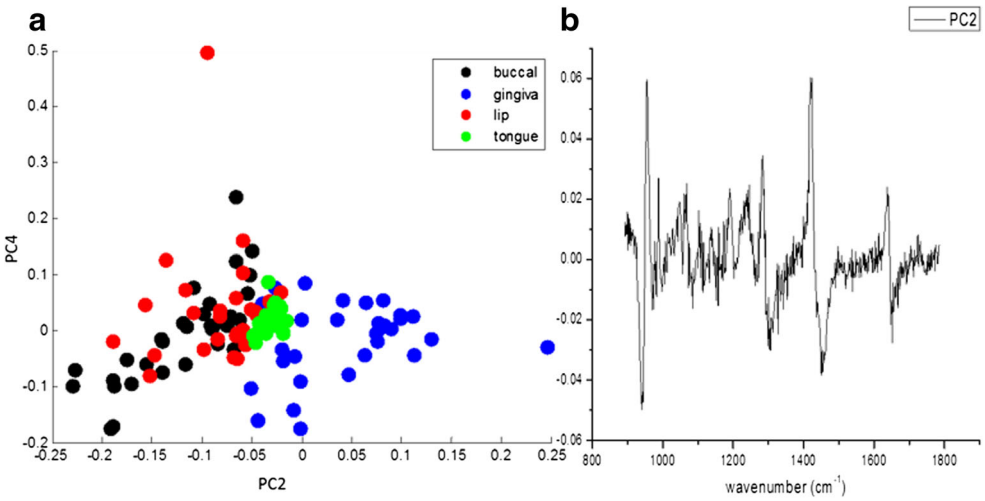

Fig. 4 PCA A—scatter plot of PC 2 × PC 4. B loading plot PC 2

composition is dominated by collagen (17%), lipids (15.5%), proteins/lipids (34%), and amino acids (4%). Phosphate (28.5%) and carbohydrate (16.5%) are major components of the gingiva, along with lipids (8%), carbonate (6%), and proteins (3.5%). The lip is rich in proteins and lipids, exhibiting 24%, 17.5%, and 31% proteins, lipids, and protein/lipid, respectively. Twenty-seven percent of the tongue composition is lipid, while collagen and proteins contribute 5 and 7%, respectively (Fig. 2). Histologically, the buccal and lip mucosa have the same tissue pattern. Since classification methods (to automatically identify tissue types) use histology results for their optimization, these results influence spectra grouping. Sahu et.al [18] have found similar features—high lipids and proteins in the buccal, with similar features in others. Importantly, they found specific collagen contribution in the tongue, which is seen in this study too. Bergholt et.al [21] also reported similar results. They used fitting of basis spectra, and found higher collagen in the lip and tongue, which is the case with our study. They found high-mineral content in the gingiva and high-lipid content in the buccal, which are also seen in this study. The mineral content found in the gingiva is due to the spectral contribution of the teeth and bone covered by the tissue. Due to the Raman penetration depth, the signal is composed also by the phosphate signal from mineralized tissues.

Table 2 Rate of classified tissues using PC-LDA with leave-one-out cross-validation. Correct classifications occur in the main diagonal of the table

Classified as	Buccal	Gingiva	Lip	Tongue
Buccal	61.29	0.00	35.48	3.23
Gingiva	3.33	83.33	6.67	6.67
Lip	25.00	0.00	40.63	34.38
Tongue	0.00	0.00	6.67	93.33

PC-LDA classification

It is well known that rapid identification of diseases based on multiple spectral characteristics can be achieved using multivariate statistical analysis. For comparison with previous studies, classification efficiency of anatomical subsites can be checked against each other to get an idea of possible differences in populations studied. With this in mind, we first used a standard method for classification in Raman spectroscopy: PC-LDA, which consists of using Principal Component Analysis (PCA) followed by linear discriminant analysis (LDA) [22–27]. PCA was performed in order to reduce the number of parameters to be used for classification by using the ones with most of the variability in the dataset. PCA generated parameters, i.e., principal components (PCs) represent data variation; with first PCs containing the maximum variation. By plotting the PCs in the Cartesian coordinate system, data variability can be visualized. Plot of PC2 and PC4 (Fig. 4A) shows a clear cluster for the gingiva and tongue, whereas the buccal and lip overlap. It is clear that PC2 results in the separation. Plot of PC2 (Fig. 4B) suggests that the bands centered at 960 cm^{-1} (phosphate), 1291 cm^{-1} (protein amide III), 1429 cm^{-1} (deoxyribose/lipids), and 1643 cm^{-1} (amide I) are the discriminating biochemicals. The clinical relevance of applying multivariate statistical analysis for spectral studies is the possibility to rapidly and clearly detect the chemical differences of each healthy tissue, and compare to another spectra of a tissue where there is some kind of doubt in the condition. In the present study, the gingiva spectra showed the high contribution of phosphate band and, if the raw spectra were isolated and analyzed, a mistake in the identification can occur. However, after PCA analysis, a clear cluster classified the gingiva and tongue (Fig. 4A).

PCA can give an idea regarding the trend of data separation, but is limited by the number of dimensions that can be plotted at a time. This can be circumvented by PC-LDA,

Table 3 Classification rates and the Area Under the Receiver Operating Characteristic curve (AUROC or ROC area) for the classification of the Raman Spectra Normalized by the Area under Spectrum (RSNAS) and the Raman Spectra Normalized by Intensity Maximum (RSNIM) using three first PCA parameters or the full spectrum. Rates of correctly classified tissues can be observed in the main diagonal of each classification matrix. They are, in most of the cases, similar between cases using full spectrum or PCA parameters with same normalization method, suggesting these parameters are sufficient to describe main features of Raman spectra used in our classification

	Classified as	K-nearest neighbors					J48 (unpruned C4.5 decision tree)					Random forest					Multilayer perceptron				
		B (%)	G (%)	L (%)	T (%)	ROC area	B (%)	G (%)	L (%)	T (%)	ROC area	B (%)	G (%)	L (%)	T (%)	ROC area	B (%)	G (%)	L (%)	T (%)	ROC area
RSNAS	Buccal (B)	74.2	0.0	22.6	3.2	0.89	61.3	3.2	35.5	0.0	0.798	71.0	0.0	29.0	0.0	0.928	77.4	0.0	22.6	0.0	0.926
	Gingiva (G)	0.0	100.0	0.0	0.0	1	0.0	93.3	6.7	0.0	0.946	0.0	100.0	0.0	0.0	1	0.0	96.7	3.3	0.0	1
	Lip (L)	43.8	6.3	40.6	9.4	0.785	31.3	6.3	53.1	9.4	0.729	21.9	0.0	62.5	15.6	0.818	18.8	0.0	65.6	15.6	0.837
	Tongue (T)	0.0	0.0	6.7	93.3	0.971	3.3	10.0	20.0	66.7	0.863	0.0	0.0	6.7	93.3	0.977	10.0	0.0	10.0	80.0	0.967
RSNAS (PCA)	Buccal (B)	64.5	0.0	32.3	3.2	0.868	51.6	0.0	45.2	3.2	0.803	51.6	0.0	45.2	3.2	0.876	74.2	0.0	22.6	3.2	0.874
	Gingiva (G)	0.0	100.0	0.0	0.0	0.995	0.0	93.3	3.3	3.3	0.953	0.0	96.7	3.3	0.0	0.999	0.0	100.0	0.0	0.0	1
	Lip (L)	40.6	9.4	37.5	12.5	0.698	31.3	6.3	43.8	18.8	0.639	40.6	0.0	40.6	18.8	0.764	34.4	3.1	46.9	15.6	0.737
	Tongue (T)	3.3	0.0	13.3	83.3	0.945	0.0	3.3	20.0	76.7	0.915	0.0	0.0	10.0	90.0	0.969	0.0	0.0	13.3	86.7	0.953
RSNIM	Buccal (B)	77.4	0.0	19.4	3.2	0.887	38.7	0.0	61.3	0.0	0.704	74.2	0.0	25.8	0.0	0.929	77.4	0.0	19.4	3.2	0.921
	Gingiva (G)	0.0	100.0	0.0	0.0	1	0.0	93.3	6.7	0.0	0.953	0.0	96.7	3.3	0.0	1	0.0	96.7	3.3	0.0	1
	Lip (L)	37.5	0.0	53.1	9.4	0.778	34.4	3.1	50.0	12.5	0.615	25.0	0.0	59.4	15.6	0.836	25.0	0.0	62.5	12.5	0.799
	Tongue (T)	0.0	0.0	10.0	90.0	0.949	0.0	16.7	20.0	63.3	0.86	0.0	0.0	6.7	93.3	0.976	0.0	0.0	16.7	83.3	0.932
RSNIM (PCA)	Buccal (B)	71.0	0.0	29.0	0.0	0.89	80.6	0.0	19.4	0.0	0.777	71.0	0.0	25.8	3.2	0.911	74.2	3.2	22.6	0.0	0.858
	Gingiva (G)	0.0	100.0	0.0	0.0	1	0.0	100.0	0.0	0.0	1	0.0	100.0	0.0	0.0	1	0.0	100.0	0.0	0.0	0.996
	Lip (L)	40.6	0.0	43.8	15.6	0.77	56.3	0.0	31.3	12.5	0.731	31.3	0.0	53.1	15.6	0.8	50.0	0.0	34.4	15.6	0.756
	Tongue (T)	3.3	0.0	6.7	90.0	0.962	6.7	0.0	0.0	93.3	0.925	3.3	0.0	10.0	86.7	0.958	0.0	0.0	10.0	90.0	0.961

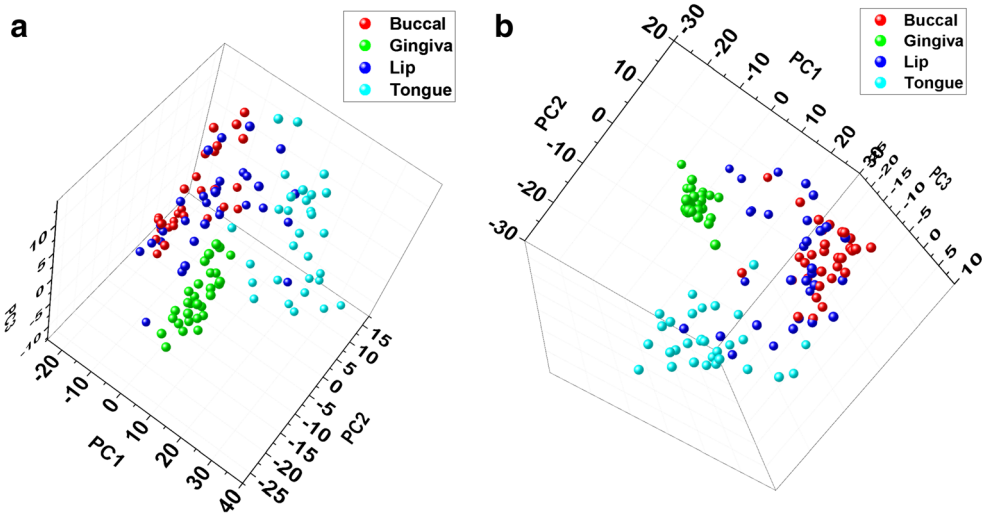

Fig. 5 PCA scores plot for Raman spectra **a** normalized by the area under the spectrum and **b** normalized by the intensity maximum. A cluster for the gingiva and tongue can be visualized in both plots. The lip group is the most heterogeneous, and a better cluster of the buccal group can be visualized in **b**

wherein important PCs are used as an input for LDA to be arranged in n-dimensional space so as to achieve maximum intergroup variability and minimum intragroup variability. The results of PC-LDA is followed by leave-one-out cross validation (LOOCV), which recreates the model by leaving one spectrum out each time.

The results of PC-LDA employing LOOCV as cross-validation method are shown in Table 2. PC-LDA LOOCV table shows 93% and 83% classification for the lip and gingiva respectively. The buccal mucosa shows 35% misclassification with the lip, while the lip shows 34 and 25% misclassification with the tongue and buccal respectively. Our data matches closely with classification results obtained by Bergholt et al. [21] despite the fact that they used eight subsites for analysis, while we used only four. With respect to the study by Sahu et al. [18], our and Bergholt et al. [21] classification efficiency was low for buccal (Sahu et.al [18] achieved a classification efficiency of 84%), but high for tongue (Sahu et.al [18] achieved classification efficiency of 81%). All three studies achieved approximately 40% classification for the lip. This further emphasizes the need for robustly defining the normal tissues spectra.

Attenuated total reflection Fourier-transform infrared spectroscopy (ATR-FTIR) study of enamel to monitor enamel erosion caused by medicaments used in the treatment of respiratory diseases was previously reported (Gomes et al. 2018). The authors combined a spectral analysis with multivariate analysis by PC-LDA to detect changes in enamel composition caused by different medicaments routinely used. Multivariate statistical analysis showed that the different medicaments were classified with efficiency from control, further highlighting the ability of ATR-FTIR to identify the degree of erosion. Thus, we can conclude that multivariate statistical analysis is a powerful tool to help in clinical diagnostic situations in both hard and soft tissues of the oral cavity.

In the present study, multivariate analysis PC-LDA allows training of statistical models that can be used to detect an incipient degree of tissue alteration by pathology of future spectrum. This will enable the clinician to obtain spectrum from patient oral tissues, immediately get valuable information of the tissue in molecular level, and have an idea of the extent of pathology stage. The dentist can use this rapid method for early diagnosis and also to monitor patient tissue conditions post-treatments such as surgeries, radiotherapy, or chemotherapy.

Table 4 Overall accuracy of the different classifiers. Random forest and multilayer perceptron classifiers allow highest accuracies

	Overall rate of correctly classified tissues			
	KNN	J48	RF	MLP
RSNAS	76.4%	68.3%	81.3%	79.7%
RSNAS (PCA)	70.7%	65.9%	69.1%	76.4%
RSNIM	79.7%	61.0%	80.5%	79.7%
RSNIM (PCA)	75.6%	75.6%	77.2%	74.0%

Accuracy improvement by other normalization and classification methods

In order to improve classification accuracy, other pre- and post-processing techniques were evaluated and compared to each other and to PC-LDA. This comparison was performed by using the full spectrum or the first three PCs after two types of normalization techniques, i.e., by the area under the curve or by its intensity maximum and four classifiers (K-nearest neighbors, unpruned C4.5 decision tree or J48, random forest, and multilayer perceptron).For the first and second types of normalization, these components account for 60.83% and

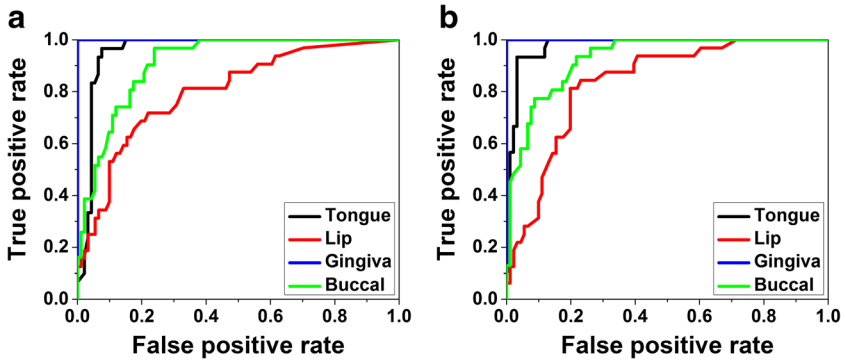

Fig. 6 ROC curves for classification of different anatomical sites when using **a** three first PCA parameters of RSNIM and **b** RSNIM parameters and random forest classifier. The area under ROC curves is very similar for each tissue type in both cases, suggesting the three first principal components are suitable to describe 515 spectral parameters without significant losses in accuracy

65.04% of the variance, respectively. Even though it may not have a high percentage of variance, it was enough to describe the main features of spectra of each anatomical site (Fig. 4 and Table 3) for classification purposes. In addition, using only three parameters allows avoidance of overfitting and a much faster classification.

The tissue classification using full spectrum led to very similar accuracies between the two normalization methods (Table 3). This happened for all classifiers and tissue types, except by the lip group when using KNN (K-nearest neighbors) and the buccal group for J48 classifier. However, when using only PCA parameters, RSNAS leads to a better classification for the lip through J48 and multilayer perceptron (MLP). On the other hand, RSNIM provides significantly higher accuracy for the buccal and tongue when using J48, and for the buccal and lip when using random forest (RF).

When comparing classification results using full spectrum or PCA parameters, a limitation in accuracy was observed for RSNAS, especially for the buccal and lip groups. For both types of normalized spectra and all classifiers, a lower accuracy was observed for the lip, which suggests features for correct classification of this type of tissue can be found in other PCs. In addition, these components may also contain important characteristics of the buccal group, in the case of RSNAS. Despite these restrictions, overall J48 classification using RSNIM was much better than using RSNAS. Moreover, the classification using RSNIM and RF achieved comparable results between full spectrum and PCA parameters (Fig. 5 and Tables 3 and 4). This indicates this combination of normalization and classification method may be the best for overall tissue discrimination results.

In terms of overall accuracy (number of correctly classified tissues/total number of tissues), MLP was the classifier with most stable results after simplifying the model using three PCs instead of the full spectrum. Also, models using RSNIM led to the most stable results among classifiers (Fig. 6). Finally, best results were achieved by random forest classifier with both highest overall accuracy and lip-identification accuracy. This overall accuracy was 8.1% higher than our results with PC-LDA and 7.8% higher than the one reported by Sahu et al.[22], suggesting our proposed classification method could improve assessment of oral normal tissues.

Any analysis to identify abnormal condition begins with rigorously defining the normal features. This logic is followed in pathology, blood, urine, and stool analysis; microbiological tests; physical parameters such as heart rate, respiratory rate, pulse rate, and so on, encompassing every aspect of medicine and disease. The same is imperative in spectroscopic diagnosis, and is the first step to make the technique acceptable in clinics. Tremendous amount of data is available on spectroscopic differences between normal and abnormal, but the literature on defining normal is comparatively scarce. In this study, we endeavor to contribute to characterization of healthy oral sub-sites. The final aim is to have a comprehensive database of spectra from different populations living in different countries, having varied range of food and oral hygiene habits. This would enable to define spectra of healthy tissues correctly and adapt multivariate analysis to encompass worldwide variations. This is of particular importance when assessing for pre-cancer changes. While cancer presents substantial biochemical change from normal, the same may not be true for early cancer biochemistry. This has been amply demonstrated by study of tobacco users, premalignant conditions in human subjects, and conditions preceding clinically detectable cancer in hamster buccal pouch model. Improper understanding of healthy spectral characteristics will invariably lead to high false-positives or negatives, undermining the application of this technology.

Conclusions

The study demonstrates that Raman spectroscopy can rapidly analyze the biochemistry of healthy oral tissues. Moreover, the study suggests the possibility of using Raman spectroscopy combined with signal processing and multivariate analysis

to identify and differentiate the oral sites in healthy conditions. Subsequently, these differences can be used to improve their contrast against pathological conditions. Further studies with larger sample sizes and different pathologies may help establish this technique as a routine tool in dental clinics.

Acknowledgments The authors would like to acknowledge Eric Marple from EmVision LLC.

Author contributions L. F. C. S. C. participated in the study design, spectra collection, manuscript writing, managed the data analysis, and paper final remarks; M.S.N. participated in manuscript writing and revision, data analysis, paper final remarks, carried out all the classification, and spectral pre- and post-processing presented in the section "Accuracy improvement by other normalization and classification methods"; T. B. participated in manuscript writing, data analysis, paper final remarks, and carried out the spectral deconvolution and fitting. L. P. M. N. carried out spectra collection, L. D. carried out spectra collection, T. O. M. participated in data analysis, R. R. carried out spectra collection, M. C. participated in the study design, A. A. M. participated in the study design, and L. E. S. S. participated in the paper final remarks.

Funding The work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP-2014/05978-1) through Luis Felipe CS Carvalho's scholarship. Luis Felipe das Chagas and Silva de Carvalho also thank FAPESP - 2018/03636-7, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) PNPD Odontologia - Universidade de Taubaté, and Centro Universitário Braz Cubas for the Scientific Initiation Program and Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES) through Tanmoy Bathacharjee's scholarship.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The study was approved by Research Ethics Committee of Universidade do Vale do Paraíba (UNIVAP) via Plataforma Brasil-Brazil (number 1132237-2015).

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. de Carvalho LFCS et al (2010) Spectral region optimization for Raman-based optical biopsy of inflammatory lesions. Photomed Laser Surg 28:S-111–S-117. https://doi.org/10.1089/pho.2009.2673
2. Carvalho LF et al (2015) Raman micro-spectroscopy for rapid screening of oral squamous cell carcinoma. Exp Mol Pathol 98:502–509. https://doi.org/10.1016/j.yexmp.2015.03.027
3. Barroso EM et al (2015) Discrimination between oral cancer and healthy tissue based on water content determined by Raman spectroscopy. Anal Chem 87:2419–2426. https://doi.org/10.1021/ac504362y
4. Bonnier F et al (2012) Analysis of human skin tissue by Raman microspectroscopy: dealing with the background. Vib Spectrosc 61:124–132. https://doi.org/10.1016/j.vibspec.2012.03.009
5. Feng X et al (2017) Raman active components of skin cancer. Biomed Optics Express 8:2835–2850. https://doi.org/10.1364/BOE.8.002835
6. Harris AT et al (2010) Raman spectroscopy in head and neck cancer. Head Neck Oncol 2:26
7. Huang Z et al (2003) Near-infrared Raman spectroscopy for optical diagnosis of lung cancer. Int J Cancer 107:1047–1052
8. Malini R et al (2006) Discrimination of normal, inflammatory, premalignant, and malignant oral tissue: a Raman spectroscopy study. Biopolymers 81:179–193
9. Oliveira AP, Bitar RA, Silveira L Jr, Zângaro RA, Martin AA (2006) Near-infrared Raman spectroscopy for oral carcinoma diagnosis. Photomed Laser Ther 24:348–353
10. Singh S, Deshmukh A, Chaturvedi P, Krishna CM (2012) Raman spectroscopy in head and neck cancers: toward oncological applications. J Cancer Res Ther 8:126
11. Singh S, Sahu A, Deshmukh A, Chaturvedi P, Krishna CM (2013) In vivo Raman spectroscopy of oral buccal mucosa: a study on malignancy associated changes (MAC)/cancer field effects (CFE). Analyst 138:4175–4182
12. Venkatakrishna K et al (2001) Optical pathology of oral tissue: a Raman spectroscopy diagnostic method. Curr Sci Bangalore 80:665–668
13. Singh S, Deshmukh A, Chaturvedi P, Krishna CM (2012) In vivo Raman spectroscopic identification of premalignant lesions in oral buccal mucosa. J Biomed Opt 17:1050021–1050029
14. de Paula Campos C et al (2017) Fluorescence spectroscopy in the visible range for the assessment of UVB radiation effects in hairless mice skin. Photodiagn Photodyn Ther 20:21–27. https://doi.org/10.1016/j.pdpdt.2017.08.016
15. Saito Nogueira M, Kurachi C (2016) Assessing the photoaging process at sun exposed and non-exposed skin using fluorescence lifetime. Spectroscopy 9703:97031W. https://doi.org/10.1117/12.2209690
16. Saito Nogueira, M. et al. (2016) Evaluation of actinic cheilitis using fluorescence lifetime spectroscopy. Proceedings of the SPIE 9703(97031U):6. https://doi.org/10.1117/12.2209689
17. D'Almeida CDP, Campos C, Saito Nogueira M, Pratavieira S, Kurachi C (2015) Time-resolved and steady-state fluorescence spectroscopy for the assessment of skin photoaging process. Proc. SPIE 9531, Biophotonics South America, 953146. https://doi.org/10.1117/12.2180975
18. Sahu A, Deshmukh A, Hole AR, Chaturvedi P, Krishna CM (2016) In vivo subsite classification and diagnosis of oral cancers using Raman spectroscopy. J Innov Opt Health Sci 9:1650017. https://doi.org/10.1142/s1793545816500176
19. Kumar P, Bhattacharjee T, Ingle A, Maru G, Krishna CM (2016) Raman spectroscopy of experimental oral carcinogenesis: study on sequential cancer progression in hamster buccal pouch model. Technol Cancer Res Treat 15:NP60–NP72
20. Kumar P et al (2016) Raman spectroscopy in experimental oral carcinogenesis: investigation of abnormal changes in control tissues. J Raman Spectrosc 47:1318–1326. https://doi.org/10.1002/jrs.4977
21. Bergholt MS, Zheng W, Huang Z (2012) Characterizing variability in in vivo Raman spectroscopic properties of different anatomical sites of normal tissue in the oral cavity. J Raman Spectrosc 43:255–262. https://doi.org/10.1002/jrs.3026
22. Crow P et al (2005) The use of Raman spectroscopy to differentiate between different prostatic adenocarcinoma cell lines. Br J Cancer 92:2166–2170. https://doi.org/10.1038/sj.bjc.6602638
23. Feng S et al (2010) Nasopharyngeal cancer detection based on blood plasma surface-enhanced Raman spectroscopy and

multivariate analysis. Biosens Bioelectron 25:2414–2419. https://doi.org/10.1016/j.bios.2010.03.033

24. Crow P et al (2005) Assessment of fiberoptic near-infrared raman spectroscopy for diagnosis of bladder and prostate cancer. Urology 65:1126–1130. https://doi.org/10.1016/j.urology.2004.12.058
25. Teh SK et al (2008) Diagnostic potential of near-infrared Raman spectroscopy in the stomach: differentiating dysplasia from normal tissue. Br J Cancer 98:457–465. https://doi.org/10.1038/sj.bjc.6604176
26. Pires L, Nogueira MS, Pratavieira S, Moriyama LT, Kurachi C (2014) Time-resolved fluorescence lifetime for cutaneous melanoma detection. Biomed Opt Express 5:3080–3089. https://doi.org/10.1364/BOE.5.003080
27. Cosci A et al (2016) Time-resolved fluorescence spectroscopy for clinical diagnosis of actinic cheilitis. Biomed Opt Express 7:4210–4219. https://doi.org/10.1364/BOE.7.004210