Analyst

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Near-infrared Raman spectroscopy for assessing biochemical changes of cervical tissue associated with precarcinogenic transformation

Shiyamala Duraipandian, B.E., Jianhua Mo, Ph.D., Wei Zheng, Ph.D., Zhiwei Huang, Ph.D*.

Optical Bioimaging Laboratory, Department of Biomedical Engineering, Faculty of Engineering, National University of Singapore, Singapore 117576

***Correspondence to**:

Dr. Zhiwei Huang

Optical Bioimaging Laboratory,

Department of Biomedical engineering,

Faculty of Engineering,

National University of Singapore,

9 Engineering Drive 1,

Singapore 117576

Tel: +65 6516 8856,

Fax: +65 6872 3069,

E-mail: biehzw@nus.edu.sg

Key words: Raman spectroscopy, cervical precancer, biochemical modeling, near-infrared

Abstract

Raman spectroscopy measures the inelastically scattered light from tissue that is capable of identifying native tissue biochemical constituents and their changes associated with disease transformation. This study aims to characterize the Raman spectroscopic properties of cervical tissue associated with multi-stage progression of cervical precarcinogenic sequence. A rapid-acquisition fiber-optic near-infrared (NIR) Raman diagnostic system was employed for tissue Raman spectral measurements at 785 nm excitation. A total of 68 Raman spectra (23 benign, 29 low-grade squamous intraepithelial lesions (LSIL) and 16 high grade squamous intraepithelial lesions (HSIL)) were measured from 25 cervical tissue biopsy specimens, as confirmed by colposcopy-histopathology. The semi-quantitative biochemical modeling based on the major biochemicals (i.e., DNA, proteins (histone, collagen), lipid (triolein) and carbohydrates (glycogen)) in cervical tissue uncovers the stepwise accumulation of biomolecular changes associated with progressive cervical precarcinogenesis. Multi-class partial least squares-discriminant analysis (PLS-DA) together with leave-one tissue site-out, cross-validation yielded diagnostic sensitivities of 95.7%, 82.8% and 81.3%; specificities of 100.0%, 92.3% and 88.5%, respectively, for discrimination among benign, LSIL and HSIL cervical tissues. This work suggests that the Raman spectral biomarkers identified have the potential to be used for monitoring the multi-stage cervical precarcinogenesis, forming the foundation of applying NIR Raman spectroscopy for early diagnosis of cervical precancer *in vivo* at the molecular level.

1. Introduction

Cervical cancer is the third most common cancer in women worldwide.[1] It follows the stepwise histological progression from low-grade squamous intra-epithelial lesions (LSIL), high-grade squamous intra-epithelial lesions (HSIL) to invasive cancer. Cervical cancer can be prevented or treated effectively if it can be detected at its dysplastic precursor stage (i.e., LSIL, HSIL). However, the standard cytology-based screening tests (i.e., Papanicolaou (Pap) smears) and the white-light colposcopy follow-up of an abnormal Pap smear could not achieve high detection sensitivity and specificity concurrently for the diagnosis of cervical precancer.[2-6] For instance, the Pap smear tests have good diagnostic specificity of 93% (95% confidence interval (CI): 89% - 97%), but comparatively a low sensitivity of 57% (95% CI: 38% - 76%).[5] On the contrary, the colposcopy has a high sensitivity of ~96%, but this is tempered by a poor detection specificity of ~48% for the diagnosis of preneoplastic cervix.[2, 5, 6] Evaluation of colposcopy-directed punch biopsies of cervix still serves as the gold standard diagnostic approach for cervical cancer and precancer diagnosis. Inter-colposcopist disagreement in the selection of biopsy sites is, however, a major source of errors that limits the diagnostic efficacy of colposcopy-guided histopathology. The high sensitivity for colposcopy-guided histopathology could be achieved by increasing the number of biopsies taken from acetowhitening areas based on visual diagnostic criterion.[7] Furthermore, histopathology is based on visual inspection of gross morphological changes such as epithelial thickening, vertical extension of abnormal cells in the cervical epithelium, without biomarker/biomolecular information.[7] Gross cutting and processing of formalin fixed specimen in histopathology may alter the biochemical state of tissue, causing sampling and interpretive errors.[8] Therefore, development of an objective optical diagnostic technique that could capture biochemical fingerprints associated with cervical precancer is of paramount clinical importance

for achieving early cancer diagnosis and targeted biopsies of suspicious lesion sites at colposcopy.

In the past few decades, the development of advanced optical modalities, especially near-infrared (NIR) Raman spectroscopy, is a topic of continuous research interest for early cancer diagnosis.[9-17] Raman spectroscopy is a unique molecular probe, capable of fingerprinting biochemical activities occurring in the tissue that reflects different histotypes with molecular and clinical heterogeneity. For instance, the diagnostic capability of Raman spectroscopy for early cancer diagnosis in a number of organs including cervix has been demonstrated with success.[9, 10, 12, 13, 16, 17] However, the clinical merit of NIR Raman spectroscopy for separating progressive cervical precarcinogenic sequence (i.e., LSIL, HSIL) has yet to be investigated in detail. Precise ascertainment of different stages of cervical precancer (e.g., LSIL and HSIL) is of significant medical relevance because LSIL has the potential to either regress to normal or progress to HSIL, but HSIL carries the substantial risk of further progressing to invasive cancer. Thus accurate grading of precancer stages could greatly reduce the prevalent occurrence of overcalled LSIL lesions or missed HSIL lesions in colposcopy/histopathology diagnosis and may influence the treatment strategies (e.g., follow-up for LSIL patients or complete excision of lesions for HSIL patients) for patients identified with cervical precancer, moving the NIR Raman spectroscopy closer to the patient's bedside. In this study, we aim to distinguish different stages of precancerous lesions in the cervix at the molecular level using the developed fiber-optic Raman spectroscopy. Multivariate algorithms based on partial least squares-discriminant analysis (PLS-DA) together with leave-one tissue site-out, cross-validation was utilized for effective discriminate among benign, low-grade and high-grade cervical precancer. We further developed a semi-quantitative biochemical model using representative basis biochemical Raman spectra for

assessing the biochemical makeup of cervical tissue and its changes with multi-step precarcinogenesis (i.e., LSIL and HSIL).

2. Materials and Methods

2.1. Raman instrumentation

A rapid NIR Raman spectroscopy system developed for cervical tissue spectral measurements has been described in detail elsewhere.[12] Briefly, the developed Raman spectroscopy system consists of an external cavity spectrum stabilized 785 nm diode laser (maximum output: 300 mW, B&W TEK Inc., Newark, DE, USA), a transmissive imaging spectrograph (Holospec f/1.8, Kaiser Optical Systems, Ann Arbor, MI, USA) equipped with a liquid nitrogen-cooled (-120°C), NIR-optimized, back-illuminated and deep-depletion charge-coupled device (CCD) camera (1340×400 pixels at 20×20 μm/pixel; Spec-10: 400BR/LN, Princeton Instruments, Trenton, NJ, USA), and a specially designed bifurcated fiber-optic Raman probe that delivers the 785 nm laser light to the tissue and collects the scattered tissue Raman signal. The 1.8 mm internally filtered fiber-optic Raman probe comprises of a bundle of 200 μm low-OH fused-silica fibers (NA=0.22) in a 32-around-1 configuration (i.e., 32 collection fibers surrounding the central laser light delivery fiber) with two stages of optical filtering. The filtering modules are incorporated at the proximal and distal ends of the probe for optimizing the collection of tissue Raman signals, while reducing the interference of Rayleigh scattered light, fiber fluorescence and silica Raman signals. The 785 nm laser light with tissue spot size of ~0.2 mm and irradiance of 1.5 W/cm^2 was utilized for tissue Raman spectral measurements. The tissue Raman spectra were acquired in the fingerprint region (800 - 1800 cm^{-1}) within an integration time of 1 sec, and the spectral resolution of ~ 9 cm^{-1}. The atomic spectral emission lines produced by mercury-argon spectral

calibration lamps (HG-1 and AR-1, Ocean Optics Inc., Dunedin, FL, USA) are used for wavelength calibration. The wavelength-calibrated Raman spectra are also corrected for wavelength-dependence of the system using a tungsten-halogen calibration lamp (RS-10, EG&G Gamma Scientific, San Diego, CA, USA). The entire system can be controlled by a personal computer using in-house developed graphical user interface (GUI) under Matlab environment.

2.2. Tissue samples

A total of 25 cervical tissue biopsies were obtained from the 15 patients recruited who were referred to the colposcopy due to abnormal Pap smears in the O&G clinic at the National University Health System (NUHS), Singapore. The recruited patients were in the ages between 18 to 70 years and were not pregnant. The study protocol was approved by the Institutional Review Board (IRB) of the National Healthcare Group (NHG) of Singapore. An informed consent form was signed by each patient undergoing colposcopy, permitting the investigative use of cervical tissue specimens. The tissue samples were subjected to Raman spectroscopic measurements immediately after the biopsies (typically within 30 min) with no pretreatment. The tissue specimens were kept moist with physiologic saline solution (pH=7.4) during Raman measurements. Note that incident laser light with a beam size of 0.2 mm was focused onto the cervical tissue specimens of ~3 mm x 3 mm x 2 mm in size to mimic the *in vivo* clinical measurements. The tissue surfaces where Raman spectral measurements were carried out were then marked and stained for pathology. Multiple Raman spectral measurements were taken from the same site of each tissue specimen and they were averaged out to one spectrum to include inter- and/or intra-tissue variability for data analysis. As a result, a total of 23 benign, 29 LSIL and 16 HSIL Raman spectra were measured from 25 cervical biopsy samples obtained from 15

patients (5 benign, 6 LSIL and 4 HSIL). After the spectral acquisition, biopsy samples were fixed in 10% formalin solution and then submitted for histopathology confirmation. The histopathology results confirmed that out of 25 cervical tissue specimens, 8 were benign, 10 were LSIL and 7 were HSIL. The benign tissue samples include chronic cervicitis and squamous metaplasia.

2.3. Biochemical modeling

The measured tissue Raman spectra were fit to a linear combination model of major tissue biochemical constituents (e.g., DNA, proteins, lipids and carbohydrates, etc.) in cervical tissue based on linear least-squares method with non-negative constraints (i.e., non-negativity-constrained least squares minimization (NNCLSM) fitting).[18-21] The developed NNCLSM modeling can be expressed as,

$$\text{Min } E(c) = ||c.\ S - d||^2, \text{ where } c \geq 0,$$

where c is the matrix of concentration coefficients to be predicted, S is the matrix of spectral components of biochemicals, d is the measured tissue spectrum of different pathologies and E is the residuals. The developed semi-quantitative biochemical diagnostic model best estimates the measured tissue Raman spectra with a very small fit-residual (E), indicating that the selected biochemicals largely represent the biochemical constituents of cervical tissue. The fitting coefficients therefore yield the relative concentration profiles of each biochemical component making up the lesions, providing an insight into tissue biochemical changes associated with disease progression. The coefficients were further constrained to a coefficient sum of 100%. In this study, we have initially measured over 50 major cervical tissue biochemicals (e.g., glycogen, phosphatidylcholine, triolein, cholesterol, actin, histones, collagen type I, albumin, DNA, RNA,

etc.). Exhaustive NNCLSM modeling with different combinations of major biochemicals based on priori insight of inter-/intra-cellular constituents of cervical tissue were carried out to determine the biomolecules best representing the measured cervical tissue Raman spectra. As a result, our NNCLSM modelling indicated that the most significant biochemicals such as DNA, proteins (histone, collagen), lipid (triolein) and carbohydrates (glycogen) make up the most basic building blocks of normal and progressive precancer cervical tissue (Figure 2).[17-22] For instance, abundant amount of glycogen is traced in the stratum spinosum zones of normal cervical squamous epithelium.[22] DNA represents the nucleic acids within the nucleus[21] and histones denote the DNA-binding proteins found in the chromatin of cervical cell nuclei that aid in DNA's packaging.[21, 23] DNA methylation and histone modification profiles have also been shown as promising candidates for improving the detection of early cervical cancer.[23] Triolein indicates the lipid droplets[24] and collagen is a representative of main structural proteins in extracellular matrix of normal cervical epithelium.[25, 26] Using the above representative Raman-active tissue biochemical components, we modeled (NNCLSM) the Raman spectroscopic profiles of benign, LSIL and HSIL cervix to analyze the relative changes of endogenous biochemical compositions associated with cervical precancerous transformation.

2.4. Data preprocessing and multi-class diagnostics

The Raman signal emanating from tissue is inherently weak, and occurs concurrently with prominent tissue autofluorescence background. Subsequent spectral preprocessing such as CCD dark-noise subtraction, system spectral response calibration, smoothing (moving average, five pixel window, Savitzky-Golay), tissue autofluorescence background subtraction (5^{th} order polynomial fit), and normalization were performed on the measured tissue spectra. The spectral

preprocessing enables a better comparison of spectral shapes and relative Raman band intensities among different tissue pathologies (i.e., benign, LSIL and HSIL). Partial least squares-discriminant analysis (PLS-DA) was further applied on the preprocessed Raman spectra to develop multi-class diagnostic algorithms for tissue pathology diagnosis and characterization. [23,24] The partial least squares based discriminant analysis is particularly useful for clinical diagnosis due to the inherent multicollinearity characteristics of biological data.[25] PLS-DA follows the principle of principal component analysis (PCA), but further rotates the components (latent variables (LVs)) to achieve maximum group separation. Consequently, PLS-DA correlates the variations in the dataset with the response variable to explain the diagnostically relevant spectral variations associated with cervical precarcinogenesis in the first few LVs.[26] The performance of the multi-class probabilistic PLS-DA model was validated in an unbiased manner using leave-one tissue site-out, cross-validation for differentiation among benign cervix, LSIL and HSIL. All the above mentioned spectral analysis was conducted using the in-house developed program running in the Matlab environment.

3. Results

Figure 1a shows the mean Raman spectra ± 1 standard deviations (SD) acquired from benign cervix (n=23), LSIL (n=29) and HSIL (n=16), respectively. The measured tissue Raman spectrum is typically a mixture of complex spectral signatures of inherent tissue biochemicals that are overlapped significantly. It contains information on the fundamental vibrational modes of tissue biomolecules and found at the following peak positions with tentative biomolecular assignments: 849 cm^{-1} (ring breathing mode of tyrosine and CCH aromatic deformation of glycogen), 933 cm^{-1} (glycogen CCH deformation and C-C stretching of proline, valine), 1004

cm^{-1}(symmetric breathing of phenylalanine), 1085 cm^{-1} (C-C stretching of phospholipids), 1125 cm^{-1} (C-C skeletal stretch in lipids), 1255 cm^{-1} and 1285 cm^{-1} (amide III – β sheet), 1339 cm^{-1} (CH_3CH_2 wagging mode of collagen/nucleic acid polynucleotide chain), 1450 cm^{-1} (CH_2 deformation in lipids), 1578 cm^{-1} (nucleic acids), 1658 cm^{-1} (amide I (α-helix)) and 1738 cm^{-1} (C=O stretching of lipids).[9, 25, 27, 28] The difference spectra ± 1 SD (Figure 1b) clearly shows the increase or decrease of distinctive molecular markers in cervical tissue related to precancer progression. For instance, the dysplastic tissue (LSIL and HSIL) showed increased Raman signals at 1004, 1339, 1578 and 1738 cm^{-1}, but exhibited much lower signal at 849, 933, 1085, 1255, 1285, and 1658 cm^{-1} (unpaired Student's t-test ($p < 0.05$)) compared to benign tissue, revealing the prominent biomolecular changes associated with premalignant progression.

To quantify the relative fractions of these prominent tissue biochemicals (i.e., proteins, lipids, DNA and carbohydrates (Figure 2)) underpinning different stages of cervical precarcinogenic lesions, a semi-quantitative NNCLSM Raman spectroscopic model of cervix was developed. A linear superposition of Raman spectra of selected basis biochemicals in NNCLSM model approximates the biochemical compositions of cervix by best fitting the measured tissue spectra with optimum fitting coefficients and minimum residuals (fit-residuals of <10%). Figure 3 shows the mean measured Raman spectra of benign cervix, LSIL and HSIL together with their corresponding reconstructed Raman spectra and very small fit residuals (measured - reconstructed). The obtained small residuals (residual variations for 3(a)-3(c): ±5E−4) is possibly due to the reason that the Raman spectral modeling may not entirely incorporate all the biochemicals present in the tissue because biological tissue constitutes of a complex mixture of various biochemical signatures. Figure 4 shows the bar-chart of mean fit coefficients ± 1 SD together with their corresponding p-values, unveiling the stepwise

accumulation in the concentration of distinctive biochemical components underpinning the progressive cervical precarcinogenic lesions. Overall, the precancerous tissue exhibited higher fitting coefficients for DNA, histones and triolein, while the coefficients were lower for collagen and glycogen. The results suggest that NIR Raman spectroscopy is able to assess the biochemical alterations in cervices associated with multi-stage progression of cervical precarcinogenesis.

The biochemical differences observed in the NIR Raman spectra of benign cervix, LSIL and HSIL were further explored using multivariate PLS-DA multi-class diagnostic algorithm to enable rapid tissue diagnosis and characterization. The measured tissue Raman spectral dataset was first mean-centered to remove magnitude dependence. The outlier spectra (i.e., blood interference, white light interference) were further discarded using PCA coupled with Hotelling's T^2 and Q-residuals statistics.[29] Following the spectral quality verification, the PLS-DA multi-class diagnostic model was developed using optimum number of components (1 LV) (Figure 5), corresponding to minimum leave-one tissue site-out cross validation error. The selected first LV accounts for 19.0% and 22.1% of total Raman spectral variations in X and Y directions, respectively, representing the prominent spectral variations around the major tissue Raman peak positions (849, 933, 1003, 1098, 1182, 1248, 1276, 1313, 1465, 1635 and 1667 cm^{-1}). The posterior probability values were further estimated and shown as a 2-D ternary scatter plot in Figure 6. The diagnostic results for Raman spectra acquired from benign cervix, LSIL and HSIL using PLS-DA together with leave-one tissue site-out, cross-validation method were summarized in Table 1. The multi-class model provided diagnostic sensitivities of 95.7%, 82.8%, and 81.3%, and specificities of 100.0%, 92.3%, and 88.5%, for the diagnosis of benign cervix, LSIL and HSIL, respectively, demonstrating the clinical utility of Raman spectroscopy to grade multi-stage progression of cervical precancer. The receiver operating characteristic (ROC) curves (Figure 7)

for each class (i.e., one-against-all) were also generated to determine correct and incorrect classifications for all tissue categories. The areas under the ROCs are 0.97, 0.87 and 0.89, respectively, for multi-class discrimination among benign cervix, LSIL and HSIL, further affirming the diagnostic efficacy of NIR Raman spectroscopy for differentiating different stages of cervical precancer.

4. Discussion

The dynamic progression of cervical carcinogenesis encompasses a cascade of events beginning with human papillomavirus (HPV) infection, viral persistence, gradual progression to LSIL and HSIL precancerous lesions and invasion.[30, 31] The definite diagnosis and grading of HPV-induced cervical precancerous lesions can be highly challenging using standard colposcopy-guided histopathology. The histopathology-based grading of cervical precancerous lesions relies on visual inspection of morphological alterations within the cells and tissues. For instance, LSIL is a morphological correlate of acute HPV infection.[7] The acceptable level of agreements is thereupon poor for multi-class prediction (i.e., benign, low-grade and high-grade precancer) compared to an acceptable level of agreements among pathologists for two major comparative group classifications (i.e., cancer, non-cancer).[8] Nevertheless, identification, surveillance and treatment at the early stage (e.g., complete excision of HSIL lesions) are the key to improving survival rate of the patients. Hence, the development of new optical diagnostic methods that can guide colposcopists to objectively detect benign cervix from precancerous lesions and further precisely grade different precancer stages is essential for cervical cancer prevention.

In this study, we utilized NIR Raman spectroscopy together with biochemical modeling for probing Raman spectral biomarkers to distinguish benign cervix from early (LSIL) and late

(HSIL) stages of cervical preoncogenesis. The distinctive and reproducible changes (e.g., 849, 933, 1004, 1085, 1125, 1255, 1285, 1339, 1450, 1578, 1658 and 1738 cm^{-1}) (Figure 1b) observed in the Raman spectral shapes and intensities among benign, LSIL and HSIL can be instantly correlated with the changes of tissue biochemicals associated with dysplastic transformation. For instance, diminished Raman intensities for collagen/glycogen (849 and 933 cm^{-1}), amide I (1658 cm^{-1}) and amide III (1255 and 1285 cm^{-1}) bands, and enhanced Raman peaks attributed to nucleic acids (1339 and 1578 cm^{-1}), phospholipids and mixture of lipid/protein bands (1339, 1450 and 1738 cm^{-1}) were observed with multi-step progression of cervical preoncogenesis. We further modeled (i.e., NNCLSM fitting) the cervical Raman spectra of different pathologies using the major biochemicals present in the cervical tissue to understand the possible mechanism behind these endogenous biochemical changes associated with precancer progression. To develop this Raman spectroscopic model, we also presumed that the Raman intensity scales linearly with the concentration of biochemical components in the tissue. The resulting model coefficients (Figure 4) showed similar variations observed in the difference spectra (Figure 1b) such as reduced glycogen and collagen, increased histones, nucleic acids and lipids with the progression of cervical precancer. For instance, the diminished glycogen content in dysplastic cervix is due to loss of cervical epithelial cell maturation[25, 28] caused by defective glycogen synthesis.[32] The defective glycogen synthesis may be associated with the low activities of key glyconeogenic enzymes phosphoglucomutase and glycogen synthetase in dysplastic tissue.[32] The amount of structural proteins (e.g., collagen) is higher in benign cervical tissue that likely due to periodic inflammation in normal cervix associated with fibrotic changes.[25, 28, 33] The elevated level of lipids with progressive dysplasia could be attributed to the presence of more cytoplasmic lipid droplets in precancer tissue compared to benign tissue.[24] The pronounced fitting

coefficients for proteins (e.g., histones) and DNA with progressive dysplasia may signify the increased epithelial cell proliferation and nuclear-to-cytoplasm ratio, respectively.[17, 33] Note that DNA methylation and histone modification profiles are the important diagnostic and prognostic biomarkers for cervical cancer.[23] Overall, the biomolecular modeling (Figure 4) showed significant stepwise increase or decrease of biochemical changes with progressive cervical dysplasia, and the changes were particularly substantial for DNA, histones and collagen. Note that the tissue biopsy specimens and *in vitro* biomolecular modeling in this study may not fully reflect true *in vivo* conditions such as vascular information, inter- and intra -molecular interactions, concentration and distribution of various Raman-active biochemicals across a single layer of tissue or different layers, etc.[34] Nevertheless, the developed biochemical model still provides us the gross estimation of the most important endogenous biochemicals that constitute the structural matrix of cervical tissue.

Following the assessment of cervical tissue biochemical compositions and their changes underpinning different grades of cervical precancer, PLS-DA algorithm was further utilized to evaluate the diagnostic efficacy of Raman spectroscopy for cervical precancer classification. We first generated PLS-DA model against each class (one-against-one). The multivariate model provided sensitivities of 62.1% (18/29) and 87.5% (14/16), and specificities of 87.0% (20/23) and 82.6% (19/23) for differentiating LSIL versus benign, and HSIL versus benign, respectively. The PLS-DA model further classified different grades of cervical precancer (i.e., LSIL versus HSIL) with a diagnostic sensitivity of 87.5% (14/16), and a specificity of 89.7% (26/29). One notes that the diagnostic sensitivity (62.1% (18/29)) is reduced for discriminating LSIL from benign lesions. The similar diagnostic difficulty of separating benign lesions from LSIL also notably appears among pathologists. This could be due to the following reasons: The benign

lesions (e.g., inflammation, squamous metaplasia) have the features such as mild nuclear enlargement, multinucleation, even chromatin coarsening, cellular disorganization, and maturation disturbances; These mild changes occurring in a background of benign lesions may lack true nuclear atypia compared to LSIL, but could be overdiagnosed as LSIL.[35] Another challenge faced in histopathologic/colposcopic evaluation of cervical precancerous lesions is that the identification of HSIL from benign (e.g., squamous metaplasia) and LSIL,[35, 36] is prone to subjective diagnosis among ccolposcopistsloposcopist/pathologists. For instance, squamous metaplasia is the cytologic mimics of HSIL, but the nuclear atypical changes hardly rise to the level of atypia found in HSIL.[36] Similarly, LSIL and HSIL lesions also have same atypical nuclear features mainly comprising hyperchromasia, irregular chromatin distribution and membrane contour irregularity, but the severity is more in HSIL.[36, 37] NIR Raman spectroscopy coupled with one-against-one algorithm separates benign from HSIL as well as differentiates among different grades of cervical precancer with notable results. But in real clinical conditions, patients may have multiple complications, substantiating the need for identifying all clinically important subgroups concurrently. To simulate the true clinical scenario, we further developed multi-class PLS-DA diagnostic model (one-against-all) to differentiate benign, low-grade and high-grade cervical precancer simultaneously. The generated multi-class PLS-DA model provided diagnostic sensitivities of 95.7%, 82.8% and 81.3%; specificities of 100.0%, 92.3% and 88.5%, respectively, for discrimination among benign, low-grade and high-grade precancerous cervix (Figure 6), illustrating the diagnostic capability of NIR Raman spectroscopy together with PLS-DA algorithms for identifying benign lesions from precancerous lesions (LSIL and HSIL) and grading of cervical precancer. Thus, NIR Raman spectroscopy coupled with PLS-DA multi-class diagnostic modeling can advantageously be utilized as a clinical diagnostic tool for

monitoring the progression of cervical precancer, pushing the Raman optical diagnostic technique closer to patient bedside.

In summary, NIR Raman diagnostic system was utilized for tissue Raman spectral measurements from benign, LSIL and HSIL cervix. Significant Raman biochemical differences were observed among benign, LSIL and HSIL cervical tissues. Raman spectroscopy in conjunction with semi-quantitative biochemical model of cervical tissue provides new insights into biomolecular origins responsible for prominent tissue Raman spectral features and their variability with progressive dysplasia. The PLS-DA based multi-class diagnostic model provides promising discrimination among benign, LSIL, and HSIL cervical lesions, laying the foundation for early diagnosis of cervical precancer *in vivo* at the molecular level.

5. Acknowledgements

This research was supported by the National Medical Research Council (NMRC), the National Research Foundation (NRF), and the Academic Research Fund from the Ministry of Education, Singapore.

References

1.D. M. Parkin, P. Pisani, and J. Ferlay, "Global cancer statistics," *CA: Cancer J. Clin.* **49**(1), 33-64 (1999).

2.M. F. Mitchell, D. Schottenfeld, G. Tortolero-Luna, S. B. Cantor, and R. R. Richards-Kortum, "Colposcopy for the diagnosis of squamous intraepithelial lesions: a meta-analysis," *Obstet. Gynecol.* **91**(4), 626-631 (1998).

3.K. Nanda, D. C. McCrory, E. R. Myers, L. A. Bastian, V. Hasselblad, J. D. Hickey, and D. B. Matchar, "Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: A systematic review," *Ann. Intern. Med.* **132**(10), 810-819 (2000).

4.J. Benavides, S. K. Chang, S. Park, R. R. Richards-Kortum, N. Mackinnon, C. Macaulay, A. Milbourne, A. Malpica, and M. Follen, "Multispectral digital colposcopy for in vivo detection of cervical cancer," *Opt. Express* **11**(10), 1223-1236 (2003).

5.M. Arbyn, R. Sankaranarayanan, R. Muwonge, N. Keita, A. Dolo, C. G. Mbalawa, H. Nouhou, B. Sakande, R. Wesley, T. Somanathan, A. Sharma, S. Shastri, and P. Basu, "Pooled analysis of the accuracy of five cervical cancer screening tests assessed in eleven studies in Africa and India," *Int. J. Cancer* **123**(1), 153-160 (2008).

6.K. Limmer, G. LoBiondo-Wood, and J.Dains, "Predictors of cervical cancer screening adherence in the United States: a systematic review." J Adv Pract Oncol. 5(1):31-41 (2014).

7.M. Schiffman, N. Wentzensen, S. Wacholder, W. Kinney, J. C. Gage, and P. E. Castle, "Human papillomavirus testing in the prevention of cervical cancer," *J. Natl. Cancer Inst.*, (2011).

8.N. Stone, C. A. Kendall, J. Smith, P. Crow, and H. Barr, "Raman spectroscopy for identification of epithelial cancers," *Faraday Discuss.* **126**, 141-157 (2004).

9.Z. Huang, A. McWilliams, H. Lui, D. I. McLean, S. Lam, and H. Zeng, "Near-infrared Raman spectroscopy for optical diagnosis of lung cancer," *Int. J. Cancer* **107**(6), 1047-1052 (2003).

10.S. K. Teh, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, and Z. Huang, "Diagnostic potential of near-infrared Raman spectroscopy in the stomach: differentiating dysplasia from normal tissue," *Br. J. Cancer* **98**(2), 457-465 (2008).

11.S. K. Teh, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, and Z. Huang, "Near-infrared Raman spectroscopy for gastric precancer diagnosis," *J. Raman Spectrosc.* **40**(8), 908-914 (2009).

12.A. Mahadevan-Jansen, M. F. Mitchell, N. Ramanujam, A. Malpica, S. Thomsen, U. Utzinger, and R. R. Richards-Kortum, "Near-infrared Raman spectroscopy for in vitro detection of cervical precancers," *Photochem. Photobiol.* **68**(1), 123-132 (1998).

13.E. Widjaja, W. Zheng, and Z. Huang, "Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines," *Int. J. Oncol.* **32**(3), 653-662 (2008).

14.N. Stone, C. A. Kendall, N. Shepherd, P. Crow, and H. Barr, "Near-infrared Raman spectroscopy for the classification of epithelial pre-cancers and cancers," *J. Raman Spectrosc.* **33**(7), 564-573 (2002).

15.D. P. Lau, Z. Huang, H. Lui, C. S. Man, K. Berean, M. D. Morrison, and H. Zeng, "Raman spectroscopy for optical diagnosis in normal and cancerous tissue of the nasopharynx—preliminary findings," *Lasers Surg. Med.* **32**(3), 210-214 (2003).

16.S. Duraipandian, W. Zheng, J. Ng, J. J. H. Low, A. Ilancheran, and Z. Huang, "In vivo diagnosis of cervical precancer using Raman spectroscopy and genetic algorithm techniques," *Analyst* **136**(20), 4328-4336 (2011).

17.S. Duraipandian, W. Zheng, J. Ng, J. J. H. Low, A. Ilancheran, and Z. Huang, "Near-infrared-excited confocal Raman spectroscopy advances in vivo detection of cervical precancer," *J. Biomed. Opt.* **18**(6), 067007 (2013).

18.G. Shetty, C. A. Kendall, N. Shepherd, N. Stone, and H. Barr, "Raman spectroscopy: elucidation of biochemical changes in carcinogenesis of oesophagus," *Br. J. Cancer* **94**(10), 1460-1464 (2006).

19.K. W. Short, S. Carpenter, J. P. Freyer, and J. R. Mourant, "Raman spectroscopy detects biochemical changes due to proliferation in mammalian cell cultures," *Biophys. J.* **88**(6), 4274-4288 (2005).

20.K. E. Shafer-Peltier, A. S. Haka, M. Fitzmaurice, J. P. Crowe, J. Myles, R. R. Dasari, and M. S. Feld, "Raman microspectroscopic model of human breast tissue: implications for breast cancer diagnosis in vivo," *J. Raman Spectrosc.* **33**(7), 552-563 (2002).

21.M. S. Bergholt, W. Zheng, K. Lin, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. Yan So, and Z. Huang, "In vivo diagnosis of esophageal cancer using image-guided Raman endoscopy and biomolecular modeling," *Technol. Cancer Res. Treat.* **10**(2), 103-112 (2011).

22.M. Hackemann, C. Grubb, and K. R. Hill, "The ultrastructure of normal squamous epithelium of the human cervix uteri," *J. Ultrastruct. Res.* **22**(5), 443-457 (1968).

23.K. Saavedra, P. Brebi, and J. Roa, "Epigenetic alterations in preneoplastic and neoplastic lesions of the cervix," *Clin. Epigenet.* **4**(1), 1-7 (2012).

24.D. Zietkowski, N. M. deSouza, R. L. Davidson, and G. S. Payne, "Characterisation of mobile lipid resonances in tissue biopsies from patients with cervical cancer and correlation with cytoplasmic lipid droplets," *NMR in Biomedicine* **26**(9), 1096-1102 (2013).

25.L. E. Kamemoto, A. K. Misra, S. K. Sharma, M. T. Goodman, H. Luk, A. C. Dykes, and T. Acosta, "Near-infrared micro-Raman spectroscopy for in vitro detection of cervical cancer," *Appl. Spectrosc.* **64**(3), 255-261 (2010).

26.M. L. Akins, K. Luby-Phelps, R. A. Bank, and M. Mahendroo, "Cervical softening during pregnancy: regulated changes in collagen cross-linking and composition of matricellular proteins in the mouse," *Biol. Reprod.* **84**(5), 1053-1062 (2011).

27.Z. Movasaghi, S. Rehman, and I. U. Rehman, "Raman spectroscopy of biological tissues," *Appl. Spectrosc. Rev.* **42**(5), 493 - 541 (2007).

28.F. M. Lyng, E. O. Faolain, J. Conroy, A. D. Meade, P. Knief, B. Duffy, M. B. Hunter, J. M. Byrne, P. Kelehan, and H. J. Byrne, "Vibrational spectroscopy for cervical cancer pathology, from biochemical analysis to diagnostic tool," *Exp. Mol. Pathol.* **82**(2), 121-129 (2007).

29.S. Duraipandian, M. S. Bergholt, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. Yan So, and Z. Huang, "Real-time Raman spectroscopy for in vivo, online gastric cancer diagnosis during clinical endoscopic examination," *J. Biomed. Opt.* **17**(8), 081418 (2012).

30.B. Albarran-Somoza, R. Franco-Topete, V. Delgado-Rizo, F. Cerda-Camacho, L. Acosta-Jimenez, M. Lopez-Botet, and A. Daneri-Navarro, "CEACAM1 in cervical cancer and precursor lesions: association with human papillomavirus infection," *J. Histochem. Cytochem.* **54**(12), 1393-1399 (2006).

31.M. L. Serrano, A. Umaña-Pérez, D. J. Garay-Baquero, and M. Sánchez-Gómez, "New Biomarkers for Cervical Cancer–Perspectives from the IGF System," (2012).

32.L. D. Kellenberger, J. E. Bruin, J. Greenaway, N. E. Campbell, R. A. Moorehead, A. C. Holloway, and J. Petrik, "The role of dysregulated glucose metabolism in epithelial ovarian cancer," *J. Oncol.* **2010**, 1-14 (2010).

33.S. Duraipandian, W. Zheng, J. Ng, J. J. H. Low, A. Ilancheran, and Z. Huang, "Simultaneous fingerprint and high-wavenumber confocal Raman spectroscopy enhances early detection of cervical precancer in vivo," *Anal. Chem.* **84**(14), 5913-5919 (2012).

34.M. S. Bergholt, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. Y. So, A. Shabbir, and Z. Huang, "Fiber-optic Raman spectroscopy probes gastric carcinogenesis in vivo at endoscopy," *J. Biophotonics* **6**(1), 49-59 (2013).

35.K. J. Park, and R. A. Soslow, "Current concepts in cervical pathology," *Arch. Pathol. Lab. Med.* **133**(5), 729-738 (2009).

36.E. S. Cibas, and B. S. Ducatman, *Cytology: diagnostic principles and clinical correlates*, Elsevier Health Sciences, Philadelphia (2009).

37.T. C. Wright, B. M. Ronnett, R. J. Kurman, and A. Ferenczy, "Precancerous lesions of the cervix," in *Blaustein's Pathology of the Female Genital Tract* R. J. Kurman, L. H. Ellenson and B. M. Ronnett, Eds., pp. 193-252, Springer US (2011).

Table 1. Classification results obtained from NIR Raman spectral prediction of benign, LSIL and HSIL cervical tissue using PLS-DA algorithms, together with leave-one tissue site-out, cross-validation method.

Raman prediction			
Tissue type	**Benign**	**LSIL**	**HSIL**
Benign	**22**	**0**	**1**
LSIL	**0**	**24**	**5**
HSIL	**0**	**3**	**13**
Sensitivity (%)	**95.7**	**82.8**	**81.3**
Specificity (%)	**100.0**	**92.3**	**88.5**

Figure captions

Fig. 1 (a) Comparison of mean tissue Raman spectra ±1 standard deviations (SD) acquired from benign (n=23), LSIL (n=29) and HSIL (n=16) cervical tissue. Note that the mean tissue Raman spectra are shifted vertically for better visualization. The shaded area represents the respective SD. (b) The corresponding difference spectra ± 1SD calculated from the mean Raman spectra between different tissue pathologies (benign, LSIL and HSIL).

Fig. 2 The basis Raman spectra (i.e., DNA, proteins (histones, collagen), lipids (triolein) and carbohydrates (glycogen)) used for biochemical modeling of precarcinogenesis progression in cervical tissue.

Fig. 3 Comparison of Raman spectra measured from benign, LSIL and HSIL cervix with the corresponding reconstructed Raman spectra using a linear combination of basis spectral set: (a) benign, (b) LSIL, and (c) HSIL. Residuals (measured spectrum minus fit spectrum) are also shown in each plot.

Fig. 4 Histograms displaying the relative biochemical concentration profiles of benign, LSIL, and HSIL cervical tissues. The one standard deviation (SD) confidence intervals are shown for each model component. Note: (*) indicates a significant differences ($p < 0.05$, one-way ANOVA) for discriminating benign cervical tissue from HSIL; and (x) indicates a significant differences ($p < 0.05$, one-way ANOVA) for discriminating benign cervix from LSIL.

Fig. 5 The diagnostically significant latent variable (LV) accounting for 19.0% and 22.1% of the total variations in the Raman spectral dataset in X direction and Y direction, respectively, revealing the diagnostically significant Raman spectral features for tissue classification.

Fig. 6 Two-dimensional ternary plot of the posterior probabilities belonging to benign, LSIL and HSIL achieved by PLS-DA multi-class modelling together with leave-one tissue site-out, cross-validation method.

Fig. 7 Receiver operating characteristic (ROC) curves of discrimination results for classification of benign, low-grade (LSIL) and high-grade (HSIL) cervical tissue using Raman spectroscopy and PLS-DA algorithms together with leave-one tissue site-out, cross-validation method.

Figure 1

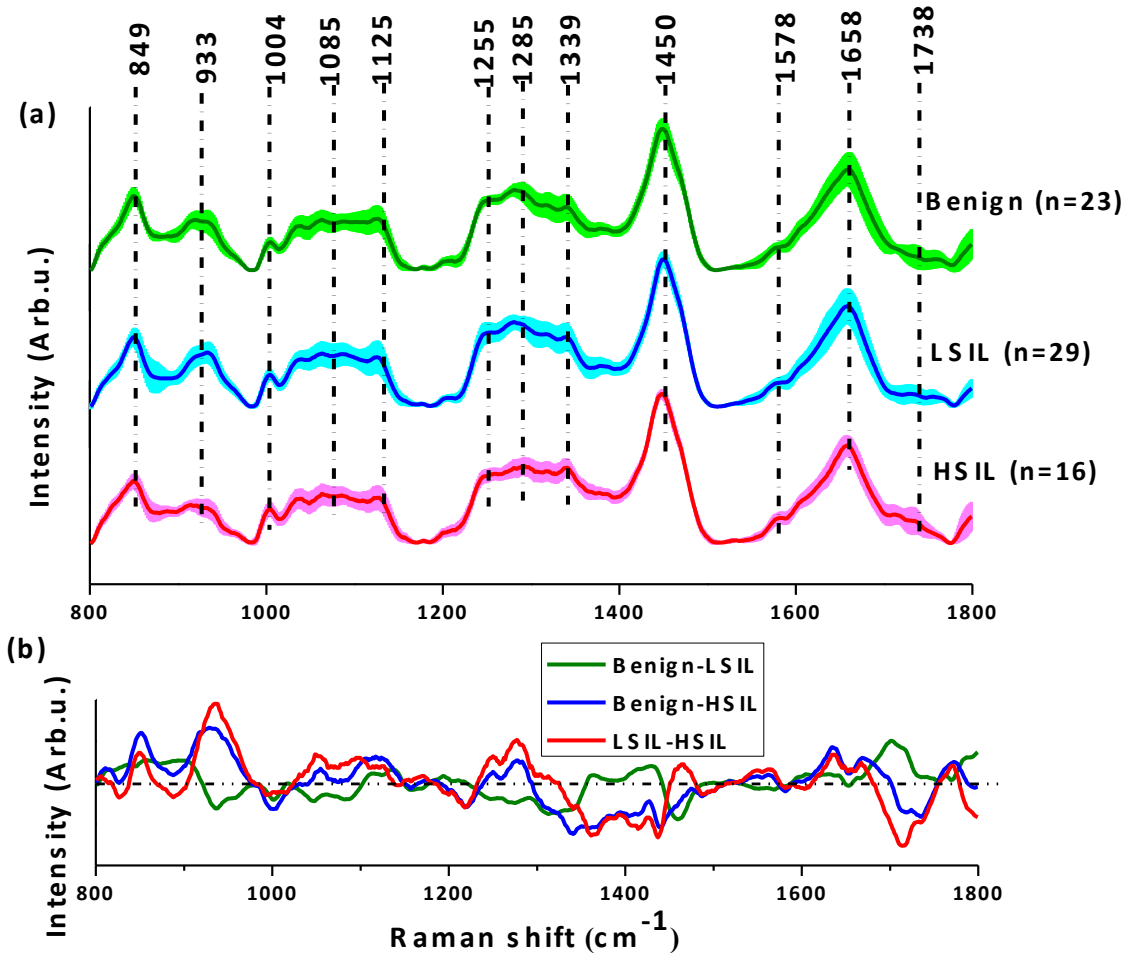


Figure 2

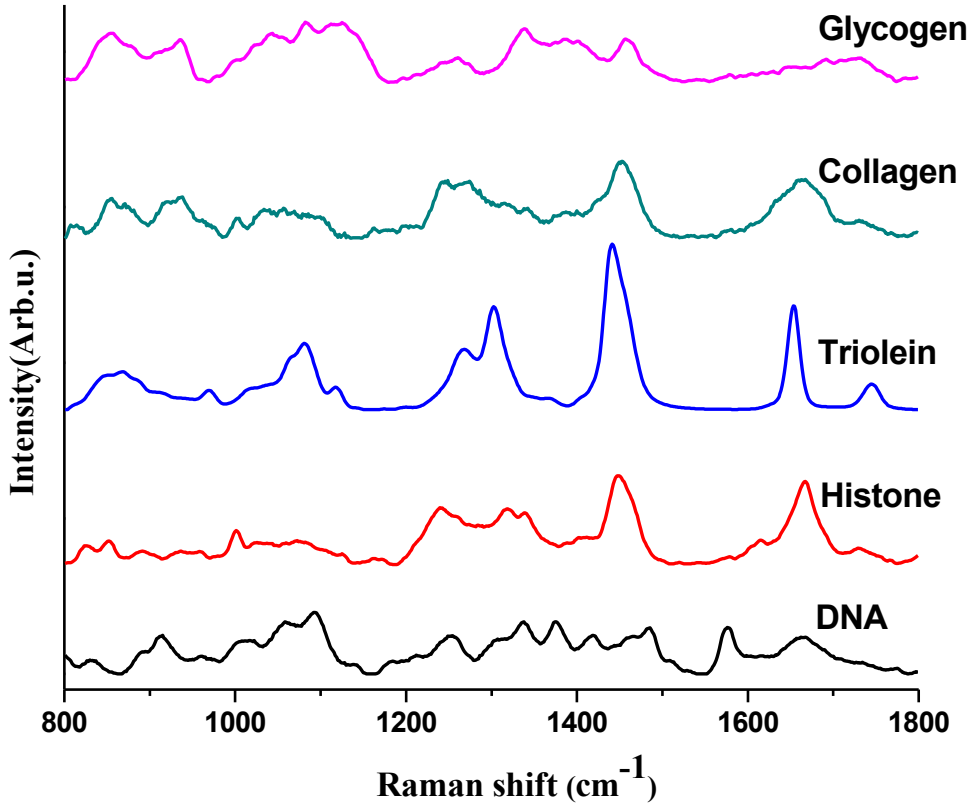


Figure 3

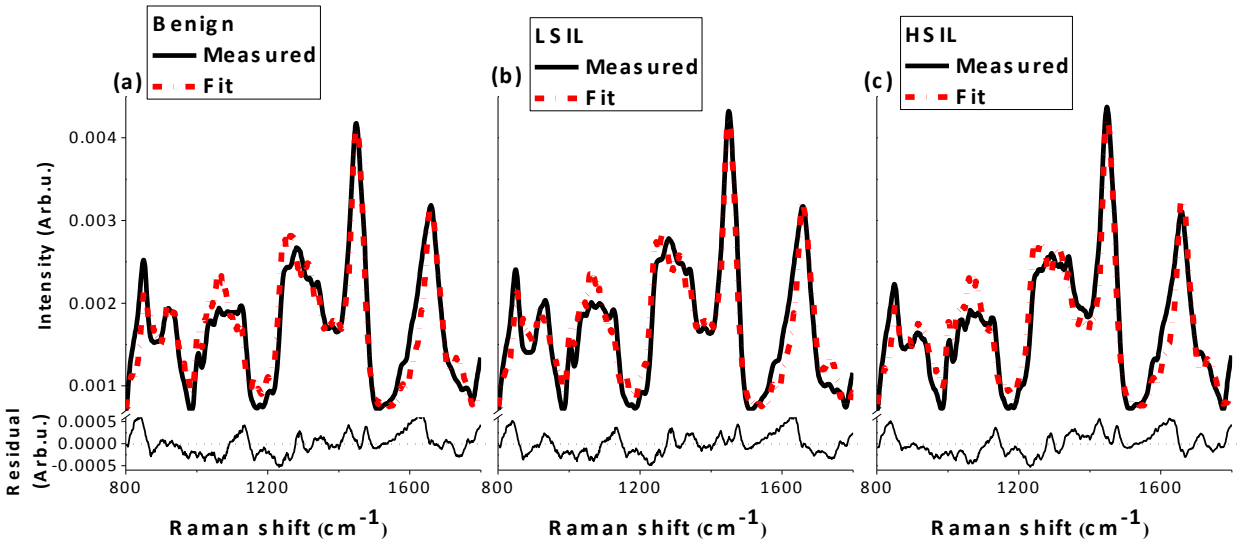


Figure 4

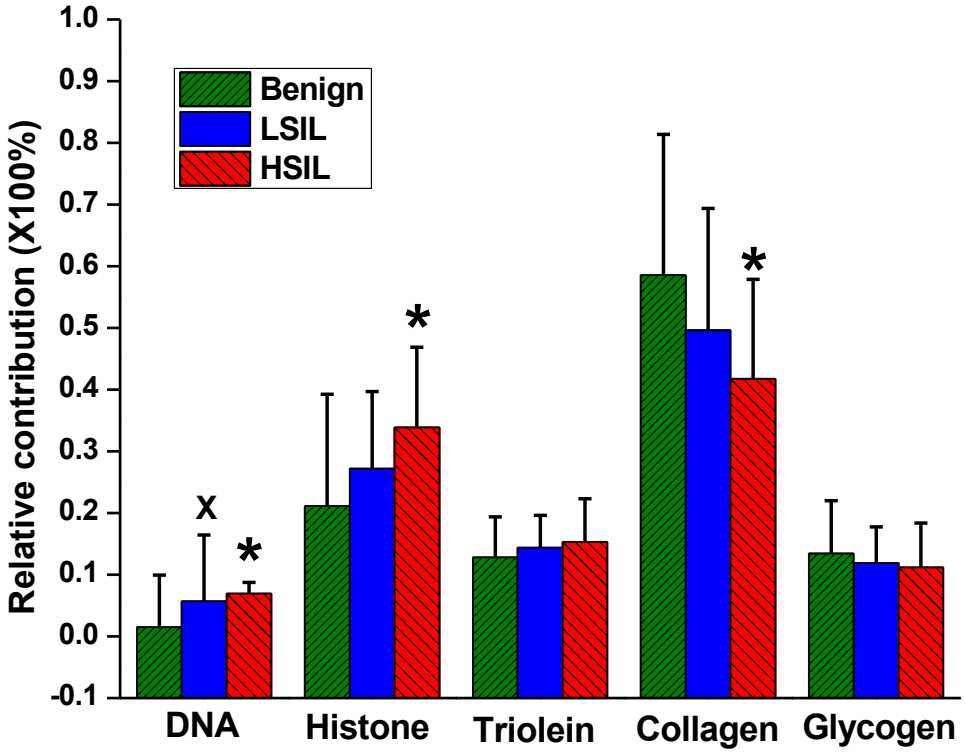


Figure 5

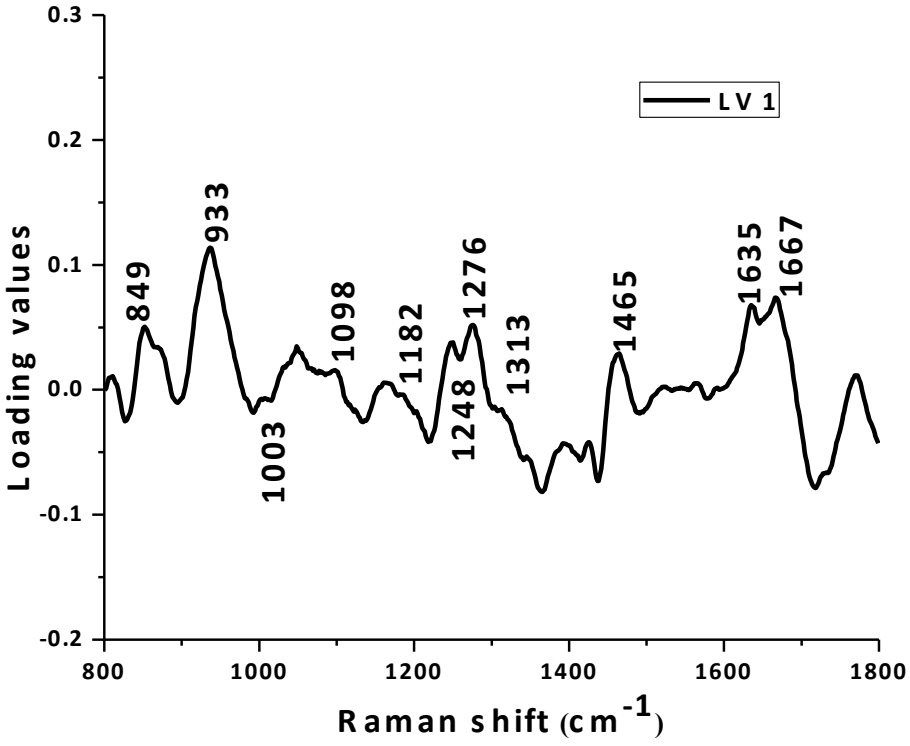


Figure 6

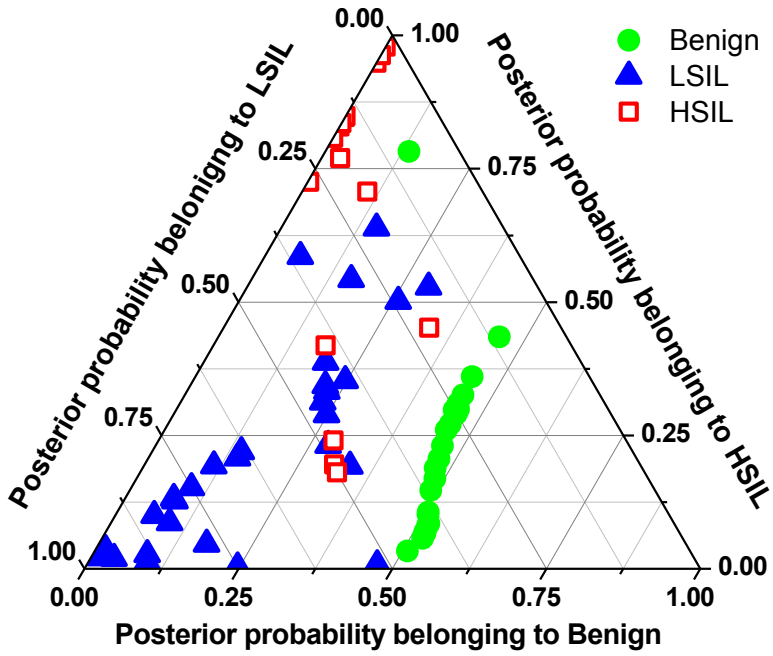
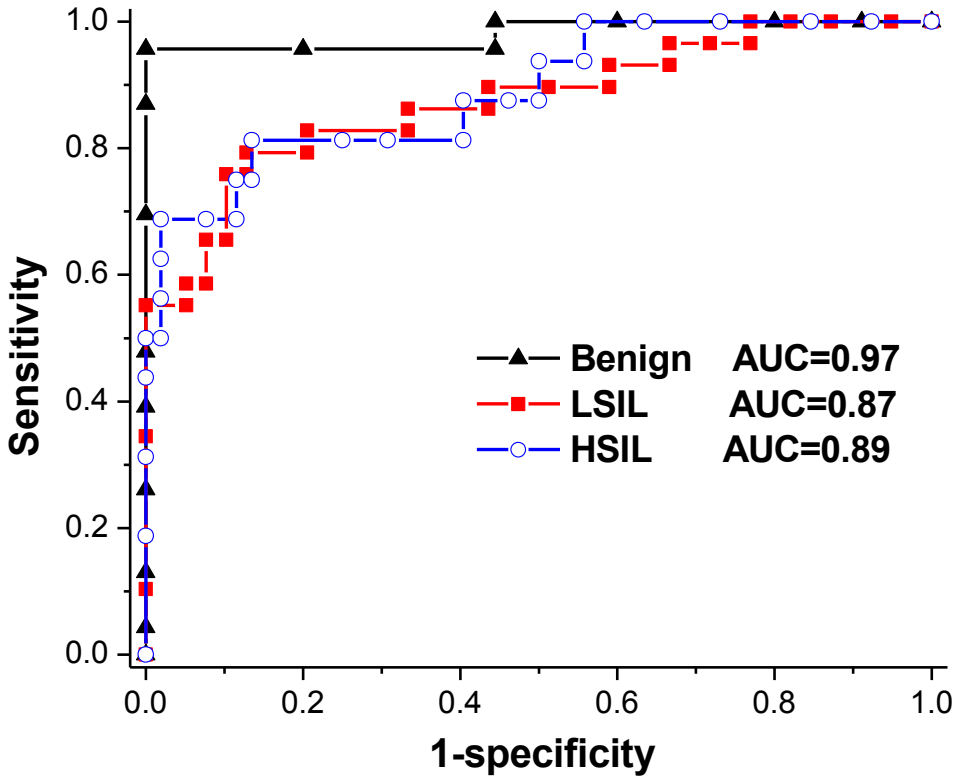
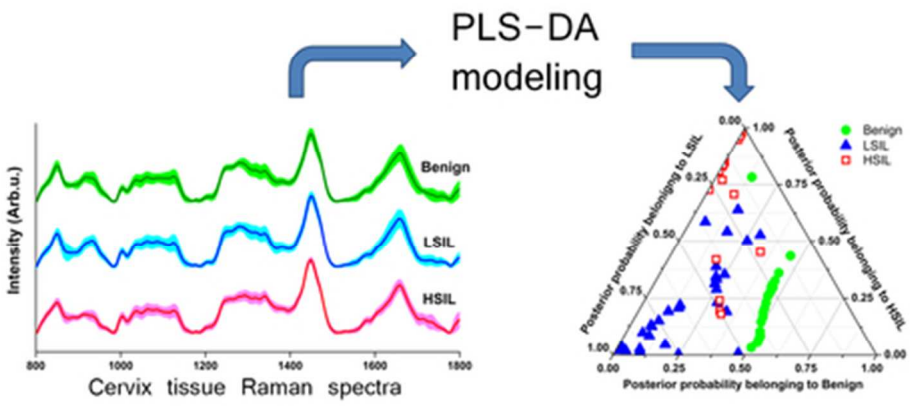


Figure 7



NIR Raman spectroscopic characterization of cervical precarcinogenic transformation



A table of contents entry
39x19mm (300 x 300 DPI)