Articles   NanoLab

Raman Microspectroscopy for the Early Detection of Pre-Malignant Changes in Cervical Tissue

Nosheen Rashid
Technological University Dublin

Haq Nawaz
Technological University Dublin

Kelvin W. Poon
Technological University Dublin

See next page for additional authors

Recommended Citation

Authors

Nosheen Rashid, Haq Nawaz, Kelvin W. Poon, Franck Bonnier, Salih Bakhiet, Cara Martin, John O'Leary, Hugh Byrne, and Fiona Lyng

1 **Raman microspectroscopy for the early detection of pre-malignant changes in**

2 **cervical tissue**

3 *Nosheen Rashid[a,1], Haq Nawaz[a,1], Kelvin W.C. Poon[a,2], Franck Bonnier[a], Salih Bakhiet[b],*

4 *Cara Martin[b], John J. O'Leary[b], Hugh J. Byrne[c], Fiona M. Lyng[a,d]*

5 [a] DIT Centre for Radiation and Environmental Science (RESC), Focas Research Institute,

6 Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland

7 [b] Department of Pathology, Coombe Women and Infants University Hospital, Dublin 8,

8 Ireland

9 [c] Focas Research Institute, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland.

10 [d] School of Physics, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland

11 *Corresponding Author :

12 Fiona M. Lyng

13 DIT Centre for Radiation and Environmental Science (RESC)

14 Focas Research Institute

15 Dublin Institute of Technology

16 Kevin Street

17 Dublin 8

18 Ireland

19 Tel. 00 353 1 4027972

20 Email:fiona.lyng@dit.ie

[1] Present address: National Institute for Biotechnology and Genetic Engineering (NIBGE), P.O.Box 577, Jhang Road Faisalabad, Pakistan.
[2] Present address: Hotchkiss Brain Institute, The University of Calgary, Canada

Abstract

Cervical cancer is the third most common cancer affecting women worldwide. The mortality associated with cervical cancer can, however, be significantly reduced if the disease is detected at the pre-malignant stage. The aim of this study was to evaluate the potential of Raman microspectroscopy for elucidation of the biochemical changes associated with the pre-malignant stages of cervical cancer. Formalin fixed paraffin preserved tissue sections from cervical biopsies classified as negative for intraepithelial lesion and malignancy (NILM), low grade squamous intraepithelial lesion (LSIL) or high grade squamous intraepithelial lesion (HSIL) were analysed by Raman spectral mapping. Raman mapping, with K-Means Cluster Analysis (KMCA), was able to differentiate the NILM cervical tissue into three layers including stroma, basal/para-basal and superficial layers, characterised by spectral features of collagen, DNA bases and glycogen respectively. In the LSIL and HSIL samples, KMCA clustered regions of the superficial layer with the basal layer. Using Principal Component Analysis (PCA), biochemical changes associated with disease were also observed in normal areas of the abnormal samples, where morphological changes were not apparent. This study has shown that Raman microspectroscopy could be useful for the early detection of pre-malignant changes in cervical tissue.

44   **Introduction**

45   Cervical cancer is the third most common cancer in women worldwide, 85% of cases

46   arising in the developing world (Jemal *et al*, 2011). Mortality associated with cervical

47   cancer can, however, be significantly reduced if the disease is detected at the early

48   stages of its development or at the pre-malignant stage (cervical intraepithelial

49   neoplasia, CIN) (Parkin *et al*, 2005).

50   Current methods for identifying neoplastic cells and differentiating them from their

51   normal counterparts are often nonspecific, slow, invasive, or a combination thereof

52   (Chan *et al*, 2006). The primary screening tool for cervical neoplasia is the

53   Papanicolaou (Pap) smear test (Papanicolaou & Traut, 1941) and colposcopy

54   normally follows an abnormal Pap smear. The incidences of the disease and

55   associated mortality have been falling in developed countries, where these screening

56   methods are widely practiced. However, these methods have several shortcomings,

57   including high false negative/positive results that could be due to the subjective

58   interpretation of the cytologist/pathologist diagnosing the disease based on

59   morphological abnormalities. There are difficulties in separation of normal, basal cell

60   hyperplasia, immature squamous metaplasia and inflammation associated changes

61   from true koilocytes, which are indicative of low grade squamous metaplasia

62   (McCluggage *et al*, 1998). There are also difficulties in differentiating low grade

63   squamous intraepithelial lesions (LSIL) from immature squamous metaplastic cells

64   with atypia, a mild abnormality, resulting in inter- and intra- observer difference of

65   opinion (McCluggage *et al*, 1998).

66   It is estimated that 90% of deaths can be prevented by early diagnosis, (Bazant-

67   Hegemark *et al*, 2008), which emphasizes the need for improved effective screening

68 methods. The molecular and biochemical changes that ultimately lead to the

69 occurrence of cancer in tissue include the disruption of the normal organisation of

70 nucleic acids, proteins, lipids, and carbohydrates (Ooi *et al*, 2008) and these changes

71 can be used as diagnostic markers of cancer. Historically, many techniques have

72 targeted the above mentioned biomolecules and sub-cellular structures, nucleus to

73 cytoplasm ratio or cellular morphology, for diagnosis as well as for the prognosis of

74 cancer. However, changes at this level are detectable only after they have already

75 caused considerable gross morphological changes in the tissue. Due to technological

76 advances, more opportunities have arisen to investigate the bio-molecular components

77 in tissue samples and to examine malignant abnormalities on this basis. As a result,

78 the mutations/ biochemical changes in bio-molecules appearing before manifestation

79 of disease on a cellular level can be detected earlier, facilitating timely and precise

80 diagnosis, leading to increased patient survival and quality of life (Ooi *et al*, 2008).

81 Potentially offering higher sensitivity and specificity based on biochemical analyses,

82 spectroscopic methods such as IR absorption and Raman scattering are attracting

83 increased attention. Investigation of 150 formalin fixed and paraffin preserved (FFPP)

84 cervical tissue sections by Raman spectroscopy was carried out and normal cervical

85 tissue was separated from malignant tissues with a sensitivity and specificity of 99.5%

86 (Krishna *et al*, 2006). An *in vivo* study on 66 patients showed that Raman

87 spectroscopy has 89% sensitivity and 81% specificity to distinguish high grade

88 dysplasia (CIN II/CIN III) from benign tissue (Robichaux-Viehoever *et al*, 2007). The

89 reported results are higher than colposcopy, which has a sensitivity of 87% and

90 specificity of 72%. In another study, the incorporation of the woman's hormonal

91 status, mainly the stage in menstrual cycle and menopausal state, into the

92 classification algorithm was shown to improve the sensitivity of Raman spectroscopy

93 to detect cervical pre-cancers from 88% to 94% (Kanter *et al*, 2009b). The same

94 group characterised the Raman spectral changes associated with the menopause as

95 well as within the menstrual cycle which led to more accurate diagnosis of LSIL *in*

96 *vivo* (Kanter *et al*, 2009a). These results render Raman spectroscopy a valid

97 alternative for clinical screening. However, few studies have delved further into the

98 biochemical origin of the diagnostic differentiation between the different tissue

99 pathologies. Our group has reported differentiation of 40 FFPP histological samples

100 into normal cervical epithelium, invasive carcinoma and CIN with nearly 100%

101 sensitivity and specificity with the differentiation of normal epithelium and carcinoma

102 *in situ* largely due to the absence of strong glycogen bands in the abnormal tissue

103 containing rapidly proliferating cells (Lyng *et al*, 2007). A recent study of de-

104 paraffinised cervical tissue with Raman mapping and hierarchical cluster analysis

105 (HCA) was able to differentiate between normal squamous epithelium and cervical

106 intraepithelial neoplasia-grade II (CIN II) and it was shown that the Raman spectra

107 associated with the CIN II lesion clustered predominantly with those of the basal

108 epithelial cells of normal squamous epithelium which suggested that the cells of these

109 regions share common biochemical profiles and spectral features responsible for their

110 differentiation are associated with the amide-I and amide-III bands (Tan *et al*, 2011).

111 In an ex vivo study of human cervix, Raman spectral profiles from the stroma of

112 tissue with HPV associated histological changes showed differences for DNA (1316

113 and 1334cm^{-1}) and glycogen (1048, 1083, 1256, 1333cm^{-1}) assignments. Further

114 differences at 1260 and 1304cm^{-1} were also observed. Kamemoto et al. (2010)

115 showed that Raman spectroscopy could distinguish normal cervical tissue from

116 cervical cancer tissue based on collagen bands in the 775-975cm^{-1} region.

117   The aim of the present study is to evaluate the potential of Raman micro spectroscopy

118   for the elucidation of the biochemical origin of the spectral markers associated with

119   the onset and progression of the disease. The initial stages will establish the

120   biochemical signatures of different layers of the normal cervical epithelium. It is

121   expected that these signatures will be distinct, as the biochemistry of each layer is

122   different. These differences in the biochemistry of each layer in the form of Raman

123   spectral markers may be helpful for the better classification of cervical intraepithelial

124   neoplasia. This would lead to the classification of cervical cancer on the basis of the

125   biochemical changes occurring during the progression of the disease instead of on

126   morphological changes. Raman micro spectroscopy will then be used to monitor the

127   complete progression of cervical cancer from normal through to low grade squamous

128   intraepithelial lesion (LSIL) and high grade squamous intraepithelial lesion (HSIL) on

129   the basis of spectral differences obtained from the different layers of the respective

130   tissue samples.

131

132 **Materials and Methods**

133 **Sample collection**

134 The samples used in this study were obtained from the Histopathology Department,

135 Coombe Women and Infants University Hospital, Dublin, Ireland. This study was

136 approved by Research Ethics Committee at the Coombe Women and Infants

137 University Hospital. Normal cervical tissue samples were taken from hysterectomy

138 specimens for a benign diagnosis such as uterine prolapse or fibroids. Further cervical

139 H&E sections from these hysterectomy specimens were microscopically examined to

140 reveal no Cervical Intraepithelial Neoplasia or significant pathological abnormality.

141 The clinical history of these cases confirmed no significant previous cervical

142 pathology. Anonymised tissue sections from cervical biopsies clinically characterised

143 as NILM, LSIL (or CIN 1) or HSIL (or CIN 2, CIN 3) were analysed. Raman spectral

144 maps from cervical tissue samples of twenty different patients, including, five NILM,

145 two LSIL (CIN 1), ten HSIL (five CIN 2, five CIN 3) and three carcinoma in situ

146 samples, were acquired. Raman spectral mapping of regions identified by the

147 pathologist as normal and diseased was carried out. For each sample, two parallel

148 tissue sections of 10µm thickness were cut from the Formalin Fixed Paraffin

149 Processed cervical tissue block using a microtome. After dewaxing with xylene

150 according to standard clinical procedures (Ó Faoláin *et al*, 2005), one section was

151 mounted on a calcium fluoride slide (Crystran Ltd., UK) and used for the Raman

152 spectroscopic analysis and the other section was mounted on a glass slide for

153 hematoxylin and eosin (H&E) staining.

154 **Raman spectral mapping**

155 Tissue sections from cervical tissue biopsies were mapped with Raman spectroscopy

156 from the regions that were identified by the pathologist as normal and as low/high

157 grade squamous lesions to elucidate the layer differentiation within the normal tissue

158 and how this differentiation is affected by the progression of CIN, based on the

159 spectroscopic signatures of the biochemical composition. The analysis of the Raman

160 maps from the cervical epithelium of different tissue sections revealed a serial change

161 associated with the progression of disease.

162 **Raman Data acquisition**

163 Raman maps were recorded using a HORIBA Jobin Yvon HR 800 Raman microscope

164 (LabSpec V5.58) with a 785 nm laser as source. An Olympus (MPLN100X) dry

165 objective, having a numerical aperture of 0.9, providing a spot size of ~1μm on the

166 sample, was used to focus on the sample and collect the Raman scattered light in a

167 backscattering geometry. Raman scattering was collected through a 100 μm confocal

168 hole onto a Synapse air-cooled CCD detector for the range of 400-1800 cm^{-1} using a

169 300 lines/mm diffraction grating, yielding a dispersion of ~1.5cm^{-1} per CCD pixel.

170 The instrument was calibrated using the 520.7 cm^{-1} peak of silicon. Raman spectral

171 mapping was performed using 2 x 15 seconds acquisitions with a step size of 18μm.

172 **Data pre-processing of Raman maps**

173 All data processing was performed directly on the spectral image using MatLab 7.2

174 and protocols which have been established in the Dublin Institute of Technology

175 (Knief, 2010). Data pre-processing included smoothing, baseline correction and

176 normalization. All spectra, including calibration and substrate backgrounds, were

177 vector normalized and smoothed using a Savitzky Golay smoothing (order 5, 13 point

178 window) method. A rubber band correction for baseline removal for all the spectra

179 was carried out and the substrate spectra were subtracted from each spectrum.

180 **Data analysis**

181 K-means cluster analysis (KMCA) and Principal Components Analysis (PCA) were

182 used to analyse the spectral data sets obtained from the Raman maps. KMCA is an

183 unsupervised non hierarchical method of clustering the cases randomly into user

184 predefined clusters (5 cluster with 10 iterations for the current study) and cycles until

185 a local minimum is found by using the Euclidean sum of squares as a descriptor

186 (Wang & Mizaikoff, 2008). The false colour map generated then shows the clusters

187 with similar spectral and hence biochemical properties. The pseudo colour map

188 generated by the KMCA leads to the identification of the distinct layers of the cervical

189 epithelium, separated on the basis of the similarities of the Raman spectral profiles

190 and the clustered data is represented by an average of all the spectra of one layer.

191 In order to further elucidate the biochemical basis of the results of the KMCA,

192 Principal component analysis (PCA) was employed. PCA is a mathematical procedure

193 involving the transformation of possibly correlated variables into a smaller number of

194 uncorrelated variables known as principal components (PC). The first principal

195 component accounts for as much variability in the data as possible, and each

196 succeeding principal component accounts for as much of the remaining variability as

197 possible. The PC loadings can describe the biochemical differences which are used by

198 the PCA to determine the variability in the Raman spectral data and hence separation

199 of different groups of spectra.

200

Results and Discussion

KMCA of NILM cervical epithelium

Figure 1a, shows the H&E stained tissue section and **Figure 1b** shows the extended video image of an unstained NILM tissue section with a 100x objective acquired using the Raman micro spectrometer. The sample was spectroscopically mapped and analysed in the area indicated by the red box. KMCA of the Raman map is shown in **Figure 1c**. The analysis employs 5 clusters which divide the cervical epithelium into three distinct layers. A comparison of the unstained tissue section with the results of the KMCA for this map suggests that the blue cluster represents the stroma, the red cluster the basal and parabasal layers, the green cluster the superficial layer and the black cluster the substrate, while the yellow cluster represents the presence of residual wax in the tissue section.

Mean Raman spectra representing the respective clusters are shown in **Figure 1d** (the spectra of clusters describing the substrate and residual wax are not shown). The Raman spectra of pure collagen and glycogen used for reference are described by Lyng *et al*, 2007, Tan *et al*, 2011.

The most distinctive Raman bands in the blue spectrum of **Figure 1d**, indicative of the blue cluster of Figure 1c, can be assigned to collagen ($853cm^{-1}$, $921cm^{-1}$, $938 cm^{-1}$ and $1245 cm^{-1}$) and phenylalanine ($1002cm^{-1}$). Collagen is a structural protein and is a major component of the connective tissue layer, termed the stroma in the case of cervical tissue. The distinctive bands in the red spectrum of **Figure 1d**, corresponding to the red cluster of **Figure 1c**, are of DNA bases, thymine ($755 cm^{-1}$), adenine ($722 cm^{-1}$) and cytosine ($782 cm^{-1}$). On the basis of this information, it can be concluded that the red cluster represents the basal and parabasal layers. The bands present at

225   480cm^{-1}, 849cm^{-1} and 938 cm^{-1} in the Raman spectrum suggest that accumulation of

226   glycogen starts in the basal and parabasal layers, which are also rich in DNA. The

227   major differentiating Raman features observed in the mean representative spectrum of

228   the green cluster can be assigned to glycogen (480cm^{-1}, 849cm^{-1} and 938 cm^{-1}), and

229   are Raman signatures of the superficial squamous epithelium of the cervix, as cervical

230   epithelial cells accumulate glycogen with maturity (Sellors & Sankaranarayanan,

231   2003).

232   It should be noted that two important Raman markers of collagen and glycogen can be

233   used to differentiate the Raman spectra of the layers of squamous epithelium and

234   stroma. Based on the analyses of (Jess *et al*, 2007; Lyng *et al*, 2007; Notingher, 2007;

235   Notingher *et al*, 2003), the Raman bands at 853cm^{-1}, 921cm^{-1}, 938cm^{-1} and 1245 cm^{-1}

236   are indicative of collagen and those at 480cm^{-1}, 849cm^{-1} and 938 cm^{-1} can be assigned

237   to glycogen. Notably, the two bands at 853 cm^{-1} (collagen) and 849 cm^{-1} (glycogen)

238   overlap and are not well differentiated by Raman spectroscopy. By considering the

239   pattern of the collagen and glycogen bands in the current study, it is observed that 921

240   cm^{-1} and 938 cm^{-1} are normally co-incident in spectra of collagen and the 480 cm^{-1}

241   and 853 cm^{-1} bands are normally present in the spectra of glycogen. These Raman

242   bands are therefore good differentiating bands between collagen and glycogen.

243   Another differentiating marker can be that at 1245 cm^{-1}, which is observed as a very

244   sharp band if associated with the presence of collagen, while otherwise it is very

245   much diminished and broadened.

246   It is also notable that the basal layer shares some features with the stromal layer, such

247   as that at 1318 cm^{-1}, assigned to guanine, and the superficial layer, such as that at

248   1084 cm^{-1}, assigned to the O-P-O stretching of DNA, and these Raman bands are

249 therefore not differentiating bands between these layers. The Raman band at 1339 cm$^-$

250 ^{1} is assigned to tryptophan and is also present in the Raman spectra of all the layers.

251 Comparison of the H&E stained tissue section, the unstained tissue section and the

252 KMCA of the Raman map demonstrates that Raman spectroscopy is able to

253 differentiate normal cervical epithelium into three layers; stroma, basal, and squamous

254 epithelium, on the basis of the prominence of Raman signatures of collagen in the

255 stroma, DNA bases in the basal and parabasal layer and glycogen in the superficial

256 layer. The presence of glycogen bands in the mean spectrum of the basal layer is an

257 indication that cells are dividing normally and gradually maturing to accumulate

258 glycogen.

259 **PCA of normal cervical epithelium [NILM]**

260 **PCA of normal basal layer vs normal stromal layer**

261 **Figure 2 A & B** represents the scatter plot and loadings of the PCA for the basal

262 versus stromal layer. These two layers are well differentiated by PC1, which accounts

263 for 67.4 % of the variance. PC2 does not differentiate these groups of Raman spectra

264 as indicated by the PCA scatter plot.

265 Notably, in the PCA scatter plot, the cluster of the Raman spectra of the basal layer is

266 distributed negatively with respect to PC1 and the Raman spectra of the stroma are

267 clustered on the positive side. This is translated to the loadings of PC1, as all the

268 negative loadings are associated with the characteristic Raman features of DNA bases

269 (Bonnier & Byrne, 2012) and are clearly indicative of the characteristics of the basal

270 layer, as it is richer in DNA as compared to the stromal layer. The features associated

271 with DNA, contributed by the basal layer, include those at 728 cm^{-1} (A), 827 cm^{-1} (O-

272 P-O of DNA), 898 cm^{-1} (deoxyribose ring breathing),1276 cm^{-1} and 1340 cm^{-1} (G).

273 Only a few features related to proteins are observed, including those at 509 cm^{-1} (S-S

274 stretching), 642 cm^{-1} (C-C stretching), 1129 cm^{-1} (C-N stretching) and 1449 cm^{-1} (CH

275 deformation). On the other hand, the positive loadings can be assigned to the

276 characteristic Raman features of the stromal layer. These loadings are assigned to

277 collagen and contribute strongly to differentiating this layer from the basal layer. The

278 features contributed by the stroma include those at 853 cm^{-1}, 921 cm^{-1}, 972 cm^{-1} and

279 1242 cm^{-1}, assigned to the collagen of the stromal layer and 1165 cm^{-1}, 1195 cm^{-1} (C-

280 N stretching) and 1655 cm^{-1} (amide-I).

281 **PCA of normal basal layer vs normal superficial layer**

282 **Figure 2 C & D** presents the PCA scatter plot and loadings of the PCA respectively

283 for the normal basal versus normal superficial layers. A good separation of the two

284 layers based on PC1, which accounts for 74.2% of the variance, is observed as two

285 well distinguished clusters of the spectra in the PCA scatter plot. In the PCA scatter

286 plot, the cluster of the Raman spectra of the normal basal layer is observed in the

287 negative PC1 axis and Raman spectra of the superficial layer are clustered in the

288 positive axis which is consistent with the loadings of the PCA as the negative

289 loadings are related to the characteristic Raman features of DNA bases which are

290 characteristic of the basal layer. The clustering of the Raman spectra of the superficial

291 layer is observed in the positive axis, which is consistent with the loadings of the

292 PCA, because the positive loadings are assigned to the characteristic Raman features

293 of glycogen, a specific marker of this layer.

294 The loadings of PC1 elucidate the discriminating Raman features which can be

295 attributed to the basal layer, including predominantly DNA related Raman bands in

296 the negative sense, and glycogen related Raman bands in the positive sense, which

297 can be attributed to the superficial layer. Specifically, the loadings of PC1 show

298 negative features at 755 cm^{-1}(T), 728 cm^{-1}, 1338 cm^{-1} (A), 775 cm^{-1} (C), 1182 cm^{-1},

299 1313 cm^{-1} (G) and 1080 cm^{-1} (O-P-O of DNA). Positive features of PC1 at 480 cm^{-1},

300 853 cm^{-1} and 938 cm^{-1} are assigned to the glycogen of the superficial layer.

301 The PCA of the basal layer of different NILM samples are presented in the

302 Supplementary Information (**figure S1**), as scatter plot and loadings respectively. The

303 scatter plot shows some differentiation of the two clusters of the Raman spectra by

304 PC1 (56.7%) and very little differentiation by PC2 (15.80%) which is attributed to

305 some degree of inter-patient variability, largely based on intensity variations of peaks

306 associated with proteins and lipids (**figure S1**).

307 Thus, while KMCA can be employed to visualise the differences of the tissue

308 architecture, the results of the PCA of the normal basal layer versus the normal

309 stromal and superficial layers demonstrate the potential to identify the specific

310 spectral features which lead to the differentiation of these layers on the basis of the

311 Raman spectral features characteristic for each layer.

312 **KMCA of HSIL tissue**

313 Raman spectral results of a HSIL (CIN 2) cervical tissue sample are presented in

314 **Figure 3,** wherein **Figure 3a** shows a H&E stained tissue section which has abnormal

315 and normal tissue areas adjacent to each other. These regions are ideal to investigate

316 how the layer differentiation is affected by the proliferation of disease. **Figure 3 b**

317 shows the optical image of the parallel unstained tissue section as viewed by the

318 Raman micro spectrometer. KMCA of the Raman map of the same sample is shown

319 in **Figure 3 c**.

320 The KMCA of the Raman map (**Figure 3 c)** clearly shows the separation of the

321 normal and abnormal regions. The normal region is separated into three layers,

322 including stroma (indicated by the yellow and blue colour clusters with a mean

323 spectrum having characteristic bands of collagen, **Figure 3d**), basal layer (indicated

324 by the red colour cluster with mean spectrum having characteristic bands of DNA

325 bases) and squamous epithelium (indicated by the green colour cluster, with mean

326 spectrum having characteristic bands of glycogen). The reason for the two clusters

327 (yellow and blue) in the stromal layer is the difference in the intensity of the collagen

328 related bands. Notably, the stromal layer also has points which are clustered with the

329 basal layer, indicating increased DNA, or reduced collagen content. This will not be

330 discussed further here, but will be the subject of a further communication.

331 In contrast, the abnormal region of the tissue section is separated into only two layers,

332 representing stromal and basal layers. The representative mean spectrum (**Figure 3d**)

333 from KMCA of the Raman map of this region indicates the presence of the

334 characteristic features of the basal cells in the superficial layer, as characterised by the

335 Raman spectral features of thymine (755 cm^{-1}), adenine (722 cm^{-1}) and cytosine (782

336 cm^{-1}).

337 Thus, the analysis indicates three distinct layers, stromal, basal and superficial layers

338 for the normal region, and two distinct layers, stromal and basal layers for the

339 abnormal region in **Figure 3**. This is consistent with the diagnosis and confirms that

340 the abnormal region is characterized by increased DNA contributions and reduced

341 glycogen contributions in the abnormal cells.

342 **KMCA of LSIL tissue**

343 **Figure 4 a** shows the H&E stained image of an LSIL (CIN 1) tissue section which

344 exhibits the normal pattern of the stroma, basal, parabasal, intermediate and

345 superficial epithelium. **Figure 4 b** shows the extended video image of the same

346 unstained tissue section with a X100 objective, acquired using the Raman micro

347 spectrometer and the red box indicates the area analysed. KMCA of the Raman map is

348 shown in **Figure 4 c**.

349 The images indicate an intact stroma, distinct from the basal layer, as confirmed by

350 the Raman bands of the collagen, including 849 cm^{-1}, 921 cm^{-1}, 938 cm^{-1} and 1245

351 cm^{-1}, in the blue mean spectrum in **Figure 4d**. The separation of the stroma and basal

352 layer is very clear, but, at the same time, the squamous epithelium has three types of

353 clusters, two with strong Raman spectral features of glycogen (green and yellow

354 clusters) and one with the characteristic Raman spectral features of the basal layer

355 (red cluster), indicating characteristics of basal cells in this layer.

356 KMCA indicates that the superficial layer shows evidence of disease, based on strong

357 DNA features throughout this layer. This is unexpected in the LSIL tissue as

358 abnormal cells would be expected to be locally confined, adjacent to the basal region,

359 and constitute only a third of the epithelium. Further investigation was thus performed

360 using PCA.

361

362 **PCA of NILM basal layer versus LSIL basal layer**

363 PCA of the Raman spectra from the basal layer of the NILM cervical epithelium

364 against the LSIL basal layer (**Figure 4**) is shown in **Figure 5 A & B** as scatter plot

365 and loadings of the PCA respectively. The PCA scatter plot shows clear

366 differentiation of the two groups by clustering the Raman spectra of the normal basal

367 layer in the negative axis and the spectra of the LSIL basal layer in the positive axis of

368 the scatter plot and hence negative and positive loadings are associated with them

369 respectively.

370 The loadings of the PCA, including those at 480cm^{-1}, 849 cm^{-1}and 939 cm^{-1}

371 (glycogen), 1346 cm^{-1} (guanine), 1056cm^{-1}, 1088 cm^{-1} (O-P-O of DNA) and those

372 associated with proteins, at 831 cm^{-1} (tyrosine), 1144cm^{-1} (C-C stretching), 1222cm^{-1}

373 (amide-III beta sheet) are contributed by the LSIL basal layer. The loadings

374 associated with the normal basal layer include those at 728cm^{-1} (adenine), 783cm^{-1}

375 (cytosine), 1375cm^{-1} (thymine), 1482cm^{-1} (guanine), 1449cm^{-1} (CH deformation) and

376 1577cm^{-1} (tryptophan ring breathing).

377 PCA shows clear differentiation of the normal basal layer and the LSIL basal layer,

378 indicating that cells of the normal basal layer are clearly different to the cells of the

379 LSIL basal layer.

380

381 **PCA results of NILM basal layer vs HSIL basal layers**

382 PCA was also performed for the NILM basal layer and the HSIL basal layers to

383 elucidate any differences between the normal and abnormal areas shown in **Figure 3**.

384 As indicated in **Figure 3c**, HSIL-basal-a is the normal area and HSIL basal-b is the

385 abnormal area which can be further subdivided into two parts, HSIL basal-bi and

386 HSIL basal-bii.

387 PCA of the normal basal layer vs the normal area of the HSIL basal layer (HSIL

388 basal-a) is presented in **Figure 5 C & D** as scatter plot and loadings respectively. It

389 may be helpful in identifying the Raman spectral features which may be taken as the

390 markers of the early stage of the disease or a pre-disease stage.

391 Interestingly, although the HSIL basal-a region was classified as normal, the PCA

392 scatter plot differentiates the two regions by PC1 (71.6%). The Raman spectra of the

393 normal basal layer are clustered in the negative axis and spectra of the HSIL basal-a

394 region in the positive axis of the scatter plot. This indicates that the loadings of the

395 PCA which are negative, including those at 480 cm^{-1}, 849 cm^{-1} and 939 cm^{-1}

396 (glycogen), and some loadings related to DNA, including the feature at 718 cm^{-1}

397 (adenine), are contributed by the normal basal layer. The positive loadings associated

398 with the DNA, including 1088 cm^{-1} (O-P-O of DNA) and 1346 cm^{-1} (guanine) and

399 1202 cm^{-1}, 1222 cm^{-1} (amide-III beta sheet) and 1675 cm^{-1} (amide-I) are contributed

400 by the HSIL basal-a layer. Thus, PCA differentiates the HSIL basal-a layer from the

401 normal basal layer and confirms that HSIL basal-a is not a true normal region in terms

402 of biochemical composition.

403 PCA analysis of the normal basal layer versus the abnormal area of the HSIL basal

404 layer (HSIL basal-bi) was performed to explore biochemical differences between

405 these layers, as HSIL basal-bi may be expected to be closer in biochemical

406 composition to the normal basal layer, as compared to HSIL basal-bii layer.

407 PCA results of the normal basal layer vs HSIL basal-bi layer are presented in **Figure**

408 **5 E & F** as scatter plot and loadings respectively. It can be seen in the PCA scatter

409 plot that the two groups are well differentiated and the Raman spectra of the normal

410 basal layer are clustered in the negative axis, while the spectra of the HSIL basal-bi

411 layer are clustered in the positive axis. Notably, the major difference which can be

412 identified at first glance is the presence of glycogen related bands at 482 cm^{-1}, 849

413 cm^{-1} and 938 cm^{-1} in the negative sense, solely contributed by the Raman spectra of

414 the normal basal layer. This indicates that the normal basal layer, as expected, has

415 more glycogen in the cells as compared to the HSIL basal-bi layer. Moreover, this

18

416 also means that the cells of the HSIL basal-bi layer are more proliferative and more

417 immature as compared to the normal basal layer and hence the glycogen content is

418 reduced due to its consumption by the cells during the process of the cell growth or

419 proliferation.

420 Other negative loadings of PC1 including those at 524 cm^{-1} (S-S stretching), 718 cm^{-1}

421 (adenine) and 1348 cm^{-1} (tryptophan ring breathing) are contributed by the normal

422 basal and the positive loadings, including those at 669 cm^{-1} (thymine), 788 cm^{-1}, 1088

423 cm^{-1} (O-P-O of DNA) and 1062 cm^{-1} (C-N stretching) correspond to the HSIL basal-

424 bi layer. This confirms that HSIL basal-bi layer is not biochemically normal and has

425 more dividing/DNA rich cells as compared to the cells of the normal basal layer.

426 Another prominent loading is that at 1222 cm^{-1} (amide-III beta sheet) which is

427 contributed by the HSIL basal-bi layer and differentiates the Raman spectra of this

428 layer from those of the normal basal layer. It should be noted that this band also

429 contributed to the differentiation of the normal basal layer versus LSIL basal layer

430 (**Figure 5 A & B**) and HSIL basal-a layer (**Figure 5 C & D**). The prominence of this

431 band has been observed using a number of NILM tissue samples for comparison, as

432 shown in supplementary information **figure S2**. This suggests that this signature could

433 be associated with the early stages of disease and, given the established aetiology,

434 with HPV infection in the cervical tissue samples.

435 PCA analysis of the normal basal layer versus the HSIL basal-bii layer was also

436 carried out to explore biochemical differences between these layers, as the HSIL

437 basal-bii layer might be expected to be different from the normal basal layer and

438 resemble the superficial layer of the normal sample.

439   PCA results of the normal basal layer vs the HSIL basal-bii layer are presented in

440   **Figure 5 G & H** as scatter plot and loadings, respectively. It can be seen in the PCA

441   scatter plot that the two groups are different from each other and the Raman spectra of

442   the normal basal layer are clustered in the negative axis and spectra of the HSIL

443   basal-bii layer are clustered in the positive axis. Once again, the major difference

444   between the two groups of the Raman spectra appearing in the form of the loadings is

445   the presence of the glycogen related bands at 482 cm^{-1}, 849 cm^{-1} and 938 cm^{-1}, which

446   are solely contributed by the Raman spectra of the normal basal layer (in the

447   negative). This confirms, as described earlier, that the cells of the normal basal layer

448   are rich in glycogen as compared to the cells of the HSIL basal-bii layer. In addition,

449   the cells of the HSIL basal-bii layer appear to be more proliferative in nature as

450   compared to those of the normal basal layer and hence glycogen content is reduced

451   due to its consumption by the cells during the process of cell growth and/or

452   proliferation. Other positive loadings contributed by the HSIL basal-bii layer include

453   those at 669 cm^{-1}, 759 cm^{-1} (thymine), 1062 cm^{-1} (C-N stretching) and by the normal

454   basal layer include 644 cm^{-1}and 1108 cm^{-1} (C-C stretching) and 1348 cm^{-1}

455   (tryptophan ring breathing). Thus, the HSIL basal-bii region is well differentiated

456   from the normal basal layer and seems to have more dividing/DNA rich cells as

457   compared to the cells of the normal basal layer. In addition, no positive loadings

458   associated with glycogen are seen, which are normally associated with the superficial

459   layer. It can be concluded that the HSIL basal-bii layer is not similar or close to the

460   biochemical composition of the superficial layer and this has also been confirmed

461   using PCA (see supplementary information **figure S3**). The loading observed at 1222

462   cm^{-1} (amide-III beta sheet), contributed by the HSIL basal-bii layer, is again the

463   predominant difference between these two groups of spectra and is consistent with the

previous PCA results, further indicating that it may be associated with the early onset of disease.

PCA of HSIL basal-a layer versus HSIL basal-bi layer versus HSIL basal-bii layer

PCA has also been performed on HSIL basal-a and HSIL basal-b, the normal and abnormal areas of the cervical tissue section presented in **Figure 3c,** respectively. On the basis of the fact that DNA related Raman features are predominant in the abnormal area as compared to the normal area, the HSIL basal-b region is expected to give more DNA related features as compared to the HSIL basal-a region. PCA of the HSIL basal a versus HSIL basal-bi and HSIL basal-bii was performed to explore biochemical differences between these layers, as HSIL basal-bii might be expected to be different to HSIL basal-a (as compared to HSIL basal-bi) and resemble the superficial layer of the true normal sample. It should be noted that HSIL basal-bii derives from the superficial layer and has been termed basal due to the classification of KMCA on the basis of the observation of the DNA related bands, in **Figure 3d**.

PCA results of the HSIL basal-a versus HSIL basal-bi are presented in **Figure 6** as scatter plot (**Figure 6A**) and loadings (**Figure 6B**) of the PCA. There is little or no differentiation between the two groups of spectra indicating that they are similar in biochemical nature. The major loadings in this case are only related to proteins and include those at 1004 cm^{-1} (phenyl alanine), 1129 cm^{-1}, 1142 cm^{-1} (C-N stretching) and 1449 cm^{-1} (CH deformation).

This finding confirms the PCA finding from **Figure 5 D** that the normal basal layer and the HSIL basal-a layer were not biochemically similar, i.e. the HSIL basal-a layer was not biochemically normal despite appearing normal morphologically.

488   PCA results of the HSIL basal bi and HSIL basal-bii, two subdivisions of the HSIL

489   basal-b layer, indicated in **Figure 3c**, which is the abnormal area of the HSIL sample,

490   are presented here in **Figure 6 C & D** as scatter plot and loadings respectively. The

491   purpose of this analysis was to establish whether these two groups are similar or

492   different in their biochemical composition, affected by the progression of the disease,

493   and on what basis the KMCA clustered the Raman data.

494   It is clear from the PCA scatter plot that the two groups are not very different from

495   each other as, although they are partially differentiated, they overlap to some extent.

496   The Raman spectra of HSIL basal-bi are largely clustered in the negative axis and

497   spectra of the HSIL basal-bii in the positive axis. This means that the loadings of the

498   PCA which are negative, including 669 cm^{-1}(thymine), 782 cm^{-1}(cytosine) and 1062

499   cm^{-1}(O-P-O of DNA), correspond to the HSIL basal-bi region and the positive ones

500   including 718 cm^{-1}, 1238 cm^{-1}(adenine), 1260 cm^{-1}(thymine) and 1278 cm^{-1}(cytosine),

501   correspond to the HSIL basal-bii region. Notably, the spectral groups are not very

502   different, as indicated by the scatter plot and reflected in the presence of DNA

503   features (thymine and cytosine) as both positive and negative loadings of the PC.

504   It should be noted that the loadings of the PCA between HSIL basal-bi and HSIL

505   basal-bii (**Figure 6 D**) are very different than those of the normal basal layer versus

506   the normal superficial layer (**Figure 2 D**) as there are no representations of glycogen

507   in the former as compared to the latter. The differentiating features are rather

508   associated with DNA, characteristic of rapidly proliferating cells. The findings are

509   also consistent with the PCA results showing clear differentiation between the normal

510   superficial layer and the HSIL basal-bii layer (supplementary information **figure S3**).

511   Under normal conditions, the basal layer provides cells to the superficial layer, but

512   those cells are under controlled growth and have the ability to store glycogen during

513   the process of the maturation, and only matured cells, having stored glycogen, move

514   into the superficial layer. This may lead to the conclusion that, during the progression

515   of cervical cancer, characteristic biochemical features of the rapidly proliferating

516   basal cells, rich in DNA and lacking in glycogen, appear in the superficial layer due to

517   the progression of the disease. The KMCA clustering of the HSIL basal bii layer with

518   the HSIL basal bi layer is a representation of this process.

519   PCA results of the HSIL basal-a versus HSIL basal-bii are presented in **Figure 6 (E)**

520   as scatter plot and **Figure 6 (F)** as loadings of the PCA. There is very good

521   differentiation between the two groups of spectra, indicating that they are different in

522   biochemical nature. In the PCA scatter plot, HSIL basal-a, is clustered in the negative

523   axis and HSIL basal-bii in the positive and they are hence associated with the negative

524   and positive loadings respectively. Loadings associated with HSIL basal-bii include

525   those at 669 cm^{-1} (thymine), 718 cm^{-1} (adenine), 782 cm^{-1} (cytosine), 888 cm^{-1}

526   (deoxyribose ring breathing) and 825 cm^{-1} (O-P-O of DNA), which are associated

527   with DNA, indicating the presence of rapidly proliferating cells having high DNA

528   content. On the other hand HSIL basal-a has contributed loadings associated only

529   with proteins, including those at 1129cm^{-1}, 1158 cm^{-1} (C-N stretching), 1172cm^{-1}

530   (tyrosine ring breathing), 1449 cm^{-1} (CH deformation) and 1675 cm^{-1} (amide-I).

531   Notably the loadings of HSIL basal-a versus HSIL basal-bii and HSIL basal-a versus

532   HSIL basal-bi are very similar, although HSIL basal-a shows some features associated

533   with glycogen (482cm^{-1}).

534

535   The repeated observation of the 1222cm^{-1} band, associated with the amide III band of

536   beta sheeted proteins, in differentiating normal and early stage disease of the cervix

suggests that it may be a spectroscopic marker of the early stages of the onset. As cervical cancer is predominantly associated with high risk HPV, it may therefore, by extension, be an indicator of infection. Should this be the case, it is more likely a marker for the physiological effects of HPV infection in tissue, rather than the spectroscopic bands associated with the virus itself, due to the very small genome size of the HPV as compared to the human genome (Diem *et al*, 2012). The intracellular targets for HPVs include a number of regulatory proteins such as cyclins, cyclin dependent kinases, cyclin inhibitors, and cell cycle-associated proteins and p16^{INK4A} overexpression has been demonstrated in cervical cancers as a result of functional inactivation of retinoblastoma protein by the HPV E7 protein (Tam *et al*, 1994). p16^{INK4A} expression levels have been shown to be correlated with degree of HPV infection in cervical cell lines (Ostrowska *et al*, 2011) and to the degree of CIN in tissue biopsies (Lesnikova *et al*, 2009). However, in an FTIR study of cervical cell lines of varying HPV infection, and correlated p16^{INK4A} expression levels, only weak variations of the cytoplasmic amide III features were observed (Ostrowska *et al*, 2011) and although a PCA analysis of Raman spectra of pellets of the same cell lines showed a good differentiation between the cell lines, any features which could be related to the amide III band were evident only in PC2 and PC3 (Ostrowska *et al*, 2010). Indeed, a pairwise PCA of the same data revealed no strong contribution of the amide III feature to the discrimination of the HPV negative cell lines from progressively HPV positive cell lines (Rashid, 2013).

Conclusion

Raman micro spectroscopy together with KMCA was able to differentiate NILM cervical tissue into three layers including stroma, basal/parabasal and superficial layers on the basis of the spectral features of collagen, DNA bases and glycogen respectively. In the HSIL tissue with both normal and abnormal regions, KMCA showed three distinct layers, stromal, basal and superficial layers for the normal region, and two distinct layers, stromal and basal layers for the abnormal region. This finding confirms that the abnormal cells are characterised by increased DNA contributions and reduced glycogen contributions. Further analysis by PCA, however showed that the normal region was not in fact normal and could be differentiated from the NILM samples. Interestingly, KMCA showed strong DNA features throughout the superficial layer of the LSIL tissue. As this tissue was classified as LSIL, abnormal cells would be expected only in the basal third of the epithelium rather than throughout the epithelium. These findings indicate that Raman microspectroscopy can identify biochemical changes in tissue where morphological changes are not yet prominent. The 1222cm^{-1} Raman feature, associated with the amide III band of beta sheeted proteins, was consistently observed in differentiating normal and pre-malignant tissue suggesting that it may be a spectroscopic marker of the early biochemical changes.

In conclusion, this study has shown that Raman microspectroscopy in combination with KMCA and PCA can be useful for the early detection of pre-malignant changes in cervical tissue.

589

590   **Conflict of Interest**

591   The authors declare no conflict of interest.

592 **References**

593 Bazant-Hegemark F, Edey K, Swingler GR, Read MD, Stone N (2008) Optical

594 Micrometer Resolution Scanning for Non-invasive Grading of Precancer in the

595 Human Uterine Cervix. *Technology in Cancer Research & Treatment* **7**(6)**:** 483-496

596

597 Bonnier F, Byrne HJ (2012) Understanding the molecular information contained in

598 principal component analysis of vibrational spectra of biological systems. *Analyst*

599 **137**(2)**:** 322-332

600

601 Chan JW, Taylor DS, Zwerdling T, Lane SM, Ihara K, Huser T (2006) Micro-Raman

602 spectroscopy detects individual neoplastic and normal hematopoietic cells.

603 *Biophysical Journal* **90**(2)**:** 648-656

604

605 Diem M, Miljkovic M, Bird B, Chernenko T, Schubert J, Marcsisin E, Mazur A,

606 Kingston E, Zuser E, Papamarkakis K, Laver N (2012) Applications of Infrared and

607 Raman Microspectroscopy of Cells and Tissue in Medical Diagnostics: Present Status

608 and Future Promises. *Spectroscopy-an International Journal* **27**(5-6)**:** 463-496

609

610 Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011) Global Cancer

611 Statistics. *Ca-a Cancer Journal for Clinicians* **61**(2)**:** 69-90

612

613 Jess PRT, Smith DDW, Mazilu M, Dholakia K, Riches AC, Herrington CS (2007)

614 Early detection of cervical neoplasia by Raman spectroscopy. *International Journal of*

615 *Cancer* **121**(12)**:** 2723-2728

616

617   Kamemoto LE, Misra AK, Sharma SK, Goodman MT, Luk H, Dykes AC, Acosta T

618   (2010) Near-Infrared Micro-Raman Spectroscopy for in Vitro Detection of Cervical

619   Cancer. *Applied spectroscopy* **64**(3)**:** 255-261

620

621   Kanter EM, Majumder S, Kanter GJ, Woeste EM, Mahadevan-Jansen A (2009a)

622   Effect of hormonal variation on Raman spectra for cervical disease detection.

623   *American journal of obstetrics and gynecology* **200**(5)**:** 512. e1-512. e5

624

625   Kanter EM, Vargis E, Majumder S, Keller MD, Woeste E, Rao GG, Mahadevan-

626   Jansen A (2009b) Application of Raman spectroscopy for cervical dysplasia

627   diagnosis. *Journal of Biophotonics* **2**(1-2)**:** 81-90

628

629   Knief P. PhD Thesis. Dublin Institute of Technology, 2010

630

631   Krishna CM, Prathima NB, Malini R, Vadhiraja BM, Bhatt RA, Fernandes DJ,

632   Kushtagi P, Vidyasagar MS, Kartha VB (2006) Raman spectroscopy studies for

633   diagnosis of cancers in human uterine cervix. *Vibrational Spectroscopy* **41**(1)**:** 136-

634   141

635

636   Lesnikova I, Lidang M, Hamilton-Dutoit S, Koch J (2009) p16 as a diagnostic marker

637   of cervical neoplasia: a tissue microarray study of 796 archival specimens. *Diagn*

638   *Pathol* **4:** 22

639

640   Lyng FM, Faolain EO, Conroy J, Meade AD, Knief P, Duffy B, Hunter MB, Byrne

641   JM, Kelehan P, Byme HJ (2007) Vibrational spectroscopy for cervical cancer

pathology, from biochemical analysis to diagnostic tool. *Experimental and Molecular Pathology* **82**(2)**:** 121-129

McCluggage WG, Walsh MY, Thornton CM, Hamilton PW, Date A, Caughley LM, Bharucha H (1998) Inter- and intra-observer variation in the histopathological reporting of cervical squamous intraepithelial lesions using a modified Bethesda grading system. *British Journal of Obstetrics and Gynaecology* **105**(2)**:** 206-210

Notingher I (2007) Raman Spectroscopy cell-based Biosensors. *Sensors* **7**(8)**:** 1343-1358

Notingher I, Verrier S, Haque S, Polak JM, Hench LL (2003) Spectroscopic study of human lung epithelial cells (A549) in culture: Living cells versus dead cells. *Biopolymers* **72**(4)**:** 230-240

Ó Faoláin E, Hunter MB, Byrne JM, Kelehan P, McNamara M, Byrne HJ, Lyng FM (2005) A study examining the effects of tissue processing on human tissue sections using vibrational spectroscopy. *Vibrational Spectroscopy* **38**(1)**:** 121-127

Ooi GJ, Fox J, Siu K, Lewis R, Bambery KR, McNaughton D, Wood BR (2008) Fourier transform infrared imaging and small angle x-ray scattering as a combined biomolecular approach to diagnosis of breast cancer. *Medical Physics* **35**(5)**:** 2151-2161

666 Ostrowska KM, Garcia A, Meade AD, Malkin A, Okewumi I, O'Leary JJ, Martin C,
667 Byrne HJ, Lyng FM (2011) Correlation of p16INK4A expression and HPV copy
668 number with cellular FTIR spectroscopic signatures of cervical cancer cells. *Analyst*
669 **136**(7)**:** 1365-1373

670

671 Ostrowska KM, Malkin A, Meade A, O'Leary J, Martin C, Spillane C, Byrne HJ,
672 Lyng FM (2010) Investigation of the influence of high-risk human papillomavirus on
673 the biochemical composition of cervical cancer cells using vibrational spectroscopy.
674 *Analyst* **135**(12)**:** 3087-3093

675

676 Papanicolaou GN, Traut HF (1941) The diagnostic value of vaginal smears in
677 carcinoma of the uterus. *American Journal of Obstetric and Gynecology* **42**(2)**:** 193-
678 206

679

680 Parkin DM, Bray F, Ferlay J, Pisani P (2005) Global cancer statistics, 2002. *Ca-a*
681 *Cancer Journal for Clinicians* **55**(2)**:** 74-108

682

683 Rashid N. Raman micro spectroscopy for the characterisation of cervical cancer. PhD,
684 Dublin Institute of Technology, Dublin, 2013

685

686 Robichaux-Viehoever A, Kanter E, Shappell H, Billheimer D, Jones H, Mahadevan-
687 Jansen A (2007) Characterization of Raman spectra measured in vivo for the detection
688 of cervical dysplasia. *Applied Spectroscopy* **61**(9)**:** 986-993

689

690   Sellors W, Sankaranarayanan R (2003) *Colposcopy and Treatment of Cervical*

691   *Intraepithelial Neoplasia: A Beginner's Manual*: Published by the International

692   Agency for Research on Cancer

693

694   Tam SW, Shay JW, Pagano M (1994) Differential expression and cell cycle

695   regulation of the cyclin-dependent kinase 4 inhibitor p16Ink4. *Cancer research*

696   **54**(22)**:** 5816-5820

697

698   Tan KM, Herrington CS, Brown CT (2011) Discrimination of normal from pre-

699   malignant cervical tissue by Raman mapping of de-paraffinized histological tissue

700   sections. *J Biophotonics* **4**(1-2)**:** 40-8

701

702   Wang LQ, Mizaikoff B (2008) Application of multivariate data-analysis techniques to

703   biomedical diagnostics based on mid-infrared spectroscopy. *Analytical and*

704   *Bioanalytical Chemistry* **391**(5)**:** 1641-1654

705

706

707

708

709

Titles and Legends to Figures

Figure 1 (a) H&E stained NILM tissue, (b) unstained tissue under Raman spectroscope, (c) Five cluster K means cluster map generated from the Raman map, (d) K-means cluster spectra, blue representing the stroma (collagen), red representing the basal layer (DNA) and green representing the superficial layer (glycogen), for the Raman map of the marked region in c.

Figure 2 PCA scatter plot and loading (PC1) for: (A, B) normal basal vs normal stromal layer and (C, D) normal basal vs normal superficial layer.

Figure 3 (a) H&E stained normal and abnormal region of HSIL (CIN 2) tissue with scale bar of 200µm, (b) unstained tissue under Raman spectroscope, (c) Five cluster KMCA map generated from the Raman map, (d) Representative mean Raman spectra from KMCA of Raman map, yellow and blue clusters corresponding to stroma, red corresponding to basal layer, and green corresponding to squamous epithelium from the marked region.

Figure 4 (a) H&E stained region of LSIL (CIN 1) tissue with scale bar of 500µm, (b) unstained tissue under Raman spectroscope, with scale bar of 100µm, (c) Five cluster KMCA map generated from the Raman map, (d) and K-means cluster spectra, blue representing stroma (collagen), red representing basal (DNA), green & yellow representing superficial (glycogen) layers, for the Raman map of the marked region.

Figure 5 PCA scatter plot and loadings (PC1) of NILM basal layer versus LSIL basal (A, B), HSIL basal-a (C, D), HSIL basal-bi (E, F), HSIL basal-bii (G, H).

Figure 6 PCA scatter plot and loadings (PC1) for HSIL basal-a layer versus HSIL basal-bi layer (A, B), HSIL basal-bi layer vs HSIL basal-bii layer (C, D), HSIL basal-a layer versus HSIL basal-bii layer (E, F).

734   **Figure S1** PCA scatter plot (A) and PC1 loadings (B) of the basal regions of two

735   NILM samples**.** PC1 accounts for only 56% of the variability, and the dominant

736   differentiating features are largely attributed to proteins and lipids.

737   **Figure S2** PCA scatter plot (A) and PC1 loadings (B) of the differentiation of the

738   basal layer of a second NILM sample from the LSIL basal layer. Again, the amide III

739   at 1222cm^{-1} is dominant.

740   **Figure S3** PCA scatter plot (A) and loadings of PC1 (B) of the normal superficial

741   layer and the HSIL basal-bii layer showing good differentiation with the loadings

742   contributed by the Raman spectra of the HSIL basal-bii layer attributed to DNA and

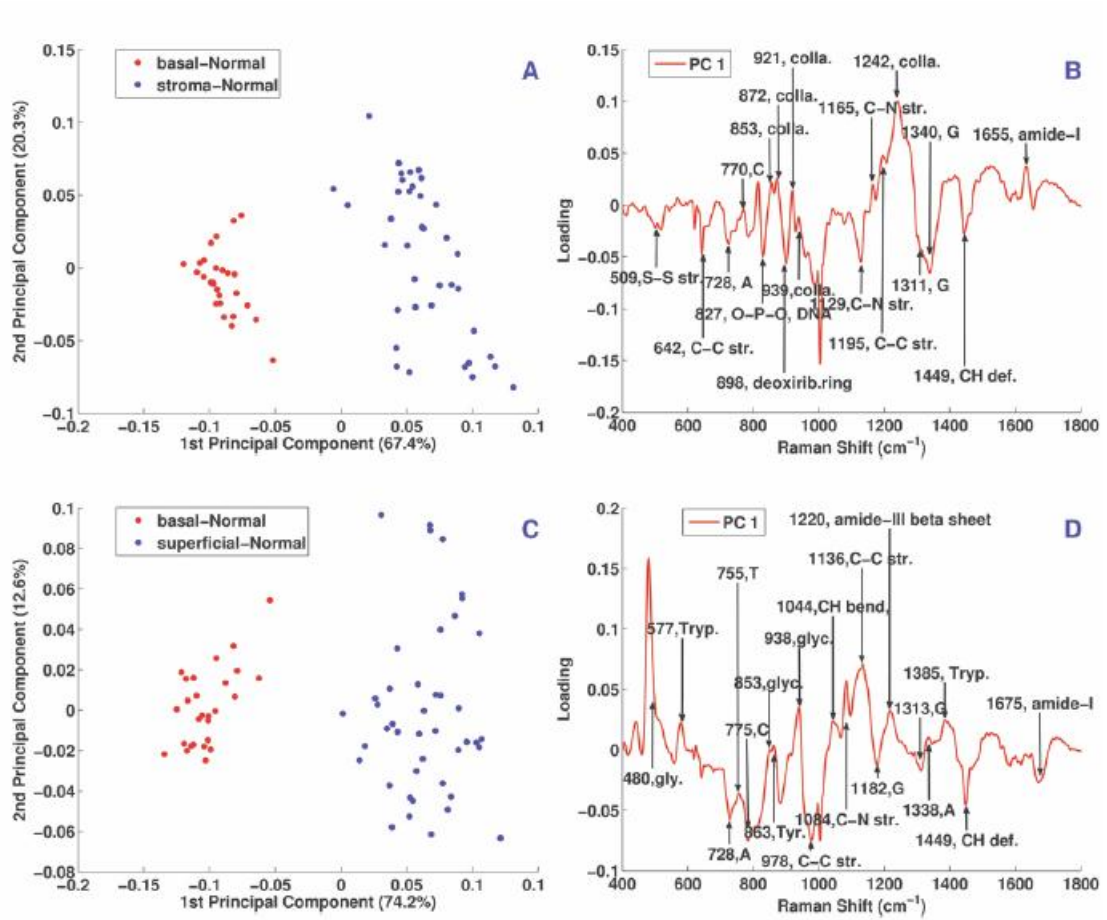743   no glycogen related loadings contributed by the cells in this layer.
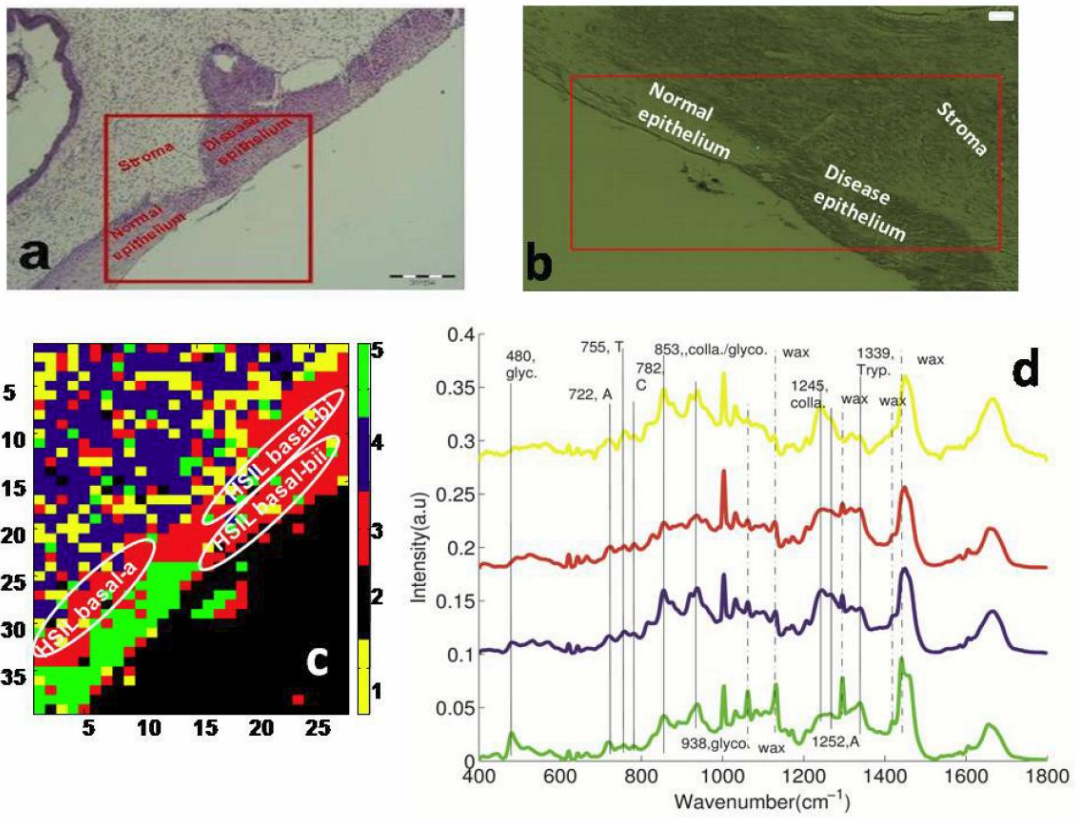
744

745

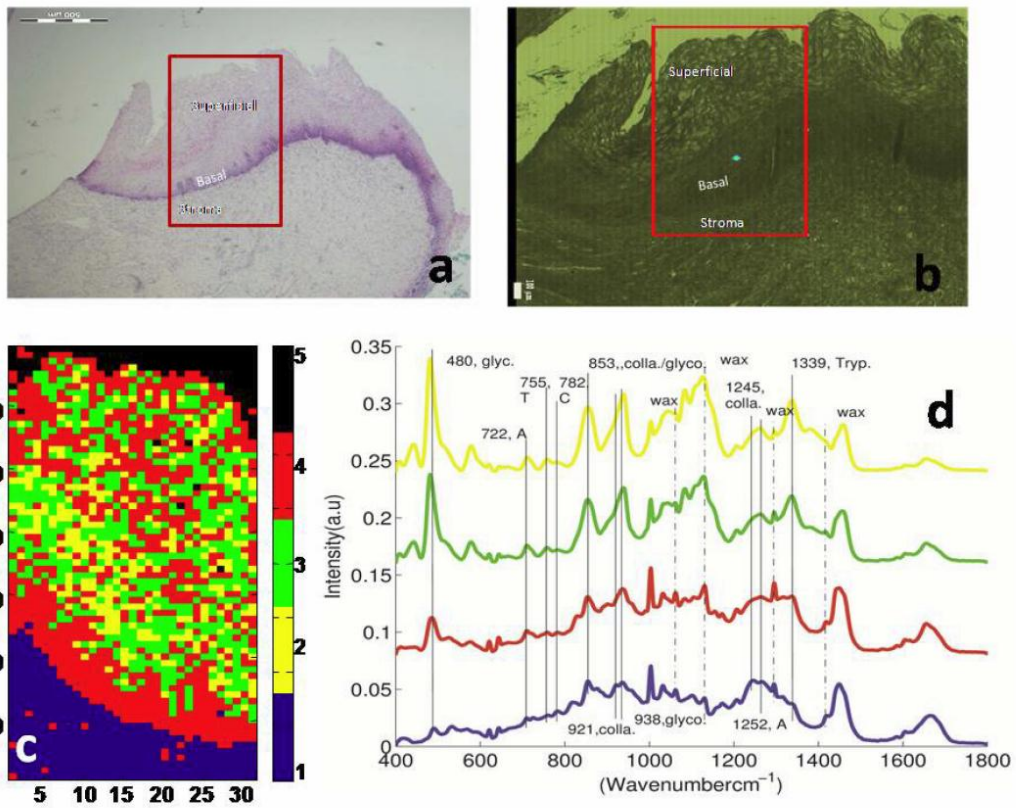746   Figure 1
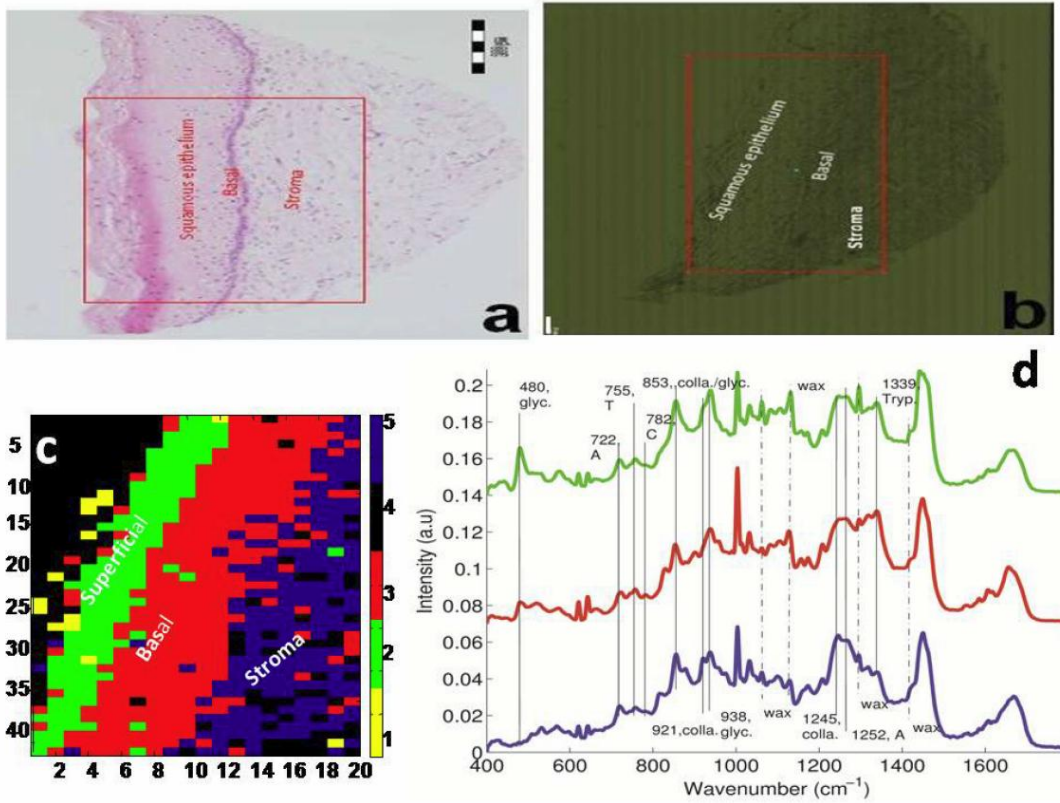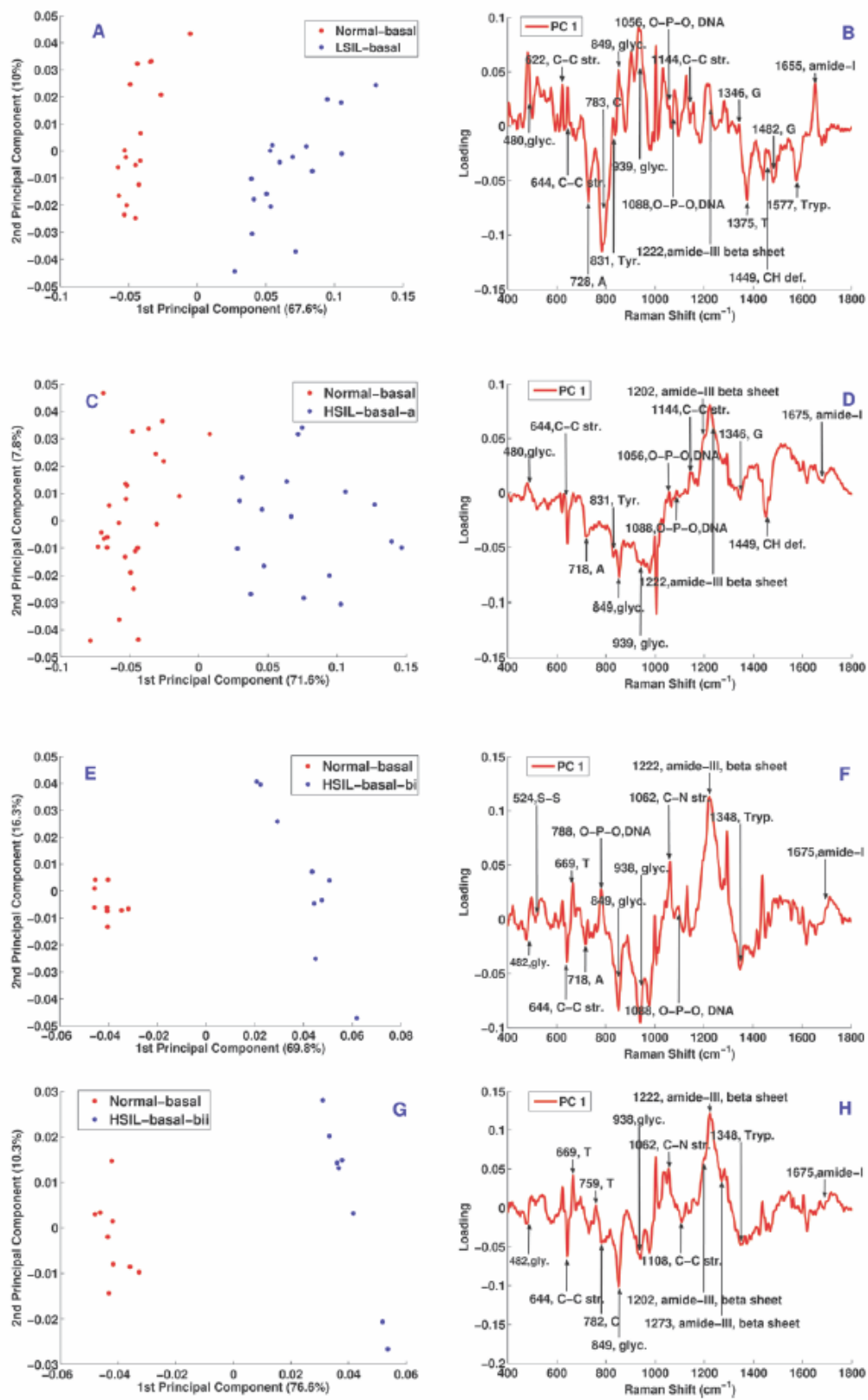

747

748

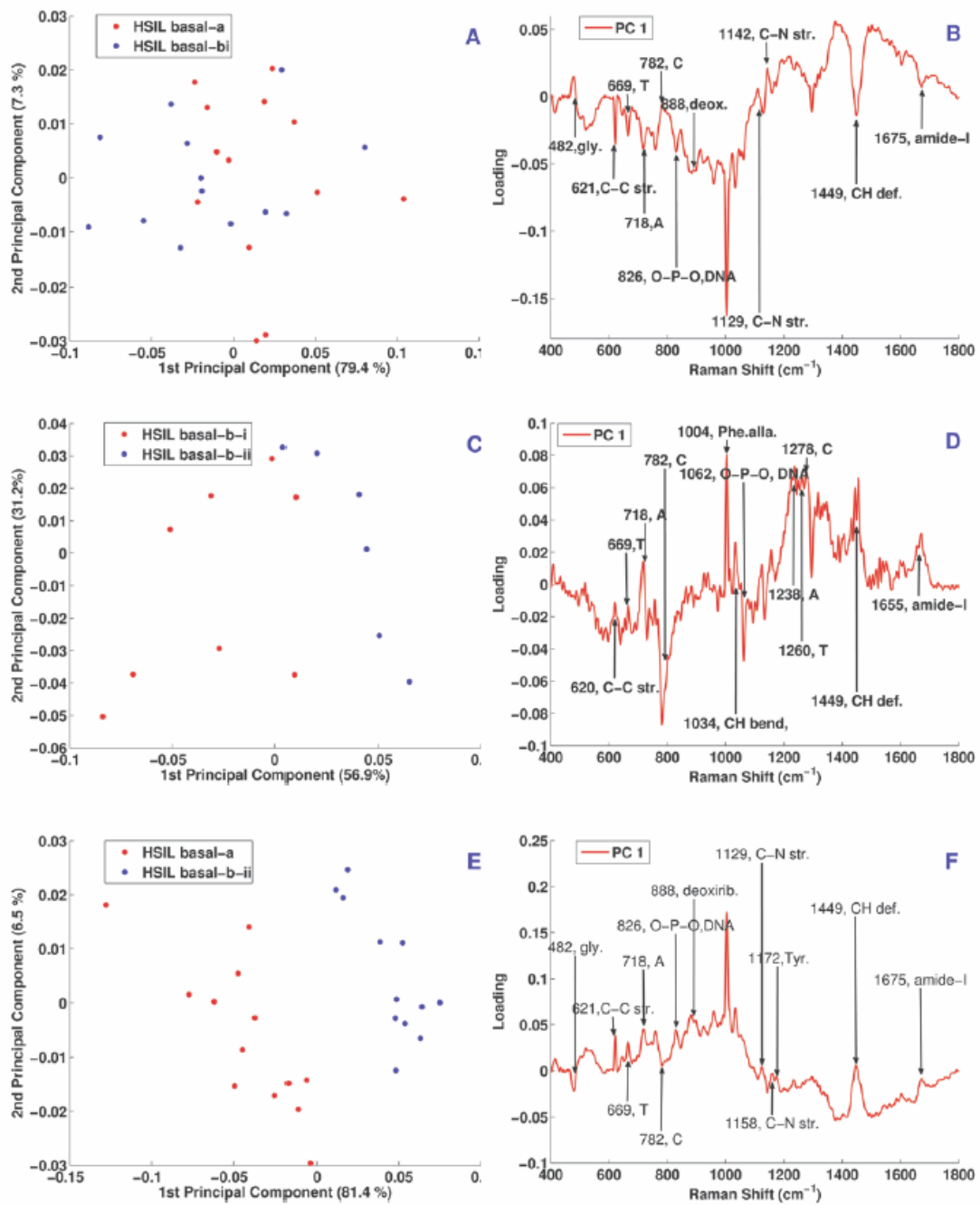749   Figure 2

750

751

752   Figure 3

753

754

755   Figure 4

756

757

Figure 5

761   Figure 6

762

763

764

765