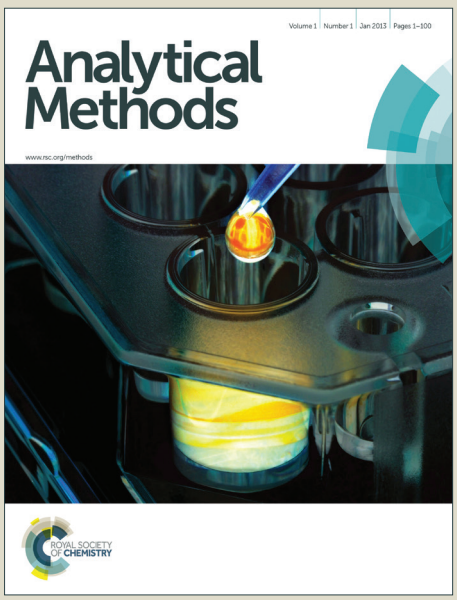
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Classification of cervical cytology for human papilloma virus (HPV) infection using biospectroscopy and variable selection techniques

Kássio M.G. Lima[1,2], Ketan Gajjar[1,3], George Valasoulis[4], Maria Nasioutziki[5], Maria Kyrgiou[6], Petros Karakitsos[7], Evangelos Paraskevaidis[4], Pierre L. Martin Hirsch[1,3], Francis L. Martin[1*]

[1]*Centre for Biophotonics, LEC, Lancaster University, Lancaster LA14YQ, UK*

[2]*Institute of Chemistry, Biological Chemistry and Chemometrics, Federal University of Rio Grande do Norte, Natal 59072-970, RN-Brazil*

[3]*Department of Obstetrics and Gynaecology, Central Lancashire Teaching Hospitals NHS Foundation Trust, Preston, UK*

[4]*Department of Obstetrics and Gynaecology, University Hospital of Ioannina, Ioannina, 45500, Greece*

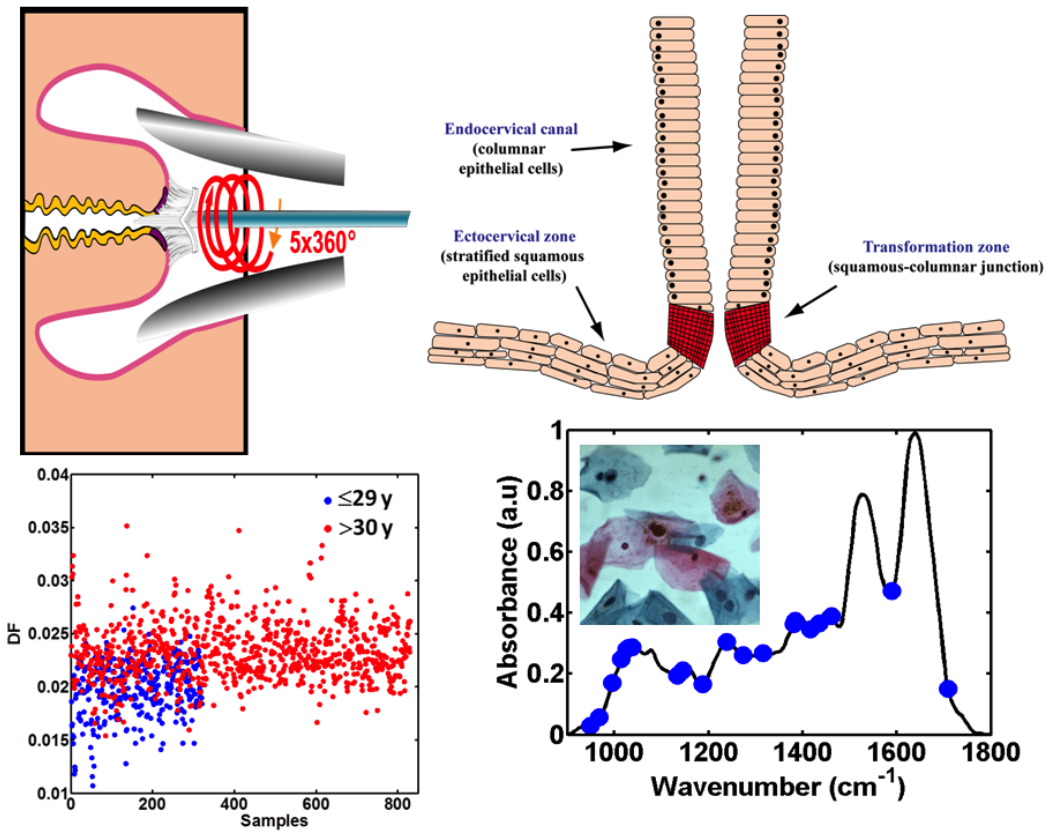
[5]*Molecular & Diagnostic Cytopathology Laboratory, Second Department of Obstetrics and Gynecology, Aristotle University of Thessaloniki, Hippokration Hospital, Thessaloniki, Greece*

[6]*Division of Surgery and Cancer, Faculty of Medicine, Imperial College, UK*

[7]*Department of Cytopathology, "Attikon" Hospital, University of Athens, Greece*

[*]**Corresponding Author:** Prof Francis L Martin PhD, Centre for Biophotonics, LEC, Lancaster University, Lancaster LA1, 4YQ, UK; Tel.: +44(0) 1524 510206; Email: f.martin@lancaster.ac.uk

ToC graphic



Cervical cytology collection towards spectral acquisition followed by variable selection for classification analysis

Abstract Cervical cancer is the second most common cancer in women worldwide. We set out to determine whether attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy combined with principal component analysis-linear discriminant analysis (PCA-LDA) or, variable selection techniques employing successive projection algorithm or genetic algorithm (GA) could classify cervical cytology according to human papilloma virus (HPV) infection [high-risk (hr) *vs.* low-risk (lr)]. Histopathological categories for squamous intraepithelial lesion (SIL) were segregated into grades (low-grade *vs.* high-grade) of cervical intraepithelial neoplasia (CIN) expressing different HPV infection (16/18, 31/35 or HPV Others). Risk assessment for HPV infection was investigated using age (≤29 y *vs.* >30 y) as the distinguishing factor. Liquid-based cytology (LBC) samples (*n*=350) were collected and interrogated employing ATR-FTIR spectroscopy. Accuracy test results including sensitivity and specificity were determined. Sensitivity in hrHPV category was high (≈87%) using a GA-LDA model with 28 wavenumbers. Sensitivity and specificity results for >30 y for HPV, using 28 wavenumbers by GA-LDA, were 70% and 67%, respectively. For normal cervical cytology, accuracy results for ≤29 y and >30 y were high (up to 81%) using a GA-LDA model with 27 variables. For the low-grade cervical cytology dataset, 83% specificity for ≤29 y was achieved using a GA-LDA model with 33 wavenumbers. HPV16/18 *vs*. HPV31/35 *vs*. HPV Others were segregated with 85% sensitivity employing a GA-LDA model with 33 wavenumbers. We show that ATR-FTIR spectroscopy of cervical cytology combined with variable selection techniques is a powerful tool for HPV classification, which would have important implications for the triaging of patients.

Keywords: Biospectroscopy; Cervical cytology; Classification; Human papilloma virus; Variable selection; Wavenumber

Introduction

Extensive laboratory and epidemiological evidence demonstrates that human papilloma virus (HPV) is a major cause of cervical squamous cell carcinoma (SCC), its precursor lesions [cervical intraepithelial neoplasia (CIN)], and other benign or malignant clinical manifestations including genital warts[1]. HPV is a small virus that is ≈55 nm in diameter and comprises a double-stranded circular DNA of nearly 8,000 bp. Its genome encodes eight proteins: early proteins E5, E6 and E7 are involved in cell proliferation and survival, whilst E6 and E7 also play a key role in HPV-associated carcinogenesis[2]. More than 200 genotypes have been identified and associated with benign (low-risk, lrHPV) or malignant (high-risk, hrHPV) cutaneous or mucosal lesions. The hrHPV subtypes 16, 18, 31, 33, and 51 have been recovered from more than 95% of cervical cancers[3]. Studies aimed at describing the distribution of HPV types in invasive cervical cancer strongly implicate subtypes 16 and 18 in approximately 70% of all cervical cancers[4-6]. Worldwide, cancer of the cervix is the second leading cause of cancer death in women: each year, an estimated 493,000 new cases are diagnosed[7].

The distribution of genital HPV types varies and is related to the degree of cervical dysplasia present[8]. HPV6 and 11 are frequently found in sexually-active adults, and are associated with low-grade (LG) squamous intraepithelial lesions (L-SIL). HPV16, 18, 31 and 45 are less frequently found and are associated with progression to invasive cancer. Detection of particular HPV types could be useful in the diagnosis and management of cervical cancer in older women, and for resolving equivocal cytology. HPV assays, which can distinguish between high-grade (HG) and LG disease, may also have a role in routine cervical screening[9].

Early detection and treatment of precancerous lesions can prevent progression to cervical cancer. Identification of precancerous lesions has been primarily achieved by cytologic screening. The modal time is 7-10 y between HPV infection occurring in the late teens or early 20's and pre-cancer peaking around 30 y of age. Invasive cancer arises over many years, even decades, in a minority of women with a peak or plateau in risk at ≈35-55 y of age. Each genotype of HPV is an independent infection, with different carcinogenic risks linked to evolutionary species[10]. Technologies for HPV DNA testing[11] and liquid-based cytology (LBC)[12] are more likely to detect cytologic abnormalities in young women who are at lrHPV for actual invasive cervical disease, opening up a requirement for better triage.

Biospectroscopy techniques include vibrational spectroscopy [infrared (IR) or Raman][13], laser-induced fluorescent spectroscopy[14], optical coherence tomography[15] and confocal imaging[16]. In particular, attenuated total reflection Fourier-transform IR spectroscopy (ATR-FTIR) has shown potential in the field of cervical cancer screening, as an inexpensive but robust technique capable of segregating grades of cytology[17,18]. The fingerprint spectra generated by ATR-FTIR spectroscopy reflects the compositional and quantitative differences of biochemical constituents in cells[19,20]. Peaks within the "biochemical-cell fingerprint" region (1800 cm^{-1} to 900 cm^{-1}) contains spectral features associated with lipids (≈1750 cm^{-1}), Amide I (≈1650 cm^{-1}), Amide II (≈1550 cm^{-1}), methyl groups of lipids and proteins (≈1400 cm^{-1}), Amide III (≈1260 cm^{-1}), asymmetric phosphate stretching vibrations ($\nu_{as}PO_2^-$; ≈1225 cm^{-1}), symmetric phosphate stretching vibrations ($\nu_sPO_2^-$; ≈1080 cm^{-1}), C-OH groups of serine, threonine and tyrosine and C-O groups of carbohydrates (≈1155 cm^{-1}), glycogen (≈1030 cm^{-1}) and protein phosphorylation (≈970 cm^{-1})[21-24].

The principle is that the "biochemical-cell fingerprint" of a liquid-based cytology (LBC) normal (benign) sample is different from that of a dysplastic one, based on alterations in DNA-, RNA-, lipid-, phosphate- and carbohydrate-associated chemical bonds. Furthermore, the spectral fingerprint of a cervical cytology sample could provide a dichotomous biomarker of LG cytology that is committed to progression[13]. The application of chemometric tools to extract discriminating variance from this spectral fingerprint is largely responsible for the advancement of biospectroscopy[25]. For the analysis of biological samples (biofluids, cells or tissues) with IR spectroscopy, principal component analysis (PCA) is often used for initial data reduction[26]; otherwise, hierarchical cluster analysis (HCA) may be applied to analyse groups in a dataset on the basis of their spectral similarities[27], or linear discriminant analysis (LDA) to classify unknown samples into predetermined groups[28]. Many studies employ the entire spectrum in the construction of these mathematical models; herein, many variables are redundant and/or non-informative. A well-developed approach to identify biomarkers or wavenumbers is the successive projection algorithm (SPA) or genetic algorithm (GA) in conjunction with LDA[29,30]. Basically, SPA-LDA and GA-LDA employ a cost function associated with the average risk of misclassification in a validation set and can also reduce the generalization problems often associated with collinearity and avoid over-fitting.

As HPV infection causes changes in expression of cervical cell-cycle regulatory proteins and nucleic acids, a non-invasive biomarker-free analytical technique for identification of alterations in LBC samples associated with lrHPV and hrHPV as a function of age in women would assist our ability to triage cytological atypia. There is a need for an automated, cost-effective tool capable of segregating grades of dysplasia related with age with higher sensitivity and specificity[31].

This study applies IR spectra, or combinations of variables, that reflect a specific biochemical feature of histopathological categories for squamous intraepithelial lesion (SIL), divided into different grades of CIN (low-grade and high-grade) containing different HPV infection (16/18, 31/35 and HPV Others) and subsequently combined into two groups: lrHPV *vs.* hrHPV. In addition, risk assessment of cervical cytology for HPV infection based on age (≤29 y *vs.* >30 y) as a distinguishing factor is an important determinant of a requirement for intervention. We employed SPA and GA to select an appropriate subset of wavenumbers for LDA, allowing the discrimination of different categories of cytology, to identify potential biomarkers and detect dysplasia stages. Cytology samples were categorised into different grades of CIN (LG *vs.* HG) containing different HPV infection (16/18, 31/35 and HPV Others) in order to elucidate altered variables in their spectral fingerprint. This novel approach as a diagnostic tool could be applied to improve accuracy and reduce subjectivity in cervical screening. Lastly, measures of test accuracy, such as sensitivity and specificity were calculated as an important quality standard in test evaluation studies.

Materials and Methods

A retrospective cross-sectional study (October 2009 and August 2012) was coordinated by the University General Hospital of Ioannina, Institutional Review Board (*i.e.*, Ethics Committee) [protocol 28/9-7-2009(s.22)], to estimate the prevalence of HPV DNA types in women with invasive cervical cancer. Ethics committee approval was also obtained from the Institutional Review Board of Hippokration Hospital at University of Thessaloniki [approval number 3715/21-03-2011] for collection of cytology samples at the Second Department of Obstetrics and

Gynaecology, Hippokration Hospital (University of Thessaloniki, Greece). Study participants were fully informed regarding the purposes of the study and consent was obtained. Participants were referred with cervical smear abnormalities or for symptoms such as post-coital bleeding. All underwent a repeat LBC sample collection prior to colposcopic assessment. Decisions regarding no treatment, punch biopsies for suspected intraepithelial lesions or treatment were made by colposcopists. In cases where both the referral cytology and colposcopy were suggestive of high-grade disease (CIN2+), punch biopsies were not considered necessary and treatment with Loop Electrosurgical Excision Procedure (LEEP) was offered to the women.

LBC samples were collected with Rovers™ Cervex-brush in a ThinPrep® solution (Cytyc, USA) and each sample underwent cytological and biomolecular analysis by resident qualified cytopathologists within quality-assured laboratories in two University Hospitals. Cervical cytology is graded as negative, atypical squamous cells of undetermined significance (ASCUS), low-grade squamous intraepithelial lesion (LSIL), high-grade squamous intraepithelial lesion (HSIL) or cancer. Specimens exhibiting viral changes without atypia were classed as HPV or koilocytosis.

In addition to cytology, HPV DNA tests (Clinical arrays HPV, Genomica, Spain) were carried out after extracting DNA from the residuum of the LBC sample using a commercial kit (Purelink, Invitrogen). The analysis for different HPV genotypes was performed with PCR amplification using the CLART® (Clinical Array Technology) HPV2 Kit. This technique is based on the amplification of specific fragments of the viral genome and their hybridization with specific probes for each HPV type. The method assessed the following hrHPV types: 16, 18, 26, 31, 33,35, 39,

43, 45, 51, 52, 53, 56, 58, 59, 66, 68, 70, 73, 82 and 85; and, lrHPV types 6, 11, 40, 42, 44, 54, 61, 62, 71, 72, 81, 83, 84 and 89.

The cytology specimens were categorised as follows: *n*=23 lrHPV and *n*=37 hrHPV types (set A); *n*=90 ≤29 y and *n*=239 >30 y for HPV infection (set B); *n*=32 ≤29 y and *n*=82 >30 y for normal cervical cytology (set C); *n*=29 ≤29 y and *n*=49 >30 y for LG cervical cytology (set D); *n*=42 HPV16/18, *n*=21 HPV31/35, *n*=50 HPV 6/33/39/45/51/52/54/58/59/61/62/66/70/83 (set E).

Samples were sent for spectroscopy analysis after cytological diagnosis was obtained. Six mL of Thin-Prep® from each specimen was analysed. Samples were centrifuged at 1500 rpm for 5 min. The resultant cell pellet, after discarding the methanol (*i.e.*, fixative in Thin-Prep®) was washed with distilled H_2O and centrifuged; this process was repeated three times. The resulting cell pellet was suspended in 0.5 mL of distilled H_2O. The suspensions were applied and left to dry on IR-reflective slides (Low-E; Kevley Technologies Inc., OH, USA). Once dry, samples were desiccated for a further 24 h. This was to remove any possibility of H_2O contaminating specimen spectra. In the event of H_2O contamination, the 3400 cm^{-1} peak tends to become more 'rounded'. In addition, the Amide I lefthand shoulder would be spikey and split with H_2O contamination. The ATR-FTIR spectra are exactly as we would have hoped in terms of being minimally influenced by aqueous and requiring minimal pre-processing (see Electronic Supplementary Information [ESI] Figs. S1 to S5). A Tensor 27 FTIR Spectrometer with Helios ATR attachment (Bruker Optik GmbH) was used to obtain IR spectra (10 per specimen). Instrument settings were 32 scans, spectral resolution of 8 cm^{-1}, and interferogram zero-filling of 2×. Prior to analysing each sample, the diamond crystal was washed and a background spectrum obtained to account for atmospheric composition.

The data import, pre-treatment and construction of chemometric classification models (PCA-LDA, SPA-LDA and GA-LDA) were implemented in MATLAB R2010a software (Mathworks Inc, Natick, MA, USA). IR spectra were pre-processed by cutting between 1,800 and 900 cm^{-1} (235 wavenumbers; a spectral resolution of 8 cm^{-1} gives a data spacing of ≈4 cm^{-1} after a 2× zero-filling of the interferogram), rubberband baseline-corrected and normalized to the Amide I peak (*i.e.*, ≈1,650 cm^{-1}).

For PCA-LDA, SPA-LDA and GA-LDA model, the samples were divided into training (70%), validation (15%) and prediction sets (15%) by applying the classic Kennard-Stone (KS) uniform sampling algorithm to the IR spectra[32]. Sample numbers in each set are presented in Table 1. Training samples were used in the modelling procedure (including variable selection for LDA), whereas the prediction set was only used in the final evaluation of the classification. The optimum number of variables for SPA-LDA and GA-LDA was determined from the minimum cost function G calculated for a given validation dataset:

$$G = \frac{1}{N_V}\sum_{n=1}^{N_V} g_{n,} \quad (1)$$

where g_n is defined as

$$g_n = \frac{r^2(x_n, m_{I(n)})}{\min_{I(m)\neq I(n)} r^2(x_n, m_{I(m)})} \quad (2)$$

and $I(n)$ is the index of the true class for the nth validation object x_n. g_n is defined as risk of misclassification of the nth validation object x_n, n=1, ..., N_V). In this definition, the numerator is the squared Mahalanobis distance between object x_n (of class index I_n) and the sample mean $m_{I(n)}$ of its true class. The denominator in Eq. (2) corresponds to the squared Mahalanobis distance between object x_n and the centre of the closest incorrect class.

The GA routine was carried out during 100 generations with 200 chromosomes each. Crossover and mutation probabilities were set to 60% and 10%, respectively. Moreover, the algorithm was repeated three times, starting from different random initial populations. The best solution (in terms of the fitness value) resulting from three realizations of the GA was employed. For this study, LDA scores, loadings and discriminant function (DF) values were obtained for the specimens. Usually, the first LDA factor (LD1) was used to visualize the alterations in the sample in 1-dimensional (D) scores plots that indicate the main biochemical alterations. SPA-LDA and GA-LDA were used to detect alterations relative to HPV infection in LBC samples based of age of participants.

Receiver-operating characteristic (ROC) analysis is recommended standard practice for test evaluation studies for non-binary tests[28]. For this study, measures of test accuracy, such as sensitivity (probability that a test result will be positive when the disease is present), specificity (probability that a test result will be negative when the disease is not present) were calculated as important quality standards in test evaluation. Both have a maximum value of 1 and a minimum of 0. Sensitivity and specificity can be calculated using the following the equations:

$$\text{Sensitivity (\%)} = \frac{TP}{TP + FN} x100$$

$$\text{Specificity (\%)} = \frac{TN}{TN + FP} x100$$

where FN is defined as a false negative and FP as a false positive. TP is defined as true positive and TN is defined as true negative.

Results

Dataset A: lrHPV *vs.* hrHPV

Figure 1*A* shows mean IR spectra obtained from all grades segregated into lrHPV *vs.* hrHPV. As can be seen, discriminating the two categories on the basis of ATR-FTIR spectral measurements is not straightforward, owing to the complexity of the dataset. Thus, pattern classification (PCA-LDA) or variable selection techniques (SPA-LDA and GA-LDA) were applied to the dataset and comparisons made between classification rates (Table 2) and interpretability. Figure 1*B* is a 2-D PCA-LDA scores plot of the derived spectral points from each category, and shows that there is 'crossover' between the two categories; this hints at minimal segregation. However, as can be seen in Table 2, the PCA-LDA models for lrHPV generated a sensitivity and specificity of 48% and 61%, respectively, using six PC scores from PCA, which account for >90% of the variance for both categories. For hrHPV, the PCA-LDA model achieved a sensitivity and specificity of 76% and 77%, respectively. Then, SPA-LDA was applied to the dataset to obtain the optimum number of variables by the minimum cost function G. Using only five selected wavenumbers (Table 3), Fisher scores were obtained and this improved segregation between classes (Figure 1*C*) when compared with PCA-LDA. The SPA-LDA model achieved a sensitivity and specificity of 50% and 50%, respectively, for lrHPV. For hrHPV, SPA-LDA, using the five wavenumbers selected, achieved a sensitivity and specificity of 76% and 76%, respectively. The GA-LDA model for comparison achieved an improvement in segregation between lrHPV *vs.* hrHPV (Figure 1*D*). The GA resulted in the selection of 28 wavenumbers (of 235 available) (Table 3).

Dataset B: ≤29 y and >30 y for HPV types

Figure 2*A* shows mean IR spectra obtained from ≤29 y and >30 y for HPV types. A PCA-LDA model was built using six PCs, together explaining 90.5% of variance in the data. In Fig. 2*B* one can see that the PC scores plot does not show category separation. The PCA-LDA model for ≤29 y obtained a sensitivity and specificity of 58% and 56%, respectively (Table 2). For >30 y, the PCA-LDA model achieved a sensitivity and specificity of 48% and 48%, respectively. Figure 2*C* is a scores plot that shows SPA-LDA generates some segregation between the two categories, ≤29 y and >30 y, for HPV; the cost function minimum point was obtained with four wavenumbers (Table 3). By using these selected wavenumbers, SPA-LDA yielded a sensitivity and specificity of 60% and 60%, respectively, for ≤29 y; for >30 y, a sensitivity and specificity of 63% and 60% were obtained, respectively. For GA-LDA (Table 2), the accuracy showed an improvement in comparison with PCA-LDA and SPA-LDA results, especially for >30 y category, using 20 selected wavenumbers (Table 3), with sensitivity and specificity of 70% and 67%, respectively. Finally, Figure 2*D* is a scores plot that shows GA-LDA (cost function minimum point obtained with 20 wavenumbers) generates better segregation for the two categories, ≤29 y *vs*. >30 y for HPV.

Dataset C: ≤29 y and >30 y based on normal cervical cytology (NCC)

Figure 3*A* shows mean IR spectra from categories divided into ≤29 y and >30 y from NCC. As before, pattern classification (PCA-LDA) and variable selection techniques (SPA-LDA and GA-LDA) were applied to this condition and comparisons were made between classification rates (Table 2) and interpretability. Figure 3*B* shows that there is a 'crossover' between ≤29 y and >30 y from NCC using the PCA-

LDA model. As can be seen in Table 2, the PCA-LDA model for ≤29 y produced a sensitivity and specificity of 48% and 47%, respectively, using seven PC scores from PCA, which accounts for >93% of the variance for both categories. For >30 y, the PCA-LDA model exhibited an improved sensitivity and specificity of 63% and 62%, respectively. The optimum number of variables for the SPA-LDA model was determined from the minimum cost function G, resulting in five wavenumbers (Table 3). Accuracy of SPA-LDA for ≤29 y was 40% and 45% for sensitivity and specificity, respectively. However, for >30 y, a sensitivity and specificity by the SPA-LDA model of 64% and 65%, respectively, was achieved. Performing LDA on the GA selected variable ≤29 y dataset, the accuracy of the model was 53% and 81% for sensitivity and specificity, respectively. The accuracy of GA-LDA for >30 y was 78% and 77% for sensitivity and specificity, respectively. The GA employed for comparison resulted in the selection of 23 wavenumbers (Table 3). Figure 3*D* shows the scores plot associated with GA-LDA variable selection, whose cost function minimum point was obtained with 20 wavenumbers, highlighting improvement over previous models.

Dataset D: ≤29 y and >30 y based on low-grade cervical cytology (LG-CC)

Figure 4*A* shows mean IR spectra following categorisation into ≤29 y and >30 y from LG-CC. Figure 4*B* details the graphical representation of Fisher scores obtained from the PCA-LDA model, using six PCs with a cumulative variance of 91%, allowing one to observe a separation of the categories albeit with some overlap. In Table 2, the PCA-LDA models for ≤29 y associated LG-CC generated a sensitivity and specificity of 53% and 58%, respectively. For >30 y from LG-CC, the PCA-LDA model achieved a sensitivity and specificity of 38% and 37%, respectively. SPA-LDA was subsequently employed to analyse the differences between two categories (≤29 y

vs. >30 y based on LG cervical cytology). Figure 4*C* is a scores plot that shows SPA-LDA results in slight segregation between the two categories, whose cost function minimum point was obtained with two wavenumbers (Table 3). By using these selected wavenumbers, SPA-LDA showed a sensitivity and specificity of 56% and 52%, respectively, for ≤29 y. For >30 y, a sensitivity and specificity of 57% and 48%, respectively, were obtained. GA was applied to the dataset and resulted in the selection of 33 variables (Table 3). Figure 4*D* is a scores plot that shows GA-LDA improved segregation between the two categories, ≤29 y and >30 y for LG-CC. Furthermore, the accuracy of GA-LDA for ≤29 y was 88% and 83% for sensitivity and specificity, respectively. On the other hand, the accuracy of GA-LDA for >30 y was 68% and 73% for sensitivity and specificity, respectively.

Dataset E: Segregate all spectra into categories HPV16/18 *vs.* HPV31/35 *vs.* HPV Others

Figure 5*A* shows mean IR spectra from the dataset split into three categories (HPV16/18 *vs.* HPV31/35 *vs.* HPV Others). Table 2 shows the accuracy tests achieved for PCA-LDA, SPA-LDA and GA-LDA models for the three categories (HPV16/18 *vs.* HPV31/35 *vs.* HPV Others). Figure 5*B* is the graphical representation of Fisher scores (DF1 × DF2) obtained by PCA-LDA from each category, using six PCs with a cumulative variance of 90%; DF1 × DF2 does not discriminate between HPV samples. As can be seen in Table 2, sensitivity and specificity of 55% and 53%, respectively, were achieved by PCA-LDA models for HPV16/18. For HPV 31/35, the sensitivity and specificity obtained were 61% and 58%, respectively. Furthermore, for HPV Others, the sensitivity and specificity obtained were 57% and 54%, respectively. SPA was applied to the dataset and resulted in the selection of four variables (Table

3). Using the four wavenumbers selected by SPA-LDA, DF1 × DF2 was obtained for all the samples in the dataset (Figure 5*C*). As can be seen, there is a positive effect of homogeneity among categories, using only the four wavenumbers selected by SPA in the LDA modelling. For HPV16/18 (Table 2), the sensitivity and specificity obtained were 64% and 58%, respectively. For HPV31/35, the sensitivity and specificity obtained were 66% and 62%, respectively. For HPV Others, the sensitivity and specificity obtained were 54% and 52%, respectively. Finally, Fig. 5D shows the scores plot associated with variable selection using GA-LDA, whose cost function minimum point was obtained with 33 wavenumbers (Table 3). There is an even larger effect of homogeneity between categories, using these 33 wavenumbers selected by GA in the LDA modelling. The accuracy of GA-LDA for the three categories (HPV16/18 *vs.* HPV31/35 *vs.* HPV Others) achieved positive values. For HPV16/18, the sensitivity and specificity obtained were 85% and 66%, respectively. For HPV31/35, the sensitivity and specificity obtained were 77% and 71%, respectively. For HPV Others, the sensitivity and specificity obtained were 56% and 55%, respectively.

Discussion

The objective of cervical cancer screening is to reduce incidence and mortality by detecting and treating precancerous lesions. Development of methods for preparing cytology specimens as well as many other screening techniques suggests that current practices may be modified in the future. The implementation of new approaches such as LBC and/or spectroscopy (IR or Raman) may permit more conservative management of women with self-limited lesions related to HPV exposure, improve detection of serious cancer precursors, and provide more cost-effective screening.

Adjunctive diagnostic procedures for the detection of HPV infection could increase the sensitivity of primary and secondary screening of cervical cancer. HPV testing could improve the specificity of screening programmes resulting in avoidance of overtreatment and saving costs for confirmatory procedures. When ATR-FTIR spectroscopy was employed to predict lrHPV and hrHPV, it was observed that using GA-LDA-associated variables (28 selected) gives better segregation than PCA-LDA and SPA-LDA together. The GA-LDA model increases the sensitivity (87%) and specificity (92%) of screening for lrHPV and hrHPV lesions. Examination of the selected wavenumbers following GA-LDA showed that the main biochemical alterations discriminating lrHPV *vs.* hrHPV were lipids, proteins, nucleic acids, carbohydrates and, to a lesser extent, DNA vibrations. Several selected wavenumbers appear to be of particular interest, namely, the variables at 1755 and 1720 cm^{-1}, associated with C=O stretching vibrations of aldehydes and lipids, respectively. These variables (1755 and 1720 cm^{-1}) appear associated with transition from normal to LSIL to HSIL and result in alterations mainly in intracellular and/or membrane proteins/lipids. Even though they are not always markedly altered, they appear consistently as distinct segregating wavenumbers. The wavenumbers between 900 and 1000 cm^{-1} represent the spectral region of DNA/RNA vibrations. Oncogenic virus particles or commitment to transformation would be expected to alter DNA/RNA as would be found in this spectral region (Figs. 2 and 5).

The natural history of HPV suggests that there is little risk of a significant precancerous lesion going undetected within the first 3-5 years from the onset of sexual activity[33]. Annual screening is recommended also by the American College of Obstetricians and Gynecologists (ACOG), although in women aged ≥30 y with negative Pap tests, screening may be conducted every 2-3 y. Herein, ATR-FTIR

spectral data was discriminated into three case studies for HPV infection (all risks, NCC and LG-CC) into ≤29 y and >30 y. Age was employed as a categorisation factor. GA-LDA was employed on all ATR-FTIR spectra (all risks, NCC and LG-CC) into ≤29 y and >30 y, it was observed that this approach results in better segregation than PCA-LDA and SPA-LDA. Several selected wavenumbers represent the spectral region of lipids, proteins, fatty acid, corresponding to the fingerprint region[34].

A variety of ancillary tests useful in the diagnosis of HPV infection are currently at the clinician's disposal. Use of laboratory-based tests is gaining popularity as an adjunctive measure, particularly in combination with Pap smears, for the detection of CIN or carcinoma. When ATR-FTIR spectroscopy was investigated within three HPV infection types (16/18, 31/35 and HPV Others), the alternative approach would be compared. Sensitivity and specificity for HPV16/18, using 33 selected wavenumbers by GA-LDA, of 85% and 66%, respectively, were achieved.

However, with the introduction of cervical cancer screening programmes, incidence and mortality has been drastically reduced. Techniques such as the traditional Pap test with/without LBC allows for the early detection of cervical abnormalities prior to the development of invasive cancer. HPV DNA testing has also been proposed as a routine screening method for the general population. Screening limitations, such as adherence, test sensitivity and specificity, access, and cost-effectiveness are reflected in current screening guidelines[35]. The metabolic fingerprint generated by ATR-FTIR spectroscopy combining with variable selection methods (SPA-LDA and GA-LDA) is a powerful adjunct for cervical screening programmes, emerging as an alternative for rapid and cost-effective identification of specimens.

Acknowledgments

We greatly acknowledge support from Lancashire Teaching Hospitals NHS Foundation Trust, UK and University Hospital of Ioannina, Aristotle University of Thessaloniki - Hippokration Hospital and Attikon Hospital, University of Athens, Greece. Kássio M.G. Lima thanks the CNPq (The National Council for Scientific and Technological Development, Brazil) for his Postdoctoral Fellowship (Ref. 246742/2012-7).

References

1. F. B. Lillo, *New Microbiol.*, 2005, **28**, 111-118.

2. J. Doorbar, *Clin. Sci. (Lond.)*, 2006, **110**, 525-41.

3. F. X. Bosch, A. N. Burchell, M. Schiffman, A. R. Giuliano, S. de Sanjose, L. Bruni, G. Tortolero-Luna, S. K. Kjaer, and N. Muñoz, *Vaccine*, 2008, **26**, K1-K16.

4. S. de Sanjose, W. G. Quint, L. Alemany, D. T. Geraets, J. E. Klaustermeier, B. Lloveras, S. Tous, A. Felix, L. E. Bravo, H.-R. Shin, C. S. Vallejos, P. A. de Ruiz, M. A. Lima, N. Guimera, O. Clavero, M. Alejo, A. Llombart-Bosch, C. Cheng-Yang, S. A. Tatti, E. Kasamatsu, E. Iljazovic, M. Odida, R. Prado, M. Seoud, M. Grce, A. Usubutun, A. Jain, G. A. H. Suarez, L. E. Lombardi, A. Banjo, C. Menéndez, E. J. Domingo, J. Velasco, A. Nessa, S. C. B. Chichareon, Y. L. Qiao, E. Lerma, S. M. Garland, T. Sasagawa, A. Ferrera, D. Hammouda, L. Mariani, A. Pelayo, I. Steiner, E. Oliva, C. J. Meijer, W. F. Al-Jassar, E. Cruz, T. C. Wright, A. Puras, C. L. Llave, M. Tzardi, T. Agorastos, V. Garcia-Barriola, C. Clavel, J. Ordi, M. Andújar, X. Castellsagué, G. I. Sánchez, A. M. Nowakowski, J. Bornstein, N. Muñoz, and F. X. Bosch, *Lancet Oncol.*, 2010, **11**, 1048-56.

5. N. Li, S. Franceschi, R. Howell-Jones, P. J. F. Snijders, and G. M. Clifford, *Int. J. Cancer*, 2011, **128**, 927-35.

6. N. Muñoz, F. X. Bosch, X. Castellsagué, M. Díaz, S. de Sanjose, D. Hammouda, K. V Shah, and C. J. L. M. Meijer, *Int. J. Cancer*, 2004, **111**, 278-85.

7. D. M. Parkin and F. Bray, *Vaccine*, 2006, **24 Suppl 3**, S3/11-25.

8. A. A. Kevin, *Lancet*, 2007, **369**, 1861-1868.

9. C. Swygart, *Br. J. Biomed. Sci.*, 1997, **54**, 299-303.

10. M. Schiff, P. E. Castle, J. Jeronimo, A. C. Rodriguez, and S. Wacholder, *Lancet*, 2007, **370**, 890-907.

11. M. Schiffman, R. Herrero, A. Hildesheim, M. E. Sherman, M. Bratti, S. Wacholder, M. Hutchinson, J. Morales, and M. D. Greenberg, *J. Am. Med. Assoc.*, 2000, **283**, 87-93.

12. A. Herbert, *Cytopathology*, 2002, **13**, 379-384.

13. K. Gajjar, A. A. Ahmadzai, G. Valasoulis, J. Trevisan, C. Founta, M. Nasioutziki, A. Loufopoulos, M. Kyrgiou, S. M. Stasinou, P. Karakitsos, E. Paraskevaidis, B. Da Gama-Rose, P. L. Martin-Hirsch, and F. L. Martin, *PLoS One*, 2014, **9**, e82416.

14. M. F. Mitchell, S. B. Cantor, N. Ramanujam, G. Tortolero-Luna, and R. Richards-Kortum, *Obstet. Gynecol.*, 1999, **93**, 462-470.

15. P. F. Escobar, J. L. Belinson, A. White, N. M. Shakhova, F. I. Feldchtein, M. V Kareta, and N. D. Gladkova, *Int. J. Gynecol. Cancer*, 2004, **14**, 470-474.

16. K. Carlson, I. Pavlova, T. Collier, M. Descour, M. Follen, and R. Richards-Kortum, *Gynecol. Oncol.*, 2005, **99**, S84-8.

17. N. C. Purandare, I. I. Patel, K. M. G. Lima, J. Trevisan, M. Ma'Ayeh, A. McHugh, G. Von Bünau, P. L. Martin Hirsch, W. J. Prendiville, and F. L. Martin, *Anal. Methods*, 2014, **6**, 4576-4584.

18. N. C. Purandare, I. I. Patel, J. Trevisan, N. Bolger, R. Kelehan, G. von Bünau, P. L. Martin-Hirsch, W. J. Prendiville, and F. L. Martin, *Analyst*, 2013, **138**, 3909-16.

19. J. G. Kelly, J. Trevisan, A. D. Scott, P. L. Carmichael, H. M. Pollock, P. L. Martin-Hirsch, and F. L. Martin, *J. Proteome Res.*, 2011, **10**, 1437-1448.

20. F. L. Martin, J. G. Kelly, V. Llabjani, P. L. Martin-Hirsch, I. I. Patel, J. Trevisan, N. J. Fullwood, and M. J. Walsh, *Nat. Protoc.*, 2010, **5**, 1748-1760.

21. J. G. Kelly, M. N. Singh, H. F. Stringfellow, M. J. Walsh, J. M. Nicholson, F. Bahrami, K. M. Ashton, M. A. Pitt, P. L. Martin-Hirsch, and F. L. Martin, *Cancer Lett.*, 2009, **274**, 208-217.

22. S. Neviliappan, L. Fang Kan, T. Tiang Lee Walter, S. Arulkumaran, and P. T. T. Wong, *Gynecol. Oncol.*, 2002, **85**, 170-4.

23. I. I. Patel, J. Trevisan, P. B. Singh, C. M. Nicholson, R. K. G. Krishnan, S. S. Matanhelia, and F. L. Martin, *Anal. Bioanal. Chem.*, 2011, **401**, 969-82.

24. M. J. Walsh, A. Hammiche, T. G. Fellous, J. M. Nicholson, M. Cotte, J. Susini, N. J. Fullwood, P. L. Martin-Hirsch, M. R. Alison, and F. L. Martin, *Stem Cell Res.*, 2009, **3**, 15-27.

25. M. J. Walsh, M. N. Singh, H. F. Stringfellow, H. M. Pollock, A. Hammiche, O. Grude, N. J. Fullwood, M. a Pitt, P. L. Martin-Hirsch, and F. L. Martin, *Biomark. Insights*, 2008, **3**, 179-189.

26. M. Cavagna, R. Dell'Anna, F. Monti, F. Rossi, and S. Torriani, *J. Agric. Food Chem.*, 2010, **58**, 39–45.

27. L. Di Giambattista, D. Pozzi, P. Grimaldi, S. Gaudenzi, S. Morrone, and A. C. Castellano, *Anal. Bioanal. Chem.*, 2011, **399**, 2771-8.

28. K. T. Cheung, J. Trevisan, J. G. Kelly, K. M. Ashton, H. F. Stringfellow, S. E. Taylor, M. N. Singh, P. L. Martin-Hirsch, and F. L. Martin, *Analyst*, 2011, **136**, 2047-2055.

29. J. S. Oliveira, T. C. Baia, R. A. Gama, and K. M. G. Lima, *Microchem. J.*, 2014, **115**, 39-46.

30. M. J. C. Pontes, R. K. H. Galvão, M. C. U. Araújo, P. N. T. Moreira, O. D. P. Neto, G. E. José, and T. C. B. Saldanha, *Chemom. Intell. Lab. Syst.*, 2005, **78**, 11-18.

31. J. G. Kelly, K. T. Cheung, C. Martin, J. J. O'Leary, W. Prendiville, P. L. Martin-Hirsch, and F. L. Martin, *Clin. Chim. Acta.*, 2010, **411**, 1027-33.

32. R. Kennard and L. Stone, *Technometric*, 1969, **11**, 137-148.

33. A. Moscicki, *Curr. Womens Heal. Rep.*, 2003, **3**, 433-437.

34. M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner and F. L. Martin, *Nat. Protoc.*, 2014, **9,** 1771-1791.

35. N. C. Purandare, J. Trevisan, I. I. Patel, K. Gajjar, A. L. Mitchell, G. Theophilou, G. Valasoulis, M. Martin, G. von Bünau, M. Kyrgiou, /E. Paraskevaidis, P. L. Martin-Hirsch, W. J. Prendiville, and F. L. Martin, *Bioanalysis*, 2013, **5,** 2697-2711.

Legends to Figures

Figure 1: Comparison of lrHPV and hrHPV cervical cytology specimens. The panel shows mean IR spectra (for standard deviation of entire spectral categories, see ESI Figs. S1A and S1B) obtained from all grades segregated into lrHPV *vs.* hrHPV (***A***). The spectra from patients with lrHPV and hrHPV are shown in blue and red, respectively. The application of principal component analysis (PCA) - linear discriminant analysis (LDA) or variable selection techniques [successive projection algorithm (SPA) and genetic algorithm (GA)] to the segregation of retrospectively categorised lrHPV and hrHPV specimens. PCA-LDA results: (***B***) DF1 × samples calculated by PCA-LDA model from lrHPV (blue) *vs.* hrHPV (red). SPA-LDA results: (***C***) DF1 × samples calculated using the 5 selected wavenumbers by SPA-LDA model from lrHPV (blue) *vs.* hrHPV (red). GA-LDA results: (***D***) DF1 × samples calculated using the 28 selected wavenumbers by GA-LDA model from lrHPV (blue) *vs.* hrHPV (red).

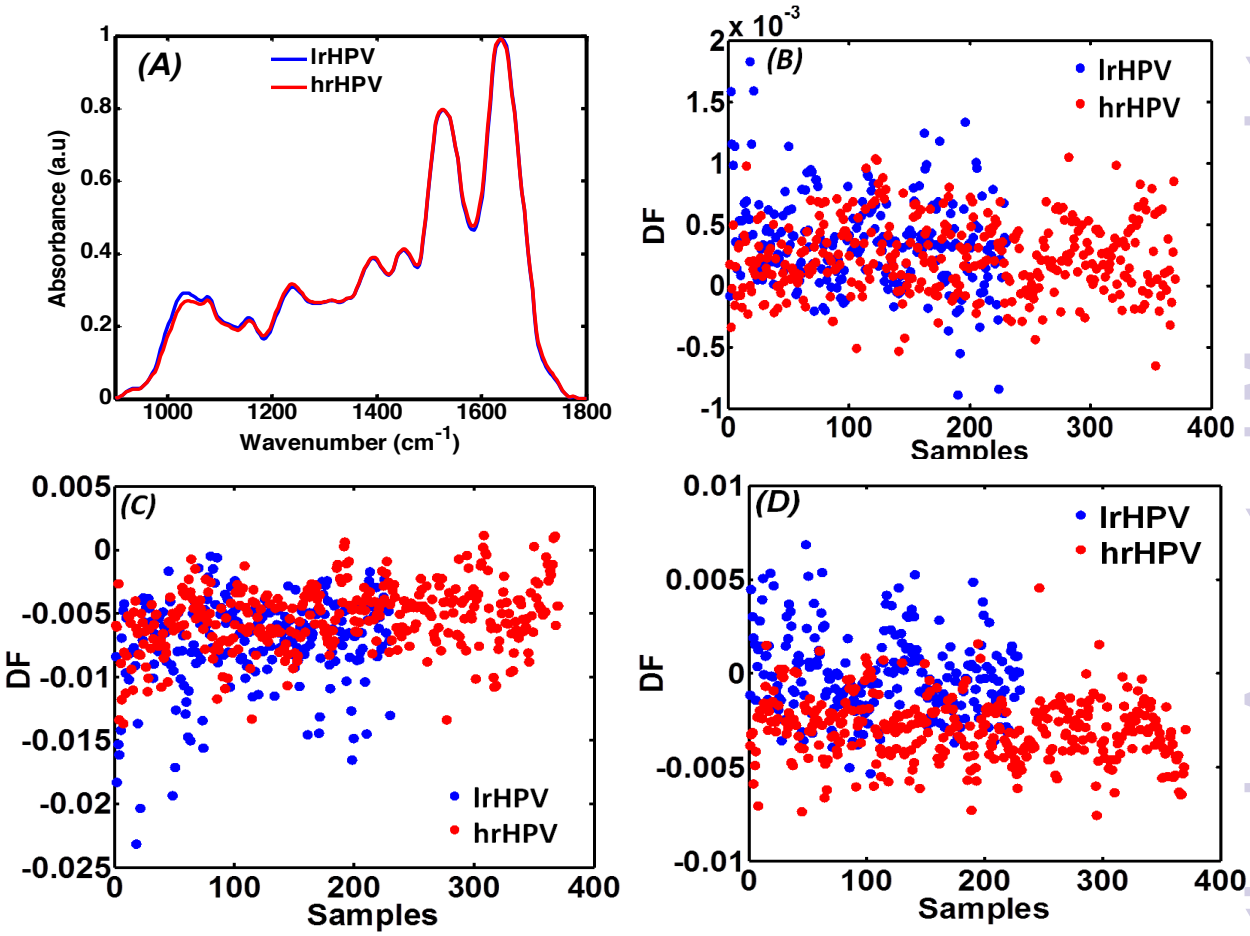
Figure 2: Comparison of ≤29 y and >30 y for HPV types. The panel shows mean IR spectra (for standard deviation of entire spectral categories, see ESI Figs. S2A and S2B) obtained from all grades segregated into ≤29 y and >30 y (***A***). The spectra from patients with ≤29 y and >30 y are shown in blue and red, respectively. The application of principal component analysis (PCA) - linear discriminant analysis (LDA) or variable selection techniques [successive projection algorithm (SPA) and genetic algorithm (GA)] to the segregation of retrospectively categorised ≤29 y and >30 y specimens. PCA-LDA results: (***B***) DF1 × samples calculated by PCA-LDA model from ≤29 y (blue) *vs.*>30 y (red). SPA-LDA results: (***C***) DF1 × samples calculated using the 5 selected wavenumbers by SPA-LDA model from ≤29 y (blue) *vs.* >30 y (red). GA-LDA results: (***D***) DF1 × samples calculated using the 28 selected wavenumbers by GA-LDA model from ≤29 y (blue) *vs.*>30 y (red).

Figure 3: Comparison of ≤29 y and >30 y based on normal cervical cytology (NCC). The panel shows mean IR spectra (for standard deviation of entire spectral categories, see ESI Figs. S3A and S3B) obtained from all grades segregated into ≤29 y and >30 y NCC (***A***). The spectra from patients with ≤29 y and >30 y NCC are shown in blue and red, respectively. The application of principal component analysis (PCA) - linear discriminant analysis (LDA) or variable selection techniques [successive projection algorithm (SPA) and genetic algorithm (GA)] to the segregation of retrospectively categorised ≤29 y and >30 y NCC specimens. PCA-LDA results: (***B***) DF1 × samples calculated by PCA-LDA model from ≤29 y (blue) *vs.*>30 y (red) NCC. SPA-LDA results: (***C***) DF1 × samples calculated using the 5 selected wavenumbers by SPA-LDA model from ≤29 y (blue) *vs.*>30 y (red) NCC. GA-LDA results: (***D***) DF1 × samples calculated using the 28 selected wavenumbers by GA-LDA model from ≤29 y (blue) *vs.* >30 y (red) NCC.

Figure 4: Comparison of ≤29 y and >30 y based on low-grade cervical cytology (LG-CC). The panel shows mean IR spectra (for standard deviation of entire spectral categories, see ESI Figs. S4A and S4B) obtained from all grades segregated into ≤29 y and >30 y LG (***A***). The spectra from patients with ≤29 y and >30 y LG-CC are shown in blue and red, respectively. The application of principal component analysis (PCA) - linear discriminant analysis (LDA) or variable selection techniques [successive projection algorithm (SPA) and genetic algorithm (GA)] to the segregation of retrospectively categorised ≤29 y and >30 y LG-CC specimens. PCA-LDA results: (***B***) DF1 × samples calculated by PCA-LDA model from ≤29 y (blue) *vs.* >30 y (red) LG. SPA-LDA results: (***C***) DF1 × samples calculated using the 5 selected wavenumbers by SPA-LDA model from ≤29 y (blue) *vs.*>30 y (red) LG-CC. GA-LDA results: (***D***) DF1 × samples calculated using the 28 selected wavenumbers by GA-LDA model from ≤29 y (blue) *vs.* >30 y (red) LG-CC.

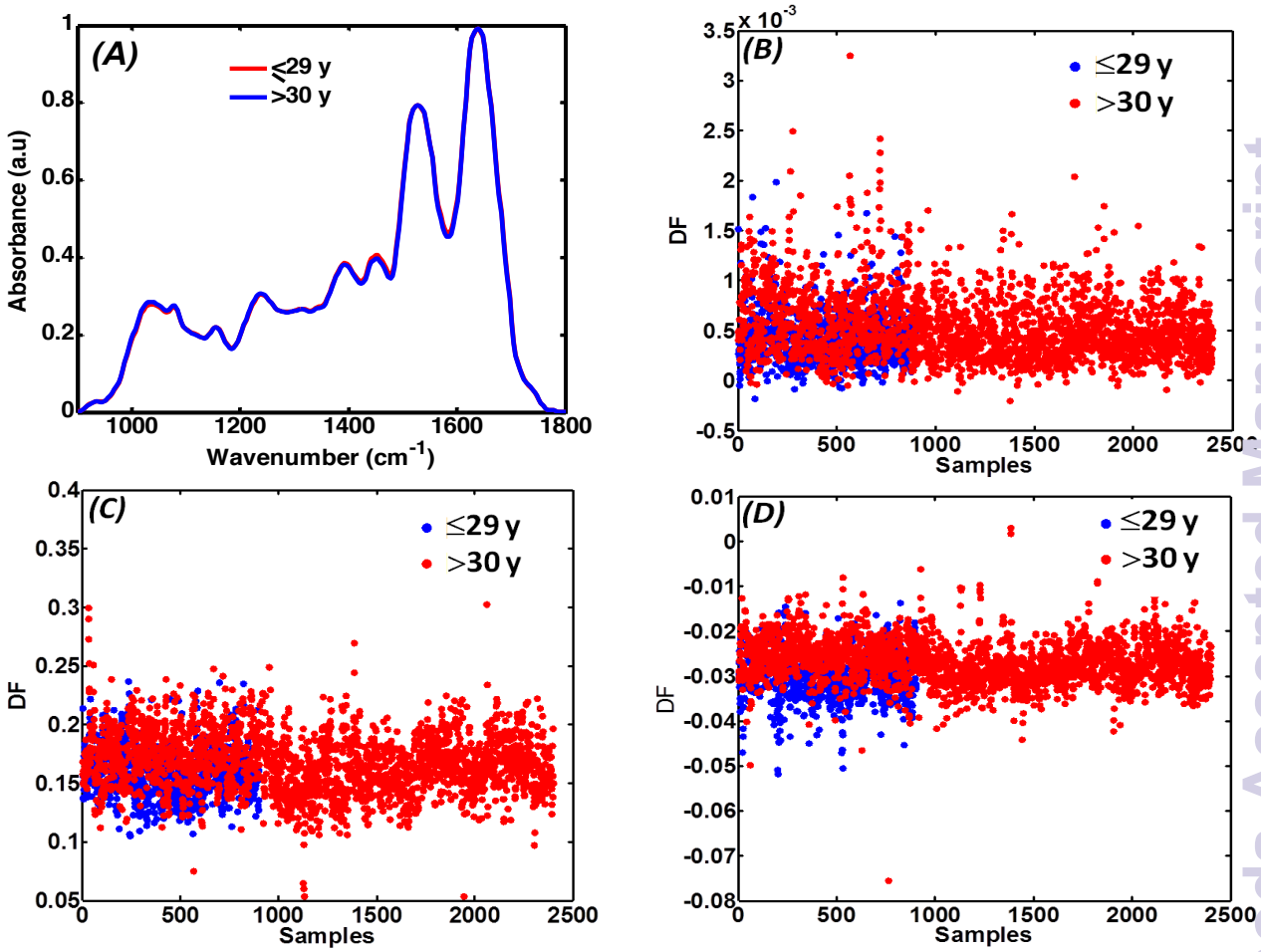
Figure 5: Comparison of HPV16/18 *vs.* HPV31/35 *vs.* HPV Others for HPV types. The panel shows mean IR spectra (for standard deviation of entire spectral categories, see ESI Figs. S5A to S5C) obtained from all HPV types segregated into HPV16/18 *vs.* HPV31/35 *vs.* HPV Others (***A***). The spectra from patients with HPV 16/18, HPV 31/35 and HPV Others are shown in red, black and blue, respectively. The application of principal component analysis (PCA) - linear discriminant analysis (LDA) or variable selection techniques [successive projection algorithm (SPA) and genetic algorithm (GA)] to the segregation of retrospectively categorised HPV16/18 *vs.* HPV31/35 *vs.* HPV Others. PCA-LDA results: (***B***) DF1 × DF2 discriminant function values calculated by PCA-LDA model into three categories: HPV16/18 (red) *vs.* HPV31/35 (black) *vs.* HPV Others (blue). SPA-LDA results (***C***) DF1 × DF2 discriminant function values calculated using the 4 selected wavenumbers by SPA-LDA model from HPV16/18 (red) *vs.* HPV31/35 (black) *vs.* HPV Others (blue) specimens. PCA-LDA results (***D***) DF1 × DF2 discriminant function values calculated using the 33 selected wavenumbers by GA-LDA model from HPV16/18 (red) *vs.* HPV31/35 (black) *vs.* HPV Others (blue) specimens.

Figure 1



601 **Figure 2**

602

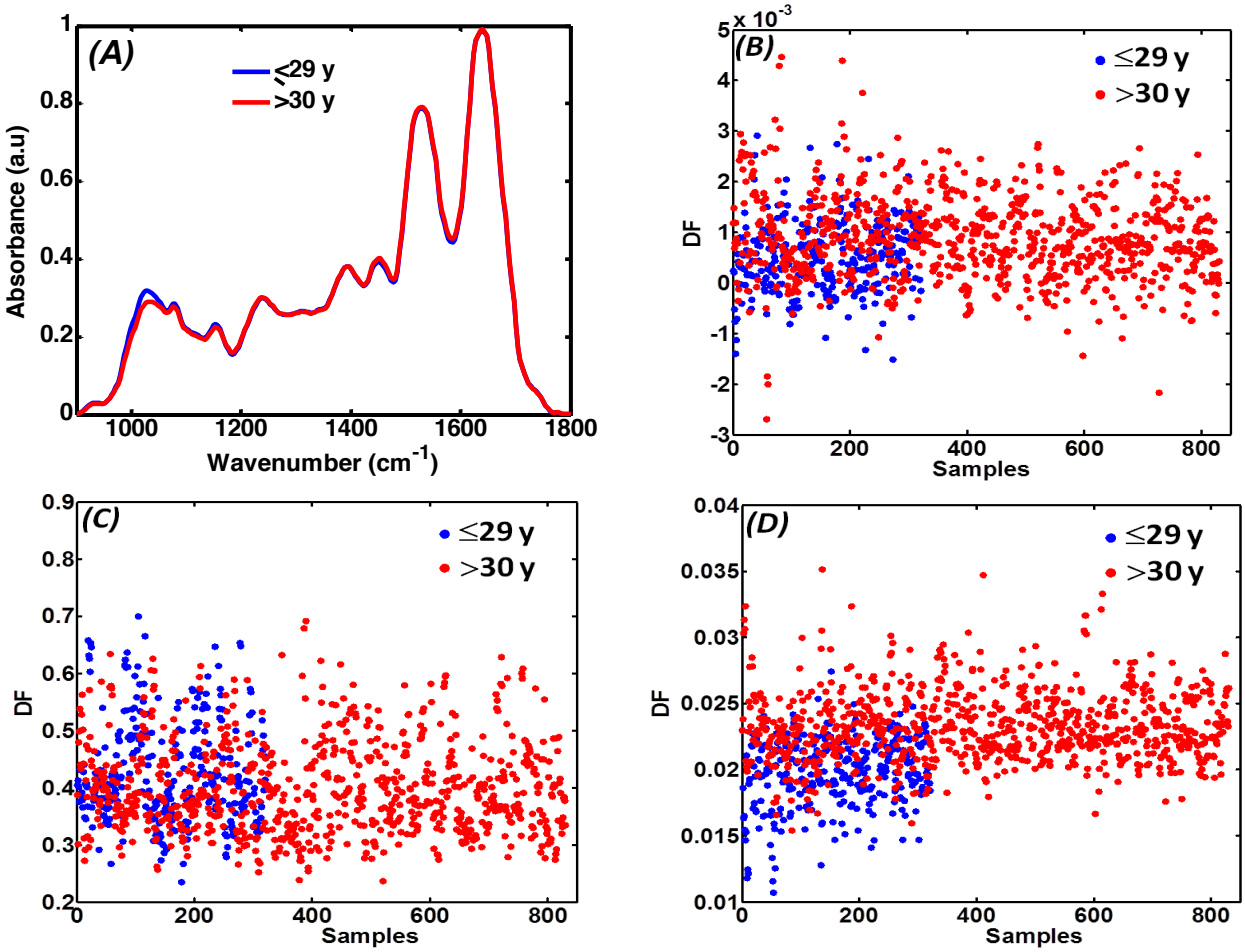


603
604

605

606 **Figure 3**

607

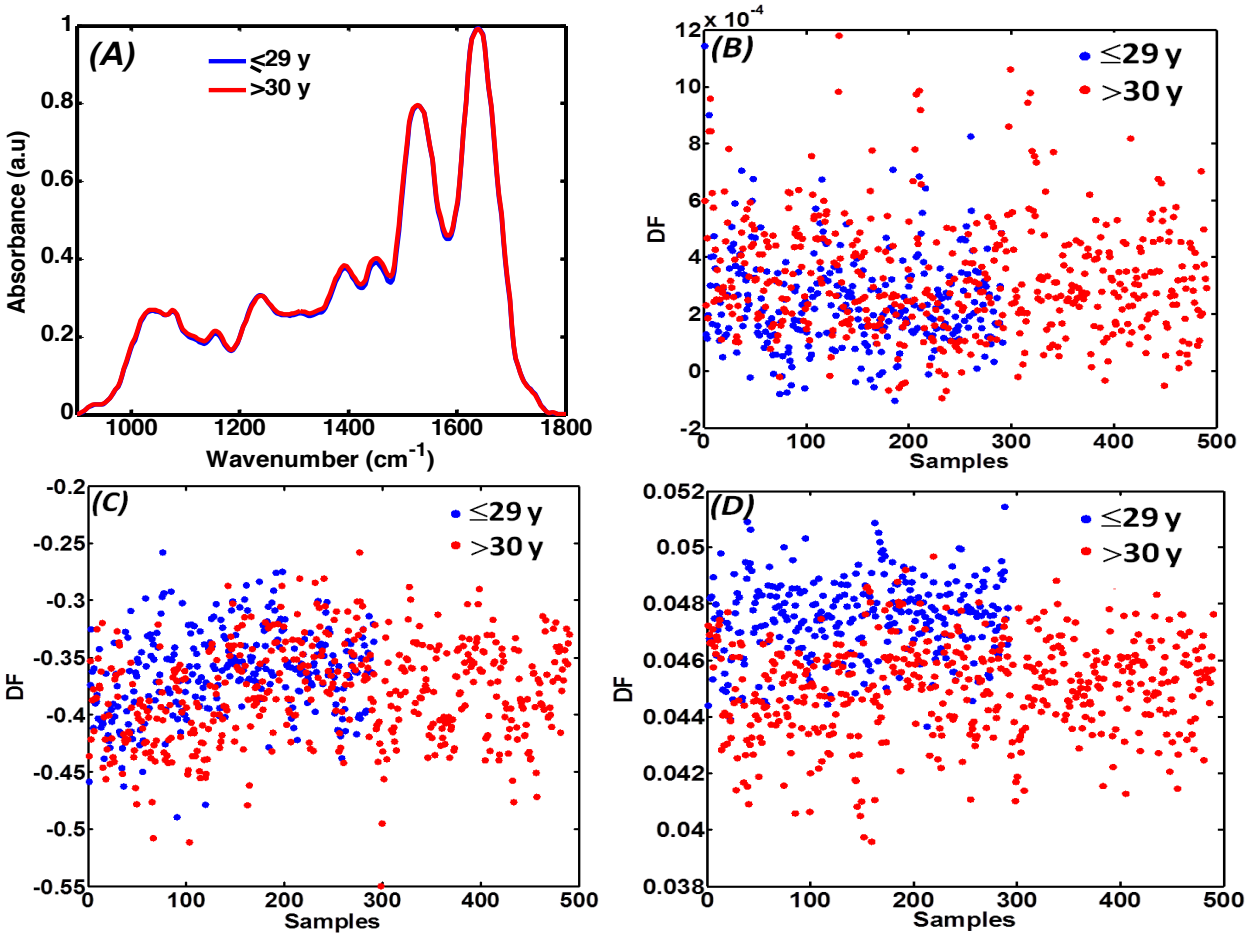


608

609

610 **Figure 4**

611



612
613

Figure 5

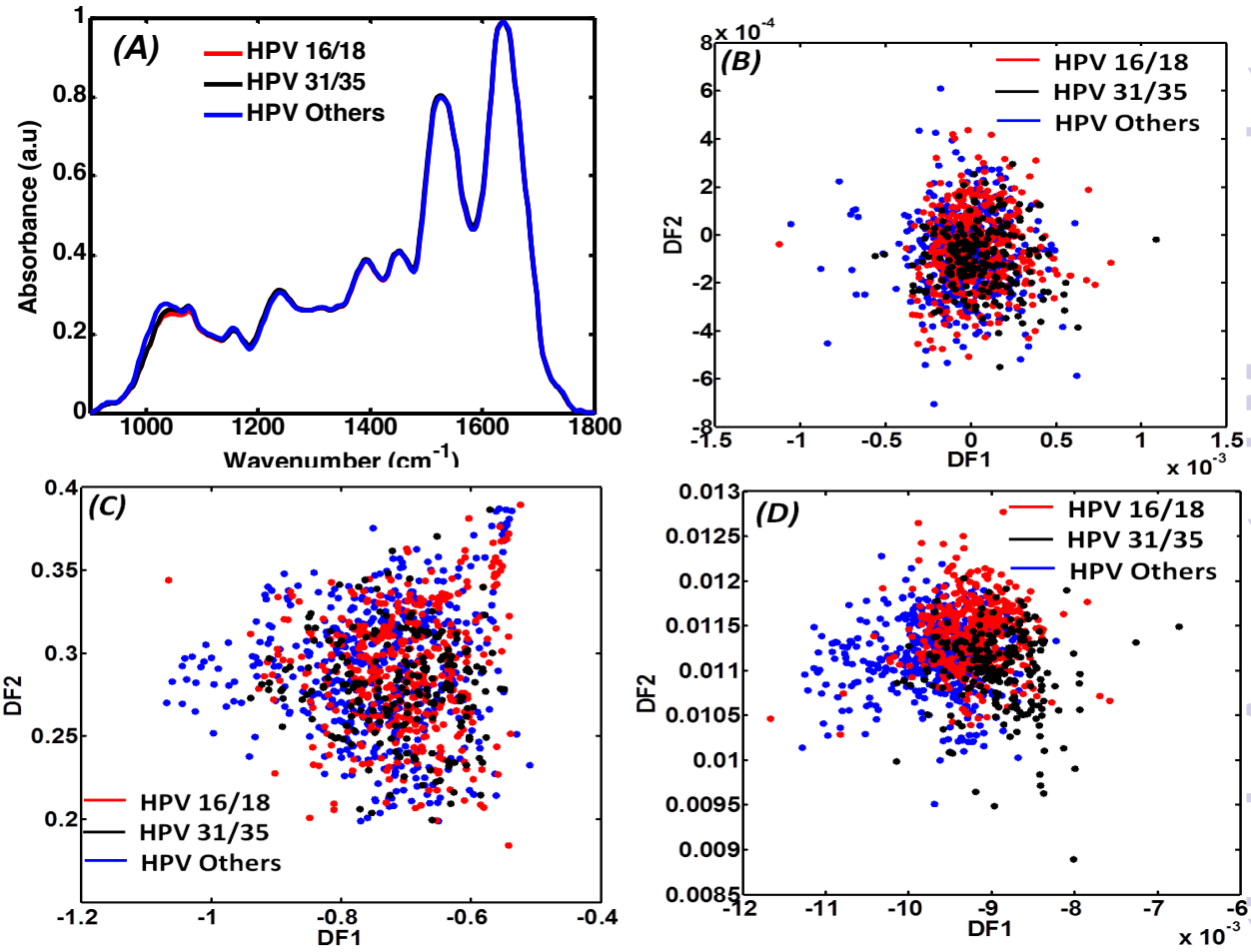


Table 1: Number of training, validation and prediction specimens (or spectra) in each category

Category	Set training	Validation	Prediction
lrHPV	160	35	35
hrHPV	260	55	55
≤29 y HPV	631	135	135
>30 y HPV	1679	360	360
≤29 y NCC	224	48	48
>30 y NCC	579	125	125
≤29 y LG-CC	201	45	45
>30 y LG-CC	340	75	75
HPV16/18	290	65	65
HPV31/35	140	35	35
HPV Others	350	75	75

Table 2: Sensibility (%) and specificity (%) together with multivariate classification methods (PCA-LDA, SPA-LDA or GA-LDA) results for lrHPV *vs.* hrHPV, ≤29 y *vs.*>30 y HPV, ≤29 y *vs.* >30 y NCC, ≤29 y *vs.* >30 y LG-CC and HPV16/18 *vs.* HPV 31/35 *vs.* HPV Others

Models	***lrHPV vs. hrHPV***		**≤29 y *vs.* >30 y HPV**		**≤29 y *vs.* >30 y NCC**	
	Sen	**Spec**	**Sen**	**Spec**	**Sen**	**Spec**
PCA-LDA	48/76	61/77	58/48	56/48	48/63	47/62
SPA-LDA	50/76	50/76	60/63	60/60	40/64	45/65
GA-LDA	54/87	54/92	65/70	60/67	53/78	81/77
Models	**≤29 y *vs.*>30 y LG-CC**		***HPV16/18 vs. HPV 31/35 vs. HPV Others***			
	Sen	**Spec**	**Sen**		**Spec**	
PCA-LDA	53/38	58/37	55/61/57		53/58/54	
SPA-LDA	56/57	52/48	64/66/54		58/62/52	
GA-LDA	88/68	83/73	85/77/56		66/71/55	

Sen = sensitivity (%); Spec = specificity (%); HPV, human papilloma virus; LG-CC, low-grade cervical cytology; NCC, normal cervical cytology; lr, low-risk; hr, high-risk

Table 3: Variables for SPA-LDA and GA-LDA determined from the minimum cost function G calculated for a given validation dataset

Computational algorithm	Minimal cost function - optimum number of variables (cm^{-1})				
	Dataset A	Dataset B	Dataset C	Dataset D	Dataset E
SPA-LDA	1018, 1064, 1504, 1597, 1643	1018, 1064, 1435, 1504	1018, 1064, 1504, 1751	1018, 1751	1018, 1500, 1589, 1620
GA-LDA	914, 921, 925, 945, 948, 979, 999, 1014, 1026, 1030, 1099, 1149, 1161, 1184, 1207, 1215, 1300, 1330, 1454, 1469, 1481, 1489, 1577, 1608, 1681, 1697, 1720, 1755	948, 968, 995, 1014, 1026, 1030, 1037, 1134, 1145, 1188, 1238, 1273, 1315, 1381, 1384, 1415, 1435, 1462, 1589, 1708	918, 925, 937, 945, 1003, 1014, 1018, 1064, 1095, 1222, 1369, 1411, 1431, 1458, 1500, 1512, 1523, 1531, 1558, 1593, 1624, 1708, 1778	910, 925, 933, 972, 1022, 1080, 1114, 1134, 1149, 1161, 1172, 1180, 1184, 1207, 1242, 1280, 1311, 1315, 1342, 1365, 1377, 1423, 1427, 1485, 1500, 1531, 1562, 1612, 1620, 1635, 1647, 1658, 1747	898, 902, 925, 948, 960, 964, 968, 1003, 1022, 1041, 1049, 1091, 1192, 1195, 1219, 1222, 1226, 1257, 1269, 1307, 1369, 1427, 1431, 1450, 1462, 1477, 1550, 1566, 1593, 1597, 1654, 1782, 1797