Journal Name

View Article Online
DOI: 10.1039/C6AN02072K

ARTICLE

Virtual Staining of Colon Cancer Tissue by Label-free Raman Micro-spectroscopy

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

D. Petersen,[a]† L. Mavarani,[a,c]† D. Niedieker,[a] E. Freier,[a,d] A. Tannapfel,[b] C. Kötting,[a] K. Gerwert[a]*, and S. F. El-Mashtoly[a]

The great capability of label-free classification of tissue via vibrational spectroscopy, like Raman or infrared imaging, is shown in numerous publications (review: Diem *et al. J. Biophotonics* 2013, **6**, 855-886). Here, we present a new approach, virtual staining, that improves the Raman spectral histopathology (SHP) images of colorectal cancer tissue by combining the integrated Raman intensity image in the C—H stretching region (2800-3050 cm^{-1}) with the pseudo-colour Raman image. This allows the display of fine structures such as the filamentous composition of muscle tissue. The morphology of the virtually stained images is in agreement with the gold standard in medical diagnosis, the haematoxylin-eosin staining. The virtual stainig image also represents the whole biochemical fingerprint, and several tissue components including carcinoma were identified automatically with high sensitivity and specificity. For fast tissue classifications, a similar approach was applied on coherent anti-Stokes Raman scattering (CARS) spectral data which is faster and therefore potentially more suitable for clinical applications.

Introduction

Colorectal cancer is among the most common cancer diseases diagnosed for humans.[1] More than 1 million individuals worldwide develop colorectal cancer each year.[2] Most colon cancers start as small benign polyps based on an adenoma sequence. The first level of detecting of colorectal carcinoma is usually performed through a visual inspection during colonoscopy. The diagnosis is performed manually by pathologists on a biopsy via histopathological examination using haematoxylin and eosin (H&E) stained thin tissue sections. In order to determine gene defects next generation gene sequencing is performed.[3] For information about the presence of certain cancer associated markers or proteins, immunohistochemical staining (IHC) is the method of choice. If colorectal cancer is diagnosed within a patient, cancer regions and their surrounding areas of the colon are resected generously.

In the last decade, several studies have shown that spectral histopathology (SHP) is capable of classifying different tissue types and especially diseased tissue such as cancer.[4–11] The measured vibrational spectra are integral signals of the proteome, genome, and metabolome. Thus, when vibrational spectra are collected from distinct regions of for example tissue sections, variations in the spectral patterns are detected and can be correlated with the tissue types or carcinoma from which the spectra were collected. For instance, colorectal and lung carcinoma were identified in this regard by infrared (IR) imaging.[12–16]

Several groups showed the application of Raman and coherent anti-Stokes Raman scattering (CARS) imaging on colon tissue.[17–22] In all cases normal and carcinoma tissue were successfully distinguished, but most of these studies lack elaborated automated bioinformatics. We have recently established a workflow that includes Raman microscopy, bioinformatics, histopathology, and IHC (Fig. S1 in Supplementary Information (SI)) to automatically classify different tissue types and cancer regions.[23] The workflow is divided into training and validation stages. In the training stage, Raman spectral imaging of thin sections of colon tissue was performed. Hierarchical cluster analysis (HCA) of the Raman spectroscopic data was performed as an unsupervised segmentation. From this segmentation similar spectra were grouped into clusters producing a pseudo-colour image. The H&E and/or IHC staining were performed on adjacent thin tissue section. Images of these staining were annotated by the pathologist and then used to identify the corresponding Raman spectral "fingerprints" of different tissue types including cancer based on the comparison with pseudo-colour

[a]*Department of Biophysics and Protein Research Unit Europe (PURE), Ruhr University Bochum, ND/04 Nord, 44780 Bochum, Germany. E-mail: gerwert@bph.ruhr-uni-bochum.de; Fax: +49 234 3214238; phone: +49 234 3224461*
[b]*Klinikum Bergmannsheil, Ruhr-University Bochum, 44780 Bochum, Germany*
†These authors have equally contributed to this work.
[c]*Current address:*
Institute for High-Frequency and Communication Technology, University of Wuppertal, 42119 Wuppertal, Germany
[d]*Current address:*
Leibniz Institute for Analytical Science (ISAS), 44227 Dortmund, Germany
Electronic Supplementary Information (ESI) available: Figures S1-S11 display workflow, images of IHC, H&E staining, Raman SHP, and Raman virtual staining as well as the Raman and CARS mean spectra of different tissue components. See DOI: 10.1039/x0xx00000x

View Article Online
DOI: 10.1039/C6AN02073K

images. These spectral "fingerprints" were used as a database to perform a supervised classification through a classifier such as random forest (RF).[24] RF classifiers are accurate and robust against over-fitting. In the validation stage, Raman spectral maps of new thin tissue sections were measured and automatically annotated by the trained RF. By using this workflow, our preliminary results of Raman based RF with 532 nm excitation displayed carcinoma regions and cells such as lymphocytes and erythrocytes in addition to an autofluorescence specific to p53 active areas in the crypt region of the *lamina propria mucosae*.[23]

This means that Raman SHP can resolve small structures like erythrocytes and lymphocytes and visualizes detailed chemical and morphological composition due to the higher spatial resolution of Raman imaging in comparison with IR imaging. This advantage allows us to detect borders and transitions between diseased and healthy tissue in an accurate way, which is of importance in clinical diagnosis.[25] Thereby, not only the carcinoma can be resected precisely, but also healthy tissue around the carcinoma is spared, which can be crucial in some organs, for example brain.[26]

Here, we present a new method for the graphical representation, virtual staining, that adds the morphological information given by the Raman intensity to the RF pseudo-colour images. These label-free images with high spatial resolution enable a direct comparison with H&E stained images, and thus can help the pathologists in their diagnosis, especially for questionable areas. Several tissue classes and carcinoma regions of colorectal carcinoma were identified and represented by highly resolved RF images. Furthermore, we extend our method to a fast tissue classification using CARS imaging of colorectal cancer tissues coupled with second harmonic generation (SHG), which is a perfect combination suitable for clinical applications.

EXPERIMENTAL SECTION

Sample preparation

Collected spontaneous Raman data sets were gathered from formalin-fixed, paraffin-embedded and native tissue sections. They were obtained from the Institute of Pathology of the Bergmann's Heil Hospital in Bochum, Ruhr-University Bochum. The research was approved by institutional review board (IRB) of the Faculty of Medicine, Ruhr-University Bochum, and complies with all applicable laws and institutional guidelines, and the institutional committee have approved the experiments. An informed consent was also obtained from the patients for use of their tissue samples.

The tissue sections were mounted on reflective silver coated microscope slides (low-emissivity slides [Kevley Technologies, Chesterland, OH]) and deparaffinised before measurements. By using formalin-fixed, deparaffinised samples, which were stable over a long period of time, we were able to use the tissue slides for long term measurements and perform several Raman measurements on the same tissue slides. Subsequent H&E staining was performed on the measured tissue sections or adjacent thin tissue section. For CARS measurements, tissue resections were first frozen in liquid nitrogen and cut with a cryotome. Afterwards, the tissue sections were mounted on glass slides (Menzel Glas, Braunschweig, Germany). These slides were first dried under dry air before and during CARS measurements, which were acquired on very short term. The subsequent H&E staining was conducted on the same tissue slide.

Data Acquisition

Raman hyperspectral data sets were acquired using a confocal Raman microscope (Alpha300AR, WITec Inc., Ulm, Germany) coupled to a frequency doubled solid state laser operating at 532 nm (WITec, Nd:YAG, max. 42 mW). A 25 µm diameter single-mode optical fiber was used to couple the laser radiation into the microscope. For all measurements 7s exposure time per pixel was used, utilizing a 100X/NA 0.90 objective (Olympus, Japan). The Raman scattered light was collected with the same objective and directed through a multi-mode optical fiber (50 µm diameter) to a spectrometer equipped with a back-illuminated electron-multiplying charge coupled device (EMCCD) camera (1600 x 200 pixels). Raman data sets were obtained with a pixel size of 0.8-1.0 µm for regions between 80-150 µm x 80-150 µm. The laser intensity was fixed to 1.5 mW at the sample position.

CARS imaging of tissue samples was performed on a commercial setup (TCS SP5 II CARS, Leica Microsystems, Heidelberg, Germany) as described previously.[27] Briefly, two picosecond-pulsed laser beams were collinearly aligned and focused on the sample through a HCX IRAPO L (25x/0.95W, Leica Microsystems) objective. Multispectral CARS and SHG datasets were acquired in a region between 2700 cm^{-1} and 3000 cm^{-1}. The datasets consist of 61 spectral images and the acquisition time for the whole dataset was ~2 ms per pixel, which is much faster than spontaneous Raman imaging by more than 100 times. Areas of roughly 300 µm x 300 µm (1024 x 1024 pixels) were scanned in epi (backward) and forward direction.

Data Analysis

The Raman raw data was processed in Matlab with the Image Processing and Statistics toolboxes (The Mathworks, Inc., Mass., USA) and algorithms developed in-house. Cosmic spikes were removed by an impulse noise filter[28] and the spectra were interpolated to a reference wavenumber scale. Hierarchical cluster analysis[29] (HCA) was performed on vector normalized data in the region between 700–1800 cm^{-1} and 2600–3100 cm^{-1}. Pseudo-colour images generated from the clustering of the spectra were compared to the annotation of a pathologist and IHC staining. The stage of colorectal cancer was not considered in the training step. The Raman spectra for training of a supervised learning algorithm, RF,[24] were extracted from these data sets. The spectra with high autofluorescence (1.1% of the total measured spectra in the present study) were removed by setting a threshold on the signal intensity of the raw data. Since Raman spectra of tissue section have different backgrounds (see Fig. S2 in SI), a fifth order polynomial was fitted to each spectrum to remove the residual spectral baseline for the classification with a RF. Supporting points were selected by applying a sweep

View Article Online
DOI: 10.1039/C6AN02072H

algorithm on the wavelet-denoised spectrum (Daubechies wavelet D4).[30] After this step, spectra were normalized between 700–3100 cm^{-1} and offset corrected. The hyperspectral data was filtered in image space with a Gaussian window (3x3, σ=1). For the following classes / tissue components spectra were gathered for training from five different patients: carcinoma (471), connective tissue (793), muscle (1538), erythrocytes (140), crypts (963), lymphocytes (593), lymph follicle (202), background (1320). The RF classification was set up as a multistep procedure as shown in Fig. 1. In the first step the background was separated from the residual classes. Therefore spectra were additionally Savitzky Golay filtered (3rd order, window size ~50 cm^{-1}). In the second step spectra were classified into the above mentioned classes. In the third step the spectra were normalized in the region between 700– 1800 cm^{-1} and 2600– 3100 cm^{-1} separately. The additional classification was run first on carcinoma and crypt classified classes only classifying the two classes. Afterwards the classification was performed on connective tissue and muscle classified components in the same way. Sensitivity and specificity of the classification were calculated by cross validation of the training data.

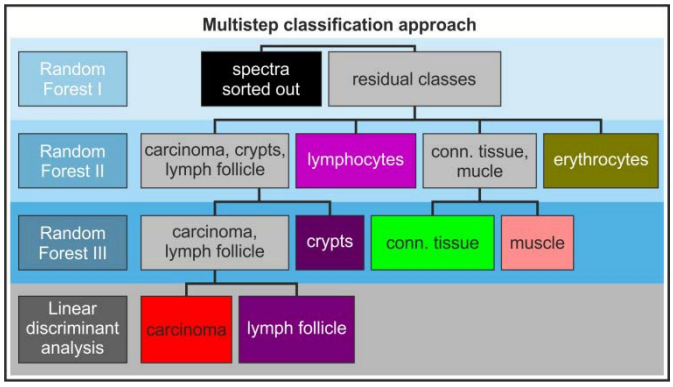

Figure 1. Multistep classification approach used for classification of different tissue components of colorectal carcinoma.

The lymph follicle and carcinoma spectra are hardly discriminable for the RF, especially for low signal to noise ratio (SNR). For this reason the carcinoma and lymph follicle class was selected from the resulting RF. The corresponding spectra were parameterized by spectral curve deconvolution.[31] Features at 1240, 1337, 1390 and 1580 cm^{-1}, were used to distinguish between both components by linear discriminant analysis (LDA). Single cell nuclei have the same spectral fingerprint as carcinoma, though they are strongly limited in their extension. Regions up to 10x10 µm are recognized as cell nuclei.

The concentration of one component is directly proportional to its Raman intensity. Utilizing this information, we created images which reflect the integrated intensity information of the CH-stretching vibration and the pseudo-colour image of the RF. In this study these images will be referred as virtual staining.

Multispectral CARS and SHG datasets were acquired in the 2700-3000 cm^{-1} region. The datasets consist of 61 spectral images. CARS spectral datasets were normalized between 2700 cm^{-1} and 3000 cm^{-1} and *k*-means clustering[29] was applied. Likewise to the virtual staining of the RF the calculated pseudo-colour images of *k*-means were weighted by a combination of the CARS and SHG intensities at 2850 cm^{-1} and 408 nm, respectively.

Immunohistochemical staining

All steps were done on a Bond maX/Bond II System (Leica miocrosystems, Wetzlar, Germany). The slides had first to be deparaffinised then the IHC staining was performed on the next adjacent slice. Afterwards, the slides were pretreated by heat for antigen-retrieval. The staining was performed by incubation of the tissue slides with primary antibodies of p53 and MiB-1 (Ki 67) for 20 minutes. The Ki-67 antigen is a large nuclear protein (345, 395 kDa) expressed during all active phases of the cell. For the Ki-67 staining a monoclonal mouse anti-human antigen (Clone MIB-1) was used (Dako, Hamburg, Germany). Monoclonal mouse anti-human p53 protein (Clone DO-7) was used for detection of wild-type and mutant-type p53 protein for the identification of p53 accumulation in human neoplasia. After staining the tissue sections were washed according to the application details with different solutions of the Bond Refine Red Kit. In a last step the tissue slides were additionally stained with haematoxylin, in order to visualize the cell nucleus and endoplasmic reticulum (see also H&E staining) and fixed in ascending ethanol series and xylene. Images were obtained by using an Olympus microscope.

H&E staining

After data acquisition the tissue slides were stained with H&E.[32] Staining of the cell nucleus and endoplasmic reticulum was achieved by incubation of the tissue with haematoxylin for 15 minutes and 1 minute for deparaffinised and native tissues, respectively. After washing and stopping the haematoxylin reaction with H_2O the cytosol was stained with eosin for 3 minutes or 50 seconds for deparaffinised and native tissues, respectively. The tissue slides were washed with H_2O and dehydrated in an ethanol gradient. The H&E stained tissue slides were evaluated by a pathologist (Department of Pathology of the Bergmannsheil Hospital in Bochum) and compared to the HCA results of Raman data in order to select spectra for the training of the RF classifier.

RESULTS AND DISCUSSION

Raman based SHP

Raman based SHP of human colorectal tissue sections was used to obtain a high quality automated annotation of different tissue types and carcinoma regions. Before automated annotation can be performed it is important to build up a diverse set of spectra for the training of a classifier (see Fig. S1 in SI). Each spectrum has to be representative for a certain tissue component, which can be distinguished by vibrational spectroscopy. In an earlier study, we showed the capability of Raman imaging with 532 nm excitation for label

View Article Online
DOI: 10.1039/C6AN02072H

free detection of carcinoma regions, lymphocytes, erythrocytes and p53 active areas in carcinoma area.[23]
Here, by using the molecular information contained within the spectra, even more tissue components or cell types were automatically identified such as carcinoma tissue, connective tissue, muscle tissue, erythrocytes, lymphocytes, lymph follicle and crypts. The classification of tissue components was enhanced in the present study by developing a new multi-step classification scheme. The scheme is divided into two parts. In the first part a multi-step RF classifier (Fig. 1) was applied and classifies the classes, connective tissue, muscle, erythrocytes, lymphocytes, crypts, carcinoma and lymph-follicle. In the second part, parameters from a curve deconvolution were calculated for spectra which were identified as carcinoma or lymph-follicle. With these parameters, both classes were successfully reclassified by a LDA. Classified carcinoma regions, which were less than 10µm x 10 µm in size, were recognized to be cell nuclei. The datasets employed for the training step were excluded from validation. The datasets shown in this study for validation were acquired from an additional thin tissue section from patient with low grade and stage I colorectal cancer.
An example of Raman based SHP results from one patient will be presented here in details. Fig. 2A displays H&E stained tissue of a colorectal adenocarcinoma. Haematoxylin stains the cell nuclei in blue/purple, while eosin stains the cytosol in different red coloured shades.[32] The annotations of tissue components were performed by an expert pathologist.

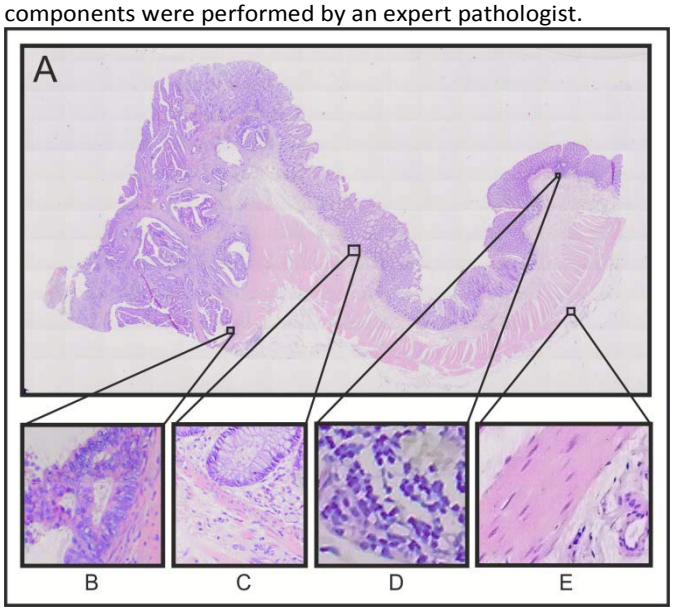

Figure 2. An image of H&E staining of a colorectal carcinoma tissue is shown in A. The colorectal adenocarcinoma is shown on the left side. Regions shown in B-E were selected for the Raman imaging and show different compositions of tissue types such as carcinoma region, muscle, connective tissue, crypt, lymphocytes, and single cell nuclei. Panel B shows carcinoma, muscle, and connective tissue. Panel C displays the mucosa containing the crypts and the submucosa separated by *lamina muscularis mucosae*. Lymph follicle is depicted in Panel D, while the transition between the *tunica muscularis* and the *tela serosa* is shown in Panel E.

These annotations were used to determine the regions of interest for Raman micro-spectroscopic measurements in the next step. The left side region of this tissue section shows the carcinoma area, whereas the right side region is non-cancerous (normal) region. In order to confirm the presence of carcinoma we performed IHC staining of the tissue with p53 and Ki-67 antibodies, which shows accumulation of p53 and Ki-67 proteins, respectively, in the left side region of the tissue (see Fig. S3 in SI).[32,33] According to the H&E staining and the IHC staining regions of interest were selected as displayed in Fig. 2. The selected regions show different tissue types characteristically for colon tissue. Panel B shows a carcinoma region, adjacent muscle and connective tissue. Panel C displays the border between the mucosa containing the colon crypts and the submucosa separated by a thin muscle layer called *lamina muscularis mucosae*, which plays an important role in the diagnosis of colon cancer. In panel D a part of a lymph follicle is presented, whereas panel E shows the transition between the *tunica muscularis* and the *tela serosa.*
For the diagnosis, it is important to differentiate between cancer, crypts and the membrane muscle layer. Fig. 3A and B show the transition from the *mucosa* containing the crypts to the *submucosa*. The two tissue types are separated by the *lamina muscularis mucosae*. We were not only able to automatically identify the nuclei part of the crypt (dark purple) but also the *lamina muscularis mucosae* (salmon). Furthermore, connective tissue (green) and even cellular shaped features like lymphocytes (pink), erythrocytes (olive) and undefined cell nuclei (blue) were automatically identified. Although a few pixels were misclassified, Raman SHP in Panel B reproduces all information from the H&E staining image (Panel A). In Fig. 3C, the p53 IHC staining for the selected carcinoma region in Fig. 2B is shown. Cancer regions (red) were identified by Raman based SHP (Fig. 3D) and these results are in an agreement with the IHC stained carcinoma area (Fig. 3C). The p53 active cancer (Fig. 3C) were obtained in the Raman SHP image (Fig. 3D) as red region. Inside the cancer region remaining goblet cells (dark purple) and infiltrating immunocompetent cells were observed (pink). The cancer region can be clearly separated towards the neighbouring muscle region (salmon). Small morphological differences were observed between IHC and SHP images because adjacent tissue slices were used.
The mean training spectra of these components are shown in Fig, S4 (see SI). The low standard-deviation of each class of the spectra, shown in grey, confirms the consistency of each class. In our continued approach towards Raman based automated SHP of colon carcinoma, we detected clear differences between carcinoma and connective tissue. Since the clinical use of the method is in focus, this clear differentiation was one of the main goals of our approach, and thus of great importance. The spectral differences between carcinoma and connective tissue are shown in details in Fig. 4a. These spectra were improved regarding their standard deviation compared to our previous study.[23] This is because new spectra were added for the training and spectra with a lower SNR were removed.
Large spectral differences between carcinoma and connective tissue are found. For example, the spectral differences can be found at 1330 and 787 cm^{-1}, indicating higher protein and DNA

View Article Online
DOI: 10.1039/C6AN02072K

content, respectively, in the carcinoma spectra. These results are similar to those reported previously.[4] A peak also appears at 1586 cm^{-1}, representing guanine and adenine (ring breathing modes of DNA bases).[4] The higher amount of DNA caused by enhanced proliferation is confirmed by the IHC staining for Ki-67 (See Fig. S3).

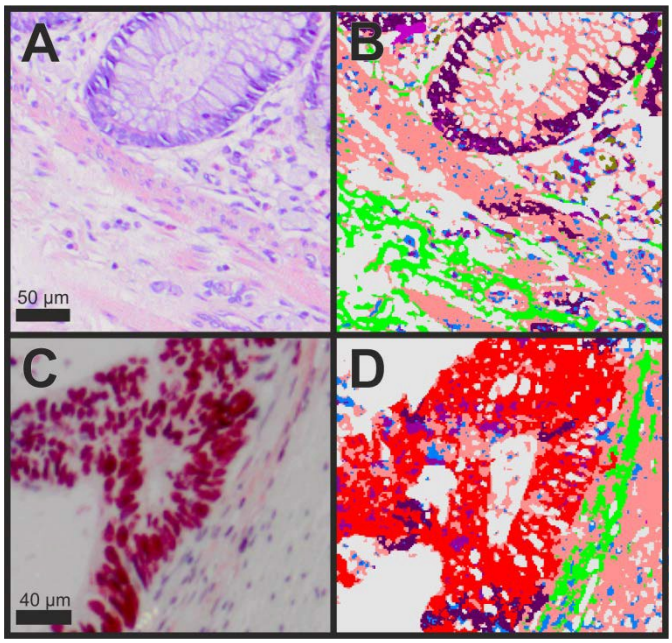

Figure 3. Comparison of H&E and IHC staining with Raman SHP of selected regions from colorectal carcinoma tissue presented in Fig. 2. (A) H&E staining of a region showing the transition from the *tunica mucosa* to the *tela submucosa*. (B) Raman SHP of the same region shown in (A). Salmon: muscle, dark purple: crypts, green: connective tissue, blue: cell nuclei, olive: erythrocytes, pink: lymphocytes. (C) IHC staining by p53 antibody of a region showing the transition between carcinoma tissue and healthy tissue. Accumulation of p53 is shown in red. (D) Raman SHP of the same region shown in (C). Red: carcinoma, green: connective tissue, salmon: muscle, dark purple: crypts, blue: cell nuclei.

The increased content of protein and DNA in carcinoma was also found by a Raman imaging study on gastric cancer[34] and in a fiber-optic approach for colon cancer.[31] On the other hand higher lipid contents were detected in the connective tissue spectra through the Raman bands at 860 and 1458 cm^{-1}. Higher lipid contents were also observed in the crypt in comparing with carcinoma through a Raman band at 1458 cm^{-1} as shown in Fig. 4b. In addition, crypts and *lamina muscularis mucosae* can be separated more clearly (Fig. 3B) in the present study. The differentiation between the crypts in mucosa and *lamina muscularis mucosae* is crucial for cancer diagnosis. This is because adenomas are formed in the mucosa, while the penetration of the *lamina muscularis mucosae* layer by carcinoma is defined as invasive cancer (see details in CARS results).

As different cell types, erythrocytes and lymphocytes were identified. The class of the erythrocytes shows the most characteristic spectra, since its hemoglobin is in resonance condition with 532 nm excitation laser. Due to normalization the enhancement caused by the resonance is not seen in the spectrum (Fig. S4 in SI). Nevertheless, the spectra shows enhanced characteristic bands for heme.[18,35,36] Differences in the spectra of lymph follicles and single lymphocyte cells in the tissue were also detected (see Fig. S5 in SI). The lymphocyte spectra have a characteristic pattern due to the large lipid content, with a strong band around 1443 cm^{-1} assigned to the CH_2 bending mode.[37] Brown et al. showed that it is possible to differentiate lymphocytes in different stage.[38] Furthermore, they reported small but significant differences in the Raman spectra of activated and non-activated T-lymphocytes. This could also be the reason for the spectral differences observed here between cells in the lymph follicle and the other lymphocytes within the tissue (see Fig. S5 in SI). The characteristic lipid bands are less intense in the spectra of the lymph follicle, where a higher content of protein bands are found. Thus, these results demonstrate the capability of Raman based SHP as a label-free method for recognition of several tissue components simultaneously and in an automated way.

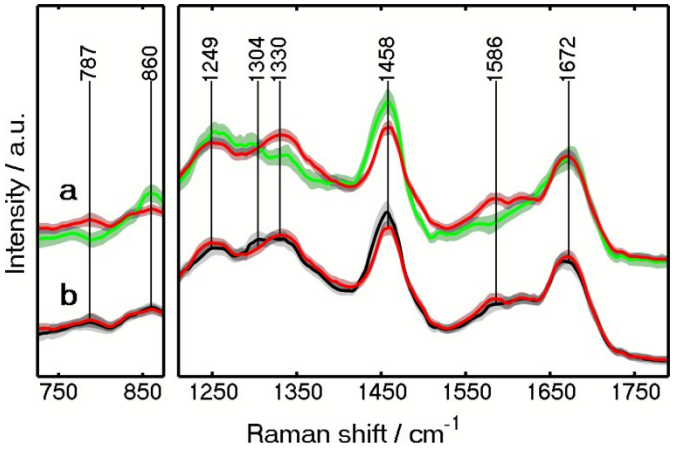

Figure 4. Wavelet-denoised Raman mean spectra of carcinoma (red), connective tissue (green), and crypt (black) used in Raman RF. The spectra are shown in 725-875 and 1210-1790 cm^{-1} regions.

Virtual staining by Raman micro-spectroscopy

Information of the local concentration of the single molecules and therewith components is lost due to the use of data correction and pre-processing. In order to get a high quality Raman SHP image with this additional information, we describe a new method to regain this structural Raman based SHP image by using the information provided by the integral Raman intensity from 2800-3050 cm^{-1} as shown in Fig. 5. The integrated Raman intensity image in the 2800-3050 cm^{-1} region is displayed in Fig. 5A, whereas Fig. 5B shows the Raman based SHP of the same region with the analysis scheme (single step RF) of the previous publication.[23]

The area shows the outer muscle layer of the colorectum (*muscularis propia*) at the top right corner, and the adjacent connective tissue (*subserosa*) at the bottom left. Especially in the muscle area the Raman based SHP (Fig. 5B) shows a problem with a lower SNR and the single step RF: a lot of misclassified pixels were recognized in the pseudo-colour image, which give the impression of a noisy image. By using a Gaussian filter (3x3 gaussian matrix, σ=1) and a multistep RF

View Article Online
DOI: 10.1039/C6AN02072K

(see data analysis section) the Raman based SHP was further improved as shown in Fig. 5C.

The image displays a large improvement to the Raman based SHP in Fig. 5B. The muscle (salmon) and the connective tissue (green) can be separated precisely. Nevertheless, the representation of the tissue appears very flat and homogenous within one component.

The concentration of molecules within a voxel is directly proportional to the integrated Raman intensities. Therefore, information of the detailed structures of a certain tissue can e.g. be obtained from the integrated signal of the CH stretching vibrations in the 2800-3050 cm^{-1} region.

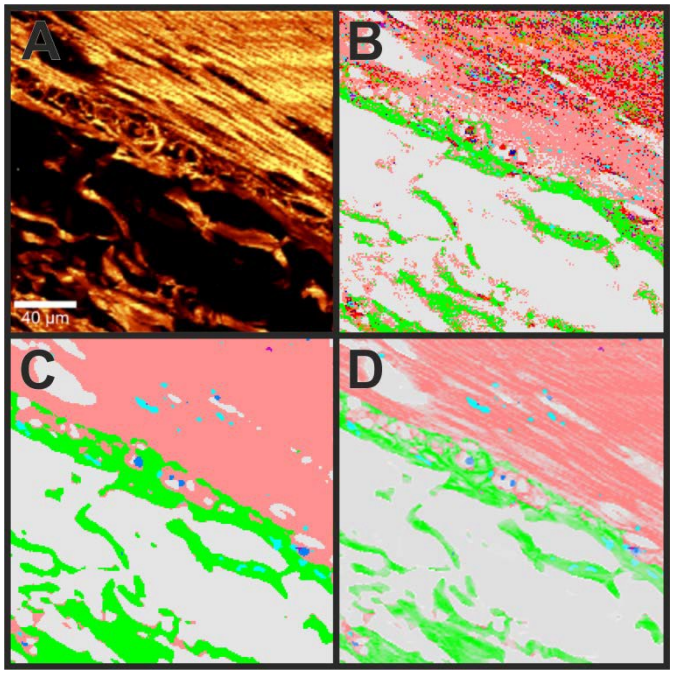

Figure 5. Raman virtual staining of tissue. (A) Integral Raman intensity image in the 2800-3050 cm^{-1} region collected from muscle and connective tissue region. (B) Raman pseudo-colour image constructed from a single step Raman RF classifier on baseline corrected and normalized data. (C) Raman pseudo-colour image constructed from a multistep RF with Gaussian filtered data image space and normalization. Salmon: muscle, green: connective tissue. (D) Raman virtual staining, constructed from Raman intensities in (A) overlaid with (C).

A combination of the pseudo-colour map of the classification and the Raman intensities allows displaying of fine structures such as the filamentous structure within muscle tissue and its orientation (Fig. 5D). This increases the information content of the presented data and gives an impression of colouring the tissue structure with our classification, which is comparable to stained tissue with dyes. Thus, instead of using a dye as in H&E staining, this method provides a label-free way to virtually stain a tissue and is therefore, a non-invasive approach. An advantage of the label-free Raman imaging in comparison with H&E is that the same tissue section can be also used in a non-invasive manner for further analysis, like next generation sequencing, proteomic analysis, or immunohistochemistry. In principle, immunohistochemistry can be performed after H&E de-staining but it is invasive method and the probability of losing the tissue sections during this process is relatively high.[39] The virtual staining approach was applied to the previously selected regions shown in Fig. 2B-E. Fig. 6 shows the H&E staining of the four selected regions (A,D,G,J) in direct comparison to the Raman SHP (B,E,H,K) and the virtually stained Raman images (C,F,I,L).

The picture clearly shows the enhanced visualization of the muscle fibres (salmon) and connective (green) tissue. The resolution of Raman imaging intensities over the CH stretching vibrations is roughly equal to the conventional imaging of the H&E, and sometimes even seems to deliver a more detailed and sharper representation of the sample. The same images can be created with black background if necessary (see Fig. S6-S8 in SI).

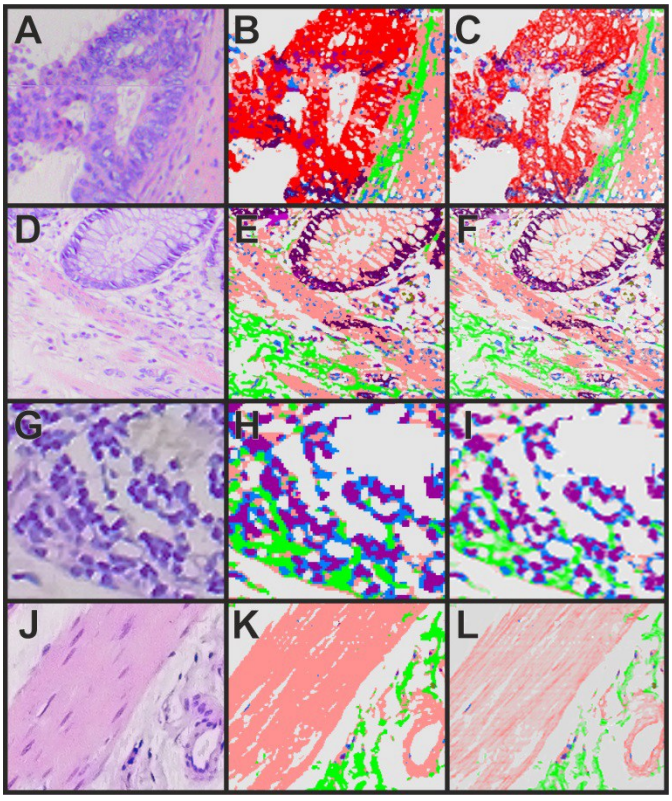

Figure 6. Comparison of H&E staining, Raman SHP and Raman virtual staining. (A,D,G,J) H&E staining of selected regions of interest shown in Fig.1. (B,E,H,K). Raman SHP of the same regions shown in (A,D,G,J). Red: carcinoma, green: connective tissue, salmon: muscle, dark purple: crypts, pink: lymphocytes, olive: erythrocytes, purple: lymph follicle. (C,F,I,L) Raman virtual staining, constructed from Raman SHP shown in (B,E,H,K) and their corresponding integrated Raman intensities in the 2800-3050 cm^{-1} region.

The sensitivity and the specificity for carcinoma recognition are at 96 % and 98 %, respectively, as shown in Table 1. This shows how precise the carcinoma can be allocated. Other approaches using Raman spectroscopy on carcinoma or basal cell carcinoma show similar results for the sensitivity and specificity.[40,41]

The virtual stained images give comparable results to the stained and labelled tissue sections. This proves Raman based SHP as a label-free supplement to the standard methods of diagnosis, as H&E and IHC staining. It does not only identify biomarker in human tissue, as shown here for colorectal

View Article Online
DOI: 10.1039/C6AN02072K

cancer, but it can be also used as a diagnostic assistant technique.

	Sensitivity %	Specificity %
Carcinoma	96	98
Crypts	96	99
Lymph follicle	86	99
Connective tissue	93	99
Muscle	98	99
Lymphocytes	99	99
Erythrocytes	100	100

Table 1. The sensitivity and specificity for recognition of different tissue classes.

HE images of tissue by multivariate statistics. This method is proposed to be a fast and precise pathological screening tool.

CARS imaging at a single wavenumber is very common in bio-spectroscopy. For instance, CARS imaging near 2850 cm^{-1} has been used to monitor lipid distribution in tissues or lipid droplets in cancer cells.[27,45–47] In addition to imaging at a single wavenumber, *Potma et al.*[45] used CARS spectra in the C—H stretching region coupled to principle component analysis to image meibomian glands. We have also used similar approach including CARS spectra in the C—H stretching region and cluster analysis to identify subcellular organelles of cancer cells.[27] Generally, clustering of CARS spectra produces pseudo-colour image, which represents the various spectral

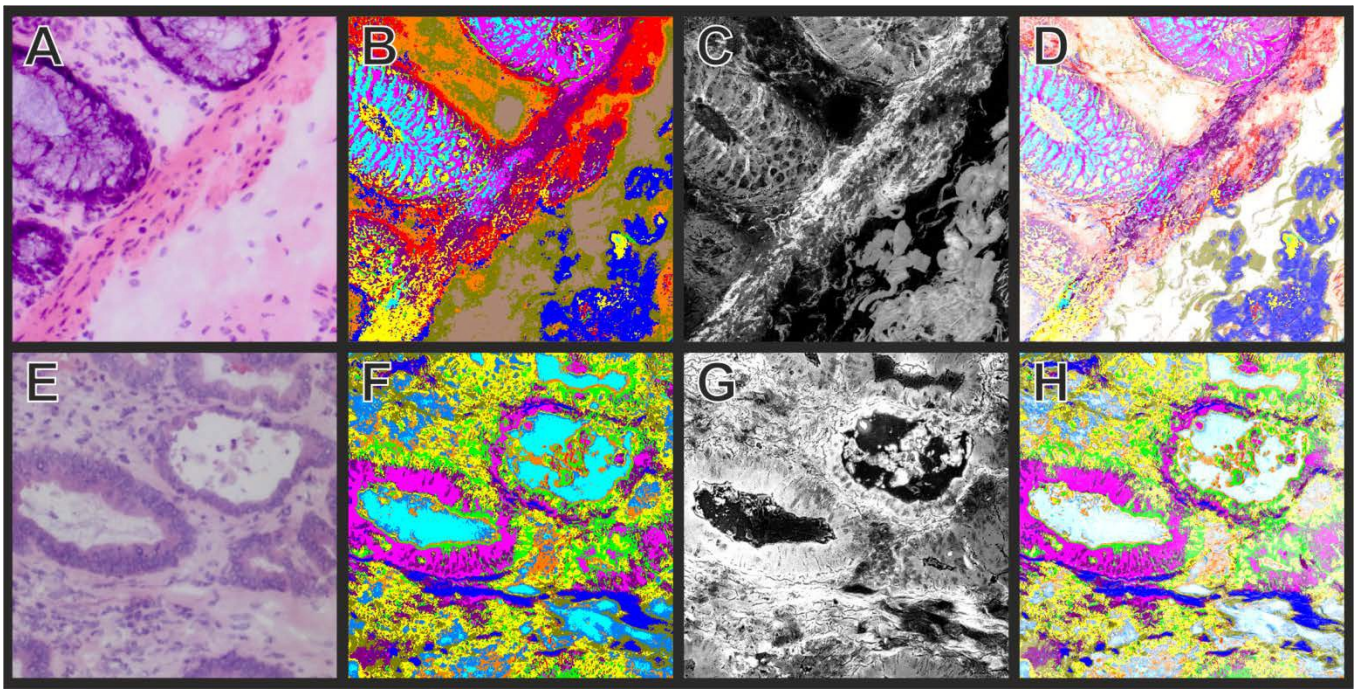

Figure 7. Comparison of H&E, SHG, and CARS imaging of benign and carcinoma regions from patient with low grade and stage IIA colorectal carcinoma. (A,E) H&E staining of selected regions of interest. (B,F) *k*-means clustering of CARS spectral datasets. (C,G) Intensity images of both SHG at 408 nm and CARS signals at 2850 cm^{-1}. (D,H) Constructed images from *k*-means clustering of CARS results (B,F) and intensities of both CARS and SHG (C,G).

Fast imaging by CARS micro-spectroscopy

One disadvantage of the conventional Raman micro-spectroscopy is that Raman measurements are slow. This is a problem for clinical applications that require generally fast measurements. For instance, the surgeon has a short time during the surgery until the pathologist determines the cancerous part of tissues that is necessary to be removed. To overcome the speed problem of the Raman measurements and make it fit with clinical applications, non-linear techniques such as CARS or stimulated Raman scattering (SRS) that can be performed at a speed up to video rate, have to be used.[42,43] Ji et al.[26] showed that SRS is able to detect brain carcinomas using a linear combination of SRS images at 2845 and 2930 cm^{-1} in a rapid and label-free way, but didn't include any bioinformatics approach to aid the pathologist. In addition, Bocklitz et al.[44] have used a combination of CARS, two photon excited autofluorescence (TPEF), and SHG to produce pseudo-distribution over the examined tissue section. Different tissue components shown in the pseudo-colour images can be identified by comparison with the corresponding an H&E stained image of the same tissue section or the next adjacent tissue slice as shown in the workflow of SHP (Fig. S1). This would provide more information than those can be obtained from CARS imaging at a single wavenumber. Here, we have used CARS spectra in the C—H stretching region in combination with SHG and cluster analysis to set the stage for automatic identification of different tissue components.

The H&E staining images shown in Fig. 7. display regions with benign (A) and carcinoma (E) morphological features. For instance, Panel A shows the transition from the *mucosa* containing intact crypts to the *submucosa* and they are separated by the *lamina muscularis*— *mucosae*. In Panel E, cancer with a low analplasia can be seen. The original structure of the crypt regions still can be observed. Fig. 7B and F display the *k*-means clustering result of CARS spectra in the 2700-3000 cm^{-1} region of the tissue sections shown in Panels A and E,

View Article Online
DOI: 10.1039/C6AN03072K

respectively. These Panels accurately reproduces the structures that are apparent in the H&E stained images: the intact crypts (cyan and pink), submucosa (dark blue), and *lamina muscularis mucosae* (purple) are clearly shown in Panel B, while carcinoma (pink) is displayed in in Panel F. Examples of the CARS mean spectra that are obtained from *k*-means clustering are shown in Fig. S9 (see SI).

Fig. 7C shows a combination of the SHG (408 nm) and CARS intensities at 2850 cm^{-1}. SHG of tissues at 408 nm visualizes mainly the fibrous collagen network, whereas the CARS intensity at 2850 m^{-1} depicts the lipid rich regions in tissue. To obtain more information about the structural details, the pseudo-colour images of *k*-means (B and F) are combined with the intensity images of both SHG and CARS (C and G) and the results are displayed in Panels D and H.

Although the concentration of one component is non-linearly proportional to its CARS or SHG intensity, such a combination improves the quality of the images as shown in details in Fig. S10 and S11 (SI). These results are comparable to the H&E staining (Fig. 7A and E, see also Fig. S10 and S11). The images (Fig. 7D and H) display an improvement in comparing with the pseudo-colour image of *k*-means clustering (Fig. 7B and F). For instance, the crypt (cyan and pink), the connective tissue (blue and olive), *lamina muscularis* (for example purple) can be separated precisely from one another (Fig. 7D). The invasive carcinoma is clearly visible in Panel H. Furthermore, the detailed structures of these components are more visible in Fig. 7 D and H. Similar to the Raman virtual imaging shown above (Fig. 6), images of high structural details can be generated with a faster imaging technique such as CARS.

CONCLUSIONS

We have presented that Raman based SHP can differentiate between several tissue components, including cancer, connective tissue, muscle, crypts, lymphocytes, lymph follicle, and erythrocytes. The information content of the pseudo-colour images can be further improved by overlaying the Raman intensities of the C—H stretching vibrations with the Raman based RF images. By this new method, virtual staining, structural details of the tissue such as fibres can be revealed and thus improves the representation of the highly resolved images. Virtual staining by Raman based SHP provides a realistic display of the tissues structure similar to conventional staining techniques like fluorescence imaging and allows a better direct comparison to the H&E staining method. The high sensitivity and the specificity for cancer recognition confirm how precise the cancer can be automatically detected. This method could therefore help the pathologist to diagnose cancer allocated regions and early stages of the disease with high precision in order to improve patients life quality.

Recent studies focus on fast measurements of tissue using non-linear techniques such as SRS, but lack the bioinformatics.[26] Pseudo-HE images of tissues was also created using CARS/TPEF/SHG to be used as a pathological screening tool.[44] On the other hand, pseudo-colour images of *k*-means of CARS spectra in the CH stretching region and SHG at 408 nm was used to produce a highly resolved pseudo-colour images that can be used to differentiate between different tissue types. A combination of the presented data evaluation and CARS measurements paves the way for fast clinical label-free diagnostics. Our next step is to perform CARS measurements of native colorectal cancer tissues from several patients to obtain a large data set that enables us to perform automatic recognition of various tissue components including carcinoma.

Acknowledgements

We thank Angela Kallenbach-Thieltges, Frederick Großerüschkamp, Claus Küpper and Melanie Horn for helpful discussions. Furthermore, we thank Lidia Janota for her expertise in tissue staining. This research was supported by the Protein Research Unit Ruhr within Europe (PURE), Ministry of Innovation, Science and Research (MIWF) of North-Rhine Westphalia, Germany.

Notes and references

1 *World Cancer Report 2014.*, World Health Organization, 2014.
2 D. Cunningham, W. Atkin, H.-J. Lenz, H. T. Lynch, B. Minsky, B. Nordlinger and N. Starling, *The Lancet*, 2010, **375**, 1030–1047.
3 B. T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J. Rogers and J. Bähler, *Nature*, 2008, **453**, 1239–1243.
4 M. Diem, J. M. Chalmers and P. R. Griffiths, *Vibrational spectroscopy for medical diagnosis*, John Wiley & Sons, Chichester, England; Hoboken, NJ, 2008.
5 R. Salzer and H. W. Siesler, *Infrared and Raman spectroscopic imaging*, Wiley-VCH, Weinheim, 2009.
6 M. Diem, M. Miljkovic, B. Bird, T. Chernenko, J. Schubert, E. Marcsisin, A. Mazur, E. Kingston, E. Zuser, K. Papamarkakis and N. Laver, *Spectrosc. Int. J.*, 2012, **27**, 463–496.
7 M. Diem, A. Mazur, K. Lenau, J. Schubert, B. Bird, M. Miljković, C. Krafft and J. Popp, *J. Biophotonics*, 2013, **6**, 855–886.
8 K. Kong, C. J. Rowlands, S. Varma, W. Perkins, I. H. Leach, A. A. Koloydenko, H. C. Williams and I. Notingher, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 15189–15194.
9 H. J. Byrne, M. Baranska, G. J. Puppels, N. Stone, B. Wood, K. M. Gough, P. Lasch, P. Heraud, J. Sulé-Suso and G. D. Sockalingum, *The Analyst*, 2015, **140**, 2066–2073.
10 N. Rashid, H. Nawaz, K. W. C. Poon, F. Bonnier, S. Bakhiet, C. Martin, J. J. O'Leary, H. J. Byrne and F. M. Lyng, *Exp. Mol. Pathol.*, 2014, **97**, 554–564.
11 K. Kong, C. Kendall, N. Stone and I. Notingher, *Adv. Drug Deliv. Rev.*, 2015, **89**, 121–134.
12 A. Kallenbach-Thieltges, F. Großerüschkamp, A. Mosig, M. Diem, A. Tannapfel and K. Gerwert, *J. Biophotonics*, 2013, **6**, 88–100.
13 C. Kuepper, F. Großerueschkamp, A. Kallenbach-Thieltges, A. Mosig, A. Tannapfel and K. Gerwert, *Faraday Discuss*, 2016, **187**, 105–118.
14 F. Großerueschkamp, A. Kallenbach-Thieltges, T. Behrens, T. Brüning, M. Altmayer, G. Stamatis, D. Theegarten and K. Gerwert, *The Analyst*, 2015, **140**, 2114–2120.
15 P. Lasch, M. Diem, W. Hänsch and D. Naumann, *J. Chemom.*, 2006, **20**, 209–220.
16 B. Bird, M. Miljković, S. Remiszewski, A. Akalin, M. Kon and M. Diem, *Lab. Invest.*, 2012, **92**, 1358–1373.

17 C. Krafft, D. Codrich, G. Pelizzo and V. Sergo, *J. Biophotonics*, 2008, **1**, 154–169.
18 C. Krafft, B. Dietzek, M. Schmitt and J. Popp, *J. Biomed. Opt.*, 2012, **17**, 40801.
19 A. Beljebbar, O. Bouché, M. D. Diébold, P. J. Guillou, J. P. Palot, D. Eudes and M. Manfait, *Crit. Rev. Oncol. Hematol.*, 2009, **72**, 255–264.
20 R. Gaifulina, A. T. Maher, C. Kendall, J. Nelson, M. Rodriguez-Justo, K. Lau and G. M. Thomas, *Int. J. Exp. Pathol.*, 2016.
21 T. W. Bocklitz, S. Guo, O. Ryabchykov, N. Vogler and J. Popp, *Anal. Chem.*, 2016, **88**, 133–151.
22 N. Vogler, T. Bocklitz, F. Subhi Salah, C. Schmidt, R. Bräuer, T. Cui, M. Mireskandari, F. R. Greten, M. Schmitt, A. Stallmach, I. Petersen and J. Popp, *J. Biophotonics*, 2016, **9**, 533–541.
23 L. Mavarani, D. Petersen, S. F. El-Mashtoly, A. Mosig, A. Tannapfel, C. Kötting and K. Gerwert, *The Analyst*, 2013, **138**, 4035–4039.
24 Leo Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
25 K. Kong, C. Kendall, N. Stone and I. Notingher, *Adv. Drug Deliv. Rev.*, 2015.
26 M. Ji, D. A. Orringer, C. W. Freudiger, S. Ramkissoon, X. Liu, D. Lau, A. J. Golby, I. Norton, M. Hayashi, N. Y. R. Agar, G. S. Young, C. Spino, S. Santagata, S. Camelo-Piragua, K. L. Ligon, O. Sagher and X. S. Xie, *Sci. Transl. Med.*, 2013, **5**, 201ra119-201ra119.
27 S. F. El-Mashtoly, D. Niedieker, D. Petersen, S. D. Krauss, E. Freier, A. Maghnouj, A. Mosig, S. Hahn, C. Kötting and K. Gerwert, *Biophys. J.*, 2014, **106**, 1910–1920.
28 G. Judith and Kumarasabapathy, *SIPIJ*, 2011, **2**, 82–92.
29 M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus and M. Diem, *The Analyst*, 2010, **135**, 2002–2013.
30 D. L. Donoho, *IEEE Trans. Inf. Theory*, 1995, **41**, 613–627.
31 M. V. P. Chowdary, K. K. Kumar, K. Thakur, A. Anand, J. Kurien, C. M. Krishna and S. Mathew, *Photomed. Laser Surg.*, 2007, **25**, 269–274.
32 G. Avwioro, *JPCS*, 2011, **1**, 24–34.
33 M. Ramael, G. Lemmens, C. Eerdekens, C. Buysse, I. Deblier, W. Jacobs and E. Van Marck, *J. Pathol.*, 1992, **168**, 371–375.
34 M. S. Bergholt, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. Y. So, A. Shabbir and Z. Huang, *J. Biophotonics*, 2013, **6**, 49–59.
35 G. Rusciano, *Phys. Medica PM Int. J. Devoted Appl. Phys. Med. Biol. Off. J. Ital. Assoc. Biomed. Phys. AIFB*, 2010, **26**, 233–239.
36 M. Asghari-Khiavi, A. Mechler, K. R. Bambery, D. McNaughton and B. R. Wood, *J. Raman Spectrosc.*, 2009, **40**, 1668–1674.
37 A. I. Mazur, J. L. Monahan, M. Miljković, N. Laver, M. Diem and B. Bird, *J. Biophotonics*, 2013, **6**, 101–109.
38 K. L. Brown, O. Y. Palyvoda, J. S. Thakur, S. L. Nehlsen-Cannarella, O. R. Fagoaga, S. A. Gruber and G. W. Auner, *J. Immunol. Methods*, 2009, **340**, 48–54.
39 M. Dardik and J. I. Epstein, *Hum. Pathol.*, 2000, **31**, 1155–1161.
40 N. Bergner, T. Bocklitz, B. F. M. Romeike, R. Reichart, R. Kalff, C. Krafft and J. Popp, *Chemom. Intell. Lab. Syst.*, 2012, **117**, 224–232.
41 A. Nijssen, T. C. Bakker Schut, F. Heule, P. J. Caspers, D. P. Hayes, M. H. A. Neumann and G. J. Puppels, *J. Invest. Dermatol.*, 2002, **119**, 64–69.
42 B. G. Saar, C. W. Freudiger, J. Reichman, C. M. Stanley, G. R. Holtom and X. S. Xie, *Science*, 2010, **330**, 1368–1370.
43 C. Krafft, B. Dietzek and J. Popp, *The Analyst*, 2009, **134**, 1046–1057.
44 T. W. Bocklitz, F. S. Salah, N. Vogler, S. Heuke, O. Chernavskaia, C. Schmidt, M. J. Waldner, F. R. Greten, R. Bräuer, M. Schmitt, A. Stallmach, I. Petersen and J. Popp, *BMC Cancer*, 2016, **16**, 534.
45 C.-Y. Lin, J. L. Suhalim, C. L. Nien, M. D. Miljković, M. Diem, J. V. Jester and E. O. Potma, *J. Biomed. Opt.*, 2011, **16**, 21104.
46 M. A. Fernandez, C. Albor, M. Ingelmo-Torres, S. J. Nixon, C. Ferguson, T. Kurzchalia, F. Tebar, C. Enrich, R. G. Parton and A. Pol, *Science*, 2006, **313**, 1628–1632.
47 P. Boström, L. Andersson, M. Rutberg, J. Perman, U. Lidberg, B. R. Johansson, J. Fernandez-Rodriguez, J. Ericson, T. Nilsson, J. Borén and S.-O. Olofsson, *Nat. Cell Biol.*, 2007, **9**, 1286–1293.

Electronic Supplementary Material (ESI) for Analyst.

Supplementary Information

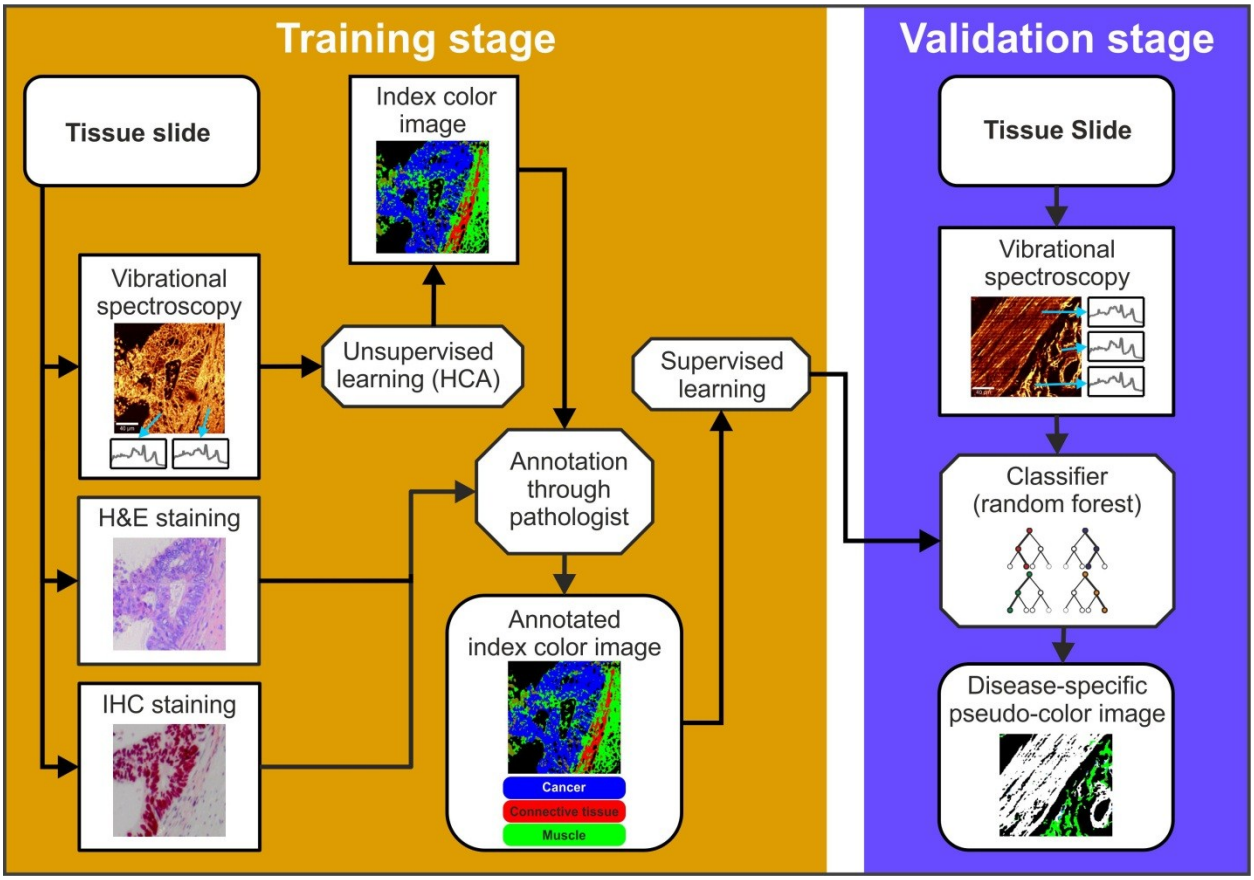

Fig. S1. Workflow describing the training and validation stages for Raman based SHP. The pseudo-colour images created by unsupervised learning algorithm are annotated by pathologist with help of H&E and IHC staining. These spectra build up a training data set for a supervised learning algorithm (RF), which can classify other tissue slides automatically.

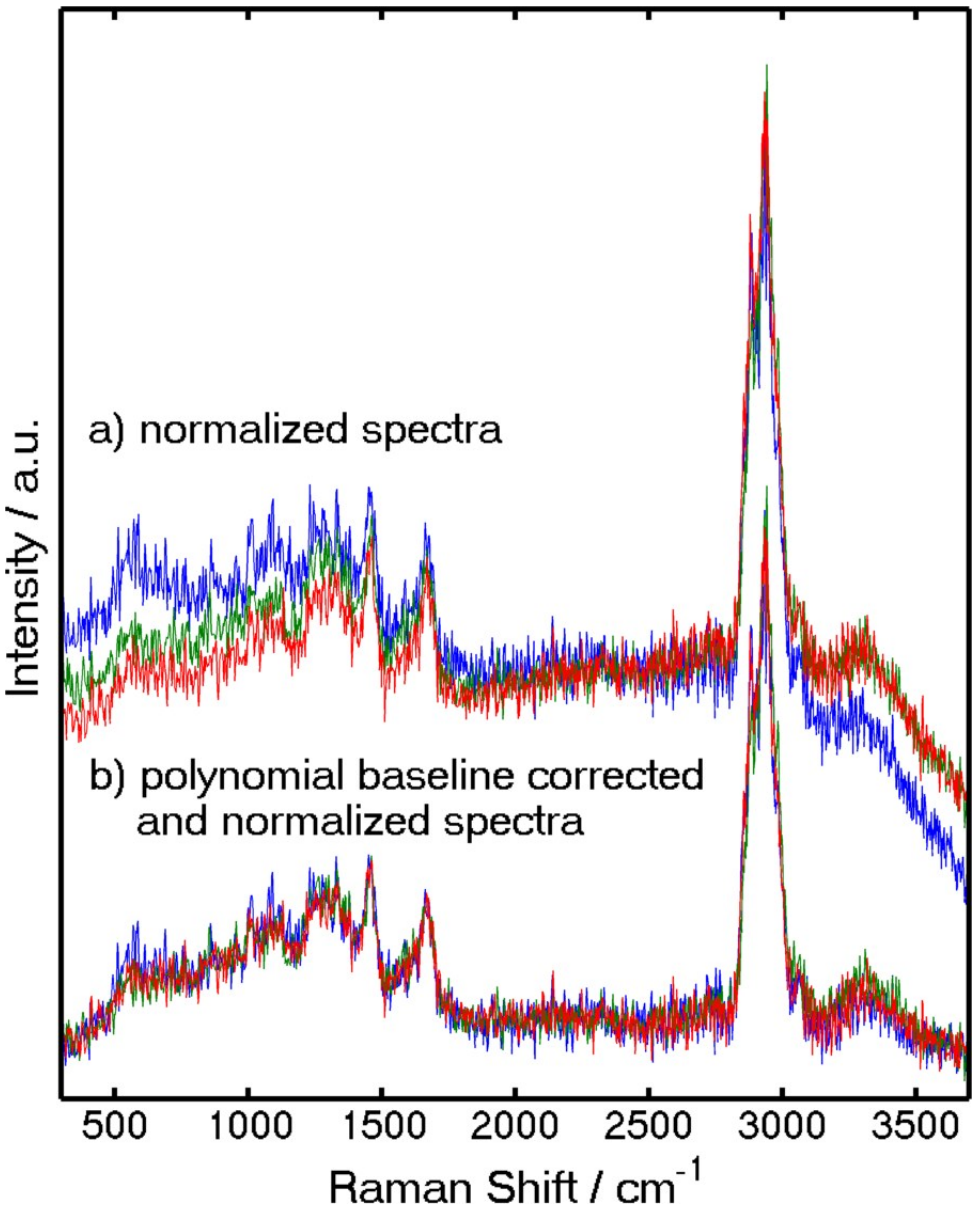

Fig. S2. Effect of baseline correction on Raman spectra. Raman spectra of tissue at different pixels without (a) and with baseline correction (b).

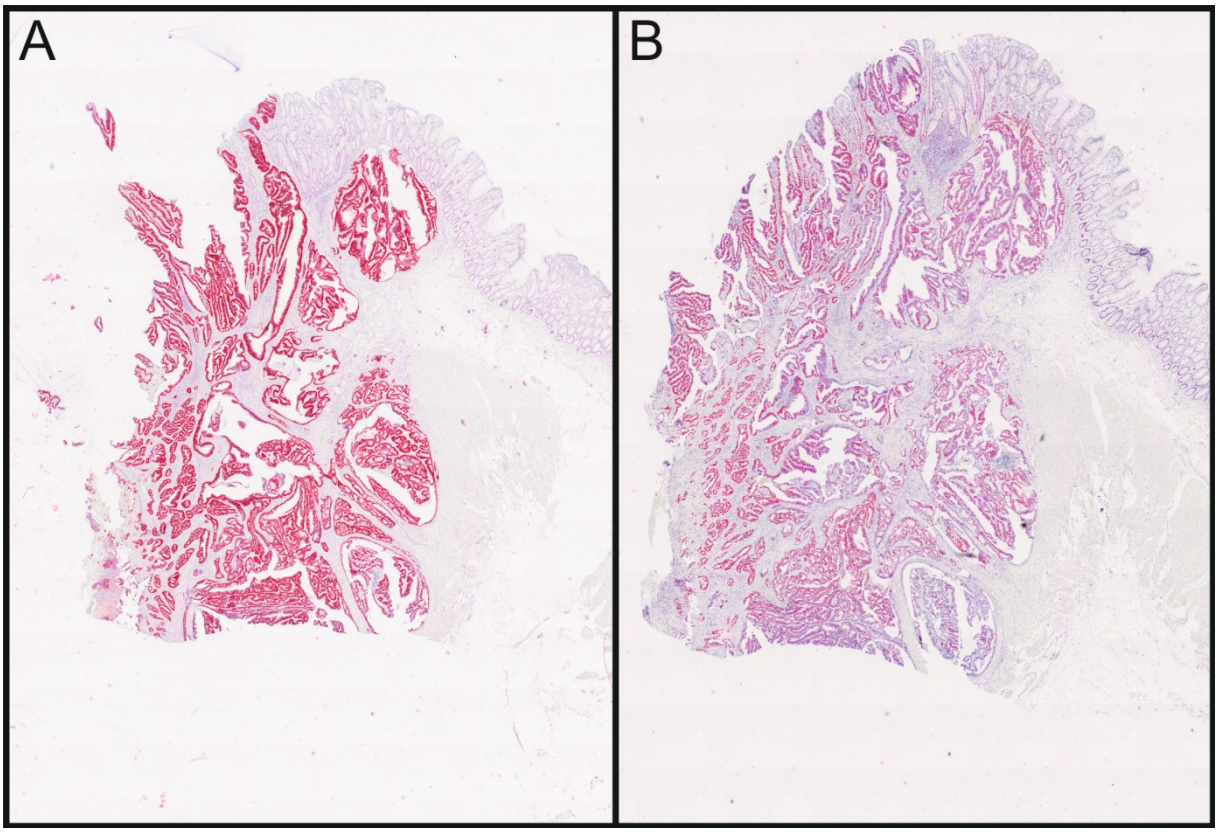

Fig. S3. Immunohistochemical staining of colorectal carcinoma tissue slide presented in Fig. 1. (A) Immunohistochemical staining of a colon tissue section by p53 antibody. Accumulation of p53 is shown in red. (B) Immunohistochemical staining of a colon tissue section by MiB-1 (Ki-67) antibody. Proliferating cells are shown in red.

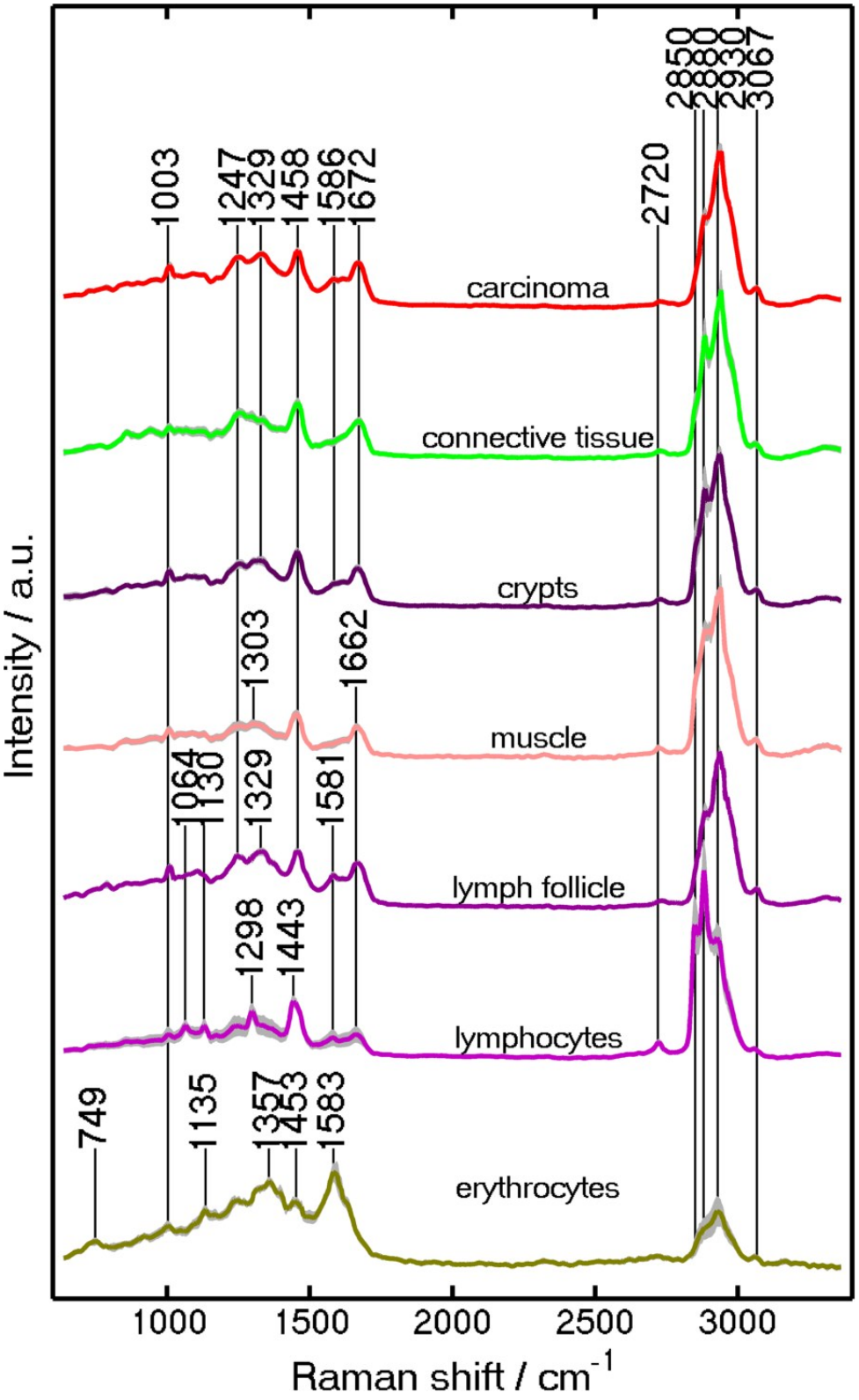

Fig. S4. Raman mean spectra of tissue components used in Raman RF classifier. The spectra are wavelet-denoised and the standard-deviation is marked in grey. The spectra are shown from 700-3500 cm^{-1}.

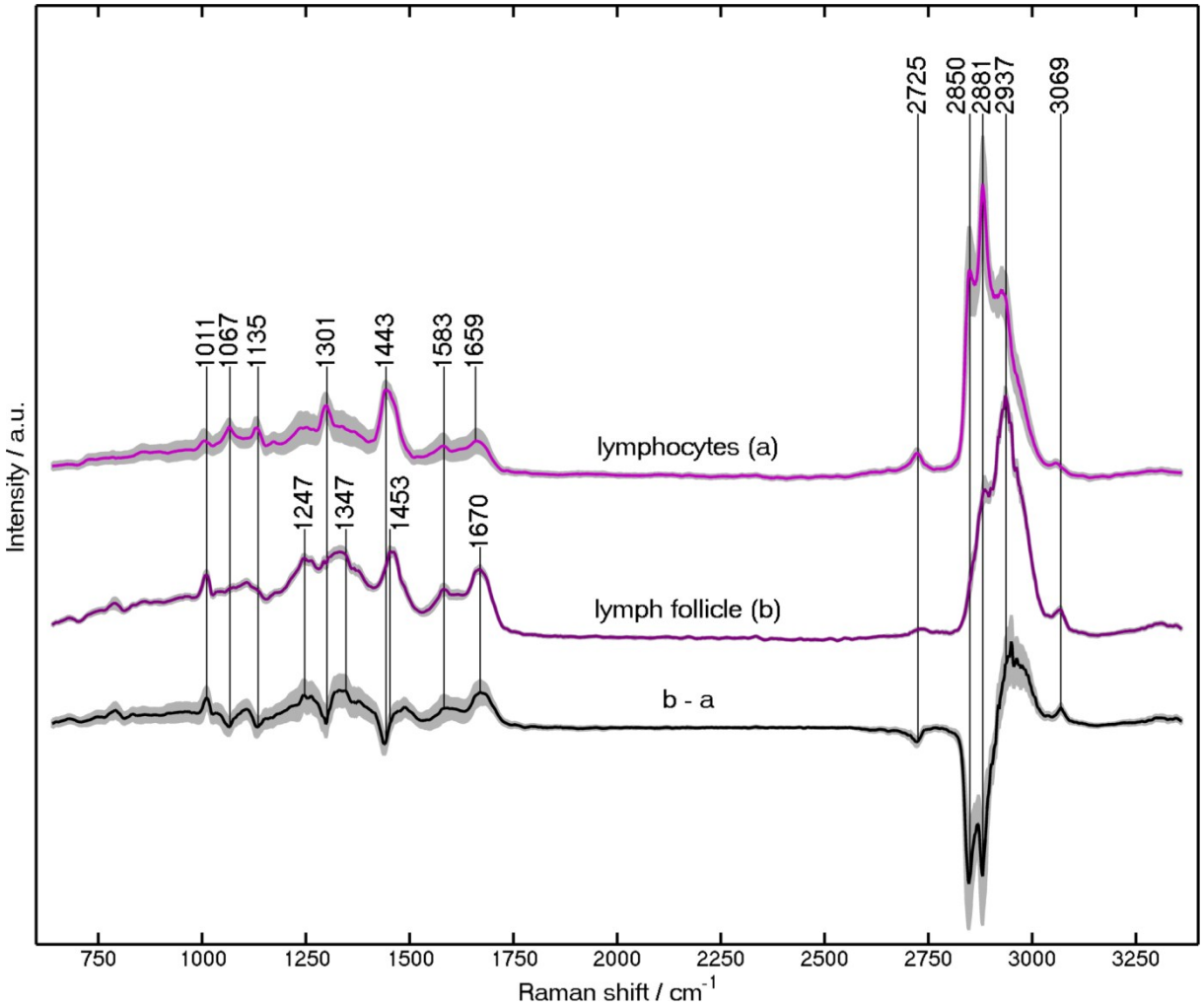

Fig. S5. Raman mean spectra of lymphocytes and lymph follicle used in Raman RF. The spectra are wavelet-denoised and the standard-deviation is marked in grey. The spectra are shown from 700-3500 cm^{-1}. The difference spectrum is showed in black.

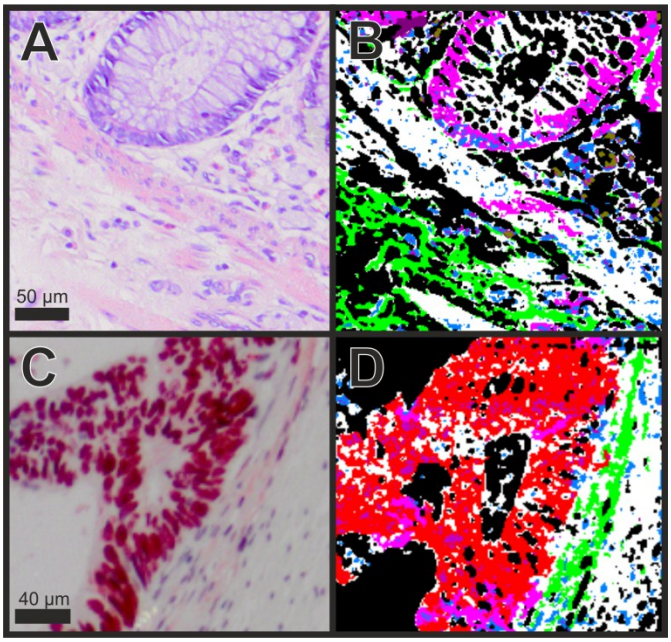

Fig. S6. Alternative representation for Fig. 3. Direct comparison of H&E and immunohistochemical staining with Raman virtual staining of selected regions from colorectal carcinoma tissue slide presented in Fig. 2.

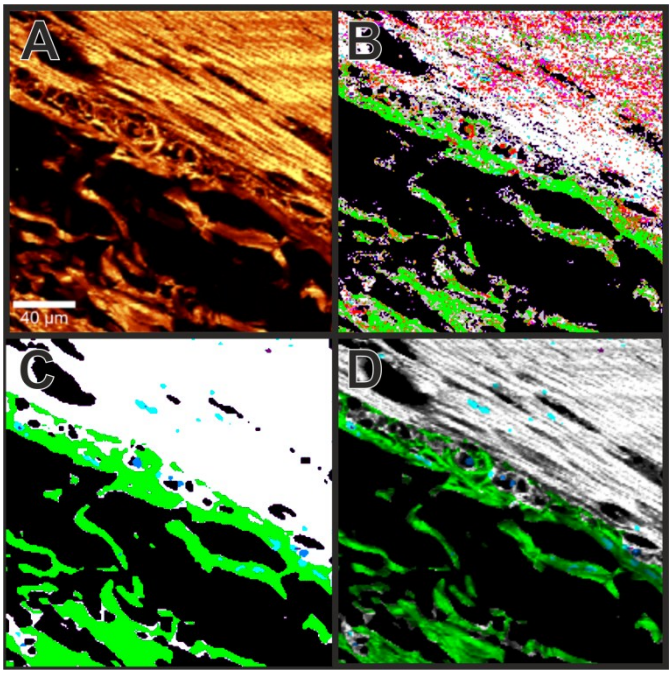

Fig. S7. Alternative representation for Fig. 5. Improvement of representation of Raman based SHP and establishment of Raman virtual staining.

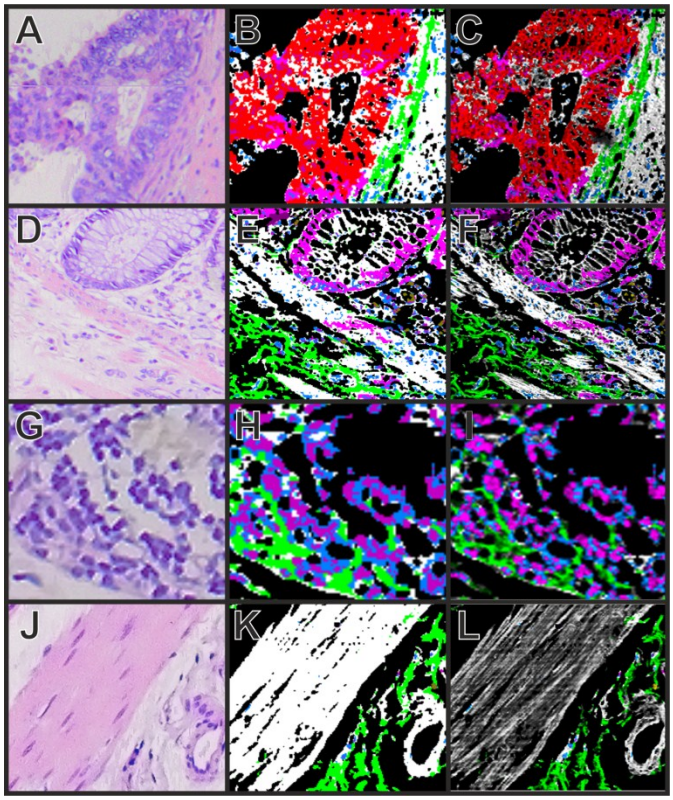

Fig. S8. Alternative representation for Fig. 6. (A,D,G,J) H&E staining of selected regions of interest shown in Fig. 2. (B,E,H,K) Raman SHP of the same regions shown in (A,D,G,J). (C,F,I,L) Raman virtual staining, constructed from Raman SHP shown in (B,E,H,K) and their corresponding integrated Raman intensities in the 2800-3050 cm^{-1} region.

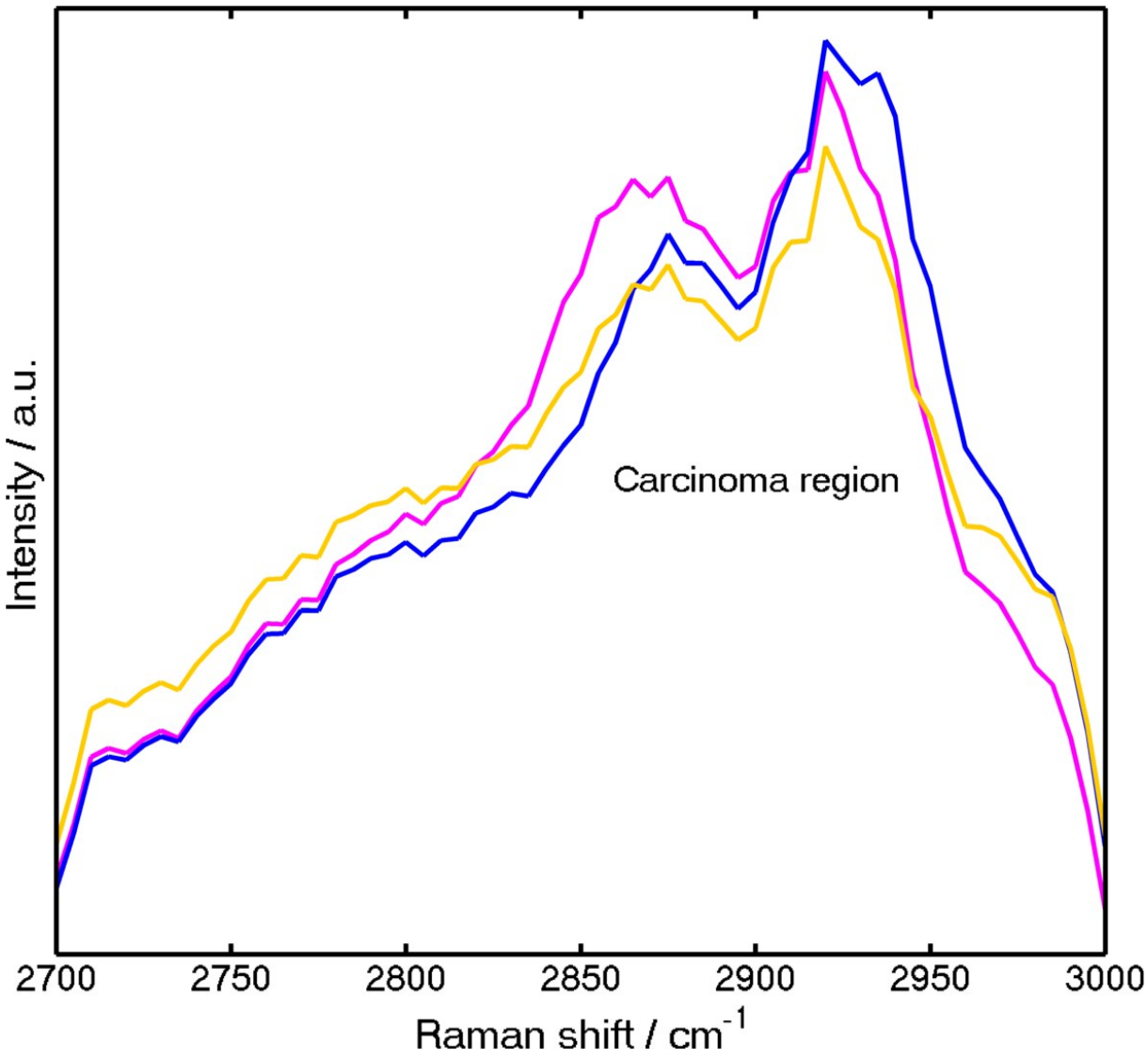

Fig. S9. CARS mean spectra of carcinoma (pink), connective tissue (blue), and tumor micro-environment (yellow). These spectra are produced using *k*-means clustering of CARS spectral dataset. For illustration the spectra are median filtered in the frequency domain.

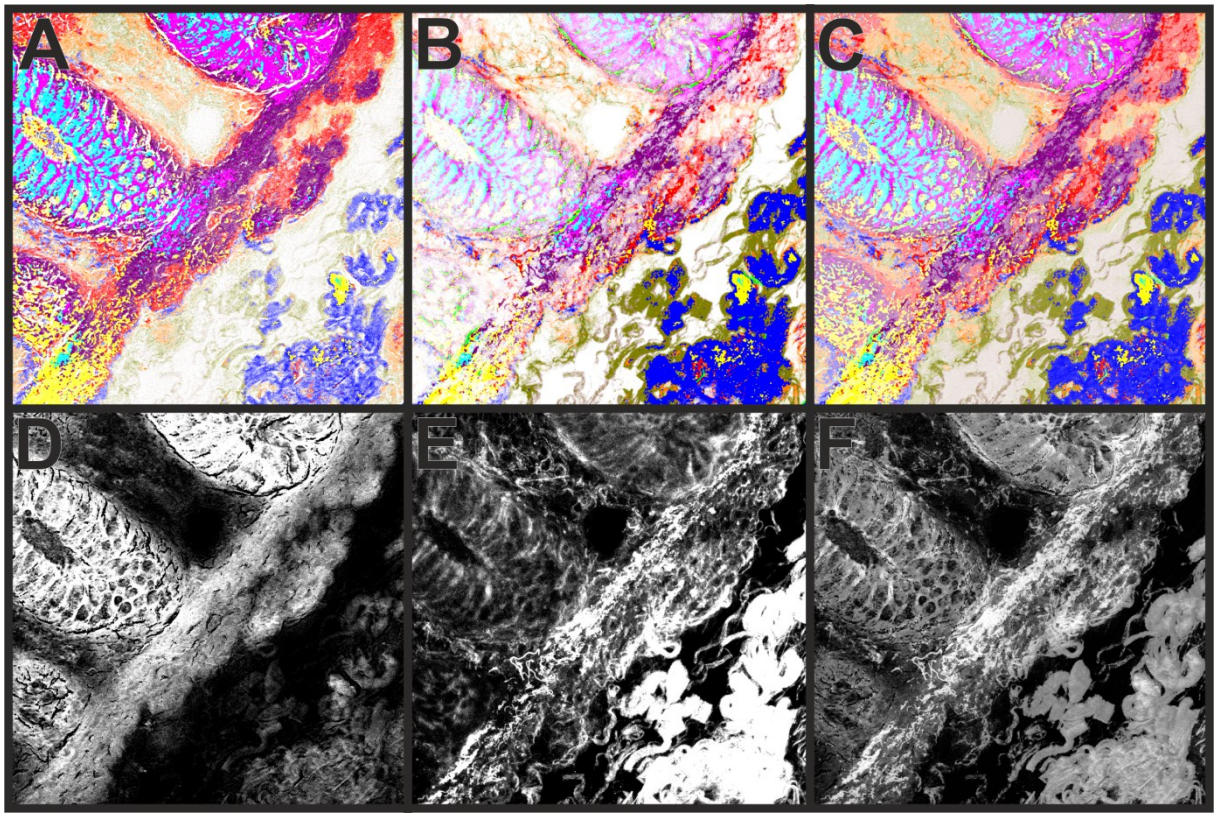

Fig. S10. Constructed images from CARS *k*-means clustering analysis weighted with intensities of CARS (A), SHG (B) and both together (C). The pure intensity images of CARS at 2850 cm^{-1} (D), SHG (E) and combined CARS and SHG (F) are also displayed.

Similar to the Raman virtual staining, we created intensity weighted pseudo-color images from *k*-means cluster analysis of CARS datasets. The pseudo-color image in Fig. 7B was weighted here with the CARS intensity at 2850 cm^{-1} (A,D), with the SHG intensity (B,E) and with a combination of both intensities (C,F). It is important to note that the intensities are non-linear in comparison to the linear Raman signal.

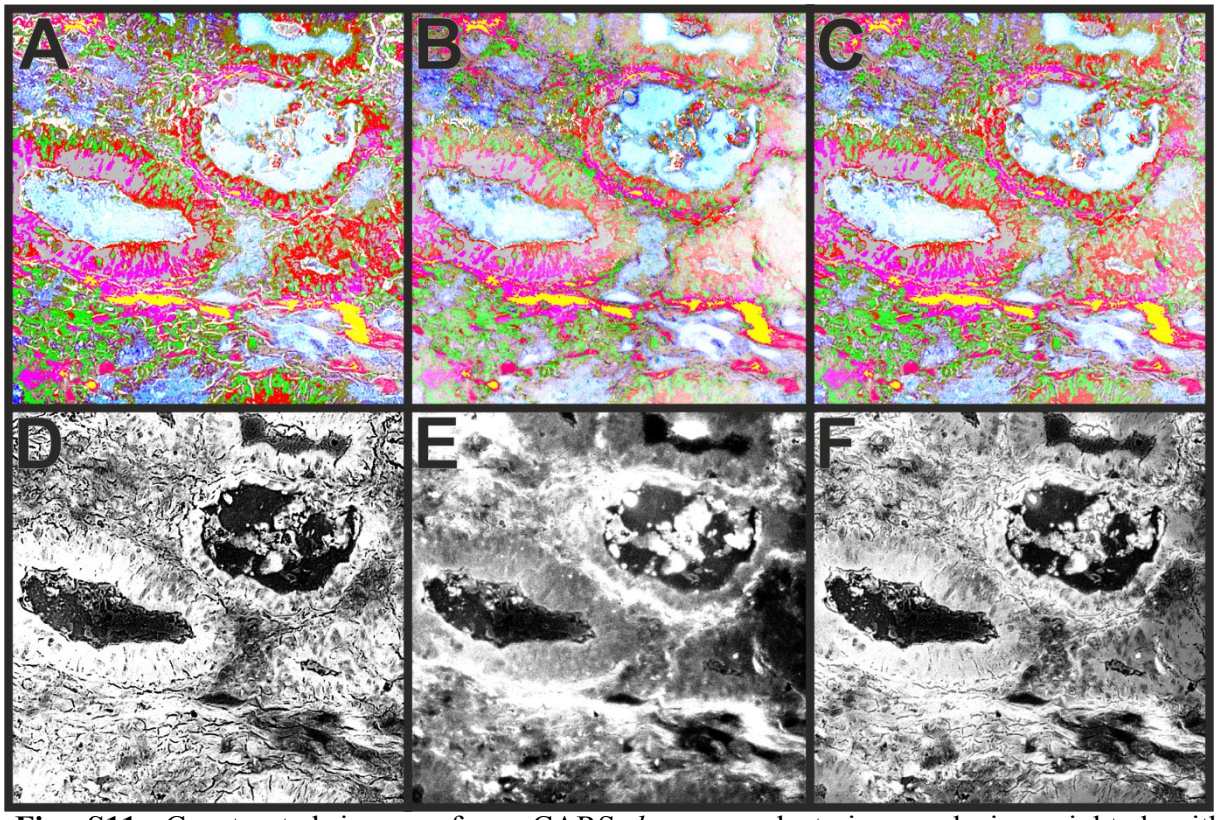

Fig. S11. Constructed images from CARS *k*-means clustering analysis weighted with intensities of CARS (A), SHG (B) and both together (C). The pure intensity images of CARS at 2850 cm^{-1} (D), SHG (E) and combined CARS and SHG (F) are also displayed.