Analyst

Accepted Manuscript

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

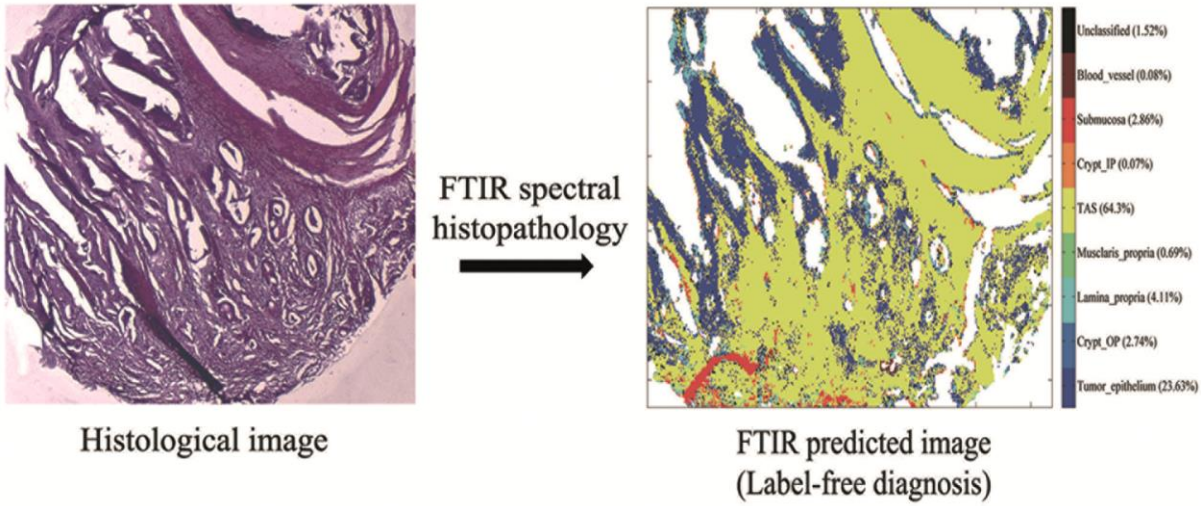
Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst



Automated and label-free colon cancer diagnosis and identification of tumor-associated features using FTIR spectral histopathology directly on paraffinized tissue arrays

Analyst Accepted Manuscript

Title:

Infrared spectral histopathology for cancer diagnosis; a novel approach for automated pattern recognition of colon adenocarcinoma

Authors and affiliations:

Jayakrupakar Nallala[a, b, e], Marie-Danièle Diebold[a, b, c], Cyril Gobinet[a, b], Olivier Bouché[a, b, d], Ganesh Dhruvananda Sockalingum[a, b], Olivier Piot[a, b], Michel Manfait[a, b*].

[a]Université de Reims Champagne-Ardenne, MéDIAN-Biophotonique et Technologies pour la Santé, UFR de Pharmacie, 51 rue Cognacq-Jay, 51096 REIMS cedex, France.

[b]CNRS UMR7369, Matrice Extracellulaire et Dynamique Cellulaire, MEDyC, Reims, France.

[c]CHU de Reims, Laboratoire Central d'Anatomo-Pathologie, 51092 Reims Cedex, France.

[d]CHU de Reims, Service d'Hépato Gastroentérologie et de Cancérologie Digestive, 51092 Reims Cedex, France.

[e]Current Address: University of Exeter, Biomedical Physics, School of Physics, Exeter EX4 4QL, United Kingdom.

Correspondence:

Michel Manfait*

michel.manfait@univ-reims.fr, Tel: +33 32 69 13 57 4, Fax: +33 32 69 13 55 0.

Analyst Accepted Manuscript

ABSTRACT:

Histopathology remains the gold standard method for colon cancer diagnosis. Novel complementary approaches for molecular level diagnosis of the disease are need of the hour. Infrared (IR) imaging could be a promising candidate method as it probes the intrinsic chemical bonds present in a tissue, and provides a "spectral fingerprint" of the biochemical composition. To this end, IR spectral histopathology, which combines IR imaging and data processing techniques, was employed on seventy seven paraffinized colon tissue samples (48 tumoral, 29 non-tumoral) in the form of tissue arrays. To avoid chemical deparaffinization, a digital neutralization of the spectral interferences of paraffin was implemented. Clustering analysis was used to partition the spectra and construct pseudo-colored images, for assigning spectral clusters to various tissue structures (normal epithelium, malignant epithelium, connective tissue etc). Based on the clustering results, linear discriminant analysis was then used to construct a stringent prediction model which was applied on samples without a priori histopathological information. The predicted spectral images not only revealed common features representative of the colonic tissue biochemical make-up, but also highlighted additional features like tumor budding, tumor-stroma association in a label-free manner. This novel approach of IR spectral imaging on paraffinized tissues showed 100 % sensitivity and allowed detection and differentiation of normal and malignant colonic features based purely on their intrinsic biochemical features. This non-destructive methodology combined with multivariate statistical image analysis appears as a promising tool for colon cancer diagnosis and opens the way to the concept of numerical spectral histopathology.

Analyst Accepted Manuscript

1. INTRODUCTION:

Colorectal cancer has one of the highest incidence and mortality among all the cancers affecting both sexes, of which the type adenocarcinoma is the most common.[1] Radiation therapy, chemotherapy and surgical intervention have improved the life expectancy of cancer patients, but the outcome of these methods is dependent upon the stage and the accuracy in diagnosis.[2] Currently different detection and screening methods are employed for colorectal cancers, including fecal occult blood test (FOBT),[3] sigmoidoscopy,[4] colonoscopy,[5] etc. However, the final diagnosis is settled upon the microscopic examination of the symptomatic tissue with the 'gold standard' histopathology in which preferential stains are used to enhance visualization of the tissue morphological alterations. Such alterations (pre-cancerous or cancerous) are the manifestations of the biomolecular changes that have already undergone the provocative changes for malignancy. However, the ongoing state of the tissue molecular changes during the onset or progression of malignancy, without any morphological signatures, poses a challenge for identification. In certain cases, immunohistochemistry (IHC) is used to identify specific proteins of interest which can give a molecular level understanding of the malignant condition. Histopathology requires precise human expertise which limits high-throughput diagnosis. Although, the histopathological diagnosis is based on morphological examination, it has successfully served in cancer diagnosis over several years. Additionally, if it is combined with approaches that could provide complementary biochemical information in a rapid, cost effective manner and reducing human involvement, the efficacy of the histopathological diagnosis can be completed.

In this regard, the optical spectroscopic approach of IR imaging appears as a potential candidate for routine tissue characterization, and has been exploited as a diagnostic tool on various tissues[6-18] which also paved the way to the concept of spectral histopathology.[19-23] IR spectroscopy probes intrinsic chemical bond vibrations of biomolecules and thus provides a biochemical fingerprint of the tissues. Combined with an imaging set-up, spectral images can be obtained rapidly in a label-free manner, in which each pixel element harbors an IR spectrum containing biochemical information at each wavenumber. Such IR images can be exploited using computer based multivariate cluster analysis to generate digitally stained morphological maps of the tissue histology. Since the constituent IR spectra of each digitally stained histological class represent its biochemical signature, such as collagen features in

Analyst Accepted Manuscript

the connective tissue, specific spectral signatures can be identified from different histological classes. Such signatures can be used to train predictive algorithms for identification of unknown tissues in a rapid and user-friendly manner. One of the important possibilities of using this methodology is automation of this protocol which can reduce human involvement and provide an objective biochemical based diagnostic approach.

In this regard, we carried out spectral histopathology based on IR imaging in conjunction with multivariate analysis. The main objectives were to digitally detect and identify malignancy and its associated features on unknown tissues without any chemical staining, constituting an automated diagnosis for colon adenocarcinoma. For this, 77 human colon tissues from normal and moderately differentiated adenocarcinoma were analyzed, in the form of paraffinized tissue arrays that were stabilized in an agarose matrix. The agarose matrix provides stability to the paraffinized tissue cores thereby reducing tissue loss during microtome sectioning, and also facilitates handling of tissue array sections. The tissue arrays are increasingly used in pathological studies since they constitute a large source of information and permit high-throughput analysis for modern histological practices.[24] An innovative process of digital deparaffinization was specially implemented to avoid chemical dewaxing, and also to reduce toxic chemical treatments and time consumption.[20] Then, a prediction model representing the main colon histological classes was constructed and its robustness was evaluated on subsequent number of tissue array cores. Digital annotation using this model facilitated characterization of malignancy, and malignancy associated features such as tumor budding, and tumor-stroma association.

2. MATERIALS AND METHODS:

2.1. Sample preparation:

Seventy seven formalin fixed paraffin embedded (FFPE) colon tissue samples (48 tumoral and 29 non-tumoral) from 32 cancer patients were obtained from the Reims University Hospital, with the approval of the Institutional Review Board. All the tumoral samples were moderately differentiated colon adenocarcinoma with the TNM grade ranging from T3N0M0 to T4N2M0. The sample details are presented in Supplementary Table 1. Several paraffinized tissue arrays that were stabilized in an agarose matrix were manually prepared from these samples. A single sample spot in the tissue array block was

Analyst Accepted Manuscript

approximately 3 mm in diameter. For each tissue array consisting around 12-16 spots, 3 and 10 µm thick sections (adjacent in most cases) were obtained. While the 3 µm section was used by the pathologist for conventional histopathological analysis via hematoxylin, phloxine, and saffron (HPS) staining, the first 10 µm unstained section was used for IR imaging analysis and the second stained section for additional histopathological comparison. The HPS stained sections were chemically deparaffinized while the unstained tissue section for IR imaging was mounted on an IR compatible calcium fluoride (CaF_2) support without any chemical deparaffinization.

2.2. Instrumentation and FTIR data collection:

IR images were acquired, by an IR imaging system (Spotlight 300, Perkin Elmer, Courtaboeuf, France) equipped with liquid nitrogen-cooled 16-element MCT detector, at 6.25 x 6.25 μm^2 pixel size, and 4 cm^{-1} spectral resolution averaged to 16 scans, in the mid-IR range of 750 to 4000 cm^{-1}. The system was continuously purged with dry air. The background spectrum from the CaF_2 support was recorded each time prior to image acquisition, using the same parameters as that of the IR image. The methodology for FTIR spectral imaging of tissue arrays is represented in Supplementary Figure 1. A total of 8 141 566 IR spectra were recorded from 77 images at an average of 105 734 per image owing to the large size of the tissue array spots, and the high spatial resolution selected for imaging.

2.3. Data pre-processing:

Raw IR data was corrected from various spectral interferences. An atmospheric correction was performed to remove contribution from water vapour and CO_2 by the built-in Perkin Elmer Spotlight software and further processing was carried out using programmes written in Matlab 7.2 (The Mathworks, Natick, MA). The spectra were reduced to the IR absorption range of 900-1800 cm^{-1} that contains several informative biochemical vibrations[25,26] as far as the tissue features are considered. Neutralization of paraffin and agarose contributions was carried out using a modified Extended Multiplicative Signal Correction (EMSC). In addition to paraffin model, a correction model for agarose was inserted into the EMSC algorithm.[20, 27] As detailed, the EMSC algorithm neutralizes the influence of their spectral variabilities by a modeling procedure rather than directly subtracting the spectral signatures of paraffin and agarose.[20] Therefore it is

Analyst Accepted Manuscript

important to note that paraffin and agarose features are not removed, but their spectral variabilities are neutralized. Therefore, in the image analysis only the spectral variabilities originating from the biochemical features are taken into account rather than those from physical features of paraffin and agarose which are no longer apparent. Furthermore, EMSC has been adapted to address the inter-patient variability using a single target spectrum (also called the model spectrum or the reference spectrum) for the all the tissue samples. Using the same target spectrum for all the samples has been an important criterion in our application in order to correct all the spectra from the same amount of baseline, paraffin and agarose, while keeping the biochemical information specific to each sample. If a different target spectrum is used for each sample (e.g. for normal and cancerous tissue), the corrected spectra of each sample will have different shapes mainly because of the different baselines and/or paraffin signals and/or agarose signals composing each of the target spectra, and not because of the biochemical differences between normal and cancer tissues.[28-29] The IR spectra were also corrected for baseline and then normalized using the same algorithm. Outliers (N=3 335 684 spectra) in the form of paraffin and agarose spectra, and spectra with poor signal-noise ratio were eliminated from the analysis and were depicted as white pixels in all the IR images.

2.4. Data processing:

The pre-processed data (N=4 805 882 spectra) was subjected to multivariate statistical prediction analysis. For this, spectral data from the non-tumoral and the tumoral samples was separated into a training group (N=9, Supplementary Table 1, sample # TG), and a validation group (N=68). While the training group, representing the IR spectral signatures indicative of malignancy and other histological structures, was used for construction of a prediction model based on linear discriminant analysis (LDA), the validation group (external validation) was used for validating the model on unknown samples for automatic recognition of tissue features, to enable identification of malignancy. LDA is a multivariate supervised statistical technique that aims at maximizing the between-class variance and minimizing the within-class variance and has been exploited in various studies.[26-27, 30-31]

Analyst Accepted Manuscript

2.4.1. Cluster analysis for LDA training:

The huge number of IR spectra from each image corresponding to the training group was subjected to unsupervised k-means clustering method owing to its capability of rapid and huge data clustering.[32] This method iteratively partitions the spectra into different clusters based on the spectral signatures from the intrinsic biochemical composition of the tissue. Therefore, spectra with similar biochemical characteristics group into the same cluster. In k-means clustering, each spectrum belongs to a unique cluster and can thus be represented by one color. K-means clustering performed using defined cluster numbers resulted in the construction of digital color-coded images. These were then compared to adjacent HPS stained sections to annotate by an expert pathologist, each spectral cluster to the tissue structural feature that it corresponds to. The spectral distance between different k-means clusters was visualized in a dendrogram obtained by hierarchical clustering analysis using Ward's linkage algorithm.

2.4.2. Prediction model:

The initially k-means clustered and annotated spectra were used as inputs for the LDA model. Training group spectra (Supplementary Table 1 # TG) from 9 samples across 6 different patients were considered for the model, to take into account the inter-patient variability. The prediction model consisted of 8 classes with different number of spectra, representing various histological features of non-tumoral and tumoral tissues: the normal epithelium defined by the crypt inner-part (Crypt-IP) (N = 8377) and the crypt outer-part (Crypt-OP) (N = 3567), the lamina propria (N = 14 106), the submucosa (N = 3964), the tumor epithelium (N = 35 083), the tumor-associated stroma (N = 16 409), the blood vessel (N = 782) and the muscularis propria (N = 4514). These spectra (N=86 802) constituting one-third of the spectra from each class were used to train the model and the other two-thirds were used for an internal validation to optimize the model. The prediction model was then applied in an external validation on different unknown samples, the spectra from which were secluded from the model, to evaluate its robustness. The external validation consisted of 68 samples encompassing a large scale spectral data base of 4 130 879 spectra. It has to be noted that if only the number of patients used in the external validation was to be considered (instead of the number of samples from all the patients as is the case in this study) the external validation group consisted of 26 patients, since several samples were obtained from a single patient

Analyst Accepted Manuscript

(Supplementary Table 1). The predictions were carried out in the IR spectral range of 1080 cm^{-1} - 1300 cm^{-1}, at a posterior probability of 0.5, wherein for each pixel a probability of belonging to each class is calculated, and the pixel showing the highest probability is assigned to a class. If the highest probably is inferior to the posterior probability of 0.5, the pixel is termed as unclassified and is not attributed to any class. The final model based diagnosis of cancer by the presence of tumor pixels was confirmed by the presence of tumoral areas in the corresponding region of the HPS stained tissue, using the gold-standard histopathological validation. Validation based on the presence of certain number of pixels (tumor pixels) was not considered as a dedicated approach in this study where heterogeneous tumoral tissue types are considered which contain varying amount of tumoral cells.

2.4.3. Spectral information to biochemical information (spectral analysis):

Since the spectral signatures are based on the biochemical properties of the tissue features, it was attempted to characterize the biochemical alterations characteristic of malignancy and the relationship of malignant tissue with the surrounding stroma. For this, the Mann-Whitney *U* test was applied to compare spectra from selected cluster groups used in the prediction model training in order to identify the most discriminant wavenumbers.

2.5. Immunohistochemistry (IHC):

IHC was used as a complementary tool (on adjacent sections) to enhance visibility of tumor budding (Anti-Human Cytokeratins-large spectrum Monoclonal Antibody, Clone KL 1, dilution 1/50, Immunotech, France) and to precise the nature of the inflammatory cells: T-lymphocytes (CD3 Rabbit anti-Human Polyclonal Antibody, dilution 1/200, Dako, France), and B-lymphocytes (CD20 Mouse antibody, clone L6 mouse, dilution 1/400, Dako, France), in order to validate some of the important observations detected by IR spectral imaging. This was performed using the fully automated IHC staining protocol (XT ultraView DAB v3).

Analyst Accepted Manuscript

3. RESULTS:

3.1. Cluster analysis:

K-means clustering was used to identify the spectral signatures characteristic of the main histological features of the non-tumoral and the tumoral colon tissues, which permitted construction of digitally stained images. For the non-tumoral as well as the tumoral tissues, this approach permitted to identify, and to recover automatically the important histological components in comparison to the adjacent HPS stained images as shown in the figure 1 (Supplementary Table 1, sample # 1D and 12C). As an example, for the non-tumoral colon tissue (figure 1A) 8 clusters permitted the observation of the important histological structures representing the colon tissue organization. They included the colon mucosa constituted by well-differentiated crypts (cluster 8 - inner part and cluster 6 - outer part); and the lamina propria (cluster 1), the supportive loose connective tissue in which the crypts are organized. The residual mucin (cluster 2) was observed to be localized within the crypt lumen while a small amount was seen secreted outside. The submucosa, attributed to clusters 4, 5 and 7 was distinguished effectively from the lamina propria by the clustering method. Finally cluster 3 appeared to represent the blood vessels. On the contrary, in the typical adenocarcinomatous tissue (figure 1B), the only important histological classes retrieved were the tumor epithelium (cluster 1) and its associated stroma in the tumor vicinity (cluster 6). Most of the other clusters represented the fibrous stromal tissue. The corresponding dendrogram showed the close spectral nature of the tumor associated stroma to its tumor where they are very closely grouped (clusters 1 and 6) while the stroma that is not in direct contact with the tumor epithelium appear more distant. A total of 11 clusters were required to identify these features. In both cases, considering the overall colon tissue organization, increasing the number of clusters did not add any further retrievable histological information. The k-means clustering is an efficient method to identify IR spectral markers specific to different histological components of non-tumoral and tumoral colon tissues. On the basis of these spectral signatures, the diagnostic potential of IR spectral imaging has been evaluated using a LDA based prediction model as schematically represented in Supplementary Figure 2.

3.2. Optimization of the prediction model - internal validation group:

Analyst Accepted Manuscript

The LDA based prediction model developed from 9 samples (6 patients) with 8 different classes comprising a total of 86802 spectra was trained, and tested in an internal validation. The sensitivity of the prediction model in the internal validation can be evaluated from the confusion matrix which shows the agreement between the histopathological class annotation (real class) and the IR spectral prediction (predicted class) (Table 1). Different spectral regions were tested and the highest sensitivity (average 89.38%) was obtained for the region between 1080 cm^{-1} to 1300 cm^{-1}. It has to be noted that for the class tumor epithelium a specificity of 96.4 % was reached, and showed no confusion with the class normal epithelium (comprising crypt inner and outer parts).

3.3. Tumor detection and tissue characterization in unknown samples - external validation group:

The external validation was performed on the remaining 68 blind samples involving a large scale spectral bank of 4 130 879 spectra and showed 100 % sensitivity for the tumor class. Along with tumor class, other histological classes were also identified with high correlation to the conventional histology.

A representative demonstration of prediction on unknown non-tumoral and tumoral samples is shown in figure 2 (Supplementary Table 1, sample # 14D and 7C). The figure 2A histologically corresponded to a non-tumoral colon tissue in which the prediction model correctly identified its characteristic features with similar morphological attributes to that of the histological image. Counterpart to the normal tissue, histologically the figure 2B corresponded to a typical moderately differentiated colon adenocarcinoma. In this, the spectral characteristics of the normal mucosa were absent and the only distinguished ones were malignant epithelial component with its associated stroma. Additionally, identification of features difficult to discern using conventional techniques, such as tumor budding was facilitated.

3.4. Detection and characterization of malignancy associated features:

3.4.1. Tumor budding:

Budding is characterized by small clusters of isolated tumor cells which become detached from the neoplastic epithelium and migrate into the stroma, and is an indication of high tumor invasiveness in colorectal cancers. Although this morphological phenomenon is detectable in conventional histopathology at high power magnification, IHC may be employed for better visualization. The IR prediction model was

Analyst Accepted Manuscript

able to clearly identify this tumor particularity even in the presence of abundant stroma as shown in the figure 3, (Supplementary Table 1, sample # 9B). In the same tumoral sample, along with the malignant epithelium, there was presence of some normal epithelial component together with normal connective tissue, and all these features were identified by the prediction model. Importantly, both the malignant and the non-malignant epithelial cells were selectively stained and discriminated using a specific color-code. The positive staining of the epithelial cells can be seen in the IHC image (see figure 3, right panel). Another tissue section obtained from different position (Supplementary Table 1, sample # 9A) of the same tumor also showed tumor budding in a stroma dominant environment, and each time it was identified by the prediction model, which was later confirmed by IHC studies (Supplementary Figure 3).

3.4.2. Tumor stroma association:

The tumor-stroma association was also reported using IR spectral imaging. The confusion matrix (table 1) highlighted the spectral proximity of tumor and its associated stroma in which, indeed 16.3 % of tumor associated stroma pixels were classified in the tumor class. Complementarily, in the predicted images these two classes appeared in geographic proximity (figure 4) (Supplementary Table 1, sample # 11B). In the same image, distinction between the tumor associated stroma and the normal connective tissue corresponding to the submucosa was attained, while in the histological stained section, this was indistinguishable. The above mentioned tumor-stroma features were also observed in the other tumoral samples (Supplementary Table 1, sample # 11A, 11C, 12A, 13A, and 15A) as shown in Supplementary Figure 4 including the cases of budding (fig 3).

3.5. Vibrational analysis of spectroscopic markers:

In this study, the k-means clustering was performed using the IR spectral range of 900 cm^{-1} - 1800 cm^{-1} that enabled identification and attribution of the important colon histological classes. For unknown sample prediction, this zone was narrowed down to 1080 cm^{-1} to 1300 cm^{-1} harboring some of the important biomolecular vibrational modes implicated in colon cancers, and which showed the best prediction outcome for all the classes together. As shown in figure 5, the most discriminant wavenumbers within this zone were identified by the Mann-Whitney *U* test performed on the individual spectra and represented on

Analyst Accepted Manuscript

the average spectra for the following pair-wise comparisons: normal epithelium with malignant epithelium (adenocarcinoma) for understanding the molecular alterations characteristic of malignancy; adenocarcinoma with its associated stroma to understand the tumor induced alterations in the stromal tissue; and the normal connective tissue with the tumor associated stroma. From the discriminant wavenumbers identified for all comparisons, a tentative correlation of IR vibrations to the biomolecular information was attempted as shown in Supplementary Table 2. Importantly, comparing the normal epithelium with the tumoral epithelium, the main differences in the IR peaks were attributed to symmetric and asymmetric PO_2^- vibrations of the nucleic acids that demonstrated relatively higher intensities in normal than the tumoral tissues. Similarly, the C-O stretching vibration corresponding to carbohydrates was relatively more intense in normal than the tumoral tissues. At the same time the hydrogen bonded C-O groups of proteins in the normal epithelium was observed to be decreased in the tumoral epithelium, while the opposite tendency was observed for the non-hydrogen bonded C-O groups of proteins. Secondly, when comparing adenocarcinoma with tumor associated stroma, and tumor associated stroma with connective tissue, the discriminating spectral features appeared to be contributed principally from collagen features.

4. DISCUSSION:

Spectral histopathology based on IR imaging has been carried out to develop an innovative label-free diagnostic methodology directly on FFPE tissue arrays embedded in an agarose matrix without any chemical pre-treatments. EMSC that has been initially developed to separate light scattering effects from light absorbance effects, has also been used for accomplishing neutralization of paraffin contributions in IR spectral analysis.[9,27,33-35] In this study, both paraffin and agarose interferences on the IR spectral images have been neutralized digitally without the use of any chemicals, using an improved EMSC algorithm. One of the important advantages of using of paraffinized tissues stabilized in an agarose matrix is that the scattering effects such as Mie scattering due to the differences in the refractive indices of the media are reduced by index matching. Additionally, resonant Mie-scattering that is related to a physical phenomenon and which can cause peak shape distortion and peak shift (e.g., the amide I peak), resulting in unreliable chemical interpretation is also reduced.[36]

Analyst Accepted Manuscript

4.1. Clustering:

K-means clustering provided a rapid way to classify the IR spectral images into their constituent histological classes in comparison to the chemically stained conventional images. While the non-tumoral colon tissues were characterized by well-differentiated architecture with both inner and the outer cryptal parts clearly distinguishable together with the connective tissue, the malignant tissues which were all of the advanced colon cancer types, were characterized by the loss of differentiation of the normal colon glands with no visible lumen; and presence of stromal tissue. The digital staining of each k-means cluster formed the basis for spectral marker assignment comprising the malignant colon characteristics, along with the normal tissue features, at different organizational levels of the colon wall. Based on this spectral database from as little as 12 % of the samples, a prediction model was trained for automatic detection of malignancy in unknown specimens independently of conventional histopathology.

4.2. Prediction:

Some of the earlier IR imaging studies have tested prediction algorithms on different tissue types.[25,26] However, the number of spectra used for constructing the model was limited compromising the robustness of the model. In our study, the relatively high resolution image acquisition parameters applied to tissue arrays (3 mm diameter) constituted a huge bank of 86 802 spectra in the prediction model, representative of the biochemical signatures of distinct colon structures, making it highly robust. Only one such IR imaging study on prostate tissues has used such a robust model for prediction on unknown tissues.[37] In this study, 8 classes were included that described the colon tissue organization in non-tumoral and tumoral samples. Even with a high sensitivity of the model (such as in the case of tumor budding), some of these histological structures may share certain similar molecular constituents with other histological classes present in the model (tumor and tumor associated stroma), or not present in the model (muscularis mucosa and tumor associated stroma). The spectral proximity arising from this leads to misclassification between such classes as shown in the Supplementary Figure 5, concerning the muscularis mucosa (visible in the HPS image) which is identified as tumor associated stroma (Supplementary Table 1, sample # 27). It has to be noted that there was no class for the muscularis mucosa in the model. This attribution can be presumed to have arisen from the residual normal

Analyst Accepted Manuscript

muscularis mucosa signatures present in the tumor associated stroma from which the corresponding class was constructed in the prediction model. This prediction error appeared predominantly in non-tumoral samples where there is an intact muscularis mucosa. Despite these misclassifications, an overall high correlation between the predicted spectral classes and the corresponding histological structures is observed in the confusion matrix.

4.3. External validation:

The remaining 88 % of IR spectral images were identified by the prediction model without any a priori knowledge on their histopathology (external validation). These blind samples constituted a huge number of 4 130 879 spectra that were scanned and annotated by the automated computer trained prediction algorithm. The diagnosis was confirmed by an expert pathologist by using the conventional histological images based on which a 100 % accuracy of the prediction model was obtained for tumor diagnosis. This high sensitivity after scanning such a huge number of unknown spectra signifies the potential of the current methodology as a diagnostic tool. The prediction analysis also facilitated simultaneously some important malignancy associated features.

4.4. Tumor budding:

The phenomenon of tumor budding is of crucial clinical importance in colorectal cancers since it has been shown to be a strong adverse prognostic marker.[38] As such, studies have correlated its occurrence with aggressiveness and lymph node metastasis.[39] In this study, the prediction model facilitated the identification of tumor budding in a stroma-dominant environment in an automated manner. This rapid and selective detection of small clusters of isolated tumor cells in an abundant stroma environment demonstrates the sensitivity and the applicability of the methodology avoiding the need of any histological or immunological markers. This envisages an important prospect since the tumor de-differentiation in the form of budding is being acknowledged as a key component in the metastatic process even in well- and moderately differentiated tumors.[40,41] At the same time, the color code based selective staining of the epithelial counter parts in the same tissue shows the discriminatory ability and the biomolecular specificity of this methodology.

Analyst Accepted Manuscript

4.5. Spectral Analysis:

The IR spectral region from 1000 cm^{-1} to 1300 cm^{-1} has been reported to carry important biochemical vibrations implicated in colon cancers and have been used for differentiating the malignant tissues from their normal counterparts.[42,43] In this study, the most discriminant spectral wavenumbers were associated with relatively decreased intensities of symmetric and asymmetric PO_2^- vibrations of the nucleic acids in the tumoral epithelium when compared to the non-tumoral tissues. On contrary to the expected increased nucleic acid intensities as shown in several studies, these spectral changes corresponding to the biochemical alterations corroborate with some of the previous studies on colon cancers where the nucleic acid intensities were shown to be reduced in malignant conditions.[32,44] It may be likely that the spectral changes involving nucleic acids are small in moderately differentiated tumors when compared to normal colon epithelial cells which themselves are highly proliferative in nature. One study has stated that decreased phosphate content in malignant colon tissues may be due to decrease in carbohydrate content,[45] which in our study was also indicated by the relatively less intense C-O stretching vibration corresponding to carbohydrates in the tumoral tissue than the normal. At the same time, the relative intensities of H-bonded C-O vibrations of proteins were observed to be more pronounced in the normal epithelium than the tumoral, while the non-H-bonded C-O bond vibrations were more pronounced in the tumor. These changes may be indicative of the molecular alterations associated with the amino acid side chains concerning tyrosine, serine and threonine.[2,32,45,46] The molecular changes involving adenocarcinoma and tumor-associated stroma, and tumor associated stroma with connective tissue appear principally due to collagen features.

4.6. Tissue inflammation influences the model specificity:

In 12 out of 29 samples histologically described as non-tumoral (Supplementary Table 1, sample # LF); tumoral characteristics (over 4 % of pixels) were observed either regionally clustered or dispersed in the lamina propria, showing a specificity of 59%. The HPS images gave insight into the regionally clustered tumor pixels as corresponding to lymphoid follicles in the colon tissue. These structures showed spectral signatures close to the tumor group relative to the other classes. However, the tumor pixels dispersed in the lamina propria could not be accounted for as no visible correspondence between them and any

Analyst Accepted Manuscript

histological feature could be found in the HPS images. Since these tissues showed high inflammatory infiltration, immuno-staining for T-lymphocytes (CD 3), B-lymphocytes (CD 20) and macrophages (KP 1) was performed to verify if the dispersed pixels corresponded to the inflammatory cells. The positive staining indicated that these pixels indeed corresponded mainly to interstitial T-lymphocytes as representatively shown in the figure 6A (Supplementary Table 1, sample # 32). In parallel, the B-lymphocytes were seen assembled in lymph follicles. Non-tumoral tissues without any marked inflammation as confirmed by the IHC showed no tumor pixels in the IR spectral images (figure 6B) (Supplementary Table 1, sample # 31). Since the model did not take into account inflammatory conditions (because of the tissue complexity arising from polymorphisms of the inflammatory infiltrates in colon cancers: polymorph predominant, mononuclear predominant, mixed or rich in lymphoid follicles, and the difficulty to have a representative spectral signature), these features were attributed to the spectrally nearest class which turned out to be the tumor class.

A recent IR imaging study on cervical cancer tissues also quoted the influence of inflammatory signatures on the prediction model sensitivity and specificity.[23] To have a broader insight into this aspect, we further looked at the spectral class attribution threshold for the tumor class. It turned out that for the attribution of spectra to tumor class, majority of the spectra corresponding to the inflammatory signatures have lesser threshold values compared to the tumor in which the majority of the spectra have the highest posterior probability values (Supplementary Figure 6). Altogether, the IR signatures from the inflammatory regions appeared to class spectrally closer to tumor than other classes of the prediction model indicating an intermediate stage between normal and malignant condition, as was shown in an earlier study.[47]

The current work of IR spectral imaging on colon tissues provides automated diagnosis of malignancy on unknown samples. Various diagnostic features associated with malignancy which provides complementary information are also characterized. Important features such as tumor budding, tumor-stroma association are dealt with in a non-destructive and label-free manner. The analysis of such a large spectral database makes the study all the more representative. All these features have never been dealt together in colon cancer diagnosis using IR spectral imaging of paraffinized tissues in any of the previous studies. IR spectral imaging presents an optimistic overture for cancer knowledge in modern histopathology.

Analyst Accepted Manuscript

The current prediction model representing the important histological features of a colon tissue certainly holds aspects for amelioration. The spectral attribution identified the inflammatory signatures classed close to the tumor. Since these specific biochemical signatures were picked up by the model, the inflammatory infiltration, which pose risk of developing into cancers, could be incorporated into the model for an automated evaluation and direct diagnostic approach for inflammatory diseases. In the same manner, classes' specific to early neoplastic condition such as dysplasia could be incorporated into the model and their spectral attribution thresholds compared to that of adenocarcinoma and normal epithelium. This can potentially provide insights, into spectral alterations in early neoplastic conditions and therefore, for early diagnosis of cancers. Aspects like genotype specific tumoral signatures and their treatment response sensibility unknown till now could open a new additional classification. Further, an automated quantification can be achieved for features like amount of tumor presence, or the amount of tumor budding, only limit being the use of adjacent tissue sections which may present slight variations from the reference tissue.

5. CONCLUSIONS:

The IR spectral imaging combined with multivariate statistical analyses appears as an optimistic diagnostic approach for colon cancers in complement to conventional histopathology. This innovative imaging approach enabled direct analysis of paraffinized tissue arrays and, via the employment of mathematical deparaffinization the need for chemical pretreatments was reduced. The prediction model permitted identification of unknown samples with a very high sensitivity, while the false positive prediction in the non-tumoral samples has put forth the influence of the inflammatory component. This very large scale spectral data base analyzed both in terms of training and validation shows the potentials of the IR spectral imaging methodology for automated diagnostic purposes. Moreover, it eliminated the need for sample staining and a priori knowledge of the sample to be analyzed. These optimistic results open a new way for developing spectral biomarkers and libraries which could be used, in complement to conventional histopathology, for early diagnosis, and also potentially for prognosis and theranostics of cancers.

Analyst Accepted Manuscript

ACKNOWLEDGEMENTS:

This study was supported by a grant of Institut National du Cancer (INCa) and Canceropôle Grand Est. We would like to thank Ligue contre le Cancer, Conférence de Coordination Interrégionale du Grand-Est, and CNRS Projets Exploratoires Pluridisciplinaires, for financial support. Plateforme IBiSA "Imagerie Cellulaire et Tissulaire", and the Tumorotheque, Champagne-Ardenne is also acknowledged. NJ is a recipient of doctoral fellowship from the Région Champagne-Ardenne.

REFERENCES:

1. J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers and D. M. Parkin, *Int. J. Cancer*, 2010, **127**, 2893-2917.

2. C. Conti, P. Ferraris, E. Giorgini, C. Rubini, S. Sabbatini, G. Tosi, J. Anastassopoulou, P. Arapantoni, E. Boukaki, S. Konstadoudakis, T. Theophanides and C. Valavanis, *J. Mol. Struct*, 2008, **881**, 46-51.

3. H. Miyoshi, M. Oka, K. Sugi, O. Saitoh, K. Katsu and K. Uchida, *Intern. Med*, 2000, **39**, 701-706.

4. T. J. Zuber, *Am. Fam. Physician*, 2001, **63**, 1375-1380.

5. D. K. Rex, *Colon tumors and colonoscopy*, 2000, **32**, 874-883.

6. A. Tfayli, O. Piot, A. Durlach, P. Bernard and M. Manfait, *Biochim. Biophys. Acta*, 2005, **1724**, 262-269.

7. H. Fabian, N. A. Thi, M. Eiden, P. Lasch, J. Schmitt and D. Naumann, *Biochim. Biophys. Acta*, 2006, **1758**, 874-882.

Analyst Accepted Manuscript

19

8. W. Steller, J. Einenkel, L. C. Horn, U. D. Braumann, H. Binder, R. Salzer and C. Krafft, Anal. *Bioanal. Che,* 2006, **384**, 45-154.

9. A. Travo, O. Piot, R. Wolthuis, C. Gobinet, M. Manfait, J. Bara, M. E. Forgue-Lafitte and P. Jeannesson, *Histopathology,* 2010, **56**, 921-931.

10. M. J. Nasse, M. J. Walsh, E. C. Mattson, R. Reininger, A. Kajdacsy-Balla, V. Macias, R. Bhargava and C. Hirschmugl, *Nat. Methods*, 2011, **8**, 413-416.

11. M. J. German, A. Hammiche, N. Ragavan, M. Tobin, L. J. Cooper, S. S. Matanhelia, A. C. Hindley, C. M. Nicholson, N. J. Fullwood, H. M. Pollock and F. L. Martin, *Biophys. J*, 2006, **90**, 3783-3795.

12. K. Yano, S. Ohoshima, Y. Gotou, K. Kumaido, T. Moriguchi and H. Katayama, *Anal. Biochem*, 2000, **287**, 218-225.

13. T. D. Wang, G. Triadafilopoulos, J. M. Crawford, L. R. Dixon, T. Bhandari, P. Sahbaie, S. Friedland, R. Soetikno and C. H. Contag, *Proc. Natl. Acad. Sci. USA*, 2007, **104**, 15864-15869.

14. X. Zhang, Y. Xu, Y. Zhang, L. Wang, C. Hou, X. Zhou, X. Ling and Z. Xu, *J. Surg. Res*, 2011, **171**, 650-6.

15. C. Krafft, S. B. Sobottka, K. D. Geiger, G. Schackert and R. Salzer, *Anal. Bioanal. Chem*, 2007, **387**, 1669-1677.

16. J. Nallala, O Piot, M. D. Diebold, C. Gobinet, O. Bouche´, M. Manfait and G. D. Sockalingum, *Cytometry Part A,* 2013, **83**, 294-300.

17. J. T. Kwak, S. M. Hewitt, S. Sinha, R. Bhargava, *BMC Cancer,* 2011, **11,** 62.

Analyst Accepted Manuscript

18. Rohit Bhargava. *Anal Bioanal Chem*, 2007, **389**, 1155-1169.

19. B. Bird, M. Miljkovic, S. Remiszewski, A. Akalin, M. Kon and M. Diem, *Lab. Invest*, 2012, 1-16.

20. J. Nallala, C. Gobinet, M. D. Diebold, V. Untereiner, O. Bouché, M. Manfait, G. D. Sockalingum and O. Piot, *J.* Biomed. Opt, 2012, **17**, 1-12.

21. M. J. Walsh, S. E. Holtona, A. Kajdacsy-Ballab and R. Bhargava, *Vibrational Spectroscopy,* 2012, **60**, 23-28.

22. J. D. Pallua, C. Pezzei, B. Zelger, G. Schaefer, L.K. Bittner, V. A. Huck-Pezzei, S. A. Schoenbichler, H. Hahn, A. Kloss-Brandstaetter, F. Kloss, G. K. Bonn and C. W. Huck, *Analyst,* 2012, **137**, 3965-3974.

23. J. Einenkel, U. D. Braumann, W. Steller, H. Binder and L. Horn, *J. Histopathology*, 2012, **60**, 1084-1098.

24. J. Kononen, L. Bubendorf, A. Kallionimeni, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M. Mihatsch, G. Sauter and O. Kallioniemi, *Nat. Med*, 1998, **4**, 844-847.

25. M. Khanmohammadi, A. B. Garmarudi, K. Ghasemi, H. K. Jaliseh and A. Kaviani, *Med. Oncol*, 2009, **26**, 292-297.

26. M. Khanmohammadi, A. B. Garmarudi, S. Samani, K. Ghasemi and A. Ashuri, *Pathol. Oncol. Res*, 2010, **17**, 435-441.

27. E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard and M. Manfait, *Analyst,* 2008, **133**, 197-205.

Analyst Accepted Manuscript

28. A. Kohler, N. K. Afseth and H. Martens, in Applications of Vibrational Spectroscopy in Food Science, ed. E. Li-Chan, P. R. Griffiths and J. M. Chalmers, John Wiley & Sons, Ltd. 2010, vol. 1, pp. 89-97.

29. N. K Afseth and A. Kohler. *Chemometrics and Intelligent Laboratory Systems*, 2012, **117**, 92-97.

30. M. Khanmohammadi, M. A. Ansari, A. B. Garmarudi, G. Hassanzadeh and G. Garoosi, *Cancer. Invest*, 2007, **25**, 397-404.

31. E. Gazi, M. Baker, J. Dwyer, N. P. Lockyer, P. Gardner, J. H. Shanks, R. S. Reeve, C. A. Hart, N. W. Clarke and M. D. Brown, *Eur. J. Urol*, 2006, **50**, 750-761.

32. P. Lasch, W. Haensch, D. Naumann and M. Diem, *Biochim. Biophys. Acta*, 2004, **1688**, 76-186.

33. H. Martens, J. P. Nielsen and S. B. Engelsen, *Anal. Chem*, 2003, **75**, 394-404.

34. A. Kohler, C. Kirschner and A. Oust, H. Martens, *Appl. Spectrosc*, 2005, **59**, 707-716.

35. D. Sebiskveradze, V. Vrabie, C. Gobinet, A. Durlach, P. Bernard, E. Ly, M. Manfait, P. Jeannesson and O. Piot, *Lab. Invest,* 2011, **91**, 799-811.

36. P. Bassan, A. Sachdeva, A. Kohler, C. Hughes, A. Henderson, J. Boyle, J. H. Shanks, M. Brown, N. W. Clarke and P. Gardner. *Analyst*, 2012, **137**, 1370-1377.

37. D. C. Fernandez, R. Bhargava, S. M. Hewitt and I. W. Levin, *Nat. Biotechnol*, 2005, **23**, 469-474.

38. L. M. Wang, D. Kevans, H. Mulcahy, J.O. Sullivan, D. Fennelly, J. Hyland, D. Donoghue and K. Sheahan*, Am. J. Surg. Pathol*, 2009, **33**, 134-141.

Analyst Accepted Manuscript

39. H. Kanazawa, H. Mitomi, Y. Nishiyama, I. Kishimoto, N. Fukui, T. Nakamura and M. Watanabe, *Colorectal. Dis*, 2008, **10**, 41-47.

40. F. Prall, *Histopathology,* 2007, **50**, 151-162.

41. H. Gabbert, *Cancer. Metastasis*. Rev, 1985, **4**, 293-309.

42. R. K. Sahu, S. Argov, S. Walfisch, E. Bogomolny, R. Moreh and S. Mordechai, *Analyst,* 2010, **135**, 538-544.

43. V. K. Katukuri, J. Hargrove, S. J. Miller, K. Rahal, J. Y. Kao, R. Wolters, E. M. Zimmermann and T. D. Wang, *Biomed. Opt. Express,* 2010, **1**, 1014-1025.

44. B. Rigas, S. Morgello, I. S. Goldman and P. T. Wong, *Proc. Natl. Acad. Sci. USA,* 1990, **87**, 8140-8144.

45. S. Argov, J. Ramesh, A. Salman, I. Sinelnikov, J. Goldstein, H. Guterman and S. J. Mordechai, *J. Biomed. Opt,* 2002, **7**, 1-7.

46. P. Wong and H. M. Yazdi, *Appl. Spectrosc,* 1993, **44**, 1830-1836.

47. S. Argov, R. K. Sahu, E. Bernshtain, A. Salman, G. Shohat, U. Zelig and S. Mordechai, *Biopolymers*, 2004, **75**, 384-392.

48. S. L. Patrick, T. T. Wong and H. M. Yazdi, *Appl. Spectrosc*, 1993, **47**, 1830-1836.

49. L. Chen, H. Y. N. Holman, H. Zhao, H. A. Bechtel, M. C. Martin, C. Wu and S. Chu, *Anal. Chem*, 2012, **84**, 4118-4125.

Analyst Accepted Manuscript

23

Figures:

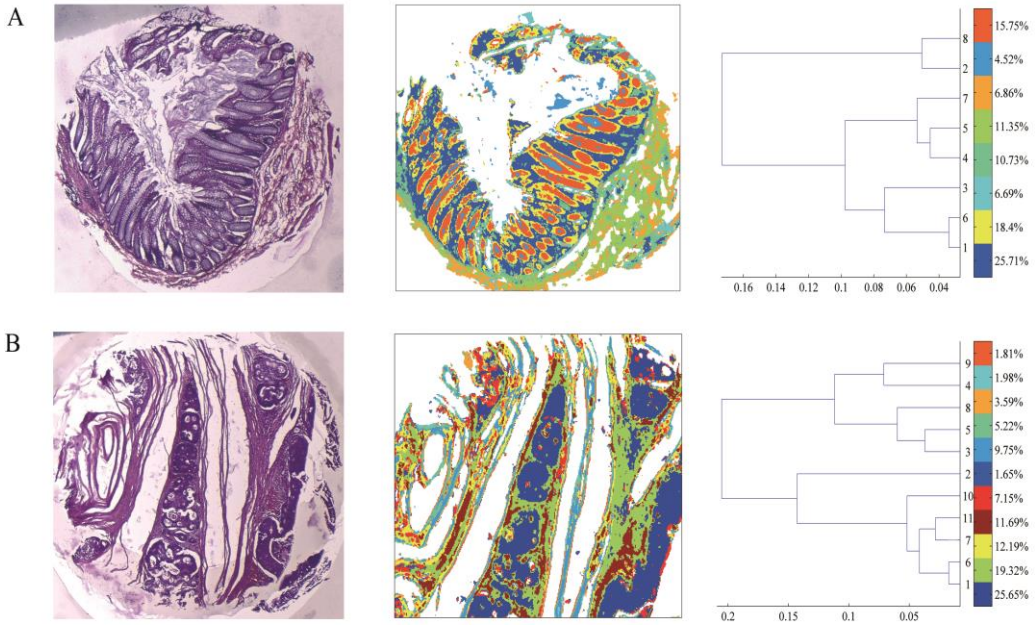


Figure 1: K-means clustering and digital staining of FTIR spectral images with random pseudo-colors. Left panel: HPS stained colon tissues (Supplementary Table S1, sample # 1D and 12C); Middle panel: K-means clustering and digital staining of FTIR spectral images with random pseudo-colors; Right panel: Dendrograms corresponding to the respective cluster images. A is a non-tumoral colonic tissue partitioned using 8 clusters representing the major normal colonic tissue features. The cluster representation is as follows: Cluster 1 - lamina propria, cluster 2 - mucous, clusters 4, 5 and 7 - submucosa, cluster 6 - crypt (outer part-OP), cluster 8 - crypt (inner part IP) and cluster 3 - undefined tissue. B is a moderately differentiated colonic adenocarcinoma partitioned using 11 clusters. The important histological classes are cluster 1 - tumor, clusters 6, 7, and 11 - tumor-associated stroma. Remaining clusters were attributed to the fibrous stroma. The HPS images are at 5X magnification.

Analyst Accepted Manuscript

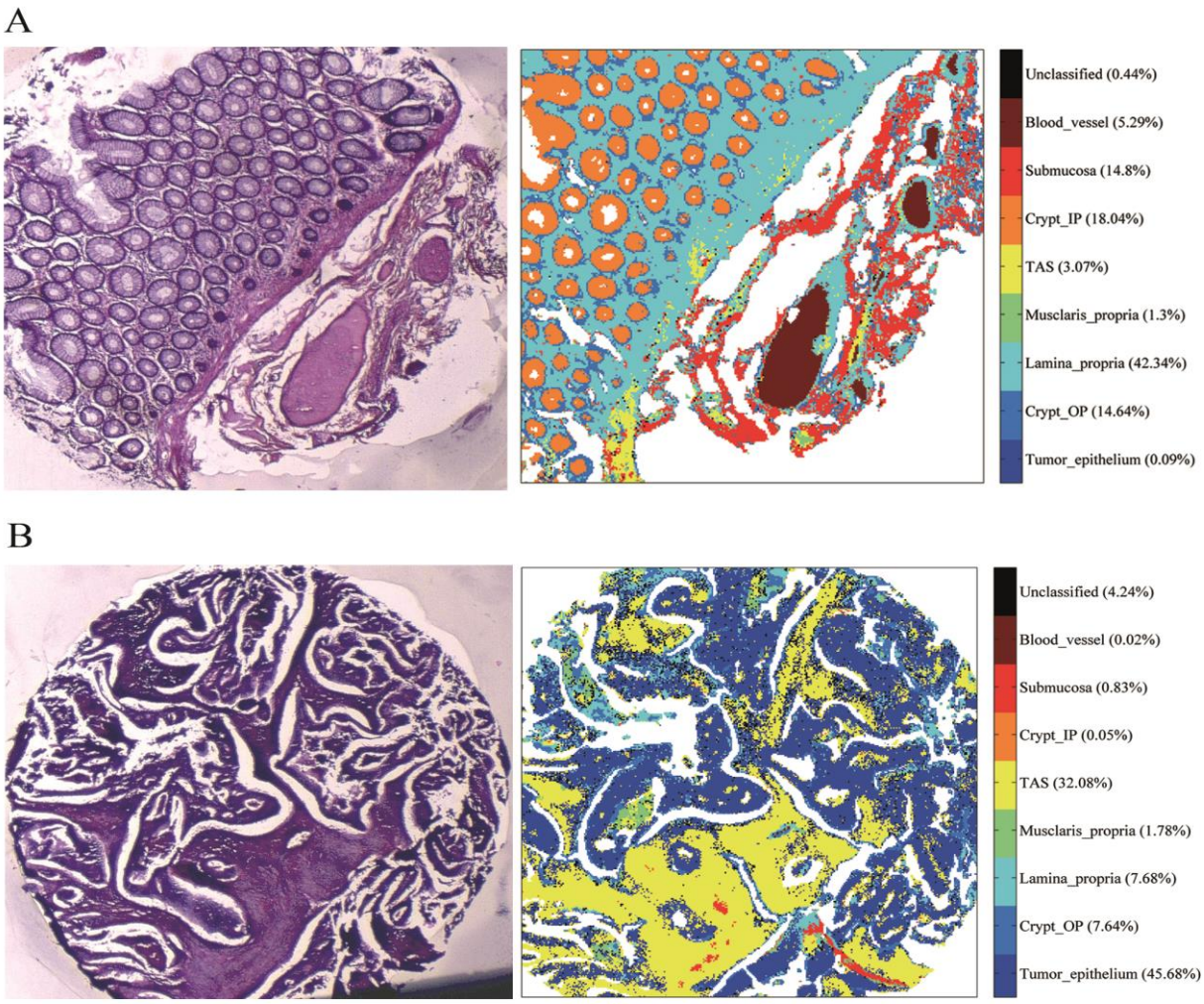


Figure 2: Performance of the prediction model: Identification of unknown colonic tissues by spectral histopathology. Left panel: HPS stained colon tissues (Supplementary Table S1, sample # 14D and 7C); Right panel: Infrared spectral predicted images. A is a non-tumoral colonic tissue section in which all the important normal colonic histological features are well-identified by the model. The important histological classes such as normal epithelium (crypt-IP and crypt-OP), connective tissue, blood vessels, etc are represented by a specific color-code. B is a moderately differentiated colon adenocarcinoma in which the tumor epithelium is together with its associated stroma are represented by the specific color-code. Note that there is a complete absence of normal epithelium. The HPS images are at 5X magnification.

Analyst Accepted Manuscript

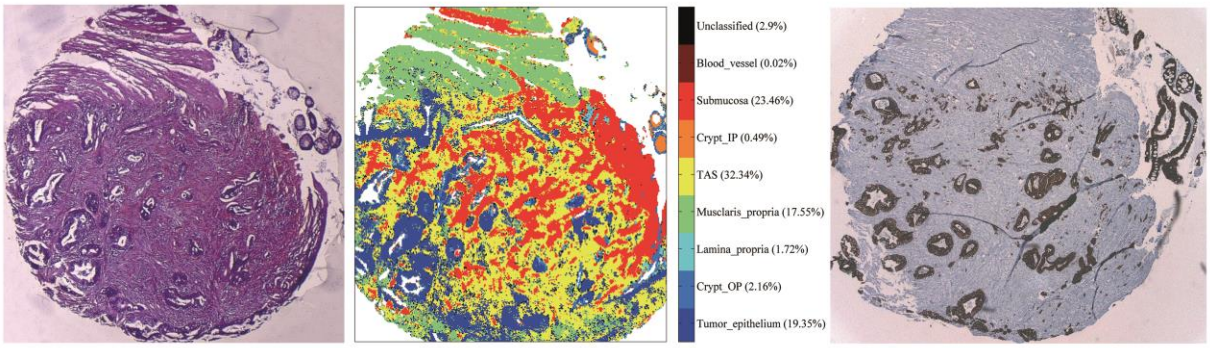


Figure 3: Identification of tumor budding in an unknown colonic tissue. Left panel: HPS stained colonic tissue (Supplementary Table S1, sample # 9B); Middle panel: Infrared spectral predicted image; Right panel: KL 1 immuno-stained image. The sample is a moderately differentiated colon adenocarcinoma in which the cancerous glands are identified along with the tumor-associated stroma. Small isolated tumor clusters representing tumor-budding are identified branching out into the stroma. The tumor-stromal boundary is also well-identified and clearly demarcated from the normal connective tissue (muscularis propria). In the same sample, few normal colonic glands are seen in the top-right position identified by presence of normal epithelium. The HPS and the IHC images are at 5X magnification.

Analyst Accepted Manuscript

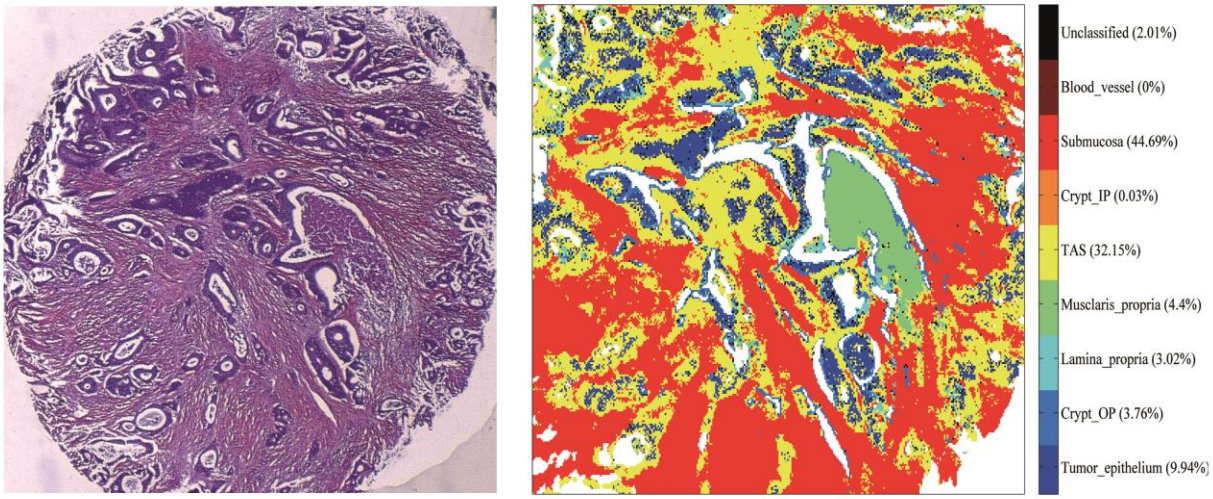
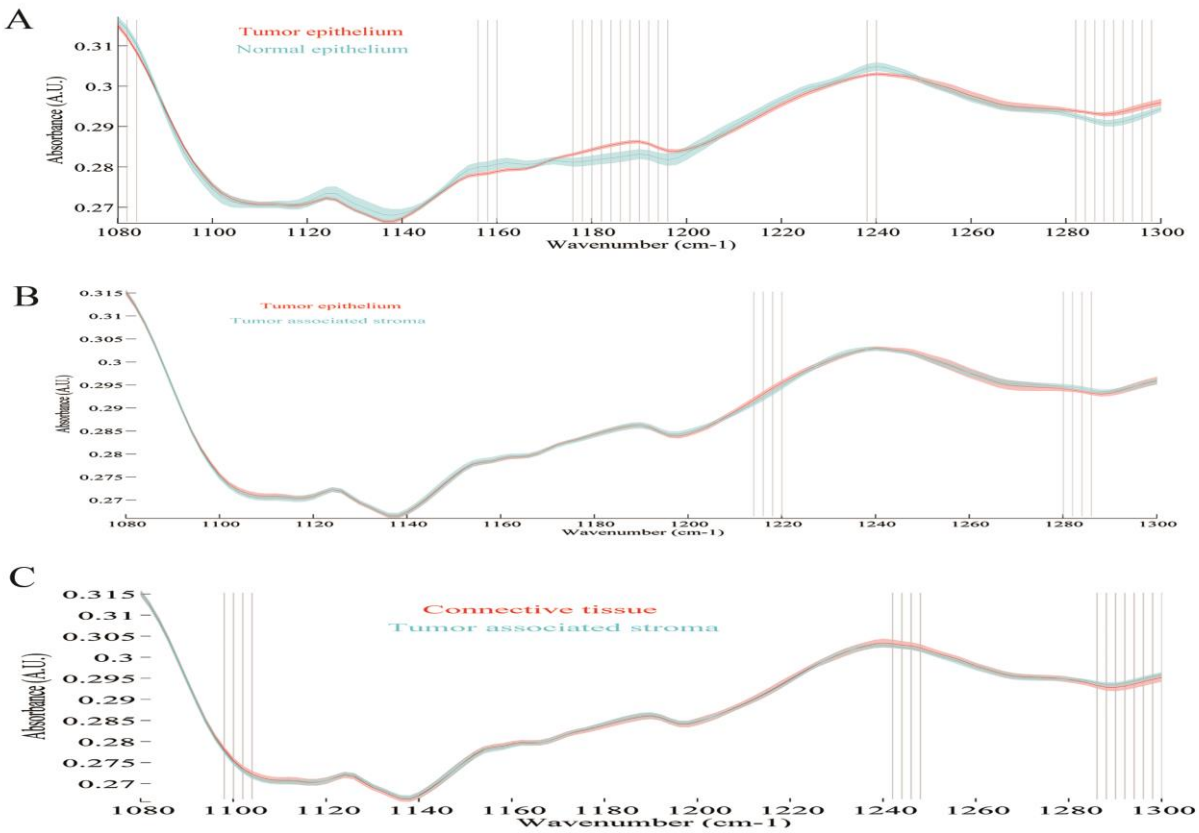


Figure 4: Tumor stroma geographical proximity. The sample is a moderately differentiated colonic adenocarcinoma with its associated stroma (Supplementary Table S1, sample # 11B). Along with the highly-correlated prediction, the nature of the connective tissue into which the tumor has infiltrated is also identified. The HPS image is at 5X magnification.

Analyst Accepted Manuscript

27



660

661

662 **Figure 5:** Most discriminant infrared spectral vibrations identified by the Mann-Whitney *U* test. The test

663 was performed for A: Tumor epithelium versus normal epithelium ($p<0.005$), B: Tumor epithelium versus

664 tumor associated stroma ($p<0.01$), and C: Connective tissue versus tumor associated stroma ($p<0.1$). For

665 each class in the figure, the mean spectrum (+/-) the standard deviation is represented.

666

667

Analyst Accepted Manuscript

28

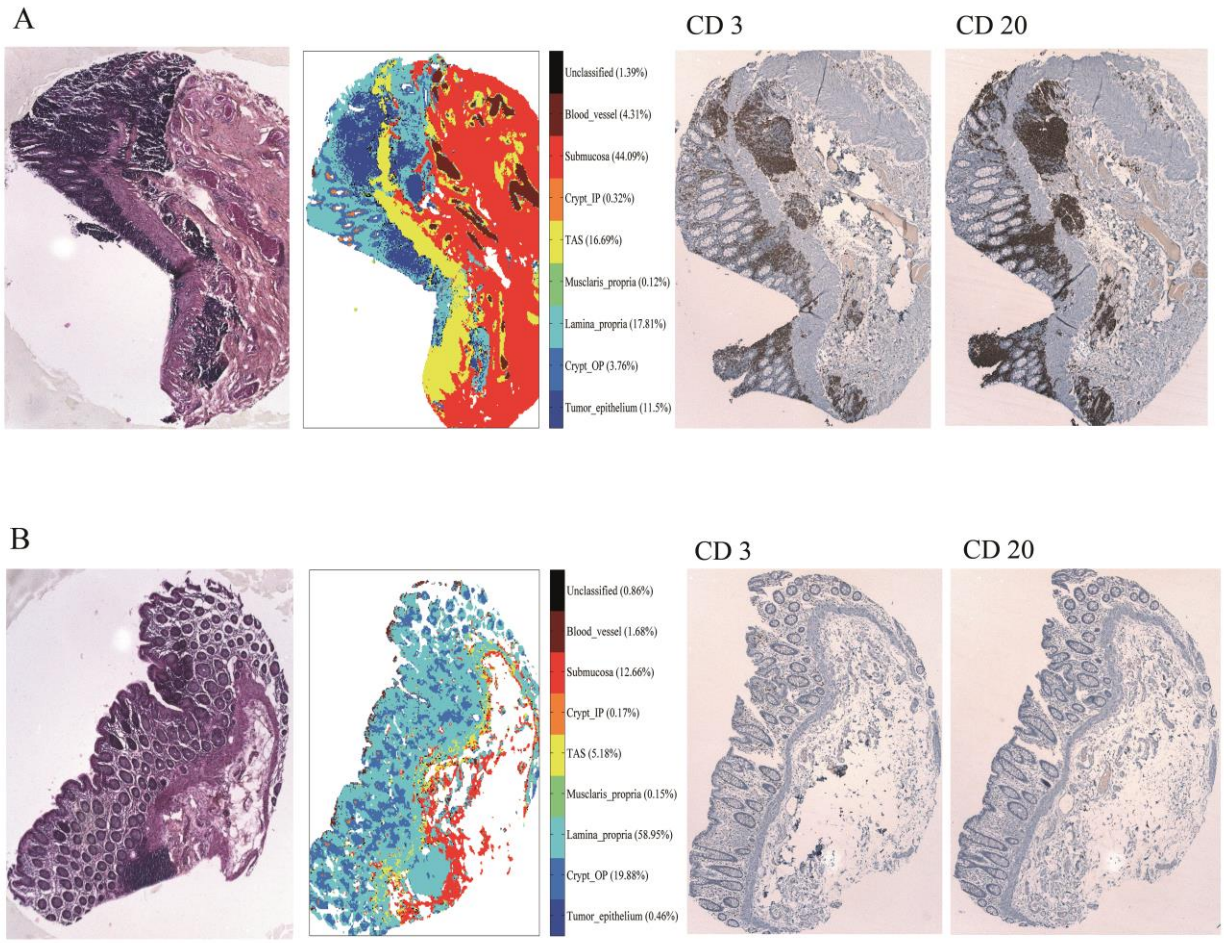


Figure 6: Influence of tissue inflammation on the prediction model. Left to right: 1. HPS stained colon tissues (Supplementary Table S1, sample # 32 and 31); 2. Infrared spectral predicted images; 3. Immuno-stained images for CD3 marker and; 4. Immuno-stained images for CD20 marker. A is a non-tumoral colonic tissue with typical normal glands. The mucosa is partially populated by lymphoid follicle as seen in the HPS image. The prediction model identified these regions in the mucosa as tumor. Immuno-staining for CD3 and CD 20 markers revealed that the tumor class in the predicted images actually corresponded to inflammatory signatures. B is another non-tumoral tissue with insignificant tumor pixels in the predicted image. Immuno-staining is negative for CD 3 and CD 20 indicating absence of inflammatory signature. The HPS and the IHC images are at 5X magnification.

Analyst Accepted Manuscript

Tables:

Table 1: The confusion matrix.

The confusion matrix representing the sensitivity of the infrared spectral imaging based prediction model, developed using 8 classes, to the gold standard histopathological attribution, in the spectral range of 1080 cm^{-1} to 1300 cm^{-1}. The table shows an average sensitivity of 89. 49 %.

Predicted class (infared imaging)

Histopathological class	'Tumor_epithelium'	'Crypt_OP'	'Lamina_propria'	'Musclaris_propria'	'TAS'	'Crypt_IP'	'Submucosa'	'Blood_vessel'
'Tumor_epithelium'	96,45	0	1,32	0,07	2,03	0	0	0,1
'Crypt_OP'	0,1	88	6,16	0,28	0	4,42	0,5	0,22
'Lamina_propria'	2	1,27	83	0,14	11,34	0,09	1,8	0,45
'Musclaris_propria'	0,1	0,04	0	98,22	1,1	0,04	0	0,04
'TAS'	16,33	0	1	0,08	81,31	0,02	1,24	0
'Crypt_IP'	0,04	6	0,04	0,04	0	93,7	0,16	0,04
'Submucosa'	0,1	0	7,5	0,05	14,42	0	77,54	0,35
'Blood_vessel'	0	0	0	0	2,3	0	0	97,7
No. of spectra used in the model	35083	3567	14106	4514	16409	8377	3964	782

Total = 86802

Analyst Accepted Manuscript