Automatic identification of paraffin pixels on FTIR images acquired on FFPE human samples

Dominique Guenot, Warda Boutegrabet, Olivier Bouché, Camille Boulagnon-Rombi, Aude Marchal Bressenot, Olivier Piot, Cyril Gobinet

Automatic identification of paraffin pixels on FTIR images acquired on FFPE human samples.

Warda Boutegrabet[1,2], Dominique Guenot[1], Olivier Bouché[2,3], Camille Boulagnon-Rombi[4,5], Aude Marchal Bressenot[2,5], Olivier Piot[2,6], Cyril Gobinet[2,*]

[1]Université de Strasbourg (Unistra), Institut National de la Santé et de la Recherche Médicale, IRFAC Inserm U1113, 3 avenue Molière, 67200 Strasbourg, France

[2]Université de Reims Champagne Ardenne, BioSpecT EA 7506, 51 rue Cognacq-Jay, 51097 Reims, France

[3]CHU de Reims, Hepato-Gastroenterology Department, rue du Général Koenig, 51092 Reims, France

[4]Université de Reims Champagne Ardenne, CNRS, MEDyC UMR 7369, 51 rue Cognacq-Jay, 51097 Reims, France

[5]CHU de Reims, Biopathology Laboratory, rue du Général Koenig, 51092 Reims, France

[6]Platform of Cellular and Tissular Imaging (PICT), 51 rue Cognacq-Jay, 51097 Reims, France

*corresponding author

Email: warda.boutegrab@gmail.com, guenot@unistra.fr, obouche@chu-reims.fr, cboulagnon-rombi@chu-reims.fr, amarchal@chu-reims.fr, olivier.piot@univ-reims.fr, cyril.gobinet@univ-reims.fr

ABSTRACT: The transfer of mid-infrared spectral histopathology to the clinic will be possible conditionally upon easily applicable in clinical practice. The rapid analysis of formalin-fixed paraffin-embedded (FFPE) tissue section is thus a prerequisite. The chemical dewaxing of these samples before image acquisition used by the majority of studies is in contradiction with this principle. Fortunately, the in silico analysis of the images acquired on FFPE samples is possible by using extended multiplicative signal correction (EMSC). However, the removal of pure paraffin pixels is essential to perform a relevant classification of tissue spectra. So far, this task was possible only if using manual and subjective histogram analysis. In this article, we thus propose a new automatic and multivariate methodology based on the analysis of optimized combinations of EMSC regression coefficients by validity indices and KMeans clustering in order to separate paraffin and tissue pixels. The validation of our method is performed using simulated infrared spectral images by measuring the Jaccard index between our partitions and the image model, with values always over 0.90 for diverse baseline complexity and signal to noise ratio. These encouraging results were also validated on real images by comparing our method with classical ones and by computing the Jaccard index between our partitions and the KMeans partitions obtained on the infrared image acquired on the same samples but after chemical dewaxing, with values always over 0.84.

Tissue microscopic imaging by mid-infrared (IR) absorption spectroscopy appears as an emerging technique to help the pathologists in the molecular characterization of tissues. Combined with statistical data processing, the approach revealed efficient to evidence various histological structures and to differentiate between physiological or pathological states, without any labelling or staining agent[1]. The analytical capability of the technique relies on the multivariate nature of the recorded spectra, reflecting the overall biochemical composition of the sample and the molecular alterations associated with physiological changes or malignancy. Among the scientific literature, this approach, named spectral histopathology (SHP) has proven to be effective in the identification of cancerous tissues in comparison with their non-tumoral counterpart. Several types of lesions were studied, e.g. skin[2], melanomas[3], prostate[4], lung[5], cervical[6], brain[7], breast[8,9] or colon[10-13] cancers. Besides the analysis of cancerous tissues, SHP was applied in other biomedical issues such as the characterization of inflammation[14] or age-related alterations in tissue such as skin[15]. These examples were proof-of-concept studies but several assets of mid-IR imaging make it possible to consider the deployment of the technique for clinical diagnostic applications. For a routine use, this technique has a low cost once the equipment is acquired, and allows to map the tissue at a microscopic scale and in a short time[1]. Importantly, data processing can be automated so as to be totally independent of the operator contrary to conventional histopathological examination. In ambiguous issues, the interpretation of histology or immunostaining can be subject to a lack of consensus between several pathologists, each interpreting with his/her own expertise and subjectivity.

The clinical transfer of this biophotonic technology requires the confirmation of the proof-of-concept studies on large-scaled retrospective cohorts coming from tumor biobanks. However, the majority of these biopsies are formalin-fixed paraffin-embedded (FFPE) for preservation, preparation and long-term storage purposes and paraffin has a significant infrared response disrupting the infrared image analysis. In previous studies[2,12,16–21], we have demonstrated the possibility to analyze FFPE tissues without chemical dewaxing, which facilitates retrospective studies on reference tissues in tumor banks. Indeed, spectral interferences of paraffin can be modelled and neutralized in order to keep only the spectral diversity associated with the molecular

composition of the tissue structures. In addition, pixels corresponding to noisy spectra or pixels with a strong contribution of paraffin are considered as outliers and can be removed from the dataset. The identification of outliers' pixels is an important step for the construction of an efficient classification model.

The procedure to remove outliers was originally developed from the extended multiplicative signal correction (EMSC) preprocessing of the data[22]. In the current version of the algorithm[16,20], the outlier's elimination proceeds by selecting manually thresholds on values of two coefficients, precisely of the reference spectrum fit and modeling error respectively. However, this operator-dependent selection is a break on the automation of the SHP.

Therefore, in this study, we describe a novel methodology which automatically removes the outlier's spectra of mid-infrared spectral images collected from FFPE colon tissues. Original tools of data partitioning and simulated data were also used to demonstrate the performance of the separation between spectra informative of tissue composition and outliers' spectra. The performance of our automatic approach has been evaluated on simulated and real infrared images and compared with other existing semi-automatic methods for detection of paraffin and outliers' spectra.

EXPERIMENTAL SECTION

Samples

Formalin-fixed paraffin-embedded tissue sections. FFPE blocks of metastatic colon carcinoma were obtained from the colon cancer surgery of 3 patients with T4N1M1 staging, at the pathology department of the Reims university hospital. The written informed patient consent was obtained according to the approved local ethics committees (n° AC-2019-3408).

Two FFPE blocks of xenografted human colon carcinoma were obtained from the INSERM U113 research group. This animal experiment was conducted in accordance with the French Ethical Approval Apafis#16125-2018030716202418 v2 according to the European guidelines.

For each block, two consecutive 6 µm thick sections were cut using a Microm HM 335 E microtome (Microm Microtech, Brignais, France). The first section was deposited on a calcium fluoride (CaF_2) window (Crystran, Dorset, UK) for mid-infrared spectral analysis. The adjacent section was mounted on a glass window and stained with hematoxylin and eosin (HE) for conventional histology, in order to serve as a reference for the spectral histopathology analysis.

Chemically dewaxed FFPE tissue sections. In the majority of published studies, FFPE tissue sections are chemically dewaxed before mid-infrared spectral analysis in order to remove the paraffin[5,6,8–10,13,23–26] which presents a parasitic signal superimposed to the tissue infrared signal[16,20,27].

In this work, 5 FFPE tissue sections (2 for mice and 3 for human patients) that were previously analyzed by mid-infrared spectral imaging were chemically dewaxed by immersion in several xylene baths. Then, mid-infrared spectral imaging was performed once again on these chemically dewaxed tissue sections which will be thus considered as gold standard in this study.

Thus, in order to objectively evaluate the performances of the different investigated approaches, the distinction between tissue and non-tissue spectra on the FFPE sections (presented below in section entitled "Identification methods of pure paraffin spectra") will be compared with the clustering outcomes obtained from the corresponding dewaxed sections.

Frozen tissue section. In order to generate simulated FFPE IR spectral images, a xenografted human colon carcinoma sample has been embedded in the Tissue-Tek optimum cutting temperature (O.C.T.) formulation to slice 6 µm-thin cryo-cross-sections using a LEICA (CM 3050 S) at −20°C. From this sample, a tissue area has been selected for FTIR spectral imaging using the same procedure as for FFPE tissue sections.

Data collection

Fourier Transform Infrared (FTIR) images were collected in transmission mode using a Spectrum Spotlight 300 FTIR imaging system coupled to a Spectrum One FTIR spectrometer (Perkin Elmer, Courtaboeuf, France), equipped with a liquid nitrogen-cooled mercury–cadmium–telluride (MCT) detector. Prior to spectral image acquisition, a visible image of the sample was collected in order to select up to 4 different tissue zones and one pure paraffin zone to be analyzed. For each pixel of these selected areas, 16 scans were averaged on the spectral range 750-4000 cm^{-1}, using a spectral resolution of 4 cm^{-1} and a pixel size of 6.25x6.25 µm². A background spectrum from the CaF_2 window was recorded using 240 accumulations and subtracted automatically from each collected image by the Spectrum Image software (Perkin Elmer).

In total, 12 FTIR spectral images were collected on the FFPE tissue sections (3 for mice and 9 for human patients), and 12 on the chemically dewaxed tissue sections (3 for mice and 9 for human patients), with 15000 pixels per image in average. In addition, 5 spectral images of paraffin composed approximately of 12000 pixels were recorded to model the spectral interference signal in the Extended Multiplicative Signal Correction (EMSC) model defined below in section entitled "Data pre-processing".

Simulated spectral images

The validation of a method on real-world datasets is the final and most valuable step. However, simulated datasets are usually constructed for the following reasons[28]. First, the ground truth is perfectly known. Second, the main variability sources that can be observed in the real datasets are completely under control. The influence of these variability sources on the results can thus be easily studied. Third, a new method is often dependent of internal parameters (such as the number of latent variables for partial least squares) that can be easily tuned on a simulated dataset. The behavior of a new proposed method can thus be fully understood using a simulated dataset.

In this sense, we constructed simulated FTIR spectral images of FFPE tissue sections using the model described in the Supporting Information. This model permits to modulate the simulated spectral images according to the following parameters. First, the ratio between the tissue and pure paraffin areas can be adjusted by specifying their respective number of pixels. Second, the baseline complexity can be modulated by specifying the order of the polynomial function used to model the baseline. Third, the signal to noise ratio (SNR) can be controlled by the standard deviation σ of the Gaussian noise added to the model.

Data pre-processing

First, the real FTIR images were corrected from the water vapor and carbon dioxyde atmospheric absorptions by the Spectrum IMAGE software (Perkin-Elmer).

Second, these spectra were cut in the 900-1800 cm^{-1} fingerprint region since it is known to be the most informative spectral range for this type of samples[17].

Third, Extended Multiplicative Signal Correction (EMSC)[22] was applied to the real and simulated IR images using the following linear model for each spectrum $\boldsymbol{s_i}$:

$$\boldsymbol{s_i} = \boldsymbol{a_i}\hat{\boldsymbol{s}} + \boldsymbol{b_i}\boldsymbol{I} + \boldsymbol{c_i}\boldsymbol{P} + \boldsymbol{e_i}. \quad (1)$$

In this work, $\hat{\boldsymbol{s}}$ is a reference spectrum chosen as the average spectrum of the dataset. On a FTIR image acquired on a pure paraffin area, the mean spectrum was computed and a principal component analysis was performed in order to find the main sources of spectral variability due to paraffin. The mean spectrum and the $N_I = 9$ first principal components (expressing 98% of variance) were pooled in a matrix $\boldsymbol{I}$, named interference matrix, in order to model the paraffin variability into the EMSC model. $\boldsymbol{P}$ is a 4th order Vandermonde matrix of wavenumbers used to model the baseline and light scattering effect. $\boldsymbol{e_i}$ is the modeling error vector. $\boldsymbol{a_i}$, $\boldsymbol{b_i} = [b_{i0}, b_{i1}, \ldots, b_{iN_I}]$ and $\boldsymbol{c_i} = [c_{i0}, c_{i1}, \ldots, c_{i4}]$ are the regression coefficients of $\hat{\boldsymbol{s}}$, $\boldsymbol{I}$ and $\boldsymbol{P}$ respectively and are estimated by ordinary least squares. Then, each spectrum is corrected using the following equation:

$$\boldsymbol{s_i^c} = \hat{\boldsymbol{s}} + \boldsymbol{e_i}/\boldsymbol{a_i}. \quad (2)$$

Thus, the EMSC preprocessing permits to perform simultaneously i) the neutralization of the variabilities of the baseline and of infrared contribution of the paraffin embedding medium, and ii) the normalization of the data spectra around the mean dataset spectrum. The reader can refer to [20,22,27] for more details about the EMSC model and its application to spectral FTIR images acquired on FFPE tissue sections.

To assess the model performance, the modeling residue $\sum_{k=1}^{N_\lambda} e_{ik}^2$ is usually computed, where N_λ is the number of wavenumbers composing each spectrum[16,17]. However, this expression being quadratic, its visualization using histogram or estimated density function is difficult. To avoid this problem, we will consider the natural logarithm of the modeling residue $r_i = ln\left(\sum_{k=1}^{N_\lambda} e_{ik}^2\right)$.

Identification methods of pure paraffin spectra

Spectral band ratio (BR). Spectral band ratio is a routinely used method to detect spectra contaminated with an unwanted compound contribution, such as water vapor, substrate, noise or preservation medium[30]. This method is based on the computation of the ratio between the integrated intensities of two bands and on the definition of a decision threshold.

In our study, this method has been applied to detect the pure paraffin pixels on the recorded FTIR tissue images by computing the ratio between the 1600–1700 cm^{-1} Amide I band associated with the tissue and the 1430–1490 cm^{-1} paraffin band. Previously to the ratio computation, each spectral band was corrected from its baseline computed as the straight line passing through the two band extreme wavenumbers, i.e. 1600 and 1700 cm^{-1} for the Amid I band, and 1430 and 1490 cm^{-1} for the paraffin band. An example of estimated baselines is shown on Figure S-1(a).

Spectra were considered as pure paraffin pixels if their band ratio was less than a threshold that was manually and differently chosen for each FTIR image.

Univariate analysis of EMSC a regression coefficient and r modeling residue (UA). Univariate analysis of EMSC regression coefficient a and modeling residue r has been developed specifically to detect pure paraffin spectra[16] and applied in the majority of studies[2,17,20,27,31] on FTIR images acquired on FFPE tissue sections. This method is based on the manual selection by the operator of two different thresholds, one for the a fitting coefficient and one for the natural logarithm of the modeling residue r estimated by EMSC. These two thresholds, named τ_a and τ_r in the following, are selected independently of each other. Spectra for which $a_i > \tau_a$ and $r_i < \tau_r$ are considered as good quality tissue spectra.

This method has also been adapted to analyze sample in tissue microarray[12,19].

Multivariate analysis of EMSC fitting coefficients (MA). In this paper, we propose a new method to identify the non-tissue spectra. After EMSC, a set of the fitting coefficients belonging to $\{a, b_0, b_1, \ldots, b_{N_I}, c_0, c_1, \ldots, c_4, r\}$ is selected and processed by KMeans clustering[32] in order to decompose the dataset into two groups, i.e. one group for the non-tissue spectra and one for the tissue spectra. Contrary to the two previously presented methods, the proposed approach is multivariate by nature since the fitting coefficients are simultaneously exploited. Furthermore, this method is automatic since based on a clustering algorithm and objective since not based on a manual threshold selection.

In order to automatize the selection of the set of fitting coefficients, validity indices were applied. A validity index objectively measures the quality of a partition, usually based on within-cluster compactness and between-cluster separation measures[18,33]. When several partitions are estimated for different values of the parameters of a clustering algorithm, e.g. the number of clusters, a validity index is useful to determine the best partition, by optimizing the values of the parameters.

In our study, the validity indices were applied on two-cluster KMeans partitions estimated on each possible combination of EMSC coefficients. The best set of fitting coefficients is thus the one resulting to the optimal validity index value, leading to the most distinct clusters of tissue and preservation medium.

Numerous validity indices have been developed in literature[18,33]. In order to have robust results, we decided to apply four different validity indices, i.e. Xie-Beni (XB), Davies-Bouldin (DB), Pakhira–Bandyopadhyay–Maulik (PBM) and Silhouette-Width-Criterion (SWC)[18,33]. For XB and DB, the optimal value is the smallest, and the highest for PBM and SWC.

Representation of pixel memberships

At the end of each above presented procedure, the pixel memberships can be represented by a binary image where 0 and 1 are affected to pixels identified as pure paraffin and tissue, respectively.

Image registration

For a given tissue section, the FTIR spectral images on the chemically dewaxed tissue section were not acquired exactly at the same orientation and position as that on the FFPE tissue section, due to slight morphological changes during the dewaxing procedure. Furthermore, the size of the spectral image acquired on the chemically dewaxed sample was chosen higher in order to include the area scanned on the FFPE tissue section.

In order to precisely compare these acquired areas, intensity-based image registration was applied between a moving image and a fixed image using a rigid transformation consisting of translation and rotation[34]. In our case, the moving image is the binary image resulting from one of the previously presented

identification methods of pure paraffin spectra, while the fixed image is the partition obtained by applying a two-cluster KMeans on the FTIR image acquired on the chemically dewaxed sample.

t-distributed stochastic neighbor embedding (t-SNE)

In this study, t-SNE was used as a tool facilitating the visualization of multivariate datasets[35]. From a high dimensional space, this nonlinear dimensionality reduction technique aims to find the best low-dimensional (usually two-dimensional) mapping preserving the local neighborhood structure of data[36]. With t-SNE, very similar data points close to each other in the original high dimensional space are kept close in the new low-dimensional space.

Gold standard and Jaccard index

In order to evaluate their performance, the identification methods of pure paraffin spectra presented above were compared to a gold standard using the Jaccard index[37]. For simulated data, the gold standard was directly accessible from the model since paraffin and tissue pixels are known. For the real data, it was defined by applying a two-cluster KMeans on the FTIR image acquired on the same tissue section after chemical dewaxing.

The Jaccard index measures the similarity between two sample sets A and B, and is defined as the ratio between the intersection of A and B and the union of A and B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

The Jaccard index is between 0 for disjoint sets and 1 for identical sets.

In this work, the set A is defined as the pixels identified as pure paraffin on a simulated FTIR image or on a real FTIR image before chemical dewaxing by one of the pure paraffin identification methods presented above. The set B corresponds to the pixels identified as paraffin on the simulated image, or as CaF_2 by a two-cluster KMeans applied on the real FTIR image acquired on the same tissue section after chemical dewaxing.

Programming environment

All the data processing presented in this study was carried out using in-house scripts written in Matlab (The Mathworks, Natick, MA).

RESULTS AND DISCUSSION

In order to compare their performance, the identification methods of pure paraffin spectra were applied on the FTIR images acquired on the FFPE tissue sections. However, to ease the reading, the results will be illustrated using a representative FTIR spectral image acquired on a human colon carcinoma FFPE sample which HE stained section is presented in Figure 1(a).

Furthermore, the classical UA and the proposed MA identification methods of pure paraffin spectra being based on pre-processing by EMSC, the model components and examples of application on this spectral image are presented in Figures S-2 and S-3. The reference spectrum $\hat{\boldsymbol{s}}$, the components of the interference matrix $\boldsymbol{I}$ and the polynomial functions composing the Vandermonde matrix $\boldsymbol{P}$ are shown on Figure S-2. Examples of raw spectra acquired on this image on paraffin and tissue pixels, and their EMSC pre-processed versions are shown on Figure S-3. The efficiency of EMSC is visible since the pre-processed paraffin spectra are mainly composed of noise, on the contrary of the pre-processed tissue spectra on which the neutralization of the paraffin signature is evident while preserving a tissue signature variability revealing the subtle biomolecular differences between tissue pixels.

Limitations of conventional identification methods of pure paraffin spectra. First, the spectral band ratio method was applied on the spectral images acquired on the FFPE tissue sections. The ratio was computed between the paraffin and Amide I bands corrected from their baseline (Figure S-1(a)) and can be represented as a ratio intensity image (Figure S-1(b)). Then, these computed ratio values were summarized by their smoothed probability density function estimated by a normal kernel function[38] (Figure S-1(c)). By visual inspection of this distribution and in function of its shape, the threshold value is selected by the operator. This density has the shape specific of a bimodal probability density function. The threshold value must thus be determined between these two modes by a visual analysis.

As an illustrative example, three different operators analyzed independently this ratio distribution. The first one selected a threshold equal to 0.3 in order to be tissue-conservative, while the second chose an intermediate threshold equal to 1.6, and the third fixed the threshold to 2.3 in order to remove all the paraffin pixels. Figures 1(b-d) correspond to the binary images obtained after thresholding by these three operators. An underspecified threshold erroneously identifies non-tissue parts of the sample as tissue (Figure 1(b)). On the contrary, an overspecified threshold confuses tissue parts with paraffin (Figure 1(d)).

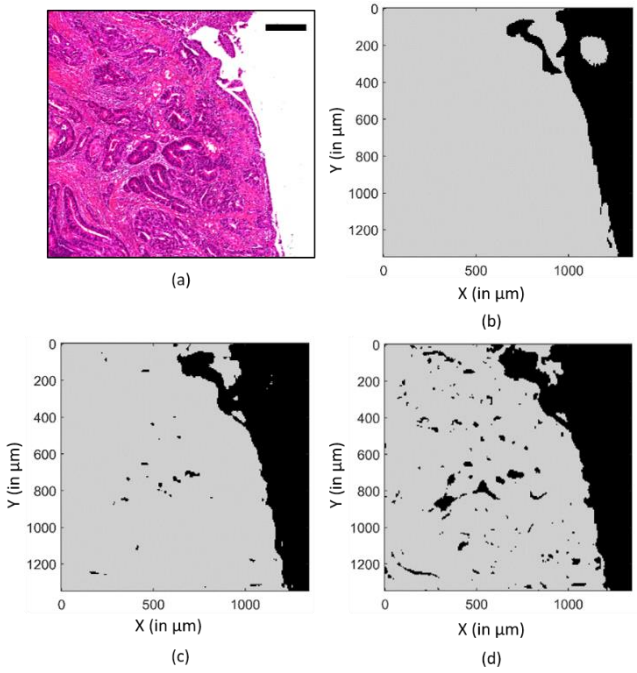

FIGURE 1: Identification of paraffin and tissue pixels by the spectral band ratio approach. (a) Human colon cancer FFPE tissue section stained with HE. The scale bar indicates 200 µm. (b-d) Binary images resulting from the thresholding of the estimated probability density using a threshold equal to 0.3, 1.6 and 2.3, respectively. Black and gray pixels correspond to estimated paraffin and tissue pixels respectively.

Second, the univariate analysis of EMSC regression coefficient a and natural logarithm of modeling residue r was tested on the same spectral image in order to detect the pure paraffin pixels. After EMSC pre-processing, the distributions of a and r

are estimated by a normal kernel function[38] (Figure S-4). The a and r distributions usually have shapes specific of bimodal probability density functions in the context of spectral image acquired on FFPE tissue sections[16]. For each distribution, a threshold value is thus determined between the two modes by a visual analysis. Pure paraffin spectra are characterized by low regression coefficient a, while tissue spectra have a high regression coefficient a. A low natural logarithm of modeling residue r is typical of paraffin or tissue spectra well fitted by the EMSC model, whereas a high value is characteristic of noisy or outlier spectra.

In order to illustrate the sensitivity of this method to threshold selection, the threshold values were selected by two different operators. The first one selected $\tau_a = 2$ and $\tau_r = 0$ in order to surely conserve all the tissue pixels. The second one chose $\tau_a = 9$ and $\tau_r = -3$ in order to completely remove the paraffin pixels. The binary images resulting from these two thresholdings are visible on Figures 2(a-b). The comparison of these figures with the image of unstained and adjacent HE tissue section (Figure 1(a)) reveals a small correlation indicating that the paraffin and tissue pixels are badly identified. Indeed, on the first hand, a significant portion of paraffin pixels are misidentified (Figure 2(a)). On the other hand, a modification of these thresholds resulted in the inverse behavior, i.e. an over identification of paraffin pixels (Figure 2(b)). The difference image (Figure 2(c)) represents the pixels that are differently identified between these two binary images. The threshold is a very sensitive parameter since for this example 6% of pixels are changing of identification. Taken together, all these results prove that the performance of this type of pure paraffin pixel identification methods based on thresholding is very sensitive to the chosen threshold values. Furthermore, identifying the optimal thresholds to completely remove the paraffin pixels is impractical in real situations because sample-, image- and user-dependent.

Towards a multivariate and automatic approach. The two previously presented thresholding methods are based on the analysis of one or two parameters. However, the EMSC model provides many supplementary information about the physical and chemical composition of the sample at each studied pixel, which can be helpful for the discrimination between paraffin and tissue pixels.

A two-dimensional t-SNE applied on the EMSC regression coefficients of data acquired on a human colon carcinoma FFPE section reveals two clearly visible data groups, suggesting the presence of two different spectral patterns which can be identified as paraffin and tissue areas respectively (Figure S-5). Paraffin and tissue areas can thus be separated from the analysis of the EMSC regression coefficients.

To confirm this intuition, images reconstructed from some EMSC regression coefficients are illustrated on Figure 3. A clear contrast between paraffin and tissue areas can be observed using a, r, b_0 and c_0 (Figures 3(a-d)). This observation is confirmed by the two-cluster KMeans partition estimated on the $\{a, r, b_0, c_0\}$ EMSC regression coefficients (Figure 4(a)). The use of a and r in the classical univariate analysis method[16] is thus justified, but this method does not exploit all the available information such as b_0 and c_0 which provide complementary information about paraffin-tissue edges.

Altogether, these results justify the multivariate exploitation of the EMSC regression coefficients which is the core of our proposed approach.

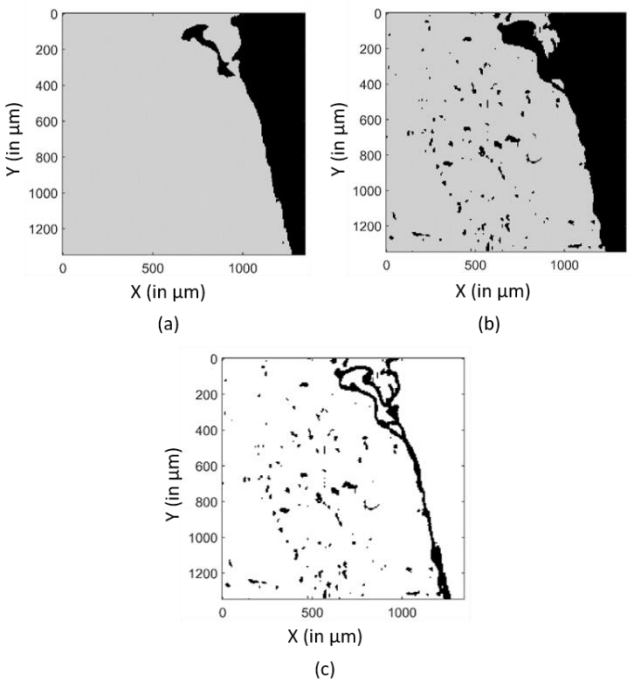

FIGURE 2: Identification of paraffin and tissue pixels by the univariate analysis of EMSC a regression coefficient and r natural logarithm of modeling residue. Binary images resulting from the thresholding of the estimated probability densities using $\tau_a = 2$ and $\tau_r = 0$ (a), and $\tau_a = 9$ and $\tau_r = -3$ (b), respectively. Black and gray pixels correspond to estimated paraffin and tissue pixels respectively. (c) Image presenting in black the pixels differently identified between (a) and (b).

Furthermore, being based on a simple application of a two-cluster KMeans in order to separate the pixels into two clusters, one for paraffin and one for tissue, our proposed method is automatic since it does not require the setting of parameters by the operator, contrary to the previously presented thresholding methods.

However, all the EMSC regression coefficients are not useful. Distinguishing between paraffin and tissue pixels is difficult, if not impossible on the images reconstructed using b_5 and c_2 (Figures 3(e-f)). This result is confirmed by the two-cluster KMeans partition estimated using the EMSC regression coefficient combination $\{b_5,c_2\}$ (Figure 4(b)) which is clearly not correlated to the sample structure (Figure 1(a)).

The relevant question is thus which combination of EMSC fitting coefficients is optimal for the distinction between tissue and paraffin spectra. In this work, validity indices were used as an objective and quantitative measure to answer this question and propose a fully automatic method.

The development of our proposed method based on the multivariate analysis of the best estimated combination of EMSC regression coefficients determined by validity indices is thus completely justified by these previous results. In order to objectively study its efficiency, we tested our method on simulated spectral images.

Evaluation of the multivariate analysis of EMSC fitting coefficients on simulated spectral images. In order to evaluate the performance of our proposed method, a total of 30 simulated spectral images were generated according to the procedure de-

scribed and the parameter setting detailed in the Supporting Information. An example of a simulated spectral image is given on Figure S-6.

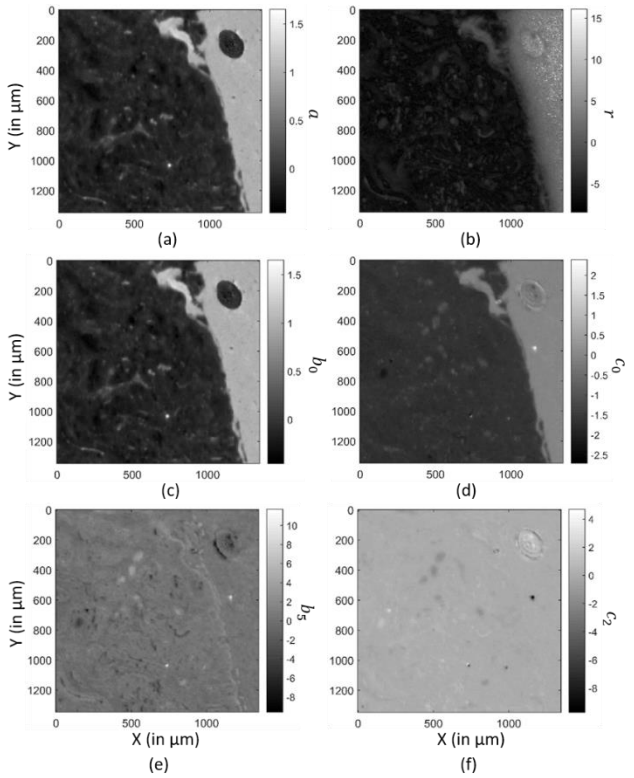

FIGURE 3: Grayscale images reconstructed from: (a) the regression coefficient a of the reference spectrum, (b) the natural logarithm of the modeling residue r, (c) the regression coefficient b_0 of the mean pure paraffin spectrum, (d) the regression coefficient c_0 corresponding the zero-order polynomial coefficient, (e) the regression coefficient b_5 of the 5th paraffin principal component, (f) the regression coefficient c_2 corresponding to the second-order polynomial coefficient, estimated by the EMSC model.

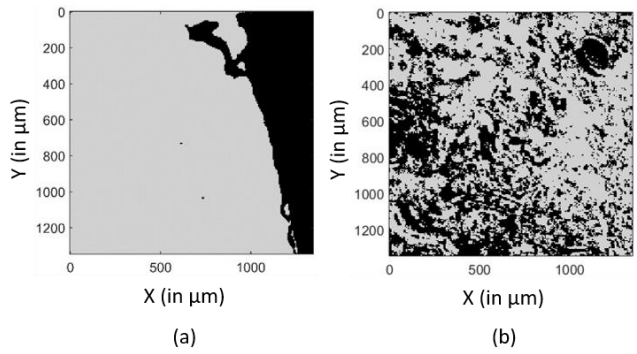

FIGURE 4: Partitions obtained by applying a two-cluster KMeans on the (a) $\{a, r, b_0, c_0\}$ and (b) $\{b_5, c_2\}$ combinations of EMSC regression coefficients. Black and gray pixels correspond to paraffin and tissue pixels on (a) respectively, and to unidentified clusters on (b).

The first part of this evaluation consisted in testing the ability of our multivariate approach coupled to validity indices to automatically estimate the optimal EMSC regression coefficient combination leading to the most compact and separated paraffin and tissue clusters. For this purpose, a noise-free (SNR = 32 dB) spectral image simulated with a first-order polynomial function was used. In order to give the same weight to all the EMSC regression coefficients, each one was normalized using the Standard Normal Variate (SNV) method before the application of our procedure. Then, each validity index was applied on the two-cluster partition estimated by the KMeans algorithm applied on each possible combination of EMSC coefficients. For a given validity index, the EMSC coefficient combinations were then ranked according to their computed validity index values (in ascending order for XB and DB, and in descending order for PBM and SWC). For each validity index, the top 5 ranked combinations of EMSC regression coefficients, i.e. giving the best compact and separated clusters, were retained and are given in Table S-1. The four used validity indices give very similar results. However, this information was summarized by considering only the EMSC coefficient combinations identified as belonging to these top 5 ranked combinations by the four validity indices simultaneously. These combinations are named consensual combinations in the remaining of the manuscript. From Table S-1, three consensual combinations are identified, namely $\{a, b_0\}$, $\{b_0, c_0\}$, $\{a, b_0, c_0\}$. The complete workflow of the proposed method is provided on Figure S-7.

These results are confirmed by the application of t-SNE to the regression coefficients $\{a, b_0, c_0\}$. Indeed, this combination is efficient to distinguish between paraffin and tissue spectra (Figure S-8 (a)). In addition, the correlation of each of these three coefficients to the ground truth classes can also be separately visualized (Figures S-8 (b-d)). Indeed, paraffin spectra are characterized by insignificant contribution of the reference spectrum (a), and by high contribution of paraffin (b_0) and baseline (c_0), while the opposite is observed for the tissue spectra.

These results are in accordance with the way that the model has been generated since the data are composed of three main sources of variability, i.e. the tissue, the paraffin and the baseline, whose contributions are mainly estimated in the EMSC model by the a, b_0 and c_0 coefficients, respectively. It has to be noticed that the identification of the paraffin and tissue pixels by our unsupervised methodology using the $\{a, b_0\}$, $\{b_0, c_0\}$, $\{a, b_0, c_0\}$ combinations is almost perfect since resulting in a Jaccard index around 0.9925. The misidentified pixels are the three with an almost zero contribution of tissue as explained in the Supporting Information and visible on Figure S-6 (b, e). On the contrary, the worst combinations identified by the validity indices, namely $\{c_1\}$, $\{b_1\}$, $\{b_3\}$, also give the worst supervised Jaccard index of 0.3166, 0.3082 and 0.3098, respectively.

However, it has to be noticed that the modeling error residue has not been identified by the validity indices as a parameter relevant for the separation of data into two distinct clusters, which can be justified by the use of a noise-free simulated spectral image, as explained above. Furthermore, a simple first order polynomial function has been used to model the baseline, while usually a higher-order polynomial function is necessary to estimate the baseline effect on FTIR images acquired on FFPE tissue sections[12,16–19,21,31,39].

To complete the characterization of our proposed method, the evaluation of its robustness to perturbations is important. It is thus interesting to study the impact of the SNR and baseline complexity on our complete methodology including the estimation of the optimal EMSC regression coefficient combination by validity indices and Jaccard index. To achieve this goal, dif-

ferent simulated FTIR spectral images were thus generated using a baseline polynomial order varying from 0 to 4, and a SNR varying from 2^0 to 2^5 (see the Supporting Information for more information). For each simulated image and for each validity index, the top 5 combinations of EMSC regression coefficients leading to the best validity index values were determined (data not shown). To summarize these results, Table 1 presents the consensual EMSC regression coefficient combinations estimated by the four validity indices for each polynomial order and SNR couple. To confront these results with the gold-standard for each polynomial order and SNR couple, the Jaccard index was computed for each of these consensual combinations. To summarize this information for each polynomial order and SNR couple without overestimating the results of our approach, only the worst consensual coefficient combination, i.e. the one giving the smallest Jaccard index value, was considered. Figure 5(a) presents these results for all the studied polynomial order and SNR couples as a map.

Globally, the Jaccard index is decreasing in function of the baseline polynomial order and of the inverse of the SNR. This result was expected since the identification of paraffin from tissue pixels becomes more difficult when the perturbation sources increase, especially noise. However, whatever the couples of SNR and baseline polynomial order, the Jaccard index remains very high (over 0.9). Our unsupervised method based on validity indices is thus objectively estimating efficient combinations of EMSC regression coefficients for the separation of tissue and paraffin pixels.

TABLE 1: Consensual EMSC regression coefficient combinations estimated by the four validity indices on the simulated spectral images in function of the baseline polynomial order and the signal to noise ratio (SNR) expressed in decibels (dB).

Polynomial order\SNR	32	16	8	4	2	1
0	a, b_0 a, c_0 a, b_0, c_0	a, r a, b_0 a, r, b_0	a, r a, b_0 a, r, b_0	a, r a, b_0 a, r, b_0	a, r a, r, b_0	a, r
1	a, b_0 b_0, c_0 a, b_0, c_0	a, b_0	a, r a, b_0 a, r, b_0	a, r a, r, b_0	a, r	a, r
2	a, b_0 b_0, c_0 a, b_0, c_0	a, c_0	a, r a, c_0 a, r, c_0	a, r a, c_0 a, r, c_0	a, r a, c_0 a, r, c_0	a, r a, c_0 a, r, c_0
3	a, b_0 b_0, c_0 a, b_0, c_0	a, b_0	a, r a, r, c_0	a, r	a, r a, r, c_0	a, r a, r, c_0
4	a, b_0 b_0, c_0 a, b_0, c_0	a, c_0	a, c_0 a, r	a, c_0 a, r	a, c_0 a, r	a, c_0 a, r

From Table 1, for high SNR (16 or 32 dB) and whatever the baseline complexity, the best combinations remain $\{a, b_0, c_0\}$ and $\{a, b_0\}$ as for the previous results obtained for a noise-free model with a first-order polynomial baseline. For increasing noise level (evaluated precisely on unshown results to SNR < 12 dB on the simulated images), the natural logarithm of modeling residue r becomes a relevant parameter, to separate paraffin from tissue pixels, in place of c_0 for low baseline polynomial order (0 or 1), and b_0 for higher baseline polynomial order. These results thus show that the set of EMSC regression coefficients must be adapted in function of the dataset characteristics. To confirm this result, the Jaccard index was computed for each couple of SNR and baseline polynomial order using only the $\{a, b_0, c_0\}$ combination (Figure 5(b)). As explained above, the $\{a, b_0, c_0\}$ combination was previously identified as optimal for noise-free spectral image.

By comparison of Figures 5(a) and 5(b), optimization of the combination of EMSC regression coefficients is thus recommended since it gives better results (Jaccard index over 0.9) than the $\{a, b_0, c_0\}$ combination (Jaccard index over 0.8) whatever SNR and baseline complexity.

These results are thus in contradiction with the classical method which is based on the immutable exploitation of the a and r coefficients. The classical method could thus be optimal only for highly noisy datasets. Furthermore, whatever the baseline complexity and noise level, the EMSC regression coefficient a of the reference spectrum is a necessary parameter to distinguish efficiently the paraffin from tissue pixels. The use of this parameter by the classical method is thus justified. Furthermore, this result showing the importance of this EMSC regression coefficient is in accordance with other studies[40] where it has been exploited for FTIR image registration.

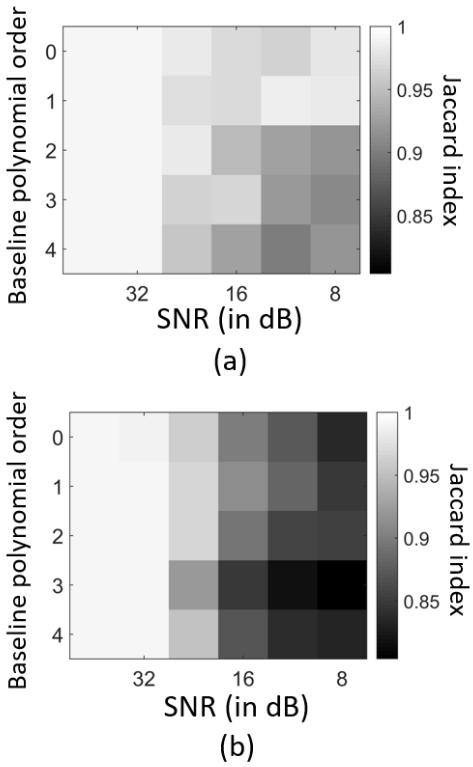

FIGURE 5: Minimum Jaccard index estimated on simulated spectral images in function of the baseline polynomial order and the signal to noise ratio (SNR) expressed in decibels (dB) for (a) the best combinations of EMSC regression coefficients estimated by validity indices, and (b) for the $\{a, b_0, c_0\}$ combination.

Altogether, these results prove that our methodology is flexible and adaptive to the main sources of distinction between paraffin and tissue spectra.

Validation of the multivariate analysis of EMSC fitting coefficients on real FTIR spectral images. The performances of our multivariate approach for identification of paraffin and tissue pixels have been successfully evaluated on simulated images. The last step consists in the validation of our methodology on real FTIR images acquired on two samples, i.e. human and xenografted FFPE colon carcinoma sections.

Such as for the simulated data, the first objective of this real data analysis was to determine the best combination of EMSC regression coefficients for the separation of paraffin and tissue pixels. For each image and for each possible combination of coefficients, a two-cluster KMeans partition was estimated and fed the four validity indices presented above. For each validity index, the top 5 combinations of EMSC regression coefficients leading to the best validity index values were determined. The obtained results are consistent with those obtained on the simulated images since the consensual combinations are $\{a, b_0\}$, $\{a, c_0\}$ and $\{a, b_0, c_0\}$. As an example, the complete results estimated on one FTIR image acquired on a FFPE human colon carcinoma sample are given in Table S-2. Extrapolating the interpretation of results obtained on simulated datasets, the obtention of these combinations necessary induces a high SNR of these real data, which is true considering the acquisition setting and the nature of the analyzed samples, and which was confirmed by the SNR of 40 dB measured on this spectral image. The same results have been obtained independently on the other acquired FTIR images (data not shown).

Such as for the simulated data, these results were confirmed by the application of t-SNE to the regression coefficients $\{a, b_0, c_0\}$. Indeed, the efficiency of our methodology to distinguish between paraffin and tissue pixels is demonstrated using a two-dimensional t-SNE by the estimation of two distinct compact and separated clusters which are completely correlated with the pixel labels estimated by our method (Figure S-9 (a)). Furthermore, the influence of each of these three coefficients on the paraffin and tissue clusters can also be separately visualized (Figures S-9 (b-d)). Such as for the simulated images, paraffin spectra are characterized by high contributions of paraffin (b_0) and baseline (c_0), while tissue spectra have a high contribution of the reference spectrum (a). Taken together, these results demonstrate the efficiency of our method on real datasets but also the realism of our simulated data generative model, where the main sources of variability of real FTIR spectral images have been efficiently incorporated.

To properly validate our methodology on real spectral datasets, it is necessary to compare our results with ground-truth defined by the chemically dewaxed FFPE tissue sections. The FTIR spectral images were acquired on chemically dewaxed samples and analyzed by a two-cluster KMeans clustering to separate substrate from tissue spectra. An example of such a gold-standard clustering image is provided on Figure 6(a).

The application of our methodology on the same section before chemical dewaxing determined $\{a, b_0, c_0\}$ as the best EMSC coefficient combination for the separation of paraffin and tissue pixels (Figure 6(b)). To compensate orientation and position differences between Figures 6(a-b), the registered version of Figure 6(b) was computed as stated above. The resulting image (Figure 6(c)) illustrates the efficiency of the used rigid registration algorithm since this image perfectly matches Figure 6(a). The paraffin and tissue areas of the studied FFPE tissue section thus visually seem to have been perfectly recognized by our proposed multivariate approach.

To objectively quantity the matching between the gold-standard and estimated registered results, the Jaccard index was computed for 3 patients (representing 9 different FTIR images) and 2 different mice (representing 3 different FTIR images). As shown in Table 2, the majority of samples have a high Jaccard index (over 0.84) with values slightly inferior to those obtained on the simulated images (around 0.99). This decrease of Jaccard index on the real data can be explained by: i) the imperfect chemical dewaxing process which is known to be incomplete, aggressive with possible sample deterioration, sensitive to chemical reagents, bath time, and histology of the analyzed tissue region[31], ii) the imperfect image registration due to the use of a simple rigid model while chemical dewaxing is well known to induce non-rigid alterations of the tissue topology. The same process, i.e. image registration and Jaccard index computation, has been applied on the same FTIR images using the BR and UA methods. For the vast majority of the considered samples, our proposed method is better than the two classical methods (third and fourth columns of Table 2). Even if the Jaccard index difference between the three methods is small, our method is automatic, contrary to the two classical ones which require time-consuming manual optimization and experienced user. Taken together, these results demonstrate the efficiency of our method, its simplicity, flexibility, automation and potential implementation in clinical routine compared to chemical dewaxing and the classical approaches.

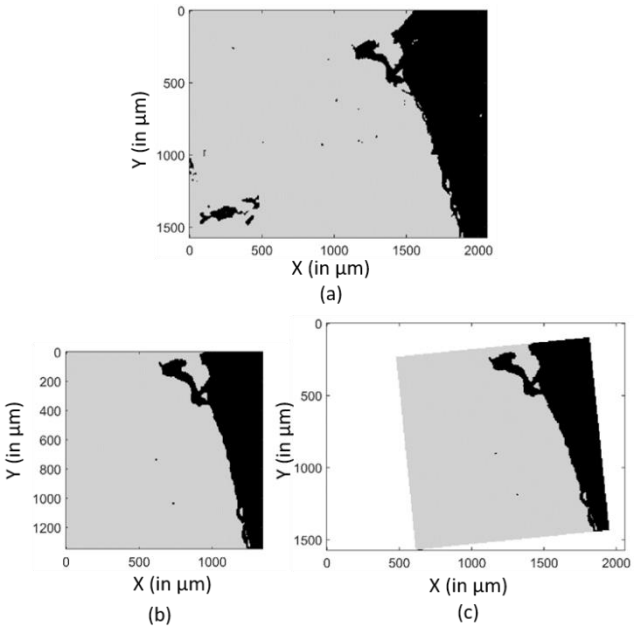

FIGURE 6: (a) Two-cluster KMeans partition estimated on the raw FTIR image acquired on a chemically dewaxed FFPE human colon carcinoma section. Black and gray pixels correspond to substrate and tissue pixels, respectively. (b) Partition obtained by our methodology on the same FFPE sample, before chemical dewaxing. Black and gray pixels correspond to paraffin and tissue pixels, respectively. (c) Image (b) after registration using image (a) as the fixed image.

Methodological discussion. Our methodology could also be applied in order to construct a specific database dedicated to the training of a supervised classification model for distinguishing paraffin and tissue pixels. The EMSC regression coefficients

combination estimated as the best by our method to identify tissue from non-tissue pixels could permit a robust calibration of the classification model.

It has to be noticed that the registration algorithm has been applied on binary images because this algorithm requires gray-scale images in inputs and these binary images are directly the results of the investigated pure paraffin pixel identification methods. However, using this strategy, the pixel identification errors may influence the registration results. In our case, this property could be in fact an advantage since the errors made by the different pure pixel identification methods will be amplified. Thus, a method making more identification errors than another will give a worse registered binary image and thus a worse Jaccard index. However, the high Jaccard indices presented in Table 2 and their small differences between methods prove that these errors little influence the registration results. However, it should be interesting to deeply study the impact of the data representation (other than binary images), of propagation of identification errors and of model complexity on the registration of FTIR images acquired on FFPE tissues sections, what is out of the scope of this article.

TABLE 2: Jaccard index computed between the paraffin pixels identified by a 2-cluster KMeans partition obtained from a FTIR image acquired on a chemically dewaxed section and by one of the following identification method of pure paraffin spectra applied on a FTIR image acquired on a FFPE section: MA, UA, and BR. Several regions of interest (ROI) were studied for 3 patients and 2 mice. For each line, the bold value indicates the best result, thus the best method.

Method / Patient and ROI	MA	UA	BR
Patient #1, ROI #1	**0.9855**	0.9796	0.9651
Patient #1, ROI #2	**0.9598**	0.9492	0.9533
Patient #1, ROI #3	**0.9797**	0.9775	0.9706
Patient #1, ROI #4	**0.9564**	0.8268	0.9011
Patient #2, ROI #1	0.8804	0.8668	**0.8988**
Patient #2, ROI #2	**0.9192**	0.9183	0.9160
Patient #3, ROI #1	0.8462	**0.8466**	0.8349
Patient #3, ROI #2	**0.9618**	0.9295	0.9038
Patient #3, ROI #3	**0.9635**	0.8518	0.8566
Mouse #1, ROI #1	**0.8454**	0.7944	0.7760
Mouse #1, ROI #2	**0.8733**	0.8613	0.7281
Mouse #2, ROI #1	**0.9417**	0.9323	0.9389

It must be noticed that optimization of the combination of EMSC regression coefficients can be time-consuming for numerous and high-dimensional FTIR spectral images. For example, for a real FTIR spectral image composed of 46656 spectra with 451 wavenumbers per spectrum, the exhaustive optimization over all EMSC coefficient combinations and using four validity indices has taken 76 hours using a computer equipped with a 3.4 GHz Intel® Core™ i7-4770 CPU, 16 Go RAM and 4 cores. However, 87% of this computational time is due to DB and PBM. Considering only XB and SWC or using other fast to compute validity indices results in a drastic reduction of this computational time.

Furthermore, the performance of our unsupervised methodology remains relatively stable in function of the SNR and baseline polynomial order on the simulated images for a fixed combination of EMSC regression coefficients since the Jaccard index is over 0.8 (Figure 5(b)). Consequently, for the processing of FTIR spectral images acquired almost in the same conditions, i.e. on tissue sections prepared according to the same standardized protocol of paraffin embedding using the same instrument with the same acquisition parameters, a good practice should thus be to optimize the combination on a unique image or on simulated data mimicking real data, and then to use the same combination on all the other images.

Furthermore, our results on real and simulated FTIR images show that the optimized EMSC coefficient combinations are always a subset of $\{a, b_0, c_0, r\}$. In real applications, another good practice should be to perform the optimization considering only the 15 possible subsets of $\{a, b_0, c_0, r\}$, i.e. $\{a\}$, $\{b_0\}$, $\{c_0\}$, $\{r\}$, $\{a, b_0\}$, $\{a, c_0\}$, $\{a, r\}$, $\{b_0, c_0\}$, $\{b_0, r\}$, $\{c_0, r\}$, $\{a, b_0, c_0\}$, $\{a, b_0, r\}$, $\{a, c_0, r\}$, $\{b_0, c_0, r\}$, $\{a, b_0, c_0, r\}$. For example, for the same real FTIR image (46656 spectra), using the 4 validity indices, this optimization has drastically reduced the computational time to 45 seconds.

The performance of KMeans algorithm is well-known to be initialization-dependent. A common practice is thus to apply KMeans several times on the same dataset in order to maximize the chances to converge to the global minimum of the KMeans objective function. The more the number of searched clusters and the complexity of the dataset (i.e. the number of dimensions and cluster overlapping), the more the necessity to repeat KMeans clustering. However, in our proposed methodology, the number of clusters is the smallest as possible, i.e. two with one for paraffin pixels and one for the tissue pixels. Furthermore, the use of only few EMSC coefficients drastically reduces the complexity of data processed by KMeans as revealed by the scatter plots of two-dimensional t-SNE applied on EMSC coefficients estimated from simulated and real datasets (Figures S-5(a), S-8(a) and S-9(a)). In the proposed methodology, repetition of KMeans algorithm was thus not considered as necessary and was thus avoided in order to reduce the computational time as much as possible. This choice was validated by the similarity of the results obtained on the simulated data and on the 12 different real images.

CONCLUSION

For complete histopathological characterization of tissue samples by FTIR imaging, the tissue area must be identified as precisely as possible. However, on FFPE tissue sections, the strong infrared signature of paraffin complicates this task by blurring the frontier between tissue and paraffin. So far, the solutions proposed in literature were based on the subjective and manual choice of thresholds from univariate histogram analysis of various quantities measured from the recorded spectra, leading to highly variable results between different operators. In this article, we proposed a new simple, objective and automatic methodology based on the multivariate exploitation of EMSC fitting coefficients. Using t-SNE, validity indices and Jaccard index on simulated and real datasets, we demonstrated the efficiency of our methodology to automatically determine the best EMSC fitting coefficients for the separation of paraffin and tissue pixels on infrared images simulated or acquired on metastatic and xenografted human colon cancer FFPE tissues. Mainly, the high similarity between FFPE tissue sections and their chemically dewaxed versions validates and confirms the

efficiency of our approach. Thus, this work confirms the efficiency and versatility of EMSC for infrared images acquired on FFPE samples since it can neutralize and normalize paraffin and baseline on tissue spectra, and combinations of its estimated regression coefficients enhance information necessary to easily distinguish paraffin from tissue pixels.

ASSOCIATED CONTENT

Supporting Information. The supporting Information is available free of charge at http://pubs.acs.org.

Description of the generative model of simulated spectral images, and its setting; supplementary figures; supplementary tables (PDF).

AUTHOR INFORMATION

Corresponding Author

*Cyril Gobinet - Université de Reims Champagne Ardenne, BioSpecT EA 7506, 51097 Reims, France.
orcid.org/0000-0002-2702-3697; Email: cyril.gobinet@univ-reims.fr

Authors

Warda Boutegrabet - Université de Strasbourg (Unistra), Institut National de la Santé et de la Recherche Médicale, IRFAC Inserm U1113, 3 avenue Molière, 67200 Strasbourg, France, and Université de Reims Champagne Ardenne, BioSpecT EA 7506, 51097 Reims, France.
orcid.org/0000-0002-6340-5967; Email: warda.boutegrab@gmail.com

Dominique Guenot - Université de Strasbourg (Unistra), Institut National de la Santé et de la Recherche Médicale, IRFAC Inserm U1113, 3 avenue Molière, 67200 Strasbourg, France.
orcid.org/0000-0003-2782-8425; Email: guenot@unistra.fr

Olivier Bouché - Université de Reims Champagne Ardenne, BioSpecT EA 7506, 51097 Reims, France, and CHU de Reims, Hepato-Gastroenterology Department, 51092 Reims, France.
Email: obouche@chu-reims.fr

Camille Boulagnon-Rombi - Université de Reims Champagne Ardenne, CNRS, MEDyC UMR 7369, 51097 Reims, France, and CHU de Reims, Biopathology Laboratory, 51092 Reims, France.
Email: cboulagnon-rombi@chu-reims.fr

Aude Marchal Bressenot - Université de Reims Champagne Ardenne, BioSpecT EA 7506, 51097 Reims, and CHU de Reims, Biopathology Laboratory, 51092 Reims, France.
Email: amarchal@chu-reims.fr

Olivier Piot - Université de Reims Champagne Ardenne, BioSpecT EA 7506, 51097 Reims, France, Université de Reims Champagne Ardenne, and Platform of Cellular and Tissular Imaging (PICT), 51097 Reims, France.
orcid.org/0000-0001-7869-5242; Email: olivier.piot@univ-reims.fr

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The authors thank the University of Strasbourg and the Région Grand Est for financial support, the Platform of Cellular and Tissular Imaging (PICT) at University of Reims Champagne-Ardenne. We also thank Nicole Bouland and Elisabeth Martin for technical support.

REFERENCES

(1) Hermes, M.; Morrish, R.B.; Huot, L.; Meng, L.; Junaid, S.; Tomko, J.; Lloyd, G.R.; Masselink, W.T.; Tidemand-Lichtenberg, P.; Pedersen, C. Mid-IR hyperspectral imaging for label-free histopathology and cytology. *J. Opt.* **2018**, 20, 023002.

(2) Ly, E.; Piot, O.; Durlach, A.; Bernard, P.; Manfait, M. Differential diagnosis of cutaneous carcinomas by infrared spectral micro-imaging combined with pattern recognition. *Analyst*. **2009**, 134, 1208-1214.

(3) Ly, E.; Cardot-Leccia, N.; Ortonne, J.-P.; Benchetrit, M.; Michiels, J.-F.; Manfait, M.; Piot, O. Histopathological characterization of primary cutaneous melanoma using infrared microimaging: a proof-of-concept study. *Br. J. Dermatol*. **2010**, 162, 1316-1323.

(4) Pezzei, C.; Pallua, J.D.; Schaefer, G.; Seifarth, C.; Huck-Pezzei, V.; Bittner, L.K.; Klocker, H.; Bartsch, G.; Bonn, G.K.; Huck, C.W. Characterization of normal and malignant prostate tissue by Fourier transform infrared microspectroscopy. *Mol. BioSyst.* **2010**, 6, 2287-2295.

(5) Akalin, A.; Mu, X.; Kon, M.A.; Ergin, A.; Remiszewski, S.H.; Thompson, C.M.; Raz, D.J.; Diem, M. Classification of malignant and benign tumors of the lung by infrared spectral histopathology (SHP). *Lab. Invest.* **2015**, 95, 406-421.

(6) Steller, W.; Einenkel, J.; Horn, L.-C.; Braumann, U.-D.; Binder, H.; Salzer, R.; Krafft, C. Delimitation of squamous cell cervical carcinoma using infrared microspectroscopic imaging. *Anal. Bioanal. Chem.* **2006**, 384, 145-154.

(7) Krafft, C.; Thümmler, K.; Sobottka, S.B.; Schackert, G.; Salzer, R. Classification of malignant gliomas by infrared spectroscopy and linear discriminant analysis. *Biopolymers*. **2006**, 82, 301-305.

(8) Benard, A.; Desmedt, C.; Smolina, M.; Szternfeld, P.; Verdonck, M.; Rouas, G.; Kheddoumi, N.; Rothé, F.; Larsimont, D.; Sotiriou, C.; Goormaghtigh, E. Infrared imaging in breast cancer: automated tissue component recognition and spectral characterization of breast cancer cells as well as the tumor microenvironment. *Analyst.* **2014**, 139, 1044-1056.

(9) Pounder, F.N.; Reddy, R.K.; Bhargava, R. Development of a practical spatial-spectral analysis protocol for breast histopathology using Fourier transform infrared spectroscopic imaging. *Faraday Discuss.* **2016**, 187, 43-68.

(10) Khanmohammadi, M.; Bagheri Garmarudi, A.; Samani, S.; Ghasemi, K.; Ashuri, A. Application of Linear Discriminant Analysis and Attenuated Total Reflectance Fourier Transform Infrared Microspectroscopy for Diagnosis of Colon Cancer. *Pathol. Oncol. Res.* **2011**, 17, 435-441.

(11) Lasch, P.; Haensch, W.; Naumann, D.; Diem, M. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. Biochim. Biophys. *Acta Mol. Basis Dis.* **2004**, 1688, 176-186.

(12) Nallala, J.; Diebold, M.-D.; Gobinet, C.; Bouché, O.; Sockalingum, G.D.; Piot, O.; Manfait, M. Infrared spectral histopathology for cancer diagnosis: a novel approach for automated pattern recognition of colon adenocarcinoma. *Analyst.* **2014**, 139, 4005-4015.

(13) Kallenbach-Thieltges, A.; Großerüschkamp, F.; Mosig, A.; Diem, M.; Tannapfel, A.; Gerwert, K. Immunohistochemistry, histopathology and infrared spectral histopathology of colon cancer tissue sections. *J. Biophotonics.* **2013**, 6, 88-100.

(14) Vuiblet, V.; Fere, M.; Gobinet, C.; Birembaut, P.; Piot, O.; Rieu, P. Renal Graft Fibrosis and Inflammation Quantification by an Automated Fourier–Transform Infrared Imaging Technique. *J. Am. Soc. Nephrol.* **2016**, 27, 2382-2391.

(15) Eklouh-Molinier, C.; Happillon, T.; Bouland, N.; Fichel, C.; Diébold, M.-D.; Angiboust, J.-F.; Manfait, M.; Brassart-Pasco, S.; Piot, O. Investigating the relationship between changes in collagen fiber orientation during skin aging and collagen/water interactions by polarized-FTIR microimaging. *Analyst.* **2015**, 140, 6260-6268.

(16) Ly, E.; Piot, O.; Wolthuis, R.; Durlach, A.; Bernard, P.; Manfait, M. Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies. *Analyst.* **2008**, 133, 197-205.

(17) Nguyen, T.N.Q.; Jeannesson, P.; Groh, A.; Piot, O.; Guenot, D.; Gobinet, C. Fully unsupervised inter-individual IR spectral histology

of paraffinized tissue sections of normal colon. *J. Biophotonics.* **2016**, 9, 521-532.

(18) Nguyen, T.N.Q.; Jeannesson, P.; Groh, A.; Guenot, D.; Gobinet, C. Development of a hierarchical double application of crisp cluster validity indices: a proof-of-concept study for automated FTIR spectral histology. *Analyst.* **2015**, 140, 2439-2448.

(19) Nallala, J.; Gobinet, C.; Diebold, M.D.; Untereiner, V.; Bouché, O.; Manfait, M.; Sockalingum, G.D.; Piot, O. Infrared spectral imaging as a novel approach for histopathological recognition in colon cancer diagnosis. *J. Biomed. Opt.* **2012**, 17, 116013.

(20) Wolthuis, R.; Travo, A.; Nicolet, C.; Neuville, A.; Gaub, M.-P.; Guenot, D.; Ly, E.; Manfait, M.; Jeannesson, P.; Piot, O. IR Spectral Imaging for Histopathological Characterization of Xenografted Human Colon Carcinomas. *Anal. Chem.* **2008**, 80, 8461-8469.

(21) de Lima, F.A.; Gobinet, C.; Sockalingum, G.D.; Garcia, S.B.; Manfait, M.; Untereiner, V.; Piot, O.; Bachmann, L. Digital de-waxing on FTIR images. *Analyst.* **2017**, 142, 1358-1370.

(22) Afseth, N.K.; Kohler, A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemom. Intell. Lab. Syst.* **2012**, 117, 92-99.

(23) Patel, I.I.; Trevisan, J.; Singh, P.B.; Nicholson, C.M.; Krishnan, R.K.G.; Matanhelia, S.S.; Martin, F.L.. Segregation of human prostate tissues classified high-risk (UK) versus low-risk (India) for adenocarcinoma using Fourier-transform infrared or Raman microspectroscopy coupled with discriminant analysis. *Anal. Bioanal. Chem.* **2011**, 401, 969-982.

(24) Ali, S.M.; Bonnier, F.; Lambkin, H.; Flynn, K.; McDonagh, V.; Healy, C.; Lee, T.C.; Lyng, F.M.; Byrne, H.J. A comparison of Raman, FTIR and ATR-FTIR micro spectroscopy for imaging human skin tissue sections. *Anal. Methods.* **2013**, 5, 2281-2291.

(25) Fernandez, D.C.; Bhargava, R.; Hewitt, S.M.; Levin, I.W. Infrared spectroscopic imaging for histopathologic recognition. *Nat. Biotechnol.* **2005**, 23, 469-474.

(26) Pallua, J.D.; Pezzei, C.; Zelger, B.; Schaefer, G.; Bittner, L.K.; Huck-Pezzei, V.A.; Schoenbichler, S.A.; Hahn, H.; Kloss-Brandstaetter, A.; Kloss, F.; Bonn, G.K.; Huck, C.W. Fourier transform infrared imaging analysis in discrimination studies of squamous cell carcinoma. *Analyst.* **2012**, 137, 3965-3974.

(27) Travo, A.; Piot, O.; Wolthuis, R.; Gobinet, C.; Manfait, M.; Bara, J.; Forgue-Laffite, M.-E.; Jeannesson, P. IR spectral imaging of secreted mucus: a promising new tool for the histopathological recognition of human colonic adenocarcinomas. *Histopathology.* **2010**, 56, 921-931.

(28) Kéry, M.; Royle, J.A. Introduction to Data Simulation. *Applied Hierarchical Modeling in Ecology*. **2016**, 1, 123-143

(29) Lieber, C.A.; Mahadevan-Jansen, A. Automated Method for Subtraction of Fluorescence from Biological Raman Spectra. *Appl. Spectrosc.* **2003**, 57, 1363-1367.

(30) Lasch, P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom. Intell. Lab. Syst.* **2012**, 117, 100-114.

(31) Nallala, J.; Lloyd, G.R.; Stone, N. Evaluation of different tissue de-paraffinization procedures for infrared spectral imaging. *Analyst.* **2015**, 140, 2369-2375.

(32) MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. **1967**, 1, 281-97.

(33) Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J.M.; Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognit.* **2013**, 46, 243-256.

(34) Oliveira, F.P.M.; Tavares, J.M.R.S. Medical image registration: a review. *Comput. Methods Biomech. Biomed. Engin.* **2014**, 17, 73-93.

(35) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, 9, 2579-2605.

(36) Cao, Y.; Wang, L. Automatic Selection of t-SNE Perplexity. *arXiv*. **2017**, 1708.03229

(37) Vorontsov, I.E.; Kulakovskiy, I.V.; Makeev, V.J. Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms Mol. Biol.* **2013**, 8, 23.

(38) Hill, P.D. Kernel estimation of a distribution function. *Commun. Stat. Theory Methods.* **1985**, 14, 605-620.

(39) Nallala, J.; Lloyd, G.R.; Hermes, M.; Shepherd, N.; Stone, N. Enhanced spectral histology in the colon using high-magnification benchtop FTIR imaging. *Vib. Spectrosc.* **2017**, 91, 83-91.

(40) Trukhan, S.; Tafintseva; V.; Tøndel, K.; Großerueschkamp, F.; Mosig, A.; Kovalev, V.; Gerwert, K.; Kohler, A. Grayscale representation of infrared microscopy images by extended multiplicative signal correction for registration with histological images. *J. Biophotonics.* **2020**, e201960223.

For Table of Contents Only