Supporting Information

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: cyril.gobinet@univ-reims.fr

Contents

1) Description of the generative model of simulated spectral images

In order to validate our approach, we constructed simulated FTIR spectral images of FFPE tissue sections where the i[th] spectrum $\boldsymbol{s_i}$ is modelled using the following linear model:

$$\boldsymbol{s_i} = \alpha_i \boldsymbol{t_i} + \beta_i \boldsymbol{p_i} + \boldsymbol{l_i} + \sigma \boldsymbol{n_i} \quad (1)$$

$\boldsymbol{t_i}$ is a spectrum randomly selected in a database composed of 2878 real FTIR spectra acquired in the tissue area of a frozen section from a xenografted human colon carcinoma. In this database, the spectra were previously corrected from the baseline using the Lieber-Mahadevan-Jansen polynomial method[30] using a 4th order polynomial function to model the baseline and normalized by Min-Max normalization. $\boldsymbol{p_i}$ is a spectrum randomly selected in a database composed of 12544 real FTIR spectra acquired in a pure paraffin area at the periphery of a human colon carcinoma FFPE section. Pure paraffin FTIR spectra being almost not distorted by a baseline, the spectra composing this database were not previously corrected from it. $\boldsymbol{l_i}$ is a baseline randomly selected in a database composed of 10905 baselines estimated by the Lieber-Mahadevan-Jansen polynomial method[30] on real FTIR spectra acquired in the tissue area of a human colon carcinoma FFPE section. $\boldsymbol{n_i}$ is a random noise vector generated using a standard normal distribution, i.e. with zero mean and unit standard deviation. α_i and β_i are the contributions of tissue and paraffin spectra respectively, which were generated by respecting the following constraints in order to approximate the topography of a real FFPE tissue section. First, $0 \leq \alpha_i \leq 1$, $0.5 \leq \beta_i \leq 1$ and $\beta_i = 1 - \alpha_i/2$ in order to limit the range of spectral intensities and to ensure the presence of a paraffin signal in each simulated spectrum. Second, the simulated FFPE tissue area is surrounded by a simulated pure paraffin area for which $\beta_i = 1$, and thus $\alpha_i = 0$ according to the second constraint. Third, the α_i coefficients in the tissue part of the simulated images were generated using the distribution of the fitting coefficient of the mean spectrum of a real human colon carcinoma FFPE section as estimated by the EMSC method (section entitled “Data pre-processing”). Fourth, these simulated α_i coefficients were sorted in ascending order, with the smallest values on the outer part of the tissue area and the highest on the inner part, in order to create a gradient of tissue. σ is the standard deviation of the Gaussian noise used to control the signal to noise ratio.

2) Parameter setting of the generative model of simulated spectral images

In order to evaluate the performance of our method described in section entitled “Multivariate analysis of EMSC fitting coefficients”, a total of 30 simulated spectral images were generated according to the procedure described in the previous section and using the following parameter setting. The simulated FFPE tissue section acquisition area corresponds to a square with 30-pixel sides, including a tissue area represented by an internal square with 20-pixel sides. The remaining outer pixels are thus paraffin pixels. This simulated sample topology is considered as the ground truth labels for the simulated spectra (Figure S-6(a)). Following this sample topology, the contributions of tissue and paraffin spectra were generated respecting the four constraints (Figure S-6(b-c)). More precisely, the distribution of the EMSC regression coefficient a (Figure S-6(d)) originating from a real FFPE tissue section was used to simulate the tissue contribution respecting the third constraint. Depending on the experiment, the order of the polynomial function varied between 0 and 4 to simulate the baseline, and the SNR varied between 1 and 1000. As examples, Figures S-6(e-g) show some simulated spectra at different locations in the tissue area using a noise-free simulation model with a first-order polynomial function for the baseline. A wide variability in paraffin, tissue and baseline contributions is clearly visible on these spectra. In particular, Figure S-6(e) shows the spectra simulated for the three pixels labelled as tissue pixels (rectangle A on Figure S-6(b)), but with a very weak contribution of tissue ($\alpha < 0.1$ and $\beta > 0.95$ in the simulated model). All the simulated spectra are depicted on Figure S-6(h).

3) Supplementary figures

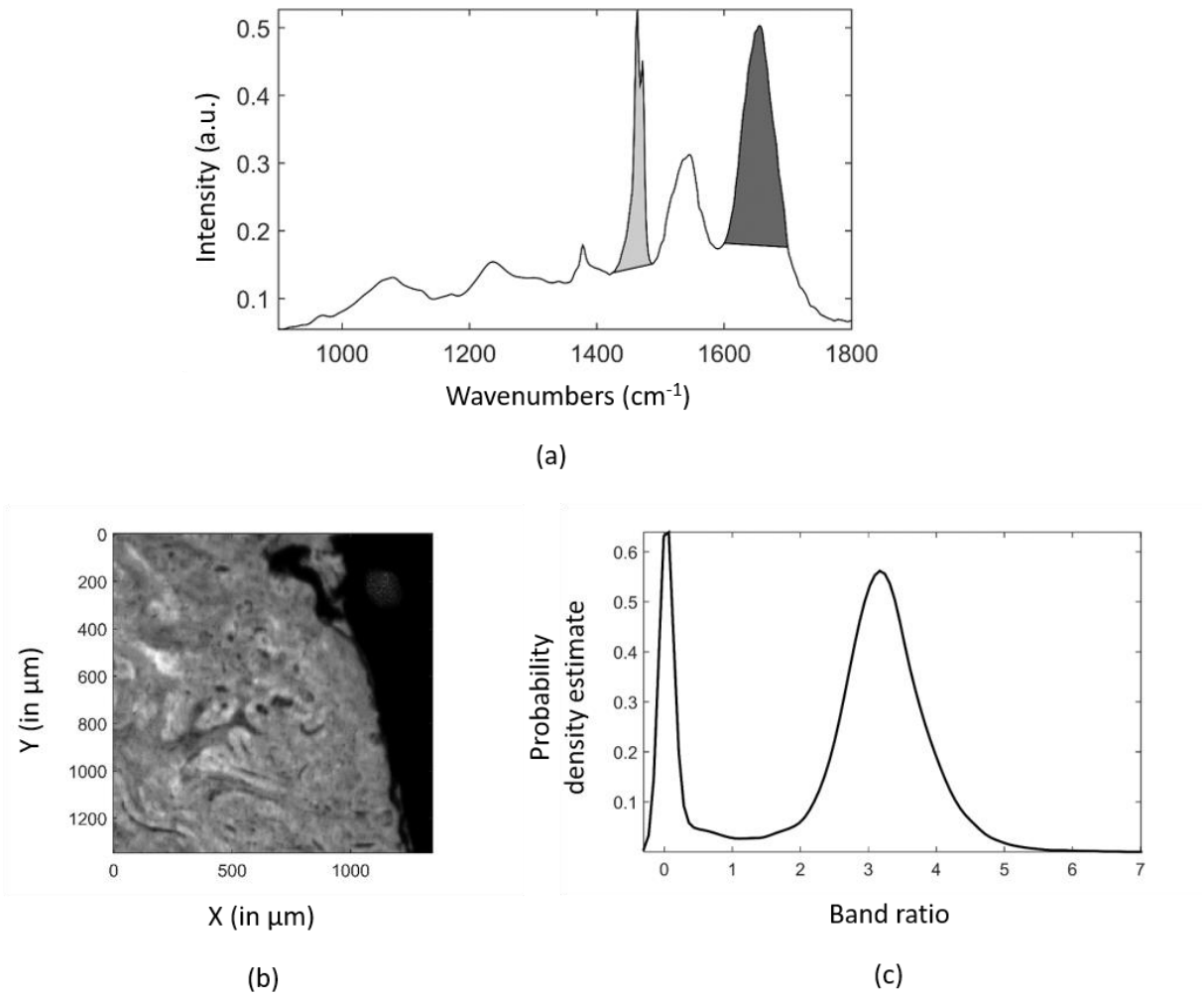


Figure S-1: Spectral band ratio computation. (a) Mean spectrum of a spectral image acquired on an FFPE tissue section. The light and dark gray areas represent the integrated paraffin (1430–1490 cm^{-1}) and Amide I (1600–1700 cm^{-1}) bands, respectively, used to compute the spectral band ratio. (b) Grayscale image reconstructed from the computed spectral band ratio values. (c) Estimated probability density of spectral band ratio.

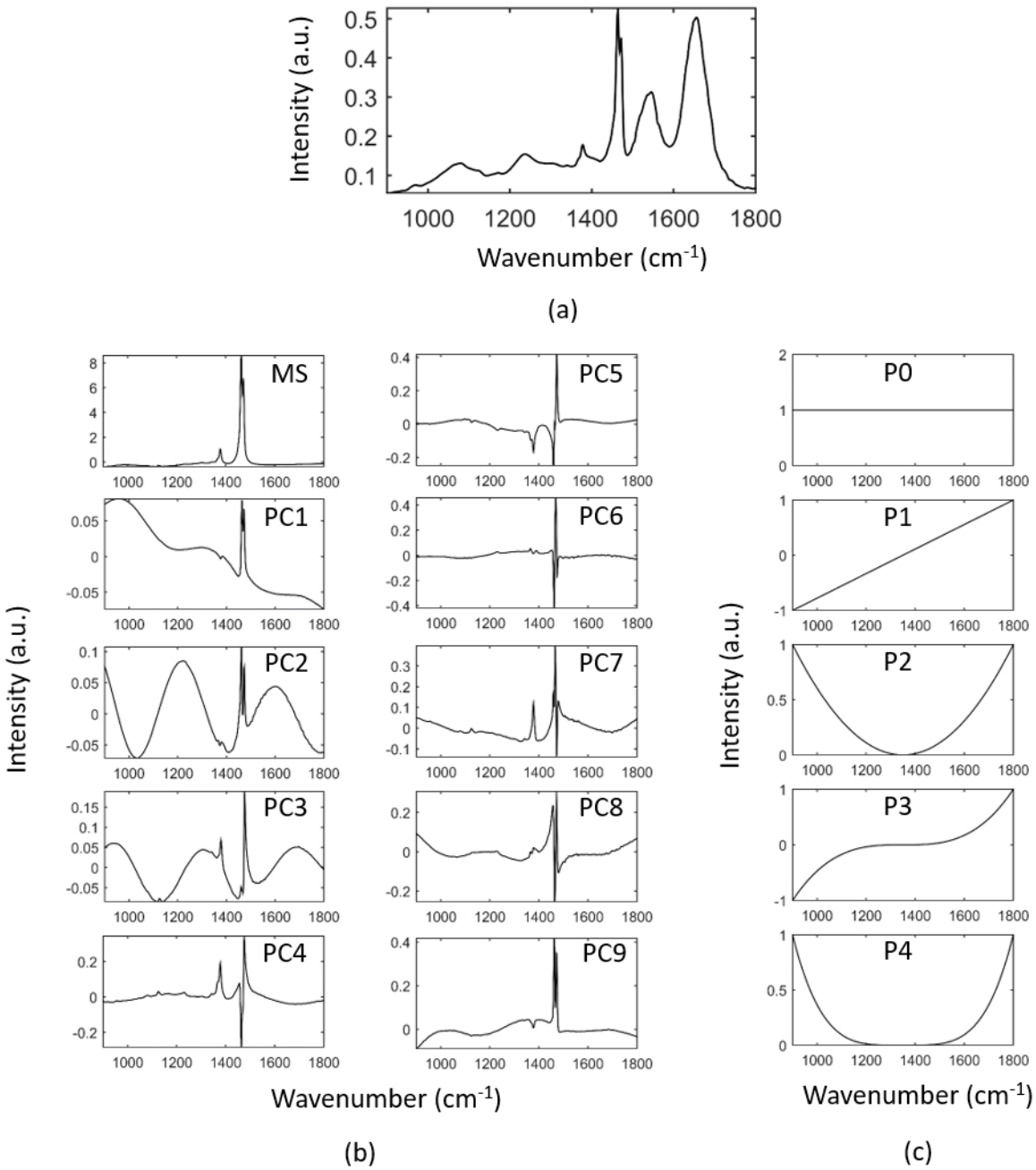


Figure S-2: Components of the EMSC model applied on a FTIR spectral image acquired on a human colon carcinoma FFPE sample. (a) The reference spectrum $\hat{\boldsymbol{s}}$ computed as the mean image spectrum. (b) The interference matrix $\boldsymbol{I}$ composed of the mean spectrum (MS) of a FTIR spectral image acquired on a pure paraffin area, and of the 9 first principal components (PC1 to PC9) computed on this pure paraffin spectral image in order to model the spectral variabilities of paraffin, such as maximum peak position, peak width, etc. (c) The Vandermonde matrix $\boldsymbol{P}$ composed of polynomial functions of order 0 to 4 (P0 to P4).

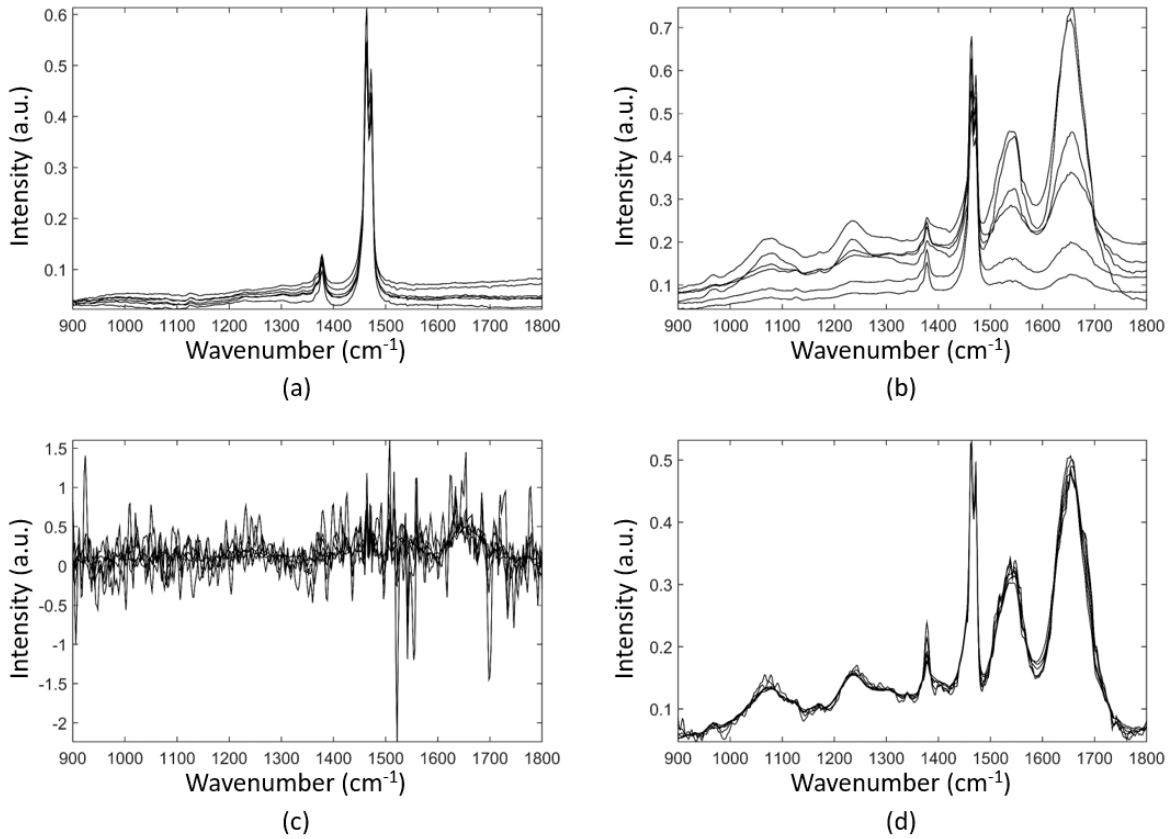
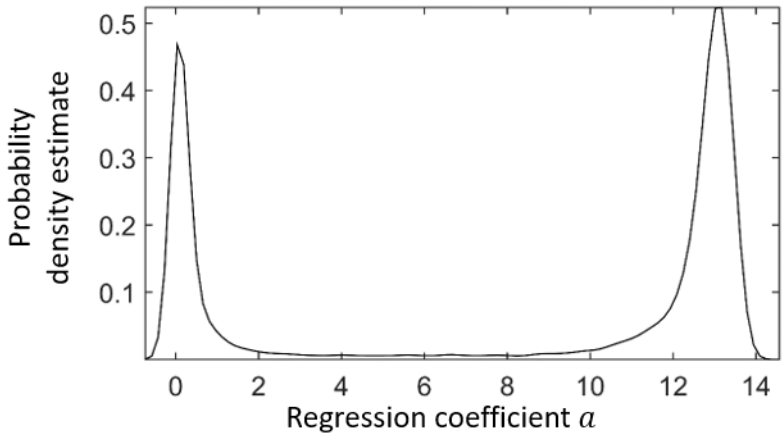
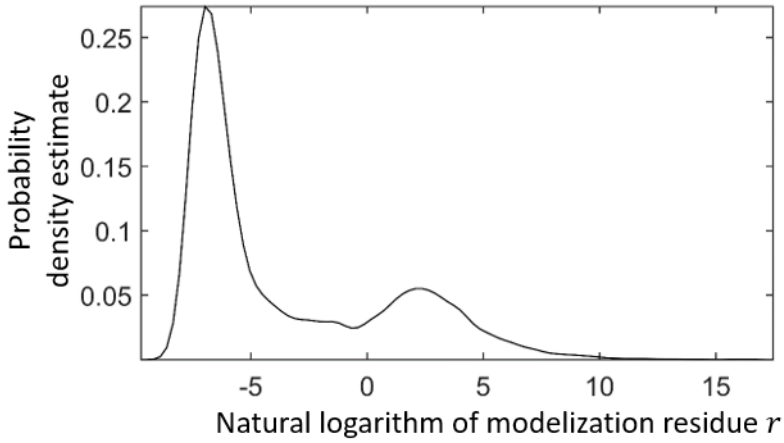


Figure S-3: Examples of application of the EMSC model to the FTIR spectral image acquired on a human colon carcinoma FFPE sample used throughout the paper as an illustrative example. Examples of raw spectra acquired on the paraffin (a) and FFPE tissue (b) areas of this sample. The same spectra after EMSC pre-processing, i.e. EMSC pre-processed spectra of paraffin (c) and FFPE tissue (d).



(a)



(b)

Figure S-4: The estimated probability densities of the regression coefficient a (a) and the natural logarithm of modelization residue r (b).

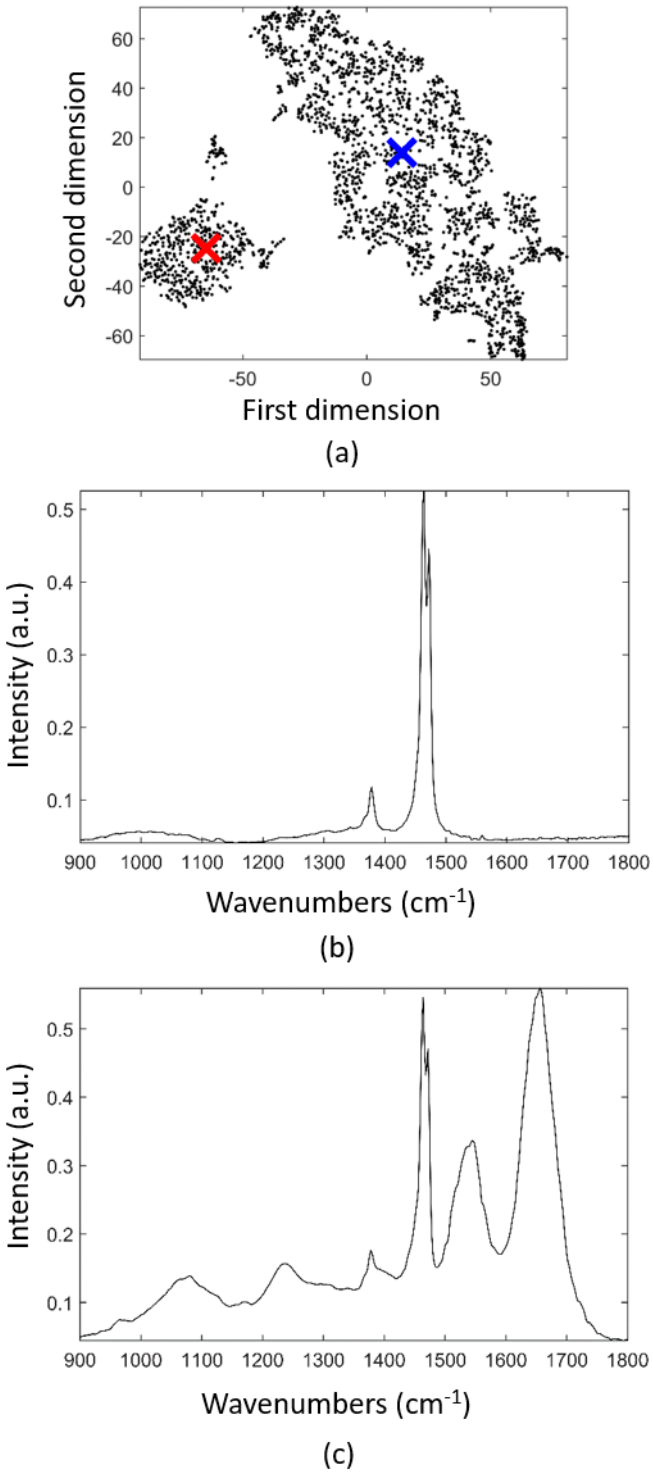


Figure S-5: Two data groups as revealed by t-SNE applied on the EMSC regression coefficients of data acquired on a human colon carcinoma FFPE section. (a) Scatter plot of a two-dimensional t-SNE. The cluster visible at the bottom left of the figure corresponds to the pure paraffin spectra, while the one at the top and at the right is typical of the tissue spectra. (b) The paraffin spectrum corresponding to the t-SNE point as marked by a red cross on (a). (c) The FFPE tissue spectrum corresponding to the t-SNE point as marked by a blue cross on (a).

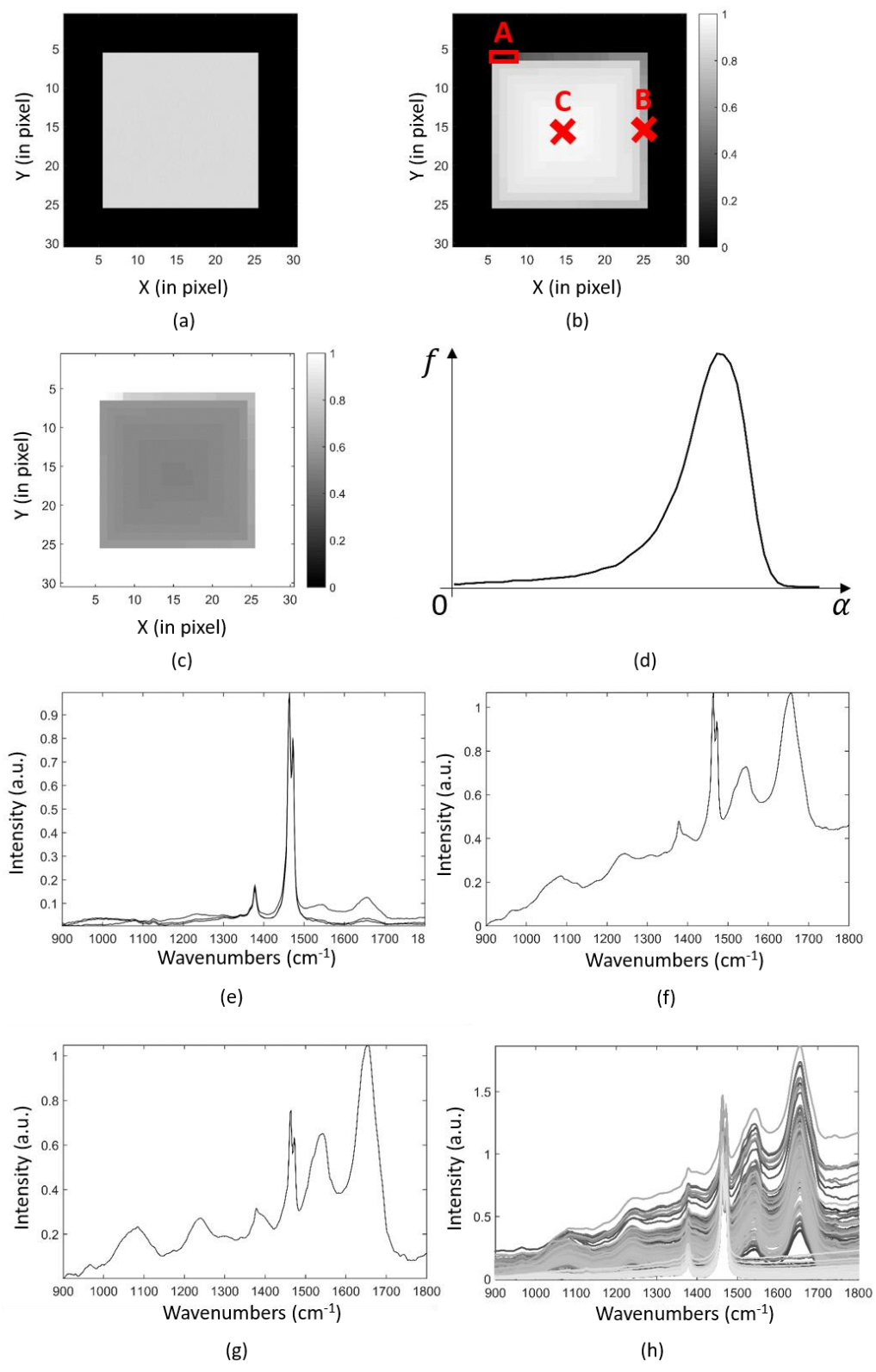


Figure S-6: An example of a simulated FTIR image acquired on a FFPE tissue section. (a) Topology of the simulated sample where black and gray pixels represent paraffin and tissue pixels respectively. (b) The spatial map of the simulated tissue concentration coefficients. (c) The spatial map of the simulated paraffin concentration coefficients. Note that on panels (a-c) a pure paraffin area is surrounding a FFPE tissue part. (d) A real distribution used to simulate the tissue concentration coefficients. (e-g) Examples of simulated spectra of pixels annotated A, B and C on panel (b). (h) The complete simulated dataset.

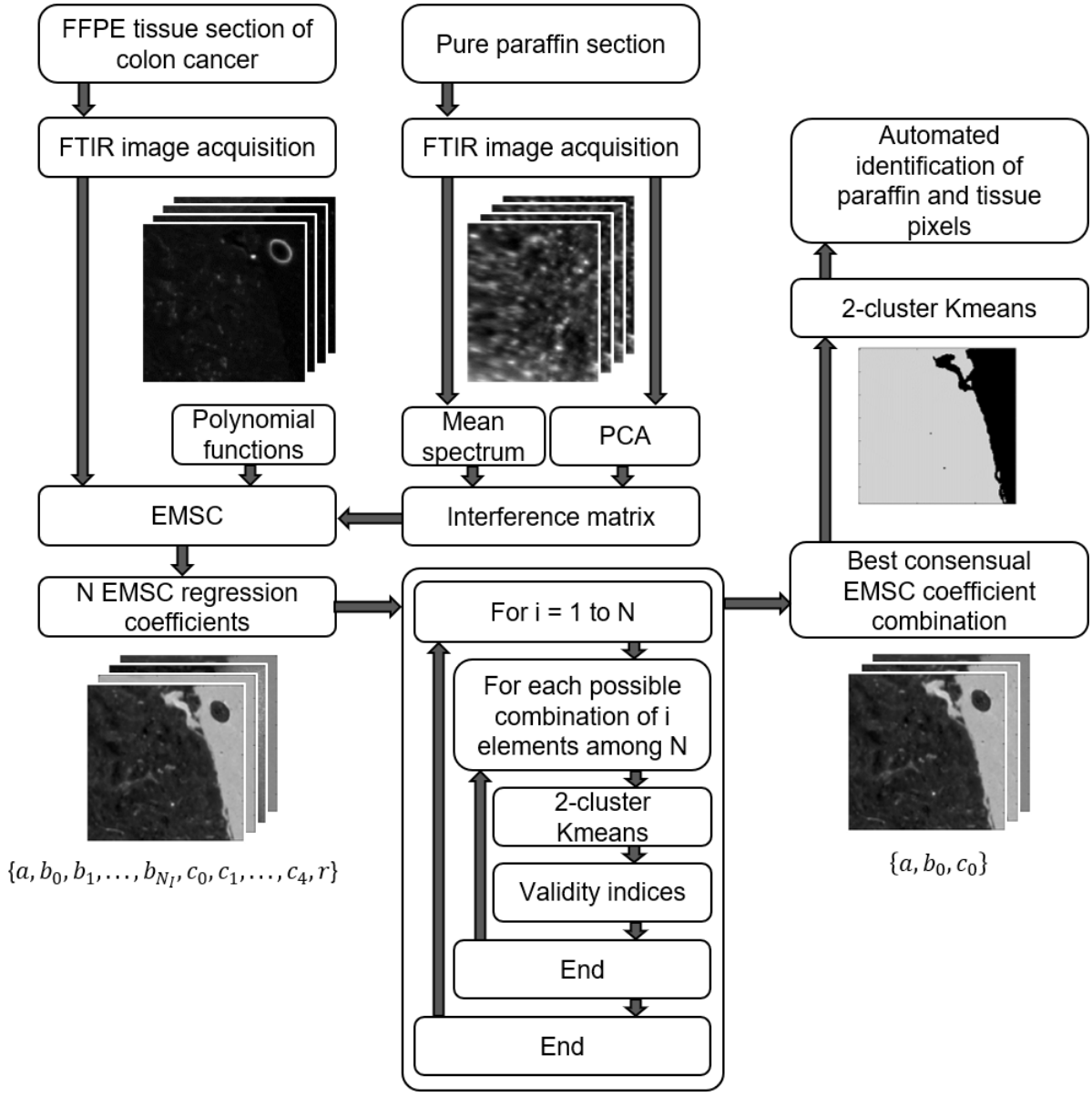


Figure S-7: Workflow of the proposed method. A FTIR image is acquired on a pure paraffin section in order to compute the mean paraffin spectrum and the principal components loadings which describe the most important sources of variation in the paraffin spectra. These components are injected into the interference matrix of the EMSC model. An example of these components if provided on Figure S-2(b). Then, a FTIR image is acquired on a FFPE tissue section which is preprocessed by EMSC as described in the experimental section. The EMSC regression coefficients are injected into an algorithm in order to determine, for each possible combination of these coefficients, the values of validity indices applied on 2-cluster KMeans partitions. The combination leading to the best consensual value of validity indices is used to run a 2-cluster KMeans in order to automatically identify the paraffin and tissue pixels.

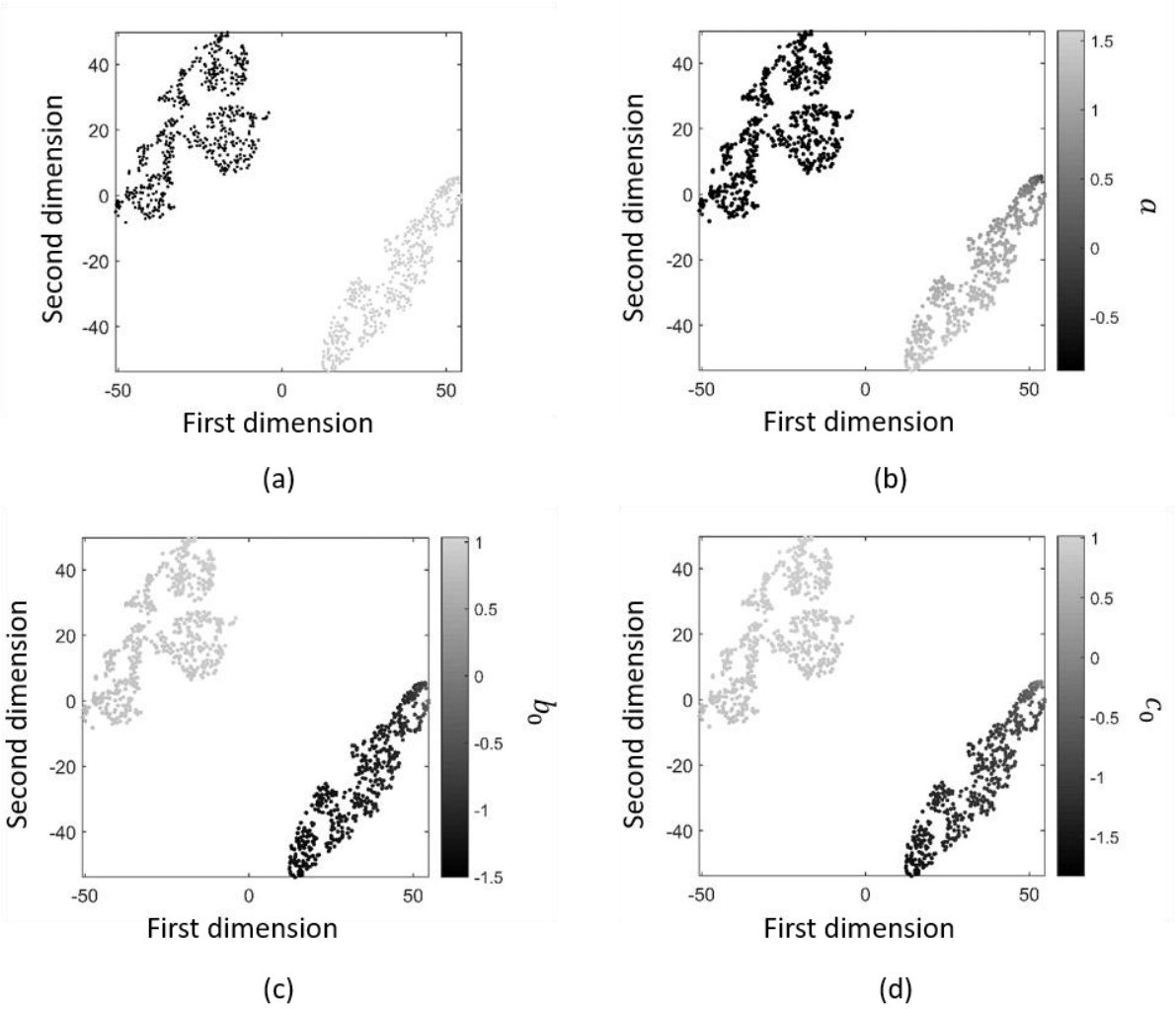


Figure S-8: Scatter plots of a two-dimensional t-SNE applied to the SNV-normalized EMSC regression coefficients $\{a, b_0, c_0\}$ estimated from simulated data with a first-order polynomial function, with point colors defined by (a) the pixel true labels (black and gray pixels correspond to paraffin and tissue pixels, respectively), (b) the a regression coefficient value, (c) the b_0 regression coefficient value, (d) the c_0 regression coefficient value.

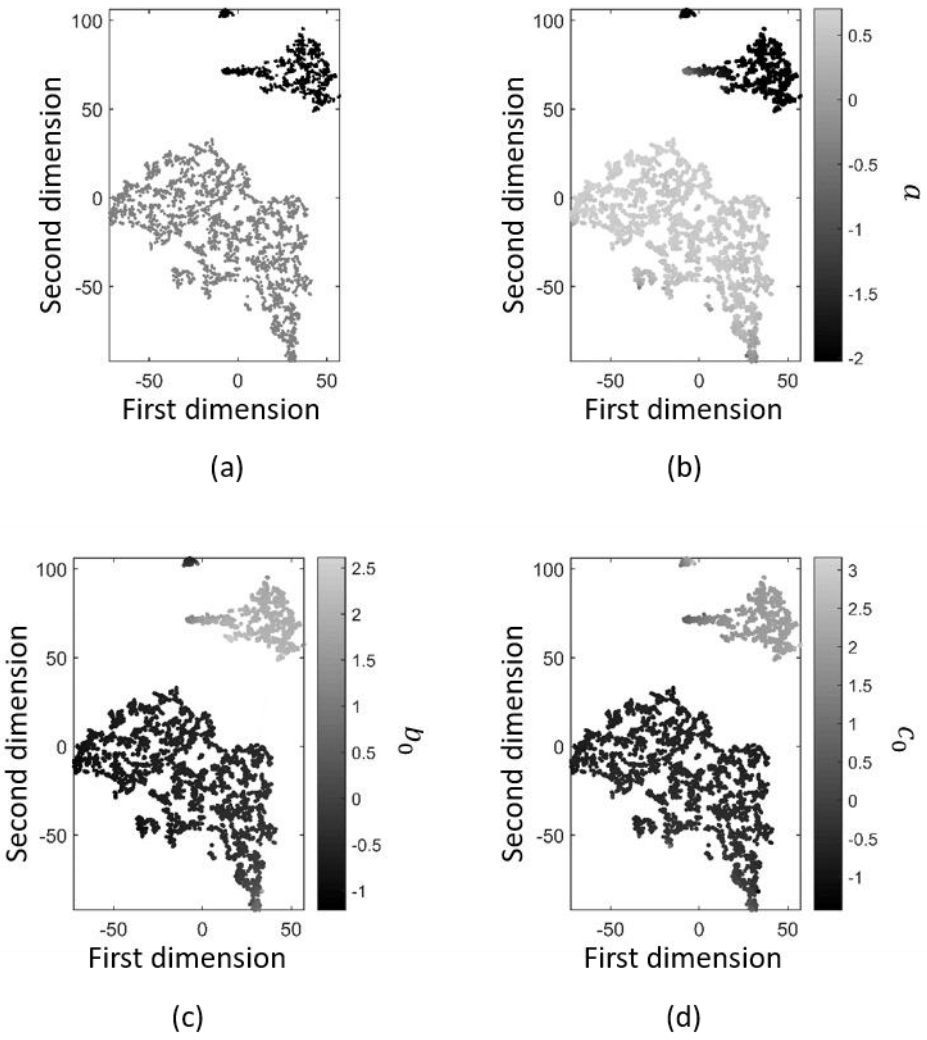


Figure S-9: Scatter plots of a two-dimensional t-SNE applied on the SNV-normalized EMSC regression coefficients $\{a,\ b_0,\ c_0\}$ estimated from a human colon cancer FFPE tissue section with point colors defined by (a) the pixel labels estimated by our proposed multivariate approach (black and gray pixels correspond to paraffin and tissue pixels, respectively), (b) the a regression coefficient value, (c) the b_0 regression coefficient value, (d) the c_0 regression coefficient value.

4) Supplementary tables

Table S-1: The top 5 combinations of SNV-normalized EMSC regression coefficients estimated on a simulated spectral image with a first-order polynomial function for four different validity indices (XB, DB, PBM, SWC) computed on their two-cluster KMeans partitions.

Rank\Validity index	**XB**	**DB**	**PBM**	**SWC**
1st	b_0	b_0	a, b_0, c_0	a, b_0
2nd	a, b_0	a	a, b_0	a, b_0, c_0
3rd	a	a, b_0	b_0, c_0	b_0, c_0
4th	b_0, c_0	b_0, c_0	a, c_0	a, c_0
5th	a, b_0, c_0	a, b_0, c_0	a, r, b_0, c_0	a, r, b_0

Table S-2: The top 5 combinations of SNV-normalized EMSC regression coefficients estimated on a FTIR image acquired on a human colon cancer sample for four different validity indices (XB, DB, PBM and SWC) computed on their two-cluster KMeans partitions.

Rank\Validity index	**XB**	**DB**	**PBM**	**SWC**
1st	a	a	a, c_0	a, c_0
2nd	a, c_0	a, c_0	a, b_0, c_0	a, b_0, c_0
3rd	a, b_0, c_0	a, b_0, c_0	a, r, b_0, c_0	a, r, b_0, c_0
4th	a, b_0	a, b_0	a, b_0	a, b_0
5th	a, r, b_0, c_0	a, r, b_0, c_0	a, r, c_0	a, c_3