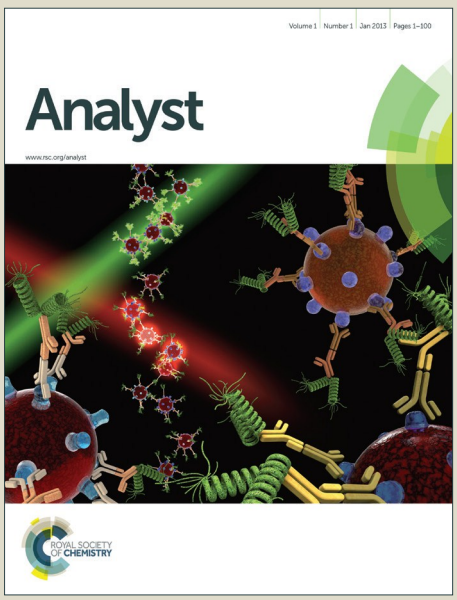
Analyst

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

FTIR imaging allows automated identification and quantification of breast tumor cells as well as investigating tumor-related stroma alterations

Characterization of human breast cancer tissues by infrared imaging

M. Verdonck[1], A. Denayer[1], B. Delvaux[1], S. Garaud[2], R. De Wind[3], C. Desmedt[4], C. Sotiriou[4], K. Willard-Gallo[2] and E. Goormaghtigh[1]

[1] Laboratory of Structure and Function of Biological Membranes, Center of Structural Biology and Bioinformatics, Université Libre de Bruxelles, Brussels, Belgium;

[2] Molecular Immunology Unit, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium;

[3] Pathological Anatomy Department, Institut Jules Bordet, Brussels, Belgium;

[4] Breast Cancer Translational Research Laboratory, Institut Jules Bordet, Brussels, Belgium.

Keywords: Infrared spectroscopy; FTIR imaging; Breast cancer; Stroma

Abbreviations:

FPA, Focal Plane Array; FFPE, Formalin-Fixed (Fixation) Paraffin-Embedded (Embedding); FN, False Negative; FP, False Positive; FTIR, Fourier Transform InfraRed; H&E, Hematoxylin and Eosin; IR, InfraRed; MANOVA, Multivariate ANalysis Of VAriance; LOOCV, Leave-One-Out Cross Validation; MCC, Matthews Correlation Coefficient; MCT, Mercury Cadmium Telluride; NANT, Non-Adjacent Non-Tumor; PC, Principal Component; PCA, Principal Component Analysis; PLS-DA, Partial Least Square Discriminant Analysis; S/N, Signal-to-Noise; TN, True Negative; TP, True Positive

Correspondence to:

Dr. Magali Verdonck

Laboratory of Structure and Function of Biological Membranes, Center of Structural Biology and Bioinformatics, CP 206/02, Université Libre de Bruxelles, Campus Plaine, Bd du Triomphe, accès 2, 1050 Brussels, Belgium.
tel.: +32-2-650-53-65 fax: +32-2-650-53-82 e-mail: mverdonc@ulb.ac.be

ABSTRACT

Fourier Transform InfraRed (FTIR) spectroscopy coupled to microscopy (IR imaging) has shown unique advantages in detecting morphological and molecular pathologic alterations in biological tissues. The aim of this study was to evaluate the potential of IR imaging as diagnostic tool to identify characteristics of breast epithelial cells and the stroma. In this study a total of 19 breast tissue samples were obtained from 13 patients. For 6 of the patients, we also obtained Non-Adjacent Non-Tumor tissue samples. Infrared images were recorded on the main cell/tissue types identified in all breast tissue samples. Unsupervised Principal Component Analyses and supervised Partial Least Square Discriminant Analyses (PLS-DA) were used to discriminate spectra. Leave-one-out cross-validation was used to evaluate the performance of PLS-DA models. Our results show that IR imaging coupled with PLS-DA can efficiently identify the main cell types present in FFPE breast tissue sections, *i.e.* epithelial cells, lymphocytes, connective tissue, vascular tissue and erythrocytes. A second PLS-DA model could distinguish normal and tumor breast epithelial cells in the breast tissue sections. A patient-specific model reached particularly high sensitivity, specificity and MCC rates. Finally, we showed that the stroma located close or at distance from the tumor exhibits distinct spectral characteristics. In conclusion FTIR imaging combined with computational algorithms could be an accurate, rapid and objective tool to identify/quantify breast epithelial cells and differentiate tumor from normal breast tissue as well as normal from tumor-associated stroma, paving the way to the establishment of a potential complementary tool to ensure safe tumor margins.

INTRODUCTION

Breast cancer is the most common carcinoma in women worldwide. This epithelial tumor accounts for 23% of all cancers in women globally and 27% in affluent countries, which is more than twice as common as cancer at any other site.[1,2] The incidence of breast cancer varies 10-fold depending on the geographical location; in Europe 430 000 new breast cancer cases are annually diagnosed.[3,1] Prognosis and survival rates vary greatly depending on the tumor type, size, grade, stage, lymph node and hormonal receptor status as well as the treatment of the patient.[4] The prognosis for patients with breast cancer is very good if detected at an early stage.[1] Therefore, early detection and accurate diagnosis remain a cornerstone to improve breast cancer patient outcome and survival.[4] Most patients with early breast cancer undergo as the first step of treatment, surgical intervention. The challenge is to remove the primary tumor with clear (negative) margins. Accurate and rapid assessment of the margins safety is essential since a positive margin leads to a re-excision procedure which may have negative consequences, including prolonged wound healing, post-operative infection, patient anxiety and poor cosmetic outcome.[3] Margin status is therefore a prognostic factor for in-breast recurrence.[5,6,3] Nonetheless, a clear definition of what constitutes an adequate histological margin of excision has been the subject of much debate.[3] Currently the absence or presence of tumor cells is assessed routinely by visual inspection of a hematoxylin and eosin (H&E) stained tissue section. Indication of changes in the microscopic morphology and composition of the breast tissue performed by pathologists is still regarded as the gold standard for the assessment of breast lesions.

Fourier Transform InfraRed (FTIR) technology shows potential as a new promising diagnostic method to detect morphological and molecular pathologic alterations in biological tissues.[7–13,4] While infrared (IR) spectroscopy reflects the chemical composition of the tissue, combination with microscopy (IR imaging) allows spatially resolved examination of complex tissues and heterogeneous samples.[14,15] Previous studies have demonstrated the possibility to discriminate normal and tumor tissue in breast,[16–19,15] colon,[20–22,15] lung,[23,15] prostate[24–27,15] and cervix[28–30,15] using IR spectroscopy. This technique can screen human tissues without any reagent, requires minimal sample preparation, is time-saving and non-destructive. Moreover this method provides spectral fingerprints which allow subsequent classification of spectra into different categories with objective computational methods.[31–35,15,4]

As recently described,[36] the role of tumor cell microenvironment appears now as critical. The stroma provides support for tumor growth, progression, invasion and metastasis.[37,38,36] Cellular tissue network involves epithelial cells, fibroblasts, myofibroblasts, endothelial cells, adipocytes and immune cells whereas the extracellular matrix, also called stroma includes collagen, laminin, metalloproteinases and fibronectin which constitute the connective tissue.[39,40] Significant differences between normal and reactive stroma have been demonstrated.[41,42] The so-called reactive stroma can determine invasiveness properties of carcinoma cells.[43,38] It was therefore of interest to evaluate whether these differences could be identified by infrared imaging and could be of potential diagnostic interest.

The purpose of this study is to assess the possibility of developing a rapid and automated tool to identify breast tumor cells (i), to discriminate tumor from normal tissue based on the IR spectral features of breast epithelial cells (ii) and to assess the tumor microenvironment (iii). More precisely, we focused on discriminating breast tumor cells and non-adjacent non-tumor (NANT) cells from the margin, referred here to as normal breast epithelial cells.

MATERIALS AND METHODS

Tissue samples selection and handling

The experimental workflow is schematically depicted in Figure 1. Formalin-Fixed Paraffin-Embedded (FFPE) tissue samples from 13 patients with breast cancer were included in this study. For 6 of the patients, non-adjacent non-tumor (NANT) tissue samples were also selected. NANT tissues are healthy breast samples located at distance from the tumor (margin). Free-tumor assessment of the NANT tissues has been double-checked by a pathologist. For better readability in this study NANT tissues are referred below to as 'normal' tissues.

For each FFPE breast tissue sample, 2 adjacent tissue sections of 4 µm were cut using a microtome. One section was deposited on a glass slide, the other on a BaF_2 slide. After 12 hours of desiccation in a 37°C-incubator, paraffin was removed from both tissue sections by incubation in 2 successive xylene baths for 20 minutes. Tissue rehydration was achieved through 3 successive ethanol baths with a decreasing gradient of ethanol (100%, 90%, 70%) for 15 minutes and 2 milliQ water baths for 10 minutes. H&E staining was performed on the tissue section deposited on the glass slide; the section on the BaF_2 slide remained unstained to perform infrared measurements.

Study Approval. All breast tissues used in this study were taken from individuals treated at the Institute Jules Bordet. All patients signed a protocol-specific consent. The protocols used for human studies were approved by the Medical Ethics Committee of the Institut Jules Bordet.

Spectral data acquisition

The IR data were collected using a Hyperion 3000 Fourier-Transform InfraRed (FTIR) imaging system (Bruker Optics, Ettlingen, Germany), equipped with a liquid nitrogen cooled 64×64 Mercury Cadmium Telluride Focal Plane Array (FPA) detector in transmission mode. The size of an image covered an area of 180 × 180 μm^2. Each unit image is composed of 4096 pixels of 2.8 × 2.8 μm^2. Each of the 4096 detectors records one IR spectrum on the whole spectral range from 3795 to 900 cm^{-1}. Recording one spectrum/pixel as the mean of 256 scans at a spectral resolution of 8 cm^{-1} took approximately 5 minutes. It should be noted that spatial resolution can be significantly lower than the pixel size, depending on the wavelength.

The H&E adjacent tissue sections were scanned using a digital slide scanner (NanoZoomer 2.0, Hamamatsu Photonics, Japan) in bright field mode (Magnitude 40x). Scans were viewed using the NDP.view software (Hamamatsu Photonics, Japan) and IR measurement areas were indicated by colored squares on the scan (Figure 2). IR measurements were recorded on the unstained tissue section with the H&E section scan used as a reference.

Our spectral database included 10 IR images recorded on each of the 5 main cell types and tissue components from the 19 breast tissue sections. The 5 main cell/tissue types in breast tissue include epithelial cells (tumor or normal), lymphocytes, erythrocytes, connective tissue (stroma) and vascular tissue. It must be noted that not all cell/tissue types were present on each tissue section. As some cell types did not cover the whole imaged area, spectra corresponding specifically to one cell type were selected by manual pixel selection.

Data analysis

Preprocessing

All spectra were preprocessed as follows: water vapor contribution was subtracted as previously described [44,45] with 1956-1935 cm^{-1} as reference peak and the CO_2 peak flattened between 2450 and 2250 cm^{-1}. The spectra were baseline-corrected. Straight lines were interpolated between the spectral points at 3620, 2995, 2800, 2395, 2247, 1765, 1724, 1480, 1355, 1144 and 950 cm^{-1} and subtracted from each spectrum. Spectra were normalized for

equal area between 1725 and 1481 cm^{-1} (Amide I and II peaks). The Signal-to-Noise ratio (S/N) was then checked on each spectrum. It was required to be higher than 300 with noise defined as the standard deviation in the 2000-1900 cm^{-1} region of the spectrum and signal defined as the maximum of the curve between 1750 and 1480 cm^{-1} after subtracting a baseline passing through these two points. Some rare spectra with normalized absorbance lower than -5 (negative lobe) and a maximum above 120 (saturation) were discarded.

Statistical analyses

Several mean spectra were computed for each image and each cell type. Each mean spectrum was the mean of all spectra or a subset of spectra recorded on one cell type image as detailed for each analysis. A spectrum that had been selected once for computing a mean spectrum was not allowed to be selected again for the computation of another mean spectrum.

In order to observe the intrinsic proximities and distances within the data set and to group IR spectra according to their similarity, Principal Component Analyses (PCA) were performed. Mean centering was applied before computing the PCA. Spectra with scores beyond 2 times the standard deviation according to PC1 were considered as outliers and discarded from further analyses. Multivariate ANalyses Of VAriance (MANOVA) were run on the spectra using for each spectrum its scores on the first 6 PCs, to evaluate the chance that the mean value of one cell population would be equal to the mean value of another cell population (p-value).

Difference spectra were produced to allow emphasizing the spectral variations between two distinct groups. These spectra were built by subtracting the mean spectrum of one group from the mean spectrum of a second group. Student t-tests were obtained at every wavenumber from a comparison of the distribution of the absorbance in the two groups. Results were reported on the spectrum obtained as the difference of the means for the two groups as a black star where the equality of the means was rejected with $p<0.01$. When the result of the t-test rejects the equality of the means, it indicates that the difference between groups is larger than variability within group.

Partial Least Square Discriminant Analysis (PLS-DA) was performed to extract latent variables that enable the construction of a factor capable of predicting a class. Bootstrapping

models and leave-one-out cross-validation (LOOCV) models were achieved using PLS-DA analysis. LOOCV models imply to train a prediction model on spectra recorded from all but one patient and to subsequently validate the model on the spectra from the remaining patient. The procedure was repeated until all patients' spectra had been challenged by the corresponding prediction model.

To evaluate PLS-DA model performances the sensitivity, the specificity and the Matthews Correlation Coefficient (MCC) were computed according to the following equations: [46]

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

Where TP is the number of True Positives, TN the number of True Negatives, FP the number of False Positives (type I error, equivalent to a false alarm) and FN the number of False Negative (type II error, equivalent to a miss). Sensitivity and specificity are statistical measures of the performance of a binary classification test. A perfect predictor would be described as 100% sensitive and 100% specific. The MCC is a measure of the quality of a binary classifier that can be used even if the classes are of very different sizes. It varies from +1 to -1, +1 indicates a perfect prediction, 0 is equal to the performance of a random classification and -1 indicates a total disagreement between prediction and observation. Models with more than two classes were evaluated by a similar pairwise analysis.

Correction of the spectra for water vapor and CO_2 contribution, baseline subtraction, normalization, quality filter application, PCA, MANOVA, Student t-test and PLS-DA analyses were carried out by Kinetics, a custom made program, running under Matlab (Matlab, Mathworks Inc).

RESULTS

Identification of different cell types within tissue sections

As described in Figure 1, each tissue section examined by FTIR imaging was associated with an adjacent tissue section that was H&E stained. The stained section was constantly used as a reference to define the areas to be analyzed by infrared imaging.

An example of areas identified as rich in epithelial cells, lymphocytes, erythrocytes, connective tissue (stroma) and vascular tissue appear on Figure 1D but the different areas are much more visible on Figure 2. Figure 2 also reports a typical view of the different regions that have been sampled. Once the region identified on the scan of the stained section was also identified on the unstained section (Figure 1G), an infrared image could be recorded (Figure 1H). Obviously, tissues are not homogenous at the scale of the images shown in Figure 2 and spectra had to be manually extracted from image regions representing a specific, well-identified cell/tissue type. This was achieved by selecting regions or individual pixels with a cursor. In some instances a PCA image was produced to guide the selection (Figure 1I). As shown in Figure 2, almost 60 unit images (ca 240,000 spectra) were recorded from this single section.

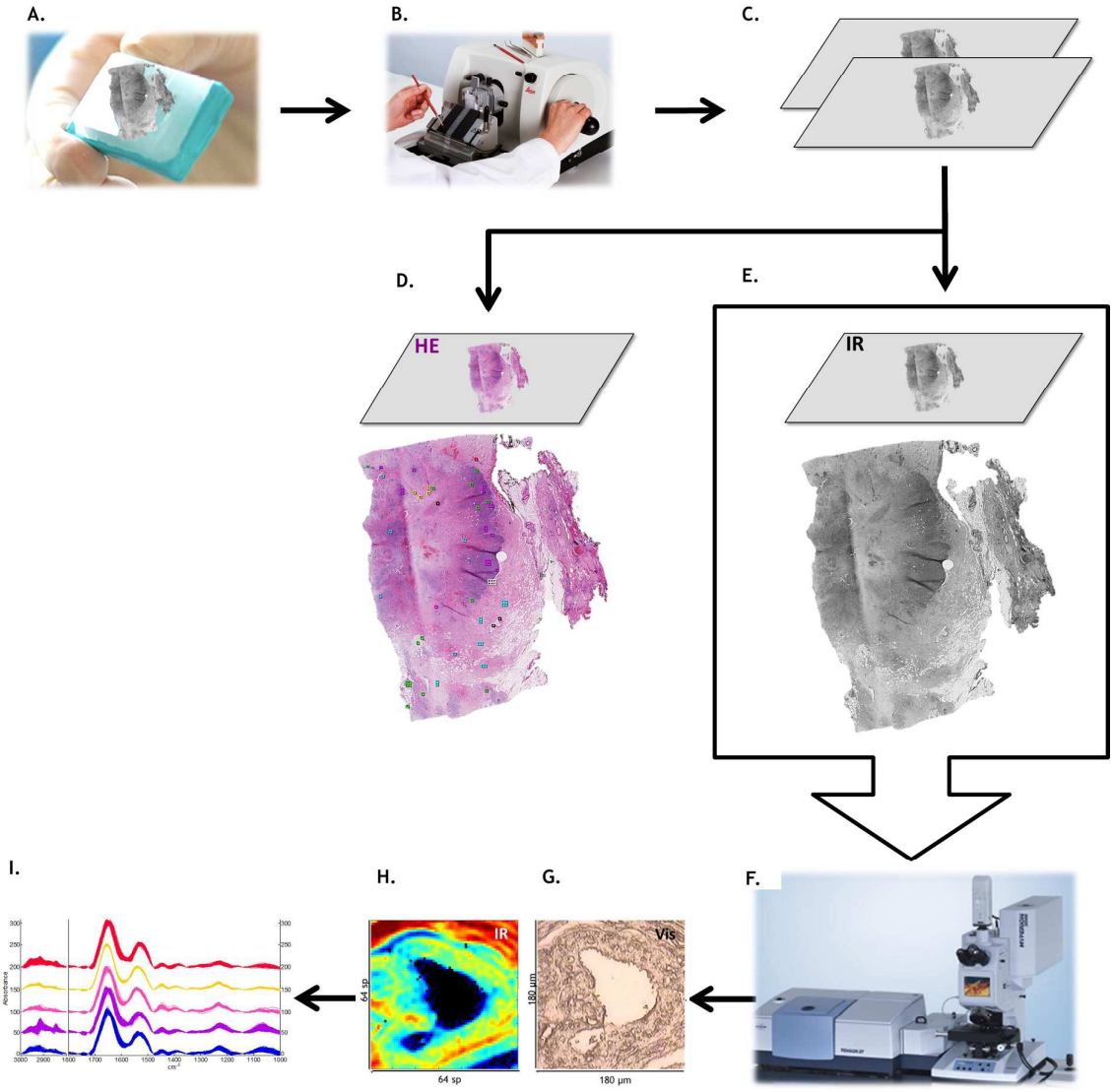


Figure 1: *Illustration of the experimental workflow. For each FFPE breast tissue sample (A), 2 adjacent tissue sections of 4 µm of thickness were cut using a microtome (B,C). The first section was H&E stained (D) while the second section remained unstained (E). Colored squares were added on the H&E stained reference sections to indicate the IR measurement positions (see also Figure 2) (D). IR spectra were recorded on the unstained sections and the IR spectra of the five main breast tissue cell types were selected by manual pixel selection (F, G, H). Spectral data were then preprocessed as described in materials and methods (I).*

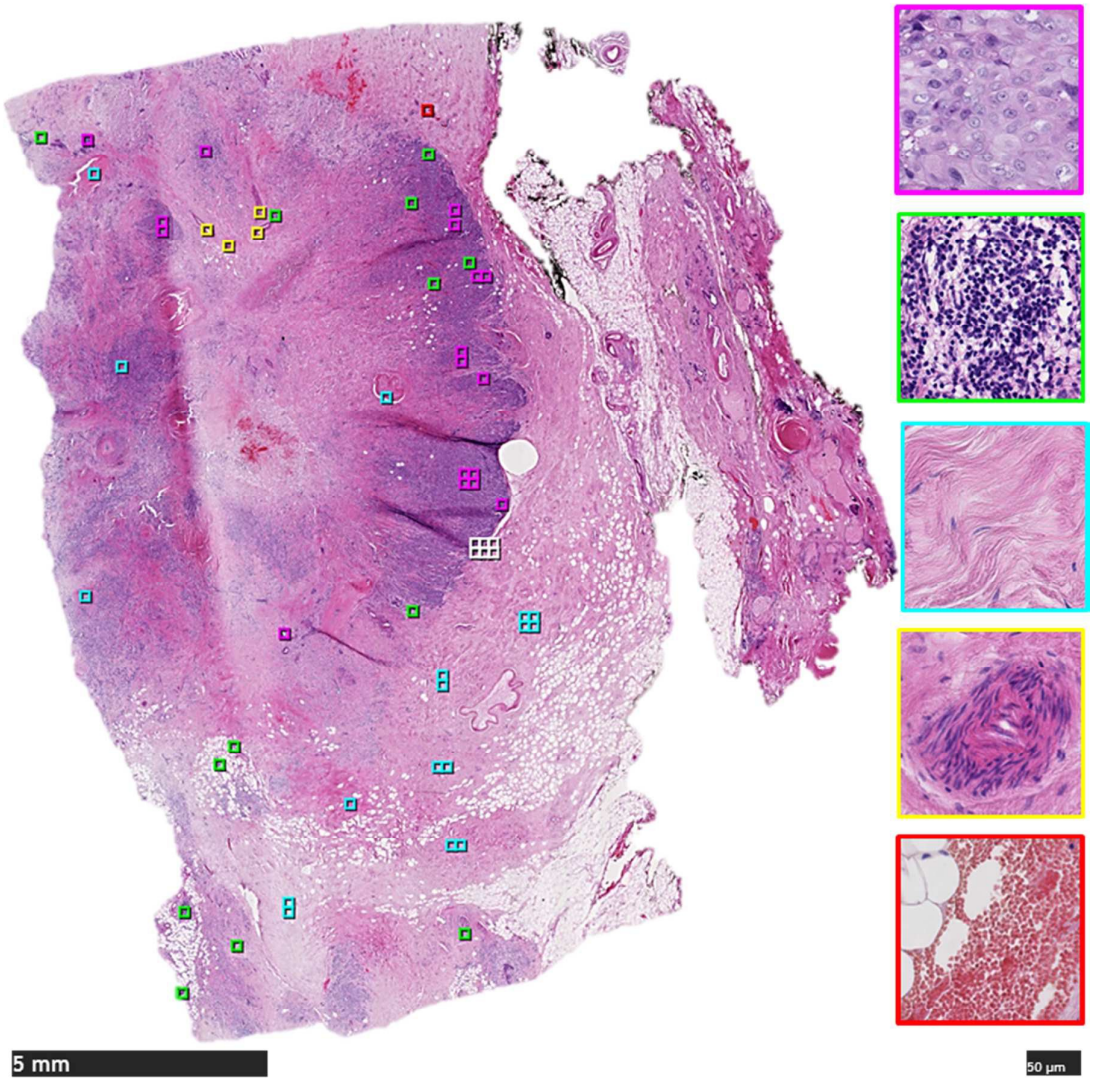


Figure 2: *Whole slide scan of an H&E stained breast tissue section annotated with colored-squares indicating the IR measurement areas (global view, left). A color-code is used for each of the 5 main breast cell/tissue types; epithelial cells, lymphocytes, connective tissue (stroma), vascular tissue and erythrocytes measurements are indicated by pink, green, cyan, yellow and red single or multiple colored-squares respectively. Images of larger areas grouping a panel of different cell/tissue types are indicated by adjacent multiple white squares. An example of each breast cell/tissue type is shown at the right side of the figure.*

Unsupervised analyses

Five main cell types and tissue components were identified in breast tumor tissue samples: epithelial cells, lymphocytes, erythrocytes, connective tissue and vascular tissue. Ten IR images were recorded on each of the main cell types, resulting in *ca.* 10^6 infrared spectra after the manual selection step described above. Due to the large volume of spectral data, in a first step 10 mean spectra were extracted from every IR measurement for computing a PCA analysis. Figure 3 illustrates the result of a PCA computed on mean spectra obtained from one breast tumor tissue section.

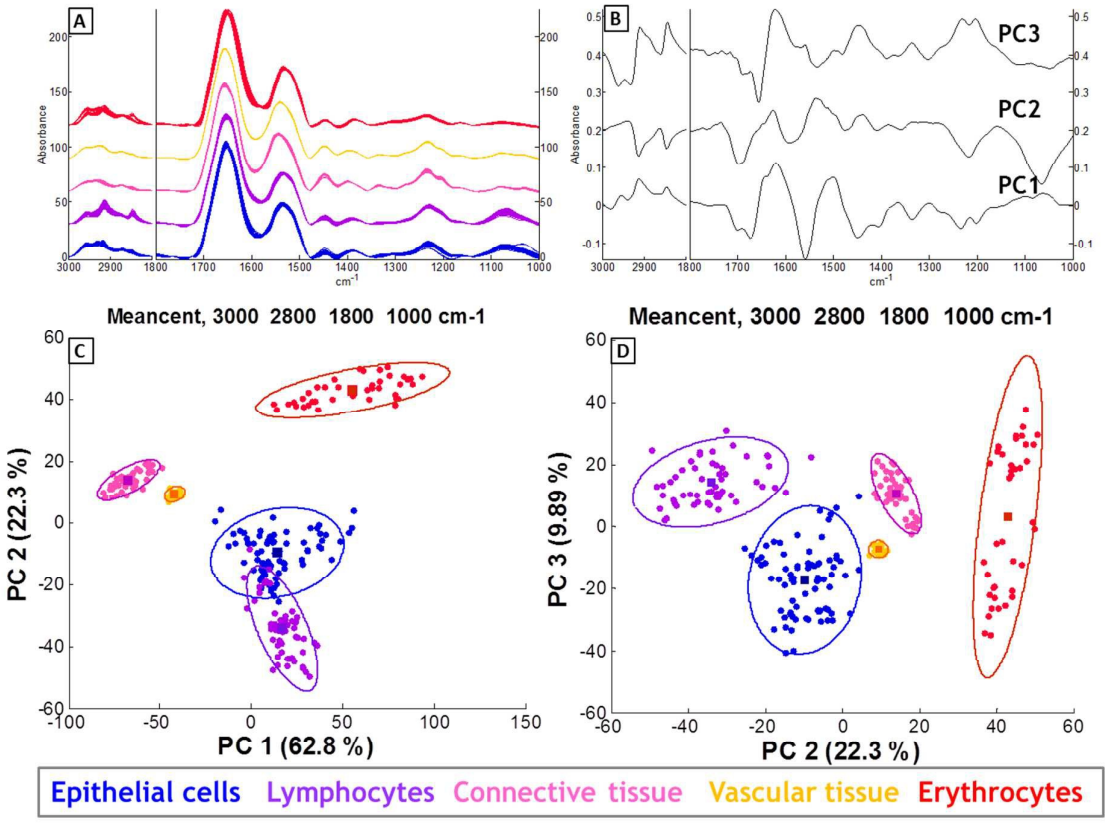


Figure 3: *PCA analysis of IR spectra recorded on the 5 main cell types and tissues found in breast tissue samples.* ***A.*** *239 mean IR spectra recorded on epithelial tumor cells (blue), lymphocytes (purple), connective tissue or stroma (pink), vascular tissue (yellow) and erythrocytes (red) of one patient.* ***B.*** *Loading vectors PC1, PC2 and PC3 of the PCA. Spectra are offset along the absorbance (Y) axis for clarity.* ***C, D.*** *PCA Score plots showing the projection of each spectrum of A in the PC1-PC2 (C) and PC2-PC3 (D) spaces. Each dot stands for one spectrum. Percentages on the axis labels indicate the variance described by PC1 (62.8%), PC2 (22.3%) and PC3 (9.89%). PCA was performed on the 3000-2800 and 1800-1000 cm^{-1} combined spectral regions.*

From the score plot reporting the score of each spectrum on PC1 and PC2 (Figure 3C) it appears that connective and vascular tissues exhibit highly similar spectral features that are not shared with the other cell types. PC1 describes the majority of the total spectral variance (62.8%) and is responsible for the separation between vascular and connective tissues on the one hand and the other cell types on the other hand (Figure 3C). PC2 (22.3% of the total spectral variance) mainly separates the erythrocyte population from the other cells. The PC2 versus PC3 score plot (Figure 3D) shows that lymphocyte spectra get finally separated from epithelial cell spectra along PC3 (9.9% of the total spectral variance). In conclusion, the first three PCs allow a good separation of the 5 main cell types and tissue components identified above. A MANOVA was run on all the spectra presented in Figure 3 using for each spectrum its scores on the first 6 PCs. The chance that the mean value of one cell population is equal to the mean value of another cell population (p-value) was always found to be lower than 10^{-20} (data not shown).

Figure 3B illustrates the first 3 PCs which show spectral features all over the IR absorbance spectral region. The double z-shape observed in the amide I - amide II region of PC1 could underline a difference in the secondary structure of the proteins. This could be understood as connective and vascular tissues both have a high content in collagens, a family of proteins characterized by unique triple helix structures. The quite intense negative band near 1210 and 1050 cm^{-1} found in PC2, tentatively assigned to nucleic acid phosphate vibrations, could be related to the lack of nucleic acids in erythrocytes, characterized on the score plot by a high positive weight of PC2 (Figure 3C).

The results reported in Figure 3 have been obtained on a relatively small number of mean spectra representing the main cell/tissue types. The question of whether spectra originating from individual pixels (voxels) of the images would result in the same type of separation according to cell type needs to be addressed. A similar PCA analysis was computed on a subset of 500 randomly selected spectra per cell type, except for the vascular tissue class where all 239 spectra recorded on the tissue section were considered for the analysis (Figure S1, supplemental data). Figure S1 indicates that separation trends are similar but with more scattering of individual spectra. Yet, identical conclusions can be drawn regarding the separation trends and the PCs shapes are very similar. Again, a MANOVA run on all the individual spectra presented in Figure S1, using for each spectrum its scores on the first 6 PCs, indicates that the chance that the mean value of one cell population would be equal to

the mean value of another cell population was always found to be lower than 10^{-20}. Similar results have been obtained for all 13 patients when studied one by one, on the tumor tissue samples as well as for the 6 normal (non-adjacent non-tumor) breast tissue samples.

Supervised analyses

A PLS-DA analysis was run on the different cell/tissue types identified above. Table 1 shows the predictions obtained with a bootstrapping approach. A model was created for each cell type using 60% of IR spectra recorded on the 13 breast tumor tissues (training set). Spectra were selected randomly. The model was then applied on the remaining 40% of the cell spectra (test set). The process was repeated 40 times with each time a different selection of the spectra belonging to the training set and to the test set. The sensitivity, specificity and MCC scores of each cell/tissue class are reported in Table 1.

		Predicted as					Sensitivity	Specificity	MCC
		Epithelial C.	Lymphocytes	Connective T.	Vascular T.	Erythrocytes			
True	Epithelial C.	**89.5±0.4**	1.1±0.1	2.3±0.2	7.0±0.4	0.1±0.1	89.5±0.4	96.6±0.8	0,85
	Lymphocytes	2.5±0.2	**88.8±0.5**	3.5±0.2	5.0±0.3	0.2±0.0	88.8±0.5	99.6±0.9	0,92
	Connective T.	5.6±0.3	0.1±0.0	**80.2±0.7**	14.1±0.6	0.1±0.0	80.2±0.7	92.5±0.8	0,70
	Vascular T.	2.0±0.2	0.0±0.0	23.3±0.7	**74.4±0.7**	0.3±0.1	74.4±0.7	92.7±0.7	0,66
	Erythrocytes	3.5±0.3	0.3±0.1	1.1±0.1	3.3±0.3	**91.8±0.3**	91.8±0.3	99.8±0.9	0,94

Table 1: *Confusion matrix obtained by PLS-DA on a subset of 50 000 randomly selected IR spectra among all 13 patients. Sensitivity, specificity and MCC scores are indicated at the right side. The model was forced to take the same number of spectra in each group, i.e. 10 000 IR spectra for each cell or tissue type. 60% of the IR spectra recorded were used to train the model (training set) the remaining 40% were used to test the model (test set). Percentages in bold indicate the rate of correct assignment after 40 bootstrap iterations (mean ± standard deviation). The analysis was performed on the combined 3000–2800 and 1800–1000 cm^{-1} spectral regions.*

As reported in Table 1, all 5 cell populations are efficiently discriminated with correct assignment scores reaching almost 90% for epithelial cells, lymphocytes and erythrocytes. Connective (stroma) and vascular tissue show lower correct assignment scores of 80 and 74% respectively. Interestingly, this is due to the fact that there is a high confusion rate between these two groups. This confusion was already highlighted by the previous unsupervised analyses and is probably due to the collagen high abundance found in both tissue types. Finally Table 1 exhibits low standard deviation values which underline the good reproducibility of the model. When applied on IR images of larger areas, supervised analyses produce pseudo-colored spectral maps highlighting the different cell types. An example is displayed in Figure 4.

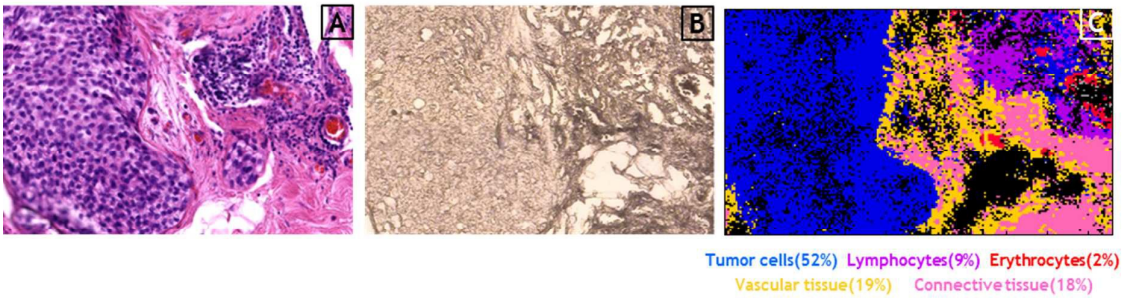


***Figure 4**: Breast tumor tissue analysis. **A.** H&E stained breast tumor tissue section of 540 x 360 µm² exhibiting tumor epithelial cells, lymphocytes, erythrocytes, vascular tissue, connective tissue. **B.** Same tissue area on the adjacent unstained tissue section. **C.** Pseudo-colored image obtained after application of the cross-validation cell type recognition PLS-DA model calibrated on a subset of 26 273 spectra from 12 patients and applied on the IR spectra recorded on the unstained tissue section (B) of the 13th patient: tumor cells, lymphocytes, erythrocytes, connective tissue and vascular tissue appear in blue, purple, red, pink and yellow respectively. Black and grey pixels correspond to spectra that do not meet the quality filter thresholds. Percentages in brackets represent the tissue portion covered by each cell type according to the recognition model.*

As described in Materials and Methods, none of the spectra of Figure 4C was included in the training set. Yet, inspection of Figure 4 reveals good agreement between the PLS-DA pseudo-colored map (Figure 4C) and the tissue features revealed by the H&E staining (Figure 4A). Interestingly, the percentage of area covered by epithelial cells, lymphocytes, erythrocytes, vascular and connective tissues is provided automatically.

This is of importance as lymphocytes are of diagnostic value and are more and more counted, manually, by the pathologists.[47–50] Similar pseudo-colored spectral maps have been obtained

for all other IR images with similar agreement with the corresponding H&E-stained reference sections.

The data acquired so far indicate that it is possible to recognize the different cell/tissue types with a high efficiency. In the next part of this work we will explore the possibility to obtain more detailed information on the epithelial cells from which the large majority of cancers originate.

Epithelial phenotype determination: tumor versus normal

Unsupervised analyses

The previous section clearly indicates that IR imaging is an effective tool to identify epithelial cells within breast tissue sections. This second section focuses on the ability of IR imaging to differentiate tumor and normal phenotypes of breast epithelial cells. Tumor and normal cells selected here are enclosed in the tumor tissue or at distance of the tumor site (margin) respectively. In a first step a PCA was performed on the epithelial spectra recorded on both the tumor tissue and the normal tissue from one patient (Figure 5).

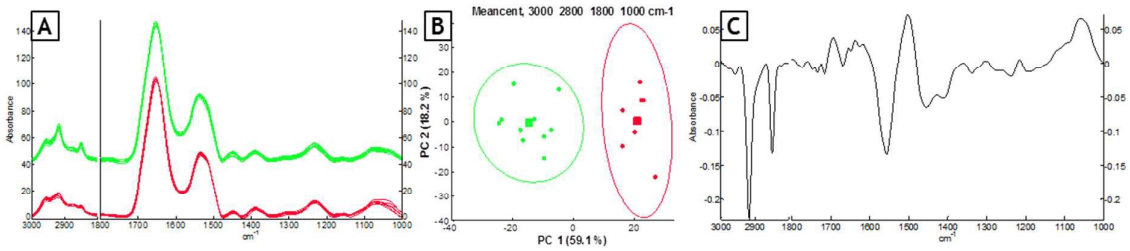


Figure 5: *PCA analysis performed on IR spectra recorded on tumor and normal breast epithelial cells of one patient.* ***A.*** *27 mean spectra from tumor (red) and normal (green) breast epithelial cells. Each spectrum is the mean of all epithelial cell spectra selected from one IR image. Spectra were shifted for the sake of clarity.* ***B.*** *PCA Score plot depicting the projection of each spectrum of A in the PC1-PC2 space, performed on the 3000-2800 and 1800-1000 cm^{-1} combined spectral regions. Each dot stands for one spectrum. The mean of each group is represented by a square surrounded by a 90% confidence ellipse. Percentages on the axis labels indicate the variance described by PC1 (59.1%) and PC2 (18.2%).* ***C.*** *Loading vector PC1 of score plot B.*

The PCA score plot presented in Figure 5B clearly shows that spectra cluster according to the epithelial phenotype along PC1. The 90% confidence ellipses do not overlap and a MANOVA performed on projections of the dataset on the first 6 PCs resulted in a p-value below 10^{-6}.

The first PC, which is the principal component that describes the majority of the spectral variance (59.1%), is illustrated in Figure 5C. PC1 displays spectral features all over the IR spectral range but most prominent features are located in the 3000-2800 cm^{-1} spectral region. This spectral region is assigned to the CH_3 antisymmetric as well as CH_2 symmetric and antisymmetric vibrational frequencies of the C-H groups present in alkyl chains, mostly contained in lipids.[51,45,52] The positive scores of the tumor cells on PC1 and the negative C-H stretching bands of this PC indicate that the tumor cells have less lipids than normal cells. Nonetheless, these peaks might also result of a difference in the paraffin uptake of both tissues. PC1 shows a second particular feature between 1600 and 1500 cm^{-1}, a region usually assigned to the protein Amide II band. In the absence of a similar feature in the amide I band, such a change cannot be safely assigned to a protein secondary structure change. Finally a band in the phosphate region (1000-1100 cm^{-1}) indicates a potential difference in the nucleic acid abundance/state between normal and tumor epithelial cells.[53] While a very good separation could be obtained when examining each patient individually, separation was less sharp when all data were analyzed simultaneously.

In a second step, a PCA was performed simultaneously on all the epithelial cell spectra recorded on the 6 patients for whom both tumor tissue and normal tissue collected away from the tumor were available (Figure 6).

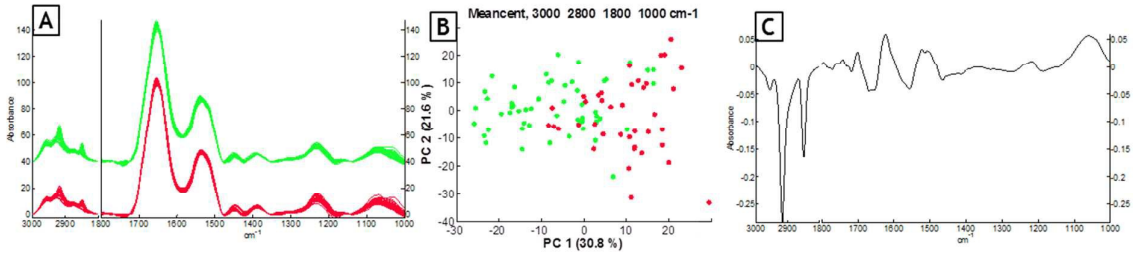


Figure 6: *PCA analysis performed on IR spectra of tumor and normal breast epithelial cells recorded on 12 tissue sections from 6 patients.* ***A.*** *93 Mean spectra from tumor (red) and normal (green) breast epithelial cells after outlier detection. Each spectrum is the mean of all epithelial cell spectra from one IR image. Spectra have been offset for visual clarity.* ***B.*** *PCA Score plot depicting the projection of*

each spectrum of A in the PC1-PC2 space, performed on the 3000-2800 and 1800-1000 cm^{-1} combined spectral regions. Each dot stands for one spectrum. Percentages on the axis labels indicate the variance described by PC1 (30.8%) and PC2 (21.6%). ***C.*** *Loading vector PC1 of score plot B. All spectra are offset along the absorbance (Y) axis for clarity.*

The PCA score plot presented in Figure 6B reveals a trend to separating tumor and normal cells based on PC1, which describes the majority of the spectral variance (30.8%), yet with a large degree of overlap. This shows that there is a relatively high portion of the variance coming from inter-patient differences that blurs the separation along PC1. A MANOVA performed on this dataset yields a p-value below 10^{-13}, confirming that even when the tissue from the 6 patients are considered together, the normal and tumor cells have distinct spectral features. The shape of PC1 computed here (Figure 6C) is very similar to the one obtained on a single patient (Figure 5C), suggesting the separation is robustly based on very similar spectral features for all patients.

Interestingly, when epithelial cell spectra recorded on all the other tumor tissue sections were added to the PCA, the tumor cell spectra of the 7 supplementary independent sections were located within the tumor cell spectra already shown at the right side of the PCA score plot of Figure 6B (not shown). Figure 7 displays a difference spectrum between the mean spectra of the two epithelial populations. The difference spectrum reveals good agreement with the PC1 loading vector shown in Figure 6C. A Student t-test computed at each wavenumber using all individual spectra confirmed that the spectral features described previously are statistically significant. Results are reported as stars on the difference spectrum presented in Figure 7.

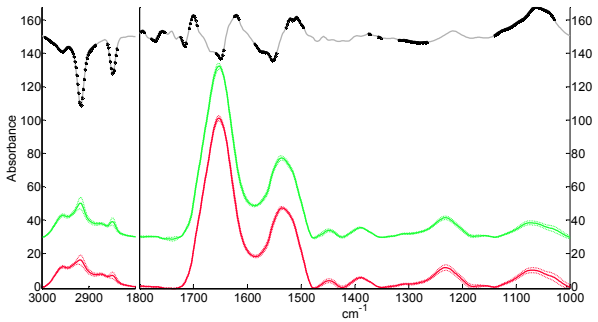


Figure 7: *Difference spectrum (grey) and mean (12 tissue sections from 6 patients) IR spectra (solid lines) ± standard deviation (dotted lines) of tumor (red) and normal (green) breast epithelial cells.. The difference spectrum was obtained by subtracting the mean spectrum of normal cells from the mean spectrum of tumor cells. Each marked wavenumber (black star) indicates a statistically*

significant difference(α=1%) between the means. Spectra are offset along the absorbance (Y) axis for clarity and a 10-fold multiplicative constant has been applied to the difference spectrum.

Supervised analyses

The unsupervised analyses described so far convincingly show that normal and tumor cells have significantly different infrared spectra. In order to develop an IR classification tool with a potential clinical value, we set up a PLS-DA bootstrapping analysis of the two populations, including all epithelial cell spectra recorded from the 12 breast tissue sections (tumor and normal cells, Table 2).

		Predicted as	
		Tumor	Normal
True	Tumor	**92.0±0.6**	8.0±0.6
	Normal	9.5±0.6	**90.5±0.6**

Table 2: *Confusion matrix obtained by performing a supervised PLS-DA statistical analysis on 11,120 IR spectra recorded on epithelial cells of 12 tumor and normal tissue sections from 6 patients. 5,560 spectra of each epithelial phenotype (tumor and normal) were considered to build the model. 60% of the IR spectra recorded were used to train the model (training set) the remaining 40% were used to test the model (test set). Percentages in bold indicate the rate of correct assignment of the spectra to the epithelial cell phenotype after 40 bootstrap iterations (mean ± standard deviation). The analysis was performed on the combined 3000–2800 and 1800–1000 cm^{-1} spectral regions.*

In this procedure, a fraction of the spectra was randomly selected to be used as a training set and another distinct fraction of the spectra was used as a test set. The operation was repeated 40 times with each time a different sampling of the training set and of the test set. The models reached very good sensitivity, specificity and MCC scores of 92.0±0.6%, 90.5±0.6% and 0.83 respectively. Nonetheless it could be argued that in this procedure the training and the test sets are not totally independent. Therefore an external validation was carried out at the next section.

When a leave-one-out cross-validation (LOOCV) procedure was set up to test the spectra from one patient at a time while leaving all the spectra from that patient out of the training set, the mean sensitivity, specificity and MCC scores dropped to 76%, 82% and 0.6 respectively. Results of the LOOCV models are illustrated in Figure 8. For all but one patient, these models

still predict correctly the overall tumor/normal phenotype of the epithelial cells when considering a threshold of 50 % for the global assignment (horizontal dotted line, Figure 8).

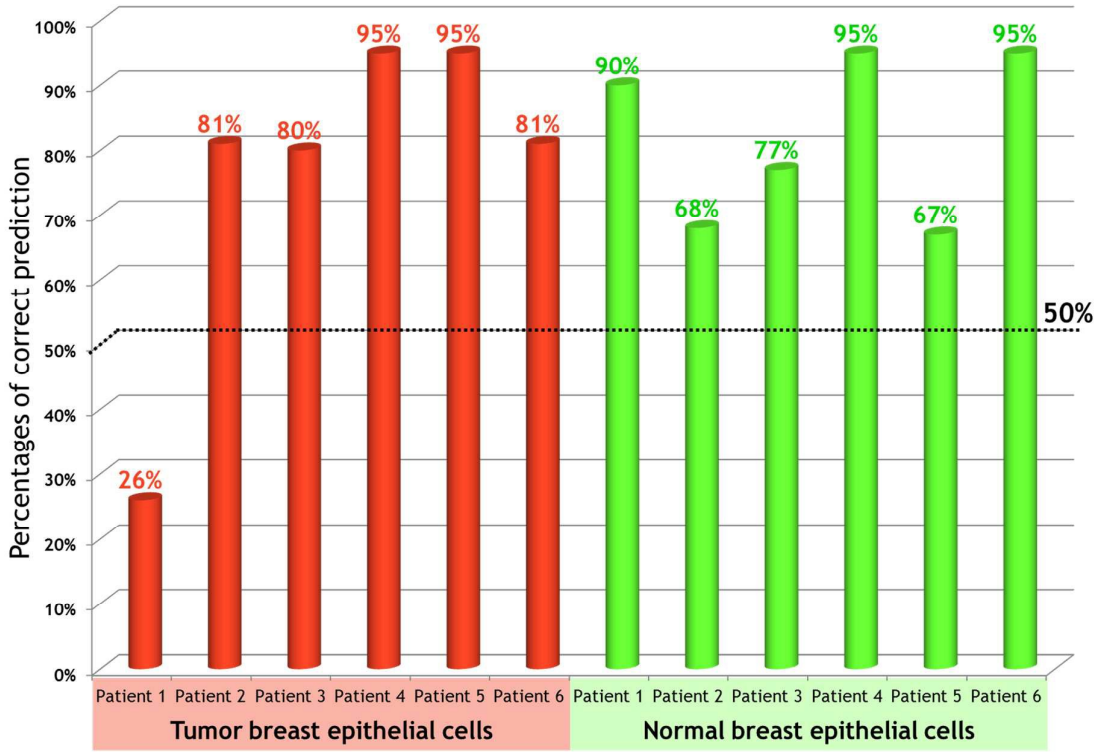


Figure 8: *Percentages of correct prediction for the status of epithelial cells. For each patient, percentages correspond to the results of correct assignments for tumor epithelial cells (red bars) and normal epithelial cells (green bars). The PLS-DA models were calibrated on 2 subsets of 5560 spectra recorded on the normal and tumor epithelial cells embedded within 12 different breast tissue samples belonging to 6 patients. The threshold indicated at 50% corresponds to a model performance equal to a random classification.*

Further improvement can be obtained if spectra of obviously normal cells and obviously cancer cells are recorded from the same patients. Such an approach could be useful to identify automatically the margins of a tumor, after calibration of the model on a set of obviously normal and obviously cancerous cells. For each patient, 3 IR images recorded on the tumor and 3 images recorded on normal epithelial cells were used to calibrate a PLS-DA model which was validated on the remaining 14 IR images obtained from the same patient. The results of the validation are displayed in Figure 9 for all patients. These patient-specific models reach dramatically higher recognition percentages (above 90% for all but one patient) and average sensitivity, specificity and MCC scores of 99%, 93% and 0.93 respectively. The

overall sensitivity value is particularly high making these models of interesting clinical utility as discussed later. Considering the 50% threshold shown by the horizontal dotted line on Figure 9 to assign the entire sample, all samples would be correctly classified.

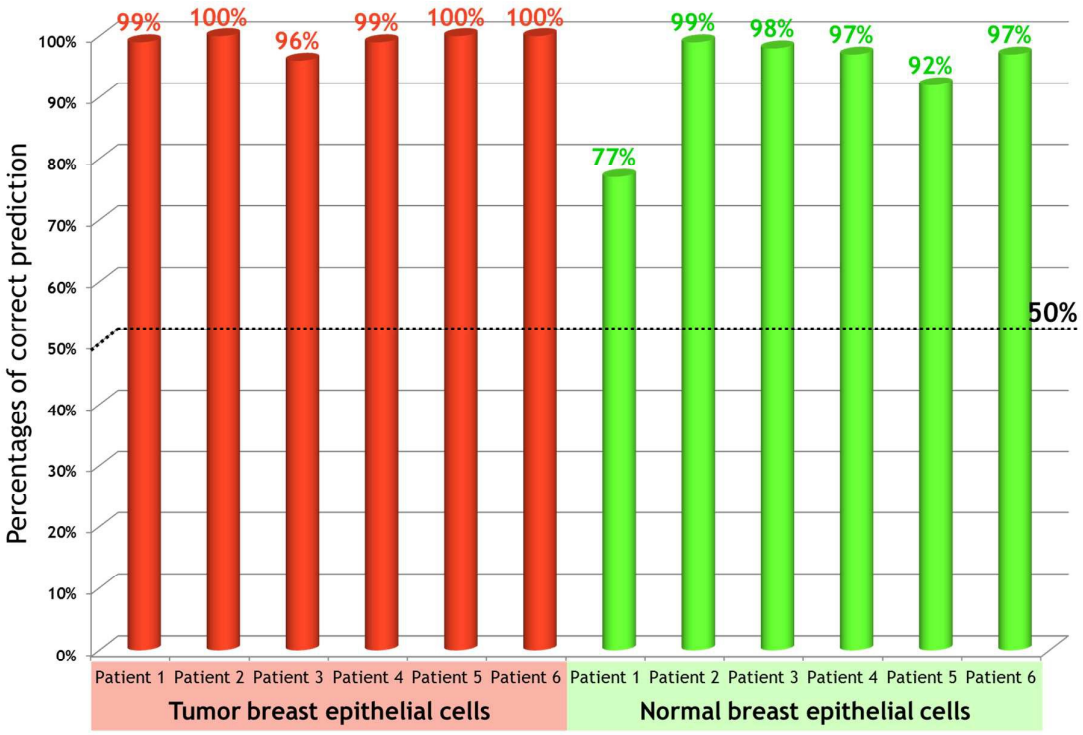


Figure 9: *Fraction of correct assignments for tumor epithelial cells (red bars) and normal epithelial cells (green bars) obtained with 6 PLS-DA models, one for each patient. Each model was calibrated on subsets of 3 IR measurements recorded on normal and tumor epithelial cells of each of the 6 patients and tested on the remaining measurements of the same patient. The threshold indicated at 50% corresponds to a model performance equal to a random classification.*

Impact of the tumor on the stroma

It has been shown before that the stroma experiences chemical changes in the vicinity of breast tumor that can be monitored by infrared imaging.[54,41] It was found that the amplitude of the chemical changes evidenced by infrared spectroscopy decreased with the distance to the tumor.

In the present work we address the question of a modification of the stroma located in the vicinity of the tumor (less than 10 mm) with respect to the stroma located at remote locations. For this purpose we collected two images (4096 spectra each) of the stroma located within a

few mm from the tumor of 11 different patients. Spectra were processed as described, checked for S/N and spectra of area corresponding to the stroma were manually extracted. In a second step, 200 spectra were extracted randomly from this first selection for each image, yielding 4,400 spectra of stroma found in a tumor environment. As a control, stroma samples from 8 patients were collected from a non-tumor location, either from mammary reduction for half of them or, from a distant location within the same breast for the other half. Again, two images were acquired for each patient, yielding finally 3,200 randomly selected spectra by the same procedure.

PCA was not able to reveal any clear distinction between the two groups of spectra when all individual spectra were considered simultaneously (not shown). Yet, some degree of separation can be observed when the mean spectra (one mean spectrum per tissue section) were considered (Figure 10B).

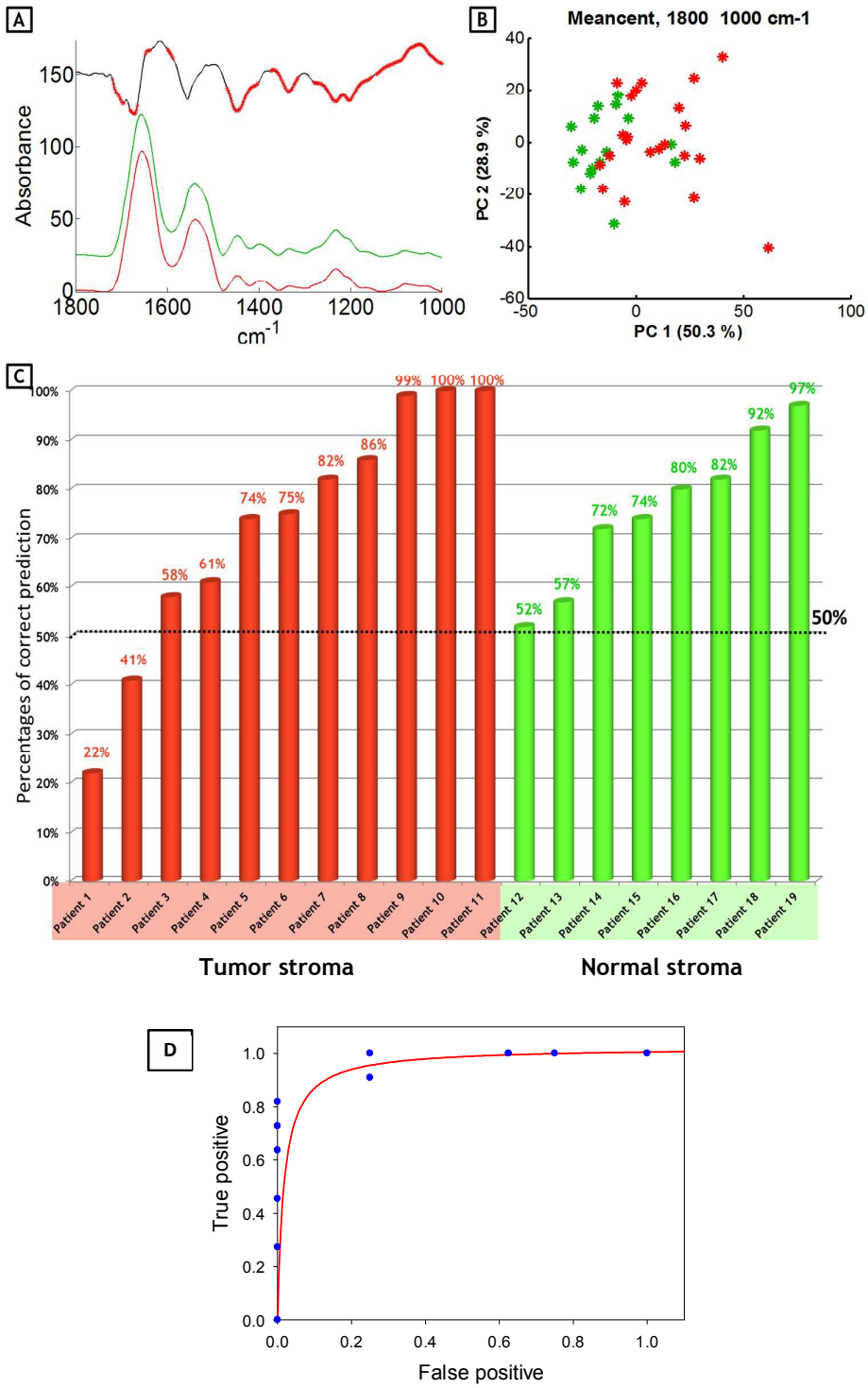


Figure 10: *Comparison of the stroma located near the tumor and in normal tissue.* ***A.*** *Difference spectrum (top, black) and mean IR spectra (below, solid lines) of the stroma found near the tumor (red, 11 patients, 22 tissue sections, one IR images of 4096 spectra each) and of the stroma found in normal tissue (green, 8 patients, 16 tissue sections and one IR image of 4096 spectra each). The difference spectrum was obtained by subtracting the mean spectrum of the stroma found in normal*

tissue from the mean spectrum of stroma found near tumor cells. Red stars indicate a statistically significant difference between the means (α=1%). Spectra were offset along the absorbance (Y) axis for clarity and a 10-fold multiplicative constant was applied to the difference spectrum. ***B.*** *Score plot of a PCA applied on the mean spectra described in A. Color code as in panel A.* ***C.*** *Result of a supervised PLS-DA analysis. The rate of correct assignment (in %) to either stroma found in a tumor environment (red) or to stroma found in a normal tissue environment (green) for each patient. A new PLS-DA model was built for each patient so that none of the tested patient's spectra was used to build the model tested on that patient. D. ROC curve built from the data presented in C.*

A Student t-test run at all wavenumbers indicates that the most significant differences between the two groups are found between 1450 and 1000 cm^{-1}. A PLS-DA was then achieved. Nineteen different models were built, each one with the spectra from 18 patients, *i.e.* without any contribution of one of the patient, and the patient left out of the model building was used as a test (Figure 10C). If assigning the stroma to a subclass (close to tumor or close to normal tissue) according to the assignment of the majority of the spectra, all "normal stroma" samples were correctly identified but 2 false negatives were found for the "tumor stroma" (Figure 10).

DISCUSSION

Intraoperative tumor-localizing techniques to guide tumor excision work on a 'macroscopic' level and fail to give information about cancer cells at the margin, what often results in re-excision surgery and local cancer recurrence. It would therefore be ideal to get clear margins at a single operation and in an intraoperative context. Nowadays visual inspection of stained tissue sections is considered as the gold standard and no perfect tool exists to assess the surgical margin intraoperatively.[3] Infrared spectroscopy and infrared imaging in particular are progressing very rapidly. Yet, today, infrared imaging cannot yet compete with the gold standard procedure used in histopathology. Conventional infrared imaging is slow, requires liquid nitrogen and large instruments that will not fit easily in surgery rooms. Pathologists are trained to examine hundreds of histological sections every day what is currently unachievable by IR imaging. Nonetheless recent advances achieved in IR whole slide imaging,[55] the advance in tunable quantum cascade lasers,[56–59] so-called discrete frequency imaging[57,14] etc. indicate that in a short time scanning and analysis may be able to compete with classical microscopic examination of tissue sections. This suggests that this spectroscopic method combined with powerful computers could soon become an interesting tool in histopathology.

The present study addresses the potential of infrared imaging to detect sufficient information in breast cancer tissue to identify the main cell types/phenotypes. Overall, cell type and phenotype discrimination remains better when analyzing tissues from the same patient. Patient by patient, difference between cell types and differences between normal and tumor epithelial cells are very clear. When mixing data from several patients, unsupervised analyses show a trend to separation and supervised PLS-DA method provides a good but not total identification of the cell type or phenotype. This observation implies that there is a significant inter-patient variance. This can be due to either true biochemical diversity or to variability in tissue processing. In the latter case, many parameters could be considered such as the post-ischemia delay before tissue fixation, the time spent in formaldehyde, the type of paraffin used and the precise temperature used for paraffin embedding, etc. Whatever the source of variance, these variations will be modelled much better when a larger cohort of patient will be used. Yet, the present study emphasizes the feasibility of using FTIR imaging as a series of spectroscopic markers to identify cell types and phenotypes. It must be stressed that the present results have been obtained on FFPE samples while tissues available in the course of surgery will be fresh. Questions about the "validity" of samples exposed to formaldehyde, ethanol, xylene and hot paraffin can be raised. Some of the lipids are indeed extracted, some of the proteins are indeed denatured.[60] Yet, the data accumulated so far in the literature demonstrate that sufficient differences between tissue and cell types are maintained. Whether the C-H stretching region (3000-2800 cm^{-1}), usually assigned mainly to lipids, can bring useful information is even more problematic as lipids could be extracted during the procedure. To address that question, we previously analyzed spectra from 8 cancerous epithelial cell lines. Before FFPE processing, these closely related cell lines could be identified uniquely on the basis of their IR spectrum. When the experiment was repeated after FFPE procedure, the IR spectra underwent small but significant changes but spectral distances between closely related cell lines were mostly conserved, including when considering the 3000-2800 cm-1 spectral range.[60] It may be surprising that lipids are not completely removed after the procedure. Examining tissue sections indicates that triglycerides present in adipocytes are fully removed (not shown) but phospholipids are not all soluble in the organic solvent used for FFPE. The sample analyzed here went through ethanol baths and then two successive xylene baths for 20 min each. In such conditions, on a 4 µm thick section, all molecules soluble in these solvent are likely to be extracted. The experiments show that some lipids are

not soluble in these conditions, in line with the previous observations carried out on cell lines.[60]

In a first step, the ability of IR imaging to identify the five main cell/tissue types in breast tumor tissue samples was evaluated. Unsupervised and supervised analyses indicated that breast cell/tissue types show different spectral features that can be used to create pseudo-colored images and provide automatically the area covered by each cell/tissue type. This confirms the potential utility of IR imaging as complementary technique to identify objectively, automatically and without staining requirement, thus rapidly, the breast cell types present in a tissue section. In agreement with the capabilities of FTIR spectroscopy underlined in previous studies[61,62,21] we showed here that an automated assignment of spectra to specific cell types is feasible and robust. This could constitute a first layer of assignments. The software could then automatically apply a specific model to each cell type identified in this first layer, as already implemented in Kinetics. In a second step, computational analyses were applied on spectra recorded on epithelial cells from breast tumors and from normal tissues. Unsupervised analyses indicated that spectral differences exist between the two epithelial phenotypes and that an IR tumor signature can be identified. Supervised PLS-DA models allowed an efficient discrimination between epithelial normal cells and tumor phenotype to be obtained. Validation of these models on independent test sets (Figure 8) identified correctly all the tissues except for one false negative. Patient-specific models (Figure 9) produce a highly reliable and powerful tumor classification tool with excellent sensitivity scores (99% in average). For such analyses, 3 IR measurements from the same patient should first be obtained both at the tumor site and from a tumor-free area. The epithelial spectra would then be extracted by the first layer model described above and encoded as reference spectra to create a specific tumor recognition model. The latter would then be applied to detect quickly and automatically tumor cells in tumor surgical margins. While it could be more labor intensive in real analyses, this method combined to the recent advances achieved in IR whole slide imaging[55] suggests IR imaging could soon become a complementary tool to assess margin status intraoperatively. This approach is in line with the current trend to set up personalized medicine in cancer. As each patient would have its own internal control, such personalized diagnostics would result in more accurate and efficient results. The patient-specific model developed in this study takes into account the specificities of the cancer developed by the patient. This is particularly important in the case of breast cancer which is now well-recognized as a heterogeneous disease, grouping more than one unique pathology.[63–65]

Quite interestingly, the last part of this study emphasizes the differences that can be observed in the stroma when it is close to the tumor or taken from a remote location. This could help identify the presence of the tumor through its effect on the stroma. As recently reviewed,[36] by comparison with invasive breast cancer which has been studied extensively, the role of the stroma is now shown to be a major player in cancer progression but lack tools for its characterization. Recent studies have identified alterations in stromal cell function that may be critical for disease progression from benign disease to invasive cancer.[36,38] Key functions of myoepithelial cells that maintain tissue structure are lost, while tissue fibroblasts become activated to produce proteases that degrade the stroma and trigger the invasive cellular phenotype.[66,67,36–39] Stromal cells express specific genes in cancer area and invasion is stimulated.[68,40] Consequently, a pressing need exists for tools that can characterize the particular chemical state responsible for stroma properties with respect to cancer progression.
The 13 tumor samples used here were selected to include a large panel of distinct clinical parameters that are routinely assessed in the diagnosis of breast cancer, *i.e.* different histological grades, tumor stages, molecular subtypes, tumor sizes, lymph node status, hormonal receptor expressions and KI67 expressions (Table S1). We performed an exploratory study to evaluate the capability of IR imaging to identify those clinical parameters based on the spectral characteristics of epithelial tumor cells from the 13 independent tumor tissue samples. No correct classifications were observed. These negative results provide an excellent control demonstrating that LOOCV procedure does not allow any classification to be obtained as soon as none of the spectra of the patient to be tested contributed to the building of the PLS-DA model.

In conclusion in this study FTIR imaging combined to computational algorithms was successfully used to distinguish the different main breast tissue cell types and normal from tumor epithelial cells. This method can be an effective aid for pathological diagnosis and more particularly to guide surgeons in the assessment of tumor margins.

Acknowledgments

The authors would like to acknowledge the patients who took part in this study.

This research has been supported by grants from the National Fund for Scientific Research (FRFC 2.4527.10, 2.4526.12 and T.0155.13), E.G. is Director of Research with the National Fund for Scientific Research (FNRS) (Belgium), M.V. is Research Fellow supported by the Fund for Research and Education within Industry and Agriculture (FRIA) from the FNRS (Belgium).

REFERENCES

1. G. Colditz and K. S. Chia, in *WHO Classification of Tumours of the Breast.*, eds. S. R. Lakhani, I. O. Ellis, S. J. Schnitt, P. H. Tan, and M. J. van de Vijver, IARC, Lyon, 4th edn., 2012, pp. 14–31.
2. J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin, in *Cancer incidence and mortality worldwide: IARC CancerBase No. 10 [internet], v1.2*, ed. GLOBOCAN (2008), Lyon, France, 2008.
3. M. Thill, K. Baumann, and J. Barinoff, *J. Surg. Oncol.*, 2014, 6.
4. P. Tian, W. Zhang, H. Zhao, Y. Lei, L. Cui, W. Wang, Q. Li, and Q. Zhu, *Int J Clin Exp Med*, 2015, **8**, 972–981.
5. D. R. McCready, *Ann Surg Oncol*, 2004, **11**, 885–7.
6. M. Morrow, E. A. Strom, W. Lawrence, D. D. Dershaw, B. Fowble, A. Giuliano, J. R. Harris, F. O. Malley, S. J. Schnitt, S. E. Singletary, and P. David, *CA Cancer J Clin*, 2002, **52**, 277–300.
7. Q. Li, X. Sun, Y. Xu, L. Yang, Y. Zhang, S. Weng, J. Shi, and J. Wu, *Clin. Chem.*, 2005, **51**, 346–350.
8. S. Salvatore, S. O. Hospital, M. M. Safety, G. Festini, and L. C-y, *Leuk. Res.*, 2003, **17**, 1670–1674.
9. X. Sun, Y. Xu, J. Wu, Y. Zhang, and K. Sun, *J. Surg. Res.*, 2013, **179**, 33–8.
10. Y. Liu, Y. Xu, Y. Zhang, D. Wang, D. Xiu, Z. Xu, X. Zhou, J. Wu, and X. Ling, *Br. J. Surg.*, 2011, **98**, 380–4.
11. S. E. Taylor, K. T. Cheung, I. I. Patel, J. Trevisan, H. F. Stringfellow, K. M. Ashton, N. J. Wood, P. J. Keating, P. L. Martin-hirsch, and F. L. Martin, *Br. J. Cancer*, 2011, **104**, 790–797.
12. C. Hughes, M. D. Brown, N. W. Clarke, K. R. Flower, and P. Gardner, *Analyst*, 2012, **137**, 4720–6.
13. A. Zwielly, S. Mordechai, G. Brkic, E. Bogomolny, I. Z. Pelly, R. Moreh, and J. Gopas, *Eur. Bophys. J.*, 2011, **40**, 795–804.
14. R. Bhargava, B. G. Wall, and J. L. Koenig, *Appl. Spectrosc.*, 2000, **54**, 470–479.
15. M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner, and F. L. Martin, *Nat. Protoc.*, 2014, **9**, 1771–91.
16. B. Bird, K. Bedrossian, N. Laver, M. Miljković, M. J. Romeo, and M. Diem, *Analyst*, 2009, **134**, 1067–76.
17. H. Fabian, P. Lasch, and D. Naumann, *J. Biomed. Opt.*, 2005, **10**, 031103.
18. M. J. Walsh, A. Kajdacsy-Balla, S. E. Holton, and R. Bhargava, *Vib. Spectrosc.*, 2012, **60**, 23–28.
19. G. J. Ooi, J. Fox, K. Siu, R. Lewis, K. R. Bambery, D. McNaughton, and B. R. Wood, *Med. Phys.*, 2008, **35**, 2151–61.
20. A. Kallenbach-Thieltges, F. Großerüschkamp, A. Mosig, M. Diem, A. Tannapfel, and K. Gerwert, *J. Biophotonics*, 2013, **6**, 88–100.
21. P. Lasch, W. Haensch, D. Naumann, and M. Diem, *Biochim. Biophys. Acta*, 2004, **1688**, 176–86.
22. J. Nallala, O. Piot, M.-D. Diebold, C. Gobinet, O. Bouché, M. Manfait, and G. D. Sockalingum, *Cytometry. A*, 2013, **83**, 294–300.
23. B. Bird, M. S. Miljković, S. Remiszewski, A. Akalin, M. Kon, and M. Diem, *Lab. Invest.*, 2012, **92**, 1358–73.
24. M. J. German, A. Hammiche, N. Ragavan, M. J. 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**, 3783–95.

25. M. J. Baker, E. Gazi, M. D. Brown, J. H. Shanks, P. Gardner, and N. W. Clarke, *Br. J. Cancer*, 2008, **99**, 1859–66.

26. M. J. Baker, E. Gazi, M. D. Brown, J. H. Shanks, N. W. Clarke, and P. Gardner, *J. Biophotonics*, 2009, **2**, 104–13.

27. 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. Urol.*, 2006, **50**, 750–60; discussion 760–1.

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

29. B. R. Wood, K. R. Bambery, C. J. Evans, M. A. Quinn, and D. Mcnaughton, *BMC Med. Imaging*, 2006, **6**, 1–9.

30. M. J. Walsh, M. J. German, M. Singh, H. M. Pollock, A. Hammiche, M. Kyrgiou, H. F. Stringfellow, E. Paraskevaidis, P. L. Martin-Hirsch, and F. L. Martin, *Cancer Lett.*, 2007, **246**, 1–11.

31. K. Gajjar, J. Trevisan, G. Owens, P. J. Keating, N. J. Wood, H. F. Stringfellow, P. L. Martin-Hirsch, and F. L. Martin, *Analyst*, 2013, **138**, 3917–26.

32. J. Ollesch, S. L. Drees, H. M. Heise, T. Behrens, T. Brüning, and K. Gerwert, *Analyst*, 2013, **138**, 4092–102.

33. E. Scaglia, G. D. Sockalingum, J. Schmitt, C. Gobinet, N. Schneider, M. Manfait, and G. Thiéfin, *Anal. Bioanal. Chem.*, 2011, **401**, 2919–25.

34. J. R. Hands, P. Abel, K. Ashton, T. Dawson, C. Davis, R. W. Lea, A. J. S. McIntosh, and M. J. Baker, *Anal. Bioanal. Chem.*, 2013, **405**, 7347–55.

35. A. Akalin, X. Mu, M. A. Kon, E. Aysegül, S. H. Remiszewski, C. M. Thompson, D. J. Raz, and M. Diem, *Lab. Investig.*, 2015, **95**, 406–421.

36. A. A. Chichon, A. C. Degnim, D. W. Visscher, and D. C. Radisky, *J. Mammary Gland Biol. Neoplasia*, 2010, 15, 389–397.

37. M.-Q. Gao, B. G. Kim, S. Kang, Y. P. Choi, H. Park, K. S. Kang, and N. H. Cho, *J. Cell Sci.*, 2010, **123**, 3507–14.

38. P. Micke and A. Ostman, *Lung Cancer*, 2004, **45 Suppl 2**, S163–75.

39. S. Selvey, L. M. Haupt, E. W. Thompson, K. I. Matthaei, M. G. Irving, and L. R. Griffiths, *BMC Cancer*, 2004, **4**, 40.

40. N. Garamszegi, S. P. Garamszegi, L. A. Shehadeh, and S. P. Scully, *Mol. Cancer Res.*, 2009, **7**, 319–29.

41. S. Kumar, C. Desmedt, D. Larsimont, C. Sotiriou, and E. Goormaghtigh, *Analyst*, 2013, **138**, 4058–65.

42. R. Kalluri and M. Zeisberg, *Nat. Rev. Cancer*, 2006, **6**, 392–401.

43. C. J. Olsen, J. Moreira, E. M. Lukanidin, and N. S. Ambartsumian, *BMC Cancer*, 2010, **10**, 444.

44. E. Goormaghtigh and J. M. Ruysschaert, *Spectrochim.Acta*, 1994, **50A**, 2137–2144.

45. E. Goormaghtigh, in *Adv. Biomed. Spectrosc. (Biological and Biomedical Infrared Spectroscopy)*, eds. A. Barth and P.I.Haris, IOS Press, 2009, vol. IOS Press, pp. 104–128.

46. C. Sandt, O. Féraud, N. Oudrhiri, M. L. Bonnet, M. C. Meunier, Y. Valogne, A. Bertrand, M. Raphaël, F. Griscelli, A. G. Turhan, P. Dumas, and A. Bennaceur-Griscelli, *PLoS One*, 2012, **7**, e30743.

47. F. Pagès, J. Galon, M.-C. Dieu-Nosjean, E. Tartour, C. Sautès-Fridman, and W.-H. Fridman, *Oncogene*, 2010, **29**, 1093–102.

48. W. H. Fridman, J. Galon, F. Pagès, E. Tartour, C. Sautès-Fridman, and G. Kroemer, *Cancer Res.*, 2011, **71**, 5601–5.

49. J. Galon, F. Pagès, F. M. Marincola, M. Thurin, G. Trinchieri, B. A. Fox, T. F. Gajewski, and P. A. Ascierto, *J. Transl. Med.*, 2012, **10**, 1.

50. R. Salgado, C. Denkert, S. Demaria, N. Sirtaine, F. Klauschen, G. Pruneri, S. Wienert, G. Van den Eynden, F. L. Baehner, E. A. Perez, E. A. Thompson, W. F. Symmans, A. L. Richardson, J. Brock, C. Criscitiello, H. Bailey, M. Ignatiadis, G. Floris, J. Sparano, Z. Kos, T. Nielsen, D. L. Rimm, K. H. Allison, J. S. Reis-Filho, S. Loibl, C. Sotiriou, G. Viale, S. Badve, S. Adams, K. Willard-Gallo, and S. Loi, *Ann. Oncol.*, 2015, **26**, 259–271.

51. C. Matthäus, B. Bird, M. Miljkovic, T. Chernenko, M. Romeo, and M. Diem, *Methods Cell Biol.*, 2008, 275–308.

52. G. Bellisola and C. Sorio, *Am. J. Cancer Res.*, 2012, **2**, 1–21.

53. D. R. Whelan, K. R. Bambery, P. Heraud, M. J. Tobin, M. Diem, D. McNaughton, and B. R. Wood, *Nucleic Acids Res.*, 2011, **39**, 5439–48.

54. S. E. Holton, M. J. Walsh, A. Kajdacsy-Balla, and R. Bhargava, *Biophys. J.*, 2011, **101**, 1513–21.

55. P. Bassan, A. Sachdeva, J. H. Shanks, M. D. Brown, N. W. Clarke, and P. Gardner, *Analyst*, 2013, **138**, 7066–9.

56. P. Bassan, M. J. Weida, J. Rowlette, and P. Gardner, *Analyst*, 2014, **139**, 3856–9.

57. K. Yeh, S. Kenkel, J.-N. Liu, and R. Bhargava, *Anal. Chem.*, 2015, **87**, 485–493.

58. N. Kröger-Lui, N. Gretz, K. Haase, B. Kränzlin, S. Neudecker, A. Pucci, A. Regenscheit, A. Schönhals, and W. Petrich, *Analyst*, 2015, **140**, 2086–2092.

59. B. Bird and M. J. Baker, *Trends Biotechnol.*, 2015, **33**, 557–8.

60. M. Verdonck, N. Wald, J. Janssis, P. Yan, C. Meyer, A. Legat, D. E. Speiser, C. Desmedt, D. Larsimont, C. Sotiriou, and E. Goormaghtigh, *Breast cancer and melanoma cell line identification by FTIR imaging after formalin-fixation and paraffin-embedding.*, 2013, vol. 138.

61. M. Diem, M. Romeo, S. Boydston-White, M. Miljkovic, and C. Matthaus, *Analyst*, 2004, **129**, 880–5.

62. A. Benard, C. Desmedt, M. Smolina, P. Szternfeld, M. Verdonck, G. Rouas, N. Kheddoumi, F. Rothé, D. Larsimont, C. Sotiriou, and E. Goormaghtigh, *Analyst*, 2014, **139**, 1044–56.

63. R. M. Neve, K. Chin, J. Fridlyand, J. Yeh, F. L. Baehner, T. Fevr, L. Clark, N. Bayani, J.-P. Coppe, F. Tong, T. Speed, P. T. Spellman, S. DeVries, A. Lapuk, N. J. Wang, W.-L. Kuo, J. L. Stilwell, D. Pinkel, D. G. Albertson, F. M. Waldman, F. McCormick, R. B. Dickson, M. D. Johnson, M. Lippman, S. Ethier, A. Gazdar, and J. W. Gray, *Cancer Cell*, 2006, **10**, 515–27.

64. K. Polyak, *J. Clin. Invest.*, 2011, **121**, 3786–8.

65. A. Prat and C. M. Perou, *Mol. Oncol.*, 2011, **5**, 5–23.

66. D. Hanahan and R. A. Weinberg, *Cell*, 2000, **100**, 57–70.

67. D. Hanahan and R. A. Weinberg, *Cell*, 2011, **144**, 646–674.

68. M. Samoszuk, J. Tan, and G. Chorn, *Breast cancer Res.*, 2005, **7**, R274–R283.