Analyst

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Journal Name

RSCPublishing

ARTICLE

Infrared imaging of primary melanoma reveals hints of regional and distant metastases

Cite this: DOI: 10.1039/x0xx00000x

N. Wald [a] and E. Goormaghtigh[a]

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Melanoma is the deadliest form of skin cancer. Metastatic melanomas are resistant to almost all existing adjuvant therapies such as chemotherapy and radiotherapy. Yet, detection of metastases remains a challenge, and no biomarkers are currently available to detect primary tumors with the highest risk of metastatization. Results presented in this paper show that Fourier Transform Infrared (FTIR) imaging of histological sections followed by supervised Partial Least Squares Discriminant Analysis (PLS-DA) can accurately (>90%) identify the main cell types commonly found in melanoma tumors. Here we define six cell types: melanoma cells, erythrocytes, connective tissue (includes blood vessel walls, dermis and collagen regions), keratinocytes, lymphocytes and necrotic cells. Interestingly, more than 98% of the melanoma cells are correctly identified. The spectra of the cells identified as melanoma were then further analyzed. First, we compared melanoma cells in primary tumors (from 26 patients) with melanoma cells from metastases (from 25 patients). Neither supervised nor unsupervised analyses revealed any significant difference. Similarly, we found no significant correlation between the infrared spectra of melanoma cells and the number of proliferative cells assessed by Ki67 immunostaining. Finally, we compared the infrared spectra of primary tumors from patients diagnosed at different stages of the disease. Infrared spectroscopy was capable to point out a difference between primary tumors of patients at stages I or II and patients at stages III or IV, even with unsupervised analysis. We then developed a supervised PLS-DA model capable of predicting whether tumor cells belonged to one of the two aggregated disease stage groups. The model predicted a high rate of true positives (sensitivity of 88.9%) and with a good rate of true negatives (specificity of 70.6%) in external validation. These results demonstrated that infrared spectroscopy can be used to help identify melanoma characteristics related to the cells' invasive capability.

Introduction

Melanoma is the deadliest form of skin cancer, accounting for 80% of deaths due to dermatologic cancer.[1] The incidence of melanoma is rising faster than that of any other solid cancer.[2] According to the American Joint Committee on Cancer (AJCC), the main criteria used for staging patients with melanoma are tumor thickness in mm (Breslow index), ulceration, mitotic rate, lymph node involvement and distant metastases status.[3,4] If diagnosed early, localized melanoma are treated by surgery. Eighty percent of the patients are dealt with in this way.[5] Survival for these patients at stages I and II (with no evidence for metastasis) is relatively high. The five-year survival rate is around 92% and 68% for stage I and stage II patients respectively.[4] By contrast, metastatic melanoma are resistant to almost all existing therapies and survival rates are poor for patients at stages III or IV, with a five-year survival rate below 10 % for stage IV patients.[4]

Biological modifications occurring in the primary tumor that lead to dissemination of cells at distant sites are now better understood. It has been shown that cells can acquire characteristics of invasiveness that allow them to detach from the primary tumor and migrate through the blood and lymphatic circulation. These characteristics include loss of adherent junctions, alteration of shape, expression of matrix-degrading enzymes, increased motility and increased resistance to apoptosis.[6,7] Currently, there are no biomarkers available to detect primary tumor with highest risk of metastatization even if some preliminary studies have shown a correlation between expression of some cadherins and the metastatization in melanoma.[8]

Fourier Transform Infrared (FTIR) spectroscopy coupled to microscopy has been recognized as an emerging tool for histopathological studies.[9–11] This technique, based on the absorption of the infrared light by vibrational transitions of covalent bonds that compose macromolecules of tissues, offers

a unique signature of the cells. FTIR imaging thus provides spatially resolved information on the chemical composition of the sample originating from all macromolecules (e.g. DNA, RNA, lipids, proteins and carbohydrates) that compose cells and tissues, without any staining required.[12] Infrared spectra are also sensitive to the conformation of molecules, in particular to the secondary structure of proteins.[13] Therefore, spectral features of cells and tissues can be correlated to biological properties.

Various studies have demonstrated the sensitivity of infrared spectral signatures for observing biological changes. In cancer research particularly, several publications have shown spectral differences between healthy and cancerous tissues.[14–16] Furthermore, numerous previous studies have demonstrated the capability of infrared spectroscopy to distinguish different cellular components of a wide diversity of tissues, based only on the spectral differences resulting from differences in the biochemical composition of cells.[14,17,18]

In this study, we investigated the potential of FTIR imaging as an automated tool for classifying the major biological cell types present in melanoma biopsies. We built supervised statistical models to identify the various cell types. We then focused on the melanoma cells and attempted to correlate their spectral properties with the presence of regional and distant metastases in the patients. Importantly, we developed a supervised statistical model that allowed a good prediction of the presence of regional and distant metastases.

Materials and Methods

TMA description

Tissue microarrays (TMA) sections were purchased from US Biomax, Inc. (Rockville, USA) and obtained from a formalin-fixed and paraffin-embedded block referred as number Me2082b. Three adjacent sections of 3-5 µm were ordered. One section was H&E stained and two were unstained. The two unstained sections were deposited respectively on a glass for immunostaining and on a BaF_2 slide for infrared spectroscopy.

An overall view of the H&E stained section of the TMA used in this work is presented in Figure 1. The TMA contains 208 cores, each measuring c.a. 1 mm in diameter, from biopsies of 208 donors. 128 cores are from primary tumors of melanoma from patients diagnosed at different stages of the disease (I, II, III and IV). 64 cores are from regional (lymph nodes) or distant metastases and 16 cores are from skin healthy tissues. Information on stage is reported according the American Joint Commission on Cancer (AJCC) Tumor-Node-Metastasis (TNM) system.[3] The TNM system contains three key pieces of information. T stands for tumor features and is assigned a number from one to four based on the tumor's thickness. N stands for spread to nearby lymph nodes. The N category is assigned a number from zero to three based on whether the melanoma cells have spread to one or several lymph nodes. The M category is based on whether the melanoma has metastasized to distant organs and depends on the site of distant metastases. Stage grouping is based on the TNM status and aims to group patients who are prognostically and therapeutically similar. The grouping must ensure homogeneous survival among each group.[4] Localized melanomas are divided into stage I and stage II, according to tumor thickness and ulceration status. Melanoma ulceration is defined as the absence of an intact epidermis above a significant part of the primary tumor.[19] Stage I is composed of patients with a primary tumor with a thickness below 2 mm and stage II is composed with patients with a primary tumor thicker than 2 mm. Stage III consists of patients with evidence for regional metastases (involving lymph node metastases). Stage IV is composed of patients with evidence for distant metastases (nonvisceral, lung metastases or visceral metastases).[3,4] TMN staging for each patient is reported in Supplementary Materials. Information about the organ where the tissue was found is also reported in Supplementary Materials.

FTIR measurements

The IR data were collected using a Hyperion 3000 IR imaging system (Bruker Optics, Ettlingen, Germany), equipped with a liquid nitrogen cooled 64x64 Mercury Cadmium Telluride (MCT) Focal Plane Array (FPA) detector, in transmission mode. The size of an image covers an area of 180x180 μm^2 composed of 4096 pixels of 2.8x2.8 μm^2 each. It must be noted that spatial resolution can be significantly lower than the pixel size, depending on the wavelength. It took about 5 minutes to record an infrared image composed of 4096 spectra at a spectral resolution of 8 cm^{-1} and where each spectrum is the average of 256 scans. Areas of measurement are shown with colored squares in the Figure 1. They were selected to encompass a wide diversity of cell types and a large number of patients. FTIR measurements were acquired on the unstained and deparaffinized tissue section deposited on a BaF_2 slide (ACM, Villiers St. Frederic, France).

Data analysis

Preprocessing

All spectra were preprocessed as follows. Water vapor contribution was subtracted as previously described[20,21] with 1956-1935 cm^{-1} as reference peak and CO_2 peak was flattened between 2450 and 2250 cm^{-1}. The spectra were baseline-corrected. Straight lines were interpolated between the spectra 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 systematically checked on every spectrum. It was required to be greater than 300 when noise was defined as the standard deviation in the 2000-1900 cm^{-1} region of the spectrum and the signal was the maximum of the curve between 1750 and 1480 cm^{-1} after subtracting a baseline passing through these two points. Finally, some rare spectra with normalized absorbance lower than -5 (negative lobe) and a maximum above 120 (saturation) were discarded. To minimize artefacts related to

abrupt refractive index changes we always selected areas of sample with contiguous cells. Visual inspection of spectra as well as systematic screening for negative lobes on the left hand side of the Amide I band did not reveal significant dispersive artifacts.

Difference spectra

Difference spectra allow emphasizing the spectral variations between two distinct conditions. Difference spectra were built by subtraction of spectra from melanoma cells of patients of stages III or IV to those from patients of stages I or II. A Student t-test was used to reveal significantly different wavenumbers between the two groups and are shown by black stars (α= 1%).

Statistical analyses

We performed unsupervised analysis in order to observe the intrinsic proximities and distances within the data set and group spectra according to their similarities. Principal component analysis (PCA) was applied, this statistical analysis is already described elsewhere.[22] Multivariate Analysis of Variance (MANOVA) is a statistical test capable to assess whether vectors of means of two or more groups are sampled from the same distribution. In the present paper, MANOVA was computed on the first three PCs scores obtained for each mean spectrum. Supervised analysis, such Partial Least Squares Discriminant Analysis (PLS-DA) was also conducted on the data set. PLS-DA was used to extract latent variables of the data set that enable the construction of a factor capable of predicting a class. Partial Least Squares Regression (PLSR) was used to correlate infrared spectrum of melanoma cells to proliferative rate. The purpose of this statistical analysis is to build a multivariate linear model that predicts a Y-variable (here, the proliferative rate) based on many X-variables which are highly correlated (IR absorbances). To build this model, new artificial variables are defined (PLS components) which are linear combinations of the X-variables. The coefficients of these PLS components are calculated taking into account matrix of variance and covariance of the X and Y variables.

Correction of the spectra for water vapor contribution, baseline subtraction and normalization, PCA and PLS-DA were carried out by Kinetics, an in-house program, running under Matlab (Matlab, Mathworks Inc).

Immunofluorescent staining targeting antigen Ki67

Ki67 protein is a proliferation marker, expressed during all phases of the cell cycle except the G0 phase.[23] An immunostaining targeting Ki67 was performed on the unstained section of the TMA deposited on the glass slide (US Biomax Inc., Rockville, USA). First, paraffin was removed by incubation in four successive baths of xylene during 20 minutes in total. The section was then rehydrated by 3 successive baths with a decreasing gradient of ethanol (100%, 90%, 70%, 5 minutes each) and followed by incubation in milliQ water (10 minutes). Slide underwent heat-induced epitope retrieval by incubation in citrate buffer diluted 10 times (stock solution: citrate buffer 10 X, pH 6, Skytek Laboratories, Logan, USA) at 95°C in a water-bath during 30 minutes and was then washed in distilled water for 5 minutes. Tissue was surrounded by a hydrophobic trait of Dakopen (Dako, Santa Clara, USA). Then non-specific sites were blocked with a 30 minute incubation in a solution of BSA (Bovine Serum Albumin; 1% w/w, Sigma, Saint-Louis, USA) in DPBS buffer (Lonza, Basel, Switzerland). Primary antibody anti-Ki67 rabbit monoclonal IgG (ab16667, Abcam, Cambridge, UK) was incubated in a moist chamber at 4°C overnight with a dilution of 1:50. Slide was washed 3 times during 5 minutes in DPBS. Secondary antibody with a fluorescent probe (Dylight™ 594) goat anti-rabbit IgG (35561, Thermo Scientific, Waltham, USA) was incubated at room temperature during two hours (dilution 1:200, concentration 5µg/mL, pH 7.2). Two tonsil sections were used as negative and positive controls. Negative control was incubated without the primary antibody. The TMA slide was scanned with a NanoZoomer 2.0-RS equipped with a fluorescence imaging module (C10730, Hamamatsu, Hamamastu, Japan). Images were analyzed by Ndpview program (Hamamastu, Hamamstu, Japan). Positive nuclei were counted with Image J program. Counts were expressed by number of positive nuclei for an area of measurement of 129,600 μm^2 and are shown in Figure S1 (supplementary material).

Results

A tissue microarrays (TMA) section composed of 208 cores obtained from 208 donors was analyzed by FTIR imaging. The adjacent section, H&E stained, was used as a reference (Figure 1). 91 areas of measurement were selected following the observation of the H&E stained section with the help of a pathologist. These areas were selected to encompass the diversity of cell types found in the tissue section and to include a large number of patients. In total, spectra from 81 patients were included in the present study.

Major cell type identification

In a first step we attempted to build a tool to identify automatically from the IR spectral features the different cell types present in the histological section described in Figure 1. For this purpose, a spectral database containing spectra from the various cell types composing the tissue cores was built. The strategy used to accumulate spectra from each cell type is schematically illustrated in Figure 2 for the erythrocytes. First, erythrocytes were identified on the adjacent H&E stained section, then the same area in the unstained section was identified and an IR image (usually 4096 spectra) was recorded. Spectra from the whole image were processed as described in Materials and Methods (subtraction of water vapor contribution, baseline subtraction, normalization) and spectra were submitted to a quality test as described. Only spectra passing the quality test thresholds were retained. The spectra of

the erythrocytes were then extracted manually from the image either in a delimited area, or one by one when necessary (around 500 spectra per measurement). It must be noted that spectra that did not pass the thresholds appear in black in Figure 2 (IR image of the unstained section) and cannot be selected. The same operation was repeated for keratinocytes, lymphocytes, dermis, blood vessel walls, melanoma cells, necrotic cells and fibrous regions (collagen). Table 1 describes the composition of the spectral database in term of histological class, number of spectra and number of patients used for each class. It must be noted that 10 patients among the 81 were used twice for a different histological class. For each of the 91 measurements achieved, a mean spectrum was calculated from the extracted spectra (c.a. 500 spectra per patient).

In a first exploratory study, the 91 mean spectra obtained for the six cell types were submitted to a principal component analysis (PCA) on the 1800-950 cm^{-1} spectral region of the IR spectra. A score plot is reported in Figure 3. The projection on the first, second and third principal components demonstrates some degree of separation (Figure 3). Some cell types are particularly well separated from the others (keratinocytes (yellow), connective tissue (pink) and melanoma cells (green)) but some other cell types such as erythrocytes (blue), lymphocytes (purple) and necrotic cells (orange) separate only partially. We also averaged spectra from each class to obtain a representative spectrum for each individual cell type. These mean spectra are represented in Figure 4. The very different chemical nature of keratinocytes must be noted. A significant shift can be observed in the amide I and amide II bands, reflecting a particular protein secondary structure. The difference in chemical composition is also very apparent in the C-H stretching region of the spectrum. We can clearly identify major spectral differences. In order to optimize separation and build a tool for cell type recognition in images, a Partial Least Squares Discriminant Analysis (PLS-DA) model was built using the individual spectra of the database (instead of the means as described in Figure 3) using two combined spectral regions, 3000-2800 and 1800-950 cm^{-1}. Recognition rates of the model were high, ranging from 91 % (for melanoma cells) to 100 % (for keratinocytes; data not shown). To evaluate the robustness of the PLS-DA model, we performed a cross-validation. Six different models were built excluding each time spectra from six patients corresponding each to an histological class. The six PLS DA models were trained with all IR spectra except those from 6 patients (one for each cell type) and validated on the spectra of the patients left. This was done six times with different patients in training sets and test sets in order that each patient was included once in the test set. Results are presented as a confusion matrix in Table 2. This table shows the average percentages of correct prediction obtained for the 6 models and the standard deviations.

A more qualitative way of validating the model was also performed. Each of the six models was applied to entire images containing different cell types from patients who did not contribute to the model. Figure 5 shows an example of application of one of the six models on six images, each mostly composed with one cell type. Each pixel is associated with a color according to the membership of the spectrum to a class. H&E equivalent sections are presented next to the false-color infrared images. At first glance, the different models have a good recognition rate. Overall, melanoma cells, keratinocytes, necrotic cells, erythrocytes and lymphocytes are well identified but some connective tissue is identified as melanoma cells. This slight confusion is also observed in the confusion matrix shown in Table 1 and are discussed later. In conclusion, these results show that overall the PLS-DA models allow a rather accurate identification of the main cell types present in these tissue cores. Now that melanoma cells can be identified on the tissue sections, we will investigate in more details biological variability occurring among tumoral cells.

Comparison of melanoma cells in the primary tumor and in metastases

In this second part of the work we attempted to compare the spectra of melanoma cells in primary tumor and in metastases coming from different patients. Around 25 biopsies from 25 patients have been considered for each class i.e. 26 primary tumors and 25 metastases. In order to evaluate if some differences could exist between spectra of melanoma from primary tumors and metastases, we carried out a Principal Component Analysis on the mean spectra, each one being the average of the ca. 15,000 spectra recorded by patients. The score plot was not able to show any separation between these two classes (data not shown). Furthermore a supervised analysis (PLS-DA) carried out with an external validation (patients who did not participate to the training set) did not result in significant separation.

It can be concluded that, at the level of experimental error characterizing our measurements, there is no significant difference between the IR spectra of melanoma cells in the primary tumors and in the metastases. This suggests that melanoma cells exhibit essentially the same phenotype in both locations. This will be discussed in more details later. It must also be pointed out that this negative result is also an interesting control demonstrating that PLS-DA does not recognize groups that do not have a significant molecular difference (overfitting).

Correlation between Ki67 expression rate and infrared signature

In this part of the work, we attempted to correlate Ki67 expression rates with infrared spectra of melanoma cells, both in primary tumors and in metastases. Ki67 is a biomarker of proliferation as it is expressed by cells during all phases of the cell cycle except the G0 phase (G1, S, G2/M). Some studies have demonstrated the role of Ki67 as a prognostic biomarker (survival and risk of metastasis) in melanoma primary tumors.[23–25] To evaluate Ki67 expression rate, a fluorescent immunostaining of the Ki67 antigen was achieved on an adjacent section of the TMA as described in Materials and Methods. The number of cells positively stained on a surface of 129,600 μm^2 has been counted for each patient's section. This

area corresponds to the area analyzed by FTIR imaging. Potential correlation was tested between the number of cells in division (i.e. expressing Ki67 antigen) and the infrared spectra. Absorbances or ratio of absorbances at different wavenumbers characteristic of nucleic acids have been tested (e.g. 1244, 1230, 1121, 1080 or 1020 cm^{-1}) but no correlation with the proliferative rate was found. Moreover, Partial Least Squares Regressions (PLSR) have been built at every wavenumber between 3000 and 900 cm^{-1}. Again no significant correlation was observed.

Melanoma spectral sub-classification according to cancer stages

In this last part of the work we addressed the question of the possibility to obtain sub-classification of melanoma cells from the primary tumor according to cancer stage. Melanoma stages are described in Materials and Methods and reported in Table S1 (supplementary material) for the different samples present on the TMA used for this study (Figure 1). The 26 mean spectra obtained from the primary tumors and described above (one mean spectrum is the average of around 15 000 spectra per patient) were subjected to PCA (Figure 6). Some limited separation occurs on the score plot between patients of stages I or II (non-metastatic stages) on the one hand (green color dots) and patients of stages III or IV (metastatic stages) on the other hand (red color dots) when the mean spectra are projected in the first three principal component space. A Multivariate Analysis of Variance (MANOVA) was computed on the first three principal components of the PCA. P-value obtained was lower than 0.002, indicating that these two groups are significantly different. A difference spectrum (non-metastatic versus metastatic) was also performed to highlight spectral difference between these two groups (Figure 7). To build this spectrum, the mean spectrum from melanoma cells of patients from stages III or IV was subtracted from the mean spectrum from melanoma patients from stages I or II. A Student t-test was applied to compare the two groups. The significantly different wavenumbers are identified by black stars (alpha= 1%). As the difference spectrum is ten times less intense than mean spectra, we can assess that the difference between the two groups are moderate. These differences mainly occurred in the protein region (amide I and II bands) and are probably due to a change in the secondary structure of proteins.

In order to create a tool able to predict the presence of metastases with only the spectral signature of the primary tumor itself, we built a PLS-DA model with the individual spectra recorded from patients at different stages of the disease. To test the predictive power of the model, we built 26 models, each time leaving out spectra from one of the 26 patients. These models were trained on subsets of spectra (c.a. 500 spectra per patient). Each model was validated on the c.a. 150,000 spectra from the patient who did not train the model (entire infrared image). Results of this validation are schematically represented in Figure 8. Green columns of this graph represent percentages occupied by pixels correctly predicted as originating from non-metastatic patients for patients with no evidence for metastases (stage I and stage II patients) and red columns represent percentages occupied by pixels correctly predicted as originating from metastatic patients for patients with evidence for regional or distant metastases (stage III and stage IV patients). A threshold of 30% of the melanoma cells characterized by pixels predicted as metastatic has been set. It is shown in Figure 8 by a dark blue line. Above this threshold patient has to be considered as metastatic. The corresponding threshold for non-metastatic patients is also represented by a dark blue line at 70%. The ROC curve presented in Figure 9 reports the true positive rate as a function of false positive rate (1-specificity). This curve allows the evaluation of the performance of a diagnostic model. The closer the curve will be to the left vertical axis, the more performant the test. Area under the curve (AUC) shows accuracy of the model and reaches here a value of 0.84 that corresponds to a good model. This value can be compared with 1 for a perfect predictive model or with 0.5 for a random model (also shown by a diagonal straight line y=x) in Figure 9.

Figure 10 shows two examples of application of one PLS-DA model on areas containing only melanoma cells from patients C5 and D5 who did not participate to the training of the model (external validation). Correct assignment to metastatic/non metastatic melanoma cells is found for 98.5 and 98.7 % of the image pixels for patients C5 and D5 respectively, which is significantly superior to random assignment (50%). Examination of Figure 10 indicates that incorrectly assigned spectra are scattered throughout the tumor. At this point it is not possible to determine whether the incorrectly assigned spectra reflect a failure of PLS-DA to achieve a perfect assignment or a real heterogeneity of the tumor. Indeed, tumors are known to be heterogeneous and a mix of phenotypes is more likely than a homogenous population.

Figure 11 presents an example of global analysis including cell type recognition followed by subclassification of melanoma cells on an entire core. Two successive PLS-DA models were applied. The tested patient did not contribute to any of the two training sets. This tissue core originated from a patient with a skin melanoma (E1) diagnosed as a stage II by pathologists (T4N0M0). Figure 11 shows the H&E stained section of this tissue core (A) and the equivalent false-color infrared image (B). The first model applied on this image allowed the identification of the six main biological classes defined as melanoma cells, connective tissue, erythrocytes, lymphocytes, keratinocytes and necrotic cells. In this case, the tissue core was predicted as being mainly composed of connective tissue (63.1%, in pink) and melanoma cells (26.8 %, in light and dark green). The second PLS-DA model was applied on spectra just assigned to melanoma cells to predict the potential of metastatization of this primary tumor. Pixels assigned as non-metastatic are shown in light green whereas pixels assigned as metastatic are shown in dark green. In this case, 82.2% of the melanoma pixels are correctly recognized as non-metastatic (light green). The figure illustrates the possibility to run

automatically successive analyses, each one taking into account the result of the previous one.

Discussion

Metastatic melanomas are refractory to most current adjuvant treatments and patients with regional and distant metastases show a quite low survival rate. The median survival rate is 6 months and less than 5% of stage IV patients reach five years.[5] Fifteen to 20% of melanoma patients will develop metastases to regional lymph node.[26] Due to the wide diversity in the outcomes and in the administered treatments, an accurate diagnosis is an essential step. Sentinel-node examination is a crucial step in the diagnosis as involvement of lymph node is the most important prognostic factor for the overall survival of patients.[27] Yet, detection of metastases remains a challenge, and no biomarkers are currently available to detect primary tumors with the highest risk of metastatization.

Results presented in this paper show that we could develop a supervised PLS-DA statistical model allowing an accurate (>90%) identification of the main cell types commonly found in melanoma tumor. Six cell types have been defined here: melanoma cells, erythrocytes, connective tissue (includes blood vessel walls, dermis and collagen regions), keratinocytes, lymphocytes and necrotic cells. Interestingly, more than 98% of the melanoma cells are correctly identified (Table 2). Yet, it must be said that there is some degree of confusion between a few percentages of the connective tissue sometimes recognized as melanoma cells. The misidentification appeared only for the subclass of the connective tissue composed of blood vessels spectra. Neither collagen nor dermis spectra were misidentified. More particularly spectra misidentified for the blood vessels were those at the inside edge of the vessel. These few misidentified spectra were not noisier but showed spectral differences throughout the spectral regions 3000-2800 and 1800-950 cm^{-1}. The interior surface of vessel is lined with a layer of endothelial cells. Therefore misidentification could be either due to the model itself or to spectral similarities occurring between melanoma cells and endothelial cells. This overall result is interesting in its own right as it provides an automated quantified analysis of the cell types, including for lymphocyte infiltration. It is also important as the "melanoma" category so identified can be submitted to further analyses.

First, we have compared melanoma cells in primary tumors (from 26 patients) with melanoma cells from metastases (from 25 patients). Neither supervised nor unsupervised analyses revealed any significant difference. This suggests that melanoma cells exhibit very similar phenotypes in both locations. The literature[6,7] about phenotypic changes observed during the invasion-metastases cascade and the colonization step may explain this result. Indeed, invasiveness is accompanied by some capability acquired through the activation of the so-called EMT (epithelial-mesenchymal transition). This transition implies activation of some traits (loss of adherent junctions, alteration of shape, expression of matrix-degrading enzymes, increased motility and increased resistance to apoptosis), all needed for the process of metastatic dissemination. At the opposite, the colonization process implies that the cells pass through the reverse process known as MET (mesenchymal-epithelial transition). This recovery of a non-invasive phenotype may result in the formation of new tumor colonies of carcinoma cells exhibiting a phenotype similar to the one found in primary tumor.[7] Moreover some genomic studies have shown that primary tumors and metastases exhibit high genomic concordance in cancer.[28] However, this result was obtained for melanoma and was achieved on 51 biopsies and with specific statistical analyses (PCA and PLS-DA mainly). A different result could be obtained for other cancer types. Furthermore, only linear correlations have been considered. Other approaches could reveal more subtle differences.

In a second step, we attempted to correlate the infrared spectral signature of different tumors (primary tumors and metastases) with the proliferation rate. The number of proliferative cells was assessed by Ki67 immunostaining. Indeed Ki67 is expressed during all phases of the cell cycle except the G0 phase. No significant correlation between these two parameters was found. An explanation could be that the fraction of cells in division is in general quite low. The average percentage of cells positively stained among all cells was generally around 3.3 %. In this study we could not identify spectra of single cells with and without Ki67 expression. Such identification would require 1) an immunostaining of the tissue section previously analyzed by FTIR imaging, 2) image registration to assign every pixel of the IR image to a specific value of the Ki67 expression. Furthermore, due to the diffraction limited resolution, the poor spatial resolution achieved in the infrared (roughly in the range of 10 µm) would not allow a precise overlap of highly localized features. The results described here rather report the average signature of all the tumoral cells of each tissue section. In addition, even if some previous publications have shown that contribution of the cycle to infrared spectra signature is identifiable, these spectral differences are quite small.[29–33] Furthermore, Ki67 positive cells remain a minor fraction of the total cell population present on a section (around 3.3%) reducing the likeliness of FTIR to detect such a small spectral difference.

The last step of this work aimed to compare the infrared spectra of primary tumors from patients diagnosed at different stages of the disease. Surprisingly we were able to observe a difference between primary tumors of patients at stages I or II and patients at stages III or IV even with unsupervised analysis (Figure 6). We then developed a supervised PLS-DA model with the capability of predicting membership to one of the two groups. The model was able to predict high rate of true positives (sensitivity of 88.9%) and with a good rate of true negatives (specificity of 70.6%) in external validation. The major biological difference between primary tumors of patients at stages I or II and patients at stages III or IV is the capability of some cells to disseminate from the primary tumor to a distant site. While the majority of the cells present in melanoma do not have a mesenchymal phenotype, infrared spectroscopy can

identify features related to their acquired potential to invade neighboring tissues. The biological nature of the markers identified by infrared spectroscopy remains to be identified. Misrecognition of some patients by the PLS-DA model could be due to either a limitation of the model or to the heterogeneity of the tumor which include cells at different step of their progression towards more malignant phenotypes. It is known that not all the cells in the primary tumor will acquire ability to migrate and invade.

Although, recent papers have shown that studies based on "tissue microarrays" are sufficiently representative of the entire tumor,[34,35] the issue of representativeness of the results remains as tumors are not homogeneous and these results were based on the study of subsets of cells present in areas of around 0.13 mm^2.

Currently, evaluation of dissemination of the primary tumor is achieved by surgical removing of the sentinel node and by the detection of melanoma cells inside the lymph node. If the lymph node is positively invaded, others nodes are then removed[36]. A promising application of our results could be to help predict the presence of metastases with only the spectral signature of the primary tumor. Yet, the results described here have been obtained on a limited number of patients (26 primary tumors distributed in the 2 classes as described in Table 1) and extending this study to a much larger number of patients is necessary.

Acknowledgements

This research has been supported by grants from the National Fund for Scientific Research Research (FRFC 2.4533.10 and 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), N.W. is Research Fellow supported by the Fund for Research and Education within Industry and Agriculture (FRIA) from the FNRS (Belgium).

References

[a] Center for Structural Biology and Bioinformatics, Laboratory for the Structure and Function of Biological Membranes; Université Libre de Bruxelles, Campus Plaine, Bld du Triomphe 2, CP206/2, B1050 Brussels, Belgium.

Abbreviations

AJCC, American Joint Committee on Cancer, FPA, Focal Plane Array; FFPE, Formalin-Fixation Paraffin-Embedding; FTIR, Fourier Transform Infrared; HCA, Hierarchical Cluster Analysis; IR, Infrared; MANOVA, Multivarite Analysis of Variance, MCT, Mercury Cadmium Telluride, PCA, Principal Component Analysis; PC, Principal Component; PLS-DA, Partial Least Squares Discriminant Analysis, PLSR, Partial Least Squares Regression, ROC, Receiver Operating Characteristic; S/N, Signal to Noise.

Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/b000000x/

1. A. J. Miller and M. C. Mihm, *N. Engl. J. Med.*, 2006, **355**, 51–65.
2. A. M. Eggermont, A. Spatz, and C. Robert, *Lancet*, 2013, **383**,816-827.
3. C. M. Balch, J. E. Gershenwald, S.-J. Soong, J. F. Thompson, M. B. Atkins, D. R. Byrd, A. C. Buzaid, A. J. Cochran, D. G. Coit, S. Ding, A. M. Eggermont, K. T. Flaherty, P. A. Gimotty, J. M. Kirkwood, K. M. McMasters, M. C. Mihm, D. L. Morton, M. I. Ross, A. J. Sober, and V. K. Sondak, *J. Clin. Oncol.*, 2009, **27**, 6199–6206.
4. F. L. Greene and L. H. Sobin, *CA. Cancer J. Clin.*, **58**, 180–190.
5. V. Gray-Schopfer, C. Wellbrock, and R. Marais, *Nature*, 2007, **445**, 851–857.
6. C. L. Chaffer and R. A. Weinberg, *Science*, 2011, **331**, 1559–1564.
7. D. Hanahan and R. A. Weinberg, *Cell*, 2011, **144**, 646–674.
8. R. Bauer, P. J. Wild, S. Meyer, F. Bataille, A. Pauer, M. Klinkhammer-Schalke, F. Hofstaedter, and A. K. Bosserhoff, *J. Clin. Pathol.*, 2006, **59**, 699–705.
9. G. Bellisola and C. Sorio, *Am. J. Cancer Res.*, 2012, **2**, 1–21.
10. R. Bhargava, *Anal. Bioanal. Chem.*, 2007, **389**, 1155–69.
11. P. Lasch, M. Diem, W. Hänsch, and D. Naumann, *J. Chemom.*, 2007, **20**, 209–220.
12. P. Lasch, L. Chiriboga, H. Yee, and M. Diem, *Technol. Cancer Res. Treat.*, 2002, **1**, 1–7.
13. E. Goormaghtigh, J.-M. Ruysschaert, and V. Raussens, *Biophys. J.*, 2006, **90**, 2946–2957.
14. B. Bird, M. S. Miljković, S. Remiszewski, A. Akalin, M. Kon, and M. Diem, *Lab. Invest.*, 2012, **92**, 1358–1373.
15. Z. Hammody, R. K. Sahu, S. Mordechai, E. Cagnano, and S. Argov, *ScientificWorldJournal.*, 2005, **5**, 173–182.
16. B. Rigas, S. Morgello, I. S. Goldman, and P. T. Wong, *Proc. Natl. Acad. Sci. U. S. A.*, 1990, **87**, 8140–8144.
17. 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**, 1044–1056.
18. D. C. Fernandez, R. Bhargava, S. M. Hewitt, and I. W. Levin, *Nat. Biotechnol.*, 2005, **23**, 469–474.
19. C. M. Balch, S.-J. Soong, M. B. Atkins, A. C. Buzaid, N. Cascinelli, D. G. Coit, I. D. Fleming, J. E. Gershenwald, A. Houghton, J. M. Kirkwood, K. M. McMasters, M. F. Mihm, D. L. Morton, D. S. Reintgen, M. I. Ross, A. Sober, J. A. Thompson, and J. F. Thompson, *CA. Cancer J. Clin.*, 2004, **54**, 131–149.

20. E. Goormaghtigh and J. . Ruysschaert, *Spectrochim. Acta*, 1994, **50**, 2137–2144.

21. E. Goormaghtigh, *Adv. Biomed. Spectrosc.*, 2009, **2**, 104–128.

22. J. H. J. Ward, *J. Am. Stat. Assoc.*, 1963, **58**, 236–244.

23. P. A. Gimotty, P. Van Belle, D. E. Elder, T. Murry, K. T. Montone, X. Xu, S. Hotz, S. Raines, M. E. Ming, P. Wahl, and D. Guerry, *J. Clin. Oncol.*, 2005, **23**, 8048–8056.

24. M. Lomuto, P. Calabrese, and A. Giuliani, *J. Eur. Acad. Dermatol. Venereol.*, 2004, **18**, 291–300.

25. S. O. Frahm, C. Schubert, R. Parwaresch, and P. Rudolph, *Hum. Pathol.*, 2001, **32**, 1376–1381.

26. D. L. Morton, J. F. Thompson, A. J. Cochran, N. Mozzillo, R. Elashoff, R. Essner, O. E. Nieweg, D. F. Roses, H. J. Hoekstra, C. P. Karakousis, D. S. Reintgen, B. J. Coventry, E. C. Glass, and H.-J. Wang, *N. Engl. J. Med.*, 2006, **355**, 1307–1317.

27. C. Voit, A. C. J. Van Akkooi, G. Schäfer-Hesterberg, A. Schoengen, K. Kowalczyk, J. C. Roewert, W. Sterry, and A. M. M. Eggermont, *J. Clin. Oncol.*, 2010, **28**, 847–852.

28. A. Brannon, E. Vakiani, B. E. Sylvester, S. N. Scott, G. McDermott, R. H. Shah, K. Kania, A. Viale, D. M. Oschwald, V. Vacic, A.-K. Emde, A. Cercek, R. Yaeger, N. E. Kemeny, L. B. Saltz, J. Shia, M. I. D Angelica, M. R. Weiser, D. B. Solit, and M. F. Berger, *Genome Biol.*, 2014, **15**, 454.

29. A. Derenne, A. Mignolet, and E. Goormaghtigh, *Analyst*, 2013, **138**, 3998–4005.

30. S. Boydston-White, M. Romeo, T. Chernenko, A. Regina, M. Miljković, and M. Diem, *Biochim. Biophys. Acta*, 2006, **1758**, 908–914.

31. A. Gaigneaux, C. Decaestecker, I. Camby, T. Mijatovic, R. Kiss, J. M. Ruysschaert, and E. Goormaghtigh, *Exp. Cell Res.*, 2004, **297**, 294–301.

32. H. Y. Holman, M. C. Martin, E. A. Blakely, K. Bjornstad, and W. R. McKinney, *Biopolymers*, 2000, **57**, 329–335.

33. R. K. Sahu, S. Argov, A. Salman, M. Huleihel, N. Grossman, Z. Hammody, J. Kapelushnik, and S. Mordechai, *Technol. Cancer Res. Treat.*, 2004, **3**, 629–638.

34. J. Boone, R. van Hillegersberg, P. J. van Diest, G. J. A. Offerhaus, I. H. M. B. Rinkes, and F. J. W. Ten Kate, *Virchows Arch.*, 2008, **452**, 507–514.

35. B. Y. Hernandez, H. F. Frierson, C. A. Moskaluk, Y. J. Li, L. Clegg, T. R. Cote, M. E. McCusker, B. F. Hankey, B. K. Edwards, and M. T. Goodman, *Hum. Pathol.*, 2005, **36**, 275–281.

36. J. F. Thompson, R. A. Scolyer, and R. F. Kefford, *Lancet*, 2009, **374**, 362–365.

Histological class	Number of individual spectra	Number of patients
Primary tumor stage I	2939	6
Primary tumor stage II	4894	11
Primary tumor stage III	2285	6
Primary tumor stage IV	1500	3
Metastases	9920	25
Collagen	2675	6
Blood vessels	1871	5
Dermis	1885	6
Erythrocytes	2019	5
Keratinocytes	2227	6
Lymphocytes	1626	7
Necrotic cells	2122	5
Total	**35963**	**91**

TABLE 1: Description of the spectral database collected from the TMA. Table 1 summarizes for each cell type class; the number of spectra manually selected and the number of patients involved in the building of the database. 10 patients were used twice for two different histological classes. 81 patients were used in the study.

		Predicted as					
True		**Melanoma cells**	**Connective tissue**	**Erythrocytes**	**Necrotic cells**	**Keratinocytes**	**Lymphocytes**
	Melanoma cells	**98.3 ± 4.6**	0.0 ± 0.0	0.1 ± 0.2	0.4 ± 0.9	0.0 ± 0.0	1.3 ± 3.4
	Connective tissue	7.7 ± 11.9	**91.5 ± 13.4**	0.4 ± 0.7	0.0 ± 0.0	0.0 ± 0.0	0.5 ± 1.1
	Erythrocytes	0.4 ± 0.5	1.6 ± 3.0	**97.8 ± 2.9**	0.0 ± 0.0	0.0 ± 0.0	0.2 ± 0.4
	Necrotic cells	0.1 ± 0.1	0.3 ± 0.6	0.0 ± 0.0	**99.7 ± 0.8**	0.0 ± 0.0	0.0 ± 0.0
	Keratinocytes	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**100.0 ± 0.0**	0.0 ± 0.0
	Lymphocytes	7.0 ± 7.8	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**93.0 ± 7.8**

TABLE 2: Validation of the cell type recognition model allowing identification of six biological classes: confusion matrix obtained from the average of the six PLS-DA supervised analyses performed on the two combined spectral regions 3000-2800 and 1800-950 cm^{-1} (mean ± standard deviation).

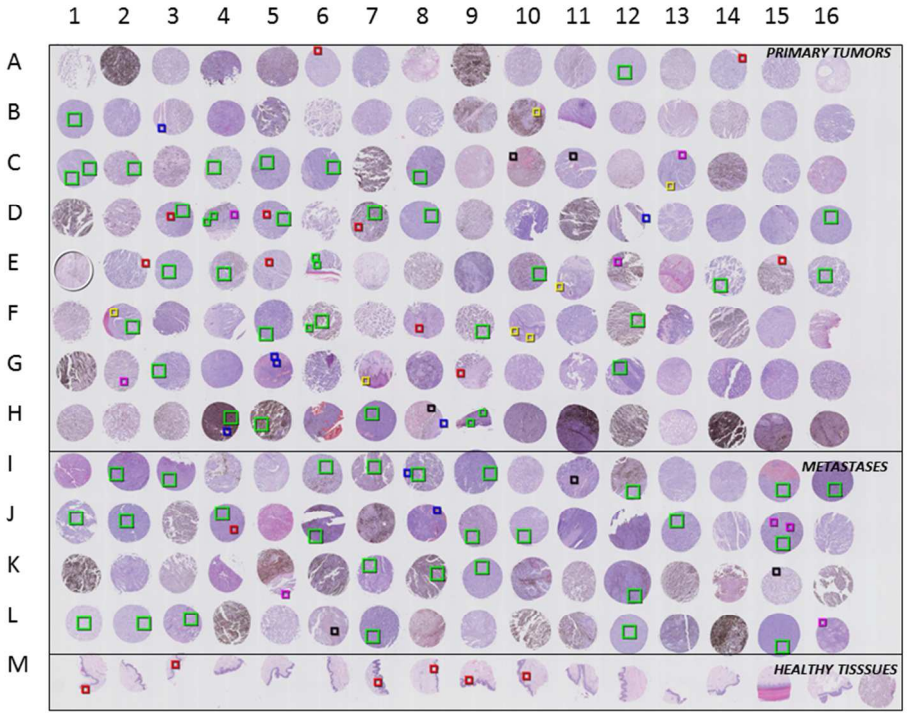

FIGURE 1: Illustration of the H&E stained "Tissue Micro Array" section with alphanumerical code associated with each core. The TMA is composed of 208 cores collected from biopsies from 208 donors. 128 cores are from primary tumors of melanoma patients at different stages of the disease (I, II, III and IV). 64 cores are from regional or distant melanoma metastases and 16 cores are from skin healthy tissues. Areas of infrared imaging measurement are represented by colored squares. In total, 122 measurements have been done, i.e. more than 1 million spectra. Green squares surround melanoma cells; red squares surround blood vessels, dermis and collagen; blue squares surround erythrocytes; yellow squares surround keratinocytes; purple squares delimit the lymphocytes and black squares delimit the necrotic cells.

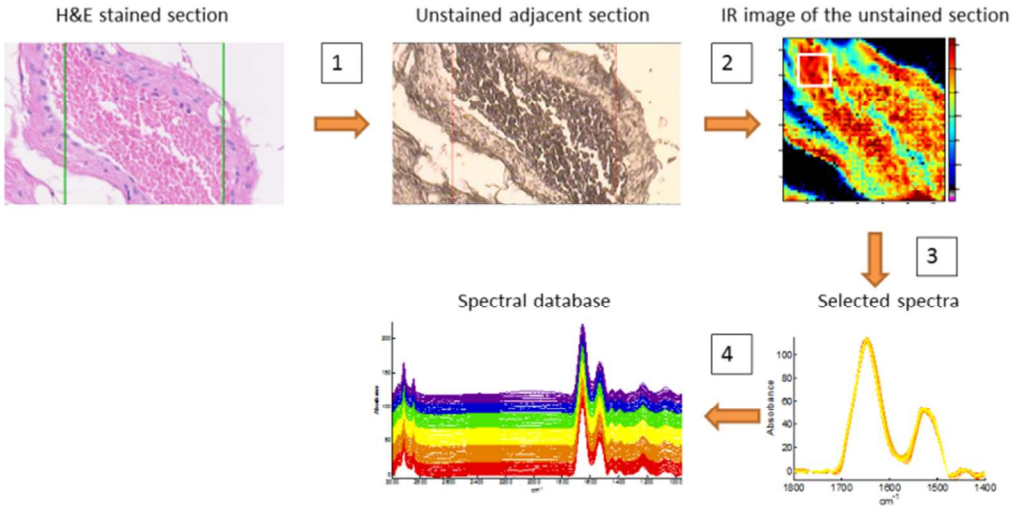

FIGURE 2: Schematic presentation of spectrum selection for building the database. 1. The first step involves the selection of the area of interest on the H&E adjacent stained section with the help of a pathologist and the localization of this equivalent area on the unstained section. 2. The second step consists in the recording of infrared images composed of at least 4096 spectra. 3. The third step is a manual selection of spectra specific to a particular cell type and with a good quality (ratio signal/noise larger than 300). 4. The last step is the compilation of spectra from all cell types or biological structure to establish a spectral database.

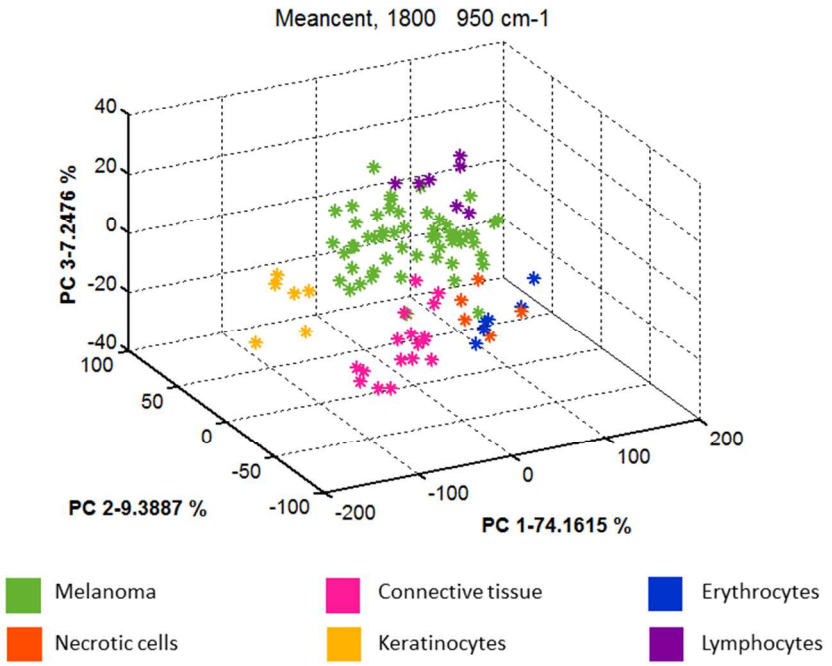

FIGURE 3: PCA score plot of the 91 mean IR spectra for the histological classes described in Table 1 from 81 different patients and representing the six main biological classes (see color code). Connective tissue is a biological class defined here as including dermis, collagen and blood vessel wall. The spectral range used for this analysis is 1800-950 cm^{-1}. Each dot plotted in the PC1-PC2-PC3 space represents a mean spectrum obtained from a different image.

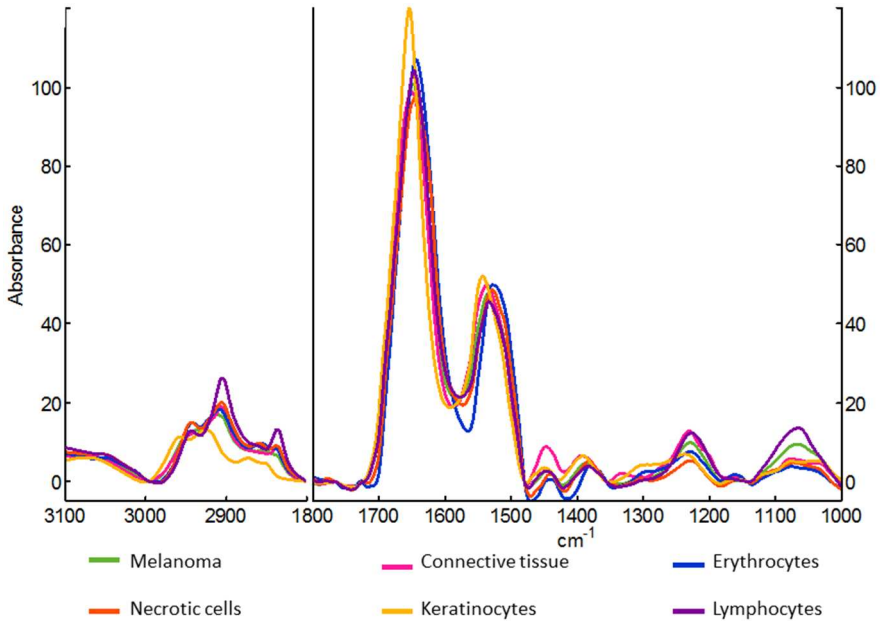

FIGURE 4: Mean IR spectra of the 6 main classes of cell types identified in melanoma tumors in this study.

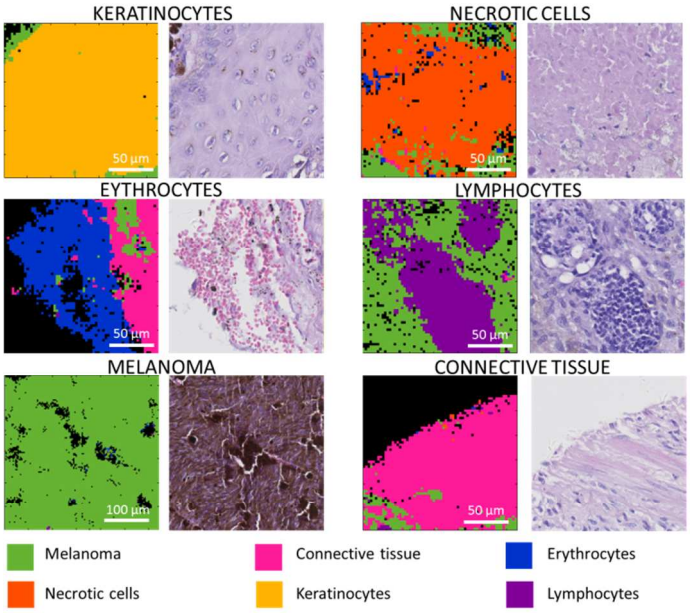

FIGURE 5: Example of application of one of the six PLS-DA models on a series of histological sections containing various cell types and coming from six patients who did not contribute to the training set used to build the model. PLS-DA models were built using two combined spectral regions composed of 3000-2800 and 1800-950 cm^{-1}. For both columns, the left side images are false-color images obtained after application of the PLS-DA model and the right side images report the H&E adjacent stained sections. For images of the left side, each pixel corresponding to a spectrum received a color according to its membership as indicated at the bottom of the figure. Black pixels correspond to spectra with a signal/noise ratio lower than 300. Pictures cover an area of 180x180 μm^2 except for the melanoma image which covers a 360x360 μm^2 area.

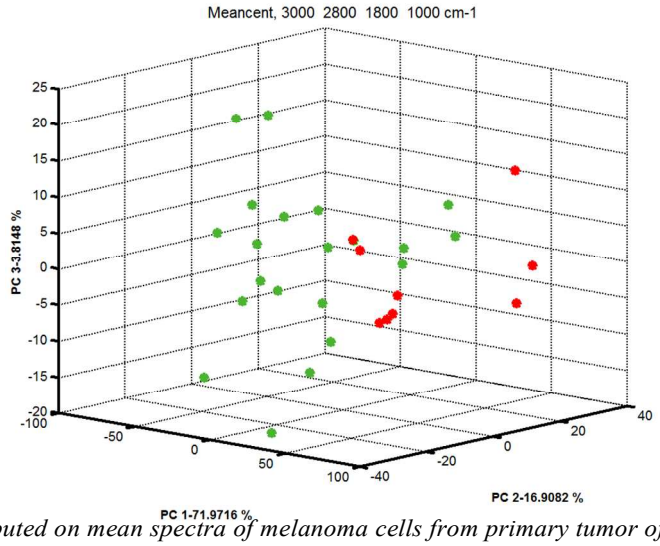

FIGURE 6: PCA score plot computed on mean spectra of melanoma cells from primary tumor of patients at different stages of the disease. The spectral ranges used for this analysis are 3000-2800 and 1800-1000 cm^{-1}. Each mean spectrum is the average of approximatively 15 000 individual spectra. Mean spectra from primary tumor from patients without any metastases (patients from stage I and stage II) are shown in green and mean spectra from patients with evidence for regional or distant metastases (patients with stage III and IV) are shown in red.

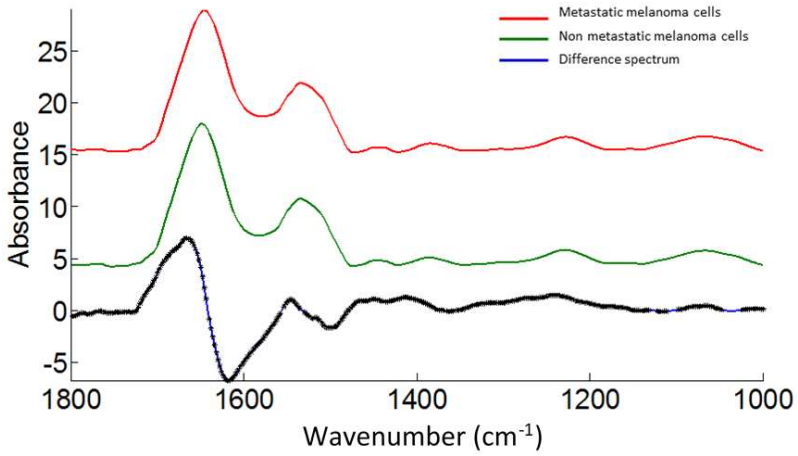

FIGURE 7: Mean spectra of primary tumors of patients diagnosed at stages III or IV (red) and at stages I or II (green). Difference spectrum obtained by subtracting the mean spectrum of primary tumor of patients of stage III and IV from the mean spectrum of patients of stage I and II. This spectrum highlights small difference that could be observed between these two groups. Black stars alongside the spectra point to wavenumber significantly different with α=1%. For the clarity of the figure, difference spectrum was amplified 10 times compared to the mean spectra and spectra are offset along the absorbance axis.

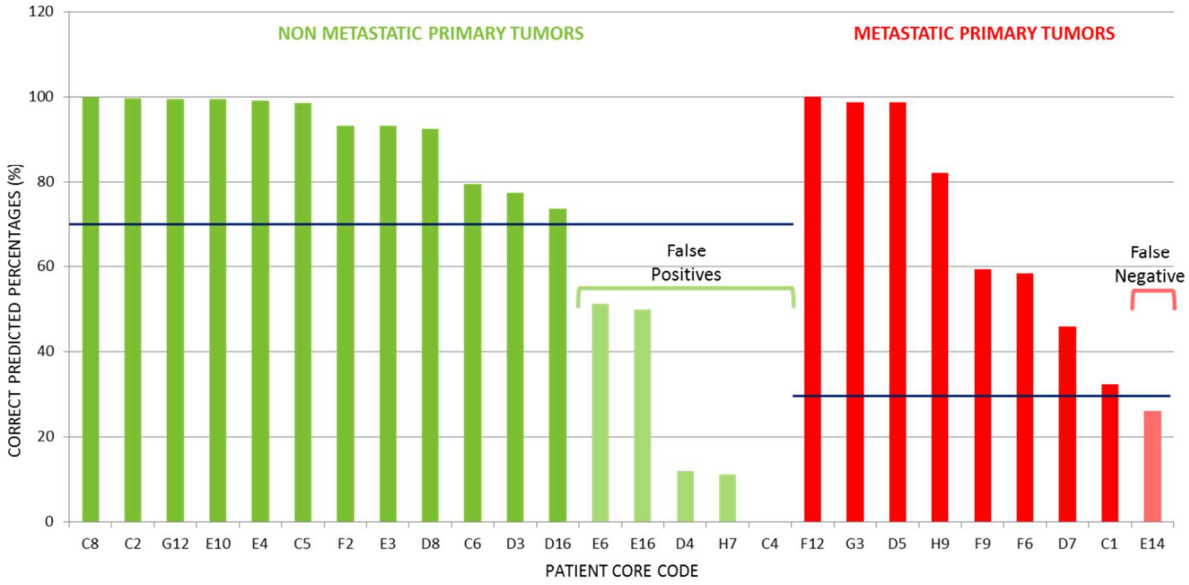

FIGURE 8: Schematic representation of the prediction of PLS DA models allowing a determination of metastatization based on spectral information present in primary tumors. Patient codes are represented in the abscissa axis. For each patient a PLS DA model was trained with the spectra of the 25 other patients (subsets of spectra) and validated on ca. 15 000 spectra from the patient considered. This cross-validation was repeated for the 26 patients. Columns in green are cores from patients at stage I and II (non-metastatic patients) and columns in red are cores from patients at stage III and IV (metastatic patients).The value indicated by the height of each column is the percentage of pixel predicted as non-metastatic (in green) or as metastatic (in red). A threshold of 30 % of metastatic pixels has been selected to reach a maximum of truly detected metastatic patients with an acceptable level of falsely positive. False positives (non-metastatic patients predicted as metastatic) are represented in light green and false negatives (metastatic patients predicted as non-metastatic) are represented in light red.

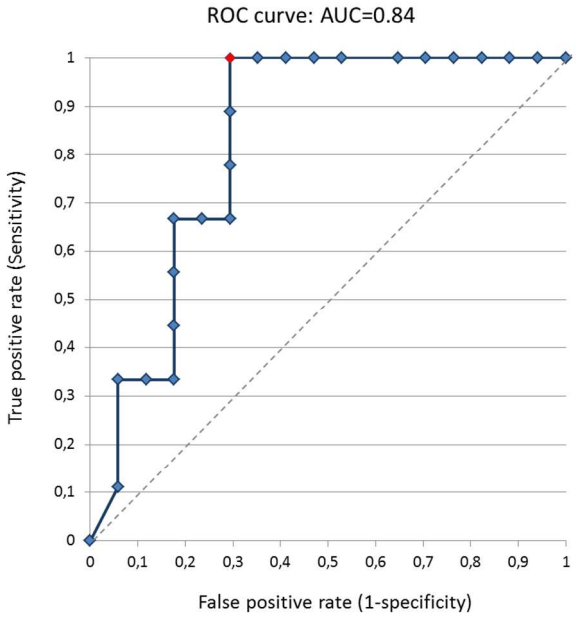

FIGURE 9: ROC curve for the PLS-DA model allowing discrimination between patients with a metastatic primary tumor or with a non-metastatic primary tumor. The red dot on the curve shows the threshold allowing the highest rate of true positives with a low rate of false positives (corresponding to a threshold of 30 % in Figure 8). Area under the curve (AUC) is 0.84.

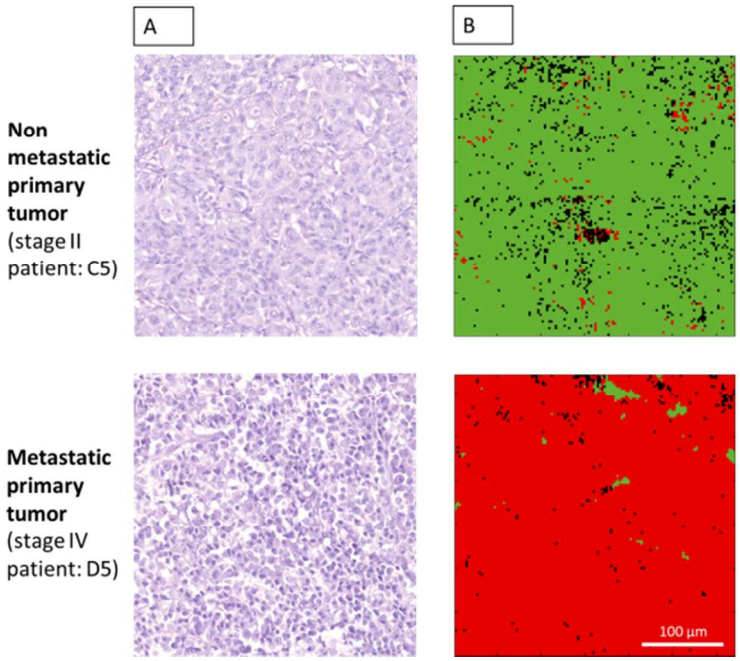

FIGURE 10: Example of application of a PLS-DA model on histological sections of primary tumors from patients graded as stage II and stage IV (B) and the equivalent H&E stained sections (A). This model allows prediction of metastatization of a primary tumor and uses two combined spectral regions: 3000-2800 and 1800-1000 cm^{-1}. Green color corresponds to cells with prediction for non-metastatic tumor and red color corresponds to cells with a potential of metastatization. The PLS-DA model was trained on the 25 patients and applied on the 26th patient for an independent validation. Pictures cover an area of 360x360 μm^2 and contain 16384 spectra. The coordinates C5, D5 refer to the TMA core/patient as indicated in Figure 1.

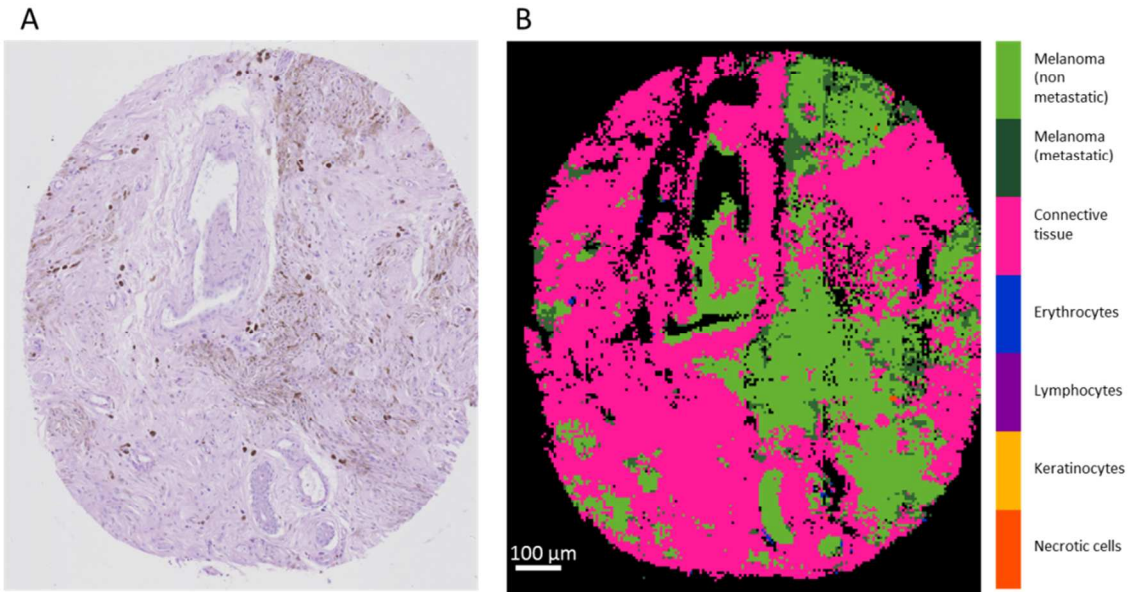

FIGURE 11: H&E stained section of an entire tissue core of the TMA (A) and false-color infrared image obtained by the application of two successive PLS DA models (B). A first model was applied to predict the membership to the main six biologicals classes (melanoma, connective tissue, erythrocytes, lymphocytes, keratinocytes, necrotic cells) and a second model was applied on the spectra assigned to melanoma cells (26.8% of the core area) to predict the potential of metastatization of the primary tumor (light green for non-metastatic and dark green for metastatic). Area of melanoma occupied by light green pixels (non-metastatic) is 82.2% and area occupied by dark green pixels (metastatic) is 17.8%. Black pixels correspond to spectra with a signal/noise ratio lower than 300.

FTIR imaging can identify the main cell types of melanoma tumors and can help identify primary melanomas with the highest risk of metastases.