# Analyst

## Accepted Manuscript

# Analyst

## ARTICLE

# The lipid-reactive oxygen species phenotype of breast cancer. Raman spectroscopy and mapping, PCA and PLSDA for invasive ductal carcinoma and invasive lobular carcinoma. Molecular tumorigenic mechanisms beyond Warburg effect

Jakub Surmacki,[a] Beata Brozek-Pluska,[a] Radzislaw Kordek[b] and Halina Abramczyk[a,*]

Vibrational signatures of human breast tissue (invasive ductal carcinoma and invasive lobular carcinoma) were used to identify, characterize and discriminate structures in the normal (noncancerous) and cancerous tissues by confocal Raman imaging, Raman spectroscopy and IR spectroscopy. The most important differences between the normal and cancerous tissues were found in regions characteristic for vibrations of carotenoids, fatty acids, proteins, and interfacial water. Particular attention was paid to the role played by unsaturated fatty acids and their derivatives. K-means clustering and basis analysis followed by PCA and PLSDA is employed to analyze Raman spectroscopic maps of human breast tissue and statistical analysis of the samples (82 patients, 164 samples). Raman maps successfully identify regions of carotenoids, fatty acids, and proteins. The intensities, frequencies and profiles of the average Raman spectra differentiate the biochemical composition of normal and cancerous tissues. The paper demonstrates that Raman imaging has reached a clinically relevant level in regard to breast cancer diagnosis applications. The sensitivity and specificity obtained directly from PLSLD and cross validation are equal to 90.5% and 84.8% for the calibration and 84.7% and 71.9% for the cross-validation respectively.

## Introduction

Substantial progress in understanding the molecular events that may contribute to cancer development made clear that many signaling pathways and metabolic alterations are responsible for tumorigenesis. Recent papers suggest that phenotypic heterogeneity in the majority of human cancers represents both genetic and nongenetic inputs.[1]

The complexity of cancer cell genotypes provokes a question if searching for individual signaling molecules is useful, because thousands of mutations, translocations, amplifications have been monitored even among histopathologically identical tumors. However, some of these channels are absolutely required for tumorigenic transformation: a) rapid ATP generation to maintain energy status; b) increased biosynthesis of macromolecules; and c) maintenance of appropriate cellular redox status.[2]

The link between cancer and the altered metabolism is usually described through the channel a) via glycolysis (glycolytic phenotype), known as the Warburg effect which replaces ATP generation through oxidative phosphorylation by ATP generation through glycolysis, even under normal oxygen concentrations.[3] There is more and more evidence[2,4-9] that metabolic alterations in tumors extends beyond the Warburg effect and the pathways b) and c) are equally important, particularly those that are crucial for macromolecular building blocks such as reduced nicotinamide adenine dinucleotide phosphate (NADPH), which is also an antioxidant forming the defense against reactive oxygen species (ROS). These paths are also associated with the production of fatty acids and are activated via multiple lipogenic enzymes affected at all levels of regulation, including transcription, translation, protein stabilization and protein phosphorylation.[10,11]

Raman and IR imaging has brought revolution in cancer detection and treatment. These methods are ideally suited to explore the cancer phenotype by monitoring the biochemistry of molecules that are necessary for survival, proliferation, differentiation, cell death, and expression of many cell-type-specific functions.[4-6,8,9,12]

Raman imaging has capability to detect even a single cancerous cell in vivo and monitor distribution of the highly toxic drugs delivered to the cancerous cells. Raman spectroscopy and imaging can identify components of extracellular matrix that

forms the milieu in which tumor cells proliferate as well as monitor the mesenchymal cells after transformation from the cancer epithelial cells. Therefore, Raman imaging, combined with other optical methods, is a powerful method of studying discrete biochemical events that lead to malignant transformation and will made the detection of genetic abnormalities in cancer cells more precise and complete.

To address these important questions on phenotypic alterations in cancer cells we have studied human breast tissue of normal and cancerous tissue (invasive ductal cancer (IDC) and invasive lobular cancer (ILC)). IDC and ILC are cancers that began growing in the duct (IDC) or the lobule (ILC) and have invaded the extracellular matrix, dominated by fatty tissue of the breast outside of the duct or the lobule. IDC and ILC are the most common forms of breast cancer, representing over 80 percent of all breast cancer diagnoses.

Several groups have explored the potential of Raman and IR spectroscopy for breast cancer detection.[4-6,13-18]

The goal of the paper is to answer how we can use the vibrational properties of biomolecules in cell and tissue for diagnosis of cancer. We will show that specific fatty acids/triglycerides/proteins/carotenoids vibrational bands are significantly altered when comparing normal and cancer tissue. Although there are many clinically significant molecular markers of breast cancer, no tumor marker now available has met requirements of perfect reproducibility, correlation with stage, grade, and response to treatment. Raman imaging and spectroscopy gives a new promise to provide a perfect tool against cancer.

We propose Raman biomarkers for breast tissue cancer diagnosis that would have the ability to accurately characterize cancer tissue and distinguish between normal, malignant and benign types. We will show that the Raman features of the lipid profile as well as carotenoids that play diverse functions including antioxidant properties to defense against ROS may be very useful for biodiagnostics of cancer. The lipid phenotype was ignored for many years by the Raman scientific community that has focused almost exclusively on the biosynthesis of proteins and the proteome profile.[14,15,19]

We demonstrate that the Raman markers based on lipids and antiROS species, such as carotenoids, are easily and reproducibly measured, and exhibit high specificity and sensitivity. We have employed statistical methods of Principal Component Analysis (PCA) and Partial Least Squares Discriminant Analysis (PLSDA) for the samples of the normal human breast tissue from the safety margin of tumor mass and the tissue from the tumor mass from 82 patients and 164 samples. To the best of our knowledge, such a large collection of human breast tissue samples was not published in any paper.

The second aim is to present the results that will be useful in practical applications for the 'optical biopsy' by Raman imaging to evaluate the area of the tumor mass and to insure negative safety margin around the tumor mass to avoid reexcision. There is an urgent need to improve the conventional methods as it has in fact been found that 70-90% of mammographically detected lesions are found to be benign upon needle biopsy.[20]

## Experimental

### Patients, samples and tissue preparation

All studies and procedures involving human tissue were approved by a protocol approved by the institutional Bioethical Committee at the Medical University of Lodz, Poland (RNN/45/14/KE/11/03/2014). Tissue samples were collected from freshly excised surgical specimens. We used fresh bulk tissue samples and cryosectioned slices from the tumor mass and the tissue from the safety margin outside of the tumor mass obtained during the breast surgery. The histological analysis was performed by professional medical doctors, board certified as pathologists, from the Medical University of Lodz, Department of Pathology, Chair of Oncology according to the standard histology protocols. The histopathology results served as the gold standard against which the Raman spectral diagnoses were compared.

Total number of patients was 82. Total number of samples was 164. For each patient the breast tissue from the safety margin and the tumor mass were measured and analyzed. The number of patients (82) is not equal to the number of histologically identified pathologies, because for 18 patients more than one pathology was identified giving total value of 100.

An essential role in spectroscopic methods is played by preprocessing of tissues, as they naturally occur at the expense of introducing some fixatives, matrices, adhesives, which may generate own vibrational spectra overlapping the signals from the native tissue. To obtain reliable results one must be sure that the method of tissue processing does not distort the vibrational spectra of the studied tissue. Detailed methodology is available elsewhere.[5]

Briefly, the bulk tissue samples required no chemical processing. To obtain slices the bulk tissue was cryosectioned at -25 °C with a microtome (Microm HM 550, Sermed) into 16 μm thick sections for histopathological and Raman analysis. The thin sections were put on microscopic glasses from $BaF_2$ substrate and stained with hematoxylin and eosin (H&E) to provide histopathological diagnosis of the suspected areas. The adjacent sections without staining have been examined by Raman spectroscopy and Raman imaging. After Raman measurements the sections were stained and histologically examined to check if the result is identical to that obtained for the adjacent sections. The samples requiring long collection times (10 s) were recorded in a home designed cuvette to maintain a proper hydration. We did not use the paraffin-embedded sections nor formaldehyde fixation for Raman measurements because: 1) wax has its own vibrational spectra of lipids that overlap those from the breast tissue, 2) the standard procedures of deparaffinization with various agents (i.e. xylene) and rehydration with aqueous solutions of ethanol lead to washing out fatty acids and artificial hydration of the tissue, which are important indicators of pathology,[21] 3) tissue

fixation may cause significant Raman spectral changes in the amino acid fingerprint region of proteins, because formaldehyde promotes the cross-linkage of amino groups in collagen.[22,23]

**Raman spectroscopy and imaging**

To avoid intratumor heterogeneity the average spectra reported in this study were acquired using a Raman spectrometer Ramanor U1000 (Jobin Yvon, JY) with double gratings of 1200 lines/mm excited with the ion Ar laser (514 nm) at 25 mW or 200 mW power focused on the sample with laser spot of 500 μm at the collection (integration) time of 0.5 s for bulk tissue and 10 s for thin slices of 16 μm, at 2 cm$^{-1}$ spectral spacing in the spectral range of 200-3600 cm$^{-1}$. Raman spectra were collected using photomultiplier (Product For Research/NC model C31034).

To explore morphological and biochemical heterogeneity within tumors Raman images were acquired using a confocal Raman microscope - WITec alpha 300 RSA (Ulm, Germany) consisting of an Olympus microscope, coupled with a 300 mm Czerny-Turner monochromator (Princeton Instruments Acton SP23000-300 mm Imaging Triple Grating Monochromator/Spectrograph) and a thermoelectrically cooled CCD Camera ANDOR Newton DU970N-UVB-353 (EMCCD chip with 1600x200 pixel format, 16 μm dimension each) camera operating in standard mode at -64°C with full vertical binning. The laser beam doubled SHG of the Nd:YAG laser (532 nm) is focused on the sample with an 40x magnification objective (NIKON CFI Plan Fluor C ELWD 40x: NA 0.60, WD 3.6- 2.8 mm; DIC-M, C.C.0-2) to the laser spot of 1 μm defined by the laser wavelength and microscope objective being used. The average laser excitation power was 10 mW, the collection (integration) time of 0.5 s with 2 cm$^{-1}$ spectral resolution in the spectral range of 200-3600 cm$^{-1}$. A piezoelectric table was used to record Raman images. Spectra were collected at one acquisition per pixel and a 1200 lines/mm diffraction grating. Raman images (80x80 μm, 60x60 points per line) from the spectral region 2850-2950 cm$^{-1}$ of the breast tissue from the tumor mass were constructed.

**Infrared spectroscopy**

IR spectra were recorded using Specord M 80, Germany. The spectra were scanned with the step of 4 cm$^{-1}$. The controlled constant humidity was maintained in a specially designed homemade cell (windows of CaF$_2$ or BaF$_2$ substrate) with a tip containing a saturated salt aqueous solution in contact with the layer of the tissue within an enclosed space. P$_2$O$_5$, CH$_3$COOK, NaCl, aqueous solutions and pure water that maintain 0%, 23%, 75%, 100% humidity, respectively.

**Data pre-processing and multivariate data analysis**

Raman average spectra collected with the photomultiplier detector were preprocessed using the Jobin Yvon LabSpec Beta 2 Version package. Raman data collected with CCD detector were preprocessed using the WITec Control/Project Plus 1.6/Project 2.10 package. The raw Raman data were corrected for cosmic rays. We used spatial and frequency filtering to remove sharp spikes attributed to cosmic rays.[24,25]

For 532 nm laser excitation and 1200 lines/mm grating, the spectral bandpass varies from about 5.5 cm$^{-1}$ per pixel at about 200 cm$^{-1}$ to about 3.3 cm$^{-1}$ per pixel at 3600 cm$^{-1}$.

The cosmic rays corrected data were smoothed by a Savitzky and Golay procedure.[26]

The raw Raman data are analyzed in the range of 200-1800 cm$^{-1}$ which is useful for control purpose, but contain Rayleigh scattering and the Raman spectrum of the CaF$_2$ or BaF$_2$ support were removed by reduction of the spectral range to 400-1800 cm$^{-1}$ before multivariate analysis. The linear backgrounds were also subtracted before further analysis.

For IR data we did not use correction for Mie scattering.[27]

The average Raman spectra were analyzed using the PCA method and MATHLAB least-squares fitting algorithm using PLS_Toolbox Version 4.0 for use with MATLAB.[28]

PCA is an unsupervised learning algorithm; which means that no previous knowledge of the samples under study is provided, used for spectral multivariate analysis of multidimensional datasets. PCA finds an alternative set of coordinates, the new orthogonal basis set, to reduce the dimensionality and complexity of the data set.[29] PCA is based on a linear transformation of spectra regarded as N dimensional vectors, where N is the number of data points in a single spectrum, in a new coordinate system in which the new vectors are orthogonal to each other. All the spectra can then be expressed in a much simpler manner through a small number of principal components (PCs) that accounts for the maximum variance in the data. In a PCA model, the matrix containing the set of spectra (X) is decomposed into two smaller matrices (the scores (T) and the loadings (P)): $X=TP^T+E$ where E is the residual containing information that is not described by the multiplication of the scores and the loadings for each spectrum. By plotting the Principal Components scores, similarities between samples (forming clusters by grouping into the same area the spectra with identical molecular properties) are revealed. In addition, plotting loadings as a function of the wavenumbers reveals the most important diagnostic vibrations in the spectra.[7,30]

The visualization of chemical similarities and differences was also demonstrated in the human breast tissue images. The 2D array of tens of thousands of individual Raman spectra recorded from the human tissue from the safety margin and the tumor mass have been used to construct Raman images using the mapping function of the WITec Control/Project Plus 1.6/Project 2.10 package. K-means clustering analysis (KMCA) and basis analysis with various metrics (Euclidan and Manhattan) have been employed to analyze Raman spectroscopic maps of human breast tissue. KMCA is a method of analysis which clusters n spectra of a dataset into $k$ clusters such that each spectrum belongs to the cluster with the nearest (closest) mean.[31] The aim of KMCA is to minimize the sum of distances between spectral vectors $S_i^J$ and cluster centroids $m_k$:

$$\sum^n (S_i^j - m_k) \tag{1}$$

where $J$ spectral vectors originally are randomly assigned to belong to a given cluster $k$ with centroid $m_k$.

The basis analysis is based on the average spectra from characteristic areas of the sample. In this method each measured spectrum of the 2D spectral array is compared to the average spectra $S_A$, $S_B$, $S_C$ using a least square to fit reach convergence to minimize the fitting error $D$ described by the equation (2)

$$D = \left( [\overrightarrow{RecordedSpectrum}] - a \times \overrightarrow{BS_A} - b \times \overrightarrow{BS_B} - \quad c \times \overrightarrow{BS_C} - \cdots \right)^2 \tag{2}$$

by varying the weighting factors $a$, $b$, $c$,... of the basis spectra $\overrightarrow{BS}$.[30] We used the average basis spectra from four characteristic regions. These values of the weighting factors are converted to a monochrome intensity map. The color code of Raman maps was based on the integrated Raman intensities in specific regions (sum option in the filter manager in the WITec Project Plus 1.60). Using a lookup table, bright yellow colors indicate the highest intensities, whereas brown colors indicate the lowest intensities of the chosen region.

Up to 9 monochrome maps for different weighting factors can be combined to yield a pseudo-color map with mixed colors indicating the intensity values of each of the weighting factors. The color maps can then be constructed to visualize the distribution of the weight factor magnitudes that provide information about biochemistry of the molecules in terms of their vibrational frequencies and intensities. We used both Euclidean and Manhattan metrics to compare the Raman images for various definition of distances between spectral vectors $S_i^J$ (and cluster centroids $m_k$) to compare the quality of the images.

The large number of peaks in the Raman spectra that are all potential cancer biomarkers create modelling and validation challenges. The PCA[32] and PLSDA[33] score plots were used to separate between the classes. The loadings of the principal components identified by PCA and PLSDA shed further light on the spectral basis for differentiation. We used permutation testing and cross model validation to assess the validation of classification models.

To assess the specificity and sensitivity we used two methods. First, we have chosen carotenoids (peaks at 1158 and 1520 cm$^{-1}$) and lipid profile (2854, 2888, 2931 cm$^{-1}$) as the dominant Raman biomarkers, because they give the dominant contribution to the principal components PC1 and PC2. Using the definitions of the sensitivity as the proportion of actual positives which are correctly identified as such (the percentage of sick people who are correctly identified as having the condition), we calculated the ratio of the samples with a positive result (in our case lack (not presence) of the characteristic Raman peaks in tissue) to the total number of the samples that are identified as tumorous tissue according to the histopathological assessment. Specificity (the true negative rate), which is defined as the proportion of negatives which are correctly identified as such (the percentage of healthy people who are correctly identified as not having the condition), has been calculated as the ratio of the samples with a negative

result (presence (not lack) of the characteristic Raman peaks in tissue) to the total number of the samples that are identified as normal tissue according to the histopathological assessment. The classification into two groups (normal tissue, tumor tissue) was based on the histological assessments.

The second method was based on PLSDA used for the classification and cross validation as well as permutation testing were used to validate the classification models.

## Results and discussion

As cancer is caused by a mutation in specific genes of the DNA molecules most diagnostic tools define cancer subtypes, recurrence of disease and response to specific therapies using protein based and DNA - based gene expression signatures in commercial microarrays. Therefore various methods, including IR and Raman spectroscopy, concentrate on proteins or DNA. Snapshot of the current state of the art on cancer research area shows that the Raman and IR spectral region in the fingerprint region of the biological systems are usually interpreted[14,15,19] in terms of the protein's secondary structure: $\alpha$ helical, $\beta$-sheet, mixed structures $\alpha/\beta$, $\alpha+\beta$.[34,35] Although substantial biochemical information on proteins is imprinted in the IR and Raman spectra, such as the Amide I at around 1660 cm$^{-1}$ in the biological tissue, it is obvious that the peak in the near vicinity at around 1655 cm$^{-1}$ represents lipids such as fatty acids and triglycerides,[4-6,13,36] the peak at around 1440 cm$^{-1}$ represents $CH_2$ or $CH_3$ deformations of fatty acids, triglycerides[36] and overlaps with bands at 1458 cm$^{-1}$ and 1461 cm$^{-1}$ corresponding to the $CH_2$ or $CH_3$ deformations of proteins[19,37] and phospholipids $CH_2$ scissoring.[38] The almost exclusive proteome interpretation of biological tissue in literature is partially related to the fact that most research use histological sections or commercially available paraffin-embedded tissues instead of using fresh tissue as presented in this paper. The standard procedures of deparaffinization with various agents (i.e. xylene) and rehydration with aqueous solutions of ethanol lead to washing out fatty acids and artificial hydration of the tissue, which are important indicators of pathology.

To avoid the proteome pitfalls and to obtain biochemical information in realistically crowded environment of the biological tissue from Raman spectra it is necessary to analyze signals from all molecules present in cells: a) proteins with the full range of amino acids, b) fatty acids and lipids, c) DNA and RNA, d) saccharides, e) primary metabolites (succinic acid, fumarate, malic acid, pyruvate, phosphoenolpyruvate, coenzyme A, acetyl coenzyme A, acetoacetate, D-fructose-6-phosphate), f) and other products that are of importance for living organisms such as carotenoids, ascorbic acid, riboflavin, glutathione[4-6,36] associated with the biological environment.

Recently we have shown that tumor is strongly heterogeneous where the distribution of various components can be easily observed by a confocal Raman microspectroscopy.[5] In the confocal microspectroscopy the laser beam is focused by an microscope objective to a very small spot offering very high spatial resolution at large NA (for a 532 nm laser with a

0.90/100x objective a diffraction limited spatial resolution is ~360 nm). We will explore the benefits of high spatial resolution later when we will discuss some molecular signatures of cancer.

On the other hand, there is increasing interest in using whole slide imaging for diagnostic purposes[39] at the expense of resolution. To avoid obstacles related to tumor heterogeneity we recorded the "average" Raman spectra at the laser spot of around 500 µm from large area of the biological sample.

Fig. 1 shows the average Raman and IR spectra of the vibration of the human breast tissue from the safety margin and from the tumor mass.

One can see from Figs 1A,1B that the Raman spectra provide many details on the vibrational features of tissues. To obtain complementary vibrational information we have recorded IR spectra of the same sample. The results are shown in Fig. 1C.

**Fig. 1** Raman average spectra of the human breast tissue from the safety margin (A) and tumor mass (B), integration time of 10 s, laser power of 200 mW, slice of 16 µm on $BaF_2$ window. IR spectra (C) of the human breast tissue from the safety margin (a) and tumor mass (b), integration time of 5.0 s, slice of 6 µm on $BaF_2$ window, controlled humidity 23%. Invasive ductal cancer G2 (patient P95).

Fig. 1C shows the typical IR spectra for the cancerous breast tissue of the same patient (P95) shown in Figs 1A,1B. The results show that the absorbance of various vibrations in the normal tissue differs markedly compared to the cancerous tissue, particularly in the region of Amide I (1660 $cm^{-1}$) and phospholipids (1098 $cm^{-1}$).

The results presented in Fig 1 show that the Raman and IR spectra contain multiple contributions from proteins, lipids, and nucleic acids, with contributions from individual RNA and DNA bases and the sugar-phosphate backbone of DNA, and other products that are of importance for living organisms such as carotenoids.

Detailed inspection into Fig. 1 shows an unambiguous discrimination between tissue from the tumor mass and the safety margin, which in this case represents the normal breast tissue with the negative safety margin identified by the histopathologists where cancer cells do not extend out to the edge of the tumor mass.

One of the most remarkable aspects of comparison are marked differences in the spectral regions characteristic for carotenoids, lipids (fatty acids, triglycerides, phospholipids) and proteins. A detailed inspection into Figs 1A, 1B demonstrates that the normal breast tissue contains a markedly higher concentration of carotenoids (1158, 1520 $cm^{-1}$) than the cancerous tissue from the tumor mass. Moreover, lipids in the safety margin have different profiles of vibrational features in the high frequency region 2800-3000 $cm^{-1}$ indicating distinct lipids composition compared to the cancerous tissue.

The Raman spectrum of the normal tissue is dominated by the peaks at around 745, 877, 1004, 1158, 1259, 1304, 1440, 1520, 1654, 1742, 2854, 2888, 2902, 2931, 3009 $cm^{-1}$, while the Raman spectrum of the cancerous tissue from the tumor mass

shows characteristic peaks at around 729, 880, 958, 997, 1089, 1165, 1440, 1458, 1660, 2888, 2931, 3067, 3311 $cm^{-1}$.

The IR spectrum of the normal tissue is dominated by the peaks at around 1098, 1250, 1351, 1374, 1455, 1591, 1742, 2854, 2921, 3009, 3320, 3450 $cm^{-1}$, while the IR spectrum of the cancerous tissue from the tumor mass shows characteristic peaks at around 1075, 1163, 1261, 1383, 1505, 1519, 1539, 1558, 1593, 1652, 1742, 2854, 2921, 3067, 3320, 3450 $cm^{-1}$.

Table 1 lists tentative assignments for the observed Raman (R) and IR bands of the human breast tissue.

A detailed inspection of Fig. 1 revealed that the vibrational spectra of the normal breast tissue are characteristic of adipose-rich and phospholipid-rich regions (Table 1). In contrast, the vibrational spectra of the cancerous breast tissue are more characteristic of protein regions (Table 1).

Table 1. Tentative assignments of the vibrational bands of the human breast tissue from the Raman (R) and IR spectra (patient P95)[5,6,19,36-38,40-46]

The Raman peaks at 1004, 1158, and 1520 $cm^{-1}$ of the normal breast tissue correspond to vibrations of the C-C coupled with $C-CH_3$ and C=C stretching modes of carotenoids, the peak at 1259 $cm^{-1}$ corresponds to the antisymmetric stretching mode of phospholipids. The peak at 1004 $cm^{-1}$ is usually attributed to phenylalanine,[34,35] but in such a characteristic triple combination we have assigned it to carotenoids like the bands at 1158, and 1520 $cm^{-1}$. Other apparent vibrations characteristic for the C-H groups of lipids (fatty acids, triglycerides and others) are: 1440 ($CH_2$ or $CH_3$ deformations, scissoring modes), 1654 (C=C stretching mode), 1742 (C=O stretching), 2854 (symmetric stretching mode), 2888 ($C-H_2$ antisymmetric stretching mode), 2931 $cm^{-1}$ ($CH_3$ symmetric stretching mode), peak at 3009 $cm^{-1}$ corresponds to (C=C)-C-H antisymmetric stretching mode.

The cancerous breast tissue has much more evident proteome profile although the lipid contribution similar to that in the normal tissue is also observed in the vibrational pattern. The cancerous breast tissue shows strong peaks at 729, 880, 958, 997, 1075, 1089, 1165, 1261, 1392, 1440, 1448, 1505, 1519, 1539, 1558, 1593, 1660, 1742, 2854, 2888, 2931, 3067 $cm^{-1}$. The spectra of proteins in the cancerous tissue are dominated by peaks corresponding to Amide I (C=O stretching, α helix) (1660 $cm^{-1}$), Amide III (coupled C-N stretching and N-H bending, α helix) (1261 $cm^{-1}$), Amide II (coupled C-N stretching and N-H bending) (1505-1593 $cm^{-1}$).[34] The peak at 1080 $cm^{-1}$ corresponds to phospholipids. The vibrations at 2854 $cm^{-1}$ are typical for the CH group of lipids, while the vibrations at 2931 $cm^{-1}$ represents a combination of the vibrations of lipids, amino acids and collagen.[16-18] The peak at 3067 $cm^{-1}$ corresponds to (C-H) - aromatic vibrations observed in aromatic amino acids (phenylalanine, tryptophan).

The low frequency peaks in the range 700-1000 $cm^{-1}$ correspond to $C-C-N^+$ symmetric stretching in phosphatidylcholine, a major component of cellular

membranes. The other peaks correspond to carbohydrates with the characteristic peaks of sugars and DNA/RNA.[36]

In the view of the results presented so far one can claim that the proteome interpretation of cancer in the human tissue is not fully justified because it does not take into account the effect of extracellular matrix, which is particularly important for invasive ductal and lobular cancers, when the epithelial cells migrate and invade through the basement membrane into the surrounding extracellular matrix consisting of connective tissue, adipose tissue and fibroblasts.[47]

The picture that emerges from our results demonstrates that tumor cells in the invasive IDC and IDL breast cancers have evidently mixed proteome-lipidic-ROS profile. Enhanced carotenoid and lipid content as well as decreased proteome were observed in normal cells compared with the cancerous cells. The normal breast tissue contains much more adipose tissue which is highly hydrophobic. The cancerous breast tissue contains much more proteins, which are usually hydrophilic.

To explore morphological and biochemical heterogeneity within tumors Raman images were acquired using a confocal Raman microscope. We will demonstrate that our interpretation of the average Raman spectra presented in Fig. 1 is fully justified.

The regions marked in Fig. 2C with blue and green crosses represent adipose tissue, the regions with other colors of crosses represent the cancerous tissue consisting of epithelial cells embedded in breast stroma. It is clear that the adipose rich regions show the characteristic peaks of carotenoids at 1161, 1527 $cm^{-1}$, fatty acids/triglycerides at 1440/1448, 1654, 2854, 2888, 2902, 2931, 3009 $cm^{-1}$. In contrast, the cancerous region has the characteristic peaks corresponding to both fatty acids/triglycerides and proteins at 1448/1458, 1662 (Amide I), 2940, 3067 $cm^{-1}$. It is interesting to notice that the frequencies of the characteristic peaks for the average Raman spectra (Fig. 1) and the microRaman spectra (Figs 2D, 2E, 2F, 2G) are shifted (e.g. average carotenoids 1520 vs 1527 $cm^{-1}$), which illustrates the effect of the extracellular matrix.

**Fig. 2** Histological image (A), microscopy stitching image 530x530 μm (B), microscopy image from the region marked with a yellow frame (C) and Raman spectra (D, E, zoomed in on F, G) corresponding to various areas of the breast tissue from the tumor mass (invasive ductal cancer GX, patient P104), colors of the Raman spectra correspond to the positions marked with crosses of the same color.

**Fig. 3** Raman images (80x80 μm, 60x60 points per line) (A, B, C, E, F) from the region marked with a orange frame on microscopy image (D) of the breast tissue from the tumor mass (invasive ductal cancer GX, patient P104). Raman images from the spectral region 1490-1582 $cm^{-1}$ (carotenoids) (A), 1427-1493 $cm^{-1}$ (fatty acids/ triglycerides) (B), 1632-1698 $cm^{-1}$ (fatty acids/ triglycerides) (C), 2850-2950 $cm^{-1}$ (fatty acids/ glycerides) (E), 2940-3000 $cm^{-1}$ (proteins) (F).

The heterogeneous distribution of carotenoids, fatty acids and proteins have been successfully identified in the noncancerous and cancerous breast tissue in Raman images presented on Fig. 3, K-means cluster analysis maps (Fig. 4B, 4C) and basis analysis (Fig. 4D).

**Fig. 4** The color Raman maps constructed from the monochrome maps for the spectral filter 1800-3200 $cm^{-1}$, (A) microscopy image, (B) K-means cluster analysis, 4 clusters, Euclidean metrics, (C) K-means cluster analysis, 4 clusters, Manhattan metrics, (D) basis analysis (4 average spectra), (E) spectra representing different clusters (or basis average spectra). The line colors correspond to the colors of the Raman maps. Blue region corresponds to the BaF$_2$ support at the edge of the tissue sample (patient P104).

To evaluate more formally the predictive validity and robustness of the vibrational spectroscopy in cancer diagnostics, PCA study was applied to the Raman spectra of 82 patients and 164 samples. In order to extract the most essential Raman features from a large amount of samples we performed PCA analysis for the average Raman spectra in the spectral range of 200-3600 $cm^{-1}$ of the bulk human breast tissue samples.

Fig. 5 shows the statistical chart for the analyzed human breast tissue pathologies: IDC (69), ILC (8), other malignant pathologies (5) (mucinous carcinoma (3), metaplastic carcinoma (1), Papillary carcinoma (1)) and benign tumors (18) (benign dysplasia (7), cystic fibrosis (10), adenosis (1)). The total number of samples was 164 from 82 patients with the total number of average Raman spectra 397 for PCA analysis.

**Fig. 5** The statistical chart with the numbers of analyzed human breast tissue pathologies.

PCA analysis leads to the reduction the dimensionality and complexity of the biological data set. We have obtained that the first three PC components accounts for the maximum variance in the data and are capable of explaining the most fundamental features in the average Raman spectra of the human breast tissues.

Fig. 6 shows the PCA score plot (model: SNV, mean center, first derivative) for all the recorded Raman spectra and all the samples. PC1 and PC2 scores accounted for 77.62% and 5.68% respectively of the total variance in the dataset. The PCA score plot shows good separation between the two arbitrary classes and clearly discriminates the samples from the safety margin and from the tumor mass as shown in Fig. 6. The samples in the left circle of the PCA plot are almost exclusively the samples from the breast tumor mass. The tissue samples from the tumor mass were grouped in the low PC2 region and the negative PC1 axis respectively. In contrast, the samples of the tissue from the safety margin appear to be much more heterogeneous, spreading through a larger area both along the PC1 and PC2 axes. In the right circle there are almost exclusively the samples from the safety margin of the tumor mass. The safety margin, also known as "margins of resection" is an area within the distance between a tumor and the edge of the surrounding tissue that is removed during surgery. Usually, the samples from the safety margin represent the normal breast tissue. However, in some cases cancer cells extend out to the edge of the tumor mass. Based on the observations of the pathologist, the margins of a tumor can be classified as: a) positive margins: cancer cells

extend out to the edge of the tissue; b) negative margins: no cancer cells are found; c) close margins: any situation that falls between positive and negative is considered.

In our analysis presented in Fig. 6 the samples where the positive margin has been assessed by the pathologists were labelled in the same way as the samples from the tumor mass.

**Fig. 6** PCA score plot (model: SNV, mean center, 1-st derivative) for the average Raman spectra of the human breast tissue samples from the tumor mass (red triangles) and the safety margin (blue circles), bulk tissue, integration time 0.5 s, laser power 25 mW, step 2 cm$^{-1}$, 200-3600 cm$^{-1}$ (A). PCA loading plot for PC1 (B) and PC2 (C).

Fig. 6A shows also the characteristic spectra corresponding to the various coordinates of the PC1-PC2 score plot. The most prominent bands that determine the coordinates on the PC1-PC2 plot are the bands in the region 2800-3000 cm$^{-1}$ corresponding to lipids and proteins (C-H symmetric and antisymmetric vibrations of the hydrocarbon chain) and the bands at 1158 cm$^{-1}$ and 1520 cm$^{-1}$ of carotenoids (C-C coupled with C-CH$_3$ (1158 cm$^{-1}$) and C=C (1520 cm$^{-1}$)) stretching vibrations of carotenoids. For the high positive PC1 (normal tissue) the Raman intensity of the bands of lipids (fatty acids and triglycerides of adipose tissue) at 2854 cm$^{-1}$ decreases along PC2 from high to low PC2 in contrast to the bands of carotenoids. For the negative PC1 (tissue from the tumor mass) the Raman intensity of the bands of lipids and carotenoids are very low and the PC2 component is close to zero. PC1 and PC2 gives the dominant account for the maximum variance in the data.

In order to simplify interpretation of the experimental observations, this study is aimed at understanding the molecular information contained in the first two PCs. Although the PCA analysis does not provide the answer what the physical meaning of the PC component is, but the loading plot presented in Figs 6B, 6C can provide some hints related to the characteristic vibrational frequencies giving the dominant contribution to the first two PC components. Fig. 6B shows the loading plot of PC1 as a function of the wavenumber. We can see that the loading plot shows the most pronounced changes around the characteristic Raman peaks: C-C coupled with CH$_3$ and C=C stretching bands of carotenoids at 1158 and 1520 cm$^{-1}$ and C-H symmetric and antisymmetric bands of lipids in the region 2800-3000 cm$^{-1}$.

Fig. 6C shows the loading plot of PC2 as a function of the wavenumber. We can see that the loading plot shows the most pronounced changes around the characteristic C-H symmetric and antisymmetric bands of lipids/proteins in the region 2900-3000 cm$^{-1}$. The loading plot for PC2 is almost entirely dominated by the lipid/protein profile in the range of 2800-3000 cm$^{-1}$ and the contribution from carotenoids is negligible.

We have calculated sensitivity and specificity. The pathology results served as the gold standard against which the Raman spectral diagnoses were compared.

Comparing the Raman spectra with the histopatological description we have evaluated the sensitivity and the specificity

of the studied samples taking into account the characteristic Raman peaks from the loading plots for the first principal components PC1 and PC2 (Fig. 6). Sensitivity (the true positive rate), which is defined as the proportion of actual positives which are correctly identified as such (the percentage of sick people who are correctly identified as having the condition), has been calculated as the ratio of the samples with a positive result (lack of the characteristic Raman peaks in tissue) to the total number of the samples that are identified as tumorous tissue according to the histopathological assessment.

We have found the sensitivity of 83% for lipid markers (2800-3000 cm$^{-1}$) and 69% for carotenoids (1158, 1520 cm$^{-1}$). Specificity (the true negative rate), which is defined as the proportion of negatives which are correctly identified as such (the percentage of healthy people who are correctly identified as not having the condition) has been calculated as the ratio of the samples with a negative result (presence of the characteristic Raman peaks in tissue) to the total number of the samples that are identified as normal tissue according to the histopathological assessment. We have found the specificity of 76% for lipid markers (2800-3000 cm$^{-1}$) and 74% for carotenoids (1158, 1520 cm$^{-1}$).

The sensitivity and specificity obtained directly from PLSLD and cross validation gives the sensitivity and the specificity of 90.5% and 84.8% for the calibration and 84.7% and 71.9% for the cross-validation respectively. The method for cross-validation was venetian blinds w/10 splits and the errors for the prediction and cross validated model were RMSEC: 0.33 and RMSECV: 0.43 respectively, showing good stability for predicting new samples.

## Conclusions

The picture that emerges from our results demonstrates that tumor cells in the invasive IDC and IDL breast cancers have evidently mixed proteome-lipidic-ROS profile. Enhanced carotenoid and lipid content as well as decreased proteome were observed in normal cells compared with the cancerous cells. The normal breast tissue contains much more adipose tissue than the cancerous breast tissue that contains much more proteins, which are usually hydrophilic.

We showed that the Raman features of the lipid profile as well as carotenoids, which play diverse functions including antioxidant properties to defense against ROS,[48,49] may be very useful for biodiagnostics of cancer. The lipid-antiROS phenotype was ignored for many years by the Raman scientific community that has focused almost exclusively on the biosynthesis of proteins and the proteome profile.[14,15,19] The proteome interpretation of cancer in the human tissue is not fully justified because it does not take into account the effect of extracellular matrix, which is particularly important for invasive ductal and lobular cancers when the epithelial cells migrate and invade through the basement membrane into the surrounding extracellular matrix consisting of connective tissue, adipose tissue and fibroblasts.

We report the Raman markers based on lipids (particularly fatty acids and triglycerides) and antiROS species, such as carotenoids, are easily and reproducibly measured, and exhibit high specificity and sensitivity. The results presented in the paper suggest that metabolic alterations in tumors extends beyond the Warburg effect and the pathways related to the production of fatty acids activated via multiple lipogenic enzymes as well as antioxidants (carotenoids) forming the defense against ROS are equally important.

The analysis based on Raman imaging, PCA and PLSDA statistical methods demonstrates that the vibrational signatures by Raman spectroscopy/microspectroscopy can accurately predict which breast tissue has normal biochemistry and morphology and clearly distinguishes cancer pathology. The results presented in the paper highlights differences between the Raman spectra of the normal breast tissue from the safety margin surrounding tumor and the cancerous breast tissue from the tumor mass. First, the comparison between the Raman spectra demonstrates that the normal tissue contains Raman peaks characteristic for glycerol monooleate derivatives at 2854, 2888, and 2926 $cm^{-1}$, which are not observed in the Raman spectra of the cancerous tissue dominated by a protein peak at 2931 $cm^{-1}$ (average Raman spectra) and 2940 $cm^{-1}$ (Raman microspectroscopy spectra of epithelial cells embedded in stroma after migration through the basement membrane into the surrounding extracellular matrix) as well as the peak at 3067 $cm^{-1}$ ((C-H)-aromatic rings of proteins). Second, a detailed analysis demonstrates that the strongest signals in the Raman spectra of the noncancerous tissue originate from carotenoids, which are not visible in the Raman spectrum of the cancerous tissue. The peaks from carotenoids are very sensitive to the effect of the extracellular matrix. The average Raman spectra of carotenoids are observed at 1158 cm-1 and 1520 $cm^{-1}$ and at 1161 $cm^{-1}$ and 1527 $cm^{-1}$ in microRaman spectra when focusing laser directly on adipose tissue. The resonance Raman enhancement of carotenoids results from the fact that the noncancerous breast tissue contains a significant contribution of glycerol monooleate derivatives contained in the adipose tissue, which acts as a dynamic reservoir for carotenoids. Third, the substantial biochemical information is imprinted in the Raman spectra of the Amide I in the biological tissue. It is obvious that the distinct peak positions of the Amide I at 1660 $cm^{-1}$ in the cancerous tissue and at 1654 $cm^{-1}$ in normal tissue must be associated with the distinct biological environment.[48,49] Fourth, marked distinctions can be observed in the lipid/protein profiles in the region of 2800-3000 $cm^{-1}$, where the contribution from the monounsaturated fatty acids, common constituents of triglycerides of the adipose tissue, dominates the average Raman spectra of the noncancerous tissue in contrast with the Raman spectra of the cancerous tissue. We have shown that the average Raman spectra of the cancerous tissue demonstrate a more protein like profile. This finding is consistent with the fact that in contrast to normal cells, abnormal cells divide in an uncontrolled process of cell growth that synthesizes large amounts of proteins.

We have evaluated the diagnostic value of the Raman biomarkers for monitoring cancer pathology by applying the principal component analysis PCA and PSLDA. Raman bands responsible for the PC1 score discrimination were the peaks at 1158, 1520 $cm^{-1}$ corresponding to carotenoids. Raman bands responsible for the PC2 score discrimination were 2854, 2888, 2902, 2931, 3009 $cm^{-1}$ related to fatty acids and 2940, 3067 $cm^{-1}$ related to proteins.

The analysis based on 164 samples from 82 patients gives the sensitivity of 83% for lipid markers (2800-3000 $cm^{-1}$) and 69% for carotenoids (1158, 1520 $cm^{-1}$). The specificity of 76% for lipid markers (2800-3000 $cm^{-1}$) and 74% for carotenoids (1158, 1520 $cm^{-1}$) was obtained. The sensitivity and specificity obtained directly from PLSLD and cross validation gives the sensitivity and the specificity of 90.5% and 84.8% for the calibration and 84.7% and 71.9% for the cross-validation respectively.

## Acknowledgements

## Notes and references

[a] Lodz University of Technology, Institute of Applied Radiation Chemistry, Laboratory of Laser Molecular Spectroscopy, Wroblewskiego 15, 93-590 Lodz, Poland

[b] Medical University of Lodz, Department of Pathology, Chair of Oncology, Paderewskiego 4, 93-509 Lodz

[*] Corresponding Author: Prof. Halina Abramczyk, Lodz University of Technology, Institute of Applied Radiation Chemistry, Laboratory of Laser Molecular Spectroscopy, Wroblewskiego 15, 93-590 Lodz, Poland, Phone: (+48 42) 631-31-88, Fax: (+48 42) 684-00-43, Email: halina.abramczyk@p.lodz.pl

1 A. Marusyk and K. Polyak, *Science*, 2013, **339**, 528-529.

2 R. A. Cairns, I. S. Harris and T. W. Mak, *Nature Reviews Cancer*, 2011, **11**, 85-95.

3 O. Warburg, *Science*, 1956, **123**, 309-314.

4 H. Abramczyk and B. Brozek-Pluska, *Chem Rev*, 2013, **113**, 5766-5781.

5 J. Surmacki, J. Musial, R. Kordek and H. Abramczyk, *Mol. Cancer*, 2013, **12**, doi:10.1186/1476-4598-12-48.

6 H. Abramczyk, B. Brozek-Pluska, J. Surmacki, J. Jablonska-Gajewicz and R. Kordek, *Prog Biophys Mol Biol*, 2012, **108**, 74-81.

7 C. Nieva, M. Marro, N. Santana-Codina, S. Rao, D. Petrov and A. Sierra, *PLOS*, 2012, **7**, DOI: 10.1371/journal.pone.0046456

8 X. Bi, B. Rexer, C. L. Arteaga, M. Guo and A. Mahadevan-Jansen, *J. Biomed Opt.*, 2004, **19**, doi: 10.1117/1.JBO.19.2.025001.

9 M. Hedegaard, Ch. Krafft, H. J. Ditzel, L. E. Johansen, S. Has-sing and J. Popp, *Anal Chem*, 2010, **82**, 2797-2802.

10 M. Hilvo, C. Denkert, L. Lehtinen, B. Muller, S. Brockmoller, T. Seppanen-Laakso, J. Budczies, E. Bucher, L. Yetukuri, S. Castillo, E. Berg, H. Nygren, M. Sysi-Aho, J. L. Griffin, O. Fiehn, S. Loibl, Ch.

Richter-Ehrenstein, C. Radke, T. Hyotylainen, O. Kallioniemi, K. Iljin and M. Oresic, *Cancer Res*, 2011, **71**, 3236-3245.

11 K. Bhalla, B. J. Hwang, R. E. Dewi, L. Ou, W. Twaddel, H. B. Fang, S. B. Vafai, F. Vazquez, P. Puigserver, L. Boros and G. D. Girnun, *Cancer Res*, 2011, **71**, 6888-98.

12 I. Notingher and L.L. Hench, *Expert Rev Med Devices*, 2006, **3**, 215-234.

13 C. J. Frank, R. L. McCreery and D. C. Redd, *Anal Chem*, 1995, **67**, 777-783.

14 A. Mahadevan-Jansen and R. R. Richards-Kortum, *J Biomed Opt*, 1996, **1**, 31-70.

15 N. J. Kline and P. J. Treado, *J Raman Spectrosc*., 1997, **28**, 119-124.

16 N. Stone, R. Baker, K. Rogers, A. W. Parker and P. Matousek, *Analyst*, 2007, **132**, 899-905.

17 A. S. Haka, Z. Volynskaya, J. A. Gardecki, J. Nazemi, R. Shenk, N. Wang, R. R. Dasari, M. Fitzmaurice and M. S. Feld, *J Biomed Opt*, 2009, **14**, doi:10.1117/1.3247154

18 M. Diem, M. Miljković, B. Bird, T. Chernenko, J. Schubert and E. Marcsisin, *Spectroscopy: An International Journal*, 2012, **27**, 463-496.

19 A. Mahadevan-Jansen, M. F. Mitchell, N. Ramanujam, A. Malpica, S. Thomsen, U. Utzinger and R. Richards-Kortumt, *Photochem Photobiol*, 1998, **68**, 123-132.

20 A. S. Haka, K. E. Shafer-Peltier, M. Fitzmaurice, J. Crowe, R. R. Dasari and M. S. Feld, *Proc Natl Acad Sci USA*, 2005, **102**, 12371-12376.

21 B. Brozek-Pluska, J. Musial, R. Kordek, E. Bailo, T. Dieing and H. Abramczyk, *Analyst*, 2012, **137**, 3773-3780.

22 Z. Huang, A. McWilliams, S. Lam, J. English, D. McLean, H. Lui and H. Zeng, *Int J Oncol*, 2003, **23**, 649-655.

23 M. G. Shim and B. C. Wilson, *Photochem Photobiol*, 1996, **63**, 662-671.

24 C. J. Behrend, C. P. Tarnowski and M. D. Morris, *Appl Spectroscopy*, 2002, **56**, 1458-1461.

25 U. B. Cappel, I. M. Bell and L. K. Pickard, *Appl Spectroscopy*, 2010, **64**, 195-200.

26 A. Savitzky and M. J. E. Golay, *Anal Chem*, 1964, **36**, 1627-1639.

27 P. Bassan, H. J. Byrne, F. Bonnier, J. Lee, P. Dumas and P. Gardner, *Analyst*, 2009, **134**, 1586-1593.

28 B. M. Wise, N. B. Gallagher, R. Bro, J. M. Shaver, W. Windig and R. S. Koch, PLS_Toolbox Version 4.0 for use with MATLAB™, Eigenvector Research, Inc.3905 West Eaglerock Drive, Wenatchee, WA 98801.

29 K. Varmuza, Introduction to multivariate statistical analysis in chemometrics. Taylor & Francis Group ed. New York: CRC Press, 2009, 59-102.

30 T. Dieing, O. Hollricher and J. Toporski, Confocal Raman Spectroscopy. Springer - Verlag, Berlin Heidelberg, 2010.

31 J. B. MacQueen, Some Methods for Classification and Analysis of Multivariate Observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, University of California Press,1967.

32 M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, Ch. Matthäus, and M. Diem, *Analyst*, 2010, **135**, 2002-2013.

33 M. Barker and W. Rayens, *Journal of Chemometrics*, 2003, **17**, 166-173.

34 A. Rugula, K. Majzner, K. M. Marzec, A. Kaczor, M. Pilarczyk and M. Baranska, *J. Raman Spectr*., 2013, **44**, 1061-1076.

35 R. Tuma, *J. Raman Spectr*, 2005, **36**, 307-319.

36 J. D. Gelder, K. D. Gussem, P. Vandenabeele and L. Moens, *J. Raman Spectr.,* 2007, **38**, 1133-1147.

37 Q. Matthews, A. Jirasek, J. Lum, X. Duan and A. G. Brolo, *Applied Spectroscopy*, 2010, **64**, 871-887.

38 J. Kneipp, T. B. Schut, M. Kliffen, M. Menke-Pluijmers and G. Puppels, *Vib Spectrosc*, 2003, **32**, 67-74.

39 L. Pantanowitz, J. H. Sinard, W. H. Henricks, L. A. Fatheree, A. B. Carter, L. Contis, B. A. Beckwith, A. J. Evans, A. Lal and A. V. Parwani, *Arch Pathol Lab Med,* 2013, **137**, 1710-1722.

40 I. Notingher, *Sensors*, 2007, **7**, 1343-1358.

41 F. S. Parker, Applications of Raman and Resonace Raman Spectroscopy in Biochemistry. Plenum Press, New York, 1983.

42 S. Verier, I. Nothinger, J. M. Polak and L. L. Hench, *Biopolymers*, 2004, **74**, 158-162.

43 H. Abramczyk, M. Kołodziejski and G. Waliszewska, *Journal of Molecular Liquids*, 1999, **79**, 223-233.

44 H. Abramczyk, B. Brozek-Pluska, J. Surmacki, J. Jablonska-Gajewicz and R.Kordek, *Journal of Biophysical Chemistry*, 2011, **2**, 159-170.

45 H. Abramczyk, B. Brozek-Pluska, M. Krzesniak, M. Kopec and A. Morawiec-Sztandera, *Spectrochimica Acta Part A Molecular and Biomolecular of Spectroscopy*, 2014,**129**, 609-623.

46 J. Kong and S. Yu, *Acta Biochimica et Biophysica Sinica*, 2007, **39**, 549-559.

47 W. G. Stetler-Stevenson, S. Aznavoorian and L. A. Liotta, *Ann Rev Cell Biology*, 1193, **9**, 541-573.

48 J. Han, S. E. Hankinson, I. De Vivo, D. Spiegelman, R. M. Tamimi, H. W. Mohrenweiser, G. A. Colditz and D. J. Hunter, *Cancer Res*, 2003, **63**, 8536-8541.

49 C. He, R. M. Tamimi, S. E. Hankinson, D. J. Hunter and J. Han, *Breast Cancer Res Treat*., 2009, **113**, 585-594.
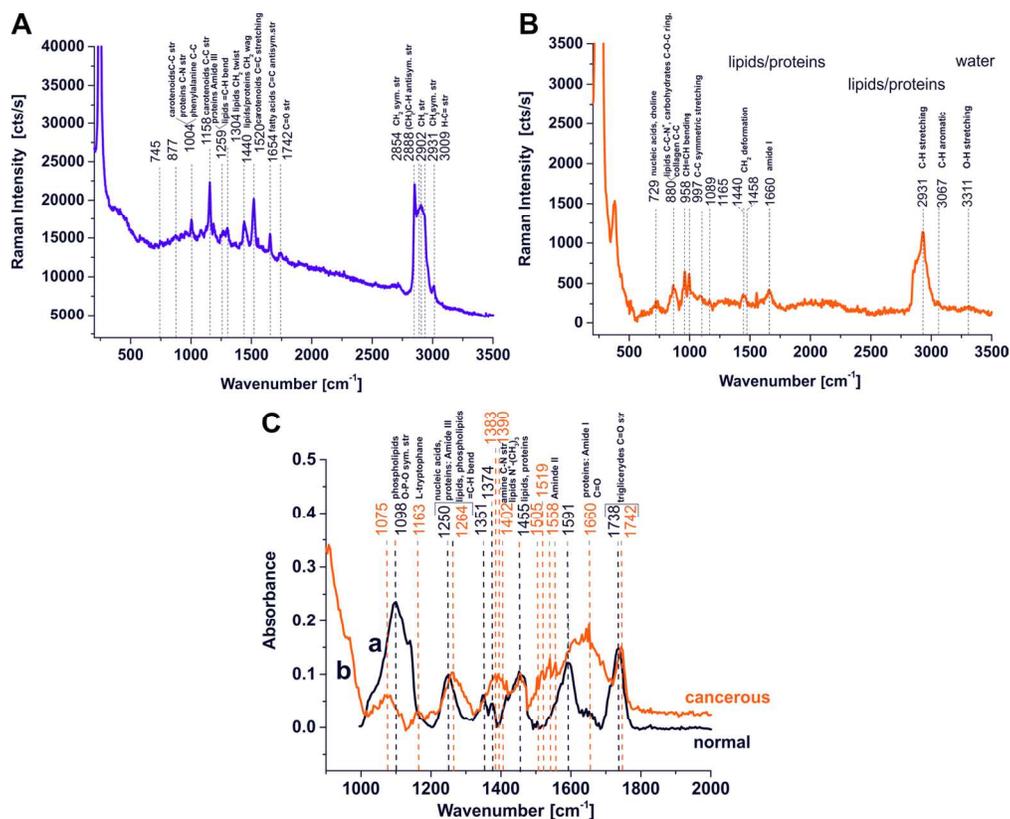
**Analyst Accepted Manuscript**

Raman average spectra of the human breast tissue from the safety margin (A) and tumor mass (B), integration time of 10 s, laser power of 200 mW, slice of 16 μm on $BaF_2$ window. IR spectra (C) of the human breast tissue from the safety margin (a) and tumor mass (b), integration time of 5.0 s, slice of 6 μm on $BaF_2$ window, controlled humidity 23%. Invasive ductal cancer G2 (patient P95).
137x111mm (300 x 300 DPI)
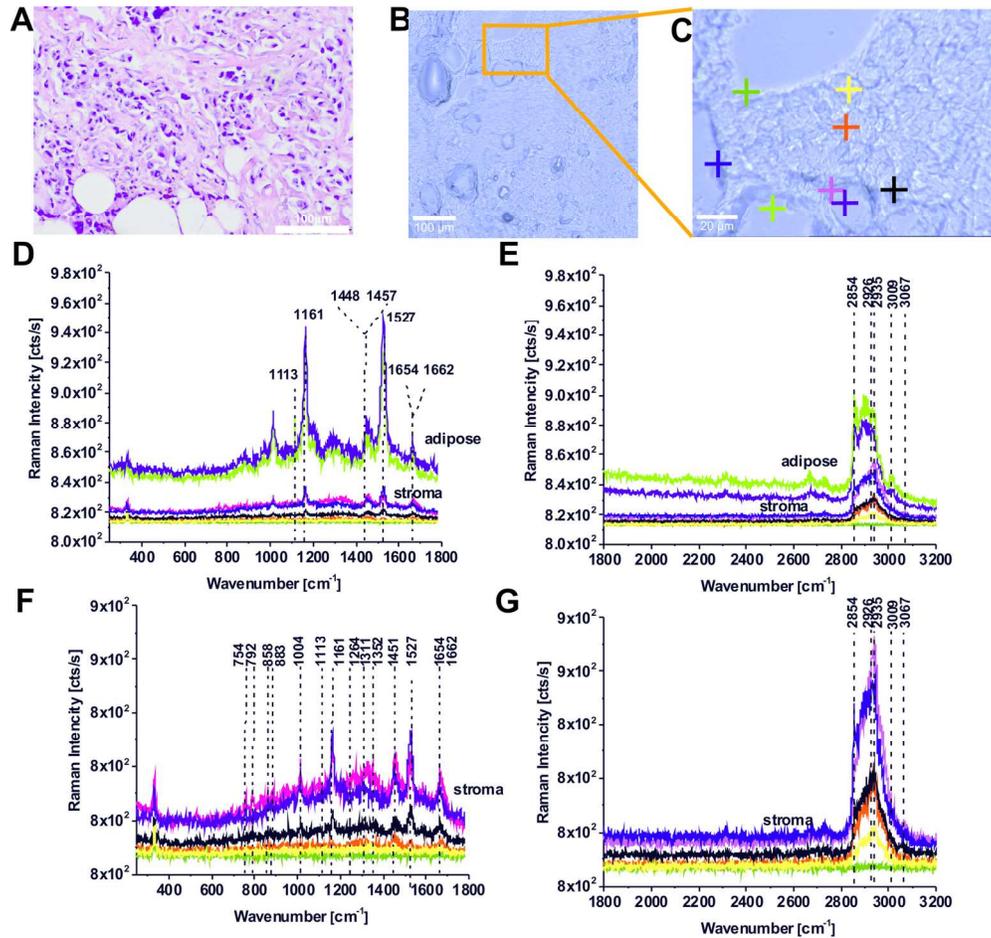
Histological image (A), microscopy stitching image 530x530 μm (B), microscopy image from the region marked with a yellow frame (C) and Raman spectra (D, E, zoomed in on F, G) corresponding to various areas of the breast tissue from the tumor mass (invasive ductal cancer GX, patient P104), colors of the Raman spectra correspond to the positions marked with crosses of the same color.

80x76mm (600 x 600 DPI)
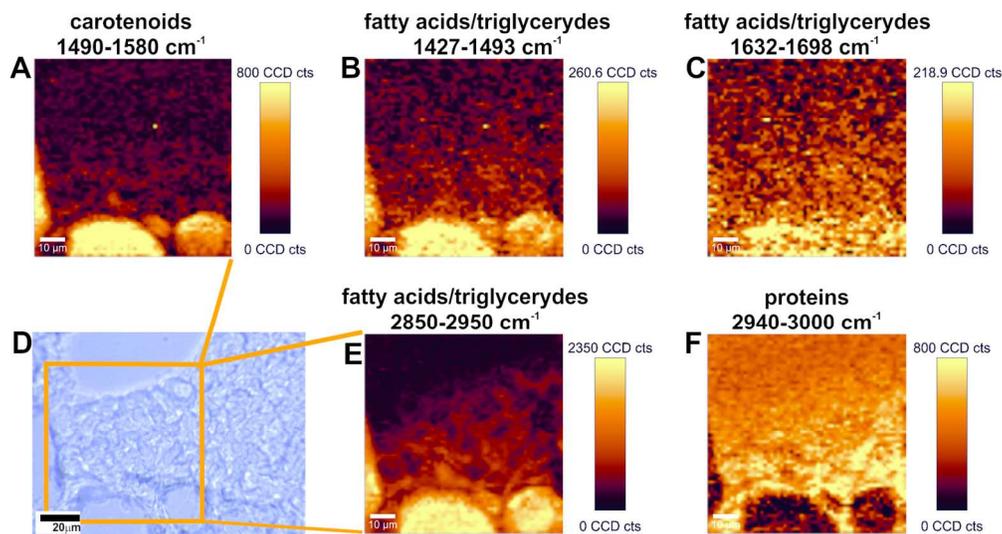
Raman images (80x80 µm, 60x60 points per line) (A, B, C, E, F) from the region marked with a orange frame on microscopy image (D) of the breast tissue from the tumor mass (invasive ductal cancer GX, patient P104). Raman images from the spectral region 1490-1582 cm$^{-1}$ (carotenoids) (A), 1427-1493 cm$^{-1}$ (fatty acids/ triglycerides) (B), 1632-1698 cm$^{-1}$ (fatty acids/ triglycerides) (C), 2850-2950 cm$^{-1}$ (fatty acids/ triglycerides) (E), 2940-3000 cm$^{-1}$ (proteins) (F).
140x73mm (300 x 300 DPI)
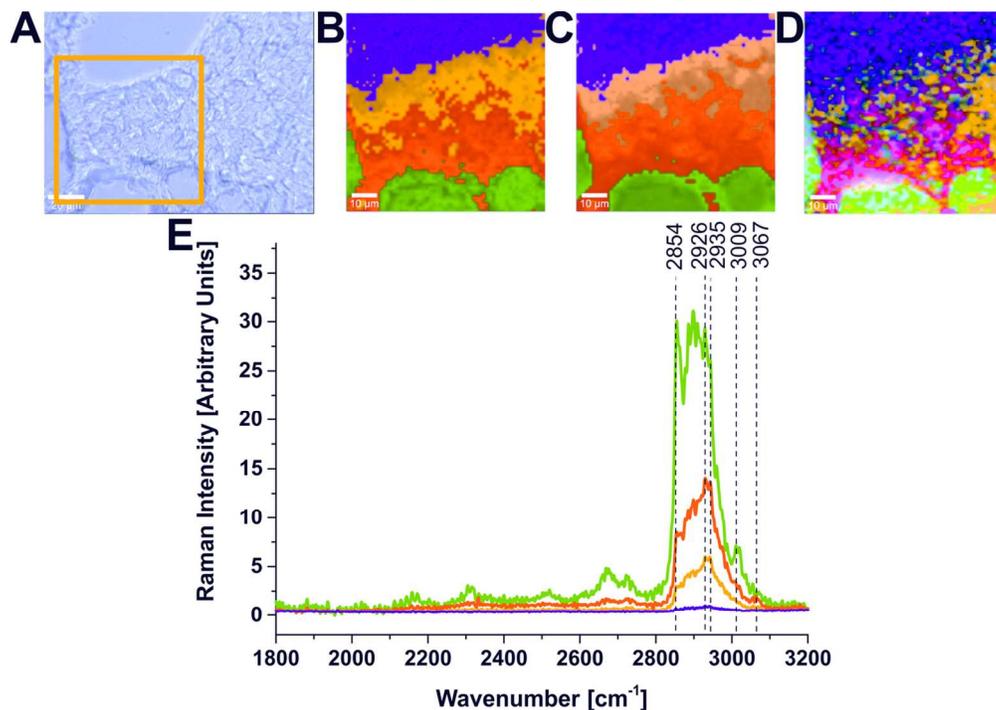
The color Raman maps constructed from the monochrome maps for the spectral filter 1800-3200 cm$^{-1}$, (A) microscopy image, (B) K-means cluster analysis, 4 clusters, Euclidean metrics, (C) K-means cluster analysis, 4 clusters, Manhattan metrics, (D) basis analysis (4 average spectra), (E) spectra representing different clusters (or basis average spectra). The line colors correspond to the colors of the Raman maps. Blue region corresponds to the BaF$_2$ support at the edge of the tissue sample (patient P104).
59x41mm (600 x 600 DPI)
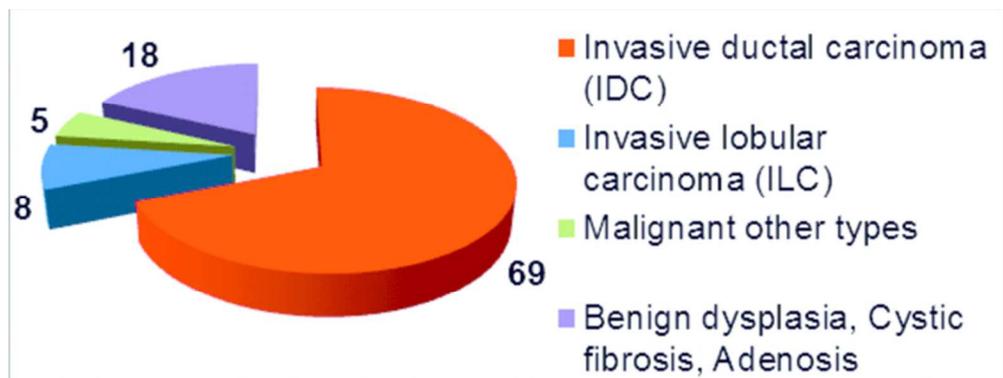
The statistical chart with the numbers of analyzed human breast tissue pathologies.
32x12mm (600 x 600 DPI)
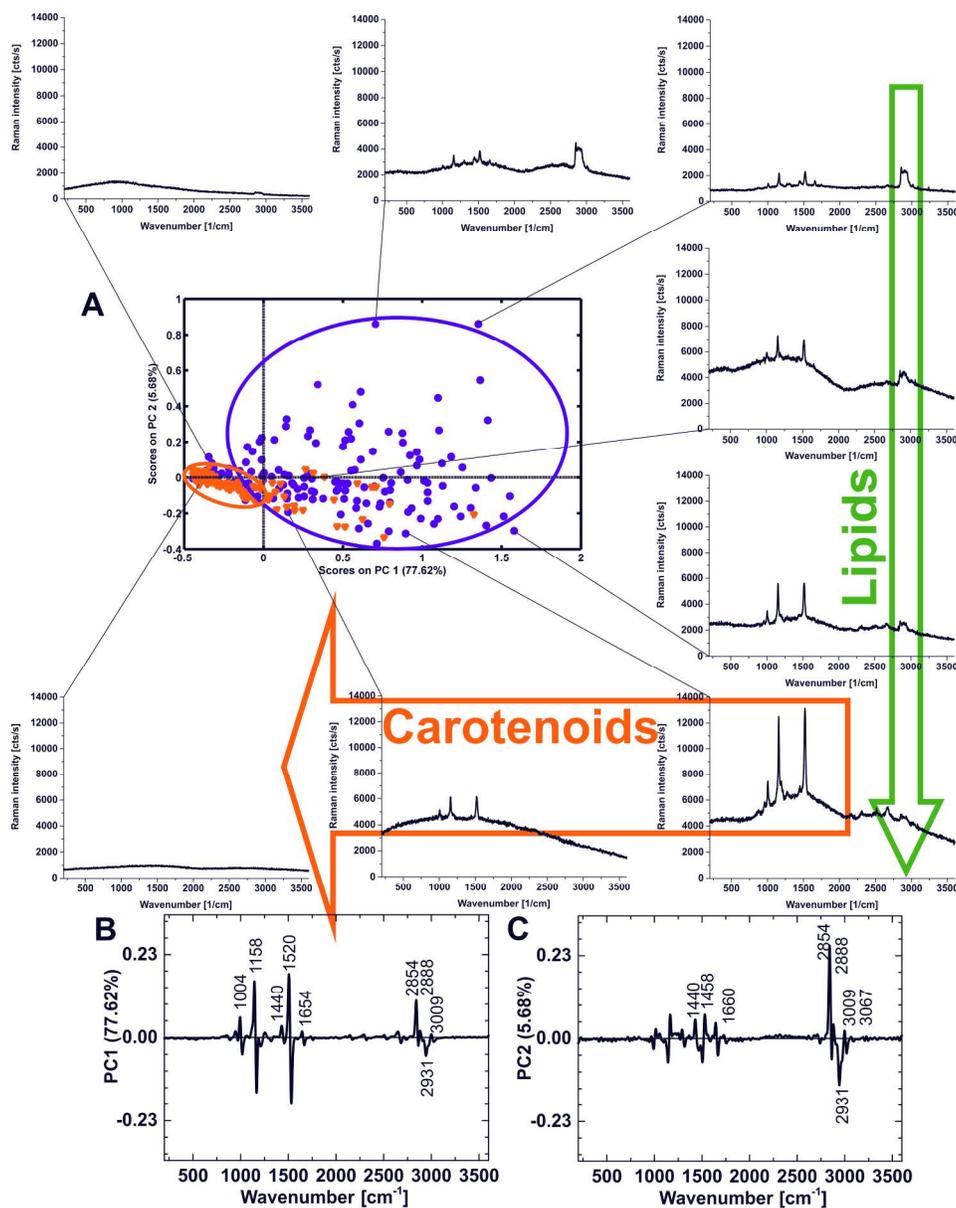
PCA score plot (model: SNV, mean center, 1-st derivative) for the average Raman spectra of the human breast tissue samples from the tumor mass (red triangles) and the safety margin (blue circles), bulk tissue, integration time 0.5 s, laser power 25 mW, step 2 cm$^{-1}$, 200-3600 cm$^{-1}$ (A). PCA loading plot for PC1 (B) and PC2 (C).
107x135mm (600 x 600 DPI)

**Table 1.** Tentative assignments of the vibrational bands of the human breast tissue from the Raman (R) and IR spectra (patient P95)[5,6,19,36-38,40-46]

| Human breast Tissue Wavenumber / cm$^{-1}$ | | Model systems Wavenumber / cm$^{-1}$ | Tentative assignments |
|---|---|---|---|
| Normal | Tumor | | |
| 745 (R) | 729 (R) | 729 | Nucleic acids, Phospholipid (choline)[40-42] |
| 877 (R) | 880 (R) | 880 | Lipids/Carbohydrates/Collagen[40] C-C-N$^+$, C-O-C ring, C-C |
| | 958 (R) | 935 | Hydroxyproline/Collagen backbone[41,42] CH=CH bending |
| | 997 (R) | | C-C symmetric stretching |
| 1004 (R) | | 1004 | Carotenoids/Phenylalanine[5,6,19] C-C stretching of carotenoids/ring breathing of protein[36] |
| | 1075 (IR) | 1068 | Lipids/Collagen[40,41] C-C str. |
| 1083 (R) 1096 (IR) | 1096 (IR) | 1096 | Phospholipids, O-P-O sym. str.[41] |
| 1158 (R) | | 1158 | Carotenoids (C-C coupled with C-CH$_3$)/Proteins (C-C/C-N str.)[37,40,43] |
| | 1160 (IR) | 1160 | L-Tryptophan[36] |
| 1189 (R) | 1199 (IR) | 1199 | C-C$_6$H$_5$ Phe, Trp[40] |
| | 1238 (IR) | 1240 | Phospholipid, O-P-O antisym. str.[41] |
| 1250 (IR) | 1264 (IR) | 1220-1285 | Nucleic acids (Try, Ala)/Proteins (Amide III), Lipid, phospholipid =C-H bend[40,41] |
| 1304(R) | | 1304 | Lipids, phospholipids[40] C-H$_2$ twist |
| | 1383 (IR) | 1383 | Lipids CH$_3$ sym. bend., lipids[41] |
| | 1402 (IR) | 1402 | Amine C-N str.[41] Lipids N$^+$(CH$_3$)$_3$ sym. bend.[38] |
| 1440 (R) | 1440 (R) 1458 (R) 1455(IR) | 1444 1461 | Fatty acids, triglycerides, CH$_2$ or CH$_3$ deformations[36] Lipids/Proteins[19,37] C-H wag, CH$_2$ or CH$_3$ def. Phospholipids, CH$_2$ scissoring[38] |
| 1520(R) | | 1528, 1523 | Carotenoids[37,43] C=C str. |
| | 1558 (IR) | 1556 | Amide II, proteins[36,41] |
| 1642 (R) | | 1642 | Water OH bend[41] |
| 1591 (IR) | 1593 (IR) | 1586 | Proteins,[36] nucleic acids[40,41] |
| 1654 (R) | | 1655 | Unsaturated fatty acids, triglycerides (C=C) str.[36] |
| | 1660 (R) | | Proteins Amide I α helix, C=O |
| | 1642-1686 (IR) | 1655-1680 | Proteins Amide I[41]/ Unsaturated fatty acids,[37,40] α helix, (C=O) str., (C-H) def./(C=C) str.,[37,40] collagen, elastin[41] |
| 1742 (R) 1742 (IR) | 1742 (IR) | 1743 | (C =O) stretching, triglycerides[36] |
| 2854 (R) | | 2854 | Fatty acids, triglycerides, C-H$_2$ sym. str. |
| 2888 (R) | 2888 (R) | 2888 | Lipids[37], C-H$_2$antisym. str. |
| | 2931 (R) | 2935 | Proteins/Lipids, CH$_3$ sym. str.[37,41] |
| 3009 (R) | | 3008 | Lipids,[37,41] =C-H str. |
| | 3067 (R) | 3060 | Nucleic acids/Proteins,[37] C-H aromatic |
| | 3311 (R) | 3311 | OH stretching of interfacial water[44,45] N-H stretching[46] |