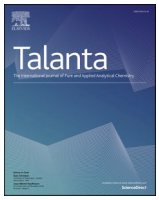
Talanta 254 (2023) 123858

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta



Rapid and low-cost liquid biopsy with ATR-FTIR spectroscopy to discriminate the molecular subtypes of breast cancer

Nikolas Mateus Pereira de Souza [a], Brenda Hunter Machado [b], Licerio Vicente Padoin [c], Daniel Prá [a,d], André Poisl Fay [g], Valeriano Antonio Corbellini [d,e,f], Alexandre Rieger [a,d,f,*]

[a] Department of Life Sciences, University of Santa Cruz do Sul, Santa Cruz do Sul, RS Brazil
[b] International affairs, International University Center, Santa Cruz do Sul, RS, Brazil
[c] Mastology Service at the Hospital of the Federal University of Santa Maria, Santa Maria, RS, Brazil
[d] Postgraduate Program in Health Promotion, University of Santa Cruz do Sul, Santa Cruz do Sul, RS, Brazil
[e] Department of Sciences, Humanities and Education, University of Santa Cruz do Sul, Santa Cruz do Sul, RS, Brazil
[f] Postgraduate Program in Environmental Technology, University of Santa Cruz do Sul, RS, Brazil
[g] Postgraduate Program in Medicine and Health Sciences, School of Medicine, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, RS, Brazil

ARTICLE INFO

Keywords:
Infrared spectroscopy
Breast cancer
Orthogonal partial least squares discriminant analysis
Chemometrics
Liquid biopsy

ABSTRACT

Breast cancer (BC) is the most prevalent cancer worldwide. The prognosis and survival of these patients are directly related to the diagnostic stage. Even so, the gold standard screening method (mammography) has a long waiting period, high rates of false positives, anxiety for patients, and consequently delays the diagnosis by core needle biopsy (invasive method). Alternatively, the Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectroscopy is a noninvasive, low-cost, rapid, and reagent-free technique that generates the spectral metabolomic profile of biomolecules. This makes it possible to assess systemic repercussions, such as the BC carcinogenesis process. Blood plasma samples (n = 56 BC and n = 18 controls) were analyzed in the spectrophotometer in the ATR-FTIR mode. For the exploratory analysis of the data, interval Principal Component Analysis (iPCA) was used, and for predictive chemometric modeling, the Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) algorithm with validation by leave-one-out cross-validation. iPCA in the region of 1118–1052 cm^{-1} (predominantly DNA/RNA bands) showed significant clustering of molecular subtypes and control. The OPLS-DA model achieved 100% accuracy with only 1 latent variable and Root Mean Square Error of Cross-Validation (RMSECV) < 0.005 for all molecular subtypes and control. The wavenumbers (cm^{-1}) with the highest iPCA peaks (loadings: 1117, 1089, 1081, 1075, 1057, and 1052) were used as input to MANOVA (Wilks' Lambda, $p < 0.001$ between molecular subtypes and control). The rapid and low-cost detection of BC molecular subtypes by ATR-FTIR spectroscopy would plausibly allow initial screening and clinical management, improving prognosis, reducing mortality and costs for the health system.

1. Introduction

Female breast cancer is the leading cause of global cancer incidence and fifth leading cause of cancer mortality worldwide [1]. The mortality scenario can be plausibly explained in the lack of understanding of the biological heterogeneity of breast cancer [2] and late diagnosis stage [3]. In turn, diagnosis stage is based on different criteria such as pathological stage, clinical stage and grade combined with molecular subtypes [4]. This work as a guide for personalized treatment, cost-cutting [5,6], and helping to understand different clinicopathological characteristics and distinct patterns of survival that affect clinical management [7].

The pathological stage is based on surgical findings and combines the characteristic of the tumor (T, from 0 to 4) with the presence of lymph node metastasis (N, from 1 to 3) or disseminated metastasis to other organs (M, from 0 to 1). Therefore, the pathological stage is directly related to tumor size and dissemination, while molecular subtypes express significant metabolomic varieties that can serve as biomarkers.

* Corresponding author. Department of Life Sciences, University of Santa Cruz do Sul, Avenida Independência, 2293, Lab 1206, Santa Cruz do Sul; CEP 96815-900, Brazil.
E-mail addresses: nikolas1@mx2.unisc.br (N.M. Pereira de Souza), brendamachado664@gmail.com (B.H. Machado), drpadoin@gmail.com (L.V. Padoin), daniel_pra@yahoo.com (D. Prá), andre.fay@pucrs.br (A.P. Fay), valer@unisc.br (V.A. Corbellini), rieger@unisc.br (A. Rieger).

https://doi.org/10.1016/j.talanta.2022.123858
Received 19 May 2022; Received in revised form 14 August 2022; Accepted 17 August 2022
Available online 21 August 2022

Clinical stage, on the other hand, combines clinical and Breast Imaging-Reporting Data System (BIRADS) findings. Grade assesses tumor aggressiveness (grades I-III) and is based on association of microscopic findings including gland differentiation, nuclear features, and mitotic activity [4].

Screening in molecular subtypes is decisive as predictive and prognostic factor [8]. The main molecular subtypes can be represented by 4 immunohistochemical types (Luminal A, Luminal B, HER2+ and Triple-negative) according to the expression of estrogen receptors (ER), progesterone receptors (PR), human epidermal growth receptor factor 2 (HER2) and cell proliferation marker (Ki-67). Thus, Luminal A is characterized by (PR+ and/or ER+, and ki67 $<$ 14%), Luminal B (PR+ and/or ER+, and ki67 $>$ 14%), HER2 (HER2+), and Triple-negative (PR-, ER- and HER2-) [7,9].

Age, histologic type, stage at diagnosis, hormonal contraceptive methods, postmenopausal hormone therapy, genetic factor, smoking, and other risk factors admittedly affect survival [10]. But stage at diagnosis is the strongest predictor of survival. Patients in stage IV have 27 to 38 times more risk of death when compared to early stages I and II depending on the molecular subtype and highlighting the need for early screening [11].

Breast cancer is a pathology that requires a diagnostic evaluation that includes self-examination, physical examination by a professional, mammography (gold standard screening technique), and when nodules are suspected (BIRADS $\geq$ 3), a biopsy is performed for definitive diagnosis. This series of steps can take a long time to wait, generating anxiety for the patient, worse prognosis, and rising costs for the health system. Furthermore, mammography has its effectiveness reduced with higher tumor densities and with smaller tumors (less than 1 mm, about 100,000 cells) and does not provide any indication of eventual disease outcome [12,13]. Core biopsy, on the other hand, has high sensitivity and specificity, but it is an invasive method that causes insecurity in the patients and it may take a long time until its realization and results [14]. Ultrasound is another common screening technique, but it is operator-dependent, excessive levels of subcutaneous fat spoil the results, and it has low resolution for very small masses [13].

Attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy is a technique capable of extracting chemical information from the vibrational energy of chemical bonds in biomolecules (nucleic acids, carbohydrates, lipids, and proteins) when applied to blood plasma [15]. Thus, complex biochemical samples can be globally analyzed quantitatively (to determine the concentration of a specific molecule) and qualitatively (through the analysis of spectral differences of the characteristic bands of biomolecules that are associated with the analyzed pathology) [16,17]. Due to its high sensitivity, specificity, and possibility to detect biochemical changes by analyzing all molecules simultaneously, ATR-FTIR becomes an excellent tool for screening numerous pathologies early. Depending on the molecular subtype and stage of breast cancer there are fluctuations of analytes such as nucleic acids, extracellular vesicles, lipids, proteins, and other biological components that are released into the bloodstream by tumor cells [18,19]. Therefore, ATR-FTIR spectroscopy can be used as a technique to perform liquid biopsy, identifying biomolecular changes in spectral bands [20,21].

Several studies have been presented to differentiate cancer from non-cancer or detect cancer stages using infrared spectroscopy [22–26]. However, this present study is a pioneer in the differentiation of molecular subtypes of BC using ATR-FTIR coupled with chemometric techniques. This makes this methodology much simpler, faster, and cheaper to be adapted as point-of-care testing or as an alternative to current screening methods for BC.

Considering metabolic fluctuations and carcinogenic variability (due to molecular subtype, tumor size, location, stage of diagnosis, among others) of BC, this study evaluates the applicability of ATR-FTIR spectroscopy coupled with chemometric techniques to discriminate the molecular subtypes of breast cancer in blood plasma samples.

2. Materials and methods

2.1. Sampling

Blood samples from women were collected at the Hospital of the Federal University of Santa Maria (UFSM), Rio Grande do Sul, Brazil. Women diagnosed with stage I, II, and III of breast cancer (n = 56) by core biopsy regardless of the histological type were included in this study. Tumor tissue obtained by core biopsy was used in immunohistochemical analysis to determine molecular subtypes. The control group [n = 18, age 41 (34–48)] was composed of random healthy women in routine appointment without a diagnosis of breast cancer. Women with benign neoplasia were not included. Blood samples were collected by venipuncture into tubes containing K3 EDTA as anticoagulant. The plasma was obtained by centrifugation at 1500 g. The study was approved by the Ethics Committee of Pontifical Catholic University of Rio Grande do Sul, RS, Brazil (CAAE: 01509918.2.0000.5336; Evaluation Report: 3101887). Only individuals who formally consent to participate by signing the Informed Consent Form were included in the study. The descriptive analysis of women with breast cancer in this study is presented in Table 1.

2.2. ATR-FTIR acquisition

The plasma aliquots of 74 samples were deposited on the reading UATR-ZnSe crystal and dehydrated in air current (60–65 °C) for 1.5 min. The readings were performed in triplicate in a spectrometer (Spectum™ 400 FT-IR/FT-NIR, PerkinElmer) in the attenuated total reflectance mode (ATR-FTIR), in the range of 650–4000 cm^{-1}, with 4 cm^{-1} of spectral resolution and 4 scan pulses. Prior to the spectral acquisition of a new sample, the crystal was cleaned with distillated water and a background spectrum was collected to minimize environmental variations and other interferences.

2.3. ATR-FTIR spectra

The mean spectrum without data pre-treatment of each molecular subtype and control was plotted to analyze the absorbance differences in each wavenumber (Fig. 1a). Fig. 3c shows the mean spectrum of control and BC (mean of the four molecular subtypes). Fig. 3b,d is the magnification of the selected region (1118-1052 cm^{-1}) used to perform principal component analysis (PCA). The main vibration of interest for breast cancer analysis selected in PCA loadings (symmetric stretching vibration of PO_2^-) is highlighted. Analysis of variance (one-way ANOVA) and Tukey's Post-Hoc test were applied to the wavenumbers (cm^{-1}) selected in the PCA loadings (1117, 1089, 1081, 1075, 1057, and 1052) to verify whether there was a significant difference in the absorption of these wavenumbers between the molecular subtypes and control. In addition, multivariate analysis of variance (one-way MANOVA) was applied considering all these wavenumbers at once. The multivariate analysis complements the univariate analysis, as it verifies the influence of the set of variables in relation to the outcome.

Table 1
Descriptive analysis of women with breast cancer.

	Breast cancer (n = 56)			
	LA (n = 31)	**LB (n = 10)**	**HER2+ (n = 12)**	**TN (n = 3)**
Age	55 (46–67)	55 (48–65)	50 (41–59)	49 (48–58)
Stage				
I.	9	1	3	1
II.	17	8	9	2
III.	5	1	0	0
Size (cm)	2.6 (1.7–3.4)	2.8 (2.5–3.2)	1.8 (1.5–2.5)	2.3 (1.7–3.1)

Age and size are represented as median (25%–75%). Abbreviations: LA: Luminal A; LB: Luminal B; TN: Triple-negative.

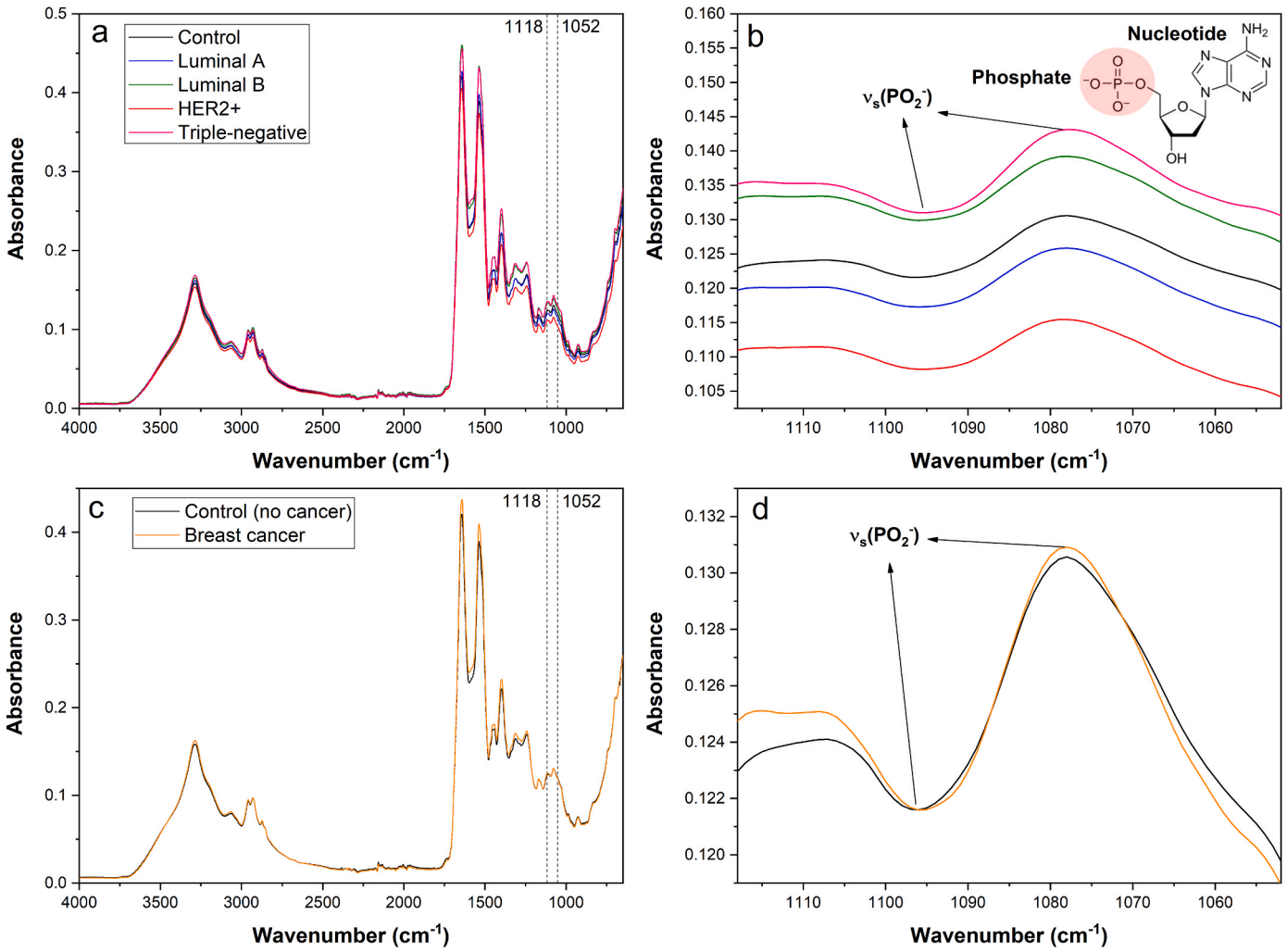


Fig. 1. Average ATR-FTIR spectra without pre-treatment with appointment of symmetric phosphate stretching modes [$\nu_s(PO_2^-)$] originate from the phosphodiester groups in nucleic acids. (a) Total spectra (4000-650 cm^{-1}) of breast cancer molecular subtypes and control. (b) Magnification of the spectral region used in PCA (1118-1052 cm^{-1}). (c) Total spectra (4000-650 cm^{-1}) of women with breast cancer (mean of all molecular subtypes) and control. (d) Magnification of the spectral region used in PCA (1118-1052 cm^{-1}). Abbreviations: ν_s: symmetric stretching vibration; PCA: principal component analysis.

2.4. Interval principal component analysis (iPCA)

PCA is an unsupervised chemometric technique for exploratory analysis of spectral data [27]. It works reducing the dimensionality of the data preserving the greatest variance of the data and projecting it into a new system of axes (principal components, PC) that retain as much information as possible, joining the groups that have the highest correlations between their variables [28]. The PCA version by intervals (iPCA) divides the ATR-FTIR spectrum into equal fragments. Therefore, the greater the number of fragments, the smaller the number of wavenumbers per interval. This means that you can find regions with higher group separation potential than using the total ATR-FTIR spectrum.

iPCA was applied aiming to find groupings of molecular subtypes and control with better separation in a specific spectral interval (Fig. 2). The total ATR-FTIR spectrum (4000-650 cm^{-1}) was empirically fragmented in 10, 20, 30, 40, 50, and 60 regions with the same number of wavenumbers in each interval. The purpose of this was to determine which biomolecular region would have the greatest potential for grouping samples belonging to the same class (molecular subtypes or control). The data was mean-centered. The region chosen was the one that demonstrated the best grouping of classes in visual inspection. The loadings (in this case, the weight of each wavenumber in determining the spatial position of each sample in the orthogonal plane of the PC) for the selected region were plotted in Fig. 2d. The wavenumbers corresponding with the peaks of the loadings are highlighted. To verify whether the spatial separation of the groups shown in the iPCA (Fig. 2) was significant, the non-parametric Kruskal-Wallis test and Dunn's Post-Hoc multiple comparisons with Holm correction (Table 2) was applied considering the PC1, PC2, PC3, and PC4 scores values. Table 3 describes the chemical correspondence of the wavenumber peaks selected in the iPCA loadings. The iPCA was performed with the ChemoStat V.2 software (Santa Cruz do Sul, RS, Brazil) and univariate statistics with Jasp 0.14.1 (Amsterdam, North Holland, Netherlands).

2.5. Orthogonal partial least squares discriminant analysis (OPLS-DA)

OPLS-DA works by maximizing the covariance between the independent and dependent variable in a new linear subspace with reduction in the number of factors (latent variables, LVs). With this, it is possible to make the prediction of new dependent variables and discriminate them [29,30].

OPLS-DA was used for supervised classification of molecular subtypes and control. The spectral data were processed in the following sequence: min-max normalization (0–1) of spectral replicates, calculation of the mean spectrum of each spectral triplicate set, derivative of Savitzky-Golay algorithm (filter width = 5, polynomial order = 2, and

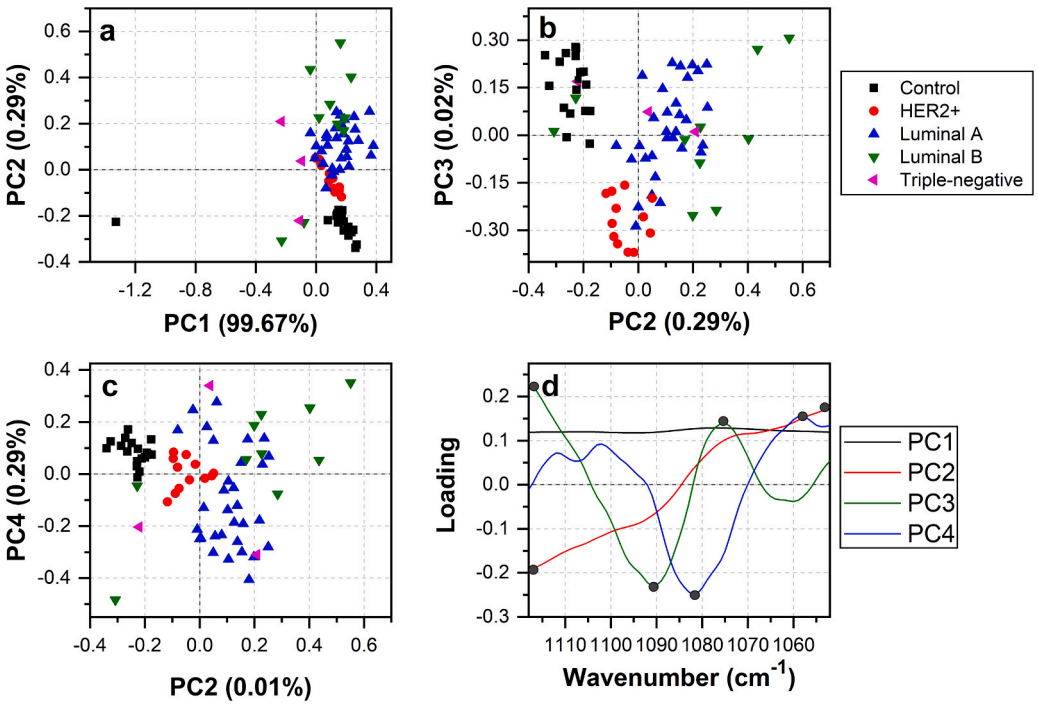


Fig. 2. PCA of spectral region of 1118–1052 cm^{-1} with data mean-centered. (a) PCA scores (PC1 x PC2). (b) PCA scores (PC2 x PC3). (c) PCA scores (PC2 x PC4). (d) Loadings of PC1, PC2, PC3, and PC4 with peak marking. Abbreviations: PCA: principal component analysis; PC: principal component.

Table 2
Kruskal-Wallis test and Dunn's Post-Hoc multiple comparisons with Holm correction for each Principal Component (PC).

	PC1	PC2	PC3	PC4
Kruskal-Wallis Test	0.004[a]	<0.001[a]	<0.001[a]	0.004[a]
Control - Luminal A	0.324	<0.001[a]	0.012[a]	0.002[a]
Control - Luminal B	0.18	<0.001[a]	0.072	0.963
Control - HER2+	0.07	0.06	<0.001[a]	0.187
Control - Triple-negative	0.006[a]	0.103	0.897	0.479
Luminal A - Luminal B	0.286	0.592	0.897	0.074
Luminal A - HER2+	0.194	0.023[a]	<0.001[a]	0.708
Luminal A - Triple-negative	0.019[a]	0.592	0.897	0.963
Luminal B - HER2+	0.392	0.022[a]	0.008[a]	0.578
Luminal B - Triple-negative	0.18	0.469	0.897	0.646
HER2+ - Triple-negative	0.194	0.592	0.024[a]	0.963

[a] Significant values ($p < 0.05$).

Table 3
Selected wavenumbers (peaks) in principal component analysis loadings and their chemical correspondence.

Peak (cm^{-1})	Assignment	Reference
1052	Phosphate I band for two different C–O vibrations of deoxyribose in DNA in A and B forms of helix	[34]
1057	Stretching C–O deoxyribose	[35]
1075	Symmetric stretching vibration of PO_2^-	[36]
1081	Symmetric stretching vibration of PO_2^- and one of the triad peaks of nucleic acids (along with 1031 and 1060)	[36]
1089	Symmetric stretching of PO_2^- in RNA	[34]
1117	C–O stretching vibration of C–OH group of ribose (RNA)	[37]

derivative order = 1) and 1 component of orthogonal signal correction (OSC) concerning each mean spectrum. Leave-one-out cross-validation (LOOCV) was used to evaluate the performance of OPLS-DA models and determine the most robust number of latent variables (Fig. 3) according to ASTM E1655-17 rule [31]. LOOCV was chosen as the validation method due to the small number of samples per class (molecular subtype). The OPLS-DA models were constructed using the total spectrum (4000-650 cm^{-1}) and performed with the Pirouette 4.0 software (Infometrix, Bothell, Washington, USA).

3. Results

3.1. Spectral analysis

In Fig. 1d it is observed that the area under the curve (AUC, integral) of the mean spectrum of women with breast cancer is greater than that of women without cancer (controls). This is important because this region selected in the iPCA (1118-1052 cm^{-1}) is predominantly composed of molecular vibrations associated with nucleic acids. There was a significant difference for all wavenumbers selected by iPCA loadings (1117, 1089, 1081, 1075, 1057, and 1052) only between control and HER2+ ($p < 0.05$). Among the other subtypes and control there was no difference. However, it was obtained significant difference ($p < 0.001$) at Wilks' Lambda of MANOVA among all molecular subtypes and control in all loadings selected in the iPCA. This highlights the importance of multivariate analysis of the joint contribution of wavenumbers. This occurs because different wavenumbers can have chemical correspondence in the same biological group of molecules. This may justifies the fact that the mean AUC of patients with BC is greater than that of the control, due to the greater contribution of absorbance of several wavenumbers with correspondence of vibrations of DNA/RNA molecules.

3.2. Chemometrics

3.2.1. iPCA

The region with the best visual separation was 1118–1052 cm^{-1} (when spectrum fragmented into 50 regions) with 67 variables (wavenumbers). Data were only mean-centered without normalization. When min-max normalization or vector normalization was applied, satisfactory results were not obtained with loss of information in the chosen region. The best combinations of PCs that demonstrated visual differentiation between molecular subtypes and control were chosen (Fig. 2). Fig. 2b (PC2 x PC3) is the one that presents the best difference between

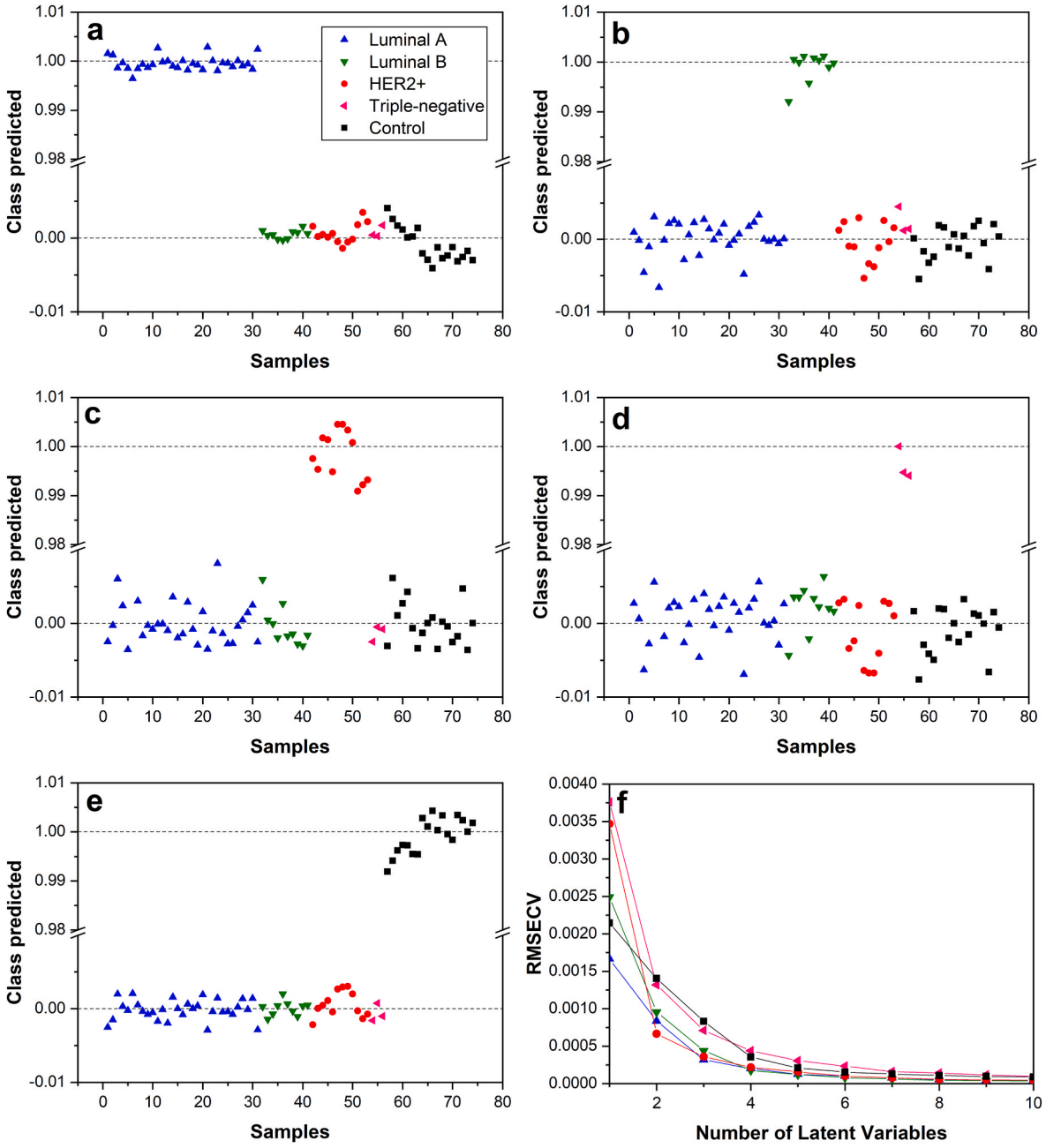


Fig. 3. OPLS-DA prediction model generated by leave-one-out cross-validation for 1 latent variable (LV) for Luminal A (a), Luminal B (b), HER2+ (c), Triple-negative (d), and control = normal (e). The value "1″ refers to the samples of interest to be predicted and the value "0″ to the other samples. The model classified 100% of the samples correctly for all 5 models. (f) The variations of RMSECV with the increase in the number of latent variables for each class (control and molecular subtypes). Abbreviations: OPLS-DA: orthogonal partial least squares discriminant analysis; RMSECV: root mean square error of cross-validation.

the molecular subtypes evidenced by the Dunn's Post Hoc test (Table 2). The combination of PC1 x PC2 x PC3 was able to significantly differentiate the control of all molecular subtypes (p-values in Table 2). PC2 discriminated the control group from Luminal A and Luminal B with a significant spatial difference ($p < 0.001$). Only PC3 was able to discriminate between control and HER2+ ($p < 0.001$) and only PC1 was able to differentiate the control from the Triple-negative ($p < 0.01$). HER2+ had significant spatial separation by PC3 for control ($p < 0.001$), Luminal A ($p < 0.001$), Luminal B ($p < 0.01$), and Triple-negative ($p < 0.05$). Luminal A and Luminal B were not separated by any PC. This probably occurred due to the similar immunohistochemical profile of both.

The chemical designations of the peaks selected in the iPCA loadings are represented in Table 3. Notably vibrations related to nucleic acid bonds have been selected.

3.2.2. OPLS-DA

Fig. 3f indicates the root mean square error of cross-validation (RMSECV) for each latent variable in each class. When the number of LVs is increased, the RMSECV decreases. However, there is a risk of overfitting the model to the intrinsic characteristics of the dataset with higher numbers of LVs. Considering that only 1 latent variable already presented RMSECV < 0.005 for all classes, only this one was used to create the OPLS-DA model. The variance conserved in LV1 for each model of prediction of molecular subtypes and control were: control (94.64%), Luminal A (96.47%), Luminal B (90.81%), HER2+ (91.39%),

and Triple-negative (73.19%). The lowest variance value for Triple-negative is due to the smallest number of samples for this class. Thus, there is less representativeness of the variation of this class in LV1 when arranging the orthogonalization process. However, for all other classes, a high variance (>90%) was obtained. Therefore, considering the high variance of LV1 and the low RMSECV, we can infer that the model is very suitable for predicting new external samples.

Fig. 3a–e shows the values predicted by the OPLS-DA model. The class to be predicted is represented by number 1 and everything else by number 0. The model was able to correctly classify 100% of all classes considered.

The effects of molecular subtype and cancer stage were differentiated by the loadings of LV1 observed in OPLS-DA models elaborated for each subtype and control (Fig. 4). This is because the molecular subtypes have different metabolomic profiles, while the stages relate more to the concentration of certain analytes than to the metabolomic variety [32,33]. Furthermore, the stages are mainly associated with the size of the tumor. In our sample of subjects (Table 1) there was no significant difference (Kruskall-Wallis test, $p > 0.05$) in tumor size between molecular subtypes. In the analysis of the OPLS-DA models loadings, can be observed the most relevant positive contributions at 1700-1650 cm^{-1} (Amide I) and 1555-1540 cm^{-1} (Amide II), and negative contributions at 1635-1620 cm^{-1} (Amide I) and 1520-1470 cm^{-1} (Amide II). The other region with the highest loadings outside the amide spectrum was at 1420-1400 cm^{-1} (positive contribution) and 1390-1375 cm^{-1} (negative contribution). This region between 1420 and 1375 cm^{-1} is a combination of the vibrations of ν(C–N), δ(C–H), and δ(N–H) [38].

4. Discussion

The major goal of differentiating molecular subtypes is early screening with the possibility of personalized treatment. The luminal A and luminal B molecular subtypes tend to have a better prognosis and the possibility of antiestrogenic treatment with tamoxifen or aromatase inhibitors [39]. The HER2+ subtype is more aggressive and characterized by the high expression of the HER2 oncoprotein and the non-expression of estrogen or progesterone receptors [40]. Thus, patients with the HER2+ subtype benefit from therapy with the monoclonal antibody trastuzumab, which acts by binding to the HER2 receptor and down-regulating it. Adjuvant trastuzumab reduces the risk of recurrence by half and mortality by one third in cancer patients primary stage breast. The Triple-negative subtype has the worst prognosis, with no possibility of therapy with trastuzumab, tamoxifen or aromatase inhibitors [41,42]. The benefits of therapy for early stages of breast cancer were estimated in the study by Burstein et al. [43]. Because of this, our methodology based on ATR-FTIR spectroscopy aims to screening molecular subtypes in early stages, considering the therapeutic and prognostic benefits.

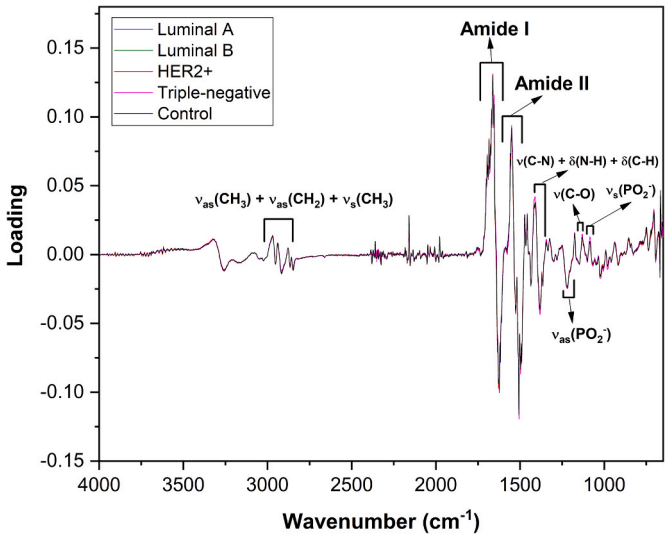


Fig. 4. Loadings of the five OPLS-DA models for breast cancer molecular subtypes and control. It can be observed the most relevant positive contributions at 1700-1650 cm^{-1} (Amide I), 1555-1540 cm^{-1} (Amide II), and 1420-1400 cm^{-1}, and negative contributions at 1635-1620 cm^{-1} (Amide I), 1520-1470 cm^{-1} (Amide II), and 1390-1375 cm^{-1}. This region between 1420 and 1375 cm^{-1} is a combination of the vibrations of ν(C–N), δ(C–H), and δ(N–H). Abbreviations: OPLS-DA: orthogonal partial least squares discriminant analysis; ν: stretching vibration; δ: deformation vibration.

The selected region (1052-1118 cm^{-1}) for principal component analysis has numerous bonds that are mainly associated with DNA and RNA molecules [38]. Emphasis on the wavenumbers 1075 and 1081 cm^{-1}, which were among the highest weights in loadings in the discrimination of molecular subtypes. These wavenumbers are related to symmetric phosphate stretching modes [$\nu_s(PO_2^-)$] originate from the phosphodiester groups in nucleic acids and suggest an increase in the nucleic acids in the malignant tissues [36,38]. In the study by Sitnikova et al. [44], for detection of patients with BC and without BC in blood serum by ATR-FTIR, a significant difference was observed in the same phosphate region highlighted in our findings. As well as previous studies that discuss oscillations in the structure of DNA and RNA in BC [45] and other cancers such as colorectal [46] and lung [47]. In the analysis of DNA of breast tissue with infrared microscope by Malins et al. [45], substantial oxidative changes in DNA base structures in mutagenesis and carcinogenesis of BC were verified. These changes were also reflected in the phosphodiester backbone and the deoxyribose moiety. In addition, the study by Zelig et al. [48] using FTIR spectroscopy in peripheral blood, they found two regions with a significant difference ($p < 0.05$, t-test) between patients with breast cancer and without cancer. The first region (1700-1450 cm^{-1}, Amide I and Amide II) corresponds to the same region that is most prominent in the loadings of our OPLS-DA models. The second region (1180-1000 cm^{-1}, mainly due to symmetric PO_2^- stretching, C–C symmetric vibrations, and C–O symmetric vibrations of proteins, nucleic acids, carbohydrates, and phospholipids) corresponds approximately to the same spectrum selected in our iPCA analysis.

Considering the intense spectral discriminatory band involving DNA/RNA between BC molecular subtypes and the control evidenced in the ATR-FTIR analysis (Fig. 1), this could also be explained by the higher concentration of circulating cell-free DNA (cfDNA) in breast cancer patients [49,50]. It is released into the blood plasma by apoptosis, necrosis, or active secretion. Many factors are associated with increased release of circulating tumor DNA (ctDNA) into the bloodstream, such as tumor volume, localization, vascularization, and antitumoral treatments (surgery, chemotherapies, radiotherapy) [51,52]. ctDNA can potentially carry deletions, translocations, methylations, different types of integrity that interfere with the structural pattern of DNA and that will likely be verified in the ATR-FTIR spectrum [53].

The OPLS-DA model proved to be an excellent supervised chemometric algorithm for dimensionality reduction and classification of BC molecular subtypes, considering 100% accuracy and RMSECV < 0.005 for all models with only 1 LV. The model was constructed using the total spectrum and not only the region selected in the iPCA. This was accomplished to avoid overfitting. Although the region selected in the iPCA presented significant separation between molecular subtypes and control, it is still more appropriate to consider the joint contributions of biomolecular regions. This was evidenced by the loadings of the OPLS-DA model, where the regions of Amide I and Amide II had the greatest contributions. This high contribution was also influenced by min-max normalization (0–1).

As limitations of the model, the OPLS-DA is invariable with the insertion of new external samples, having to create the model again to improve it. However, the model is extremely functional with few samples. As a solution to this problem, there is the possibility of using artificial neural networks (ANN) with the backpropagation algorithm

associated with PCA or PLS to reduce dimensionality and serve as input data. In the ANN, for each new sample inserted, the model recalculates the weights and rebuilds the model, being able to create a permanent and self-adjusting classifying model. However, ANN tend to require a greater number of samples and computational demand [54,55].

The OPLS-DA model was developed only with plasma from patients with an indication for biopsy, in this case, all patients had BIRADS ≥ 3. In this condition, the model is more adequate to differentiate molecular subtypes of patients with indication for biopsy. The inclusion of patients classified as benign tumors would be more appropriate for the development of a model for screening cancer patients. A proposal to be developed later.

5. Conclusion

The OPLS-DA model with only 1 LV (RMSECV < 0.005 for all classes) and high variance in this LV obtained 100% accuracy for discrimination of molecular subtypes and control. The selected region in iPCA (1052-1118 cm^{-1}, mainly composed of DNA/RNA vibrations) was able to significantly differentiate the molecular subtypes and control. Therefore, our study obtained exciting results towards the translation of ATR-FTIR spectroscopy to the clinic. The methodology proposed in this study is simple, fast, low cost, and reagent-free. Nonetheless, the operation flowchart still has some limitations, such as the centrifugation step to obtain blood plasma and the sample drying step to eliminate interference from the water bands. Even so, this methodology would plausibly allow molecular subtypes of BC screening and consequently improve the prognosis by instituting early treatment.

Author contributions

Nikolas Mateus Pereira de Souza: Conceptualization, Methodology, Formal Analysis, and Writing – Original draft. **Brenda Hunter Machado:** Formal analysis and Writing – Review and Editing. **Valeriano Antonio Corbellini:** Conceptualization, Methodology, Investigation, and Writing – Review and Editing. **Andreia Koche:** Resources and Writing – Review and Editing. **Lucia Beatriz Fernandes da Silva Furtado:** Resources and Writing – Review and Editing. **Débora Becker:** Resources and Writing – Review and Editing. **Alexandre Rieger:** Conceptualization and Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

The authors acknowledge the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, modality PIBITI/CNPq) for scholarship.

References

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global Cancer Statistics 2020 : GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, vol. 71, 2021, pp. 209–249, https://doi.org/10.3322/caac.21660.

[2] A. Prat, E. Pineda, B. Adamo, P. Galv, L. Gaba, M. Díez, M. Viladot, A. Fern, A. Arance, Clinical Implications of the Intrinsic Molecular Subtypes of Breast Cancer, vol. 24, 2015, pp. 26–35, https://doi.org/10.1016/j.breast.2015.07.008.

[3] K. Unger-saldaña, Challenges to the early diagnosis and treatment of breast cancer in developing countries, 5, 2014, pp. 465–478, https://doi.org/10.5306/wjco.v5.i3.465.

[4] A.E. Giuliano, S.E. Edge, G.N. Hortobagyi, Eighth edition of the AJCC cancer staging manual, Breast Cancer 25 (2018) 1783–1785, https://doi.org/10.1245/s10434-018-6486-6.

[5] H. Blumen, K. Fitch, V. Polkus, Comparison of Treatment Costs for Breast Cancer , by Tumor Stage and Type of Service, 2016, pp. 23–32.

[6] S. Broekx, E. Den Hond, R. Torfs, The Costs of Breast Cancer Prior to and Following Diagnosis, 2011, pp. 311–317, https://doi.org/10.1007/s10198-010-0237-3.

[7] S. Park, J. Seung, M. Suk, H. Seok, J. Sang, J. Seok, S. Il, B. Park, Characteristics and outcomes according to molecular subtypes of breast cancer as classified by a panel of four biomarkers using immunohistochemistry, Breast 21 (2012) 50–57, https://doi.org/10.1016/j.breast.2011.07.008.

[8] M.T. Weigel, M. Dowsett, Current and Emerging Biomarkers in Breast Cancer : Prognosis and Prediction, 2010, pp. 245–262, https://doi.org/10.1677/ERC-10-0136.

[9] S. Dawood, R. Hu, M.D. Homes, G.A. Colditz, R.M. Tamimi, Defining Breast Cancer Prognosis Based on Molecular Phenotypes : Results from a Large Cohort Study, 2011, pp. 185–192, https://doi.org/10.1007/s10549-010-1113-7.

[10] Z. Momenimovahed, H. Salehiniya, Epidemiological characteristics of and risk factors for breast cancer in the world, Breast Cancer 11 (2019) 151–164, https://doi.org/10.2147/BCTT.S176070.

[11] A.B. Mph, Breast Cancer Survival by Molecular Subtype: a Population- Based Analysis of Cancer Registry Data, 2017, https://doi.org/10.9778/cmajo.20170030.

[12] T.M. Kolb, J. Lichy, J.H. Newhouse, Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations, Radiology 225 (2002) 165–175, https://doi.org/10.1148/radiol.2251011667.

[13] L. Wang, Early diagnosis of breast cancer, Sensors (2017) 17, https://doi.org/10.3390/s17071572.

[14] B.C. Calhoun, Core needle biopsy of the breast: an evaluation of contemporary data, Surg. Pathol. Clin. 11 (2018) 1–16, https://doi.org/10.1016/j.path.2017.09.001.

[15] K. Naseer, S. Ali, J. Qazi, ATR-FTIR spectroscopy as the future of diagnostics: a systematic review of the approach using bio-fluids, Appl. Spectrosc. Rev. 56 (2021) 85–97, https://doi.org/10.1080/05704928.2020.1738453.

[16] D.L. Pavia, G.M. Lampman, G.S. Kriz, J.R. Vyvyan, Introducão à Espectroscopia – Tradução da 4ª Edição Norte-Americana, 4ª ed., Cengage Learning, Brazil, 2010.

[17] Y. Li, F. Li, X. Yang, L. Guo, F. Huang, Z. Chen, X. Chen, S. Zheng, Quantitative analysis of glycated albumin in serum based on ATR-FTIR spectrum combined with SiPLS and SVM, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 201 (2018) 249–257, https://doi.org/10.1016/j.saa.2018.05.022.

[18] C. Dna, Liquid Biopsy of Methylation Biomarkers in, Trends Mol. Med. 27 (n.d.) 482–500. https://doi.org/10.1016/j.molmed.2020.12.011.

[19] K. Pantel, C. Alix-panabi, Liquid biopsy and minimal residual disease — latest advances and implications for cure, (n.d.). https://doi.org/10.1038/s41571-019-0187-3.

[20] K.A. Brown, Metabolic pathways in obesity-related breast cancer, Nat. Rev. Endocrinol. 17 (2021) 350–363, https://doi.org/10.1038/s41574-021-00487-0.

[21] A.A. Bunaciu, V.D. Hoang, H.Y. Aboul-enein, Critical Reviews in Analytical Chemistry Applications of FT-IR Spectrophotometry in Cancer Diagnostics Applications of FT-IR Spectrophotometry in Cancer Diagnostics, 2015, p. 8347, https://doi.org/10.1080/10408347.2014.904733.

[22] J. Backhaus, R. Mueller, N. Formanski, N. Szlama, H.G. Meerpohl, M. Eidt, P. Bugert, Diagnosis of breast cancer with infrared spectroscopy from serum samples, Vib. Spectrosc. 52 (2010) 173–177, https://doi.org/10.1016/j.vibspec.2010.01.013.

[23] F. Elmi, A. Fayyaz, M. Mitra, H. Alinezhad, N. Nikbakhsh, Spectrochimica acta Part A : molecular and biomolecular spectroscopy application of FT-IR spectroscopy on breast cancer serum analysis, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 187 (2017) 87–91, https://doi.org/10.1016/j.saa.2017.06.021.

[24] V.E. Sitnikova, M.A. Kotkova, T.N. Nosenko, T.N. Kotkova, D.M. Martynova, M. V. Uspenskaya, Breast cancer detection by ATR-FTIR spectroscopy of blood serum and multivariate data-analysis, Talanta 214 (2020), 120857, https://doi.org/10.1016/j.talanta.2020.120857.

[25] I.C.C. Ferreira, E.M.G. Aguiar, A.T.F. Silva, L.L.D. Santos, L. Cardoso-Sousa, T. G. Araújo, D.W. Santos, L.R. Goulart, R. Sabino-Silva, Y.C.P. Maia, C.J. Li, Attenuated total reflection-fourier transform infrared (ATR-FTIR) spectroscopy analysis of saliva for breast cancer diagnosis, JAMA Oncol. 2020 (2020), https://doi.org/10.1155/2020/4343590.

[26] H.F. Nargis, H. Nawaz, A. Ditta, T. Mahmood, M.I. Majeed, N. Rashid, M. Muddassar, H.N. Bhatti, M. Saleem, K. Jilani, F. Bonnier, H.J. Byrne, Raman spectroscopy of blood plasma samples from breast cancer patients at different stages, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 222 (2019), 117210, https://doi.org/10.1016/j.saa.2019.117210.

[27] L.C. Lee, C.Y. Liong, A.A. Jemain, A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum, Chemometr. Intell. Lab. Syst. 163 (2017) 64–75, https://doi.org/10.1016/j.chemolab.2017.02.008.

[28] M.M.C. Ferreira, Quimiometria: conceitos, métodos e aplicações, Editora da Unicamp, 2015, ISBN 978-85-268-1471-4, p. 493, https://doi.org/10.7476/9788526814714.

[29] L.C. Lee, C.Y. Liong, A.A. Jemain, Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary

practice strategies and knowledge gaps, Analyst 143 (2018) 3526–3539, https://doi.org/10.1039/c8an00599k.
[30] C.K. Muro, K.C. Doty, L.D.S. Fernandes, I.K. Lednev, Forensic body fluid identification and differentiation by Raman spectroscopy, Forensic Chem. 1 (2016) 31–38, https://doi.org/10.1016/j.forc.2016.06.003.
[31] A. E1655-17, Standard Practices for Infrared Multivariate Quantitative Analysis, ASTM International, West Conshohocken, PA, 2017.
[32] Y. Fan, X. Zhou, T.S. Xia, Z. Chen, J. Li, Q. Liu, R.N. Alolga, Y. Chen, M. De Lai, P. Li, W. Zhu, L.W. Qi, Human plasma metabolomics for identifying differential metabolites and predicting molecular subtypes of breast cancer, Oncotarget 7 (2016) 9925–9938, https://doi.org/10.18632/oncotarget.7155.
[33] L. Díaz-Beltrán, C. González-Olmedo, N. Luque-Caro, C. Díaz, A. Martín-Blázquez, M. Fernández-Navarro, A.L. Ortega-Granados, F. Gálvez-Montosa, F. Vicente, J. P. Del Palacio, P. Sánchez-Rovira, Human plasma metabolomics for biomarker discovery: targeting the molecular subtypes in breast cancer, Cancers 13 (2021) 1–18, https://doi.org/10.3390/cancers13010147.
[34] G.I. Dovbeshko, V.I. Chegel, N.Y. Gridina, O.P. Repnytska, Y.M. Shirshov, V. P. Tryndiak, I.M. Todor, G.I. Solyanik, Surface enhanced IR absorption of nucleic acids from tumor cells: FTIR reflectance study, Biopolym. - Biospectroscopy Sect. 67 (2002) 470–486, https://doi.org/10.1002/bip.10165.
[35] G.I. Dovbeshko, N.Y. Gridina, E.B. Kruglova, O.P. Pashchuk, FTIR spectroscopy studies of nucleic acid damage, Talanta 53 (2000) 233–246, https://doi.org/10.1016/S0039-9140(00)00462-8.
[36] N. Fujioka, Y. Morimoto, T. Arai, M. Kikuchi, Discrimination between normal and malignant human gastric tissues by Fourier transform infrared spectroscopy, Cancer Detect. Prev. 28 (2004) 32–36, https://doi.org/10.1016/j.cdp.2003.11.004.
[37] 3 IDUNA FICHTNER; and HENRY H, MANTSCH'+ HEINZ FABIAN,'* MICHAEL JACKSON,' LEIGH MURPHY,* PETER H. WATSON, breast tumors and breast tumor cell xenografts, Breast 1 (1995) 37–45.
[38] Z. Movasaghi, S. Rehman, I.U. Rehman, Fourier transform infrared (FTIR) spectroscopy of biological tissues, Appl. Spectrosc. Rev. 43 (2008) 134–179.
[39] A.K. Dunbier, H. Anderson, Z. Ghazoui, J. Salter, J.S. Parker, C.M. Perou, I. E. Smith, M. Dowsett, Association between breast cancer subtypes and response to neoadjuvant anastrozole, Steroids 76 (2011) 736–740, https://doi.org/10.1016/j.steroids.2011.02.025.
[40] L. Braun, F. Mietzsch, P. Seibold, A. Schneeweiss, P. Schirmacher, J. Chang-Claude, H. Peter Sinn, S. Aulmann, Intrinsic breast cancer subtypes defined by estrogen receptor signalling - prognostic relevance of progesterone receptor loss, Mod. Pathol. 26 (2013) 1161–1171, https://doi.org/10.1038/modpathol.2013.60.
[41] M. Dowsett, M. Procter, W. McCaskill-Stevens, E. De Azambuja, U. Dafni, J. Rueschoff, B. Jordan, S. Dolci, M. Abramovitz, O. Stoss, G. Viale, R.D. Gelber, M. Piccart-Gebhart, B. Leyland-Jones, Disease-free survival according to degree of HER2 amplification for patients treated with adjuvant chemotherapy with or without 1 year of trastuzumab: the HERA trial, J. Clin. Oncol. 27 (2009) 2962–2969, https://doi.org/10.1200/JCO.2008.19.7939.
[42] J. Wang, B. Xu, Targeted therapeutic options and future perspectives for her2-positive breast cancer, Signal Transduct. Targeted Ther. 4 (2019), https://doi.org/10.1038/s41392-019-0069-2.
[43] H.J. Burstein, G. Curigliano, S. Loibl, P. Dubsky, M. Gnant, P. Poortmans, M. Colleoni, C. Denkert, M. Piccart-Gebhart, M. Regan, H.J. Senn, E.P. Winer, B. Thurlimann, Estimating the benefits of therapy for early-stage breast cancer: the St. Gallen International Consensus Guidelines for the primary therapy of early breast cancer 2019, Ann. Oncol. 30 (2019) 1541–1557, https://doi.org/10.1093/annonc/mdz235.
[44] V.E. Sitnikova, M.A. Kotkova, T.N. Nosenko, T.N. Kotkova, D.M. Martynova, M. V. Uspenskaya, Breast cancer detection by ATR-FTIR spectroscopy of blood serum and multivariate data-analysis, Talanta 214 (2020), 120857, https://doi.org/10.1016/j.talanta.2020.120857.
[45] D.C. Malins, N.L. Polissar, K. Nishikida, E.H. Holmes, H.S. Gardner, S. J. Gunselman, The etiology and prediction of breast cancer. Fourier transform-infrared spectroscopy reveals progressive alterations in breast DNA leading to a cancer-like phenotype in a high proportion of normal women, Cancer 75 (1995) 503–517, https://doi.org/10.1002/1097-0142(19950115)75:2<503::AID-CNCR2820750213>3.0.CO;2-0.
[46] D. Lin, S. Feng, J. Pan, Y. Chen, J. Lin, G. Chen, S. Xie, H. Zeng, R. Chen, Colorectal cancer detection by gold nanoparticle based surface-enhanced Raman spectroscopy of blood serum and statistical analysis, Opt Express 19 (2011), 13565, https://doi.org/10.1364/oe.19.013565.
[47] X. Wang, X. Shen, D. Sheng, X. Chen, X. Liu, FTIR spectroscopic comparison of serum from lung cancer patients and healthy persons, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 122 (2014) 193–197, https://doi.org/10.1016/j.saa.2013.11.049.
[48] U. Zelig, E. Barlev, O. Bar, I. Gross, F. Flomen, S. Mordechai, J. Kapelushnik, I. Nathan, H. Kashtan, N. Wasserberg, O. Madhala-Givon, Early detection of breast cancer using total biochemical analysis of peripheral blood components: a preliminary study, BMC Cancer 15 (2015) 1–10, https://doi.org/10.1186/s12885-015-1414-7.
[49] T. Shibayama, S.K. Low, M. Ono, T. Kobayashi, K. Kobayashi, I. Fukada, Y. Ito, T. Ueno, S. Ohno, Y. Nakamura, S. Takahashi, Clinical significance of gene mutation in ctDNA analysis for hormone receptor-positive metastatic breast cancer, Breast Cancer Res. Treat. 180 (2020) 331–341, https://doi.org/10.1007/s10549-019-05512-5.
[50] Q. Chen, Z.H. Zhang, S. Wang, J.H. Lang, Circulating cell-free DNA or circulating tumor dna in the management of ovarian and endometrial cancer, OncoTargets Ther. 12 (2019) 11517–11530, https://doi.org/10.2147/OTT.S227156.
[51] S. Volik, M. Alcaide, R.D. Morin, C. Collins, Cell-free DNA (cfDNA): clinical signi fi cance and utility in cancer shaped by, Emerg. Technol. 14 (2016) 898–909, https://doi.org/10.1158/1541-7786.MCR-16-0044.
[52] D. Fernandez-Garcia, A. Hills, K. Page, R.K. Hastings, B. Toghill, K.S. Goddard, C. Ion, O. Ogle, A.R. Boydell, K. Gleason, M. Rutherford, A. Lim, D.S. Guttery, R. C. Coombes, J.A. Shaw, Plasma cell-free DNA (cfDNA) as a predictive and prognostic marker in patients with metastatic breast cancer, Breast Cancer Res. 21 (2019) 1–13, https://doi.org/10.1186/s13058-019-1235-8.
[53] E. Tzanikou, E. Lianidou, The potential of ctDNA analysis in breast cancer, Crit. Rev. Clin. Lab Sci. 57 (2020) 54–72, https://doi.org/10.1080/10408363.2019.1670615.
[54] M. Saikat, Y. Jun, S. Maitra, J. Yan, Principle component analysis and partial least squares : two dimension reduction techniques for reegression, Casualty Actuar. Soc. (2008) 79–90.
[55] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.