# Accepted Manuscript


Multimodal chemometric approach for the analysis of human exhaled breath in lung cancer patients by TD-GC × GC-TOFMS

R. Pesesse, P.-H. Stefanuto, F. Schleich, R. Louis, J.-F. Focant

PII: S1570-0232(18)31635-0
DOI: https://doi.org/10.1016/j.jchromb.2019.01.029
Reference: CHROMB 21512

To appear in: *Journal of Chromatography B*

Received date: 31 October 2018
Revised date: 18 December 2018
Accepted date: 17 January 2019

Please cite this article as: R. Pesesse, P.-H. Stefanuto, F. Schleich, et al., Multimodal chemometric approach for the analysis of human exhaled breath in lung cancer patients by TD-GC × GC-TOFMS, Journal of Chromatography B, https://doi.org/10.1016/j.jchromb.2019.01.029

# Multimodal chemometric approach for the analysis of human exhaled breath in lung cancer patients by TD-GC×GC-TOFMS

R. Pesesse[1], P.-H. Stefanuto[1], F. Schleich[2], R. Louis[2], J.-F. Focant[1*]

[1]Organic and Biological Analytical Chemistry Group, MolSys research unit, University of Liège, B6c, Agora District, 4000 Liège, Belgium

[2]Pneumology and Allergology, GIGA Research Group, CHU of Liège, University of Liege, B35, Hospital District, Liege, Belgium

* Corresponding Author
Professor Jean-François Focant
University of Liège
Chemistry Department – CART, Organic & Biological Analytical Chemistry Group
Allée du 6 Août B6c, B-4000 Liège, Belgium
Phone: +32 (0)4 366 35 31
Fax: +32 (0)4 366 43 87
email: JF.Focant@uliege.be

**Keywords:** Breath analysis, Lung cancer, Thermal desorption (TD), Comprehensive two-dimensional gas chromatography (GC×GC), Time-of-flight mass spectrometry (TOFMS), Volatile organic compounds (VOCs), Statistics

**Abstract**

Lung cancer is the deadliest cancer in developed countries. To reduce its mortality rate, it is important to enhance our capability to detect it at earlier stages by developing early diagnostic methods. In that context, the analysis of exhaled breath is an interesting approach because of the simplicity of the medical act and its non-invasiveness. Thermal desorption comprehensive two-dimensional gas chromatography time of flight mass spectrometry (TD-GC×GC-TOFMS) has been used to characterize and compare the volatile content of human breath of lung cancer patients and healthy volunteers. On the sampling side, the contaminations induced by the bags membrane and further environmental migration of VOCs during and after the sampling have also been investigated. Over a realistic period of 6 h, the concentration of contaminants inside the bag can increase from 2 to 3 folds based on simulated breath samples. On the data processing side, Fisher ratio (FR) and random forest (RF) approaches were applied and compared in regards to their ability to reduce the data dimensionality and to extract the significant information. Both approaches allow to efficiently smooth the background signal and extract significant features (27 for FR and 17 for RF). Principal component analysis (PCA) was used to evaluate the clustering capacity of the different models. For both approaches, a separation along PC-1 was obtained with a variance score around 35%. The combined model provides a partial separation with a PC-1 score of 52%. This proof-of-concept study further confirms the potential of breath analysis for cancer detection but also underlines the importance of quality control over the full analytical procedure, including the processing of the data.

## 1. Introduction

Cancer is one of the major causes of death in Europe and the Western world [1]. Although the most prevalent cancers are prostate cancer for men and breast cancer for women [1, 2], the highest death rate is observed for lung cancer [3-7]. In 2012, more than 1.6 million people died because of lung cancer worldwide [8]. In 2018, for the United States only, lung and bronchus cancer deaths are estimated to more than 154,000 [9]. The main cause is the lack of specific symptoms of this cancer at the early stage, leading to late stage diagnosis and consequent low average 5-years survival rate (<15%) as treatment efficiency decreases with the disease progression [10-13]. If lung cancer would be detected early, using alternative diagnostic instrumentation for example, the 5-years survival rate could increase from around 60% and up to 90% [13-17]. Currently, the most common screening method is chest X-ray. Other confirmatory methods such as sputum cytology, computer assisted tomography (CT), fluorescence bronchoscopy, positron emission tomography (PET), and magnetic resonance imaging (MRI) are also used [18-24]. Even if CT helps to detect lung cancer more specifically and hence reduces by 20% its mortality [25], this screening method still suffers from significant false positive responses. Moreover, those methods are expensive, need experienced operating personnel, and expose patients and medical staff to ionizing radiation [26]. Thus, there is still a need for the development of alternative screening tests allowing the detection of lung cancer at a more curable stage [27].

Human exhaled breath contains several hundreds of volatile organic compounds (VOCs) that can be seen as a fingerprint that could possibly be used to differentiate between individuals exhibiting various health status [28]. Breath analysis has shown to be usable to highlight possible markers of specific diseases in these individuals [29-34]. Such an approach is particularly adapted to potential early diagnosis of cancer because its low level of invasiveness and relative ease of implementation on a large scale basis. The carrying out of an

early diagnostic procedure for cancer screening by means of breath analysis could contribute to increase the survival rate of diagnosed patients. Breath analysis has several advantages. It is non-invasive, it does not require experienced operating personnel to pose medical acts for the collection of samples. Furthermore, breath collection is a relatively inexpensive, rapid, painless and safe sampling process [19, 33, 35-37]. However, breath analysis has also some drawbacks. The very low concentration (from nanomolar to picomolar) of volatile organic compounds (VOCs) requires pre-concentration steps (e.g., solid phase microextraction (SPME), thermal desorption (TD), purge and trap) to ensure proper analysis of such diluted air samples [19, 31, 35, 38]. This is true for mixed expiratory air sampling where total breath including dead space air is sampled, resulting in dilution of endogenous VOCs, but also true for alveolar air sampling where endogenous VOC concentrations are only 2-3 times higher [39].

Even when compounds are isolated and reported in the literature, results are rarely reproducible and display high dispersion between studies, leading to low reliability of the approach. The origin of this issue is partly due to the fact that the essential chemical information is hidden under massive amounts of irrelevant signals that make the isolation of putative markers of disease from breath a real analytical challenge. Indeed, such irrelevant signals are made of significant amounts of endogenous VOCs issued from the basic metabolism of the individual and exogenous VOCs related to factors such as food habits, hygiene, tobacco consumption, and ambient air [18, 32, 33]. Moreover, other contaminants might also come from the materials used during the sampling. Commonly used Tedlar® bags are suspected to generate cross contaminations, leaching, and leaking [40]. Phenol and N-N-dimethyl acetamide are commonly cited in the literature to be the main compounds released from those bags [35, 36, 41]. The concentration of the VOCs trapped inside the bags also decreases over the time due to the permeability of the membrane [42]. In addition, the lack of

standardization and normalization are the main limitation and ongoing challenges of breath analysis [10, 18, 19].

To resolve such mixtures of VOCs, the breath content is typically analyzed using gas chromatography coupled to mass spectrometry (GC-MS) [18, 31, 43, 44]. Based on such GC-MS approaches, a limited number of VOCs has been tentatively identified as part of a volatile lung cancer profile [19, 31, 43, 45-50]. However, because of the limited peak capacity and sensitivity of GC-MS, but also because of the lack of robustness, quality control, and validation of sampling and analysis, the approach has not yet found its way to clinical application [51]. Peak capacity and sensitivity can be enhanced by using comprehensive two-dimensional gas chromatography coupled to time-of-flight MS (GC×GC-TOFMS), a known efficient separation technique for complex sample analysis [51-53]. Basically, GC×GC relies on the use of two different GC phases connected in series via a modulator [54] that, when using cryogens to operate, not only enhances the separation power but also provides better limits of detection by cryogenic zone compression of chromatographic peaks [55]. When coupled to full scan high acquisition speed TOFMS, GC×GC peaks (<200 ms of peak width at half height) are accurately described and further deconvoluted in the spectral domain if required [56, 57]. Broad dynamic range while allowing mass spectral deconvolution. Proof of concept early reports have shown the superiority of GC×GC-TOFMS over GC-MS for the separation and identification of VOCs in breath analysis [32, 40, 53, 58-61].

With the aim of further supporting the use of GC×GC-TOFMS for breath analysis for lung cancer screening, we developed and optimized a TD-GC×GC-TOFMS method based on mixed expiratory air sampling. We also studied the importance of control and reliability during the sampling of the breath. On the processing side, we investigated a multimodal data treatment approach on data sets resulting from the analyses of 29 individuals (15 lung cancer patient and from 14 healthy volunteers).

## 2. Experimental

### *2.1. Patient information*

A total of 29 individuals including 15 lung cancer patients and 14 healthy volunteers were included in this study. All subjects were at least 18 years old and they all signed an informed consent to participate at this study after being informed about its goals. The study was approved by the Ethics Committee of the University of Liège (BECT B707201420493) and conducted in conformity with the Declaration of Helsinki. Patients with abnormal chest X-rays who were scheduled for bronchoscopy, in addition to age and sex matching controls, were sampled using Tedlar® bags of 5 L at the pneumology unit of the university hospital of Liège, in Belgium, between January and May 2014 in a series of three sampling campaigns (January, March and May). None of the patients had received any form of anticancer therapy or medication before the sampling. Prior to the sampling, Tedlar® bags were flushed twice with nitrogen (purity >99.99%, Air liquid, BE) to decrease residual contaminants. Characteristics of the study population are reported in Table 1. Exhaled breath were transferred from the bag to a sorbent tube containing Tenax GR and Carbopack B (Markes International Ltd, UK) with a pump at a flow rate of 300 mL/min directly after the sampling to avoid alteration of the samples [62, 63].

### *2.2. Tedlar® bags permeability testing*

12 Tedlar® bags of 1 L were filled with high purity nitrogen. All Bags were placed in a box with a saturated atmosphere of toluene, methanol, hexane, and dichloromethane. Every 2 hours, three bags were pull out the box and the content was transferred onto thermal desorption tube following the same protocol than for exhaled human breath analysis.

*2.3. Analytical instrumentation and parameters*

Thermal desorption tubes were stored at room temperature (20 °C) before being desorbed onto a Unity 2 series thermal desorber (Markes International Ltd.) coupled to a Pegasus 4D (LECO, Corp., St. Joseph, MI). The modulator was mounted in an Agilent 7890A gas chromatograph equipped with a secondary oven and a quad-jet dual stage modulator working with liquid nitrogen as cryofluid [64]. Details regarding the system have been reported elsewhere [65]. The column set used was a combination of a Rxi-5Sil (30 m × 0.25 mm i.d x 0.25 μm df) (Restek Corp., Bellefonte, PA, USA) in the first dimension (1D) and a BPX- 50 (1.2 m × 0.10 mm i.d x 0.10 μm df) (SGE, Austin, TX, USA) in the second dimension (2D). This column combination is classic but offers several advantages for non-targeted screening (e.g. structured separation). The use of this classic combination is also useful for study to study comparison since, it is the most common used combination. This column set was already successfully used in previous VOC mixtures untargeted analysis [63, 66, 67]. During the thermal desorption, samples were first purged with dry nitrogen during 1 min to remove water. Then, tubes were heated at 300 °C during 5 min and VOCs samples were recollected on the general purposed cold trap (Tenax TA/Carbograph 1TD sorbent bed) at -10 °C. Samples were injected in the system by heating of the cold trap at 300 °C during 3 min. Helium was used as carrier gas with a constant flow rate of 1 mL/min. The main oven had an initial temperature of 35 °C during 5 min and then increased until 240°C at a rate of 5 °C/min. The temperature offset for the secondary oven was 5 °C above the main oven. The modulation period ($P_M$) was 4 s with a hot pulse duration set at 700 ms and a cooling time between stages of 1300 ms. The modulator temperature offset was 10 °C above the temperature of the

secondary GC oven. 70 eV electron ionization was used. The data acquisition rate was set at a frequency of 100 Hz for a mass range from 29 to 450 m/z. Tuning and mass calibration were performed daily with perfluorotributylamine (PFTBA).

*2.4. Chromatographic alignment and feature identification*

Data were acquired and processed with the LECO ChromaTOF® 4.5 software (LECO Corp.). Peak finding, mass deconvolution, integration peak and library searching were performed by this software. Mass spectral identification used Wiley (2011) and NIST (2014) databases with a match factor threshold >800. Statistical compare option of ChromaTOF® 4.5 software was used to align 2D chromatograms and built a peak table which contains every peak found in each samples with a signal to noise ratio of 100 [68, 69]. Peak tables created were extracted in .csv files for further data process.

*2.5. Multivariate analysis*

Multivariate statistical analyses were conducted using R 3.4.3 using the Rstudio interface (Free Software Foundation's GNU project). All the packages are provided in supplementary information. First, all the data were normalized using probabilistic quotient normalization (PQN) and log transformed [70]. For specific features detection two approached were compared. In an univariate approach, Fisher ratio (FR) calculation was performed in order to identify specific compounds differentiating between the two groups [71, 72]. The compounds with a FR value above the critical F value (Fcrit) were considered as significant. In a multivariate approach, Random Forest algorithm and variable importance ranking were used to select the significant features [21]. The resulting data clustering and classification efficiency was visualized using principal component analysis (PCA).

## 3. Results and discussion

*3.1. Sample integrity investigation during storage in Tedlar® bag*

It has been reported that the concentration of compounds trapped inside Tedlar® bags decreases over time, underlying limitations of the permeability of the membrane and limited storage potential [73]. However, as far as we know, no studies have yet investigated the resistance of the bag membrane against outside environmental contaminations. To evaluate this effect, 12 Tedlar® bags filled with nitrogen were placed in a box in which the atmosphere was saturated in toluene, methanol, hexane, and dichloromethane. Every two hours, three bags were pulled out of the box and deflated on TD tubes to be analyzed. The kinetic study illustrated in Figure 1 shows how relative intensities of solvents peaks increase as a function of time. This time-trend study shows that the relative intensity of each solvent inside the bags increased according to the exposure time. The membrane of Tedlar® bag is thus also prone to permeation of chemicals from the environment to the bag. It can be concluded that the residency time of the sampled breath inside the bags should be kept to a minimum to ensure low impact of the sampling procedure on sample integrity. Furthermore, storage conditions should carefully be described in studies using such bags.

*3.2. Influence of exogenous VOCs during the sampling process*

The TD-GC×GC-TOFMS analysis of the exhaled breath samples of 29 lung cancer patients and healthy volunteers, conducted to the detection of an average of 1,078 features for each chromatogram. After chromatographic alignment of all samples, the composite peak table contained a total of 1,350 robust features. Features screening for chromatographic artifacts, multiple peak identifications, and columns bleeding allowed the reduction of the data set to 1,019 features. A non-supervised PCA was performed based on this data set and the resulting plot can be seen on Figure 2. The visualization of such an unsupervised processing showed a clustering of the data according three apparent batches, each of which

appeared to be related to a sampling period (January, March and April) independently of the nature of the samples (patients and controls). This phenomenon demonstrated that the influence of the presence of various levels of background of exogenous VOCs during the sampling was higher than any possible differences related to the health status of the sampled patients, despite the fact that all samples were taken in the same room at the same hospital, with the same method, by the same operating staff.

Different approaches were investigated to smooth the environmental effect on the background. A possible approach to reduce the impact of the presence of background exogenous VOCs is to perform more complex alveolar air sampling [74]. Having patients breathe medical air for lung washout is another option but it is time consuming [9] and our own testing in that direction was not conclusive. As shown on the unsupervised dendrogram displayed in Fig SI-1, the samples taken from three individuals in three different locations with or without medical air washout, does not display any particular clustering. This demonstrates that the washout was not able to remove the environment background and do not represent a way to go for this study.

In a second time, data pre-processing and batch effect correction was implemented to reduce the impact of exogenous VOCs. Each batch was individually mean-centered in order to smooth the impact of the sampling dates (Figure 2 bottom). This step was possible due to the parallel sampling between patients and controls. This means that the correction was affecting the two classes in the exact same way and it doesn't generate any overfitting. The corrected data allow to more efficiently extract putative biomarkers from the initial raw data. Moreover, it maintains a sampling, which involves spontaneously breathing subjects. Following this pre-processing step, two different data processing approaches were investigated: 1) a univariate feature selection tool based on Fisher Ratio calculation; 2) a multivariate approach using Random Forest algorithms and feature importance ranking.

*3.3. Univariate feature selection*

The use of Fisher ratio (FR) to select features of interest from biological data sets is widely spread among GC×GC non-target studies [62, 71]. Due to the large amount of data generated, the supervised FR approach was used to decrease the data dimensionality and highlight possible chemical differences between the two classes of samples (lung cancer patients versus healthy volunteers) by extracting portions of data where class-to-class variations were greater than within-class variations. Furthermore, we applied a critical FR cutoff defined for a 1% significance level, in order to even further reduce data dimensionality [75]. From the 1,019 features of the cleaned data set, this 1% significance level FR approach permitted to extract a list of 27 features (Figure 3). Based on the remaining 27 features, PCA and clustering analysis (Figure 4) resulted on a clear separation between lung cancer patients and healthy volunteers, without any remaining influence of the sampling period. This demonstrates the efficiency of such univariate feature selection approach for the extraction of biologically relevant information.

*3.4. Multivariate feature selection*

The second statistical approach for feature selection was based on the use of Random Forest algorithm (RF), a multivariate machine learning approach build on a decision tree approach. Hence, multiple decision trees were created and merged together to obtain a more accurate and stable prediction. After the construction of the classification trees, the variables were ranked according to their importance and effect on the classification accuracy. The significant features were selected based on this ranking and a cut-off value of 0.1 was set in

mean decrease accuracy. This resulted in the selection of a set of 17 significant features for the separation of the two populations. Like for the univariate FR approach, this RF demonstrates the potential of this multivariate feature selection approach to properly cluster the two populations in the PCA space based on these 17 features (Figure 5).

*3.5. Comparison of the feature selection approaches*

The main difference between Random Forest and Fisher Ratio for feature selection is the multivariate dimension. Indeed, the combination of decision tree allows performing classification based on combined information from different features. Moreover, random forest allows obtaining direct classification performance information in addition to the feature selection possibilities. Both uni- and multivariate methods allowed to reduce the impact of the presence of variable amount of exogenous VOCs related to time of sampling to a level that did not significantly interfere anymore with the extraction of biologically relevant VOC signatures.

The 37 features selected by the two approaches were sorted according to their chemical family (i.e., alcohol, aldehyde, fatty acid methyl ester (FAME), hydrocarbon, ketone, nitrogen containing compounds). From these chemical families, average intensities were calculated (Figure 6) and the ratio Patients-intensity on Controls-intensity were evaluated. For FR features, the highest ratios were obtained for FAME and ketone compounds. The overexpression of these compounds in lung patient samples could be explained by inflammation processes inside the lungs. The same family classification process was applied to the 17 features highlighted by the random forest approach. Interestingly, the two major ratios were also coming from the ketone and the FAME. This observation could indicate that the major processes involved in the production of VOCs in the lung of cancer patient could be linked to FAME and ketone. These chemical families were already identified in previous

studies and they are supposed to come from oxidative stress reactions. This family based approach in non-targeted screening provides insights regarding the general trends of the samples, which is already informative. Indeed, for non-targeted studies, the full identification of thousands of features is practically impossible, which made the study-to-study comparison highly complicated. However, if a group of compounds are found to be specific in different studies, it could orient the future research in a predefined group of molecules.

From the lists of 27 and 17 features, a set of 7 features was common to both approaches (Table S1) and would be considered as the most representative markers of differentiation between the two classes. As illustrated in Figure 7, a PCA based on these 7 markers results in a separation trend between the two classes but the clustering is not as clear as when either FR or RF models are applied separately. It can however be noticed that the percentage of explained variance (67%) is higher than in both separated models. Even if this PCA is not providing extra information, it demonstrates the importance of the technique used for model building and the interest of applying different models in order to validate the data processing approach.

Only four of the reported features (both models included) where common to compounds previously reported as a lung cancer human breath biomarker in the literature : Cyclopentane, methyl- [76, 77], 2,5-Cyclohexadiene-1,4-dione, 2,6-bis(1,1-dimethylethyl)- [47], Hexadecane [50], and eicosane [78]. However, further discussion on compounds identification will require identity validation with high resolution detector and standard injections.

## 4. Conclusion

This work demonstrates the capacity of exhaled breath to discriminate between cancer patients and healthy individuals. Even using a straightforward sampling method, using

sampling (Tedlar®) bags, the combination of a powerful analytical strategy and a robust statistical approach provides good classification performances, overpassing limitations such as environmental contamination and sampling variability. It was possible to extract the significant information on breath VOC content. The membrane permeability of the sampling bags was shown to permit migration of VOCs from the environment to the inside of the bag, altering the sample integrity. Moreover, the background suppression effect of medical air washout did not provide any reduction of the environmental contamination. Nevertheless, it was demonstrated that the influence of the exogenous VOCs could be corrected by using a proper pre-processing step. Finally, two different data analysis strategies (FR and RF) were applied to extract significant features. With Both methods, a total of 37 features were detected and allow distinguishing the two populations of individuals. Among them, seven features were detected by both statistical approaches and allowed to separate the two populations. This study underlines the need for analytical controls from sampling to data processing in untargeted biological studies. Moreover, the utilization of chemical family profiling could represent an alternative to full identification of large data matrix. Confidence in compound identifications can be enhanced using high resolution MS and individual standards but was out of the scope of the study. The biological interest of the molecules highlighted using our approach has to be confirmed by larger cohort studies.

**Acknowledgments**

The university hospital of Liège (CHU) for providing samples; Fund for Industry and Agricultural Research (F.R.I.A) for financial support of Ph.D. We would like to thank MARKES® and Supelco Sigma-Aldrich® for their support by providing us with GC columns and various GC consumables. We also would like to thank LECO® for providing us technical support.

**References**

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. CA Cancer J. Clin. 67 (2017) 7-30

[2] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int. J. Cancer 136 (2015) E359-386.

[3] Siegel R, Naishadham, D., Jemal, A.,. Cancer statistics. CA Cancer J. Clin. 63 (2013) 11-30.

[4] Liu H, Li C, Wang H, Huang Z, Zhang P, Pan Z, et al. Characterization of Volatile Organic Metabolites in Lung Cancer Pleural Effusions by SPME–GC/MS Combined with an Untargeted Metabolomic Method. Chromatographia 77 (2014) 1379-1386.

[5] Dent AG, Sutedja TG, Zimmerman PV. Exhaled breath analysis for lung cancer. J. Thorac. Dis. 5 (2013) 540-550.

[6] Torre LA, Siegel RL, Jemal A. Lung Cancer Statistics. Advances in experimental medicine and biology 893 (2016) 1-19.

[7] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J. Clin. 66 (2016)7-30

[8] Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J. Clin. 65 (2015) 87-108.

Risbya TH, Sehnerta SS. Clinical applicati of breath biomarkers of oxidative stress status. Free radical biology and Medicine 27 (2000) 1182-1192

[9] Risbya TH, Sehnerta SS. Clinical applicati of breath biomarkers of oxidative stress status. Free radical biology and Medicine 27 (2000) 1182-1192

[10] Boedeker E, Friedel G, Walles T. Sniffer dogs as part of a bimodal bionic research approach to develop a lung cancer screening. Interactive CardioVascular and Thoracic Surgery 14 (2012) 511-515

[11] Huo D, Xu Y, Hou C, Yang M, Fa H. A novel optical chemical sensor based AuNR-MTPP and dyes for lung cancer biomarkers in exhaled breath identification. Sensors and Actuators B 199 (2014) 446-456.

[12] Schmekel B, Winquist F, Vikstrom A. Analysis of breath samples for lung cancer survival. Analytica chimica acta 840 (2014) 82-86.

[13] Pastorino U. Early detection of lung cancer. Respiration 73 (2006) 5-13.

[14] Flehinger BJ KM, Melamed MR. Survival from early lung cancer: implications for screening. Chest 101 (1992) 113-118.

[15] Shah R SS, Richardson J, Means AJ, Goulden C. Results of surgical treatment of stage I and II lung cancer. J. Cardiovasc. Surg. 37 (1996) 169-172.

[16] Sobue T, Suzuki R, Matsuda M, Kuroishi T, Ikedi S, Naruke T. Survival for clinical stage I lung cancer not surgically treated. Cancer biomarkers : section A of Disease markers 69 (1992) 685-692.

[17] Patz EF, Rossi S, Harpole DH, Herndon JE, Goodman PC. Correlation of tumor size and survival in patients with stage Ia non-small cell lung cancer. Chest (2000) 1568–1571

[18] Song G, Qin T, Liu H, Xu GB, Pan YY, Xiong FX, et al. Quantitative breath analysis of volatile organic compounds of lung cancer patients. Lung cancer 67 (2010) 227-231.

[19] Hakim M, Broza YY, Barash O, Peled N, Phillips M, Amann A, et al. Volatile organic compounds of lung cancer and possible biochemical pathways. Chem. Rev. 112 (2012) 5949-5966 [20] Filipiak W, Filipiak A, Sponring A, Schmid T, Zelger B, Ager C, et al. Comparative analyses of volatile organic compounds (VOCs) from patients, tumors and transformed cell lines for the validation of lung cancer-derived breath markers. J. Breath Res. 8 (2014).

[21] Purcaro G, Stefanuto PH, Franchina F, Beccaria M, Wieland-Alter WF, Wright PF, et al. SPME-GCxGC-TOF Ms fingerprint of virally-infected cell culture: Sample preparation optimization and data processing evaluation. Anal. Chem. 1027 (2018) 158-167

[22] Gohagan J, Marcus P, Fagerstrom R, Pinsky P, Kramer B, Prorok P, et al. Baseline findings of a randomized feasibility trial of lung cancer screening with spiral CT scan vs chest radiograph: the Lung Screening Study of the National Cancer Institute. Chest 126 (2004) 114-121.

[23] Donna E. Maziak; Gail E. Darling RIIKYG, ; Albert A. Driedger, ; Yee C. Un, ; John D. Miller; Chu-Shu Gu; Kathryn J. Cline; William K. Evans; and Mark N. Levine. Positron Emission Tomography in Staging Early Lung Cancer A Randomized Trial. Ann. Intern. Med. 151 (2009) 221-228.

[24] Silvestri GA, Gould MK, Margolis ML, Tanoue LT, McCrory D, Toloza E, et al. Noninvasive staging of non-small cell lung cancer: ACCP evidenced-based clinical practice guidelines (2nd edition). Chest 132 (2007) 178S-201S.

[25] Jett J. Screening for lung cancer: Who should be screened? Arch. Pathol. Lab. Med. 136 (2012) 1511-1514.

[26] Rebecca Smith-Bindman DLM, Eric Johnson, Choonsik Lee, Heather Spencer Feigelson, Michael Flynn, Robert T. Greenlee, Randell L. Kruger, Mark C. Hornbrook, Douglas Roblin, Leif I. Solberg, Nicholas Vanneman, Sheila Weinmann, Andrew E. Williams,. Use of Diagnostic Imaging Studies and Associated Radiation Exposure for Patients Enrolled in Large Integrated Health Care Systems, 1996-2010. Jama 307 (2012) 2400-2409.

[27] Gasparri R, Romano R, Sedda G, Borri A, Petrella F, Galetta D, et al. Diagnostic biomarkers for lung cancer prevention. J. Breath Res. 12 (2018) 027111.

[28] van Oort PM, Povoa P, Schnabel R, Dark P, Artigas A, Bergmans D, et al. The potential role of exhaled breath analysis in the diagnostic process of pneumonia-a systematic review. J. Breath Res. 12 (2018) 024001.

[29] Horvath I, Lazar Z, Gyulai N, Kollai M, Losonczy G. Exhaled biomarkers in lung cancer. European respiratory journal 34 (2009) 261-275.

[30] L Linus Pauling ABR, Roy Teranishi, and Paul Cary. Quantitative Analysis of Urine Vapor and Breath by Gas-Liquid Partition Chromatography. Proc. Nat. Acad. Sci. 68 (1971) 2374-2376.

[31] Bajtarevic A, Ager C, Pienz M, Klieber M, Schwarz K, Ligor M, et al. Noninvasive detection of lung cancer by analysis of exhaled breath. BMC Cancer 9 (2009) 348.

[32] Juan M. Sanchez aRDS. Development of a Multibed Sorption Trap, Comprehensive Two-Dimensional Gas Chromatography, and Time-of-Flight Mass Spectrometry System for the Analysis of Volatile Organic Compounds in Human Breath. Anal. Chem. 78 (2006) 3046-3054.

[33] Shirasu M, Touhara K. The scent of disease: volatile organic compounds of the human body related to disease and disorder. J. Biochem. 150 (2011) 257-266.

[34] Raed A Dweik aAA. Exhaled breath analysis: the new frontier in medical testing. J Breath Res. 2 (2008);2.

[35] Buszewski B, Kesy M, Ligor T, Amann A. Human exhaled air analytics: biomarkers of diseases. Biomed. Chromatogr. 21 (2007) 553-566.

[36] Kusano M, Mendez E, Furton KG. Development of headspace SPME method for analysis of volatile organic compounds present in human biological specimens. Anal. Bioanal. Chem. 400 (2011) 1817-1826.

[37] Kim KH, Jahan SA, Kabir E. A review of breath analysis for diagnosis of human health. Trends in Analytical Chemistry 33 (2012) 1-8.

[38] Ma W, Gao P, Fan J, Hashi Y, Chen Z. Determination of breath gas composition of lung cancer patients using gas chromatography/mass spectrometry with monolithic material sorptive extraction. Biomed. chromatogr. 29 (2015) 961-965.

[39] Miekisch W, Schubert JK, Noeldge-Schomburg GF. Diagnostic potential of breath analysis--focus on volatile organic compounds. Clinica chimica acta; international journal of clinical chemistry 347 (2004) 25-39.

[40] Ma H, Li X, Chen J, Wang H, Cheng T, Chen K, et al. Analysis of human breath samples of lung cancer patients and healthy controls with solid-phase microextraction (SPME) and flow-modulated comprehensive two-dimensional gas chromatography (GC × GC). Anal. Methods 6 (2014) 6841-6849.

[41] Paschke KM, Mashir A, Dweik RA. Clinical applications of breath testing. F1000 Medicine Reports 2 (2010) 56.

[42] Szylak-Szydlowski M. Odour Samples Degradation During Detention in Tedlar Bags. Water, air, and soil pollution (2015) 226-227.

[43] Fuchs P, Loeseken C, Schubert JK, Miekisch W. Breath gas aldehydes as biomarkers of lung cancer. Int. J. Cancer 126 (2010) 2663-2670

[44] Poli D, Goldoni M, Corradi M, Acampa O, Carbognani P, Internullo E, et al. Determination of aldehydes in exhaled breath of patients with lung cancer by means of on-fiber-derivatisation SPME-GC/MS. J. Chromatogr. B 878 (2010) 2643-2651.

[45] Chan HP, Lewis C, Thomas PS. Exhaled breath analysis: novel approach for early detection of lung cancer. Lung cancer 63 (2009) 164-168..

[46] McCulloch M, Jezierski T, Broffman M, Hubbard A, Turner K, Janecki T. Diagnostic accuracy of canine scent detection in early- and late-stage lung and breast cancers. Integrative cancer therapies 5 (2006) 30-39.

[47] Phillips M AN, Austin JH, Cameron RB, Cataneo RN, Greenberg J, Kloss R, Maxfield RA, Munawar MI, Pass HI, Rashid A, Rom WN, Schmitt P. Prediction of lung cancer using volatile biomarkers in breath. Cancer biomarkers 3 (2007) 95-109

[48] Fu X-A, Li M, Knipp RJ, Nantz MH, Bousamra M. Noninvasive detection of lung cancer using exhaled breath. Cancer medicine 3 (2014) 174-181.

[49] Rudnicka J, Walczak M, Kowalkowski T, Jezierski T, Buszewski B. Determination of volatile organic compounds as potential markers of lung cancer by gas chromatography–mass spectrometry versus trained dogs. Sensors and Actuators B 202 (2014) 615-621.

[50] Wang C, Dong R, Wang X, Lian A, Chi C, Ke C, et al. Exhaled volatile organic compounds as lung cancer biomarkers during one-lung ventilation. Scientific reports 4(2014) 7312.

[51] Beccaria M, Mellors TR, Petion JS, Rees CA, Nasir M, Systrom HK, et al. Preliminary investigation of human exhaled breath for tuberculosis diagnosis by multidimensional gas chromatography - Time of flight mass spectrometry and machine learning. J. Chromatogr. B 1074-1075 (2018) 46-50.

[52] Dimandja J-MD, Clouden GC, Colón I, Focant J-F, Cabey WV, Parry RC. Standardized test mixture for the characterization of comprehensive two-dimensional gas chromatography columns: the Phillips mix. J. Chromatogr. A 1019 (2003) 261-272.

[53] Libardoni M, Stevens PT, Waite JH, Sacks R. Analysis of human breath samples with a multi-bed sorption trap and comprehensive two-dimensional gas chromatography (GCxGC). J. Chromatogr. B 842 (2006) 13-21.

[54] J.F. Focant ASdaDGP, Jr., In: W. Niessen Encyclopedia of Mass Spectrometry. The Netherlands: Elsevier; (2006).

[55] Patterson DG, Jr., Welch SM, Turner WE, Sjodin A, Focant JF. Cryogenic zone compression for the measurement of dioxins in human serum by isotope dilution at the

attogram level using modulated gas chromatography coupled to high resolution magnetic sector mass spectrometry. J. Chromatogr. A 1218 (2011) 3274-3281.

[56] Bings NH, Costa-Fern´andez JeM, Jr. JPG, Leach AM, HieftjeU GM. Time-of-flight mass spectrometry as a tool for speciation analysis. Spectrochimica Acta Part B 55 (2000) 767-778..

[57] Bristow T. Evolution and Revolution In Time-Of-Flight Mass Spectrometry And Its Impact On Research Withon The Parmaceutical Industry. European Pharmaceutical Review 16 (2011) 13-15..

[58] Michael Phillips M, Urvish Patel, Renee N. Cataneo, Xiang Zhang. Detection of an Extended Human Volatome with Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry. PLoS One 8 (2013) e75274.

[59] Maren Mieth JKS, Thomas Groger, Bastian Sabel, Sabine Kischkel,, Patricia Fuchs DH, Ralf Zimmermann, and Wolfram Miekisch. Automated Needle Trap Heart-Cut GC/MS and Needle Trap Comprehensive Two-Dimensional GC/TOF-MS for Breath Gas Analysis in the Clinical Environment. Anal. Chem. 82 (2010) 2541–2551.

[60] Koek MM, van der Kloet FM, Kleemann R, Kooistra T, Verheij ER, Hankemeier T. Semi-automated non-target processing in GC x GC-MS metabolomics analysis: applicability for biomedical studies. Metabolomics 7 (2011) 1-14.

[61] Das MK, Bishwal SC, Das A, Dabral D, Varshney A, Badireddy VK, et al. Investigation of gender-specific exhaled breath volatome in humans by GCxGC-TOF-MS. Anal. Chem. 86 (2014) 1229-1237.

[62] Stefanuto PH, Perrault KA, Stadler S, Pesesse R, LeBlanc HN, Forbes SL, et al. GC x GC-TOFMS and supervised multivariate approaches to study human cadaveric decomposition olfactive signatures. Anal. Bioanal. Chem. (2015) 4767-4778.

[63] Stadler S, Stefanuto PH, Brokl M, Forbes SL, Focant JF. Characterization of volatile organic compounds from human analogue decomposition using thermal desorption coupled to comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry. Anal. Chem. 85 (2013) 998-1005.

[64] Focant J, Sjödin A, Turner W, Don Patterson J. Measurement of Selected Halogenated Contaminants in Human Serum and Milk using GCxGC-IDTOFMS. ORGANOHALOGEN COMPOUNDS 66 (2004) 804-811.

[65] Dimandja J-MD. Comprehensive 2-D GC provides high-performance separations in terms of selectivity, sensitivity, speed, and structure. Anal. Chem. (2004) 167-174.

[66] Dekeirsschieter J, Stefanuto PH, Brasseur C, Haubruge E, Focant JF. Enhanced characterization of the smell of death by comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry (GCxGC-TOFMS). PLoS One 7 (2012) e39005.

[67] Brasseur C, Dekeirsschieter J, Schotsmans EM, de Koning S, Wilson AS, Haubruge E, et al. Comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry for the forensic study of cadaveric volatile organic compounds released in soil by buried decaying pig carcasses. J. Chromatogr. A 1255 (2012) 163-170..

[68] Perrault KA, Stefanuto PH, Stuart BH, Rai T, Focant JF, Forbes SL. Detection of decomposition volatile organic compounds in soil following removal of remains from a surface deposition site. Forensic Sci. Med. Pathol. 11 (2015) 376-387.

[69] Forbes SL, Perrault KA, Stefanuto PH, Nizio KD, Focant JF. Comparison of the decomposition VOC profile during winter and summer in a moist, mid-latitude (Cfb) climate. PLoS One 9 (2014) e113681.

[70] Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. Anal. Chem. 78 (2006) 4281-4290.

[71] Karisa M. Pierce JCH, Janiece L. Hope, Petrie M. Rainey, Andrew N. Hoofnagle, Rhona M. Jack, Bob W. Wright, and Robert E. Synovec. Fisher Ratio Method Applied to Third-Order Separation Data To Identify Significant Chemical Components of Metabolite Extracts. Anal. Chem. 78 (2006) 5068-5075.

[72] Heim J. Using the Statistical Compare and Fisher Ratio ChromaTOF Features to Define Variance Prior to Multivariate Analysis in the Small Metabolite Profile of Diabetic Versus Non-Diabetic Urine by GCxGC-TOFMS.

[73] Mochalski P, King J, Unterkofler K, Amann A. Stability of selected volatile breath constituents in Tedlar, Kynar and Flexfilm sampling bags. The Analyst 138 (2013)1405-1418.

[74] Schubert JK, Spittler K-H, Braun G, Geiger K, GuttmannJ. CO2-controlled sampling of alveolar gas in mechanically ventilated patients. J Appl Physiol 90 (2001) 486–492.

[75] P-H S, Perrault KA, Dubois L, L'Homme B, Loughnane AC, Ochiai, et al. Advanced method optimization for volatile aroma profiling of beer using two-dimensioanal gas chromatography time-of-flight mass spectrometry. Journal of chromatography A 1507 (2017) 45-52.

[76] Chen X, Xu F, Wang Y, Pan Y, Lu D, Wang P, et al. A study of the volatile organic compounds exhaled by lung cancer cells in vitro for breath diagnosis. CANCER 110 (2007) 835-844.

[77] Phillips M, Gleeson K, Hughes JMB, Greenberg J, Cataneo RN, Baker L, et al. Volatile organic compounds in breath as markers of lung cancer: a cross-sectional study. The Lancet 353 (1999) 1930-1933.

[78] Wang Y, Hu Y, Wang D, Yu K, LingWang, Zoa Y, et al. The analysis of volatile organic compounds biomarkers for lung cancer in exhaled breath, tissues and cell lines. Cancer Biomarkers 11 (2012) 129-137.

**Table 1:** Characteristics of study subjects

| | Lung cancer patients | Healthy volunteers |
|---|---|---|
| Number (%) | 15 (52%) | 14 (48%) |
| Age (year), mean (± SD) | 62 ± 7 | 58 ± 11 |
| Gender (M/F) | 12/3 | 9/5 |
| Smoker/ Ex-smoker/ Non-smoker | 3/12/0 | 1/6/7 |
| Sampling period (January/March/May) | 8/4/3 | 8/3/3 |

**Figure legends**

**Figure 1**. Kinetic study of the permeability of Tedlar® bag membrane to different solvents.

**Figure 2.** ***Top:*** Non-supervised PCA score plot with all features detected in the exhaled breath of 29 individuals (15 lung cancer patients in red and 14 healthy volunteers in blue). ***Bottom****:* Non-supervised PCA score plot after individual mean-centering according to the batch. Both PCAs were performed on the 1019 features from post pre-processing.

**Figure 3.** Scheme of the general workflow used for the data treatment and feature selections.

**Figure 4.** PCA score plot based on the 27 features extracted by the 1% significance level Fisher Ratio approach. The PCA is based on the first two PCs, displaying 44.24% of the total variance.

**Figure 5.** PCA score plot based on the 17 features extracted by the Random Forest approach. The PCA is based on the first two PCs, displaying 45.87% of the total variance.

**Figure 6.** Relative chemical family contributions for the selected features isolated fot patients and control using both methods (left), Fisher Ratio only (center), Random forest only (right). Chemical families: Alcohol (light blue), Aldehyde (orange), FAME (grey), Hydrocarbon (yellow), Ketone (dark blue), Nitrogen containing compounds (green).

**Figure 7.** PCA score plot of the healthy volunteers (negative) and lung cancer patients (positive) by using the six features highlight with the Random forest and Fisher Ratio approaches. The PCA is based on the first two PCs, displaying 67.49% of the total variance.

**Highlights:**

- Fisher ratio and random forest statistical succeed to extract putative markers from high dimensional data sets
- Multimodal statistical approaches allow to identify independent marker-compounds
- Sample integrity depends on residency time in sampling bag

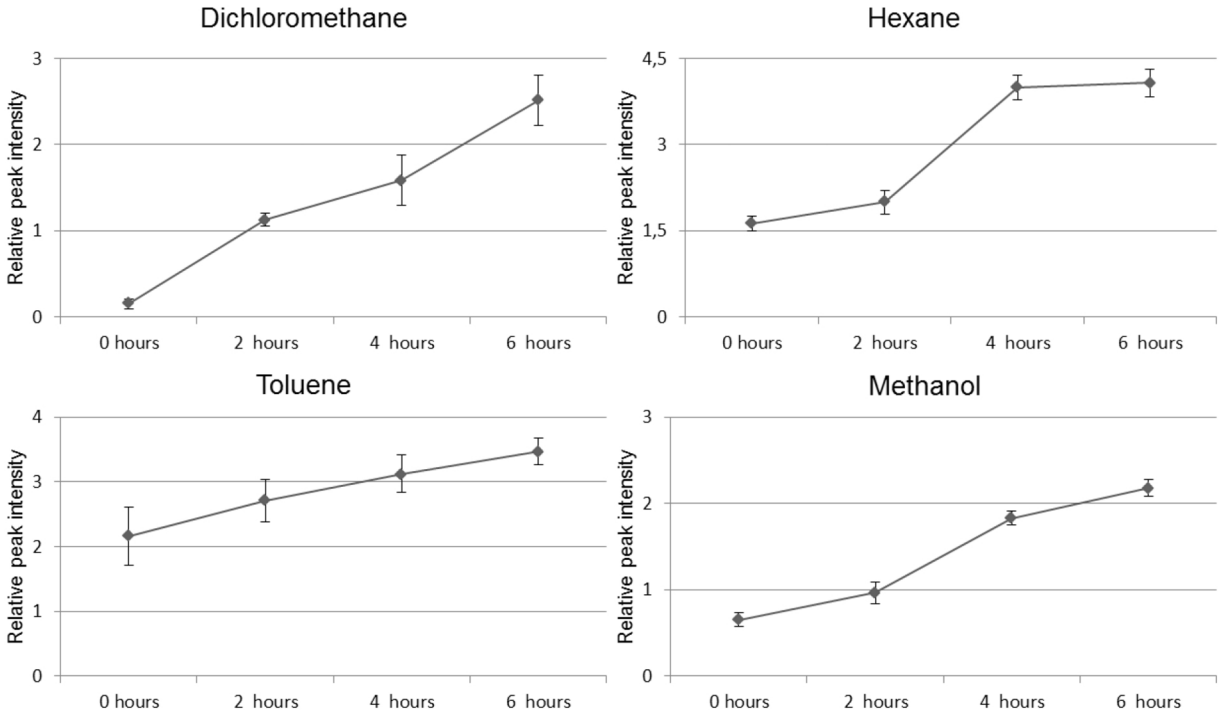
Dichloromethane
Relative peak intensity
3
2
1
0
0 hours
2 hours
4 hours
6 hours
Hexane
Relative peak intensity
4,5
3
1,5
0
0 hours
2 hours
4 hours
6 hours
Toluene
Relative peak intensity
4
3
2
1
0
0 hours
2 hours
4 hours
6 hours
Methanol
Relative peak intensity
3
2
1
0
0 hours
2 hours
4 hours
6 hours

Figure 1

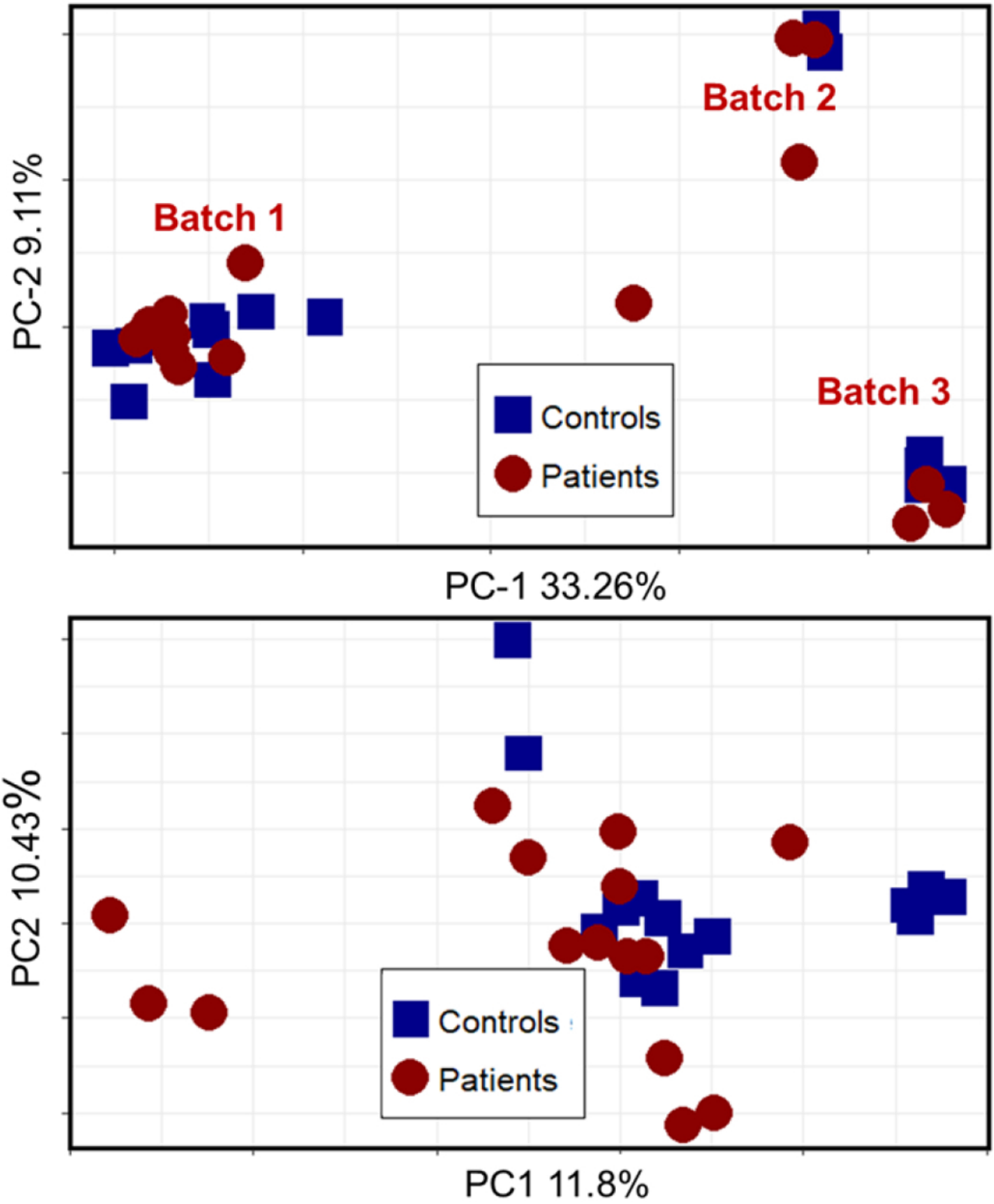
Batch 2
Batch 1
Batch 3
Controls
Patients
PC-2 9.11%
PC-1 33.26%
Controls
Patients
PC2 10.43%
PC1 11.8%

Figure 2

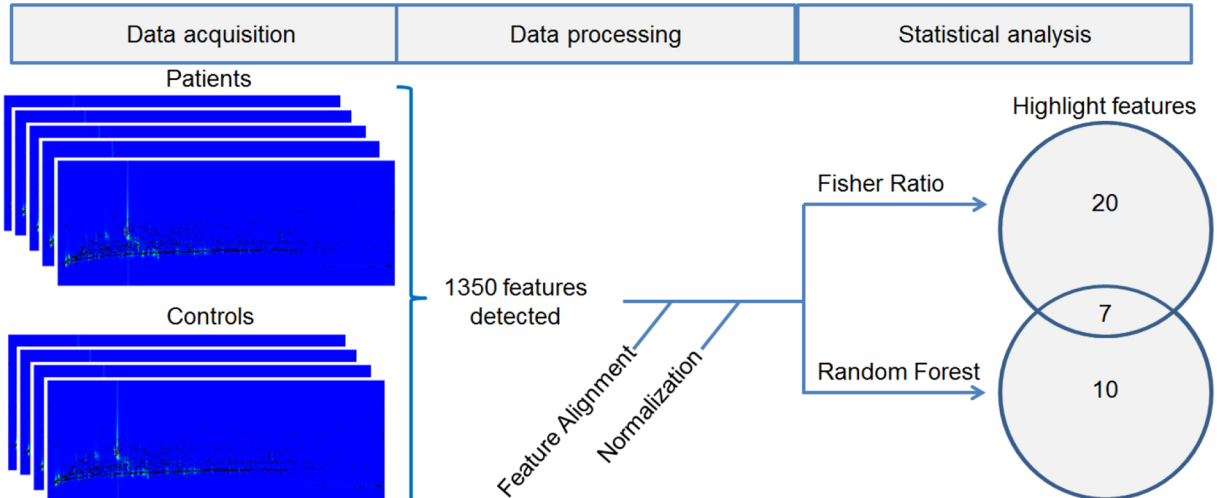
Data acquisition
Data processing
Statistical analysis
Patients
Controls
1350 features detected
Feature Alignment
Normalization
Fisher Ratio
Random Forest
Highlight features
20
7
10

Figure 3

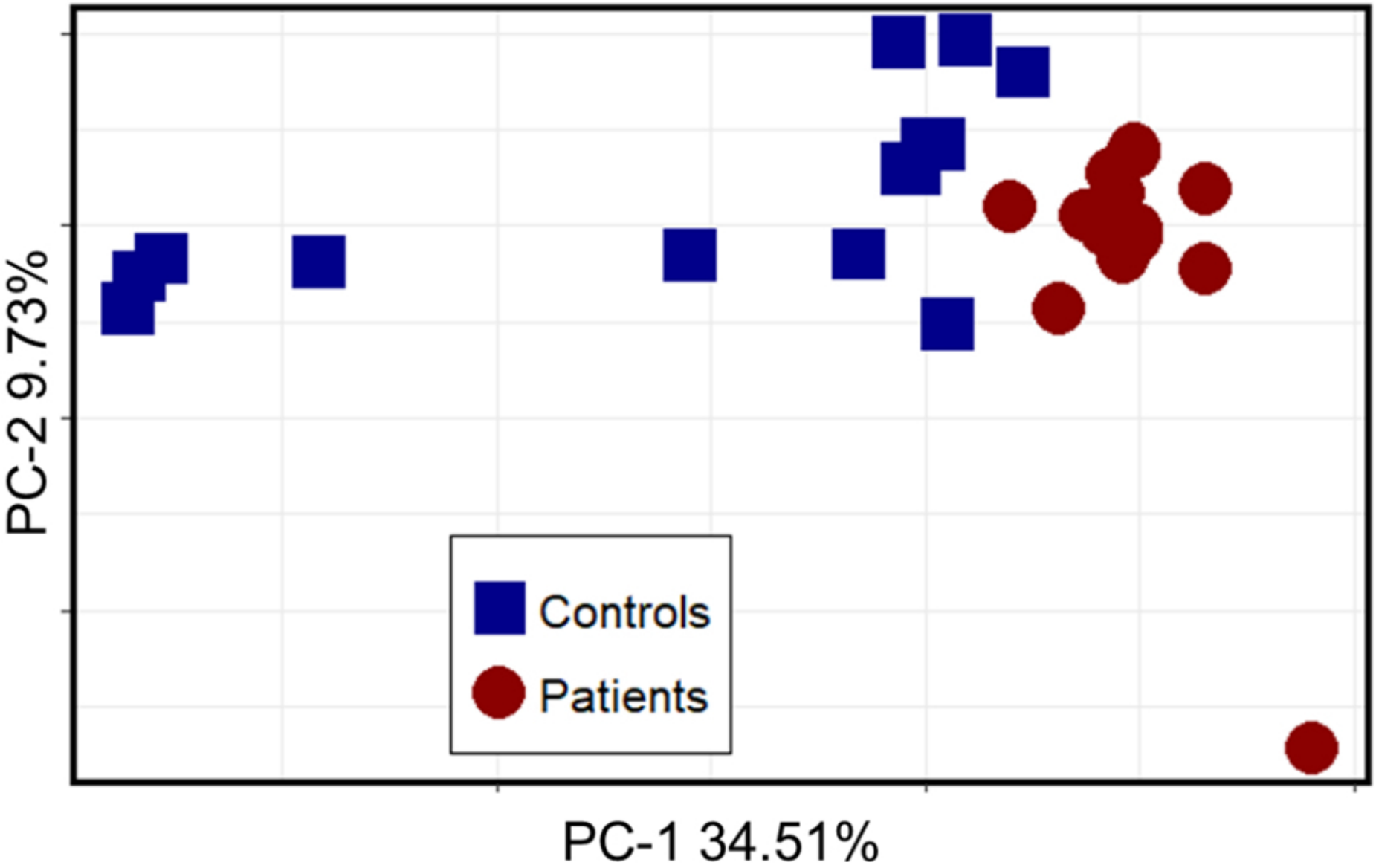


Figure 4

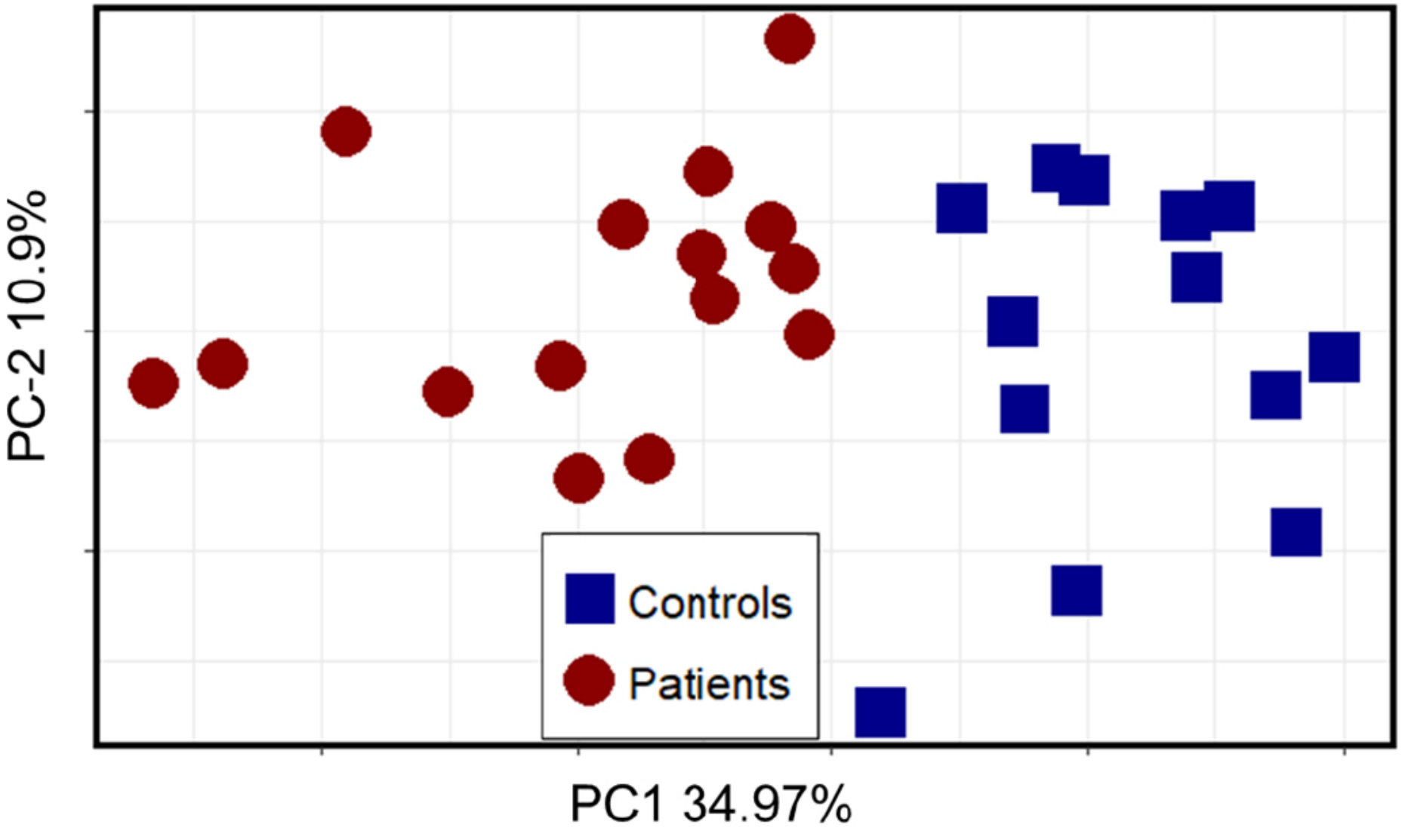


Figure 5

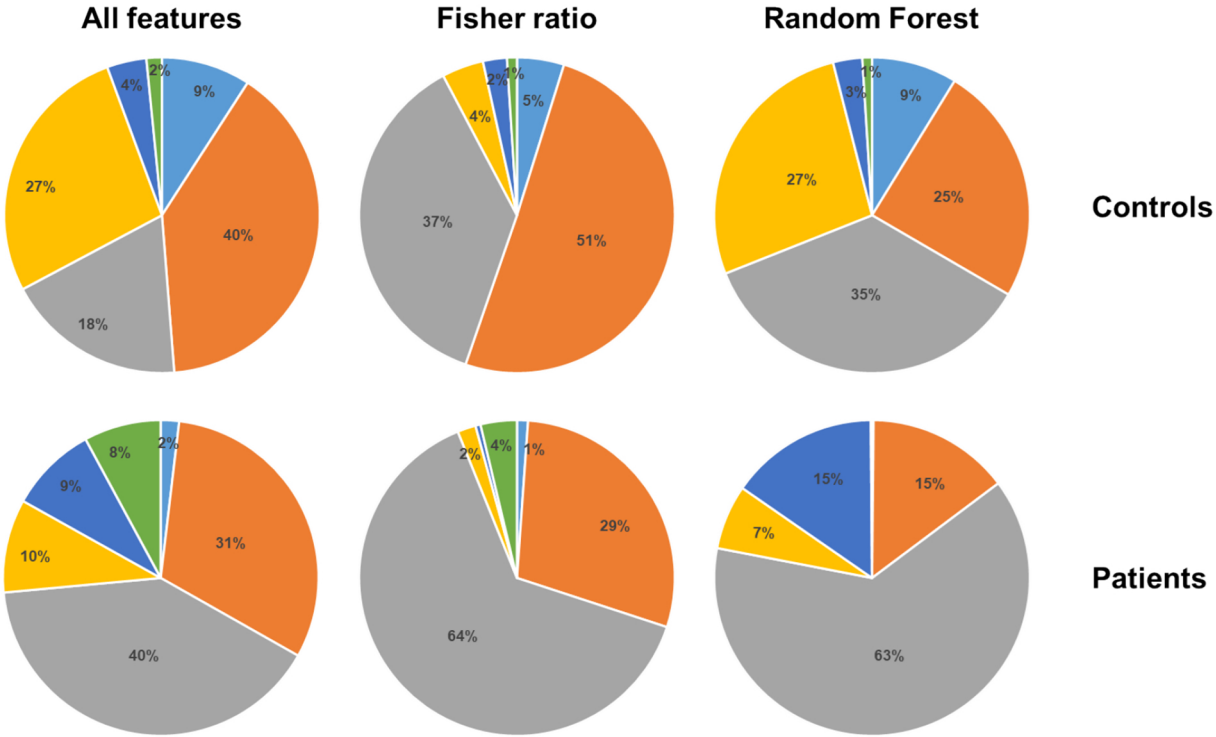
All features
Fisher ratio
Random Forest
Controls
Patients
9%
40%
18%
27%
4%
2%
5%
51%
37%
4%
2%
1%
9%
25%
35%
27%
3%
1%
2%
31%
40%
10%
9%
8%
1%
29%
64%
2%
4%
15%
15%
63%
7%

Figure 6

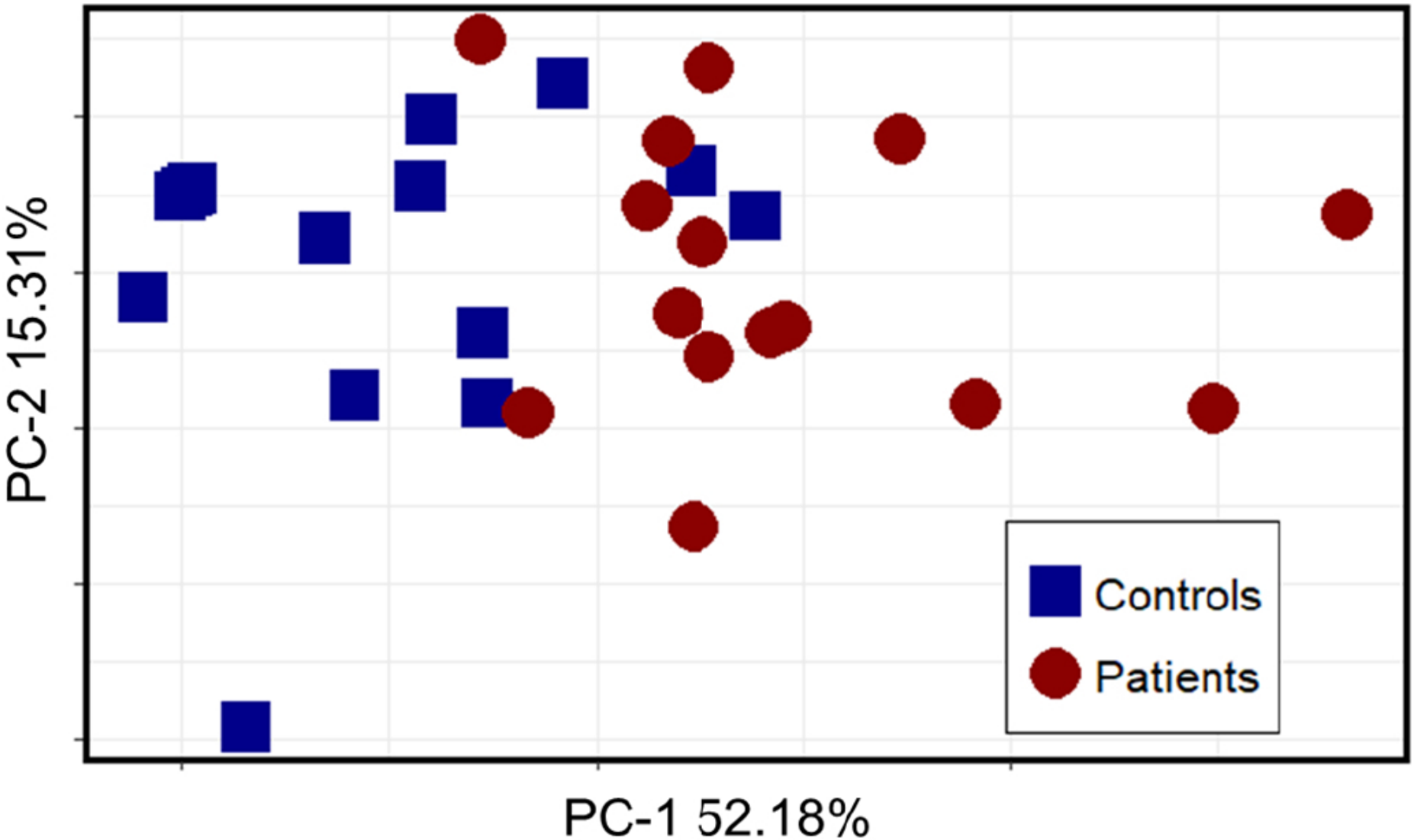


Figure 7