Secondary electrospray ionization-mass spectrometry and a novel statistical bioinformatic approach identifies a cancer-related profile in exhaled breath of breast cancer patients:...

Article *in* Journal of Breath Research · September 2015

DOI: 10.1088/1752-7155/9/3/031001

CITATIONS
7

READS
217

9 authors, including:

Pablo M-L Sinues
University Children's Hospital Basel, University of Basel
78 PUBLICATIONS **789** CITATIONS
SEE PROFILE

Elena Landoni
Fondazione IRCCS Istituto Nazionale dei Tumori di Milano
20 PUBLICATIONS **53** CITATIONS
SEE PROFILE

Rosalba Miceli
Fondazione IRCCS Istituto Nazionale dei Tumori di Milano
249 PUBLICATIONS **8,469** CITATIONS
SEE PROFILE

Matteo Dugo
Fondazione IRCCS Istituto Nazionale dei Tumori di Milano
111 PUBLICATIONS **430** CITATIONS
SEE PROFILE

Journal of Breath Research

RECEIVED
5 May 2015

REVISED
5 August 2015

ACCEPTED FOR PUBLICATION
17 August 2015

PUBLISHED
21 September 2015

NOTE

Secondary electrospray ionization-mass spectrometry and a novel statistical bioinformatic approach identifies a cancer-related profile in exhaled breath of breast cancer patients: a pilot study

Pablo Martinez-Lozano Sinues[1,8], Elena Landoni[2,3], Rosalba Miceli[3], Vincenza F Dibari[4], Matteo Dugo[5], Roberto Agresti[6], Elda Tagliabue[7], Simone Cristoni[4] and Rosaria Orlandi[7,9]

[1] National Research Council, Institute for Biomedical Technologies, Segrate, Milan, Italy
[2] Department of Clinical Sciences and Community Health, University of Milan, Italy
[3] Medical Statistics, Biometry and Bioinformatics, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[4] ISB, Ion Source & Biotechnologies, Gerenzano, Varese, Italy
[5] Functional Genomics DOSMM, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[6] Breast Surgery, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[7] Molecular Targeting Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

E-mail: rosaria.orlandi@istitutotumori.mi.it

Keywords: breast cancer, breath analysis, SESI, mass spectrometry, class prediction, machine learning

Supplementary material for this article is available online

Abstract

Breath analysis represents a new frontier in medical diagnosis and a powerful tool for cancer biomarker discovery due to the recent development of analytical platforms for the detection and identification of human exhaled volatile compounds. Statistical and bioinformatic tools may represent an effective complement to the technical and instrumental enhancements needed to fully exploit clinical applications of breath analysis. Our exploratory study in a cohort of 14 breast cancer patients and 11 healthy volunteers used secondary electrospray ionization-mass spectrometry (SESI–MS) to detect a cancer–related volatile profile. SESI–MS full-scan spectra were acquired in a range of 40–350 mass-to-charge ratio (m/z), converted to matrix data and analyzed using a procedure integrating data pre-processing for quality control, and a two-step class prediction based on machine-learning techniques, including a robust feature selection, and a classifier development with internal validation. MS spectra from exhaled breath showed an individual-specific breath profile and high reciprocal homogeneity among samples, with strong agreement among technical replicates, suggesting a robust responsiveness of SESI–MS. Supervised analysis of breath data identified a support vector machine (SVM) model including 8 features corresponding to m/z 106, 126, 147, 78, 148, 52, 128, 315 and able to discriminate exhaled breath from breast cancer patients from that of healthy individuals, with sensitivity and specificity above 0.9.

Our data highlight the significance of SESI–MS as an analytical technique for clinical studies of breath analysis and provide evidence that our noninvasive strategy detects volatile signatures that may support existing technologies to diagnose breast cancer.

Introduction

High-throughput platforms for cancer biomarker discovery are currently focused largely on genomic and proteomic studies. A complementary approach consists in comparing the entire metabolome profile of clinical samples to detect the significant metabolic changes occurring in cancer cells [1]. Metabolomics reflects changes in phenotype and thus function, thereby representing a powerful tool in addition to genomic and proteomic-based approaches to detect cancer development [2]. Cancer metabolites can be studied in virtually all body fluids including human breath. In this context, 'breathomics' has recently been defined as the metabolomic study of exhaled air mainly focusing on the characterization of health-related volatile organic compounds (VOCs) [3, 4]. Human breath analysis-based diagnosis has unique advantages, including

[8] Current address: Department of Chemistry and Applied Biosciences, ETH Zurich, 8093 Zurich, Switzerland.
[9] Author to whom any correspondence should be addressed.

simplicity, safety, minimal invasiveness, painlessness and readily acceptance by patients. Once fully developed, the use of such techniques may be particularly appropriate not only in early diagnosis, but also in on- and off-line management of pediatric patients, in all medical conditions requiring frequent diagnostic assessments and in monitoring therapeutic protocols or during the surgical procedures [5, 6]. In breast cancer (BC), about 60% of diagnosed invasive BCs remain localized at the time of diagnosis and the 5-year survival is nearly 100%; however, survival rates drop to 85% and 25% if regional or distal tissue invasion occurs, respectively [7]. Noninvasive identification of molecular markers that pinpoint small lesions, invisible by imaging techniques, could greatly improve the cure rate of BC and reduce its related mortality. Indeed, reports indicate that BC can be detected by canine olfaction [8] and by gas chromatography (GC)/ mass spectrometry (MS) [9].

Interest in studies aimed at identifying clinically relevant exhaled compounds, as pioneered by Pauling *et al* [10], has led to a significant development of appropriate analytical techniques for the detection and identification of human exhaled VOCs [11–14]. One such technique, secondary electrospray ionization MS (SESI–MS) [15, 16], has been shown to efficiently detect trace gas-phase compounds in breath or in any other matrix in real time [17]. While this technology has recently been used to identify bacterial pathogens [18, 19] and to characterize potential differences between patients with chronic obstructive pulmonary disease [20], dedicated statistical and bioinformatic tools that complement the technical and instrumental enhancements in analyzing breath-derived data are still lacking [21].

Our present study, exploring the value of SESI–MS technology together with novel statistical analysis tools in cancer biomarker discovery, supports the notion that this approach can identify a cancer-related volatile signature able to discriminate exhaled breath of BC patients from that of healthy controls.

Materials and methods

Subjects

A total of 25 women participated to this study, including 14 BC patients (cases) and 11 healthy volunteers (controls). Before surgery, patients were diagnosed with BC following the standard procedure in Fondazione IRCCS Istituto Nazionale dei Tumori; none of the patients received pharmacological treatment before breath sampling. Participants were asked not to smoke, eat, drink (except water), brush their teeth or use lipstick for at least 2 h before analysis. The study was approved by the Medical Ethics Committee of Fondazione IRCCS Istituto Nazionale dei Tumori (INT 122/14).

Sample collection

Breath samples were collected on 4 different days into 2 L inert plastic bags with a valve and disposable mouthpiece (ISB, Gerenzano, Italy) previously sterilized at 40 °C and 500 mTorr in the presence of H_2O_2. For 22 subjects (12 cases and 10 controls), two replicates were sampled within 5 min. To minimize variability in sample collection, storage and processing, human breath was sampled within 10 d in the same conditions for cases and controls, and plastic bags containing human breath were kept at 10ºC until analysis by SESI within 2 h of collection [22] using a mass spectrometer dedicated exclusively to analysis of breath samples.

Mass spectrometry

Mass spectrometer HCT Ion Trap (Bruker Daltonics, Billerica, MA, USA) coupled to a laboratory-built SESI source [17] was operated in the positive ion mode. Full-scan spectra were acquired in a range of 40–350 m z^{-1}; ion source parameters were capillary 3800 V, dry gas 2 L min^{-1} and temperature 40 °C. The MS instrument was slightly modified to allow admission of exhaled breath as described [22]. ES buffer (0.1% formic acid in H_2O) was infused at a flow rate of 130 nL min^{-1} by a syringe pump located outside the instrument [23].

Data acquisition and conversion to matrix

Hystar software (Bruker Daltonics, Breme, Germany) was used for data acquisition. MS spectra data of the volatile fraction were converted to the ascii xy format using Data Analysis software (Bruker Daltonics, Breme, Germany) and analyzed using the NIST database approach for pattern recognition [24]. Custom Perl (www.perl.org/) script served to optimize absolute values of each MS signal by approximating m/z and eliminating decimal places, with values then included in a [feature x sample] matrix.

Statistical and bioinformatic methods

Analyses were carried out using R software, version 2.15.2 (www.r-project.org/) and Bioconductor (www.bioconductor.org/). Test results were considered significant at p-value <0.05.

Data pre-processing

Data pre-processing (figure 1) included five steps. Steps I, III, IV refer to analytical procedures for data quality control using statistical and bioinformatic methods imported from gene expression studies. Step I involved assessment of concordance between technical replicates using concordance and Bland–Altman plots [25] and estimation of the Lin's concordance correlation coefficient (CCC) [26] with the corresponding bootstrap 95% confidence interval [27], based on a user-defined R function (*f. concordance,* supplementary file S1A), which exploits the *epi.ccc* function included in the *epiR* package. The Bland–Altman plot identified possible outliers, i.e. replicates outside the confidence bands, the mean of which could be a biased estimate of the true value. In step II, features not detected in at least 50% of samples were filtered and discarded. Step III was

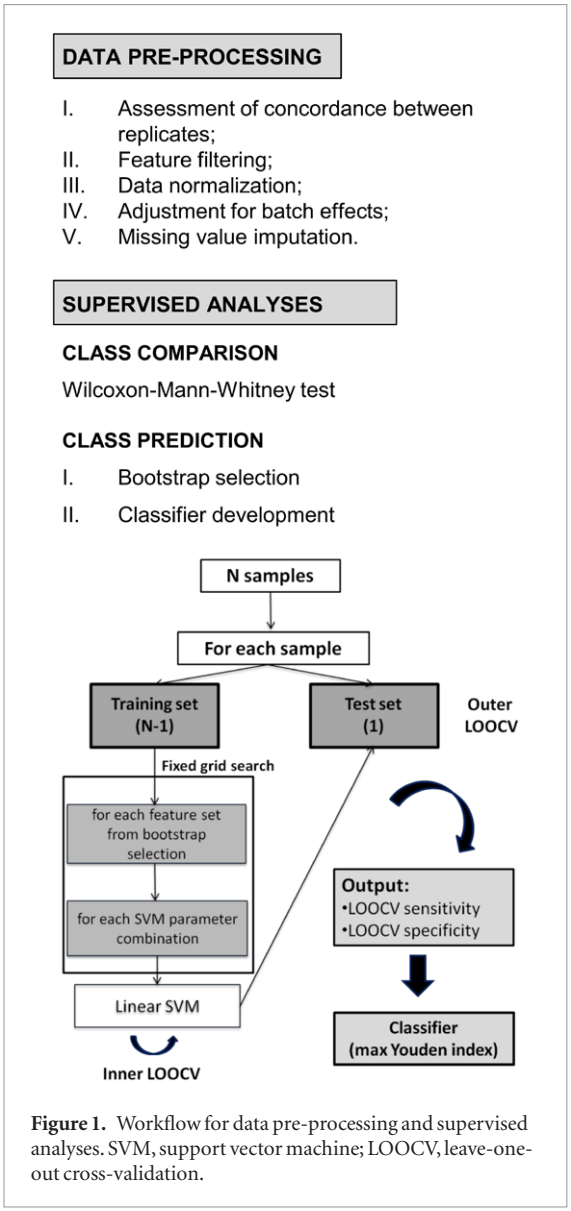

Figure 1. Workflow for data pre-processing and supervised analyses. SVM, support vector machine; LOOCV, leave-one-out cross-validation.

data normalization [28] using the quantile method (*normalize between arrays* function in the Limma package) to impose the same empirical feature distribution to each subject. Batch effects arising from different dates of sample collection (step IV) were corrected using the ComBat method (combating batch effects when combining batches of gene expression microarray data) [29]. The *ComBat* function in the *sva* R package was used for this task. Dendrograms based on hierarchical clustering of subjects before and after ComBat correction were generated to visualize the effectiveness of adjustment for batch effects. Subjects were identified according to dates of sample collection (batches); average linkage was used as the linkage criterion to construct the hierarchical cluster tree and distance was measured as one minus the Spearman correlation coefficient [30] such that two subjects exhibiting a strong positive correlation are closer, possibly reflecting the same feature profile between paired subjects. The *heatmap.2* function in the *gplots* package was used to perform hierarchical clustering. In step V, missing values in feature replicates were imputed after data normalization as described by Karpievitch *et al* [31].

Supervised analyses

Class comparison was performed using the nonparametric Wilcoxon–Mann–Whitney test, adjusting *p*-values for multiple testing by the Benjamini–Hochberg method [32] to control for false-discovery rate (FDR); class prediction involved 'bootstrap feature selection' [33], followed by classifier development and its internal validation (figure 1). In the first step, 1000 bootstrap samples were drawn from the original dataset and the features were robustly ranked according to the proportion of bootstrap samples in which they were jointly identified as independent class predictors by three different classification algorithms, i.e. prediction analysis for microarrays (PAM) [34], random forest (RF) with Boruta feature selection [35] and SVM algorithm with L1-penalisation (L1 SVM) [36] or, alternatively, elastic smoothly clipped absolute deviation (SCAD) SVM [37]. An egg-shaped plot was initially used to summarize the bootstrap-derived feature occurrences (nodes) and co-occurrences (edge thickness). The larger the node, the more often the corresponding feature occurred in the bootstrap samples; the thicker the edge between two nodes/features, the more often they were selected together in the bootstrap samples. Bootstrap selection was performed using modified *doBS* and *importance igraph* functions in the *bootfs* package. To develop the cross-validated classifier, we applied linear SVM models, well-established machine-learning techniques used for high-dimensional data such as 'omics' data [38, 39]. A linear SVM model, implemented using the function *svm* in the *e1071* package, requires the tuning of only two parameters (cost parameter and class weights). Models were fitted by varying parameters and number of included features, forwardly selected according to the bootstrap-generated list. Each model was then internally validated using a leave-one-out cross-validation (LOOCV) procedure to optimize parameters and estimate the model classification ability by computing sensitivity, specificity and Youden index (sensitivity + specificity − 1). The false-positive rate (FPR, 1—specificity) and true-positive rate (TPR, sensitivity) were graphically represented in the 'ROC space' plot, which can be seen as a generalization of the ROC curve representing the classification performance of the different linear SVM models. The final model used to develop the classifier was chosen based on both best classification performance, as indicated by the highest Youden index, and smallest number of features included in the model. The heatmap was generated by clustering feature values and using the 'one minus the Spearman correlation coefficient' as distance metric and the average linkage as linkage criterion.

'Feature importance analysis' was performed by generating 1000 permuted data sets and running the L1 SVM-based bootstrap selection procedure on each

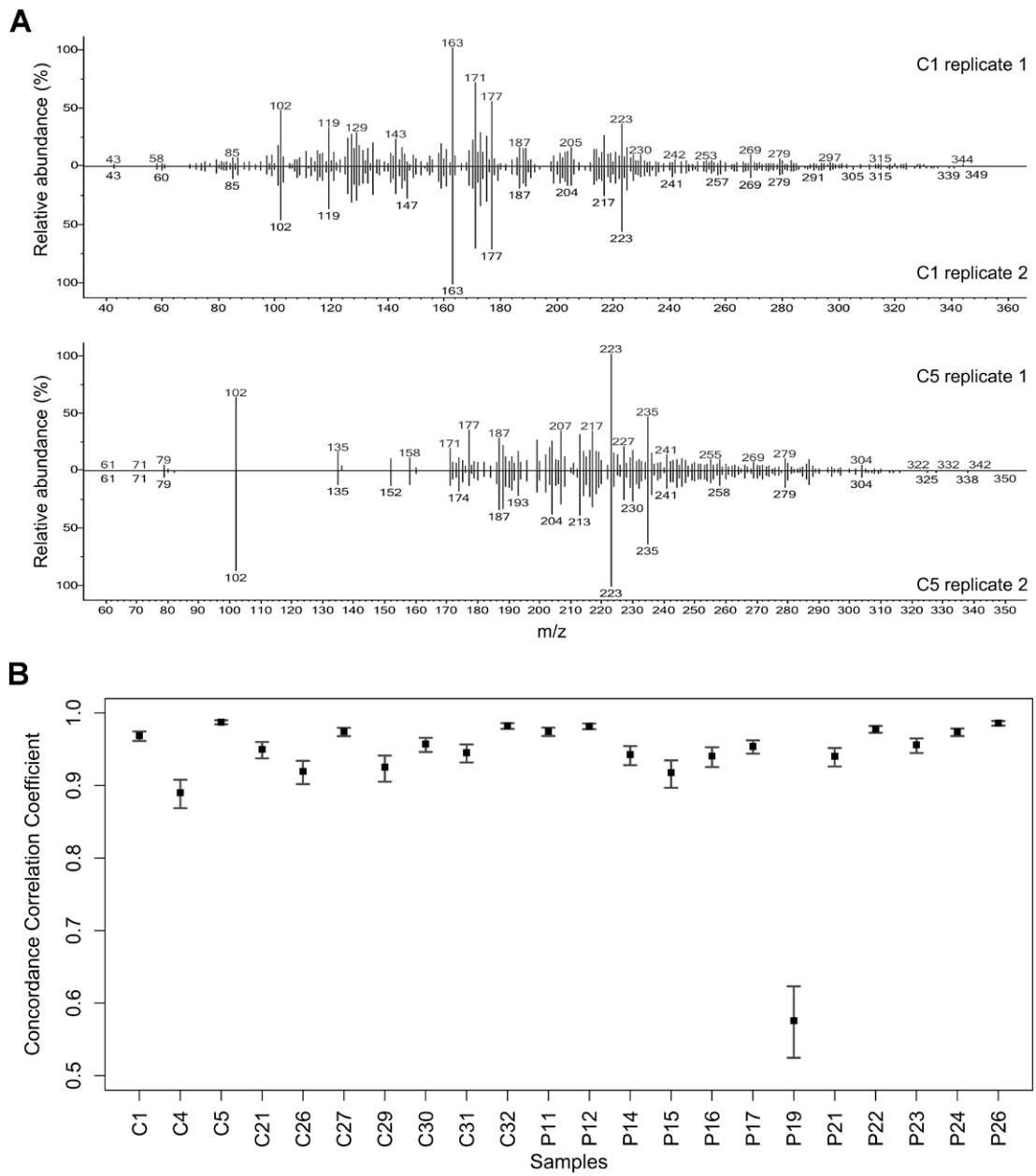

Figure 2. Quality control of pre-analytical and analytical procedures in data pre-processing. (A) Evaluation of MS spectra quality based on NIST comparison of MS spectra from exhaled breath of two replicates from randomly selected subjects. (B) Concordance analysis of replicates based on concordance correlation coefficient (CCC) and corresponding 95% bootstrap confidence interval (CI) for each subject.

random data set. The best three features of each selection were extracted, compared to the features bootstrap-selected on not permuted original data and the co-occurrence in permuted and original datasets were calculated.

Results

Exhaled breath from BC patients and healthy controls were sampled in duplicate within 2 h before SESI/MS analysis. Patients' breath samples were collected 3–24 h before surgery. Histology confirmed the presence of a tumor mass at the time of sampling. Supplementary file S2 lists the clinical and pathological characteristics of patients, revealing consistency in BC consecutive cohorts, i.e. average tumor size 2 cm, 64% node-negative tumors, 71% ER (estrogen receptor) positive, and 71% grade I-II.

Data pre-processing

MS spectra from exhaled breath of a subject were highly similar in replicates and each participant showed an individual-specific breath profile (figure 2(A)), consistent with previous studies [22, 40, 41]. Breath mass spectra were processed and converted to a final matrix including 351 features and 47 samples (16497 total values). About 17% of values (2838/16497) were missing and equally distributed between the first and second replicates and between cases and controls, suggesting that missing values were missing at random

(MAR, supplementary file S3). In step I of data pre-processing (figure 1), the concordance correlation coefficient (CCC, figure 2(B)) for the 22 subjects with technical replicates, together with the concordance plots and the Bland–Altman plots (supplementary file S1B), confirmed the agreement between technical replicates (CCC range: 0.89–0.99). Overall, variability of SESI–MS measurements was higher for signals in the low-abundance region. Based on the concordance analysis, the mean of the two replicate values for each feature in each sample was calculated and, after filtering (step II), 296 features remained for subsequent analyses. Based on the above results and considering the quality and characteristics of SESI breath data and their matrix data structure, statistical and bioinformatic tools for data pre-processing and supervised analyses of gene expression data were applied. After quantile normalization (step III, supplementary file S4A) and ComBat correction (step IV), the dendrogram obtained from hierarchical clustering indicated grouping of samples independent of daily batches (supplementary file S4B). Seventy values missing in both replicates and occurring in 18 samples and 30 features were imputed using the median of each feature (step V) under the MAR assumption.

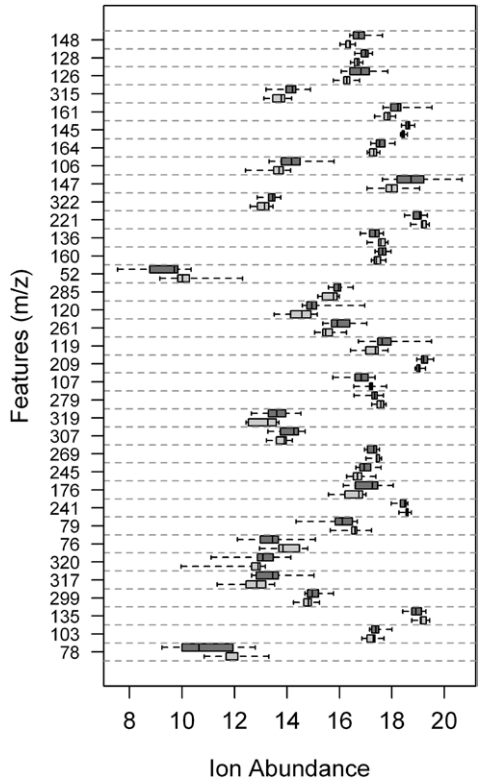

Figure 3. Class comparison analysis. Boxplots showing the distribution of the ion abundance values of the 35 significantly differentially detected features in cases (dark grey) versus controls (light grey), sorted in ascending order according to the Wilcoxon test *p*-value.

Classifier development

The resulting data matrix [296 features × 25 subjects] was further statistically analyzed in a supervised setting (figure 1) in an effort to identify signals reflecting exhaled compounds that may be valuable in identifying BC, based on the mass spectrometric fingerprints. Class comparison revealed 35 features in breath that differed significantly between cases and controls (figure 3), 24 (69%) of which were present at higher levels in the exhaled breath from cases. This subset included the features with m/z 148 and 128, showing FDR-adjusted *p*-values of 0.0259 and 0.0770, respectively (nominal *p*-values: 9e-05 and 5e-04). In class prediction analysis, we first attempted to identify the most informative signals using the bootstrap feature selection strategy, based on 3 different classification algorithms: PAM, RT and L1 SVM Bootstrap results are represented by an egg-shaped plot (figure 4(A)) that provides an immediate overview of the feature relevance in terms of bootstrap occurrence (node size) and co-occurrence (edge thickness); the latter can suggest possible structural and/or biological links among molecules. The signal detected at m/z 106, selected in 346 of 1000 bootstrap samples was the most discriminative, followed in decreasing order by signals at m/z 126 and 147 (345 of 1000), 78 (331 of 1000), 148 and 52 (322 of 1000) and 128 (259 of 1000). Figure 4(B) shows the frequency of bootstrap occurrences and co-occurrences of the features represented in the egg-shaped plots. The most frequent co-occurrences involved feature corresponding to m/z 126 and were jointly selected with feature 147 (202 of 1000 bootstrap samples), 148 (190 of 1000), 128 (185 of 1000) and 315 (160 of 1000); note that signal at m/z 148 co-occurred with its isotope at m/z 147 (161 of 1000). Signals selected in at least one bootstrap sample were used in a multivariable context to develop a cross-validated linear SVM classifier. The 'ROC space' in figure 5(A) shows the results of the different SVM models in terms of cross-validated classification performance. The model with sensitivity = 0.93, specificity = 0.91 and Youden index = 0.84, including eight features (m/z 106, 126, 147, 78, 148, 52, 128, 315) was used to develop the classifier, characterized by the best performance in discriminating exhaled breath from BC patients and by the smallest number of features among the models with equal discriminating performance. Figure 5(B) shows a heat map representing the abundance of the 8 features in each sample. Overall, supervised analysis indicated that the mass regions, 147–148, 126–128, 106 and 315 included signals originating from molecules possibly chemically and/or functionally related and differentially over-detected in exhaled breath of BC patients.

The bootstrap feature selection was also performed by applying the elastic SCAD SVM, in conjunction with PAM and RF, obtaining similar results as those achieved using L1 SVM (supplementary file S5). Both feature selection algorithms clearly tends to pick isotopic features (e.g. m/z 147–148), suggesting that they efficiently selects interconnected signals.

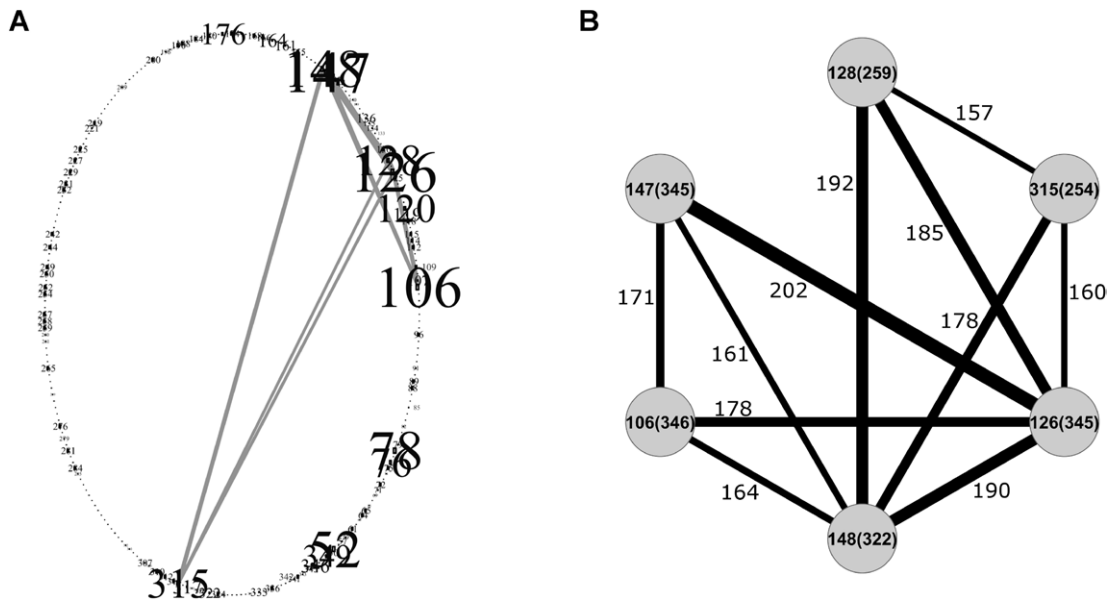

Figure 4. Bootstrap feature selection. (A) Egg-shaped plot representing bootstrap results and showing features (m/z) with at least 150 co-occurrences. (B) Frequency of occurrences (in brackets) and co-occurrences (next to edge) of the features represented in the egg-shaped plot. Node size and edge thickness are proportional to the frequency of bootstrap occurrences and co-occurrences, respectively.

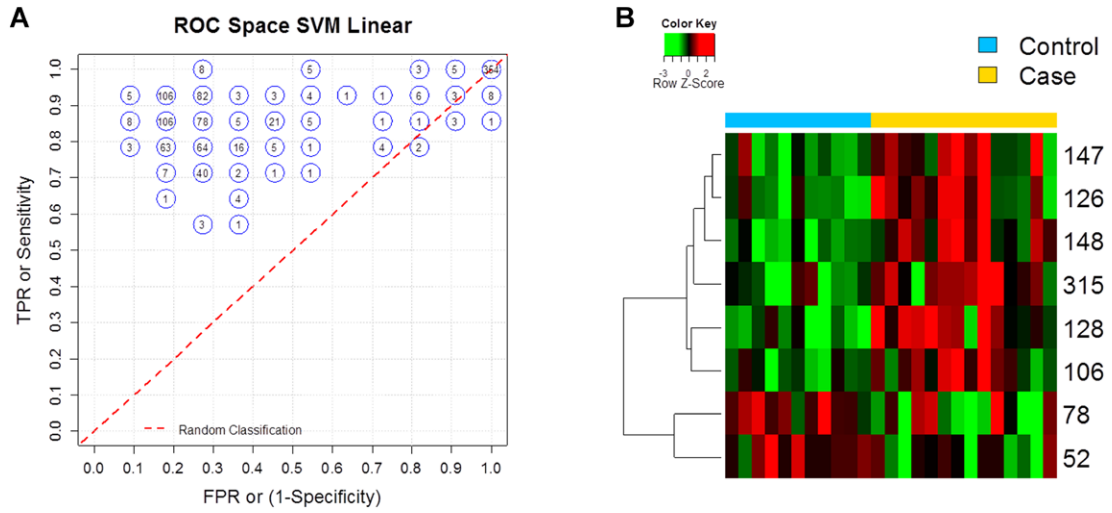

Figure 5. Classifier development. The 'ROC space' plot represents the LOOCV classification performance of different linear SVM models for class prediction in terms of false-positive rate (FPR; 1-specificity) and true-positive rate (TPR; sensitivity). B) Heatmap showing color-coded ion abundance values of the 8 features of the final classifier. Green and red represent under- and over-detected signals in cases, respectively.

Finally, to ensure that the discriminant features were not selected merely by chance, we calculated how often the eight features of the final classifier were the top three in the bootstrap selection classifications obtained from 1000 permuted data sets, in which a random association between features and classes (cases and controls) was generated. None of the eight features was selected as top three in 848 permuted data sets, one of the eight features was top three in 141 selections, and a couple of features appeared jointly in 11 selections. The most predictive feature, i.e. 106, appeared 2% of the times as top three; the corresponding percentages for the other features were: 0.4% for 126 and 147, 5.9% for 78, 0.1% for 148, 7.2% for 52, 0.2% for 128 and 0.1% for 315. This indicates that the eight BC–related features were randomly bootstrap-selected in permuted datasets, supporting the robustness of the classifier.

Discussion

The recent dramatic improvement of analytical platforms for metabolic profiling has provided evidence encouraging the use of metabolic biomarkers as a valuable tool for cancer detection [1]. In the case of biomarker discovery in exhaled breath, sample handling, chemical analysis and subsequent data mining await standardization after further exploration

of novel and suitable approaches. In particular, there is a well-identified need for defining data pre-processing techniques and supervised methods in breath analysis, as recently highlighted by Smolinska and coworkers [21]. Here, we explored the clinical usefulness of a combination of a promising breath analytical tools, i.e. SESI–MS, and a statistical and bioinformatic tool for data pre-processing and classifier development. Using this novel strategy, we identified a BC-related signature able to classify patient and control breathprints with sensitivity and specificity above 0.9.

Because all 'omics' studies are generally affected by several sources of variability, including inherent biological variation as well as data noise due to intrinsic technical variability, there is a need for efficient governance of non-biological variability in the pre-analytical and analytical phases to minimize data misinterpretation. One of the main advantages of the analytical platform used herein resides in the rapid screening of breath metabolites made possible by SESI–MS, resulting in rich MS fingerprints without any sample preparation. Indeed, the complete SESI–MS analysis of two replicate samples collected in appropriate bags was typically accomplished in less than one minute. The accuracy of replicate profiles and high reciprocal homogeneity among all samples suggest the robustness of SESI–MS measurements. In addition, data produced using SESI–MS analysis were high-dimensional data of quality comparable to gene expression analysis output, prompting us to use methods originally developed for gene expression data in analyzing SESI–MS breathprints. Together, the properties of SESI–MS suggest its particular suitability for large-scale clinical breath analyses.

In the post-analytical step, we used data pre-processing techniques to correct for residual systematic technical or non-biological experimental variation. The ComBat correction, in particular, was used to adjust for batch effects arising from the breath sample collection procedure and analysis performed on different days.

With the aim of developing a VOC signature that accurately discriminates between cases and controls, we used a two-step procedure involving current and new approaches for data analysis and representation that promises to ensure generalization of results and provide insights into feature interconnections. Bootstrap feature selection raised a robust and specific feature ranking, as supported by the random selection of the BC-related features when the bootstrap procedure was applied to 1000 permuted datasets. The bootstrap feature selection was based on conceptually different machine-learning algorithms: PAM, RF and two alternative SVM algorithms, i.e. L1 elastic SCAD. The above algorithms were chosen because they can overcome the 'curse of dimensionality' typical of 'omics' data, i.e. feature numbers much larger than subject numbers, and are representative of methodological categories using different decision rules for classification. PAM provides simplicity and interpretability, while RF and SVMs algorithms are suitable for complex classification patterns. RF is nonparametric, which is desirable especially when outliers occur, and also deals with data overfitting; however, RF only outputs importance measures, the interpretation of which is controversial in the presence of correlated features [42]. The recent elastic SCAD SVM [37] has been proposed as an effective method for considering the correlation structures in the input data (grouping effect) and applied to develop a miRNA-based classifier able to discriminate hemolyzed and not hemolyzed plasma samples [43]. The L1 SVM showed high prediction accuracy for models in which the sparsity assumption (small number of nonzero parameters) is tenable [36], as in the case of 'omics' data characterized by few predictive variables. In our application the two alternative SVM algorithms generated comparable bootstrap feature rankings, but the computation time of bootstrap process was dramatically reduced using L1 SVM jointed to PAM and RT, addressing this setting of our pipeline to biomarker discovery in large cohort of patients.

We summarized the bootstrap feature selection in an egg-shaped plot, allowing immediate visualization of the most discriminative features and their co-occurrences and highlighting the possible interconnections underlying the structural/biological framework usually hidden in massive data. Lastly, we derived a molecular signature associated with cancer patient breath samples using a linear SVM model based on the 'rule' of best classification performance, as indicated by the highest Youden index, and smallest number of features included in the model. Linear SVM models require the tuning of only two parameters but, like all SVM models, do not allow probability estimation or ROC curve generation; nevertheless, a binary classifier can be easily derived from SVM predictions, since they cluster around two different values.

While conclusions from our study are limited by our small sample size, our statistical and bioinformatic strategy for breath analysis has already been adapted to other 'omics' data, including microRNA data [43] and a complex LC-MS dataset from high-resolution Orbitrap analysis of plasma samples [43]. This indicates that our procedure is flexible enough to adapt the algorithm to different questions, data structure or knowledge domain.

The present study designed a general strategy effective in discovering potential volatile biomarkers to be developed for early diagnosis of cancer. The classification performance of our volatile signature awaits confirmation in larger cohorts of BC patients and non-cancer control subjects. Another important aspect is the identification of the most discriminative exhaled compounds to gain insights into the disease. This task can be undertaken by combining SESI with mass spectrometers of high resolution and fragmentation capabilities, thereby enabling unambiguous identification [44, 45].

The translation of the biomarker discovery phase to the clinical practice, beyond the validation phase,

will still require a significant development of the analytical and bioinformatics procedures, tailored on the definitive classifier, to finally deliver a user-friendly tool for the rapid, simple and precise detection of cancer-related signals by breath analysis.

Conclusion

Overall, our study supports the value of SESI–MS as an analytical technique for clinical studies, since it allows rapid collection of rich metabolic breathprints, and underscores the importance of sample quality assessment and quality control of raw data from breath analysis using a robust data pre-processing techniques to address unbiased pattern discovery. Our identification of a potential cancer-related volatile signature that identifies BC patients based on their exhaled metabolic breathprint provides the foundation and rationale for further analyses aimed at developing a noninvasive diagnostic tool for prediction of BC.

Acknowledgments

We thank all healthy volunteers, patients, physicians and nurses participating in this study as well as L Zingaro for excellent technical assistance.

This work was partially supported by a Marie Curie Intra-European Fellowship within the 7th European Community Framework Programme (PIEF-GA-2008–220511 to PMLS).

The authors have declared no conflicts of interest.

References

[1] Sreekumar A *et al* 2009 Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression *Nature* **457** 910–4
[2] Nicholson J K and Lindon J C 2008 Systems biology: metabonomics *Nature* **455** 1054–6
[3] Rattray N J, Hamrang Z, Trivedi D K, Goodacre R and Fowler S J 2014 Taking your breath away: metabolomics breathes life in to personalized medicine *Trends Biotechnol.* **32** 538–48
[4] Haick H, Broza Y Y, Mochalski P, Ruzsanyi V and Amann A 2014 Assessment, origin, and implementation of breath volatile cancer markers *Chem. Soc. Rev.* **43** 1423–49
[5] Amann A and Smith D 2013 *Volatile Biomarkers* 1st edn (Amsterdam: Elsevier) vol 1
[6] Martinez-Lozano Sinues P, Zenobi R and Kohler M 2013 Analysis of the exhalome: a diagnostic tool of the future *Chest* **144** 746–9
[7] National Cancer Institute 2015 SEER Stat Fact Sheets: Breast Cancer (http://seer.cancer.gov/statfacts/html/breast.html)
[8] McCulloch M, Jezierski T, Broffman M, Hubbard A, Turner K and Janecki T 2006 Diagnostic accuracy of canine scent detection in early- and late-stage lung and breast cancers *Integr. Cancer Ther.* **5** 30–9
[9] Phillips M *et al* 2014 Rapid point-of-care breath test for biomarkers of breast cancer and abnormal mammograms *PLoS.One* **9** e90226
[10] Pauling L, Robinson A B, Teranishi R and Cary P 1971 Quantitative analysis of urine vapor and breath by gas–liquid partition chromatography *Proc. Natl Acad. Sci. USA* **68** 2374–6
[11] Lovett A M, Reid N M, Buckley J A, French J B and Cameron D M 1979 Real-time analysis of breath using an atmospheric pressure ionization mass spectrometer *Biomed. Mass Spectrom.* **6** 91–7
[12] Benolt F M, Davidson W R, Lovett A M, Nacson S and Ngo A 1983 Breath analysis by atmospheric pressure ionization mass spectrometry *Anal. Chem.* **55** 805–7
[13] Smith D and Spanel P 2005 Selected ion flow tube mass spectrometry (SIFT-MS) for on-line trace gas analysis *Mass Spectrom. Rev.* **24** 661–700
[14] Lindinger W and Jordan A 1998 Proton-transfer-reaction mass spectrometry (PTR–MS): on-line monitoring of volatile organic compounds at pptv levels *Chem. Soc. Rev.* **27** 347–75
[15] Dillon L A, Stone V N, Croasdell L A, Fielden P R, Goddard N J and Thomas C L 2010 Optimisation of secondary electrospray ionisation (SESI) for the trace determination of gas-phase volatile organic compounds *Analyst* **135** 306–14
[16] Reynolds J C *et al* 2010 Detection of volatile organic compounds in breath using thermal desorption electrospray ionization-ion mobility-mass spectrometry *Anal. Chem.* **82** 2139–44
[17] Martinez-Lozano P and Fernandez de la Mora J 2008 Direct analysis of fatty acid vapors in breath by electrospray ionization and atmospheric pressure ionization-mass spectrometry *Anal. Chem.* **80** 8210–5
[18] Ballabio C, Cristoni S, Puccio G, Kohler M, Sala M R, Brambilla P and Martinez-Lozano Sinues P 2014 Rapid identification of bacteria in blood cultures by mass-spectrometric analysis of volatiles *J. Clin. Pathol.* **67** 743–6
[19] Zhu J, Bean H D, Jimenez-Diaz J and Hill J E 2013 Secondary electrospray ionization-mass spectrometry (SESI–MS) breathprinting of multiple bacterial lung pathogens, a mouse model study *J. Appl. Physiol.* **114** 1544–9
[20] Martinez-Lozano Sinues P, Meier L, Berchtold C, Ivanov M, Sievi N, Camen G, Kohler M and Zenobi R 2014 Breath analysis in real time by mass spectrometry in chronic obstructive pulmonary disease *Respiration* **87** 301–10
[21] Smolinska A, Hauschild A C, Fijten R R, Dallinga J W, Baumbach J and van Schooten F J 2014 Current breathomics—a review on data pre-processing techniques and machine learning in metabolomics breath analysis *J. Breath Res.* **8** 027105
[22] Martinez-Lozano P, Zingaro L, Finiguerra A and Cristoni S 2011 Secondary electrospray ionization-mass spectrometry: breath study on a control group *J. Breath Res.* **5** 016002
[23] Martinez-Lozano Sinues P, Criado E and Vidal G 2012 Mechanistic study on the ionization of trace gases by an electrospray plume *Int. J. Mass Spectrom.* **313** 21–9
[24] Sinues P M, Alonso-Salces R M, Zingaro L, Finiguerra A, Holland M V, Guillou C and Cristoni S 2012 Mass spectrometry fingerprinting coupled to national institute of standards and technology mass spectral search algorithm for pattern recognition *Anal. Chim. Acta.* **755** 28–36
[25] Bland J M and Altman D G 1986 Statistical methods for assessing agreement between two methods of clinical measurement *Lancet* **1** 307–10
[26] Lin L I 1989 A concordance correlation coefficient to evaluate reproducibility *Biometrics* **45** 255–68
[27] Efron B 1979 Bootstrap methods: another look at jackknife *Ann. Stat.* **7** 1–26
[28] Callister S J, Barry R C, Adkins J N, Johnson E T, Qian W J, Webb-Robertson B J, Smith R D and Lipton M S 2006 Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics *J. Proteome Res.* **5** 277–86
[29] Johnson W E, Li C and Rabinovic A 2007 Adjusting batch effects in microarray expression data using empirical Bayes methods *Biostatistics* **8** 118–27
[30] Key M 2012 A tutorial in displaying mass spectrometry-based proteomic data using heat maps *BMC Bioinformatics* **13** S10
[31] Karpievitch Y V, Dabney A R and Smith R D 2012 Normalization and missing value imputation for label-free LC-MS analysis *BMC Bioinformatics* **13** S5

[32] Benjamini Y and Hochberg Y 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing *J. R. Stat. Soc.* B **57** 289–300
[33] Austin P C and Tu J V 2004 Bootstrap methods for developing predictive models *Am. Stat.* **58** 131–7
[34] Tibshirani R, Hastie T, Narasimhan B and Chu G 2002 Diagnosis of multiple cancer types by shrunken centroids of gene expression *Proc. Natl Acad. Sci. USA* **99** 6567–72
[35] Kursa M B and Rudnicki W R 2010 Feature selection with the Boruta package *J. Stat. Softw.* **36** 1–13
[36] Hastie T, Tibshirani R and Wainwright M 2015 *Statistical Learning with Sparsity: The Lasso and Generalizations* 1st edn (London: Chapman and Hall)
[37] Becker N, Toedt G, Lichter P and Benner A 2011 Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data *BMC Bioinformatics* **12** 138
[38] Cortes C and Vapnik V 1995 Support-vector networks *Mach. Learn.* **20** 273–97
[39] Furey T S, Cristianini N, Duffy N, Bednarski D W, Schummer M and Haussler D 2000 Support vector machine classification and validation of cancer tissue samples using microarray expression data *Bioinformatics* **16** 906–14
[40] Martinez-Lozano Sinues P, Kohler M and Zenobi R 2013 Human breath analysis may support the existence of individual metabolic phenotypes *PLoS One* **8** e59909
[41] Wang X R, Lizier J T, Berna A Z, Bravo F G and Trowell S C 2015 Human breath-print identification by E-nose, using information-theoretic feature selection prior to classification *Sensors Actuators* B **217** 165–74
[42] Grömping U 2009 Variable importance assessment in regression: Linear regression versus random forest *Am Stat.* **63** 308–19
[43] Landoni L *et al* 2015 Proposal of supervised data analysis strategy of plasma miRNAs from hybridisation array data with an application to assess hemolysis-related deregulation *BMC Bioinformatics* In Press
[44] García-Gómez D, Bregy L, Barrios-Collado C, Vidal-de-Miguel G and Zenobi R 2015 Real-time high-resolution tandem mass spectrometry identifies furan derivatives in exhaled breath *Anal. Chem.* **87** 6919–24
[45] García-Gómez D, Martínez-Lozano Sinues P, Barrios-Collado C, Vidal-De-Miguel G, Gaugg M and Zenobi R 2015 Identification of 2-alkenals, 4-hydroxy-2-alkenals, and 4-hydroxy-2,6-alkadienals in exhaled breath condensate by UHPLC-HRMS and in breath by real-time HRMS *Anal. Chem.* **87** 3087–93