The metabolomic detection of lung cancer biomarkers in sputum

Simon J. S. Cameron[1], Keir E. Lewis[2,3], Manfred Beckmann[1], Gordon G. Allison[1], Robin Ghosal[2], Paul D. Lewis[3], and Luis A. J. Mur[1]*.

[1]Institute of Biological, Environmental and Rural Sciences, Edward Llywd Building, Penglais Campus, Aberystwyth, Ceredigion, SY23 3FG, UK.

[2]Department of Respiratory Medicine, Prince Phillip Hospital, Llanelli, SA14 8LY, UK.

[3]College of Medicine, Swansea University, Swansea, SA2 8PP, UK.

Running Title: Sputum Metabolomics for Lung Cancer Biomarkers

*** Corresponding Authors**: Professor Luis A. J. Mur. Institute of Biological, Environmental and Rural Sciences, Edward Llywd Building, Aberystwyth University, Penglais Campus, Aberystwyth, Ceredigion, SY23 3FG, Wales, UK. Phone: 01970 622981. Fax: 01970 622350. Email Address: lum@aber.ac.uk

Abstract

Objectives: Developing screening and diagnosis methodologies based on novel biomarkers should allow for the detection of the lung cancer (LC) and possibly at an earlier stage and thereby increase the effectiveness of clinical interventions. Here, our primary objective was to evaluate the potential of spontaneous sputum as a source of non-invasive metabolomic biomarkers for LC status.

Materials and Methods: Spontaneous sputum was collected and processed from 34 patients with suspected LC, alongside 33 healthy controls. Of the 34 patients, 23 were subsequently diagnosed with LC (LC^+, 16 NSCLC, six SCLC, and one radiological diagnosis), at various stages of disease progression. The 67 samples were analysed using flow infusion electrospray ion mass spectrometry (FIE-MS) and gas-chromatography mass spectrometry (GC-MS).

Results: Principal component analysis identified negative mode FIE-MS as having the main separating power between samples from healthy and LC. Discriminatory metabolites were identified using ANOVA and Random Forest. Indications of potential diagnostic accuracy involved the use of receiver operating characteristic / area under the curve (ROC/AUC) analyses. This approach identified metabolites changes that were only observed with LC. Metabolites with AUC values of greater than 0.8 which distinguished between LC^+/LC^- binary classifications where identified and included Ganglioside GM1 which has previously been linked to LC.

Conclusion: This study indicates that metabolomics based on sputum can yield metabolites that can be used as a diagnostic and/or discriminator tool. These could aid clinical intervention and targeted diagnosis of LC within an 'at risk' LC^- population group. The use of sputum as a non-invasive source of metabolite biomarkers may aid in the development of an at-risk population screening programme for lung cancer or enhanced clinical diagnostic pathways.

Key Words: Lung cancer, metabolomics, biomarkers, sputum, polyamines, gangliosides

1 Introduction

Lung cancer (LC) is the most prevalent cancer in the world; responsible for 1.3 million deaths annually [1]. The last 30 years has seen little improvement in the overall five year survival rate for LC; with only 15% of patients living for at least five years after their initial diagnosis [2]. These relatively poor survival rates are primarily a result of the late detection of a malignancy; reducing the success of clinical interventions. Clinicians currently rely on three main tools for LC diagnosis: X-ray, computerised tomography (CT) scans, and bronchoscopy. These methods have improved our ability to detect lung cancer, but have nevertheless failed to improve the rate of early LC detection [3]. Another aspect of this poor early detection is the association of LC with smoking, which masks some of the disease's early symptoms, which has been linked to approximately 90% of LC tumours [4] .

An alternative screening methodology to radiography, which is currently the most widely used approach, is the utilisation of molecular markers, both genetic and metabolomic, in biofluids. For example, microRNAs have been suggested as biomarkers for NSCLC in sputum [5], plasma [6], and serum [7]. Previous work by members of this research group has demonstrated that chemometric analysis combined with Fourier transform infrared spectroscopy is a non-invasive approach that allows for the discrimination of LC positive patients. This demonstrated that sputum could be used as a non-invasive source of biomarkers for LC [8]. However, analysis of mid-IR spectra only provides information on broad changes in classes of chemicals, and has a poor ability to resolve changes to particular chemicals. By comparison, metabolite profiling based on sample screening using Mass Spectrometry (MS) can resolve changes in individual chemicals and thus, could more readily identify biomarkers linked to LC detection.

The aim of this study was to employ MS metabolomic profiling to identify clinically relevant biomarkers in sputum that could be used for detect LC (diagnosis) as well as provide some pathophysiological insights based on the characteristics of the chemical biomarkers. We utilised two MS approaches in this study, Gas Chromatography MS (GC-MS) and Flow Infusion Electrospray MS (FIE-MS). Our rationale for this approach is that both MS technologies are widely used in biomarker discovery, but have differing levels of sensitivities and different approaches in regards to sample preparation and analysis. For example, GC-MS requires chemical

derivatization of sample metabolites prior to analysis whilst FIE-MS requires no pre-treatment [9]. Although, our study employed both univariate and multivariate approaches our study sought to conform to the demands of the TRIPOD (The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement by adhering to the recommended checklist [10]. We employed assessments of diagnostic accuracy based on receiver operating characteristic (ROC)/ Area under the Curve (AUC) that suggest that our approach could be used in clinical context to inform the detection of the disease. To the best of our knowledge, metabolomic profiles have not been reported using sputum as a biofluid from clinical patients. Thus, beyond, the detection of biomarkers, a description of the LC sputum metabolome offers a novel insight into the pathology of LC.

2 Materials and Methods

2.1 Ethics Statement

The MedLung observational study (UKCRN ID 4682) received loco-regional ethical approval from the Hywel Dda Health Board (05/WMW01/75). Written informed consent was obtained from all participants at least 24 hours before sampling, at a previous clinical appointment, and all data was link anonymised before analysis.

2.2 Study Design

This study aimed to compare the metabolomes of three groups of sputum samples. Two sets of sputum samples were obtained from patients referred to the access LC clinic at the Prince Phillip Hospital, Wales, UK; a site of primary care. Lung cancer status was subsequently assessed as part of the Medlung observational study (UKCRN ID 4682) and patients were classified as either LC^+ or diagnosed with another pulmonary disease (LC^-) based on histological assessments of sputum bronchoscopy derived samples (Table S1). Metadata including comorbidities, smoking history and drug history are given in Table S1. Additionally, spontaneous sputum samples were collected from staff members at Swansea University who had no previous history of cancer or lung disease, other than asthma. These non-clinical samples were designated as a control (CON) group. The design expensively exploited pair-wise analyses between LC+ and LC- groups and the CON group. As this project was seen as a pilot project, no external validation set, comprising, for example, testing on another set of patients samples was used. Further, the danger of over-fitting the derived data was reduced through the extensive use of simple two-way ANOVA in our pairwise comparisons. Sampling occurred between 2012-2013 to align with the Medlung study timeline and this, rather than an *a priori* design target, governed the number of samples analysed.

2.3 Patient Recruitment and Sampling

Spontaneous sputum was collected from referrals to our rapid access LC clinic at the Prince Phillip Hospital, Wales, UK or volunteers from staff members at Swansea University. No *a priori* criteria were applied to the selection of patients or volunteers other than their ability to produce sputum. Patients were asked to cough into sterile, 50 mL polypropylene tubes (Greiner Bio-One Ltd, UK) prior to bronchoscopy, to at total volume of

2-3 mL. A 100 µL aliquot of all samples, including the CON group, was taken to create a second pellet that was subsequently formalin fixed and wax embedded prior to sectioning and staining with haemotoxylin and eosin (H&E). To confirm samples were of bronchial origin, H&E stained sections were assessed by a consultant histopathologist for presence of bronchial epithelial cells. Histological assessments of the LC^+ class allowed the recording of LC type and stage. Thus, NSCLC classifications were obtained for sixteen samples and six were SCLC. Only in one case (LC06) was no classification obtained. Within the NSCLC samples, seven could be sub-classified as adenocarcinoma type and five squamous cell types. Considering the LC^- classified samples, three were diagnosed with chronic obstructive pulmonary disease (COPD) and two with pneumonia. Amongst the LC^+ group, only two (LC07, LC20) were diagnosed with COPD which could be considered a LC co-morbidity and none with pneumonia.

2.4 Processing of Raw Sputum

In line Raw sputum samples were frozen at -80 °C and defrosted in ice for approximately two hours when required. Sputum cells were isolated by adding 0.5 mL of a working solution of dithiothreitol (DTT), made up by adding 2.5 g of DTT to 31 mL of 30% aqueous methanol, and 5 mL of 30% aqueous methanol. The samples were then placed on a vortex mixer for 15 minutes and underwent centrifugation at 1,800 x g for 10 minutes. The supernatant was removed and the pellet used in subsequent metabolomic profiling.

2.5 Flow Infusion Electrospray Mass Spectrometry (FIE-MS)

After processing, 20 µL of the sputum pellet was added to 20 µL of ultrapure water and 40 µL of ice-cold HPLC grade acetone. Samples were vortex mixed for five seconds, cooled on ice for 30 minutes, and then underwent centrifugation at 11,000 x g for five minutes. After centrifugation, 50 µL of the supernatant was removed and 250 µL of 70% methanol (made up using HPLC grade methanol and ultrapure water) was added. Glass vials were capped and analysed in random order on a LTQ linear ion trap (Thermo Electron Corporation). Data were acquired in alternating positive and negative ionization modes over 4 scan ranges (15–110, 100–220, 210–510, and 500–1200 *m/z*), with an acquisition time of five minutes. The resulting mass spectrum was the mean of 20 scans about the apex of the infusion profile.

2.6 Gas Chromatography Mass Spectrometry (GC-MS)

The sputum pellet was processed as described in section 2.4 and 50 µL of the supernatant after centrifugation removed and dried using a DNA SpeedVac (Savant, USA) at 40°C. After removal of all liquid, 30 µL of a 20 mg/ml solution of methoxyamine dissolved in pyridine was added and each sample was transferred to a 11 mm diameter glass GC vials which were capped with Teflon crimp caps and incubated at 90°C for 15 minutes. After cap removal, 20 µL of N,O-Bis(trimethylsilyl)trifluoroacetamide (BSTFA) was added to the sample, alongside 5 µL of an alkane standard mix. This mixture comprised of C_{10}, C_{13}, C_{15}, C_{18}, C_{19}, C_{23}, C_{28}, C_{32} and C_{36} alkanes dissolved in pyridine each at a concentration of 2 µL/mL (for alkanes liquid at room temperature) or 2 mg/mL (for alkanes solid at room temperature). The vials were recapped and incubated at 90°C for 15 minutes. Samples were analysed by duplicate injection on a 6890N GC linked to a 5973N mass analyser and a 7683 auto-sampler (Agilent Technologies) fitted with a Thermo Scientific TR-Pesticide II fused silica column (30m x 0.25 mm ID x 0.25 µm film thickness). Helium carrier gas was supplied at a constant flow rate of 1 mL per minute and following the injection of 1 µl of sample the GC oven was held at 80°C for three minutes, increased to 280°C at a rate of 15°C per minute, and then to 330°C at a rate of 50°C per minute. The inlet temperature was 280°C and samples were split with a 2:1 split ratio. The temperature of the MS transfer line was 330°C.

2.7 Accurate Mass Determination

Selected discriminatory nominal mass signals were investigated further by targeted nano-flow Fourier Transform-Ion Cyclotron Resonance Ultra-Mass-Spectrometry (FT-ICR-MS) using TriVersa NanoMate (Advion BioSciences Ltd) on a LTQ-FT-ULTRA (Thermo Scientific) to obtain ultra-high accurate mass information and MSn ion-trees [11]. Resulting accurate mass values were used to interrogate the Human Metabolome Database [12]. Based on an accuracy of 1 ppm for the FT-ICR-MS, the top ranking metabolite with this range indicated as the identification for each discriminatory negative ionisation mode FIE-MS metabolite.

2.8 Data and Statistical Analysis

All GC-MS data pre-treatment procedures, including baseline correction, chromatogram alignment, and data compression were performed by using custom scripts in Matlab version 6.5.1 (The Math Works Inc). Targeted peak lists were generated, and peak apex intensities of each characteristic mass in a retention time window

were saved in an intensity matrix (run x metabolite). FIE-MS data was normalised with the total ion count for each sample used to transform the intensity value for each metabolite in to a percentage of the total ion count, after the removal of metabolites below 50 m/z. Principal Component Analyses (PCA). Briefly, this involves projecting a ("X") matrix formed from set of (N) spectra with a given mass range (p) (i.e. a N x p matrix) onto multi-dimensional space. Principal components (PC) are linear combinations of original variables (known as loadings) which are used in the projection of the X matrix. Individual PCs are ranked (PC1, PC2…etc) on the basis of the variance within the original data-set that is explained. PCA is an unsupervised method where no *a priori* knowledge of experimental structure is given. Thus, if there is clustering of either 2D or 3D projections of PCA from replicate data this indicates that the original experimental parameters are the sources of maximal variation. Hierarchical Cluster Analyses (HCA) with heat maps, and Random Forest (RF) multivariate analysis were completed using the PyChem (Version 3.0.5g Beta) package [13] and/or MetaboAnalyst 2.0 [14]. ROC (Receiver Operating Characteristic) curve analyses plot the true positive rate (Sensitivity) in function of the false positive rate (Specificity) and the validity of the fit is indicated based on area under curve (AUC) calculations. ROC-AUC analyses used the ROC Curve Explorer and Tester (ROCCET) online platform [15] to assess our standard binary classification tests. Due to the exploratory nature of this pilot study, no external validation set consisting of an independent population of samples was available to be included in (e.g.) the ROC analyses.

3 Results

Patients and participants sampled as part of this study are summarised in Table 1, with individual sample data in Supplementary Tables 1a and 1b. A total of 34 patients with suspected LC were recruited, with 23 confirmed with LC (LC^+) (16 NSCLC (nine Stage 4, three Stage 3A, one Stage 3B, three Stage 2B, and one Stage 1B), six SCLC (three extensive and three limited), and one receiving a clinic-radiological diagnosis made by the LC multidisciplinary team), and 11 had no diagnosis of LC after extensive testing and follow up for at least one year (LC^-). In addition, a total of 33 non-clinical controls (CON) were collected from participants with no history of clinical lung disease.

Metabolomic profiles of the sputum samples were acquired using FIE-MS (in negative and positive ionization modes) and GC-MS platforms and analysed using PCA. Both MS platforms were examined as although GC-MS is widely employed in metabolomics profiling, it lacks the sensitivity of FIE-MS and thus, the latter could yield a more comprehensive data set [16]. PCA indicated that the metabolomic profile acquired in negative ionisation FIE-MS mode (Figure 1a) showed the greatest degree of separation between the three sample groups (LC^+/LC^-/CON). Such a separation was not evident in positive FIE-MS mode (Figure 1b), and only partially exhibited in the analyses of the GC-MS profiles (Figure 1c) suggestive of the value of the greater sensitivity of the FIE-MS approach and platform. The FIE-MS^- metabolites were then analysed using one-way ANOVA which identified the top 25 metabolites based on their discriminatory ability whose levels significantly differed between the sample groups. Derivation of a HCA with heat map based on these top 25 metabolites also demonstrated that the LC+ and LC- could be readily separated from the CON group (Figure 2). Furthermore, many LC^+ samples clustered together.

Whilst simple analyses such as PCA or ANOVA could distinguish between the LC^+/LC- class and CON, supervised analyses, where *a priori* information of the sample classes was required, would be need to identify variable between the LC^+ and LC^- classes. Due to the separation shown with FIE-MS^- metabolites into clinically relevant classes these datasets were used to identify clinical relevant metabolomic biomarkers. Random Forest (RF) analyses were then used to indicate a number of metabolites which differentiated between the experimental

classes (Figure 3). Metabolites which were either increased or decreased in the LC^+ or CON classes compared to the LC^- class which were taken forwards to identification by high resolution MS.

ROC -AUC analyses were also used to identify discriminatory metabolites. The top five metabolites for each differential comparison are listed in Table 2 with the AUC figure and box and whisker distributions of the data for the top differential metabolites shown in Figure 4. *t*-tests of the targeted metabolites indicated a high level of significance in each comparison. These identified a number of metabolites that had a high AUC value (>0.99) for differentiating between non-clinically (CON- class) and clinically acquired (LC^-/LC^+ classes) samples. Four metabolites were identified with an AUC value of greater than 0.80, a threshold for clinically useful prediction.

To identify the mass-ions targeted by RF and ROC-AUC analyses, high resolution MS using FT-ICR-MS was employed. Metabolites were identified, where possible, based on this accurate mass profiling and database interrogations, (Table 3) and these were used to annotate the analyses shown in Figures 3 and 4. Examination of the metabolites listed in Supplementary Table 2 includes those involved in polyamine (putrescine), amino acid, and lipid metabolism. Clinical samples (LC^+/LC^-) appeared to be separated from CON sample through differential processing of polyamine metabolites; putrescine and N,N,N-Trimethylethenaminium, and lipid metabolites, including glycerophospholipids of the cardiolipin (PC) class, and isobutyl decanoate and diethyl glutarate. Separation between the clinical samples (LC^+ and LC^- classes) appeared to be due to elevated levels of metabolites identified as hexanal, cysteic acid, hydroxypyruvic acid, and the cholesterol ester with an acyl group CE (22:5(4Z,7Z,10Z,13Z,16Z)). The mass-ion 1496.72 showing the highest AUC value (0.85) was identified as the ganglioside GM1 (18:1/12:0).

4 Discussion

Since the 'Warburg Effect' was first described in 1956 [17], the alterations that cells undergo during carcinogenesis has been a focus of both basic and applied clinical research. To date, the majority of metabolomic lung cancer studies appear to have focussed on the cancerous tumours themselves or serum from affected patients, using a limited range of MS techniques [18]. Here, we suggest that the sputum of patients can be used as a non-invasive source of biomarkers for the identification of LC status.

Sputum represents a biofluid that could be readily accessed from the target group and the results of this study indicate it could be used as a biofluid matrix for an efficient LC screen. We used two mass-spectrometry platforms; the widely employed GC-MS and also FIE-MS on the same sample set to allow comparison of the discriminatory power of both. These results suggested that derivatisation (in the case of GC-MS) or the wide range of adducts formed with positive ionisation using FIE-MS (as opposed to negative, ionisation where simple proton loss [M^-—H^+] is predominant) can obscure screens of sputum.

Analyses of FIE-MS^- data allowed identification of clinically relevant groupings and both PCA and HCA could separate a "healthy" control samples from samples taken from clinically-referred patients. Although, not all of these patients were subsequently confirmed to be LC^+, the LC^- group had symptoms necessitating referral and thus, should be considered to be "unwell". Even at this level, a non-invasive and rapid test of lung health would be useful to the medical community.

Random Forest analysis appeared to be particularly effective in discriminating between LC^+ and CON samples; with LC^- samples between these extremes. We coupled RF analyses with assessment of ROC using AUC analysis; which has been widely used to determine the diagnostic value of biomarkers. Here, the False Discovery rate vs. True Discovery rate compared a series of binary tests between our three sample groups. The CON group was highly distinctive, with ROC-AUC analyses detecting metabolites with extremely high AUC values. Crucially, a number of FIE-MS^- metabolites that had AUC values greater than 0.80 when comparing LC+/LC, a cut-off for discrimination that may be useful in a clinical setting were identified. This equates to a false discovery rate of

under 20% for these metabolites although care needs to be taken with this figure and it requires confirmation with external validation datasets to remove any danger of "overfitting" i.e. deriving a model which describes random error or noise rather than any true relationship. Our LC^+ group consisted of a range of LC stages and histology, suggesting that biomarkers established through metabolomic profiling techniques could have utility as a preliminary screen, identifying patients for clinical follow-up for LC confirmation, histology and staging.

Considering the identities of metabolites separating the CON and LC- class, the increases in putrescine were interesting because polyamines are essential for normal mammalian cell growth. Polyamine metabolism is frequently dysregulated in cancer and has emerged as a target for therapeutic intervention [19]. However, as polyamines did not discriminate between the CON and LC+ or LC- and LC+ classes, we were not able to associate these polyamine changes with LC in this study. Therefore, changes in polyamines may have reflected changes linked to an inflammatory response and/or cell death; which may reflect pathogen attack or polyamine catabolism which can generate reactive oxygen species (ROS) [20].

Also prominent in the clinical samples (LC+/LC-) compared to the CON class were lipid metabolites, including glycerophospholipids of the cardiolipin (PC) class as well as isobutyl decanoate and diethyl glutarate. Cardiolpins are major components of the inner mitochondrial and is particularly susceptible to ROS attack due to its high content of unsaturated fatty acids. Increased ROS would affect mitochondrial membrane fluidity, possibly resulting in cardiolipin release and possibly leading to the greater than two fold increases that we have detected in our study. Cardiolipin-associated changes in membrane fluidity have been associated with reduced mitochondrial oxidative phosphorylation efficiency and apoptosis [21]. In this context, it is relevant that isobutyl decanoate and diethyl glutarate, as potential phospholipid fragments, could represent the products of lipid peroxidation and as they exhibited a 4.69 and 3 fold increase, respectively, in the LC^+ class compared to CON.

Identifying the metabolite changes in the LC^+ samples targeted by RF and ROC-AUC, there appeared to be higher levels of hexanal, cysteic acid, hydroxypyruvic acid, and one metabolite without accurate mass identification, and eleven metabolites with lower levels. Hexanal has previously been shown to be elevated in blood samples from lung cancer patients [22], suggesting its validity as a LC biomarker. To our knowledge there have been no

reports of cysteic acid or hydroxypyruvic acid being targeted as LC biomarkers. These could suggest alterations in cysteine metabolism (in the case of cysteate) or glycolysis (in the case of the pyruvic acid derivative) are being targeted in our metabolomic analyses. The potential relevance of the cholesteryl docosapentaenoate; CE(22:5(4Z,7Z,10Z,13Z,16Z)) is unknown, but its increase could reflect membrane disruption. Of particular interest was ganglioside GM1 (18:1/12:0) which represents a glycosphingolipid inked to a single sialic acid through its sugar group. Gangliosides have primarily been studied in neural tissues, but can be found in most cell types where they are involved in cell–cell recognition, cell–matrix attachment, cell growth and cell differentiation. Interestingly, ganglioside GM1 has already been associated with LC and particularly with SCLC due to a tendency to arise from neuroectodermal tissue [23]. Indeed, GM1 ganglioside-fused to hemocyanin has been used to specifically target SCLC tissue in patients [24]. Cholera toxin which is known to target GM1 ganglioside was found to suppress the growth of 9 out of 15 SCLC cell lines with those resistant to the toxin exhibiting reduced GM1 ganglioside expression [25]. Taken together with our results, ganglioside GM1 could be a good candidate for biomarker based LC screens.

5 Conclusions

As far as we can ascertain, this is the first study to report on the metabolomic profiling of sputum acquired from LC patients. The use of sputum, the production of which is symptomatic of LC, as a biofluid for screening carries the benefit of being non-invasive, high-throughput, and low-cost, compared to current conventional methods such as CT scan [26]. It may be that a combination of metabolomic biomarkers and other types, such as circulating miRNAs, would allow for an integrated approach to LC screening, as has been suggested for other cancers [27]. Here, we have shown the power of using metabolomics to identify biomarkers with potential clinical application for LC. Further work, using a larger patient cohort, will be required to ascertain the utility of metabolomic biomarkers for LC stage and histological subtype.

Acknowledgements: SJSC is grateful for studentship from Aberystwyth University. The MedLung Study is funded by a grant from the National Institute of Social Care and Health Research (NISCHR), Wales. The sponsor was Hywel Dda University Health Board and neither the funders – Aberystwyth University or NISCHR - nor sponsor had any input into the design or reporting of the study. We wish to thank Dr Paul Griffiths, Consultant histopathologist, for sputum samples assessment, Dr Sion Bayliss for collection of healthy age-matched control samples, and Kathleen Tailliart for running samples through mass spectrometry. We are also highly appreciative of the constructive criticisms provided by the anonymous reviewers which improved this manuscript.

References

[1] WHO, Cancer Factsheet, WHO Fact Sheets, Number 2970. (2013). http://www.who.int/mediacentre/factsheets/fs297/en/ (accessed May 29, 2013).

[2] A. Jemal, R. Siegel, J.Q. Xu, E. Ward, Cancer Statistics, 2010, CA A Cancer J. Clin. 60 (2010) 277–300. doi:10.1002/caac.20073.

[3] G. Sutedja, New Techniques for Early Detection of Lung Cancer., Eur. Respir. J. 39 (2003) 57s–66s. http://www.ncbi.nlm.nih.gov/pubmed/12572703 (accessed February 25, 2014).

[4] P. Jha, R. Peto, W. Zatonski, J. Boreham, M.J. Jarvis, A.D. Lopez, Social Inequalities in Male Mortality, and in Male Mortality from Smoking: Indirect Estimation from National Death Rates in England and Wales, Poland, and North America., Lancet. 368 (2006) 367–70. doi:10.1016/S0140-6736(06)68975-7.

[5] Y. Xie, N.W. Todd, Z. Liu, M. Zhan, H. Fang, H. Peng, et al., Altered miRNA Expression in Sputum for Diagnosis of Non-Small Cell Lung Cancer., Lung Cancer. 67 (2010) 170–6. doi:10.1016/j.lungcan.2009.04.004.

[6] J. Shen, N.W. Todd, H. Zhang, L. Yu, X. Lingxiao, Y. Mei, et al., Plasma microRNAs as Potential Biomarkers for Non-Small-Cell Lung Cancer., Lab. Investig. 91 (2011) 579–87. doi:10.1038/labinvest.2010.194.

[7] K.M. Foss, C. Sima, D. Ugolini, M. Neri, K.E. Allen, G.J. Weiss, miR-1254 and miR-574-5p: Serum-Based microRNA Biomarkers for Early-Stage Non-Small Cell Lung Cancer., J. Thorac. Oncol. 6 (2011) 482–8. doi:10.1097/JTO.0b013e318208c785.

[8] P.D. Lewis, K.E. Lewis, R. Ghosal, S. Bayliss, A.J. Lloyd, J. Wills, et al., Evaluation of FTIR Spectroscopy as a Diagnostic Tool for Lung Cancer Using Sputum, BMC Cancer. 10 (2010). doi:64010.1186/1471-2407-10-640.

[9] J. Draper, A.J. Lloyd, R. Goodacre, M. Beckmann, Flow infusion electrospray ionisation mass spectrometry for high throughput, non-targeted metabolite fingerprinting: a review, Metabolomics. 9 (2012) 4–29. doi:10.1007/s11306-012-0449-x.

[10] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G.M. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement., BMC Med. 13 (2015) 1. doi:10.1186/s12916-014-0241-z.

[11] A.J. Lloyd, G. Favé, M. Beckmann, W. Lin, K. Tailliart, L. Xie, et al., Use of Mass Spectrometry Fingerprinting to Identify Urinary Metabolites after Consumption of Specific Foods., Am. J. Clin. Nutr. 94 (2011) 981–91. doi:10.3945/ajcn.111.017921.

[12] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y. Liu, et al., HMDB 3.0: The Human Metabolome Database in 2013., Nucleic Acids Res. 41 (2013) D801–7. doi:10.1093/nar/gks1065.

[13] R.M. Jarvis, D. Broadhurst, H. Johnson, N.M. O'Boyle, R. Goodacre, PYCHEM: A Multivariate Analysis Package for Python., Bioinformatics. 22 (2006) 2565–6. doi:10.1093/bioinformatics/btl416.

[14] J. Xia, R. Mandal, I. V Sinelnikov, D. Broadhurst, D.S. Wishart, MetaboAnalyst 2.0: A Comprehensive Server for Metabolomic Data Analysis., Nucleic Acids Res. 40 (2012) W127–33. doi:10.1093/nar/gks374.

[15] J. Xia, D.I. Broadhurst, M. Wilson, D.S. Wishart, Translational Biomarker Discovery in Clinical Metabolomics: An Introductory Tutorial., Metabolomics. 9 (2013) 280–299. doi:10.1007/s11306-012-0482-9.

[16] L.W. Sumner, P. Mendes, R.A. Dixon, Plant metabolomics: large-scale phytochemistry in the functional genomics era, Phytochemistry. 62 (2003) 817–836. doi:10.1016/S0031-9422(02)00708-2.

[17] O. Warburg, On the Origin of Cancer Cells, Science (80-.). 123 (1956) 309–314. doi:10.1126/science.123.3191.309.

[18] S. Hori, S. Nishiumi, K. Kobayashi, M. Shinohara, Y. Hatakeyama, Y. Kotani, et al., A Metabolomic Approach to Lung Cancer., Lung Cancer. 74 (2011) 284–92. doi:10.1016/j.lungcan.2011.02.008.

[19] A.E. Pegg, Mammalian Polyamine Metabolism and Function., IUBMB Life. 61 (2009) 880–94. doi:10.1002/iub.230.

[20] M.H. Park, K. Igarashi, Polyamines and their Metabolites as Diagnostic Markers of Human Diseases., Biomol. Ther. (Seoul). 21 (2013) 1–9. doi:10.4062/biomolther.2012.097.

[21] G. Paradies, G. Petrosillo, V. Paradies, F.M. Ruggiero, Role of Cardiolipin Peroxidation and Ca2+ in Mitochondrial Dysfunction and Disease., Cell Calcium. 45 (2009) 643–50. doi:10.1016/j.ceca.2009.03.012.

[22] C. Deng, X. Zhang, N. Li, Investigation of volatile biomarkers in lung cancer blood using solid-phase microextraction and capillary gas chromatography-mass spectrometry., J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci. 808 (2004) 269–77. doi:10.1016/j.jchromb.2004.05.015.

[23] T. Brezicka, B. Bergman, S. Olling, P. Fredman, Reactivity of Monoclonal Antibodies with Ganglioside Antigens in Human Small Cell Lung Cancer Tissues, Lung Cancer. 28 (2000) 29–36. doi:10.1016/S0169-5002(99)00107-5.

[24] L.M. Krug, G. Ragupathi, C. Hood, M.G. Kris, V.A. Miller, J.R. Allen, et al., Vaccination of patients with small-cell lung cancer with synthetic fucosyl GM-1 conjugated to keyhole limpet hemocyanin., Clin. Cancer Res. 10 (2004) 6094–100. doi:10.1158/1078-0432.CCR-04-0482.

[25] R. Fuentes, R. Allman, M.. Mason, Ganglioside expression in lung cancer cell lines, Lung Cancer. 18 (1997) 21–33. doi:10.1016/S0169-5002(97)00049-4.

[26] M.P. Rivera, A.C. Mehta, M.M. Wahidi, Establishing the Diagnosis of Lung Cancer: Diagnosis and Management of Lung Cancer, Chest. 143 (2013) e142S–65S. doi:10.1378/chest.12-2353.

[27] B. Laxman, D.S. Morris, J. Yu, J. Siddiqui, J. Cao, R. Mehra, et al., A First-Generation Multiplex Biomarker Analysis of Urine for the Early Detection of Prostate Cancer., Cancer Res. 68 (2008) 645–9. doi:10.1158/0008-5472.CAN-07-3224.

Tables and Legends

Table 1 | Summarised Patient and Participant Information

Summarised patient information detailing clinical data. Full clinical data for clinically acquired samples, and information collected for healthy control participants, are fully detailed in Supplementary Table 1.

	Non-Clinical Controls (CON)	LC Negative (LC-)	LC Positive (LC+)
Number	33	11	23
Age	55.3 (14.6)	66.5 (14.3)	66.6 (8.1)
Gender			
Male	20	10	11
Female	13	1	12
Smoking Status			
Current	15	3	10
Ex	0	8	10
Never	18	0	3
Smoking Pack Years	NC	49.0 (34.9)	39.3 (18.9)
Infection Present			
Yes	NC	3	1
No	NC	8	22
CO Level (ppm)	NC	3.7 (1.3)	4.2 (2.8)

Table 2 | Top Five Area Under Curve Values for Negative FIE-MS Mode Metabolites

Using the online ROCCET platform, the top five metabolites, based on AUC values, for each differential group comparison were identified. For clinical and non-clinical comparisons, high AUC values were obtained, and for the LC negative and positive comparison, a number of metabolites were identified with AUC values greater than 0.8. AUC range refers to the 95% confidence intervals of the true AUC value as given by the ROCCET platform.

Differential	Metabolite	AUC Value	True AUC Range	t-Test	Fold Change
CON Vs LC-	N,N,N-Trimethylethenaminium /CL(16:1(9Z)/18:1(11Z)/16:1(9Z)/18:1(9Z))	1.00	0.989 - 1.000	4.47×10^{-15}	-2.38
	N,N,N-Trimethylethenaminium /1560.81	1.00	0.983 – 1.000	3.10×10^{-15}	-2.26
	Putrescine/CL(16:1(9Z)/18:1(11Z)/16:1(9Z)/18:1(9Z))	0.99	0.975 – 1.000	6.09×10^{-15}	-2.25
	Putrescine/1560.81	0.99	0.975 – 1.000	3.64×10^{-15}	-2.14
	53.27/1209.45	0.99	0.967 – 1.000	1.79×10^{-14}	-2.28
LC+ Vs CON	53.27/Isobutyl decanoate	1.00	0.993 – 1.000	8.74×10^{-24}	4.71
	Putrescine/Isobutyl decanoate	1.00	0.992 – 1.000	7.42×10^{-24}	4.69
	189.09	1.00	0.987 – 1.000	6.12×10^{-20}	2.57
	Diethyl glutarate	0.99	0.979 – 1.000	1.42×10^{-18}	3.00
	Cysteamine	0.99	0.980 – 1.000	3.03×10^{-20}	2.35
LC+ Vs LC -	Ganglioside GM1 (18:1/12:0)	0.85	0.709 – 0.953	2.93×10^{-3}	-0.03
	957.36	0.83	0.680 – 0.953	4.57×10^{-3}	0.31
	1382.45	0.83	0.668 – 0.957	5.93×10^{-4}	0.07
	CE(22:5(4Z,7Z,10Z,13Z,16Z))	0.82	0.644 – 0.947	1.28×10^{-2}	0.00
	1434.00	0.81	0.621 – 0.947	9.84×10^{-4}	0.14

Figures

Figure 1 | Principal Component Analysis Plots for FIE-MS and GC-MS Metabolites

PCA, based on metabolites acquired in (a) FIE-MS negative mode, (b) FIE-MS positive mode, and (c) GC-MS, clearly differentiates between the clinically and non-clinically acquired samples, though separation of the two clinical groups, lung cancer and symptom controls, does not occur. For (c), coordinate markers are means of individually calculated coordinates from duplicate GC-MS runs.

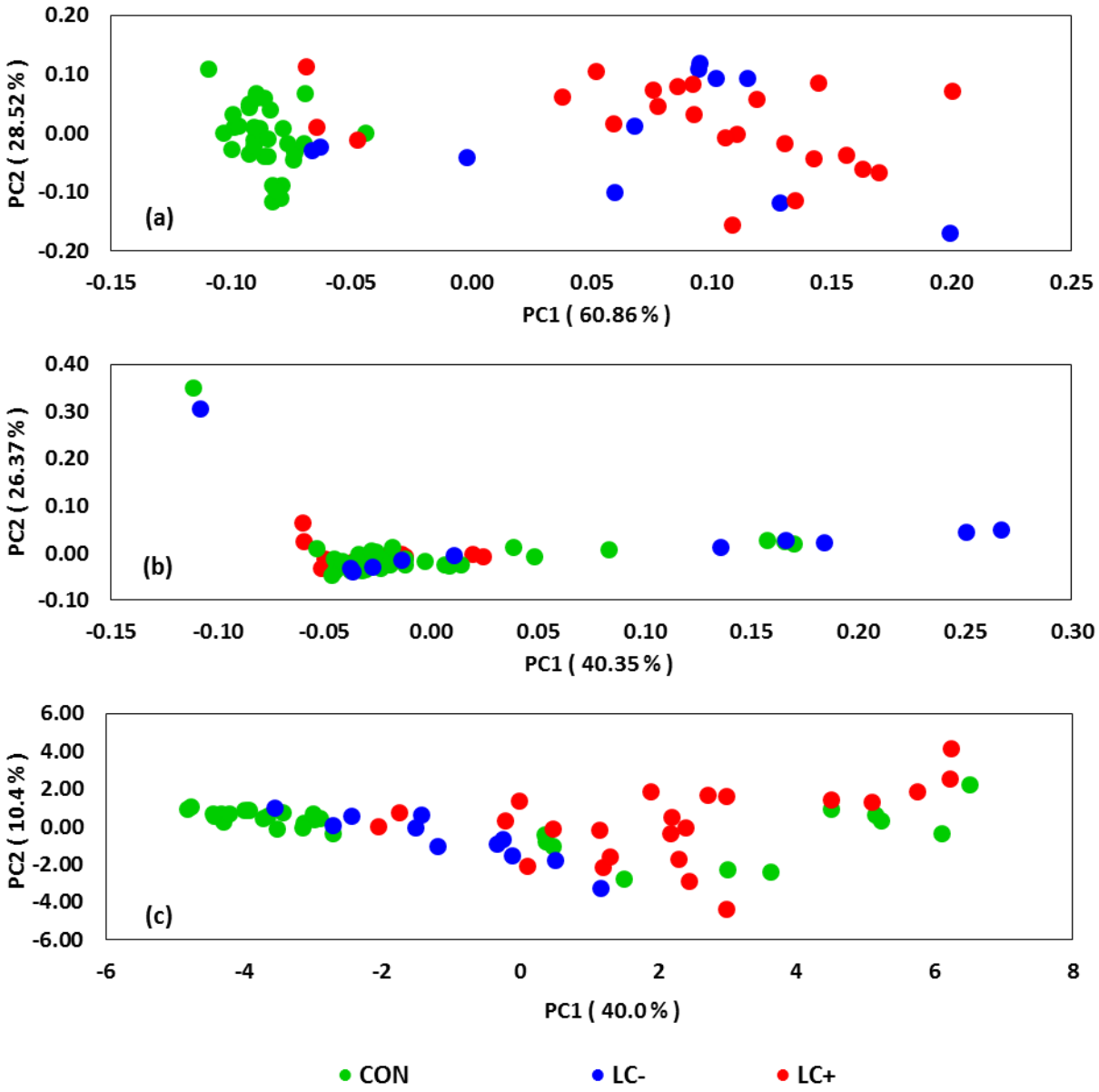

Figure 2 |Hierarchical Cluster Analysis with Heat Mapping for Negative Ionisation FIE-MS

Hierarchical cluster analysis and corresponding heat maps were constructed, based on the top 25 metabolites identified through one-way ANOVAs, for metabolites identified in FIE-MS negative ionisation mode. Similarly to PCA plots, separation between the clinically and non-clinically acquired samples was clear, but separation between LC positive and negative samples was not evident.

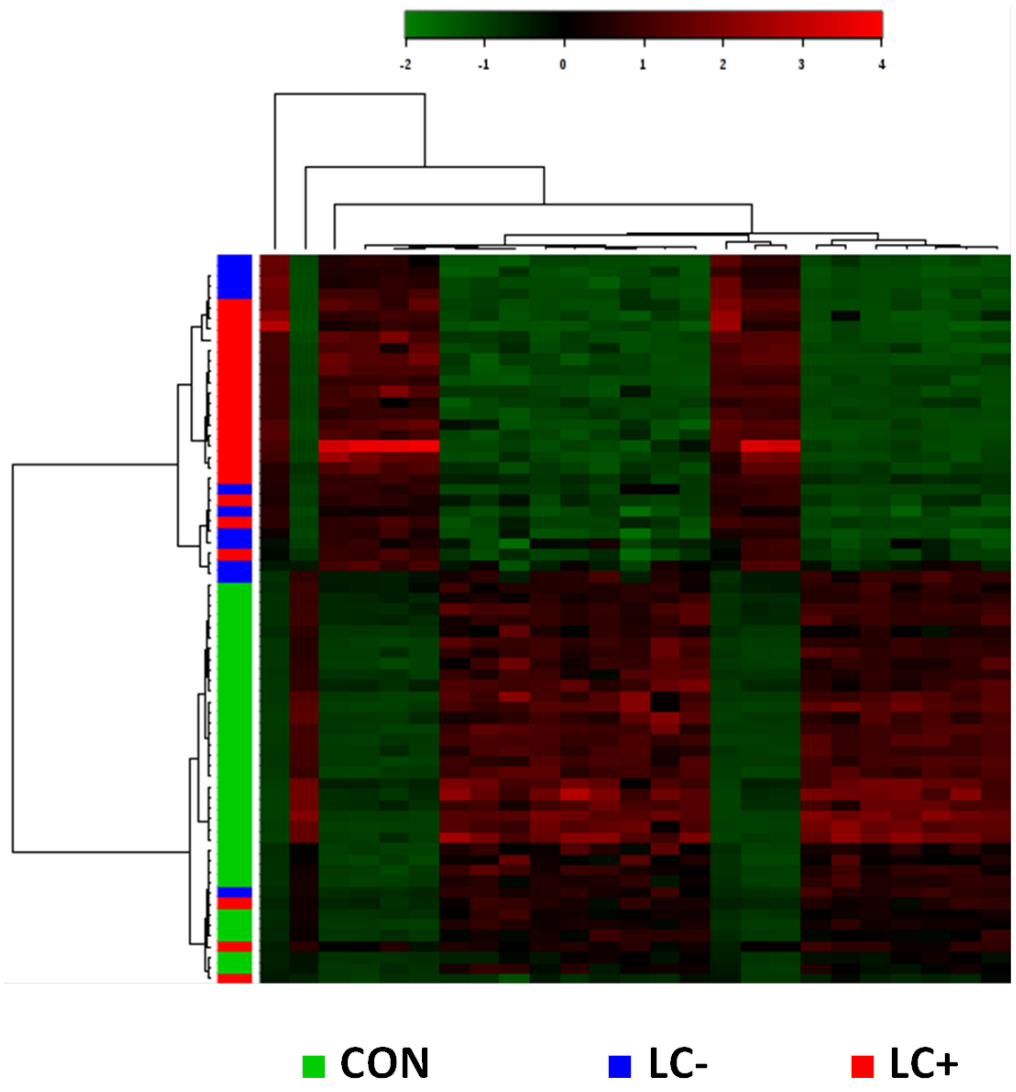

Figure 3| Random Forest Plots for Identification of Key FIE-MS Metabolites

Random forests plots were constructed, using MetaboAnalyst 2.0 for negative ionisation FIE-MS mode, which revealed a number of metabolites which may have potential in terms of diagnostic markers, particularly those that are either higher or lower in the LC positive group.

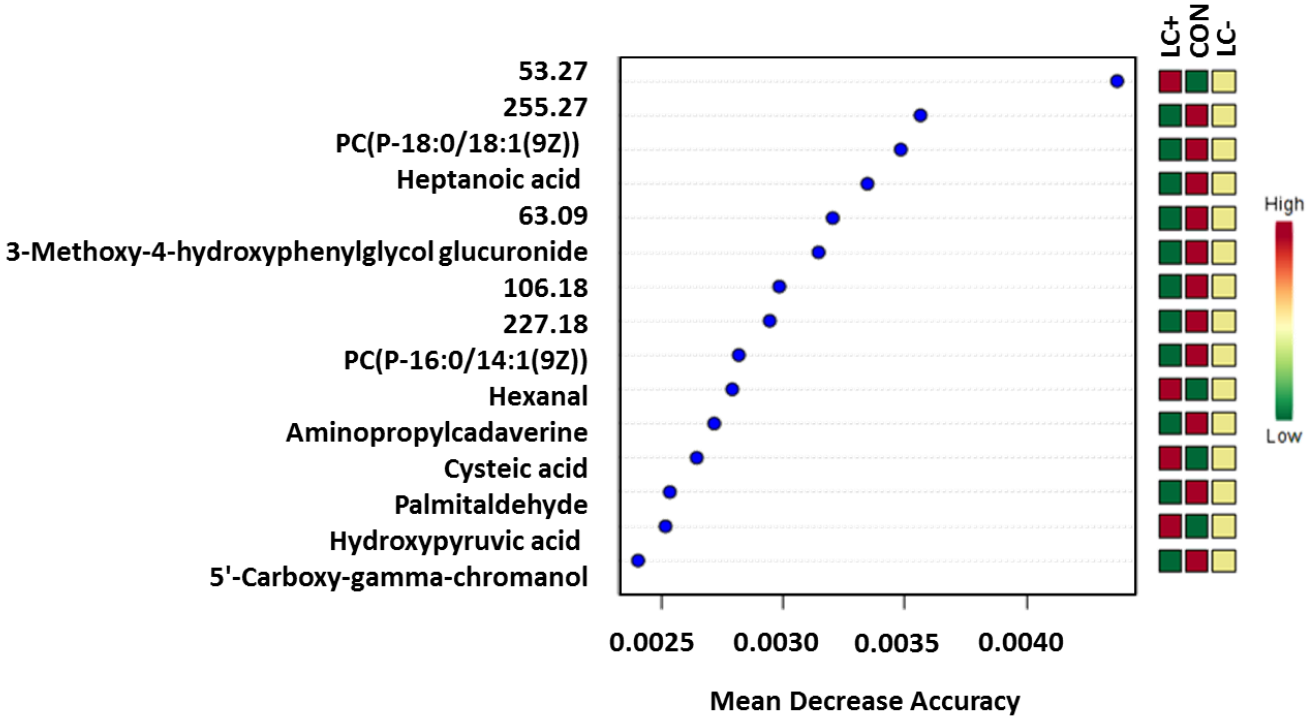

Figure 4| Univariate Receiver Operating Characteristic Curve Analyses for Biomarker Identification

Using the online facility, ROCCET, univariate receiver operating characteristic curves (ROC) were created, and plotted to create area under the curve (AUC) figures for metabolites identified in negative ionisation FIE-MS mode. The metabolite with the highest AUC value for each differential group is plotted.