Electronic Supplementary Information for

Combining MALDI-MS with Machine Learning for Metabolomic Characterization of Lung Cancer Patient Sera

Xiaopin Lai[a, †], Kunbin Guo[b, †], Wei Huang[a, †], Yang Su[a], Siyu Chen[a], Qiongdan Li[a], Kaiqing Liang[a], Wenhua Gao[a], Xin Wang[b], Yuping Chen[b*], Hongbiao Wang [b], Wen Lin[b*], Xiaolong Wei[b], Wenxiu Ni[c], Yan Lin[d], Daizhi Jiang[e], Yu-Hong Cheng[f], Chi-Ming Che[f] and Kwan-Ming Ng[a*]

Corresponding authors:

kwanming@stu.edu.cn; stchenyp@hotmail.com; 1263811129@qq.com

[†] **These authors contribute equally in this work.**

[a] Department of Chemistry and Key Laboratory for Preparation and Application of Ordered Structural Materials of Guangdong Province, Shantou University, Guangdong, 515063, P. R. China. E-mail: kwanming@stu.edu.cn

[b] The Cancer Hospital of Shantou University Medical College, Guangdong, 515041, P. R. China. E-mail: stchenyp@hotmail.com; 1263811129@qq.com

[c] Department of Medical Chemistry, Shantou University Medical College, Shantou, Guangdong, 515041, P. R. China.

[d] The Second Affiliated Hospital of Shantou University Medical College, Guangdong, 515041, P. R. China.

[e] Department of Computer Science, College of Engineering, Shantou University, Guangdong, 515063, P. R. China.

[f] Department of Chemistry and State Key Laboratory of Synthetic Chemistry, The University of Hong Kong, Hong Kong S.A.R., P. R. China.

Table of Contents

Text S1: Supplementary Methods

Mass Spectrometer Settings and Acquisition Parameters. The mass spectrometric measurements were performed with an Autoflex Speed MALDI TOF/TOF mass spectrometer equipped with a 355 nm solid-state smartbeam Nd:YAG laser. The linear positive mode was adopted to improve the detection sensitivity. The instrumental parameters were optimized as follows: ion source 1 at 19.50 kV; ion source 2 at 17.90 kV; lens voltage at 6.00 kV; detector voltage at 2.90 kV; pulsed ion extraction, 20 ns; Acquisition mode, random walk; total laser shot number at each sample well, 40,000; laser shot number at each raster position, 100; laser shot frequency, 500 Hz; and acquired mass range, m/z 0-1500. The laser was adjusted to 65% of the maximum power. The vacuum pressure was kept at around 10^{-6} to 10^{-7} mbar in the source and 10^{-7} to 10^{-8} mbar in the analyzer. The instrument was controlled via Bruker Daltonics flexControl 3.4 software. The reflectron positive mode was adopted to measure the accurate mass of distinctive features. Mass calibration was performed with the internal standard calibration mixture with mass precision of 30 ppm, and the mass resolution (at m/z 361) is approximately 7000 (FWHM).

MALDI-TOF Data Processing. MALDI-TOF raw data were converted to mzML with software ProteoWizard MSConvert and then processed with R packages MALDIquant and MALDIquantForeign. The log2 transformation was applied, followed by SavitzkyGolay smoothing, and SNIP baseline correction. The mass value alignment was performed with the alignSpectra command. Before peak detection, the six technical replicates were averaged with the averageMassSpectra command. Then, the peak detection was conducted with a signal-to-noise ratio of 3 and a half window size of 20. Peaks were binned with the binPeaks command with a tolerance of 0.0009. Peak filtration was applied with the filterPeaks command to keep the peaks with frequency ≥25% in all spectra of a group (lung cancer patients or healthy controls). Finally, the obtained data matrices were used for the following analysis.

Feature selection. The matrix of peak intensities was subjected to normalization with MSTUS (MS total useful signal) method with a "home-built" macro in Excel. The partial least squares-discriminant analysis (PLS-DA) was performed using Metaboanalyst 5.0 (McGill University, Montreal, Canada). The variable importance for the projection (VIP) identified by PLS-DA showed the contribution of each feature, and the peaks with top 20 VIP scores were selected as features of lung cancer. The Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Light Gradient

Boosting Machine (LightGBM) was performed using Python Scikit-learn 0.23 package in Jupyter Notebook 6.1.4 and the feature importance function was performed to calculate the feature importance parameter. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance. The higher the value, the more important the feature. The 17 peaks with the highest normalized feature importance in RF, 14 peaks with the highest normalized feature importance in XGBoost, 23 peaks with the highest normalized feature importance in LightGBM were selected as features of lung cancer. Finally, a two-sample t-test was performed to check the significance of altered levels of these features in the serum samples of lung cancer patients versus healthy controls ($p < 0.001$) using Metaboanalyst 5.0 (McGill University, Montreal, Canada).

Internal validation. The principal component analysis (PCA) and Hierarchical Clustering Heatmap were performed using Metaboanalyst 5.0 (McGill University, Montreal, Canada). The K-nearest Neighbors (KNN) was performed using the sklearn.neighbors.KNeighborsClassifier package in Jupyter Notebook 6.1.4 and the n_neighbors of parameters was 1. The Support Vector Machine (SVM) was performed using the sklearn.svm.SVC package in Jupyter Notebook 6.1.4 and the kernel of parameters was poly. The Logistic Regression (LR) was performed using the sklearn.linear_model.LogisticRegression package in Jupyter Notebook 6.1.4 with default parameters. The Extremely Randomized Trees (ExtraTree) were performed using the sklearn.ensemble.ExtraTreesClassifier package in Jupyter Notebook 6.1.4 with default parameters.

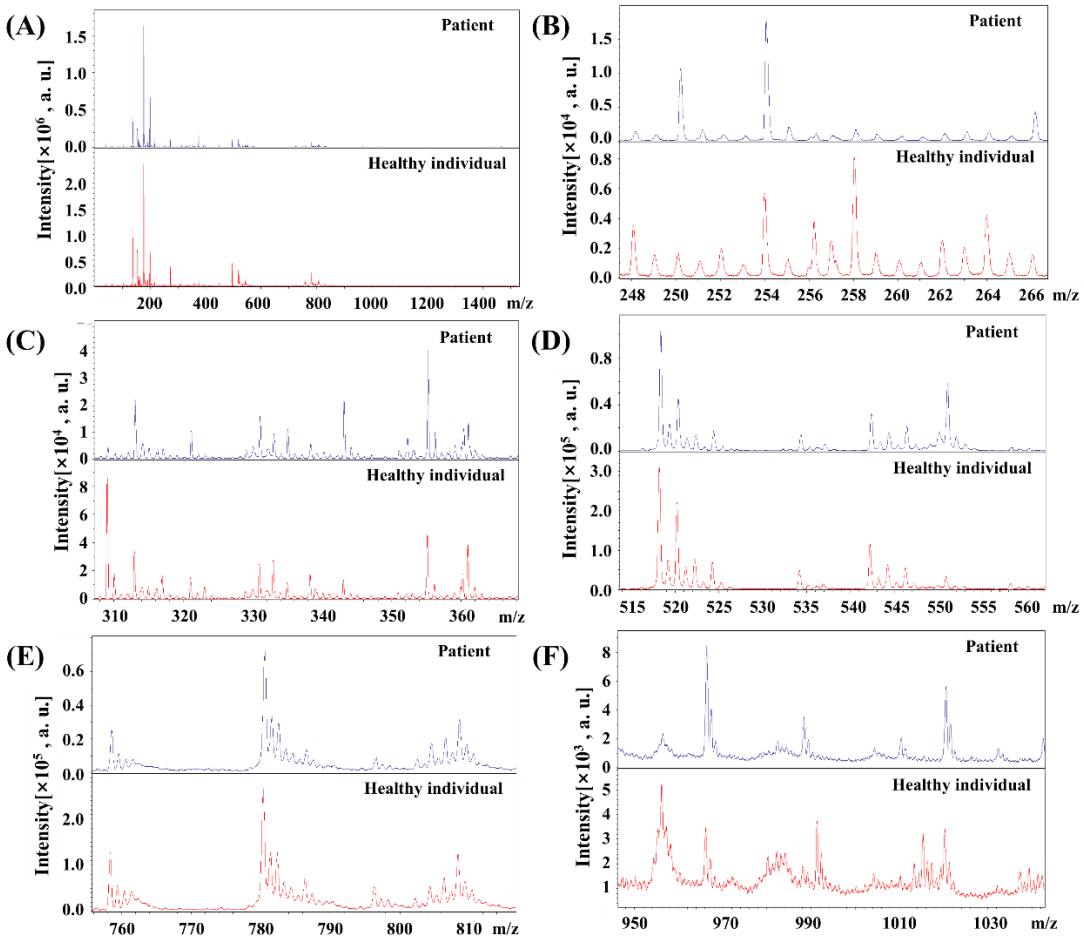


Figure S1. The representative MALDI-TOF MS spectra of samples from lung cancer patients and healthy control subjects. (A) MALDI-TOF mass spectra of serum samples from lung cancer patients and healthy controls; (B-F) Partial enlarged view of (A).

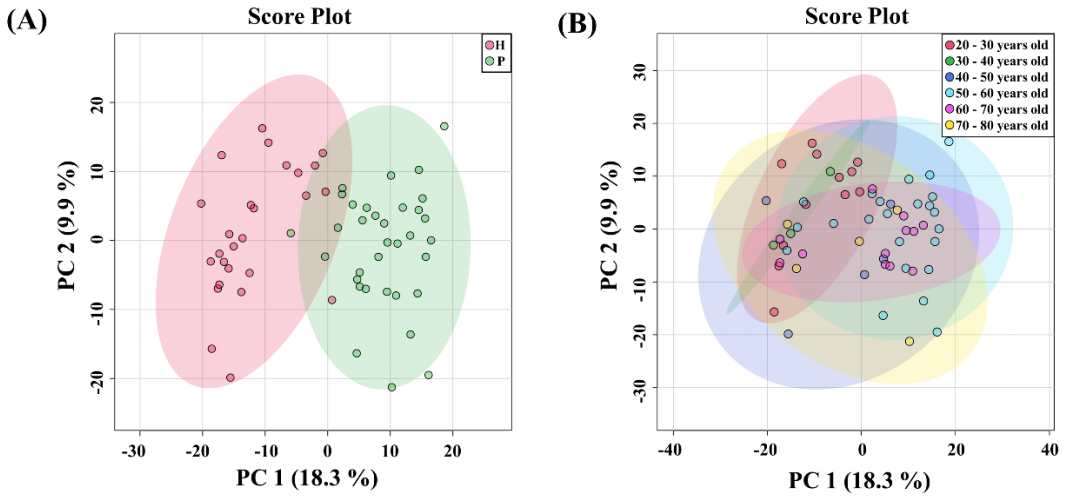


Figure S2. Unsupervised Learning between lung cancer patients and healthy controls. PCA score plots of all 783 processed ion peaks, with the first two PCs explaining 28.2% of the total variance. (A) The label colours - green represented lung cancer patients and red represented healthy controls; (B) The label colours - red represented samples from 20 to 30 years old, green represented samples from 30 to 40 years old, purple represented samples from 40 to 50 years old, blue represented samples from 50 to 60 years old, blue represented samples from 50 to 60 years old, aubergine represented samples from 60 to 70 years old and yellow represented samples from 70 to 80 years old. The data cannot be found a trend to be grouped according to age.

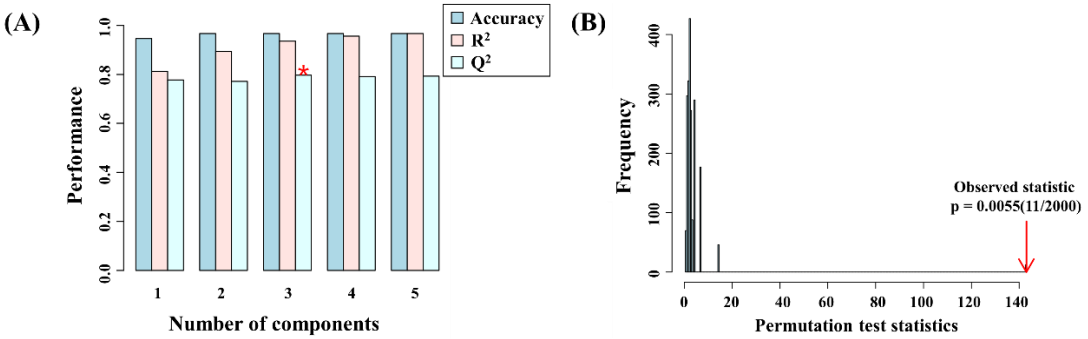


Figure S3. PLS-DA model based on the MALDI-TOF data set. (A) Supervised PLS-DA classification of lung cancer patients and healthy controls using different number of components. Red asterisk indicates the best classifier, R2=0.94, Q2=0.80, accuracy=0.97; (B) Permutation tests based on separation distance of PLS-DA, indicating that the discriminatory power of the PLS-DA model is robust and is associated with a statistically significant p value < 5E-04 (0/2000).

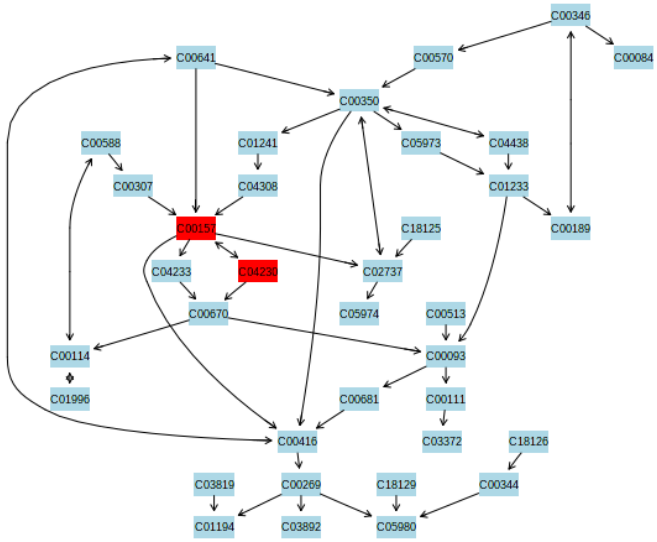
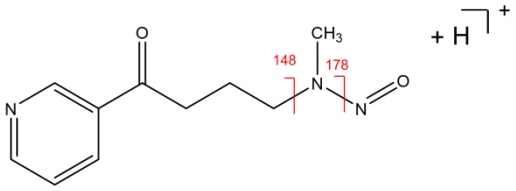
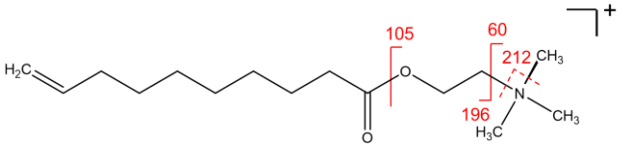


Figure S4. Schematic illustration of Glycerophospholipid metabolism. The compound colours within the pathway - blue represented metabolites that were not in the data and used as background in topology analysis with total importance of 0.89; and red represented the two metabolite (LysoPC(18:2(9Z,12Z)/0:0) and PC(14:0/20:2(11Z,14Z))) that was in the data and used in topology analysis with total importance of 0.11.

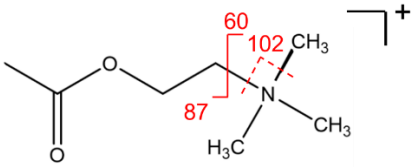
4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone



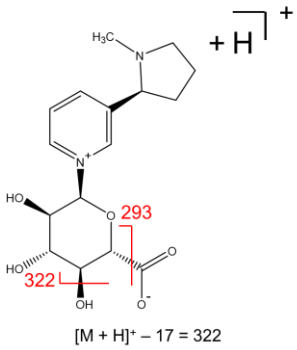
9-Decenoylcholine



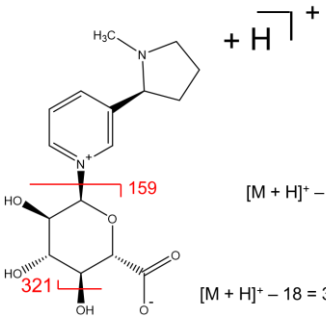
Acetylcholine
(Analogous compound of 9-Decenoylcholine)



Nicotine glucuronide



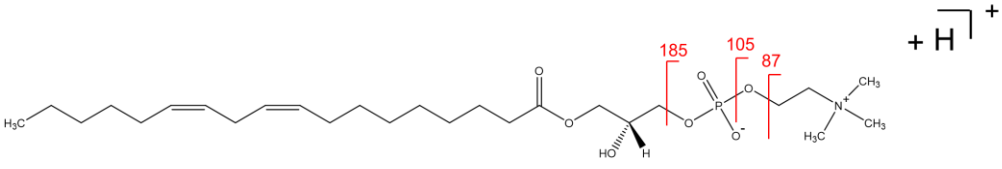
[M + H]⁺ – 17 = 322



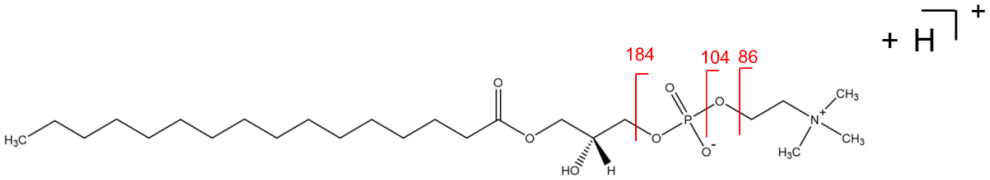
[M + H]⁺ – 18 - 162 = 159

[M + H]⁺ – 18 = 321

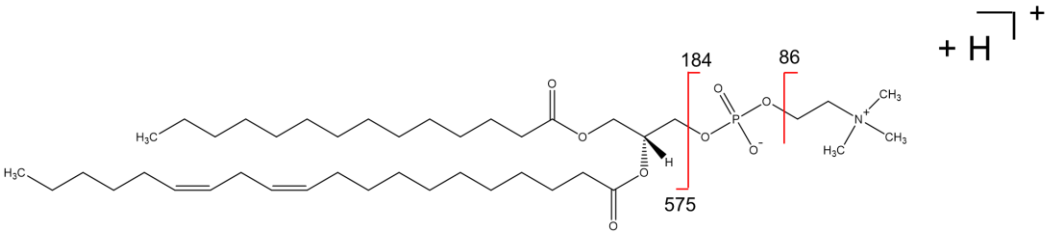
LysoPC(18:2(9Z,12Z)/0:0)



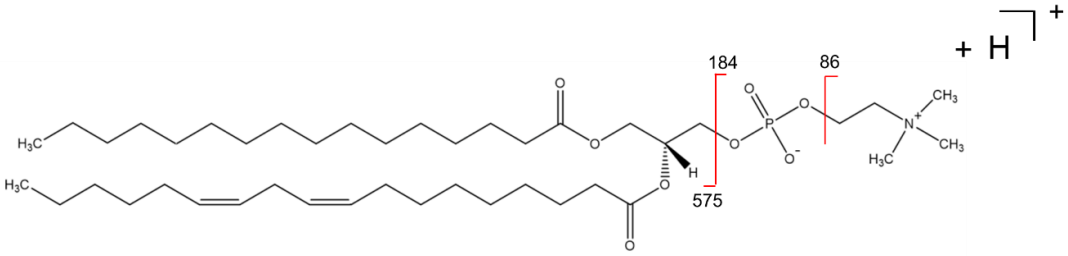
1-PalMitoyl-2-hydroxy-sn-glycero-3-phosphocholine (P-lysoPC, LPC)
(Analogous compound of LysoPC(18:2(9Z,12Z)/0:0)



PC(14:0/20:2(11Z,14Z))



Lecithin (Analogous compound of PC(14:0/20:2(11Z,14Z))



Disialosyl galactosyl globoside

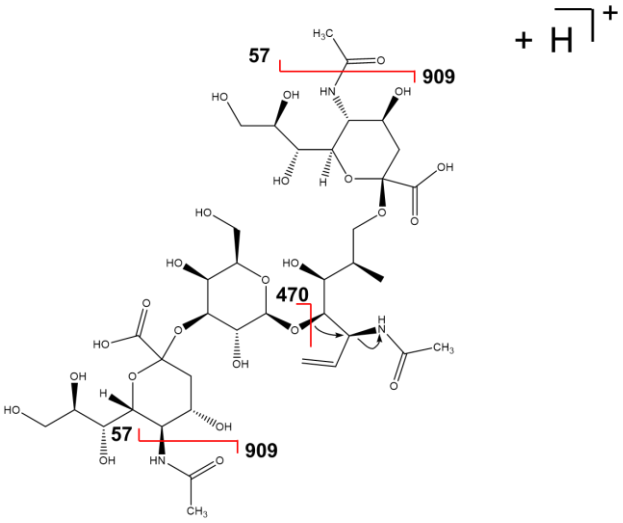


Figure S5. The structural assignment of fragment ions recorded in MS/MS spectra.

Table S1. Metabolites selected as biomarkers to distinguish lung cancer patients from healthy controls

Metabolite name	HMDB ID	Monoisotopic molecular weight (Da)	Adduct ion signals m/z		Selected ion peaks from MS/MS spectra m/z (relative abundance)	MS/MS spectra of corresponding standard
Disialosyl galactosyl globoside[2]	HMDB0006588	965.333	$[M+H]^+$	**966.335**[a]	56.6(18.8), 184.1(14.2), 470.1(100), 781.3(26.2), 909.4(79.7), 948.8(6.9)	---
			$[M+Na]^+$	988.312	184.1(4.4), 496.3(100), 804.4(7.9), 970.3(4.88)	---
			$[M+K]^+$	1004.277	184.1(21.7), 496.5(100), 976.7(18.1), 987.2(32.8)	---
PC(14:0/20:2(11Z,14Z))[1]	HMDB0007880	757.562	$[M+H]^+$	**758.568**[a]	86.0(0.9), 184.0(100), 575.1(3.1)	86.1(0.1), 184.0(100), 575.5(0.1)
			$[M+Na]^+$	780.542	86.1(9.2), 184.0(100), 575.5(13.6)	86.1(0.1), 184.1(100), 575.6(0.3)
			$[M+K]^+$	796.515	86.1(6.9), 184.1(100)	86.1(0.7), 184.1(100)
LysoPC(18:2(9Z,12Z)/0:0)[1]	HMDB0010386	519.332	$[M+H]^+$	**520.335**[a]	86.0(1.1), 104.1(37.6), 184.0(100)	86.0(3.3), 104.1(67.7), 184.0(100)
			$[M+Na]^+$	**542.312**[a]	86.1(4.4), 104.1(100), 184.0(19.3)	86.0(9.7), 104.1(100), 184.0(1.2)
			$[M+K]^+$	**558.285**[a]	86.1(8.6), 104.1(100)	86.1(10.9), 104.1(100)
Nicotine glucuronide[2]	HMDB0001272	338.148	$[M+H]^+$	339.030	159.0(15.3), 292.9(11.2), 321.1(16.2), 322.2(100)	---
			$[M+Na]^+$	**361.007**[a]	160.4(68.4), 204.6(74.2), 343.0(100)	---
			$[M+K]^+$	**376.985**[a]	No obvious fragment ions	---
Unidentified biomarker	---	---	$[M+H]^+$	299.130	84.6(19.6), 123.2(45), 137.0(22.5), 180.5(67.8), 240.1(100)	---
			$[M+Na]^+$	**321.107**[a]	84.6(22.1), 116.9(26.9), 141.8(16), 180.8(100), 261.0(78.3)	---
			$[M+K]^+$	337.088	No obvious fragment ions	---
Unidentified biomarker	---	---	$[M+H]^+$	287.212	55.9(1.87), 59.0(15.2), 70.1(1.34), 84.2(100), 105.4(6.41), 143.4(5.3), 176.1(1.8), 228.2(48.6), 231.8(3.5), 265.9(6.1)	---
			$[M+Na]^+$	**309.156**[a]	55.8(2.4), 58.9(3.5), 70.0(2.0), 84.1(28.6), 103.3(2.9), 1445(3.5), 176.2(8.6), 228.4(1.7), 232.1(1.13), 250.1(100), 266.0(5.2)	---
			$[M+K]^+$	325.112	56.2(26.1), 58.3(14), 70.4(20.1), 84.6(46.5), 103.9(28.4), 176.0(20.1), 227.7(100), 230.7(55.9), 267.1(50.4)	---
Unidentified biomarker	---	---	$[M+H]^+$	228.198	85.0(67.0), 146.5(60.8), 187.0(97.3), 200.0(100)	---
			$[M+Na]^+$	**250.179**[a]	84.1(17.7), 137.9(38.3), 188.4(68.3), 199.3(100)	---
			$[M+K]^+$	**266.164**[a]	83.9(12.8), 138.5(100), 192.8(23.4), 203.1(18.7)	---
9-Decenoylcholine[1]	HMDB0013206	256.228	$[M]^+$	**256.295**[a]	59.6(1.0),104.8(17.4), 196.0(5.0), 212.1(100)	59.8(30.2), 87.0(100), 102.1(0.1)
4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone	HMDB0011603	207.100	$[M+H]^+$	**208.108**[a]	148.1(86.6), 178.0(100.0)	148.1(51.9), 178.1(100.0)
			$[M+Na]^+$	230.101	No obvious fragment ions	No obvious fragment ions
			$[M+K]^+$	246.063	No obvious fragment ions	No obvious fragment ions

[a] 13 distinctive features.

[1] The reference compound used for MS/MS identification is an analogous compound which contain the similar functional group. The fragmentation patterns are shown in Figure S5 for comparison.

[2] The assignment of the fragmentation patterns are illustrated in Figure S5.

Table S2. Demographic information and clinical feature of lung cancer patients and healthy controls

Characteristics	Lung cancer patients (n=34)	Healthy controls (n=26)
Age (mean ± SD, years)	58±6.55	40±17.22
Sex		
Male	20	16
Female	14	10
Smoking		
Never	25	26
Previous	3	0
Current	6	0
Stage		
I	3	
II	3	
III	4	
IV	24	