Research Article

For reprint orders, please contact reprints@future-science.com

Bioanalysis

Multiplatform serum metabolic phenotyping combined with pathway mapping to identify biochemical differences in smokers

Aim: Determining perturbed biochemical functions associated with tobacco smoking should be helpful for establishing causal relationships between exposure and adverse events. **Results:** A multiplatform comparison of serum of smokers (n = 55) and never-smokers (n = 57) using nuclear magnetic resonance spectroscopy, UPLC–MS and statistical modeling revealed clustering of the classes, distinguished by metabolic biomarkers. The identified metabolites were subjected to metabolic pathway enrichment, modeling adverse biological events using available databases. Perturbation of metabolites involved in chronic obstructive pulmonary disease, cardiovascular diseases and cancer were identified and discussed. **Conclusion:** Combining multiplatform metabolic phenotyping with knowledge-based mapping gives mechanistic insights into disease development, which can be applied to next-generation tobacco and nicotine products for comparative risk assessment.

Manuja R Kaluarachchi[‡,1], Claire L Boulangé[‡,1], Isabel Garcia-Perez[‡,1], John C Lindon[1] & Emmanuel F Minet*[,2]

[1]Metabometrix Ltd, Bioincubator, Prince Consort Road, South Kensington, London, SW7 2BP, UK

[2]British American Tobacco, Research & Development, Regents Park Road, Southampton, SO15 8TL, UK

*Author for correspondence: Tel.: +44 0 2380 588 997 emmanuel_minet@bat.com

[‡]Authors contributed equally

First draft submitted: 25 May 2015; Accepted for publication: 3 August 2016; Published online: 16 September 2016

Keywords: biomarker • metabonomics/metabolomics • xenobiotics

One of the undisputed approaches to determine the potential health risk and diseases associated with lifestyle, and exposure to toxicological substances is epidemiology. In the early 1950s and 1960s, epidemiological evidence linked for the first time tobacco smoking with lung cancer and subsequently with other respiratory diseases and cardiovascular diseases [1–3]. Next-generation tobacco and nicotine products, such as electronic cigarettes and heated tobacco devices, are emerging across markets with differing levels of regulation having limited scientific evidence regarding their relative risk for health [4,5]. Since epidemiology is retrospective in nature and requires a marketed product and a study population in which the adverse biological outcome may take years or decades to develop it cannot at this point be used to guide regulatory decisions on new tobacco and nicotine products. Furthermore, epidemiology does not give mechanistic information about the events leading to a disease. Therefore, biomarkers that are predictive of biological adverse effects can inform the decisions and actions of policy makers and product developers. A toxic stress or disease will trigger a tissue response, such as the secretion of inflammatory mediators, and can cause cell damage with leakage of cell material in the blood stream that can be detected by targeted assays. Such protein, miRNA and metabolite markers found in serum and urine have been used to assess tobacco product toxicity [6,7]. Targeted toxicological end points in biofluids, including single or multiplexed markers of inflammation and genotoxicity, give a useful, yet narrow and truncated view of the overall perturbations caused by a toxic stress. Thus, adverse biological events that might be of importance when comparing the risk of novel tobacco products against conventional tobacco and nicotine replacement therapies could be missed. However, the dra-

FUTURE SCIENCE part of fsg

matic advance in detection methods, statistical models and computer sciences offers new perspective to apply global screens (omics) to assess adverse biological responses for product risk assessment. Such a multidisciplinary approach applied to product risk assessment is often referred as systems toxicology [8,9].

In systems toxicology, a biological entity can be simplified to a network of genes, proteins, lipids and metabolites, in cells and organs interacting with each other in equilibrium to perform diverse functions [10]. A disease or a toxicological stress can be viewed as temporary or permanent perturbation of this network homeostasis evolving over a certain period of time [10,11]. Theoretical pathways and disease networks can be built by leveraging the information present in medical and biological databases and queried using the calculating power of modern computers [12,13]. Thus if a detailed signature of biological perturbations caused by an exposure event can be obtained and mapped to biological pathways it potentially becomes possible to model the causal relationship between exposure and adverse events potentially leading to diseases [13]. In the context of risk assessment, there is an interest in creating a comprehensive map of the perturbed biological functions caused by tobacco smoking to use as a benchmark to contrast and compare conventional and next-generation tobacco and nicotine products both *in vitro* and in clinical samples.

Metabolic phenotyping (also known as metabonomics or metabolomics) is an approach that allows the collection of a comprehensive signature of the metabolic profile of human subjects, which can be altered by factors, such as lifestyle, diet, medical intervention or diseases, for example [14]. This paradigm is particularly well suited for tobacco risk assessment in clinical studies since it can be applied to biofluid samples not requiring invasive biopsy procedures that are not ethical in healthy subjects. Only a handful of studies have conducted a metabolic analysis of either urine, saliva or serum of smokers using a single- or a two-platform approach, such as GC–MS, capillary electrophoresis mass spectrometry or UPLC–MS [15–19]. Each platform provides complementary coverage of the metabolic space, and is built to preferentially target polar, lipophilic or charged metabolites; therefore a single platform can only acquire a fraction of a metabolic profile.

The objective of this study was to perform the first multiplatform study using a range of both combined NMR and MS assays to establish key metabolic perturbations when the serum of smokers and nonsmokers is compared. Thus, within each of the two main platforms we used two types of NMR spectra, and four types of UPLC–MS, namely reverse phase (RP) and hydrophilic (HILIC) UPLC–MS in both positive and negative ESI to determine the serum metabolome/lipidome of smokers and nonsmokers. Identification of biological adverse events was performed by mapping the metabolic perturbations to functional pathways using biological network knowledge-based applications. The impact on sphingolipids, glycerophospholipid metabolism, levels of HDL and antioxidants, such as glutathione, is reported and discussed in light of the current literature. This work highlights the potential benefit of applying a multiplatform metabolic phenotyping approach combined with knowledge-based tools for future evaluation of next-generation tobacco products.

Experimental section

Samples

Sixty-seven smokers and 61 never-smokers from the Hamburg (Germany) area were enrolled in the study for a period of 183 and 164 days, respectively [20]. Inclusion criteria were a minimum weight of 52 kg for men and 45 kg for women and a BMI within the defined healthy range with no clinical history of heart, lung diseases and chronic diseases. Pregnant women were excluded from the study. Additional requirements for smokers were to be a regular consumer of 10–30 6–8 mg ISO Tar yield cigarettes/day, to be aged between 28 and 55 years old, to have been a smoker for at least 5 years and with a urinary cotinine (a major nicotine metabolite) level above 100 ng/ml at the point of screening [20]. The smokers were provided with a standard 7-mg ISO tar yield, 0.6-mg nicotine king-size commercial product (Lucky Strike Silver) to smoke for the duration of the study [20]. The never-smoker group was defined as never having smoked more than 100 cigarettes during his/her lifetime, with no cigarettes in the past 5 years, no regular exposure to second-hand smoke and to be aged between 28 and 55 years old. The urinary cotinine level measured at screening and during the course of the study had to remain below 30 ng/ml in never-smokers. Further details on the inclusion and exclusion criteria (e.g., medicines, medical conditions) are given in the published protocol by Shepperd *et al.* [20,21]. For the duration of the study, participants were given an electronic diary to record cigarette consumption, diet, exercise, medications and any health-related event. The diary entries were checked during regular ambulatory visits for protocol violations (days 31, 62, 95, 124 and for smokers at day 165). Breaches of protocol were assessed for severity in order to evaluate continued participation to the study. In-clinic evaluations with biofluid collection (saliva, blood, urine) were performed on days 14, 45, 76, 108 and 183 for smokers and on days 3, 80 and 164 for never-smokers [20]. A period of 48 h of

clinical confinement with controlled diet preceded the biofluid sample collection with exclusion of caffeine, alcohol, grilled, fried and smoked food. Adherence to these dietary restrictions was also required 48 h before the visit to the clinic. A variety of biomarkers of tobacco smoke exposure and biological effects listed in Shepperd *et al.* [20] were measured in blood, serum, saliva and urine and the data were published in Shepperd *et al.*, Haswell *et al.* and Banerjee *et al.* [21–23]. For this study, we only used serum samples collected at day 183 (n = 55; 28 males, 27 females) for smokers and day 164 (n = 57; 29 males, 28 females) for never-smokers who completed the study. The total nicotine equivalent quantification (TNEQ) was performed in urine as previously described [21]. Table 1 presents the summary demographic information on the participants who completed the study and the detailed anonymized metadata including recent past medical history are accessible on the MetaboLights database [24] (accession number: MTBLS364). The study protocol and informed consent forms were approved by the Ethics Committee of the Ärztekammer Hamburg, Germany and the clinical study was conducted in accordance with the World Medical Association Declaration of Helsinki (World Medical Association, 2004) and ICH Guidelines for Good Clinical Practice (ICH, 1996). The study was registered in the Current Controlled Trials database under the reference ISRCTN81286286. Cessation counseling was provided during the course and at the end of the study through workshops, and lectures assisted by a trained psychologist to develop individual plans for quitting. Voluntary-free enrollment was provided for the Smoke-Free counselling Programme of the Institut für Gesundheitsförderung, which also comprises a relapse prevention phase [20].

Chemicals

LC–MS grade solvents were used throughout the application. Acetonitrile, isopropanol, formic acid, ammonium formate, sodium phosphate, trimethylsilyl-[2,2,3,3,-$^{2}H_{4}$]-propionate, deionized water and deuterated water were all purchased from Sigma Aldrich, UK.

Sample preparation

Frozen serum samples (-80°C) were thawed and then centrifuged at 2700 × *g* for 10 min to remove particulates and precipitated proteins. Serum samples were prepared for metabolic profiling analysis by RP and HILIC UPLC–MS as follows: 200 µl of supernatant was treated (1:3) with isopropanol, incubated at -20°C for 24 h, centrifuged at 2700 × *g* for 20 min and aliquoted for HILIC and RP methods. QC samples were prepared by pooling 50 µl volumes of each sample. During the analysis, the samples were maintained at 4°C in the autosampler and separated and analyzed using both RP and HILIC chromatographic methods [25]. Alternatively, 300 µl of individual serum samples were prepared with pH 7.4 phosphate buffer, as described previously for high-resolution ^{1}H-NMR spectroscopy [26].

Instrumentation

RP and HILIC UPLC–MS metabolic profiling experiments were performed using a Waters Acquity Ultra Performance LC system (Waters, MA, USA) coupled to a Xevo G2 Q-TOF mass spectrometer (Waters, MA, USA) with an electrospray source. ^{1}H-NMR metabolic profiling analysis of blood serum was performed at 300K on a Bruker Avance spectrometer at 600 MHz (Bruker Biospin, Rheinstetten, Germany).

Table 1. Summary demographic for subjects who completed the clinical study and provided samples used in this metabolic phenotyping study.

		Smokers, n = 55	Never-smokers, n = 57
Age (years)	Mean ± SD	39.7 ± 8.8	42.4 ± 7.2
	Median (min, max)	43 (24, 54)	44 (28, 55)
Gender (n)	Females	27	28
	Males	28	29
BMI (kg/m^2)	Mean ± SD	24.7 ± 2.5	25.0 ± 2.7
	Median (min, max)	24.4 (18.6, 30)	25.0 (18.8, 30)
Ethnicity	Caucasians	55	55
	NonCaucasians	0	2
Exposure TNEQ (µg/ml)	Mean ± SD	8.4 ± 4.9	BLOQ
Cigarettes per day	Mean ± SD (over study duration)	21.6 ± 7.5	0

BLOQ: Below limit of quantification; SD: Standard deviation of the mean; TNEQ: Total nicotine equivalent.

Reverse phase & HILIC UPLC–MS conditions

The serum samples were first subjected to analysis using UPLC–MS, with an RP chromatographic method with both positive and negative MS ionization detection. Second, to separate and detect more polar molecules, an HILIC chromatographic stage was used with both positive and negative MS ionization detection. Separation of lipid components by the RP method was performed according to the Waters application publication [27].

HILIC separation was performed as previously described [25] and samples were analyzed in a random order. Capillary and cone voltages were set at 1.5 kV and 30 V, respectively. The desolvation gas was set to 1000 l/h at a temperature of 600°C; the cone gas was set to 50 l/h and the source temperature was set to 120°C. For mass accuracy a lock–spray interface was used with leucine–enkephalin (556.27741/554.2615 amu) solution at a concentration of 2000 ng/ml and at a flow rate of 15 µl/min as the lock mass.

^{1}H-NMR conditions

The serum samples were analyzed using ^{1}H-NMR spectroscopy with two different pulse sequences [26]. The first, the so-called noesy–presat sequence provides an overview of all proton-containing species and yields sharp peaks for small molecule species, broad bands from the lipoproteins (used later for lipoprofile analysis, see below) and a broad largely featureless background from proteins, the most abundant being albumin. The second-pulse sequence, the so-called Carr-Purcell-Meiboom-Gill (CPMG) sequence, takes advantage of the different nuclear spin relaxation times between large and small molecules to attenuate the peaks from large molecules (those with shorter spin–spin relaxation times) to leave mainly small molecule metabolite peaks. The main data analysis was performed using the CPMG spectral data and the lipoprotein profiling was carried out using the noesy–presat spectral data. For the CPMG spectra a relaxation delay of 4 s, a mixing time of 0.01 s, a spin–echo delay of 0.3 ms, 128 loops and a free induction decay acquisition time of 3.067 s were used. A total of 32 scans were recorded into 96 k data points with a spectral width of 20 ppm.

Data treatment

All the raw spectra for the different platform have been uploaded and are accessible on the MetaboLights database [24] (accession number: MTBLS364). The raw mass spectrometric data acquired were processed using xcms in R [28] and the centwave peak picking methods were used to detect chromatographic peaks. The xcms-centwave parameters were dataset specific. Data were normalized using probabilistic quotient normalization [29] using the median spectrum as the reference.

The acquired ^{1}H-NMR spectra were zero-filled to 128 k points, Fourier transformed, phase and baseline corrected using Bruker Topspin 3.1 (Bruker Biospin, Rheinstetten, Germany). The serum spectra were referenced to the anomeric proton assigned to α-glucose at δ5.22 and imported to MATLAB™ (R2014a, The Mathworks, MA, USA) for further analysis. Regions corresponding to the water peak (δ4.3–4.925) were removed. All spectra were normalized using probabilistic quotient normalization [28] using the mean spectrum as the reference.

Lipoprotein profiles from deconvolution of NMR peak shapes

Quantification of lipoprotein subclasses was obtained from deconvolution of the methyl peak near δ0.89 using a Bruker (Bruker Biospin, Rheinstetten, Germany) procedure based on the method of Petersen *et al.* [30]. For QC purposes, the Bland–Altman prediction errors and correlations coefficients were calculated using the conventional values and the Bruker NMR measurement of total HDL, LDL, triglycerides and cholesterol. The measurement quality is in line with and meets the Bruker routine standard, therefore the complete analysis of 105 lipoprotein subclasses was performed including different chemical components of IDL (density 1.006–1.019 kg/l), VLDL (0.950–1.006 kg/l), LDL (density 1.09–1.63 kg/l) and HDL (density 1.063–1.210 kg/l). The LDL subfraction was separated in six density classes (LDL-1 1.019–1.031 kg/l, LDL-2 1.031–1.034 kg/l, LDL-3 1.034–1.037 kg/l, LDL-4 1.037–1.040 kg/l, LDL-5 1.040–1.044 kg/l, LDL-6 1.044–1.063 kg/l) and the HDL subfraction in four density classes (HDL-1 1.063–1.100 kg/l, HDL-2 1.100–1.125 kg/l, HDL-3 1.125–1.175 kg/l, HDL-4 1.175–1.210 kg/l). Based on their lipoprotein fractionation protocol, Bruker has implemented a specific nomenclature for the 105 lipoprotein components (see Supplementary Table 1).

Statistical analysis

Multivariate statistical analysis was used to examine the datasets and to observe clustering in the results according to predefined classes. The dominant source of variations in each data matrix and the outliers were identified by PCA.

Univariate logistic regression was performed, adjusting for confounding factors between never-smokers and smokers with class variables (BMI, gender, administered medications and age) for each NMR and MS feature in the datasets. Linear regression of the meta-

bolic data was performed against urinary TNEQ as the dependent variable. For each NMR variable/MS feature, a regression coefficient was obtained from the logistic regression model, and was converted to a t-score and subsequently to a p-value. These p-values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (pFDR) method. All variables/features that have a pFDR ≤ 5% are considered to be significantly associated with the separation of the never-smokers or smokers groups and are subsequently represented in a Manhattan plot.

Metabolite identifications

Metabolite identification of significant MS features employed online database searches (METLIN) [31], human metabolome database (HMDB) [32] and Lipid MAPS® [33]. Confirmation of such metabolite identifications was obtained using further MS–MS experiments and comparison to MS fragmentation patterns in the literature.

Confirmation of metabolite identities in the NMR data was obtained using 1D and 2D NMR experiments, spike-in of chemical standards, JRES (J-Resolved spectroscopy), (TOCSY) TOtal Correlation SpectroscopY and HSQC (Hetero-nuclear ^{1}H-[13]C Single Quantum Coherence) spectroscopy.

The identification of a metabolite was characterized by a level of assignment (LoA) score that describes how the identification was made (adapted from Sumner *et al.* [34]). The LoA used for the molecules identified by MS are LoA 1: identified compound, confirmed by comparison to an authentic chemical reference, LoA 2: MS/MS spectrum matched to database or literature to putatively annotate compound, LoA 3: accurate mass (*m/z*) matched to database and detection of common fragment ion to putatively characterize the compound classes, LoA 4: accurate mass (*m/z*) matched to database to make a tentative assignment. The LoA of the ^{1}H-NMR peaks are reported as follows: LoA 1: identified compound, confirmed by adding the authentic chemical compound to the plasma samples (spike-in experiments) LoA 2: ^{1}H-NMR chemical shifts and their multiplicity matched to database or literature to putatively annotate the compound.

Pathway enrichment analysis

HMDB [32] and Lipid MAPS® [33] were searched to obtain an overview of the classes of metabolites that were most affected by smoking using the tentatively identified metabolites in our dataset. Functional enrichment analysis of statistically significant metabolites was performed using MetaboAnalyst 3.0 [35,36] to interpret some of the observed metabolic changes. A cutoff value of pFDR ≤ 0.05 was used for MetaboAnalyst unfiltered for fold change, and filtered to only include metabolites with a fold change of 10% or more.

Results

Metabolic phenotype data

Smokers' and never-smokers' metabolic phenotype data were obtained by analyzing the serum samples with three different analytical platforms. The acquired data were organized into six separate datasets, namely CPMG NMR spectra, lipoprotein fraction data, RP UPLC–MS data (positive and negative ionization separately) and HILIC UPLC–MS data (positive and negative ionization separately) that were used for further statistical analysis. The lipoprotein fraction data were based on deconvolution of the complex lineshape of the methyl peak arising from lipoproteins in the noesy–presat NMR spectra,

The noesy–presat and CPMG ^{1}H-NMR spectra and representative RP and HILIC UPLC–MS profiles of smokers are shown in Figure 1A–F. Table 2 summarizes the number of features detected with each analytical platform.

Statistical analysis of the metabolic phenotype data

To discover any natural clustering by class end point (smoking status as the target end point, and gender as a potential confounder end point) in the metabolic data, PCA analysis was applied to all six datasets ^{1}H NMR, NMR-lipoprotein fraction data, HILIC UPLC–MS (ESI+), HILIC UPLC–MS (ESI-), RP UPLC–MS (ESI+) and RP UPLC–MS (ESI-)). The information from this analysis is displayed for illustrative purposes only as score plots in which each colored spot represents one individual. The PCA score plots of PC1 versus PC2 for the six datasets are shown in Supplementary Figure 1. When the data points are color-coded for smoking status, examination of all of the PCs for all six datasets show some degree of clustering with the MS-derived data giving greater clustering when assessed visually. From the PCA score plots of the MS data in particular it was shown that smoking status is the major source of data variance contributing to the first PCs and for the HILIC UPLC MS data there is separation in PC1. As a typical possible confounder variable, the gender effect was also investigated using the PCA data. When the points are color-coded by gender it can be seen for all six datasets there is a degree of clustering by gender with the variation due to this end point contributing to PC2. In this study HILIC UPLC–MS showed the best smoker/nonsmoker separation in PCA (Supplementary Figure 1).

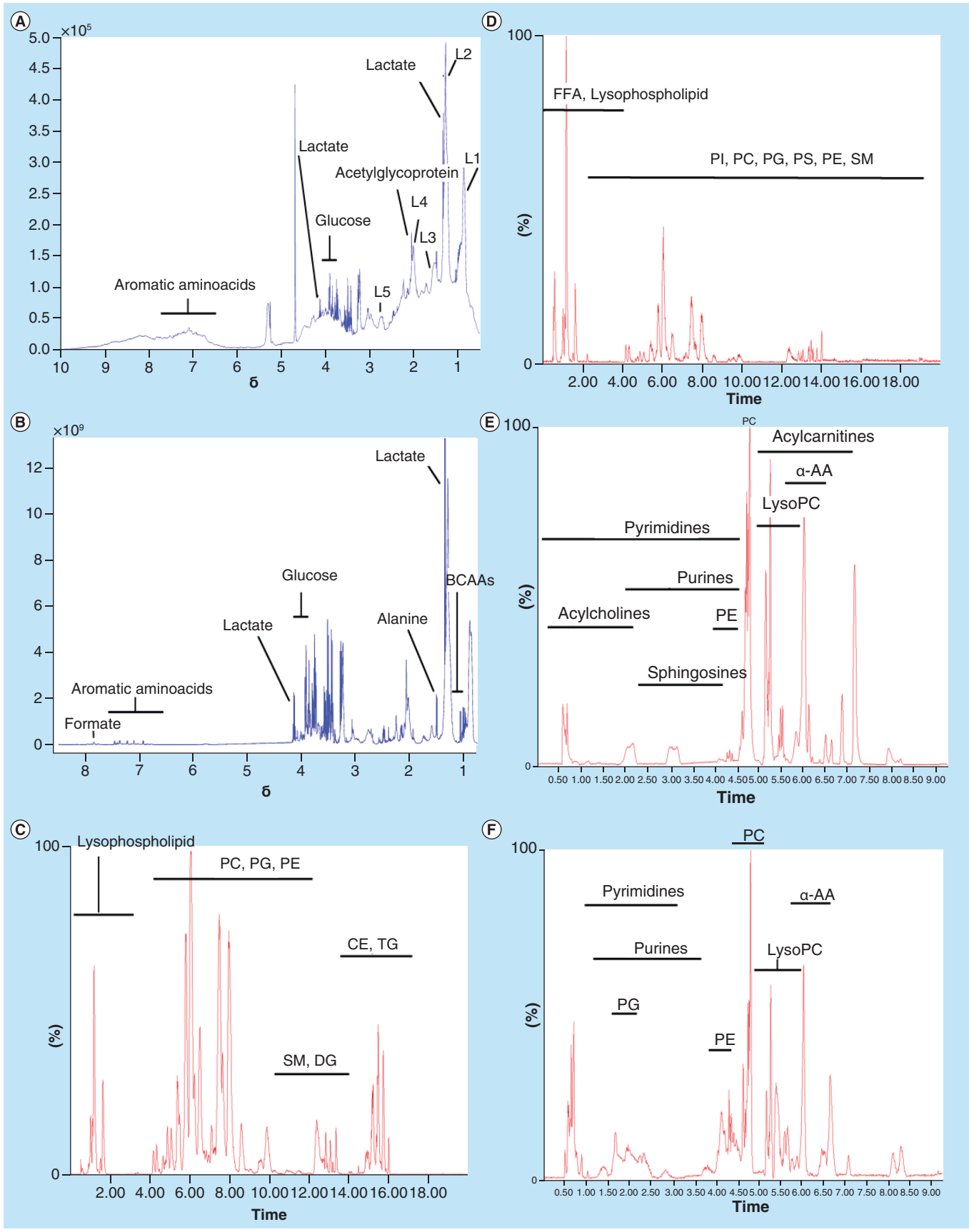
A
×10^5
Lactate
L2
Lactate
L1
Acetylglycoprotein
L4
Glucose
L3
L5
Aromatic aminoacids
δ
B
×10^9
Lactate
Glucose
BCAAs
Lactate
Alanine
Aromatic aminoacids
Formate
δ
C
Lysophospholipid
PC, PG, PE
CE, TG
SM, DG
(%)
Time
D
FFA, Lysophospholipid
PI, PC, PG, PS, PE, SM
(%)
Time
E
PC
Acylcarnitines
α-AA
LysoPC
Pyrimidines
Purines
Acylcholines
PE
Sphingosines
(%)
Time
F
PC
Pyrimidines
α-AA
Purines
LysoPC
PG
PE
(%)
Time

Figure 1. Metabolic phenotyping data of serum samples from smokers (see facing page). Median ^{1}H NMR **(A)** Noesy–presat and **(B)** CPMG spectra. Base peak intensity (BPI) chromatograms from **(C)** positive ESI mode (ESI+), **(D)** negative ESI mode (ESI-) from the RP-UPLC–MS analysis; **(E)** ESI+ and **(F)** ESI- from the HILIC UPLC–MS analysis. L1: Lipids methyl group: **CH_3**-$(CH_2)_n$ or **CH_3**-CH_2-CH_2C=, L2: Lipids : methylene group: CH_3-**CH_2**-CH_2, **$(CH_2)_n$** or CH_3-**CH_2**$(CH_2)_n$, L3: Lipids: methylene group: **CH2**-CH2-CH2-CO, L4: Lipids: methylene group: **CH_2**-CH_2-CO or **CH_2**-C=C, L5: Lipids: methylene group: **CH_2**-C=O, α–AA: Alpha amino acids.
CE: Cholesterol ester; DG: Diacylglycerol; PC: Phosphatidylcholine; PE: Phophatidylethanolamine; PG: Phosphatidylglycerol; SM: Sphingomyelin; TG: Triacylglycerol.

Univariate logistical regression and multiple factor correction analysis Figure 2A–C was applied to all six metabolic datasets to compare the metabolic profiles of never-smokers and smokers, taking into account variation caused by the variables that would potentially confound the data, such as gender variation that was already observed in the PCA.

As a result of applying this multiplatform approach, significantly different metabolites between the two classes were obtained after accounting for any variation caused by differences in age, gender and drug intake. The total number of discriminating features is shown in Table 2 with their corresponding analytical platform. The tentatively identified metabolites significant at pFDR ≤ 0.05 with at least one identifier in HMDB [32], Pubchem [37], CHEbi [38], KEGG [39] or Lipid MAPS® [33] are listed in Table 3. The fold changes in these metabolites between smokers and never-smokers, the pFDR values, analytical platform and mass are also summarized in Table 3.

Among the identified metabolites, glutathione and citrate decrease in the CPMG NMR serum profiles (Figure 2A & Table 3) of smokers while glutamate increases in the serum of smokers.

Similar results are evident for the four UPLC–MS datasets as seen in Figure 2B based on the retention times and observed *m/z* values for metabolites that differentiate smokers and never-smokers at pFDR ≤ 0.05. The identified candidates are listed in Table 3. Those tentatively identified lipids that were detected using more than one ionization mode or platform include LysoPC(18:0), LysoPC(16:0), SM(18:1/20:4) and SM(d18:1/23:0) (Table 3).

Using normalized intensity values box plot examples of six tentatively identified metabolites that are significantly different between smokers and nonsmokers are presented in Figure 3. These six metabolites are further examined in the discussion due to their potential link with cardiovascular diseases or emphysema.

The univariate logistic regression approach was also used to analyze the results from the lipoprotein fraction analysis from the NMR data. The main components (cholesterol, free cholesterol, phospholipids and triglycerides) of LDL subclasses 5 and 6 were found to be increased in serum of smokers compared with the never-smokers, see Figure 2C. These LDL subfractions constitute the most HDL from the analyzed

Table 2. Summary of platforms and features.

Method	Detected	Significant S v NS	Filtering hits for acceptance criteria†	Included in pathway enrichment‡,§
^{1}H-NMR CPMG	187,998 data points	767 data points	767 data points. Among these, three metabolites were identified	3 metabolites
Lipoprotein fraction data	105 lipoprotein components	12	12	0
RP UPLC–MS (ESI+)	739 MS features	230 MS features	175 MS features (84 metabolites of which 31 remain unassigned)	49 metabolites
RP UPLC–MS (ESI-)	529 MS features	115 features	82 MS features (30 metabolites of which 16 remain unassigned)	2 metabolites
HILIC UPLS–MS (ESI+)	732 MS features	336 features	146 MS features (44 metabolites of which 8 remain unassigned)	20 metabolites
HILIC UPLC–MS (ESI-)	220 MS features	101 MS features	28 MS features (18 metabolites)	8 metabolites

† All significant hits (features) discovered using UPLC–MS are manually extracted from the raw data to assign the nature of the ion (parent, isotope or adduct). At this stage it is also possible to identify features that are false positives and in the noise. These are removed from the list.
‡ Metabolites with an HMDB, PubChem, CHEbi, KEGG or lipidmaps ID.
§ Some metabolites were detected with multiple platforms and are highlighted in Table 3.
CPMG: Carr-Purcell-Meiboom-Gill; HILIC: Hydrophilic; NS: Never-smokers; RP: Reverse phase; S: Smokers.

Table 3. Metabolites discovered using NMR and UPLC–MS at pFDR ≤ 0.05 and with a public database identifier.

Generic name[†,‡]	Ion detected	*m/z* (parent)[§]	Platform	FC S vs NS[#]	pFDR	LoA[¶]	HMDB/ PubChem/CHEbi	KEGG	Lipid maps
(25R)-3alpha,7alpha-dihydroxy-5beta-cholestan-27-oyl taurine	Isotope of [M-H]-	543.0 (540.3)	HILIC-	0.59	0.003	4			LMST05040008
1-(6-[3]-ladderane-hexanyl)-2-(8-[3]-ladderane-octanyl)-sn-glycerophosphocholine	Isotope of [M+H]+	776.6 (774.6)	RP+	0.79	0.0006	3			LMGP01040090
13-methyl-4,4-bisnor-8,11,13-podocarpatrien-3-one	[M+H]+	229.2	HILIC+	0.79	0.007	4	CHEbi:25212		
2-aminobutyric acid	[M+H]+	104.1	HILIC+	1.2	0.01	4	HMDB00650	C02261	
6-Methoxyquinoline	[M+H]+	160.1	HILIC+	1.17	0.006	3	PubChem: 14860		
Aminohippuric acid	[M+H]+	195.1	HILIC+	1.18	0.0005	4	HMDB01867	D06890	
Androsterone sulfate	[M-H]-	369.2	HILIC-	0.56	0.004	2	HMDB02759		
Arginine	[M+H]+	175.1	HILIC+	1.18	0.04	2	HMDB00517	C00062	
Bilirubin	Isotope of [M+H]+	587.3 (585.3)	RP+	0.58	0.0092	3	HMDB00054	C00486	
Carboxylic acid	[M+H]+	130	HILIC+	0.64	0.02	2		C00060	
Carnitine	Isotope of [M+H]+	165.1 (162.1)	HILIC+	1.18	0.0002	2	HMDB00062	C00318	
Cer(d18:1/24:0)	[M+Na]+	672.6 (650.6)	RP+	1.2	0.0002	4	HMDB04956	C00195	LMSP02010012
Cer(d18:1/24:1(15Z))	[M+H]+	648.6	RP+	1.5	0.008	4	HMDB04953	C00195	LMSP02010009
Citrate	NMR	NA	NMR	0.9	0.02	2	HMDB00094	C00158	
FMC-5(d18:1/22:0)	[M+H]+	994.7	HILIC+	1.17	2.42E-05	3			LMSP05010034
GalCer(d18:1/24:0)	[M+H]+	812.7	RP+	1.53	0.0006	4			LMSP0501AC05
GlcCer(d14:1/22:1)	[M+H]+	742.6	RP+	0.86	0.0007	4			LMSP0501AA67
GlcCer(d18:1/23:0)	[M+H]+	798.7	RP+	1.29	0.005	4			LMSP0501AA32
Glutamate	NMR	NA	NMR	1.1	0.03	2	HMDB03339	C0025	
Gluthatione	NMR	NA	NMR	0.86	0.0005	2	HMDB00125	C00051	
Guanidinoacetic acid	Isotope of [M-H]-	117.1 (116.1)	HILIC-	0.65	0.0001	4	HMDB00128	C00581	
Hypoxanthine	[M-H]-	135	HILIC-	2.26	0.009	4	HMDB00157	C00262	
L-Argininosuccinic acid	[M-H]-	289.1	HILIC-	0.37	0.016	4	HMDB00052	C03406	
L-Argininosuccinic anhydride	[M-H]-	271.1	HILIC-	0.88	0.00013	4		C03406	

[†]Metabolites identified in more than one dataset are highlighted in bold.
[‡]Lipids key: Cer: Ceramide; GlcCer: Glyceryl ceramide; GalCer: Galactosylceramide; LysoPC: Lysophosphatidylcholine; LysoPE: Lysophosphatidylethanolamine; PC: Phosphatidylcholine; PE: Phosphatidylethanolamine; PE-Cer: Phosphatidylethanolamine-ceramide; PG: Phosphatidylglycerol; SM: Sphingomyelin; TG: Tri(acyllalkyl)glycerols.
[§]*m/z* mass charge ratio of detected feature including isotopes and salts, *m/z* of the parent in ().
[#]FC S vs NS: Fold-change smokers versus never-smokers.
[¶]LoA: Level of assignment.
NS: Never-smokers; S: Smokers.

Table 3. Metabolites discovered using NMR and UPLC–MS at pFDR ≤ 0.05 and with a public database identifier (cont.).

Generic name[†,‡]	Ion detected	*m/z* (parent)[§]	Platform	FC S vs NS[#]	pFDR	LoA[¶]	HMDB/ PubChem/CHEbi	KEGG	Lipid maps
L-Kynurenine	[M+H]+	209.2	HILIC+	2.47	0.0004	4	HMDB00684	C00328	
L-Phenylalanine	[M-H]-	164.1	HILIC-	0.93	0.0008	3	HMDB00159	C00079	
Lysine	[M+H]+	147.1	HILIC+	0.83	0.0003	4	HMDB00182	C00047	
LysoPC(15:0)	Isotope of [M+H]+	509.4 (508.4)	HILIC+	1.32	0.00009	2	HMDB10381	C04230	LMGP01050016
LysoPC(16:0)	[M+H]+	496.4	HILIC+	1.6	0.0002	2	HMDB10382	C04230	LMGP01050018
LysoPC(16:0)	[M+H]+	496.3	RP+	1.18	0.0015	2	HMDB10382	C04230	LMGP01050018
LysoPC(18:0)	[2M+H]+	1047.7 (524.3)	RP+	0.85	0.0039	2	HMDB10384	C04317	LMGP01050026
LysoPC(18:0)	[M+H]+	524.3	HILIC+	1.47	7.60E-06	3	HMDB10384	C04230	LMGP01050026
LysoPC(18:2)	[2M+H]+	1039.7 (520.3)	RP+	1.17	0.0003	3	HMDB10386	C04230	LMGP01050035
LysoPC(20:0)	[M+H]+	552.4	RP+	0.72	0.0043	3	HMDB10390	C04230	LMGP01050045
LysoPC(20:1)	[M+H]+	550.4	RP+	0.9	0.0063	3	HMDB10391	C04230	LMGP01050131
LysoPC(20:4)	[M+K]+	582.3 (544.3)	RP+	0.84	0.0037	2	HMDB10396	C04230	LMGP01050140
LysoPC(22:6)	[M+K]+	606.3 (568.3)	RP+	0.72	0.0011	3	HMDB10404	C04230	LMGP01050056
LysoPC(P-16:0)	[M+H]+	480.3	RP+	0.89	0.0072	3	HMDB10407	C04230	LMGP01070006
LysoPC(P-18:0)	[M+Na]+	530.4 (508.4)	RP+	0.8	0.0012	3	HMDB13122	C04230	LMGP01070009
LysoPE(22:6)	[M+H]+	526.3	RP+	0.67	0.0071	2	HMDB11526		LMGP02050013
PC(15:1/20:4)	[M+H]+	766.6	RP+	0.81	0.002	3			LMGP01011454
PC(16:0/18:1)	[M+IsoProp+H]+	820.6 (760.6)	RP+	0.83	0.0009	2	HMDB07972	C00157	LMGP01010005
PC(16:0/18:2(9Z,12Z))	[M+IsoProp+H]+	818.6 (758.6)	RP+	0.79	0.0043	4	HMDB07973	C00157	LMGP01010591
PC(17:1/20:4)	Isotope of [M+H]+	795.6 (794.6)	RP+	0.78	0.003	3			LMGP01011543
PC(18:0/16:0)	Isotope of [M+H]+	763.6 (762.6)	RP+	1.18	0.0093	2	HMDB08034	C00157	LMGP01010742
PC(18:0/22:6)	Isotope of [M+H]+	837.6 (834.5)	RP+	0.71	0.0038	2	HMDB08057	C00157	LMGP01010821

[†]Metabolites identified in more than one dataset are highlighted in bold.
[‡]Lipids key: Cer: Ceramide; GlcCer: Glyceryl ceramide; GalCer: Galactosylceramide; LysoPC: Lysophosphatidylcholine; LysoPE: Lysophosphatidylethanolamine; PC: Phosphatidylcholine; PE: Phosphatidylethanolamine; PE-Cer: Phosphatidylethanolamine-ceramide; PG: Phosphatidylglycerol; SM: Sphingomyelin; TG: Tri(acyllalkyl)glycerols.
[§]*m/z* mass charge ratio of detected feature including isotopes and salts, *m/z* of the parent in ().
[#]FC S vs NS: Fold-change smokers versus never-smokers.
[¶]LoA: Level of assignment.
NS: Never-smokers; S: Smokers.

Table 3. Metabolites discovered using NMR and UPLC–MS at pFDR ≤ 0.05 and with a public database identifier (cont.).

Generic name[†,‡]	Ion detected	*m/z* (parent)[§]	Platform	FC S vs NS[#]	pFDR	LoA[¶]	HMDB/ PubChem/CHEbi	KEGG	Lipid maps
PC(18:1(9Z)/0:0)	[M+H]+	522.3	HILIC+	1.23	0.0007	3	HMDB02815	C04230	LMGP01050032
PC(18:2/18:1)	[M+H]+	784.6	RP+	0.86	0.0068	2	HMDB08137	C00157	LMGP01011624
PC(20:4/22:6)	[M+H]+	854.6	RP+	0.97	0.002	4	HMDB08452	C00157	LMGP01011925
PC(O-16:0/18:1)	Isotope of [M+H]+	748.5 (746.5)	RP+	0.84	0.0003	3			LMGP01020003
PC(O-16:0/22:6)	[M+H]+	792.6	RP+	0.77	0.0052	3	HMDB13409		LMGP01020064
PC(O-18:0/0:0)	[M+Na]+	532.4 (510.4)	RP+	0.79	0.0042	3	HMDB11149	C04317	LMGP01060014
PC(O-18:1(11Z)/0:0)	[M+H]+	508.4	HILIC+	1.26	0.0002	3			LMGP01060034
PC(O-20:0/18:2(9Z,12Z))	[M+H]+	800.6	RP+	0.83	0.007	2			LMGP01020228
PC(P-18:0/16:0)	Isotope of [M+H]+	747.6 (746.6)	RP+	0.87	0.0004	3			LMGP01030052
PE(18:0/20:2)	[M+H]+	772.6	RP+	0.8	0.0012	4	HMDB09000	C00350	LMGP02010124
PE-Cer(15:2/24:0)	[M+H]+	729.6	RP+	0.86	0.0045	4			LMSP03020046
PG(20:0/22:0)	[M+H]+	863.6	RP+	1.26	0.003	4			LMGP04010947
Proline	[M+H]+	70.1	HILIC+	0.85	0.02	4	HMDB00162	C00148	
Serine	[M-H]-	104	HILIC-	0.86	0.0008	2	HMDB00187	C00065	
SM(18:1/16:0)	Isotope of [M+H]+	704.6 (703.5)	RP+	1.2	0.0007	2		C00550	LMSP03010003
SM(d18:1/14:0)	[M+Na]+	697.5 (675.5)	RP+	0.97	0.0041	4	HMDB12097		LMSP03010028
SM(d18:1/18:1)	[M+H]+	729.6	RP+	0.87	0.0007	4	HMDB12101	C00550	LMSP03010029
SM(d18:1/22:0)	[M+FA-H]-	831.7 (785.7)	RP-	1.4	0.00086	2	HMDB12103	C00550	LMSP03010092
SM(d18:1/23:0)	[M+Na]+	824.7 (801.7)	RP+	1.6	0.012	2	HMDB12105	C00550	LMSP03010078
SM(d18:1/23:0)	[M+FA-H]-	846.7 (799.7)	RP-	1.5	0.0001	2	HMDB12105	C00550	LMSP03010078
SM(d18:1/24:0)	Isotope of [M+H]+	818.7 (815.7)	RP+	1.3	0.0024	2	HMDB11697		LMSP03010008
SM(d18:1/24:1)	[M+Na]+	835.7 (813.7)	RP+	0.97	0.004	2	HMDB12107	C00550	LMSP03010007
SM(d18:2/14:0)	[M+H]+	673.5	RP+	0.96	0.0004	3		C00550	LMSP03010034
SM(d18:2/14:0)	Isotope of [M+H]+	674.5 (673.5)	HILIC+	1.07	0.0006	3			LMSP03010034
SM(d18:2/15:0)	[M+H]+	687.5	RP+	0.83	0.0009	3			LMSP03010036
SM(d18:2/21:0)	[M+H]+	771.6	RP+	0.89	0.004	3			LMSP03010064
SM(d18:2/22:0)	[M+ACN+Na]+	848.7 (785.6)	RP+	0.77	0.002	2			LMSP03010092
SM(d18:2/24:0)	Isotope of [M+H]+	814 (813.7)	HILIC+	1.67	0.00004	3			LMSP03010081

[†]Metabolites identified in more than one dataset are highlighted in bold.
[‡]Lipids key: Cer: Ceramide; GlcCer: Glyceryl ceramide; GalCer: Galactosylceramide; LysoPC: Lysophosphatidylcholine; LysoPE: Lysophosphatidylethanolamine; PC: Phosphatidylcholine; PE: Phosphatidylethanolamine; PE-Cer: Phosphatidylethanolamine-ceramide; PG: Phosphatidylglycerol; SM: Sphingomyelin; TG: Tri(acyllalkyl)glycerols.
[§]*m/z* mass charge ratio of detected feature including isotopes and salts, *m/z* of the parent in ().
[#]FC S vs NS: Fold-change smokers versus never-smokers.
[¶]LoA: Level of assignment.
NS: Never-smokers; S: Smokers.

Table 3. Metabolites discovered using NMR and UPLC–MS at pFDR ≤ 0.05 and with a public database identifier (cont.).

Generic name[†,‡]	Ion detected	*m/z* (parent)[§]	Platform	FC S vs NS[#]	pFDR	LoA[¶]	HMDB/ PubChem/CHEbi	KEGG	Lipid maps
SM(d18:2/24:1)	[M+H]+	811.7	RP+	0.77	0.0011	4			LMSP03010080
TG(12:0/18:2(9Z,12Z)/20:5(5Z,8Z,11Z,14Z,17Z))	[M+H]+	821.7	RP+	1.59	0.0007	2			LMGL03013492
TG(14:0/18:2(9Z,12Z)/20:5(5Z,8Z,11Z,14Z,17Z))	Isotope of [M+H]+	850.7 (869.7)	RP+	1.24	0.0007	2			LMGL03014392
TG(15:0/18:1/20:5)	[M+H]+	865.7	RP+	1.26	0.0006	2	HMDB43275		LMGL03015157
Tricosanamide	[M+Na]+	658.6 (636.6)	RP+	1.5	0.001	4	HMDB00950		LMSP02010021
Trimethyllysine	Isotope of [M+H]+	190.2 (189.2)	HILIC+	0.86	0.03	4	HMDB01325	C03793	
Vaccenyl carnitine	[M+Na]+	448.3 (426.4)	HILIC+	0.6	0.00003	4	HMDB06351		

[†]Metabolites identified in more than one dataset are highlighted in bold.
[‡]Lipids key: Cer: Ceramide; GlCer: Glyceryl ceramide; GalCer: Galactosylceramide; LysoPC: Lysophosphatidylcholine; LysoPE: Lysophosphatidylethanolamine; PC: Phosphatidylcholine; PE: Phosphatidylethanolamine; PE-Cer: Phosphatidylethanolamine-ceramide; PG: Phosphatidylglycerol; SM: Sphingomyelin; TG: Tri(acyllalkyl)glycerols.
[§]*m/z* mass charge ratio of detected feature including isotopes and salts, *m/z* of the parent in ().
[#]FC S vs NS: Fold-change smokers versus never-smokers.
[¶]LoA: Level of assignment.
NS: Never-smokers; S: Smokers.

LDL pool. (LDL-5 10.10–1.044 kg/l and LDL-6: 1.044–1.063 kg/l). In addition, Apolipoprotein-B in total serum and the overall LDL subfraction were both found higher in smokers compared with never-smokers.

Linear regression analysis against the TNEQ measure & BMI

Linear regression was also carried out using the results from all six sets of the metabolic phenotyping data against a measure of the TNEQ present in the urine [34]. TNEQ is an independent measure of how much nicotine was absorbed into the system. The smoking group only is considered for this analysis, since nicotine is not detected in never-smokers. This calculation was performed uncorrected and corrected for age, gender and BMI. The regression of the NMR data to TNEQ was not significant.

In addition, from the RP UPLC–MS (ESI-) data, the isotope of the parent ([M+H]+) ion of PE(O-18:1/20:4) was detected and this was negatively associated with TNEQ with a false discovery p-value of pFDR = 0.00596. The pFDR p-value of the parent ion was not significant. In the HILIC UPLC–MS (ESI+) data, there was a feature positively associated with TNEQ with a false discovery p-value of pFDR = 0.00287. On first inspection, this feature could be an isotope of the sodiated ([M+Na]+) ion of PE(O-18:1/20:4). The pFDR p-value of the sodiated ion itself was not significant. Thus, in both of these datasets the assignment of PE(O-18:1/20:4) discovered using UPLC–MS was based solely on the detection of an isotope ion. This equivocal result suggests that there may be coelution of another molecule in either or both of the datasets and that this molecule may be misassigned. Therefore, it was not used for further functional pathways analyses.

Linear regression of the lipoprotein fraction and RP UPLC MS (ESI+) data to TNEQ was not significant. To investigate a possible major confounder, a similar calculation was carried using BMI in the univariate model, uncorrected and corrected for age and gender. The ^{1}H-NMR data analysis showed no significant features being correlated with BMI. The lipoprotein analysis did reveal an increase in cholesterol (pFDR = 0.02905), phospholipids (pFDR = 0.02905) and apolipoprotein-B (pFDR = 0.02905) of the LDL subclass 5 (particle size: 1.040–1.044 kg/l) when regressed against BMI but the correlation was weak. Further information and statistics for these models can be found in Supplementary Figures 2 & 3.

Pathway enrichment analysis

A list of tentatively identified metabolites is shown in Table 3, and Supplementary Table 2 presents the metabolite ontologies based on HMDB and LIPID Maps®

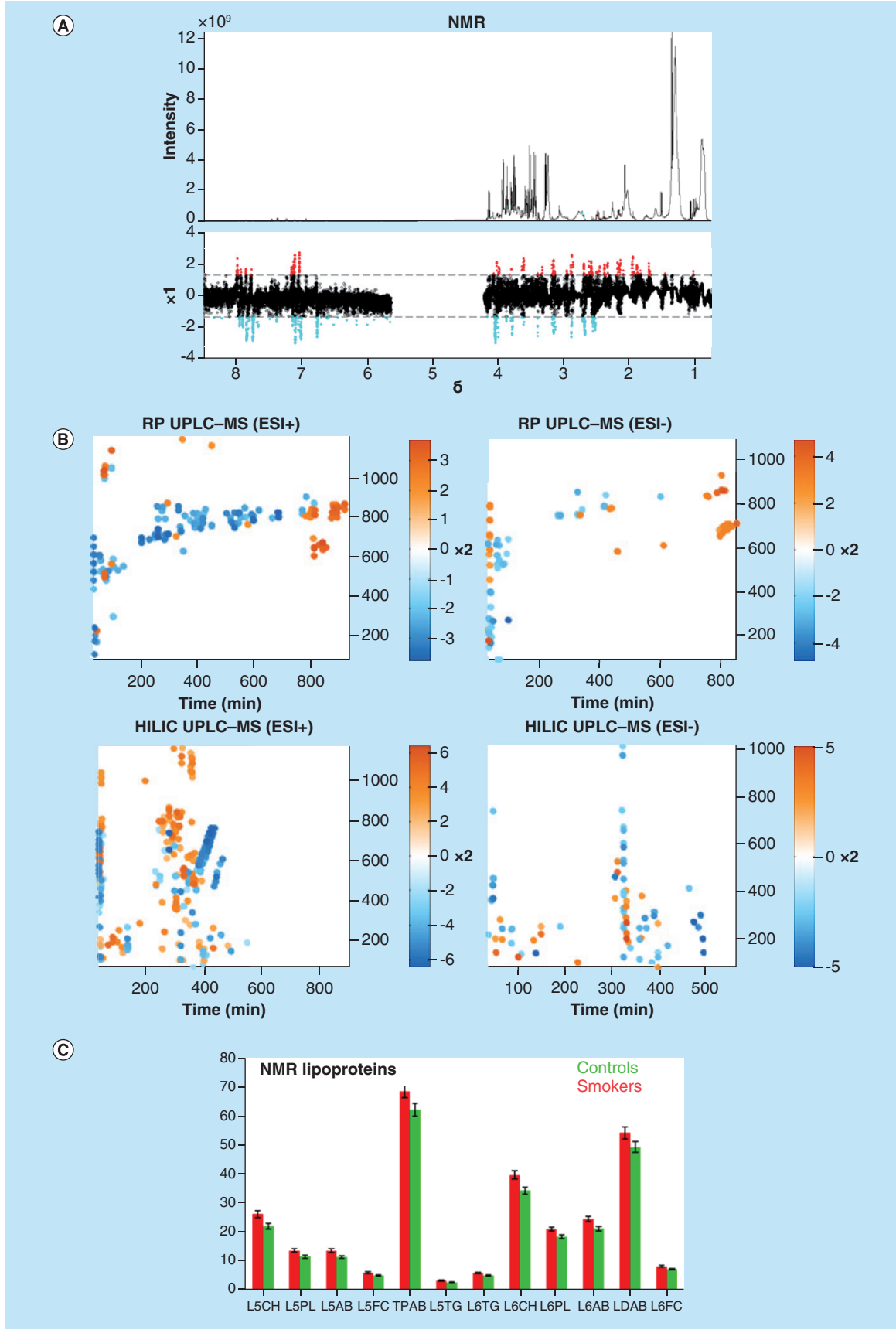
A
NMR
×10⁹
Intensity
×1
δ
B
RP UPLC–MS (ESI+)
RP UPLC–MS (ESI-)
HILIC UPLC–MS (ESI+)
HILIC UPLC–MS (ESI-)
Time (min)
×2
C
NMR lipoproteins
Controls
Smokers
L5CH L5PL L5AB L5FC TPAB L5TG L6TG L6CH L6PL L6AB LDAB L6FC

Figure 2. Differences in the metabolic phenotyping data between smoker and never-smoker groups after correction for confounders (BMI, gender age and drug intake) (cont.). **(A)** Manhattan plot of the NMR CPMG data. A significant p-value threshold of 0.05 was chosen after calculating the false discovery rate (FDR) and is marked with the dotted line. The red spots relate to NMR data points that highly correlate with the individuals in the smokers group, the blue spots are the data points that highly correlate to the never-smokers group. **(B)** Manhattan plots for all four UPLC–MS datasets showing the retention time on the x-axis and the *m/z* value on the y-axis for each significant (pFDR ≤ 0.05) feature. The color of the spot relates to the direction of the correlation (smokers, red and never-smokers, blue) and the shading shows the strength of the correlation (the darker the color the more significant). **(C)** Histogram showing the 12 lipoprotein subclasses (mean ± standard error) being significantly different (pFDR < 0.05) between never-smokers (green) and smokers (red). L5PL: phospholipids in LDL subclass 5, L5AB, apolipoprotein in LDL subclass 5, L5CH: cholesterol in LDL subclass 5, L5FC, free cholesterol in LDL subclass 5, L5TG: triglycerides in LDL subclass 5, TPAB: apolpoliprotein-B in total plasma, L6TG: triglycerides in LDL subclass 6, L6CH: cholesterol in LDL subclass 6, L6PL: phospholipids in LDL subclass 6, L6AB: apolipoprotein-B in LDL subclass-6, L6FC: free cholesterol in LDL subclass 6, LDAB: apolipoprotein-B in total LDL subclass. Particle size of LDL subclass 5 (LDL-5): 1.040–1.044 kg/l. Particle size of LDL subclass 6 (LDL-6): 1.044–1.063 kg/l. The lipoproteins were fractionated according to their density size by a sedimentation method adapted from literature by Bruker. The lipoprotein nomenclature is specific to the Bruker lipoprotein fractionation protocol list of 105 lipoprotein measurements in given in **Supplementary Table 1**.
CPMG: Carr-Purcell-Meiboom-Gill.

classification for the main compounds that were changed in smokers compared with never-smokers.

A functional enrichment analysis was conducted using MetaboAnalyst [35,36] and the tentatively identified metabolites at pFDR ≤ 0.05 in smokers relative to never-smokers with no fold-change filter criteria (50 single metabolites with HMDB IDs). MetaboAnalyst allowed the plotting of the enrichment profile as a function of the enrichment $-\log_2$ (p-value) and the pathway impact (%) (Figure 4). Perturbations in eight pathways were highlighted on the MetaboAnalyst plot with the four most significant features contained in the limits of a $-\log_2$ (p-value) threshold of 3 ($p < 0.05$) and a minimum of 10% impact (Figure 4). The enrichment for sphingolipid metabolism, glycerophospholipid metabolism and the various amino acids pathways were well aligned with the major HMDB and Lipid MAPS® ontology classes found in our dataset (Supplementary Table 2). If a fold-change filter criterion of 10% had been applied to conduct the MetaboAnalyst enrichment, the resulting 44 metabolites would have returned a very similar result in terms of enriched categories (results not shown). Overall similarly enriched categories are also obtained if only the metabolites tentatively identified with an LoA of 3 or lower are used, however, the reported pathway impact and significance is reduced (Supplementary Figure 4).

Discussion

Systems toxicology consists of the integrative study of the biological perturbations caused by toxicants using untargeted *in vitro* and *in vivo* screens combined with computational tools to determine possible toxicological outcomes. The development of next-generation tobacco and nicotine products offers a significant opportunity to reduce the burden on population health caused by tobacco use, but epidemiological data are currently lacking. In this context, systems toxicology could form part of a weight-of-evidence approach for next-generation tobacco product risk assessment. The first step would be to create a comprehensive map of tobacco smoking-related biological perturbations that can be interrogated for comparison against novel potentially reduced risk tobacco and nicotine products. The objective of the current study was to use a multiplatform metabolic phenotyping approach to gain a detailed insight into perturbed metabolic markers in healthy human smokers and apply a pathway enrichment analysis to identify functional changes associated with chronic smoking.

Both NMR and MS approaches were used in this study. The relative merits and sensitivities of the two approaches have been well documented previously and it is well known that the various platforms are highly complementary [40,41]. MS is generally more sensitive than NMR but the coverage of metabolic space is highly assay specific whereas, although NMR is less sensitive, a metabolite will give an NMR spectrum if it contains hydrogens [40]. For instance one of the NMR profiles gives the comprehensive information on lipoprotein subfractions while RP UPLC–MS is optimum for the individual lipid species that are embedded in many of the different lipoproteins [42], while HILIC UPLC–MS targets mainly polar molecules [40]. A broad metabolic phenotyping approach was chosen here to provide an exploratory overview of as wide a coverage as possible and this methodology is known to be semi-quantitative and reproducible [25–26,43]. It is possible to employ assays targeted at specific metabolites that would be more quantitative but this was not done herein as those would bias any subsequent functional enrichment analysis based on the selected targets. The multiplatform approach we used has found widespread application in large cohort studies [44] as multiple analytical techniques provide detection of a broad range of metabolites with diverse physicochemical properties.

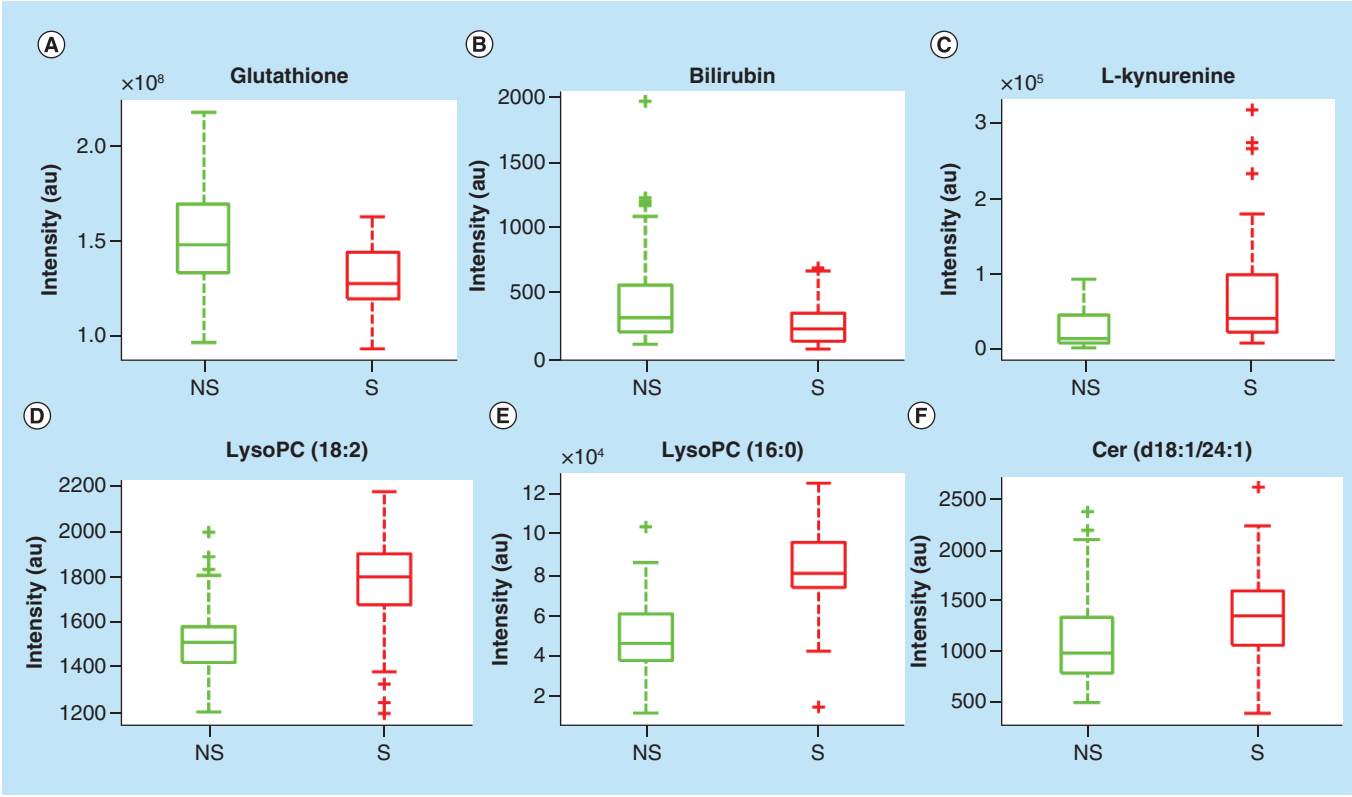

Figure 3. Changes in smoking-related metabolites in current and never-smokers. Box plots of the S and NS groups are illustrated for **(A)** glutathione, **(B)** bilirubin, **(C)** L-kynurenine, **(D)** LysoPC(18:2), **(E)** LysoPC(16:0) and **(F)** Cer(18:1/24:1). For each group, the five parameters are the lowest intensity of the metabolite, lower quartile, median, upper quartile and highest intensity. The points outside the quartiles are outliers.
The data used herein are normalized intensities expressed in arbitrary units (au).
NS: Never-smokers; S: Smokers.

Moreover, the complementarity of the various platforms is illustrated by the natural variance in the data shown in the PCA scores plots when color-coded by the various desired and confounder end points (Supplementary Figure 1). For example, when the points are color-coded by gender it can be seen for all six datasets there is a degree of clustering by gender with the variation due to this end point contributing to PC2. In this particular study HILIC UPLC–MS showed the best smoker/never-smoker separation in PCA (Supplementary Figure 1). It cannot be concluded, however, that this platform is sufficient to conduct a detailed comparison of smokers and never-smokers since all of the technology platforms provide some degree of separation for both the target end point and confounder end points based on complementary coverage of the metabolic space.

In terms of standardization and validation, the approach already in the literature was followed for NMR spectroscopy [26] and this has been shown to be highly reproducible. Similarly for the MS assays, protocols developed for high-throughput large-scale studies were followed with comprehensive use of QC samples as defined by Lewis *et al.* [43].

In this study, we used the Probabilistic Quotient Normalization method, which accounts for dilution effects of complex matrices by calculating the most probable dilution factors [29]. Because this normalization approach is not influenced by the change of single-peak intensities or by baseline distortion, it has been shown to perform better than two other commonly used metabolic phenotyping normalization methods in complex matrices, particularly in serum samples which contain large broad signals from lipids and proteins [29]. It is therefore ideal for subsequent data analysis.

In terms of validation of metabolite assignments, we adapted the approach proposed by Sumner *et al.* [34], which uses a 'LoA' confidence scale. A decreasing score value from 4 to 2 indicates increased confidence in the molecular identity of a metabolite by matching a combination of accurate mass, common fragment ions and common spectrum to repository databases. A score of 1 denote identification using a synthetic standard. Due to the size of our datasets non-novel metabolites were putatively identified, and those with an LoA of 2 had the highest degree of assignment short of full

identification by authentic standard. LoA assignment combined with fold change and statistical significance offers the advantage of enabling the prioritization of metabolites for further identification with authentic standards as this can be a time consuming step. All multivariate and univariate statistical models were validated using appropriate statistical measures, correcting for false positive and confounding variables, as is conventional in the literature, as for example shown in Elliot *et al.* [44].

Metabolites of biological interest were discovered from each of the datasets (^{1}H NMR, HILIC UPLC–MS, RP UPLC–MS and Bruker NMR lipoprotein profiles). The tentatively identified metabolites and LoA are presented in Table 3. An overview of the underlying biological changes occurring between smokers and never-smokers was obtained by functional enrichment analysis. Two annotation databases, namely HMDB, and Lipid MAPS®, were used to search for metabolites ontology using the list of tentatively identified metabolites (Supplementary Table 2). The enrichment analysis was performed with the metabolites significant at pFDR ≤ 0.05 with or without a fold-change criterion of more than 10% in smokers compared with never-smokers. The enriched functional pathways were visualized using MetaboAnalyst [35,36] and are shown in Figure 4. From these different analyses multiple pathways emerged as the top enriched categories impacted by smoking: sphingolipid and glycerophospholipid metabolism, amino acid metabolism (D-glutamate, glycine, serine, threonine, lysine, arginine) and aminoacyl t-RNA biosynthesis were highlighted. Those can be further categorized based on the current weight of evidence for their association with specific diseases. Sphingolipids and glycerophospholipids are modified by oxidative stress, which is involved in inflammation processes leading to cardiovascular diseases and emphysema [45,46]. Furthermore, other metabolic phenotyping studies in smokers using single analytical platforms or targeted approaches have also identified a subset of the pathways found in our analysis. The KORA (Cooperative Health Research in the region of Augsburg) study, which used a targeted metabolite approach screened 140 and 198 metabolites in serum, and identified a strong impact of smoking on glycerophospholipids, sphingolipids, amino acids, such as arginine and glycine and aminoacyl-tRNA metabo-

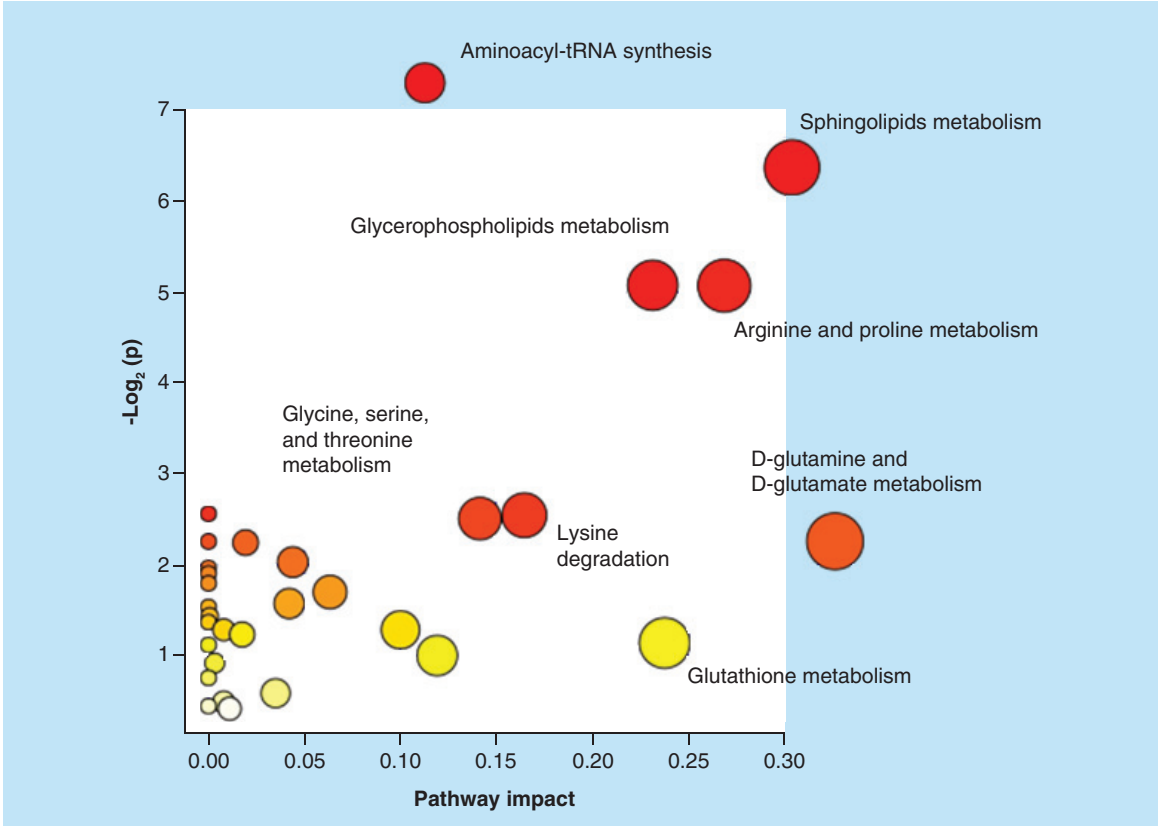

Figure 4. Functional pathway and ontologies enrichment analyses. MetaboAnalyst enrichment plot for metabolites significant at pFDR < 0.05. Top perturbed pathways based on impact and -log$_2$ (p-value) are labeled on the graph. The -log$_2$ (p-value) is the enrichment score. The impact score (0–1) indicated the pathway topological importance of the metabolites.

lism [19,47]. Monoacylglycerophosphocholine was one of the glycerophospholipid class proportionally more impacted (11 [14.5%] out of a total of 76 listed in Lipid Maps) by smoking. In this subclass of lipids LysoPC(16:0) is one example for which an increased level has been shown to be linked with carotid intima thickness in smokers [48]. In agreement with this observation LysoPC(16:0) was increased by 60% in our analysis (Figure 3). We also detected an increase of LysoPC(18:2) (Figure 3), one glycerophospholipid that has been demonstrated to have an association with the risk of coronary heart disease and mortality [45,49]. However, the reported association with risk of coronary heart disease was inversely related to the level of LysoPC(18:2) in serum and age-dependent [45]. On the other hand higher levels of LysoPCs have also been observed as a result of LDL oxidation which converts them into atherogenic particles [50]. This illustrates the complexity of understanding the dynamics of individual lipid markers with regards to risk. Yet, in general, glycerophospholipids are thought to have protective qualities for the cardiovascular system [49] and we observed a reduction of more than 10% for a total of 19 of them (Lipid MAPS®) across different subfamilies. Alteration of glycerophospholipids and glutamate were also reported in a UPLC-TOF-MS metabolomics cross-sectional clinical study with smokers [17], which also reported an effect on carbohydrate metabolism.

In addition, we observed an increase of small higher density LDL subclasses (LDL-5, 1.040–1.044 kg/l, LDL-6, 1.044–1.063 kg/l) and APO-B measured by NMR in our smoker group (Figure 2). Increased levels of small higher density LDL has been linked to higher risk of coronary heart diseases and carotid intima thickness [51,52]. The small and dense LDL particles are more likely to be glycated or oxidized than normal LDL particles. They can easily access the subendothelial space where they trigger inflammation and undergo a transformation into plaque, thus promoting atherosclerosis [53]. Finally, we also report a marked increase of L-kynurenine in our smoker group (Figure 3). L-kynurenine has been recently investigated as a marker of cardiovascular risk in particular in obese people who had an increase serum ratio L-kynurenine/tryptophan and is therefore potentially a promising biomarker of cardiovascular risk. There remain however some severe challenges to the identification of a panel of lipids and metabolites in smokers predictive of cardiovascular diseases, which would require multiple cohorts and a prospective design [54].

Our analysis also pointed toward a change in the abundance of sphingolipids including changes in the ceramide and sphingomyelin subclasses (Figure 4; Supplementary Table 2). More specifically, the ceramide Cer(d18:1/24:1[15Z]) appeared to be a strong responder to cigarette smoke with a 1.5-fold increase (Figure 3). The expression of COL4A3BP, a candidate gene in the etiology of COPD has been correlated with the level of Cer(d18:1/24:1(15Z)) in COPD patients [55], but to this date it is not known whether this sphingolipid is predictive of the onset of COPD in healthy smokers. Cer(d18:1/24:1(15Z)) is, therefore, an interesting candidate biomarker to follow in a longitudinal study to establish whether this lipid has a predictive value in early COPD diagnosis. Lower levels of the sphingomyelins SM(d18:1/14:0) and SM(d18:2/14:01) have also been reported in the serum of COPD patients [55]. We also detected lower levels of those two sphingolipids in our screen of healthy smokers, albeit those reductions were modest (Table 3) and there is currently no evidence that those lipids are predictive of disease. Sphingolipids are signaling molecules involved in processes, such as apoptosis, cell cycle and inflammation [43,44], which contribute to tobacco-specific diseases development, however, the balance between apoptotic and mitotic promotion by sphingolipids, such as ceramides and sphingomyelins in lung disease is not well understood. This has been reviewed in detail by Goldkorn *et al.* who suggested the involvement of ceramide-dependent exosome excretion of miRNA that regulate EGFR, a cell growth regulator [56] that plays a role in both COPD and lung cancer. For example they mentioned let-7a as a candidate miRNA associated with lung cancer poor clinical outcome and others reported that miR-124 was also correlated with tumor metastasis in non-small-cell lung cancer [56,57]. Interestingly, in a screen of 80 miRNA that we performed in the same serum samples, those were the two miRNAs that we found altered as a function of smoking status [22,58].

A recent global metabolic phenotyping study (using C18-UPLC–MS/MS and GC–MS) of blood (EDTA-plasma and serum) from 892 men and women from four studies identified metabolites related to cigarette smoking behavior in current smokers [59]. Twenty-four metabolites were statistically significant after Bonferroni correction based on p-values of fixed-effect meta-analyses ($0.05/700 = 7.1 \times 10^{-5}$). Fifteen metabolites were derived from xenobiotics possibly originating from tobacco smoking and coffee consumption. The endogenous metabolites implicated phenylalanine and tyrosine, benzoate and tryptophan metabolism [59]. Our study (Table 3) and the study from Gu *et al.* [59] both identified bilirubin as significantly lower in smokers. Bilirubin is a well-known blood antioxidant that is inversely correlated with increased risk of coronary heart disease and lung cancer [60–62]. Glu-

tathione was another antioxidant highlighted in our study to be significantly lower in the serum of smokers (Figure 3), which other studies have linked to a higher risk of cardiovascular disease [63]. The smaller number of metabolites identified in the large study described above can originate from the use of the highly stringent Bonferroni false discovery rate and the combination of samples obtained from multiple studies with different designs and objectives. Furthermore, our study included methods aimed at lipid profiling, which in our case formed the majority of features differentiating smokers and never-smokers.

We also conducted linear regression analysis of the metabolic data to TNEQ and BMI. Linear regression of the metabolic and lipoprotein data in our study has shown no correlation with nicotine dose (TNEQ) while three lipoproteins (cholesterol, phospholipids and apolipoprotein-B of the LDL subclass 5) were weakly correlated with BMI (Supplementary Figures 2 & 3). The lack of correlation with nicotine dose is not entirely surprising since TNEQ was measured in urine not in serum and TNEQ is only representative of the smoking behavior over the 24 h prior to sampling due to its short half-life. Endogenous metabolic perturbations are likely to accumulate over a prolonged period of time on a different scale compared with TNEQ. Therefore, correlation with TNEQ might not yield meaningful results regarding the dose–response relationship. A larger study group might be required to establish clear dose–response relationships or the use of a blood marker of smoke exposure, such as the acrylonitrile hemoglobin adduct, which is representative of cigarette consumption over a period of several months.

The strength of our study lies in the use of multiple platforms giving extensive coverage of the serum metabolome with robust statistical analysis but it also has limitations. The way in which the study was conducted did not lead itself to trying to identify whether each individual could be predicted to be a smoker or a never-smoker. We were interested in finding metabolites that statistically could distinguish the groups, smoker versus never-smoker. Given the modest sample numbers and the large number of features we took care to validate the multivariate models carefully and robustly. This precluded the generation of training and tests sets that could be used for class prediction or individual status prediction. The multivariate models were used to show class trends but the significant metabolites were identified by using univariate modeling but with comprehensive removal of variance caused by confounding variables. Of course, if the subject had smoked within the previous 48 h of giving a sample, then the nicotine metabolites present would be the best way to identify an individual's status. This is another reason why we did not do individual predictions in this study.

Our enrichment analysis using MetaboAnalyst is a qualitative assessment and therefore it does not predict whether a pathway is up or down regulated, but is more representative of the topological perturbations of a pathway. Tools, such as IPA Ingenuity®, offer the possibility to perform downstream causal prediction based on reference pathways [64] and inclusive of fold change, however the knowledge-base of Ingenuity currently includes very limited data on lipids. The relevance of pathways identified in our enrichment analysis, such as aminoacyl-tRNA biosynthesis, should be considered carefully. Indeed, there is a clear overlap between the metabolites found in those and pathways relating to amino acids metabolism, and furthermore the number of signature metabolites mapping to some pathways can be low. Other key factors to consider in interpreting the data are the fold change of the tentatively identified metabolites, and the level of confidence in the metabolite identification (LoA rank). We opted for a pFDR < 0.05 and a 10% fold change filter to conduct part of our enrichment analysis and we cannot exclude that a number of included metabolites might be false positives. Filters are arbitrary in nature and it is not uncommon to find metabolic phenotyping studies where, for example, only a statistical significance filter is used [65,66]. A higher fold-change filter and lower pFDR value might yield greater confidence in the metabolites used to predict perturbed pathways and functions with a trade-off of potentially eliminating others that might turn out to be biologically relevant. Finally, in this study no confirmation of metabolite identification using synthetic standard materials was performed and therefore it is possible that some metabolites especially with a high LoA might be reclassified. Therefore, this study primarily illustrates a multiplatform strategy from samples acquisition, analysis, selection of features of interests with an example of pathway enrichment application. The identification stage is in itself a significant piece of work and as shown in Table 2, many features remain unassigned. Addressing this would increase the robustness and granularity of any subsequent enrichment analysis. As more and more detailed spectra populate public repositories it is expected that the confidence in the identification of metabolic features will be increasingly facilitated.

Finally, our study was performed with a relatively small number of subjects with an overrepresentation of the Caucasian ethnic background and no clinical confinement, our results are in close agreement with the KORA study, which was conducted on a large

group of Korean smokers [19]. This indicates that those metabolic alterations are present independently of the ethnic group. Furthermore, in light of our results and the KORA study a full clinical confinement to control for confounding factors, such as diet, does not appear to be required to detect smoking-specific events in serum. However, as a minimum, a diary of taken medications should be kept for possible confounder corrections.

Conclusion

Applying a multiplatform, metabolic-phenotyping approach produced a rich dataset that in combination with a knowledge-based enrichment analysis gave mechanistic insights into biological alterations potentially associated with tobacco-related diseases in healthy smokers. The dataset and enrichment analysis are in agreement with previously published single platform metabolomics, lipidomics and biomarkers studies pointing toward an impact of smoking on antioxidants, such as glutathione, bilurubin and lipids involved in apoptosis and cardiovascular diseases. Our study together with the few other related studies in smokers contributes to strengthen the weight of evidence of functional alterations caused by smoking. The reversibility of these alterations in metabolic pathways should be investigated for their use as risk assessment benchmarks in studies where smokers are switched to next-generation tobacco and nicotine products, such as electronic cigarettes. Further validation of the candidate metabolites could be performed by evaluating whether a panel of identified metabolites can predict the smoking status of test subjects and which ones are predictive of disease onset in prospective clinical studies.

Future perspective

The development of next-generation tobacco and nicotine products offers a significant opportunity to reduce the burden of tobacco use on population health, but epidemiological data are currently lacking. In this context, systems toxicology could form part of a weight of evidence approach for tobacco product risk assessment. The use of metabolic phenotyping tools in clinical studies of smokers who switched to those novel devices or quit smoking altogether should provide further insights into the reversibility of those pathway perturbations and potential benefits of switching. Application of these approaches to *in vitro* cell systems, such as reconstituted airway epithelium exposed to aerosols, could further complement the risk assessment of next-generation tobacco and nicotine devices.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science/doi/full/10.4155/bio-2016-0108

Acknowledgements

We thank J Shepperd, M McEwan, N Newland and A Eldridge for their assistance with the clinical sample collection and biobanking. We thank Bruker Biospin for generous provision of the lipoprotein analysis of the NMR spectra and for detailed discussions and help in their interpretation

Financial & competing interests disclosure

This study was funded by British American Tobacco to Metabometrix Ltd. as a commercial contract. EF Minet was an employee of British American Tobacco during the conduct of this study. British American Tobacco manufactures the products used by the smokers in this study. The authors have no other relevant affiliations or financial involvement with any

Executive summary

- Perturbed biochemical functions associated with tobacco smoking can be investigated using omics platforms coupled with knowledge-based bioinformatics tools.

Methods

- A comprehensive multiplatform metabolic phenotyping and lipidomics comparison was applied to the serum of smokers (n = 55) and never-smokers (n = 57) with multivariate and univariate analyses.

Results

- Seventy-one unique metabolites (pFDR ≤ 0.01) were tentatively identified of which 50 varied by 10% or more.
- The metabolites were used as input for pathway enrichment to model adverse biological events.
- Changes in lipids and amino acid metabolism are reported.

Discussion & perspective

- Multiplatform metabolic phenotyping generates a comprehensive metabolic profile allowing in depth group comparison.
- Over-represented classes of compounds and affected biological functions can be modeled with knowledge-based analyses.
- Such approaches can be used to generate benchmark metabolic profiles in risk assessment of next-generation nicotine products.

organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Open access

References

1 Alberg AJ, Samet JM. Epidemiology of lung cancer. *Chest* 123(Suppl. 1), S21–S49 (2003).

2 Diaz-Guzman E, Mannino DM. Epidemiology and prevalence of chronic obstructive pulmonary disease. *Clin. Chest Med.* 35(1), 7–16 (2014).

3 Paffenbarger RS Jr, Hyde RT, Wing AL, Hsieh C. Cigarette smoking and cardiovascular diseases. *IARC Sci. Publ.* (74), 45–60 (1986).

4 Hall W, Gartner C, Forlini C. Ethical issues raised by a ban on the sale of electronic nicotine devices. *Addiction* 110(7), 1061–1067 (2015).

5 Protano C, Di Milia LM, Orsi GB, Vitali M. Electronic cigarette: a threat or an opportunity for public health? State of the art and future perspectives. *Clin. Ter.* 166(1), 32–37 (2015).

6 Lowe FJ, Luettich K, Gregg EO. Lung cancer biomarkers for the assessment of modified risk tobacco products: an oxidative stress perspective. *Biomarkers* 18(3), 183–195 (2013).

7 Valavanidis A, Vlachogianni T, Fiotakis C. 8-hydroxy-2′ -deoxyguanosine (8-OHdG): a critical biomarker of oxidative stress and carcinogenesis. *J. Environ. Sci. Health C. Environ. Carcinog. Ecotoxicol. Rev.* 27(2), 120–139 (2009).

8 Ansari S, Baumer K, Boue S *et al.* Comprehensive systems biology analysis of a 7 month cigarette smoke inhalation study in C57BL/6 mice. *Sci. Data* 3, 150077 (2016).

9 Phillips B, Veljkovic E, Boue S *et al.* An 8-Month Systems Toxicology Inhalation/Cessation Study in Apoe-/- mice to investigate cardiovascular and respiratory exposure effects of a candidate modified risk tobacco product, THS 2. 2, compared with conventional cigarettes. *Toxicol. Sci.* 151(2), 462–464 (2016).

10 Zhu X, Gerstein M, Snyder M. Getting connected: analysis and principles of biological networks. *Genes Dev.* 21(9), 1010–1024 (2007).

11 Han JD. Understanding biological functions through molecular networks. *Cell Res.* 18(2), 224–237 (2008).

12 Almaas E. Biological impacts and context of network theory. *J. Exp. Biol.* 210(Pt 9), 1548–1558 (2007).

13 Vallabhajosyula RR, Raval A. Computational modeling in systems biology. *Methods Mol. Biol.* 662, 97–120 (2010).

14 Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* 17(7), 451–459 (2016).

15 Dunn WB, Lin W, Broadhurst D *et al.* Molecular phenotyping of a UK population: defining the human serum metabolome. *Metabolomics* 11, 9–26 (2015).

16 Garcia-Perez I, Lindon JC, Minet E. Application of CE–MS to a metabonomics study of human urine from cigarette smokers and non-smokers. *Bioanalysis* 6(20), 2733–2749 (2014).

17 Hsu PC, Zhou B, Zhao Y *et al.* Feasibility of identifying the tobacco-related global metabolome in blood by UPLC-QTOF-MS. *J. Proteome. Res.* 12(2), 679–691 (2013).

18 Muller DC, Degen C, Scherer G, Jahreis G, Niessner R, Scherer M. Metabolomics using GC-TOF-MS followed by subsequent GC-FID and HILIC–MS/MS analysis revealed significantly altered fatty acid and phospholipid species profiles in plasma of smokers. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 966, 117–126 (2014).

19 Xu T, Holzapfel C, Dong X *et al.* Effects of smoking and smoking cessation on human serum metabolite profile: results from the KORA cohort study. *BMC Med.* 11, 60 (2013).

20 Shepperd CJ, Newland N, Eldridge A, Graff D, Meyer I. A single-blinded, single-centre, controlled study in healthy adult smokers to identify the effects of a reduced toxicant prototype cigarette on biomarkers of exposure and of biological effect versus commercial cigarettes. *BMC Public Health* 13, 690 (2013).

21 Shepperd CJ, Newland N, Eldridge A *et al.* Changes in levels of biomarkers of exposure and biological effect in a controlled study of smokers switched from conventional cigarettes to reduced-toxicant-prototype cigarettes. *Regul. Toxicol. Pharmacol.* 72(2), 273–291 (2015).

22 Banerjee A, Waters D, Camacho OM, Minet E. Quantification of plasma microRNAs in a group of healthy smokers, ex-smokers and non-smokers and correlation to biomarkers of tobacco exposure. *Biomarkers* 20(2), 123–131 (2015).

23 Haswell LE, Papadopoulou E, Newland N, Shepperd CJ, Lowe FJ. A cross-sectional analysis of candidate biomarkers of biological effect in smokers, never-smokers and ex-smokers. *Biomarkers* 19(5), 356–367 (2014).

24 European Bioinformatics Institute. www.ebi.ac.uk/metabolights/

25 Want EJ, Wilson ID, Gika H *et al.* Global metabolic profiling procedures for urine using UPLC–MS. *Nat. Protoc.* 5(6), 1005–1018 (2010).

26 Dona AC, Jimenez B, Schafer H *et al.* Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal. Chem.* 86(19), 9887–9894 (2014).

27 Isaac G, McDonald S, Astarita G. Waters application note. www.waters.com

28 Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78(3), 779–787 (2006).

29 Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. *Anal. Chem.* 78(13), 4281–4290 (2006).

30 Petersen M, Dyrby M, Toubro S *et al.* Quantification of lipoprotein subclasses by proton nuclear magnetic resonance-based partial least-squares regression models. *Clin. Chem.* 51(8), 1457–1461 (2005).

31 Smith CA, O'Maille G, Want EJ *et al.* METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* 27(6), 747–751 (2005).

32 Wishart DS, Jewison T, Guo AC *et al.* HMDB 3.0 – the human metabolome database in 2013. *Nucleic Acids Res.* 41, D801–D807 (2013).

33 LIPID MAPS, LIPID metabolites and Pathway Strategy, Welcome Trust. www.lipidmaps.org/

34 Sumner LW, Amberg A, Barrett D *et al.* Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* 3(3), 211–221 (2007).

35 Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. MetaboAnalyst 2.0 – a comprehensive server for metabolomic data analysis. *Nucleic Acids Res.* 40, W127–W133 (2012).

36 Xia J, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0 – making metabolomics more meaningful. *Nucleic Acids Res.* 43(W1), W251–W257 (2015).

37 PubChem, NIH. www.pubchem.ncbi.nlm.nih.gov/

38 ChEBI (Chemical Entities of Biological Interest), The European Molecular Biology Laboratory. www.ebi.ac.uk/chebi/

39 KEGG (Kyoto Encyclopedia of Genes and Genomres, GenomeNet. www.genome.jp/kegg/

40 Lindon JC, Nicholson JK. Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)* 1, 45–69 (2008).

41 Buscher JM, Czernik D, Ewald JC, Sauer U, Zamboni N. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. *Anal. Chem.* 81(6), 2135–2143 (2009).

42 Kofeler HC, Fauland A, Rechberger GN, Trotzmuller M. Mass spectrometry based lipidomics: an overview of technological platforms. *Metabolites* 2(1), 19–38 (2012).

43 Lewis MR, Pearce JTM, Spagou K *et al*. Development and applicationof UPLC-TOF MS for precision large scale urinarymetabolic phenotyping. *Anal. Chem.* doi:10.1021/acs.analchem.6b01481 (2016) (Epub ahead of print).

44 Elliott P, Posma JM, Chan Q *et al.* Urinary metabolic signatures of human adiposity. *Sci. Transl. Med.* 7, 285 (2015).

45 Ganna A, Salihovic S, Sundstrom J *et al.* Large-scale metabolomic profiling identifies novel biomarkers for incident coronary heart disease. *PLoS. Genet.* 10, 12 (2014).

46 Telenga ED, Hoffmann RF, Ruben t *et al.* Untargeted lipidomic analysis in chronic obstructive pulmonary disease. Uncovering sphingolipids. *Am. J. Respir. Crit. Care Med.* 190(2), 155–164 (2014).

47 Wang-Sattler R, Yu Y, Mittelstrass K *et al.* Metabolic profiling reveals distinct variations linked to nicotine consumption in humans – first results from the KORA study. *PLoS ONE* 3(12), e3863 (2008).

48 Fratta PA, Stranieri C, Pasini A *et al.* Lysophosphatidylcholine and carotid intima-media thickness in young smokers: a role for oxidized LDL-induced expression of PBMC lipoprotein-associated phospholipase A2? *PLoS ONE* 8(12), e83092 (2013).

49 Sigruener A, Kleber ME, Heimerl S, Liebisch G, Schmitz G, Maerz W. Glycerophospholipid and sphingolipid species and mortality: the Ludwigshafen Risk and Cardiovascular Health (LURIC) study. *PLoS ONE* 9(1), e85724 (2014).

50 Rozenberg O, Shih DM, Aviram M. Human serum paraoxonase 1 decreases macrophage cholesterol biosynthesis: possible role for its phospholipase-A2-like activity and lysophosphatidylcholine formation. *Arterioscler. Thromb. Vasc. Biol.* 23(3), 461–467 (2003).

51 Carmena R, Duriez P, Fruchart JC. Atherogenic lipoprotein particles in atherosclerosis. *Circulation* 109(23 Suppl. 1), III2–III7 (2004).

52 Liu ML, Ylitalo K, Nuotio I, Salonen R, Salonen JT, Taskinen MR. Association between carotid intima-media thickness and low-density lipoprotein size and susceptibility of low-density lipoprotein to oxidation in asymptomatic members of familial combined hyperlipidemia families. *Stroke* 33(5), 1255–1260 (2002).

53 Kwiterovich PO Jr. Lipoprotein heterogeneity: diagnostic and therapeutic implications. *Am. J. Cardiol.* 90(8A), 1i–10i (2002).

54 Stegemann C, Pechlaner R, Willeit P *et al.* Lipidomics profiling and risk of cardiovascular disease in the prospective population-based Bruneck study. *Circulation* 129(18), 1821–1831 (2014).

55 Bowler RP, Jacobson S, Cruickshank C *et al.* Plasma sphingolipids associated with chronic obstructive pulmonary disease phenotypes. *Am. J. Respir. Crit. Care Med.* 191(3), 275–284 (2015).

56 Goldkorn T, Chung S, Filosto S. Lung cancer and lung injury: the dual role of ceramide. *Handb. Exp. Pharmacol.* (216), 93–113 (2013).

57 Zhang Y, Li H, Han J, Zhang Y. Down-regulation of microRNA-124 is correlated with tumor metastasis and

poor prognosis in patients with lung cancer. *Int. J. Clin. Exp. Pathol.* 8(2), 1967–1972 (2015).

58 Uhlmann S, Mannsperger H, Zhang JD *et al.* Global microRNA level regulation of EGFR-driven cell-cycle protein network in breast cancer. *Mol. Syst. Biol.* 8, 570 (2012).

59 Gu F, Derkach A, Freedman ND *et al.* Cigarette smoking behaviour and blood metabolomics. *Int. J. Epidemiol.* pii: dyv330 (2015)(Epub ahead of print).

60 Lin JP, O'Donnell CJ, Schwaiger JP *et al.* Association between the *UGT1A1*28* allele, bilirubin levels, and coronary heart disease in the Framingham Heart Study. *Circulation* 114(14), 1476–1481 (2006).

61 Stocker R, Yamamoto Y, McDonagh AF, Glazer AN, Ames BN. Bilirubin is an antioxidant of possible physiological importance. *Science* 235(4792), 1043–1046 (1987).

62 Wen CP, Zhang F, Liang D *et al.* The ability of bilirubin in identifying smokers with higher risk of lung cancer: a large cohort study in conjunction with global metabolomic profiling. *Clin. Cancer Res.* 21(1), 193–200 (2015).

63 Shimizu H, Kiyohara Y, Kato I *et al.* Relationship between plasma glutathione levels and cardiovascular disease in a defined population: the Hisayama study. *Stroke* 35(9), 2072–2077 (2004).

64 Kramer A, Green J, Pollard J Jr., Tugendreich S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* 30(4), 523–530 (2014).

65 Lai YS, Chen WC, Kuo TC *et al.* Mass-spectrometry-based serum metabolomics of a C57BL/6J mouse model of high-fat-diet-induced non-alcoholic fatty liver disease development. *J. Agric. Food Chem.* 63(35), 7873–7884 (2015).

66 Schicho R, Shaykhutdinov R, Ngo J *et al.* Quantitative metabolomic profiling of serum, plasma, and urine by (1)H NMR spectroscopy discriminates between patients with inflammatory bowel disease and healthy individuals. *J. Proteome Res.* 11(6), 3344–3357 (2012).