Effective Carbon Number and Inter-Class Retention Time Conversion Enhances Lipid Identifications in Untargeted Clinical Lipidomics

Jake B. White[1,2,3], Paul J. Trim[1,3] ,Thalia Salagras[2], Aaron Long[1], Peter J. Psaltis[1,2], Johan W. Verjans[1,2], Marten F. Snel* [1,3]

1. Adelaide Medical School, Faculty of Health and Medical Sciences, University of Adelaide, South Australia, Australia.

2. Vascular Research Centre, South Australian Health and Medical Research Institute, South Australia, Australia;
3. Proteomics, Metabolomics and MS-Imaging Core Facility, South Australian Health and Medical Research Institute, South Australia, Australia.

* Corresponding Author - Email Address: marten.snel@sahmri.com

Abstract

Chromatography is often used as a method for reducing sample complexity prior to analysis by mass spectrometry, the use of retention time (RT) is becoming increasingly popular to add valuable supporting information in lipid identification.

The RT of lipids with the same headgroup in reverse-phase separation can be predicted using the effective carbon number (ECN) model. This model describes the effect of acyl chain length and degree of saturation on lipid RT, which increases predictably with acyl chain length and degree of saturation. Furthermore, we have found a robust correlation in the chromatographic separation of lipids with different headgroups that share the same fatty acid motive. By measuring a small number of lipids from each subclass it is possible to build a model that allows for the prediction of the RT of one lipid subclass based on another.

Here, we utilise ECN modelling and inter-class retention time conversion (IC-RTC) to build a glycerophospholipid RT library with 481 entries based on 136 MS/MS characterised lipid RTs from NIST SRM-1950 plasma and lipid standards.

The library was tested on a patient cohort undergoing coronary artery bypass grafting surgery (n=37). A total of 129 unique circulating glycerophospholipids were identified, of which, 57 (4 PC, 24 PE, 4 PG, 15 PI, 10 PS) were detected with IC-RTC, thereby demonstrating the utility of this technique for the identification of lipid species not found in commercial standards.

Introduction

Lipids are a complex class of molecules that play a pivotal role in the structure and function of living cells, including membrane structure, energy storage, inflammation, protein folding and aspects of second messenger signaling [reviewed in (1–3)]. The well-established role of circulating cholesterol contained in low-density lipoproteins (LDL) has been essential to developing our current understanding of atherosclerotic cardiovascular disease and continues to be used as a surrogate measurement for cardiovascular disease risk in clinical practice (4, 5). Improvements in mass spectrometry (MS) technology and data analysis methods have caused an influx of clinical cardiovascular lipidomics research (6). Among other findings, lipidomics has identified lipoprotein subclass-specific moieties which may contribute to aggregation or inflammation (7), identified total- and LDL-cholesterol independent changes in the circulating lipidome that are correlated with atherosclerotic cardiovascular disease (8–12), and has been shown to improve recurrent event prediction in patients with clinical cardiovascular disease (13).

Traditional analysis of the lipidome often relies on the generation of fragmentation information from nominated masses to identify compounds with MS techniques. This is rigorous and allows for thorough compound characterization; however, targeted acquisition methods (e.g., multiple reaction monitoring (MRM)) require the selection of target lipids before analysis, and any compounds not specified are not measured. Conversely, the identification of compounds from untargeted data, such as MS-only, time of flight (ToF) data, is difficult without MS/MS data for verification. Therefore, most MS approaches must consider the advantages and disadvantages between targeted and untargeted methodologies before acquisition.

In recent years, major advances have improved the sensitivity and selectivity of MS based methods for the analysis of lipids. Alongside this, utilizing the chromatographic dimension of separation allows for the prediction of retention times (RT) of compounds based on their physicochemical properties, which has been explored extensively in metabolomics and lipidomics previously (14–21). Reverse-phase separation allows for the prediction of RT by using the effective carbon number (ECN) model, which describes the tendency of species to elute earlier with fewer carbons in the acyl chain and as the degree of unsaturation increases (22–24)

Despite these advances, there is a great degree of heterogeneity between measurement techniques, sample preparation, and use of internal standards in MS-based lipidomics protocols (25, 26). Therefore, appropriate method validation must be performed to determine the reliability and accuracy of measurements. The most widely applicable methods for verification involve the use of stable isotope-labelled internal standards for quantification, and standard reference materials to check measurements against established consensus.

Here we describe an approach that leverages hydrophobic interaction (HILIC) chromatography, which is selective for lipid headgroup rather than acyl chain, to identify abundant phosphatidylcholine (PC) species in the National Institute of Standards (NIST) Standard Reference Material-1950 (SRM 1950) pooled blood plasma. Abundant species were analyzed with Quadrupole-ToF (QToF)-MS/MS with reverse phase chromatography to build a retention time library. We then utilize the predictable RT behavior of lipids in reverse-phase liquid chromatography to predict the RT of lysophosphatidylglycerol (LPG), lysophosphatidylethanolamine (LPE), lysophosphatidylcholine (LPC), phosphatidylethanolamine (PE), phosphatidylglycerol (PG), phosphatidylinositol (PI) and phosphatidylserine (PS) species from the gradient conditions of a previously published method (19) to create a glycerophospholipid (GPL) RT library. The use of interclass-retention time conversion (IC-RTC), demonstrated for the first time here, also allows for resolution of acyl chain composition for several lipid species without using MS/MS, where they are separated in the chromatographic dimension. Additionally, this method preserves the data from unannotated compounds detected by ToF-MS, allowing for retrospective identification of new compounds. By combining ECN and IC-RTC, it is possible to generate large, accurate lipid databases from a much smaller set of thoroughly characterized measurements. This can be used to expand the number of lipid species measured from a sample cohort post acquisition. To determine the efficacy of IC-RTC in a clinical context, we applied this technique to detect circulating lipid species that could not be detected in the NIST SRM-1950 plasma sample.

Materials and Methods

Chemicals

EquiSPLASH and UltimateSPLASH ONE lipidomics standards were purchased from Avanti Polar Lipids (Alabaster, AL). Formic acid, ammonium formate, 2-propanol, acetonitrile, and water were purchased from Honeywell (Charlotte, NC). All reagents used were LC-MS grade.

A stock solution of EquiSPLASH solution was prepared in a 1:173 v/v dilution in 1:1 v/v isopropyl alcohol: acetonitrile for lipid extraction (Extraction Buffer).

Plasma Samples

NIST SRM-1950 blood plasma standard was purchased from Merck (Darmstadt, Germany). Human blood plasma was collected from patients undergoing coronary artery bypass grafting (CABG) surgery at the Royal Adelaide Hospital (n = 37). Clinical data were gathered from patient records and case notes at the time of consent. Specimen and data collection were in accordance with approved human ethics (CALHN ethics number R20180206). All patients were fasted for 8 h prior to sample collection. Patient

characteristics were as follows: the mean age was 65.7 ± 9.5 years, 73% of participants were male, and 54% of patients had diabetes.

Clinical samples were assigned an internal ID and injected in the following sequence: six reagent blank injections to bed in the column, two injections of extracted NIST SRM 1950 plasma sample (one spiked with internal standard, one without), one injection of EquiSPLASH internal standard spiked extraction buffer, followed by the clinical samples in randomized order. NIST SRM 1950 plasma was used as a quality control (QC) sample and injected after every 10 samples, immediately followed by a reagent blank.

Lipid Extraction

Plasma samples were removed from -80°C storage and were thawed at 4°C overnight prior to extraction. 174 µL of internal standard spiked extraction buffer was added to 6 µL of blood plasma. Samples were vortexed for 3 s and centrifuged for 15 min *ca* 16,000 × *g* and the supernatant aliquoted into QuanRecovery plates (Waters, Milford, MA). NIST SRM-1950 spiked with EquiSPLASH aliquots were used for library building. For the identification of retention characteristics of other GPL sub-classes, 1 µL of UltimateSPLASH ONE was spiked into 173 µL of extraction buffer, which was used to extract an aliquot of NIST plasma. For characterization of the effects of double bond position and stereochemistry on RT, three separate 1:500 dilutions of 1 mg/mL Avanti PC 18:1(9E)/18:1(9E) (850376), PC 18:1(9Z)/18:1(9Z) (850375), and PC 18:1(6Z)/18:1(6Z) (850374) standards (in methanol) were spiked into EquiSPLASH spiked extraction buffer.

HILIC-MS Candidate Discovery

Samples were mass analyzed on a Xevo G2-XS Q-ToF (Waters, Milford, MA) coupled to an Acquity I-Class-FTN UPLC system (Waters, Milford, MA) in negative ion mode. Lipids were separated with a 100 mm Waters Acquity BEH HILIC column (2.1 mm ID, 1.7 µm particle size). Sample chamber temperature was set to 8 °C and the column heater was set to 50 °C. 10 µL of extracted NIST SRM 1950 plasma was injected on column. Mobile Phase A was a 10 mM ammonium acetate solution in 95:5 acetonitrile:water v/v. Mobile Phase B was a 10 mM ammonium acetate solution made up in 50:50 acetonitrile:water v/v. Gradient conditions are described in **Table 1**.

Samples were mass analyzed in negative ion MS^E acquisition mode, where collision voltage was switched between elevated and low energy every 0.1 s and the resultant data written to two different functions. Capillary voltage was set to 3.0 kV. Collision voltage was set to 6 V in the low CE channel and set to 30 V in the elevated CE channel. Prior to data acquisition the mass spectrometer was calibrated between 50-2000 Da with sodium formate, with an RMS residual mass error of 0.4 ppm. Mass calibration was

maintained using leucine enkephalin lockspray which was sampled for 0.2 s every 30 s. Purge solution was composed of 1:1:1:1 water:acetonitrile:isopropyl alcohol:methanol. Seal wash solution was 10% methanol in water. Needle wash solution was 50% acetonitrile in water.

Time (min)	Flow Rate (mL/min)	% Mobile Phase B
0.0	0.4	90
0.1	0.4	43
5.0	0.4	50
5.2	0.4	54
7.2	0.4	70
7.4	0.4	99

Table 1: LC gradient for HILIC method.

RP-LC-MS & MS/MS Conditions

Lipid extracts were separated using a 100 mm Waters Acquity Premier CSH C18 column (2.1 mm ID, 1.7 µm particle size) heated to 55 °C. Mobile phase A was prepared by combining 1200 mL of acetonitrile, 800 mL of water, 2 mL of formic acid, and 1260.3 mg of ammonium formate. Mobile phase B was prepared by combining 1800 mL of isopropyl alcohol with 200 mL of acetonitrile, 2 mL of formic acid and 1260.3 mg of aqueous ammonium formate (dissolved in 2 mL of water). Sample chamber temperature was set to 8 °C. Conditions for the LC gradient are specified in **Table 2**.

Time (min)	Flow Rate (mL/min)	% Mobile Phase B
0.0	0.4	40
2.0	0.4	43
2.1	0.4	50
12.0	0.4	54
12.1	0.4	70
18.0	0.4	99
20.4	0.4	99
20.5	0.4	40
22.5	0.4	40

Table 2: Gradient conditions for reverse-phase LC method.

The mass spectrometer was calibrated between 50-2000 Da with 20 μg/mL sodium formate in 50% acetonitrile with 0.1% formic acid. RMS residual mass error was 0.7 ppm before acquisition. For MS/MS analysis, a quadrupole isolation window of *ca* ±0.6 Da was set using the LM/HM settings of 20/15. Candidate masses were analyzed in sets of 15 unscheduled MS/MS transitions with 30 V collision voltage and 0.019 s scan time, where one transition in each injection always contained the PC 18:1(d7)/15:0 internal standard present in the extraction buffer. Mass calibration was maintained using leucine enkephalin lockspray which was sampled for 0.2 s every 30 s. A 2 μL EquiSPLASH-spiked extraction of NIST SRM-1950 plasma was used for library building. For clinical samples, analysis was performed using negative ion MS^E acquisition mode with 0.1 s scan time. Capillary voltage was set to 3.0 kV. Collision voltage was turned off in the low CE channel and 30 V in the elevated CE channel. A 2 μL injection of UltimateSPLASH ONE spiked NIST SRM-1950 plasma was used for retention prediction with the MS^E method.

Selection of Candidate Ions for Spectral Library generation

Mass spectral peaks were extracted from a RT range of 2.9 to 3.3 minutes from the HILIC experiment. This chromatographic region contains the PC 18:1(d7)/15:0 standard and several other common PC candidate masses (**Figure 1**). An MS/MS-candidate list was generated using the criteria that: the nominal *m/z* of the ^{12}C isotope was even; the peak height was over $1x10^3$ counts and that the Kendrick mass defect fell between .4 and .8 Da. Putative PC were identified in descending precursor ion intensity order based on intensities in the NIST SRM-1950 sample. The criteria set for identification are listed in **Table 3**. Other GPL species were selected for MS/MS analysis based on the NIST community consensus paper (20). LPC species were identified based on exact mass and the neutral loss of the acetate ion in the elevated collision energy channel of the MS^E data.

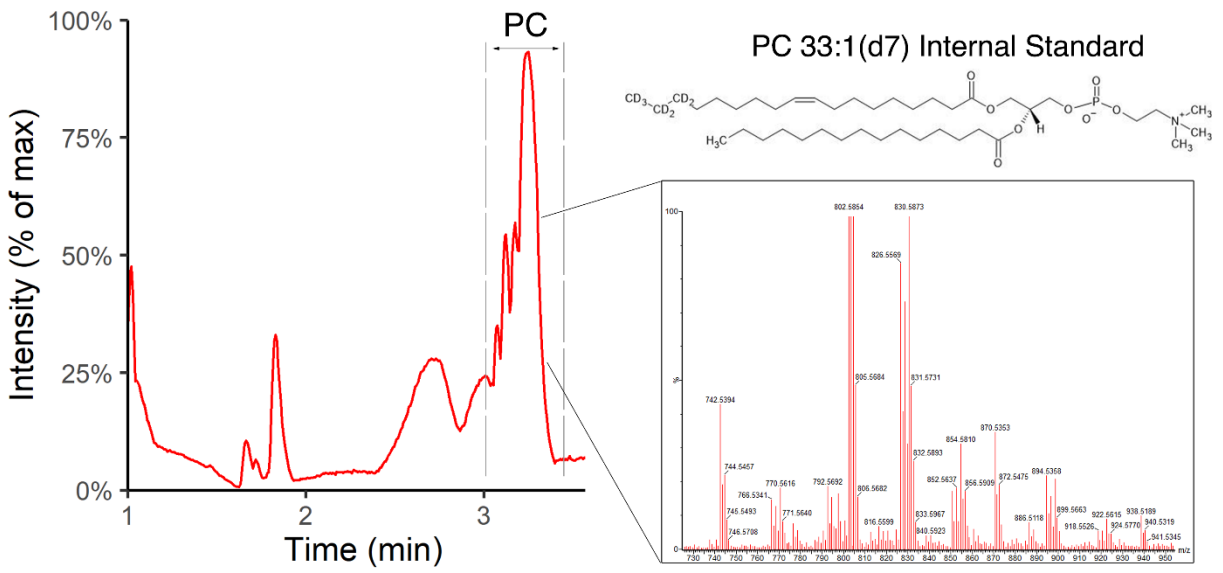

Figure 1: Typical mass chromatogram of NIST SRM-1950 plasma obtained using HILIC chromatography with the region containing the PC 33:1 (d7) internal standard highlighted. *M/z* values found in the highlighted region were used as precursor masses for MS/MS in subsequent reverse-phase chromatography experiments

Raw Data Preprocessing

An overview of data processing methods is given in **Figure 2**. Raw feature lists were extracted using MZMine2 version 2.52 (26). A mass list was generated from spectra using the selection criteria that peaks had to have peak heights greater than 500 counts and fall in the 100 to 1000 *m/z* range. Chromatograms were built from mass lists using the ADAP chromatogram builder module, where a minimum peak height of 1000 counts was required for detection, and a minimum of 5 points with height greater than 500 counts were required for inclusion. An *m/z* tolerance of ±0.015 Da was set for inclusion of points to the same peak. Extracted peaks were smoothed with an 11-point smooth and deconvoluted with the wavelets (ADAP) chromatogram deconvolution module (27). A signal-to-noise (S/N) threshold of 10 was set with intensity window S/N estimation. A minimum feature height of 1000 counts was selected with a coefficient/area threshold minimum of 50. Peak duration range was set between 0.05 min and 0.50 min. RT wavelet range was set from 0.00 - 0.15.

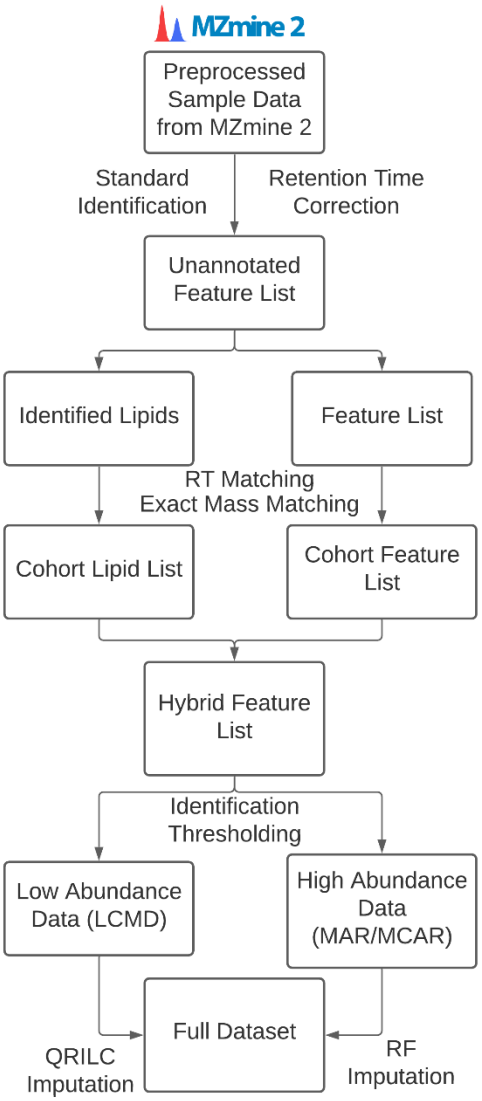

Figure 2: Overview of Data Processing Methodology. Chromatograms extracted with MZmine2 are retention corrected based on exact mass and RT of deuterated internal standard mixture. Lipids are identified based on mass charge and RT. Unidentified features are aggregated based on exact mass and RT. Peak areas from both datasets are normalized to the peak area of the PC 33:1 (d7) internal standard, where missing peaks in data are imputed with either QRILC or RF imputation, depending on their peak area. Lipid identifications are converted back to peak area and normalized to the intensity of their respective subclass internal standards.

RT Calibration & Lipid Identification Logic for Clinical Samples

Retention calibration was performed individually on the two linear sections of the solvent gradient corresponding to the elution of lyso- and diacyl lipid species respectively. The RTs of the lysolipid-region of chromatograms were calibrated with a flat calibration using stable isotope labelled LPC and LPE internal standards (0.5 min - 2.0 min). The diacyl-region RTs were similarly calibrated using labelled PC, PG, PI, PE, and PS internal standards (2.1 min - 12.1 min). Library matching criteria for RT correction standards were: m/z within ±10 mDa of theoretical value and chromatographic RT within 7.5% of the library value.

RT correction factor was calculated as the relative change in RT for the internal standards between all standard peaks. A linear regression of time *vs* change in RT was calculated to identify a drift coefficient, which was used to standardize RTs to those in the library in the diacyl-region of the gradient.

This correction was applied on a per-sample basis to account for any potential inter- and intra-batch RT variation. Potential sources of variability include: chromatographic column equilibration, variable rates of solvent evaporation from mobile phase reservoirs and measurement inaccuracies in the preparation of new batches of mobile phases.

There are two potential situations that can occur when identifying species with the same sum-composition. In the case of lipids that are baseline-separated chromatographically in the acquired MS/MS data peaks, were annotated as distinct measurements, *e.g.* PC 36:4 can be split into PC 18:2_18:2 and PC 18:1_18:3, with RTs of 5.00 min and 5.15 min respectively. In the case where multiple lipids co-elute within 3 s of other species with the same sum-composition, an aggregate measurement was recorded and assigned only to the total carbon and double bond level. For example, PC 32:0 with a single chromatographic peak at 7.12 min cannot be further resolved with MS-only data, although from MS/MS data is known to be a combination of PC 16:0_16:0 or PC 18:0_14:0. All lipids were normalized to the peak area of the internal standard in the same subclass. From RT-aligned data, lipids were identified based on the RT and identification criteria specified in **Table 3**. Mass tolerance was set to ±15 mDa, and corrected RT tolerance was set to ±3 s. Lipid identifications were binned to the nearest .7 min to simplify complex regions of the diacyl chromatography and aggregate *sn-1* and *sn-2* lyso-lipids.

Headgroup	Adduct	Candidate Masses for MS/MS	Characteristic Fragment Ions
PC	[M+HCOO]-	105	[FA1-H]-, [FA2-H]-, [M-CH3]-
PI	[M-H]-	12	[FA1-H]-, [FA2-H]-, [IP - H2O - H]-
PE	[M-H]-	18	[FA1-H]-, [FA2-H]-
PG	[M-H]-	3	[FA1-H]-, [FA2-H]-
PS	[M-H]-	1	[FA1-H]-, [FA2-H]-

Table 3: Identification characteristics for glycerophospholipid species in negative ion mode. FA1 and FA2 are fatty acids. IP is inositol mono-phosphate.

Untargeted Data Processing

Detected peaks from samples were normalized to the peak area of the PC 33:1 (d7) internal standard. Normalized peak areas were sorted in descending order and iteratively identified with ± 0.015 m/z and ± 0.05 min tolerances.

Missing Value Handling

To account for missing values in data, an approach was adopted from *Wei et al*, which uses a hybrid of quantile regression imputation of left-censored data (QRILC) and random forest imputation and to handle values which were below LOD and missing at random, respectively (28).

After compound identification, raw peak areas of identified lipids were normalized to the PC 33:1(d7) internal standard, to allow comparison with the untargeted data processing. The complete dataset was combined, and missing values were handled based on their minimum recorded intensity (see **Figure 2**). For features that had a minimum intensity below 5% of the peak area of the internal standard and were detected in more than 70% of samples, missing values were assumed to be left-censored missing data (*i.e.* below LOD), where missing data points were imputed with QRILC, using the impute.QRILC function from the imputeLCMD R package (29).

For features detected in more than 70% of samples and minimum intensity higher than 5% of the internal standard, missing values were attributed as missing at random or missing completely at random, where random forest imputation was used to calculate missing values using the MissForest package in R (30) Fully processed data was mean-centered and scaled using the scale function in base R.

Data Analysis and Visualization

PCA plots were generated using the base prcomp function in R and visualized using the ggplot2 package (31).

Results

Lipid Identifications

A total of 74 PC species were identified using MS/MS from 105 candidate masses (**Table 4**). Of these, 35 were resolvable to the acyl-chain composition level (*i.e.*, no evidence of other fatty acids with the same elution profile). An additional 6 structural isomers were identified that separated chromatographically. For non-PC subclasses, 14 PE and 5 PI species were observed, along with a total of 22 LPC, 14 LPE, and 5 LPG species.

Headgroup	IC-RTC Formula	Observed (MSMS)	Predicted	Total Lipids	IC-RTC difference from observed value (Mean (min) ± SD)	Adduct
LPC	-	22	0	22	-	[M+HCOO]-
LPE	x * 1.0794 – 0.0337	14	8	22	0.01 ± 0.01	[M-H]-
LPG	x * 0.9796 + 0.0119	5	17	22	0.00 ± 0.02	[M-H]-
PC	-	74	9	83	-	[M+HCOO]-
PE	x * 1.0804 - 0.0175	14	69	83	0.02 ± 0.03	[M-H]-
PG	x * 0.7652 + 0.3578	0	83	83	-	[M-H]-
PI	x * 0.7071 + 0.375	5	78	83	0.01 ± 0.03	[M-H]-
PS	x * 0.7466 + 0.282	0	83	83	-	[M-H]-
Total	-	134	347	481	-	-

Table 4: Identified phospholipids in NIST SRM 1950 blood plasma. Predicted PC species were calculated based on their fit to the ECN model. RTs for other species were predicted based on the PC and LPC identifications using IC-RTC.

Linear Elution of PC species in RP-LC-MS

The elution of PC species was observed to be predictable with a second order polynomial trendline, as previously reported (17,18). Here, we observed an increase in RT as the carbon number increased, while keeping the number of double bonds consistent **(Figure 3)**. The inverse relationship was also true when plotting RTs against double bonds, while keeping the carbon number consistent. Low abundance lipids

were resolved by utilizing this relationship to provide narrowed RT windows for identification **(Figure 3)**. Furthermore, we also analyzed three PC standards that only differed in their double bond position and/or stereochemistry **(Figure 4)**. We found that PCs containing the 18:1(9Z) were baseline separated from the 18:1(6Z) and 18:1(9E) species, which co-elute in reverse-phase LC-MS/MS.

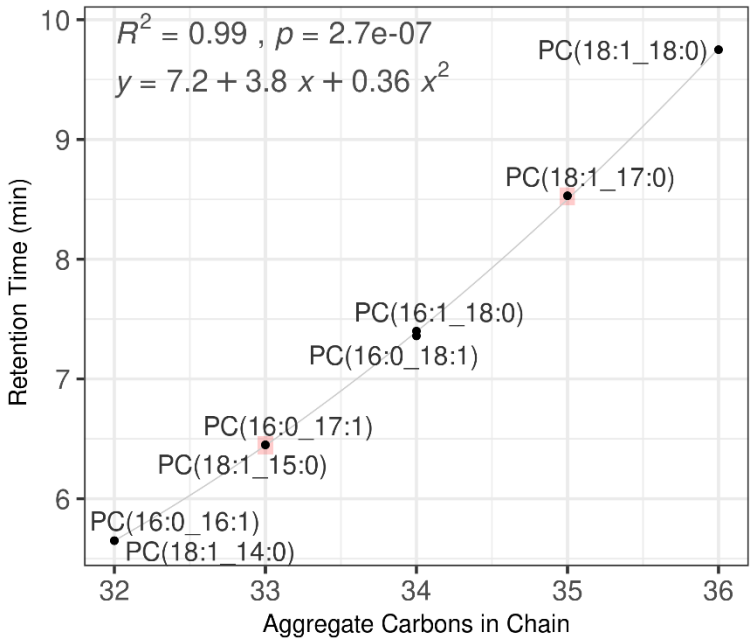

Figure 3: Retention behavior of monounsaturated PC species. Graph demonstrating the predictable RTs of monounsaturated species in RP-LC-MS from NIST SRM-1950 plasma standard. All species were verified by MS/MS. Red boxes highlight the theoretical RTs of PC 33:1 and 35:1. Fitted to a second-order polynomial trendline.

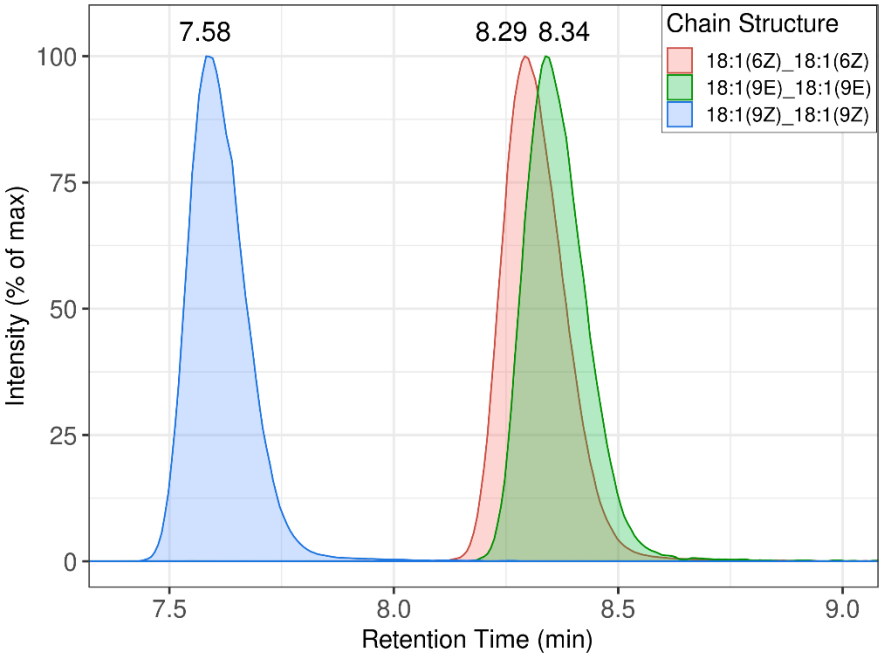

Figure 4: Elution of Stereoisomers in Reverse-Phase LCMS. Elution of PC stereoisomers in RP-LC-MS from 3 sequential injections of PC (18:1_18:1) standards with known stereochemistry. PC 18:1(9Z) elutes

with complete baseline separation from 9E and 6Z counterparts. The 9E and 9Z species coelute and are unresolvable using LC-MS/MS.

Prediction of Non-PC GPLs in Reverse-Phase Chromatography with IC-RTC

The RT of different GPL subclasses relative to PCs with the same acyl chain composition was plotted to study the effect of differing headgroups on RT. The resulting trendline was linear with an R-squared value of 1.0 for all observed GPL subclasses (**Figure 5**). This relationship was used to predict RTs of GPL subclasses from PC identifications (**Table 4).** To predict the RT of a non-PC GPL, the equations listed in **Table 4** were used with the RT of the PC with the same acyl chain composition as the lipid of interest *e.g* to calculate the RT of PE 16:0_20:4, the known RT of PC 16:0_20:4 (5.71 min) was substituted into the equation $RT_{PE} = RT_{PC} * 1.0804 - 0.0175$, giving a predicted RT of 6.15 min. Characteristics of species predicted solely from the RTs of other species are listed in **Table 5**.

Headgroup	Mean RT Difference (min ± SD)	Lipids Identified with IC-RTC and ECN RT prediction	Mean Normalised Peak Area (Mean% ± SD of IS Area)	Min (% of IS Area)	Max (% of IS Area)
PC	0.03 ± 0.02	4	36.8% ± 35.5%	1.3%	120.0%
PE	0.02 ± 0.02	24	6.4% ± 8.3%	0.3%	80.2%
PG	0.02 ± 0.01	4	0.6% ± 0.1%	0.2%	0.9%
PI	0.02 ± 0.01	15	7.3% ± 1.4%	0.1%	110.0%
PS	0.03 ± 0.01	10	1.8% ± 1.9 %	0.4%	9.1%

Table 5: Difference between IC-RTC and observed RT of ECN matched lipids in clinical samples. A total of 1,513 features were ECN and exact-mass matched in clinical samples. Only species which were quantifiable to their internal standard and predicted using IC-RTC were included.

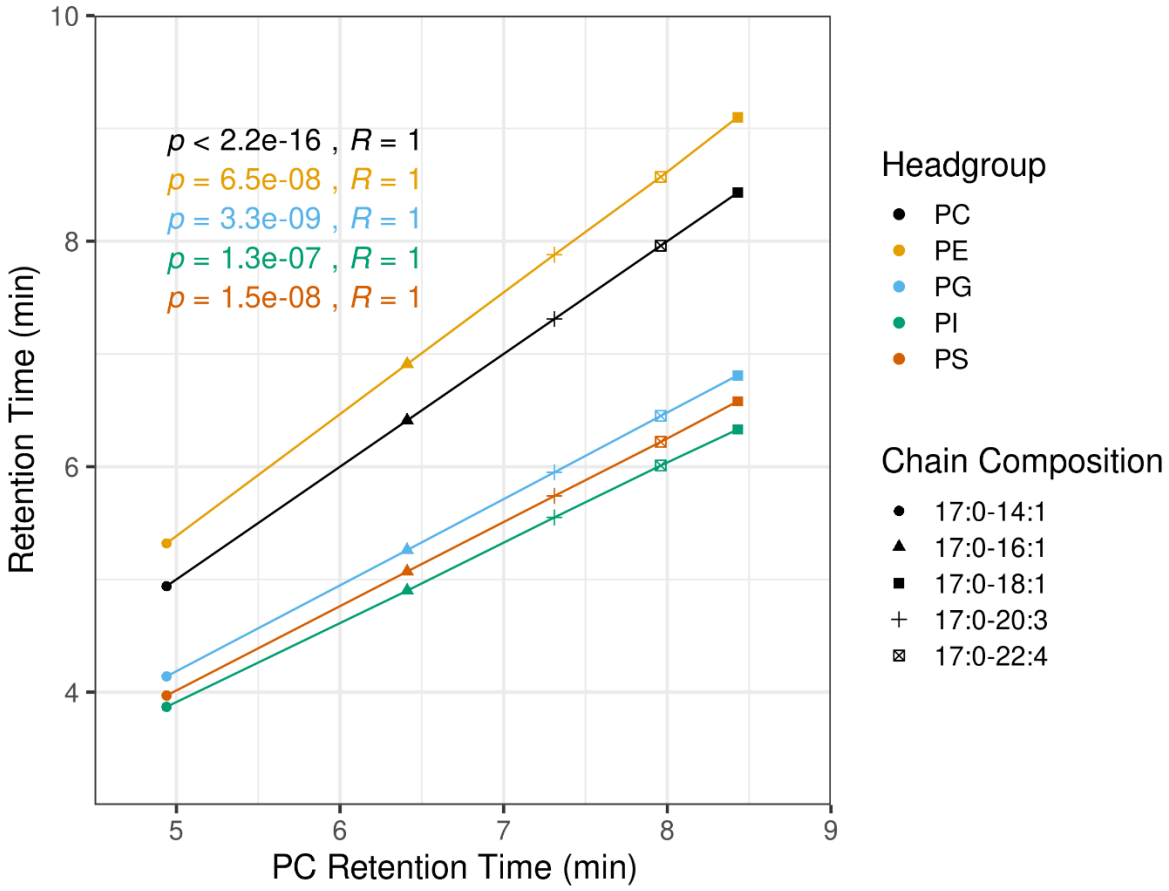

Figure 5: Relative RT indices of differing GPL subclasses. Distribution of varying chain length GPLs compared across multiple headgroup subclasses. All results are plotted relative to PC RT (x axis). Data acquired in negative ion mode with UltimateSPLASH One standards spiked into NIST SRM 1950 plasma.

Clinical Sample Results

A sum-total of 98,821 features were detected in 37 clinical samples and a NIST QC, a mean of 2,352 features were detected per sample. A total of 129 lipid species were identified, where 77 were present in >70% of samples (details in **Supplementary Table 1**). A total of 1,198 features were detected in >70% of samples. To highlight the heterogeneity of the lipidome, PCA was used to visualize identified lipid data (**Figure 6, A)** and unannotated feature data (**Figure 6, B)**. A total of 57 lipids unique to clinical samples were detected based on ECN and interclass retention prediction, primarily belonging to PI and PE subclasses **(Table 5)**. When stratifying samples between clinical diagnoses of recent myocardial infarction or stable angina, six unidentified compounds were significantly different between groups ($p < 0.05$, unpaired t-test); however, these features were not assessed further due to low abundance and high variation.

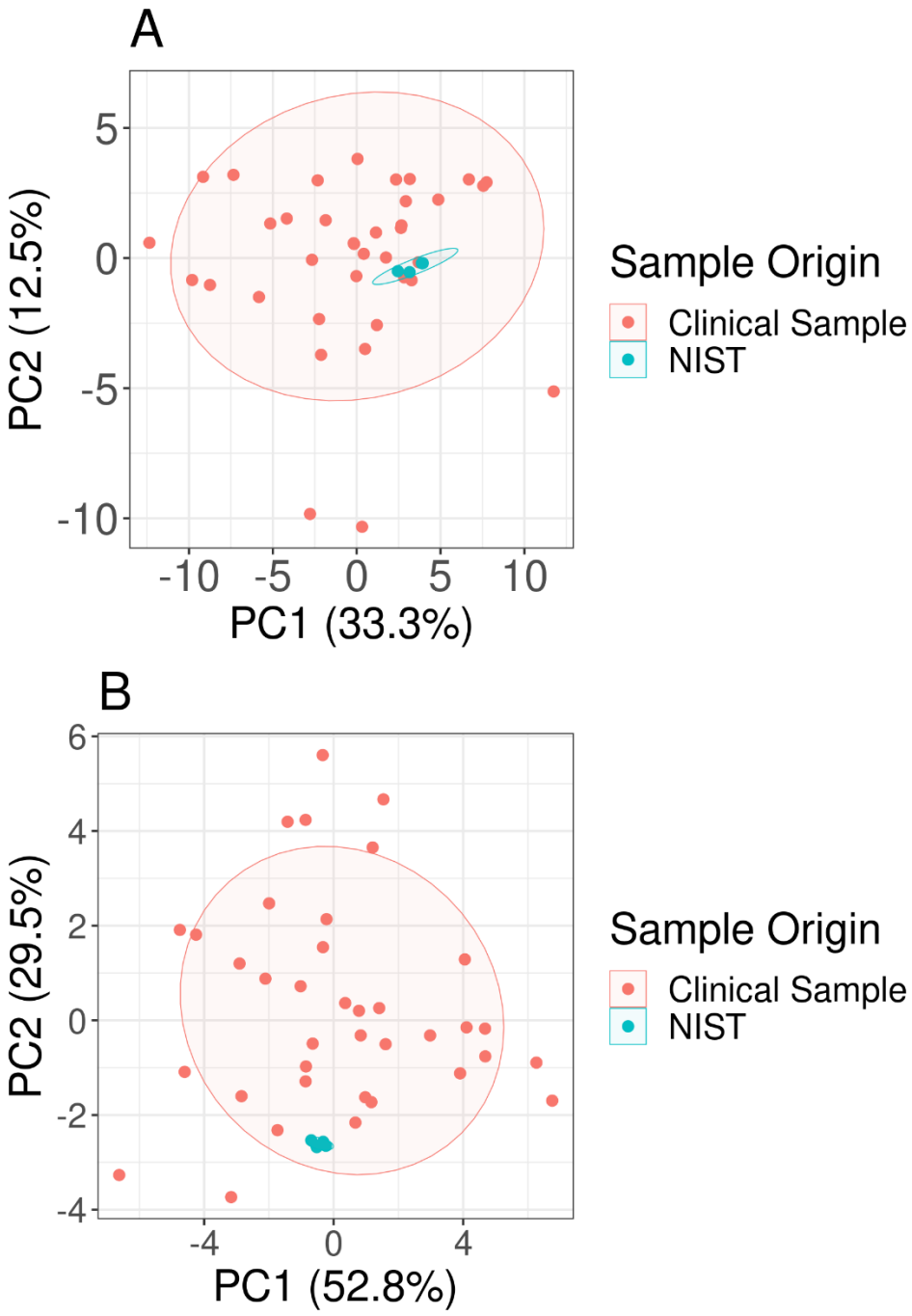

Figure 6: Dimensionality reduction of targeted lipid data (A) and unidentified feature data (B) for features with greater than 70% identification rate between NIST SRM-1950 plasma standard and clinical plasma samples. Ellipses indicate 95% confidence intervals for each group. NIST SRM-1950 plasma was injected from the same well after every 10 clinical samples.

Discussion

The data shown here illustrates the utility of RT-based characterization of lipid species from a matrix containing a large amount of unknown lipid species. The prediction of RT based on lipid physicochemical properties has been previously described in an intra-subclass context but has not yet been demonstrated on an inter-subclass basis to our knowledge. Our small clinical dataset also highlights the heterogeneity of the lipidome in human patient samples and the necessity of additional identification mechanisms besides exact-mass matching in clinical workflows, where IC-RTC and ECN matching reduces false-positive annotations.

RT Prediction of Lipid Species with ECN

The ECN model has been previously described as an additional mechanism to assess the validity of results in reverse-phase chromatography (22–24). By utilizing the relationship between RT and carbon chain length, it was possible to predict the RT of less abundant species to provide additional identifications (**Figure 7**).

A source of RT variability, potentially impacting RT prediction, is the double bond positioning and stereochemistry of unsaturated fatty acids. This variability is illustrated in **Figure 4,** where the 6Z, 9Z and 9E variants of PC (18:1/18:1) elute with slightly different measured peak tops. This effect has been reported in the context of PC, PG and TAG species(22, 32). In mammalian systems, there is limited diversity in abundant unsaturated fatty acids, these are commonly 16:1, 18:1, 18:2, with either 7Z/E or 9Z/E double bond positioning and stereochemistry (33). This diversity may be reflected in **Figure 7**, where multiple peaks are observed for lipids with the same MS/MS fragmentation *e.g.* 18:1_20:4 and 18:1_22:6, although this cannot be conclusively proven with the equipment used. The utilization of post-ionisation structural elucidation techniques such as ozonolysis may be able to resolve this further (34).

By measuring only one fatty acid and mapping the effect of differing combinations on the other acyl chain **(Figure 7 A)**, the accuracy of the ECN RT prediction is significantly improved **(Figure 7 B)**. Provided that there is enough biological diversity in samples to detect all combinations of species, a combination of aggregated and single fatty-acid prediction is useful to improve lipidome coverage. RT error *i.e.* the difference between predicted and observed measurements, is described in **Supplementary Table 1**.

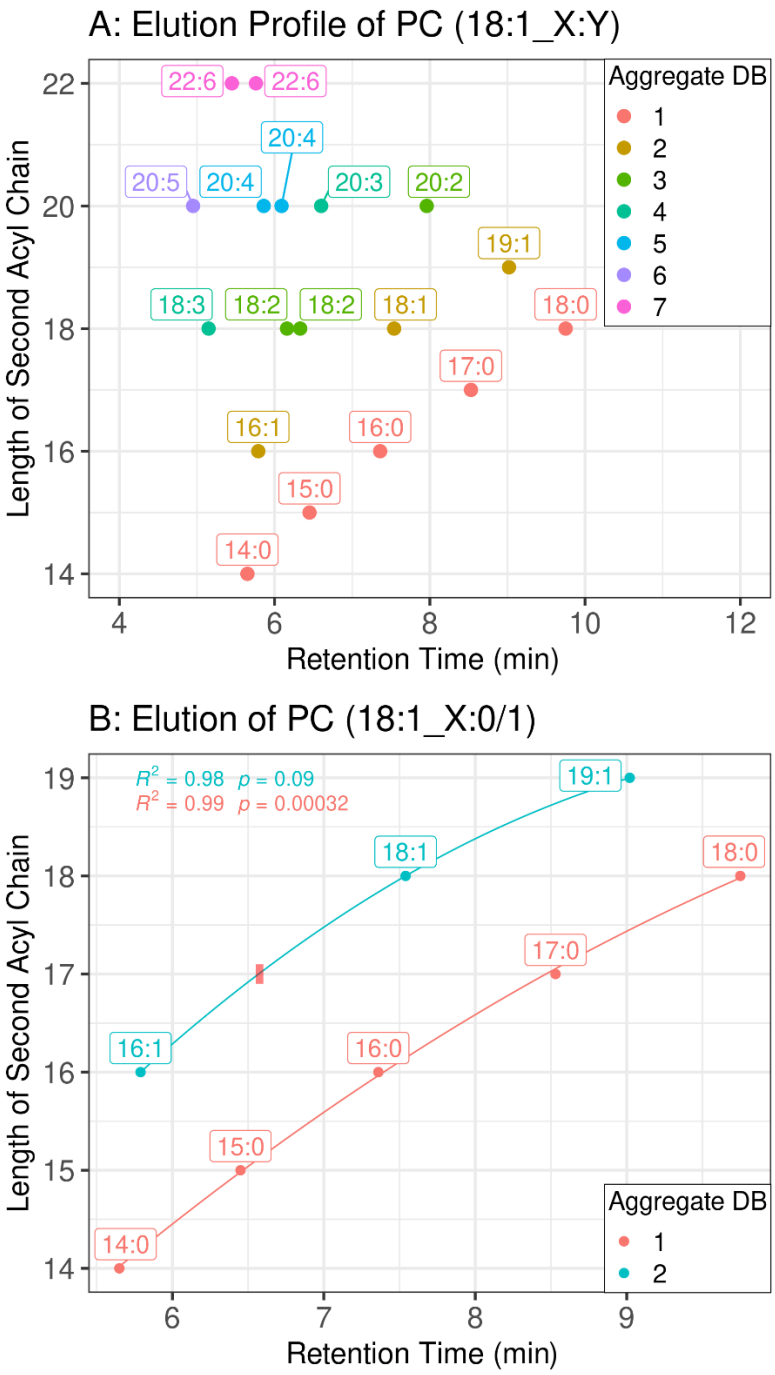

Figure 7: Elution profile of PC 18:1_X with different fatty acid conjugations. A total of 18 fatty acid combinations containing the 18:1 fatty acid were identified with MS/MS, of which 3 species were chromatographically resolvable isomers with the same fatty acid combination (A). It is possible to predict the RT of the PC 18:1_17:1 by studying the elution of other fatty acid species, illustrated by the annotated red box (B).

Inter-Class Retention Time Conversion of GPLs

To our knowledge, the mechanism which allows the conversion of RT between lipids of different subclasses has not been described previously. By utilizing the conversion method described here, it is possible to leverage the high abundance of PC species in blood plasma for the discovery of less abundant lipids from other GPL subclasses. While non-PC lipids are generally much lower in abundance, it is valuable to improve lipidome coverage. In all assessed lipids, the mean error in RTs obtained through IC-RTC was less than 1.5 s for compounds manually verified with MS/MS, and between 0 s and 5 s for compounds processed with the automated retention correction. As shown in **Table 5**, the abundance of predicted non-PC GPLs is considerably lower than that of other subclasses, as previously described in the context of the NIST SRM-1950 blood plasma standard (20).

An additional strength of the method described here is that it can assist in deconvolution of species from MS1 data; for instance, the neutral loss of the acetate ion of PC produces species which are isobaric with PE species with two fewer carbons in the aggregate acyl chain. This fragmentation also occurs in-source, which would complicate identification of compounds from MS-Only data if IC-RTC was not used. However, by using IC-RTC as an isolating mechanism for isobaric peaks, overlaps in RT and mass are restricted to lipid species with the same aggregate carbon chain composition in the diacyl region of chromatography (**Figure 8**).

For broader lipidome analysis in non-plasma research contexts, IC-RTC streamlines identification, as theoretical RTs for lipid species have already been recorded for an additional 347 GPL species could not be routinely detected in blood plasma here.

Despite the high accuracy of the IC-RTC, it is crucial that significant findings involving the detection of predicted species are assessed with MS/MS to confirm headgroup fragmentation and acyl chain composition before any assumptions about the underlying biology are made, particularly in cohorts with low statistical power.

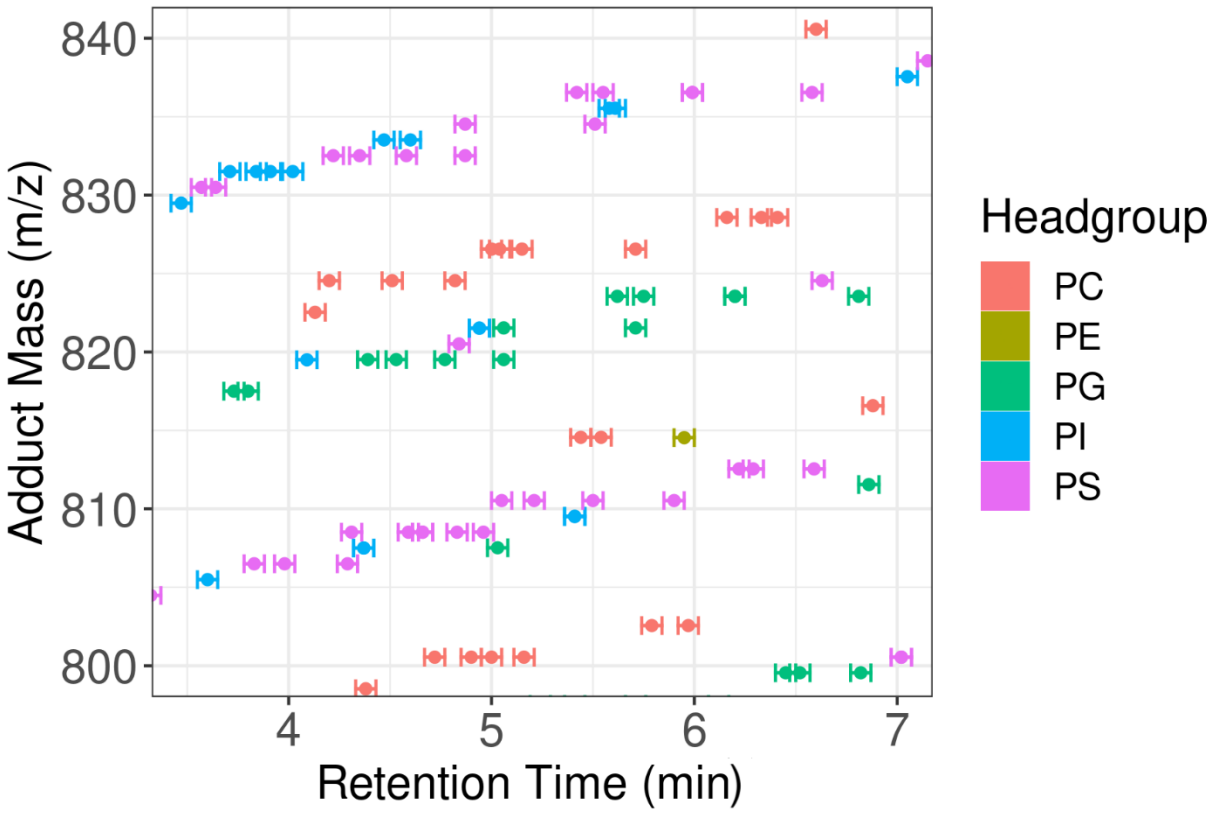

Figure 8: Visual representation of theoretical GPL elution in diacyl range of chromatography. Points correspond to the measured peak top of the resolved lipid species. Horizontal bars indicate RT range used for identification of each lipid. Mass tolerance for all peaks was set to 15 ± mDa. All overlaps of mass and RT were localized to species with the same aggregate carbon-chain length / double bond count. No overlaps of mass and RT occurred between lipids of different subclasses.

Clinical Data

Given the heterogeneity of human blood plasma, the primary use of clinical samples was to detect lipids not reported in the NIST 1950 standard using the ECN and test IC-RTC on a pilot cohort for future studies. We were able to detect a total of 57 candidate precursors that follow the two methods with mass error below 15 mDa that were not identified previously with MS/MS, primarily belonging to PI and PE subclasses **(Table 5)**. The consensus measurements of other GPLs are several times lower than their PC counterparts, therefore their rare detection in clinical samples is not surprising but allows for extra coverage of unexpected lipid species from ToF data (20).

The sample preparation protocol took approximately one hour to complete for 37 samples, because of its simplicity it is highly reproducible from sample to sample. The maximum sample throughput using this method is 40 samples per hour, making data acquisition time the bottleneck at 22.5 min per sample. In positive ion mode the triacylglycerol, sphingomyelin and ceramide species elute in the 12.1 min to 16.0 min range of the reverse-phase gradient (data not shown); however, this step could be omitted if GPLs are the only analytes of interest. Using the sample preparation method all internal standards were observed with 100% success and a 0% false positive rate **(Supplementary Figure 1)**. Experimental reproducibility was high, as illustrated by the low scatter in the dimensionality reduction of targeted and untargeted data of the NIST QC sample **(Figure 6)**.

Eight significantly altered compounds ($p < 0.05$, students t-test) were detected in the unannotated data when stratifying between by clinical diagnosis of myocardial infarction vs stable angina. These results were excluded from any subsequent analysis due to having maximum abundance below 1% of the PC internal standard in all cases.

When considering the normalized peak areas of all identified lipids, a significant difference in the distribution of lipids containing 5, 6 and 7 double bonds was noted when comparing clinical samples *vs* technical replicates of the NIST SRM-1950 plasma **(Supplementary Figure 2)**. Due to the high innate biological variance and expected differences in the targeted patient sampling of this study (patients undergoing cardiothoracic surgery for vascular complications) and the participants who donated plasma for the NIST standard, differences were expected when looking at simplified measures of the lipidome.

Utility of ECN and IC-RTC for Lipidomics Method Development

ECN provides an additional degree of structural identification confidence that cannot otherwise be attained without fragment ion information. Furthermore, with the exceptionally high accuracy of IC-RTC, maintaining theoretical RTs of all lipid species is recommended to increase lipidome coverage. This

advantage extends particularly well to instruments utilizing MS-only ToF or similar technologies, where no additional fragmentation data is generated.

Conclusion

The use of ECN to confirm lipid RTs provides an additional degree of confidence to compound identification in untargeted lipidomics, particularly when dealing with complex matrices, such as human blood plasma.

Furthermore, the use of one lipid subclass (in this instance, PCs) as the basis for the prediction of RT for other subclasses with IC-RTC demonstrates a substantial advantage of reverse-phase chromatography. Nominal RTs for other GPL species can be mined from untargeted lipidomics data, with a mechanism more rigorous than exact mass alone.

The combination of ECN and IC-RTC accelerates compound library building, where several theoretical RTs are generated from only one observed measurement.

For use with a clinical cohort, this method was used to identify a substantial amount of previously undetected lipid species, present at low abundance. For large scale clinical datasets using untargeted methodology, the methods discussed here allow for the more rigorous identification of lipids from reverse-phase LCMS data.

References

1. Hong, H. 2015. *In* Lipids in Protein Misfolding Advances in Experimental Medicine and Biology (Gursky, O., ed.). pp. 1–31. , Springer International Publishing, Cham

2. Dowhan, W., M. Bogdanov, and E. Mileykovskaya. 2016. *In* Biochemistry of Lipids, Lipoproteins and Membranes (Sixth Edition) (Ridgway, N. D., and McLeod, R. S., eds.). pp. 1–40. , Elsevier, Boston.

3. Chiurchiù, V., A. Leuti, and M. Maccarrone. 2018. Bioactive Lipids and Chronic Inflammation: Managing the Fire Within. *Front. Immunol.* **9**: 38.

4. Goldstein, J. L., and M. S. Brown. 2015. A Century of Cholesterol and Coronaries: From Plaques to Genes to Statins. *Cell*. **161**: 161–172.

5. Duncan, M. S., R. S. Vasan, and V. Xanthakis. 2019. Trajectories of Blood Lipid Concentrations Over the Adult Life Course and Risk of Cardiovascular Disease and All-Cause Mortality: Observations From the Framingham Study Over 35 Years. *J. Am. Heart Assoc.* **8**

6. Ding, M., and K. M. Rexrode. 2020. A Review of Lipidomics of Cardiovascular Disease Highlights the Importance of Isolating Lipoproteins. *Metabolites*. **10**.

7. Chapman, M. J., A. Orsoni, R. Tan, N. A. Mellett, A. Nguyen, P. Robillard, P. Giral, P. Thérond, and P. J. Meikle. 2020. LDL subclass lipidomics in atherogenic dyslipidemia: effect of statin therapy on bioactive lipids and dense LDL. *J. Lipid Res.* **61**: 911–932.

8. Sigruener, A., M. E. Kleber, S. Heimerl, G. Liebisch, G. Schmitz, and W. Maerz. 2014. Glycerophospholipid and Sphingolipid Species and Mortality: The Ludwigshafen Risk and Cardiovascular Health (LURIC) Study. *PLoS ONE*. **9**.

9. Cheng, J. M., M. Suoniemi, I. Kardys, T. Vihervaara, S. P. M. de Boer, K. M. Akkerhuis, M. Sysi-Aho, K. Ekroos, H. M. Garcia-Garcia, R. M. Oemrawsingh, E. Regar, W. Koenig, P. W. Serruys, R.-J. van Geuns, E. Boersma, and R. Laaksonen. 2015. Plasma concentrations of molecular lipid species in relation to coronary plaque characteristics and cardiovascular outcome: Results of the ATHEROREMO-IVUS study. *Atherosclerosis*. **243**: 560–566.

10. Havulinna, A. S., M. Sysi-Aho, M. Hilvo, D. Kauhanen, R. Hurme, K. Ekroos, V. Salomaa, and R. Laaksonen. 2016. Circulating Ceramides Predict Cardiovascular Outcomes in the Population-Based FINRISK 2002 Cohort. *Arterioscler. Thromb. Vasc. Biol.* **36**: 2424–2430.

11. Anroedh, S., M. Hilvo, K. M. Akkerhuis, D. Kauhanen, K. Koistinen, R. Oemrawsingh, P. Serruys, R.-J. van Geuns, E. Boersma, R. Laaksonen, and I. Kardys. 2018. Plasma concentrations of molecular lipid species predict long-term clinical outcome in coronary artery disease patients. *J. Lipid Res.* **59**: 1729–1737.

12. Alshehry, Z. H., P. A. Mundra, C. K. Barlow, N. A. Mellett, G. Wong, M. J. McConville, J. Simes, A. M. Tonkin, D. R. Sullivan, E. H. Barnes, P. J. Nestel, B. A. Kingwell, M. Marre, B. Neal, N. R. Poulter, A. Rodgers, B. Williams, S. Zoungas, G. S. Hillis, J. Chalmers, M. Woodward, and P. J. Meikle. 2016. Plasma Lipidomic Profiles Improve on Traditional Risk Factors for the Prediction of Cardiovascular Events in Type 2 Diabetes Mellitus. *Circulation.* **134**: 1637–1650.

13. Mundra, P. A., C. K. Barlow, P. J. Nestel, E. H. Barnes, A. Kirby, P. Thompson, D. R. Sullivan, Z. H. Alshehry, N. A. Mellett, K. Huynh, K. S. Jayawardana, C. Giles, M. J. McConville, S. Zoungas, G. S. Hillis, J. Chalmers, M. Woodward, G. Wong, B. A. Kingwell, J. Simes, A. M. Tonkin, and P. J. Meikle. 2018. Large-scale plasma lipidomic profiling identifies lipids that predict cardiovascular events in secondary prevention. *JCI Insight*. **3**: e121326.

14. Tsugawa, H., T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. VanderGheynst, O. Fiehn, and M. Arita. 2015. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods*. **12**: 523–526.

15. Tsugawa, H., T. Bamba, M. Shinohara, S. Nishiumi, M. Yoshida, and E. Fukusaki. 2011. Practical non-targeted gas chromatography/mass spectrometry-based metabolomics platform for metabolic phenotype analysis. *J. Biosci. Bioeng.* **112**: 292–298.

16. Kind, T., G. Wohlgemuth, D. Y. Lee, Y. Lu, M. Palazoglu, S. Shahbaz, and O. Fiehn. 2009. FiehnLib – mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.* **81**: 10038–10048.

17. Aicheler, F., J. Li, M. Hoene, R. Lehmann, G. Xu, and O. Kohlbacher. 2015. Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches. *Anal. Chem.* **87**: 7698–7704.

18. Aalizadeh, R., N. A. Alygizakis, E. L. Schymanski, M. Krauss, T. Schulze, M. Ibáñez, A. D. McEachran, A. Chao, A. J. Williams, P. Gago-Ferrero, A. Covaci, C. Moschet, T. M. Young, J. Hollender, J. Slobodnik, and N. S. Thomaidis. 2021. Development and Application of Liquid Chromatographic Retention Time Indices in HRMS-Based Suspect and Nontarget Screening. *Anal. Chem.* **93**: 11601–11611.

19. Feng, C., L. Xue, D. Lu, Y. Jin, X. Qiu, F. J. Gonzalez, G. Wang, and Z. Zhou. 2021. Novel Strategy for Mining and Identification of Acylcarnitines Using Data-Independent-Acquisition-Based Retention Time Prediction Modeling and Pseudo-Characteristic Fragmentation Ion Matching. *J. Proteome Res.* **20**: 1602–1611.

20. Yang, Q., H. Ji, H. Lu, and Z. Zhang. 2021. Prediction of Liquid Chromatographic Retention Time with Graph Neural Networks to Assist in Small Molecule Identification. *Anal. Chem.* **93**: 2200–2206.

21. Haddad, P. R., M. Taraji, and R. Szücs. 2021. Prediction of Analyte Retention Time in Liquid Chromatography. *Anal. Chem.* **93**: 228–256.

22. Damen, C. W. N., G. Isaac, J. Langridge, T. Hankemeier, and R. J. Vreeken. 2014. Enhanced lipid isomer separation in human plasma using reversed-phase UPLC with ion-mobility/high-resolution MS detection. *J. Lipid Res.* **55**: 1772–1783.

23. Ovčačíková, M., M. Lísa, E. Cífková, and M. Holčapek. 2016. Retention behavior of lipids in reversed-phase ultrahigh-performance liquid chromatography–electrospray ionization mass spectrometry. *J. Chromatogr. A.* **1450**: 76–85.

24. Lísa, M., E. Cífková, and M. Holčapek. 2011. Lipidomic profiling of biological tissues using off-line two-dimensional high-performance liquid chromatography–mass spectrometry. *J. Chromatogr. A.* **1218**: 5146–5156.

25. Lipidomics needs more standardization. 2019. *Nat. Metab.* **1**: 745–747.

26. Pluskal, T., S. Castillo, A. Villar-Briones, and M. Orešič. 2010. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*. **11**: 395.

27. Smirnov, A., Y. Qiu, W. Jia, D. I. Walker, D. P. Jones, and X. Du. 2019. ADAP-GC 4.0: Application of Clustering-Assisted Multivariate Curve Resolution to Spectral Deconvolution of Gas Chromatography–Mass Spectrometry Metabolomics Data. *Anal. Chem.* **91**: 9069–9077.

28. Wei, R., J. Wang, M. Su, E. Jia, S. Chen, T. Chen, and Y. Ni. 2018. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci. Rep.* **8**: 663.

29. Lazar, C. 2015. imputeLCMD: A collection of methods for left-censored missing data imputation.

30. Stekhoven, D. J., and P. Bühlmann. 2012. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinforma. Oxf. Engl.* **28**: 112–118.

31. Wickham, H. 2016. ggplot2: elegant graphics for data analysis. Second edition. Springer, Switzerland.

32. Bird, S. S., V. R. Marur, I. G. Stavrovskaya, and B. S. Kristal. 2012. Separation of Cis–Trans Phospholipid Isomers Using Reversed Phase LC with High Resolution MS Detection. *Anal. Chem.* **84**: 5509–5517.

33. Zlatanos, S. N., K. Laskaridis, and A. Sagredos. 2008. Conjugated linoleic acid content of human plasma. *Lipids Health Dis.* **7**: 34.

34. Thomas, M. C., T. W. Mitchell, D. G. Harman, J. M. Deeley, J. R. Nealon, and S. J. Blanksby. 2008. Ozone-Induced Dissociation: Elucidation of Double Bond Position within Mass-Selected Lipid Ions. *Anal. Chem.* **80**: 303–311.