

**An unbiased lipid phenotyping approach to study the genetic determinants and mechanisms of coronary heart disease risk factors**

Eric L. Harshfield \*, Albert Koulman \*, Daniel Ziemek, Luke Marney, Eric Fauman, Dirk S. Paul, David Stacey, John Danesh, Danish Saleheen, Adam S. Butterworth, Angela M. Wood, Julian L. Griffin

\* Joint first authors

Correspondence: Eric Harshfield  
Department of Public Health and Primary Care  
University of Cambridge  
Strangeways Research Laboratory  
Cambridge CB1 8RN  
UK  
  
eh457@medschl.cam.ac.uk  
Tel: +44 1223 748 669  
Fax: +44 1223 748 658


Short title: Harshfield – Lipid phenotyping coronary heart disease

Word count: 295 (abstract), 4742 (main text), 9041 (total)


1 table and 4 figures


1 Appendix and 13 Supplementary Figures and Tables

## 1 ABSTRACT

2  
3 **Background** Studies of the association of lipids with CHD have predominantly assessed  
4 major circulating lipids  our understanding of the diversity of lipid species, underlying  
5 metabolic pathways, and their genetic determinants is limited. Direct infusion high-  
6 resolution mass spectrometry (DIHRMS) is a novel high-throughput approach to rapidly  
7 and accurately profile hundreds of lipids in human serum without prior chromatography,  
8 allowing rapid and detailed lipid phenotyping for large cohort studies and providing detailed  
9 information about the association of lipid subfractions with CHD risk factors.

10  
11 **Methods** Intact lipid profiling by DIHRMS was performed on 5,551 serum samples from  
12 non-myocardial infarction participants in the Pakistan Risk of Myocardial Infarction Study  
13 (PROMIS). A novel peak-picking algorithm was developed to identify and record signals at  
14 mass-to-charge ratios corresponding to 444 known lipids in positive and negative  
15 ionization mode. The differences between the lipid phenotypes were assessed for a  
16 common polymorphism in the lipoprotein lipase (*LPL*) gene.

17  
18 **Results** The DIHRMS method provided information from five lipid categories (fatty acyls,  
19 glycerolipids, glycerophospholipids, sphingolipids, and sterol lipids) on 444 lipids t  
20 covered by other high-throughput metabolite profiling methods, with an average CV of  
21 13.44%. This wide lipid coverage showed that individuals with genetically lower *LPL*  
22 activity have increased levels of circulating triglycerides—especially those containing  
23 mono-unsaturated fatty acids—and decreased levels of sphingomyelins and cholesterol  
24 esters. This is the first report showing a link between *LPL* activity and circulating  
25 sphingomyelin levels.

26  
27 **Conclusions** This newly developed open-profiling method is highly suitable to provide  
28 detailed lipid profiles in large-scale epidemiological studies, with a wider coverage of lipids  
29 than most other high-throughput  profiling methods. This platform can provide many novel

insights into the effect of physiology and diet on lipid metabolism, the genetic determinants of lipids, and detailed information about the relationship between lipid subfractions and CHD.

**Keywords** lipidomics, mass spectrometry, protocol, coronary heart disease

CONFIDENTIAL

## 1 INTRODUCTION

2 Lipids have many different functions, from membrane components to cell signaling, and  
3 are associated with a number of chronic diseases, including cardiovascular diseases  
4 (CVD).<sup>1-3</sup> Atherosclerosis is associated with lipid accumulation and aggregation,  
5 particularly of cholesterol and its derivatives, while excess-energy intake can lead to  
6 hyperlipidemia, a major risk factor for CVD. Many polymorphisms in genes involved in lipid  
7 metabolism and lipid circulation are associated with CVD risk.<sup>4</sup> Improving our  
8 understanding of the mechanisms that underlie these associations will facilitate the  
9 development of pharmacological interventions to reduce the risk and burden of CVD.

10  
11 The development of new analytical technologies for the analysis of lipids, particularly in  
12 mass spectrometry (MS) and data processing,<sup>5</sup> has led to the rise of lipidomics,<sup>6</sup> with the  
13 aim of capturing information on a wide range of lipid metabolites in a given biological  
14 sample.<sup>1</sup> In particular, lipidomic analysis of human blood has the potential to identify the  
15 role of specific lipids in diseases, including CVD.<sup>7</sup> There are important prerequisites for any  
16 method that uses profiles of serum lipids to study lipid metabolism. Since an open-profiling  
17 method does not target particular lipid classes or species within each class, it must have  
18 the ability to measure and discriminate a wide range of lipids with minimal bias across  
19 different lipid species and classes.<sup>8</sup>

20  
21 This study provides a detailed description of the method used for a large-scale lipidomics  
22 analysis in the Pakistan Risk of Myocardial Infarction Study (PROMIS),<sup>9</sup> a case-control  
23 study of myocardial infarction (MI) in Pakistan with lipidomics measurements for 5,500  
24 individuals, which is currently among the largest MS-based metabolomics studies. While  
25 the vast majority of metabolomics studies have focused on Western populations, our study  
26 is one of the first to profile lipid levels in a Pakistani population. Chronic diseases in  
27 Pakistan are responsible for 50% of the total disease burden.<sup>10</sup> Ischaemic heart disease  
28 (IHD) and stroke are two of the top ten causes of years of life lost (YLL) in Pakistan, and  
29 observed YLLs due to IHD are nearly twice as high as expected based on socio-

demographic index.<sup>11</sup> Pakistan is also among the top ten countries in the world with the highest number of people living with diabetes.<sup>12</sup>

Our newly developed method used direct nanospray-infusion coupled to high-resolution mass spectrometry (DIHRMS), which yielded the intensities of several thousand features. Although the platform itself is non-targeted and measures all signals within a specified mass-to-charge ratio ( $m/z$ ) window, we used a customized peak-picking approach, which is scalable to studies with large-scale epidemiological data, to obtain signals for a targeted set of lipids that can reasonably be expected from human serum.

To demonstrate the utility of the platform to inform lipid research and the genetic determinants of human metabolism, we examined levels of circulating lipids associated with variants in the well-studied enzyme lipoprotein lipase, encoded by the *LPL* gene. It has been shown that variants in *LPL* alter circulating levels of triglycerides, HDL cholesterol, and chylomicron remnants, and are associated with increased risk of CVD.<sup>13</sup> The rs328 variant in *LPL* introduces a premature stop codon, making a protein two amino acids shorter. This is a gain-of-function mutation as the truncated protein shows increased activity,<sup>14</sup> and the variant is associated with reduced triglyceride levels.<sup>15</sup> However, studies to date have only looked at *LPL* variants in relation to overall triglycerides, so this lipidomics platform provides much greater detail about the nature of the association in relation to specific triglyceride species. One of the lead SNPs at the *LPL* locus in our study, rs117199990, is in perfect LD with the gain-of-function variant rs328 ( $r^2 = 1.0$ ). This variant is associated with genome-wide significant decreases ( $P < 8.9 \times 10^{-10}$ ) in levels of nine different diglyceride and triglyceride species and increases in the levels of four other lipid species. Thus, using *LPL* as an example provided validity for the DIHRMS platform and for open-profiling lipidomics as a whole.

We also used principal component analysis (PCA) to determine the main differences in lipid profiles across the participants and investigated the association of these principal

- 1 components with several established coronary heart disease (CHD) risk factors and genetic
- 2 loci.
- 3
- 4

CONFIDENTIAL

## METHODS

*Study description.* PROMIS is a case-control study of first-ever acute MI in patients from nine centers in urban Pakistan.<sup>9</sup> Anthropometry measurements, including systolic and diastolic blood pressure (SBP and DBP, respectively), height, and weight, were measured using standardized procedures and equipment. Non-fasting blood samples were drawn from each participant, centrifuged within 45 minutes of venipuncture, frozen, and transported on dry ice to Cambridge, UK, from which standard biomarkers including fasting plasma glucose (FPG) and hemoglobin (HbA<sub>1c</sub>) were measured. Samples were stored at –80 °C until use.

*Lipidomics sample selection and batch design.* A total of 5,674 PROMIS participants were selected who had genetic data and complete information on age, sex, ethnicity, center, and date of survey completion. Only non-MI participants were selected to avoid possible confounding from the metabolic factors of patients who had recently experienced an MI. The samples were analyzed on 72 plates, each with a maximum of 80 samples per plate, according to a randomized block design that was obtained using the “blockTools” package<sup>16</sup> in R v3.1.2. The factors used were sex, age, ethnicity, center, and time in years since date of survey; the distance between blocks was minimized for all factors. A quality control (QC) sample was created by pooling 100 µL of serum from 200 randomly selected samples, which was mixed and aliquoted for use on each plate. A subset of the QC sample was diluted with phosphate-buffered saline (PBS) solution to two different concentrations, giving three different QC samples per plate (QC1 was undiluted, QC2 was 1:1 diluted, and QC3 was 1:3 diluted). For all samples, including QC samples, 15 µL was aliquoted into 1.2 mL Cryovial tubes.

*Lipidomics sample extraction and mass spectrometry.* We adapted a method for open-profiling of lipids by direct infusion high-resolution mass spectrometry (DIHRMS).<sup>17, 18</sup> The method used automated sample preparation with an eight-head liquid handler, compatible with a 96-well plate format, resulting in an organic phase upper layer that was used for

lipid profiling and an aqueous bottom layer containing salts, proteins, and polar metabolites. The method relied on chip-based nanospray where each sample had a dedicated tip and nozzle, which limited the chance of any carry-over of highly hydrophobic lipids and ensured reproducibility. Further details are provided in the **Supplementary Methods (Appendix 1)**.

*Lipidomics data processing and DIHRMS peak picking.* For the 259 samples that had to be repeated due to poor quality signal data, we retained only the result file with the latest timestamp and/or highest quality mass spectral data, which was determined from visual inspection of the raw spectra. Once a clean list of 96 raw files was obtained for each of the 72 plates, including blanks and QC samples, the files were decompressed and converted to mzXML format using the "msconvert" tool in ProteoWizard.<sup>19</sup> For each infusion an average spectrum was calculated from the user-defined time window (inset **Figure 1a**). The R package "xcms"<sup>5</sup> was used to average fifty spectra per mode using an  $m/z$  window of 185–1000 and a retention time window of 20–70 seconds for positive ionization mode and 95–145 seconds for negative ionization mode (**Figure 1b**).

A list of  $m/z$  identity pairs, based on expected and possible lipids in human serum, was used to extract small windows of data around the target  $m/z$  in the average spectrum (**Figure 1c**). The peak maximum was measured and the two closest points to the half-height on either side were found, giving four points. The points with which a horizontal line at half-height intersected a line connecting the two points on either side of the peak (one above the half-height and one below) was used for a peak-width calculation (distance of the line) and a more accurate  $m/z$  value for the peak maximum (midpoint of the line). For all  $m/z$  identity pairs, the maximum intensity was recorded as well as the deviation of the peaks' accurate  $m/z$  (see **Figure 1d**). A major advantage of this approach is that it could be performed for each sample independently and run in parallel. The final step was the combination of all the signals (**Figure 1e**) and deviations (**Figure 1f**) into their



1 respective files. The technical set-up yielded average deviations of less than 4 ppm for the  
2 detected lipid species.

3  
4 *Lipidomics data cleaning and quality control.* The peak-picking algorithm initially selected  
5 all lipids from a list containing 1,305 lipids in positive ionization mode and 3,772 lipids in  
6 negative ionization mode, corresponding to the expected major ions of all known lipids  
7 within the  $m/z$  range used.<sup>20</sup> Quality control was performed to remove lipid signals that  
8 were not reliably detectable or did not show a linear response. Lipid signals present in  
9 fewer than 80% of all QC samples or that had a poor correlation with concentration within  
10 the dilution range of QC samples (Pearson correlation  $r < 0.95$ ) were removed. The  
11 coefficient of variation (CV) for each lipid signal was then determined and all lipids with  
12 CVs of more than 25% were omitted (**eFigure 1**). This ultimately yielded 444 lipids, 207  
13 from positive ionization mode and 237 from negative mode (**Table 1**).

14  
15 For each sample, the sum of the signals of all lipids within each ionization mode that  
16 passed the QC steps were calculated. Participants were excluded from analysis in a  
17 particular ionization mode if the total signal for the lipids in that mode was less than  
18 5,000,000 (relative units), signifying poor infusion of the sample. Each lipid species was  
19 normalized by expressing the signal as a proportion of the total signal for each participant.  
20 Since the distributions of most of the lipid signals showed approximate log-normality,  
21 natural log-transformation was applied to each lipid signal. Lipid signals for individual  
22 participants were considered outliers and excluded if the normalized, log-transformed  
23 signal was more than 10 standard deviations (SD) from the mean for that lipid across all  
24 participants.

25  
26 *Genotyping and imputation.* DNA was extracted from leucocytes in Pakistan and genotyped  
27 at the Wellcome Trust Sanger Institute in Cambridge, UK, on either the Illumina 660-Quad  
28 genome-wide array or the Illumina HumanOmniExpress array. Details of the genotyping,  
29 imputation, and QC have been described previously.<sup>21</sup>

1  
2 *Statistical analyses.* Normalized relative intensities for each of the lipids from each  
3 participant were plotted across the  $m/z$  spectrum (**eFigure 2**). Pearson correlation  
4 coefficients were calculated to determine lipid cross-correlations within each ionization  
5 mode. Ward's minimum variance agglomerative clustering method,<sup>22</sup> which minimizes the  
6 sum of the squared distance of pairs of objects, was applied iteratively to determine the  
7 hierarchy of the entire set of lipids (**eFigure 3**).

8  
9 *Gaussian Graphical Modelling.* In order to better resolve lipid cross-correlations, we  
10 estimated a Gaussian Graphical Model (GGM) on the normalized relative intensities of the  
11 lipids. The GGM resulted in a set of edges in which each edge connected two detected  
12 lipids if their cross-correlation conditioned on all other lipids was significantly different  
13 from zero. Subjects with more than 10% missing lipids as well as lipids with more than  
14 20% missing subjects were removed from the analysis. The "genenet" R package was  
15 used to infer the GGM.<sup>23</sup> A similar approach for metabolomics data has been suggested  
16 previously.<sup>24</sup> To focus on strong effects we retained only edges in the model that met an  
17 FDR cutoff of 0.05 and had a partial correlation coefficient greater than 0.2.

18  
19 *Fatty acid chain enrichment analysis.* We manually annotated detected lipids with their  
20 constituent fatty acid chains. To test whether edges from the GGM were enriched for any  
21 combination of fatty acid chains, we permuted the annotation 1000 times, keeping the  
22 number of annotations per lipid and fatty acid chain constant using the R package  
23 "BiRewire".<sup>25</sup> For each combination of fatty acid chains, we counted the number of GGM  
24 edges connecting lipids with that specific combination using the true annotation as well as  
25 the permuted versions. Using these counts, we directly estimated  $P$ -values of enrichment  
26 and depletion, respectively (**Figure 2**).

27  
28 *Phenome scans.* A phenome scan was conducted to analyze the nature of the association  
29 of lipid metabolites with key variants in the *LPL* and *APOA5* regions. The association of

each lipid with a gain-of-function variant in the *LPL* region (rs328) and a common polymorphism in the *APOA5* region (rs662799) revealed that the magnitude and direction of the association of different lipids with these variants varied significantly. The most strongly associated lipids within each lipid class [TG(53:3) (*m/z* 888.8016) for triglycerides, FreeFA(24:0) (*m/z* 367.3582) for fatty acids, PC-O(39:3) or PC-P(39:2) (*m/z* 812.6532) for glycerophospholipids, SM(42:3) (*m/z* 811.6688) for sphingolipids, and CE(20:3) (*m/z* 692.6339) for sterol lipids] are shown in **Figure 4** and **eFigure 4** for their association with a wide range of major lipids and other circulating biomarkers, including total cholesterol, LDL-C, HDL-C, HbA<sub>1c</sub>, apolipoprotein B, and C-reactive protein. Additionally, the association of each triglyceride with circulating VLDL is shown in **eFigure**

## 5.

EH, AK, JG, JD, and DS designed the study; EH had full access to all data in the study and takes responsibility for its integrity and the data analysis. Analyses were performed using Stata v13.1<sup>26</sup> and R v3.2.0<sup>27</sup> except where noted otherwise. Two-sided *P*-values and 95% confidence intervals (CIs) are presented.

## RESULTS

Lipid profiles obtained using DIHRMS were available for 5,551 PROMIS participants following data processing, cleaning, and QC. Only 2% of samples did not pass the quality control checks. Demographic and clinical characteristics of these participants are shown in **eTable 1**. Although all participants were without MI at recruitment, 56% of the participants were overweight, 17% were obese, 18% had hypertension, and the proportion with diabetes was 38%. The subset of PROMIS controls selected for the lipidomics assay were comparable to all PROMIS controls; however, PROMIS participants were older on average, and a higher proportion consumed tobacco and were overweight, compared with the head of household in the general Pakistani population, based on the latest available data obtained from the Demographic and Health Survey for Pakistan (**eTable 1**).<sup>28</sup>


Our DIHRMS method for lipidomics covers a wide range of lipids, including fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, and sterol lipids (**Table 1**), and does not require a prior selection of specific lipids or lipid classes, in contrast to fragmentation-based approaches using tandem mass spectrometry. The high throughput of the method means that with an analysis time of just over two minutes per sample, it is possible to run a full plate within four hours. The automated sample preparation of one plate is possible in 1.5 hours, which makes this approach especially useful for large-scale lipid profiling.

The DIHRMS method included measurement of neutral lipids such as triglycerides and cholesterol esters, which are not covered by the commercial platforms that are currently most widely used in large-scale metabolite phenotyping for genome-wide association studies.<sup>29</sup> We analysed approximately 125 metabolic features that have not yet been assessed in any of the major genome-wide association studies of human metabolism. Additionally, we included lipids that contain odd-chain fatty acids, which are generally ignored in metabolic profiling methods, while we have shown that these are important to human metabolism.<sup>30</sup>

Following QC filtering, 207 lipids in positive ionization mode and 238 lipids in negative ionization mode remained for analysis, all with unique mass-charge ratios and identifiers. The coefficients of variation (CVs) for the lipids that were retained in each ionization modes are shown in **eFigure 1**. The precision was higher in positive mode (average CV 13.44%, median CV 11.61%) than in negative mode (average CV 23.45%, median CV 22.01%). However, the CVs demonstrated that simple normalization yielded reproducible data on a par with other high-throughput metabolic profiling methods.<sup>31, 32</sup> The normalized relative intensity recorded for each participant are shown for each lipid in **eFigure 2**.

A heat map of Pearson correlation coefficients with hierarchical clustering revealed a distinct clustering of lipids (**eFigure 3**). Although the lipids could be clustered broadly into three overall groups, they were correlated with other lipids both within and across lipid classes and subclasses. The GGM approach used partial correlations to determine if specific lipids were still strongly correlated after adjusting for all other lipids, and if lipid signals that were assigned to specific lipids were in fact dominated by isotopologues of other lipids or signals present as artifacts. From the 314 GGMs, there were ten correlations that were purely M+1 isotopes of other lipid signals and four correlations that were purely M+2 isotopes, based on very high correlations ( $r > 0.997$ ) and correct isotope ratios. There were 26 correlations where the M+1 isotope contributed considerably to the signal and four where the M+2 isotope contributed predominantly to the signal. However, those signals also showed contributions of different lipid signals, for which the correlations were not as high ( $r < 0.997$ ) or the isotope ratio was incorrect. There were 36 correlations where the signals came from the same lipid in both positive and negative ionization modes, and two sets of lipids for which the signals overlapped and the peak-picking algorithm was unable to distinguish the signals. The remaining 222 significant correlations confirmed biological explanations (**Figure 2**).

*Phenome scan of LPL region.* We used the detailed lipid profiles to study the effect of the gain-of-function mutation in the *LPL* gene (rs328) on lipid levels. All lipids that showed a

significant association with the polymorphism are shown in **Figure 3**, while results for the association of lipids with a common polymorphism in the *APOA5* region (rs662799) are shown in **eFigure 11**. lycerides, triglycerides, cholesterol esters, sphingomyelins, and phosphocholines were affected by the *LPL* polymorphism. Within the triglycerides, those containing mono-unsaturated fatty acids had the greatest magnitude of effect and significance. For the most significantly associated lipids within these classes we determined their cross-correlation with circulating biomarkers. Triglyceride TG(53:3) (**Figure 4**) showed a significant positive correlation with major circulating triglycerides, ApoC3 and ApoE, while there was a significant negative correlation with HDL and ApoA1. For sphingomyelins and cholesterol esters (**eFigures 4a & 4b**) the strongest inverse associations were with major circulating triglycerides, ApoC3 and ApoE. Sphingomyelins had the strongest correlation with HDL-C, while cholesterol esters had the strongest association with LDL-C.

*Principal component analysis of lipids.* We assessed the overall differences in lipid metabolism in the cohort using PCA. Scatter plots showing the matrix loadings of the first four principal components for each lipid, colored according to overall lipid class, are shown in **eFigure 6**. A subset of the PCA results for the triglycerides is shown in **eFigure 7**, and a subset of the PCA for the lipids that were significant associated with *LPL* is shown in **eFigure 8**.

As is often observed with microarray data, the first principal component (eigenvalue 135.7) was mainly driven by batch effects; in this case differences between the positive and negative ionization modes were the dominant trends in the first component. The dynamic range of the negative mode data was more limited than the positive mode data, and due to the lower ionization efficiency, the data were more prone to ion suppression. These differences between the ionization modes were amplified when the data were expressed relative to total signal intensity. We therefore excluded the first principal component from further data analysis.

The second component (eigenvalue 49.9) was dominated by triglycerides containing shorter and more saturated fatty acids, which had the highest positive loadings, versus fatty acids (e.g. free linoleic acid) and cholesterol esterified with polyunsaturated fatty acids [e.g. CE(18:2)], which had the strongest negative loadings (**eFigure 6a**). The third component (eigenvalue 29.2) differentiated saturated phosphatidylcholine [e.g. PC aa (32:0), PC aa (34:0), PC ae (32:0)] from triglycerides containing longer, unsaturated fatty acids [e.g. TG(54:5), TG(54:7), TG(56:7)] (**eFigure 6b**). The fourth component (eigenvalue 20.4) differentiated the odd-chain fatty acid containing sphingomyelins [e.g. SM(39:1), SM(41:1), SM(37:1)] with the highest positive loadings versus saturated free fatty acids and triglycerides [e.g. TG(52:2), TG(54:2)] with the strongest negative loadings (**eFigure 6b**).

The association between the second, third, and fourth principal components with several CHD risk factors (i.e. overweight, obesity, hypertension, and diabetes) are shown in **eFigure 9**. Individuals who were overweight or diabetic—i.e. with high levels of CHD risk factors—were more likely to have lipid profiles similar to those corresponding to the second principal component, whereas individuals who were not overweight or diabetic and did not have hypertension—i.e. at reduced risk of CHD—were more likely to have lipid profiles matching the third or fourth principal components. For example, a 1-SD increase in the loading scores of the lipids that made up the third principal component resulted in a 20% reduction in the risk of being overweight (OR = 0.80, 95% CI 0.76-0.84) and a 33% reduced risk of having diabetes according to levels of HbA<sub>1c</sub> (OR = 0.77, 95% CI 0.72-0.81), which is a reflection of long-term blood glucose levels rather than short-term fluctuations.

A GWAS of the principal components determined that there were 74 variants from two loci (*APOA5-APOC3* and *FADS1-2-3*) that reached genome-wide significance ( $P < 5 \times 10^{-8}$ ) for association with the second, third, and fourth principal components of the lipids. The

1 second principal component (eigenvalue 49.9) was only associated with a single variant  
2 from the *APOA5-APOC3* region (rs662799), the third principal component was associated  
3 with 22 variants from the *APOA5-APOC3* region, and the fourth principal component was  
4 associated with 74 variants from both loci. The association results for the most strongly  
5 associated SNP within each locus that was associated with each principal component are  
6 summarized in **eTable 2**, and the overall results for each principal component are  
7 displayed in the Manhattan plots and regional association plots in **eFigure 10**.

CONFIDENTIAL



## DISCUSSION

We used DIHRMS and developed a novel peak-picking algorithm that enabled us to measure and characterize 444 lipids in 5,551 participants from PROMIS. This robust approach to obtaining a detailed overview of lipid metabolism in individuals is a stepping-stone in the application of lipidomics to molecular epidemiology. The approach we developed has three practical advantages over most published approaches: (1) The method is extremely fast, so that with an analysis time of just over two minutes per sample it is possible to run several hundred samples per day; (2) as a consequence of the high throughput, the cost per sample is greatly reduced; (3) by virtue of the simplicity of the method it is also robust with low CVs achievable for many of the lipid species detected. This opens up the possibility of applying the method to much larger studies or exploring possible applications in routine assessment of patients and health screenings.

A phenome scan of the most common polymorphism in the *LPL* gene revealed that the difference in lipoprotein lipase activity leads to several specific changes in the lipid profile (**Figure 3**). The major differences are that decreased *LPL* activity leads to an increase in levels of specific triglycerides, mainly those containing mono-unsaturated fatty acids, which provides more detail than the previously reported substrate specificity.<sup>33</sup> These findings highlight the opportunities of this approach in larger studies, which can include sufficient coverage of rare genetic variants, to help understand how polymorphisms in *LPL* can lead to changes in substrate specificity for triglycerides. We also determined the correlation of each triglyceride with VLDL concentration (**eFigure 5**), which showed that the lipids affected by the polymorphism are not specific to the VLDL fraction, which corresponds with the fact that *LPL* also hydrolyses triglycerides from different lipoprotein fractions.

The other major differences were decreased concentrations of cholesterol esters and sphingomyelins. It is important to note that there were not differences in phosphocholine lipids or other lipid classes found to be associated with this *LPL* polymorphism. This shows

1 that the decreases of cholesterol esters and sphingomyelins were not just the result of an  
2 increase in triglycerides, because if this was an unspecific effect on all lipid classes,  
3 phosphocholines would also been affected as these are the major polar lipids in the  
4 lipoproteins.<sup>34</sup> Several studies have speculated that sphingomyelins and cholesterol esters  
5 can reduce *LPL* activity.<sup>35</sup> It is therefore possible that sphingomyelins and cholesterol  
6 esters are used as moderator of *LPL* activity and that the increase of *LPL* activity is  
7 compensated by increasing sphingomyelin levels in lipoprotein particles to retain lipase  
8 acitivity at an optimal level. We currently do not know what mechanism is responsible for  
9 the increased production of these lipid classes when *LPL* becomes more active. These  
10 observations are important to take into account when *LPL* agonists are studied as  
11 pharmacological treatments for dyslipidemia. Just targeting *LPL* activity alone might have  
12 a more limited physiological impact than interventions that also counteract the mediation  
13 of *LPL* activity by sphingomyelins and cholesterol esters.

14  
15 From the results of the PCA, the second principal component revealed a contrast between  
16 free fatty acid levels versus small saturated triglycerides. Several factors could have  
17 contributed to this differentiation. Volunteers were recruited at different hospitals and  
18 blood samples were taken directly after consent. This means that there was significant  
19 variation in the time since participants had eaten their last meal, which would have  
20 strongly affected both the free fatty acid and triglyceride pool. This could obscure the  
21 genetic associations, and indeed, in the GWAS there was only one SNP that showed a  
22 genome-wide significant association with the second principal component. Variants in the  
23 *APOA5-APOC3* region have been associated with type 2 diabetes,<sup>36</sup> fatty liver disease,<sup>37</sup>  
24 hypertriglyceridemia,<sup>38, 39</sup> and dyslipidaemia.<sup>40</sup> This gene region has also been associated  
25 with metabolites such as 1-linoleoylglycerol in previous metabolomics studies.<sup>41</sup> However,  
26 as only one variant was significantly associated with the second principal component  
27 compared to 74 variants that were associated with the third and fourth principal  
28 components, this suggests that the second principal component was largely driven by diet  
29 and eating patterns, rather than genetic differences. The lipid species that contributed

most to the second principal component are also affected by obesity and insulin secretion/sensitivity. At the same time, adiposity and insulin secretion will have an effect on free fatty acids levels as well. The second principal component also demonstrated a significant positive association with the relative likelihood for being overweight and having diabetes, according to levels of both HbA<sub>1c</sub> and FPG.

Although the third principal component was most closely characterized by unsaturated triglycerides, it did not show any significant associations with variants in the *FADS1-2-3* locus, and was only associated with variants in the *APOA5-APOC3* region. In contrast, variants in both the *FADS1-2-3* and *APOA5-APOC3* regions were significantly associated with the fourth principal component. The *APOA5-APOC3* locus has been previously highlighted in other lipidomics studies<sup>42</sup> and previous GWAS.<sup>43-45</sup> The loadings of the fourth principal component showed that the triglycerides containing linoleic acid (18:2), as well as linoleic acid as a free fatty acid, had negative loading scores, while sphingomyelins containing odd-chain fatty acids and desaturated phospholipids had positive loading scores. The association between SNPs within the *FADS1-2-3* region with sphingomyelins has been described previously<sup>46</sup>, although not explained, and has not previously been described for odd-chain fatty acid-containing sphingomyelins. The effect on the triglycerides also explained the association with SNPs in the *APOA5-APOC3* locus for the fourth principal component. Both the third and fourth principal components showed a negative association with the relative risk for being overweight and having diabetes, while only the fourth principal component also showed a negative association with the relative risk for hypertension. This last observation is striking as sphingomyelins have thus far been implicated in hypertension as precursors to ceramide production, but odd-chain fatty acid-containing sphingomyelins have mostly been unexplored. It is possible that the higher levels of these sphingomyelins [e.g. SM(39:1), SM(41:1) and SM(37:1)] compete with other sphingomyelins in ceramide biosynthesis and therefore reduce the production of ceramides, and in this way play a protective role in the development of hypertension.

1 Future work resulting from this lipidomics platform will focus on a univariate GWAS of each  
2 of the individual lipids and the subsequent identification of novel loci for CHD risk. These  
3 analyses will be based on the current dataset described here as well as on newer, even  
4 larger studies for which lipid profiling is currently underway, optimized for using 384-well  
5 plates. This lipidomics platform, then, is a useful tool that may prove valuable to identify  
6 lipids that could become new biomarkers used for clinical application in areas such as CVD  
7 screening,<sup>47</sup> risk prediction,<sup>48</sup> and drug development.<sup>49</sup>

8  
9 In conclusion, we show here that fast and reproducible lipid profiling is possible on very  
10 large datasets, revealing that lipid profiles are influenced by both lifestyle and genetic  
11 factors. The open-profiling approach revealed information on lipids that have not  
12 previously been implicated in relation to CHD risk and discovered a novel link between *LPL*  
13 activity and circulating sphingomyelin levels.


## 1 **ACKNOWLEDGEMENTS**

2 We would like to acknowledge the contributions of the following individuals: Asif Rasheed  
3 for preparing and shipping the samples, Michael Eiden for his assistance with the batch  
4 design, Philip Haycock and Nasir Sheikh for aliquoting and ordering the samples on the  
5 plates, and Lee Matthews for performing the lipid profiling.

## 7 **SOURCES OF FUNDING**

8 The Cardiovascular Epidemiology Unit is underpinned by programme grants from the UK  
9 Medical Research Council (G0800270), British Heart Foundation (SP/09/002), UK National  
10 Institute for Health Research Cambridge Biomedical Research Centre, European Research  
11 Council (268834), and European Commission Framework Programme 7 (HEALTH-F2-  
12 2012-279233). AK and JLG are funded by the Medical Research Council under the Lipid  
13 Dynamics and Regulation supplementary grant (MC\_PC\_13030) and Lipid Programming  
14 and Signalling programme grant (MC\_UP\_A090\_1006) and Cambridge Lipidomics  
15 Biomarker Research Initiative (G0800783).

## 17 **DISCLOSURES**

18 JD received research funding from the British Heart Foundation, the National Institute for  
19 Health Research Cambridge Comprehensive Biomedical Research Centre, the Bupa  
20 Foundation, diaDexus, the European Research Council, the European Union, the Evelyn  
21 Trust, the Fogarty International Centre, GlaxoSmithKline, Merck, the National Heart, Lung,  
22 and Blood Institute, the National Institute for Health Research, the National Institute of  
23 Neurological Disorders and Stroke, NHS Blood and Transplant, Novartis, Pfizer, the UK  
24 Medical Research Council, and the Wellcome Trust. All other authors declare no competing  
25 interests. 

## 27 **AUTHOR AFFILIATIONS**

28 Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care,  
29 University of Cambridge (E.L.H., D.S.P., D.St., J.D., A.S.B., A.M.W.); Medical Research

1 Council Elsie Widdowson Laboratory (A.K., J.L.G.); Core Metabolomics and Lipidomics  
2 Laboratory, National Institute for Health Research Cambridge Biomedical Research Centre  
3 (A.K.); Pfizer Worldwide Research & Development, Cambridge, Massachusetts, United  
4 States of America (D.Z., E.F.); College of Science and Engineering, Seattle University  
5 (L.M.); Department of Biostatistics and Epidemiology, University of Pennsylvania,  
6 Philadelphia, PA & Center for Non-Communicable Diseases, Karachi, Pakistan (D.Sa.);  
7 Department of Biochemistry and Cambridge Systems Biology Centre, University of  
8 Cambridge (J.L.G.).  
9

## REFERENCES

1. Griffin JL, Atherton H, Shockcor J and Atzori L. Metabolomics as a tool for cardiac research. *Nat Rev Cardiol.* 2011;8:630-643.
2. Shah SH, Kraus WE and Newgard CB. Metabolomic profiling for the identification of novel biomarkers and mechanisms related to common cardiovascular diseases: form and function. *Circulation.* 2012;126:1110-20.
3. Pechlaner R, Kiechl S and Mayr M. Potential and Caveats of Lipidomics for Cardiovascular Disease. *Circulation.* 2016.
4. Global Lipids Genetics Consortium, Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, Beckmann JS, Bragg-Gresham JL, Chang HY, Demirkan A, Den Hertog HM, Do R, Donnelly LA, Ehret GB, Esko T, Feitosa MF, Ferreira T, Fischer K, Fontanillas P, Fraser RM, Freitag DF, Gurdasani D, Heikkila K, Hypponen E, Isaacs A, Jackson AU, Johansson A, Johnson T, Kaakinen M, Kettunen J, Kleber ME, Li X, Luan J, Lyytikainen LP, Magnusson PK, Mangino M, Mihailov E, Montasser ME, Muller-Nurasyid M, Nolte IM, O'Connell JR, Palmer CD, Perola M, Petersen AK, Sanna S, Saxena R, Service SK, Shah S, Shungin D, Sidore C, Song C, Strawbridge RJ, Surakka I, Tanaka T, Teslovich TM, Thorleifsson G, Van den Herik EG, Voight BF, Volcik KA, Waite LL, Wong A, Wu Y, Zhang W, Absher D, Asiki G, Barroso I, Been LF, Bolton JL, Bonnycastle LL, Brambilla P, Burnett MS, Cesana G, Dimitriou M, Doney AS, Doring A, Elliott P, Epstein SE, Eyjolfsson GI, Gigante B, Goodarzi MO, Grallert H, Gravito ML, Groves CJ, Hallmans G, Hartikainen AL, Hayward C, Hernandez D, Hicks AA, Holm H, Hung YJ, Illig T, Jones MR, Kaleebu P, Kastelein JJ, Khaw KT, Kim E, Klopp N, Komulainen P, Kumari M, Langenberg C, Lehtimaki T, Lin SY, Lindstrom J, Loos RJ, Mach F, McArdle WL, Meisinger C, Mitchell BD, Muller G, Nagaraja R, Narisu N, Nieminen TV, Nsubuga RN, Olafsson I, Ong KK, Palotie A, Papamarkou T, Pomilla C, Pouta A, Rader DJ, Reilly MP, Ridker PM, Rivadeneira F, Rudan I, Ruukonen A, Samani N, Scharnagl H, Seeley J, Silander

K, Stancakova A, Stirrups K, Swift AJ, Tired L, Uitterlinden AG, van Pelt LJ, Vedantam S, Wainwright N, Wijmenga C, Wild SH, Willemsen G, Wilsngaard T, Wilson JF, Young EH, Zhao JH, Adair LS, Arveiler D, Assimes TL, Bandinelli S, Bennett F, Bochud M, Boehm BO, Boomsma DI, Borecki IB, Bornstein SR, Bovet P, Burnier M, Campbell H, Chakravarti A, Chambers JC, Chen YD, Collins FS, Cooper RS, Danesh J, Dedoussis G, de Faire U, Feranil AB, Ferrieres J, Ferrucci L, Freimer NB, Gieger C, Groop LC, Gudnason V, Gyllenstein U, Hamsten A, Harris TB, Hingorani A, Hirschhorn JN, Hofman A, Hovingh GK, Hsiung CA, Humphries SE, Hunt SC, Hveem K, Iribarren C, Jarvelin MR, Jula A, Kahonen M, Kaprio J, Kesaniemi A, Kivimaki M, Kooner JS, Koudstaal PJ, Krauss RM, Kuh D, Kuusisto J, Kyvik KO, Laakso M, Lakka TA, Lind L, Lindgren CM, Martin NG, Marz W, McCarthy MI, McKenzie CA, Meneton P, Metspalu A, Moilanen L, Morris AD, Munroe PB, Njolstad I, Pedersen NL, Power C, Pramstaller PP, Price JF, Psaty BM, Quertermous T, Rauramaa R, Saleheen D, Salomaa V, Sanghera DK, Saramies J, Schwarz PE, Sheu WH, Shuldiner AR, Siegbahn A, Spector TD, Stefansson K, Strachan DP, Tayo BO, Tremoli E, Tuomilehto J, Uusitupa M, van Duijn CM, Vollenweider P, Wallentin L, Wareham NJ, Whitfield JB, Wolffenbuttel BH, Ordovas JM, Boerwinkle E, Palmer CN, Thorsteinsdottir U, Chasman DI, Rotter JI, Franks PW, Ripatti S, Cupples LA, Sandhu MS, Rich SS, Boehnke M, Deloukas P, Kathiresan S, Mohlke KL, Ingelsson E and Abecasis GR. Discovery and refinement of loci associated with lipid levels. *Nat Genet.* 2013;45:1274-83.

5. Smith CA, Want EJ, O'Maille G, Abagyan R and Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry.* 2006;78:779-87.

6. Shevchenko A and Simons K. Lipidomics: coming to grips with lipid diversity. *Nature reviews Molecular cell biology.* 2010;11:593-8.

7. Stegemann C, Pechlaner R, Willeit P, Langley SR, Mangino M, Mayr U, Menni C, Moayyeri A, Santer P, Rungger G, Spector TD, Willeit J, Kiechl S and Mayr M.



- 1 Lipidomics Profiling and Risk of Cardiovascular Disease in the Prospective  
2 Population-Based Bruneck Study. *Circulation*. 2014;129:1821-1831.
- 3 8. DeJaegher B and Heyden YV. Ruggedness and robustness testing. *Journal of*  
4 *chromatography A*. 2007;1158:138-57.
- 5 9. Saleheen D, Zaidi M, Rasheed A, Ahmad U, Hakeem A, Murtaza M, Kayani W,  
6 Faruqui A, Kundi A, Zaman KS, Yaqoob Z, Cheema LA, Samad A, Rasheed SZ,  
7 Mallick NH, Azhar M, Jooma R, Gardezi AR, Memon N, Ghaffar A, Fazal ur R, Khan  
8 N, Shah N, Ali Shah A, Samuel M, Hanif F, Yameen M, Naz S, Sultana A, Nazir A,  
9 Raza S, Shazad M, Nasim S, Javed MA, Ali SS, Jafree M, Nisar MI, Daood MS,  
10 Hussain A, Sarwar N, Kamal A, Deloukas P, Ishaq M, Frossard P and Danesh J.  
11 The Pakistan Risk of Myocardial Infarction Study: a resource for the study of  
12 genetic, lifestyle and other determinants of myocardial infarction in South Asia.  
13 *Eur J Epidemiol*. 2009;24:329-38.
- 14 10. Abegunde DO, Mathers CD, Adam T, Ortegon M and Strong K. The burden and  
15 costs of chronic diseases in low-income and middle-income countries. *Lancet*.  
16 2007;370:1929-38.
- 17 11. GBD Mortality and Causes of Death Collaborators. Global, regional, and national  
18 life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of  
19 death, 1980-2015: a systematic analysis for the Global Burden of Disease Study  
20 2015. *Lancet*. 2016;388:1459-1544.
- 21 12. Ghaffar A, Reddy KS and Singhi M. Burden of non-communicable diseases in  
22 South Asia. *BMJ*. 2004;328:807-10.
- 23 13. Nordestgaard BG. Triglyceride-Rich Lipoproteins and Atherosclerotic  
24 Cardiovascular Disease: New Insights From Epidemiology, Genetics, and Biology.  
25 *Circulation research*. 2016;118:547-63.
- 26 14. Kozaki K, Gotoda T, Kawamura M, Shimano H, Yazaki Y, Ouchi Y, Orimo H and  
27 Yamada N. Mutational analysis of human lipoprotein lipase by carboxy-terminal  
28 truncation. *Journal of lipid research*. 1993;34:1765-72.

15. Myocardial Infarction Genetics and CARDIoGRAM Exome Consortia Investigators, Stitzel NO, Stirrups KE, Masca NG, Erdmann J, Ferrario PG, König IR, Weeke PE, Webb TR, Auer PL, Schick UM, Lu Y, Zhang H, Dube MP, Goel A, Farrall M, Peloso GM, Won HH, Do R, van Iperen E, Kanoni S, Kruppa J, Mahajan A, Scott RA, Willenberg C, Braund PS, van Capelleveen JC, Doney AS, Donnelly LA, Asselta R, Merlini PA, Duga S, Marziliano N, Denny JC, Shaffer CM, El-Mokhtari NE, Franke A, Gottesman O, Heilmann S, Hengstenberg C, Hoffman P, Holmen OL, Hveem K, Jansson JH, Jöckel KH, Kessler T, Kriebel J, Laugwitz KL, Marouli E, Martinelli N, McCarthy MI, Van Zuydam NR, Meisinger C, Esko T, Mihailov E, Escher SA, Alver M, Moebus S, Morris AD, Müller-Nurasyid M, Nikpay M, Olivieri O, Lemieux Perreault LP, AlQarawi A, Robertson NR, Akinkunle KO, Reilly DF, Vogt TF, Yin W, Asselbergs FW, Kooperberg C, Jackson RD, Stahl E, Strauch K, Varga TV, Waldenberger M, Zeng L, Kraja AT, Liu C, Ehret GB, Newton-Cheh C, Chasman DI, Chowdhury R, Ferrario M, Ford I, Jukema JW, Kee F, Kuulasmaa K, Nordestgaard BG, Perola M, Saleheen D, Sattar N, Surendran P, Tregouet D, Young R, Howson JM, Butterworth AS, Danesh J, Ardisino D, Bottinger EP, Erbel R, Franks PW, Girelli D, Hall AS, Hovingh GK, Kastrati A, Lieb W, Meitinger T, Kraus WE, Shah SH, McPherson R, Orho-Melander M, Melander O, Metspalu A, Palmer CN, Peters A, Rader D, Reilly MP, Loos RJ, Reiner AP, Roden DM, Tardif JC, Thompson JR, Wareham NJ, Watkins H, Willer CJ, Kathiresan S, Deloukas P, Samani NJ and Schunkert H. Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. *N Engl J Med*. 2016;374:1134-44.
16. blockTools: Blocking, Assignment, and Diagnosing Interference in Randomized Experiments [computer program]. Version 0.6-1; 2014.
17. Han X and Gross RW. Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics. *Journal of lipid research*. 2003;44:1071-9.

18. Graessler J, Schwudke D, Schwarz PE, Herzog R, Shevchenko A and Bornstein SR. Top-down lipidomics reveals ether lipid deficiency in blood plasma of hypertensive patients. *PLoS one*. 2009;4:e6261.
19. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak MY, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL and Mallick P. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol*. 2012;30:918-20.
20. Quehenberger O, Armando AM, Brown AH, Milne SB, Myers DS, Merrill AH, Bandyopadhyay S, Jones KN, Kelly S, Shaner RL, Sullards CM, Wang E, Murphy RC, Barkley RM, Leiker TJ, Raetz CR, Guan Z, Laird GM, Six DA, Russell DW, McDonald JG, Subramaniam S, Fahy E and Dennis EA. Lipidomics reveals a remarkable diversity of lipids in human plasma. *Journal of lipid research*. 2010;51:3299-305.
21. Saleheen D, Soranzo N, Rasheed A, Scharnagl H, Gwilliam R, Alexander M, Inouye M, Zaidi M, Potter S, Haycock P, Bumpstead S, Kaptoge S, Di Angelantonio E, Sarwar N, Hunt SE, Sheikh N, Shah N, Samuel M, Haider SR, Murtaza M, Thompson A, Gobin R, Butterworth A, Ahmad U, Hakeem A, Zaman KS, Kundi A, Yaqoob Z, Cheema LA, Qamar N, Faruqui A, Mallick NH, Azhar M, Samad A, Ishaq M, Rasheed SZ, Jooma R, Niazi JH, Gardezi AR, Memon NA, Ghaffar A, Rehman FU, Hoffmann MM, Renner W, Kleber ME, Grammer TB, Stephens J, Attwood A, Koch K, Hussain M, Kumar K, Saleem A, Daood MS, Gul AA, Abbas S, Zafar J, Shahid F, Bhatti SM, Ali SS, Muhammad F, Sagoo G, Bray S, McGinnis R, Dudbridge F, Winkelmann BR, Boehm B, Thompson S, Ouwehand W, Marz W, Frossard P, Danesh J and Deloukas P. Genetic determinants of major

- 1 blood lipids in Pakistanis compared with Europeans. *Circ Cardiovasc Genet*.  
2 2010;3:348-57.
- 3 22. Ward JH. Hierarchical Grouping to Optimize an Objective Function. *Journal of the*  
4 *American Statistical Association*. 1963;58:236-244.
- 5 23. Opgen-Rhein R and Strimmer K. From correlation to causation networks: a simple  
6 approximate learning algorithm and its application to high-dimensional plant gene  
7 expression data. *BMC Syst Biol*. 2007;1:37.
- 8 24. Krumsiek J, Suhre K, Illig T, Adamski J and Theis FJ. Gaussian graphical modeling  
9 reconstructs pathway reactions from high-throughput metabolomics data. *BMC*  
10 *Syst Biol*. 2011;5:21.
- 11 25. Gobbi A, Iorio F, Dawson KJ, Wedge DC, Tamborero D, Alexandrov LB, Lopez-  
12 Bigas N, Garnett MJ, Jurman G and Saez-Rodriguez J. Fast randomization of large  
13 genomic datasets while preserving alteration counts. *Bioinformatics*.  
14 2014;30:i617-23.
- 15 26. Stata Statistical Software: Release 13 [computer program]. College Station,  
16 Texas: StataCorp LP; 2013.
- 17 27. R: A Language and Environment for Statistical Computing [computer program].  
18 Version 3.2.0. Vienna, Austria: R Foundation for Statistical Computing; 2015.
- 19 28. National Institute of Population Studies (NIPS) Pakistan and ICF International.  
20 Pakistan Demographic and Health Survey 2012-13. 2013.
- 21 29. Kastenmuller G, Raffler J, Gieger C and Suhre K. Genetics of human metabolism:  
22 an update. *Hum Mol Genet*. 2015.
- 23 30. Jenkins BJ, Seyssel K, Chiu S, Pan PH, Lin SY, Stanley E, Ament Z, West JA,  
24 Summerhill K, Griffin JL, Vetter W, Autio KJ, Hiltunen K, Hazebrouck S,  
25 Stepankova R, Chen CJ, Alligier M, Laville M, Moore M, Kraft G, Cherrington A,  
26 King S, Krauss RM, de Schryver E, Van Veldhoven PP, Ronis M and Koulman A.  
27 Odd Chain Fatty Acids; New Insights of the Relationship Between the Gut  
28 Microbiota, Dietary Intake, Biosynthesis and Glucose Intolerance. *Sci Rep*.  
29 2017;7:44845.

31. Suhre K, Shin SY, Petersen AK, Mohny RP, Meredith D, Wagle B, Altmaier E, Deloukas P, Erdmann J, Grundberg E, Hammond CJ, de Angelis MH, Kastenmuller G, Kottgen A, Kronenberg F, Mangino M, Meisinger C, Meitinger T, Mewes HW, Milburn MV, Prehn C, Raffler J, Ried JS, Romisch-Margl W, Samani NJ, Small KS, Wichmann HE, Zhai G, Illig T, Spector TD, Adamski J, Soranzo N and Gieger C. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*. 2011;477:54-60.
32. Illig T, Gieger C, Zhai G, Romisch-Margl W, Wang-Sattler R, Prehn C, Altmaier E, Kastenmuller G, Kato BS, Mewes HW, Meitinger T, de Angelis MH, Kronenberg F, Soranzo N, Wichmann HE, Spector TD, Adamski J and Suhre K. A genome-wide perspective of genetic variation in human metabolism. *Nat Genet*. 2010;42:137-41.
33. Dugi KA, Dichek HL, Talley GD, Brewer HB, Jr. and Santamarina-Fojo S. Human lipoprotein lipase: the loop covering the catalytic site is essential for interaction with lipid substrates. *J Biol Chem*. 1992;267:25086-91.
34. Osei M, Griffin JL and Koulman A. Hyphenating size-exclusion chromatography with electrospray mass spectrometry; using on-line liquid-liquid extraction to study the lipid composition of lipoprotein particles. *Rapid communications in mass spectrometry : RCM*. 2015;29:1969-76.
35. Arimoto I, Saito H, Kawashima Y, Miyajima K and Handa T. Effects of sphingomyelin and cholesterol on lipoprotein lipase-mediated lipolysis in lipid emulsions. *Journal of lipid research*. 1998;39:143-51.
36. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D, Parish S, Shen H, Galan P, Meneton P, Hercberg S, Zelenika D, Chen WM, Li Y, Scott LJ, Scheet PA, Sundvall J, Watanabe RM, Nagaraja R, Ebrahim S, Lawlor DA, Ben-Shlomo Y, Davey-Smith G, Shuldiner AR, Collins R, Bergman RN, Uda M, Tuomilehto J, Cao A, Collins FS, Lakatta E, Lathrop GM, Boehnke M, Schlessinger D, Mohlke KL and

- 1 Abecasis GR. Newly identified loci that influence lipid concentrations and risk of  
2 coronary artery disease. *Nat Genet.* 2008;40:161-9.
- 3 37. Feng Q, Baker SS, Liu W, Arbizu RA, Aljomah G, Khatib M, Nugent CA, Baker RD,  
4 Forte TM, Hu Y and Zhu L. Increased apolipoprotein A5 expression in human and  
5 rat non-alcoholic fatty livers. *Pathology.* 2015;47:341-8.
- 6 38. Pennacchio LA, Olivier M, Hubacek JA, Cohen JC, Cox DR, Fruchart JC, Krauss RM  
7 and Rubin EM. An apolipoprotein influencing triglycerides in humans and mice  
8 revealed by comparative sequencing. *Science.* 2001;294:169-73.
- 9 39. Talmud PJ, Hawe E, Martin S, Olivier M, Miller GJ, Rubin EM, Pennacchio LA and  
10 Humphries SE. Relative contribution of variation within the APOC3/A4/A5 gene  
11 cluster in determining plasma triglycerides. *Hum Mol Genet.* 2002;11:3039-46.
- 12 40. Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, Ban MR, Martins RA,  
13 Kennedy BA, Hassell RG, Visser ME, Schwartz SM, Voight BF, Elosua R, Salomaa  
14 V, O'Donnell CJ, Dallinga-Thie GM, Anand SS, Yusuf S, Huff MW, Kathiresan S and  
15 Hegele RA. Excess of rare variants in genes identified by genome-wide  
16 association study of hypertriglyceridemia. *Nat Genet.* 2010;42:684-7.
- 17 41. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, Arnold M, Erte  
18 I, Forgetta V, Yang TP, Walter K, Menni C, Chen L, Vasquez L, Valdes AM, Hyde  
19 CL, Wang V, Ziemek D, Roberts P, Xi L, Grundberg E, Multiple Tissue Human  
20 Expression Resource C, Waldenberger M, Richards JB, Mohnen RP, Milburn MV,  
21 John SL, Trimmer J, Theis FJ, Overington JP, Suhre K, Brosnan MJ, Gieger C,  
22 Kastenmuller G, Spector TD and Soranzo N. An atlas of genetic influences on  
23 human blood metabolites. *Nat Genet.* 2014;46:543-50.
- 24 42. Li F, Jiang C, Larsen MC, Bushkofsky J, Krausz KW, Wang T, Jefcoate CR and  
25 Gonzalez FJ. Lipidomics reveals a link between CYP1B1 and SCD1 in promoting  
26 obesity. *Journal of proteome research.* 2014;13:2679-87.
- 27 43. Talmud PJ, Drenos F, Shah S, Shah T, Palmen J, Verzilli C, Gaunt TR, Pallas J,  
28 Lovering R, Li K, Casas JP, Sofat R, Kumari M, Rodriguez S, Johnson T, Newhouse  
29 SJ, Dominiczak A, Samani NJ, Caulfield M, Sever P, Stanton A, Shields DC,

- 1 Padmanabhan S, Melander O, Hastie C, Delles C, Ebrahim S, Marmot MG, Smith  
2 GD, Lawlor DA, Munroe PB, Day IN, Kivimaki M, Whittaker J, Humphries SE,  
3 Hingorani AD, investigators A, investigators N and Consortium B. Gene-centric  
4 association signals for lipids and apolipoproteins identified via the HumanCVD  
5 BeadChip. *Am J Hum Genet.* 2009;85:628-42.
- 6 44. Plaisier CL, Horvath S, Huertas-Vazquez A, Cruz-Bautista I, Herrera MF, Tusie-  
7 Luna T, Aguilar-Salinas C and Pajukanta P. A systems genetics approach  
8 implicates USF1, FADS3, and other causal candidate genes for familial combined  
9 hyperlipidemia. *PLoS Genet.* 2009;5:e1000642.
- 10 45. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, Kaplan  
11 L, Bennett D, Li Y, Tanaka T, Voight BF, Bonnycastle LL, Jackson AU, Crawford G,  
12 Surti A, Guiducci C, Burt NP, Parish S, Clarke R, Zelenika D, Kubalanza KA,  
13 Morken MA, Scott LJ, Stringham HM, Galan P, Swift AJ, Kuusisto J, Bergman RN,  
14 Sundvall J, Laakso M, Ferrucci L, Scheet P, Sanna S, Uda M, Yang Q, Lunetta KL,  
15 Dupuis J, de Bakker PI, O'Donnell CJ, Chambers JC, Kooner JS, Hercberg S,  
16 Meneton P, Lakatta EG, Scuteri A, Schlessinger D, Tuomilehto J, Collins FS, Groop  
17 L, Altshuler D, Collins R, Lathrop GM, Melander O, Salomaa V, Peltonen L, Orho-  
18 Melander M, Ordovas JM, Boehnke M, Abecasis GR, Mohlke KL and Cupples LA.  
19 Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet.*  
20 2009;41:56-65.
- 21 46. Draisma HH, Pool R, Kobl M, Jansen R, Petersen AK, Vaarhorst AA, Yet I, Haller T,  
22 Demirkan A, Esko T, Zhu G, Bohringer S, Beekman M, van Klinken JB, Romisch-  
23 Margl W, Prehn C, Adamski J, de Craen AJ, van Leeuwen EM, Amin N, Dharuri H,  
24 Westra HJ, Franke L, de Geus EJ, Hottenga JJ, Willemsen G, Henders AK,  
25 Montgomery GW, Nyholt DR, Whitfield JB, Penninx BW, Spector TD, Metspalu A,  
26 Eline Slagboom P, van Dijk KW, t Hoen PA, Strauch K, Martin NG, van Ommen GJ,  
27 Illig T, Bell JT, Mangino M, Suhre K, McCarthy MI, Gieger C, Isaacs A, van Duijn  
28 CM and Boomsma DI. Genome-wide association study identifies novel genetic

1 variants contributing to variation in blood metabolite levels. *Nat Commun.*  
2 2015;6:7208.

3 47. Roberts LD, Koulman A and Griffin JL. Towards metabolic biomarkers of insulin  
4 resistance and type 2 diabetes: progress from the metabolome. *The Lancet*  
5 *Diabetes & Endocrinology*. 2014;2:65-75.

6 48. Meikle PJ, Wong G, Barlow CK and Kingwell BA. Lipidomics: Potential role in risk  
7 prediction and therapeutic monitoring for diabetes and cardiovascular disease.  
8 *Pharmacology & therapeutics*. 2014;143:12-23.

9 49. Wishart DS. Applications of metabolomics in drug discovery and development.  
10 *Drugs R D*. 2008;9:307-22.

11

CONFIDENTIAL



## FIGURE LEGENDS

### Tables

Table 1. Categorization of lipids in positive and negative ionization mode measured by direct infusion high resolution mass spectrometry in PROMIS

(+) denotes lipids measured in positive ionization mode; (–) denotes lipids measured in negative ionization mode.

### Figures

Figure 1. Schematic of the peak-picking process

(a) XCMS was used to average 50 spectra in positive and negative ionization modes, yielding (b) the average mass spectrum for that particular polarity, for which signals were obtained using a peak-picking algorithm that determined the (c) peak signal at the midpoint of a line drawn at half-height for peaks near signals that corresponded to known lipids. Signals and deviations from known lipids were then (d) combined in a database, and separated into separate files for (e) signals and (f) deviations for each lipid.

Figure 2. Heat map showing relationships between lipid subclasses based on partial correlations derived using Gaussian Graphical Modelling

This heat map shows the relationships between the lipid subclasses. Each cell shows the ratio of the observed number of GGM edges for all lipids within that subclass to the total number of connections between lipids within that subclass. The cells are coloured red or blue according to whether the observed number of GGM edges is more or less than

1 expected due to chance alone, and a box is drawn around the cell if there is a significant  
2 difference between the number of observed versus expected GGM edges.

3  
4 Figure 3. Phenome scan of lipids that are significantly associated with *LPL*

5  
6 Association results are shown for each lipid that is significantly associated with rs328  
7 (chr8:19819724) in the *LPL* region at  $P < 5 \times 10^{-8}$ .

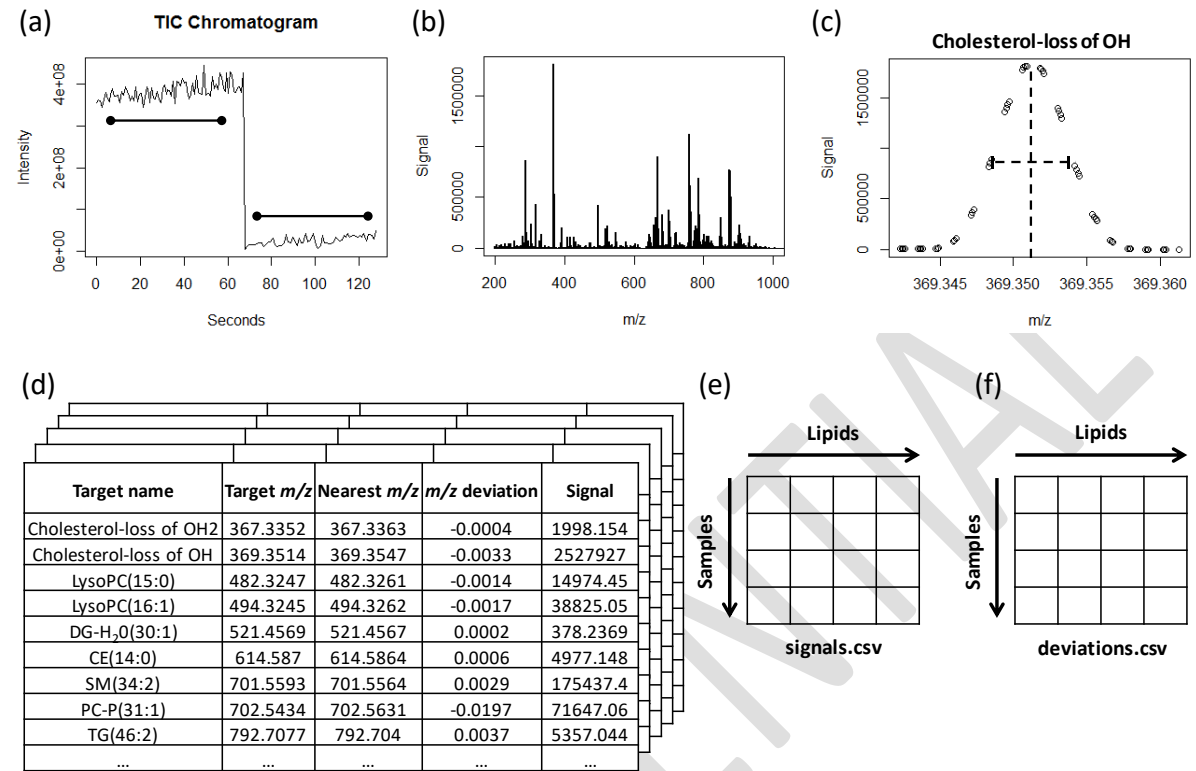
8  
9 Figure 4. Cross-correlations of circulating biomarkers with the triglyceride most  
10 strongly associated with *LPL*

11  
12 The triglyceride that was most strongly associated with rs328 (chr8:19819724) in the  
13 *LPL* region was TG(53:3) ( $m/z$  888.8016). The cross-correlations of this triglyceride with  
14 a range of circulating biomarkers is shown. Analyses were adjusted for age and sex.

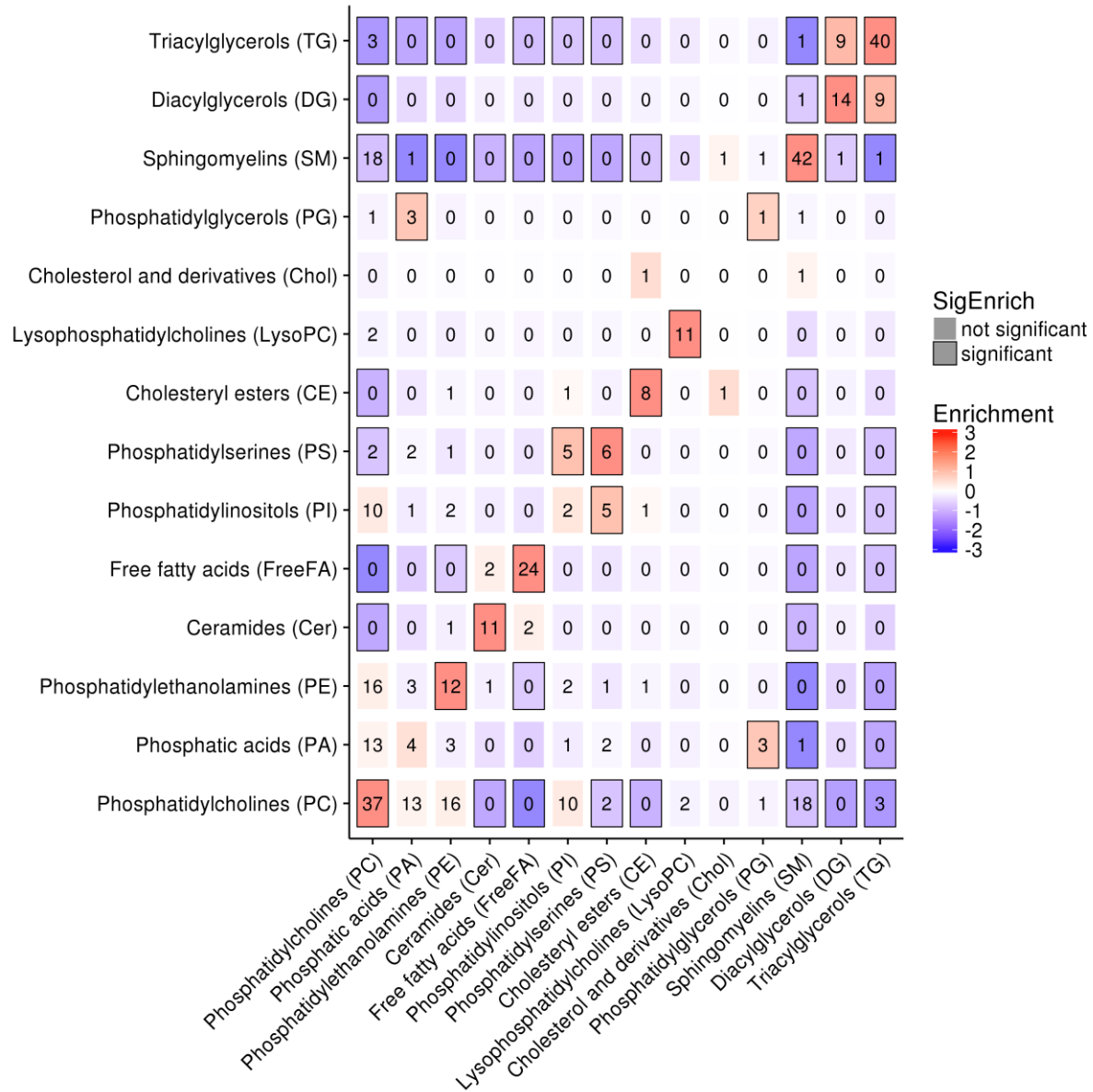
**Table 1.** Categorization of lipids in positive and negative ionization mode measured by direct infusion high resolution mass spectrometry in PROMIS

Overall lipid category	Lipid main class	Lipid subclass	No. (%) of lipids
Fatty acyls (FA)	Fatty acids & conjugates	Free fatty acids (FreeFA) (–)	22 (5.0%)
Glycerolipids (GL)	Diradylglycerols	Diacylglycerols (DG) (+)	19 (4.3%)
	Triradylglycerols	Triacylglycerols (TG) (+)	56 (12.6%)
Glycerophospholipids (GP)	Glycerophosphates	Phosphatic acids (PA) (–)	20 (4.5%)
		Phosphatic acids (PA) (+)	13 (2.9%)
	Glycerophosphocholines	Lysophosphocholines (LysoPC) (+)	8 (1.8%)
		Phosphatidylcholines (PC) (–)	52 (11.7%)
		Phosphatidylcholines (PC) (+)	54 (12.2%)
	Glycerophosphoethanolamines	Phosphatidylethanolamines (PE) (–)	24 (5.4%)
		Phosphatidylethanolamines (PE) (+)	16 (3.6%)
	Glycerophosphoglycerols	Phosphatidylglycerols (PG) (–)	5 (1.1%)
	Glycerophosphoinositols	Phosphatidylinositols (PI) (–)	25 (5.6%)
	Glycerophosphoserines	Phosphatidylserines (PS) (–)	22 (5.0%)
Sphingolipids (SP)	Ceramides	Ceramides (Cer) (–)	16 (3.6%)
	Phosphosphingolipids	Sphingomyelins (SM) (–)	51 (11.5%)
		Sphingomyelins (SM) (+)	27 (6.1%)
Sterol lipids (ST)	Sterols	Cholesterols & derivatives (Chol) (+)	2 (0.5%)
		Cholesteryl esters (CE) (+)	12 (2.7%)
Total lipids			444 (100%)

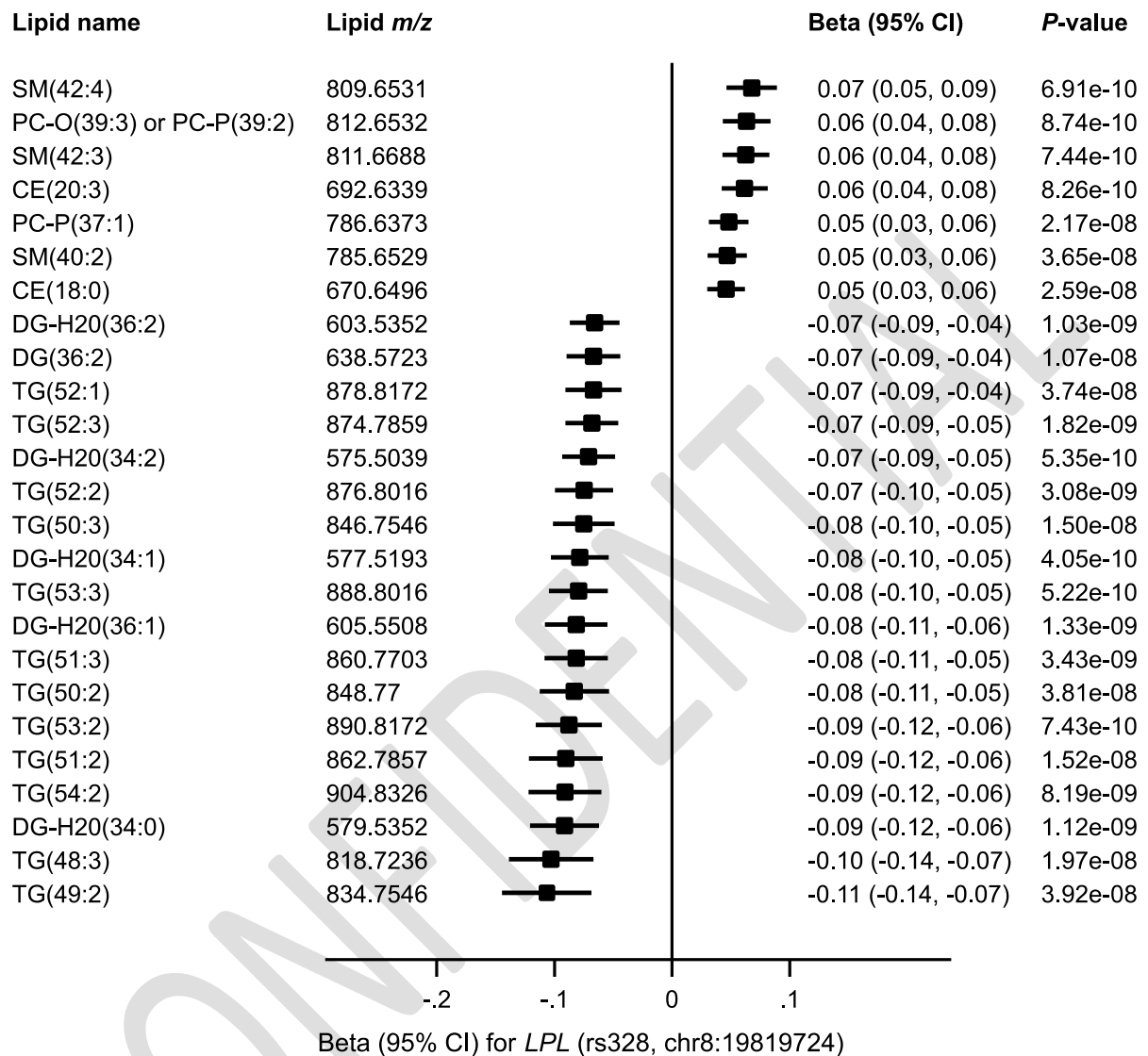
**Figure 1.** Schematic of the peak-picking process



**Figure 2.** Heat map showing relationships between lipid subclasses based on partial correlations derived using Gaussian Graphical Modelling



**Figure 3** Phenome scan of lipids that are significantly associated with *LPL*



Significance threshold:  $P < 5e-8$

**Figure 4** Cross-correlations of circulating biomarkers with the triglyceride most strongly associated with *LPL*: TG(53:3) (*m/z* 888.8016)

