Guida, F., Tan, Y., Corbin, L. J., Bull, C. J., Vincent, E. E., Timpson, N. J., & Johansson, M. (Accepted/In press). The blood metabolome of incident kidney cancer: A case-control study nested within the MetKid consortium. *PLoS Medicine*.

Title: "The blood metabolome of incident kidney cancer: A case-control study nested within the MetKid consortium"

Florence Guida[1*], Vanessa Y Tan[2,3*], Laura J Corbin[2,3*], Karl Smith-Byrne[1*], Karine Alcala[1], Claudia Langenberg[4], Isobel D Stewart[4], Adam Butterworth[5,6,7,8,9], Praveen Surendran[10,11,7,12], David Achaintre[1], Jerzy Adamski[13,14,15,16], Pilar Amiano Exezarreta[17], Manuela M Bergmann[18], Caroline J Bull[2,3,19], Christina C Dahm[20], Audrey Gicquiau[1], Graham G Giles[21,22,23], Marc Gunter[1], Toomas Haller[24], Arnulf Langhammer[25,26], Tricia L Larose[1,25,27], Börje Ljungberg[28], Andres Metspalu[24], Roger Milne[21,29,30], David C Muller[31], Therese H Nøst[32], Elin Pettersen Sørgjerd[25], Cornelia Prehn[13], Elio Riboli[31], Sabina Rinaldi[1], Joseph Rothwell[33], Augustin Scalbert[1], Julie A Schmidt[34], Gianluca Severi[33,35], Sabina Sieri[36], Roel Vermeulen[37], Emma E Vincent[2,3,19], Melanie Waldenberg[13], Nicholas J Timpson[2,3*], Mattias Johansson[1*]

* these authors contributed equally to this work

[1]International Agency for Research on Cancer (IARC), Lyon, France,

[2]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom,

[3]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom,

[4]MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom,

[5]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom,

[6]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, United Kingdom,

[7]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, United Kingdom,

[8]National Institute for Health Research Cambridge Biomedical Research Center, University of Cambridge and Cambridge University Hospitals, Cambridge, United Kingdom,

[9]National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, United Kingdom,

[10]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom,

[11]British Heart Foundation Centre of Research Excellence, Cambridge, United Kingdom,

[12]Rutherford Fund Fellow, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom,

[13]Research Unit of Molecular Endocrinology and Metabolism, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Germany,

[14]German Center for Diabetes Research (DZD), Neuherberg, Germany,

[15]Chair of Experimental Genetics, School of Life Science, Weihenstephan, Technische Universität München, Freising, Germany,

[16]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore,

[17]Public Health Division of Gipuzkoa, BioDonostia Research Institute, Donostia-San Sebastian, CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain,

[18]German Institute of Human Nutrition Potsdam-Rehbrücke, Nuthetal, Germany,

[19]School of Cellular and Molecular Medicine, University of Bristol, Bristol, United Kingdom,

[20]Department of Public Health, Aarhus University, Aarhus, Denmark,

[21]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia,

[22]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Australia,

[23]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Australia,

[24]Institute of Genomics, University of Tartu, Tartu, Estonia,

[25]HUNT Research Centre, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Levanger, Norway,

[26]Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway,

[27]Department of Community Medicine and Global Health, Institute of Health and Society, University of Oslo, Oslo, Norway,

[28]Department of surgical and perioperative sciences, urology and andrology, Umeå University, Umeå, Sweden,

[29]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Australia,

[30]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Melbourne, Australia,

[31]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, St. Mary's Campus, London, United Kingdom,

[32]Department of Community Medicine, Faculty of Health Sciences, UiT - The Arctic University of Norway, Tromsø, Norway,

[33]Université Paris-Saclay, UVSQ, Inserm, Gustave Roussy, Équipe "Exposome et Hérédité", CESP UMR1018, Inserm, Villejuif, France,

[34]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom,

[35]Department of Statistics, Computer Science and Applications (DISIA), University of Florence, Florence, Italy,

[36]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milano, Italy,

[37]Institute for Risk Assessment Sciences (IRAS),, Utrecht University, Utrecht, the Netherlands

Corresponding author details:

Mattias Johansson, johanssonm@iarc.fr & Nicholas J. Timpson n.j.timpson@bristol.ac.uk

Short title: The blood metabolome of incident kidney cancer

Disclaimer:

F.G, K.S.B, K.A. and M.J are identified as personnel of the International Agency for Research on Cancer / World Health Organization. These authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

Abstract

Background: Excess bodyweight and related metabolic perturbations have been implicated in kidney cancer aetiology, but the specific molecular mechanisms underlying these relationships are poorly understood. In this study we sought to identify circulating metabolites that predispose kidney cancer and to evaluate the extent to which they are influenced by body-mass index (BMI).

Methods and Findings: We assessed the association between circulating levels of 1,416 metabolites and incident kidney cancer using pre-diagnostic blood samples from up to 1,305 kidney cancer case-control pairs from five prospective cohort studies. Cases were diagnosed on average eight years after blood collection. We found 25 metabolites robustly associated with kidney cancer risk. In particular, 14 glycerophospholipids (GPL) were inversely associated with risk, including eight phosphatidylcholines (PC) and two plasmalogens. The PC with the strongest association was PC ae C34:3 with an odds-ratio (OR) for one standard deviation (SD) increment of 0.75 (95% CI: 0.68 to 0.83, p=2.6x10^{-8}). In contrast, four amino acids, including glutamate (OR for 1 SD=1.39, 95% CI: 1.20 - 1.60, p=1.6x10^{-5}), were positively associated with risk. Adjusting for BMI partly attenuated the risk association for some – but not all – metabolites, whereas other known risk factors of kidney cancer, such as smoking and alcohol consumption, had minimal impact on the observed associations. A Mendelian randomization analysis of the influence of BMI on the blood metabolome highlighted that some metabolites associated with kidney cancer risk are influenced by BMI. Specifically, elevated BMI appeared to decrease levels of several GPLs that were also found inversely associated with kidney cancer risk (e.g -0.17 standard deviation change [$ß_{BMI}$] in 1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2) levels per SD change in BMI, $p=3.4x10^{-5}$). BMI was also associated with increased levels of glutamate ($ß_{BMI}$: 0.12, $p=1.5x10^{-3}$).

Whilst our results were robust across the participating studies, they were limited to study participants of European descent and it will, therefore, be important to evaluate if our findings can be generalized to populations with different genetic backgrounds.

Conclusions: This study suggests a potentially important role of the blood metabolome in kidney cancer aetiology by highlighting a wide range of metabolites associated with the risk of developing kidney cancer, and the extent to which changes in levels of these metabolites are driven by BMI - the principal modifiable risk factor of kidney cancer.

Author summary

Why was this study done?

- Several modifiable risk factors have been established for kidney cancer, amongst which elevated BMI and obesity are central.
- The biological mechanisms underlying these relationships are poorly understood, but obesity-related metabolic perturbations may be important.

What did the researchers do and find?

- We looked at the association between kidney cancer and the levels of 1,416 metabolites measured in blood on average eight years before the disease onset. The study included 1,305 kidney cancer cases and 1,305 healthy controls.
- We found 25 metabolites robustly associated with kidney cancer risk.
- Specifically, multiple glycerophospholipids were inversely associated with risk, while several amino acids were positively associated with risk.
- Accounting for body-mass index (BMI) highlighted that some – but not all – metabolites associated with kidney cancer risk are influenced by BMI.

What do these findings mean?

- These findings illustrate the potential utility of prospectively measured metabolites in helping us to understand the aetiology of kidney cancer.
- By examining overlap between the metabolomic profile of prospective risk of kidney cancer and that of modifiable risk factors for the disease – in this case BMI – we can begin to identify biological pathways relevant to disease onset.

Introduction

Kidney cancer is the 14th most common cancer worldwide with renal cell carcinoma (RCC) making up the majority of cases[1]. There are important geographical variations in kidney cancer incidence that are only partly understood [2]. Excess bodyweight and related conditions, such as hypertension, diabetes, and related metabolic perturbations, are among the most robustly implicated risk factors for kidney cancer, with support from both traditional observational studies and genetic studies [2-7]. For instance, in the UK, an estimated 24% of kidney cancer cases are attributable to overweight and obesity, making this the leading modifiable risk factor for the disease [8]. Germline mutations responsible for an inherited predisposition to kidney cancer (a small proportion of kidney cancer cases) have a key role in regulating cellular metabolism [9] and this, together with evidence of extensive metabolic reprogramming within tumours themselves [10], have led to the characterisation of kidney cancer as a metabolic disease. However, the molecular mechanisms predisposing kidney cancer remain largely unknown. Given the likely metabolic underpinnings of kidney cancer, studies of circulating metabolites, the downstream products of cellular regulatory processes, may improve our understanding into pathways relevant to kidney cancer aetiology [11].

Metabolite variations are the result of genetic and non-genetic factors and provide a read-out of physiological functions [12]. Metabolomics technologies based on mass spectrometry (MS) and nuclear magnetic resonance (NMR) have enabled the systematic quantification of hundreds of metabolites (the 'metabolome') from a single biological sample. The analysis of metabolites has enabled a more thorough exploration of an individual's metabolic status, providing important insights into the biological pathways leading to diseases such as cancer [11,13,14] and has enabled the discovery and development of new drug targets[15]. Already, global metabolic profiling of blood,[16-19] urine[20-24] and tissue samples[24-27] has been

used to characterise kidney cancer and identify novel potential diagnostic biomarkers. However, because of the cross-sectional or retrospective design of these studies, they could not inform the identification of biomarkers for incident disease development. Prospective cohort studies, where healthy individuals initially donate blood at recruitment and are longitudinally followed over time for incident disease, can circumvent many of the problems of retrospective study designs - particularly where the focus is on identifying risk factors for disease onset.

The aim of this study was to identify circulating metabolites associated with the development of kidney cancer in a prospective case-control framework. We used two complementary metabolomics platforms [28] to quantify over 1000 metabolites in blood samples donated by research participants later diagnosed with kidney cancer along with matched control subjects. In a series of follow-up analyses, including a two-sample Mendelian randomization (MR) analysis, which uses genetic variants as proxies for an exposure of interest [29], we evaluated the extent to which the metabolomic signature of disease risk could be explained by body mass index (BMI), the leading modifiable risk factor for kidney cancer.

Methods

Analytical strategy (Figure 1)

The primary analysis was pre-defined and involved investigating the association between circulating levels of metabolites and kidney cancer risk using pre-diagnostic metabolomics measurements in a case-control study nested within multiple large-scale prospective cohorts (the MetKid consortium). Adjustment for known risk factors for kidney cancer (BMI, hypertension, alcohol consumption and smoking)[2] was then carried out to evaluate the

extent to which these could explain the associations between blood metabolites and kidney cancer risk.

A natural complementary analysis would have been to interrogate the potentially causal role for the identified risk-associated metabolites in kidney cancer aetiology through Mendelian randomization (MR) analyses. However, given the methodological constraints of MR in this context, specifically, widespread pleiotropic instruments, which would violate the MR assumptions, we chose not to pursue this analysis. Our analysis plan was therefore revised, and as a secondary analysis, we rather used a two-sample MR approach to estimate the causal effect of BMI on the blood metabolome. This analysis complemented the main risk analysis by quantifying the extent to which BMI – the central risk factor of kidney cancer – influenced the identified risk metabolites. This study is reported as per the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and STROBE-MR guidelines (**Supplementary Tables S1 and S2)** [30,31].

Study population, sample collection and follow-up

Our study population consisted of kidney cancer nested case-control studies drawn from 5 independent cohorts: the European Prospective Investigation into Cancer and Nutrition (EPIC), The Melbourne Collaborative Cohort Study (MCCS), Northern Sweden Health and Disease study (NSHDS), University of Tartu - Estonian Biobank (Estonian BB) and The Trøndelag Health Study (HUNT) (**Supplementary Table S3;** details of the cohorts are described in the Supplementary Methods). Cases were defined as participants diagnosed with incident malignant neoplasm of the kidney or renal pelvis (ICD-O3 code C64/C65) who gave a blood sample at recruitment. In each independent cohort, one randomly selected control without history of kidney cancer was matched to each case based on age, sex and date of blood collection. There were small variations between the cohorts in the tightness by which controls

were matched to cases according to their age and date of blood draw (see Supplementary Methods), owing to inherent differences in demography and availability of controls. The study was approved by the International Agency for Research on Cancer (IARC) Ethics Committee.

Metabolite data acquisition and quality control (QC)

Plasma and serum samples from 2,614 participants (1,307 cases and 1,307 controls) were analysed. Samples from all cohorts were analysed using the Biocrates targeted mass spectrometry assay. Samples from EPIC and NSHDS (n=1,596) were additionally analysed using Metabolon's untargeted mass spectrometry platform. Samples from matched case-control pairs were assayed in adjacent wells (in random order), and in the same analytical batch. Laboratory personnel were blinded to case-control status of the samples.

An overview of the QC pipeline is shown in **Supplementary Figure S1.** All the QC steps were performed for each cohort separately before pooling the data.

Targeted metabolomics - Biocrates

All samples from EPIC and MCCS were assayed at the International Agency for Research on Cancer (IARC), while samples from NSHDS, HUNT and the Estonian BB were assayed by the Metabolomics Core Facility of the Genome Analysis Center of the Helmholtz Zentrum München [32]. The targeted metabolomics approach was based on LC-ESI-MS/MS and FIA-ESI-MS/MS measurements using the Absolute*IDQ* p180 Kit (BIOCRATES Life Sciences AG, Innsbruck, Austria). The assay allows simultaneous quantification of 188 metabolites using 10 µL plasma or serum. Sample preparation and mass spectrometry measurements were performed as described in **Supplementary Methods**. The median intra- and inter-batch coefficients of variation (CV) were 5.6% and 6.9% respectively (interquartile range = 1.7% and

2.8%, respectively). The lower limits of detection (LODs) were set to three times the values of the zero samples (phosphate buffered saline solution).

Values lower than the LLOQ, or higher than the ULOQ, as well as lower than batch-specific LOD (for compounds semi-quantified: acylcarnitines, glycerophospholipids (GPL), sphingolipids), were imputed with half of the LOD/LLOQ, or the ULOQ. For NSHDS, metabolites with internal standard out of range were left as missing (n=205). Metabolites with less than 100 values above LOD/LLOQ in any individual cohort were excluded from the analyses. In our samples, a total of 164 metabolites were retained for statistical analyses (30 acylcarnitines, 21 amino acids, 10 biogenic amines, 88 GPLs, 14 sphingolipids and the sum of hexoses). In addition to individual metabolites, 22 ratios or sums selected for their capacity to provide detailed insight into a wide range of disorders of the metabolic disease spectrum were computed (listed in **Supplementary Table S4**). Among them, the Fischer's ratio, a clinical indicator of liver metabolism and function, was calculated as the molar ratio of branched chain amino acids (leucine + isoleucine + valine) to aromatic amino acids (phenylalanine + tyrosine). Lower Fischer's ratio values are associated with liver dysfunction.

Untargeted metabolomics – Metabolon

Untargeted metabolomic analyses were performed at Metabolon, Inc. (Durham, North Carolina, USA) on a platform consisting of four independent ultra-high performance liquid chromatography-tandem mass spectrometry (UPLC-MS/MS) methods. Detailed descriptions of the platform and workflow to identify features, including extraction of raw data, peak-identification, and internal quality control (QC) processes can be found in the **Supplementary Methods** and in published work [33-35]. Samples from EPIC and NSHDS were processed as two independent experimental batches. The median intra-batch CV were 5% and 4% for EPIC and NSHDS, respectively while the median inter-batch CV were 11% for both EPIC and NSHDS.

A variety of curation procedures were carried out by Metabolon, Inc. to ensure that a high-quality data set was made available for statistical analysis and data interpretation (**Supplementary Methods**). Each metabolite was rescaled to set the median equal to 1 and missing values imputed with the minimum observed value. Data returned for EPIC comprised a total of 1308 metabolite features, 982 of known identity (named biochemicals) and 326 compounds of unknown structural identity (unnamed biochemicals). Data returned for NSHDS comprised a total of 1302 metabolite features, 979 of known identity (named biochemicals) and 323 compounds of unknown structural identity (unnamed biochemicals). A total of 1275 metabolites were available across the two datasets with the total number of unique metabolites reaching 1335. Metabolites were categorised by Metabolon, Inc. as belonging to one of eight mutually exclusive chemical classes: amino acids and amino acid derivatives (subsequently referred to as 'amino acids'), carbohydrates, cofactors and vitamins, energy metabolites, lipids, nucleotides, peptides, or xenobiotics. An asterisk (*) at the end of the metabolite name indicates the metabolite identity has not been confirmed by comparison with an authentic chemical standard. After the exclusion of metabolites for which less than 100 participants had values recorded (86 and 176 for EPIC and NSHDS, respectively), 1230 metabolite features remained for analysis (1222 and 1126 for EPIC and NSHDS, respectively; 1118 in common).

Statistical analysis

Primary statistical analysis: prospective observational analysis of circulating metabolites and kidney cancer risk

Log-transformed and standardised (z-score) metabolite concentrations were used in all analyses. Crude conditional logistic regressions were performed to estimate the odds ratio (OR) for kidney cancer per one standard deviation (SD) increment in log-transformed

metabolite concentrations, conditioning on the individual case-control sets. To consider multiple comparisons whilst accounting for the correlation between the different metabolites, we estimated the effective number of independent tests performed (ENT) as the number of principal components explaining more than 95% of the variance in our metabolite matrices. Metabolites with p-values equal or below 0.05/ENT in the pooled analyses and equal or below 0.05 in at least two cohorts independently, were deemed robustly associated with kidney cancer risk. For these metabolites, we carried out additional conditional logistic regressions adjusted for BMI, smoking history (smoking status: never, former, current smokers and pack years of smoking), lifetime alcohol consumption (in g/day) and hypertension (ever/never). To avoid comparing different sets of participants due to missingness in risk factor data, we restricted these analyses to study participants with complete risk factor information.

To further characterise the epidemiological properties of the association between metabolites and kidney cancer risk, we also carried out conditional logistic regression stratified by age at blood collection, sex, country, BMI, waist-to-hip ratio, smoking status, alcohol consumption, hypertension and time to diagnosis (number of years between blood draw and diagnosis).

Secondary statistical analysis: Mendelian randomization and profile comparison analyses

We initially investigated pleiotropy among potential SNP instruments for the circulating metabolites associated with kidney cancer risk in prospective analyses (Biocrates and Metabolon) with a view to conducting a two-sample MR analysis for metabolites (as the exposure) and kidney cancer risk (as the outcome). SNP-metabolite associations were extracted from the largest GWASs currently available for circulating metabolites and included summary statistics for 174 Biocrates metabolites [36] (N=ranged from 8,569 to 56,040 for

different metabolites, depending on the platform used in each contributing study) and 913 Metabolon metabolites (N=14,296). Specifically, pleiotropy was assessed by estimating the variance explained in all metabolites by the single nucleotide polymorphisms (SNPs) (i.e. the potential 'instruments') associated with each of our candidate risk metabolites (see Supplementary methods for more details of how instruments were selected). Where the variance explained in other metabolites (i.e. those not associated with risk in the prospective analysis) was similar to that explained in the candidate risk metabolite we inferred low metabolite-specificity for current GWAS results, and thus violation of the MR assumptions necessary to infer potential single exposure causality.

To evaluate the extent to which the metabolomic signature of disease risk could be explained by BMI we first conducted a two-sample MR analysis to provide estimates of the causal relationships between BMI and circulating metabolites (Biocrates and Metabolon). 549 independent SNPs (R^2<0.01) that were robustly associated with BMI at genome-wide significance were selected as instruments from the largest GWAS meta-analysis for BMI from the Genetic Investigation of Anthropometric Traits (GIANT) consortium (n= approximately 700,000[37] see **Supplementary Table S5**). SNP-exposure associations were extracted from the BMI GWAS meta-analysis[37] and SNP-outcome associations were extracted from the metabolite GWAS described above. A BMI effect estimate was generated for each metabolite measured and calculated as an SD unit increase in log-transformed metabolite level per SD increment in BMI. The primary MR analysis was conducted using the inverse-variance weighted (IVW) method[38]. We performed the following sensitivity analyses to attempt to account for potential unbalanced horizontal pleiotropy: 1) MR-Egger regression to test overall directional pleiotropy and provide a valid causal estimate, taking into account the presence of pleiotropy[39] and 2) weighted median,[40] which provides a consistent estimate of causal

effect if at least 50% of the information in the analysis comes from variants that are valid instrumental variables. To account for multiple testing, we used the same *p* value threshold as used in our observational analyses ($p<8.3x10^{-4}$ and $p<1x10^{-4}$ for Biocrates and Metabolon, respectively).

To examine the extent to which kidney cancer-associated metabolites are driven by BMI, we assessed the correlation between the kidney cancer-associated metabolite profile (metabolites associated with kidney cancer risk in the prospective observational analyses) and the BMI-associated metabolite profile (metabolites associated with BMI levels in the MR analyses) using Spearman rank correlation analyses. Effect estimates from both the prospective and MR analyses were divided by the standard error of the estimate before conducting the correlation analyses.

Negative control analyses

The presence or absence of overlap between metabolite profiles flagged by prospective analysis and those derived from BMI MR is only informative in the context of a null, or negative control comparator. To allow this, we repeated the profile comparison analysis described above (with BMI as the exposure) in an analysis in which we used dental disease as a negative control exposure (i.e. an exposure not likely to be a risk factor for kidney cancer) and one that we would therefore expect to deliver a null. This strategy of repeating an experiment under conditions which are expected to deliver a null result has previously been advocated within observational epidemiology [41]. In our analysis of the causal relationship between dental disease and circulating metabolites, 47 independent ($R^2<0.01$) SNPs that were robustly associated at genome-wide significance ($p<5x10^{-8}$) were selected from the largest GWAS for dental disease (n=487,823) (detailed information for instrumental variables for dental disease are presented in **Supplementary Table S6**). SNP-exposure associations

were extracted from the largest dental disease GWAS meta-analysis[42] and SNP-outcome associations were extracted from the metabolite GWAS described above. Effect estimates were calculated as SD unit increase in metabolite levels per logOR increase in dental disease. Methods used in the two sample MR analyses were as described above.

All MR analyses were performed using the TwoSample MR R package version 0.4.13 (http://github.com/MRCIEU/TwoSampleMR) [43].

Results

Population characteristics and metabolites overview

Demographic and baseline characteristics for the 1,305 cases and 1,305 matched controls are presented in **Table 1**. The mean age at diagnosis for cases was 65.6 years (SD=9.79) and cases were diagnosed on average 8 years after blood collection. The majority (58%) of samples were collected after fewer than 6 hours of fasting. Overall, 186 metabolites or ratios/sums of metabolites were measured using the Biocrates assay on 2,610 samples (all cohorts), and 1,230 metabolites were measured using the Metabolon platform on 1,596 samples (EPIC and NSHDS cohorts). Mean concentrations of the 1,416 metabolites by case-control status are shown in **Supplementary Table S7**.

Prospective observational analysis of circulating metabolites and kidney cancer risk

We identified 25 metabolites robustly associated with kidney cancer risk (i.e. metabolites associated with risk after correction for multiple testing in the pooled analysis and nominally significant in at least 2 cohorts; **Figure 2 and Table 2**). Amongst these metabolites, 12 were measured with the Biocrates assay and 13 were measured with the Metabolon platform. Two metabolites - glutamate and 1-linoleoyl-GPC (18:2) (known as lysoPC a C18:2 in Biocrates) - were measured on both platforms and resulted in similar risk association estimates (for

glutamate OR: 1.34 in Biocrates and 1.39 in Metabolon; for 1-linoleoyl-GPC (18:2), OR: 0.77 in Biocrates and 0.76 in Metabolon). Pearson correlations amongst risk-metabolites are displayed in **Supplementary Figure S2**.

We found that increased concentrations of 14 individual GPLs were associated with reduced kidney cancer risk. These included 8 phosphatidylcholines (PC; overall p-values ranging from $6x10^{-4}$ to $3x10^{-8}$), amongst which PC ae C34:3 had the strongest association (OR=0.75, 95% CI: 0.68 to 0.83, p=$2.61x10^{-8}$). Similar associations were identified for the lysophosphatidyl-cholines, lysoPC a C18:1, and lysoPC a C18:2 (labelled as 1-linoleoyl-GPC (18:2) in Metabolon) (p-values between $1.60x10^{-5}$ and $9.65x10^{-7}$). Two plasmalogens were also inversely associated with risk, 1-(1-enyl-palmitoyl-2-oleoyl-GPC (P-16:0/18:1) (p=$1.27x10^{-5}$) and 1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2) (p=$2.79x10^{-5}$), as well as the lysoplasmalogen 1-(1-enyl-palmitoyl)-GPC (P-16:0) (p=$8.32x10^{-6}$).

Amongst 274 metabolites involved in amino acid metabolism, we found four positively associated with kidney cancer risk, including glutamate, formiminoglutamate, hydantoin-5-propionate and the Fischer's ratio (p-values between $1.25x10^{-4}$ and $5.11x10^{-7}$). For example, the relative odds of kidney cancer associated with a standard deviation increment in log-transformed glutamate levels was estimated at 1.39 (95% CI: 1.20 - 1.60) when measured on the Metabolon platform. Another amino acid, cysteine-glutathione disulphide, was inversely associated with risk (OR: 0.77, 95% CI: 0.69 - 0.86, p=$7.42x10^{-6}$). The two peptides gamma-glutamylvaline (p=$1.22x10^{-7}$) and gamma glutamylisoleucine (p=$1.07x10^{-6}$), were positively associated with risk. Finally, we found beta-cryptoxanthin negatively associated with kidney cancer risk (OR: 0.73, 95%CI: 0.65, 0.83, p=$4.83x10^{-7}$) while an unidentified metabolite (X-12096) was positively associated (OR: 1.33, 95%CI: 1.17, 1.51, p=$9.97x10^{-6}$). Adjusting for the

fasting status of the samples (more vs less than 6 hours) did not modify the OR estimates for the identified risk metabolites (**Supplementary Table S8)**.

Associations with risk of kidney cancer for all metabolites analysed are presented in **Supplementary Table S9**.

The influence of kidney cancer risk factors on kidney cancer-associated metabolites

We assessed the extent to which known modifiable risk factors could explain the observed associations by multivariable analyses. For all 25 metabolites found to be associated with risk in the primary analysis, we found that adjustments for BMI partly attenuated the OR estimates for some metabolites, although they all remained at least nominally significant (i.e. p-value below 0.05, **Table 2**). The association most modified by adjustment for BMI was that of glutamate (from 1.34, 95%CI: 1.17-1.53, p=1.62x10^{-5} to 1.24, 95%CI: 1.08-1.42, p=2.46x10^{-3}), followed by PC ae C42:3 and PC aa C42:1 (OR increased by 6% for both metabolites: from 0.82, 95%CI: 0.74-0.92, p=4.17x10^{-4} to 0.87, 95%CI: 0.78-0.98, p=1.75x10^{-2} and 0.83, 95%CI: 0.75-0.93, p=6.27x10^{-4} to 0.88, 95%CI: 0.79-0.99, p=2.59x10^{-2} for PC ae C42:3 and PC aa C42:1, respectively). Conversely, association for PC ae C38:6 was not influenced by adjustment for BMI (OR:0.85, 95%CI: 0.77-0.93, p=5.06x10^{-4} to 0.86, 95%CI: 00.78-0.95, p=1.85x10^{-3}). Results adjusted for all individual risk factors on participants with complete information on these risk factors are shown in **Supplementary Table S10** (N=1,162 and 996 for Biocrates and Metabolon, respectively). Adjustment for smoking and alcohol consumption did not modify any OR by more than 1.5% and 1.2%, respectively, whereas adjusting for hypertension partly attenuated the associations of lysoPC a C18:1 and lysoPC a C18:2, albeit to a lesser extent than BMI (5% change for both). In fully adjusted models, risk associations remained nominally significant (p-value below 0.05) for 10 out of 25 metabolites with all effect estimates in the

same direction as in the primary analysis, although, due to missing data for some risk factors, this analysis included only 581 and 498 case-control pairs for Biocrates and Metabolon, respectively.

In stratified risk analyses by time to diagnosis (**Supplementary Figures S3 to S27)**, several metabolites appeared to display a stronger risk-association closer to diagnosis, including 1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2) (heterogeneity p=0.02) (**Supplementary Figure S15**) and the metabolite of unknown structural identity X-12096 (heterogeneity p=0.02) that was measured on the Metabolon platform (**Supplementary Figure S27**). The lysophosphatidyl-choline lysoPC a C18:2, as measured by Biocrates, showed a stronger association when alcohol consumption was above the median compared to lower (heterogeneity p=0.03) (**Supplementary Figure S6**); this pattern was evident for the same metabolite measured in Metabolon but was not statistically significant (heterogeneity p=0.3) (**Supplementary Figure S18**).

Two sample Mendelian randomization and profile comparison analyses

We identified genetic instruments for 17 of the 25 risk metabolites but observed substantial pleiotropy for the instruments defined for 16 of the 17 instrumented metabolites. The total variance explained from a risk-metabolite's instruments was typically similar across classes of metabolite (lipids and 1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2), for example), and far from specific to the given risk-metabolite being instrumented. Further, the variance explained was often higher for an alternative metabolite compared to the risk-metabolite (**see Supplementary Figures S28 to S44**). Following these observations, we chose not to carry out a formal MR analysis of the relation between individual metabolites and kidney cancer risk because the profound pleiotropy across metabolites clearly violates the MR assumptions.

Rather, to complement the risk analyses, and to gain further understanding of how BMI – the leading modifiable risk factor of kidney cancer – might explain our findings, we conducted a two-sample MR analysis to evaluate the extent to which the measured metabolites are driven by differences in BMI. Using the IVW method, 60 metabolites (22 Biocrates and 38 Metabolon) were associated with BMI. In an MR framework, there was consistent evidence between both platforms that BMI was associated with decreased concentrations of many GPLs and increased concentrations of several amino acids and nucleotides, as well as acylcarnitines, sphingomyelins and several metabolites of unknown identity (**Figure S45**). Estimates from MR-Egger and weighted median analyses were consistent with the IVW estimates (**Supplementary Table S11 and S12**).

When comparing the metabolic profile of kidney cancer (metabolites associated with kidney cancer risk in the prospective analyses) and BMI (metabolites associated with BMI levels in the MR analyses), we observed moderate correlation between the BMI-driven metabolite profile and metabolite profile associated with kidney cancer risk (**Figure 3)** (r=0.53, p=2.2x10^{-6} for Biocrates metabolites and r=0.36, p=2.2x10^{-6} for Metabolon metabolites). Specifically, elevated BMI appeared to decrease levels of several GPLs that were also found inversely associated with kidney cancer risk, including 1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2)*, 1-linoleoyl-GPC (18:2) (lysoPC a C18:2), lysoPC a C18:1 and PC ae C34:3. For instance, one SD increment in BMI was associated with a 0.17 SD decrease in 1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2) levels ([$ß_{BMI}$], p=3.4x10^{-5}). We also found that BMI was associated with increased levels of glutamate ($ß_{BMI}$: 0.12, p=1.5x10^{-3}), which was positively associated with kidney cancer risk**.** Several metabolites associated with kidney cancer risk in our prospective analysis did not appear to be strongly influenced by BMI, but we note that for all but two metabolites (PC ae 32:2 and PC ae 42:3), estimates were

directionally concordant (i.e. positively correlated) but with the effect size estimates from the BMI MR being closer to the null than those seen in the observational analysis. Conversely, some of the metabolites that were most strongly affected by BMI (e.g. phenylalanine and valine), were not associated with kidney cancer risk.

Negative control analyses

There was little evidence that genetic predisposition to dental disease influenced circulating metabolite levels with no metabolites reaching our pre-determined threshold for a statistically significant association (Supplementary **Table S13 and S14**). We observed low correlation between the dental disease-metabolite estimates from MR analyses and the kidney cancer-metabolite estimates from the prospective analysis for both Biocrates metabolites (r=0.15, *p*=0.06) and Metabolon (r=0.12, *p*=0.002) (Supplementary **Figure S46).** None of the 25 metabolites that were associated with kidney cancer risk in prospective analyses were associated with dental disease from the MR analyses (Supplementary **Figure S46**). These findings suggest that when the profile comparison analysis is conducted using a hypothetically unrelated exposure (dental disease) we see no meaningful relationship between metabolite associations from the prospective analysis and the MR.

Discussion

This study describes the relationship between the pre-diagnostic blood-metabolome and risk of developing kidney cancer based on data from five longitudinal population cohorts. This is the first comprehensive metabolomics analysis of incident kidney cancer to be conducted using a prospective design, and as such, complements existing work characterising the metabolic profile (in tissue and biofluids) of the disease itself [16-26]. We investigated 1,416

metabolites in relation to the occurrence of kidney cancer using two complementary analytical methods and observed 25 metabolites to be robustly associated with risk. These metabolites included 14 GPLs inversely associated with risk, five amino acids positively associated, and one inversely associated with risk, as well as risk associations for a carotenoid, two peptides, a nucleotide and an unidentified feature. Results of an MR analysis designed to evaluate the extent to which BMI influences the key risk-associated metabolites, suggest that differences in BMI may be responsible for part of the metabolite profile associated with the development of kidney cancer.

The majority of metabolites found to be associated with kidney cancer risk in this study can be classified as glycerophospholipids (GPLs). GPLs are the main component of cell membranes and are essential for maintaining cellular structure and for regulating cell signalling. The circulating metabolite associations we see here pre-diagnosis appear to intersect with the known cellular metabolic programming observed within kidney tumour tissue. For example, it has been proposed that clear cell RCC cells use exogenous lipids for membrane formation and cell signalling [44]. The relationship between lipid metabolites and prospective kidney cancer risk reported in our study could, theoretically, be capturing increased uptake of lipid metabolites by preclinical kidney carcinogenesis.

GPLs can be broadly classified into two types based on their biochemical structure – diacyl (aa) or acyl-alkyl (ae) – and can be further characterised according to their lipid side-chain composition, specifically the number of carbons and their degree of (un)saturation (number of double bonds). The association of a subset of long chain unsaturated (mainly acyl-alkyl) phosphatidylcholines (PCs), lysophophatidylcholines (LPCs) and plasmalogens with reduced kidney cancer risk is consistent with some limited existing literature. Specifically, lower levels of total phosphatidylcholine/choline have been reported in the serum of diagnosed kidney

cancer patients compared to control participants,[17] and numerous studies have found decreased LPCs in both tumour and normal kidney tissues,[27,45,46] as well as in the circulation of kidney cancer patients [18,47]. The mechanisms underpinning these associations are not well-understood, but some of these molecules (e.g. plasmalogens) have been proposed as antioxidants[48]. Low levels of plasmalogens in cancer patients have been proposed as a potential mechanism by which increased oxidative stress could drive cancer progression [49].

We assessed the extent to which known risk factors could explain the observed metabolite associations and observed that adjusting for BMI – the main modifiable risk factor for kidney cancer – partially attenuated (less than 9% change in OR) the risk association for some specific metabolites. To further understand the relation with BMI for the kidney cancer risk-associated metabolites, we estimated the causal influence of BMI on metabolite levels using Mendelian randomization. This analysis clearly demonstrated that some – but not all – metabolites inversely associated with kidney cancer risk are also decreased by elevated BMI (e.g. several GPLs), whereas other metabolites positively associated with risk (e.g. glutamate), are also increased by elevated BMI. The association of long chain unsaturated (mainly acyl-alkyl) GPLs with both lower risk of RCC and lower BMI is consistent with extensive literature linking lower levels of these and similar molecules to a range of common diseases that include a metabolic component such as obesity and hypertension,[50-52] type 2 diabetes,[53] type 1 diabetes development[54] and non-alcoholic fatty liver disease [55].

Glutamate was found to be positively associated with both kidney cancer risk and BMI and was also the metabolite for which adjusting for BMI resulted in the greatest attenuation in its OR estimate. Glutamate and glutamine are both found to be increased in kidney tumour tissue [44]. This observation provides further evidence of overlap between metabolites

relevant to disease development and those whose levels are perturbed in the disease state [10,56]. Consistent with our findings, glutamate has previously been shown to be increased in visceral obesity[57,58] and glutamine-derived glutamate has been linked to tumour cell metabolism[59] with renal cell carcinoma being no exception [60]. α-Ketoglutarate, generated from glutamine-derived glutamate, enters the tricarboxylic acid (TCA) cycle providing both energy and biosynthetic intermediates [61]. A large intracellular glutamate pool is also important for nonessential amino acid synthesis in addition to cellular redox regulation [61]. Two previous nuclear magnetic resonance (NMR)-based studies found lower levels of glutamine in serum of kidney cancer cases taken at diagnosis compared to controls [16,17]. Whilst we did not identify a robust association of glutamine in our study, the point estimate was consistent with a weak inverse association with risk of kidney cancer.

A final overarching observation was that in comparison with previously published prospective metabolomics analyses on other cancer sites,[62-64] the sheer number of metabolites found to be associated with risk in the current study suggests that the blood metabolome is particularly important in the aetiology of kidney cancer.

Strengths, limitations and prospects for future studies

The chief strength of our study was the design of the primary risk analysis wherein control subjects were individually matched to incident kidney cancer cases with pre-diagnostic blood samples from five independent population cohorts, a design that minimized differential bias and allowed for identification of novel and robust risk metabolites of kidney cancer. The use of two complementary metabolomics platforms also increased the overall coverage of the metabolome. The well-characterized cohorts offered the opportunity to carefully assess the influence of known kidney cancer risk factors (i.e. potential confounders) on identified risk-

associated metabolites, as well as the robustness of their risk associations across the independent cohort studies. Well-designed prospective studies can provide compelling evidence in favour of a role of molecular risk factors in cancer aetiology, but residual confounding from imperfectly measured risk factors may still bias the association estimates. We therefore complemented the main risk analysis with a genetic analysis to assess the influence of BMI on the identified risk metabolites. We believe that this independent analysis provided important independent evidence when interpreting the relation between the identified risk metabolites and kidney cancer risk in the context of BMI – the principal risk factor of kidney cancer.

Limitations of our study include the presence of measurement error in the (semi-) quantification of metabolites. However, by using well-established platforms with built-in validation procedures along with randomisation schemes to ensure any batch variation was orthogonal to the outcome of interest (in this case kidney cancer case status), we can be confident there was no systematic bias in our estimates as a result of measurement error. In addition, the consistency in estimates we see for metabolites that appear on both platforms provides increased confidence in our results, but we note that statistical power to identify risk metabolites exclusive to the Metabolon platform was lower than for metabolites exclusive to the Biocrates platform due to the lower sample size. In this study, we focused on those metabolites that demonstrated consistency in risk associations across the five participating cohorts. Whilst this approach ensured the robustness of the estimates, any risk marker present in specific populations would not be highlighted. Although we only measured metabolite levels at a single time point, we do not believe this represents a major limitation as the majority of measured metabolites have a high within person stability over time (stable over 4 months to 2 years) [65-67]. Another limitation of our study is the lack of detailed data

on body composition. It is possible that some individual risk markers may reflect a certain adiposity distribution that is specifically strongly associated with kidney cancer risk. Whilst the current literature on kidney cancer aetiology does not highlight any specific aspect of obesity as being particularly important in kidney cancer aetiology, evaluating the identified risk markers in relation to detailed body composition (e.g. using DEXA scan data) represents an appealing future focus of our kidney cancer research. The remaining limitations relate to the generalisability of our findings. Given evidence for specific metabolic alterations by kidney cancer histotype [10], it is possible that kidney cancer subtypes have different dependencies on circulating metabolites. In this case, findings from this study are likely most relevant to the major histological subtype – clear cell RCC – which made up 71% of kidney cancer cases. Furthermore, our study does not inform on the extent to which the identified risk markers translate to populations of non-European descent. Addressing these limitations should constitute an important focus for future studies addressing the role of the blood metabolome in the aetiology of kidney cancer.

Whilst the results of our prospective risk analysis are consistent with circulating metabolites playing an important role in kidney cancer aetiology, it is appealing to complement such observational analyses with MR studies to further inform causal inference. However, we chose not to carry out an MR analysis on kidney cancer risk for individual metabolites for a number of reasons related to characteristics specific to circulating metabolites. Firstly, owing to high correlational structure of many metabolites, few SNPs have been found associated with specific metabolites, leading to pleiotropic instruments for most metabolites [36]. Secondly, there is a high degree of pleiotropy for metabolite-associated SNPs with modifiable risk factors and other disease endpoints. That few metabolites have a sufficient number of instruments is particularly problematic as applying statistical methods aiming to correct for

these biases is not possible (e.g. MR-Egger and MR-PRESSO), nor is the use of techniques designed to evaluate the effect of multiple correlated exposures (e.g. multivariable MR [68]). Whilst the genetic architecture of blood metabolites is complicated for the reasons outlined above, there are hundreds of independent SNPs robustly associated with BMI [37] and this gave us greater confidence in the application of this analysis [69]. Better characterizing of the genetic architecture of circulating metabolites together with methodological advancements may allow for more robust causal inference in future metabolomics studies.

Conclusions

This study points to a particularly important role of the blood metabolome in kidney cancer aetiology, specifically by identifying positive risk associations for several amino acids, as well as negative risk associations with multiple lipids, including PCs, LPCs and plasmalogens. Downstream analyses indicated that some – but not all – risk metabolites are influenced by BMI, which partly explains their associations with kidney cancer risk, whereas the risk associations for other metabolites could not be explained by known risk factors. These results provide important insight into the metabolic pathways underpinning the central role of obesity in kidney cancer aetiology, and clues to novel pathways involved in kidney cancer aetiology.

Acknowledgments

We thank Julia Scarpa, Werner Römisch-Margl, and Silke Becker for supporting the metabolomics measurements performed at the Helmholtz Zentrum München, Genome Analysis Center, Metabolomics Core Facility. We thank Audrey Brunet-Manquat for supporting the metabolomics measurements performed in the Biomarkers group at IARC. The EPIC-Norfolk study (https://doi.org/10.22025/2019.10.105.00004) has received funding from the Medical Research Council (MR/N003284/1 MC-UU_12015/1 and MC_UU_00006/1) and Cancer Research UK (C864/A14136). The genetics work in the EPIC-Norfolk study was funded by the Medical Research Council (MC_PC_13048). Metabolite measurements in the EPIC-Norfolk study were supported by the MRC Cambridge Initiative in Metabolic Science (MR/L00002/1) and the Innovative Medicines Initiative Joint Undertaking under EMIF grant agreement no. 115372. We are grateful to all the participants who have been part of the project and to the many members of the study teams at the University of Cambridge who have enabled this research. Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping were co-funded by the National Institute for Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk) and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) [The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care]. The academic coordinating centre for INTERVAL was supported by core funding from the: NIHR Blood and Transplant Research Unit in Donor Health and Genomics (NIHR BTRU-2014-10024), UK Medical Research Council (MR/L003120/1), British Heart Foundation (SP/09/002; RG/13/13/30194; RG/18/13/33946) and NIHR Cambridge BRC (BRC-1215-

20014). A complete list of the investigators and contributors to the INTERVAL trial is provided in reference 19 of Supplementary methods (*Di Angelantonio, et al.)*. The academic coordinating centre would like to thank blood donor centre staff and blood donors for participating in the INTERVAL trial.

Author contributions

Conceived and designed the experiments: NJT, MJ, LJC, FG, VYT, KSB

Performed the experiments: DA, AG, CP, MW, JA

Analyzed the data: FG, VYT, IDS

Contributed materials/analysis tools: SR, AS, CL, IDS, AB, PS, BL, GS, RM, GG, AL, TLL, EPS, TH, AM, JR, SS, PAE, JAS, RV, MMB, THN, CCD, DM, ER

Wrote the paper: all

References

1. Ferlay J EM, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F. Global Cancer Observatory: Cancer Today. 2018 [Accessed on 13 November 2020]. Available from: https://gco.iarc.fr/today.
2. Scelo G, Larose TL. Epidemiology and Risk Factors for Kidney Cancer. J Clin Oncol. 2018:JCO2018791905. Epub 2018/10/30. doi: 10.1200/JCO.2018.79.1905. PMID: 30372394.
3. Haggstrom C, Rapp K, Stocks T, Manjer J, Bjorge T, Ulmer H, et al. Metabolic factors associated with risk of renal cell carcinoma. PLoS One. 2013;8(2):e57475. doi: 10.1371/journal.pone.0057475. PMID: 23468995.
4. Stocks T, Bjorge T, Ulmer H, Manjer J, Haggstrom C, Nagel G, et al. Metabolic risk score and cancer risk: pooled analysis of seven cohorts. Int J Epidemiol. 2015;44(4):1353-63. Epub 2015/02/06. doi: 10.1093/ije/dyv001. PMID: 25652574.
5. Laaksonen MA, MacInnis RJ, Canfell K, Giles GG, Hull P, Shaw JE, et al. The future burden of kidney and bladder cancers preventable by behavior modification in Australia: A pooled cohort study. Int J Cancer. 2020;146(3):874-83. Epub 2019/05/21. doi: 10.1002/ijc.32420. PMID: 31107541.
6. Johansson M, Carreras-Torres R, Scelo G, Purdue MP, Mariosa D, Muller DC, et al. The influence of obesity-related factors in the etiology of renal cell carcinoma-A mendelian randomization study. PLoS Med. 2019;16(1):e1002724. Epub 2019/01/04. doi: 10.1371/journal.pmed.1002724. PMID: 30605491.
7. Wang F, Xu Y. Body mass index and risk of renal cell cancer: a dose-response meta-analysis of published cohort studies. Int J Cancer. 2014;135(7):1673-86. Epub 2014/03/13. doi: 10.1002/ijc.28813. PMID: 24615287.
8. Brown KF, Rumgay H, Dunlop C, Ryan M, Quartly F, Cox A, et al. The fraction of cancer attributable to modifiable risk factors in England, Wales, Scotland, Northern Ireland, and the United Kingdom in 2015. Br J Cancer. 2018;118(8):1130-41. Epub 2018/03/24. doi: 10.1038/s41416-018-0029-6. PMID: 29567982.
9. Linehan WM, Srinivasan R, Schmidt LS. The genetic basis of kidney cancer: a metabolic disease. Nat Rev Urol. 2010;7(5):277-85. Epub 2010/05/08. doi: 10.1038/nrurol.2010.47. PMID: 20448661.
10. DiNatale RG, Sanchez A, Hakimi AA, Reznik E. Metabolomics informs common patterns of molecular dysfunction across histologies of renal cell carcinoma. Urol Oncol. 2020;38(10):755-62. Epub 2019/06/04. doi: 10.1016/j.urolonc.2019.04.028. PMID: 31155438.
11. Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. Nat Rev Genet. 2014;15(1):34-48. Epub 2013/12/04. doi: 10.1038/nrg3575. PMID: 24296534.
12. Homuth G, Teumer A, Volker U, Nauck M. A description of large-scale metabolomics studies: increasing value by combining metabolomics with genome-wide SNP genotyping and transcriptional profiling. J Endocrinol. 2012;215(1):17-28. Epub 2012/07/12. doi: 10.1530/JOE-12-0144. PMID: 22782382.
13. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. Nat Rev Mol Cell Biol. 2016;17(7):451-9. Epub 2016/03/17. doi: 10.1038/nrm.2016.25. PMID: 26979502.
14. Sullivan LB, Gui DY, Vander Heiden MG. Altered metabolite levels in cancer: implications for tumour biology and cancer therapy. Nat Rev Cancer. 2016;16(11):680-93. Epub 2016/10/25. doi: 10.1038/nrc.2016.85. PMID: 27658530.

15. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. Nat Rev Drug Discov. 2016;15(7):473-84. Epub 2016/03/12. doi: 10.1038/nrd.2016.32. PMID: 26965202.
16. Zira AN, Theocharis SE, Mitropoulos D, Migdalis V, Mikros E. (1)H NMR metabonomic analysis in renal cell carcinoma: a possible diagnostic tool. Journal of proteome research. 2010;9(8):4038-44. Epub 2010/06/10. doi: 10.1021/pr100226m. PMID: 20527959.
17. Gao H, Dong B, Liu X, Xuan H, Huang Y, Lin D. Metabonomic profiling of renal cell carcinoma: high-resolution proton nuclear magnetic resonance spectroscopy of human serum with multivariate data analysis. Anal Chim Acta. 2008;624(2):269-77. Epub 2008/08/19. doi: 10.1016/j.aca.2008.06.051. PMID: 18706333.
18. Lin L, Huang Z, Gao Y, Yan X, Xing J, Hang W. LC-MS based serum metabonomic analysis for renal cell carcinoma diagnosis, staging, and biomarker discovery. J Proteome Res. 2011;10(3):1396-405. Epub 2010/12/29. doi: 10.1021/pr101161u. PMID: 21186845.
19. Lin L, Yu Q, Yan X, Hang W, Zheng J, Xing J, et al. Direct infusion mass spectrometry or liquid chromatography mass spectrometry for human metabonomics? A serum metabonomic study of kidney cancer. Analyst. 2010;135(11):2970-8. Epub 2010/09/22. doi: 10.1039/c0an00265h. PMID: 20856980.
20. Ganti S, Taylor SL, Kim K, Hoppel CL, Guo L, Yang J, et al. Urinary acylcarnitines are altered in human kidney cancer. Int J Cancer. 2012;130(12):2791-800. Epub 2011/07/07. doi: 10.1002/ijc.26274. PMID: 21732340.
21. Ganti S, Weiss RH. Urine metabolomics for kidney cancer detection and biomarker discovery. Urol Oncol. 2011;29(5):551-7. Epub 2011/09/21. doi: 10.1016/j.urolonc.2011.05.013. PMID: 21930086.
22. Kind T, Tolstikov V, Fiehn O, Weiss RH. A comprehensive urinary metabolomic approach for identifying kidney cancerr. Anal Biochem. 2007;363(2):185-95. Epub 2007/02/24. doi: 10.1016/j.ab.2007.01.028. PMID: 17316536.
23. Perroud B, Lee J, Valkova N, Dhirapong A, Lin PY, Fiehn O, et al. Pathway analysis of kidney cancer using proteomics and metabolic profiling. Mol Cancer. 2006;5:64. Epub 2006/11/25. doi: 10.1186/1476-4598-5-64. PMID: 17123452.
24. Niziol J, Bonifay V, Ossolinski K, Ossolinski T, Ossolinska A, Sunner J, et al. Metabolomic study of human tissue and urine in clear cell renal carcinoma by LC-HRMS and PLS-DA. Anal Bioanal Chem. 2018;410(16):3859-69. Epub 2018/04/17. doi: 10.1007/s00216-018-1059-x. PMID: 29658093.
25. Wettersten HI, Hakimi AA, Morin D, Bianchi C, Johnstone ME, Donohoe DR, et al. Grade-Dependent Metabolic Reprogramming in Kidney Cancer Revealed by Combined Proteomics and Metabolomics Analysis. Cancer Res. 2015;75(12):2541-52. doi: 10.1158/0008-5472.CAN-14-1703. PMID: 25952651.
26. Saito K, Arai E, Maekawa K, Ishikawa M, Fujimoto H, Taguchi R, et al. Lipidomic Signatures and Associated Transcriptomic Profiles of Clear Cell Renal Cell Carcinoma. Sci Rep. 2016;6:28932. Epub 2016/07/01. doi: 10.1038/srep28932. PMID: 27357243.
27. Schaeffeler E, Buttner F, Reustle A, Klumpp V, Winter S, Rausch S, et al. Metabolic and Lipidomic Reprogramming in Renal Cell Carcinoma Subtypes Reflects Regions of Tumor Origin. Eur Urol Focus. 2019;5(4):608-18. Epub 2018/02/18. doi: 10.1016/j.euf.2018.01.016. PMID: 29452772.
28. Yet I, Menni C, Shin SY, Mangino M, Soranzo N, Adamski J, et al. Genetic Influences on Metabolite Levels: A Comparison across Metabolomic Platforms. PLoS One.

2016;11(4):e0153672. Epub 2016/04/14. doi: 10.1371/journal.pone.0153672. PMID: 27073872.
29. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32(1):1-22. Epub 2003/04/12. doi: 10.1093/ije/dyg070. PMID: 12689998.
30. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296. Epub 2007/10/19. doi: 10.1371/journal.pmed.0040296. PMID: 17941714.
31. Davey Smith G, Davies NM, Dimou N, Egger M, Gallo V, Golub R, et al. STROBE-MR: Guidelines for strengthening the reporting of Mendelian randomization studies. . PeerJ Preprints 2019;7:e27857v1. doi: https://doi.org/10.7287/peerj.preprints.27857v1.
32. Zukunft S, Sorgenfrei M, Prehn C, Moller G, Adamski J. Targeted Metabolomics of Dried Blood Spot Extracts. Chromatographia. 2013;76(19-20):1295-305. doi: 10.1007/s10337-013-2429-3. PMID: WOS:000324825100011.
33. Evans AM, DeHaven CD, Barrett T, Mitchell M, Milgram E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. Anal Chem. 2009;81(16):6656-67. Epub 2009/07/25. doi: 10.1021/ac901536h. PMID: 19624122.
34. Dehaven CD, Evans AM, Dai H, Lawton KA. Organization of GC/MS and LC/MS metabolomics data into chemical libraries. J Cheminform. 2010;2(1):9. Epub 2010/10/20. doi: 10.1186/1758-2946-2-9. PMID: 20955607.
35. Zierer J, Jackson MA, Kastenmuller G, Mangino M, Long T, Telenti A, et al. The fecal metabolome as a functional readout of the gut microbiome. Nat Genet. 2018;50(6):790-5. Epub 2018/05/29. doi: 10.1038/s41588-018-0135-7. PMID: 29808030.
36. Lotta LA, Pietzner M, Stewart ID, Wittemans LBL, Li C, Bonelli R, et al. A cross-platform approach identifies genetic regulators of human metabolism and health. Nat Genet. 2021;53(1):54-64. Epub 2021/01/09. doi: 10.1038/s41588-020-00751-5. PMID: 33414548.
37. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. Hum Mol Genet. 2018;27(20):3641-9. Epub 2018/08/21. doi: 10.1093/hmg/ddy271. PMID: 30124842.
38. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol. 2013;37(7):658-65. Epub 2013/10/12. doi: 10.1002/gepi.21758. PMID: 24114802.
39. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015;44(2):512-25. Epub 2015/06/08. doi: 10.1093/ije/dyv080. PMID: 26050253.
40. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet Epidemiol. 2016;40(4):304-14. Epub 2016/04/12. doi: 10.1002/gepi.21965. PMID: 27061298.
41. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010;21(3):383-8. Epub 2010/03/26. doi: 10.1097/EDE.0b013e3181d61eeb. PMID: 20335814.

42. Shungin D, Haworth S, Divaris K, Agler CS, Kamatani Y, Keun Lee M, et al. Genome-wide analysis of dental caries and periodontitis combining clinical and self-reported data. Nat Commun. 2019;10(1):2773. Epub 2019/06/27. doi: 10.1038/s41467-019-10630-1. PMID: 31235808.
43. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7. Epub 2018/05/31. doi: 10.7554/eLife.34408. PMID: 29846171.
44. Lucarelli G, Loizzo D, Franzin R, Battaglia S, Ferro M, Cantiello F, et al. Metabolomic insights into pathophysiological mechanisms and biomarker discovery in clear cell renal cell carcinoma. Expert Rev Mol Diagn. 2019;19(5):397-407. Epub 2019/04/16. doi: 10.1080/14737159.2019.1607729. PMID: 30983433.
45. Cifkova E, Holcapek M, Lisa M, Vrana D, Gatek J, Melichar B. Determination of lipidomic differences between human breast cancer and surrounding normal tissues using HILIC-HPLC/ESI-MS and multivariate data analysis. Anal Bioanal Chem. 2015;407(3):991-1002. Epub 2014/10/30. doi: 10.1007/s00216-014-8272-z. PMID: 25352274.
46. Du Y, Wang Q, Zhang X, Wang X, Qin C, Sheng Z, et al. Lysophosphatidylcholine acyltransferase 1 upregulation and concomitant phospholipid alterations in clear cell renal cell carcinoma. J Exp Clin Cancer Res. 2017;36(1):66. Epub 2017/05/13. doi: 10.1186/s13046-017-0525-1. PMID: 28494778.
47. Sullentrop F, Moka D, Neubauer S, Haupt G, Engelmann U, Hahn J, et al. 31P NMR spectroscopy of blood plasma: determination and quantification of phospholipid classes in patients with renal cell carcinoma. NMR Biomed. 2002;15(1):60-8. Epub 2002/02/13. doi: 10.1002/nbm.758. PMID: 11840554.
48. Engelmann B. Plasmalogens: targets for oxidants and major lipophilic antioxidants. Biochem Soc Trans. 2004;32(Pt 1):147-50. Epub 2004/01/30. doi: 10.1042/bst0320147. PMID: 14748736.
49. Messias MCF, Mecatti GC, Priolli DG, de Oliveira Carvalho P. Plasmalogen lipids: functional mechanism and their involvement in gastrointestinal cancer. Lipids Health Dis. 2018;17(1):41. Epub 2018/03/09. doi: 10.1186/s12944-018-0685-9. PMID: 29514688.
50. Graessler J, Schwudke D, Schwarz PE, Herzog R, Shevchenko A, Bornstein SR. Top-down lipidomics reveals ether lipid deficiency in blood plasma of hypertensive patients. PLoS One. 2009;4(7):e6261. Epub 2009/07/16. doi: 10.1371/journal.pone.0006261. PMID: 19603071.
51. Pietilainen KH, Sysi-Aho M, Rissanen A, Seppanen-Laakso T, Yki-Jarvinen H, Kaprio J, et al. Acquired obesity is associated with changes in the serum lipidomic profile independent of genetic effects--a monozygotic twin study. PLoS One. 2007;2(2):e218. Epub 2007/02/15. doi: 10.1371/journal.pone.0000218. PMID: 17299598.
52. Wahl S, Yu Z, Kleber M, Singmann P, Holzapfel C, He Y, et al. Childhood obesity is associated with changes in the serum metabolite profile. Obes Facts. 2012;5(5):660-70. Epub 2012/10/31. doi: 10.1159/000343204. PMID: 23108202.
53. Razquin C, Toledo E, Clish CB, Ruiz-Canela M, Dennis C, Corella D, et al. Plasma Lipidomic Profiling and Risk of Type 2 Diabetes in the PREDIMED Trial. Diabetes Care. 2018;41(12):2617-24. Epub 2018/10/18. doi: 10.2337/dc18-0840. PMID: 30327364.
54. Oresic M, Simell S, Sysi-Aho M, Nanto-Salonen K, Seppanen-Laakso T, Parikka V, et al. Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. J Exp Med. 2008;205(13):2975-84. Epub 2008/12/17. doi: 10.1084/jem.20081800. PMID: 19075291.

55. Oresic M, Hyotylainen T, Kotronen A, Gopalacharyulu P, Nygren H, Arola J, et al. Prediction of non-alcoholic fatty-liver disease and liver fat content by serum molecular lipids. Diabetologia. 2013;56(10):2266-74. Epub 2013/07/05. doi: 10.1007/s00125-013-2981-2. PMID: 23824212.
56. Lucarelli G, Rutigliano M, Sallustio F, Ribatti D, Giglio A, Lepore Signorile M, et al. Integrated multi-omics characterization reveals a distinctive metabolic signature and the role of NDUFA4L2 in promoting angiogenesis, chemoresistance, and mitochondrial dysfunction in clear cell renal cell carcinoma. Aging (Albany NY). 2018;10(12):3957-85. Epub 2018/12/13. doi: 10.18632/aging.101685. PMID: 30538212.
57. Takashina C, Tsujino I, Watanabe T, Sakaue S, Ikeda D, Yamada A, et al. Associations among the plasma amino acid profile, obesity, and glucose metabolism in Japanese adults with normal glucose tolerance. Nutr Metab (Lond). 2016;13:5. Epub 2016/01/21. doi: 10.1186/s12986-015-0059-5. PMID: 26788116.
58. Boulet MM, Chevrier G, Grenier-Larouche T, Pelletier M, Nadeau M, Scarpa J, et al. Alterations of plasma metabolite profiles related to adipose tissue distribution and cardiometabolic risk. Am J Physiol Endocrinol Metab. 2015;309(8):E736-46. Epub 2015/08/27. doi: 10.1152/ajpendo.00231.2015. PMID: 26306599.
59. Li T, Le A. Glutamine Metabolism in Cancer. Heterogeneity of Cancer Metabolism. 2018;1063:13-32. doi: 10.1007/978-3-319-77736-8_2. PMID: WOS:000451318000003.
60. Shroff EH, Eberlin LS, Dang VM, Gouw AM, Gabay M, Adam SJ, et al. MYC oncogene overexpression drives renal cell carcinoma in a mouse model through glutamine metabolism. Proc Natl Acad Sci U S A. 2015;112(21):6539-44. Epub 2015/05/13. doi: 10.1073/pnas.1507228112. PMID: 25964345.
61. DeBerardinis RJ, Chandel NS. Fundamentals of cancer metabolism. Sci Adv. 2016;2(5):e1600200. Epub 2016/07/08. doi: 10.1126/sciadv.1600200. PMID: 27386546.
62. His M, Viallon V, Dossus L, Gicquiau A, Achaintre D, Scalbert A, et al. Prospective analysis of circulating metabolites and breast cancer in EPIC. BMC Med. 2019;17(1):178. Epub 2019/09/25. doi: 10.1186/s12916-019-1408-4. PMID: 31547832.
63. Shu X, Xiang YB, Rothman N, Yu D, Li HL, Yang G, et al. Prospective study of blood metabolites associated with colorectal cancer risk. Int J Cancer. 2018;143(3):527-34. Epub 2018/02/27. doi: 10.1002/ijc.31341. PMID: 29479691.
64. Shu X, Zheng W, Yu D, Li HL, Lan Q, Yang G, et al. Prospective metabolomics study identifies potential novel blood metabolites associated with pancreatic cancer risk. Int J Cancer. 2018;143(9):2161-7. Epub 2018/05/03. doi: 10.1002/ijc.31574. PMID: 29717485.
65. Carayol M, Licaj I, Achaintre D, Sacerdote C, Vineis P, Key TJ, et al. Reliability of Serum Metabolites over a Two-Year Period: A Targeted Metabolomic Approach in Fasting and Non-Fasting Samples from EPIC. PLoS One. 2015;10(8):e0135437. Epub 2015/08/15. doi: 10.1371/journal.pone.0135437. PMID: 26274920.
66. Floegel A, Drogan D, Wang-Sattler R, Prehn C, Illig T, Adamski J, et al. Reliability of serum metabolite concentrations over a 4-month period using a targeted metabolomic approach. PLoS One. 2011;6(6):e21103. Epub 2011/06/24. doi: 10.1371/journal.pone.0021103. PMID: 21698256.
67. Townsend MK, Clish CB, Kraft P, Wu C, Souza AL, Deik AA, et al. Reproducibility of metabolomic profiles among men and women in 2 large cohort studies. Clin Chem. 2013;59(11):1657-67. Epub 2013/07/31. doi: 10.1373/clinchem.2012.199133. PMID: 23897902.

68. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. Int J Epidemiol. 2018. Epub 2018/12/12. doi: 10.1093/ije/dyy262. PMID: 30535378.
69. Burgess S, Davies NM, Thompson SG. Bias due to participant overlap in two-sample Mendelian randomization. Genet Epidemiol. 2016;40(7):597-608. Epub 2016/10/19. doi: 10.1002/gepi.21998. PMID: 27625185.

Table 1. Population characteristics of the 2,610 kidney cancer cases and controls from 5 independent cohorts with pre-diagnostic blood samples included in our analyses.

	Cases mean (SD) or N (%)	Controls mean (SD) or N (%)
Total	1305	1305
Age at blood collection (years)	57.6 (10.1)	57.6 (10.1)
Length of follow-up from blood collection (years)	7.95 (4.98)	-
Histology		
Clear Cell	931 (71.3)	-
Other	282 (21.6)	-
Unknown	92 (7.1)	-
Sex		
Male	725 (55.6)	725 (55.6)
Female	580 (44.4)	580 (44.4)
Cohort		
EPIC	634 (48.6)	634 (48.6)
Estonian Biobank	115 (8.8)	115 (8.8)
HUNT	254 (19.5)	254 (19.5)
MCCS	139 (10.6)	139 (10.6)
NSHDS	163 (12.5)	163 (12.5)
Education		
None	43 (3.3)	52 (4)
Primary School	468 (35.9)	456 (34.9)
Technical School	233 (17.9)	222 (17)
Secondary School	239 (18.3)	236 (18.1)
University	216 (16.6)	242 (18.5)
Unknown	106 (8.1)	97 (7.4)
Body Mass Index (BMI)		
mean (SD)	27.79 (4.62)	26.95 (4.28)
BMI classes		
<18.5	6 (0.5)	6 (0.5)
[18.5-25[	364 (27.9)	458 (35.1)
[25-30[	596 (45.7)	581 (44.5)
>=30	335 (25.7)	254 (19.5)
Unknown	4 (0.3)	6 (0.5)
Smoking status		

Never	553 (42.4)	603 (46.2)
Former	418 (32)	445 (34.1)
Current	315 (24.1)	233 (17.9)
Unknown	19 (1.5)	24 (1.8)
Smoking quantity		
Pack-years; mean (SD)	11.77 (17.13)	9.63 (15.34)
min-max	0.00-153.45	0.00-100.00
Alcohol consumption (g/d)		
mean (SD)	13.85 (25.14)	14.87 (29.61)
Diabetes		
No	1069 (81.9)	1099 (84.2)
Yes	80 (6.1)	54 (4.1)
Unknown	156 (12)	152 (11.7)
Hypertension		
No	612 (46.9)	718 (55)
Yes	433 (33.2)	333 (25.5)
Unknown	260 (19.9)	254 (19.5)
Fasting status		
Fasting for less than 6 hours	768 (58.8)	759 (58.2)
Fasting for 6 hours or more	476 (36.5)	497 (38.1)
Unknown	61 (4.7)	49 (3.7)

BMI: Body Mass Index; EPIC: The European Prospective Investigation into Cancer and Prevention; Estonian BB: University of Tartu- Estonian Biobank; HUNT: The Trøndelag Health Study; MCCS: The Melbourne Collaborative Cohort Study; NSHDS: Northern Sweden Health and Disease study; d:days; g:grams; N: number of participants; OR: Odds Ratio; SD: Standard Deviation.

Table 2. Metabolites robustly[Ψ] associated with kidney cancer risk.

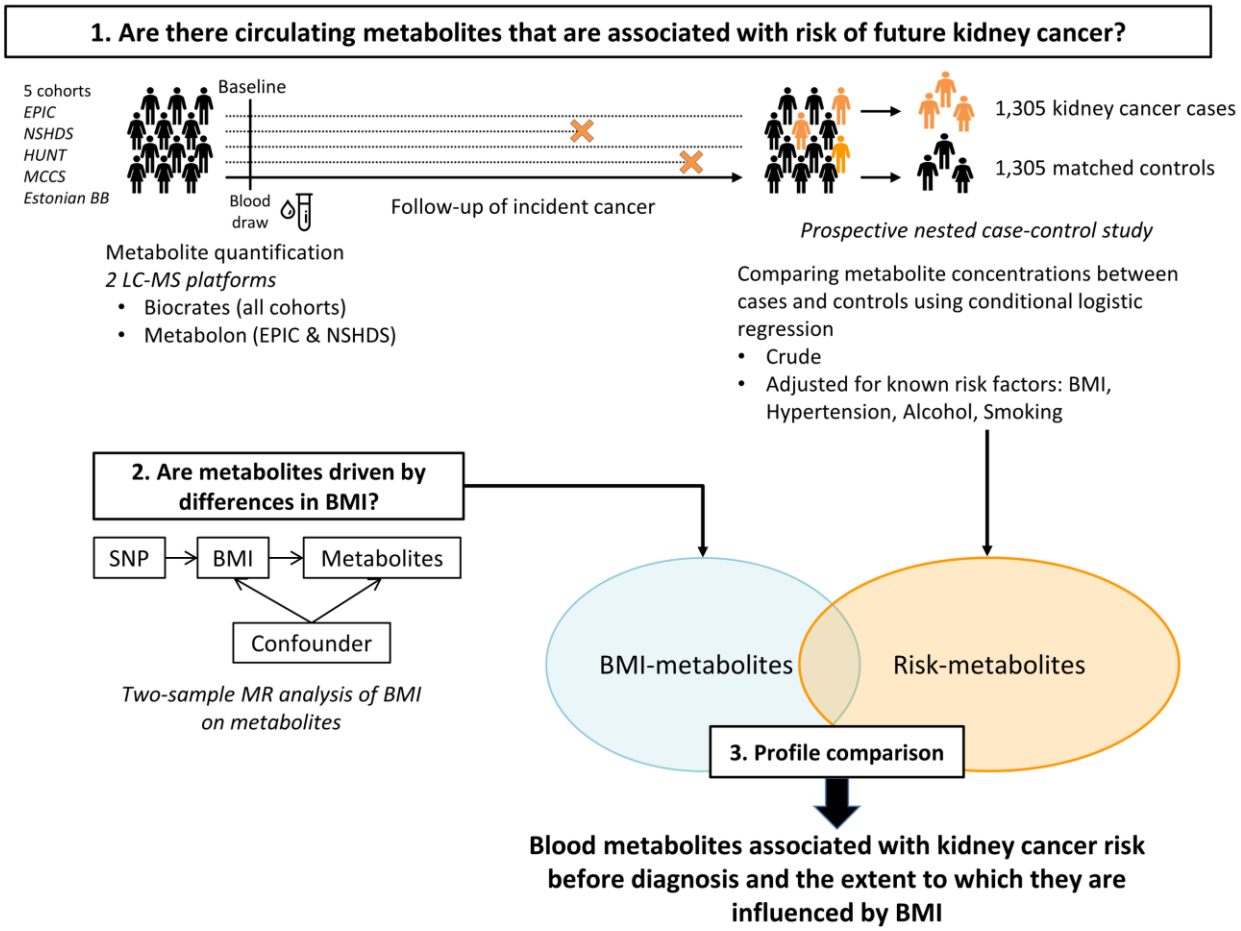
		Crude[a]				Adjusted for BMI[b]			
Metabolite Name	**Class**	**N_{pairs}**	**OR**	**95%CI**	**P-value**	**N_{pairs}**	**OR**	**95%CI**	**P-value**
Biocrates									
Glutamate	Amino Acid	1300	1.34	1.17-1.53	1.62E-05	1290	1.24	1.08-1.42	2.46E-03
Fischer's ratio	Amino Acid (ratio)	1300	1.18	1.09-1.29	1.25E-04	1290	1.14	1.04-1.24	5.02E-03
PC ae C34:3	Glycerophospholipids	1304	0.75	0.68-0.83	2.61E-08	1294	0.79	0.71-0.88	1.05E-05
lysoPC a C18:2	Glycerophospholipids	1304	0.77	0.70-0.86	9.65E-07	1294	0.81	0.73-0.90	1.35E-04
PC ae C34:2	Glycerophospholipids	1304	0.78	0.70-0.87	8.47E-06	1294	0.82	0.73-0.91	4.00E-04
lysoPC a C18:1	Glycerophospholipids	1304	0.77	0.69-0.87	1.60E-05	1294	0.81	0.72-0.92	8.04E-04
PC ae C40:1	Glycerophospholipids	1304	0.81	0.73-0.90	4.57E-05	1294	0.84	0.76-0.93	8.96E-04
PC ae C32:2	Glycerophospholipids	1304	0.78	0.69-0.89	1.27E-04	1294	0.81	0.72-0.92	1.31E-03
PC ae C36:3	Glycerophospholipids	1304	0.82	0.73-0.91	2.12E-04	1294	0.85	0.76-0.95	3.24E-03
PC ae C42:3	Glycerophospholipids	1304	0.82	0.74-0.92	4.17E-04	1294	0.87	0.78-0.98	1.75E-02
PC ae C38:6	Glycerophospholipids	1304	0.85	0.77-0.93	5.06E-04	1294	0.86	0.78-0.95	1.85E-03
PC aa C42:1	Glycerophospholipids	1304	0.83	0.75-0.93	6.27E-04	1294	0.88	0.79-0.99	2.59E-02
Metabolon									
Formiminoglutamate	Amino Acid	798	1.34	1.20-1.50	5.11E-07	794	1.28	1.14-1.45	4.23E-05
Glutamate	Amino Acid	798	1.39	1.20-1.60	5.79E-06	794	1.30	1.11-1.51	8.02E-04
Cysteine-glutathione disulfide	Amino Acid	798	0.77	0.69-0.86	7.42E-06	794	0.79	0.70-0.89	6.99E-05
Hydantoin-5-propionate	Amino Acid	798	1.25	1.12-1.39	6.17E-05	794	1.22	1.09-1.36	3.76E-04
Beta-cryptoxanthin	Cofactors and Vitamins	798	0.73	0.65-0.83	4.83E-07	794	0.76	0.67-0.86	1.81E-05
1-linoleoyl-GPC (18:2)	Glycerophospholipids	798	0.76	0.67-0.86	7.03E-06	794	0.79	0.70-0.89	2.04E-04
1-(1-enyl-palmitoyl)-GPC (P-16:0)*	Glycerophospholipids	798	0.73	0.64-0.84	8.32E-06	794	0.77	0.67-0.88	1.71E-04
1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)*	Glycerophospholipids	798	0.79	0.71-0.88	1.27E-05	794	0.83	0.74-0.93	1.41E-03
1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2)*	Glycerophospholipids	798	0.80	0.72-0.89	2.79E-05	794	0.84	0.76-0.94	1.61E-03
N1-methyladenosine	Nucleotide	798	1.40	1.23-1.60	6.50E-07	794	1.35	1.18-1.55	8.74E-06
Gamma-glutamylvaline	Peptide	798	1.38	1.23-1.56	1.22E-07	794	1.32	1.17-1.49	1.24E-05
Gamma-glutamylisoleucine*	Peptide	798	1.40	1.22-1.61	1.07E-06	794	1.33	1.15-1.53	1.01E-04
X – 12096	Unknown	798	1.33	1.17-1.51	9.97E-06	794	1.27	1.12-1.45	2.40E-04

BMI: Body Mass Index; CI: Confidence Interval; ENT: Effective Number of Test; N_{pairs}: number of case control pairs included in the analyses; OR: Odds Ratio; * metabolite identity not yet confirmed by comparison with an authentic chemical standard

a: Odds ratios and confidence intervals were estimated for 1 SD of log transformed metabolite levels by logistic regression conditioned on case set; b: Odds ratios and confidence intervals were estimated for 1 SD of log transformed metabolite levels by logistic regression conditioned on case set and adjusted for Body Mass Index; Estimated ENT are 60 and 499 for Biocrates and Metabolon metabolites, respectively. P-values threshold are thus 8.33E-04 and 1.00E-04 for Biocrates and Metabolon metabolites, respectively; $^{\Psi}$p-values below 0.05/ENT in the pooled analyses and at least nominally significant in two cohorts independently

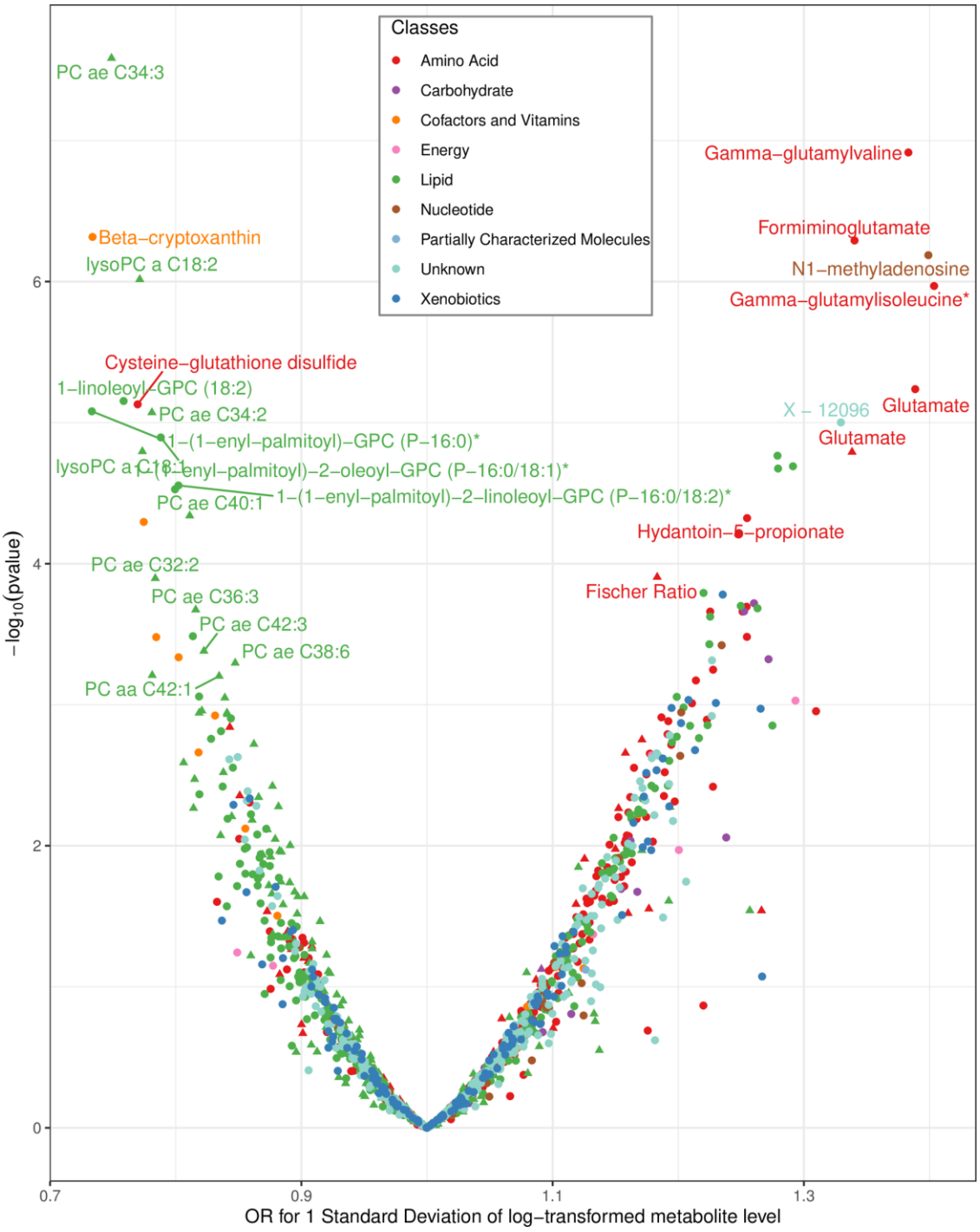
Figures Legends

Figure 1. Conceptual framework of the study design. This study includes three main analytical steps: *i)* the investigation of the associations between circulating levels of metabolites and kidney cancer risk using pre-diagnostic measurements in a case-control study nested within multiple large-scale prospective cohorts; *ii)* the assessment of the causal effect of body mass index, the leading modifiable risk factor for kidney cancer, on circulating metabolites levels; *iii)* the evaluation of the overlap between the metabolic footprint of BMI and that of kidney cancer risk.



BMI: Body Mass Index; EPIC: The European Prospective Investigation into Cancer and Prevention; Estonian BB: University of Tartu-Estonian Biobank; HUNT: The Trøndelag Health Study; LC-MS: liquid chromatography-tandem mass spectrometry; MCCS: The Melbourne Collaborative Cohort Study; MR: Mendelian Randomization; NSHDS: Northern Sweden Health and Disease study; SNP: single-nucleotide polymorphism.
The orange X's indicate the time at which a subject is diagnosed with kidney cancer when his follow-up is stopped. Controls have been selected amongst subjects free of cancer at the time their matched case was diagnosed.
Metabolites from all samples have been measured on the Biocrates platform while only samples from EPIC and NSHDS cohorts have been measured with Metabolon platform.

Figure 2. Volcano plot depicting the association between circulating metabolites measured by either Biocrates (triangle) or Metabolon (dots) with kidney cancer risk in five prospective cohorts. Metabolites that are labelled have a *p* value below the threshold (*p*<0.05/Effective number of tests (ENT)) in the pooled analyses and are nominally significant in at least 2 cohorts separately.

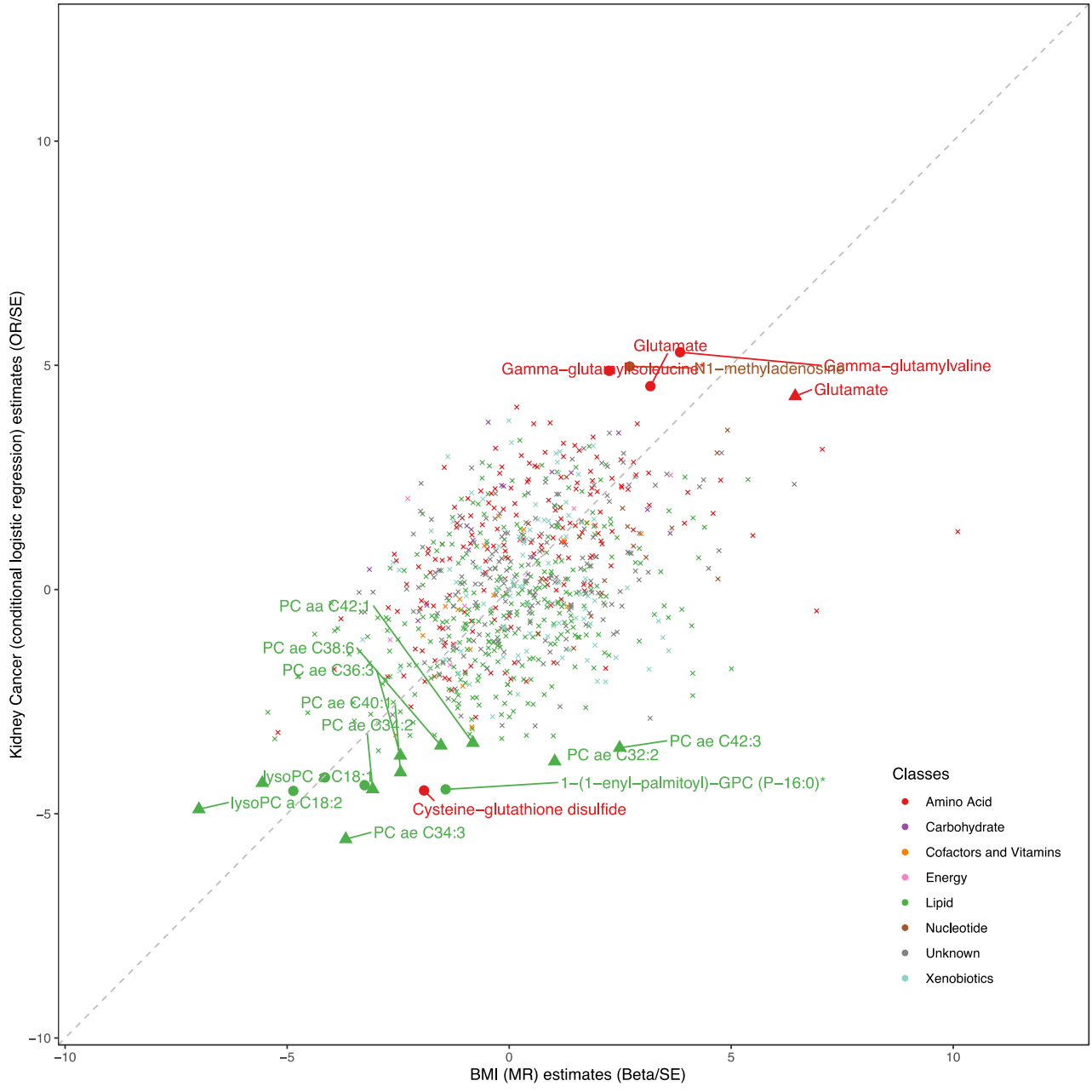


ENT: Effective Number of Test; OR: Odds Ratio; SD: Standard Deviation.

* metabolite identity not yet confirmed by comparison with an authentic chemical standard

Odds ratios and confidence intervals were estimated for 1 SD of log transformed metabolite levels by logistic regression conditioned on case set. Estimated ENT are 60 and 499 for Biocrates and Metabolon metabolites, respectively. P-values threshold are thus 8.33E-04 and 1.00E-04 for Biocrates and Metabolon metabolites, respectively.

Figure 3. Scatter plot comparing the metabolite profile associated with kidney cancer from prospective observational analyses with the BMI-driven metabolite profile from MR analyses. Metabolites that are labelled have a *p* value below the threshold ($p<0.05$/Effective number of tests (ENT)) in the prospective pooled analyses and are nominally significant in at least 2 cohorts separately. Metabolites measured by the Biocrates platform that are below the *p* value threshold are represented by triangles, those measured by the Metabolon platform that are below the *p* value threshold are represented by dots and those that are measured by either the Biocrates or the Metabolon platform that are above the *p* value threshold are represented by an x.



MR: Mendelian Randomization; OR: Odds Ratio; SE: Standard Error.

* metabolite identity not yet confirmed by comparison with an authentic chemical standard

On the y-axis, the OR and SE were derived from the logistic regression analyses conditioned on case set estimating the associations between circulating metabolites and kidney cancer risk in five prospective cohorts.

On the x-axis, the beta and SE were derived from the mendelian randomization analyses evaluating the effect of BMI on circulating metabolites levels.

Supplementary Table 1. STROBE Statement—Checklist of items that should be included in reports of *case-control studies*

	Item No	Recommendation	Paragraph number
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	Title
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	Abstract
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	Introduction, Paragraph 1 and 2
Objectives	3	State specific objectives, including any prespecified hypotheses	Introduction, Paragraph 3
Methods			
Study design	4	Present key elements of study design early in the paper	Methods, Section "Analytical strategy"
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Methods, Section "Study population, sample collection and follow-up";
			Supplementary methods, Section "Study population"
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls	Methods, Section "Study population, sample collection and follow-up";
			Supplementary methods, Section "Study population"
		(*b*) For matched studies, give matching criteria and the number of controls per case	Methods, Section "Study population, sample collection and follow-up";
			Supplementary methods, Section "Study population"

Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	Methods, Section “Study population, sample collection and follow-up” and Section “Metabolite data acquisition and quality control (QC)”;
			Supplementary methods, Section “Study population” and “Metabolite data acquisition”
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Methods, Section “Study population, sample collection and follow-up” and Section “Metabolite data acquisition and quality control (QC)”;
			Supplementary methods, Section “Study population” and “Metabolite data acquisition”
Bias	9	Describe any efforts to address potential sources of bias	Statistical analysis, Section “Primary statistical analysis: prospective observational analysis of circulating metabolites and kidney cancer” Paragraph 1 and 2
Study size	10	Explain how the study size was arrived at	Supplementary Methods, Section “Study population” and Supplementary Figure S1
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	Methods, Section “Primary statistical analysis: prospective observational analysis of circulating metabolites and kidney cancer risk” Paragraph 1 and 2
		(*b*) Describe any methods used to examine subgroups and interactions	Methods, Section “Primary statistical analysis: prospective observational analysis of circulating metabolites and kidney cancer risk” Paragraph 1 and 2
		(*c*) Explain how missing data were addressed	Analyses only included the complete data and those with missing data were excluded
		(*d*) If applicable, explain how matching of cases and controls was addressed	Methods, Section “Primary statistical analysis: prospective observational analysis of circulating metabolites and kidney cancer risk” Paragraph 1

		(*e*) Describe any sensitivity analyses	Methods, Section "Primary statistical analysis: prospective observational analysis of circulating metabolites and kidney cancer risk" Paragraph 1 and 2
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	Supplementary Methods, Section "Study population".
		(b) Give reasons for non-participation at each stage	Supplementary Methods, Section "Study population"
		(c) Consider use of a flow diagram	NA
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	Table 1
		(b) Indicate number of participants with missing data for each variable of interest	Table 1 and Supplementary Table S3
Outcome data	15*	Report numbers in each exposure category, or summary measures of exposure	Table 1
Discussion			
Key results	18	Summarise key results with reference to study objectives	Discussion, Paragraph 1
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	Discussion, Section "Strengths, limitations and prospects for future studies" Paragraph 1 and 2

Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	Discussion, Paragraph 2, 3 and 4
Generalisability	21	Discuss the generalisability (external validity) of the study results	Discussion, Section “Strengths, limitations and prospects for future studies”
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Funding

Supplementary Table 2. STROBE-MR checklist

Item	Complete/location
1. Title and Abstract: "Mendelian randomization" is named both in the title and the abstract	As the focus of this project is the prospective case-control analysis of the association between metabolites and kidney cancer, we did not include the term “Mendelian randomization” in the title. Mendelian randomization is named in the abstract, Section “Methods and Findings”
Introduction	
2.Background: Explain the scientific background and rationale for the reported study. Is causality between exposure and outcome plausible? Justify why MR is a helpful method to address the study question.	Introduction, Paragraph 1-3
3.Objectives: State specific objectives clearly, including pre-specified causal hypotheses (if any).	Introduction, Paragraph 3 and Methods, Section “Analytical strategy”
4. **Study design and data sources:** Present key elements of study design early in the paper. Consider including a table listing sources of data for all phases of the study. For each data source contributing to the analysis, describe the following: a) Describe the study design and the underlying population from which it was drawn. Describe also the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection, if available. b) Give the eligibility criteria, and the sources and methods of selection of participants. c) Explain how the analyzed sample size was arrived at. d) Describe measurement, quality and selection of genetic variants. e) For each exposure, outcome and other relevant variables, describe methods of assessment and, in the case of diseases, the diagnostic criteria used. f) Provide details of ethics committee approval and participant informed consent, if relevant.	Available information about the GWAS studies is provided in the Methods, Section “Secondary statistical analysis: Mendelian randomization and profile comparison analyses” and Supplementary Methods, Section “Data sources for Mendelian randomization analyses”. Further information is provided in each of the original GWAS publications. Selection of genetic variants is described in Section “Secondary statistical analysis: Mendelian randomization and profile comparison analyses”

5. Assumptions: Explicitly state assumptions for the main analysis (e.g. relevance, exclusion, independence, homogeneity) as well assumptions for any additional or sensitivity analysis.	Methods, Section "Secondary statistical analysis: Mendelian randomization and profile comparison analyses"; paragraph 1 and 2
6. **Statistical methods main analysis** Describe statistical methods and statistics used. a) Describe how quantitative variables were handled in the analyses (i.e., scale, units, model). b) Describe the process for identifying genetic variants and weights to be included in the analyses (i.e, independence and model). Consider a flow diagram. c) Describe the MR estimator, e.g. two-stage least squares, Wald ratio, and related statistics. Detail the included covariates and, in case of two-sample MR, whether the same covariate set was used for adjustment in the two samples. d) Explain how missing data were addressed. e) If applicable, say how multiple testing was dealt with.	a, b, c) Methods, Section "Secondary statistical analysis: Mendelian randomization and profile comparison analyses" d, e) Not applicable to our study
7. Assessment of assumptions: Describe any methods used to assess the assumptions or justify their validity.	Methods, Section "Secondary statistical analysis: Mendelian randomization and profile comparison analyses"; paragraph 1 and 2
8.Sensitivity analyses: Describe any sensitivity analyses or additional analyses performed.	Methods, Section "Secondary statistical analysis: Mendelian randomization and profile comparison analyses"; paragraph 1 and 2
9. **Software and pre-registration** a) Name statistical software and package(s), including version and settings used. b) State whether the study protocol and details were pre-registered (as well as when and where).	a) Methods, Section "Secondary statistical analysis: Mendelian randomization and profile comparison analyses" b) Methods, Section "Analytical strategy"
Results	

10. **Descriptive data** a) Report the numbers of individuals at each stage of included studies and reasons for exclusion. Consider use of a flow-diagram. b) Report summary statistics for phenotypic exposure(s), outcome(s) and other relevant variables (e.g. means, standard deviations, proportions). c) If the data sources include meta-analyses of previous studies, provide the number of studies, their reported ancestry, if available, and assessments of heterogeneity across these studies. Consider using a supplementary table for each data source. d) For two-sample Mendelian randomization: i. Provide information on the similarity of the genetic variant-exposure associations between the exposure and outcome samples. ii. Provide information on extent of sample overlap between the exposure and outcome data sources.	a) Methods, Section "Secondary statistical analysis: Mendelian randomization and profile comparison analyses" b) Supplementary Table S5 and S6 c) Methods, "Secondary statistical analysis: Mendelian randomization and profile comparison analyses" and Supplementary Methods, Section "Data sources for Mendelian randomization analyses" d) Supplementary Methods, Section "Data sources for Mendelian randomization analyses"
11. **Main results** a) Report the associations between genetic variant and exposure, and between genetic variant and outcome, preferably on an interpretable scale (e.g. comparing 25th and 75th percentile of allele count or genetic risk score, if individual-level data available). b) Report causal effect estimate between exposure and outcome, and the measures of uncertainty from the MR analysis. Use an intuitive scale, such as odds ratio, or relative risk, per standard deviation difference. c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time-period. d) Consider any plots to visualize results (e.g. forest plot, scatterplot of associations between genetic variants and outcome versus between genetic variants and exposure).	Our results are given in terms of betas and confidence intervals throughout the results section. We visualize results using a volcano plot in Supplementary Figure S45 and a scatter plot in Figure 3 and Supplementary Figure S46

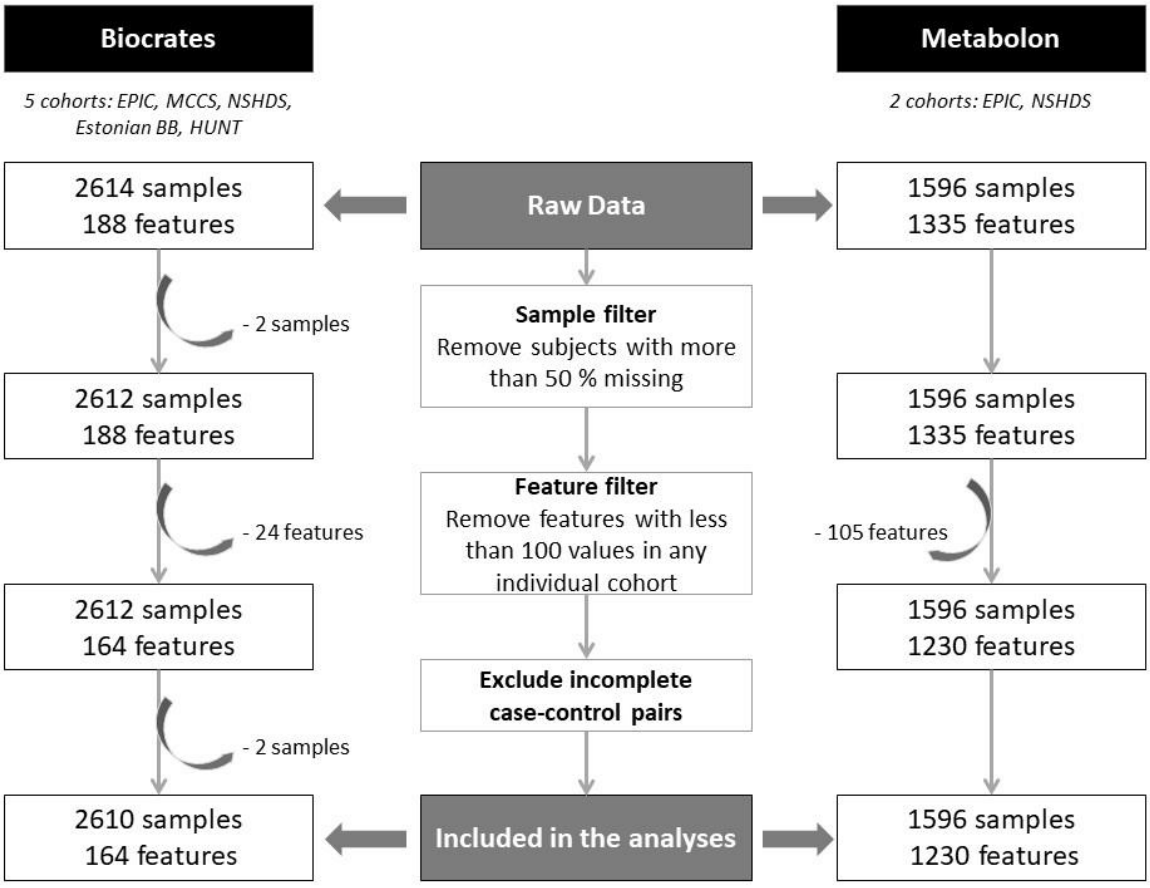
12. **Assessment of assumptions** a) Assess the validity of the assumptions. b) Report any additional statistics (e.g., assessments of heterogeneity, such as I2, Q statistic).	We assessed the validity using sensitivity analyses such as MR-Egger and weighted median analyses, described in the "Two sample Mendelian randomization and profile comparison analyses" section of the results section and presented in Supplementary Table S11 and S12.
13. **Sensitivity and additional analyses** a) Use sensitivity analyses to assess the robustness of the main results to violations of the assumptions. b) Report results from other sensitivity analyses (e.g., replication study with different dataset, analyses of subgroups, validation of instrument(s), simulations, etc.). c) Report any assessment of direction of causality (e.g., bidirectional MR). d) When relevant, report and compare with estimates from non-MR analyses. e) Consider any additional plots to visualize results (e.g., leave-one-out analyses).	Results, "Two sample Mendelian randomization and profile comparison analyses" and Supplementary Table S11 and S12.
Discussion	
14. Key results	Discussion, Paragraph 1
15. **Limitations** Discuss limitations of the study, taking into account the validity of the MR assumptions, other sources of potential bias, and imprecision. Discuss both direction and magnitude of any potential bias, and any efforts to address them.	Discussion, Section “Strengths, limitations and prospects for future studies”

16. **Interpretations** a) Give a cautious overall interpretation of results considering objectives and limitations. Compare with results from other relevant studies. b) Discuss underlying biological mechanisms that could be modelled by using the genetic variants to assess the relationship between the exposure and the outcome. c) Discuss whether the results have clinical or policy relevance, and whether interventions could have the same size effect.	a, b) Discussion, Paragraph 2-4 c) Conclusion
17. Generalizability:	Discussion, Section "Strengths, limitations and prospects for future studies"
18. Funding:	Funding
19. Data and data sharing:	Data availability statement
20. Conflicts of Interest:	Conflict of interest statement

Supplementary Figures

Table of Contents

Supplementary Figure S1. Overview of the quality control pipelines used for the metabolite measurements pre-processing



EPIC: The European Prospective Investigation into Cancer and Prevention; Estonian BB: University of Tartu- Estonian Biobank; HUNT: The Trøndelag Health Study; MCCS: The Melbourne Collaborative Cohort Study; NSHDS: Northern Sweden Health and Disease study.

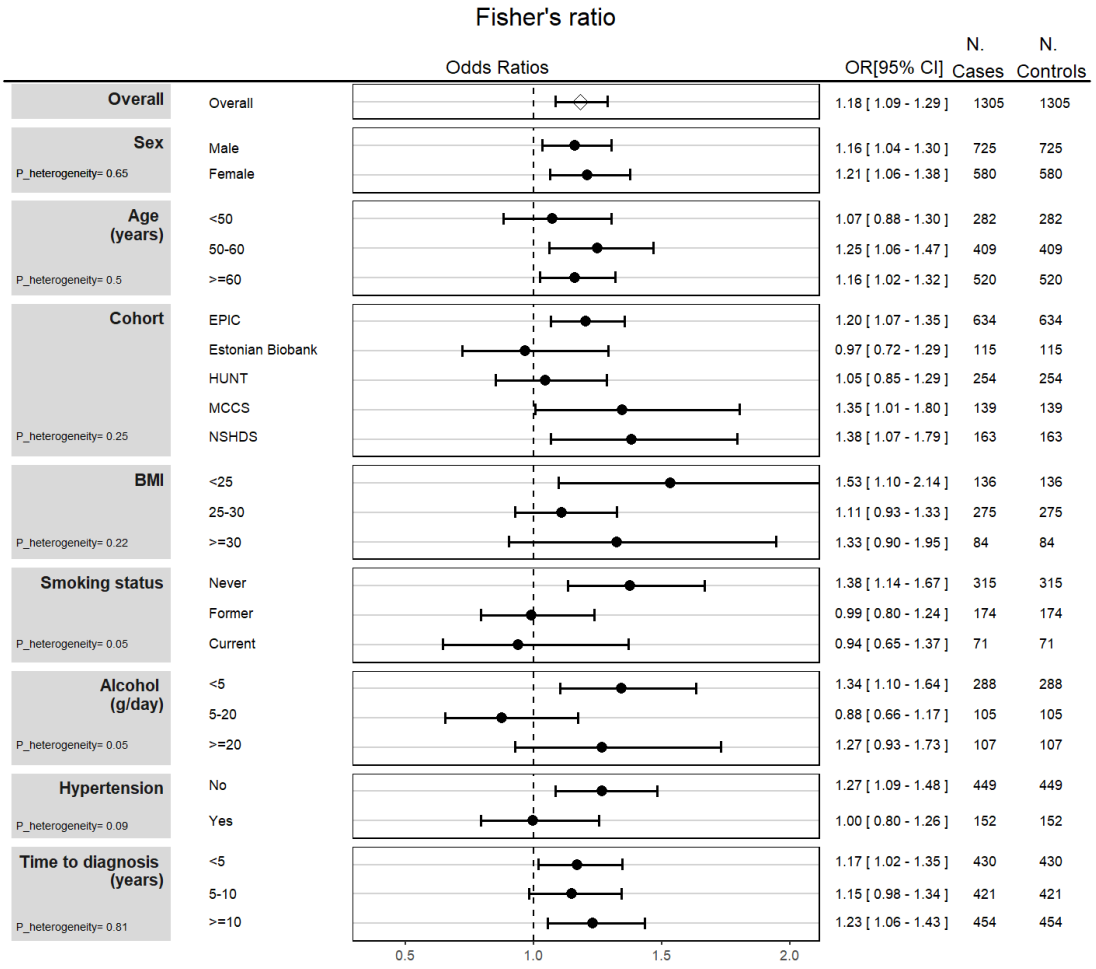
The number of features is the maximum number of features measured across all cohorts. Some are only measured in one cohort.

Supplementary Figure S2. Heatmap of Pearson correlation coefficients for selected Biocrates (left) and Metabolon (right) metabolites.

Corr: Pearson correlation coefficient

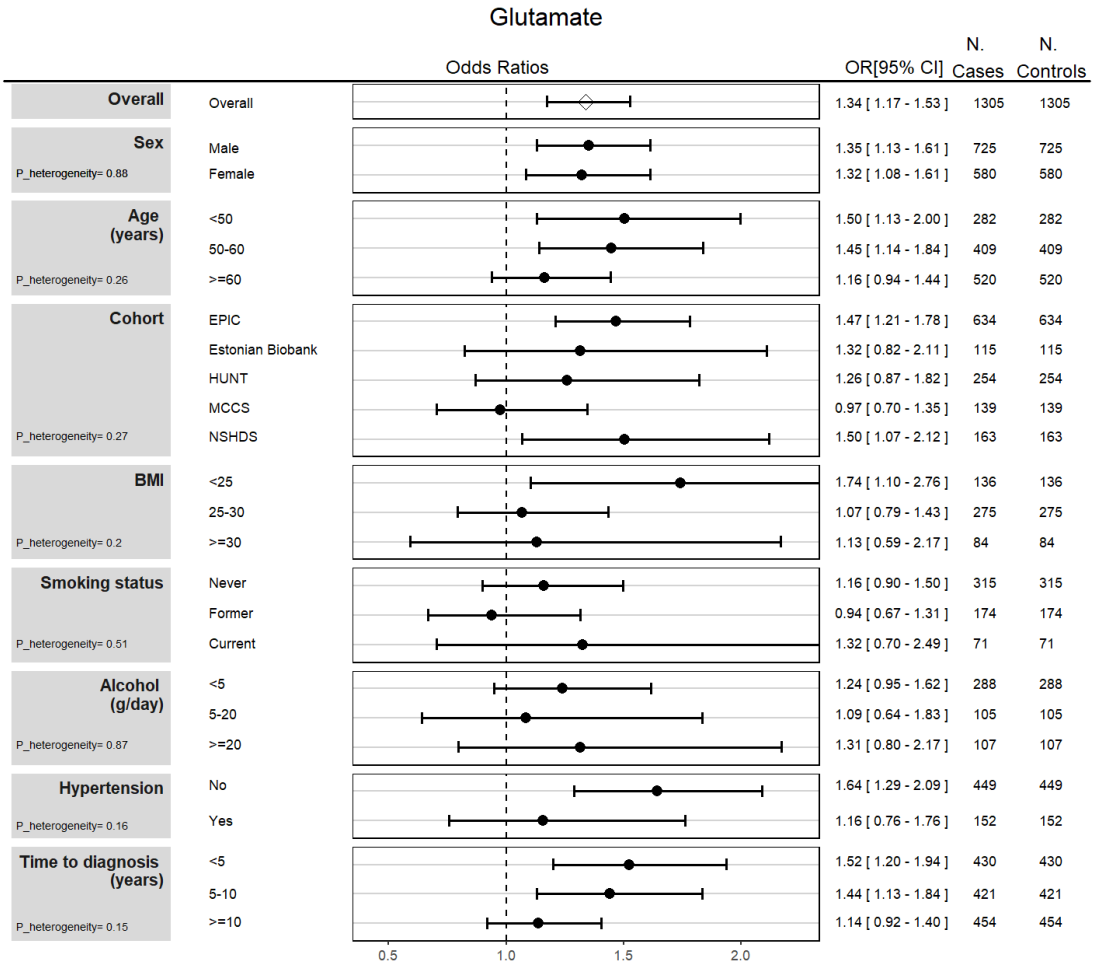
Supplementary Figures S3 to S27. Forest plots depicts the kidney cancer risk association for each metabolite deemed robustly associated with kidney cancer risk, stratified by specific risk factors.

Supplementary Figure S3. Forest plots depicts the kidney cancer risk association for the Fischer's ratio stratified by kidney cancer risk factors.



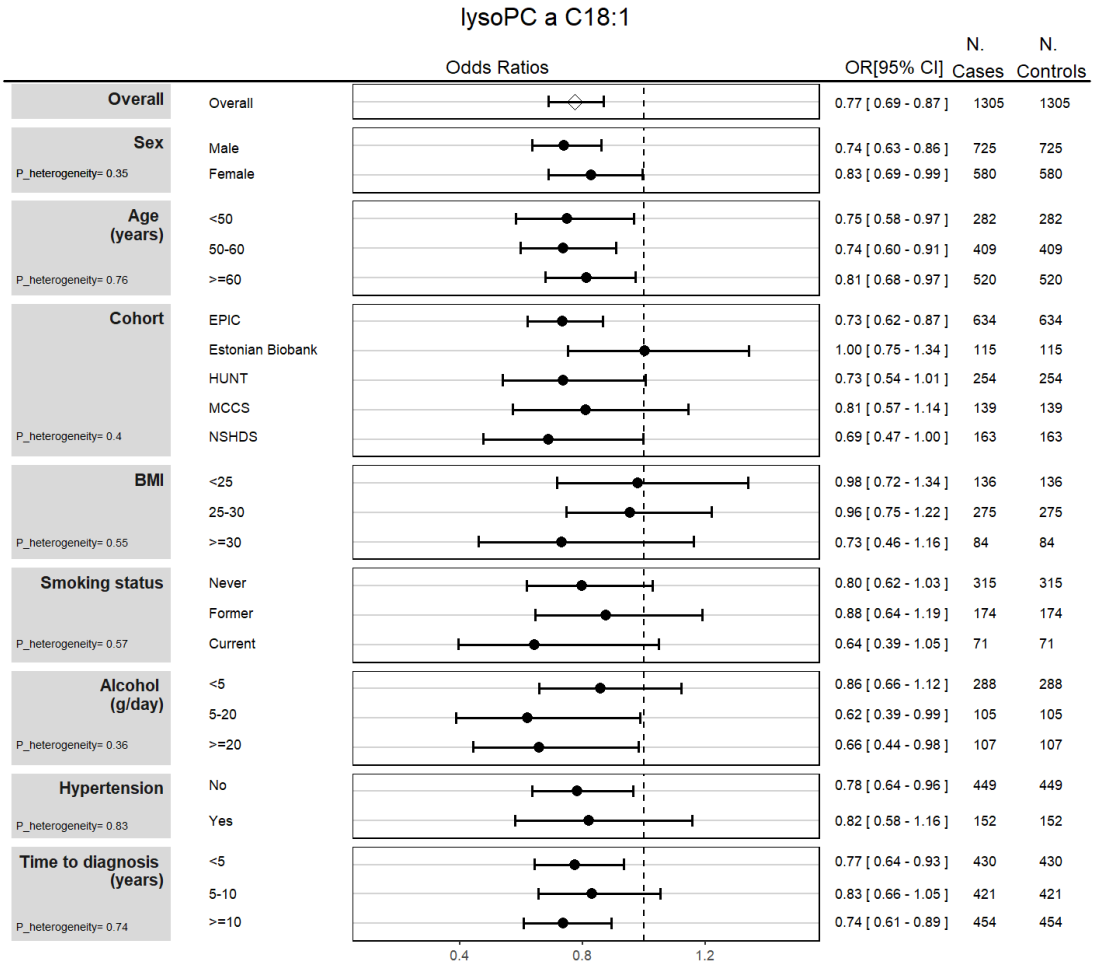
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.
The Fischer's ratio is a clinical indicator of liver metabolism and function, was calculated as the molar ratio of branched chain amino acids (leucine + isoleucine + valine) to aromatic amino acids (phenylalanine + tyrosine). Lower Fischer's ratio values are associated with liver dysfunction.

Supplementary Figure S4. Forest plots depicts the kidney cancer risk association for glutamate (Biocrates), stratified by risk factors.



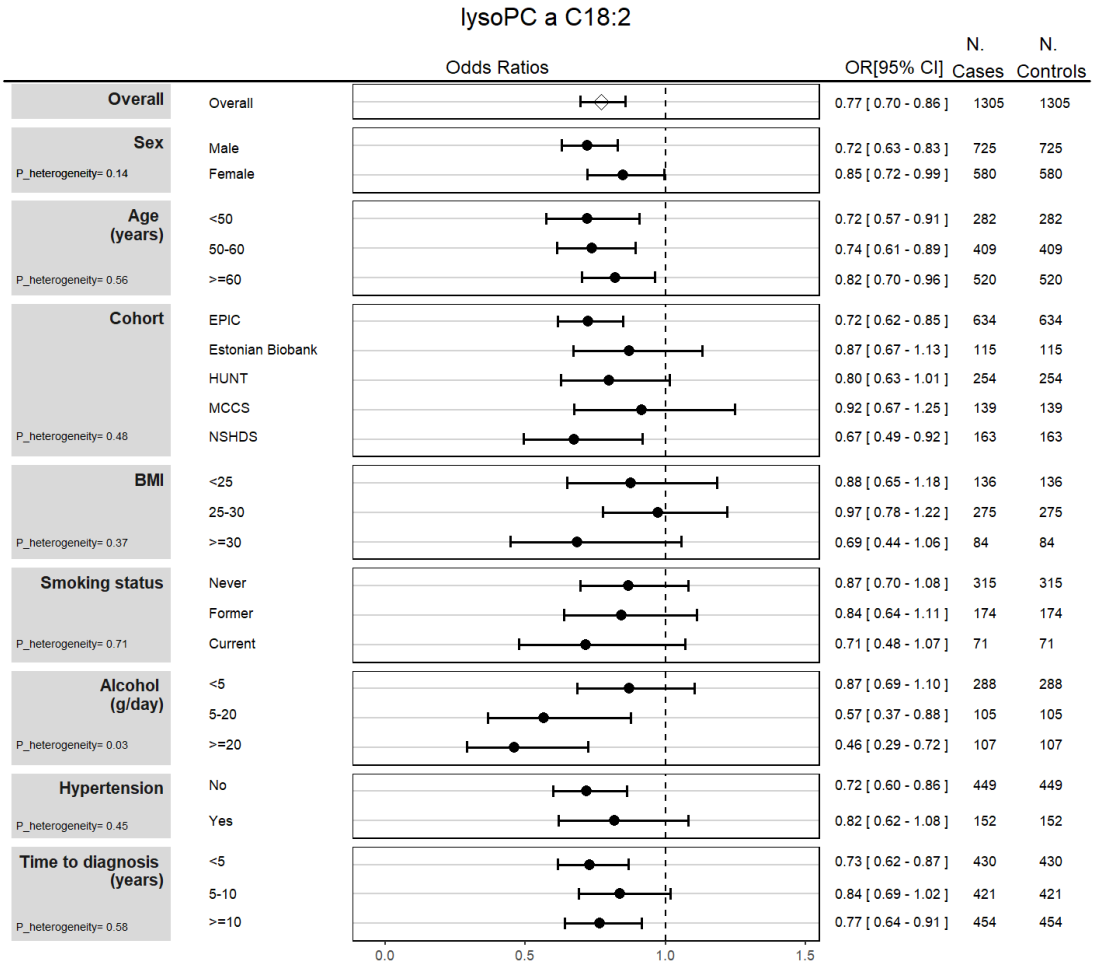
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S5. Forest plots depicts the kidney cancer risk association for lysoPC a C18:1, stratified by risk factors.



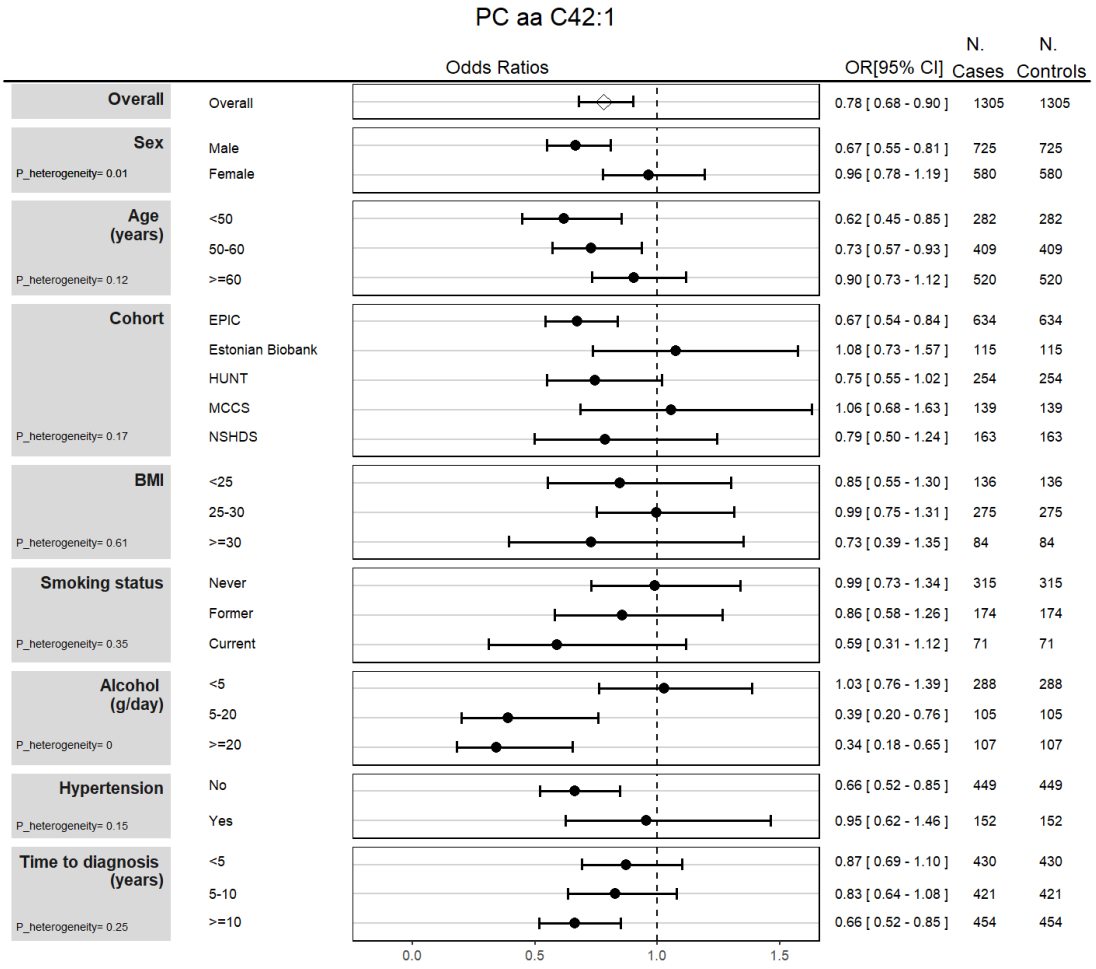
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S6. Forest plots depicts the kidney cancer risk association for lysoPC a C18:2, stratified by risk factors.



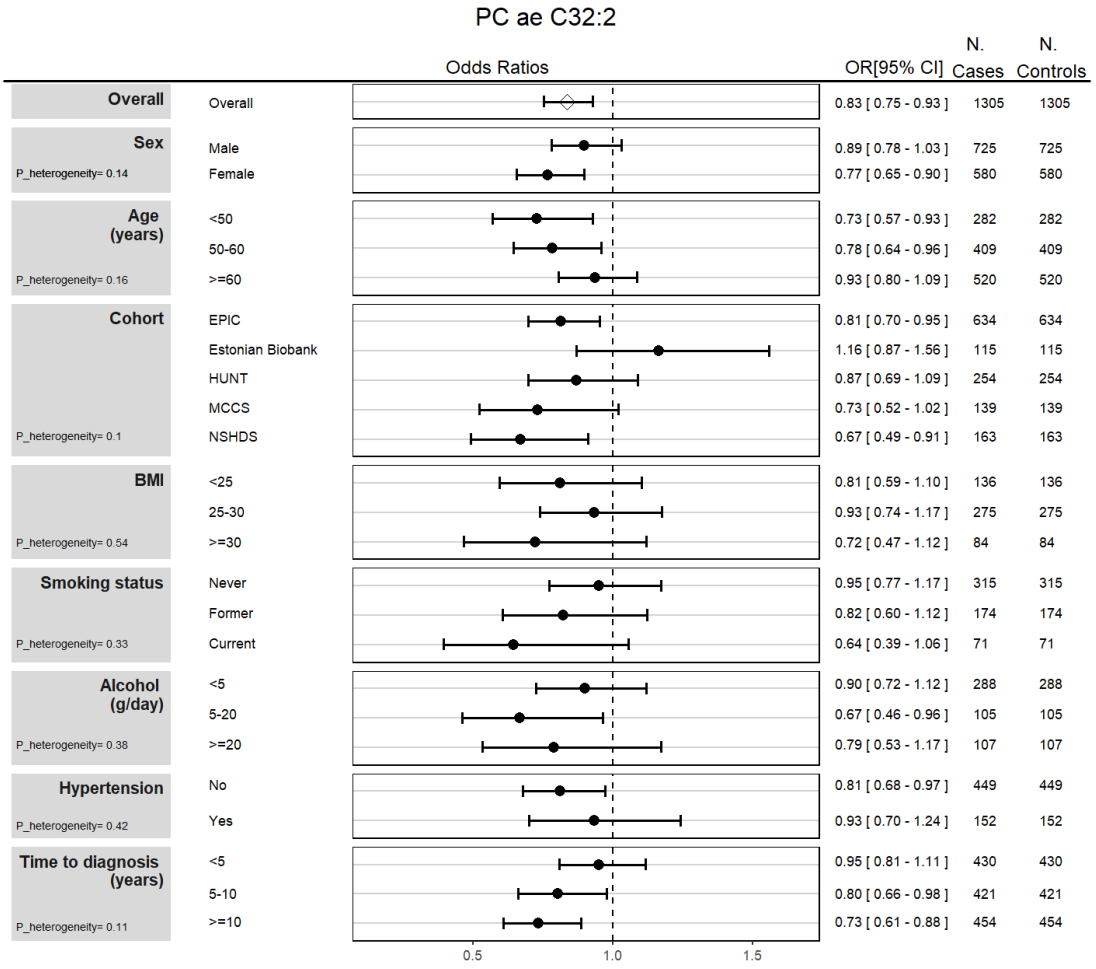
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S7. Forest plots depicts the kidney cancer risk association for PC aa C42:1, stratified by risk factors.



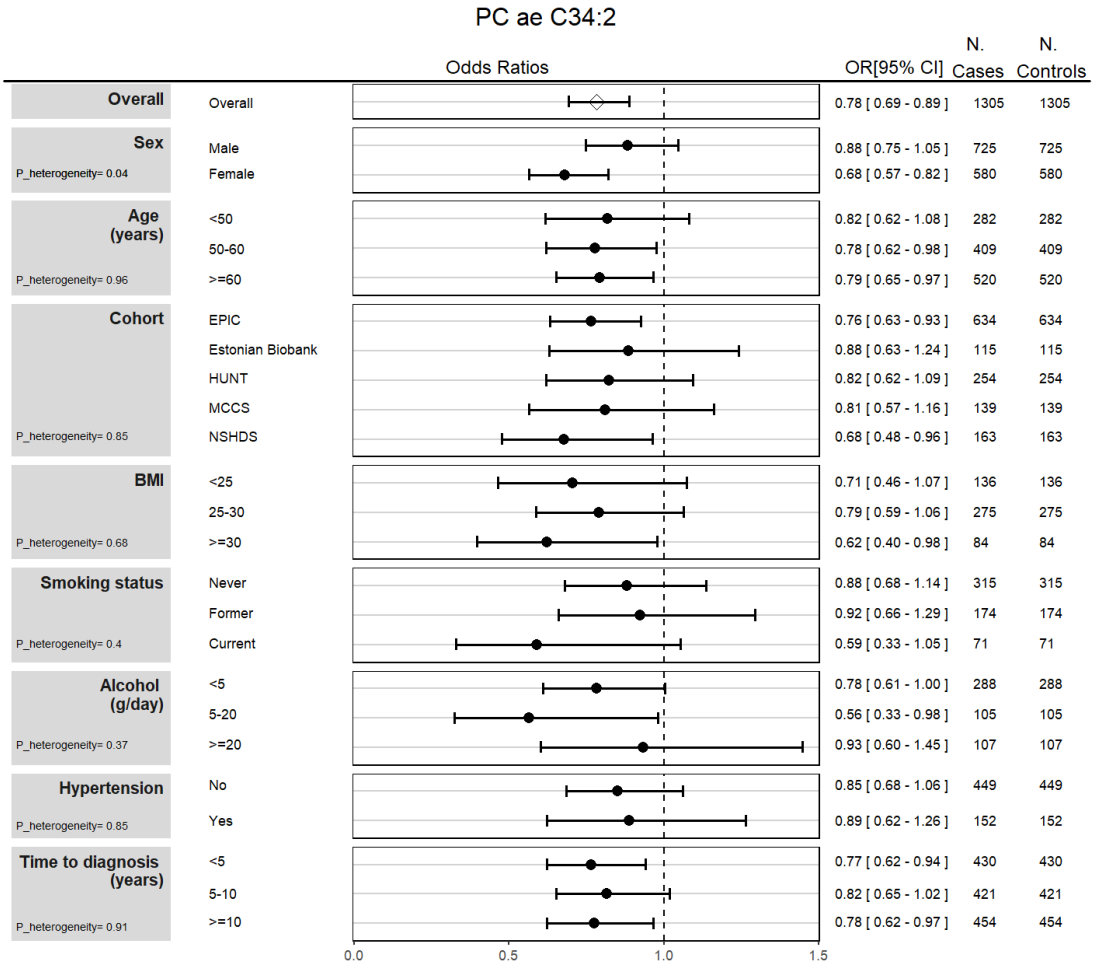
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S8. Forest plots depicts the kidney cancer risk association for PC ae C32:2, stratified by risk factors.



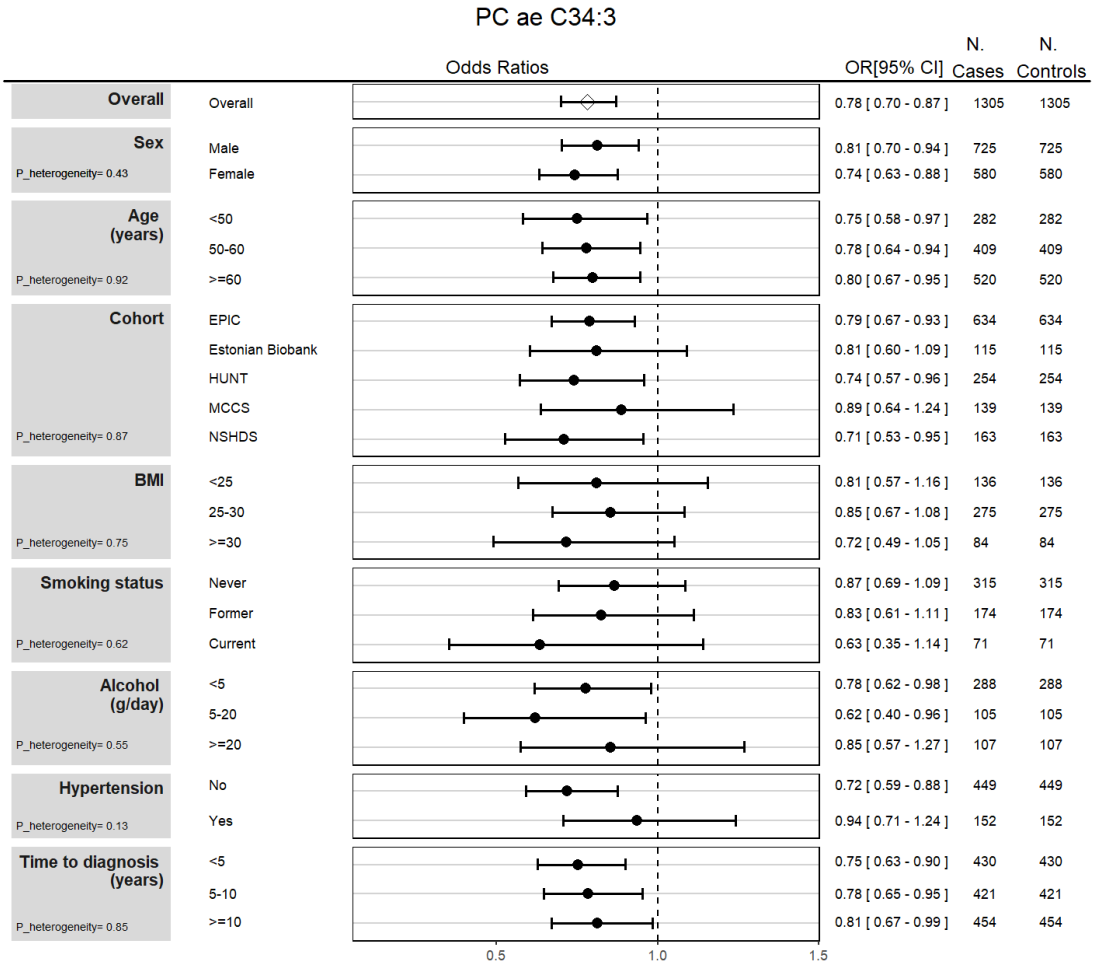
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S9. Forest plots depicts the kidney cancer risk association for PC ae C34:2, stratified by risk factors.



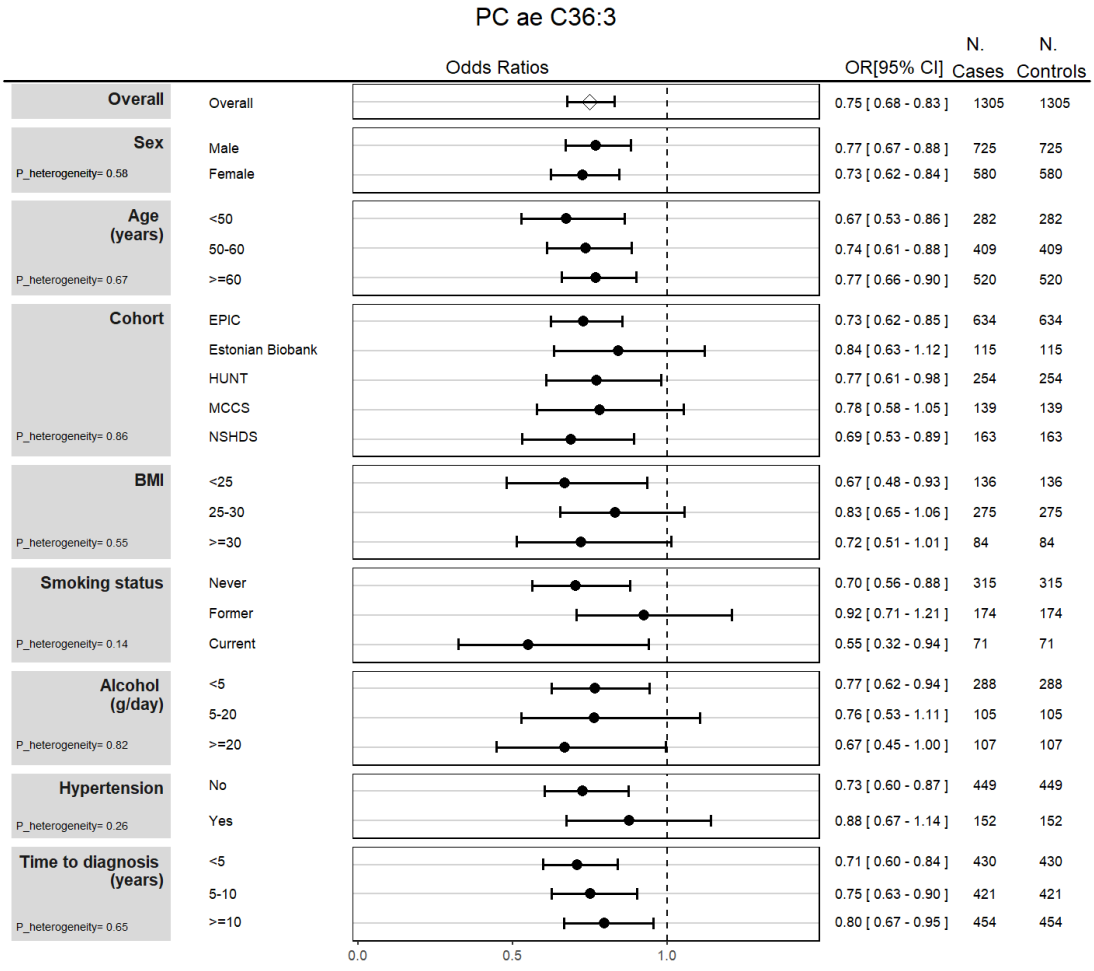
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S10. Forest plots depicts the kidney cancer risk association for PC ae C34:3, stratified by risk factors.



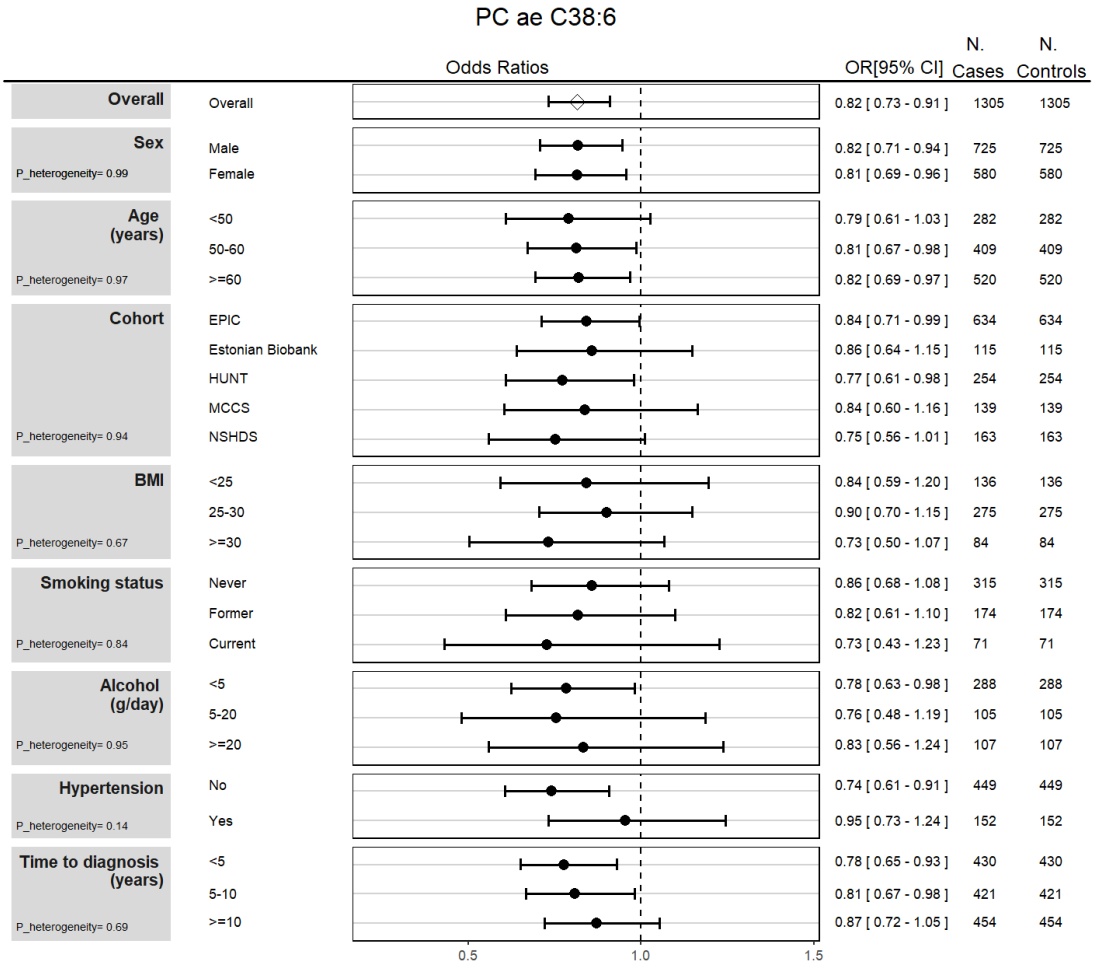
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S11. Forest plots depicts the kidney cancer risk association for PC ae C36:3, stratified by risk factors.



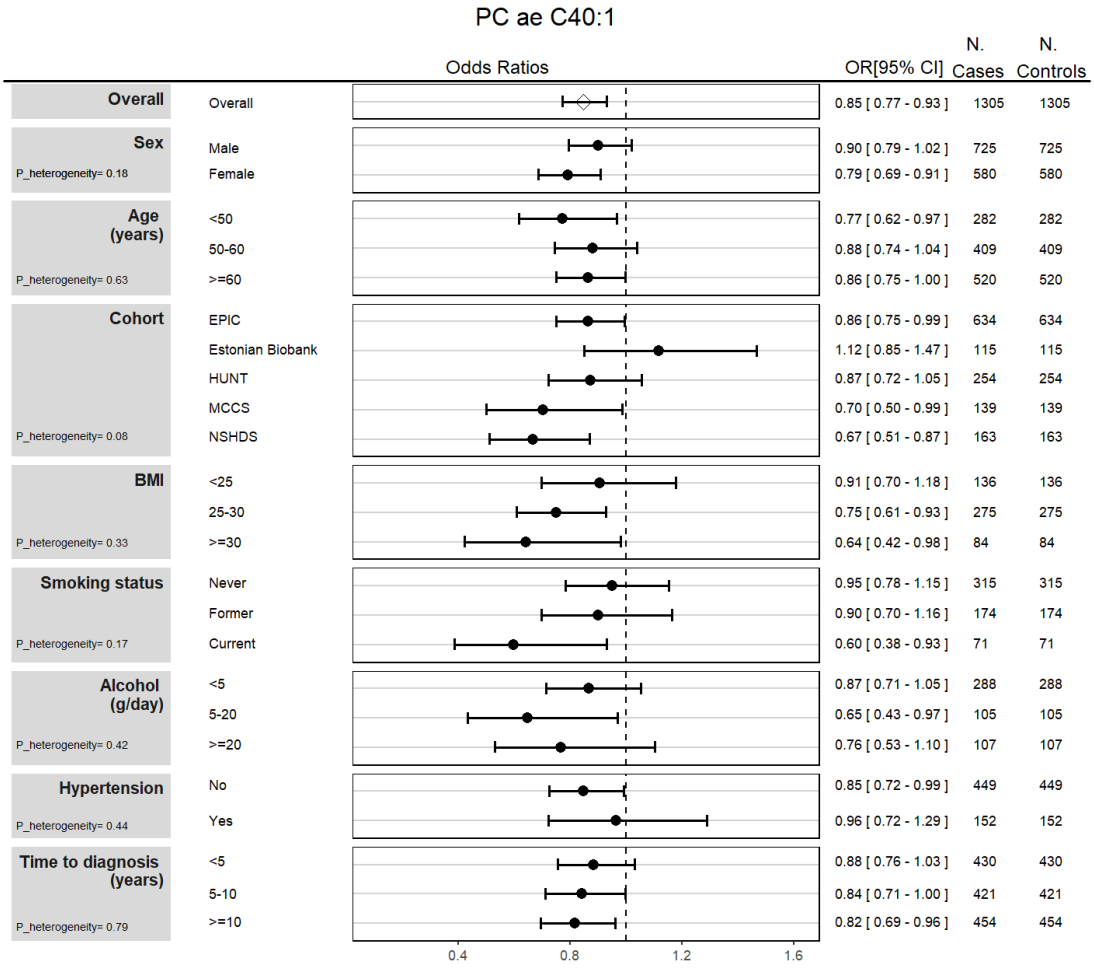
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S12. Forest plots depicts the kidney cancer risk association for PC ae C38:6, stratified by risk factors.



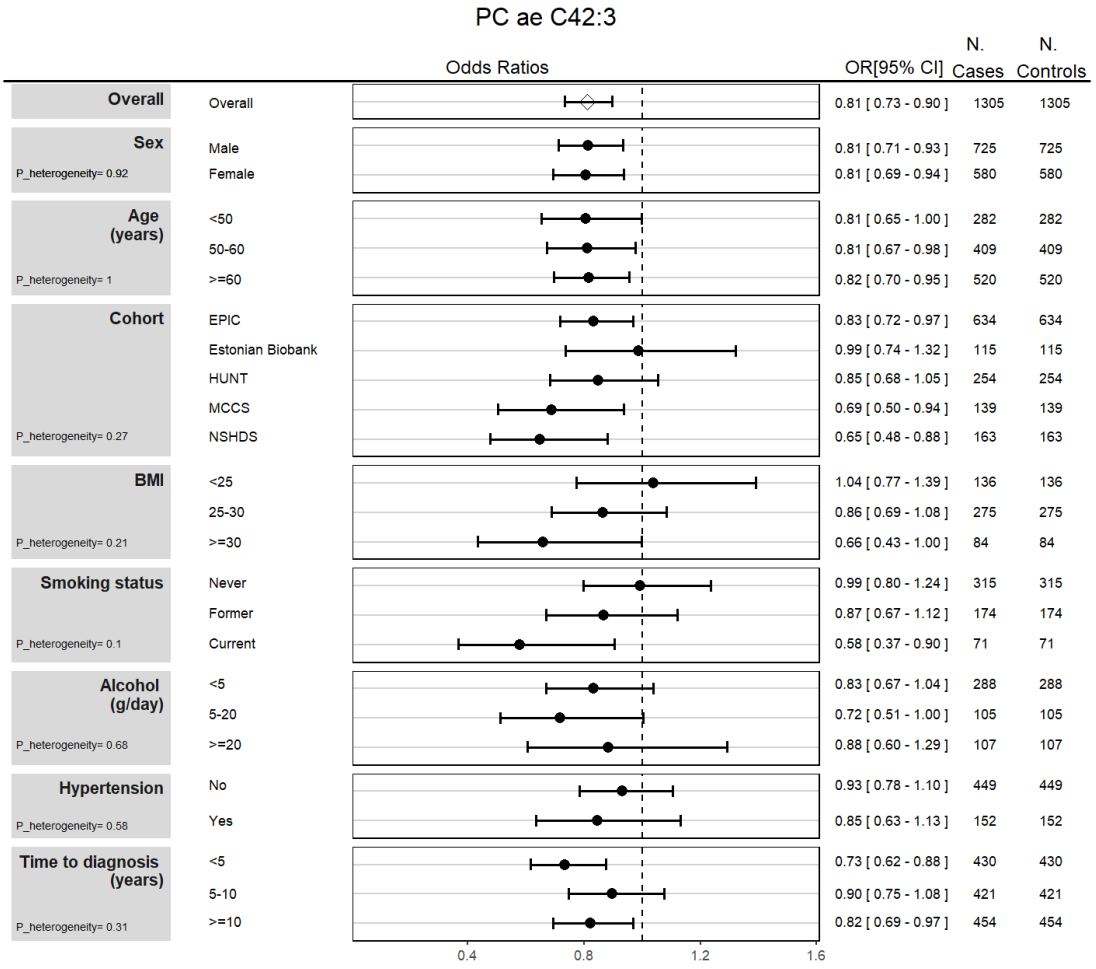
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S13. Forest plots depicts the kidney cancer risk association for PC ae C40:1, stratified by risk factors.



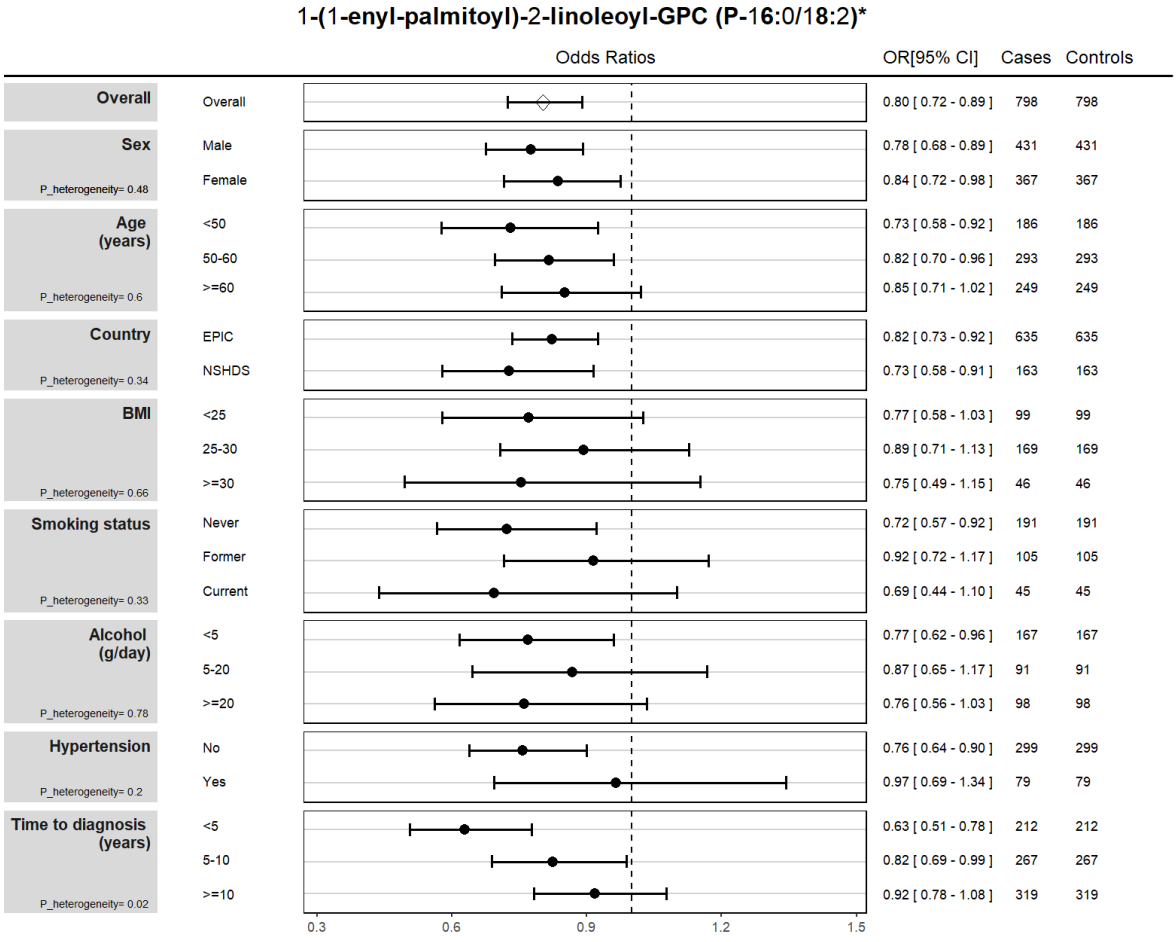
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S14. Forest plots depicts the kidney cancer risk association for PC ae C42:3, stratified by risk factors.



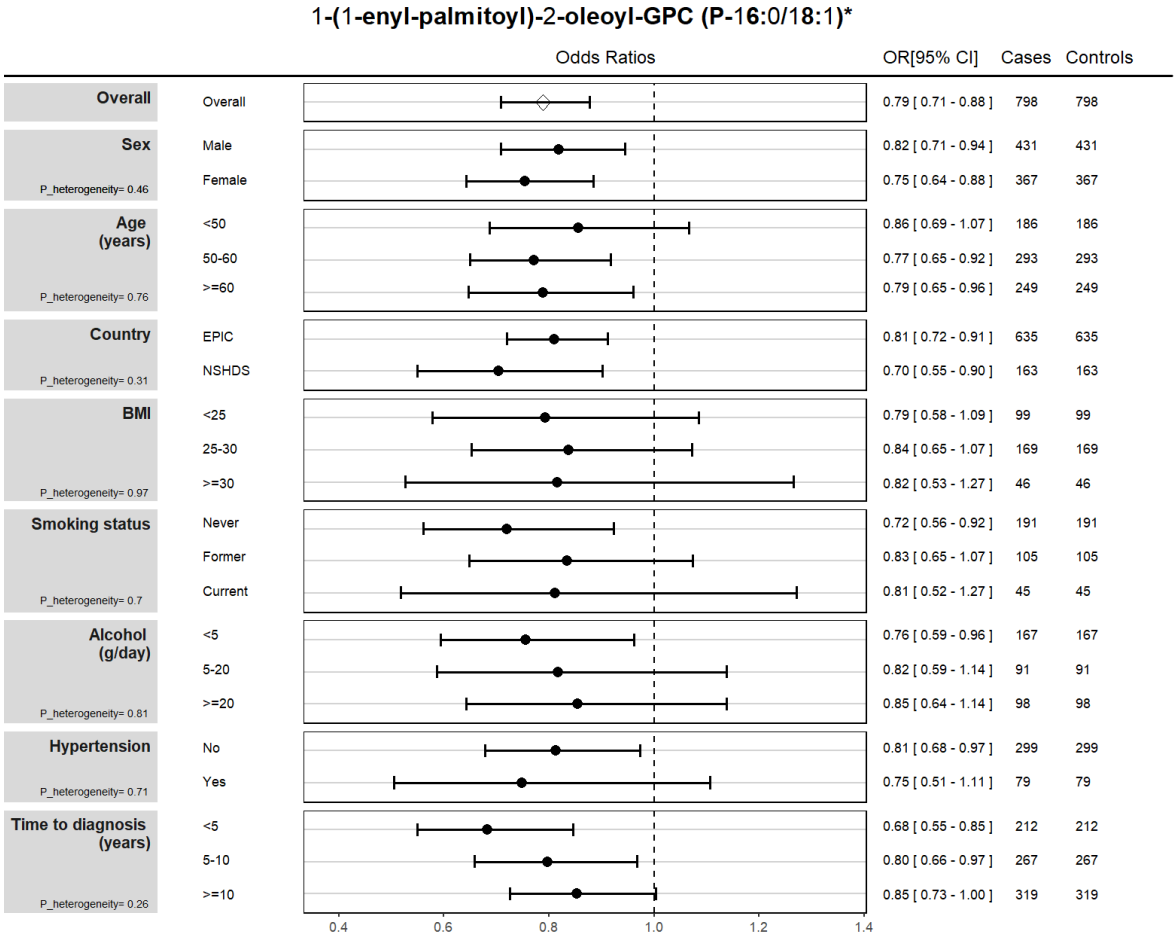
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S15. Forest plots depicts the kidney cancer risk association for 1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2)*, stratified by risk factors.



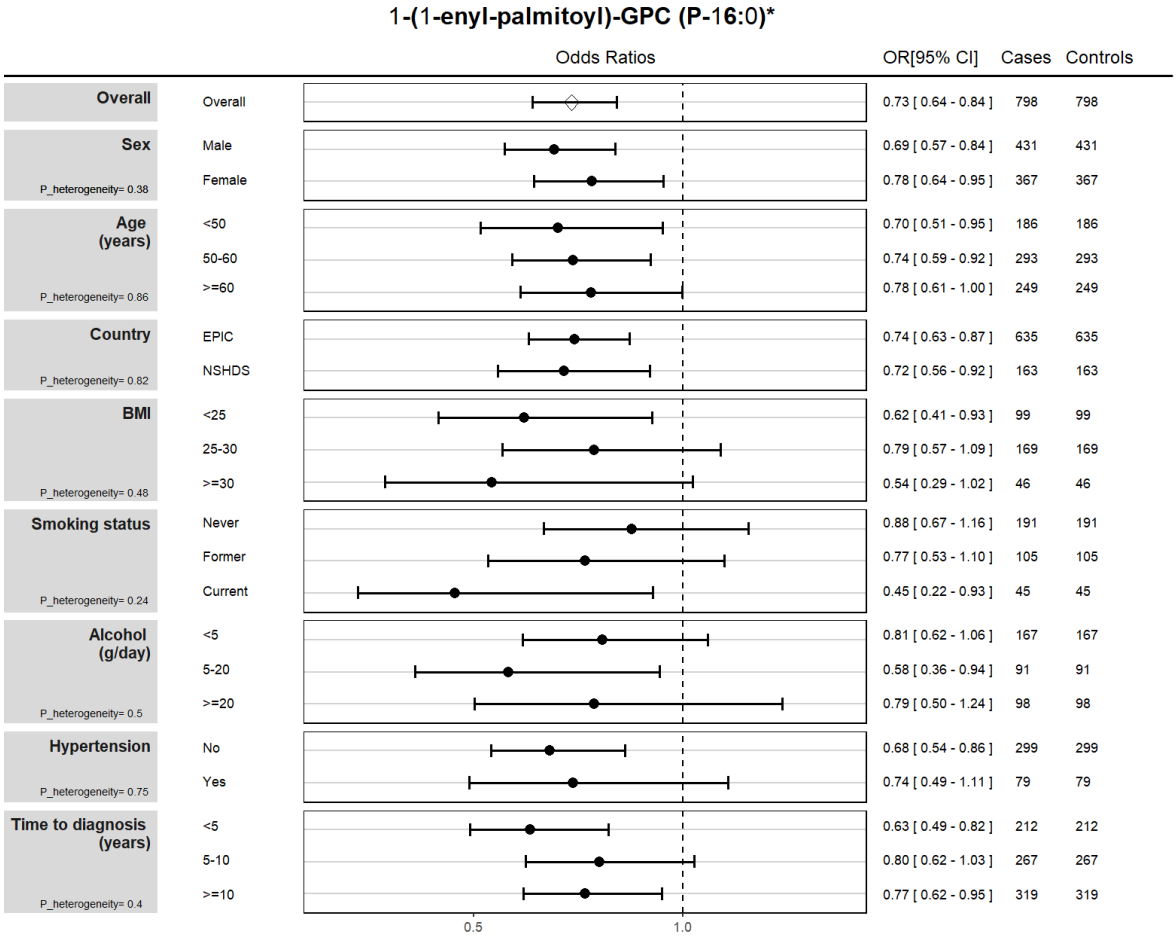
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.
* metabolite identity not yet confirmed by comparison with an authentic chemical standard

Supplementary Figure S16. Forest plots depicts the kidney cancer risk association for 1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)*, stratified by risk factors.



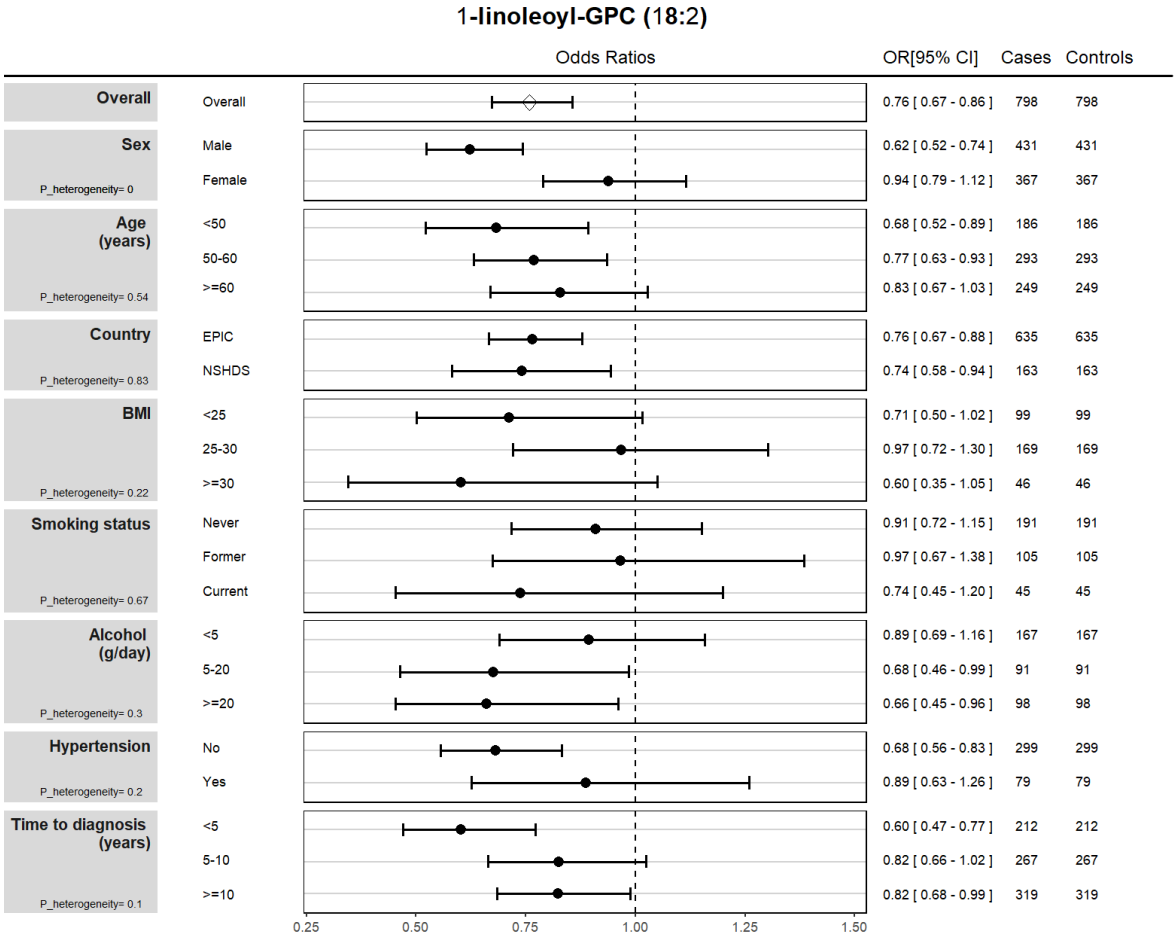
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.
* metabolite identity not yet confirmed by comparison with an authentic chemical standard

Supplementary Figure S17. Forest plots depicts the kidney cancer risk association for 1-(1-enyl-palmitoyl)-GPC (P-16:0)*, stratified by risk factors.



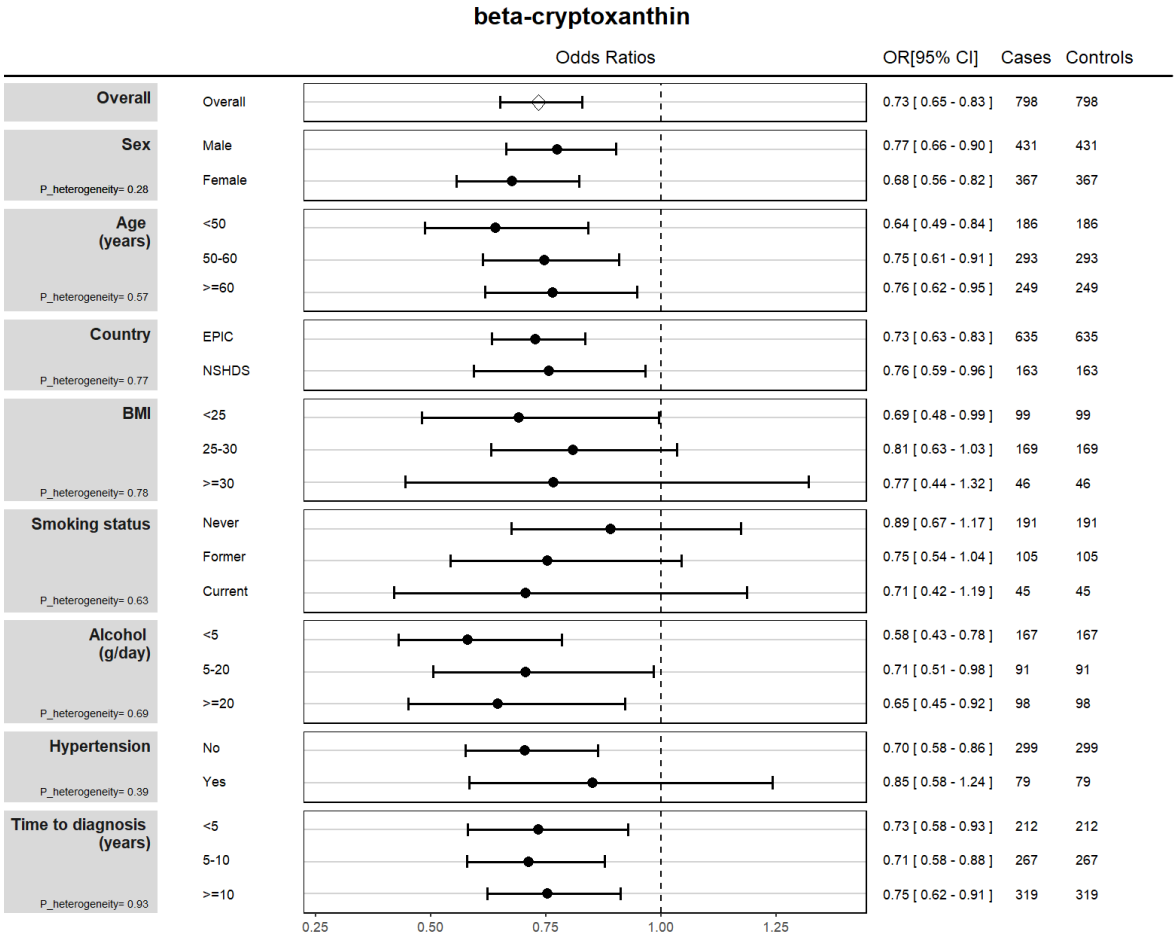
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.
* metabolite identity not yet confirmed by comparison with an authentic chemical standard

Supplementary Figure S18. Forest plots depicts the kidney cancer risk association for 1-linoleoyl-GPC (18:2), stratified by risk factors.



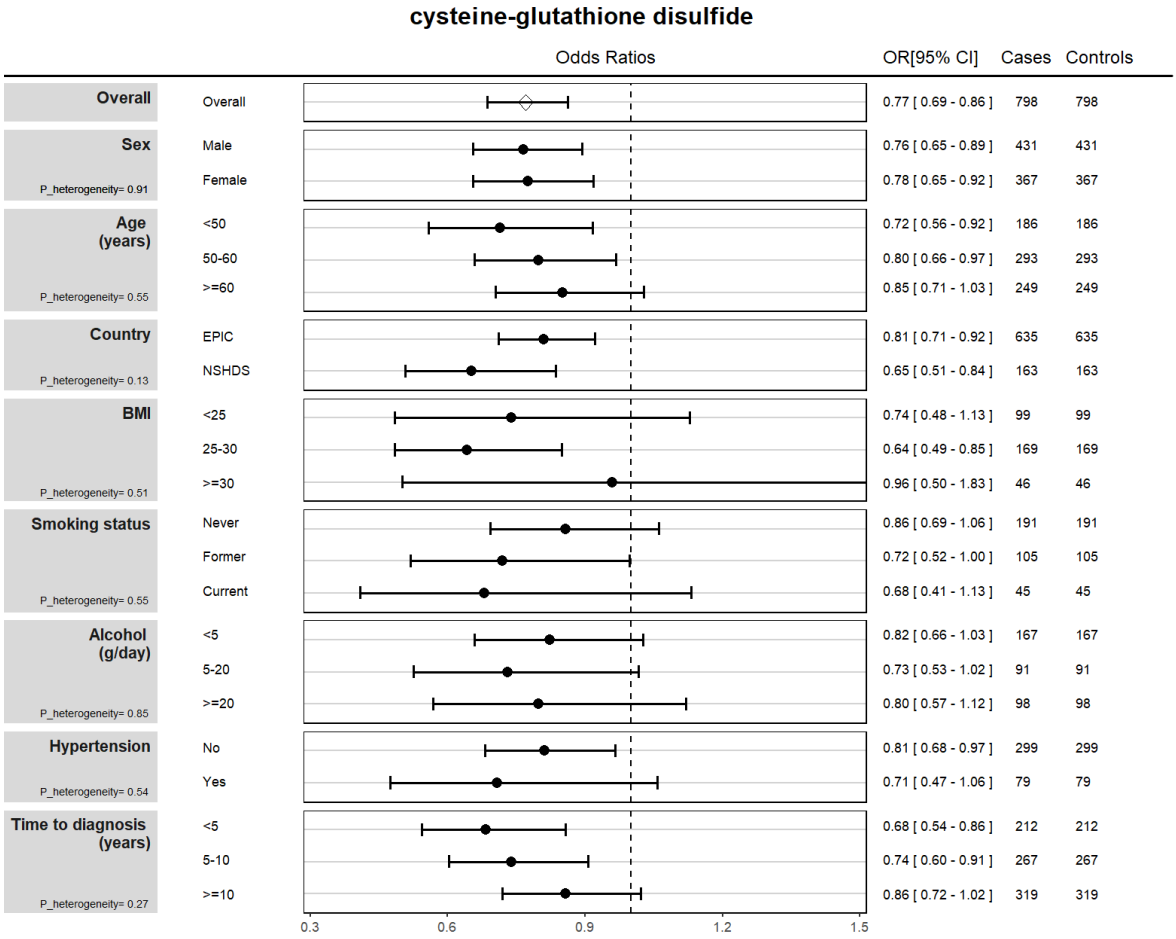
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S19. Forest plots depicts the kidney cancer risk association for beta-cryptoxanthin, stratified by risk factors.



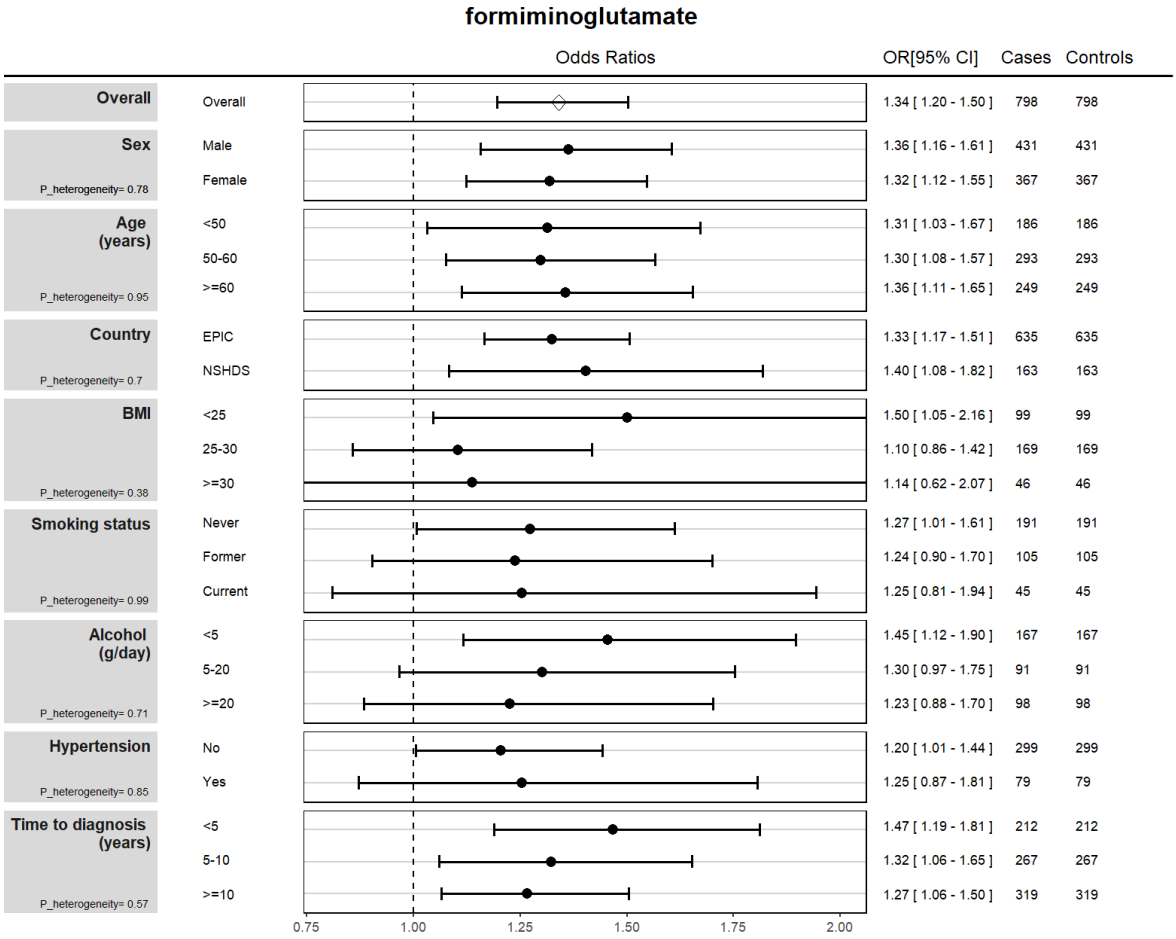
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S20. Forest plots depicts the kidney cancer risk association for cysteine-glutathione disulfide, stratified by risk factors.



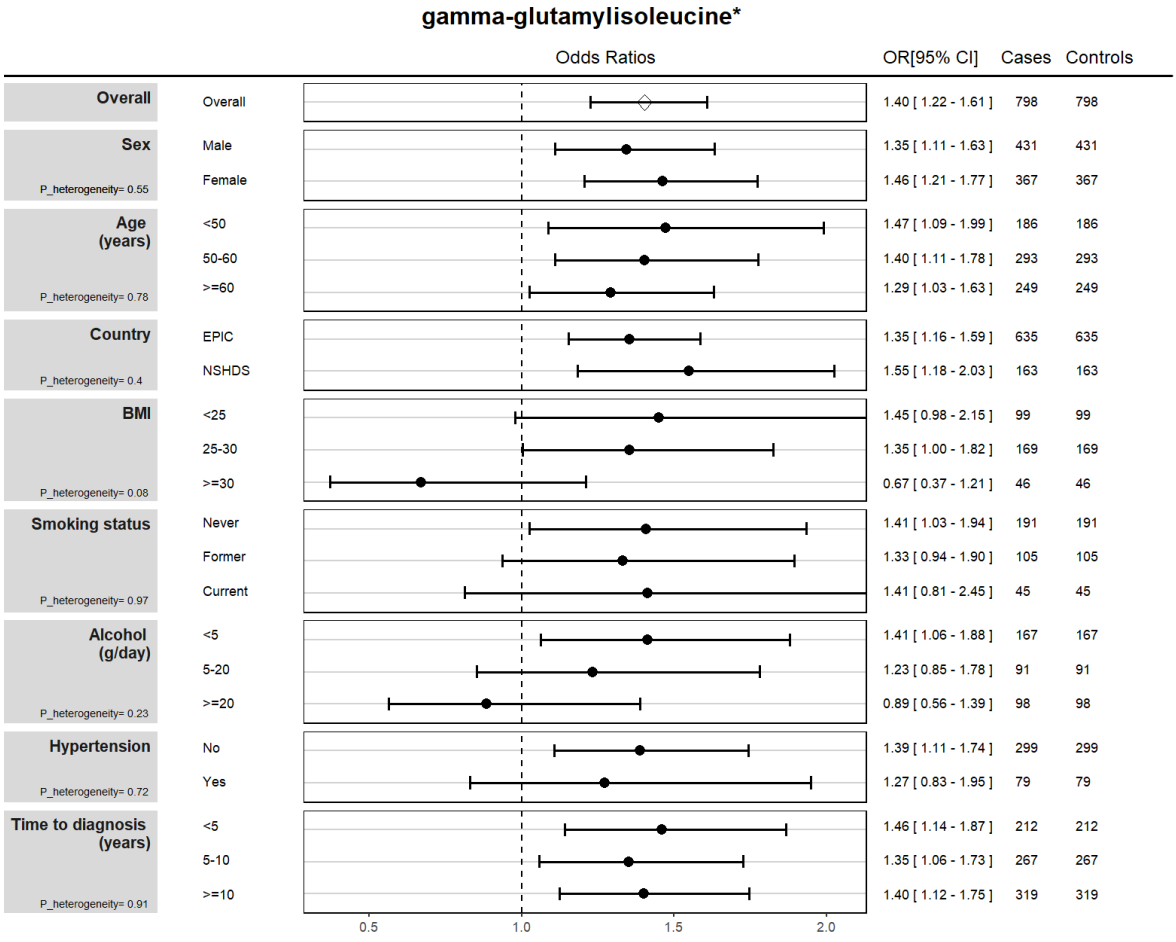
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S21. Forest plots depicts the kidney cancer risk association for formiminoglutamate, stratified by risk factors.



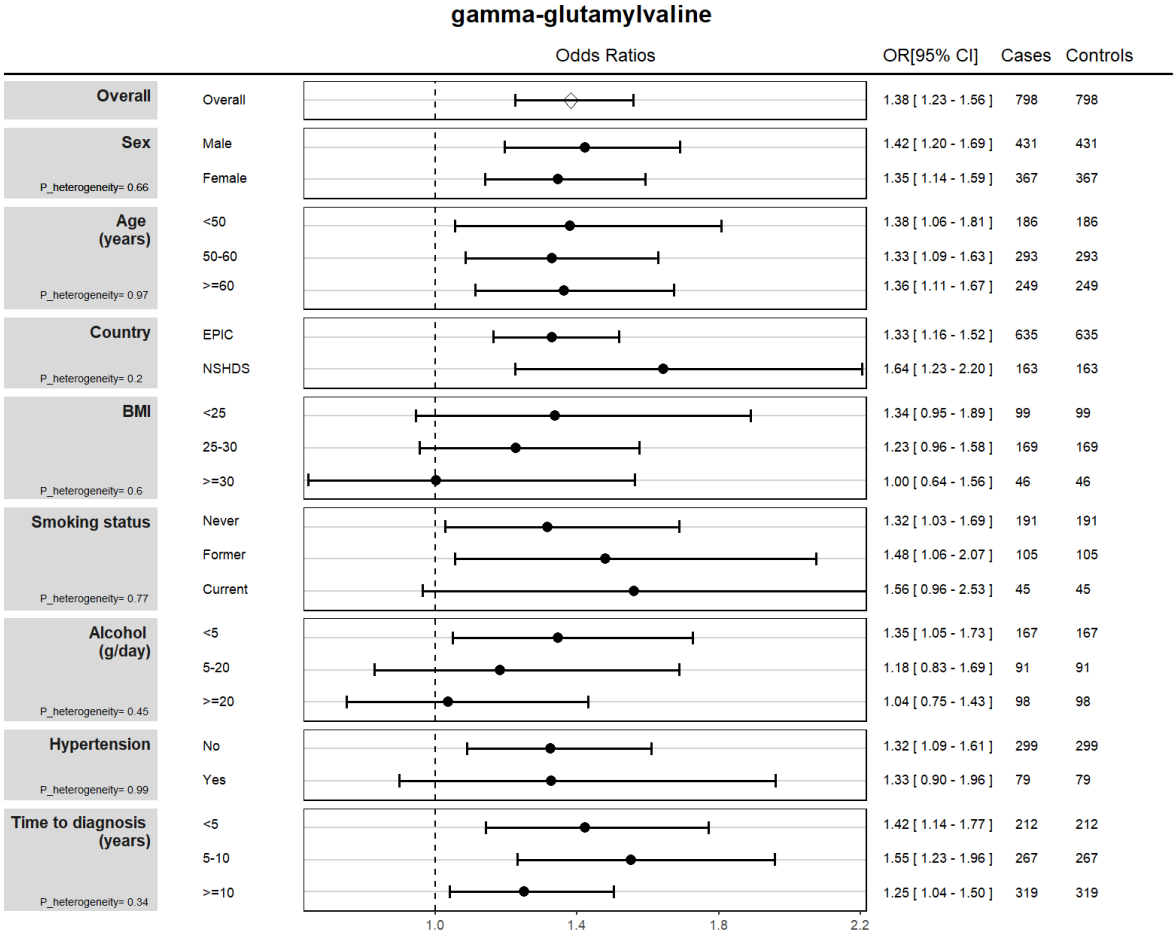
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S22. Forest plots depicts the kidney cancer risk association for gamma-glutamylisoleucine*, stratified by risk factors.



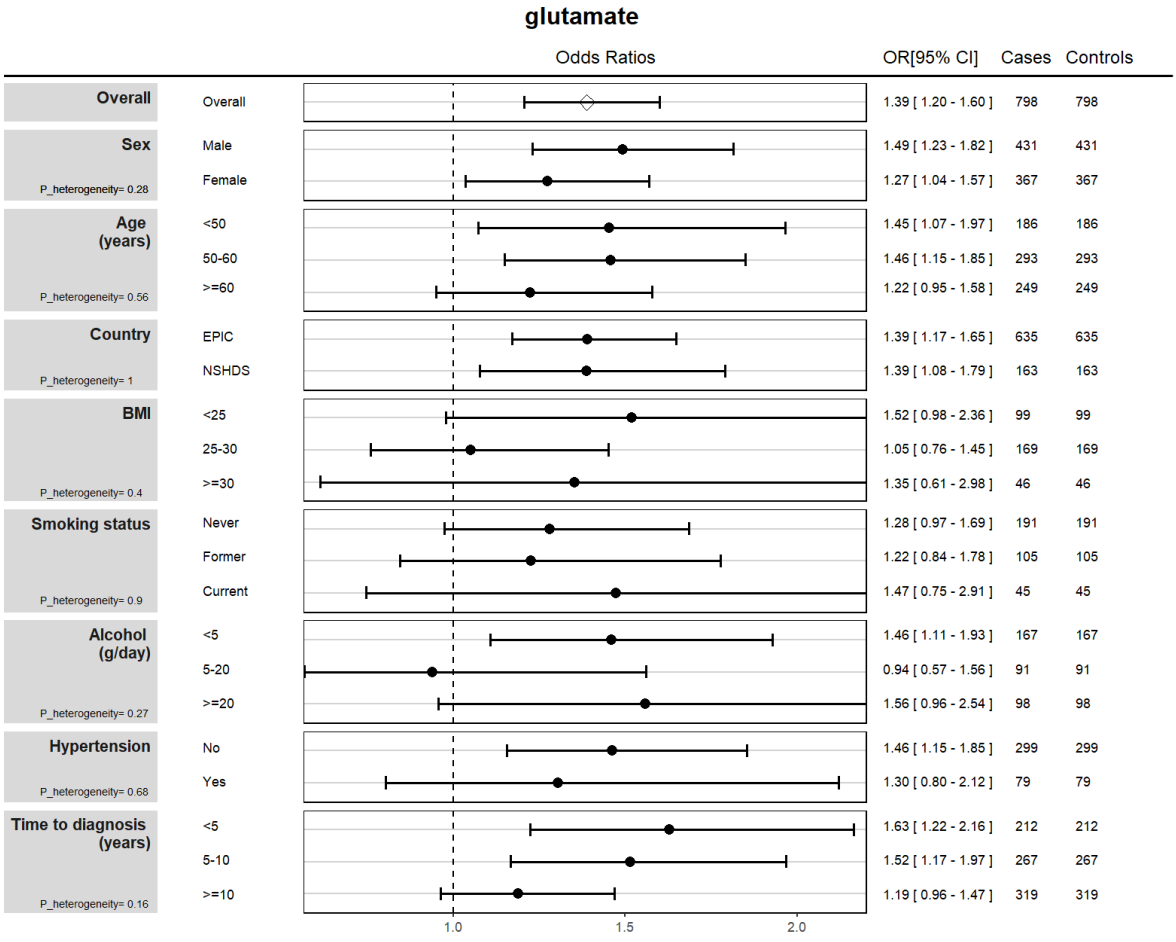
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.
* metabolite identity not yet confirmed by comparison with an authentic chemical standard

Supplementary Figure S23. Forest plots depicts the kidney cancer risk association for gamma-glutamylvaline, stratified by risk factors.



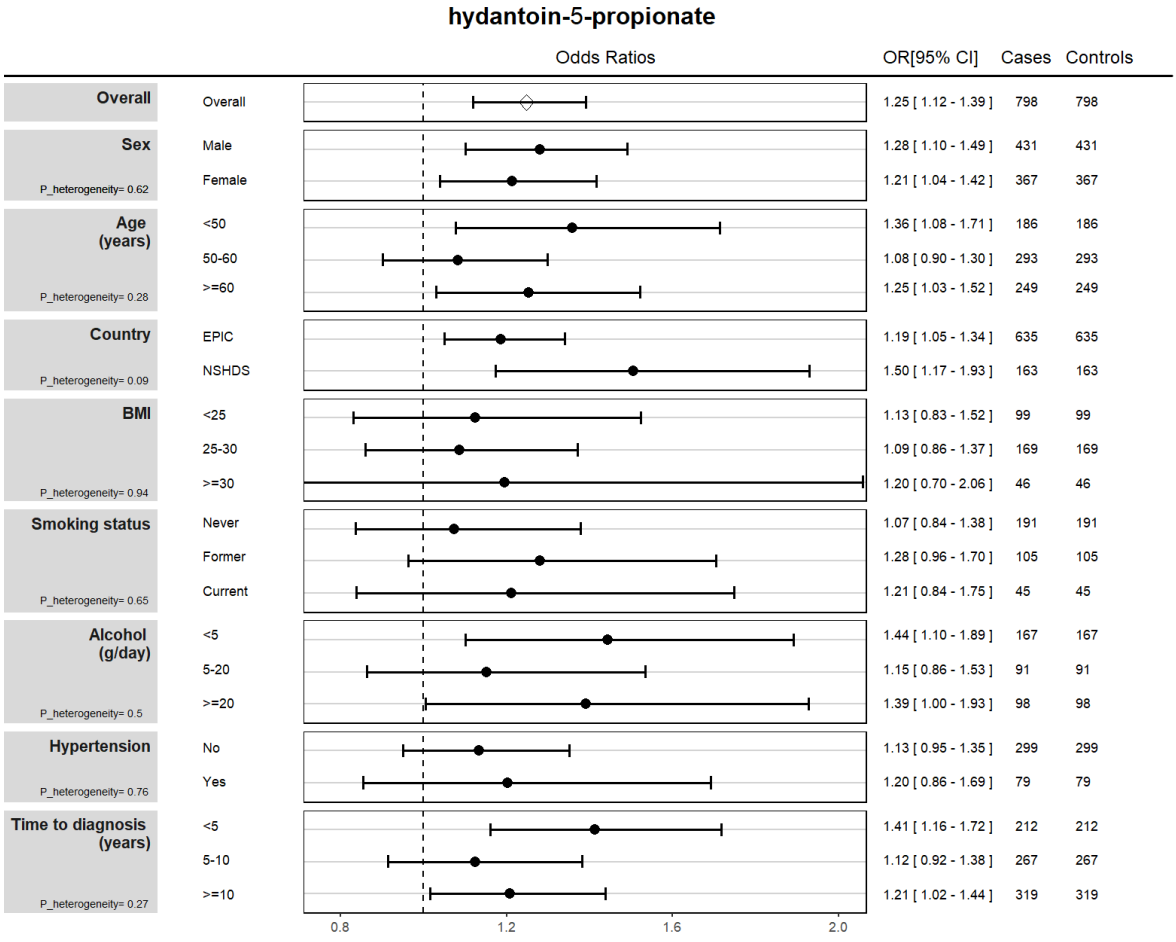
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S24. Forest plots depicts the kidney cancer risk association for glutamate (Metabolon), stratified by risk factors.



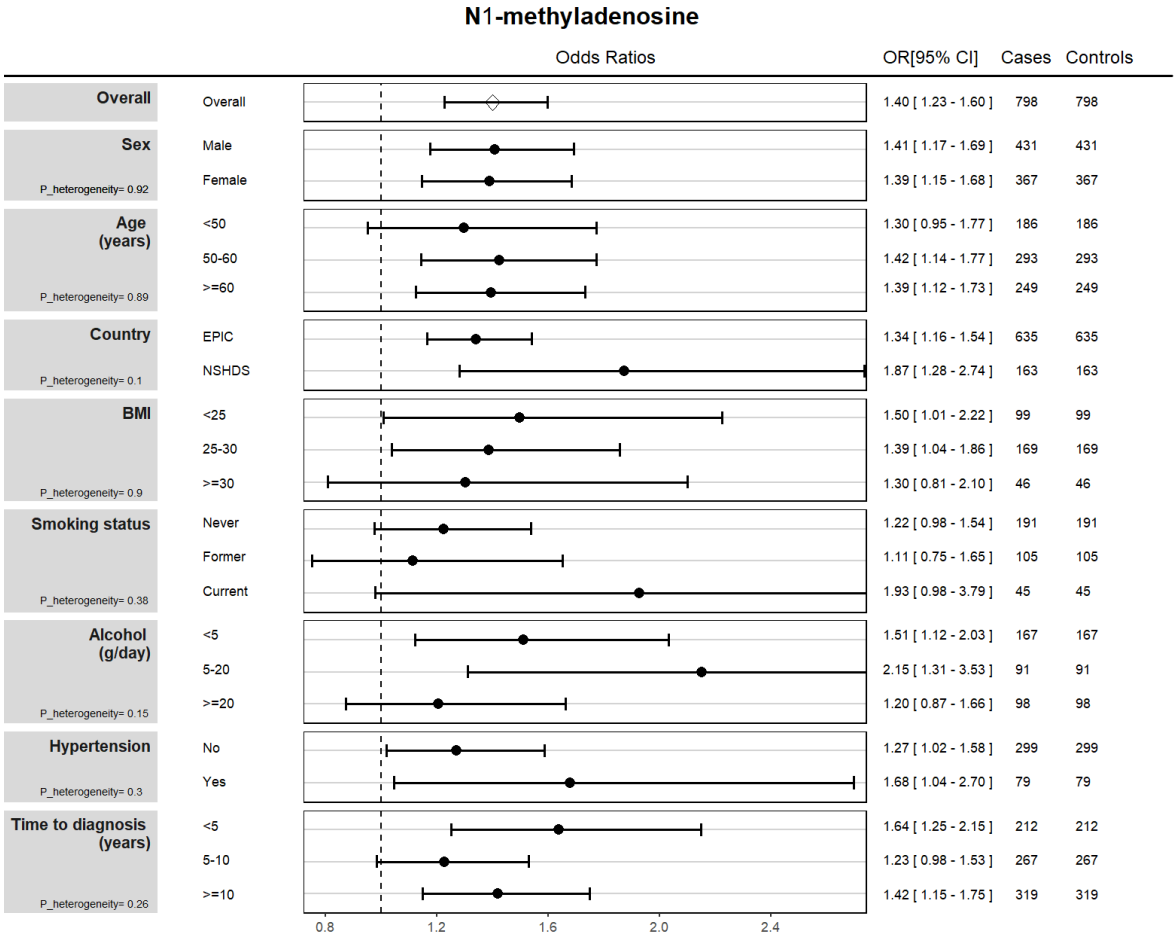
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S25. Forest plots depicts the kidney cancer risk association for hydantoin-5-propionate, stratified by risk factors.



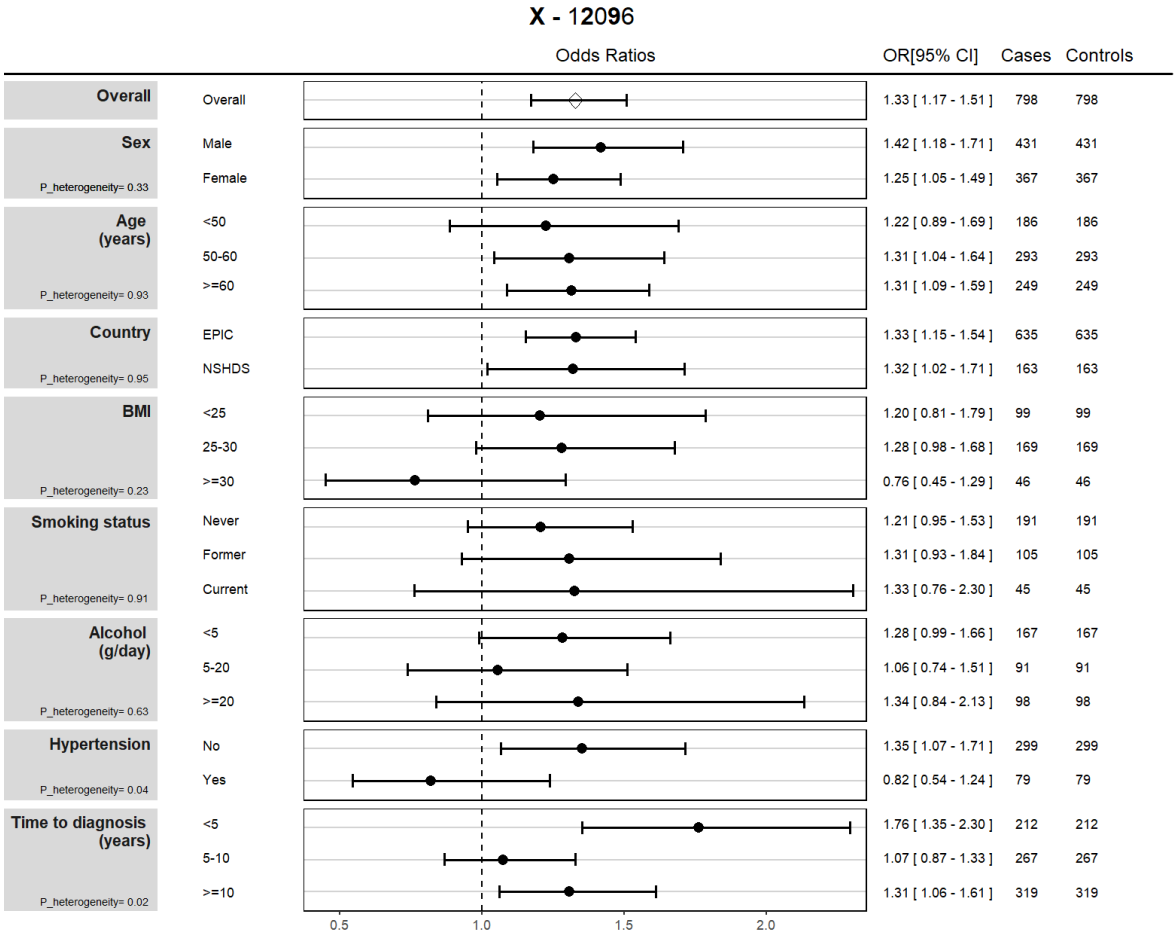
BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S26. Forest plots depicts the kidney cancer risk association for N1-methyladenosine, stratified by risk factors.



BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figure S27. Forest plots depicts the kidney cancer risk association for X-12096, stratified by risk factors.

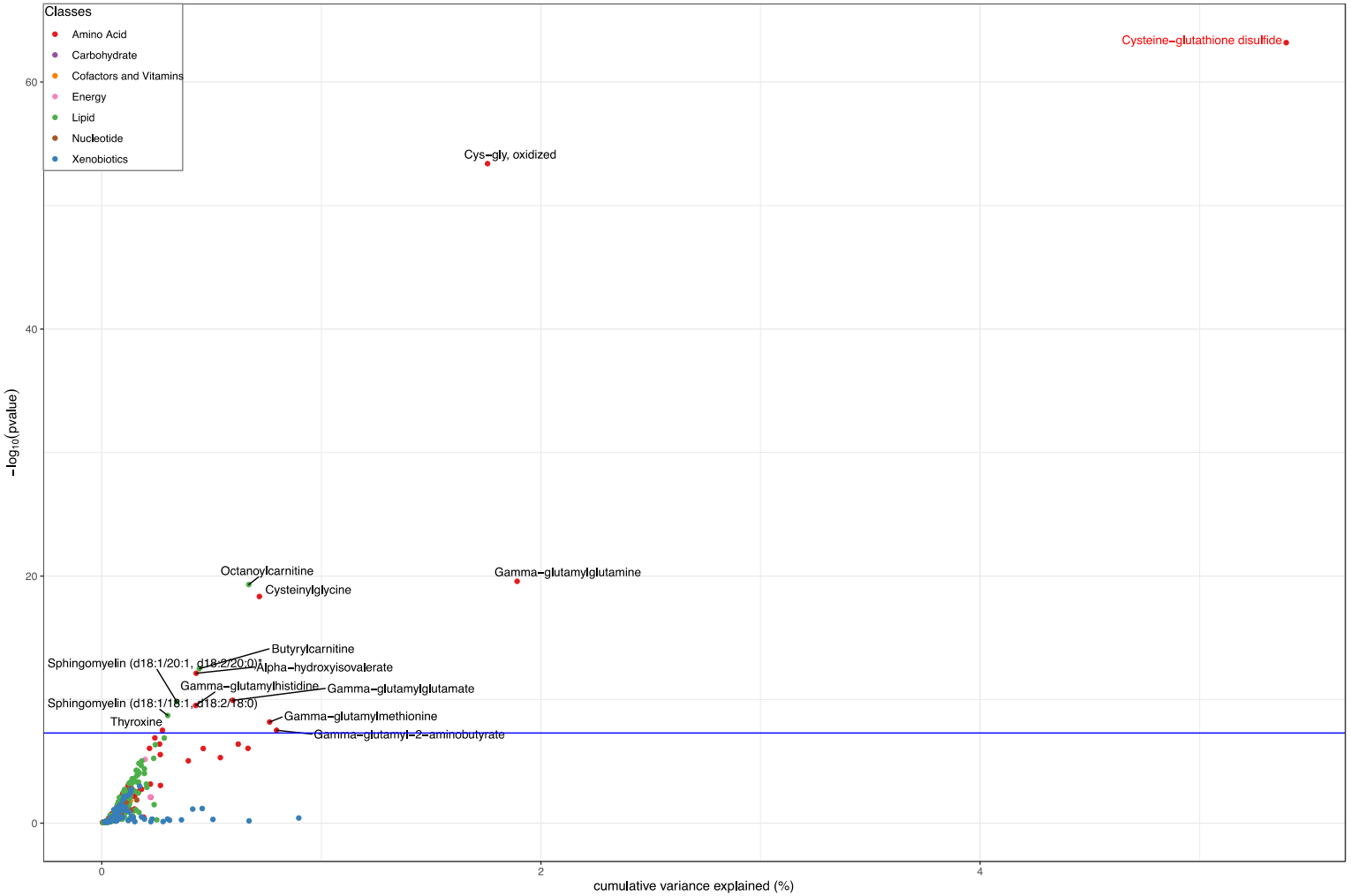


BMI: Body Mass Index; CI: Confidence Interval; d: days; g: grams; N.: number of participants; OR: Odds Ratio.

Supplementary Figures S28-44. Scatter plots of the cumulative variance explained in the Metabolon/Biocrates metabolites by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for the specified risk metabolite (labelled in red)

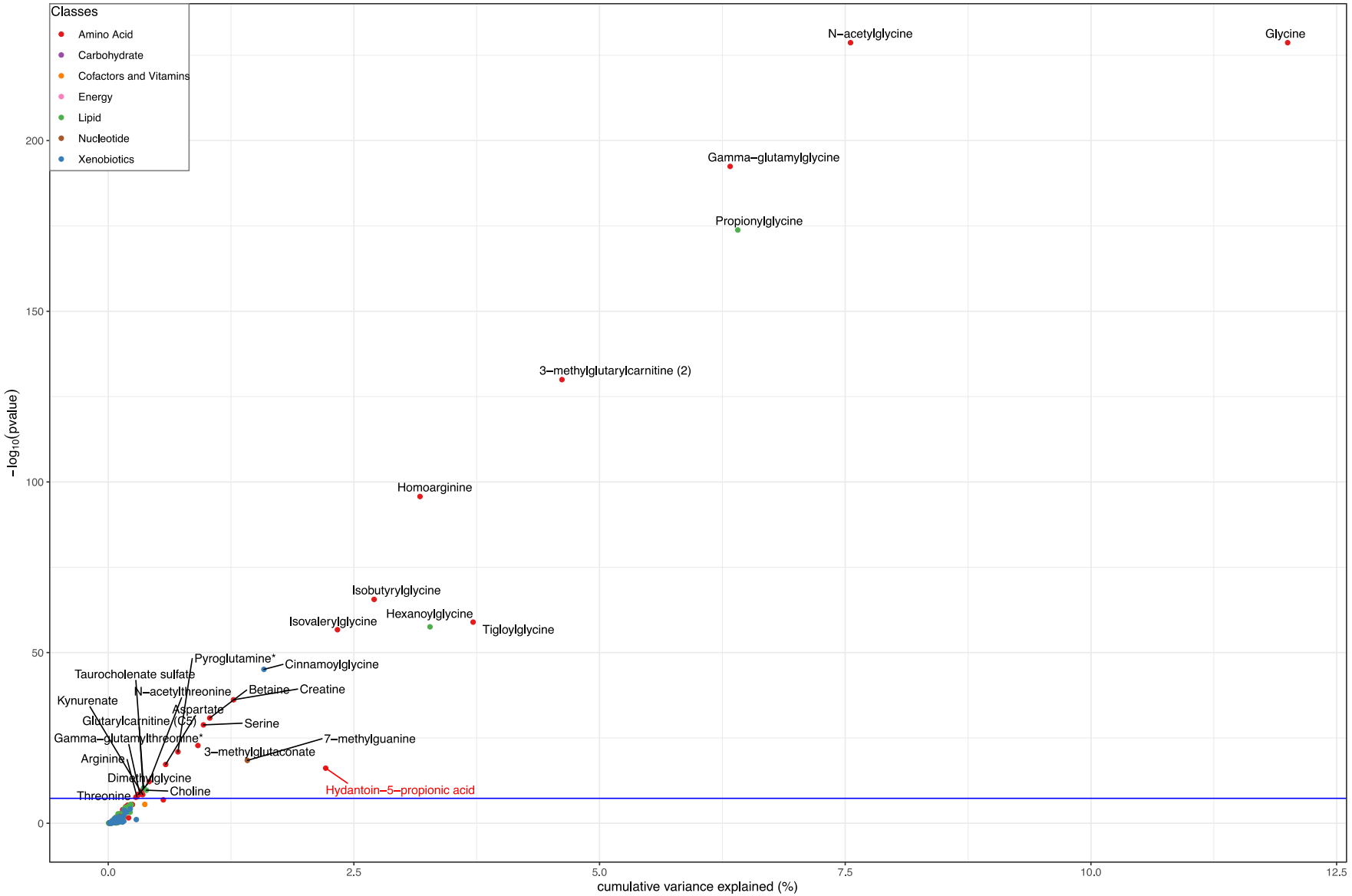
Supplementary Figure S28. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for cysteine-glutathione disulfide (Metabolon).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-08).



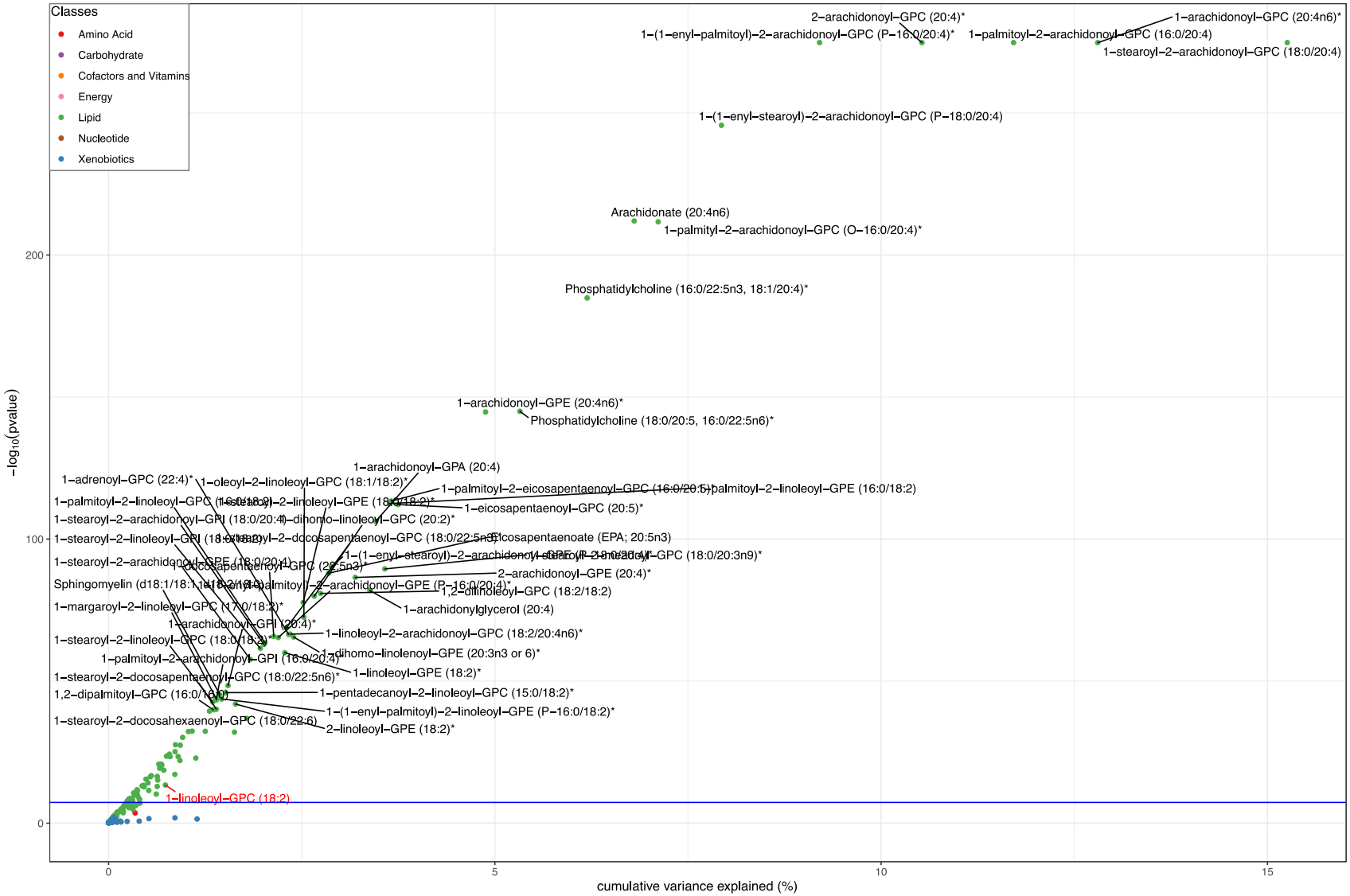
Supplementary Figure S29. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for Hydantoin-5-propionate (Metabolon).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-08).



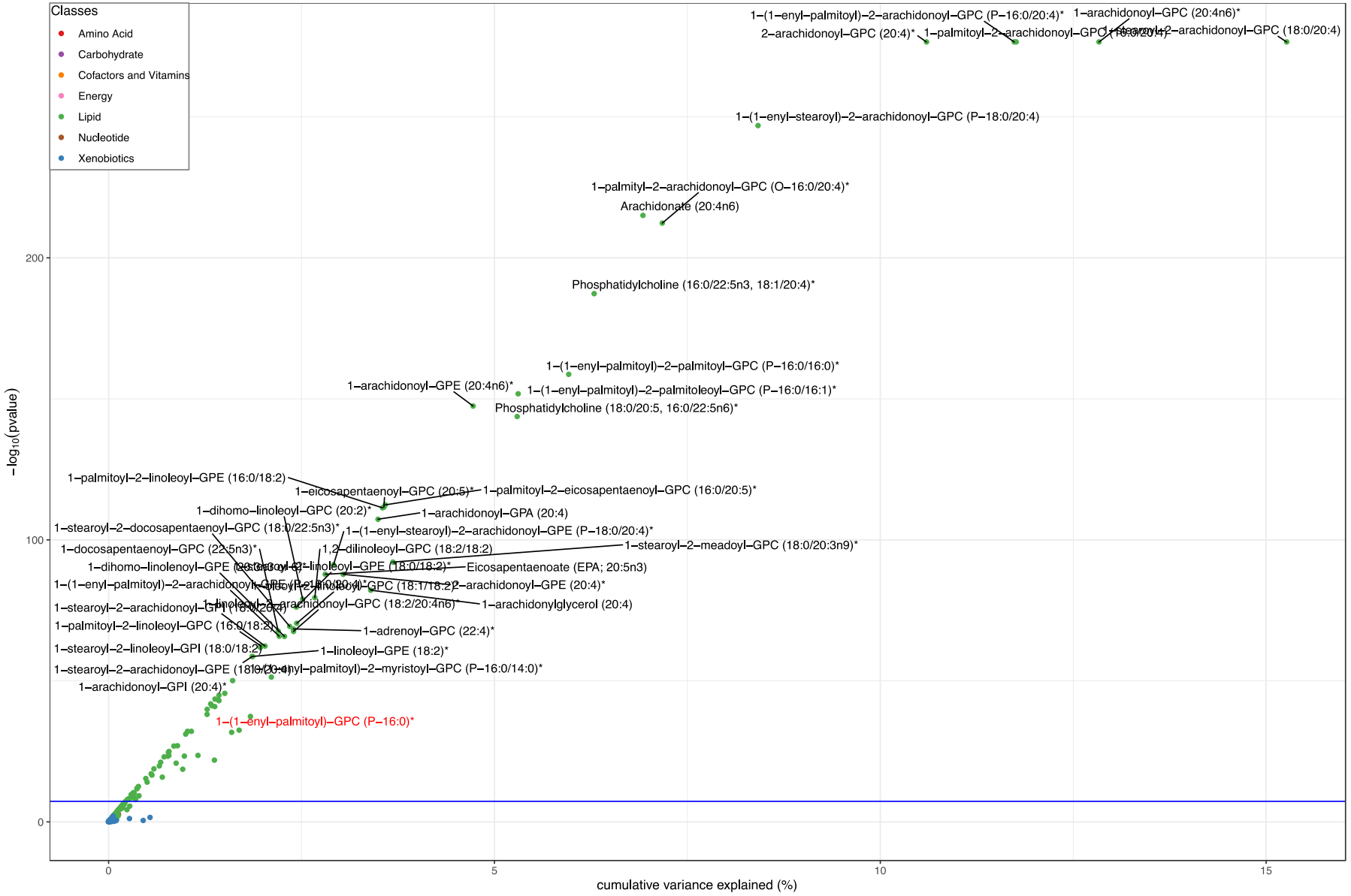
Supplementary Figure S30. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for 1-linoleoyl-GPC (18:2) (Metabolon).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-40).


Classes
Amino Acid
Carbohydrate
Cofactors and Vitamins
Energy
Lipid
Nucleotide
Xenobiotics
2−arachidonoyl−GPC (20:4)*
1−arachidonoyl−GPC (20:4n6)*
1−(1−enyl−palmitoyl)−2−arachidonoyl−GPC (P−16:0/20:4)*
1−palmitoyl−2−arachidonoyl−GPC (16:0/20:4)
1−stearoyl−2−arachidonoyl−GPC (18:0/20:4)
1−(1−enyl−stearoyl)−2−arachidonoyl−GPC (P−18:0/20:4)
Arachidonate (20:4n6)
1−palmityl−2−arachidonoyl−GPC (O−16:0/20:4)*
Phosphatidylcholine (16:0/22:5n3, 18:1/20:4)*
1−arachidonoyl−GPE (20:4n6)*
Phosphatidylcholine (18:0/20:5, 16:0/22:5n6)*
1−arachidonoyl−GPA (20:4)
1−adrenoyl−GPC (22:4)*
1−oleoyl−2−linoleoyl−GPC (18:1/18:2)*
1−palmitoyl−2−eicosapentaenoyl−GPC (16:0/20:5)
palmitoyl−2−linoleoyl−GPE (16:0/18:2)
1−palmitoyl−2−linoleoyl−GPC (16:0/18:2)*
1−stearoyl−2−linoleoyl−GPE (18:0/18:2)*
1−eicosapentaenoyl−GPC (20:5)*
1−stearoyl−2−arachidonoyl−GPI (18:0/20:4)
dihomo−linoleoyl−GPC (20:2)*
1−stearoyl−2−linoleoyl−GPI (18:0/18:2)
1−stearoyl−2−docosapentaenoyl−GPC (18:0/22:5n3)*
eicosapentaenoate (EPA; 20:5n3)
1−stearoyl−2−arachidonoyl−GPE (18:0/20:4)
1−(1−enyl−stearoyl)−2−arachidonoyl−GPC (P−18:0/20:4)*
docosapentaenoyl−GPC (22:5n3)*
2−arachidonoyl−GPE (20:4)*
Sphingomyelin (d18:1/18:1(11Z)/18:2(9,12))
1−palmitoyl−2−arachidonoyl−GPE (P−16:0/20:4)*
1,2−dilinoleoyl−GPC (18:2/18:2)
1−margaroyl−2−linoleoyl−GPC (17:0/18:2)*
1−arachidonylglycerol (20:4)
1−arachidonoyl−GPI (20:4)*
1−linoleoyl−2−arachidonoyl−GPC (18:2/20:4n6)*
1−stearoyl−2−linoleoyl−GPC (18:0/18:2)
1−dihomo−linolenoyl−GPE (20:3n3 or 6)*
1−palmitoyl−2−arachidonoyl−GPI (16:0/20:4)
1−linoleoyl−GPE (18:2)*
1−stearoyl−2−docosapentaenoyl−GPC (18:0/22:5n6)*
1,2−dipalmitoyl−GPC (16:0/16:0)
1−pentadecanoyl−2−linoleoyl−GPC (15:0/18:2)*
1−(1−enyl−palmitoyl)−2−linoleoyl−GPE (P−16:0/18:2)*
1−stearoyl−2−docosahexaenoyl−GPC (18:0/22:6)
2−linoleoyl−GPE (18:2)*
1−linoleoyl−GPC (18:2)
200
100
0
0
5
10
15
−log10(pvalue)
cumulative variance explained (%)

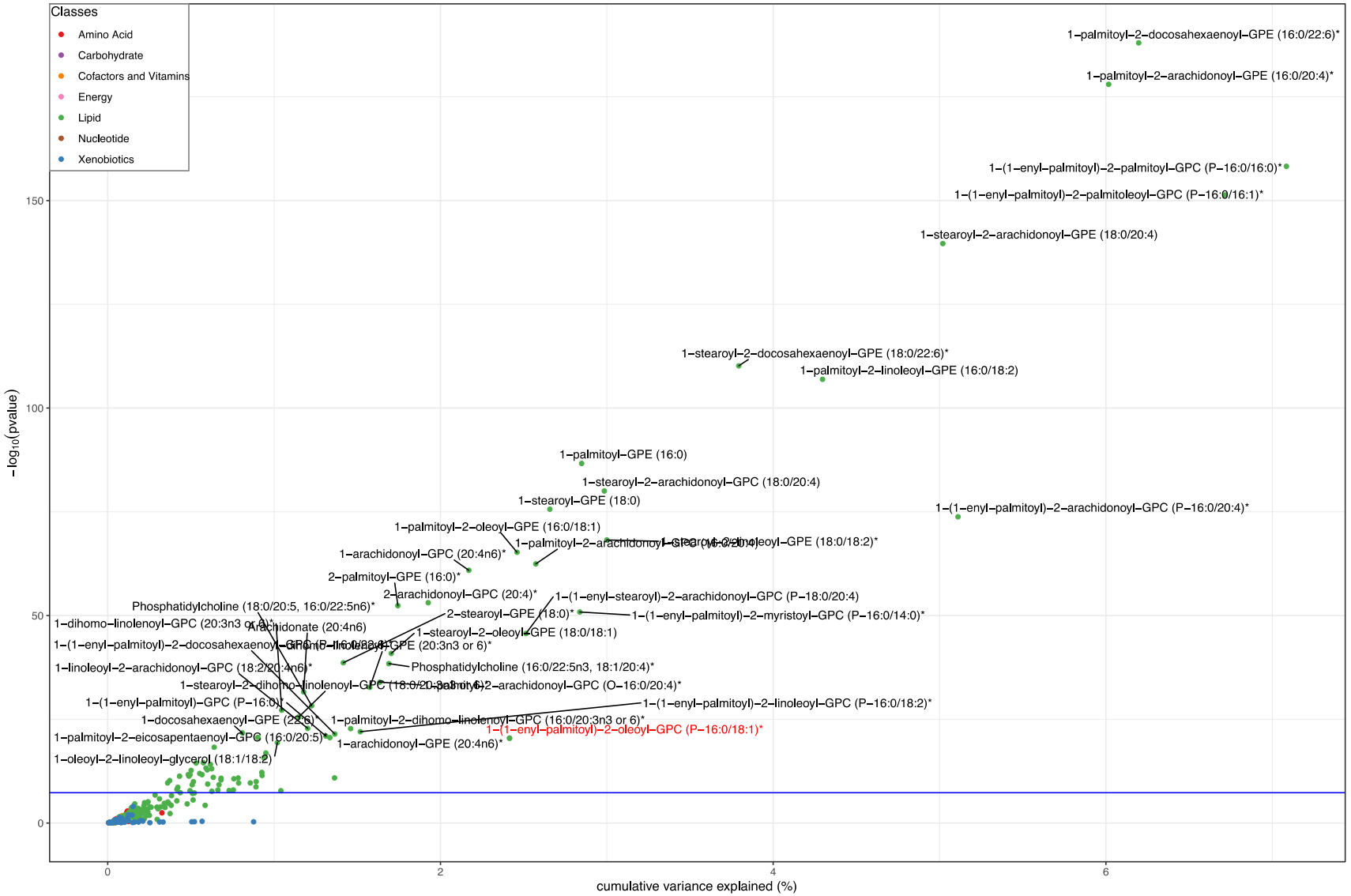
Supplementary Figure S31. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for 1-(1-enyl-palmitoyl)-GPC (P-16:0) (Metabolon).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-50).



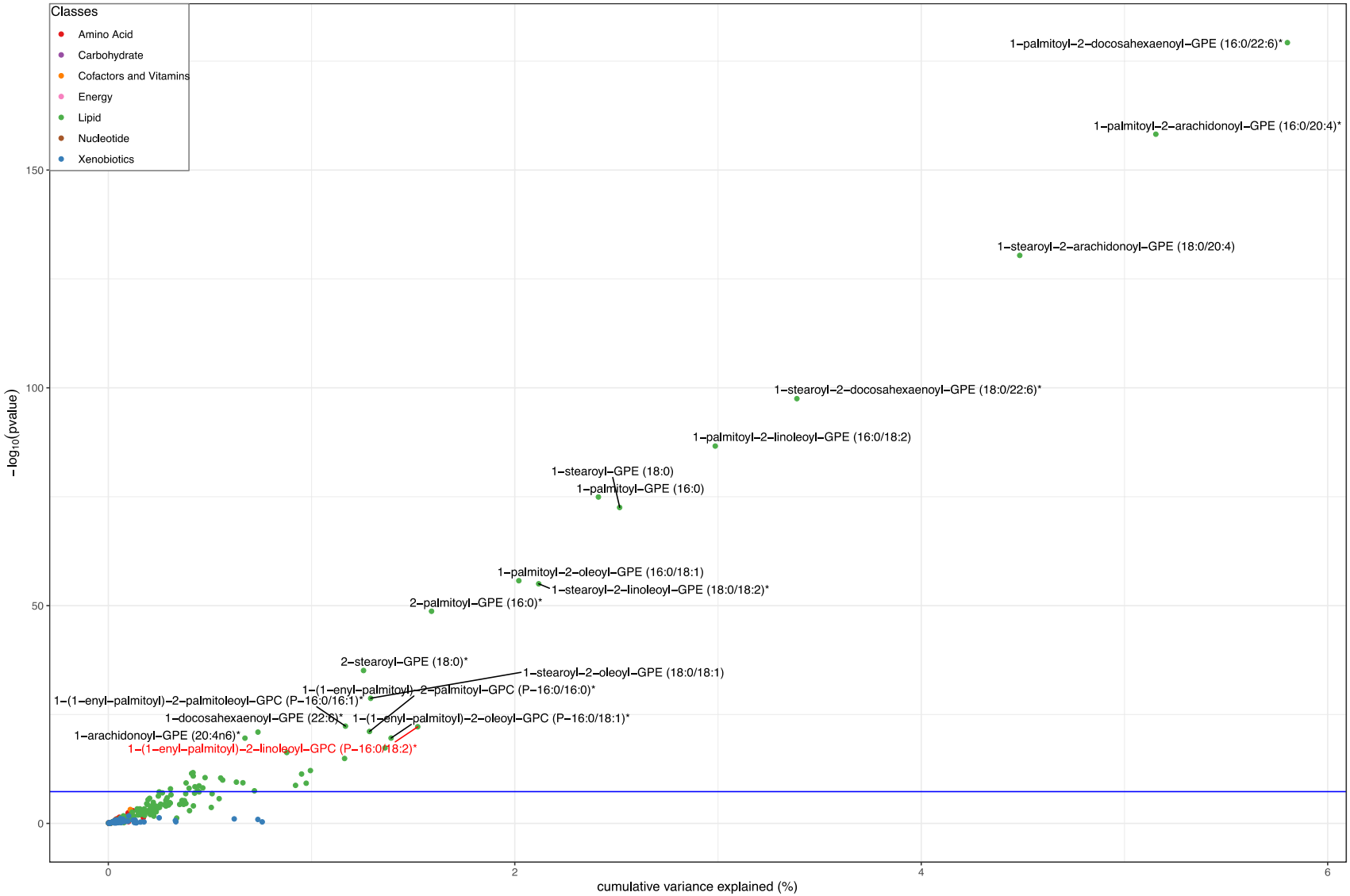
Supplementary Figure S32. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for 1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1) (Metabolon).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-20).



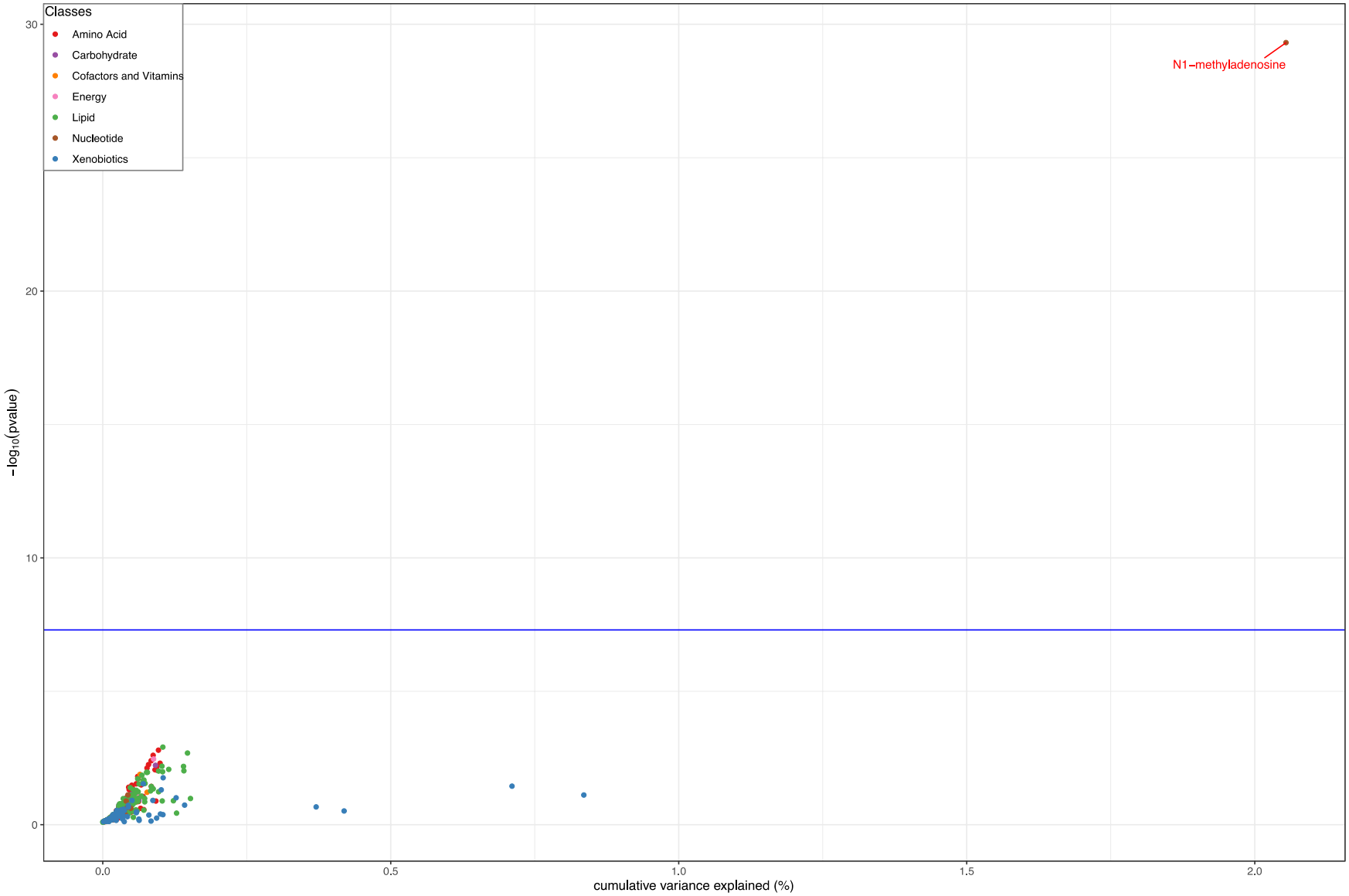
Supplementary Figure S33. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for 1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2) (Metabolon).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-20).



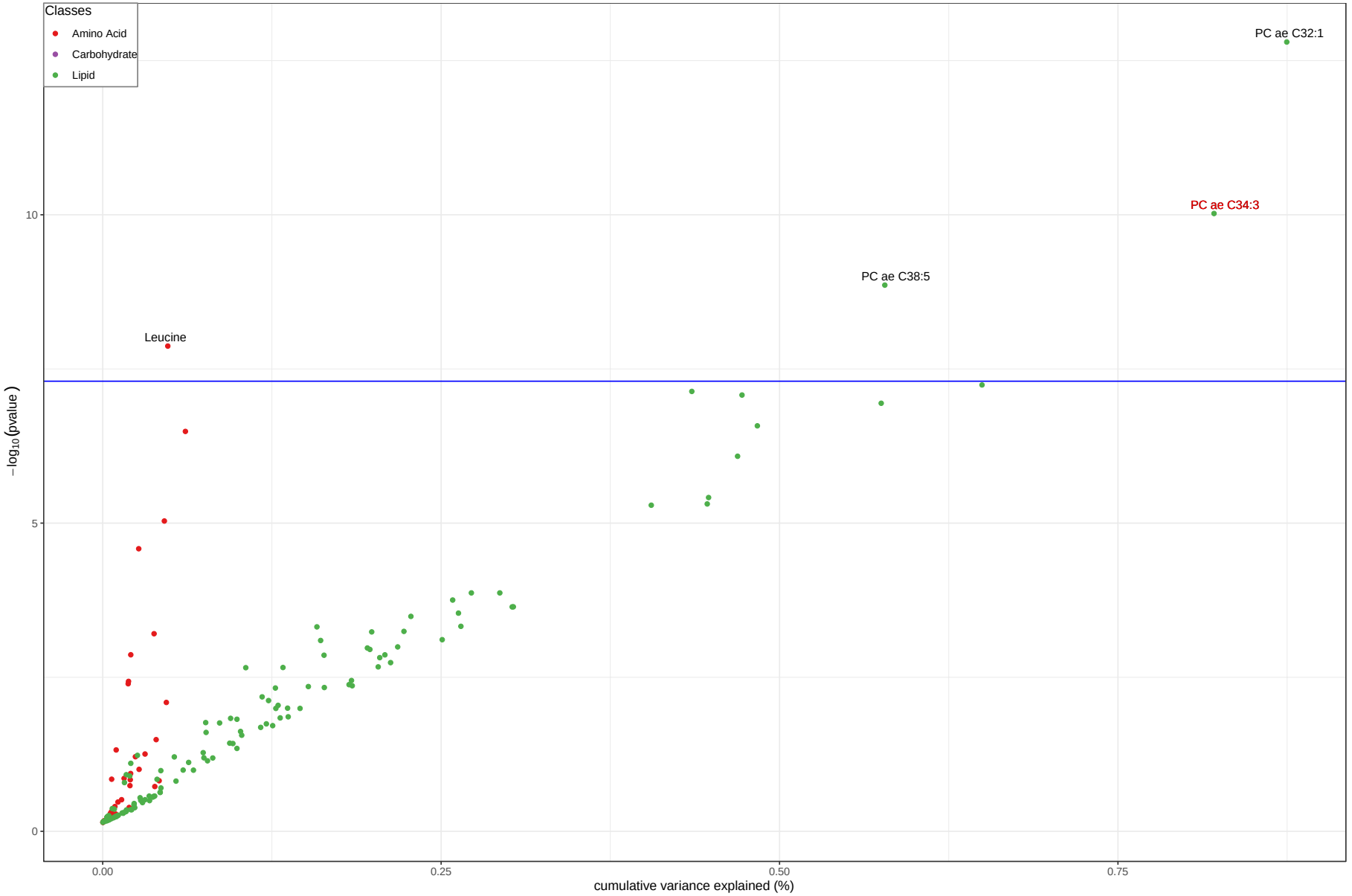
Supplementary Figure S34. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for N1-methyladenosine (Metabolon).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-08).


Classes
Amino Acid
Carbohydrate
Cofactors and Vitamins
Energy
Lipid
Nucleotide
Xenobiotics
N1−methyladenosine
30
20
10
0
−log10(pvalue)
0.0
0.5
1.0
1.5
2.0
cumulative variance explained (%)

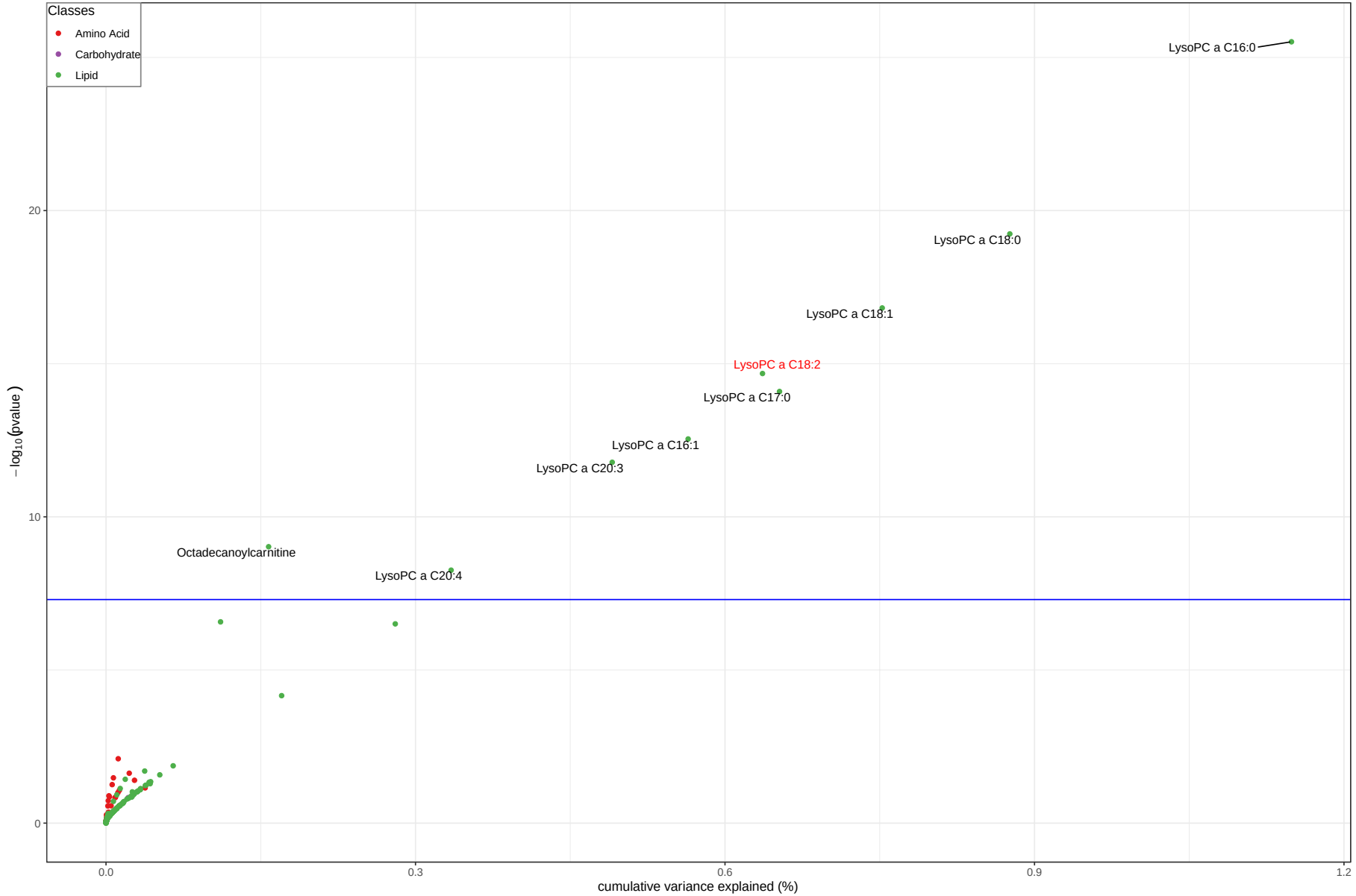
Supplementary Figure S35. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for PC ae C34:3 (Biocrates).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-08).


Classes
Amino Acid
Carbohydrate
Lipid
PC ae C32:1
PC ae C34:3
PC ae C38:5
Leucine
10
5
0
−log10(pvalue)
0.00
0.25
0.50
0.75
cumulative variance explained (%)

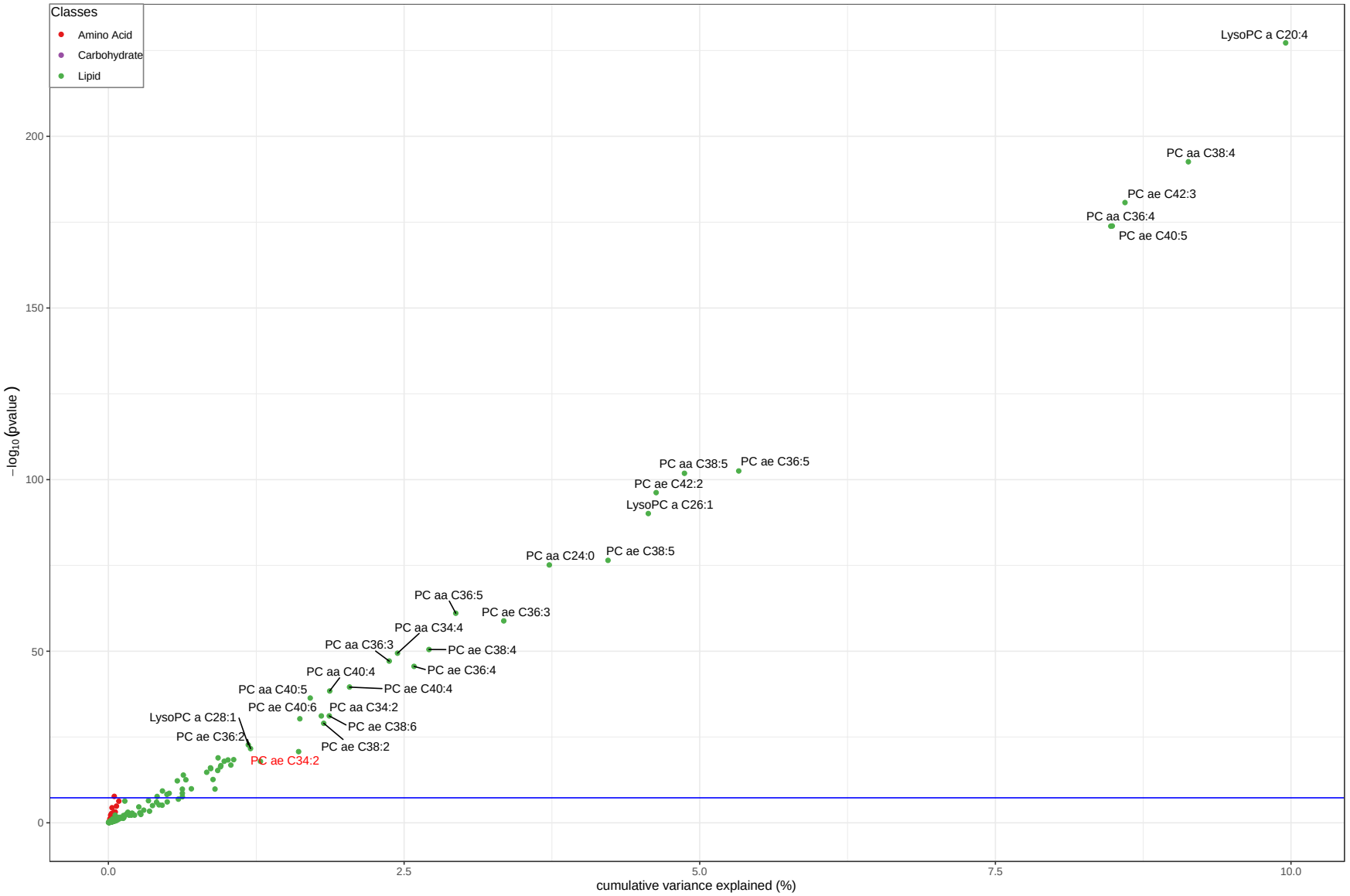
Supplementary Figure S36. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for lysoPC a C18:2 (Biocrates).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-08).



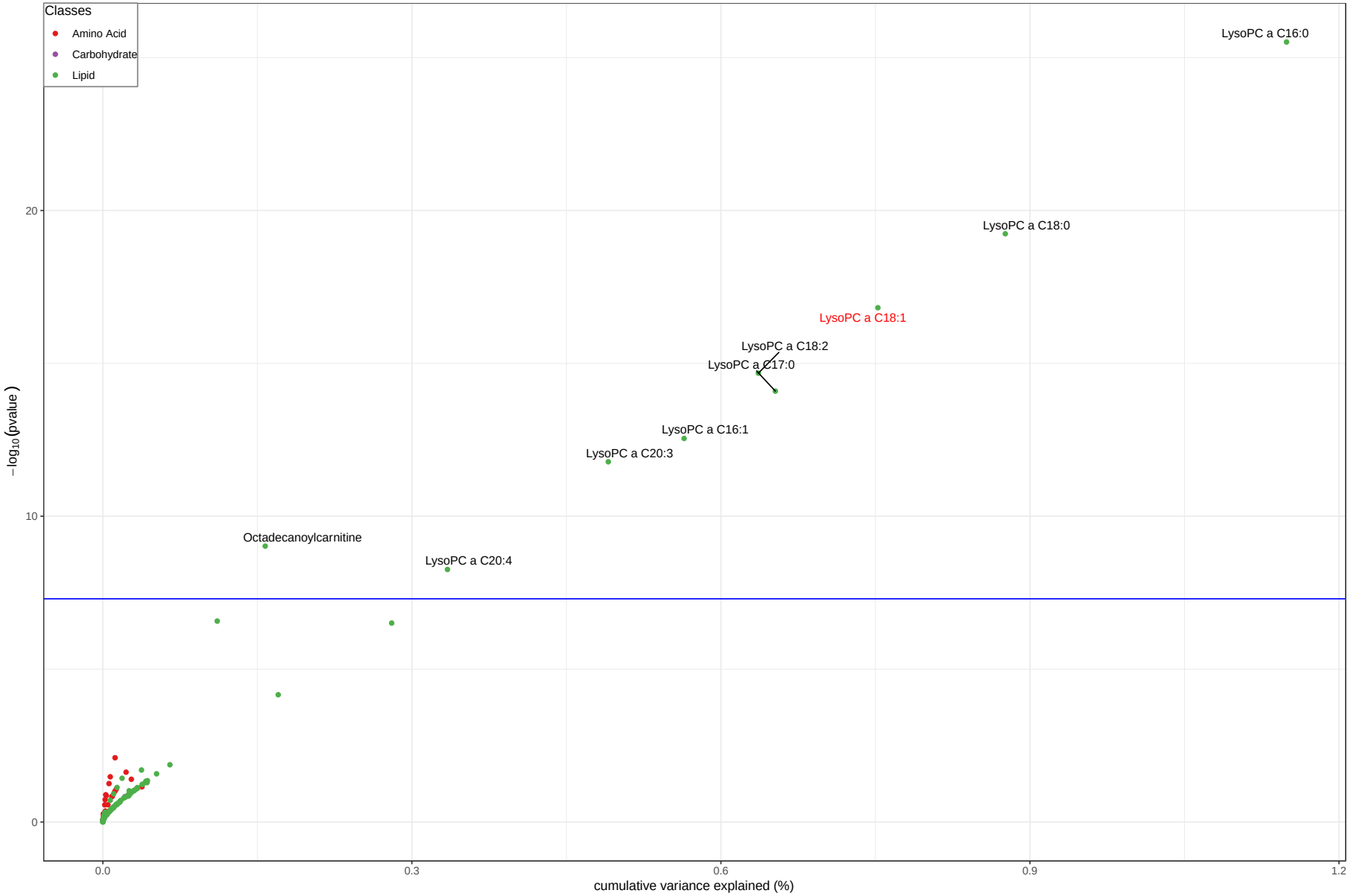
Supplementary Figure S37. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for PC ae C34:2 (Biocrates).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-20).



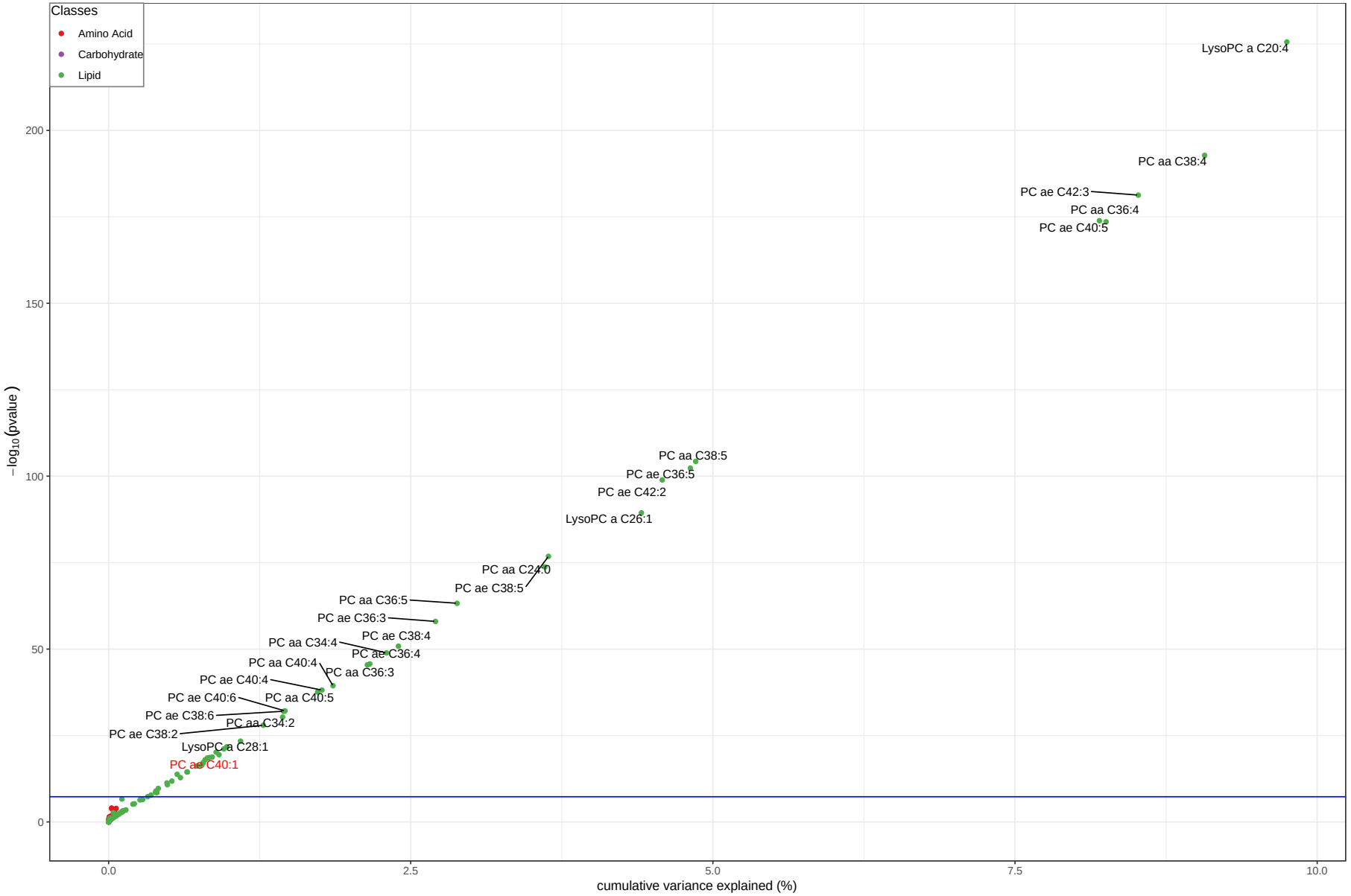
Supplementary Figure S38. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for lysoPC a C18:1 (Biocrates).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-08).


Classes
Amino Acid
Carbohydrate
Lipid
LysoPC a C16:0
LysoPC a C18:0
LysoPC a C18:1
LysoPC a C18:2
LysoPC a C17:0
LysoPC a C16:1
LysoPC a C20:3
Octadecanoylcarnitine
LysoPC a C20:4
20
10
0
0.0
0.3
0.6
0.9
1.2
−log10 (pvalue)
cumulative variance explained (%)

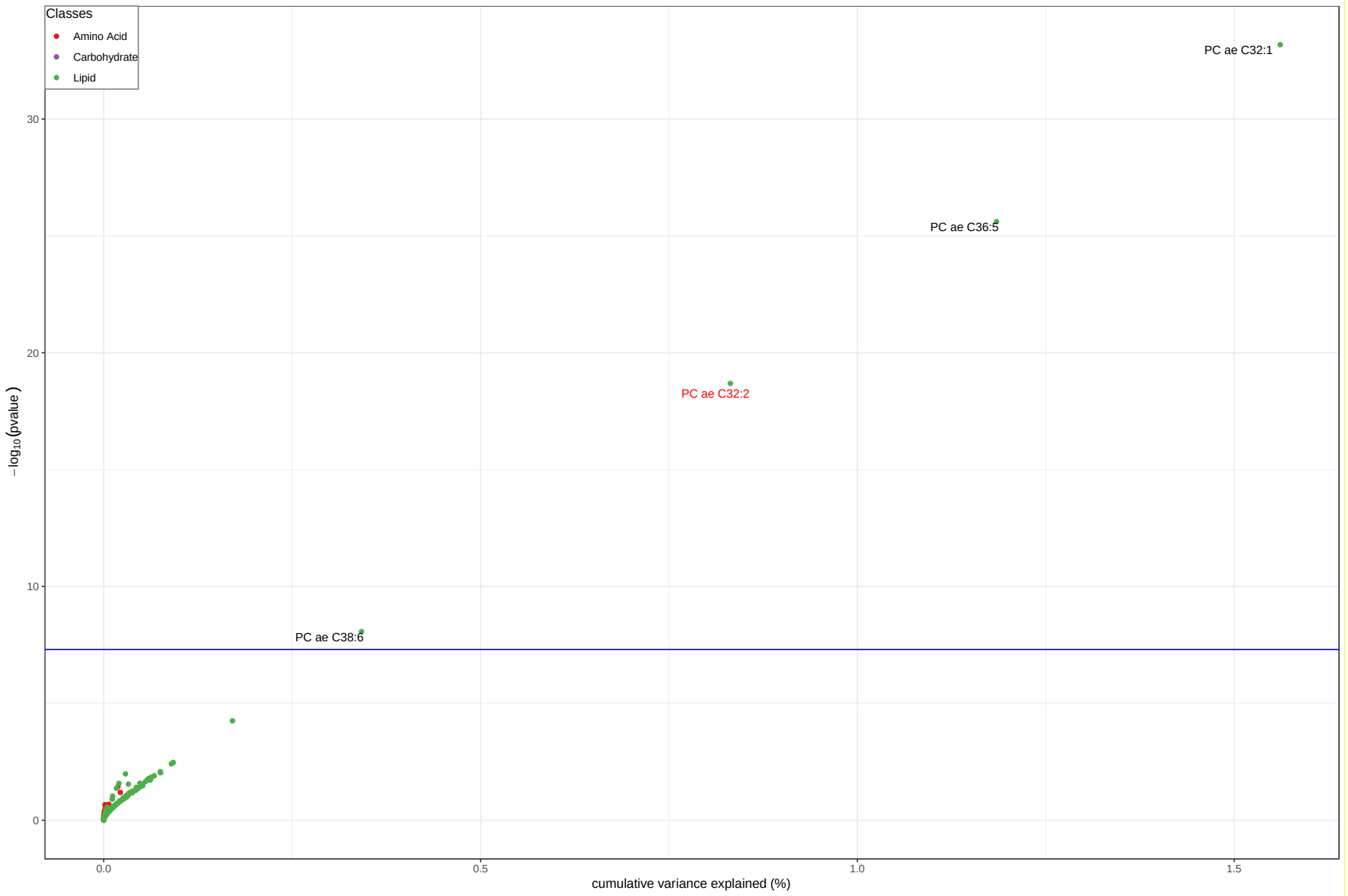
Supplementary Figure S39. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for PC ae C40:1 (Biocrates).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-10).


Classes
Amino Acid
Carbohydrate
Lipid
LysoPC a C20:4
PC aa C38:4
PC ae C42:3
PC aa C36:4
PC ae C40:5
PC aa C38:5
PC ae C36:5
PC ae C42:2
LysoPC a C26:1
PC aa C24:0
PC ae C38:5
PC aa C36:5
PC ae C36:3
PC aa C34:4
PC ae C38:4
PC ae C36:4
PC aa C40:4
PC aa C36:3
PC ae C40:4
PC ae C40:6
PC aa C40:5
PC ae C38:6
PC aa C34:2
PC ae C38:2
LysoPC a C28:1
PC aa C40:1
200
150
100
50
0
0.0
2.5
5.0
7.5
10.0
−log10(pvalue)
cumulative variance explained (%)

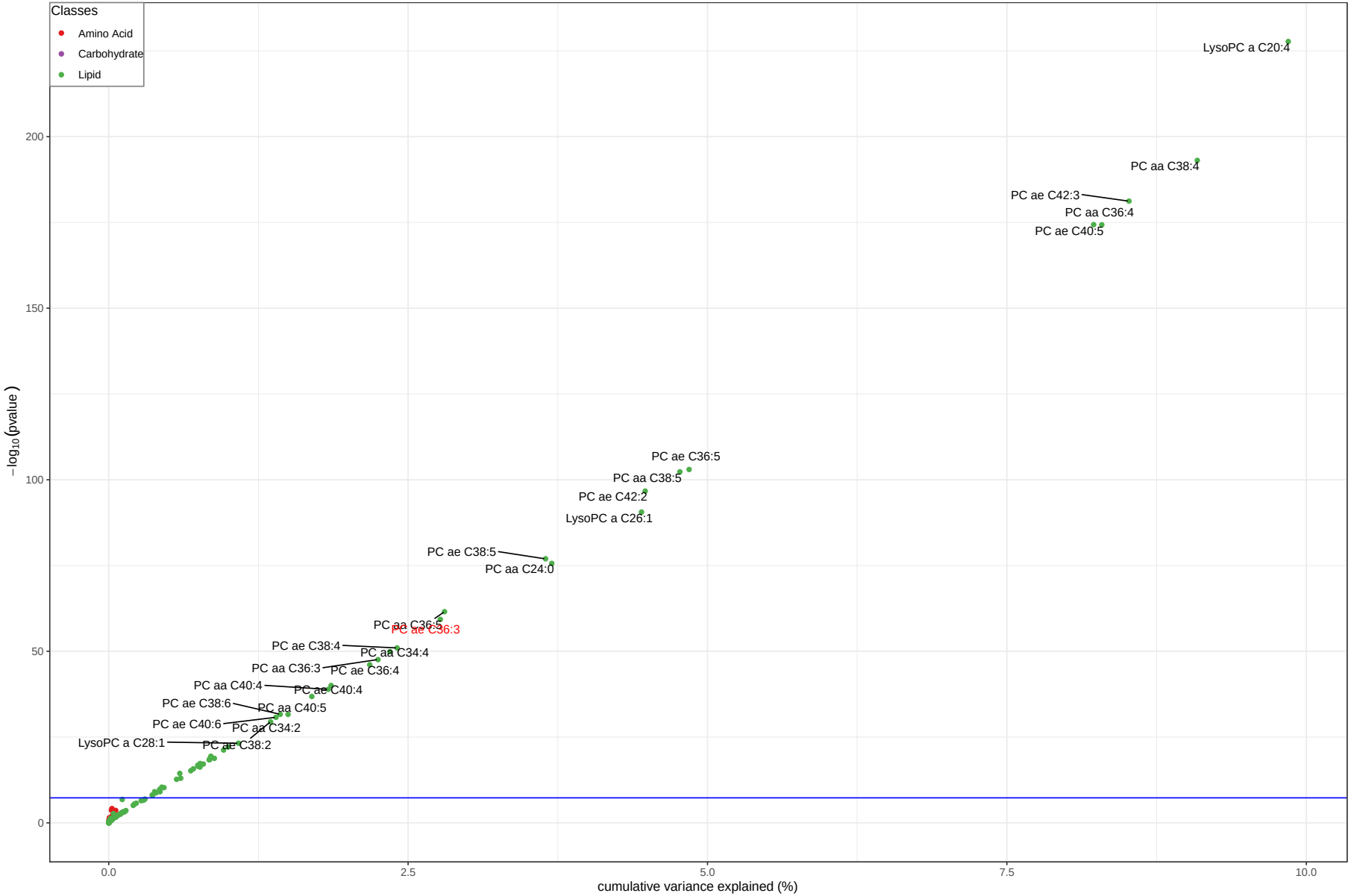
Supplementary Figure S40. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for PC ae C32:2 (Biocrates).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-08).


Classes
Amino Acid
Carbohydrate
Lipid
PC ae C32:1
PC ae C36:5
PC ae C32:2
PC ae C38:6
30
20
10
0
−log10 (pvalue)
0.0
0.5
1.0
1.5
cumulative variance explained (%)

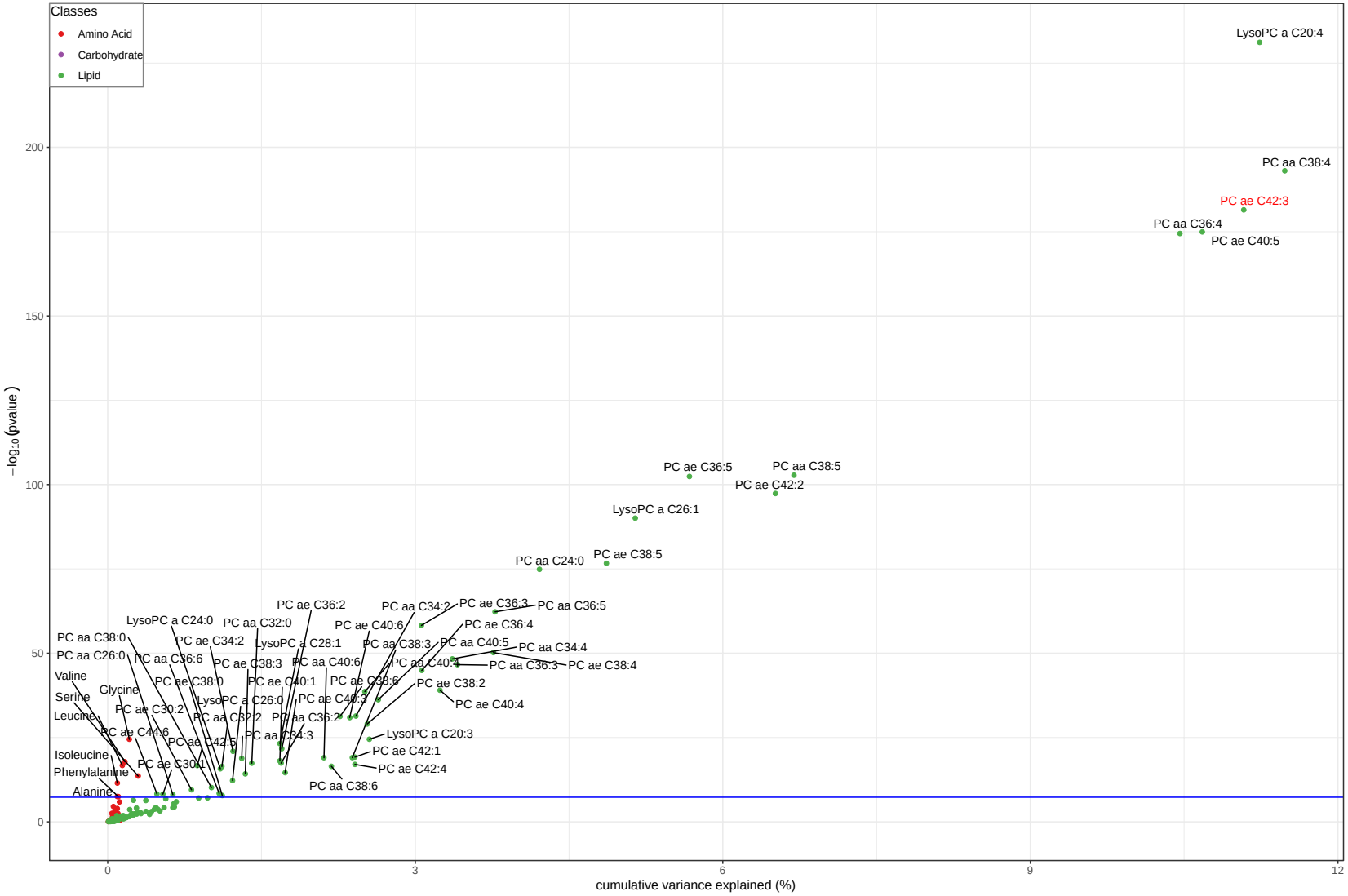
Supplementary Figure S41. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for PC ae C36:3 (Biocrates).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-08).



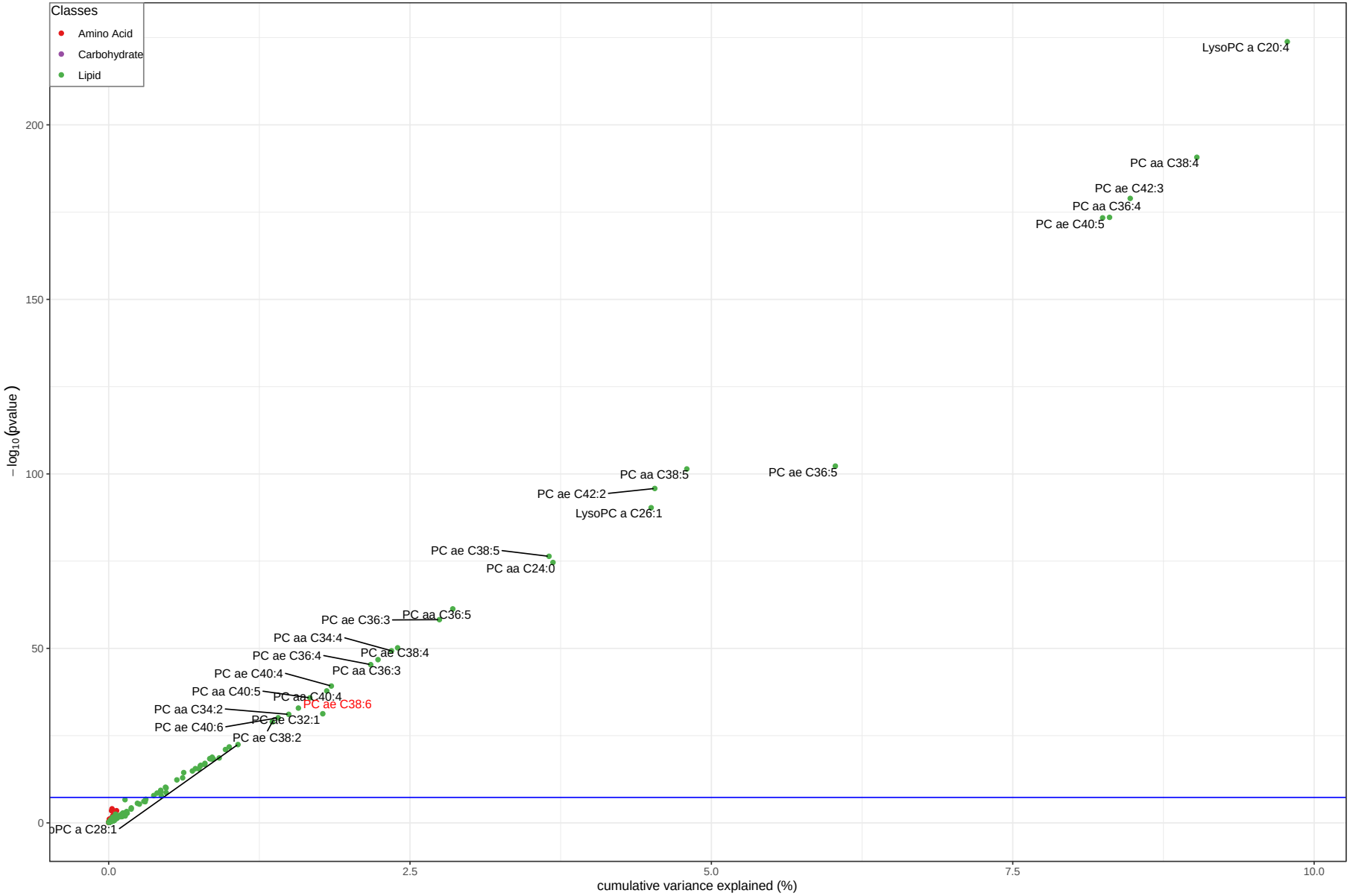
Supplementary Figure S42. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for PC ae C42:3 (Biocrates).

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-08).



Supplementary Figure S43. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for PC ae C38:6 (Biocrates).

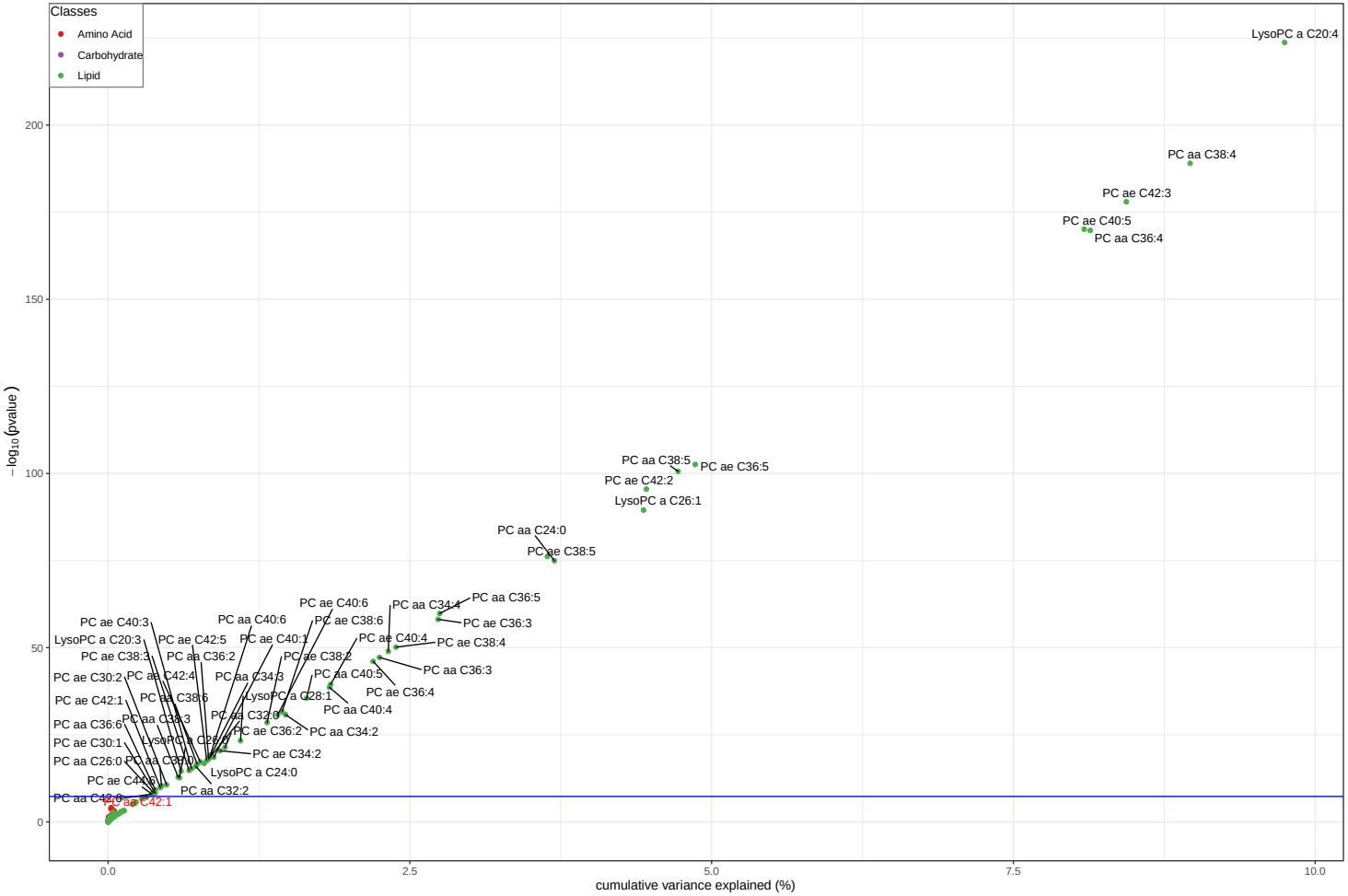
Metabolites that are labelled have a p value below the genom



e-wide significance threshold (p<5E-08).

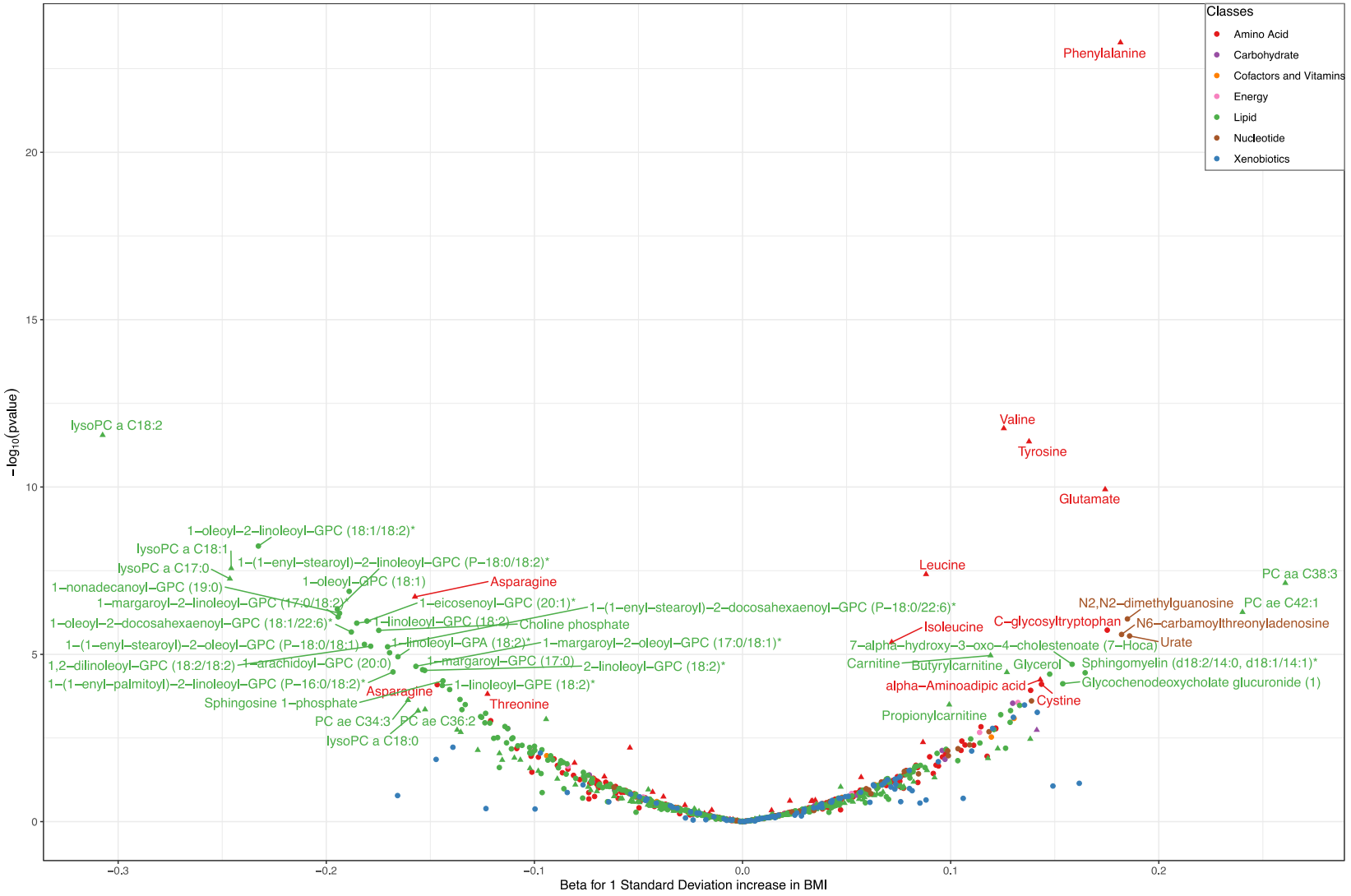
Supplementary Figure S44. Scatter plots of the cumulative variance explained by the genome-wide significant (p<5x10-8) independent (R2<0.01) single nucleotide polymorphisms (SNPs) for PC aa C42:1 (Biocrates)

Metabolites that are labelled have a p value below the genome-wide significance threshold (p<5E-08).



Supplementary Figure S45. Volcano plots representing the association between BMI and circulating Biocrates metabolites (triangle) and Metabolon metabolites (dots) from MR analyses.

Metabolites that are labelled have a *p* value below the significance threshold (*p*<0.0003 for Biocrates and *p*<5.48x10-5 for Metabolon*).*

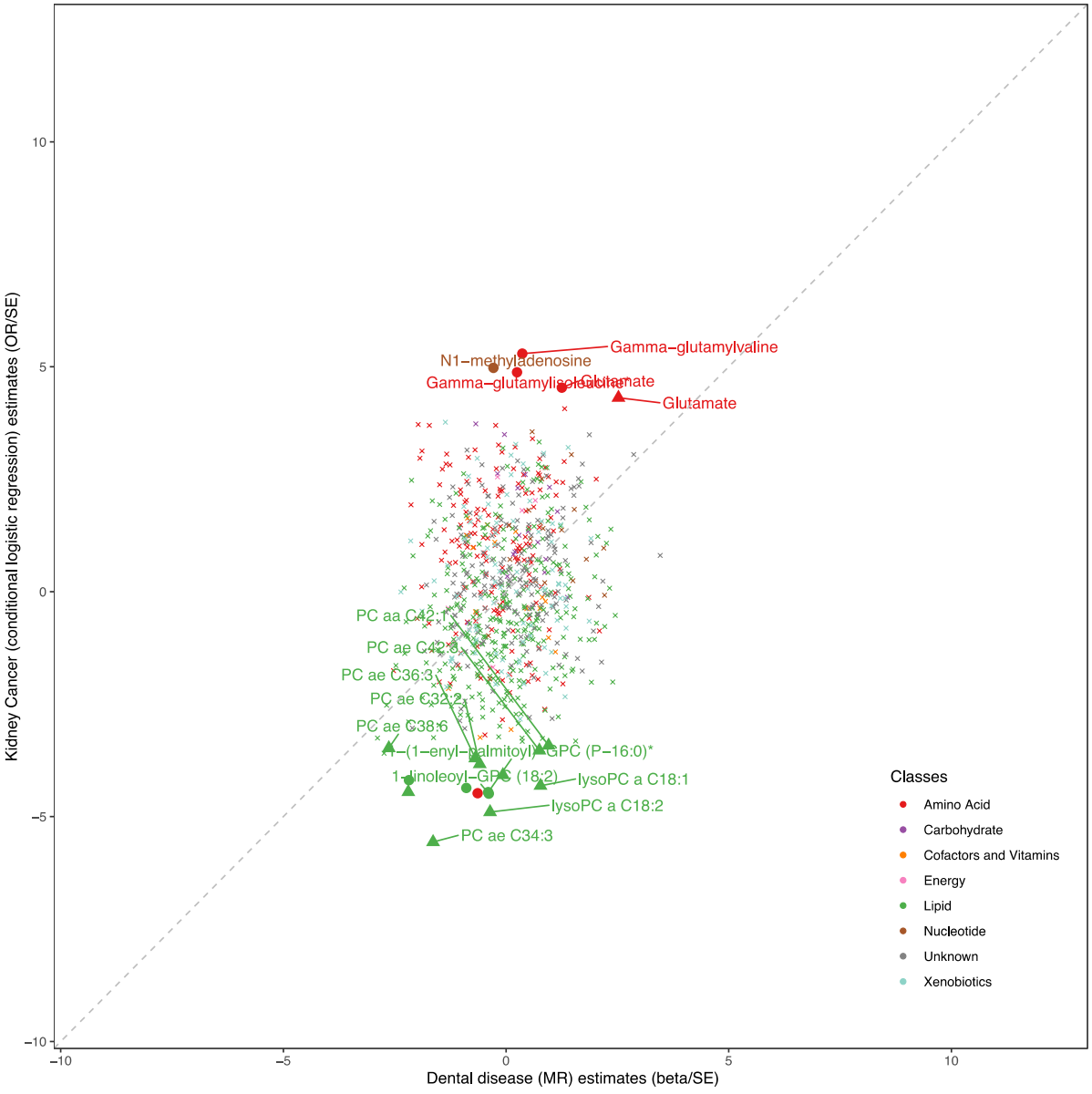


BMI: Body Mass Index

* metabolite identity not yet confirmed by comparison with an authentic chemical standard

Supplementary Figure S46. Scatter plots comparing the metabolite profile associated with kidney cancer from prospective observational analyses with the dental disease-driven metabolite profile from MR analyses.

Z score was calculated by dividing the effect estimate (log OR or beta) by the standard errors. Metabolites that are labelled have a *p* value below the threshold (*p*<0.05/Effective number of tests (ENT)) in the pooled analyses and are nominally significant in at least 2 cohorts separately. Metabolites measured by the Biocrates platform that are below the *p* value threshold are represented by triangles, those measured by the Metabolon platform that are below the *p* value threshold are represented by dots and those that are measured by either the Biocrates or the Metabolon platform that are above the *p* value threshold are represented by an x.



MR: Mendelian Randomization; OR: Odds Ratio; SE: Standard Error.

* metabolite identity not yet confirmed by comparison with an authentic chemical standard

On the y-axis, the OR and SE were derived from the logistic regression analyses conditioned on case set estimating the associations between circulating metabolites and kidney cancer risk in five prospective cohorts.

On the x-axis, the beta and SE were derived from the mendelian randomization analyses evaluating the effect of BMI on circulating metabolites levels.

Supplementary Methods

Table of Contents

Study population

Acronym	Country/Region	Cohort name	N case-control pairs	Biocrates metabolites	Metabolon metabolites
EPIC	EUROPE	European Prospective Investigation into Cancer and Prevention	635	x	x
NSHDS	SWEDEN	Northern Sweden Health and Disease Study	163	x	x
HUNT	NORWAY	The Nord-Trøndelag Health Study	254	x	-
MCCS	AUSTRALIA	The Melbourne Collaborative Cohort Study	140	x	-
Estonian BB	ESTONIA	Estonian Genome Center - Estonian Biobank	115	x	-

Overview of the cohorts included in the study population

European Prospective Investigation into Cancer and Nutrition (EPIC)

The European Prospective Investigation into Cancer and Nutrition (EPIC) is an ongoing multicenter prospective cohort study designed primarily to investigate the relationship between nutrition and cancer. The recruitment and baseline assessment of the EPIC cohort are described in detail elsewhere[1,2]. Between 1992 and 2000, 521,330 individuals from 10 European countries. In this project we included participants from France, Germany, Greece, Italy, the Netherlands, Norway, Spain and the United Kingdom were recruited. Participants completed self-administered questionnaires on their diet, lifestyle and medical history. Height and weight of individuals were measured using standard protocols. Of the 521,330 individuals, 385,747 individuals provided a blood sample. Blood fractions were aliquoted into 0.5mL straws, which were heat sealed and stored in liquid nitrogen tanks at -196^{o}C. All participants gave written informed consent. The study was approved by the ethics committee at the International Agency for Research on Cancer (Lyon, France) and the local ethics committee of the study centres.

Incident cancer cases were identified via linkage to population-based cancer registries (in Italy (except Naples), the Netherlands, Norway, Spain, and the United Kingdom) or by active follow-up (in France, Germany, Greece, and Naples), which involved a combination of methods, including review of health insurance records and cancer and pathology registries, as well as direct contact with participants and their next of kin. Participants were followed

up from study entry until cancer diagnosis (except nonmelanoma skin cancer), death, emigration, or the end of follow-up.

We identified 635 eligible kidney cancer cases defined as participants who were diagnosed with Kidney cancer (with *International Classification of Diseases for Oncology, Second Edition*, code C64 and C65), excluding prevalent cases and cases with a history of another cancer (except nonmelanoma skin cancer). For each case, 1 control was chosen randomly from risk sets consisting of all cohort members who were alive and free of cancer (except nonmelanoma skin cancer) at the time of diagnosis of the index case. Matching criteria were country, sex, date of blood collection (±1 month, relaxed to ±5 months for sets without available controls), and date of birth (±1 year relaxed to ±5 years for sets without available controls). Written informed consent was obtained from all participants. In total, 635 matched case-controls pairs were included in our study.

Northern Sweden Health and Disease study (NSHDS)

The Northern Sweden Health and Disease Study (NSHDS) includes several prospective cohorts[3]. The current study included study participants from the Västerbotten Intervention Project (VIP) , which is a sub-cohort within NSHDS. The ongoing VIP prospective cohort is an intervention study aimed at health promotion of the general population of the Västerbotten County in Sweden. In 1985, when VIP was started, all residents in the Västerbotten County were invited to participate by attending a health check-up at 40, 50 and 60 years of age. Participants were asked to complete a self-administered questionnaire that inquired about various population characteristics such as education, smoking habits, physical activity, diet, height and weight. Fasting blood samples were collected from participants during a medical examination. Blood specimens were collected and processed by centrifugation and separation and frozen at -80^{o}C within 1 hr of collection. Plasma samples were stored in the Medical Biobank (Umea, Sweden).

Newly identified cancer cases were identified through linkage with the Swedish Cancer Registry and the local Northern Sweden Cancer Registry. Eligible controls were selected among those who were alive and cancer-free at the time of the case's diagnosis and matched on birthdate (within 2.5 years), sex, blood draw date (within the same year), and fasting status. This study was approved by the Ethics Committee of the Faculty of medicine at Umea University, Umea Sweden. Written informed consent was obtained from all

participants. In total, 163 incident kidney cancer cases and 163 individually matched controls were included in our study.

The Trøndelag Health Study (HUNT)

The Trondelag Health Study (HUNT) includes repeated surveys of a large population-based cohort in Norway[4]. Data from 570 individuals aged 20 years and older from HUNT2 (1995 to 1997, n=416) and HUNT3 (2006 to 2008, n=154) were used in this study. Individuals who participated in both HUNT2 and HUNT3 were included as part of HUNT3. Blood samples were collected at the health examination stations and stored in the HUNT biobank at -70^{o}C for later use. The self-administered questionnaires used in HUNT included medical history, smoking, alcohol consumption. Weight (kg) and height (cm) were measured in a standardized manner in HUNT2 and HUNT3. Body mass index (BMI) was calculated as weight (kg) divided by height (m) squared. Blood samples were collected at the time of participation as described in in the Cohort paper[4] and earlier in this section. The study was approved by the Regional Committee for Ethics in Medical Research, the National Directorate of Health, and by the Norwegian Data Inspectorate.

The mandatory reporting of cancer by physicians and hospitals to the Cancer Registry of Norway (www.krefregisteret.no) provides information on incident cases of kidney cancer that occurred during follow-up. Incident kidney cancer cases were identified using ICD10 codes (C 64) and we acquired information on date of first diagnosis of participants from the Cancer Registry of Norway. All participants with previous cancer diagnosis were excluded. One randomly selected control, matched by sex, age ±2 years, date of blood collection (± 2 months) and time since last meal when blood sample was collected (fasting status). Controls were alive and did not have a cancer diagnosis at the diagnosis time of their index case. Written informed consent was obtained from all participants. In total, 254 matched case-controls pairs were included in our study.

The Melbourne Collaborative Cohort Study (MCCS)

The Melbourne Collaborative Cohort Study (MCCS) is a prospective study of 41,513 healthy adult volunteers (24,469 women) aged between 27 and 76 years (99.3% aged 40-69) when recruited between 1990 and 1994[5,6]. At baseline, demographic characteristics and lifestyle factors were collected by interviewer-administered questionnaires (including

smoking and alcohol consumption) while height, weight, and waist and hip circumferences were measured. Peripheral blood was drawn at recruitment (1990-1994) or at subsequent follow-up (2003-2007). The study was approved by Cancer Council Victoria's Human Research Ethics Committee and performed in accordance with the institution's ethical guidelines.

Cases of kidney cancer were identified by record linkage with the Victorian Cancer Registry that receives mandatory notification of all new cancer cases in Victoria, Australia. Diagnostic pathology reports were reviewed and classified according to the International Classification of Disease (ICD-0-3 WHO classification). Subjects with any history of kidney cancer before blood collection were excluded. Controls were individually matched to cases by age, sex and country of birth. Study participants provided informed consent in accordance with the Declaration of Helsinki. In total, 140 incident kidney cancer cases and 140 individually matched controls were included in our study.

University of Tartu - Estonian Biobank (Estonian BB)

The Estonian Genome Center, The University of Tartu (EGCUT), cohort is a population biobank containing 5% of the Estonian adult population. Detailed description of the Estonian cohort was described previously[7]. The age, sex and geographical distribution of the 152,000 participants closely reflect those of the Estonian adult population. EGCUT can link its own database with the national electronic databases (eight total) to constantly update the phenotype information of the participants. Every entry in the biobank consists of: (i) biological samples, (ii) answers to the questions of a computer-assisted personal interview conducted at the doctor's office (including questions about smoking and alcohol consumption), (iii) objective measurements performed at the doctor's office (including weight, height, waist and hip circumferences and blood pressure), (iv) electronic health data from various databases, (v) genotype data from array genotyping, exome sequencing, or whole-genome sequencing, and (vi) biomedical data obtained by performing various assays on the material collected. Written informed consent was obtained from all participants for the baseline and follow-up investigations.

Kidney cancer cases were identified through national cancer registries and through independent review of medical records. For diagnosis of kidney cancer, we used the ICD-10 C64.0 code. For each case, we selected 1 random control, matching on age at sample

collection, sex and time of blood collection. Controls were individuals who were alive and without a diagnosis of kidney cancer at time of the case's diagnosis date. In total, 115 matched case-controls pairs were included in our study

Metabolite data acquisition

Biocrates

The targeted metabolomics approach was based on LC-ESI-MS/MS and FIA-ESI-MS/MS measurements by AbsoluteIDQ p180 Kit (BIOCRATES Life Sciences AG, Innsbruck, Austria).
The assay allows simultaneous quantification of 188 metabolites out of 10 µL plasma or serum, and includes free carnitine, 39 acylcarnitines (Cx:y), 21 amino acids (19 proteinogenic + citrulline + ornithine), 21 biogenic amines, hexoses (sum of hexoses – about 90-95 % glucose), 90 glycerophospholipids (14 lysophosphatidylcholines (lysoPC) and 76 phosphatidylcholines (PC)), and 15 sphingolipids (SMx:y). The abbreviations Cx:y are used to describe the total number of carbons and double bonds of all chains, respectively (for more details see 1). The method of AbsoluteIDQ p180 Kit has been proven to be in conformance with the EMEA-Guideline "Guideline on bioanalytical method validation (July 21st 2011") [8], which implies proof of reproducibility within a given error range. The long-time stability of plasma metabolites during storage at -80 °C and the performance of the targeted-metabolomics platform using the AbsoluteIDQ p180 Kit have been evaluated in [9].

In the IARC laboratory, a liquid chromatography-tandem mass spectrometry system (Agilent UHPLC-1290/Sciex QTRAP5500 (AB Sciex, Framingham, MA, USA) was used to measure metabolites levels. For the LC-part, compound identification and quantification were based on scheduled multiple reaction monitoring measurements (sMRM). Sample preparation and LC-MS/MS measurements were performed as described in the manufacturer in manual UM-P180-Sciex-13. Analytical specifications for the limit of detection (LOD) and evaluated quantification ranges, further LOD for semiquantitative measurements, identities of quantitative and semiquantitative metabolites, specificity, potential interferences, linearity, precision and accuracy, reproducibility and stability were described in Biocrates manual AS-P180. The LODs were set to three times the values of the zero samples (phosphate buffered

saline solution). The lower and upper limits of quantification were determined experimentally by Biocrates.

In the Helmholtz Zentrum München, an API4000 mass spectrometer (Sciex Deutschland GmbH, Darmstadt, Germany) was used to measure metabolites levels. The assay procedures of the AbsoluteIDQ p180 Kit as well as the metabolite nomenclature have been described in detail previously[10]. Sample handling was performed by a Hamilton Microlab STARTM robot (Hamilton Bonaduz AG, Bonaduz, Switzerland) and a Ultravap nitrogen evaporator (Porvair Sciences, Leatherhead, U.K.), beside standard laboratory equipment. Mass spectrometric analyses were done on an API 4000 triple quadrupole system (Sciex Deutschland GmbH, Darmstadt, Germany) equipped with a 1200 Series HPLC (Agilent Technologies Deutschland GmbH, Böblingen, Germany) and a HTC PAL auto sampler (CTC Analytics, Zwingen, Switzerland) controlled by the software Analyst 1.6.2. Data evaluation for quantification of metabolite concentrations and quality assessment was performed with the software MultiQuant 3.0.1 (Sciex) and the MetIDQ™ software package, which is an integral part of the AbsoluteIDQ Kit. Metabolite concentrations were calculated using internal standards and reported in µM.

Metabolon

All samples were maintained at -80^{o}C until processed. Samples were prepared with use of an automated MicroLab STAR system (Hamilton Company, Reno, NV, USA). For quality control (QC), a pooled sample from all experimental samples was used throughout the experiment, and a mixture of Metabolon QC standards were spiked into all experimental samples to monitor instrument performance and chromatographic alignment. Samples were randomised prior to experimentation. Experiments were conducted on Waters Acuity ultra-performance liquid chromatography (UPLC) systems (Waters Corporation, Milford, MA, USA) using Thermo Scientific Q- Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyser (Thermo Fisher Scientific, MA, USA). The analysis platform used four methods for Ultrahigh Performance Liquid Chromatography- Tandem Mass Spectroscopy (UPLC-MS/MS) including a) positive ion mode electrospray ionisation (ESI), b) positive ion mode optimised for hydrophobic compounds, c) negative ion mode ESI and d) negative ionisation following elution from a hydrophilic interaction chromatography (HILIC) column. Scan time varied

between methods and covered 70- 1000m/z.: Raw data was extracted, peak-identified and QC processed using Metabolon's hardware and software. Metabolites were identified by comparison to the in-house Metabolon standard library using retention time, mass (m/z), adducts and MS/MS spectra. As experiments were conducted over multiple consecutive days, a data normalization step was performed to correct variation resulting from instrument inter-day tuning differences.

The cases and their matched controls were assayed within the same batches in order to avoid any effect of batch differences on the risk estimates.

Instrument variability was determined by calculating the median relative standard deviation (RSD) for the internal standards that were added to each sample prior to injection into the mass spectrometers. Overall process variability was determined by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in the MTRX5 technical replicates (a large pool of human plasma maintained by Metabolon that has been characterized extensively).

Values for instrument and process variability meet Metabolon's acceptance criteria: median RSD for internal standards were 5% and 4% for EPIC and NSHDS samples, respectively; median RSD for endogenous biochemicals were 11% for both EPIC and NSHDS.

Data sources for Mendelian randomization analyses

BMI GWAS

Summary-level GWAS data for BMI was obtained from a 2018 meta-analysis of GWASs of BMI [11] (downloaded from: https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files#2018_GIANT_and_UK_BioBank_Meta-analysis). This analysis was a fixed-effects meta-analysis combining results from a GWAS of BMI performed among 456,426 participants from the UK Biobank (adjusted for age, sex, recruitment center, genotyping batch and 10 genetic principal components) and results from a BMI GWAS published by the GIANT (Genetic Investigation of Anthropometric Traits)[12] consortium, which included 253,288 participants from 79 studies (adjusted for age, sex, and study specific covariates). For UK Biobank, BMI (weight in kg per height in metres squared) was measured during the initial assessment centre visit whereas for the BMI GWAS conducted by the GIANT consortium, BMI was either measured or self-reported.

Metabolite GWAS

Summary-level GWAS data for 174 Biocrates metabolites [13] and 913 Metabolon metabolites were used. The metabolite GWAS data used in the MR analyses, are available via www.omicscience.org for all Biocrates and a subset of Metabolon metabolites. Metabolite associations for the BMI-associated and dental disease-associated SNPs used in the Mendelian randomization analyses are available to download from:[**data.bris.url to be inserted on acceptance of manuscript**].

Biocrates:

The GWAS meta-analysis for the 174 Biocrates metabolites was a fixed-effects meta-analysis combining results from the Fenland cohort [14] (maximum N= 9736, available at: https://omicscience.org/apps/crossplatform/) (metabolites profiled by the Biocrates p180 kit and measured using mass spectrometry) with those from the EPIC-Norfolk [15] (maximum N=5841) and INTERVAL studies [16] (maximum N=40,818) (metabolites were profiled using mass spectrometry (Metabolon Discovery HD4 platform) and proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy). Ten of the 174 Biocrates metabolites were covered across all platforms, while 38 were available on the Biocrates and Metabolon platforms and 126 were unique to Biocrates. An overall z-score meta-analysis was also conducted by further integrating publicly available summary statistics from GWAS of the same metabolites measured using mass spectrometry (with Biocrates or Metabolon platforms[17,18]) or ^{1}H-NMR spectroscopy [18] (N=ranged from 8,569 to 86,507 for different metabolites, available at: https://omicscience.org/apps/crossplatform/).

Genotyping in Fenland was performed using Affymetrix SNP5.0 and Affymetrix Axiom and genotype imputation was performed using 1000 Genomes Phase 1v3 or phase 3 reference panels. In EPIC-Norfolk, genotype imputation was performed using 1000 Genomes Phase 3 reference panels. Genotyping in INTERVAL was performed using Affymetrix Axiom and imputation was performed using the 1000 Genomes Phase 3 (May 2013)-UK10K reference imputation panel. For Fenland and EPIC-Norfolk, GWAS analyses were carried out using BOLT-LMM and SNPTEST adjusting for age, sex and study-specific covariates in mixed linear models. For the GWAS conducted in INTERVAL, phenotype residuals were corrected for age,

gender, metabolon batch, INTERVAL centre, plate number, appointment month, the lag time between the blood donation appointment and sample processing, and the first 5 ancestry principal components.

For the pleiotropy analyses, SNPs associated with metabolites at $p<4.9x10^{-10}$ (conventional threshold of genome-wide significance corrected for 102 tests which corresponded to the number of principal components that explained 95% of the variance of the 174 metabolites in the Fenland cohort) were identified from the overall z-score meta-analysis. The estimated effect sizes for each of the metabolite-associated SNPs were then obtained from the three-cohort meta-analysis (Fenland and, when available, EPIC-Norfolk and/or INTERVAL) and only metabolite-associated SNPs with a $p<5x10^{-08}$ in the three-cohort meta-analysis was included in the pleiotropy analyses. For the MR analyses, the metabolite associations for the BMI-associated SNPs or dental disease-associated SNPs were obtained from the three-cohort meta-analysis.

Metabolon:

A GWAS of metabolon metabolite levels was performed using samples from the EPIC-Norfolk [15] and INTERVAL studies [19]. 14,296 participants were included in a *discovery* set (5,841 from EPIC-Norfolk; 8,455 from INTERVAL) and 5,698 from EPIC-Norfolk in a *validation* set. Metabolites were measured using the Metabolon DiscoveryHD4 platform (Metabolon, Inc., Durham, USA), from plasma samples collected at baseline. A total of 913 metabolites measured in at least 100 participants in each study were taken forward for GWAS analysis. Metabolite measures were median normalised for run day, log transformed, winsorised to 5 standard deviations, before being regressed against age, sex and study specific variables using linear regression. Residuals from this regression were standardised (mean 0, standard deviation 1) and used for further analysis. Genotyping was performed using the Affymetrix Axiom UK Biobank genotyping array. In INTERVAL, genotype imputation was performed using the combined UK10K+1000 Genomes Phase 3 reference panel. In EPIC-Norfolk, imputation was performed using the Haplotype Reference Consortium reference panel, with additional variants imputed using the UK10K+1000 Genomes Phase 3 reference panel.

Association analyses were performed using BOLT-LMM [20] or SNPTEST [21,22] separately in each study and combined using inverse variance weighted fixed effect meta-analysis methods implemented in METAL [23]. Genome-wide significant ($p < 5 \times 10^{-8}$) lead regional associations that were directionally consistent and significant at $p < 0.01$ in both studies were considered validated if they were significant at $p < 5.48 \times 10^{-11}$ ($p < 5 \times 10^{-8}$ Bonferroni corrected for 913 metabolites) and directionally consistent in a meta-analysis including the independent validation samples, as described above. To identify independent associations, exact conditional analyses were then performed using forward stepwise regression with a significance threshold of $p < 1.25 \times 10^{-8}$. For the current study, unconditional effect estimates for both primary and conditionally independent associations were used. In analyses to assess pleiotropy of potential instruments, we obtained the effect estimates from the unconditional analysis and all SNPs used had a $p<5$ $\times 10^{-08}$ in the unconditional analysis.

Dental disease GWAS

Summary-level data for dental disease was obtained from a 2019 meta-analysis of GWASs of dental disease (DMFS (Decayed, Missing and Filled tooth Surfaces); N=26,792 from 9 studies) and dentures (n_{case}= 77,714 and $n_{controls}$ = 383,317) [24] (downloaded from: https://data.bris.ac.uk/data/dataset/2j2rqgzedxlq02oqbb4vmycnc2). This analysis was a fixed effects meta-analysis combining results from a GWAS of dental disease performed in the UK Biobank (adjusted for age, age-squared, sex, genotyping batch) and a GWAS of dental disease conducted by GLIDE (Gene-Lifestyle Interactions in Dental Endpoints) (adjusted for age, age-squared, genetic principal components and other study-specific covariates). Dental disease. Self-reported measures of oral health were characterised in UK Biobank while clinical dental records were used to calculate DMFS/dentures in GLIDE.

References

1. Riboli E, Kaaks R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. Int J Epidemiol. 1997;26 Suppl 1:S6-14. Epub 1997/01/01. PMID: 9126529.
2. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. 2002;5(6B):1113-24. Epub 2003/03/18. doi: 10.1079/PHN2002394. PMID: 12639222.
3. Hallmans G, Agren A, Johansson G, Johansson A, Stegmayr B, Jansson JH, et al. Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort - evaluation of risk factors and their interactions. Scand J Public Health Suppl. 2003;61:18-24. doi: 10.1080/14034950310001432. PMID: 14660243.
4. Krokstad S, Langhammer A, Hveem K, Holmen TL, Midthjell K, Stene TR, et al. Cohort Profile: the HUNT Study, Norway. Int J Epidemiol. 2013;42(4):968-77. Epub 2012/08/11. doi: 10.1093/ije/dys095. PMID: 22879362.
5. Milne RL, Fletcher AS, MacInnis RJ, Hodge AM, Hopkins AH, Bassett JK, et al. Cohort Profile: The Melbourne Collaborative Cohort Study (Health 2020). Int J Epidemiol. 2017;46(6):1757-i. Epub 2017/06/24. doi: 10.1093/ije/dyx085. PMID: 28641380.
6. Giles GG, English DR. The Melbourne Collaborative Cohort Study. IARC Sci Publ. 2002;156:69-70. PMID: 12484128.
7. Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol. 2015;44(4):1137-47. Epub 2014/02/13. doi: 10.1093/ije/dyt268. PMID: 24518929.
8. Guideline on bioanalytical method validation. Committee for Medicinal Products for Human Use (CHMP). 2011;EMEA/CHMP/EWP/192217/2009 (Rev. 1 Corr. 2). Epub 21 July 2011.
9. Haid M, Muschet C, Wahl S, Romisch-Margl W, Prehn C, Moller G, et al. Long-Term Stability of Human Plasma Metabolites during Storage at -80 degrees C. J Proteome Res. 2018;17(1):203-11. Epub 2017/10/25. doi: 10.1021/acs.jproteome.7b00518. PMID: 29064256.
10. Zukunft S, Sorgenfrei M, Prehn C, Moller G, Adamski J. Targeted Metabolomics of Dried Blood Spot Extracts. Chromatographia. 2013;76(19-20):1295-305. doi: 10.1007/s10337-013-2429-3. PMID: WOS:000324825100011.
11. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. Hum Mol Genet. 2018;27(20):3641-9. Epub 2018/08/21. doi: 10.1093/hmg/ddy271. PMID: 30124842.
12. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197-206. Epub 2015/02/13. doi: 10.1038/nature14177. PMID: 25673413.
13. Lotta LA, Pietzner M, Stewart ID, Wittemans LBL, Li C, Bonelli R, et al. A cross-platform approach identifies genetic regulators of human metabolism and health. Nat Genet. 2021;53(1):54-64. Epub 2021/01/09. doi: 10.1038/s41588-020-00751-5. PMID: 33414548.
14. Lindsay T, Westgate K, Wijndaele K, Hollidge S, Kerrison N, Forouhi N, et al. Descriptive epidemiology of physical activity energy expenditure in UK adults (The Fenland

study). Int J Behav Nutr Phys Act. 2019;16(1):126. Epub 2019/12/11. doi: 10.1186/s12966-019-0882-6. PMID: 31818302.
15. Day N, Oakes S, Luben R, Khaw KT, Bingham S, Welch A, et al. EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. Br J Cancer. 1999;80 Suppl 1:95-103. Epub 1999/08/31. PMID: 10466767.
16. Moore C, Sambrook J, Walker M, Tolkien Z, Kaptoge S, Allen D, et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. Trials. 2014;15:363. Epub 2014/09/19. doi: 10.1186/1745-6215-15-363. PMID: 25230735.
17. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. Nat Genet. 2014;46(6):543-50. Epub 2014/05/13. doi: 10.1038/ng.2982. PMID: 24816252.
18. Draisma HHM, Pool R, Kobl M, Jansen R, Petersen AK, Vaarhorst AAM, et al. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. Nat Commun. 2015;6:7208. Epub 2015/06/13. doi: 10.1038/ncomms8208. PMID: 26068415.
19. Di Angelantonio E, Thompson SG, Kaptoge S, Moore C, Walker M, Armitage J, et al. Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. Lancet. 2017;390(10110):2360-71. Epub 2017/09/25. doi: 10.1016/S0140-6736(17)31928-1. PMID: 28941948.
20. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47(3):284-90. Epub 2015/02/03. doi: 10.1038/ng.3190. PMID: 25642633.
21. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11(7):499-511. Epub 2010/06/03. doi: 10.1038/nrg2796. PMID: 20517342.
22. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661-78. Epub 2007/06/08. doi: 10.1038/nature05911. PMID: 17554300.
23. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1. Epub 2010/07/10. doi: 10.1093/bioinformatics/btq340. PMID: 20616382.
24. Shungin D, Haworth S, Divaris K, Agler CS, Kamatani Y, Keun Lee M, et al. Genome-wide analysis of dental caries and periodontitis combining clinical and self-reported data. Nat Commun. 2019;10(1):2773. Epub 2019/06/27. doi: 10.1038/s41467-019-10630-1. PMID: 31235808.

Table of Contents

Funding

The metabolomics analysis of this study was supported by World Cancer Research Fund (reference: 2014/1193, MJ) and the European Commission (FP7: BBMRI-LPC; reference: 313010, MJ). The work was supported by a Cancer Research UK Programme Grant [The Integrative Cancer Epidemiology Programme, ICEP] (C18281/A19169, NJT). This research was funded in whole, or in part, by the Wellcome Trust (202802/Z/16/Z, NJT). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

The coordination of EPIC is financially supported by International Agency for Research on Cancer (IARC, MJ) and also by the Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London which has additional infrastructure support provided by the NIHR Imperial Biomedical Research Centre (BRC, ER). The national cohorts are supported by: Danish Cancer Society (Denmark, CCD); Ligue Contre le Cancer (GS), Institut Gustave Roussy (GS), Mutuelle Générale de l'Education Nationale (GS), Institut National de la Santé et de la Recherche Médicale (INSERM) (France, GS); German Cancer Aid, German Cancer Research Center (DKFZ, MMB), German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE, MMB), Federal Ministry of Education and Research (BMBF) (Germany, MMB); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy, Compagnia di SanPaolo and National Research Council (Italy, SC); Dutch Ministry of Public Health, Welfare and Sports (VWS, RV), Netherlands Cancer Registry (NKR, RV), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland, RV), World Cancer Research Fund (WCRF, RV), Statistics Netherlands (The Netherlands, RV); Health Research Fund (FIS) - Instituto de Salud Carlos III (ISCIII), Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, and the Catalan Institute of Oncology - ICO (Spain, PAE); Swedish Cancer Society, Swedish Research Council and County Councils of Skåne and Västerbotten (BJ). We thank the National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands, for their contribution and ongoing support to the EPIC Study (RV).

The EPIC-Norfolk study (https://doi.org/10.22025/2019.10.105.00004) has received funding from the Medical Research Council (MR/N003284/1 MC-UU_12015/1 and MC_UU_00006/1, JAS) and Cancer Research UK (C864/A14136, JAS). The genetics work in the EPIC-Norfolk study was funded by the Medical Research Council (MC_PC_13048, CL). Metabolite measurements in the EPIC-Norfolk study were supported by the MRC Cambridge Initiative in Metabolic Science (MR/L00002/1, CL) and the Innovative Medicines Initiative Joint Undertaking under EMIF grant agreement no. 115372 (CL).

Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping were co-funded by the National Institute for Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk) and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014, AB) . [The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care]. The academic coordinating centre for INTERVAL was supported by core funding from the: NIHR Blood and Transplant Research Unit in Donor Health and Genomics (NIHR BTRU-2014-10024, AB), UK Medical Research Council (MR/L003120/1, AB), British Heart Foundation (SP/09/002; RG/13/13/30194; RG/18/13/33946, AB) and NIHR Cambridge BRC (BRC-1215-20014, AB). A complete list of the investigators and contributors to the INTERVAL trial is provided in reference 19 of Supplementary methods (Di Angelantonio, et al.).

Melbourne Collaborative Cohort Study (MCCS) cohort recruitment was funded by VicHealth and Cancer Council Victoria (GG, RM). The MCCS was further augmented by Australian National Health and Medical Research Council grants 209057, 396414 and 1074383 (GG, RM) and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry and the Australian Institute of Health and Welfare, including the National Death Index and the Australian Cancer Database (ACD). New South Wales (NSW) cancer registry data were obtained via the ACD with the assistance of the NSW Ministry of Health.

NJT, VYT, FG and KSB are supported by the Cancer Research UK (CRUK) Integrative Cancer Epidemiology Programme (C18281/A29019). NJT, LJC and VYT work in the MRC IEU at the University of Bristol which is supported by the MRC (MC_UU_00011) and the University of Bristol. NJT is a Wellcome Trust Investigator (202802/Z/16/Z) and works within the University of Bristol National Institute for Health Research (NIHR) Biomedical Research Centre (BRC). LJC is supported by NJT's Wellcome Trust Investigator grant (202802/Z/16/Z). EEV is supported by Diabetes UK (17/0005587). EEV and CJB are supported by the World Cancer Research Fund (WCRF UK), as part of the World Cancer Research Fund International grant programme (IIG_2019_2009). University of Tartu - Estonian Biobank was supported by NIH grant no 5R01 DK07 57 87 -13, under subward-agreement no GENFDOOO1B52751, the European Union through Horizon 2020 research and innovation programme under grant no 633589, the European Union through the European Regional Development Fund (Project No. 2014-2020.4.01.16-0125), the Estonian Research Council grant PUT (PRG687). The work of TLL was supported by Research Council of Norway Grant No. 267776/H10 within the framework of an agreement between the Research Council of Norway and the Norwegian University of Science and Technology. MMB was funded by the German Institute of Human Nutrition Potsdam-Rehbrücke, a government-financed organization.

No funders had role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests statement

I have read the journal's policy and the authors of this manuscript have the following competing interests: CL is an Academic Editor on PLOS Medicine's editorial board; AB reports grants outside of this work from AstraZeneca, Biogen, BioMarin, Bioverativ, Merck, Novartis, Pfizer and Sanofi and personal fees from Novartis; during the course of this project, PS became a full-time employee of GSK. No other conflicts of interest have been declared by the authors.