Pan-cancer analysis of pre-diagnostic blood metabolite concentrations in the European Prospective Investigation into Cancer and Nutrition

Marie Breeur[1], Pietro Ferrari[1], Laure Dossus[1], Mazda Jenab[1], Mattias Johansson[2], Sabina Rinaldi[1], Ruth C. Travis[3], Mathilde His[1], Tim J. Key[3], Julie A. Schmidt[3,4], Kim Overvad[5], Anne Tjønneland[6], Cecilie Kyrø[6], Joseph A. Rothwell[7], Nasser Laouali[7], Gianluca Severi[7], Rudolf Kaaks[8], Verena Katzke[8], Matthias B. Schulze[9], Fabian Eichelmann[9,10], Domenico Palli[11], Sara Grioni[12], Salvatore Panico[13], Rosario Tumino[14], Carlotta Sacerdote[15], Bas Bueno-de-Mesquita[16], Karina Standahl Olsen[17], Torkjel Manning Sandanger[17], Therese Haugdahl Nøst[17], J. Ramón Quirós[18], Catalina Bonet[19], Miguel Rodríguez Barranco[20,21,22], María-Dolores Chirlaque[22,23], Eva Ardanaz[22,24,25], Malte Sandsveden[26], Jonas Manjer[27], Linda Vidman[28], Matilda Rentoft[28], David Muller[29], Kostas Tsilidis[29], Alicia K. Heath[29], Hector Keun[30], Jerzy Adamski[31,32,33], Pekka Keski-Rahkonen[1], Augustin Scalbert[1], Marc J. Gunter[1], Vivian Viallon[1]

[1]Nutrition and Metabolism Branch, International Agency for Research on Cancer, 69372 Lyon CEDEX 08, France.

[2]Genetics Branch, International Agency for Research on Cancer, 69372 Lyon CEDEX 08, France.

[3]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK.

[4]Department of Clinical Epidemiology, Department of Clinical Medicine, Aarhus University Hospital and Aarhus University, DK-8200 Aarhus N, Denmark.

[5]Aarhus University Department of Public Health, DK-8000 Aarhus C, Denmark.

[6]Danish Cancer Society Research Center Diet, Genes and Environment Nutrition and Biomarkers, DK-2100 Copenhagen

[7]Université Paris-Saclay, UVSQ, Inserm, CESP U1018, "Exposome and Heredity" team, Gustave Roussy, 94800 Villejuif, France.

[8]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

[9]German Institute of Human Nutrition, Dept. of Molecular Epidemiology, 14558 Nuthetal, Germany

[10]German Center for Diabetes Research (DZD), 85764 Neuherberg, Germany

[11]Institute of Cancer Research, Prevention and Clinical Network (ISPRO), 50139 Florence, Italy

[12]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, 20133 Milan, Italy

[13]Dipartimento di Medicina Clinica e Chirurgia, Federico II University, 80131 Naples, Italy

[14]Hyblean Association for Epidemiological Research, AIRE-ONLUS, 97100 Ragusa, Italy
[15]Unit of Cancer Epidemiology Città della Salute e della Scienza University-Hospital, 10126 Turin, Italy
[16]Centre for Nutrition, Prevention and Health Services, National Institute for Public Health and the Environment (RIVM), PO Box 1, 3720 BA Bilthoven, The Netherlands
[17]Department of Community Medicine, UiT The Arctic University of Norway, N-9037 Tromsø, Norway
[18]Public Health Directorate, 33006 Asturias, Spain
[19]Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), 08908 L'Hospitalet de Llobregat, Barcelona, Spain.
[20]Escuela Andaluza de Salud Pública (EASP), 18011 Granada, Spain
[21]Instituto de Investigación Biosanitaria ibs.GRANADA, 18012 Granada, Spain
[22]Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), 28029 Madrid, Spain
[23]Department of Epidemiology, Regional Health Council, IMIB-Arrixaca, Murcia University, 30003 Murcia, Spain.
[24]Navarra Public Health Institute, 31003 Pamplona, Spain.
[25]IdiSNA, Navarra Institute for Health Research, 31008 Pamplona, Spain.
[26]Department of Clinical Sciences Malmö Lund University, SE-214 28 Malmö, Sweden.
[27]Departement of Surgery Skåne University Hospital Malmö Lund University, SE-214 28 Malmö, Sweden.
[28]Department of Radiation Sciences, Oncology Umeå University SE-901 87 Umeå, Sweden.
[29]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London W2 1PG, UK.
[30]Cancer Metabolism and Systems Toxicology Group, Division of Cancer, Department of Surgery and Cancer, Imperial College London, SW7 2AZ London, UK
[31]Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany.
[32]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore.
[33]Institute of Biochemistry, Faculty of Medicine, University of Ljubljana, 1000 Ljubljana, Slovenia.

Corresponding Author: Vivian Viallon, NME Branch, IARC; viallonv@iarc.fr

Abstract (251 words)

Background: Epidemiological studies of associations between metabolites and cancer risk have typically focused on specific cancer types separately. Here, we designed a multivariate pan-cancer analysis to identify metabolites potentially associated with multiple cancer types, while also allowing the investigation of cancer type-specific associations.

Methods: We analyzed targeted metabolomics data available for 5,828 matched case-control pairs from cancer-specific case-control studies on breast, colorectal, endometrial, gallbladder, kidney, localized and advanced prostate cancer, and hepatocellular carcinoma nested within the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort. From pre-diagnostic blood levels of an initial set of 117 metabolites, 33 cluster representatives of strongly correlated metabolites, and 17 single metabolites were derived by hierarchical clustering. The mutually adjusted associations of the resulting 50 metabolites with cancer risk were examined in penalized conditional logistic regression models adjusted for body mass index, using the data shared lasso penalty.

Results: Out of the 50 studied metabolites, *(i)* six were inversely associated with risk of most cancer types: glutamine, butyrylcarnitine, lysophosphatidylcholine a C18:2 and three clusters of phosphatidylcholines (PCs); *(ii)* three were positively associated with most cancer types: proline, decanoylcarnitine and one cluster of PCs; and *(iii)* 10 were specifically associated with particular cancer types, including histidine that was inversely associated with colorectal cancer risk, and one cluster of sphingomyelins that was inversely associated with risk of hepatocellular carcinoma and positively with endometrial cancer risk.

Conclusions: These results could provide novel insights for the identification of pathways for cancer development, in particular those shared across different cancer types.

Keywords: metabolomics, cancer, breast, colorectal, endometrial, kidney, liver, prostate, lasso, EPIC

Background

Metabolomics allows the simultaneous measurement of a large variety of compounds present in biological samples, such as human blood[1,2]. Circulating metabolite levels can reflect both endogenous and exogenous processes, providing a snapshot of biological activity[3,4]. As a result, metabolomics may facilitate the identification of biological mechanisms involved in the development of chronic diseases. For example, prior metabolomics studies have identified metabolites associated with the risk of various chronic conditions, including type-2 diabetes (T2D)[5–7], cardiovascular diseases (CVD)[8–10], and different site-specific cancers, including cancers of the breast[11], prostate[12,13], endometrium[14], kidney[15], colorectum[16–18], hepatocellular carcinoma (HCC)[19], and others[20,21].

Several shared biological mechanisms are known to underlie multiple chronic diseases. Obesity, physical inactivity and adherence to a Western-type diet, as well as chronic inflammation and insulin resistance, are recognized risk factors for cardio-metabolic diseases, including T2D, CVD, and several site-specific cancers[22–24]. Metabolomics may help uncover novel etiological mechanisms that are common to several chronic diseases as well as those that are disease-specific. One recent study identified metabolites associated with the risk of multimorbidity, defined as the simultaneous presence of multiple chronic conditions within one individual. Focusing on a pre-defined panel of metabolites, a targeted metabolomics study of breast, prostate and colorectal cancers in a German population found that circulating levels of the phosphatidylcholine PC ae C30:0 and several lysophosphatidylcholines, including lysoPC a C18:0, were predictive of the development of any of these three cancers[25], suggesting that some etiological mechanisms could be shared across multiple cancer types.

In this work, we extended this concept by leveraging targeted metabolomics data available within nested case-control studies on eight cancer types (breast, colorectal, endometrial, gallbladder and biliary tract, kidney, localized prostate and advanced prostate cancers, and HCC) previously acquired in the European Prospective Investigation into Cancer and Nutrition (EPIC)[11,12,14,15,19]. The data shared lasso[26–28], a penalized multivariate approach specifically designed for the investigation of a set of shared risk factors across different disease outcomes, was used to carry out a multivariate pan-cancer analysis to identify mutually adjusted metabolites associated with cancer risk and to identify those metabolites with consistent or heterogeneous patterns of associations across the eight cancer types.

Methods

Study population. EPIC is an ongoing multicentric prospective study with over 500,000 men and women recruited between 1992 and 2000 from 23 centers in 10 European countries[29], originally designed to study the relationship between diet and cancer risk. Incident cancer cases were identified through a combination of methods, including

health insurance records, cancer and pathology registries and active follow-up through study participants and their next-of-kin. At recruitment, information on diet and lifestyle was collected via self-administered questionnaires. Blood samples were collected from around 386,000 participants according to a standardized protocol. In France, Germany, Greece, Italy, the Netherlands, Norway, Spain, and the UK, serum (except in Norway), plasma, erythrocytes, and buffy coat aliquots were stored in liquid nitrogen (− 196 °C) in a centralized biobank at the International Agency for Research on Cancer (IARC). In Denmark, blood fractions were stored locally in the vapor phase of liquid nitrogen containers (− 150 °C), and in Sweden, they were stored locally at − 80 °C in standard freezers. Fasting was not required.

Our analyses used a set of metabolomics measurements from 15,948 EPIC participants from seven cancer-specific matched case-control studies nested within EPIC (Table 1). In each study, each case was matched to one control selected among cancer-free participants (other than non-melanoma skin cancer) by risk set sampling, using matching factors that included study center, sex, age at blood collection, time of the day of blood collection, fasting status, and use of exogenous hormones for women. All participants provided written informed consent to participate in the EPIC study. The cancer-specific case-control studies were all approved by the ethics committee of IARC and participating EPIC centers.

Laboratory analysis. As summarized in Table 1, pre-diagnostic blood samples were assayed at the Helmholtz Zentrum (München, Germany) for the second colorectal cancer study, at Imperial College London (UK) for the endometrial cancer study, and at IARC for all other studies. Data for a total of 171 metabolites were acquired by tandem mass spectrometry using either the AbsoluteIDQ p150 (for the second colorectal cancer study) or the AbsoluteIDQ p180 commercial kit (Biocrates Life Science AG, Innsbruck Austria). Two successive assays were used, liquid chromatography-tandem mass spectrometry (LC-MS/MS) for amino acids and biogenic amines, and flow injection analysis-tandem mass spectrometry (FIA-MS/MS) for the other metabolites. Samples were either serum or citrate plasma, and samples within each study were all from the same type of blood matrix, except for the breast cancer study (Table 1).

Selection of the metabolites, data pre-processing. Data were pre-processed following an established procedure[30]. Briefly, metabolites with more than 25% missing values in any study were excluded. Samples with more than 25% missing values overall were excluded, as were those detected as outliers by a principal component analysis (PCA)-based approach applied within each study separately. Then, for all metabolites measured by FIA with a semi-quantitative method (acylcarnitines, glycerophospholipids, sphingolipids, hexoses), measurements below the batch-specific limit of detection (LOD) were imputed to half the LOD. When the batch-specific LOD was unknown, LOD was first set to study-specific medians of known batch-specific LODs. For the metabolites measured with a fully quantitative approach (amino acids and biogenic amines), measurements below the lower limit of

quantification (LLOQ) or above the upper limit of quantification (ULOQ) were imputed to half the LLOQ or to the ULOQ, respectively. For all metabolites, other missing values were imputed to the batch-specific median of the non-missing measurements. The resulting measurements were then log-transformed to improve symmetry.

Cancer types and exclusion criteria. We focused on eight cancer types, namely breast, colorectal, endometrial, kidney, gallbladder and biliary tract cancers, HCC, advanced and localized prostate. As detailed in Section 1 of the Supplementary Material (Additional file 1), matched case-control pairs for HCC and gallbladder and biliary tract cancer were extracted from the liver cancer study, while matched case-control pairs for advanced and localized prostate cancer were extracted from the prostate cancer study. Since hormones could affect metabolite levels and their association with cancer risk[11], women using exogenous hormones (either hormone replacement therapy or oral contraceptive) at baseline were excluded.

Statistical analyses. All analyses were performed using R software. Characteristics of cases and controls for the eight studied cancer types were described using mean and standard deviation or frequency. Pearson correlations between the metabolites were computed in controls only to reduce collider bias.

Clustering of metabolites: The most strongly correlated metabolites were grouped together by applying the hierarchical clustering approach implemented in the ClustOfVar R package[31] to the control samples. For each cluster, the method defined its representative as the first principal component in the PCA of the metabolites grouped into that cluster. In our figures and tables, cluster representatives were labeled as "xxx_clus", with "xxx" representing one particular metabolite that composed that cluster. We retained the model with lowest number of clusters such that representatives explained at least 80% of the total variation in each cluster. Cluster representatives and metabolites left isolated after the clustering were simply referred to as metabolites hereafter.

Multivariate analyses: Given the number of studied metabolites, penalized conditional logistic regression models were used to estimate mutually adjusted associations with cancer risk. Since body-mass index (BMI) could be a strong confounder of the relationship between several of the examined metabolites[32,33] and cancers[34–38], metabolite-specific linear models were used to compute residuals on BMI. To account for the large number of metabolites and leverage possible commonalities among the metabolic disorders preceding cancer development for different cancer types, estimation was based on the data shared lasso[26–28], an extension of the lasso[39] allowing the analysis of case-control studies with multiple disease types. For each metabolite, the data shared lasso decomposes its type-specific odds-ratio as the product of *(i)* an overall odds-ratio capturing the overall association with cancer, and *(ii)* type-specific deviations from this overall odds-ratio. Then, the method identifies whether its overall (mutually adjusted) association with cancer is null or not, and also whether some of

its type-specific associations deviate from its (possibly null) overall association with cancer. Compared to more standard approaches, the data shared lasso was shown to perform particularly well for the identification of features with a consistent non-null association with multiple disease types, while also allowing for the identification of type-specific associations[28].

To assess the robustness of the identified associations, the data shared lasso was applied repeatedly on 100 bootstrap samples generated from the original sample[40]. Moreover, following the rationale of the lasso-OLS hybrid[41], associations identified by the data shared lasso were further inspected using unpenalized conditional logistic regression models, *(i)* to quantify their strength and investigate possible heterogeneity among the type-specific associations beyond those identified by the data shared lasso (see Section 3 in Additional file 1 for details); *(ii)* to assess possible departure from linearity by comparing models with natural cubic splines to models with linear terms only; and *(iii)* to assess possible attenuation after excluding, in turn, first two and first seven years of follow-up (to examine potential reverse causation and more generally assess the impact of time to diagnosis on our findings), and after adjustment for additional factors (education level, waist circumference, height, physical activity, smoking status, alcohol intake, use of non-steroidal anti-inflammatory drugs, and, for women, menopausal status and phase of menstrual cycle in premenopausal women). Finally, effect modification by BMI was assessed under standard (i.e., non-conditional) logistic regression models after breaking the matching and correcting metabolite measurements for batch and study effects[30].

Univariate analyses: For comparison, non-mutually adjusted associations with cancer risk were estimated for each metabolite in conditional logistic regression models adjusted for BMI. Each cancer type was first modelled separately, and then jointly, via one global conditional logistic regression model. Heterogeneity of associations across cancer types was tested by comparing the difference in log-likelihood between the global model and a model with interaction terms between each metabolite and cancer type to a chi-square distribution with 8-1=7 degrees of freedom. To account for multiple comparisons, associations and heterogeneities with a False Discovery Rate (FDR) inferior to 5% were considered as statistically significant[42].

Analysis of additional metabolites: The 16 metabolites (Table S1, Additional file 2) that were not acquired in the second colorectal cancer study (AbsoluteIDQ p150 kit) were not included in our main analysis and were examined in a reduced sample, using the methods described above.

Results

Description of the study population. After the exclusions of subjects detailed in Figure 1, 11,656 EPIC participants were included in the analysis comprising 5,828 matched case-control pairs. Cases were diagnosed at an average age of 64.4 years, 8.4 years after blood collection. The main characteristics of cases and controls in each study are

displayed in Table 2. The main analysis focused on 117 metabolites that were retained after the pre-processing step (Table S1, Additional file 2). As displayed in Figure S1 in Additional file 2, strong positive correlations were observed between some metabolites, particularly between some of the glycerophospholipids (phosphatidylcholines, PCs, and lysophosphatidylcholines, lysoPCs), and sphingomyelins (SMs).

Clustering of metabolites. The hierarchical clustering applied to controls grouped 100 metabolites into 33 clusters of size ranging from 2 to 6 metabolites per cluster, while 17 metabolites remained isolated. As displayed in Figure 2, clusters comprised metabolites of the same chemical class, and correlations between metabolites and their representative were consistently greater than 0.83. On average, clusters' representatives explained 86% of the total variation of their cluster (range: 80%-95%), and the 33 + 17 = 50 studied metabolites together explained more than 88% of the total variation of the original 117 metabolites.

Multivariate analyses. As displayed in Figures 3 and 4, the data shared lasso identified nine metabolites with a non-null overall association with cancer: butyrylcarnitine (acylcarnitine C4), glutamine, lysoPC a C18:2, and three clusters of PCs (those containing PC aa C32:2, PC aa C36:0, and PC aa C36:1, respectively), with an inverse overall association with cancer risk, and decanoylcarnitine (acylcarnitine C10), proline and the cluster of PCs that included PC aa C28:1 with a positive overall association. Cancer type-specific deviations from the overall association with cancer risk were identified for three of these metabolites: the association between proline and breast cancer risk was inverse or null, while the associations between lysoPC a C18:2 and the cluster containing PC aa C36:0 with localized prostate cancer were positive or null.

Several cancer type-specific associations were identified among the remaining 41 metabolites. Specifically, positive associations were observed between breast cancer risk and two clusters, that included tetradecenoylcarnitine (acylcarnitine C14:1) and PC aa C36:5, respectively. Risk of colorectal cancer was positively associated with arginine and PC ae C36:0, and inversely associated with the cluster that included histidine. Risk of HCC was positively associated with the cluster containing PC aa C40:2, and inversely associated with the two clusters that included lysoPC a C20:3 and SM C16:0, respectively. This latter cluster was also positively associated with endometrial cancer risk. The cluster that included octadecenoylcarnitine (acylcarnitine C18:1) was inversely associated with risk of advanced prostate cancer. Finally, risk of localized prostate cancer was inversely associated with hexoses (H1).

The strength of the associations identified by the data shared lasso was similar after excluding, in turn, the first two and the first seven years of follow-up (Figure S2, Additional file 2). Likewise, models adjusted for additional factors produced similar associations (Figure S2, Additional file 2), except for the overall association with cancer for the cluster that included PC aa C28:1, whose odds-ratio (OR) was attenuated from 1.09 (95% confidence interval: 1.01-1.17) to 1.04 (0.98-1.12), and for the association

between endometrial cancer risk and the cluster that included SM C16:0, whose OR decreased from 1.51 (1.19-1.93) to 1.20 (0.97-1.47). For each overall association and type-specific deviation identified by the data shared lasso, linearity and absence of effect modification by BMI were compatible with our data (Figure S3, Additional file 2). Focusing on the nine metabolites that had a non-null overall association with cancer, the analysis presented in Figure S4 in Additional file 2 suggested possible cancer type-specific deviations from the overall associations beyond the three ones identified by the data shared lasso, in particular for HCC (with acylcarnitine C4, proline and the cluster that comprises PC aa C36:1) and for kidney cancer (with acylcarnitines C10 and C4 and the cluster that comprises PC aa C36:1). However, none of the comparisons between the models identified by the data shared lasso and the nine "extended" models used to derive these fully cancer-type specific associations reached statistical significance (Figure S4, Additional file 2).

As displayed in Table 3 (third column), 15 out of the 22 associations identified by data shared lasso were replicated in more than 50% of the bootstrap samples. As displayed in Table 4, three inverse cancer type-specific associations that were not identified by the data shared lasso on the original sample were identified in more than 55% of the bootstrap samples: the cluster comprising glycine with endometrial cancer risk (identified in 65% of the bootstrap samples), the cluster containing decenoylcarnitine (acylcarnitine C10:1) with risk of kidney cancer (56%) and lysoPC a C16:1 with risk of localized prostate cancer (84%). Positive associations between arginine and kidney cancer risk (74%) and between the cluster containing lysoPC a C16:0 and localized prostate cancer risk (86%) were also observed in more than 55% of the bootstrap samples.

Analysis of the extended list of metabolites. After excluding 2,134 samples from the second colorectal cancer study which used a different platform that measured a lower number of metabolites, 16 additional metabolites could be evaluated (Table S1, Additional file 2). Among them, the clustering step grouped leucine and isoleucine together. The analysis of this extended list of metabolites then focused on 65 metabolites (31 isolated metabolites and 34 cluster representatives), measured in 9,522 participants. As displayed in Table 3, 11 out of the 22 associations identified in the main analysis presented above were again replicated in more than 50% of the bootstrap samples generated from this reduced sample. Four associations that were not identified in our previous analyses were identified in more than 55% of these new bootstrap samples (Table 4): an overall positive association between cancer risk and glutamate (55% of the bootstrap samples), an overall inverse association between cancer risk with spermine (78%), as well as two cancer type-specific associations between glutamate with breast cancer risk (inverse, 56%) and between serotonin and colorectal cancer risk (positive, 84%).

Univariate analyses. As displayed in Figure S5 in Additional file 2, the cancer type-specific univariate analyses identified associations with risks of breast cancer (two positive associations and four inverse), colorectal cancer (three inverse), endometrial

cancer (two inverse), kidney cancer (one inverse), HCC (four positive associations and 16 inverse) and advanced prostate cancer (seven inverse associations). The univariate pooled analysis identified 15 inverse associations, and there was no evidence of heterogeneity across cancer type for four of them (butyrylcarnitine, and the three clusters containing PC aa C32:2, PC ae C36:4, and PC ae C38:2, respectively).

Discussion

Using available metabolomics data from eight cancer-specific matched case-control studies nested within the EPIC cohort, we investigated the relationship between pre-diagnostic blood levels of over one hundred metabolites and risks of breast cancer, colorectal cancer, endometrial cancer, gallbladder and biliary tract cancer, HCC, kidney cancer, and localized and advanced prostate cancers. In our main analysis, we found nine metabolites associated with cancer risk across different cancer types, suggesting the existence of shared metabolic pathways, as well as fourteen cancer-type specific associations. These identified associations were found to be robust after extensive sensitivity analyses: in particular, they were not attenuated after exclusion of the first years of follow-up, hence were less likely to be due to reverse causality, were not attenuated after adjustment for relevant cancer risk factors, were not modified by BMI, and did not deviate significantly from linearity. In additional analyses, in particular those based on bootstrap samples, we identified several additional metabolites possibly associated with risk of specific cancer types or with cancer risk across different cancer types.

Our results suggested that concentrations of glycerophospholipids (phosphatidylcholines and lysophosphatidylcholines) could be linked to the risk of cancer overall as well as to specific cancer types. The role of glycerophospholipids in carcinogenesis is not fully understood but could be related to their documented anti-inflammatory properties, protection from oxidative stress, inhibition of cell proliferation and induction of apoptosis[43–45]. We observed a consistent inverse association between cancer risk with lysoPC a C18:2 as well as three clusters of phosphatidylcholines across all studied cancer types, except localized prostate cancer for which the association with lysoPC a C18:2 and one cluster of phosphatidylcholines was absent, or positive. An inverse association was previously reported between lysoPC a C18:2 with T2D in different studies[7,46] as well as with risks of breast, colorectal and prostate cancers in the pan-cancer analysis conducted in the EPIC Heidelberg study[25]. Our results regarding the three clusters of phosphatidylcholines were in line with many previously reported inverse associations between cancer and phosphatidylcholines[11,12,15,16,20,47]. Besides, we identified a positive association between the cluster that included PC aa C28:1 and cancer risk across all studied cancer types. This cluster also comprised PC ae C30:0, for which a positive association was reported with risks of breast, colorectal and prostate cancers in the EPIC Heidelberg study[25]. Cancer type-specific positive associations were found for the cluster containing PC aa

C36:5 with breast cancer, PC ae C36:0 with colorectal cancer, and the cluster containing PC aa C40:2 with HCC. These three clusters were correlated with one another (Pearson correlation greater than 0.48), indicating that higher levels of these phosphatidylcholines might contribute to the development of these three cancer types.

We also observed robust associations between specific circulating amino acids and cancer risk. Our results suggested that proline was positively related to cancer risk across all studied cancer types, except breast cancer and possibly HCC (see Figure S4 in Additional file 2). A positive association between proline and prostate cancer risk was previously reported in EPIC[12]. In addition, a drosophila model of high-sugar diet[48] recently highlighted the possible role of proline in tumour growth, and proline was also found to distinguish colorectal cancer patients from those with adenomas[49], and to be associated with metastasis formation[50]. Glutamine was inversely associated with overall cancer risk in our analysis, while glutamate, a metabolite of glutamine, was positively related to the risk of all cancer types except for breast cancer. Although prior studies of the French E3N and SU.VI.MAX cohorts reported a positive association between glutamine and premenopausal breast cancer[51,52], our results regarding glutamine and glutamate were consistent with those of many previous studies that reported inverse associations between glutamine and risk of colorectal cancer[18], HCC[19,53] and T2D[7,54], and positive associations between glutamate and risk of premenopausal breast cancer[52], kidney cancer[15], HCC[19,53], as well as T2D[7]. Lower serum levels of glutamine were also observed in kidney cancer[55] and ovarian cancer[56] cases compared to controls. Glutamine is an energy substrate for cancer cells and makes a major contribution to nitrogen metabolism. Alterations in glutamine-glutamate equilibrium often reflect energetic processes related to cancer metabolism[57]. It is possible that altered levels of glutamine and glutamate in individuals subsequently diagnosed with cancer may reflect ongoing metabolic processes related to cancer development and as such may serve as an early biomarker of cancer risk. However, the inverse association between glutamine levels and overall cancer risk observed in our analysis was only slightly attenuated after excluding, in turn, the first two and the first seven years of follow-up suggesting that changes in the glutamine-glutamate may precede cancer development.

Our analysis additionally identified two positive and two inverse cancer type-specific associations with circulating amino acids. We observed an inverse association between colorectal risk and the cluster containing histidine, for which previous studies reported inverse associations with risks of colorectal cancer and T2D[54], while a positive association was reported with breast cancer[52]. Also, lower serum levels of histidine were previously reported in ovarian cancer cases compared to controls[58]. Our results further suggested an inverse association between endometrial cancer risk and the cluster composed of glycine and serine, in line with previous results from the EPIC cancer-specific study of endometrial cancer[14]. Previous studies also reported inverse associations between glycine and/or serine with risks of T2D[54]. Finally, our analysis suggested a positive association between arginine with risks of colorectal and kidney

cancers (Table 4). Arginine was previously found to be positively associated with breast cancer in the E3N cohort[52], while an inverse association with breast cancer was reported in EPIC[11].

Regarding the biogenic amines, we found a positive association between serotonin levels and colorectal cancer risk, consistent with previous results from the CORSA case-control study and a previous EPIC analysis of colon cancer[59]. We also found a consistent inverse association between spermine and risk of the eight studied cancer types. Like other polyamines, spermine is involved in cell proliferation and differentiation and has antioxidant properties[60], and dysregulation of polyamines metabolism is characteristic of multiple types of tumours[61]. It was previously reported that polyamine supplementation, in particular spermidine, which acts as an intermediate in the conversion of putrescine to spermine, could be related to reduced overall and cancer-specific mortality[62–64].

In our analysis, localized and advanced prostate cancers were considered as two different outcomes as previous results suggested that metabolic dysregulation might be predictive of advanced or aggressive prostate cancers only[12]. In fact, we observed some differences between the metabolites associated with risks of localized and advanced prostate cancers, respectively. Specifically, and as previously reported[12,13], our results suggested that hexoses, glycerophospholipids, octadecenoylcarnitine and/or octadecadienylcarnitine could help differentiate the respective mechanisms involved in the development of aggressive and localized prostate tumours.

Some metabolites identified in our study were previously associated with established cancer risk factors, such as obesity [32,33]. In particular, a recent metabolomics study of BMI reported inverse associations with glutamine, lysophosphatidylcholine a C18:2 and phosphatidylcholine PC aa C38:0 (which was clustered with PC aa C36:0 in our analysis), and a positive association with glutamate[33]. Directions of the associations with BMI were consistent with those identified in our study with cancer risk after adjustment for BMI, indicating that these metabolites might be mediators of the obesity-cancer relationship.

Our study has several strengths. First, it relied on a large sample of pre-diagnostic metabolomics data acquired among 5,828 case-control pairs in nested studies on eight cancer types within a large prospective cohort, on average 6.4 years before cases developed cancer. Second, in a context where some metabolites might be predictive of cancer risk for multiple cancer types, the data shared lasso used in our analysis automatically accounted for or ignored cancer types when assessing the association between each metabolic feature with cancer risk, depending on whether heterogeneity among the cancer type-specific associations was supported by the data for that particular feature. The comparison of results produced by the standard univariate analyses and the data shared lasso illustrated the interest of the latter. First, the data shared lasso benefited from the increased statistical power of the pooled analysis for the identification of metabolites that could be involved in cancer development for

multiple cancer types: for example, butyrylcarnitine (acylcarnitine C4) was not associated with cancer risk in any of the cancer type-specific univariate analyses, while it was in the univariate pooled analysis and in the data shared lasso analysis. Moreover, unlike the simple pooled analysis, the data shared lasso would not necessarily mask cancer type-specific associations: for example, the data shared lasso identified a positive association between the cluster containing tetradecenoylcarnitine (acylcarnitine C14:1) and breast cancer risk, as the univariate analysis of the breast cancer study did, while the univariate pooled analysis could not. Another key difference between the standard univariate analyses and the data shared lasso is that the latter allowed the investigation of mutually adjusted associations, hence the identification of metabolites or clusters of metabolites whose association with cancer risk could not be explained away by other metabolites included in our analysis. Further, mutual adjustment revealed associations that could not be detected in minimally-adjusted models, such as the one between arginine and colorectal cancer risk, which was not apparent in models not adjusted for glutamine and histidine. Another strength of our study stemmed from the extensive sensitivity analyses that we carried out.

On the other hand, identifying cancer risk factors is particularly challenging when candidate risk factors are strongly correlated with one another. Here, we clustered the most strongly correlated metabolites together prior to applying the data shared lasso. As a sensitivity analysis, the data shared lasso was applied to the original set of 117 metabolites, thus ignoring the clustering step, and results were largely consistent with those of our main analysis (Figure S6, Additional file 2). Moreover, because strong correlations remained among some of the metabolites produced by the hierarchical clustering (Figure S7, Additional file 2), we applied the data shared lasso to multiple bootstrap samples to gauge the robustness and specificity of the associations identified in our main analysis. Although most of the identified associations were replicated in a large proportion of bootstrap samples, a few of them were less robust, hence more questionable. For example, the identified inverse association between HCC risk and the cluster that included lysoPC a C20:3 was replicated in 32% of the bootstrap samples only. This lack of robustness could be due to the strong correlation between this cluster and the other three studied metabolites related to lysoPCs (Pearson correlation greater than 0.65; Figure S7 in Additional file 2). As a matter of fact, an inverse association between HCC risk and at least one of the four metabolites related to lysoPCs was identified in 78% of the bootstrap samples. Overall, these results were suggestive of a stronger inverse association with features related to lysoPCs for HCC compared to the other cancer types, but our analysis failed to unambiguously identify which specific lysoPCs might underlie this stronger inverse association. An additional limitation for interpreting the lipid results is the lack of specificity for lipids measured with the AbsoluteIDQ p180/p150 kits as a result of the FIA method[65,66]. Moreover, the limited sample size for some of the studied cancer types (in particular gallbladder and biliary tract cancer and HCC) was a limitation for the identification of cancer type-specific

deviations. In this respect, we complemented our analysis by the inspection of estimates computed under models derived from the one identified by the data shared lasso but that further allowed fully type-specific associations (Figure S4, Additional file 2). Another potential limitation of our study was the lack of repeated measurements, yet previous studies suggested that blood levels of metabolites were relatively stable and that a single measurement might be sufficient to capture medium term exposure[67–69].

Conclusions

Our results confirmed the complex link between metabolism and cancer risk and highlighted the potential of metabolomics to identify possible informative markers associated with cancer risk and to gain insights into the biological mechanisms leading to cancer development. Our study indicated that specific metabolite families might be related to the risk of multiple cancer types. Some of these metabolites could reflect biological mechanisms underlying the carcinogenic effects of some established cancer risk factors, including obesity.

List of abbreviations

Adv.PrC: advanced prostate cancer; BMI: body mass index; BrC: breast cancer; CRC: colorectal cancer; CVD: cardio-vascular diseases; EnC: endometrial cancer; EPIC: European Prospective Investigation into Cancer and nutrition.; FIA: flow injection analysis; FDR: false discovery rate; GBC: gallbladder and biliary tract cancer; HCC: hepatocellular carcinoma; HZM: Helmholtz Zentrum; IARC: International Agency for Research on Cancer; ICL: Imperial College London; KiC: kidney cancer; Lasso: least absolute shrinkage and selection operator; LC: liquid chromatography; LLOQ: lower limit of quantification; Loc.PrC: localized prostate cancer; LOD: limit of detection; lysoPC: lysophosphatidylcholine; MS/MS: tandem mass spectrometry; OLS: ordinary least square regression; OR: odds ratio; PC: phosphatidylcholine; PCA: principal component analysis; SM: sphingomyelin; T2D: type 2 diabetes; ULOQ: upper limit of quantification.

Additional files

Additional file 1: Supplementary material regarding (i) the definition of cancer cases for HCC, GBC, Adv.PrC and Loc.PrC; (ii) the definition and implementation of the data-shared lasso; and (iii) the models used to derive point estimates and confidence intervals from the model selected by the data-shared lasso. (.docx 42 kb)

Additional file 2: Supplementary tables and figures. Figure S1: Pearson correlation between the 117 original metabolites. Figure S2: sensitivity analyses of mutually

adjusted ORs for the overall associations and cancer type-specific deviations. Figure S3: p-values of tests for departure from linearity and effect modification by BMI. Figure S4: ORs for the overall associations identified by the data-shared lasso with (i) the original model (ii) the extended type-specific model. Figure S5: results from the univariate analyses. Figure S6: Comparison of the associations identified by the data shared lasso when working with the 50 features (as in our main analysis) or with the original 117 metabolites. Figure S7: Pearson correlation between the 50 clusters. Table S1: list of the 117 metabolites studied in the main analysis, and of the 16 additional metabolites studied when excluding the second colorectal study. (.docx 2,698 kb)

Declarations

Ethics approval and consent: The EPIC study, and in particular the seven case-control studies nested within EPIC, were conducted according to the Declaration of Helsinki, and approved by the ethics committee at the International Agency for Research on Cancer (IARC): on 10 April 2008 (IEC 08-06) and on 11 February 2016 (IEC 16-06) for the liver cancer study, on 7 April 2014 (IEC 14-07) for the breast cancer study, on 7 April 2014 (IEC 14-08) for the two colorectal cancer studies, on 7 April 2014 (IEC 14-09) for the prostate cancer study, on 25 February 2015 (IEC 15-06) for the kidney cancer study, on 28 April 2016 (IEC 16-20) for the endometrial cancer study. Written informed consent was obtained from all subjects involved in the study.

Availability of data and materials: The R scripts developed to implement the analyses will be made available on the GitHub platform, for easy access to all interested scientists. The EPIC data is not publicly available, but access requests can be submitted to the Steering Committee (https://epic.iarc.fr/access/submit_appl_access.php).

Competing interest: The authors declare that they have no competing interests.

Funding: The coordination of EPIC is financially supported by International Agency for Research on Cancer (IARC) and 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).
The national cohorts are supported by: Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); German Cancer Aid, German Cancer Research Center (DKFZ), German Institute of Human Nutrition Potsdam- Rehbruecke (DIfE), Federal Ministry of Education and Research (BMBF) (Germany); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy, Compagnia di SanPaolo and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer

Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); 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); Swedish Cancer Society, Swedish Research Council and County Councils of Skåne and Västerbotten (Sweden); Cancer Research UK (14136 to EPIC-Norfolk; C8221/A29017 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk; MR/M012190/1 to EPIC-Oxford) (United Kingdom). IDIBELL acknowledges support from the Generalitat de Catalunya through the CERCA Program. The breast cancer study was funded by the French National Cancer Institute (grant number 2015-166). The colorectal cancer studies were funded by World Cancer Research Fund (reference: 2013/1002; www.wcrf.org/), the European Commission (FP7: BBMRI-LPC; reference: 313010; https://ec.europa.eu/). The endometrial cancer study was funded by Cancer Research UK (grant number C19335/A21351). The kidney study was funded by the World Cancer Research Fund (MJ; reference: 2014/1193; www.wcrf.org/) and the European Commission (FP7: BBMRI-LPC; reference: 313010; https://ec.europa.eu/). The liver cancer study was supported in part by the French National Cancer Institute (L'Institut National du Cancer; INCa; grant numbers 2009-139 and 2014-1-RT-02-CIRC-1) and by internal funds of the IARC. For the participants in the prostate cancer study, sample retrieval and preparation, and assays of metabolites were supported by Cancer Research UK (C8221/A19170), and funding for grant 2014/1183 was obtained from the World Cancer Research Fund (WCRF UK), as part of the World Cancer Research Fund International grant programme. Mathilde His' work reported here was undertaken during the tenure of a postdoctoral fellowship awarded by the International Agency for Research on Cancer, financed by the Fondation ARC. The funders were not involved in designing the study; collecting, analyzing and interpreting results; or in writing and submitting the manuscript for publication.

Authors contributions: The authors responsibilities were as follows: PF, MJG and VV conceived, designed and supervised the research. MB and VV analyzed the data. MB, PF, MJG and VV were responsible for drafting the manuscript. LD, MJen, MJoh, SR, RCT and MJG conducted and supervised metabolomics analyses. LD, MJen, Mjoh, SR, RCT, MH, TJK, JAS, KO, AT, CK, JAR, NL, GS, RK, VK, MSB, FE, DP, SG, SP, RT, CS, BBdM, KSO, TMS, THN, JRQ, CB, MRB, MDC, EA, MS, JM, LV, MR, DM, KT, AKH, HK, JA, PKR, AS and MJG provided the original data, information on the respective populations, and advice on the study design, analysis and interpretation of the results. All authors read and approved of the final manuscript.

IARC disclaimer: Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent

659 the decisions, policy, or views of the International Agency for Research on
660 Cancer/World Health Organization.
661
662
663 **References**
664 1. Beger RD. A review of applications of metabolomics in cancer. *Metabolites*.
665 2013;3(3):552-574. doi:10.3390/metabo3030552
666 2. Scalbert A, Huybrechts I, Gunter MJ. The Food Exposome. In: Dagnino S,
667 Macherone A, eds. *Unraveling the Exposome*. Springer International Publishing;
668 2019:217-245. doi:10.1007/978-3-319-89321-1_8
669 3. Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. The Blood
670 Exposome and Its Role in Discovering Causes of Disease. *Environ Health Perspect*.
671 2014;122(8):769-774. doi:10.1289/ehp.1308015
672 4. González-Domínguez R, Jáuregui O, Queipo-Ortuño MI, Andrés-Lacueva C.
673 Characterization of the Human Exposome by a Comprehensive and Quantitative
674 Large-Scale Multianalyte Metabolomics Platform. *Anal Chem*. 2020;92(20):13767-13775.
675 doi:10.1021/acs.analchem.0c02008
676 5. Gonzalez-Franquesa A, Burkart AM, Isganaitis E, Patti ME. What have
677 metabolomics approaches taught us about Type 2 Diabetes? *Curr Diab Rep*.
678 2016;16(8):74. doi:10.1007/s11892-016-0763-1
679 6. Ahola-Olli AV, Mustelin L, Kalimeri M, et al. Circulating metabolites and the
680 risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish
681 cohorts. *Diabetologia*. 2019;62(12):2298-2309. doi:10.1007/s00125-019-05001-w
682 7. Sun Y, Gao HY, Fan ZY, He Y, Yan YX. Metabolomics Signatures in Type 2
683 Diabetes: A Systematic Review and Integrative Analysis. *The Journal of Clinical*
684 *Endocrinology & Metabolism*. 2020;105(4):1000-1008. doi:10.1210/clinem/dgz240
685 8. McGarrah RW, Crown SB, Zhang GF, Shah SH, Newgard CB. Cardiovascular
686 Metabolomics. *Circ Res*. 2018;122(9):1238-1258. doi:10.1161/CIRCRESAHA.117.311002
687 9. Cavus E, Karakas M, Ojeda FM, et al. Association of Circulating Metabolites
688 With Risk of Coronary Heart Disease in a European Population: Results From the
689 Biomarkers for Cardiovascular Risk Assessment in Europe (BiomarCaRE)
690 Consortium. *JAMA Cardiology*. 2019;4(12):1270-1279. doi:10.1001/jamacardio.2019.4130
691 10. Müller J, Bertsch T, Volke J, et al. Narrative review of metabolomics in
692 cardiovascular disease. *J Thorac Dis*. 2021;13(4):2532-2550. doi:10.21037/jtd-21-22
693 11. His M, Viallon V, Dossus L, et al. Prospective analysis of circulating metabolites
694 and breast cancer in EPIC. *BMC Med*. 2019;17(1):178. doi:10.1186/s12916-019-1408-4
695 12. Schmidt JA, Fensom GK, Rinaldi S, et al. Pre-diagnostic metabolite
696 concentrations and prostate cancer risk in 1077 cases and 1077 matched controls in the
697 European Prospective Investigation into Cancer and Nutrition. *BMC Med*.
698 2017;15(1):122. doi:10.1186/s12916-017-0885-6
699 13. Schmidt JA, Fensom GK, Rinaldi S, et al. Patterns in metabolite profile are
700 associated with risk of more aggressive prostate cancer: A prospective study of 3,057
701 matched case-control sets from EPIC. *Int J Cancer*. 2020;146(3):720-730.

doi:10.1002/ijc.32314

14. Dossus L, Kouloura E, Biessy C, et al. Prospective analysis of circulating metabolites and endometrial cancer risk. *Gynecologic Oncology*. Published online June 5, 2021. doi:10.1016/j.ygyno.2021.06.001

15. Guida F, Tan VY, Corbin LJ, et al. The blood metabolome of incident kidney cancer: A case–control study nested within the MetKid consortium. *PLOS Medicine*. 2021;18(9):e1003786. doi:10.1371/journal.pmed.1003786

16. Shu X, Xiang YB, Rothman N, et al. Prospective study of blood metabolites associated with colorectal cancer risk. *Int J Cancer*. 2018;143(3):527-534. doi:10.1002/ijc.31341

17. Harlid S, Gunter MJ, Van Guelpen B. Risk-Predictive and Diagnostic Biomarkers for Colorectal Cancer; a Systematic Review of Studies Using Pre-Diagnostic Blood Samples Collected in Prospective Cohorts and Screening Settings. *Cancers*. 2021;13(17):4406. doi:10.3390/cancers13174406

18. Rothwell JA, Bešević J, Dimou N, et al. Circulating amino acid levels and colorectal cancer risk in the European Prospective Investigation into Cancer and Nutrition and UK Biobank cohorts (In preparation).

19. Stepien M, Duarte-Salles T, Fedirko V, et al. Alteration of amino acid and biogenic amine metabolism in hepatobiliary cancers: Findings from a prospective cohort study. *Int J Cancer*. 2016;138(2):348-360. doi:10.1002/ijc.29718

20. Shu X, Zheng W, Yu D, et al. Prospective metabolomics study identifies potential novel blood metabolites associated with pancreatic cancer risk. *Int J Cancer*. 2018;143(9):2161-2167. doi:10.1002/ijc.31574

21. Zeleznik OA, Clish CB, Kraft P, Avila-Pacheco J, Eliassen AH, Tworoger SS. Circulating Lysophosphatidylcholines, Phosphatidylcholines, Ceramides, and Sphingomyelins and Ovarian Cancer Risk: A 23-Year Prospective Study. *J Natl Cancer Inst*. 2020;112(6):628-636. doi:10.1093/jnci/djz195

22. Deng T, Lyon CJ, Bergin S, Caligiuri MA, Hsueh WA. Obesity, Inflammation, and Cancer. *Annu Rev Pathol*. 2016;11:421-449. doi:10.1146/annurev-pathol-012615-044359

23. Wiebe N, Stenvinkel P, Tonelli M. Associations of Chronic Inflammation, Insulin Resistance, and Severe Obesity With Mortality, Myocardial Infarction, Cancer, and Chronic Pulmonary Disease. *JAMA Netw Open*. 2019;2(8):e1910456. doi:10.1001/jamanetworkopen.2019.10456

24. Li Y, Schoufour J, Wang DD, et al. Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: prospective cohort study. *BMJ*. Published online January 8, 2020:l6669. doi:10.1136/bmj.l6669

25. Kühn T, Floegel A, Sookthai D, et al. Higher plasma levels of lysophosphatidylcholine 18:0 are related to a lower risk of common cancers in a prospective metabolomics study. *BMC Med*. 2016;14:13. doi:10.1186/s12916-016-0552-3

26. Gross SM, Tibshirani R. Data Shared Lasso: A Novel Tool to Discover Uplift. *Comput Stat Data Anal*. 2016;101:226-235. doi:10.1016/j.csda.2016.02.015

27. Ollier E, Viallon V. Regression modelling on stratified data with the lasso.

Biometrika. 2017;104(1):83-96. doi:10.1093/biomet/asw065

28. Ballout N, Garcia C, Viallon V. Sparse estimation for case-control studies with multiple disease subtypes. *Biostatistics*. Published online January 24, 2020. doi:10.1093/biostatistics/kxz063

29. Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr*. 2002;5(6B):1113-1124. doi:10.1079/PHN2002394

30. Viallon V, His M, Rinaldi S, et al. A New Pipeline for the Normalization and Pooling of Metabolomics Data. *Metabolites*. 2021;11(9):631. doi:10.3390/metabo11090631

31. Chavent M, Kuentz-Simonet V, Liquet B, Saracco J. ClustOfVar: An R Package for the Clustering of Variables. *Journal of Statistical Software*. 2012;50:1-16. doi:10.18637/jss.v050.i13

32. Carayol M, Leitzmann MF, Ferrari P, et al. Blood Metabolic Signatures of Body Mass Index: A Targeted Metabolomics Study in the EPIC Cohort. *J Proteome Res*. 2017;16(9):3137-3146. doi:10.1021/acs.jproteome.6b01062

33. Kliemann N, Viallon V, Murphy N, et al. Metabolic signatures of greater body size and their associations with risk of colorectal and endometrial cancers in the European Prospective Investigation into Cancer and Nutrition. *BMC Med*. 2021;19(1):101. doi:10.1186/s12916-021-01970-1

34. Pischon T, Nimptsch K. *Obesity and Cancer. Recent Results in Cancer Research*. Spinger; 2016.

35. Fortner RT, Katzke V, Kühn T, Kaaks R. Obesity and Breast Cancer. *Recent Results Cancer Res*. 2016;208:43-65. doi:10.1007/978-3-319-42542-9_3

36. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat Rev Gastroenterol Hepatol*. 2019;16(12):713-732. doi:10.1038/s41575-019-0189-8

37. Capitanio U, Bensalah K, Bex A, et al. Epidemiology of Renal Cell Carcinoma. *Eur Urol*. 2019;75(1):74-84. doi:10.1016/j.eururo.2018.08.036

38. Dashti SG, English DR, Simpson JA, et al. Adiposity and Endometrial Cancer Risk in Postmenopausal Women: A Sequential Causal Mediation Analysis. *Cancer Epidemiol Biomarkers Prev*. 2021;30(1):104-113. doi:10.1158/1055-9965.EPI-20-0965

39. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996;58(1):267-288. doi:10.1111/j.2517-6161.1996.tb02080.x

40. Bach FR. Bolasso: model consistent Lasso estimation through the bootstrap. In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Association for Computing Machinery; 2008:33-40. doi:10.1145/1390156.1390161

41. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *The Annals of Statistics*. 2004;32(2):407-499. doi:10.1214/009053604000000067

42. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995;57(1):289-300.

788 43. Treede I, Braun A, Sparla R, et al. Anti-inflammatory effects of
789 phosphatidylcholine. *J Biol Chem*. 2007;282(37):27155-27164.
790 doi:10.1074/jbc.M704408200
791 44. Hannun YA, Obeid LM. Principles of bioactive lipid signalling: lessons from
792 sphingolipids. *Nat Rev Mol Cell Biol*. 2008;9(2):139-150. doi:10.1038/nrm2329
793 45. Beloribi-Djefaflia S, Vasseur S, Guillaumond F. Lipid metabolic reprogramming
794 in cancer cells. *Oncogenesis*. 2016;5:e189. doi:10.1038/oncsis.2015.49
795 46. Klein MS, Shearer J. Metabolomics and Type 2 Diabetes: Translating Basic
796 Research into Clinical Application. *J Diabetes Res*. 2016;2016:3898502.
797 doi:10.1155/2016/3898502
798 47. Stepien M, Keski-Rahkonen P, Kiss A, et al. Metabolic perturbations prior to
799 hepatocellular carcinoma diagnosis: Findings from a prospective observational cohort
800 study. *International Journal of Cancer*. 2021;148(3):609-625. doi:10.1002/ijc.33236
801 48. Newton H, Wang YF, Camplese L, et al. Systemic muscle wasting and
802 coordinated tumour response drive tumourigenesis. *Nat Commun*. 2020;11:4653.
803 doi:10.1038/s41467-020-18502-9
804 49. Gumpenberger T, Brezina S, Keski-Rahkonen P, et al. Untargeted Metabolomics
805 Reveals Major Differences in the Plasma Metabolome between Colorectal Cancer and
806 Colorectal Adenomas. *Metabolites*. 2021;11(2):119. doi:10.3390/metabo11020119
807 50. Elia I, Broekaert D, Christen S, et al. Proline metabolism supports metastasis
808 formation and could be inhibited to selectively target metastasizing cancer cells. *Nat*
809 *Commun*. 2017;8(1):15267. doi:10.1038/ncomms15267
810 51. Lécuyer L, Dalle C, Lyan B, et al. Plasma Metabolomic Signatures Associated
811 with Long-term Breast Cancer Risk in the SU.VI.MAX Prospective Cohort. *Cancer*
812 *Epidemiol Biomarkers Prev*. 2019;28(8):1300-1307. doi:10.1158/1055-9965.EPI-19-0154
813 52. Jobard E, Dossus L, Baglietto L, et al. Investigation of circulating metabolites
814 associated with breast cancer risk by untargeted metabolomics: a case-control study
815 nested within the French E3N cohort. *Br J Cancer*. 2021;124(10):1734-1743.
816 doi:10.1038/s41416-021-01304-1
817 53. Fages A, Duarte-Salles T, Stepien M, et al. Metabolomic profiles of
818 hepatocellular carcinoma in a European prospective cohort. *BMC Med*. 2015;13:242.
819 doi:10.1186/s12916-015-0462-9
820 54. Pietzner M, Stewart ID, Raffler J, et al. Plasma metabolites to profile pathways
821 in noncommunicable disease multimorbidity. *Nature Medicine*. Published online
822 March 11, 2021:1-9. doi:10.1038/s41591-021-01266-0
823 55. Gao H, Dong B, Liu X, Xuan H, Huang Y, Lin D. Metabonomic profiling of renal
824 cell carcinoma: High-resolution proton nuclear magnetic resonance spectroscopy of
825 human serum with multivariate data analysis. *Analytica Chimica Acta*. 2008;624(2):269-
826 277. doi:10.1016/j.aca.2008.06.051
827 56. Plewa S, Horała A, Dereziński P, et al. Usefulness of Amino Acid Profiling in
828 Ovarian Cancer Screening with Special Emphasis on Their Role in Cancerogenesis. *Int*
829 *J Mol Sci*. 2017;18(12):E2727. doi:10.3390/ijms18122727
830 57. Yi H, Talmon G, Wang J. Glutamate in cancers: from metabolism to signaling. *J*

Biomed Res. 2019;34(4):260-270. doi:10.7555/JBR.34.20190037

58. Plewa S, Horała A, Dereziński P, Nowak-Markwitz E, Matysiak J, Kokot ZJ. Wide spectrum targeted metabolomics identifies potential ovarian cancer biomarkers. *Life Sci*. 2019;222:235-244. doi:10.1016/j.lfs.2019.03.004

59. Papadimitriou N, Gunter MJ, Murphy N, et al. Circulating tryptophan metabolites and risk of colon cancer: Results from case-control and prospective cohort studies. *Int J Cancer*. 2021;149(9):1659-1669. doi:10.1002/ijc.33725

60. Muñoz-Esparza NC, Latorre-Moratalla ML, Comas-Basté O, Toro-Funes N, Veciana-Nogués MT, Vidal-Carou MC. Polyamines in Food. *Frontiers in Nutrition*. 2019;6:108. doi:10.3389/fnut.2019.00108

61. Moinard C, Cynober L, de Bandt JP. Polyamines: metabolism and implications in human diseases. *Clin Nutr*. 2005;24(2):184-197. doi:10.1016/j.clnu.2004.11.001

62. Vargas AJ, Ashbeck EL, Wertheim BC, et al. Dietary polyamine intake and colorectal cancer risk in postmenopausal women. *Am J Clin Nutr*. 2015;102(2):411-419. doi:10.3945/ajcn.114.103895

63. Pietrocola F, Castoldi F, Kepp O, Carmona-Gutierrez D, Madeo F, Kroemer G. Spermidine reduces cancer-related mortality in humans. *Autophagy*. 2018;15(2):362-365. doi:10.1080/15548627.2018.1539592

64. Fan J, Feng Z, Chen N. Spermidine as a target for cancer therapy. *Pharmacological Research*. 2020;159:104943. doi:10.1016/j.phrs.2020.104943

65. Koelmel JP, Ulmer CZ, Jones CM, Yost RA, Bowden JA. Common cases of improper lipid annotation using high-resolution tandem mass spectrometry data and corresponding limitations in biological interpretation. *Biochim Biophys Acta*. 2017;1862(8):766-770. doi:10.1016/j.bbalip.2017.02.016

66. Köfeler HC, Ahrends R, Baker ES, et al. Recommendations for good practice in MS-based lipidomics. *J Lipid Res*. 2021;62:100138. doi:10.1016/j.jlr.2021.100138

67. Floegel A, Drogan D, Wang-Sattler R, et al. Reliability of serum metabolite concentrations over a 4-month period using a targeted metabolomic approach. *PLoS One*. 2011;6(6):e21103. doi:10.1371/journal.pone.0021103

68. Townsend MK, Clish CB, Kraft P, et al. Reproducibility of metabolomic profiles among men and women in 2 large cohort studies. *Clin Chem*. 2013;59(11):1657-1667. doi:10.1373/clinchem.2012.199133

69. Carayol M, Licaj I, Achaintre D, 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. doi:10.1371/journal.pone.0135437

Tables and Figures

Table 1. Description of the original seven cancer-specific matched case-control studies nested within EPIC

Cancer site	Number of Samples	Matrix	Laboratory	Kit Used
Breast	3,172	Citrate plasma[1]	IARC	p180
Colorectal (Study 1)	946	Citrate plasma	IARC	p180
Colorectal (Study 2)	2,295	Serum	HZM[2]	p150
Endometrial	1,706	Citrate plasma	ICL[3]	p180
Liver	662	Serum	IARC	p180
Kidney	1,213	Citrate plasma	IARC	p180
Prostate	6,020	Citrate plasma	IARC	p180

[1]except Swedish participants (n=101; EDTA plasma). [2]Helmhotz Zentrum München. [3]Imperial College London

Table 2. Main characteristics of the control (Ctrl) and case (Case) sub-populations in the eight cancer type-specific EPIC studies

	BrC study		CRC study		EnC study		KiC study		GBTC study		HCC study		Adv. PrC study		Loc. PrC study	
	N = 1,088 pairs		N = 1,500 pairs		N = 689 pairs		N = 511 pairs		N = 85 pairs		N = 121 pairs		N = 533 pairs		N = 1301 pairs	
	Ctrl	Case	Ctrl	Case	Ctrl	Case	Ctrl	Case	Ctrl	Case	Ctrl	Case	Ctrl	Case	Ctrl	Case
Age at blood collection																
Mean (SD)	51.8 (8.31)	51.8 (8.33)	57.0 (7.58)	57.1 (7.57)	54.3 (7.83)	54.3 (7.84)	55.8 (8.47)	55.8 (8.46)	58.7 (7.13)	58.7 (7.08)	59.9 (7.01)	59.9 (6.98)	57.6 (7.18)	57.6 (7.18)	57.9 (6.80)	58.0 (6.80)
Age at cancer diagnosis																
Mean (SD)	- -	60.4 (8.83)	- -	64.9 (8.18)	- -	62.7 (8.16)	- -	64.5 (8.83)	- -	64.9 (7.60)	- -	66.1 (7.49)	- -	66.3 (7.02)	- -	67.1 (6.36)
Sex																
Female	1088 (100%)	1088 (100%)	769 (51.3%)	769 (51.3%)	689 (100%)	689 (100%)	197 (38.6%)	197 (38.6%)	48 (56.5%)	48 (56.5%)	35 (28.9%)	35 (28.9%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
BMI (kg/m²)																
Mean (SD)	25.7 (4.32)	26.2 (4.80)	26.5 (3.88)	27.2 (4.34)	26.0 (4.29)	28.2 (5.52)	26.7 (3.84)	27.8 (4.47)	26.9 (4.38)	27.3 (3.98)	26.9 (3.72)	28.4 (4.73)	26.7 (3.48)	27.0 (3.20)	27.5 (3.49)	27.2 (3.37)
Education																
None	56 (5.1%)	62 (5.7%)	124 (8.3%)	136 (9.1%)	67 (9.7%)	76 (11.0%)	42 (8.2%)	34 (6.7%)	6 (7.1%)	6 (7.1%)	6 (5.0%)	7 (5.8%)	28 (5.3%)	32 (6.0%)	106 (8.1%)	136 (10.5%)
Primary school completed	397 (36.5%)	377 (34.7%)	574 (38.3%)	526 (35.1%)	269 (39.0%)	231 (33.5%)	184 (36.0%)	186 (36.4%)	33 (38.8%)	39 (45.9%)	49 (40.5%)	52 (43.0%)	160 (30.0%)	166 (31.1%)	444 (34.1%)	411 (31.6%)
Technical/professional school	245 (22.5%)	254 (23.3%)	333 (22.2%)	334 (22.3%)	124 (18.0%)	118 (17.1%)	107 (20.9%)	109 (21.3%)	20 (23.5%)	15 (17.6%)	31 (25.6%)	38 (31.4%)	140 (26.3%)	127 (23.8%)	305 (23.4%)	306 (23.5%)
Secondary school	158 (14.5%)	178 (16.4%)	188 (12.5%)	227 (15.1%)	100 (14.5%)	127 (18.4%)	66 (12.9%)	72 (14.1%)	11 (12.9%)	8 (9.4%)	11 (9.1%)	5 (4.1%)	58 (10.9%)	59 (11.1%)	103 (7.9%)	91 (7.0%)
Longer education (incl. University deg.)	211 (19.4%)	195 (17.9%)	241 (16.1%)	227 (15.1%)	100 (14.5%)	100 (14.5%)	96 (18.8%)	93 (18.2%)	15 (17.6%)	17 (20.0%)	22 (18.2%)	17 (14.0%)	132 (24.8%)	124 (23.3%)	306 (23.5%)	324 (24.9%)
Not specified	21 (1.9%)	22 (2.0%)	40 (2.7%)	50 (3.3%)	29 (4.2%)	37 (5.4%)	16 (3.1%)	17 (3.3%)	0 (0%)	0 (0%)	2 (1.7%)	2 (1.7%)	15 (2.8%)	25 (4.7%)	37 (2.8%)	33 (2.5%)

Table 3: Robustness of the associations identified in the main analysis. For each identified association, the proportion of bootstrap samples on which it was replicated is reported (in bold when ≥50%).

Feature	Cancer Type*	Proportion of bootstrap samples[1]	Proportion of bootstrap samples[2]
Overall associations with cancer risk			
c10	Overall	**62%**	**59%**
c4	Overall	47%	39%
gln	Overall	**73%**	**76%**
pro	Overall	**65%**	**50%**
lysopc_a_c18_2	Overall	**57%**	47%
pc_aa_c28_1_Clus	Overall	**57%**	**64%**
pc_aa_c32_2_Clus	Overall	49%	**71%**
pc_aa_c36_0_Clus	Overall	**86%**	**95%**
pc_aa_c36_1_Clus	Overall	**50%**	40%
Cancer type-specific associations			
c14_1_Clus	BrC	**80%**	**76%**
pro	BrC	**77%**	**70%**
pc_aa_c36_5_Clus	BrC	36%	47%
arg	CRC	**88%**	19%
his_Clus	CRC	**81%**	**72%**
pc_ae_c36_0	CRC	**80%**	46%
sm_c16_0_Clus	EnC	**85%**	**87%**
lysopc_a_c20_3_Clus	HCC	32%	47%
pc_aa_c40_2_Clus	HCC	**61%**	34%
sm_c16_0_Clus	HCC	**90%**	**78%**
c18_1_Clus	Adv.PrC	40%	49%
lysopc_a_c18_2	Loc.PrC	14%	23%
pc_aa_c36_0_Clus	Loc.PrC	49%	41%

* BrC stands for breast cancer, CRC for colorectal cancer, EnC for endometrial cancer, HCC for hepatocellular carcinoma, and Adv.Prc Loc.PrC for advanced and localized prostate cancers, respectively.

[1] Bootstrap samples were generated from the original sample of 5,828 matched case-control pairs with information on 117 metabolites (corresponding to 50 features after the clustering step).

[2] Bootstrap samples were generated from the original sample which comprised 4,761 matched case-control pairs with information on 133 metabolites (corresponding to 65 features after the clustering step) after excluding the participants of the second CRC study.

Table 4: Other associations identified in a large proportion of the bootstrap samples. Associations identified in at least 55% of both bootstrap analyses are reported, along with the proportion of bootstrap samples in which they were identified, and the corresponding average log odds-ratio (as estimated by the data shared lasso on each bootstrap sample).

Feature	Cancer Type*	Proportion of bootstrap samples[1]	Average log-OR[1]	Proportion of bootstrap samples[2]	Average log-OR[2]
Overall associations with cancer risk					
glu	Overall	--	--	55%	0.09
spermine	Overall	--	--	78%	-0.10
Type-specific associations					
gly_Clus	EnC	65%	-0.17	78%	-0.14
c10_1_Clus	KiC	56%	-0.18	56%	-0.17
lysopc_a_c16_1	Loc.PrC	84%	-0.19	78%	-0.18
arg	KiC	74%	0.23	71%	0.21
lysopc_a_c16_0_Clus	Loc.PrC	86%	0.24	79%	0.22
glu	BrC	--	--	56%	-0.14
serotonin	CRC	--	--	84%	0.35

* BrC stands for breast cancer, CRC for colorectal cancer, EnC for endometrial cancer, KiC for Kidney cancer, and Loc.PrC for localized prostate cancer.

[1] Bootstrap samples were generated from the original sample of 5,828 matched case-control pairs with information on 117 metabolites (corresponding to 50 features after the clustering step).

[2] Bootstrap samples were generated from the original sample which comprised 4,761 matched case-control pairs with information on 133 metabolites (corresponding to 65 features after the clustering step) after excluding the participants of the second CRC study.

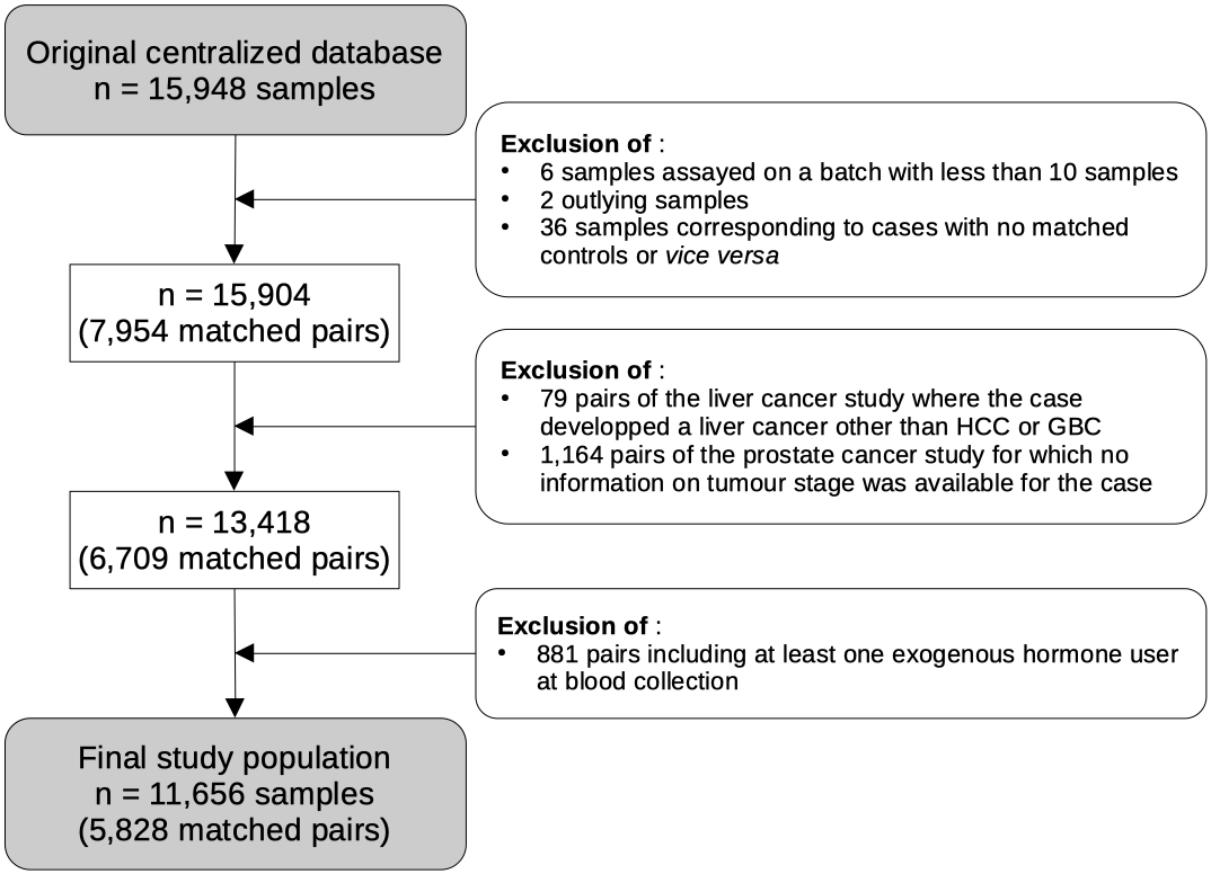

Figure 1. Flowchart summarizing the exclusion criteria to derive the final sample used in our main analysis. GBC stands for gallbladder and biliary tract cancer and HCC for hepatocellular carcinoma.

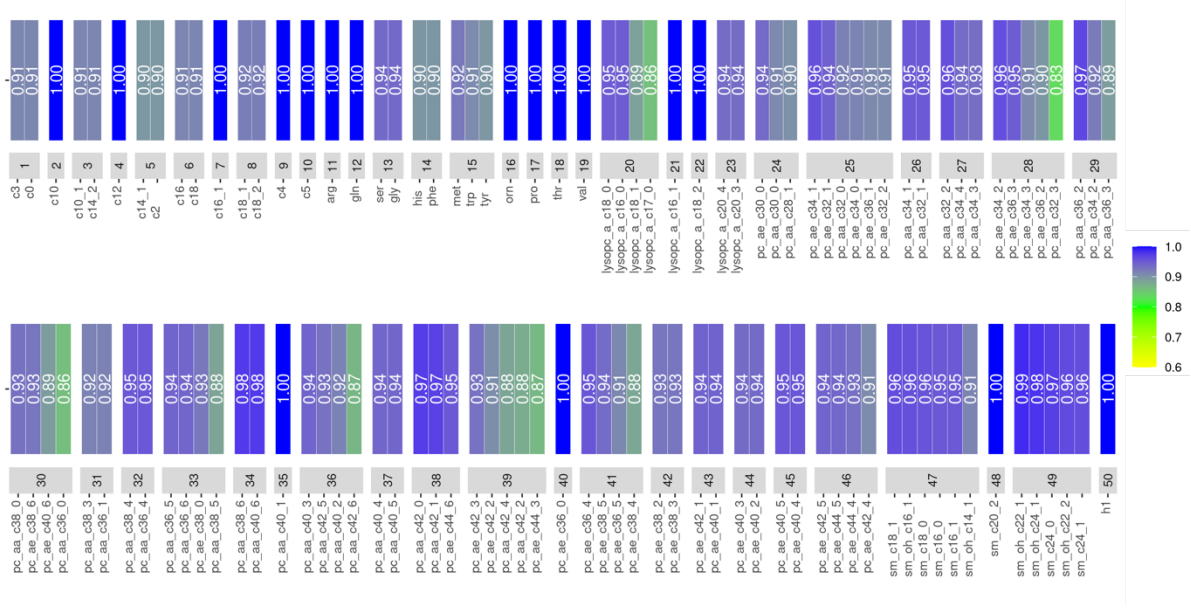

Figure 2. Description of the 50 "metabolites" retained for the main analysis, which includes 33 clusters of strongly correlated metabolites and 17 "isolated" metabolites. For example, the 19th metabolite is an isolated metabolite (valine), while the 26th one is a cluster made of two phosphatidylcholines. For each cluster, correlations between its representative and the individual metabolites that compose that cluster are represented as a heat map (this correlation is 1 when the "cluster" is actually reduced to an isolated metabolite)

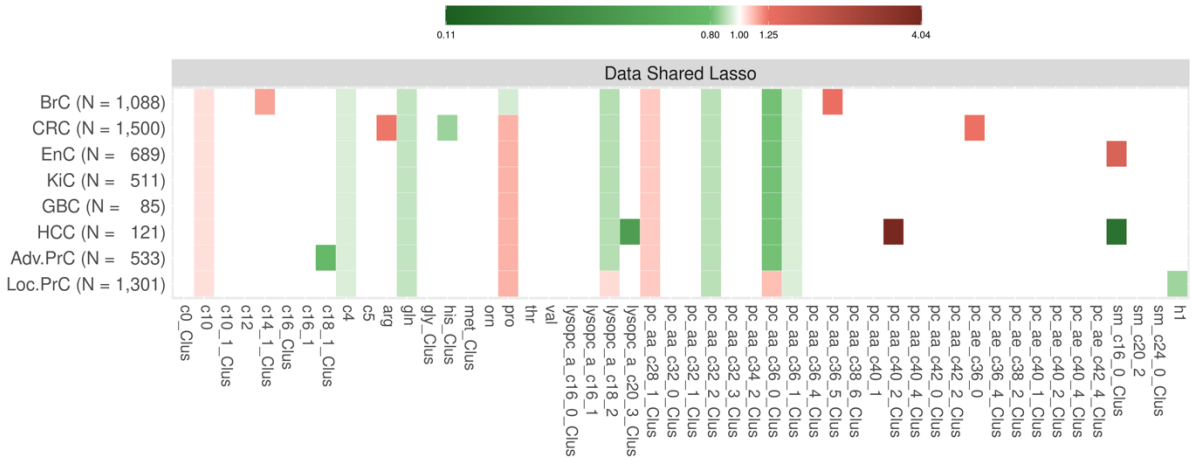

Figure 3. Summary of the main results from our the multivariate pan-cancer analysis, which evaluated mutually adjusted associations between each feature (more precisely, their residuals after adjustment for BMI) and the risk of the eight cancer types, using a data shared lasso penalty. White entries correspond to the absence of identified associations, while green and red entries correspond to inverse and positive associations, respectively. The more intense the colour, the larger the absolute value of the log-odds-ratio (that were re-estimated in multivariate unpenalized conditional regression models; see Section 3.a in the Supplementary Material for details). The x-axis represents the 50 features (33 cluster representatives and 17 isolated metabolites). In the labels of the y-axis, numbers correspond to numbers of pairs for each type-specific cancer (and in total for the pooled analysis), while BrC stands for breast cancer, CRC for colorectal cancer, EnC for endometrial cancer, KiC for Kidney cancer, GBC for gallbladder and biliary tract cancer, HCC for hepatocellular carcinoma, and Adv.PrC and Loc.PrC for advanced and localized prostate cancers, respectively.

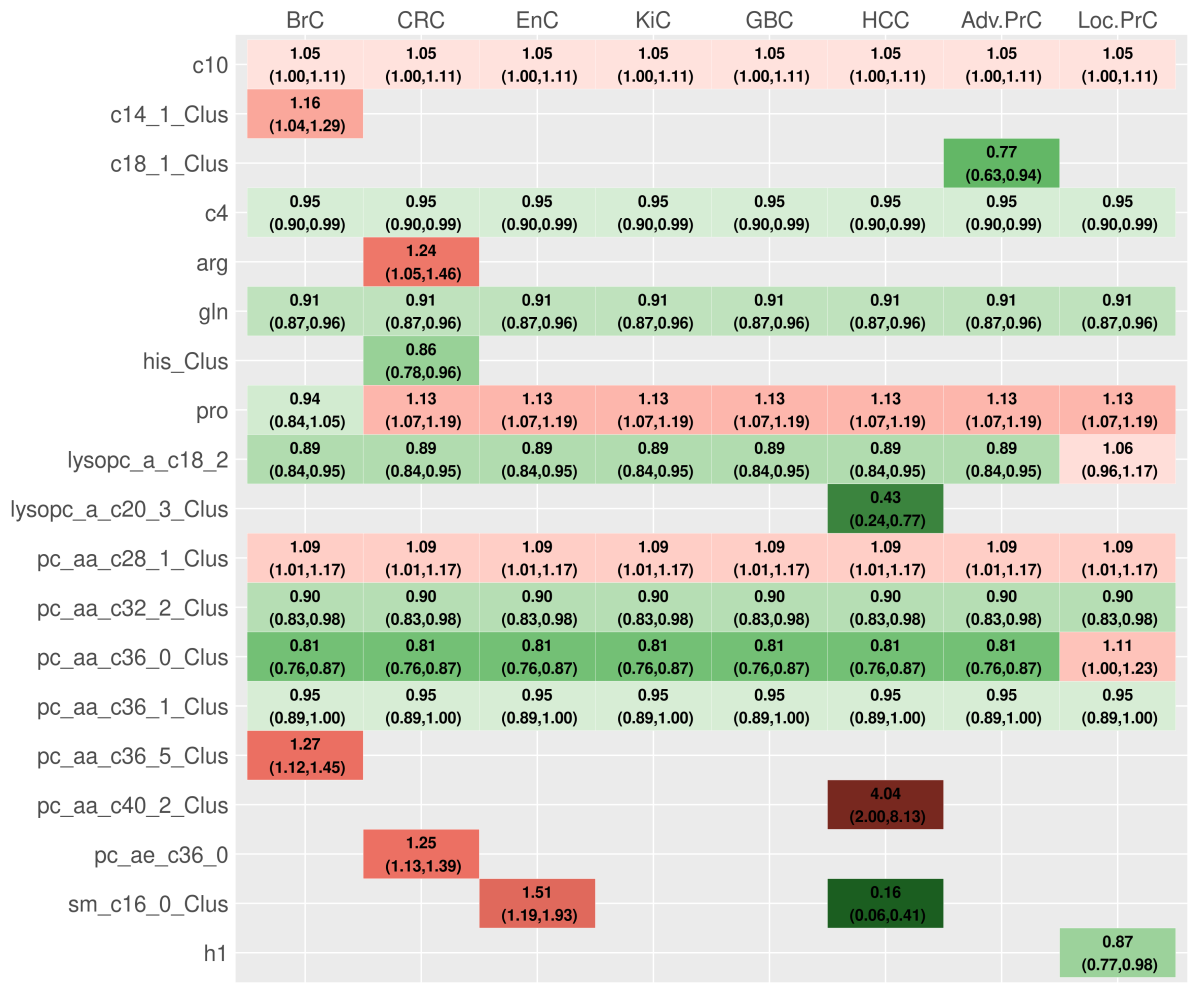

Figure 4. Summary of the mutually adjusted associations between the 50 metabolic features and risks of the eight cancer types, as identified by the data shared lasso. Only the 19 features (8 isolated metabolites and 11 cluster representatives) for which the data shared lasso identified an association with at least one cancer type are presented on the y axis. Point estimates and 95% confidence intervals of the corresponding odds-ratios were obtained through non-penalized conditional logistic regression models using the design matrix derived from the positions of the non-zero components in the data shared lasso vector estimate $(\hat{\boldsymbol{\mu}},\ \hat{\boldsymbol{\delta}}_1, \cdots, \hat{\boldsymbol{\delta}}_K)$; see Section 3.a in the Supplementary Material for details. They have to be interpreted with caution since they are the result of post-selection inference. In the labels of the columns, BrC stands for breast cancer, CRC for colorectal cancer, EnC for endometrial cancer, KiC for Kidney cancer, GBC for gallbladder and biliary tract cancer, HCC for hepatocellular carcinoma, and Adv.PrC and Loc.PrC for advanced and localized prostate cancers, respectively.