Untargeted Metabolic Profiling Identifies Altered Serum Metabolites of Type 2 Diabetes Mellitus in a Prospective, Nested Case Control Study

Dagmar Drogan,[1]* Warwick B. Dunn,[2,3,4,5] Wanchang Lin,[2,3,4] Brian Buijsse,[1] Matthias B. Schulze,[6] Claudia Langenberg,[7] Marie Brown,[2,3,4] Anna Floegel,[1] Stefan Dietrich,[1] Olov Rolandsson,[8] David C. Wedge,[4,9] Royston Goodacre,[4] Nita G. Forouhi,[7] Stephen J. Sharp,[7] Joachim Spranger,[10] Nick J. Wareham,[7] and Heiner Boeing[1]

BACKGROUND: Application of metabolite profiling could expand the etiological knowledge of type 2 diabetes mellitus (T2D). However, few prospective studies apply broad untargeted metabolite profiling to reveal the comprehensive metabolic alterations preceding the onset of T2D.

METHODS: We applied untargeted metabolite profiling in serum samples obtained from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam cohort comprising 300 individuals who developed T2D after a median follow-up time of 6 years and 300 matched controls. For that purpose, we used ultraperformance LC-MS with a protocol specifically designed for large-scale metabolomics studies with regard to robustness and repeatability. After multivariate classification to select metabolites with the strongest contribution to disease classification, we applied multivariable-adjusted conditional logistic regression to assess the association of these metabolites with T2D.

RESULTS: Among several alterations in lipid metabolism, there was an inverse association with T2D for metabolites chemically annotated as lysophosphatidylcholine(dm16:0) and phosphatidylcholine(O-20:0/O-20:0). Hexose sugars were positively associated with T2D, whereas higher concentrations of a sugar alcohol and a deoxyhexose sugar reduced the odds of diabetes by approximately 60% and 70%, respectively. Furthermore,

there was suggestive evidence for a positive association of the circulating purine nucleotide isopentenyladenosine-5′-monophosphate with incident T2D.

CONCLUSIONS: This study constitutes one of the largest metabolite profiling approaches of T2D biomarkers in a prospective study population. The findings might help generate new hypotheses about diabetes etiology and develop further targeted studies of a smaller number of potentially important metabolites.

© 2014 American Association for Clinical Chemistry

Type 2 diabetes mellitus (T2D)[11] is a serious metabolic disorder whose prevalence in adults has doubled over nearly 3 decades *(1)*. Insulin resistance and impaired insulin secretion are common features of T2D, but the complex metabolic alterations contributing to T2D are still not completely understood. Broad metabolite profiling of human biospecimens may improve our understanding of these alterations, because the metabolome is the ultimate manifestation of environmental and genetic risk factors *(2)*.

Recent applications of metabolite profiling in cross-sectional or case control studies have linked alterations in hexoses, amino acids, ketone bodies, fatty acids, acylcarnitines, and phospholipids with T2D *(3–16)*. To date, however, there have been just a few applications of metabolite profiling in prospective studies, which are more suitable to distinguish between metabolic causes and the
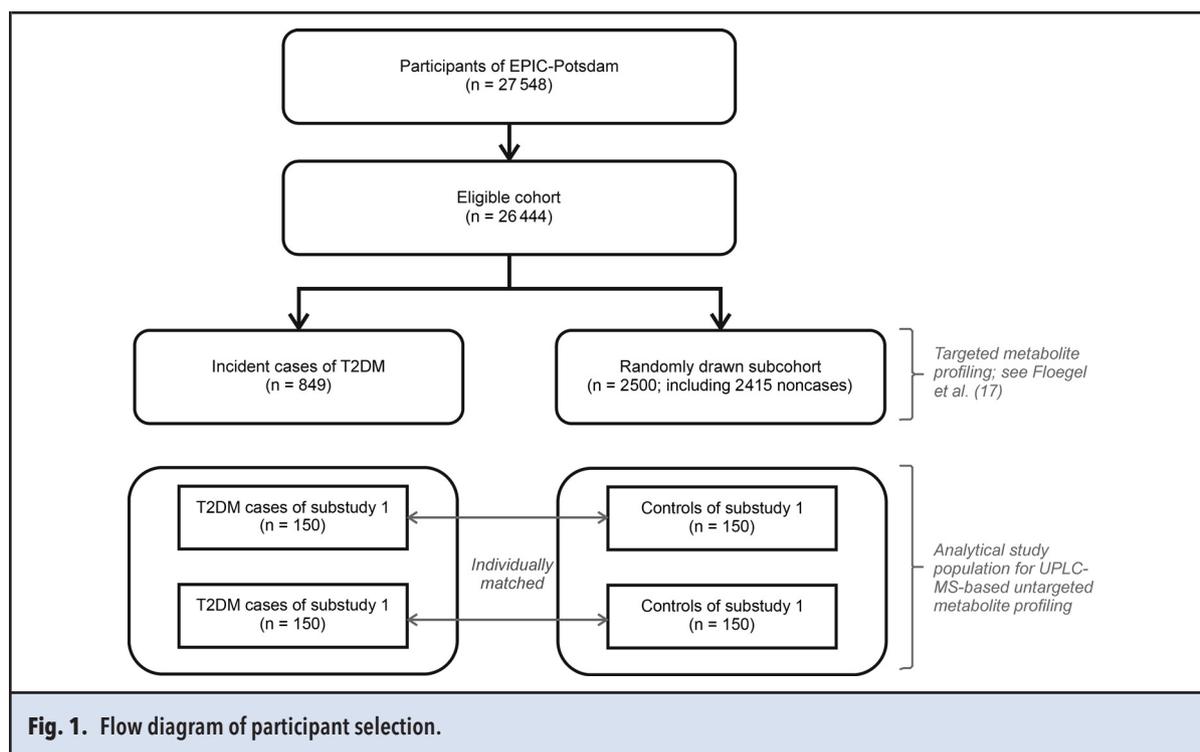
Fig. 1. Flow diagram of participant selection.

actual consequences of T2D *(17–22)*. Of note, these studies applied targeted metabolomic approaches, in which measurement was restricted to a limited number (typically <200) of selected and biochemically annotated metabolites. However, few prospective studies use an untargeted approach that may allow for a broader assessment of metabolic alterations preceding the clinical stage of T2D. Recently, an ultraperformance LC (UPLC)-MS–based method has been designed to fulfill the unique requirements of applying high-throughput analytical methodologies in large-scale epidemiological studies, specifically with regard to robustness and repeatability *(23)*. We used this untargeted metabolite profiling method to analyze metabolic alterations linked to incident T2D in a case control study embedded into the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam cohort.

Materials and Methods

STUDY POPULATION

The EPIC-Potsdam study included 27 548 adults aged 35–65 years. Baseline examination (1994–1998) included blood sampling, measurements of blood pressure (BP) and anthropometric parameters, dietary and life-style questionnaires, and personal computer-assisted interviews *(24)*. The study procedures were approved by the Ethics Committee of the Medical Association of the State of Brandenburg (Germany). All procedures fol-

lowed the Helsinki Declaration for human rights, and participants provided written informed consent. Potentially incident cases with self-reported diabetes diagnosis, relevant medication, or dietary treatment owing to diabetes were verified by questionnaires mailed to the diagnosing physician for confirming diagnosis.

Our case control study was embedded into a case-cohort study of EPIC-Potsdam comprising 849 individuals with incident T2D and a random subcohort of 2500 individuals that was recently used for targeted metabolite profiling *(17)* (Fig. 1). For untargeted metabolite profiling, we randomly selected 300 cases of incident T2D. Each case was individually matched to 1 nondiabetic control according to the following matching criteria: age (±6 months), sex, fasting time (<3 h, ≥3 h, <6 h, and ≥6 h before blood draw), time of day of blood sampling (±2 h), and season at blood sampling. Furthermore, each nondiabetic control participant was chosen to have at least the diabetes-free follow-up time of the respective case. The study population was randomly split into 2 independent substudies, each containing 150 matched case-control pairs to allow 2 separate metabolite profiling studies to be performed to reduce the false-positive rate in any single study.

ASSESSMENT OF BASELINE COVARIATES AND CLINICAL BIOMARKERS

Anthropometric parameters were measured by trained personnel. Smoking habits, physical activity, educational

attainment, and medical history were assessed during a standardized interview. Dietary habits during the previous year were assessed by use of a validated, self-administered food frequency questionnaire (25). Systolic and diastolic BP were measured after a resting period of 15–30 min (26). Prevalent hypertension was defined as systolic BP ≥140 mmHg, diastolic BP ≥90 mmHg, self-reported hypertension diagnosis, or use of antihypertensive medication. Additionally, 30 mL peripheral blood was obtained by venipuncture according to a standardized protocol (27). Before centrifugation, 20 mL was filled into monovettes containing sodium citrate as anticoagulant and the remaining 10 mL into monovettes without anticoagulant. After centrifugation at 1500g for 20 min, blood fractions (serum, plasma, buffy coat, and erythrocytes) were placed in 0.5-mL plastic straws and stored at −196 °C. We measured plasma concentrations of glucose, glycated hemoglobin (Hb A_{1C}), alanine aminotransferase (ALT), γ-glutamyltransferase (GGT), high-sensitivity C-reactive protein (hs-CRP), triglycerides, and HDL cholesterol with the automatic Advia 1650 chemistry system (Siemens Medical Solutions). We measured plasma adiponectin by ELISA (Linco Research).

METABOLITE PROFILING

Metabolite profiling data of 600 serum samples was acquired separately for 2 independent substudies (substudy 1 and substudy 2; 300 samples per study). We prepared serum samples (140-μL aliquots) by deproteinization and drying in a randomized order and according to a protocol described in Supplemental Methods, which accompanies the online version of this article at http://www.clinchem.org/content/vol61/issue3, and elsewhere (23, 28). We pooled sera from each participant (60 μL) to create a single pooled QC sample, from which 140-μL aliquots were deproteinized and dried as described above. These QC samples were applied for conditioning of the analytical system, signal correction, and quality assurance as previously described (23, 29).

All samples were reconstituted in 70 μL of 50:50 methanol/water followed by centrifugation for 15 min at 13 363g. We analyzed samples in substudies 1 and 2 in a random order, with each substudy consisting of 3 analytical batches. Batches 1–3 were applied in substudy 1 and batches 4–6 were applied in substudy 2. The respective batches were composed of 100 subject samples and 28 intermediate QC injections. We carried out analytical UPLC-MS measurements on an Acquity UPLC system (Waters) that was interfaced with a LTQ-Orbitrap XL hybrid mass spectrometer equipped with an electrospray ionization source (Thermo Fisher Scientific). The UPLC method was operated over a 22-min run time as described in online Supplemental Methods and as described previously (28). The same gradient was applied for positive and negative ion mode to enable accurate integration of data for metabolite identification for both ion modes. Fifty percent of the column eluent was introduced to the electrospray source of the LTQ-Orbitrap XL mass spectrometer. The mass spectrometer was operated in positive and negative ion modes separately, and data were acquired in the range 50–1000 m/z with a scan rate of 0.4 s. After each analytical batch, the column was washed for 30 min with 100% methanol.

RAW DATA PREPROCESSING

Raw data acquired in each analytical batch were converted from the instrument-specific format to the NetCDF format applying the file converter program in XCalibur software (Thermo Fisher Scientific). Deconvolution was performed with XCMS software according to settings described previously (30, 31). We carried out signal correction and quality assurance with the QC samples present in substudies 1 and 2 separately and in positive and negative ion modes. As described previously (23), we applied univariate local polynomial regression fitting to each ion detected in QC samples and regressed to the same ion in subject samples. Metabolites were retained only if at least 60% of the QC peaks were observed in each of the 3 batches for a single study and differed from the mean QC peak area by ≤20%. Metabolites with a retention time <40 or >1100 s were excluded from the analyses.

METABOLITE IDENTIFICATION

Putative annotation of metabolites or metabolite groups was performed by applying the PUTMEDID-LCMS workflows operating in the Taverna workflow environment (23). We applied 5 ppm mass error and a retention time range of ±2 s in feature grouping and molecular formula and metabolite matching. Because different metabolites can be detected with the same accurate m/z (for example, isomers with the same molecular formula), multiple annotations could be observed for a single detected metabolite feature. Also, a single metabolite could be detected as multiple molecules, particularly as a different type of ion (e.g., protonated and sodiated ions). Throughout this article, the term "metabolite" refers to either single metabolites or groups of molecules with the same retention time and the same accurate m/z. All molecules were annotated according to guidelines for reporting of chemical analysis results, specifically to Metabolomics Standards Initiative level 2 (32). Phosphatidylcholine (PC), phosphatidylethanolamine, and lysophosphatidyl-choline (LysoPC) are characterized by x:y, with x indicating the number of carbon atoms and y indicating the number of double bonds. The O and P prefixes indicate the presence of an alkyl or an alkenyl ether substituent, respectively.

For metabolite selection in the 2 substudies, we first applied multivariate classification-based feature selection with random forests (RFs) *(33)* and partial least squares discriminant analysis (PLS-DA) *(34)*, using the R packages randomForest and pls. A detailed description is given in online Supplemental Methods. Briefly, for RF, ntree and mtry were fixed at 500 and one-third of predictor variables, respectively. In total, we selected 100 bootstrap samples with replacement on the matched case-control pairs. Metabolites were ranked according to the RF importance scores reflecting the difference between oob error of a variable randomly permuted and the original variable. For PLS-DA, metabolites of each substudy were ranked according to the regression coefficients representing the importance of each metabolite in classification. The results of both methods were aggregated to a ranking list by the mean of each ranking score, with a higher score indicating higher importance of a given metabolite with regard to classification of T2D cases and controls. We selected a subset of 60 predictors per ion mode and substudy for further statistical analyses, because larger subsets did not improve discrimination accuracy of the models (data not shown).

We used multivariable-adjusted conditional logistic regression to estimate odds ratios (ORs) and 95% CIs for the association of each metabolite with T2D. For that purpose, the participants were divided into tertiles on the basis of the study-specific distribution of the selected metabolites among controls. The regression models were adjusted for age, body mass index (BMI) (kg/m^2), waist circumference (cm), physical activity (inactive, moderately inactive, moderately active, active), smoking (never, former, current <20 U/day, current ≥20 U/day), education (vocational school or less, technical school, university), prevalent hypertension (yes/no), and consumption of alcoholic beverages (women: nonconsumer, >0–6 g/day, >6–12 g/day, >12 g/day; men: nonconsumer, >0–12 g/day, >12–24 g/day, >24 g/day), coffee (cups/day), red meat (g/day), and whole-grain bread (g/day). In 265 case-control pairs with complete data on traditional T2D biomarkers, we additionally adjusted for the plasma concentrations of HDL cholesterol, triglycerides, ALT, GGT, CRP, adiponectin, glucose, and Hb A$_{1c}$. For the set of metabolites selected by the multivariate classification methods, we also calculated a false discovery rate–adjusted *P* value for linear trend across tertiles within the whole study population by using the participants' assigned tertile as a continuous variable in the respective regression model. Spearman correlation coefficients were calculated to investigate the association between metabolites with selected risk factors of T2D. The correlation analyses were restricted to only those metabolites selected for multivariable-adjusted regression models, as described above. All regression and correlation analyses

were carried out with SAS software, version 9.2 (SAS Institute).

Results

Both substudies were comparable in terms of baseline characteristics (Table 1). Incident cases of T2D had higher mean BMI and higher mean waist circumference. Furthermore, individuals with T2D were more likely to be hypertensive and physically inactive compared with individually matched controls (Table 1). Because fasting time was one of the matching variables, cases and controls were identical in this respect, and the percentage of participants having a meal or beverage within 3 h before blood draw was 61% in substudy 1 and 59% in substudy 2. After signal correction and/or quality assurance processes, 3139 and 1381 metabolites were detected on the UPLC-MS positive ion mode platform and the UPLC-MS negative ion mode platform, respectively. Metabolites were ranked according to their ability for separating cases and controls with multivariate classification. Online Supplemental Table 1 displays the study-specific ranking scores for the top 60 metabolites per ion mode. In total, 206 metabolites or metabolite groups ranked high in at least 1 study population, and 34 metabolites ranked high in both.

Metabolites lacking chemical annotation were not considered for subsequent analyses. Among the chemically annotated metabolites, several nonlipids were significantly associated with T2D in both study populations after adjustment for relevant dietary and lifestyle factors (i.e., *P* value for the comparison between third and first tertile <0.05 in both substudies) (Fig. 2; online Supplemental Table 2). Specifically, higher concentrations of the purine nucleotide isopentenyladenosine-5′-monophosphate were associated with an approximate 5-fold increase in the odds of T2D. For the large group of hexose sugars and derivatives, both positive and negative associations were detected. For instance, metabolites annotated as "tetrahydroxyhexanoic acid and/or hexose sugar (e.g., glucose, fructose, inositol)" exhibited a strongly increased OR for the comparison between extreme tertiles. In contrast, individuals in the highest tertile of a deoxyhexose sugar (e.g., anhydro-glucitol, deoxygalactose, or deoxyglucose) had ORs of 0.23 (95% CI 0.07–0.72) and 0.23 (0.08–0.65) in the first and second substudy, respectively. Furthermore, higher concentrations of a sugar alcohol, 1-*O*-β-D-glucopyranosyl-D-mannitol or maltitol, were associated with a substantially reduced OR of T2D.

In total, 6 of the selected annotated lipids exhibited a significant association with T2D in both study populations, when multivariable-adjusted logistic regression was applied (Fig. 3). Specifically, individuals in the highest tertile of PC(22:4/dm18:0) or PC(O-18:0/22:5) had

Table 1. Baseline characteristics of incident cases of T2D and matched controls of 2 randomly selected case control studies nested within the EPIC-Potsdam cohort.[a]

Characteristic	Substudy 1			Substudy 2		
	Controls	Cases	*P*	Controls	Cases	*P*
n	150	150		150	150	
Age, years[b]	54.2 (7.5)	54.2 (7.5)	0.982	55.1 (7.6)	55.1 (7.6)	0.996
Male sex[b]	51.3	51.3	1.000	51.3	51.3	1.000
BMI, kg/m^2	26.2 (3.5)	30.3 (4.8)	<0.0001	26.2 (3.6)	30.6 (4.6)	<0.0001
Waist circumference, cm	87.8 (11.1)	99.4 (13.1)	<0.0001	88.0 (11.8)	99.4 (12.6)	<0.0001
Red meat, g/d	0.3 (0.2)	0.3 (0.2)	0.479	0.3 (0.2)	0.4 (0.2)	0.003
Whole-grain bread, g/d	1.0 (1.2)	0.6 (0.8)	0.003	1.0 (1.1)	0.8 (1.0)	0.119
Coffee, g/d	2.6 (1.8)	2.5 (2.0)	0.633	3.0 (2.0)	2.6 (1.9)	0.145
Alcohol, g/d	17.9 (21.1)	14.1 (19.1)	0.106	14.5 (15.4)	17.3 (24.3)	0.228
Physical inactivity	16.0	30.0	0.004	20.0	29.3	0.061
University degree	37.3	33.3	0.470	40.0	28.0	0.028
Smoker	20.7	20.0	0.886	19.3	23.3	0.399
Hypertension	40.0	63.3	<0.0001	38.7	68.0	<0.0001

[a] Data are mean (SD) or %.
[b] Matching factors.

a >4-fold higher chance of developing diabetes compared with individuals in the lowest tertile. In contrast, substantially lower odds of T2D were observed for a metabolite annotated as PC(O-20:0/O-20:0). For this glycerophospholipid, the OR between extreme tertiles was 0.19 (95% CI 0.06–0.59) in the first substudy and 0.28 (0.08–0.92) in the second substudy. Likewise, individuals in the highest tertile of LysoPC(dm16:0) had an OR of 0.19 (0.07–0.60) or 0.27 (0.10–0.75).

All selected metabolites displayed in Figs. 2 and 3 were still significantly associated with T2D after adjusting for the plasma concentrations of adiponectin, CRP, GGT, or ALT (data not shown). Furthermore, PC(O-20:0/O-20:0) was the only metabolite not significantly associated with T2D after adjusting for circulating HDL cholesterol and triglycerides. In contrast, adjustment for plasma glucose considerably weakened the metabolite–disease associations for molecules annotated as "tetrahydroxyhexanoic acid and/or hexose sugar (e.g., glucose, fructose, inositol)," isopentenyladenosine-5′-monophosphate, and hydroxy-methylbutenyl diphosphate.

When performing subgroup analyses in individuals fasting <3 h or ≥3 h before blood sampling, all selected metabolites exhibited the same direction of association with T2D as observed in the overall study population (data not shown). Among these metabolites, tetrahydroxyhexanoic acid was the only feature for which effect modification of fasting status on the association between metabolite and T2D was apparent (*P* for interaction < 0.05).

We also conducted correlation analyses between the selected serum metabolites and T2D risk factors including established biomarkers (Table 2). Plasma glucose correlated strongly with hydroxyl-methylbutenyl diphosphate (r_s = 0.52), isopentenyladenosine-5′-monophosphate (r_s = 0.44), and the detected hexose sugars. Hydroxyl-methylbutenyl diphosphate and isopentenyladenosine-5′-monophosphate also correlated positively with each other (r_s = 0.50) (see online Supplemental Table 3). In comparison, the correlation of the selected lipids with each other and with established T2D biomarkers was low (Table 2; online Supplemental Table 4).

Discussion

With coverage of >4500 metabolite features, this study constitutes one of the largest metabolite profiling approaches of T2D biomarkers in a prospective study population to date. Using UPLC-MS, we detected alterations in serum carbohydrates, purines, and phospholipids preceding the onset of T2D by about 6 years. Just recently, findings from well-powered prospective studies have provided deeper insights into early metabolic alterations predictive for future diabetes (17–22). A range of biologically important plasma metabolites were measured in these studies, but they applied a targeted metabolite profiling approach. The findings of our untargeted approach may thus complement current knowledge about diabetes pathogenesis.
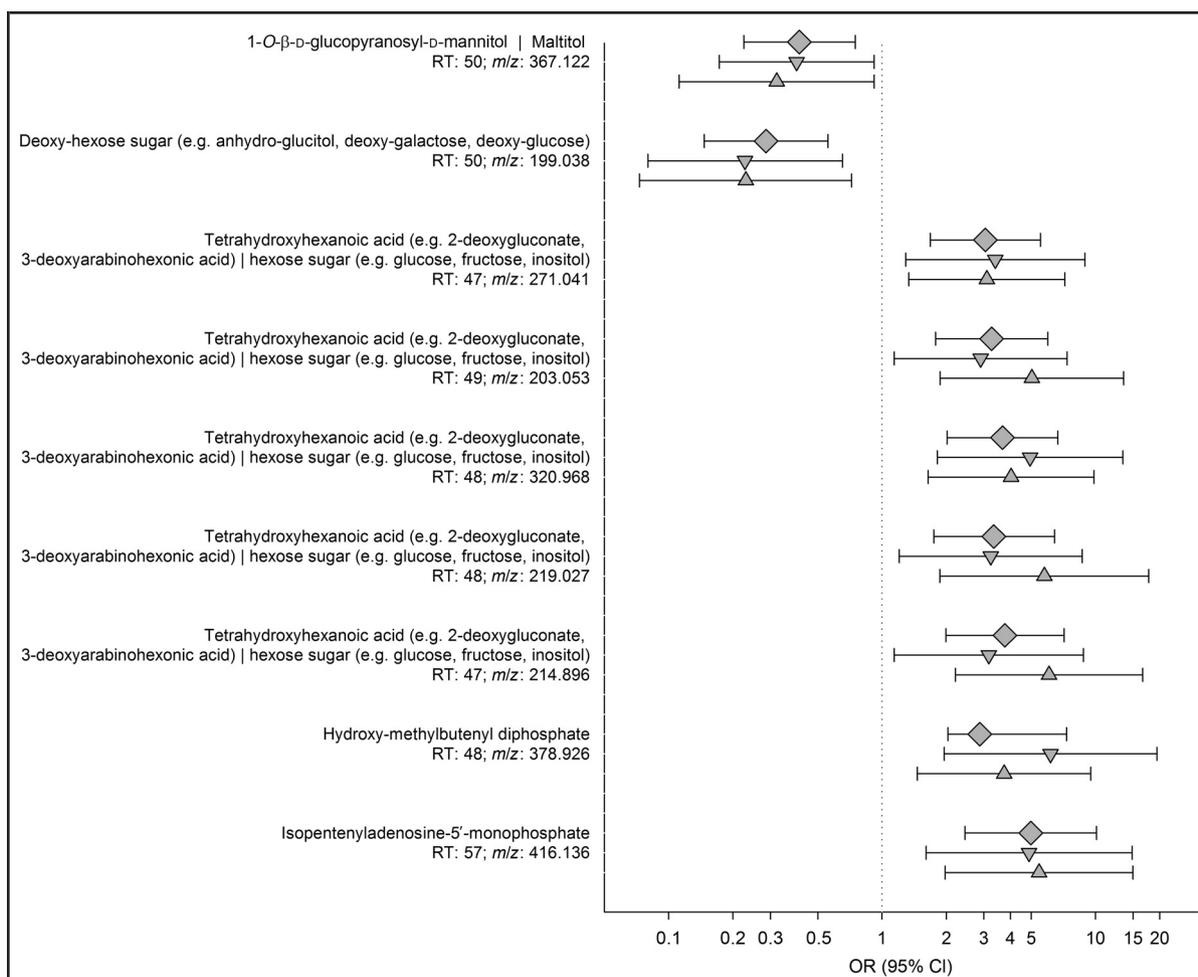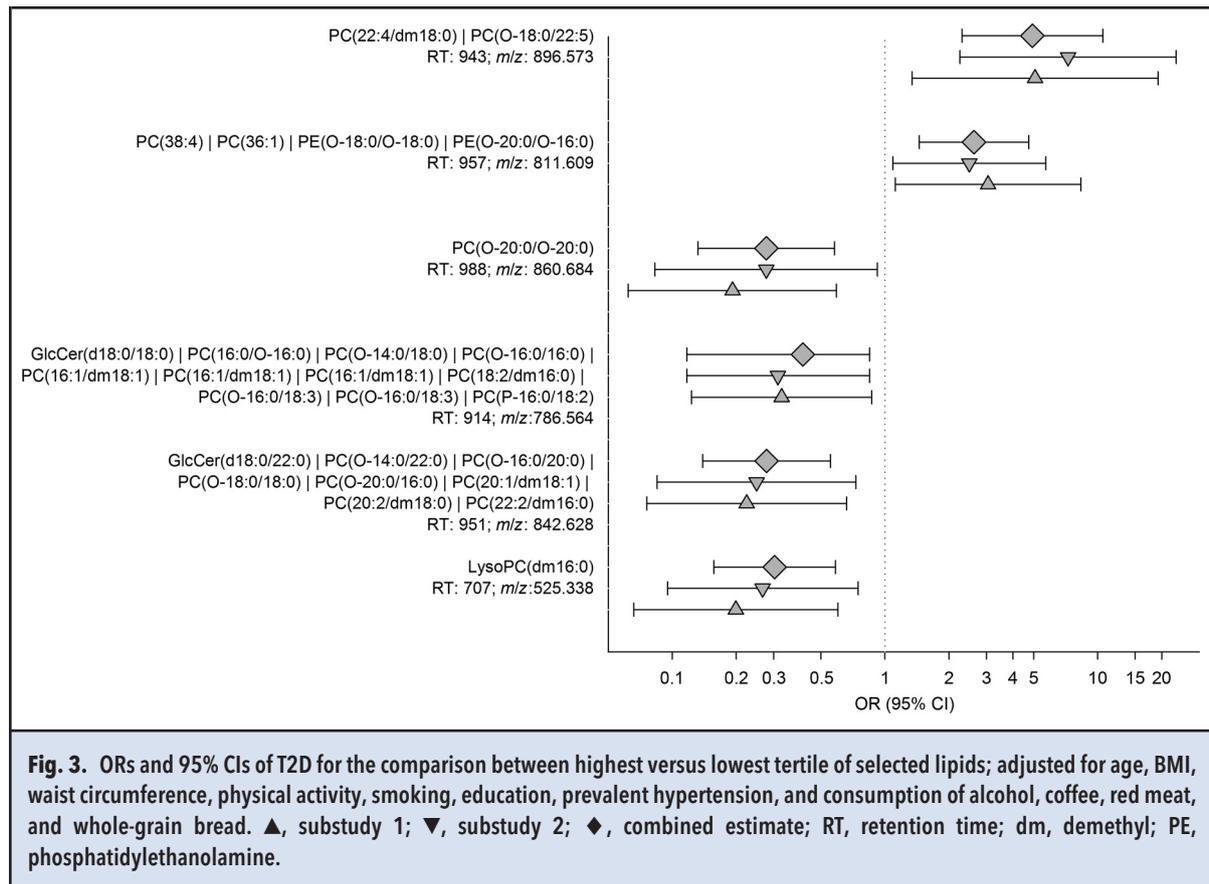
Fig. 2. ORs and 95% CIs of T2D for the comparison between highest versus lowest tertile of selected nonlipids; adjusted for age, BMI, waist circumference, physical activity, smoking, education, prevalent hypertension, and consumption of alcohol, coffee, red meat, and whole-grain bread. ▲, substudy 1; ▼, substudy 2; ◆, combined estimate; RT, retention time; dm, demethyl; PE, phosphatidylethanolamine.

Alterations in glycerophospholipids are common in the (pre)diabetic state *(5, 9, 14, 35)*, possible reflecting the impact of glycerophospholipids on membrane properties, cell integrity, and cellular signaling *(36)*. In our study, we observed a lower OR of T2D for some glycerophospholipids including LysoPC(dm16:0) and PC(O-20:0/O-20:0). However, neither metabolite has been linked to T2D before because of incomplete coverage of glycerophospholipids in recent profiling studies. Ether linkage–containing phospholipids may act as serum antioxidants to prevent lipoprotein oxidation, thereby potentially modifying diabetes risk *(37)*. Furthermore, previous targeted approaches highlighted differences in T2D incidence with regard to linkage and composition of fatty acid residues in glycerophospholipids *(17, 21)*.

T2D is characterized by hyperglycemia resulting from impaired insulin sensitivity and relative insulin deficiency *(38)*. Because glucose is the major monosaccharide in human blood, our observation of strong positive associations of several hexoses with T2D is in the expected direction. Adjustment for plasma glucose substantially attenuated these associations, suggesting that glucose contributes strongly to these hexoses. Furthermore, our results may reflect the impact of hyperglycemia on monosaccharides with the same accurately measured m/z, e.g., fructose *(3)* and mannose *(5)*.

Interestingly, we also observed an inverse association of deoxyhexose sugars (e.g., anhydro-glucitol, deoxygalactose, deoxyglucose) with T2D. Of note, a higher risk of T2D has been observed for increasing concentra-

Fig. 3. ORs and 95% CIs of T2D for the comparison between highest versus lowest tertile of selected lipids; adjusted for age, BMI, waist circumference, physical activity, smoking, education, prevalent hypertension, and consumption of alcohol, coffee, red meat, and whole-grain bread. ▲, substudy 1; ▽, substudy 2; ◆, combined estimate; RT, retention time; dm, demethyl; PE, phosphatidylethanolamine.

tions of deoxyhexose *(5)* and 3,6-anhydrogalactose *(3)*. However, Suhre et al. detected decreased concentrations of the hexose sugar 1,5-anhydroglucitol (1,5-AG) in diabetic patients in a multiplatform metabolomics study *(5)*. Because of renal loss during hyperglycemic conditions, low concentrations of 1,5-AG can be used as a marker of short-term glycemic control *(39)*. It is possible that this well-known metabolite contributes to the inverse association between the deoxyhexose sugar and T2D in our own study population. Yet, 1,5-AG has been shown to correlate strongly with Hb A_{1c} *(39)*, which is not the case with the deoxyhexose sugar detected in our data (Table 2).

Moreover, a sugar alcohol comprising 1-*O*-β-D-glucopyranosyl-D-mannitol and/or maltitol was negatively related to T2D in our data. *O*-β-D-glucopyranosyl-D-mannitol (Isomalt) and maltitol are widely used sweeteners with prebiotic properties, thereby potentially modifying obesity and diabetes risk *(40)*. To our knowledge, increased concentrations of *O*-β-D-glucopyranosyl-D-mannitol or maltitol have not been associated with a lower risk of T2D before, but the lack of unique metabolic annotations hampers the interpretation of our findings. Importantly, the deoxyhexose sugar and the above

sugar alcohol displayed only poor correlations with glucose and other biomarkers of T2D-risk ($r_s \leq |0.2|$) (Table 2). Moreover, adjustment for plasma glucose did not strongly affect the association of both metabolite groups with T2D, suggesting that they may not solely reflect hyperglycemia-induced metabolic shifts.

One purine nucleotide annotated as isopentenyl-adenosine-5′-monophosphate was positively associated with T2D. This metabolite is a precursor of cytokines in plants and bacteria *(41)*. Little is known about its role in humans, but the correlation with plasma glucose was $r_s = 0.44$, and its association with T2D lost statistical significance after adjustment for plasma glucose. Given the structural similarity with adenosine monophosphate, this metabolite might reflect the consequences of high glucose on nucleoside triphosphate diphosphohydrolase (NTPDase) activity and thus on breakdown of adenosine nucleotides *(42)*.

We also observed a yet-unexplored positive association between hydroxy-methylbutenyl disphosphate and T2D. Of note, an isomer of this metabolite, (*E*)-4-hydroxy-3-methyl-but-2-enyl pyrophosphate (HMBPP), is a precursor of the nonmevalonate pathway of isoprenoid biosynthesis in the majority of human pathogens *(43)*.

Table 2. Spearman rank-order correlation coefficients for the correlation of chemically annotated metabolites with a consistent association with T2D and selected baseline characteristics in nondiabetic participants of the EPIC-Potsdam cohort.[a]

Metabolite	Platform and ion mode	m/z	RT[b]	Age	BMI	SBP	Glc	Hb A$_{1c}$	ALT	GGT	HDL	TG	CRP
1-O-β-D-glucopyranosyl-D-mannitol||Maltitol	UPLC-MS (+)	367.122	50	-0.10	0.002	0.10	0.14	0.03	0.10	0.13	0.02	0.02	-0.04
Deoxyhexose sugar (e.g. anhydro-glucitol, deoxygalactose, deoxy-glucose)	UPLC-MS (-)	199.038	50	0.03	0.05	0.20	0.07	0.05	0.17	0.20	-0.11	0.09	-0.03
Tetrahydroxyhexanoicacid (e.g. 2-deoxygluconate, 3-deoxyarabinohexonic acid)|hexose sugar (e.g. glucose, fructose, inositol)	UPLC-MS (+)	271.041	47	0.04	0.08	0.20	0.46	0.04	0.13	-0.01	-0.17	0.04	0.01
Tetrahydroxyhexanoicacid (e.g. 2-deoxygluconate, 3-deoxyarabinohexonic acid)|hexose sugar (e.g. glucose, fructose, inositol)	UPLC-MS (+)	203.053	49	0.04	0.05	0.14	0.56	0.10	0.08	0.03	-0.05	0.02	0.04
Tetrahydroxyhexanoicacid (e.g. 2-deoxygluconate, 3-deoxyarabinohexonic acid)|hexose sugar (e.g. glucose, fructose, inositol)	UPLC-MS (+)	320.968	48	0.04	-0.01	0.14	0.58	0.03	0.11	0.11	-0.04	-0.07	0.01
Tetrahydroxyhexanoicacid (e.g. 2-deoxygluconate, 3-deoxyarabinohexonic acid)|hexose sugar (e.g. glucose, fructose, inositol)	UPLC-MS (+)	219.027	48	0.07	-0.09	-0.04	0.27	0.07	-0.01	-0.04	0.07	-0.07	-0.01
Tetrahydroxyhexanoicacid (e.g. 2-deoxygluconate, 3-deoxyarabinohexonic acid)|hexose sugar (e.g. glucose, fructose, inositol)	UPLC-MS (-)	214.896	47	0.11	0.08	0.14	0.64	0.12	0.19	0.10	-0.08	-0.01	0.04
Hydroxy-methylbutenyl diphosphate	UPLC-MS (+)	378.926	48	0.04	-0.03	0.14	0.52	0.09	-0.04	-0.06	0.06	-0.12	0.06
Isopentenyladenosine-5'-monophosphate	UPLC-MS (+)	416.136	243	0.03	0.05	0.13	0.44	0.09	0.09	0.03	-0.02	-0.02	0.07
PC(22:4/dm18:0)|PC(O-18:0/22:5)	UPLC-MS (+)	896.573	943	0.08	0.27	0.12	0.04	0.12	0.27	0.25	-0.17	0.31	0.21
PC(38:4)|PC(36:1)|PE(O-18:0/O-18:0)|PE(O-20:0/O-16:0)	UPLC-MS (+)	811.609	957	0.05	-0.05	0.03	0.01	0.07	0.12	0.11	0.03	0.06	-0.01
PC(O-20:0/O-20:0)	UPLC-MS (-)	860.684	988	-0.00	0.12	0.06	-0.02	0.02	-0.03	0.05	0.12	-0.10	0.13
GlcCer(d18:0/18:0)|PC(16:0/O-16:0)|PC(O-14:0/18:0)|PC(O-16:0/16:0)|PC(16:1/dm18:1)|PC(18:2/dm16:0)|PC(O-16:0/18:3)|PC(P-16:0/18:2)	UPLC-MS (-)	786.564	914	-0.04	0.09	0.08	0.04	0.02	-0.07	0.03	-0.03	0.03	-0.10
GlcCer(d18:0/22:0)|PC(O-14:0/22:0)|PC(O-16:0/20:0)|PC(O-18:0/18:0)|PC(O-20:0/16:0)|PC(20:1/dm18:1)|PC(20:2/dm18:0)|PC(22:2/dm16:0)	UPLC-MS (-)	842.628	951	-0.02	0.004	0.01	-0.03	-0.12	-0.23	-0.23	0.12	-0.36	-0.06
LysoPC(dm16:0)	UPLC-MS (-)	525.338	707	0.002	-0.08	-0.01	-0.07	-0.08	-0.15	-0.08	-0.03	-0.12	-0.24

[a] Data on age, BMI, and SBP are based on 300 controls of the 2 substudies. For the biomarkers, analyses were restricted to 265 controls with available data.
[b] RT, retention time; SBP, systolic BP; HDL, HDL cholesterol; TG, triglycerides; PE, phosphatidylethanolamine.

This metabolite is the most potent stimulant of human Vγ2Vδ2 T cells, thereby triggering immune response. Interestingly, certain bacteria also use HMBPP as an isopentenyl donor to produce substituted purine derivatives such as isopentenyladenosine-5′-monophosphate (44). To our knowledge, no such pathway exists in humans, but the correlation of 0.50 between hydroxymethylbutenyl disphosphate and isopentenyladenosine-5′-monophosphate in our data (see online Supplemental Table 3) points to a shared metabolic pathway or source.

The current work expands on a previous targeted profiling approach in EPIC-Potsdam, which profited from unambiguous chemical annotation and quantification of 127 serum metabolites in an even larger study population (17). However, 75% of the covered metabolites were choline-containing phospholipids, whose close structural and metabolic interrelationship is also reflected by strong correlation coefficients between metabolites. Findings of that study reinforce the utility of targeted approaches for in-depth analysis of specific metabolic pathways. Our untargeted approach, however, resulted in a larger coverage of metabolites and metabolite classes. Specifically, sugar alcohols, deoxyhexose sugars, and the ether glycerophospholipids PC(O-20/O-20) and Lyso-PC(dm16:0) were not covered by the previous targeted approach and were all inversely related to T2D in this study. The same is true for some metabolites that were positively associated with the end point, e.g., hydroxymethylbutenyl disphosphate and isopentenyladenosine-5′-monophosphate. To our knowledge, the above metabolites have not been linked to diabetes incidence before and point to an even broader range of metabolic alterations preceding T2D.

Some limitations of our study need to be acknowledged. Specifically, the UPLC-MS method applied was limited to specific classes of metabolites, including lipophilic metabolites, and was not expected to detect all metabolites present in human biofluids. The application of other analytical platforms (e.g., GC-MS, nuclear magnetic resonance) and other UPLC-MS methods (e.g., hydrophilic interaction chromatography) would be appropriate to define a more global definition of metabolism in the subjects studied. Furthermore, the applied UPLC-MS system did not distinguish between metabolites with the same accurate m/z and retention time, although it would discriminate between metabolites with the same m/z and a different retention time. Therefore, interpretation of our findings is challenging and mainly limited to those molecules with unambiguous chemical annotation. Thus, further research is necessary to derive unique metabolite identifications with high confidence. Given the exploratory nature, our findings may allow further targeted studies of a smaller number of biologically important metabolites or metabolite classes to be developed.

It is possible that a small number of unstable metabolites will degrade during sample collection, storage, and processing. However, all samples were collected according to a standardized protocol and stored at −196 °C (26). Because of the prospective design of this study, the future disease status had no influence on sample handling. Furthermore, plasma metabolites detected by UPLC-TOF-MS have been shown to be stable over a period of 13–17 years, even if stored at −80 °C (45). We thus believe that bias related to sample handling or storage is not the driving force behind our study findings.

Generally, untargeted metabolite profiling results in large data sets, requiring data mining approaches to reduce the dimensionality of the data. Indeed, we applied a strict data mining process. Because multivariate classification methods were used to select a subset of potentially relevant metabolites whose association to T2D was then confirmed in 2 substudies, we believe that the chance of false-positive findings is substantially reduced. However, because of dividing the data into 2 subpopulations, statistical power is limited. Therefore we cannot rule out the possibility that potentially relevant disease markers have been dismissed. The failure to replicate our recent finding of a positive association between branched-chain amino acids and T2D in EPIC-Potsdam may be related to the above issue (17), or may be caused by mass spectrometer signal saturation. Furthermore, as biological and technical variability affects the strength of metabolite–disease associations, metabolites with high reproducibility might have been preferentially selected.

False-positive findings are a common problem in association studies that use data derived from high-throughput omics technologies. We aimed to deal with this problem by randomly dividing our study population into 2 independent subcohorts. After a first feature selection process, we retained only metabolites that were significantly associated with T2D in multivariable-adjusted logistic regression models in both substudies (Fig. 1). Each substudy was a separate metabolomics study, and therefore each was independent of the other. If findings cannot be validated in external study populations, internal validation is an often-used alternative to check for consistency of results. Our internal validation approach is comparable to partitioning of the data into 1 subset used for statistical analyses (training set) and 1 complementary subset for testing (test set). As only consistent findings were considered, the training and test sets are exchangeable. Still, validation of findings in independent external cohorts would be desirable.

Our study profits from a large metabolite profiling approach to investigate metabolic alterations associated with T2D in a prospective study population, thereby reducing the possibility that metabolite concentrations change as a result of the outcome. Because of the detailed baseline phenotyping of our participants, we were able to

adjust for a wide range of T2D risk factors. However, only a small proportion of participants provided fasting blood samples. If systematically linked to the disease status, fasting time or type of food eaten before blood sampling may bias metabolite–disease associations. Although such a systematic error is unlikely owing to the prospective study design and careful matching for fasting periods, we cannot exclude the possibility that random variation in the most recent diet contributed to random variation in serum metabolites and thus to a higher proportion of false-negative findings. Consequently, a restriction or control of food intake before sampling is desirable in metabolomics studies.

Because we did not match for BMI, an important T2D risk factor, metabolites correlating with BMI might have been preferentially selected by the applied multivariate classification methods. Indeed, the selected 206 high-ranking metabolites exhibited a mean absolute correlation with BMI of 0.11, which ranged from 0 to 0.38 (data not shown). Although weak on average, these correlations were slightly stronger then the absolute correlations of BMI with metabolites not selected by the classification methods (mean 0.05; range 0–0.30). These findings do not question the validity of the ORs displayed in Figs. 2 and 3, because we adjusted for BMI in logistic regression models. However, we cannot rule out the possibility that potentially important metabolites might have been dismissed by the classification-based variable selection, owing to a higher ranking of metabolites correlating slightly more strongly with BMI.

In summary, the applied metabolomic platform detected diverse classes of metabolites, and we observed altered serum concentrations of glycerophospholipids, a purine nucleotide, and sugar metabolites including de-oxyhexose sugars and sugar alcohols that preceded the onset of overt T2D by a median of 6 years. Several of the candidates have not been linked to T2D incidence before, which underlines the potential of untargeted metabolomic approaches for biomarker discovery in prospective studies. Particularly, new hypotheses about T2D pathophysiology could be generated for future external validation.

References

1. Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. Lancet 2011;378: 31–40.

2. Dunn WB, Goodacre R, Neyses L, Mamas M. Integration of metabolomics in heart disease and diabetes research: current achievements and future outlook. Bioanalysis 2011;3:2205–22.

3. Fiehn O, Garvey WT, Newman JW, Lok KH, Hoppel CL, Adams SH. Plasma metabolomic profiles reflective of glucose homeostasis in non-diabetic and type 2 diabetic obese African-American women. PLoS One 2010; 5:e15234.

4. Li X, Xu Z, Lu X, Yang X, Yin P, Kong H, et al. Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry for metabonomics: biomarker discovery for diabetes mellitus. Anal Chim Acta 2009; 633:257–62.

5. Suhre K, Meisinger C, Doring A, Altmaier E, Belcredi P, Gieger C, et al. Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. PLoS One 2010;5:e13953.

6. Menni C, Fauman E, Erte I, Perry JR, Kastenmuller G, Shin SY, et al. Biomarkers for type 2 diabetes and impaired fasting glucose using a non-targeted metabolomics approach. Diabetes 2013;62:4270–6.

7. Xu F, Tavintharan S, Sum CF, Woon K, Lim SC, Ong CN. Metabolic signature shift in type 2 diabetes mellitus revealed by mass spectrometry-based metabolomics. J Clin Endocrinol Metab 2013;98:E1060–5.

8. Bao Y, Zhao T, Wang X, Qiu Y, Su M, Jia W, Jia W. Metabonomic variations in the drug-treated type 2 diabetes mellitus patients and healthy volunteers. J Proteome Res 2009;8:1623–30.

9. Ha CY, Kim JY, Paik JK, Kim OY, Paik YH, Lee EJ, Lee JH. The association of specific metabolites of lipid metabolism with markers of oxidative stress, inflammation and arterial stiffness in men with newly diagnosed type 2 diabetes. Clin Endocrinol (Oxf) 2012;76:674–82.

10. van Doorn M, Vogels J, Tas A, van Hoogdalem EJ, Burggraaf J, Cohen A, van der Greef J. Evaluation of metabolite profiles as biomarkers for the pharmacological effects of thiazolidinediones in type 2 diabetes mellitus patients and healthy volunteers. Br J Clin Pharmacol 2007;63:562–74.

11. Zhang J, Yan L, Chen W, Lin L, Song X, Yan X, et al. Metabonomics research of diabetic nephropathy and type 2 diabetes mellitus based on UPLC-OATOF-MS system. Anal Chim Acta 2009;650:16–22.

12. Wurtz P, Tiainen M, Makinen VP, Kangas AJ, Soininen P, Saltevo J, et al. Circulating metabolite predictors of glycemia in middle-aged men and women. Diabetes Care 2012;35:1749–56.

13. Yi LZ, He J, Liang YZ, Yuan DL, Chau FT. Plasma fatty acid metabolic profiling and biomarkers of type 2 diabetes mellitus based on GC/MS and PLS-LDA. FEBS Lett 2006; 580:6837–45.

14. Liu L, Wang M, Yang X, Bin M, Na L, Niu Y, et al. Fasting serum lipid and dehydroepiandrosterone sulfate as important metabolites for detecting isolated postchallenge diabetes: serum metabolomics via ultra-high-performance liquid chromatography/ mass spectrometry. Clin Chem 2013;59:1338–48.

15. Adams SH, Hoppel CL, Lok KH, Zhao L, Wong SW, Minkler PE, et al. Plasma acylcarnitine profiles suggest incomplete long-chain fatty acid beta-oxidation and al-

tered tricarboxylic acid cycle activity in type 2 diabetic African-American women. J Nutr 2009;139:1073–81.

16. Zhu C, Liang QL, Hu P, Wang YM, Luo GA. Phospholipidomic identification of potential plasma biomarkers associated with type 2 diabetes mellitus and diabetic nephropathy. Talanta 2011;85:1711–20.

17. Floegel A, Stefan N, Yu Z, Muhlenbruch K, Drogan D, Joost HG, et al. Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. Diabetes 2013;62:639–48.

18. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, et al. Metabolite profiles and the risk of developing diabetes. Nat Med 2011;17:448–53.

19. Wang-Sattler R, Yu Z, Herder C, Messias AC, Floegel A, He Y, et al. Novel biomarkers for pre-diabetes identified by metabolomics. Mol Sys Biol 2012;8:615.

20. Ferrannini E, Natali A, Camastra S, Nannipieri M, Mari A, Adam KP, et al. Early metabolic markers of the development of dysglycemia and type 2 diabetes and their physiological significance. Diabetes 2013;62:1730–7.

21. Rhee EP, Cheng S, Larson MG, Walford GA, Lewis GD, McCabe E, et al. Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. J Clin Invest 2011;121: 1402–11.

22. Wang TJ, Ngo D, Psychogios N, Dejam A, Larson MG, Vasan RS, et al. 2-Aminoadipic acid is a biomarker for diabetes risk. J Clin Invest 2013;123:4309–17.

23. Dunn WB, Broadhurst D, Begley P, Zelena E, FrancisMcIntyre S, Anderson N, et al. Procedures for largescale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc 2011;6:1060–83.

24. Boeing H, Korfmann A, Bergmann MM. Recruitment procedures of EPIC-Germany. European Investigation into Cancer and Nutrition. Ann Nutr Metab 1999;43: 205–15.

25. Bohlscheid-Thomas S, Hoting I, Boeing H, Wahrendorf J. Reproducibility and relative validity of food group intake in a food frequency questionnaire developed for the German part of the EPIC project. European Prospective Investigation into Cancer and Nutrition. Int J Epidemiol 1997;26:S59–70.

26. Kroke A, Bergmann MM, Lotze G, Jeckel A, KlipsteinGrobusch K, Boeing H. Measures of quality control in the German component of the EPIC study. European Prospective Investigation into Cancer and Nutrition. Ann Nutr Metab 1999;43:216–24.

27. Boeing H, Wahrendorf J, Becker N. Epic-Germany: a source for studies into diet and risk of chronic diseases. European Investigation into Cancer and Nutrition. Ann Nutr Metab 1999;43:195–204.

28. Zelena E, Dunn WB, Broadhurst D, Francis-McIntyre S, Carroll KM, Begley P, et al. Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. Anal Chem 2009; 81:1357–64.

29. Dunn WB, Wilson ID, Nicholls AW, Broadhurst D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. Bioanalysis 2012;4:2249–64.

30. Dunn WB, Broadhurst D, Brown M, Baker PN, Redman CW, Kenny LC, Kell DB. Metabolic profiling of serum using ultra performance liquid chromatography and the LTQ-Orbitrap mass spectrometry system. J Chromatogr B Analyt Technol Biomed Life Sci 2008;871: 288–98.

31. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 2006;78:779–87.

32. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis. Metabolomics 2007;3: 211–21.

33. Breiman L. Random forests. Machine Learning 2001; 45:5–32.

34. Barker M, Rayens W. Partial least squares for discrimination. J Chemometrics 2003:166–73.

35. Gall WE, Beebe K, Lawton KA, Adam KP, Mitchell MW, Nakhle PJ, et al. Alpha-hydroxybutyrate is an early biomarker of insulin resistance and glucose intolerance in a nondiabetic population. PLoS One 2010;5:e10883.

36. Cole LK, Vance JE, Vance DE. Phosphatidylcholine biosynthesis and lipoprotein metabolism. Biochim Biophys Acta 2012;1821:754–61.

37. Wallner S, Schmitz G. Plasmalogens: the neglected regulatory and scavenging lipid species. Chem Phys Lipids 2011;164:573–89.

38. Kahn BB. Type 2 diabetes: when insulin secretion fails to compensate for insulin resistance. Cell 1998;92: 593–6.

39. McGill JB, Cole TG, Nowatzke W, Houghton S, Ammirati EB, Gautille T, et al. Circulating 1,5-anhydroglucitol levels in adult patients with diabetes reflect longitudinal changes of glycemia: a U.S. trial of the Glycomark assay. Diabetes Care 2004;27:1859–65.

40. Beards E, Tuohy K, Gibson G. A human volunteer study to assess the impact of confectionery sweeteners on the gut microbiota composition. Br J Nutr 2010;104: 701–8.

41. Kakimoto T. Identification of plant cytokinin biosynthetic enzymes as dimethylallyl diphosphate:ATP/ADP isopentenyltransferases. Plant Cell Physiol 2001;42: 677–85.

42. Lunkes GI, Lunkes DS, Leal D, Araujo Mdo C, Correa M, Becker L, et al. Effect of high glucose levels in human platelet NTPDase and 5`-nucleotidase activities. Diabetes Res Clin Pract 2008;81:351–7.

43. Heuston S, Begley M, Gahan CG, Hill C. Isoprenoid biosynthesis in bacterial pathogens. Microbiology 2012; 158:1389–401.

44. Sakakibara H, Kasahara H, Ueda N, Kojima M, Takei K, Hishiyama S, et al. Agrobacterium tumefaciens increases cytokinin production in plastids by modifying the biosynthetic pathway in the host plant. Proc Natl Acad Sci U S A 2005;102:9972–7.

45. Hebels DG, Georgiadis P, Keun HC, Athersuch TJ, Vineis P, Vermeulen R, et al. Performance in omics analyses of blood samples in long-term storage: opportunities for the exploitation of existing biobanks in environmental health research. Environ Health Perspect 2013;121: 480–7.