Diabetologia (2018) 61:1315–1324
https://doi.org/10.1007/s00125-018-4599-x

ARTICLE

Metabolomics insights into early type 2 diabetes pathogenesis and detection in individuals with normal fasting glucose

Jordi Merino[1,2] · Aaron Leong[2,3] · Ching-Ti Liu[4] · Bianca Porneala[3] · Geoffrey A. Walford[1,2] · Marcin von Grotthuss[2] · Thomas J. Wang[5] · Jason Flannick[1,2] · Josée Dupuis[4,6] · Daniel Levy[6,7] · Robert E. Gerszten[8,9] · Jose C. Florez[1,2,10] · James B. Meigs[2,3,10]

Received: 24 October 2017 / Accepted: 26 February 2018 / Published online: 6 April 2018

Abstract

Aims/hypothesis Identifying the metabolite profile of individuals with normal fasting glucose (NFG [<5.55 mmol/l]) who progressed to type 2 diabetes may give novel insights into early type 2 diabetes disease interception and detection.

Methods We conducted a population-based prospective study among 1150 Framingham Heart Study Offspring cohort participants, age 40–65 years, with NFG. Plasma metabolites were profiled by LC-MS/MS. Penalised regression models were used to select measured metabolites for type 2 diabetes incidence classification (training dataset) and to internally validate the discriminatory capability of selected metabolites beyond conventional type 2 diabetes risk factors (testing dataset).

Results Over a follow-up period of 20 years, 95 individuals with NFG developed type 2 diabetes. Nineteen metabolites were selected repeatedly in the training dataset for type 2 diabetes incidence classification and were found to improve type 2 diabetes risk prediction beyond conventional type 2 diabetes risk factors (AUC was 0.81 for risk factors vs 0.90 for risk factors + metabolites, $p = 1.1 \times 10^{-4}$). Using pathway enrichment analysis, the nitrogen metabolism pathway, which includes three prioritised metabolites (glycine, taurine and phenylalanine), was significantly enriched for association with type 2 diabetes risk at the false discovery rate of 5% ($p = 0.047$). In adjusted Cox proportional hazard models, the type 2 diabetes risk per 1 SD increase in glycine, taurine and phenylalanine was 0.65 (95% CI 0.54, 0.78), 0.73 (95% CI 0.59, 0.9) and 1.35 (95% CI 1.11, 1.65), respectively. Mendelian randomisation demonstrated a similar relationship for type 2 diabetes risk per 1 SD genetically increased glycine (OR 0.89 [95% CI 0.8, 0.99]) and phenylalanine (OR 1.6 [95% CI 1.08, 2.4]).

Conclusions/interpretation In individuals with NFG, information from a discrete set of 19 metabolites improved prediction of type 2 diabetes beyond conventional risk factors. In addition, the nitrogen metabolism pathway and its components emerged as a potential effector of earliest stages of type 2 diabetes pathophysiology.

Keywords Metabolomics · Normoglycaemia · Type 2 diabetes pathophysiology · Type 2 diabetes prediction

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00125-018-4599-x) contains peer-reviewed but unedited supplementary material, which is available to authorised users.

✉ James B. Meigs
JMEIGS@mgh.harvard.edu

[1] Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

[2] Programs in Metabolism and Medical & Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[3] Division of General Internal Medicine, Massachusetts General Hospital, 100 Cambridge St, Boston, MA 02114, USA

[4] Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA

[5] Division of Cardiovascular Medicine, Vanderbilt University, Nashville, TN, USA

[6] The Framingham Heart Study, National Heart, Lung and Blood Institute, National Institutes of Health, Framingham, MA, USA

[7] The Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD, USA

[8] Division of Cardiovascular Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

[9] Broad Institute of MIT and Harvard Program in Metabolism, Cambridge, MA, USA

[10] Department of Medicine, Harvard Medical School, Boston, MA, USA

Research in context

What is already known about this subject?

- Plasma metabolite abnormalities are present several years before the development of type 2 diabetes
- Prior studies have identified metabolites that are associated with the development of type 2 diabetes among initially non-diabetic individuals
- Identified metabolites do not substantively improve type 2 diabetes risk prediction beyond clinical risk factors in general samples that include a mix of normoglycaemic and dysglycaemic individuals

What is the key question?

- Do specific metabolites predict type 2 diabetes and improve the prediction of the disease beyond conventional clinical risk factors among people with normoglycaemia?

What are the new findings?

- In individuals with normal fasting glucose who progressed to type 2 diabetes, we identified 19 metabolites associated with type 2 diabetes incidence
- Information from these 19 metabolites improved the predictive capability of incident type 2 diabetes beyond conventional clinical risk factors
- These metabolites identify alterations in nitrogen metabolism pathways that may contribute to the earliest stages of type 2 diabetes pathogenesis

How might this impact on clinical practice in the foreseeable future?

- If these findings are confirmed in other populations, we would have confidence that measurement of a defined set of metabolites could help identify individuals at risk for future type 2 diabetes, even when they appear clinically at low risk

Abbreviations

FHS	Framingham Heart Study
HDLc	HDL-cholesterol
IFG	Impaired fasting glucose
IGT	Impaired glucose tolerance
IVW	Inverse-variance weighted
LASSO	Least absolute shrinkage and selection operator
LDLc	LDL-cholesterol
NFG	Normal fasting glucose
ROC	Receiver operator characteristic
TAG	Triacylglycerol

Introduction

Type 2 diabetes is epidemic, affecting the health of millions of people worldwide. The number of years that people with type 2 diabetes are living has increased by 32% over the last few decades due to the rise in age-specific prevalence and population growth and ageing [1]. Consequently, type 2 diabetes ranks sixth among leading causes of the burden of disease globally [2]. Previous studies have shown that type 2 diabetes incidence can be prevented or delayed [3, 4] and that people at risk for developing type 2 diabetes can be identified through measuring common clinical risk factors [5].

Higher fasting plasma glucose levels, even in the non-diabetic range, can predict future type 2 diabetes. Individuals with impaired fasting glucose (IFG [5.6–7.0 mmol/l]) have an annual relative risk of type 2 diabetes of 4.7% (95% CI 2.5, 6.9) compared with normoglycaemic individuals [6]. Impaired glucose tolerance (IGT) and elevated HbA_{1c} (39–46 mmol/mol [5.7–6.4%]) are also associated with an increased risk of type 2 diabetes compared with those with completely normal glycaemia [7, 8]. However, around 5–10% of middle-aged individuals of European descent with normal fasting glucose (NFG [<5.55 mmol/l]) develop type 2 diabetes over a 5–10 year period [6–8]. The incidence rate is much higher in middle-aged individuals from other ethnic backgrounds; Asian Indians have one of the highest incidence rates of diabetes, with rapid conversion from normoglycaemia to dysglycaemia (19.4% over a 9 year follow-up period) [9].

Prior studies have identified plasma metabolites that are associated with the development of future type 2 diabetes in individuals with both normoglycaemia and prevalent dysglycaemia (IFG and/or IGT) [10–18]. Alterations in these metabolites likely signal changes in relevant biological pathways, including amino acid catabolism [10, 11, 13–18], lipid oxidation [12, 13, 15, 17] and hexose metabolism [15, 17]. However, in terms of

population-level prediction of future type 2 diabetes, identified metabolites have little value beyond clinical risk factors, such as fasting glucose, one of the most robust predictors of future type 2 diabetes [16]. Further, because previous studies were conducted in a mixture of individuals with normoglycaemia and prevalent dysglycaemia, they were unable to discern whether early dysglycaemia preceded changes in metabolite levels or whether identified metabolites were harbingers of early dysglycaemia. We therefore tested the hypothesis that a metabolomics analysis in people with NFG who developed type 2 diabetes could identify new markers and pathways that elucidate early type 2 diabetes pathogenesis and improve prediction of incident type 2 diabetes beyond clinical risk factors.

Methods

Study participants

We included participants from the Framingham Heart Study (FHS) Offspring cohort, a prospective, observational, community-based cohort including 3799 attendees, age 40–65 years, at the fifth quadrennial examination cycle 1991–1995 (baseline examination) [19]. Participants at the fifth and subsequent quadrennial examination cycles underwent a physician-administered physical examination and medical history and routine laboratory tests. For the current analyses, we excluded individuals without profiling of metabolites (n = 1326) and those with prevalent diabetes or cardiovascular events (n = 346), fasting plasma glucose ≥5.6 mmol/l (n = 967) or 2 h glucose ≥11 mmol/l (n = 10). The final study population included 1150 individuals with NFG and no diabetes. All participants provided written informed consent and the study protocol was approved by the Boston University Medical Center Institutional Review Board.

Metabolite profiling

At baseline, after participants had fasted overnight, plasma samples were collected in EDTA, processed immediately and stored at −80°C until assayed. Plasma samples were collected at the fifth quadrennial examination, which took place between 1991 and 1995, and were processed in 2008. A previous study has documented concordance in several metabolite measures between archived samples from the Framingham Offspring Study and freshly obtained samples [20]. Targeted metabolite profiling was performed using liquid chromatography with tandem mass spectrometry (LC-MS/MS) as previously described [11, 12]. Additional details, including accuracy of the methodology used in analyses, calibration and annotation, are provided in the ESM Methods. Metabolites at high missing rate (>20%) were excluded from this analysis, which includes 220 metabolites.

Ascertainment of incident type 2 diabetes

The primary endpoint of this study was incident type 2 diabetes. Incident type 2 diabetes was ascertained during the follow-up at every quadrennial examination and was defined as follows: fasting glucose ≥7 mmol/l, non-fasting blood glucose ≥11 mmol/l or the use of glucose-lowering medications, including insulin. Time to type 2 diabetes incidence was derived from the time of the baseline examination. Chart review was conducted to identify and exclude two participants with type 1 diabetes mellitus.

Clinical covariates

Demographic, lifestyle and clinical characteristics were assessed at baseline. BMI was calculated as weight divided by height squared (kg/m^2). The HOMA-IR was calculated [21] and was log-transformed due to a skewed distribution. Total cholesterol, HDL-cholesterol (HDLc) and triacylglycerols (TAGs) were measured, in individuals who had fasted overnight, using standard methods. LDL-cholesterol (LDLc) was indirectly calculated using the Friedewald formula when TAG concentrations were lower than 4.52 mmol/l [22]. We used conventional type 2 diabetes risk factors to estimate risk of new onset of type 2 diabetes for each participant, including sex and parental history of diabetes as categorical variables and age, fasting glucose, BMI, HDLc, TAG and blood pressure as continuous variables. We also considered HOMA-IR and 2 h glucose as continuous variables.

Statistical analysis

Differences in clinical characteristics between participants with and without incident type 2 diabetes were analysed in generalised estimating equations models accounting for familial correlation among participants.

The analytical plan flow-chart for metabolite selection, prediction performance and complementary analyses is summarised in Fig. 1. First, plasma metabolite concentrations were log-transformed and standardised. Next, a random binomial variable was used to split the sample into a testing dataset and a training dataset (4:6) and to serve as an internal validation and avoid inflation of the discrimination estimates. For the retained training dataset (60% of the sample), we conducted least absolute shrinkage and selection operator-penalised regressions (LASSO) with tenfold cross validation to select metabolites predictive of type 2 diabetes incidence based on the criteria giving minimum mean cross-validated error [23]. We then assessed the predictive capability of type 2 diabetes risk factors alone (including age, sex, parental history of diabetes, fasting glucose, BMI, HDLc, TAG and blood pressure) and the predictive capability of type 2 diabetes risk

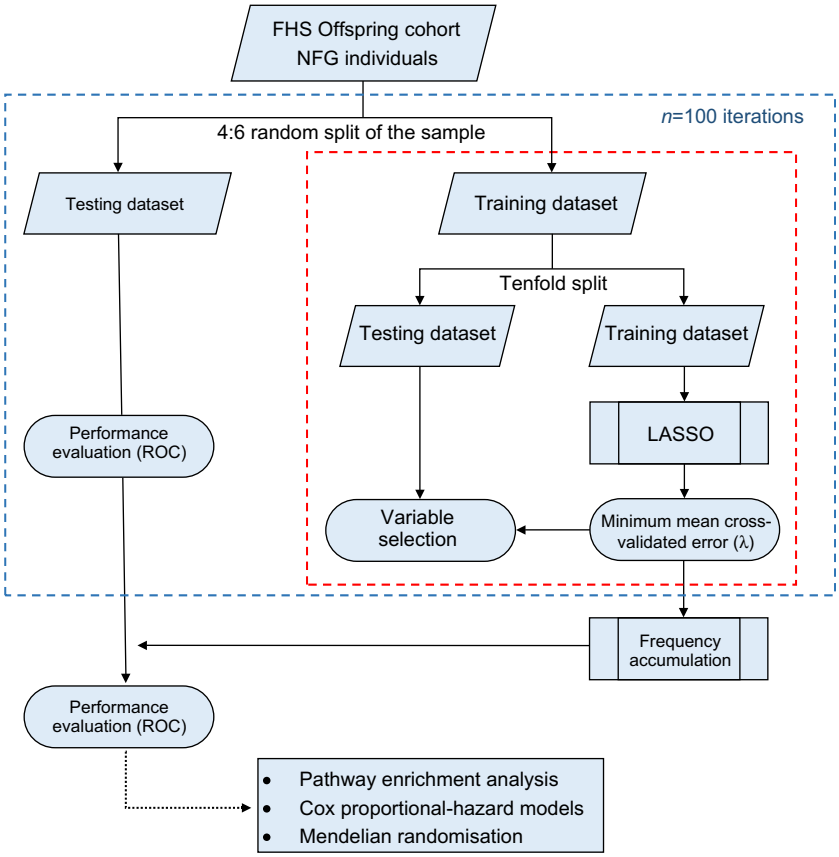

Fig. 1 Flow-chart summarising the main steps in the analytical plan. We first split the sample into a testing dataset and a training dataset (at a 4:6 ratio). For the retained training dataset (60% of the sample, red dashed line), we conducted LASSO regression with tenfold cross validation [22] to select metabolites predictive of type 2 diabetes incidence based on the criteria giving minimum mean cross-validated error. Next, we assessed the predictive capability of type 2 diabetes risk factors alone and the predictive capability of type 2 diabetes risk factors plus selected metabolites in the testing set (40% of the sample, blue dashed line) by generating the AUC of the ROC curve in each of the 100 iterations. Next, we evaluated the capability of the final set of metabolites selected ten or more times in 100 iterations to improve prediction of type 2 diabetes when added to a model that included traditional type 2 diabetes risk factors, now using the entire sample. Although we did not have an entirely independent data sample as a testing set, the 95% CIs around the AUC from 100 permutations in the 40% testing set gives the lower bounds that might be expected in testing in an independent sample. Testing predictive improvement in the entire sample also permitted better power for use of several complimentary approaches. These included multivariable-adjusted Cox proportional hazard models to evaluate model robustness, and pathway enrichment analysis and Mendelian randomisation to investigate biological insights and test causality in the metabolite–type 2 diabetes risk association

factors plus selected metabolites in the testing set (40% of the sample) by generating the area under the receiver operator characteristic (ROC) curve. We used a nonparametric approach (DeLong's test) to compare the discriminatory capability of the two highly correlated ROC curves [24]. We repeated this process 100 times and accumulated the selection frequency across 100 iterations for each metabolite separately and used a cut-off of ten selections to prioritise the top predictors of incident type 2 diabetes. Next, we evaluated the capability of the metabolites selected ten or more times in 100 iterations to improve prediction of type 2 diabetes over conventional type 2 diabetes risk factors in the entire cohort. As a sensitivity analysis, we included HOMA-IR and 2 h glucose in the model for type 2 diabetes risk factors and repeated the same methodological approach. These analyses were performed using glmnet (https://cran.r-project.org/web/packages/glmnet/index.html) and pROC (https://cran.r-project.org/web/packages/pROC/index.html) packages implemented in R v3.2.0 program (https://www.r-project.org/). We took two-sided $p < 0.05$ to denote evidence against the null hypothesis of no type 2 diabetes risk prediction improvement when adding metabolites to the prediction model.

Finally, Cox proportional hazard models were used to investigate the association between prioritised metabolites and type 2 diabetes risk after adjusting for age, sex, BMI, fasting glucose and fasting TAG at baseline. SAS v9.3 (SAS Institute, Cary, NC, USA) was used for the association analyses. We took Bonferroni-corrected threshold for significance at two-

sided $p < 2.63 \times 10^{-3}$ (0.05/19 metabolites) to denote evidence against the null hypothesis of no association between prioritised metabolites and type 2 diabetes risk.

Bioinformatics methods

Pathway analysis We applied pathway enrichment analysis and metabolite set enrichment analysis to identify enriched metabolic pathways using MetaboAnalyst 3.0 [25] for the set of 19 prioritised metabolites. Pathway enrichment analysis at the false discovery rate of 5% was set for significance.

Mendelian randomisation Mendelian randomisation was conducted for causal inference analyses between components of the nitrogen metabolism pathway and type 2 diabetes risk. Genetic determinants of plasma metabolites were extracted from the MAGNETIC Consortium (n = 24,925) [26]. In the MAGNETIC Consortium, we identified genetic variants associated with glycine and phenylalanine at genome-wide significance ($p < 5 \times 10^{-8}$) (taurine was not available, but all metabolite meta-analysis results are available through www.computationalmedicine.fi/data/NMR_GWAS/), For each independent variant, we gathered summary-level association results for type 2 diabetes from the GoT2D diabetes dataset (www.type2diabetesgenetics.org/projects/got2d; n = 11,645 cases and 32,769 controls) since these variants were not available in other type 2 diabetes genetics consortia [27]. The Mendelian randomisation overall instrumental estimated effect size of the exposure on the outcome, referred to as the inverse-variance weighted (IVW) estimator [28], was calculated using the Genetics ToolboX package (GTX; available at http://cran.r-project.org/web/packages/gtx) (detailed in the ESM Methods). Instrumental heterogeneity was assessed using the Q statistic and reported as a heterogeneity p value. The presence of unbalanced horizontal pleiotropy was assessed by using Mendelian randomisation–Egger when the set of variants in the genetic instrument allowed us to conduct the analysis [29]. We used individual-level data from FHS participants to estimate the variance explained in metabolite levels by the genetic variants. We used genotyped variants with genotyping success rate ≥0.95 and variants in Hardy–Weinberg equilibrium ($p > 1 \times 10^{-4}$). When not directly genotyped, we included variants at high-quality imputation ratio (r^2 value threshold of 0.85, representing an approximate correlation with the true genotype higher than 0.9). A linear mixed-effect model with covariates age, sex and random effects to account for familial correlation, including five variants for glycine and three variants for phenylalanine fit individually in an additive genetic model, was used to estimate the variance in plasma metabolite concentrations explained by genetic variants.

Results

Over a 20 year follow-up period, 95 individuals with NFG (8.3%) developed type 2 diabetes. Baseline characteristics according to type 2 diabetes incidence are presented in Table 1. Individuals who developed type 2 diabetes did not differ in age and parental history of type 2 diabetes distribution from those who did not develop type 2 diabetes, but diabetes incidence was higher in men and in individuals who had significantly higher BMI and slightly higher glycaemic trait measurements (fasting and 2 h glucose) at metabolomics sampling. Still, individuals with NFG who progressed to type 2 diabetes were normoglycaemic at baseline, as indicated by 2 h glucose and HbA_{1c} values being in the normal range.

Overall, 67 metabolites were selected at least once in the training set for type 2 diabetes incidence classification (ESM Table 1). Among them, two metabolites, sphingomyelin C24:0 and diacylglycerol C36:1, were selected in every one of the 100 iterations. The median change in the AUC in the internal validation set upon adding the metabolites selected within each of the 100 iterations to a model that included traditional type 2 diabetes risk factors alone was 0.088 (p = 0.013). Nineteen metabolites were prioritised by LASSO ten or more times in the training dataset (Table 2). The subset of 19 metabolites significantly improved type 2 diabetes prediction when added to a model that included traditional type 2 diabetes risk factors alone using the entire sample (AUC was 0.810 [95% CI 0.77, 0.86] for type 2 diabetes risk factors and 0.902 [95% CI 0.87, 0.94] for type 2 diabetes risk factors + metabolites, $p = 1.1 \times 10^{-4}$) (Fig. 2). In a sensitivity analysis including HOMA-IR and 2 h glucose as additional type 2 diabetes risk factors, metabolites still significantly improved type 2 diabetes prediction (AUC was 0.828 [95% CI 0.78, 0.88] for type 2 diabetes risk factors and 0.906 [95% CI 0.87, 0.94] for type 2 diabetes risk factors + metabolites, $p = 2 \times 10^{-4}$) (ESM Fig. 1).

Next, we used the set of 19 metabolites to identify enriched metabolic pathways. A significant enrichment for association was observed for the nitrogen metabolism pathway at the false discovery rate of 5% (p = 0.047) (Table 3). This pathway is composed of 39 species, three of which (glycine, taurine and phenylalanine) were prioritised by LASSO in the training set ten or more times (ESM Table 2). In separate Cox proportional hazard models for metabolites in the nitrogen metabolism pathway, type 2 diabetes risk was lower per 1 SD increase in plasma glycine (HR 0.65 [95% CI 0.54, 0.78]) and taurine (HR 0.73 [95% CI 0.59, 0.90]) and higher for 1 SD increase in phenylalanine (HR 1.35 [95% CI 1.11, 1.65]) after adjusting for confounders (Table 3). The associations between other prioritised plasma metabolites or conventional risk factors and type 2 diabetes risk is detailed in ESM Tables 3 and 4.

Finally, we investigated whether genetically increased metabolites in the nitrogen metabolism pathway have a causal

Table 1 Baseline characteristics of participants

Characteristic	All ($n = 1150$)	Incident diabetes ($n = 95$)	No diabetes ($n = 1055$)	p value[a]
Age, years	53 ± 10	54 ± 9	53 ± 10	0.44
Female sex, n (%)	679 (59)	43 (45.3)	636 (60.3)	0.004
Parental history of T2D, n (%)	191 (16.6)	20 (21)	171 (16.3)	0.22
BMI, kg/m^2	26.46 ± 4.46	29.66 ± 5.18	26.18 ± 4.28	<0.001
Glycaemic traits				
Fasting glucose, mmol/l	4.99 ± 0.35	5.19 ± 0.26	4.97 ± 0.35	<0.001
HbA_{1c}, mmol/mol (%)	33 ± 4 (5.2 ± 0.6)	34 ± 4 (5.3 ± 0.6)	34 ± 4 (5.2 ± 0.6)	0.042
2 h Glucose, mmol/l	5.51 ± 1.38	6.53 ± 1.72	5.42 ± 1.31	<0.001
HOMA-IR[b]	6.20 ± 2.24	8.08 ± 2.71	6.03 ± 2.12	<0.001
Lipids				
HDLc, mmol/l	1.35 ± 0.4	1.08 ± 0.29	1.37 ± 0.4	<0.001
LDLc, mmol/l	3.26 ± 0.88	3.37 ± 0.89	3.25 ± 0.88	0.16
TAG, mmol/l	1.53 ± 0.18	2.10 ± 1.06	1.47 ± 1.19	<0.001

Values are given as the mean±SD, except for qualitative variables which are expressed as n (%). During a median follow-up of 20 years, 95 (8.3%) individuals with normoglycaemia at baseline developed type 2 diabetes

[a] p values were obtained by generalised estimating equation models

T2D, type 2 diabetes

role in type 2 diabetes risk (ESM Table 5). Using the IVW estimator method, we found that for every 1 SD genetically increased glycine, the odds of type 2 diabetes was reduced by 11% (OR 0.89 [95% CI 0.80, 0.99]; heterogeneity $p = 0.08$, Fig. 3a). The genetic variance in glycine metabolite levels was 11.1% in FHS. The genetic variance attributed to *CPS1* was 10% and the allele associated with higher glycine concentrations is also associated with lower risk of type 2 diabetes. The adjusted causal effect estimate was similar when applying the bootstrap method in Mendelian randomisation–Egger regression, showing a trend towards statistical significance (OR 0.87 [95% CI 0.74, 1.00], $p = 0.074$) (ESM Table 6). The estimate for the intercept in the Mendelian randomisation–Egger regression suggested no evidence of presence of unbalanced pleiotropy ($\beta_{intercept} = 0.01$; SE = 0.02; $p = 0.118$). In a Mendelian randomisation analysis for phenylalanine, which included three phenylalanine risk-increasing variants (variance explained = 16.5% in FHS), the estimate using the IVW estimator method was 1.6 type 2 diabetes higher odds per 1 SD genetically increased phenylalanine (95% CI 1.08, 2.04; heterogeneity $p = 0.19$) (Fig. 3b). We did not conduct Mendelian randomisation–Egger regression for phenylalanine given the low number of variants in this analysis.

Table 2 Metabolites prioritised ≥10 times in the training set for type 2 diabetes incidence differentiation

Name	HMDB-ID	Selected times (out of 100 runs)
SM C24:0	HMDB11697	100
DAG C36:1	HMDB07216	100
TAG C58:11	HMDB10531	89
5-Hydroxyindoleacetic acid	HMDB00763	86
PC C36:4	HMDB07983	85
3-Methyladipic acid	HMDB00555	81
D-Glucose	HMDB00122	76
2-Aminodipate	HMDB00510	71
Isocitrate	HMDB00193	64
L-Phenylalanine	HMDB00159	49
LPC C18:2	HMDB10386	47
Glycine	HMDB00123	46
TAG C52:1	HMDB05367	39
LPC C18:1	HMDB02815	22
TAG C48:1	HMDB05359	21
TAG C48:0	HMDB05356	21
CE C20:3	HMDB06736	13
Taurine	HMDB00251	11
TAG C54:8	HMDB10518	11

Nineteen metabolites were selected ten or more times out of 100 iterations in the training set with tenfold cross validation at minimum mean cross-validated error

CE, cholesteryl ester; DAG, diacylglycerol; HMDB-ID, Human Metabolome Database identification; LPC, lysophosphatidylcholine; SM, sphingomyelin; PC, phosphatidylcholine

Discussion

We conducted a population-based prospective study in individuals with NFG at baseline, of whom 95 progressed to type 2

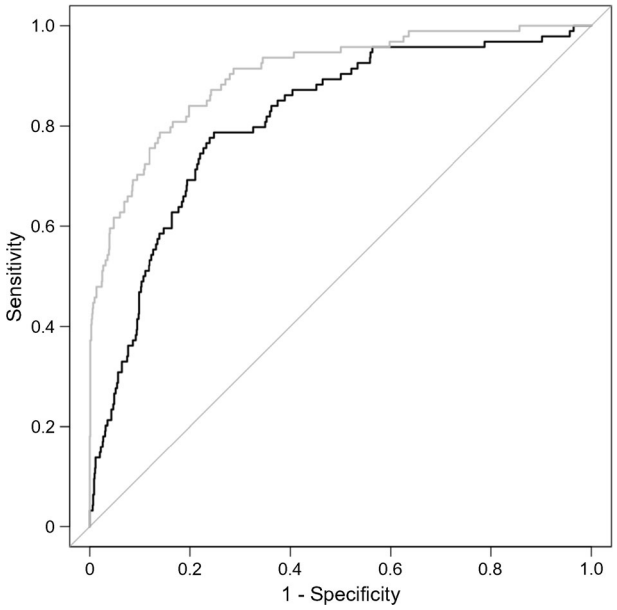

Fig. 2 ROC curves for models predicting incident type 2 diabetes with and without selected metabolites. ROC curves were derived from a Cox proportional hazard model in the testing dataset. Graphs plot the sensitivity vs (1 – specificity) for diabetes at each possible model cut point. The area under the ROC curve corresponds to the *C* statistic of that model. Black line, type 2 diabetes risk factors, including age, sex, parental history of diabetes, fasting glucose, BMI, HDLc, TAG and blood pressure; grey line, type 2 diabetes risk factors + LASSO selected metabolites (selected ten or more times in 100 iterations with tenfold cross validation at minimum mean cross-validated error). AUC was 0.81 (95% CI 0.77, 0.86) for the type 2 diabetes risk score and 0.902 (95% CI 0.87, 0.94) for the type 2 diabetes risk score + metabolites ($p = 1.1 \times 10^{-4}$)

diabetes during the follow-up period of 20 years. Using information from a discrete set of 19 metabolites associated with type 2 diabetes incidence, we improved the capability of predicting

Table 3 Nitrogen metabolism pathway metabolite associations with type 2 diabetes risk

Metabolite	Type 2 diabetes risk	
	HR (95% CI)	*p* value
Glycine	0.65 (0.54, 0.78)	2.84×10^{-6}
Taurine	0.73 (0.59, 0.90)	0.003
Phenylalanine	1.35 (1.11, 1.65)	0.002

Data show HR and 95% CI per SD unit of $\log_2$-transformed standardised metabolite increase on type 2 diabetes risk

The nitrogen metabolism pathway was enriched for association at the false discovery rate of 5% ($p = 0.047$). The nitrogen metabolism pathway is composed of 39 metabolite species (detailed in ESM Table 2) and three (glycine, taurine and phenylalanine) were prioritised ten or more times in the training set for type 2 diabetes incidence differentiation. Next, we tested the association between glycine, taurine, phenylalanine and type 2 diabetes risk. Each metabolite represents a separate Cox proportional hazard model adjusted for age, sex, BMI, fasting glucose and TAGs at baseline

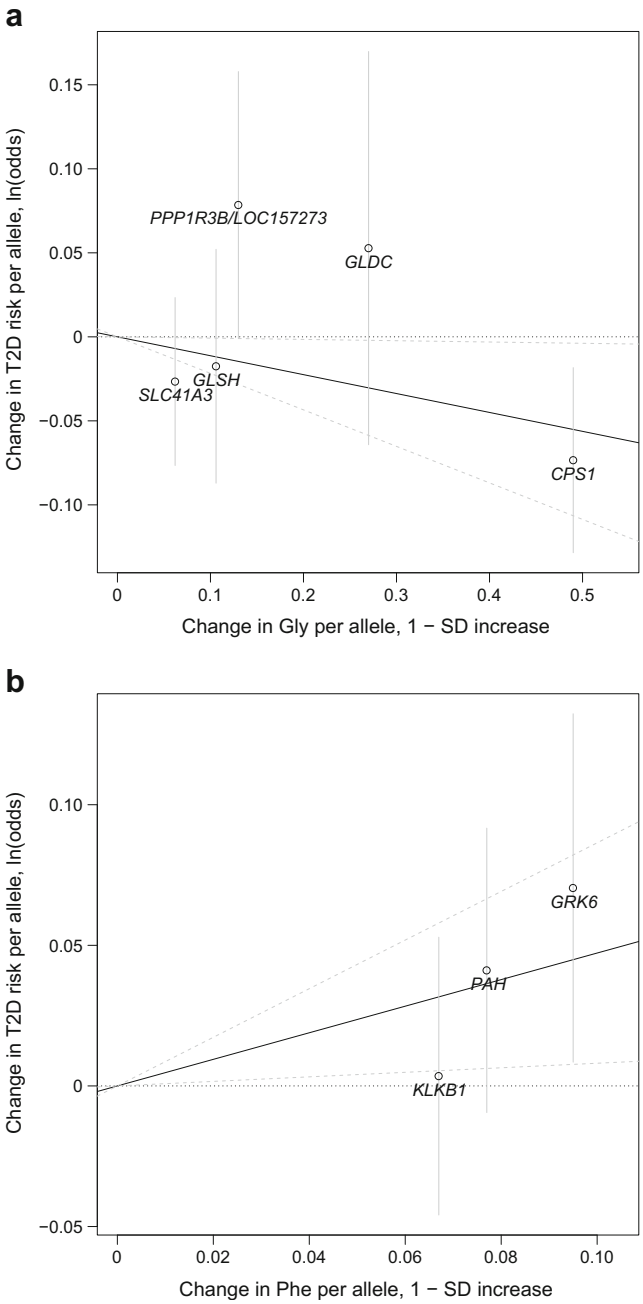

Fig. 3 Mendelian randomisation estimate of raised glycine and phenylalanine on type 2 diabetes. Effect of glycine- and phenylalanine-raising genetic variants on type 2 diabetes odds obtained from publicly available data from MAGNETIC Consortium ($n = 24{,}925$) and GoT2D ($n = 44{,}414$). Each white circle represents a known locus for glycine (**a**) or phenylalanine (**b**). The association of each variant with type 2 diabetes [base *e* logarithmic OR transformation, ln(odds)] is denoted on the *y*-axis (white circle and grey error bars around the white circles) while the association with glycine and phenylalanine is denoted on the *x*-axis. The black lines illustrate regression of glycine and phenylalanine on type 2 diabetes and the dashed grey lines show the 95% CI. The effect of the polygenic instrument comprising all five glycine-increasing variants reduced the odds of type 2 diabetes risk per 1 SD increase in metabolite concentration (OR 0.89 [95% CI 0.80, 0.99]). Genetically driven phenylalanine increased odds of type 2 diabetes risk per 1 SD increase in metabolite concentration (OR 1.6 [95% CI 1.08, 2.37]). T2D, type 2 diabetes

incident type 2 diabetes beyond the predictions made using only the clinical risk factors obtained in routine care. A significant biological finding is the enrichment in the nitrogen metabolism pathway with type 2 diabetes risk. Further, the genetic approach provides additional evidence that markers identified in the nitrogen metabolism pathway—glycine and phenylalanine—may be causal rather than only associative, suggesting that alterations in this pathway and its components may contribute to the earliest stages of type 2 diabetes pathogenesis.

While prior clinical metabolomics studies have focused on a mixture of individuals with normoglycaemia and prevalent dysglycaemia [10–18], our work is novel in its study of participants who were normoglycaemic yet progressed to type 2 diabetes. By studying individuals who progressed from having normal glucose metabolism to having type 2 diabetes, we eliminate confounding of our results by processes that occur in response to development of dysmetabolism (IFG or IGT). The main clinical finding of the present study is that selected metabolites substantially improved the ability to predict type 2 diabetes, beyond the prediction achieved using conventional risk factors, in individuals classified as normoglycaemic based on fasting glucose. Our findings are slightly different from those of previous metabolomics studies, which were not able to show that metabolites materially improved type 2 diabetes risk prediction over other clinical risk factors [11, 15, 16]. A possible explanation might be related to the study population: in normoglycaemic individuals, blood glucose and other traditional type 2 diabetes risk factors may not be as strong predictors of future type 2 diabetes as in those with dysglycaemia [30]. Thus, novel risk factors like metabolites could show stronger predictive capability. Nevertheless, the AUC of clinical risk factors in normoglycaemic individuals was still >80%, slightly higher than reported in other recent studies [31, 32]. This suggests that differences in the length of follow-up or the inclusion of different ethnic groups could affect the predictive capability of traditional type 2 diabetes risk factors. Another possible explanation for the increased predictive ability of metabolites might be related to the methodological approach implemented in this study. The predictive capability observed in this study is aligned with prediction performance observed in recent studies using similar machine learning approaches to prioritise metabolites [31, 32].

The normoglycaemic individuals included in this study were selected not only because of their normal fasting glucose but also because of their normal glucose tolerance and HbA_{1c}, under current accepted definitions. Although different, perhaps more stringent, thresholds for 'normal' might have been chosen (especially as there were subtle elevations in fasting glucose among those who developed type 2 diabetes vs those who did not), we think that the data provide insight into early type 2 diabetes pathogenesis among those currently considered clinically normoglycaemic. The convergence of selected metabolites in the nitrogen metabolism pathway, serving as nitrogen donors for the urea cycle [33], suggests that this pathway may influence the early pathogenesis of type 2 diabetes. Data from the Mendelian randomisation experiments further support the notion that components within this pathway may have a causal role in type 2 diabetes development; therefore, confounding effects by obesity or lipid abnormalities are less likely. In our study, genetically increased glycine reduced the odds of type 2 diabetes, consistent with findings from previous epidemiological studies in terms of directionality and effect sizes [15, 34]. However, this conflicts with a previous Mendelian randomisation analysis that showed no association between a single genetic variant for glycine or glycine-to-serine ratio and diabetes-related traits [35]. More precisely estimated effect sizes derived using data from the largest metabolites meta-genome-wide association studies currently available and the increase in the number of genetic variants used to proxy glycine here are likely to explain the difference between these two Mendelian randomisation studies. With regard to phenylalanine, different studies have reported a direct association between this metabolite and type 2 diabetes risk [11, 15, 34], although no Mendelian randomisation analysis for phenylalanine on type 2 diabetes risk has been conducted yet.

In the present study, we also documented that particular biomarkers previously associated with type 2 diabetes risk (e.g. TAGs with lower carbon number or 2-aminoadipate [12, 14]) are likely to be relevant even before initial glycaemic perturbations. In contrast, other metabolites such as branched chain amino acids, associated with type 2 diabetes incidence in previous studies, were not prioritised by our selection algorithm. Notably, our findings, which highlight early metabolite changes in the pathogenesis of type 2 diabetes, are consistent with a recent Mendelian randomisation analysis finding that elevations in branched chain amino acids occur after the development of insulin resistance [36].

We acknowledge that the results of our population-based analysis should be interpreted with caution since several limitations such as unmeasured factors (e.g. changes in lifestyle factors, medications or insulin secretion and resistance over time) might have influenced our findings. Using a Mendelian randomisation approach for available metabolites in the nitrogen metabolism pathway partially mitigates this concern, suggesting that genetically driven glycine and phenylalanine are indeed related to the risk of type 2 diabetes independently of potential confounders. While our results were internally validated, we recognise that they were not confirmed in a separate prospective cohort. The lack of independent validation is due to the lack of availability of comparable cohorts of normoglycaemic individuals who developed type 2 diabetes for whom the necessary data were available. However, the internal validation approach, using 40% of the sample and running 100 iterations, allowed us to rule out other potential conflicting issues, such as compatibility between

metabolomics platforms or available standards in libraries, even when similar platforms were used. Five TAGs were prioritised by our methodological approach but we did not find significant enrichment of pathways associated with these species. This might be because the software we used for pathway enrichment analysis may provide poor reporting for lipid classes and pathways or because only five TAGs in a particular metabolic pathway are likely to be less than expected by chance for enrichment. In addition, most prioritised metabolites correlate with baseline clinical risk factors such as BMI, 2 h glucose, HOMA-IR, HDLc and TAGs (ESM Table 7) but associations of metabolites with type 2 diabetes in normoglycaemic individuals remained after risk factor adjustment. Last, we recognise that participants in this study were all of European descent. Further work is needed to determine whether our findings can be replicated in an independent cohort of the same ethnicity and to extend the study to other racial/ethnic groups.

In conclusion, our study identifies a discrete set of metabolites that signal increased risk for type 2 diabetes among normoglycaemic individuals; these metabolites are involved and may play a causal role in the early stages of type 2 diabetes pathogenesis.

Acknowledgements This research was conducted in part using data and resources from the FHS of the National Heart Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine. The analyses reflect intellectual input and resource development from the FHS investigators participating in the SNP Health Association Resource (SHARe) project. The authors wish to thank the GoT2D Consortium for access to their data.

Data availability Metabolomics data that support the findings of this study have been deposited in dbGaP with the study accession number phs000007.v29.p10 and dataset phenotypic identifiers 'pht002234.v5.p10:' (Metabolomics – HILIC), 'pht002894.v1.p10:' (Central Metabolomics – HILIC), 'pht002343.v4.p10:' (Metabolomics - Lipid Platform).

Funding This work was partially supported by the National Heart, Lung and Blood Institute's FHS (contract no. N01-HC-25195 and HHSN268201500001I) and its contract with Affymetrix, Inc. for genotyping (contract no. N02-HL-6-4278) and metabolomic services (R01-HL081572) and supported by U01 DK078616 and NIDDK K24DK080140 (JBM). JM was supported by a postdoctoral fellowship funded by the European Commission Horizon 2020 program and Marie Skłodowska-Curie actions (H2020-MSCA-IF- 2015-703787). JCF is a Massachusetts General Hospital Research Scholar and is supported by NIDDK K24 DK110550.

Duality of interest JCF has received consulting honoraria from Boehringer-Ingelheim, Merck and Intarcia Therapeutics. All other authors declare that there is no duality of interest associated with their contribution to this manuscript.

Contribution statement JM, GAW, CTL, JD, JCF and JBM participated in the design and conception of the study. JM, BP, MG and JF acquired and analysed the data. All authors participated in the interpretation of data, drafting of the manuscript and its revisions and approved the final version. JM and JBM are the guarantors of this work and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. NCD Risk Factor Collaboration (NCD-RisC) (2016) Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. Lancet 387:1513–1530
2. GBD (2015) Disease and Injury Incidence and Prevalence Collaborators (2016) Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 388:1545–1602
3. Knowler WC, Barrett-Connor E, Fowler SE et al (2002) Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 346:393–403
4. Diabetes Prevention Program Research Group (2015) Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the Diabetes Prevention Program Outcomes Study. Lancet Diabetes Endocrinol 3:866–875
5. Wilson PWF, Meigs JB, Sullivan L et al (2007) Prediction of incident diabetes mellitus in middle-aged adults. Arch Intern Med 167:1068
6. Tirosh A, Shai I, Tekes-Manova D et al (2005) Normal fasting plasma glucose levels and type 2 diabetes in young men. N Engl J Med 353:1454–1462
7. de Vegt F, Dekker JM, Jager A et al (2001) Relation of impaired fasting and postload glucose with incident type 2 diabetes in a Dutch population: The Hoorn Study. JAMA 285:2109–2113
8. Choi SH, Kim TH, Lim S et al (2011) Hemoglobin A1c as a diagnostic tool for diabetes screening and new-onset diabetes prediction: a 6-year community-based prospective study. Diabetes Care 34:944–949
9. Anjana RM, Shanthi Rani CS, Deepa M et al (2015) Incidence of diabetes and prediabetes and predictors of progression among Asian Indians: 10-year follow-up of the Chennai Urban Rural Epidemiology Study (CURES). Diabetes Care 38:1441–1448
10. Newgard CB, An J, Bain JR et al (2009) A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. Cell Metab 9:311–326
11. Wang TJ, Larson MG, Vasan RS et al (2011) Metabolite profiles and the risk of developing diabetes. Nat Med 17:448–453
12. Rhee EP, Cheng S, Larson MG et al (2011) Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. J Clin Invest 121:1402–1411
13. Newgard CB (2012) Interplay between lipids and branched-chain amino acids in development of insulin resistance. Cell Metab 15: 606–614
14. Wang TJ, Ngo D, Psychogios N et al (2013) 2-Aminoadipic acid is a biomarker for diabetes risk. J Clin Invest 123:4309–4317
15. Floegel A, Stefan N, Yu Z et al (2013) Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. Diabetes 62:639–648
16. Walford GA, Porneala BC, Dauriz M et al (2014) Metabolite traits and genetic risk provide complementary information for the prediction of future type 2 diabetes. Diabetes Care 37:2508–2514
17. Drogan D, Dunn WB, Lin W et al (2015) Untargeted metabolic profiling identifies altered serum metabolites of type 2 diabetes mellitus in a prospective, nested case control study. Clin Chem 61:487–497
18. Walford GA, Ma Y, Clish C et al (2016) Metabolite profiles of diabetes incidence and intervention response in the Diabetes Prevention Program. Diabetes 65:1424–1433
19. Kannel WB, Feinleib M, McNamara PM et al (1979) An investigation of coronary heart disease in families. The Framingham offspring study. Am J Epidemiol 110:281–290

20. Shaham O, Wei R, Wang TJ et al (2008) Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. Mol Syst Biol 4:214
21. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC (1985) Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 28:412–419
22. Friedewald WT, Levy RI, Fredrickson DS (1972) Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin Chem 18:499–502
23. Tibshirani R (1996) Regression shrinkage and selection via the Lasso on JSTOR. J R Stat Soc 58:267–288
24. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845
25. Xia J, Sinelnikov IV, Han B, Wishart DS (2015) MetaboAnalyst 3.0–making metabolomics more meaningful. Nucleic Acids Res 43:W251–W257
26. Kettunen J, Demirkan A, Würtz P et al (2016) Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nat Commun 7:11122
27. Fuchsberger C, Flannick J, Teslovich TM et al (2016) The genetic architecture of type 2 diabetes. Nature 536:41–47
28. Burgess S, Butterworth A, Thompson SG (2013) Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol 37:658–665
29. Bowden J, Davey Smith G, Burgess S (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol 44:512–525
30. Tabák AG, Herder C, Rathmann W et al (2012) Prediabetes: a high-risk state for diabetes development. Lancet 379:2279–2290
31. Sun L, Liang L, Gao X et al (2016) Early prediction of developing type 2 diabetes by plasma acylcarnitines: a population-based study. Diabetes Care 39:1563–1570
32. Peddinti G, Cobb J, Yengo L et al (2017) Early metabolic markers identify potential targets for the prevention of type 2 diabetes. Diabetologia 60:1740–1750
33. Kikuchi G (1973) The glycine cleavage system: composition, reaction mechanism, and physiological significance. Mol Cell Biochem 1:169–187
34. Guasch-Ferre M, Hruby A, Toledo E et al (2016) Metabolomics in prediabetes and diabetes: a systematic review and meta-analysis. Diabetes Care 39:833–846
35. Xie W, Wood AR, Lyssenko V et al (2013) Genetic variants associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes. Diabetes 62:2141–2150
36. Mahendran Y, Jonsson A, Have CT et al (2017) Genetic evidence of a causal effect of insulin resistance on branched-chain amino acid levels. Diabetologia 60:873–878