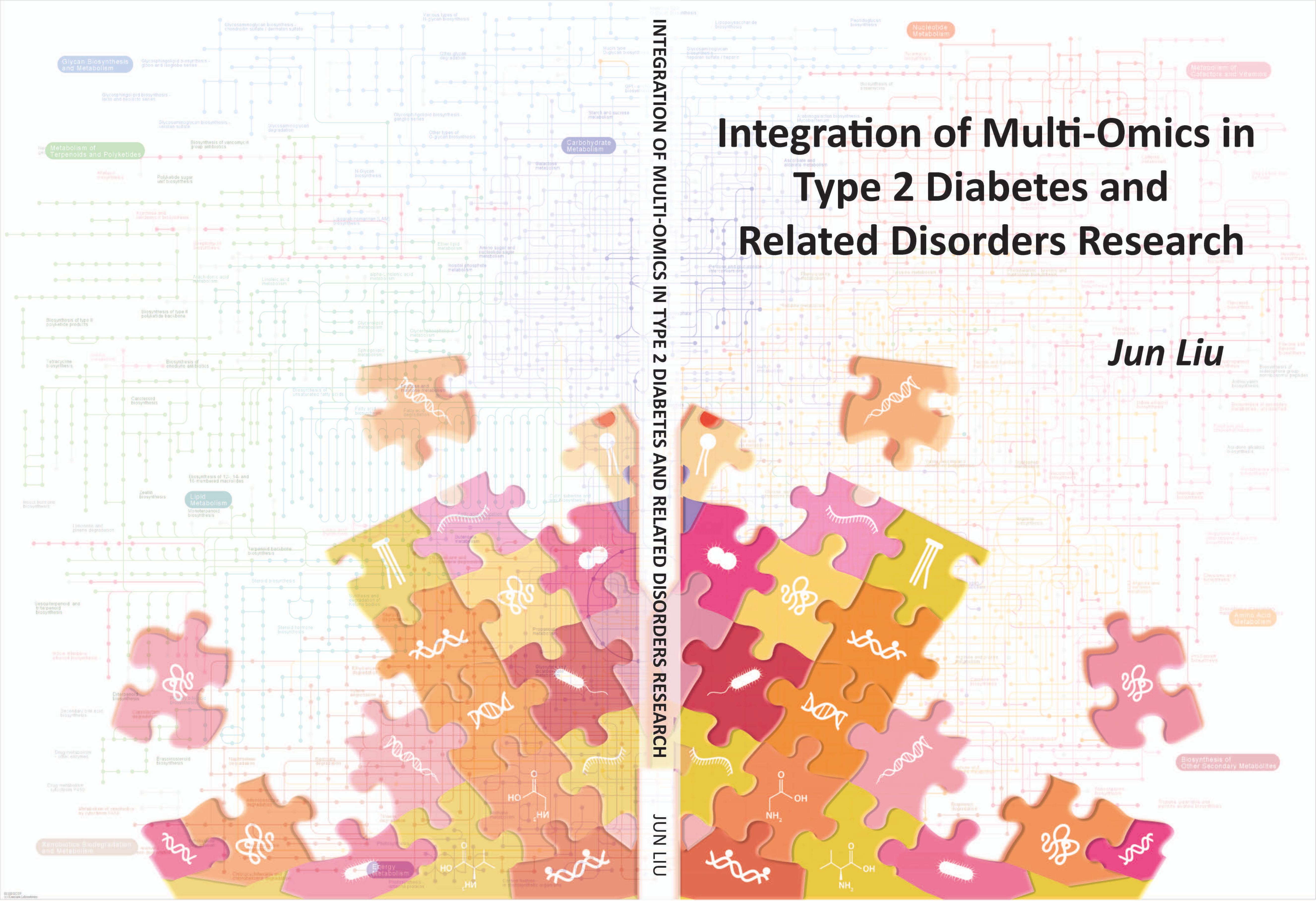

Glycan Biosynthesis and Metabolism
Metabolism of Terpenoids and Polyketides
Carbohydrate Metabolism
Lipid Metabolism
Xenobiotics Biodegradation and Metabolism
Energy Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Amino Acid Metabolism
Biosynthesis of Other Secondary Metabolites
HO
NH
OH
NH2

INTEGRATION OF MULTI-OMICS IN TYPE 2 DIABETES AND RELATED DISORDERS RESEARCH

JUN LIU

Integration of Multi-Omics in Type 2 Diabetes and Related Disorders Research

Jun Liu

STELLINGEN BEHORENDE BIJ HET PROEFSCHRIFT:

Integration of Multi-Omics in Type 2 Diabetes and Related Disorders Research

1. DNA differential methylation explains at least 16.9% of the association between obesity and insulin. *(this thesis)*
2. Among the low risk populations, a multi-metabolite risk prediction model predicts future type 2 diabetes better than fasting glucose. *(this thesis)*
3. Large high-density-lipoprotein (HDL) particles and small HDL particles have distinct causal associations with glucose. *(this thesis)*
4. The blood pressure-related phosphatidylethanolamines (PE) 38:3, PE 38:4 and Phosphatidylcholines (PC) 40:5 are under control of fatty acid desaturase pathways which links them to lipoproteins, blood cell counts (white, red and platelet) and pulse rate. *(this thesis)*
5. The metabolic atlas shows a link between proton pump inhibitors, the gut microbiome, circulating metabolites and liver function. *(this thesis)*
6. Omics studies, by their nature, rely on large numbers of comparisons, tailored statistical analyses, and a considerable investment of time, skilled manpower, and money. *(Hasin, et al., Genome Biology, 2017)*
7. It is important to understand which “at-risk” individuals are most likely to progress to overt disease. *(Wang, et al., Nat Med, 2011)*
8. The tens of thousands of small molecules circulating in the blood can reflect many causal chains of events between genes, traits, and critically, the environment. *(Edward Lau, et al., Circulation Research, 2018)*
9. The focus of metabolomic studies is shifting from cataloguing chemical structures to finding biological stories. *(Baker M., Nat Methods, 2011)*
10. It is now possible to define metabolic signatures of drug exposure that can identify pathways involved in both drug efficacy and adverse drug reactions. *(Kaddurah-Daouk, et al., Clin Pharmacol Ther, 2014)*
11. A man should look for what is, and not for what he thinks should be. *(Albert Einstein)*

Jun Liu

17 January, 2020

Integration of Multi-Omics in Type 2 Diabetes and Related Disorders Research

Jun Liu

Acknowledgments

The work described in this thesis was conducted at the Genetic Epidemiology unit, Department of Epidemiology at the Erasmus Medical Center, Rotterdam, the Netherlands.

Erasmus Rucphen Family (ERF) was supported by the Consortium for Systems Biology (NCSB), both within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO). ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme “Quality of Life and Management of the Living Resources” of 5th Framework Programme (no. QLG2-CT-2002-01254) as well as FP7 project EUROHEADPAIN (nr 602633). The Rotterdam Study is funded by Erasmus Medical Centre and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. High-throughput analysis of the ERF and Rotterdam Study data was supported by joint grant from Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). High throughput metabolomics measurements of the ERF and Rotterdam Study has been supported by Biobanking and Biomolecular Resources Research Infrastructure of the Netherlands (BBMRI-NL, 184.021.007). The methylation array data of Rotterdam Study received funding from the European Union’s Horizon 2020 research and innovation programme (733206, LIFECYCLE). Lipidomics data of ERF was supported by the European Commission FP7 grant LipidomicNet (2007-202272).

The research described in this thesis is a part of the Veni grant (2015) from ZonMw (VENI 91616165), the Personalized pREvention of Chronic DIseases consortium (PRECeDI) (H2020-MSCA-RISE-2014), CardioVasculair Onderzoek Nederland (CVON 2012-03), the Common mechanisms and pathways in Stroke and Alzheimer's disease (CoSTREAM) project (www.costream.eu, grant agreement No 667375), Memorabel program (project number 733050814), Netherlands X-omics Research Infrastructure and U01-AG061359 NIA and European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947).

I gratefully thank to the BBMRI-NL consortium, Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium and European Special Populations Research Network (EUROSPAN). I also thank all the participants from the community for their interest and willingness to contribute to all the studies in the thesis.

Printing of this thesis was supported by: Erasmus University Rotterdam, the department of epidemiology,Erasmus MC and ChipSoft. Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

Layout: Jun Liu
Cover design: Jun Liu (image adapted from https://www.genome.jp/kegg and nipic.com)
Printed by: ProefschriftMaken

ISBN: 978-94-6380-611-4

Integration of Multi-Omics in Type 2 Diabetes and Related Disorders Research

Integratie van multi-omics in type 2 diabetes en gerelateerde aandoeningen

Proefschrift

ter verkrijging van de graad van doctor aan de

Erasmus Universiteit Rotterdam

op gezag van de

rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

vrijdag 17 januari 2020 om 13.30 uur

door

Jun Liu

geboren te Hunan, China

Erasmus University Rotterdam

Promotiecommissie:

Promotor:	Prof.dr.ir C.M. van Duijn
Overige leden:	Prof.dr. E. Sijbrands
	Prof.dr. I. Prokopenko
	Prof.dr. K.W. van Dijk
Copromotor:	Dr. A. Demirkan

Paranimfen:	Hata Karamujić-Čomić
	Jinluan Chen

To my parents

献给我的父母

Table of Contents

Publications and manuscripts in this thesis

Chapter 2.1

Liu J*, Carnero-Montoro E*, van Dongen J, Lent S, Nedeljkovic I, Ligthart S, Tsai PC, Martin TC, Mandaviya PR, Jansen R, Peters MJ, Duijts L, Jaddoe VWV, Tiemeier H, Felix JF, Willemsen G, de Geus EJC, Chu AY, Levy D, Hwang SJ, Bressler J, Gondalia R, Salfati EL, Herder C, Hidalgo BA, Tanaka T, Moore AZ, Lemaitre RN, Jhun MA, Smith JA, Sotoodehnia N, Bandinelli S, Ferrucci L, Arnett DK, Grallert H, Assimes TL, Hou L, Baccarelli A, Whitsel EA, van Dijk KW, Amin N, Uitterlinden AG, Sijbrands EJG, Franco OH, Dehghan A, Spector TD, Dupuis J, Hivert MF, Rotter JI, Meigs JB, Pankow JS, van Meurs JBJ, Isaacs A, Boomsma DI, Bell JT, Demirkan A**, van Duijn CM**: **An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis**. Nat Commun 2019;10:2581
*These authors contributed equally to this work
**These senior authors contributed equally to this work

Chapter 3.1

Liu J, Semiz S, van der Lee SJ, van der Spek A, Verhoeven A, van Klinken JB, Sijbrands E, Harms AC, Hankemeier T, van Dijk KW, van Duijn CM, Demirkan A: **Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study**. Metabolomics 2017;13:104

Chapter 3.2

Liu J, van Klinken JB, Semiz S, van Dijk KW, Verhoeven A, Hankemeier T, Harms AC, Sijbrands E, Sheehan NA, van Duijn CM, Demirkan A: **A Mendelian Randomization Study of Metabolite Profiles, Fasting Glucose, and Type 2 Diabetes**. Diabetes 2017;66:2915-2926

Chapter 3.3

Liu J*, Vries PSd*, Greco FD, Johansson Å, Schraut KE, Hayward C, Franco OH, Hicks AA, Vitart V, Rudan I, Campbell H, Polasek O, Pramstaller PP, Wilson JF, Gyllensten U, van Duijn CM, Dehghan A**, Demirkan A**: **A multi-omics study of circulating phospholipid markers of blood pressure.** (In preparation)
*These authors contributed equally to this work
**These senior authors contributed equally to this work

Chapter 4.1

Liu J, Lahousse L, Nivard MG, Bot M, Chen L, van Klinken JB, Thesing CS, Beekman M, van der Akker EB, Slieker RC, Waterham E, van der Kallen CJH, der Boer I, Li-Gao R, Vojinovic D, Amin N, Radjabzadeh D, Kraaij R, Alferink LJM, Murad SD, Uitterlinden AG, Willemsen G, Pool R, Milaneschi Y, van Heemst D, Suchiman HED, Rutters F, Elders PJM, Beulens JWJ, van der Heijden AAWA, van Greevenbroek MMJ, Arts ICW, Onderwater GLJ, van der Maagdenberg AMJM, Mook-Kanamori DO, Hankemeier T, Terwindt GM, Stehouwer CDA, Geleijnse JM, Hart LMt, Slagboom PE, van Dijk KW, Zhernakova A, Fu J, Penninx BWJH, Boomsma D, Demirkan A, Stricker BHC, van Duijn CM: **Integration of epidemiologic, pharmacologic, genetic and gut microbiome data in a drug-metabolite atlas**. Nature Medicine; in press

Chapter 2

Epigenomics of glucose and insulin homeostasis

Chapter 2.1

An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis

Jun Liu[§], Elena Carnero-Montoro[§], Jenny van Dongen, Samantha Lent, Ivana Nedeljkovic, Symen Ligthart, Pei-Chien Tsai, Tiphaine C. Martin, Pooja R. Mandaviya, Rick Jansen, Marjolein J. Peters, Liesbeth Duijts, Vincent W.V. Jaddoe, Henning Tiemeier, Janine F. Felix, Gonneke Willemsen, Eco J.C. de Geus, Audrey Y. Chu, Daniel Levy, Shih-Jen Hwang, Jan Bressler, Rahul Gondalia, Elias L. Salfati, Christian Herder, Bertha A. Hidalgo, Toshiko Tanaka, Ann Zenobia Moore, Rozenn N. Lemaitre, Min A Jhun, Jennifer A. Smith, Nona Sotoodehnia, Stefania Bandinelli, Luigi Ferrucci, Donna K. Arnett, Harald Grallert, Themistocles L. Assimes, Lifang Hou, Andrea Baccarelli, Eric A. Whitsel, Ko Willems van Dijk, Najaf Amin, André G. Uitterlinden, Eric J.G. Sijbrands, Oscar H. Franco, Abbas Dehghan, Tim D. Spector, Josée Dupuis, Marie-France Hivert, Jerome I. Rotter, James B. Meigs, James S. Pankow, Joyce B.J. van Meurs, Aaron Isaacs, Dorret I. Boomsma, Jordana T. Bell, Ayşe Demirkan* and Cornelia M. van Duijn*

[§] These authors contributed equally to this work.
* These authors jointly supervised this work.

This chapter was published in *Nature Communications* (2019) 10:2581.

The supplemental information for this paper is available online at
https://doi.org/10.1038/ s41467-019-10487-4

Abstract

Despite existing reports on differential DNA methylation in type 2 diabetes (T2D) and obesity, our understanding of its functional relevance remains limited. Here we show the effect of differential methylation in the early phases of T2D pathology by a blood-based epigenome-wide association study of 4,808 non-diabetic Europeans in the discovery phase and 11,750 individuals in the replication. We identify CpGs in *LETM1*, *RBM20*, *IRS2*, *MAN2A2* and the 1q25.3 region associated with fasting insulin, and in *FCRL6*, *SLAMF1*, *APOBEC3H* and the 15q26.1 region with fasting glucose. *In silico* cross-omics analyses highlight the role of differential methylation in the crosstalk between the adaptive immune system and glucose homeostasis. The differential methylation explains at least 16.9% of the association between obesity and insulin. Our study sheds light on the biological interactions between genetic variants driving differential methylation and gene expression in the early pathogenesis of T2D.

Introduction

Type 2 diabetes (T2D) is a common metabolic disease, characterized by disturbances in glucose and insulin metabolism. The pathogenesis of T2D is driven by inherited and environmental factors[1]. There is increasing interest in differential DNA methylation in the development of T2D as well as with glucose and insulin metabolism[2, 3, 4, 5, 6]. Depending on the region, DNA methylation may result in gene silencing and thus regulate gene expression and subsequent cellular functions[7]. Differential methylation in the circulation may predict the development of future T2D beyond traditional risk factors such as age and obesity[3, 8], but it may also be part of the biological mechanism that links age and/or obesity to glucose, insulin metabolism and/or T2D. A recent longitudinal study with multiple visits reported that most DNA methylation changes occur 80–90 days before detectable glucose elevation[9], suggesting that differential DNA methylation evokes changes in glucose and is involved in the early stage(s) of diabetes. Differential DNA methylation is further associated with obesity, which is an important driver of the T2D risk and also precedes the increase in glucose and insulin level in persons developing T2D[8]. A key question to answer is whether the differential methylation associated with glucose and insulin metabolism is an irrelevant epiphenomenon that is related to obesity acting as a statistical confounder or whether there are functional effects of the differential methylation relevant of obesity that is associated to metabolic pathology.

Here, we determine the relation of differential DNA methylation and fasting glucose and insulin metabolism as markers of early stages of diabetes pathology in non-diabetic subjects, accounting for obesity measured as body mass index (BMI). We identify and replicate nine CpG sites associated with fasting glucose (in *FCRL6*, *SLAMF1*, *APOBEC3H* and the 15q26.1 region) and insulin (in *LETM1*, *RBM20*, *IRS2*, *MAN2A2* and the 1q25.3 region). Using cross-omics analyses, we present *in silico* evidence supporting the functional relevance of the CpG sites on the development and progression of diabetes, in terms of their effect on expression paths and elucidate the genetic networks involved.

Results

Epigenome-wide association analysis and replication

In the discovery phase, we performed a blood-based epigenome-wide association study (EWAS) meta-analysis of four cohorts including 4,808 non-diabetic individuals of European ancestry (Supplementary Data 1), which revealed differential DNA methylation at

28 unique CpG sites in either the baseline model without BMI adjustment or in the second model with BMI adjustment (Table 1 and Supplementary Table 1). The summary statistic results of the EWAS are provided as a Source Data file. These include three CpG sites associated with both insulin and glucose, eight CpG sites associated with fasting glucose only and 17 with fasting insulin (P-value < 1.3×10^{-7} in meta-analysis). Of these 28 CpG sites, 13 were identified by earlier EWAS studies of either T2D or related traits, including glucose, insulin, hemoglobin A1c (HbA1c), and homeostatic model assessment-insulin resistance (HOMA-IR)[2, 3, 4, 5, 8, 10, 11] (Supplementary Table 1). The known CpG sites include three sites located in *SLC7A11*, *CPT1A* and *SREBF1* that are associated with both glucose and insulin. The remaining ten CpG sites, located in *DHCR24*, *CPT1A*, *RNF145*, *ASAM*, *KDM2B*, *MYO5C*, *TMEM49*, *ABCG1* (harboring two CpG sites) and the 4p15.33 region, are associated with insulin only. All of the previously reported CpG sites with glycemic traits are also associated with BMI in previous EWAS[8, 10, 12, 13, 14, 15] (Supplementary Table 1).

The 15 novel CpG sites were tested using the same statistical models in 11 independent cohorts including 11,750 non-diabetic participants from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium (Supplementary Data 1). Nine unique CpG-trait associations were replicated when correcting for multiple testing using Bonferroni (15 tests, P-value threshold for significance < 3.3×10^{-3}) and were investigated in the further analyses (Table 2). These include five sites (in *LETM1*, *RBM20*, *IRS2*, *MAN2A2* and the 1q25.3 region) associated with fasting insulin and one site (in *FCRL6)* associated with fasting glucose in the baseline model without adjusting for BMI, and three (in *SLAMF1*, *APOBEC3H* and the 15q26.1 region, all associated with fasting glucose) emerging in the BMI-adjusted model. Of note, no locus was found to be associated with fasting insulin in the BMI-adjusted model.

Because the replication cohorts also included individuals of African ancestry (AA, n = 4,355) and Hispanic ancestry (HA, n = 577), we also performed the replication stratified by ancestry (Supplementary Data 2). Two CpG sites (cg13222915 and cg18247172) were replicated in the AA population when corrected by the number of tests and two (cg00936728 and cg06229674) replicated with nominal significance. In the HA population, cg20507228 was replicated at nominal significance. Two CpG sites (cg18881723 and cg13222915) show the opposite direction for the effect estimate in HA ancestry population as compared to the other two populations. However, the estimates of effect size are not significantly different from zero (P-value = 0.63 in cg18881723 and P-value = 0.092 in cg13222915).

Glycemic differential DNA methylation and transcriptomics

To determine whether the differential DNA methylation has functional effects on gene expression and subsequent cellular functions, we conducted three series of analyses. Figure 1 shows the overview of the cross-omics analyses. First, we explored the Genotype-Tissue Expression (GTEx)[16] database for the expression levels of the genes which annotated to the novel CpG sites. We found that the genes are expressed in a wide range of tissues, including whole blood and spleen (in particular *MAN2A2* and *RBM20*), but also other tissues relevant for glucose and insulin metabolism such as adipose subcutaneous, adipose visceral omentum, liver (in particular *SLAMF1, APOBEC3H, FCRL6* and *RBM20*), pancreas and skeletal muscle (in particular *SLAMF1*, *APOBEC3H*, *FCRL6* and *MAN2A2*) and small intestine terminal ileum (in particular *MAN2A2*, *RBM20*, *FCRL6* and *APOBEC3H*; Supplementary Figure 1).

Table1 CpG sites associated with glycemic traits in discovery phase.

Locus	CpG	Chr: Pos	Trait	Beta$_{M1}$	P-value$_{M1}$	Beta$_{M2}$	P-value$_{M2}$
FCRL6	cg00936728	1:159772194	Glucose	-1.79	$9.1 \times 10^{-8‡}$	-1.60	1.9×10^{-7}
SLAMF1	cg18881723	1: 160616870	Glucose	1.16	$7.5 \times 10^{-8‡}$	1.25	$3.4 \times 10^{-10‡}$
1q25.3	cg13222915	1: 184598594	Insulin	-1.69	$2.6 \times 10^{-9‡}$	-1.06	4.1×10^{-6}
BRE	cg20657709	2: 28509570	Glucose	-1.42	2.7×10^{-6}	-1.53	$4.1 \times 10^{-8‡}$
LRPPRC	cg01913188	2: 44223249	Glucose	1.18	9.4×10^{-6}	1.38	$5.7 \times 10^{-9‡}$
IRAK2	cg14527942	3: 10276383	Insulin	2.44	$3.4 \times 10^{-10‡}$	2.14	$2.9 \times 10^{-11‡}$
LETM1	cg13729116	4: 1859262	Insulin	2.38	$4.3 \times 10^{-8‡}$	1.64	4.5×10^{-6}
RBM20	cg15880704	10: 112546110	Insulin	2.50	$3.8 \times 10^{-9‡}$	1.38	6.7×10^{-5}
IRS2	cg25924746	13: 110432935	Insulin	2.11	$3.0 \times 10^{-9‡}$	1.32	4.9×10^{-6}
SPTB	cg07119168	14: 65225253	Glucose	-1.64	4.4×10^{-7}	-1.63	$4.9 \times 10^{-8‡}$
15q26.1	cg18247172	15: 91370233	Glucose	-1.05	4.9×10^{-6}	-1.18	$2.8 \times 10^{-8‡}$
MAN2A2	cg20507228	15: 91460071	Insulin	1.18	$5.5 \times 10^{-8‡}$	0.87	9.0×10^{-7}
FAM92B	cg06709610	16: 85143924	Insulin	6.22	$6.5 \times 10^{-9‡}$	6.30	$5.8 \times 10^{-13‡}$
CD300A	cg08087047	17: 72461209	Glucose	-1.35	5.9×10^{-6}	-1.45	$1.1 \times 10^{-7‡}$
APOBEC3H	cg06229674	22: 39492189	Glucose	-1.62	1.8×10^{-6}	-1.70	$4.7 \times 10^{-8‡}$

Novel epigenome-wide significant results in the discovery phase (n = 4,808) are shown. ‡ Significant results (P-value $< 1.3 \times 10^{-7}$). Model 1 (M1) indicates inverse variance-weighted fixed effect meta-analysis of effect estimates in four cohorts. Each cohort performed a regression model adjusting for age, sex, technical covariates, white blood cell, and smoking status, and accounting for family structure in family-based cohorts. Model 2 (M2) indicates the meta-analysis of the same studies, adjusting for body mass index (BMI) additionally. Locus: the cytogenetic location or the gene symbol of the CpG sites from Illumina annotation. Beta: effect estimate of the meta-analysis. P-value shown is genomic controlled after meta-analysis. The effect refers to the increase/ decrease in fasting glucose/ insulin as the outcome in the model.

Table2 CpG sites associated with glycemic traits in replication.

Locus	CpG	Chr: Pos	Trait	Beta$_{M1}$	P-value$_{M1}$	Beta$_{M2}$	P-value$_{M2}$
FCRL6	cg00936728	1:159772194	Glucose	-1.55×10^{-3}	$9.6 \times 10^{-5\ddagger}$	NP	NP
SLAMF1	cg18881723	1: 160616870	Glucose	1.17×10^{-3}	7.7×10^{-3}	1.48×10^{-3}	$1.2 \times 10^{-3\ddagger}$
1q25.3	cg13222915	1: 184598594	Insulin	-3.77×10^{-3}	$3.3 \times 10^{-16\ddagger}$	NP	NP
BRE	cg20657709	2: 28509570	Glucose	NP	NP	-9.40×10^{-4}	0.036
LRPPRC	cg01913188	2: 44223249	Glucose	NP	NP	1.64×10^{-5}	0.90
IRAK2	cg14527942	3: 10276383	Insulin	-6.49×10^{-5}	0.48	-7.72×10^{-5}	0.45
LETM1	cg13729116	4: 1859262	Insulin	1.92×10^{-3}	$7.0 \times 10^{-7\ddagger}$	NP	NP
RBM20	cg15880704	10: 112546110	Insulin	3.05×10^{-3}	$8.6 \times 10^{-12\ddagger}$	NP	NP
IRS2	cg25924746	13: 110432935	Insulin	3.38×10^{-3}	$3.0 \times 10^{-11\ddagger}$	NP	NP
SPTB	cg07119168	14: 65225253	Glucose	NP	NP	-7.18×10^{-4}	0.070
15q26.1	cg18247172	15: 91370233	Glucose	NP	NP	-1.77×10^{-3}	$5.1 \times 10^{-4\ddagger}$
MAN2A2	cg20507228	15: 91460071	Insulin	6.11×10^{-3}	$2.3 \times 10^{-15\ddagger}$	NP	NP
FAM92B	cg06709610	16: 85143924	Insulin	2.08×10^{-5}	0.81	5.37×10^{-5}	0.59
CD300A	cg08087047	17: 72461209	Glucose	NP	NP	-4.92×10^{-4}	0.28
APOBEC3H	cg06229674	22: 39492189	Glucose	NP	NP	-2.09×10^{-3}	$1.4 \times 10^{-6\ddagger}$

Novel epigenome-wide significant results in the replication (n = 11,750) are shown. ‡ Significant results (P-value < 3.3×10^{-3}). Replication was not performed in the non-significant associated model or trait (NP). Model 1 (M1) indicates inverse variance-weighted fixed effect meta-analysis of effect estimates in 11 cohorts. Each study performed a regression model adjusting for age, sex, technical covariates, white blood cell, and smoking status, and accounting for family structure in family-based cohorts. Model 2 (M2) indicates the meta-analysis of the same studies, adjusting for body mass index (BMI) additionally. Locus: the cytogenetic location or the gene symbol of the CpG sites from Illumina annotation. Beta: effect estimate of the meta-analysis. The effect refers to the increase/ decrease in methylation as the outcome in the model.

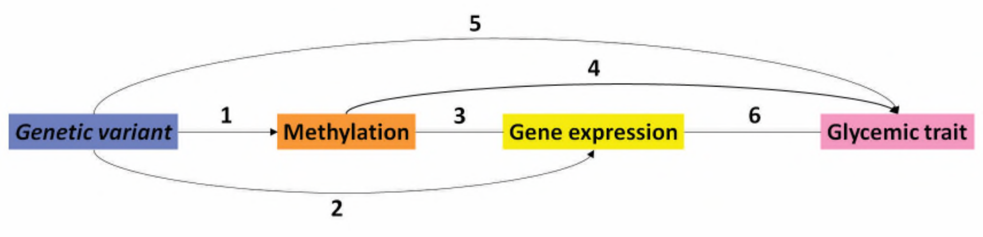


Figure 1 Overview of the cross-omics analysis. 1. Methylation quantitative trait loci (meQTL). 2. Expression quantitative trait loci (eQTL). 3. Expression quantitative trait methylation (eQTM). 4. Epigenome-wide association study (EWAS) and Mendelian randomization (MR). 5. Genome-wide association study (GWAS). 6. The association of gene expression expressed in the glucose or insulin metabolism related tissues and glycemic traits. Results in 1, 2, 3 were extracted from the summary statistics from Biobank-based Integrative Omics Study (BIOS) database (n = 3,814). Results in 4 was the results in the current EWAS (discovery phase, n = 4,808, replication phase, n = 11,750) and the two-sample Mendelian randomization based on the BIOS database (n = 3,814) and GWAS results of Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC). Results in 5 was from the GWAS results of MAGIC or the DIAbetes Genetics Replication And Meta-analysis consortium (DIAGRAM, n = 96,496 ~ 452,244). Results in 6 was based on the summary statistics of Genotype-Tissue Expression project (GTEx) and MAGIC or DIAGRAM (n = 153 ~ 491).

Second, the effect on gene expression in blood of the previously identified 11 independent CpG sites (cg00574958 in *CPT1A* and cg06500161 in *ABCG1* were used) and the nine novel sites from our current study was examined in the Biobank-based Integrative Omics Study (BIOS) database that is part of the Biobanking and BioMolecular Infrastructure of the Netherlands (BBMRI-NL)[17] (indicated in Figure 2 in the orange boxes). We found that five CpG sites, i.e. cg00936728 (*FCRL6*), cg18881723 (*SLAMF1*), cg00574958 (*CPT1A*), cg11024682 (*SREBF1*) and cg06500161 (*ABCG1*), are expression quantitative trait methylations (eQTMs), i.e. there is correlation between gene expression and methylation[18]. In most cases, the differential methylation levels are associated with the expression (indicated in Figure 2 in the yellow boxes) of their respective genes. Cg18881723 (*SLAMF1*) is also associated with the expression of two other genes near *SLAMF1*, i.e. *SLAMF7* and *CD244* (Supplementary Table 2).

Third, we investigated whether the genetically regulated expression of the annotated genes in specific tissues is altered in T2D or related traits, such as glucose, insulin and HbA1c. To answer this question, we mined in the MetaXcan database for genome-wide association studies (GWAS) of T2D, fasting glucose, HbA1c, insulin and HOMA-IR[19, 20, 21, 22, 23] as a genetic proxy for the traits[24]. No association was found between glycemic traits and the DNA expression in adipose subcutaneous, adipose visceral omentum and small intestine terminal ileum. Supplementary Table 3 gives the significant findings for tissues known to be implicated in glucose and insulin metabolism including blood, liver, pancreas and skeletal muscle (P-value < 0.05 for MetaXcan). As described earlier, we associated the increased expression of *SREBF1* with decreased risk of T2D and decreased HbA1c levels in the whole blood[25]. The increased expression in the whole blood of *ABOBEC3H*, a methylation locus we identified in the present study, is associated with increased HOMA-IR level, a measure of insulin resistance. In skeletal muscle, the increased fasting glucose is associated with the increased expression of *KDM2B* and decreased expression of *MAN2A2*. Moreover, we discovered that increased hepatic expression of *FCRL6*, which was annotated to the methylation locus associated with fasting glucose in the present study, is associated with the risk of T2D. In the pancreas, the increased expression of the methylation loci *MYO5C* and *RBM20* are associated with increased fasting glucose levels.

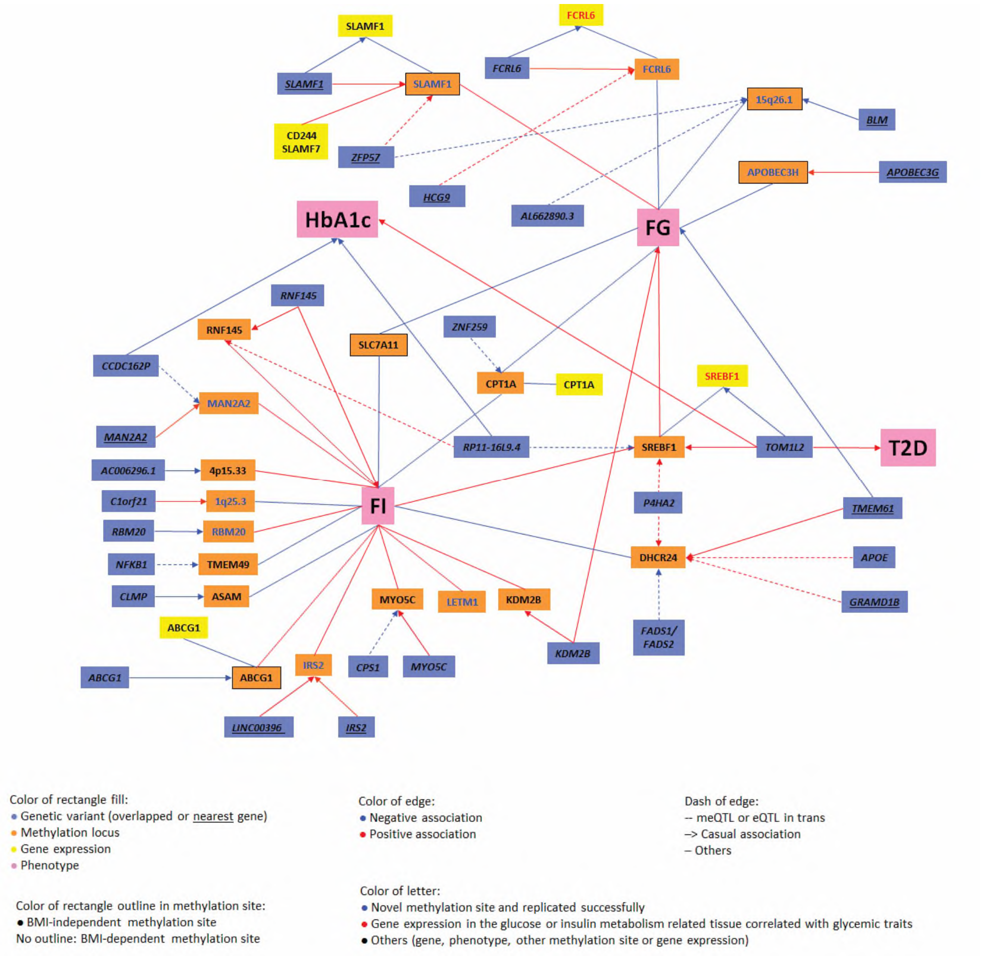


Figure 2 Significant associations of the cross-omics integration. The effect allele is standardized across all associations. Only the significant associations which passed the specific P-value threshold in each association step and the direction of effects consistent were shown in the figure. FG: fasting glucose. FI: fasting insulin. T2D: type 2 diabetes. HbA1c: hemoglobin A1c.

Glycemic differential DNA methylation and genomics

Although differential DNA methylation may be the result of environmental exposures, the process is often (partly) heritable with genetic variants (co-)determining the process[26]. Therefore, we next set out to find whether the differential methylation associated with fasting glucose and insulin levels is driven by genetic variants which referred to as methylation quantitative trait loci (meQTLs). Using the BIOS database (blood-based data)[17], we were able to study 18 out of the 20 unique CpG sites in this respect. We associated 2,991 single-nucleotide polymorphisms (SNPs) in 29 unique meQTLs (indicated in Figure 2 in the blue boxes) with differential methylation either in *cis* or *trans* (for details see Supplementary Data 3). Six of these meQTLs (4 *cis* and 2 *trans*-acting) are also associated with T2D, fasting glucose, fasting Insulin, or HbA1c in earlier studies[21, 22, 27, 28, 29], and the directions of the effect between the SNP, methylation and glycemic traits are consistent (shown in Figure 2 in the pink boxes, for details see Supplementary Table 4). A genetic locus near *TMEM61* is a common genetic driver affecting the differential methylation at nearby CpG cg17901584 (*DHCR24*) in our study and fasting glucose levels in an earlier study[22]. Further, the *RNF145* locus was found to be a common driver affecting the differential methylation at cg26403843 (*RNF145*) and fasting insulin levels[21]. The *KDM2B* locus affects differential methylation at cg13708645 (*KDM2B*) and fasting glucose levels[22], and the *TOM1L2*/*RAI1* locus affects the differential methylation at cg11024682 (*SREBF1*) as well as HbA1c and T2D[27, 28]. Two *trans*-acting loci involve a genetic locus in *CCDC162P* that is affecting differential methylation at cg20507228 (*MAN2A2*) and HbA1c[27] and the genetic locus in *RP11-16L9.4* affecting the differential methylation at cg11024682 (*SREBF1*) and HbA1c[27].

We next explored if these genetic variants associated with differential methylation (meQTLs) are also associated with gene expression, i.e. quantitative trait loci (eQTLs; see the integrated outline of analyses in Figure 2 and detailed in Supplementary Table 5). We searched specifically for expression profiles earlier associated with glycemic CpG sites in blood (listed in Supplementary Table 2). We associated three genetic variants with both differential methylation and gene expression in blood. These include that: 1) rs11265282 in *FCRL6* is positively associated with the differential methylation at cg00936728 (*FCRL6*) and decreased the expression of *FCRL6* in blood, 2) rs1577544 near *SLAMF1* is associated with decreased differential methylation at cg18881723 (*SLAMF1*) and decreased *SLAMF1* expression in blood, and 3) rs6502629 in *TOM1L2* is associated with increased differential methylation at cg11024682 (*SREBF1*) and decreased *SREBF1* expression in blood.

As we observed that the genes driving glycemic CpG sites overlapped with genetic determinants of T2D or related traits, we studied the causal effect of differential methylation on glucose and insulin metabolism with a generalized summary statistic-based Mendelian randomization (MR) test[30]. Up to eight independent genetic variants include in the genetic risk score were used as the instrumental variable for each CpG. Thirteen CpG sites out of the initial 20 met the present MR criteria and were tested by MR (Supplementary Data 4). No significant association was detected when adjusting for multiple testing accounting for 13 independent tests (P-value threshold for significance < 3.8×10^{-3}). The genetic risk score for cg15880704 (*RBM20*) methylation levels is nominally significantly associated with fasting insulin levels (P-value = 0.04), and the genetic risk score for cg18881723 (*SLAMF1*) levels is nominally associated with fasting glucose levels (P-value = 0.05) in the MR tests.

Multi-omics integration and functional annotation

To understand the biological relevance of our findings, we first integrated the cascade of associations into genomics, epigenomics, transcriptomics and glycemic traits through EWAS, eQTM, meQTL and eQTL. There are three pathways emerging when considering the consistency of the direction of the effects between the associations. One pathway involves *SREBF1* which in part was reported earlier[3, 25, 31] but substantially extended in the current report. The other two involve differential methylation of *FCRL6* and *SLAMF1* (Figure 3). The C allele of rs11265282 in *FCRL6* is associated with increased methylation, which turns down the *FCRL6* expression in blood. In addition, the genetically decreased *FCRL6* expression in the liver is also associated with a decreased risk of T2D. The T allele of rs1577544 near *SLAMF1* increases the differential methylation levels in the blood, which decreases *SLAMF1* expression in the circulation, consistent with the negative association between the genetic variant and gene expression levels.

To understand the correlation of the findings, we clustered the normalized differential methylation values of the nine novel CpG sites including those not annotated to a gene. Two clusters emerge, one including *IRS2, MAN2A2*, 1q25.3 locus (intergenic), *RBM20, LETM1* and *SLAMF1* and the second one including *FCRL6*, 15q26.1 (intergenic) and *APOBEC3H* (Figure 4 and Supplementary Table 6). Four CpGs in *FCRL6*, 15q26.1 (intergenic), *APOBEC3H* and *SLAMF1* are highly correlated with each other, in which the absolute correlation coefficients are bigger than 0.6, while they are located in different chromosomes, suggesting a common biological mechanism: *SLAMF1* and *FCRL6* from chromosome 1, *APOBEC3H* from chromosome 22 and 15q26.1 from chromosome 15. We next performed gene set enrichment analysis in different pathway databases, including KEGG pathways[32], Reactome Pathway Knowledgebase[33] and Gene Ontology (GO) biological process classification[34]. We found that the genes in the first cluster are highly enriched together in multiple pathways, including regulation of leukocyte proliferation, protein secretion and cell activation (*SLAMF1* and *IRS2*), hexose, monosaccharide and carbohydrate metabolism (*IRS2* and *MAN2A2*). Further, *SLAMF1* (cluster 1) and *APOBEC3H* (cluster 2) are both enriched in immune effector processes and innate immune response (Supplementary Table 7).

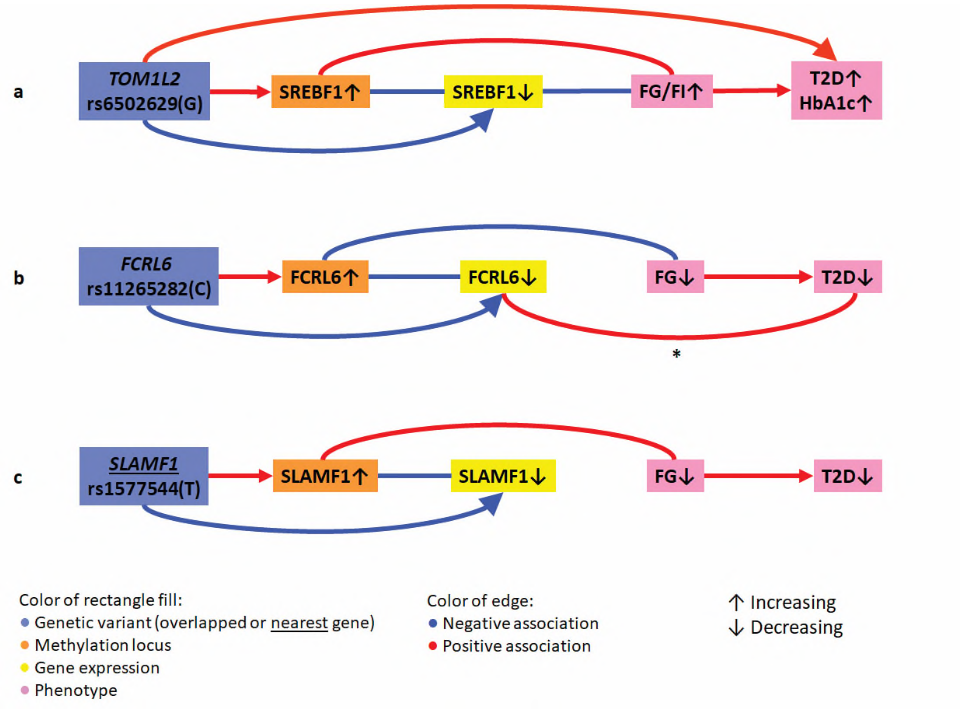


Figure 3 The cross-omics integration of CpGs in *SREBF1* (a), *FCRL6* (b) *and SLAMF1* (c). Cascading associations cross multi-omics were integrated in the network. * The association happens in the FCRL6 expression in liver. All other differential methylation or gene expression was measured in blood. FG: fasting glucose. FI: fasting insulin. T2D: type 2 diabetes. HbA1c: hemoglobin A1c.

BMI in the association of methylation and glycemic traits

Of note, among the 20 methylation loci associated with glycemic metabolism in the present analyses, 11 are associated with BMI in the previous EWAS[8, 10, 12, 13, 14, 15]. These 11 loci are all associated with insulin metabolism (Supplementary Table 1). Based on the bi-direction MR findings performed as part of the previous EWAS of BMI[8], we found that BMI appears to drive methylation for cg06500161 (*ABCG1*, P-value = 6.4×10^{-5}), a CpG that we associated with insulin levels. Using a marginal P-value of 0.05 in their MR results, the differential methylation appears to be a consequence of obesity rather than a cause for three other CpG sites: cg110244682 (*SREBF1*; P-value = 4.1×10^{-3}), cg17901584 (*DHCR24*; P-value = 4.1 × 10^{-3}) and cg26403843 (*RNF145*; P-value = 0.011)[8]. Taken together (Supplementary Table 1), our results raise the question whether BMI is driving differential methylation, which subsequently raises insulin level in the circulation. Such a pathway would predict that the association between BMI and insulin changes when adjusting for differential methylation at *ABCG1*, *SREBF1, DHCR24* and *RNF145*. We tested this hypothesis in the non-diabetic individuals of the Rotterdam Study by comparing the relationship between BMI and fasting insulin with and without adjusting for the methylation levels at the four CpG sites. The variance explained (R^2) by the linear regression model improves significantly from 0.40 to 0.43 (P-value = 1.2 × 10^{-13} by analysis of variance (ANOVA) testing) when adjusting for the CpG effect, while the effect estimates for BMI decrease by 9.2% (beta: 0.065, standard error (SE): 0.003, P-value = 1.2 × 10^{-82} for the model without CpG adjustment compared to beta: 0.059, SE: 0.003, P-value = 2.9 × 10^{-70} adjusting for the four CpGs). When we extended the adjustment to the 16 independent CpG sites associated with circulating insulin levels, the variance explained by the model improves further (R^2=0.46, P-value = 2.1 × 10^{-18}) and the beta for BMI reduces further by 16.9% (beta: 0.054, SE: 0.003, P-value = 4.6 × 10^{-58} for the model adjusting for 16 CpG sites).

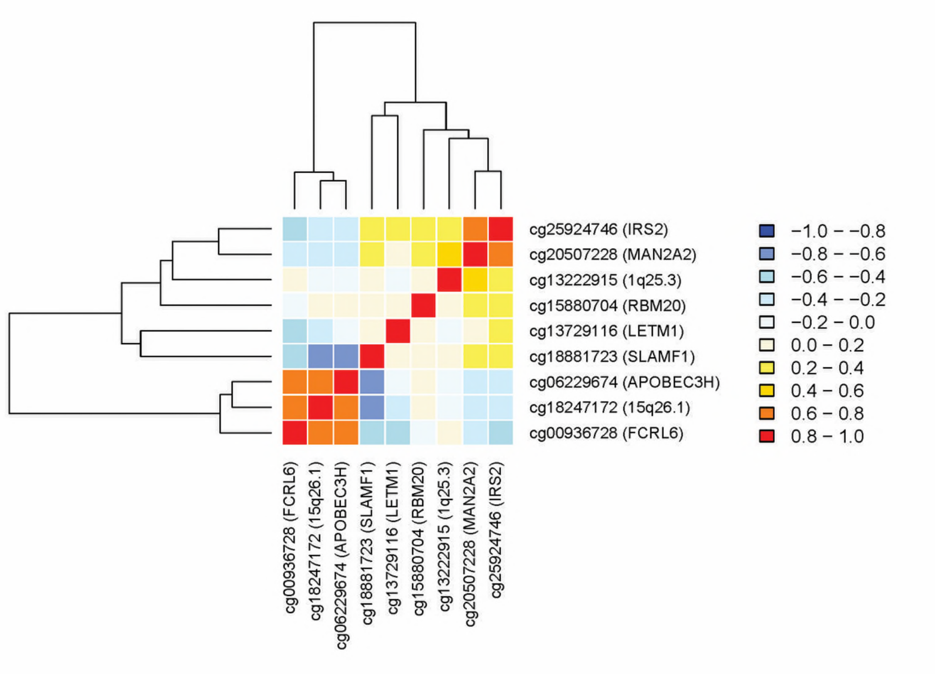


Figure 4 Clustered correlation of the nine novel glycemic CpGs. The correlation of the novel CpG sites was checked by Pearson's correlation test (n = 1,544). The hierarchical cluster analysis was used in the clustering.

Discussion

The current large-scale EWAS identify and replicate nine CpG sites associated with fasting glucose (in *FCRL6*, *SLAMF1*, *APOBEC3H* and the 15q26.1 region) or insulin (in *LETM1*, *RBM20*, *IRS2*, *MAN2A2* and the 1q25.3 region). When we adjust for BMI as a potential confounder, three CpG sites (in *SLAMF1*, *APOBEC3H* and the 15q26.1 region) are associated with fasting glucose only after adjustment for BMI. We validate 13 previously reported CpG sites from 11 independent genetic loci[2, 3, 4, 5, 6, 8, 10, 12, 13, 14, 15] and complement the understanding on why these CpG sites are associated with T2D and/or glycemic traits based on comprehensive cross-omics analyses. We present *in silico* evidence supporting the functional relevance of the CpG sites, in terms of their effect on expression paths and elucidate the genetic networks involved.

Our data shows that differential methylation plays a key role in understanding the immunological changes observed in glucose metabolism[35]. *SLAMF1* and *APOBC3H* are both enriched in immune function and the innate immune response. The differential methylation level at *FCRL6*, 15q26.1 (intergenic), *APOBEC3H* and *SLAMF1* were highly correlated though they on three different chromosomes. This finding suggests a common pathway. *SLAMF1* belongs to the immunoglobulin gene superfamily and is involved in T-cell stimulation[36]. *APOBEC3H* proteins are part of an intrinsic immune defense that has potent activity against a variety of retroelements[36] and its expression in whole blood is positively associated with HOMA-IR from the current study. *FCRL6* is a distinct indicator of cytotoxic effector T-lymphocytes that is upregulated in diseases characterized by chronic immune stimulation[36]. Meanwhile, we show that decreased *FCRL6* differential methylation increased expression of *FCRL6* and fasting glucose in the blood. A key finding that links *FCRL6* to glucose metabolism is that the genetically determined *FCRL6* expression in the liver is also associated with decreased risk of T2D. In line with a role in immune relation and pathology[37, 38], the HLA region (6p22.1 region) is a key meQTLs of *FCRL6* (rs2523946), 15q26.1 (rs3129055 and rs4324798) and *SLAMF1 (rs3129055)*. Of interest is that in the population of non-diabetic individuals, we found strong signals of the immune system particularly when we adjust the effects attributed to BMI. Remarkably, three out of the four methylation loci at *SLAMF1*, *APOBEC3H* and the 15q26.1 region emerged in the BMI-adjusted model, suggesting that these associations were masked by confounding noise of BMI on methylation in opposite effects to that of insulin.

We studied the interplay between BMI, fasting glucose and insulin levels and differential methylation in the circulation. On the one hand, we find evidence that the differential methylation of the insulin-related CpG sites together explained up to 16.9% of the association between obesity and insulin levels. These findings are in line with the Nature paper on the EWAS of BMI that found that the methylation patterns in blood predict future diabetes[8]. Our study reveals that insulin is a key player underlying the association reported earlier[8]. On the other hand, we find evidence that the association between differential methylation and insulin metabolism is attenuated up to 62%, e.g. CpG sites in *SREBF1* (62%), *ASAM* (56%), *CPT1A* (54%) and *TMEM49* (52%), when BMI is accounted for in the model, suggesting that the interplay between BMI, differential methylation and insulin metabolism is extremely complex and differs across CpG sites. BMI may be a confounder of associations for some CpGs but may be in the causal pathway for others.

To our knowledge, we report for the first time that in blood, differential methylation of *IRS2* was associated with fasting insulin level. Expression level of *IRS2* (insulin receptor substrate 2) in β-cells in the pancreas are associated with the onset of diabetes[39, 40, 41]. Though the expression level of *IRS2* is low in blood, we find its blood-based differential methylation was associated with fasting insulin. We also find an insulin-related genetic locus, *MAN2A2* (mannosidase alpha class 2A member 2) in our EWAS. *MAN2A2* encodes an enzyme that forms intermediate asparagine-linked carbohydrates (N-glycans)[42]. It is related to the hexose/monosaccharide metabolism. In addition, the expression of *MAN2A2* in skeletal muscle is negatively associated with fasting glucose level and the meQTL (rs9374080*)* of *MAN2A2* associates with HbA1c[27]. Together, these findings suggest that regulating the differential methylation level or expression level of *MAN2A2* may be relevant to the development of insulin resistance. Another interesting gene that emerged is the familial cardiomyopathy related gene *RBM20,* which may play a role in cardiovascular complications of diabetes via mediating insulin damage in cardiac tissues[43]. The expression of *RBM20* in the pancreas is also associated with fasting glucose. The meQTL for *RBM20* is associated with pulse rate (P-value = 4.6 × 10^{-5}) in UKBIOBANK GWAS[44], and its mRNA is highly expressed in cardiac tissues[45].

One limitation of our study is that the main findings are based on data from blood which was the only accessible tissue in our epidemiological studies and may not be representative of more disease-relevant tissues. However, the concordance of differential methylation between blood and adipose is high for certain pathways[46]. DNA methylation globally is considered a relatively stable epigenetic mark that can be inherited through multiple cell divisions[47, 48]. However, some changes can be dynamic reflected by recent environmental exposures. This phenomenon could be site-specific. While our study provides a snapshot of associations specific to the fasting state, instant methylation of different CpG sites in the vicinity of *IRS2* and *KDM2B* have been reported earlier[49]. Such effects may also occur at the loci presented in the present study. Our present MR analyses yield no evidence for the causal effects of CpG sites on fasting glucose or insulin. One limitation in the interpretation of the findings is that low power of the MR due to the fact we lack insight in the genes driving differential methylation. For instance, seven of the 13 performed CpG sites have instrumental variables which explain less than 5% of the exposure. Further studies are needed to include additional biologically relevant tissues and perform MR based on the tissue-specific meQLTs. Last but not least, cg19693031 in *TXNIP* has been repeatedly associated with type 2 diabetes case-control status earlier[3, 50, 51]. Although it did not pass our predefined EWAS significance threshold, *TXNIP* is associated with fasting glucose in the non-diabetic population (P-value = 7.6×10^{-7} in the BMI adjustment model) if we take the current study aiming to replicate earlier findings. Of note is that cg19693031 is not associated with fasting insulin (in BMI unadjusted model, p-value = 0.30; in the BMI adjusted model, p-value = 0.37).

In conclusion, our large-scale EWAS and replication identifies nine differentially methylated sites associated with fasting glucose or insulin, and shows that differential methylation explains part of the association between obesity and insulin metabolism. The integrative *in silico* cross-omics analyses provide insights of glycemic loci into the genetics, epigenetics and transcriptomics pathways. We also highlight that differential methylation is a key point in the involvement of the adaptive immune system in glucose homeostasis. Further studies in the future will benefit from tissue-specific methylation and meQTL databases which are currently the missing piece of the *in silico* data integration framework.

Methods

Study population

The discovery samples consisted of 4,808 European individuals without diabetes from four non-overlapped cohorts, recruited by Rotterdam Study III-1 (RS III-1, n = 626), Rotterdam Study II-3 and Rotterdam Study III-2 (called as RS-BIOS, n = 705), Netherlands Twin Register (NTR, n = 2,753) and UK adult Twin registry (TwinsUK, n = 724). The replication sets contained up to 11,750 individuals from 11 independent cohorts from CHARGE, including up to 6,818 individuals from European ancestry, 4,355 from African ancestry and 577 from Hispanic ancestry (Supplementary Data 1). They are from Atherosclerosis Risk in Communities (ARIC) Study, Baltimore Longitudinal Study of Aging (BLSA), Cardiovascular Health Study (CHS), Framingham Heart Study Cohort (FHS), The Genetic Epidemiology Network of Arteriopathy (GENOA), Genetics of Lipid Lowering Drugs and Diet Network (GOLDN), Hypertension Genetic Epidemiology Network (HyperGEN), Invecchiare in Chianti Study (InCHIANTI), Kooperative Gesundheitsforschung in der Region Augsburg (KORA), Women's Health Initiative - Broad Agency Award 23 (WHI-BAA23) and Women's Health Initiative - Epigenetic Mechanisms of PM-Mediated CVD (WHI-EMPC). We excluded individuals with known diabetes and/or fasting glucose ≥ 7mmol/l and/or those on anti-diabetic treatment. All studies were approved by their respective Institutional Review Boards, and all participants provided written informed consent. Details about the studies have been reported previously, and the key references as well as the summary of the design of each study are reported in Supplementary Note 1.

Glycemic traits and covariates

Venous blood samples were obtained after an overnight fast in all discovery and replication cohorts. BMI was calculated as weight over height squared (kg m^{-2}) based on clinical examinations. Smoking status was divided into current, former and never, based on questionnaires. White blood cell counts were quantified using standard laboratory techniques or predicted from methylation data using the Houseman method[52]. The cohort specific measurement of glycemic traits and covariates are shown inSupplementary Note 1.

DNA methylation quantification

The Illumina© Human Methylation450 array was used in all discovery and replication cohorts to quantify genome-wide DNA methylation in blood samples. We obtained DNA methylation levels reported as β values, which represents the cellular average methylation level ranging from 0 (fully unmethylated) to 1 (fully methylated). Study-specific details regarding DNA methylation quantification, normalization and quality control procedures are provided in the Supplementary Note 1.

Epigenome-wide association analysis and replication

All statistical analyses were performed using *R* statistical software and the two-tailed test was considered. Insulin was natural log transformed. In the discovery analysis, we first performed EWAS in each cohort separately. Linear regression analysis was used to test the association between glucose and insulin with each CpG site in the Rotterdam Study samples. Linear mixed models were used in NTR and TwinsUK accounting for the family structure. We fitted two models for each cohort: 1) the baseline model adjusting for age, sex, technical covariates (chip array number and position on the array), white blood cell counts (lymphocytes, monocytes, and granulocytes) and smoking status, and 2) a second model additionally adjusting for BMI. We removed probes that have evidence of multiple mapping or contain a genetic variant in the CpG site[53]. All cohort-specific EWAS results for each model were then meta-analysed using inverse variance-weighted fixed effect meta-analysis as implemented in the metafor R package[54]. In total, we meta-analysed 393,183 CpG sites that passed quality control in all four discovery cohorts. The details of the quality control for each cohort could be found in the Supplementary Note 1. The association was later corrected by the genomic control factor (λ) in each meta-EWAS[55]. We produced quantile-quantile (QQ) plots of the $-\log_{10}$ (P) to evaluate inflation in the test statistic (Supplementary Figure 2). A Bonferroni correction was used to correct for multiple testing and identify epigenome-wide significant results ($P < 1.3 \times 10^{-7}$). We did not correct the number of glycemic traits and models, as they are highly correlated and not independent. The genome coordinates were provided by Illumina (GRCh37/hg19). The CpG sites were annotated to genes using Infinium HumanMethylation 450 BeadChip annotation resources. The correlation of the CpG sites located in the same gene was further checked in the overall RS III-1 and RS-BIOS samples by Pearson's correlation test (n = 1,544) to find the independent top CpG sites.

For the associations discovered in the meta-EWAS that have not been reported previously, we attempted replication in independent samples using the same traits and regression models as in the discovery analyses. Study-specific details of replication cohorts are provided in Supplementary Data 1 and Supplementary Note 1. Results from each replication cohort were meta-analysed using the same methods as in the discovery analyses. Bonferroni P-value $< 3.3 \times 10^{-3}$ (0.05 corrected by 15 loci tested for associations) was considered significant.

Glycemic differential DNA methylation and transcriptomics

To explore whether the differential CpG sites were associated with gene expression level in blood, we explored eQTMs[17] from the European blood-based BIOS database[17] from BBMRI-NL which captured meQTLs, eQTLs and eQTMs from genome-wide database of 3,841 Dutch blood samples (See resources of the database in URLs). The associated gene expression probes of the known and replicated CpG sites were searched. We then tested whether the expression of the genes that harbor the identified methylation sites was associated with T2D and related traits in glucose metabolism-related tissues (adipose subcutaneous, adipose visceral omentum, liver, whole blood, pancreas, skeletal muscle and small intestine terminal ileum) using *MetaXcan* package[24, 56]. MetaXcan associates the expression of the genes with the phenotype by integrating functional data generated by large-scale efforts, e.g. GTEx project[16] with that of the GWAS of the trait. MetaXcan is trained on transcriptome models in 44 human tissues from GTEx and is able to estimate their tissue-specific effect on phenotypes from GWAS. For this study we used the GWAS studies of T2D[19], fasting glucose traits[21, 22], fasting insulin[22], HbA1c[23] and HOMA-IR[20]. We used the nominal P-value threshold (P-value threshold for significance < 0.05) as we had separate assumptions for each terminal pathway between gene expressions and phenotype. The associations with genes in low prediction performance were excluded, i.e. the association of the tissue model's correlation to the gene's measured transcriptome is not significant (P-value > 0.05).

Glycemic differential DNA methylation and genomics

We identified the genetic determinants of the significant CpG sites known or replicated through the current EWAS using the results of the *cis* and *trans* meQTLs from the European blood-based BIOS database[17] (See resources of the database in URLs). All the reported SNPs with P-value adjusted for false discovery rate (FDR) less than 0.05 in the database were treated as the target genetic variants in the present study. The SNPs were annotated based on the information in the BIOS study[17] or the nearest protein-coding gene list from SNPnexus[57] on GRCh37/hg19. We also explored the associations of these DNA methylation-related genetic variants with T2D or related traits, i.e. fasting glucose, insulin, HbA1c and HOMA-IR, based on public GWAS datasets in European ancestry[20, 21, 22, 27, 28, 29]. Meanwhile, we checked the effect direction consistency of the association between the SNPs, CpG sites and T2D or related traits. That is the direction of the association between SNP and T2D or related traits should be a combination of the direction of SNP with CpG sites and CpG sites with T2D or related traits. A multiple-testing correction was performed by Bonferroni adjustment (P-value significant threshold < 1.8×10^{-3}, 0.05 corrected by the 29 genetic loci shown in Supplementary Data 3). The associations of the DNA methylation-related genetic variants and the gene expression were also looked up in the BIOS database[17]. This is limited to the expression profiles earlier associated with glycemic CpG sites in blood.

For the significant CpG sites known or replicated through EWAS, we attempted to evaluate the causality effect of CpG sites on their significant traits, either fasting glucose or fasting insulin, using two-sample MR approach as described in detail before by Dastani *et al*[30, 58] based on the summary statistic GWAS results from the BIOS database and the Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) database[17, 21] (Supplementary Figure 3). Briefly, we constructed a weighted genetic risk score for individual CpG on phenotype using independent SNPs as the instrument variables of the CpG, implemented in the R-package gtx. The effect of each score on phenotype was calculated as

$$ahat = \frac{\sum (\omega_i \beta_i / s_i^2)}{\sum (\omega_i^2 / s_i^2)}$$

, where β_i is the effect of the CpG-increasing alleles on phenotype, s_i its corresponding standard error and ω_i the SNP effect on the respective CpG. Because the genetic variants might be close (*cis*) or far (*trans*) from the methylated site, we also performed MR test in the *cis* only SNPs if the CpG has both *cis* and *trans* genetic markers. All SNPs were mapped to the human genome build hg19. For each test (one CpG with one trait), we extracted all the genetic markers of the CpG in the fasting glucose or insulin GWAS from the MAGIC dataset (n = 96,496)[21] with their effect estimate and standard error on fasting glucose or insulin. Within the overlapped SNPs, we removed SNPs in potential linkage disequilibrium (LD, pairwise $R^2 \geq 0.05$) in 1-Mbp window based on the 1000 Genome imputed genotype dataset from the general population: Rotterdam Study I (RS I, n = 6,291)[59]. We managed to exclude the genetic loci which were genome-wide associated with glycemic traits, but none of the genetic loci meet this exclusion criterion. The instrumental variables that explain more than 1% of the variance in exposure (DNA methylation) were taken forward for MR test. The Bonferroni P-value threshold was used to correct for the 13 CpG sites available for MR (P-value < 3.8×10^{-3}).

Functional annotation

Further, we integrated the cascade of associations as above among the results of EWAS, eQTM, meQTL and eQTL and showed in Figure 3. We checked the effect direction consistency of the association between the SNPs, CpG sites, gene expression in blood and glycemic traits. The correlation of the novel CpG sites was checked in the overall RS III-1 and RS-BIOS samples by Pearson's correlation test (n = 1,544). The hierarchical cluster analysis was used in the clustering. Gene set enrichment analyses were performed in the genes of new CpG sites[60]. We tested if genes of interest were overrepresented in any of the pre-defined gene sets from KEGG pathway database[32], Reactome Pathway Knowledgebase[33] and GO biological process[34]. Multiple test correction was performed in the tests. Gene sets of KEGG pathway database, Reactome Pathway Knowledgebase were obtained from Molecular Signatures Database (MsigDB) c2 and GO biological process was obtained from MsigDB c5[60]. We used the platform of Functional Mapping and Annotation of Genome-wide Association Studies (FUMA GWAS)[61] and GENE2FUNC function to perform the gene set enrichment analysis and the tissue-specific gene expression patterns based on GTEx v6[16]. Besides, the tools Ensembl Human Genes[62] (see URLs) and UCSC GRCh37/hg19[63] (see URLs) were also used in interpreting genetic determinants, CpG sites and genes.

BMI in the association of methylation and glycemic traits

We used linear regression to check the effect of CpGs on the relationship between BMI and fasting insulin in the non-diabetic individuals in Rotterdam study. The initial model used BMI as the independent variable and the natural log transformed insulin as the dependent variable. The covariates included age, sex, technical covariates (chip array number and position on the array), white blood cell counts, smoking status and dataset (RS III-1 and RS-BIOS). The normalized differential methylation values of CpG sites were added as covariates in the advanced model. The differences of the models were compared by ANOVA testing using anova function in *R* (P-value < 0.05).

URLs. BIOS database, https://genenetwork.nl/biosqtlbrowser/; SNPnexus, http://snp-nexus.org/index.html ; GWAS database of glycemic traits, https://www.magicinvestigators.org/ ; GWAS database of T2D, http://diagram-consortium.org/ ; MetaXcan, https://s3.amazonaws.com/imlab-open/Data/MetaXcan/results/metaxcan_results_database_v0.1.tar.gz; NHGRI-EBI Catalog, https://www.ebi.ac.uk/gwas/ ; Ensembl, https://www.ensembl.org/Homo_sapiens/Info/Index; FUMA, http://fuma.ctglab.nl; UCSC, https://genome.ucsc.edu/cgi-bin/hgGateway (available: 1st Jan 2019)

References

1. American Diabetes A. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2018. *Diabetes care* 41, S13-S27 (2018).
2. Hidalgo B*, et al.* Epigenome-wide association study of fasting measures of glucose, insulin, and HOMA-IR in the Genetics of Lipid Lowering Drugs and Diet Network study. *Diabetes* 63, 801-807 (2014).
3. Chambers JC*, et al.* Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *Lancet Diabetes Endocrinol* 3, 526-534 (2015).
4. Kriebel J*, et al.* Association between DNA Methylation in Whole Blood and Measures of Glucose Metabolism: KORA F4 Study. *PLoS One* 11, e0152314 (2016).
5. Kulkarni H*, et al.* Novel epigenetic determinants of type 2 diabetes in Mexican-American families. *Hum Mol Genet* 24, 5330-5344 (2015).
6. Ding J*, et al.* Alterations of a Cellular Cholesterol Metabolism Network Are a Molecular Feature of Obesity-Related Type 2 Diabetes and Cardiovascular Disease. *Diabetes* 64, 3464-3474 (2015).
7. Ehrlich M, Lacey M. DNA methylation and differentiation: silencing, upregulation and modulation of gene expression. *Epigenomics* 5, 553-568 (2013).
8. Wahl S*, et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* 541, 81-86 (2017).
9. Chen R*, et al.* Longitudinal personal DNA methylome dynamics in a human with a chronic condition. *Nat Med*, (2018).
10. Al Muftah WA*, et al.* Epigenetic associations of type 2 diabetes and BMI in an Arab population. *Clin Epigenetics* 8, 13 (2016).
11. Walaszczyk E*, et al.* DNA methylation markers associated with type 2 diabetes, fasting glucose and HbA1c levels: a systematic review and replication in a case-control sample of the Lifelines study. *Diabetologia* 61, 354-368 (2018).

12. Demerath EW*, et al.* Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Hum Mol Genet* 24, 4464-4479 (2015).
13. Aslibekyan S*, et al.* Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference. *Obesity (Silver Spring)* 23, 1493-1501 (2015).
14. Wang B*, et al.* Methylation loci associated with body mass index, waist circumference, and waist-to-hip ratio in Chinese adults: an epigenome-wide analysis. *Lancet* 388 Suppl 1, S21 (2016).
15. Wilson LE, Harlid S, Xu Z, Sandler DP, Taylor JA. An epigenome-wide study of body mass index and DNA methylation in blood using participants from the Sister Study cohort. *Int J Obes (Lond)* 41, 194-199 (2017).
16. Consortium GT*, et al.* Genetic effects on gene expression across human tissues. *Nature* 550, 204-213 (2017).
17. Bonder MJ*, et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet* 49, 131-138 (2017).
18. Jones MJ, Fejes AP, Kobor MS. DNA methylation, genotype and gene expression: who is driving and who is along for the ride? *Genome Biol* 14, 126 (2013).
19. Replication DIG*, et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* 46, 234-244 (2014).
20. Dupuis J*, et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 42, 105-116 (2010).
21. Manning AK*, et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* 44, 659-669 (2012).
22. Scott RA*, et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat Genet* 44, 991-1005 (2012).
23. Soranzo N*, et al.* Common variants at 10 genomic loci influence hemoglobin A1C levels via glycemic and nonglycemic pathways. *Diabetes* 59, 3229-3239 (2010).
24. Barbeira AN*, et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* 9, 1825 (2018).
25. Grarup N*, et al.* Association of variants in the sterol regulatory element-binding factor 1 (SREBF1) gene with type 2 diabetes, glycemia, and insulin resistance: a study of 15,734 Danish subjects. *Diabetes* 57, 1136-1142 (2008).
26. Jin B, Li Y, Robertson KD. DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes Cancer* 2, 607-617 (2011).
27. Wheeler E*, et al.* Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med* 14, e1002383 (2017).
28. Mahajan A*, et al.* Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat Genet* 50, 559-571 (2018).
29. Scott RA*, et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* 66, 2888-2902 (2017).
30. Dastani Z*, et al.* Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet* 8, e1002607 (2012).
31. Mendelson MM*, et al.* Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. *PLoS Med* 14, e1002215 (2017).
32. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 45, D353-D361 (2017).
33. Fabregat A*, et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* 46, D649-D655 (2018).
34. The Gene Ontology C. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* 45, D331-D338 (2017).
35. Pickup JC, Crook MA. Is type II diabetes mellitus a disease of the innate immune system? *Diabetologia* 41, 1241-1248 (1998).
36. M.D. GIB. National Library of Medicine (US), National Center for Biotechnology Information (2004).
37. Di Bernardo MC*, et al.* Risk of developing chronic lymphocytic leukemia is influenced by HLA-A class I variation. *Leukemia* 27, 255-258 (2013).
38. Thomson G, Bodmer WF. The genetics of HLA and disease associations. In: *Measuring selection in natural populations.* Springer (1977).

39. Withers DJ*, et al.* Disruption of IRS-2 causes type 2 diabetes in mice. *Nature* 391, 900-904 (1998).

40. Lingohr MK*, et al.* Decreasing IRS-2 expression in pancreatic beta-cells (INS-1) promotes apoptosis, which can be compensated for by introduction of IRS-4 expression. *Mol Cell Endocrinol* 209, 17-31 (2003).

41. Hennige AM*, et al.* Upregulation of insulin receptor substrate-2 in pancreatic beta cells prevents diabetes. *J Clin Invest* 112, 1521-1532 (2003).

42. Akama TO*, et al.* Germ cell survival through carbohydrate-mediated interaction with Sertoli cells. *Science* 295, 124-127 (2002).

43. Guo W*, et al.* RBM20, a gene for hereditary cardiomyopathy, regulates titin splicing. *Nat Med* 18, 766-773 (2012).

44. Bycroft C*, et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203-209 (2018).

45. Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45, 580-585 (2013).

46. Huang YT*, et al.* Epigenome-wide profiling of DNA methylation in paired samples of adipose tissue and blood. *Epigenetics* 11, 227-236 (2016).

47. Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* 16, 6-21 (2002).

48. Kim M, Costello J. DNA methylation: an epigenetic mark of cellular memory. *Exp Mol Med* 49, e322 (2017).

49. Rask-Andersen M*, et al.* Postprandial alterations in whole-blood DNA methylation are mediated by changes in white blood cell composition. *Am J Clin Nutr* 104, 518-525 (2016).

50. Soriano-Tarraga C*, et al.* Epigenome-wide association study identifies TXNIP gene associated with type 2 diabetes mellitus and sustained hyperglycemia. *Hum Mol Genet* 25, 609-619 (2016).

51. Florath I*, et al.* Type 2 diabetes and leucocyte DNA methylation: an epigenome-wide association study in over 1,500 older adults. *Diabetologia* 59, 130-138 (2016).

52. Houseman EA*, et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 13, 86 (2012).

53. Chen YA*, et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 8, 203-209 (2013).

54. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 36, 1-48 (2010).

55. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 55, 997-1004 (1999).

56. Gamazon ER*, et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 47, 1091-1098 (2015).

57. Dayem Ullah AZ, Oscanoa J, Wang J, Nagano A, Lemoine NR, Chelala C. SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Res* 46, W109-W113 (2018).

58. Liu J*, et al.* A Mendelian Randomization Study of Metabolite Profiles, Fasting Glucose, and Type 2 Diabetes. *Diabetes* 66, 2915-2926 (2017).

59. Ikram MA*, et al.* The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol* 32, 807-850 (2017).

60. Subramanian A*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102, 15545-15550 (2005).

61. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 8, 1826 (2017).

62. Zerbino DR*, et al.* Ensembl 2018. *Nucleic Acids Res* 46, D754-D761 (2018).

63. Haeussler M*, et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* 47, D853-D858 (2019).

Chapter 3

Metabolomics in type 2 diabetes

Chapter 3.1

Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study

Jun Liu, Sabina Semiz, Sven J. van der Lee, Ashley van der Spek, Aswin Verhoeven, Jan B. van Klinken, Eric Sijbrands, Amy C. Harms, Thomas Hankemeier, Ko Willems van Dijk, Cornelia M. van Duijn, Ayşe Demirkan

This chapter was published in Metabolomics (2017) 13:104.

The supplemental information for this paper is available online at https://doi.org/10.1007/s11306-017-1239-2

Abstract

Background The growing field of metabolomics has opened up new opportunities for prediction of type 2 diabetes (T2D) going beyond the classical biochemistry assays.

Objectives We aimed to identify markers from different pathways which represent early metabolic changes and test their predictive performance for T2D, as compared to the performance of traditional risk factors (TRF).

Methods We analyzed 2,776 participants from the Erasmus Rucphen Family study from which 1,571 disease free individuals were followed up to 14-years. The targeted metabolomics measurements at baseline were performed by three different platforms using either nuclear magnetic resonance spectroscopy or mass spectrometry. We selected 24 T2D markers by using Least Absolute Shrinkage and Selection operator (LASSO) regression and tested their association to incidence of disease during follow-up.

Results The 24 markers i.e. high-density, low-density and very low-density lipoprotein sub-fractions, certain triglycerides, amino acids, and small intermediate compounds predicted future T2D with an area under the curve (AUC) of 0.81. The performance of the metabolic markers compared to glucose was significantly higher among the young (age < 50 years) (0.86 *vs* 0.77, p-value < 0.0001), the female (0.88 *vs* 0.84, p-value = 0.009), and the lean (BMI < 25 kg/m^2) (0.85 *vs* 0.80, p-value = 0.003). The full model with fasting glucose, TRFs, and metabolic markers yielded the best prediction model (AUC = 0.89).

Conclusions Our novel prediction model increases the long-term prediction performance in combination with classical measurements, brings a higher resolution over the complexity of the lipoprotein component, increasing the specificity for individuals in the low risk group.

Introduction

Early lifestyle intervention is a cost-effective recommendation to reduce the incidence of type 2 diabetes[1, 2, 3], asking for informative, sensitive and specific markers. Although the standard laboratory tests, such as fasting glucose, 2-hour postprandial glucose, and glycated hemoglobin A1c (HbA1c), provide strong evidence for the risk of type 2 diabetes[4, 5, 6], these predictors emerge after years of subclinical metabolic dysfunction[7]. Traditional risk factors (TRFs) such as age, sex, body mass index (BMI), and waist circumference also explain considerable part of future risk[8, 9], but fail to capture the full complexity of the etiology and their predictive performance vary between different risk groups[10]. BMI has been put forward as the modifiable risk factor but, there are also metabolically unhealthy normal weight (MUHNW) as well as metabolically healthy obese (MHO) individuals, raising the question to what extent BMI explain the mechanisms of the underlying metabolic disease[11]. Therefore, there is an increasing interest in finding informative markers that indicate the particular metabolic dysfunctions before the manifestation of the disease. Hence, people identified at high risk would be able to take preventive lifestyle interventions or treatments targeted to their individual molecular profile, eventually personalizing their health care.

High throughput metabolomics offers an opportunity to test multiple metabolic markers in large settings. Such approach led to the discovery of five amino acids by the prospective Framingham Heart Study (FHS) using a 12-year follow-up[12]. Branched chain amino acids (BCAA) from this panel were previously pointed out in a case-control setting [13] and later in a follow-up study of limited size [14, 15]. Other metabolites including phospholipids, triglycerides, acyl-carnitines, organic acids and small molecular weight compounds were also added to the list of metabolomics based predictors [13, 14, 15, 16, 17, 18], covering the glucose and phospholipid metabolism. However, lipoprotein metabolism, which is one of the key components of metabolic dysfunction, has not been addressed.

In the present study, we aimed to identify novel metabolic markers using a total of 261 metabolic features measured by either targeted mass spectrometry (MS) or by targeted nuclear magnetic resonance (NMR). The chemical classes of tested molecules include sub-fractions of lipoproteins, triglycerides, phospholipids, amino acids, and small intermediate compounds. We estimated the predictive performance of the selected marker set in comparison to other well-known predictors, including fasting glucose, TRFs, and the validated panel of amino acids.

Research Design and Methods

Study population

The Erasmus Rucphen Family genetic isolate study (ERF) is a prospective family based study located in Southwest of the Netherlands. This young genetic isolate was founded in the mid-eighteenth century and minimal immigration and marriages occurred between surrounding settlements due to social and religious reasons. The ERF study population includes 3,465 individuals that are living descendants of 22 couples with at least six children baptized. Informed consent has been obtained from patients where appropriate. The study protocol was approved by the medical ethics board of the Erasmus Medical Center Rotterdam, the Netherlands[19].

The baseline demographic data and measurements of the ERF participants were collected around 2002 to 2006. All the participants filled out questionnaires on socio-demographics, diseases and medical history and lifestyle factors, and were invited to the research center for an interview and blood collection for biochemistry and physical examinations including blood pressure and anthropometric measurements have been performed. The participants were asked to bring all their current medications for registration during the interview. Venous blood samples were collected after at least eight hours fasting. Hypertension was defined as systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg or treatment for hypertension. The family history was coded as 0, 1, 2 based on no first-degree relatives has type 2 diabetes, one has type 2 diabetes and more than one have type 2 diabetes. Baseline type 2 diabetes was defined according to the fasting plasma glucose ≥ 7.0 mmol/L and/or anti-diabetic treatment, yielding 212 cases and 2,564 controls, totaling up to 2,776. The follow-up data collection of the ERF study took place from March 2015 to May 2016 (9 to 14 years after baseline visit). During the follow up a total of 1,935 participants' records were scanned for incidence of type 2 diabetes in general practitioner’s databases. Additionally, a questionnaire on type 2 diabetes medication surveyed on 1,232 participants in June 2010 (4 to 8 years after baseline visit) was referred if a participant were not included in May 2016 follow-up. This effort yielded the inclusion of 18 otherwise missed extra cases. To summarize, out of the 2,564 controls at baseline, 1,571 were followed-up for a mean 11.3 years (inter quartile range: 11.0 - 12.2). Among those, 137 developed type 2 diabetes, whereas 1,434 did not, comprising together the analytical sample for prediction analysis.

Metabolomics measurements

In total 261 metabolic marker molecules including sub-fractions of lipoproteins, triglycerides, phospholipids, amino acids and small intermediate compounds, which throughout this article will be referred as “*metabolites*”, were measured by three different targeted platforms, either by NMR spectrometry or MS at baseline. The samples included in metabolomics measurements were not selected based on any disease. The platforms used in this research are: (1) Liquid Chromatography-MS (LC-MS, 116 positively charged lipids, comprising of 39 triglycerides (TG), 47 phosphatidylcholines (PC), 8 phosphatidylethanolamines (PE), 20 sphingolipids (SM), and 2 ceramides (Cer), available in up to 2,638 participants) measured in Netherlands Metabolomics Center, Leiden using the method described before[20], (2) small molecular compounds window based NMR spectroscopy (NMR-COMP, 41 molecules comprising of 29 low-molecular weight molecules and 12 amino acids available in up to 2,639 participants) measured in Center for Proteomics and Metabolomics, Leiden University Medical Center[21, 22], (3) lipoprotein window based NMR spectroscopy (NMR-LIPO, 104 lipoprotein particles size sub-fractions comprising of 28 very low-density lipoprotein (VLDL) components, 30 high-density lipoprotein (HDL) components, 35 low-density lipoprotein (LDL) components, 5 IDL components and 6 plasma totals, available in up to 2,609 participants) measured in Proteomics and Metabolomics, Leiden University Medical Center and lipoprotein sub-fraction concentrations were determined by the Bruker algorithm (Bruker BioSpin GmbH, Germany) details were given previously[23]. Details over the quality control of samples in these platforms can be found in the Supplementary Information. The laboratories had no access to phenotype information and the data pre-filtering and quality control for measurement errors were based on internal controls and duplicates.

Metabolite identification

The compounds measured by LC-MS and NMR-COMP were identified according to the metabolomics standards initiative (MSI) level 1 using information coming from at least 2 different sources [24]. The available ChEBI ID were shown in Supplementary Table 1.

For metabolites measured by LC-MS, the identities of the lipids were assigned on the basis of accurate mass, fragmentation pattern, and retention times matched to authentic standards where available. The detail of the metabolite identification can be found in previous publications[25].

For metabolites measured by NMR-COMP, the identities of the small components and molecules were assigned by the peaks which are annotated using the combined information from chemical shift databases, spiking experiments, and correlation behaviors. The detail of the metabolite identification can be found in the methodological paper[22].

For lipoproteins measured by NMR-LIPO, the method is based on the analysis of signals in the ^{1}H-NMR spectrum which are related to the lipoproteins. Differences in lipoprotein composition, size and density translate into respective signal line shape differences, which can be used to extract information on lipoprotein main- and subclasses. As these are not real metabolites, the MSI criteria do not apply.

Statistical methods

The distributions of individual metabolites were checked for non-normality by eye and outlying values that were more than four times standard deviation away from the mean were excluded from analysis. Non-normally distributed measurements were natural logarithm transformed, or rank transformed. Figure 1 shows the procedure that we followed for the selection of metabolites.

Firstly, we tested the association between the 261 individual metabolites and prevalent type 2 diabetes using a logistic regression model adjusting for age, sex, and lipid-lowering medication. Residuals from the polygenic model (using “polygenic” function in the R package *GenABEL*), were used in all analysis to account for family relations among the ERF participants[26]. To control for multiple testing, we applied a Bonferroni correction based on the effect number of independent vectors in the data which were estimated to be 81 independent equivalents using Matrix Spectral Decomposition (MSD)[27]. Thus, a p-value less than 6.18×10^{-4} (0.05/81) was used as the threshold for metabolome-wide significance. We repeated analysis stratifying the cases into medicated and non-medicated cases to test if the associations were attributed to the effect of anti-diabetic medication. Metabolites that did not differentiate (p-value > 0.05) between non-medicated diabetics (n = 68) *vs* controls (n = 2,564) were not taken forward. These metabolite levels were assumed to be different due to the post medication metabolic changes in the diabetics. The remaining metabolites (n = 88, the list is given in Supplementary Table 1) and the TRFs (age, sex, family history, BMI, waist circumference, hypertension, HDL-cholesterol, and triglycerides) with scaled around 0 and standard deviation as 1 were included in the prior to LASSO (Least Absolute Shrinkage and Selection Operator) regression to select the set of predictors that maximize the prediction performance. The LASSO regression was performed using *glmnet* package in R[28]. We imputed these missing data points (i.e. 9.6% ~ 18.5% missing values) before selecting the independent predictors by LASSO regression which requires all the variables to be complete measurements. In order to select the best imputation method suitable for our data, we first generated a training dataset with 20% missing values at random and compared three methods: (1) deterministic imputation, (2) random regression imputation, (3) multiple imputation with R package "mice"[29]. After comparing the results to the initial correlation with glucose and the means between the imputed values and real values for each method, multiple imputation was selected. The sum of predicted values from the multiple random regression model divided by the number of imputations (n = 20) was used to replace the missing data. The outliers more or less than four times standard deviation were removed after imputation. With the selected independent type 2 diabetes metabolic predictors, we assessed their associations with fasting glucose by linear regression analysis in the non-diabetic participants at baseline. To account for multiple testing in these 24 linear regression sets, a p-value less than 0.003 (0.05/16) was used as the threshold after MSD of the 24 metabolites that yielded 16 independent components.

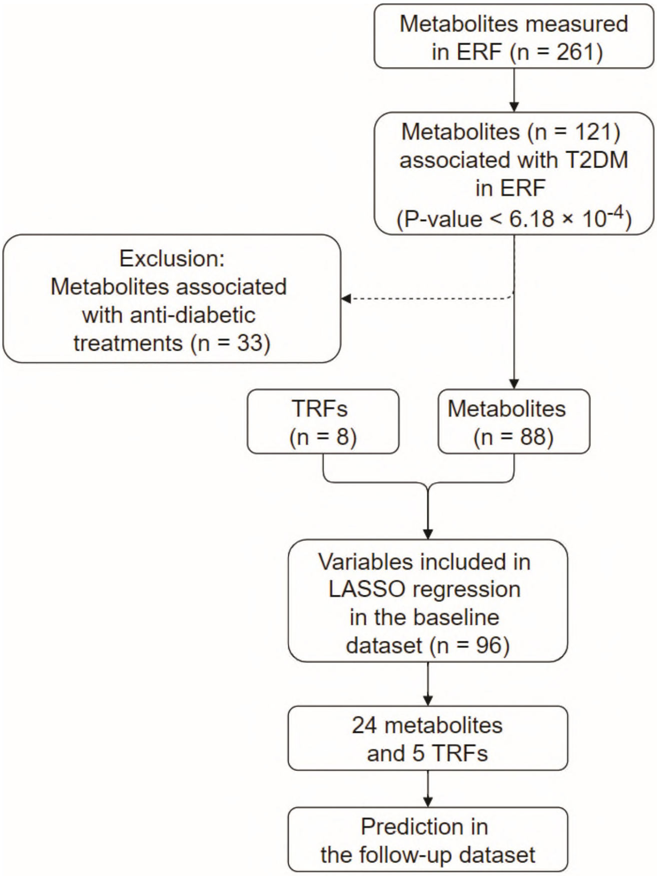

Metabolites measured in ERF (n = 261)
Metabolites (n = 121) associated with T2DM in ERF (P-value < 6.18 × 10-4)
Exclusion: Metabolites associated with anti-diabetic treatments (n = 33)
TRFs (n = 8)
Metabolites (n = 88)
Variables included in LASSO regression in the baseline dataset (n = 96)
24 metabolites and 5 TRFs
Prediction in the follow-up dataset

Figure 1 Flow chart of the metabolite selection.

Prediction of incident type 2 diabetes

The metabolites selected from the baseline population were tested to predict the incidence of type 2 diabetes during the follow-up time. Area under the receiver operator characteristics (ROC) curve (AUC) of logistic regression together with continuous Net Reclassification Improvement (NRI) was performed to estimate the discrimination and reclassification in different models[30]. The models compared were: *ERF metabolite model* (the metabolites those selected in the current study only), *FHS metabolite model,* (the amino acids reported by the FHS research: isoleucine, leucine, valine, tyrosine, and phenylalanine), and the *TRF model* (age, sex, family history, BMI, waist circumference, hypertension, HDL-cholesterol, and triglycerides) and glucose only model (fasting plasma glucose measured at baseline) and combination of those. As some of the previous studies showed the association between metabolites and covariates, i.e. age, sex and BMI[31, 32], we also tested the models in subgroups of sex, age (< 50 y vs ≥ 50 y), and BMI (< 25 kg/m^2 *vs* ≥ 25 kg/m^2). A p-value < 0.05 here was used as a cut off for significance improvement across the models. Meanwhile, the specificity with fixed 80% sensitivity in different prediction models is compared. Analyses were conducted using *R* (version 3.2.3).

Results

Table 1 displays the baseline characteristics of the participants stratified by prevalent cases at baseline and incident cases in the follow-up. Compared to the participants who did not develop type 2 diabetes, those with type 2 diabetes were older, more often had a family history of the disease, suffered from hypertension, and have been using lipid-lowering medication. They had higher levels of BMI, waist circumference, blood pressure, triglycerides, fasting glucose, and lower levels of HDL-cholesterol. The participants with incident type 2 diabetes had higher fasting glucose at baseline compared to the individuals who did not develop type 2 diabetes during the follow up.

Metabolites associated with type 2 diabetes at baseline

We identified 24 independent metabolites together with five TRFs (age, sex, family history, waist circumference, and HDL-cholesterol) from LASSO regression maximizing the discrimination at baseline. These metabolites and their associations with prevalent and incident type 2 diabetes, as well as fasting glucose at baseline are listed in Table 2. Four of them (i.e. PC(O-34:2), L-HDL-free cholesterol, XXL-LDL-phospholipids and L-LDL-cholesterol) are associated with decreased risk of type 2 diabetes, whereas twenty of them associated with increased risk; including three triglycerides, seven lipoprotein particles, three amino acids, and seven small intermediate compounds. Among the seven lipoprotein particles, two are sub-fractions of HDL, two are of LDL, and three are of VLDL (See details in Table 2). Out of the 24 metabolites, PC(O-34:2), XXL-LDL-triglycerides, HDL-triglycerides, L-HDL-ApoA2, and M-HDL-ApoA2 are not associated with fasting glucose in the non-diabetic population at baseline and incident type 2 diabetes.

Table 1 Characteristics of the study population.

	Baseline (n=2,776)		Follow-up (n=1,571)	
	Controls	Cases	Controls	Cases
	(n = 2,564)	(n = 212)	(n = 1,434)	(n = 137)
Male [n (%)]	1132 (44.1)	108 (50.9)	595 (41.5)	78 (56.9) *
Age (years)	48.2 ± 14.3	59.8 ± 11.8*	47.7 ± 13.9	57 ± 10.7*
Diabetes in first-degree relatives				
0 individuals [n (%)]	1711 (76.6)	71 (55.0)	966 (76.4)	63 (53.8)
1 individual [n (%)]	428 (19.2)	37 (28.7)	248 (19.6)	38 (32.5)
≥ 2 individuals [n (%)]	95 (4.3)	21 (16.3) *	50 (4.0)	16 (13.7) *
Body mass index (kg/m^2)	26.7 ± 4.6	30.0 ± 5.9*	26.6 ± 4.4	30.1 ± 5.1*
Waist circumference (cm)	86.7 ± 13.1	99.3 ± 14.2*	86.2 ± 12.8	98.9 ± 13.4*
Systolic blood pressure (mmHg)	139 ± 20	154 ± 21*	137.7 ± 19.6	152.4 ± 21.8*
Diastolic blood pressure (mmHg)	80.3 ± 10.0	82.9 ± 9.9	79.7 ± 9.6	84.8 ± 9.8*
Hypertension [n (%)]	1282 (50)	170 (80.2) *	674 (47.0)	111 (81.0) *
HDL-cholesterol (mmol/l)	1.3 ± 0.4	1.1 ± 0.3*	1.3 ± 0.4	1.1 ± 0.3*
Triglycerides (mmol/l)	1.2 (0.8, 1.6)	1.6 (1.1, 1.9) *	1.2 (0.8, 1.6)	1.7 (1.1, 2.1) *
Fasting glucose (mmol/l)	4.5 ± 0.7	7.4 ± 2.2*	4.4 ± 0.6	5.3 ± 0.7*
Lipid-lowering medication [n (%)]	265 (10.3)	99 (46.7) *	136 (9.5)	42 (30.9) *

Data are means ± standard deviations (SD), medians (inter-quartile range), or n (%). Triglycerides were natural logarithm transformed prior to analysis. * P-value < 0.05 after adjusting age, sex and/or lipid-lowering medication.

Predicting incident type 2 diabetes

Figure 2 shows the AUC comparisons across the different prediction models. The ERF metabolites discriminate future type 2 diabetes with an AUC [95% confidence interval] of 0.81 [0.77, 0.85]. The AUC of the ERF metabolite model was significantly higher than of the FHS metabolite model (AUC 0.81 [0.77, 0.85] *vs* 0.77 [0.73, 0.81], NRI = 0.42, p-value < 0.0001). It is of note that tyrosine and isoleucine, which were previously selected by FHS, were also selected in the ERF metabolite model. The AUC for the model including both ERF and FHS metabolites together was significantly higher than the AUC for models with either set of predictors (AUC 0.83 [0.79, 0.86] *vs* 0.77 [0.73, 0.81], NRI = 0.67, p-value < 0.0001 for ERF and FHS metabolites *vs* only FHS metabolites; AUC 0.83 [0.79, 0.86] *vs* 0.81 [0.77, 0.85], NRI = 0.29, p-value = 0.0015 for ERF and FHS metabolites *vs* only ERF metabolites). The AUC of the ERF and FHS combined metabolite model did not differ from that of fasting glucose (AUC 0.83 [0.79, 0.86] *vs* 0.84 [0.81, 0.88], p-value = 0.45). However, combining the ERF metabolites and fasting glucose together in a model improved the predictive performance significantly over the performance of fasting glucose (AUC 0.88 [0.84, 0.91] *vs* 0.84 [0.81, 0.88], NRI = 0.66, p-value < 0.0001). Adding TRFs to fasting glucose and metabolite model maximized the AUC to 0.89 [0.86, 0.92]. The specificity with fixed 80% sensitivity increases from 70% to 80% when metabolites are added to the glucose only model (Supplementary Figure 1).

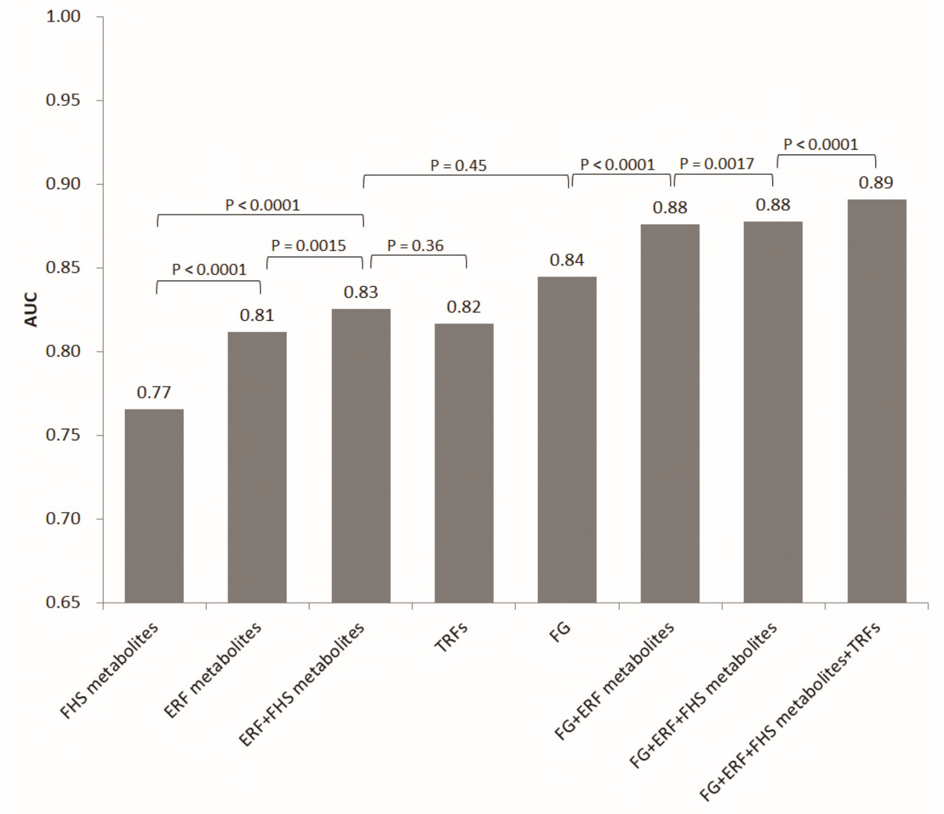


Figure 2 AUC comparisons in different prediction models. Continuous Net Reclassification Improvement (NRI) indices were performed to compare different prediction models. FG: fasting glucose. TRFs: all traditional risk factors: age, sex, family history, BMI, waist circumference, hypertension, HDL-cholesterol, triglycerides.

Table 2 Association of LASSO regression selected metabolites with type 2 diabetes and fasting glucose.

		Prevalent cases *vs* controls		Incident cases *vs* controls		Fasting glucose	
Metabolites	ChEBI ID	OR [95%CI]	P-value	OR [95%CI]	P-value	Effect	P-value
PC(O-34:2)	CHEBI:64544	0.6 [0.5, 0.7]	1.3×10^{-7}	0.9 [0.7, 1.1]	0.19	-0.01	0.28
Isoleucine	CHEBI:24898	2.4 [2.0, 2.9]	2.7×10^{-20}	2.0 [1.6, 2.5]	4.4×10^{-9}	0.09	3.6×10^{-8}
Methionine	CHEBI:16811	1.4 [1.2, 1.6]	1.2×10^{-4}	1.3 [1.1, 1.6]	7.4×10^{-3}	0.05	2.6×10^{-4}
Tyrosine	CHEBI:18186	1.5 [1.2, 1.7]	1.6×10^{-5}	2.0 [1.6, 2.5]	5.3×10^{-10}	0.13	6.0×10^{-18}
2-hydroxybutyrate	CHEBI:64552	2.0 [1.7, 2.5]	2.5×10^{-13}	2.0 [1.6, 2.6]	2.6×10^{-10}	0.15	2.8×10^{-27}
1,5-AG	CHEBI:16070	2.3 [1.9, 2.7]	5.0×10^{-19}	1.5 [1.2, 1.8]	3.3×10^{-4}	0.09	4.5×10^{-10}
2-oxoglutaric acid	CHEBI:30915	1.5 [1.3, 1.8]	2.70×10^{-6}	1.8 [1.4, 2.2]	6.0×10^{-7}	0.13	8.9×10^{-20}
Glycine betaine	CHEBI:17750	2.2 [1.8, 2.6]	2.50×10^{-17}	1.5 [1.2, 1.9]	2.3×10^{-4}	0.12	1.8×10^{-14}
Glycerol	CHEBI:17754	2.3 [1.8, 2.8]	2.1×10^{-14}	1.7 [1.3, 2.1]	1.5×10^{-5}	0.13	9.1×10^{-18}
Lactate	CHEBI:24996	1.7 [1.4, 1.9]	4.9×10^{-11}	1.5 [1.2, 1.7]	3.1×10^{-5}	0.11	3.1×10^{-15}
Pyruvate	CHEBI:15361	1.6 [1.4, 1.8]	3.0×10^{-9}	1.5 [1.3, 1.8]	3.3×10^{-6}	0.14	1.3×10^{-25}
TG (48:0)	CHEBI:85870	1.4 [1.2, 1.6]	2.3×10^{-5}	1.6 [1.3, 1.9]	9.3×10^{-7}	0.08	2.0×10^{-8}
TG (48:1)	CHEBI:85726	1.4 [1.2, 1.6]	1.3×10^{-4}	1.5 [1.3, 1.9]	8.0×10^{-6}	0.07	5.1×10^{-8}
TG (50:5)	CHEBI:90301	1.3 [1.1, 1.4]	2.3×10^{-3}	1.5 [1.2, 1.7]	6.5×10^{-6}	0.06	1.4×10^{-5}
VLDL-free cholesterol	-	1.4 [1.2, 1.7]	7.2×10^{-7}	1.6 [1.4, 1.9]	8.2×10^{-8}	0.08	1.7×10^{-9}
XXL-VLDL-cholesterol	-	1.3 [1.1, 1.5]	4.9×10^{-4}	1.5 [1.3, 1.7]	2.9×10^{-6}	0.08	1.1×10^{-8}
VLDL-triglycerides	-	1.4 [1.2, 1.6]	3.2×10^{-6}	1.5 [1.3, 1.8]	1.0×10^{-6}	0.09	3.3×10^{-10}
XXL-LDL-phospholipids	-	0.6 [0.5, 0.7]	4.4×10^{-9}	0.7 [0.6, 0.9]	2.9×10^{-3}	-0.06	6.4×10^{-5}
XXL-LDL-triglycerides	-	1.4 [1.2, 1.6]	2.4×10^{-4}	0.9 [0.7, 1.1]	0.16	0.01	0.34
L-LDL-cholesterol	-	0.5 [0.5, 0.6]	8.1×10^{-14}	0.7 [0.6, 0.9]	1.7×10^{-3}	-0.05	1.9×10^{-4}
XS-LDL-ApoB	-	1.4 [1.2, 1.7]	3.6×10^{-6}	1.6 [1.3, 1.9]	3.7×10^{-7}	0.05	2.2×10^{-4}
L-HDL-ApoA2	-	1.4 [1.2, 1.6]	2.1×10^{-4}	1.0 [0.8, 1.2]	0.94	0.02	0.14
L-HDL-free cholesterol	-	0.5 [0.4, 0.6]	3.9×10^{-12}	0.7 [0.5, 0.8]	1.5×10^{-4}	-0.09	3.2×10^{-10}
M-HDL-ApoA2	-	1.4 [1.2, 1.7]	5.8×10^{-5}	1.1 [0.9, 1.3]	0.61	0.04	4.1×10^{-3}

Odds ratio (OR) and 95% confidence interval (CI) estimates provided from logistic regression and Effect from linear regression with age- sex- and lipid-lowering medication-adjusted in the standardized metabolite variables.

Predicting incident type 2 diabetes in different baseline risk groups

The AUC of the combined ERF, and FHS metabolite models and of fasting glucose model in subpopulations stratified by age, sex, and BMI is shown in Figure 3. In the group with age < 50 years, the AUC of the combined metabolite model is significantly higher than that of fasting glucose model (AUC 0.86 [0.78, 0.94] *vs* 0.77 [0.67, 0.87], NRI = 0.72, p-value < 0.0001), whereas the AUCs of these two models are not statistically different in the elderly group (AUC 0.83 [0.78, 0.87] *vs* 0.84 [0.80, 0.88], p-value = 0.06). The AUC of the metabolite model is significantly higher than that of fasting glucose in the female group (AUC 0.88 [0.83, 0.92] *vs* 0.84 [0.79, 0.90], NRI = 0.44, p-value = 0.001), whereas in the male group there is an opposite trend (0.78 [0.72, 0.84] *vs* 0.83 [0.79, 0.88], NRI = -0.40, p-value = 0.001). Similarly, in the group with normal BMI, the AUC of metabolite model is significantly higher than that of fasting glucose model (AUC 0.85 [0.75, 0.95] *vs* 0.80 [0.66, 0.93], NRI = 0.49, p-value = 0.04). In the overweight and obese group, the trend is opposite but not significantly different (AUC 0.81 [0.76, 0.85] *vs* 0.83 [0.79, 0.87], p-value = 0.13). When the sensitivity is fixed to 80%, the specificity rises from 59% (glucose only model) to 87% (glucose and metabolite model) in the young (age < 50y), which is much higher increase than in the old (age ≥ 50y, from 66% to 82%). The specificity also grows when we add metabolites or TRFs to the prediction model. (Supplementary Figure 1) The ROC curves for the models and subgroups are given in the Supplementary Figure 2 and Supplementary Figure 3. The separation shown by time to event curves across different risk groups are given in Supplementary Figure 4.

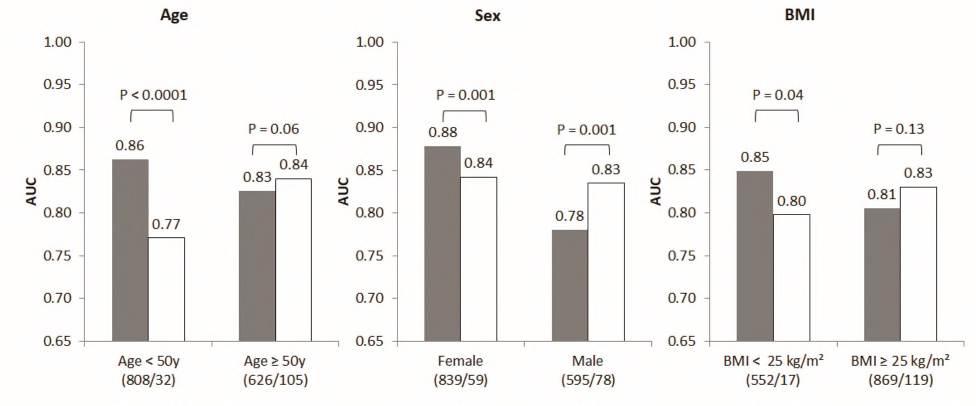


Figure 3 AUC comparisons in different subgroups. Continuous Net Reclassification Improvement (NRI) indices were performed to compare different prediction models. Black bars: metabolite model; white bars: fasting glucose model. (/): Number of controls and incident cases analyzed in the follow-up.

Discussion

In the present study, we showed that the combined effect of 24 metabolites including ten lipoprotein sub-fractions yield a powerful discrimination model for predicting future type 2 diabetes. The ERF metabolite model significantly improved the prediction performance of FHS metabolite model and fasting glucose. We showed that combined metabolite model predicts future type 2 diabetes better than fasting glucose in the population who are female, younger than 50 years, or those with normal weight. In addition, we confirmed the conclusion from the FHS that isoleucine and tyrosine are predictors of type 2 diabetes independent of other factors[12].

The ERF metabolite model includes molecules from five classes: triglycerides, amino acids, lipoproteins, phospholipids and small intermediate compounds. Among those, metabolites such as 1,5-anhydro-D-glucitol (1,5-AG), 2-hydroxybutyrate, pyruvate, phosphatidylcholines, betaine, some triglycerides, and BCAA have been previously reported to be potential predictive and diagnostic markers for type 2 diabetes[3, 12, 18, 33, 34, 35]. Despite the fact that LASSO regression method is used to select independent components of our model, various metabolites from the same biochemical class were selected, supporting the view that the sub-fractions of some classical measurements play independent functions in the pathogenesis of type 2 diabetes[36]. In line with this, the ERF metabolite model points out lipid perturbations evident in the very early stage of the disease. For example, levels of different triglycerides (e.g. TG (48:0), TG (48:1)) show independent effects. Our results on HDL and LDL sub-fractions are particularly interesting. We found them associated with both increased and decreased risk. L-HDL-ApoA2, M-HDL-ApoA2, XS-LDL-ApoB and XXL-LDL-triglycerides are associated with increased risk of type 2 diabetes, whereas L-HDL-free cholesterol, XXL-LDL-phospholipids and L-LDL-cholesterol are associated with decreased risk of type 2 diabetes. This suggests different roles for HDL and LDL particles and their content. Our results highlight the importance of reclassifying lipoproteins of clinical value into sub-fractions of HDL, LDL and VLDL, as the measurement techniques develop in the coming decade.

We also demonstrated that PC(O-34:2) is inversely associated with type 2 diabetes, which is in line with a recent study performed in the population based KORA study that showed decrease in PC(O-34:2) levels in patients with impaired glucose tolerance[18]. Phosphatidylcholine is a key element in lipoproteins[34]. Elevated plasma levels of choline and betaine mark cardiovascular risk in diabetes[37], while increased level of isoleucine was significantly associated with an increased risk of hypertriglyceridemia[38]. 2-hydroxybutyrate appears to be useful as an early indicator of insulin resistance in non-diabetic subjects[39], and its elevated serum levels have recently been indicated to predict worsening of glucose tolerance[40].

Among the other ERF metabolites, our results on two (1,5-AG and glycerol) are inconsistent with the previous studies in terms of direction of association: Suhre et al. studied on 40 diabetes cases and 60 healthy male controls in the German population[13]; Lu J. et al.'s study included 22 Chinese cases and 22 healthy controls[41], and the study by Shaham O. et al. was done in 47 healthy academic students [42]. Considering the larger sample size, our study should have yielded more reliable estimates compared to the above studies. It has been shown that levels of 1,5-AG metabolite reflect glycemic changes, and recent clinical studies demonstrated significant differences in 1,5-AG levels between diabetic patients receiving different treatments, consistent with their individual glucose profiles[43, 44].

As shown in Table 2, all metabolites are associated with prevalent diabetes, but some are not associated with incident case control status. We kept those in the ERF prediction model as this could be due to their small effect sizes which need more sample size (power) to be detected since the effect estimates were in the expected direction. Another explanation could be that some metabolite levels change depending on the duration and progression of the disease that we cannot control for in the statistical model. A third explanation is that it could as well be due to anti-diabetic medication effect but we have already filtered the associations controlling for that. The addition of ERF metabolites can complement the type 2 diabetes prediction by fasting glucose and TRFs, yielding the best model when combined. This is partly a result of our selection approach performed independently of TRFs but may also be due to the assumption that high resolution metabolites reflect different possible etiologies of type 2 diabetes. Thus, improvement of predictive performance with additional metabolites implies that potential metabolic ramifications may extend far beyond and prior to impaired glucose metabolism. It is of note that each metabolite contributed equally to the improvement of the AUC except tyrosine, exclusion of which dropped the AUC significantly. The AUC of the model without tyrosine is 0.79, and is significantly lower than the AUC of ERF metabolite model which is 0.81 (NRI = 0.38, p-value < 0.0001), suggesting that tyrosine is an important component of the model.

In the present study, we found a higher AUC of the metabolite model in lower risk population as female, younger, or leaner subgroups. For the optimum cut-off value of the ROC curve, we observed the biggest difference in specificity especially in the young age group, such that if the sensitivity of the prediction model is set to 80%, the metabolite only model yielded a specificity of 0.82, whereas the glucose model is as low as 0.59. This suggests that the metabolomics information may have better utility for type 2 diabetes prediction specifically in those without the risk condition, which is in agreement with previous study from Walford et al[17]. Interestingly, low risk population that develop type 2 diabetes were reported to have higher risk of mortality[45], raising the importance of more specific predictors suited for different underlying mechanisms. Markers which reflect the metabolic condition both dependent and independent of BMI that may partially help to address the different active pathways underlying to the MUHNW and MHO phenotypes[11].

Two additional platforms measured among subsets of the ERF population which were not included in our main analysis due to sample size restrictions gave us the opportunity to compare some of the associations using these different measurement methods. These were electrospray-Ionization MS, measured in 878 participants, using the method described before[46] and AbsoluteIDQTM p150 Kit of Biocrates Life Sciences AG measured in 989 participants as details mentioned before[47]. Supplementary Figure 5 shows the x-y plots of the effect estimates per standard error (i.e. Z score) in the 62 lipids and 9 amino acids that were measured in duplication. The Z scores between these platforms are strongly correlated with correlation coefficients ranging from 0.74 to 0.87.

The present study has a strong design such that the new cases develop among the control group in the baseline. However, due to the wide metabolite spectrum in the present study, validation of the full model in an external sample is not available yet. One limitation can be that in the present study, 46.7 % of the type 2 diabetes patients at baseline took lipid-lowering medication compared 10.3 % in the non-diabetics. To reduce the bias, all the participants were fasted overnight before taking the blood sample and we adjusted for lipid-lowering medication in each step of statistical analysis. It also needs mentioning that the metabolite set that predicts type 2 diabetes is assumed to point out the biochemical pathways disrupted before the disease onset. However, these metabolites may not be necessarily in the causal pathway. We have previously shown that most these metabolites are partially heritable[21, 23, 47] and our increasing knowledge about their genetic determinants opens up new opportunities for testing causal inference using Mendelian randomization[23].

Conducting a 14-years prospective study with comparably large sample size and wide metabolite spectrum, we developed a novel prediction model which includes informative markers of dyslipidemia, and which also increases the specificity for the young individuals. Importantly, this model has a high potential to result with better understanding of the biological mechanisms leading to glycemic deterioration in prediabetes and diabetes.

References

1. Knowler WC, *et al.* Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 346, 393-403 (2002).
2. Li R, Zhang P, Barker LE, Chowdhury FM, Zhang X. Cost-effectiveness of interventions to prevent and control diabetes mellitus: a systematic review. *Diabetes Care* 33, 1872-1894 (2010).
3. Nanditha A, *et al.* Early improvement predicts reduced risk of incident diabetes and improved cardiovascular risk in prediabetic Asian Indian men participating in a 2-year lifestyle intervention program. *Diabetes Care* 37, 3009-3015 (2014).
4. Haffner SM, Stern MP, Mitchell BD, Hazuda HP, Patterson JK. Incidence of type II diabetes in Mexican Americans predicted by fasting insulin and glucose levels, obesity, and body-fat distribution. *Diabetes* 39, 283-288 (1990).
5. Shaw JE, *et al.* Impaired fasting glucose or impaired glucose tolerance. What best predicts future diabetes in Mauritius? *Diabetes Care* 22, 399-402 (1999).
6. Droumaguet C, *et al.* Use of HbA1c in Predicting Progression to Diabetes in French Men and Women Data from an Epidemiological Study on the Insulin Resistance Syndrome (DESIR). *Diabetes Care* 29, 1619-1625 (2006).
7. Tabak AG, Jokela M, Akbaraly TN, Brunner EJ, Kivimaki M, Witte DR. Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study. *Lancet* 373, 2215-2221 (2009).
8. Gray LJ, *et al.* The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diabetic medicine : a journal of the British Diabetic Association* 27, 887-895 (2010).
9. Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB, Sr. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Arch Intern Med* 167, 1068-1074 (2007).
10. Kengne AP, *et al.* Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. *Lancet Diabetes Endocrinol* 2, 19-29 (2014).
11. Mathew H, Farr OM, Mantzoros CS. Metabolic health and weight: Understanding metabolically unhealthy normal weight or metabolically healthy obese patients. *Metabolism* 65, 73-80 (2016).
12. Wang TJ, *et al.* Metabolite profiles and the risk of developing diabetes. *Nat Med* 17, 448-453 (2011).
13. Suhre K, *et al.* Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PLoS One* 5, e13953 (2010).
14. Lu Y, *et al.* Metabolic signatures and risk of type 2 diabetes in a Chinese population: an untargeted metabolomics study using both LC-MS and GC-MS. *Diabetologia*, (2016).
15. Yu D, *et al.* Plasma metabolomic profiles in association with type 2 diabetes risk and prevalence in Chinese adults. *Metabolomics* 12, (2016).
16. Floegel A, *et al.* Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes* 62, 639-648 (2013).
17. Walford GA, *et al.* Metabolite traits and genetic risk provide complementary information for the prediction of future type 2 diabetes. *Diabetes Care* 37, 2508-2514 (2014).
18. Wang-Sattler R, *et al.* Novel biomarkers for pre-diabetes identified by metabolomics. *Mol Syst Biol* 8, 615 (2012).
19. Santos RL, *et al.* Heritability of fasting glucose levels in a young genetically isolated population. *Diabetologia* 49, 667-672 (2006).
20. Gonzalez-Covarrubias V, *et al.* Lipidomics of familial longevity. *Aging Cell* 12, 426-434 (2013).
21. Demirkan A, *et al.* Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. *PLoS Genet* 11, e1004835 (2015).
22. Verhoeven A, Slagboom E, Wuhrer M, Giera M, Mayboroda OA. Automated quantification of metabolites in blood-derived samples by NMR. *Analytica Chimica Acta*, (2017).
23. Kettunen J, *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* 7, 11122 (2016).
24. Sansone SA, *et al.* The metabolomics standards initiative. *Nature biotechnology* 25, 846-848 (2007).
25. Hu C, *et al.* RPLC-ion-trap-FTMS method for lipid profiling of plasma: method validation and application to p53 mutant mouse model. *J Proteome Res* 7, 4982-4991 (2008).

26. Aulchenko YS, de Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 177, 577-585 (2007).

27. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* 95, 221-227 (2005).

28. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* 33, 1-22 (2010).

29. Andrew G, Jennifer H. Data Analysis Using Regression and Multilevel/Hierarchical Models. *Cambridge University Press*, 529-543 (2006).

30. Pencina MJ, D'Agostino RB, Sr., Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Stat Med* 31, 101-113 (2012).

31. Dunn WB*, et al.* Molecular phenotyping of a UK population: defining the human serum metabolome. *Metabolomics* 11, 9-26 (2015).

32. Lawton KA*, et al.* Analysis of the adult human plasma metabolome. *Pharmacogenomics* 9, 383-397 (2008).

33. Kim YJ*, et al.* Association of Metabolites with Obesity and Type 2 Diabetes Based on FTO Genotype. *PloS one* 11, e0156612 (2016).

34. Park S, Sadanala KC, Kim EK. A Metabolomic Approach to Understanding the Metabolic Link between Obesity and Diabetes. *Mol Cells* 38, 587-596 (2015).

35. Yousri NA*, et al.* A systems view of type 2 diabetes-associated metabolic perturbations in saliva, blood and urine at different timescales of glycaemic control. *Diabetologia* 58, 1855-1867 (2015).

36. Kotronen A*, et al.* Serum saturated fatty acids containing triacylglycerols are better markers of insulin resistance than total serum triacylglycerol concentrations. *Diabetologia* 52, 684-690 (2009).

37. Lever M*, et al.* Betaine and Trimethylamine-N-Oxide as Predictors of Cardiovascular Outcomes Show Different Patterns in Diabetes Mellitus: An Observational Study. *PloS one* 9, e114969 (2014).

38. Mook-Kanamori DO*, et al.* Increased amino acids levels and the risk of developing of hypertriglyceridemia in a 7-year follow-up. *J Endocrinol Invest* 37, 369-374 (2014).

39. Gall WE*, et al.* alpha-hydroxybutyrate is an early biomarker of insulin resistance and glucose intolerance in a nondiabetic population. *PloS one* 5, e10883 (2010).

40. Ferrannini E*, et al.* Early metabolic markers of the development of dysglycemia and type 2 diabetes and their physiological significance. *Diabetes* 62, 1730-1737 (2013).

41. Lu J*, et al.* Serum metabolic signatures of fulminant type 1 diabetes. *J Proteome Res* 11, 4705-4711 (2012).

42. Shaham O*, et al.* Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. *Mol Syst Biol* 4, 214 (2008).

43. McGill JB*, 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* 27, 1859-1865 (2004).

44. Moses AC, Raskin P, Khutoryansky N. Does serum 1,5-anhydroglucitol establish a relationship between improvements in HbA1c and postprandial glucose excursions? Supportive evidence utilizing the differential effects between biphasic insulin aspart 30 and insulin glargine. *Diabetic medicine : a journal of the British Diabetic Association* 25, 200-205 (2008).

45. Carnethon MR*, et al.* Association of weight status with mortality in adults with incident diabetes. *JAMA* 308, 581-590 (2012).

46. Demirkan A*, et al.* Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. *PLoS Genet* 8, e1002490 (2012).

47. Draisma HH*, et al.* Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. In: *Nat Commun* (ed^(eds) (2015).

48. Haug K*, et al.* MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res* 41, D781-786 (2013).

Chapter 3.2

A Mendelian randomization study of metabolite profiles, fasting glucose and type 2 diabetes

Jun Liu, Jan Bert van Klinken, Sabina Semiz, Ko Willems van Dijk, Aswin Verhoeven, Thomas Hankemeier, Amy C. Harms, Eric Sijbrands, Nuala A. Sheehan, Cornelia M. van Duijn, Ayşe Demirkan

This chapter was published in Diabetes (2017) 66: 2915-2926.

The supplemental information for this paper is available online at
https://doi.org/10.2337/db17-0199

Abstract

Mendelian randomization (MR) provides us the opportunity to investigate the causal paths of metabolites in type 2 diabetes and glucose homeostasis. We developed and tested an MR approach based on genetic risk scoring for plasma metabolite levels, utilizing a pathway-based sensitivity analysis to control for non-specific effects. We focused on 124 circulating metabolites which correlate with fasting glucose in the Erasmus Rucphen Family study (n = 2,564) and tested the possible causal effect of each metabolite with glucose and type 2 diabetes and vice versa. We detected fourteen paths with potential causal effects by MR, following pathway based sensitivity analysis. Our results suggest that elevated plasma triglycerides might be partially responsible for increased glucose level and type 2 diabetes risk, which is consistent with previous reports. Additionally, elevated high-density lipoprotein (HDL) components i.e. S-HDL-triglycerides might have a causal role of elevating glucose levels. In contrast, large (L) and extra-large (XL) HDL lipid components i.e. XL-HDL-cholesterol, XL-HDL-free cholesterol, XL-HDL-phospholipids, L-HDL-cholesterol and L-HDL-free cholesterol as well as HDL-cholesterol seem to be protective against increasing fasting glucose, but not against type 2 diabetes. Finally, we demonstrate that genetic predisposition to type 2 diabetes associates with increased levels of alanine, and decreased levels of phosphatidylcholine alkyl-acyl C42:5 and phosphatidylcholine alkyl-acyl C44:4. Our MR results provide novel insight into promising causal paths to and from glucose and type 2 diabetes and underline the value of additional information from high-resolution metabolomics over classical biochemistry.

Introduction

Type 2 diabetes is a progressive metabolic disease characterized by hyperglycemia, initially as a result of insulin resistance and in later stages also as a result of insulin insufficiency. Type 2 diabetes is also associated with dyslipidemia, including higher circulating concentrations of triglycerides and lower concentrations of high-density lipoprotein (HDL) cholesterol. In addition, several circulating molecules have previously been shown to be dysregulated in type 2 diabetes, including phospholipids, branched-chain amino acids, keto-acid metabolites and other metabolites such as acyl-carnitines[1, 2, 3]. However, the causal paths between these metabolites and glucose/type 2 diabetes in human remain unclear from observational studies, and require randomized controlled trials that are difficult to conduct.

As an alternative, Mendelian randomization (MR) is an instrumental variable method that has gained in popularity over the last decade, to investigate causal effects of traits using genetic predictors. MR uses the principle that the allocation of genetic variants that affect a specific trait, from parents to offspring is random and unrelated to factors other than the trait [4]. Furthermore, associations between the genotype and the outcome will not be affected by reverse causation because disease will occur after the meiosis. MR has previously been used to determine whether metabolic markers such as classical blood lipids are causally involved in type 2 diabetes[5, 6, 7, 8, 9, 10, 11] and has yielded contradicting results. One reason for this could be that these studies are affected by the heterogeneous nature of the metabolic markers chosen, such as in the example of total HDL cholesterol, which in reality consists of a collection of different sized HDL particles possibly with different functions. This may dilute the causal effects of SNPs when only combined (total) HDL is considered. However, false signals may also be due to pleiotropic effects of the chosen genetic variants leading to possibly invalid instrumental variables. As high-throughput analyses techniques improve, the quantification of circulating molecules is becoming ever more detailed and precise. For instance, instead of LDL-cholesterol, HDL-cholesterol and total triglycerides (TG) determined by routine clinical biochemistry, lipoprotein particle size distribution and content as well as tens of biochemical components can now be measured using Nuclear Magnetic Resonance (NMR) spectroscopy and Mass Spectrometry (MS) based approaches[12, 13]. These additional measures offer an opportunity to gain novel insight into the pathogenesis of diseases like type 2 diabetes. With the knowledge of genetic determinants of metabolites gained from genome-wide association studies (GWAS)[14, 15, 16], one can use MR for causal inference given the specific conditions encoded in Figure 1. In the present study, with the aim of unraveling potentially causal metabolic paths that underlie the observed associations, we used genetic predictors from published metabolite

GWAS, guided by pathway-based evidence to select instrumental variables, and performed MR between selected metabolic markers and glucose/type 2 diabetes.

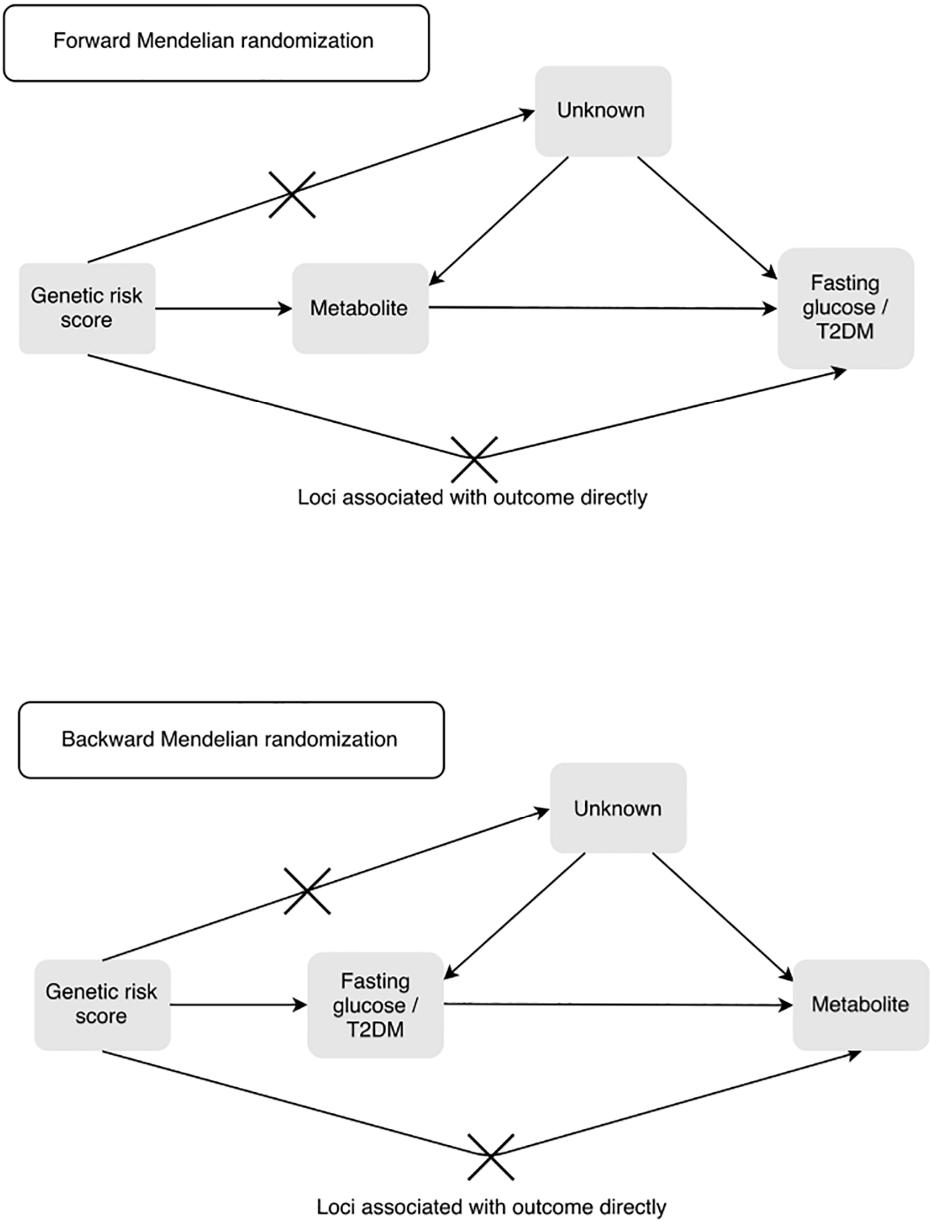

Forward Mendelian randomization
Unknown
Genetic risk score
Metabolite
Fasting glucose / T2DM
Loci associated with outcome directly
Backward Mendelian randomization
Unknown
Genetic risk score
Fasting glucose / T2DM
Metabolite
Loci associated with outcome directly

Figure 1 Overview of the Mendelian randomization process.

Materials and Methods

Study population

The observational associations between metabolites and fasting glucose/type 2 diabetes were tested in the Erasmus Rucphen Family (ERF) study which is a prospective family based study with 3,465 individuals in the Southwest of the Netherlands. The study protocol for ERF was approved by the medical ethics board of the Erasmus Medical Center Rotterdam, the Netherlands[17]. The baseline demographic data and measurements of the ERF participants were collected between 2002 and 2006. Venous blood samples were collected after at least eight hours fasting. The detailed description of the ERF study and related measurements were reported previously[17]. Type 2 diabetes was defined according to a fasting plasma glucose ≥ 7·0 mmol/L and/or anti-diabetic treatment. The analytical sample included 2,564 non-diabetic and 212 diabetic participants.

Metabolite measurements

Metabolic markers were measured by five different metabolomics platforms using the methods which have been described in earlier publications[15, 16, 18, 19]. In total 562 metabolic markers including sub-fractions of lipoproteins, triglycerides, phospholipids, ceramides, amino acids, acyl-carnitines and small intermediate compounds, which throughout this article will be referred as "*metabolites*", were measured either by NMR spectrometry or by MS. The platforms used in this research are: (1) Liquid Chromatography-MS (LC-MS, 116 positively charged lipids, comprising of 39 triglycerides, 47 phosphatidylcholines, 8 phosphatidylethanolamines, 20 sphingolipids, and 2 ceramides, available in up to 2,638 participants) measured in the Netherlands Metabolomics Center, Leiden using the method described before[18]; (2) Electrospray-Ionization MS (ESI-MS, in total 148 phospholipids and sphingolipids comprising of 16 plasmologens, 72 phosphatidylcholines, 27 phosphatidylethanolamines, 24 sphingolipids, 9 ceramides, available in up to 878 participants), measured in the Institute for Clinical Chemistry and Laboratory Medicine, University Hospital Regensburg, Germany using the method described previously[14]; (3) Small molecular compounds window based NMR spectroscopy (NMR-COMP, 41 molecules comprising of 29 low-molecular weight molecules and 12 amino acids available in up to 2,639 participants) measured in the Center for Proteomics and Metabolomics, Leiden University Medical Center[19, 20]; (4) Lipoprotein window-based NMR spectroscopy (NMR-LIPO, 104 lipoprotein particles size sub-fractions comprising of 28 VLDL components, 30 HDL components, 35 LDL components, 5 IDL components and 6 plasma totals, available in up to 2,609 participants) measured in the Center for Proteomics and Metabolomics, Leiden University Medical Center and lipoprotein sub-fraction concentrations were determined by the Bruker algorithm (Bruker BioSpin GmbH, Germany) as detailed in Kettunen *et al*[16]; (5) AbsoluteIDQTM p150 Kit of Biocrates Life Sciences AG (153 molecules comprising of 14 amino acids, 91 phospholipids, 14 sphingolipids, 33 acyl-carnitines and hexose available in up to 989 participants) measured as detailed in publication from Draisma *et al*[15] and the experiments were carried out at the Metabolomics Platform of the Genome Analysis Center at the Helmholtz Zentrum München, Germany as per the manufacturer's instructions. The laboratories had no access to phenotype information.

Statistical methods

We assessed the pairwise partial correlation between each metabolite and each glycemic trait (i.e. fasting glucose, fasting insulin, HOMA-IR, BMI and WHR) in the non-diabetic participants group. We included age, sex and lipid lowering medication as covariates. Bonferroni correction was applied based on the number of independent vectors in the data. By the Matrix Spectral Decomposition (MSD) method[21], we estimated 191 independent vectors using the pairwise bivariate correlation matrix of the 562 metabolites. A P-value $< 5.24 \times 10^{-5}$ (0.05/191/5) adjusted by number of independent vectors and number of outcomes was used as the threshold for metabolome-wide significance. The metabolites associated with glucose in the ERF study were taken forward (n = 124) as candidates for MR. In this set of 124 metabolites, we also tested the association with type 2 diabetes using logistic regression.

Mendelian randomization

For each metabolite associated with glucose, we performed two-sample bi-directional MR. The same method on two-sample MR has been performed in the previous MR studies on type 2 diabetes[6, 9, 22]. We tested if genetically varying levels of a particular metabolite affect the risk for elevated glucose and type 2 diabetes (we call this the *forward* approach) and if genetically increased risk of type 2 diabetes or elevated glucose is associated with circulating levels of a particular metabolite (we call this the *backward* approach). The associations between the instrumental variables and the exposure and the outcome are estimated from different studies, either the metabolite GWAS[14, 15, 16, 19] or fasting glucose/type 2 diabetes GWAS published by MAGIC and DIAGRAM consortium[23, 24], using the genetic risk score method. The effect of the genetic risk score was constructed by summing up the weighted effects of genome wide significant SNPs on the exposure variable, in relation to their effects on the outcome, as detailed in a previous publication[6]. This was performed using summary statistics level data utilizing the method described by Dastani *et al*[25] and implemented in the R-package "*gtx*". Figure 2 shows the overview of the instrumental variable construction. All SNPs were mapped to human genome build hg19. Given that MR assumes no pleiotropic effect beyond that on the risk factor of interest (i.e. exposure), we excluded the top SNPs from previously published body-mass index (BMI) and WHR GWAS[26, 27], and any SNPs within a 1 Mbp window distance of these, from the genetic score. We additionally excluded the genetic loci (1 Mbp window) of glucose, type 2 diabetes, insulin and HOMA-IR extracted from previous publications[23, 24, 28] in the *forward* MR and the genetic loci (1 Mbp window) of the particular metabolite of interest, using the published GWAS information in the *backwards* MR (see Supplemental Table 1 for list of genetic loci excluded at this stage). We restricted the SNP lists to a set of independent SNPs in low linkage disequilibrium (pairwise $R^2 < 0.05$) for each test[29] based on the genotype data in ERF. SNPs with disproportionate effects in the risk score were excluded to reduce pleiotropy (see Supplemental Table 2 for list of SNPs excluded at this stage). Genetic risk scores comprising > 5 SNPs which explain > 1% of variance in exposure were taken forward. This effort yielded 20 metabolite-glucose/type 2 diabetes sets in the *forward* MR and 76 glucose-metabolite sets and 79 type 2 diabetes-metabolite sets in the *backward* MR. A false discovery rate (FDR) of 0.05 was used as the significance threshold for the four series (i.e. metabolite-glucose, metabolite-type 2 diabetes, glucose-metabolite and type 2 diabetes-metabolite series).

Pathway-based sensitivity analysis

Although we applied several restrictions on the SNPs in the genetic risk scores as explained above, the instrumental variable assumption that the locus is associated with the outcome only via the association with the exposure (Figure 1) is still hard to justify in practice. We harnessed the extensive background biological knowledge available to make the additional semi parametric assumption to get the MR estimates of the causal effect. That is, for each set, we evaluated whether we could identify the gene in proximity to the locus that could explain the change in exposure levels. If a gene codes for an enzyme that catalyzes the exposure or a related compound, or if it is present in a signaling cascade that affects the exposure, we assumed that the link between the instrumental variable and the exposure was direct and not mediated by the outcome. For the *forward* approach, we checked the biological link between the locus and the target metabolite and for the *backward* approach the link with glucose. As the pathway in the disease type 2 diabetes is complex, we did not check the biological link with type 2 diabetes in the *backward* approach. If the gene directly links to the exposure, the related SNPs are taken forward to calculate the genetic risk score. Then, MR is performed for any genetic risk score (comprising > 5 SNPs) which explains > 1% of variance in exposure. To explore potential mechanistic links between the locus and the exposure, we used an automated workflow that was developed in house to gather gene-specific knowledge of all genes in proximity to each locus. In detail, we downloaded a number of online databases from the respective ftp servers and integrated them offline in

Matlab®. Subsequently, for each SNP we selected genes within a window of 100kbp, with coordinates based on the dbSNP[30], and NCBI Gene (http://www.ncbi.nlm.nih.gov/gene; GRCh37), and genes whose expression is affected by the locus (GTEx-eQTL). Then, for each gene we gathered protein-related knowledge from UniProtKB[31] and affected pathways from ConsensusPathDB[32]. Finally, for each protein we investigated metabolic activity by checking if it concerned a transporter protein in TCDB[33], or enzyme in ExPASy[34], and, if so, what the catalyzed metabolic reaction was in the KEGG database[35]. For the KEGG database the last freely available version was used. The database integration pipeline generated one HTML file for each locus, containing gene-specific knowledge and hyperlinks to the original database entries, which was then inspected for finding a mechanistic link with the exposure. The strength of this approach is that it identifies loci for which the instrumental variable assumption can be validated using genetic and biochemical evidence from online databases. We have successfully applied this workflow in earlier studies[19, 36, 37, 38]. Heemskerk *et al* gives the best example of the power of our method[37], where we re-analyzed published results of a GWAS on metabolite levels[39] and confirmed the annotation by an *in vitro* experiment.

Results

Observed associations

Characteristics of the present study population are given in Table 1. Participants with type 2 diabetes were older, tended to be more often male, and more likely to be on lipid-lowering medication. They had higher BMI, WHR, systolic blood pressure, triglycerides, fasting glucose, insulin, HOMA-IR, and lower HDL-cholesterol, adiponectin and LDL-cholesterol.

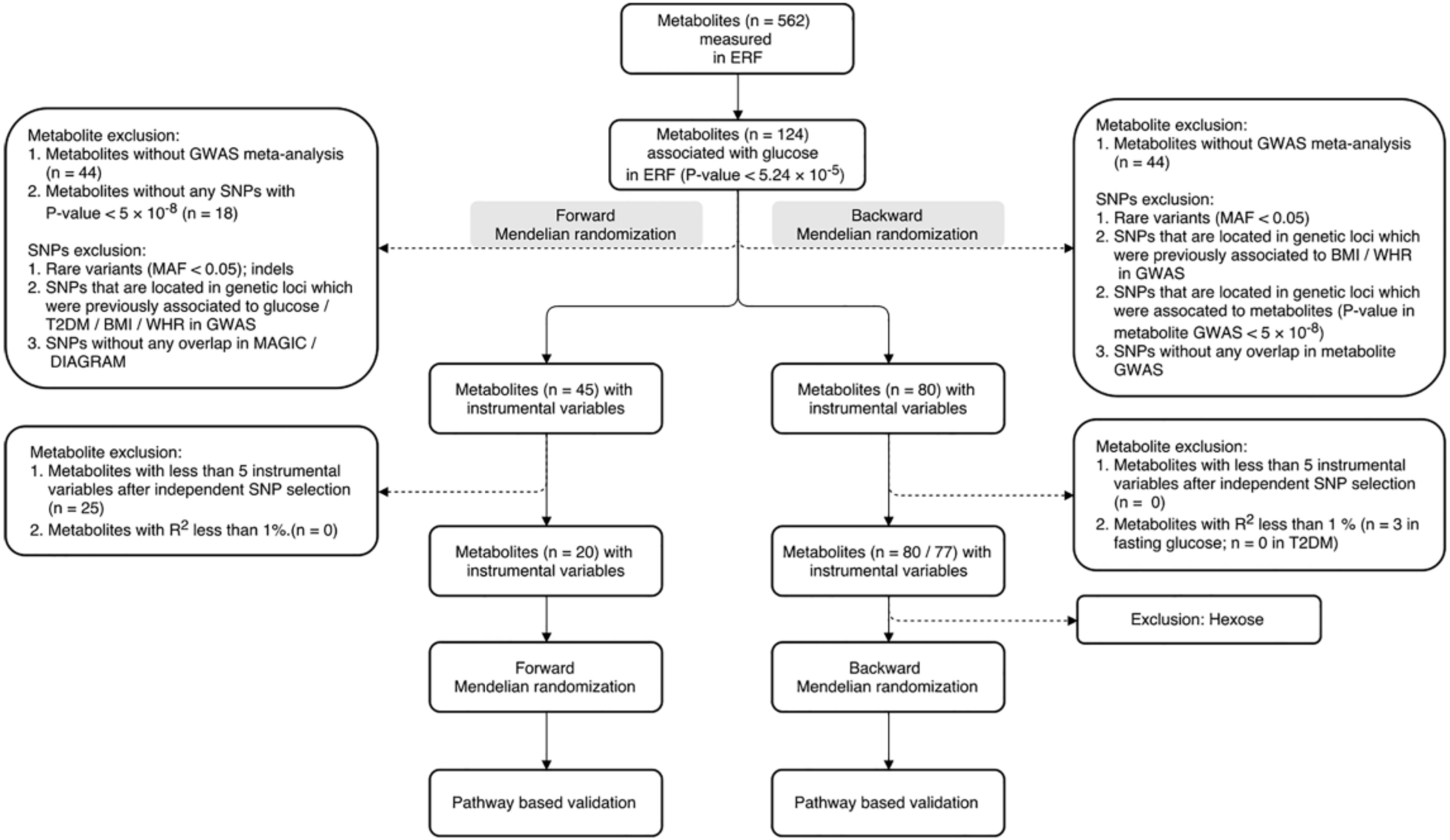
We identified 124 metabolites that observationally associate with fasting glucose in the control population with P-value < 5.24×10^{-5} (Figure 3). These consisted of 36 phospholipids (Figure 3A), 20 triglycerides (Figure 3B), 24 small molecular compounds (Figure 3C) and 44 lipoprotein particle sub-fractions (Figure 3D). Correlation coefficients, P-values as well as the overlap with previous research for all 124 metabolites are given in Supplemental Table 3. A clustered heatmap of correlation structure in-between the 124 selected metabolites are shown in Supplemental Figure 1. Among the 124, 112 of them also associated with type 2 diabetes (P < 0.05), and their associations with type 2 diabetes and glucose were in the same direction. In addition to that, their associations with BMI, WHR, fasting insulin and HOMA-IR were in line with the direction of their associations with glucose. Out of the 124 metabolites, 90 of them correlated positively and 34 correlated negatively with fasting glucose. We observed negative correlation between glucose and alkyl-acyl and diacyl phosphatidylcholines, mostly of the poly-unsaturated type, lysophosphatidylcholines, mostly of the saturated type and parts of the lipoprotein sub-fractions from LDL and HDL. These lipoprotein sub-fractions particularly consisted of lipid components of extra-large (XL) and large (L) LDL particles, XL-HDL and L-HDL particles as well as total HDL measurements. The second cluster of metabolites which we observed to correlate positively with glucose included several phospholipids; phosphatidylethanolamines, and lysophosphatidylcholines. Amino acids and low-molecular weight compounds also correlated positively with glucose, in addition to lipid side-groups, and triglycerides. Finally, from the lipoprotein sub-fractions, small (S), extra-small (XS), medium (M) and XL-VLDL particles, as well as the total VLDL components, followed by IDL and LDL-triglycerides, XS-LDL to M-LDL particle components, as well as the ApoA1 and triglyceride content of S-HDL particles were correlated positively with fasting glucose in the non-diabetic population.

Table 1 Characteristics of the study population.

	Controls n=2,564	Cases n=212	P-value	P-value*
Male [n (%)]	1132 (44.1)	108 (50.9)	0.059	0.20
Age (years)	48.2 ± 14.3	59.8 ± 11.8	6.4×10^{-32}	2.1×10^{-12}
Body mass index (kg/m2)	26.7 ± 4.6	30.0 ± 5.9	3.4×10^{-13}	3.7×10^{-12}
Waist-to-hip ratio	0.86 ± 0.10	0.95 ± 0.10	9.5×10^{-27}	2.6×10^{-17}
Systolic blood pressure (mmHg)	139 ± 20	154 ± 21	7.3×10^{-19}	8.2×10^{-6}
Diastolic blood pressure (mmHg)	80.3 ± 10.0	82.9 ± 9.9	4.5×10^{-4}	0.11
LDL-cholesterol (mmol/l)	3.8 ± 1.0	3.2 ± 1.0	4.8×10^{-15}	1.0×10^{-9}
Triglycerides (mmol/l)	1.2 (0.8, 1.6)	1.6 (1.1, 1.9)	2.0×10^{-10}	5.1×10^{-6}
HDL-cholesterol (mmol/l)	1.3 ± 0.4	1.1 ± 0.3	2.7×10^{-11}	5.6×10^{-8}
Fasting glucose (mmol/l)	4.5 ± 0.7	7.4 ± 2.2	9.4×10^{-44}	1.5×10^{-54}
Fasting insulin (mU/L)	11 (8, 15)	16 (11, 22)	1.2×10^{-7}	9.0×10^{-7}
HOMA-IR	2.3 (1.6, 3.1)	5.0 (3.7, 7.4)	1.5×10^{-23}	2.5×10^{-24}
Lipid lowering medication [n (%)]	265 (10.3)	99 (46.7)	7.2×10^{-20}	1.5×10^{-22}

Data are means ± standard deviations (SD), medians (inter-quartile range) or percentages. Triglycerides, fasting insulin, adiponectin, and HOMA-IR were natural logarithm transformed prior to analysis. P-value: T-test and Chi-squire test were used in continuous variables and categorical variables, respectively. P-value*: Logistic regression was used with adjusting age, sex and lipid lowering medication.

Figure 2 Data handling, quality checks and exclusions during Mendelian randomization. *MAGIC and DIAGRAM sets are imputed based on HapMap2 and therefore do not include indels.


Metabolites (n = 562) measured in ERF
Metabolites (n = 124) associated with glucose in ERF (P-value < 5.24 × 10^{-5})
Forward Mendelian randomization
Backward Mendelian randomization
Metabolite exclusion:
1. Metabolites without GWAS meta-analysis (n = 44)
2. Metabolites without any SNPs with P-value < 5 × 10^{-8} (n = 18)
SNPs exclusion:
1. Rare variants (MAF < 0.05); indels
2. SNPs that are located in genetic loci which were previously associated to glucose / T2DM / BMI / WHR in GWAS
3. SNPs without any overlap in MAGIC / DIAGRAM
Metabolite exclusion:
1. Metabolites without GWAS meta-analysis (n = 44)
SNPs exclusion:
1. Rare variants (MAF < 0.05)
2. SNPs that are located in genetic loci which were previously associated to BMI / WHR in GWAS
2. SNPs that are located in genetic loci which were assocated to metabolites (P-value in metabolite GWAS < 5 × 10^{-8})
3. SNPs without any overlap in metabolite GWAS
Metabolites (n = 45) with instrumental variables
Metabolites (n = 80) with instrumental variables
Metabolite exclusion:
1. Metabolites with less than 5 instrumental variables after independent SNP selection (n = 25)
2. Metabolites with R^2 less than 1%.(n = 0)
Metabolite exclusion:
1. Metabolites with less than 5 instrumental variables after independent SNP selection (n = 0)
2. Metabolites with R^2 less than 1 % (n = 3 in fasting glucose; n = 0 in T2DM)
Metabolites (n = 20) with instrumental variables
Metabolites (n = 80 / 77) with instrumental variables
Exclusion: Hexose
Forward Mendelian randomization
Backward Mendelian randomization
Pathway based validation
Pathway based validation

Mendelian randomization

Table 2 shows the significant results from the association of the relevant metabolites with fasting glucose using MR. Among the 20 eligible metabolite-glucose/type 2 diabetes sets, genetically decreased levels of eight metabolites associated significantly with fasting glucose (FDR < 0.05). These include XL-HDL-cholesterol (FDR = 0.03), XL-HDL-phospholipids (FDR = 2.76 × 10^{-3}), XS-VLDL-phospholipids (FDR = 0.04), XL-HDL-free cholesterol (FDR = 0.01), L-HDL-cholesterol (FDR = 0.01), L-HDL-free cholesterol (FDR = 2.76 × 10^{-3}), HDL-cholesterol (FDR = 0.04), and IDL-phospholipids (FDR = 0.04). After the pathway-based subset analysis, a causal role for IDL-phospholipids was not supported (FDR = 0.17). At the same time, pathway-based sensitivity analysis revealed possibly causal roles for three additional metabolic markers, including S-VLDL-triglycerides (FDR = 0.04), S-HDL-triglycerides (FDR = 0.04), and plasma-triglycerides (FDR = 0.04). Table 3 shows the suggested causal effects of metabolites on type 2 diabetes, i.e. XS-VLDL-phospholipids, IDL-phospholipids and plasma-triglycerides. Interestingly, the statistical significance for both XS-VLDL-phospholipids and IDL-phospholipids in the initial results are filtered out after the sensitivity analysis (FDR: XS-VLDL-phospholipids 0.03 *vs* 0.31; IDL-phospholipids 0.01 *vs* 0.24), while plasma-triglycerides shift to being borderline significant (FDR = 0.07 *vs* 0.046). The results from the full lists of performed *forward* MR tests are given in Supplemental Table 4 and the SNPs included in the all the genetic risk scores are given in Supplemental Table 5.

The significant results of the *backward* MR are shown in Table 4. We found that genetic predisposition to type 2 diabetes is associated with lower levels of phosphatidylcholine alkyl-acyl 42:5 (FDR = 0.02) and phosphatidylcholine alkyl-acyl 44:4 (FDR = 0.02) and higher levels of alanine (FDR = 0.02). The details of all the tested SNP sets are shown in Supplemental Table 6 and Supplemental Table 7. No possible causal role for glucose was supported. As the genetic risk scores of the glucose explained less than 1% of variance, the *backward* MR with pathway analysis is not performed. Figure 4 displays the suggested paths discovered by the MR approach after the pathway-based sensitivity analysis. Overall, the associations estimated by MR were in the consistent direction with the observed associations in ERF.

Table 2 Mendelian randomization of metabolites (exposure) on fasting glucose (outcome).

	Fasting glucose				Fasting glucose*			
Exposure	R^2 (%)	n	β	FDR	R^2 (%)	n	β	FDR
S-VLDL-triglycerides	4.80	13	0.06	0.08	3.92	10	0.08	0.04
XS-VLDL-phospholipids	7.97	23	-0.06	0.04	6.30	15	-0.07	0.04
IDL-phospholipids	7.16	26	-0.06	0.04	4.84	15	-0.05	0.17
XL-HDL-cholesterol	4.25	10	-0.09	0.03	4.25	10	-0.09	0.03
XL-HDL-free cholesterol	6.48	16	-0.09	0.01†	6.48	16	-0.09	0.01†
XL-HDL-phospholipids	10.21	22	-0.08	2.76×10^{-3}†	9.61	19	-0.09	1.72×10^{-3}†
L-HDL-cholesterol	7.58	17	-0.08	0.01†	7.41	16	-0.08	0.01†
L-HDL-free cholesterol	7.58	18	-0.09	2.76×10^{-3}†	7.27	16	-0.10	1.72×10^{-3}†
HDL-cholesterol	4.84	10	-0.07	0.04	4.67	9	-0.07	0.04
S-HDL-triglycerides	3.97	11	0.07	0.08	3.52	9	0.09	0.04
Plasma-triglycerides	3.93	11	0.07	0.08	2.78	7	0.10	0.04

The Mendelian randomization sets with FDR < 0.05 with respect to either outcome is shown in this table. R^2 (%): the percentage of explained variance in the exposure by genetic risk score. n: the number of SNPs in the genetic risk score. β: the weighted effect of the genetic risk score of exposure on outcome FDR: A false discovery rate on the number of Mendelian randomization sets adjusted P-value. * results of pathway-based analysis. † the Mendelian randomization sets with P-value < Bonferroni P-value 2.5×10^{-3} (0.05/20).

Table 3 Mendelian randomization of metabolites (exposure) on type 2 diabetes (outcome)

	Type 2 diabetes				Type 2 diabetes*			
Exposure	R^2 (%)	n	β	FDR	R^2 (%)	n	β	FDR
XS-VLDL-phospholipids	8.02	23	-0.08	0.03	6.34	15	-0.06	0.31
IDL-phospholipids	7.18	26	-0.09	0.01	4.86	15	-0.07	0.24
Plasma-triglycerides	4.21	12	0.08	0.07	3.16	8	0.12	0.046†

The Mendelian randomization sets with either FDR < 0.05 are shown in this table. R^2 (%): the percentage of explained variance in the exposure by genetic risk score. n: the number of SNPs in the genetic risk score. β: the weighted effect of the genetic risk score of exposure on outcome. FDR: A false discovery rate on the number of Mendelian randomization sets adjusted P-value. * results of pathway-based analysis. † the Mendelian randomization sets with P-value < Bonferroni P-value 2.5×10^{-3} (0.05/20).

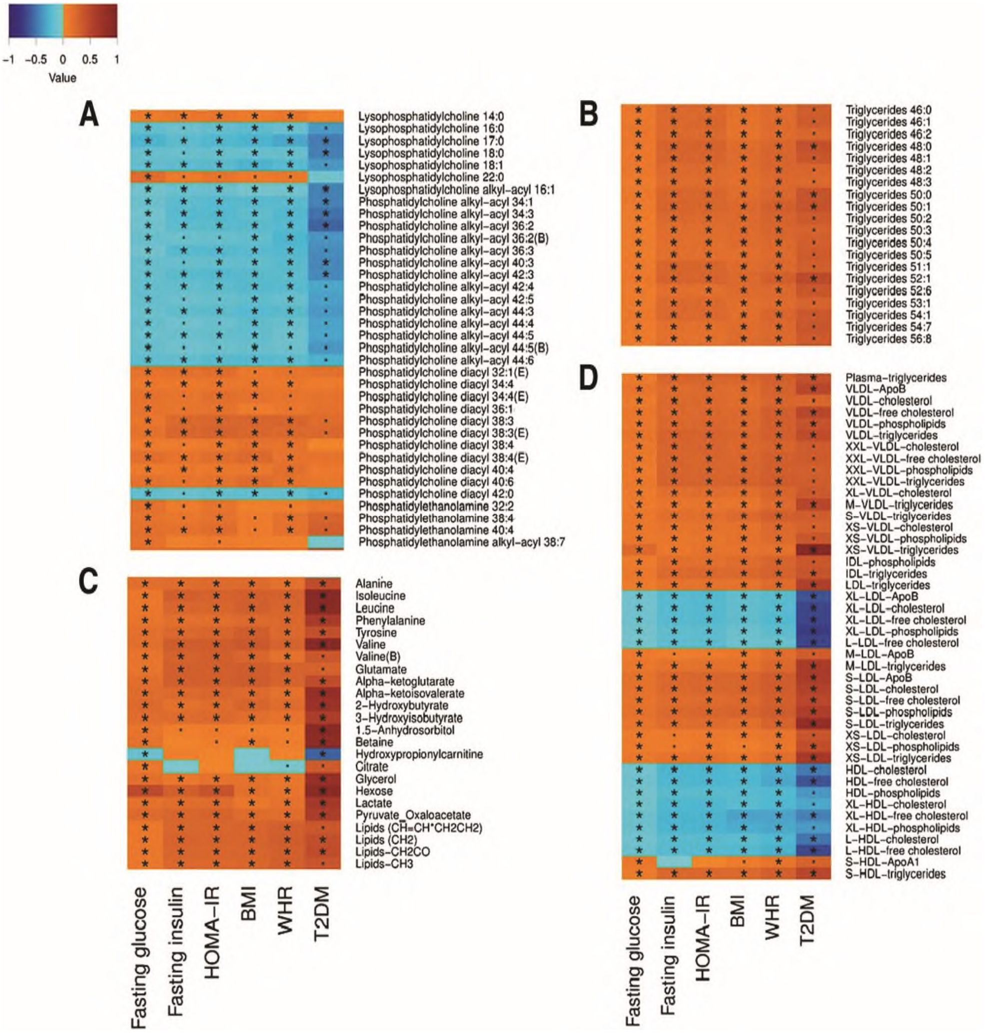


Figure 3 Metabolites correlated with markers of T2DM and anthropometric risk factors. A: Phosphatidylcholines; B: Triglycerides; C: small molecules and amino acids; D: lipoproteins. The associations between metabolites and continuous variables were performed by partial correlation in the non-diabetic population. The color in the figure displays the value of correlation coefficient. The associations between metabolites and type 2 diabetes status were performed by logistic regression. The color in the figure displays the standardized effect of metabolites on type 2 diabetes. Age, sex and lipid lowering medication are considered as covariates. FG: fasting glucose. FI: fasting insulin. HOMA-IR: homeostasis model assessment of insulin resistance. BMI: body mass index. WHR: waist-to-hip ratio. *: P-value < 5.24 × 10^{-5} (0.05/191/5). •: P-value < 0.05 and P-value ≥ 5.24 × 10^{-5}. (B): Selected measurement is from the Biocrates platform when the same metabolite is also captured by the LC-MS/NMR-COMP/NMR-LIPO platform. (E): Selected measurement is from the ESI-MS platform when the same metabolite is also captured by the LC-MS platform.

Table 4 Mendelian randomization of fasting glucose/type 2 diabetes (exposure) on metabolites (outcome)

	Fasting glucose				Type 2 diabetes			
Outcome	R^2 (%)	n	β	FDR	R^2 (%)	n	β	FDR
PC alkyl-acyl C42:5	0.83	13	NP	NP	1.51	32	-0.08	0.02†
PC alkyl-acyl C44:4	1.10	15	0.02	0.95	1.51	32	-0.08	0.02
Alanine	1.06	14	0.06	0.48	1.48	31	0.08	0.02

The Mendelian randomization sets with either FDR < 0.05 are shown in this table. PC: Phosphatidylcholine. R^2 (%): the percentage of explained variance in the exposure by genetic risk score. n: the number of SNPs in the genetic risk score. β: the weighted effect of the genetic risk score of exposure on outcome. FDR: A false discovery rate on the number of Mendelian randomization sets adjusted P-value. NP: Not performed. † the Mendelian randomization sets with P-value < Bonferroni P-value 6.33×10^{-4} (0.05/79).

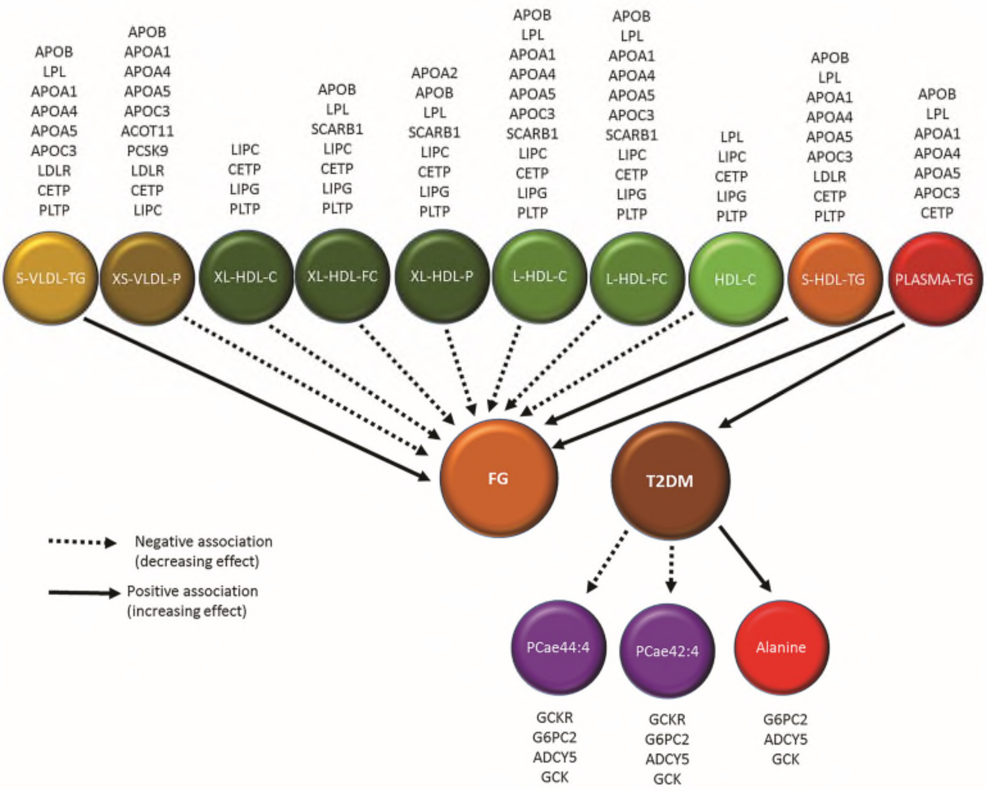


Figure 4 Suggested causal paths for glucose homeostasis and type 2 diabetes after pathway-based sensitivity analysis. FG: fasting glucose; TG: triglycerides; C: cholesterol; FC: free cholesterol; P: phospholipids; PCae: phosphatidylcholine alkyl-acyl. The gene names above the metabolite names indicate the loci where the SNPs used in the genetic risk score are located.

Discussion

We selected 124 metabolites that are correlated with glucose in the non-diabetic population and, using MR, we tested if this metabolic profile points to any causal paths involved in glucose level or type 2 diabetes. Combining metabolomics and MR, we detected fourteen candidate causal associations; ten metabolites influencing fasting glucose, one influencing type 2 diabetes and three influenced by type 2 diabetes.

Our initial observational association tests yielded correlation estimates within the expected range of power calculations for the 124 glucose-associated metabolites. To our knowledge, 35 of these metabolites were previously published to be associated with glucose or type 2 diabetes, including 31 concordant and 4 discordant results (Supplemental Table 3) in studies with very limited sample size[40, 41]. Our significant results on subfractions of lipoproteins yielded resolution on the established association of dyslipidemia, especially for the HDL subfractions.

One of our main findings is that genetically increased cholesterol, free cholesterol and phospholipid content of circulating XL-HDL and L-HDL particles together with XS-VLDL-phospholipids associate with decreased glucose level. Our second finding is that triglyceride content of S-HDL and S-VLDL particles as well as total plasma triglycerides seem to have a glucose increasing causal effect and considering the total triglycerides, this effect has been extended to the outcome type 2 diabetes. Finally, we showed that genetic predisposition to type 2 diabetes associates with lower levels of two alkyl-acyl phosphatidylcholines and higher level of alanine. Our report is the first one using higher resolution (metabolomics driven) lipoprotein based exposure variables. Hence no other study exists for comparison except for HDL-cholesterol, LDL-cholesterol and total triglycerides which, from routine biochemistry, have been previously studied as exposure variables for MR to understand their causal effects on type 2 diabetes and glucose (An overview is given in Table 5). Our method is similar to the method of White *et al*[9] and Fall *et al*[6] in terms of the application of the genetic risk score function utilizing the DIAGRAM/MAGIC datasets. White *et al*[9] showed that high levels of all three blood lipids (HDL-cholesterol, LDL-cholesterol and plasma triglycerides) were genetically associated with a lower risk of diabetes, although the results for triglycerides were inconsistent. However, the study did not consider the genetic variants that might be involved in the confounding phenotypes such as BMI, WHR, nor did they exclude the SNPs that are involved in type 2 diabetes directly. Fall *et al*[6] showed that the association between total HDL-cholesterol risk score and low fasting glucose was attenuated when adjusted for the effects of SNPs on LDL-cholesterol, triglycerides and surrogates of adiposity. Different from the two studies mentioned above, current MR was done in a broad spectrum of metabolites included a detailed sub-classification of lipoproteins that have not been tested before. Using such high resolution phenotypes, we demonstrate that decreasing effect of HDL-cholesterol on fasting glucose is more specific to the L-HDL or XL-HDL subclasses, whereas for S-HDL-triglycerides, an increasing effect exists. These results advocate that a higher resolution of high density in lipoproteins may reveal the observed epidemiological associations or biological functions of HDL-cholesterol more accurately and will uncover the mystery of complex lipids such as HDL. Certain HDL sub-fractions and characteristics of these sub-fractions may have independent associations with glucose, particularly for the small vs large size particles. Such a different role for HDL-triglycerides and HDL-large fractions occurred upon sleeve gastrectomy of obese patients and was associated with reduced insulin resistance and HDL remodeling[42]. In addition, as experimentally shown, HDL indeed may mediate glucose regulation in the pathophysiology of T2DM[43]. Suggested mechanisms include[44]: (i) Insulin secretion from pancreatic beta cells combating cellular lipid accumulation and lipotoxicity[45], endoplasmic reticulum stress and apoptosis[46, 47]; (ii) Insulin-independent glucose uptake by muscle via the AMP-activated protein kinase[48], calcium/calmodulin activated protein kinase[49]; (iii) and insulin sensitivity[50]. The ILLUMINATE trial[51] demonstrated that in a subgroup of diabetic participants statin treatment led to increased glucose levels, while this effect was not observed in participants treated with combination of statin and CETP-inhibitor torcetrapib, suggesting that CETP inhibition and consequent HDL cholesterol elevation may improve glycemic control in diabetic patients. It is of note that *CETP* gene is a major determinant of XL-HDL and was included in our MR experiment.

We have detected three associations potentially pointing out an influence of type 2 diabetes over the metabolome. The first two are long chain polyunsaturated alkyl-acyl phosphatidylcholines which are decreased in type 2 diabetes. This is interesting considering our previous report which showed that three shorter chain polyunsaturated alkyl-acyl phosphatidylcholines are increased in type 2 diabetes patients and decreased in patients using the glucose lowering drug metformin[52]. The other molecule affected by diabetes was alanine, which is a non-essential amino acid and can be synthesized in the body from pyruvate and branched chain amino acids such as valine, leucine, and isoleucine. Alanine has been previously implicated in glucose response[53]. The enzyme alanine aminotransferase (ALT) catalyzes the conversion of alanine to pyruvate and glutamate and high levels of ALT indicate liver damage.

Our study differs from previous reports in three ways. Firstly, we used a bidirectional approach and included a wide range of molecular markers to be tested, using high resolution phenotypes, measured by MS or NMR. Secondly, we exploited pathway knowledge that was gathered through an automated workflow to perform subset analysis in MR. Statistical methods that deal with pleiotropy in MR analyses, such as the Egger regression method[9, 54, 55], exist but are still being refined. They all rely on additional strong assumptions about the unobserved pleiotropy, such as the InSIDE assumption, and are sensitive to violations of these assumptions. They also suffer from a lack of power. However, one can harness the available genetic and biological knowledge in online databases in order to maximize the uniqueness of the genetic risk score for the exposure variable for this purpose and to validate the chosen instruments. It has to be mentioned that although powerful for most metabolites, our approach with the genetic and biological knowledge is also firstly conservative because it ultimately relies on the comprehensiveness of the content of the databases that are included. As a consequence, all loci for which no strong evidence is present that a nearby gene directly affects the exposure, e.g. because the involved gene is affected through a yet unknown regulatory mechanism, are excluded. Considering glucose for which the instrumental strength was initially lower compared to the others, the pathway approach yielded lower explained variance in exposure ($R^2 < 1\%$). While one can argue that this would lead to lack of power, it may also reflect the fact that such polygenic traits like glucose may not be the most suitable exposure variables for an MR analysis. To limit this, we utilized the large-population-based GWAS of broad-spectrum metabolites and fasting glucose/type 2 diabetes with the combined instrument MR approach[25]. We want to point out that although we controlled the pleiotropic effects between the outcome and exposure by (1) excluding the known predictors, (2) heterogeneity tests and (3) finally by pathway analysis, we cannot exclude a correlation between the genetic instruments tested, especially for the HDL subfractions, for which the genes coding overlap. Whilst effect alleles were weighted by their original effects estimates from each GWAS (of exposure variables), there was strong overlap in the SNPs used for different lipid subfractions, meaning the genetic instruments were not highly specific to these subfractions.

In conclusion, using MR, the present study provide evidence for potentially causal metabolic paths of glucose homeostasis and type 2 diabetes. Our results indicate that increase of large HDL particles might have decreasing effect on glucose, while increase of small HDL particles have increasing effect, refining earlier MR findings suggesting a possible causal effect of HDL on glucose levels, as well as pointing these particles out as targets for glucose management. We further found evidence that type 2 diabetes may alter levels of alkyl-acyl phosphatidylcholines and alanine which also here can be translated into prevention of disease complications and prognosis.

Table 5 Review of the previous MR in metabolites or lipids and type 2 diabetes (T2D) or glucose.

Study	Methods	Exposure	Outcome	OR/β (95%CI)	P-value	Instrumental variables and pleiotropy control
Lotta, 2016[22]	Two-sample MR	Isoleucine	T2D (n=315,571)	1.44 (1.22, 1.17)	2.0×10^{-5}	1) Independent SNPs from GWAS meta-analysis. 2)Control for pleiotropy.
		Leucine		1.73 (1.28, 2.34)	3.4×10^{-4}	
		Valine		1.45 (1.18, 1.77)	3.4×10^{-5}	
Marott SC, 2016[8]	Two-stage least-squares regression	HDL-C	T2D (n=93,097)	0.86 (0.43, 1.72)	0.68	3 variants from *ABCA1, CETP*
		TG	T2D (n=97,199)	1.05 (0.88, 1.24)	0.60	4 variants from *TRIB1, APOA5, LPL*.
White J, 2016 [9]	Conventional two-sample MR; Multivariate MR; MR-Egger	LDL-C	T2D (DIAGRAM)	0.79 (0.71, 0.88)	P < 0.05	1) Independent SNPs from GLGC GWAS. 2) *gtx* package with pleiotropic control.
		HDL-C		0.83 (0.76, 0.90)	P < 0.05	
		TG		0.83 (0.72, 0.95)*	P < 0.05	
Haase CL, 2015[5]	Two-stage least-squares regression	HDL-C	T2D (n=47,627)	0.93 (0.78, 1.11)	0.42	9 variants from *ABCA1, CETP, LCAT, LIPC, APOA1.*
Fall T, 2015[6]	Two-sample MR	LDL-C	T2D (DIAGRAM)	-0.03 (-0.19, 0.12)*	0.67	1) Independent SNPs with large effect on the lipid and smaller effect on other lipid fractions from GLGC GWAS. 2) *gtx* package with pleiotropic control.
			FG (MAGIC)	0 (-0.03, 0.03)*	0.85	
		HDL-C	T2D (DIAGRAM)	-0.19 (-0.38, -0.01)*	0.04	
			FG (MAGIC)	-0.02 (-0.06, 0.01)*	0.24	
Andersson C, 2015[10]	Two-stage least-squares regression	LDL-C	Incident T2D	0.85 (0.76, 0.96)	0.009	*GRS* from 37 LDL-C related SNPs without any pleiotropic control.
Islam M, 2012[11]	Two-stage least-squares regression	TG	T2D (n=2,111)	0.04 (0.014, 0.072)*	0.004	Included 10 independent SNPs from previous studies (excluded *FADS1* and *GCKR*).
De Silva NM, 2011[7]	Two-stage least-squares regression	TG	T2D (n=8,335)	0.99 (0.97, 1.01)	0.26	Included 10 independent SNPs from previous studies (excluded *FADS1* and *GCKR*).
			FG (n=8,271)	0 (-0.01, 0.01)*	0.88	

* β (95%CI).

References

1. Floegel A*, et al.* Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes* **62**, 639-648 (2013).
2. Kotronen A*, et al.* Serum saturated fatty acids containing triacylglycerols are better markers of insulin resistance than total serum triacylglycerol concentrations. *Diabetologia* **52**, 684-690 (2009).
3. Roberts LD, Koulman A, Griffin JL. Towards metabolic biomarkers of insulin resistance and type 2 diabetes: progress from the metabolome. *Lancet Diabetes Endocrinol* **2**, 65-75.
4. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* **27**, 1133-1163 (2008).
5. Haase CL, Tybjaerg-Hansen A, Nordestgaard BG, Frikke-Schmidt R. HDL Cholesterol and Risk of Type 2 Diabetes: A Mendelian Randomization Study. *Diabetes* **64**, 3328-3333 (2015).
6. Fall T*, et al.* Using Genetic Variants to Assess the Relationship Between Circulating Lipids and Type 2 Diabetes. *Diabetes* **64**, 2676-2684 (2015).
7. De Silva NM*, et al.* Mendelian randomization studies do not support a role for raised circulating triglyceride levels influencing type 2 diabetes, glucose levels, or insulin resistance. *Diabetes* **60**, 1008-1018 (2011).
8. Marott SC, Nordestgaard BG, Tybjaerg-Hansen A, Benn M. Components of the Metabolic Syndrome and Risk of Type 2 Diabetes. *J Clin Endocrinol Metab* **101**, 3212-3221 (2016).
9. White J*, et al.* Association of Lipid Fractions With Risks for Coronary Artery Disease and Diabetes. *JAMA cardiology* **1**, 692-699 (2016).
10. Andersson C, Lyass A, Larson MG, Robins SJ, Vasan RS. Low-density-lipoprotein cholesterol concentrations and risk of incident diabetes: epidemiological and genetic insights from the Framingham Heart Study. *Diabetologia* **58**, 2774-2780 (2015).
11. Islam M*, et al.* Multiple genetic variants explain measurable variance in type 2 diabetes-related traits in Pakistanis. *Diabetologia* **55**, 2193-2204 (2012).
12. Moreno-Gordaliza E vdLS, Demirkan A, van Duijn CM, Kuiper J, Lindenburg PW, Hankemeier T. A novel method for serum lipoprotein profiling using high performance capillary isotachophoresis. *Analytica Chimica Acta* **944**, 57-59.
13. Suna T*, et al.* 1H NMR metabonomics of plasma lipoprotein subclasses: elucidation of metabolic clustering by self-organising maps. *NMR Biomed* **20**, 658-672 (2007).
14. Demirkan A*, et al.* Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. *PLoS Genet* **8**, e1002490 (2012).
15. Draisma HH*, et al.* Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. In: *Nat Commun* (ed^(eds) (2015).
16. Kettunen J*, et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* **7**, 11122 (2016).
17. Santos RL*, et al.* Heritability of fasting glucose levels in a young genetically isolated population. *Diabetologia* **49**, 667-672 (2006).
18. Gonzalez-Covarrubias V*, et al.* Lipidomics of familial longevity. *Aging Cell* **12**, 426-434 (2013).
19. Demirkan A*, et al.* Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. *PLoS Genet* **11**, e1004835 (2015).
20. Verhoeven A, Slagboom E, Wuhrer M, Giera M, Mayboroda OA. Automated quantification of metabolites in blood-derived samples by NMR. *Analytica Chimica Acta*, (2017).
21. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* **95**, 221-227 (2005).
22. Lotta LA*, et al.* Genetic Predisposition to an Impaired Metabolism of the Branched-Chain Amino Acids and Risk of Type 2 Diabetes: A Mendelian Randomisation Analysis. *PLoS Med* **13**, e1002179 (2016).

23. Scott RA*, et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat Genet* **44**, 991-1005 (2012).

24. Morris AP*, et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**, 981-990 (2012).

25. Dastani Z*, et al.* Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet* **8**, e1002607 (2012).

26. Locke AE*, et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).

27. Shungin D*, et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187-196 (2015).

28. Prasad RB, Groop L. Genetics of type 2 diabetes-pitfalls and possibilities. *Genes (Basel)* **6**, 87-123 (2015).

29. Purcell S*, et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).

30. Sherry ST*, et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311 (2001).

31. Magrane M, UniProt Consortium. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, (2011).

32. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res* **39**, D712-D717 (2011).

33. Saier J, M.H., Tran CV, Barabote RD. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res* **34**, D181-D186 (2006).

34. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31**, 3784-3788 (2003).

35. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27-30 (2000).

36. van Leeuwen EM*, et al.* Meta-analysis of 49 549 individuals imputed with the 1000 Genomes Project reveals an exonic damaging variant in ANGPTL4 determining fasting TG levels. *J Med Genet* **53**, 441-449 (2016).

37. Heemskerk MM, van Harmelen VJ, van Dijk KW, van Klinken JB. Reanalysis of mGWAS results and in vitro validation show that lactate dehydrogenase interacts with branched-chain amino acid metabolism. *Eur J Hum Genet* **24**, 142-145 (2016).

38. Draisma HHM*, et al.* Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nature communications* **6**, (2015).

39. Suhre K*, et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54-60 (2011).

40. Suhre K*, et al.* Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PLoS One* **5**, e13953 (2010).

41. Cheng S*, et al.* Metabolite profiling identifies pathways associated with metabolic risk in humans. *Circulation* **125**, 2222-2231 (2012).

42. Eickhoff H, Guimaraes A, Louro TM, Seica RM, Castro ESF. Insulin resistance and beta cell function before and after sleeve gastrectomy in obese patients with impaired fasting glucose or type 2 diabetes. *Surg Endosc* **29**, 438-443 (2015).

43. Drew BG, Rye KA, Duffy SJ, Barter P, Kingwell BA. The emerging role of HDL in glucose metabolism. *Nat Rev Endocrinol* **8**, 237-245 (2012).

44. Siebel AL, Heywood SE, Kingwell BA. HDL and glucose metabolism: current evidence and therapeutic potential. *Front Pharmacol* **6**, 258 (2015).

45. Hao M, Head WS, Gunawardana SC, Hasty AH, Piston DW. Direct effect of cholesterol on insulin secretion: a novel mechanism for pancreatic beta-cell dysfunction. *Diabetes* **56**, 2328-2338 (2007).
46. Petremand J*, et al.* Involvement of 4E-BP1 in the protection induced by HDLs on pancreatic beta-cells. *Mol Endocrinol* **23**, 1572-1586 (2009).
47. Rutti S*, et al.* Low- and high-density lipoproteins modulate function, apoptosis, and proliferation of primary human and murine pancreatic beta-cells. *Endocrinology* **150**, 4521-4530 (2009).
48. Dalla-Riva J, Stenkula KG, Petrlova J, Lagerstedt JO. Discoidal HDL and apoA-I-derived peptides improve glucose uptake in skeletal muscle. *J Lipid Res* **54**, 1275-1282 (2013).
49. Drew BG*, et al.* High-density lipoprotein modulates glucose metabolism in patients with type 2 diabetes mellitus. *Circulation* **119**, 2103-2111 (2009).
50. Briand F*, et al.* Raising HDL with CETP inhibitor torcetrapib improves glucose homeostasis in dyslipidemic and insulin resistant hamsters. *Atherosclerosis* **233**, 359-362 (2014).
51. Barter PJ*, et al.* Effect of torcetrapib on glucose, insulin, and hemoglobin A1c in subjects in the Investigation of Lipid Level Management to Understand its Impact in Atherosclerotic Events (ILLUMINATE) trial. *Circulation* **124**, 555-562 (2011).
52. Xu T*, et al.* Effects of metformin on metabolite profiles and LDL cholesterol in patients with type 2 diabetes. *Diabetes Care* **38**, 1858-1867 (2015).
53. Wurtz P*, et al.* Circulating metabolite predictors of glycemia in middle-aged men and women. *Diabetes Care* **35**, 1749-1756 (2012).
54. Tyrrell J*, et al.* Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. *Bmj* **352**, i582 (2016).
55. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol*, (2017).

C3.2

Chapter 4

Drug-metabolite atlas

Chapter 4.1

Integration of epidemiologic, pharmacologic, genetic and gut microbiome data in a drug-metabolite atlas

Jun Liu, Lies Lahousse, Michel G. Nivard, Mariska Bot, Lianmin Chen, Jan Bert van Klinken, Carisha S. Thesing, Marian Beekman, Erik Ben van den Akker, Roderick C. Slieker, Eveline Waterham, Carla J.H. van der Kallen, Irene de Boer, Ruifang Li-Gao, Dina Vojinovic, Najaf Amin, Djawad Radjabzadeh, Robert Kraaij, Louise J.M. Alferink, Sarwa Darwish Murad, André G. Uitterlinden, Gonneke Willemsen, Rene Pool, Yuri Milaneschi, Diana van Heemst, H. Eka D. Suchiman, Femke Rutters, Petra J.M. Elders, Joline W.J. Beulens, Amber A.W.A. van der Heijden, Marleen M.J. van Greevenbroek, Ilja C.W. Arts, Gerrit L.J. Onderwater, Arn M.J.M. van den Maagdenberg, Dennis O. Mook-Kanamori, Thomas Hankemeier, Gisela M. Terwindt, Coen D.A. Stehouwer, Johanna M. Geleijnse, Leen M. 't Hart, P. Eline Slagboom, Ko Willems van Dijk, Alexandra Zhernakova, Jingyuan Fu, Brenda W.J.H. Penninx, Dorret I. Boomsma, Ayse Demirkan, Bruno H.C. Stricker, Cornelia M. van Duijn

This chapter was accepted by Nature Medicine.

The extended data and supplemental information for this paper is available online at https://www.dropbox.com/sh/olnvvg6fl5vxfz0/AADpyTmW6q6ZOv0NFAv-lxzwa?dl=0

Abstract

Progress in high-throughput metabolic profiling provides unprecedented opportunities to obtain insights into the effects of drugs on human metabolism. The Biobanking BioMolecular Research Infrastructure of the Netherlands (BBMRI-NL) has constructed an atlas of drug-metabolite associations for 87 commonly prescribed drugs and 150 clinically relevant plasma-based metabolites assessed by proton nuclear magnetic resonance (^{1}H-NMR). The atlas involves a meta-analysis of ten cohorts (18,873 persons) and uncovers 1,071 drug-metabolite associations after evaluating confounders including co-treatment. We show the effect estimates of statins on metabolites from the cross-sectional study are comparable to those from intervention and genetic observational studies. Further data integration links proton pump inhibitors to circulating metabolites, liver function, hepatic steatosis and the gut microbiome. Our atlas provides a tool for targeted experimental pharmaceutical research and clinical trials to improve drug efficacy, safety and repurposing. We provide a web-based resource for visualization of the atlas (http://bbmri.researchlumc.nl/atlas/).

In the past decade, metabolomics technology has developed rapidly[1], facilitating large-scale studies which highlighted the importance of differential molecular dynamics captured in a wide range of common complex diseases, including diabetes, cardiovascular disease, asthma and dementia[2-9]. The human metabolome is in part driven by the human genome and new genetic drivers of these metabolites continue to be revealed.[10-13]. The past decade has also seen major successes in understanding the relation of the human metabolome to the exposome, e.g. lifestyle, nutrition, environment and microbiome[14-16]. Although the use of drugs is recognized to have a major effect on the metabolism, our knowledge on the drug-metabolite associations incomplete and limited to the most commonly prescribed drugs, e.g. statins, metformin and antihypertensives[17-22]. In addition, even for the commonly prescribed drugs, the metabolic and physiologic effects, including on- or off-target effects, are virtually unexplored. Mapping these unexplored drug-metabolite associations is crucial for pharmaco-epidemiological research and practice as it may offer new avenues to improve drug efficacy, enable repurposing of drugs[23-25] and improve our understanding of the off-target effects of drugs occurring in an individual patient[26,27]. However, pointing out such associations is complicated since confounding may occur due to the metabolic changes that are the cause or the consequence of the pathology for which the drug is prescribed. Furthermore, many patients are treated with multiple drugs for multiple diseases, raising the important question of whether drug-metabolite associations are confounded by co-treatment[28]. Last not but least, longitudinal observations are often lacking for relatively rare off-target effects, forcing clinical decision making to be based on cross-sectional data.

The aim of the present study was to develop a comprehensive atlas of the associations between a wide range of commonly prescribed drugs (Supplementary Table 1) and 150 plasma-based metabolites as measured by proton nuclear magnetic resonance (^{1}H-NMR) platform of Nightingale Health (Supplementary Table 2). The platform allows rapid and cost-effective characterization of metabolites in human blood and it has been successfully used globally to discover and validate disease-metabolite associations[29], such as diabetes[30], dementia[6], cardiovascular diseases[31,32], migraine[33], graves' disease[34] and mortality[35,36]. Nightingale Health is now being validated for use in clinical care, which makes it timely to develop a pharmacological metabolomics atlas for this platform that can be used in research as well as clinical care. The term “metabolite” throughout the manuscript does not refer to the products of drug metabolism but to endogenous metabolites that are naturally produced by an organism and in this context includes lipoprotein particles as well. In the present paper, we work through a series of examples of applications of the atlas, including disentangling the disease effect of the drug-metabolite associations and exploring in-depth the interaction of metabolites and two drugs, statins and proton pump inhibitors (PPIs).

Results

Overall drug-metabolite associations

We meta-analyzed 12 datasets of ten Dutch cohorts (Supplementary Table 3, 18,873 individuals) from Biobanking and BioMolecular resources Research Infrastructure of the Netherlands (BBMRI-NL). We discovered 2,087 significant associations out of 13,050 meta-analyzed tests involving 87 drugs and 150 metabolites in model 1 with adjustment for age and sex (Bonferroni P-value threshold = 1.9×10^{-5}). The number of drug users ranged from 3,023 (16.0%, for lipophilic statin) down to 20 (0.11%, for leukotriene receptor antagonists). Among the 13,050 tests, 543 (4%) showed heterogeneity across datasets and for these we used the random-effect model to pool data across datasets. Supplementary Table 4 shows all drug-metabolite associations tested across different models, as well as disease-metabolite associations. Over the metabolites studied, effect estimates derived from different datasets agreed convincingly (pairwise P-values ranging from 1.67×10^{-11} to 1.0×10^{-318} of pairwise correlation tests) (Supplementary Figure 1 and Supplementary Table 5). Figure 1 shows the associations of model 1 for the top 15 drugs that were associated with the largest number of metabolites. The 15 drugs belong to five clinical pharmacological groups: (1) six antihypertensives, i.e. selective beta-blockers, angiotensin II antagonists, ACE inhibitors, high-ceiling diuretics, low-ceiling diuretics and potassium-sparing agents, (2) two glucose-lowering drugs, i.e. metformin and sulfonamides-urea derivatives, (3) two lipid-modifying drugs, i.e. lipophilic statin and hydrophilic statin, (4) three other cardiovascular-related drugs, i.e. vitamin K antagonists, antithrombotic agents-acetylsalicylic acid and digoxin, and (5) two others including PPIs and selective serotonin reuptake inhibitors (SSRIs). Thirteen of the top 15 drugs that were associated with the largest number of metabolites were cardio-metabolic related drugs, which may be for a large part explained by the fact that the numbers of users were large and the current metabolome spectrum contains mainly lipids and correlated with each other (Extended Data Figure 1).

Effects of BMI, smoking and co-treatment as major confounders

Next, we studied the potential confounding effect of BMI and smoking. In total, 1,640 of the 2,087 significant associations (78.6%) in model 1 were still significant after the adjustment for BMI and smoking in model 2 (Extended Data Figure 2 and Figure 1). The drugs for which the evidence for association was most dramatically impacted by adjustment for BMI and smoking were SSRIs: 59 of the initial 65 SSRIs-metabolite associations (90.8%) were no longer significant after adjustment for BMI and smoking. A major impact of adjustment was also seen for two antihypertensives: 56 (60.9%) associations with high-ceiling-diuretics were no longer significant, and 53 (49.1%) associations with angiotensin-II-antagonists lost their significance. After we additionally excluded the confounding of other drugs by adjustment for co-treatments (Extended Data Figure 3 and Figure 2), 1,071 significant associations remained to be investigated. For five out of six antihypertensives, the associations with LDL and IDL particles were explained by co-treatments (Figure 2). Notably, statin use was correlated with antihypertensives and associated with LDL and IDL particles, which leads to a false discovery association of LDL and IDL particles and antihypertensives. Most antihypertensives associations disappeared after adjusting for co-treatment including statins, except for 14.3% of the selective beta-blocker and all of the angiotensin II antagonists which remained significantly associated with LDL and IDL particles, suggesting that these associations are independent of co-treatments. In our epidemiological study, metformin was co-prescribed with hydrophilic statins and both drugs were associated to similar circulating metabolites, i.e. there were 85 metabolites associated with metformin, and 59 of these were also associated with hydrophilic statins. However, none of the metformin-metabolite associations were explained by hydrophilic statins, suggesting that metformin and hydrophilic statins are independently associated with the metabolites (Figure 2). These results above were confirmed by our sensitivity analysis from sub-samples of patients who use one drug only: all significant associations in the sensitivity analyses remained significant in the model with co-treatment adjusted for (Extended Data Figure 4).

Examples of applications of the atlas

Effect of indicated disease: drug-metabolite associations explained by the indication

First, we tested whether indicated diseases causally related to the drug-related metabolites using genetic risk score of the disease as an instrumental variable in Mendelian randomization (MR) (Supplementary Table 6, 7). Second, we associated the drug-related metabolites with the indicated disease in those who were not receiving the treatment, i.e. the on-target-treatment-naive population (Supplementary Table 4). For instance, in the current study, metformin use is associated with increasing alanine, but we also know that type 2 diabetes (causally by MR) increases alanine levels in the blood[4]. This finding raises the question of whether the disease (type 2 diabetes) or its endophenotype partially or fully explain the association of metformin and alanine. This hypothesis was supported by the finding that after excluding all metformin users, type 2 diabetes was still associated with increasing alanine levels (beta = 0.42, P-value = 8.3×10^{-19}). Integration of the findings on drug-metabolite and disease-metabolite associations suggests that alanine levels in blood are most likely raised by type 2 diabetes effect rather than by metformin effect.

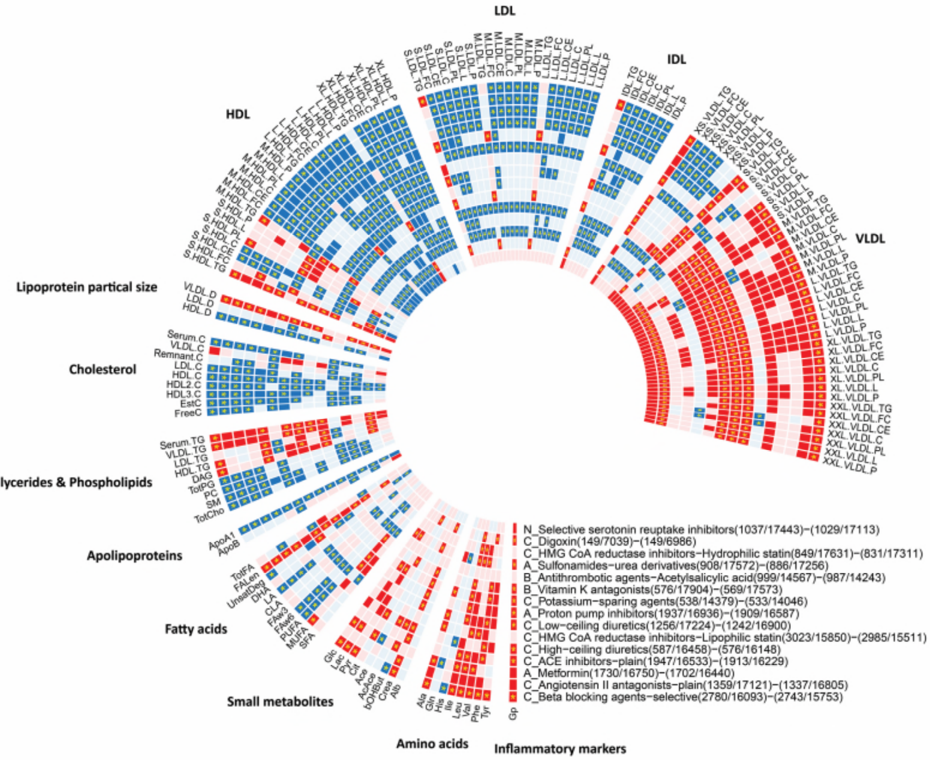


Figure 1 Drug-metabolite associations in baseline model versus model 2 with adjustment for BMI and smoking. The top 15 drugs that were associated with the largest number of metabolites in the baseline model are ordered and shown in the figure. The first letter of the Anatomical Therapeutic Chemical (ATC) code is shown in front of the drug names to identify different categories. N: nervous system; C: cardiovascular system; A: alimentary tract and metabolism; B: blood and blood forming organs. Sample sizes of the drug users and non-users in the baseline model and BMI and smoking adjusted model are shown behind the drug names, respectively. Red: positive significant associations in the baseline model ($P < 1.9 \times 10^{-5}$). Light red: positive non-significant associations in the baseline model ($P \geq 1.9 \times 10^{-5}$). Blue: negatively significant associations in the baseline model ($P < 1.9 \times 10^{-5}$). Light blue: negatively non-significant associations in the baseline model ($P \geq 1.9 \times 10^{-5}$). Star in boxes (*): The direction and significance status did not change between baseline model and model 2 ($P < 1.9 \times 10^{-5}$).

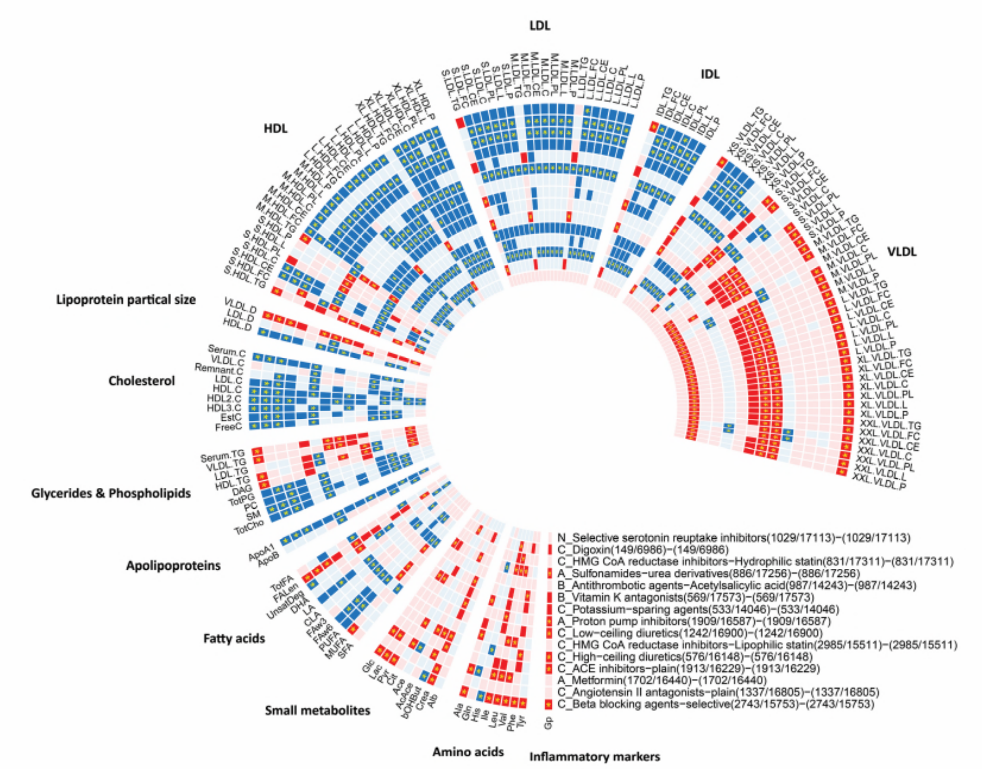


Figure 2 Drug-metabolite associations in model 2 versus model 3 with adjustment for co-treatments. The top 15 drugs that were associated with the largest number of metabolites are ordered and shown in the figure. The first letter of the Anatomical Therapeutic Chemical (ATC) code is shown in front of the drug names to identify different categories. N: nervous system; C: cardiovascular system; A: alimentary tract and metabolism; B: blood and blood forming organs. Sample sizes of the drug users and non-users in model 2 and model 3 are shown behind the drug names, respectively. Red: positive significant associations in model 2 ($P < 1.9 \times 10^{-5}$). Light red: positive non-significant associations in in model 2 ($P \geq 1.9 \times 10^{-5}$). Blue: negatively significant associations in in model 2 ($P < 1.9 \times 10^{-5}$). Light blue: negatively non-significant associations in in model 2 ($P \geq 1.9 \times 10^{-5}$). Star in boxes (*): The direction and significance status did not change between model 2 and model 3 (P-value threshold is multiple testing corrected per drug).

Following the line of research outlined above, we noticed that hypertension or high blood pressure partially or fully explained the associations of very-low-density lipoprotein (VLDL) particles and various triglycerides with beta-blockers and low-ceiling diuretics. Depression partially or fully explained the association of estimated degree of unsaturation of fatty acids and SSRIs, but not for those high-density lipoprotein (HDL) particles. Notably, type 2 diabetes or its endophenotype, fasting glucose, partially or fully explained a substantial part of associations, including 98.8% associations with metformin and 100% with sulfonamides-urea derivatives, based on a nominal significance level in the disease-metabolite associations (P-value < 0.05, Figure 3). With such a strict exclusion of effect of the indicated disease, we still found acetate was negatively associated with metformin effect, and there is no evidence that the relationship is resulted from the effect of type 2 diabetes or fasting glucose levels.

Effects of drugs in cross-sectional and longitudinal studies

We compared our results on statin-metabolite associations in the present cross-sectional study with that of the longitudinal study published earlier by Wurtz and co-workers[17]. In their paper, the changes of metabolite concentrations in blood (two time points per individual) were compared between 716 patients who started statin therapy during follow-up and 4,874 persistent non-users[17]. There are 48 metabolites that overlapped with our study[17], in which metabolite and statin use were assessed at the same time in 3,023 individuals who were using lipophilic statin and 15,850 non-users, providing a cross-sectional snapshot. Twenty-nine (60%) of the metabolites showed consistently significant results between the two studies (Figure 4). We further checked the metabolite associations with genetic variant rs12916-T located in gene *HMGCR* (3-Hydroxy-3-Methylglutaryl-CoA Reductase). This genetic variant was used as an instrumental variable for the effect of statins as the protective T allele results in low functioning HMG-CoA reductase, which is one of the pharmacological targeted effects of statins[17,37]. Figure 4 shows that 20 of the 29 associations (69.0%) were consistently and significantly associated with rs12916-T in both the cross-sectional and longitudinal analyses. The 20 statin-metabolite associations involved mainly fatty acids (30.0%) and non-HDL cholesterols and lipoprotein particles (50.0%). Meanwhile, 15 of the 19 metabolites (80%) that were inconsistently associated with statins between our study and the previous study[17] were not associated with rs12916-T.

We additionally identified 35 of the tested 55 statin-related metabolites (63.6%) associated with rs12916-T in the same direction as with lipophilic statins (Figure 4 and Supplementary Table 8). Twenty-five of them are new and complement the findings of the above-mentioned study by Wurtz and co-workers[17]. The new metabolites emerging, by association with rs12916-T in our cross-sectional analyses, involved very small to medium VLDL particles, IDL particles, LDL particles and the total phosphatidylcholine and other choline.

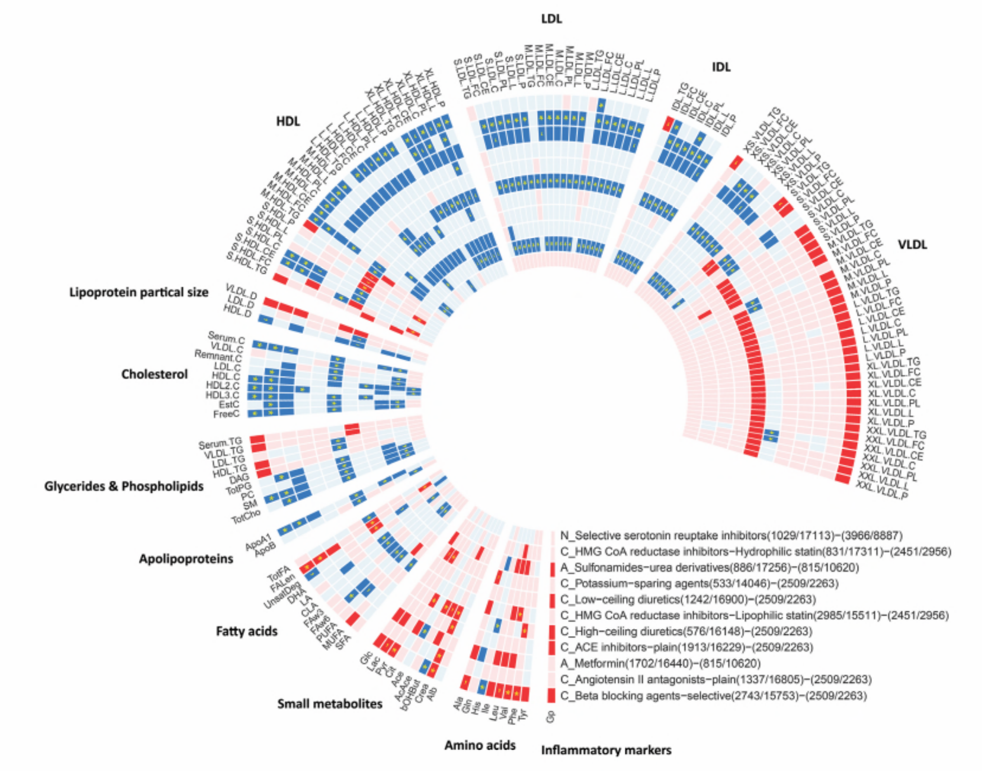


Figure 3 Drug-metabolite associations in model 3 versus significance after disentangling the indicated disease/endophenotype effect. The drugs in the top 15 drugs that were associated with the largest number of metabolites are ordered and shown in the figure. The first letter of the Anatomical Therapeutic Chemical (ATC) code is shown in front of the drug names to identify the different categories. N: nervous system; C: cardiovascular system; A: alimentary tract and metabolism. Sample size of the drug users and non-users in model 3 and sample size of the cases and controls in the disease-metabolite associations are shown behind the drug names, respectively. The sample size of the association of fasting glucose and metabolites in the non-diabetes participants is 5,871. The sample size of the association of blood pressure and metabolites in the non-hypertension participants is 2,263. Red: positive significant associations in model 3. Light red: positive non-significant associations in model 3. Blue: negatively significant associations in model 3. Light blue: negatively non-significant associations in model 3. Star in boxes (*): The significant associations confirmed after disentangling the disease/endophenotype effect (P < 0.05 in the disease-metabolite associations). Common in boxes (,): The associations confirmed after disentangling the disease/endophenotype effect (P-value threshold after multiple testing corrected per disease ≤ P < 0.05).

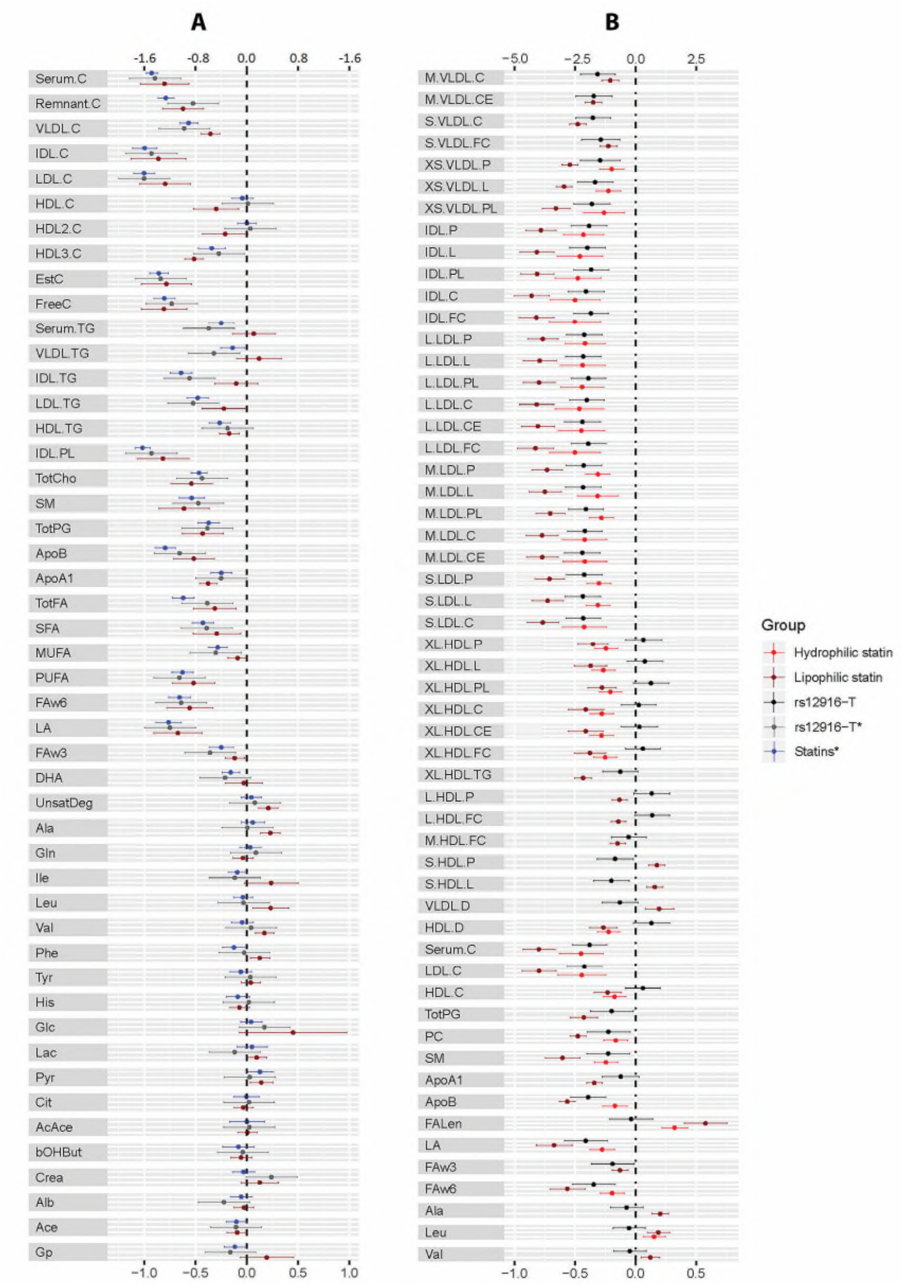


Figure 4 Comparison of statin-metabolite associations between cross-sectional study, longitudinal study and genetic study. Figure 5A shows the comparison of statin-metabolite associations between the current cross-sectional study, longitudinal study[17] and genetic study. The results of statin-metabolite associations in the longitudinal study are shown as the previous study in standard deviation (SD)-scaled metabolite concentration units (top axis) and related to the lowering effect on LDL cholesterol (bottom axis). The results of rs12916-T-metabolite associations in Figure 5A are shown in effect estimate per SD and related to the lowering effect on LDL

cholesterol (bottom axis). Figure 5B shows the comparison of significant statin-metabolite associations in the cross-sectional study and genetic study. The results of statin-metabolite associations are shown in the effect estimate (standardized metabolite concentration unit; bottom axis) and the results of rs12916-T-metabolite associations are shown in five times of the effect estimate (standardized metabolite concentration unit; top axis). The error bars of the Figure 5 are 95% confidence intervals (CI) which were statistically corrected for multiple testing. It means that if the error bar crosses the line zero, the association is not significant at the multiple testing significance level. * Statistical data were extracted from the previous longitudinal study[17].

Cross-omics analysis exploring the association of PPIs, circulating metabolites, liver function and gut microbiome

In our study, PPIs were found to be associated with 55 metabolites after adjustment of co-treatment (Figure 5A), involving small to extremely large VLDL, large HDL, triglycerides particles, monounsaturated fatty acids, isoleucine, creatinine and glycoprotein acetyls mainly a1-acid glycoprotein (glycoprotein). These associations were validated by drug-dose-metabolite associations. Analysis in the population-based cohort, Rotterdam Study (n = 700), shows a high consistency of the association between PPI (yes/no) and metabolites and the Defined Daily Dose in PPI users and metabolites (Extended Data Figure 5).

PPIs are often used by patients with cirrhosis and in these patients PPIs are associated with infections and worsening prognosis[38]. We next studied in Rotterdam Study (n = 3,436) whether the PPI-associated metabolites are also associated to liver function, including biochemical variables of liver function test and hepatic steatosis. Figure 5A and Figure 5B show a high consistency of the patterns of association between PPIs and metabolites and between metabolites and liver function (Supplementary Table 9). The consistency of associations in terms of the number of significant associations overlapping is for hepatic steatosis 98.2%, gamma-glutamyl transferase (GGT) 80.0% and alanine transaminase (ALT) 81.8% (positively associated), and 90.9% for the ratio of aspartate transaminase and ALT (AST/ALT) and 69.1% for total bilirubin (inversely associated). Of these liver function variables, total bilirubin and GGT were significantly associated with reported PPI use in Rotterdam Study (Supplementary Table 10).

We then studied the PPI-associated metabolites in relation to microbial diversity and the abundance of microbiota that are pharmacologically driven by PPI use in population[39-43]. We found that 94.5% of the metabolites associated with PPIs are also associated with gut microbial (alpha) diversity in a meta-analyses of 2,305 participants that did not use antibiotics (Figure 5C and Supplementary Table 11). Of the 92 gut microbiota of which the abundances were associated with PPI use[39], 45 were available to test the association with metabolites (Supplementary Table 12). We found that three common microbiota (phylum *Tenericutes*, class *Mollicutes* and family *Ruminococcaceae*) which showed reduced abundance in PPI users had a consistent metabolite association pattern with the PPI-metabolite association pattern but in the opposite direction (Figure 5D and Supplementary Table 13). The genera of *Scardovia* showed an increased abundance in the gut of patients using PPIs. Although the genera of *Scardovia* showed a similar metabolite association pattern to PPIs, of note is that only the association to glycoprotein reaches statistical significance when adjusted for multiple testing.

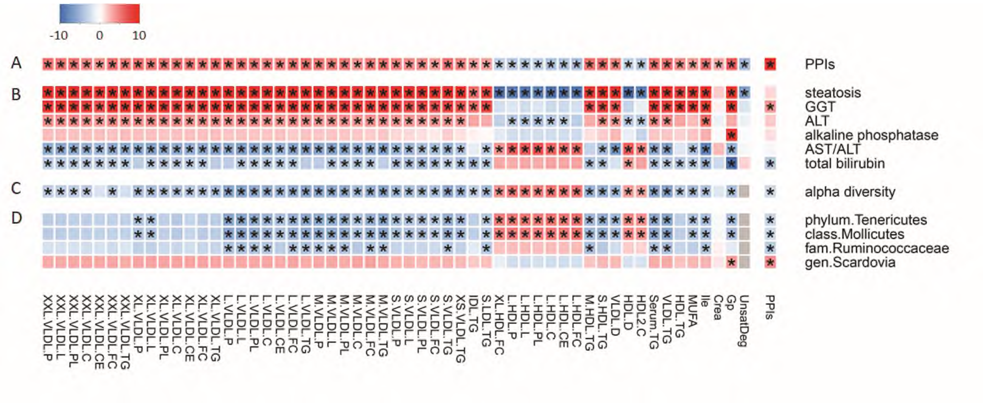


Figure 5 Integrating data of proton pump inhibitors (PPIs), metabolites, liver function measurements and gut microbiome. The figure shows the significant results after integrating the directions of the associations of between PPIs, metabolites, liver function measurements and gut microbiome. GGT: Gamma-glutamyl transferase. ALT: alanine transaminase. AST/ALT: ratio of aspartate transaminase and ALT. Red: positive association. Blue: negative association. The depth of red and blue presents the value of effect estimate per standard error. Grey: associations were not performed. Star in boxes (*): significance of the associations.

Discussion

To our knowledge, we performed the most comprehensive analysis of the interaction between 87 commonly prescribed drugs and as many as 150 circulating metabolites measured by ^{1}H-NMR in 18,873 individuals. We uncovered 1,071 drug-metabolite associations after adjustment for age, sex, BMI, smoking and co-treatment, covering a wide range of drug-metabolite associations which were not studied before. We also demonstrated three examples of applications of the atlas, disentangling disease (e.g. type 2 diabetes) and therapy (e.g. metformin) effects, aligning longitudinal and genetic analysis with our large-scale cross-sectional findings, and ultimately, linking PPI-metabolite interactions to the gut microbiome abundance and liver function.

Although many of the metabolites cluster strongly in populations (Extended Data Figure 1), our analysis shows the direction and significance of drug-metabolite associations are not always the same among different metabolites in the same cluster. This especially true for VLDL and HDL particles. This is consistent with previous studies of the role of lipid particle profiles and diseases[4,6,31-34,44,45]. This is also true for amino acids. In the Rotterdam Study, histone is clustering strongly with leucine, valine and isoleucine (P-values of correlation tests 3.3×10^{-23}). But histone is negatively associated with selective beta-blocker use (Figure 3), and leucine, valine and isoleucine are positively associated with selective beta-blocker use._We showed that BMI is a major confounder of associations with SSRIs. The high proportion of elimination in the SSRIs-metabolite associations (90.8%) after adjustment for smoking and BMI may be explained by the fact that body weight is a strong determinant of circulating metabolites and significant weight loss when not dieting or weight gain is part of the diagnostic criteria for depression[46]. After adjustment for co-treatment, the similar significant association patterns between different drugs (e.g. angiotensin II antagonists and metformin) may imply that drug-metabolite associations are independently associated with a similar shift in metabolism, but this is only true if the pathology for which the two drugs are prescribed does not explain the drug-metabolite association. For instance, if metabolic syndrome is associated with a shift in circulating metabolites, this may result in a false discovery association with drugs often prescribed to these patients (e.g. statins, antihypertensives and metformin). This type of confounding was further addressed by investigating whether drug-metabolite associations are related to the pathology (e.g. diabetes, hypertension, dyslipidemia) that indicated prescription. As a typical metabolic disorder, evidences show that type 2 diabetes explains a substantial glucose-lowering-drug-metabolite associations. The validation of the effects awaits clinical trials or prospective studies, but our example illustrates how the drug-metabolite atlas can be used in combination with disease-metabolite studies to tease out drug and disease effects and generate testable hypothesis for future trials. We further showed that to some extent the statin-metabolite associations in a large-scale cross-sectional study can mimic that of longitudinal effect of statin administration, which are preferred from a methodological perspective. It is strengthened as the two studies are benchmarked by MR. These findings suggest that the atlas does yield informative associations that may be tested in future trials and follow-up studies.

C4.1

The third and by far the most exciting example integrates the atlas data into state-of-the-art research questions. The finding that PPIs are associated with lower gut microbial diversity and a shift of the composition of the gut microbiome has been long recognized[39,41,47]. Interestingly, a recent study[48] reported that non-diabetic obese patients with hepatic steatosis have low microbial gene richness and increased genetic potential for processing of dietary lipids and dysregulation of branched-chain amino acid (BCAA) metabolism, which is very much consistent with our findings. Zooming in into oral bacteria, genus *Scardovia* is found to be increased in the gut microbiome of PPI users[39]. This raises the hypothesis that due to the PPI related changes of the gastric acid secretion in stomach, these microbiotas are reaching the gastrointestinal tract, very similar to the mechanism described in mice[40] and in the study of human gut microbiome in patients with liver cirrhosis[49]. Genus *Scardovia* was most strongly and significantly associated with glycoprotein, which is an intriguing metabolite from a clinical and epidemiological perspective, as this acute phase glycoprotein is synthesized in the liver[50] and associated to a wide spectrum of incident diseases[51], such as cardiovascular disease[52], type 2 diabetes[53], cognition[6] and all-cause mortality[36]. A key question to answer in future studies is to what extent glycoprotein plays a mediating role in the relation of gut microbiome and morbidity. Our analysis validated the previous findings that human gut microbiome is changed in patients with liver cirrhosis[49] and withdrawal of PPIs in the cirrhosis patients decrease oral-origin taxa[38] in a general population study which has very low prevalence of severe diseases such as advanced liver or kidney disease (less than 3%). Our study also showed associations of PPIs with liver function variables, gut microbiota and metabolites in the blood circulation. Again, a longitudinal or intervention study is still required to examine this hypothesis.

Another note of interest is that the experimental study of the effect of PPIs on the gut microbiome in patients with cirrhosis is based on omeprazole[38]. If we compared the different drugs that are included in the PPI category, we found that omeprazole is indeed associated to the metabolites identified in the drug category analysis (Extended Data Figure 6). However, we also found that other drugs such as lansoprazole are even more strongly and significantly associated, while the association to rabeprazole and esomeprazole is less strong and non-significant. Also these are interesting findings to follow-up.

This first comprehensive drug-metabolite atlas provides a basis for future exploration of drug-metabolite interactions, using the omics-based approach as we used or other (un)targeted experimental and longitudinal pharmaceutical research in the future. Our study includes examples of how to use the atlas which can be extended to other settings. We have limited the atlas to the most common drugs, but the atlas can be extended in the future for more rare drugs as such data for this platform are generated in larger cohorts such as UK Biobank. These “mega” cohorts would also allow studying the interaction of multiple drugs intake with sufficient statistical power systematically. On the other hand, the current atlas can be a starting point for the future researches which focus on certain limited number of drugs with metabolomics to check drug interactions. Another future challenge is to extend the atlas to a wider range of metabolites measured by other platforms (e.g. mass spectrometry) and tissues (e.g. urine). The use of MR is a strength of the current study, as it enables us to disentangle the effect of drugs and indicated diseases. However, we are not always able to capture strong instruments for the MR test, which may reduce the power of our analyses aiming to exclude the disease effects. Since our knowledge of the genes mimicking effects of drugs and diseases is rapidly growing, we are optimistic that more powerful genetic instrumental variables will be identified in the near future, opening windows of opportunities to MR analyses in pharmacometabolomics research and in clinical trials.

Our comprehensive *in vivo* reference atlas will empower future clinical and pharmacological research in a number of areas. These not only advance knowledge on the mechanisms of on-target drug effects as well as off-target drug effects but may also provide evidence for the discovery of novel therapeutic applications of known drugs. By making the atlas freely available through a web-based browser with downloadable datasets (http://bbmri.researchlumc.nl/atlas/), we hope to facilitate the use of the data by pharmacists, drug developers and clinical researchers on their drug or disease of interest.

Methods

Study population

The research was performed within the BBMRI-NL. The study included 18,873 individuals from 12 datasets of ten Dutch cohorts who had metabolites measured by Nightingale Health, drug information based on the Anatomical Therapeutic Chemical (ATC) Classification and clinical phenotypes which allow us to control for confounders. These cohorts included Rotterdam Study with three datasets (RS Dataset 1: n = 2,975, RS Dataset 2: n = 729, RS Dataset 3: n = 1,487)[54], Netherlands Twin Register (NTR, n = 3,563)[55], Netherlands Study of Depression and Anxiety (NESDA, n = 2,914)[56], Leiden Longevity Study (LLS, n = 1,873)[57], LifeLines DEEP cohort (n = 1,435)[58], Hoorn Diabetes Care System Cohort (Hoorn DCS, n = 995)[59], Alpha Omega Cohort (n = 877)[60], The Maastricht Study (TMS, n = 854)[61], Erasmus Rucphen Family study (ERF, n = 778)[62] and Leiden University MIgraine Neuro-Analysis (LUMINA, n = 393)[63].

In the examples of application atlas, we additionally involved Netherlands Epidemiology of Obesity Study (NEO; n = 6,603)[64], which is an obese cohort but have adjusted for BMI in type 2 diabetes-metabolite associations by inverse probability weighting on BMI to make the results comparable with the Dutch general population. Cohort descriptions, specific data processing and ethical compliance can be found in **Supplementary Table 3**. All studies have been approved by their respective Institutional Review Boards Local research ethics committees, and all participants have provided written informed consent to the original study.

Metabolite measurements

The present study included 150 absolute-value-based metabolites measured by high-throughput ^{1}H-NMR metabolomics (Brainshake Ltd./Nightingale Health, Helsinki, Finland). The explanation of the metabolites was shown in Supplementary Table 2. The metabolites include the quantitative molecular data on 14 lipoprotein subclasses, apolipoprotein A-I and B, multiple cholesterol and triglyceride measures, albumin, various fatty acids as well as on numerous low-molecular-weight metabolites, including amino acids, glycolysis-related measures and ketone bodies. The 14 lipoprotein subclasses included IDL, six VLDL subclasses, three LDL subclasses and four HDL subclasses based on the particle diameters. The components of these lipoprotein subclasses were quantified on total lipids (L), total cholesterol (C), particle concentration (P), phospholipids (PL), triglycerides (TG), free cholesterol (FC) and cholesterol esters (CE). The values of the representative coefficients of variations (CVs) for the metabolites ranged between 0.3% and 19.5% (mean 4.5%) and most values are comparable to the clinical chemistry assays[11,65].

The blood samples of different cohorts have been centralized in Leiden University Medical Center (LUMC) and were shipped to and analyzed by Nightingale Health as part of a national initiative. A standardized protocol of metabolite measurement was applied for all the cohorts following the comprehensive quantitative platform generated by Nightingale Health and described originally by Soininen et al[11,65,66]. The protocol includes sample quality control and sample preparation, data storage and automated spectral analyses. The metabolite values which were suggested to be uncreditable in the quality control provided by Nightingale Health during the measurement procedure were treated as missing. Within the consortium, we checked and reported the distribution of zero values in our previous study by van den Akker, et al[67]. The quality control was unified and included an in-depth evaluation of the consistency of findings across datasets, a metabolite correlation matric and the principal component analysis on cohorts with different population structure. Pearson's correlation test was used to check the pairwise correlation of the overall estimate values of drug-metabolite associations in model 1 between datasets. We also checked the correlation matrix of metabolites in a population-based cohort, Rotterdam study (n = 5,191), by Pearson's correlation and hierarchical cluster analysis, reporting that the distinct clustering groups were in accordance with the biochemical pathways (Extended Data Figure 1 and Supplementary Table 14). The effect of population structure on metabolite clustering was checked by principal component analysis using joint data from four cohorts that differ extremely in population: 1) one population-based study, Rotterdam Study[54], 2) one family-based study, ERF[62], 3) one disease-based study, TMS[61], which includes only patients with type 2 diabetes in the current dataset, and 4) a case-control study, Alpha Omega Cohort[60], including patients with cardiovascular disease and non-disease controls (details in Supplementary Table 3). The obvious difference between Alpha Omega Cohort and TMS underscore that meta-analysis should be performed instead of a joint analysis with pooled data (Extended Data Figure 7): the fixed-effect meta-analysis assumes a similar effect and structure over cohorts, while the random-effect meta-analysis allows for high heterogeneity across cohorts.

As some distributions of metabolites were skewed, we transformed the metabolite values in each cohort to normal distribution. We first added the value of one to all the metabolites before doing the natural logarithm transformation, to include samples labeled zero that had metabolite levels below the detectable value. Then we scaled these transformed values to standard deviation units.

Drug categories

The drug information was classified by ATC codes in each cohort. In brief, the drug information per cohort was obtained either from the pharmacy records or from the questionnaires during the interview. Details on drug information of each cohort can be found in Supplementary Table 3. We used the drug category instead of the individual compound in all the analyses. We merged drugs with similar chemical, pharmacodynamics, pharmacokinetics and/or therapeutic characteristics into one category. For the ATC codes used for combinations of active ingredients, we categorized them into separate categories if possible. We excluded categories with five or fewer users in each cohort or less than 20 users in total from all the cohorts. Thus, we ended up with 87 drug categories (Supplementary Table 1). The drug categorization was confirmed by two experienced pharmacologists: Lies Lahousse and Bruno H.C. Stricker. Throughout the text, the term *drug category* is further referred to *drug*. The individuals with metabolite and drug information available were included in the analysis.

Statistical analysis

All statistical analyses were performed using *R* statistical software and the two-tailed test was considered.

Association between drug and metabolite

To check for drug-metabolite associations, linear regression was performed in each cohort with drug use as an independent variable and metabolite as a dependent variable. Linear regression was used in the individual cohorts. The specific family relationship was considered in the three family-based cohorts (see details in Supplementary Table 3). In the baseline analysis, we used age and sex as the covariates (model 1). we additionally adjusted for smoking (current smoking: yes/no) which is a major common risk factor of pathology[68] and body mass index (BMI, kg/m^2) which is a major determinant of circulating metabolites that captures the effects of diet and physical activity[69] (model 2). Meta-analysis was performed with either the inverse-variance weighted fixed-effect model (no heterogeneity between cohorts) or a maximum likelihood random-effect model (significant heterogeneity between cohorts). The degree of heterogeneity was based on Cochran's Q test. The P-value threshold of both the Cochran's Q test and the meta-analysis was Bonferroni corrected with 30 independent equivalents of the 150 metabolites and 87 drugs tested (P-value < 1.9×10^{-5}). Matrix Spectral Decomposition was used to calculate the number of independent equivalents[70] in the largest population-based dataset: RS Dataset 1. R-package '*metafor*" was used for the meta-analysis[71].

Effects of co-treatment: drugs prescribed together

We next checked the potential confounding of drugs which were prescribed together (model 3) in each significant drug-metabolite pair. A co-treatment matrix with Spearman's correlation was made in the two population-based cohorts (Rotterdam Study and LifeLines DEEP, n = 6,631) separately and meta-analyzed. Potential confounding co-treatment for each drug-metabolite pair was defined if: (1) a drug was positively correlated with the target drug (explained as prescribed together, Extended Data Figure 8 and Supplementary Table 15), and (2) this drug and the target drug were associated with the target metabolite in the same direction. We used the Bonferroni P-value correction with 85 drugs available in the co-treatment matrix (P-value < 5.9×10^{-4}). We then performed the same regression analysis as above in each dataset (12 datasets) and meta-analyzed with age, sex, BMI, smoking and all the available confounding co-treatments as covariates in each significant drug-metabolite pair (model 3). A sensitivity analysis was performed in the sub-samples of patients who use one drug only (one-drug-users) and all-treatment-naive controls adjusting for age, sex, BMI and smoking. We used the Bonferroni P-value threshold by correcting the independent equivalents of the number of tested significant metabolites for each drug.

Mendelian randomization to check the effect of indicated disease on metabolites

We further focused on the drugs in the top 15 drug lists that had the largest number of related metabolites and the metabolite associations after adjustment for co-treatments. We explored the confounding effect of the disease indicating the prescription of the drug by MR. MR is a statistical method which uses the effect of genetic variants determining an exposure and test its association with the outcome under study, based on the assumption that the genetic variant is inherited independent of the confounding variables[72]. Thus, we tested whether the genetic determinants driving indicated diseases are also related to metabolites, using the genetic risk score of the disease as an instrumental variable of exposure. Genetic risk scores comprising > 5 genetic single nucleotide polymorphisms (SNPs) and explaining > 1% of variance in exposure were taken forward. For type 2 diabetes, we looked up the results from our previous well-organized MR research[4], and 16 metabolites were found to be associated with either metformin or sulfonamides-urea derivatives. In brief, this MR research was a two-sample bi-direction MR study checking the causation of metabolites and type 2 diabetes and fasting glucose, following by biological knowledge-based sensitivity analysis to control for the pleiotropic effect of the SNPs in the instrumental variables[4]. We currently used the results of the backward MR that was checking the association of the genetic score of type 2 diabetes and metabolites.

For hypertension and depression, we performed two-sample MR based on the previous GWAS results on blood pressure[73] (n = 317,754), major depression[74] (n = 135,458 cases and n = 344,901 controls) and NMR metabolite GWAS[11] (n = 24,925). Among the 123 metabolites associated with either antihypertensives, 96 metabolites were available to perform MR with systolic and diastolic blood pressure. We also performed MR of major depression with six metabolites associated with SSRIs. We did not perform MR of dyslipidemia over the statin-associated metabolites because most of the metabolites were lipoproteins which are part of the dyslipidemia definition.

The R-package *TwoSampleMR* was used for the two-sample MR tests[75]. Genetic loci of major depression were extracted from previous paper as its original GWAS was not available[74]. The default pipeline in the *TwoSampleMR* package[75] was used. In brief, the genetic score was based on the top genetic determinant SNPs (P-value $< 5 \times 10^{-8}$) with linkage disequilibrium (LD) $R^2 < 0.001$ within 10,000bps clumping distance. Proxy SNPs were searched for if SNPs not available in the metabolite GWAS ($R^2 > 0.8$). The palindromic SNPs with minor allele frequency less than 0.3 were excluded. It resulted in 161 independent SNPs for systolic blood pressure (R2 = 2.6%), 174 SNPs for diastolic blood pressure (R2 = 2.8%) and 40 SNPs for major depression (R2 = 1.1%). Inverse variance weighted MR, Maximum likelihood MR, MR Egger analysis and median-based estimator were also performed to check the significant results[75]. We used the Bonferroni P-value threshold by correcting the independent equivalents of the number of tests per disease: P-value $< 2.3 \times 10^{-3}$ for blood pressure, and P-value < 0.025 for depression.

Indicated disease-metabolite associations: effect of indicated disease

We associated the drug-related metabolites with the indicated disease in those who were not receiving the drug under study, i.e. the on-target-treatment-naive population. This was focused on type 2 diabetes, dyslipidemia, hypertension and depression. The type 2 diabetes analyses were performed based on Rotterdam Study and NEO. Type 2 diabetes was defined as fasting glucose ≥7.0 mmol/L, and the cases who used glucose-lowering drugs were excluded in the analysis (815 cases and 10,619 non-diabetics controls in meta-analysis). We performed a regression model with type 2 diabetes status as an independent variable, glucose-lowering-drug-related metabolite as the dependent variable. Covariates included age, sex, BMI, smoking and lipid-modifying drugs.

Dyslipidemia and hypertension were tested in ERF and Rotterdam Study. We tested the association of 87 lipid-modifying-drug-related metabolites and dyslipidemia. Dyslipidemia was defined according to the National Cholesterol Education Program-Adult Treatment Panel III as either total cholesterol ≥ 240 mg/dL, LDL-C ≥160 mg/dL, HDL-C < 40 mg/dL, or triglyceride ≥200 mg/dL[76] (2,451 cases and 2,956 controls in meta-analysis). We excluded the subjects with lipid-modifying drugs and adjusted for age, sex, BMI and smoking in the model. The associations of 123 antihypertensives-related metabolites and hypertension were performed. Hypertension was defined as either systolic blood pressure ≥140 mmHg or diastolic blood pressure ≥90 mmHg (2,506 cases and n = 2,263 controls in meta-analysis). We excluded the subjects with antihypertensives and adjusted for age, sex, BMI, smoking and lipid-modifying drugs in the model.

For depression, we tested the associations between the six SSRIs-related metabolites and depressed mood in the participants without any antidepressant drug (ATC code as N06A)[77]. Depressed mood was measured by either diagnostic interviews or validated depression questionnaires (3,966 cases and 8,887 controls in the meta-analysis). The detailed definition of cases and control in cohorts was descripted in our previous publication[77]. We adjusted for age, sex, fasting status, lipid-modifying drug and current smoking status.

In addition, we checked the association of fasting glucose and glucose-lowering-drug-related metabolites in the non-diabetes population (n = 5,871) and the association of systolic and diastolic blood pressure and antihypertensives-related metabolites in the non-hypertension population (n = 2,263) in ERF and Rotterdam Study. The non-diabetes population were those fasting glucose ≤ 6.9 mg/dl and without any anti-diabetics treatment; the non-hypertension population were those systolic blood pressure less than 140 mmHg, diastolic blood pressure less than 90 mmHg and without any antihypertensives. Linear regression was performed with adjustment for age, sex, BMI, smoking and lipid-modifying drugs in the model. The P-value threshold for significance of associations was corrected for the number of independently tested metabolite equivalents per disease or endophenotype. Nominal significance between disease/endophenotype and metabolite was also considered (P-value < 0.05).

A comparison of cross-sectional and longitudinal studies and benchmarking findings by genetics: using statin as an example

Forty-eight metabolites in the current cross-sectional study were also studied in the previous longitudinal study by Wurtz and co-workers which also quantified the ^{1}H-NMR metabolic profiles in blood samples but focused on the change of metabolite concentrations of two time points: baseline and follow-up[17]. As the longitudinal study only adjusted for age and sex, we used the same model in the present cross-sectional study to allow a fair comparison. Since the effects of lipophilic statin and hydrophilic statin are similar in the current study, we used the results of lipophilic statin which had the largest sample size to do the comparison. The results of MR analysis, association of rs12916-T and metabolites, from Wurtz and co-workers were also used in the comparison[17].

We then compared the significant statin-metabolite associations in the current cross-sectional study with the associations of rs12916-T and metabolites. We used the GWAS results of the NMR metabolites from our previous paper which included 24,925 individuals without lipid-modifying drug usage[11]. It resulted in 55 metabolites in the comparison.

PPIs, circulating metabolite and liver function

We studied biochemical variables in liver function test, i.e. ALT, AST, GGT, AST/ALT, total bilirubin and alkaline phosphatase, and hepatic steatosis. The liver function test used automatic enzymatic procedures (Roche Diagnostics GmbH, Mannheim, Germany)[78]. Abdominal ultrasonography was performed by a certified and experienced technician (Pavel Taimr) on Hitachi HI VISION 900 (Highland Heights, OH). Images were stored digitally and re-evaluated by a single hepatologist with more than ten years of experience in ultrasonography. The diagnosis of steatosis was determined by the ultrasound technician according to the protocol by Hamaguchi et al[79].

Linear regression was performed in Rotterdam Study (n = 3,436) with liver function measurements as an independent variable and metabolite levels as a dependent variable. The covariates included age, sex, BMI, smoking, lipid-modifying drugs, PPIs and alcohol intake per day calculated from questionnaires. The P-value threshold was Bonferroni corrected with 10 independent equivalents of 55 PPI-related metabolites and six independent equivalents of the seven liver function measurements (P-value $< 8.3 \times 10^{-4}$). We further checked the association of PPI use and liver function measurements by linear regression with adjustment for age, sex, BMI, smoking and alcohol intake per day (P-value $< 8.3 \times 10^{-3}$).

PPIs, circulating metabolites and gut microbiome

We extracted the associations of PPIs with gut microbiota and (alpha) diversity from our previous paper by Imhann and co-workers[39]. Age, sex, BMI, antibiotics use and sequence read depth were corrected in the association analysis[39]. In total, 92 bacterial taxa abundance assessed by tag sequencing of the 16S rRNA gene[58] and Shannon's diversity index (alpha diversity) were reported to be significantly different between PPI users and non-users (211 PPI users and 1,594 non-users, FDR < 0.05). Forty-five of the 92 bacterial taxa abundance and alpha diversity were also tested association with metabolites measured by Nightingale Health in our previous study[80]. In brief, it included 2,309 individuals who were not using antibiotics from Rotterdam Study (n = 1,390) and LifeLines DEEP (n = 915)[47,58]. Age, sex, BMI, technical covariates (time in mail and storage time) and medication use (lipid-modifying drugs, metformin and PPIs) were adjusted in the association analysis. The P-value threshold for gut microbiota was Bonferroni corrected with 10 independent equivalents of 55 PPI-related metabolites and 15 independent equivalents of the 45 gut microbiota (P-value < 3.3×10^{-4}). P-value threshold for alpha diversity was 5.0×10^{-3}.

References

1. Patti, G.J., Yanes, O. & Siuzdak, G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol* **13**, 263-269 (2012).
2. Park, J.E., Lim, H.R., Kim, J.W. & Shin, K.H. Metabolite changes in risk of type 2 diabetes mellitus in cohort studies: A systematic review and meta-analysis. *Diabetes Res Clin Pract* **140**, 216-227 (2018).
3. McGarrah, R.W., Crown, S.B., Zhang, G.F., Shah, S.H. & Newgard, C.B. Cardiovascular Metabolomics. *Circ Res* **122**, 1238-1258 (2018).
4. Liu, J.*, et al.* A Mendelian Randomization Study of Metabolite Profiles, Fasting Glucose, and Type 2 Diabetes. *Diabetes* **66**, 2915-2926 (2017).
5. Wang, T.J.*, et al.* Metabolite profiles and the risk of developing diabetes. *Nat Med* **17**, 448-453 (2011).
6. van der Lee, S.J.*, et al.* Circulating metabolites and general cognitive ability and dementia: Evidence from 11 cohort studies. *Alzheimers Dement* **14**, 707-722 (2018).
7. Mapstone, M.*, et al.* Plasma phospholipids identify antecedent memory impairment in older adults. *Nat Med* **20**, 415-418 (2014).
8. Thorburn, A.N.*, et al.* Evidence that asthma is a developmental origin disease influenced by maternal diet and bacterial metabolites. *Nat Commun* **6**, 7320 (2015).
9. Mabalirajan, U.*, et al.* Linoleic acid metabolite drives severe asthma by causing airway epithelial injury. *Sci Rep* **3**, 1349 (2013).
10. Illig, T.*, et al.* A genome-wide perspective of genetic variation in human metabolism. *Nat Genet* **42**, 137-141 (2010).
11. Kettunen, J.*, et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* **7**, 11122 (2016).
12. Draisma, H.H.*, et al.* Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat Commun* **6**, 7208 (2015).
13. Suhre, K.*, et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54-60 (2011).
14. Beger, R.D.*, et al.* Metabolomics enables precision medicine: "A White Paper, Community Perspective". *Metabolomics* **12**, 149 (2016).

15. Rappaport, S.M., Barupal, D.K., Wishart, D., Vineis, P. & Scalbert, A. The blood exposome and its role in discovering causes of disease. *Environ Health Perspect* **122**, 769-774 (2014).

16. Schutte, B.A.*, et al.* The effect of standardized food intake on the association between BMI and (1)H-NMR metabolites. *Sci Rep* **6**, 38980 (2016).

17. Wurtz, P.*, et al.* Metabolomic Profiling of Statin Use and Genetic Inhibition of HMG-CoA Reductase. *J Am Coll Cardiol* **67**, 1200-1210 (2016).

18. Altmaier, E.*, et al.* Metabolomics approach reveals effects of antihypertensives and lipid-lowering drugs on the human metabolism. *Eur J Epidemiol* **29**, 325-336 (2014).

19. Elbadawi-Sidhu, M.*, et al.* Pharmacometabolomic signature links simvastatin therapy and insulin resistance. *Metabolomics* **13**(2017).

20. Kaddurah-Daouk, R.*, et al.* Lipidomic analysis of variation in response to simvastatin in the Cholesterol and Pharmacogenetics Study. *Metabolomics* **6**, 191-201 (2010).

21. Xu, T.*, et al.* Effects of metformin on metabolite profiles and LDL cholesterol in patients with type 2 diabetes. *Diabetes Care* **38**, 1858-1867 (2015).

22. t Hart, L.M.*, et al.* Blood Metabolomic Measures Associate With Present and Future Glycemic Control in Type 2 Diabetes. *J Clin Endocrinol Metab* **103**, 4569-4579 (2018).

23. Moosavinasab, S.*, et al.* 'RE:fine drugs': an interactive dashboard to access drug repurposing opportunities. *Database (Oxford)* **2016**(2016).

24. Voora, D. & Shah, S.H. Pharmacometabolomics meets genetics: a “natural” clinical trial of statin effects. (Journal of the American College of Cardiology, 2016).

25. Wishart, D.S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat Rev Drug Discov* **15**, 473-484 (2016).

26. Van Norman, G.A. Drugs, devices, and the FDA: Part 1: an overview of approval processes for drugs. *JACC: Basic to Translational Science* **1**, 170-179 (2016).

27. (FDA), U.S.F.D.A. 22 Case studies where phase 2 and phase 3 trials had divergent results *https://www.fda.gov/* (2017).

28. Brahma, D.K., Wahlang, J.B., Marak, M.D. & Ch Sangma, M. Adverse drug reactions in the elderly. *J Pharmacol Pharmacother* **4**, 91-94 (2013).

29. Wurtz, P.*, et al.* Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies. *Am J Epidemiol* **186**, 1084-1096 (2017).

30. Ahola-Olli, A.V.*, et al.* Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish cohorts. *Diabetologia* (2019).

31. Ference, B.A.*, et al.* Association of Triglyceride-Lowering LPL Variants and LDL-C-Lowering LDLR Variants With Risk of Coronary Heart Disease. *JAMA* **321**, 364-373 (2019).

32. Holmes, M.V.*, et al.* Lipids, Lipoproteins, and Metabolites and Risk of Myocardial Infarction and Stroke. *J Am Coll Cardiol* **71**, 620-632 (2018).

33. Onderwater, G.L.J.*, et al.* Large-scale plasma metabolome analysis reveals alterations in HDL metabolism in migraine. *Neurology* **92**, e1899-e1911 (2019).

34. Struja, T.*, et al.* Metabolomics for Prediction of Relapse in Graves' Disease: Observational Pilot Study. *Front Endocrinol (Lausanne)* **9**, 623 (2018).

35. Deelen, J.*, et al.* A metabolic profile of all-cause mortality risk identified in an observational study of 44,168 individuals. *Nat Commun* **10**, 3346 (2019).

36. Fischer, K.*, et al.* Biomarker profiling by nuclear magnetic resonance spectroscopy for the prediction of all-cause mortality: an observational study of 17,345 persons. *PLoS Med* **11**, e1001606 (2014).

37. Teslovich, T.M.*, et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-713 (2010).

38. Bajaj, J.S.*, et al.* Proton Pump Inhibitor Initiation and Withdrawal affects Gut Microbiota and Readmission Risk in Cirrhosis. *Am J Gastroenterol* **113**, 1177-1186 (2018).

39. Imhann, F.*, et al.* Proton pump inhibitors affect the gut microbiome. *Gut* **65**, 740-748 (2016).

40. Llorente, C.*, et al.* Gastric acid suppression promotes alcoholic liver disease by inducing overgrowth of intestinal Enterococcus. *Nat Commun* **8**, 837 (2017).

41. Jackson, M.A.*, et al.* Proton pump inhibitors alter the composition of the gut microbiota. *Gut* **65**, 749-756 (2016).

42. Liu, R., *et al.* Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nat Med* **23**, 859-868 (2017).
43. Pedersen, H.K., *et al.* Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535**, 376-381 (2016).
44. Kontush, A. HDL particle number and size as predictors of cardiovascular disease. *Front Pharmacol* **6**, 218 (2015).
45. Ahola-Olli, A.V., *et al.* Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish cohorts. *BioRxiv*, 513648 (2019).
46. Mitchell, A.J., Vaze, A. & Rao, S. Clinical diagnosis of depression in primary care: a meta-analysis. *Lancet* **374**, 609-619 (2009).
47. Bajaj, J.S., *et al.* Systems biology analysis of omeprazole therapy in cirrhosis demonstrates significant shifts in gut microbiota composition and function. *Am J Physiol Gastrointest Liver Physiol* **307**, G951-957 (2014).
48. Hoyles, L., *et al.* Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. *Nat Med* **24**, 1070-1080 (2018).
49. Qin, N., *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59-64 (2014).
50. Bates, C., Adams, W. & Handschumacher, R. Control of the formation of uridine diphospho-N-acetyl-hexosamine and glycoprotein synthesis in rat liver. *Journal of Biological Chemistry* **241**, 1705-1712 (1966).
51. Kettunen, J., *et al.* Biomarker Glycoprotein Acetyls Is Associated With the Risk of a Wide Spectrum of Incident Diseases and Stratifies Mortality Risk in Angiography Patients. *Circ Genom Precis Med* **11**, e002234 (2018).
52. Akinkuolie, A.O., Buring, J.E., Ridker, P.M. & Mora, S. A novel protein glycan biomarker and future cardiovascular disease events. *J Am Heart Assoc* **3**, e001221 (2014).
53. Akinkuolie, A.O., Pradhan, A.D., Buring, J.E., Ridker, P.M. & Mora, S. Novel protein glycan side-chain biomarker and risk of incident type 2 diabetes mellitus. *Arterioscler Thromb Vasc Biol* **35**, 1544-1550 (2015).
54. Ikram, M.A., *et al.* The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol* **32**, 807-850 (2017).
55. Boomsma, D.I., *et al.* Netherlands Twin Register: from twins to twin families. *Twin Res Hum Genet* **9**, 849-857 (2006).
56. Penninx, B.W., *et al.* The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int J Methods Psychiatr Res* **17**, 121-140 (2008).
57. Schoenmaker, M., *et al.* Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *Eur J Hum Genet* **14**, 79-84 (2006).
58. Tigchelaar, E.F., *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
59. van der Heijden, A.A., *et al.* The Hoorn Diabetes Care System (DCS) cohort. A prospective cohort of persons with type 2 diabetes treated in primary care in the Netherlands. *BMJ Open* **7**, e015599 (2017).
60. Geleijnse, J.M., *et al.* Effect of low doses of n-3 fatty acids on cardiovascular diseases in 4,837 post-myocardial infarction patients: design and baseline characteristics of the Alpha Omega Trial. *Am Heart J* **159**, 539-546 e532 (2010).
61. Schram, M.T., *et al.* The Maastricht Study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. *Eur J Epidemiol* **29**, 439-451 (2014).
62. Sayed-Tabatabaei, F.A., *et al.* Heritability of the function and structure of the arterial wall: findings of the Erasmus Rucphen Family (ERF) study. *Stroke* **36**, 2351-2356 (2005).
63. van Oosterhout, W.P., *et al.* Validation of the web-based LUMINA questionnaire for recruiting large cohorts of migraineurs. *Cephalalgia* **31**, 1359-1367 (2011).
64. de Mutsert, R., *et al.* The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. *Eur J Epidemiol* **28**, 513-523 (2013).
65. Soininen, P., Kangas, A.J., Wurtz, P., Suna, T. & Ala-Korpela, M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet* **8**, 192-206 (2015).
66. Inouye, M., *et al.* Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol Syst Biol* **6**, 441 (2010).

67. van den Akker, E.*, et al.* Predicting biological age based on the BBMRI-NL 1H-NMR metabolomics repository. *bioRxiv*, 632919 (2019).

68. Sturm, R. The effects of obesity, smoking, and drinking on medical problems and costs. *Health Aff (Millwood)* **21**, 245-253 (2002).

69. Van Gaal, L.F., Mertens, I.L. & De Block, C.E. Mechanisms linking obesity with cardiovascular disease. *Nature* **444**, 875-880 (2006).

70. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* **95**, 221-227 (2005).

71. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* **36**, 1-48 (2010).

72. Lawlor, D.A., Harbord, R.M., Sterne, J.A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* **27**, 1133-1163 (2008).

73. lab, N. Details and considerations of the UK Biobank GWAS, available 10 Dec, 2018. *http://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-the-uk-biobank-gwas* (2017).

74. Wray, N.R.*, et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* **50**, 668-681 (2018).

75. Hemani, G.*, et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**(2018).

76. National Cholesterol Education Program Expert Panel on Detection, E. & Treatment of High Blood Cholesterol in, A. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* **106**, 3143-3421 (2002).

77. Bot, M.*, et al.* Metabolomics profile in depression: a pooled analysis of 230 metabolic markers in 5,283 cases with depression and 10,145 controls. *Biological Psychiatry* (2019).

78. Koehler, E.M.*, et al.* Presence of diabetes mellitus and steatosis is associated with liver stiffness in a general population: The Rotterdam study. *Hepatology* **63**, 138-147 (2016).

79. Hamaguchi, M.*, et al.* The severity of ultrasonographic findings in nonalcoholic fatty liver disease reflects the metabolic syndrome and visceral fat accumulation. *Am J Gastroenterol* **102**, 2708-2715 (2007).

80. Dina Vojinovic*, et al.* Relationship between gut microbiota and circulating metabolites in population-based cohorts. *Nature communications* **In press**(2019).

Chapter 5

General discussion

The aim of this thesis is to elucidate the pathogenesis and pathological progression of type 2 diabetes (T2D) and related traits including blood pressure, glucose and insulin metabolism. I have used a broad multiple molecular (omics) approaches, including genomics, epigenomics, transcriptomics and metabolomics. Building upon classical epidemiological and genetic approaches, I have explored a new avenue to integrate data from multi-omics levels for their use in prediction of future cases, as well as for their use in understanding the biology behind the disease. During the course of my PhD, it became evident that medication use is a major confounder in observational omics studies, with the exception of genomics. Without a doubt, drugs may confound omics studies. Therefore, as part of my thesis, I decided to develop a comprehensive atlas on drug and circulating metabolites association. This chapter summarizes the main findings of the thesis, discusses the implication towards the understanding of molecular processes and pathways and makes comments with regards to future research.

Findings of this thesis

Epigenomics of glucose and insulin homeostasis

In **Chapter 2**, I study epigenetic effects related to glucose and insulin metabolism. Historically, epigenetics is considered to be the result of environmental interactions with the genome. However, thanks to methylation quantitative trait loci (meQTL) studies, the evidence is increasing that a large part of these effects is also under strong control from genetics. **Chapter 2** focuses on identifying the DNA methylation sites associated with two endophenotypes of T2D, glucose and insulin homeostasis, and accounts for obesity. The chapter integrates the epigenomics findings by cross-omics analysis including a wealth of genomic and transcriptomic-based data. Although some previous studies reported differential DNA methylation in T2D and obesity[1-5], our understanding of the role of differential methylation as a causal factor of T2D is very limited. A case study of a single individual whose global methylation pattern was measured at 57-time points spanning 36 months reported changes in methylation preceding changes in glucose levels[6]. In addition, there is no evidence-based data to answer the question if the association of differential methylation and glucose and/or insulin metabolism is an irrelevant epiphenomenon that is related to obesity, the major determinant of T2D. Obesity has been associated with 187 DNA methylation sites, including 62 of them which are also associated with T2D[7]. The disturbances in the obesity-related DNA methylation could also predict future type-2 diabetes[7]. The earlier epigenome-wide association study (EWAS) research on T2D case-control status[2,8,9], established only one CpG locus in *TXNIP* replicated multiple times, and the association is independent of BMI[2,8,9]. Along with limitation of power as a function of sample size, lack of replicable findings can be attributable to confounding factors namely obesity, medication and comorbidity among the cases which in theory can have a strong influence on differential methylation.

As BMI is a major driver of both epigenetics and T2D[7], I performed a blood-based epigenome-wide association study (GWAS) of fasting glucose and insulin, with and without body mass index (BMI) adjustment. The study was conducted in the context of the Biobanking and BioMolecular Infrastructure of the Netherlands (BBMRI-NL), in which 4,084 participants of three studies were characterized for the Illumina© Human Methylation450 array and included in the discovery phase of the current study, and a total of 3,841 participants underwent RNA sequencing and were included in the previous studies[10,11]. For all participants genome-wide data for common genetic variants was also available. Meanwhile, the T2D working group of the Cohorts for Heart and Aging Research in Genomic Epidemiolog*y* (CHARGE) served as a control including 11,750 participants. To exclude that the epigenetic effects were the consequence of the T2D or its treatment, analysts from each center studied only participants without diabetes. By meta-analysing the discovery cohorts and replicating the findings, I identified and successfully replicated nine CpGs not implicated in T2D and/or related traits previously and validated 11 known loci. The newly identified CpGs in *LETM1*, *RBM20*, *IRS2*, *MAN2A2* and the 1q25.3 region are associated with fasting insulin, the ones in *FCRL6*, *SLAMF1*, *APOBEC3H* and the 15q26.1 region are with fasting glucose. Three CpG sites (in *SLAMF1*, *APOBEC3H* and the 15q26.1 region) were associated with fasting glucose only after adjustment for BMI.

Overall, obesity has been associated with over 11 out of 20 identified CpGs, remarkably all associated with insulin. After a closer look at these associations, I found that differential methylation of the insulin-related CpG sites explained up to 16.9% of the association between obesity and insulin levels. These findings are in line with the report on the EWAS of BMI that found that the methylation patterns in blood predict future diabetes[8] and suggest that BMI may be the driver of differential methylation which subsequently raises insulin level in the circulation. Meanwhile, the fasting glucose-related methylation loci are either enriched in immune function and innate immune response (*SLAMF1* and *APOBC3H*) or have genetic determinants in the immune-related HLA region (*FCRL6*, 15q26.1 and *SLAMF1*). These findings provide us with strong epigenetic signals of the immune system on glucose metabolism in the population of non-diabetic individuals.

Of further note is that the well-established locus, *TXNIP*, which has been repeatedly associated with T2D in multiple EWAS researches[2,8,9]was also associated with the fasting glucose in the non-diabetic population (P-value = 7.6 × 10^{-7} in the BMI adjustment model) in our discovery setting, indicating that the differential methylation in *TXNIP* already takes place before the onset of diabetes. Although we have conducted the largest study to date and have doubled the number of CpGs, the power of the present study is still limited. At present, we are collaborating on an analysis of the CHARGE consortium. This analysis includes 13,500 non-diabetes participants and has identified 43 CpGs associated with fasting glucose and 89 associated with fasting insulin, suggesting that we are not facing the end of the discovery of epigenetic loci implicated in T2D.

It is also of note that tissue-specific experiments are important for the omics analysis but hard to achieve. In **Chapter 2**, I systematically integrated the multi-omics (genomics, epigenomics and transcriptomics) based on the selected glycaemic related CpGs. However, most of the data was based on the concentration in blood circulation. Blood is not the main relevant tissue involved in the pathogenesis of T2D but is the easiest sample to obtain. Given consideration for the relatively high statistic power in large-scale population, and the correlation of CpG levels and gene expression between blood and tissues, using blood data to screen the target loci and then validating them in a smaller sample size with specific-tissue-based experiments can be a cost-effective method to find targets and investigate the mechanism of T2D. I thus call for future studies in tissue-specific methylation and expression to fill in the current integration framework in **Chapter 2**.

Metabolomics in T2D

In **Chapter 3**, I focused on the associations of metabolites and T2D or its endophenotypes and directionality of possible causal effects. To start with, in **Chapter 3.1**, I investigated the predictive performance of metabolomics on future T2D and compared it with the well-known predictor, fasting glucose, in the whole population, but also in subgroups stratified by the traditional risk factors of T2D. I analysed 261 metabolites in 2,776 participants from the Erasmus Rucphen Family study (ERF) in the baseline and 1,571 diabetes-free individuals who were followed for 14-years. Using the Least Absolute Shrinkage and Selection Operator (LASSO) regression, I selected 24 metabolites, i.e. high-density lipoprotein (HDL), low-density lipoprotein (LDL) and very-low-density lipoprotein (VLDL) sub-fractions, certain triglycerides, amino acids, and small intermediate compounds, as the T2D markers in the baseline data. These markers together can predict future T2D independently with an area under the curve (AUC) of 0.81. Together with the traditional risk factor and previous reported BCAA, the AUC can be maximized to 0.89. I also found that in the young (age < 50 years) and lean (BMI < 25 kg/m^2) female group of participants, the predictive performance of the metabolites is significantly better than fasting glucose compared with the old and overweight male groups. In **Chapter 3.1**, I generated a novel prediction model that increases the long-term predictive performance, and brings a higher resolution over the complexity of the lipoprotein component. I provided evidence that metabolomics has better utility for T2D prediction specifically in the young, the female and the lean population with the stratification analysis based on age, gender and overweight status. Although our knowledge of the metabolic signature of T2D is far from complete, **Chapter 3.1** suggests that metabolomics may help to address the pathways underlying metabolically unhealthy normal weight (MUHNW) and metabolically healthy obese (MHO) phenotypes[12]. Recent research from the PREDIMED cohort study[13] also reported improved predictive ability by a selection of metabolites for high HOMA-IR levels after 1, 2 and 3-year follow-up with a maximum AUC of 0.81 when compared to classical risk factors alone (AUC of 0.69). The same metabolite risk score associated with significantly increased T2D risk with a hazard ratio of 2.00 for each SD of the multi-metabolite risk score in the analysis of 152 incident cases and 548 control participants (median follow-up: 3.8 years)[13].

In **Chapter 3.2**, I explored the evidence for causation of the association between metabolites and T2D or fasting glucose. To this end, I conducted a bi-directional two-sample Mendelian randomization (MR) study. As described in **Chapter 1** and **Chapter 3.2**, MR uses the additive effects of genetic variants as instrumental variables for exposure. The basic rationale is that if exposure is causally related to T2D, the genetic variants that are associated with the exposure (as instrumental variables) should also show association to the disease. We tested the following hypothesis:

1. Do the genes that determine the metabolite levels in the blood circulation also associate with T2D, implying these metabolites are causally associated with T2D.
2. Do the genes that determine T2D risk also determine circulating metabolites, implying that the metabolites are a consequence of the disease process or progression (or medication).

Although MR has been successfully used for elucidating directionality in several disorders [14-17], a major assumption is that the genetic variants do not have pleiotropic effects. As described in **Chapter 1,** this is quite an assumption for complex disorders such as T2D, which is driven by common genetic variants which are typically involved in multiple processes and thus have pleiotropic effects. In **Chapter 3.2,** I therefore carefully evaluated pleiotropic effects based on the biological pathway-based sensitivity analysis. I detected 14 candidate causal associations: ten metabolites influencing fasting glucose, one influencing T2D and three influenced by T2D. Firstly, genetically increased cholesterol, free cholesterol and phospholipid content of circulating XL-HDL and L-HDL particles together with XS-VLDL-phospholipids associated with decreased plasma glucose level. Secondly, triglyceride content of S-HDL and S-VLDL particles as well as total plasma triglycerides increase plasma glucose. Thirdly, total triglycerides increase the risk of T2D. Finally, genetic variants implicated in T2D are found to be associated with lower levels of two alkyl-acyl phosphatidylcholines (PC) and a higher level of alanine, suggesting the change in levels of these metabolites are a consequence of the disease process. The findings add information to the previous MR results that increased LDL-cholesterol, HDL-cholesterol, and possibly TG levels are associated with a lower risk of diabetes[18,19]. **Chapter 3.2** shows that a higher resolution metabolic phenotypes can be achieved in a cost-effective way and demonstrates that the decreasing effect of HDL-cholesterol on fasting glucose is specific to the L-HDL or XL-HDL subclasses since the effect of genetic risk scores of L-HDL or XL-HDL subclasses (cholesterol, free cholesterol, phospholipids) have larger weighted effect than the HDL-cholesterol, while no significant negative association was found with M-HDL, S-HDL or XS-HDL particles. These results advocate for further metabolic studies in order to improve our understanding of the pathogenesis of T2D and glucose metabolism.

Hypertension is a key feature of metabolic syndrome (**Chapter 1**), which is strongly related to the obesity epidemic and one of the main drivers of cardiovascular disease (see **Chapter 1**). While there have been numerous studies of the metabolomics in obesity, type 2 diabetes and cardiovascular disease[20,21], a few have studied blood pressure and hypertension, especially for the detailed characterized phospholipids. In **Chapter 3.3**, I examined the association of the plasma phospholipids that are characterized with systolic and diastolic blood pressure. I discovered and replicated five associations between metabolites and systolic blood pressure (PC 32:1, PC 40:5, phosphatidylethanolamines (PE) 38:3, PE 40:5, and PE 40:6), as well as six associations between metabolites and diastolic blood pressure (PE 38:3, PE:38:4, PE38:6, PE 40:4, PE:40:5 and PE 40:6). Ten of the associations were validated in a third sample by data-mining. The multi-metabolic blood pressure profile only added a small increase on top of the traditional risk score for incident hypertension. However, from a biological point of view as tested by MR, blood pressure seems to be causal in elevating the level of a distinct fatty acid type phospholipids PE 40:5, yet the pathways responsible for this effect still need to be understood. The multi-metabolomic blood pressure profile clustered with both triglycerides and other cardiometabolic traits, and are likely to share genetic determinants of other lipoproteins, blood cell counts and pulse rate. The associations with medication use is interesting, however needs to be separately investigated in a larger omics framework which we established in **Chapter 4** of this thesis.

Drug-metabolite atlas

One of the major problems encountered in observational clinical and epidemiological metabolic studies is the use of medication by the study participants. The drug may be prescribed in relation to the disease under study, to a related disorder, e.g.

patients with diabetes or hypertension are often co-medicated to dyslipidaemia, or co-occurring disease, e.g. elder patients often suffer from multi-morbidity. Within the BBMRI Metabolomics Consortium[22], confounding by unknown medication was identified as a major issue. The BBMRI Metabolomics Consortium has conducted a joint metabolomics study involving 25 cohorts and over 25,000 participants. As comprehensive data involving all medication used by a participant is not always available in the cohorts studied, joint analysis of data was hampered by the fact that results may be confounded by medication not registered by cohorts.

To overcome this problem, in **Chapter 4**, I have constructed an atlas of the metabolic effects of drugs prescribed in T2D and other disorders. I uncovered 1,071 drug-metabolite associations after evaluating confounding including age, sex, BMI, smoking and co-treatment, covering a wide range of drug-metabolite associations which were not studied before. As the atlas is based on well-characterized cohorts, I was able to adjust for confounders of the drug-metabolite association. I found that BMI and smoking accounted for 21.4% of the significant associations between the drug and the metabolites, indicating that these two factors should be considered when studying such associations.

The major problem of developing a drug-metabolite atlas is that the target disease for which the drug is prescribed may explain the drug association. For instance, up to 100% of the metabolic associations with sulfonamides-urea derivatives and 84.1% with metformin are partially or fully explained by T2D. From the previous T2D study in BBMRI Metabolomics Consortium, 24 of the 26 metabolites significantly associated with either glucose-lowering drug in the T2D patients are also associated with HbA1c level[23]. The study also reported that the metabolic associations with insufficient glycaemic control were similar between different treatment groups[23]. However, if the association of metabolites and HbA1c/insufficient glycaemic control is explained by T2D or by the drug effects cannot be concluded in their study.

In addition to being a toolbox to explore drug metabolites associations, the atlas can also be used to explore on-target and off-target effects of drugs. However, the relationships in the atlas are based on cross-sectional data. An important finding of **Chapter 4** is that the statin findings in a large-scale cross-sectional study are very similar to that of longitudinal study for pharmacometabolomics research. More importantly, we could benchmark the similarity of the study design by using MR. This finding in statins does not warrant that we can extrapolate the similar associations between cross-sectional data and longitudinal data to other drugs. But, our results at least show that cross-sectional atlas data can yield information on future drug effects, which can be a starting point for proper prospective studies and trials. As such, the atlas opens avenues to the pharmacometabolomics studies in large epidemiological biobanks.

Exploring off-target effects, I also found a similar metabolic association pattern among proton pump inhibitors (PPIs), certain biochemical liver-parameters, hepatic steatosis, diversity of gut microbiome and specific gut microbiota. Again, this yields a starting point for future research on the mechanism through which the metabolites associated with PPI are associated with liver function, hepatic steatosis and gut microbiome. To make the data available to the wider community, we have created an analysis tool via a web browser http://bbmri.researchlumc.nl/atlas/.

Challenges and future developments

Challenges in large biobank studies

My thesis has been one of the first unique examples of integration of biobank data with data mining in genetic epidemiology, going beyond the golden standard "meta-analysis-discovery-replication-validation" path. By the help of new biobanks and unique omics datasets emerging, exponentially increasing a number of similar research studies in epidemiology is inevitable. One of the growing biobank resources, BBMRI-NL, has been founded with the mission to maximize the use of biosamples, images and data for health research on the prevention, diagnosis and treatment of diseases[24]. BBMRI-NL has provided researchers access to biosamples, images and data, tools to capture, integrate and analyse data, as well as support on ethical, legal and societal implications. Another remarkable biobank that has recently emerged is the UK Biobank, which has been a pillar in modern epidemiological research not only by its sample size (N = 500,000) but also by the accessibility of its wide-range of phenotypes, including questionnaires (diet, cognitive function, work history and digestive health), image (brain, heart, abdomen, bones & carotid artery), electronic health records (cancer, death, hospital episodes, general practice), blood biochemistry and genetic data[25]. Moreover, the GWAS of selected lists of traits from the UK Biobank which were analysed by independent research groups and summary statistic files were then deposited in public repositories such as PheWeb[26] thus allowing me to perform phenome-wide association studies (PheWAS), exploring which phenotypes are associated with a genetic variant (SNP). However, the amount of readily analysed/collected data comes with a price. One and perhaps the only major problem with these types of publicly available bulk amount of data is the quality of the phenotypic data collected/analysed. Although a standard genotyping quality check (QC) would be sufficient to filter out outlying individuals, detect contaminated samples and unknown familial relatedness, for the phenotypic data especially when analysed according to "one solution fits all" manner, no such quality criterion exists. This is in line with the fact that we are living in a decade where scientific activities produce more data than can be QCed and analysed by qualified researchers. This fact, therefore, emerges as a growing challenge in epidemiological research that currently has no solution.

Opportunities and challenges in MR

Since MR methodology was published by *Smith, et al.* in 2003[27], it has provided an exciting promise for epidemiological studies of gene-disease associations especially with the continuing success of large scale GWAS in identifying robust genetic associations and development of multiple MR methods utilizing GWAS summary data. These multi-instrument MR methods aggregate estimates from multiple instrumental variables, testing for a causal relationship between a given exposure and outcome in a linear regression framework in which the variants' effects on the outcome are regressed on the same variants' effects on the exposure[28,29].

In the last five years, using keywords (Mendelian randomization) OR 'Mendelian randomization' while searching in PubMed, the number of MR papers has increased from 63 in 2014 to 369 in 2018. By July 2019, the number has reached 304 (Figure 1). An automated MR, called MR-base with a web interface was even developed for non-specialists[30]. MR-base uses a strategy known as two-sample Mendelian randomization (2SMR), bypassing the need for individual-level data. The software includes several sensitivity analyses for assessing the impact of horizontal pleiotropy and other violations of assumptions[30]. Ruling out horizontal pleiotropy is the fundamental assumption of MR which requires that the instrumental variable acts on the target outcome *exclusively* through the exposure[31]. Horizontal pleiotropy occurs when the variant has an effect on other traits outside of the pathway of the impact of exposure on the target outcome, or when the variant has a direct effect on the target outcome[32]. Horizontal pleiotropy can lead to inaccurate causal estimates, loss of statistical power and potential false-positive causal relationships. There are several methods to detect this[33]: the MR-PRESSO global test, Q test[34,35], Q (modified) test[36], Q' test[34,35], Q' (modified) test[36] and several methods to correct for it, including MR-Egger regression[37,38] and multi-variable MR[39,40]. However, PheWAS and genetic correlation results of large epidemiological studies proved that pleiotropy is the rule rather than the exception and accounted for over 48% of significant causal relationships in MR[32].

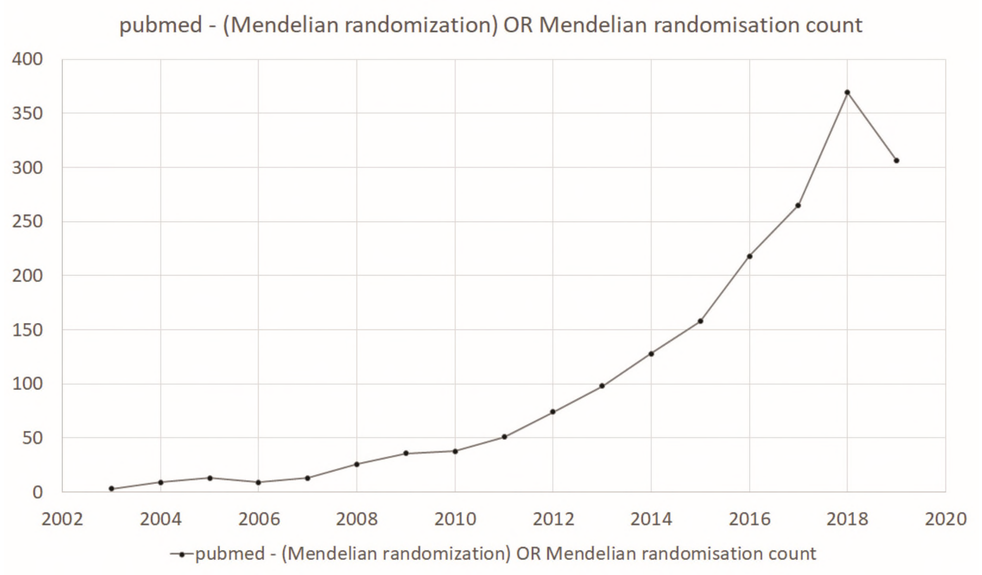


Figure 1 Trend of the papers on MR in PubMed.

Altogether, automated MR results should be interpreted with caution after careful consideration of conditions of MR as we mentioned earlier in **Chapter 1**. For instance, some of the associations between metabolites and fasting glucose in the MR-base have inconsistent results when compared with the results from our study presented in **Chapter 2.3**. The knowledge of gene function and pleiotropic effects are limited, and the association of SNPs in the instrument and confounders between exposure and outcome is not always very clear.

Meanwhile, MR is often described as naturally occurring randomized trials as the genetic factors are randomly assigned by nature[41]. But the frequently made analogy between MR studies and randomized trails also has key threats which should be considered during the MR study design. These threats include: (1) the population stratification and linkage disequilibrium make the case and control groups not as exchangeable as in the randomized trials, (2) the MR studies have no clear definition of the time for setting eligibility criteria as in the randomized trials, (3) the genetic variant is not always causal indicating that the measurement of treatment is not always correct, (4) the “adherence” in the MR is not clear, for instance, not all the persons with the mutation of FTO gene get obesity.

The development of MR methodology is challenging, but it is the preferred epidemiological source for estimating the causation of traits when the trial data is not available, such as in metabolomics and proteomics. MR opens up opportunities to investigate the causation in and between the omics data as there is a lot that is known of the genomic drivers of metabolome and proteome. However, a weak instrument may lead to a false negative association in the MR studies, especially when the identified genetic associations of metabolome and proteome are limited and only explain a small variance of the exposure because of the small sample size function in the GWAS. Some of the metabolite GWAS have reached a sample size as large as 25,000[42] in one platform to 80,000[43] in combined multiple platforms. However, the sample size of the GWAS of the proteome is much lower. The largest one was done in 8,293 Finns of 48 cytokines[44]. High-throughput, multiplex and high sensitivity proteomics technology has developed fast in the last five years. For instance, SOMAScan® array can measure 5,000 proteins across eight orders of magnitude of abundance and converts protein measurements as straightforward as DNA measurement[45]. On the other hand, Olink® can measure 350 human proteins[46]. These proteomics platforms have been used in large epidemiological studies such as INTERVAL[47], the Cooperative Health Research in the Region of Augsburg (KORA) F4 study[48] and Framingham Heart Study (FHS)[48], suggesting a bright future in identifing genetic drivers of high resolution proteins while supporting less weak instrument in the MR studies of proteomics.

Tissue specificity of methylome and transcriptome

One of the scientific challenges we face today in a large biobank is the choice of tissue for the omics research to be conducted. This is perhaps a lesser problem for metabolomics as liver, the metabolite conversion warehouse of the body, is directly connected to the bloodstream, as compared to the gene methylation and transcriptomics studies. This is a major challenge for the interpretation of transcriptomics and methylation studies as gene expression and methylation are not only age and environment-dependent, but also tissue-specific. On one hand, the most pragmatic solution is to assay these type of omics data and to use accessible tissue biobanks such as circulating white blood cells, maximizing the sample size for all cohorts, as it is currently done today. On the other hand, the research results are more difficult to interpret in the biological context when the involved tissue is other than blood such as brain or liver. While GTEX project database which provides gene expression profiles from several tissues and related genetically regulated expression prediction tools, can predict the genetic component of variance in human transcriptome[49,50], the methods have the obvious limitation in that they are independent of aging, gender and environmental factors which are also contributors of diseases. Therefore, such predictions would be weak when modelling the effect of environmental factors on diseases; in such scenarios, environment affects the gene expression which in turn leads to disease. Additionally, to our knowledge, the same scale training set for genome-wide methylation does not exists, hampering research of tissue-specific methylation patterns in different outcomes. Methods harvesting omics information from single cells are also very promising in terms of biological validity. However, how they will be placed in a population-based biobanking is a good question as these types of trends obviously will come with their own technical challenges in both material collection and quality control as well as price.

Metabolomics platforms to date

Ten years after the first report on population-based metabolomics research[51], a handful of platforms are validated and widely used in biobanks, utilising either Mass-Spectrometry (MS) or Nuclear Magnetic Resonance (NMR) Spectrometry. The well-knowns among them include (1) detailed analysis of circulating lipoprotein particles as developed by Bruker® and Nightingale® companies, (2) phospholipids and carnitine assays that were mainly developed by Biocrates® but also by numerous in-house frameworks, (3) and finally, the so-called "complete metabolomic" assays, initially developed by Metabolon® which focuses on the products of enzymatic reactions. Although all three types of platforms significantly contributed to gene discovery by their GWASs[42,52-55], their value in disease prediction and aetiology research has been particularly limited to certain disease groups. For example, the lipoprotein particle analysis has great potential to replace the classical cholesterol measurements in the clinic and complement routine biochemistry tests leading to improvement in precision medicine[56]. On the other hand, platforms enriched in membrane lipid particles have potential for neurologic and psychiatric phenotypes such as Alzheimer disease and depressive disorders[57,58]. Although their predictive and diagnostic ability can be high, these platforms measure rather complex molecules which were technologically defined rather than biologically and which makes it very difficult to plug them into enzymatic pathways and elementary flux models. The third and final platform advancing MS can profile up to 1,000 metabolites with less than 5% process variability and has been a success[59] but the availability for biobanks is rather low due to the high costs of the experiments.

Microbiome

My personal interest in the microbiome and its relationship to metabolites in the circulation and organs and their effects on health has grown. In particular, the gut microbiome is large and includes more than 10^{14} bacteria that reside in the gut. It has even larger genetic material than human genome and is referred to as the second genome of human body[60]. As a fast-growing field, the human microbiome is enormously complex. Genome is static, while microbiome is dynamic. The composition of microbiota during the course of a human life is different and dynamic, depending on age, birth and development, diet, disease and other environmental factors. Some previous studies on gut microbiome and T2D or glucose have reported various potential mechanisms of T2D, ranging from endocrine and metabolic pathways to mechanisms on a cellular and genetic level through gut microbiome[61]. Well-designed targeted diet can regulate the structure of gut microbiome[62] and supplementation of beneficial bacteria such as *Akkermansia muciniphila* was shown to decrease insulin resistance and cause weight loss in obese volunteers[63]. Although the expectations from gut microbiota research have been very high for cardiometabolic diseases, population-based meta-analysis studies suffer from inconsistent association and the replication rate is very low[64]. This is due to the very heterogeneous nature and per-individual complexity of the gut microbiome under continuous control of the exposome. Unexplored areas such as microbiome and genome-wide methylation/transcriptome in time-matched samples will help to elucidate the exposomal pathways in which human gut microbiota take place. However, strong confounding factors such as medication and diet need to be controlled, requiring well-characterized populations. Recent MR research on microbiome uncovered some short fatty acid chains related to causal effects between the microbiome and metabolic diseases[65], but overall MR of microbiome is still underpowered due to the weak instrumental variables for the microbiota exposure.

Concluding Remarks

In this thesis, I have used multi-omics data to provide insight into the pathophysiology of T2D and its related traits and therapeutics information. I have integrated cross-omics to discover novel biomarkers and pathways of T2D, elucidate the causality of them and link the information to therapeutics. As the high-throughput era in omics is progressing, integration of multi-omics will be much more popular and effective, yielding insights into the pathophysiology, prevention and treatment of T2D and related disorders.

References

1. Hidalgo, B., *et al.* Epigenome-wide association study of fasting measures of glucose, insulin, and HOMA-IR in the Genetics of Lipid Lowering Drugs and Diet Network study. *Diabetes* **63**, 801-807 (2014).
2. Chambers, J.C., *et al.* Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *Lancet Diabetes Endocrinol* **3**, 526-534 (2015).
3. Kriebel, J., *et al.* Association between DNA Methylation in Whole Blood and Measures of Glucose Metabolism: KORA F4 Study. *PLoS One* **11**, e0152314 (2016).

4. Kulkarni, H., *et al.* Novel epigenetic determinants of type 2 diabetes in Mexican-American families. *Hum Mol Genet* **24**, 5330-5344 (2015).
5. Ding, J., *et al.* Alterations of a Cellular Cholesterol Metabolism Network Are a Molecular Feature of Obesity-Related Type 2 Diabetes and Cardiovascular Disease. *Diabetes* **64**, 3464-3474 (2015).
6. Chen, R., *et al.* Longitudinal personal DNA methylome dynamics in a human with a chronic condition. *Nat Med* (2018).
7. Wahl, S., *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81-86 (2017).
8. Soriano-Tarraga, C., *et al.* Epigenome-wide association study identifies TXNIP gene associated with type 2 diabetes mellitus and sustained hyperglycemia. *Hum Mol Genet* **25**, 609-619 (2016).
9. Florath, I., *et al.* Type 2 diabetes and leucocyte DNA methylation: an epigenome-wide association study in over 1,500 older adults. *Diabetologia* **59**, 130-138 (2016).
10. Bonder, M.J., *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet* **49**, 131-138 (2017).
11. Zhernakova, D.V., *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet* **49**, 139-145 (2017).
12. Mathew, H., Farr, O.M. & Mantzoros, C.S. Metabolic health and weight: Understanding metabolically unhealthy normal weight or metabolically healthy obese patients. *Metabolism* **65**, 73-80 (2016).
13. Papandreou, C., *et al.* Plasma metabolites predict both insulin resistance and incident type 2 diabetes: a metabolomics approach within the Prevencion con Dieta Mediterranea (PREDIMED) study. *Am J Clin Nutr* **109**, 626-634 (2019).
14. Millwood, I.Y., *et al.* Conventional and genetic evidence on alcohol and vascular disease aetiology: a prospective study of 500 000 men and women in China. *Lancet* **393**, 1831-1842 (2019).
15. Millwood, I.Y., *et al.* Association of CETP Gene Variants With Risk for Vascular and Nonvascular Diseases Among Chinese Adults. *JAMA cardiology* **3**, 34-43 (2018).
16. Lu, L., *et al.* Association of vitamin D with risk of type 2 diabetes: A Mendelian randomisation study in European and Chinese adults. *PLoS Med* **15**, e1002566 (2018).
17. Sliz, E., *et al.* Metabolomic consequences of genetic inhibition of PCSK9 compared with statin treatment. *Circulation* **138**, 2499-2512 (2018).
18. White, J., *et al.* Association of Lipid Fractions With Risks for Coronary Artery Disease and Diabetes. *JAMA cardiology* **1**, 692-699 (2016).
19. Fall, T., *et al.* Using Genetic Variants to Assess the Relationship Between Circulating Lipids and Type 2 Diabetes. *Diabetes* **64**, 2676-2684 (2015).
20. Guasch-Ferre, M., *et al.* Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis. *Diabetes Care* **39**, 833-846 (2016).
21. Ussher, J.R., Elmariah, S., Gerszten, R.E. & Dyck, J.R. The Emerging Role of Metabolomics in the Diagnosis and Prognosis of Cardiovascular Disease. *J Am Coll Cardiol* **68**, 2850-2870 (2016).
22. https://www.bbmri.nl/Omics-metabolomics. *Available: 1st, July 2019*.
23. t Hart, L.M., *et al.* Blood Metabolomic Measures Associate With Present and Future Glycemic Control in Type 2 Diabetes. *J Clin Endocrinol Metab* **103**, 4569-4579 (2018).
24. https://www.bbmri.nl/. *Available: 1st, July 2019*.
25. https://www.ukbiobank.ac.uk/about-biobank-uk/. *Available: 1st, July 2019*.
26. http://pheweb.sph.umich.edu/. *Available: 1st, July 2019*.
27. Smith, G.D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**, 1-22 (2003).
28. Pierce, B.L. & Burgess, S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am J Epidemiol* **178**, 1177-1184 (2013).
29. Burgess, S., *et al.* Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol* **30**, 543-552 (2015).
30. Hemani, G., *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**(2018).
31. Ebrahim, S. & Davey Smith, G. Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Human genetics* **123**, 15-33 (2008).

32. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. & Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* **14**, 483-495 (2013).

33. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature genetics* **50**, 693 (2018).

34. Greco M, F.D., Minelli, C., Sheehan, N.A. & Thompson, J.R. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Statistics in medicine* **34**, 2926-2940 (2015).

35. Bowden, J.*, et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in medicine* **36**, 1783-1802 (2017).

36. Bowden, J.*, et al.* Improving the accuracy of two-sample summary data Mendelian randomization: moving beyond the NOME assumption. *BioRxiv*, 159442 (2018).

37. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512-525 (2015).

38. Bowden, J.*, et al.* Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I 2 statistic. *International journal of epidemiology* **45**, 1961-1974 (2016).

39. Do, R.*, et al.* Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nature genetics* **45**, 1345 (2013).

40. Corbin, L.J.*, et al.* BMI as a modifiable risk factor for type 2 diabetes: refining and understanding causal estimates using Mendelian randomization. *Diabetes* **65**, 3002-3007 (2016).

41. Swanson, S.A., Tiemeier, H., Ikram, M.A. & Hernan, M.A. Nature as a Trialist?: Deconstructing the Analogy Between Mendelian Randomization and Randomized Trials. *Epidemiology* **28**, 653-659 (2017).

42. Kettunen, J.*, et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* **7**, 11122 (2016).

43. Wittemans, L.B.L.*, et al.* Assessing the causal association of glycine with risk of cardio-metabolic diseases. *Nat Commun* **10**, 1060 (2019).

44. Ahola-Olli, A.V.*, et al.* Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *Am J Hum Genet* **100**, 40-50 (2017).

45. https://somalogic.com/technology/our-platform/. *Available: 1st, July 2019*.

46. https://www.olink.com/. *Available: 1st, July 2019*.

47. Joshi, A. & Mayr, M. In Aptamers They Trust: The Caveats of the SOMAscan Biomarker Discovery Platform from SomaLogic. *Circulation* **138**, 2482-2485 (2018).

48. Yao, C.*, et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat Commun* **9**, 3268 (2018).

49. Consortium, G.T.*, et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).

50. Barbeira, A.N.*, et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* **9**, 1825 (2018).

51. Yap, I.K.*, et al.* Metabolome-wide association study identifies multiple biomarkers that discriminate north and south Chinese populations at differing risks of cardiovascular disease: INTERMAP study. *J Proteome Res* **9**, 6647-6654 (2010).

52. Demirkan, A.*, et al.* Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. *PLoS Genet* **8**, e1002490 (2012).

53. Draisma, H.H.*, et al.* Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat Commun* **6**, 7208 (2015).

54. Shin, S.Y.*, et al.* An atlas of genetic influences on human blood metabolites. *Nat Genet* **46**, 543-550 (2014).

55. Yousri, N.A.*, et al.* Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population. *Nat Commun* **9**, 333 (2018).

56. Soininen, P., Kangas, A.J., Würtz, P., Suna, T. & Ala-Korpela, M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circulation: Genomic and Precision Medicine* **8**, 192-206 (2015).

57. Demirkan, A.*, et al.* Plasma phosphatidylcholine and sphingomyelin concentrations are associated with depression and anxiety symptoms in a Dutch family-based lipidomics study. *J Psychiatr Res* **47**, 357-362 (2013).

58. van der Lee, S.J.*, et al.* Circulating metabolites and general cognitive ability and dementia: Evidence from 11 cohort studies. *Alzheimers Dement* **14**, 707-722 (2018).

59. https://www.metabolon.com/who-we-are/blog/metabolomics-needs-precision-case-mass-spectrometry. *Available: 1st, July 2019*.

60. Zhu, B., Wang, X. & Li, L. Human gut microbiome: the second genome of human body. *Protein Cell* **1**, 718-725 (2010).

61. Hartstra, A.V., Bouter, K.E., Backhed, F. & Nieuwdorp, M. Insights into the role of the microbiome in obesity and type 2 diabetes. *Diabetes Care* **38**, 159-165 (2015).

62. Ercolini, D. & Fogliano, V. Food Design To Feed the Human Gut Microbiota. *J Agric Food Chem* **66**, 3754-3758 (2018).

63. Depommier, C.*, et al.* Supplementation with Akkermansia muciniphila in overweight and obese human volunteers: a proof-of-concept exploratory study. *Nature medicine*, 1 (2019).

64. Wang, J.*, et al.* Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative. *Microbiome* **6**, 101 (2018).

65. Sanna, S.*, et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat Genet* **51**, 600-605 (2019).

Chapter 6

Summary/Samenvatting

Summary

A substantial burden of mortality and disability worldwide is caused by diabetes and its complications, including cardiovascular diseases, retinopathy, nephropathy and neuropathy. Among the diabetic patients, 90-95% are suffering type 2 diabetes (T2D), which is age-related and determined by a combination of genetic, epigenetic and environmental factors, including overweight, obesity, diet, physical activity, smoking and other factors. Meanwhile, T2D is also highly co-occurrent with the components of metabolic syndrome, including high blood pressure, lipid dyslipidaemia and abdominal obesity. However, the molecular processes and pathways underlying T2D and these related risk factors and its endophenotypes are still not clear. Understanding the pathophysiology of T2D and glycaemic traits can help to facilitate the prevention and treatment strategies. In this thesis, I aim to provide novel insights into the pathophysiology of T2D by several omics approaches and integration of multi-omics, centering on genomics, epigenomics and metabolomics.

In **Chapter 2**, I found nine novel differential DNA methylation loci associated with fasting glucose and insulin by a large-scale epigenome-wide association study (EWAS) and its replication, which greatly complemented the previous knowledge on epigenomics of T2D. I also validated 11 loci which was previously reported. After following our pipeline of multi-omics integration, 29 genetic determinants of the glycaemic related methylation sites were identified. Two novel pathways in *FCRL6* and *SLAMF1* under the pipeline involving genomics, epigenomics, transcriptomics and fasting glucose were identified. Strong signals of adaptive immune system on glucose metabolism was surfaced in this chapter. Meanwhile, I also found that the insulin-related methylation sites would explain 17% of the association of obesity and fasting insulin.

Chapter 3 focused on the association of metabolomics and T2D or related traits. In **Chapter 3.1**, I found a combination of 24 glucose-related circulating metabolites can effectively predict the future T2D with a high predictive performance (the area under the curve (AUC) up to 0.81). Adding them to the traditional risk factors of T2D significantly improved the predictive performance to 0.89. Notably, in the low risk groups, the predictive performance of the metabolites alone is even better than the baseline fasting glucose. In **Chapter 3.2**, a bi-direction two-sample Mendelian randomization (MR) was performed to explore the causality of circulating metabolites and T2D and glucose. I detected 14 candidate causal associations: ten metabolites influencing fasting glucose, one influencing T2D and three influenced by T2D: increased total cholesterol, L-HDL-cholesterol, XL-HDL-cholesterol, L-HDL-free cholesterol, XL-HDL-free cholesterol, L-HDL-phospholipid and XL-HDL-phospholipid and XS-VLDL-phospholipids effect decreased glucose level; S-VLDL-

triglycerides, S-HDL-triglycerides and plasma triglycerides increase the fasting glucose; lower levels of two alkyl-acyl phosphatidylcholines and higher level of alanine are probably the consequence of T2D. In **Chapter 3.3**, we identified eight circulating phospholipid molecules which associated with blood pressure and have a predictive value for incident hypertension. There is a very strong association of the selected phospholipids and triglycerides, waist, waist-hip ratio, glucose, total fat percentage, BMI and leptin. The genetically increased systolic blood pressure level associate with the increase of circulating PE 40:5 and PE 38:3. The genetic variants of the selected phospholipids are also strongly linked to HDL-cholesterol, subfractions and diameter, blood cell counts (white, red and platelet) and pulse rate.

Finally, in **Chapter 4**, I built a comprehensive drug-metabolite atlas, involving on-target and off-target effects of drugs prescribed in T2D and other disorders. It remained 1071 drug-metabolite associations after evaluating the effect of age, sex, body mass index (BMI), smoking and co-treatment. With the genomics approach, I provided evidence that cross-sectional atlas data can yield information on future statin effects. I further linked proton pump inhibitors to circulating metabolites, liver function related variables, hepatic steatosis and the gut microbiome with data integration. A web-based resource for visualization of the atlas (http://bbmri.researchlumc.nl/atlas/) was also established.

Samenvatting

Een aanzienlijke sterfte- en invaliditeitslast wereldwijd wordt veroorzaakt door diabetes en de complicaties ervan, waaronder hart- en vaatziekten, retinopathie, nefropathie en neuropathie. Van de diabetespatiënten heeft 90-95% diabetes type 2 (T2D), dat leeftijdsgebonden is en wordt bepaald door een combinatie van genetische, epigenetische en omgevingsfactoren, waaronder overgewicht, obesitas, voeding, fysieke activiteit, roken en andere factoren. Ondertussen is T2D ook in hoge mate gelijktijdig aanwezig met de componenten van het metabool syndroom, waaronder hoge bloeddruk, lipidedyslipidemie en abdominale obesitas. De moleculaire processen en routes die ten grondslag liggen aan T2D en deze gerelateerde risicofactoren en de endofenotypen ervan zijn echter nog steeds niet duidelijk. Het begrijpen van de pathofysiologie van T2D en glycemische eigenschappen kan de preventie- en behandelingsstrategieën helpen vergemakkelijken. In dit proefschrift probeer ik nieuwe inzichten te verschaffen in de pathofysiologie van T2D door verschillende omics-benaderingen en integratie van multi-omics, met de nadruk op genomics, epigenomics en metabolomics.

In **Hoofdstuk 2** vond ik negen nieuwe differentiële DNA-methylatie-loci geassocieerd met nuchter glucose en insuline door een grootschalige epigenoombrede associatiestudie (EWAS) en de replicatie ervan, die in grote mate de eerdere kennis over epigenomica van T2D aanvulde. Ik valideerde ook 11 loci die eerder werden gerapporteerd. Na het volgen van onze pijplijn van multi-omics integratie, werden 29 genetische determinanten van de glycemisch gerelateerde methylatieplaatsen geïdentificeerd. Twee nieuwe routes in FCRL6 en SLAMF1 onder de pijplijn met genomica, epigenomica, transcriptomics en nuchtere glucose werden geïdentificeerd. In dit hoofdstuk zijn sterke signalen van adaptief immuunsysteem op het glucosemetabolisme naar voren gekomen. Ondertussen vond ik ook dat de aan insuline gerelateerde methylatieplaatsen 17% van de associatie van obesitas en nuchter insuline zouden verklaren.

Hoofdstuk 3 concentreerde zich op de associatie van metabolomics en T2D of gerelateerde kenmerken. In **Hoofdstuk 3.1** vond ik dat een combinatie van 24 glucosegerelateerde circulerende metabolieten de toekomstige T2D met een hoge voorspellende prestatie (het oppervlak onder de curve (AUC) tot 0,81) effectief kan voorspellen. Door ze toe te voegen aan de traditionele risicofactoren van T2D is de voorspellende prestatie aanzienlijk verbeterd tot 0,89. Met name in de groepen met een laag risico is de voorspellende werking van de metabolieten alleen nog beter dan de basislijn nuchtere glucose. In **Hoofdstuk 3.2** is een tweerichtings Mendeliaanse randomisatie (MR) in twee richtingen uitgevoerd om de causaliteit van circulerende metabolieten en T2D en

glucose te onderzoeken. Ik ontdekte 14 mogelijke causale associaties: tien metabolieten beïnvloeden nuchtere glucose, één beïnvloedde T2D en drie beïnvloedden door T2D: verhoogd totaal cholesterol, L-HDL-cholesterol, XL-HDL-cholesterol, L-HDL-vrij cholesterol, XL-HDL-vrij cholesterol, L-HDL-fosfolipide en XL-HDL-fosfolipide en XS-VLDL-fosfolipiden hebben een verlaagd glucosegehalte; S-VLDL-triglyceriden, S-HDL-triglyceriden en plasma-triglyceriden verhogen de nuchtere glucose; lagere niveaus van twee alkyl-acyl fosfatidylcholines en een hoger niveau van alanine zijn waarschijnlijk het gevolg van T2D. In **hoofdstuk 3.3** hebben we acht circulerende fosfolipidemoleculen geïdentificeerd die geassocieerd zijn met bloeddruk en een voorspellende waarde hebben voor incidente hypertensie. Er is een zeer sterke associatie van de geselecteerde fosfolipiden en triglyceriden, taille, taille-heupratio, glucose, totaal vetpercentage, BMI en leptine. Het genetisch verhoogde systolische bloeddrukniveau hangt samen met de toename van circulerend PE 40: 5 en PE 38: 3. De genetische varianten van de geselecteerde fosfolipiden zijn ook sterk gekoppeld aan HDL-cholesterol, subfracties en diameter, bloedcelaantallen (wit, rood en bloedplaatjes) en polsslag.

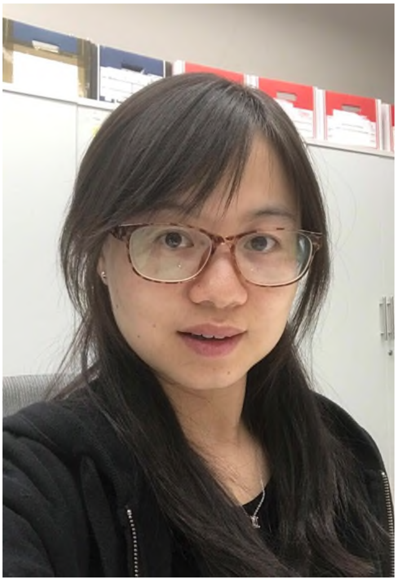
Ten slotte heb ik in **Hoofdstuk 4** een uitgebreide drug-metaboliet atlas gebouwd, met on-target en off-target effecten van medicijnen voorgeschreven in T2D en andere aandoeningen. Het bleef 1071 drug-metaboliet associaties na evaluatie van het effect van leeftijd, geslacht, body mass index (BMI), roken en co-behandeling. Met de genomics-aanpak heb ik aangetoond dat cross-sectionele atlasgegevens informatie over toekomstige statine-effecten kunnen opleveren. Verder heb ik protonpompremmers gekoppeld aan circulerende metabolieten, leverfunctie gerelateerde variabelen, hepatische steatosis en het darmmicrobioom met data-integratie. Een web-gebaseerde bron voor visualisatie van de atlas (http://bbmri.researchlumc.nl/atlas/) werd ook opgericht.

Chapter 7

Appendix

Chapter 7.4

About the author



Jun Liu (刘俊) was born in Hunan, China, on April 1991. In 2008 she completed her high school in Hunan, China and started her study in clinical medicine at Shanghai Jiao Tong University, School of Medicine, Shanghai, China, where she obtained her medical degree in 2013. The same year, she started her master research training in clinical epidemiology at Department of Endocrine and Metabolic Diseases, Rui-Jin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. From July 2014, she joined Genetic Epidemiology unit, Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands and got a full scholarship from the department for her master at Netherlands Institute of Health Sciences (NIHES) under Dr. Ayşe Demirkan's supervision. After obtaining her Master of Science degree in Genetic Epidemiology in 2016, she continued her education as a PhD candidate in the same unit and department under the supervision of Prof. Cornelia M. van Duijn. In this period, she has lead multiple projects in (inter)national consortium. She was one of the convener of the Biobanking and BioMolecular resources Research Infrastructure of the Netherlands (BBMRI-NL) medication working group. From early 2019, she becomes a convener in the Cohorts of Heart and Ageing Research in Genomic Epidemiology (CHARGE) metabolomics working group. In 2016 and 2017, as a part of Marie Curie Research and Innovation Staff Exchange, Jun spent three months at Better Value Healthcare, Oxford, UK and Linkcare Health Services, Barcelona, Spain where she worked on personalized prevention of chronic diseases. From start of 2019, Jun joined Nuffield Department of Population Health, University of Oxford, UK to finish her PhD thesis. After obtaining her PhD degree, Jun will continue working in multi-omics field at Nuffield Department of Population Health, University of Oxford, UK.