

**TECHNISCHE UNIVERSITÄT MÜNCHEN**  
Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt  
Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

**Metabolites:  
implications in type 2 diabetes and the effect of  
epigenome-wide interaction with genetic variation**

Sophie Claudia Molnos

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende: Univ.-Prof. Dr. C. C. Schön

Prüfende/-r der Dissertation: 1. Univ.-Prof. Dr. Dr. F. Theis

2. Univ.-Prof. Dr. W. Wurst

Die Dissertation wurde am 22.08.2017 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt 16.01.2018 angenommen.



---

## Danksagung (Acknowledgements)

An dieser Stelle möchte ich mich bei allen bedanken, die mich in den letzten Jahren zum Gelingen dieser Arbeit unterstützt haben.

Als erstes gilt mein besonderer Dank meinem Doktorvater Herrn Prof. Dr. Dr. Fabian Theis für die Übernahme der Betreuung an der TUM und der Ermöglichung dieser Arbeit. Außerdem danke ich dem zweiten Prüfer Herrn Prof. Dr. Wolfgang Wurst für das Lesen und Bewerten der Doktorarbeit.

Ebenso möchte ich mich ganz besonders bei Herrn Dr. Harald Grallert sowie bei Herrn Dr. Christian Gieger für ihre Unterstützung und Betreuung dieser Arbeit bedanken.

Mein besonderer Dank geht an Jennifer Kriebel und Annika Wahl für ihre unermüdliche Motivation und Korrektur der Doktorarbeit. Sie gaben mir den nötigen Druck, ohne den ich es nie geschafft hätte, diese Doktorarbeit fertig zu schreiben. Ich danke Clemens Baumbach für seine Geduld, mir das Programmieren der R-packages zu erklären, für seine Unterstützung bei der Optimierung der Codes und beim Korrigieren des Papers. Rory Wilson, Ruhi Phaltane, Carola Marzi, Kathi Schramm und Elisabeth Altmaier danke ich ebenso herzlichst für die Korrektur der Doktorarbeit. Auch möchte ich mich für die statistische Hilfe bei Rory Wilson bedanken. Ivan Kondofersky danke ich für die hilfreichen Ratschläge für meine Doktorarbeit. Mein Dank geht auch an Yakov Tsepilov und Sodbo Sharapov für die Einblicke in die verschiedenen Tools von *GenABEL*. Matthias Heinig danke ich für die Einführung in *MatrixEQTL*, der schnellsten linearen Regressionsberechnung in R, die ich durch ihn kennengelernt habe, und die ich sonst nicht zum Vergleich meines R Paketes verwendet hätte. Simone Wahl und Leen 't Hart danke ich für die Ermöglichung und Unterstützung des Projektes in IMI DIRECT.

Ich möchte mich bei meiner Familie herzlichst bedanken, die mir immer emotional zur Seite stand, bei all den Hochs und Tiefs, die ich durchlaufen habe. Insbesondere danke ich meiner großen Schwester Sonja - ohne ihr unnachgiebiges Nachhaken über meinen Stand der Doktorarbeit, ihr lockeres Rangehen an das Selbstprogrammieren - hätte ich wahrscheinlich immer noch versucht mit den bisherigen Softwares die Interaktionsanalyse zum Laufen zu bringen. Und an alle, die ich noch vergessen haben sollte: Danke!



## Contents

Danksagung (Acknowledgements) .....	I
Contents .....	III
Summary .....	VII
Zusammenfassung .....	IX
List of abbreviations .....	XI
1. Introduction .....	1
1.1. Scientific question of this thesis .....	1
1.2. Overview of this thesis .....	3
1.3. Scientific contribution .....	4
2. Background .....	7
2.1. Human metabolism in the field of common diseases – explained by genetics and epigenetics .....	8
2.2. Type 2 diabetes .....	9
2.3. Human metabolism and type 2 diabetes .....	9
2.4. Genomics .....	10
2.5. Missing heritability .....	11
2.6. Epigenomics .....	12
2.7. The interplay between DNA methylation, genetic variants, and environment .....	14
2.8. Software analyzing linear regression with interaction term .....	17
2.9. What is after GWAS? .....	18
3. Material .....	19
3.1. Study populations .....	19
3.1.1. Cooperative health research in the region of Augsburg .....	19
3.1.2. European prospective investigation into cancer and nutrition - Potsdam study .....	21
3.1.3. Leiden longevity study .....	22
3.1.4. Netherlands twin register .....	23
3.2. Metabolomic measurements and quality control .....	24
3.2.1. Biocrates platform .....	24
3.2.2. Metabolon platform .....	25

3.2.3.	Imputation .....	26
3.3.	Genotyping and quality control.....	27
3.4.	Array-based DNA methylation & quality control.....	28
3.5.	Assessment of diabetes-status .....	29
4.	Methods .....	31
4.1.	Linear regression analysis .....	32
4.1.1.	Testing hypotheses on individual regression coefficients .....	34
4.1.2.	Coefficient of determination .....	34
4.2.	Logistic regression analysis .....	35
4.3.	Survival analysis .....	36
4.3.1.	Cox proportional hazards regression analysis.....	38
4.3.2.	Time-dependent receiver operating characteristic curve .....	41
4.3.3.	Net reclassification improvement .....	45
4.4.	Confounding.....	46
4.5.	Violation of the assumption of the underlying regression model .....	46
4.5.1.	Box-Cox transformation .....	47
4.5.2.	Diagnostic plots .....	48
4.6.	Multiple testing .....	51
4.7.	Meta-analysis .....	51
4.8.	P-gain .....	52
4.9.	Pulver .....	52
4.9.1.	Linear regression with interaction term .....	53
4.9.2.	Theory underlying pulver .....	53
4.9.3.	Avoiding redundant computations.....	58
4.9.4.	Programming language and general information about the program .....	59
4.9.5.	Comparison with other R tools for running linear regressions .....	60
4.9.6.	Feature reduction .....	63
5.	Results and discussion.....	65
5.1.	T2D and metabolite ratios analysis .....	65
5.1.1.	Results.....	65
5.1.2.	Discussion.....	72

---

5.2. Interaction analysis.....	77
5.2.1. Results.....	77
5.2.2. Discussion.....	83
6. Summary and future perspectives.....	91
References.....	XIII
Appendix.....	XXIX





## Summary

Metabolomics can provide deep insights into the underlying biochemical mechanisms of diseases. Supporting this hypothesis, studies have shown that metabolic measurements in blood can reflect metabolic changes during the development of type 2 diabetes (T2D). To further investigate this relationship, the first part of this thesis addresses the association between blood metabolites and diabetes. Since ratios of metabolite concentrations can serve as proxies for enzymatic reaction rates pairwise metabolite ratios were used to identify associations with T2D. This analysis aimed to validate previous analyses studying the association between pairwise metabolite ratios and insulin response. In this study, four cohorts from the Netherlands and Germany were used to analyze the associations between incident and prevalent T2D and metabolite ratios. Out of 9,045 analyzed ratios one novel association was detected, that of the valine to phosphatidylcholine acyl-alkyl C32:2 (PC ae C32:2) ratio.

Furthermore, in the Cooperative Health Research in the Region of Augsburg (KORA) study, the performance of different models was compared by measures of the time-dependent receiver operating characteristics and net reclassification improvement. Modest improvements were observed after the addition of this metabolite ratio to a model based on a set of established risk factors.

Such variations in metabolite levels can be influenced by genetic and epigenetic changes. Thus, many genome-wide association studies (GWAS) have been conducted. However, so far only a minor part of the total metabolite variation can be explained by common sequence variants. This leads to the assumption that the analysis of multiple “omics” layers (such as the combination of genomics, epigenomics, proteomics....) – and their interactions – help to further explain variations in metabolite levels. Previous interaction analyses in GWAS mainly considered interactions between single-nucleotide polymorphisms (SNPs) and ignored possible interactions between different “omics” layers. Thus, the second part of this thesis concerns the development of a statistical tool to analyze different “omics” layers, e.g., DNA methylation and genetic variants, in a reasonable amount of time. This novel R package, called *pulver*, is the first tool to allow the computation of p-values of billions of linear regressions with an interaction term within only a few days.

The rapid running time was achieved by using the correlation coefficient to test the null-hypothesis, i.e., whether the coefficient of the interaction term significantly differs from zero. Usually, for that the time intensive computation of matrix inversion is used.

To further accelerate the run time, the order of the matrices, when to iterate through which “omics” layer, was taken into account and the tool was implemented in the fast programming language C++. Given the possible interplay between different “omics” layers in biological processes, *pulver* can be used to conduct comprehensive screenings that are beyond the capabilities of existing tools.

This R package was applied to real-world data from the KORA study, comparing metabolite levels with the interaction of genetic variants and DNA methylation sites. Hereby, one significant locus, near the *ACADS* gene, was found. Genetic and epigenetic interaction at this locus was found to influence levels of the metabolite butyrylcarnitine.

Furthermore, this R package was also applied to the previously identified ratio, valine to PC ae 32:2. However, no statistically significant associations were identified after strict correction for multiple testing.

In conclusion, this thesis provides analyses and tools to build a more holistic picture of human metabolism. This is achieved by identifying disease (T2D) related metabolite changes and by combining different “omics” layers (metabolomics, genetics, and epigenetics) to detect interactions in a more efficient way.

## Zusammenfassung

Die Metabolomik kann zum grundlegenden Verständnis biochemischer Erkrankungsmechanismen beitragen. So haben Metabolitenmessungen im Blut gezeigt, dass sie die metabolischen Änderungen während der Entwicklung zum Typ 2 Diabetes (T2D) widerspiegeln können. Aus diesem Grund behandelt der erste Teil der Dissertation die Assoziation zwischen Metaboliten im Blut und Diabetes. Da Verhältnisse der Konzentration verschiedener Metabolite auch als enzymatische Reaktionsrate dienen können, wurden paarweise Metabolitenverhältnisse verwendet, um Assoziationen zu T2D zu identifizieren. Diese Analyse wurde durchgeführt, um vorherige Analysen zu validieren, die die Assoziation zwischen paarweisen Metabolitenverhältnissen und der Insulinantwort untersuchten.

Für die Assoziationsberechnung zwischen inzidentem bzw. prävalentem T2D und der Metabolitenverhältnisse wurden vier Kohorten aus den Niederlanden und aus Deutschland verwendet. Unter 9.045 analysierten Verhältnissen wurde eine neue Assoziation entdeckt: Das Verhältnis Valin zu Phosphatidylcholin-Acyl-Alkyl C32:2 (PC ae C32:2). Zusätzlich wurde in der "Kooperative Gesundheitsforschung in der Region Augsburg" (KORA) Studie die Güte der Modelle durch Messung der zeitabhängigen Receiver-Operating-Characteristic-Kurve und net reclassification improvement verglichen. Fügt man zum Model, das auf etablierten Risikofaktoren basiert, das Metabolitenverhältnis hinzu, wurden kleine Verbesserungen in den Messungen beobachtet.

Die Variation der Metabolitenwerte kann von genetischen und epigenetischen Änderungen beeinflusst werden. Daher wurden viele genomweite Assoziationsstudien (GWAS) durchgeführt. Jedoch wurde bisher nur ein geringfügiger Teil der totalen Metabolitenvariation durch häufige Varianten erklärt. Dies führt zu der Annahme, dass die Analyse von vielen „Omics“-Ebenen (wie zum Beispiel die Kombination von Genetik, Epigenetik,...) – und ihren Interaktionen –hilft, die Variation der Metabolitenwerte besser zu erklären. In bisherigen Interaktionsanalysen in GWAS wurden größtenteils Interaktionen zwischen Einzelnukleotid-Polymorphismen (SNPs) betrachtet und daher mögliche Interaktionen zwischen verschiedenen „Omics“-Ebenen ignoriert.

Der zweite Teil dieser Doktorarbeit beschäftigt sich mit der Entwicklung eines statistischen Programms, um die verschiedenen „Omics“-Ebenen in einer angemessenen Zeit zu analysieren. Das neue R Paket, genannt *pulver*, ist das erste Programm, welches die Berechnung von p-Werten von Milliarden von linearen Regressionen mit einem Interaktionsterm innerhalb weniger

Tage ermöglicht. Die schnelle Laufzeit wurde durch die Verwendung von Korrelationskoeffizienten erreicht, die die Null-Hypothese testet, ob sich der Koeffizient vom Interaktionsterm signifikant von Null unterscheidet. Traditionell wird dazu die sehr zeitintensive Berechnung der Matrixinversion verwendet.

Um die Laufzeit weiter zu verkürzen, wurde die Reihenfolge, wann durch welche „Omics“-Ebene iteriert werden soll, berücksichtigt und das Programm in der schnellen Programmiersprache C++ implementiert. Angesichts des möglichen Zusammenspiels der verschiedenen „Omics“-Ebenen in biologischen Prozessen, kann *pulver* verwendet werden, um umfangreiche Untersuchungen durchzuführen, die außerhalb der Kapazitäten von existierenden Programmen liegen.

Dieses R Paket wurde mit echten Daten der KORA Studie getestet und die Metabolitenwerte mit der Interaktion von genetischen Varianten und DNA Methylierungsstellen verglichen.

Ein signifikanter Locus, in der Nähe des Gens *ACADS*, wurde dabei gefunden. Genetische und epigenetische Interaktionen in diesem Locus haben einen Einfluss auf die Metabolitenwerte von Butyrylcarnitine. Weiterhin wurde das R Paket auf das vorherige identifizierte Verhältnis, Valin zu PCae 32:2 angewendet. Jedoch wurden keine statistisch signifikanten Assoziationen nach striktem Korrigieren für multiples Testen gefunden.

Diese Doktorarbeit liefert Analysen und Programme, um ein holistisches Bild vom menschlichen Metabolismus zu erhalten. Dies wurde durch die Identifizierung von krankheitsbezogenen (T2D) Metabolitenänderungen und durch die Kombination von verschiedenen „Omics“-Ebenen (Metabolomik, Genetik und Epigenetik) erreicht, um die Interaktionen in einem effizienteren Weg aufzudecken.

---

## List of abbreviations

3-HIB	3-hydroxyisobutyrate
ACADS	Acyl-CoA dehydrogenase, C-2 to C-3 short chain
AUC	Area under the curve
BCAA	Branched-chained amino acid
CpG	Cytosine-phosphate-guanine
DNA	Deoxyribonucleic acid
EPIC-Potsdam	European prospective investigation into cancer and nutrition - Potsdam study
EWAS	Epigenome-wide association analysis
FIA-MS/MS	Flow Injection Analysis tandem MS
FPR	False positive rate
GC/MS	Gas chromatography/mass spectrometry
GRN	Gene regulatory networks
GWAS	Genome-wide association study
GxE	Gene-environment interactions
HMDB	Human Metabolome database
IMI DIRECT	Innovative medicines initiative diabetes research on patient stratification
KORA	Cooperative health research in the region of Augsburg
KM	Kaplan-Meier
LC-MS/MS	Liquid chromatography tandem mass spectrometry
LLS	Leiden longevity study
LOD	Limit of detection
MAF	Minor allele frequency
meQTL	Methylation quantitative trait locus

MAR	Missing at random
MCAR	Missing completely at random
<i>mice</i>	Multivariate imputation by chained equations
MNAR	Missing not at random
NRI	Net reclassification improvement
NTR	Netherlands twin register
OGTT	Oral glucose tolerance test
OLS	Ordinary least squares
OR	Odds ratio
PC	Phosphatidylcholine
PC aa	Phosphatidylcholine acyl-acyl
PC ae	Phosphatidylcholine acyl-alkyl
<i>pulver</i>	Parallel ultra-rapid p-value computation for linear regression interaction term
ROC	Receiver operating characteristic
SCAD	Short-chain acyl-CoA dehydrogenase
SCADD	Acyl-Coenzyme A Dehydrogenase Deficiency
SNP	Single nucleotide polymorphism
T2D	Type 2 diabetes
TPR	True positive rate
TRF	Traditional risk factors
UHPLC/MS/MS2	Ultrahigh-performance liquid chromatography/tandem mass spectrometry
Val	Valine
WGS	Whole genome sequencing

# 1. Introduction

## *1.1. Scientific question of this thesis*

A major cause of death in developed countries are diseases due to disorders of lipid and sugar metabolism, such as cerebrovascular and cardiovascular diseases [1]. For example, Suhre et al. investigated the changes of metabolites within subjects with self-reported type 2 diabetes (T2D) in comparison to controls and identified 420 different metabolite levels from the sera of these subjects [2]. They identified perturbations in pathways related to kidney dysfunction, lipid metabolism, and gut microflora interactions [2].

Moreover, these perturbations of the metabolome are a good representation of the activities of the cell at a functional level, as metabolomics are the result of gene expression and complement other “omics”, such as transcriptomics and proteomics [1].

Since most metabolite levels are highly impacted by enzymes, it is reasonable to consider the ratios of two metabolite levels, which may mirror the specific enzyme activities [3]. This in turn can be extrapolated to examine potential perturbations in pathways to gain further insight into the pathophysiology of T2D. Furthermore, as previously demonstrated by different studies [2, 4, 5], the analysis of metabolite ratios increases the power in genetic and disease association studies. Thus, the main research question of this thesis was to determine which metabolite ratios changes are observed during T2D. This additional knowledge may help to improve individual risk prediction for T2D and may even lead to prevention of the manifestation of the disease.

In addition, integrating other “omics” data to the analysis, such as genomics, transcriptomics, and proteomics, can provide a greater understanding of the global system biology and therefore the pathophysiology of cardiovascular diseases [1, 6, 7]. For example, a detailed picture of the pathways of the different “omics” layers or tissues can be drawn by computing correlation analysis and Gaussian graphical models to build a large-scale map of statistical associations [8]. For example, Yousri et al. [9] constructed a metabolic network, using Gaussian graphical modelling, that links diabetes-associated metabolites from saliva, blood plasma, and urine. This network reflected the biochemical dysregulation of metabolites from different pathways of diabetes pathology [9]. By using more complex models such as similarity network fusion, subtypes of complex diseases can be identified [7]. This method is robust to noise and data heterogeneity because it constructs first a sample-similarity network for each data type and then

integrates them into one single similarity network [7]. Another possibility is to include different “omics” data into a linear regression model, either additively or even as an interaction. For example, in a genome-wide approach genetic and epigenetic influences on metabolite levels in human blood were identified. Petersen et al. [10] showed that there is an interplay between DNA methylation and metabolite concentrations; further studies showed a similar relationship between single nucleotide polymorphism (SNPs) and metabolites [4, 5, 11]. Moreover, studies have observed associations between specific epigenetic-genetic interactions and a phenotype [12-14].

Furthermore, the biological mechanisms of those identified loci can be further elucidated by experiments in the lab. For example, since most associated variants are located in non-coding DNA regions it is assumed that some affect transcriptional regulation [15]. Therefore, Lee et al. [15] investigated the relationship between identified variants affecting T2D and the binding of transcription factors and co-regulators at the T2D associated *PPARG* locus. They observed that cis-regulatory variants contribute to the pathophysiology of T2D [15].

However, the technology and the measurements of “omics” data have rapidly increased to the point that the development of databases and methods for efficient storage, retrieval, integration, and analysis of massive data becomes necessary [16]. For example, recently, an open-source, scalable framework for exploring and analyzing genomic data called *Hail* (see <https://github.com/hail-is/hail>) was developed [17]. This software package is able to efficiently perform quality control, annotation, and analysis of large-scale sequencing data [17].

Nevertheless, existing tools either consider only linear association analysis without interaction term [17-19] or genetic variant interactions [20, 21], and no practical tool is available for large-scale investigations of the interactions between pairs of arbitrary quantitative variables.

Thus, the second part of this thesis aimed to develop such a software tool and thirdly, to apply this software to investigate the interplay among DNA methylation, genetic variants, and metabolite levels. Furthermore, this tool was also applied to the metabolite ratios found to be associated with T2D to elucidate possible influence of DNA methylation and genetic variants on these ratios.

In summary, the overall aim of this thesis is to provide a more holistic picture of human metabolism, particularly with respect to T2D, by investigating metabolite changes related to this



disease and by combining metabolomics data with interactions between genetic and epigenetic data.

### ***1.2. Overview of this thesis***

Two studies will be described in this thesis to provide a more holistic picture of human metabolism: The first study the associations between the changes of metabolite ratios in T2D including studies from the Netherlands and Germany is investigated (section 5.1). The second study combined metabolite data with interactions between genetic and epigenetic data from the Cooperative health research in the region of Augsburg (KORA) study using a self-implemented R package to enable a huge amount of regression analyses (section 5.2). For simplicity, I denote the first part “T2D and metabolite ratios analysis” and the second part “interaction analysis”.

Prior to elaborate on these studies, the biological background relevant for this thesis is introduced in section 2. An overview over basic concepts and principles of the human metabolism (section 2.1.), particularly with respect to type 2 diabetes (section 2.2 – 2.3), genomics association studies, the missing heritability and epigenomics association studies (section 2.4-2.6, respectively), as well as interactions between genomics, epigenomics and metabolomics (section 2.7), are given. In section 2.8, software which is used to analyze linear regressions using huge data sets is presented. Finally, an outlook for GWAS is provided in section 2.9.

In section 3, I introduce the study populations from the Netherlands and Germany (section 3.1) and the measurements and quality control of the “omics” data, i.e., the metabolite data (section 3.2), the genotyping (section 3.3), and DNA methylation (section 3.4). Finally, in section 3.5 the assessment of the diabetes status of the different study populations is described in further detail.

The statistical methods used in this thesis are provided in section 4: The linear regression, which is basic for the implementation of the R package for running linear regressions with interaction terms (section 4.1), the logistic regression to analyze prevalent T2D and metabolite ratios (section 4.2), and the survival analysis (section 4.3), more precisely, the Cox proportional hazard regression, to include the time of the onset of T2D in the model. The inclusion of appropriate covariates to these models to avoid confounding is further explained in section 4.4.

All models are parametric models, which assume a certain distribution: especially the linear regression model assumes a normally distributed error term, which is further addressed in sections 4.4 and 4.5. In section 4.7 the meta-analysis, i.e., the combined analysis of the results from different studies, and in section 4.8 the p-gain, i.e. the measurement of the improvement using metabolite ratios compared to single metabolites, are explained. Finally, the theory and the implementation of the R package *pulver* are described in detail in section 4.9.

Section 6 concludes this work by giving a summary of the key results of section 5 and a discussion in the context of future perspectives of analyzing different “omics” layers.

### ***1.3. Scientific contribution***

The major scientific contributions discussed in this thesis are listed in the following.

- The association analysis of prevalent/incident T2D and metabolite ratios leads to the identification of one novel association, that of the valine to phosphatidylcholine acyl-alkyl C32:2 (PC ae C32:2) ratio
- Development and preparation of the R software package *pulver* for computing p-values for the interactions term in a very large number of linear regression models which includes benchmarking different test scenarios and comparison with other R software packages
- Applying *pulver* to analyze the regression between the interaction of a genetic variant and DNA methylation and metabolite levels, only one locus achieved significance suggesting that the power was not sufficient or that these interactions play only a minor role in the determination of metabolite levels in the blood
- There is no evidence that the metabolite ratio valine to PCae32:2 is associated to the interaction of DNA methylation and genetic variants

Parts of these contributions were already published in peer-reviewed journals. Some parts of this thesis will therefore correspond to these publications:

**Sophie Molnos**, Simone Wahl, Mark Haid, E Marelise W Eekhoff, René Pool, Anna Floegel, Joris Deelen, Daniela Much, Cornelia Prehn, Michaela Breier, Harmen H. Draisma, Nienke van Leeuwen, Annemarie M.C. Simonis-Bik, Anna Jonsson, Gonneke Willemsen, Wolfgang

Bernigau, Rui Wang-Sattler, Karsten Suhre, Annette Peters, Barbara Thorand, Christian Herder, Wolfgang Rathmann, Michael Roden, Christian Gieger, Diana van Heemst, Helle Krogh Pedersen, Valborg Gudmundsdottir, Matthias B Schulze, Tobias Pischon, Eco JN de Geus, Heiner Boeing, Dorret I Boomsma, Anette G Ziegler, P Eline Slagboom, Sandra Hummel, Marian Beekman, Harald Grallert, Søren Brunak, Mark I McCarthy, Ramneek Gupta, Ewan R Pearson, Jerzy Adamski, Leen M. 't Hart (2017). The ratio of the metabolites valine and phosphatidylcholine acyl-alkyl C32:2 associates with increased risk of type 2 diabetes; a DIRECT study. *Diabetologia*. *In press*.

**Sophie Molnos**, Clemens Baumbach, Simone Wahl, Martina Müller-Nurasyid, Konstantin Strauch, Rui Wang-Sattler, Melanie Waldenberger, Thomas Meitinger, Jerzy Adamski, Gabi Kastenmüller, Karsten Suhre, Annette Peters, Harald Grallert, Fabian J Theis, Christian Gieger. pulver: An R package for parallel ultra-rapid p-value computation for linear regression interaction terms. *BMC Bioinformatics*. *Under review*.

### **Further scientific contributions**

Furthermore, the author of this thesis was involved in several other research projects, which were not directly connected to the focus of the thesis. The findings in these projects were also published in peer-reviewed journals:

Melanie Heitkamp, Monika Siegrist, **Sophie Molnos**, Simone Wahl, Helmut Langhof, Harald Grallert, Martin Halle. Impact of obesity genes on weight loss during lifestyle intervention in children with obesity. *In preparation*.

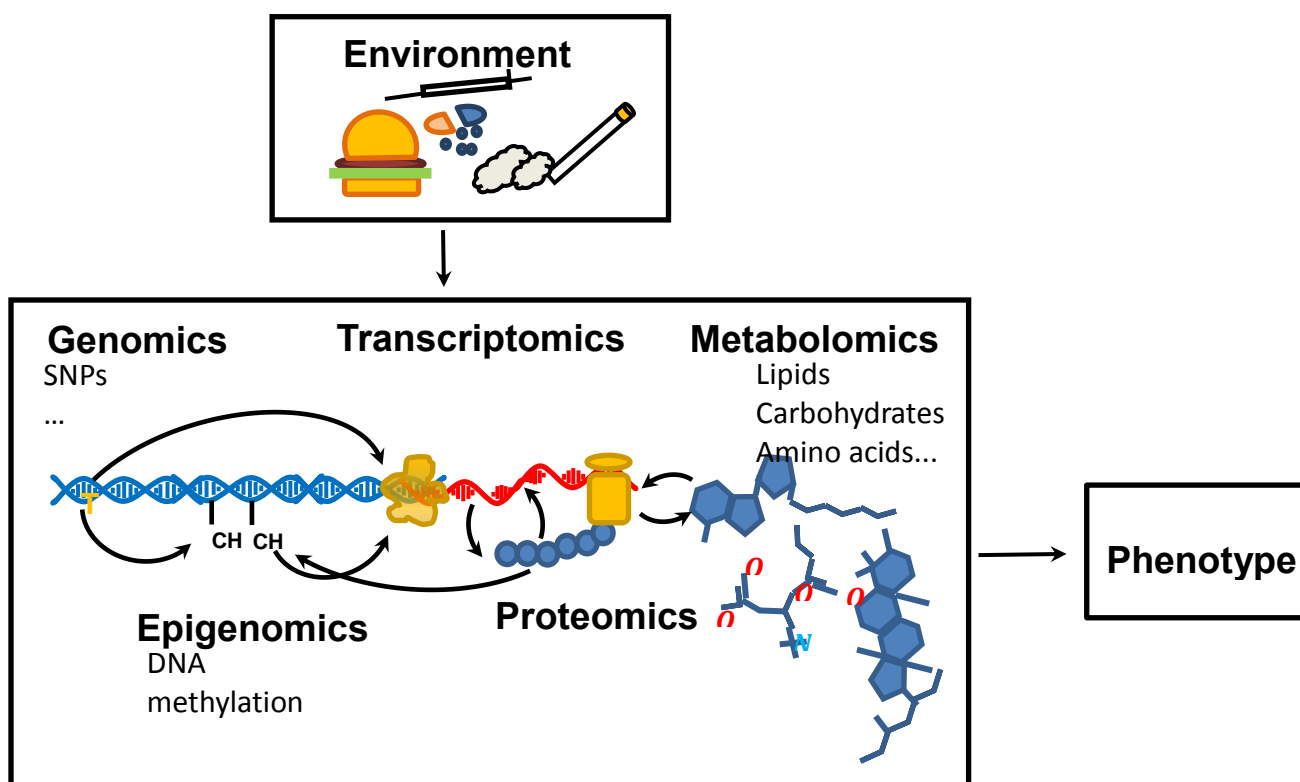
Heekyoung Lee, Kun Qian, Christine von Toerne, Lena Hoerburger, Melina Claussnitzer, Christoph Hoffmann, Viktoria Glunk, Simone Wahl, Michaela Breier, Franziska Eck, Leili Jafari, **Sophie Molnos**, Harald Grallert, Ingrid Dahlman, Peter Arner, Cornelia Brunner, Hans Hauner, Stefanie M. Hauck, Helmut Laumen (2017). Allele-specific quantitative proteomics unravels molecular mechanisms modulated by cis-regulatory PPAR $\gamma$  locus variation. *Nucleic acids research*, 45(6), 3266-3279.

Jennifer Kriebel, Christian Herder, Wolfgang Rathmann, Simone Wahl, Sonja Kunze, **Sophie Molnos**, Nadezda Volkova, Katharina Schramm, Maren Carstensen-Kirberg, Melanie Waldenberger, Christian Gieger, Annette Peters, Thomas Illig, Holger Prokisch, Michael Roden, Harald Grallert. (2016). Association between DNA Methylation in Whole Blood and Measures of Glucose Metabolism: KORA F4 Study. *PLoS one*, 11(3), e0152314.

Peter Reitmeir, Birgit Linkohr, Margit Heier, **Sophie Molnos**, Ralf Strobl, Holger Schulz, Michaela Breier, Theresa Faus, Dorothea M. Küster, Andrea Wulff, Harald Grallert, Eva Grill, Annette Peters, Jochen Graw (2016). Common eye diseases in older adults of southern Germany: results from the KORA-Age study. *Age and ageing*, 46(3), 481-486.

## 2. Background

The recent development of high-throughput measurement technologies has facilitated the analysis of different “omics” layers, such as variations of the deoxyribonucleic acids (DNA) sequence (genomics), functionally relevant modifications of the genome that leave the DNA sequence unchanged (epigenomics), gene expression (transcriptomics), protein abundances, and modifications (proteomics), as well as metabolite profiles (metabolomics) in thousands of samples in large well-powered studies (see Figure 2.1). The interaction of the different “omics” layers and the environment, e.g. smoking, can lead to different phenotypes, such as developing T2D. By integrating multi-omics datasets it is now possible to draw a more holistic picture of the biological system in health and disease [22].



**Figure 2.1:** The interplay between different “omics” layers, environment, and the phenotype adapted from [8]. This chart represents a simplified view of an information flow which is still the subject of active debate.

There are several ways how to integrate different “omics” layers, such as using association or correlation analysis and Gaussian graphical models [8] or using more complex models such as similarity network fusion [7] as mentioned in the introduction.

### ***2.1. Human metabolism in the field of common diseases – explained by genetics and epigenetics***

Metabolomics is the study of low molecular weight compounds (< 2,000 Dalton) that are intermediate or end products of enzymatic reactions [23]. Metabolite levels represent a combination of biological pathways in the organism, including, besides many others, genetically determined processes, environmental exposures, and the gut microbiome [24]. Thus, they reflect the human health status and are therefore ideal candidates to identify dysregulations of enzymes which may lead to diseases [8]. Several studies have already shown that metabolite levels can be used as biomarkers for certain diseases [25-27]. For example, the concentration of phenylalanine and the ratio of phenylalanine to tyrosine are measured to diagnose the inborn error of phenylketonuria [28], a disease which is characterized by intellectual disability, microcephaly, and seizures [29]. In addition, there is evidence in literature that the study of metabolite ratios might reveal important biological processes [5]. It was for instance shown that the ratio of 3-methyl-2-oxobutanoate and alpha-hydroxyisovalerate levels is under strong genetic control [30]. Both metabolites are products of valine catabolism and recently it was shown that genetic variation influencing the lactate dehydrogenase gene is likely responsible for the changed alpha-hydroxyisovalerate level, coding for an enzyme which catalyzes multiple reactions [31].

There are two main approaches available to measure levels of metabolites, namely targeted and non-targeted techniques. The non-targeted approach detects and identifies as many metabolites as possible, which includes even unknown molecules whereas in the targeted approach only selected metabolites can be quantified [32]. The targeted approach is more sensitive and accurate regarding the detection of a limited number of metabolites. With the untargeted approach it is possible to detect and identify unknown metabolites but with no possibility for quantification and high-quality precision. Furthermore, the time required for accurate metabolite identification and quantification can be significant [33, 34].

## **2.2. *Type 2 diabetes***

T2D is one of the most wide-spread diseases, with rising prevalences and incidences worldwide [35]. The International Diabetes Federation estimated that the total number of people having diabetes will increase from 415 million (8.8%) in 2015 to 642 million (10.4%) in 2040 [36]. In Germany T2D rose from 8.5% in 2009 to 9.5% in 2015 [37]. Additionally, there is a high estimated number of unknown cases, e.g., in the study of Rathmann et al. [38] about 50% of the cases with T2D were undiagnosed. T2D is characterized by increased blood glucose levels caused by pancreatic  $\beta$ -cell dysfunction and/or insulin resistance [27]. In contrast to type 1 diabetes, the insulin deficiency is caused by non-autoimmune etiology [39, 40]. T2D is related to increasing adiposity, inactivity, and age [41]. Furthermore, there is strong evidence for a genetic susceptibility [42] and the heritability is estimated at 25-80% [43]. However, all T2D associated genetic variants identified to date explain less than 20% of T2D heritability [43]. There are several severe long term micro- and macrovascular clinical consequences of T2D such as myocardial infarction [44] or eye diseases [45]. The global estimated costs of treating diabetes and its consequences are expected to increase from 673 billion in 2015 up to 802 billion US dollars in 2040 [36]. Therefore, it is necessary to understand the pathogenic mechanisms to find new biomarkers for early detection/ risk prevention or therapies in fighting this disease.

## **2.3. *Human metabolism and type 2 diabetes***

Circulating metabolites have been shown to reflect metabolic changes during the development from a healthy glucose metabolism to the clinical manifestation of T2D [27, 46]. It was shown that branched-chain amino acids (BCAAs), valine, leucine, and isoleucine, as well as several phospholipids can have an impact on disease progression [47-50]. Wang-Sattler et al. [27], for instance, investigated the association between metabolites and individuals with impaired and normal glucose tolerance. They identified three novel pre-diabetes-specific markers glycine, lysophosphatidylcholine (LPC) (18:2), and acetylcarnitine. Floegel et al. [46] observed that hexose, phenylalanine, and diacyl-phosphatidylcholines C32:1, C36:1, C38:3, and C40:5 were independently associated with increased risk of T2D while glycine, sphingomyelin C16:1, acyl-alkyl-phosphatidylcholines C34:3, C40:6, C42:5, C44:4, and C44:5; and lysophosphatidylcholine C18:2 were associated with decreased risk. Furthermore, elevated free

fatty acids in the plasma has also been observed to be associated with insulin resistance and T2D, which might reflect global impaired fatty acid oxidation [51, 52].

#### **2.4. Genomics**

Genomics is the generic name of the study of genes and their function encoded in the DNA sequence [53]. The DNA consists of four bases, namely cytosine, guanine, adenine, and thymine, which are bound to sugar deoxyribose and a phosphate groups, thereby forming nucleotides. Nucleotides are connected to one another, building a strand [54]. Two strands intertwine to a double helix. Cytosine and guanine as well as adenine and thymine form pairs of complementary bases. The double helices are further wrapped on proteins called histone octamers resulting in a structure called nucleosome [55]. The histone octamer consists of histones H2A, H2B, H3 and H4 [56]. DNA, the histones, and proteins together form the chromatin. The gene expression depends on the chromatin organization. Various covalent modifications on specific residues of histones such as methylation, phosphorylation, acetylation, and ubiquitination can alter the chromatin organization [56].

About 99.9% of the human genome is identical between individuals [57]. Thus, only few genetic variations lead to individual differences between the subjects. The type of genetic variations most commonly studied are single base exchanges, the single nucleotide polymorphisms (SNPs) [58, 59]. SNPs which are in neighboring loci tend to be co-inherited. This observation is also known as linkage disequilibrium (LD) [60]. The different states of a SNP are called alleles or genotypes. A person can be homozygous in the major allele, i.e. both chromosome copies carry the base of the more frequent SNP, or homozygous in the minor allele, which means, that both chromosomes carry the less frequent base [61]. If the person carries different bases, it is called heterozygotic [61]. If the true state is unknown, SNPs can also be presented as the probabilities of a subject being homozygotic in the major allele, heterozygotic or homozygotic in the minor allele. If a SNP is genotyped, the value of one of these states is 1 and the others are 0. However, since the International Haplotype Map Project started to sequence individuals from eleven different populations [62] following the 1,000 Genomes project which aimed to sequence the genome of 2,504 subjects from 26 populations [63], it has become possible to estimate (= impute) SNPs that are not measured based on a specific set of measured SNPs using information on correlations between SNPs [64]. Thus, if a



SNP is imputed, the probability of each the three states varies between 0 and 1, adding up to 1 in total.

To check whether there are problems with genotyping or population structure, it is common practice to calculate the exact test of the “Hardy-Weinberg equilibrium” (HWE). To test if HWE is valid in the absence of migration, mutation, natural selection, and assortative mating, genotype frequencies at any locus are calculated as a simple function of allele frequencies [65]. Analyzing millions of SNPs has become feasible due to better and cheaper high throughput techniques [66] as well as faster software to identify linear associations. For example, *OmicABEL* [18] efficiently exploits the structure of the data and the R package *MatrixEQTL* [19] computes linear regressions very quickly based on matrix operations. In genome-wide association studies (GWAS), in which thousands of genetic associations are calculated, it is possible to open new perspectives in understanding and treatment of diseases. For example, before the GWAS era, the only known robust association between DNA sequence variations and body mass index (BMI) were low-frequency variants in the *MC4R* gene [67]. But with the advent of GWAS the number of identified associations has increased and in 2015 Locke et al. [68] found 634 associations within 97 loci using a genome-wide approach.

## **2.5. Missing heritability**

Although many associations between genetic variations and traits have been found, it soon became obvious that in the cases of complex multifactorial diseases (e.g., T2D) or traits, such as BMI, and levels of different metabolites, the so far identified SNPs only explained a small fraction of variation, much smaller than the expected heritability estimated from twin studies. This phenomenon has become to known as the “missing heritability” [69-71]. Increasing the sample size apparently does not solve the disparity completely.

Moreover, the distribution of research on genes is biased and tends to be especially dense on few genes, probably because researchers tend to work on genes they perceive to be important [72]. However, even after two decades of intense and sophisticated molecular and genetic analyses, there are still more than a third protein-coding genes with negligible literature or known function which is also known as “ignorome”[72].

One way to tackle the “missing heritability” and “ignorome” is the development of methods analyzing polygenic variants, i.e., the same variants are jointly associated with the complex disease, [70, 71], or pleiotropic variants, i.e., the same variants are associated with multiple

traits [67, 73]. Recent work has indicated the potential power of jointly analyzing multiple phenotypes and several different analytical tools have been developed [74-76]. For example, Shen et al. [77] performed a multi-trait meta-GWAS using summary statistics and discovered 359 novel loci significantly associated with six anthropometric traits (BMI, height, weight, hip circumference, waist circumference, and waist-hip ratio).

As previously stated in section 2.1., metabolites are linked closer to the genetics as compared to the typically analyzed phenotypes, such as BMI and diseases [8]. Therefore, they have the potential to identify genetic mutations or environmental influences associated with the underlying disease pathways. Several GWAS have been conducted to find the influence of genetics on human metabolism [4, 5, 11]. Those analyses are used to get a deeper understanding of how single enzymes are determined by genetic and environmental factors, and how they are involved in the metabolic pathways eventually leading to a major disease. For example, recently, Draisma et al. [11] found a previously unidentified association between SNP rs7582179 in the *AGPS* gene and the choline plasmalogen PC ae C44:5. The authors mentioned that this gene encoded the enzyme alkylglycerone phosphate synthase and that mutations of this gene may lead to a rare autosomal recessive disorder, named rhizomelic chondrodysplasia punctata type 3 (*RCDP3*). However, the variance in serum metabolite levels explained by significantly associated SNPs (< 10%) is less than the heritability estimated in a monozygotic twin sample (< 80%) [11].

## **2.6. Epigenomics**

Epigenomics is the study of factors which influence the expression of genes in a cell or entire organism and are not caused by changes in the DNA sequence. There are different types of epigenetic mechanisms, such as DNA methylation, histone modification, chromatin remodeling, and ribonucleic acid (RNA) interference [78]. Yet, the epigenetic modification most often studied is DNA methylation, which denotes the attachment of a methyl group to a DNA base. DNA methylation is involved in genome stability, the regulation of gene expression, imprinting, and X-chromosome inactivation in females [79, 80].

Mostly, human DNA methylation is observed on the cytosine nucleotides preceding a guanine nucleotide, also called CpG sites. Hereby the methyl group is catalyzed by DNA methyltransferases (DNMT) to the 5' carbon of the cytosine, leading to a 5' methylcytosine [79].

The maintenance of DNA methylation within the cells occurs with the help of DNMT1. The new attachment of the methyl group happens with DNMT3a, DNMT3b, and the regulatory factor DNMT3L. The Tet family proteins are responsible for actively demethylating the methylated cytosine [81]. 5'-hydroxymethyl-cytosines are produced as an intermediate on the pathway. 5'-hydroxymethyl-cytosines are found to be positively correlated with gene activity and with methylation of histone H3 at lysine 4 (H3K4me1) and acetylation of histone 3 at lysine 27 (H3K27ac) [81].

In vertebrate genomes, most CpG sites are methylated, except so called CpG islands, which are DNA sequences that are on average 1000 base pairs long, GC-rich, CpG-rich, and predominantly nonmethylated [82-84]. These islands are usually rare and often near to promoters. Usually these promoters are not wrapped up in nucleosomes and the regions are flanked by nucleosomes which are marked with trimethylation of histone H3 at lysine 4 (H3K4me3) [79].

In general, methylated CpG sites are underrepresented because methylated cytosine tends to spontaneously deaminate and convert to thymine, resulting in a thymine:guanine mismatch [85]. In contrast, unmethylated cytosine can deaminate to uracil, an RNA-base, which is then recognized and fixed to cytosine. The methylation of DNA near to gene promoters usually leads to decreased gene expression [86]. Generally it was thought that expression can be stimulated when the gene bodies of actively transcribed genes are methylated [79, 87, 88]. However, it has been shown that this is not always true. For example, CTCF, a DNA binding factor, can be blocked by DNA methylation within the gene body. CTCF is responsible for slowing down the elongation rate of polymerase II on the human *CD45* gene by serving as a roadblock, and thus elevating the probability of the inclusion of the alternative exon [89]. In contrast, the multifunctional protein MeCP2 binds to methylated CpG sites. MeCP2 also slows the polymerase elongation resulting in an elevated inclusion of exons [89].

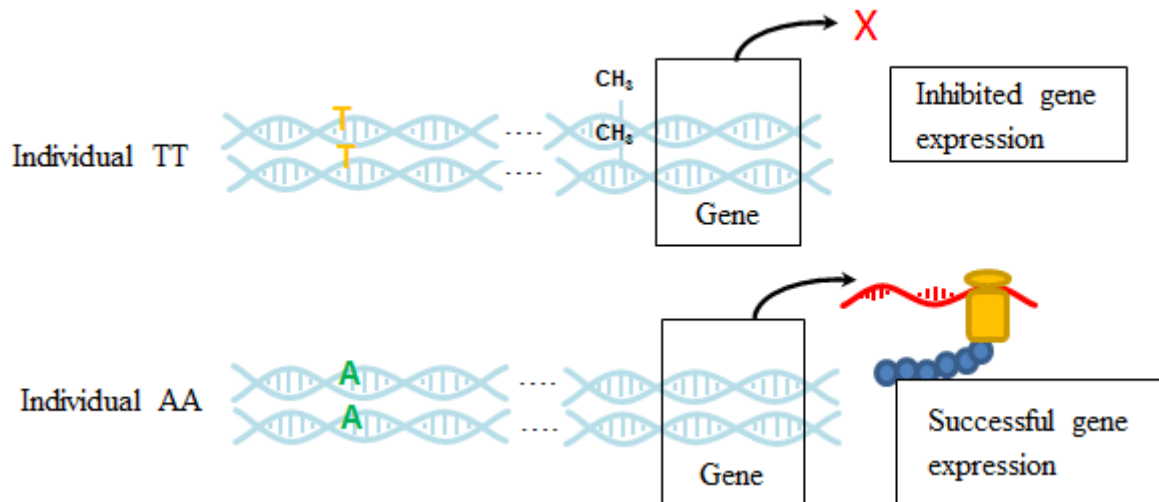
Due to cheaper and massively parallel techniques, it has become feasible to measure DNA methylation across the whole genome [86, 90]. For example, recently, Wahl et al. [91] conducted an epigenome-wide analysis study (EWAS) of 5,387 individuals and observed that increased BMI appears to lead to changes in DNA methylation. These alterations might also eventually lead to T2D independently of traditional risk factors [91].

With the help of GWAS lots of loci have been found to be associated with certain metabolite levels, but, usually, they can only explain a minor proportion of the variance of the analyzed phenotype (section 2.4) [70]. This missing heritability might be caused by epigenetics. Further, a greater understanding of human diseases could be achieved by analyzing the association between DNA methylation and metabolites, since metabolites are the connection between the genotype and the phenotype [8]. The first EWAS with metabolites was conducted by Petersen et al. [10]. They identified that certain CpG loci are associated with 4-vinylphenol sulfate (4-vs). These CpG sites resided within a region which was previously identified to be associated with tobacco smoking [92]. Furthermore, they observed associations, which disappeared after adjustment for SNPs in the neighborhood, thus demonstrating that the associations were potentially driven by genetic factors. The major difference of EWAS to GWAS is that causality cannot be inferred as easily. Altered DNA methylation profiles could be the cause or the consequence or part of a complex network of interactions of the observed environmental factor. For example, Etchegaray and Mostoslavsky described, that the activity of most enzymes involved in dynamic chromatin modifications is dependent on intermediary metabolites, such as acetyl-CoA, SAM, ATP, NAD<sup>+</sup>, flavin adenine dinucleotide (FAD),  $\alpha$ -KG, and uridine diphosphate (UDP) [93]. Alternatively, DNA methylation may even just serve as biomarker without being directly a causal part of the disease [94].

### ***2.7. The interplay between DNA methylation, genetic variants, and environment***

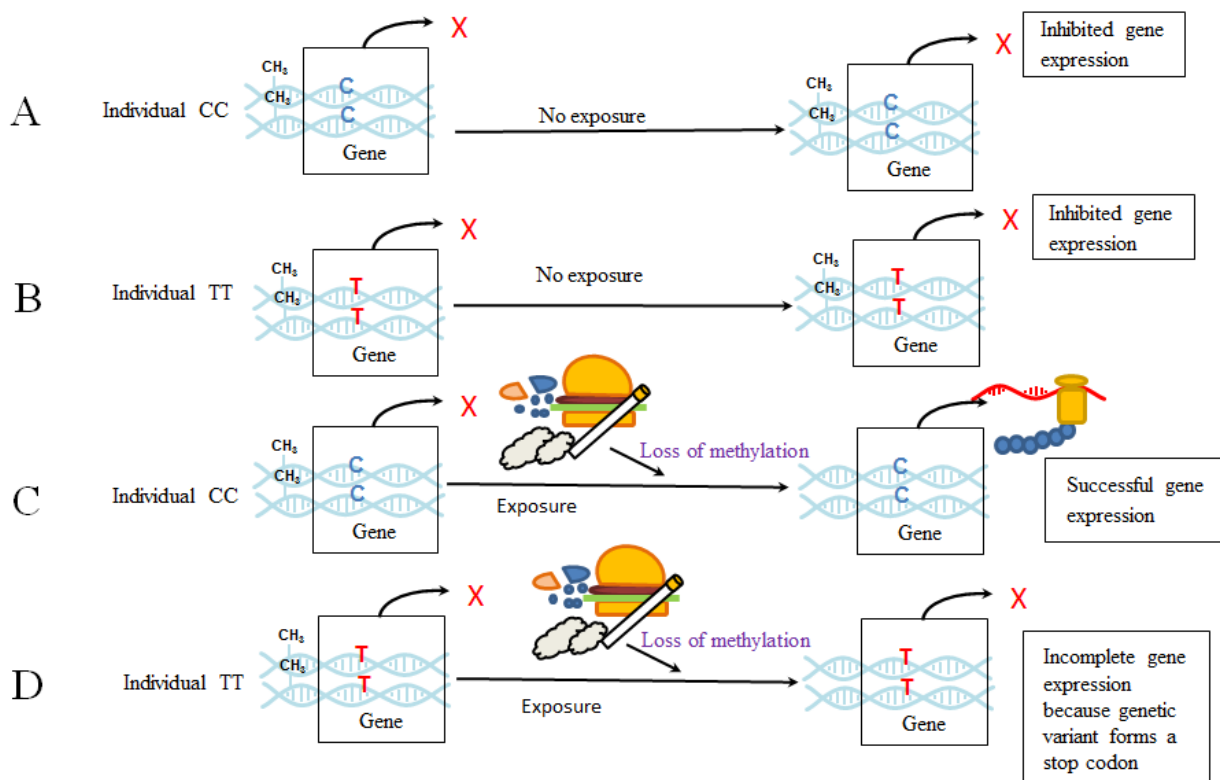
Epigenetics may provide mechanistic insights into the complex interplay of genetic and environmental risk factors for disease or may serve as a biomarker of exposure or disease. There are several biological theories which aim to explain the observed associations between DNA methylation, the genetic variants and environment, respectively. For example, DNA methylation can act as so called epigenetic mediation. This means that the methylation state of a specific locus is driven by a nearby genetic variant or environmental factor and eventually can lead to disease [94]. The loci, usually SNPs, where genotype is associated with methylation level at a given locus are called methylation quantitative trait loci (meQTL) [12, 95], see also Figure 2.2 for epigenetic mediation due to a genetic variant. Most of the identified meQTLs have been observed with CpG sites within 3000 base pairs of the genetic variant [94]. These relationships

are perhaps realized due to intermediate mediators, such as DNA binding factors or secondary chromatin structures [12, 96]. Furthermore, meQTLs may provide a possible explanation for the identified associations between intergenic and intronic SNPs and the phenotype [12, 96]. For example, Heyn et al. [12] investigated the interplay between SNPs and DNA methylation to improve the interpretation of risk alleles in human tumors identified in GWAS. They observed that 21% of the interrogated cancer risk polymorphisms are also associated with DNA methylation.



**Figure 2.2:** Example of an epigenetic mediation due to genetic variation adapted from [94]. The individual with the TT genotype is methylated at a particular CpG site and hence the associated gene is silenced. However, the individual with the genotype AA is not methylated at this CpG site, which leads to the transcription of the associated gene. The differences in the DNA methylation are driven by genotype differences and thus the epigenetic state mediates the relationship between the genotype and the phenotype and can potentially lead to a disease.

From a more complicated point of view there are also epigenetic mechanisms within gene-environment interactions (GxE) [94]. In this concept, the gene expression, and, consequently, the phenotype is influenced by the combination of genetics and environmental factors. For instance, healthy humans who have a disease-causing variant but are protected through DNA methylation can suddenly suffer from the disease because of the deletion of this DNA methylation due to environmental factors. Figure 2.3 shows an example for GxE with DNA methylation.



**Figure 2.3:** Example of gene-environment interaction (GxE) adapted from [94]. The expression of a gene may be influenced by both, the underlying genetic variant and a particular environmental exposure. Individuals with genotype CC and TT in panels A and B show no gene expression because the genes are silenced due to DNA methylation at a particular locus. In contrast, if the DNA methylation is removed due to environmental exposure the gene is transcribed. However, because the individual with the genotype TT forms a stop codon the protein is not produced (panel D).

As already illustrated in Figure 2.1, many interdependencies within and between different “omics” data can be observed. As mentioned in section 2.1, metabolite levels result from a combination of genetically determined processes and environmental exposures. Thus, an alternative to analyzing the association between genetic-epigenetic interaction and environmental factors, is considering metabolite levels. Recent literature suggests that specific epigenetic modifying enzymes are dependent on the availability of metabolites [29, 93, 97]. On the other hand, levels of metabolites can represent the activity of enzymes which were influenced by SNPs and DNA methylation [25]. It is very likely that there even exists an interaction between genetic, epigenetic data, and metabolite levels. For example, Ma et al. [13] explored whether the interaction between a genetic variant and a fatty acid, is mediated by DNA methylation leading to the observed blood lipids. They found some evidence that these interactions act on blood lipids through DNA methylation. For example, higher plasma HDL

cholesterol was associated with fewer C alleles at SNP rs2246293 (region *ABCA1*) and higher circulating eicosapentaenoic acid (EPA). Furthermore, they showed that the CpG site cg14019050 was significantly associated with HDL cholesterol and that SNP rs2246293, EPA and the interaction between both were also associated with the DNA methylation of cg14019050 indicating a mediation of the CpG site.

However, due to computational costs to run the exhaustive search of all pairs, to this point, no association analysis between three layers, i.e., DNA methylation, genetic variants, and metabolite levels in a genome-wide view has been conducted.

### ***2.8. Software analyzing linear regression with interaction term***

As introduced in section 2.4, it is possible to analyze millions of linear regressions using rapid software, such as *OmicABEL* [18] and the R package *MatrixEQTL* [19]. With the recently established software framework *Hail* (see <https://github.com/hail-is/hail>) [17], it is possible to efficiently analyze gigabyte-scale data on a laptop or terabyte-scale data on a cluster. In addition, several fast tools have been implemented analyzing the interactions between SNPs in genome-wide studies. For example, *BiForce* [98] is a stand-alone Java program that integrates bitwise computing with multithreaded parallelization, *SPHINX* [99] is a framework for genome-wide association mapping which finds SNPs as well as SNP-SNP-interactions using a piecewise linear model, and *epiGPU* [20] is a software that calculates contingency table-based approximate tests using consumer level graphics cards. However, they are specialized for the case of analyzing SNP-SNP-interactions. In case of analyzing interactions between any kinds of “omics” data, e.g., between DNA methylation and mRNA expression, it is only possible to use standard software, such as R’s build-in function *lm* or the software *OmicABEL* [18]. However, before calling the function *OmicABEL*, it is necessary to compute the interaction of these “omics” layers first. The reason therefore is this function computes only linear regressions without interaction term. Taking into consideration the time required for first loading, then computing the interaction, and finally run the linear regression, this process can be very time consuming. Furthermore, some tools may require the data in a special format which can add to the processing time. Thus, there is a need for implementing fast tools for models analyzing interactions between different “omics” layers.

### **2.9. *What is after GWAS?***

GWAS could not resolve the genetic bases of common diseases [70, 100, 101]. Still, there is a disparity between the explained and expected heritability of complex diseases, as introduced in section 2.5. In most GWAS only additive genetic variation are computed, but underlying molecular networks are highly nonlinear [101]. For example, Zuk et al. [70] argued that estimates of total heritability ignore the genetic interactions (epistasis) among loci often observed in studies. Also, the proportion of heritability is computed from the ratio of the significant associated variants and the total heritability, inferred indirectly from population data [70]. Thus, when accounting for epistasis, the total heritability may be much smaller and thus the proportion of heritability explained much larger [70]. In this study the authors referred to the missing heritability for Crohn's disease, where 80% of the currently missing heritability could be due to genetic interactions [70]. Therefore, as conducted in this thesis, different models including interaction terms as well as combining different "omics" data such as DNA methylation, transcriptomics, or metabolomics, should be analyzed to further examine the biological system. Furthermore, results from GWAS can be used to infer biological networks, such as gene regulatory networks (GRN) [101]. Thus, the combination of population genetic models and molecular biological knowledge may help the fitting of experimental data to very complex models, as well as allow accurate in capturing of the uncertainty of resulting inference [101]. For example, Frau et al. [102] used the loci identified in GWAS for T2D to conduct network and pathway enrichment analyses to understand mechanisms of action and clinical relevance of these variants.

Experiments and epidemiological studies can complement each other to further examine the biological background underlying complex diseases. For example, in this thesis, the observed association between metabolite ratios and insulin secretion is validated by computing the association between T2D and insulin secretion [103].



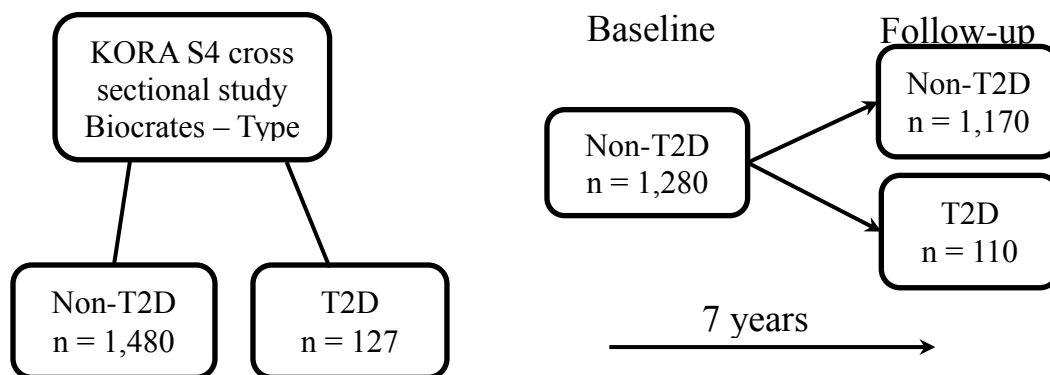
### 3. Material

#### 3.1. Study populations

In this section I describe the study populations of four cohorts from two different European countries (the Netherlands and Germany) which were analyzed in this thesis. The studies were approved by their local Ethics Committee and all participants signed an informed consent.

##### 3.1.1. Cooperative health research in the region of Augsburg

Cooperative health research in the region of Augsburg (KORA) is a research platform that comprises a population-based set of epidemiological surveys and follow-up studies in the region of Augsburg in southern Germany [104]. For this thesis, subsets of the data from the survey KORA S4 (1999/2000) comprising 4,261 subjects aged 25-74 years and the follow-up study KORA F4 (2006–2008) comprising 3,080 participants were included. Detailed information about study design, sampling method and data collection has been described elsewhere [104]. For the IMI DIRECT project the baseline characteristics for KORA S4 are shown in Table 3.1, for KORA F4 in Table 3.2. See Figure 3.1 for visualization of number of subjects from KORA S4 with information of T2D as used for analysis of IMI DIRECT (see section 5.1. for further information). For the interaction project, a subset of 1,613 or 1,643 participants were analyzed including the intersection of DNA methylation, genotyping, and metabolite levels, measured from the platforms Metabolon or Biocrates, respectively.



**Figure 3.1:** Schematic design of the KORA S4 (left) and KORA S4 to F4 studies.

**Table 3.1:** The baseline characteristics of the KORA S4 incident T2D sample.

Number of participants (n)	1610	
	n missing	
Age (Years)	0	64.1 ± 5.5
Gender (n male (%))	0	827 (51.4)
BMI (kg/m <sup>2</sup> )	13	28.6 ± 4.4
Fasting Glucose (mmol/l)	225	5.69 ± 0.95
Fasting Insulin (pmol/l)	144	113 ± 162
Physical activity (active)	8	670 (41.8)
Alcohol intake (g/day)	7	16±20.9
Smoking (smoker (n (%)))	2	206 (14.9)
Systolic blood pressure (mmHg)	6	136.6±20.6
HDL cholesterol (mg/dl)	1	57.5±16.4
Incident Type 2 diabetes (n (%))*	110 ( 9.4)	
Prevalent Type 2 diabetes (n (%)) **	127 (7.9)	
Lipid medication (Yes), (n (%))	195 (12.1)	
Fasting (Yes) (n (%))	1,349 (86.7)	

Data are means ± SD or number (n). \* Developed T2D during the on average seven year follow-up from the baseline S4 measurement till the follow-up F4 measurements (denoted in the text as KORA S4 to F4 sample including 110 incident T2D cases and 1,170 non-diabetic controls) (27).

\*\* excluded in this study

**Table 3.2:** The baseline characteristics of the KORA F4 prevalent T2D sample.

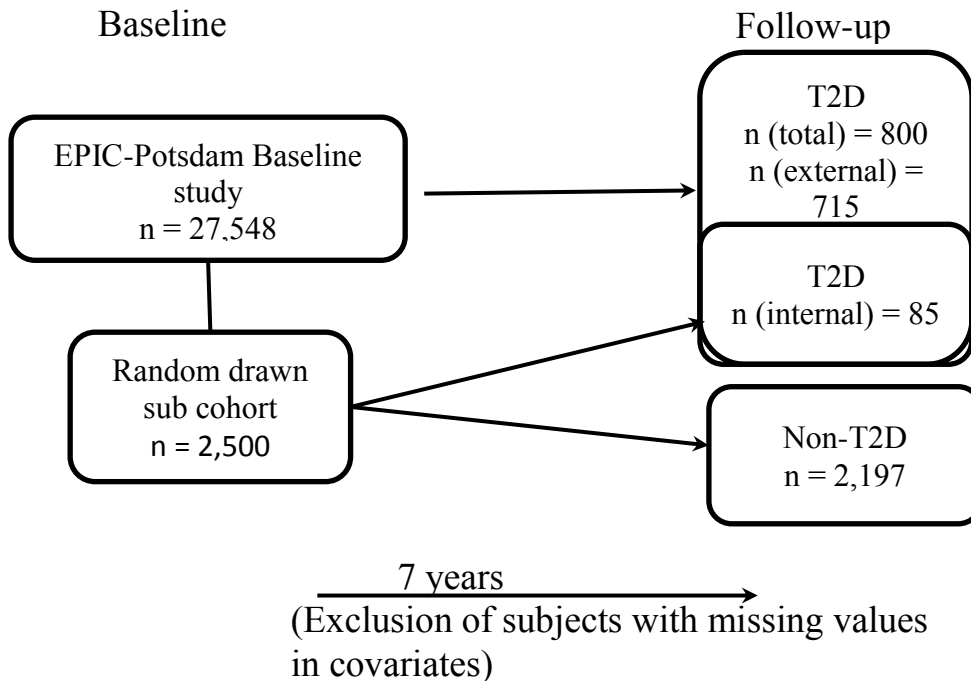
Number of participants (n)	3044	
	n missing	
Age (Years)	0	56.0 ± 13.3
Gender (n male (%))	0	1,472 (48.4)
BMI (kg/m <sup>2</sup> )	15	27.6 ± 4.8
Fasting Glucose (mmol/l)	34	5.45 ± 1.05
Fasting Insulin (pmol/l)	19	54 ± 205
Incident Type 2 diabetes (n (%))	NA	

Prevalent Type 2 diabetes (n (%))	213 (7.0)
Lipid medication (Yes), (n (%))	392 (12.9)
Current smoking (n (%))	466 (18.7)
Fasting (Yes) (n (%))	3,026 (99.4)

Data are means  $\pm$  SD or number (n). NA not available

### 3.1.2. European prospective investigation into cancer and nutrition - Potsdam study

European prospective investigation into cancer and nutrition - Potsdam study (EPIC-Potsdam) is part of the multicenter EPIC study and comprises 27,548 participants aged between 35-65 years, recruited in 1994-1998 from the general population in the area of Potsdam in eastern Germany [105]. A case-cohort study within EPIC-Potsdam was constructed by randomly drawing subjects from the EPIC-Potsdam study population [46]. Hereby all incident cases of T2D identified up to 31 August 2005 (n = 849, mean follow-up seven years) were included. The controls were drawn from a sub cohort and controlled for matched age and sex (n = 2,500). See Figure 3.2 for visualization of structure of EPIC-Potsdam. The baseline characteristics of EPIC-Potsdam are shown in Table 3.3.



**Figure 3.2:** Schematic design of the EPIC-Potsdam study.

**Table 3.3:** The baseline characteristics of the EPIC-Potsdam study population.

Number of participants (n)	2997	
	n missing	
Age (Years)	0	50.7 ± 8.8
Gender (n male (%))	0	1289 (43.0)
BMI (kg/m <sup>2</sup> )	0	27.0 ± 4.7
Random Glucose (mmol/l)	0	5.14 ± 1.50
Fasting Insulin (pmol/l)	2555	53 ± 40
Coffee (cups/d)	0	2.8 ± 2.1
Whole Grain Bread (g/d)	0	44.2 ± 52.5
Red Meat (g/d)	0	44.6 ± 30.7
Waist circumference (cm)	0	88.9 ± 14.0
Incident Type 2 diabetes (n (%))	800 (26.7)	
Prevalent Type 2 diabetes (n (%))	NA	
Prevalent hypertension (n (%))	1,630 (54.4)	
Lipid medication (Yes), (n (%))	186 (6.2)	
Current smoking (n (%))	617 (20.6)	
Fasting (Yes) (n (%))	431 (14.3)	

Data are means ± SD or number (n). NA not available.

The EPIC-Potsdam sample uses a case-cohort design including all incident cases of the whole cohort (n=27,548) and a randomly sample sub cohort (n=2,500) of which 2,197 healthy controls with Biocrates data were used in the current study (7).

### 3.1.3. Leiden longevity study

Leiden longevity study (LLS) is a family-based cohort that recruited 420 families [106]. Families were chosen if at least two long-lived siblings were alive and fulfilled the age-criterion of age 89 and older for men and 91 and older for women. There were no selection criteria on health or demographic characteristics [107]. Baseline characteristics of LLS are shown in Table 3.4.

**Table 3.4:** The baseline characteristics of the LLS study population.

Number of participants (n)	558	
	n missing	
Age (Years)	0	63.0 ± 6.5
Gender (n male (%))	0	267 (47.8)
BMI (kg/m <sup>2</sup> )	0	25.6 ± 3.6
Fasting Glucose (mmol/l)	327	5.09 ± 0.50
Fasting Insulin (pmol/l)	327	45 ± 32
Incident Type 2 diabetes (n (%))	NA	
Prevalent Type 2 diabetes (n (%))	42 (7.5)	
Lipid medication (Yes), (n (%))	47 (8.4)	
Current smoking (n (%))	76 (13.6)	
Fasting (Yes) (n (%))	239 (42.8)	

Data are means ± SD or number (n). NA not available.

### 3.1.4. Netherlands twin register

Netherlands twin register (NTR) is a family-based twin registry that recruited twin families between 2004 and 2008. Detailed information about study design, sampling method and data collection has been described elsewhere [108]. Baseline characteristics of NTR are shown in Table 3.5.

**Table 3.5:** The baseline characteristics of the NTR T2D study sample.

Number of participants (n)	1326	
	n missing	
Age (Years)	0	51.4 ± 14.0
Gender (n male (%))	0	888 (66.7)
BMI (kg/m <sup>2</sup> )	7	26.0 ± 3.8
Fasting Glucose (mmol/l)	1	5.71 ± 1.14
Fasting Insulin (pmol/l)	1326	NA

Incident Type 2 diabetes (n (%))	NA
Prevalent Type 2 diabetes (n (%))	51 (3.9)
Lipid medication (Yes), (n (%))	167 (12.6)
Current smoking (n (%))	NA
Fasting (Yes) (n (%))	1,255 (94.7)

Data are means  $\pm$  SD or number (n). NA not available

### 3.2. *Metabolomic measurements and quality control*

Two different platforms for measurements of metabolites provided by Metabolon and Biocrates are described in detail in the following two sections. Moreover, missing values of both measurements were imputed using the R package “mice” which is explained in section 3.2.3 [109].

#### 3.2.1. **Biocrates platform**

For all five cohorts (KORA S4/F4, EPIC-Potsdam, LLS, NTR), metabolite concentrations were measured using the Biocrates Absolute*IDQ*p150 kit or p180kit (BIOCRATES Life Sciences AG, Innsbruck, Austria) at the Metabolomics Platform of the Genome Analysis Center at the Helmholtz Zentrum München, Germany, following the instructions described in the manufacturers’ manual [5, 110-112]. Briefly, Biocrates uses a targeted Flow Injection Analysis tandem mass spectrometry (FIA-MS/MS) technique for quantification of known metabolites, 163 for kit p150 or 186 for kit p180, measurements. The p180 kit is an extension of p150 that uses additional liquid chromatography tandem mass spectrometry (LC-MS/MS) separation. The analytical process was performed with the Analyst 1.4 software and the Met*IQ*<sup>TM</sup> software package, integrated in the Absolute*IDQ*<sup>TM</sup> kits. Internal standards serve as reference for the calculation of metabolite concentrations. With the used analytical technique, it is not possible to determine the precise position of the double bonds and the distribution of carbon atoms between two fatty acid side chains. Thus, the lipid side chain composition is abbreviated as Cx:y, where x denotes the number of carbons in the side chain and y the number of double bonds [110]. Metabolite concentrations are given in  $\mu\text{mol/l}$ . See in Appendix Table A.1 for a list of measured metabolites.

For quality control, for each metabolite and plate the coefficient of variation was calculated. The coefficient of variation for metabolite *i* and plate *j* is defined as:

$$CV_{i,j} = \frac{sd_{ij}}{mean_{i,j}},$$

where  $sd_{ij}$  is the standard deviation (sd) and  $mean_{i,j}$  is the mean over all reference measurements per plate  $j$  and metabolite  $i$ . If the coefficient of variation averaged over all plates of one metabolite exceeds 25% it was excluded from the dataset.

Outlying metabolite concentration values and outlying samples were also removed. An outlier was defined as a metabolite concentration of one subject which is greater or less than the mean plus five standard deviations. The resulting dataset was naturally log-transformed to obtain a normal distribution. All missing values were imputed with the R package *mice* which uses a linear regression approach [109] (see section 3.2.3 for details).

### 3.2.2. Metabolon platform

For KORA F4 subjects' metabolomics measurements were performed on two separate ultrahigh-performance liquid chromatography/tandem mass spectrometry (UHPLC/MS/MS2) injections and one gas chromatography/mass spectrometry (GC/MS) injections per sample. Measurements on the platform at the company Metabolon Inc. (Durham, NC, USA) is described in detail elsewhere [30, 32].

After the relative quantifications, 325 of the in total 517 compounds could so far be identified based on a standard library of MS/MS spectra. The identified molecules belong to a variety of metabolite classes, namely, amino acids, peptides, carbohydrates, fatty acids, glycerophospholipids, acylcarnitines, sphingolipids, steroids, ketone bodies, bile acid metabolites, nucleotide metabolites, vitamins and xenobiotics, see in section Appendix Table A.2 for a list of measured metabolites.

For quality control and normalization, the metabolite levels were divided by the median value of samples measured on the same day for each metabolite independently. Furthermore, metabolite levels having a value greater than four times the standard deviations from the mean of the respective metabolite on natural log scale were set to missing. Metabolites were excluded, if they had more than 40% missing values leading to a total of 406 remaining metabolites. All missing values were imputed with the R package *mice* (Multivariate imputation by chained equations) [109] (see section 3.2.3 for details).

### 3.2.3. Imputation

Missing values are frequently observed in high-throughput mass spectrometry-based metabolomics measurements. However, for running association analyses, as conducted in this thesis, a complete data set is required. Therefore, such missing values either need to be handled during statistical analyses, e.g. using complete cases only, or by imputation approaches prior to analysis. In general, missing of values can occur for various reasons and are categorized into three groups [113-115]: Missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In MCAR, the missing of an outcome is independent from any observed or unobserved variables. MCAR can occur because of technical reasons, e.g., the metabolite was not measured properly due to matrix or contamination effects preventing the quantification of a metabolite in a sample. In case of MAR, the missing values depend only on observed values and not on unobserved values. For example, the probability of missing values for the metabolite caffeine is increased in children, because it is less likely that children drink caffeinated beverages. In MNAR, the missing values depend on unobserved data, conditional on the observed data. For example, if the metabolite level is below the instrument sensitivity thresholds, it might not be detectable in a sample (limit of detection, LOD).

Based on these categorizations, there are different ways to handle the missing values. For example, if we assume that they are MCAR, it is sufficient to use single imputation, i.e., one data set is generated by filling up the missing values through mean imputation or regression techniques. Otherwise in case of MAR or MNAR, it might be useful to compute and analyze multiple complete data sets [116] (at least five data sets [109]) and combine them based on rules by Rubin [117]. However, as huge data sets are used in this thesis, it would take too long to compute each data set separately. Therefore, we used a single imputation based on the R package *mice* [109].

The chained equation process which is applied in the R package *mice* can be divided in several steps as introduced by Azur et al. [118]:

1. The missing values are filled with arbitrary start values, such as mean
2. The imputed values for one variable “var” are set back to missing



3. The variable “var” is then regressed on the other variables using only the observed variable from variable “var”, i.e., “var” is the dependent variable in a regression model and all the other variables are the independent variables.
4. Subsequently, the missing values for “var” are replaced with values drawn from the conditional posterior distribution of the missing values. The conditional posterior distribution is computed from the estimates obtained from the regression. Then, the same procedure is applied for other variables, using “var” with imputed values as an independent variable in the regression model
5. Steps 2-4 are repeated for each variable with missing values. In addition, they are repeated for several cycles, depending how many imputed data sets are wanted

In this thesis, for the Biocrates as well as for the Metabolon measurements, five data sets were generated which were then averaged to obtain one data set.

### ***3.3. Genotyping and quality control***

Genotyping for the KORA F4 study was performed on the Affymetrix Axiom chip [119] and called with the Affymetrix software and annotated to NCBI 37 of 1000g phase 1. Genotypes were imputed against 1000g phase 1 integrated haplotypes reference set using IMPUTE v2.3.0 [120], with SHAPEIT v2 [121] as a pre-phasing tool. For quality control, all subjects were removed who have discordances between phenotypic and genetic sex. Furthermore, all observations were removed which did not cluster with HapMap CEU population in a joint plot of the first two principle components or deviated by at least five standard deviations from the mean heterozygosity rate.

In addition, a so called observation-wise call rate and a SNP-wise call rate was calculated. The observation-wise call rate is the proportion of SNPs per subject for which a reliable genotype assignment could be made based on the fluorescence signals obtained for the two alleles of the SNP. All subjects having a call rate below 97% were removed from the data set. The SNP-wise call rate is the proportion of observation per SNP for which a reliable genotype assignment could be made. All SNPs having a call rate below 98% were removed from the dataset. Furthermore, SNPs having a HWE p-value below  $5 \times 10^{-6}$  or having an imputation information score below 0.5 were also excluded from the dataset. Finally the data were transformed to dosages (i.e. estimated counts) of the reference allele, calculated as:

$Dosage = 2 \cdot p(AA) + 1 \cdot p(AB)$ , where  $p(AA)$  is defined as the probability to be homozygous to the reference allele, and  $p(AB)$  the probability to be heterozygous.

### 3.4. Array-based DNA methylation & quality control

In KORA F4 genome-wide DNA methylation measurement at 485,577 genomic sites was performed using the Infinium HumanMethylation450K BeadChip (Illumina, Inc., CA, USA) and is described in detail elsewhere [92]. Briefly, the single-stranded genomic DNA underwent a bisulfite treatment using the EZ-96 DNA Methylation Kit (Zymo Research, Orange, CA, USA). Afterwards, samples were whole genome amplified, fragmented, resuspended, and hybridized to the Bead Chips. Finally, arrays were stained with fluorescence and scanned with the Illumina HiScan SQ scanner resulting into a methylated and an unmethylated signal count per CpG site. These counts are then transformed into  $\beta$ -values, which represent the proportion of methylation at a given CpG site and are defined as the ratio of methylated signal intensity divided by the overall signal intensity [122]:

$$\beta - value = \frac{\max(M, 0)}{\max(M, 0) + \max(U, 0) + \alpha}$$

The  $\alpha$  is set to 100 as a regularization in case both methylated signal ( $M$ ) and unmethylated signal ( $U$ ) are too low [123].

For quality control, the DNA methylation was corrected for possible background noise using the R package *minfi*, version 1.6.0 [124]. In addition, detection p-values, i.e. the probability of a signal being detected above the background signal level, are estimated from negative control probes. Thus, all signals having p-values  $\geq 0.01$  were removed. Furthermore, signals which came from less than three functional beads on the chip are potentially unreliable signals and therefore removed from the dataset.

To reduce the non-biological variability between observations, data were normalized using the quantile-normalization (QN) with the R package *limma*, version 3.16.5 [125].

Furthermore, to avoid false positive associations, all CpG sites which were listed by Chen et al. [126] as cross-reactive probes were removed. Cross-reactive probes bind on repetitive sequences or co-hybridize to alternate sequences which are highly homologous to the intended targets and could lead to false signals.

Moreover, an open challenge is the measurement of CpG sites from whole blood, and thus coming from a mixture of different cell types. To diminish cell type confounding, the model was adjusted by the so-called Houseman variables which reflect the blood cell proportions [127]. Thus, CpG sites were represented by their residuals after regressing on age, sex, body mass index (BMI), Houseman variables, and the first 20 principal components of the principal component analysis control probes from 450K Illumina arrays. The control probes were used to adjust for technical confounding. According to Lehne et al. [128], the first 30 principal components are sufficient to almost entirely remove test statistic inflation consistent with effective correction for the batch and technical effects. However, in KORA it was observed that already 20 principal components were enough to remove technical confounding.

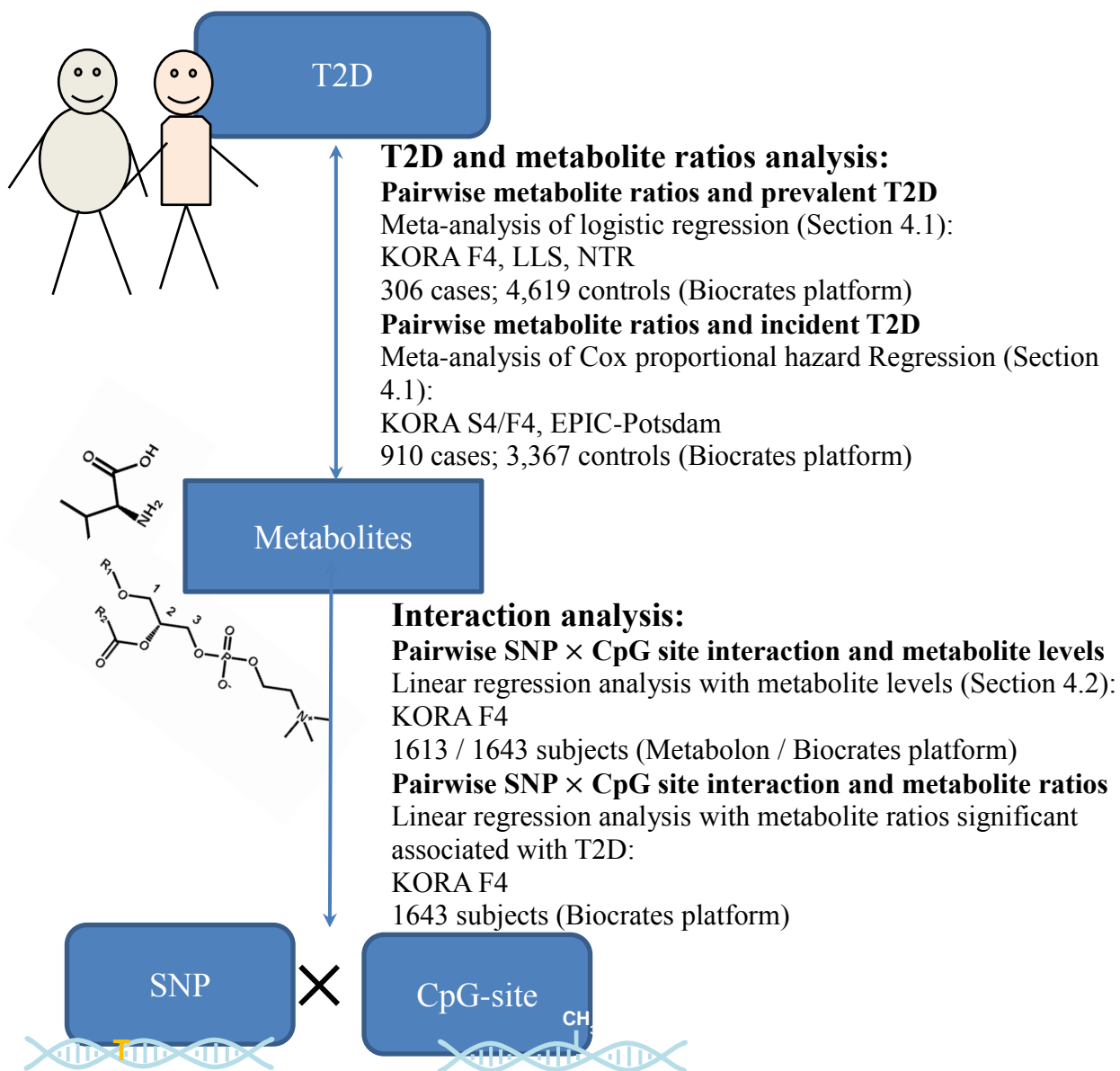
### ***3.5. Assessment of diabetes-status***

Incident and prevalent diabetes information in KORA were obtained by self-reported diabetes type and date of diagnosis. In addition, cases were validated by contacting their physician. All other subjects underwent an oral glucose tolerance test (OGTT) after overnight fasting. Thus, incident T2D was then defined based on either validation by physician diagnosis or newly diagnosed diabetes by the OGTT ( $\geq 7.0$  mmol/l fasting or  $\geq 11.1$  mmol/l 2-h glucose) [129]. In EPIC-Potsdam follow-up questionnaires were sent to participant every 2-3 years to identify incident cases of T2D. Cases were further verified by medical records [46]. Information about T2D in LLS was requested from the participants' treating physicians [106]. T2D information from the NTR was retrieved by questions in a survey asking whether a doctor ever diagnosed diabetes and whether any diabetes-related medication was used [42].



## 4. Methods

In this thesis, several statistical tests were applied to analyze the association between T2D and metabolite ratios as well as the association between metabolite levels and the interaction of SNPs and DNA methylation. For the analysis of T2D and metabolite ratios, a logistic regression and a Cox proportional hazard regression is used. For the interaction analysis, a linear regression with an interaction term is used. Here, they are briefly introduced. An overview is given in Figure 4.1.



**Figure 4.1:** Schematic overview of the study design used in this thesis. Further details on the study samples can be found in the section 3.

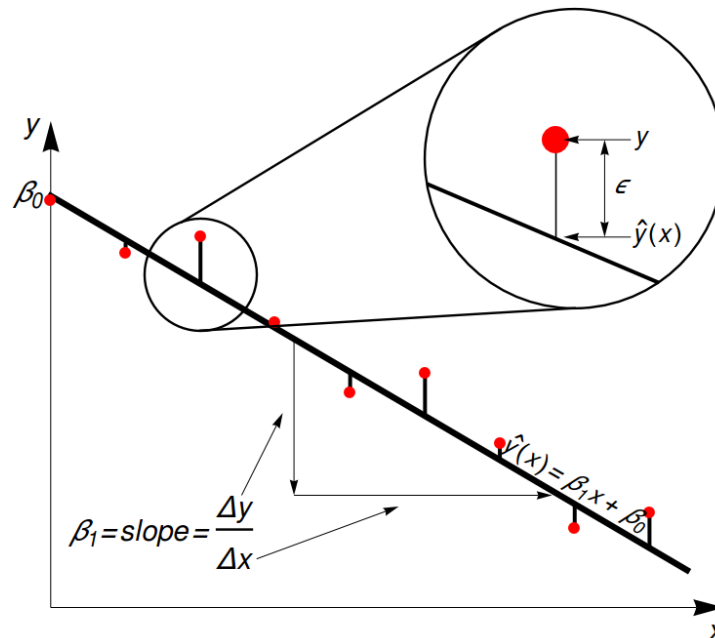
### 4.1. Linear regression analysis

The most common statistical test used in epidemiological studies is based upon the linear regression model. In this thesis, the univariate linear regression is used to examine the relationship between metabolite levels and the interaction between DNA methylation and genetic variants.

The simple linear regression is the model that tries to find a linear relationship between one independent variable  $x$  and the dependent variable  $y$ . In brief, a line between two variables gets fitted, that aims at minimizing the vertical differences (residuals) between  $n$  predictions of outcome  $y$  and  $n$  observations of  $x$  (see Figure 4.2) [130]:

$$y = \beta_0 + x\beta_1 + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n),$$

where  $\beta_0$  is the intercept,  $\beta_1$  is the slope of the regression line, and  $\sigma^2$  is the variance of the error term  $\epsilon$ , which is independent and identical normally distributed with mean 0. Mostly, we are interested in obtaining the slope  $\beta_1$  which can be estimated by minimizing the error term  $\epsilon$ . Conventionally, the minimization of the error term  $\epsilon$  is achieved by minimizing the sum of the squared error term leading to the name ordinary least squares (OLS) regression [130].



**Figure 4.2:** Schematic illustration of simple linear regression adapted from [130]. The regression line,  $\hat{y} = \beta_0 + \beta_1 x$ , is estimated by minimizing the sum of the squared vertical differences (the residuals, here shown as  $\epsilon$ ) between the points and the regression line.

In a multiple linear regression, more than one independent variable is included into this model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon = X\beta + \epsilon, \quad \epsilon \sim i.i.d. N(0, \sigma^2 I_N), \quad (4.1)$$

where  $p$  is the number of independent variables,  $\beta^T = (\beta_0, \beta_1, \dots, \beta_p)$  denotes the regression coefficients, i.e. slopes, of the corresponding independent matrix  $X$  which contains  $p + 1$  independent variables  $(1, x_1, \dots, x_p)$ .

If the data is fitted with a regression line as illustrated in Figure 4.2 assuming model (4.1), and defining the error term as the difference between the observed value and the predicted value  $\epsilon_i = y_i - \hat{y}(x_i)$  for each pair  $(x_i, y_i)$ , the estimation of the regression coefficients in  $\beta$  and the standard deviation  $\sigma^2$  are then [130]:

$$\beta = \arg \min_{\beta} \sum_{i=1}^N \epsilon^2 = \arg \min_{\beta} \sum_{i=1}^N (y_i - \hat{y}(x_i))^2 = \arg \min_{\beta} \sum_{i=1}^N (y_i - (\alpha + x_i \beta^T))^2$$

The estimators (based on OLS) are thus [131]:

$$\hat{\beta} = (X^T X)^{-1} X^T y, \quad (4.2)$$

$$\hat{\sigma}^2 = \frac{\epsilon^T \epsilon}{N-p-1} = \frac{(y-\hat{y})^T (y-\hat{y})}{N-p-1}, \quad (4.3)$$

where  $p$  is the number of independent variables and  $n$  the number observations.

Instead of computing the variance directly from the formula given in Equation (4.3), it is more common to use a form based on the relationship,  $SST = SSR + SSE$ .  $SST$ , in simple terms, may be regarded as the total variation of the dependent variable  $y$ . One part of its variation is reflected by the regression ( $SSR$ ) and the other part by its residuals ( $SSE$ ) [130].  $SST$  describes the total sum of squares and is defined as:

$$SST = \sum_{i=1}^N (y_i - \bar{y})^2$$

$SSR$ , the regression sum of squares, can be calculated as follows:

$$SSR = \sum_{i=1}^N [\hat{y}(x_i) - \bar{y}]^2,$$

and the sum of squared errors is:

$$SSE = \sum_{i=1}^N \epsilon_i^2.$$

#### 4.1.1. Testing hypotheses on individual regression coefficients

Testing a statistical hypothesis means always testing a null hypothesis  $H_0$  against an alternative hypothesis  $H_1$  and eventually failing to reject or rejecting the null hypothesis [132]. It is possible to come up with an incorrect decision, i.e., to reject  $H_0$  although it is true (Type I error or  $\alpha$  error), or to accept  $H_0$  although it is false (Type II error). Thus, a common practice to keep the Type I error low, is to set the significance level of the test to  $\alpha = 0.05$ .

After estimating the coefficients  $\beta$ , we want to know whether the coefficients significantly differ from zero. Thus, we are testing whether to reject the null hypothesis  $H_0: \beta_j = 0$  for  $j = 1, \dots, p$ .

The student's  $t$  test statistic is applied to test the null hypothesis [133]:

$$t_0 = \frac{\beta_j}{se(\beta_j)} = \frac{\beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}, \text{ for } j = 1, \dots, p,$$

where  $C_{jj}$  is the diagonal element of  $(X^T X)^{-1}$  corresponding to  $\beta_j$ . The null hypothesis  $H_0: \beta_j = 0$  is rejected if  $|t_0| > t_{\alpha/2, n-p-1}$ , meaning that  $t_0$  is drawn from a  $t$  distribution with  $n - p - 1$  degrees of freedom. The null hypothesis is rejected when its value is higher than the  $t$  value corresponding to the same  $t$  distribution when  $\alpha$  is set to 0.05.

#### 4.1.2. Coefficient of determination

The coefficient of determination ( $R^2$ ) is a measurement describing how good the linear model fit to a given data set. It is defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$R^2$  can also be interpreted as the proportion of variance of the predicted variable explained by the estimated model. Thus, if the regression line fits the data perfectly  $R^2$  is equal to 1, and if the model does not fit the data at all  $R^2$  is equal to 0 [130].



However, the  $R^2$  increases with every additional variable added to the model [134]. To avoid this problem the adjusted variable has been introduced which is given by:

$$R_a^2 = R^2 - r(1 - R^2), r = \frac{p}{n-p-1} > 0, n \text{ is the sample size and } p \text{ the number of parameters.}$$

#### 4.2. Logistic regression analysis

To analyze relationships with a binary response, e.g., whether or not a patient suffers from diabetes, a logistic regression analysis can be applied [133]. The aim of a logistic regression is to model the expected value  $\mathbb{E}(y)$  or the probability of a certain outcome of  $y$  in the presence of covariates [135]:

$$\mathbb{E}(y) = P(y = 0) \cdot 0 + P(y = 1) \cdot 1 = P(y = 1 | x_1, \dots, x_k) = \pi$$

To avoid problems caused by meeting appropriate requirements such as probabilities larger than 0 or lower than 1, following model is assumed:

$$\pi_i = P(y_i = 1) = F(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}),$$

where  $x = (1, x_1, \dots, x_n)$  is the independent variable,  $\beta = (\beta_0, \beta_1, \dots, \beta_p)$  denotes the regression coefficients, the response  $y$  takes the value 0 or 1 and function  $F$  is restricted to be in interval  $[0,1]$ . Using the equation

$$F(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)},$$

yields the logit model

$$\pi_i = P(y_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)},$$

with the linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} \dots + \beta_p x_{ip}.$$

In section 5.1, four independent epidemiological studies were used to study the association with prevalent T2D (LLS [106, 107], NTR [11, 108], KORA F4 [136, 137]). In total, 306 subjects with prevalent T2D and 4619 non-diabetic controls were included. Logistic regression models

were adjusted for age, sex, BMI, use of lipid lowering medication, study specific covariates and fasting status (where appropriate) as covariates.

### **4.3. Survival analysis**

In this section, the analysis of survival times is introduced. For this analysis, the time of the onset of T2D is analyzed as well as which factors influence this time point.

In section 4.3.1 the Cox proportional hazards regression is described. This method is used to estimate the association between metabolite ratios and the onset of T2D. I assess 5-year event risk for models fitting T2D with different independent variables listed in Table 4.1 in section 4.3.1. Their performances are estimated using two measurements: the time-dependent area under the receiver operating characteristic curve (time-dependent AUC) using Bayes' theorem (section 4.3.2), and the net reclassification improvement (NRI) to measure the differences between two models (section 4.3.3).

In contrast to a logistic regression where the actual time point when the event occurs is irrelevant, the Cox proportional hazards regression analyzes data by taking into account the time until the event [138]. For the analyses, we need the survival distribution, which can be described by two functions, the survival function  $S(t)$  and the hazard function  $h(t)$  [139].

The survival function,  $S(t) = P(T > t), 0 < t < \infty$ , defines the probability that the event does not occur up to a time point  $t$ . Time point  $t$  represents a specific value of interest given that the event of the person's event time  $T$  does not occur before, i.e.  $t$  is smaller than  $T$  when the event occur and takes values between 0 and 1.

In this analysis, the event is T2D. That is,  $S(t)$  is the probability that T2D did not occur before time  $t$ .  $S(t) = 1$  at time point 0, i.e. all patients have not yet developed T2D at time point 0. The survival function must be non-increasing (monotone decreasing) over time.

A characteristic in survival analysis is, in addition to the time being included, censoring. Censoring arises when the ending events are not observed, meaning, we know that the event does not occur up to a time point  $T$ , however we do not know if and when exactly the event will occur after time point  $T$  [140].

To construct the survival distribution  $S(t)$ , unparametric estimations, such as Kaplan-Meier estimator (KM estimator) [141], become popular. This estimator is the product of conditional

probability terms. That is, each term in the product is the probability that the event does not occur in a specific ordered time point  $t_i$  given that a subject has not that event to that time point.

$$\begin{aligned}
\hat{S}_{KM}(t_i) &= \hat{S}_{KM}(t_{i-1}) \cdot P\left(\underbrace{T > t_i}_{\text{event does not occur before time point } t_i} \mid \underbrace{T \geq t_i}_{\text{event does not occur at least to time point } t_i}\right) \\
&= \prod_{k=1}^{i-1} P(T > t_k | T \geq t_k) \cdot P(T > t_i | T \geq t_i) \\
&= \prod_{k=1}^i P(T > t_k | T \geq t_k) \\
&= \prod_{k=1}^i \left(1 - P\left(\underbrace{T = t_k}_{\text{event occurs at time point } t_k}\right)\right) \\
&= \prod_{k=1}^i \left(1 - \frac{\sum_{j=1}^k \mathbf{1}(Z_j = t_k) \delta_j}{\sum_{j=1}^k \mathbf{1}(Z_j \geq t_k)}\right)
\end{aligned}$$

$$\delta_j = \begin{cases} 1 & , \text{ subject } j \text{ has event up to time } t_k \\ 0 & , \text{ otherwise} \end{cases}$$

where  $Z_j = \min(T_j, C_j)$  is the time when the event occurs and  $C_j$  is the censoring time for subject  $j$ , and  $\delta_j$  ensures that the numerator  $\sum_{j=1}^k \mathbf{1}(Z_j = t_k) \delta_j$  only counts the subjects which have the event, i.e., excluding subjects which are censored. The survival function can be also used to describe the distribution function of events, which is defined as  $F(t) = P(T > t) = 1 - S(t)$ .

The hazard function  $h(t)$ , also called instantaneous event (death, failure) rate, is defined as the probability that the event will occur in the next small time interval  $\Delta t$ , given that this event has not occurred before this interval

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

### 4.3.1. Cox proportional hazards regression analysis

The standard Cox proportional hazards model assumes a hazard function for individual  $i$  of the form  $h_i(t) = h_0(t)\exp(\beta X_i) = h_0(t)\exp(\beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \dots + X_{ip}\beta_p)$  [142], where  $h_0(t)$  is an unspecified nonnegative function of time, also called the baseline hazard,  $X_i$  are the covariates, and  $\beta$  is a  $p \times 1$  column vector of coefficients. The aim is now to estimate the hazard function and/or assess how the covariates affect it.

The hazard ratio  $\widehat{HR}$  between two subjects with fixed covariates  $X_i$  and  $X_j$  with

$$\widehat{HR} = \frac{h_i(t)}{h_j(t)} = \frac{h_0(t) \exp(\beta X_i)}{h_0(t) \exp(\beta X_j)} = \frac{\exp(\beta X_i)}{\exp(\beta X_j)} = \exp(\beta(X_i - X_j))$$

is constant over time, leading to the name proportional hazards model. Estimation of  $\beta$  is based on the partial likelihood function developed by D.R. Cox [143] and therefore is often referred to the Cox proportional hazards model. The likelihood function is called “partial” because we do not consider probabilities for all subjects, i.e., only probabilities for those subjects who undergo the event were considered and not explicitly for those who are censored [140].

For every time point  $t_j$  a risk set is formed which denoted by  $R_j$  containing all subjects for whom the event has not occurred yet and calculate the conditional probability that one of them, say subject  $i$ , among the subjects in the risk set  $R_j$  has an event at time point  $t_j$ . The log-likelihood is then the sum of these log-transformed conditional probabilities over all  $k$  events [140]:

$$L(\beta) = L_1 + L_2 + \dots + L_k = \sum_{j=1}^k L_j, \quad (4.1)$$

$$L_j = \log\left(\frac{\exp(X_i\beta)}{\sum_{r \in R_j} \exp(X_r\beta)}\right) = \log(\exp(X_i\beta)) - \log\left(\sum_{r \in R_j} \exp(X_r\beta)\right), \quad (4.2)$$

where the expression  $r \in R_j$  denotes the sum is taken over all subjects in the risk set  $R_j$  at time point  $t_j$ .

Incident T2D was obtained from the KORA S4 to F4 prospective follow-up, and the EPIC-Potsdam study. Both the KORA S4 to F4 and the EPIC-Potsdam study, have on average seven years of follow-up.

Because EPIC-Potsdam used case-cohort data from a random sub cohort  $C$  of size  $m$  and all cases from the entire cohort, the relative risk parameter  $\beta$  is estimated by maximizing the function suggested by Prentice [144]:

$$\tilde{L}(\beta) = \tilde{L}_1 + \tilde{L}_2 + \dots + \tilde{L}_k = \sum_{j=1}^k \tilde{L}_j,$$

$$\tilde{L}_j = \log \left( \frac{\exp(X_i \beta)}{\sum_{r \in \tilde{R}_j} \exp(X_r \beta)} \right) = \log(\exp(X_i \beta)) - \log \left( \sum_{r \in \tilde{R}_j} \exp(X_r \beta) \right),$$

where the expression  $r \in \tilde{R}_j$  denotes the sum which is taken over all subjects at risk in the sub cohort  $C$  and all cases from the entire cohort if they have an event at time point  $t_j$ . Thus, all subjects of the sub cohort are included in the analysis, while cases outside the sub cohort are only included in the risk set at their event time [145, 146]. To account for the case-cohort design in EPIC-Potsdam robust sandwich covariance estimates were used [147, 148]. The log-likelihood function is assumed to be the correct model and  $\beta_0$  the true value of  $\beta$ . Then, the log-likelihood function can be expanded in a Taylor series around  $\beta_0$  as introduced by Freedman [149]:  $L(\beta) = L(\beta_0) + L'(\beta_0)(\beta - \beta_0) + \frac{1}{2}(\beta - \beta_0)^T L''(\beta_0)(\beta - \beta_0) + \dots$ . Eventually, higher-order terms are ignored, resulting to quadratic log-likelihood functions, whose maximum can found by solving the likelihood equation  $L'(\beta) = 0$ :

$$L'(\beta_0) + (\beta - \beta_0)^T L''(\beta_0) = 0.$$

Thus,  $\hat{\beta} - \beta_0 \approx [-L''(\beta_0)]^{-1} L'(\beta_0)^T$ . Then the covariance is a symmetric  $p \times p$  matrix  $cov_{\beta_0} \hat{\beta} \approx [-L''(\beta_0)]^{-1} [cov_{\beta_0} L'(\beta_0)] [-L''(\beta_0)]^{-1}$ . The sandwich idea is to estimate  $L''(\beta_0)$  directly from the sample data as well as  $cov_{\beta_0} L'(\beta_0)$ . Thus,  $cov_{\beta_0} \hat{\beta}$  can be estimated with the ‘‘Huber sandwich estimator’’  $\hat{V} = (-A)^{-1} B (-A)^{-1}$ , with  $A = L''(\hat{\beta})$  and  $B = cov_{\beta_0} L'(\beta_0)$ . The square roots of the diagonal elements of  $\hat{V}$  are ‘‘robust standard errors’’

[149]. For the analysis of incident diabetes, we included 910 participants who were free of diabetes at baseline when blood was drawn but who developed T2D during follow-up, and 3,367 non-diabetic controls. For this association, a Cox proportional hazards regression analysis were performed with covariates as described by Wang-Sattler et al. [27] and Floegel et al. [46]. See Table 4.1 for details on the covariates included. The above described base model was expanded to include the ratio of valine to PC ae C32:2. The resulting model reflects a well-established prediction model for incident T2D and has been validated in several independent cohort studies [150-152]. Regression was performed using either the function *coxph* of the R package *survival* [153] or PROC PHREG in SAS [154].

**Table 4.1** Covariates used for adjustment of Cox proportional hazards regression for KORA S4 to F4 and EPIC-Potsdam studies.

<b>Variable</b>	<b>KORA S4 to F4</b>	<b>EPIC-Potsdam</b>
<b>Age</b>	years	years
<b>BMI</b>	kg/m <sup>2</sup>	kg/m <sup>2</sup>
<b>Sex (male/female)</b>	0/1	0/1
<b>Whole-grain bread intake</b>	-	g/day
<b>Waist circumference</b>	-	cm
<b>Physical activity</b>	Active/inactive	h/week
<b>Alcohol intake</b>	g/day	from beverages (nonconsumers; women >0–6,6–12,and >12 g/day; and men >0–12,12–24,and >24 g/day)
<b>Smoking</b>	Smoker/non-smoker	(never, former, current ≤20 cigarettes/day, current >20 cigarettes/day)
<b>Education</b>	-	low, medium, high
<b>Coffee intake</b>	-	cups/day
<b>Red meat intake</b>	-	g/day
<b>Prevalent hypertension</b>	-	Yes/no
<b>Systolic blood pressure</b>	mm Hg	-
<b>Use of lipid lowering medication</b>	-	Yes/no
<b>HDL cholesterol</b>	mg/dl	-
<b>Additional adjustment</b>		
<b>fasting glucose</b>	mg/dl*	mg/dl*

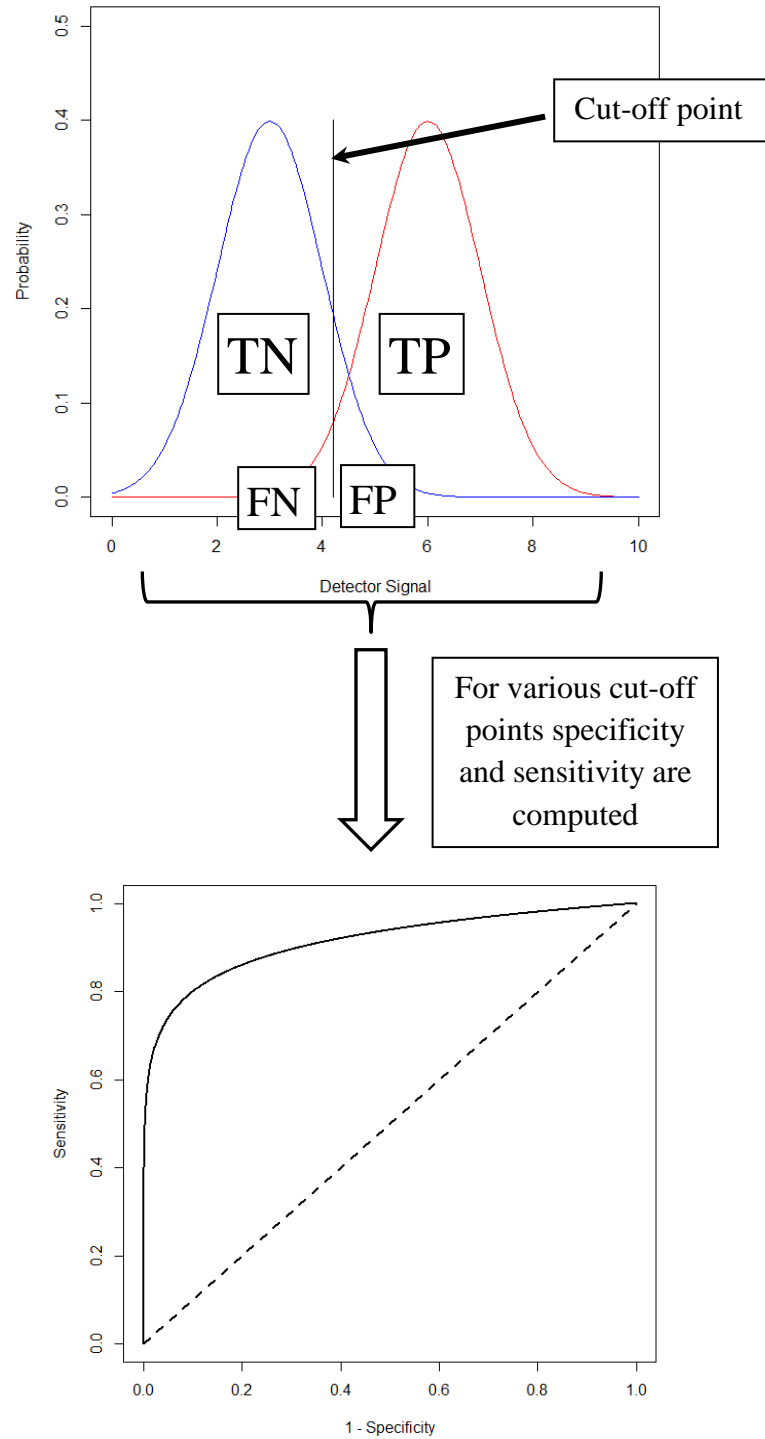
Covariate selection based on previously published paper from Wang-Sattler et al. [27] and Floegel et al. [46]. Presented are the units of the included continuous covariates or definition of categories for categorical covariates in the two studies. \* only used for the model with additional adjustment for glucose.

### 4.3.2. Time-dependent receiver operating characteristic curve

Originating from the area of signal processing, the receiver operating characteristics (ROC) analysis has become a standard evaluation tool in medical sciences to compare the true positive rate ( $TPR$ ) and the false positive rate ( $FPR$ ) [155]. Given a model which distinguishes between two classes, e.g. diabetes vs. non-diabetes, then  $TPR$  is defined as the number of positives which were correctly classified (assigned to diabetes by the model and being diabetic in reality) divided by the number of total positives (assigned to diabetes by the model no matter whether or not diabetic in reality). This is also called sensitivity, i.e. the conditional probability that the diagnostic test is positive given the subject has the disease  $P(X > c|D = 1)$ , where  $X$  is the diagnostic marker,  $c$  is the cut-off point, and  $D$  is the binary disease variable. In contrast, the  $FPR$  denotes the positive incorrectly (i.e. assigned to diabetic by the model but non-diabetic in real) classified divided by the total number of positives (all samples assigned to diabetic).

The  $FPR$  is the same as 1- specificity, which itself denotes the conditional probability that the diagnostic test is negative given the subject does not have the disease ( $P(X \leq c|D = 0)$ ). Visualization of these rates is achieved by the receiver operating characteristic (ROC) curve, which plots the  $TPR$  against the  $FPR$  for several possible cut-off points [156] (see Figure 4.3 for visualization). A ROC curve is thus a visualization of how well a diagnostic marker can classify groups with or without prevalent disease [157].

A good model shows a line close to the top-left corner whereas random decision making is reflected by a  $45^\circ$  line [156]. To compare different methods the area under the ROC curve (AUC) is computed. Thereby, an AUC close to 1 indicates a good model. However, many disease outcomes are time dependent. Thus it would be useful to estimate ROC curves that vary as a function of time. This was first introduced by Heagerty et al. [158]. In general, we want to assess how well a diagnostic marker measured at baseline can differentiate between subjects who become diseased and subjects who do not up to a time point  $t$ . Heagerty et al. presented two methods to estimate the time-dependent ROC curves [158]. The first is based on the Kaplan-Meier (KM) estimator [141] and Bayes' theorem. However, due to possible problems with respect to non-monotonicity of specificity and sensitivity Heagerty et al. provide an alternative based on a nearest neighbor estimator for bivariate distribution functions [159].



**Figure 4.3:** Two-gaussian model for the Receiver Operator Characteristic (ROC). Data are classified into positives and negatives based on a cut-off value for a biomarker: the classification results in true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (upper Figure was adapted from [160]). The true negative probability is indicated by the blue Gaussian curve and the true positive probability by the red Gaussian curve. For various cut-off points TPR and FPR are estimated and plotted into a ROC curve (right).



Therefore the definition of sensitivity and specificity are rewritten involving the time dependency:

$$\text{sensitivity}(c, t) = P(X > c | D(t) = 1)$$

$$\text{specificity}(c, t) = P(X \leq c | D(t) = 0)$$

For the KM estimator, both functions were transformed using Bayes' theorem [132]:

$$P(X > c | D(t) = 1) = \frac{(1 - S(t|X > c))P(X > c)}{(1 - S(t))}$$

$$P(X \leq c | D(t) = 0) = \frac{S(t|X \leq c)P(X \leq c)}{S(t)},$$

where  $S(t)$  is the survival function and  $S(t|X > c)$  the conditional survival function for the subset defined by  $X > c$  (i.e. classification of the disease for a given threshold  $c$ ). A simple estimator for sensitivity and specificity at a particular time point  $t$  is then given by combining the KM estimator and the empirical distribution function of the marker covariate,  $X$ , as

$$\hat{P}_{KM}(X > c | D(t) = 1) = \frac{(1 - \hat{S}_{KM}(t|X > c))(1 - \hat{F}_X(c))}{(1 - \hat{S}_{KM}(t))}$$

$$\hat{P}_{KM}(X \leq c | D(t) = 0) = \frac{\hat{S}_{KM}(t|X \leq c)\hat{F}_X(c)}{\hat{S}_{KM}(t)}$$

$$\text{with } \hat{F}_X(c) = \sum_i \mathbf{1}(X_i \leq c)/n,$$

where  $X_i$  indicates the marker covariate of subject  $i$ . Because there is no guarantee of monotonicity of specificity or sensitivity, the second approach presented by Heagerty et al. [158] involves nearest neighbour estimation of a bivariate distribution. The bivariate function can be defined as:

$$F(c, t) = P(X \leq c, T \leq t),$$

or equivalently

$$S(c, t) = P(X > c, T > t).$$

The estimator is based on the representation  $S(c, t) = \int_c^\infty S(t|X = s)dF_X(s)$ , where  $F_X(s)$  is the distribution function for  $X$ . Shown by Akritas et al. [159],  $S(c, t)$  can be estimated by

$$\hat{S}_{\lambda_n}(c, t) = \frac{1}{n} \sum_i \hat{S}_{\lambda_n}(t|X = X_i) \mathbf{1}(X_i > c),$$

where  $\hat{S}(t|X = X_i)$  denotes the estimator of the conditional survival function characterized by the smoothing parameter  $\lambda_n$  and  $X_i$  is the covariate value for subject  $i$ .

The weighted Kaplan-Meier estimator is

$$\hat{S}_{\lambda_n}(t|X = X_i) = \prod_{k=1}^i \left( 1 - \frac{\sum_{j=1}^k K_{\lambda_n}(X_j, X_i) \mathbf{1}(Z_j = t_k) \delta_j}{\sum_{j=1}^k K_{\lambda_n}(X_j, X_i) \mathbf{1}(Z_j \geq t_k)} \right).$$

Here,  $\delta_j$  is a censor indicator with  $\delta_j = 1$  if subject  $j$  has this event up to time  $T$ ,  $\delta_j = 0$  otherwise,  $Z_j$  is the follow-up time and  $K_{\lambda_n}(X_j, X_i)$  is a kernel function that depends on the smoothing parameter  $\lambda_n$ :

$$K_{\lambda_n}(X_i, X_j) = \mathbf{1}(-\lambda_n < \hat{F}_X(X_i) - \hat{F}_X(X_j) < \lambda_n).$$

Consequently,  $2\lambda_n \in (0,1)$  represents the percentage of observations included in each neighbourhood, excluding the boundaries of the distribution of  $X$ .

The estimated sensitivity and specificity are then:

$$\hat{P}_{\lambda_n}(X > c|D(t) = 1) = \frac{(1 - \hat{F}_X(c)) - \hat{S}_{\lambda_n}(c, t)}{1 - \hat{S}_{\lambda_n}(t)}$$

$$\hat{P}_{\lambda_n}(X \leq c|D(t) = 0) = 1 - \frac{\hat{S}_{\lambda_n}(c, t)}{\hat{S}_{\lambda_n}(t)},$$

where  $\hat{S}_{\lambda_n}(t) = \hat{S}_{\lambda_n}(-\infty, t)$ .

The performance of the proportional hazards regression can now likewise be assessed by the area under the receiver-operating characteristic curves (AUCs). In my project, they were calculated using the R package *survivalROC*, v1.0.3 [161] in KORA S4 to F4.

In general, when the same data are used to generate the model and to compute the AUC, the performance is quite high. This is also called “overfitting” [162]. Thus, it is recommended to estimate the performance of a model by using independent data drawn from the same underlying population. This means, to estimate the predictive performance by running internal validation approaches. In KORA, a 10-fold cross-validation was used. Here, the data got divided into ten parts (i.e. folds). The model was fitted on nine folds and evaluation was carried out on the remaining fold. This procedure of drawing the ten folds was repeated randomly 100 times to increase stability. Finally, performance estimates were averaged [163]. The R package *cvTools*, version 0.3.2 [164] was used for cross-validation.

### 4.3.3. Net reclassification improvement

The net reclassification improvement (NRI) was first introduced by Pencina et al. [165]. It describes a procedure to evaluate whether a new prediction is an improvement over an existing prediction model. This method is regarded a reasonable alternative to the comparison of AUCs [166] which can be very conservative in detecting clinically significant risk differences [165, 167]. Instead of comparing the difference in AUCs, reclassification methods stratify the estimated absolute risk into categories, and then compare if the altered model of interest can classify subjects into higher or lower risk categories more accurately than the basic model [167]. Thus, NRI distinguishes between two different risk prediction algorithms, here defined as “new” and “old” [168]. For each algorithm, the predicted probabilities are classified to a set of meaningful ordinal categories of absolute risk. If the predicted probability of the new algorithm changed the classification of a single subject into a higher category, it is defined as upward movement (up) and if the change is in the opposite direction, downward movement (down) [168]. For survival data the NRI at time point  $t$  is then computed as [167]:

$$\begin{aligned} NRI(t) &= P(up|D(t) = 1) - P(down|D(t) = 1) + P(down|D(t) = 0) - P(up|D(t) = 0) \\ &= \text{relative improvement among cases} + \text{relative improvement among controls} \end{aligned}$$

Predicted probabilities can be estimated using the KM estimator as described in section 4.3.

In our analysis, the analyzed disease was T2D. For various covariates, including the metabolite ratio valine to PC ac C32:2, the “new” model was defined, while the basic model contained only traditional covariates as used in the analysis from Wang-Sattler et al. [27]. NRI was used to assess the goodness of the models. The categories for the predicted probabilities to become

diabetic were set to 0–3.0%, 3.1–8.0%, 8.1–15.0%, and 15% [165, 169]. The analysis was performed using the R-package *nricens* [153] and compared the new model with the model using traditional risk factors and glucose as covariates.

#### 4.4. Confounding

While computing associations it is very important to include appropriate covariates to ensure that model assumptions are met, to reduce the noise in the response variable, and to avoid confounding of the investigated association. Confounders are variables that are not under investigation but are related to both the response variable and independent variable [156].

Thus, to avoid false associations between response and independent variable it is necessary to adjust for confounders, i.e., by including them as covariates in the model or computing the residuals of the association between the variable of interest and the confounder. For example, a specific age-related metabolite as a response variable is associated with age-related BMI. However, this association diminished as soon we include age into this model. Another example is the DNA methylation data which were measured from whole blood samples. Blood is composed of different cell types that differ strongly in DNA methylation [127]. Therefore, the data of 500 most cell type-specific CpG sites was used to infer cell proportions from the whole blood data [127].

#### 4.5. Violation of the assumption of the underlying regression model

As described in section 4.1 and 4.2, the linear and logistic regression models are parametric models and thus, assume a certain distribution. Particularly, the linear model assumes a normally distributed error term, which is the same as assuming a normally distributed response conditioning on the covariates:

$$\epsilon \sim N(0, \sigma^2 I_n)$$

$$y = X\beta + \epsilon \sim N(X\beta, \sigma^2 I_n).$$

However, the metabolite levels which are used for the linear association analysis are not always normally distributed. Thus, to achieve approximate normality, the response variable is transformed in several ways, such as log transformation, inverse normal rank transformation [170] or Box-Cox transformation [171]. Log transformation does not always ensure normality

and inverse normal rank transformation is very conservative and can also reduce statistical power in some circumstances [170]. Thus, an alternative is the Box-Cox transformation, which is explained in more detail in the section 4.5.1.

In section 4.5.2 the diagnostic plots are introduced. To check whether the assumptions for linear regression are still valid for the significant associations in the interaction analysis, four diagnostic plots are depicted: a plot of residuals against fitted values, a Scale-Location plot of the square root of the absolute values of the residuals ( $\sqrt{|\text{residuals}|}$ ) against fitted values, a Normal Q-Q plot, and a plot of residuals against leverages.

Furthermore, the proportional hazards assumption, that the hazard functions are proportional over time (see section 4.3.1), is the key to construct the partial likelihood, because of which the baseline hazard function is canceled out from the partial likelihood factors [139]. However, in practice the assumption is an approximation, and minor violations are unlikely to have major effects on inferences on model parameters [139]. For the Cox proportional hazards regression analysis, we used the plot of the Schoenfeld residuals against the time.

#### 4.5.1. Box-Cox transformation

Often transformation of data is necessary to ensure a symmetric distribution to use familiar and traditional statistical techniques, such as linear regression.

One of the common transformations is the Box-Cox transformation introduced by G.E. Box and D.R. Cox in 1964 [171]. The aim of the Box-Cox transformation is to obtain normal distribution of the dependent variable  $y$ , thus trying to induce normally distributed residuals.

The transformation has following form:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & , \lambda \neq 0 \\ \ln(y) & , \lambda = 0 \end{cases}, \text{ with } \lambda \in \mathbb{R}.$$

To accommodate also negative  $y$  values, an extended method includes  $\lambda_2$ :

$$y(\lambda) = \begin{cases} \frac{(y + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & , \lambda_1 \neq 0 \\ \ln(y + \lambda_2) & , \lambda_1 = 0 \end{cases}, \text{ with } \lambda = (\lambda_1, \lambda_2).$$

It is common to choose  $\lambda_2$  such that  $\lambda_2 + y > 0$  for any  $y$ . Thus, an appropriate  $\lambda_1$  can achieve a rough symmetry of the data. The parameter  $\lambda_1$  can be estimated by maximizing the log-likelihood function for the Gaussian distribution:

$$L(\lambda_1) = -\frac{n}{2} \ln(s^2(\lambda_1)) + (\lambda_1 - 1) \sum_{i=1}^n \ln(x_i), \quad (4.4)$$

with  $n$  the sample size and  $s^2(\lambda_1)$  the sample variance of the data after transformation with the parameter  $\lambda_1$ . The sum in the second term of Equation (4.4) uses the untransformed data.

Hence, to ensure normal distribution, in the interaction analysis metabolite levels and metabolite ratios are represented by their Box–Cox transformed residuals after regressing on age, sex, and BMI. The R package *car* [172] was applied to compute the Box–Cox transforms.

#### 4.5.2. Diagnostic plots

It is not sufficient to just compute the linear regression or Cox proportional hazards regression and accept the results uncritically. Unsatisfied assumptions underlying the computations might cause in wrong results [130]. Thus, to check that the assumptions of the linear regression and the Cox proportional hazard regression are valid, four diagnostic plots are introduced for the linear regression (section 4.5.2.1 - 4.5.2.4) and one additional for the Cox proportional hazard regression (section 4.5.2.5).

##### 4.5.2.1. Residuals vs. fitted values plot

The most frequently created plot is a scatter plot of residuals on the y-axis and fitted values, i.e., estimated response on the x-axis. The plot is used to detect non-linearity, unequal error variances, and outliers. Since the assumptions of a linear regression pertain to the residuals, it is important to examine the residuals for consistency. Ideally, this plot should show a horizontal line with equally spread points. If the scatterplots between residuals and fitted values show a pattern, then the relationship may be nonlinear and the model or the response variable will need to be modified accordingly. Thus, in this case might be "heteroscedasticity" in the errors, i.e., the variance of the residuals may not be constant [130].

**4.5.2.2. Scale-Location plot of the square root of the absolute values of the residuals ( $\sqrt{|\text{residuals}|}$ ) against fitted values**

Scale-Location plot is also called Spread-Location plot and is similar to the plot mentioned above. But instead of using linear residuals, the square root of the absolute values of the residuals is computed. Thus, it might be easier to reveal trends in the magnitudes of residuals. For a good model, the values should be randomly distributed. Thus, this plot shows if residuals are spread equally along the ranges of predictors [173] and is useful to check whether the assumption of equal variance (homoscedasticity) is valid.

**4.5.2.3. Normal Q-Q plot**

Normal Q-Q plot or Normal quantile-quantile plot is an easy graphical method to see whether the residuals are normally distributed. The idea is that if the residuals are normally distributed the empirical quantiles should correspond to the theoretical quantiles. Thus, this plot compares the shape of the distribution of the data to a normal (or bell-shaped) distribution [174].

If we have  $n$  observations, we can compute  $n$  empirical quantiles. The quantile for the  $i$ th observation ( $i = 1, \dots, n$ ) of the sorted data is computed as  $p_i = \frac{i-0.5}{n}$ . Since in the normal distribution the 100% quantile is infinity, 0.5 is subtracted in the equation. The observations are then standardized with mean  $\mu = 0$  and standard variance  $\sigma^2 = 1$ . The corresponding theoretical quantile is computed from the inverse normal distribution. The empirical quantiles are on the y-axis while the theoretical quantiles are on the x-axis. The data points fall along an approximately straight line when the data are from a normal distribution, otherwise the data are from other distributions [174].

**4.5.2.4. Plot of residuals against leverages**

The influence of one particular observation on the estimation results in a linear model is measured by the leverage  $h_{ii}$  [175] which is the  $i$ th diagonal element of the hat matrix  $H = X(X^T X)^{-1} X^T$ . The leverage ranges from  $\frac{1}{n}$  to 1. If a leverage is large, i.e., close to 1, it has a considerable influence on the estimation results and indicate some unusual covariate values in  $x_i$  [135]. Thus, the plot of residuals against leverages indicates potential influential cases. In addition, the Cook's distance is plotted as a dashed red line. The Cook's distance is defined by  $D_i = \frac{(\hat{y}_i - y)^T (\hat{y}_i - \hat{y})}{p \cdot \hat{\sigma}^2}$ , where  $\hat{y}_i$  denotes the estimator for  $y$  that uses all observations with exception of the  $i$ th observation,  $\hat{y}$  the estimator for  $y$  that uses all observations,  $p$  is the

number of parameters, and  $\hat{\sigma}^2$  is the estimated standard variance [135]. According to Fahrmeir et al [135], observations that have a Cook's distance  $D_i > 1$  should be always examined.

#### 4.5.2.5. *Plot of Schoenfeld residuals against the time*

The plot of the Schoenfeld residuals against the time is an approach to check the proportional hazards assumption, i.e., that the hazard ratio comparing any two specification of independent variables is constant over time [140]. If this assumption is correct, a plot of the residuals of all individuals against the time for the examined covariate will yield a pattern of points that are centered at zero [139].

The residuals are derived from the partial log-likelihood function as described in section 4.3.1, Equation 4.1, and 4.2. The score function is then the derivative of Equation 4.1 (see section 4.3.1):

$$L'(\beta) = \sum_{j=1}^k X_i - \log \left( \sum_{r \in R_j} X_r \cdot \frac{\exp(X_h \beta)}{\sum_{h \in R_r} \exp(X_h \beta)} \right),$$

where  $i$  denotes subject  $i$ ,  $k$  is the number of events, and  $R_j$  is a risk set containing all subjects for whom the event has not occurred yet at time point  $t_j$ ,  $R_r$  is a risk set containing all subjects for whom the event has not occurred yet for time point  $t_r$  [140]. The Schoenfeld residuals are the individual terms of the score function, and each term is the observed value of the covariate for patient  $i$  minus the expected value  $E(\bar{X}_j)$ , which is a weighted sum, with weights given by  $\frac{\exp(X_h \beta)}{\sum_{h \in R_r} \exp(X_h \beta)}$ , of the covariate values for subjects at risk at that time. Each weight may be viewed as the probability of selecting a particular person from the risk set at time  $t_j$  [139]. For an estimate  $\hat{\beta}$ , the residual for the  $i$ th failure time is

$$\hat{r}_j = X_i - \sum_{k \in R_j} X_r \cdot \frac{\exp(X_h \hat{\beta})}{\sum_{h \in R_r} \exp(X_h \hat{\beta})} = X_i - \bar{X}_i.$$

Thus, the Schoenfeld residual for a particular variable is the observed value of the variable minus a weighted average of the variable for the other subjects still at risk at time  $t_j$  [140].



### 4.6. Multiple testing

Testing many associated hypotheses simultaneously leads to a so called multiple testing problem [132]. This means that increasing the number of independent tests also increases the probability of Type I error:

$$P(\text{Type I error}) = 1 - P(\text{no false rejection}) = 1 - \prod_{j=1}^m 1 - \alpha = 1 - (1 - \alpha)^m,$$

where  $m$  is the number of tests. To adjust for that, a “family-wise” error-rate is used, i.e. the p-value threshold is set to  $p \leq \frac{\alpha}{m}$ , for each individual test, also known as the Bonferroni correction.

### 4.7. Meta-analysis

Meta-analysis is the procedure of combining and pooling results from different studies [176]. Thereby, a common true parameter  $\beta$  underlying all studies is assumed which has to be estimated.

In the fixed-effect meta-analysis all  $K$  independent studies’ estimates  $\hat{\beta}_k, k = 1, \dots, K$  are assumed to have a common mean  $\beta$  and a common error variance  $\sigma^2$  [177]. Thus, to obtain the pooled effect, all effect sizes across all studies are averaged:

$$\hat{\beta}_{unweighted} = \sum_{k=1}^K \frac{1}{K} \hat{\beta}_k$$

However, to control for different contributions of studies, the estimated  $\beta$  is calculated with different weights per study:

$$\hat{\beta}_{weighted} = \frac{\sum_{k=1}^K w_k \hat{\beta}_k}{\sum_{k=1}^K w_k}.$$

To increase the power to detect small metabolite ratio effects on T2D, results across the three studies, KORA, LLS, and NTR, are combined in a meta-analysis using the weighted mean of the individual  $\hat{\beta}_k$ , the coefficients from the logistic regressions. The weights are chosen to be the inverse variance from the coefficients  $w_k = 1/se(\hat{\beta}_k)^2$  (also called pooled inverse variance-weighted beta coefficient) as described by de Bakker et al. [178].

In section 5.1, meta-analysis was only performed on those metabolites that were successfully measured in all three cohorts. Metabolite ratios are represented as the z-score  $z_{meta}$ , i.e. a value following the  $\chi^2$  – distribution with mean  $\mu = 0$  and standard error  $\sigma^2 = 1$ , through following

formula:  $z_{meta} = \frac{\beta}{se(\beta)}$ . The fixed-effects meta-analysis was done using the R package *Meta* v4.3-2 [179]. For the association between the ratios and incident diabetes a meta-analysis between KORA S4 to F4 and EPIC-Potsdam was conducted. To allow comparison across cohorts and facilitate meta-analysis metabolite level data were log-transformed followed by z-scaling before analysis.

### 4.8. P-gain

As outlined in the introduction all possible combinations of ratios between metabolite pairs are analyzed in a hypothesis-free approach. To measure the improvement of using metabolite ratios instead of single metabolite concentrations in section 5.1 a so-called p-gain is computed [4, 180]. It compares the change of the p-value when using the ratio and the smaller of the two p-values when using the two single metabolite concentrations individually.

$$p - gain\left(\frac{M1}{M2} \middle| X\right) := \frac{\min(P(M1|X), P(M2|X))}{P\left(\frac{M1}{M2} \middle| X\right)},$$

where  $P(M1|X)$  denotes the p-value of the association between trait  $X$  and metabolite  $M1$ ,  $P(M2|X)$  the p-value of the association regression between trait  $X$  and metabolite  $M2$ , and  $P\left(\frac{M1}{M2} \middle| X\right)$  is the p-value computed using the ratio between metabolites  $M1$  and  $M2$ . In this thesis,  $X$  is binary and denotes the occurrence of T2D.

A p-gain is considered to be significant if it is above the significance level  $\text{gain} \geq \frac{B}{2 \cdot \alpha}$ , with  $B$ =number of metabolites tested and significance level  $\alpha$  (set to 0.05). If it is significant, it suggests that the metabolite ratio can explain the outcome better than the two single metabolites.

### 4.9. Pulver

In this section the R package *pulver*, which is implemented to compute linear regressions with interaction term for a very large number of linear regression models, is introduced in more detail.

### 4.9.1. Linear regression with interaction term

As previously mentioned, I want to examine possible interactions between two variables, i.e., examine each combination of covariates from given matrices  $X$  and  $Z$  and their impact on  $Y$ .

Assuming following model based on multiple linear regression:

$$y = \beta_0 + \beta_1 xz + \beta_2 x + \beta_3 z + \epsilon, \quad \epsilon \sim i.i.d. N(0, \sigma^2),$$

where  $y$  is the outcome variable,  $x$  and  $z$  are covariates, and  $xz$  is the interaction (product) of covariates  $x$  and  $z$ . All variables are quantitative and vectors. As outlined in the section 2.8 there is still a lack of fast software computing the significance of the interaction term. For this reason, the R package *pulver* was implemented. The acronym *pulver* denotes parallel ultra-rapid p-value computation for linear regression interaction terms. This R package tests the null-hypothesis  $\beta_1 = 0$  against the alternative  $\beta_1 \neq 0$ . Estimating the coefficients  $\beta_2$  and  $\beta_3$  is not of relevance for this matter. Thus, it is possible to take a computational shortcut. By centering, such that  $\sum_i y_i = \sum_i x_i = \sum_i z_i = \sum_i xz_i = 0$ , and orthogonalizing the variables, the multiple linear regression problem gets converted into a simple linear regression without intercept. Then, the Student's  $t$ -test statistic was computed for the coefficient  $\beta_1$  as a function of the Pearson correlation coefficient  $r$  between  $y$  and the orthogonalized  $xz$ :  $t_{\beta_1} = r\sqrt{(DF/(1-r^2))}$ , where  $DF$  are the degrees of freedom of the linear regression model. By computing the  $t$ -test statistic based on the correlation coefficient, which itself has a very simple expression in the simplified model, fitting the entire model including estimation of coefficients  $\beta_2$  and  $\beta_3$  is avoided. Consequently, only the interaction's regression coefficient is taken into account. The inputs  $X$  and  $Z$  for the interaction analysis can be either vectors or matrices. If matrices are given, the algorithm will iterate through all columns in order to compute the interaction analysis.

### 4.9.2. Theory underlying pulver

Let  $X = \begin{pmatrix} 1 & x_1 & z_1 & w_1 = x_1 \cdot z_1 \\ 1 & x_2 & z_2 & w_2 = x_2 \cdot z_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & z_n & w_n = x_n \cdot z_n \end{pmatrix}$ , with  $x_i, z_i, w_i \in \mathbb{R}$ , and the unknown regression

coefficients  $\beta^T = (\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3)$ . Then, the general linear model given in Equation 4.1 reduces to the following linear regression model

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2),$$

with  $\epsilon^T = (\epsilon_1, \dots, \epsilon_n)$  being independent and identical distributed (*i. i. d.*).

The null hypothesis that  $\beta_3 = 0$  against the alternative hypothesis that  $\beta_3 \neq 0$  is tested, where  $\beta_3$  is the regression coefficient of  $w$ . To eliminate the intercept  $\beta_0$ , all variables are centered, such that  $\sum_i y_i = \sum_i x_i = \sum_i z_i = \sum_i w_i = 0$ , to obtain the following simplified regression model:

$$y = \beta_1 x + \beta_2 z + \beta_3 w + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \text{ i. i. d.}$$

(For simplicity, the notations from above are retained for the simplified model (for variable names  $y, x, z$ , and  $w$  for the centered variables, regression coefficients, error term).)

The vectors  $x, z$  and  $w$  span a subspace  $S$  of  $\mathbb{R}^n$ . The ordinary least-squares (OLS) estimates  $\hat{\beta}$  of  $\beta$  are found by minimizing the residual sum of squares over  $y - X\beta$ :

$$\hat{\beta} = \arg \min_{\beta} (y - X\beta)^T (y - X\beta).$$

Geometrically, this means that  $\hat{\beta}_1, \hat{\beta}_2$ , and  $\hat{\beta}_3$  must be selected such that

$$y' = \hat{\beta}_1 x + \hat{\beta}_2 z + \hat{\beta}_3 w \quad (4.5)$$

is the orthogonal projection of  $y$  onto  $S$ , the subspace spanned by  $w, x$ , and  $z$ . It can be shown that if  $w, x$ , and  $z$  form an orthogonal basis of  $S$ , the coefficients of the orthogonal projection  $y'$  of  $y$  onto  $S$  are given by

$$\hat{\beta}_1 = \frac{\langle y, x \rangle}{\langle x, x \rangle}, \quad \hat{\beta}_2 = \frac{\langle y, z \rangle}{\langle z, z \rangle}, \quad \hat{\beta}_3 = \frac{\langle y, w \rangle}{\langle w, w \rangle} \quad [181].$$

Unlike the usual formula for computing OLS coefficient estimates ( $\hat{\beta} = (X^T X)^{-1} X^T y$ , see section 4.1), this new formula does not involve an expensive matrix inversion, but instead it is easy and fast to compute.

In general,  $w, x$ , and  $z$  do not form an orthogonal basis. Thus, to test the null hypothesis, the following steps are carried out:

1. Create an orthogonal basis  $v_1, v_2$ , and  $v_3$  for  $S$  based on  $x, z$ , and  $w$ , respectively.
2. Compute  $y'$ , the orthogonal projection of  $y$  onto  $S$ , using the orthogonal basis created in step 1.

3. Deduce the estimate of the regression coefficient  $\hat{\beta}_3$  for  $w$  from the regression coefficients for  $y'$ .
4. Compute the Student's  $t$ -test statistic to test  $\hat{\beta}_3 = 0$  as a function of the correlation coefficient  $r$  between  $y'$  and  $\hat{\beta}_3 v_3$ .

1. Create an orthogonal basis for  $S$

Let

$$v_1 = x,$$

$$v_2 = z - \text{proj}(z, v_1), \text{ and}$$

$$v_3 = w - \text{proj}(w, v_1) - \text{proj}(w, v_2),$$

with

$$\text{proj}(a, b) = \frac{\langle a, b \rangle}{\langle b, b \rangle} b$$

is the orthogonal projection of  $a$  onto  $b$ . The vectors  $v_1$ ,  $v_2$ , and  $v_3$  form an orthogonal basis of  $S$ . By construction, we clearly observe that  $v_1$  is dependent on  $x$  only,  $v_2$  is dependent on  $z$  and  $x$  and  $v_3$  depends on  $x$ ,  $z$ , and  $w$ .

2. Orthogonally project  $y$  onto  $S$

The orthogonal projection  $y'$  of  $y$  onto  $S$  has the form

$$y' = \beta'_1 v_1 + \beta'_2 v_2 + \beta'_3 v_3 \quad (4.6)$$

where

$$\beta'_i = \frac{\langle y, v_i \rangle}{\langle v_i, v_i \rangle} \quad (i = 1, 2, 3).$$

3. Deduce the estimate of  $w$ 's regression coefficient

We want to estimate the regression coefficient  $\beta_3$  of the vector  $w$  given in Equation 4.5 using Equation 4.6. The vector  $w$  occurs in the calculation of  $v_3$  but not in  $v_1$  or  $v_2$ . This allows us to write  $y'$  as

$$\begin{aligned}
 y' &= \beta'_1 v_1 + \beta'_2 v_2 + \beta'_3 v_3 \\
 &= \beta'_1 v_1 + \beta'_2 v_2 + \beta'_3 (w - \text{proj}(w, v_1) - \text{proj}(w, v_2)) \\
 &= \beta'_1 v_1 + \beta'_2 v_2 + \beta'_3 w - \beta'_3 \text{proj}(w, v_1) - \beta'_3 \text{proj}(w, v_2) \\
 &= \beta'_3 w + \beta'_1 v_1 + \beta'_2 v_2 - \beta'_3 \text{proj}(w, v_1) - \beta'_3 \text{proj}(w, v_2) \\
 &= \beta'_3 w + \beta'_1 v_1 + \beta'_2 v_2 - \underbrace{\beta'_3 \frac{\langle w, v_1 \rangle}{\langle v_1, v_1 \rangle}}_{\text{scalar}} v_1 - \underbrace{\beta'_3 \frac{\langle w, v_2 \rangle}{\langle v_2, v_2 \rangle}}_{\text{scalar}} v_2 \\
 &= \beta'_3 w + c \left( \begin{array}{c} v_1, v_2 \\ c(x) \quad c(x, z) \end{array} \right) \\
 &= \beta'_3 w + c(x, z)
 \end{aligned}$$

where  $c(\dots)$  represents a linear combination of  $x$  or  $x$  and  $z$ , accordingly. This allows us to identify  $\beta_3$ , and we estimate the regression coefficient of  $w$  in Equation 4.5:

$$\hat{\beta}_3 = \beta'_3 = \frac{\langle y, v_3 \rangle}{\langle v_3, v_3 \rangle}.$$

4. Compute the Student's  $t$ -test statistic to test  $\beta_3 = 0$  as function of the correlation coefficient  $r$  between  $y'$  and  $\beta_3 v_3$

In this last step, it is presented how the Pearson's correlation coefficient  $r$  can be used to test  $\beta_3 = 0$  in a linear regression model instead of using the Student's  $t$ -test statistic, i.e. testing  $t \geq t^*$  for significant threshold  $t^*$ . The Pearson's correlation coefficient  $r$  between  $y'$  and  $v_3$  (both centered) is computed as follows:

$$r = \frac{\sum_{i=1}^N y_i' v_{3i}}{\|y'\| \|v_3\|} = \frac{\sum_{i=1}^N y_i' v_{3i}}{\sqrt{\sum_{i=1}^N y_i'^2} \sqrt{\sum_{i=1}^N v_{3i}^2}}$$

$$\text{with } \|a\| = \sqrt{\langle a, a \rangle},$$

and  $\langle a, b \rangle = \sum_{i=1}^n a_i b_i$  being the inner product of vectors  $a$  and  $b$  in  $\mathbb{R}^n$ .

Then it follows that the null hypothesis  $\hat{\beta}_3 = 0$  can be rejected if  $r \geq t^* \cdot \sqrt{\frac{1}{DF+t^{*2}}}$ .

The fact that  $v_1, v_2$ , and  $v_3$  are orthogonal means that  $\hat{\beta}_3$  is actually the OLS estimate of the correlation coefficient  $r$  in the simple linear regression

$$y' = \hat{\beta}_3 v_3 + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \text{ i. i. d.}$$

The Student's  $t$  statistic to test for coefficient  $\beta_3 = 0$  is given by

$$t = \frac{\hat{\beta}_3}{se(\hat{\beta}_3)}$$

and it has a Student's  $t$  distribution with  $DF = n - 4$  degrees of freedom. Subtracting four results from the number of regression coefficients in the initial model and the estimated variance of  $\epsilon$ :  $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, s^2$ .

From the theory of simple linear regression, the following relationships are known (e.g., see Snedecor and Cochran 1967 [182], chapter 7.3, p. 175 ff.):

$$\text{a) } \hat{\beta}_3 = r \frac{se(y')}{se(v_3)}$$

$$\text{b) } se(\hat{\beta}_3) = s / \sqrt{\sum_{i=1}^n v_{3i}^2}$$

$$\text{c) } s^2 = \frac{1-r^2}{DF} \sum_{i=1}^N y_i'^2$$

$$\text{d) } se(a) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n a_i^2}, \text{ with } a \in \mathbb{R}^n \text{ and } \sum_i a_i = 0,$$

where  $\hat{\beta}_3$  is the OLS estimate of Equation 4.5;  $se(y')$  and  $se(v_3)$  are the sample estimates of the standard deviations of  $y$  and  $v_3$ , respectively;  $se(\hat{\beta}_3)$  is the estimate of the standard deviation of  $\hat{\beta}_3$ ;  $s^2$  is the OLS estimate of  $\sigma^2$ , the variance of the error term  $\epsilon$ ;  $r$  is the Pearson's correlation coefficient of  $y'$  and  $v_3$ ; and  $DF$  is the degree of freedom.

After plugging Equations a–d into the formula for the Student's  $t$ , we obtain the following:

$$\begin{aligned}
 t &= \frac{\hat{\beta}_3}{se(\hat{\beta}_3)} \\
 &= \frac{r se(y') \sqrt{\sum_{i=1}^N v_{3i}^2}}{se(v_3) \sqrt{\frac{(1-r^2)}{DF} \sum_{i=1}^N y_i'^2}} \\
 &= \frac{r \sqrt{DF}}{\sqrt{(1-r^2)}} \cdot \frac{se(y') \sqrt{\sum_{i=1}^n v_{3i}^2}}{se(v_3) \sqrt{\sum_{i=1}^n y_i'^2}} \\
 &= \frac{r \sqrt{DF}}{\sqrt{(1-r^2)}}
 \end{aligned}$$

Thus, as stated is above,  $H_0$  can be rejected if  $r \geq t^* \cdot \sqrt{\frac{1}{DF+t^{*2}}}$ , for significance threshold  $t^*$ .

### 4.9.3. Avoiding redundant computations

Computation time can be saved by sophisticated arrangement of computations. The naïve approach would be to iterate through three nested for-loops – one for each matrix – with all computations happening in the innermost loop. However, the pseudocode below shows how computations can be moved out of the innermost loop to avoid redundant computations (Figure 4.4).



```

Input: matrices
X (number of observations×number of variables),
Y (number of observations×number of variables),
Z (number of observations×number of variables),
p-value threshold

Output: table with p-values < p-value threshold and columns
matching variable names in X, Y, Z
(number of identified associations × 4 (= p-value, X, Y, Z))

Compute r-value threshold using p-value threshold
for x in variables in matrix X
  for z in variables in matrix Z
    Orthogonalize z wrt x
    Compute interaction xz
    Center xz
    Orthogonalize xz wrt x and z
    Compute  $\|xz\|$  (norm of xz)
    for y in variables in matrix Y
      Orthogonalize y wrt x and z
      Compute  $\|y\|$ 
      Compute  $r = \langle y, xz \rangle / (\|y\| \cdot \|xz\|)$ 
      if  $r > r\text{-value threshold}$ 
        Calculate p-value

```

**Figure 4.4:** Pseudocode of the function *pulverize*. Here we choose as test statistic the sample correlation between response variable  $y$  and the interaction of variables  $x$  and  $z$ . Prior all variables must be orthogonal to each other to build the orthogonal projection of  $y$  onto  $S$  and thus only using the correlation as test statistic as explained in section 4.9.2. To avoid redundant computations, if possible, some computations were moved out of the innermost loop.

#### 4.9.4. Programming language and general information about the program

The algorithm is provided in an R package called *pulver*. Due to speed considerations, the core of the algorithm is implemented in C++. *pulver* was implemented in R version 3.3.1 and the

C++ code was compiled with the gcc compiler version 4.4.7. To integrate C++ into R the R package *Rcpp* [183] (version 0.12.7) is used. Additionally, to find out whether C/Fortran can improve on the performance of C++ the algorithm was also implemented in a combination of C and Fortran via the C interface provided by R.

Parallelization of the middle loop was realized by OpenMP version 3.0 [184]. Furthermore, the order through which the matrices  $X$  and  $Z$  will be iterated, is changed when the number of columns of matrix  $X$  is greater than matrix  $Z$ . The middle loop therefore runs over more variables than the outer loop, aiming at minimizing the amount of time required to coordinate parallel tasks. Thus the amount of work per thread was maximized. For efficiency reasons, the program does not allow additional covariates beyond  $x$  and  $z$ . If additional covariates are desired, the outcome  $y$  must be replaced by the residuals from the regression of  $y$  on the additional covariates. Missing values in input matrices are replaced by the respective column mean. The *pulver* package can be used as a screening tool for scenarios where the number of models (number of variables in matrix  $X \times$  number of variables in matrix  $Z$  for several outcome variables) is too big for conventional tools. By specifying a p-value threshold saved results can be limited to models with interaction term p-values below the threshold. Thereby, the size of the output can be largely reduced. After initial screening, additional model characteristics for the significant models, e.g., effect estimates and standard errors, can be obtained via traditional methods such as R's *lm* function.

The user has access to *pulver*'s functionality via two functions: *pulverize* and *pulverize\_all*. The *pulverize* function expects three numeric matrices and returns a table with p-values for models with interaction term p-values below the (optionally specified) p-value threshold. The wrapper function *pulverize\_all* expects the names of files containing  $X$ ,  $Y$ ,  $Z$  matrices, calls *pulverize* to do the actual computation, and returns a table in the same format as *pulverize*.

#### **4.9.5. Comparison with other R tools for running linear regressions**

As illustrated in Figure 4.5, the inputs for the interaction analysis can be vectors or matrices. Currently, *pulver* is the only available option for users who want all the inputs to be matrices. It is possible to adapt other tools to all-matrix inputs, but the resulting code is not optimized for this use and will be too slow for practical purposes. Thus, the speed of this package was

compared to that of R's built-in *lm* function and the R package *MatrixEQTL* [19] (version 2.1.1).

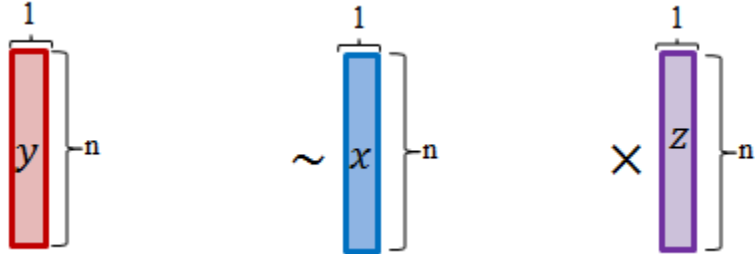
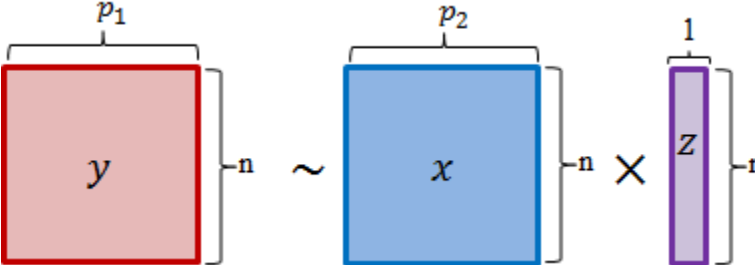
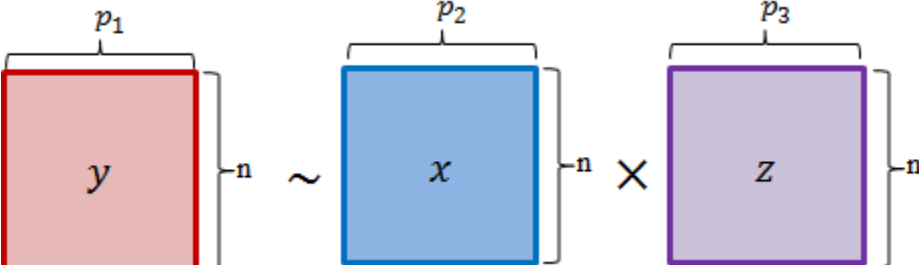
The R package *MatrixEQTL*, written in R, uses for computation matrix operations leading to a rapid computation of linear regressions. Similar to the R package *pulver* it computes the correlation between the scaled and orthogonalized outcome and variable of interest. The p-value is only computed if the correlation exceeds a specified p-value. However, as illustrated in Figure 4.5, *MatrixEQTL* currently only assesses the significance of the interaction between all variables of one matrix  $X$  and the last variable of the other matrix  $Z$ . Thus, for the benchmark, over all variables in matrix  $Z$  are iterated.

R's built-in *lm* function, written in Fortran, uses the standard approach for computation, thus estimates the beta coefficient via matrix inversion ( $\hat{\beta} = ((x \cdot z)^T(x \cdot z))^{-1}(x \cdot z)^T y$ ). By calling the *summary.lm* function, written in R, the p-value and many other statistical variables, such as coefficient of determination, F-statistic, and square root of the estimated variance of the random error are computed. The *lm* function is only able to compute one outcome variable and one interaction term. For computation, the interaction of each variable of each matrix, the *lm* function is called in the inner most loop of the three for-loops, one for-loop for each matrix.

The parallelization feature of the function *pulverize* that is part of the R package *pulver* was not used because it is not available in R's *lm* function or *MatrixEQTL* and thus would lead to biased results if only the speed of the functions is considered. However, parallelization is possible and can lead to speedups, although sublinear. Each scenario was run 200 times using the R package *microbenchmark* (version 1.4-2.1, <https://CRAN.R-project.org/package=microbenchmark>) and only results with a p-value below 0.05 were written into a file.

The complexity of *pulver* in asymptotic notation is the product of the number of variables of the matrices  $X$  (x-columns),  $Y$  (y-columns),  $Z$  (z-columns), and the number of samples ( $n$ ):  $O(\text{x-columns} \cdot \text{y-columns} \cdot \text{z-columns} \cdot n)$ . This is similar to the R package *MatrixEQTL*.

In contrast, the complexity of R's build-in function *lm* is  $O(\text{x-columns} \cdot \text{y-columns} \cdot \text{z-columns} \cdot n^2)$  as this function computes the linear regression by solving the Ordinary Least Squares problem  $\hat{\beta} = ((x \cdot z)^T(x \cdot z))^{-1}(x \cdot z)^T y$  [18].

	R <i>lm</i>	R <i>MatrixEQTL</i>	R <i>pulver</i>
	✓	✓	✓
	---	✓	✓
	---	---	✓

**Figure 4.5:** Input types required for different R tools to run linear regressions. Comparison of different input types handled by the R tools *lm*, *MatrixEQTL*, and *pulver* for computation of the linear regression with interaction term. By the braces the dimensions of the matrices are depicted. The R's build-in function *lm* can only compute the linear regression with interaction term using one variable with  $n$  observations per call. The R package *MatrixEQTL* can compute simultaneously the linear regression for each of  $p_1$  variables from the outcome matrix  $Y$  and the interaction term of the matrix  $X$  with  $p_2$  variables and the vector  $Z$ . In contrast, *pulver* in addition iterates through  $p_3$  variables of the matrix  $Z$  and finally computes the linear regression for each column of matrices  $Y$ ,  $X$ , and  $Z$ .

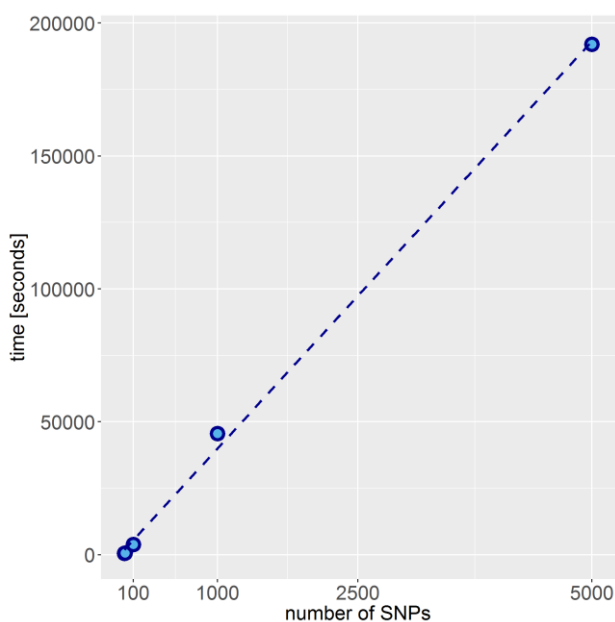
$p_1$ ,  $p_2$ , and  $p_3$  are  $\in \mathbb{N}$

### 4.9.6. Feature reduction

This R package is in a first step applied to the KORA data set comprising 345,372 CpG sites, 9,143,401 SNPs (coded as values between 0 and 2 according to an additive genetic model), and 557 metabolites measured by the Biocrates and Metabolon platform.

The whole analysis of  $1.8 \cdot 10^{15}$  models would have taken a very long time even with *pulver*.

For the estimation of the time required to analyze the whole dataset, scenarios using all CpG sites, all metabolites, and different numbers of SNPs (100, 1,000, 2,000, 4,000, and 5,000; see Figure 4.6) was run. Subsequently, the run time was extrapolated to estimate the required time to analyze all SNPs. Due to time limitations, each of the scenarios defined above was run only once. The required runtime to analyze the complete dataset with parallelizing the work across 40 processors was estimated to take 1.5 years.



**Figure 4.6:** Time required to analyze 100, 1,000, 2,000, 4,000, and 5,000 SNPs and 345,372 CpGs. A line was fitted through all time points.

Therefore, only SNPs that had previously shown significant associations with at least one metabolite were selected [5, 185]. Subsequently, interaction terms DNA methylation and SNPs were added into the models.

The final data set comprised 345,372 CpG sites, 117 SNPs, and 16 metabolites from the Biocrates and 345,372 CpG sites, 6,406 SNPs, and 376 metabolites from the Metabolon

platform. Only associations were considered that had a p-value less than the p-value threshold after adjusting for multiple testing, so a p-value threshold of  $\frac{0.05}{345372 \cdot 117 \cdot 16 + 345372 \cdot 6406 \cdot 376} = 6.01 \cdot 10^{-14}$  according to Bonferroni correction was used.

Eventually, the number of associations was further reduced by taking the correlation among SNPs into account. SNPs were clustered together if they have a correlation greater than 60 %. For each cluster the top hit, i.e., the SNP-metabolite association with the lowest p-value, is included into the interaction analysis.

Finally, *pulver* is applied to the metabolite ratios which are significant associated to T2D as described in section 5.1. To reduce the number of tests a two-stage analysis as introduced by Kooperberg and LeBlanc is applied [186]: only SNPs, having a minor allele frequency greater than 0.05, and CpG sites are included that are weakly associated to the corresponding metabolite ratio ( $p < 0.05$ ). This computation is conducted using the R package *MatrixEQTL* [19]. Finally, 22,112 CpG sites and 314,643 SNPs were included into the analysis. Associations with a p-value less than the Bonferroni significance threshold  $\frac{0.05}{22,112 \cdot 314,643} \approx 7.19 \cdot 10^{-12}$  are significant.

## 5. Results and discussion

### 5.1. T2D and metabolite ratios analysis

#### 5.1.1. Results

For the T2D and metabolite ratio analysis, the associations between metabolites and prevalent and incident T2D were computed using logistic regression and proportional hazard Cox regression models, respectively. This analysis is part of the IMI (Innovative Medicines Initiative) DIRECT (Diabetes Research on Patient Stratification) study [187, 188]. One of the aims of DIRECT is to use deep phenotyping to identify biomarkers for risk stratification and response to diabetes therapy.

The analysis served as a validation of two prior studies from the Netherlands (LLS, Postpartum Outcomes in mothers with Gestational diabetes and their Offspring study (POGO) [189]) [103]. Both studies investigated the association of metabolite ratios and insulin secretion in a hyperglycaemic clamp study sample and an OGTT. In both analyses, the metabolite ratio valine to PC ae C32:2 was more strongly associated with insulin than the single metabolites, as indicated by a significant p-gain.

##### 5.1.1.1. Pairwise metabolite ratios and prevalent T2D

In three different independent epidemiological studies, KORA F4, LLS, and NTR, the associations between pairwise metabolite ratios and prevalent T2D were analyzed (306 cases and 4,619 controls in total). Results were jointly analyzed in a fixed-effects meta-analysis using models adjusted for age, sex, BMI, and lipid lowering medications. Because this analysis is a validation of the clamp study that tested glucose-stimulated insulin response and metabolite ratios, only the p-gains from those metabolite ratios were considered which were also significantly associated with insulin secretion in the clamp study. Nine of the ten ratios were significantly associated with prevalent T2D (see Table 5.1,  $p < \frac{0.05}{135+10} \approx 3.4 \cdot 10^{-4}$ ). However, only the ratio of valine to PC ae C32:2 had a stronger association with prevalent T2D ( $OR_{Val,PC\ ae\ C32:2} = 2.64$  (beta(SE) = 0.97 (0.09));  $p = 1.0 \cdot 10^{-27}$ ) than each of the two metabolites alone ( $p_{valine} = 2.22 \cdot 10^{-16}$ ;  $p_{PC\ ae\ C32:2} = 1.31 \cdot 10^{-13}$ ; p-gain=2.2 · 10<sup>11</sup>; see

Table 5.2 for all significant single metabolite associations). The association of the ratio valine to PC ae

C32:2 being stronger was indicated by the p-gain being above the significance threshold ( $p\text{-gain} \geq \frac{135}{2 \cdot 0.05} = 1,350$ , see methods).

**Table 5.1:** Results of logistic regression of metabolite ratios and prevalent T2D in LLS, NTR, and KORA F4.

Metabolite ratio	LLS	NTR	KORA S4	Meta-analysis	
	Beta (SE) p-value	Beta (SE) p-value	Beta (SE) p-value	Beta (SE) p-value	P-gain
Val_PC aa C34:4	0.387 (0.198) $5.11 \cdot 10^{-2}$	0.399 (0.160) $1.29 \cdot 10^{-2}$	0.381 (0.094) $4.62 \cdot 10^{-5}$	0.386 (0.075) $2.69 \cdot 10^{-7}$	0
xLeu_PC aa C34:3	0.499 (0.220) $2.28 \cdot 10^{-2}$	0.632 (0.180) $4.56 \cdot 10^{-4}$	0.677 (0.100) $1.03 \cdot 10^{-11}$	0.644 (0.081) $2.44 \cdot 10^{-15}$	0
Val_PC aa C34:3	0.654 (0.238) $6.04 \cdot 10^{-3}$	0.565 (0.177) $1.44 \cdot 10^{-3}$	0.657 (0.107) $7.77 \cdot 10^{-10}$	0.635 (0.085) $1.07 \cdot 10^{-13}$	0
Ser_PC ae C32:2	0.537 (0.237) $2.34 \cdot 10^{-2}$	0.227 (0.171) 0.18	0.505 (0.088) $1.11 \cdot 10^{-8}$	0.456 (0.074) $8.65 \cdot 10^{-10}$	0
Val_PC ae C32:2	1.022 (0.283) $2.99 \cdot 10^{-4}$	0.609 (0.180) $7.10 \cdot 10^{-4}$	1.100 (0.110) $2.33 \cdot 10^{-23}$	0.972 (0.089) $1.01 \cdot 10^{-27}$	2.2 $\cdot 10^{11}$
Val_PC ae C36:0	0.922 (0.255) $2.96 \cdot 10^{-4}$	0.270 (0.166) 0.10	0.593 (0.101) $4.95 \cdot 10^{-9}$	0.548 (0.082) $1.93 \cdot 10^{-11}$	0
Gln_PC ae C32:2	0.747 (0.265) $4.82 \cdot 10^{-3}$	0.221 (0.144) 0.12	0.467 (0.093) $5.46 \cdot 10^{-7}$	0.423 (0.075) $1.68 \cdot 10^{-8}$	0
PC aa C32:3_PC ae C34:3	0.345 (0.199) $8.33 \cdot 10^{-2}$	0.018 (0.201) 0.93	0.313 (0.081) $1.04 \cdot 10^{-4}$	0.281 (0.070) $6.42 \cdot 10^{-5}$	0
Val_lysoPC a C18:1	0.528 (0.243) $3.00 \cdot 10^{-2}$	0.311 (0.174) $7.40 \cdot 10^{-2}$	0.526 (0.092) $9.17 \cdot 10^{-9}$	0.484 (0.077) $3.50 \cdot 10^{-10}$	0
PC ae C36:5_PC ae C38:4	-0.212 (0.205) $2.99 \cdot 10^{-4}$	-0.307 (0.157) $5.11 \cdot 10^{-2}$	-0.193 (0.080) $1.70 \cdot 10^{-2}$	-0.216(0.067) $1.33 \cdot 10^{-3}$	0

Model: T2D = standardized metabolite ratio + age + sex + BMI + lipid lowering medication + study specific covariates. Beta: estimated effect of the ratio, SE: standard error, p-value: p-value of the effect estimate of ratio in this model. The p-gain was calculated by dividing the lowest p-value of the single metabolites by the p-value of the ratio [180]. Fixed effect meta-analysis was applied to calculate the common effect size and p-value across the three studies.



**Table 5.2:** Results of logistic regression of standardized metabolite levels and prevalent T2D in LLS, NTR, and KORA F4.

Metabolite	LLS	NTR	KORA F4	Meta-analysis
	Beta (SE) p-value	Beta (SE) p-value	Beta (SE) p-value	Beta (SE) p-value
H1	0.766 (0.166) $4.03 \cdot 10^{-6}$	1.915 (0.240) $4.62 \cdot 10^{-16}$	1.258 (0.083) $2.08 \cdot 10^{-52}$	1.226 (0.071) $6.06 \cdot 10^{-67}$
xLeu*	0.120 (0.156) 0.44	0.426 (0.135) $1.56 \cdot 10^{-3}$	0.744 (0.086) $6.26 \cdot 10^{-18}$	0.558 (0.066) $2.29 \cdot 10^{-17}$
Gln	-0.166 (0.184) 0.37	0.067 (0.147) 0.65	-0.244 (0.085) $3.87 \cdot 10^{-3}$	-0.166 (0.068) 0.015
Gly	-0.286 (0.256) 0.26	-0.536 (0.245) 0.028	-0.340 (0.101) $7.16 \cdot 10^{-4}$	-0.359 (0.088) $4.31 \cdot 10^{-5}$
Ser	-0.137 (0.198) 0.49	0.057 (0.150) 0.70	-0.044 (0.084) 0.61	-0.034 (0.069) 0.62
Val	0.309 (0.166) 0.063	0.410 (0.133) $1.96 \cdot 10^{-3}$	0.753 (0.096) $5.06 \cdot 10^{-15}$	0.577 (0.070) $2.22 \cdot 10^{-16}$
PC aa C32:3	-0.649 (0.299) 0.030	-0.423 (0.206) 0.040	-0.312 (0.093) $8.15 \cdot 10^{-4}$	-0.354 (0.082) $1.38 \cdot 10^{-5}$
PC aa C34:3	-0.441 (0.311) 0.16	-0.222 (0.211) 0.29	-0.180 (0.098) 0.065	-0.207 (0.085) 0.016
PC aa C34:4	-0.176 (0.248) 0.48	-0.153 (0.172) 0.37	-0.042 (0.087) 0.63	-0.089 (0.072) 0.22
PC ae C32:2	-0.884 (0.330) $7.34 \cdot 10^{-3}$	-0.329 (0.211) 0.12	-0.717 (0.103) $3.76 \cdot 10^{-12}$	-0.660 (0.089) $1.31 \cdot 10^{-13}$
PC ae C34:3	-0.923 (0.363) 0.011	-0.222 (0.222) 0.32	-0.550 (0.102) $7.66 \cdot 10^{-8}$	-0.519 (0.090) $7.43 \cdot 10^{-9}$
PC ae C36:0	-0.698 (0.315) 0.027	0.000 (0.180) 0.998	-0.150 (0.083) 0.072	-0.155 (0.073) 0.035
PC ae C36:5	-0.496 (0.295) 0.092	-0.434 (0.178) 0.015	-0.390 (0.090) $1.38 \cdot 10^{-5}$	-0.406 (0.077) $1.60 \cdot 10^{-7}$
PC ae C38:4	-0.359 (0.223) 0.11	-0.267 (0.147) 0.070	-0.279 (0.091) $2.11 \cdot 10^{-3}$	-0.285 (0.073) $9.87 \cdot 10^{-5}$
LysoPC a C18:1	-0.277 (0.297) 0.35	-0.057 (0.187) 0.76	-0.143 (0.092) 0.12	-0.137 (0.080) 0.085

Model: T2D = standardized metabolite concentration + age + sex + BMI + lipid lowering medication + study specific covariates. Beta: estimated effect of the ratio, SE: standard error, p-value: p-value of the effect estimate of ratio in this model. \* The AbsoluteIDQ<sup>im</sup> p150 kit does not distinguish between leucine and isoleucine, xLeu represents their combined levels.

### 5.1.1.2. Pairwise metabolite ratios and incident T2D

The Cox proportional hazards regression was first conducted in two independent studies, KORA S4 to F4 and EPIC Potsdam, and the results were subsequently combined in a meta-

analysis. Altogether, 910 incident T2D cases and 3,367 controls were included. The analyses were adjusted for the covariates shown in Table 4.1 in section 4.3.1. Results of the single analyses as well as the meta-analysis are shown in Table 5.3. Similarly, to the observations on prevalent T2D described above, a significant association was observed between incident T2D and the ratio valine to PC aa C32:2 (Table 5.3;  $HR_{Val\_PC\ aa\ C32:2} = 1.57$  (Beta(SE) = 0.45 (0.06));  $p = 1.3 \cdot 10^{-15}$ ).

**Table 5.3:** Results of Cox proportional hazards regression of metabolite ratios and prevalent T2D in EPIC-Potsdam and KORA S4 to F4.

Metabolite ratio	KORA-S4 to F4	EPIC-Potsdam	Meta-analysis	
	Beta (SE) p-value	Beta (SE) p-value	Beta (SE) p-value	P-gain
Ile_PC aa C34:3	0.309 (0.121) $1.07 \cdot 10^{-2}$	na		$3^a$
Ile_PC aa C34:4	0.175 (0.118) 0.14	na		$0^a$
Val_PC aa C34:4	0.085 (0.114) 0.46	0.147 (0.058) $1.05 \cdot 10^{-2}$	0.135 (0.051) $8.85 \cdot 10^{-3}$	0
Leu_PC aa C34:3	0.211 (0.116) $7.01 \cdot 10^{-2}$	na		$3^a$
Ile_PC aa C32:3	0.406 (0.130) $1.80 \cdot 10^{-3}$	na		$19^a$
Ile_PC aa C36:4	0.210 (0.114) $6.61 \cdot 10^{-2}$	na		$1^a$
Val_PC aa C34:3	0.202 (0.113) $7.36 \cdot 10^{-2}$	0.152 (0.054) $4.99 \cdot 10^{-3}$	0.161 (0.049) $9.32 \cdot 10^{-4}$	0
Ser_PC ae C32:2	-0.042 (0.108) 0.70	0.182 (0.055) $8.48 \cdot 10^{-4}$	0.137 (0.049) $5.01 \cdot 10^{-3}$	0
Val_PC ae C32:2	0.403 (0.132) $2.26 \cdot 10^{-3}$	0.463 (0.065) $9.41 \cdot 10^{-13}$	0.451 (0.058) $7.10 \cdot 10^{-15}$	$1.29 \cdot 10^6$
Val_PC ae C36:0	0.184 (0.117) 0.11	0.204 (0.057) $3.77 \cdot 10^{-4}$	0.151 (0.052) $3.40 \cdot 10^{-3}$	0
Gln_PC ae C32:2	0.050 (0.109) 0.65	0.090 (0.044) $3.95 \cdot 10^{-2}$	0.084 (0.041) $3.77 \cdot 10^{-2}$	0
Ile_PC ae C36:0	0.285 (0.122) $1.92 \cdot 10^{-2}$	na		$2^a$
PC aa C34:4_PC aa C38:1	0.080 (0.100) 0.43	na		$1^a$
Ala_Gly	0.541 (0.111) $1.11 \cdot 10^{-6}$	na		$378^a$

PC aa C32:3_PC ae C34:3	0.146 (0.105) 0.17	0.293 (0.054) $7.59 \cdot 10^{-8}$	0.262 (0.048) $5.73 \cdot 10^{-8}$	0
Ala_lysoPC a C18:1	0.395 (0.1183) $7.97 \cdot 10^{-4}$	na		11 <sup>a</sup>
Val_lysoPC a C18:1	0.271 (0.119) $2.27 \cdot 10^{-2}$	0.317 (0.055) $8.24 \cdot 10^{-9}$	0.309 (0.050) $5.52 \cdot 10^{-10}$	65
PC ae C36:5_PC ae C38:4	0.157 (0.102) 0.13	-0.076 (0.055) 0.17	-0.023 (0.048) 0.63	0

Model: T2D = standardized metabolite ratio + study specific covariates as given in Table 4.1 (in section 4.3.1.). The p-gain was calculated by dividing the lowest p-value of the single metabolites by the p-value of the ratio [180]. Fixed effect meta-analysis was applied to calculate the common effect size and p-value. na not available.

<sup>a</sup> Only calculated for the KORA data.

Beta: estimated effect of the ratio, SE: standard error, p-value: p-value of the estimated effect of ratio in this model. na not available.

As described above for prevalent T2D, the p-value of the association with the metabolite ratio was again significantly stronger than those of the associations between T2D and the two metabolites alone ( $p_{\text{valine}} = 3.57 \cdot 10^{-8}$ ;  $p_{\text{PC ae C32:2}} = 9.16 \cdot 10^{-9}$ ;  $p\text{-gain} = 1.3 \cdot 10^6$ ; see Table 5.4 for all significant single metabolite associations). In addition, for all significant associations two diagnostic plots as described in section 4.5.2.5-4.5.2.6 were plotted to visually examine whether the assumptions for a Cox proportional regression are met. No violation for the metabolite ratio valine to PCaeC32:2 as well as for most of the other metabolite ratios was observed (see Figure A.5 in Appendix).

In a next step, the association of incident T2D and the ratio between valine and PC ae C32:2 was further investigated with respect to additional covariates in KORA. Models were compared using time-dependent AUC and NRI, as shown in Table 5.5. Adjusting the model for glucose levels at baseline only marginally affected the results and the association remained highly significant ( $HR_{\text{Val}_{\text{PC ae C32:2}}} = 1.45$  (Beta(SE) = 0.37 (0.06));  $p = 1.4 \cdot 10^{-9}$ ). When the valine to PC ae C32:2 ratio was added to the traditional baseline prediction model comprising widely accepted traditional risk factors (TRF) as shown in Table 4.1 (in section 4.3.1.), the AUC estimated from the time-dependent ROC improved from 0.780 to 0.801 in the KORA S4 to F4 study ( $p = 3.2 \cdot 10^{-2}$  for the ratio, Table 5.5). The area under the curve was marginally larger in a model comprising the metabolite ratio compared to that assessed in a model with the two single metabolites instead of the ratio ( $AUC = 0.793$ ). The computed NRI compared the model with traditional risk factors and the model with additional covariates and supported these results. The

NRI showed a slightly improvement of the model with the metabolite ratio (NRI = 0.013) and a slightly worsening of the model with the single metabolites (NRI = -0.016).

Because both studies, KORA S4 to F4 as well as EPIC Potsdam adjusted for different covariates, accuracy of the predictive models was assessed using cross-validation. The cross-validation yielded comparable results to the estimated model using the total data, suggesting little overfitting in the present situation with large sample sizes and few added covariates (Table 5.5).

**Table 5.4:** Results of Cox proportional hazards regression of standardized metabolite levels and prevalent T2D in EPIC-Potsdam and KORA S4 to F4.

Metabolite	KORA S4 to F4	EPIC-Potsdam	Meta-analysis
	Beta (SE) P-value	Beta (SE) P-value	Beta (SE) P-value
H1	0.896 (0.084) $3.00 \cdot 10^{-26}$	0.674 (0.056) $3.46 \cdot 10^{-33}$	0.741 (0.046) $2.04 \cdot 10^{-57}$
Ala	0.247 (0.094) $9.03 \cdot 10^{-3}$	na	
Ile	0.220 (0.104) $3.39 \cdot 10^{-2}$	na	
Leu	0.133 (0.104) 0.20	na	
Gln	-0.229 (0.106) $3.11 \cdot 10^{-2}$	-0.177 (0.051) $4.72 \cdot 10^{-4}$	-0.187 (0.046) $4.38 \cdot 10^{-5}$
Gly	-0.449 (0.127) $4.20 \cdot 10^{-4}$	-0.301 (0.063) $1.84 \cdot 10^{-6}$	-0.330 (0.056) $5.02 \cdot 10^{-9}$
Ser	-0.298 (0.109) $6.27 \cdot 10^{-3}$	-0.030 (0.057) 0.59	-0.087 (0.050) $8.25 \cdot 10^{-2}$
Val	0.132 (0.105) 0.21	0.298 (0.054) $3.29 \cdot 10^{-8}$	0.263 (0.048) $3.57 \cdot 10^{-8}$
PC aa C32:3	-0.263 (0.124) $3.42 \cdot 10^{-2}$	-0.108 (0.056) 0.055	-0.135 (0.051) $8.79 \cdot 10^{-3}$
PC aa C34:3	-0.104 (0.109) 0.34	0.050 (0.053) 0.35	0.021 (0.048) 0.66
PC aa C34:4	0.026 (0.114) 0.82	0.060 (0.057) 0.29	0.053 (0.051) 0.30
PC aa C38:1	-0.098 (0.102) 0.33	na	
PC ae C32:2	-0.469 (0.141) $8.74 \cdot 10^{-4}$	-0.275 (0.057) $1.30 \cdot 10^{-6}$	-0.302 (0.053) $9.16 \cdot 10^{-9}$
PC ae C34:3	-0.531 (0.144) $2.20 \cdot 10^{-4}$	-0.452 (0.064) $1.12 \cdot 10^{-12}$	-0.465 (0.058) $1.22 \cdot 10^{-15}$
PC ae C36:0	-0.059 (0.114) 0.60	0.038 (0.049) 0.44	0.023 (0.045) 0.62
PC ae C36:5	-0.025 (0.112) 0.82	-0.156 (0.051) $2.39 \cdot 10^{-3}$	-0.125 (0.047) $7.63 \cdot 10^{-3}$
PC ae C38:4	-0.139 (0.114)	-0.117 (0.054)	-0.121 (0.049)

	0.22	$3.03 \cdot 10^{-2}$	$1.32 \cdot 10^{-2}$
LysoPC a C18:1	-0.205 (0.123)	-0.176 (0.057)	-0.181 (0.052)
	$9.61 \cdot 10^{-2}$	$2.09 \cdot 10^{-3}$	$4.81 \cdot 10^{-4}$

Model: T2D = standardized metabolite concentration + age + sex + BMI + lipid lowering medication + study specific covariates. The P-gain was calculated by dividing the lowest p-value of the single metabolites by the p-value of the ratio [180]. Fixed effect meta-analysis was applied to calculate the common effect size and p-value.

Beta: estimated effect of the ratio, SE: standard error, p-value: p-value of the estimated effect of ratio in this model. na not available.

**Table 5.5:** Apparent and cross-validated model performance for incident T2D in KORA S4 to F4.

model	KORA S4 to F4			
	AUC ROC	NRI	Beta (SE)*	P-value*
<b>Val_PC ae C32:2</b>	0.697	-0.206	0.651 (0.102)	$2.05 \cdot 10^{-10}$
<i>Cross validation result</i>	0.693	-0.193		
<b>Glucose + Val_PC ae C32:2</b>	0.782	0.036	0.562 (0.113)	$6.43e \cdot 10^{-7}$
<i>Cross validation result</i>	0.779	0.047		
<b>Glucose + TRF</b>	0.780	- (Reference)	-	-
<i>Cross validation result</i>	0.766	- (Reference)		
<b>Glucose + TRF + Val + PC ae C32:2</b>	0.793	-0.016	Val: 0.05 (0.12) PC ae C32:2: -0.48 (0.15)	Val: 0.66 PC ae C32:2: $2.1 \cdot 10^{-3}$ Joint effect: $8.8 \cdot 10^{-3}$
<i>Cross validation result</i>	0.774	-0.004		

<b>Glucose + TRF + Val_PC ae C32:2</b>	0.801	0.013	0.311 (0.145)	$3.19 \cdot 10^{-2}$
<b>Cross validation result</b>	0.781	0.019		

TRF = traditional risk factors as shown in Table 4.1. in section 4.3.1

\* Betas (SE) and p-values are provided for the ratio of valine and PC ae C32:2. NRI was calculated comparing the model in the given row and the model Glucose + TRF.

AUC ROC: Area under curve of the Receiver Operating Characteristic Curve, Beta: estimated effect of the ratio, SE: standard error, p-value: p-value of the estimated effect of ratio in this model.

### 5.1.2. Discussion

In this analysis the association of the metabolite ratios with prevalent and incident T2D was investigated in four independent cohorts from the Netherlands and Germany.

The analysis served as a validation of two prior studies investigating the association of metabolite ratios and insulin secretion [103]. In this thesis, the follow-up investigation, the association between the metabolite ratios and prevalent and incident T2D, was conducted.

It has been shown that metabolite ratios can reveal perturbations in pathways relevant for a certain phenotype and may thus reveal stronger and more meaningful associations than associations with single metabolite levels [2, 190]. However, even if not directly involved in the associated pathways, metabolite ratios can serve as good biomarkers with predictive ability beyond that of the single constituents. It has been shown that this approach can reduce the variance of the single metabolite levels and increases the statistical power [180]. Thus, the focus of the current study was to determine whether metabolite ratios improve the prediction of T2D when combined with known traditional risk factors. Note that this does not imply an optimal set of predictors for T2D. In the framework of this study, the p-gains of associations between metabolite ratios that were significantly associated with glucose-stimulated insulin in the clamp study and prevalent or incident T2D were computed. In both analyses only the metabolite ratio valine to PC ae C32:2 displayed a significant p-gain, an indicator that the metabolite ratio is more strongly associated than the metabolites individually.

In T2D, the BCAA valine has been shown to be increased and responsive to glucose stimulation in several studies [46, 191, 192] but also to the glucose lowering drugs glipizide

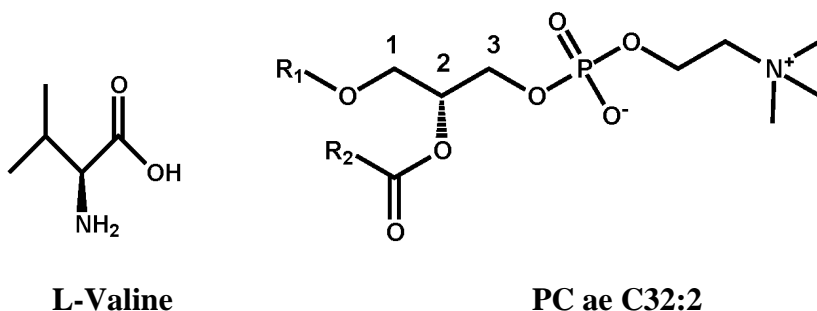
and metformin [48, 193]. In general, it is known that BCAAs associate with insulin sensitivity [194, 195] and development of diabetes [50].

The observed increase of BCAA levels may be the result of impaired metabolism, i.e., impairments of the branched-chain keto acid dehydrogenase complex, causing the accumulation of potentially toxic intermediates (branched-chain  $\alpha$ -ketoacids) that may lead to  $\beta$ -cell mitochondrial dysfunction [196, 197]. A second hypothesis for the observed association between elevated BCAA levels and insulin resistance might be the permanent activation of the serine kinases S6K1 and the mammalian target of rapamycin complex 1 through BCAAs and together with insulin. This leads to serine phosphorylation of insulin receptor substrates-1 and insulin receptor substrates-2, which impede insulin signalling, such as inducing gluconeogenesis in the liver and translocation of glucose transporter type 4 in the skeletal muscle, resulting in increased insulin demand [40].

However, other studies indicate that BCAA levels are probably not sufficient to trigger disease and rather being the consequence of impaired insulin action, as insulin resistance is associated to reduced expression of mitochondrial BCAA catabolic enzymes [196, 198]. Thus, BCAAs might serve only as a marker of increased insulin resistance [196, 198].

PC species are defined as a class of phospholipids with a choline as a head group. Like valine, PCs have been found to be associated with T2D [27]. As PCs cannot be detected by all metabolomics platforms, replication is less frequent compared to the BCAAs [27, 46, 50, 199]. However, PC ae C32:2 has been shown to be associated with prevalent [199] and incident T2D [46] and was also found to respond to glucose stimulation during OGTT [200]. Furthermore, the Biocrates kit which was used to detect the metabolites in the present study does not allow for a detailed analysis of the exact lipid composition of metabolites such as PC ae C32:2 (see Figure 5.1. for a skeletal formula). Thus, interpretation of results and literature searches for mechanistic validation are limited. According to the Human Metabolome data (HMDB) PC ae C32:2 as quantified by the Biocrates Kit is composed of either the fatty acids C16:1/C16:1, C18:1/C14:1 or C18:2/C14:0 ([www.HMDB.org](http://www.HMDB.org)) [201].

In general, phosphatidylcholines are constituents of cellular membranes and are suspected of playing an important role in cellular signal transduction [46, 202]. In addition, plasmalogens which are a subclass of acyl-alkyl-phosphatidylcholines, have been found to prevent lipoprotein oxidation [46, 190, 202, 203].



**Figure 5.1:** Structure of PC ae C32:2 and L-valine. Chemical formulas of PC ae 32:2 and L-valine are shown. Please note that for PC ae 32:2 the carbon number is calculated as  $R1 + R2 + 1 = 32$ , and the double bonds number is 2 for  $R1 + R2$ . It is not known which specific chain lengths are quantified by Biocrates.

In previous studies by Floegel et al. and Wang-Sattler et al., phosphatidylcholines and sphingomyelins were found to be associated with increased risk of impaired glucose tolerance (IGT) and incident T2D [27, 46]. However, acyl-alkyl-phosphatidylcholines, such as PC ae C32:2, were associated with decreased risk of incident T2D [46]. Furthermore, it was observed that acyl-alkyl-phosphatidylcholine concentrations are significantly lower in the obese state in children and in adults compared to normal weight participants [190, 204]. Thus, the observed negative association between acyl-alkyl phosphatidylcholine levels and T2D or obesity may reflect an increased consumption of plasmalogens during oxidative stress [190].

In murine adipose tissue, several studies have shown that the BCAAs and lipogenesis are related [205-207]. The catabolism of the BCAAs contributes to the synthesis of odd-chain and even-chain fatty acids, like C14, C16, and C18 chains which are components of PC ae C32:2 [206]. Jang et al [208] showed that 3-hydroxyisobutyrate (3-HIB), a catabolic intermediate of valine, is a paracrine regulator of trans-endothelial fatty acid transport. In particular, 3-HIB stimulates muscle fatty acid uptake and promotes lipid accumulation in muscle, eventually leading to insulin resistance in mice [208]. The insulin resistance is potentially driven by the increased fatty acid flux from the blood resulting to an increase in myocellular diacylglycerol in the muscle which leads to a decreased insulin-stimulated glucose-transport activity after several activation of enzymes [209]. Thus, the regulation of the fatty acid flux might be another link to the relation of fatty acids and BCAA catabolism, providing a new mechanistic explanation for how increased BCAA catabolic flux can cause diabetes [208].

At present it is not clear whether the metabolite ratio is causally related to diabetes risk. However, the time sequence implied in the Cox proportional hazards model is a first indicator for



a putative causal relationship. Furthermore, a recent Mendelian randomization study suggested a causal relationship between increased BCAA levels, such as valine, and T2D risk [210]. This does, however, not imply that the ratio is causal as well.

Further research is necessary to investigate the possible functional relationship between valine and PC ae C32:2 and whether there is a direct causal relation underlying the observed associations with glucose-stimulated insulin secretion and risk of developing diabetes.

A limitation of this study is that KORA S4 to F4 and EPIC-Potsdam used different covariates in the Cox proportional hazards regression because not all covariates were available in both cohorts. However, both sets of covariates comprised well established risk factors which had been previously used in similar metabolomic studies [27, 46] and results of those studies were validated in independent replication panels [150-152]. Furthermore, consistent results of both studies show that, despite the differences, the associations are robust and reliable.

Moreover, because in KORA the same data was used to generate the model and to compute the AUC, it is expected that the performance is quite high [162]. Therefore, the accuracies of the predictive models in KORA were assessed using a cross-validation approach. Both, the apparent and cross-validated model yielded comparable results concerning the time-dependent AUC and NRI, and thus suggesting little overfitting.

In this study, the valine to PC ae C32:2 metabolite ratio improved the prediction of incident T2D extending previous evidence of an association between the two constituent metabolites and T2D [27, 46]. Furthermore, it is important to note that in all analyses conducted in the present study the estimated accuracy of the prediction model containing the ratio was slightly larger than that of a model comprising the two constituent metabolite levels alone, suggesting that the use of ratios improves risk prediction. Large prospective studies aiming to identify the best set of predictors are needed to evaluate the clinical relevance of metabolite ratios in individual T2D risk prediction. The simple and relatively low invasive nature of metabolomics measurements and the fact that alterations in metabolite profiles can be detected years before disease manifestation, indicate that metabolomics might prove to be a useful instrument in personalizing prevention and treatment strategies for T2D.

In conclusion, a novel association between the ratio of two metabolites, valine to PC ae C32:2, was identified and validated with both prevalent and incident T2D. This ratio significantly improved the prediction of future T2D manifestation beyond established traditional risk factors.

These findings open opportunities for future functional studies investigating the causality of the association as well as its clinical relevance as an early biomarker for T2D.

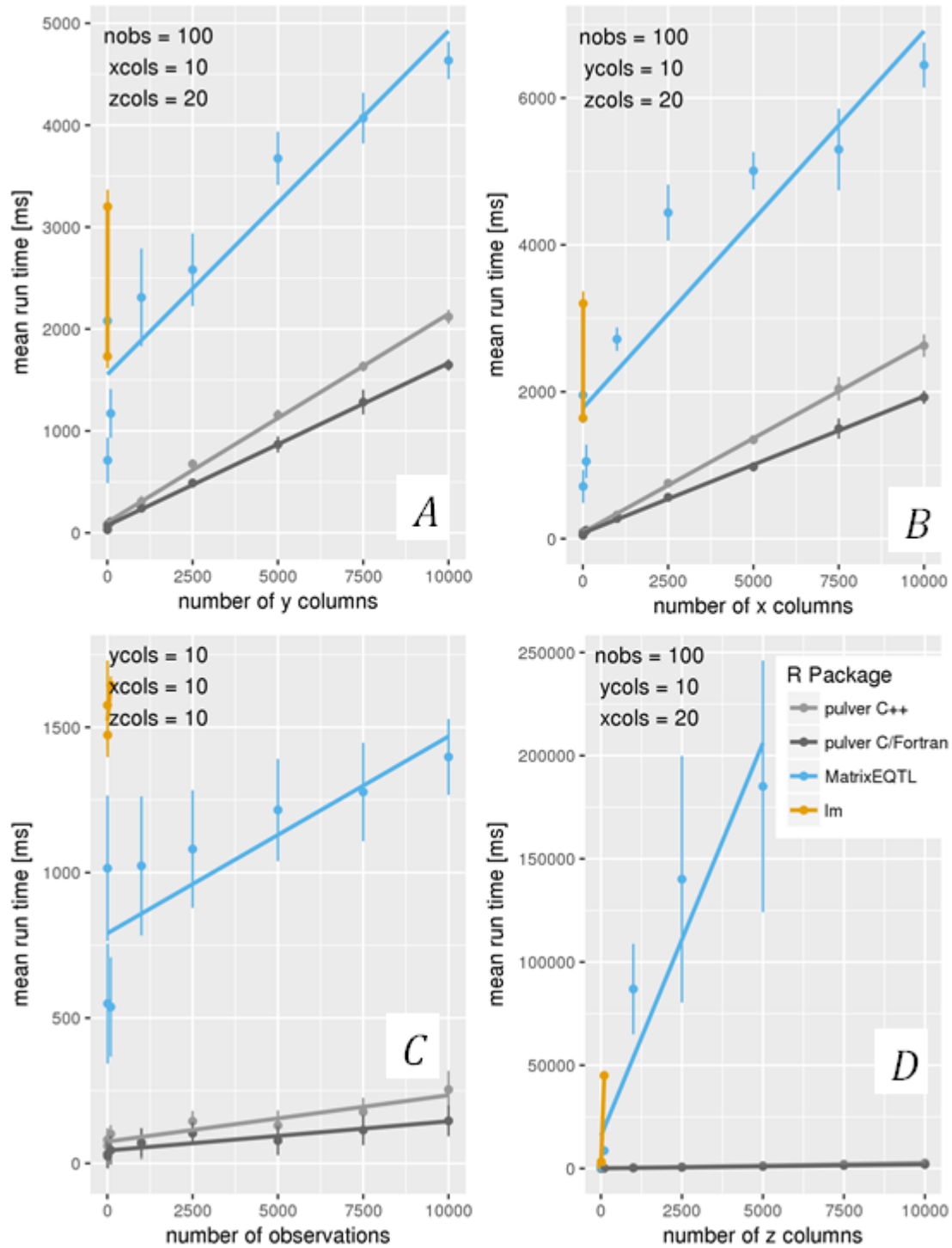
## 5.2. Interaction analysis

### 5.2.1. Results

The overall aim of this study was to implement and evaluate the R package *pulver*, which computes billions of linear regressions with an interaction terms in a reasonable amount of time. In the first part of this study, the performance of this R package was benchmarked against two other programs using simulated  $X$ ,  $Y$ , and  $Z$  matrices with different numbers of observations and variables. In the second part of this study, *pulver* was applied to real data from the KORA study, i.e., the interaction of SNPs and CpG sites on metabolite levels measured by the Biocrates platform and the Metabolon platform. Finally, the metabolite ratio valine to PC ae C32:2, identified in the analysis with incident and prevalent T2D, was investigated for potential associations with SNPs and CpG sites.

#### 5.2.1.1. Performance comparison using simulated data

For the comparison of the R packages *pulver* using the C++ and Fortran version and *MatrixEQTL* and R's built-in *lm* function four different scenarios were conducted. In the scenarios the number of columns in the  $X$ ,  $Y$ , and  $Z$  matrices, and the number of subjects were varied and the mean run times computed. Figure 5.2 shows the mean run times for all different scenarios (A-D). For all benchmark sets, *pulver* performed better than the alternatives. In all benchmark scenarios the results obtained for the *lm* function were so slow that only the first mean runtime was included into the corresponding chart for comparability. Most striking, in the scenario involving varying numbers of variables in matrix  $Z$  (see Figure 5.2 D), *pulver* outperformed the other methods by several orders of magnitude. Even the run times of *MatrixEQTL* are so slow that they are only partly included in the chart. The poor performance of *MatrixEQTL* is obtained because only one  $Z$  variable can be included into this function, forcing the user to repeatedly call *MatrixEQTL* for every variable in the  $Z$  matrix. This type of iteration is known to be slow in R. Thus, benchmark D reflects the intended user case for *pulver* where all input matrices contain many variables and *pulver* can be utilized optimally.



**Figure 5.2:** Mean run times and standard deviations for the interaction analysis using R's *lm* function (orange), *MatrixEQTL* (blue), and *pulver* (black/dark grey). The execution times are in milliseconds. A line through the time points for each package was fitted. R's *lm* function was

very inefficient for this type of interaction analysis, and only the first two points are shown for every benchmark.

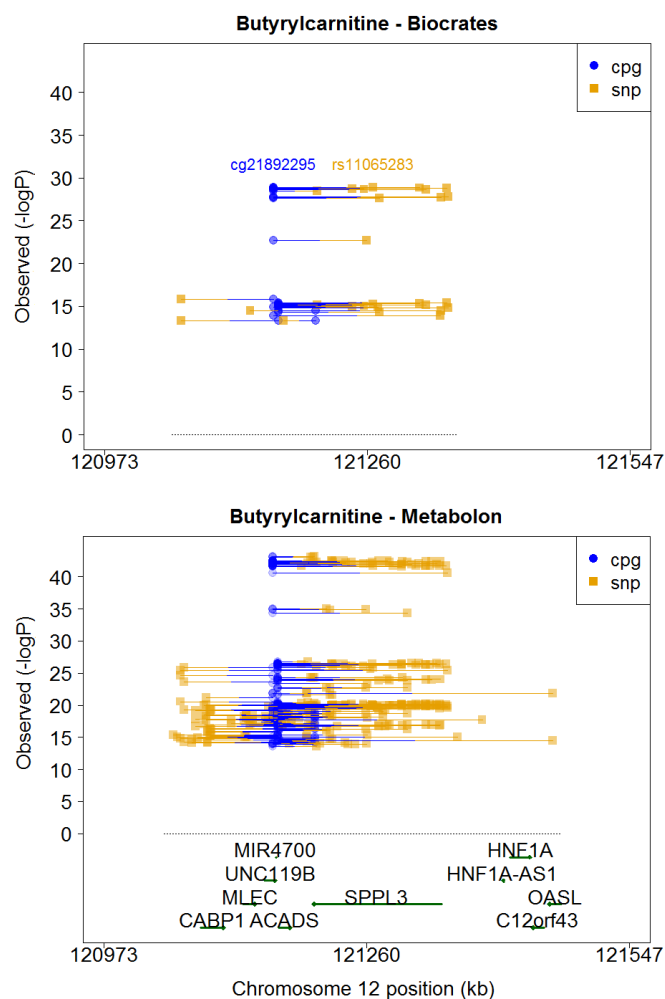
nobs: number of observations, xcols: number of columns in the  $X$  matrix, ycols: number of columns in the  $Y$  matrix, zcols: number of columns in the  $Z$  matrix.

### 5.2.1.2. *SNP-CpG interaction and metabolite levels*

In the real world scenario, the association between the interaction of SNPs and CpG sites with metabolite levels measured by the Biocrates platform and the Metabolon platform were analyzed. The final data sets comprised 345,372 CpG sites, 117 SNPs, and 16 metabolites from the Biocrates kit and 345,372 CpG sites, 6,406 SNPs, and 376 metabolites from the Metabolon platform.

Altogether, 27 significant associations for metabolites from the Biocrates platform (p-values ranging from  $1.28 \cdot 10^{-29}$  to  $5.17 \cdot 10^{-14}$ ) and 286 significant associations for metabolites from the Metabolon platform (p-values ranging from  $1.15 \cdot 10^{-42}$  to  $3.73 \cdot 10^{-14}$ ) were identified. All of the significant associations involved the metabolite butyrylcarnitine and SNPs and CpG sites located on chromosome 12 in close proximity to the *ACADS* gene (+ strand, see Figure 5.3, and Table A.3/A.4 in Appendix). More precisely, in total five CpG sites, cg06793505, cg21721566, cg21892295, cg23907586, and cg02419362, were associated with butyrylcarnitine. Comparing correlations between these five CpG sites revealed that CpG sites cg21721566 and cg21892295 as well as cg23907586 and cg21892295 were highly correlated with each other (Pearson correlation:  $r^2 \approx 0.7$ ,  $r^2 \approx 0.6$ , respectively).

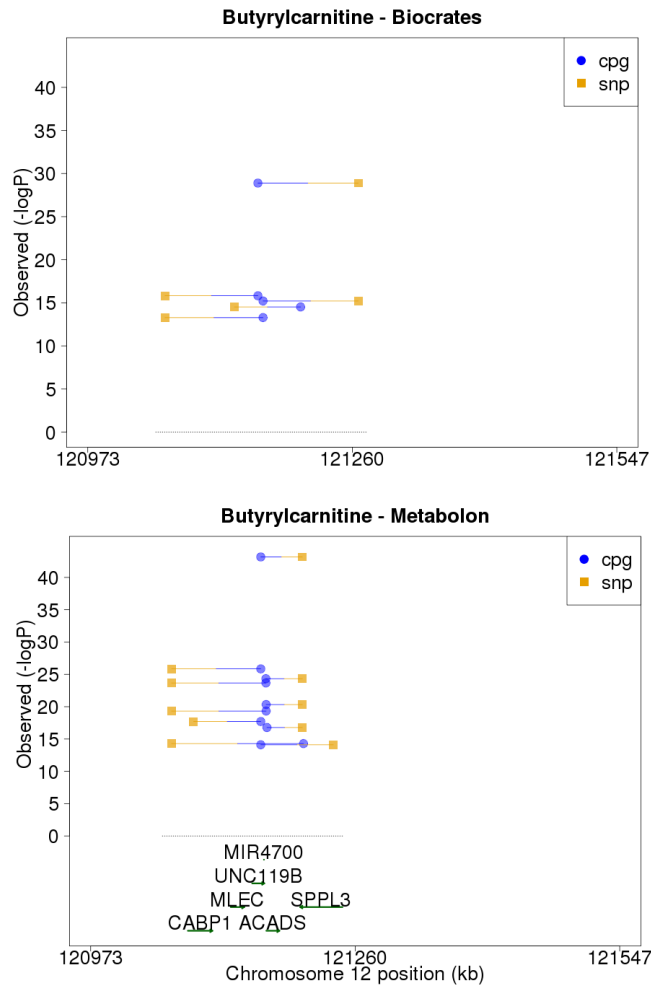
After reducing the number of SNPs as described in section 4.9.6, there were ten significant associations with butyrylcarnitine in the Metabolon platform data. These associations involved four SNPs and five CpG sites (see Figure 5.4 lower panel): rs10774563, rs9431, rs1039302, rs7965649 and cg02419362, cg06793505, cg21721566, cg21892295, cg23907586.



**Figure 5.3:** Regional plot with significant associations among SNPs (circles), CpGs (squares), and butyrylcarnitine for the Biocrates platform (top) and Metabolon platform (bottom). Significant interactions between SNPs and CpGs are visualized by lines connecting SNPs and CpGs.

For butyrylcarnitine measured using the Biocrates platform, five associations involving three SNPs (rs7964786, rs11065283, rs2001133) and three CpG sites (cg02419362, cg21721566, cg21892295) were significant (see Figure 5.4 upper panel). The scatterplots for all 16 significant associations are shown in Appendix Figure A.1 and A.2 in the left panel, whereas in the upper right the adjusted coefficients of determination are shown for different models: SNP  $\sim$  CpG, metabolite  $\sim$  CpG, metabolite  $\sim$  SNP, metabolite  $\sim$  SNP + CpG, and metabolite  $\sim$  SNP  $\cdot$  CpG + SNP + CpG. The different models illustrate how the inclusion of an interaction term in the model increased the adjusted coefficient of determination  $R^2$  (computed using the

*summary.lm* function in R). For models including the interaction term the coefficient of determination ranged from 0.2491 to 0.3777 in the data assessed using the Metabolon platform and from 0.1689 to 0.2491 in the data assessed using the Biocrates platform. In the lower right panel, the beta coefficients of the CpG site, SNP and their interaction, computed by the linear regression model with interaction term, are depicted. In all associations the beta coefficient of the SNP alone was small compared to the beta coefficient of the interaction term or CpG site and had thus less impact on the outcome. For associations involving the metabolite butyrylcarnitine measured by the Biocrates platform the largest effect was observed for the interaction term, whereas for associations with butyrylcarnitine measured by the Metabolon platform the CpG site had, in some cases, the largest beta coefficient.

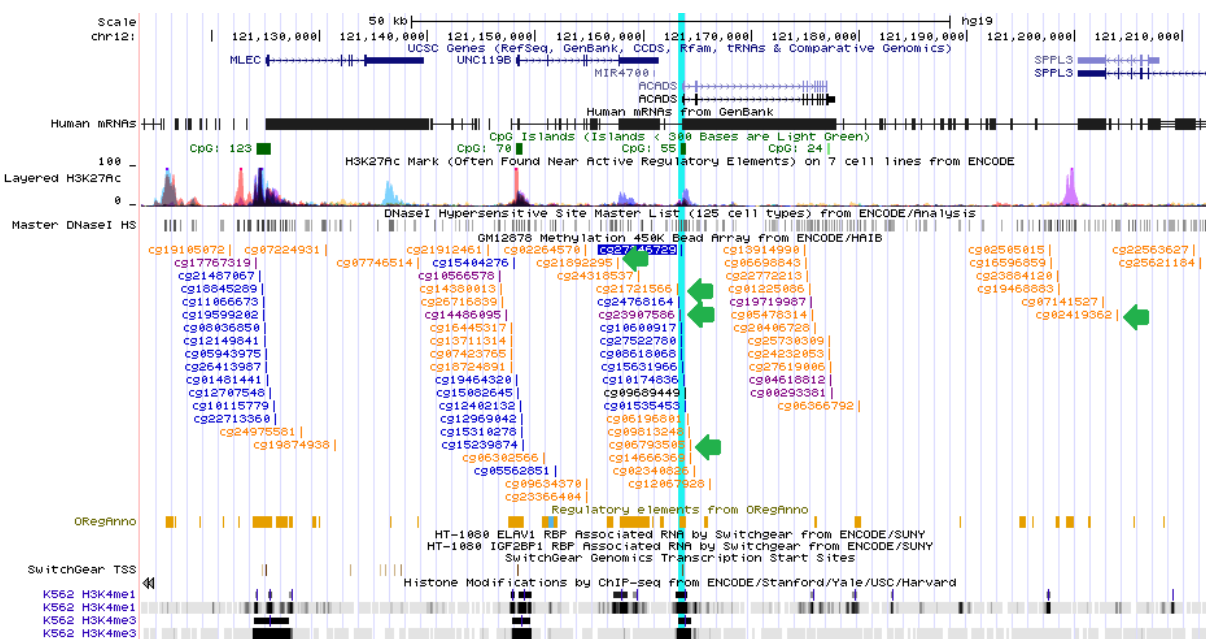


**Figure 5.4:** Regional plot with significant associations between SNPs with correlation coefficient below 60% (circles), CpGs (squares), and butyrylcarnitine for data of the Biocrates

platform (top) and of the Metabolon platform (bottom). Significant interactions between SNPs and CpGs are visualized by lines connecting SNPs and CpGs.

The positions of the five CpG sites are depicted in Figure 5.5. They indicate that the CpG sites are near to a CpG island but also to transcription factor binding sites. This CpG island is followed by the presence of the overlay of H3K4Me1 and H3K27Ac as well the overlay of H3K4Me3. H3K4Me1 denotes the methylation of histone H3 at lysine 4, H3K27Ac the acetylation of histone 3 at lysine 27 [81], and H3K4Me3 the trimethylation of histone H3 at lysine 4 [79]. H3K4me3 is primarily associated with active promoters and H3K27Ac is associated with both active promoters and enhancers [211].

Furthermore, in this region a hypersensitivity of DNase enzyme was observed. All information was obtained from the UCSC Genome Browser (<https://genome.ucsc.edu/>).



**Figure 5.5:** Regional plot of CpG sites on chromosome 12 in the *ACADS* region, created by the UCSC Genome Browser (<https://genome.ucsc.edu/>). Positions of exons and introns of genes, CpG islands, histone modifications, transcription factor binding sites, and CpG sites are shown. CpG sites for which the interaction with a SNP is significantly associated with butyrylcarnitine are marked with an arrow. The colors of the names of the CpG site represent different degrees of methylation: Orange indicates a high degree of methylation ( $\beta \cdot 1000 \geq 600$ ), purple indicates a modest degree of methylated ( $200 < \beta \cdot 1000 \leq 600$ ), and blue indicates a low degree to no methylation ( $0 < \beta \cdot 1000 \leq 200$ ), CpG sites where the methylated status is unknown are colored black.



In addition, four diagnostic plots associations as described in section 4.5.2.1-4.5.2.4 for all identified 15 significant associations were plotted to visually examine whether the assumptions for a normal distribution are met. No violation of the observations was observed (see Figure A.3 and A.4 in Appendix).

### 5.2.1.3. SNP-CpG interaction and metabolite ratio

In section 5.1 the metabolite ratio valine to PC ae C32:2 was identified to be significantly associated with incident and prevalent T2D. To elaborate potential epigenetic and genetic influence beyond this metabolite ratio, the R package *pulver* was applied for this ratio as well. No association could be identified that had a p-value less than the Bonferroni significance threshold of  $7.19 \cdot 10^{-12}$ , i.e., there were no significant associations. The first ten associations with the lowest p-value between this metabolite ratio and the interaction SNP  $\times$  CpG are depicted in Table 5.6. In contrast to the significant associations between butyrylcarnitine and the SNP-CpG interaction, some SNPs are located on different chromosomes than the CpG sites.

**Table 5.6:** The first ten associations with the lowest p-value between metabolite ratio valine to PC ae C32:2 measured with the Biocrates platform and the interaction SNP  $\times$  CpG.

CpG	SNP	P-value	Position (SNP)	Chr. (SNP)	Position (CpG)	Chr. (CpG)
cg13562917	rs61933106	$9.95 \cdot 10^{-10}$	115040096	12	90776873	15
cg13562917	rs61933105	$1 \cdot 10^{-9}$	115039967	12	90776873	15
cg27633287	rs142854002	$1.06 \cdot 10^{-9}$	130298443	12	130766243	12
cg14669379	rs199825610	$1.08 \cdot 10^{-9}$	31378517	3	112058559	1
cg14132884	rs10087554	$1.16 \cdot 10^{-9}$	70346293	8	75230010	14
cg23850377	rs33922594	$1.18 \cdot 10^{-9}$	19614509	8	20251576	2
cg03404572	rs62254320	$1.27 \cdot 10^{-9}$	53228655	3	61329252	18
cg03404572	rs12490555	$1.29 \cdot 10^{-9}$	53228836	3	61329252	18
cg03404572	rs12490645	$1.44 \cdot 10^{-9}$	53228919	3	61329252	18
cg03404572	rs60520044	$1.47 \cdot 10^{-9}$	53238942	3	61329252	18

Chr. = Chromosome.

## 5.2.2. Discussion

The interplay between metabolite levels, genetic variants, and CpG sites was examined from an omics-wide perspective. Previous interaction analyses in GWAS mainly considered interactions

between SNPs and thus ignored possible interactions between different “omics” layers, which are in general measured on a continuous scale. Thus, in this study the R package *pulver* was implemented.

#### 5.2.2.1. Performance comparison using simulated data

Compared to its competitors, the R package *MatrixEQTL* [19] and R’s built-in *lm*, it is the fastest implementation available for calculating p-values for the interaction term of two quantitative variables given a huge number of linear regression models. In particular, in the case where interaction terms need to be calculated for many pairs of variables, *pulver* performs far better (see Figure 5.2). The run time differences between the programming languages Fortran and C++ were negligible. Time savings are achieved by avoiding redundant calculations, as also implemented in *omicABEL* [18]. However, this latter software is not flexible enough to compute regression with an interaction term.

Similar to *MatrixEQTL*, computationally expensive p-values are only computed at the very end and only for results below the significance threshold determined using the (computationally cheap) Pearson’s correlation coefficient.

To maximize the speedup, it is always worthwhile to specify a p-value threshold and use *pulver* as a filter to find models with significant or near-significant interaction terms. If a p-value threshold is not specified, the time savings will be suboptimal and the number of results will be very high. Thus, we recommend using a p-value threshold to adjust for multiple testing, such as Bonferroni correction, i.e.  $\frac{0.05}{\text{number of tests}}$ .

The core algorithm of *pulver* was implemented in two languages namely C++ and C/Fortran to examine different performances due to programming languages. However, comparing the two different implementation of *pulver* reveals no striking differences. Thus, we continued to use the C++ version as it offered some useful implemented functions such as those implemented in the C++ Standard Library algorithms [212].

However, comparing the complexity between *pulver* and *MatrixEQTL* revealed that in theory for large matrices the two algorithms might have similar processing time. In practice, differences are seen, as we need to iterate through each variable of the *Z* matrix to call the function *MatrixEQTL*. In *pulver*, this is already implemented with the C++ programming language which is known to be faster than the script language R [213].

The package *pulver* imputes missing values based on their column means. If this is not desired, then one can use other more sophisticated methods, such as the *mice* package in R [109], in order to impute missing values before applying *pulver*. Furthermore, *pulver* can typically only be used as a screening tool because it simply returns p-values. Other information regarding the fitted models, such as slope coefficients, standard errors, or residuals, must be computed in a second step using traditional tools.

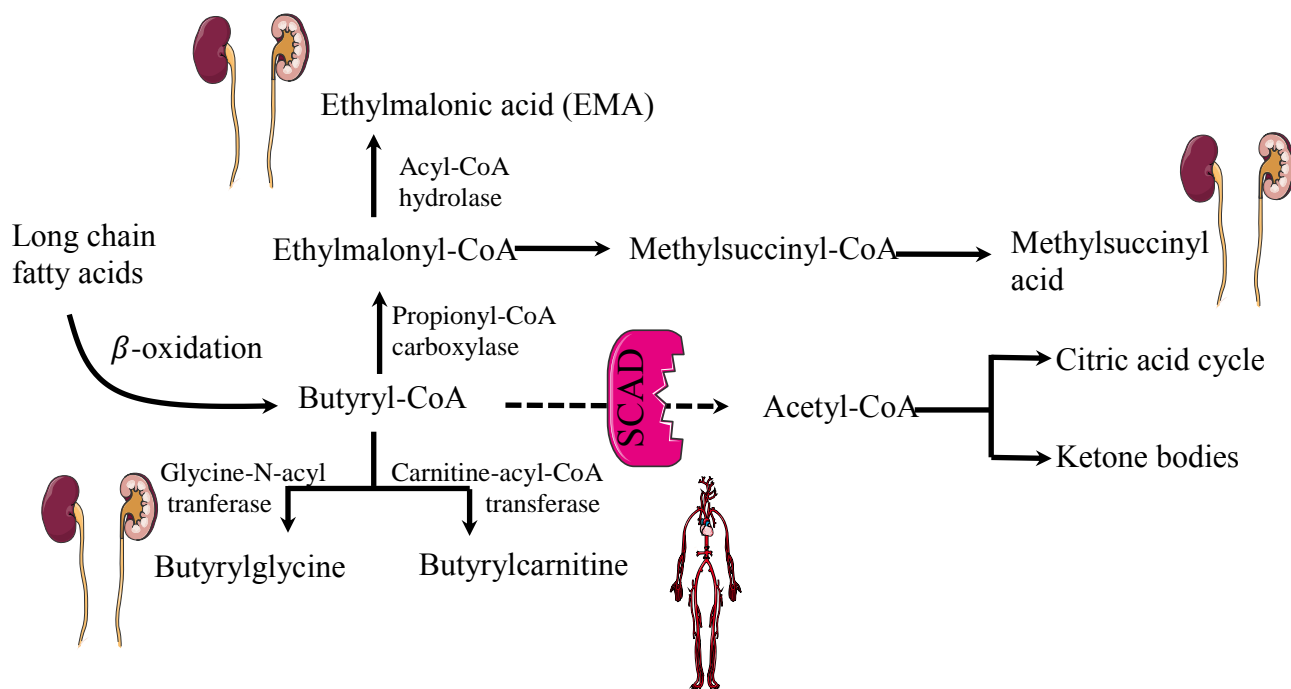
Moreover, further improvement might be achieved by integrating math libraries for computation of matrix operations into this package as used in MatrixEQTL. However, since through each variable of each matrix is iterated, only matrix-vector multiplication could be computed. According to Fabregat-Traver et al. for that kind of operations on small matrices using BLAS kernels an efficiency of approximately five percent could be achieved [18]. Thus, depending on sample sizes there might be only small benefits.

In conclusion, *pulver* can be part of a processing pipeline focused on interaction terms in linear regression models and its main value is allowing users to conduct comprehensive screenings that are beyond the capabilities of existing tools regarding different “omics” layers.

#### 5.2.2.2. *SNP-CpG interaction and metabolite levels*

Subsequently, the R package *pulver* was applied to data from the KORA data base to run just above 800 billion regressions. Only results with a p-value less than the Bonferroni-corrected threshold  $6.01 \cdot 10^{-14}$  were stored, a total of 313 significant associations. To further reduce the number of associations, only the SNP with the lowest p-value was selected from clusters of correlated SNPs. This effort reduced the number of significant associations to 15. All SNPs and CpG sites reside within the chromosome 12 near the *ACADS* gene (OMIM: 606885), with butyrylcarnitine being the related metabolite in both platforms, Metabolon and Biocrates. The *ACADS* gene encodes the enzyme short-chain acyl-CoA dehydrogenase (SCAD), which catalyzes the initial reaction in the mitochondrial  $\beta$ -oxidation of short-chain fatty acids, i.e., C4 and C6 fatty acids [214, 215]. The *ACADS* gene is located in the terminal region of the long arm of chromosome 12, and covers approximately 14.2 kb, containing ten exons [216, 217].  $\beta$ -oxidation represents an important source of energy for the human body, particularly during times of fasting and metabolic stress generating an alternative source of different kinds of acetyl-CoA such as propionyl-, butyryl-, and crotonyl-CoA [218, 219] (Figure 5.6). Over 60 inactivating mutations of the *ACADS* gene are known [220]. Inactivation results in an

accumulation of the encoded enzyme's substrate, which includes butyryl-CoA. This substrate can be converted into different intermediates, such as butyrylcarnitine, butyrylglycine, ethylmalonic acid (EMA), and methylsuccinic acid in blood, urine, and cells [215, 221] (Figure 5.6).



**Figure 5.6:** Biochemical pathways for alternative metabolism of butyryl-CoA adapted from [215]. Metabolite levels of butyrylglycine, ethylmalonic acid, and methylsuccinic acid may be elevated in urine and butyrylcarnitine in the blood for SCAD deficient patients. Illustration was prepared using a template on the Servier medical art website (<http://www.servier.com/Powerpoint-image-bank>).

Most of the gene variants in *ACADS* are missense variants impairing the correct folding and/or stability of the native protein structure [215]. Abnormally folded SCAD proteins and organic acids result in cellular toxicity and risk of acute metabolic acidosis and physiologic stress [215, 222].

Corydon et al. postulated that an interplay of genetic, cellular, and environmental factors leads to reduced catalytic activity of the enzyme SCAD [214]. Enzyme activity below a critical threshold may cause the onset of diverse clinical symptoms such as developmental delay, hyper- and hypotonia, ketotic hypoglycemia, and epilepsy due to a disease named Acyl-Coenzyme A Dehydrogenase Deficiency (SCADD) [221]. One of the accumulated metabolites in SCADD is butyric acid [215]. In high concentration butyric acid can inhibit the activity of

the histone deacetylase and thus prevent gene expression [223]. The association between the manifestation of SCADD and a mutation of the enzyme SCAD is still under discussion because many individuals having genetic variations in the *ACADS* gene do not develop SCADD [215]. Thus, SCADD is not part of the newborn screening programs in most countries [215].

Butyrylcarnitine is a metabolite involved in BCAA catabolism and fatty acid metabolism and has been shown to be associated with higher visceral fat mass [224], obesity, and insulin resistance [225]. In addition, butyrylcarnitine is a known marker of excessive fatty acid oxidation [226] and its increased concentration in the plasma might reflect impaired mitochondrial function, which is characterized by an accumulation of intermediates of fatty acid oxidation [52].

Previous findings had already suggested that butyrylcarnitine levels are strongly influenced by genetic variations [227, 228] and it had been shown that in the *ACADS* gene region SNPs and CpG sites (potentially also driven by SNPs in the neighborhood) are associated with butyrylcarnitine [5, 10, 185].

In this thesis, the significant association between the interacting SNP and CpG site and the levels of butyrylcarnitine in addition to an increased coefficient of determination suggests that the strength of association of the methylation with the metabolite level varies depending on the number of alleles of the SNP (Figures A.1 and A.2). However, the identified SNPs reside within a cluster of SNPs in strong linkage disequilibrium. Thus, further studies are warranted to determine the truly functional SNP within this gene cluster. In this study, five CpG sites accounted for a total of 15 significant associations. Four of these (cg06793505, cg21721566, cg21892295, cg23907586) were located near a CpG island, suggesting an influence on the transcription start site of the *ACADS* gene. This is supported by the presence of major chromatin features at this locus, such as H3K4Me1, H3K27Ac, and H3K4Me3, which are often found in regions containing active enhancers or promoters [81].

The interaction terms of the 15 associations showed both positive and negative associations with the metabolite butyrylcarnitine. As shown in Figures A.1 and A.2 in the Appendix the highest slope were mainly determined by the CpG site and the interaction  $CpG \cdot SNP$ . For associations measured with the Biocrates platform the interaction term had the highest impact on the association.

Since associations do not allow to deduce causality it is possible that levels of butyrylcarnitine influence the degree of DNA methylation or that the degree of DNA methylation has an influence on the metabolite levels. However, the locations of all but one of the identified CpG sites being near a promotor region of the *ACADS* gene may indicate that the direction of causality is more likely from methylation to metabolite. It is assumed that CpG sites of the promotor region of the *ACADS* gene have an influence on transcription binding sites. The *ACADS* gene codes for the enzyme SCAD, which in turn influences the concentration of butyrylcarnitine [214, 215].

One of the identified CpG sites, cg02419362, is located in the intronic region of the *SPPL3* gene. However, associations involving this CpG site should be considered carefully as an annotated SNP, rs35950819, encoding a deletion resides 49 bp upstream of this CpG site. Unfortunately, no further information about the allele frequency is available for this SNP and it was not measured in the KORA study.

The correlation coefficient between butyrylcarnitine levels across different platforms was 72%. Yet et al. compared the platforms by comparing correlation coefficients and results of GWAS on metabolite levels measured using the Biocrates platform with those on metabolite levels measured using the Metabolon platform [229]. The mean correlation coefficient of the 43 common metabolites was 0.44. Thus, the correlation of 0.72 is quite high. The differences between same metabolites measured from different platforms might result because of different types of measurements as the Biocrates kit uses a targeted approach whereas the Metabolon uses an untargeted. As mentioned in section 3.2.2, the Metabolon kit determines relative concentrations of as many metabolites as possible, using UHPLC/MS/MS2 injections and one GC/MS without absolute quantification. In contrast, the Biocrates kit uses a quantitative FIA-MS/MS method and in addition, internal standards serve as reference for the calculation of metabolite concentrations. Moreover, as described in section 3.2.1 and 5.1.2, in the Biocrates kit it is not possible to determine the precise position of the double bonds and the distribution of carbon atoms between the two fatty acid side chains. Thus, the Biocrates kit quantify the sum of different forms which might cause different correlations [229]. The lower correlation might also be due to different quality of measurement for different metabolites [229].

Furthermore, between the GWAS, Yet et al. found seven common loci which were associated with 16 metabolites including butyrylcarnitine, which were all located within the locus on chromosome 12 near the *ACADS* gene.

Similarly to SNP-SNP interactions, CpG-SNP interactions are difficult to identify because of the increasing complexity and computational burden [58, 230]. Only one region within the genome was identified where the interaction between SNPs and CpG sites plays a role in metabolite levels. Some significant associations of previous studies analyzing SNP-SNP interactions on a genome-wide scale [231, 232] could be reduced to only one SNP association through complex linkage disequilibrium patterns [233]. In contrast, this is not possible in CpG-SNP studies as DNA methylation is not lying in linkage disequilibrium and can change during time [79]. Furthermore, one cannot rule out other potential confounding for example SNPs not listed in currently available common databases or other regulatory elements/functionally relevant genomic features like microRNA within the probe-binding site of a CpG site. However, since the other identified CpG sites were located near a CpG island in the promotor region of the *ACADS* gene, it is rather unlikely that the CpG site is confounded by other regulatory elements/functionally relevant genomic features.

### 5.2.2.3. *SNP-CpG interaction and metabolite ratio*

In addition to the association analysis between metabolite levels and SNPs and CpG sites, an analysis regarding the metabolite ratio valine to PC ae C32:2 and SNPs and CpG sites was conducted. Unfortunately, no associations were identified that reached the Bonferroni threshold ( $p\text{-value} \leq 7.19 \cdot 10^{-12}$ ). In contrast to the significant associations of the metabolite butyrylcarnitine and SNPs and CpG sites are located near the *ACADS* gene that codes for an enzyme using butyryl-CoA as a substrate, it is rather unlikely that only one enzyme is responsible for the concentration level of this metabolite ratio. Valine is catalyzed in BCAA catabolism and PC ae C32:2 in fatty acid metabolism as well as in the metabolism of phosphatidylcholines [207]. As mentioned in section 5.1.2, recently, potential links between fatty acids and BCAA catabolism has been found, as 3-HIB, a catabolic intermediate of valine, is needed to enable the uptake of blood-borne lipid which initially traverse the blood vessel wall to the skeletal muscle stems [208]. Thus, multiple enzymes are responsible for the catabolism of valine and the synthesis of phosphatidylcholines and therefore, potentially the power is not enough to detect gene regions which are associated to the metabolite ratio.

**5.2.2.4. Conclusion**

Since the results of association analysis without an interaction term, i.e. only linear regression with metabolites and SNPs or metabolites and CpGs, indicate small influence of the SNPs and CpGs individually, it might be plausible that the power to identify significant associations of the interaction term within the same data set is even more limited. It might also be that interactions between SNPs and CpG sites play only a minor role.

Further large-scale studies with an increased sample size or studies following a candidate approach, i.e. analyzing only certain SNPs and CpG sites underlying a particular hypothesis, might be promising in the future.



## 6. Summary and future perspectives

In this thesis, the associations between metabolite ratios and incident and prevalent T2D, as well as the associations between metabolites and the interaction of SNPs and CpG sites, were computed.

The first analysis revealed an up-to-now unreported significant association between prevalent and incident T2D and the metabolite ratio valine to PC ae 32:2. These findings open opportunities for future functional studies investigating the causality of the association as well as its clinical relevance as an early biomarker for T2D. Furthermore, this result might serve as a starting point for finding prediction models incorporating combinations of metabolites and to test if such combinations show improved prediction compared to single metabolites or ratios.

To better understand the genetic and environmental mechanisms which influence the metabolite levels in the blood, an interaction analysis between metabolite levels, CpG sites, and SNPs was conducted. To achieve this, an R package *pulver* was implemented. It was shown that this R package can compute interaction analyses faster than the existing R package *MatrixEQTL* and R's built-in function *lm*.

Applying *pulver* to the CpG sites, SNPs, and metabolite levels, only one locus, near the *ACADS* gene, achieved significance for an effect of an interaction between SNPs and CpG sites on a metabolite, here butyrylcarnitine. This suggests that the power to detect interactions was not sufficient or that interactions between SNPs and CpG sites play only a minor role in the determination of metabolite levels in the blood.

In addition, the interaction between DNA methylation and SNPs and the metabolite ratio valine to PCae32:2 was investigated. However, no Bonferroni-corrected significant association was identified.

Because the manifestation of complex diseases such as T2D is influenced by several genetic and environmental factors it is difficult to find genetic risk factors. Future approaches for analyzing T2D and understanding its pathophysiology might include identification of causal genes and variants through denser catalogues of variations, improved imputation methods, and resequencing approaches. In total, even common variants associated with T2D explain only a fraction of the heritability of this disease [234]. In addition, the majority of identified diabetes variants are in noncoding intronic or intergenic regions of the genome, and thus these variants

may not be true causal variants, but only proxies in linkage disequilibrium with the causal variants [235]. It might be that the true causal variants have larger effect sizes and therefore may improve prediction of the cumulative genetic information.

Fuchsberger et al. assessed rarer variation that may not be well-tagged by GWAS arrays to test the hypothesis that lower-frequency variants explain much of the missing heritability [234]. However, they could not confirm that lower-frequency variants have a major role in predisposition to T2D [234]. This suggests that larger GWAS find more likely common than low-frequency SNPs related to T2D.

Due to cheaper costs and advancements in the whole genome sequencing (WGS), it might become possible to analyze the full genomic sequence of many subjects. This can further improve our knowledge in biological mechanisms in health and disease. However, increasing data calls for advanced softwares and computer systems to process it. First attempts were to divide the data sets into smaller parts and analyze them one after the other. However, it is now becoming common to use a supercomputer which can run analyses with thousands of cores and have gigabytes of memory space. An alternative is the recently developed *Hail* framework (see <https://github.com/hail-is/hail>) [17]. *Hail* can query the collection of all genotypes in the dataset, e.g., from the Genome Aggregation Database (<http://gnomad.broadinstitute.org/variant/11-31809070-C-T>) that has about 5 trillion unique genotypes that does not fit in one computer memory.

Yet, in GWAS, only small effect sizes have been found for complex diseases. However, smaller effect sizes of variants do not mean that the biological mechanisms underlying these associations are less important. Thus, there are good reasons to perform even larger GWAS meta-analyses than those that have already been performed [235]. Analyses which include multiple ethnic groups, such as African Americans, Hispanics, and Asians, would help to localize the causing variant given that T2D prevalence varies by ancestry and the most useful biomarkers may also vary across ethnicities [235].

Furthermore, the genes related to T2D could be validated by eQTL studies, i.e. associations of gene variants from GWAS studied in relation to gene transcripts from the region, or by targeted proteomics, i.e. proteins encoded by genes in the GWAS loci examined for associations with the outcome [235]. Recently, Suhre et al. conducted a GWAS with intermediate phenotypes, such as

---

changes in metabolite and protein levels, that might provide functional evidence to map disease associations and translate them into clinical applications [236]. To understand the pathophysiology underlying the genotype-T2D association it would be beneficial to combine animal and human studies to examine tissue-specific effects, and thus achieving a more complete understanding, as the methods are complementary [235]. For example, Adam et al. compared the effect of metformin, a first-line oral medication to increase insulin sensitivity in patients with T2D, on humans and multiple murine tissues, which corroborated and complemented the findings from the human cohort [237].

Further characterization of gene functions could be obtained by involving combinations of other “omics” data as done in this thesis by including three different “omics” layers in one model or analyzing many polygenic variants with the complex disease [70, 71]. As conducted in this thesis for the interaction analysis, the number of associations can be reduced by using the identified GWAS loci in order to conduct interaction analysis or pathway analysis. In previous studies, interactions between SNPs and environment, such as physical activity, BMI, and waist circumference, and T2D or fasting insulin and glucose, have been studied [238-240]. They showed some evidence of gene-environment interactions in T2D.

Moreover, GWAS loci can be used to infer the causality between diseases such as T2D and environment through Mendelian randomization. For example, recently, Wahl et al. conducted a Mendelian randomization study relating DNA methylation and BMI and observed that changes in BMI appear to lead to changes in the DNA methylation, rather than the reverse [91]. Causality analysis of the ratio identified in this thesis and T2D could be similarly done using Mendelian randomization to find potential underlying biological mechanisms.

Probably it will not be possible to eliminate complex diseases such as T2D, but with further understanding of the underlying biological mechanisms, the focus of the investigations will turn from treatment to prevention.



## References

1. Putri, S.P., et al., *Current metabolomics: practical applications*. Journal of bioscience and bioengineering, 2013. **115**(6): p. 579-589.
2. Suhre, K., et al., *Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting*. PLoS. ONE, 2010. **5**(11): p. e13953.
3. Suhre, K., J. Raffler, and G. Kastenmuller, *Biochemical insights from population studies with genetics and metabolomics*. Arch Biochem Biophys, 2016. **589**: p. 168-76.
4. Gieger, C., et al., *Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum*. PLoS Genet, 2008. **4**(11): p. e1000282.
5. Illig, T., et al., *A genome-wide perspective of genetic variation in human metabolism*. Nat Genet, 2010. **42**(2): p. 137-41.
6. Kirk, P., et al., *Bayesian correlated clustering to integrate multiple datasets*. Bioinformatics, 2012. **28**(24): p. 3290-3297.
7. Wang, B., et al., *Similarity network fusion for aggregating data types on a genomic scale*. Nature methods, 2014. **11**(3): p. 333-337.
8. Krumsiek, J., J. Bartel, and F.J. Theis, *Computational approaches for systems metabolomics*. Current opinion in biotechnology, 2016. **39**: p. 198-206.
9. Yousri, N.A., et al., *A systems view of type 2 diabetes-associated metabolic perturbations in saliva, blood and urine at different timescales of glycaemic control*. Diabetologia, 2015. **58**(8): p. 1855-1867.
10. Petersen, A.K., et al., *Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits*. Hum Mol Genet, 2014. **23**(2): p. 534-45.
11. Draisma, H.H., et al., *Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels*. Nat Commun, 2015. **6**: p. 7208.
12. Heyn, H., et al., *Linkage of DNA methylation quantitative trait loci to human cancer risk*. Cell reports, 2014. **7**(2): p. 331-338.
13. Ma, Y., et al., *Interaction of methylation-related genetic variants with circulating fatty acids on plasma lipids: a meta-analysis of 7 studies and methylation analysis of 3 studies in the Cohorts for Heart and Aging Research in Genomic Epidemiology consortium*. The American journal of clinical nutrition, 2016. **103**(2): p. 567-578.
14. Bell, C.G., et al., *Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus*. PLoS One, 2010. **5**(11): p. e14040.
15. Lee, H., et al., *Allele-specific quantitative proteomics unravels molecular mechanisms modulated by cis-regulatory PPARG locus variation*. Nucleic acids research, 2017. **45**(6): p. 3266-3279.
16. Afendi, F.M., et al., *Data mining methods for omics and knowledge of crude medicinal plants toward big data biology*. Computational and Structural Biotechnology Journal, 2013. **4**(5): p. 1-14.
17. Ganna, A., et al., *Ultra-rare disruptive and damaging mutations influence educational attainment in the general population*. Nature neuroscience, 2016. **19**(12): p. 1563.
18. Fabregat-Traver, D., et al., *High-Performance Mixed Models Based Genome-Wide Association Analysis with omicABEL software*. F1000Res, 2014. **3**: p. 200.
19. Shabalin, A.A., *Matrix eQTL: ultra fast eQTL analysis via large matrix operations*. Bioinformatics, 2012. **28**(10): p. 1353-8.

- 
20. Hemani, G., et al., *EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards*. *Bioinformatics*, 2011. **27**(11): p. 1462-5.
  21. Sluga, D., et al., *Heterogeneous computing architecture for fast detection of SNP-SNP interactions*. *BMC Bioinformatics*, 2014. **15**: p. 216.
  22. Hasin, Y., M. Seldin, and A. Lusis, *Multi-omics approaches to disease*. *Genome Biology*, 2017. **18**(1): p. 83.
  23. Wishart, D.S., et al., *HMDB 3.0—the human metabolome database in 2013*. *Nucleic acids research*, 2012: p. gks1065.
  24. Beger, R.D., et al., *Metabolomics enables precision medicine: “A white paper, community perspective”*. *Metabolomics*, 2016. **12**(10): p. 149.
  25. Kastenmüller, G., et al., *Genetics of human metabolism: an update*. *Human molecular genetics*, 2015. **24**(R1): p. R93-R101.
  26. LeWitt, P.A., et al., *Metabolomic biomarkers as strong correlates of Parkinson disease progression*. *Neurology*, 2017: p. 10.1212/WNL.0000000000003663.
  27. Wang-Sattler, R., et al., *Novel biomarkers for pre-diabetes identified by metabolomics*. *Mol Syst Biol*, 2012. **8**: p. 615.
  28. Hughes, I., *Pediatric Endocrinology and Inborn Errors of Metabolism*. *Journal of Paediatrics and Child Health*, 2009. **45**(7-8): p. 478-478.
  29. Fearnley, L.G. and M. Inouye, *Metabolomics in epidemiology: from metabolite concentrations to integrative reaction networks*. *International journal of epidemiology*, 2016: p. dyw046.
  30. Suhre, K., et al., *Human metabolic individuality in biomedical and pharmaceutical research*. *Nature*, 2011. **477**(7362): p. 54-60.
  31. Heemskerk, M.M., et al., *Reanalysis of mGWAS results and in vitro validation show that lactate dehydrogenase interacts with branched-chain amino acid metabolism*. *European Journal of Human Genetics*, 2016. **24**(1): p. 142-145.
  32. DeHaven, C.D., et al., *Organization of GC/MS and LC/MS metabolomics data into chemical libraries*. *Journal of cheminformatics*, 2010. **2**(1): p. 9.
  33. Gorrochategui, E., et al., *Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow*. *TrAC Trends in Analytical Chemistry*, 2016. **82**: p. 425-442.
  34. Sas, K.M., et al., *Metabolomics and diabetes: analytical and computational approaches*. *Diabetes*, 2015. **64**(3): p. 718-732.
  35. Berezin, A.E., *Biomarkers for cardiovascular risk in patients with diabetes*. 2016, BMJ Publishing Group Ltd and British Cardiovascular Society.
  36. Ogurtsova, K., et al., *IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040*. *Diabetes Research and Clinical Practice*, 2017.
  37. Goffrier, B., M. Schulz, and J. Bätzing-Feigenbaum, *Administrative Prävalenzen und Inzidenzen des Diabetes mellitus von 2009 bis 2015*. *Zentralinstitut für die kassenärztliche Versorgung in Deutschland (Zi)*. 2017. **17**(03).
  38. Rathmann, W., et al., *High prevalence of undiagnosed diabetes mellitus in Southern Germany: target populations for efficient screening. The KORA survey 2000*. *Diabetologia*, 2003. **46**(2): p. 182-9.
  39. Park, C., et al., *Blood sugar level follows perceived time rather than actual time in people with type 2 diabetes*. *Proceedings of the National Academy of Sciences*, 2016: p. 201603444.

40. Muoio, D.M. and C.B. Newgard, *Molecular and metabolic mechanisms of insulin resistance and  $\beta$ -cell failure in type 2 diabetes*. Nature reviews Molecular cell biology, 2008. **9**(3): p. 193-205.
41. McCarthy, M.I., *Genomics, type 2 diabetes, and obesity*. N Engl J Med, 2010. **363**(24): p. 2339-50.
42. Willemsen, G., et al., *The concordance and heritability of type 2 diabetes in 34,166 twin pairs from international twin registers: The discordant twin (DISCOTWIN) consortium*. Twin Research and Human Genetics, 2015. **18**(06): p. 762-771.
43. Prasad, R.B. and L. Groop, *Genetics of type 2 diabetes—pitfalls and possibilities*. Genes, 2015. **6**(1): p. 87-123.
44. Capes, S.E., et al., *Stress hyperglycaemia and increased risk of death after myocardial infarction in patients with and without diabetes: a systematic overview*. The Lancet, 2000. **355**(9206): p. 773-778.
45. Reitmeir, P., et al., *Common eye diseases in older adults of southern Germany: results from the KORA-Age study*. Age and Ageing, 2016.
46. Floegel, A., et al., *Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach*. Diabetes, 2013. **62**(2): p. 639-48.
47. Roberts, L.D., A. Koulman, and J.L. Griffin, *Towards metabolic biomarkers of insulin resistance and type 2 diabetes: progress from the metabolome*. Lancet Diabetes Endocrinol, 2014. **2**(1): p. 65-75.
48. Suhre, K., *Metabolic profiling in diabetes*. J Endocrinol, 2014. **221**(3): p. R75-85.
49. Sas, K.M., et al., *Metabolomics and diabetes: analytical and computational approaches*. Diabetes, 2015. **64**(3): p. 718-32.
50. Guasch-Ferre, M., et al., *Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis*. Diabetes Care, 2016. **39**(5): p. 833-46.
51. Boden, G. and G. Shulman, *Free fatty acids in obesity and type 2 diabetes: defining their role in the development of insulin resistance and  $\beta$ -cell dysfunction*. European journal of clinical investigation, 2002. **32**(s3): p. 14-23.
52. Mihalik, S.J., et al., *Increased levels of plasma acylcarnitines in obesity and type 2 diabetes and identification of a marker of glucolipotoxicity*. Obesity, 2010. **18**(9): p. 1695-1700.
53. Research, W.H.O.A.C.o.H., *Genomics and world health: Report of the Advisory Committee on Health Research*. 2002: World Health Organization.
54. Klug, A., *The discovery of the DNA double helix*. DNA, Changing Science and Society, 2004.
55. Saha, A., J. Wittmeyer, and B.R. Cairns, *Chromatin remodelling: the industrial revolution of DNA around histones*. Nature reviews Molecular cell biology, 2006. **7**(6): p. 437-447.
56. Kimmins, S. and P. Sassone-Corsi, *Chromatin remodelling and epigenetic features of germ cells*. Nature, 2005. **434**(7033): p. 583-589.
57. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing*. Nature, 2010. **467**(7319): p. 1061-73.
58. Uppu, S., A. Krishna, and R. Gopalan, *A review of machine learning and statistical approaches for detecting SNP interactions in high-dimensional genomic data*. IEEE/ACM transactions on computational biology and bioinformatics, 2016.

- 
59. Hinds, D.A., et al., *Common deletions and SNPs are in linkage disequilibrium in the human genome*. *Nature genetics*, 2006. **38**(1): p. 82-85.
  60. Ardlie, K.G., L. Kruglyak, and M. Seielstad, *Patterns of linkage disequilibrium in the human genome*. *Nature Reviews Genetics*, 2002. **3**(4): p. 299-309.
  61. Wang, D.G., et al., *Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome*. *Science*, 1998. **280**(5366): p. 1077-1082.
  62. Frazer, K.A., et al., *A second generation human haplotype map of over 3.1 million SNPs*. *Nature*, 2007. **449**(7164): p. 851-861.
  63. Consortium, G.P., *An integrated map of genetic variation from 1,092 human genomes*. *Nature*, 2012. **491**(7422): p. 56-65.
  64. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies*. *Nat Rev Genet*, 2010. **11**(7): p. 499-511.
  65. Wigginton, J.E., D.J. Cutler, and G.R. Abecasis, *A note on exact tests of Hardy-Weinberg equilibrium*. *The American Journal of Human Genetics*, 2005. **76**(5): p. 887-893.
  66. Hirschhorn, J.N. and M.J. Daly, *Genome-wide association studies for common diseases and complex traits*. *Nature Reviews Genetics*, 2005. **6**(2): p. 95-108.
  67. Visscher, P.M., et al., *Five years of GWAS discovery*. *The American Journal of Human Genetics*, 2012. **90**(1): p. 7-24.
  68. Locke, A.E., et al., *Genetic studies of body mass index yield new insights for obesity biology*. *Nature*, 2015. **518**(7538): p. 197-206.
  69. Kutalik, Z., et al., *Novel method to estimate the phenotypic variation explained by genome-wide association studies reveals large fraction of the missing heritability*. *Genetic epidemiology*, 2011. **35**(5): p. 341-349.
  70. Zuk, O., et al., *The mystery of missing heritability: Genetic interactions create phantom heritability*. *Proc Natl Acad Sci U S A*, 2012. **109**(4): p. 1193-8.
  71. Power, R.A., J. Parkhill, and T. de Oliveira, *Microbial genome-wide association studies: lessons from human GWAS*. *Nature Reviews Genetics*, 2016.
  72. Pandey, A.K., et al., *Functionally enigmatic genes: a case study of the brain ignorome*. *PLoS One*, 2014. **9**(2): p. e88889.
  73. Solovieff, N., et al., *Pleiotropy in complex traits: challenges and strategies*. *Nature Reviews Genetics*, 2013. **14**(7): p. 483-495.
  74. Casale, F.P., et al., *Efficient set tests for the genetic analysis of correlated traits*. *Nature methods*, 2015. **12**(8): p. 755-758.
  75. Zhou, X. and M. Stephens, *Efficient multivariate linear mixed model algorithms for genome-wide association studies*. *Nature methods*, 2014. **11**(4): p. 407-409.
  76. Bolormaa, S., et al., *A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle*. *PLoS Genet*, 2014. **10**(3): p. e1004198.
  77. Shen, X., et al., *Simple multi-trait analysis identifies novel loci associated with growth and obesity measures*. *bioRxiv*, 2015: p. 022269.
  78. Slotkin, R.K. and R. Martienssen, *Transposable elements and the epigenetic regulation of the genome*. *Nature Reviews Genetics*, 2007. **8**(4): p. 272-285.
  79. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond*. *Nature Reviews Genetics*, 2012. **13**(7): p. 484-492.



80. Robertson, K.D., *DNA methylation and human disease*. Nature Reviews Genetics, 2005. **6**(8): p. 597-610.
81. Calo, E. and J. Wysocka, *Modification of enhancer chromatin: what, how, and why?* Molecular cell, 2013. **49**(5): p. 825-837.
82. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription*. Genes & development, 2011. **25**(10): p. 1010-1022.
83. Schenkel, L.C., et al., *DNA methylation analysis in constitutional disorders: clinical implications of the epigenome*. Critical reviews in clinical laboratory sciences, 2016. **53**(3): p. 147-165.
84. Weber, M., et al., *Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome*. Nature genetics, 2007. **39**(4): p. 457-466.
85. Antequera, F., *Structure, function and evolution of CpG island promoters*. Cellular and Molecular Life Sciences, 2003. **60**(8): p. 1647-1658.
86. Laird, P.W., *Principles and challenges of genome-wide DNA methylation analysis*. Nature Reviews Genetics, 2010. **11**(3): p. 191-203.
87. Ball, M.P., et al., *Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells*. Nature biotechnology, 2009. **27**(4): p. 361-368.
88. Jones, P.A., *The DNA methylation paradox*. Trends in Genetics, 1999. **15**(1): p. 34-37.
89. Maor, G.L., A. Yearim, and G. Ast, *The alternative role of DNA methylation in splicing regulation*. Trends in Genetics, 2015. **31**(5): p. 274-280.
90. Ehrlich, M., et al., *Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry*. Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(44): p. 15785-15790.
91. Wahl, S., *Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity*. Nature, 2016.
92. Zeilinger, S., et al., *Tobacco smoking leads to extensive genome-wide changes in DNA methylation*. PLoS One, 2013. **8**(5): p. e63812.
93. Etchegaray, J.-P. and R. Mostoslavsky, *Interplay between metabolism and epigenetics: a nuclear adaptation to environmental changes*. Molecular cell, 2016. **62**(5): p. 695-711.
94. Ladd-Acosta, C. and M.D. Fallin, *The role of epigenetics in genetic and environmental epidemiology*. 2015.
95. Smith, A.K., et al., *Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type*. BMC genomics, 2014. **15**(1): p. 145.
96. McDaniell, R., et al., *Heritable individual-specific and allele-specific chromatin signatures in humans*. Science, 2010. **328**(5975): p. 235-239.
97. Lu, C. and C.B. Thompson, *Metabolic regulation of epigenetics*. Cell metabolism, 2012. **16**(1): p. 9-17.
98. Gyenesei, A., et al., *BiForce Toolbox: powerful high-throughput computational analysis of gene-gene interactions in genome-wide association studies*. Nucleic Acids Res, 2012. **40**(Web Server issue): p. W628-32.
99. Lee, S., et al., *An Efficient Nonlinear Regression Approach for Genome-wide Detection of Marginal and Interacting Genetic Variations*. Journal of Computational Biology, 2016. **23**(5): p. 372-389.

- 
100. Gershon, E.S., N. Alliey-Rodriguez, and C. Liu, *After GWAS: searching for genetic risk for schizophrenia and bipolar disorder*. American Journal of Psychiatry, 2011. **168**(3): p. 253-256.
  101. Marjoram, P., A. Zubair, and S. Nuzhdin, *Post-GWAS: where next? More samples, more SNPs or more biology?* Heredity, 2014. **112**(1): p. 79.
  102. Frau, F., et al., *Type-2 diabetes-associated variants with cross-trait relevance: Post-GWAs strategies for biological function interpretation*. Molecular Genetics and Metabolism, 2017. **121**(1): p. 43-50.
  103. Molnos, S., et al., *The ratio of the metabolites valine and phosphatidylcholine acyl-alkyl C32:2 associates with measures of insulin secretion and increased risk of type 2 diabetes; a DIRECT study*. Diabetologia, 2017. **In press**.
  104. Holle, R., et al., *KORA--a research platform for population based health research*. Gesundheitswesen, 2005. **67 Suppl 1**: p. S19-25.
  105. Boeing, H., A. Korfmann, and M.M. Bergmann, *Recruitment procedures of EPIC-Germany. European Investigation into Cancer and Nutrition*. Ann Nutr Metab, 1999. **43**(4): p. 205-15.
  106. Westendorp, R.G., et al., *Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: The Leiden Longevity Study*. J. Am. Geriatr. Soc, 2009. **57**(9): p. 1634-1637.
  107. Rozing, M.P., et al., *Favorable glucose tolerance and lower prevalence of metabolic syndrome in offspring without diabetes mellitus of nonagenarian siblings: the Leiden longevity study*. J. Am. Geriatr. Soc, 2010. **58**(3): p. 564-569.
  108. Willemsen, G., et al., *The Netherlands Twin Register biobank: a resource for genetic epidemiological studies*. Twin Res Hum Genet, 2010. **13**(3): p. 231-45.
  109. Buuren, S. and K. Groothuis-Oudshoorn, *mice: Multivariate imputation by chained equations in R*. Journal of statistical software, 2011. **45**(3).
  110. Draisma, H.H., et al., *Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels*. Nature communications, 2015. **6**.
  111. Römisch-Margl, W., et al., *Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics*. Metabolomics, 2012. **8**(1): p. 133-142.
  112. Jourdan, C., et al., *Body fat free mass is associated with the serum metabolite profile in a population-based study*. PLoS One, 2012. **7**(6): p. e40009.
  113. Higgins, J.P., I.R. White, and A.M. Wood, *Imputation methods for missing outcome data in meta-analysis of clinical trials*. Clinical Trials, 2008. **5**(3): p. 225-239.
  114. White, I.R., P. Royston, and A.M. Wood, *Multiple imputation using chained equations: issues and guidance for practice*. Statistics in medicine, 2011. **30**(4): p. 377-399.
  115. Schafer, J.L. and J.W. Graham, *Missing data: our view of the state of the art*. Psychological methods, 2002. **7**(2): p. 147.
  116. Rässler, S., D.B. Rubin, and E.R. Zell, *19 Incomplete Data in Epidemiology and Medical Statistics*. Handbook of Statistics, 2007. **27**: p. 569-601.
  117. Rubin, D.B., *Multiple imputation for nonresponse in surveys*. Vol. 81. 2004: John Wiley & Sons.
  118. Azur, M.J., et al., *Multiple imputation by chained equations: what is it and how does it work?* International journal of methods in psychiatric research, 2011. **20**(1): p. 40-49.

119. Livshits, G., et al., *An omics investigation into chronic widespread musculoskeletal pain reveals epiandrosterone sulfate as a potential biomarker*. Pain, 2015. **156**(10): p. 1845.
120. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies*. PLoS Genet, 2009. **5**(6): p. e1000529.
121. O'Connell, J., et al., *A general approach for haplotype phasing across the full spectrum of relatedness*. PLoS Genet, 2014. **10**(4): p. e1004234.
122. Bibikova, M., et al., *High density DNA methylation array with single CpG site resolution*. Genomics, 2011. **98**(4): p. 288-95.
123. Du, P., et al., *Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis*. BMC bioinformatics, 2010. **11**(1): p. 587.
124. Aryee, M.J., et al., *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays*. Bioinformatics, 2014. **30**(10): p. 1363-1369.
125. Smyth, G.K., *Limma: linear models for microarray data*, in *Bioinformatics and computational biology solutions using R and Bioconductor*. 2005, Springer. p. 397-420.
126. Chen, Y.A., et al., *Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray*. Epigenetics, 2013. **8**(2): p. 203-9.
127. Houseman, E.A., et al., *DNA methylation arrays as surrogate measures of cell mixture distribution*. BMC Bioinformatics, 2012. **13**: p. 86.
128. Lehne, B., et al., *A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies*. Genome biology, 2015. **16**(1): p. 37.
129. Rathmann, W., et al., *Prediction models for incident type 2 diabetes mellitus in the older population: KORA S4/F4 cohort study*. Diabetic Medicine, 2010. **27**(10): p. 1116-1123.
130. Wilks, D.S., *Statistical methods in the atmospheric sciences*. Vol. 100. 2011: Academic press.
131. Hocking, R.R., *Methods and applications of linear models: regression and the analysis of variance*. 2013: John Wiley & Sons.
132. Ewens, W.J. and G.R. Grant, *Statistical methods in bioinformatics: an introduction*. 2006: Springer Science & Business Media.
133. Myers, R.H., et al., *Generalized linear models: with applications in engineering and the sciences*. Vol. 791. 2012: John Wiley & Sons.
134. Srivastava, A.K., V.K. Srivastava, and A. Ullah, *The coefficient of determination and its adjusted version in linear regression models*. Econometric Reviews, 1995. **14**(2): p. 229-240.
135. Fahrmeir, L., et al., *Regression: models, methods and applications*. 2013: Springer Science & Business Media.
136. Holle, R., et al., *KORA--a research platform for population based health research*. Gesundheitswesen, 2005. **67 Suppl 1**: p. S19-25.
137. Rathmann, W., et al., *Incidence of Type 2 diabetes in the elderly German population and the effect of clinical and lifestyle risk factors: KORA S4/F4 cohort study*. Diabet Med, 2009. **26**(12): p. 1212-9.
138. Harrell, F., *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. 2015: Springer.
139. Moore, D.F., *Applied Survival Analysis Using R*. Use R! 2016, Cham.

- 
140. Kleinbaum, D.G. and M. Klein, *Survival analysis: a self-learning text*. 2006: Springer Science & Business Media.
  141. Kaplan, E.L. and P. Meier, *Nonparametric estimation from incomplete observations*. Journal of the American statistical association, 1958. **53**(282): p. 457-481.
  142. Therneau, T.M. and P.M. Grambsch, *Modeling survival data: extending the Cox model*. 2013: Springer Science & Business Media.
  143. Cox, D.R., *Regression models and life-tables*, in *Breakthroughs in statistics*. 1992, Springer. p. 527-541.
  144. Prentice, R.L., *A case-cohort design for epidemiologic cohort studies and disease prevention trials*. Biometrika, 1986. **73**(1): p. 1-11.
  145. Consortium, I., *Design and cohort description of the InterAct Project: an examination of the interaction of genetic and lifestyle factors on the incidence of type 2 diabetes in the EPIC Study*. Diabetologia, 2011. **54**(9): p. 2272.
  146. Onland-Moret, N.C., et al., *Analysis of case-cohort data: a comparison of different methods*. Journal of clinical epidemiology, 2007. **60**(4): p. 350-355.
  147. Hoffmann, K., et al., *A statistical test for the equality of differently adjusted incidence rate ratios*. American journal of epidemiology, 2008. **167**(5): p. 517-522.
  148. Lin, D.Y. and L.-J. Wei, *The robust inference for the Cox proportional hazards model*. Journal of the American statistical Association, 1989. **84**(408): p. 1074-1078.
  149. Freedman, D.A., *On the so-called "Huber sandwich estimator" and "robust standard errors"*. The American Statistician, 2006. **60**(4): p. 299-302.
  150. Abbasi, A., et al., *External validation of the KORA S4/F4 prediction models for the risk of developing type 2 diabetes in older adults: the PREVEND study*. Eur J Epidemiol, 2012. **27**(1): p. 47-52.
  151. Kengne, A.P., et al., *Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models*. Lancet Diabetes Endocrinol, 2014. **2**(1): p. 19-29.
  152. Schulze, M.B., et al., *An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes*. Diabetes Care, 2007. **30**(3): p. 510-5.
  153. Therneau, T. and T. Lumley, *Survival: Survival analysis, including penalised likelihood. R package version 2.36-5*. Survival: Survival analysis, including penalised likelihood. R package version, 2011: p. 2.36-2.2010.
  154. Allison, P., *Estimating Cox regression models with PROC PHREG*. Anonymous Survival Analysis Using the SAS System. Cary, NC: SAS Institute Inc, 2000. **11184**.
  155. Powers, D.M., *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. 2011.
  156. Chap, T.L., *Introductory biostatistics*. A John Wiley & Sons Publication, 2003.
  157. Cook, N.R., *Use and misuse of the receiver operating characteristic curve in risk prediction*. Circulation, 2007. **115**(7): p. 928-935.
  158. Heagerty, P.J., T. Lumley, and M.S. Pepe, *Time-dependent ROC curves for censored survival data and a diagnostic marker*. Biometrics, 2000. **56**(2): p. 337-344.
  159. Akritas, M.G., *Nearest neighbor estimation of a bivariate distribution under random censoring*. The Annals of Statistics, 1994: p. 1299-1327.
  160. Morse, E.C., *Analytical methods for nonproliferation*. 2016: Springer.

161. Heagerty, P.J. and Y. Zheng, *Survival model predictive accuracy and ROC curves*. Biometrics, 2005. **61**(1): p. 92-105.
162. Hawkins, D.M., *The problem of overfitting*. J Chem Inf Comput Sci, 2004. **44**(1): p. 1-12.
163. Steyerberg, E.W., et al., *Internal validation of predictive models: efficiency of some procedures for logistic regression analysis*. J Clin Epidemiol, 2001. **54**(8): p. 774-81.
164. Alfons, A., *cvTools: Cross-validation tools for regression models*. package version 0.3.0, 2012.
165. Pencina, M.J., R.B. D'Agostino, and R.S. Vasan, *Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond*. Statistics in medicine, 2008. **27**(2): p. 157-172.
166. Jewell, E.S., et al., *Net Reclassification Improvement*. Anesthesia & Analgesia, 2016. **122**(3): p. 818-824.
167. French, B., et al., *Development and evaluation of multi-marker risk scores for clinical prognosis*. Statistical methods in medical research, 2016. **25**(1): p. 255-271.
168. Pencina, M.J., R.B. D'Agostino, and E.W. Steyerberg, *Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers*. Statistics in medicine, 2011. **30**(1): p. 11-21.
169. Herder, C., et al., *Immunological and cardiometabolic risk factors in the prediction of type 2 diabetes and coronary events: MONICA/KORA Augsburg case-cohort study*. PLoS One, 2011. **6**(6): p. e19852.
170. Beasley, T.M., S. Erickson, and D.B. Allison, *Rank-based inverse normal transformations are increasingly used, but are they merited?* Behavior genetics, 2009. **39**(5): p. 580.
171. Box, G.E. and D.R. Cox, *An analysis of transformations*. Journal of the Royal Statistical Society. Series B (Methodological), 1964: p. 211-252.
172. Fox, J. and S. Weisberg, *An R Companion to Applied Regression*. Second ed. 2011: Sage.
173. Wollschläger, D., *Grundlagen der Datenanalyse mit R: eine anwendungsorientierte Einführung*. 2015: Springer-Verlag.
174. Vittinghoff, E., et al., *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. 2011: Springer Science & Business Media.
175. Hedderich, J. and L. Sachs, *Angewandte statistik: methodensammlung mit r*. 2015: Springer-Verlag.
176. Cheung, M.W.-L., *Meta-analysis: A structural equation modeling approach*. 2015: John Wiley & Sons.
177. Normand, S.-L.T., *Tutorial in biostatistics meta-analysis: formulating, evaluating, combining, and reporting*. Statistics in medicine, 1999. **18**(3): p. 321-359.
178. De Bakker, P.I., et al., *Practical aspects of imputation-driven meta-analysis of genome-wide association studies*. Human molecular genetics, 2008. **17**(R2): p. R122-R128.
179. Hartung, J. and G. Knapp, *On tests of the overall treatment effect in meta-analysis with normally distributed responses*. Stat Med, 2001. **20**(12): p. 1771-82.
180. Petersen, A.K., et al., *On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies*. BMC Bioinformatics, 2012. **13**: p. 120.
181. Saville, D. and G.R. Wood, *Statistical methods: The geometric approach*. 2012: Springer Science & Business Media.
182. Snedecor, G. and W. Cochran, *Statistical methods, ed. 6, Ames, Iowa, 1967*. Iowa State University Press, Section. **12**: p. 349-352.

- 
183. Eddelbuettel, D., et al., *Rcpp: Seamless R and C++ integration*. Journal of Statistical Software, 2011. **40**(8): p. 1-18.
  184. OpenMP, A., *OpenMP Application Program Interface V3. 0*. OpenMP Architecture Review Board, 2008.
  185. Shin, S.Y., et al., *An atlas of genetic influences on human blood metabolites*. Nat Genet, 2014. **46**(6): p. 543-50.
  186. Kooperberg, C. and M. LeBlanc, *Increasing the power of identifying gene× gene interactions in genome-wide association studies*. Genetic epidemiology, 2008. **32**(3): p. 255-263.
  187. Goldman, M., *The innovative medicines initiative: a European response to the innovation challenge*. Clinical Pharmacology & Therapeutics, 2012. **91**(3): p. 418-425.
  188. Koivula, R.W., et al., *Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: rationale and design of the epidemiological studies within the IMI DIRECT Consortium*. Diabetologia, 2014. **57**(6): p. 1132-1142.
  189. Hummel, S., et al., *Postpartum outcomes in women with gestational diabetes and their offspring: POGO study design and first-year results*. Rev Diabet Stud, 2013. **10**(1): p. 49-57.
  190. Wahl, S., et al., *Childhood obesity is associated with changes in the serum metabolite profile*. Obes Facts, 2012. **5**(5): p. 660-70.
  191. Zheng, Y., et al., *Cumulative consumption of branched-chain amino acids and incidence of type 2 diabetes*. International Journal of Epidemiology, 2016: p. dyw143.
  192. Menni, C., et al., *Biomarkers for type 2 diabetes and impaired fasting glucose using a non-targeted metabolomics approach*. Diabetes, 2013: p. DB\_130570.
  193. Walford, G.A., et al., *Branched chain and aromatic amino acids change acutely following two medical therapies for type 2 diabetes mellitus*. Metabolism, 2013. **62**(12): p. 1772-8.
  194. Wurtz, P., et al., *Branched-chain and aromatic amino acids are predictors of insulin resistance in young adults*. Diabetes Care, 2013. **36**(3): p. 648-55.
  195. Wurtz, P., et al., *Metabolic signatures of insulin resistance in 7,098 young adults*. Diabetes, 2012. **61**(6): p. 1372-80.
  196. Lynch, C.J. and S.H. Adams, *Branched-chain amino acids in metabolic signalling and insulin resistance*. Nature Reviews Endocrinology, 2014. **10**(12): p. 723-736.
  197. Olson, K.C., et al., *Alloisoleucine differentiates the branched-chain aminoacidemia of Zucker and dietary obese rats*. Obesity, 2014. **22**(5): p. 1212-1215.
  198. Mahendran, Y., et al., *Genetic evidence of a causal effect of insulin resistance on branched-chain amino acid levels*. Diabetologia, 2017. **60**(5): p. 873-878.
  199. Knebel, B., et al., *Specific metabolic profiles and their relationship to insulin resistance in recent-onset type-1 and type-2 diabetes*. J Clin Endocrinol Metab, 2016: p. jc20154133.
  200. Wahl, S., et al., *Comparative analysis of plasma metabolomics response to metabolic challenge tests in healthy subjects and influence of the FTO obesity risk allele*. Metabolomics, 2013. **10**(3): p. 386-401.
  201. Wishart, D.S., et al., *HMDB: a knowledgebase for the human metabolome*. Nucleic Acids Res, 2009. **37**(Database issue): p. D603-10.
  202. Wallner, S. and G. Schmitz, *Plasmalogens the neglected regulatory and scavenging lipid species*. Chemistry and Physics of Lipids, 2011. **164**(6): p. 573-589.

203. Colas, R., et al., *Increased lipid peroxidation in LDL from type-2 diabetic patients*. *Lipids*, 2010. **45**(8): p. 723-731.
204. Pietiläinen, K.H., et al., *Acquired obesity is associated with changes in the serum lipidomic profile independent of genetic effects—a monozygotic twin study*. *PloS one*, 2007. **2**(2): p. e218.
205. Green, C.R., et al., *Branched-chain amino acid catabolism fuels adipocyte differentiation and lipogenesis*. *Nat Chem Biol*, 2016. **12**(1): p. 15-21.
206. Crown, S.B., N. Marze, and M.R. Antoniewicz, *Catabolism of Branched Chain Amino Acids Contributes Significantly to Synthesis of Odd-Chain and Even-Chain Fatty Acids in 3T3-L1 Adipocytes*. *PLoS One*, 2015. **10**(12): p. e0145850.
207. Halama, A., et al., *Metabolic switch during adipogenesis: From branched chain amino acid catabolism to lipid synthesis*. *Arch Biochem Biophys*, 2016. **589**: p. 93-107.
208. Jang, C., et al., *A branched-chain amino acid metabolite drives vascular fatty acid transport and causes insulin resistance*. *Nature medicine*, 2016. **22**(4): p. 421-426.
209. Shulman, G.I., *Ectopic fat in insulin resistance, dyslipidemia, and cardiometabolic disease*. *New England Journal of Medicine*, 2014. **371**(12): p. 1131-1141.
210. Lotta, L.A., 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*, 2016. **13**(11): p. e1002179.
211. McVicker, G., et al., *Identification of genetic variants that affect histone modifications in human cells*. *Science*, 2013. **342**(6159): p. 747-749.
212. Stroustrup, B., *Programming: principles and practice using C++*. 2014: Pearson Education.
213. Eddelbuettel, D., *Seamless R and C++ integration with Rcpp*. 2013: Springer.
214. Corydon, M.J., et al., *Role of common gene variations in the molecular pathogenesis of short-chain acyl-CoA dehydrogenase deficiency*. *Pediatric research*, 2001. **49**(1): p. 18-23.
215. Nochi, Z., R.K.J. Olsen, and N. Gregersen, *Short-chain acyl-CoA dehydrogenase deficiency: from gene to cell pathology and possible disease mechanisms*. *Journal of Inherited Metabolic Disease*, 2017: p. 1-15.
216. Corydon, M.J., et al., *Structural organization of the human short-chain acyl-CoA dehydrogenase gene*. *Mammalian genome*, 1997. **8**(12): p. 922-926.
217. Chen, Y. and Z. Su, *Reveal genes functionally associated with ACADS by a network study*. *Gene*, 2015. **569**(2): p. 294-302.
218. Khan, A., J.B. Bridgers, and B.D. Strahl, *Expanding the Reader Landscape of Histone Acylation*. *Structure*, 2017. **25**(4): p. 571-573.
219. Sabari, B.R., et al., *Metabolic regulation of gene expression through histone acylations*. *Nature Reviews Molecular Cell Biology*, 2016.
220. Tonin, R., et al., *Clinical relevance of short-chain acyl-CoA dehydrogenase (SCAD) deficiency: Exploring the role of new variants including the first SCAD-disease-causing allele carrying a synonymous mutation*. *BBA clinical*, 2016. **5**: p. 114-119.
221. van Maldegem, B.T., et al., *Clinical, biochemical, and genetic heterogeneity in short-chain acyl-coenzyme A dehydrogenase deficiency*. *Jama*, 2006. **296**(8): p. 943-952.
222. Schuck, P.F., et al., *Promotion of lipid and protein oxidative damage in rat brain by ethylmalonic acid*. *Neurochemical research*, 2010. **35**(2): p. 298-305.

- 
223. Davie, J.R., *Inhibition of histone deacetylase activity by butyrate*. The Journal of nutrition, 2003. **133**(7): p. 2485S-2493S.
224. Pallister, T., et al., *Untangling the relationship between diet and visceral fat mass through blood metabolomics and gut microbiome profiling*. International Journal of Obesity, 2017.
225. Lustgarten, M.S., et al., *Serum glycine is associated with regional body fat and insulin resistance in functionally-limited older adults*. PLoS One, 2013. **8**(12): p. e84034.
226. Moore, S.C., et al., *Human metabolic correlates of body mass index*. Metabolomics, 2014. **10**(2): p. 0.
227. Long, T., et al., *Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites*. Nature Genetics, 2017.
228. van Maldegem, B.T., et al., *Flavin adenine dinucleotide status and the effects of high-dose riboflavin treatment in short-chain acyl-CoA dehydrogenase deficiency*. Pediatric research, 2010. **67**(3): p. 304-308.
229. Yet, I., et al., *Genetic influences on metabolite levels: a comparison across metabolomic platforms*. PloS one, 2016. **11**(4): p. e0153672.
230. Wei, W.-H., G. Hemani, and C.S. Haley, *Detecting epistasis in human complex traits*. Nature Reviews Genetics, 2014. **15**(11): p. 722-733.
231. Hemani, G., et al., *Detection and replication of epistasis influencing transcription in humans*. Nature, 2014. **508**(7495): p. 249-253.
232. Brown, A.A., et al., *Genetic interactions affecting human gene expression identified by variance association mapping*. Elife, 2014. **3**: p. e01381.
233. Wood, A.R., et al., *Another explanation for apparent epistasis*. Nature, 2014. **514**(7520): p. E3-E5.
234. Fuchsberger, C., et al., *The genetic architecture of type 2 diabetes*. Nature, 2016.
235. Florez, J.C., *The Genetics of Type 2 Diabetes and Related Traits*. 2016: Springer Nature. 571.
236. Suhre, K., et al., *Connecting genetic risk to disease end points through the human blood plasma proteome*. Nature communications, 2017. **8**: p. 14357.
237. Adam, J., et al., *Metformin effect on non-targeted metabolite profiles in patients with type 2 diabetes and multiple murine tissues*. Diabetes, 2016: p. db160512.
238. Langenberg, C., et al., *Gene-lifestyle interaction and type 2 diabetes: the EPIC interact case-cohort study*. PLoS Med, 2014. **11**(5): p. e1001647.
239. Brito, E.C., et al., *Previously associated type 2 diabetes variants may interact with physical activity to modify the risk of impaired glucose regulation and type 2 diabetes*. Diabetes, 2009. **58**(6): p. 1411-1418.
240. Manning, A.K., et al., *A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance*. Nat Genet, 2012. **44**(6): p. 659-669.



## Appendix

**Table A.1:** List of metabolites measured with the Absolute*IDQ*<sup>tm</sup> p180 Kit  
\* not measured with p150 Kit.

<b>Acylcarnitines (40)</b>			
C0	Carnitine	C10:1	Decenoylcarnitine
C2	Acetylcarnitine	C10:2	Decadienylcarnitine
C3	Propionylcarnitine	C12	Dodecanoylcarnitine
C3:1	Propenoylcarnitine	C12:1	Dodecenoylcarnitine
C3-OH	Hydroxypropionylcarnitine	C12-DC	Dodecanedioylcarnitine
C4	Butyrylcarnitine	C14	Tetradecanoylcarnitine
C4:1	Butenoylcarnitine	C14:1	Tetradecenoylcarnitine
C4-OH (C3-DC)	Hydroxybutyrylcarnitine	C14:1-OH	Hydroxytetradecenoylcarnitine
C5	Valerylcarnitine	C14:2	Tetradecadienylcarnitine
C5:1	Tiglylcarnitine	C14:2-OH	Hydroxytetradecadienylcarnitine
C5:1-DC	Glutaconylcarnitine	C16	Hexadecanoylcarnitine
C5-DC (C6-OH)	Glutaryl carnitine (Hydroxyhexanoylcarnitine)	C16:1	Hexadecenoylcarnitine
C5-M-DC	Methylglutaryl carnitine	C16:1-OH	Hydroxyhexadecenoylcarnitine
C5-OH (C3-DC-M)	Hydroxyvalerylcarnitine (Methylmalonylcarnitine)	C16:2	Hexadecadienylcarnitine
C6 (C4:1-DC)	Hexanoylcarnitine (Fumaryl carnitine)	C16:2-OH	Hydroxyhexadecadienylcarnitine
C6:1	Hexenoylcarnitine	C16-OH	Hydroxyhexadecanoylcarnitine
C7-DC	Pimelylcarnitine	C18	Octadecanoylcarnitine
C8	Octanoylcarnitine	C18:1	Octadecenoylcarnitine
C9	Nonanoylcarnitine	C18:1-OH	Hydroxyoctadecenoylcarnitine
C10	Decanoylcarnitine	C18:2	Octadecadienylcarnitine
<b>Amino Acids (21)</b>			
Ala*	Alanine	Lys*	Lysine
Arg	Arginine	Met	Methionine
Asn*	Asparagine	Orn	Ornithine
Asp*	Aspartate	Phe	Phenylalanine
Cit*	Citrulline	Pro	Proline

Gln	Glutamine	Ser	Serine
Glu*	Glutamate	Thr	Threonine
Gly	Glycine	Trp	Tryptophan
His	Histidine	Tyr	Tyrosine
Ile*	Isoleucine	Val	Valine
Leu*	Leucine	xLeu (p150 only)	Leucine/Isoleucine
Monosaccharides (1)			
Sum of Hexoses (including Glucose)			
Glycerophospholipids (90)			
lysoPC=lysoPhosphatidylCholine; PC=PhosphatidylCholine; a=acyl; aa=diacyl; ae=acyl-alkyl)			
lysoPC a C14:0	PC aa C34:1	PC aa C42:0	PC ae C38:2
lysoPC a C16:0	PC aa C34:2	PC aa C42:1	PC ae C38:3
lysoPC a C16:1	PC aa C34:3	PC aa C42:2	PC ae C38:4
lysoPC a C17:0	PC aa C34:4	PC aa C42:4	PC ae C38:5
lysoPC a C18:0	PC aa C36:0	PC aa C42:5	PC ae C38:6
lysoPC a C18:1	PC aa C36:1	PC aa C42:6	PC ae C40:1
lysoPC a C18:2	PC aa C36:2	PC ae C30:0	PC ae C40:2
lysoPC a C20:3	PC aa C36:3	PC ae C30:1	PC ae C40:3
lysoPC a C20:4	PC aa C36:4	PC ae C30:2	PC ae C40:4
lysoPC a C24:0	PC aa C36:5	PC ae C32:1	PC ae C40:5
lysoPC a C26:0	PC aa C36:6	PC ae C32:2	PC ae C40:6
lysoPC a C26:1	PC aa C38:0	PC ae C34:0	PC ae C42:0
lysoPC a C28:0	PC aa C38:1	PC ae C34:1	PC ae C42:1
lysoPC a C28:1	PC aa C38:3	PC ae C34:2	PC ae C42:2
PC aa C24:0	PC aa C38:4	PC ae C34:3	PC ae C42:3
PC aa C26:0	PC aa C38:5	PC ae C36:0	PC ae C42:4
PC aa C28:1	PC aa C38:6	PC ae C36:1	PC ae C42:5
PC aa C30:0	PC aa C40:1	PC ae C36:2	PC ae C44:3
PC aa C30:2	PC aa C40:2	PC ae C36:3	PC ae C44:4
PC aa C32:0	PC aa C40:3	PC ae C36:4	PC ae C44:5
PC aa C32:1	PC aa C40:4	PC ae C36:5	PC ae C44:6
PC aa C32:2	PC aa C40:5	PC ae C38:0	
PC aa C32:3	PC aa C40:6	PC ae C38:1	

Sphingolipids (15) SM=Sphingomyelin			
SM (OH) C14:1	SM C18:0	SM (OH) C22:1	SM (OH) C24:1
SM C16:0	SM C18:1	SM (OH) C22:2	SM C26:0
SM C16:1	SM C20:2	SM C24:0	SM C26:1
SM (OH) C16:1	SM C22:3	SM C24:1	
Biogenic Amines (21)			
Ac-Orn	Acetylorithine	PEA'	Phenylethylamine
ADMA*	Asymmetric dimethylarginine	OH-Pro*	4-Hydroxyproline
alpha-AAA*	alpha-Aminoadipic acid	Putrescine*	Putrescine
Carnosine*	Carnosine	Sarcosine*	Sarcosine
Creatinine*	Creatinine	SDMA*	Symmetric dimethylarginine
DOPA*	DOPA	Serotonin*	Serotonin
Dopamine*	Dopamine	Spermidine*	Spermidine
Histamine*	Histamine	Spermine*	Spermine
Kynurenine*	Kynurenine	Taurine*	Taurine
Met-SO*	Methionine sulfoxide	total DMA*	Total dimethylarginine
Nitro-Tyr*	Nitrotyrosine		

**Table A.2:** List of metabolites measured with Metabolon

Amino Acids (71)			
M00053	glutamine	M22138	homocitrulline
M00054	tryptophan	M27513	indoleacetate
M00059	histidine	M27672	3-indoxyl sulfate
M00060	leucine	M27710	N-acetylglycine
M00064	phenylalanine	M27718	creatine
M00513	creatinine	M31453	cysteine
M01125	isoleucine	M31454	cystine

M01284	threonine	M32197	3-(4-hydroxyphenyl)lactate
M01299	tyrosine	M32315	serine
M01301	lysine	M32319	trans-4-hydroxyproline
M01302	methionine	M32322	glutamate
M01444	pipecolate	M32338	glycine
M01493	ornithine	M32339	alanine
M01494	5-oxoproline	M32348	2-aminobutyrate
M01558	4-acetamidobutanoate	M32405	indolepropionate
M01585	N-acetylalanine	M32553	phenol sulfate
M01638	arginine	M32672	pyroglutamine*
M01649	valine	M32675	C-glycosyltryptophan*
M01670	urea	M32709	N-[3-(2-Oxopyrrolidin-1-yl)propyl]acetamide
M01898	proline	M33131	methylcysteine
M02132	citrulline	M33441	isobutyrylcarnitine
M02342	serotonin (5HT)	M33515	hydroxytryptophane*
M03141	betaine	M33937	alpha-hydroxyisovalerate
M12017	3-methoxytyrosine	M33939	N-acetylthreonine
M12129	beta-hydroxyisovalerate	M34283	asparagine
M15140	kynurenine	M34407	isovalerylcarnitine
M15630	N-acetylornithine	M35126	phenylacetylglutamine
M15676	3-methyl-2-oxovalerate	M35159	cysteine-glutathione disulfide
M15749	3-phenylpropionate (hydrocinnamate)	M35428	tiglylcarnitine
M15996	aspartate	M35431	2-methylbutyrylcarnitine
M16822	3,4-dihydroxybutyrate*	M35433	hydroxyisovaleroylcarnitine
M18349	indolelactate	M35439	glutaroylcarnitine

M21044	2-hydroxybutyrate (AHB)	M36103	p-cresol sulfate
M21047	3-methyl-2-oxobutyrate	M36808	dimethylarginine (SDMA + ADMA)
M22030	2-hydroxyisobutyrate	M37097	tryptophan betaine
M22116	4-methyl-2-oxopentanoate		
<b>Carbohydrate (12)</b>			
M00527	lactate	M15964	arabitol
M00577	fructose	M20489	glucose
M00584	mannose	M20675	1,5-anhydroglucitol (1,5-AG)
M00599	pyruvate	M27722	erythrose
M01572	glycerate	M33477	erythronate*
M15335	mannitol	M35854	threitol
<b>Cofactors and Vitamins (12)</b>			
M01508	pantothenate	M31555	pyridoxate
M01561	alpha-tocopherol	M32586	bilirubin (E;E)*
M01640	ascorbate (Vitamin C)	M32593	heme*
M02137	biliverdin	M32910	O-methylascorbate*
M27716	bilirubin (Z;Z)	M33138	oxidized bilirubin*
M27738	threonate	M34106	bilirubin (E;Z or Z;E)*
<b>Energy (6)</b>			
M01303	malate	M15488	acetylphosphate
M01564	citrate	M33453	alpha-ketoglutarate
M11438	phosphate	M37058	succinylcarnitine

**Lipids (114)**

M00063	cholesterol	M33883	12-hydroxyeicosatetraenoate (12-HETE)
M00542	3-hydroxybutyrate (BHBA)	M33884	5,8-tetradecadienoate
M01105	linoleate (18:2n6)	M33936	octanoylcarnitine
M01110	arachidonate (20:4n6)	M33941	decanoylcarnitine
M01114	deoxycholate	M33955	1-palmitoylglycerophosphocholine
M01121	margarate (17:0)	M33957	1-heptadecanoylglycerophosphocholine
M01336	palmitate (16:0)	M33960	1-oleoylglycerophosphocholine
M01356	nonadecanoate (19:0)	M33961	1-stearoylglycerophosphocholine
M01358	stearate (18:0)	M33968	5-dodecenoate (12:1n7)
M01359	oleate (18:1n9)	M33969	stearidonate (18:4n3)
M01361	pentadecanoate (15:0)	M33971	10-heptadecenoate (17:1n7)
M01365	myristate (14:0)	M33972	10-nonadecenoate (19:1n9)
M01481	inositol 1-phosphate (I1P)	M33973	epiandrosterone sulfate
M01642	caprate (10:0)	M34035	linolenate [alpha or gamma; (18:3n3 or 6)]
M01644	heptanoate (7:0)	M34214	1-arachidonoylglycerophosphoinositol*
M01645	laurate (12:0)	M34409	stearoylcarnitine
M01712	cortisol	M34416	1-stearoylglycerophosphoethanolamine

M01769	cortisone	M34419	1-linoleoylglycerophosphocholine
M12035	pelargonate (9:0)	M34534	laurylcarnitine
M12067	undecanoate (11:0)	M34674	docosapentaenoic acid (n6-DPA)
M15122	glycerol	M34732	isovalerate
M15365	glycerol 3-phosphate (G3P)	M34878	stearamide
M15500	carnitine	M35160	oleoylcarnitine
M15506	choline	M35186	1-arachidonoylglycerophosphoethanolamine*
M15990	glycerophosphorylcholine (GPC)	M35189	nonanoylcarnitine*
M17805	dihomo-linoleate (20:2n6)	M35253	2-palmitoylglycerophosphocholine*
M17945	2-hydroxystearate	M35254	2-oleoylglycerophosphocholine*
M18467	eicosapentaenoate (EPA; 20:5n3)	M35255	2-stearoylglycerophosphocholine*
M18476	glycocholate	M35257	2-linoleoylglycerophosphocholine*
M19323	docosahexaenoate (DHA; 22:6n3)	M35305	1-palmitoylglycerophosphoinositol*
M19324	1-stearoylglycerophosphoinositol	M35472	2-tetradecenoylcarnitine
M19934	myo-inositol	M35626	1-myristoylglycerophosphocholine

M21127	1-palmitoylglycerol (1-monopalmitin)	M35628	1-oleoylglycerophosphoethanolamine
M21184	1-oleoylglycerol (1-monoolein)	M35631	1-palmitoylglycerophosphoethanolamine
M21188	1-stearoylglycerol (1-monostearin)	M35675	2-hydroxypalmitate
M22189	palmitoylcarnitine	M35718	dihomo-linolenate (20:3n3 or n6)
M22842	cholate	M36754	octadecanedioate
M27447	1-linoleoylglycerol (1-monolinolein)	M36776	7-alpha-hydroxy-3-oxo-4-cholestenoate (7-Hoca)
M27531	hyodeoxycholate	M36802	n-butyl oleate
M31591	androsterone sulfate	M36850	tauroithocholate 3-sulfate
M31787	3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF)	M37190	5alpha-androstan-3beta,17beta-diol disulfate
M32198	acetylcarnitine	M37202	4-androsten-3beta,17beta-diol disulfate 1*
M32328	hexanoylcarnitine	M37203	4-androsten-3beta,17beta-diol disulfate 2*
M32346	glycochenodeoxycholate	M37253	2-hydroxyglutarate
M32379	scyllo-inositol	M37506	palmitoyl sphingomyelin
M32412	butyrylcarnitine	M38178	cis-4-decenoylcarnitine
M32418	myristoleate (14:1n5)	M38768	15-methylpalmitate (isobar with 2-methylpalmitate)
M32425	dehydroisoandrosterone sulfate (DHEA-S)	M33230	1-palmitoleoylglycerophosphocholine*
M32452	propionylcarnitine	M33443	valerate
M32455	linoleamide (18:2n6)	M33447	palmitoleate (16:1n7)



M32458	oleamide	M33587	eicosenoate (20:1n9 or 11)
M32489	caproate (6:0)	M33821	1-eicosatrienoylglycerophosphocholine*
M32492	caprylate (8:0)	M33822	1-docosahexaenoylglycerophosphocholine*
M32497	10-undecenoate (11:1n1)	M33871	1-eicosadienoylglycerophosphocholine*
M32504	docosapentaenoate (n3 DPA; 22:5n3)	M32762	5alpha-androstan-3beta,17beta-diol disulfate
M32635	1-linoleoylglycerophosphocholine*	M32980	adrenate (22:4n6)
M32654	3-dehydrocarnitine*	M33228	1-arachidonoylglycerophosphocholine*
<b>Nucleotide (11)</b>			
M00606	uridine	M15650	N1-methyladenosine
M01123	inosine	M32739	xanthine
M01573	guanosine	M33442	pseudouridine
M01604	urate	M33510	N1-methyl-3-pyridone-4-carboxamide
M03127	hypoxanthine	M35114	7-methylguanine
M03147	xanthine		
<b>Peptide (21)</b>			
M02730	gamma-glutamylglutamine	M33422	gamma-glutamylphenylalanine
M02734	gamma-glutamyltyrosine	M33801	ADpSGEGDFXAEGGGVR*

M18357	glycylvaline	M34420	bradykinin; des-arg(9)
M18369	gamma-glutamylleucine	M35127	pro-hydroxy-pro
M22175	aspartylphenylalanine	M36115	leucylalanine
M31548	DSGEGDFXAEGGGVR *	M36131	alpha-glutamyltyrosine
M32393	gamma-glutamylvaline	M36134	phenylalanylserine
M32836	HWESASXX*	M36230	leucylalanine
M33084	ADSGEGDFXAEGGGV R*	M36376	phenylalanylleucine
M33363	gamma-glutamylmethionine*	M38150	phenylalanylphenylalanine
M33364	gamma-glutamylthreonine*		
<b>Xenobiotics (21)</b>			
M00569	caffeine	M33935	piperine
M15753	hippurate	M34384	stachydrine
M15778	benzoate	M34389	1-methylxanthine
M18254	paraxanthine	M34390	7-methylxanthine
M18335	quininate	M34395	1-methylurate
M18392	theobromine	M35320	catechol sulfate
M18394	theophylline	M36098	4-vinylphenol sulfate
M20699	erythritol	M36099	4-ethylphenylsulfate
M27728	glycerol 2-phosphate	M37004	vanillin
M32445	3-methylxanthine	M37459	ergothioneine
M33009	homostachydrine*		
<b>Unknown (138)</b>			

M12593	X-02973	M33140	X-11795
M12626	X-03003	M33144	X-11799
M12768	X-03088	M33150	X-11805
M12770	X-03090	M33154	X-11809
M12774	X-03094	M33163	X-11818
M16634	X-04357	M33165	X-11820
M16816	X-04494	M33190	X-11845
M16818	X-04495	M33192	X-11847
M16821	X-04498	M33194	X-11849
M17807	X-18601	M33195	X-11850
M18283	X-05426	M33204	X-11859
M18929	X-05907	M33221	X-11876
M19362	X-06226	M33225	X-11880
M19363	X-06227	M33380	X-12029
M19364	X-06246	M33389	X-12038
M19368	X-06267	M33390	X-12039
M19396	X-06307	M33408	X-12056
M19414	X-06350	M33415	X-12063
M21630	X-08402	M33507	X-12092
M22481	X-08988	M33509	X-12094
M22548	X-09026	M33627	X-12206
M22649	X-09108	M33633	X-12212
M24074	X-09706	M33637	X-12216
M25459	X-10395	M33638	X-12217
M25599	X-10429	M33652	X-12230
M27256	X-10500	M33653	X-12231

---

M27273	X-10506	M33666	X-12244
M27278	X-10510	M33675	X-12253
M28354	X-10675	M33833	X-12405
M30805	X-10810	M33885	X-12443
M32518	X-11204	M33901	X-12456
M32549	X-02269	M33910	X-12465
M32557	X-06126	M34040	X-12510
M32564	X-11247	M34062	X-12524
M32578	X-11261	M34112	X-12544
M32587	X-02249	M34123	X-12556
M32616	X-11299	M34221	X-12627
M32632	X-11315	M34244	X-12644
M32634	X-11317	M34289	X-12680
M32644	X-11327	M34306	X-12696
M32651	X-11334	M34359	X-12749
M32691	X-11374	M34453	X-12776
M32698	X-11381	M34469	X-12786
M32729	X-11412	M34481	X-12798
M32735	X-01911	M34499	X-12816
M32740	X-11423	M34527	X-12844
M32753	X-09789	M34533	X-12850
M32754	X-11437	M34761	X-13069
M32755	X-11438	M35072	X-13372
M32757	X-11440	M35187	X-13429
M32758	X-11441	M35193	X-13435
M32759	X-11442	M35240	X-13477

---

M32761	X-11444	M35270	X-13496
M32769	X-11452	M35326	X-13548
M32786	X-11469	M35327	X-13549
M32787	X-11470	M35331	X-13553
M32802	X-11485	M35397	X-13619
M32808	X-11491	M35464	X-13671
M32814	X-11497	M35551	X-13741
M32838	X-11521	M35754	X-13859
M32846	X-11529	M35977	X-14056
M32847	X-11530	M35978	X-14057
M32854	X-11537	M36009	X-14086
M32855	X-11538	M36300	X-14374
M32857	X-11540	M36399	X-14473
M32863	X-11546	M36515	X-14588
M32867	X-11550	M36552	X-14625
M32869	X-11552	M36553	X-14626
M33132	X-11787	M36673	X-14745

**Table A.3:** Significant associations between metabolite butyrylcarnitine measured with the Biocrates platform and the interaction SNP  $\times$  CpG in chromosome 12.

CpG	SNP	P-value	Position (CpG)	Minor allele frequency	Reference allele	Effect allele	Position (SNP)
cg02419362	rs3794214	4.97E-14	121168245	0.42539	T	C	121203948
cg02419362	rs7964786	2.99E-15	121132100	0.36557	C	T	121203948
cg21721566	rs11065283	6.17E-16	121266453	0.47831	T	C	121163144
cg21721566	rs11065324	3.72E-16	121346795	0.47883	G	C	121163144
cg21721566	rs1696357	1.55E-15	121348649	0.4815	T	C	121163144
cg21721566	rs11065311	6.42E-16	121324115	0.47994	T	A	121163144
cg21721566	rs2001133	5.17E-14	121056754	0.49814	T	A	121163144
cg21721566	rs3213572	6.99E-16	121205078	0.4783	G	A	121163144
cg21721566	rs2047568	1.01E-15	121243790	0.47938	G	A	121163144
cg21721566	rs1151849	3.11E-15	121340599	0.48635	G	A	121163144
cg21721566	rs3883901	9.16E-16	121256520	0.47936	T	C	121163144
cg21721566	rs4454799	4.52E-16	121317075	0.47891	G	T	121163144
cg21721566	rs531782	4.64E-15	121273247	0.48666	T	C	121163144
cg21892295	rs11065283	1.28E-29	121266453	0.47831	T	C	121157589
cg21892295	rs1696357	1.61E-28	121348649	0.4815	T	C	121157589
cg21892295	rs1151851	1.26E-14	121340139	0.40487	A	T	121157589
cg21892295	rs11065286	1.14E-15	121271734	0.43872	T	C	121157589
cg21892295	rs2047568	1.84E-29	121243790	0.47938	G	A	121157589
cg21892295	rs11065311	2.34E-29	121324115	0.47994	T	A	121157589
cg21892295	rs2001133	1.51E-16	121056754	0.49814	T	A	121157589
cg21892295	rs11065324	1.57E-29	121346795	0.47883	G	C	121157589
cg21892295	rs4454799	1.40E-29	121317075	0.47891	G	T	121157589
cg21892295	rs1151849	1.79E-28	121340599	0.48635	G	A	121157589
cg21892295	rs3883901	2.06E-29	121256520	0.47936	T	C	121157589
cg21892295	rs3213572	3.62E-29	121205078	0.4783	G	A	121157589
cg21892295	rs2062507	2.03E-23	121259051	0.38659	C	T	121157589
cg21892295	rs531782	2.26E-28	121273247	0.48666	T	C	121157589

**Table A.4:** Significant associations between metabolite butyrylcarnitine measured with the Metabolon platform and the interaction SNP  $\times$  CpG in the Metabolon platform in chromosome 12.

CpG	SNP	P-value	Position (CpG)	Minor allele frequency	Reference allele	Effect allele	Position (SNP)
cg02419362	rs556001	3.09E-20	121152967	0.44425	A	G	121203948
cg02419362	rs10774563	2.93E-14	121060780	0.45882	G	A	121203948
cg02419362	rs555379	1.75E-16	121090498	0.41397	C	A	121203948
cg02419362	rs17847	1.05E-18	121149475	0.46543	A	G	121203948
cg02419362	rs513175	2.57E-14	121086234	0.37259	A	C	121203948
cg02419362	rs7976497	7.19E-19	121135467	0.46449	T	C	121203948
cg02419362	rs10431385	4.46E-18	121128882	0.41883	T	C	121203948
cg02419362	rs2239760	3.77E-18	121163518	0.41888	C	A	121203948
cg02419362	rs2001133	1.15E-14	121056754	0.49814	T	A	121203948
cg02419362	rs526007	2.83E-14	121085884	0.37264	A	G	121203948
cg02419362	rs10431384	2.67E-15	121127347	0.33097	A	G	121203948
cg02419362	rs11065202	1.25E-17	121112429	0.42711	T	C	121203948
cg02419362	rs3914	9.93E-19	121174899	0.4638	T	C	121203948
cg02419362	rs3794214	5.83E-19	121168245	0.42539	T	C	121203948
cg02419362	rs10431386	2.46E-15	121128926	0.33102	C	T	121203948
cg02419362	rs3794215	5.89E-19	121168083	0.42539	T	C	121203948
cg02419362	rs2005455	4.45E-18	121128699	0.41869	A	G	121203948
cg02419362	rs4766975	1.21E-14	121069126	0.36972	G	A	121203948
cg02419362	rs3829290	1.16E-18	121126438	0.46357	T	C	121203948
cg02419362	rs696337	6.80E-19	121159380	0.46329	T	C	121203948
cg02419362	rs473121	1.93E-16	121089739	0.41381	T	C	121203948
cg02419362	rs9204	7.03E-18	121177778	0.3488	A	G	121203948
cg02419362	rs522632	1.67E-16	121088886	0.41427	G	C	121203948
cg02419362	rs7964786	6.26E-18	121132100	0.36557	C	T	121203948
cg06793505	rs1542859	8.55E-20	121336766	0.47882	A	G	121164278
cg06793505	rs1151862	7.82E-17	121331285	0.48632	G	A	121164278
cg06793505	rs10849791	8.37E-20	121235280	0.47753	T	C	121164278
cg06793505	rs4454799	8.38E-20	121317075	0.47891	G	T	121164278
cg06793505	rs2393717	1.14E-14	121220375	0.46065	G	C	121164278
cg06793505	rs610694	8.21E-17	121304826	0.48619	T	C	121164278
cg06793505	rs12824150	7.66E-16	121220031	0.38613	G	C	121164278
cg06793505	rs11065311	1.11E-19	121324115	0.47994	T	A	121164278
cg06793505	rs10774572	8.13E-20	121309561	0.4789	A	G	121164278
cg06793505	rs11065292	8.00E-20	121289155	0.48002	T	C	121164278
cg06793505	rs520753	9.35E-17	121298993	0.48661	T	A	121164278

---

cg06793505	rs661647	8.49E-17	121282659	0.48537	C	T	121164278
cg06793505	rs11065282	1.04E-19	121264415	0.47675	G	T	121164278
cg06793505	rs11065324	1.39E-19	121346795	0.47883	G	C	121164278
cg06793505	rs11065300	7.90E-20	121298512	0.47881	G	C	121164278
cg06793505	rs11611087	7.85E-20	121267126	0.48125	C	T	121164278
cg06793505	rs1151849	7.65E-17	121340599	0.48635	G	A	121164278
cg06793505	rs11065283	7.05E-20	121266453	0.47831	T	C	121164278
cg06793505	rs525425	1.01E-19	121195625	0.4779	A	G	121164278
cg06793505	rs11065301	5.22E-16	121304368	0.38637	C	T	121164278
cg06793505	rs10849807	1.09E-19	121314056	0.47997	T	C	121164278
cg06793505	rs4767935	6.49E-16	121216531	0.38657	T	C	121164278
cg06793505	rs7137504	8.40E-20	121221628	0.47752	C	T	121164278
cg06793505	rs531782	7.39E-17	121273247	0.48666	T	C	121164278
cg06793505	rs1696357	2.36E-19	121348649	0.4815	T	C	121164278
cg06793505	rs1177585	8.25E-17	121325321	0.48625	C	T	121164278
cg06793505	rs4767937	1.10E-19	121223244	0.47875	G	C	121164278
cg06793505	rs471688	7.57E-17	121285424	0.48647	A	G	121164278
cg06793505	rs13746	9.71E-17	121201167	0.48824	C	T	121164278
cg06793505	rs3809313	1.11E-19	121329967	0.47992	T	C	121164278
cg06793505	rs3213570	8.40E-20	121222411	0.47766	C	G	121164278
cg06793505	rs869781	1.13E-19	121340246	0.47994	T	C	121164278
cg06793505	rs3213572	1.05E-19	121205078	0.4783	G	A	121164278
cg06793505	rs3883901	9.25E-20	121256520	0.47936	T	C	121164278
cg06793505	rs12822123	1.04E-19	121298644	0.47986	C	T	121164278
cg06793505	rs909053	1.07E-19	121237668	0.47858	G	A	121164278
cg06793505	rs2047568	9.31E-20	121243790	0.47938	G	A	121164278
cg06793505	rs494632	1.51E-19	121189116	0.47928	C	T	121164278
cg06793505	rs3213566	1.14E-14	121222578	0.46062	T	C	121164278
cg06793505	rs3897746	1.11E-19	121324727	0.47994	A	G	121164278
cg06793505	rs2062507	5.35E-16	121259051	0.38659	C	T	121164278
cg06793505	rs3901854	8.40E-20	121225526	0.47753	C	T	121164278
cg06793505	rs9431	1.12E-16	121202664	0.48786	A	C	121164278
cg06793505	rs508595	9.65E-17	121198891	0.48823	C	G	121164278
cg06793505	rs12580949	7.66E-20	121298164	0.47873	A	G	121164278
cg21721566	rs11065292	1.42E-25	121289155	0.48002	T	C	121163144
cg21721566	rs10774572	8.61E-26	121309561	0.4789	A	G	121163144
cg21721566	rs1177585	2.01E-23	121325321	0.48625	C	T	121163144
cg21721566	rs3914	2.62E-19	121174899	0.4638	T	C	121163144
cg21721566	rs1168067	2.16E-18	121082469	0.42925	C	T	121163144
cg21721566	rs1542859	7.99E-26	121336766	0.47882	A	G	121163144
cg21721566	rs2239760	8.05E-19	121163518	0.41888	C	A	121163144



cg21721566	rs525425	5.01E-26	121195625	0.4779	A	G	121163144
cg21721566	rs610578	4.44E-15	121194565	0.33012	A	G	121163144
cg21721566	rs11065324	1.31E-25	121346795	0.47883	G	C	121163144
cg21721566	rs482522	1.88E-15	121194862	0.33074	C	T	121163144
cg21721566	rs12824150	5.70E-22	121220031	0.38613	G	C	121163144
cg21721566	rs11065282	6.89E-25	121264415	0.47675	G	T	121163144
cg21721566	rs3794214	3.82E-17	121168245	0.42539	T	C	121163144
cg21721566	rs3213566	2.88E-21	121222578	0.46062	T	C	121163144
cg21721566	rs668622	6.38E-19	121198299	0.42714	G	A	121163144
cg21721566	rs11065301	4.98E-22	121304368	0.38637	C	T	121163144
cg21721566	rs674240	2.14E-15	121048935	0.36908	G	A	121163144
cg21721566	rs11065283	1.03E-25	121266453	0.47831	T	C	121163144
cg21721566	rs10774563	4.33E-23	121060780	0.45882	G	A	121163144
cg21721566	rs7137504	1.08E-25	121221628	0.47752	C	T	121163144
cg21721566	rs2062507	5.18E-22	121259051	0.38659	C	T	121163144
cg21721566	rs1168070	7.73E-20	121084654	0.42587	C	A	121163144
cg21721566	rs11065311	1.41E-25	121324115	0.47994	T	A	121163144
cg21721566	rs10431385	1.35E-18	121128882	0.41883	T	C	121163144
cg21721566	rs2001133	5.67E-25	121056754	0.49814	T	A	121163144
cg21721566	rs556001	7.36E-18	121152967	0.44425	A	G	121163144
cg21721566	rs11065300	9.32E-26	121298512	0.47881	G	C	121163144
cg21721566	rs3213570	1.08E-25	121222411	0.47766	C	G	121163144
cg21721566	rs1151849	1.78E-23	121340599	0.48635	G	A	121163144
cg21721566	rs11065202	4.86E-18	121112429	0.42711	T	C	121163144
cg21721566	rs509152	2.27E-18	121186549	0.43951	C	G	121163144
cg21721566	rs4767935	4.61E-22	121216531	0.38657	T	C	121163144
cg21721566	rs584001	6.94E-17	121228055	0.40535	G	T	121163144
cg21721566	rs520753	9.29E-24	121298993	0.48661	T	A	121163144
cg21721566	rs1696357	1.09E-24	121348649	0.4815	T	C	121163144
cg21721566	rs13746	9.06E-24	121201167	0.48824	C	T	121163144
cg21721566	rs522632	2.22E-19	121088886	0.41427	G	C	121163144
cg21721566	rs555379	1.75E-19	121090498	0.41397	C	A	121163144
cg21721566	rs12580949	9.56E-26	121298164	0.47873	A	G	121163144
cg21721566	rs2005455	1.35E-18	121128699	0.41869	A	G	121163144
cg21721566	rs3829290	8.41E-18	121126438	0.46357	T	C	121163144
cg21721566	rs558275	3.62E-17	121196891	0.40505	A	G	121163144
cg21721566	rs2708081	2.46E-14	121463288	0.48605	T	C	121163144
cg21721566	rs4454799	8.08E-26	121317075	0.47891	G	T	121163144
cg21721566	rs3883901	1.85E-25	121256520	0.47936	T	C	121163144
cg21721566	rs2859263	8.67E-18	121072799	0.42399	C	T	121163144
cg21721566	rs6553	8.49E-17	121202362	0.4051	T	C	121163144

---

cg21721566	rs9204	5.57E-18	121177778	0.3488	A	G	121163144
cg21721566	rs1151862	1.87E-23	121331285	0.48632	G	A	121163144
cg21721566	rs11611087	3.61E-25	121267126	0.48125	C	T	121163144
cg21721566	rs504403	1.20E-17	121133037	0.42635	G	C	121163144
cg21721566	rs4767938	5.66E-18	121261162	0.36469	C	T	121163144
cg21721566	rs3901854	1.08E-25	121225526	0.47753	C	T	121163144
cg21721566	rs2393717	2.88E-21	121220375	0.46065	G	C	121163144
cg21721566	rs12822123	1.60E-25	121298644	0.47986	C	T	121163144
cg21721566	rs4767918	4.94E-15	121053044	0.39216	T	C	121163144
cg21721566	rs10849807	1.42E-25	121314056	0.47997	T	C	121163144
cg21721566	rs2047568	2.18E-25	121243790	0.47938	G	A	121163144
cg21721566	rs3213572	1.93E-25	121205078	0.4783	G	A	121163144
cg21721566	rs531782	2.68E-23	121273247	0.48666	T	C	121163144
cg21721566	rs494632	1.12E-25	121189116	0.47928	C	T	121163144
cg21721566	rs17847	8.39E-18	121149475	0.46543	A	G	121163144
cg21721566	rs3897746	1.41E-25	121324727	0.47994	A	G	121163144
cg21721566	rs661647	1.52E-23	121282659	0.48537	C	T	121163144
cg21721566	rs3809313	1.39E-25	121329967	0.47992	T	C	121163144
cg21721566	rs7976497	5.84E-18	121135467	0.46449	T	C	121163144
cg21721566	rs610694	2.28E-23	121304826	0.48619	T	C	121163144
cg21721566	rs508595	9.50E-24	121198891	0.48823	C	G	121163144
cg21721566	rs2686552	4.33E-15	121075426	0.3732	C	T	121163144
cg21721566	rs11065286	9.67E-20	121271734	0.43872	T	C	121163144
cg21721566	rs1151851	5.10E-17	121340139	0.40487	A	T	121163144
cg21721566	rs869781	1.38E-25	121340246	0.47994	T	C	121163144
cg21721566	rs4767937	1.86E-25	121223244	0.47875	G	C	121163144
cg21721566	rs7964786	9.35E-17	121132100	0.36557	C	T	121163144
cg21721566	rs909053	1.86E-25	121237668	0.47858	G	A	121163144
cg21721566	rs471688	2.69E-23	121285424	0.48647	A	G	121163144
cg21721566	rs473121	2.13E-19	121089739	0.41381	T	C	121163144
cg21721566	rs4766975	2.28E-14	121069126	0.36972	G	A	121163144
cg21721566	rs625228	6.63E-19	121278266	0.42616	G	A	121163144
cg21721566	rs10849791	1.06E-25	121235280	0.47753	T	C	121163144
cg21721566	rs532703	7.37E-17	121273143	0.40521	C	G	121163144
cg21721566	rs3794215	3.83E-17	121168083	0.42539	T	C	121163144
cg21721566	rs9431	8.46E-24	121202664	0.48786	A	C	121163144
cg21721566	rs558314	7.02E-18	121171803	0.42718	C	G	121163144
cg21721566	rs696337	5.55E-18	121159380	0.46329	T	C	121163144
cg21721566	rs594507	7.29E-17	121280522	0.40518	A	G	121163144
cg21721566	rs513175	4.54E-16	121086234	0.37259	A	C	121163144
cg21721566	rs526007	5.67E-16	121085884	0.37264	A	G	121163144

cg21892295	rs1039302	8.49E-15	121236258	0.15562	C	T	121157589
cg21892295	rs610578	1.35E-21	121194565	0.33012	A	G	121157589
cg21892295	rs12580949	2.20E-41	121298164	0.47873	A	G	121157589
cg21892295	rs508595	1.37E-42	121198891	0.48823	C	G	121157589
cg21892295	rs11065300	2.23E-41	121298512	0.47881	G	C	121157589
cg21892295	rs10774563	1.19E-25	121060780	0.45882	G	A	121157589
cg21892295	rs10849791	5.70E-42	121235280	0.47753	T	C	121157589
cg21892295	rs625228	4.24E-18	121278266	0.42616	G	A	121157589
cg21892295	rs12822123	4.36E-41	121298644	0.47986	C	T	121157589
cg21892295	rs10774568	1.20E-14	121239696	0.15577	A	G	121157589
cg21892295	rs10774572	1.93E-41	121309561	0.4789	A	G	121157589
cg21892295	rs2859263	3.97E-19	121072799	0.42399	C	T	121157589
cg21892295	rs11065311	3.44E-41	121324115	0.47994	T	A	121157589
cg21892295	rs661647	4.64E-42	121282659	0.48537	C	T	121157589
cg21892295	rs10849778	2.93E-18	121135508	0.13239	G	A	121157589
cg21892295	rs10849786	1.06E-14	121209302	0.15656	A	G	121157589
cg21892295	rs11611087	3.78E-41	121267126	0.48125	C	T	121157589
cg21892295	rs11065360	6.60E-18	121386532	0.38257	A	G	121157589
cg21892295	rs4767938	4.12E-23	121261162	0.36469	C	T	121157589
cg21892295	rs11065301	8.92E-34	121304368	0.38637	C	T	121157589
cg21892295	rs10849807	3.51E-41	121314056	0.47997	T	C	121157589
cg21892295	rs1542859	1.87E-41	121336766	0.47882	A	G	121157589
cg21892295	rs1696357	6.49E-40	121348649	0.4815	T	C	121157589
cg21892295	rs1151851	1.49E-19	121340139	0.40487	A	T	121157589
cg21892295	rs494632	3.68E-41	121189116	0.47928	C	T	121157589
cg21892295	rs11065292	4.14E-41	121289155	0.48002	T	C	121157589
cg21892295	rs11065259	2.96E-14	121205604	0.15695	T	C	121157589
cg21892295	rs1151862	6.92E-42	121331285	0.48632	G	A	121157589
cg21892295	rs11065282	1.68E-41	121264415	0.47675	G	T	121157589
cg21892295	rs9431	1.15E-42	121202664	0.48786	A	C	121157589
cg21892295	rs2001133	1.14E-24	121056754	0.49814	T	A	121157589
cg21892295	rs2393717	3.96E-18	121220375	0.46065	G	C	121157589
cg21892295	rs3213570	5.72E-42	121222411	0.47766	C	G	121157589
cg21892295	rs11065283	9.70E-42	121266453	0.47831	T	C	121157589
cg21892295	rs1151849	6.95E-42	121340599	0.48635	G	A	121157589
cg21892295	rs2047568	1.13E-41	121243790	0.47938	G	A	121157589
cg21892295	rs11065286	8.39E-20	121271734	0.43872	T	C	121157589
cg21892295	rs12824150	2.76E-34	121220031	0.38613	G	C	121157589
cg21892295	rs6553	1.34E-19	121202362	0.4051	T	C	121157589
cg21892295	rs594507	1.82E-19	121280522	0.40518	A	G	121157589
cg21892295	rs2062507	2.66E-34	121259051	0.38659	C	T	121157589

---

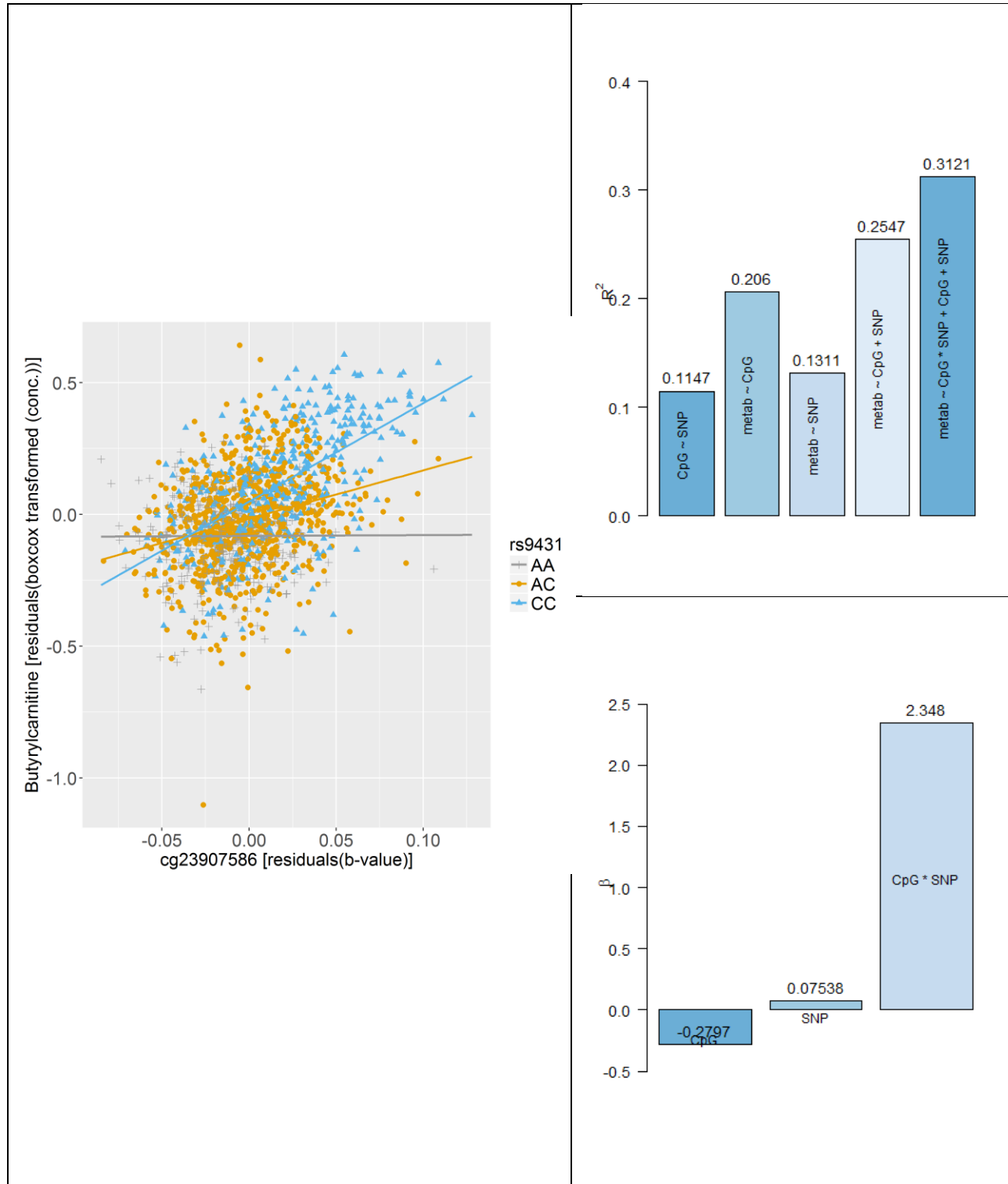
cg21892295	rs3213566	3.96E-18	121222578	0.46062	T	C	121157589
cg21892295	rs11065324	5.30E-41	121346795	0.47883	G	C	121157589
cg21892295	rs4767937	1.11E-41	121223244	0.47875	G	C	121157589
cg21892295	rs13746	1.35E-42	121201167	0.48824	C	T	121157589
cg21892295	rs2708081	8.19E-22	121463288	0.48605	T	C	121157589
cg21892295	rs695950	1.10E-14	121175560	0.10176	G	C	121157589
cg21892295	rs1168067	6.81E-20	121082469	0.42925	C	T	121157589
cg21892295	rs3213572	1.15E-41	121205078	0.4783	G	A	121157589
cg21892295	rs7965649	1.37E-18	121084678	0.16728	C	T	121157589
cg21892295	rs531782	4.68E-42	121273247	0.48666	T	C	121157589
cg21892295	rs4454799	1.75E-41	121317075	0.47891	G	T	121157589
cg21892295	rs471688	8.95E-42	121285424	0.48647	A	G	121157589
cg21892295	rs4767941	6.77E-16	121359586	0.36177	G	A	121157589
cg21892295	rs532703	1.75E-19	121273143	0.40521	C	G	121157589
cg21892295	rs3809313	3.58E-41	121329967	0.47992	T	C	121157589
cg21892295	rs610694	8.31E-42	121304826	0.48619	T	C	121157589
cg21892295	rs1177585	6.90E-42	121325321	0.48625	C	T	121157589
cg21892295	rs4767935	2.11E-34	121216531	0.38657	T	C	121157589
cg21892295	rs3901854	5.72E-42	121225526	0.47753	C	T	121157589
cg21892295	rs584001	1.57E-19	121228055	0.40535	G	T	121157589
cg21892295	rs668622	2.24E-18	121198299	0.42714	G	A	121157589
cg21892295	rs520753	6.24E-42	121298993	0.48661	T	A	121157589
cg21892295	rs1168070	4.20E-21	121084654	0.42587	C	A	121157589
cg21892295	rs7137504	5.72E-42	121221628	0.47752	C	T	121157589
cg21892295	rs909053	1.26E-41	121237668	0.47858	G	A	121157589
cg21892295	rs482522	1.53E-21	121194862	0.33074	C	T	121157589
cg21892295	rs3897746	3.43E-41	121324727	0.47994	A	G	121157589
cg21892295	rs525425	1.57E-41	121195625	0.4779	A	G	121157589
cg21892295	rs7135147	1.17E-14	121259984	0.15474	T	C	121157589
cg21892295	rs3883901	1.61E-41	121256520	0.47936	T	C	121157589
cg21892295	rs6489782	1.31E-14	121247491	0.15568	C	A	121157589
cg21892295	rs869781	3.68E-41	121340246	0.47994	T	C	121157589
cg21892295	rs558275	2.91E-19	121196891	0.40505	A	G	121157589
cg23907586	rs471688	1.67E-19	121285424	0.48647	A	G	121163367
cg23907586	rs10849791	1.56E-19	121235280	0.47753	T	C	121163367
cg23907586	rs1151862	1.77E-19	121331285	0.48632	G	A	121163367
cg23907586	rs10774572	1.39E-19	121309561	0.4789	A	G	121163367
cg23907586	rs11065324	1.81E-19	121346795	0.47883	G	C	121163367
cg23907586	rs11065300	1.43E-19	121298512	0.47881	G	C	121163367
cg23907586	rs12580949	1.37E-19	121298164	0.47873	A	G	121163367
cg23907586	rs11065283	1.31E-19	121266453	0.47831	T	C	121163367

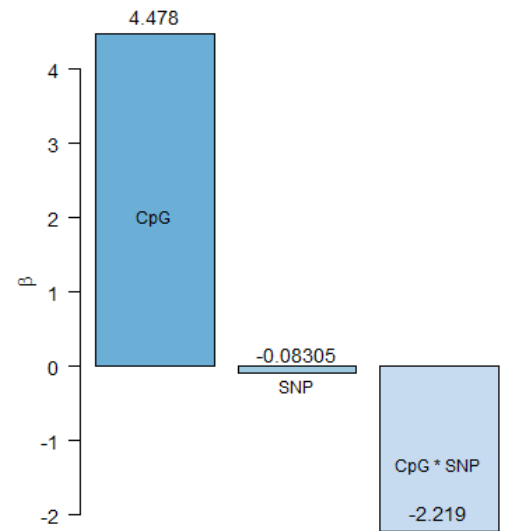
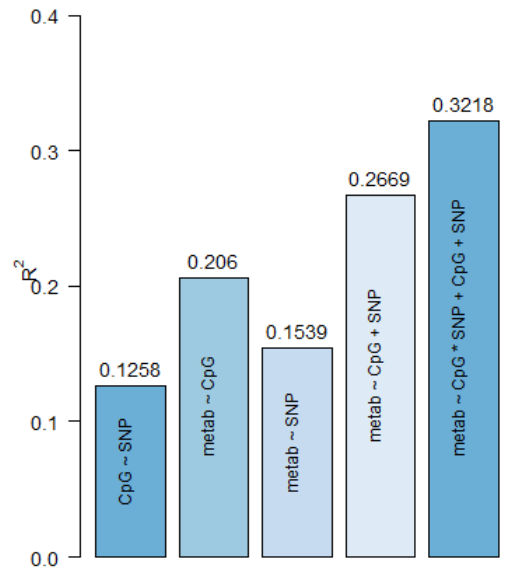
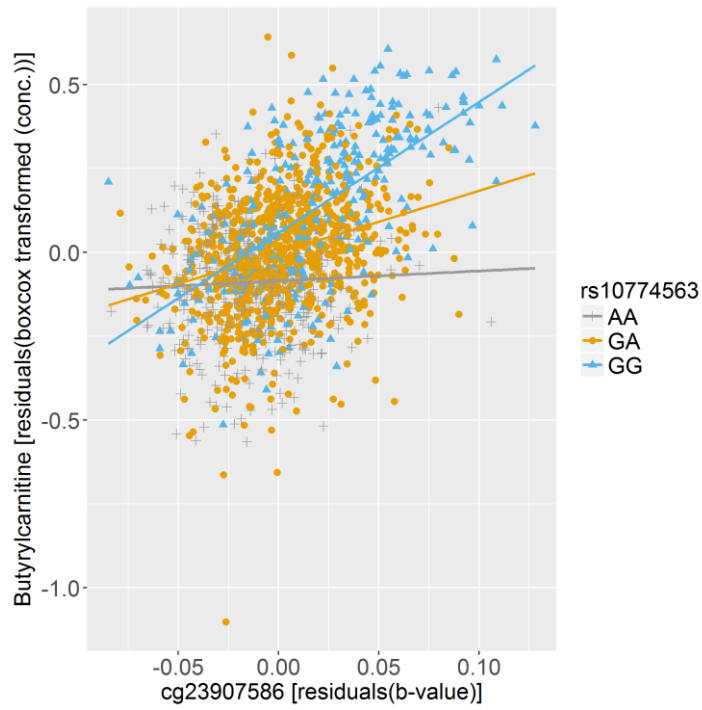
cg23907586	rs12824150	5.33E-18	121220031	0.38613	G	C	121163367
cg23907586	rs12822123	1.22E-19	121298644	0.47986	C	T	121163367
cg23907586	rs7137504	1.57E-19	121221628	0.47752	C	T	121163367
cg23907586	rs4767935	4.84E-18	121216531	0.38657	T	C	121163367
cg23907586	rs11065301	3.92E-18	121304368	0.38637	C	T	121163367
cg23907586	rs1696357	2.56E-19	121348649	0.4815	T	C	121163367
cg23907586	rs3213570	1.57E-19	121222411	0.47766	C	G	121163367
cg23907586	rs525425	6.11E-20	121195625	0.4779	A	G	121163367
cg23907586	rs2062507	4.03E-18	121259051	0.38659	C	T	121163367
cg23907586	rs473121	3.02E-15	121089739	0.41381	T	C	121163367
cg23907586	rs1151849	1.77E-19	121340599	0.48635	G	A	121163367
cg23907586	rs1168067	1.18E-16	121082469	0.42925	C	T	121163367
cg23907586	rs3901854	1.57E-19	121225526	0.47753	C	T	121163367
cg23907586	rs10849807	1.14E-19	121314056	0.47997	T	C	121163367
cg23907586	rs3213572	1.39E-19	121205078	0.4783	G	A	121163367
cg23907586	rs909053	1.37E-19	121237668	0.47858	G	A	121163367
cg23907586	rs11065311	1.13E-19	121324115	0.47994	T	A	121163367
cg23907586	rs522632	3.89E-15	121088886	0.41427	G	C	121163367
cg23907586	rs494632	8.00E-20	121189116	0.47928	C	T	121163367
cg23907586	rs508595	2.00E-19	121198891	0.48823	C	G	121163367
cg23907586	rs610578	2.73E-14	121194565	0.33012	A	G	121163367
cg23907586	rs11065282	2.24E-19	121264415	0.47675	G	T	121163367
cg23907586	rs11611087	1.74E-19	121267126	0.48125	C	T	121163367
cg23907586	rs4767938	4.06E-15	121261162	0.36469	C	T	121163367
cg23907586	rs1542859	1.47E-19	121336766	0.47882	A	G	121163367
cg23907586	rs2047568	1.19E-19	121243790	0.47938	G	A	121163367
cg23907586	rs520753	1.79E-19	121298993	0.48661	T	A	121163367
cg23907586	rs3883901	1.13E-19	121256520	0.47936	T	C	121163367
cg23907586	rs482522	3.73E-14	121194862	0.33074	C	T	121163367
cg23907586	rs661647	1.75E-19	121282659	0.48537	C	T	121163367
cg23907586	rs2859263	2.00E-16	121072799	0.42399	C	T	121163367
cg23907586	rs11065292	1.86E-19	121289155	0.48002	T	C	121163367
cg23907586	rs13746	1.99E-19	121201167	0.48824	C	T	121163367
cg23907586	rs3809313	1.21E-19	121329967	0.47992	T	C	121163367
cg23907586	rs4767937	1.33E-19	121223244	0.47875	G	C	121163367
cg23907586	rs610694	1.87E-19	121304826	0.48619	T	C	121163367
cg23907586	rs2001133	2.12E-20	121056754	0.49814	T	A	121163367
cg23907586	rs1168070	6.48E-18	121084654	0.42587	C	A	121163367
cg23907586	rs1177585	1.71E-19	121325321	0.48625	C	T	121163367
cg23907586	rs10774563	4.59E-19	121060780	0.45882	G	A	121163367
cg23907586	rs4454799	1.34E-19	121317075	0.47891	G	T	121163367

---

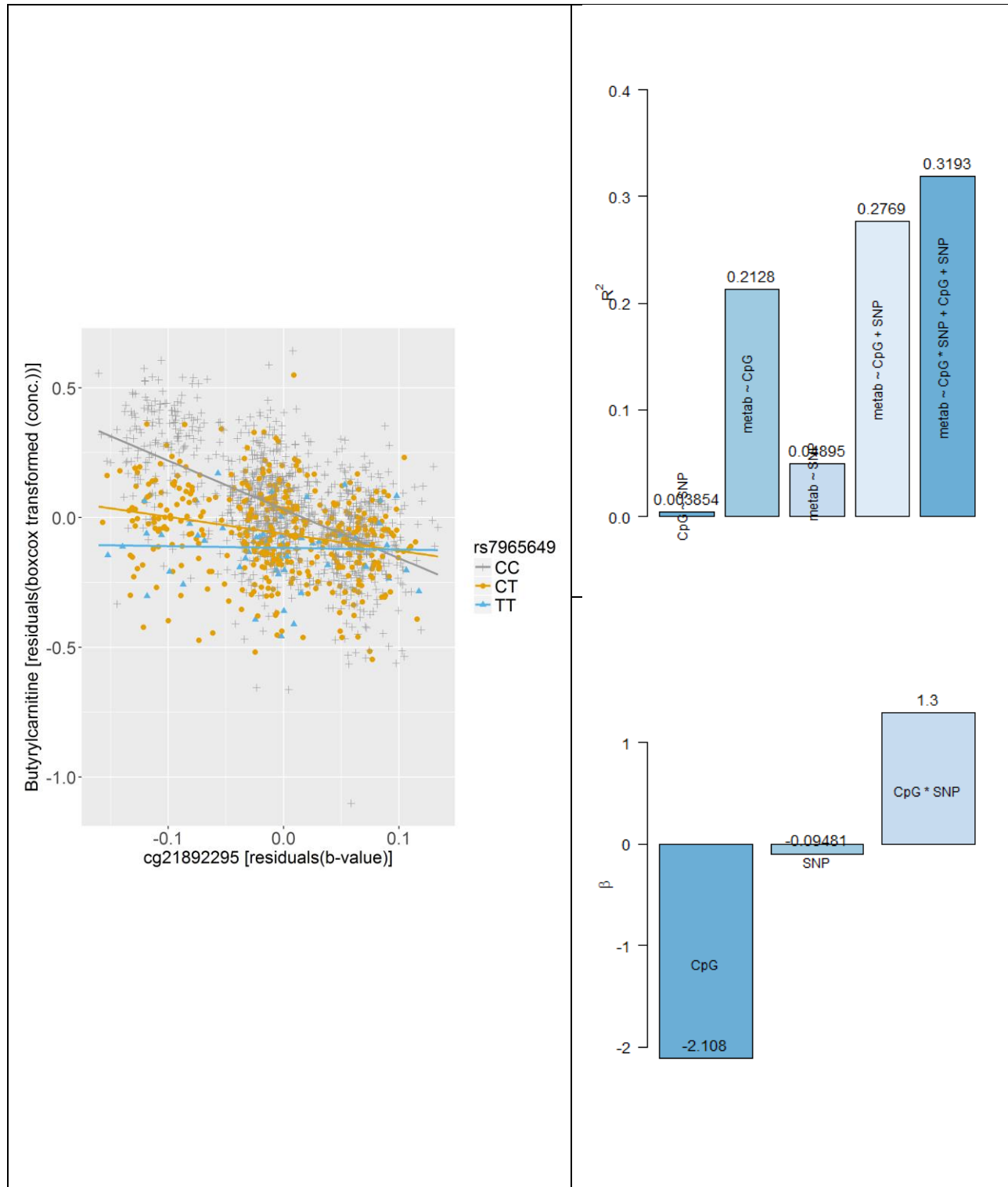
cg23907586	rs3897746	1.13E-19	121324727	0.47994	A	G	121163367
cg23907586	rs531782	1.73E-19	121273247	0.48666	T	C	121163367
cg23907586	rs555379	2.51E-15	121090498	0.41397	C	A	121163367
cg23907586	rs9431	7.64E-20	121202664	0.48786	A	C	121163367
cg23907586	rs869781	1.25E-19	121340246	0.47994	T	C	121163367

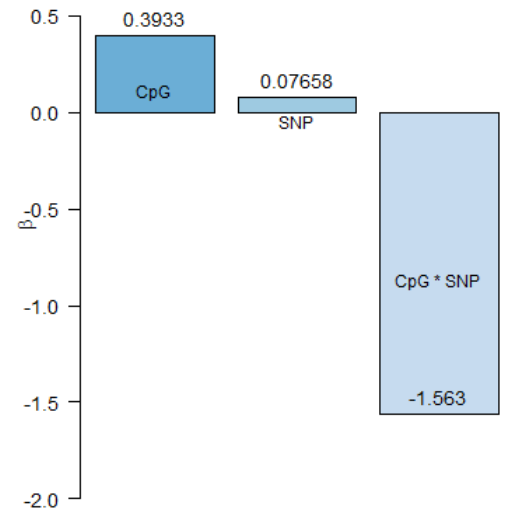
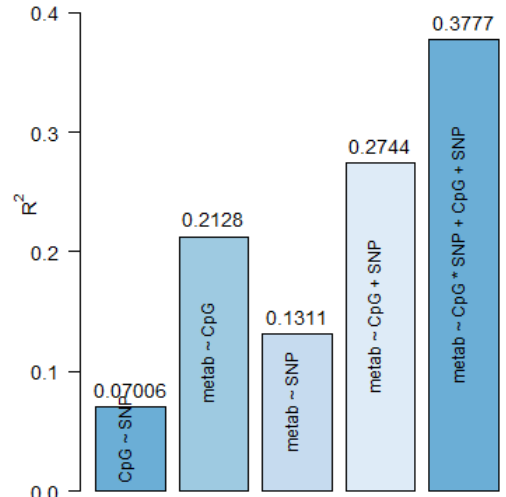
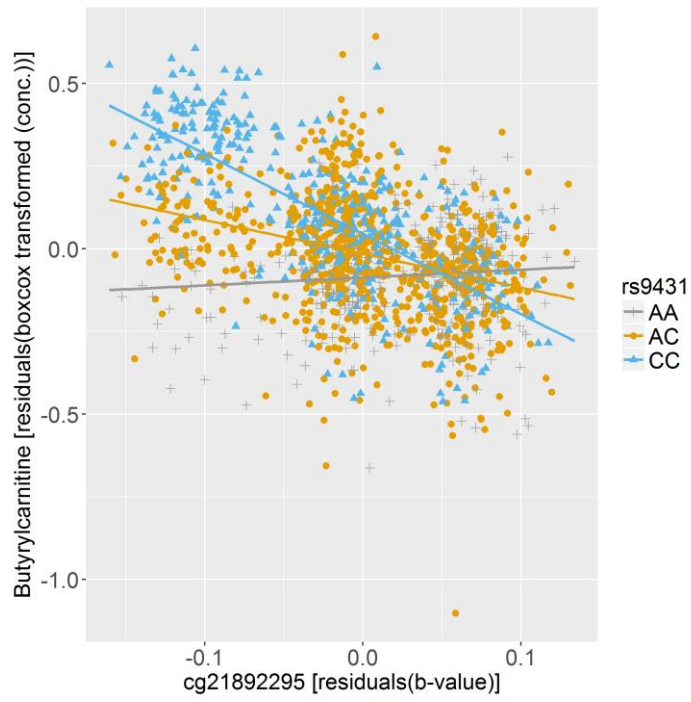
Metabolon

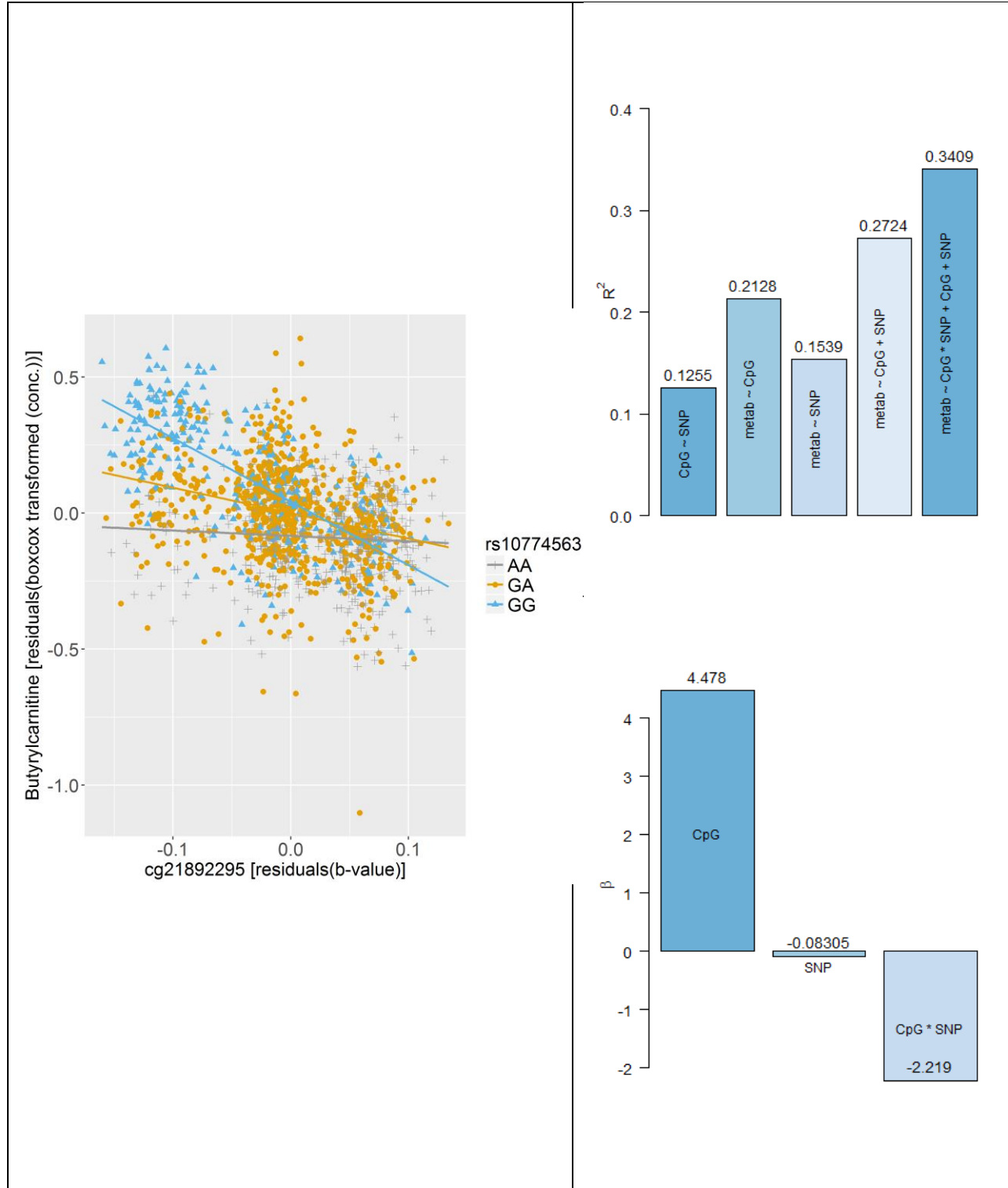


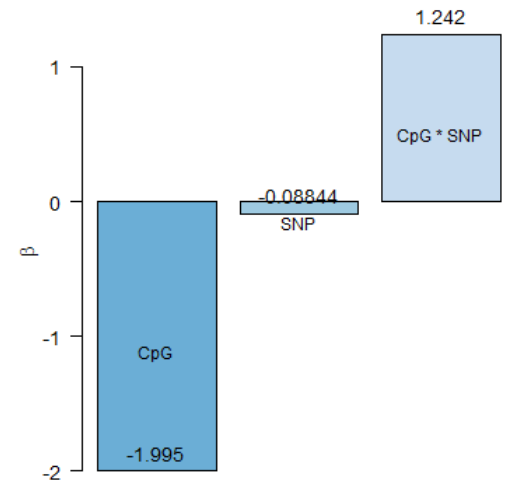
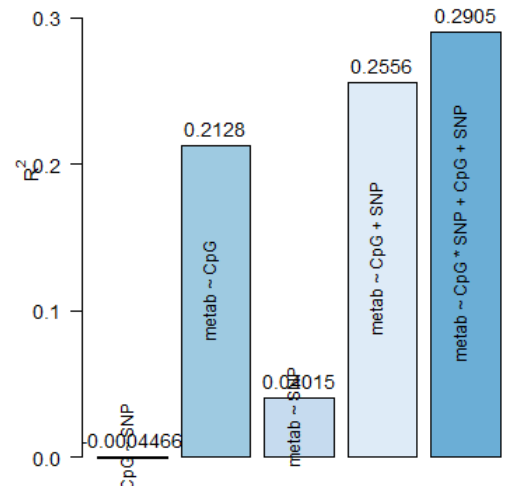
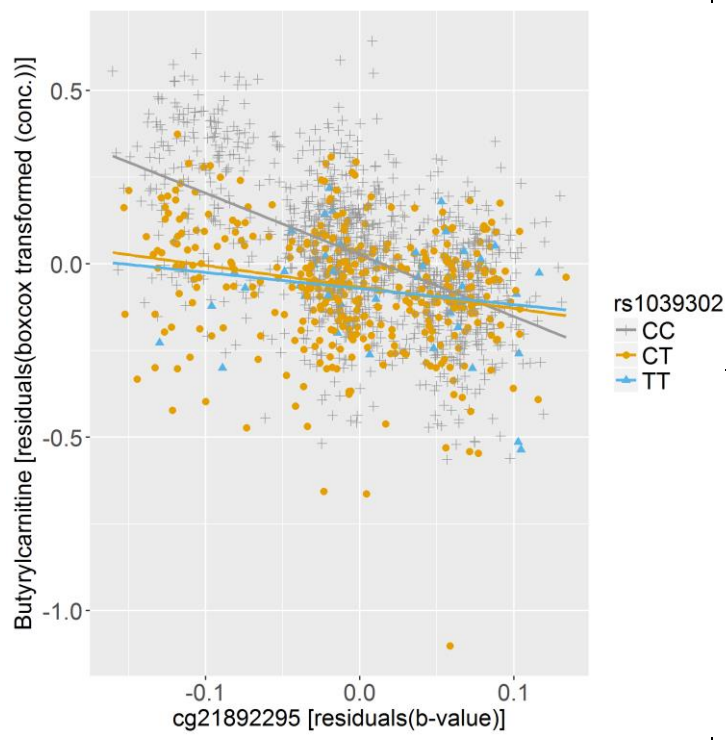


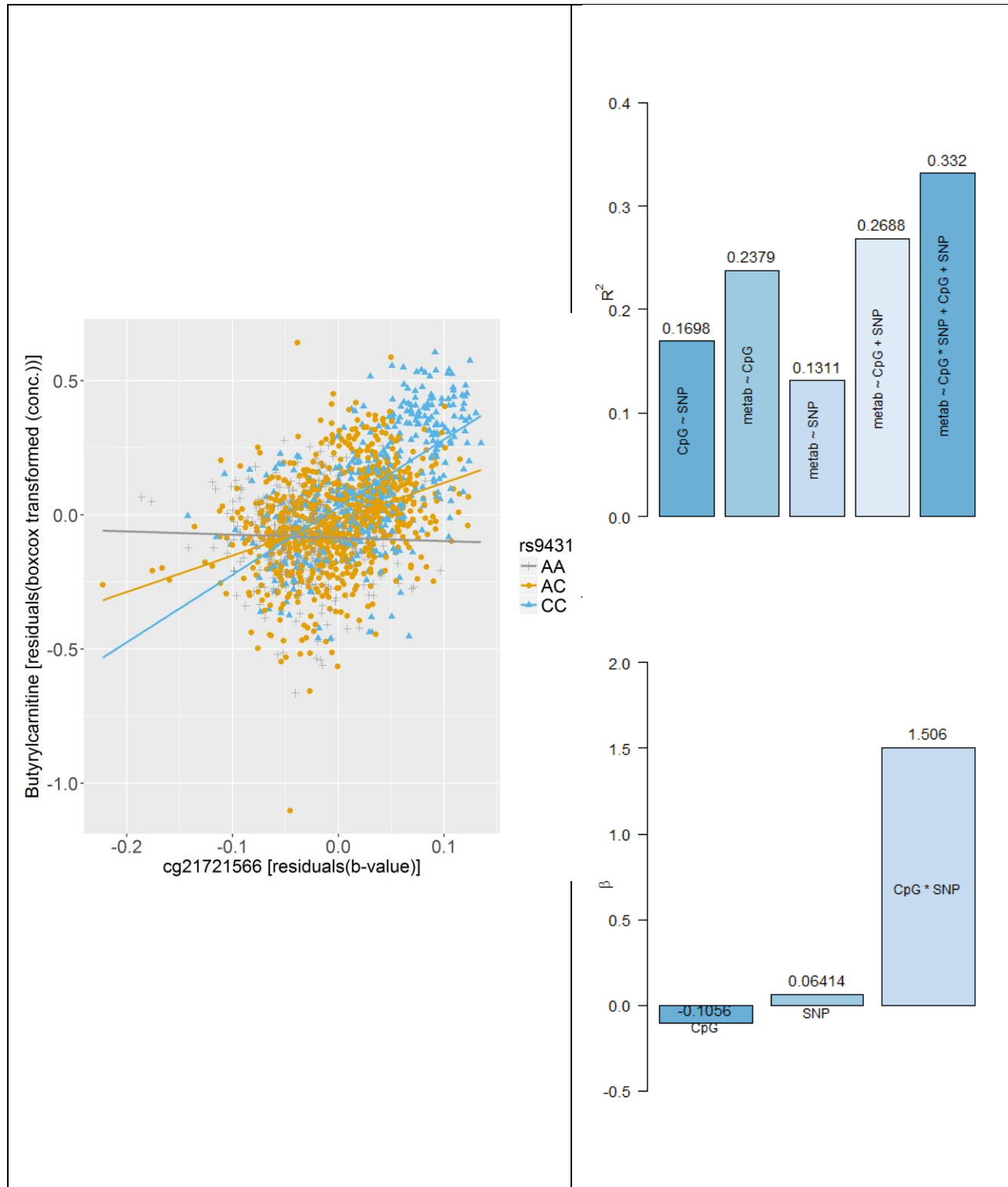


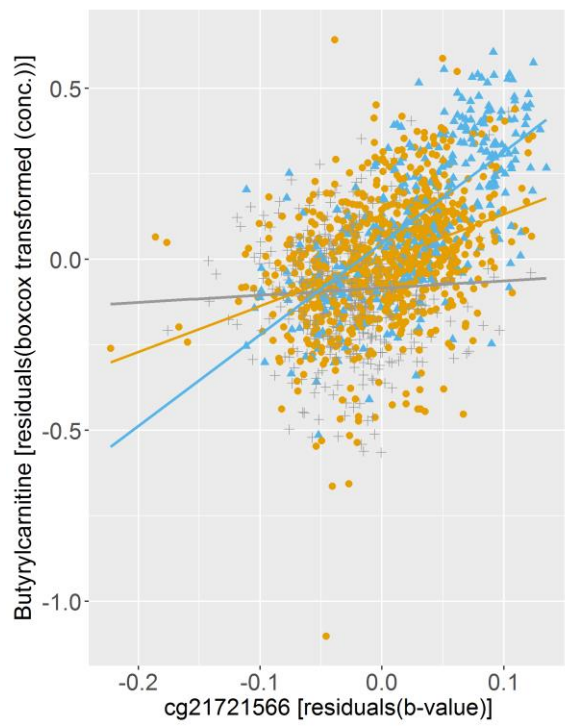




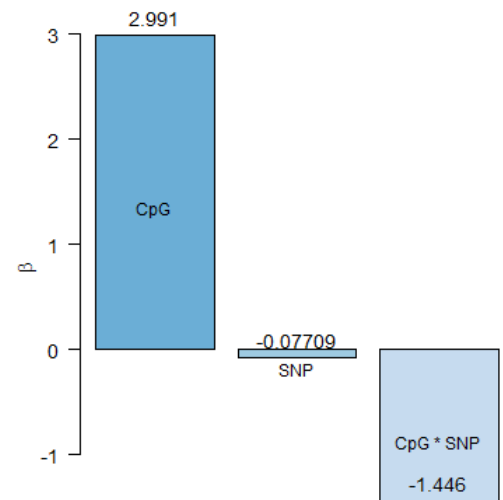
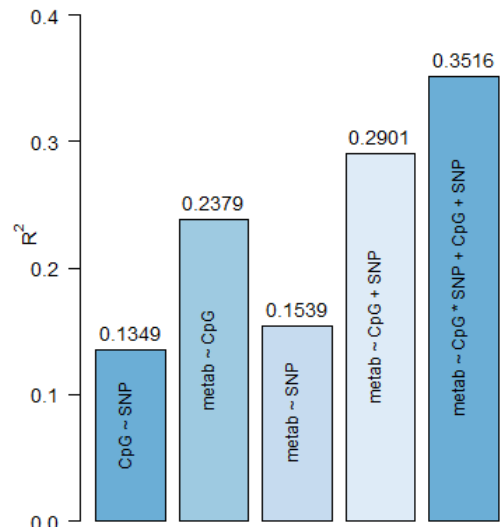


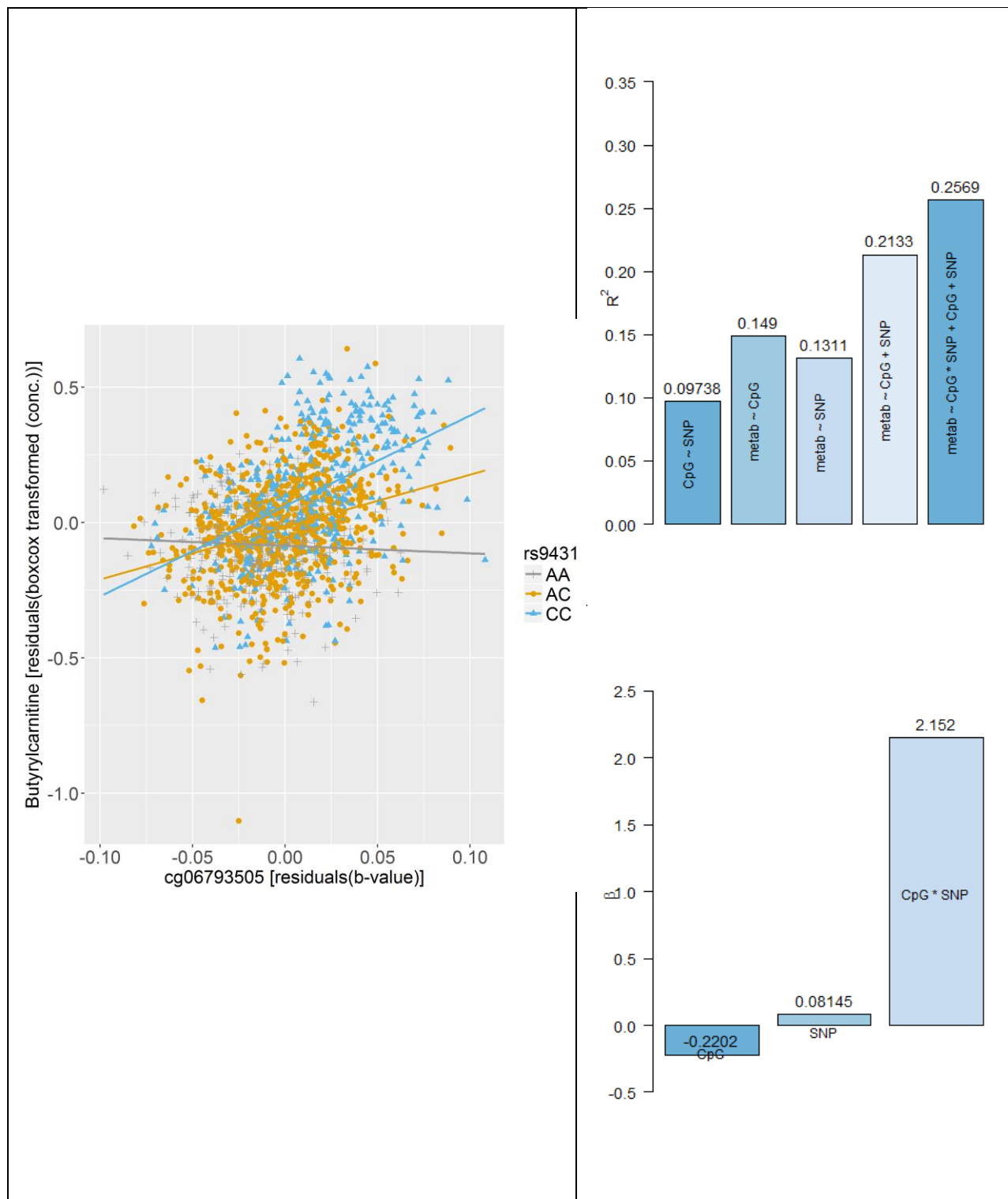


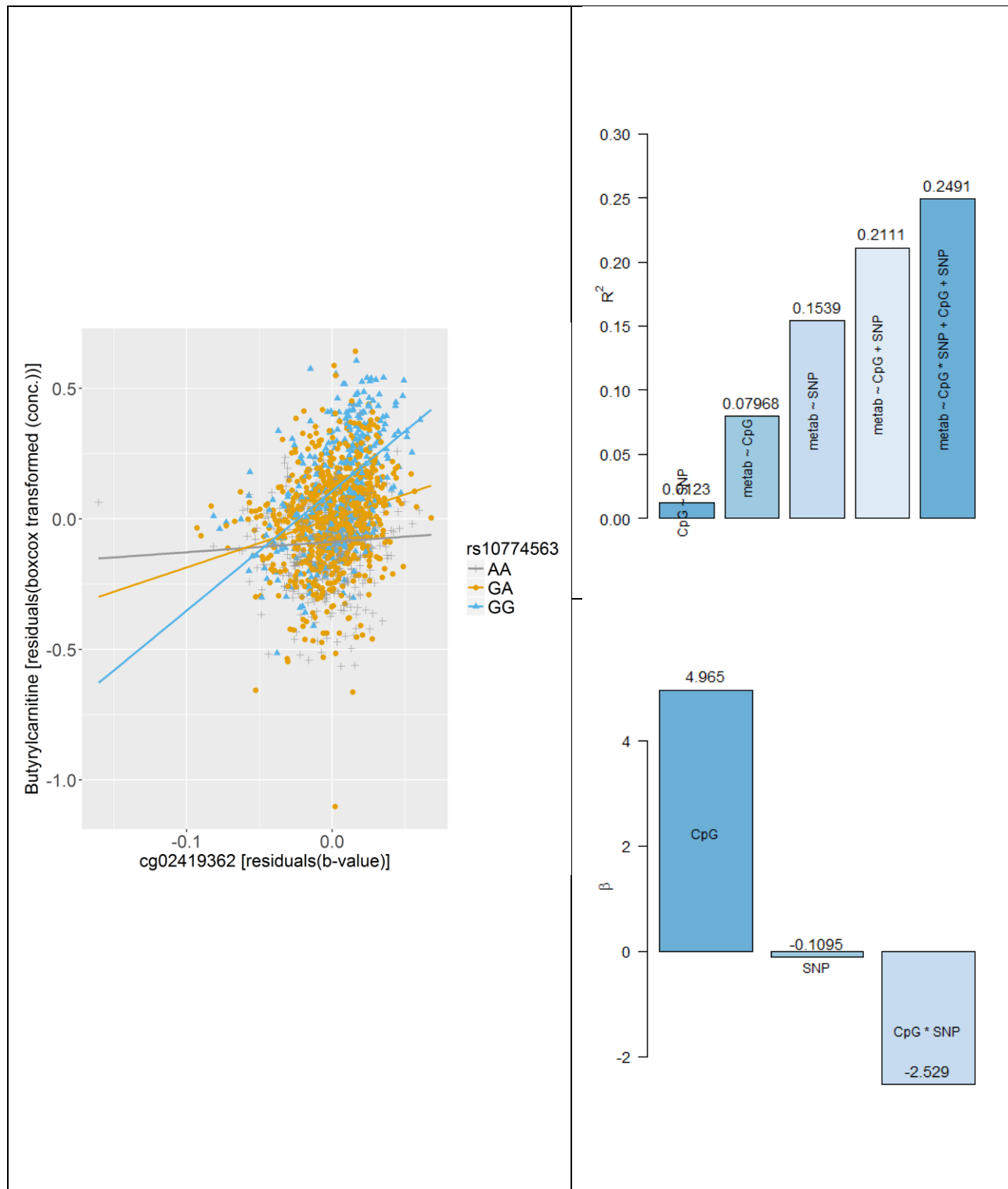




rs10774563  
 + AA  
 ● GA  
 ▲ GG



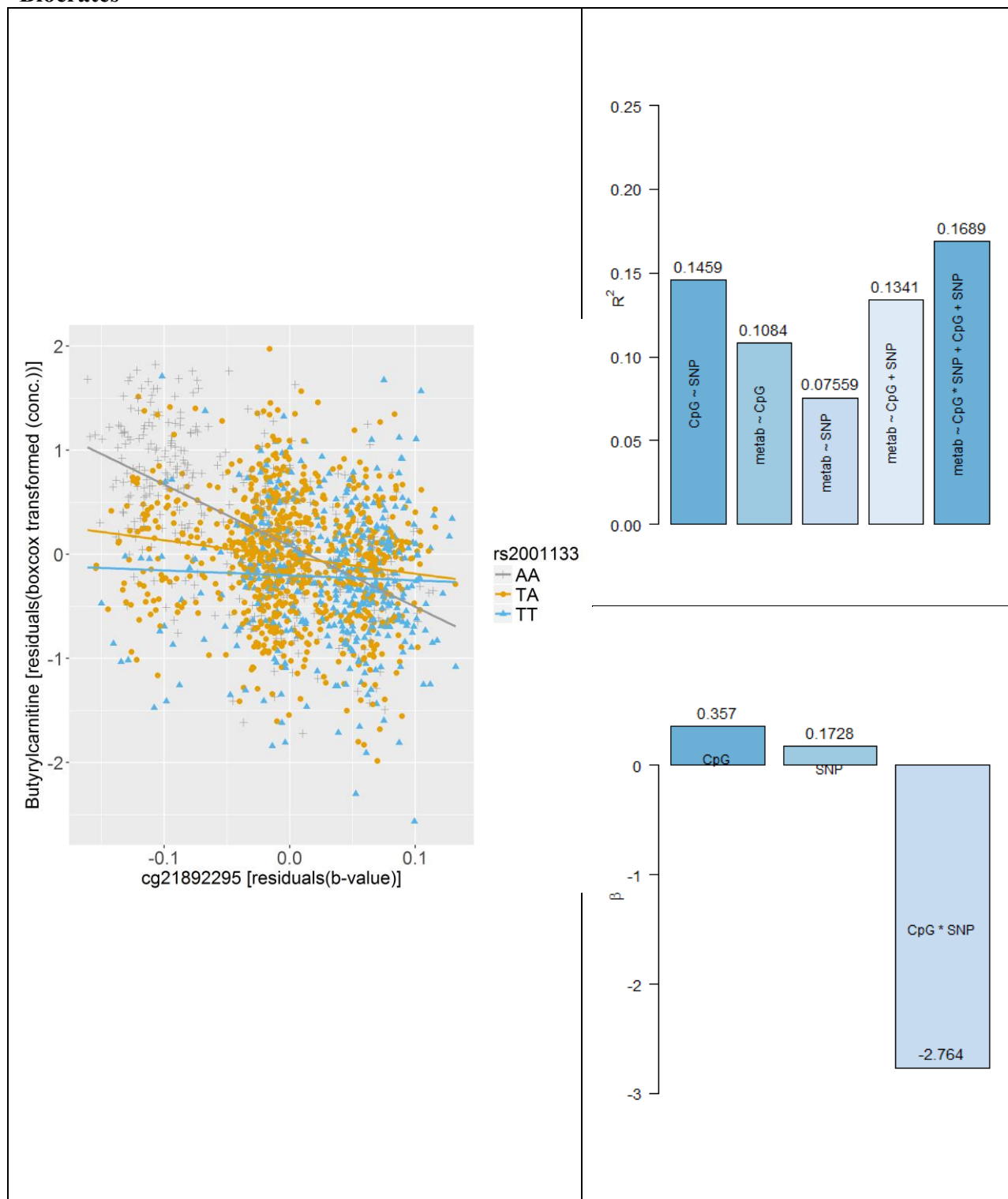


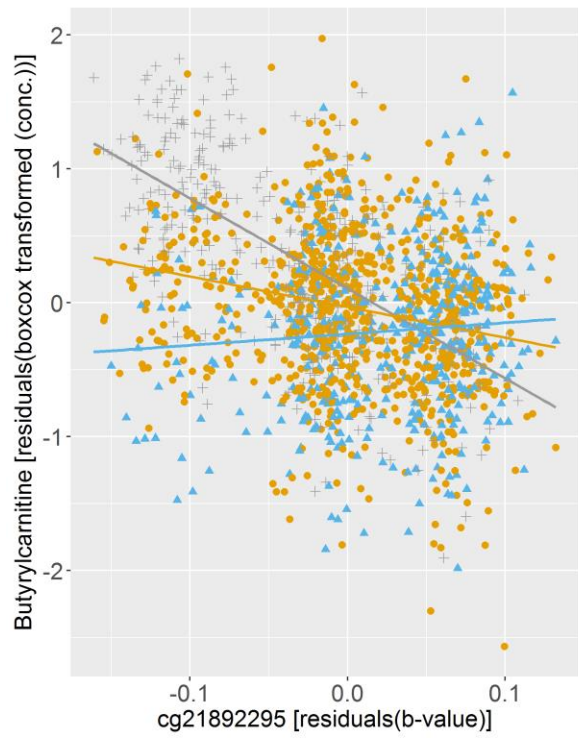


**Figure A.1:** Scatterplot of CpG site against metabolite levels of butyrylcarnitine measured in the Metabolon platform, genotypes are color-coded (left) and the improvement of the coefficient of determination  $R^2$  in the models with and without the interaction term (right) as well as the  $\beta$  coefficient of the linear regression with interaction term (right).

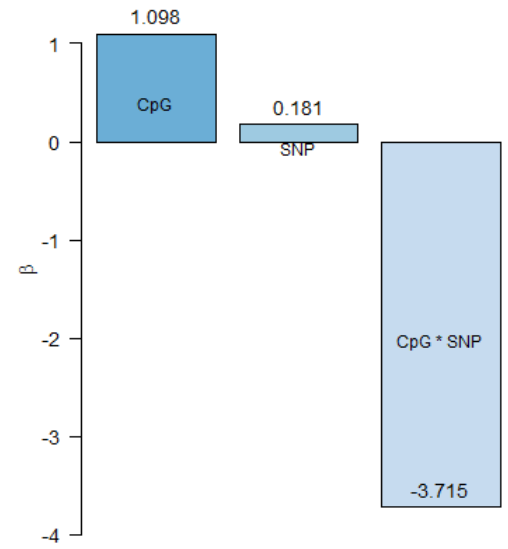
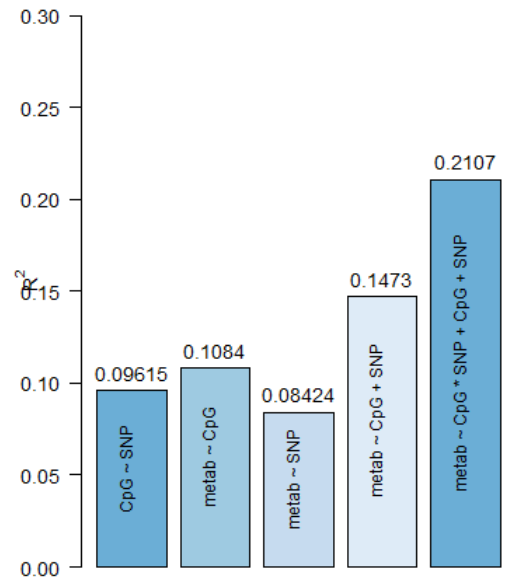


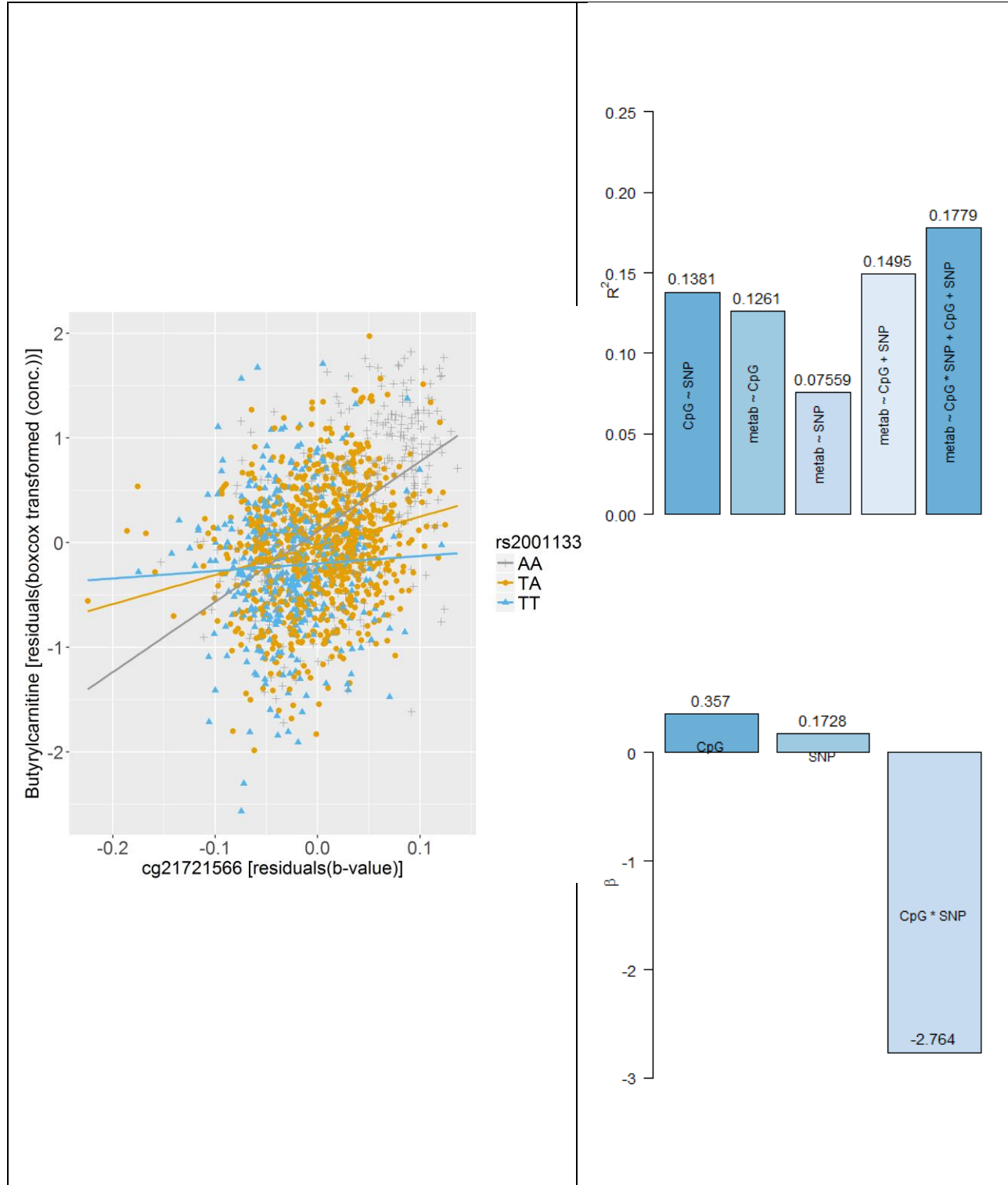
Biocrates

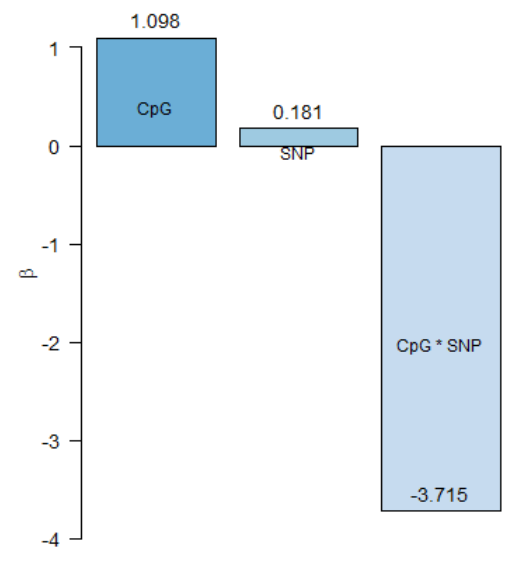
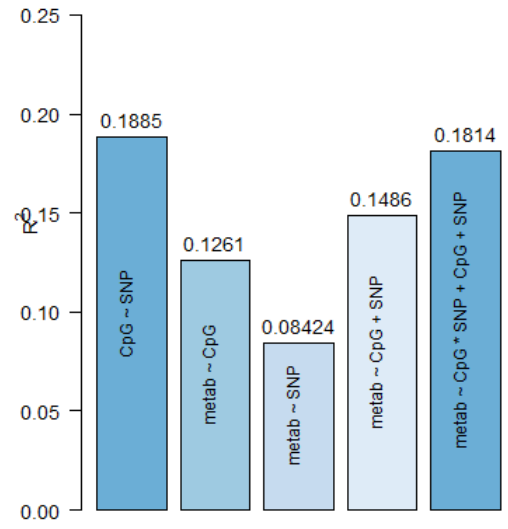
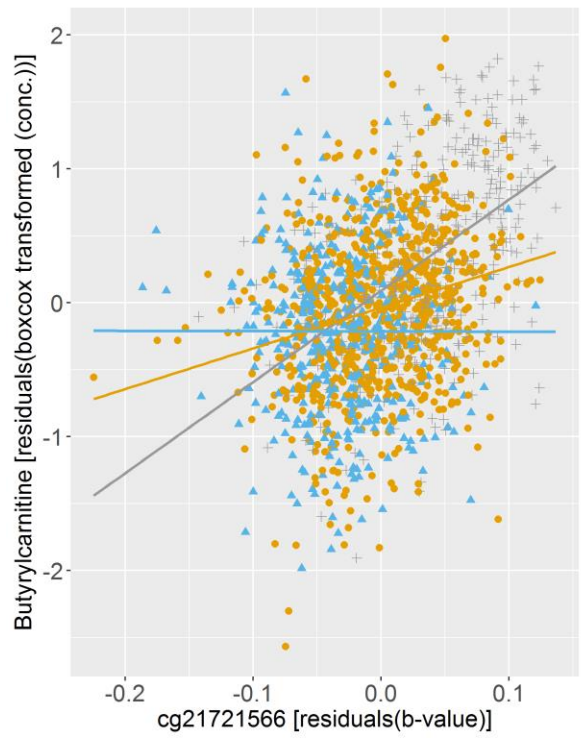


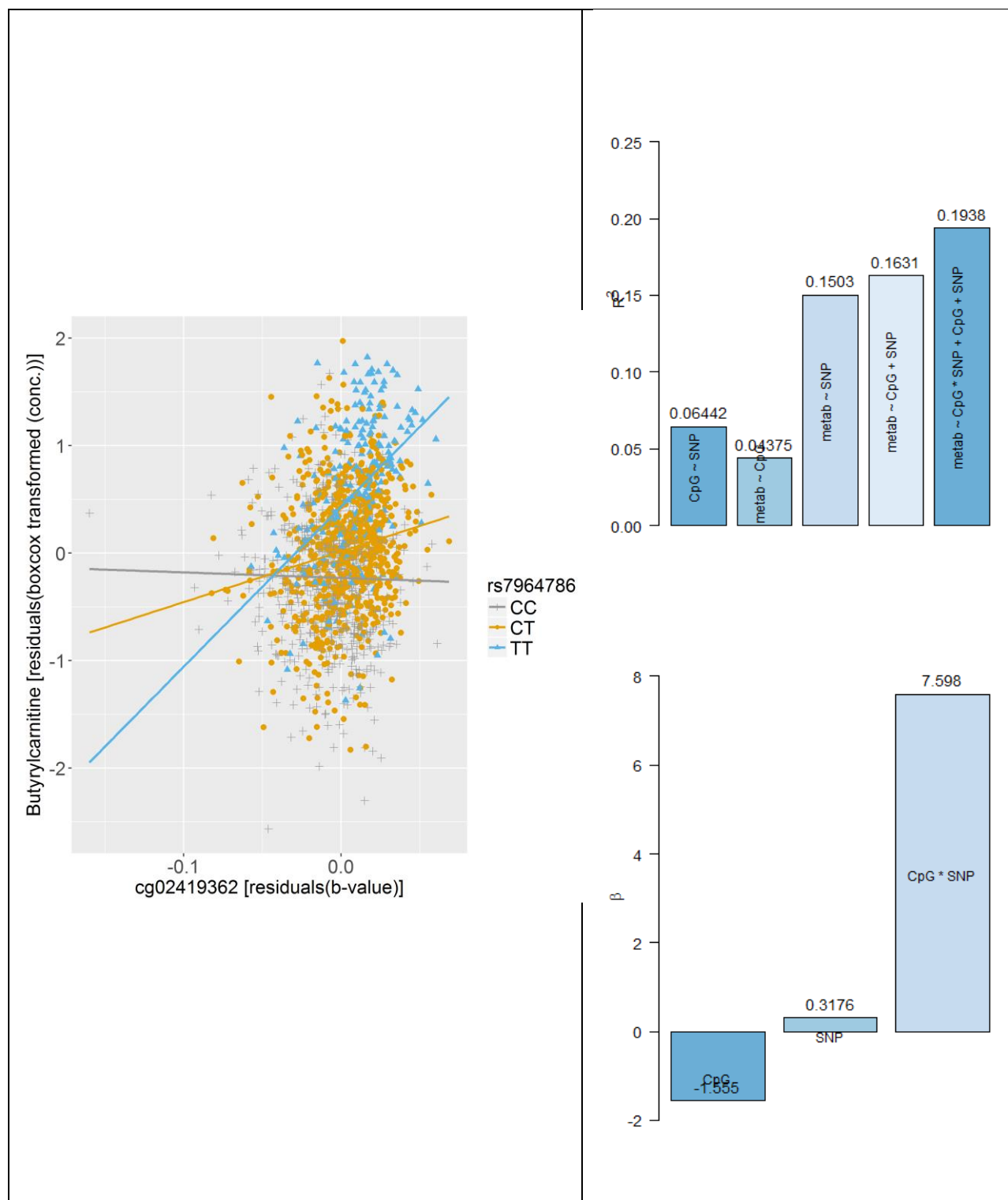


rs11065283  
 + CC  
 ● TC  
 ▲ TT





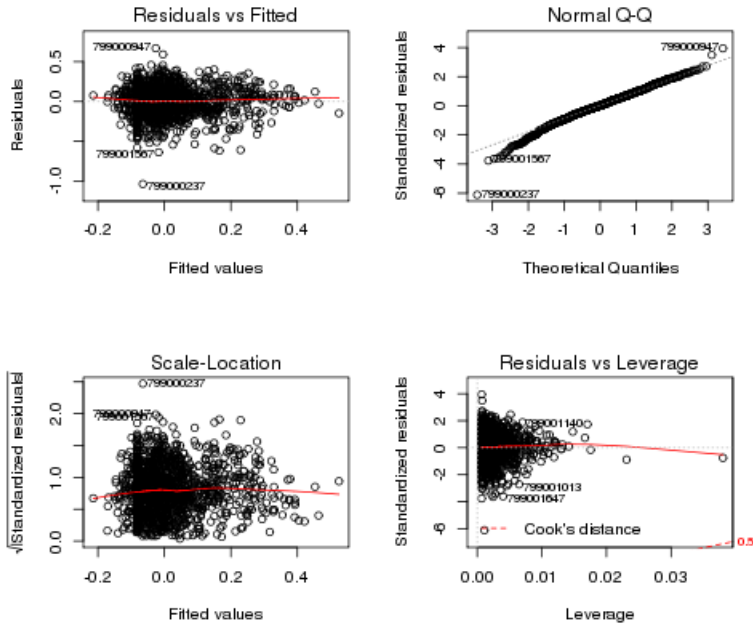




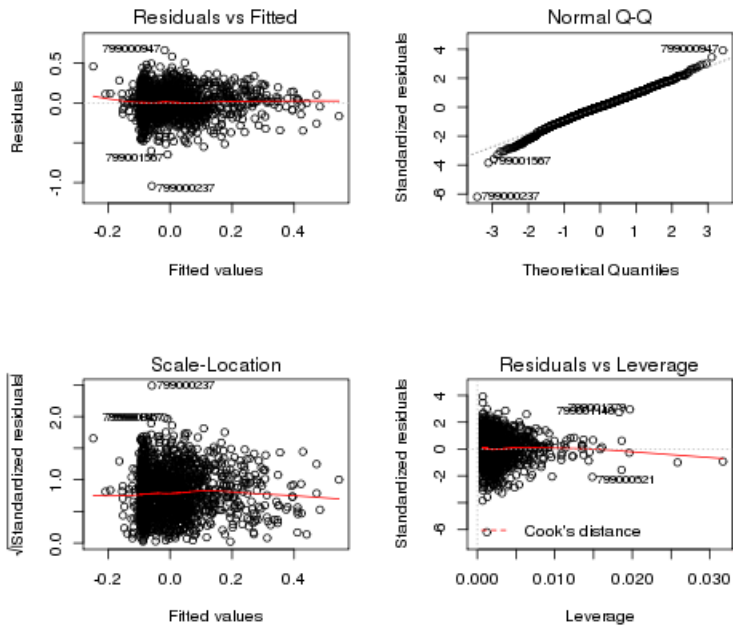
**Figure A.2:** Scatterplot of CpG sites against metabolite levels of butyrylcarnitine measured in the Biocrates platform, genotypes are color-coded (left) and the improvement of the coefficient of determination  $R^2$  in the models with and without the interaction term (right).

# Metabolon

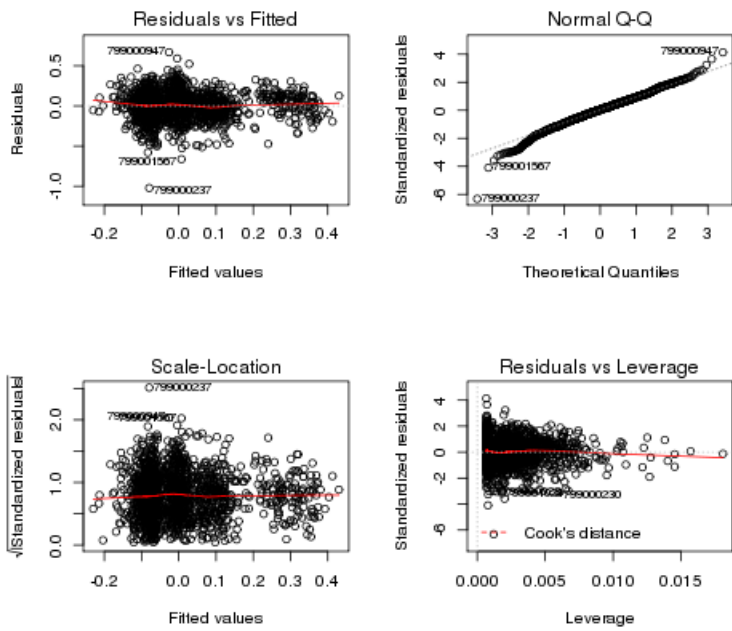
Model:  $C4 \sim rs9431 * cg23907586 + rs9431 + cg23907586$



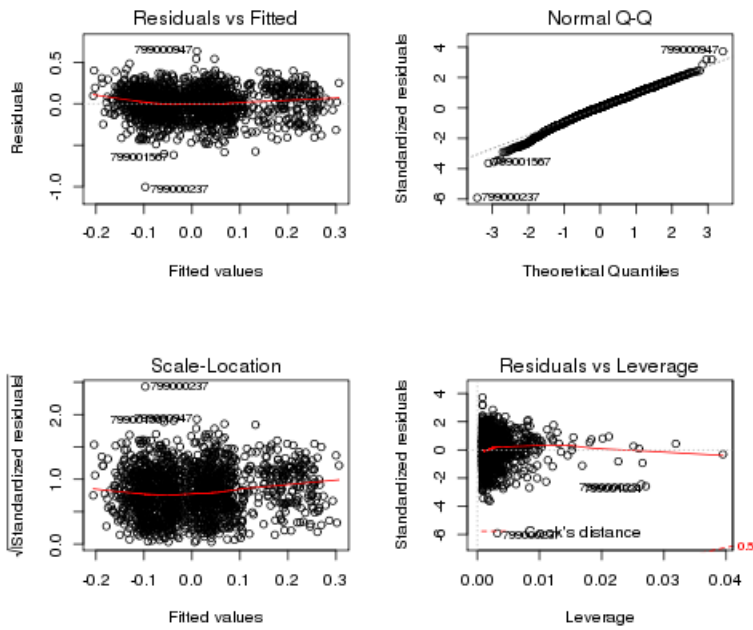
Model:  $C4 \sim rs10774563 * cg23907586 + rs10774563 + cg23907586$



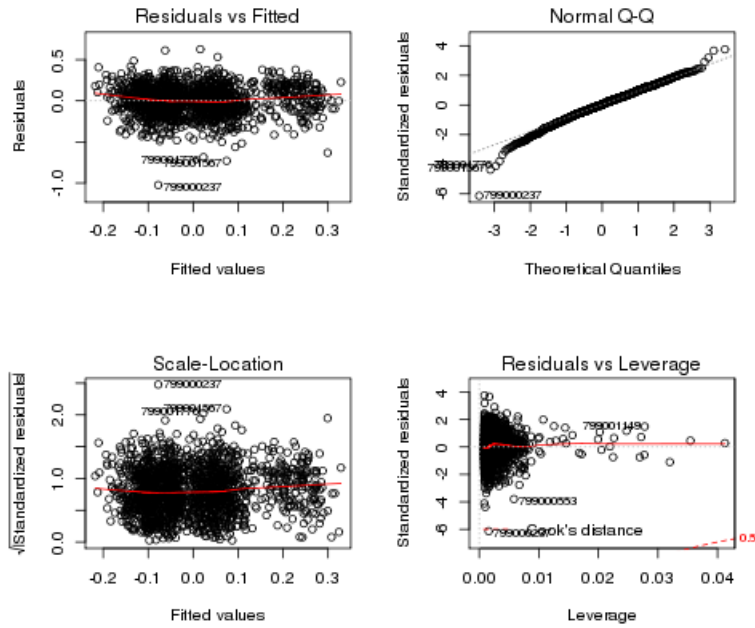
Model: C4 ~ rs9431 \* cg21892295 + rs9431 + cg21892295



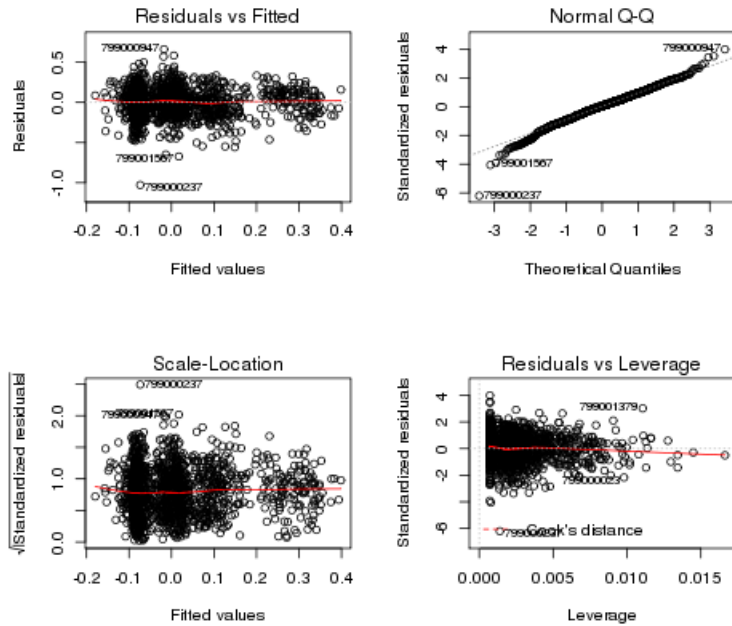
Model: C4 ~ rs1039302 \* cg21892295 + rs1039302 + cg21892295



Model: C4 ~ rs7965649 \* cg21892295 + rs7965649 + cg21892295

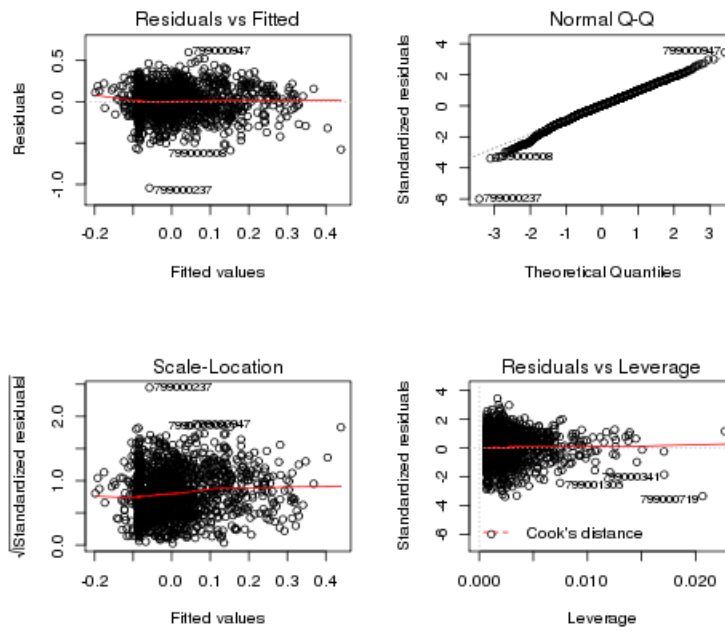


Model: C4 ~ rs10774563 \* cg21892295 + rs10774563 + cg21892295

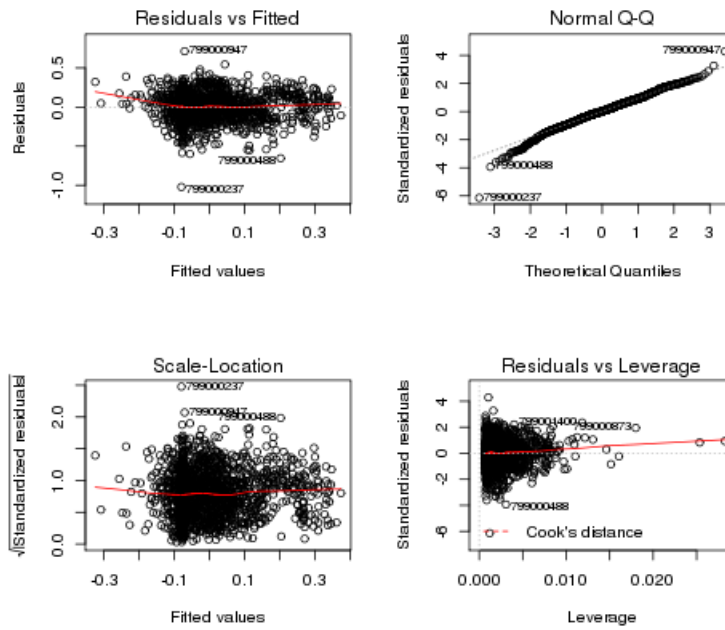




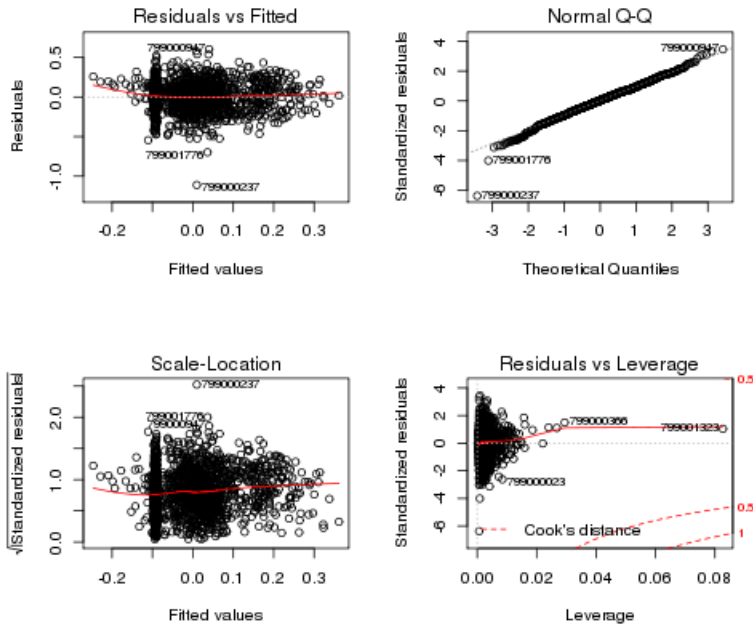
Model: C4 ~ rs9431 \* cg06793505 + rs9431 + cg06793505



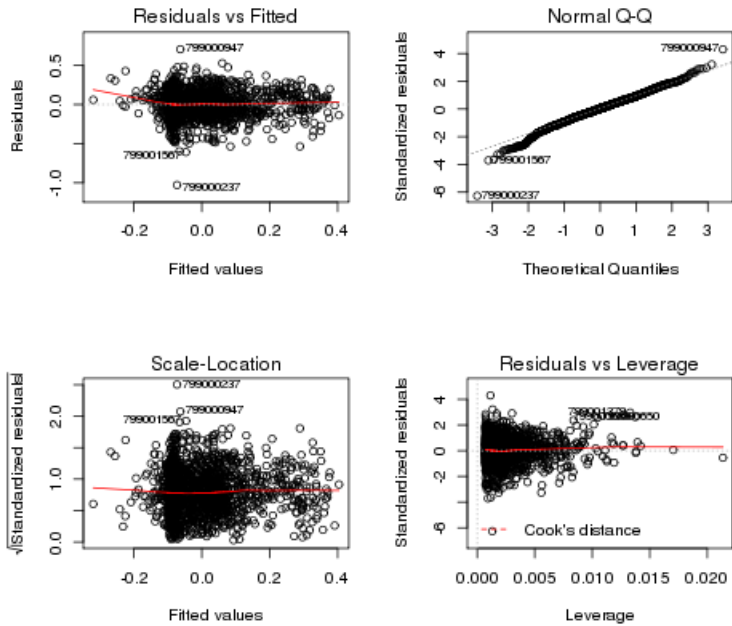
Model: C4 ~ rs9431 \* cg21721566 + rs9431 + cg21721566



Model: C4 ~ rs10774563 \* cg02419362 + rs10774563 + cg02419362



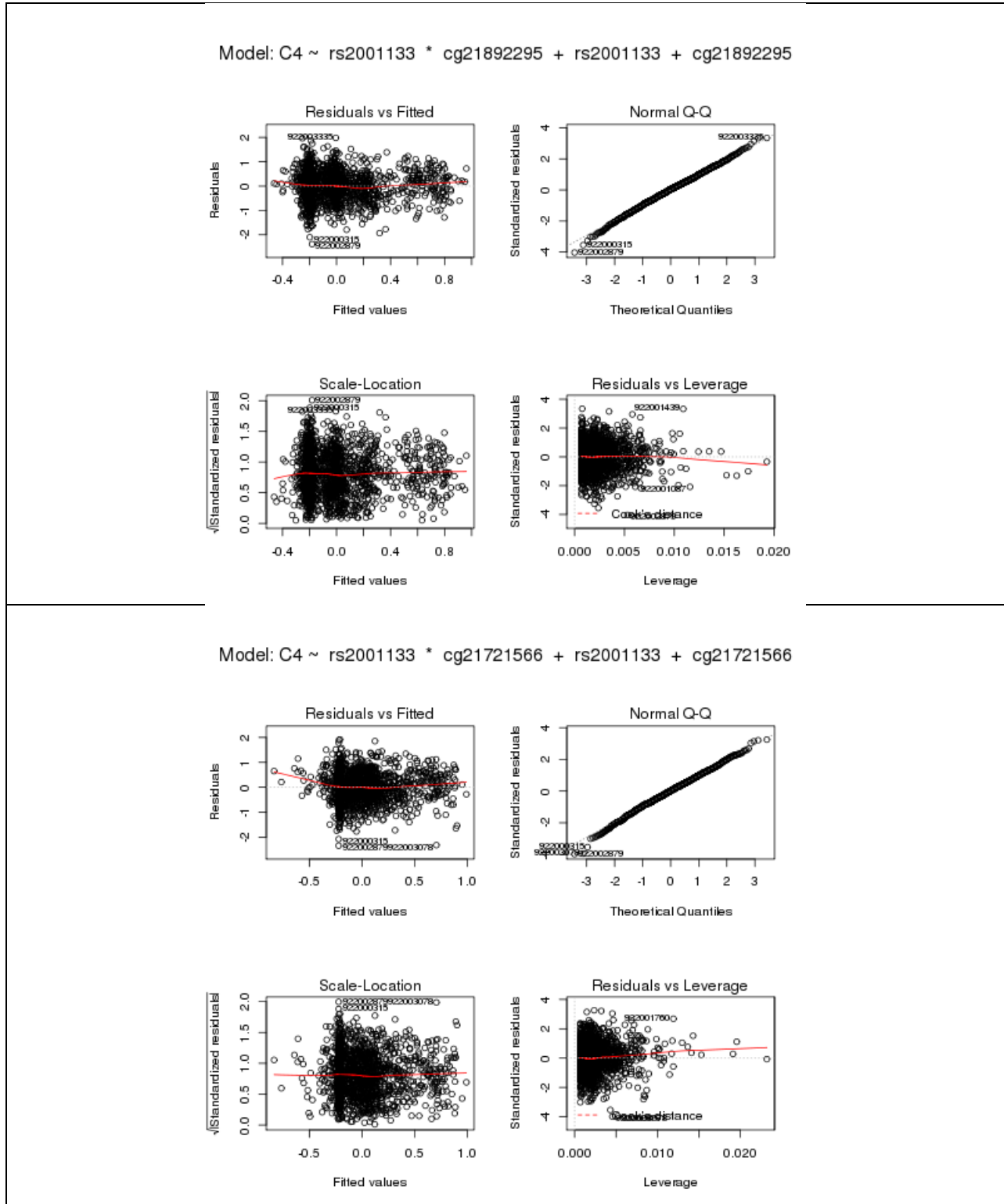
Model: C4 ~ rs10774563 \* cg21721566 + rs10774563 + cg21721566



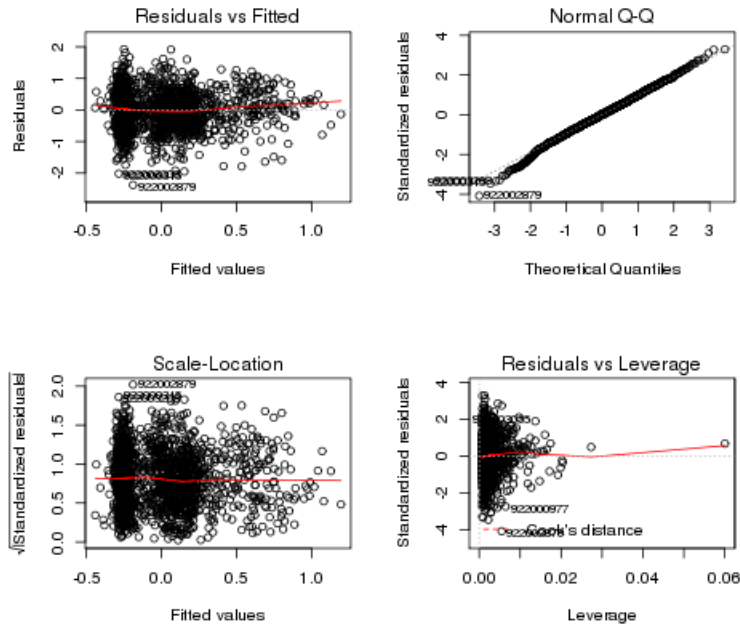
**Figure A.3:** Four plots for each significant association between butyrylcarnitine measured in the Metabolon platform and the SNP-CpG-interaction to examine whether the assumptions for the linear regression are satisfied: a plot of residuals against fitted values, a Scale-Location plot of

sqrt(| residuals |) against fitted values, a Normal Q-Q plot, and a plot of residuals against leverages.

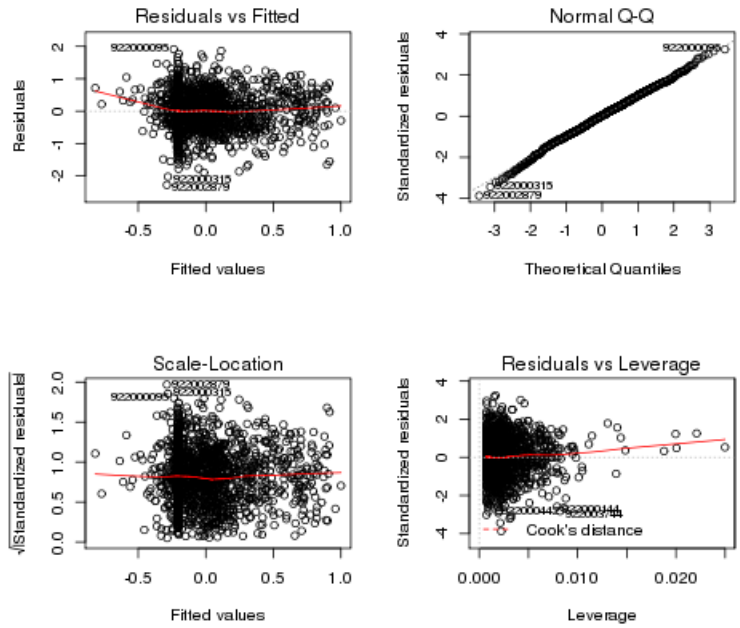
**Biocrates**

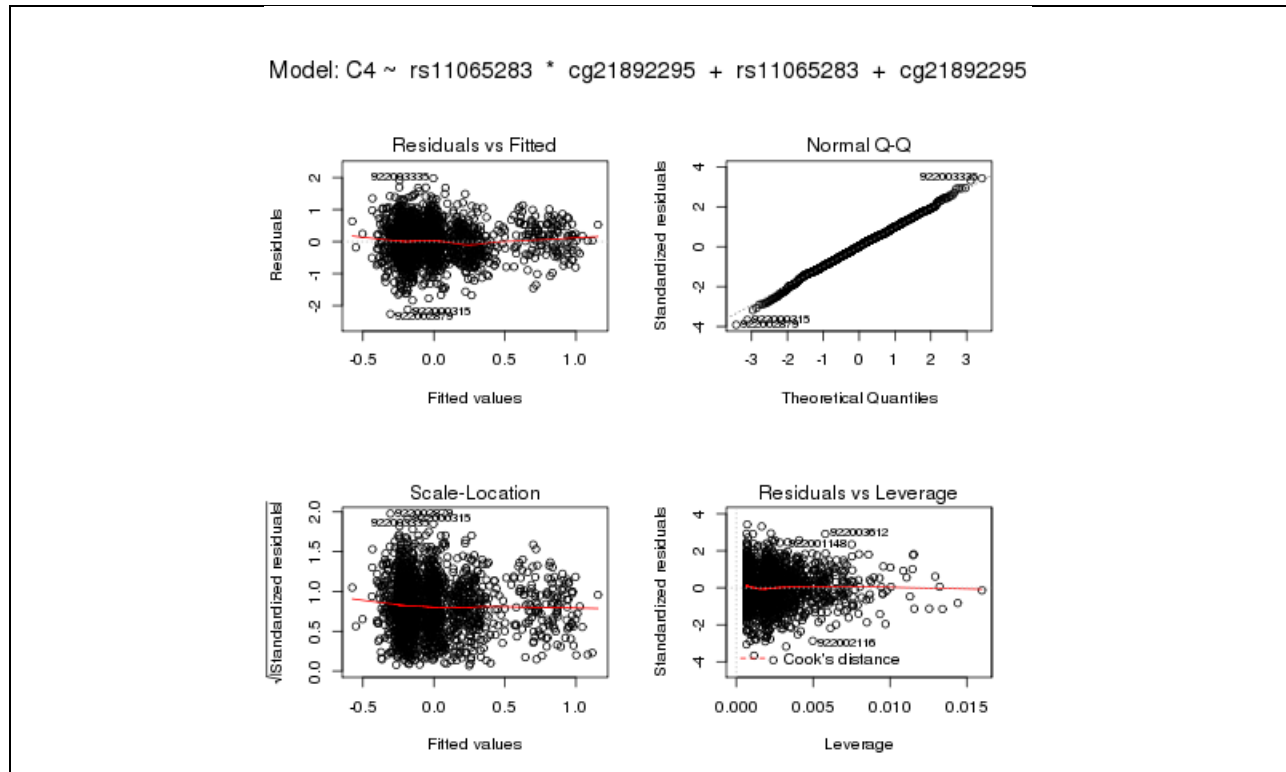


Model: C4 ~ rs7964786 \* cg02419362 + rs7964786 + cg02419362

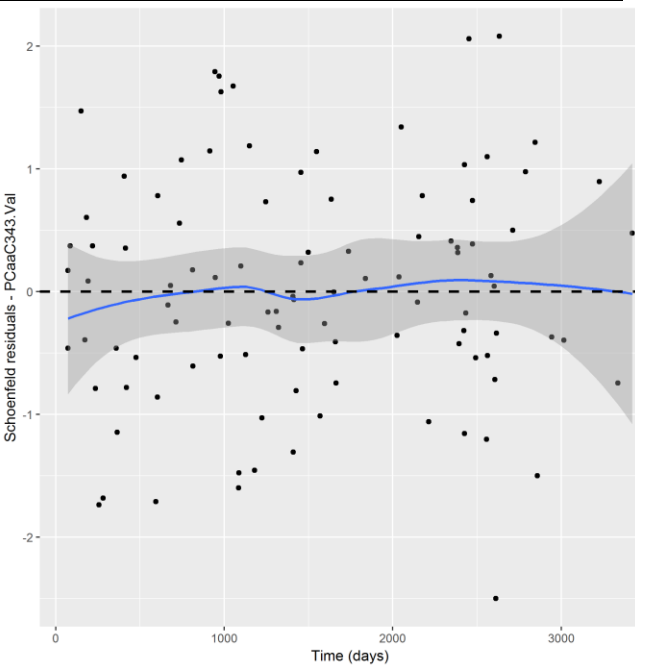
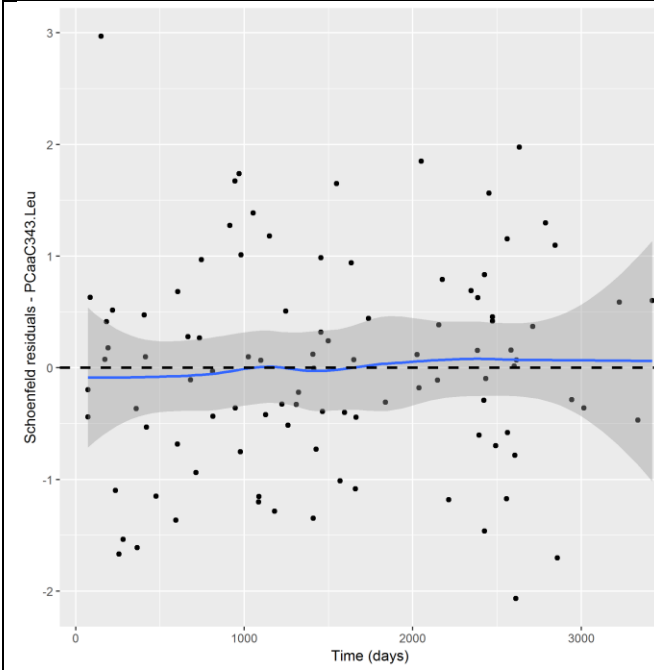
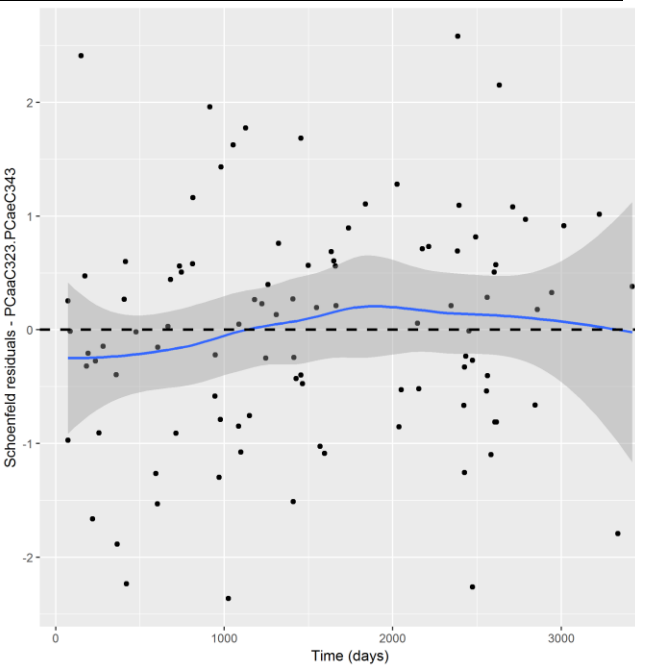
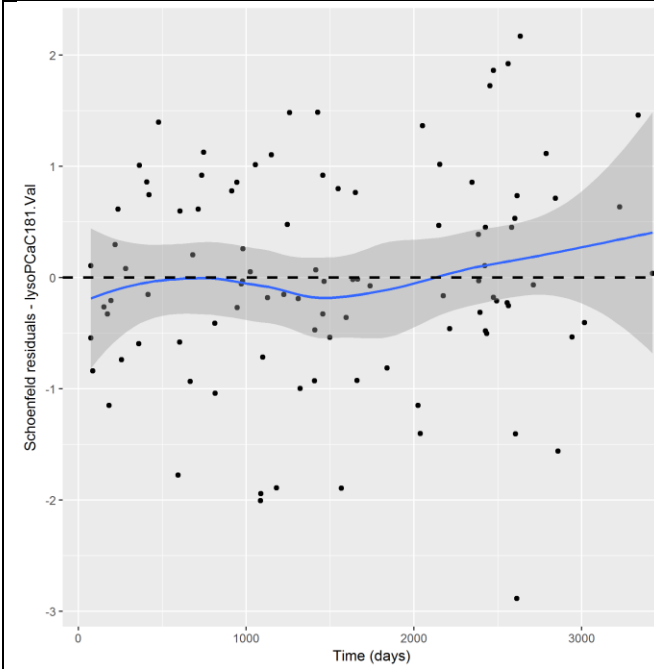


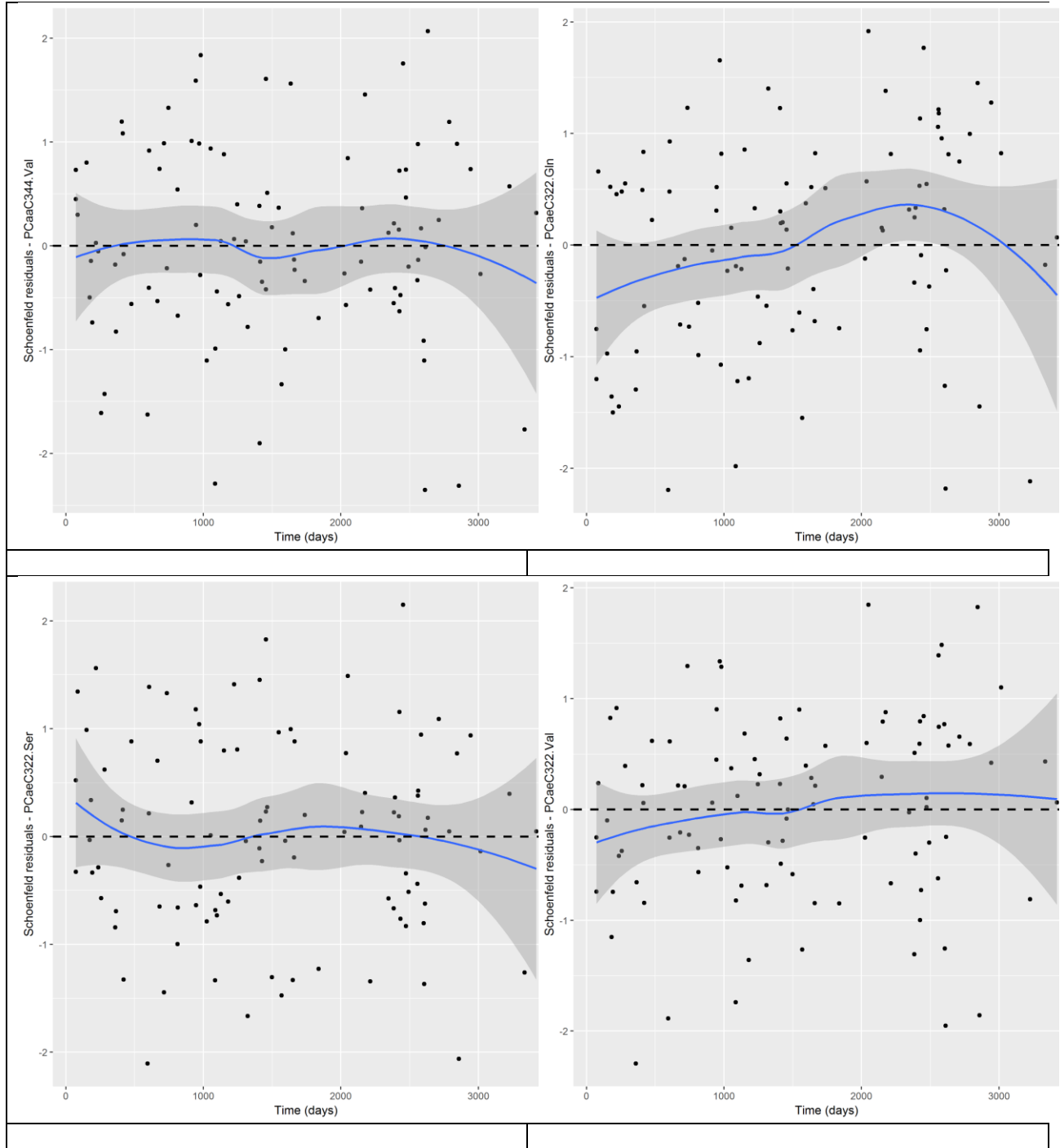
Model: C4 ~ rs11065283 \* cg21721566 + rs11065283 + cg21721566

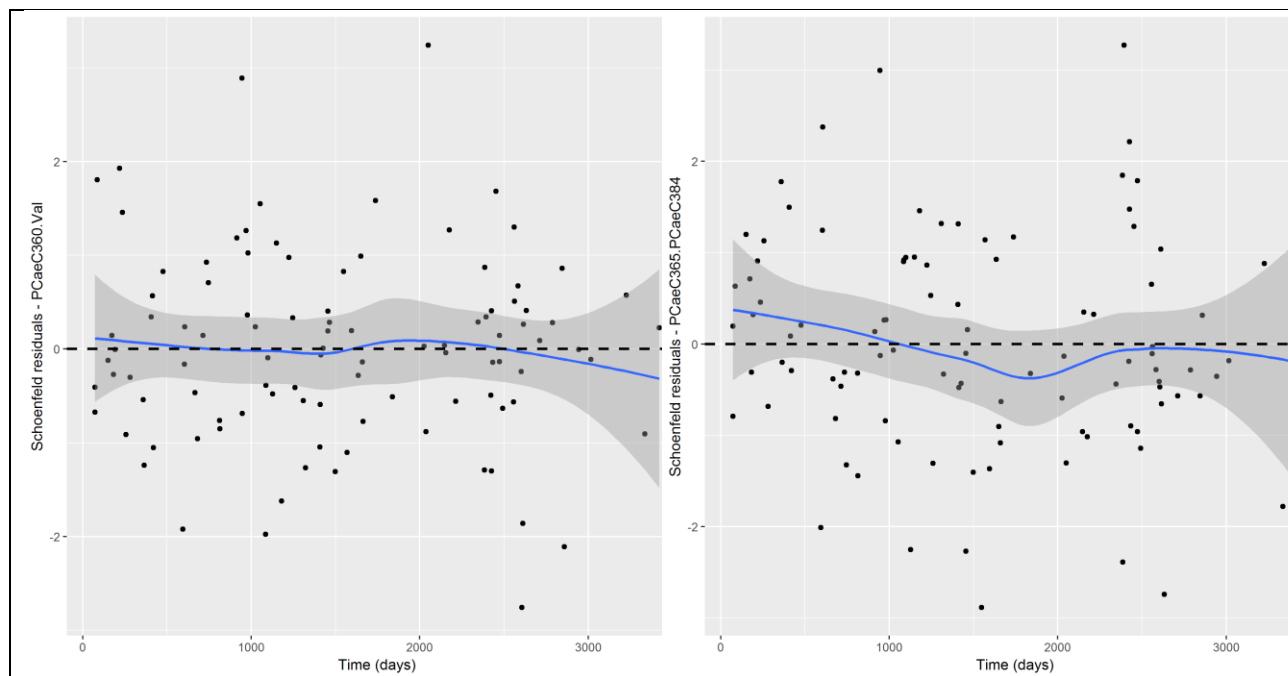




**Figure A.4:** Four plots for each significant association between butyrylcarnitine measured in the Biocrates platform and the SNP-CpG-interaction to examine whether the assumptions for the linear regression are satisfied: a plot of residuals against fitted values, a Scale-Location plot of  $\sqrt{|\text{residuals}|}$  against fitted values, a Normal Q-Q plot, and a plot of residuals against leverages.







**Figure A.5:** Plot of Schoenfeld residuals against the time for each significant association between T2D and the metabolite ratios measured in the Biocrates platform to examine whether the assumptions for the Cox proportional hazards regression are satisfied.