Cover Page

Universiteit Leiden

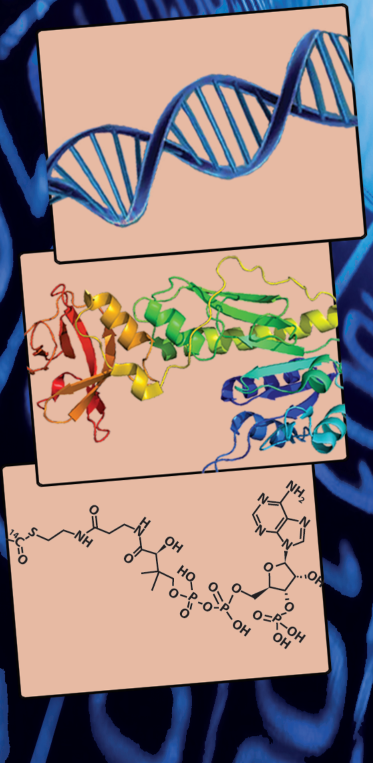
Leiden University
Repository

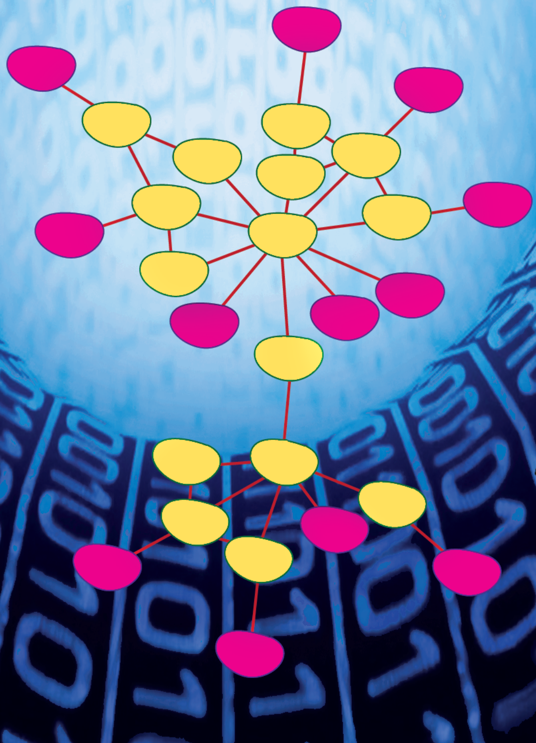
The handle http://hdl.handle.net/1887/38350 holds various files of this Leiden University dissertation.

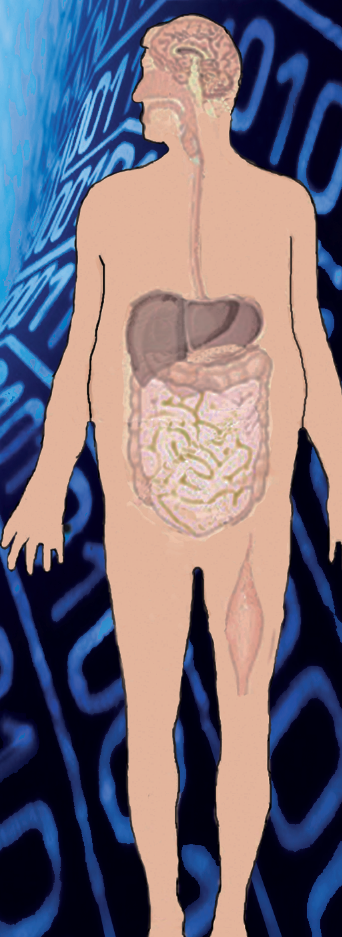
Author: Dharuri, Harish
Title: Bioinformatic approaches to identify genomic, proteomic and metabolomic biomarkers for the metabolic syndrome
Issue Date: 2016-03-02

Bioinformatic Approaches to Identify Genomic, Proteomic and Metabolomic Biomarkers for the Metabolic Syndrome

HARISH DHARURI







Bioinformatic Approaches to Identify Genomic, Proteomic and Metabolomic Biomarkers for the Metabolic Syndrome

by

Harish Dharuri

Cover design and layout by: Aria publishers, Fremont, USA and Harish Dharuri
Printed by: Gildeprint
ISBN: 978-94-6233-229-4
© 2016, Harish Dharuri

Bioinformatic Approaches to Identify Genomic, Proteomic and Metabolomic Biomarkers for the Metabolic Syndrome

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker
volgens besluit van het College voor Promoties
te verdedigen op woensdag 2 maart 2016
klokke 15:00 uur
door

Harish Dharuri
geboren te Dharur, India
in 1969

promotiecommissie

promotor: prof. dr. ir. J.A.P Willems van Dijk
co-promotor: dr. P.A.C. ’t Hoen
overige leden: prof.dr. A.M.J.M. van den Maagdenberg
prof.dr. J.T. den Dunnen
prof.dr. C van Duijn(1)
prof.dr.ir C Evelo(2)

(1) Department of Epidemiology, Erasmus MC, Rotterdam
(2) Department of Bioinformatics, Maastricht University, Maastricht

“One who sees inaction in action, and action in inaction, is intelligent among humans”

Bhagavad Gita
-ancient Indian scripture

TABLE OF CONTENTS

Chapter 1: General introduction and outline 1

Chapter 2: Automated workflow-based exploitation of pathway databases provides new insights into genetic associations of metabolite profiles 11

Chapter 3: Structuring research methods and data with the Research Object model: genomics workflows as a case study 55

Chapter 4: Down-regulation of the acetyl-CoA metabolic network in adipose tissue of obese diabetic individuals and recovery after weight loss 89

Chapter 5: Differential allele specific expression of KLRK1 in subcutaneous and visceral adipose tissue of very obese individuals with and without type 2 diabetes mellitus 129

Chapter 6: Proteomic Analysis in Type 2 Diabetes Patients before and after a Very Low Calorie Diet Reveals Potential Disease State and Intervention Specific Biomarkers 145

Chapter 7: Prolonged niacin treatment leads to increased adipose tissue PUFA synthesis and an anti-inflammatory lipid and oxylipin plasma profile 191

Chapter 8: Genetics of the human metabolome, what is next? 223

Chapter 9: General Discussion 251

Summary 259

Samenvatting 265

Curriculum Vitae 269

Publications... 271

Acknowledgements.. 273

Chapter 1: General introduction and outline

The elucidation of the DNA structure by Watson and Crick marked the beginning of a new era in biological sciences [1]. This structure revealed that the basis of heredity lies in the arrangement of just four nucleotides: adenine, thymine, guanine and cytosine. This digital feature of DNA engendered a new view of biology as an information science. The discovery of the DNA structure spawned the field of molecular biology to investigate the molecular genetic basis of life. The central dogma of molecular biology put forth by Francis Crick in 1958 holds that genes - ordered sequences of nucleotides along the DNA molecule - are transcribed into messenger RNAs which are then translated into polypeptide chains [2]. This directional transfer of information reinforced the notion that life can be interpreted as a molecular process regulated by genetic information.

The reductionist method of dissecting biological systems into their constituent parts has provided a wealth of information pertaining to molecular and cellular processes. However, in recent years the limits of the reductionist approach have become increasingly evident. Under question is not the value of these investigations, but rather that life can be fully understood at the molecular and genetic level applying the reductionist approach. Biological systems are extremely complex [3], are composed of many intricately connected components and have emergent properties, i.e. the whole is much more than the sum of the parts [4]. Recent developments in high throughput data measurement, processing and storage have exposed the limitations of hypothesis-driven research. It appears nearly impossible to manipulate a single component of a biological system, without simultaneously affecting many other components.

Exactly fifty years after the discovery of the structure of DNA, another epochal event took place in biological research: the sequence of the human genome was completed, which provided a genetic blueprint of a human being [5]. It is now possible for researchers to simultaneously investigate all genes. Faced with massive data, the hypothesis driven approach to science, while still valid, is increasingly being seen as one of two approaches. Discovery science, based on inductive reasoning that uncovers important rules through careful observations, is being embraced as a second approach, to make sense of the data deluge in the post-genomic era.

OMICS-Revolution

Technological advancements in the biological sciences have facilitated a paradigm shift from exclusively hypothesis-driven to hypothesis- and data-driven scientific exploration. In contrast to hypothesis-driven research, a

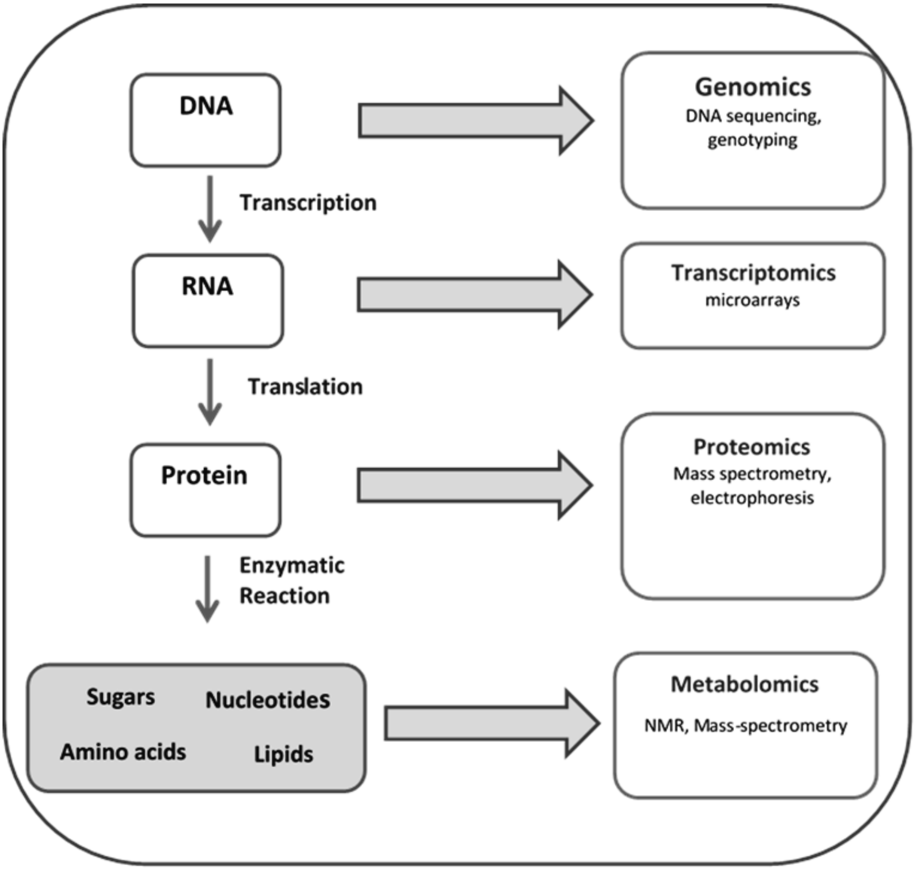


Figure 1 OMICs data. The components and process of information flow in biological systems is depicted on the left and the OMIC technologies to measure the components are shown on the right side of the figure.

data-driven approach allows rapid evaluation of additional hypotheses followed by refining candidates into a smaller set of testable hypothesis. The data rich environment necessary for such exploration is in large part driven by high-throughput "OMICS" experiments, as shown in Figure 1, which routinely generate "genome-wide" data. High-throughput "OMICS" technologies include methods to identify and quantify DNA and RNA (genomics), proteins (proteomics), metabolites (metabolomics) and other biologically relevant entities.

Data analyses

Statistical Analysis

A wide spectrum of analytic techniques has been proposed to analyze high-throughput omics data. Application of univariate statistical tests, for example, t-tests and ANOVA, is a common approach to assess group wise differences. However, the high dimensionality of a typical omics dataset poses a serious challenge to the validity of univariate tests. An omics approach often leads to high dimensional and low-sample size data settings where the number of variables measured (e.g., mRNA, proteins, metabolites) exceeds the number of samples by far. Application of univariate tests to such datasets may result in a high number of false positives, known as the multiple testing problem. Moreover, the predominant approach of p-value correction to account for these false positive (eg., Benjamini and Hochberg's false discovery rate (FDR), Bonferroni correction) may be a bit too conservative and are associated with significant losses in statistical power. As an alternative, multivariate statistical techniques, for example principal component analysis (PCA) and partial least squares regression (PLS) are being employed for integration and interpretation of omics datasets [6]. Often, several distinct approaches to investigate the same dataset are needed to come to a proper interpretation.

Pathway Analysis

Even as data generation is proceeding at an unprecedented pace, translation of this data into actionable biological insight remains a critical challenge. To address this issue, pathway analysis that combines analytical tools and *a priori* biological knowledge is increasingly being recognized as an important strategy to gain a deeper and broader understanding of biological underpinnings of experimental observations. Pathway-based approaches examine test statistics for a group of genes in contrast to single-marker analysis [7]. The 'group of genes' is an expert defined set that is functionally related to the phenotype. The term 'pathway' in a pathway analysis is usually referring to a set of functionally related genes participating in a common biological process. An important example of pathway analysis is Gene Set Enrichment Analysis (GSEA) [8]that was initially proposed for microarray analysis and has subsequently been modified and applied to GWAS data [7]. The goal of GSEA and other pathway-based methods is to examine the behaviour of gene sets rather than single genes across the biological conditions investigated.

The resources of prior knowledge that are commonly used in pathway analysis include controlled vocabularies like Gene Ontology [9], manually curated gene sets from MSigDB [8] and the pathway databases like KEGG [10], BioCyc [11] and REACTOME [12]. However, biological data resources in general and pathway databases in particular have low consensus [13] that mandates interrogation and integration of multiple databases in order to ensure comprehensive data collection. Crowd sourcing efforts in building pathway databases such as WikiPathways [14] address this problem through a community resource that allows contributions from users towards building pathways in addition to integration of publicly available data and customization of information content.

Integration of heterogeneous and disparate data resources remains a key bioinformatics challenge. Recent developments in workflow technologies in general and scientific workflow tools like Taverna [15] and Galaxy [16] in particular, have facilitated an easy interface for integration of disparate biological data resources. In addition these technologies make data analysis routines reusable and reproducible. A closely associated concept to scientific workflows is the idea of the Semantic Web. The latter is an extension of the Web built on standards laid out by the World Wide Web consortium (W3C) (www.w3.org). Semantic Web facilitates the integration of heterogonous data on the World Wide Web by making the semantics of the data explicit through formal ontologies [17]. The inclusion of semantic web technologies into scientific workflows enables *in silico* experimentation and is increasingly being recognized as a promising platform for integrative biology [18].

This thesis combines statistical and bioinformatic techniques to extract greater value from high-throughput datasets than is possible using traditional data analysis and interpretation approaches. This work is in line with the paradigm of e-science, in that the overarching research theme is to promote scientific discovery through analysis of data over distributed environments. More specifically, we demonstrate the utility of scientific workflows in facilitating knowledge discovery in high-throughput datasets like Genome-Wide Association Studies (GWAS), Next-Generation Sequencing (NGS) and microarray datasets.

Biological problem

The metabolic syndrome (MetS) is defined as a cluster of metabolic abnormalities including central obesity, hypertension, hyperglycemia and dyslipidemia [19]. It is associated with increased risk of type 2 diabetes, cardiovascular disease and stroke. The increasing prevalence of MetS is driven by the obesity epidemic and poses a serious health problem worldwide. Effective prevention and intervention requires improved understanding of factors that contribute to MetS. It is now understood that the syndrome results from a complex interplay of environmental and genetic components. Epidemiological studies have shown that social and lifestyle issues like physical inactivity, western-style diet and age increase the risk for MetS. On the other hand, family and twin studies indicate that the genetic component also plays an important role.

From 2007 on, Genome-Wide Association studies (GWAS) have helped identify common genetic variants associated with obesity and other metabolic syndrome traits [20]. However, the cumulative contribution of common variants, as accounted for by GWAS, to the heritability of these traits is quite modest. In addition, the biological context of candidate genes detected by means of GWAS frequently remains unclear. Often, the assigned gene is located at a significant distance from the associated variant and causality between gene and variant is not known. To gain additional insight in the relation between genetic variants, metabolic traits and outcome, GWAS analysis of metabolite levels has recently sparked interest. These intermediate phenotypes generally demonstrate larger effect sizes and potentially point at pathways relevant to disease [21]. Metabolite GWAS results are proving to be excellent starting points for functional studies as well as bioinformatics and systems biology approaches to unravel novel biochemical pathways underlying complex traits like type 2 diabetes and related disorders. In this thesis, we explore pathway and network analysis of high-throughput datasets to gain further mechanistic insight into complex traits like type 2 diabetes.

OUTLINE OF THE THESIS

The aim of the present thesis is to identify biomarkers in genomic, proteomic and metabolomics datasets using novel bioinformatic techniques. In Chapter 2 we demonstrate the utility of automated exploitation of background knowledge present in pathway databases for the analysis of GWAS datasets of metabolomics phenotypes. This research work explores a strategy to identify novel and biologically relevant SNP-metabolite pairs in Genome-Wide Association Studies (GWAS) of metabolite profiles. We demonstrate the utility of an automated workflow approach that utilizes prior knowledge of biochemical pathways present in databases like KEGG and BioCyc to generate a smaller SNP set relevant to the metabolite. In addition to reporting novel loci, this chapter presents the opportunities and challenges in the analysis of GWAS of metabolomic phenotypes.

Encapsulating all aspects of an *in silico* analysis and communicating it to the scientific community is a key challenge in a computational experiment. **Chapter 3** explores the utility of semantic web technologies in the preservation of computational experiments. More specifically, the chapter discusses the Research Object (RO) model, where a research object is defined as a resource that aggregates other resources, e.g. datasets, software, spreadsheets, text, etc. The RO model was applied to a study where the goal was to facilitate the interpretation of the results of a GWAS of metabolite profiles.

Obesity is a growing world-wide epidemic and is associated with decreased life expectancy due to associated metabolic and cardiovascular disorders. The expanded adipose tissue is thought to serve as the pathogenic link between obesity and type-2 diabetes. While a majority of obese individuals develop insulin resistance and type-2 diabetes (T2DM), some remain metabolically healthy or Normal Glucose Tolerant (NGT). **Chapter 4** presents a study designed to investigate the role of the adipose tissue in development of T2DM in severely obese subjects by performing RNA-Sequencing of the subcutaneous (SAT) and visceral adipose tissue (VAT) samples. We demonstrate a bioinformatic network-based approach that helped identify an important biochemical feature in the pathophysiology of type 2 diabetes in obese individuals. **Chapter 5** addresses the issue of Allelic imbalance which is the uneven expression of a transcript from its two allelic copies in heterozygous individuals. A growing number of studies have shown that a genetic variation in non-coding regions of the genome has important consequences for phenotypic variation. The objective of this study was to identify, from a panel of known diabetes and obesity susceptible loci, as reported by published GWA studies, the subset of genes that are under the control of *cis*-regulatory elements. RNA-Seq data from the SAT and VAT of T2DM and NGT subjects mentioned in the earlier chapter was utilized to determine if there was a tissue-specific allelic imbalance in known obesity associated loci. The bioinformatic and statistical investigation helped identify a novel locus that displays a differential allele-specific expression between the two tissues.

Chapter 6 and **Chapter 7** demonstrate the utility of pathway analysis in extracting biological meaning from proteomic and microarray datasets. **Chapter 6** pertains to studies in obese T2DM patients subjected to Very low calorie diets (VLCD) with and without exercise programs that lead to major metabolic improvements in these subjects. Proteomic analysis using blood samples from these subjects was used to uncover novel biomarkers for these interventions. In addition to statistical analysis, we employed text mining and pathway-based approaches to gain an understanding of intervention-specific biomarkers. In **Chapter 7** pathway analysis is applied to microarray data obtained from adipose tissue of mice treated with or without niacin. The conclusion from the bioinformatics and statistical analysis was used to guide *in vivo* and *in vitro* investigation into biochemical feature of prolonged niacin treatment in mice.

In **Chapter 8**, we present a global review of the current status of metabolomic GWAS (mGWAS) and sketch future directions towards enhanced interpretation of these studies. Finally, **Chapter 9** provides a general discussion of topics mentioned in chapters 4-7.

References

1. Watson JD, Crick FHC: **Molecular structure of nucleic acids**. *Nature* 1953, **171**:737–738.
2. CRICK FH: **On protein synthesis.** *Symp Soc Exp Biol* 1958, **12**:138–163.
3. Kitano H: **Systems biology: a brief overview.** *Science (80-)* 2002, **295**:1662–4.
4. Bhalla US, Iyengar R: **Emergent properties of networks of biological signaling pathways.** *Science* 1999, **283**:381–387.
5. International Human Genome Sequencing Consortium.: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431**:931–945.
6. Chadeau-Hyam M, Campanella G, Jombart T, Bottolo L, Portengen L, Vineis P, Liquet B, Vermeulen RCH: **Deciphering the complex: Methodological overview of statistical models to derive OMICS-based biomarkers**. *Environmental and Molecular Mutagenesis* 2013:542–557.
7. Wang K, Li M, Hakonarson H: **Analysing biological pathways in genome-wide association studies.** *Nat Rev Genet* 2010, **11**:843–854.
8. Subramanian A, Subramanian A, Tamayo P, Tamayo P, Mootha VK, Mootha VK, Mukherjee S, Mukherjee S, Ebert BL, Ebert BL, Gillette M a, Gillette M a, Paulovich A, Paulovich A, Pomeroy SL, Pomeroy SL, Golub TR, Golub TR, Lander ES, Lander ES, Mesirov JP, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102**:15545–50.
9. Gene T, Consortium O: **Gene Ontology : tool for the**. *Nat Genet* 2000, **25**(may):25–29.

10. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Res* 2009, **38**.
11. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD: **Computational prediction of human metabolic pathways from the complete human genome.** *Genome Biol* 2005, **6**:R2.
12. D'Eustachio P: **Pathway databases: Making chemical and biological sense of the genomic data flood**. *Chemistry and Biology* 2013:629–635.
13. Stobbe MD, Houten SM, Jansen G a, van Kampen AHC, Moerland PD: **Critical assessment of human metabolic pathway databases: a stepping stone for future integration.** *BMC Syst Biol* 2011, **5**:165.
14. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR: **WikiPathways: building research communities on biological pathways.** *Nucleic Acids Res* 2012, **40**(Database issue):D1301–7.
15. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P, Bhagat J, Belhajjame K, Bacall F, Hardisty A, Nieva de la Hidalga A, Balcazar Vargas MP, Sufi S, Goble C: **The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud.** *Nucleic Acids Res* 2013, **41**(Web Server issue).
16. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11**:R86.
17. Berners-Lee T, Hendler J, Lassila O: **The Semantic Web**. *Scientific American* 2001:34–43.
18. Chen H, Yu T, Chen JY: **Semantic web meets integrative biology: A survey**. *Briefings in Bioinformatics* 2013:109–125.
19. Huang PL: **A comprehensive definition for metabolic syndrome**. *Dis Model Mech* 2009, **2**:231–237.
20. Fall T, Ingelsson E: **Genome-wide association studies of obesity and metabolic syndrome**. *Molecular and Cellular Endocrinology* 2014:740–757.
21. Gieger C, Geistlinger L, Altmaier E, De Angelis MH, Kronenberg F, Meitinger T, Mewes HW, Wichmann HE, Weinberger KM, Adamski J, Illig T, Suhre K: **Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum**. *PLoS Genet* 2008, **4**.

Chapter 2: Automated workflow-based exploitation of pathway databases provides new insights into genetic associations of metabolite profiles

Harish Dharuri
Peter Henneman
Ayse Demirkan
Jan Bert van Klinken
Dennis Owen Mook-Kanamori
Rui Wang-Sattler
Christian Gieger
Jerzy Adamski
Kristina Hettne
Marco Roos
Karsten Suhre
Cornelia M. Van Duijn
EUROSPAN consortia
Ko Willems van Dijk
Peter A.C. ’t Hoen

BMC Genomics. 2013; 14:865

Abstract

Background: Genome-wide association studies (GWAS) have identified many common single nucleotide polymorphisms (SNPs) that associate with clinical phenotypes, but these SNPs usually explain just a small part of the heritability and have relatively modest effect sizes. In contrast, SNPs that associate with metabolite levels generally explain a higher percentage of the genetic variation and demonstrate larger effect sizes. Still, the discovery of SNPs associated with metabolite levels is challenging since testing all metabolites measured in typical metabolomics studies with all SNPs comes with a severe multiple testing penalty. We have developed an automated workflow approach that utilizes prior knowledge of biochemical pathways present in databases like KEGG and BioCyc to generate a smaller SNP set relevant to the metabolite. This paper explores the opportunities and challenges in the analysis of GWAS of metabolomic phenotypes and provides novel insights into the genetic basis of metabolic variation through the re-analysis of published GWAS datasets.

Results: Re-analysis of the published GWAS dataset from Illig *et al* (Nature Genetics, 2010) using a pathway-based workflow (http://www.myexperiment.org/packs/319.html), confirmed previously identified hits and identified a new locus of human metabolic individuality, associating Aldehyde dehydrogenase family1 L1 (*ALDH1L1*) with serine / glycine ratios in blood. Replication in an independent GWAS dataset of phospholipids (Demirkan *et al*, PLoS Genetics, 2012) identified two novel loci supported by additional literature evidence: *GPAM (*Glycerol-3 phosphate acyltransferase) and *CBS* (Cystathionine beta-synthase). In addition, the workflow approach provided novel insight into the affected pathways and relevance of some of these gene-metabolite pairs in disease development and progression.

Conclusions: We demonstrate the utility of automated exploitation of background knowledge present in pathway databases for the analysis of GWAS datasets of metabolomic phenotypes. We report novel loci and potential biochemical mechanisms that contribute to our understanding of the genetic basis of metabolic variation and its relationship to disease development and progression.

Background

GWAS have resulted in the identification of novel genetic loci associated with a variety of diseases and clinical phenotypes. However, a disease or clinical phenotype is the end point of the behaviour of numerous genes and pathways in addition to environmental influences. This at least partly explains the general observation that the effect size of genetic association with clinical phenotypes is rather small. Spurred by recent technological developments in the field of metabolomics, interest in genome wide association studies with metabolite levels in blood [1,2,3,4] is gathering momentum. Metabolites are intermediate phenotypes, entities that lie between genes and clinical end points [5,6]. Due to their proximity to an enzyme/gene, metabolites may offer greater effect sizes for GWAS than clinical phenotypes [7]. Moreover, the pathways in which the metabolite plays a role may provide insight into the underlying biological mechanism responsible for the development of the associated disease.

Typically, in metabolomics GWAS, hundreds of metabolites are tested for genetic association. However, association of all SNPs with all measured metabolites comes with considerable multiple testing problems. Recent publications have also shown that testing ratios of metabolites for genetic association results in much larger effect sizes; however this further exacerbates the multiple testing problem which precludes genuine SNP-metabolite pairs from reaching genome-wide significance. Several approaches like gene based tests [8,9] and pathway analysis [10] have been proposed to overcome this limitation of inadequate statistical power in GWAS. All these approaches have been suggested in the context of GWAS with clinical phenotypes but genetic association with metabolites presents its own set of unique opportunities and challenges. Herewith, we explore the utility of background knowledge present in metabolic pathway databases to increase the power in identification of metabolite Quantitative Trait Loci (mQTL).

Our approach involves selective testing of SNPs near genes in pathways supposedly relevant to the metabolite levels, as a way to reduce the multiple testing burden in GWAS. Background knowledge pertaining to a metabolite is retrieved through systematic interrogation of metabolic pathway databases which describe biochemical pathways, reactions, and enzymes relevant to human metabolism. Several pathway databases have been created by groups around the world, while the intent of these efforts remains the elucidation of biological mechanism, the databases however, differ quite significantly in their content, size, user accessibility, download formats and most importantly availability and type of web services for machine-enabled interrogation of the database [11]. In this publication, as a proof of principle, we have chosen to focus on two important metabolic pathway databases,

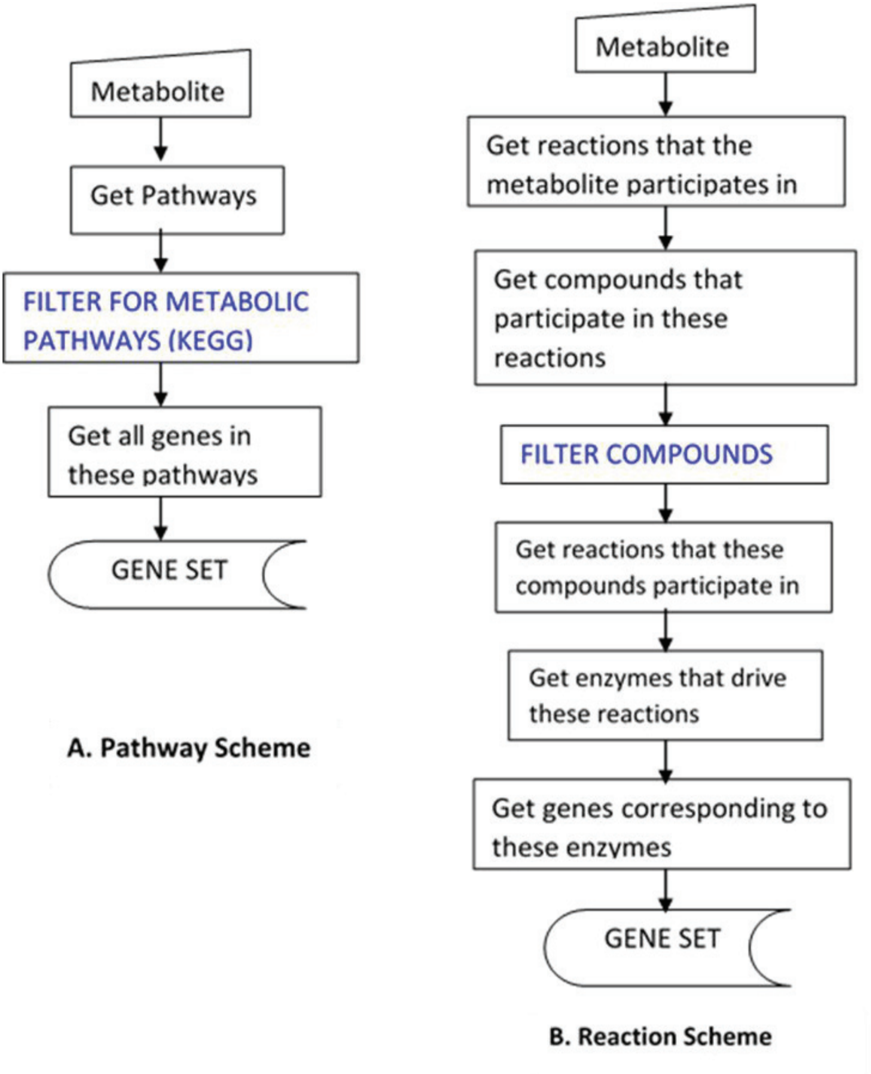

Metabolite
Get Pathways
FILTER FOR METABOLIC PATHWAYS (KEGG)
Get all genes in these pathways
GENE SET
A. Pathway Scheme
Metabolite
Get reactions that the metabolite participates in
Get compounds that participate in these reactions
FILTER COMPOUNDS
Get reactions that these compounds participate in
Get enzymes that drive these reactions
Get genes corresponding to these enzymes
GENE SET
B. Reaction Scheme

Figure 1 The database interrogation schemes. The two interrogation schemes: pathway scheme (A) and reaction scheme (B) are shown. The blue color indicates the intermediate steps to filter out certain pathways/compounds from the two schemes to avoid non-specific connections.

KEGG [12] and BioCyc [13]. KEGG is an integrated database resource of seventeen databases which provide system, genomic and chemical information. The pathway database consists of both metabolic and non-

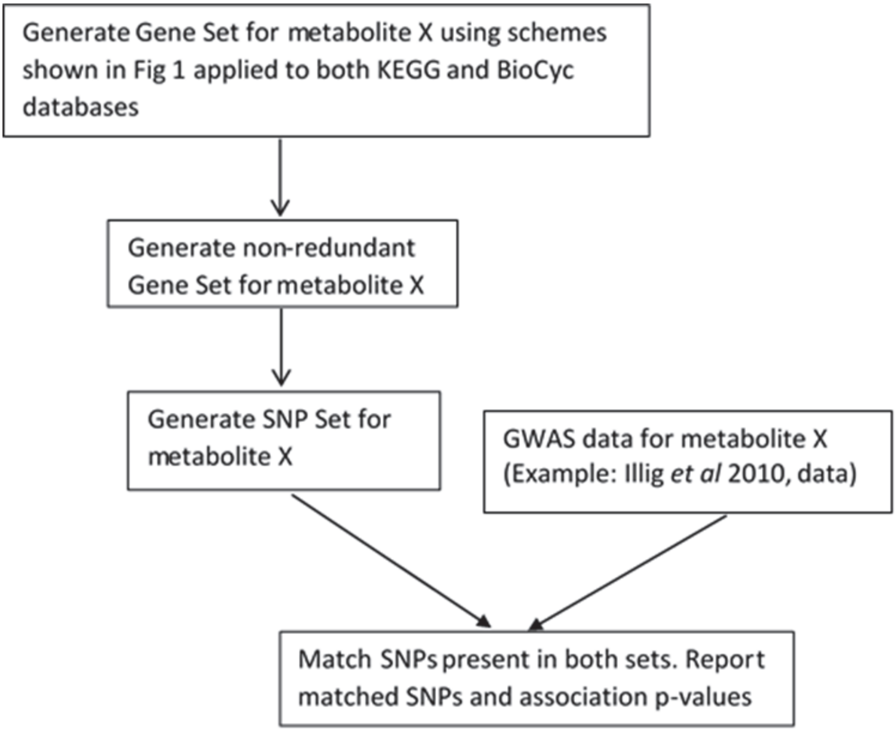

Generate Gene Set for metabolite X using schemes shown in Fig 1 applied to both KEGG and BioCyc databases
Generate non-redundant Gene Set for metabolite X
Generate SNP Set for metabolite X
GWAS data for metabolite X (Example: Illig *et al* 2010, data)
Match SNPs present in both sets. Report matched SNPs and association p-values

Figure 2 Strategy to find biologically relevant SNP-metabolite pairs in published GWAS datasets. Background knowledge pertaining to a metabolite is collected from the pathway databases KEGG and BioCyc in an automated fashion to generate a gene/SNP set relevant to the synthesis and degradation of the metabolite.

metabolic pathways and is constructed by a team of curators based on information available in the literature. BioCyc is a collection of pathway/genome databases that describe the genome and metabolic pathways of several organisms. The database that describes human genomes and pathways, HumanCyc was interrogated in this study. In our approach, for every metabolite under consideration, genes acting in the vicinity of the metabolite are determined using knowledge present in databases mentioned above. We thus generate an integrated set of genes that represent entities with influence over the metabolite. A workflow management system called Taverna [14] was used to generate these gene sets and the SNPs associated with these genes. The workflows that were designed for this purpose have been submitted to a workflow repository at http://www.myexperiment.org/packs/319.html [15].

A previously published metabolomics dataset by Illig *et al* 2010 [2] was analyzed to evaluate the sensitivity of the method in picking true positives and to identify novel SNP-metabolite pairs that had hitherto been obscured in the GWA list given the stringent threshold for significance. In addition to validating a novel bioinformatics workflow analysis tool, we identified a new locus of human metabolic individuality, Aldehyde dehydrogenase family1 L1 (*ALDH1L1*). This locus was found associated with serine/glycine ratios, a metabolic trait that functionally matches the gene function.

Candidate genes identified through the analysis of Illig *et al* dataset were taken up for replication in a separate study published by Demirkan *et al* [4]. We report *GPAM* (Glycerol-3 phosphate acyltransferase) and *CBS* (Cystathionine beta-synthase) as novel loci associated with phosphatidylcholine moieties.

Results

Our approach can be divided into three stages: (i) Generate a non-redundant gene set for every metabolite considered using knowledge in pathway databases like KEGG and BioCyc applying interrogation schemes as shown in Fig 1 and outlined below. (ii) For every gene in the set, generate the set of SNPs within the gene and 50 kb flanking sequences, and create a SNP set for each metabolite (iii) Match SNPs generated for a metabolite with the GWAS for the same metabolite and store the matches with the p-values reported for the association (Fig 2).

Analysis strategy of databases and Interrogation schemes

To retrieve a prioritized list of candidate genes associated with metabolite levels, gene sets were generated for each metabolite through the pathway scheme and the reaction scheme [Fig 1A and 1B] for the KEGG and BioCyc databases (see Method). The pathway scheme generates a list of genes that participate in pathways relevant to the synthesis or degradation of the metabolite. In the reaction scheme, the metabolite is used as a seed node and shells of reactions around the metabolite are explored. The list of genes that catalyse the reactions are retrieved and form the gene set for the given metabolite. For every gene set, a corresponding SNP set is generated by retrieving SNPs within the flanking 50 kb of every gene. In the final step, the SNP set for a metabolite is matched with the GWAS dataset for the same

Table 1 Gene and SNP sets generated by the database: interrogation schemes for each of the metabolites

Metabolite	BioCyc	BioCyc	KEGG	KEGG	Size of unique	Size of unique	Number of
Arginine	20	104	57	179	257	10788	10788
Glutamine	51	132	100	282	388	15591	15591
Glycine	90	192	173	432	523	20767	20767
Histidine	8	9	45	155	181	7126	7126
Leucine	8	0	44	83	117	5037	5037
Methionine	27	104	35	243	284	11532	11532
Ornithine	16	150	103	159	247	10089	10089
Phenylalanine	6	113	25	163	196	8419	8419
Proline	10	12	57	83	119	5075	5075
Serine	37	135	152	219	360	14996	14996
Threonine	1	11	39	49	75	2633	2633
Tryptophan	15	19	78	221	261	10419	10419
Tyrosine	14	106	61	158	219	9365	9365
Valine	15	93	80	137	211	9365	9365
Carnitine	32	206	81	94	263	11239	460799
Phosphatidylcholine	188	361	312	343	640	31676	2914192
Sphingomyelin	160	331	189	241	460	21290	319350
Sum	698	2078	1631	3241	4801	205407	3835543
Unique Set	399	806	703	768	1246	55952	55952

The number of genes for each metabolite and the corresponding database:interrogation scheme is shown. [1] The size of the union of the gene set obtained from all the four database:interrogation schemes. [2] The size of the corresponding SNP set. [3] The number of tests is the same as the size of the SNP set for the amino acids whereas for aggregated entities like the lipids and carnitine the SNP set is multiplied by the number of compounds present in that class.

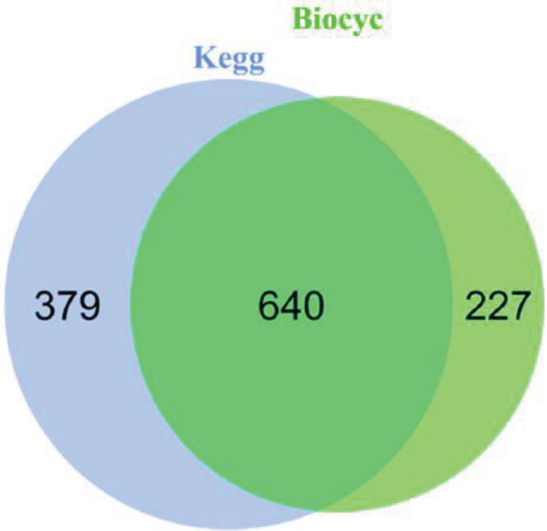


Figure 3 Gene set overlap for the KEGG and BioCyc databases. The Venn diagram depicts the overlap between the non-redundant gene set for KEGG and the BioCyc metabolic pathway database. These genes correspond to the combined set from the pathway and reaction interrogation schemes. The total number of unique genes that our method yields is 1246.

metabolite. At this stage, the sensitivity of the method is evaluated and potential novel discoveries are explored.

Results for each of three classes of metabolites (14 amino acids, 1 carnitine and 2 lipids) are shown in Table 1. For example, for glycine, interrogation of the KEGG database identified 173 and 432 genes using the pathway and reaction schemes respectively, whereas the corresponding numbers of genes were 90 and 192 for the BioCyc database. The union of all the four interrogation schemes results in a gene set consisting of 523 genes relevant to glycine metabolism (Table 1). For all the three classes of metabolites, 1246 unique genes were found, 640 are common to KEGG and BioCyc, the number of genes unique to each of the two databases are 379 and 227 respectively (Fig. 3).

Statistical Threshold

The number of unique SNPs generated for each of the metabolites is shown in Table 1. For aggregated metabolites like phosphatidylcholines, sphingomyelins and carnitines the size of the unique SNP set is multiplied by

Table 2 Performance of the database:interrogation schemes in GWAS dataset analysis

Database: interrogation scheme	Size of gene set[1]	Top hits from Illig et al. study identified by the method[2]	Sensitivity[3]
BioCyc pathway	399	*ACADL*, *ACADM*, *ACSL1*, *CPS1*, *FADS1*,	0.53
BioCyc reaction	806	*ACADM*, *ACADS*, *ACSL1*, *CPS1*, *FADS1*,	0.47
KEGG pathway	703	*ACADL*, *ACADM*, *ACADS*, *ACSL1*, *CPS1*,	0.67
KEGG reaction	768	*ACADL*, *ACADM*, *ACADS*, *ACSL1*, *CPS1*,	0.53
Pooled set	1246	*ACADL*, *ACADM*, *ACADS*, *ACSL1*, *CPS1*,	0.67

Snapshot of the matches between our method and the association data from the Illig et al. 2010 study for each of the database:interrogation scheme. [1]corresponds to the unique set of genes generated for all the metabolites for the given database:interrogation scheme. [2]corresponds to the top hits in the Illig et al. publication that were present in the gene set for the given database:interrogation scheme. [3]Sensitivity is a measure of the actual positives that have been captured by our method and is equal to the ratio of the number of top hits identified by the method over the total number of top hits in the Illig et al. publication which is 15.

the number of metabolites that fall within each class to yield the total number of tests. For example, the size of the unique SNP set for carnitine is 11,239; this is multiplied by the number of carnitines which is 41, to yield a total number of 460,799 tests for these compounds, as shown in the last column of Table 1. The sum of all SNPs derived from our set of metabolites is 3,835,543. The multiple testing threshold for metabolite concentrations using a Bonferroni correction at a nominal p-value of 0.05 is 1.3E-08 (0.05/3,835,543). In contrast, the p-value threshold for significant association of SNPs with the same metabolite concentrations in the Illig *et al* study would be 5.96E-10 (0.05/162*517,840). This represents a reduction of the multiple testing burden by about two orders of magnitude, regardless of the dependency between the SNPs or metabolites.

It has been demonstrated that GWAS of metabolite ratios offer robust statistical associations and point to biological mechanisms related to the interconversion of metabolite pairs. To investigate the association of SNPs with metabolite ratios, we generated the union of SNP sets for all combinations of metabolites (Table S3). In the case of aggregated metabolites like the lipids and carnitines, the union of the SNP set is multiplied by the number of compounds that fall within each class. For example, the union of the SNP set for arginine and carnitine is 20,000, this is multiplied by 41 to yield the total number of 820,000 tests for this group of ratios. The number of tests for ratios of compounds within classes such as phosphatidylcholines is equal to the size of the unique SNP set multiplied by the number of combinations, n*(n-1)/2, which in this case would be 92*91/2=4186. In choosing combinations of ratios, we have assumed that the association p-value for a linear regression model using a metabolite ratio of A/B is equivalent to that computed using it's reciprocal, B/A. The evidence for lack of independence of a ratio and its reciprocal is provided by the Illig *et al* study where a comparison of associations computed using untransformed and log-scaled ratios did not detect significant differences. This implies that we may consider the p-values computed using A/B and B/A to be approximately equal.

The sum of the number of tests for all ratios is 423,645,558 as shown in Table S3**.** The multiple testing threshold for the ratios using Bonferroni correction at nominal p-value of 0.05 is 1.18E-10. This represents a multiple threshold reduction by two orders of magnitude over the genome-wide threshold estimated by Illig *et al* which is 3.63E-12.

Proof of principle: Sensitivity

The sensitivity of the method was evaluated based on its ability to identify the top hits in the previously published Illig *et al* genome-wide association study. The overall sensitivity of the method as well as the interrogation specific breakdown is shown in Table 2. For example, for the BioCyc pathway scheme the size of the unique gene set generated for all the metabolites is shown to be 399. The number of genes that are among the 15 top hits in the Illig *et al* study for this database:interrogation scheme is 8 which results in a sensitivity measure of 0.53. A metabolite specific breakdown of each of these schemes and the genes with a p-value cut-off of 1E-02 is shown in supplementary table S5. Overall, combining the results from the four database:interrogation schemes helped identify 10 of the 15 top associations (67% sensitivity) published by Illig *et al*.

Novel Discovery in the Illig et al dataset

Analysis of the first stage or the "discovery stage" dataset of 1029 samples from the Illig *et al* dataset yielded several associations with p-values indicative for association, but that did not meet the significance threshold applied by Illig *et al*. Associations with p-value less than 1E-02 were evaluated in the combined "replication stage" dataset with 1809 samples. Analysis of SNPs in the *ALDH1L1* (aldehyde dehydrogenase family 1 L1) gene locus lowered the p-value of association with serine/glycine ratio from 4.83E-09 in the discovery dataset to 5.13E-12 in the combined dataset. This is well below

Table 3 Replication of candidate genes in the Demirkan et al. dataset

Gene	Trait	SNP from dataset	p-value[1]	SNP from the et al. dataset	p-value[2]	Combined p-value[3]
ADCY8	PC ae C40:6	rs11786743	4.03E-05	rs913819	6.73E-04	2.15E-07
CBS[*]	PC ae C40:6	rs2839631	5.67E-06	rs378376	5.17E-04	2.90E-08
CNR1	PC ae C38:2	rs10485168	2.42E-04	rs9359765	4.61E-04	7.54E-07
GPAM[*]	PC ae C34:3	rs2246253	1.25E-04	rs2419603	1.76E-04	1.56E-07
HSD17B12	PC aa C34:4	rs2862999	2.66E-05	rs11037685	6.13E-04	1.35E-07
MBOAT1	PC ae C40:6	rs9465673	1.11E-04	rs694094	4.47E-04	3.53E-07
PECR	PC aa C38:0	rs3770536	5.55E-04	rs3770562	9.43E-05	3.79E-07
PLCB1	PC aa C30:0	rs6056188	9.55E-06	rs17363114	1.96E-03	2.06E-07
TECR	PC aa C32:0	rs7252966	1.69E-05	rs7254215	2.09E-03	3.57E-07

Top hits from the meta-analysis of candidate genes identified in the Illig et al. study and replicated in the Demirkan et al. dataset. [1,2,3]p-value of association of the SNP with the trait in the Illig et al., Demirkan et al. and combined p-value respectively. *indicates genes for which further evidence was found.

our threshold of 1.18E-10, but above the threshold to be applied when considering all associations between SNPs and metabolite ratios. Furthermore, the original publication did not select this association for replication because of the threshold set in the first stage of the analysis. This is an example of the method pointing to potential true positives in a genome-wide scan and the association of *ALDH1L1* with the trait is being reported as a novel discovery.

Statistical threshold in the replication study

The analysis of the Illig *et al* dataset identified several biologically relevant candidate genes with p-values less than 1E-02. A list of 56 of these genes associated with phosphatidylcholines and sphingomyelins were investigated in an independent study in the GWAS dataset of phospholipids published by Demirkan *et al*. The number of matches between the two datasets was: 56 phosphatidylcholines and 6 sphingomyelins. Demirkan *et al* also performed GWAS for within class molar proportions for these moieties. We took these into consideration in addition to the GWAS of absolute concentrations. Therefore, the total number of metabolites and proportions investigated in the Demirkan *et al* GWAS dataset was 124. A principal component analysis based on the method proposed by Li *et al* [16] was performed on this set of metabolites resulting in 51 effectively independent variables. As we considered 2413 independent SNPs in the candidate loci for these metabolites, the statistical threshold, applying Bonferroni correction at a nominal p-value of 0.05, for the replication study was 4.06E-07 (0.05/2413*51).

Novel discoveries in the replication study

Table 3 shows the top hits in the meta-analysis of candidate genes identified in the Illig *et al* dataset for replication. The meta-analysis was performed using Stouffer's Z-score based method of combining p-values [17]. Since the SNPs in the loci replicated in the Demirkan *et al* dataset had relatively low r^2 values with the SNPs reported in the Illig *et al* dataset, we could not perform a traditional meta-analysis where strict linkage disequilibrium criteria are applied. Therefore, we combined the lowest p-value per gene and sought additional supporting evidence for potential allelic heterogeneity (see Discussion). As mentioned earlier, the p-value threshold for the replication study is set at 4.06E-07. SNPs in the vicinity of the genes *CBS, GPAM, ADCY8*, *CNR1*, *HSD17B12*, *MBOAT1*, *PECR*, *PLCB1* and *TECR* pass this threshold.

Discussion

Genome wide association studies with metabolites as phenotypes have identified several loci that explain human metabolic individuality. However, the large metabolite panel being tested results in a severe multiple testing burden that precludes genuine SNP-metabolite pairs from consideration when they fail to reach the stringent threshold for statistical significance. Our method aims to address this problem by selectively testing genes that operate in reactions and pathways relevant to the metabolite. The goal is to reduce the severity of the multiple testing burden and identify potential true positives in the list of genome-wide associations. Taverna, a workflow management system was used to generate the SNP-metabolite pairs. We have deposited the workflows at a repository called myexperiment.org, making it easier for the scientific community to interpret, repeat and reproduce the result. The sensitivity of the method, defined as retrieval of previously identified associations, is high, as evident from the proof of principle study carried out on the genome scan published by Illig *et al*. Replication studies on some of the promising SNP-metabolite pairs identified by the method pointed to a novel and statistically significant association at the *ALDH1L1* locus with serine/glycine ratios. Additional replication studies of phosphatidylcholines and sphingomyelins uncovered significant gene-wise associations with *CBS, GPAM, ADCY8*, *CNR1*, *HSD17B12*, *MBOAT1*, *PECR*, *PLCB1* and *TECR.*

Databases, interrogation schemes and software tool

The pathway databases have technical and conceptual differences [11] that mandate interrogation of multiple databases and integration of the results. Interpretation of these results requires a close coordination between biologists and computer scientists. Workflow management systems in general and Taverna [Supplementary section, S2] in particular is an example of a software tool that is intuitive enough for the biologist, while at the same time offering the flexibility for exploring the algorithmic aspects for the computer scientist [18]. In using Taverna as a software tool and depositing the workflows in the repository myexperiment.org, we have attempted to make the method and the rationale transparent to users, thus facilitating its retrieval, reuse and reproduction by other independent scientists [19].

Sensitivity of the method

As a sensitivity measure of our method, we evaluated its ability to pick the top hits in the Illig *et al* publication [2]. Some 60% of the top associations were identified successfully. A similar analysis of GWAS dataset published by Suhre *et al* [3] yielded a sensitivity of 54 % (20 out of 37 hits) (data not shown). However, 4 of the “misses” in the Suhre *et al* dataset were peptide fragments that do not have an entry in the pathway databases, which is a prerequisite for our method to work.

We interpret the high sensitivity of our method in three ways; first it reinforces the rationale that GWAS with metabolomic phenotypes provides a functional approach to the study of human genetic variation [1]. In other words, the known function of the associated gene and the biochemical characteristics of the affected metabolite support each other in ways that lends itself to a narrative on the underlying biological mechanism. Second, while the pathway databases have a long way to go in achieving a comprehensive annotation and delineation of biological processes, they, however, are a good resource of information in so far as the top hits in a GWAS with metabolomic phenotypes are concerned. Only two out of the 15 top hits in the study by Illig *et al* were genes with unknown functions (*PLEKHH1*, *SYNE2*), and two others were hitherto uncharacterized solute transporters (*SLC16A9*, *SLC22A4*). Third, a good sensitivity measure is a validation of our method and reflects its comprehensive data collection ability through integration of disparate data sources and utilization of appropriate interrogation strategies.

Novel Discoveries

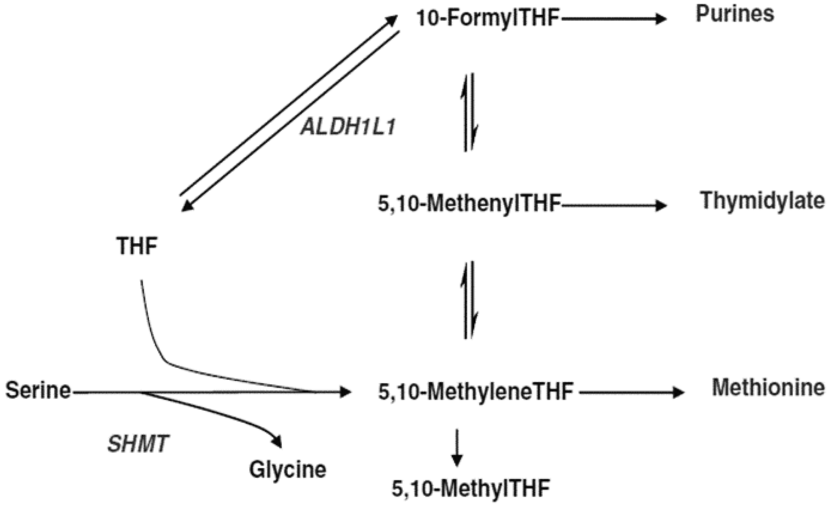


Figure 4 Role of ALDH1L1 in the cytosolic one-carbon pool metabolism. A simplified schematic of the one-carbon pool metabolism in the cytosol is depicted. ALDH1L1: Aldehyde Dehydrogenase 1 Family, Member L1; THF: tetrahydrofolate; SHMT: Serine hydrxymethyltransferase.

Our analysis of the GWAS dataset of the Illig *et al* publication based on the first step of the "discovery design" yielded several interesting associations that had not been reported among the top hits in the publication. We selected a few of the promising associations for replication in the combined dataset of 1809 subjects. One of the genes, Aldehyde dehydrogenase family 1 L1 (*ALDH1L1*) was found associated with the ratio of serine/glycine with a p-value of 5.13E-12 in the combined set of 1809 subjects. *ALDH1L1* also known as 10-formyltetrahydrofolate dehydrogenase (*10-FTHFDH*, *FDH*) catalyzes the NADP$^+$ dependent oxidation of 10-formyltetrahydrofolate to CO_2 and tetrahydrofolate (THF) [20] as shown in Figure 4. It plays an important role in folate metabolism [21, 22, 23, 24, 25]. Among other functions, *ALDH1L1* has been known to deplete cellular 10-formyltetrahydrofolate pool resulting in a loss of *de novo* purine biosynthesis [23], maintain cellular folate concentrations by regulating the availability of THF [22], but most importantly, it has been shown to compete with the enzyme serine hydroxymethyl transferase (*SHMT*) for the polyglutamyltetrahydrofolates [25] . The latter enzyme catalyzes the conversion of serine to glycine as shown in Fig 4. It has also been shown that glycine to serine inter-conversion by *SHMT* accounts for approximately 41% of whole body glycine flux inclusive of both mitochondrial and cytoplasmic processes [26].

To further investigate the potential of our approach to uncover novel genetic associations, we extended the analysis to an additional independent GWAS dataset [4]. Candidate genes identified in the Illig *et al* dataset in association with phosphatidylcholines and sphingomyelins were considered for replication in the dataset provided by Demirkan *et al* [4]. We discuss here two novel findings for which additional evidence was obtained.

SNPs near glycerol-3 phosphate acyltransferase (*GPAM)* are associated with PC ae C34:3 moieties in the Illig *et al* and Demirkan *et al* datasets with p-values of 1.25E-04 and 1.75E-04, respectively, with a meta-analysis p-value of 1.56E-07. *GPAM* encodes a mitochondrial protein that esterifies the acyl group from acyl-coA to the sn-1 position of glycerol-3-phosphate. It is a rate-limiting enzyme that catalyzes the initial step in the biosynthesis of triacylglycerols and phospholipids [27]. A recent study showed that in breast cancer, *GPAM* expression is strongly correlated with survival rates, clinico-pathological features as well as metabolomic and lipidomic profiles [28]. Interestingly, the study identified the metabolite PC C34:3 as the most significantly altered metabolite with respect to GPAM expression in breast cancer patients. This suggests that, for this particular example, genetic control is primarily at the level of gene expression, with secondary effects on enzyme levels and metabolic conversion rates. The example also highlights the potential influence of genetic variation of metabolic pathways on disease.

A large number of genes identified by our method in the context of phospholipids participate in fatty acid metabolism and are therefore likely to affect the levels of groups of phosphatidylcholines and sphingomyelins. For example, *GPAM* esterifies the acyl group from acyl-ACP to the sn-1 position of glycerol-3-phosphate, and is therefore relevant to both acyl-acyl and acyl-alkyl moieties. The lowest p-value of association, at this locus, with a phosphatidylcholine moiety in the Illig *et al* study is with PC ae C36:3, while in the Demirkan *et al* study it is PC aa C36:3. Since both associations make biological sense, future work should incorporate joint modelling of suitable phospholipid moieties to help identify loci that are biologically relevant but fail to reach the statistical threshold in GWAS analysis. We have reported such best case associations for phosphatidylcholines in Table S6.

SNPs near Cystathionine beta-synthase (*CBS)* are associated with PC ae C40:6 moieties in the Illig *et al* and Demirkan *et al* datasets with p-values of 5.67E-06 and 5.17E-04, respectively, with a meta-analysis p-value of 2.9E-08. Mutations in *CBS* cause hyperhomocysteinemia [29], which is marked by elevated levels of homocysteine. Several studies have associated altered phosphatidylcholine biosynthesis with hyperhomocysteinemia/*CBS* deficiency [30,31,32,33]. In one of the studies [30], phosphatidylcholine levels and the activity of the enzyme lecithin-cholesterol acyltransferase (*LCAT*) were significantly lower in *CBS* deficient mice than in wild type mice. While there is considerable literature evidence for the role of *CBS* in phosphatidylcholine metabolism, the stringent p-value threshold obscures this association in the list of GWAS results.

The low r^2 values for significant SNPs in *GPAM*, *CBS* and other loci between the Illig *et al* and Demirkan *et al* datasets could be explained by allelic heterogeneity. The latter is a phenomenon where multiple alleles from one gene influence a trait. However, in some cases it may be that the two apparently independent SNPs are tagging a third SNP [34]. This may be the case for the two SNPs (rs2839631, rs378376) near *CBS* which have an r^2 of 0.067 and are associated with C40:6 phosphatidylcholines in both the datasets. However, both SNPs are in LD with *cis*-eQTLs in the region (for example, rs719037, $r^2 \sim 0.4$). This is suggestive of the SNPs exerting their effect through the expression levels of the *CBS* enzyme, as was suggested for *GPAM*. Apparent allelic heterogeneity may preclude identification in a standard meta-analysis, but would justify further investigation of independent or dependent signals at loci showing this phenomenon.

Challenges and future direction

In general, our effort was directed at exploring the utility of machine-enabled interrogation of metabolic pathway databases in prioritizing SNP-metabolite associations in a GWAS dataset. While the method's sensitivity and ability to make novel discovery are encouraging, considerable progress needs to be made in metabolite disambiguation to achieve a relevant and comprehensive gene set for a given metabolite. This problem is particularly acute for phospholipids like phosphatidylcholines and sphingomyelins and various forms of the fatty acid transporters of L-carnitine. For example, the metabolomics technology used in the Illig *et al* study differentiated more than 90 forms of phosphatidylcholines based on alkyl or acyl bonds and single or double bonds on the side chains. However, the pathway databases do not yet contain information for the complex structures. This forces users to analyze these metabolites at a higher aggregation level. Another issue that requires attention is the bias introduced in selecting genes for inclusion in the gene set. We have formulated simple rules for interrogation [supplementary section, S1] that facilitates unbiased generation of gene sets for any given metabolite.

Another challenge arises due to the high correlation between metabolites, particularly the phospholipids like phosphatidylcholines and sphingomyelins. These moieties are associated with loci relevant to fatty acid metabolism. While the variation at these loci effects the levels of fatty acids and thereby the phospholipid pool, to a large extent, these loci are not specific for any particular phospholipid moiety. As a result, several loci exhibit a pleiotropic effect for biologically related metabolic phenotypes in general and phospholipids in particular [Shown in supplementary table S7] We have demonstrated that background knowledge and evidence-based approach is ideally suited to identify such candidate genes, however future work should focus on statistical methodologies with sufficient power to detect such pleiotropic loci in GWAS of intermediate phenotypes. In summary, future work includes integration of more pathway databases, metabolite disambiguation, consideration of allelic heterogeneity and multivariate statistical techniques that take into account the high degree of correlation between the metabolites.

Methods

GWAS data set for proof of principle studies

The GWAS dataset published by Illig *et al* 2010 [2] was used to evaluate the validity of the method. Illig *et al* employed a two-stage discovery design in the KORA F4 population cohort with 1029 male and female individuals in the first stage and 780 individuals in the second stage. Loci with p-value of association $<10^{-7}$ for metabolite concentrations and p-value $< 10^{-9}$ for concentration ratios were taken up for the second stage independent testing in 780 individuals. The joint p-values of association for all the 1809 individuals were then computed and 15 loci were reported whose strength of association increased after the second stage of the discovery process. The authors note that “although this approach is less well powered than a full genome-wide joint analysis, it reflects the historical way in which [they] selected SNPs for follow-up“. This means that if we can identify potential true positives using the 1029 samples, we can validate them in the full dataset, since this has not been done in the Illig *et al*. study for all hits with p-value $> 10^{-7}$ for metabolite concentrations and p-value $> 10^{-9}$ for concentration ratios.

Therefore, the GWAS dataset based on 1029 samples was analyzed for our proof of principle studies. Additionally, to evaluate novel associations identified by the method in the discovery stage dataset, the strength of the signal was assessed in the combined GWAS dataset for 1809 subjects.

GWAS dataset for follow-up studies

Candidate loci identified in the Illig et al dataset by our method were taken up for follow-up studies in the dataset published by Demirkan *et al*. The latter conducted a meta-analysis of GWAS on plasma levels of ceramides, phosphatidylcholines, lysophosphatidylcholines, sphingomyelins, phosphatidylethanolamines and plasmalogens in five European populations: the Erasmus Rucphen Family (ERF) study, conducted in the Netherlands, (2) the MICROS study from the Tyrol region in Italy, (3) the Northern Swedish Population Health Survey (NSPHS) in

Norrbotten, Sweden, (4) the Orkney Complex Disease Study (ORCADES) in Scotland, and (5) the CROAS (CROATIA_Vis) study conducted on Vis Island, Croatia. Broadly, the metabolite overlap between the Illig *et al* dataset and Demirkan A *et al* dataset was confined to the class of phosphatidylcholines, lysophosphatidylcholines and sphingomyelins. More specifically, the overlap represented 62 phospholipid moieties. Also, 56 candidate genes were identified for follow up in the Illig *et al* dataset. We choose to focus on SNPs in the flanking 50 kb region of these genes for the follow-up study in the Demirkan A *et al* dataset.

Metabolites considered for the generation of gene sets

Gene sets are defined as entities that participate in pathways and reactions relevant to the metabolite and hence hold the potential to influence its levels. The goal was to generate gene sets for the compounds that were measured in the Illig *et al* 2010 publication: 14 amino acids (Arginine, Glutamine, Glycine, Histidine, Methionine, Ornithine, Phenylalanine, Proline, Serine, Threonine, Tryptophan, Tyrosine, Valine, and Leucine), 41 Carnitines, 92 Phosphatidylcholines and 15 Sphingomyelins. In addition to the metabolites mentioned above, Illig *et al* also measured Hexose. We did not consider this metabolite for investigation because pathway information surrounding hexose is lacking. While metabolites like glucose and fructose could have been considered as proxies, we did not pursue this because of the enormous size of the resulting gene set, combined with a lack of confidence in the relevance of many of these genes to the metabolite measured by the metabolomics platform.

Pathway databases and interrogation schemes

The metabolic pathway databases KEGG (release 63) [12] and BioCyc (version 16) [13] were accessed for retrieving background knowledge surrounding metabolites. Two interrogation schemes were employed: pathway scheme and reaction scheme (Fig 1). In a pathway scheme, for a given metabolite, all the pathways that it participates in are determined followed by the retrieval of all the genes that participate in these pathways (Fig 1A). In a reaction scheme, given a metabolite, all the reactions that it is part of and the compounds that participate in these reactions are determined. The compounds obtained at this point are subjected to the same strategy as in the previous step in that all the reactions that these compounds participate in are determined. This can be visualized as expanding by a radius of 2 steps in the reaction space of every metabolite. Finally, the enzymes that drive all these reactions are determined (Fig 1B). As an intermediate step certain compounds were filtered out in order to avoid non-specific connections. The details about the filtration step and the compounds that were filtered are provided in the supplementary material, S4. In all there are four schemes: kegg:pathway, kegg:reaction, biocyc:pathway, and biocyc:reaction. The set of non-redundant genes combined from all the schemes then forms the gene set for any given metabolite.

Software used to generate gene and SNP sets

Taverna version 2.4 [14], a workflow management system was used to generate metabolite specific gene sets as well as for the generation of SNPs present in the 50kb flanking region of each gene. Taverna allows users access to remote data resources like KEGG, BioCyc, Ensembl, NCBI etc and data management systems like Biomart through implementation of web services. Each component in a workflow is responsible for a particular function and many such components need to be chained together in a pipeline to create a workflow that performs a certain task. The pipeline depicted in Fig 1 is implemented in a Taverna workflow through appropriate linking of remote web services and local scripts. Web services are software systems that facilitate machine to machine interaction over a network. Taverna allows the inclusion of different kinds of web services like Web Services Description Language (WSDL) and REpresentational State Transfer (REST). The services provided by the KEGG database were implemented using the REST services made available in the Taverna workbench. The BioCyc database was accessed through the REST interface using the BioVelo language. The latter is a query language designed to let the users write precise queries against the pathway/genome databases, available at BioCyc, to retrieve pathways, reactions, compounds, genes *etc*. All the workflows were designed following best practices for workflow design [35].

Workflow accessibility

To facilitate retrieval and reproducibility, the workflows have been deposited in a repository at http://www.myexperiment.org/packs/319.html. While the focus of this paper was on a specific set of metabolites; using appropriate identifiers from the KEGG or BioCyc database users will be able to generate gene sets for other metabolites. To generate a gene set for any metabolite using the KEGG or BioCyc database, users have to input the metabolite identifier for that database and the output is a text file containing the entrez gene identifiers. For example, to generate a gene set for the metabolite Arginine, for either the pathway or reaction scheme using the KEGG database, users input the KEGG identifier for Arginine: C00062. Similarly, to obtain a gene set using the BioCyc database, the input for the same metabolite is “L-arginine”. The workflows may also be repurposed to suit other objectives, for example, to filter out non-specific connections, we remove hub metabolites like ATP, NADP and other entities like co-enzymes; however, users may change the filtration criteria if they find it too stringent for their objectives. A detailed tutorial on how to access and run these workflows is provided in the supplementary section, S2.

Acknowledgements

This study is funded by the European Community’s Seventh Framework Programme (FP7/2007-2013) ENGAGE, the Centre for Medical Systems Biology (CMSB) and Netherlands Consortium for Systems Biology (NCSB), both within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) and the European Commission Seventh Framework Programme Wf4Ever (Digital Libraries and Digital Preservation area ICT-2009.4.1 project reference 270192). EUROSPAN consortium members are: Ayşe Demirkan, Cornelia M van Duijn, Peter Ugocsai, Aaron Isaacs, Peter P Pramstaller,Gerhard Liebisch, James F Wilson, Åsa Johansson, Igor Rudan, Yurii S Aulchenko, Anatoly V Kirichenko, A Cecile JW Janssens, Ritsert C Jansen, Carsten Gnewuch, Francisco S Domingues, Cristian Pattaro, Sarah H Wild, Inger Jonasson, Ozren Polasek, Irina V Zorkoltseva, Albert Hofman, Lennart Karssen, Maksim Struchalin, James Floyd, Wilmar Igl, Zrinka Biloglav, Linda Broer, Arne Pfeufer, Irene Pichler,Susan Campbell, Ghazal Zaboli, Ivana Kolcic, Fernando Rivadeneira, Jennifer Huffman, Nicholas D Hastie, Andre Uitterlinden, Lude

Franke, Christopher S Franklin, Veronique Vitart, Jacqueline CM Witteman, Tatiana Axenovich, Ben A Oostra, Thomas Meitinger, Andrew A Hicks, Caroline Hayward, Alan F Wright, Ulf Gyllensten, Harry Campbell, Gerd Schmitz.

References

1. Gieger C, Geistlinger L, Altmaier E, Hrabe de Angelis M, Kronenberg F, Meitinger T, Mewes HW, Wichmann HE, Weinberger KM, Adamski J *et al*: **Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum**. In: *PLoS Genet.* vol. 4, 2008/12/02 edn; 2008: e1000282.
2. Illig T, Gieger C, Zhai G, Romisch-Margl W, Wang-Sattler R, Prehn C, Altmaier E, Kastenmuller G, Kato BS, Mewes HW *et al*: **A genome-wide perspective of genetic variation in human metabolism**. *Nat Genet* 2010, **42**(2):137-141.
3. Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wagele B, Altmaier E, Deloukas P, Erdmann J, Grundberg E *et al*: **Human metabolic individuality in biomedical and pharmaceutical research**. *Nature* 2011, **477**(7362):54-60.
4. Demirkan A, van Duijn CM, Ugocsai P, Isaacs A, Pramstaller PP, Liebisch G, Wilson JF, Johansson A, Rudan I, Aulchenko YS *et al*: **Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations**. *PLoS Genet* 2012, **8**(2):e1002490.
5. Wang-Sattler R, Yu Z, Herder C, Messias AC, Floegel A, He Y, Heim K, Campillos M, Holzapfel C, Thorand B *et al*: **Novel biomarkers for pre-diabetes identified by metabolomics**. *Mol Syst Biol* 2012, **8**:615.
6. Xu T, Holzapfel C, Dong X, Bader E, Yu Z, Prehn C, Perstorfer K, Jaremek M, Roemisch-Margl W, Rathmann W *et al*: **Effects of smoking and smoking cessation on human serum metabolite profile: results from the KORA cohort study**. *BMC Med* 2013, **11**:60.
7. Suhre K, Gieger C: **Genetic variation in metabolic phenotypes: study designs and applications.** Nat Rev Genet. 2012 Nov;13(11):759-69.
8. Li M, Wang K, Grant SF, Hakonarson H, Li C: **ATOM: a powerful gene-based association test by combining optimally weighted markers**. *Bioinformatics* 2009, **25**(4):497-503.

9. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Hayward NK, Montgomery GW, Visscher PM, Martin NG *et al*: **A versatile gene-based test for genome-wide association studies**. *Am J Hum Genet* 2010, **87**(1):139-145.
10. Wang K, Li M, Bucan M: **Pathway-based approaches for analysis of genomewide association studies**. *Am J Hum Genet* 2007, **81**(6):1278-1283.
11. Stobbe MD, Houten SM, Jansen GA, van Kampen AH, Moerland PD: **Critical assessment of human metabolic pathway databases: a stepping stone for future integration**. *BMC Syst Biol* 2011, **5**:165.
12. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Res* 2010, **38**(Database issue):D355-360.
13. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD: **Computational prediction of human metabolic pathways from the complete human genome**. *Genome Biol* 2005, **6**(1):R2.
14. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P *et al*: **The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud**. *Nucleic Acids Res* 2013, **41**(Web Server issue):W557-561.
15. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P *et al*: **myExperiment: a repository and social network for the sharing of bioinformatics workflows**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W677-682.
16. Li J, Ji L: **Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix**. *Heredity (Edinb)* 2005, **95**(3):221-227.
17. Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM Jr.: **Adjustment during army life**. *Princeton*, *NJ*, *Princeton University Press*, 1949.
18. Moore JH, Asselbergs FW, Williams SM: **Bioinformatics challenges for genome-wide association studies**. *Bioinformatics* 2010, **26**(4):445-455.
19. Mesirov JP: **Computer science. Accessible reproducible research.** *Science* 2010 **327**(5964):415-416.
20. Cook RJ, Lloyd RS, Wagner C: **Isolation and characterization of cDNA clones for rat liver 10-formyltetrahydrofolate dehydrogenase. *J Biol Chem* 1991, 266(8):4965-4973.**

21. Anguera MC, Field MS, Perry C, Ghandour H, Chiang EP, Selhub J, Shane B, Stover PJ: **Regulation of folate-mediated one-carbon metabolism by 10-formyltetrahydrofolate dehydrogenase**. *J Biol Chem* 2006, **281**(27):18335-18342.
22. Krebs HA, Hems R, Tyler B: **The regulation of folate and methionine metabolism**. *Biochem J* 1976, **158**(2):341-353.
23. Fu TF, Maras B, Barra D, Schirch V: **A noncatalytic tetrahydrofolate tight binding site is on the small domain of 10-formyltetrahydrofolate dehydrogenase**. *Arch Biochem Biophys* 1999, **367**(2):161-166.
24. Krupenko SA, Oleinik NV: **10-formyltetrahydrofolate dehydrogenase, one of the major folate enzymes, is down-regulated in tumor tissues and possesses suppressor effects on cancer cells. *Cell Growth Differ* 2002, 13(5):227-236.**
25. Kim DW, Huang T, Schirch D, Schirch V: **Properties of tetrahydropteroylpentaglutamate bound to 10-formyltetrahydrofolate dehydrogenase**. *Biochemistry* 1996, **35**(49):15772-15783.
26. Lamers Y, Williamson J, Gilbert LR, Stacpoole PW, Gregory JF, 3rd: **Glycine turnover and decarboxylation rate quantified in healthy men and women using primed, constant infusions of [1,2-(13)C2]glycine and [(2)H3]leucine**. *J Nutr* 2007, **137**(12):2647-2652.
27. Wendel AA, Lewin TM, Coleman RA: **Glycerol-3-phosphate acyltransferases: rate limiting enzymes of triacylglycerol biosynthesis**. *Biochim Biophys Acta* 2009, **1791**(6):501-506.
28. Brockmoller SF, Bucher E, Muller BM, Budczies J, Hilvo M, Griffin JL, Oresic M, Kallioniemi O, Iljin K, Loibl S *et al*: **Integration of metabolomics and expression of glycerol-3-phosphate acyltransferase (GPAM) in breast cancer-link to patient survival, hormone receptor status, and metabolic profiling**. *J Proteome Res* 2012, **11**(2):850-860.
29. Beard RS, Jr., Bearden SE: **Vascular complications of cystathionine beta-synthase deficiency: future directions for homocysteine-to-hydrogen sulfide research**. *Am J Physiol Heart Circ Physiol* 2011, **300**(1):H13-26.
30. She QB, Hayakawa T, Tsuge H: **Alteration in the phosphatidylcholine biosynthesis of rat liver microsomes caused by vitamin B6 deficiency.** *Biosci Biotechnol Biochem* 1995, **59**(2):163-167.
31. Namekata K, Enokido Y, Ishii I, Nagai Y, Harada T, Kimura H: **Abnormal lipid metabolism in cystathionine beta-synthase-deficient mice, an animal model for hyperhomocysteinemia**. *J Biol Chem* 2004, **279**(51):52961-52969.

32. Devlin AM, Singh R, Wade RE, Innis SM, Bottiglieri T, Lentz SR: **Hypermethylation of Fads2 and altered hepatic fatty acid and phospholipid metabolism in mice with hyperhomocysteinemia**. *J Biol Chem* 2007, **282**(51):37082-37090.
33. Ikeda K, Kubo A, Akahoshi N, Yamada H, Miura N, Hishiki T, Nagahata Y, Matsuura T, Suematsu M, Taguchi R *et al*: **Triacylglycerol/phospholipid molecular species profiling of fatty livers and regenerated non-fatty livers in cystathionine beta-synthase-deficient mice, an animal model for homocysteinemia/homocystinuria**. *Anal Bioanal Chem* 2011, **400**(7):1853-1863.
34. Wood AR, Hernandez DG, Nalls MA, Yaghootkar H, Gibbs JR, Harries LW, Chong S, Moore M, Weedon MN, Guralnik JM *et al*: **Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association**. Hum Mol Genet. 2011 Oct 15;20(20):4082-92.
35. Hettne KM, Wolstencroft K, Belhajjame K, Goble CA, Mina E, Dharuri H, De Roure D, Verdes-Montenegro L, Garrido J, Roos M: Best Practices for Workflow Design: How to Prevent Workflow Decay. Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences, Paris, France, November 28-30, 2012,CEUR-WS.org, Volume 952.

Supplementary section

S1.

Rules to generate Metabolite-Gene sets:

1. Metabolic pathway databases accessed via Taverna workflows: KEGG, Biocyc
2. Interrogation Scheme:

a. Pathway Scheme (KEGG:Pathway and Biocyc:Pathway): Given a metabolite determine all the pathways it participates in and pull all the genes that participate in these pathways (Fig 2A).
    i. For KEGG consider only metabolic pathways.
    ii. Phosphatidylcholine and sphingomyelins contain fatty acids in their side chains. Genes that are involved in fatty acid metabolism alter the levels of phospholipids. Previously published GWAS datasets have shown that most of the significant genes associated with phospholipids, for example *FADS1,* are involved in fatty acid metabolism. To incorporate such genes we ran the pathway and the reaction scheme on various fatty acids and incorporated them into the gene set for phosphatidylcholines and sphingomyelins. For the biocyc:pathway scheme, gene set for phosphatidylcholine and sphingomyelin contained genes generated for: arachidonate, a fatty acid, laurate, linoleate, a lipid, a long chain fatty acid, octanoate, oleate, palmitate, a 2,3,4 saturated fatty acid and stearate. Similarly, for the kegg:pathway scheme the gene sets for the following compounds were included in the set for phosphatidylcholine and sphingomyelin: arachidonic acid (C00219) and palmitic acid (C00249). The choice of fatty acids was based on whether pathway information was available for the compounds for the database being considered.
    iii. Biocyc:Pathway for carnitines included the following compounds: L-Carnitines, Palmitoylcarnitine, L-Ocatnoylcarnitine, O-acetylcarnitine, butanoyl-CoA, decanoyl-CoA, lauroyl-CoA, myristoyl-CoA,octanoyl-CoA,

palmitoyl-CoA, and stearoyl-CoA. The rationale behind this inclusion is provided below in the “Reaction Scheme”.

b. Reaction Scheme (KEGG:Reaction and Biocyc:Reaction): Given a compound find all the reactions and the compounds that participate in these reactions. The reactions that these compounds participate in and the enzymes that drive these reactions are determined (Fig 2B).

    i. Compounds that make too general connections are filtered out. List of filtered compounds for Kegg and Biocyc database interrogation provided.

    ii. For the biocyc:reaction scheme gene sets for the following compounds were included in the set for phosphatidylcholine and sphingomyelin: (9Z)-12,13-dihydroxyoctadeca-9-enoate, octanoate, laurate, decanoate, stearate, palmitate, oleate, myristate, linoleate, arachidonate, arachidate, a long-chain fatty acid, a fatty acid, a phospholipid, and a lipid. For the kegg:reaction scheme the following fatty acids were included for the purpose of generating a gene set for phosphatidylcholines and sphingomyelins: arachidonic acid (C00219), linoleic acid (C01595), palmitic acid (C00249) and stearic acid (C01530). The choice of fatty acids was based on whether reaction information was available for the compounds for the database being considered.

    iii. Biocyc compounds are structured as classes. We have established the following rules for interrogation:

        a. In general for amino acids we do not consider the super class and we don’t have to deal with child terms.

        b. Phosphatidylcholine and Sphingomyelin, are considered as “class” terms. Sphingomyelin does not have child terms, phosphatidylcholine does have child terms but the reactions in the database are not considered at the level of the latter. For both phosphatidylcholine and sphingomyelin the parent terms (a phosphoglyceride and a sphingolipid respectively) are not considered for interrogation.

c. For carnitines, we combine results for the L-carnitine, L-Octanoylcarnitine,L-palmitoylcarnitine, O-acetylcarnitine (these are the only carnitines present in the database, there is considerable overlap in terms of genes returned for the four compounds). Two of the compounds returned for these interrogations: palmitoyl-CoA and octanoyl-CoA are instances of “a 2,3,4 saturated fatty acid” and since some of the reactions are given at the level of the parent class, we have also included “a 2,3,4 saturated fatty acid” in the Biocyc:Reaction interrogation scheme.

The various instances of “a 2,3,4 saturated fatty acid” are: butanoyl-CoA,decanoyl-CoA,lauroyl-CoA,myristoyl-CoA,octanoyl-CoA,palmitoyl-CoA,stearoyl-CoA. These are all the acyl fatty acids that are transported by carnitine for mitochondrial fatty acid beta oxidation. The corresponding esters were all measured by Illig et al, therefore it was decided to consider all the above instances of “a 2,3,4 saturated fatty acid” to generate the gene set for Carnitine.

In a nutshell then, we have four schemes that yield genes that operate in the vicinity of a given metabolite: Kegg:Pathway, Kegg:Reaction, Biocyc:Pathway, Biocyc:Reaction. As an example, Kegg:Pathway means employing the pathway scheme as mentioned above on the Kegg database. Table 1 displays the yield of genes for each database:interrogation scheme. The gene set for a metabolite is the integration of all the genes coming out of the four schemes into a non-redundant set as shown in the last column of Table 1 in the publication. The sum of all such non-redundant set equals 4801 genes. The total number of unique genes that came out of all the schemes and all the metabolites is 1246 with the number of unique genes from Kegg being 379 and those from Biocyc being 227 and 640 genes present in both databases (Fig 3 of the publication).

S2
Taverna workflow management system

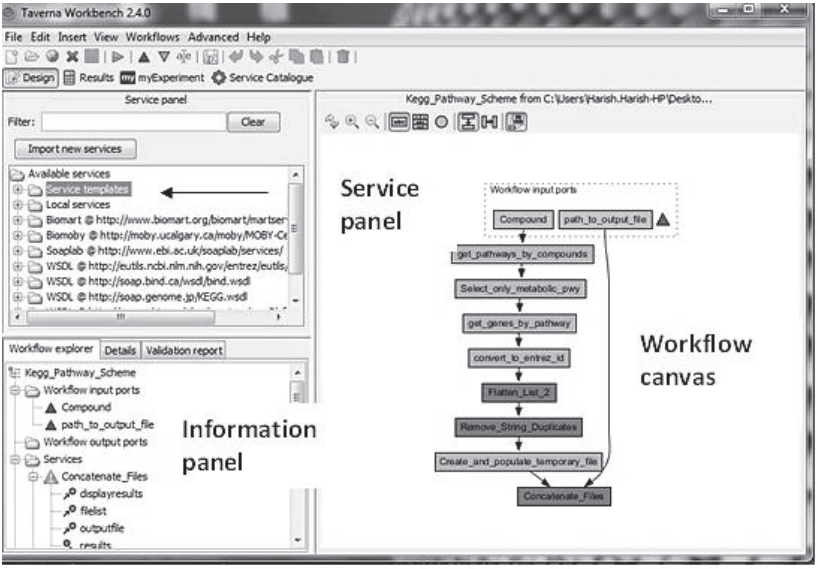


Fig 1 Snapshot of the Taverna workbench which consists of three panels as pointed to in the figure

Workflow management system is a software environment designed to compose and execute a series of computational or data manipulation steps. Taverna workbench is an example of a workflow management system that provides a desktop environment for accomplishing bioinformatic tasks. Taverna allows users access to data sources and analysis tools made available by institutions like NCBI, DDBJ, EBI etc through web services. In addition to making available third-party services, Taverna offers a suite of shim services that run on the local computer and are essentially used for data manipulation.

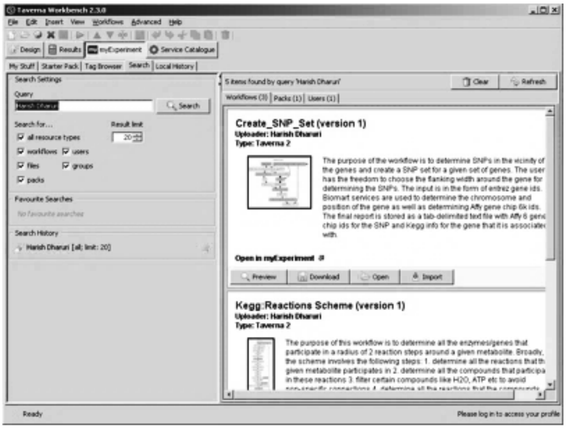
A Taverna workflow is a directed acyclic graph consisting of components (web or shim services) having various functionalities chained together appropriately to perform a useful task. Figure 1 shows a snapshot of the Taverna workbench that consists of three panels: a service panel at the top left that makes available third-party services and also a few local services that are included by default, the panel on the right showing the workflow diagram is a space where workflows can be created by pulling services from the left panel in a drag and drop fashion. Existing workflows can be opened in the workflow canvas using the open tab. Later in this section, a tutorial on how to open workflows stored on myExperiments.org is provided. The panel in the bottom left is known as the workflow explorer which depicts the workflow in a tree like fashion and allows editing of properties of the components of the workflow.

Downloading the Taverna workflow management system

Instructions for downloading Taverna can be found at: http://www.taverna.org.uk/download/
The tutorials to learn about features and how to run Taverna are available at the Taverna web site: http://www.taverna.org.uk/documentation/taverna-2-x/quick-start-guide/

Tutorial: Download workflow from myExperiment.org, learn more about workflow functionality and run the workflows

1. Click on myExperiment tab → followed by the 'Search' button.
2. Type in the name 'Harish Dharuri' in the query and click the Search button
3. This should display the models that we have submitted to myExperiment in the right panel.
4. Choose the model that you would like to run and click the 'Open' button.



5 items found by query 'Harish Dharuri'

Create_SNP_Set (version 1)
Uploader: Harish Dharuri
Type: Taverna 2

The purpose of the workflow is to determine SNPs in the vicinity of the genes and create a SNP set for a given set of genes. The user has the freedom to choose the flanking width around the gene for determining the SNPs. The input is in the form of entrez gene ids. Biomart services are used to determine the chromosome and position of the gene as well as determining Affy gene chip 6k ids. The final report is stored as a tab-delimited text file with Affy 6 gene chip ids for the SNP and Kegg info for the gene that it is associated with.

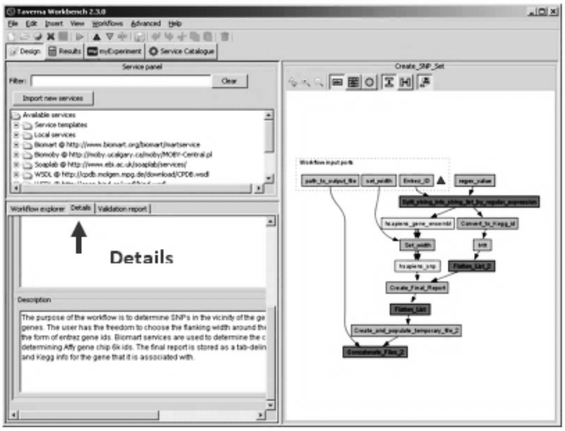
Open in myExperiment

Kegg:Reactions Scheme (version 1)
Uploader: Harish Dharuri
Type: Taverna 2

The purpose of this workflow is to determine all the enzymes/genes that participate in a radius of 2 reaction steps around a given metabolite. Broadly, the scheme involves the following steps: 1. determine all the reactions that th
given metabolite participates in 2. determine all the compounds that participa
in these reactions 3. filter certain compounds like H2O, ATP etc to avoid

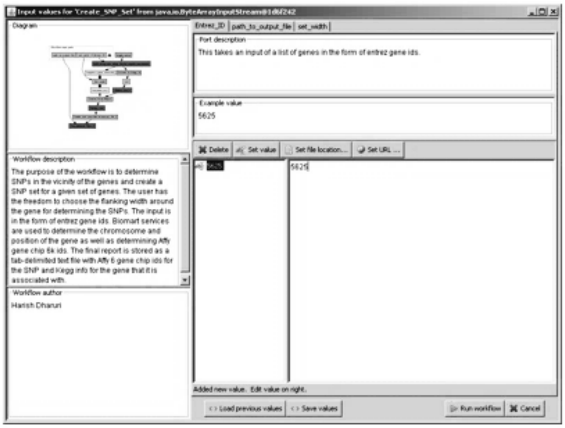
Please log in to access your profile

5. This will open the workflow in the 'Design' mode of Taverna.
6. Click the 'Details' to read the description of the workflow.
7. Click on any input ports and read the annotation in 'Details' to know about input type required.
8. Run the workflow by clicking this button ▶



The purpose of the workflow is to determine SNPs in the vicinity of the ge
genes. The user has the freedom to choose the flanking width around th
the form of entrez gene ids. Biomart services are used to determine the c
determining Affy gene chip 6k ids. The final report is stored as a tab-delin
and Kegg info for the gene that it is associated with.

9. This will open the 'input dialog' to enter the input values.
10. Click on the tabs at the top-right of the panel for each of input to know the example value.
11. Click 'Set value' and change/enter a value.
12. Press the 'Run workflow' button when you are done entering all the values.



Port description

This takes an input of a list of genes in the form of entrez gene ids.

Example value

5625

Workflow description

The purpose of the workflow is to determine SNPs in the vicinity of the genes and create a SNP set for a given set of genes. The user has the freedom to choose the flanking width around the gene for determining the SNPs. The input is in the form of entrez gene ids. Biomart services are used to determine the chromosome and position of the gene as well as determining Affy gene chip 6k ids. The final report is stored as a tab-delimited text file with Affy 6 gene chip ids for the SNP and Kegg info for the gene that it is associated with.

Workflow author

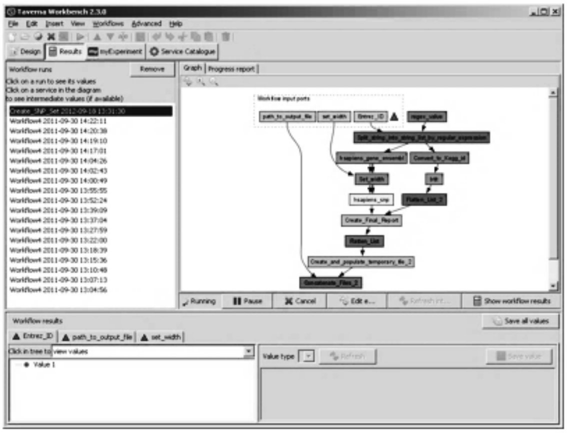
Harish Dharuri

5625

5625

Added new value. Edit value on right.

13. This will open the workflow in the 'Result' mode of Taverna.
14. Watch the progress of your run in the default 'Graph' mode as shown in the picture on the right.



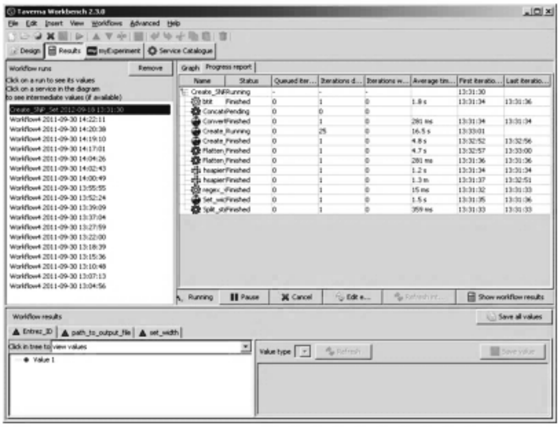
Click on a run to see its values
Click on a service in the diagram
to see intermediate values (if available)

Create_SNP_Set 2012-09-18 13:31:30
Workflow4 2011-09-30 14:22:11
Workflow4 2011-09-30 14:20:38
Workflow4 2011-09-30 14:19:10
Workflow4 2011-09-30 14:17:01
Workflow4 2011-09-30 14:04:26
Workflow4 2011-09-30 14:02:43
Workflow4 2011-09-30 14:00:49
Workflow4 2011-09-30 13:55:55
Workflow4 2011-09-30 13:52:24
Workflow4 2011-09-30 13:39:09
Workflow4 2011-09-30 13:37:04
Workflow4 2011-09-30 13:27:59
Workflow4 2011-09-30 13:22:00
Workflow4 2011-09-30 13:18:39
Workflow4 2011-09-30 13:15:36
Workflow4 2011-09-30 13:10:48
Workflow4 2011-09-30 13:07:13
Workflow4 2011-09-30 13:04:56

Click in tree to view values

Value 1

15. Or click the 'Progress Report' tab to see the progress of the run in tabular mode of the Result mode as shown in the picture at the right.
16. At the end of the run the results will be stored as a text file in a path provided as input by the user.



Click on a run to see its values
Click on a service in the diagram
to see intermediate values (if available)

Create_SNP_Set 2012-09-18 13:31:30
Workflow4 2011-09-30 14:22:11
Workflow4 2011-09-30 14:20:38
Workflow4 2011-09-30 14:19:10
Workflow4 2011-09-30 14:17:01
Workflow4 2011-09-30 14:04:26
Workflow4 2011-09-30 14:02:43
Workflow4 2011-09-30 14:00:49
Workflow4 2011-09-30 13:55:55
Workflow4 2011-09-30 13:52:24
Workflow4 2011-09-30 13:39:09
Workflow4 2011-09-30 13:37:04
Workflow4 2011-09-30 13:27:59
Workflow4 2011-09-30 13:22:00
Workflow4 2011-09-30 13:18:39
Workflow4 2011-09-30 13:15:36
Workflow4 2011-09-30 13:10:48
Workflow4 2011-09-30 13:07:13
Workflow4 2011-09-30 13:04:56

Name	Status	Queued iter...	Iterations d...	Iterations w...	Average tim...	First iteratio...	Last iteratio...
Create_SNP	Running	-	-	-		13:31:30	
bbt	Finished	0	1	0	1.8 s	13:31:34	13:31:36
Concat	Pending	0	0	0			
Convert	Finished	0	1	0	281 ms	13:31:34	13:31:34
Create_	Running	0	25	0	16.5 s	13:33:01	
Create_	Finished	0	1	0	4.8 s	13:32:52	13:32:56
Flatten_	Finished	0	1	0	4.7 s	13:32:57	13:33:00
Flatten_	Finished	0	1	0	281 ms	13:31:36	13:31:36
hsapien	Finished	0	1	0	1.2 s	13:31:34	13:31:34
hsapien	Finished	0	1	0	1.3 m	13:31:37	13:32:51
regex_	Finished	0	1	0	15 ms	13:31:32	13:31:33
Set_wi	Finished	0	1	0	1.5 s	13:31:35	13:31:36
Split_st	Finished	0	1	0	359 ms	13:31:33	13:31:33

Click in tree to view values

Value 1

Alternate way: The workflows may be downloaded from: http://www.myexperiment.org/packs/319.html and run in Taverna.

S3: SNP set generated for ratios of metabolites

Metabolite Ratio	Union Set[1]	Number of tests[2]
Arginine/Carnitine	20000	820000
Arginine/Glutamine	20340	20340
Arginine/Glycine	26234	26234
Arginine/Histidine	14743	14743
Arginine/Leucine	13133	13133
Arginine/Methionine	18982	18982
Arginine/Ornithine	14049	14049
Arginine/Phenylalanine	14577	14577
Arginine/Phosphatidylcholine	44228	4068976
Arginine/Proline	11897	11897
Arginine/Serine	23269	23269
Arginine/Sphingomyelin	33739	506085
Arginine/Threonine	12362	12362
Arginine/Tryptophan	16681	16681
Arginine/Tyrosine	15865	15865
Arginine/Valine	15865	15865
Carnitine/Glutamine	23428	960548
Carnitine/Glycine	27500	1127500
Carnitine/Histidine	16172	663052
Carnitine/Leucine	13595	557395
Carnitine/Methionine	20612	845092
Carnitine/Ornithine	19817	812497
Carnitine/Phenylalanine	17971	736811
Carnitine/Phosphatidylcholine	37658	142045976
Carnitine/Proline	14282	585562
Carnitine/Serine	23453	961573
Carnitine/Sphingomyelin	27892	17153580
Carnitine/Threonine	13388	548908
Carnitine/Tryptophan	19031	780271
Carnitine/Tyrosine	17279	708439
Carnitine/Valine	17279	708439
Glutamine/Glycine	28589	28589
Glutamine/Histidine	19521	19521
Glutamine/Leucine	17644	17644

Glutamine/Methionine	23790	23790
Glutamine/Ornithine	19818	19818
Glutamine/Phenylalanine	17948	17948
Glutamine/Phosphatidylcholine	46176	4248192
Glutamine/Proline	17411	17411
Glutamine/Serine	27052	27052
Glutamine/Sphingomyelin	35706	535590
Glutamine/Threonine	17943	17943
Glutamine/Tryptophan	21007	21007
Glutamine/Tyrosine	19595	19595
Glutamine/Valine	19595	19595
Glycine/Histidine	23648	23648
Glycine/Leucine	23462	23462
Glycine/Methionine	24339	24339
Glycine/Ornithine	24918	24918
Glycine/Phenylalanine	25261	25261
Glycine/Phosphatidylcholine	46133	4244236
Glycine/Proline	23758	23758
Glycine/Serine	28932	28932
Glycine/Sphingomyelin	38072	571080
Glycine/Threonine	22578	22578
Glycine/Tryptophan	24651	24651
Glycine/Tyrosine	25633	25633
Glycine/Valine	25633	25633
Histidine/Leucine	8881	8881
Histidine/Methionine	13622	13622
Histidine/Ornithine	14270	14270
Histidine/Phenylalanine	12238	12238
Histidine/Phosphatidylcholine	38491	3541172
Histidine/Proline	9457	9457
Histidine/Serine	21504	21504
Histidine/Sphingomyelin	29802	447030
Histidine/Threonine	9257	9257
Histidine/Tryptophan	12229	12229
Histidine/Tyrosine	13034	13034
Histidine/Valine	13034	13034
Leucine/Methionine	13718	13718

Leucine/Ornithine	12389	12389
Leucine/Phenylalanine	10449	10449
Leucine/Phosphatidylcholine	37936	3490112
Leucine/Proline	7614	7614
Leucine/Serine	19786	19786
Leucine/Sphingomyelin	27194	407910
Leucine/Threonine	7201	7201
Leucine/Tryptophan	12209	12209
Leucine/Tyrosine	9862	9862
Leucine/Valine	9862	9862
Methionine/Ornithine	18044	18044
Methionine/Phenylalanine	16695	16695
Methionine/Phosphatidylcholine	41910	3855720
Methionine/Proline	14314	14314
Methionine/Serine	23487	23487
Methionine/Sphingomyelin	33290	499350
Methionine/Threonine	13097	13097
Methionine/Tryptophan	16107	16107
Methionine/Tyrosine	17375	17375
Methionine/Valine	17375	17375
Ornithine/Phenylalanine	13651	13651
Ornithine/Phosphatidylcholine	42666	3925272
Ornithine/Proline	11728	11728
Ornithine/Serine	23324	23324
Ornithine/Sphingomyelin	32150	482250
Ornithine/Threonine	11974	11974
Ornithine/Tryptophan	16502	16502
Ornithine/Tyrosine	14655	14655
Ornithine/Valine	14655	14655
Phenylalanine/Phosphatidylcholine	41835	3848820
Phenylalanine/Proline	11052	11052
Phenylalanine/Serine	21438	21438
Phenylalanine/Sphingomyelin	31347	470205
Phenylalanine/Threonine	10338	10338
Phenylalanine/Tryptophan	13228	13228
Phenylalanine/Tyrosine	12489	12489
Phenylalanine/Valine	12489	12489

Phosphatidylcholine/Proline	38810	3570520
Phosphatidylcholine/Serine	42395	3900340
Phosphatidylcholine/Sphingomyelin	34731	47928780
Phosphatidylcholine/Threonine	35955	3307860
Phosphatidylcholine/Tryptophan	41184	3788928
Phosphatidylcholine/Tyrosine	41940	3858480
Phosphatidylcholine/Valine	41940	3858480
Proline/Serine	20006	20006
Proline/Sphingomyelin	28006	420090
Proline/Threonine	7155	7155
Proline/Tryptophan	13016	13016
Proline/Tyrosine	11811	11811
Proline/Valine	11811	11811
Serine/Sphingomyelin	31738	476070
Serine/Threonine	16657	16657
Serine/Tryptophan	22076	22076
Serine/Tyrosine	22311	22311
Serine/Valine	22311	22311
Sphingomyelin/Threonine	25277	379155
Sphingomyelin/Tryptophan	32681	490215
Sphingomyelin/Tyrosine	31476	472140
Sphingomyelin/Valine	31476	472140
Threonine/Tryptophan	11676	11676
Threonine/Tyrosine	11305	11305
Threonine/Valine	11305	11305
Tryptophan/Tyrosine	15561	15561
Tryptophan/Valine	15561	15561
Tyrosine/Valine	9633	9633
Carnitine/Carnitine	11239	9215980
Phosphatidylcholine/Phosphatidylcholine	31676	132595736
Sphingomyelin/Sphingomyelin	21290	2235450
Total	**2969397**	**423645558**

[1] is the union of the SNP set generated for the metabolites in the numerator and denominator of the corresponding ratio. [2] In the case of aggregated compounds, the SNP set is multiplied by the number of compounds present in that class.

S4
Compounds filtered for the Kegg:Reaction Scheme

cpd:C00001 H2O; Water

cpd:C00002 ATP; Adenosine 5'-triphosphate

cpd:C00003 NAD+; NAD; Nicotinamide adenine dinucleotide; DPN; Diphosphopyridine nucleotide; Nadide

cpd:C00004 NADH; DPNH; Reduced nicotinamide adenine dinucleotide

cpd:C00005 NADPH; TPNH; Reduced nicotinamide adenine dinucleotide phosphate

cpd:C00006 NADP+; NADP; Nicotinamide adenine dinucleotide phosphate; beta-Nicotinamide adenine dinucleotide phosphate; TPN; Triphosphopyridine nucleotide

cpd:C00007 Oxygen; O2

cpd:C00008 ADP; Adenosine 5'-diphosphate

cpd:C00009 Orthophosphate; Phosphate; Phosphoric acid; Orthophosphoric acid

cpd:C00010 CoA; Coenzyme A; CoA-SH

cpd:C00011 CO2; Carbon dioxide

cpd:C00012 Peptide

cpd:C00013 Diphosphate; Diphosphoric acid; Pyrophosphate; Pyrophosphoric acid; PPi

cpd:C00014 NH3; Ammonia

cpd:C00015 UDP; Uridine 5'-diphosphate

cpd:C00016 FAD; Flavin adenine dinucleotide

cpd:C00019 S-Adenosyl-L-methionine; S-Adenosylmethionine; AdoMet; SAM

cpd:C00020 AMP; Adenosine 5'-monophosphate; Adenylic acid; Adenylate; 5'-AMP; 5'-Adenylic acid; 5'-Adenosine monophosphate; Adenosine 5'-phosphate

cpd:C00024 Acetyl-CoA; Acetyl coenzyme A

cpd:C00027 Hydrogen peroxide; H2O2; Oxydol

cpd:C00028 Acceptor; Hydrogen-acceptor; A; Oxidized donor

cpd:C00030 Reduced acceptor; AH2; Hydrogen-donor; Donor

cpd:C00033 Acetate; Acetic acid; Ethanoic acid

cpd:C00035 GDP; Guanosine 5'-diphosphate; Guanosine diphosphate

cpd:C00040 Acyl-CoA; Acyl coenzyme A

cpd:C00044 GTP; Guanosine 5'-triphosphate

cpd:C00046 RNA; RNAn; RNAn+1; RNA(linear); (Ribonucleotide)n; (Ribonucleotide)m; (Ribonucleotide)n+m; Ribonucleic acid

cpd:C00055 CMP; Cytidine-5'-monophosphate; Cytidylic acid

cpd:C00063 CTP; Cytidine 5'-triphosphate; Cytidine triphosphate

cpd:C00067 Formaldehyde; Methanal; Oxomethane; Oxomethylene; Methylene oxide; Formalin
cpd:C00075 UTP; Uridine 5'-triphosphate; Uridine triphosphate
cpd:C00080 H+; Hydron
cpd:C00084 Acetaldehyde; Ethanal
cpd:C00086 Urea; Carbamide
cpd:C00091 Succinyl-CoA; Succinyl coenzyme A
cpd:C00105 UMP; Uridylic acid; Uridine monophosphate; Uridine 5'-monophosphate; 5'Uridylic acid
cpd:C00106 Uracil
cpd:C00112 CDP; Cytidine 5'-diphosphate; Cytidine diphosphate
cpd:C00125 Ferricytochrome c; Cytochrome c3+
cpd:C00126 Ferrocytochrome c; Cytochrome c2+; Reduced cytochrome c
cpd:C00131 dATP; 2'-Deoxyadenosine 5'-triphosphate; Deoxyadenosine 5'-triphosphate; Deoxyadenosine triphosphate
cpd:C00144 GMP; Guanosine 5'-phosphate; Guanosine monophosphate; Guanosine 5'-monophosphate; Guanylic acid
cpd:C00147 Adenine; 6-Aminopurine
cpd:C00151 L-Amino acid; L-2-Amino acid
cpd:C00161 2-Oxo acid; 2-Oxocarboxylate
cpd:C00162 Fatty acid
cpd:C00177 Cyanide; Prussiate; CN-; Cyano
cpd:C00212 Adenosine
cpd:C00178 Thymine; 5-Methyluracil
cpd:C00206 dADP; 2'-Deoxyadenosine 5'-diphosphate
cpd:C00214 Thymidine; Deoxythymidine
cpd:C00239 dCMP; Deoxycytidylic acid; Deoxycytidine monophosphate; Deoxycytidylate; 2'-Deoxycytidine 5'-monophosphate
cpd:C00240 rRNA; Ribosomal RNA
cpd:C00227 Acetyl phosphate
cpd:C00242 Guanine; 2-Amino-6-hydroxypurine
cpd:C00286 dGTP; 2'-Deoxyguanosine 5'-triphosphate; Deoxyguanosine 5'-triphosphate; Deoxyguanosine triphosphate
cpd:C00288 HCO3-; Bicarbonate; Hydrogencarbonate; Acid carbonate
cpd:C00299 Uridine
cpd:C00330 Deoxyguanosine; 2'-Deoxyguanosine
cpd:C00360 dAMP; 2'-Deoxyadenosine 5'-phosphate; 2'-Deoxyadenosine 5'-monophosphate; Deoxyadenylic acid; Deoxyadenosine monophosphate
cpd:C00361 dGDP; 2'-Deoxyguanosine 5'-diphosphate

cpd:C00362 dGMP; 2'-Deoxyguanosine 5'-monophosphate; 2'-Deoxyguanosine 5'-phosphate; Deoxyguanylic acid; Deoxyguanosine monophosphate
cpd:C00363 dTDP; Deoxythymidine 5'-diphosphate
cpd:C00364 dTMP; Thymidine 5'-phosphate; Deoxythymidine 5'-phosphate; Thymidylic acid; 5'-Thymidylic acid; Thymidine monophosphate; Deoxythymidylic acid; Thymidylate
cpd:C00365 dUMP; Deoxyuridylic acid; Deoxyuridine monophosphate; Deoxyuridine 5'-phosphate; 2'-Deoxyuridine 5'-phosphate
cpd:C00380 Cytosine
cpd:C00387 Guanosine
cpd:C00458 dCTP; Deoxycytidine 5'-triphosphate; Deoxycytidine triphosphate; 2'-Deoxycytidine 5'-triphosphate
cpd:C00459 dTTP; Deoxythymidine triphosphate; Deoxythymidine 5'-triphosphate; TTP
cpd:C00460 dUTP; 2'-Deoxyuridine 5'-triphosphate
cpd:C00475 Cytidine
cpd:C00512 S-Benzoate coenzyme A; Benzoyl-CoA
cpd:C00526 Deoxyuridine; 2-Deoxyuridine; 2'-Deoxyuridine
cpd:C00533 Nitric oxide; NO; Nitrogen monoxide
cpd:C00559 Deoxyadenosine; 2'-Deoxyadenosine
cpd:C00575 3',5'-Cyclic AMP; Cyclic adenylic acid; Cyclic AMP; Adenosine 3',5'-phosphate; Adenosine 3',5'-cyclic phosphate; cAMP
cpd:C00698 Cl-; Chloride; Chloride ion
cpd:C00705 dCDP; 2'-Deoxycytidine diphosphate; 2'-Deoxycytidine 5'-diphosphate
cpd:C00821 DNA adenine
cpd:C00856 DNA cytosine; Cytosine (in DNA)
cpd:C00881 Deoxycytidine; 2'-Deoxycytidine
cpd:C00941 3',5'-Cyclic CMP; Cytidine 3',5'-cyclic monophosphate
cpd:C00942 3',5'-Cyclic GMP; Guanosine 3',5'-cyclic monophosphate; Guanosine 3',5'-cyclic phosphate; Cyclic GMP; cGMP
cpd:C00943 3',5'-Cyclic IMP; Inosine 3',5'-cyclic monophosphate
cpd:C00968 3',5'-Cyclic dAMP
cpd:C01021 Aromatic amino acid; Aromatic L-amino acid
cpd:C01346 dUDP; 2'-Deoxyuridine 5'-diphosphate
cpd:C01352 FADH2
cpd:C01642 tRNA(Gly)
cpd:C01647 tRNA(Met)

cpd:C01764 tRNA containing uridine at position 54; tRNA UpsiC
cpd:C01794 Choloyl-CoA; 3alpha,7alpha,12alpha-Trihydroxy-5beta-cholanoyl-CoA; 3alpha,7alpha,12alpha-Trihydroxy-5beta-cholan-24-one-CoA
cpd:C01977 tRNA guanine
cpd:C02353 2',3'-Cyclic AMP
cpd:C02354 2',3'-Cyclic CMP
cpd:C02355 2',3'-Cyclic UMP
cpd:C02412 Glycyl-tRNA(Gly)
cpd:C02430 L-Methionyl-tRNA; L-Methionyl-tRNA(Met)
cpd:C02507 3',5'-Cyclic dGMP
cpd:C03110 DNA N4-methylcytosine
cpd:C03391 DNA 6-methylaminopurine
cpd:C03446 tRNA containing ribothymidine at position 54; tRNA TpsiC
cpd:C03395 Fatty acid methyl ester
cpd:C04152 rRNA containing N1-methylguanine
cpd:C04153 rRNA containing N2-methylguanine
cpd:C04154 rRNA containing N6-methyladenine; rRNA(N6-methyladenine)
cpd:C04156 tRNA containing N1-methyladenine
cpd:C04157 tRNA containing N1-methylguanine
cpd:C04158 tRNA containing N2-methylguanine
cpd:C04159 tRNA containing N6-methyladenine
cpd:C04160 tRNA containing N7-methylguanine
cpd:C04268 dTDP-4-amino-4,6-dideoxy-D-glucose
cpd:C04545 tRNA containing 2'-O-methylguanosine
cpd:C04728 tRNA containing 5-methylaminomethyl-2-thiouridylate; tRNA containing mnm5s2U
cpd:C04779 rRNA containing a single residue of 2'-O-methyladenosine
cpd:C05167 alpha-Amino acid
cpd:C05198 5'-Deoxyadenosine
cpd:C05337 Chenodeoxycholoyl-CoA; 3alpha,7alpha-Dihydroxy-5beta-cholanoyl-CoA
cpd:C05338 4-Hydroxyphenylacetyl-CoA
cpd:C05777 Coenzyme F430; Factor F430
cpd:C05359 e-; Electron
cpd:C06194 2',3'-Cyclic GMP
cpd:C11378 Ubiquinone-10; Ubidecarenone; Coenzyme Q10
cpd:C15670 Heme A
cpd:C15672 Heme O
cpd:C15817 Heme C

cpd:C11478 tRNA containing 5-aminomethyl-2-thiouridine; tRNA containing nm5s2U
cpd:C17023 Sulfur donor; S-donor
cpd:C17322 tRNA containing 2-thiouridine; tRNA containing s2U
cpd:C17323 tRNA containing 5-carboxymethylaminomethyl-2-thiouridine; tRNA containing cnm5s2U
cpd:C17324 tRNA adenine
cpd:C19637 Coenzyme M; 2,2'-Dithiodiethanesulfonic acid

S5. Metabolite specific break-up of the performance of database:interrogation schemes

Arginine	Size of Gene Set	Sensitivity	Genes with SNPs < 1E-02	Number of Genes with SNPs < 1E-02	Percentage of genes at 1E-02
Biocyc_Pathway	20			0	0
Biocyc_Reaction	104		SLC32A1	1	0.96
Kegg_Pathway	57			0	0
Kegg_Reaction	179			0	0
Carnitine	Size of Gene Set	Sensitivity	Genes with SNPs < 1E-02	Number of Genes with SNPs < 1E-02	Percentage of genes at 1E-02
Biocyc_Pathway	32	0.3	ACADL,ACADM,ACSL1	3	9.38
Biocyc_Reaction	206	0.4	ABHD6,ACADM,ACADS,ACSL1,AGPAT4,CRAT,DHTKD1,FADS1,FADS2,GPAM,HMGCS2,IDH3B,MCCC1,NMT2,P4HA2,PLA2G2A,PLA2G2E,SCD,SLC27A6,XYLT1	20	9.71
Kegg_Pathway	81	0.4	ACADL,ACADM,ACADS,ACSL1,ADH1A,ADH1B,ADH1C,ADH7	8	9.88

Kegg_Reaction	94	0.3	ACADL,ACADM,ACSL1,BAAT,CRAT,DHTKD1,IDH3B,P4HA2,PECR,PHGDH	10	10.64
Glycine	Size of Gene Set	Sensitivity	Genes with SNPs < 1E-02	Number of Genes with SNPs < 1E-02	Percentage of genes at 1E-02
Biocyc_Pathway	90	0.0		0	0
Biocyc_Reaction	192	0.0	ALDH1L1	1	0.52
Kegg_Pathway	173	1.0	CPS1	1	0.57
Kegg_Reaction	432	0.0	ALDH1L1	1	0.23
Ornithine	Size of Gene Set	Sensitivity	Genes with SNPs < 1E-02	Number of Genes with SNPs < 1E-02	Percentage of genes at 1E-02
Biocyc_Pathway	16			0	0
Biocyc_Reaction	150			0	0
Kegg_Pathway	103			0	0
Kegg_Reaction	159	1.0	PHGDH	1	0.63

Phosphatidylcholine	Size of Gene Set	Sensitivity	Genes with SNPs < 1E-02	Number of Genes with SNPs < 1E-02	Percentage of genes at 1E-02
Biocyc_Pathway	188	0.4	ACSL1,AGPAT1,FADS1,FADS2,GPAM,LRAT,MOGAT1,PLA2G4E,PLCB1,PLD1,PLD2,PNLIP,PPAPDC1A,PTGIS,RBP4,RLBP1,SCD	17	9.04
Biocyc_Reaction	361	0.4	ACSL1,ADH7,ADPRM,AGPAT1,ATP8A1,ATP8A2,ATP8B4,CBS,DGKQ,FADS1,FADS2,GPAM,LRAT,MBOAT1,MOGAT1,PLA2G4E,PLA2G7,PLCB1,PLD1,PLD2,PNLIP,PNPLA6,PPAPDC1A,PPT2,PTDSS1,PTGIS,RBP4,RLBP1,SCD,SLC27A6,SPTLC3,XYLT1	32	8.86
Kegg_Pathway	312	0.6	ACADM,ACOT1,ACSL1,ADCY8,ADCY9,ADH7,ADPRM,AGPAT1,BAAT,CACNA1C,CNR1,DAGLA,DGKQ,ELOVL2,FADS1,FADS2,GABRB1,GABRB2,GABRR3,GNB4,GNGT1,GPAM,HSD17B12,KCNJ3,KCNJ6,MBOAT1,PECR,PLA2G4E,PLCB1,PLD1,PLD2,PNPLA	40	12.82

			6,PPT2,PRKCA,PRKCB,PTDSS1,PTGIS,SCD,SLC32A1,TECR		
Kegg_Reaction	343	0.2	AADAC,ACADM,ACOT1,ACSL1,ADH7,ADPRM,AGPAT1,AHCYL2,BAAT,CBS,CERS4,DGKQ,ENPP2,GPAM,LRAT,MBOAT1,MLL,MOGAT1,PECR,PLA2G4E,PLCB1,PLD1,PLD2,PNLIP,PNLIPRP1,PNPLA6,PPT2,PTGIS,SCD,SOAT1,SOAT2,SPTLC3	32	9.33
Serine	Size of Gene Set	Sensitivity	Genes with SNPs < 1E-02	Number of Genes with SNPs < 1E-02	Percentage of genes at 1E-02
Biocyc_Pathway	37	1.0	PHGDH	1	2.7
Biocyc_Reaction	135			0	0
Kegg_Pathway	152	1.0	PHGDH,PSPH	2	1.32
Kegg_Reaction	219		ACOT1,PECR,PSPH	3	1.37

Sphingomyelin	Size of Gene Set	Sensitivity	Genes with SNPs < 1E-02	Number of Genes with SNPs < 1E-02	Percentage of genes at 1E-02
Biocyc_Pathway	160		ACSL1,FADS1,FADS2,LRAT,PLCB1,PLD1,PLD2,PNLIP,PTGIS,SCD	10	6.25
Biocyc_Reaction	331	1.0	ACSL1,ATP8A1,DGKQ,FADS1,FADS2,LRAT,PLCB1,PLD1,PLD2,PNLIP,PPAPDC1A,PTGIS,SCD,SPTLC3	14	4.23
Kegg_Pathway	189	1.0	ACSL1,CERS4,ELOVL2,FADS1,FADS2,HSD17B12,PTGIS,SCD,SPTLC3,TECR	10	5.29
Kegg_Reaction	241	1.0	ACSL1,CERS4,DGKQ,ENPP2,LRAT,PLCB1,PLD1,PLD2,PNLIP,PNLIPRP1,PTGIS,SCD,SOAT1,SPTLC3	14	5.81

Chapter 3: Structuring research methods and data with the Research Object model: genomics workflows as a case study

Kristina M Hettne
Harish Dharuri
Jun Zhao
Katherine Wolstencroft
Khalid Belhajjame
Stian Soiland-Reyes
Eleni Mina
Mark Thompson
Don Cruickshank
Lourdes Verdes-Montenegro
Julian Garrido
David de Roure
Oscar Corcho
Graham Klyne
Reinout van Schouwen
Peter A.C. 't Hoen
Sean Bechhofer
Carole Goble
Marco Roos

Journal of Biomedical Semantics. 2014; 5(1):41

Abstract

Background

One of the main challenges for biomedical research lies in the computer-assisted integrative study of large and increasingly complex combinations of data in order to understand molecular mechanisms. The preservation of the materials and methods of such computational experiments with clear annotations is essential for understanding an experiment, and this is increasingly recognized in the bioinformatics community. Our assumption is that offering means of digital, structured aggregation and annotation of the objects of an experiment will provide necessary meta-data for a scientist to understand and recreate the results of an experiment. To support this we explored a model for the semantic description of a workflow-centric Research Object (RO), where an RO is defined as a resource that aggregates other resources, e.g., datasets, software, spreadsheets, text, etc. We applied this model to a case study where we analysed human metabolite variation by workflows.

Results

We present the application of the workflow-centric RO model for our bioinformatics case study. A set of workflows were produced following recently defined Best Practices for workflow design. By modelling the experiment as an RO, we were able to automatically query the experiment and answer questions such as “which particular data was input to a particular workflow to test a particular hypothesis?”, and “which particular conclusions were drawn from a particular workflow?”.

Conclusions

Applying a workflow-centric RO model to aggregate and annotate the resources used in a bioinformatics experiment, allowed us to retrieve the conclusions of the experiment in the context of the driving hypothesis, the executed workflows and their input data. The RO model is an extendable reference model that can be used by other systems as well.

Background

One of the main challenges for biomedical research lies in the integrative study of large and increasingly complex combinations of data in order to understand molecular mechanisms, for instance to explain the onset and progression of human diseases. Computer-assisted method- ology is needed to perform these studies, posing new challenges for upholding scientific quality standards for the reproducibility of science. The aim of this paper is to describe how the research data, methods and metadata related to a workflow-centric computational experiment can be aggregated and annotated using standard Semantic Web technologies, with the purpose of helping scientists performing such experiments in meeting requirements for understanding, sharing, reuse and repurposing.

The workflow paradigm is gaining ground in bioinformatics as the technology of choice for recording the steps of computational experiments [1-4]. It allows scientists to delineate the steps of a complex analysis and expose this to peers using workflow design and execution tools such as Taverna [5], and Galaxy [6], and workflow sharing platforms such as myExperiment [7] and crowdLabs [8]. In a typical workflow, data outputs are generated from data inputs via a set of (potentially distributed) computational tasks that are coordinated following a workflow definition. However, workflows do not provide a complete solution for aggregating all data and all meta-data that are necessary for understanding the full context of an experiment. Consequently, scientists often find it difficult (or impossible) to reuse or repurpose existing workflows for their own analyses [9]. In fact, insufficient meta-data has been listed as one of the main causes of workflow decay in a recent study of Taverna workflows on myExperiment [9]. Workflow decay is the term used when the ability to re-execute a workflow after its inception has been compromised.

We will be able to better understand scientific workflows if we are able to capture more relevant data and meta-data about them; including the purpose and context of the experiment, sample input and output datasets, and the provenance of workflow executions. Moreover, if we wish to publish and exchange these resources as a unit, we need a mechanism for aggregation and annotation that would work in a broad scientific community. Semantic Web technology seems a logic choice of technology, given its focus on capturing the meaning of data in a machine readable format that is extendable and supports interoperability. It allows defining a Web-accessible reference model for the annotation of the aggregation and the aggregated resources that is independent of how data are stored in repositories. Examples of other efforts where Semantic Web technology has been used for the biomedical data integration includes the Semantic Enrichment of the Scientific Literature (SESL) [10] and Open PHACTS [11] projects. We applied the recently developed Research Object (RO) family of tools and ontologies [12,13] to preserve the scientific assets and their annotation related to a computational experiment. The concept of the RO was first proposed as an abstraction for sharing research investigation results [14]. Later, the potential role for ROs in facilitating not only the sharing but also the reuse of results, in order to increase the reproducibility of these results, was envisioned [15]. Narrowing down to workflow- centric ROs, preservation aspects were explored in [16], and their properties as first class citizen structures that aggregate resources in a principled manner in [13]. We also showed the principle of describing a (text mining) workflow experiment and its results by Web Ontology Language (OWL) ontologies [17]. The OWL ontologies were custom built, which we argue is now an unnecessary bottleneck for exchange and interoperability. These studies all contributed to the understanding and implementation of the concept of an RO, but the data used were preliminary, and the studies were focused on describing workflows with related datasets and provenance information, rather than from the viewpoint of describing a scientific experiment of which workflows are a component.

A workflow-centric RO is defined as a resource that aggregates other resources, such as workflow(s), provenance, other objects and annotations. Consequently, an RO represents the method of analysis and all its associated materials and meta-data [13,15], distinguishing it from other work mainly focusing on provenance of research data [18,19]. Existing Semantic Web frameworks are used, such as (i) the Object Exchange and Reuse (ORE) model [20]; (ii) the Annotation Ontology (AO) [21]; and (iii) the W3C-recommended provenance exchange models [22]. ORE defines the standards for the description and exchange of aggregations of Web resources and provides the basis for the RO ontologies. AO is a general model for annotating resources and is used to describe the RO and its constituent resources as well as the relationships between them. The W3C provenance exchange models enable the interchange of provenance information on the Web, and the Provenance Ontology (PROV-O) forms the basis for recording the provenance of scientific workflow executions and their results.

In addition, we used the minimal information model “Minim”, also in Semantic Web format, to specify which elements in an RO we consider “must haves”, “should haves” and “could haves” according to user-defined requirements [23]. A checklist service subsequently queries the Minim annotations as an aid to make sufficiently complete ROs [24]. The idea of using a checklist to perform quality assessment is inspired by related checklist based approaches in bioinformatics, such as the Minimum Information for Biological and Biomedical Information (MIBBI)-style models [25].

Case study: genome wide association studies

As real-world example we aggregate and describe the research data, methods and metadata of a computational experiment in the context of studies of genetic variation in human metabolism. Given the potential of genetic variation data in extending our understanding of genetic diseases, drug development and treatment, it is crucial that the steps leading to new biological insights can be properly recorded and understood. Moreover, bioinformatics approaches typically involve aggregation of disparate online resources into complex data parsing pipelines. This makes this a fitting test case for an instantiated RO. The biological goal of the experiment is to aid in the interpretation of the results of a Genome-Wide Association Study (GWAS) by relating metabolic traits to the Single Nucleotide Polymorphisms (SNPs) that were identified by the GWAS. GWA studies have successfully identified genomic regions that dispose individuals to diseases (see for example [26], for a review see [27]). However, the underlying biological mechanisms often remain elusive, which led the research community to evince interest in genetic association studies of metabolites levels in blood (see for example [28-30]). The motivation is that the biochemical characteristics of the metabolite and the functional nature of affected genes can be combined to unravel biological mechanisms and gain functional insight into the aetiology of a disease. Our specific experiment involves mining curated pathway databases and a specific text mining method called concept profile matching [31,32].

In this paper we describe the current state of RO ontologies and tools for the aggregation and annotation of a computational experiment that we developed to elucidate the genetic basis for human metabolic variation.

Methods

We performed our experiment using workflows developed in the open source Taverna Workflow Management System version 2.4 [5]. To improve the understanding of the experiment, we have added the following additional resources to the RO, using the RO-enabled myExperiment [33]: 1) the hypothesis or research question (what the experiment was designed to test); 2) a workflow-like sketch of the overall experiment (the overall data flow and workflow aims); 3) one or more workflows encapsulating the computational method; 4) input data (a record of the data that were used to reach the conclusions of an experiment); 5) provenance of workflow runs (the data lineage paths built from the workflow outputs to the originating inputs); 6) the results (a compilation of output data from workflow runs); 7) the conclusions (interpretation of the results from the workflows against the original hypothesis). Such an RO was then stored in the RO Digital Library [34]. RO completeness evaluation is checked from myExperiment with a tool implementing the Minim model [24]. Detailed description of the method follows.

Workflow development

We developed three workflows for interpreting SNP- metabolite associations from a previously published genome-wide association study, using pathways from the KEGG metabolic pathway database [35] and Gene Ontology (GO) [36] biological process associations from text mining of PubMed. To understand an association of a SNP with a metabolite, researchers would like to know the gene in the vicinity of the SNP that is affected by the polymorphism. Then, researchers examine the functional nature of the gene and evaluate if it makes sense given the biochemical characteristics of the metabolite with which it is associated. This typically involves interrogation of biochemical pathway databases and mining existing literature. We would like to evaluate the utility of background knowledge present in the databases and literature in facilitating a biological interpretation of the statistically significant SNP-metabolite pairs. We do this by first determining the genes closest to the SNPs, and then reporting the pathways that these genes participate in. We implemented two main workflows for our experiment. The first one mines the manually curated KEGG database of metabolic pathway and gene associations that are available via the KEGG REST Services [37]. The second workflow mines the text-mining based database of associations between GO biological processes and genes behind the Anni 2.1 tool [31] that are available via the concept profile mining Web services [38]. We also created a workflow to list all possible concept sets in the concept profile database, to encourage reuse of the concept profile-based workflow for matching against other concept sets than GO biological processes. The workflows were developed following the 10 Best Practices for workflow design [39]. The Best Practices were developed to encourage re-use and prevent workflow decay, and briefly consists of the following steps:

1) Make a sketch workflow to help design the overall data flow and workflow aims, and to identify the tools and data resources required at each

stage. The sketch could be created using for example flowchart symbols or empty beanshells in Taverna.

2) Use modules, i.e. implement all executable components as separate, runnable workflows to make it easier for other scientists to reuse parts of a workflow at a later date.

3) Think about the output. A workflow has the potential to produce masses of data that need to be visualized and managed properly. Also, workflows can be used to integrate and visualise data as well as for analysing it, so one should consider how the results will be presented easily to the user.

4) Provide input and output examples to show the format of input required for the workflow and the type of output that should be produced. This is crucial for the understanding, validation, and maintenance of the workflow.

5) Annotate, i.e. choose meaningful names for the workflow title, inputs, outputs, and for the processes that constitute the workflow as well as for the interconnections between the components, so that annotations are not only a collection of static tags but capture the dynamics of the workflow. Accurately describing what individual services do, what data they consume and produce, and the aims of the workflow are all essential for use and reuse.

6) Make it executable from outside the local environment by for example using remote Web services, or platform independent code/plugins. Workflows are more reusable if they can be executed from anywhere. If there is need to use local services, library or tools, then the workflow should be annotated in order to define its dependencies.

7) Choose services carefully. Some services are more reliable or more stable than others, and examining which are the most popular can assist with this process.

8) Reuse existing workflows by for example searching collaborative platforms such as myExperiment for workflows using the same Web service. If a workflow has been tried, tested and published, then reusing it can save a significant amount of time and resource.

9) Test and validate by defining test cases and implementing validation mechanisms in order to understand the limitations of workflows, and to monitor changes to underlying services.

10) Advertise and maintain by publishing the workflow on for example myExperiment, and performing frequent testing of the workflow and

monitoring of the services used. Others can only reuse it if it is accessible and if it is updated when required, due to changes in underlying services.

The RO core model

The RO model [12,13] aims at capturing the elements that are relevant for interpreting and preserving the results of scientific investigations, including the hypothesis investigated by the scientists, the data artefacts used and generated, as well as the methods and experiments employed during the investigation. As well as these elements, to allow third parties to understand the content of the RO, the RO model caters for annotations that describe the elements encapsulated by the ROs, as well as the RO as a whole. Therefore, two main constructs are at the heart of the RO model, namely aggregation and annotation. The work reported on in this article uses version 0.1 of the RO model, which is documented online [12].

Following myExperiment packs [7], ROs use the ORE model [20] to represent aggregation. Using ORE, an RO is defined as a resource that aggregates other resources, e.g., datasets, software, spreadsheets, text, etc. Specifically, the RO extends ORE to define three new concepts: i) ro: ResearchObject is a sub-class of ore:Aggregation which represents an aggregation of resources. ii) ro:Resource is a sub-class of ore:AggregatedResource representing a resource that is aggregated within an RO. iii) ro:Manifest is a sub-class of ore:ResourceMap, representing a resource that is used to describe the RO.

To support the annotation of ROs, their constituent resources, as well as their relationship, we use the Annotation Ontology [21]. Several types of annotations are supported by the Annotation Ontology, e.g., comments, textual annotations (classic tags) and semantic annotations, which relate elements of the ROs to concepts from underlying domain ontologies. We make use of the following Annotation Ontology terms: i) ao:Annotation, which acts as a handle for the annotation. ii) ao:annotatesResource, which represents the resource(s)/RO(s) subjects to annotation. iii) ao:body, which describes the target of the annotation. The body of the annotation takes the form of a set of Resource Description Framework (RDF) statements. Note that it is planned for later revisions of the RO model to use the successor of AO, the W3C Community Open Annotation Data Model (OA) [40]. For our purposes, OA annotations follows a very similar structure using oa: Annotation, oa:hasTarget and oa:hasBody.

Support for workflow-centric ROs

A special kind of ROs that are supported by the model is what we call workflow-centric ROs, which, as indicated by the name, refer to those ROs that contain resources that are workflow specifications. The structure of the workflow in ROs is detailed using the wfdesc vocabulary [41], and is defined as a graph in which the nodes refers to steps in the workflow, which we call wfdesc:Process, and the edges representing data flow dependencies, wfdesc:DataLink, which is a link between the output and input parameters (wfdesc:Parameter) of the processes that compose the workflow. As well as the description of the workflow, workflow centric ROs support the specification of the workflow runs, wfprov:WorkflowRun, that are obtained as a result of enacting workflows. A workflow run is specified using the wfprov ontology [42], which captures information about the input used to feed the workflow execution, the output results of the workflow run, as well as the constituent process runs, wfprov: ProcessRun, of the workflow run, which are obtained by invoking the workflow processes, and the input and outputs of those process runs.

Support for domain-specific information

A key aspect of the RO model design is the freedom to use any vocabulary. This allows for inclusion of very domain-specific information about the RO if that serves the desired purpose of the user. We defined new terms under the name space roterms [43]. These new terms serve two main purposes. They are used to specify annotations that are, to our knowledge, not catered for by existing ontologies, e.g., the classes roterms:Hypothesis and roterms:Conclusion to annotate the hypothesis and conclusions part of an RO, and the property roterms: exampleValue to annotate an example value for a given input or output parameter given as an roterms:WorkflowValue instance. The roterms are also used to specify shortcuts that make the ontology easy to use and more accessible. For example, roterms:inputSelected associates a wfdesc:WorkflowDefinition to an ro:Resource to state that a file is meant to be used with a given workflow definition, without specifying at which input port or in which workflow run.

Minim model for checklist evaluation

When building an RO in myExperiment users are provided with a mechanism of quality insurance by our so-called checklist evaluation tool, which is built upon the Minim checklist ontology [23,44] and defined using Web Ontology Language. Its basic function is to assess that all required information and

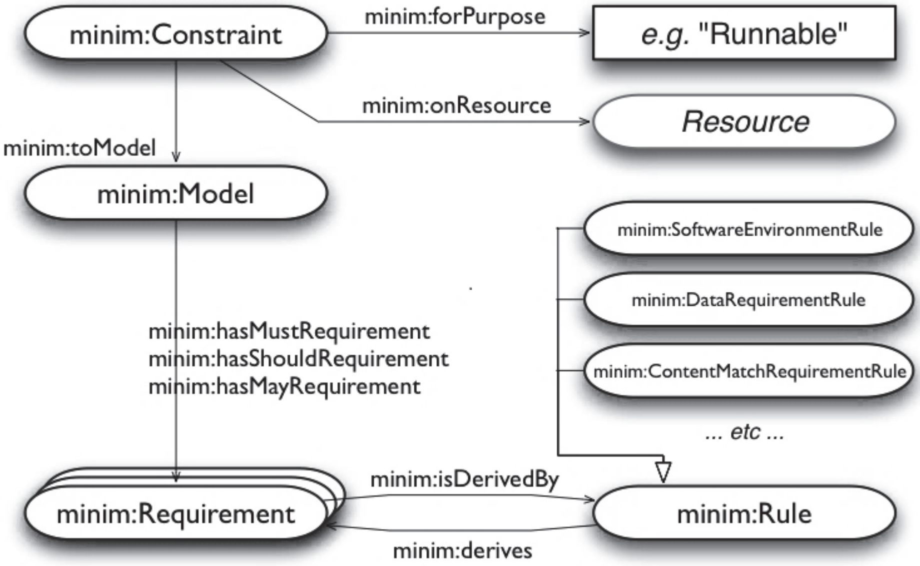
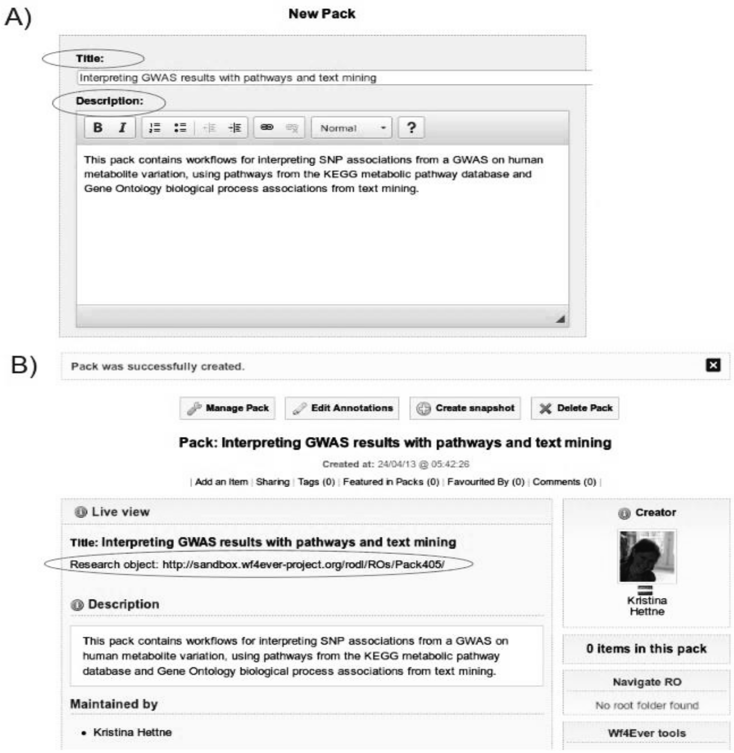


Figure 1 An overview of the Minim model. An overview of the four components: a constraint, a model, a requirement, and a rule.

descriptions about the aggregated resources are present and complete. Additionally, according to explicit requirements defined in a checklist, the tool can also assess the accessibility of those resources aggregated in an RO, in order to increase the trust on the understanding of the RO. The Minim model has four key components, as illustrated by Figure 1: 1) a Constraint, which associates a model (checklist) to use with an RO, for a specific assessment purpose, e.g. reviewing an RO containing sufficient information before being shared; 2) a Model, which enumerates of the set of requirements to be considered, which may be declared at levels of MUST, SHOULD or MAY be satisfied for the model as a whole; 3) a Requirement, which is the key part for expressing the concrete quality requirements to an RO, for example, the presence of certain information about an experiment, or liveness (accessibility) of a data server; 4) a Rule, which can be a SoftwareRequirementRule, to specify the software to be present in the operating environment, a ContentMatchRequirementRule, to specify the presence of certain pattern in the assessed data, or a DataRequirementRule, for specifying data resource to be aggregated in an RO.



Title:
Interpreting GWAS results with pathways and text mining

Description:
This pack contains workflows for interpreting SNP associations from a GWAS on human metabolite variation, using pathways from the KEGG metabolic pathway database and Gene Ontology biological process associations from text mining.

Pack was successfully created.

Pack: Interpreting GWAS results with pathways and text mining

Created at: 24/04/13 @ 05:42:26

Title: Interpreting GWAS results with pathways and text mining

Research object: http://sandbox.wf4ever-project.org/rodl/ROs/Pack405/

Description

This pack contains workflows for interpreting SNP associations from a GWAS on human metabolite variation, using pathways from the KEGG metabolic pathway database and Gene Ontology biological process associations from text mining.

Maintained by

- Kristina Hettne

Kristina Hettne

No root folder found

Figure 2 - Screenshots from myExperiment illustrating the process of creating a Research Object placeholder. Before pressing the “create” button the user can enter a title and description (A), while pressing the “create” button will result in a placeholder Research Object with an identifier (B).

RO digital library

While myExperiment acts mainly as front-end to users, the RO Digital Library [34] acts as a back-end, with two complementary storage components: a digital repository to keep the content, as a triple store to manage the meta-data content. The ROs in the repository can be accessed via a Restful API [45] or via a public SPARQL endpoint [46]. All the ROs created in the myExperiment. org are also submitted to the RO Digital Library.

Workflow-centric RO creation process

Below we describe the steps that we conducted when creating the RO for our case study in an “RO-enabled” version of myExperiment [33]. The populated

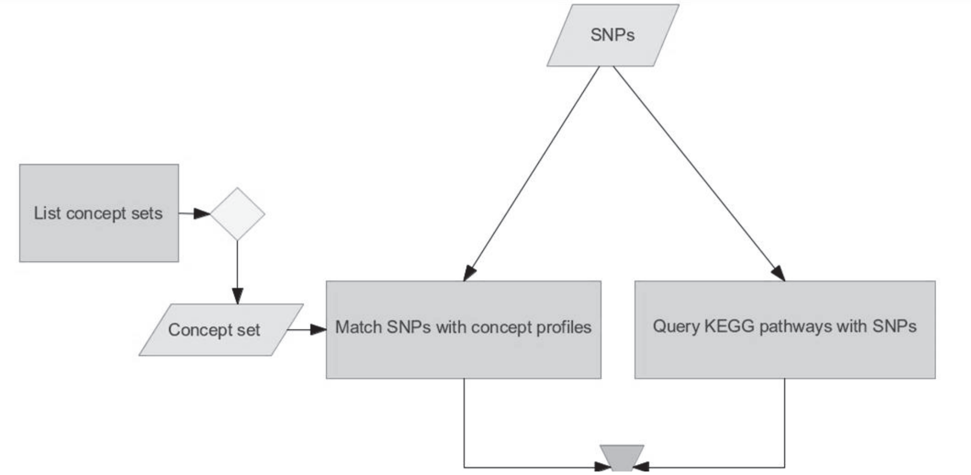

SNPs
List concept sets
Concept set
Match SNPs with concept profiles
Query KEGG pathways with SNPs

Figure 3 - Workflow sketch. A workflow sketch showing that our experiment follows two paths to interpret genome wide association study results: matching with concept profiles and matching with KEGG pathways.

RO is intended to contain all the information required to re- run the experiment, or understand the results presented, or both.

Creating an RO

The action of creating an RO consists of generating the container for the items that will be aggregated, and getting a resolvable identifier for it. In myExperiment the action of creating an RO is similar to creating a pack. We filled in a title and description of the RO at the point of creation and got a confirmation that the RO had been created and had been assigned a resolvable identifier in the RO Digital Library (Figure 2).

Adding the experiment sketch

Using a popular office presentation tool, we made an experiment sketch and saved it as a PNG image. We then uploaded the image to the pack, selecting the type “Sketch”. As a result, the image gets stored in the Digital Library and aggregated in the RO. In addition, an annotation was added to the RO to specify that the image is of type “Sketch”. A miniature version of the sketch is shown within the myExperiment pack (Figure 3).

Adding the hypothesis

To specify the hypothesis, we created a text file that describes the hypothesis, and then upload it to the pack as type “Hypothesis”. The file gets stored in the Digital Library and aggregated in the RO, this time annotated to be of type “Hypothesis”.

Adding workflows

We saved the workflow definitions to files and uploaded them to the pack as type “Workflow”. MyExperiment then automatically performed a workflow-to-RDF transformation in order to extract the workflow structure according to the RO model, which includes user descriptions and metadata created within the Taverna workbench. The descriptions and the extracted structure gets stored in the RO Digital Library and associated with the workflow files as annotations.

Adding the workflow input file

The data values were stored in files that were then uploaded into the pack as “Example inputs”. Such files gets stored in the RO Digital Library and aggregated in the RO, and as “Example inputs”.

Adding the workflow provenance

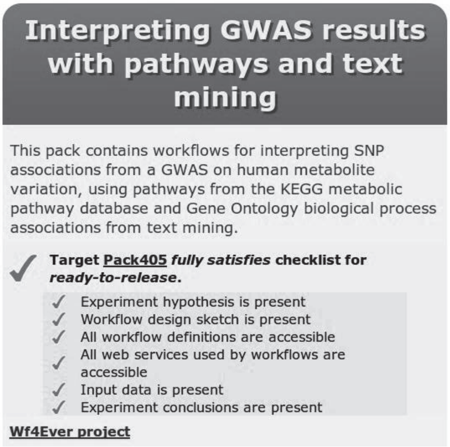
Using the Taverna-Prov [47] extension to Taverna, we exported the workflow run provenance to a file that we uploaded to the pack as type “Workflow run”. Similar to other resources, the provenance file gets stored in the digital library with the type “Workflow run”, however as the file is in the form of RDF according to the wfprov [42] and W3C PROV-O [22] ontologies; it is also integrated into the RDF store of the digital library and available for later querying.

Adding the results

We made a compilation of the different workflow outputs to a result file in table format, uploaded to the pack as type “Results”. The file gets stored in the digital library and aggregated in the RO, annotated to be of the type “Results”.

Adding the conclusions

To specify the hypothesis, we created a text file that describes the hypothesis, and then uploaded it to the pack as type “Hypothesis”. The file gets stored in the digital library and aggregated in the RO, annotated to be of type “Conclusions”.



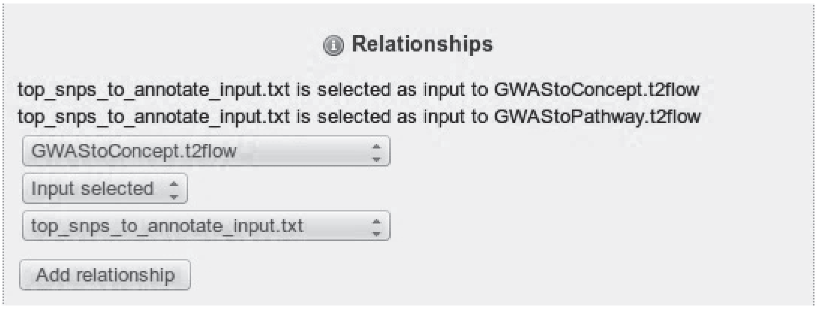
Interpreting GWAS results with pathways and text mining

This pack contains workflows for interpreting SNP associations from a GWAS on human metabolite variation, using pathways from the KEGG metabolic pathway database and Gene Ontology biological process associations from text mining.

✓ Target Pack405 *fully satisfies* checklist for *ready-to-release*.

✓ Experiment hypothesis is present
✓ Workflow design sketch is present
✓ All workflow definitions are accessible
✓ All web services used by workflows are accessible
✓ Input data is present
✓ Experiment conclusions are present

Figure 4 - Screenshot of the results from the second check with the checklist evaluation service. The results from checklist evaluation service show that the Research Object satisfies the defined checklist for a Research Object.Intermediate step: checklist evaluation



Relationships

top_snps_to_annotate_input.txt is selected as input to GWAStoConcept.t2flow
top_snps_to_annotate_input.txt is selected as input to GWAStoPathway.t2flow

GWAStoConcept.t2flow
Input selected
top_snps_to_annotate_input.txt

Figure 5 - Screenshot of the relationships in the RO in myExperiment. The relationships between example inputs and workflows in the Research Object have been defined in myExperiment.At this point we checked how far we were from satisfying the Minim model, and were informed by the tool that the RO now fully satisfies the checklist (Figure 4).

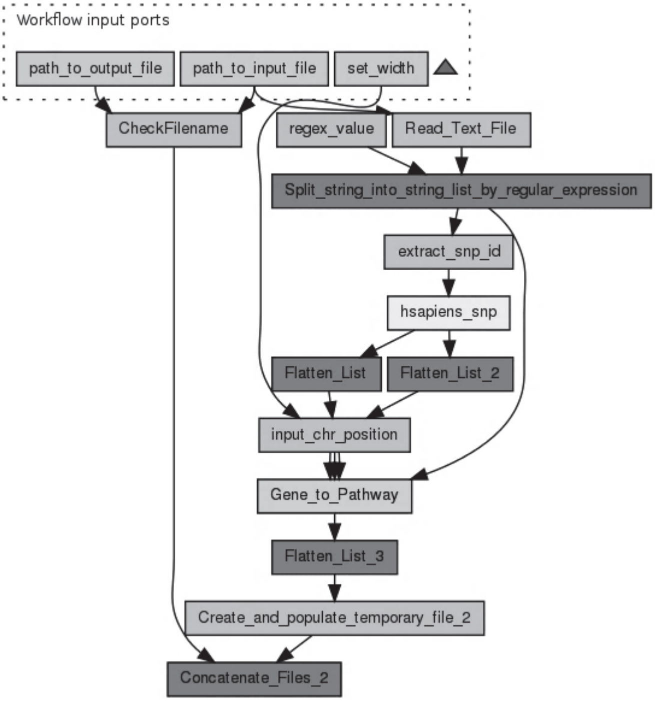


Figure 6 - Taverna workflow diagram for the KEGG workflow. Blue boxes are workflow inputs, brown boxes are scripts, grey boxes are constant values, green boxes are Web services, purple boxes are Taverna internal services, and pink boxes are nested workflows.

Annotating and linking the resources

We linked the example input file to the workflows that used the file by the property “Input_selected” (Figure 5). In this particular case, both workflows have the same inputs but they need to be configured in different ways. This is described in the workflow description field in Taverna.

Results

The RO for our experiment is the container for the items that we wished to aggregate. In terms of RDF, we first instantiated an ro:ResearchObject in an RO-enabled version of myExperiment [33]. We thereby obtained a unique and resolvable Uniform Resource Identifier (URI) from the RO store that underlies this version of myExperiment. In our experimental setup this was

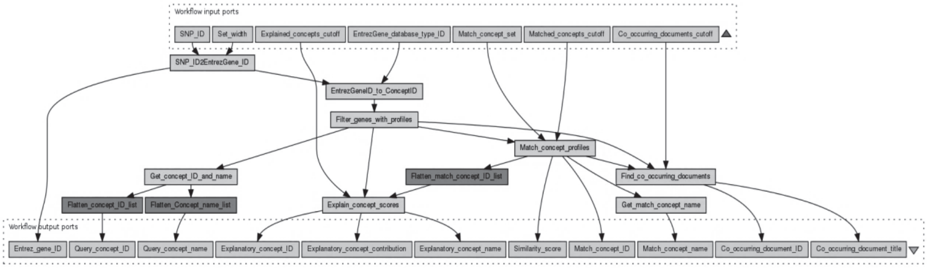
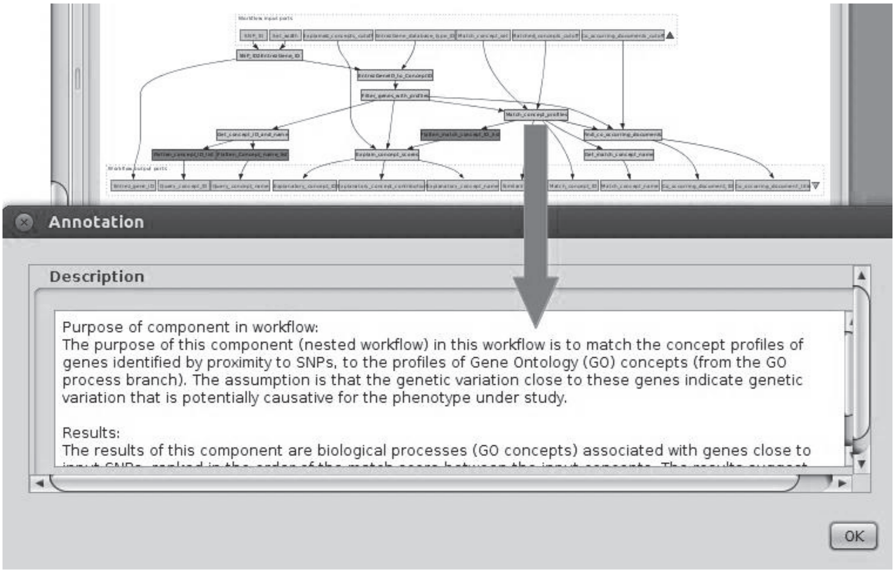


Figure 7 - Taverna workflow diagram for the concept profile mining workflow. Blue boxes are workflow inputs, purple boxes are Taverna internal services, and pink boxes are nested workflows.



Purpose of component in workflow:
The purpose of this component (nested workflow) in this workflow is to match the concept profiles of genes identified by proximity to SNPs, to the profiles of Gene Ontology (GO) concepts (from the GO process branch). The assumption is that the genetic variation close to these genes indicate genetic variation that is potentially causative for the phenotype under study.

Results:
The results of this component are biological processes (GO concepts) associated with genes close to

Figure 8 Taverna workflow annotation example. An example of an annotation of the purpose of a nested workflow in Taverna.

http:// sandbox.wf4ever-project.org/rodl/ROs/Pack405/. It is accessible from myExperiment [48]. Each of the subsequent items in the RO was aggregated as an ro:Resource, indicating that the item is considered a constituent member of the RO from the point of view of the scientist (the creator of the RO).

Aggregated resources

We aggregated the following items: 1) the hypothesis (roterms:Hypothesis): we hypothesized that SNPs can be functionally annotated using metabolic pathway information complemented by text mining, and that this will lead to formulating new hypotheses regarding the role of genomic variation in biological processes; 2) the sketch (roterms:Sketch) shows that our experiment follows two paths to interpret SNP data: matching with concept profiles and matching with Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Figure 3); 3) the workflows (wfdesc:Workflow): Figure 6 shows the workflow diagram for the KEGG workflow and Figure 7 shows the workflow diagram for the concept profile matching workflow. In Taverna, we aimed to provide sufficient annotation of the inputs, outputs and the functions of each part of the workflow to ensure a clear interpretation and to ensure that scientists know how to replay the workflows using the same input data, or re-run them with their own data. We provided textual descriptions in Taverna of each step of the workflow, in particular to indicate their purpose within the workflow (Figure 8);

4) the input data (roterms:exampleValue) that we aggregated in our RO was a list of example SNPs derived from the chosen GWAS [28]; 5) the workflow run provenance (roterms:WorkflowRunBundle): a ZIP archive that contains the intermediate values of the workflow run, together with its provenance trace expressed using wfprov:WorkflowRun and subsequent terms from the wfprov ontology. We thus stored process information from the input of the workflow execution to its output results, including the information for each constituent process run in the workflow run, modelled as wfprov:ProcessRun. The run data is: 3 zip files containing 2090 intermediate values as separate files totalling 9.7 MiB, in addition to 5 MiB of provenance traces; 6) the results (roterms:Result) were compiled from the different workflow outputs to one results file (see result document in the RO [49] Additional file 1). For 15 SNPs it lists the associated gene name, the biological annotation from the GWAS publication, the associated KEGG pathway, and the most strongly associated biological process according to concept profile matching. Our workflows were able to compute a biological

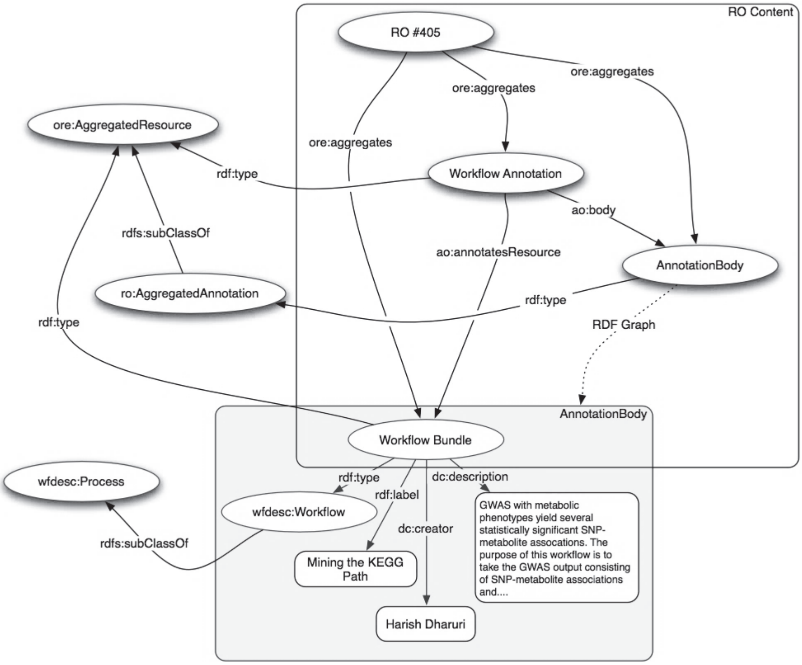


Figure 9 Simplified diagram showing part of the Research Object for our experiment. The Research Object contains the items that were aggregated by the "Research Object-enabled" version of myExperiment. Shown is the part of the RDF graph that aggregates and annotates the KEGG pathway mining workflow.

annotation from KEGG for 10 out of 15 SNPs and 15 from mining PubMed. All KEGG annotations and most text mining annotations corresponded to the annotations by Illig et al [28]. An important result of the text mining workflow was the SNP-annotation "rs7156144stimulation of tumor necrosis factor production", which represents a hypothetical relation that to our knowledge was not reported before; 7) the conclusions (roterms: Conclusion): we concluded that our KEGG and text mining workflows were successful in retrieving biological annotations for significant SNPs from a GWAS experiment, and predicting novel annotations.

As an example of our instantiated RO, Figure 9 provides a simplified view of the RDF graph that aggregates and annotates the KEGG mining workflow. It shows the result of uploading our Taverna workflow to myExperiment, as it initiated an automatic transformation from a Taverna 2 t2flow file to a Taverna 3 workflow bundle, while extracting the workflow structure and user

Table 1 RO items checklist. RO items for a workflow-based experiment annotated with the appropriate term from the Minim vocabulary.

Research Object item	Requirement	RO ontology term
Hypothesis or Research question	Should	roterms:Hypothesis/ roterms:ResearchQuestion
Workflow sketch	Should	roterms:Sketch
One or more workflows	Must	wfdesc:Workflow
Web services of the workflow	Must	wfdesc:Process
Example input data	Must	roterms:exampleValue
Provenance of workflow runs	Must	wfprov:WorkflowRun
Example results	Must	roterms:Result
Conclusions	Must	roterms:Conclusion

descriptions in terms of the wfdesc model [41]. The resulting RDF document was aggregated in the RO and used as the annotation body of a ao:Annotation on the workflow, thus creating a link between the aggregated workflow file and its description in RDF. The Annotation Ontology uses named graphs for semantic annotation bodies. In the downloadable ZIP archive of an RO each named graph is available as a separate RDF document, which can be useful in current RDF triple stores that do not yet fully support named graphs. The other workflows were aggregated and annotated in the same way. The RO model further uses common Dublin Core vocabulary terms [50] for basic metadata such as creator, title, and description.

In some cases we manually inserted specified relations between the RO resources via the myExperiment user interface. An example is the link between input data and the appropriate workflow for cases when an RO has multiple workflows and multiple example inputs. In our case, both workflows have the same inputs, but they need to be configured in different ways. This was described in the workflow description field in Taverna which becomes available from an annotation body in the workflow upload process.

Checking for completeness of an RO: application of the Minim model

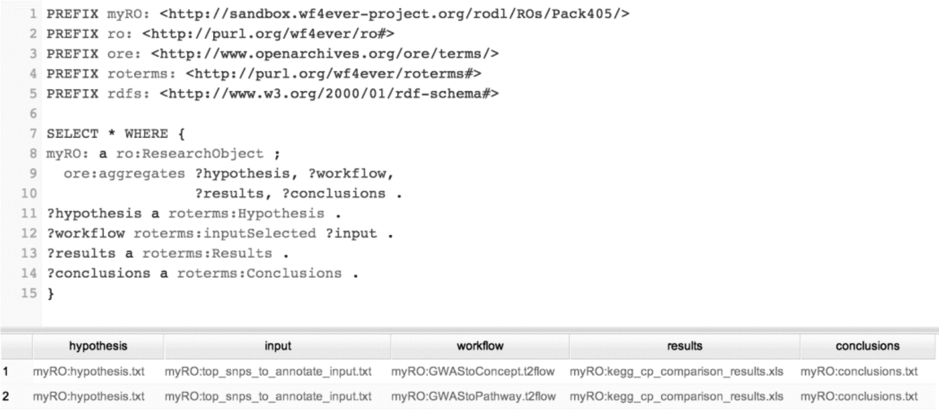
We also applied Semantic Web technology for checking the completeness of our RO. We implemented a checklist for the items that we consider essential or desirable for understanding a workflow-based experiment by annotating the corresponding parts of the RO model with the appropriate term from the Minim vocabulary (Table 1).

Thus, some parts were annotated as “MUST have” with the property minim:hasMustRequirement (e.g. at least one workflow definition), and others as “SHOULD have” with the property minim:hasShouldRequirement (e.g. the overall sketch of the experiment). The complete checklist document can be found online in RDF format [51] and in a format based on the spreadsheet description of the workflow [52]. We subsequently used a checklist service that evaluates if an RO is complete by executing SPARQL queries on the Minim mappings. The overall result is a summary of the requirement levels associated with the individual items; e.g. a missing MUST requirement is a more serious omission than a missing SHOULD (or COULD) requirement. We justified the less strict requirements for some items to accommodate cases when an RO is used to publish a method as such. We found that treating the requirement levels as mutually exclusive (hence not sub properties) simplifies the implementation of checklist evaluation, and in particular the generation of results when a checklist item is not satisfied.

Discussion

In this paper we explored the application of the Semantic Web encoded RO model to provide a container data model for preserving sufficient information for researchers to understand a computational experiment. We found that the model indeed allowed us to aggregate the necessary material together with sufficient annotation (both for machines and humans). Moreover, mapping of selected RO model artefacts to the Minim vocabulary allowed us to check if the RO was complete according to our own predefined criteria. The checklist service can be configured to accommodate different criteria. Research groups may have different views on what is essential, but also libraries or publishers may define their own standards, enabling partial automation of the process of checking a submission against specific instructions to authors. Furthermore, the service can be run routinely to check for workflow decay, in particular decay related to references that go missing.

In using the RO model, we sought to meet requirements for sharing, reuse and repurposing, as well as interoperability and reproducibility. This fits with current trends to enhance reproducibility and transparency of science (e.g. see [53-55]). Reproducibility in computational science has been defined as a spectrum [55], where a computational experiment that is described only by a publication is not seen as reproducible, while adding code, data, and finally the linked data and execution data will move the experiment towards full replication. Adhering to this definition, our RO-enabled computational



```
PREFIX myRO: <http://sandbox.wf4ever-project.org/rodl/ROs/Pack405/>
PREFIX ro: <http://purl.org/wf4ever/ro#>
PREFIX ore: <http://www.openarchives.org/ore/terms/>
PREFIX roterms: <http://purl.org/wf4ever/roterms#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT * WHERE {
myRO: a ro:ResearchObject ;
  ore:aggregates ?hypothesis, ?workflow,
                 ?results, ?conclusions .
?hypothesis a roterms:Hypothesis .
?workflow roterms:inputSelected ?input .
?results a roterms:Results .
?conclusions a roterms:Conclusions .
}
```

	hypothesis	input	workflow	results	conclusions
1	myRO:hypothesis.txt	myRO:top_snps_to_annotate_input.txt	myRO:GWAStoConcept.t2flow	myRO:kegg_cp_comparison_results.xls	myRO:conclusions.txt
2	myRO:hypothesis.txt	myRO:top_snps_to_annotate_input.txt	myRO:GWAStoPathway.t2flow	myRO:kegg_cp_comparison_results.xls	myRO:conclusions.txt

Figure 10 Screenshot showing a SPARQL query and its results. Query to obtain a reference to the data that was used as input to our workflows and the conclusions that we drew from evaluating the workflow results.

experiment comes close to fulfilling the ultimate golden standard of full replication, but falls short because it has not been analyzed using independently collected data. The benefit offered by the RO in terms of reproducibility is that it provides a context (RO) within which an evaluation of reproducibility can be performed. It does this by providing an enumerated and closed set of resources that are part of the experiment concerned, and by providing descriptive metadata (annotations) that may be specific to that context. This is not necessarily the complete solution to reproducible research, but at least an incremental step in that direction. We have used RDF as the underlying data model for exchanging ROs. One of advantages is the ability to query the data, which becomes clear when we want to answer questions about the experiment, such as: 1) which conclusions were drawn from a given workflow? 2) Which workflow (run) supports a particular conclusion and which datasets did it use as inputs?; 3) Which different workflows used the same dataset X as input?; 4) Who can be credited for creating workflows that use GWAS data? The answers for the first two questions can readily be found using a simple SPARQL [56] query. Figure 10 shows the SPARQL query and the results as returned by the SPARQL endpoint of the RO Digital Library. Note that in our case we got two result rows, one for each of the workflows that were used to confirm the hypothesis. We emphasize that queries could also be constructed to answer more elaborate questions such as question 3 and 4. Without adding any complexity to the query or the infrastructure, it is possible to query over the entire repository of research objects. This effectively integrates all meta-data of any workflow-based experiment that was uploaded to the RO Digital Library via myExperiment. When more ROs have become available that use the same annotations as described in this paper, then we can start sharing queries that can act as templates. We did not explore further formalization in terms of rejecting or accepting hypotheses, since formulating such a hypothesis model properly would be very domain specific, such as current efforts in neuromedicine [57]. However, the RO model does not exclude the possibility to do so.

Applying the RO model in genomic working environments

An important criterion for our evaluation of the RO model and tools is that it should support researchers in preparing their digital methods and results for publication. We have shown that the RO model can be applied in an existing framework for sharing computational workflows (myExperiment). We used Taverna to create our workflows, and the wf4ever toolkit [58], including dLibra [59] that was extended with a triple store as a back end to store the ROs. The RO features of the test version of myExperiment that we used are currently under development for migration to the production version of myExperiment [60]. Creating an RO in the test version of myExperiment is not any different to a user than the action of creating a pack, completely hiding the creation of RDF objects under the hood. The difference lies in the support of the RO model, which allows the user to add data associated with a computational experiment in a structured way (a sketch representing the experimental setup, the hypothesis document, result files, etc.), and metadata in the form of annotations. Every piece of data in an RO can be annotated, either in a structured or machine-generated way like the automatic annotation of a wfdesc description of a workflow as provided by the workflow-to-RDF transformation service, or manually by the user at the time of resource upload, such as the annotation of an experiment overview as “Sketch”. Since RO descriptions are currently not a pre-requisite to publishing workflow results in journal, we hope that this support and streamlining of the annotation process will act as an incentive for scientists to start using the RO technology.

The representation of an RO in myExperiment as presented in this paper should be seen as a proof-of-concept. Crucial elements of a computational experiment are handled, but there is room for improvement. For example, the hypothesis and conclusions are at the moment only shown as downloadable text files and the content and provenance of a workflow run is not shown to the user. We found that more tooling is needed to make practical use of the provenance trace. It is detailed and focus is on data lineage, rather than the biological meaning of the recorded steps. Nevertheless, we regard this raw workflow data as highly valuable as the true record of what exactly was executed. It allows introspection of the data lineage, such as which service was invoked with exactly which data. By providing this proof-of-concept and the RO model as a reference model, we hope to stimulate developers of other genomic working environments such as Galaxy [6] and Genome Space [61] to start implementing the RO model as well, thus enabling scientists to share their investigation results as a complete knowledge package. Similarly, workflow systems use different workflow languages [62,63], and by presenting the workflow-to-RDF transformation service that handles the t2flow serialization format to transform a workflow to an RO, we hope to encourage systems that use other workflow languages to develop similar services to transform their workflows to ROs. This would allow for a higher-level understanding of workflow-based experiments regardless of the type of workflow system used.

It should be noted that although our ROs fully capture the individual data items of individual steps within workflow runs, this approach is not applicable to all scientific workflows. In fact, we have since further developed the provenance support for Taverna so that larger pieces of data are only recorded as URI references and not bundled within the ZIP file. The Taverna workflow system already supports working with such references; however many bioinformatics Web services still only support working directly with values. When dealing with references, the workflow run data only capture the URI and its metadata, and full access to the run data therefore would also depend on the continued availability (or mirroring) of those referenced resources, and their consistency would therefore later need to be verified against metadata such as byte size and Secure Hash Algorithm checksums.

Generalization to other domains

We acknowledge that apart from enabling the structured aggregation and annotation of digital ROs technically, scientists appreciate guidelines and Best Practices for producing high quality ROs. In fact, the minimal requirements for a complete RO that we implemented via the Minim model, were inspired by the 10 Best Practices that we defined for creating workflows [39]. An RO may be evaluated using different checklists for different purposes. A checklist description is published as linked data, and may be included in the RO, though we anticipate more common use will be for it to be published separately in a community web site. In our work to date, we have used checklist definitions published via Github (e.g. [64]), and are looking to create a collection of example checklist definitions to seed creation of checklists for different domains or purposes [65]. We envision that instructions to authors of ROs may differ between research communities, and publishers who wish to adopt RO technology for digital submissions may develop their own 'Instructions to Authors' for ROs. This could be implemented by different mappings of the Minim model.

Related work

The RO model was implemented as a Semantic Web model to provide a general, domain-agnostic reference that can be extended by domain specific ontologies. For instance, while the RO model offers terms pertaining to experimental science such as "hypothesis" and "conclusion", extensions to existing models that also cover this area and are already in use in the life science domain could be considered. It is beyond the scope of this article to exhaustively review related ontologies and associated tools, but we wish to mention six that in our view are prime candidates to augment the RO family of ontologies and tools. The first is the Ontology for Biomedical Investigations (OBI) that aims to represent all phases of experimental processes, such as study designs, protocols, instrumentation, biological material, collected data and analyses performed on that data [66]. OBI is used for the ontological representation of the results of the Investigation-Study-Assay (ISA) metadata tools [67] that is the next on our list of candidates. ISA, developed by the ISA commons community, facilitates curation, management, and reuse of omics datasets in a variety of life science domains [68]. It puts spreadsheets at the heart of its tooling, making it highly popular for study capture in the omics domain [69]. The third candidate is the ontology for scientific experiments EXPO [70]. EXPO is defined by OWL-DL axioms and is grounded in upper ontologies. Its coverage of experiment terms is good, but we are unsure about its uptake by the community. Perhaps unfortunate for a number of good ontologies, we consider this an important criterion for interoperability. Four and five on our list relate to the annotation of Web Services (or bioinformatics operations in general): the EMBRACE Data and Methods (EDAM) ontology encompasses over 2200 terms for annotating tool or workflow functions, types of data and identifiers, application domains and data formats [71]. It is developed and maintained by the European Bioinformatics Institute and has been adopted for annotation of for instance the European Molecular Biology Open Software Suite. The myGrid-BioMoby ontology served as a starting point for the development of EDAM. This will facilitate the adoption of EDAM by for instance BioCatalogue,org and service-oriented tools such as Taverna, which would further broaden its user base and thereby its use for inter- operability. The Semantic Automated Discovery and Integration (SADI) framework [72] takes semantic annotation of Web Services one step further. A SADI Web Service describes itself in terms of OWL classes, and produces and consumes instances of OWL classes. This enables instant annotation in a machine readable format when a workflow is built from SADI services. In addition, via a SADI registry suggestions can be made about which services to connect to which. SADI has clear advantages as an annotation framework. However, not all bioinformatics services are available as SADI services, while the conversion is not trivial without training in Semantic Web modelling. Therefore, SADI and RO frameworks could be strongly complementary for workflows that use a heterogeneous mix of service types. This would be further facilitated when both are linked to common ontologies such as EDAM. Finally, we highlight the recent development of models for microattribution and nanopublication that aim to provide a means of getting credit for individual assertions and making these available in a machine readable format [73-75]. Taking nanopublications as an example, we could “nanopublish” specific results from our experiment, such as the text mining-based association that we found between the SNP “rs7156144” and the biological process “stimulation of tumor necrosis factor production”. In addition to an assertion, a nanopublication consists of provenance meta-data (to ensure trust in the assertion) and publication information (providing attribution to authors and curators). Nanopublication and RO complement each other in two ways. On the one hand, nanopublications can be used to publish and expose valuable results from workflows and included in the RO aggregate. On the other hand, an RO could be referenced as part of the provenance of a nanopublication, serving as a record of the method that led to assertion of the nanopublication. Similar to the nanopublication and microattribution models, the Biotea and Elsevier Smart Content Initiative data models also aim to model scientific results, but are focused on encapsulating a collection of information that are related to the results reported in publications [76,77]. The relationship between an RO and these datasets is not much different from an RO with a nanopublication statement. An RO can be referenced by, e.g. the Biotea dataset, by its URI, which can provide detailed experimental information or provenance information about the results described by the Biotea dataset. In the meanwhile, an RO can also reference a Biotea dataset or an Elsevier linked dataset.

Summarizing, the RO model provides a general framework with terms for aggregating and annotating the components of digital research experiments, by which it can complement related frameworks that are already used in the life science domain such as EXPO, OBI, ISA, EDAM, SADI and nanopublication. We observe that models are partly complementary and partly overlapping in scope. Therefore, we stimulate collaboration towards the development of complementarity frameworks. For instance, we initiated an investigation of the combination of ISA, RO, and Nanopublication as a basis for general guidelines for publishing digital research artefacts (Manuscript in preparation).

Uptake by the research community

Beyond the RO presented in this paper, the RO model has been used to generate ROs within the domains of musicology [78] and astronomy using AstroTaverna [79]. In addition, we recently explored how an RO could be referenced as part of the provenance of nanopublications of genes that are differentially expressed in Huntington's Disease (HD) with certain genomic regions [80,81]. The results from the in silico analysis of the differentially expressed genes were obtained from a Taverna data integration workflow and the RO itself was stored in the Digital Library. Using the PROV-O ontology, the nanopublication provenance was modelled to link to the workflow description in the RO. Since the RO was mostly automatically generated by the procedure described in this paper, the nanopublication refers to detailed provenance information without requiring additional modelling effort. To encourage further uptake by the research community we have developed the Web resource ResearchObject.org [82]. ResearchObject.org lists example ROs [83], presents the ongoing activities of the open RO community, and gathers knowledge about related developments and adoptions.

Conclusions

Applying the workflow-centric RO model and associated models such as Minim provides a digital method to increase the understanding of bioinformatics experiments. Crucial meta-data related to the experiment is preserved in a Digital Library by structured aggregation and anno- tation of hypothesis, input data, workflows, workflow runs, results, and conclusions. The Semantic Web representation provides a reference model for life scientists who perform computational analyses and for systems that support this, and can complement related annotation frameworks that are already in use in the life science domain.

References

1. Chen H, Yu T, Chen JY: **Semantic Web meets Integrative Biology: a survey**. *Br Bioinform* 2012, **14**:109–125.
2. Sneddon TP, Li P, Edmunds SC: **GigaDB: announcing the GigaScience database**. *Gigascience* 2012, **1**:11.
3. Ghosh S, Matsuoka Y, Asai Y, Hsin K-Y, Kitano H: **Software for systems biology: from tools to integrated platforms**. *Nat Rev Genet* 2011, **12**:821–832.
4. Beaulah SA, Correll MA, Munro REJ, Sheldon JG: **Addressing informatics challenges in Translational Research with workflow technology**. *Drug Discov Today* 2008, **13**:771–777.
5. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P: **The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud**. *Nucleic Acids Res* 2013, **41**(Web Server issue):W557–W561.
6. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11**:R86.
7. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P: **myExperiment: a repository and social network for the sharing of bioinformatics workflows**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W677–W682.
8. Mates P, Santos E, Freire J, Silva CT: **CrowdLabs: Social Analysis and Visualization for the Sciences**. *Sci Stat Database Manag Vol 6809* 2011:555–564.
9. Zhao J, Gomez-Perez JM, Belhajjame K, Klyne G, Garcia-Cuesta E, Garrido A, Hettne K, Roos M, De Roure D, Goble C: **Why workflows break - Understanding and combating decay in Taverna workflows**. *2012 IEEE 8th Int Conf E-Science* 2012:1–9.
10. Rebholz-Schuhmann D, Grabmuller C, Kavaliauskas S, Croset S, Woollard P, Backofen R, Filsell W, Clark D: **A case study: semantic integration of gene-disease associations for type 2 diabetes mellitus from literature and biomedical data resources**. *Drug Discov Today* 2013, **7**:882–889.

11. Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C, Mons B: **Open PHACTS: semantic interoperability for drug discovery**. *Drug Discov Today* 2012, **17**:1188–1198.
12. **Wf4Ever Research Object model**. http://wf4ever.github.io/ro .
13. Belhajjame K, Corcho O, Garijo D, Zhao J, Missier P, Newman DR, Palma R, Bechhofer S, Garcia Cuesta E, Gomez-Perez JM, Klyne G, Page K, Roos M, Enrique Ruiz J, Soiland-Reyes S, Verdes-Montenegro L, De Roure D, Goble C: **Workflow-Centric Research Objects: A First Class Citizen in the Scholarly Discourse**. *Proc 2nd Work Semant Publ Vol 903* 2012.
14. Bechhofer S, De Roure D, Gamble M, Goble CA, Buchan I: **Research objects: Towards exchange and reuse of digital knowledge**. *Futur Web Collab Sci* 2010.
15. Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, Bhagat J, Couch P, Cruickshank D, Delderfield M, Dunlop I, Gamble M, Michaelides D, Owen S, Newman D, Sufi S, Goble C: **Why linked data is not enough for scientists**. *Futur Gener Comput Syst* 2013, **29**:599–611.
16. De Roure D, Missier P, Manuel J, Hettne K, Klyne G, Goble C: **Towards the Preservation of Scientific Workflows**. .
17. Roos M, Marshall MS, Gibson AP, Schuemie M, Meij E, Katrenko S, van Hage WR, Krommydas K, Adriaans PW: **Structuring and extracting knowledge for the support of hypothesis generation in molecular biology**. *BMC Bioinformatics* 2009, **10 Suppl 1**(Suppl 10):S9.
18. Livingston KM, Bada M, Hunter LE, Verspoor K: **Representing annotation compositionality and provenance for the Semantic Web**. *J Biomed Semant* 2013, **4**:38.
19. Ciccarese P, Soiland-Reyes S, Belhajjame K, Gray AJ, Goble C, Clark T: **PAV ontology: provenance, authoring and versioning**. *J Biomed Semant* 2013, **4**:37.
20. **Object Exchange and Reuse (ORE) model**. http://www.openarchives.org/ ore/1.0/primer.html.
21. Ciccarese P, Ocana M, Garcia Castro LJ, Das S, Clark T: **An open annotation ontology for science on web 3.0**. *J Biomed Semant* 2011, **2**(Suppl 2):S4.
22. Missier P, Belhajjame K, Cheney J: **The W3C PROV family of specifications for modelling provenance metadata**. *Proc 16th Int Conf Extending Database Technol - EDBT '13* 2013:773.

23. Zhao J, Klyne G, Gamble M, Goble CA: **A Checklist-Based Approach for Quality Assessment of Scientific Information**. *Proc Third Linked Sci Work co-located Int Semant Web Conf* 2013.
24. **Minim checklist service**. https://github.com/wf4ever/ro-manager/blob/master/Minim/Minim-description.md.
25. Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T, Brazma A, Brinkman RR, Michael Clark A, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Le Novere N: **Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project**. *Nat Biotechnol* 2008, **26**:889–896.
26. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Bostrom KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N: **Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes**. *Nat Genet* 2008, **40**:638–645.
27. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges**. *Nat Rev Genet* 2008, **9**:356–369.
28. Illig T, Gieger C, Zhai G, Romisch-Margl W, Wang-Sattler R, Prehn C, Altmaier E, Kastenmuller G, Kato BS, Mewes HW: **A genome-wide perspective of genetic variation in human metabolism**. *Nat Genet* 2010, **42**:137–141.
29. Gieger C, Geistlinger L, Altmaier E, De Angelis MH, Kronenberg F, Meitinger T, Mewes HW, Wichmann HE, Weinberger KM, Adamski J, Illig T, Suhre K: **Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum**. *PLoS Genet* 2008, **4**.
30. Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wagele B, Altmaier E, Deloukas P, Erdmann J, Grundberg E: **Human metabolic individuality in biomedical and pharmaceutical research**. *Nature* 2011, **477**:54–60.
31. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LCJ, Jenster G, Kors JA: **Anni 2.0: a multipurpose text-mining tool for the life sciences**. *Genome Biol* 2008, **9**:R96.

32. Hettne KM, Boorsma A, van Dartel DA, Goeman JJ, de Jong E, Piersma AH, Stierum RH, Kleinjans JC, Kors JA: **Next-generation text-mining mediated generation of chemical response-specific gene sets for interpretation of gene expression data**. *BMC Med Genomics* 2013, **6**:2.
33. **myExperiment alpha**. http://alpha.myexperiment.org .
34. Palma R, Corcho O, Hotubowicz P, Perez S, Page K, Mazurek C: **Digital libraries for the preservation of research methods and associated artifacts**. *Proc 1st Int Work Digit Preserv Res Methods Artefacts - DPRMA '13* 2013:8–15.
35. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic Acids Res* 2012, **40**(Database issue):D109–14.
36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**:25–29.
37. **KEGG REST services**. http://www.kegg.jp/kegg/rest/keggapi.html.
38. **Concept Profile Mining Web services**. https://www.biocatalogue.org/services/3559 .
39. Hettne KM, Wolstencroft K, Belhajjame K, Goble CA, Mina E, Dharuri H, De Roure D, Verdes-Montenegro L, Garrido J, Roos M: **Best Practices for Workflow Design: How to Prevent Workflow Decay**. *Proc 5th Int Work Semant Web Appl Tools Life Sci Paris, Fr Novemb 28-30, 2012, Vol 952* 2012.
40. Sanderson R, Ciccarese P, de Sompel H: **Designing the W3C open annotation data model**. *Proc 5th Annu ACM Web Sci Conf - WebSci '13* 2013:366–375.
41. **wfdesc vocabulary**. https://github.com/wf4ever/ro/blob/master/wfdesc.owl.
42. **wfprov ontology**. http://purl.org/wf4ever/wfprov#.
43. **RO terms vocabulary**. http://purl.org/wf4ever/roterms.
44. **Minim checklist ontology**. http://purl.org/minim/ .
45. **Research Object Digital Library Restful API**. http://www.wf4ever-project. org/wiki/display/docs/RO+API+6.
46. **Research Object Digital Library SPARQL endpoint**. http://sandbox.wf4ever- project.org/portal/sparql?1.

47. Alper P, Belhajjame K, Goble CA, Karagoz P: **Enhancing and abstracting scientific workflow provenance for data publishing**. *Proc Jt EDBT/ICDT 2013 Work - EDBT '13* 2013:313.
48. **Research Object in myExperiment**. http://www.myexperiment.org/packs/428.
49. **Research Object results**. http://alpha.myexperiment.org/packs/405/ resources/kegg_cp_comparison_results.xls .
50. **DCMI Usage Board (2012): DCMI Metadata Terms**. http://dublincore.org/ documents/2012/06/14/dcmi-terms/ .
51. **RO checklist document in RDF**. https://github.com/wf4ever/ro-catalogue/ blob/master/minim/minim-workflow-demo.rdf.
52. **Spreadsheet-based RO checklist document**. https://github.com/wf4ever/ ro-catalogue/blob/master/minim/minim-workflow-demo.pdf .
53. **Enhancing reproducibility**. *Nat Methods* 2013, **10**:367.
54. Ince DC, Hatton L, Graham-Cumming J: **The case for open computer programs**. *Nature* 2012, **482**:485–488.
55. Peng RD: **Reproducible research in computational science**. *Science (80-)* 2011, **334**:1226–1227.
56. **SPARQL Protocol and RDF Query Language**. http://www.w3.org/TR/ sparql11-overview/ .
57. Cheung K-H, Kashyap V, Luciano JS, Chen H, Wang Y, Stephens S, Ciccarese P, Wu E, Wong G, Ocana M, Kinoshita J, Ruttenberg A, Clark T: **The SWAN biomedical discourse ontology**. *J Biomed Inf* 2008, **41**:739–751.
58. Page K, Palma R, Holubowicz P, Klyne G, Soiland-Reyes S, Cruickshank D, Cabero RG, Cuesta EG, De Roure D, Zhao J: **From workflows to Research Objects: an architecture for preserving the semantics of science**. *Proc 2nd Int Work Linked Sci* 2012.
59. **dLibra**. http://dlab.psnc.pl/dlibra/ .
60. **myExperiment release schedule**. http://wiki.myexperiment.org/index.php/ Developer:ReleaseSchedule .
61. **Genome Space**. http://www.genomespace.org/ .
62. Tiwari A, Sekhar AKT: **Workflow based framework for life science informatics**. *Comput Biol Chem* 2007, **31**:305–319.
63. Romano P: **Automation of in-silico data analysis processes through workflow management systems**. *Br Bioinform* 2008, **9**:57–68.

64. **Example Minim checklist definition**. https://github.com/wf4ever/ro-catalogue/blob/master/v0.1/Y2Demo-test/workflow-experiment-checklist.rdf .
65. **Collection of example Minim checklist definitions**. https://github.com/wf4ever/ro-catalogue/tree/master/minim .
66. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, Malone J, Parkinson H, Peters B, Rocca-Serra P, Ruttenberg A, Sansone S-A, Soldatova LN, Stoeckert CJ, Turner JA, Zheng J: **Modeling biomedical experimental processes with OBI**. *J Biomed Semant* 2010, **1**(Suppl 1):S7.
67. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, Field D, Harris S, Hide W, Hofmann O, Neumann S, Sterk P, Tong W, Sansone S-A: **ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level**. *Bioinformatics* 2010, **26**:2354–2356.
68. Sansone S-A, Rocca-Serra P, Brandizi M, Brazma A, Field D, Fostel J, Garrow AG, Gilbert J, Goodsaid F, Hardy N, Jones P, Lister A, Miller M, Morrison N, Rayner T, Sklyar N, Taylor C, Tong W, Warner G, Wiemann S: **The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?"** *OMICS* 2008, **12**:143–149.
69. Maguire E, Gonzalez-Beltran A, Whetzel PL, Sansone S-A, Rocca-Serra P: **OntoMaton: a bioportal powered ontology widget for Google Spreadsheets**. *Bioinformatics* 2013, **29**:525–527.
70. Soldatova LN, King RD: **An ontology of scientific experiments**. *J R Soc Interface* 2006, **3**:795–803.
71. Ison J, Kalas M, Jonassen I, Bolser D, Uludag M, McWilliam H, Malone J, Lopez R, Pettifer S, Rice P: **EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics, and formats**. *Bioinformatics* 2013, **29**:1325–1332.
72. Wilkinson MD, Vandervalk B, McCarthy L: **The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern**. *API Ref Implement J Biomed Semant* 2011, **2**:8.
73. Patrinos GP, Cooper DN, van Mulligen E, Gkantouna V, Tzimas G, Tatum Z, Schultes E, Roos M, Mons B: **Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain**. *Hum Mutat* 2012, **33**:1503–1512.
74. Mons B, Van Haagen H, Chichester C, Hoen 't P-B, Dunnen JT D, Van Ommen G, Mulligen EM V, Singh B, Hooft R, Roos M, Hammond J, Kiesel B, Giardine B, Velterop J, Groth P, Schultes E, Den Dunnen JT: **The value of data**. *Nat Genet* 2011, **43**:281–283.

75. **Nanopublication schema**. http://nanopub.org/nschema .
76. Garcia Castro L, McLaughlin C, Garcia A: **Biotea: RDFizing PubMed Central in support for the paper as an interface to the Web of Data**. *J Biomed Semant* 2013, **4**(Suppl 1):S5.
77. **data.elsevier.com**. http://data.elsevier.com/documentation/index.html .
78. Page KR, Fields B, De Roure D, Crawford T, Downie JS: **Capturing the workflows of music information retrieval for repeatability and reuse**. *J Intell Inf Syst* 2013, **41**:435–459.
79. Garrido J, Soiland-Reyes S, Enrique Ruiz J, Sanchez S: **AstroTaverna: Tool for Scientific Workflows in Astronomy**. *Astrophys Source Code Libr* 2013.
80. Mina E, Thompson M, Zhao J, Hettne K, Schultes E, Roos M: **Nanopublications for exposing experimental data in the life-sciences: a Huntington's Disease case study**. *SWAT4LS, Vol 1114 CEUR Work Proceedings, CEUR-WS.org* 2013.
81. **Huntington's Disease study Research Object**. http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2/ .
82. **ResearchObject.org**. http://www.researchobject.org/ .
83. Research Object examples. http://www.researchobject.org/initiative/.

Chapter 4: Down-regulation of the acetyl-CoA metabolic network in adipose tissue of obese diabetic individuals and recovery after weight loss

Harish Dharuri
Peter A.C. ’t Hoen
Jan B. van Klinken
Peter Henneman
Jeroen F.J. Laros
Mirjam A Lips
Fatiha el Bouazzaoui
Gert-Jan van Ommen
Ignace Janssen
Bert van Ramshorst
Bert A. van Wagensveld
Hanno Pijl
Ko Willems van Dijk
Vanessa van Harmelen

Diabetologia. 2014;57(11):2384-92

ABSTRACT

Aims/Hypothesis

Not all obese individuals develop type-2 diabetes. Why some obese individuals remain normal glucose tolerant (NGT) is not well understood. We hypothesize that the biochemical mechanisms that underlie the function of adipose tissue can help explain the difference between obese individuals with NGT and those with type 2 diabetes.

Methods

RNA-sequencing was used to analyse the transcriptome of samples extracted from visceral adipose tissue (VAT) and subcutaneous adipose tissue (SAT) of obese women with NGT or type 2 diabetes who were undergoing bariatric surgery. The gene expression data was analysed by bioinformatic visualization and statistical analyses techniques.

Results

A network-based approach to distinguish obese individuals with NGT from obese individuals with type 2 diabetes identified acetyl-CoA metabolic network down-regulation as an important feature in the pathophysiology of obese individuals with type 2 diabetes. In general, genes within two reaction steps of acetyl-CoA were found to be down-regulated in the VAT and SAT of individuals with type 2 diabetes. Upon weight loss and amelioration of metabolic abnormalities three months following bariatric surgery, the expression level of these genes recovered to levels seen in NGT individuals. We report four novel genes associated with type-2 diabetes and recovery upon weight loss: acetyl-CoA acetyltransferase 1 (*ACAT1*), acetyl-CoA carboxylase alpha (*ACACA*), aldehyde dehydrogenase 6 family, member A1 (*ALDH6A1*) and methylenetetrahydrofolate dehydrogenase (*MTHFD1*).

Conclusion/Interpretation

Down-regulation of the acetyl-CoA network in VAT and SAT is an important feature in the pathophysiology of type 2 diabetes in obese individuals. ACAT1, ACACA, ALDH6A1 and MTHFD1 represent novel biomarkers in adipose tissue associated with type 2 diabetes in obese individuals.

INTRODUCTION

Obesity is associated with increased risk of premature death and has reached epidemic proportions in modern societies [1]. Obesity results in decreased life expectancy due to associated metabolic and cardiovascular disorders, as well as several types of cancer [2, 3]. A majority of obese individuals develop insulin resistance and type-2 diabetes. However, approximately 10-25% of these individuals seem to remain insulin sensitive and metabolically "healthy" [4]. Studies have shown that the expanded adipose tissue serves as an important pathogenic site in the development of type 2 diabetes [5]. Furthermore, the prevalence of metabolically "healthy" obese has been attributed to a normal adipose tissue function [5]. A criterion for distinguishing the obese subtypes is of crucial importance to develop appropriate intervention and prevention strategies for these individuals [6]. Most studies have focussed on developing risk scores based on blood pressure, lipid levels, glucose homeostasis, and inflammatory parameters to distinguish the metabolically "healthy" from the metabolically abnormal [7, 8]. However, the biological mechanisms underlying the phenotypic differences observed among obese individuals are not fully understood. In view of the central role of adipose tissue in the manifestation of obesity pathology, we investigated gene expression and biochemical pathway profiles in visceral adipose tissue (VAT) and subcutaneous adipose tissue (SAT) in a human cohort comprised of very obese individuals (BMI>40 kg/m^2) who had normal glucose tolerance (NGT) or who had type-2 diabetes.

Whole genome expression profiling of both SAT and VAT presents an opportunity to study the development of disease in the adipose tissue depots and to delineate biological processes explaining the dysregulation of metabolism in these tissues. Earlier studies used microarray analyses to compare gene expression profiles in the SAT and VAT of obese individuals and found co-regulation of immune and metabolic genes with insulin resistance and metabolic syndrome [9-11]. We have employed next-generation RNA sequencing technology as it offers extensive coverage, precise quantitation of transcripts, and a large dynamic range [12-14].

The current study applied bioinformatic visualization and statistical analyses techniques to the gene expression data and showed dysregulated acetyl-CoA metabolism as a distinguishing feature of obese individuals with type 2 diabetes. Multiple genes in the immediate vicinity of the acetyl-CoA reaction network were down-regulated in diabetic obese individuals. To ascertain if the down-regulation of these genes was correlated to health status, we studied expression levels of these genes before and three months after bariatric surgery associated with significant weight loss and improvement of morbidity.

Table 1 Characteristics of participants with NGT and type 2 diabetes at baseline and 3 months post-bariatric surgery

Characteristic	NGT		T2DM		*p* value			
	Baseline	3 months post-surgery	Baseline	3 months post-surgery	T2DM vs NGT (baseline)	T2DM vs NGT (3 months)	NGT 3 months vs baseline	T2DM 3 months vs baseline
n	17	17	15	15				
Age (years)	49±6	49±6	53±5	53±5	NS	NS	NS	NS
BMI (kg/m^2)	42.9±3.2	36.9±3.3	43.4±4.4	35.9±4.0	NS	NS	1.38×10^{-15}	8.21×10^{-16}
Weight (kg)	122.2±3.1	105.0±2.8	118.9±4.5	98.9±3.7	NS	NS	1.23×10^{-14}	2.11×10^{-14}
HOMA-IR	2.79±2.05	1.72±1.62	4.25±3.26	1.68±0.91	0.06	NS	0.09	0.001
Fasting glucose (mmol/l)	5.08±0.54	5.08±0.76	9.28±2.61	5.87±1.21	2.03×10^{-10}	NS	NS	6.9×10^{-8}
Fasting insulin (pmol/l)	72.5±49.9	42.9±34.9	59.4±40.0	38.9±19.7	NS	NS	0.006	0.09
HbA_{1c}								
mmol/mol	37.6±2.3	34.1±0.9	55.0±4.3	40.2±1.8	8.2×10^{-6}	0.10	NS	0.0002
%	5.6	5.3	7.2	5.8	8.2×10^{-6}	0.10	NS	0.0002
Triacylglycerol (mmol/l)	1.49±0.17	1.30±0.13	2.02±0.19	1.32±0.14	0.03	NS	0.03	1.13×10^{-7}
NEFA (mmol/l)	0.99±0.07	1.16±0.08	1.18±0.11	1.14±0.09	NS	NS	0.06	NS
Total cholesterol (mmol/l)	4.84±0.25	4.20±0.18	4.34±0.22	3.49±0.20	NS	0.03	0.002	0.0006
HDL-cholesterol (mmol/l)	1.14±0.07	1.05±0.05	1.10±0.09	1.05±0.07	NS	NS	0.050	NS
LDL-cholesterol (mmol/l)	3.03±0.21	2.42±0.20	2.33±0.17	1.84±0.19	0.02	0.050	0.005	0.053
CRP (mg/l)	7.74±1.90	6.16±2.26	8.30±1.95	4.13±1.00	NS	NS	NS	0.004

Data are means±SD

Statistical differences between NGT and T2DM and pre- and post-intervention groups were determined with a mixed-effects model, where subject- specific deviances were modelled with random intercepts

CRP, C-reactive protein; T2DM, type 2 diabetes

RESEARCH DESIGN AND METHODS

Participants

The study group consisted of 17 obese women with NGT (with normal fasting glucose levels) and 15 obese women with type 2 diabetes (classified according to WHO standards). The groups were matched for age, weight and BMI (Table 1). All the women had been morbidly obese (BMI>40 kg/m2) for at least five years. Participants who reported the use of weight loss medications within 90 days prior to enrolment in the study were excluded. Body weight of all participants had been stable for at least 3 months prior to inclusion. The participants were investigated in the morning after an overnight fast. A venous blood sample was taken for the determination of plasma glucose (by the routine chemistry laboratory at the hospital) and insulin (by IRMA; Medgenix, Fleurus, Belgium). Thereafter, SAT was obtained from the parumbilical region by needle aspiration under local anesthesia using lidocaine. Around four weeks after the first examination all individuals underwent bariatric surgery (gastric bypass/banding). Within 1h after opening the abdominal wall adipose tissue specimens were taken from the epigastric region of the abdominal wall (SAT) and from the major omentum (VAT). One piece of these adipose tissues was immediately put in RNA-later (Ambion®, Life Technologies, Bleiswijk, The Netherlands) and subsequently stored at -80°C. Another piece of adipose tissue was used for the isolation of adipocytes using collagenase treatment, as described [15]. Three months after the operation, the participants were investigated again after an overnight fast. Plasma glucose and insulin was determined and another SAT needle biopsy was taken. The participants were not calorie restricted in the period prior to the bariatric surgery.The study was approved by the Ethics Committee of Leiden University. All participants gave informed consent to participate in the study.

Medication

For obvious reasons we could not restrict to obese participants not using any type of medication. All participants were allowed to use cholesterol lowering statins and antihypertensive medication. The use of drugs such as statins and antihypertensive drugs was slightly higher in the diabetic participants. At baseline, statins were used by 60% of patients with type 2 diabetes and 18% of patients with NGT. Of the diabetic patients 75% used anti-hypertensives against 40% in individuals with NGT. A substantial proportion of patients with type 2 diabetes received treatment with metformin (n=9; 60%) or sulfonylurea derivatives (n=4; 25%).

Table 2 Top 25 genes up- or downregulated in VAT of diabetic individuals

Gene	Coefficient NGT vs T2DM	p- value NGT vs T2DM	Adjusted p-value NGT vs T2DM
ALDH6A1	-0.670	1.49E-06	0.005502
C14orf45	-0.462	1.59E-06	0.005502
ECHS1	-0.521	1.48E-06	0.005502
IRS1	-0.601	3.41E-07	0.005502
STBD1	-0.615	6.74E-07	0.005502
IARS2	-0.311	2.73E-06	0.006958
NAT8L	-0.745	2.81E-06	0.006958
AIFM2	-0.452	3.24E-06	0.007013
ATPAF1	-0.349	3.71E-06	0.007141
ACAD9	-0.311	8.28E-06	0.010501
GPI	-0.285	8.25E-06	0.010501
HADH	-0.575	8.49E-06	0.010501
HSPD1	-0.299	7.74E-06	0.010501
MTHFD1	-0.423	6.16E-06	0.010501
ACACA	-0.560	9.14E-06	0.010554
MAP3K15	-0.433	1.19E-05	0.012882
HK2	-0.712	1.32E-05	0.01298
PARVG	0.654	1.5E-05	0.01298
PDHA1	-0.375	1.48E-05	0.01298
PRKAR2B	-0.716	1.39E-05	0.01298
ACAT1	-0.406	1.81E-05	0.012994
ATP9A	-0.400	2.1E-05	0.012994
CEBPA	-0.566	1.97E-05	0.012994
DARS2	-0.379	1.64E-05	0.012994
NXPH4	-1.002	1.89E-05	0.012994

Coefficient NGT vs T2DM: log fold change of NGT vs T2DM; a negative value reflects downregulation whereas a positive value reflects upregulation of the gene in type 2 diabetic individuals

For the complete list of up- or downregulated genes in VAT of type 2 diabetic individuals see ESM Table 2

The adjusted *p* value NGT vs T2DM is the *p* value after Benjamini– Hochberg FDR correction

Table 3 Top 25 genes up- or downregulated in SAT of diabetic individuals

Gene	Coefficient NGT vs T2DM	p- value NGT vs T2DM	Adjusted p-value NGT vs T2DM
DHTKD1	-0.39953	3.38E-06	0.027658
DPEP2	0.941324	3.63E-06	0.027658
S100A11	0.389024	4.79E-06	0.027658
IRS1	-0.64306	7.26E-06	0.027696
BIVM	-0.32809	8E-06	0.027696
CRABP2	0.889426	1.15E-05	0.033234
PXMP2	-0.46718	1.65E-05	0.03571
LSP1	0.826079	1.53E-05	0.03571
RNF14	-0.29276	2.01E-05	0.038745
FXYD5	0.508216	3E-05	0.041435
TYROBP	0.789776	2.74E-05	0.041435
CYBA	0.573909	2.8E-05	0.041435
THNSL1	-0.48462	3.11E-05	0.041435
ALDH6A1	-0.59541	5.12E-05	0.042281
C14orf45	-0.39723	0.000107	0.042281
HADH	-0.45138	0.000145	0.042281
MTHFD1	-0.3727	7.95E-05	0.042281
MAP3K15	-0.39465	9.36E-05	0.042281
SLC2A4	-0.73171	0.000105	0.042281
ME1	-0.45845	9.99E-05	0.042281
LDHD	-0.53027	9.59E-05	0.042281
FAN1	-0.26323	5.17E-05	0.042281
TMEM218	-0.39528	0.000128	0.042281
EEPD1	-0.45794	0.000156	0.042281
IL2RG	0.835802	0.000114	0.042281

Coefficient NGT vs T2DM: log fold change of NGT vs T2DM; a negative value reflects downregulation whereas a positive value reflects upregulation of the gene in type 2 diabetic individuals

The adjusted *p* value NGT vs T2DM is the *p* value after Benjamini– Hochberg FDR correction

Isolation of RNA

Total RNA was isolated using the Nucleospin RNA kit (Macherey-Nagel, Düren, Germany) according to the instructions of the manufacturer. The quality of each mRNA sample was examined using the Agilent 2100 Bioanalyzer (Santa Clara, CA). All RNA samples had a RIN value >7.

RNA Deep Sequencing

RNA (fifty µg) of the adipose tissue samples obtained during bariatric surgery were used for RNA deep sequencing which was performed at the Beijing Genomics Institute (BGI) using RNA-Seq (Transcriptome) sequencing on the HiSeq2000 with 90 nucleotide long Paired End reads, resulting in a minimum of 3Gb clean data per sample. The reads were aligned to the Human reference genome build 19 (hg19) to obtain a histogram of coverage per exon and the associated count data (**ESM Methods 1**). Differential expression analysis was done on exon, gene and transcript levels as described in **ESM Methods 1**.

Bioinformatic analysis

The bioinformatic analysis was performed as described in **ESM Methods 2.**

Quantitative Real Time PCR for comparison of pre and post-surgery gene expression data for select members of acetyl-CoA gene set

The RNA of the needle biopsies obtained pre and post bariatric surgery as well as the RNA obtained from the adipocytes during bariatric surgery were used for quantitative real-time PCR (**See ESM Methods 3**).

RESULTS

Characteristics of participants at baseline and three months post-bariatric surgery

Characteristics of the participants are shown in **Table 1**. At baseline fasting glucose, HbA1c and triglyceride levels were significantly higher in individuals with type 2 diabetes than in those with NGT. Three months post-surgery, individuals with NGT and type 2 diabetes showed the same weight-reduction. Fasting glucose, HbA1c and triglyceride levels were significantly reduced in the diabetic individuals and similar to levels in the individuals with NGT.

Gene expression analysis

We utilized RNA-sequencing to analyse the transcriptome of samples extracted from VAT and SAT of 32 (15 with type 2 diabetes, 17 with NGT)

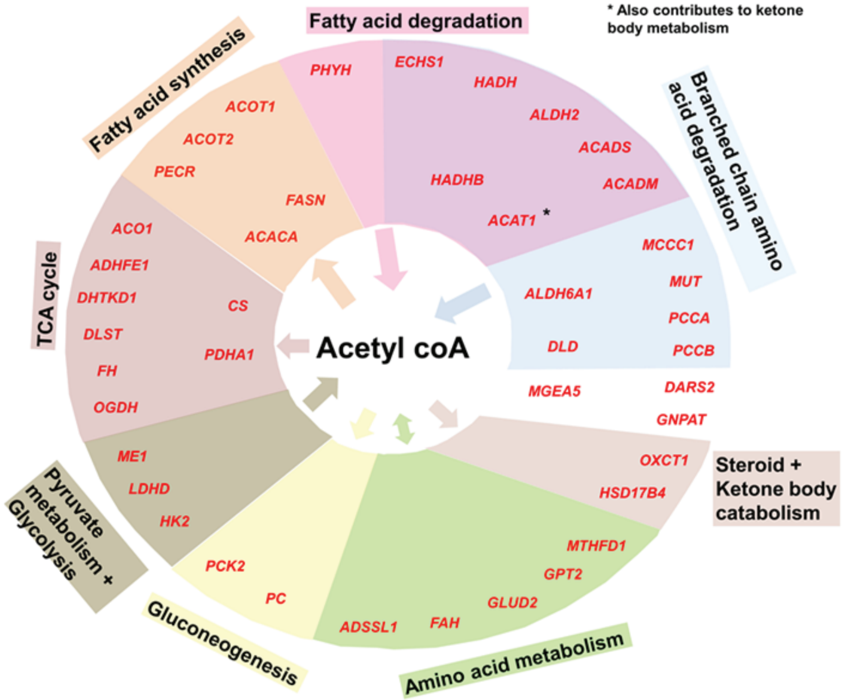


Figure 1 Downregulation of the acetyl-CoA gene network in type 2 diabetes. Forty-two genes that are among the top differentially expressed genes in VAT are also members of the acetyl-CoA gene set. The genes within the inner circle act directly on acetyl-CoA while the genes in the outer circle participate one reaction step away from acetyl-CoA. All the genes were downregulated in VAT. *Also contributes to ketone body metabolism. TCA, tricarboxylic acid cycle

obese female individuals undergoing bariatric surgery (Table 1). We first determined whether the overall gene expression profiles differed between obese women with type 2 diabetes and those with NGT and applied the global test [16] on all expressed genes. The global test on VAT and SAT yielded a p-value of 3.7E-03 and 9.4E-04 respectively indicating a significant association of gene expression with health status.

Gene-level analysis with the limma package in R identified 168 genes differentially expressed in VAT ($p<0.05$, after Benjamin-Hochberg FDR correction) between obese individuals with NGT and those with type 2 diabetes (Table 2 and ESM Table 2). Applying the same method on SAT yielded 121 genes that were significantly differentially expressed between obese individuals with NGT and those with type 2 diabetes (Table 3). There was an overlap of 24 of the differentially expressed genes between the two tissues.

Bioinformatic analysis to identify sub-networks in gene expression data

We further investigated biological mechanisms underlying the differential health status among the participants. Statistically significant differentially expressed genes ($p<0.05$ after FDR correction) in VAT and SAT were used as an input to a pathway-based over-representation analysis tool made available by ConsensusPathDB (http://cpdb.molgen.mpg.de/, accessed 14 January 2013). This analysis of genes from VAT identified pathways relevant to carbon, amino acid and fatty acid metabolism (**ESM Table 3**). A similar analysis strategy for SAT identified pathways relevant to several bacterial infections, regulation of actin cytoskeleton and Fc-Gamma R-mediated phagocytosis (ESM Table 4). The overlap between significant (q-value<0.05) pathways identified for the two tissues is limited to insulin-signalling, branched-chain amino acid degradation and pyruvate metabolism. Furthermore, to determine if significantly differentially expressed genes in each of the two tissues operate in close proximity in network space, we utilized “Network neighbourhood-based entity sets” (NEST) a software tool made available by ConsensusPathDB. ESM Table 5 shows the result for an input of top differentially expressed genes in VAT (168 genes, $p<0.05$ after multiple test correction). This analysis indicated that the differentially expressed genes in VAT operate in a network neighbourhood at the intersection of carbohydrate, amino acid and fatty acid metabolism. Importantly, a majority of the genes mapped onto these pathways were present in close proximity in network space to acetyl-CoA metabolism (Figure 1). A similar approach using NEST with the significant hits from SAT did not yield any statistically significant sets.

The acetyl-CoA metabolic network is down-regulated in diabetic obese individuals

The enriched network neighbourhood-based sets described above hinted at the possibility of acetyl-CoA metabolic network being a common feature of the statistically significant differentially expressed genes in VAT. To evaluate if genes within two reaction steps of acetyl-CoA metabolism were significantly represented among the top hits in VAT, a gene-set was generated using the Taverna workflow management system and the KEGG pathway database (ESM Methods 4). This approach involved finding all the genes that participate within a radius of 2 steps in the reaction space surrounding acetyl-CoA. This algorithm was implemented in Taverna and the pathway information present in the KEGG database was used to generate the gene set. The total number of genes in the acetyl-CoA set is 419.

We then performed statistical tests to determine if members of the acetyl-CoA gene set were significantly represented among top hits in VAT. The number of genes among the 168 top hits in VAT that are also members of the acetyl-CoA gene set is 42 (ESM Table 2), ten times more than expected by chance (*p*=1E-63, permutation test), indicating that the presence of the members of acetyl-CoA gene set among the top hits due to chance alone is negligible. All these 42 genes were down regulated in VAT of obese individuals with type 2 diabetes (ESM Table 2). Additionally, the global test to evaluate the acetyl-CoA gene set as a predictor of health status in VAT and SAT yielded a p-value of 2.4E-02 and 8.4E-03 respectively. The network-neighbourhood test did not yield a significant set for SAT, yet the acetyl-CoA gene set is more significant in SAT than in VAT because most of the genes in the acetyl-CoA gene set are borderline significant in SAT. These genes fail to make the cut-off necessary to be included for network neighbourhood tests. However, the global test takes into account the p-value of all the entities in the gene set, and since most genes have modest p-values in SAT, the overall p-value generated for the acetyl-CoA gene set in that tissue type is lower than we would expect by examining the network neighbourhood of the most significant genes. In conclusion, genes in the acetyl-coA reaction network displayed a general down-regulation in both VAT and SAT of individuals with type 2 diabetes.

Analysis at the transcript or exon level

We investigated possible differential splicing events, comparing obese individuals with NGT and type 2 diabetes, for the 42 genes in the acetyl-CoA gene set. To do so, we analysed differences at the 1) transcript level, 2) expression level of individual exons. Of the 42 genes, there were 16 genes with multiple annotated transcripts. All of the transcript variants were significantly down-regulated in the individuals with type 2 diabetes as compared with the individuals with NGT (data not shown).

At the exon level, we did not identify any exon that deviated significantly from the overall gene expression pattern and did not obtain any evidence for alternative splicing between individuals with NGT and those with type 2 diabetes (data not shown).

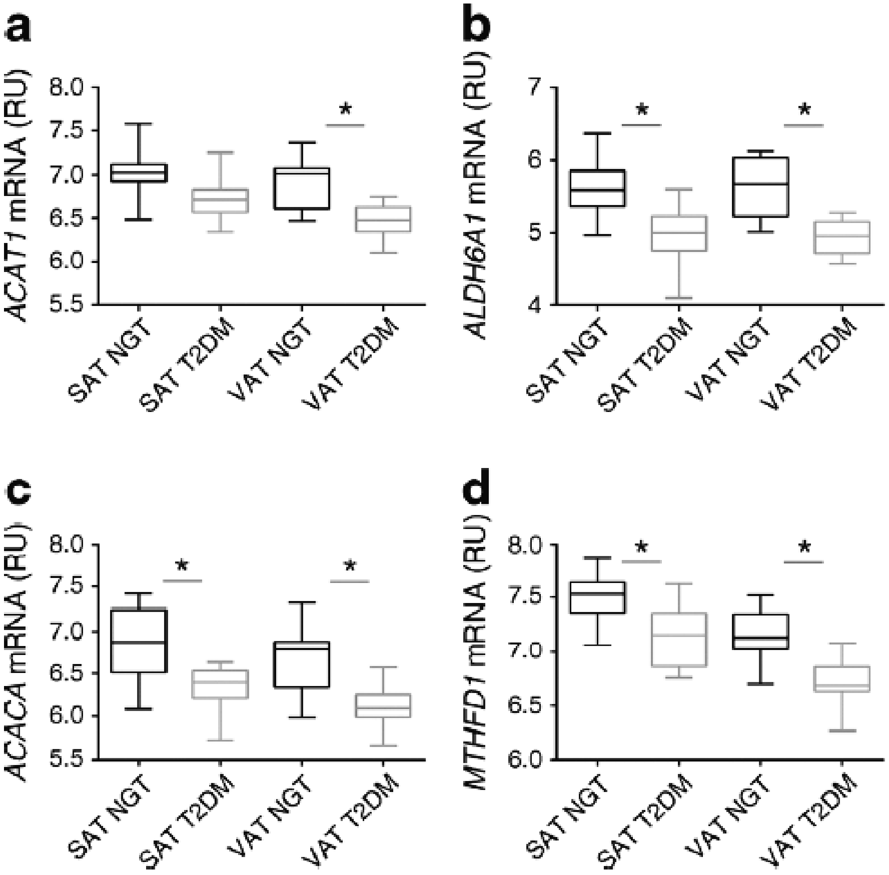


Figure 2 Gene expression of acetyl-CoA network genes in VAT and SAT. Box plots of normalised gene expression profiles (relative units [RU]: log2-scale) of a few representative genes, ACAT1 (a), ALDH6A1 (b), ACACA (c), MTHFD1 (d), in the acetyl-CoA reaction network that are downregulated (*adjusted p value <0.05 for indicated comparison) in both VAT and SAT of obese individuals with type 2 diabetes (grey bars) compared with NGT (black bars). The whiskers in the boxplots represent the upper and lower limits of the data. T2DM, type 2 diabetes

Down-regulation of genes in the acetyl-CoA reaction network recovers after weight loss

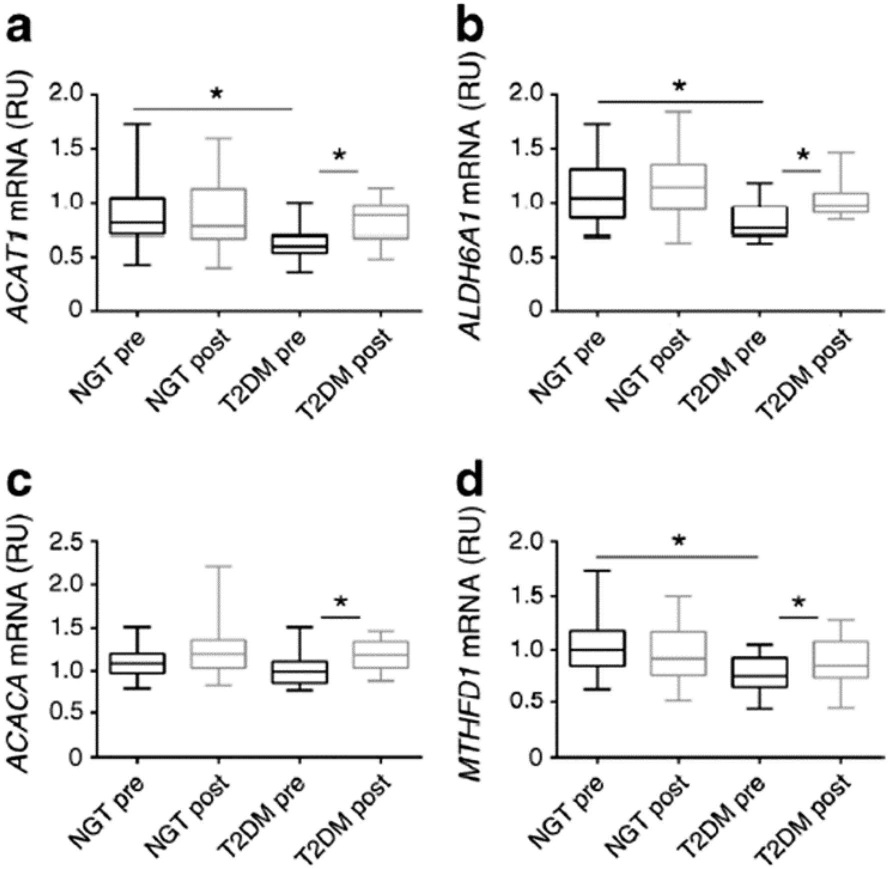


Figure 3 Gene expression of acetyl-CoA network genes in obese individuals with type 2 diabetes are normalised after bariatric surgery. Box plots of expression levels of four representative genes, ACAT1 (a), ALDH6A1 (b), ACACA (c), MTHFD1 (d) (as determined by quantitative PCR, corrected for housekeeping gene, linear scale: relative units [RU]), in type 2 diabetes and NGT before (black bars) and after bariatric surgery (grey bars). T2DM, type 2 diabetes. *p<0.05 (mixed-model-analysis). The whiskers in the boxplots represent the upper and lower limits of the data.

Among the 24 genes that overlapped between the statistically significant top hits in VAT and SAT, 9 genes are members of the acetyl-CoA gene set *(ACACA, ALDH6A1, MTHFD1, HADH, ME1, PC, LDHD, DHTKD1, and GNPAT)*. The gene expression profile of all the 9 genes from the RNA-Seq experiments shows a

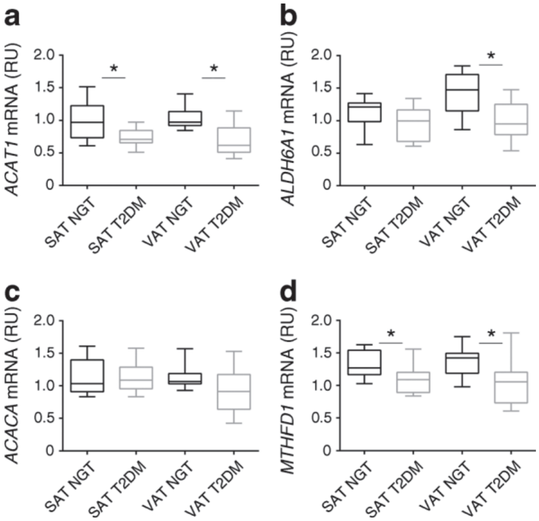


Figure 4 Gene expression of acetyl-CoA network genes in adipocytes. Adipocytes were isolated from SAT and VAT of individuals with type 2 diabetes (grey bars) and NGT (black bars). Gene expression of four representative genes, ACAT1 (a), ALDH6A1 (b), ACACA (c), MTHFD1 (d), was measured using quantitative PCR, corrected for housekeeping gene expression and plotted on a linear scale (RU). The whiskers in the boxplots represent the upper and lower limits of the data. T2DM, type 2 diabetes. *p<0.05 (t test) NGT vs T2DM

consistent down-regulation among individuals with type 2 diabetes in both adipose tissues. The boxplot depicting the expression levels in each of the tissues for both health types is shown for some of the acetyl-CoA genes in Figure 2.

To ascertain whether the down-regulation of the acetyl-CoA genes was correlated to type 2 diabetes, we compared the pre and post-surgery (3 months after) expression levels of these genes in SAT by qPCR. At this time the majority of diabetic obese women had a significantly improved metabolic health status as evidenced by lower fasting glucose levels (Table 1). We observed a statistically significant up-regulation of acetyl-CoA carboxylase alpha (*ACACA*) (p=9.3E-03), aldehyde dehydrogenase 6 family, member A1 (*ALDH6A1*) (p=4.1E-05) and methylenetetrahydrofolate dehydrogenase (*MTHFD1*) (p=4.7E-02) post-surgery in individuals with type 2 diabetes when compared with the changes in expression level observed in individuals with NGT (Fig 3). Also acetyl-CoA acetyltransferase 1 (*ACAT1*) which is at the intersection of the acetyl-CoA network (Fig 1) was up-regulated post-surgery in type 2 diabetes (p=2.3E-03). Three other genes, encoding dehydrogenase E1 and transketolase domain (*DHTKD1*), lactate dehydrogenase (*LDHD*) and pyruvate carboxylase (*PC*) displayed a similar up-regulation post-surgery among individuals with type 2 diabetes but did not reach the statistical p-value threshold of 0.05. This indicates that the improved health status of diabetic individuals post-surgery is associated with a reversal of the disturbance in the acetyl-CoA metabolic network.

Gene expression of acetyl-CoA network in isolated adipocytes

As adipose tissue not only consists of adipocytes but is a mixture of cells, including endothelial cells and leukocytes we determined whether the down-regulation of the acetyl-CoA network in diabetic individuals specifically takes place in the adipocytes of the diabetic individuals. Indeed isolated adipocytes of diabetic individuals showed reduced gene expression levels for *ALDH6A1, ACAT1* and *MTHFD1* (Figure 4).

DISCUSSION

We have performed an in depth comparison of gene expression in SAT and VAT of severely obese women with and without type 2 diabetes. Network analyses revealed that the acetyl-CoA network was dysregulated in type 2 diabetes and that specific genes directly associated with acetyl-CoA metabolism were down-regulated in both VAT and SAT. Importantly, upon weight loss and amelioration of metabolic abnormalities, the expression of these genes in SAT recovered to the corresponding level among NGT women. These results imply that down-regulation of the acetyl-CoA network in VAT and SAT is a marker for the metabolic dysregulation characteristic of type 2 diabetes and, moreover, that it is reversible.

Network-based approaches have emerged as a powerful tool to unravel the mechanisms underlying complex traits [17-19]. Biological networks consist of molecular entities called nodes and functional interconnections between them called edges. An important property of these networks is that they are “scale-free” in that some nodes called “hubs” are connected to a substantially large number of other nodes and therefore considered essential for maintaining the integrity of the cell [18]. In general, these systems are robust against random mutations but are vulnerable to attacks against the hub [17]. Acetyl-CoA is a key hub metabolite of the metabolic network and plays a critical role in maintaining cellular homeostasis [20]. Previous studies have implicated branched-chain amino acid degradation (BCAD) [21], fatty-acid oxidation [22, 23], and citrate cycle [22, 23] dysregulation as a characteristic feature of type 2 diabetes and related traits. In this study, in addition to confirming the previous findings, we argue that the acetyl-CoA reaction network is a unifying principle and that its dysregulation distinguishes between obese women with type 2 diabetes and those with NGT.

Acetyl-CoA lies at the crossroads of glycolysis, citrate cycle, ketogenesis, lipid synthesis, amino acid and fatty acid metabolism, suggesting that the metabolite may play a key role as an energy sensor in the cell [20]. Carbon skeletons of sugars, amino acids and fatty acids are degraded to the acetyl group to form acetyl-CoA that enters the citric acid cycle for energy generation. In addition, it is known to modulate gene expression through its role as a co-factor of histone acetyl-transferases (HAT) which enable the transcription of genes through histone acetylation at chromatin structures [24]. Cai et al argue that the primordial role of protein acetylation could have been to enable a cell to modulate gene expression/protein function in tune with the carbon source availability [25]. In other words, the acetyl-CoA is likely to serve as a fundamental and widely conserved gauge of metabolic state. A disturbance in this gauge may contribute to metabolic diseases such as type 2 diabetes as a consequence of altered cell metabolism and transcriptional regulation.

We report four genes associated with type 2 diabetes and recovery in the SAT of obese individuals: *ACAT1, ACACA*, *ALDH6A1* and *MTHFD1*. These genes all participate in the immediate vicinity of acetyl-CoA metabolism and are known hotspots of human metabolism, with *ACAT1*, *ALDH6A1* and *ACACA* recorded among inborn errors of metabolism (IEM) (OMIM: 203750, 614105 613933 respectively). IEMs are congenital metabolic defects arising due to single or multiple enzyme deficiencies. Recently [26], IEMs have been mapped onto a mathematical reconstruction of human metabolism [27]. Analyses of IEMs in the context of network topology led to the observation that the IEMs are adjacent to each other with acetyl-CoA acting as the central metabolite. This clearly suggests that the vicinity of acetyl-CoA in the network topology is a hub where abnormalities in individual genes potentially accumulate and upon reaching a certain risk threshold lead to the manifestation of disease.

The genes reported in this study function at critical decision points in cellular biochemical pathways as illustrated by *ACAT1.* The latter enzyme mediates the reversible conversion of 2 molecules of acetyl-CoA to acetoacetyl-CoA

[28]. This enzyme catalyzes the final step in branched-chain amino acid and fatty acid degradation pathways and the acetyl-CoA produced here is used as an input for the citric acid cycle (http://www.genome.jp/dbget-bin/www_bget?hsa:38). When energetics favors the production of acetoacetyl-CoA in this reaction step, the metabolite is used for ketone body synthesis [28]. *ACAT1* also mediates the first step in the mevalonate pathway whose end-product Farnesyl-PP is a precursor for cholesterol among other several important metabolites (http://www.genome.jp/dbget-bin/www_bget?hsa:38). Therefore, the *ACAT1* enzyme is strategically placed at the intersection of important cellular pathways that respond to the energy status of the cell.

Intriguingly, additional genetic evidence for a role of *ACAT1* in type 2 diabetes is provided by a genome-wide association study (GWAS) in a UK prospective diabetes study that investigated the glycemic response to metformin and reported a Single Nucleotide Polymorphism (SNP), rs11212617, associated with metformin success [29]. Based on the proximity to the polymorphism, the study concluded *ATM* (ataxia telangiectasia mutated) as the causal gene that plays a role in metformin success and that the variation at this gene alters the glycemic response to metformin. However, re-analyzing the polymorphism rs11212617, we found that the polymorphism is in fact an eQTL for the nearby *ACAT1* gene and not *ATM*. The confirmation for this eQTL is provided by two independent studies; Zeller et al who studied the monocyte transcriptome to determine eQTLs of relevance to human disease [31] (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/) and the data from the GEUVADIS consortium [31, 32], where the SNP was found to be an eQTL for *ACAT1* (nominal p-value=1.1e-6). This means that the variation in the expression level of *ACAT1* alters the glycemic response to metformin and therefore plays a role in the success of metformin treatment. Furthermore, this clearly suggests that *ACAT1* plays a role in type 2 diabetes. Individuals with the polymorphism that alters its expression level may represent a sub-type among individuals with type 2 diabetes, perhaps with different response to metformin.

There were differences in the usage of medication between obese women NGT and type 2 diabetes, especially in the usage of metformin, which was not used by any of the NGT women and by 60% of the women with type 2 diabetes. As metformin acts on enzymes within the acetyl-coA network and affects lipid and glucose metabolism, the usage of metformin may have confounded our results, but we have not found any evidence for this: 1) There was no difference in gene expression of *ACAT1*, *ALDH6A1*, *ACACA* and

MTHFD1 between metformin and no metformin users (ESM Fig. 1). 2) When metformin users were excluded from the comparison between individuals with NGT and those with type 2 diabetes, there was still a down-regulation of *ACAT1*, *ALDH6A1*, *ACACA* and *MTHFD1* in the women with type 2 diabetes (ESM Fig. 2).

Our cohort consisted of severely obese women. We do not know whether the observed differences were a consequence of the metabolic defects that occur in type 2 diabetes (i.e. hyperglycemia) or represented the underlying etiology of type 2 diabetes. However, a previous study that used microarrays to analyse gene expression in adipose tissue showed that during the progression from the lean to the obese state and then further towards the metabolic syndrome the genes involved in metabolic processing were gradually down-regulated [10]. These data suggest that the down-regulation of metabolic pathways underlie the pathology of type 2 diabetes.

Previous studies have postulated that low-grade inflammation of the adipose tissue plays an important role in the development of insulin resistance [33-36]. For example, a recent study in monozygotic twins discordant for obesity showed that SAT transcript profile in the metabolically healthy obese is characterized by the maintenance of mitochondrial function and absence of inflammation [35]. This is in line with the results in our study, where we observe an inverse correlation pattern of differential expression of genes that are down-regulated in metabolic and up-regulated in inflammatory pathways in VAT and SAT of individuals with type 2 diabetes.

In summary, our results demonstrate that the acetyl-CoA network is dysregulated in VAT and SAT of obese women with type 2 diabetes. We find significant down-regulation of several genes in the immediate vicinity of acetyl-CoA and report a statistically significant recovery for 4 genes after amelioration of the metabolic abnormalities in SAT. Further research into the causal role of down-regulation of the acetyl-CoA network in type 2 diabetes should indicate whether direct intervention in the acetyl-CoA network will provide novel therapeutic approaches.

References

1. Malik VS, Willett WC, Hu FB: **Global obesity: trends, risk factors and policy implications.** *Nat Rev Endocrinol* 2013, **9**:13–27.

2. Mokdad AH, Ford ES, Bowman BA, Dietz WH, Vinicor F, Bales VS, Marks JS: **Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001**. *JAMA : the journal of the American Medical Association* 2003, **289**(1):76-79.
3. Van Gaal LF, Mertens IL, De Block CE: **Mechanisms linking obesity with cardiovascular disease**. *Nature* 2006, **444**(7121):875-880.
4. Bluher M: **Are there still healthy obese patients?** *Current opinion in endocrinology, diabetes, and obesity* 2012, **19**(5):341-346.
5. Bluher M: **Adipose tissue dysfunction contributes to obesity related metabolic diseases**. *Best practice & research Clinical endocrinology & metabolism* 2013, **27**(2):163-177.
6. Kantartzis K, Machann J, Schick F, Rittig K, Machicao F, Fritsche A, Haring HU, Stefan N: **Effects of a lifestyle intervention in metabolically benign and malign obesity**. *Diabetologia* 2011, **54**(4):864-868.
7. Lindstrom J, Tuomilehto J: **The diabetes risk score: a practical tool to predict type 2 diabetes risk**. *Diabetes care* 2003, **26**(3):725-731.
8. Pajunen P, Kotronen A, Korpi-Hyovalti E, Keinanen-Kiukaanniemi S, Oksa H, Niskanen L, Saaristo T, Saltevo JT, Sundvall J, Vanhala M *et al*: **Metabolically healthy and unhealthy obesity phenotypes in the general population: the FIN-D2D Survey**. *BMC public health* 2011, **11**:754.
9. Wolfs MG, Rensen SS, Bruin-Van Dijk EJ, Verdam FJ, Greve JW, Sanjabi B, Bruinenberg M, Wijmenga C, van Haeften TW, Buurman WA *et al*: **Co-expressed immune and metabolic genes in visceral and subcutaneous adipose tissue from severely obese individuals are associated with plasma HDL and glucose levels: a microarray study**. *BMC medical genomics* 2010, **3**:34.
10. Klimčáková E, Roussel B, Márquez-Quiñones A, Kováčová Z, Kováčiková M, Combes M, Šiklová-Vítková M, Hejnová J, Šrámková P, Bouloumié A, Viguerie N, Štich V, Langin D: **Worsening of obesity and metabolic status yields similar molecular adaptations in human subcutaneous and visceral adipose tissue: Decreased metabolism and increased immune response**. *J Clin Endocrinol Metab* 2011, **96**.
11. Qatanani M, Tan Y, Dobrin R, Greenawalt DM, Hu G, Zhao W, Olefsky JM, Sears DD, Kaplan LM, Kemp DM: **Inverse regulation of inflammation and mitochondrial function in adipose tissue defines extreme insulin sensitivity in morbidly obese patients**. *Diabetes* 2013, **62**:855–863.

12. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nature methods* 2008, **5**(7):621-628.

13. Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities**. *Nature reviews Genetics* 2011, **12**(2):87-98.

14. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature reviews Genetics* 2009, **10**(1):57-63.

15. Van Harmelen V, Lonnqvist F, Thorne A, Wennlund A, Large V, Reynisdottir S, Arner P: **Noradrenaline-induced lipolysis in isolated mesenteric, omental and subcutaneous adipocytes from obese subjects**. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity* 1997, **21**(11):972-979.

16. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome**. *Bioinformatics* 2004, **20**(1):93-99.

17. Barabasi AL, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease**. *Nature reviews Genetics* 2011, **12**(1):56-68.

18. Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, Kohane IS, Kasif S: **Network-based analysis of affected biological processes in type 2 diabetes models**. *PLoS genetics* 2007, **3**(6):e96.

19. Chan SY, Loscalzo J: **The emerging paradigm of network medicine in the study of human disease**. *Circulation research* 2012, **111**(3):359-374.

20. Naimi M, Arous C, Van Obberghen E: **Energetic cell sensors: a key to metabolic homeostasis**. *Trends in endocrinology and metabolism: TEM* 2010, **21**(2):75-82.

21. Herman MA, She P, Peroni OD, Lynch CJ, Kahn BB: **Adipose tissue branched chain amino acid (BCAA) metabolism modulates circulating BCAA levels**. *The Journal of biological chemistry* 2010, **285**(15):11348-11356.

22. Dahlman I, Forsgren M, Sjogren A, Nordstrom EA, Kaaman M, Naslund E, Attersand A, Arner P: **Downregulation of electron transport chain genes in visceral adipose tissue in type 2 diabetes independent of obesity and possibly involving tumor necrosis factor-alpha**. *Diabetes* 2006, **55**(6):1792-1799.

23. Dahlman I, Mejhert N, Linder K, Agustsson T, Mutch DM, Kulyte A, Isaksson B, Permert J, Petrovic N, Nedergaard J *et al*: **Adipose tissue pathways involved in weight loss of cancer cachexia**. *British journal of cancer* 2010, **102**(10):1541-1548.

24. Cai L, Sutter BM, Li B, Tu BP: **Acetyl-CoA induces cell growth and proliferation by promoting the acetylation of histones at growth genes**. *Molecular cell* 2011, **42**(4):426-437.

25. Cai L, Tu BP: **On acetyl-CoA as a gauge of cellular metabolic state**. *Cold Spring Harbor symposia on quantitative biology* 2011, **76**:195-202.

26. Sahoo S, Franzson L, Jonsson JJ, Thiele I: **A compendium of inborn errors of metabolism mapped onto the human metabolic network**. *Molecular bioSystems* 2012, **8**(10):2545-2558.

27. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD *et al*: **A community-driven global reconstruction of human metabolism**. *Nature biotechnology* 2013, **31**(5):419-425.

28. Haapalainen AM, Merilainen G, Pirila PL, Kondo N, Fukao T, Wierenga RK: **Crystallographic and kinetic studies of human mitochondrial acetoacetyl-CoA thiolase: the importance of potassium and chloride ions for its structure and function**. *Biochemistry* 2007, **46**(14):4305-4321.

29. GoDarts, Group UDPS, Wellcome Trust Case Control C, Zhou K, Bellenguez C, Spencer CC, Bennett AJ, Coleman RL, Tavendale R, Hawley SA *et al*: **Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes**. *Nature genetics* 2011, **43**(2):117-120.

30. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H *et al*: **Genetics and beyond--the transcriptome of human monocytes and disease susceptibility**. *PloS one* 2010, **5**(5):e10693.

31. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG *et al*: **Transcriptome and genome sequencing uncovers functional variation in humans**. *Nature* 2013, **501**(7468):506-511.

32. t Hoen PA, Friedlander MR, Almlof J, Sammeth M, Pulyakhina I, Anvar SY, Laros JF, Buermans HP, Karlberg O, Brannvall M *et al*: **Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories**. *Nature biotechnology* 2013, **31**(11):1015-1022.

33. Soronen J, Laurila PP, Naukkarinen J, Surakka I, Ripatti S, Jauhiainen M, Olkkonen VM, Yki-Jarvinen H: **Adipose tissue gene expression analysis reveals changes in inflammatory, mitochondrial respiratory and lipid metabolic pathways in obese insulin-resistant subjects**. *BMC medical genomics* 2012, **5**:9.

34. Karelis AD, Faraj M, Bastard JP, St-Pierre DH, Brochu M, Prud'homme D, Rabasa-Lhoret R: **The metabolically healthy but obese individual presents a favorable inflammation profile**. *The Journal of clinical endocrinology and metabolism* 2005, **90**(7):4145-4150.

35. Naukkarinen J, Heinonen S, Hakkarainen A, Lundbom J, Vuolteenaho K, Saarinen L, Hautaniemi S, Rodriguez A, Fruhbeck G, Pajunen P *et al*: **Characterising metabolically healthy obesity in weight-discordant monozygotic twins**. *Diabetologia* 2014, **57**(1):167-176.

36. Phillips CM, Perry IJ: **Does inflammation determine metabolic health status in obese and nonobese adults?** *The Journal of clinical endocrinology and metabolism* 2013, **98**(10):E1610-1619.

Supplementary Section

ESM Methods 1: Methods describing RNA deep sequencing, Alignment and Gene annotation and Differential Gene Expression Analysis

RNA Deep Sequencing

The experimental pipeline followed by BGI consisted of enriching mRNA with the help of oligo(dT) beads. Fragmentation buffer was added to generate short mRNA fragments. Taking these short fragments as templates, random hexamer primers were used to synthesize the first strand cDNA. The second strand cDNA was synthesized using buffer, dNTPs, RNase H and DNA polymerase I. Short fragments were purified with QiaQuick PCR extraction kit and resolved with EB-buffer for end reparation and adding poly(A). The short fragments were then connected with sequencing adaptors. Suitable fragments were then selected for amplification by PCR.

Alignment and gene annotation

After assessing the quality of the raw data using FastQC, version: 0.9.3 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), we aligned the reads to the Human reference genome build 19 (hg19, GRCh37) using GSNAP [1] with the novel splicing option (-N1) enabled. The aligned data was further converted to a sorted BAM file using SAMTools, version: 0.1.18 [2]. For the quantification of the number of nucleotides that were mapped per exon, we used BEDTools, version: 2.13.2 [3] in conjunction with an in-house program (https://git.lumc.nl/lgtc-bioinformatics/ngs-misc/blob/master/src/hist2count.py) to obtain a histogram of coverage per exon and the associated count data. Gene annotation (RefSeq version v54) was retrieved from the UCSC (http://genome.ucsc.edu/cgi-bin/hgTables?db=hg19, retrieved July 9, 2012).

Differential Gene Expression

Differential expression analysis was done on exon, gene and transcript levels. For exon level analyses, we summed the coverage values of all nucleotides in an exon for all unique exons annotated in Ensembl. For transcript and gene level analysis, the coverage in all exonic regions of a transcript gene were summed. Only genes expressed in 75% or more of the samples were retained in the statistical analysis as a filter for low abundant genes. To account for differences in number or reads per sample, count data were normalized with the TMM function from the edgeR package [4]. Data were log-transformed with the voom function from the limma-package

[5]. Weights from the voom transformation were taken into subsequent linear models. A hierarchical linear model was fit with the voom transformed expression data as dependent variables and health status and tissue as the independent variables, using the lmFit function from the limma package. P-values were corrected for multiple testing using Benjamini-Hochberg false discovery rate. Entrez Gene identifiers were retrieved using the biomaRt package v2.12.0 in R.

ESM Methods 2: Bioinformatic analysis to identify sub-networks in gene expression data

Bioinformatic visualization tools

Over-representation analysis tools made available by ConsensusPathDB (http://cpdb.molgen.mpg.de/) were used to investigate the relationship among top differentially expressed genes in VAT and SAT. To determine if significantly differentially expressed genes in each of the two tissues operate in close proximity in network space, we utilized "Network neighbourhood-based entity sets" (NEST).

Acetyl-CoA gene set generation

The Taverna version 2.4 [6] workflow management system was used to generate the gene set for acetyl-CoA. We employed a reaction scheme [7] that can be visualized as expanding by a radius of 2 steps in the reaction space of acetyl-CoA. Specifically in this scheme, the reactions that acetyl-CoA is part of and the compounds that participate in these reactions is determined using information present in the KEGG-database [release 63] [8]. As an intermediate step certain compounds like ATP, ADP, NADP, NADPH were filtered out in order to avoid non-specific connections.

Global test

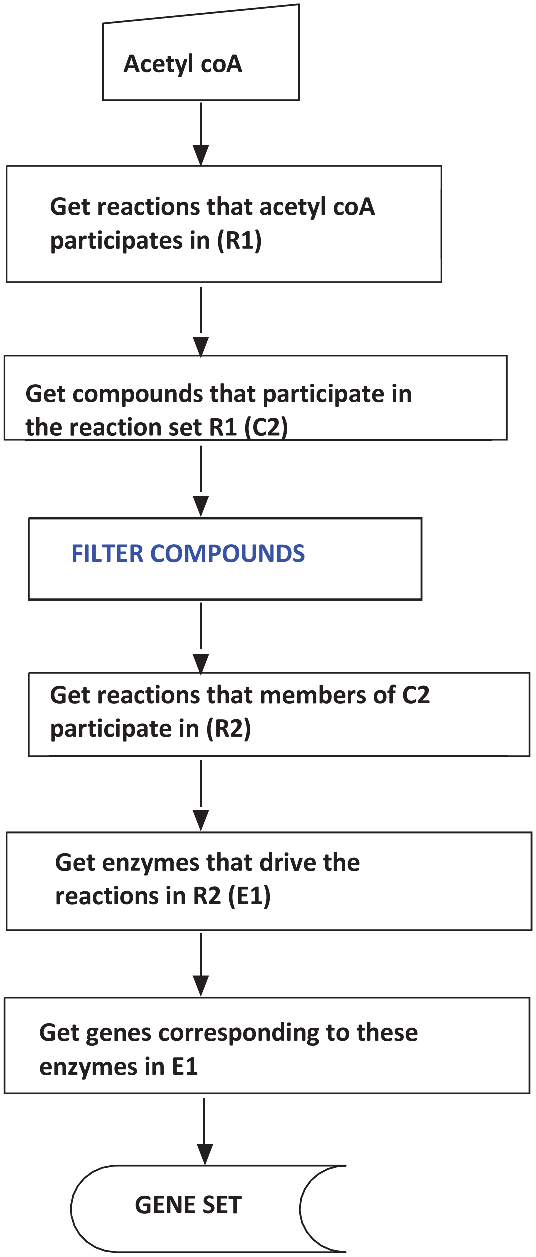
Global test is a statistical method to determine if global expression pattern of a group of genes is significantly related to the phenotype of interest. The global test is available as an R-package at http://www.bioconductor.org/packages/2.13/bioc/html/globaltest.html.
The voom-transformed gene expression data as mentioned earlier was used to determine the association of all the genes as well as to evaluate the association of the acetyl-CoA gene set with T2DM.

ESM Methods 3: Quantitative Real Time PCR for comparison of pre and post-surgery gene expression data for select members of acetyl-CoA gene set

The RNA of the needle biopsies obtained pre and post bariatric surgery as well as the RNA obtained from the adipocytes during bariatric surgery were used for quantitative real-time PCR. Six hundred ng of total RNA was reverse-transcribed with iScript cDNA synthesis kit (Bio-Rad, Hercules, CA) and obtained cDNA was purified with Nucleospin Extract II kit (Macherey-Nagel). Real-Time PCR was carried out on the IQ5 PCR machine (Biorad) using the Sensimix SYBR Green RT-PCR mix (Quantace, London, UK) and QuantiTect SYBR Green RT-PCR mix (Qiagen, Valencia, CA). mRNA levels were calculated and normalized to mRNA levels of the housekeeping gene *LRP10* using Bio-Rad CFX Manager 3.0 software (Bio-Rad). Primer sequences are listed in ESM Table 1.

ESM Methods 4: Reaction Scheme implemented in Taverna, a workflow management system.

Genes within two reaction steps of acetyl coA are identified in the KEGG pathway database. These form the gene set for acetyl coA metabolism. The method is shown in the workflow below and further discussed in Dharuri et al[7].


Acetyl coA
Get reactions that acetyl coA participates in (R1)
Get compounds that participate in the reaction set R1 (C2)
FILTER COMPOUNDS
Get reactions that members of C2 participate in (R2)
Get enzymes that drive the reactions in R2 (E1)
Get genes corresponding to these enzymes in E1
GENE SET

ESM Table 1: Sequences of primers used in the qPCR

Gene	Fw primer	Rv primer	Annealing Temperature
LRP10	CAGACTGTCACCATCAGGTTC	GAGAGGGGAGCGTAGGGTTA	60
ACACA	TTTAAGGGGTGAAGAGGGTGC	CCAGAAAGACCTAGCCCTCAAG	56
ACAT1	CAATTGGGATGTCTGGAGC	TAGCATGGCAGAAGCACCTC	58
ALDH6A1	GTGCTTCTGGGCAGTAGAG	TCACCTTGGAAGAAACCTGC	58
MTHFD1	AGGTGTCCCTACAGGCTTCA	GCATTGTGCTCATCGTTCCT	61

ESM Table 2: Genes significantly up or down-regulated in VAT of T2DM subjects

gene_id	Coefficient VAT NGT vs T2DM	p- value VAT NGT vs T2DM	p-value adj VAT NGTvs T2DM	Coefficient SAT NGTvs T2DM	p-value SAT NGT vs T2DM	p-value adj SAT NGTvs T2DM
ALDH6A1	-0.67033	1.49E-06	0.005502	-0.59541	5.12E-05	0.042281
C14orf45	-0.46222	1.59E-06	0.005502	-0.39723	0.000107	0.042281
ECHS1	-0.52103	1.48E-06	0.005502	-0.31088	0.003448	0.078286
IRS1	-0.6011	3.41E-07	0.005502	-0.64306	7.26E-06	0.027696
STBD1	-0.61468	6.74E-07	0.005502	-0.32893	0.005018	0.089566
IARS2	-0.31061	2.73E-06	0.006958	-0.18008	0.001517	0.060769
NAT8L	-0.74468	2.81E-06	0.006958	-0.44867	0.001498	0.060769
AIFM2	-0.45243	3.24E-06	0.007013	-0.24872	0.002298	0.070283
ATPAF1	-0.34906	3.71E-06	0.007141	-0.29532	0.003426	0.078286
ACAD9	-0.31106	8.28E-06	0.010501	-0.26532	0.001228	0.0603
GPI	-0.28503	8.25E-06	0.010501	-0.16923	0.008524	0.10884
HADH	-0.57479	8.49E-06	0.010501	-0.45138	0.000145	0.042281
HSPD1	-0.29857	7.74E-06	0.010501	-0.16624	0.029418	0.183414
MTHFD1	-0.42341	6.16E-06	0.010501	-0.3727	7.95E-05	0.042281
ACACA	-0.56009	9.14E-06	0.010554	-0.47629	0.000297	0.0487
MAP3K15	-0.43265	1.19E-05	0.012882	-0.39465	9.36E-05	0.042281
HK2	-0.71165	1.32E-05	0.01298	-0.38422	0.005958	0.094515
PARVG	0.654257	1.5E-05	0.01298	0.744737	0.000608	0.055969
PDHA1	-0.37534	1.48E-05	0.01298	-0.28852	0.001045	0.057525
PRKAR2B	-0.71637	1.39E-05	0.01298	-0.37036	0.02795	0.179313
ACAT1	-0.4062	1.81E-05	0.012994	-0.27406	0.002336	0.070807
ATP9A	-0.40037	2.1E-05	0.012994	-0.34628	0.005481	0.091691
CEBPA	-0.5664	1.97E-05	0.012994	-0.37387	0.002787	0.075329

DARS2	-0.37947	1.64E-05	0.012994	-0.30113	0.000532	0.055836
NXPH4	-1.0023	1.89E-05	0.012994	-0.72237	0.039951	0.213164
OXCT1	-0.5273	1.99E-05	0.012994	-0.41823	0.000444	0.053704
SLC2A4	-0.88429	2.09E-05	0.012994	-0.73171	0.000105	0.042281
TMEM120B	-0.39756	2.09E-05	0.012994	-0.40465	0.001103	0.058403
HIBADH	-0.34092	2.41E-05	0.014395	-0.2844	0.00054	0.055836
MME	-0.87028	2.53E-05	0.014586	-0.24391	0.077446	0.290709
ATP5B	-0.29223	2.73E-05	0.014753	-0.14885	0.01547	0.136867
CST7	0.804504	2.9E-05	0.014753	0.57403	0.009737	0.114841
GPT2	-0.59523	2.68E-05	0.014753	-0.4677	0.000991	0.057525
UQCRC2	-0.3105	2.85E-05	0.014753	-0.18736	0.001774	0.064807
PHYH	-0.4099	3.1E-05	0.015179	-0.25588	0.002172	0.068621
SORBS1	-0.52573	3.16E-05	0.015179	-0.3601	0.002971	0.076138
FXYD5	0.462931	3.38E-05	0.015638	0.508216	3E-05	0.041435
ME1	-0.44801	3.52E-05	0.015638	-0.45845	9.99E-05	0.042281
SDHC	-0.32374	3.52E-05	0.015638	-0.20976	0.009145	0.111578
LOC401052	-0.5606	3.92E-05	0.016969	-0.46015	0.003091	0.076138
ABHD14A-	-0.31203	4.04E-05	0.017066	-0.12479	0.067336	0.274133
CS	-0.40868	4.37E-05	0.017595	-0.25358	0.003252	0.07663
FASN	-0.86173	4.3E-05	0.017595	-0.709	0.000487	0.055836
PECR	-0.52655	4.55E-05	0.017913	-0.39495	0.002181	0.068769
LOC80054	-0.5741	4.98E-05	0.018747	-0.36113	0.006328	0.096929
NEK9	-0.24342	4.92E-05	0.018747	-0.25243	0.000163	0.043349
ADCY6	-0.39224	5.76E-05	0.018983	-0.34851	0.001307	0.060323
CDO1	-0.60039	5.76E-05	0.018983	-0.34148	0.013689	0.128261
DMGDH	-0.51115	5.65E-05	0.018983	-0.39788	0.001753	0.064445
GCOM1	-0.42785	5.8E-05	0.018983	-0.2221	0.020122	0.156278

KCNN4	0.791172	5.35E-05	0.018983	0.719036	0.004726	0.087804
LETMD1	-0.31904	5.81E-05	0.018983	-0.21855	0.002345	0.070845
PEX19	-0.34994	5.62E-05	0.018983	-0.30875	0.000527	0.055836
MLXIPL	-0.47102	6.33E-05	0.019469	-0.39568	0.00155	0.060769
MUT	-0.36496	6.25E-05	0.019469	-0.29517	0.000441	0.053704
NDUFS1	-0.33767	6.41E-05	0.019469	-0.26825	0.005722	0.092817
PC	-0.51421	6.25E-05	0.019469	-0.33894	0.000302	0.0487
ATP5A1	-0.26726	6.8E-05	0.020287	-0.15425	0.012799	0.125416
FNTA	-0.23067	7.3E-05	0.021412	-0.15997	0.000901	0.057525
GYG2	-0.49552	7.54E-05	0.021771	-0.37986	0.000567	0.055836
C12orf35	0.482851	8.21E-05	0.022143	0.211704	0.110692	0.347015
LDHD	-0.7308	7.85E-05	0.022143	-0.53027	9.59E-05	0.042281
MCCC1	-0.44783	8.31E-05	0.022143	-0.32769	0.002073	0.067862
MYOM1	-0.74875	7.95E-05	0.022143	-0.57005	0.001222	0.0603
SLC25A33	-0.55683	8.31E-05	0.022143	-0.21109	0.106244	0.340839
CD3D	1.057606	9.32E-05	0.024452	0.689572	0.015962	0.139
ITGA7	-0.43821	9.49E-05	0.024532	-0.29145	0.000602	0.055969
HADHB	-0.28151	9.67E-05	0.024621	-0.12942	0.071027	0.280078
FAH	-0.36888	0.000102	0.025337	-0.21549	0.015303	0.136221
PGM1	-0.43899	0.000102	0.025337	-0.32841	0.006897	0.100804
FAN1	-0.24678	0.000104	0.025465	-0.26323	5.17E-05	0.042281
MYZAP	-0.49368	0.00011	0.026527	-0.31503	0.002982	0.076138
MGEA5	-0.34359	0.000114	0.026968	-0.30296	0.001829	0.065041
ALDH2	-0.47219	0.000126	0.027353	-0.40667	0.001386	0.060323
FBXO27	-0.47275	0.00012	0.027353	-0.32289	0.006789	0.100256
KCNIP2	-0.57813	0.000126	0.027353	-0.33357	0.026098	0.174084
PDHX	-0.3518	0.000117	0.027353	-0.25083	0.005747	0.092817
PFKFB1	-0.61156	0.000124	0.027353	-0.57186	0.00039	0.051411
SLC6A6	0.448635	0.000126	0.027353	0.388987	0.020459	0.157549

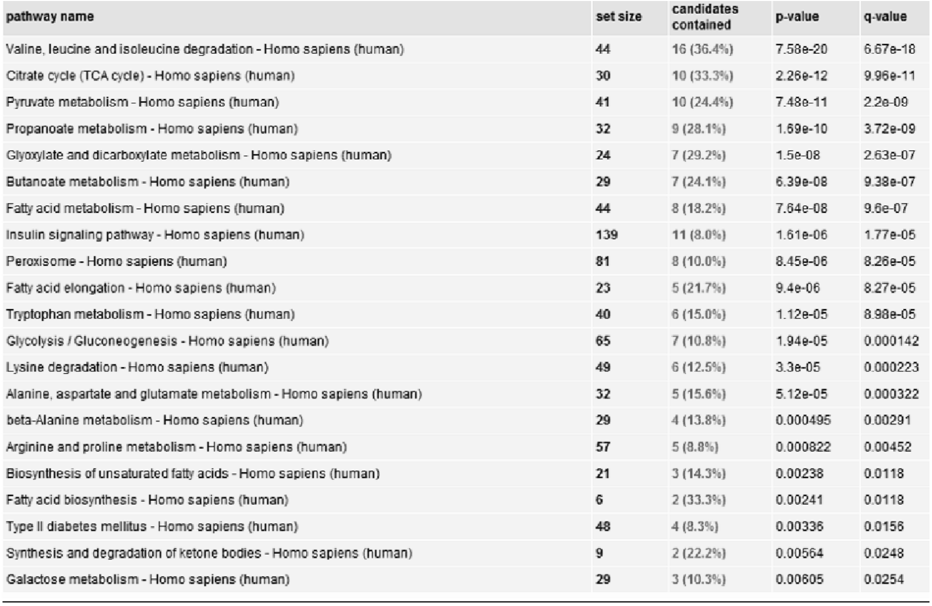
TM7SF2	-0.65447	0.000121	0.027353	-0.5652	0.000245	0.048507
DLD	-0.36203	0.00013	0.027768	-0.22936	0.002916	0.075807
PEX11A	-0.54283	0.000134	0.028208	-0.28022	0.029285	0.18285
ALDH5A1	-0.41815	0.000143	0.028227	-0.37903	0.002663	0.074601
BNIP3	-0.46674	0.000136	0.028227	-0.26732	0.005206	0.090496
CPS1	-0.592	0.000147	0.028227	-0.45495	0.002242	0.069444
FH	-0.25037	0.00014	0.028227	-0.08348	0.209234	0.476835
GABRE	-0.41347	0.00014	0.028227	-0.44716	0.00082	0.056752
LETM1	-0.28644	0.000144	0.028227	-0.15596	0.025489	0.17185
PCCB	-0.39374	0.000144	0.028227	-0.19472	0.022988	0.165424
PHF13	-0.39433	0.000147	0.028227	0.050692	0.625383	0.816692
ARG2	-0.29061	0.000149	0.028265	-0.14506	0.020686	0.158062
BTBD6	-0.40376	0.00015	0.028313	-0.22028	0.015406	0.136581
CA8	-0.72509	0.000153	0.028427	-0.52113	0.011228	0.120415
LST1	0.597583	0.000156	0.028427	0.73995	0.000566	0.055836
YWHAG	-0.34916	0.000156	0.028427	-0.1706	0.053963	0.247066
APCDD1	-0.43512	0.000159	0.028617	-0.20187	0.148309	0.399287
ETFDH	-0.36711	0.000161	0.02868	-0.20541	0.008033	0.106776
MMD	-0.62533	0.000165	0.029125	-0.41409	0.003817	0.082012
AMICA1	0.708154	0.00017	0.029539	0.760944	0.00161	0.061115
OGDH	-0.21233	0.000171	0.029539	-0.12606	0.024143	0.168181
ACO1	-0.38915	0.000173	0.029635	-0.27544	0.000833	0.056752
ACOT1	-0.53314	0.000176	0.029652	-0.34982	0.007398	0.103802
COQ6	-0.30529	0.000178	0.029652	-0.23084	0.001309	0.060323
TUSC5	-0.39626	0.000178	0.029652	-0.31553	0.01315	0.126493
AIFM1	-0.25849	0.000182	0.02994	-0.12482	0.055699	0.250486
AKAP1	-0.45989	0.000183	0.02994	-0.22612	0.069932	0.278155
CAT	-0.33619	0.000193	0.031147	-0.25952	0.001106	0.058403
DLST	-0.24018	0.000195	0.031147	-0.08557	0.070547	0.278997

PLIN5	-0.77331	0.000196	0.031147	-0.6064	0.004128	0.084327
GLUD2	-0.2456	0.000205	0.032341	-0.11297	0.038981	0.211289
IMMT	-0.19286	0.000209	0.032599	-0.13107	0.006316	0.096929
PCCA	-0.38327	0.000211	0.032658	-0.30718	0.004828	0.088137
ACADS	-0.47336	0.000218	0.033182	-0.22351	0.012826	0.125416
CHST11	0.474785	0.000218	0.033182	0.587619	0.001251	0.060323
RETSAT	-0.45639	0.000221	0.033265	-0.28288	0.016783	0.142511
PDE3B	-0.53384	0.000231	0.034471	-0.31326	0.031749	0.189882
AKR1B10	4.423196	0.000239	0.034715	2.274977	0.094247	0.321015
BCL2L13	-0.2233	0.000243	0.034715	-0.09604	0.143732	0.393576
CHCHD10	-0.43065	0.000236	0.034715	-0.07441	0.439305	0.683267
DIS3L	-0.22639	0.00024	0.034715	-0.16482	0.026102	0.174084
ORMDL3	-0.3648	0.000241	0.034715	-0.35641	0.00098	0.057525
ADORA3	0.787044	0.000248	0.03485	0.849999	0.005258	0.090592
BOK	-0.5296	0.000248	0.03485	-0.42262	0.00056	0.055836
CORO1A	0.588905	0.000253	0.03485	0.57774	0.001825	0.065041
EPB41L4B	-0.5678	0.000254	0.03485	-0.39557	0.005131	0.0902
NDN	-0.29789	0.000253	0.03485	-0.14583	0.056384	0.251827
ACADM	-0.45711	0.000264	0.035079	-0.37292	0.002267	0.069893
ACADSB	-0.33375	0.000265	0.035079	-0.3195	0.000556	0.055836
ADSSL1	-0.75955	0.000259	0.035079	-0.67785	0.000229	0.048025
HSDL2	-0.33571	0.000262	0.035079	-0.24675	0.006545	0.098355
PHKA2	-0.30828	0.000259	0.035079	-0.35695	0.000548	0.055836
KLRK1	0.713466	0.000269	0.035309	0.429783	0.072961	0.283239
STRADB	-0.36152	0.000283	0.036577	-0.23321	0.032869	0.193406
TYRO3	-0.31518	0.000282	0.036577	-0.30272	0.003522	0.078997
CYB5A	-0.32699	0.000289	0.036589	-0.343	0.001823	0.065041
KIAA0368	-0.18082	0.00029	0.036589	-0.14045	0.020676	0.158062
PXMP2	-0.44048	0.000289	0.036589	-0.46718	1.65E-05	0.03571

ACOT2	-0.5115	0.000294	0.036657	-0.26945	0.03003	0.184965
C1orf162	0.679527	0.000292	0.036657	0.756212	0.000302	0.0487
UQCC	-0.26188	0.000297	0.036689	-0.25136	0.000325	0.048769
CD8B	1.149373	0.000323	0.039612	0.889473	0.031652	0.18956
HCAR1	-0.49796	0.000332	0.040514	-0.5247	0.023562	0.16643
HSD17B4	-0.19447	0.000345	0.041461	-0.12918	0.04775	0.231962
ZNF436	-0.33341	0.000343	0.041461	-0.01462	0.852844	0.93309
ADHFE1	-0.52698	0.000359	0.041829	-0.42806	0.0025	0.072607
MRPL30	-0.23547	0.000358	0.041829	-0.14517	0.014305	0.131674
PHLDB2	-0.37206	0.000354	0.041829	-0.29726	0.016369	0.140724
RRM1	-0.22802	0.000354	0.041829	-0.15169	0.044719	0.225816
TNS1	-0.42061	0.00036	0.041829	-0.35763	0.005129	0.0902
CENPK	1.086756	0.000365	0.042038	0.528342	0.076188	0.289028
KANK1	-0.39156	0.000367	0.042038	-0.29404	0.007223	0.102463

Coefficient VAT NGT vs T2DM: log fold change of NGT vs T2DM in visceral adipose tissue; a negative value reflects down-regulation whereas a positive value reflects up-regulation of the gene in T2DM subjects. Adj p-value NGT vs T2DM: p-value after Benjamin-Hochberg FDR correction. Also the log fold change and p-values for subcutaneous tissue (SAT) are shown. The genes highlighted in grey are the 42 genes that are members of the acetyl-CoA network. All these 42 genes are significantly down-regulated in VAT of T2DM.

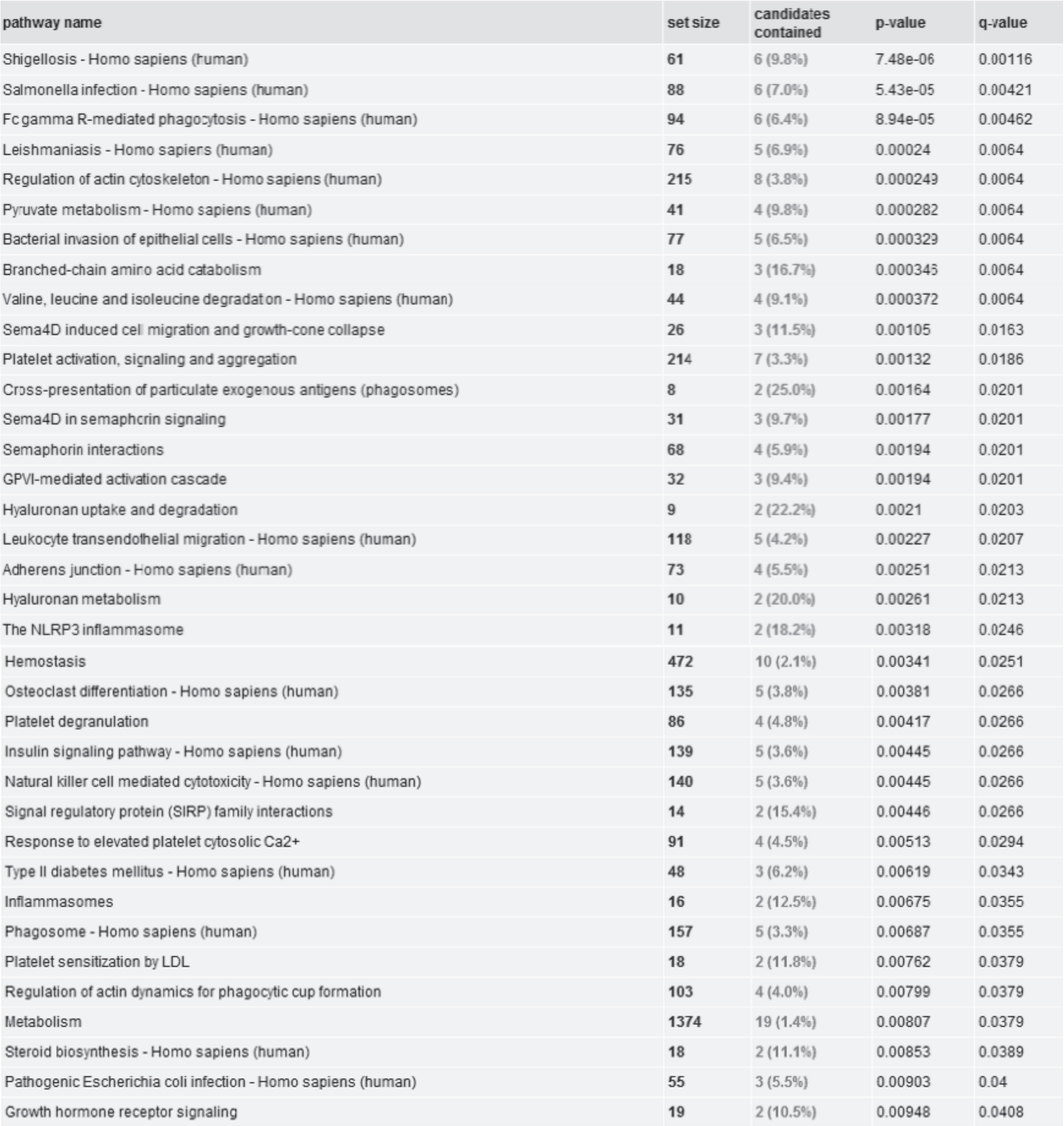
ESM Table 3: KEGG Pathway over-representation analysis among significantly differentially expressed genes in the VAT



pathway name	set size	candidates contained	p-value	q-value
Valine, leucine and isoleucine degradation - Homo sapiens (human)	44	16 (36.4%)	7.58e-20	6.67e-18
Citrate cycle (TCA cycle) - Homo sapiens (human)	30	10 (33.3%)	2.26e-12	9.96e-11
Pyruvate metabolism - Homo sapiens (human)	41	10 (24.4%)	7.48e-11	2.2e-09
Propanoate metabolism - Homo sapiens (human)	32	9 (28.1%)	1.69e-10	3.72e-09
Glyoxylate and dicarboxylate metabolism - Homo sapiens (human)	24	7 (29.2%)	1.5e-08	2.63e-07
Butanoate metabolism - Homo sapiens (human)	29	7 (24.1%)	6.39e-08	9.38e-07
Fatty acid metabolism - Homo sapiens (human)	44	8 (18.2%)	7.64e-08	9.6e-07
Insulin signaling pathway - Homo sapiens (human)	139	11 (8.0%)	1.61e-06	1.77e-05
Peroxisome - Homo sapiens (human)	81	8 (10.0%)	8.45e-06	8.26e-05
Fatty acid elongation - Homo sapiens (human)	23	5 (21.7%)	9.4e-06	8.27e-05
Tryptophan metabolism - Homo sapiens (human)	40	6 (15.0%)	1.12e-05	8.98e-05
Glycolysis / Gluconeogenesis - Homo sapiens (human)	65	7 (10.8%)	1.94e-05	0.000142
Lysine degradation - Homo sapiens (human)	49	6 (12.5%)	3.3e-05	0.000223
Alanine, aspartate and glutamate metabolism - Homo sapiens (human)	32	5 (15.6%)	5.12e-05	0.000322
beta-Alanine metabolism - Homo sapiens (human)	29	4 (13.8%)	0.000495	0.00291
Arginine and proline metabolism - Homo sapiens (human)	57	5 (8.8%)	0.000822	0.00452
Biosynthesis of unsaturated fatty acids - Homo sapiens (human)	21	3 (14.3%)	0.00238	0.0118
Fatty acid biosynthesis - Homo sapiens (human)	6	2 (33.3%)	0.00241	0.0118
Type II diabetes mellitus - Homo sapiens (human)	48	4 (8.3%)	0.00336	0.0156
Synthesis and degradation of ketone bodies - Homo sapiens (human)	9	2 (22.2%)	0.00564	0.0248
Galactose metabolism - Homo sapiens (human)	29	3 (10.3%)	0.00605	0.0254

Significantly differentially expressed genes in the VAT were mapped onto the KEGG pathway for over- representation analysis using the software tool made available by ConsensusPathDB. ‘Pathway names’ contains the names of the significant pathways, ‘set size’ is the number of genes in the pathway, ‘candidates contained’ is the number of genes in the input that are members of the pathway. The p- value is calculated according to the hypergeometric test based on the number of genes present in both the predefined set and list of significant genes from VAT provided as input. The p-values are corrected for multiple testing using false discovery rate and are shown as q-values above. The results provided in the table above are for a q-value cut-off of < 0.05.

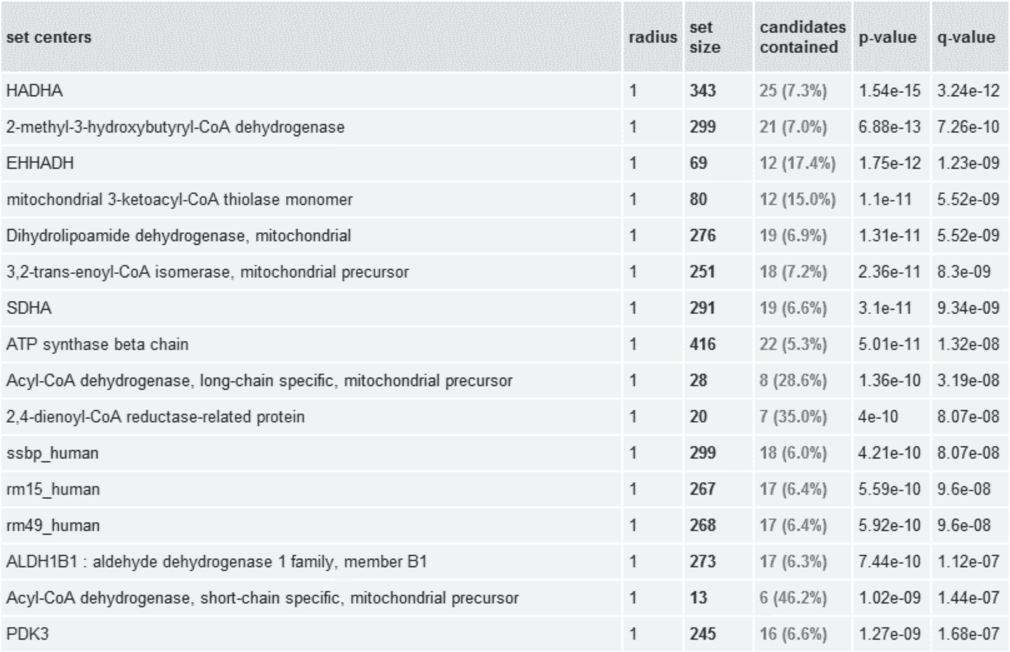
ESM Table 4: KEGG Pathway over-representation analysis among significantly differentially expressed genes in the SAT



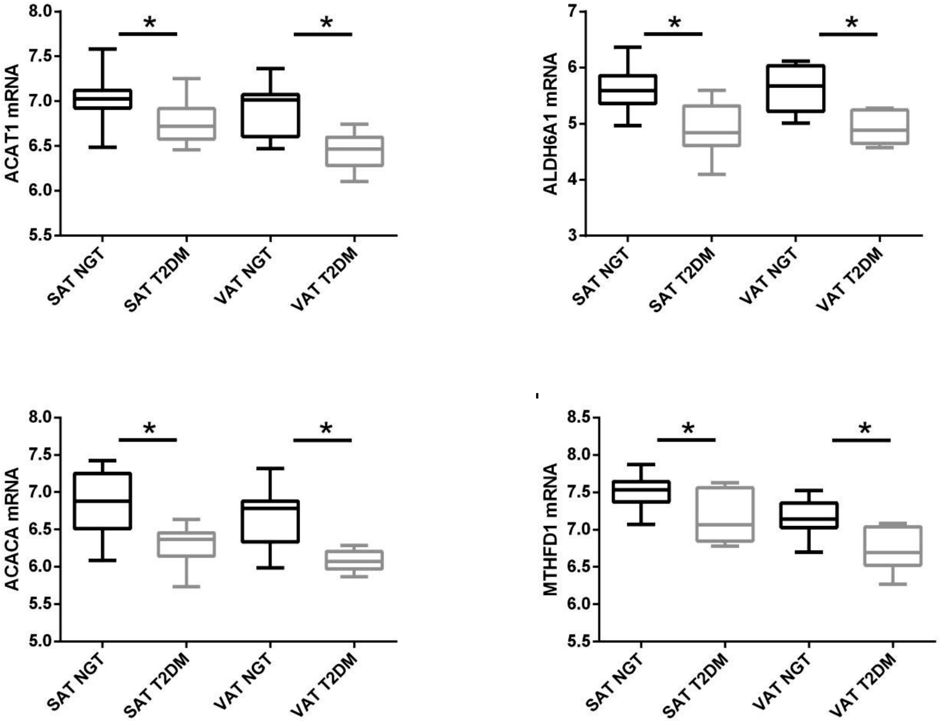
pathway name	set size	candidates contained	p-value	q-value
Shigellosis - Homo sapiens (human)	**61**	6 (9.8%)	7.48e-06	0.00116
Salmonella infection - Homo sapiens (human)	**88**	6 (7.0%)	5.43e-05	0.00421
Fc gamma R-mediated phagocytosis - Homo sapiens (human)	**94**	6 (6.4%)	8.94e-05	0.00462
Leishmaniasis - Homo sapiens (human)	**76**	5 (6.9%)	0.00024	0.0064
Regulation of actin cytoskeleton - Homo sapiens (human)	**215**	8 (3.8%)	0.000249	0.0064
Pyruvate metabolism - Homo sapiens (human)	**41**	4 (9.8%)	0.000282	0.0064
Bacterial invasion of epithelial cells - Homo sapiens (human)	**77**	5 (6.5%)	0.000329	0.0064
Branched-chain amino acid catabolism	**18**	3 (16.7%)	0.000345	0.0064
Valine, leucine and isoleucine degradation - Homo sapiens (human)	**44**	4 (9.1%)	0.000372	0.0064
Sema4D induced cell migration and growth-cone collapse	**26**	3 (11.5%)	0.00105	0.0163
Platelet activation, signaling and aggregation	**214**	7 (3.3%)	0.00132	0.0186
Cross-presentation of particulate exogenous antigens (phagosomes)	**8**	2 (25.0%)	0.00164	0.0201
Sema4D in semaphorin signaling	**31**	3 (9.7%)	0.00177	0.0201
Semaphorin interactions	**68**	4 (5.9%)	0.00194	0.0201
GPVI-mediated activation cascade	**32**	3 (9.4%)	0.00194	0.0201
Hyaluronan uptake and degradation	**9**	2 (22.2%)	0.0021	0.0203
Leukocyte transendothelial migration - Homo sapiens (human)	**118**	5 (4.2%)	0.00227	0.0207
Adherens junction - Homo sapiens (human)	**73**	4 (5.5%)	0.00251	0.0213
Hyaluronan metabolism	**10**	2 (20.0%)	0.00261	0.0213
The NLRP3 inflammasome	**11**	2 (18.2%)	0.00318	0.0246
Hemostasis	**472**	10 (2.1%)	0.00341	0.0251
Osteoclast differentiation - Homo sapiens (human)	**135**	5 (3.8%)	0.00381	0.0266
Platelet degranulation	**86**	4 (4.8%)	0.00417	0.0266
Insulin signaling pathway - Homo sapiens (human)	**139**	5 (3.6%)	0.00445	0.0266
Natural killer cell mediated cytotoxicity - Homo sapiens (human)	**140**	5 (3.6%)	0.00445	0.0266
Signal regulatory protein (SIRP) family interactions	**14**	2 (15.4%)	0.00446	0.0266
Response to elevated platelet cytosolic Ca2+	**91**	4 (4.5%)	0.00513	0.0294
Type II diabetes mellitus - Homo sapiens (human)	**48**	3 (6.2%)	0.00619	0.0343
Inflammasomes	**16**	2 (12.5%)	0.00675	0.0355
Phagosome - Homo sapiens (human)	**157**	5 (3.3%)	0.00687	0.0355
Platelet sensitization by LDL	**18**	2 (11.8%)	0.00762	0.0379
Regulation of actin dynamics for phagocytic cup formation	**103**	4 (4.0%)	0.00799	0.0379
Metabolism	**1374**	19 (1.4%)	0.00807	0.0379
Steroid biosynthesis - Homo sapiens (human)	**18**	2 (11.1%)	0.00853	0.0389
Pathogenic Escherichia coli infection - Homo sapiens (human)	**55**	3 (5.5%)	0.00903	0.04
Growth hormone receptor signaling	**19**	2 (10.5%)	0.00948	0.0408

Significantly differentially expressed genes in the SAT were mapped onto the KEGG pathway for over- representation analysis using the software tool made available by ConsensusPathDB. ‘Pathway names’ contains the names of the significant pathways, ‘set size’ is the number of genes in the pathway, ‘candidates contained’ is the number of genes in the input that are members of the pathway. The p-value is calculated according to the hypergeometric test based on the number of genes present in both the predefined set and list of significant genes from SAT provided as input. The p-values are corrected for multiple testing using false discovery rate and are shown as q-values above. The results provided in the table above are for a q-value cut-off of < 0.05.

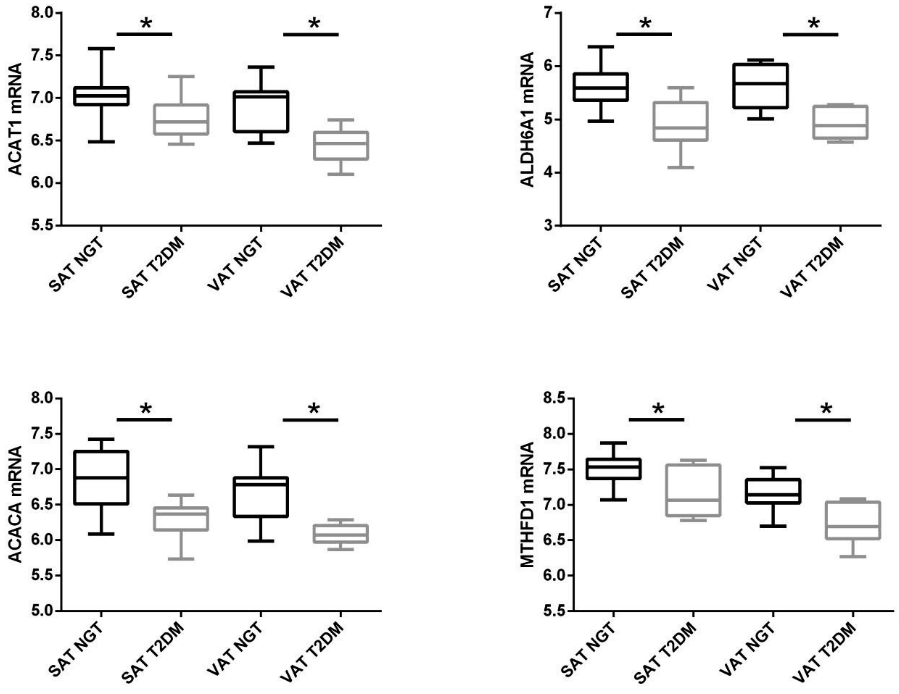
ESM Table 5: Top 15 Enriched Network-based Sets (NESTs) for an input of top hits from the visceral adipose tissue differentially expressed between diabetic and healthy subjects.



set centers	radius	set size	candidates contained	p-value	q-value
HADHA	1	**343**	25 (7.3%)	1.54e-15	3.24e-12
2-methyl-3-hydroxybutyryl-CoA dehydrogenase	1	**299**	21 (7.0%)	6.88e-13	7.26e-10
EHHADH	1	**69**	12 (17.4%)	1.75e-12	1.23e-09
mitochondrial 3-ketoacyl-CoA thiolase monomer	1	**80**	12 (15.0%)	1.1e-11	5.52e-09
Dihydrolipoamide dehydrogenase, mitochondrial	1	**276**	19 (6.9%)	1.31e-11	5.52e-09
3,2-trans-enoyl-CoA isomerase, mitochondrial precursor	1	**251**	18 (7.2%)	2.36e-11	8.3e-09
SDHA	1	**291**	19 (6.6%)	3.1e-11	9.34e-09
ATP synthase beta chain	1	**416**	22 (5.3%)	5.01e-11	1.32e-08
Acyl-CoA dehydrogenase, long-chain specific, mitochondrial precursor	1	**28**	8 (28.6%)	1.36e-10	3.19e-08
2,4-dienoyl-CoA reductase-related protein	1	**20**	7 (35.0%)	4e-10	8.07e-08
ssbp_human	1	**299**	18 (6.0%)	4.21e-10	8.07e-08
rm15_human	1	**267**	17 (6.4%)	5.59e-10	9.6e-08
rm49_human	1	**268**	17 (6.4%)	5.92e-10	9.6e-08
ALDH1B1 : aldehyde dehydrogenase 1 family, member B1	1	**273**	17 (6.3%)	7.44e-10	1.12e-07
Acyl-CoA dehydrogenase, short-chain specific, mitochondrial precursor	1	**13**	6 (46.2%)	1.02e-09	1.44e-07
PDK3	1	**245**	16 (6.6%)	1.27e-09	1.68e-07



ESM Fig. 1a-d: No difference in adipose tissue expression of a) ACAT1, b) ACACA, c) ALDH6A1 or d) MTHFD1 between T2DM subjects that use metformin and those that do not use metformin. Boxplots of normalized gene expression profiles (log2-scale) are shown. (SAT = subcutaneous adipose tissue, VAT = visceral adipose tissue)



ESM Fig. 2a-d: The comparison of adipose tissue gene expression of ACAT1, ACACA, ALDH6A1 and MTHFD1 between NGT and T2DM subjects when all metformin users are excluded. The acetyl coA genes are lower in the T2DM subjects. Boxplots of normalized gene expression profiles (log2-scale) are shown. (SAT = subcutaneous adipose tissue, VAT = visceral adipose tissue)

References (Supplemental section)

1 Wu TD, Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads**. *Bioinformatics* 2010, **26**(7):873-881.

2. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

3. Dale RK, Pedersen BS, Quinlan AR: **Pybedtools: a flexible Python library for manipulating genomic datasets and annotations**. *Bioinformatics* 2011, **27**(24):3423-3424.

4. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.

5. Law CW, Chen Y, Shi W, Smyth GK: **Voom: precision weights unlock linear model analysis tools for RNA-seq read counts.** *Genome Biol* 2014, **15**:R29.

6. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P *et al*: **The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud**. *Nucleic acids research* 2013, **41**(Web Server issue):W557-561.

7. Dharuri H, Henneman P, Demirkan A, van Klinken JB, Mook-Kanamori DO, Wang-Sattler R, Gieger C, Adamski J, Hettne K, Roos M *et al*: **Automated workflow-based exploitation of pathway databases provides new insights into genetic associations of metabolite profiles**. *BMC genomics* 2013, **14**(1):865.

8. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic acids research* 2010, **38**(Database issue):D355-360.

Chapter 5: Differential allele specific expression of *KLRK1* in subcutaneous and visceral adipose tissue of very obese individuals with and without type 2 diabetes mellitus

Harish Dharuri
Irina Pulyakhina
Vanessa van Harmelen
Ko Willems van Dijk
Peter A.C. ’t Hoen

In preparation

Abstract

Background: Obesity is associated with reduced life expectancy due to increased rates of cardiovascular diseases and type-2 diabetes. Studies have shown that visceral adipose tissue plays a more critical role than the subcutaneous adipose tissue in the development of insulin resistance and the metabolic syndrome. This has been attributed to functional differences between the two tissues. However, the molecular basis for these intra-depot differences and inter-individual differences in the functioning of the two tissues is mostly lacking. Next generation RNA-sequencing technology has made it possible to quantify gene expression but also to call haplotypes of an individual based on heterozygosity of expressed loci. So called allele-specific expression studies help to understand the cis-regulatory basis of variation in gene expression. Here, we investigate the hypothesis that cis-regulatory variants differentially affect gene expression in visceral and subcutaneous adipose tissue. To this end, we investigated differential allele-specific expression between visceral and subcutaneous adipose tissue of very obese individuals (BMI>40) with and without type 2 diabetes mellitus with the aim of identifying regulatory variants that could explain the pathophysiological differences observed in the two tissues.

Results: A protocol to identify “high-confidence” heterozygous sites yielded a total of 1115 SNPs in the RNA-sequencing data obtained from the two tissues. A quasi-binomial test at each heterozygous site identified a polymorphism, rs1049174 in the *KLRK1* gene, in the Natural Killer complex region, with a significant differential allele-specific expression between visceral and subcutaneous adipose tissue. The allelic imbalance for KLRK1 was highest in subcutaneous adipose tissue. Individuals homozygous for the alternative allele demonstrated lower expression than heterozygous individuals. Interestingly, the expression of KLRK1 was higher in the visceral adipose tissues of very obese individuals with type 2 diabetes mellitus, in particular in heterozygous individuals, when compared to NGT individuals.

Conclusion: The differential allele-specific expression of *KLRK1* (NKGD2) between visceral and subcutaneous adipose tissue and the increased expression of KLRK1 in visceral adipose tissue of very obese individuals with type 2 diabetes provides evidence for a role of *KLRK1* (NKGD2) in the susceptibility to type 2 diabetes.

Introduction

Obesity has reached epidemic proportions in modern societies [1] and results in reduced life expectancy due to associated metabolic and cardiovascular disorders, as well as several types of cancer [2, 3]. The central role of the expanded adipose tissue in obesity-related complications like type-2 diabetes and coronary artery disease has been well documented [4]. In addition, there is clear evidence for functional differences between the visceral adipose tissue (VAT) and subcutaneous adipose tissue (SAT) [5, 6] in its involvement in the disease. Studies have shown that the VAT secretes more and a larger variety of pro-inflammatory cytokines called adipocytokines than SAT. These adipocytokines induce insulin resistance and endothelial dysfunction. This is thought to explain that, in comparison to SAT, the VAT has a stronger association with obesity related complications like type 2 diabetes (T2D) and cardiovascular disorders. In a previous study, we investigated gene expression differences between VAT and SAT that could help explain the differences between normal glucose tolerant (NGT) individuals and individuals with type-2 diabetes in a study cohort of very obese individuals (BMI > 40 kg/m^2) [7]. We analyzed RNA-sequencing data obtained from the VAT and SAT obtaining during bariatric surgery. This study confirmed large differences between VAT and SAT gene expression profiles, identified a distinct diabetes signature in VAT, with larger aberrations in the expression of metabolic genes. However, the molecular origin of the differences in function between VAT and SAT are still mostly unknown.

In the present study, we investigated to what extent genetic variation plays a role in functional differences between VAT and SAT. RNA-sequencing offers a unique opportunity to study this. In individuals heterozygous for an expressed SNP, there are paired observations of the expression of both alleles. The uneven expression of the two allelic copies of a transcript is commonly known as allelic imbalance (AI) or allele specific expression (ASE) and can be identified from the read counts for the two alleles. ASE likely represents a difference in the genetic regulation of gene expression at the locus, and may help to explain genetic associations between the locus and the phenotype. The analysis of ASE allows for the analysis of the genetic component of gene expression in much smaller numbers of individuals than in traditional expression quantitative trait loci (eQTL) studies, where the genetic variation is usually only a minor contributor to the total degree of variation in gene expression between individuals [8]. However, it should be realized that AI may not be purely genetic, but also caused by epigenetic factors [9, 10].

Recent studies have shown that genetic variants associated with disease susceptibility may regulate gene expression in a tissue-dependent manner [11, 12], and that differences in regulation between tissues are reflected by differences in ASE [13, 14]. These differences may be a result of the differential expression of tissue-specific transcription or other regulatory factors. Here, we studied the differential ASE between VAT and SAT to find clues on the differences in genetic regulation of gene expression between the two tissues, and their relation to the type 2 diabetes (T2D) phenotype. By focusing on the differences in ASE between the tissues, we were not affected by the reference bias that may result in false positive identifications of ASE events, because the reference bias would be similar for the two tissues. In order to reduce the search space, we zoomed in on a panel of SNPs downloaded from the Genome Wide Association Studies (GWAS) catalog [15], aiming for a further functional characterization of these GWAS hits. Therefore, our objective is to identify from a panel of known genome-wide association hits the subset of common variants that are under the control of cis-regulatory elements and to assess the consequence of such variants on the T2D phenotype. We identified a single nucleotide polymorphism (SNP) rs1049174, in the 3’ untranslated region (3’ UTR) in *KLRK1* (Killer cell lectin like receptor subfamily K, family member 1) gene that displays a significant differential allelic expression between VAT and SAT, and for which expression is different between individuals with normal glucose tolerance (NGT) and T2D.

Material and Methods

Subjects

The subjects (all with BMI >40), isolation and characterization of subcutaneous adipose tissue (SAT) and visceral adipose tissue (VAT), isolation of adipose tissue RNA and deep sequencing have been extensively described previously [7, 16]. In short, the study group consisted of 17 obese women with normal glucose tolerance and 15 obese women with type 2 diabetes classified according to WHO standards based on fasting glucose levels. The groups were matched for age, BMI and waist circumference. All individuals underwent bariatric surgery (gastric bypass or banding), during which procedure a piece of VAT and SAT were obtained. RNA was isolated using a standard kit.

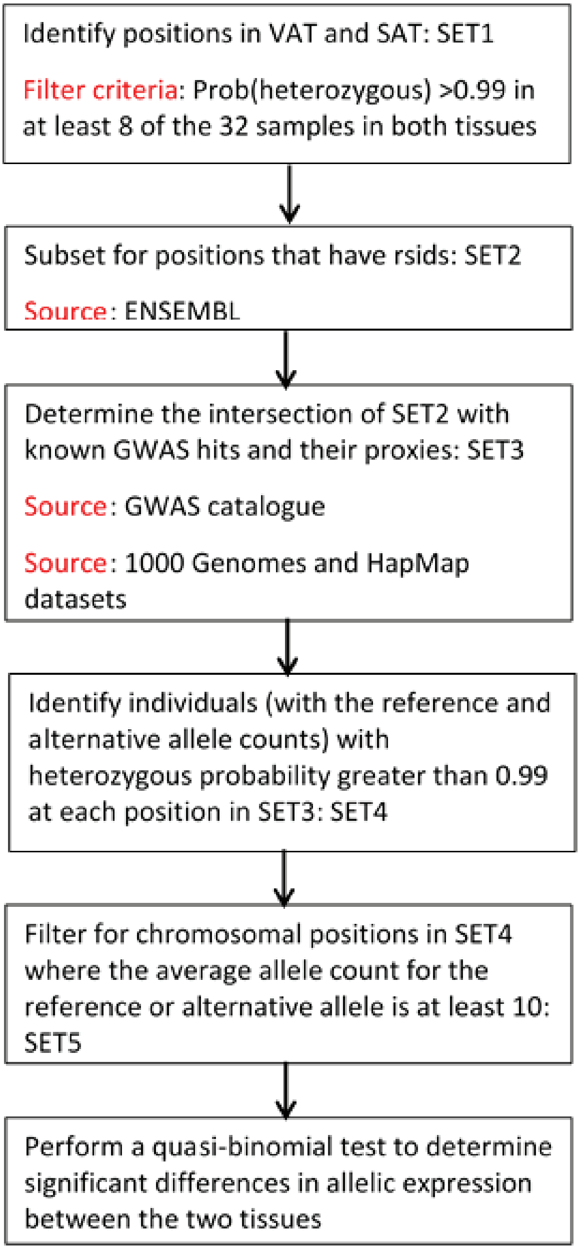

Identify positions in VAT and SAT: SET1
Filter criteria: Prob(heterozygous) >0.99 in at least 8 of the 32 samples in both tissues
Subset for positions that have rsids: SET2
Source: ENSEMBL
Determine the intersection of SET2 with known GWAS hits and their proxies: SET3
Source: GWAS catalogue
Source: 1000 Genomes and HapMap datasets
Identify individuals (with the reference and alternative allele counts) with heterozygous probability greater than 0.99 at each position in SET3: SET4
Filter for chromosomal positions in SET4 where the average allele count for the reference or alternative allele is at least 10: SET5
Perform a quasi-binomial test to determine significant differences in allelic expression between the two tissues

Figure 1: Strategy for identification of relevant SNPs and determination of allele-specific expression in VAT and SAT.

Deep sequencing and bioinformatics analysis of RNA

RNA Deep Sequencing was performed on mRNA enriched using oligo(dT) beads at the Beijing Genomics Institute (BGI, Beijing, China). The sequencing was done using IlluminaHiSeq 2000 with 90-nucleotide long Paired-end reads, resulting in a minimum of 3GB clean data per sample. The reads were aligned to the Human reference genome build 19 (hg19) to obtain a histogram of coverage per exon and the associated count data. Further details related to alignment and gene annotation can be obtained in our previous publication [7]. SNPs were called in each sample individually using SNVMix2 version 0.12.2-rc1 [SNVMix] with default settings. SNPs called in both VAT and SAT samples from the same subject were extracted (around 5 million SNPs per subject).

SNP selection criteria

The protocol for selection of heterozygous sites in the VAT and SAT is shown in Figure 1. In the first step, chromosomal positions were selected whose confidence score for a heterozygous genotype estimated by SNVMix2 was determined to be greater than 99% in at least 8 of the 32 individuals in each of SAT and VAT. Next, only positions that had designated “rsids” as determined by the ENSEMBL database (www.ensembl.org) served as further filtration criteria. The genome-wide association study (GWAS) catalogue was downloaded and the proxies for the SNPs were determined using Broad Institutes “SNAP” web portal (http://www.broadinstitute.org/mpg/snap/). The criteria to select proxies were the following: $r^2>0.8$ in population panel of Caucasian and European origin (CEU). SNP datasets from 1000 Genomes Pilot 1 as well as HapMap were used to identify relevant proxies to SNPs from the GWAS catalogue. The “rsids” within our study cohort as described above were matched with the GWAS catalogue and its proxies and the intersection of these sets was taken up for investigating differential ASE in the VAT and SAT. For each of these positions, individual level reference and alternate allele counts were extracted. Therefore, for each position determined by the filter criteria described above, we created a table of reference and alternate allele counts in the two tissues for individuals that had passed the criteria for the selection of the SNP.

Statistical analysis

To determine differential ASE between the VAT and SAT, a quasi-binomial test was performed at each heterozygous site identified by the SNP selection protocol. In a quasi-binomial test, an additional overdispersion parameter φ

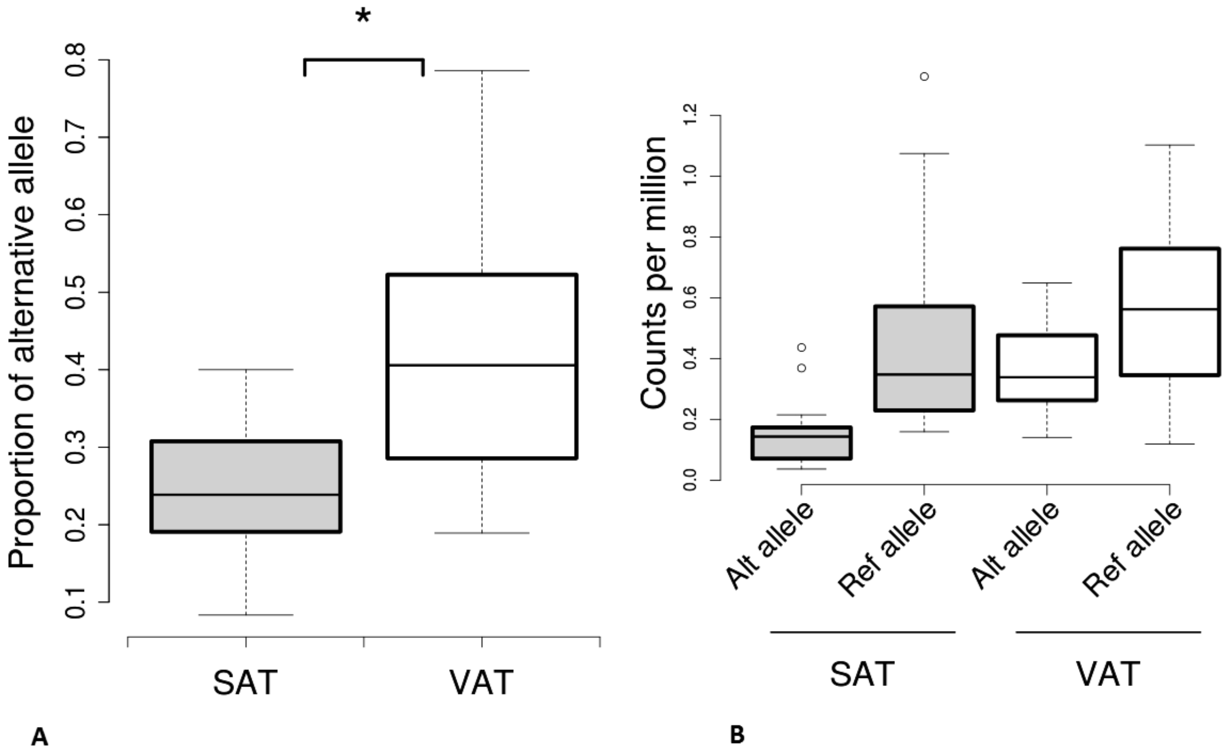


Figure 2 Proportion of the alternative allele (C) count in the SAT and VAT (A). Reference (G) and alternative (C) allele counts in the SAT and VAT for the SNP rs1049174 (B).

is incorporated in the variance for binomial distribution: variance = φnp(1-p) where n and p are defined in this study as the total number of counts and probability of identifying an alternate allele respectively. The additional parameter accounts for overdispersion caused due to biological or measurement variation between subjects. This test is a post hoc adjustment to the variance of a binomial model which generally results in inflated standard errors and consequently a more conservative test result. The quasi-binomial model was implemented using the glm function in the R programming language and specifying family=quasibinomial. The protocol for selecting heterozygous sites yielded a total of 1115 SNPs in the RNA-Seq data. Broad institute's web portal, SNPsnap [17] was used to determine the number of independent SNPs in this set. For the 648 independent tests determined by SNPsnap, multiple testing threshold for statistical significance was established using Bonferroni correction.

RESULTS

Significant allelic imbalance between the two tissues observed at rs1049174 chr12:10525365)

To identify common SNPs displaying a difference in allelic expression between SAT and VAT, only SNPs were evaluated associated with relevant traits in GWAS studies or in close LD (r2 > 0.8) with these SNPs. 1115 SNPs were identified in the RNA-Seq data that were heterozygous in at least 8 individuals with at least 10 times coverage. This set consisted of 648 independent SNPs, as determined by SNPsnap. The multiple testing threshold using a Bonferroni correction at a nominal p-value of 0.05 for 648 independent SNPs is 0.05/648 = 7.71e-05. A quasi-binomial test used to determine differential allelic expression between the two tissues identified one significant hit, rs1049174 (chr12:10525365), with a p-value of 4.21e-05. As shown in Figure 2, the proportion of the alternative allele (C) of rs1049174 is significantly lower in the SAT.

The alternative allele of rs1049174 is associated with lower expression of KLRK1 in SAT

The SNP, rs1049174 is in the 3' UTR region of KLRK1 Killer cell lectin like receptor subfamily K, family member 1 (KLRK1) gene that codes for the NKG2D protein that is a receptor on natural killer and other inflammatory cells. To determine the effect of the genotype on the expression of the KLRK1 gene, we evaluated the expression of the gene for individuals who were homozygous and heterozygous for the alternative allele as our initial

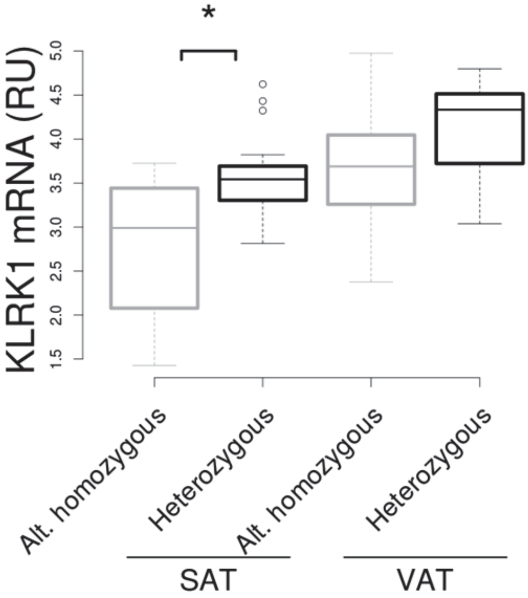


Figure 3 Gene expression profile of the *KLRK1* in the SAT and VAT for homozygous alternative allele and heterozygous individuals at chr12:10525365 (rs1049174). Boxplot of normalized gene expression profile (relative units [RU]: $\log_2$-scale) of *KLRK1* for obese heterozygous (black) and homozygous alternative (gray) individuals (* adjusted p value < 0.05 for indicated comparison).

evaluation did not include individuals who were homozygous for the reference allele. As shown in Fig 3, the alternative allele is associated with lower expression of KLRK1. This difference is statistically significant in the SAT (p-value: 0.009), and a similar trend is observed in the VAT.

KLRK1 is significantly differentially expressed in VAT between T2DM and NGT subjects

We have previously reported gene expression analysis of RNA-Seq data from the transcriptome extracted from SAT and VAT of these subjects [7]. Gene-level analysis with the limma package in R had identified 168 genes differentially expressed in VAT (p-value < 0.05 after Benjamin-Hochberg FDR correction) between obese individuals with NGT and those with type 2 diabetes. The same method on SAT yielded 121 genes that were differentially

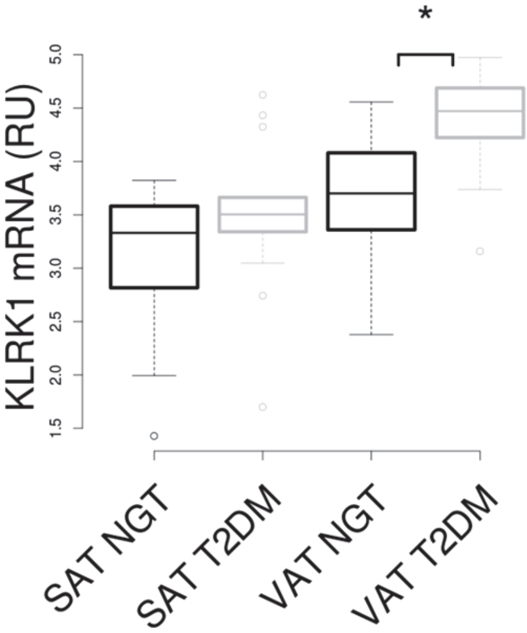


Figure 4 Gene expression profile of the *KLRK1* in the SAT and VAT. Boxplot of normalized gene expression profile (relative units [RU]: $\log_2$-scale) of *KLRK1* for obese NGT (black) and T2DM (gray) individuals (* adjusted p value < 0.05 for indicated comparison).

expressed between NGT and type 2 diabetes subjects. KLRK1, with an adjusted p-value of 0.03, is among the 168 genes differentially expressed in the VAT based on health status and is significantly up-regulated among type 2 diabetes subjects (Figure 4). While it is not among the top hits in the SAT, its expression profile as shown in Fig 4 is indicative of an association between gene expression and diabetes status.

Individuals heterozygous for rs1049174 are significantly enriched for T2DM

The SNP rs1049174 is in Hardy-Weinberg equilibrium in the Dutch population [18] with an expected frequency for heterozygous individuals is 0.408. In our study, among the 15 T2DM subjects, 11 are heterozygous and 4 are alternative homozygous individuals (2pq=0.465); a binomial test was performed to test a null hypothesis of finding 11 heterozygous individuals, by chance out of a total of 15 when the expected frequency is 0.408. This yielded a p-value of 0.011, which suggests that heterozygous individuals are significantly enriched among T2DM subjects. The frequency of homozygous and heterozygous in 17 NGT subjects was not different 2pq=0.39) from the expected frequency.

Discussion

The visceral adipose tissue is metabolically and functionally distinct from the subcutaneous adipose tissue and these differences are thought to play a role in obesity related complications like type 2 diabetes. Previous efforts have focused on differential gene expression profile in order to elucidate mechanisms that serve as the basis for functional differences seen in the two tissues. In this study, we examined differential allele-specific expression with the aim of identifying genetic variants that regulate gene expression and contribute to the differences in VAT and SAT. To decrease the multiple testing burden and to zoom in on clinically relevant loci, we restricted the search space to known GWAS hits and their proxies and detected significant allelic imbalance at chr12:10525365.

Chr12:10525365 (rs1049174) is in the 3' UTR region of (Killer cell lectin like receptor subfamily K, family member 1 *(KLRK1*) gene that codes for the NKG2D protein, a receptor on the natural killer cells, CD8$^+$ αβ T cells, γδ T cells, and activated macrophages [19, 20]. In addition to significant differential ASE, we observed that the alternate allele (C) at this locus is associated with lower expression of the *KLRK1* gene. Interestingly, independent studies have pointed to the genetic control exerted by this locus. For example, Veyrieras JB et al [21] have reported that the polymorphism, rs1049174 is an eQTL for *KLRK1*. Additionally, in the BBMRI-BIOS study we verified that the polymorphism is an eQTL for the read-through transcript, KLRC4-KLRK1 that codes for the *KLRK1* gene (BIOS consortium, manuscript in preparation). Furthermore, a case-control study [22] to identify genetic factors associated with natural cytotoxic activity reported two NKG2D haplotypes: high natural killer (*HNK1*) activity and low natural killer (*LNK1*) activity that could be assessed based on 5 SNPs in tight linkage disequilibrium in the noncoding region of the gene. Interestingly, rs1049174 (G/C) is one of the SNPs; the genotype CC is associated with *LNK1* and GG with *HNK1*. Furthermore, investigations in this study eliminated the possibility of “as-yet undiscovered” SNPs in the coding region, thus implying that the SNPs may be involved in the transcriptional regulation of the *KLRK1* gene and affect its expression levels.

Results from this and earlier studies confirm the role of SNP rs1049174 in immune function; from immunosurveillance in cancer [23] to improving immunity and outcomes among subjects receiving bone marrow transplantation [24]. Furthermore, rs1049174 was included in the SNP set based on its proximity (r^2>0.8) to the GWAS hit, rs2617170 a SNP that has been associated with the Behcet's syndrome. The latter is a chronic disorder involving inflammation of the blood vessels throughout the body. Taken together, these results point to the control of the variant on gene expression and consequently on immune/inflammation response. The latter is of particular interest to us given its implications in obesity related complications like type 2 diabetes. The expression of *KLRK1* was significantly correlated with the expression of *CD3*, *CD8* and *CD14,* immune cell markers for T-cell, Cytotoxic T-cells and macrophages. This suggests that the expression of *KLRK1* is linked to the activity of multiple immune cells. The higher expression in VAT compared to SAT reflects the higher inflammatory activity in VAT among type 2 diabetes subjects.

In our previous study we had observed that approximately 8,000 genes were differentially expressed between VAT and SAT. *KLRK1* was one of the genes that were significantly higher expressed in VAT than in SAT. The fact that we found several thousand genes differentially expressed between the two tissues and only one gene displaying differential ASE bears some explanation. First, our search space was restricted to known GWAS hits. Furthermore, we employed strict filter criteria for selection of SNPs. Second, the regulation of the expression for the vast majority of genes may be largely similar for both alleles.

We observed a statistically significant enrichment of heterozygous individuals among T2DM subjects and no such association between the alternate homozygous and disease status. However, replication in an independent study [25] did not point to enrichment of heterozygous individuals among obese T2DM individuals. However, due to the lack of statistical power in our study as well as the replication cohort further investigation is required to understand the influence of the genotype at this locus on obesity related complications. While we could not replicate the role of this polymorphism in T2DM, this study identified rs1049174 as a *cis*-eQTL for yet another set of genes (*KLRC1*, *KLRC2*, *KLRC3*) in the NK gene complex region on chromosome 12. It is likely that the complexity of the genomic region and the uncertainty in the assignment of RNA-seq reads or microarray probes to the different (read-through) transcripts originating from this locus contributes to the lack of congruence between studies.

In conclusion, our investigation into allele-specific expression between visceral and subcutaneous adipose tissue points to a variant in the NK gene complex region of chromosome 12. This study provides evidence for a role of *KLRK1* (NKGD2) in the susceptibility to type-2 diabetes among very obese individuals.

References

1. Malik VS, Willett WC, Hu FB: **Global obesity: trends, risk factors and policy implications.** *Nat Rev Endocrinol* 2013, **9**:13–27.

2. Mokdad AH, Ford ES, Bowman BA, Dietz WH, Vinicor F, Bales VS, Marks JS: **Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001.** *JAMA* 2003, **289**:76–9.

3. Van Gaal LF, Mertens IL, De Block CE: **Mechanisms linking obesity with cardiovascular disease**. *Nature* 2006, **444**:875–880.

4. Blüher M: **Adipose tissue dysfunction contributes to obesity related metabolic diseases.** *Best Pract Res Clin Endocrinol Metab* 2013, **27**:163–77.

5. Fox CS, Massaro JM, Hoffmann U, Pou KM, Maurovich-Horvat P, Liu C-Y, Vasan RS, Murabito JM, Meigs JB, Cupples LA, D’Agostino RB, O’Donnell CJ: **Abdominal visceral and subcutaneous adipose tissue compartments: association with metabolic risk factors in the Framingham Heart Study.** *Circulation* 2007, **116**:39–48.

6. Hamdy O, Porramatikul S, Al-Ozairi E: **Metabolic obesity: the paradox between visceral and subcutaneous fat.** *Curr Diabetes Rev* 2006, **2**:367–73.

7. Dharuri H, ’t Hoen PAC, van Klinken JB, Henneman P, Laros JFJ, Lips MA, el Bouazzaoui F, van Ommen G-JB, Janssen I, van Ramshorst B, van Wagensveld BA, Pijl H, Willems van Dijk K, van Harmelen V: **Downregulation of the acetyl-CoA metabolic network in adipose tissue of obese diabetic individuals and recovery after weight loss**. *Diabetologia* 2014, **57**:2384–2392.

8. Hasin-Brumshtein Y, Hormozdiari F, Martin L, van Nas A, Eskin E, Lusis AJ, Drake T a: **Allele-specific expression and eQTL analysis in mouse adipose tissue.** *BMC Genomics* 2014, **15**:471.

9. Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, et al.: **Transcriptome and genome sequencing uncovers functional variation in humans.** *Nature* 2013, **501**:506–11.

10. Lagarrigue S, Martin L, Hormozdiari F, Roux PF, Pan C, van Nas A, Demeure O, Cantor R, Ghazalpour A, Eskin E, Lusis AJ: **Analysis of allele-specific expression in mouse liver by RNA-seq: A comparison with Cis-eQTL identified using genetic linkage**. *Genetics* 2013, **195**:1157–1166.

11. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, Dermitzakis ET: **Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations**. *PLoS Genet* 2010, **6**.

12. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K, Hedman ÅK, Bataille V, Bell J, Surdulescu G, Dimas AS, Ingle C, Nestle FO, Meglio P, Min JL, Wilk A, Hammond CJ, Hassanali N, Yang TP, Montgomery SB, O’Rahilly S, Lindgren CM, Zondervan KT, Soranzo N, Barroso I, Durbin R, et al.: **The architecture of gene regulatory variation across multiple human tissues: The muTHER study**. *PLoS Genet* 2011, **7**.

13. GTEx Consortium: **Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.** *Science* 2015, **348**:648–660.

14. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, et al.: **The Genotype-Tissue Expression (GTEx) project.** *Nat Genet* 2013, **45**:580–5.

15. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H: **The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.** *Nucleic Acids Res* 2014, **42**(Database issue):D1001–6.

16. Lips MA, Van Klinken JB, van Harmelen V, Dharuri HK, ’t Hoen PAC, Laros JFJ, van Ommen G-J, Janssen IM, Van Ramshorst B, Van Wagensveld BA,

Swank DJ, Van Dielen F, Dane A, Harms A, Vreeken R, Hankemeier T, Smit JWA, Pijl H, Willems van Dijk K: **Roux-en-Y gastric bypass surgery, but not calorie restriction, reduces plasma branched-chain amino acids in obese women independent of weight loss or the presence of type 2 diabetes.** *Diabetes Care* 2014, **37**:3150–6.

17. Pers TH, Timshel P, Hirschhorn JN: **SNPsnap: a Web-based tool for identification and annotation of matched SNPs.** *Bioinformatics* 2015, **31**:418–20.
18. Collection S, Genome T: **Whole-genome sequence variation, population structure and demographic history of the Dutch population.** *Nat Genet* 2014:1–95.
19. Bauer S, Groh V, Wu J, Steinle A, Phillips JH, Lanier LL, Spies T: **Activation of NK cells and T cells by NKG2D, a receptor for stress-inducible MICA.** *Science* 1999, **285**:727–729.
20. Wu J, Song Y, Bakker AB, Bauer S, Spies T, Lanier LL, Phillips JH: **An activating immunoreceptor complex formed by NKG2D and DAP10.** *Science* 1999, **285**:730–732.
21. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK: **High-resolution mapping of expression-QTLs yields insight into human gene regulation**. *PLoS Genet* 2008, **4**.
22. Hayashi T, Imai K, Morishita Y, Hayashi I, Kusunoki Y, Nakachi K: **Identification of the NKG2D haplotypes associated with natural cytotoxic activity of peripheral blood lymphocytes and cancer immunosurveillance**. *Cancer Res* 2006, **66**:563–570.
23. Furue H, Matsuo K, Kumimoto H, Hiraki A, Suzuki T, Yatabe Y, Komori K, Kanemitsu Y, Hirai T, Kato T, Ueda M, Ishizaki K, Tajima K: **Decreased risk of colorectal cancer with the high natural killer cell activity NKG2D genotype in Japanese**. *Carcinogenesis* 2008, **29**:316–320.
24. Espinoza JL, Takami A, Onizuka M, Sao H, Akiyama H, Miyamura K, Okamoto S, Inoue M, Kanda Y, Ohtake S, Fukuda T, Morishima Y, Kodera Y, Nakao S: **NKG2D gene polymorphism has a significant impact on transplant outcomes after HLA-fully-matched unrelated bone marrow transplantation for standard risk hematologic malignancies**. *Haematologica* 2009, **94**:1427–1434.
25. Wolfs MGM, Rensen SS, Bruin-Van Dijk EJ, Verdam FJ, Greve J-W, Sanjabi B, Bruinenberg M, Wijmenga C, van Haeften TW, Buurman WA, Franke L,

Hofker MH: **Co-expressed immune and metabolic genes in visceral and subcutaneous adipose tissue from severely obese individuals are associated with plasma HDL and glucose levels: a microarray study.** *BMC Med Genomics* 2010, **3**:34.

Chapter 6: Proteomic Analysis in Type 2 Diabetes Patients before and after a Very Low Calorie Diet Reveals Potential Disease State and Intervention Specific Biomarkers

Maria A. Sleddering*
Albert J. Markvoort*
Harish K. Dharuri
Skhandhan Jeyakar
Marieke Snel
Peter Juhasz
Moira Lynch
Wade Hines
Xiaohong Li
Ingrid M. Jazet
Aram Adourian
Peter A. J. Hilbers
Johannes W. A. Smit
Ko Willems Van Dijk

*Authors contributed equally

PLoS One. 2014; 9(11):e112835

Abstract

Very low calorie diets (VLCD) with and without exercise programs lead to major metabolic improvements in obese type 2 diabetes patients. The mechanisms underlying these improvements have so far not been elucidated fully. To further investigate the mechanisms of a VLCD with or without exercise and to uncover possible biomarkers associated with these interventions, blood samples were collected from 27 obese type 2 diabetes patients before and after a 16-week VLCD (Modifast ,450 kcal/day). Thirteen of these patients followed an exercise program in addition to the VCLD. Plasma was obtained from 27 lean and 27 obese controls as well. Proteomic analysis was performed using mass spectrometry (MS) and targeted multiple reaction monitoring (MRM) and a large scale isobaric tags for relative and absolute quantitation (iTRAQ) approach. After the 16-week VLCD, there was a significant decrease in body weight and HbA1c in all patients, without differences between the two intervention groups. Targeted MRM analysis revealed differences in several proteins, which could be divided in diabetes-associated (fibrinogen, transthyretin), obesity-associated (complement C3), and diet- associated markers (apolipoproteins, especially apolipoprotein A-IV). To further investigate the effects of exercise, large scale iTRAQ analysis was performed.

However, no proteins were found showing an exercise effect. Thus, in this study, specific proteins were found to be differentially expressed in type 2 diabetes patients versus controls and before and after a VLCD. These proteins are potential disease state and intervention specific biomarkers.

Trial Registration: Controlled-Trials.com ISRCTN76920690

Introduction

The incidence of insulin resistant states, such as the metabolic syndrome and type 2 diabetes (T2DM), has increased dramatically in recent years [1, 2]. T2DM is a chronic multifactorial disease characterized by insulin resistance of the liver, skeletal muscle and adipose tissue and the progressive failure of pancreatic b-cells [3, 4]. Furthermore, research has shown that T2DM is associated with inflammation, oxidative stress and vascular dysfunction [3, 5].

Over 80% of T2DM patients is overweight or obese [6, 7], nevertheless T2DM develops in only about one-third of obese, insulin-resistant individuals.

Simultaneously, some 30% of obese (BMI.30) individuals seem metabolically healthy. Whether these patients are protected from, or merely have a delayed risk for developing T2DM is not known [8, 9]. Because of the contribution of obesity to insulin resistance, it is essential for obese T2DM patients to reduce body weight. The most fundamental aspect of the treatment of obesity is life-style change, i.e. reduction of caloric intake and increase of physical activity. Very low calorie diets (VLCD) have been shown to lead to a substantial amount of weight loss and subsequently result in major metabolic improvements in obese T2DM patients [10]. Recently, we have shown that a 16-week VLCD in T2DM patients leads to a decrease in pericardial fat volume and an increase in quality of life (QoL) [11, 12]. In addition, adding an exercise program to the VLCD in these patients has been shown to have moderate additional favorable effects [13].

In the past decade large scale proteome analysis, also referred to as 'proteomics', has been used to identify new biomarkers for the risk prediction of various diseases, such as cancer, Alzheimer's disease, cardiovascular disease and diabetes. Proteomics can also be used to further elucidate disease mechanism and molecular processes and to investigate the response of the body to interventions [14, 15]. In diabetes research, proteomics have been analysed in various bodily fluids, cell-lines and tissues, such as blood, urine, saliva, semen, vitreous fluid, b- cells, adipocytes, hepatocytes and skeletal muscle [16–22]. However, most of the proteomics studies are cross-sectional and there are currently no studies on proteomic analysis in obese T2DM patients before and after a diet, the hallmark of their treatment.

To gain more insight into the pathophysiology of type 2 diabetes we performed plasma proteomics on the obese T2DM patients, before and after a VLCD with or without exercise, for which clinical and metabolic improvements after the VLCD were published before. [11–13, 23–25] Furthermore, we compared these T2DM patients before and after the diet with obese and lean controls. Because of the drastic weight loss and major improvements in glycemic control after such a diet, we hypothesized that differences in proteins can be found that might be involved in the development of, and recovery from, T2DM. By comparing the patients to controls, we aim to uncover proteins differentially expressed in T2DM patients as compared to lean and obese controls, and changes in these differences after the intervention. In addition, by comparing the groups with and without exercise, we aim to uncover possible biomarkers associated with the additional favorable effects of adding an exercise

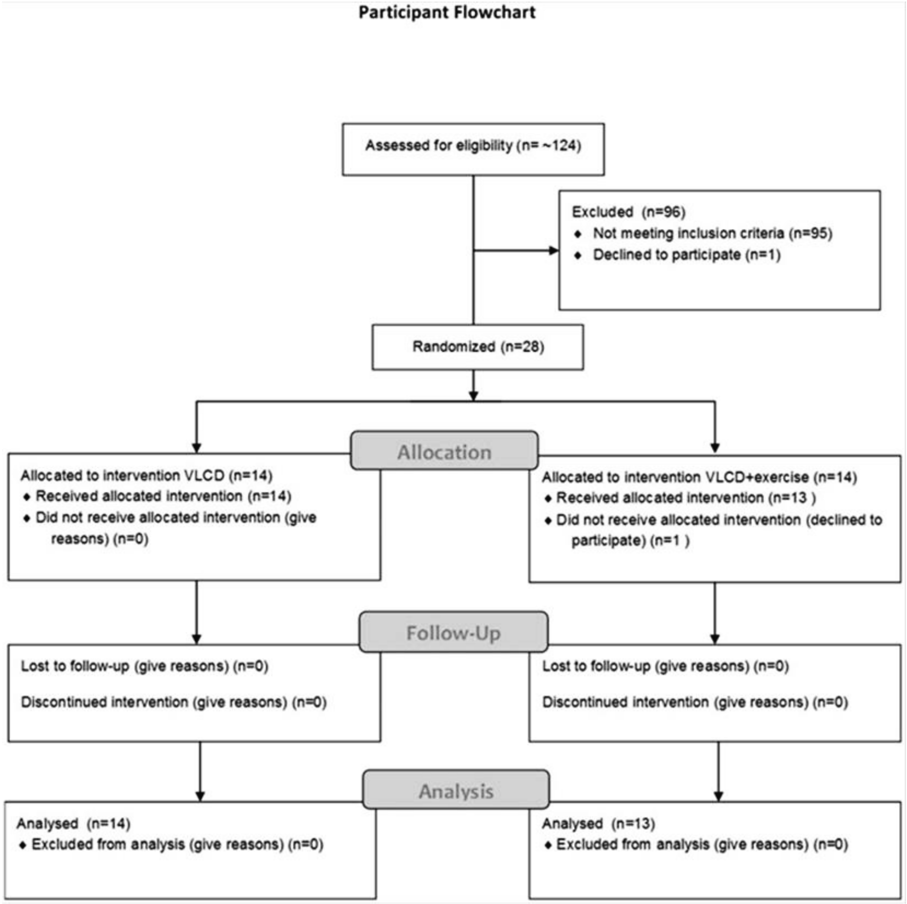

Participant Flowchart
Assessed for eligibility (n= ~124)
Excluded (n=96)
Not meeting inclusion criteria (n=95)
Declined to participate (n=1)
Randomized (n=28)
Allocation
Allocated to intervention VLCD (n=14)
Received allocated intervention (n=14)
Did not receive allocated intervention (give reasons) (n=0)
Allocated to intervention VLCD+exercise (n=14)
Received allocated intervention (n=13)
Did not receive allocated intervention (declined to participate) (n=1)
Follow-Up
Lost to follow-up (give reasons) (n=0)
Discontinued intervention (give reasons) (n=0)
Lost to follow-up (give reasons) (n=0)
Discontinued intervention (give reasons) (n=0)
Analysis
Analysed (n=14)
Excluded from analysis (give reasons) (n=0)
Analysed (n=13)
Excluded from analysis (give reasons) (n=0)

Figure 1 Participant flowchart

program to the VLCD. Firstly, we conducted a targeted MRM analysis of 13 abundant proteins hypothesized to be associated with T2DM and obesity, including apolipoproteins and markers of inflammation and coagulation. Subsequently, we performed a large scale iTRAQ analysis in samples of the T2DM patients before and after the diet to uncover differences between the VLCD with and without exercise groups also for less abundant proteins.

Materials and Methods

Patients

The protocol of this study has been described previously [13]. In short, twenty- seven (14 men, 13 women) T2DM patients were included in the

Table 1. Clinical characteristics of type 2 diabetes patients before and after a 16-week VLCD +/- exercise and obese and lean controls.(These data were previously published REF VLCD, VLCD QoL, VLCD inflamm).

	VLCD + excercise				**VLCD only**				**controls**		
	baseline		16 weeks		baseline		16 weeks		obese		lean
Sex (M/F)	8/5				6/8				14/13		14/13
Age (years)	53 ± 2				56 ± 2				55 ± 2		56 ± 2
Weight (kg)	113.5 ± 5.1	*	86.3 ± 4.2	#*†	112.7 ± 5.6	*	89.0 ± 4.3	#*†	117 ± 4	*	73 ± 2
BMI (kg/m²)	36.4 ± 1.1	*	27.7 ± 1	#*†	37.9 ± 1.4	*	30.0 ± 1.1	#*†	38.8 ± 1.2	*	24.2 ± 0.4
Waist (cm)	123 ± 3	*	98 ± 3	#*†	122 ± 3	*	103 ± 3	#*†	120 ± 2	*	88 ± 2
Fat mass (kg)	45.4 ± 3.2	*	23.5 ± 2.2	#*†$	49.9 ± 3.6	*	33.2 ± 2.8	#†	43.0 ± 3.4		35.3 ± 3.1
Systolic BP (mmHg)	145 ± 5	$	132 ± 5	#†	160 ± 4	*	140 ± 4	#†	153 ± 4	*	139 ± 4
Diastolic BP (mmHg)	81 ± 3	†	75 ± 2	†	87 ± 3		78 ± 2	#†	89 ± 2	*	81 ± 2
Glucose (mmol/L)	10.9 ± 0.7	*†	6.6 ± 0.8	#*	12.1 ± 0.5	*†	7.7 ± 0.6	#*†	5.3 ± 0.2	*	4.8 ± 0.1
Insulin (mU/L)	25 ± 2.2	*†	9 ± 0.8	*†#	24 ± 4.3	*†	13 ± 2	#*	13 ± 1.1	*	5 ± 0.7
HbA1c (%)	7.8 ± 0.4	*†	6.3 ± 0.4	*#	7.8 ± 0.3	*†	6.7 ± 0.3	#*†	5.5 ± 0.1	*	5.2 ± 0.04
Total cholesterol (mmol/L)	5.4 ± 0.4		4.5 ± 0.3	#*†$	6.1 ± 0.4		5.5 ± 0.3		6.2 ± 0.2		6.2 ± 0.2
Triglycerides (mmol/L)	2.5 ± 0.5	*	1.2 ± 0.1	#†	2.3 ± 0.2	*†	1.5 ± 0.2	#*	1.7 ± 0.2	*	1.1 ± 0.1
LDL cholesterol (mmol/L)	3.6 ± 0.3		3.0 ± 0.2	#*†	4.4 ± 0.4		3.7 ± 0.3		4.0 ± 0.2		3.8 ± 0.2

HDL cholesterol (mmol/L)	1.1 ± 0.0 *†	1.2 ± 0.1 *†	1.2 ± 0.1 *†	1.2 ± 0.1 *	1.4 ± 0.1 *	1.8 ± 0.1
Diabetes duration (years)	7.9 ± 1.2		9.8 ± 1.1		N/A	N/A
Insulin dose (units/day)	77	0	86	0	N/A	N/A
Metformin (n)	10	0	9	0	N/A	N/A
SU-derivatives (n)	3	0	1	0	N/A	N/A

Mean ± SEM, unless otherwise specified. # Significant difference within group vs. baseline; * significant difference vs. lean controls; † significant difference vs. obese controls; $ significant difference VLCD only vs. VLCD+exercise. BMI: body mass index; BP: blood pressure; HDL: high density lipoprotein; LDL: low density lipoprotein.

study (Figure 1). Diabetes duration was 8.9±0.8 years (mean ±SEM) and patients were obese with an average BMI of 37.2±0.9 kg/m2. All patients were on insulin therapy (average insulin dose 82±11 units/day) with or without additional oral glucose-lowering medication. Smoking, recent weight change (past 3 months), a history of cardiovascular disease or any other chronic disease were reasons for exclusion.

Two control subjects were recruited via advertisements for every T2DM patient, one lean and one obese subject. Control subjects were matched for gender, age, race and geographical area. In addition, obese control subjects were matched for BMI as well. Clinical characteristics are shown in Table 1.

Ethics statement

This study was conducted in accordance with the Declaration of Helsinki. The study protocol was approved by the local ethics committee (Commissie Medische Ethiek, Leiden University Medical Center) and written informed consent was obtained from all subjects. The study was registered under ISRCTN76920690 (http://www.controlled-trials.com/isrctn/). The study was conducted between 2006 and 2009. The proteomics analysis was performed in 2010–2011. The proteomic analysis was not planned when the study was approved by the ethics committee, but was added later. The proteomics protocol is described in detail below.

Study design

All T2DM patients followed a VLCD for a period of 16 weeks. We randomly assigned 13 of the 27 patients to simultaneously follow an exercise program. All patients were provided with the same instruction forms and were all willing to be randomized to either intervention. We then assigned the first 13 fit candidates to the VLCD with exercise intervention. The following fit candidates were assigned to the VLCD-only intervention. The patients were not aware of the randomization order. Patients were studied before and after the VLCD intervention. Oral glucose-lowering medication was discontinued three weeks before the start of the study and insulin therapy was stopped the day before. During the 16-week intervention period, all glucose-lowering medication, including insulin, remained discontinued.

VLCD

The VLCD consisted of three sachets of Modifast (Nutrition & Sante´, Antwerp, Belgium), containing a total of 450 kcal per day. It provides about 50 to 60 grams of carbohydrate, 50 grams of protein, 7 to 9 grams of lipid, 10 grams of dietary fibers and all necessary vitamins and micronutrients. During the whole intervention period, patients visited the outpatient clinic weekly for measurement of body weight, to check glucoregulation and to confirm compliance with the diet.

Exercise program

Thirteen of the 27 T2DM patients simultaneously participated in an exercise program. This program comprised a minimum of 4 days training at home for 30 min at 70% of maximum aerobic capacity on a cyclo-ergometer. Furthermore, patients participated in a weekly one-hour aerobic exercise training under supervision of a physiotherapist. Compliance was assessed by reading the heart rate monitor worn during exercise sessions both at home and in the hospital (Polar S610 itm, Polar Electro Oy, Finland). Patients in the VLCD-only group were instructed to maintain their normal pattern of physical activity during the study.

Anthropometric and laboratory measurements

At baseline and after the 16-week intervention period patients were studied after an overnight fast and after 2 days without any exercise. All T2DM patients completed the 16-week VLCD and no patients were lost to follow-up. The lean and obese control subjects were studied only once.

Height, weight, BMI and waist circumference were measured according to the World Health Organization recommendations. Blood pressure was measured with an Omron 705IT blood pressure device (Omron Matsusaka Co., Ltd., Japan) and recorded within the limits of 1 mmHg. Fat mass was assessed by bioelectrical impedance analysis (BIA, Bodystat 1500 MDD, Bodystat Ltd., Douglas, Isle of Man, United Kingdom). Blood samples were drawn for the measurement of fasting plasma levels of glucose, insulin, hemoglobin A1c (HbA1c), total cholesterol (TC), high density lipoprotein (HDL)-cholesterol, low density lipoprotein (LDL)-cholesterol and triglycerides (TG).

Proteomics measurements

Targeted protein assays through multiple reaction monitoring (MRM)

Ten µL of plasma aliquots were processed in 1.5-mL screw cap tubes. One hundred ninety five µL of 100 mM TEAB/2M urea/10% acetonitrile/1% n-octyl- glucoside/10 mM TCEP was added to the plasma samples. Samples were incubated at room temperature for one hour for complete reduction. Four µL of 0.5 M iodoacetamide (Sigma-Aldrich) was added and alkylation

Table 2. VLCD effect for proteins in the MRM dataset as compared to obese and lean controls. The numerical entries represent ratio measurements relative to a pooled reference sample.

	T2DM				**controls**		
Protein description[a]	baseline		16 weeks		obese		lean
Alpha-1-acid glycoprotein	1.02 ± 0.08		0.91 ± 0.05		1.06 ± 0.07		0.84 ± 0.06
Antitrypsin	1.01 ± 0.03		1.11 ± 0.04	#*†	0.95 ± 0.05		0.97 ± 0.04
Apolipoprotein A-I	0.95 ± 0.03	*	0.88 ± 0.04	*†	1.03 ± 0.05		1.17 ± 0.06
Apolipoprotein A-IV	1.33 ± 0.08	*†	0.71 ± 0.06	#*†	1.04 ± 0.06		1.06 ± 0.06
Apolipoprotein B-100	1.20 ± 0.07	*†	1.00 ± 0.05	#	0.98 ± 0.04		0.92 ± 0.04
Apolipoprotein C-III	1.36 ± 0.14	*	0.85 ± 0.05	#	1.02 ± 0.07		0.87 ± 0.06
Apolipoprotein E	1.24 ± 0.10	*	0.96 ± 0.05	#	1.01 ± 0.04		0.88 ± 0.05
Beta-2-glycoprotein 1	1.10 ± 0.03	*†	1.02 ± 0.04		0.97 ± 0.05		0.94 ± 0.04
Complement C3	1.17 ± 0.03	*	0.97 ± 0.04	#*	1.08 ± 0.04	*	0.85 ± 0.04
Fibrinogen alpha chain	1.04 ± 0.05	*†	1.08 ± 0.05	*†	0.87 ± 0.05		0.79 ± 0.09
Fibrinogen beta chain	1.06 ± 0.04	*†	1.12 ± 0.05	*†	0.91 ± 0.05		0.82 ± 0.07
Fibrinogen gamma chain	1.06 ± 0.04	*†	1.13 ± 0.05	*†	0.89 ± 0.04		0.82 ± 0.07
Transthyretin	0.87 ± 0.04	*†	0.85 ± 0.04	*†	1.07 ± 0.06		1.04 ± 0.04

Mean ± SEM. # significant difference within group *vs.* baseline; * significant difference *vs.* lean controls; † significant difference *vs.* obese controls.

[a] For gene symbols and accession numbers see Supplementary Table S1 in the Supporting Information.

was completed for 30 minutes at room temperature. Forty µL of each aliquot of reduced/alkylated plasma sample was digested with 12 µg sequencing grade trypsin. Digestion was stopped after overnight incubation at room temperature by adding 45 µL of 2 M urea/1% formic acid. To monitor LC/MS instrument trending, 0.3 µg fibrinopeptide A standard (AnaSpec, Fremont, CA) was spiked into each sample vial. Twenty µL of digested samples were injected for quantitative analysis.

LC-MRM analysis was performed on 4000QTrap instrument (AB/SCIEX, Concord, ON) interfaced with a U3000 HPLC system (Dionex, Sunnyvale, CA). Peptides were separated on a Targa C18 (5 mm) 15061.0 mm column (Higgins Analytical, Mountain View, CA) utilizing a 200- µL /min flow rate. Peptides were eluted carried out over a 21-min gradient from 2% B to 32%B (A: 5% acetonitrile, 0.1% formic acid, B: 95% acetonitrile, 0.1% formic acid). The HPLC column compartment was kept at 50˚C during analysis.

Two peptides and two fragments from each were carefully selected to represent the target proteins to be assayed. Thirteen target proteins were analyzed: apolipoproteins A-I, A-IV, B100, C-III, E, Beta-2-glycoprotein 1 alpha-I- antitrypsin, complement C3, fibrinogen alpha, beta, gamma chains, alpha-1-acid glycoprotein and transthyretin. Accession numbers of these proteins are given in Table 2, while Table S2 in File S1 shows the used peptide sequences.

Specimens from all T2DM and control subjects (114 samples) were analyzed in three acquisition batches. Primary samples (following every four) were interleaved with QC reference samples. MRM signals (ion intensities of fragments) from the primary samples were normalized to the median signal from the same fragments in the QC samples. This accurate relative quantification could be achieved without the need of using isotope labeled peptide standards. MRM signals were integrated using the Multiquant v1.1 software tool (AB/SCIEX).

iTRAQ Discovery Proteomics

Proteomic analysis was carried out by utilizing the 8-plex iTRAQ reagent for relative quantification [26]. In this workflow a single 2D LC-MS/MS experiment is used for the quantification of peptides (and proteins) from up to eight samples. Eight-plex experiments were configured to profile six primary samples and two replicates of reference (QC) sample that was created by combining a fraction of the primary samples. By normalizing peptide measurements from the primary samples to those in the QC samples it is feasible to compare large numbers of primary samples analyzed in different experiments. The study - 54 primary samples, 18 reference QC samples - consisted of nine such iTRAQ experiments.

One hundred µL plasma samples were delipidated by diluting with 400 µL 1XPBS (Sigma-Aldrich, St. Louis, MO) and 250 µL tetrachloroethylene (Sigma- Aldrich), vortexing thoroughly and spinning at 14,000 rpm for 10 minutes at 4˚C. The resulting top aqueous phase was transferred to a new tube for further processing.

Abundant proteins were removed from delipidated plasma in two stages utilizing IgY14 5-mL and Supermix 2-mL columns (Sigma) on a Vision HPLC Workstation (Applied Biosystems, Foster City, CA) as described earlier [27]. The protein fraction corresponding to the depletion flow-through was recovered on a Poros R1 reversed-phase column, eluted with 95% acetonitrile and dried down in a SpeedVac. Only this fraction was used for discovery proteomics. Dried protein fractions were re-suspended in 22 µL 2 M urea, 1 M TEAB, 1% n-octyl-glucoside buffer (pH 8.5) and reduced with 5 mM TCEP for one hour at room temperature. Reduced samples were alkylated by adding 1 µL 84 mM iodoacetamide and incubating in the dark for 30 minutes at room temperature. Trypsin digestion was completed overnight at a 1:10 enzyme/substrate ratio (w/w) at room temperature by adding 5 µL 1 mg/mL sequencing grade trypsin (Promega, Madison, WI) in 4 mM N-acetyl cysteine (to quench remaining iodoacetamide). Digested samples were labeled by the 8-plex iTRAQ reagents following the manufacturer's protocols (Applied Biosystems) using an amount of digest pool containing approximately 40 µg material. Primary samples were labeled with the reagents yielding the m/z 114, 115, 116, 118, 119, 121 reporter fragments in the MS/MS scans. QC samples (replicates from the reference pool) were labeled with the 113 and 117 reagents. iTRAQ labeling was quenched by the addition of 1 M ammonium bicarbonate.

Eight samples were combined to an iTRAQ mix, desalted, and fractionated by strong cation exchange (SCX) chromatography using a Poly Sulfoethyl Strong Cation Exchange Column (PolyLC, Columbia, MD) on an Agilent 1200 instrument (Agilent, Santa Clara, CA). Peptides were collected into nine SCX fractions through eluting with a gradient of 10 mM KH2PO4 to 10 mM KH2PO4/ 1M KCl at pH 3.5. SCX fractions dried and re-suspended in 50 µL 95:5:0.1 water- acetonitrile-trifluoroacetic acid (TFA) (Buffer A for HPLC). Reversed-phase separation was performed on a Dionex U3000 HPLC (Dionex, Sunnyvale, CA) with a 60-min gradient from 5% solvent B (10%

H_2O/90% ACN/0.1% TFA) to 38% B. Eleven-second HPLC fractions were collected onto MALDI plates through a Probot fraction collector (Dionex). MALDI matrix and mass calibration standard were co-infused with a syringe pump at 2-µL/min flow rate. MALDI plates were analyzed on an AB4800 mass spectrometer (Applied Biosystems/MDS SCIEX, Concord, ON, Canada) utilizing internally developed scripts for MS/MS precursor selection that was optimized to select and measure a reproducible set of peptides from each iTRAQ mix.

Peptide quantification was carried out by calculating the average iTRAQ ion intensity ratios relative to the m/z 113 and 117 peaks. Protein ratios were determined as the medians of all peptide ratios matching to the same protein. Peptide mappings are shown in Table S5 in File S1. Peptide sequences were identified from MS/MS fragmentation spectra using the Mascot search engine (Matrix Science, UK) the IPI sequence database (v3.72 of human sequences). For peptide matching trypsin specificity was used with up to two missed cleavage sites. iTRAQ modification, cysteine-alkylation, methionine oxidation, asparagine deamidation, and N-terminal pyro-Gly and pyro-cmc formation were considered as variable modifications. Precursor ion mass tolerance was 50 ppm and fragment ion tolerance was 0.4 Da. Peptide matches were validated by an internally developed procedure with an estimated rate of false peptide identification of less than 1%, as explained by Juhasz et al [27]. Once all the study samples were analyzed, the complete set of identified peptides was re-mapped to a minimum, non-redundant protein set through an internally developed procedure. During this process proteins that had unique peptides matching to them were kept separate from protein groups that shared peptides. Measured values of protein expression were normalized using a procedure based on Vandesompele et al [28].

Assays

Plasma glucose, TC, HDL-cholesterol and TG concentrations were analyzed as previously described [13] with a fully automated P-800 module (Roche, Almere, The Netherlands). Serum insulin was measured with an immunoradiometric assay (Biosource, Nivelles, Belgium). HbA1c was detected with a semi-automated HPLC machine Primus Ultra 2 (Kordia, Leiden, The Netherlands).

Statistics

The data of the two intervention groups, i.e., both the clinical data and protein expression levels measured from the MRM as well as iTRAQ platforms, were studied using a linear mixed effect model for repeated measures in order to study the influence of the VLCD and the additional exercise program. The model was fitted by Maximum Likelihood (ML). The initial model included the random patient effects to account for the correlation between 2 repeated measures within the same patient, and age, gender, treatment and time as fixed effects, for each outcome variable separately. The model was tested for significance of each individual factor and the interaction effect of time and treatment. On doing so, it was found that the effects of age and gender were not significant. The final model consisted of the random patient effects, the fixed treatment and time effects, and the interaction between treatment and time. The influence of the VLCD was tested by studying the effect of time on the model and the additional influence of exercise with VLCD was tested by studying the effect of treatment and time interaction on the model. The p values for each of the tests are reported.

Differences between all groups in the clinical dataset as well as in the MRM dataset, i.e., two intervention groups and the lean and obese control groups, were analysed using t-tests, where paired t-tests were used when comparing two time points for the same group and independent t-tests for all other comparisons.

Adjustment for multiple hypothesis testing has been performed in all proteomics analyses using the Benjamini-Hochberg (BH) method (unless otherwise stated in the text). A significance level of $p=0.05$ was used (unless otherwise stated in the text). Data are presented as mean ±SEM. The statistical analyses were conducted using the free software R version 2.10.1 with the lme4 and multcomp libraries [29–31].

Results

Effect on body weight and glucoregulation

Anthropometric and laboratory results were published previously. [11–13, 23–25] As shown in Table 1, there were no significant differences in clinical characteristics, except for systolic blood pressure, between the VLCD+exercise and the VLCD-only group at baseline. Furthermore, the control groups were well matched with both intervention groups with respect to age and gender and for the obese control group with both intervention groups at baseline with respect to weight, BMI and waist circumference. Both control groups had significantly lower levels of glucose, insulin and HbA1c.

After the 16-week VLCD there was a significant decrease in body weight in both intervention groups (-27.2±1.9 kg VLCD+exercise; -23.7±1.6 kg VLCD-only). Patients also lost a significant amount of fat mass and waist circumference. Moreover, the 16-week VLCD resulted in an impressive improvement in glycaemic control as shown by a significant decrease in HbA1c in both treatment groups (VLCD + exercise 7.8±0.4 vs. 6.3±0.4%; VLCD-only 7.8±0.3 vs. ±0.3%), despite the discontinuation of all glucose-lowering medication. In both treatment groups, plasma TG were significantly decreased to near normal values. After the 16-week intervention period the VLCD+exercise group had significantly less fat mass and a significantly lower total cholesterol level as compared to the VLCD-only group. There was no significant difference in glucoregulation between the groups after the 16-week intervention period (Table 1)

Targeted MRM analysis

A total of 15 proteins, including 2 internal control proteins (not shown), were quantified using MRM and mass spectrometry in the VLCD groups, with and without exercise and before and after the intervention. These proteins were also quantified in the obese and lean controls.

Intervention effects

After 16 weeks, there was a significant decrease in concentrations of apolipoproteins A-IV, B-100, C-III and E as well as of Complement C3 in both intervention groups. These effects were however not significantly different between the two intervention groups (see Table S1 in File S1). Since no additional influence of exercise with VLCD was observed for any of the proteins in the MRM set, the VLCD+exercise and VLCD-only groups were combined for further analysis into one group of T2DM patients. Table 2 shows the full comparison of the combined T2DM group at the two time points (T2DM0 and T2DM16, respectively) with the two control groups (lean and obese).

Apolipoprotein A-IV showed the most significant effect of VLCD among all proteins considered in the MRM dataset. Apolipoprotein A-IV concentration did not differ between both control groups (lean 1.06±0.06 vs. obese 1.04±0.06 A.U., p=0.90), however, the level for T2DM patients was significantly higher (1.33±0.08 A.U.) compared to both control groups (p=0.01 for lean and p=0.04 for obese) before the diet, whereas the level for T2DM patients was significantly lower to those of the controls (p=0.002 for lean and p=0.003 for obese) after the diet.

Also for apolipoproteins E, C-III and B-100 the concentration levels were significantly higher for T2DM patients at baseline compared to the lean control group, and VLCD resulted in significant decrease in their concentration levels. Contrary to Apolipoprotein A-IV, these decreased levels after the diet are not significantly different from the control groups.

Disease state discriminating proteins

Obesity associated markers - Only Complement C3 showed a significant difference between the lean control group and all three other groups. These differences were highly significant for lean against obese (0.85±0.04 vs 1.08±0.04 A.U., p=0.001) as well as for lean against the T2DM group at baseline (0.85±0.04 vs 1.17±0.03 A.U., p,0.001). Upon the VLCD, the concentrations of C3 decreased (from 1.17±0.03 to 0.97±0.04 A.U., p,0.001), and, although still significantly different (p=0.04), approached the concentration in the lean control group.

Diabetes associated markers - The fibrinogens alpha, beta and gamma chains all showed the same behaviour. Namely, all three showed a significantly increased level for T2DM patients as compared to both the lean and the obese controls, both at baseline and after 16 weeks of VLCD, although these increased levels for T2DM patients at baseline as compared to the obese were only just significant (p=0.04 for all three). Moreover, all three showed no significant difference between obese and lean (p=0.58 for all three) nor between the T2DM patients before and after the diet (p>0.12). Transthyretin showed a very similar behaviour as the fibrinogens except for the fact that the transthyretin level was lower for T2DM patients as compared to the controls.

Large scale iTRAQ analysis

A total of 635 proteins were quantified using iTRAQ and mass spectrometry in the VLCD groups, with and without exercise and before and after the intervention. Only 234 of those proteins could be measured for all 27 patients on both time- points (i.e., at baseline and after the 16-week VLCD). These included two proteins added as internal controls. The data analysis was applied on the remaining 232 proteins.

Exercise associated markers

Table 3. Top 15 proteins identified from iTRAQ experiments showing a VLCD effect. The numerical entries represent ratio measurements relative to a pooled reference sample.

	T2DM		
Protein description[a]	Baseline	16 weeks	adj. p-value
Biotinidase	0.96 ± 0.02	0.86 ± 0.02	6.0E-07
Selenoprotein P	0.80 ± 0.02	0.95 ± 0.02	1.0E-06
Insulin-like growth factor-binding protein 2	0.68 ± 0.04	1.04 ± 0.06	1.0E-06
Isoform 2 of Inter-alpha-trypsin inhibitor heavy chain H4	0.79 ± 0.02	0.97 ± 0.02	1.0E-06
Isoform 1 of Sex hormone-binding globulin	0.72 ± 0.05	1.18 ± 0.09	1.7E-06
Isoform 3 of Interleukin-1 receptor accessory protein	0.75 ± 0.02	0.88 ± 0.03	2.0E-06
Afamin precursor	0.95 ± 0.04	0.78 ± 0.03	2.6E-05
Apolipoprotein A-IV precursor	1.23 ± 0.09	0.78 ± 0.06	2.6E-05
Leucine-rich alpha-2-glycoprotein precursor	0.72 ± 0.02	0.92 ± 0.04	2.7E-05
Beta-Ala-His dipeptidase	1.12 ± 0.05	0.91 ± 0.03	3.4E-05
Leucine-rich alpha-2-glycoprotein precursor	0.65 ± 0.06	0.90 ± 0.09	3.4E-05
Apolipoprotein A-IV precursor	1.21 ± 0.12	0.77 ± 0.08	5.9E-05
Lysozyme C precursor	0.81 ± 0.04	0.92 ± 0.04	6.6E-05
Pigment epithelium-derived factor precursor	1.22 ± 0.05	0.95 ± 0.04	6.8E-05
Fructose-bisphosphate aldolase B	1.05 ± 0.06	0.80 ± 0.03	2.5E-04

Mean ± SEM.

Of the 232 proteins 18 showed a significant exercise effect when considering the unadjusted p-value measured by the interaction of treatment and time from the model (see Table S3 in File S1). Amongst these, for two proteins SHBG and MASP-1, the p-value was lower than 0.005. For SHBG, the mean of the measurements for VLCD+exercise at 16 weeks showed a stronger increase from the measurements at baseline (VLCD+exercise at 16 weeks: 1.32±0.16 A.U. vs VLCD+exercise at baseline: 0.69±0.09 A.U.) in comparison to the increase for VLCD-only (VLCD-only at 16 weeks: 1.04±0.08 A.U. vs VLCD-only at baseline: 0.74±0.05 A.U.). For MASP-1, the level for the VLCD+exercise group was decreased after 16 weeks (VLCD+exercise at 16 weeks: 0.86±0.04 A.U. vs VLCD+exercise at baseline: 0.99±0.03 A.U.) whereas the level for the VLCD-only group hardly changed (VLCD at 16 weeks: 0.93±0.02 A.U. vs VLCD at baseline: 0.92±0.03 A.U.). However, on applying the multiple testing correction, none of the analytes were found to be significant.

VLCD associated markers

Of the 232 proteins, 87 showed a significant VLCD effect, where the effect is considered to be statistically significant if the unadjusted p-value from the model for the effect of time was less than 0.05 and the Benjamini-Hochberg-adjusted p- value was less than 0.10. Fourtysix proteins from these significant cases were up- regulated after treatment, i.e., the measured expression levels were higher after 16 weeks of VLCD than at baseline, while the other 41 proteins were down-regulated after 16 weeks of VLCD. The top 13 proteins (based on p-value) identified from iTRAQ experiments showing a VLCD effect are shown in Table 3. A list of all proteins and their changes after 16 weeks of VLCD are shown in Table S4 in File S1. Fourtyfour of the significantly changed proteins could be traced to pathways in KEGG, with 17 of them being present in the Complement and Coagulation cascade.

Discussion

Using a targeted MRM analysis we showed that several proteins differ between T2DM patients before and after a VLCD and between T2DM patients and lean and obese controls. As shown in Figure 2, these proteins can be divided in subgroups based on similar patterns of differences between the groups. Thereby a distinction can be made between potential biomarkers that are intervention (diet) or disease state (diabetes or obesity) associated.

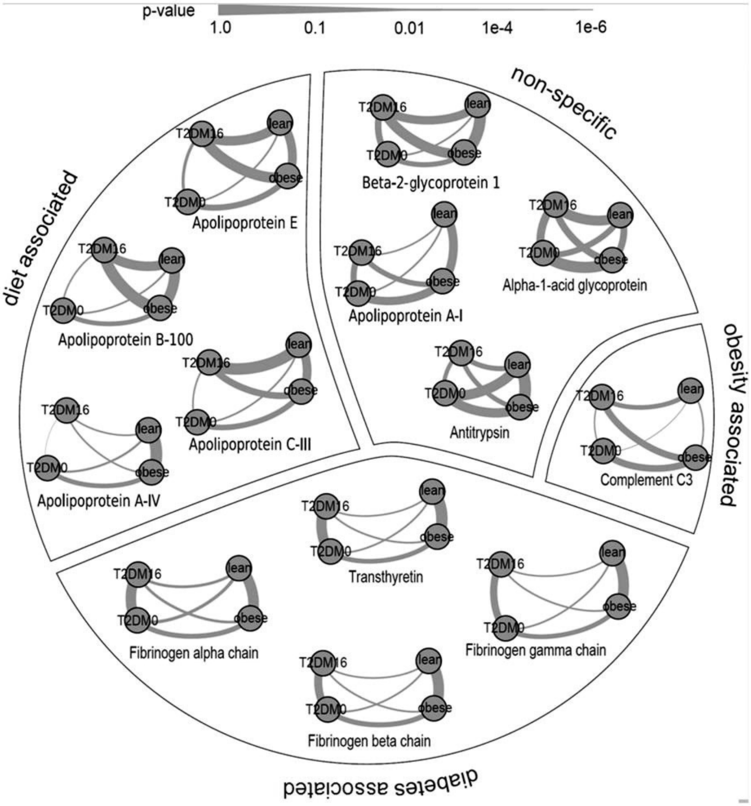


Figure 2 Graph representation of group wise comparisons for the proteins in the MRM data set. Comparisons between all pairs of the four groups, i.e., the diabetes patients at baseline (T2DM0) and after 16 weeks of VLCD (T2DM16) as well as the obese and lean control groups, are represented by edges, where the thickness of the edge represents the p-value. Groups that hardly can be discerned are thus connected by thick edges and located close together, whereas groups that can be well distinguished are connected by thin edges and are slightly further distinct. Furthermore, the proteins have been clustered into groups (i.e., obesity associated, diabetes associated, diet associated, and non- associated) based on similarity in patterns of differences between the groups.

Diet associated markers

The proteins showing a diet effect most evidently in this study were the apolipoproteins, especially apolipoprotein A-IV (APOA-IV), as shown in Figure 2. APOA-IV is synthesized by the enterocytes of the small intestine in response to fat absorption [32, 33]. Although the precise role of APOA-IV has not been fully elucidated, studies suggest that it has anti-atherogenic

[34] and anti- inflammatory [35] properties and that it serves as a satiety factor [32, 36].

Interestingly, APOA-IV levels were significantly higher in T2DM patients before the diet as compared to controls, which was also found in earlier studies [37, 38]. In contrast to our study, some also showed higher APOA-IV levels in obese, non- diabetic mice and humans [39, 40], though others did not find an association between APOA-IV and BMI [41]. An explanation for the higher APOA-IV levels, which is counter-intuitive, has not yet been identified. Shen et al. showed that obese mice, although peripheral APOA-IV levels were high, have lower APOA-IV levels in the hypothalamus, the site where APOA-IV is thought to exert its effect on satiety [39]. It has also been hypothesized that the high APOA-IV levels reflect a state of APOA-IV resistance [42], as is the case for leptin, which is also been thought to regulate APOA-IV [39]. APOA-IV also showed the highest MFC in response to the VLCD, resulting in significantly lower APOA-IV levels in T2DM patients after the diet than in controls. This decrease was consistently shown in 100% of the subjects, indicating that a decrease in APOA-IV might be a marker for weight loss. On the other hand it is known that APOA-IV levels are influenced by changes in dietary fat content [40, 43] and the observed decrease might thus be more reflective of the low amount of fat intake and caloric restriction during the VLCD. It would be interesting to investigate APOA-IV levels in patients back in a eucaloric state to elucidate this further. Furthermore, it has been hypothesized that, as APOA-IV serves as a satiety factor, lower APOA-IV levels can be a signal for stimulating feeding behavior [44]. Low APOA-IV levels may therefore contribute to the difficulties in maintaining achieved weight loss over longer periods of time. In this context it is interesting that in a study by Culnan et al., using iTRAQ proteomic analysis, an increase in APOA-IV levels was shown after weight loss induced by Roux-en-Y gastric bypass surgery (RYGB), which is known to result in more sustained weight loss as compared to diets [42]. The contrasting APOA-IV levels may, however, be explained by the fact that those after-surgery levels were measured after a mean follow-up of 19.2 months post-RYGB, and by the altered anatomy of the small intestine, the production site of APOA-IV.

Complement C3 - an obesity associated marker

Accumulating evidence shows that both T2DM and obesity are associated with a chronic inflammatory state [3]. Complement C3 (C3) has an important role in the immune system and is produced by the liver, adipose tissue and macrophages [45]. Our MRM analysis showed higher concentrations of C3 in obese T2DM patients and healthy obese subjects as compared to lean controls. This agrees with several other studies, that also showed such elevated levels of C3 in patients with obesity [45, 46]. Furthermore, a significant decrease in C3 levels was seen after the VLCD, whereas no differences were shown between obese subjects with or without T2DM, indicating that C3 might be a marker of obesity rather than T2DM. However, other studies have demonstrated C3 levels to be increased in lean versus obese T2DM patients and to be associated with diabetes development independently of body weight [47, 48]. C3 has also proved to be higher in young adults with type 1 diabetes and a decrease in HbA1c in this group has been associated with a decrease in C3 levels [49]. These data indicate that C3 level and changes therein are dependent on the pathophysiology of the patient.

Diabetes-associated markers

The fibrinogens were found to be elevated in T2DM patients as compared to both lean controls and obese controls, before as well as after the diet. Furthermore, concentrations did not differ between lean and obese controls, suggesting that fibrinogen is more diabetes than obesity associated. A high fibrinogen level is thought to reflect a hypercoagulable state and is suggested to be a strong independent cardiovascular risk factor [50, 51]. Other studies also found high fibrinogen levels in T2DM patients [52, 53] and this may contribute to the increased risk of cardiovascular events in type 2 diabetes [54, 55]. However, not all studies showed an increased fibrinogen level in T2DM patients [56]. After the VLCD we did not observe differences in the fibrinogen levels, although weight loss has been associated with a decrease in fibrinogen in literature [57].

Another interesting diabetes-associated marker found in this study is transthyrethin (TTR). TTR, previously known as pre-albumin, is a carrier protein for thyroid hormones and retinol-binding protein and is produced in the liver, choroid plexus and pancreatic islets [58]. TTR has been used as a biomarker for malnutrition [59–61] and has been shown to decrease in response to a VLCD [62– 64]. However, it has been shown by Afolabi et al. that after an initial decrease at 5% weight loss, TTR levels returned back to baseline upon further weight loss [65]. In our study, where the average weight loss is 22%, also no differences were found after the VLCD.

Exercise effect

No significant exercise effect was observed for any of the proteins in the MRM analysis. Therefore, we performed a large scale iTRAQ proteomic analysis to reveal candidate pathways involved in the additional beneficial effects of adding an exercise program that we have shown before [13]. Without correcting for multiple testing, concentrations were significantly different between the two VLCD groups for a few proteins, of which especially sex hormone-binding globulin (SHBG, P04278) and mannose-binding lectin (MBL)-associated serine protease (MASP-1, P48740) could be interesting. MASP-1 is a protease that contributes to the activation of the lectin complement pathway [66]. SHBG has been related to exercise before, although the influence of exercise on SHBG levels is less clear [67– 70]. Moreover, SHBG levels are known to be inversely associated with insulin resistance and are thought to predict the risk on T2DM [71]. After correction for multiple testing, however, none of the proteins showed significant differences between the groups any more. Further research on these specific proteins is needed to uncover possible pathways involved in the beneficial effects of exercise.

Strengths and limitations

The major strength of our study is the VLCD intervention. By studying T2DM patients before and after the diet, we showed that several proteins change with weight loss and improved glycemic control. By comparing the patients to obese and lean controls, these proteins could further be discerned between diabetes- associated and obesity-associated markers.

Limitations of our study are the lack of a control (non-diabetic obese) VLCD group as well as the absence of lean and obese control groups in the iTRAQ analysis. In addition, because of the many comparisons and the consequentially required correction for multiple testing, no significant differences were found using the iTRAQ analysis. The study would benefit from quantification of one or more promising candidates by an independent complementary technology (e.g. ELISA). This was, however, beyond the scope of the current study.

In conclusion, using proteomic analysis several potential disease state and intervention associated markers were found distinguishing T2DM patients from obese and lean controls and showing a VLCD effect. Although no specific exercise markers were discovered, the iTRAQ analysis indicated some proteins as potential interesting targets for further research.

References

1. Danaei G, Finucane M, Lu Y, Singh G, Cowan M: **National, regional, and global trendsin fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surverys and epidemiological studies with 370 country-years and 2.7 million participants**. *Lancet* 2011, **2**:31–40.
2. Finucane MM, Stevens G a, Cowan MJ, Danaei G, Lin JK, Paciorek CJ, Singh GM, Gutierrez HR, Lu Y, Bahalim AN, Others: **National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9{\textperiodcentered} 1 million participants**. *Lancet* 2011, **377**:557–567.
3. Donath MY, Shoelson SE: **Type 2 diabetes as an inflammatory disease**. *Nat Rev Immunol* 2011, **11**:98–107.
4. Matthaei S, Stumvoll M, Kellerer M, Häring HU: **Pathophysiology and pharmacological treatment of insulin resistance.** *Endocr Rev* 2000, **21**:585–618.
5. Dandona P, Aljada A: **A rational approach to pathogenesis and treatment of type 2 diabetes mellitus, insulin resistance, inflammation, and atherosclerosis.** *Am J Cardiol* 2002, **90**:27G–33G.
6. Mokdad a H, Bowman B a, Ford ES, Vinicor F, Marks JS, Koplan JP: **The continuing epidemics of obesity and diabetes in the United States.** *JAMA* 2001, **286**:1195–1200.
7. CDC: **Prevalence of overweight and obesity among adults with diagnosed diabetes--United States, 1988-1994 and 1999-2002.** *MMWR Morb Mortal Wkly Rep* 2004, **53**:1066–1068.
8. Karelis AD: **Metabolically healthy but obese individuals**. *Lancet* 2008, **372**:1281–1283.
9. Kramer CK, Zinman B, Retnakaran R: **Are Metabolically Healthy Overweight and Obesity Benign Conditions?A Systematic Review and Meta-analysis**. *Ann Intern Med* 2013, **159**:758–769.
10. Jazet IM, de Craen AJ, van Schie EM, Meinders a. E: **Sustained beneficial metabolic effects 18 months after a 30-day very low calorie diet in severely obese, insulin-treated patients with type 2 diabetes**. *Diabetes Res Clin Pract* 2007, **77**:70–76.
11. Snel M, Sleddering M a, Inge D, Romijn J a, Pijl H, Meinders a E, Jazet IM: **European Journal of Internal Medicine Quality of life in type 2 diabetes mellitus after a very low calorie diet and exercise**. *Eur J Intern Med* 2012, **23**:143–149.

12. Snel M, Jonker JT, Hammer S, Kerpershoek G, Lamb HJ, Meinders a. E, Pijl H, de Roos A, Romijn J a., Smit JW a., Jazet IM: **Long-Term Beneficial Effect of a 16-Week Very Low Calorie Diet on Pericardial Fat in Obese Type 2 Diabetes Mellitus Patients**. *Obesity* 2012, **20**:1572–1576.
13. Snel M, Gastaldelli A, Ouwens DM, Hesselink MKC, Schaart G, Buzzigoli E, Frölich M, Romijn J a., Pijl H, Meinders a. E, Jazet IM: **Effects of adding exercise to a 16-week very low-calorie diet in obese, insulin-dependent type 2 diabetes mellitus patients**. *J Clin Endocrinol Metab* 2012, **97**:2512–2520.
14. Herder C, Karakas M, Koenig W: **Biomarkers for the prediction of type 2 diabetes and cardiovascular disease.** *Clin Pharmacol Ther* 2011, **90**:52–66.
15. Lyons TJ, Basu A: **Biomarkers in diabetes: Hemoglobin A1c, vascular and tissue markers**. *Transl Res* 2012, **159**:303–312.
16. Garcia-Ramirez M, Canals F, Hernandez C, Colome N, Ferrer C, Carrasco E, Garcia-Arumi J, Simo R, Mesa J: **Proteomic analysis of human vitreous fluid by DIGE: a new strategy for identifying potential candidates in the pathogenesis of proliferative diabetic retinopathy**. *Diabetologia* 2006, **55**:158.
17. Kim H-J, Cho E-H, Yoo J-H, Kim P-K, Shin J-S, Kim M-R, Kim C-W: **Proteome analysis of serum from type 2 diabetics with nephropathy.** *J Proteome Res* 2007, **6**:735–743.
18. Kriegel TM, Heidenreich F, Kettner K, Pursche T, Hoflack B, Grunewald S, Poenicke K, Glander HJ, Paasch U: **Identification of diabetes- and obesity-associated proteomic changes in human spermatozoa by difference gel electrophoresis**. *Reprod Biomed Online* 2009, **19**:660–670.
19. Maris M, Overbergh L, Mathieu C: **Type 2 diabetes: Gaining insight into the disease process using proteomics**. *Proteomics - Clin Appl* 2008, **2**:312–326.
20. Rao P V, Reddy AP, Lu X, Dasari S, Biggs E, Jr CTR, Nagalla SR, Krishnaprasad A, Roberts CT: **Article Proteomic Identification of Salivary Biomarkers of Type-2 Diabetes Proteomic Identification of Salivary Biomarkers of Type-2 Diabetes**. *J Proteome Res* 2009, **8**(January):239–245.
21. Riaz S, Alam S, Srai S, Skinner V, Riaz A, Akhtar M: **Proteomic Identification of Human Serum Biomarkers in Diabetes Mellitus Type 2**. *J Pharm Biomed Anal* 2010, **51**:1103–1107.

22. Riaz S, Alam S, Srai S, Skinner V, Riaz a, Akhtar M: **Proteomic Identification of Human Urinary Biomarkers in Diabetes Mellitus Type 2**. *Diabetes Technol Ther* 2010, **12**:979–988.
23. Hammer S, Snel M, Lamb HJ, Jazet IM, van der Meer RW, Pijl H, Meinders E a., Romijn J a., de Roos A, Smit JW a: **Prolonged Caloric Restriction in Obese Patients With Type 2 Diabetes Mellitus Decreases Myocardial Triglyceride Content and Improves Myocardial Function**. *J Am Coll Cardiol* 2008, **52**:1006–1012.
24. Snel M, van Diepen J a., Stijnen T, Pijl H, Romijn J a., Meinders a. E, Voshol P, Jazet IM: **Immediate and long-term effects of addition of exercise to a 16-week very low calorie diet on low-grade inflammation in obese, insulin-dependent type 2 diabetic patients**. *Food Chem Toxicol* 2011, **49**:3104–3111.
25. Wang Y, Snel M, Jonker JT, Hammer S, Lamb HJ, De Roos A, Meinders a. E, Pijl H, Romijn J a., Smit JW a, Jazet IM, Rensen PCN: **Prolonged caloric restriction in obese patients with type 2 diabetes mellitus decreases plasma CETP and increases apolipoprotein AI levels without improving the cholesterol efflux properties of HDL**. *Diabetes Care* 2011, **34**:2576–2580.
26. Choe L, Ascenzo MD, Relkin NR, Pappin D, Ross P, Williamson B, Guertin S, Pribil P, Lee KH: **8-Plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer ' s disease**. *Proteomics* 2007, **7**:3651–3660.
27. Juhasz P, Lynch M, Sethuraman M, Campbell J, Hines W, Paniagua M, Song L, Kulkarni M, Adourian A, Guo Y, Li X, Martin S, Gordon N: **Semi-targeted plasma proteomics discovery workflow utilizing two-stage protein depletion and off-line LC-MALDI MS/MS**. *J Proteome Res* 2011, **10**:34–45.
28. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F: **Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes.** *Genome Biol* 2002, **3**:RESEARCH0034.
29. Computing, R Foundation for Statistical Vienna A: **R: A language and environment for statistical computing.** 2008.
30. Bates D, Maechler M, Bolker B: **lme Linear mixed-effects models using S4 classes.** *R Packag version 0* 2011, **4 SRC - G**:999342–999375.
31. Hothorn T, Bretz F, Westfall P: **Simultaneous inference in general parametric models**. *Biometrical J* 2008, **50**:346–363.

32. Tso P, Liu M: **Apolipoprotein A-IV, food intake, and obesity.** *Physiol Behav* , **83**:631–643.
33. Green PH, Glickman RM, Riley JW, Quinet E: **Human apolipoprotein A-IV. Intestinal origin and distribution in plasma.** *J Clin Invest* 1980, **65**:911–919.
34. Kronenberg F, Stühlinger M, Trenkwalder E, Geethanjali FS, Pachinger O, von Eckardstein a, Dieplinger H: **Low apolipoprotein A-IV plasma concentrations in men with coronary artery disease.** *J Am Coll Cardiol* 2000, **36**:751–757.
35. Quilliot D, Walters E, Guerci B, Fruchart JC, Duriez P, Drouin P, Ziegler O: **Effect of the inflammation, chronic hyperglycemia, or malabsorption on the apolipoprotein A-IV concentration in type 1 diabetes mellitus and in diabetes secondary to chronic pancreatitis**. *Metabolism* 2001, **50**:1019–1024.
36. Fujimoto K, Cardelli J a, Tso P: **Increased apolipoprotein A-IV in rat mesenteric lymph after lipid meal acts as a physiological signal for satiation.** *Am J Physiol* 1992, **262**(6 Pt 1):G1002–G1006.
37. Sun Z, Larson I a., Ordovas JM, Barnard JR, Schaefer EJ: **Effects of age, gender, and lifestyle factors on plasma apolipoprotein A-IV concentrations**. *Atherosclerosis* 2000, **151**:381–388.
38. Vergès BL, Vaillant G, Goux A, Lagrost L, Brun JM, Gambert P: **Apolipoprotein A-IV levels and phenotype distribution in NIDDM**. *Diabetes Care* 1994, **17**:810–817.
39. Shen L, Tso P, Woods SC, Sakai RR, Davidson WS, Liu M: **Hypothalamic apolipoprotein A-IV is regulated by leptin**. *Endocrinology* 2007, **148**:2681–2689.
40. Vergès B, Guerci B, Durlach V, Galland-Jos C, Paul JL, Lagrost L, Gambert P: **Increased plasma apoA-IV level is a marker of abnormal postprandial lipemia: a study in normoponderal and obese subjects**. *J Lipid Res* 2001, **42**:2021–2029.
41. Ehnholm C, Tenkanen H, De Knijff P, Havekes L, Rosseneu M, Menzel HJ, Tiret L: **Genetic polymorphism of apolipoprotein A-IV in five different regions of Europe. Relations to plasma lipoproteins and to history of myocardial infarction: The EARS study**. *Atherosclerosis* 1994, **107**:229–238.
42. Culnan DM, Cooney RN, Stanley B, Lynch CJ: **Apolipoprotein A-IV, a putative satiety/antiatherogenic factor, rises after gastric bypass.** *Obesity (Silver Spring)* 2009, **17**:46–52.

43. Weinberg RB, Dantzker C, Patton CS: **Sensitivity of serum apolipoprotein A-IV levels to changes in dietary fat content.** *Gastroenterology* 1990, **98**:17–24.
44. Bertile F, Schaeffer C, Maho YL, Raclot T, Van Dorsselaer A: **A proteomic approach to identify differentially expressed plasma proteins between the fed and prolonged fasted states**. *Proteomics* 2009, **9**:148–158.
45. Onat A, Can G, Rezvani R, Cianflone K: **Complement C3 and cleavage products in cardiometabolic risk**. *Clin Chim Acta* 2011, **412**:1171–1179.
46. Hernández-Mijares a, Jarabo-Bueno MM, López-Ruiz a, Solá-Izquierdo E, Morillas-Ariño C, Martínez-Triguero ML: **Levels of C3 in patients with severe, morbid and extreme obesity: its relationship to insulin resistance and different cardiovascular risk factors.** *Int J Obes (Lond)* 2007, **31**:927–932.
47. Yang Y, Lu HL, Zhang J, Yu HY, Wang HW, Zhang MX, Cianflone K: **Relationships among acylation stimulating protein, adiponectin and complement C3 in lean vs obese type 2 diabetes.** *Int J Obes (Lond)* 2006, **30**:439–446.
48. Engström G, Hedblad B, Eriksson K-F, Janzon L, Lindgärde F: **Complement C3 is a risk factor for the development of diabetes: a population-based cohort study.** *Diabetes* 2005, **54**:570–575.
49. Hess K, Alzahrani SH, Mathai M: **A novel mechanism for hypofibrinolysis in diabetes : the role of complement C3**. *Diabetologia* 2011, **55**:1103–1113.
50. Barazzoni R, Kiwanuka E, Zanetti M, Cristini M, Vettore M, Tessari P: **Insulin acutely increases fibrinogen production in individuals with type 2 diabetes but not in individuals without diabetes**. *Diabetes* 2003, **52**:1851–1856.
51. Ceriello a, Mercuri F, Fabbro D, Giacomello R, Stel G, Taboga C, Tonutti L, Motz E, Damante G: **Effect of intensive glycaemic control on fibrinogen plasma concentrations in patients with Type II diabetes mellitus. Relation with beta-fibrinogen genotype.** *Diabetologia* 1998, **41**:1270–1273.
52. Kannel WB, D'Agostino RB, Wilson PW, Belanger AJ, Gagnon DR: **Diabetes, fibrinogen, and risk of cardiovascular disease: the Framingham experience.** *Am Heart J* , **120**:672–676.
53. Zhao Y, Zhang J, Zhang J, Wu J: **Diabetes mellitus is associated with shortened activated partial thromboplastin time and increased fibrinogen values**. *PLoS One* 2011, **6**.

54. Becker a, van der Does FEE, van Hinsbergh VWM, Heine RJ, Bouter LM, Stehouwer CD a: **Improvement of glycaemic control in type 2 diabetes: favourable changes in blood pressure, total cholesterol and triglycerides, but not in HDL cholesterol, fibrinogen, Von Willebrand factor and (pro)insulin.** *Neth J Med* 2003, **61**:129–136.
55. Jae SY, Heffernan KS, Lee MK, Fernhall B, Park WH: **Relation of Cardiorespiratory Fitness to Inflammatory Markers, Fibrinolytic Factors, and Lipoprotein(a) in Patients With Type 2 Diabetes Mellitus**. *Am J Cardiol* 2008, **102**:700–703.
56. Missov RM, Stolk RP, Van Der Bom JG, Hofman A, Bots ML, Pols H a P, Grobbee DE: **Plasma fibrinogen in NIDDM the Rotterdam study**. *Diabetes Care* 1996, **19**:157–159.
57. Ernst E: **Does weight loss reduce plasma fibrinogen?** *Br Heart J* 1993, **70**:116–118.
58. Jacobsson B, Collins VP, Grimelius L, Pettersson T, Sandstedt B, Carlström a: **Transthyretin immunoreactivity in human and porcine liver, choroid plexus, and pancreatic islets.** *J Histochem Cytochem* 1989, **37**:31–37.
59. Kelleher PC, Phinney SD, Sims E a, Bogardus C, Horton ES, Bistrian BR, Amatruda JM, Lockwood DH: **Effects of carbohydrate-containing and carbohydrate-restricted hypocaloric and eucaloric diets on serum concentrations of retinol-binding protein, thyroxine-binding prealbumin and transferrin.** *Metabolism* 1983, **32**:95–101.
60. Ritz P, Becouarn G, Douay O, Sallé A, Topart P, Rohmer V: **Gastric Bypass is not Associated with Protein Malnutrition in Morbidly Obese Patients**. *Obes Surg* 2009, **19**:840–844.
61. Ramalho R, Guimarães C, Gil C, Neves C, Guimarães JT, Delgado L: **Morbid Obesity and Inflammation: A Prospective Study after Adjustable Gastric Banding Surgery**. *Obes Surg* 2009, **19**:915–920.
62. Scalfi L, Contaldo F, Borrelli R, De Caterina M, Spagnuolo G, Alfieri R, Mancini M: **Protein balance during very-low-calorie diets for the treatment of severe obesity.** *Ann Nutr Metab* 1987, **31**:154–159.
63. Pasquali R, Casimirri F, Melchionda N: **Protein metabolism in obese patients during very low-calorie mixed diets containing different amounts of proteins and carbohydrates.** *Metabolism* 1987, **36**:1141–1148.
64. Hoffer LJ, Bistrian BR, Young VR, Blackburn GL, Wannemacher RW: **Metabolic effects of carbohydrate in low-calorie diets.** *Metabolism* 1984, **33**:820–825.

65. Afolabi PR, Jahoor F, Jackson A a, Stubbs J, Johnstone AM, Faber P, Lobley G, Gibney E, Elia M: **The effect of total starvation and very low energy diet in lean men on kinetics of whole body protein and five hepatic secretory proteins.** *Am J Physiol Endocrinol Metab* 2007, **293**:E1580–E1589.
66. Takahashi M, Iwaki D, Kanno K, Ishida Y, Xiong J, Matsushita M, Endo Y, Miura S, Ishii N, Sugamura K, Fujita T: **Mannose-binding lectin (MBL)-associated serine protease (MASP)-1 contributes to activation of the lectin complement pathway.** *J Immunol* 2008, **180**:6132–6138.
67. Tymchuk CN, Tessler SB, Barnard RJ: **Changes in sex hormone-binding globulin, insulin, and serum lipids in postmenopausal women on a low-fat, high-fiber diet combined with exercise.** *Nutr Cancer* 2000, **38**:158–162.
68. Tymchuk CN, Tessler SB, Aronson WJ, Barnard RJ: **Effects of diet and exercise on insulin, sex hormone-binding globulin, and prostate-specific antigen.** *Nutr Cancer* 1998, **31**:127–131.
69. Caballero MJ, Mahedero G, Hernández R, Alvarez JL, Rodríguez J, Rodríguez I, Maynar M: **Effects of physical exercise on some parameters of bone metabolism in postmenopausal women.** *Endocr Res* 1996, **22**:131–138.
70. Caballero MJ, Mena P, Maynar M: **Changes in sex hormone binding globulin, high density lipoprotein cholesterol and plasma lipids in male cyclists during training and competition**. *Eur J Appl Physiol Occup Physiol* 1992, **64**:9–13.
71. Ding EL, Song Y, Manson JE, Hunter DJ, Lee CC, Rifai N, Buring JE, Gaziano JM, Liu S: **Sex hormone-binding globulin and risk of type 2 diabetes in women and men.** *N Engl J Med* 2009, **361**:1152–1163.

Supplementary section

Supplementary Table S1. Exercise effect for proteins in the MRM dataset.

	VLCD+exercise		VLCD only		
	baseline	16 weeks	baseline	16 weeks	adj. p-value
Alpha-1-acid glycoprotein	0.99 ± 0.06	0.88 ± 0.09	1.05 ± 0.14	0.93 ± 0.07	0.96
Antitrypsin	1.02 ± 0.05	1.13 ± 0.05	1.01 ± 0.03	1.09 ± 0.07	0.80
Apolipoprotein A-I	0.93 ± 0.03	0.84 ± 0.04	0.96 ± 0.05	0.93 ± 0.07	0.60
Apolipoprotein A-IV	1.33 ± 0.10	0.68 ± 0.08	1.34 ± 0.13	0.74 ± 0.09	0.80
Apolipoprotein B-100	1.15 ± 0.11	0.92 ± 0.07	1.25 ± 0.11	1.07 ± 0.06	0.78
Apolipoprotein C-III	1.33 ± 0.25	0.72 ± 0.06	1.40 ± 0.16	0.97 ± 0.07	0.78
Apolipoprotein E	1.29 ± 0.19	0.86 ± 0.06	1.20 ± 0.10	1.06 ± 0.07	0.43
Beta-2-glycoprotein 1	1.14 ± 0.05	0.99 ± 0.05	1.08 ± 0.04	1.05 ± 0.06	0.43
Complement C3	1.22 ± 0.05	0.94 ± 0.06	1.14 ± 0.04	1.00 ± 0.05	0.43
Fibrinogen alpha chain	1.09 ± 0.08	1.09 ± 0.09	0.99 ± 0.06	1.07 ± 0.07	0.72
Fibrinogen beta chain	1.11 ± 0.07	1.11 ± 0.08	1.01 ± 0.05	1.12 ± 0.05	0.43
Fibrinogen gamma chain	1.11 ± 0.07	1.13 ± 0.08	1.00 ± 0.05	1.13 ± 0.06	0.43
Transthyretin	0.92 ± 0.07	0.87 ± 0.07	0.83 ± 0.04	0.82 ± 0.04	0.78

Mean ± SEM.

Supplementary Table S2. Parameters for the protein MRM measurements

Protein	Gene Symbol	Peptide Sequence	Transition	Q1 (m/z)	Q3 (m/z)
Apolipoprotein A-I	APOA1	THLAPYSDELR	MH_3^{3+} -> b_4^+	434.6	423.3
			MH_3^{3+} -> y_5^+	434.6	619.3
		ATEHLSTLSEK	MH_3^{3+} -> y_3^+	405.9	363.3
			MH_3^{3+} -> y_9^{2+}	405.9	522.3
Apolipoprotein A-IV	APOA4	IDQNVEELK	MH_2^{2+} -> y_4^+	544.3	518.3
			MH_2^{2+} -> y_6^+	544.3	731.4
		SLAPYAQDTQEK	MH_2^{2+} -> y_9^{2+}	675.8	575.8
			MH_2^{2+} -> y_{10}^{2+}	675.8	540.3
Apolipoprotein B-100	APOB	TEVIPPLIENR	MH_2^{2+} -> y_7^+	640.8	838.4
			MH_2^{2+} -> y_7^{2+}	640.8	419.8
		FPEVDVLTK	MH_2^{2+} -> y_7^+	524.3	803.5
			MH_2^{2+} -> y_4^+	524.3	450.8
Apolipoprotein C-III	APOC3	ADALSSVQESQVAQQAR	MH_3^{3+} -> b_4^+	572.9	672.4
			MH_3^{3+} -> y_5^+	572.9	800.4
		GWVTDGFSSLK	MH_2^{2+} -> y_6^+	598.8	638.4
			MH_2^{2+} -> y_8^+	598.8	854.4
Apolipoprotein E	APOE	LAVYQAGAR	MH_2^{2+} -> y_6^+	474.8	665.3
			MH_2^{2+} -> y_7^+	474.8	764.4
		LGPLVEQGR	MH_2^{2+} -> y_5^+	484.8	588.3
			MH_2^{2+} -> y_7^{2+}	484.8	399.7
Beta-2-glycoprotein 1	APOH	ATVVYQGER	MH_2^{2+} -> y_6^+	511.8	652.3
			MH_2^{2+} -> y_7^+	511.8	751.4
		VCPFAGILENGAVR	MH_3^{3+} -> y_5^+	501.6	516.3
			MH_3^{3+} -> y_6^+	501.6	645.3
Complement C3	C3	SSLSVPYVIVPLK	MH_2^{2+} -> y_3^+	467.9	357.3
			MH_2^{2+} -> y_5^+	467.9	569.4
		TGLQEVEVK	MH_2^{2+} -> y_6^+	501.8	603.3
			MH_2^{2+} -> y_7^+	501.8	731.4
Fibrinogen Alpha Chain	FGA	NSLFEYQK	MH_2^{2+} -> b_3^+	514.8	315.2
			MH_2^{2+} -> y_5^+	514.8	714.4
		HPDEAAFFDTASTGK	MH_3^{3+} -> y_7^+	531.9	621.3
			MH_3^{3+} -> y_3^+	531.9	679.3
Fibrinogen Beta Chain	FGB	AHYGGFTVQNEANK	MH_2^{2+} -> y_5^+	512.6	575.3
			MH_2^{2+} -> y_6^+	512.6	703.3
		HGTDDGVVWMNWK	MH_3^{3+} -> b_7^+	515.8	682.3
			MH_3^{3+} -> y_6^+	515.8	432.2
Fibrinogen Gamma Chain	FGG	YEASILTHDSSIR	MH_3^{3+} -> b_5^+	497.9	564.8
			MH_3^{3+} -> y_{11}^{2+}	497.9	600.3
		IHLISTQSAIPYALR	MH_3^{3+} -> b_3^+	561.6	364.2
			MH_3^{3+} -> y_5^+	561.6	619.4

Alpha-1-acid glycoprotein	ORM1	YVGGQEHFAHLLILR	MH_3^{3+} -> y_{13}^{2+}	584.9	745.9
			MH_3^{3+} -> y_{14}^{2+}	584.9	795.5
		NWGLSVYADKPETTK	MH_3^{3+} -> y_{11}^{2+}	570.3	619.8
			MH_3^{3+} -> y_{13}^{2+}	570.3	704.9
Alpha-1-Antitrypsin	SERPIN A1	LSITGTYDLK	MH_2^{2+} -> y_6^{+}	555.8	696.4
			MH_2^{2+} -> y_7^{+}	555.8	797.5
		AVLTIDEK	MH_2^{2+} -> y_5^{+}	444.7	605.3
			MH_2^{2+} -> y_6^{+}	444.7	718.4
Transthyretin	TTR	GSPAINVAVHVFR	MH_3^{3+} -> y_{11}^{2+}	456.3	611.9
			MH_3^{3+} -> y_{13}^{2+}	456.3	408.3
		AADDTWEPFASGK	MH_2^{2+} -> y_6^{+}	697.8	606.3
			MH_2^{2+} -> y_8^{+}	697.8	921.4

Supplementary Table S3. Exercise effect for proteins identified from iTRAQ experiments.

	p-value	adj. p-value	MD
Complement-activating component of Ra-reactive factor precursor	0.002	0.337	-0.14
Isoform 1 of Sex hormone-binding globulin	0.003	0.337	0.34
Cartilage oligomeric matrix protein	0.013	0.553	-0.11
Isoform 2 of Inter-alpha-trypsin inhibitor heavy chain H4	0.016	0.553	0.10
Cathepsin D	0.017	0.553	-0.14
Isoform 1 of CD166 antigen	0.017	0.553	-0.13
Carboxypeptidase N subunit 2 precursor	0.018	0.553	-0.09
Apolipoprotein B-100	0.027	0.553	-0.11
Isoform 1 of Pregnancy zone protein	0.029	0.553	0.22
Alpha-amylase 2B	0.029	0.553	-0.29
Ig kappa chain V-IV region	0.030	0.553	-0.37
Isoform 2 of Vascular non-inflammatory molecule 3	0.030	0.553	-0.16
Immunoglobulin superfamily containing leucine-rich repeat protein precursor	0.031	0.553	-0.13
Complement C5 precursor	0.036	0.557	-0.18
Cadherin-13 precursor	0.040	0.557	-0.25
Ig lambda chain V-I region NIG-64	0.041	0.557	-0.34
Ig kappa chain V-I region CAR	0.041	0.557	-0.25
Ig lambda chain V-IV region Hil	0.043	0.557	-0.33
Muscle type neuropilin 1	0.050	0.576	0.09

Isoform 1 of Coagulation factor XI	0.056	0.576	-0.07
ADP-ribosyl cyclase 2 precursor	0.057	0.576	-0.07
Complement C5 precursor	0.057	0.576	-0.12
Cholinesterase precursor	0.060	0.576	-0.08
Isoform 1 of Ectonucleotide pyrophosphatase/phosphodiesterase family member 2	0.061	0.576	-0.13
apolipoprotein A-IV precursor	0.063	0.576	0.72
cDNA FLJ55673, highly similar to Complement factor B	0.065	0.576	-0.05
Monocyte differentiation antigen CD14 precursor	0.069	0.576	0.07
72 kDa type IV collagenase	0.071	0.576	0.08
Coagulation factor X precursor	0.073	0.576	-0.07
Alpha-2-macroglobulin precursor	0.076	0.576	0.19
Dopamine beta-hydroxylase	0.077	0.576	-0.11
Isoform 1 of Vitamin K-dependent protein Z precursor	0.081	0.579	-0.10
Alpha-2-macroglobulin precursor	0.082	0.579	0.16
Isoform 1 of Contactin-1 precursor	0.089	0.585	-0.08
Carboxypeptidase N catalytic chain precursor	0.094	0.585	-0.18
Inter-alpha-trypsin inhibitor heavy chain H2	0.094	0.585	-0.08
Ig kappa chain V-I region AU	0.097	0.585	-0.17
Complement C4-A	0.098	0.585	-0.09
Isoform 1 of Phosphatidylinositol-glycan-specific phospholipase D precursor	0.098	0.585	-0.10
AMBP protein precursor	0.101	0.585	-0.10
Reticulon-4 receptor-like 2 precursor	0.108	0.601	-0.19
Apolipoprotein B-100 precursor	0.109	0.601	-0.08
Isoform 2 of Collagen alpha-1(XVIII) chain precursor	0.113	0.605	0.08
Membrane copper amine oxidase	0.115	0.605	-0.11
tropomyosin 1 alpha chain isoform 2	0.124	0.637	0.31
Ig mu heavy chain disease protein	0.126	0.637	-0.33
Procollagen C-endopeptidase enhancer 1	0.137	0.678	-0.07
Complement component 6 precursor	0.144	0.683	-0.06
Retinoic acid receptor responder protein 2 precursor	0.144	0.683	-0.06
Cholinesterase precursor	0.176	0.789	-0.07
Coagulation factor XII precursor	0.181	0.789	-0.10
Vitamin D-binding protein precursor	0.181	0.789	-0.05

immunoglobulin J chain	0.183	0.789	-0.19
similar to complement component 3	0.186	0.789	-0.18
Glutathione peroxidase 3 precursor	0.191	0.789	-0.09
SPARC-like protein 1	0.199	0.789	-0.09
Hepatocyte growth factor activator precursor	0.203	0.789	-0.04
Leucine-rich alpha-2-glycoprotein precursor	0.205	0.789	-0.11
Transferrin receptor protein 1	0.205	0.789	-0.07
Isoform 1 of Collagen alpha-3(VI) chain	0.206	0.789	-0.05
Complement C4-A	0.215	0.789	-0.08
Xaa-Pro dipeptidase	0.216	0.789	-0.13
Fructose-bisphosphate aldolase B	0.222	0.789	-0.12
Vitamin K-dependent protein S	0.230	0.789	0.05
Alpha-1-antitrypsin	0.234	0.789	0.09
Isoform 1 of Sulfhydryl oxidase 1 precursor	0.236	0.789	-0.05
Uncharacterized protein FETUB	0.237	0.789	-0.10
Complement component C7	0.251	0.789	-0.03
Intercellular adhesion molecule 1	0.256	0.789	-0.05
Isoform 2 of Multiple inositol polyphosphate phosphatase 1	0.258	0.789	0.04
Isoform Gamma-B of Fibrinogen gamma chain	0.260	0.789	-0.17
Clusterin precursor	0.262	0.789	0.08
Transforming growth factor-beta-induced protein ig-h3 precursor	0.263	0.789	-0.06
Vitamin K-dependent protein C	0.265	0.789	0.06
Apolipoprotein A-II precursor	0.274	0.789	-0.04
Insulin-like growth factor-binding protein 2	0.274	0.789	0.10
Serum paraoxonase/arylesterase 1	0.281	0.789	-0.07
Transforming growth factor-beta-induced protein ig-h3	0.288	0.789	-0.05
Beta-Ala-His dipeptidase	0.288	0.789	-0.07
Procollagen C-endopeptidase enhancer 1 precursor	0.290	0.789	-0.05
Histidine-rich glycoprotein precursor	0.291	0.789	0.04
Plasma glutamate carboxypeptidase	0.292	0.789	-0.09
Isoform 3 of Neural cell adhesion molecule 1	0.296	0.789	-0.05
Follistatin-related protein 1	0.296	0.789	0.07
Isoform 1 of Low affinity immunoglobulin gamma Fc region receptor II-a	0.302	0.789	0.05

Isoform 1 of Vascular cell adhesion protein 1 precursor	0.304	0.789	0.05
Isoform 1 of Extracellular matrix protein 1	0.305	0.789	0.07
Vitronectin precursor	0.306	0.789	0.07
Dopamine beta-hydroxylase	0.307	0.789	-0.08
Fibrinogen beta chain precursor	0.311	0.789	-0.21
Corticosteroid-binding globulin precursor	0.314	0.789	-0.04
Hemopexin precursor	0.316	0.789	-0.07
Thrombospondin-4 precursor	0.320	0.789	-0.08
Uncharacterized protein KLKB1	0.325	0.789	-0.04
Angiotensinogen	0.326	0.789	0.05
Fc-gamma receptor IIIb	0.326	0.789	-0.04
CD5 antigen-like precursor	0.335	0.800	-0.06
Ig lambda chain V-I region NIG-64	0.345	0.808	-0.09
Pantetheinase precursor	0.347	0.808	-0.05
Plastin-2	0.349	0.808	0.06
Isoform 1 of Fibronectin	0.353	0.808	0.13
Isoform 2 of Neural cell adhesion molecule L1-like protein	0.358	0.808	0.04
Isoform 1 of Mannan-binding lectin serine protease 2 precursor	0.365	0.808	-0.10
Coagulation factor IX	0.370	0.808	0.05
alpha-2-glycoprotein 1, zinc	0.370	0.808	0.05
Complement factor I	0.371	0.808	-0.04
Ceruloplasmin	0.372	0.808	-0.06
Isoform XB of Tenascin-X	0.382	0.820	-0.05
GUGU beta form	0.398	0.847	-0.05
Afamin precursor	0.410	0.858	-0.05
Phosphatidylcholine-sterol acyltransferase precursor	0.412	0.858	-0.02
Hepatocyte growth factor-like protein	0.417	0.858	-0.02
Fibrinogen beta chain precursor	0.418	0.858	-0.15
Coagulation factor X	0.423	0.860	-0.06
Serum amyloid P-component precursor	0.429	0.865	0.08
von Willebrand factor	0.437	0.868	0.03
Endothelial protein C receptor precursor	0.438	0.868	0.03
Isoform HMW of Kininogen-1	0.448	0.881	-0.03
Corticosteroid-binding globulin precursor	0.463	0.886	0.03
Serotransferrin precursor	0.463	0.886	-0.07
Gamma-glutamyl hydrolase precursor	0.468	0.886	-0.02

Coagulation factor IX precursor	0.472	0.886	0.04
alpha-2-glycoprotein 1, zinc	0.473	0.886	0.02
Complement C5 precursor	0.474	0.886	-0.07
Isoform LAMP-2A of Lysosome-associated membrane glycoprotein 2	0.482	0.890	-0.03
Insulin-like growth factor-binding protein complex acid labile chain	0.483	0.890	-0.03
cDNA FLJ55606, highly similar to Alpha-2-HS-glycoprotein	0.493	0.901	-0.07
Plasma serine protease inhibitor precursor	0.510	0.918	-0.05
Serotransferrin	0.511	0.918	-0.07
Prothrombin (Fragment)	0.517	0.919	-0.05
Prothrombin precursor (Fragment)	0.521	0.919	-0.04
Lumican precursor	0.523	0.919	-0.03
Pigment epithelium-derived factor precursor	0.534	0.932	-0.06
Peroxiredoxin-1	0.541	0.936	0.08
Apolipoprotein C-I precursor	0.546	0.936	0.05
Tetranectin precursor	0.556	0.936	-0.03
Complement component C7	0.558	0.936	-0.05
Plasma protease C1 inhibitor	0.559	0.936	0.03
Selenoprotein P	0.562	0.936	0.02
Complement component C9 precursor	0.566	0.936	0.05
Reticulon-4 receptor-like 2 precursor	0.573	0.936	-0.02
Complement component C8 alpha chain precursor	0.574	0.936	-0.02
Pigment epithelium-derived factor precursor	0.577	0.936	-0.04
Isoform 1 of Gelsolin precursor	0.583	0.938	0.04
Basement membrane-specific heparan sulfate proteoglycan core protein	0.586	0.938	-0.03
Coagulation factor XIII A chain	0.592	0.941	0.04
Isoform 1 of Fibrinogen alpha chain	0.602	0.942	-0.08
Ribonuclease pancreatic precursor	0.606	0.942	-0.03
Coagulation factor V	0.608	0.942	0.04
Hemopexin	0.617	0.942	-0.04
Isoform 1 of C-reactive protein	0.618	0.942	0.13
Isoform 1 of Isocitrate dehydrogenase [NAD] subunit alpha, mitochondrial	0.623	0.942	0.04
Isoform 2 of Carboxypeptidase B2	0.626	0.942	0.02
cDNA FLJ55673, highly similar to Complement factor B	0.628	0.942	-0.01

Apolipoprotein A-IV precursor	0.629	0.942	-0.07
Protein AMBP	0.636	0.943	-0.04
Isoform 3 of Mannan-binding lectin serine protease 1	0.642	0.943	-0.01
Cadherin-5	0.644	0.943	0.02
Plastin-2	0.648	0.943	0.02
Complement C1s subcomponent	0.650	0.943	-0.02
Mannose-binding protein C precursor	0.661	0.944	-0.03
Isoform 1 of Peptidase inhibitor 16 precursor	0.662	0.944	-0.02
Alpha-1-acid glycoprotein 2	0.666	0.944	-0.07
Isoform 1 of N-acetylmuramoyl-L-alanine amidase precursor	0.670	0.944	0.07
Biotinidase	0.677	0.944	-0.01
4F2 cell-surface antigen heavy chain	0.686	0.944	-0.02
Leucine-rich alpha-2-glycoprotein precursor	0.693	0.944	0.03
Isoform 1 of Multiple inositol polyphosphate phosphatase 1 precursor	0.696	0.944	-0.02
Isoform 3 of Interleukin-1 receptor accessory protein	0.703	0.944	0.01
Isoform 1 of Carboxypeptidase B2 precursor	0.707	0.944	-0.02
Properdin precursor	0.709	0.944	-0.02
Cystatin-C precursor	0.712	0.944	0.03
Ceruloplasmin precursor	0.713	0.944	-0.03
Serpin peptidase inhibitor, clade D (Heparin cofactor), member 1	0.716	0.944	-0.05
Putative uncharacterized protein ALB	0.719	0.944	-0.21
MAN1A1 protein	0.720	0.944	-0.03
Thyroxine-binding globulin precursor	0.720	0.944	0.02
HP protein	0.727	0.944	-0.02
Complement C1s subcomponent	0.729	0.944	0.03
Coagulation factor XIII B chain precursor	0.733	0.945	0.02
Plasma serine protease inhibitor	0.740	0.948	-0.02
Insulin-like growth factor IA	0.751	0.949	-0.04
Galectin-3-binding protein	0.752	0.949	-0.02
Lysozyme C precursor	0.754	0.949	-0.01
Mannosyl-oligosaccharide 1,2-alpha-mannosidase IA	0.758	0.949	-0.03
Isoform 1 of Inter-alpha-trypsin inhibitor heavy chain H3	0.761	0.949	0.01

Kallistatin precursor	0.767	0.951	-0.02
Protein Z-dependent protease inhibitor precursor	0.772	0.951	-0.01
Apolipoprotein A-I precursor	0.775	0.951	-0.03
Isoform 1 of Cartilage acidic protein 1 precursor	0.798	0.959	-0.01
Serum paraoxonase/arylesterase 1	0.800	0.959	0.02
Isoform 1 of Attractin	0.801	0.959	-0.01
Aminopeptidase N	0.805	0.959	-0.01
Insulin-like growth factor-binding protein 3	0.807	0.959	-0.01
Inter-alpha-trypsin inhibitor heavy chain H1 precursor	0.813	0.959	0.01
30 kDa protein	0.816	0.959	0.01
Alpha-1-acid glycoprotein 2 precursor	0.816	0.959	-0.03
HSPA5 protein	0.823	0.959	-0.01
Complement factor D preproprotein	0.828	0.959	-0.01
Isoform A of Coagulation factor VII	0.829	0.959	0.01
Carboxypeptidase N catalytic chain precursor	0.831	0.959	-0.02
Complement component C1q receptor	0.848	0.971	-0.01
Hemoglobin subunit epsilon	0.850	0.971	-0.03
Apolipoprotein C-III precursor	0.854	0.971	-0.02
45 kDa protein	0.865	0.976	0.01
Apolipoprotein E	0.872	0.976	-0.01
Complement C1r subcomponent precursor	0.878	0.976	-0.01
Vasorin precursor	0.881	0.976	0.00
Complement component C8 beta chain precursor	0.882	0.976	-0.02
Angiogenin precursor	0.887	0.976	-0.01
Apolipoprotein A-I precursor	0.889	0.976	-0.02
Antithrombin III variant	0.894	0.976	-0.01
Carbonic anhydrase 1	0.903	0.976	0.02
Cadherin-2	0.904	0.976	0.01
Isoform 1 of Vinculin	0.906	0.976	-0.01
Complement C1r subcomponent-like protein	0.909	0.976	0.01
Insulin-like growth factor-binding protein 5	0.920	0.976	-0.01
Isoform 1 of Phosphatidylinositol-glycan-specific phospholipase D precursor	0.920	0.976	0.01
Protein AMBP	0.921	0.976	0.01
Flavin reductase	0.937	0.978	0.01

Insulin-like growth factor-binding protein 6 precursor	0.937	0.978	0.00
Apolipoprotein A-IV precursor	0.937	0.978	-0.01
Isoform 2 of Inter-alpha-trypsin inhibitor heavy chain H4	0.940	0.978	0.01
Lumican precursor	0.951	0.985	0.00
Isoform 1 of Insulin-like growth factor II	0.965	0.986	0.00
Isoform B of Fibulin-1	0.965	0.986	0.00
Insulin-like growth factor-binding protein 4 precursor	0.972	0.986	0.00
Apolipoprotein C-II precursor	0.972	0.986	0.00
Alpha-1B-glycoprotein	0.973	0.986	0.00
Isoform 1 of EGF-containing fibulin-like extracellular matrix protein 1	0.982	0.987	0.00
Isoform 1 of Cell surface glycoprotein MUC18 precursor	0.987	0.987	0.00
Angiotensinogen precursor	0.987	0.987	0.00

MD = mean difference (mean of concentrations for group with exercise minus mean for group without exercise)

Supplementary Table S4. VLCD effect for proteins identified from iTRAQ experiments.

	p-value	Adj. p-value	MD
Biotinidase	2.59E-09	6.00E-07	-0.10
Selenoprotein P	1.56E-08	1.02E-06	0.15
Insulin-like growth factor-binding protein 2	1.73E-08	1.02E-06	0.36
Isoform 2 of Inter-alpha-trypsin inhibitor heavy chain H4	1.76E-08	1.02E-06	0.18
Isoform 1 of Sex hormone-binding globulin	3.58E-08	1.66E-06	0.46
Isoform 3 of Interleukin-1 receptor accessory protein	5.27E-08	2.04E-06	0.13
Afamin precursor	8.02E-07	2.61E-05	-0.18
Apolipoprotein A-IV precursor	9.00E-07	2.61E-05	-0.45
Leucine-rich alpha-2-glycoprotein precursor	1.04E-06	2.68E-05	0.20
Beta-Ala-His dipeptidase	1.44E-06	3.35E-05	-0.21
Leucine-rich alpha-2-glycoprotein precursor	1.61E-06	3.39E-05	0.25
Apolipoprotein A-IV precursor	3.03E-06	5.86E-05	-0.43
Lysozyme C precursor	3.67E-06	6.55E-05	0.11
Pigment epithelium-derived factor precursor	4.10E-06	6.80E-05	-0.27
Fructose-bisphosphate aldolase B	1.60E-05	2.48E-04	-0.25
Cholinesterase precursor	2.30E-05	3.34E-04	-0.13
Pigment epithelium-derived factor precursor	2.45E-05	3.35E-04	-0.20
Cathepsin D	3.09E-05	3.99E-04	-0.15
Protein AMBP	3.29E-05	4.02E-04	-0.13
Aminopeptidase N	1.01E-04	1.17E-03	0.08
Isoform 2 of Neural cell adhesion molecule L1-like protein	1.34E-04	1.48E-03	0.09
Complement component C9 precursor	1.80E-04	1.90E-03	0.18
Isoform 1 of EGF-containing fibulin-like extracellular matrix protein 1	2.26E-04	2.23E-03	0.10
Isoform 3 of Mannan-binding lectin serine protease 1	2.30E-04	2.23E-03	-0.06
Coagulation factor X precursor	2.62E-04	2.35E-03	-0.08
Isoform 1 of Attractin	2.63E-04	2.35E-03	-0.06
Monocyte differentiation antigen CD14 precursor	3.02E-04	2.53E-03	0.08
Isoform 1 of Phosphatidylinositol-glycan-specific phospholipase D precursor	3.06E-04	2.53E-03	-0.13
Complement component C8 alpha chain precursor	3.67E-04	2.94E-03	0.07

Alpha-1-antitrypsin	3.90E-04	3.01E-03	0.15
alpha-2-glycoprotein 1, zinc	4.25E-04	3.18E-03	0.07
Pantetheinase precursor	4.88E-04	3.54E-03	-0.10
Coagulation factor X	5.18E-04	3.64E-03	-0.13
Isoform 1 of Contactin-1 precursor	6.64E-04	4.53E-03	0.09
Isoform 2 of Vascular non-inflammatory molecule 3	7.75E-04	5.14E-03	-0.15
Isoform 1 of Vascular cell adhesion protein 1 precursor	1.02E-03	6.56E-03	0.08
Isoform 1 of Inter-alpha-trypsin inhibitor heavy chain H3	1.13E-03	7.06E-03	0.07
Protein Z-dependent protease inhibitor precursor	1.20E-03	7.31E-03	-0.06
Transforming growth factor-beta-induced protein ig-h3	1.28E-03	7.61E-03	-0.08
Plastin-2	1.45E-03	8.42E-03	0.08
Isoform 1 of Isocitrate dehydrogenase [NAD] subunit alpha, mitochondrial	1.54E-03	8.73E-03	0.15
Isoform 2 of Inter-alpha-trypsin inhibitor heavy chain H4	1.64E-03	9.04E-03	0.15
Cartilage oligomeric matrix protein	2.00E-03	1.08E-02	-0.08
Vitamin K-dependent protein C	2.20E-03	1.16E-02	0.09
Apolipoprotein C-III precursor	2.74E-03	1.41E-02	-0.17
Plasma serine protease inhibitor precursor	2.99E-03	1.51E-02	-0.13
Alpha-2-macroglobulin precursor	3.25E-03	1.59E-02	0.15
Corticosteroid-binding globulin precursor	3.30E-03	1.59E-02	0.07
Cystatin-C precursor	3.45E-03	1.61E-02	0.12
Insulin-like growth factor-binding protein 6 precursor	3.47E-03	1.61E-02	-0.07
Corticosteroid-binding globulin precursor	4.06E-03	1.85E-02	0.07
Isoform 1 of Pregnancy zone protein	4.76E-03	2.13E-02	0.16
Plastin-2	5.27E-03	2.31E-02	0.09
Retinoic acid receptor responder protein 2 precursor	5.60E-03	2.40E-02	-0.07
HSPA5 protein	5.98E-03	2.46E-02	0.04
30 kDa protein	6.03E-03	2.46E-02	0.07
Serum amyloid P-component precursor	6.11E-03	2.46E-02	-0.14
Isoform 1 of Cartilage acidic protein 1 precursor	6.16E-03	2.46E-02	0.05
Isoform 1 of Low affinity immunoglobulin gamma Fc region receptor II-a	6.38E-03	2.51E-02	-0.06
Antithrombin III variant	6.63E-03	2.56E-02	0.11
Isoform 1 of Vinculin	6.78E-03	2.58E-02	0.10

Fc-gamma receptor IIIb	6.94E-03	2.60E-02	-0.06
Coagulation factor XII precursor	7.45E-03	2.74E-02	-0.11
Muscle type neuropilin 1	8.98E-03	3.25E-02	0.07
72 kDa type IV collagenase	9.38E-03	3.35E-02	0.06
Clusterin precursor	1.07E-02	3.77E-02	0.10
Alpha-2-macroglobulin precursor	1.28E-02	4.42E-02	0.15
Complement factor I	1.34E-02	4.58E-02	-0.06
SPARC-like protein 1	1.39E-02	4.68E-02	0.10
Complement component C1q receptor	1.55E-02	5.06E-02	0.05
Isoform 1 of Coagulation factor XI	1.55E-02	5.06E-02	-0.05
apolipoprotein A-IV precursor	1.57E-02	5.06E-02	-0.51
Histidine-rich glycoprotein precursor	1.75E-02	5.57E-02	0.05
Isoform 1 of Sulfhydryl oxidase 1 precursor	1.88E-02	5.90E-02	-0.05
Isoform 1 of C-reactive protein	1.95E-02	6.02E-02	-0.33
Isoform 1 of Ectonucleotide pyrophosphatase/phosphodiesterase family member 2	1.97E-02	6.02E-02	-0.09
Complement-activating component of Ra-reactive factor precursor	2.00E-02	6.03E-02	-0.06
Ceruloplasmin	2.39E-02	7.11E-02	0.09
Intercellular adhesion molecule 1	2.43E-02	7.12E-02	-0.06
Xaa-Pro dipeptidase	2.45E-02	7.12E-02	-0.13
Complement C1s subcomponent	2.54E-02	7.21E-02	-0.10
Apolipoprotein B-100	2.55E-02	7.21E-02	0.06
Tetranectin precursor	2.98E-02	8.24E-02	0.06
Isoform 1 of Vitamin K-dependent protein Z precursor	2.98E-02	8.24E-02	-0.06
Cadherin-13 precursor	3.18E-02	8.68E-02	0.14
Isoform 1 of Phosphatidylinositol-glycan-specific phospholipase D precursor	3.25E-02	8.78E-02	-0.06
Thyroxine-binding globulin precursor	3.43E-02	9.15E-02	0.06
Isoform 1 of Insulin-like growth factor II	3.90E-02	1.03E-01	0.05
Alpha-1-acid glycoprotein 2 precursor	4.02E-02	1.05E-01	-0.14
Isoform A of Coagulation factor VII	4.51E-02	1.16E-01	-0.05
Isoform 1 of Fibrinogen alpha chain	4.68E-02	1.18E-01	0.16
Fibrinogen beta chain precursor	4.69E-02	1.18E-01	0.19
Isoform 1 of Carboxypeptidase B2 precursor	4.77E-02	1.18E-01	0.05
Uncharacterized protein KLKB1	4.80E-02	1.18E-01	-0.04
Membrane copper amine oxidase	4.98E-02	1.22E-01	-0.07
Apolipoprotein C-I precursor	5.17E-02	1.25E-01	-0.08

Putative uncharacterized protein ALB	5.60E-02	1.34E-01	0.58
Basement membrane-specific heparan sulfate proteoglycan core protein	5.89E-02	1.39E-01	0.05
Ceruloplasmin precursor	5.92E-02	1.39E-01	0.08
Vasorin precursor	6.03E-02	1.40E-01	0.03
HP protein	6.34E-02	1.46E-01	-0.04
cDNA FLJ55673, highly similar to Complement factor B	6.80E-02	1.55E-01	-0.02
Cholinesterase precursor	7.70E-02	1.73E-01	-0.04
Insulin-like growth factor IA	8.08E-02	1.80E-01	0.11
Vitamin D-binding protein precursor	8.85E-02	1.96E-01	0.03
alpha-2-glycoprotein 1, zinc	9.34E-02	2.03E-01	0.05
Phosphatidylcholine-sterol acyltransferase precursor	9.38E-02	2.03E-01	0.03
Isoform 1 of CD166 antigen	1.07E-01	2.30E-01	0.05
Plasma protease C1 inhibitor	1.13E-01	2.40E-01	0.04
Ig lambda chain V-I region NIG-64	1.17E-01	2.46E-01	-0.14
Ig lambda chain V-IV region Hil	1.18E-01	2.46E-01	-0.14
Complement C1s subcomponent	1.22E-01	2.51E-01	-0.03
Complement C5 precursor	1.22E-01	2.51E-01	-0.08
Isoform 1 of Mannan-binding lectin serine protease 2 precursor	1.40E-01	2.85E-01	-0.08
Alpha-1-acid glycoprotein 2	1.46E-01	2.94E-01	-0.13
Prothrombin precursor (Fragment)	1.48E-01	2.94E-01	-0.04
Complement component 6 precursor	1.49E-01	2.94E-01	0.03
Plasma serine protease inhibitor	1.49E-01	2.94E-01	-0.04
Hepatocyte growth factor activator precursor	1.53E-01	2.99E-01	-0.02
Hemoglobin subunit epsilon	1.56E-01	3.03E-01	0.11
ADP-ribosyl cyclase 2 precursor	1.59E-01	3.04E-01	0.03
Hemopexin precursor	1.64E-01	3.11E-01	-0.05
Gamma-glutamyl hydrolase precursor	1.69E-01	3.19E-01	-0.02
Isoform Gamma-B of Fibrinogen gamma chain	1.80E-01	3.37E-01	0.11
Thrombospondin-4 precursor	1.88E-01	3.50E-01	-0.05
Carbonic anhydrase 1	1.96E-01	3.62E-01	0.13
Angiotensinogen precursor	1.99E-01	3.64E-01	-0.05
Cadherin-2	2.02E-01	3.66E-01	0.03
Ig kappa chain V-IV region	2.07E-01	3.70E-01	-0.11
Coagulation factor IX	2.07E-01	3.70E-01	0.03
Isoform 1 of N-acetylmuramoyl-L-alanine amidase precursor	2.15E-01	3.81E-01	-0.10

Isoform 1 of Fibronectin	2.19E-01	3.84E-01	-0.09
Serum paraoxonase/arylesterase 1	2.20E-01	3.84E-01	0.04
MAN1A1 protein	2.34E-01	4.03E-01	-0.06
Complement C5 precursor	2.35E-01	4.03E-01	-0.04
Kallistatin precursor	2.43E-01	4.14E-01	-0.04
45 kDa protein	2.46E-01	4.14E-01	0.03
Glutathione peroxidase 3 precursor	2.46E-01	4.14E-01	0.04
Serpin peptidase inhibitor, clade D (Heparin cofactor), member 1	2.48E-01	4.15E-01	-0.08
tropomyosin 1 alpha chain isoform 2	2.56E-01	4.25E-01	0.12
Isoform 2 of Multiple inositol polyphosphate phosphatase 1	2.76E-01	4.45E-01	-0.02
Isoform 2 of Collagen alpha-1(XVIII) chain precursor	2.77E-01	4.45E-01	-0.03
Apolipoprotein A-I precursor	2.78E-01	4.45E-01	0.05
Apolipoprotein A-II precursor	2.78E-01	4.45E-01	0.02
Vitamin K-dependent protein S	2.78E-01	4.45E-01	0.02
Inter-alpha-trypsin inhibitor heavy chain H1 precursor	2.85E-01	4.54E-01	0.02
Complement C1r subcomponent-like protein	2.90E-01	4.57E-01	-0.03
Carboxypeptidase N subunit 2 precursor	2.94E-01	4.62E-01	-0.02
Inter-alpha-trypsin inhibitor heavy chain H2	3.00E-01	4.63E-01	0.02
Carboxypeptidase N catalytic chain precursor	3.00E-01	4.63E-01	-0.06
Apolipoprotein A-I precursor	3.15E-01	4.85E-01	0.06
Apolipoprotein C-II precursor	3.41E-01	5.20E-01	-0.04
Isoform 1 of Cell surface glycoprotein MUC18 precursor	3.48E-01	5.26E-01	-0.04
Isoform HMW of Kininogen-1	3.49E-01	5.26E-01	0.02
Complement component C7	3.53E-01	5.28E-01	-0.04
Dopamine beta-hydroxylase	3.74E-01	5.57E-01	-0.03
Complement component C7	3.90E-01	5.77E-01	-0.01
Fibrinogen beta chain precursor	3.95E-01	5.80E-01	0.09
Peroxiredoxin-1	4.06E-01	5.92E-01	0.06
Hemopexin	4.09E-01	5.94E-01	-0.03
Flavin reductase	4.20E-01	6.05E-01	0.06
Apolipoprotein B-100 precursor	4.28E-01	6.08E-01	0.02
cDNA FLJ55673, highly similar to Complement factor B	4.32E-01	6.08E-01	-0.01
Isoform 1 of Extracellular matrix protein 1	4.32E-01	6.08E-01	-0.03
Isoform 1 of Peptidase inhibitor 16 precursor	4.33E-01	6.08E-01	-0.02
Prothrombin (Fragment)	4.43E-01	6.20E-01	-0.03

CD5 antigen-like precursor	4.55E-01	6.32E-01	0.02
Coagulation factor XIII B chain precursor	4.61E-01	6.36E-01	-0.02
Isoform 1 of Collagen alpha-3(VI) chain	4.67E-01	6.41E-01	0.02
Complement C1r subcomponent precursor	4.73E-01	6.42E-01	-0.02
Carboxypeptidase N catalytic chain precursor	4.74E-01	6.42E-01	-0.03
Isoform XB of Tenascin-X	4.76E-01	6.42E-01	-0.02
Lumican precursor	4.89E-01	6.55E-01	0.02
Angiogenin precursor	5.06E-01	6.74E-01	-0.02
Uncharacterized protein FETUB	5.12E-01	6.78E-01	0.03
Isoform 3 of Neural cell adhesion molecule 1	5.47E-01	7.19E-01	0.01
Isoform 1 of Gelsolin precursor	5.49E-01	7.19E-01	-0.02
4F2 cell-surface antigen heavy chain	5.67E-01	7.39E-01	0.02
Coagulation factor V	5.80E-01	7.50E-01	-0.02
Complement component C8 beta chain precursor	5.83E-01	7.50E-01	-0.04
Mannose-binding protein C precursor	5.85E-01	7.50E-01	0.02
Follistatin-related protein 1	5.98E-01	7.62E-01	0.02
Galectin-3-binding protein	6.09E-01	7.70E-01	-0.02
Angiotensinogen	6.11E-01	7.70E-01	-0.01
Serum paraoxonase/arylesterase 1	6.15E-01	7.71E-01	0.02
Ig kappa chain V-I region CAR	6.30E-01	7.86E-01	-0.03
Complement factor D preproprotein	6.38E-01	7.92E-01	0.01
Dopamine beta-hydroxylase	6.64E-01	8.14E-01	-0.02
Lumican precursor	6.64E-01	8.14E-01	0.01
Alpha-amylase 2B	6.67E-01	8.14E-01	0.03
Vitronectin precursor	6.70E-01	8.14E-01	-0.02
Ig kappa chain V-I region AU	6.75E-01	8.16E-01	-0.02
cDNA FLJ55606, highly similar to Alpha-2-HS-glycoprotein	6.87E-01	8.23E-01	0.02
AMBP protein precursor	6.88E-01	8.23E-01	0.01
Ig lambda chain V-I region NIG-64	6.91E-01	8.23E-01	-0.02
Properdin precursor	7.20E-01	8.50E-01	-0.01
Insulin-like growth factor-binding protein 3	7.22E-01	8.50E-01	-0.01
Isoform LAMP-2A of Lysosome-associated membrane glycoprotein 2	7.26E-01	8.50E-01	0.01
Isoform 2 of Carboxypeptidase B2	7.42E-01	8.65E-01	0.01
similar to complement component 3	7.47E-01	8.66E-01	-0.02
Transforming growth factor-beta-induced protein ig-h3 precursor	7.59E-01	8.72E-01	-0.01
Coagulation factor XIII A chain	7.59E-01	8.72E-01	0.01

Cadherin-5	7.72E-01	8.82E-01	-0.01
Ig mu heavy chain disease protein	7.84E-01	8.86E-01	-0.03
Plasma glutamate carboxypeptidase	7.85E-01	8.86E-01	0.01
Serotransferrin	7.87E-01	8.86E-01	-0.01
Procollagen C-endopeptidase enhancer 1	7.91E-01	8.86E-01	0.01
Apolipoprotein E	8.27E-01	9.22E-01	0.01
Complement C4-A	8.30E-01	9.22E-01	0.01
Procollagen C-endopeptidase enhancer 1 precursor	8.37E-01	9.25E-01	-0.01
immunoglobulin J chain	8.48E-01	9.33E-01	0.01
Ribonuclease pancreatic precursor	8.58E-01	9.39E-01	0.00
Protein AMBP	8.76E-01	9.50E-01	-0.01
Insulin-like growth factor-binding protein 4 precursor	8.76E-01	9.50E-01	0.01
Reticulon-4 receptor-like 2 precursor	8.92E-01	9.58E-01	0.01
Mannosyl-oligosaccharide 1,2-alpha-mannosidase IA	8.94E-01	9.58E-01	0.01
Complement C4-A	8.96E-01	9.58E-01	0.00
Alpha-1B-glycoprotein	9.01E-01	9.59E-01	0.00
Complement C5 precursor	9.19E-01	9.74E-01	0.00
Serotransferrin precursor	9.27E-01	9.77E-01	0.00
von Willebrand factor	9.34E-01	9.80E-01	0.00
Reticulon-4 receptor-like 2 precursor	9.41E-01	9.83E-01	0.00
Hepatocyte growth factor-like protein	9.53E-01	9.91E-01	0.00
Transferrin receptor protein 1	9.66E-01	9.93E-01	0.00
Immunoglobulin superfamily containing leucine-rich repeat protein precursor	9.68E-01	9.93E-01	0.00
Insulin-like growth factor-binding protein 5	9.71E-01	9.93E-01	0.00
Endothelial protein C receptor precursor	9.71E-01	9.93E-01	0.00
GUGU beta form	9.77E-01	9.94E-01	0.00
Coagulation factor IX precursor	9.89E-01	9.94E-01	0.00
Insulin-like growth factor-binding protein complex acid labile chain	9.90E-01	9.94E-01	0.00
Isoform B of Fibulin-1	9.90E-01	9.94E-01	0.00
Isoform 1 of Multiple inositol polyphosphate phosphatase 1 precursor	9.97E-01	9.97E-01	0.00

MD = mean difference (mean of concentrations after VLCD minus mean at baseline). Analytes are considered significant if the p-value is less than 0.05 and the adjusted p-value is less than 0.1

Chapter 7: Prolonged niacin treatment leads to increased adipose tissue PUFA synthesis and an anti-inflammatory lipid and oxylipin plasma profile

Mattijs M. Heemskerk*
Harish K. Dharuri*
Sjoerd A.A. van den Berg
Hulda S. Jónasdóttir
Dick-Paul Kloos
Martin Giera
Ko Willems van Dijk
Vanessa van Harmelen

* Both authors contributed equally

Journal of Lipid Research. 2014; 55(12):2532-40

Abstract

Prolonged niacin treatment elicits beneficial effects on the plasma lipid and lipoprotein profile that are associated with a protective cardiovascular disease (CVD) risk profile. Acute niacin treatment inhibits non-esterified fatty acid (NEFA) release from adipocytes and stimulates prostaglandin release from skin Langerhans cells, but the acute effects diminish upon prolonged treatment, while the beneficial effects remain. To gain insight in the prolonged effects of niacin on lipid metabolism in adipocytes, we used a mouse model with a human-like lipoprotein metabolism and drug response (female APOE*3-Leiden.CETP mice) treated with and without niacin for 15 weeks. The gene expression profile of gonadal white adipose tissue (gWAT) from niacin treated mice showed an up- regulation of the “biosynthesis of unsaturated fatty acid (PUFA)” pathway, which was corroborated by qPCR and analysis of the FA ratios in gWAT. Also, adipocytes from niacin treated mice secreted more of the PUFA docosahexaenoic acid (DHA) *ex vivo*. This resulted in an increased DHA/arachidonic acid (AA) ratio in the adipocyte FA secretion profile and in plasma of niacin treated mice. Interestingly, the DHA metabolite 19,20-dihydroxy docosapentaenoic acid (19,20-diHDPA) was increased in plasma of niacin treated mice. Both an increased DHA/AA ratio and increased 19,20-diHDPA are indicative for an anti-inflammatory profile and may indirectly contribute to the atheroprotective lipid and lipoprotein profile associated with prolonged niacin treatment.

Introduction

Niacin (vitamin B3) treatment reduces cardiovascular disease and atherosclerosis development [1]. These beneficial effects are mediated, in part, by lowering circulating levels of LDL-cholesterol, VLDL-TG and lipoprotein(a) [2] as well as by increasing HDL-cholesterol [3]. In addition, prolonged niacin treatment also decreases plasma, adipose tissue and vascular inflammation [4, 5], which might contribute to reducing CVD. The induction of these beneficial effects after prolonged niacin treatment are in striking contrast to the unwanted acute niacin effects.

Acutely, niacin binds to the inhibitory hydroxycarboxylic acid receptor 2 (HCA2) (previously known as GPR109A). In adipocytes this leads to an inhibition of adipocyte lipolysis followed by an acute reduction of plasma non-esterified fatty acid (NEFA) levels. Lowering NEFA levels causes metabolic stress [6, 7], which increases stress hormone levels [8–12] after niacin treatment. In the skin Langerhans cells and keratinocytes, acute niacin binding to the HCA2 receptor leads to a release of arachidonic acid (AA) and subsequent cyclooxygenase-mediated oxylipin synthesis (mostly prostaglandins) causing flushing [13] and a decrease in blood pressure [14]. Intriguingly, these acute effects decrease upon prolonged niacin treatment. Adipocyte lipolysis normalizes [15, 16] and flushing diminishes [17].

The fact that certain acute niacin effects decrease over time whereas the beneficial lipid lowering and anti-inflammatory effects remain, suggests differences between the induction of intracellular signaling pathways upon acute and prolonged niacin treatment. In the current study we set out to characterize changes in signaling regulation upon prolonged niacin treatment. We specifically investigated effects of niacin on adipose tissue as adipose tissue has been shown to be the most affected organ at the gene expression level after 7h of niacin treatment [18].

We treated mice with 0.3% niacin mixed through the diet and isolated gonadal white adipose tissue (gWAT) after 15 weeks of intervention. The mice used in this study were female APOE*3-Leiden.CETP mice [19] which-in contrast to wild type mice-have a human like lipoprotein profile and respond similarly to atheroprotective drugs like niacin [20]. A microarray was used to compare gene expression profiles in the adipose tissue. We applied bio-informatic and statistical analyses to the gene expression data and showed that prolonged niacin treatment led to an increase in the unsaturated FA synthesis pathways. To investigate whether PUFA levels and possible derivatives thereof (i.e. oxylipins) were functionally affected we determined the fatty acid (FA) composition in the adipose tissue by gas chromatography mass spectrometry (GC-MS) and measured PUFA and oxylipin profiles in plasma by liquid chromatography tandem mass spectrometry (LC-MS/MS).

Materials and methods

Mouse experiments

Female APOE*3-Leiden.CETP mice were bred at the Leiden University Medical Center. At age 15 ± 1 week, mice were fed a western type diet (Diet T with 0.1 g% cholesterol, which consisted of 16 kcal% protein, 43 kcal% carbohydrate and 41 kcal% fat. AB Diets, Woerden, the Netherlands) with or without niacin (0.3 g%, Sigma Aldrich, St Louis, MO, USA). Supplementary table SII shows the fatty acid composition of the diet. Body weight was registered weekly. Animals were housed in a controlled environment (21°C, 40- 50% humidity) with a daily 12h photoperiod (07h00-

19h00). Food and tap water were available ad libitum during the whole experiment. Food intake was determined weekly by weighing the food in the cages at t=0 and at t= 1 days. The difference between these time points was equal to 24h food intake of the mice. The mice in this study are the same as in our previously published study (16). All experiments were performed after a 15 week dietary intervention period. All animals (n=14 per group) were anaesthetized and sacrificed in the fed state between 08h00 and 9h30 by cardiac puncture. Organs and plasma were collected and stored at -80°C. Fresh gonadal white adipose tissue (gWAT) was harvested and kept in PBS with or without niacin. One niacin treated animal did not have sufficient gWAT for the analyses. All animal experiments were performed in accordance with the regulations of Dutch law on animal welfare. The institutional scientific committee and ethics committee for animal procedures from the Leiden University Medical Center, Leiden, The Netherlands approved the protocols.

gWAT gene expression analysis

RNA was isolated from gWAT using the Nucleospin RNA/Protein kit (MACHEREY-NAGEL GmbH & Co. KG, Düren, Germany) after which RNA quality was assessed by NanoDrop (NanoDrop) and 2100 BioAnalyzer (Agilent). All samples had an RNA Integrity Number of >7.5. cRNA was synthesized using the TotalPrep RNA Amplification Kit (Ambion, Illumina). cRNA levels were normalized to 150ng/µL and loaded onto MouseWG-6 v2.0 Expression BeadChips by Service XS (Leiden, The Netherlands). Each BeadChip contains eight arrays. Hybridization and washing were performed according to the Illumina manual. Image analysis and extraction of raw expression data was performed with Illumina GenomeStudio v2011.1 gene expression software with default settings.

Lumi [21] module in the R-based Bioconductor package was used to read in the combined (average) signal intensities per probe. A variance-stabilizing transformation (lumiT) available in the R package was used to stabilize the expression variance based on the bead level expression variance and mean relations. Expression data were normalized using the function lumiN available within the lumi package. We used limma [22] an R-based Bioconductor package to calculate the level of differential gene expression. In addition to determining significant differentially expressed genes, gene set analysis based on KEGG pathway and Gene Ontology was performed using the Bioconductor package "GlobalTest" [23].

Quantitative PCR

RNA was isolated from gWAT and liver using the Nucleospin RNA/Protein kit (MACHEREY- NAGEL GmbH & Co. KG, Düren, Germany). Subsequently, 1µg of RNA was used for cDNA synthesis by iScript (BioRad, Hercules, CA, USA), which was purified by the Nucleospin Gel and PCR clean-up kit (Machery Nagel). Real-Time PCR was carried out on the IQ5 PCR machine (BioRad) using the Sensimix SYBR Green RT-PCR mix (Quantace, London, UK) and QuantiTect SYBR Green RT-PCR mix (Qiagen, Venlo, the Netherlands). Target mRNA levels were normalized to *Rplp0* & *Ppia* mRNA levels. Primer sequences and PCR conditions can be found in Supplementary table SI.

gWAT, liver and diet fatty acid composition

FA composition analysis of gWAT, liver and diet was carried out as described recently by Kloos et al. [24]. Briefly: triplicate samples were weighed of approximately 10 mg diet or organ from niacin treated and control mice. 1 mL of water, 3 mL MeOH and 1 mL 10M NaOH were added, the samples flushed with argon and hydrolyzed for 1 h at 90 °C. After acidification with 2 mL of 6M HCl, 10 µL of an internal standard solution ([$^2H_{31}$]palmitic acid and ergosterole 10 µg/mL each) was added. The samples were extracted twice with 3 mL *n*-hexane and the combined organic extracts were dried under a gentle stream of nitrogen. Dried samples were derivatized using 25 µL of *N*-*tert*.-butyldimethylsilyl-*N*-methyltrifluoroacetamide (Sigma Aldrich, Schnelldorf, Germany) for 10 min at 21 °C, subsequently 25 µL of N,O-bis(trimethylsilyl)trifluoroacetamide containing 1% trimethylchlorosilane (Thermo Scientific, Waltham, MA, USA) and 2.5 µL of pyridine were added and the sample was heated for 15 min to 50 °C. Next, 947.5 µL of n- hexane, containing 10 µg/mL octadecane (C_{18}) as system monitoring component, was added. Samples were analyzed in SIM mode on a Scion TQ GC-MS (Bruker, Bremen, Germany) equipped with a 15 m × 0.25 mm × 0.25 mm BR5MS column (Bruker). The injection volume was 1 µL, the injector was operated in splitless mode at 280 °C and the oven program was as follows: 90 °C kept constant for 0.5 min, then ramped to 180 °C with 30 °C/min then to 250 °C with 10 °C/min then to 266 °C with 2 °C/min and finally to 300 °C with 120 °C/min, kept constant for 2 min. Helium (99.9990%, Air Products, The Netherlands) was used as carrier gas. For data analysis a total area correction was applied and triplicates were averaged.

Gonadal adipocyte PUFA release assay

Fresh gonadal adipose tissue was minced and digested in 0.5 g/L collagenase type I in HEPES buffer (pH 7.4) with 20 g/L of dialyzed bovine serum albumin

Table 1: Differentially expressed gene hits from microarray analysis of gWAT after niacin treatment.

Gene symbol	Gene ID	Gene name	Adjusted P	Log(fold change)
Pdzk1ip1	67182	PDZK1 interacting protein 1	0.002	1.190
Orm2	18406	Orosomucoid 2	0.002	0.857
Orm1	18405	Orosomucoid 1	0.003	0.550
Elovl6	170439	Elongation of long chain fatty	0.004	1.371
Lctl	235435	Lactase-like	0.004	1.107
Rdh11	17252	Retinol dehydrogenase 11	0.007	0.887
Nudt7	67528	Nudix (nucleoside diphosphate linked moiety X)- type motif 7	0.012	0.537
Acat2	110460	Acetyl-Coenzyme A	0.013	0.695
Mup3	17842	Major urinary protein 3	0.013	1.047
1500017E21Rik	668215	RIKEN cDNA 1500017E21 gene	0.013	0.612
Clstn3	232370	Calsyntenin 3	0.013	0.607
Apoc1	11812	Apolipoprotein C-I	0.013	0.523
Comt	12846	Catechol-O-methyltransferase	0.014	0.536
Zfp385b	241494	Zinc finger protein 385B	0.014	-0.436
Tecr	106529	Trans-2,3-enoyl-CoA reductase	0.029	0.498
G6pdx	14381	Glucose-6-phosphate	0.029	0.454
Elovl5	68801	Elongation of long chain fatty	0.033	0.405
Pkm2	18746	Pyruvate kinase, muscle	0.034	0.521
D430019H16Rik	268595	RIKEN cDNA D430019H16 gene	0.034	-0.505
Aacs	78894	Acetoacetyl-CoA synthetase	0.035	0.524
Lpcat3	14792	Lysophosphatidylcholine	0.035	0.504
Kcnj15	16516	Potassium inwardly-rectifying channel, subfamily J, member	0.035	0.450
Cyp51	13121	Cytochrome P450, family 51	0.039	0.747
Aard	239435	Alanine and arginine rich domain containing protein	0.039	-0.547
Fasn	14104	Fatty acid synthase	0.126	0.487
Acly	104112	ATP citrate lyase	0.139	0.691

(BSA, fraction V, Sigma Aldrich) for 1 h at 37°C. The disaggregated WAT was filtered through a nylon mesh with a pore size of 236 µm. For the isolation of mature adipocytes, cells were obtained from the surface of the filtrate and washed several times. Adipocytes (~10,000 cells/mL) were incubated in triplicate in a 96 well plate at 37°C in 200µL per well of DMEM/F12 medium with 2%w/w BSA with or without niacin 10^{-6} M) for 2 hours. The adipocyte conditioned medium (100 µL) was frozen at -20°C until further analysis.

Plasma PUFA and oxylipins measurement

Protein precipitation was performed on adipocyte conditioned medium (80 µL) or plasma (20 µL) by the addition of methanol (233.6 µL for medium and 53.6 µl for plasma) and 6.4 µL of internal standard solution containing ($[^2H_8]$15-HETE, $[^2H_4]PGE_2$, $[^2H_4]LTB_4$ and $[^2H_5]$DHA, each 50 ng/mL in methanol), which was left to equilibrate for 20 minutes at -20°C. The samples were spun down for 10 min, 16200g at 4°C. Supernatant (240 µL for medium and 30 µL for plasma) was pipetted into a deactivated glass insert (Agilent, CA, USA). Plasma supernatant was diluted in 30 µL of H_2O, while medium supernatant was dried by Speedvac at room temperature. The dried medium sample was dissolved in 60 µL 1:2 methanol/H_2O. For both sample types, 20 µL was injected for LC-MS/MS analysis as described previously [25, 26].

LC-MS/MS analysis is carried out on a QTrap 6500 mass spectrometer (AB Sciex, Nieuwerkerk aan den Ijssel, The Netherlands), coupled to a Dionex Ultimate 3000 LC-system including auto-sampler and column oven (Dionex part of Thermo, Oberschleiβheim, Germany). The employed column was a Kinetex C18 50 × 2.1 mm, 1.7 µm, protected with a C_8 pre-column (Phenomenex, Utrecht, The Netherlands). H_2O (A) and methanol (B) both with 0.01% acetic acid were used. The gradient program started at 40% eluent B and was kept constant for 1 min, then linearly increased to 45% B at 1.1 min, then to 53.5% B at 4 min, to 55% B at 6.5 min, then to 90% B at 12 min and finally to 100% B at 12.1 min, kept constant for 3 min. The flow rate was set to 250.0 µL/min. The MS was operated under the following conditions: the collision gas flow was set to medium, the drying temperature was 400 ºC, the needle voltage -4500 V, the curtain gas was 30 psi, ion source gas 1 was 40 psi and the ion source gas 2 was 30 psi (air was used as drying gas and nitrogen as curtain gas). For quantitation, the multiple reaction monitoring (MRM) transitions and collision energies (CE) given in supplementary table SV were used combined with calibration lines. All substances used as standards were from Cayman Chemicals (Ann Arbor, MI, USA) if not stated otherwise, except RvE1, RvE2 18S-RvE3 and 18R-RvE3 (gifts from Dr. Makoto Arita, Tokyo, Japan). Metabolite identification in

Table 2: Pathways regulated on gene expression level by niacin in the gWAT according to global test.

KEGG ID	KEGG pathway name	p-value	FDR q-value
map01040	Biosynthesis of unsaturated fatty	1.81E-05	0.00381
map00310	Lysine degradation	7.97E-04	0.16654
map00900	Terpenoid backbone biosynthesis	1.02E-03	0.21273
map00620	Pyruvate metabolism	1.09E-03	0.22603
map00100	Steroid biosynthesis	1.32E-03	0.27205

plasma was verified by MS/MS spectral comparison with standards, of which leukotriene E4, thromboxane B2 and 19,20-diHDPA are included in the supplements (Supplementary figure SIV until SVI).

Statistics

Mean values and standard deviations are reported in all figures. The gene expression data were statistically analyzed by using the multiple test correction method of Benjamin-Hochberg for control of false discovery rate (FDR) for both differentially expressed individual genes and for KEGG pathways. An adjusted p-value < 0.05 was considered significant. Calculations for the lipid measurements were performed in Prism version 6 (GraphPad Software, La Jolla, USA).

Multiple t-tests were performed and a 5% FDR value was applied. An F-test was applied to test whether linear regression lines were significantly non-zero. The levels of significance were set at $p < 0.05$.

Results

gWAT gene expression analysis

Female APOE3.Leiden.CETP mice (n=14 per group) were fed a Western type diet (containing 0.1% cholesterol) with and without niacin for 15 weeks. As previously published [16], niacin treatment did not lead to differences in body weight nor gonadal white adipose tissue weight in these mice. However, plasma lipids, i.e. total cholesterol, triglycerides and phospholipids were all decreased [16]. Gene expression analysis generated 24 differentially expressed genes due to niacin treatment after multiple test correction (adjusted $p<0.05$, see Table 1). The global test was applied to identify KEGG pathways affected by niacin treatment. Table 2 depicts the top 5 pathways identified by global test, however only "biosynthesis of

Table 3: Gene expression level of significant genes in the "Biosynthesis of unsaturated fatty acids" pathway and the associated enzymatic substrate/product ratio of FAs.

Gene name	Symbol	Adjusted P	Fold change	$\frac{Substrate}{product}$ ratio	p-value	Fold change
Trans-2,3-enoyl-CoA reductase	Tecr	0.029	0.498(↑)	General FA elongation		
Elongation of long chain fatty acids 6	Elovl6	0.004	1.371(↑)	C16:0 / C18:0	0.416	-0.152(↓)
				C16:1n-9 / C18:1n-9	0.019	-0.370(↓)
Elongation of long chain fatty acids 5	Elovl5	0.033	0.405(↑)	C18:3n-3 / C20:3n-3	0.007	-0.619(↓)
				C18:4n-3 / C20:4n-3	Sub & prod not measured	
				C18:2n-6 / C20:2n-6	0.028	-0.540(↓)
				C18:3n-6 / C20:3n-6	0.049	-0.390(↓)
Elongation of long chain fatty acids 5/2	Elovl5/ Elovl2			C20:5n-3 / C22:5n-3	0.155	0.529(↑)
				C20:4n-6 / C22:4n-6	0.032	0.387(↑)

unsaturated fatty acids" remained significant after correction for false discovery rate (q<0.05). The differentially expressed genes from Table 1 were clustered and highlighted according to KEGG pathways. The top-hits from the "biosynthesis of unsaturated fatty acids" (*Elovl6*, *Tecr* and *Elovl5*) were all specifically involved in FA elongation, not FA desaturation, and were all up-regulated. Quantitative PCR measurements of *Elovl6* and *Elovl5* in gWAT confirmed up-regulation of mRNA levels of these enzymes after niacin treatment (Fig. 1). The rate-limiting desaturase enzyme of PUFA synthesis encoded by *Fads2* (Fatty acid desaturase 2) showed a trend towards increased expression after niacin.

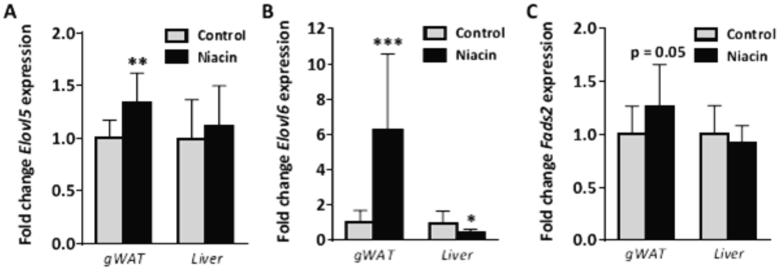


Figure 1: Gene expression by qPCR of gWAT and liver tissue isolated from unfasted control and niacin treated mice. A) Elovl5, B) Elovl6 and C) Fads2 mRNA levels expressed as fold change from control. *p<0.05, **p<0.01, ***p<0.001 compared to control

gWAT fatty acid composition and adipocyte PUFA secretion

To investigate whether the increased mRNA levels of genes in the "biosynthesis of unsaturated fatty acids" translated to adipose tissue FA metabolism changes, we examined the FA composition of the gWAT by GC-MS. In the adipose tissue the fractions of the substrates for PUFA synthesis, the essential fatty acids α- linolenic acid (ALA, n-3) and linoleic acid (LA, n-6), were decreased after niacin treatment while their down- stream products were not fractionally different (Supplementary figure SI and table SII). As the only source of essential FAs was the diet, of which the consumption was equal (data not shown), an increased enzymatic processing of essential FAs towards down-stream elongated and desaturated PUFAs would be plausible. To examine enzymatic processing, we investigated the substrate/product ratios for the enzymes in the PUFA synthesis pathway. We exclusively found differential elongase ratios and no desaturase ratios between control and niacin treatment (data not shown). Furthermore, the differential ratios that were decreased were the C18 to C20 elongation ratios, while the C20 to C22 ratios were increased indicating a possible increase in the metabolism and processing of essential FAs towards down-stream PUFAs in gWAT from niacin treated mice (Table 3). Given that niacin did no elevate the fractional content of the down-stream PUFAs of the essential FAs, we studied whether niacin treatment increased PUFA secretion from freshly isolated adipocytes.

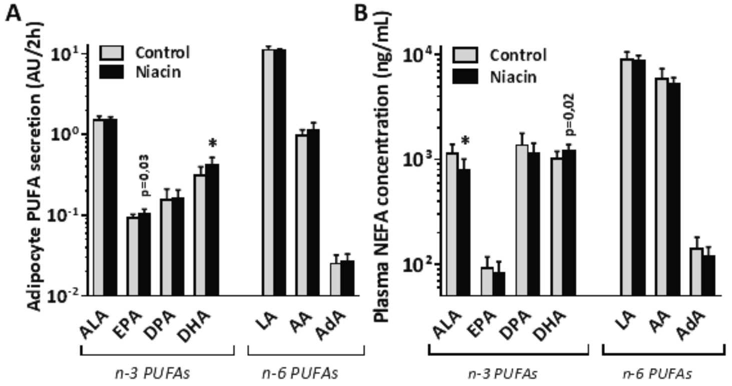


Figure 2: A) PUFA release from ex vivo isolated adipocytes from control and niacin treated mice incubated for two hours in DMEM/F12 medium. B) PUFA concentration in unfasted plasma of control and niacin treated mice. Mean±SD, n=14 for Control/n=13 for Niacin. *p<0.05 compared to control gWAT after FDR correction. P-values listed were before FDR correction.

Although the fraction of medium chain fatty acids (MCFA, C10:0 / C12:0 / C14:0) was also increased in gWAT after niacin, adipocyte release of these MCFA was not different (Supplemental figure S VII). Of the PUFAs, both ALA and LA were secreted in equal amounts for control and niacin treated adipocytes (Figure 2A). Interestingly, down-stream metabolic products of the essential n-3 fatty acid ALA, namely EPA (non- significant after FRD correction) and DHA, were secreted to a greater extent after niacin treatment.

Liver PUFA biosynthesis gene expression and fatty acid composition

As adipose tissue and the liver are the main sites of NEFA processing, we also examined the effects of prolonged niacin on the liver. We found by using qPCR that Elovl5 and Fads2 expression were unaffected by niacin treatment, while Elovl6 expression was down-regulated (Figure 1). Liver fatty acid composition did not differ between control and niacin treated mice (supplementary figure SII and table SIII), neither did the substrate/product ratios relevant for PUFA biosynthesis (Data not shown). Although the PUFA fractions of the livers from niacin treated mice went in the inverse direction as seen in gWAT, this effect was non-significant.

Plasma PUFAs and oxylipins

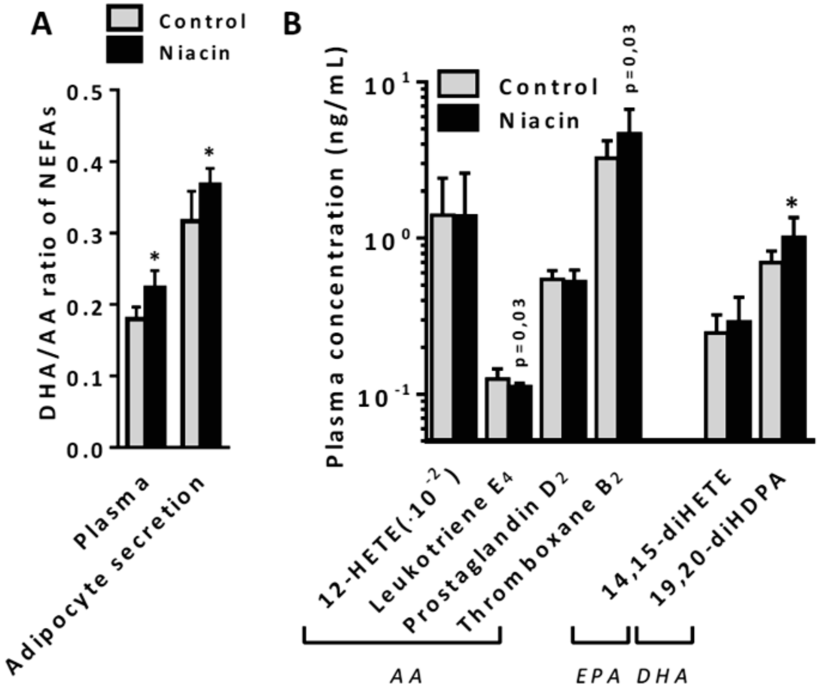


Figure 3: A) Docosahexaenoic acid over arachidonic acid ratio in adipocyte secreted medium and in plasma. B) Oxylipin concentration in plasma of control and niacin treated mice. Mean±SD, n=14 per group. *p<0.05 compared to control gWAT after FDR correction. P-values listed were before FDR correction.

In addition to measuring PUFA levels in adipocyte medium *ex vivo* we also examined PUFA levels in plasma by LC-MS/MS. Niacin reduced circulating levels of ALA and tended to increase the levels of its down-stream product DHA (Figure 2B and supplementary table SIV, DHA was NS after FRD correction). EPA levels were not affected by niacin. We next examined the ratio of DHA over AA as a surrogate marker for PUFA associated cardiovascular risk [27–29] and found that the ratio was shifted towards DHA, both in adipocyte medium and in plasma (Figure 3A). PUFA derived oxylipin signaling molecules were also measured in the plasma (Figure 3B and supplementary table SIV). Arachidonic acid metabolite prostaglandin D_2 was not affected by niacin treatment, whereas thromboxane B_2 levels

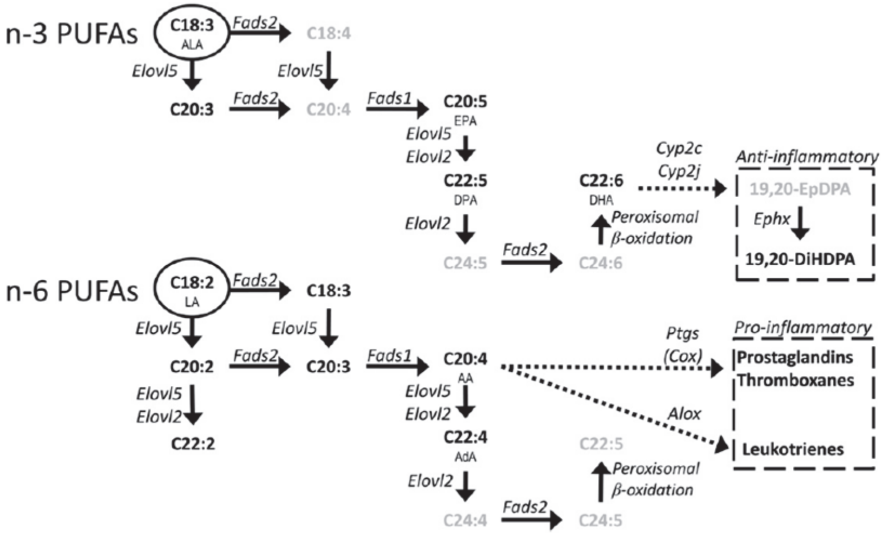


Figure 4: Schematic overview of the synthesis of poly unsaturated fatty acids and the subsequent conversion to a selection of oxylipins. Genes are in italic, metabolites in bold and essential FAs are encircled. Metabolites in grey were not measured. Based on the review by Guillou et al.[31]

increased (NS after FDR correction). AA metabolite leukotriene E4 decreased after niacin treatment (NS after FDR correction), whereas 12- hydroxy eicosatetraenoic acid (12-HETE) levels remained unchanged. The n-3 PUFA derived diol metabolite 19,20-dihydroxy docosapentaenoic acid (19,20-diHDPA) produced by cytochrome P450 was significantly increased. Due to the increase in DHA levels we investigated the presence of DHA derived resolvins [30], which could however not be detected by our approach.

Discussion

The current study demonstrates for the first time that prolonged niacin treatment results in an up- regulation of the n-3 PUFA synthesis pathway in adipose tissue. Gene expression analysis of gWAT showed that our hyperlipidemic mouse model responded to niacin by up-regulating genes involved in the unsaturated FA biosynthesis. Fatty acid composition analysis corroborated the increased PUFA synthesis. A higher degree of n-3 PUFA secretion from prolonged niacin treated adipocytes was seen, which was also reflected in increased n-3 PUFA plasma levels. Markedly, the plasma levels of n-3 PUFA derived oxylipins produced by cytochrome P450 and hydrolyzed by soluble epoxy hydrolases were increased. Oxylipins produced by cytochrome P450 from n-3 PUFAs and the n-3 PUFAs themselves suggest a beneficial vascular health profile, which might contribute to the prolonged niacin-induced atheroprotective effect.

Gene expression analysis of the gonadal white adipose tissue of hyperlipidemic mice treated with niacin for 15 weeks demonstrated an up-regulation of the "biosynthesis of unsaturated fatty acid" pathway, mostly by up-regulation of *Elovl6*, *Tecr* and *Elovl5*. All three genes are involved in FA elongation, not desaturation (as shown in figure 4 and table 3). This discovery was confirmed by qPCR, but also by gWAT FA composition and FA ratio analysis, which all pointed towards PUFA elongation. This increase in PUFA elongation was seen in adipose tissue, but not in liver tissue, where a more inverse trend towards PUFA accumulation could be seen in the fatty acid composition. When examining the PUFA secretion of adipocytes isolated from these mice, we found that specifically end-products of n-3 PUFA biosynthesis were secreted to a higher degree, as seen by DHA (C22:6) and also by EPA (C20:5) secretion. As the genes involved in PUFA biosynthesis are the same for n-3 PUFAs as for n-6 PUFAs, the specificity for increased n-3 PUFA secretion was puzzling. It is conceivable that the PUFA biosynthesis enzymes have a higher affinity for n-3 PUFAs, as was already shown for zebrafish desaturase enzymes [32]. The rat elongase 5 enzyme possesses a higher affinity for n-3 substrates than for n-6 substrates [33], and the mouse equivalent was found to be up-regulated in our study. Selective DHA biosynthesis, unlike AA or EPA, requires partial peroxisomal beta oxidation (Figure 4). Although the microarray did not point towards this pathway, increased peroxisomal beta oxidation after niacin could lead to preferential DHA synthesis. Asides from preferential n-3 PUFA biosynthesis, preferential mobilization from adipose tissue would also explain an increased n-3 PUFA release. A well-documented phenomenon is selective PUFA release from adipocytes [34], exemplified by fasting-induced preferential n-3 PUFA depletion of adipose tissue triglycerides [35]. Our preliminary results also indicate preferential n-3 PUFA release from adipocytes when (fasting-induced) lipolysis is stimulated by 8Br-cAMP (Supplemental figure S VIIIc). Other potential mechanisms for preferential n-3 PUFA release might be phospholipid hydrolysis, as it has been previously shown that cytosolic PLA2 releases AA and EPA from phospholipids whereas the release of DHA from phospholipids requires calcium-independent PLA2 [36]. Although 99% of the fatty acids are located in the triglyceride fraction, the contribution of the 1% fatty acids contained in the phospholipid fraction to n-3 PUFA release cannot be excluded. Additional research is required to investigate the underlying mechanisms for the preferential n-3 PUFA release after prolonged niacin treatment.

Adipocyte lipolysis contributes to the free fatty acid pool in the circulation. In the plasma of the niacin treated animals, we found a tendency for increased levels of the n-3 PUFA DHA in the NEFA pool. Although we do not have direct proof, our data suggest that DHA secretion by adipocytes is the main source of DHA in the plasma. Interestingly we did not find up-regulation of gene expression levels of *Elovl5*, *Elovl6* nor *Fads2*, or any change in fatty acid composition in the livers of the niacin treated mice, indicating that the niacin induced PUFA synthesis is selective for adipose tissue.

The n-3 PUFAs have been reported to confer CVD protective abilities via their conversion to anti- inflammatory oxylipins. For example, DHA can be converted to the oxylipin 19(20)-epoxy docosapentaenoic acid (19(20)-EpDPA) by cytochrome P450 (CYP) as can be seen in figure 4. Likewise, the n-3 PUFA EPA can be converted to 14(15)-epoxy eicosatetraenoic acid (14(15)-EpETE) by CYP. These epoxide metabolites have powerful biological effects on cardiovascular health. This was shown by previous studies where the epoxide metabolism pathway was genetically manipulated [37] or its compounds were pharmacologically elevated [38]. These studies showed the importance of epoxy metabolites in resolving inflammation, preserving vascular tone and general vascular homeostasis. The biologically active 19(20)-EpDPA and 14(15)-EpETE can be hydrolyzed by soluble epoxy hydrolases (encoded by the *Ephx2* gene in mice) to their respective diol metabolites 19,20-diHDPA and 14,15-dihydroxy eicosatetraenoic acid (14,15-diHETE). The levels of both these diol products were increased in plasma of niacin treated animals. The hydrolyzed diol metabolites have a far lower biological effect than their epoxide metabolites, but are more stable and can be detected in plasma by LC-MS/MS. Although we did not directly measure whether the levels of the bioactive epoxy metabolites 19(20)-EpDPA or 14(15)-EpETE were increased after niacin treatment, we found a positive correlation in plasma between the precursor and diol metabolite of 19(20)-EpDPA (DHA and 19,20- diHDPA) in niacin treated mice (Supplementary figure SIII). This correlation suggests that the levels of 19(20)-EpDPA must also have increased after niacin treatment.

In general, the anti-inflammatory oxylipins such as epoxy metabolites produced by CYP (high affinity for n-3 PUFAs), are balanced by the pro-inflammatory oxylipins such as those produced by cyclooxygenases (COX) and arachidonate lipoxygenases (ALOX) (both with high affinity for n-6 PUFAs, such as AA) [39]. Prolonged niacin treatment did not dramatically affect AA derived oxylipin levels, although there was a tendency towards decreased levels of leukotriene E4, a lipoxygenase pathway product stimulating inflammation, and towards increased levels of thromboxane B2, a cyclooxygenase product stimulating coagulation.

Acute treatment of mouse adipocytes with niacin did not lead to an increased release of DHA or AA, nor a change in the ratio of DHA/AA in the adipocyte conditioned medium (Supplemental figure S IX). Acute niacin treatment however, is a well-known trigger for AA-derived oxylipin synthesis in the skin. Irritative subcutaneous skin flushing is a common acute side-effect of niacin, induced by cyclooxygenase product prostaglandin D2 [40] in Langerhans cells and keratinocytes. As mentioned above, we did not see an increase in pro-inflammatory prostaglandins after prolonged niacin treatment. These results are in line with results by Stern et al. [17] and suggest tolerance for flushing after prolonged niacin treatment. It is possible that the tolerance for flushing after prolonged niacin is mediated via n-3 PUFAs as suggested by vanHorn et al. [41]. Whether there is a role for anti-inflammatory n-3 PUFA derived oxylipins after acute niacin remains unclear. Inceoglu et al. [42] have acutely administered niacin to mice being treated with a soluble epoxide hydrolase inhibitor, which resulted in a blunted flushing response compared to wild type mice, while acute prostaglandin D2 treatment did not blunt flushing. These results support a role for cytochrome P450 epoxide metabolites not only after prolonged niacin treatment, but also acutely in inhibiting the flushing response by niacin. Flushing severity also suggests an important balance between pro- and anti-inflammatory oxylipins, which can be modulated by niacin treatment. Most likely, the n-6 derived oxylipins prevail during acute niacin treatment, while after prolonged niacin treatment the n-3 derived oxylipins prevail.

Plasma DHA/AA ratio has been shown to be a diagnostic marker for PUFA associated cardiovascular health [27–29]. In addition to being metabolized to anti-inflammatory oxylipins, n-3 PUFA confer their CVD protective abilities by direct competition with n-6 PUFAs. Vanhorn et al. [41] have described that DHA supplementation increases the DHA/AA ratio in membrane phospholipids of Langerhans cells, thereby diminishing the relative availability of AA for pro-inflammatory prostaglandin synthesis. As a low n-3/n-6 ratio is associated with a risk for cardiovascular disease, increasing the ratio by supplementary n-3 PUFAs has been posed as a treatment target [43]. In our study, we see that the DHA/AA ratio has increased towards the anti-inflammatory DHA side without supplementary n-3 PUFAs. We have seen this increased DHA/AA ratio in both the *ex vivo* adipocyte PUFA secretion profile and in the *in vivo* plasma NEFA profile of niacin treated mice. These effects of niacin on adipose tissue and plasma PUFAs and oxylipins pose a potential contributing mechanism by which niacin treatment reduces cholesterol levels and CVD risk. Although we used mice in this study which are human like with respect to lipoprotein profile it remains to be investigated whether there are changes in the plasma DHA/AA ratio in humans treated with niacin.

In conclusion, prolonged niacin treatment of our hyperlipidemic mouse model with niacin resulted in up-regulation the entire pathway of PUFA biosynthesis in gWAT, increased n-3 PUFA secretion from the adipocytes and an increased plasma level of n-3 PUFAs and their anti-inflammatory oxylipins, which together point towards an atheroprotective plasma profile induced by prolonged niacin treatment.

Acknowledgements

This work was supported by grants from Center of Medical Systems Biology (CMSB), the Netherlands Consortium for Systems Biology (NCSB) established by The Netherlands Genomics Initiative/Netherlands Organization for Scientific Research (NGI/NWO), Leiden University Medical Center. Hulda Jónasdóttir is partially funded by the Stichting Prof. Jan Veltkamp Fonds (LUMC). This study was performed within the framework of the Center for Translational Molecular Medicine (http://www.ctmm.nl);

project PREDICCt (grant 01C-104). We thank Dr. Makoto Arita from the department of Health Chemistry Graduate School of Pharmaceutical Sciences, University of Tokyo, Japan, for the kind gift of RvE1, RvE2, 18S-RvE3 and 18R-RvE3 used as standard material in the employed LC-MS/MS platform. The authors declare that there is no duality of interest associated with this manuscript.

References

1. Bruckert E, Labreuche J, Amarenco P: **Meta-analysis of the effect of nicotinic acid alone or in combination on cardiovascular events and atherosclerosis**. *Atherosclerosis* 2010:353–361.

2. Morgan JM, Capuzzi DM, Baksh RI, Intenzo C, Carey CM, Reese D, Walker K: **Effects of extended-release Niacin on lipoprotein subclass distribution**. *Am J Cardiol* 2003, **91**:1432–1436.
3. Hernandez M, Wright SD, Cai TQ: **Critical role of cholesterol ester transfer protein in nicotinic acid-mediated HDL elevation in mice**. *Biochem Biophys Res Commun* 2007, **355**:1075–1080.
4. Wanders D, Graff EC, White BD, Judd RL: **Niacin Increases Adiponectin and Decreases Adipose Tissue Inflammation in High Fat Diet-Fed Mice**. *PLoS One* 2013, **8**:e71285.
5. Wu BJ, Chen K, Barter PJ, Rye K -a.: **Niacin Inhibits Vascular Inflammation via the Induction of Heme Oxygenase-1**. *Circulation* 2012, **125**:150–158.
6. Chen X, Iqbal N, Boden G: **The effects of free fatty acids on gluconeogenesis and glycogenolysis in normal subjects**. *J Clin Invest* 1999, **103**:365–372.
7. Kasalický J, Konopková M, Melichar F: **18F-fluorodeoxyglucose accumulation in the heart, brain and skeletal muscle of rats; the influence of time after injection, depressed lipid metabolism and glucose-insulin.** *Nucl Med Rev Cent East Eur J Bulg Czech, Maced Polish, Rom Russ Slovak, Yugosl Soc Nucl Med Ukr Soc Radiol* 2001, **4**:39–42.
8. Kaushik S V., Plaisance EP, Kim T, Huang EY, Mahurin a. J, Grandjean PW, Mathews ST: **Extended-release niacin decreases serum fetuin-A concentrations in individuals with metabolic syndrome**. *Diabetes Metab Res Rev* 2009, **25**:427–434.
9. O'Neill M, Watt MJ, Heigenhauser GJF, Spriet LL: **Effects of reduced free fatty acid availability on hormone-sensitive lipase activity in human skeletal muscle during aerobic exercise.** *J Appl Physiol* 2004, **97**:1938–1945.
10. Oh YT, Oh K-S, Choi YM, Jokiaho A, Donovan C, Choi S, Kang I, Youn JH: **Continuous 24-h nicotinic acid infusion in rats causes FFA rebound and insulin resistance by altering gene expression and basal lipolysis in adipose tissue.** *Am J Physiol Endocrinol Metab* 2011, **300**:E1012–E1021.
11. Quabbe HJ, Luyckx AS, L'age M SC: **Growth Hormone, Cortisol, and Glucagon Concentrations during Plasma Free Fatty Acid Depression: Different Effects of Nicotinic Acid and an Adenosine Derivative (BM 11.189).** *J Clin Endocrinol Metab* 1983, **57**:410–4.
12. Watt MJ, Holmes AG, Steinberg GR, Mesa JL, Kemp BE, Febbraio M a: **Reduced plasma FFA availability increases net triacylglycerol degradation, but not GPAT or HSL activity, in human skeletal muscle.** *Am J Physiol Endocrinol Metab* 2004, **287**:E120–E127.

13. Hanson J, Gille A, Zwykiel S, Lukasova M, Clausen BE, Ahmed K, Tunaru S, Wirth A, Offermanns S: **Nicotinic acid- and monomethyl fumarate-induced flushing involves GPR109A expressed by keratinocytes and COX-2-dependent prostanoid formation in mice**. *J Clin Invest* 2010, **120**:2910–2919.
14. Gadegbeku C a., Shrayyef MZ, Ullian ME: **Hemodynamic effects of chronic hemodialysis therapy assessed by pulse waveform analysis**. *Am J Hypertens* 2003, **16**:814–817.
15. Wang W, Basinger A, Neese RA, Christiansen M, Hellerstein MK: **Effects of nicotinic acid on fatty acid kinetics, fuel selection, and pathways of glucose production in women.** *Am J Physiol Endocrinol Metab* 2000, **279**:E50–E59.
16. Heemskerk MM, van den Berg S a a, Pronk ACM, van Klinken J-B, Boon MR, Havekes LM, Rensen PCN, van Dijk KW, van Harmelen V: **Long-term niacin treatment induces insulin resistance and adrenergic responsiveness in adipocytes by adaptive downregulation of phosphodiesterase 3B.** *Am J Physiol Endocrinol Metab* 2014, **306**:E808–13.
17. Stern RH, Spence JD, Freeman DJ, Parbtani a: **Tolerance to nicotinic acid flushing.** *Clin Pharmacol Ther* 1991, **50**:66–70.
18. Choi S, Yoon H, Oh KS, Oh YT, Kim YI, Kang I, Youn JH: **Widespread effects of nicotinic acid on gene expression in insulin-sensitive tissues: Implications for unwanted effects of nicotinic acid treatment**. *Metabolism* 2011, **60**:134–144.
19. Westerterp M, Van Der Hoogt CC, De Haan W, Offerman EH, Dallinga-Thie GM, Jukema JW, Havekes LM, Rensen PCN: **Cholesteryl ester transfer protein decreases high-density lipoprotein and severely aggravates atherosclerosis in APOE*3-Leiden mice**. *Arterioscler Thromb Vasc Biol* 2006, **26**:2552–2559.
20. Kühnast S, Louwe MC, Heemskerk MM, Pieterman EJ, van Klinken JB, van den Berg S a a, Smit JW a, Havekes LM, Rensen PCN, van der Hoorn JW a, Princen HMG, Jukema JW: **Niacin Reduces Atherosclerosis Development in APOE*3Leiden.CETP Mice Mainly by Reducing NonHDL-Cholesterol**. *PLoS One* 2013, **8**.
21. Du P, Kibbe W a., Lin SM: **lumi: A pipeline for processing Illumina microarray**. *Bioinformatics* 2008, **24**:1547–1548.
22. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3**:Article3.

23. Goeman JJ: **A global test for association of a group of genes with a clinical outcome**. *Statistics (Ber)* 2003, **20**:93–99.
24. Kloos DP, Gay E, Lingeman H, Bracher F, Müller C, Mayboroda O a., Deelder AM, Niessen WM a, Giera M: **Comprehensive gas chromatography-electron ionisation mass spectrometric analysis of fatty acids and sterols using sequential one-pot silylation: Quantification and isotopologue analysis**. *Rapid Commun Mass Spectrom* 2014, **28**:1507–1514.
25. Giera M, Ioan-Facsinay A, Toes R, Gao F, Dalli J, Deelder AM, Serhan CN, Mayboroda O a.: **Lipid and lipid mediator profiling of human synovial fluid in rheumatoid arthritis patients by means of LC-MS/MS**. *Biochim Biophys Acta - Mol Cell Biol Lipids* 2012, **1821**:1415–1424.
26. Yang R, Chiang N, Oh SF, Serhan CN: **Metabolomics-lipidomics of eicosanoids and docosanoids generated by phagocytes**. *Curr Protoc Immunol* 2011, **Chapter 14**(SUPPL. 95):Unit 14.26.
27. Nozue T, Yamamoto S, Tohyama S, Fukui K, Umezawa S, Onishi Y KT, Sato A, Nozato T, Miyake S, Takeyama Y, Morino Y, Yamauchi T, Muramatsu T H, K, Terashima M MI: **Low serum docosahexaenoic acid is associated with progression of coronary atherosclerosis in statin-treated patients with diabetes mellitus: results of the treatment with statin on atheroma regression evaluated by intravascular ultrasound with virtual hi**. *Cardiovasc Diabetol* 2014, **13**.
28. Nishizaki Y, Shimada K, Tani S, Ogawa T, Ando J, Takahashi M, Yamamoto M, Shinozaki T, Miyauchi K, Nagao K, Hirayama A, Yoshimura M, Komuro I, Nagai R, Daida H: **Significance of imbalance in the ratio of serum n-3 to n-6 polyunsaturated fatty acids in patients with acute coronary syndrome.** *Am J Cardiol* 2014, **113**:441–5.
29. Dohi T, Miyauchi K, Okazaki S, Yokoyama T, Tamura H, Kojima T, Yokoyama K, Kurata T, Daida H: **Long-term impact of mild chronic kidney disease in patients with acute coronary syndrome undergoing percutaneous coronary interventions**. *Nephrol Dial Transplant* 2011, **26**:2906–2911.
30. Serhan CN, Petasis N a.: **Resolvins and protectins in inflammation resolution**. *Chem Rev* 2011, **111**:5922–5943.
31. Guillou H, Zadravec D, Martin PGP, Jacobsson A: **The key roles of elongases and desaturases in mammalian fatty acid metabolism: Insights from transgenic mice**. *Prog Lipid Res* 2010, **49**:186–199.

32. Hastings N, Agaba M, Tocher DR, Leaver MJ, Dick JR, Sargent JR, Teale a J: **A vertebrate fatty acid desaturase with Delta 5 and Delta 6 activities.** *Proc Natl Acad Sci U S A* 2001, **98**:14304–14309.
33. Gregory MK, Gibson R a, Cook-Johnson RJ, Cleland LG, James MJ: **Elongase reactions as control points in long-chain polyunsaturated fatty acid synthesis.** *PLoS One* 2011, **6**:e29662.
34. Raclot T: **Selective mobilization of fatty acids from adipose tissue triacylglycerols**. *Prog Lipid Res* 2003, **42**:257–288.
35. Nieminen P, Mustonen A-M, Kärjä V, Asikainen J, Rouvinen-Watt K: **Fatty acid composition and development of hepatic lipidosis during food deprivation--mustelids as a potential animal model for liver steatosis.** *Exp Biol Med (Maywood)* 2009, **234**:278–286.
36. Strokin M, Sergeeva M, Reiser G: **Docosahexaenoic acid and arachidonic acid release in rat brain astrocytes is mediated by two separate isoforms of phospholipase A2 and is differently regulated by cyclic AMP and Ca2+.** *Br J Pharmacol* 2003, **139**:1014–1022.
37. Imig JD: **Epoxides and Soluble Epoxide Hydrolase in Cardiovascular Physiology**. *Physiol Rev* 2012, **92**:101–130.
38. Askari A a., Thomson S, Edin ML, Lih FB, Zeldin DC, Bishop-Bailey D: **Basal and inducible anti-inflammatory epoxygenase activity in endothelial cells**. *Biochem Biophys Res Commun* 2014, **446**:633–637.
39. Fischer R, Konkel A, Mehling H, Blossey K, Gapelyuk A, Wessel N, von Schacky C, Dechend R, Muller DN, Rothe M, Luft FC, Weylandt K, Schunck W-H: **Dietary Omega-3 Fatty Acids Modulate the Eicosanoid Profile in Man Primarily via the CYP-epoxygenase Pathway**. *J Lipid Res* 2014, **55**:1150–1164.
40. Serebruany V, Malinin A, Aradi D, Kuliczkowski W, Norgard NB, Boden WE: **The in vitro effects of niacin on platelet biomarkers in human volunteers**. *Thromb Haemost* 2010, **104**:311–317.
41. VanHorn J, Altenburg JD, Harvey K a., Xu Z, Kovacs RJ, Siddiqui R a.: **Attenuation of niacin-induced prostaglandin D 2 generation by omega-3 fatty acids in THP-1 macrophages and Langerhans dendritic cells**. *J Inflamm Res* 2012, **5**:37–50.
42. Inceoglu AB, Clifton HL, Yang J, Hegedus C, Hammock BD, Schaefer S: **Inhibition of Soluble Epoxide Hydrolase Limits Niacin-induced Vasodilation in Mice**. *J Cardiovasc Pharmacol* 2012, **60**:70–75.
43. Simopoulos AP: **The importance of the omega-6/omega-3 fatty acid ratio in cardiovascular disease and other chronic diseases.** *Exp Biol Med (Maywood)* 2008, **233**:674–688.

Supplementary section

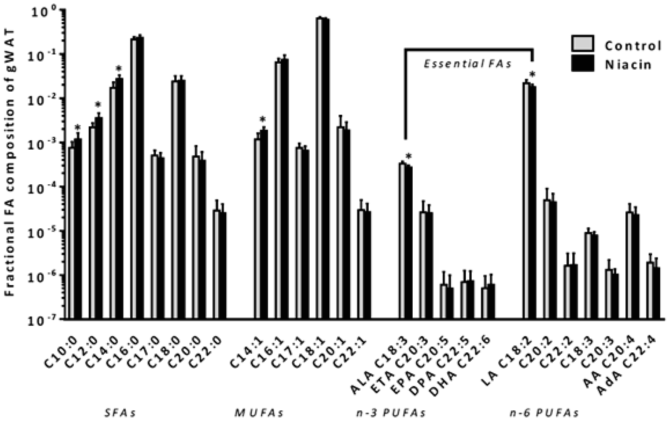


Figure S I : Adipose tissue fatty acid composition of gWAT from APOE*3-Leiden.CETP mice fed a western type diet with 0.1% cholesterol with and without niacin. Mean±SD, N=14 for Control/N=13 for Niacin, *p<0.05 compared to control gWAT after false discovery rate correction

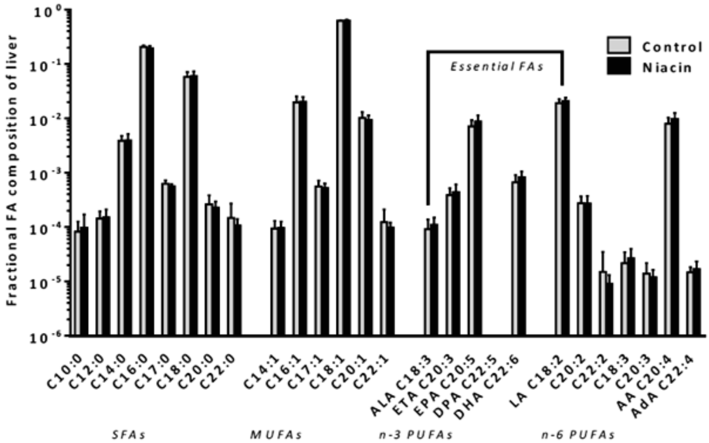


Figure S II: Liver fatty acid composition from APOE*3-Leiden.CETP mice fed a western type diet with 0.1% cholesterol with and without niacin. Fraction of total area corrected sum. Mean±SD, N=14 for Control/N=13 for Niacin. *p<0.05 comparing control gWAT to niacin gWAT after false discovery rate correction.

Table S III : Liver fatty acid composition from APOE*3-Leiden.CETP mice fed a western type diet with 0.1% cholesterol with and without niacin. Fraction of total area corrected sum. Mean±SD, N=14 for Control/N=13 for Niacin. (*)Significant finding after false discovery rate correction.

	Control Liver		**Niacin Liver**		**Control vs Niacin**
SFA	**Average**	**SD**	**Average**	**SD**	**P-value**
C10:0	8,23E-05	*4,42E-05*	9,74E-05	*7,38E-05*	0,5248
C12:0	0,000143	*5,17E-05*	0,000152	*5,92E-05*	0,6948
C14:0	0,003877	*0,000907*	0,003968	*0,001167*	0,8245
C16:0	0,20753	*0,013181*	0,196168	*0,020302*	0,0989
C17:0	0,000625	*9,85E-05*	0,000565	*5,39E-05*	0,0739
C18:0	0,057714	*0,014338*	0,059939	*0,013182*	0,6860
C20:0	0,000262	*0,000122*	0,000229	*6,67E-05*	0,4218
C22:0	0,000147	*0,000125*	0,000107	*3,45E-05*	0,2909
MUFA					
C14:1	9,46E-05	*3,57E-05*	9,7E-05	*2,96E-05*	0,8528
C16:1	0,019737	*0,00563*	0,020314	*0,004366*	0,7759
C17:1	0,00056	*0,000161*	0,000524	*0,000112*	0,5268
C18:1	0,621529	*0,018559*	0,628717	*0,023243*	0,3893
C20:1	0,010283	*0,002869*	0,009425	*0,001852*	0,3834
C22:1	0,000124	*8,64E-05*	9,79E-05	*2,31E-05*	0,3213
PUFA n-3					
ALA C18:3	9,17E-05	*4,6E-05*	0,00011	*3,92E-05*	0,2796
ETA C20:3	0,000388	*0,000133*	0,000438	*0,000173*	0,4111
EPA C20:5	0,00721	*0,00204*	0,00886	*0,002503*	0,0763
DPA C22:5	-	-	-	-	
DHA C22:6	0,000669	*0,000237*	0,000828	*0,000228*	0,0945
PUFA n-6					
LA C18:2	0,019119	*0,003479*	0,020887	*0,003278*	0,1973
C20:2	0,000276	*9,29E-05*	0,000271	*0,000102*	0,9115
C22:2	1,49E-05	*2,02E-05*	9,08E-06	*4,02E-06*	0,3334
C18:3	2,17E-05	*1,29E-05*	2,69E-05	*1,31E-05*	0,3204
C20:3	1,39E-05	*7,87E-06*	1,19E-05	*4,41E-06*	0,4437
AA C20:4	0,008023	*0,002318*	0,009784	*0,002791*	0,0916
AdA C22:4	1,47E-05	*3,65E-06*	1,69E-05	*6,41E-06*	0,3014

Table S IV: Unfasted plasma PUFA and oxylipin concentrations of APOE*3-Leiden. CETP mice fed a western type diet with 0.1% cholesterol with and without niacin. Mean±SD, N=14 for Control/N=13 for Niacin. (*)Significant finding after false discovery rate correction.

	Control		**Niacin**			**Control vs Niacin**
PUFA	**Average (ng/mL)**	**SD**	**Average (ng/mL)**	**SD**		**P-value**
ALA	1139,98	*69,76*	**784,61**	*61,24*		(*)0,0007
EPA	91,51	*7,10*	82,03	*6,51*		0,3342
DPA	1373,47	*118,48*	1133,20	*81,11*		0,0996
DHA	1012,96	*52,65*	1194,23	*52,00*		0,0226
LA	9033,60	*427,04*	8875,20	*273,66*		0,7614
AA	5833,59	*417,18*	5342,29	*192,80*		0,2949
AdA	139,25	*11,33*	119,86	*7,21*		0,1608
Oxylipins						
12-HETE	140,60	*100,95*	138,77	*121,74*		0,9666
Leukotriene E_4	0,125	*0,020*	0,112	*0,005*		0,0324
Prostaglandin D_2	0,545	*0,073*	0,528	*0,097*		0,6066
Thromboxane B_2	3,25	*0,95*	4,67	*2,00*		0,0252
14,15-diHETE	0,247	*0,075*	0,291	*0,127*		0,2891
19,20-diHDPA	0,696	*0,131*	**1,011**	*0,345*		(*)0,0065

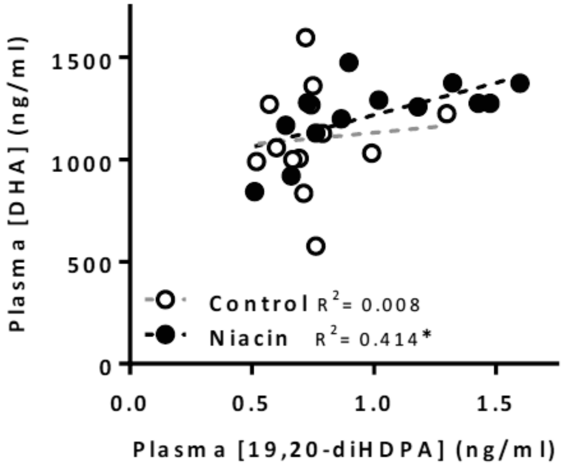


Figure S III: Correlation between the plasma concentrations of 19,20-dihydroxydocosapentaenoic acid and docosahexaenoic acid. N=14 mice per group, $^*p<0.05$ compared to a slope of zero.

Table S V: Multiple Reaction Monitoring setup for ion transitions of the target compounds. Symbols in bold refer to internal standards. RT retention time, Q1 quadrupole 1 ion selection, Q3 quadrupole 3 ion selection, EP entrance potential, CE, collision energy, CCEP collision cell exit potential. HODEs, HOTrEs, HETEs, HEPEs, diHETEs and diHDPAs are given without chiral descriptors

Symbol	**Lipid Maps ID**	**RT (min)**	**Q1 (m/z)**	**Q3 (m/z)**	**DP (Volts)**	**EP (Volts)**	**CE (Volts)**	**CCEP (Volts)**
RvE1	LMFA03070019	4.0	349.1	195.0	-95	-10	-22	-13
20-hydroxy LTB_4	LMFA03020018	4.4	351.1	195.0	-60	-10	-24	-17
8-iso-$PGF_2\alpha$	LMFA03110001	5.1	353.1	193.0	-135	-10	-34	-11
15-keto-PGE_2	LMFA03010030	5.1	349.0	234.9	-65	-10	-20	-13
TxB_2	LMFA03030002	5.2	369.1	169.0	-55	-10	-24	-15
8-iso-PGE_2	LMFA03110003	5.3	351.1	271.0	-5	-10	-24	-19
13,14-dihydro-15-keto-	LMFA03010031	5.6	351.1	235.0	-45	-10	-30	-13
PGE_2-d_4	LMFA03010008	5.6	355.1	193.0	-50	-10	-26	-17
PGE_2	LMFA03010003	5.7	351.2	271.1	-50	-10	-22	-21
PGD_2	LMFA03010004	5.8	351.1	233.0	-30	-10	-16	-13
LXB_4	LMFA03040002	6.0	351.1	220.9	-60	-10	-22	-13
$PGF_{2\alpha}$	LMFA03010002	6.1	353.1	193.0	-80	-10	-34	-11
RvD2	LMFA04000007	6.2	375.1	277.1	-60	-10	-18	-15
LXA_4	LMFA03040001	6.5	351.1	114.8	-40	-10	-20	-11
13,14-dihydro-15-keto-	LMFA03010027	6.6	353.1	195.0	-110	-10	-32	-11

AT-RvD1	LMFA04000074	6.7	375.0	215.0	-50	-10	-26	-11
RvD1	LMFA04000006	6.7	375.1	215.0	-50	-10	-26	-11
epi-LXA$_4$	LMFA03040003	6.8	351.1	114.9	-20	-10	-22	-11
RvE2	LMFA03070036	7.8	333.1	114.9	-35	-10	-18	-15
18S-RvE3	LMFA03070048	8.8	333.1	245.2	-25	-10	-16	-17
6-trans-LTB$_4$	LMFA03020013	8.9	335.1	194.9	-105	-10	-22	-11
8S,15S-diHETE	LMFA03060050	8.9	335.1	207.9	-55	-10	-22	-17
LTD$_4$	LMFA03020006	9.0	495.1	177.0	-70	-10	-28	-19
6-trans-12-epi-LTB$_4$	LMFA03020014	9.1	335.1	194.9	-80	-10	-22	-25
10S,17S-diHDHA (PDX)	LMFA04000047	9.2	359.1	153.0	-70	-10	-22	-9
18R-RvE$_3$	LMFA03070049	9.2	333.1	245.0	-55	-10	-18	-23
7S-MaR1	n.a.	9.3	359.1	249.9	-20	-10	-20	-19
MaR1	LMFA04000048	9.4	359.2	250.2	-65	-10	-20	-13
LTB$_4$-d$_4$	LMFA03020030	9.4	339.1	196.9	-70	-10	-22	-19
LTB$_4$	LMFA03020001	9.4	335.1	195.0	-65	-10	-22	-21
14,15-diHETE	LMFA03060077	9.5	335.1	207.0	-65	-10	-24	-21
7,17-diHDPA	n.a.	9.5	361.1	198.9	-45	-10	-26	-23
LTE$_4$	LMFA03020002	9.6	438.1	333.1	-55	-10	-26	-15
19,20-diHDPA	LMFA04000043	10.2	361.1	273.0	-55	-10	-22	-15
9-HOTrE	LMFA02000024	10.2	293.0	170.9	-75	-10	-20	-15
13-HOTrE	LMFA02000051	10.3	293.0	195.0	-45	-10	-24	-19
18-HEPE	LMFA03070038	10.4	317.1	259.0	-5	-10	-16	-7
15-HEPE	LMFA03070009	10.5	317.1	219.0	-65	-10	-18	-19
13-HODE	LMFA02000228	10.8	295.0	194.9	-110	-10	-24	-21
9-HODE	LMFA02000188	10.8	295.0	171.0	-130	-10	-22	-7
15-HETE-d$_8$	LMFA03060080	10.9	327.2	226.0	-85	-10	-18	-11
15-HETE	LMFA03060001	11.0	319.1	219.1	-55	-10	-18	-9
11-HETE	LMFA03060003	11.1	319.1	167.0	-70	-10	-22	-15
17-HDHA	LMFA04000072	11.1	343.1	245.0	-65	-10	-16	-15
12-HETE	LMFA03060007	11.2	319.1	179.0	-65	-10	-20	-23
8-HETE	LMFA03060006	11.2	319.1	154.9	-70	-10	-20	-19
5-HETE	LMFA03060002	11.3	319.1	115.0	-65	-10	-18	-11
ALA	LMFA01030152	12.4	277.0	233.0	-90	-10	-22	-29
EPA	LMFA01030759	12.4	301.0	202.9	-125	-10	-18	-21
DHA-d$_5$	LMFA01030762	12.4	332.0	288.1	-75	-10	-16	-13
DHA	LMFA01030185	12.7	327.1	229.2	-115	-10	-18	-11
AA	LMFA01030001	12.7	303.0	205.1	-155	-10	-20	-11
LA	LMFA01030120	12.8	279.0	261.0	-115	-10	-28	-13
DPA n-3	LMFA04000044	13.0	329.1	231.1	-50	-10	-20	-17
AdA	LMFA01030178	13.1	331.1	233.0	-130	-10	-22	-11

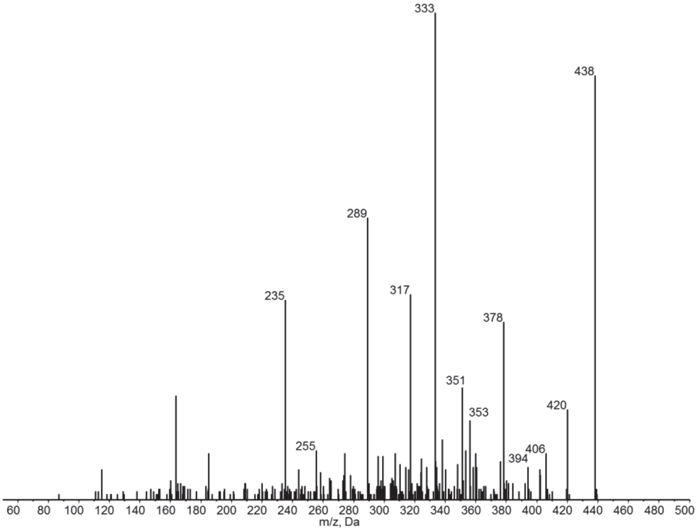


Figure S IVa: MS/MS of 0.1 ng/mL standard sample at Relative RT 1.016 (Leukotriene E4)

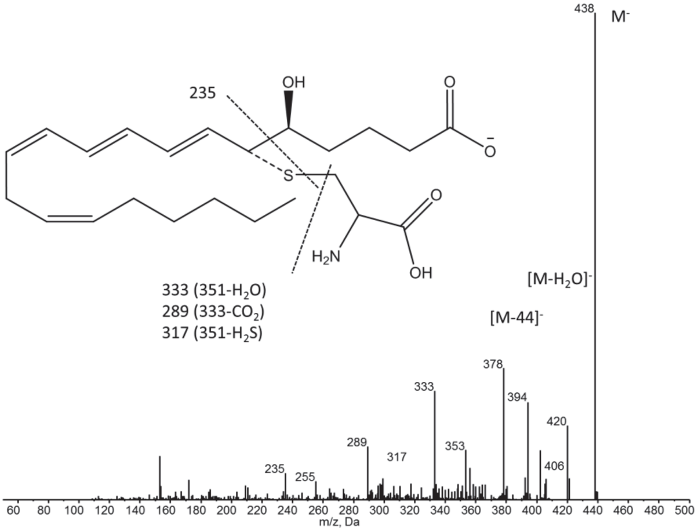


Figure S IVb: MS/MS spectra of representative sample at Relative RT 1.015 (Leukotriene E4)

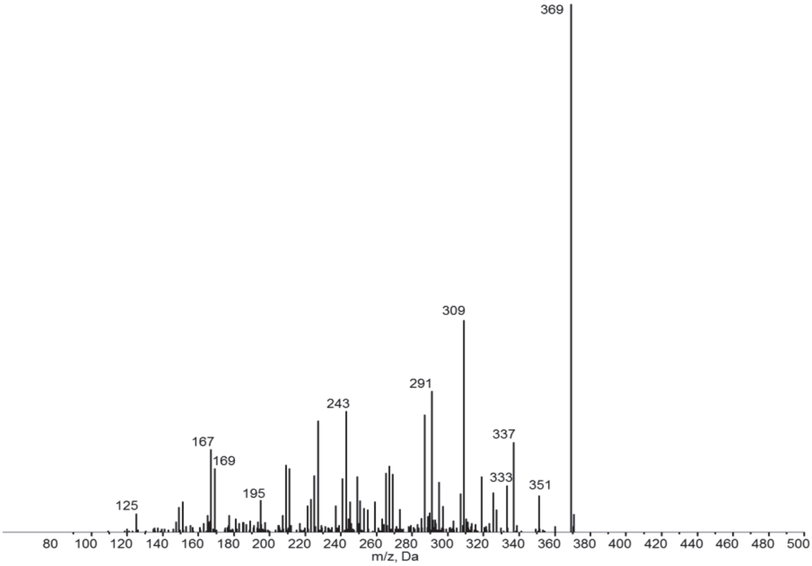


Figure S Va: MS/MS spectra of 0.1 ng/mL standard sample at Relative RT 0.925 (Thromboxane B2)

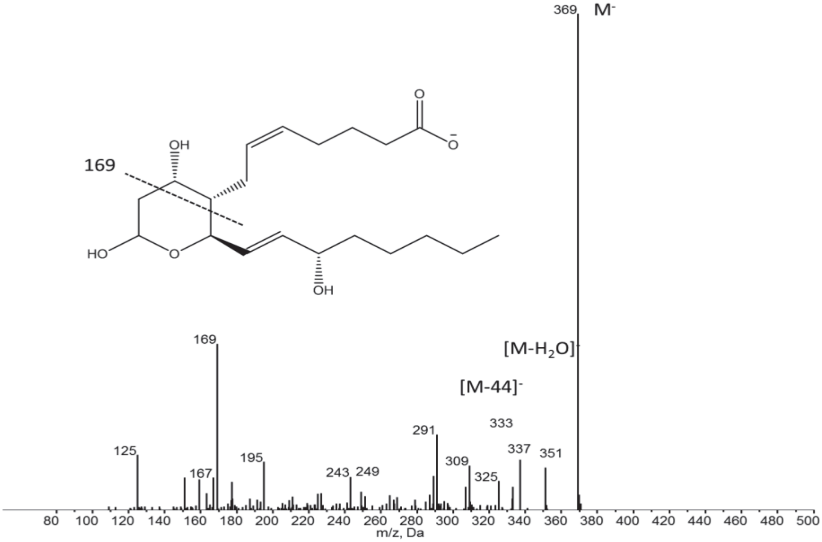


Figure S Vb: MS/MS spectra of representative sample at Relative RT 0.927 (Thromboxane B2)

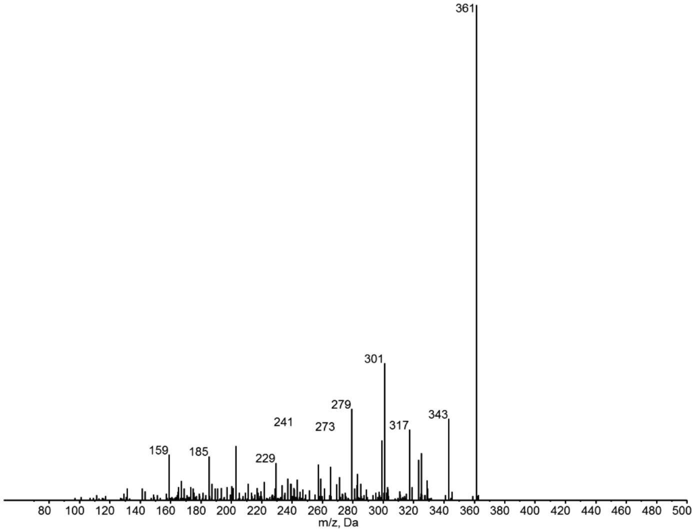


Figure S VIa: MS/MS spectra of 0.1 ng/mL standard sample at Relative RT 1.087 (19,20-diHDPA)

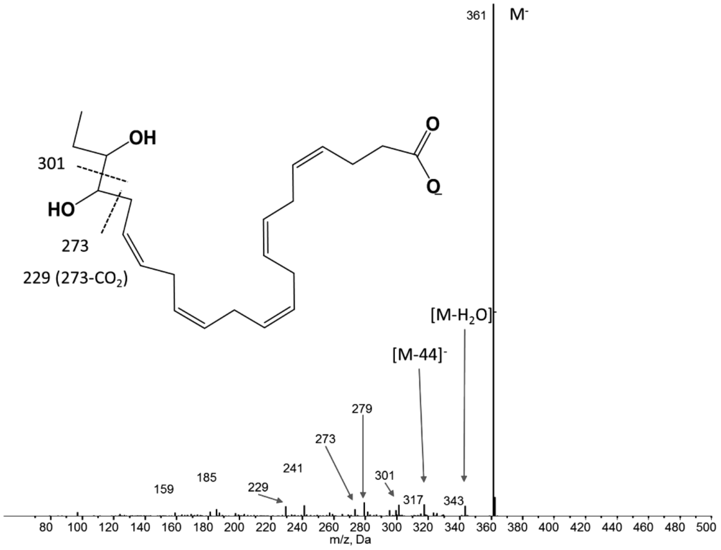


Figure S VIb: MS/MS spectra of representative sample at Relative RT 1.087 (19,20-diHDPA)

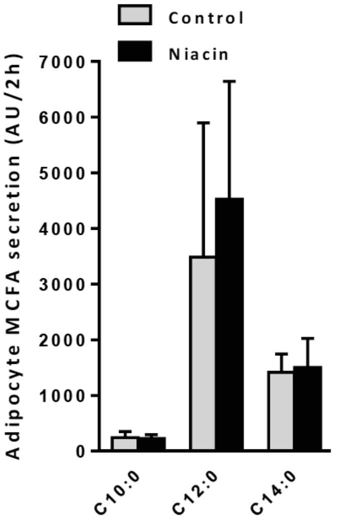


Figure S VII: Release of medium chain saturated fatty acids from adipocytes isolated from APOE*3- Leiden. CETP mice fed a western type diet with 0.1% cholesterol with and without niacin. Fatty acid release in arbitrary units during a 2 hour ex vivo basal incubation. Mean±SD, N=14 for Control/N=13 for Niacin.

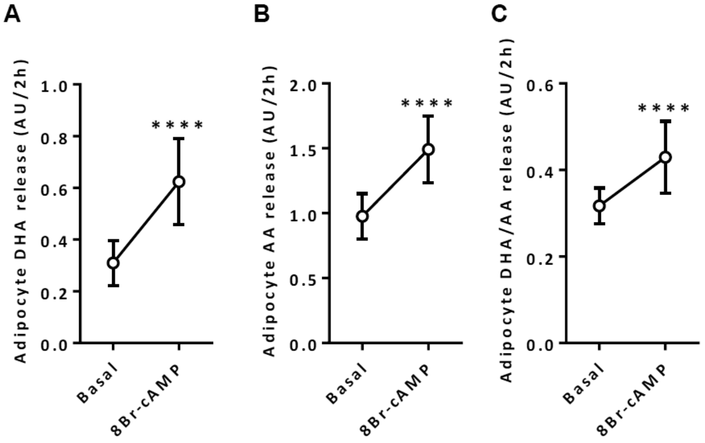


Figure S VIII: Release of DHA and AA from adipocytes isolated from APOE*3-Leiden. CETP mice fed a western type diet with 0.1% cholesterol without niacin. Fatty acid release in arbitrary units during a 2 hour ex vivo incubation in basal and 8Bromo-cAMP stimulated conditions. Mean±SD, N=14 for Control/N=13 for Niacin. **** p<0,0001 for Basal vs 8Br-cAMP.

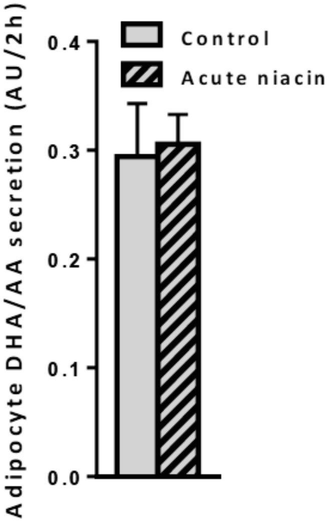


Figure S IX: Ratio of DHA/AA released fatty acids from adipocytes isolated from APOE*3-Leiden. CETP mice fed a western type diet with 0.1% cholesterol without niacin. Fatty acid release in arbitrary units during a 2 hour ex vivo incubation under basal and acute niacin conditions. Mean±SD, N=14 for Control/N=10 for Acute niacin.

Chapter 8: Genetics of the human metabolome, what is next?

Harish Dharuri

Ayşe Demirkan

Jan Bert van Klinken

Dennis Owen Mook-Kanamori

Cornelia M. Van Duijn

Peter A.C. ’t Hoen

Ko Willems van Dijk

Biochim Biophys Acta. 2014; 1842(10):1923-1931

Abstract

Increases in throughput and decreases in costs have facilitated large scale metabolomics studies, the simultaneous measurement of large numbers of biochemical components in biological samples. Initial large scale studies focused on biomarker discovery for disease or disease progression and helped to understand biochemical pathways underlying disease. The first population-based studies that combined metabolomics and genome wide association studies (mGWAS) have increased our understanding of the (genetic) regulation of biochemical conversions. Measurements of metabolites as intermediate phenotypes are a potentially very powerful approach to uncover how genetic variation affects disease susceptibility and progression. However, we still face many hurdles in the interpretation of mGWAS data. Due to the composite nature of many metabolites, single enzymes may affect the levels of multiple metabolites and, conversely, levels of single metabolites may be affected by multiple enzymes. Here, we will provide a global review of the current status of mGWAS. We will specifically discuss the application of prior biological knowledge present in databases to the interpretation of mGWAS results and discuss the potential of mathematical models. As the technology continuously improves to detect metabolites and to measure genetic variation, it is clear that comprehensive systems biology based approaches are required to further our insight in the association between genes, metabolites and disease.

Introduction

The “inborn errors of metabolism” as defined by Garrod at the beginning of the twentieth century depict the first clearly recognized examples of specific genetic defects leading to the accumulation of metabolites in body fluids [1]. For example, in alkaptonuria, a genetic defect in the enzyme homogentisate 1,2-dioxygenase leads to the accumulation of homogentisic acid and its oxide alkapton in plasma and urine. Detection of alkapton in urine is relatively simple in that exposure of urine from affected patients to air results in black discoloration that is readily detected by eye. Alkaptonuria is transmitted as a recessive Mendelian trait with near complete penetrance and is an example of a rare metabolic disease caused by rare genetic variants [2].

Changes in plasma metabolites are also pathogenic hallmarks of common metabolic diseases such as type-2 diabetes. The defining metabolic marker for type 2 diabetes is glucose, but hyperglycemia co-occurs with changes in a variety of additional metabolites including amino acids, lipids and lipoproteins. The high heritability of type 2 diabetes is not explained by rare genetic variants segregating in families, but is thought to be caused by a variety of, and presumably combination of common genetic variants. This paradigm is referred to as “common disease-common variant” hypothesis and is pursued in so-called genome wide association studies (GWAS). In GWAS, genome wide genotyping platforms measure genotypes for hundred thousand to millions of single nucleotide polymorphisms (SNPs) with minor allele frequencies (MAF) generally larger than 0.05 and test each of those SNPs for association with a specific trait [3]. A large number of GWAS have been performed with a variety of both binary traits (e.g. type 2 diabetes) and quantitative traits (e.g. fasting glucose levels). These studies have successfully uncovered genetic variants that contribute to disease risk and also to the variation in quantitative phenotypes [4]. For example, for type-2 diabetes, thus far, more than 60 risk loci have been identified, giving novel insights into the complex pathophysiology of the disease. However, the risk attributed to individual SNPs in the vicinity of even the strongest candidate gene, transcription factor 7-like 2 (*TCF7L2*), are relatively modest (odds ratios of 1.5-1.7) [5]. Moreover, the combined genetic loci discovered to date explain only a small proportion (less than 5%) of the observed heritability of type 2 diabetes. Thus, a significant proportion of the observed heritability remains to be uncovered [6].

Since a large proportion of the SNPs discovered through GWAS are intergenic or lie within the intronic regions of genes, rather than in the protein coding sequences, the genetic basis for the association is often not obvious. It is possible that the SNPs discovered through GWAS are in linkage disequilibrium (LD) with the real causal variant that is not captured by the platform. This hypothesis to uncover “missing heritability” is currently being tested by many labs using next generation deep sequencing approaches to screen the whole genome or whole exome to locate the functional variants. Unfortunately, thus far, these approaches have met with relatively limited success. This lack of success may be associated with our inability to recognize the causative variants among the many detected variants. Alternatively, GWAS hits may constitute expression quantitative trait loci (eQTLs) influencing the expression level of one or more genes nearby (*cis*-eQTLs), or at a distant physical location (*trans*-eQTLs) [7, 8]. Recently, a combination of RNA and genome sequencing has provided in-depth insight into the relation between genetic variation and transcriptome variation and their association with functional variation [9].

Whereas it is often difficult to determine the effect of GWAS-discovered SNPs on nearby or distant genes, it is clear that many different genes and loci are involved in the pathogenesis of complex diseases such as type 2 diabetes. In addition, it is also clear that environmental factors including lifestyle (i.e. diet and physical activity) affect the development of diabetes. Therefore, it may be more appropriate to consider common metabolic disorders such as diabetes as the outcome of a variety and often combination of mild “inborn errors of metabolism” in conjunction with the environment. These mild “inborn errors of metabolism” would be reflected by differences in the concentrations of metabolites in cells and/or body fluids and could provide insight into the “missing heritability”. The terms “genetically determined metabotype” (GDM) [10] and “genetically influenced metabotype” (GIM) have been coined for this [11]. GIM has been defined as relatively prevalent genetic variants that lead to substantial modification in the efficiency of metabolic conversions [12]. The combination of GIMs in any given individual determines his metabolic individuality and thus, in combination with environment and lifestyle, the risk for metabolic disorders such as type 2 diabetes.

Metabolomics measurements

The detection of GIMs has been facilitated by technological developments in the field of metabolomics, where it is now possible to simultaneously measure hundreds of metabolites in large sets of biological samples using automated procedures, and at relatively low cost (10s of euros per sample). A variety of metabolomics platforms are available, all having their own characteristics. Generally speaking, the metabolomics techniques can be divided in two types of platforms and two types of approaches. Metabolomics platforms based on mass spectrometry (MS) in general require extensive sample preparation and are used in-line with gas or liquid chromatography (GC-MS and LC-MS). In contrast, nuclear magnetic resonance (NMR) based platforms require relatively limited sample preparation and the samples can be analyzed without prior separation procedures. MS and NMR based platforms can be employed for targeted and/or non-targeted approaches. In a targeted approach, the platform is optimized for detection of a set of predefined metabolites and absolute or relative concentrations are determined using internal standards. In contrast, in a non-targeted approach, the platform is optimized to capture global snapshots of the test and reference samples and reports the differences. To subsequently identify the metabolites underlying the differential signal in the untargeted approach, additional analyses are required that are frequently challenging. Therefore, metabolomics datasets from a non-targeted approach often contain a large number of ‘unknown’ compounds. The main characteristic of all metabolomics platforms is that a subset of compounds can be detected based on common chemical properties of these compounds rather than their biological relatedness. No single analytical technique exists that is suitable for the identification and quantification of all endogenous metabolites in a sample.

Excellent reviews on the possibilities and challenges of the different metabolomics platforms and approaches are available [13-15]. In general, NMR spectroscopy is highly reproducible and quantitative. However, NMR spectroscopy is relatively insensitive and metabolite identification relies on specialized and mostly proprietary spectral deconvolution algorithms. These algorithms may not always identify the same metabolites and may not always base the identification of a specific metabolite on the same spectral signal. In contrast, MS based platforms provide highly precise information on metabolite mass from which identity can often be inferred. However, metabolite quantification requires spiked internal standards. Thus, a common challenge in metabolomics on any platform is the reproducibility of reported metabolite levels across different laboratories. In addition to these platform-specific challenges, additional variability may be caused by differences in instrumentation and experimental setup conditions such as sample preparation and extraction method, collection protocols, source material (plasma, serum, urine, etc), but also sample storage conditions and batch effects. These aspects all require careful consideration when replicating observations and pooling metabolomics data for meta-analyses.

Genome wide association studies of metabolomics data

Since metabolomics data are (semi)quantitative, they are suited for metabolomics GWAS (mGWAS), uncovering genetic variants that affect metabolite levels. One of the first studies employed an MS-based platform that could identify and quantify up to 363 metabolites in 284 individuals [10]. The study reported that common SNPs explained up to 12% of the observed variance in metabolite levels. Moreover, the study determined that the explained variance could be dramatically increased by considering ratios of metabolites. This is because analyzing ratios of metabolite concentrations potentially reduces the variation in the dataset when the pair of metabolites is related to the substrate and product of a given enzymatic reaction. Furthermore, where a SNP impacts such a metabolic reaction, consideration of ratios leads to a dramatic reduction in p-value of association. For example, rs174548, a SNP in an intron of the fatty acid delta-5 desaturase 1 (*FADS1*) gene is associated with a phosphatidylcholine moiety, PC C36:4 (36 denotes the number of carbons in the side chains and 4 denotes the number of double bonds) levels with a p-value of 4.52×10^{-8}, slightly above the genome-wide threshold. However, association of the same SNP with the ratio of PC C36:4 / PC C36:3 has a p-value of 2.4×10^{-22}, a reduction by 14 orders of magnitude. The FADS1 enzyme introduces a double bond in long chain polyunsaturated fatty acids and the moities PC C36:3 and PC C36:4 are related to the substrate and product of this enzymatic reaction.

A consistent theme that has emerged from mGWAS is that significant SNP-metabolite associations point to the underlying biological mechanism. This is in contrast to GWAS of clinical endpoints where unravelling the underlying mechanism is often much more challenging. In addition to FADS1, several other associations have shown that the functional nature of the gene matches with the biochemical characteristics of the associated metabolite. For example, SNPs in the gene *GLS2* (glutamine synthase 2) have been found associated with glutamine [16, 17].This is a biologically plausible association because the enzyme GLS2 catalyses the hydrolysis of glutamine. Furthermore, genome-wide hits with unknown gene function offer an opportunity to infer novel biological mechanism underlying the SNP-metabolite association. For example, as a proof of principle, Suhre et al experimentally investigated the association of the SNP rs7094971 in the solute carrier family 16, member 9 (*SLC16A9*) with carnitine. The study validated that the hitherto uncharacterized protein was indeed a carnitine transporter in *Xenopus* oocytes [17]. This result underscores the utility of mGWAS in uncovering novel functions and identifying candidate genes for further study.

Table 1 provides an overview of published mGWAS, their characteristics and main findings. It is obvious that the number of highly significant associations is overwhelming and that many of these associations have yet to be interpreted in their proper pathophysiological context. The heritability of small metabolites and amino acids has been reported to vary between 23% and 55%. The heritability of lipids and lipoproteins is somewhat higher ranging, respectively, from 48% to 62% and 50% to 76% [16]. A recent report from a community based cohort indicates that for the majority of metabolites, heritability explains > 20 % of inter-individual variation and that variation attributable to heritable factors is greater than that attributable to clinical factors [18]. The non-heritable proportion of the variation in

Table 1. Overview of published mGWAS datasets. Characteristic features of the study are shown in the table. PC: phosphatidylcholine, SM: sphingomyelin, PE: phosphatidylethanolamine, HDL: high-density lipoprotein, LDL: low-density-lipoprotein, VLDL: very-low-density-lipoprotein, TG: triglycerides, PL: phospholipid, LPC: lysophosphatidylcholine, TC: total cholesterol, PUFA: polyunsaturated fatty acid, Fischer's ratio: (valine + leucine + isoleucine)/(phenylalanine + tyrosine), uk: unknown (not-assigned) peak . *not defined, **reported by two different study groups at the same time.

Study	Platform	Metabolites	MAF	Genotypes	Sample	Discovery	Replication	Novel hits	Known hits	Correction	mGWAS P-value
[10] Gieger, 2008	ESI-MS/MS	metabolites and ratios (n=363)	>0.05	Affy GeneChip Human Mapping 500K	serum	KORAF3 (n=284)	none	reported suggestively significant: PC–*FADS1*, SM–*FADS1*, PE–*LIPC*, SM–*LIPC*, SM–*PLEK*, lysine–*PARK2*, SM–*ANKRD30A*, propionylcarnitine/butyrylcarnitine–*SCAD*, lauroylcarnitine/octanoylcarnitine–*MCAD*	none	Bonferroni	1.3×10^{-9}
[69] Tanaka, 2008	GC	n-3 and n-6 fatty acids (n=6)	*	Illumina Infinium HumanHap550	plasma	InCHIANTI (n=1075)	GOLDN (n=1076)	none	*FADS1* and *ELOVL2*	none	1.0×10^{-7}
[70] Hicks, 2009	ESI-MS/MS	sphingolipids and proportions (n=76)	*	Illumina Infinium HumanHap300	plasma, serum	EUROSPAN (n=4110)	none	ceramide–*ATP10D*, SM–*SGPP1*, SM–*LASS4*, ceramide–*LASS4*, ceramide–*SPTLC3*	SM–*FADS1*	Bonferroni (n=76)	1.0×10^{-10}
[71] Chasman, 2009	1H-NMR (LipoProtein-I and II assays) and direct assay	Lipoproteins (n=22)	>0.01	HumanHap300 Duo chips	plasma	WGHS (n=17296)	PROCARDIS (n=200), FHS (n=2700)	small HDL–*PCCB*, TG–*BTNL2*, medium VLDL–*PPP1R3B*, small LDL–*KLF14*, total HDL–*PAH*, large HDL–*DNAH10*, medium HDL–*WIPI1*	apoB–*PCSK9*, VLDL–*ANGPTL3*, apoB–*CELSR2/PSRC1/SPRT1*, medium HDL–*APOA2*, small VLDL–*APOB*, TG–*GCKR*, LDL cholesterol–*ABCG5*, HDL cholesterol–*COBLL1*/*GRB14*, LDL cholesterol–*HMGCR*, TG–*BTNL2*, TG–*MLXIPL*, medium VLDL–*PPP1R3B*, medium VLDL–*LPL*, small LDL–*TRIB1*, apoA1–*ABCA1*, small VLDL–*ABO*, large HDL–*FADS1-3*, medium VLDL–*APOA1-5*, LDL cholesterol–*HNF1A*, large HDL–*LIPC*, large HDL–*CETP*, apoA1–*LIPG*, LDL cholesterol–*LDLR*, total LDL–*APOC1-APOE*, apoA1–*HNF4A*, small HDL–*PLTP*	none	5.0×10^{-8}
[19] Illig, 2010	Biocrates	metabolites (n=163) and ratios	>0.1	Affymetrix 6.0 GeneChip and Illumina Hap317K	serum	KORAF4 (n=1029)	KORA F4 (n=780),female TWINSUK (422)-3 steps design	reported suggestive loci : glycine/PC–*CPS1*, valine/Isovalerylcarnitine–*SLC22A4*, PC and PC ratios–*SYNE2(SGPP1)* and significant loci : PC and PC ratios–*FADS1*, PC and PC ratios–*ELOVL2*, carnitine ratios–*SCAD(ACADS)*, carnitine	none	Bonferroni*	3.6×10^{-12}

								ratios–*MCAD(ACADM)*, carnitine ratios–*ACADL(LCAD)*, PC and PC ratios *PLEKHH1*, SM and SM ratios–*SPTLC3*, carnitine ratios–*ETFHD*, carnitine–*SLC16A9*			
[17] Suhre, 2011	Metabolon (UHPLC/MS/MS 2 and GC-MS)	>250 metabolites and ratios yielding (n=37,000)	>0.01	Affymetrix 6.0 GeneChip (discovery), HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M(replication) followed by HapMap2 imputations	serum	KORAF4 (n=1768)	TWINSUK (N=1052)	5-oxoproline–*OPLAH*, myristate/myritsoleate–*SCD*, androsterone sulphate–*CYP3A4*, 10-nonadeceonate/ 10-undeceonate–*CYP4A*, glycine–*CPS1*, succinylcarnitine–*LACTB*, isobutyrlcarnitine–*SLC22A1*, serine–*PHGDH*, fibrinogen cleavage peptide ratio–*ENPEP*, andosterone sulphate/epiandrosterone sulphate–*AKR1C*, inosine–*NT5E*, proline–*PRODH*, alpha-hydroxyisovalerate–*HPS5*, fibrinogen cleavage peptine ratio–*ALPL*, bradykinin–*KLKB1*, glutamine–*GLS2*, cafeine/quinate–*AHR*, lactate/isovelarylcarnitine–*IVD*, decanoylcarnitine–*ETFDH*, glutaorylcarnitine/lysine–*SLC7A9*, docosahexaenoic acid/eicosapentaenoic acid–*ELOVL2*, carnitine–*SLC16A9*, isoleucine/tyrosine–*SLC16A10*	butyrylcarnitine/propinylcarnitine–*ACADS*, N-acetylornithine–*NAT8*, PC and PC ratios–*FADS1*, bilirubin/oleolylcarnitine–*UGT1A*, hexaonlycarnitine/oleate–*ACADM*, glucose/mannose–*GCKR*, methylxanthine/4-acetamidobutanoate–*NAT2*, fibrinogen cleavage peptide ratio–*ABO*, urate–*SLC2A9*, eicosenoate/tetradecanedionate–*SLC01B1*, fibrinogen cleavage peptide ratio–*FUT2*, aspartyl-phenyalanine–*ACE*, isovalerylcarnitine–*SLC22A4*, LPC and LPC ratios–*PDXDC1*	Bonferroni (n=37,000)	2.0×10^{-12}
[72]Nicholson, 2011	1H-NMR and Biocrates (FIA-MS)	526 redundant NMR peaks and Biocrates metabolites (n=163)	>0.01	Hapmap 2	Plasma, urine	MolTWIN (n=142) females, longitudinal	MolOBB (n=69)	trimethylamine–*PYROXD2*, dimethylamine–*PYROXD2*, N-acetylated compound–*NAT8*	3-aminoisobutyrate–*AGXT2***		
[73] Suhre, 2011	1H-NMR followed by CHENOMX NMR suit 6.1	56 metabolites and 1661 ratios	*	Affymetrix Human SNP array 6.0	urine	SHIP (males, n=862)	SHIP(n=1032), Longitudinal sample(n=170), KORAF4 (n=992)	formate/succinate–*NAT2*, lysine/valine–*SLC7A9*	3-aminoisobutyrate–*AGXT2***, 2-hydroxyisobutyrate–*WDR66*, alanine/dimethylglycine–*SLC6A20*	Bonferroni (n=1720)	4.5×10^{-11}
[74] Lemaitre 2011	GC	n-3 fatty acids (n=4)	>0.01	HapMap 2	plasma	CHARGE (n=8866)	none	docosapentanoic acid–*GCKR*	alpha-linolenic acid–*FADS1-2*, eicosapentaenoic acid–*FADS1-2*, docosapentanoic acid–*FADS1-2*, eicosapentaenoic acid, docosapentanoic acid, docosahexaenoic acid–*ELOVL2*	none	5.0×10^{-8}

[20] Demirkan, 2012	ESI-MS/MS	Sphingolipids, Phospholipids, (n=115) and their proportions	none	Hapmap 2	plasma, serum	EUROSPAN (n=4034)	none	PE–*PAQR9*, PC–*AGPAT1*, LPC–*PKD2L1*, SM–*PLD2*, SM–*APOE*	PC–*GCKR*, PC–*ELOVL2*, PC–*FADS1-2*, PC–*APOA5*, PC–*PLEKHH1*, PE–*LIPC*, ceramide–*ATP10D*, SM–*SGPP1*, SM–*LASS4*, SM–*SPTLC3*, LPC–*PDXDC1*	Bonferroni (n=23 independent vectors)	2.2×10^{-9}
[16] Kettunen, 2012	1H-NMR, LIPO and LMWM extraction windows	Lipoproteins, lipids, small metabolites (n=117) and 99 selected ratios.	*	7.7 M imputed SNPs from a custom made reference set of HapMap 3 and 1000G	serum	NFBC1966, YF, HBCS, GenMets, DILGOM, Twins (n=8330)	none	alanine/valine–*SLC1A4*, Fischer's ratio–*PPM1K*, phenylalanine–*F12*, glutamine/histidine–*DHDPSL*, phenylalanine/tyrosine–*TAT*, Fischer's ratio–*SLC2A4*, citrate–*SLC25A1*, x-large HDL–*FCGR2B*, albumin–*ALB*, xx-large VLDL–*PPP1R11*, linoleic acid/PUFA–*CPT1A*.	alanine/glutamine–*GCKR*, glucose–*G6PC2*, histidine/valine–*KLKB1*, alanine/tyrosine–*SLC16A10*, glucose–*MTNR1B*, glutamine/glucose–*GLS2*, large LDL free cholesterol–*PCSK9*, mobile lipids–*ANGPTL3*, VLDL diameter–*MLXIPL*, medium VLDL PL–*LPL*, TC/esterified cholesterol–*ABCA1*, linoleic acid/PUFA–*FADS1*, valine/TG–*APOA1*, x-large HDL TG–*LPL*, linoleic acid/PUFA–*PDXDC1*, HDL cholesterol–*CETP*, x-large HDL TG–*LIPG*, medium LDL cholesterol/medium LDL PL–*LDLR*, large LDL free cholesterol–*APOE*, large HDL lipid content/medium HDL lipid content–*PLTP*	Bonferroni (n=216)	2.3×10^{-10}
[75] Wu, 2013	GC	fatty acids (n=4)	>0.01	HapMap 2	plasma	CHARGE Consortium (n=8961)	none	palmitic acid, stearic acid–*ALG14*, palmitoleic acid , oleic acid and stearic acid–*FADS1*, stearic acid–*LPAGT1*, palmitoleic acid–*GCKR*, palmitoleic acid –*PKD2L1* and a locus on chromosome 2.	none	none	5.0×10^{-8}
[76] Rueedi, 2014	1H-NMR	redundant NMR features(n=1276)	*	HapMap 2 (discovery) Illumina OmniQuad (replication)	urine	Colaus (n=835)	TasteSensomics (n=601)	uk–*PYROXD2*, fucose–*FUT*	N-acetylated compounds–*ALMS1(NAT8)*, uk–*ACADL*, 3-aminoisobutyrate–*AGXT2*, uk–*NAT2*, uk–*ABO*, trimethylamine–*PYROXD2*, uk–*PYROXD2*, uk–*ACADS*, 2-hydroxyisobutyrate–*PSDM9*, lysine–*SLC7A9*	Bonferroni (n=125)	5.7×10^{-10}

metabolite levels is likely due to factors such as age, gender, menopause, medication, smoking, nutrition and underlying diseases. The relative contribution and interplay of each of these factors requires larger mGWAS and modeling of gene x environment interactions.

Challenges associated with mGWAS

Metabolomics platforms generally yield information on the levels of one to several hundreds of metabolites. Consideration of all metabolites results in a severe multiple testing burden. This precludes genuine SNP-metabolite pairs from being considered when they fail to reach the stringent statistical threshold for significance. This problem is further exacerbated when considering metabolite ratios. The p-value threshold for a single outcome GWAS is determined by the number of independent genomic loci. Due to the intricate LD structure of the human genome, this p-value is typically set at p < 5 × 10-8. Similar to SNPs in LD, a significant proportion of the metabolites are highly correlated to other often similar metabolites and cannot be considered as independent. To account for multiple test correction, some groups have computed the Bonferroni correction by counting all the metabolites [10, 17, 19], while a few other groups have adopted a less stringent strategy by taking into account the number of independent metabolites as determined by a principle component analysis [20]. A standardized approach to deal with the multiple testing issue in mGWAS remains to be formulated. Another issue relates to the reporting of novel hits. In conventional GWAS, a hit for a specific phenotype is novel if it is independent from previously reported SNPs that are associated with the phenotype. In mGWAS, some of the hits associate with closely related yet non-identical metabolites/phenotypes. In these cases, the association but not the SNP is novel.The mGWAS that have been reported so far followed the classical GWAS approach to uncover genetic variants affecting metabolites and metabolite ratios. The selection of metabolite ratios for GWAS has been done based on selected prior knowledge or simply by analyzing all possible combinations. For example, Illig et al. analyzed the whole ratio matrix of 163 metabolites quantified by a commercial targeted array designed to capture a selection of sugars, amino-acids, acyl-carnitines and phospholipids. Despite the burden of multiple testing inherent to this approach, they still were able to capture associations below the significance threshold, particularly for the *FADS* locus [19]. This is likely due to the fact that both the substrate and the product of the FADS enzymes were present in the platform, which may not be always the case for other metabolites and enzymes. Our group followed a similar approach and performed an mGWAS for phospholipids and sphingolipids. To decrease the burden of multiple testing we used the proportion of each metabolite within its own class, in addition to its absolute concentration and reported additional 6 new loci for these molecules [20]. However, these unbiased but naive approaches seem insufficient to fully exploit the data generated by the metabolomics platforms. Although increasing the sample size will reveal novel genes affecting metabolite levels, additional novel approaches that utilize knowledge of biological relatedness between the molecules are required to take mGWAS one step further.

Various genes that have been identified thus far to affect metabolite levels have also been identified in GWAS of conventional metabolic traits, such as glucose and total plasma lipids. For example, variation in the FADS gene cluster is associated with the fatty acid composition of phospholipids, but also fasting glucose levels, triglycerides and total cholesterol (table 1 and [21]). In addition, the FADS gene cluster has also been associated with the intermediary outcome intima media thickness [20]. These data are in agreement with the notion that phospholipids are somehow causally involved in one of the first steps leading to disturbances in glucose and/or lipid metabolism and subsequent cardio-metabolic disease. However, numerous hypotheses can be formulated to link phospholipids with cardio-metabolic disease. These hypotheses include changes in cellular membrane properties and thus receptor function, but also changes in lipoprotein surface properties and function. Whether any or all of these potential mechanisms play a role in the link between the FADS gene cluster and disease remains to be determined and experimentally validated. However, detailed insight into the specific pathways that are affected by variation in phospholipids is a required first step to select the most likely hypotheses.

Pathway analysis of mGWAS data

Pathway analysis is exquisitely suited to increase the statistical power to identify biologically plausible loci and simultaneously improve our understanding of the underlying biological mechanisms. Pathway-based approaches examine test statistics for a group of genes in contrast to single-marker analysis. The ‘group of genes’ is an expert defined set that is functionally related to the phenotype. The utility of this technique to identify novel and biologically meaningful loci has already been shown in GWAS with clinical endpoints [22-25]. Furthermore, pathway based approaches are uniquely suited to mGWAS owing to the abundance of knowledge on proteins involved in metabolite conversion and secretion, as captured in various databases of metabolic pathways and reactions.

The term 'pathway' in a pathway analysis is usually referring to a set of functionally related genes participating in a common biological process. The resources of prior knowledge that are commonly used in pathway analysis include controlled vocabularies like Gene Ontology [26], manually curated gene sets from MSigDB [27] and the pathway databases like KEGG[28], BioCyc[29] and REACTOME [30]. Metabolic pathways offer the ideal knowledge resource for pathway analysis in mGWAS due to the direct relationship between entities represented in these databases and compounds measured on metabolomics platforms.

Metabolic pathways consist of three tiers of information: 1) metabolites at the lowest level; 2) reactions built from metabolites and the enzymes that drive these reactions; and 3) pathways built upon reactions [31]. Pathway databases like KEGG, BioCyc and Reactome have extended our knowledge of human metabolism. However, no single database captures all relevant biochemical knowledge and conceptual differences between the databases pose a serious challenge to knowledge integration efforts [31, 32]. For example, a study [33] published in 2011 found that the consensus among five major pathway databases at the level of the genes is 13%, at the level of enzyme commission (EC) numbers is 18%, at the level of metabolites is 9% and at the level of the reactions is merely 3%. The lack of consensus in metabolite specific databases extends to resources like HMDB [34] and ChEBI[35] due to differing representation of common metabolites and reactions. Three recent efforts namely BKM-react [36], MetRxn [37], and MNXref [32] attempt to automate the reconciliation of metabolite and reaction information.

Pathway analysis entails selecting a pre-defined set of genes or pathways to test for enrichment. This selection is generally based on the relevance of the test set to the phenotype being assessed by the GWAS. The generation of gene sets relevant to metabolites requires a systematic interrogation of metabolite databases and depends heavily on the accessibility and download formats made available by the database. Furthermore, it is important that the software developed to generate such gene sets is easy to use. To address these issues, we have developed tools to systematically interrogate on-line databases using Taverna [38], a workflow-based management system. Taverna allows users access to remote data resources like KEGG, BioCyc, Ensembl [39] and NCBI [http://www.ncbi.nlm.nih.gov/] and data management systems like Biomart [40] through implementation of web services. To generate a gene set for each of the metabolites measured on a metabolomics platform, we designed workflows to interrogate pathway databases and retrieve genes from pathways and reactions relevant to the metabolite [41]. A corresponding SNP set (SNPs present in ±25 kb flanking region of the genes) was generated for each of the metabolites. As a proof of principle, we investigated the utility of the reduced and biologically relevant SNP set to identify known and novel association from a published GWA dataset by Illig et al [19]. The smaller SNP set reduced the multiple-testing threshold by around two orders of magnitude. This reduction helped us discover novel SNP-metabolite associations in the Illig et al GWAS datasets [41]. For example, a SNP in the gene *ALDH1L1* (aldehyde dehydrogenase 1 L1) was found associated with the ratio of serine/glycine. The original study missed this association because the p-value cut-off in the discovery stage of the study precluded this association from being considered in the replication stage. ALDH1L1 is an important component of the one-carbon pool pathway and acts upstream of SHMT (serine hydroxy methyl transferase) enzyme that mediates the bulk of glycine to serine conversion in the cell. This reaffirms the notion that a method that relies on background knowledge present in pathway databases has the ability to reduce the multiple test burden and thereby facilitate the discovery of true positives in GWAS results. It should be noted that assignment of SNPs to genes represents a challenge in itself. It is common to include only SNPs in the coding region of the gene or within a certain, more or less arbitrary, distance threshold. However, Hong et al [42] note that the reliable conversion of SNPs to representative genes is not trivial and that positional gene clustering if not corrected for can lead to spurious results in a pathway analysis. Properly accounting for LD structure and knowledge on eQTLs will help to link SNPs to the right genes.

Our pathway analysis approach to alleviate the multiple-testing burden through selective testing of SNPs can be seen as complementary to conventional GWAS analysis. However, pathway analysis can also be used in a post-GWAS setting to identify enriched pathways within the identified significantly associated SNPs. We have reported [20] a pathway analysis designed to identify enriched pathways using web accessible software made available by ConsensusPathDB [43]. The latter is a database that integrates pathways and interaction resources made available by databases like KEGG, BioCyc and Reactome. The study reported the enrichment of the following pathways for phospholipid traits: glycerolipid metabolism, chylomicron-mediated lipid transport, triglyceride biosynthesis and metabolism of lipids and lipoproteins. The list of enriched pathways functionally matches the traits, thus reinforcing the importance of pathway analysis in such studies.

Pathway analysis approaches for GWAS can be categorized based on the type of input data and the specific null hypothesis that is being tested [44]. With regard to input data, there are two types of approaches; one approach uses SNP p-values and the other approach uses the effect sizes derived from SNP phenotype data (beta's) to calculate pathway-level statistics. With regard to the null hypothesis being tested, two approaches are available: competitive tests and self-contained tests. A competitive test compares the test statistic of a gene set to a standard defined by its complement. In contrast, a self-contained test compares the test statistic of the gene set to a fixed standard and does not take into account genes in other gene sets. The issues and solutions to SNP-to-gene mapping and tests for gene set enrichment are common to all GWAS and we would like to direct the readers to other excellent reviews [44-46].

Gaussian Graphical Modelling

Gaussian Graphical Modelling (GGM) is an unbiased and database independent approach to reconstruct metabolic networks from large-scale metabolomics data sets [47]. GGMs are undirected probabilistic graphical models, in which pairwise correlations between metabolites are conditioned against the correlations with all other metabolites in the dataset. Krumsiek et al. [47] demonstrated that the high partial correlations represent direct interactions and that groups of metabolites that score highly in the correlation matrix can be attributed to reaction steps in known pathways. As indicated earlier, non-targeted metabolomics platforms also quantify many 'unknown' metabolites. This issue was addressed in a recent work by Krumsiek et al [48] who demonstrated that unknown metabolites can be identified by integrating GGMs with mGWAS results. Their method exploits partial correlations between known and unknown metabolites in addition to their association to specific loci in order to generate a hypothesis regarding the identity of the unknown metabolites. Through experimental validation the study provided genetic and biochemical evidence for classification of several unknown metabolites. These studies demonstrate that GGMs in combination with mGWAS could potentially facilitate metabolite classification and also provide a more comprehensive elucidation of enzyme-metabolite relationship and metabolic pathways.

Pleiotropy in mGWAS

Most metabolomics platforms measure numerous metabolites that are highly related and correlated to each other. For example, the Biocrates Absolute IDQ© p150 mass-spectrometry based platform measures up to 163 metabolites belonging to the classes of amino acids, carnitines, and phospholipids. Of the phospholipids, 90 different PCs are measured that only differ based on alkyl/acyl bonds, number of single/double bonds and length of the side chains. Genes that affect the levels or degree of saturation of fatty acids also influence the phospholipid pool. Hence, several loci that participate in fatty acid metabolism associate with multiple phosphatidylcholines [10, 19, 20].

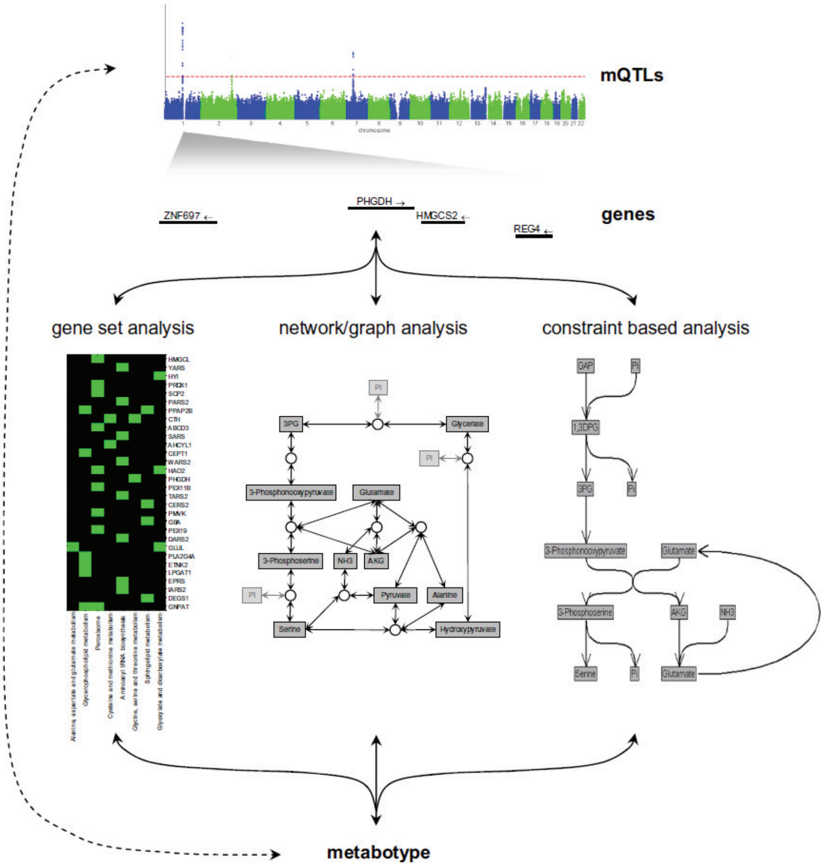
Pleiotropy, the association of a genotype with multiple phenotypes, represents an opportunity to increase the power to identify novel loci and gain insight in metabolic pathways. However, GWAS based on univariate statistical analysis does not take pleiotropy into account. A few groups have developed algorithms and software to exploit the potential of increased statistical power using multivariate statistical analysis [49-54]. Ried et al. [49] developed a method called "Phenotype Set Enrichment Analysis" (PSEA) for the analysis of gene effects on iron-related and blood count traits. The aim of PSEA is to test if a predefined set of phenotypes is associated with a gene. The advantage of such a joint analysis is two-fold: first, the combined analysis of multiple phenotypes can provide insight into the underlying genetic basis and second, it leads to improved statistical power in comparison to association analysis of single phenotypes. PSEA is based on the idea of gene set enrichment analysis for the investigation of phenotype sets. The analysis consists of four steps: i) generate a gene-wise test statistic per phenotype; ii) determine an enrichment score for each combination of phenotype set and gene; iii) a permutation test to determine the enrichment of a phenotype set; and iv) determine the statistical significance of the phenotype set and account for multiple test correction. In another study, Stephens et al. [50] report a unified framework that extensively relies on Bayesian statistics for association analysis of multiple related phenotypes. The utility of the method is illustrated with an application to a genome-wide association study of blood lipid traits from the Global Lipids consortium. To identify novel associations the study applied a two-stage process where in the first stage promising SNPs were identified by applying univariate and multivariate tests to every SNP and in the second stage a Bayesian analysis was performed on the set of promising SNPs. The method could identify 18 potentially novel genetic associations that were not identified by the traditional univariate analysis. In general, however, a limitation of multivariate algorithms is that they operate only for a modest number of phenotypes. Inouye et al [54] report a multivariate analysis that utilizes the correlation structure of the 130 metabolites measured on their NMR platform. An unsupervised algorithm is used to identify metabolic networks and in the next step a multivariate test of association for each of the networks with the SNP panel is performed. The authors report 7 new loci using this method. These results indicate that mGWAS analyses profit from a shift from the univariate analysis paradigm to joint modeling of phenotypes to improve the power in identification of novel loci as well as to improve our understanding of the biological function for known loci.

Towards mechanistic models

To completely understand the relations between different metabolites in the various tissues and cell types, it is essential to have a full description of all relevant metabolic reactions and the involved enzymes and transporter proteins. This knowledge can be utilized to develop mathematical models that describe the fluxes through the metabolic system. Furthermore, these models can then be used to predict how fluxes and metabolite levels change as a consequence of genetic variation. This modelling approach is generally referred to as ‘bottom-up’ systems biology [55].

Recently, in a global research effort several genome-scale metabolic models (GSMMs) have been merged into a consensus model for human metabolism [56]. The key difference between this model and pathway databases is that GSMMs are represented mathematically and have typically undergone additional curation steps that enable mathematical analysis of these models. Most importantly, curation consists of 1) ensuring that all reactions are mass balanced, 2) filling the gaps in the model such that the network is fully connected, and 3) checking that the model is functionally valid, i.e. it has to faithfully predict which metabolic conversions an organism (or tissue) is capable of. A detailed protocol for the reconstruction and curation of organism and tissue-specific GSMMs has recently been described by Thiele and Palsson [57].

Given a fully functional GSMM, its behavior can be analyzed in terms of the space of feasible steady state fluxes through the network, using a set of techniques commonly referred to as constraint-based analysis (CBA) [58-61]. An application of CBA that is of particular interest to metabolic research in



Box 1.

Pathway analysis of mGWAS results. The first step of pathway analysis consists of mapping the locus associated with the metabolite level to a set of candidate genes. Candidate gene selection may be based on LD, vicinity to the associated locus or the fact that the locus affects the expression of a gene at some distance (eQTL). An alternative approach for selecting candidate genes from mGWAS results is to aggregate the *p*-values of all SNPs that lie close to a gene into a gene-wise *p*-value and subsequently consider significant genes. The next step involves integrating the selected genes with knowledge from pathway databases and/or metabolic models. Three separate approaches can be distinguished at this point: (1) gene set analysis, (2) network or graph analysis and (3) constraint based analysis. (1) Gene set analysis employs expert derived gene sets representing biological pathways and processes to determine whether certain sets are statistically enriched for the selected genes. (2) Network analysis uses the topology of a biological network to identify enriched submodules. The most commonly used biological networks are Protein–Protein Interaction (PPI) networks and metabolic networks that consist of graphs with edges between metabolites and enzymatically catalyzed reactions (represented by squares and circles, respectively, in the diagram). (3) Constraint based analysis (CBA) provides a set of mathematical techniques that characterize the functional capacity of a metabolic network in terms of the feasible fluxes through the network. In contrast to traditional network analysis, CBA takes into account the steady state and thermodynamic constraints that are imposed by the set of reactions. That is, internal metabolites may not be net produced or consumed and the flux through irreversible reactions must be non-negative [57–61]. CBA requires well curated genome scale models such as developed by Thiele et al. [56]. In the diagram the Manhattan plot of a GWAS on serine levels is shown, focusing on the locus inside the PHGDH gene that was first discovered by Sühre et al. [17]. The enzyme encoded by PHGDH catalyzes the conversion of 3-phospho-D-glycerate (3PG) to 3-phosphonooxypyruvate. Mapping this gene to the pathway gene sets defined in KEGG shows that it occurs in the “glycine, serine and threonine metabolism” pathway (rn00260), which provides a direct link to serine. Using network analysis, several possible paths are found between the reaction catalyzed by PHGDH and serine that could explain the association between gene and metabolite. Finally, CBA gives a more specific result and shows that PHGDH plays a role in serine biosynthesis humans is to simulate changes in the flux distribution in response to perturbations that reflect pathological or drug treated states. Shlomi et al [62] and Thiele et al [56] have used this method to predict metabolite biomarkers for inborn errors of metabolism. Their approach consisted of predicting the variation in metabolite concentrations and comparing this variation between the healthy case, in which fluxes could pass through the reaction associated with the gene of interest, and the disease case, for which this reaction was blocked. Applying this method on the consensus model of human metabolism, Thiele et al [56] were able to predict directional changes in metabolite biomarkers with an accuracy of 77%. See Box 1 for a comprehensive overview of the different approaches to perform pathway analyses of mGWAS results.

Recently, the use of human GSMMs as a scaffold for the integration and interpretation of omics data has been pioneered by Lewis et al [63], Jerby and Ruppin [64], and Mardinoglu et al [65]. However, GSMMs have not yet been used in the analysis and interpretation mGWAS results. The main advantage of CBA is that it goes beyond traditional methods of pathway analysis where pathways are either represented as pre-defined gene sets or as reaction chains that follow from graph-based searches. Therefore, its application to the mGWAS setting has great potential for providing true mechanistic insight into the links between genetic loci and metabolic phenotypes and constitutes a promising direction for future research. Ultimately integration of GSMMs with genetic data and expression and clinical phenotypes will help unravel disease patho-physiology and identify optimal individualized treatment strategies [66, 67].

Conclusions

The first waves of metabolomics and genetic analyses by mGWAS have provided a wealth of insight into the genetic basis of metabolic individuality and risk factors for common metabolic disorders, even with modest sample sizes and conventional and conservative statistical approaches. However, true understanding of the interrelation between common metabolic disorders, metabolites and genetic variation requires in depth insight into the associated pathways and their regulation. One approach to gain this insight is to mine available pathway databases using statistical tools and this approach has already proven its value in mGWAS. The next step in pathway analyses is to include stoichiometric and kinetic parameters and complement the statistical analyses with a more comprehensive systems biology bases approach using mathematical modelling. GSMMs are a first step towards that direction, but thus far lack quantitative information. Inclusion of quantitative data on the regulation of enzyme activity and reaction kinetics will be vital for developing more accurate predictive models [68]. The combined efforts of numerous research groups around the world to address these issues will pave the way for the application of comprehensive systems biology based approaches to gain insight into the genetics of the human metabolome and especially its relation to disease.

Acknowledgements

The authors are receiving funds from the European Community's Seventh Framework Programme (FP7/2007–2013) 202272 ENGAGE (201413), the Centre for Medical Systems Biology (CMSB) and the Netherlands Consortium for Systems Biology (NCSB), both within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) and the European Commission Seventh Framework Programme Wf4Ever (Digital Libraries and Digital Preservation area ICT-2009.4.1 project reference 270192). DOMK is supported by a personal VENI-grant fromNWO (016.146.023).

References

1. Garrod SAE: **Inborn errors of metabolism**; 1909.
2. Scriver CR: **Garrod's Croonian Lectures (1908) and the charter 'Inborn Errors of Metabolism': albinism, alkaptonuria, cystinuria, and pentosuria at age 100 in 2008**. *J Inherit Metab Dis* 2008, **31**(5):580-598.
3. Siva N: **1000 Genomes project**. *Nature biotechnology* 2008, **26**(3):256.
4. Stranger BE, Stahl EA, Raj T: **Progress and promise of genome-wide association studies for human complex trait genetics**. *Genetics* 2011, **187**(2):367-383.
5. Groves CJ, Zeggini E, Minton J, Frayling TM, Weedon MN, Rayner NW, Hitman GA, Walker M, Wiltshire S, Hattersley AT *et al*: **Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk**. *Diabetes* 2006, **55**(9):2640-2644.

6. Hara K, Fujita H, Johnson TA, Yamauchi T, Yasuda K, Horikoshi M, Peng C, Hu C, Ma RC, Imamura M *et al*: **Genome-wide association study identifies three novel loci for type 2 diabetes**. *Hum Mol Genet* 2013.
7. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ: **Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS**. *PLoS genetics* 2010, **6**(4):e1000888.
8. Zhernakova DV, de Klerk E, Westra HJ, Mastrokolias A, Amini S, Ariyurek Y, Jansen R, Penninx BW, Hottenga JJ, Willemsen G *et al*: **DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts**. *PLoS genetics* 2013, **9**(6):e1003594.
9. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG *et al*: **Transcriptome and genome sequencing uncovers functional variation in humans**. *Nature* 2013, **501**(7468):506-511.
10. Gieger C, Geistlinger L, Altmaier E, Hrabe de Angelis M, Kronenberg F, Meitinger T, Mewes HW, Wichmann HE, Weinberger KM, Adamski J *et al*: **Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum**. *PLoS genetics* 2008, **4**(11):e1000282.
11. Suhre K, Gieger C: **Genetic variation in metabolic phenotypes: study designs and applications**. *Nature reviews Genetics* 2012, **13**(11):759-769.
12. Adamski J, Suhre K: **Metabolomics platforms for genome wide association studies--linking the genome to the metabolome**. *Curr Opin Biotechnol* 2013, **24**(1):39-47.
13. Serkova NJ, Standiford TJ, Stringer KA: **The emerging field of quantitative blood metabolomics for biomarker discovery in critical illnesses**. *Am J Respir Crit Care Med* 2011, **184**(6):647-655.
14. Kell DB: **Systems biology, metabolic modelling and metabolomics in drug discovery and development**. *Drug Discov Today* 2006, **11**(23-24):1085-1092.
15. Griffiths WJ, Koal T, Wang Y, Kohl M, Enot DP, Deigner HP: **Targeted metabolomics for biomarker discovery**. *Angew Chem Int Ed Engl* 2010, **49**(32):5426-5445.

16. Kettunen J, Tukiainen T, Sarin AP, Ortega-Alonso A, Tikkanen E, Lyytikainen LP, Kangas AJ, Soininen P, Wurtz P, Silander K *et al*: **Genome-wide association study identifies multiple loci influencing human serum metabolite levels**. *Nature genetics* 2012, **44**(3):269-276.

17. Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wagele B, Altmaier E, CardioGram, Deloukas P, Erdmann J *et al*: **Human metabolic individuality in biomedical and pharmaceutical research**. *Nature* 2011, **477**(7362):54-60.

18. Rhee EP, Ho JE, Chen MH, Shen D, Cheng S, Larson MG, Ghorbani A, Shi X, Helenius IT, O'Donnell CJ *et al*: **A genome-wide association study of the human metabolome in a community-based cohort**. *Cell Metab* 2013, **18**(1):130-143.

19. Illig T, Gieger C, Zhai G, Romisch-Margl W, Wang-Sattler R, Prehn C, Altmaier E, Kastenmuller G, Kato BS, Mewes HW *et al*: **A genome-wide perspective of genetic variation in human metabolism**. *Nature genetics* 2010, **42**(2):137-141.

20. Demirkan A, van Duijn CM, Ugocsai P, Isaacs A, Pramstaller PP, Liebisch G, Wilson JF, Johansson A, Rudan I, Aulchenko YS *et al*: **Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations**. *PLoS genetics* 2012, **8**(2):e1002490.

21. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(23):9362-9367.

22. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, Sleiman PM, Imielinski M, Glessner J, Hou C *et al*: **Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease**. *American journal of human genetics* 2009, **84**(3):399-405.

23. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE: **Integrating pathway analysis and genetics of gene expression for genome-wide association studies**. *American journal of human genetics* 2010, **86**(4):581-591.

24. Torkamani A, Topol EJ, Schork NJ: **Pathway analysis of seven common diseases assessed by genome-wide association**. *Genomics* 2008, **92**(5):265-272.

25. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC: **Using genome-wide pathway analysis to unravel the etiology of complex diseases**. *Genetic epidemiology* 2009, **33**(5):419-431.
26. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nature genetics* 2000, **25**(1):25-29.
27. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics* 2011, **27**(12):1739-1740.
28. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic acids research* 2012, **40**(Database issue):D109-114.
29. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA *et al*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic acids research* 2012, **40**(Database issue):D742-753.
30. D'Eustachio P: **Pathway databases: making chemical and biological sense of the genomic data flood**. *Chemistry & biology* 2013, **20**(5):629-635.
31. Altman T, Travers M, Kothari A, Caspi R, Karp PD: **A systematic comparison of the MetaCyc and KEGG pathway databases**. *BMC bioinformatics* 2013, **14**:112.
32. Bernard T, Bridge A, Morgat A, Moretti S, Xenarios I, Pagni M: **Reconciliation of metabolites and biochemical reactions for metabolic networks**. *Briefings in bioinformatics* 2012.
33. Stobbe MD, Houten SM, Jansen GA, van Kampen AH, Moerland PD: **Critical assessment of human metabolic pathway databases: a stepping stone for future integration**. *BMC systems biology* 2011, **5**:165.
34. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S *et al*: **HMDB: a knowledgebase for the human metabolome**. *Nucleic acids research* 2009, **37**(Database issue):D603-610.

35. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M *et al*: **The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013**. *Nucleic acids research* 2013, **41**(Database issue):D456-463.
36. Lang M, Stelzer M, Schomburg D: **BKM-react, an integrated biochemical reaction database**. *BMC biochemistry* 2011, **12**:42.
37. Kumar A, Suthers PF, Maranas CD: **MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases**. *BMC bioinformatics* 2012, **13**:6.
38. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P *et al*: **The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud**. *Nucleic acids research* 2013, **41**(Web Server issue):W557-561.
39. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S *et al*: **Ensembl 2013**. *Nucleic acids research* 2013, **41**(Database issue):D48-55.
40. Kasprzyk A: **BioMart: driving a paradigm change in biological data management**. *Database : the journal of biological databases and curation* 2011, **2011**:bar049.
41. Dharuri H, Henneman P, Demirkan A, van Klinken JB, Mook-Kanamori DO, Wang-Sattler R, Gieger C, Adamski J, Hettne K, Roos M *et al*: **Automated workflow-based exploitation of pathway databases provides new insights into genetic associations of metabolite profiles**. *BMC genomics* 2013, **14**:865.
42. Hong MG, Pawitan Y, Magnusson PK, Prince JA: **Strategies and issues in the detection of pathway enrichment in genome-wide association studies**. *Human genetics* 2009, **126**(2):289-301.
43. Kamburov A, Stelzl U, Lehrach H, Herwig R: **The ConsensusPathDB interaction database: 2013 update**. *Nucleic acids research* 2013, **41**(Database issue):D793-800.
44. Wang K, Li M, Hakonarson H: **Analysing biological pathways in genome-wide association studies**. *Nature reviews Genetics* 2010, **11**(12):843-854.

45. Sun YV: **Integration of biological networks and pathways with genetic association studies**. *Human genetics* 2012, **131**(10):1677-1686.
46. Khatri P, Sirota M, Butte AJ: **Ten years of pathway analysis: current approaches and outstanding challenges**. *PLoS computational biology* 2012, **8**(2):e1002375.
47. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ: **Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data**. *BMC systems biology* 2011, **5**:21.
48. Krumsiek J, Suhre K, Evans AM, Mitchell MW, Mohney RP, Milburn MV, Wagele B, Romisch-Margl W, Illig T, Adamski J *et al*: **Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information**. *PLoS genetics* 2012, **8**(10):e1003005.
49. Ried JS, Doring A, Oexle K, Meisinger C, Winkelmann J, Klopp N, Meitinger T, Peters A, Suhre K, Wichmann HE *et al*: **PSEA: Phenotype Set Enrichment Analysis--a new method for analysis of multiple phenotypes**. *Genetic epidemiology* 2012, **36**(3):244-252.
50. Stephens M: **A unified framework for association analysis with multiple related phenotypes**. *PloS one* 2013, **8**(7):e65245.
51. van der Sluis S, Posthuma D, Dolan CV: **TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies**. *PLoS genetics* 2013, **9**(1):e1003235.
52. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin MR, Coin LJ: **MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS**. *PloS one* 2012, **7**(5):e34861.
53. Park SH, Lee JY, Kim S: **A methodology for multivariate phenotype-based genome-wide association studies to mine pleiotropic genes**. *BMC systems biology* 2011, **5 Suppl 2**:S13.
54. Inouye M, Ripatti S, Kettunen J, Lyytikainen LP, Oksala N, Laurila PP, Kangas AJ, Soininen P, Savolainen MJ, Viikari J *et al*: **Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis**. *PLoS genetics* 2012, **8**(8):e1002907.
55. Bruggeman FJ, Westerhoff HV: **The nature of systems biology**. *Trends in microbiology* 2007, **15**(1):45-50.
56. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD *et al*: **A community-**

driven global reconstruction of human metabolism. *Nature biotechnology* 2013, **31**(5):419-425.

57. Thiele I, Palsson BO: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nature protocols* 2010, **5**(1):93-121.
58. Varma A, Palsson BO: **Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110**. *Applied and environmental microbiology* 1994, **60**(10):3724-3731.
59. Edwards JS, Palsson BO: **Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions**. *BMC bioinformatics* 2000, **1**:1-10.
60. Schuster S, Fell DA, Dandekar T: **A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks**. *Nature biotechnology* 2000, **18**(3):326-332.
61. Schilling CH, Letscher D, Palsson BO: **Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective**. *Journal of theoretical biology* 2000, **203**(3):229-248.
62. Shlomi T, Cabili MN, Ruppin E: **Predicting metabolic biomarkers of human inborn errors of metabolism**. *Molecular systems biology* 2009, **5**:263.
63. Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, Cheng JK, Patel N, Yee A, Lewis RA, Eils R *et al*: **Large-scale in silico modeling of metabolic interactions between cell types in the human brain**. *Nature biotechnology* 2010, **28**(12):1279-1285.
64. Jerby L, Ruppin E: **Predicting drug targets and biomarkers of cancer via genome-scale metabolic modeling**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2012, **18**(20):5572-5584.
65. Mardinoglu A, Agren R, Kampf C, Asplund A, Nookaew I, Jacobson P, Walley AJ, Froguel P, Carlsson LM, Uhlen M *et al*: **Integration of clinical data with a genome-scale metabolic model of the human adipocyte**. *Molecular systems biology* 2013, **9**:649.
66. Mardinoglu A, Nielsen J: **Systems medicine and metabolic modelling**. *Journal of internal medicine* 2012, **271**(2):142-154.

67. Varemo L, Nookaew I, Nielsen J: **Novel insights into obesity and diabetes through genome-scale metabolic modeling**. *Frontiers in physiology* 2013, **4**:92.

68. Jamshidi N, Palsson BO: **Formulating genome-scale kinetic models in the post-genome era**. *Molecular systems biology* 2008, **4**:171.

69. Tanaka T, Shen J, Abecasis GR, Kisialiou A, Ordovas JM, Guralnik JM, Singleton A, Bandinelli S, Cherubini A, Arnett D *et al*: **Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study**. *PLoS genetics* 2009, **5**(1):e1000338.

70. Hicks AA, Pramstaller PP, Johansson A, Vitart V, Rudan I, Ugocsai P, Aulchenko Y, Franklin CS, Liebisch G, Erdmann J *et al*: **Genetic determinants of circulating sphingolipid concentrations in European populations**. *PLoS genetics* 2009, **5**(10):e1000672.

71. Chasman DI, Pare G, Mora S, Hopewell JC, Peloso G, Clarke R, Cupples LA, Hamsten A, Kathiresan S, Malarstig A *et al*: **Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis**. *PLoS genetics* 2009, **5**(11):e1000730.

72. Nicholson G, Rantalainen M, Li JV, Maher AD, Malmodin D, Ahmadi KR, Faber JH, Barrett A, Min JL, Rayner NW *et al*: **A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection**. *PLoS genetics* 2011, **7**(9):e1002270.

73. Suhre K, Wallaschofski H, Raffler J, Friedrich N, Haring R, Michael K, Wasner C, Krebs A, Kronenberg F, Chang D *et al*: **A genome-wide association study of metabolic traits in human urine**. *Nature genetics* 2011, **43**(6):565-569.

74. Lemaitre RN, Tanaka T, Tang W, Manichaikul A, Foy M, Kabagambe EK, Nettleton JA, King IB, Weng LC, Bhattacharya S *et al*: **Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium**. *PLoS genetics* 2011, **7**(7):e1002193.

75. Wu JH, Lemaitre RN, Manichaikul A, Guan W, Tanaka T, Foy M, Kabagambe EK, Djousse L, Siscovick D, Fretts AM *et al*: **Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: results from the Cohorts for Heart and Aging Research in**

Genomic Epidemiology (CHARGE) consortium. *Circ Cardiovasc Genet* 2013, **6**(2):171-183.

76. Rueedi R, Ledda M, Nicholls AW, Salek RM, Marques-Vidal P, Morya E, Sameshima K, Montoliu I, Da Silva L, Collino S *et al*: **Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links**. *PLoS genetics* 2014, **10**(2):e1004132.

Chapter 9: General Discussion

In **Chapter 8**, the current state of Genome-Wide association studies of metabolite profiles was reviewed and this served as a discussion of **Chapters 2** and **3**. This chapter provides further perspectives on **Chapters 4, 5, 6**, and **7**.

In **Chapter 4** we report that the downregulation of acetyl-CoA metabolic network is an important feature in the pathophysiology of obese type-2 diabetes patients. This work is in line with the network medicine paradigm that aims to address the problem that a disease is rarely caused by malfunction of one individual gene product, but instead depends on multiple gene products that interact in a complex network [1–3]. Through a network-based bioinformatics analysis, we argue that acetyl-CoA metabolic network is the unifying principle behind previously implicated pathways such as branched-chain amino acid degradation, fatty acid oxidation and citrate cycle dysregulation, in the pathophysiology of type-2 diabetes. The vicinity of acetyl-coA metabolism represents a hotspot where abnormalities in individual genes potentially accumulate and upon reaching a certain risk threshold, lead to the manifestation of type-2 diabetes. Furthermore, disease heterogeneity may arise when affected individuals contribute different genes in this network topography, or variants within these genes, to the manifestation of the phenotype.

Type-2 diabetes is currently believed to be a multifactorial, complex disease. While patients may all exhibit similar clinical manifestations like hyperglycemia and insulin resistance, the underlying etiology is heterogeneous [4]. However, the current disease classification paradigm overgeneralizes pathophenotypes, and does not consider individualized nuances in disease expression [3]. More importantly, these patients are often treated similarly, with little consideration of individual characteristics that might affect clinical outcome and therapeutic response. Therefore, there is growing recognition that personalized approaches to treating type-2 diabetes might bring substantial benefits for the patient as well as the pharmaceutical companies. The application of network medicine to pharmacology and clinical trials paves the way for individualized or “precision” medicine. In the context of our research, the fact that a central pathway in energy metabolism is dysregulated as a whole indicates that pharmacological treatment of individual targets in this metabolic pathway is unlikely to succeed. Chen and Butte [5] point out that the best approach to treating complex disorders such as type-2 diabetes may be to “modulate the disease network by targeting multiple components using a designed polypharmacological ligand or a combination of drugs” rather than using single target drugs. Furthermore, taking individual nuances in the network into consideration in the design of such compounds should pave the way for better and effective drugs. Further research into the causal role of downregulation of the acetyl-CoA network in type 2 diabetes should indicate whether direct intervention in the acetyl-CoA network will provide novel therapeutic approaches.

Chapter 5 explores differential Allele-Specific Expression (ASE) with the aim of identifying genetic variants that are associated with or affect gene expression and contribute to the functional differences observed in visceral and subcutaneous adipose tissues. ASE studies help to understand the cis-regulatory basis of variation in gene expression. The analysis of ASE allows for the analysis of the genetic component of gene expression in much smaller numbers of individuals than in traditional expression quantitative trait loci (eQTL) studies, where the genetic variation is usually only a minor contributor to the total degree of variation in gene expression between individuals [6]. However, it should be realized that allelic imbalance may not be purely genetic, but also caused by epigenetic factors [7, 8]. In general, ASE is a promising technique and has potential for clinical applications. For example, it has been used for tumor type classification [9] and cancer diagnosis [10]. Another application for ASE is in interrogating gene-environment interactions [11]. While environmental factors have been shown to substantially affect human disease risk, this interaction has not been well characterized in genome wide studies owing to small genetic effect sizes and the steep multiple testing burden. By associating risk factors such as diet, exercise, lipid levels, drug usage etc with an individual’s allele-specific expression, it will be possible to understand and treat type-2 diabetes more effectively.

Very low calorie diets (VLCD) with and without exercise programs lead to major metabolic improvements in obese T2DM patients. However, the biological mechanisms underlying these improvements have so far not been elucidated fully. **Chapter 6** describes the effects of VLCD with or without exercise in obese T2DM patients through proteomic analysis of plasma obtained from these subjects as lean controls. This study shows that proteomic analysis reveals many proteins that exhibit significantly different levels in type 2 diabetes patients versus controls and before and after a VLCD. Although this gives us an insight into the proteins affected by obesity, insulin resistance, T2DM and diet-induced weight loss, further studies are needed to establish if these proteins are causally related to these conditions or the success of the intervention. Dense longitudinal sampling could potentially resolve the issue of causality by providing mechanistic insight into the changes occurring in these subjects as a result of VLCD. A recent study [12] showed that a differential metabolic adaptation of mice to a high fat diet is associated with striking differences in gene expression patterns. After an initial state where high fat feeding induces coherent changes in gene expression in liver in all mice tested, there is subsequent modification of gene expression towards patterns characteristic of each phenotype (obese and diabetic, lean and diabetic, lean and non-diabetic). This shows that there is a phased response to an intervention and that mechanisms causing insulin resistance may vary over time. High fat diet-induced obesity/diabetes in the mouse is considered a good model for the pathogenesis of the human conditions. Therefore, future studies should consider integrative, longitudinal "omic" assessments to monitor intervention specific adaptations [13].

Prolonged niacin treatment elicits beneficial effects on the plasma lipid and lipoprotein profiles that are associated with a beneficial cardiovascular disease (CVD) risk profile. However, niacin also elicits unwanted effects which include a severe flushing response. In **Chapter 7** we explore the prolonged effects of niacin on lipid metabolism in adipocytes of a hyperlipidemic mouse model. Prolonged niacin treatment resulted in upregulation of the biosynthesis of unsaturated fatty acids pathway in gonadal white adipose tissue (gWAT), increased n-3 PUFA secretion from the adipocytes, and an increased plasma level of n-3 PUFAs and their anti-inflammatory oxylipins, which together point towards an atheroprotective plasma profile. Niacin (also known as nicotinic acid or vitamin B3) has been widely used in the prevention of cardiovascular disease. However, the majority of patients experience the aforementioned flushing response that is characterized by severe reddening of the skin, itching, and tingling. Studies have shown that the flushing response is due to the vasodilatory effects of prostaglandin D2 (PGD2) and prostaglandin E2 (PGE2) which are formed by the enzymatic action of cyclooxygenase (COX) on arachidonic acid (AA) that is released from membrane phospholipids as a result of niacin action. Interestingly, schizophrenia is associated with a blunted flush response to niacin and there is evidence for the relevance of n-6 PUFA pathway to the phenotype in these subjects [14]. Interestingly, both niacin and omega-3 PUFA have shown clinical potential for the treatment of psychosis in schizophrenia patients [15, 16]. Therefore, a clear mechanistic understanding of the biochemical response to these supplements can benefit both CVD and schizophrenia.

It should be noted that PUFA supplementation as a preventive strategy for CVD has shown mixed results. While a few studies suggested improvements of risk factors [17], a more systematic clinical trial showed no improvements for CVD endpoints [18]. A recent article reported that the Inuit population in Greenland evolved unique genetic adaptations for metabolizing omega-3 and other fatty acids [19]. This discovery raises questions about whether omega-3 fats are really good for everyone despite the recommended guidelines that have been in place for several decades. Alternatively, endogenous induction of PUFA or modulation of PUFA derived oxylipin profile may be explored as therapeutic strategies. However, outside the context of metabolic disease, increased PUFA biosynthesis might be harmful due to their potential oxidation to lipoperoxide inflammatory triggers. Similarly, modulating PUFA conversion to specific oxylipins may have unintended consequences for the inflammatory pathways that play an important role in cancer progression. Therefore these therapeutic strategies must be explored with caution while taking into account the many functions of PUFA metabolites. Future studies should explore novel bioinformatics and systems biology approaches to build a network model that predicts phenotypic traits and outcomes for various perturbations such as niacin and omega-3 PUFA. Furthermore, this model must incorporate individual variation in response, to better understand the genetic underpinnings of these complex pathways.

References

1. Piro RM: **Network medicine: Linking disorders**. *Hum Genet* 2012, **131**:1811–1820.
2. Barabási A-L, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease.** *Nat Rev Genet* 2011, **12**:56–68.
3. Chan SY, Loscalzo J: **The emerging paradigm of network medicine in the study of human disease**. *Circ Res* 2012, **111**:359–374.
4. Malandrino N, Smith RJ: **Personalized medicine in diabetes**. *Clin Chem* 2011, **57**:231–240.
5. Chen B, Butte a J: **Network medicine in disease analysis and therapeutics.** *Clin Pharmacol Ther* 2013, **94**:627–9.

6. Hasin-Brumshtein Y, Hormozdiari F, Martin L, van Nas A, Eskin E, Lusis AJ, Drake T a: **Allele-specific expression and eQTL analysis in mouse adipose tissue.** *BMC Genomics* 2014, **15**:471.
7. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, et al.: **Transcriptome and genome sequencing uncovers functional variation in humans.** *Nature* 2013, **501**:506–11.
8. Lagarrigue S, Martin L, Hormozdiari F, Roux PF, Pan C, van Nas A, Demeure O, Cantor R, Ghazalpour A, Eskin E, Lusis AJ: **Analysis of allele-specific expression in mouse liver by RNA-seq: A comparison with Cis-eQTL identified using genetic linkage**. *Genetics* 2013, **195**:1157–1166.
9. Mayba O, Gilbert HN, Liu J, Haverty PM, Jhunjhunwala S, Jiang Z, Watanabe C, Zhang Z: **MBASED: allele-specific expression detection in cancer tissues and cell lines**. *Genome Biol* 2014, **15**:405.
10. De Lellis L, Aceto GM, Curia MC, Catalano T, Mammarella S, Veschi S, Fantini F, Battista P, Stigliano V, Messerini L, Mareni C, Sala P, Bertario L, Radice P, Cama A: **Integrative analysis of hereditary nonpolyposis colorectal cancer: The contribution of allele-specific expression and other assays to diagnostic algorithms**. *PLoS One* 2013, **8**:1–12.
11. **Allele-specific expression reveals interactions between genetic variation and environment** [http://biorxiv.org/content/early/2015/09/13/025874]
12. De Fourmestraux V, Neubauer H, Poussin C, Farmer P, Falquet L, Burcelin R, Delorenzi M, Thorens B: **Transcript profiling suggests that differential metabolic adaptation of mice to a high fat diet is associated with changes in liver to muscle lipid fluxes**. *J Biol Chem* 2004, **279**:50743–50753.
13. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, et al.: **Personal omics profiling reveals dynamic molecular and medical phenotypes**. *Cell* 2012, **148**:1293–1307.
14. Messamore E, Hoffman WF, Yao JK: **Niacin sensitivity and the arachidonic acid pathway in schizophrenia.** *Schizophr Res* 2010, **122**:248–56.

15. Emsley R, Oosthuizen P, van Rensburg SJ: **Clinical potential of omega-3 fatty acids in the treatment of schizophrenia.** *CNS Drugs* 2003, **17**:1081–91.
16. Durmaz O: **Niacin supplement in schizophrenia: Hit two birds with one stone.** *Eur Rev Med Pharmacol Sci* 2015, **19**:2325.
17. Lorente-Cebrián S, Costa AG V, Navas-Carretero S, Zabala M, Martínez JA, Moreno-Aliaga MJ: **Role of omega-3 fatty acids in obesity, metabolic syndrome, and cardiovascular diseases: A review of the evidence**. *J Physiol Biochem* 2013, **69**:633–651.
18. Rizos EC, Ntzani EE, Bika E, Kostapanos MS, Elisaf MS: **Association between omega-3 fatty acid supplementation and risk of major cardiovascular disease events: a systematic review and meta-analysis.** *JAMA* 2012, **308**:1024–1033.
19. Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jørgensen ME, Korneliussen TS, Gerbault P, Skotte L LA, Christensen C, Brandslund I, Jørgensen T, Huerta-Sánchez E, Schmidt EB, Pedersen O, Hansen T, Albrechtsen A NR: **Greenlandic Inuit show genetic signatures of diet and climate adaptation.** *Science (80-)* 2015, **349**::1343–47.

Summary

Advances in technology have turned modern biology into a data-intensive enterprise. The advent of high-output technologies like Microarrays and Next-generation sequencing technologies has resulted in researchers grappling not just with huge volumes but also multiple types of data. While generation and storage of high-quality data are an important research focus, it is increasingly recognized that translating data into actionable information and insight is a critical research challenge. To infer reliable conclusions from the data, it is often necessary to integrate large amounts of heterogeneous data with different formats and semantics. Given the breadth and volume of data involved, this goal is best achieved through automated methods and tools for data integration and workflow management. This thesis presents automated strategies that combine bioinformatics and statistical methods to identify novel biomarkers in high-throughput OMICs datasets pertaining to the metabolic syndrome and to gain mechanistic insight into the underlying biological processes. An underlying theme in this thesis is data-driven approaches that generate plausible hypothesis which is followed by experimental verification.

The main findings in each of the chapters are summarized below.

Genome-wide association studies of metabolite profiles explain a higher percentage of genetic variation and have larger effect sizes than clinical phenotypes and traits. However, given the large number of metabolites measured, these studies come with a large multiple testing penalty. In **Chapter 2** we present an automated workflow approach that utilizes prior knowledge of biochemical pathways present in databases like KEGG and BioCyc to generate smaller gene sets relevant to the metabolite. To retrieve a prioritized list of candidate genes associated with metabolite levels, gene sets were generated for each metabolite by identifying genes that participate in pathways and reactions relevant to the synthesis or degradation of the metabolite. For every gene set, a corresponding SNP set was generated by retrieving SNPs within the flanking 50 kb of every gene. Re-analysis of a published GWAS dataset using the metabolite specific SNP sets confirmed previously identified hits and identified a new locus of human metabolic individuality, associating Aldehyde dehydrogenase family1 L1 (*ALDH1L1*) with serine / glycine ratios in blood. The workflow paradigm described in chapter 2 is gaining ground in bioinformatics as the technology of choice for

recording the steps of computational experiments. In a typical workflow, data outputs are generated from data inputs via a set of (potentially distributed) computational tasks that are coordinated following a workflow definition. However, workflows do not provide a complete solution for aggregating all data and all meta-data that are necessary for understanding the full context of an experiment. Encapsulating all aspects of an *in silico* analysis and communicating it to the scientific community is a key challenge in a computational experiment. **Chapter 3** explores the utility of semantic web technologies in the preservation of computational experiments. Semantic web technologies facilitate the integration of heterogonous data on the World Wide Web by making the semantics of the data explicit through formal ontologies. More specifically, the chapter discusses the Research Object (RO) model, where a research object is defined as a resource that aggregates other resources, e.g. datasets, software, spreadsheets, text, etc. The overarching goal of the RO model is to facilitate transparency and reproducibility of scientific studies. The RO model was applied to a study where the goal was to facilitate the interpretation of the results of a GWAS of metabolite profiles. Applying a workflow-centric RO model to aggregate and annotate the resources used in the bioinformatics experiment, allowed us to retrieve the conclusions of the experiment in the context of the driving hypothesis, the executed workflows and their input data.

Obesity results in decreased life expectancy due to associated metabolic and cardiovascular disorders, as well as several types of cancer. A majority of obese individuals develop insulin resistance and type-2 diabetes. However, approximately 10-25% of these individuals seem to remain insulin sensitive and Normal Glucose Tolerant (NGT). Studies have shown that the expanded adipose tissue serves as an important pathogenic site in the development of type 2 diabetes. **Chapter 4** presents a study designed to investigate the role of the adipose tissue in development of T2DM in severely obese subjects, by performing RNA-Sequencing of the subcutaneous (SAT) and visceral adipose tissue (VAT) samples obtained during bariatric surgery. The sets of expressed genes were subjected to a gene network-based approach to distinguish obese individuals with NGT from obese individuals with type 2 diabetes. This identified acetyl-CoA metabolic network down-regulation as an important feature in the pathophysiology of obese individuals with type 2 diabetes. In general, genes within two reaction steps of acetyl-CoA were found to be down-regulated in the VAT and SAT of individuals with type 2 diabetes. Upon weight loss and amelioration of metabolic abnormalities three months following bariatric surgery, the expression level of these genes recovered to levels seen in NGT individuals. We report four novel genes associated with type-2 diabetes and recovery upon weight loss: acetyl-CoA acetyltransferase 1 (*ACAT1*), acetyl-CoA carboxylase alpha (*ACACA*), aldehyde dehydrogenase 6 family, member A1 (*ALDH6A1*) and methylenetetrahydrofolate dehydrogenase (*MTHFD1*). In addition to confirming earlier findings by other groups on the role of branched-chain amino acid degradation, fatty acid oxidation and citrate cycle in type-2 diabtes, we show through a network analysis that acetyl-CoA metabolism is the unifying principle and that its dysregulation distinguishes between obese women with type-2 diabetes and those with NGT.

Next generation RNA-sequencing technology has made it possible to quantify gene expression but also to use the sequence itself to identify expressed alleles by calling haplotypes of an individual based on heterozygosity of SNPs in expressed loci. Allele-specific expression studies help to understand the *cis*-regulatory basis of variation in gene expression. In **Chapter 5**, we investigate the hypothesis that cis-regulatory variants differentially affect gene expression in visceral and subcutaneous adipose tissue. We investigated differential allele-specific expression between visceral and subcutaneous adipose tissue of very obese individuals (BMI>40) with and without type 2 diabetes mellitus with the aim of identifying regulatory variants that could explain the pathophysiological differences observed in the two tissues. The objective of the study was to identify from a panel of known genome-wide association hits the subset of common variants that are under the control of cis-regulatory elements and to assess the consequence of such variants on the T2DM phenotype. We identified a single nucleotide polymorphism (SNP) rs1049174, in the 3’ untranslated region (3’ UTR) in *KLRK1* (Killer cell lectin like receptor subfamily K, family member 1) gene that displays a significant differential allelic expression between VAT and SAT, and for which expression is different between individuals with normal glucose tolerance (NGT) and T2D. The differential allele-specific expression of *KLRK1* between visceral and subcutaneous adipose tissue and the increased expression of *KLRK1* in visceral adipose tissue of very obese individuals with type 2 diabetes provides evidence for a role of *KLRK1* in the susceptibility to type-2 diabetes.

Very low calorie diets (VLCD) with and without exercise programs lead to major metabolic improvements in obese type 2 diabetes patients. In **Chapter 6**, we investigate the mechanisms of a VLCD with or without exercise to uncover possible biomarkers associated with these interventions. In the first step, targeted multiple reaction monitoring (MRM) analysis was conducted for 13 abundant proteins hypothesized to be associated with T2DM and obesity, including apolipoproteins and markers of inflammation and coagulation. Subsequently, a large scale isobaric tag for relative and absolute quantification (iTRAQ) approach was utilized to uncover differences between the VLCD with and without exercise groups for less abundant proteins. Using proteomic analysis several potential disease state and intervention associated markers were found distinguishing T2DM patients from obese and lean controls and showing a VLCD effect.

In **Chapter 7** we explore the prolonged effects of niacin on gene expression profile in adipocytes of a hyperlipidemic mouse model. We applied bioinformatic and statistical analyses to the gene expression data and showed that prolonged niacin treatment led to an increase in the poly unsaturated fatty acid (PUFA) synthesis pathways. To investigate whether PUFA levels and possible derivatives thereof (i.e. oxylipins) were functionally affected, we determined the fatty acid composition in the adipose tissue. These analyses revealed increased n-3 PUFA secretion from the adipocytes and an increased plasma level of n-3 PUFAs and their anti-inflammatory oxylipins. Together with the up-regulation of the PUFA biosynthesis pathway in gWAT, this point towards an atheroprotective plasma profile induced by prolonged niacin treatment.

In Chapter 8, we present a global review of the current status of metabolomic GWAS (mGWAS). The first waves of metabolomics and genetic analyses by mGWAS have provided a wealth of insight into the genetic basis of metabolic individuality and risk factors for common metabolic disorders. However, we still face many hurdles in the interpretation of mGWAS data. Metabolomics platforms generally yield information on the levels of one to several hundreds of metabolites. Consideration of all metabolites results in a severe multiple testing burden. This precludes genuine SNP-metabolite pairs from being considered when they fail to reach the stringent statistical threshold for significance. Pathway analysis is exquisitely suited to increase the statistical power to identify biologically plausible loci and simultaneously improve our understanding of the underlying biological mechanisms. In addition, the next step in pathway analysis is to include stoichiometric and kinetic parameters and complement the statistical analysis with a more comprehensive systems biology based approach using mathematical modelling. The application of *a priori* knowledge present in databases and the potential of mathematical models in enhancing the interpretation of mGWAS are presented.

The thesis concludes with **Chapter 9** where future developments in the discipline are outlined.

Samenvatting

Technologische vooruitgang heeft de biologie tot een data-intensieve wetenschap gemaakt. Met de introductie van high-throughput technologieën zoals microarrays en next generation sequencing worstelen biomedisch onderzoekers in toenemende mate met grote data volumina en verschillende data formats. Naast de generatie en de opslag van data, is de grootste uitdaging van de moderne biologie om deze grote hoeveelheden data om te zetten in nieuwe biologische inzichten en informatie waar je in de praktijk wat mee kunt. Dit vereist nieuwe methoden om heterogene data met elkaar te combineren en te integreren. Daarnaast is het noodzakelijk om de data en de analyse van de data op een zodanige manier in te richten dat de resultaten reproduceerbaar zijn. Dit proefschrift laat zien hoe geautomatiseerde strategieën op het gebied van bioinformatica en biostatistiek gebruikt kunnen worden om nieuwe biomarkers voor metabool syndroom uit high-throughput –omics data te destilleren en om mechanistische inzichten te verwerven in de biologische processen die ten grondslag liggen aan dit syndroom. Een gerelateerd thema in dit proefschrift is dat een data gedreven aanpak kan leiden tot plausibele hypotheses die vervolgens experimenteel kunnen worden geverifieerd.

De belangrijkste bevindingen uit dit proefschrift worden hieronder beschreven.

Genoom-wijde associatie studies (GWAS) van metaboliet profielen laten in het algemeen grotere effecten zien dan vergelijkbare studies met klinische fenotypes. Genetische variatie verklaart ook een groter deel van de interindividuele variatie van metaboliet niveaus dan van conventionele klinische parameters. Echter, door het grote aantal metabolieten wat doorgaans wordt gemeten krijgen deze studies te maken met een relatief zware correctie voor het aantal statistische associatie testen dat wordt uitgevoerd (multiple testing correctie). In **hoofdstuk 2** presenteren we een automatische workflow die beschikbare informatie over metabole routes (zoals gedocumenteerd in de KEGG en BioCyc databases) gebruikt om alleen de relevante gen-metaboliet combinaties te identificeren en te testen. Relevante combinaties zijn gebaseerd op genen waarvan de eiwitproducten direct of indirect betrokken zijn bij omzetting of productie van de betreffende metaboliet. Her-analyse van reeds gepubliceerde GWAS met specifieke gen-metaboliet combinaties bevestigde eerder gevonden genetische associaties en leidde tot de ontdekking van een nieuw genetisch locus in het *ALDH1L* gen die de glycine / serine ratio's in het bloed beïnvloedt.

De analyse routines die gebruikt zijn in hoofdstuk 2 zijn opgeslagen in zogenaamde workflows. Via een workflow worden resultaten gegenereerd door het automatisch doorlopen van een aantal gedefinieerde, en mogelijk gedistribueerde, stappen. Workflows vormen echter niet een complete oplossing voor het aggregeren van alle data en meta-data die nodig is om een experiment te omschrijven. Daarom onderzoekt **hoofdstuk 3** de bruikbaarheid van semantisch web technologieën om computationele experimenten voor langere tijd te bewaren. Semantisch web technologieën maken het mogelijk om heterogene data op het internet te integreren door de data te beschrijven met formele ontologieën. In dit hoofdstuk wordt het "Research Object" model toegelicht. Een Research Object is gedefinieerd als een entiteit dat andere entiteiten zoals datasets, software, spreadsheets en tekst, omvat. Het gebruik van het Research Object model bevordert de transparantie en reproduceerbaarheid van wetenschappelijke studies. Het Research Object model is toegepast in een studie waarin de resultaten van GWAS voor metaboliet profielen zijn geïnterpreteerd. Het Research Object archiveert de conclusies van een bioinformatisch experiment in de context van de werkhypothese, de data analyse routines en de uitgangsdata.

Zwaarlijvigheid (obesitas) resulteert in een verminderde levensverwachting als gevolg van het optreden van metabole en cardiovasculaire problemen en een verhoogde incidentie van kanker. Een meerderheid van de individuen met obesitas ontwikkelt insuline resistentie en type II diabetes. Echter, 10-25% van deze individuen blijft gevoelig voor insuline en tolerant voor glucose. Meerdere wetenschappelijke studies hebben aangetoond dat vetweefsel een belangrijke rol speelt in de ontwikkeling van type II diabetes. De rol van het vetweefsel in individuen met een extreme vorm van obesitas wordt nader bestudeerd in **Hoofdstuk 4**. Van deze patiënten zijn tijdens een maagoperatie monsters verzameld van subcutaan (SC) en visceraal (VC) vetweefsel en hierin is genexpressie bepaald door toepassing van RNA-sequencing. Gen netwerk analyse is gebruikt om individuen met normale glucose tolerantie te onderscheiden van individuen met type 2 diabetes. Hierbij werd gevonden dat het acetyl-CoA metabole netwerk minder actief is in individuen met type 2 diabetes. Na gewichtsverlies en verbetering in de metabole status in de drie maanden na de maagverkleining operatie was ook de activiteit van het acetyl-CoA netwerk weer grotendeels genormaliseerd. We rapporteren vier nieuwe genen in het vetweeefsel die geassocieerd zijn met type 2 diabetes en herstel na gewichtsverlies: acetyl-CoA acetyltransferase 1 (*ACAT1*), acetyl-CoA carboxylase alpha (*ACACA*), aldehyde dehydrogenase 6 family, member A1 (*ALDH6A1*) en methylenetetrahydrofolate dehydrogenase (*MTHFD1*). Hiermee tonen we aan dat een verstoord acetyl-CoA metabolisme een belangrijk aspect van obesitas met type 2 diabetes vormt, naast eerder gevonden verschillen in vertakte keten aminozuur metabolisme, vet oxidatie en een verstoorde citroenzuurcyclus.

RNA-sequencing technologie maakt het niet alleen mogelijk om gen expressie te kwantificeren maar ook om onderscheid te maken tussen de expressie van de twee chromosomen van een chromosoom paar. Dit gebeurt op basis van heterozygote genetische varianten (“single nucleotide polymorphisms” of SNPs) die worden afgelezen en geteld tijdens de sequencing van RNA. Door toepassing van deze zogenaamde allel-specifieke expressie analyse kan meer inzicht worden verkregen in de regulatie van gen expressie. In **hoofdstuk 5** onderzochten we de hypothese dat genetische variatie rond een gen op een van de twee chromosomen tot gevolg heeft dat expressie van dat gen anders is tussen SC en VC vetweefsel. Dit zou dan ook mede kunnen verklaren waarom het VC een belangrijker rol lijkt te spelen in de ontwikkeling van het metabool syndroom dan het SC vet. We gebruikten hierbij de data gegenereerd voor hoofdstuk 4. We identificeerden een SNP in de 3’ ongetransleerde regio van het *KLRK1* (Killer cell lectin like receptor subfamily K, family member 1) gen. Dit gen vertoont verschillen in allel-specifike expressie tussen VC en SC vet en ook verschillen in expressie tussen insuline sensitieve en insuline gevoelige individuen. We suggereren daarom dat *KLRK1* een rol speelt bij de ontwikkeling van type 2 diabetes en dat genetische variatie in dit gen de gevoeligheid voor het ontwikkelen van type 2 diabetes deels kan verklaren.

Diëten met een heel laag aantal calorieën, gecombineerd met bewegingstraining, kunnen belangrijke metabole verbeteringen in type 2 diabetes patiënten met obesitas bewerkstelligen. In **hoofdstuk 6** onderzochten we de mechanismen die deze verbeteringen veroorzaken en biomarkers (moleculaire detectoren) die deze verbeteringen kunnen laten zien. In de eerste stap is een massa spectrometrie-gebaseerde methode (multiple reaction monitoring, MRM) toegepast om 13 eiwitten te kunnen meten die in verband gebracht zijn met type 2 diabetes en obesitas, waaronder apolipoproteïnen en ontstekings- en bloedstollingseiwitten. Vervolgens is de “isobaric tag for relative and absolute quantification” (iTRAQ) methode toegepast om ook eiwitten die in lagere concentraties aanwezig zijn, te kunnen meten. Met behulp van deze proteomics technologieën vonden we verschillende eiwitten die type 2 diabetes patiënten konden onderscheiden van zwaarlijvige individuen en individuen met normaal gewicht en eiwitten die reageerden op het volgen van een calorie-arm dieet.

In **hoofdstuk 7** onderzochten we het effect van langdurige toediening van niacine op het gen expressie profiel van adipocyten in een hyperlipidemisch muismodel. We pasten bioinformatische en statistische analyses toe op de gen expressie data en toonden aan dat langdurige niacine toediening de synthese van meervoudig onverzadigde vetzuren verhoogt. Daarnaast vonden we aanwijzingen dat de adipocyten een hogere concentratie van omega-3 vetzuren en daarvan afgeleide, anti-inflammatoire oxylipines uitscheiden. De combinatie van deze bevindingen suggereert dat niacine kan beschermen tegen atherosclerose.

In **hoofdstuk 8** presenteren we een globaal overzicht van de stand van onderzoek rondom metaboliet GWAS (mGWAS). De eerste mGWAS hebben veel nieuwe inzichten verschaft in de genetische basis van interindividuele verschillen in metaboliet profielen. Toch moeten nog vele hordes worden genomen in dit soort analyses. Analyse op het niveau van complete metabole routes in plaats van op het niveau van individuele metabolieten kan de interpretatie helpen en zorgen voor een minder strenge multiple testing correctie. Een volgende stap in dit soort analyses is om ook stoichiometrische en kinetische parameters mee te nemen en meer kwantitatieve modellen te ontwikkelen.

Het proefschrift wordt afgesloten met **hoofdstuk 9**, waarin de toekomstige ontwikkelingen van het vakgebied worden geschetst.

Curriculum Vitae

Harish Dharuri was born on August 17th 1969 in the southern part of India. He completed his undergraduate degree in Chemical Engineering from Bangalore University, India in 1994. Upon graduation, he developed an interest in biotechnology and was admitted to a Master's degree in the same subject at the Jawaharlal Technological University (JNTU) in Hyderabad, India. As part this program, he worked on biochemical engineering research project that aimed to produce 2,3 Butanediol using *Klebsiella Oxytoca* in a membrane recycled bioreactor. After graduating with distinction in 1997, he worked in the chemical industry. Also, at this time, he taught courses in biochemical engineering fundamentals, bioreactor design, and downstream processing to graduate students at JNTU. In 2000, he was admitted to a professional master's degree at the Keck Graduate Institute (KGI) for Applied Life Sciences (part of consortium of Claremont colleges) near Los Angeles in the United States. The degree program at KGI combined business management with the biosciences. The goal of the two year Master of Biosciences degree was to produce leaders who could catalyze development of basic life sciences research into useful new products, processes and services. Upon graduating in 2002, along with four other students from the graduating class, he started a bioinformatics based company called Zuyder Corporation.

In 2006, he went on to work at the California Institute of Technology (Caltech) at Pasadena in the department of Control and Dynamical Systems. Here, he worked on the Biomodels database, a repository of dynamic models of biochemical phenomena, and Systems Biology MarkUp Language (SBML), an xml based standard for the exchange of computational models of biological processes. This was a collaborative effort involving Caltech and the European Bioinformatics Institute (EBI) at Hinxton, UK. At the same time, he also worked on the Virtual Cell modeling and analysis software at the University of Connecticut Health Center (UCHC), Connecticut, USA. While working at Caltech, he pursued a Master's degree in Biostatistics at the California State University, East Bay, in a program that was geared towards working adults. He completed his course requirements and graduated in 2010.

In 2011, he moved to the Netherlands to join the Leiden University Medical Center as a Scientific Researcher in the Center for Human and Clinical

Genetics under prof. Ko Willems van Dijk and dr. Peter-Bram ‘t Hoen. He worked on a bioinformatic approach that utilized prior knowledge to identify novel SNP-metabolite associations in GWAS of metabolite profiles. The pathway-based approach was also applied to other high-output technologies such as RNA-Seq to gain biological insight into the disease focus of his research, metabolic syndrome. His work was presented at the American Society and International Conference of Human Genetics (ASHG/ICHG) and other conferences. He was formally admitted to the PhD program in the year 2013. During the course of his work at LUMC, he gained expertize in the area of pathway bioinformatics and he taught courses on the topic at workshops conducted by the department of epidemiology, Erasmus MC. The result of his work at LUMC are presented in this thesis.

He is currently employed at Illumina Inc, at their Santa Clara office in the San Francisco bay area where he is working as a biomedical information scientist in the bioinformatics and data science group.

Publications

- Draisma H, Pool R, Kobl M,…**Dharuri H** *et al*. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. **Nature Communications** 2015 Jun 12;6:7208. doi: 10.1038/ncomms8208.
- Demirkan A, Henneman P, Verhoeven A, **Dharuri H** *et al*. Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. **PLoS Genet**. 2015 Jan 8;11(1):e1004835. doi: 10.1371/journal.pgen.1004835. eCollection 2015 Jan.
- Sleddering MA, Markvoort AJ, **Dharuri HK** *et al*. Proteomic analysis in type 2 diabetes patients before and after a very low calorie diet reveals potential disease state and intervention specific biomarkers. **PLoS One**. 2014 Nov 21;9(11):e112835. doi: 10.1371/journal.pone.0112835. eCollection 2014.
- Heemskerk MM, **Dharuri HK** *et al*. Prolonged niacin treatment leads to increased adipose tissue PUFA synthesis and anti-inflammatory lipid and oxylipin plasma profile. **J Lipid Res**. 2014 Dec;55(12):2532-40. doi: 10.1194/jlr.M051938. Epub 2014 Oct 15.
- Lips MA, Van Klinken JB, van Harmelen V, **Dharuri HK** *et al*. Roux-en-Y gastric bypass surgery, but not calorie restriction, reduces plasma branched-chain amino acids in obese women independent of weight loss or the presence of type 2 diabetes. **Diabetes Care**. 2014 Dec; 37(12):3150-6. doi: 10.2337/dc14-0195. Epub 2014 Oct 14.
- Hettne KM, **Dharuri H** *et al*. Structuring research methods and data with the research object model: genomics workflows as a case study. **J Biomed Semantics**. 2014 Sep 18;5(1):41. doi: 10.1186/2041-1480-5-41. eCollection 2014.
- **Dharuri H** *et al*. Downregulation of the acetyl-CoA metabolic network in adipose tissue of obese diabetic individuals and recovery after weight loss. **Diabetologia**. 2014 Nov;57(11):2384-92. doi: 10.1007/s00125-014-3347-0.
- **Dharuri H** *et al*. Genetics of the human metabolome, what is next? **Biochim Biophys Acta**. 2014 ct;1842(10):1923-1931. doi: 10.1016/j.bbadis.2014.05.030.

- **Dharuri H** *et al*. Automated workflow-based exploitation of pathway databases provides new insights into genetic associations of metabolite profiles. **BMC Genomics**. 2013 Dec 9;14:865. doi: 10.1186/1471-2164-14-865.

- Li C, Donizelli M, Rodriguez N, **Dharuri H** *et al*. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. **BMC Syst Biol**. 2010 Jun 29;4:92. doi: 10.1186/1752-0509-4-92.

- Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, **Dharuri H** *et al*. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. **Nucleic Acids Res**. 2006 Jan 1; 34(Database issue):D689-91.

Acknowledgements

It is never easy pursuing a PhD program at any age, but to pursue it at a later stage in life with a family that is thousands of miles away and in a country that you have never been to before, let us just say, it can be challenging! I would not have been on this home stretch without the support and understanding of my supervisors, Ko Willems van Dijk and Peter-Bram. I would like to thank them for giving me an opportunity to work on interesting projects, providing guidance and importantly in facilitating a work-life balance without which none of this would have been possible. Thank you very much for your hard work, kindness and encouragement.

I would like to thank Vanessa for the countless hours of discussion that helped me understand the biology behind some of the projects we worked on together. Jan, it was really nice knowing you. Your help and encouragement meant a lot to me. Peter Henneman, right at the beginning of my stay in the Netherlands, you made me feel at home and introduced me to the world of GWAS. Thank you for being a good colleague and friend. Sjoerd, we knew each other briefly, but it was enough for me to see what a wonderful human being you are. Ayse, you helped me at very critical moments, I am very thankful to you for that. Mattjis, I would like to thank you for all the help, both personal as well as professional. Amanda, you always bought cheer and happiness to the group. I will remember your cheerful demeanor for a very long time. I would like to thank the other lipidos Lianne, Fatiha, Sam, Saeed and Lisa for wonderful times together.

I have seen very few with the kind of enthusiasm that Marco has for his work. It was a privilege knowing you, Marco. Throughout this journey I interacted closely with the biosemantics group, in particular, I would like to thank Kristina for the co-operation and the work we could accomplish together. My friendship with Kostas was one of the high points of my stay in Leiden.

The discussions with fellow bioinformaticians were very helpful in shaping my own thoughts and it helped that these were extremely friendly people. Thanks very much to Yahya, Jeroen, Irina, Eleni and many others.

I would also like to thank Hanno Pijl for his words of encouragement and Mirjam for all the help.

My short stay in the Netherlands has changed me forever. I am happy I followed this opportunity. Thank you, Netherlands!

Finally, all this would not have been possible without the support of my family. I would like to thank my parents for helping me to be the person I am today. I would like to thank my wife, Mahalakshmi for holding the fort. To my children, Pranav and Hasini, I missed you very much through those days of separation. Life takes us to unexpected places but love brings us home.